Submitted to the Annals of Statistics
arXiv: arXiv:0000.0000

ROP: MATRIX RECOVERY VIA RANK-ONE PROJECTIONS*

By T. Tony Cai and Anru Zhang

University of Pennsylvania

Estimation of low-rank matrices is of significant interest in a range of contemporary applications. In this paper, we introduce a rank-one projection model for low-rank matrix recovery and propose a constrained nuclear norm minimization method for stable recovery of low-rank matrices in the noisy case. The procedure is adaptive to the rank and robust against small perturbations. Both upper and lower bounds for the estimation accuracy under the Frobenius norm loss are obtained. The proposed estimator is shown to be rate-optimal under certain conditions. The estimator is easy to implement via convex programming and performs well numerically.

The techniques and main results developed in the paper also have implications to other related statistical problems. An application to estimation of spiked covariance matrices from one-dimensional random projections is considered. The results demonstrate that it is still possible to accurately estimate the covariance matrix of a high-dimensional distribution based only on one-dimensional projections.

1. Introduction. Accurate recovery of low-rank matrices has a wide range of applications, including quantum state tomography [1, 23], face recognition [3, 11], recommender systems [26], and linear system identification and control [35]. For example, a key step in reconstructing the quantum states in low-rank quantum tomography is the estimation of a low-rank matrix based on Pauli measurements [23, 41]. And phase retrieval, a problem which arises in a range of signal and image processing applications including X-ray crystallography, astronomical imaging, and diffraction imaging, can be reformulated as a low-rank matrix recovery problem [14, 11]. See Recht, et al. [35] and Candès and Plan [12] for further references and discussions.

Motivated by these applications, low-rank matrix estimation based on a small number of measurements has drawn much recent attention in several fields, including statistics, electrical engineering, applied mathematics, and computer science. For example, Candès and Recht [13], Candès and Tao [15]

*The research was supported in part by NSF FRG Grant DMS-0854973, NSF Grant DMS-1208982 and NIH Grant R01 CA127334-05.

Keywords and phrases: Constrained nuclear norm minimization, low-rank matrix recovery, optimal rate of convergence, rank-one projection, restricted uniform boundedness, spiked covariance matrix

and Recht [34] considered the exact recovery of a low-rank matrix based on a subset of uniformly sampled entries. Negahban and Wainwright [29] investigated matrix completion under a row/column weighted random sampling scheme. Recht, et al. [35], Candès and Plan [12], and Cai and Zhang [8, 9, 10] studied matrix recovery based on a small number of linear measurements in the framework of restricted isometry property (RIP), and Koltchinskii, et al. [25] proposed the penalized nuclear norm minimization method and derived a general sharp oracle inequality under the condition of restrict isometry in expectation.

The basic model for low-rank matrix recovery can be written as

$$y = \mathcal{X}(A) + z, \tag{1.1}$$

where $\mathcal{X} : \mathbb{R}^{p_1 \times p_2} \to \mathbb{R}^n$ is a linear map, $A \in \mathbb{R}^{p_1 \times p_2}$ is an unknown low-rank matrix, and z is a noise vector. The goal is to recover the low-rank matrix A based on the measurements $(\mathcal{X}, y)$. The linear map $\mathcal{X}$ can be equivalently specified by n $p_1 \times p_2$ measurement matrices $X_1, \cdots, X_n$ with

$$\mathcal{X}(A) = (\langle X_1, A\rangle, \langle X_2, A\rangle, \cdots, \langle X_n, A\rangle)^\intercal, \tag{1.2}$$

where the inner product of two matrices of the same dimensions is defined as $\langle X, Y\rangle = \sum_{i,j} X_{ij} Y_{ij}$. Since $\langle X, Y\rangle = \mathrm{Trace}(X^\intercal Y)$, (1.1) is also known as trace regression.

A common approach to low-rank matrix recovery is the constrained nuclear norm minimization method which estimates A by

$$\hat{A} = \arg\min_M \{\|M\|_* : y - \mathcal{X}(M) \in \mathcal{Z}\}. \tag{1.3}$$

Here $\|X\|_*$ is the nuclear norm of the matrix X which is defined to be the sum of its singular values, and $\mathcal{Z}$ is a bounded set determined by the noise structure. For example, $\mathcal{Z} = \{0\}$ in the noiseless case and $\mathcal{Z}$ is the feasible set of the error vector z in the case of bounded noise. This constrained nuclear norm minimization method has been well studied. See, for example, [35, 12, 30, 8, 9, 10].

Two random design models for low-rank matrix recovery have been particularly well studied in the literature. One is the so-called "Gaussian ensemble" [35, 12], where the measurement matrices $X_1, \cdots, X_n$ are random matrices with i.i.d. Gaussian entries. By exploiting the low-dimensional structure, the number of linear measurements can be far smaller than the number of entries in the matrix to ensure stable recovery. It has been shown that a matrix A of rank r can be stably recovered by nuclear norm minimization with high

probability, provided that $n \gtrsim r(p_1+p_2)$ [12]. One major disadvantage of the Gaussian ensemble design is that it requires $O(np_1p_2)$ bytes of storage space for $\mathcal{X}$, which can be excessively large for the recovery of large matrices. For example, at least 45 TB of space is need to store the measurement matrices M_i in order to ensure accurate reconstruction of 10000×10000 matrices of rank 10. (See more discussion in Section 5.) Another popular design is the "matrix completion" model [13, 15, 34], under which the individual entries of the matrix A are observed at randomly selected positions. In terms of the measurement matrices X_i in (1.2), this can be interpreted as

$$\mathcal{X}(A) = (\langle e_{i_1} e_{j_1}^\intercal, A\rangle, \langle e_{i_2} e_{j_2}^\intercal, A\rangle, \cdots, \langle e_{i_n} e_{j_n}^\intercal, A\rangle)^\intercal \tag{1.4}$$

where $e_i = (0, \cdots, 0, \overbrace{1}^{i\text{th}}, 0, \cdots, 0)$ is the ith standard basis vector, and $i_1, \cdots, i_n$ and $j_1, \cdots, j_n$ are randomly and uniformly drawn with replacement from $\{1, \cdots, p_1\}$ and $\{1, \cdots, p_2\}$, respectively. However, as pointed out in [13, 34], additional structural assumptions, which are not intuitive and difficult to check, on the unknown matrix A are needed in order to ensure stable recovery under the matrix completion model. For example, it is impossible to recover spiked matrices under the matrix completion model. This can be easily seen from a simple example where the matrix A has only one non-zero row. In this case, although the matrix is only of rank one, it is not recoverable under the matrix completion model unless all the elements on the non-zero row are observed.

In this paper we introduce a *"Rank-One Projection" (ROP)* model for low-rank matrix recovery and propose a constrained nuclear norm minimization method for this model. Under the ROP model, we observe

$$y_i = (\beta^{(i)})^\intercal A \gamma^{(i)} + z_i, \quad i = 1, ..., n \tag{1.5}$$

where $\beta^{(i)}$ and $\gamma^{(i)}$ are random vectors with entries independently drawn from some distribution $\mathcal{P}$, and z_i are random errors. In terms of the linear map $\mathcal{X} : \mathbb{R}^{p_1 \times p_2} \to \mathbb{R}^n$ in (1.1), it can be defined as

$$[\mathcal{X}(A)]_i = (\beta^{(i)})^\intercal A \gamma^{(i)}, \quad i = 1, \cdots, n. \tag{1.6}$$

Since the measurement matrices $X_i = \beta^{(i)}(\gamma^{(i)})^\intercal$ are of rank-one, we call the model (1.5) a *"Rank-One Projection" (ROP)* model. It is easy to see that the storage for the measurement vectors in the ROP model (1.5) is $O(n(p_1+p_2))$ bytes which is significantly smaller than $O(np_1p_2)$ bytes required for the Gaussian ensemble.

We first establish a sufficient identifiability condition in Section 2 by considering the problem of exact recovery of low-rank matrices in the noiseless case. It is shown that, with high probability, ROP with $n \gtrsim r(p_1 + p_2)$ random projections is sufficient to ensure exact recovery of all rank r matrices through the constrained nuclear norm minimization. The required number of measurements $O(r(p_1 + p_2))$ is rate optimal for any linear measurement model since a rank r matrix $A \in \mathbb{R}^{p_1+p_2}$ has the degree of freedom $r(p_1 + p_2 - r)$. The Gaussian noise case is of particular interest in statistics. We propose a new constrained nuclear norm minimization estimator and investigate its theoretical and numerical properties in the Gaussian noise case. Both upper and lower bounds for the estimation accuracy under the Frobenius norm loss are obtained. The estimator is shown to be rate-optimal when the number of rank-one projections satisfies either $n \gtrsim (p_1 + p_2)\log(p_1 + p_2)$ or $n \sim r(p_1 + p_2)$. The lower bound also shows that if the number of measurements $n < r\max(p_1, p_2)$, then no estimator can recover rank-r matrices consistently. The general case where the matrix A is only approximately low-rank is also considered. The results show that the proposed estimator is adaptive to the rank r and robust against small perturbations. Extensions to the sub-Gaussian design and sub-Gaussian noise distribution are also considered.

The ROP model can be further simplified by taking $\beta^{(i)} = \gamma^{(i)}$ if the low-rank matrix A is known to be symmetric. This is the case in many applications, including low-dimensional Euclidean embedding [37, 35], phase retrieval [14, 11], and covariance matrix estimation [16, 5, 6]. In such a setting, the ROP design can be simplified to symmetric rank-one projections (SROP)

$$[\mathcal{X}(A)]_i = (\beta^{(i)})^\intercal A \beta^{(i)}.$$

We will show that the results for the general ROP model continue to hold for the SROP model when A is known to be symmetric. Recovery of symmetric positive definite matrices in the noiseless and ℓ_1-bounded noise settings has also been considered in a recent paper by Chen et al [16] which was posted on arXiv at the time of the writing of the present paper. Their results and techniques for symmetric positive definite matrices are not applicable to the recovery of general low-rank matrices. See Section 6 for more discussions.

The techniques and main results developed in the paper also have implications to other related statistical problems. In particular, the results imply that it is possible to accurately estimate a spiked covariance matrix based only on one-dimensional projections. Spiked covariance matrix model has been well studied in the context of principal component analysis (PCA) based on i.i.d. data where one observes p-dimensional vectors

$X^{(1)}, \cdots, X^{(n)} \overset{iid}{\sim} N(0, \Sigma)$ with $\Sigma = I_p + \Sigma_0$ and Σ_0 being low-rank [24, 4, 5, 6]. This covariance structure and its variations have been used in many applications including signal processing, financial econometrics, chemometrics, and population genetics. See, for example, [20, 28, 32, 33, 42]. Suppose that the random vectors $X^{(1)}, \cdots, X^{(n)}$ are not directly observable. Instead, we observe only one-dimensional random projections of $X^{(i)}$,

$$\xi_i = \langle \beta^{(i)}, X^{(i)} \rangle, \quad i = 1, ..., n,$$

where $\beta^{(i)} \overset{iid}{\sim} N(0, I_p)$. It is somewhat surprising that it is still possible to accurately estimate the spiked covariance matrix Σ based only on the one-dimensional projections $\{\xi_i : \ i = 1, ..., n\}$. This covariance matrix recovery problem is also related to the recent literature on covariance sketching [17, 18], which aims to recover a symmetric matrix A (or a general rectangular matrix B) from low-dimensional projections of the form $X^\intercal AX$ (or $X^\intercal BY$). See Section 4 for further discussions.

The proposed methods can be efficiently implemented via convex programming. A simulation study is carried out to investigate the numerical performance of the proposed nuclear norm minimization estimators. The numerical results indicate that ROP with $n \geq 5r \max(p_1, p_2)$ random projections is sufficient to ensure the exact recovery of rank r matrices through constrained nuclear norm minimization and show that the procedure is robust against small perturbations, which confirm the theoretical results developed in the paper. The proposed estimator outperforms two other alternative procedures numerically in the noisy case. In addition, the proposed method is illustrated through an image compression example.

The rest of the paper is organized as follows. In Section 2, after introducing basic notation and definitions, we consider exact recovery of low-rank matrices in the noiseless case and establish a sufficient identifiability condition. A constrained nuclear norm minimization estimator is introduced for the Gaussian noise case. Both upper and lower bounds are obtained for estimation under the Frobenius norm loss. Section 3 considers extensions to sub-Gaussian design and sub-Gaussian noise distributions. An application to estimation of spiked covariance matrices based on one-dimensional projections is discussed in detail in Section 4. Section 5 investigates the numerical performance of the proposed procedure through a simulation study and an image compression example. A brief discussion is given in Section 6. The main results are proved in Section 7 and the proofs of some technical lemmas are given in the Appendix.

2. Matrix Recovery under Gaussian Noise. In this section, we first establish an identifiability condition for the ROP model by considering exact recovery in the noiseless case, and then focus on low-rank matrix recovery in the Gaussian noise case.

We begin with the basic notation and definitions. For a vector $\beta \in \mathbb{R}^n$, we use $\|\beta\|_q = \sqrt[q]{\sum_{i=1}^n |\beta_i|^q}$ to define its vector q-norm. For a matrix $X \in \mathbb{R}^{p_1 \times p_2}$, the Frobenius norm is $\|X\|_F = \sqrt{\sum_{i=1}^{p_1} \sum_{j=1}^{p_2} X_{ij}^2}$ and the spectral norm $\|\cdot\|$ is $\|X\| = \sup_{\|\beta\|_2 \le 1} \|X\beta\|_2$. For a linear map $\mathcal{X} = (X_1, ..., X_n)$ from $\mathbb{R}^{p_1 \times p_2}$ to $\mathbb{R}^n$ given by (1.2), its dual operator $\mathcal{X}^* : \mathbb{R}^n \to \mathbb{R}^{p_1 \times p_2}$ is defined as $\mathcal{X}^*(z) = \sum_{i=1}^n z_i X_i$. For a matrix $X \in \mathbb{R}^{p_1 \times p_2}$, let $X = \sum_i a_i u_i v_i^\intercal$ be the singular value decomposition of X with the singular values $a_1 \ge a_2 \ge \cdots \ge 0$. We define $X_{\max(r)} = \sum_{i=1}^r a_i u_i v_i^\intercal$ and $X_{-\max(r)} = X - X_{\max(r)} = \sum_{i \ge r+1} a_i u_i v_i^\intercal$. For any two sequences $\{a_n\}$ and $\{b_n\}$ of positive numbers, denote by $a_n \gtrsim b_n$ when $a_n \ge C b_n$ for some uniform constant C and denote by $a_n \sim b_n$ if $a_n \gtrsim b_n$ and $b_n \gtrsim a_n$.

We use the phrase "rank-r matrices" to refer to matrices of rank at most r and denote by $\mathbb{S}^p$ the set of all $p \times p$ symmetric matrices. A linear map $\mathcal{X} : \mathbb{R}^{p_1 \times p_2} \to \mathbb{R}^n$ is called ROP from distribution $\mathcal{P}$ if $\mathcal{X}$ is defined as in (1.6) with all the entries of $\beta^{(i)}$ and $\gamma^{(i)}$ independently drawn from the distribution $\mathcal{P}$.

2.1. *RUB, Identifiability, and Exact Recovery in the Noiseless Case.* An important step towards understanding the constrained nuclear norm minimization is the study of exact recovery of low-rank matrices in the noiseless case which also leads to a sufficient identifiability condition. A widely used framework in the low-rank matrix recovery literature is the Restricted Isometry Property (RIP) in the matrix setting. See [35, 12, 36, 8, 9, 10]. However, the RIP framework is not well suited for the ROP model and would lead to sub-optimal results. See Section 2.2 for more discussions on the RIP and other conditions used in the literature. See also [14]. In this section, we introduce a Restricted Uniform Boundedness (RUB) condition which will be shown to guarantee the exact recovery of low-rank matrices in the noiseless case and stable recovery in the noisy case through the constrained nuclear norm minimization. It will also be shown that the RUB condition are satisfied by a range of random linear maps with high probability.

Definition 2.1 (Restricted Uniform Boundedness). *For a linear map $\mathcal{X} : \mathbb{R}^{p_1 \times p_2} \to \mathbb{R}^n$, if there exist uniform constants C_1 and C_2 such that for*

all nonzero rank-r matrices $A \in \mathbb{R}^{p_1 \times p_2}$

$$C_1 \leq \frac{\|\mathcal{X}(A)\|_1/n}{\|A\|_F} \leq C_2,$$

where $\|\cdot\|_1$ *means the vector* ℓ_1 *norm, then we say that* $\mathcal{X}$ *satisfies the restricted uniform boundedness (RUB) condition of order* r *and constants* C_1 *and* C_2.

In the noiseless case, we observe $y = \mathcal{X}(A)$ and estimate the matrix A through the constrained nuclear norm minimization

$$A_* = \arg\min_M \{\|M\|_* : \mathcal{X}(M) = y\}. \tag{2.1}$$

The following theorem shows that the RUB condition guarantees the exact recovery of all rank-r matrices.

THEOREM 2.1. *Let* $k \geq 2$ *be an integer. Suppose* $\mathcal{X}$ *satisfies RUB of order* kr *with* $C_2/C_1 < \sqrt{k}$*, then the nuclear norm minimization method recovers all rank-r matrices. That is, for all rank-r matrices* A *and* $y = \mathcal{X}(A)$*, we have* $A_* = A$*, where* A_* *is given by* (2.1).

Theorem 2.1 shows that RUB of order kr with $C_2/C_1 < \sqrt{k}$ is a sufficient identifiability condition for the low-rank matrix recovery model (1.1) in the noisy case. The following result shows that the RUB condition is satisfied with high probability under the ROP model with a sufficient number of measurements.

THEOREM 2.2. *Suppose* $\mathcal{X} : \mathbb{R}^{p_1 \times p_2} \to \mathbb{R}^n$ *is ROP from the standard normal distribution. For integer* $k \geq 2$*, positive numbers* $C_1 < \frac{1}{3}$ *and* $C_2 > 1$*, there exist constants* C *and* δ*, not depending on* p_1, p_2*, and* r*, such that if*

$$n \geq Cr(p_1 + p_2), \tag{2.2}$$

then with probability at least $1 - e^{-n\delta}$*,* $\mathcal{X}$ *satisfies RUB of order* kr *and constants* C_1 *and* C_2.

REMARK 2.1. The condition $n \geq O(r(p_1 + p_2))$ on the number of measurements is indeed necessary for $\mathcal{X}$ to satisfy non-trivial RUB with $C_1 > 0$. Note that the degree of freedom of all rank-r matrices of $\mathbb{R}^{p_1 \times p_2}$ is $r(p_1 + p_2 - r) \geq \frac{1}{2}r(p_1 + p_2)$. If $n < \frac{1}{2}r(p_1 + p_2)$, there must exist a non-zero rank-r matrix $A \in \mathbb{R}^{p_1 \times p_2}$ such that $\mathcal{X}(A) = 0$, which leads to the failure of any non-trivial RUB for $\mathcal{X}$.

As a direct consequence of Theorems 2.1 and 2.2, ROP with the number of measurements $n \geq Cr(p_1+p_2)$ guarantees the exact recovery of all rank-r matrices with high probability.

COROLLARY 2.1. *Suppose $\mathcal{X} : \mathbb{R}^{p_1 \times p_2} \to \mathbb{R}^n$ is ROP from the standard normal distribution. There exist uniform constants C and δ such that, whenever $n \geq Cr(p_1 + p_2)$, the nuclear norm minimization estimator A_* given in (2.1) recovers all rank-r matrices $A \in \mathbb{R}^{p_1 \times p_2}$ exactly with probability at least $1 - e^{-n\delta}$.*

Note that the required number of measurements $O(r(p_1 + p_2))$ above is rate optimal, since the degree of freedom for a matrix $A \in \mathbb{R}^{p_1+p_2}$ of rank r is $r(p_1 + p_2 - r)$, and thus at least $r(p_1 + p_2 - r)$ measurements are needed in order to recover A exactly using any method.

2.2. *RUB, RIP and Other Conditions.* We have shown that RUB implies exact recovery in the noiseless and proved that the random rank-one projections satisfy RUB with high probability whenever the number of measurements $n \geq Cr(p_1+p_2)$. As mentioned earlier, other conditions, including the restricted isometry property (RIP), RIP in expectation, and spherical section property (SSP), have been introduced for low-rank matrix recovery based on linear measurements. Among them, RIP is perhaps the most widely used. A linear map $\mathcal{X} : \mathbb{R}^{p_1 \times p_2} \to \mathbb{R}^n$ is said to satisfy RIP of order r with positive constants C_1 and C_2 if

$$C_1 \leq \frac{\|\mathcal{X}(A)\|_2/\sqrt{n}}{\|A\|_F} \leq C_2$$

for all rank-r matrices A. Many results have been given for low-rank matrices under the RIP framework. For example, Recht et al [35] showed that Gaussian ensembles satisfy RIP with high probability under certain conditions on the dimensions. Candès and Plan [12] provided a lower bound and oracle inequality under the RIP condition. Cai and Zhang [8, 9, 10] established the sharp bounds for the RIP conditions that guarantee accurate recovery of low-rank matrices.

However, the RIP framework is not suitable for the ROP model considered in the present paper. The following lemma is proved in the Appendix.

LEMMA 2.1. *Suppose $\mathcal{X} : \mathbb{R}^{p_1 \times p_2} \to \mathbb{R}^n$ is ROP from the standard normal distribution. Let*

$$C_1 = \min_{A:\mathrm{rank}(A)=1} \frac{\|\mathcal{X}(A)\|_2/\sqrt{n}}{\|A\|_F} \quad \textit{and} \quad C_2 = \max_{A:\mathrm{rank}(A)=1} \frac{\|\mathcal{X}(A)\|_2/\sqrt{n}}{\|A\|_F}.$$

Then for all $t > 1$, $C_2/C_1 \geq \sqrt{p_1p_2/(4tn)}$ *with probability at least* $1 - e^{-p_1/4} - e^{-p_2/4} - \frac{8}{n(t-1)^2}$.

Lemma 2.1 implies that at least $O(p_1p_2)$ number of measurements are needed in order to ensure that $\mathcal{X}$ satisfies the RIP condition that guarantees the recovery of only rank-one matrices. Since $O(p_1p_2)$ is the degree of freedom for all matrices $A \in \mathbb{R}^{p_1\times p_2}$ and it is the number of measurements needed to recover all $p_1 \times p_2$ matrices (not just the low-rank matrices), Lemma 2.1 shows that the RIP framework is not suitable for the ROP model. In comparison, Theorem 2.2 shows that if $n \geq O(r(p_1 + p_2))$, then with high probability $\mathcal{X}$ satisfies the RUB condition of order r with bounded C_2/C_1 , which ensures the exact recovery of all rank-r matrices.

The main technical reason for the failure of RIP under the ROP model is that RIP requires an upper bound for

$$\max_{A\in\mathcal{C}} \|\mathcal{X}(A)\|_2^2/n = \max_{A\in\mathcal{C}} \left(\sum_{j=1}^{n} \left((\beta^{(j)})^\intercal A\gamma^{(j)}\right)^2\right)/n \tag{2.3}$$

where $\mathcal{C}$ is a set containing low-rank matrices. The right-hand side of (2.3) involves the 4th power of the Gaussian (or sub-Gaussian) variables $\beta^{(j)}$ and $\gamma^{(j)}$. A much larger n than the bound given in (2.2) is needed in order for the linear map $\mathcal{X}$ to satisfy the required RIP condition, which would lead to sub-optimal result.

Koltchinskii, et al. [25] uses RIP in expectation, which is a weaker condition than RIP. A random linear map $\mathcal{X} : \mathbb{R}^{p_1\times p_2} \to \mathbb{R}^n$ is said to satisfy RIP in expectation of order r with parameters $0 < \mu < \infty$ and $0 \leq \delta_r < 1$ if

$$(1 - \delta_r)\|A\|_F^2 \leq \mu\frac{1}{n}E\|\mathcal{X}(A)\|_2^2 \leq (1 + \delta_r)\|A\|_F^2$$

for all rank-r matrices $A \in \mathbb{R}^{p_1\times p_2}$. This condition was originally introduced by Koltchinskii, et al. [25] to prove an oracle inequality for the estimator they proposed and a minimax lower bound. The condition is not sufficiently strong to guarantee the exact recovery of rank-r matrices in the noiseless case. To be more specific, the bounds in Theorems 1 and 2 in [25] depend on $\mathbf{M} = \left\|\frac{1}{n}\sum_{i=1}^{n}(y_iX_i - E(y_iX_i))\right\|$, which might be non-zero even in the noiseless case. In fact, in the ROP model considered in the present paper, we have

$$\begin{aligned}\frac{1}{n}E\|\mathcal{X}\|_2^2 =& \frac{1}{n}\sum_{i=1}^{n} E\left(\beta^{(i)T}A\gamma^{(i)}\right)^2 = E(\beta^TA\gamma\gamma^TA^T\beta)\\ =& E\mathrm{tr}(A\gamma\gamma^TA^T\beta\beta^T) = \mathrm{tr}(AA^T) = \|A\|_F^2\end{aligned}$$

which means RIP in expectation is met for $\mu = 1$ and $\delta_r = 0$ for any number of measurements n. However, as we discussed earlier in this section that at least $O(r(p_1 + p_2))$ measurements are needed to guarantee the model identifiability for recovery of all rank-r matrices, we can see that RIP in expectation cannot ensure recovery.

Dvijotham and Fazel [19] and Oymak et al [31] used a condition called the spherical section property (SSP) which focuses on the null space of $\mathcal{X}$. $\text{Null}(\mathcal{X})$ is said to satisfy Δ-SSP if for all $Z \in \text{Null}(\mathcal{X})\backslash\{0\}$, $\|Z\|_*/\|Z\|_F \geq \sqrt{\Delta}$. Dvijotham and Fazel [19] showed that if $\mathcal{X}$ satisfies Δ-SSP, $p_1 \leq p_2$ and $\text{rank}(A) < \min\left(3p_1/4 - \sqrt{9p_1^2/16 - p_1\Delta/4}, p_1/2\right)$, the nuclear norm minimization (2.1) recovers A exactly in the noiseless case. However, the SSP condition is difficult to utilize in the ROP framework since it is hard to characterize the matrices $Z \in \text{Null}(\mathcal{X})$ when $\mathcal{X}$ is rank-one projections.

2.3. *Gaussian Noise Case.* We now turn to the Gaussian noise case where $z_i \overset{iid}{\sim} N(0, \sigma^2)$ in (1.5). We begin by introducing a constrained nuclear norm minimization estimator. Define two sets

$$\mathcal{Z}_1 = \{z : \|z\|_1/n \leq \sigma\} \quad \text{and} \quad \mathcal{Z}_2 = \{z : \|\mathcal{X}^*(z)\| \leq \eta\} \tag{2.4}$$

where $\eta = \sigma\left(12\sqrt{\log n}(p_1 + p_2) + 6\sqrt{2n(p_1 + p_2)}\right)$, and let

$$\mathcal{Z}_G = \mathcal{Z}_1 \cap \mathcal{Z}_2. \tag{2.5}$$

Note that both $\mathcal{Z}_1$ and $\mathcal{Z}_2$ are convex sets and so is $\mathcal{Z}_G$. Our estimator of A is given by

$$\hat{A} = \arg\min_M \{\|M\|_* : y - \mathcal{X}(M) \in \mathcal{Z}_G\}. \tag{2.6}$$

The following theorem gives the rate of convergence for the estimator $\hat{A}$ under the squared Frobenius norm loss.

THEOREM 2.3 (Upper Bound). *Let $\mathcal{X}$ be ROP from the standard normal distribution and let $z_1, \cdots, z_n \overset{iid}{\sim} N(0, \sigma^2)$. Then there exist uniform constants C, W and δ such that, whenever $n \geq Cr(p_1 + p_2)$, the estimator $\hat{A}$ given in (2.6) satisfies*

$$\|\hat{A} - A\|_F^2 \leq W\sigma^2 \min\left(\frac{r\log n(p_1 + p_2)^2}{n^2} + \frac{r(p_1 + p_2)}{n}, 1\right) \tag{2.7}$$

for all rank-r matrices A, with probability at least $1 - 11/n - 3\exp(-\delta(p_1 + p_2))$.

Moreover, we have the following lower bound result for ROP.

THEOREM 2.4 (Lower Bound). *Assume that $\mathcal{X}$ is ROP from the standard normal distribution and that $z_1, \cdots, z_n \overset{iid}{\sim} N(0, \sigma^2)$. There exists a uniform constant C such that, when $n > Cr\max(p_1, p_2)$, with probability at least $1 - 26n^{-1}$,*

(2.8)
$$\inf_{\hat{A}} \sup_{A \in \mathbb{R}^{p_1 \times p_2}: \mathrm{rank}(A)=r} P_z\left(\|\hat{A} - A\|_F^2 \geq \frac{\sigma^2 r(p_1+p_2)}{32n}\right) \geq 1 - e^{-(p_1+p_2)r/64}$$

$$\inf_{\hat{A}} \sup_{A \in \mathbb{R}^{p_1 \times p_2}: \mathrm{rank}(A)=r} E_z\|\hat{A} - A\|_F^2 \geq \frac{\sigma^2 r(p_1+p_2)}{4n} \tag{2.9}$$

where E_z, and P_z are the expectation and probability with respect to the distribution of z.

When $n < r\max(p_1, p_2)$, then

$$\inf_{\hat{A}} \sup_{A \in \mathbb{R}^{p_1 \times p_2}: \mathrm{rank}(A)=r} E_z\|\hat{A} - A\|_F^2 = \infty. \tag{2.10}$$

Comparing Theorem 2.3 and Theorem 2.4, our proposed estimator is rate optimal in the Gaussian noise case when $n \gtrsim \log n(p_1+p_2)$ (which is equivalent to $n \gtrsim (p_1+p_2)\log(p_1+p_2)$) or $n \sim r(p_1+p_2)$. Since $n \gtrsim r(p_1+p_2)$, this condition is also implied by $r \gtrsim \log(p_1+p_2)$. Theorem 2.4 also shows that no method can recover matrices of rank r consistently if the number of measurements n is smaller than $r\max(p_1, p_2)$.

The result in Theorem 2.3 can also be extended to the more general case where the matrix of interest A is only approximately low-rank. Let $A = A_{\max(r)} + A_{-\max(r)}$.

PROPOSITION 2.1. *Under the assumptions of Theorem 2.3, there exist uniform constants C, W_1, W_2 and δ such that, whenever $n \geq Cr(p_1+p_2)$, the estimator $\hat{A}$ given in (2.6) satisfies*

(2.11)
$$\|\hat{A} - A\|_F^2 \leq W_1\sigma^2 \min\left(\frac{r\log n(p_1+p_2)^2}{n^2} + \frac{r(p_1+p_2)}{n}, 1\right) + W_2\frac{\|A_{-\max(r)}\|_*^2}{r}$$

for all matrices $A \in \mathbb{R}^{p_1 \times p_2}$, with probability at least $1 - 11/n - 3\exp(-\delta(p_1+p_2))$.

If the matrix A is approximately of rank r, then $\|A_{-\max(r)}\|_*$ is small, and the estimator $\hat{A}$ continues to perform well. This result shows that the constrained nuclear norm minimization estimator is adaptive to the rank r and robust against perturbations of small amplitude.

REMARK 2.2. All the results remain true if the Gaussian design is replaced by the Rademacher design where entries of $\beta^{(i)}$ and $\gamma^{(i)}$ are i.i.d. ± 1 with probability $\frac{1}{2}$. More general sub-Gaussian design case will be discussed in Section 3.

REMARK 2.3. The estimator $\hat{A}$ we propose here is the minimizer of the nuclear norm under the constraint of the intersection of two convex sets $\mathcal{Z}_1$ and $\mathcal{Z}_2$. Nuclear norm minimization under either one of the two constraints, called "ℓ_1 constraint nuclear norm minimization" ($\mathcal{Z} = \mathcal{Z}_1$) and "matrix Dantzig Selector" ($\mathcal{Z} = \mathcal{Z}_2$), has been studied before in various settings [12, 35, 8, 9, 10, 16]. Our analysis indicates the following.

1. The ℓ_1 constraint minimization performs better than the matrix Dantzig Selector for small n ($n \sim r(p_1 + p_2)$) when $r \ll \log n$;
2. The matrix Dantzig Selector outperforms the ℓ_1 constraint minimization for large n as the loss of the matrix Dantzig Selector decays at the rate $O(n^{-1})$;
3. The proposed estimator $\hat{A}$ combines the advantages of the two estimators.

See Section 5 for a comparison of numerical performances of the three methods.

2.4. *Recovery of Symmetric Matrices.* For applications such as low-dimensional Euclidean embedding [37, 35], phase retrieval[14, 11], and covariance matrix estimation [16, 5, 6]), the low-rank matrix A of interest is known to be symmetric. Examples of such matrices include distance matrices, Gram matrices, and covariance matrices. When the matrix A is known to be symmetric, the ROP design can be further simplified by taking $\beta^{(i)} = \gamma^{(i)}$.

Denote by $\mathbb{S}^p$ the set of all $p \times p$ symmetric matrices in $\mathbb{R}^{p\times p}$. Let $\beta^{(1)}, \beta^{(2)}, \cdots, \beta^{(n)}$ be independent p-dimensional random vectors with i.i.d. entries generated from some distribution $\mathcal{P}$. Define a linear map $\mathcal{X} : \mathbb{S}^p \to \mathbb{R}^n$ by

$$[\mathcal{X}(A)]_i = (\beta^{(i)})^\intercal A \beta^{(i)}, \quad i = 1, \cdots, n.$$

We call such a linear map $\mathcal{X}$ "Symmetric Rank-One Projections" (SROP) from the distribution $\mathcal{P}$.

Suppose we observe

$$y_i = (\beta^{(i)})^\intercal A \beta^{(i)} + z_i, \quad i = 1, ..., n \tag{2.12}$$

and wish to recover the symmetric matrix A. As for the ROP model, in the noiseless case we estimate A under the SROP model by

$$A_* = \arg\min_{M \in \mathbb{S}^p} \{\|M\|_* : \ y = \mathcal{X}(M)\}. \tag{2.13}$$

PROPOSITION 2.2. *Let $\mathcal{X}$ be SROP from the standard normal distribution. Similar to Corollary 2.1, there exist uniform constants C and δ such that, whenever $n \geq Crp$, the nuclear norm minimization estimator A_* given by (2.13) recovers exactly all rank-r symmetric matrices $A \in \mathbb{S}^p$ with probability at least $1 - e^{-n\delta}$.*

For the noisy case, we propose a constraint nuclear norm minimization estimator similar to (2.6). Define the linear map $\tilde{\mathcal{X}} : \mathbb{R}^{p_1 \times p_2} \to \mathbb{R}^{\lfloor \frac{n}{2} \rfloor}$ by

$$[\tilde{\mathcal{X}}(A)]_i = [\mathcal{X}(A)]_{2i-1} - [\mathcal{X}(A)]_{2i}, \quad i = 1, \cdots, \lfloor \frac{n}{2} \rfloor \tag{2.14}$$

and define $\tilde{y} \in \mathbb{R}^{\lfloor n/2 \rfloor}$ by

$$\tilde{y}_i = y_{2i-1} - y_{2i}, \quad i = 1, \cdots, \lfloor \frac{n}{2} \rfloor. \tag{2.15}$$

Based on the definition of $\tilde{\mathcal{X}}$, the dual map $\tilde{\mathcal{X}}^* : \mathbb{R}^{\lfloor \frac{n}{2} \rfloor} \to \mathbb{S}^p$ is

$$\tilde{\mathcal{X}}^*(z) = \sum_{i=1}^{\lfloor \frac{n}{2} \rfloor} z_i \left(\beta^{(2i-1)} \beta^{(2i-1)\intercal} - \beta^{(2i)} \beta^{(2i)\intercal} \right) \tag{2.16}$$

Let $\eta = 24\sigma \left(\sqrt{pn} + 2p\sqrt{2 \log n} \right)$. The estimator $\hat{A}$ of the matrix A is given by

$$\hat{A} = \arg\min_{M \in \mathbb{S}^p} \left\{ \|M\|_* : \ \|y - \mathcal{X}(M)\|_1 / n \leq \sigma, \ \|\tilde{\mathcal{X}}^*(\tilde{y} - \tilde{\mathcal{X}}(M))\| \leq \eta \right\}. \tag{2.17}$$

REMARK 2.4. An important property in the ROP model considered in Section 2.3 is that $E\mathcal{X} = 0$, i.e., $EX_i = 0$ for all the measurement matrices X_i. However, under the SROP model $X_i = \beta^{(i)}(\beta^{(i)})^\intercal$ and so $E\mathcal{X} \neq 0$. The step of taking the pairwise differences in (2.14) and (2.15) is to ensure that $E\tilde{\mathcal{X}} = 0$.

The following result is similar to the upper bound given in Proposition 2.1 for ROP.

PROPOSITION 2.3. *Let $\mathcal{X}$ be SROP from the standard normal distribution and let $z_1, \cdots, z_n \overset{iid}{\sim} N(0, \sigma^2)$. There exist constants C, W_1, W_2 and δ such that, whenever $n \geq Crp$, the estimator $\hat{A}$ given in (2.17) satisfies*

$$\|\hat{A} - A\|_F^2 \leq W_1 \sigma^2 \min\left(\frac{rp^2 \log n}{n^2} + \frac{rp}{n}, 1\right) + W_2 \frac{\|A_{-\max(r)}\|_*^2}{r} \tag{2.18}$$

for all matrices $A \in \mathbb{S}^p$, with probability at least $1 - 15/n - 5\exp(-p\delta)$.

In addition, we also have lower bounds for SROP, which show that the proposed estimator is rate-optimal when $n \gtrsim p \log n$ or $n \sim rp$, and no estimator can recover a rank-r matrix consistently if the number of measurements $n < \lfloor \frac{r}{2} \rfloor \cdot \lfloor \frac{p}{2} \rfloor$.

PROPOSITION 2.4 (Lower Bound). *Assume that $\mathcal{X}$ is SROP from the standard normal distribution and that $z_1, \cdots, z_n \overset{iid}{\sim} N(0, \sigma^2)$. Then there exists a uniform constant C such that, when $n > Crp$ and $p, r \geq 2$, with probability at least $1 - 26n^{-1}$,*

$$\inf_{\hat{A}} \sup_{A \in \mathbb{S}^p : \mathrm{rank}(A) = r} P_z\left(\|\hat{A} - A\|_F^2 \geq \frac{\sigma^2 rp}{192n}\right) \geq 1 - e^{-pr/192}$$

$$\inf_{\hat{A}} \sup_{A \in \mathbb{S}^p : \mathrm{rank}(A) = r} E_z \|\hat{A} - A\|_F^2 \geq \frac{\sigma^2 rp}{24n}$$

where $\hat{A}$ is any estimator of A, E_z, P_z are the expectation and probability with respect to z.

When $n < \lfloor \frac{r}{2} \rfloor \cdot \lfloor \frac{p}{2} \rfloor$ and $p, r \geq 2$, then

$$\inf_{\hat{A}} \sup_{A \in \mathbb{S}^p : \mathrm{rank}(A) = r} E_z \|\hat{A} - A\|_F^2 = \infty.$$

3. Sub-Gaussian Design and Sub-Gaussian Noise. We have focused on the Gaussian design and Gaussian noise distribution in Section 2. These results can be further extended to more general distributions. In this section we consider the case where the ROP design is from a symmetric sub-Gaussian distribution $\mathcal{P}$ and the errors z_i are also from a sub-Gaussian distribution. We say the distribution of a random variable Z is sub-Gaussian with parameter τ if

$$P(|Z| \geq t) \leq 2\exp(-t^2/(2\tau^2)), \quad \text{for all } t > 0. \tag{3.1}$$

The following lemma provides a necessary and sufficient condition for symmetric sub-Gaussian distributions.

LEMMA 3.1. *Let $\mathcal{P}$ be a symmetric distribution and let the random variable $X \sim \mathcal{P}$. Define*

$$\alpha_{\mathcal{P}} = \sup_{k\geq 1} \left(\frac{EX^{2k}}{(2k-1)!!} \right)^{\frac{1}{2k}}. \tag{3.2}$$

Then the distribution $\mathcal{P}$ is sub-Gaussian if and only if $\alpha_{\mathcal{P}}$ is finite.

For the sub-Gaussian ROP design and sub-Gaussian noise, we estimate the low-rank matrix A by the estimator $\hat{A}$ given in (1.3) with

$$\begin{aligned} \mathcal{Z}_G =& \{z : \|z\|_1/n \leq 6\tau\} \\ & \cap \left\{ z : \|\mathcal{X}^*(z)\| \leq 6\alpha_{\mathcal{P}}^2 \tau \left(\sqrt{6n(p_1+p_2)} + 2\sqrt{\log n}(p_1+p_2) \right) \right\} \end{aligned} \tag{3.3}$$

where $\alpha_{\mathcal{P}}$ is given in (3.2).

THEOREM 3.1. *Suppose $\mathcal{X} : \mathbb{R}^{p_1\times p_2} \to \mathbb{R}^n$ is ROP from a symmetric and variance 1 sub-Gaussian distribution $\mathcal{P}$. Assume that z_i are i.i.d. sub-Gaussian with parameter τ and $\hat{A}$ is given by (1.3) with $\mathcal{Z} = \mathcal{Z}_G$ defined in (3.3). Then there exist constants C, W_1, W_2, δ which only depend on $\mathcal{P}$, such that if $n \geq Cr(p_1+p_2)$, we have*

$$\|\hat{A}-A\|_F^2 \leq W_1\tau^2 \min\left(\frac{r\log n(p_1+p_2)^2}{n^2} + \frac{r(p_1+p_2)}{n}, 1 \right) + W_2 \frac{\|A_{-\max(r)}\|_*^2}{r} \tag{3.4}$$

with probability at least $1 - 2/n - 5e^{-\delta(p_1+p_2)}$.

An exact recovery result in the noiseless case for the sub-Gaussian design follows directly from Theorem 3.1. If $z = 0$, then, with high probability, all rank-r matrices A can be recovered exactly via the constrained nuclear minimization (2.1) whenever $n \geq C_{\mathcal{P}} r(p_1+p_2)$ for some constant $C_{\mathcal{P}} > 0$.

REMARK 3.1. For the SROP model considered in Section 2.4, we can similarly extend the results to the case of sub-Gaussian design and sub-Gaussian noise. Suppose $\mathcal{X}$ is SROP from a symmetric variance 1 sub-Gaussian distribution $\mathcal{P}$ (other than the Rademacher ± 1 distribution) and z satisfies (3.1). Define the estimator of the low-rank matrix A by

$$\hat{A} = \arg\min_{M\in\mathbb{S}^p} \left\{ \|M\|_* : \|y - \mathcal{X}(M)\|_1/n \leq 6\tau, \|\tilde{\mathcal{X}}^*(\tilde{y} - \tilde{\mathcal{X}}(M))\| \leq \eta \right\} \tag{3.5}$$

where $\eta = C_{\mathcal{P}} \left(\sqrt{np} + \sqrt{\log np} \right)$ with $C_{\mathcal{P}}$ some constant depending on $\mathcal{P}$.

PROPOSITION 3.1. *Suppose $\mathcal{X} : \mathbb{R}^{p\times p} \to \mathbb{R}^n$ is SROP from a symmetric sub-Gaussian distribution $\mathcal{P}$ with variance 1. Also, assume that* $\mathrm{Var}(\mathcal{P}^2) > 0$ *(i.e.* $\mathrm{Var}(w^2) > 0$ *where $w \sim \mathcal{P}$). Let $\hat{A}$ be given by (3.5). Then there exist constants $C, C_{\mathcal{P}}, W_1, W_2$, and δ which only depend on $\mathcal{P}$, such that for $n \geq Crp$,*

$$\|\hat{A} - A\|_F^2 \leq W_1 \tau^2 \min\left(\frac{rp^2 \log n}{n^2} + \frac{rp}{n}, 1\right) + W_2 \frac{\|A_{-\max(r)}\|_*^2}{r} \tag{3.6}$$

with probability at least $1 - 2/n - 5e^{-\delta p}$.

By restricting $\mathrm{Var}(\mathcal{P}^2) > 0$, Rademacher ± 1 is the only symmetric and variance 1 distribution that has been excluded. The reason why the Rademacher ± 1 distribution is an exception for the SROP design is as follows. If $\beta^{(i)}$ are i.i.d. Rademacher ± 1 distributed, then

$$[\mathcal{X}(A)]_i = (\beta^{(i)})^\intercal A \beta^{(i)} = \sum_{j=1}^p a_{jj} + \sum_{j\neq k} \beta_j^{(i)} \beta_k^{(i)} a_{jk}, \quad i = 1, \cdots, n.$$

So the only information contained in $\mathcal{X}(A)$ about $\mathrm{diag}(A)$ is $\mathrm{trace}(A)$, which makes it impossible to recover the whole matrix A.

4. Application to Estimation of Spiked Covariance Matrix. In this section, we consider an interesting application of the methods and results developed in the previous sections to estimation of a spiked covariance matrix based on one-dimensional projections. As mentioned in the introduction, spiked covariance matrix model has been used in a wide range of applications and it has been well studied in the context of PCA based on i.i.d. data where one observes i.i.d. p-dimensional random vectors $X^{(1)}, \cdots, X^{(n)}$ with mean 0 and covariance matrix Σ, where $\Sigma = I_p + \Sigma_0$ and Σ_0 being low-rank. See, for example, [24, 4, 5, 6]. Here we consider estimation of Σ_0 (or equivalently Σ) based only on one-dimensional random projections of $X^{(i)}$. More specifically, suppose that the random vectors $X^{(1)}, \cdots, X^{(n)}$ are not directly observable and instead we observe

$$\xi_i = \langle \beta^{(i)}, X^{(i)} \rangle = \sum_{j=1}^p \beta_j^{(i)} X_j^{(i)}, \quad i = 1, ..., n, \tag{4.1}$$

where $\beta^{(i)} \overset{iid}{\sim} N(0, I_p)$. The goal is to recover Σ_0 from the projections $\{\xi_i,\ i = 1, \cdots, n\}$.

Let $y = (y_1, ..., y_n)^\intercal$ with $y_i = \xi_i^2 - \beta^{(i)\intercal}\beta^{(i)}$. Note that

$$E(\xi^2|\beta) = E\left(\sum_{i,j} \beta_i\beta_j X_i X_j|\beta\right) = \sum_{i,j} \beta_i\beta_j\sigma_{i,j} = \beta^\intercal\Sigma\beta$$

and so $E(\xi^2 - \beta^\intercal\beta|\beta) = \beta^\intercal\Sigma_0\beta$. Define a linear map $\mathcal{X} : \mathbb{S}^p \to \mathbb{R}^n$ by

$$[\mathcal{X}(A)]_i = \beta^{(i)\intercal}A\beta^{(i)}. \tag{4.2}$$

Then y can be formally written as

$$y = \mathcal{X}(\Sigma_0) + z \tag{4.3}$$

where $z = y - \mathcal{X}(\Sigma_0)$. We define the corresponding $\tilde{\mathcal{X}}$ and $\tilde{y}$ as in (2.14) and (2.15) respectively, and apply the constraint nuclear norm minimization to recover the low-rank matrix Σ_0 by

$$\hat{\Sigma}_0 = \arg\min_M \left\{\|M\|_* : \ \|y - \mathcal{X}(M)\| \le \eta_1, \|\tilde{\mathcal{X}}(\tilde{y} - \tilde{\mathcal{X}}(M))\| \le \eta_2\right\}. \tag{4.4}$$

The tuning parameters η_1 and η_2 are chosen as

$$\eta_1 = c_1\sum_{i=1}^n \xi_i^2 \quad \text{and} \quad \eta_2 = 24c_2\sqrt{p\sum_{i=1}^n \xi_i^4} + 48c_3 p\log n \max_{1\le i\le n} \xi_i^2 \tag{4.5}$$

where $c_1 > \sqrt{2}$, $c_2, c_3 > 1$ are constants.

We have the following result on the estimator (4.4) for spiked covariance matrix estimation.

THEOREM 4.1. *Suppose $n \ge 3$, we observe $\xi_i, i = 1, \cdots, n$, as in (4.1), where $\beta^{(i)} \overset{iid}{\sim} N(0, I_p)$ and $X^{(1)}, \cdots, X^{(n)} \overset{iid}{\sim} N(0, \Sigma)$ with $\Sigma = I_p + \Sigma_0$ and Σ_0 positive semidefinite and rank$(\Sigma_0) \le r$. Let $\hat{\Sigma}_0$ be given by (4.4). Then there exist uniform constants C, D, δ such that when $n \ge Drp$,*

$$\|\hat{\Sigma}_0 - \Sigma_0\|_F^2 \le C\min\left(\frac{rp}{n}\|\Sigma\|_*^2 + \frac{rp^2\log^4 n}{n^2}(\|\Sigma\|_*^2 + \log^2 n\|\Sigma\|^2),\ \|\Sigma\|_*^2\right) \tag{4.6}$$

with probability at least $1 - O(1/n) - 4\exp(-p\delta) - \frac{2}{\sqrt{2\pi\log n}}$.

REMARK 4.1. We have focused estimation of spiked covariance matrices on the setting where the random vectors $X^{(i)}$ are Gaussian. Similar to the discussion in Section 3, the results given here can be extended to more general distributions under certain moment conditions.

REMARK 4.2. The problem considered in this section is related to the so-called covariance sketching problem considered in Dasarathy et al. [17]. In covariance sketching, the goal is to estimate the covariance matrix of high-dimensional random vectors $X^{(1)}, \cdots, X^{(n)}$ based on the low dimensional projections

$$y^{(i)} = QX^{(i)}, \quad i = 1, ...n,$$

where Q is a fixed $m \times p$ projection matrix with $m < p$. The main differences between the two settings are that the projection matrix in covariance sketch is the same for all $X^{(i)}$ and the dimension m is still relatively large with $m \geq C\sqrt{p}\log^3 p$ for some $C > 0$. In our setting, $m = 1$ and Q is random and varies with i. The techniques for solving the two problems are very different. Comparing to [17], the results in this section indicate that there is a significant advantage to have different random projections for different random vectors $X^{(i)}$ as opposed to having the same projection for all $X^{(i)}$.

5. Simulation Results. The constrained nuclear norm minimization methods can be efficiently implemented. The estimator $\hat{A}$ proposed in Section 2.3 can be implemented by the following convex programming:

(5.1)

$$\begin{aligned} &\text{minimize} && \mathrm{Tr}(B_1) + \mathrm{Tr}(B_2) \\ &\text{subject to} && \begin{bmatrix} B_1 & A \\ A^T & B_2 \end{bmatrix} \succeq 0, \quad \|y - \mathcal{X}(A)\|_1 \leq \lambda_1, \quad \|\mathcal{X}^*(y - \mathcal{X}(A))\| \leq \lambda_2, \end{aligned}$$

with optimization variables $B_1 \in \mathbb{S}^{p_1}, B_2 \in \mathbb{S}^{p_2}$, $A \in \mathbb{R}^{p_1 \times p_2}$. We use the CVX package [21, 22] to implement the proposed procedures. In this section, a simulation study is carried out to investigate the numerical performance of the proposed procedures for low-rank matrix recovery in various settings.

We begin with the noiseless case. In this setting, Theorem 2.2 and Corollary 2.1 show that the nuclear norm minimization recovers a rank r matrix exactly whenever

$$n \geq Cr \max(p_1, p_2). \tag{5.2}$$

A similar result holds for the Gaussian ensemble [12]. However, the minimum constant C that guarantees the exact recovery with high probability is not specified in either case. It is of practical interest to find the minimum constant C. For this purpose, we randomly generate $p_1 \times p_2$ rank r matrices A as $A = X^\intercal Y$, where $X \in \mathbb{R}^{r \times p_1}$, $Y \in \mathbb{R}^{r \times p_2}$ are i.i.d. Gaussian matrices. We compare ROP from the standard Gaussian distribution and the Gaussian ensemble, with the number of measurements $n = Cr \max(p_1, p_2)$ from a

range of values of C using the constrained nuclear norm minimization (2.1). A recovery is considered successful if $\|\hat{A} - A\|_F / \|A\|_F \leq 10^{-4}$. Figure 1 shows the rate of successful recovery when $p_1 = p_2 = 100$ and $r = 5$.

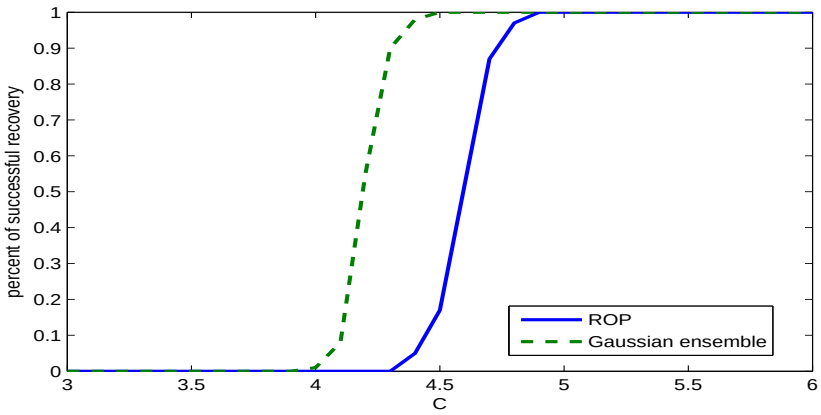

FIG 1. *Rates of successful recovery for the ROP and Gaussian ensemble with* $p_1 = p_2 = 100$, $r = 5$, *and* $n = Cr\max(p_1, p_2)$ *for* C *ranging from 3 to 6.*

The numerical results show that for ROP from the Gaussian distribution, the minimum constant C to ensure exact recovery with high probability is slightly less than 5 in the small scale problems ($p_1,\ p_2 \leq 100$) we tested. The corresponding minimum constant C for the Gaussian ensemble is about 4.5. Matrix completion requires much larger number of measurements. Based on the theoretical analyses given in [13, 34], the required number of measurements for matrix completion is $O(\mu r(p_1 + p_2)\log^2(p_1 + p_2))$, where $\mu \geq 1$ is some coherence constant describing the "spikedness" of the matrix A. Hence for matrix completion, the factor C in (5.2) needs to grow with the dimensions p_1 and p_2 and it requires $C \gtrsim \mu \log^2(p_1 + p_2)$, which is much larger than what is needed for the ROP or Gaussian ensemble. The required storage space for the Gaussian ensemble is much greater than that for the ROP. In order to ensure accurate recovery of $p \times p$ matrices of rank r, one needs at least $4.5rp^3$ bytes of space to store the measurement matrices, which could be prohibitively large for the recovery of high-dimensional matrices. In contrast, the storage space for the projection vectors in ROP is only $10rp^2$ bytes, which is far smaller than what is required by the Gaussian ensemble in the high-dimensional case.

We then consider the recovery of approximately low-rank matrices to investigate the robustness of the method against small perturbations. To this

end, we randomly draw 100×100 matrix A as $A = U\cdot\mathrm{diag}(1, 2^{-1/2}, \cdots, r^{-1/2})\cdot V^{\intercal}$, where $U \in \mathbb{R}^{100\times r}$ and $V \in \mathbb{R}^{100\times r}$ are random matrices with orthonormal columns. We then observe $n = 2000$ random rank-one projections with the measurement vectors being i.i.d. Gaussian. Based on the observations, the nuclear minimization procedure (2.1) is applied to estimate A. The results for different values of r are shown in Figure 2. It can be seen from the plot that in this setting one can exactly recover a matrix of rank at most 4 with 2000 measurements. However, when the rank r of the true matrix A exceeds 4, the estimate is still stable. The theoretical result in Proposition 2.1 bounds the loss (blue line) at $O(\|A_{-\max(4)}\|_*^2/4)$ (shown in the dashed green line) with high probability, which corresponds to Figure 2.

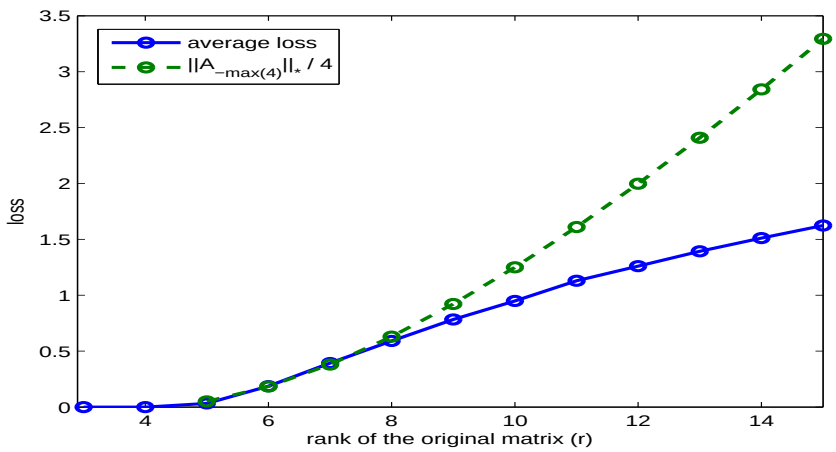

FIG 2. *Recovery accuracy (solid blue line) for approximately low-rank matrices with different values of r, where $p_1 = p_2 = 100$, $n = 2000$, $\sigma(A) = (1, 1/\sqrt{2}, \cdots, 1/\sqrt{r})$. The dashed green line is the theoretical upper bound.*

We now turn to the noisy case. The low-rank matrices A are generated by $A = X^{\intercal}Y$, where $X \in \mathbb{R}^{r\times p_1}$ and $Y \in \mathbb{R}^{r\times p_2}$ are i.i.d. Gaussian matrices. The ROP $\mathcal{X}$ is from the standard Gaussian distribution and the noise vector $z \sim N_n(0, \sigma^2)$. Based on $(\mathcal{X}, y)$ with $y = \mathcal{X}(A)+z$, we compare our proposed estimator $\hat{A}$ with the ℓ_1 constraint minimization estimator $\hat{A}^{\ell_1}$ [16] and the

matrix Dantzig Selector $\hat{A}^{DS}$ [12], where

$$\hat{A} = \arg\min_M \{\|M\|_* : y - \mathcal{X}(M) \in \mathcal{Z}_1 \cap \mathcal{Z}_2\},$$
$$\hat{A}^{\ell_1} = \arg\min_M \{\|M\|_* : y - \mathcal{X}(M) \in \mathcal{Z}_1\},$$
$$\hat{A}^{DS} = \arg\min_M \{\|M\|_* : y - \mathcal{X}(M) \in \mathcal{Z}_2\},$$

with $\mathcal{Z}_1 = \{z : \|z\|_1/n \le \sigma\}$ and $\mathcal{Z}_2 = \{z : \|\mathcal{X}(z)\| \le \sigma(\sqrt{\log n}(p_1 + p_2) + \sqrt{n(p_1 + p_2)})\}$. Note that $\hat{A}^{\ell_1}$ is similar to the estimator proposed in Chen et al [16], except their estimator is for symmetric matrices under the SROP but ours is for general low-rank matrices under the ROP. Figure 3 compares the performance of the three estimators. It can be seen from the left panel that for small n, ℓ_1 constrained minimization outperforms the matrix Dantzig Selector, while our estimator outperforms both $\hat{A}^{\ell_1}$ and $\hat{A}^{DS}$. When n is large, our estimator and $\hat{A}^{DS}$ are essentially the same and both outperforms $\hat{A}^{\ell_1}$. The right panel of Figure 3 plots the ratio of the squared Frobenius norm loss of $\hat{A}^{\ell_1}$ to that of our estimator. The ratio increases with n. These numerical results are consistent with the observations made in Remark 2.3.

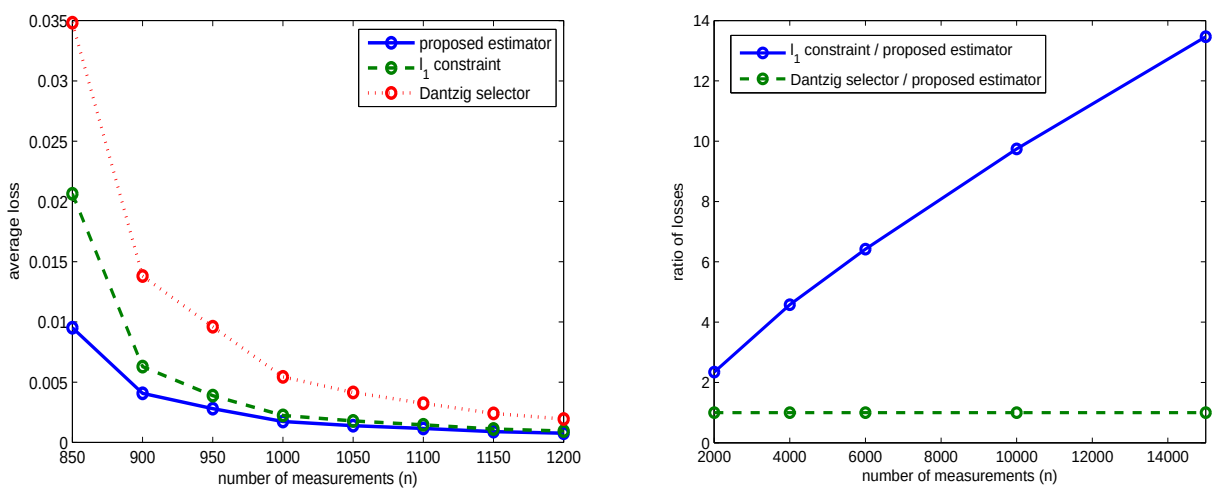

FIG 3. *Left Panel: Comparison of the proposed estimator with $\hat{A}^{\ell_1}$ and $\hat{A}^{DS}$ for $p_1 = p_2 = 50$, $r = 4$, $\sigma = 0.01$, and n ranging from 850 to 1200. Right Panel: Ratio of the squared Frobenius norm loss of $\hat{A}^{\ell_1}$ to that of the proposed estimator for $p_1 = p_2 = 50$, $r = 4$, and n varying from 2000 to 15000.*

We now turn to the recovery of symmetric low-rank matrices under the

SROP model (2.12). Let $\mathcal{X}$ be SROP from the standard normal distribution. We consider the setting where $p = 40$, n varies from 50 to 600, $z_i \sim \sigma \cdot \mathcal{U}[-1, 1]$ with $\sigma = .1, .01, .001$ or $.0001$, and A is randomly generated as rank-5 matrix by the same procedure discussed above. The setting is identical to the one considered in Section 5.1 of [16]. Although we cannot exactly repeat the simulation study in [16] as they did not specify the choice of the tuning parameter, we can implement both our procedure

$$\hat{A} = \arg\min_M \Big\{ \|M\|_* : \|y - \mathcal{X}(M)\|_1 \le \frac{n\sigma}{2}, \\ \|\tilde{\mathcal{X}}^*(\tilde{y} - \tilde{\mathcal{X}}(M))\| \le \frac{\sigma(\sqrt{\log n}p + \sqrt{np})}{3} \Big\}$$

and the estimator $\hat{A}^{\ell_1}$ with only the ℓ_1 constraint which was proposed by Chen et al [16]

$$\hat{A}^{\ell_1} = \arg\min_M \Big\{ \|M\|_* : \ \|y - \mathcal{X}(M)\|_1 \le \frac{n\sigma}{2} \Big\}.$$

The results are given in Figure 4. It can be seen that our estimator $\hat{A}$ outperforms the estimator $\hat{A}^{\ell_1}$.

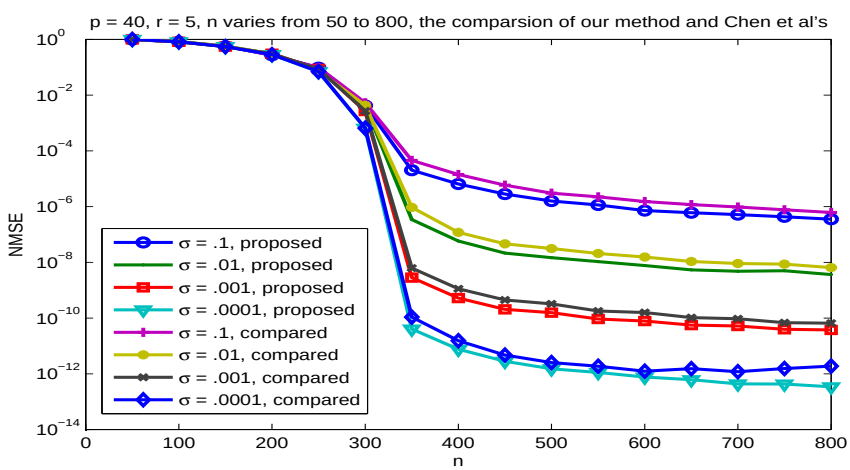

FIG 4. *Comparison of the proposed estimator $\hat{A}$ with the $\hat{A}^{\ell_1}$. Here $p = 40$, $r = 5$, $\sigma = 0.1, 0.01, 0.001, 0.0001$, and n ranges from 50 to 800.*

5.1. *Data Driven Selection of Tuning Parameters.* We have so far considered the estimators

$$\hat{A} = \arg\min_{B}\{\|B\|_* : \ \|y - \mathcal{X}(B)\|_1/n \le \lambda, \ \|\mathcal{X}^*(y - \mathcal{X}(B))\| \le \eta\}, \tag{5.3}$$

$$\hat{A} = \arg\min_{M}\{\|M\|_* : \ \|y - \mathcal{X}(M)\|_1/n \le \lambda, \ \|\tilde{\mathcal{X}}^*(\tilde{y} - \tilde{\mathcal{X}}(M))\| \le \eta\} \tag{5.4}$$

for the ROP and SROP, respectively. The theoretical choice of the tuning parameters λ and η depends on the knowledge of the error distribution such as the variance. In real applications, such information may not be available and/or the theoretical choice may not be the best. It is thus desirable to have a data driven choice of the tuning parameters. We now introduce a practical method for selecting the tuning parameters using K-fold cross-validation.

Let $(\mathcal{X}, y) = \{(X_i, y_i), i = 1, \cdots, n\}$ be the observed sample and let T be a grid of positive real values. For each $t \in T$, set

$$(\lambda, \eta) = (\lambda(t), \eta(t)) = \begin{cases} \left(t, t\left(\sqrt{\log n}(p_1 + p_2) + \sqrt{n(p_1 + p_2)}\right)\right) & \text{for ROP;} \\ \left(t, t\left(\sqrt{\log n}p + \sqrt{np}\right)\right) & \text{for SROP.} \end{cases} \tag{5.5}$$

Randomly split the n samples $(X_i, y_i), i = 1, \cdots, n$ into two groups of sizes $n_1 \sim \frac{(K-1)n}{K}$ and $n_2 \sim \frac{n}{K}$ for I times. Denote by $J_1^i, J_2^i \subseteq \{1, \cdots, n\}$ the index sets for Group 1 and Group 2 respectively for the i-th split. Apply our procedure ((5.3) for ROP and(5.4) for SROP, respectively) to the sub-samples in Group 1 with the tuning parameters $(\lambda(t), \eta(t))$ and denote the estimators by $\hat{A}^i(t)$, $i = 1, \cdots, I$. Evaluate the prediction error of $\hat{A}^i(t)$ over the sub-sample in Group 2 and set

$$\hat{R}(t) = \sum_{i=1}^{I} \sum_{j \in J_2^i} |y_j - \langle A^i(t), X_j\rangle|^2, \quad t \in T.$$

We select

$$t_* = \arg\min_{T} \hat{R}(t)$$

and choose the tuning parameters $(\lambda(t_*), \eta(t_*))$ as in (5.5) with $t = t_*$ and the final estimator $\hat{A}$ based on (5.3) or (5.4) with the chosen tuning parameters.

We compare the numerical result by 5-fold cross-validation with the result based on the known σ by simulation in Figure 5. Both the ROP and SROP are considered. It can be seen that the estimator with the tuning parameters chosen through 5-fold cross-validation has the same performance as or outperforms the one with the theoretical choice of the tuning parameters.

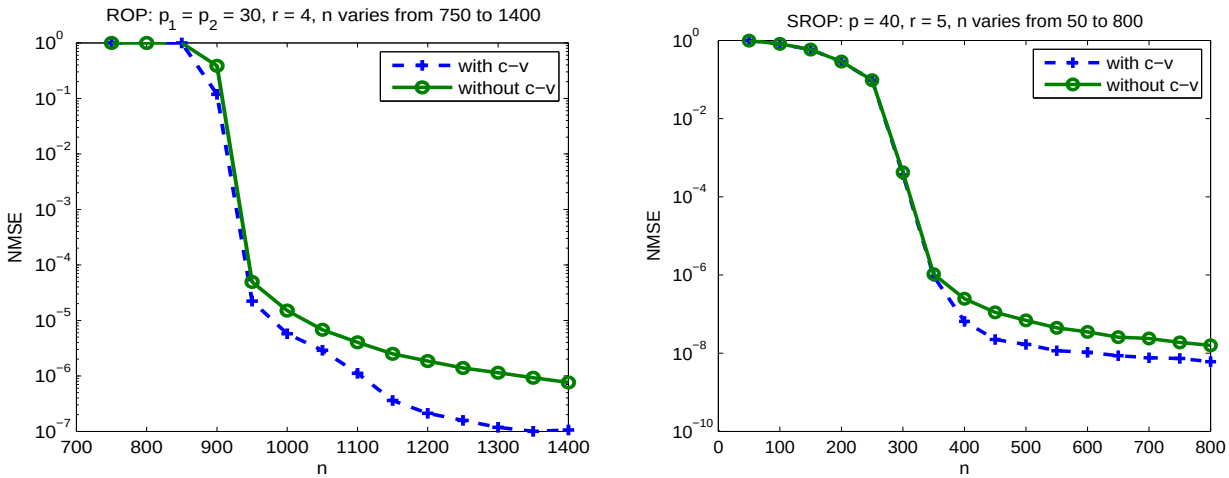

FIG 5. *Comparison of the performance with cross validation and without cross-validation in both ROP and SROP. Left panel: ROP, $p_1 = p_2 = 30$, $r = 4$, n varies from 750 to 1400. Right panel: SROP, $p = 40$, $r = 5$, n varies from 50 to 800.*

5.2. *Image Compression.* Since a two-dimensional image can be considered as a matrix, one approach to image compression is by using low-rank matrix approximation via the singular value decomposition. See, for example, [2, 35, 39]. Here we use an image recovery example to further illustrate the nuclear norm minimization method under the ROP model.

For a grayscale image, let $A = (a_{i,j}) \in \mathbb{R}^{m\times n}$ be the intensity matrix associated with the image, where a_{ij} is the grayscale intensity of the (i, j) pixel. When the matrix A is approximately low-rank, the ROP model and nuclear norm minimization method can be used for image compression and recovery. To illustrate this point, let us consider the following grayscale MIT Logo image (Figure 6).

FIG 6. *Original grayscale MIT logo*

The matrix associated with MIT logo is of the size 50×80 and of rank 6. We take rank-one random projections $\mathcal{X}(A)$ as the observed sample, with various sample sizes. Then the constrained nuclear norm minimization method is applied to reconstruct the original low-rank matrix. The recovery results are shown in Figure 7. The results show that the original image can be compressed and recovered well via the ROP model and the nuclear norm minimization.

FIG 7. *recovery of MIT logo based on different number of measurements. Left: 900; Middle: 1000; Right: 1080.*

6. Discussions. This paper introduces the ROP model for the recovery of general low-rank matrices. A constrained nuclear norm minimization method is proposed and its theoretical and numerical properties are studied. The proposed estimator is shown to be rate-optimal when the number of rank-one projections $n \gtrsim \log n(p_1+p_2)$ or $n \sim r(p_1+p_2)$. It is also shown that the procedure is adaptive to the rank and robust against small perturbations. The method and results are applied to estimation of a spiked covariance matrix. It is somewhat unexpected that it is possible to accurately recover a spiked covariance matrix from only one-dimensional projections. An interesting open problem is to estimate the principal components/subspace based on the one-dimensional random projections. We leave this as future work.

In a recent paper, Chen et al [16] considered quadratic measurements for the recovery of symmetric positive definite matrices, which is similar to the special case of SROP that we studied here. The paper was posted on arXiv as we finish writing the present paper. They considered the noiseless and ℓ_1 bounded noise cases and introduced the so-called "RIP-ℓ_2/ℓ_1" condition. The "RIP-ℓ_2/ℓ_1" condition is similar to RUB in our work. But these two conditions are not identical as the RIP-ℓ_2/ℓ_1 condition can only be applied to symmetric low-rank matrices as only symmetric operators are considered in the paper. In contrast, RUB applies to all low-rank matrices.

Chen et al. ([16] version 4) considered ℓ_1-bounded noise case under the SROP model and gave an upper bound in their Theorem 3 (after a slight

change of notation)

$$\|\hat{\Sigma} - \Sigma\|_F \leq C_1 \frac{\|\Sigma - \Sigma_\Omega\|_*}{\sqrt{r}} + C_2 \frac{\varepsilon}{n}. \tag{6.1}$$

This result for ℓ_1 bounded noise case is not applicable to the i.i.d. random noise setting. When the entries of the noise term $\eta \in \mathbb{R}^n$ are of constant order, which is the typical case for i.i.d. noise with constant variance, one has $\|\eta\|_1 \sim Cn$ with high probability. In such a case, the term $C_2 \frac{\epsilon_1}{n}$ on the right hand side of (6.1) does not even converge to 0 as the sample size $n \to \infty$.

In comparison, the bound (3.6) in Proposition 3.1 can be equivalently rewritten as

$$\|\hat{A} - A\|_F \leq W_2 \frac{\|A_{-\max(r)}\|_*}{\sqrt{r}} + W_1 \tau \min\left(\frac{\sqrt{r \log n}\, p}{n} + \sqrt{\frac{rp}{n}}, 1\right) \tag{6.2}$$

where the first term $W_2 \frac{\|A_{-\max(r)}\|_*}{\sqrt{r}}$ is of the same order as $C_1 \frac{\|\Sigma - \Sigma_\Omega\|_*}{\sqrt{r}}$ in (6.1) while the second term decays to 0 as $n \to \infty$. Hence, for the recovery of rank-r matrices, as the sample size n increases our bound decays to 0 but the bound (6.1) given in Chen et al. [16] does not. The main reason of this phenomenon lies in the difference in the two methods: we use nuclear norm minimization under two convex constraints (See Remark 2.3), but Chen et al. [16] used only the ℓ_1 constraint. Both theoretical results (see Remark 2.3) and numerical results (Figure 3 in Section 5) show that the additional constraint $\mathcal{Z}_2$ improves the performance of the estimator.

Moreover, the results and techniques in [16] for symmetric positive definite matrices are not applicable to the recovery of general non-symmetric matrices. This is due to the fact that for a non-symmetric square matrix $A = (a_{ij})$, the quadratic measurements $(\beta^{(i)})^\intercal A \beta^{(i)}$ satisfy

$$(\beta^{(i)})^\intercal A \beta^{(i)} = (\beta^{(i)})^\intercal A^s \beta^{(i)},$$

where $A^s = \frac{1}{2}(A + A^\intercal)$. Hence, for a non-symmetric matrix A, only its symmetrized version A^s can be possibly identified and estimated based on the quadratic measurements, the matrix A itself is neither identifiable nor estimable.

7. Proofs. We prove the main results in this section. We begin by collecting a few important technical lemmas that will be used in the proofs of the main results. The proofs of some of these technical lemmas are involved and are postponed to the Appendix.

7.1. *Technical Tools.* Lemmas 7.1 and 7.2 below are used for deriving the RUB condition (see Definition 2.1) from the ROP design.

LEMMA 7.1. *Suppose $A \in \mathbb{R}^{p_1 \times p_2}$ is a fixed matrix and $\mathcal{X}$ is ROP from a symmetric sub-Gaussian distribution $\mathcal{P}$, i.e.*

$$[\mathcal{X}(A)]_j = \beta^{(j)T} A \gamma^{(j)}, \quad j = 1, \cdots, n$$

where $\beta^{(j)} = (\beta_1^{(j)}, \cdots, \beta_{p_1}^{(j)})^T, \gamma^{(j)} = (\gamma_1^{(j)}, \cdots, \gamma_{p_2}^{(j)})^T$ are random vectors with entries i.i.d. generated from $\mathcal{P}$. Then for $\delta > 0$, we have

$$\left(\frac{1}{3\alpha_{\mathcal{P}}^4} - 2\alpha_{\mathcal{P}}^2\delta - \alpha_{\mathcal{P}}^2\delta^2\right) \|A\|_F \leq \|\mathcal{X}(A)\|_1/n \leq \left(1 + 2\alpha_{\mathcal{P}}^2\delta + \alpha_{\mathcal{P}}^2\delta^2\right) \|A\|_F$$

with probability at least $1 - 2\exp(-\delta^2 n)$. Here $\alpha_{\mathcal{P}}$ is defined by (3.2).

LEMMA 7.2. *Suppose $A \in \mathbb{R}^{p_1 \times p_2}$ is a fixed matrix. $\beta = (\beta_1, \cdots, \beta_{p_1})^T, \gamma = (\gamma_1, \cdots, \gamma_{p_2})^T$ are random vectors such that $\beta_1, \cdots, \beta_{p_1}, \gamma_1, \cdots, \gamma_{p_2} \overset{iid}{\sim} \mathcal{P}$, where $\mathcal{P}$ is some symmetric variance 1 sub-Gaussian distribution, then we have*

$$\frac{\|A\|_F}{3\alpha_{\mathcal{P}}^4} \leq E\left|\beta^T A \gamma\right| \leq \|A\|_F$$

where $\alpha_{\mathcal{P}}$ is given by (3.2).

Let $z \in \mathbb{R}^n$ be i.i.d sub-Gaussian distributed. By measure concentration theory, $\|z\|_p^p/n$, $1 \leq p \leq \infty$, are essentially bounded; Specifically, we have the following lemma.

LEMMA 7.3. *Suppose $z \in \mathbb{R}^n$ and $z_i \overset{iid}{\sim} N(0, \sigma^2)$, we have*

$$P(\|z\|_1 \geq \sigma n) \leq \frac{9}{n}$$

$$P\left(\|z\|_2 \geq \sigma\sqrt{n + 2\sqrt{n \log n}}\right) \leq \frac{1}{n}$$

$$P(\|z\|_\infty \geq 2\sigma\sqrt{\log n}) \leq \frac{1}{n\sqrt{2\pi \log n}}.$$

More general, when z_i are i.i.d. sub-Gaussian distributed such that (3.1) holds, then

$$P\left(\|z\|_1 \geq Cn\right) \leq \exp\left(-\frac{n(C - 2\sqrt{2\pi}\gamma)^2}{2\gamma^2}\right), \quad \forall C > 2\sqrt{2\pi}\gamma$$

$$P(\|z\|_2 \geq \sqrt{Cn}) \leq \exp\left(-\frac{n(C-4\gamma^2)^2}{8\gamma^2 C}\right), \quad \forall C > 4\gamma^2$$

$$P(\|z\|_\infty \geq C\gamma\sqrt{\log n}) \leq 2n^{-C^2/2-1}, \quad \forall C > 0$$

Lemma 7.4 below presents an upper bound for the spectral norm of $\mathcal{X}(z)$ for a fixed vector z.

LEMMA 7.4. *Suppose $\mathcal{X}$ is ROP from some symmetric sub-Gaussian distribution $\mathcal{P}$ and $z \in \mathbb{R}^n$ is some fixed vector, then for $C > \log 7$, we have*

$$\|\mathcal{X}^*(z)\| \leq 3\alpha_{\mathcal{P}}^2 \left(C(p_1+p_2)\|z\|_\infty + \sqrt{2C(p_1+p_2)}\|z\|_2\right)$$

with probability at least $1 - 2\exp\left(-(C-\log 7)(p_1+p_2)\right)$. Here $\alpha_{\mathcal{P}}$ is defined by (3.2).

We are now ready to prove the main results of the paper.

7.2. *Proof of Theorem 2.1.*. We introduce the following two technical lemmas that will be used in the proof of theorem.

The *null space property* below is a well-known result in affine rank minimization problem (See [31]). It provides a necessary, sufficient and easier-to-check condition for exact recovery in the noiseless setting.

LEMMA 7.5 (Null Space Property). *Using (2.1) one can recover all matrices A of rank at most r if and only if for all $R \in \mathcal{N}(\mathcal{X})\backslash\{0\}$,*

$$\|R_{\max(r)}\|_* < \|R_{-\max(r)}\|_*.$$

The following lemma is given in [10], which provides a way to decompose the general vectors to sparse ones.

LEMMA 7.6 (Sparse Representation of A Polytope). *Suppose s is a non-negative integer, $v \in \mathbb{R}^p$ and $\theta \geq 0$. Then $\|v\|_\infty \leq \theta, \|v\|_1 \leq s\theta$ if and only if v can be expressed as a weighted mean,*

$$v = \sum_{i=1}^N \lambda_i u_i, \quad 0 \leq \lambda_i \leq 1, \quad \sum_{i=1}^N \lambda_i = 1$$

where u_i satisfies:

$$u_i \text{ is } s\text{-sparse}, \quad \operatorname{supp}(u_i) \subseteq \operatorname{supp}(v), \quad \|u_i\|_1 = \|v\|_1, \quad \|u_i\|_\infty \leq \theta \tag{7.1}$$

For the proof of Theorem 2.1, by null space property (Lemma 7.5), we only need to show for all non-zero R with $\mathcal{X}(R) = 0$, we must have $\|R_{\max(r)}\|_* < \|R_{-\max(r)}\|_*$.

If this does not hold, suppose there exists non-zero R with $\mathcal{X}(R) = 0$ and $\|R_{\max(r)}\|_* \geq \|R_{-\max(r)}\|_*$. We denote $p = \min(p_1, p_2)$ and assume the singular value decomposition of R is

$$R = \sum_{i=1}^{p} \sigma_i u_i v_i^T = U \mathrm{diag}(\vec{\sigma}) V^T,$$

where u_i, v_i are orthogonal basis in $\mathbb{R}^{p_1}$, $\mathbb{R}^{p_2}$, respectively and $\vec{\sigma}$ is the singular value vector such that $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_p \geq 0$. Without loss of generality, we can assume $p \geq kr$, otherwise we can set the undefined entries of σ as 0.

Consider the singular value vector $\vec{\sigma} = (\sigma_1, \sigma_2, \cdots, \sigma_p)$, we note that $\vec{\sigma}_{-\max(kr)}$ satisfies

$$\|\vec{\sigma}_{-\max(kr)}\|_\infty \leq \sigma_{kr},$$

$$\begin{aligned}\|\vec{\sigma}_{-\max(kr)}\|_1 =& \|\vec{\sigma}_{-\max(r)}\|_1 - (\sigma_{r+1} + \cdots + \sigma_{kr}) \\ \leq& \|\vec{\sigma}_{-\max(r)}\|_1 - (k-1)r\sigma_{kr} \leq \|\vec{\sigma}_{\max(r)}\|_1 - (k-1)r\sigma_{kr}.\end{aligned}$$

Denote $\theta = \max\left\{\sigma_{kr}, (\|\vec{\sigma}_{\max(r)}\|_1 - r(k-1)\sigma_{kr})/(kr)\right\}$, by the two inequalities above we have $\|\vec{\sigma}_{-\max(kr)}\|_\infty \leq \theta$ and $\|\vec{\sigma}_{-\max(kr)}\|_1 \leq kr\theta$. Now apply Lemma 7.6, we can get $b^{(i)} \in \mathbb{R}^p, \lambda_i \geq 0, i = 1, \cdots, N$ such that $\sum_{i=1}^N \lambda_i = 1$, $\vec{\sigma}_{-\max(kr)} = \sum_{i=1}^N \lambda_i b^{(i)}$ and

$$\begin{aligned}&\mathrm{supp}(b^{(i)}) \subseteq \mathrm{supp}(\vec{\sigma}_{-\max(kr)}), \quad \|b^{(i)}\|_0 \leq kr, \\ &\|b^{(i)}\|_1 = \|\vec{\sigma}_{-\max(kr)}\|_1, \quad \|b^{(i)}\|_\infty \leq \theta.\end{aligned} \tag{7.2}$$

which leads to

$$\|b^{(i)}\|_2 \leq \sqrt{\|b^{(i)}\|_1 \cdot \|b^{(i)}\|_\infty} \leq \sqrt{(\|\vec{\sigma}_{\max(r)}\|_1 - r(k-1)\sigma_{kr}) \cdot \theta}$$

If $\theta = \sigma_{kr}$, we have

$$\begin{aligned}\|b^{(i)}\|_2 &\leq \sqrt{(\|\vec{\sigma}_{\max(r)}\|_1 - r(k-1)\sigma_{kr})\sigma_{kr}} \\ &\leq \sqrt{\left(\|\vec{\sigma}_{\max(r)}\|_1 - r(k-1)\frac{\|\vec{\sigma}_{\max(r)}\|_1}{2r(k-1)}\right)\frac{\|\vec{\sigma}_{\max(r)}\|_1}{2r(k-1)}} \\ &\leq \frac{\|\vec{\sigma}_{\max(r)}\|_1}{\sqrt{4r(k-1)}} \leq \frac{\|\vec{\sigma}_{\max(r)}\|_2}{\sqrt{4(k-1)}}\end{aligned}$$

If $\theta = (\|\vec{\sigma}_{\max(r)}\|_1 - r(k-1)\sigma_{kr})/(kr)$, we have

$$\|b^{(i)}\|_2 \le \sqrt{\frac{1}{kr}}(\|\vec{\sigma}_{\max(r)}\|_1 - r(k-1)\sigma_{kr}) \le \sqrt{\frac{1}{kr}}\|\vec{\sigma}_{\max(r)}\|_1 \le \frac{\|\vec{\sigma}_{\max(r)}\|_2}{\sqrt{k}}$$

Since $k \ge 2$, we always have $\|b^{(i)}\|_2 \le \|\vec{\sigma}_{\max(r)}\|_2/\sqrt{k}$. Finally, we define $B_i = U\mathrm{diag}(b^{(i)})V^T$, then the rank of B_i are all at most kr and $\sum_{i=1}^N \lambda_i B_i = R_{-\max(kr)}$ and

$$\|B_i\|_F = \|b^{(i)}\|_2 \le \|\vec{\sigma}_{\max(r)}\|_2/\sqrt{k} = \|R_{\max(r)}\|_F/\sqrt{k}$$

Hence,

$$\begin{aligned}
0 = \|\mathcal{X}(R)\|_1 &\ge \|\mathcal{X}(R_{\max(kr)})\|_1 - \|\mathcal{X}(R_{-\max(kr)})\|_1 \\
&\ge C_1\|R_{\max(kr)}\|_F - \sum_{i=1}^N \|\mathcal{X}(\lambda_i B_i)\|_1 \\
&\ge C_1\|R_{\max(r)}\|_F - \sum_{i=1}^N \lambda_i C_2\|B_i\|_F \\
&\ge C_1\|R_{\max(r)}\|_F - C_2\|R_{\max(r)}\|_F/\sqrt{k} > 0
\end{aligned}$$

Here we used the RUB condition. The last inequality is due to $C_2/C_1 < \sqrt{k}$ and $R \ne 0$ (so $R_{\max(r)} \ne 0$). This is a contradiction, which finishes the proof of the theorem.

7.3. *Proof of Theorem 2.2.*. Notice that for $\mathcal{P}$ as standard Gaussian distribution, the constant $\alpha_{\mathcal{P}}$ (defined as (3.2)) equals 1. We will prove the following more general result than Theorem 2.2 instead. The proof is provided in the Appendix.

PROPOSITION 7.1. *Suppose $\mathcal{X} : \mathbb{R}^{p_1\times p_2} \to \mathbb{R}^n$ is ROP from some variance 1 symmetric sub-Gaussian distribution $\mathcal{P}$. For integer $k \ge 2$, positive $C_1 < \frac{1}{3\alpha_{\mathcal{P}}^4}$ ($\alpha_{\mathcal{P}}$ is defined as (3.2)) and $C_2 > 1$, there exists constants C and δ, only depending on $\mathcal{P}, C_1, C_2$ but not on p_1, p_2, r, such that if $n \ge Cr(p_1+p_2)$, then with probability at least $1 - e^{-n\delta}$, $\mathcal{X}$ satisfies RUB of order kr and constants C_1 and C_2.*

7.4. *Proof of Theorem 2.3, Proposition 2.1 and Theorem 3.1.*. In order to prove the result, we introduce the following technical lemma as an extension of Null Space Property (Lemma 7.5) from exact low-rank into the approximate low-rank setting.

LEMMA 7.7. *Suppose $A_*, A \in \mathbb{R}^{p_1 \times p_2}$, $R = A_* - A$. If $\|A_*\|_* \leq \|A\|_*$, we have*

$$\|R_{-\max(r)}\|_* \leq \|R_{\max(r)}\|_* + 2\|A_{-\max(r)}\|_* \tag{7.3}$$

The following two lemmas described the separate effect of constraint $\mathcal{Z}_1 = \{z : \|z\|_1/n \leq \lambda_1\}$ and $\mathcal{Z}_2 = \{z : \|\mathcal{X}^*(z)\| \leq \lambda_2\}$ on the estimator.

LEMMA 7.8. *Suppose $\mathcal{X}$ satisfies RUB condition of order kr with constants C_1, C_2 such that $C_1 > C_2/\sqrt{k}$. Assume that $A_*, A \in \mathbb{R}^{p_1 \times p_2}$ satisfy $\|A_*\|_* \leq \|A\|_*$, $\|\mathcal{X}(A_* - A)\|_1/n \leq \lambda_1$. Then we have*

$$\|A_* - A\|_F \leq \frac{2}{C_1 - C_2/\sqrt{k}}\lambda_1 + \left(\frac{3}{\sqrt{k}C_1/C_2 - 1} + \frac{1}{\sqrt{k-1}}\right)\frac{\|A_{-\max(r)}\|_*}{\sqrt{r}}$$

LEMMA 7.9. *Suppose $\mathcal{X}$ satisfies RUB condition of order kr with constants C_1, C_2 such that $C_1 > C_2/\sqrt{k}$. Assume that $\hat{A}^{DS}$ satisfies $\|\mathcal{X}^*\mathcal{X}(A_* - A)\| \leq \lambda_2$. Then we have*

$$\|A_* - A\|_F \leq \frac{4}{(C_1 - C_2/\sqrt{k})^2} \cdot \frac{\sqrt{r}\lambda_2}{n} + \left(\frac{5}{\sqrt{k}C_1/C_2 - 1} + \frac{1}{\sqrt{k-1}} + 1\right)\frac{\|A_{-\max(r)}\|_*}{\sqrt{r}}$$

The proof of Lemma 7.7, 7.8 and 7.9 are listed in the Appendix. Now we prove Theorem 2.3 and Proposition 2.1. We only need to prove Proposition 2.1 since Theorem 2.3 is a special case of Proposition 2.1. By Lemma 7.3 and Lemma 7.4, we have

$$P_z(\|z\|_1 \leq \sigma n) \leq \frac{9}{n},$$

$$\begin{aligned}
&P_{\mathcal{X},z}\left(\|\mathcal{X}^*(z)\| \geq \sigma\left(12(p_1+p_2)\sqrt{\log n} + 6\sqrt{2(p_1+p_2)n}\right)\right) \\
\leq &P_{\mathcal{X}}\left(\|\mathcal{X}^*(z)\| \geq \left(6(p_1+p_2)\|z\|_\infty + 6\sqrt{p_1+p_2}\|z\|_2\right)\right) \\
&+ P_z\left(\|z\|_\infty \geq 2\sigma\sqrt{\log n}\right) + P_z\left(\|z\|_2 \geq \sigma\sqrt{2n}\right) \\
\leq &2\exp(-(2-\log 7)(p_1+p_2)) + \frac{1}{n\sqrt{2\pi\log n}} + \frac{1}{n}
\end{aligned}$$

Here $P_{\mathcal{X}}$ (P_z or $P_{\mathcal{X},z}$) means the probability with respect to $\mathcal{X}$ (z or $(\mathcal{X}, z)$). Hence, we have

$$P(z \in \mathcal{Z}_1 \cap \mathcal{Z}_2) \geq 1 - 2\exp(-(2-\log 7)(p_1+p_2)) - \frac{11}{n}.$$

Under the event that $z \in \mathcal{Z}_1 \cap \mathcal{Z}_2$, A is in the feasible set of the programming (2.6), which implies $\|\hat{A}\|_* \leq \|A\|_*$ by the definition of $\hat{A}$. Moreover, we have

$$\begin{aligned}\|\mathcal{X}(\hat{A}-A)\|_1/n \leq & \|y-\mathcal{X}(A)\|_1/n + \|y-\mathcal{X}(\hat{A})\|_1/n \\ \leq & \|z\|_1/n + \|y-\mathcal{X}(\hat{A})\|_1/n \leq 2\sigma\end{aligned}$$

$$\begin{aligned}\|\mathcal{X}^*\mathcal{X}(\hat{A}-A)\| \leq & \|\mathcal{X}^*(y-\mathcal{X}(\hat{A}))\| + \|\mathcal{X}^*(y-\mathcal{X}(A))\| \\ \leq & \|\mathcal{X}^*(y-\mathcal{X}(\hat{A}))\| + \|\mathcal{X}^*(z)\| \leq 2\eta\end{aligned}$$

On the other hand, suppose $k = 10$, by Theorem 2.2, we can have find a uniform constant C and δ such that if $n \geq Crk(p_1+p_2)$, $\mathcal{X}$ satisfies RUB of order $10r$ and constants $C_1 = 0.32, C_2 = 1.02$ with probability at least $1-e^{-n\delta'}$. Hence, we have $D(= Ck)$ and δ' such that if $n \geq Dr(p_1+p_2)$, $\mathcal{X}$ satisfies RUB of order $10r$ and constants C_1, C_2 satisfying $C_2/C_1 < \sqrt{10}$ with probability at least $1-e^{-n\delta'}$.

Now under the event that

1. $\mathcal{X}$ satisfies RUB of order $10r$ and constants C_1, C_2 satisfying $C_2/C_1 < \sqrt{10}$,
2. $z \in \mathcal{Z}_1 \cap \mathcal{Z}_2$,

Apply Lemma 7.8 and Lemma 7.9 with $A_* = \hat{A}$, we can get (2.11). The probability that these two events both happen is at least $1 - 2\exp(-(2-\log 7)(p_1+p_2)) - \frac{11}{n} - \exp(-\delta' n)$. Set $\delta = \min(2-\log 7, \delta')$, we finished the proof of Proposition 2.1.

For Theorem 3.1, the proof is similar. We apply the latter part of Lemma 7.3 and Lemma 7.4 and get

$$\begin{aligned}&P(z \notin \mathcal{Z}_1 \cap \mathcal{Z}_2) \\ \leq & P(\|z\|_1/n > 6\tau) + P\left(\|\mathcal{X}(z)\| > \tau\alpha_{\mathcal{P}}^2\left(6\sqrt{6n(p_1+p_2)} + 12\sqrt{\log n}(p_1+p_2)\right)\right) \\ \leq & P(\|z\|/n > 6\tau) + P(\|z\|_2 > \sqrt{6n}\tau) + P(\|z\|_\infty > 2\sqrt{\log n}\tau) \\ & + P_{\mathcal{X}}(\|\mathcal{X}(z)\| > \alpha_{\mathcal{P}}^2(6(p_1+p_2)\|z\|_\infty + 6\sqrt{p_1+p_2}\|z\|_2)) \\ \leq & \exp\left(-n(6-2\sqrt{2\pi})^2/2\right) + \exp(-n/12) + \frac{2}{n} + 2\exp(-(2-\log 7)(p_1+p_2))\end{aligned}$$

Besides, we choose $k > (3\alpha_{\mathcal{P}}^4)^2$, then we can find $C_1 < 1/(3\alpha_{\mathcal{P}}^4)$ and $C_2 > 1$ such that $C_2/C_1 < \sqrt{k}$. Apply Proposition 7.1, there exists C, δ' only depending on $\mathcal{P}$, C_1, C_2 such that if $n \geq Ckr(p_1+p_2)$, $\mathcal{X}$ satisfies RUB of order kr with constants C_1 and C_2 with probability at least $1-\exp(-\delta'(p_1+p_2))$. Note that C_1, C_2 only depends on $\mathcal{P}$, we can conclude that there exist

constants $D(=Ck), \delta'$ only depending on $\mathcal{P}$ such that if $n \geq Dr(p_1+p_2)$, $\mathcal{X}$ satisfies RUB of order kr with constants C_1, C_2 satisfying $C_2/C_1 \leq \sqrt{k}$.

Similarly to the proof of Proposition 2.1, under the event that

1. $\mathcal{X}$ satisfies RUB of order kr and constants C_1, C_2 satisfying $C_2/C_1 < \sqrt{k}$;
2. $z \in \mathcal{Z}_1 \cap \mathcal{Z}_2$;

we can get (3.4) (we shall note that W_1 depends on $\mathcal{P}$, so its value can also depend on $\alpha_{\mathcal{P}}$). The probability that those events happen is at least $1 - 2/n - 5\exp(-\delta(p_1+p_2))$ for $\delta \leq \min((6-2\sqrt{2\pi})^2/2, 1/12, 2-\log 7, \delta')$. □

7.5. *Proof of Theorem 2.4.*. Without loss of generality, we assume that $p_1 \leq p_2$. We consider the class of rank-r matrices

$$\mathcal{F}_c = \{A \in \mathbb{R}^{p_1 \times p_2} : A_{ij} = 0, \text{ whenever } i \geq r+1\}$$

namely the matrices with all non-zeros entries in the first r rows. The model (1.1) become

$$y_i = \beta_{1:r}^{(i)T} A_r \gamma^{(i)} + z_i, \quad i = 1, \cdots, n$$

where $\beta_{1:r}^{(i)}$ is the vector of the first to the r-th entries of $\beta^{(i)}$. Note that this is a linear regression model with variable $A_r \in \mathbb{R}^{r \times p_2}$, by Lemma 3.11 in [12], we have

$$\inf_{\hat{A}} \sup_{A \in \mathcal{F}_c} E\|\hat{A}(y) - A\|_F^2 = \sigma^2 \text{trace}\left[(\mathcal{X}_r^* \mathcal{X}_r)^{-1}\right] \tag{7.4}$$

$$\inf_{\hat{A}} \sup_{A \in \mathcal{F}_c} E\|\hat{A}(y) - A\|_F^2 = \infty, \quad \text{when } \mathcal{X}_r^* \mathcal{X}_r \text{ is singular,} \tag{7.5}$$

where $\mathcal{X}_r : \mathbb{R}^{r \times p_2} \to \mathbb{R}^n$ is the $\mathcal{X}$ constrained on $\mathcal{F}_c$, Then $\mathcal{X}_r$ sends A_r to $\left(\beta_{1:r}^{(1)} A_r \gamma^{(1)}, \cdots, \beta_{1:r}^{(n)} A_r \gamma^{(n)}\right)^T$. When $n < p_2 r$, $\mathcal{X}_r$ is singular, hence we have (2.10).

When $n \geq p_2 r$, we can see in order to show (2.9), we only need to show $\text{trace}(\mathcal{X}_r^* \mathcal{X}_r) \geq \frac{p_2 r}{2n}$ with probablity at least $1 - 26n^{-1}$. Suppose the singular value of $\mathcal{X}_r$ are $\sigma_i(\mathcal{X}_r)$, $i = 1, \cdots, rp_2$, then $\text{trace}(\mathcal{X}_r^* \mathcal{X}_r) = \sum_{i=1}^{p_2 r} \sigma^{-2}(\mathcal{X}_r)$.

Suppose $\mathcal{X}$ is ROP while $B \in \mathbb{R}^{r \times p_2}$ is i.i.d. standard Gaussian random matrix (both $\mathcal{X}$ and B_r are random). Then by some calculation we can see

$$E_{B,\mathcal{X}_r}\|\mathcal{X}_r(B)\|_2^2 = nE_{B,\beta,\gamma}\left(\beta_{1:r}^T B \gamma\right)^2 = n\sum_{j=1}^{r}\sum_{k=1}^{p_2} E\left(\beta_j B_{jk} \gamma_k\right)^2 = np_2 r$$

Note (7.20) in the proof of Lemma 7.1, we know $E\left(\beta_{1:r}^{(i)T}B\gamma^{(i)}\|_2^4\middle|B\right) \leq 9\|B\|_F^4$. Hence,

$$\begin{aligned}
&E\|\mathcal{X}_r(B)\|_2^4\\
=&\sum_{i=1}^n E\left(\beta_{1:r}^{(i)T}B\gamma^{(i)}\right)^4 + 2\sum_{1\leq i<l\leq n} E\sum_{j=1}^n\left(\beta_{1:r}^{(i)T}B\gamma^{(i)}\right)^2\cdot E\sum_{j=1}^n\left(\beta_{1:r}^{(l)T}B\gamma^{(l)}\right)^2\\
=&n\cdot 9E\|B\|_F^4 + n(n-1)(p_2r)^2 = 9nE\left(\chi^2(p_2r)\right)^2 + n(n-1)p_2^2r^2\\
=&9n(p_2^2r^2+2p_2r)+n(n-1)p_2^2r^2\\
=&n^2p_2^2r^2+2np_2r(4p_2r+9)\leq n^2p_2^2r^2+26np_2^2r_2^2
\end{aligned}$$

Besides,

$$E\|\mathcal{X}_r(B_r)\|_2^2 = E\left(E\left(\|\mathcal{X}_r(B_r)\|_2^2\middle|\mathcal{X}_r\right)\right) = E\left(\sum_{i=1}^{rp_2}\sigma_i^2(\mathcal{X}_r)\right)$$

$$\begin{aligned}
&E\|\mathcal{X}_r(B_r)\|_2^4 = E\left(E\left(\|\mathcal{X}_r(B_r)\|_2^4|\mathcal{X}_r\right)\right)\\
=&E\left(\sum_{i=1}^{rp_2}3\sigma_i^4(\mathcal{X}_r)+2\sum_{1\leq i<j\leq rp_2}\sigma_i^2(\mathcal{X}_r)\sigma_j^2(\mathcal{X}_r)\right)\geq E\left(\sum_{i=1}^{rp_2}\sigma_i^2(\mathcal{X}_r)^2\right)^2
\end{aligned}$$

Hence,

$$E\left(\sum_{i=1}^{rp_2}\sigma_i^2(\mathcal{X}_r)^2\right) = np_2r$$

$$\mathrm{Var}\left(\sum_{i=1}^{rp_2}\sigma_i^2(\mathcal{X}_r)^2\right) = E\left(\sum_{i=1}^{rp_2}\sigma_i^2(\mathcal{X}_r)^2\right)^2 - \left(E\sum_{i=1}^{rp_2}\sigma_i^2(\mathcal{X}_r)^2\right)^2 \leq 26np_2^2r^2$$

Then by Chebyshev's inequality, we have

$$\sum_{i=1}^{rp_2}\sigma_i^2(\mathcal{X}_r)\leq 2np_2r \tag{7.6}$$

with probability at least $1-\frac{26np_2^2r^2}{(npr)^2} = 1-\frac{26}{n}$. By Cauchy-Schwarz's inequality, we have

$$\mathrm{trace}\left((\mathcal{X}_r^*\mathcal{X}_r)^{-1}\right) = \sum_{i=1}^{rp_2}\sigma_i^{-2}(\mathcal{X}_r)\geq\frac{(p_2r)^2}{\sum_{i=1}^{rp_2}\sigma_i^2(\mathcal{X}_r)}$$

Therefore, we have

$$\mathrm{trace}\left((\mathcal{X}_r^*\mathcal{X}_r)^{-1}\right)\geq\frac{p_2r}{2n}$$

with probability at least $1-26/n$, which shows (2.9).

Finally we consider (2.8). Suppose inequality (7.6) holds, then

$$\left|\{i:\sigma_i^2(X_r)\geq 4n\}\right|\leq \frac{p_2 r}{2}$$

$$\Rightarrow \left|\{i:\sigma_i^{-2}(X_r)\leq \frac{1}{4n}\}\right|\geq \frac{p_2 r}{2}$$

$$\Rightarrow \left|\{i:\sigma_i^{-2}(X_r)\geq \frac{1}{4n}\}\right|\geq \frac{p_2 r}{2} \tag{7.7}$$

By Lemma 3.12 in [12], we know

$$\begin{aligned}
&\inf_{\hat{A}}\sup_{A\in\mathcal{F}_c} P_z\left(\|\hat{A}-A\|_F^2\geq \frac{p_2 r\sigma^2}{16n}\right)\\
=&\inf_{\hat{A}}\sup_{A\in\mathcal{F}_c} E_z 1_{\{x\geq \frac{p_2 r\sigma^2}{16n}\}}(\|\hat{A}-A\|_F^2)\\
=&E_z 1_{\{x\geq \frac{p_2 r\sigma^2}{16n}\}}(\|(\mathcal{X}_r^*\mathcal{X}_r)^{-1}\mathcal{X}_r^*(z)\|_F^2)\\
=&P_z\left(\|(\mathcal{X}_r^*\mathcal{X}_r)^{-1}\mathcal{X}_r^*(z)\|_F^2\geq \frac{p_2 r\sigma^2}{16n}\right)
\end{aligned}$$

where $1_{\{x\geq \frac{p_2 r\sigma^2}{16n}\}}(\cdot)$ is the indicator function. Note that when $z\overset{iid}{\sim}N(0,\sigma^2)$, $\|(\mathcal{X}_r^*\mathcal{X}_r)^{-1}\mathcal{X}_r^*(z)\|_F^2$ is identical distributed as $\sum_{i=1}^{rp_2}\frac{y_i^2}{\sigma_i^2(\mathcal{X}_r)}$, where $y_1,\cdots,y_{rp_2}\overset{iid}{\sim}N(0,\sigma^2)$, hence,

$$\begin{aligned}
&P\left(\|(\mathcal{X}_r^*\mathcal{X}_r)^{-1}\mathcal{X}_r^*(z)\|_F^2\leq \frac{p_2 r\sigma^2}{16n}\right)\\
=&P\left(\sum_{i=1}^{rp_2}\frac{y_i^2}{\sigma_i^2(\mathcal{X}_r)}\leq \frac{p_2 r\sigma^2}{16n}\right)\\
\leq&P\left(\sum_{i:\sigma_i^{-2}(\mathcal{X}_r)\geq 1/(4n)} y_i^2\sigma_i^{-2}(\mathcal{X}_r)\leq \frac{p_2 r\sigma^2}{16n}\right)\\
\leq&P\left(\sum_{i:\sigma_i^{-2}(\mathcal{X}_r)\geq 1/(4n)} \frac{y_i^2}{4n}\leq \frac{p_2 r\sigma^2}{16n}\right)\leq P\left(\chi^2(\lceil\frac{rp_2}{2}\rceil)\leq \frac{p_2 r}{4}\right)\\
\leq&\exp\left(-\frac{rp_2}{32}\right).
\end{aligned}$$

The last inequality is due to the tail bound of χ^2 distribution given by Lemma 1 in [27]; the second last inequality is due to (7.7). In summary, when (7.6) holds, we have

$$\inf_{\hat{A}} \sup_{A\in\mathcal{F}_c} P_z\left(\|\hat{A}-A\|_F^2 \geq \frac{p_2 r\sigma^2}{16n}\right) \leq \exp(-\frac{rp_2}{32})$$

Finally since $p_2 \geq (p_1+p_2)/2$, we showed that with probability at least $1-26n^{-1}$, $\mathcal{X}$ satisfies (2.8). □

7.6. *Proof of Theorem 4.1.*. We first introduce the following lemma about the upper bound of $\|z\|_1, \|z\|_2, \|z\|_\infty$.

LEMMA 7.10. *Suppose z is defined as (4.3), then for constants $C_1 > \sqrt{2}$, $M_1 > 1$, we have*

$$\begin{aligned} &P\left(\|z\|_1/n \leq \frac{C_1}{n}\sum_{i=1}^n \xi_i^2\right) \geq 1-\frac{9C_1^2+6}{n(C_1-\sqrt{2})^2}, \\ &P\left(\frac{C_1}{n}\sum_{i=1}^n \xi_i^2 \leq M_1 C_1\|\Sigma\|_*\right) \geq 1-\frac{9}{n(M_1-1)^2}; \end{aligned} \tag{7.8}$$

for constants $C_2 > 1$, $M_2 > 9$,

$$\begin{aligned} &P\left(\|z\|_2^2/n \leq \frac{C_2^2\sum_{i=1}^n \xi_i^4}{n}\right) \geq 1-\frac{105\left(105C_2^4+60\right)}{n(3C_2^2-2)^2}, \\ &P\left(\frac{C_2^2\sum_{i=1}^n \xi_i^4}{n} \leq M_2 C_2^2\|\Sigma\|_*^2\right) \geq 1-\frac{105^2}{n(M_1-9)^2}; \end{aligned} \tag{7.9}$$

for constants $C_3 > 1$, $M_3 > 1$,

$$\begin{aligned} &P\left(\|z\|_\infty \leq C_3\log n \max_{1\leq i\leq n}\xi_i^2\right) \geq 1-\frac{2}{\sqrt{2\pi C_3\log n}}, \\ &P\left(C_3\log n\max_{1\leq i\leq n}\xi_i^2 \leq 2C_3M_3\log^2 n\left(\sqrt{\|\Sigma\|_*}+\sqrt{2M_3\log n\|\Sigma\|}\right)^2\right) \\ \geq& 1-2n^{-M_3+1}. \end{aligned} \tag{7.10}$$

The proof of Lemma 7.10 is listed in the Appendix. The rest of the proof is basically the same as Proposition 2.3. Suppose $\mathcal{X}_1, \mathcal{X}_2$ and $\tilde{z}$ are given by (7.36), (7.37) and (7.39), then $\mathcal{X}_1$, $\mathcal{X}_2$ are ROP. By Lemma 7.4,

$$\|\mathcal{X}_1^*(\tilde{z})\| \leq 6\left(2p\|\tilde{z}\|_\infty + \sqrt{2p}\|\tilde{z}\|_2\right) \tag{7.11}$$

$$\|\mathcal{X}_2^*(\tilde{z})\| \le 6\left(2p\|\tilde{z}\|_\infty + \sqrt{2p}\|\tilde{z}\|_2\right) \tag{7.12}$$

with probability at least $1-4\exp\left(-2(2-\log 7)p\right)$. Hence there exists $\delta > 0$ such that,

$$\begin{aligned}
&P\left(\Sigma_0 \text{ is NOT in the feasible set of (4.4)}\right) = P\left(\|z\|_1/n > \eta_1, \text{ or } \|\tilde{\mathcal{X}}^*(\tilde{z})\| > \eta_2\right)\\
\le& P\left(\|z\|_1/n > \frac{c_1}{n}\sum_{i=1}^n \xi_i^2\right) + P\left(\|\tilde{z}\|_\infty > 2c_3\log n \max_{1\le i\le n}\xi_i^2\right) + P\left(\|\tilde{z}\|_2 > c_2\sqrt{2\sum_{i=1}^n \xi_i^4}\right)\\
&+ P\left(\|\tilde{\mathcal{X}}^*(z)\| > 24p\|\tilde{z}\|_\infty + 12\sqrt{2p}\|\tilde{z}\|_2\right)\\
\le& P\left(\|z\|_1/n > \frac{c_1}{n}\sum_{i=1}^n \xi_i^2\right) + P\left(\|z\|_\infty > c_3\log n \max_{1\le i\le n}\xi_i^2\right) + P\left(\|z\|_2 > c_2\sqrt{\sum_{i=1}^n \xi_i^4}\right)\\
&+ P\left(\|\mathcal{X}_1^*(\tilde{z})\| > 12p\|\tilde{z}\|_\infty + 6\sqrt{2p}\|\tilde{z}\|_2\right) + P\left(\|\mathcal{X}_2^*(z)\| > 12p\|\tilde{z}\|_\infty + 6\sqrt{2p}\|\tilde{z}\|_2\right)\\
\le& O(1/n) + 4\exp\left(-2(2-\log 7)p\right) + \frac{2}{\sqrt{2\pi c_3\log n}}.
\end{aligned}$$

Here we used the fact that $\tilde{\mathcal{X}}^* = \mathcal{X}_1^* + \mathcal{X}_2^*$,

$$\|\tilde{z}\|_2 = \sqrt{\sum_{i=1}^{\lfloor n/2\rfloor}(z_{2i-1}-z_{2i})^2} \le \sqrt{\sum_{i=1}^{\lfloor n/2\rfloor} 2(z_{2i-1}^2 + z_{2i}^2)} \le \sqrt{2}\|z\|_2,$$

$$\|\tilde{z}\|_\infty = \max_i |z_{2i-1} - z_{2i}| \le 2\max_i |z_i| \le 2\|z\|_\infty.$$

Similarly to the proof of Proposition 2.3, since $\mathcal{X}_1$ is ROP, there exists constants D and δ' such that if $n \ge Drp$, $\mathcal{X}_1$ satisfies RUB of order $10k$ with constants C_1, C_2 satisfying $C_2/C_1 < \sqrt{10}$ with probability at least $1-e^{-n\delta'}$.

Now under the event that

1. A is feasible in (4.4),
2. $\mathcal{X}_1$ satisfies RUB of order $10k$ with constants C_1, C_2 satisfying $C_2/C_1 < \sqrt{10}$,
3. The latter part of (7.8), (7.9) and (7.10) hold for some $M_1 > 1, M_2 > 9, M_3 > 2$,

we can prove (4.6) similarly as the proof of Proposition 2.3, which we omit the proof here. □

Acknowledgments. We thank the Associate Editor and the referees for their thorough and useful comments which have helped to improve the presentation of the paper.

References.

[1] Alquier, P., Butucea, C., Hebiri, M., and Meziani, K. (2012). Rank penalized estimation of a quantum system. *Physical Review A*, to appear.

[2] Andrews, H. C. and Patterson, C. L., III (1976). Singular value decomposition (SVD) image coding. *IEEE Trans. Commun.* **24** 425-432.

[3] Basri, R. and Jacobs, D. W. (2003). Lambertian reflectance and linear sub-spaces. *IEEE Trans. Pattern Anal. Mach. Intell.* **25** 218-233.

[4] Birnbaum, A., Johnstone, I.M., Nadler, B., and Paul, D. (2013). Minimax bounds for sparse PCA with noisy high-dimensional data. *Ann. Statist.* **41** 1055-1084.

[5] Cai, T., Ma, Z. and Wu, Y. (2013). Sparse PCA: Optimal rates and adaptive estimation. *Ann. Statist.* **41** 3074-3110

[6] Cai, T., Ma, Z. and Wu, Y. (2013). Optimal estimation and rank detection for sparse spiked covariance matrices. *Probability Theory and Related Fields*, to appear.

[7] Cai, T., Xu. G. and Zhang, J. (2009). On recovery of sparse signal via l1 minimization. *IEEE Trans. Inf. Theory* **55** 3388-3397.

[8] Cai, T. and Zhang, A. (2013). Sharp RIP bound for sparse signal and low-rank matrix recovery. *Appl. Comput. Harmon. Anal.* **35** 74-93.

[9] Cai, T. and Zhang, A. (2013). Compressed sensing and affine rank minimization under restricted isometry. *IEEE Trans. Signal Process.* **61** 3279-3290.

[10] Cai, T. and Zhang, A. (2014). Sparse Representation of a Polytope and Recovery of Sparse Signals and Low-rank Matrices. *IEEE Trans. Inf. Theory* **60** 122-132.

[11] Candès, E., Li, X., Ma, Y. and Wright, J. (2011). Robust principal component analysis. *J. ACM* **58** 1-37.

[12] Candès, E. and Plan, Y. (2011). Tight oracle for low-rank matrix recovery from a minimal number of random measurements. *IEEE Trans. Inf. Theory* **57** 2342-2359.

[13] Candès, E. and Recht, B. (2009). Exact matrix completion via convex optimization. *Foundations of Computational Mathematics* **9** 717-772.

[14] Candès, E., Strohmer, T. and Voroninski, V. (2013). Phaselift: Exact and stable signal recovery from magnitude measurements via convex programming. *Communications on Pure and Applied Mathematics* **66** 1241-1274.

[15] Candès, E. and Tao, T. (2010). The power of convex relaxation: Near-optimal matrix completion, *IEEE Trans. Inform. Theory* **56** 2053 - 2080.

[16] Chen, Y., Chi, Y. and Goldsmith, A. (2013) Exact and Stable Covariance Estimation from Quadratic Sampling via Convex Programming. *arxiv preprint arXiv:1310.0807*.

[17] Dasarathy, G., Shah, P., Bhaskar, B. N., and Nowak, R. (2012). Covariance sketching. *Communication, Control, and Computing (Allerton), 2012 50th Annual Allerton Conference on. IEEE*.

[18] Dasarathy, G., Shah, P., Bhaskar, B. N., and Nowak, R. (2013). Sketching Sparse Matrices. *arXiv preprint arXiv:1303.6544*.

[19] Dvijotham, K., and Fazel, M. (2010). A nullspace analysis of the nuclear norm heuristic for rank minimization. In *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*. 3586-3589.

[20] Fan, J., Fan, Y., and Lv, J. (2008). High dimensional covariance matrix estimation using a factor model. *Journal of Econometrics* **147** 186-197.

[21] Grant, M. and Boyd, S. (2012). CVX: Matlab software for disciplined convex programming, version 2.0 beta. http://cvxr.com/cvx.

[22] Grant, M. and Boyd, S. (2008). Graph implementations for nonsmooth convex programs, Recent Advances in Learning and Control (a tribute to M. Vidyasagar). (V. Blondel, et al eds.), *Lecture Notes in Control and Information Sciences*, Springer,

95-110.

[23] Gross, D., Liu, Y. K., Flammia, S. T., Becker, S., and Eisert, J. (2010). Quantum state tomography via compressed sensing. *Physical review letters* **105** 150401-150404.

[24] Johnstone, I. M. (2001). On the distribution of the largest eigenvalue in principal component analysis. *Ann. Statist.* **29** 295-327.

[25] Koltchinskii, V., Lounici, K. and Tsybakov, A. B. (2011). Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *Ann. Statist.* **39** 2302-2329.

[26] Koren, Y., Bell, R., and Volinsky, C. (2009). Matrix factorization techniques for recommender systems. *Computer* **42** 30-37.

[27] Laurent, B. and Massart, P. (2000). Adaptive Estimation of a Quadratic Functional by Model Selection. *Ann. Statist.* **28** 1302-1338.

[28] Nadler, B. (2010). Nonparametric detection of signals by information theoretic criteria: performance analysis and an improved estimator. *IEEE Trans. Signal Process.* **58** 2746-2756.

[29] Negahban, S. and Wainwright, M. J. (2011). Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *Ann. Statist.* **39** 1069-1097.

[30] Oymak, S. and Hassibi, B. (2010). New Null Space Results and Recovery Thresholds for Matrix Rank Minimization. *arXiv preprint arXiv:1011.6326.*

[31] Oymak, S., Mohan, K., Fazel, M. and Hassibi, B. (2011). A simplified approach to recovery conditions for low-rank matrices. *Proc. Intl. Sympo. Information Theory (ISIT)* 2318-2322.

[32] Patterson, N., Price, A. L., and Reich, D. (2006). Population structure and eigenanalysis. *PLoS genetics* **2** e190.

[33] Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics* **38** 904-909.

[34] Recht, B. (2011). A Simpler Approach to Matrix Completion. *Journal of Machine Learning Research* **12** 3413-3430.

[35] Recht, B., Fazel, M. and Parrilo, P. (2010). Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Review* **52** 471-501.

[36] Rohde, A. and Tsybakov, A. B. (2011). Estimation of high-dimensional low-rank matrices. *Ann. Statist.* **39** 887-930.

[37] Trosset, M. W. (2000). Distance matrix completion by numerical optimization. *Comput. Optim. Appl.* **17** 11-22.

[38] Vershynin, R. (2011). Spectral norm of products of random and deterministic matrices. *Probability Theory and Related Fields* **150** 471-509.

[39] Wakin, M., Laska, J., Duarte, M., Baron, D., Sarvotham, S., Takhar, D., Kelly, K. and Baraniuk, R. (2006). An architecture for compressive imaging. *Proceedings of the International Conference on Image Processing (ICIP 2006)* 1273-1276.

[40] Wang H. and Li, S. (2010). Restricted Isometry Constants for Nuclear Norm Minimization. Preprint.

[41] Wang, Y. (2013). Asymptotic Equivalence of Quantum State Tomography and Trace Regression. *Ann. Statist.* **41** 2462-2504.

[42] Wax, M. and Kailath, T. (1985). Detection of signals by information theoretic criteria. *IEEE Transactions on Acoustics, Speech and Signal Processing* **33** 387-392.

Appendix: Proof of Technical Lemmas. We prove in this appendix the technical lemmas used in the proofs of the main results.

Proof of Lemma 2.1. Suppose $\mathcal{X} : \mathbb{R}^{p_1 \times p_2} \to \mathbb{R}^n$ is given by (1.6), we consider rank-1 matrices:

$$A_1 = e_1^{(p_1)} e_1^{(p_2)T}, \quad A_2 = \frac{\beta^{(1)}\gamma^{(1)}}{\|\beta^{(1)}\|_2 \|\gamma^{(1)T}\|_2}$$

Here $e_1^{(p_1)}, e_2^{(p_2)}$ are the p_1- and p_2- dimensional vectors with first entry 1 and others 0, respectively. Then we have $\|A_1\|_F = \|A_2\|_F = 1$.

$$E\|\mathcal{X}(A_1)\|_2^2 = E\sum_{i=1}^n (\beta_1^{(i)})^2 (\gamma_1^{(i)})^2 = n$$

$$\mathrm{Var}\|\mathcal{X}(A_1)\|_2^2 = \mathrm{Var}\sum_{i=1}^n (\beta_1^{(i)})^2 (\gamma_1^{(i)})^2 = n\mathrm{Var}\left(\beta_1^2\gamma_1^2\right) = 8n.$$

By Chebyshev's inequality, we have for all $t > 1$,

$$P\left(\|\mathcal{X}(A_1)\|_2^2 \geq tn\right) \leq \frac{8}{n(t-1)^2} \tag{7.13}$$

On the other hand,

$$\begin{aligned}\|\mathcal{X}(A_2)\|_2^2 =& \|\beta^{(1)}\|_2^2 \|\gamma^{(1)}\|_2^2 + \sum_{i=2}^n \left(\frac{\beta^{(i)T}\beta^{(1)}}{\|\beta^{(1)}\|_2}\right)^2 \left(\frac{\gamma^{(i)T}\gamma^{(1)}}{\|\gamma^{(1)}\|_2}\right)^2 \\ \geq& \|\beta^{(1)}\|_2^2 \|\gamma^{(1)}\|_2^2 \sim \chi^2(p_1)\cdot\chi^2(p_2)\end{aligned}$$

By Lemma 1 in [27], we know

$$P\left(\chi^2(p_1) \geq p_1/2\right) \leq \exp(-p_1/4), \quad P\left(\chi^2(p_2) \geq p_2/2\right) \leq \exp(-p_2/4)$$

Hence,

$$P\left(\|\mathcal{X}(A_2)\|_2^2 \geq p_1p_2/4\right) \leq \exp(-p_1/4) + \exp(-p_2/4). \tag{7.14}$$

Combining (7.13) and (7.14), we can see

$$C_2/C_1 \geq \frac{\|\mathcal{X}(A_2)\|_2/(\sqrt{n}\|A_2\|_F)}{\|\mathcal{X}(A_1)\|_2/(\sqrt{n}\|A_1\|_F)} = \sqrt{\frac{\|\mathcal{X}(A_2)\|_2^2}{\|\mathcal{X}(A_1)\|_2^2}} \geq \sqrt{\frac{p_1p_2/4}{tn}} \tag{7.15}$$

holds with probability at least $1 - e^{-p_1/4} - e^{-p_2/4} - \frac{8}{n(t-1)^2}$. □

Proof of Lemma 3.1. First, the common used definition for sub-Gaussian distribution of random variable X include the following two.

$$\exists c, C > 0, \text{ such that } P(|X| \geq t) \leq C\exp(-ct^2) \tag{7.16}$$

$$\exists c > 0, \text{ such that } Ee^{tX} \leq \exp(c^2t^2/2) \tag{7.17}$$

Suppose $\alpha_{\mathcal{P}}$ is finite, then we have

$$Ee^{tX} = \sum_{k=0}^{\infty} \frac{t^{2k}}{(2k)!} EX^{2k} \leq \sum_{k=0}^{\infty} \frac{t^{2k}}{(2k)!} \alpha^{2k}(2k-1)!! = \sum_{k=0}^{\infty} \frac{(t\alpha)^{2k}}{2^k k!} = \exp((\alpha t)^2/2)$$

namely X is sub-Gaussian. Now suppose X is sub-Gaussian, then

$$\begin{aligned} EX^{2k} &= (2k)\int_0^{\infty} P(|X| > t)t^{2k-1}dt \leq 2kC\int_0^{\infty} t^{2k-1}e^{-ct^2}dt \\ &= \frac{kC}{c^k}\int_0^{\infty}(ct^2)^{k-1}e^{-ct^2}d(ct^2) = \frac{k!C}{c^k} \leq \left(\frac{\max(C,1)}{c}\right)^k (2k-1)!! \end{aligned}$$

which implies that $\alpha_{\mathcal{P}} \leq \sqrt{\max(C,1)/c}$ is finite. □

Proof of Lemma 7.1. Without confusion, we simply use α to represent $\alpha_{\mathcal{P}}$ in the proof. Note that we can multiply A by a scale without loss of generality. So we assume throughout the proof that $\|A\|_F = 1$. We'll prove this lemma by steps. First, we show an inequality on the even moments of $|\beta^T A\gamma|$; next, we give a bound on the moment generation function of $|\beta^T A\gamma|$. Finally, we give the desired tail bound.

1. **Step 1: Even moments of $|\beta^T A\gamma|$.**
   Assume that $x = (x_1, \cdots, x_{p_1}), y = (y_1, \cdots, y_{p_2})$ are two random i.i.d. standard normal distributed vectors. Based on the definition of $\alpha_{\mathcal{P}}$ in (3.2), we know

$$\begin{aligned} E\beta_i^{2k} = E\gamma_j^{2k} \leq \alpha^{2k}Ex_i^{2k} = \alpha^{2k}Ey_j^{2k}; \\ E\beta_i^{2k-1} = E\gamma_j^{2k-1} = Ex_i^{2k-1} = Ey_j^{2k-1} = 0 \end{aligned} \tag{7.18}$$

   Consider the expansion of $E(\beta^T A\gamma)^{2k}$, where the non-zero terms can be written as

$$\prod_{l=1}^{2k} A_{i_l,j_l} \cdot \prod_{i=1}^{p_1} E\beta_i^{2s_i} \cdot \prod_{j=1}^{p_2} E\gamma_j^{2t_j}.$$

Here $s_1 + \cdots + s_{p_1} = t_1 + \cdots t_{p_2} = k$. By (7.18), this term can be bounded as

$$\left| \prod_{l=1}^{2k} A_{i_l,j_l} \cdot \prod_{i=1}^{p_1} E\beta_i^{2s_i} \cdot \prod_{j=1}^{p_2} E\gamma_j^{2t_j} \right| \leq \prod_{l=1}^{2k} |A_{i_l,j_l}| \cdot \alpha^{4k} \cdot \prod_{i=1}^{p_1} Ex_i^{2s_i} \cdot \prod_{j=1}^{p_2} Ey_j^{2t_j}$$

The right hand side is exact the term in the expansion of $\alpha^{4k} E(x^T A_{abs} y)^{2k}$, where A_{abs} is the the element-wise absolute value of A. Therefore, we have

$$E[\beta^T A\gamma]^{2k} \leq \alpha^{4k} E[x^T A_{abs} y]^{2k}. \tag{7.19}$$

Now we suppose A_{abs} has singular value decomposition

$$A_{abs} = \sum_{i=1}^{p} a_i u_i v_i^T = U \mathrm{diag}(a) V^T$$

where U, V are orthogonal and $a = (a_1, \cdots, a_p)$ is the singular value vector of A_{abs}. A well-known fact is that $\sum_i a_i^2 = \|A_{abs}\|_F^2 = \|A\|_F^2$. Since x, y are standard normal distributed, we can see that $x^T A_{abs} y$ and $x^T \mathrm{diag}(a) y$ has the same distribution. So

$$E[x^T A_{abs} y]^{2k} = E[\sum_{i=1}^{p} a_i x_i y_i]^{2k}$$

Next, we note

$$z = \sum_{i=1}^{p} x_i \frac{a_i y_i}{\sqrt{\sum_{j=1}^{p} a_j^2 y_j^2}},$$

then z is standard normal distributed and independent of $\sqrt{\sum_{j=1}^{p} a_j^2 y_j^2}$ since

$$\sum_{i=1}^{p} \left(\frac{a_i y_i}{\sqrt{\sum_{j=1}^{p} a_j^2 y_j^2}} \right)^2 = 1.$$

and z given $y_1, \cdots, y_p$ is always standard normal distributed.

For integer $k \geq 1$,

$$
\begin{aligned}
E[x^T A_{abs} y]^{2k} =& E\left| z \cdot \sqrt{\sum_{j=1}^{p} a_j^2 y_j^2} \right|^{2k} = E|z|^{2k} \cdot E(\sum_{j=1}^{p} a_j^2 y_j^2)^k \\
=& (2k-1)!! \cdot \sum_{\substack{k_1,k_2,\cdots,k_p \geq 0, \\ k_1+\cdots+k_p=k}} \frac{k!}{\prod_{j=1}^{p} k_j!} \prod_{i=1}^{p} E(a_i^2 y_i^2)^{k_i} \\
=& \sum_{\substack{k_1,k_2,\cdots,k_p \geq 0, \\ k_1+\cdots+k_p=k}} \frac{k!}{\prod_{j=1}^{p} (k_j)!} \prod_{i=1}^{p} a_i^{2k_i} \cdot \prod_{i=1}^{p} (2k_i-1)!! \\
\leq& \sum_{\substack{k_1,k_2,\cdots,k_p \geq 0, \\ k_1+\cdots+k_p=k}} \frac{k!}{\prod_{j=1}^{p} k_j!} \prod_{i=1}^{p} a_i^{2k_i} \cdot (2(k_1+k_2+\cdots k_p)-1)!! \\
=& ((2k-1)!!)^2 \cdot \sum_{\substack{k_1,k_2,\cdots,k_p \geq 0, \\ k_1+\cdots+k_p=k}} \frac{k!}{\prod_{j=1}^{p} k_j!} \prod_{i=1}^{p} (a_i^2)^{k_i} \\
=& ((2k-1)!!)^2 (\sum_{i=1}^{p} a_i^2)^k = ((2k-1)!!)^2 \|A\|_F^{2k}
\end{aligned}
$$

Together with (7.19), we have

$$
E[\beta^T A\gamma]^{2k} \leq \alpha^{4k}((2k-1)!!)^2 \|A\|_F^{2k} \tag{7.20}
$$

2. **Log-moment generation function of $|\beta^T A\gamma|$.**
By the bound of the even moments of $|\beta^T A\gamma|$, we can also give the estimate of odd moments, for integer $k \geq 1$,

$$
\begin{aligned}
0 \leq E\left|\beta^T A\gamma\right|^{2k+1} &\leq \sqrt{E[\beta^T A\gamma]^{2k} \cdot E[\beta^T A\gamma]^{2k+2}} \\
&\leq \alpha^{4k+2}(2k-1)!! \cdot (2k+1)!! \leq \alpha^{4k+2}(2k+1)!
\end{aligned}
$$

Also,

$$
E\left[\beta^T A\gamma\right]^{2k} \leq \alpha^{4k}((2k-1)!!)^2 \leq \alpha^{4k}(2k)!
$$

So for all $k \geq 2$, $E|\beta^T A\gamma|^k \leq \alpha^{2k} k!$. Denote $\mu = E|\beta^T A\gamma|$, then for $0 \leq t < \frac{1}{\alpha^2}$,

$$
Ee^{t|\beta^T A\gamma|} = 1 + \sum_{k=1}^{\infty} \frac{t^k}{k!} E|\beta^T A\gamma|^k \leq 1 + t\mu + \sum_{k=2}^{\infty} t^k \alpha^{2k} = 1 + t\mu + \frac{t^2\alpha^4}{1-t\alpha^2}
$$

For $-\frac{1}{\alpha^2} < t < 0$, we have

$$\begin{aligned} Ee^{t|\beta^T A\gamma|} =& 1 + t\mu + \sum_{k=1}^{\infty} t^{2k}\alpha^{4k} - \sum_{k=1}^{\infty} \frac{|t|^{2k+1}}{(2k+1)!} E|\beta^T A\gamma|^{2k+1} \\ \leq& 1 + t\mu + \frac{t^2\alpha^4}{1-|t|^2\alpha^4} \leq 1 + t\mu + \frac{t^2\alpha^4}{1-|t|\alpha^2} \end{aligned}$$

Hence, we have for all $-1/\alpha^2 < t < 1/\alpha^2$,

$$Ee^{t|\beta^T A\gamma|} \leq 1 + t\mu + \frac{t^2\alpha^4}{1-|t|\alpha^2}.$$

Note that $\log(1+x) \leq 1+x$ for all $-1 < x < \infty$, we have

(7.21)

$$\log E\exp(t(|\beta^T A\gamma| - \mu)) = \log E\exp(t|\beta^T A\gamma|) - t\mu \leq \frac{t^2\alpha^4}{1-|t|\alpha^2}.$$

for all $-1/\alpha^2 < t < 1/\alpha^2$.

3. **The tail bound of $\|\mathcal{X}(A)\|_1/n$.**
Finally, we estimate the tail bound of $\|\mathcal{X}(A)\|1/n$. Note that

$$\|\mathcal{X}(A)\|_1/n = \left(\sum_{j=1}^{n} \left|\beta^{(j)T} A\gamma^{(j)}\right|\right)/n,$$

based on (7.21), the logarithm of moment generating function of $\|\mathcal{X}(A)\|_1/n$ satisfies

$$\log E\exp\left(t(\|\mathcal{X}(A)\|_1/n - \mu)\right) = n\log E\exp\left(\frac{t}{n}(|\beta^T A\gamma| - \mu)\right) \leq \frac{t^2\alpha^4/n}{1-|t|\alpha^2/n}$$

By the proof of Lemma 1 in [27], we know $\|\mathcal{X}(A)\|_1/n - \mu$ has the tail bound,

$$P(\|\mathcal{X}(A)\|_1/n - \mu \geq \alpha^2(x/n + 2\sqrt{x/n})) \leq \exp(-x)$$

$$P(\|\mathcal{X}(A)\|_1/n - \mu \leq \alpha^2(x/n + 2\sqrt{x/n})) \leq \exp(-x)$$

Finally we set $\delta = \sqrt{x/n}$, by Lemma 7.2, $1/(3\alpha^4) \leq \mu \leq 1$, we finish the proof of Lemma. □

Proof of Lemma 7.2. Since $\mathcal{P}$ is symmetric and of variance 1, we have $E\beta_i = E\gamma_j = 0$, $E\beta_i^2 = E\gamma_j^2 = 1$, $E\beta_i^4 = E\gamma_j^4 \leq 3\alpha_{\mathcal{P}}^4$ for all i, j. Then by some expansions and calculations,

$$E\left(\beta^T A\gamma\right)^2 = E\left(\sum_{i,j}\beta_i A_{ij}\gamma_j\right)^2 = \sum_{i,j} E\beta_i^2 A_{ij}^2\gamma_j^2 = \sum_{i,j} A_{ij}^2 = \|A\|_F^2$$

By the first part in the proof of Lemma 7.1, we have

$$E|\beta^T A\gamma|^4 \leq 9\alpha_{\mathcal{P}}^8\|A\|_F^4$$

By Hölder's inequality,

$$E|\beta^T A\gamma| \leq \sqrt{E|\beta^T A\gamma|^2} = \|A\|_F$$

which gives the right of the original inequality. For the left, note that

$$E|\beta^T A\gamma|^2 \leq \left(E|\beta^T A\gamma|\right)^{2/3}\cdot\left(E|\beta^T A\gamma|^4\right)^{1/3}$$

So

$$E|\beta^T A\gamma| \geq \sqrt{\frac{\left(E|\beta^T A\gamma|^2\right)^3}{E|\beta^T A\gamma|^4}} \geq \frac{\|A\|_F}{3\alpha_{\mathcal{P}}^4}. \quad \square$$

Proof of Lemma 7.3.

- We first prove the sub-Gaussian part of the lemma. The moment generating function of $|z_i|$ $(t \geq 0)$ and z_i^2 $(0 \leq t < 1/(2\gamma))$ satisfy

$$\begin{aligned}
Ee^{t|z_i|} &= -\int_0^\infty \exp(t\lambda)dP(|X| \geq \lambda) = 1 + \int_0^\infty P(|X| \geq \lambda)d\exp(t\lambda) \\
&\leq 1 + \int_0^\infty 2t\exp(t\lambda - \lambda^2/(2\gamma^2))d\lambda \\
&\leq 1 + 2t\exp(t^2\gamma^2/2)\int_0^\infty \exp(-(\lambda-\gamma^2)^2/(2\gamma^2))d\lambda \\
&\leq 1 + 2t\exp\left(t^2\gamma^2/2\right)\sqrt{2\pi}\gamma \\
&\leq \exp(t^2\gamma^2/2)\left(1 + 2t\sqrt{2\pi}\gamma\right) \\
&\leq \exp\left(t^2\gamma^2/2 + 2\sqrt{2\pi}\gamma t\right)
\end{aligned}$$

$$\begin{aligned}
Ee^{tz_i^2} &= -\int_0^\infty \exp(t\lambda^2)dP(|X| \geq \lambda) = 1 + \int_0^\infty P(|X| \geq \lambda)d\exp(t\lambda^2) \\
&\leq 1 + \int_0^\infty 4t\lambda\exp\left(-\lambda^2(1/(2\gamma^2) - t)\right)d\lambda \\
&= 1 + \frac{2t}{1/(2\gamma^2) - t}
\end{aligned}$$

Then the moment generating function of $\|z\|_1/n$ and $\|z\|_2^2/n$ satisfies

$$Ee^{t\|z\|_1/n} = \left(Ee^{t\|z_i\|/n}\right)^n \le \exp\left(t^2\gamma^2/(2n) + 2\sqrt{2\pi}t\gamma\right)$$

$$Ee^{t\|z\|_2^2/n} = \left(1 + \frac{2t/n}{1/(2\gamma^2) - t/n}\right)^n \le \exp\left(\frac{2t}{1/(2\gamma^2) - t/n}\right)$$

Hence for $C \ge 0$,

$$P\left(\|z\|_1/n \ge C\right) \le \frac{E\exp(t\|z\|_1/n)}{\exp\left(tC\right)} \le \exp\left(t^2\gamma^2/(2n) + t\left(2\sqrt{2\pi}\gamma - C\right)\right)$$

$$= \exp\left(\frac{\gamma^2}{2n}\left(t + \frac{n(2\sqrt{2\pi}\gamma - C)}{\gamma^2}\right)^2 - \frac{n(2\sqrt{2\pi}\gamma - C)^2}{2\gamma^2}\right)$$

For $C > 2\sqrt{2\pi}\gamma$, we can set $t = \frac{\gamma^2}{n(C-2\sqrt{2\pi}\gamma)}$, then

$$P\left(\|z\|_1/n \ge C\right) \le \exp\left(-\frac{n(C - 2\sqrt{2\pi}\gamma)^2}{2\gamma^2}\right)$$

Now we consider the tail bound of $\|z\|_2^2/n$. For $C > 4\gamma^2$,

$$P(\|z\|_2^2/n \ge C) = \frac{E\exp(t\|z\|_2^2/n)}{\exp(tC)} = \exp\left(\frac{2t^2\gamma^2 C/n - t(C - 4\gamma^2)}{1 - 2\gamma^2 t/n}\right).$$

We set $t = \frac{C-4\gamma^2}{4C\gamma^2/n}$,

$$P\left(\|z\|_2^2/n \ge C\right) \le \exp\left(-\frac{n(C - 4\gamma^2)^2}{8\gamma^2 C(1 - 2\gamma^2 t/n)}\right) \le \exp\left(-\frac{n(C - 4\gamma^2)^2}{8\gamma^2 C}\right)$$

Finally we consider $\|z\|_\infty$,

$$P(\|z\|_\infty \le C\sqrt{\log n}\gamma) \le 2n\exp(-C^2\log n\gamma^2/(2\gamma^2)) = 2n^{-(C^2/2-1)}$$

- Next, we consider the Gaussian part of the lemma. The bound of $\|z\|_2$ is already given by Lemma 5.1 in [7]. For $\|z\|_1$, we can see $E|z_i|^2 = \sigma^2$,

$$E|z_i| = \frac{\sigma}{\sqrt{2\pi}}\int_0^\infty xe^{-x^2/(2)}dx = \sigma\sqrt{2/\pi}$$

Hence, $E(\|z\|_1/n) = \sigma\sqrt{2/\pi}$, $\mathrm{Var}(\|z\|_1/n) = \mathrm{Var}(|z_i|)/n = (1 - 2/\pi)\sigma^2/n$. By Chebyshev' inequality,

$$P(\|z\|_1/n \ge \sigma) \le P\left(\left|\|z\|_1/n - \sigma\sqrt{2/\pi}\right| \ge \sigma(1 - \sqrt{2/\pi})\right)$$

$$\le \frac{\mathrm{Var}(\|z\|_1/n)}{\sigma^2(1 - \sqrt{2/\pi})^2} = \frac{1 + \sqrt{2/\pi}}{(1 - \sqrt{2/\pi})n}$$

For the bound of $\|z\|_\infty$, we have

$$\begin{aligned} P(\|z\|_\infty \geq 2\sqrt{\log n}\sigma) &\leq \sum_{i=1}^{n} P(|z_i| \leq 2\sqrt{\log n}\sigma) \\ &\leq n \cdot \frac{2}{2\sqrt{2\pi \log n}} \exp(-\frac{1}{2} \cdot 4 \log n) = \frac{1}{n\sqrt{2\pi \log n}} \end{aligned}$$

□

Proof of Lemma 7.4. Again without confusion, we simply use α to represent $\alpha_{\mathcal{P}}$ in the proof. The proof also requires some knowledge of moment generation function and ε-net method. We'll prove by steps.

- **Moment Generation Function of $a^T\mathcal{X}^*(z)b$.** Suppose $a \in \mathbb{R}^{p_1}$, $b \in \mathbb{R}^{p_2}$ are fixed unit vectors. In order to handle the operator norm of $\mathcal{X}^*(z)$, we first consider $a^T\mathcal{X}^*(z)b$. Note that

$$a^T\mathcal{X}^*(z)b = \sum_{i=1}^{n} z_i a^T \beta^{(i)} \gamma^{(i)T} b \tag{7.22}$$

Denote $X_i = a^T\beta^{(i)}$, $Y_i = b^T\gamma^{(i)}$, then $\{X_i\}_{i=1}^{p_1}, \{Y_i\}_{i=1}^{p_2}$ are two independent sets of i.i.d. sub-Gaussian samples. Moreover by $\beta^{(i)}, \gamma^{(i)}$ are i.i.d. from symmetric distribution $\mathcal{P}$, one can show

$$E(X_i)^{2k-1} = E(\sum_j a_j\beta_j)^{2k-1} = 0,$$

$$\begin{aligned} E(X_i)^{2k} = E(\sum_{j=1}^{p_1} a_j\beta_j)^{2k} &= \sum_{k_1+\cdots k_{p_1}=k} \frac{k!}{k_1!\cdots k_{p_1}!} \left(\prod_{i=1}^{p_1} a_i^{2k_i} E(\beta_i^{2k_i})\right) \\ &\leq \sum_{k_1+\cdots k_{p_1}=k} \frac{k!}{k_1!\cdots k_{p_1}!} \left(\prod_{i=1}^{p_1} a_i^{2k_i} E(x_i^{2k_i})\alpha^{2k}\right) \\ &= \alpha^{2k} E(\sum_{i=1}^{p_1} a_i x_i)^{2k} \leq \alpha^{2k}(2k-1)!! \end{aligned}$$

Here $x_i \overset{iid}{\sim} N(0,1)$. Similarly, $E(Y_i)^{2k-1} = 0$, $E(Y_i)^{2k} \le \alpha^{2k}(2k-1)!!$. Then for $|t| < 1/\alpha^2$,

(7.23)
$$\begin{aligned} E\exp(tX_iY_i) &= \sum_{k=0}^{\infty} \frac{t^k E(X_iY_i)^k}{k!} \le \sum_{k=0}^{\infty} \frac{t^{2k}(\alpha^{2k}(2k-1)!!)^2}{(2k)!} \\ &= \sum_{k=0}^{\infty} \frac{(t\alpha^2)^{2k}(2k-1)!!}{2^k k!} = \sum_{k=0}^{\infty} (t\alpha^2)^{2k} \cdot (-1)^k \binom{-1/2}{k} \\ &= \frac{1}{\sqrt{1-t^2\alpha^4}} \end{aligned}$$

Now for fixed $z \in \mathbb{R}^n$, the logarithm of the moment generating function of $a^T\mathcal{X}^*(z)b$ satisfies

$$\begin{aligned} \log E\exp(ta^T\mathcal{X}^*(z)b) &= \sum_{i=1}^{n} \log E\exp(tz_iX_iY_i) \le \sum_{i=1}^{n} -\frac{1}{2}\log(1-t^2z_i^2\alpha^4) \\ &\le \sum_{i=1}^{n} \frac{t^2z_i^2\alpha^4}{2(1-t^2z_i^2\alpha^4)} \le \frac{t^2\|z\|_2^2\alpha^4}{2(1-t^2\|z\|_\infty^2\alpha^4)} \\ &\le \frac{t^2\|z\|_2^2\alpha^4}{2(1-|t|\|z\|_\infty\alpha^2)} \end{aligned}$$

for any $|t| < 1/(\|z\|_\infty\alpha^2)$. Here we used the fact that

$$-\log(1-x) = \sum_{i=1}^{\infty} \frac{x^i}{i} \le \sum_{i=1}^{\infty} x^i = \frac{x}{1-x}$$

for $0 \le x < 1$.

- **Tail Bound of $a^T\mathcal{X}^*(z)b$**
  By the proof of Lemma 1 in [27], we know for fixed $z \in \mathbb{R}^n$, $a^T\mathcal{X}^*(z)b$ has tail bound: for $x > 0$ and fixed a, b, we have

$$P\left(a^T\mathcal{X}^*(z)b \ge \alpha^2\left(\|z\|_\infty x + \|z\|_2\sqrt{2x}\right)\right) \le \exp(-x);$$

$$P\left(a^T\mathcal{X}^*(z)b \le -\alpha^2\left(\|z\|_\infty x + \|z\|_2\sqrt{2x}\right)\right) \le \exp(-x).$$

  Set $x = C(p_1+p_2)$, we have

(7.24)
$$\begin{aligned} &P\left(|a^T\mathcal{X}^*(z)b| \ge \alpha^2\left(\|z\|_\infty C(p_1+p_2) + \|z\|_2\sqrt{2C(p_1+p_2)}\right)\right) \\ \le& 2\exp(-C(p_1+p_2)). \end{aligned}$$

For convenience, We denote

$$T = \alpha^2 \left(\|z\|_\infty (C(p_1+p_2)) + \|z\|_2 \sqrt{2C(p_1+p_2)} \right). \tag{7.25}$$

- **ε-net and the upper bound of $\|\mathcal{X}^*(z)\|$.**
  In this step, we still fix z. We use the ε-net method to derive the upper bound of $\|\mathcal{X}^*(z)\|_2$, which is given by

$$\|\mathcal{X}^*(z)\|_2 = \sup_{a\in\mathbb{R}^{p_1}, b\in\mathbb{R}^{p_2}} a^T \mathcal{X}^*(z) b$$

  From Lemma 2.5 in [38], we can find an ε-net A in the unit sphere of $\mathbb{R}^{p_1}$, i.e. for all a in the unit sphere of $\mathbb{R}^{p_1}$, there exists $a' \in A$ such that $\|a'-a\|_2 \le \epsilon$. Besides, $|A| \le (1+2/\varepsilon)^{p_1}$. Similarly, there exists ϵ-net B of the unit ball of $\mathbb{R}^{p_2}$ such that $|B| \le (1+2/\varepsilon)^{p_2}$.
  By (7.24), we have

$$P\left(\left|a^T\mathcal{X}^*(z)b\right| \ge T, \exists a\in A, b\in B\right) \le 2(1+2/\varepsilon)^{p_1+p_2}\exp(-C(p_1+p_2)) \tag{7.26}$$

  Now we consider under the event that $\left|a^T\mathcal{X}^*(z)b\right| \le T, \forall a\in A, b\in B$. Suppose $\mu = \|\mathcal{X}^*(z)\|_2 = \max_{\|a\|_2=\|b\|_2=1} a^T\mathcal{X}^*(z)b$, $(a^*,b^*) = \arg\max_{a,b} a^T\mathcal{X}^*(z)b$, then we can find $a'\in A, b'\in B$ such that $\|a'-a^*\| \le \varepsilon$, $\|b'-b^*\| \le \varepsilon$. Then,

$$\begin{aligned}
\mu &= \left|a^{*T}\mathcal{X}^*(z)b^*\right| \\
&= \left|a'^T\mathcal{X}^*(z)b'\right| + \left|(a'-a^*)^T\mathcal{X}^*(z)b'\right| + \left|a^{*T}\mathcal{X}^*(z)(b^*-b')\right| \\
&\le T + (\|a'-a^*\|_2 + \|b'-b^*\|_2)\cdot\|\mathcal{X}^*(z)\| \le T + 2\varepsilon\mu
\end{aligned}$$

  This means $\mu \le T/(1-2\varepsilon)$. Therefore, when $\left|a^T\mathcal{X}^*(z)b\right| \le T, \forall a\in A, b\in B$, we have $\|\mathcal{X}^*(z)\| \le T/(1-2\varepsilon)$.
- Finally, we set $\varepsilon = 1/3$, under the event that

$$\begin{aligned}
&\left|a^T\mathcal{X}^*(z)b\right| \le T \\
=&\alpha^2\left(\|z\|_\infty(C(p_1+p_2)) + \|z\|_2\sqrt{2C(p_1+p_2)}\right), \forall a\in A, b\in B
\end{aligned}$$

  we have

$$\|\mathcal{X}^*(z)\| \le T/(1-2\varepsilon) \le 3\alpha^2\left(\|z\|_\infty C(p_1+p_2) + \|z\|_2\sqrt{2C(p_1+p_2)}\right)$$

  By (7.26), the probability that all the event happen is at least

$$1 - 2\exp\left(-(C-\log 7)(p_1+p_2)\right)$$

  This finished the proof of the lemma. □

Proof of Proposition 7.1.. In the proof, we will use α to represent $\alpha_{\mathcal{P}}$ without any confusion. The ideas of the proof of Proposition 7.1 follows from [35], [12].

Denote $S_{kr} = \{X \in \mathbb{R}^{p_1 \times p_2} : \text{rank}(X) \le kr, \|X\|_F = 1\}$. By Lemma 3.1 in [12], for any $\varepsilon > 0$, there exists ε-net S'_{kr} such that $|S'_{kr}| \le (9/\varepsilon)^{(p_1+p_2+1)kr}$.

For given C_1, C_2 such that $C_1 < 1/(3\alpha^4)$, $C_2 > 1$, we set $C'_1 = \frac{C_1 + 1/(3\alpha^4)}{2}$, $C'_2 = \frac{C_2+1}{2}$. We can choose δ_0 small enough such that

$$\alpha^2 \left(2\delta_0 + \delta_0^2\right) \le \min\left(\frac{1/(3\alpha^4) - C_1}{2}, \frac{C_2 - 1}{2}\right)$$

then by Lemma 7.1, for any given $A \in \mathbb{R}^{p_1 \times p_2}$, we have $C'_1 \le \|\mathcal{X}(A)\|_1/n \le C'_2$ with probability at least $1 - 2\exp(-\delta_0^2 n)$. Hence,

$$P\left(C'_1 \le \|\mathcal{X}(A)\|_1/n \le C'_2, \text{ for all } A \in S'_r\right) \ge 1 - 2(9/\varepsilon)^{kr(p_1+p_2+1)} \cdot \exp(-\delta_0^2 n)$$

Next, we'll estimate the bound of $\|\mathcal{X}(A)\|_1/n$ on the whole set S_{kr} provided that $C'_1 \le \|\mathcal{X}(A)\|_1/n \le C'_2$ for all $A \in S'_{kr}$. Define

$$\kappa_1 = \inf_{A \in S_{kr}} \|\mathcal{X}(A)\|_1/n \text{ and } \kappa_2 = \sup_{A \in S_{kr}} \|\mathcal{X}(A)\|_1/n.$$

For any $A \in S_{kr}$, there exists $A' \in S'_{kr}$ such that $\|A - A'\|_F \le \varepsilon$. So

$$\|\mathcal{X}(A)\|_1/n \le \|\mathcal{X}(A')\|_1/n + \|\mathcal{X}(A - A')\|_1/n \le C'_2\|A\| + \kappa_2\|A - A'\|_F \le C'_2 + \kappa_2\varepsilon$$

$$\|\mathcal{X}(A)\|_1/n \ge \|\mathcal{X}(A')\|_1/n - \|\mathcal{X}(A - A')\|_1/n \ge C'_1\|A\| - \kappa_2\|A - A'\|_F \ge C'_1 - \kappa_2\varepsilon$$

which mean

$$\kappa_2 = \sup_{A \in S_{kr}} \|\mathcal{X}(A)\|_F \le C'_2 + \varepsilon\kappa_2, \quad \kappa_1 = \inf_{A \in S_{kr}} \|\mathcal{X}(A)\|_F \ge C'_1 - \varepsilon\kappa_2$$

namely, $\kappa_2 \le C'_2/(1 - \varepsilon)$, $\kappa_1 \ge C'_1 - \varepsilon\kappa_2$. We choose

$$\varepsilon \le \min\left(\frac{C_2 - 1}{2C_2}, \frac{1/(3\alpha^4) - C_1}{2C_2}\right),$$

by some calculations we can see $\kappa_1 \ge C_1$, $\kappa_2 \le C_2$.

To sum up, we can choose δ_0, ε only depending on C_1, C_2, α, to ensure

$$C_1 \le \kappa_1 = \inf_{A \in S_{kr}} \|\mathcal{X}(A)\|_1/n \le \sup_{A \in S_{kr}} \|\mathcal{X}(A)\|_1/n = \kappa_2 \le C_2$$

with probability at least $1 - 2(9/\varepsilon)^{kr(p_1+p_2+1)}\exp(-\delta_0^2 n)$. The last step is to estimate the probability above. We choose $D \geq 8k\log(9/\varepsilon)/\delta_0^2$, then for $n \geq Dr(p_1+p_2)$, we have

$$\delta_0^2 n/2 \geq 4\log(9/\varepsilon)kr(p_1+p_2) \geq \log 2 + 2\log(9/\varepsilon)kr(p_1+p_2+1),$$

$$\begin{aligned}&1 - 2(9/\varepsilon)^{kr(p_1+p_2+1)}e^{-\delta_0^2 n} = 1 - \exp(-\delta_0^2 n + \log 2 + kr(p_1+p_2+1)\log(9/\varepsilon))\\ \geq& 1 - \exp(-\delta_0^2 n/2).\end{aligned}$$

Finally, we finish the proof of the Theorem by choosing $\delta \leq \delta_0^2/2$. □

Proof of Lemma 7.7.. The idea of the proof is originated from [40, 30]. We provide the proof here for the completeness of the paper. Note $p = \min(p_1, p_2)$, suppose for any matrix B, $\sigma_i(B)$ is the i-th largest singular value of B. By Lemma 2 in [30], we have

$$\begin{aligned}&\|A_{\max(r)}\|_* + \|A_{-\max(r)}\|_*\\ =&\|A\|_* \geq \|A_*\|_* = \|A - (-R)\|_* \geq \sum_{i=1}^{p}|\sigma_i(A) - \sigma_i(-R)|\\ \geq& \sum_{i=1}^{r}(\sigma_i(A) - \sigma_i(R)) + \sum_{i=r+1}^{p}(\sigma_i(R) - \sigma_i(A))\\ =&\|A_{\max(r)}\|_* - \|A_{-\max(r)}\|_* + \|R_{-\max(r)}\|_* - \|R_{\max(r)}\|_*\end{aligned}$$

which implies (7.3). □

Proof of Lemma 7.8.. Suppose $R = A_* - A$, then we have

$$\|\mathcal{X}(R)\|_1/n \leq \lambda_1 \tag{7.27}$$

Since $\|A_*\|_* \leq \|A\|_*$. By Lemma 7.7, we must have (7.3). Suppose $p = \min(p_1, p_2)$ and R has the singular value decomposition, $R = \sum_{i=1}^{p}\sigma_i u_i v_i^T = U\mathrm{diag}(\vec{\sigma})V^T$, then $\vec{\sigma}_{-\max(kr)}$ satisfies

$$\|\vec{\sigma}_{-\max(kr)}\|_\infty \leq \sigma_{kr},$$

$$\begin{aligned}\|\vec{\sigma}_{-\max(kr)}\|_1 =&\|\vec{\sigma}_{-\max(r)}\|_1 - (\sigma_{r+1} + \cdots + \sigma_{kr})\\ \leq&\|\vec{\sigma}_{\max(r)}\|_1 + 2\|A_{-\max(r)}\|_* - (k-1)r\sigma_{kr}\end{aligned}$$

Set

$$\theta = \max\left(\sigma_{kr}, (\|\vec{\sigma}_{\max(r)}\|_1 + 2\|A_{-\max(r)}\|_* - (k-1)r\sigma_{kr})/(kr)\right),$$

then $\|\vec{\sigma}_{-\max(kr)}\|_\infty \le \theta$, $\|\vec{\sigma}_{-\max(kr)}\|_1 \le kr\theta$. Similarly to the proof of Theorem 2.1, apply Lemma 7.6, we can get $b^{(i)} \in \mathbb{R}^n$, $\lambda_i \ge 0$, $i = 1, \cdots, N$ such that $\vec{\sigma}_{-\max(kr)} = \sum_{i=1}^N \lambda_u b^{(i)}$ and (7.2). Hence,

$$\|b^{(i)}\|_2 \le \sqrt{\|b^{(i)}\|_1 \cdot \|b^{(i)}\|_\infty} \le \sqrt{\theta(\|\vec{\sigma}_{\max(r)}\|_1 + 2\|A_{-\max(r)}\|_* - (k-1)r\sigma_{kr})}$$

If $\theta = \sigma_{kr}$, we can optimize over σ_{kr} in the inequality,

$$\begin{aligned}\|b^{(i)}\|_2 \le& \sqrt{\sigma_{kr}(\|\vec{\sigma}_{\max(r)}\|_1 + 2\|A_{-\max(r)}\|_* - (k-1)r\sigma_{kr})}\\ \le& \frac{\|\vec{\sigma}_{\max(r)}\|_1 + 2\|A_{-\max(r)}\|_*}{2\sqrt{r(k-1)}};\end{aligned}$$

if $\theta = (\|\vec{\sigma}_{\max(r)}\|_1 + 2\|A_{-\max(r)}\|_* - (k-1)r\sigma_{kr})/(kr)$, we have

$$\begin{aligned}\|b^{(i)}\|_2 \le& \frac{\|\vec{\sigma}_{\max(r)}\|_1 + 2\|A_{-\max(r)}\|_* - (k-1)r\sigma_{kr}}{\sqrt{kr}}\\ \le& \frac{\|\vec{\sigma}_{\max(r)}\|_1 + 2\|A_{-\max(r)}\|_*}{\sqrt{kr}}.\end{aligned} \tag{7.28}$$

Since $k \ge 2$, we always have (7.28). Next, we define $B_i = U\mathrm{diag}(b^{(i)})V^T$, then the rank of B_i are at most kr, $\sum_{i=1}^N \lambda_i B_i = R_{-\max(kr)}$ and $\|B_i\|_F = \|b^{(i)}\|_2$. Then

$$\begin{aligned}\lambda_1 \ge& \|\mathcal{X}(R)\|_1/n \ge \|\mathcal{X}(R_{\max(kr)})\|_1/n - \|\mathcal{X}(R_{-\max(kr)})\|_1/n\\ \ge& C_1\|R_{\max(kr)}\|_F - \sum_{i=1}^N \|\mathcal{X}(\lambda_i B_i)\|_1/n\\ \ge& C_1\|R_{\max(kr)}\|_F - \sum_{i=1}^N \lambda_i C_2\|B_i\|_F\\ \ge& C_1\|R_{\max(kr)}\|_F - C_2\frac{\|R_{\max(r)}\|_* + 2\|A_{-\max(r)}\|_*}{\sqrt{kr}}\\ \ge& C_1\|R_{\max(kr)}\|_F - \frac{C_2}{\sqrt{k}}\|R_{\max(kr)}\|_F - \frac{2C_2\|A_{-\max(r)}\|_*}{\sqrt{kr}},\end{aligned} \tag{7.29}$$

where the last inequality is due to $\|R_{\max(kr)}\|_F \ge \|R_{\max(r)}\|_F \ge \sqrt{r}\|R_{\max(r)}\|_*$. Therefore,

$$\|R_{\max(kr)}\|_F \le \frac{\lambda_1}{C_1 - C_2/\sqrt{k}} + \frac{2\|A_{-\max(r)}\|_*}{\sqrt{r}(\sqrt{k}C_1/C_2 - 1)} \tag{7.30}$$

Finally,

$$\begin{aligned}\|R_{-\max(kr)}\|_F =& \|\vec{\sigma}_{-\max(kr)}\|_2 \le \sqrt{\|\vec{\sigma}_{-\max(kr)}\|_1 \cdot \|\vec{\sigma}_{-\max(kr)}\|_\infty} \\ \le& \sqrt{\sigma_{kr} \cdot (\|\vec{\sigma}_{-\max(r)}\|_1 - r(k-1)\sigma_{kr})} \\ \le& \frac{\|\vec{\sigma}_{-\max(r)}\|_1}{2\sqrt{r(k-1)}} \le \frac{\|\vec{\sigma}_{\max(r)}\|_1 + 2\|A_{-\max(r)}\|_*}{2\sqrt{r(k-1)}} \\ \le& \frac{\|R_{\max(r)}\|_F}{2\sqrt{k-1}} + \frac{\|A_{-\max(r)}\|_*}{\sqrt{r(k-1)}}\end{aligned}$$

Therefore,

$$\begin{aligned}\|R\|_F =& \sqrt{\|R_{\max(kr)}\|_F^2 + \|R_{-\max(kr)}\|_F^2} \\ \le& \sqrt{\|R_{\max(kr)}\|_F^2 + \left(\frac{\|R_{\max(r)}\|_F}{2\sqrt{k-1}} + \frac{\|A_{-\max(r)}\|_*}{\sqrt{r(k-1)}}\right)^2} \\ \le& \sqrt{1 + \frac{1}{4(k-1)}}\|R_{\max(kr)}\|_F + \frac{\|A_{-\max(r)}\|_*}{\sqrt{r(k-1)}} \\ \le& \left(1 + \frac{1}{8(k-1)}\right)\|R_{\max(kr)}\|_F + \frac{\|A_{-\max(r)}\|_*}{\sqrt{r(k-1)}} \\ \le& \frac{2}{C_1 - C_2/\sqrt{k}}\lambda_1 + \left(\frac{3}{\sqrt{k}C_1/C_2 - 1} + \frac{1}{\sqrt{k-1}}\right)\frac{\|A_{-\max(r)}\|_*}{\sqrt{r}}\end{aligned} \tag{7.31}$$

Proof of Lemma 7.9.. The proof of this theorem is similar to the proof of Lemma 7.8. Suppose $R = A_* - A$. In this case we have

$$\|\mathcal{X}^*\mathcal{X}(R)\| \le \lambda_2 \tag{7.32}$$

instead of (7.27). Besides, since $\|A_*\|_* \le \|A\|_*$ and Lemma 7.7, we still have (7.3). With the similar argument as (7.29), we have

(7.33)

$$\begin{aligned}\lambda_2\|R\|_* \ge& \langle R, \mathcal{X}^*\mathcal{X}(R)\rangle = \|\mathcal{X}(R)\|_2^2 \ge \|\mathcal{X}(R)\|_1^2/n \\ \ge& n\left(C_1\|R_{\max(kr)}\|_F - \frac{C_2}{\sqrt{k}}\|R_{\max(kr)}\|_F - \frac{2C_2\|A_{-\max(r)}\|_*}{\sqrt{kr}}\right)_+^2\end{aligned}$$

Here $(x)_+$ means $\max(x, 0)$. Besides,

$$\begin{aligned}\lambda_2\|R\|_* \le& \lambda_2\left(\|R_{\max(r)}\|_* + (\|R_{\max(r)}\|_* + 2\|A_{-\max(r)}\|_*)\right) \\ \le& 2\lambda_2\left(\sqrt{r}\|R_{\max(r)}\|_F + \|A_{-\max(r)}\|_*\right) \\ \le& 2\lambda_2\left(\sqrt{r}\|R_{\max(kr)}\|_F + \|A_{-\max(r)}\|_*\right)\end{aligned}$$

Suppose $x = \|R_{\max(kr)}\|_F$, $y = \|A_{-\max(r)}\|_*/\sqrt{r}$. Based on the previous two inequalities, we have

$$n\left((C_1 - C_2/\sqrt{k})x - \frac{2C_2}{\sqrt{k}}y\right)_+^2 \le 2\sqrt{r}(x+y)\lambda_2$$

When $x \ge \frac{2C_2 y}{\sqrt{k}(C_1 - C_2/\sqrt{k})}$, the inequality above leads to

(7.34)
$$n(C_1 - C_2/\sqrt{k})^2 x^2 - \left(2n(C_1 - C_2/\sqrt{k})\frac{2C_2}{\sqrt{k}}y + 2\sqrt{r}\lambda_2\right)x - 2\sqrt{r}\lambda_2 y \le 0.$$

Note that for second order inequality $ax^2 - bx - c \le 0$, $a > 0$, $b, c \ge 0$, we have $x \le \frac{b+\sqrt{b^2+4ac}}{2a} \le b/a + \sqrt{c/a}$. Hence we can get an upper bound of x from (7.34).

$$\begin{aligned} x &\le \frac{2\sqrt{r}\lambda_2}{n(C_1 - C_2/\sqrt{k})^2} + \frac{4C_2/\sqrt{k}y}{(C_1 - C_2/\sqrt{k})} + \frac{\sqrt{2\sqrt{r}\lambda_2 y}}{\sqrt{n}(C_1 - C_2/\sqrt{k})} \\ &\le \frac{2\sqrt{r}\lambda_2}{n(C_1 - C_2/\sqrt{k})^2} + \frac{4C_2/\sqrt{k}y}{(C_1 - C_2/\sqrt{k})} + \frac{\sqrt{r}\lambda_2}{n(C_1 - C_2/\sqrt{k})^2} + \frac{1}{2}y. \end{aligned}$$

Hence whenever $x \ge \frac{2C_2 y}{\sqrt{k}(C_1 - C_2/\sqrt{k})}$ or not,

(7.35)
$$\begin{aligned} &\|R_{\max(kr)}\|_F = x \\ &\le \max\left\{\frac{2C_2 y}{\sqrt{k}(C_1 - C_2/\sqrt{k})}, \frac{3\sqrt{r}\lambda_2}{n(C_1 - C_2/\sqrt{k})^2} + \left(\frac{4C_2/\sqrt{k}}{(C_1 - C_2/\sqrt{k})} + \frac{1}{2}\right)y\right\} \\ &\le \frac{3\sqrt{r}\lambda_2}{n(C_1 - C_2/\sqrt{k})^2} + \left(\frac{4C_2/\sqrt{k}}{(C_1 - C_2/\sqrt{k})} + \frac{1}{2}\right)\frac{\|A_{-\max(r)}\|_*}{\sqrt{r}}. \end{aligned}$$

Finally, similarly to (7.31) in Lemma 7.8, we can get the upper bound of $\|R\|_F$,

$$\begin{aligned} \|R\|_F &\le \left(1 + \frac{1}{8k-8}\right)\|R_{\max(kr)}\|_F + \frac{\|A_{-\max(r)}\|_*}{\sqrt{r(k-1)}} \\ &\le \frac{4}{(C_1 - C_2/\sqrt{k})^2}\cdot\frac{\sqrt{r}(\varepsilon+\eta)}{n} + \left(\frac{5}{\sqrt{k}C_1/C_2 - 1} + \frac{1}{\sqrt{k-1}} + 1\right)\frac{\|A_{-\max(r)}\|_*}{\sqrt{r}} \end{aligned}$$

which finished the proof of lemma 7.9. □

7.7. *Proof of Propositions 2.2, 2.3 and 2.4..* We first show Proposition 2.2. Denote $\mathcal{X}_1$, $\mathcal{X}_2 : \mathbb{R}^{p\times p} \to \mathbb{R}^{\lfloor \frac{n}{2} \rfloor}$ such that

$$[\mathcal{X}_1]_i(B) = (\frac{\beta^{(2i-1)} + \beta^{(2i)}}{\sqrt{2}})^T B(\frac{\beta^{(2i-1)} - \beta^{(2i)}}{\sqrt{2}}), \quad i = 1, \cdots, \lfloor \frac{n}{2} \rfloor \tag{7.36}$$

$$[\mathcal{X}_2](B) = (\frac{\beta^{(2i-1)} - \beta^{(2i)}}{\sqrt{2}})^T B(\frac{\beta^{(2i-1)} + \beta^{(2i)}}{\sqrt{2}}), \quad i = 1, \cdots, \lfloor \frac{n}{2} \rfloor \tag{7.37}$$

Note that $\frac{1}{\sqrt{2}}\left(\beta^{(2i-1)} + \beta^{(2i)}\right)$ and $\frac{1}{\sqrt{2}}\left(\beta^{(2i-1)} - \beta^{(2i)}\right)$ are independent i.i.d. standard normal samples, so both $\mathcal{X}_1$ and $\mathcal{X}_2$ are ROP design (see (1.6)). By Corollary 2.1, we know there exists uniform constant C such that whenever $\lfloor \frac{n}{2} \rfloor \geq Cr \cdot 2p$, $\mathcal{X}_1$ satisfies the following property with probability at least $1 - \exp(-n\delta)$,

$$\forall A \in \{A \in \mathbb{R}^{p\times p} : rank(A) \leq r\}, \quad A = \arg \min_{B\in\mathbb{R}^{p\times p}} \|B\|_* \quad \text{subject to} \quad \mathcal{X}_1(B) = \mathcal{X}_1(A). \tag{7.38}$$

Now we consider the event that (7.38) holds. We note that for any symmetric matrix B,

$$[\mathcal{X}_1]_i(B) = \frac{1}{2}\beta^{(2i-1)T} B\beta^{(2i-1)} - \frac{1}{2}\beta^{(2i)T} B\beta^{(2i)} = \frac{1}{2}\left([\mathcal{X}]_{2i-1}(B) - [\mathcal{X}]_{2i}(B)\right)$$

So $\mathcal{X}(B) = \mathcal{X}(A)$ implies $\mathcal{X}_1(B) = \mathcal{X}_1(A)$ for symmetric A and B. Also, since A is feasible in programming (2.13), we have

$$\begin{aligned}\|A\|_* &\geq \min_{B\in\mathbb{S}^p} \|B\|_* \quad \text{subject to} \quad \mathcal{X}(B) = \mathcal{X}(A)\\ &\geq \min_{B\in\mathbb{R}^{p\times p}} \|B\|_* \quad \text{subject to} \quad \mathcal{X}_1(B) = \mathcal{X}_1(A)\\ &= \|A\|_*,\end{aligned}$$

So we can conclude that A can be exactly recovered by (2.13) given (7.38) holds. In summary, for $n \geq 6Crp$, with probability at least $1 - \exp(-n\delta)$, $\mathcal{X}$ satisfies (7.38), then programming (2.13) can recover all $A \in \mathbb{S}^p$ of rank at most r. □

Next, we consider Proposition 2.3. The idea of the proof is similar to Proposition 2.1. Define $\tilde{z} \in \mathbb{R}^{\lfloor \frac{n}{2} \rfloor}$ such that

$$\tilde{z}_i = z_{2i-1} - z_{2i}, \quad i = 1, \cdots, \lfloor \frac{n}{2} \rfloor. \tag{7.39}$$

Then $\tilde{z} \overset{iid}{\sim} N(0,2)$. We shall also point out two facts, $\tilde{z} = \tilde{y} - \tilde{\mathcal{X}}(A)$ and $\mathcal{X}^* = \mathcal{X}_1^* + \mathcal{X}_2^*$ (defined as (7.36), (7.37)). By Lemma 7.3, we know

$$P(\|z\|_1/n > \sigma) \le \frac{9}{n}$$

$$P(\|\tilde{z}\|_2 > \sigma\sqrt{2n}) \le \frac{1}{\lfloor n/2 \rfloor}$$

$$P(\|\tilde{z}\|_\infty > 2\sigma\sqrt{2\log n}) \le \frac{1}{\lfloor n/2 \rfloor}$$

Hence,

$$\begin{aligned}
&P\left(A \text{ is NOT in the feasible set of programming (2.17)}\right)\\
=&P\left(\|z\|_1/n > \sigma, \text{ or } \|\tilde{\mathcal{X}}^*(\tilde{z})\| > \eta\right)\\
=&P\left(\|z\|_1/n > \sigma, \text{ or } \|\tilde{\mathcal{X}}^*(\tilde{z})\| > 24\sigma\sqrt{pn} + 48\sigma p\sqrt{2\log n}\right)\\
\le&P(\|z\|_1/n > \sigma) + P_{\mathcal{X}}(\|\tilde{z}\|_2 > \sigma\sqrt{2n}) + P(\|\tilde{z}\|_\infty > 2\sigma\sqrt{2\log n})\\
&+ P_{\mathcal{X}}\left(\|\tilde{\mathcal{X}}^*(\tilde{z})\| > 24p\|\tilde{z}\|_\infty + 12\sqrt{2p}\|\tilde{z}\|_2\right)\\
\le&\frac{9}{n} + \frac{2}{\lfloor n/2 \rfloor} + P_{\mathcal{X}}\left(\|\mathcal{X}_1(\tilde{z})\| > 12p\|\tilde{z}\|_\infty + 6\sqrt{2p}\|\tilde{z}\|_2\right)\\
&+ P_{\mathcal{X}}\left(\|\mathcal{X}_2(\tilde{z})\| > 12p\|\tilde{z}\|_\infty + 6\sqrt{2p}\|\tilde{z}\|_2\right)\\
\le&\frac{15}{n} + 4\exp(-2p(2-\log 7))
\end{aligned}$$

When A is in the feasible set of programming (2.17), we have $\|\hat{A}\|_* \le \|A\|_*$ and

$$\begin{aligned}
\|\tilde{\mathcal{X}}(\hat{A} - A)\|_1 &= \sum_{i=1}^{\lfloor \frac{n}{2} \rfloor} \left|[\mathcal{X}]_{2i-1}(\hat{A} - A) - [\mathcal{X}]_{2i}(\hat{A} - A)\right|\\
&\le \|\mathcal{X}(\hat{A} - A)\|_1 \le \|y - \mathcal{X}(\hat{A})\|_1 + \|\mathcal{X}(A) - y\|_1\\
&\le \|y - \mathcal{X}(\hat{A})\|_1 + \|z\|_1 \le 2n\sigma
\end{aligned} \tag{7.40}$$

$$\begin{aligned}
&\|\tilde{\mathcal{X}}^*\tilde{\mathcal{X}}(\hat{A} - A)\| \le \|\tilde{\mathcal{X}}^*(\tilde{y} - \tilde{\mathcal{X}}(\hat{A}))\| + \|\tilde{\mathcal{X}}^*(\tilde{\mathcal{X}}(A) - \tilde{y})\|\\
=&\|\tilde{\mathcal{X}}^*(\tilde{y} - \tilde{\mathcal{X}}(\hat{A}))\| + \|\tilde{\mathcal{X}}^*(\tilde{z})\| \le 2\eta
\end{aligned} \tag{7.41}$$

Similarly as the proof to Proposition 2.1, by Theorem 2.2, there exists constant D, δ' such that if $n \ge Drp$, $\mathcal{X}_1$ satisfies RUB of order $10r$ with constants C_1, C_2 such that $C_2/C_1 < \sqrt{10}$ with probability at least $1 - \exp(-n\delta')$. Now we suppose the following two events happen,

1. $\mathcal{X}_1$ satisfies RUB of order $10r$ and constants C_1, C_2 satisfying $C_2/C_1 < \sqrt{10}$,
2. A is feasible in the programming (2.17).

Since $\tilde{\mathcal{X}}(B) = 2\mathcal{X}_1(B)$ for any symmetric matrix B, by $\mathcal{X}_1$ satisfies RUB condition, we have $\tilde{\mathcal{X}}$ satisfies RUB for symmetric matrices of order $10r$ and constants $2C_1, 2C_2$ satisfying $(2C_2)/(2C_1) < \sqrt{10}$. We note that the proof of Lemmas 7.8 and 7.9 still apply for $\tilde{\mathcal{X}}$ in the symmetric matrices class, so we can get (2.18) based on (7.40) and (7.41) under those two events happen. Finally the probability that these events happen is at least $1 - 15/n - 4\exp(-p\delta) - \exp(-n\delta)$ for $\delta < \min(2(2 - \log 7), \delta')$, which finished the proof of Proposition 2.3. □

Finally we consider Proposition 2.4. Denote $p' = \lfloor p/2 \rfloor$, $r' = \lfloor r/2 \rfloor$. By $r, p \geq 2$, we have $r' \geq r/3, p' \geq p/3$. Define a sub-class of the class rank-r symmetric matrices,

$$\mathcal{G} = \left\{ A \in \mathbb{S}^p : A = \begin{array}{c} \begin{array}{cc} p' & p-p' \end{array} \\ \begin{bmatrix} 0 & B \\ B^T & 0 \end{bmatrix} \end{array} \begin{array}{c} p' \\ p-p' \end{array}, B \in \mathbb{R}^{p' \times (p-p')}, \operatorname{rank}(B) \leq r' \right\}$$

we can see $\forall A \in \mathcal{G}$,

$$[\mathcal{X}(A)]_i = \beta^{(i)T} A \beta^{(i)} = 2(\beta_1^{(i)}, \cdots, \beta_{p'}^{(i)}) B (\beta_{p'+1}^{(i)}, \cdots, \beta_p^{(i)})^T,$$

so in $\mathcal{G}$ the SROP model becomes

$$\frac{y_i}{2} = (\beta_1^{(i)}, \cdots, \beta_{p'}^{(i)}) B (\beta_{p'+1}^{(i)}, \cdots, \beta_p^{(i)})^T + \frac{z_i}{2}, \quad \frac{z}{2} \overset{iid}{\sim} N(0, \sigma^2/4)$$

which is an ROP model which we already discussed in section 2. We omit the rest of the proof as it can be followed by the proof of Theorem 2.4. □

Proof of Proposition 3.1.. The proof can follow from the proof of Proposition 2.3 and Theorem 3.1 once we can prove that in high probability, $\mathcal{X}_1$ (defined in (7.36)) satisfies RUB condition with C_1, C_2 such that C_2/C_1 is bounded. This can be proved similarly as Proposition 7.1, where we only need to edit the proof that we use the following Lemma 7.11 instead of Lemma 7.1.

LEMMA 7.11. *Suppose $A \in \mathbb{R}^{p \times p}$ is a fixed matrix (not necessarily symmetric) and $\mathcal{X}_1$ is given by (7.36). $\beta^{(i)}$ is a set of p-dimensional vectors such*

that $\overset{iid}{\sim} \mathcal{P}$*, where* $\mathcal{P}$ *is some symmetric variance 1 sub-Gaussian distribution except Rademacher* ± 1 *distribution. Then for* $\delta > 0$*, we have*

$$\left(\frac{\min^{3/2}(\mathrm{Var}(\mathcal{P}^2)/2, 1)}{3(2\alpha_{\mathcal{P}})^4} - 8\alpha_{\mathcal{P}}^2\delta - 4\alpha_{\mathcal{P}}^2\delta^2\right)\|A\|_F \leq \frac{\|\mathcal{X}_1(A)\|_1}{\lfloor n/2 \rfloor} \tag{7.42}$$
$$\leq \left(\sqrt{3/2}\alpha_{\mathcal{P}}^2 + 8\alpha_{\mathcal{P}}^2\delta + 4\alpha_{\mathcal{P}}^2\delta^2\right)\|A\|_F.$$

with probability at least $1 - 2\exp(-\delta^2\lfloor n/2 \rfloor)$*.*

The proof of Lemma 7.11 is in the Appendix right after this paragraph. Note that provided $\mathcal{P}$ is symmetric and with variance 1, $\mathrm{Var}(\mathcal{P}^2) = 0$ if and only $\mathcal{P}$ is Rademacher ± 1 and A is diagonal, in which the lower bound of (7.42) becomes meaningless. So we only exclude Rademacher ± 1 distribution from the result. □

Proof of Lemma 7.11.. The proof of Lemma 7.11 is basically the same to Lemma 7.1. We only need to redo two parts of the proof, where there are major differences.

1. Part 1. "Step 1. Even moments of $|\frac{1}{2}(\beta^{(1)} + \beta^{(2)})^T A(\beta^{(1)} - \beta^{(2)})|$.". First, based on $\mathcal{P}$ is symmetric and with variance 1, we have $E\mathcal{P}^{2k+1} = 0$, $E\mathcal{P}^{2k} \leq \alpha^{2k}Ex^{2k} = \alpha^{2k}(2k-1)!!$.Then we can calculate that

$$E(\beta_i^{(1)} + \beta_i^{(2)})^{2k+1} = E(\beta_i^{(1)} - \beta_i^{(2)})^{2k+1} = 0$$

$$\begin{aligned}
&E(\beta_i^{(1)} + \beta_i^{(2)})^{2k} \\
=&\sum_{l=0}^{k}\binom{2k}{2l}E(\beta_i^{(1)})^{2l}E(\beta_i^{(2)})^{2(k-l)} \leq \alpha^{2k}\sum_{l=0}^{k}\binom{2k}{2l}(2l-1)!!(2(k-l)-1)!! \\
=&\alpha^{2k}\sum_{l=0}^{k}\frac{(2k-1)!!2^k k!}{2^l l! 2^{k-l}(k-l)!} = \alpha^{2k}(2k-1)!!\sum_{l=0}^{k}\binom{k}{l} = 2^k\alpha^{2k}(2k-1)!!
\end{aligned}$$

Similarly, $E(\beta_i^{(1)} - \beta_i^{(2)})^{2k} \leq 2^k\alpha^{2k}(2k-1)!!$. Next, we can similarly consider the expansion of $E\left(\frac{1}{2}(\beta^{(1)} + \beta^{(2)})A(\beta^{(1)} - \beta^{(2)})\right)^{2k}$, where the non-zero terms can be written as

$$\frac{1}{2^{2k}}\prod_{l=1}^{2k}A_{i_l,j_l}\prod_{i=1}^{p}E(\beta_i^{(1)} + \beta_i^{(2)})^{2s_i}\prod_{j=1}^{p}E(\beta_j^{(1)} - \beta_j^{(2)})^{2t_j}$$

Here $s_1 + \cdots + s_p = t_1 + \cdots + t_p = k$. This term can be bounded as

$$\begin{aligned}
&\left| \frac{1}{2^{2k}} \prod_{l=1}^{2k} A_{i_l,j_l} \prod_{i=1}^{p} E(\beta_i^{(1)} + \beta_i^{(2)})^{2s_i} (\beta_i^{(1)} - \beta_i^{(2)})^{2t_i} \right| \\
\leq & \frac{1}{2^{2k}} \prod_{l=1}^{2k} |A_{i_l,j_l}| \cdot \prod_{i=1}^{p} \left(\frac{s_i}{s_i+t_i} E\left(\beta_i^{(1)} + \beta_i^{(2)}\right)^{2s_i+2t_i} + \frac{t_i}{s_i+t_i} E\left(\beta_i^{(1)} - \beta_i^{(2)}\right)^{2s_i+2t_i} \right) \\
\leq & \frac{1}{2^{2k}} \prod_{l=1}^{2k} |A_{i_l,j_l}| \prod_{i=1}^{p} 2^{s_i+t_i} \alpha^{2(s_i+t_i)} (2(s_i+t_i)-1)!! \\
= & \alpha^{4k} \prod_{l=1}^{2k} |A_{i_l,j_l}| \prod_{i=1}^{p} \frac{(2(s_i+t_i))!}{2^{s_i+t_i}(s_i+t_i)!} \leq \alpha^{4k} \prod_{l=1}^{2k} |A_{i_l,j_l}| \prod_{i=1}^{p} \frac{((2s_i+2t_i)!!)^2}{2^{s_i+t_i} s_i! t_i!} \\
\leq & \alpha^{4k} \prod_{l=1}^{2k} |A_{i_l,j_l}| \prod_{i=1}^{p} \frac{2^{2(s_i+t_i)} \left((s_i+t_i)!\right)^2}{2^{s_i+t_i} s_i! t_i!} \leq \alpha^{4k} 2^{4p} \prod_{l=1}^{2k} |A_{i_l,j_l}| \prod_{i=1}^{p} \frac{(2s_i)!(2t_i)!}{(2s_i)!!(2t_i)!!} \\
= & \alpha^{4k} 2^{4p} \prod_{l=1}^{2k} |A_{i_l,j_l}| \prod_{i=1}^{p} E x_i^{2s_i} \cdot \prod_{i=1}^{p} E y_i^{2t_i}
\end{aligned}$$

Here we assume that $x_i, y_i \overset{iid}{\sim} N(0,1)$. The right hand side of the inequality above is exactly the term in the expansion of $(2\alpha)^{4k} E(x^T A_{abs} y)^{2k}$, where A_{abs} is the element-wise absolute value of A. Therefore, we have

$$E\left(\frac{1}{2}\left(\beta^{(1)} + \beta^{(2)}\right)^T A \left(\beta^{(1)} - \beta^{(2)}\right)\right)^{2k} \leq (2\alpha)^{4k} E[x^T A_{abs} y]^{2k}$$

Now we follow the same argument of the rest part of Step 1 in the proof of Lemma 7.1, we can prove that

(7.43)
$$E\left(\frac{1}{2}\left(\beta^{(1)} + \beta^{(2)}\right)^T A \left(\beta^{(1)} - \beta^{(2)}\right)\right)^{2k} \leq (2\alpha)^{4k} ((2k-1)!!)^2 \|A\|_F^{2k}$$

2. Part 2. The upper and lower bound of $\mu = E\left|\frac{1}{2}\left(\beta^{(1)} + \beta^{(2)}\right)^T A \left(\beta^{(1)} - \beta^{(2)}\right)\right|$. To follow the argument of the proof of Step 3 in Lemma 7.1, we need to derive a new bound for $\mu = E|\frac{1}{2}\left(\beta^{(1)} + \beta^{(2)}\right)^T A \left(\beta^{(1)} - \beta^{(2)}\right)|$. First,

we denote $M = |\frac{1}{2}\left(\beta^{(1)}+\beta^{(2)}\right)^T A\left(\beta^{(1)}-\beta^{(2)}\right)|$, then

$$\begin{aligned}
\mu =& EM \le \sqrt{EM^2} \\
=& \sqrt{E\left(\frac{1}{2}(\beta^{(1)}+\beta^{(2)})^T A(\beta^{(1)}-\beta^{(2)})\right)^2} \\
=& \sqrt{\frac{1}{4}\sum_{i,j} E(\beta_i^{(1)}+\beta_i^{(2)})^2 A_{ij}^2(\beta_j^{(1)}-\beta_j^{(2)})^2} \\
=& \sqrt{\sum_{i\neq j} A_{ij}^2 + \frac{1}{2}\mathrm{Var}(\mathcal{P}^2)\sum_i A_{ii}^2} \le \sqrt{\sum_{i\neq j} A_{ij}^2 + \frac{3}{2}\alpha_{\mathcal{P}}^4 \sum_i A_{ii}^2} \\
\le& \sqrt{\frac{3}{2}}\alpha_{\mathcal{P}}^2 \|A\|_F
\end{aligned}$$

On the other hand,

$$\begin{aligned}
EM^2 =& E\left(\frac{1}{2}(\beta^{(1)}+\beta^{(2)})^T A(\beta^{(1)}-\beta^{(2)})\right)^2 \\
=& \sum_{i\neq j} A_{ij}^2 + \frac{1}{2}\mathrm{Var}(\mathcal{P}^2)\sum_i A_{ii}^2 \ge \min(\frac{1}{2}\mathrm{Var}(\mathcal{P}^2), 1)\|A\|_F^2,
\end{aligned}$$

By (7.43), we also have $EM^4 \le 9(2\alpha)^8\|A\|_F^{2k}$. By Hölder's inequality, $EM^2 \le (EM)^{2/3}(EM^4)^{1/3}$. Hence,

$$\mu \ge \sqrt{\frac{(EM^2)^3}{EM^4}} \ge \frac{\min^{3/2}(\mathrm{Var}(\mathcal{P}^2)/2, 1)\|A\|_F}{3(2\alpha)^4}$$

To sum up, instead of Lemma 7.2, we have the bound of μ as follows,

$$\frac{\min^{3/2}(\mathrm{Var}(\mathcal{P}^2)/2, 1)\|A\|_F}{3(2\alpha_{\mathcal{P}})^4} \le \mu \le \sqrt{3/2}\alpha_{\mathcal{P}}^2\|A\|_F \tag{7.44}$$

The rest of the proof follows the proof of Lemma 7.1 with modifications of constants, which we do not go into details. □

Proof of Lemma 7.10.. Note that $\xi_i\Big|\beta^{(i)} \sim N(0, \beta^{(i)T}\Sigma\beta^{(i)})$, we can assume that

$$\xi_i^2 = \beta^{(i)T}\Sigma\beta^{(i)} \cdot Z_i, \tag{7.45}$$

where $Z_i \overset{iid}{\sim} (N(0,1))^2$ and $Z_i, \beta^{(i)}$ are independent. Based on the definition of z in (4.3), we have

$$z_i = y_i - [\mathcal{X}(\Sigma_0)]_i = \xi_i^2 - \beta^{(i)\intercal}\Sigma\beta^{(i)} = \beta^{(i)T}\Sigma\beta^{(i)}\left(Z_i - 1\right), \quad i = 1, \cdots, n. \tag{7.46}$$

We also denote

$$Q_1 = \frac{C_1}{n}\sum_{i=1}^{n}\xi_i^2, \quad Q_2 = \frac{C_2^2}{n}\sum_{i=1}^{n}\xi_i^4, \quad Q_3 = C_3 \cdot \log n \max_{1\le i\le n}\xi_i^2.$$

- We'll first consider the former part of (7.8), (7.9) and (7.10). Suppose $Z \sim (N(0,1))^2$. It is well known that the non-central m-th moment of Z is $(2m-1)!!$, so we have

$$E\left(C_1 Z - |Z-1|\right) \ge C_1 - \sqrt{E\left|Z-1\right|^2} = C_1 - \sqrt{2} \tag{7.47}$$

$$E\left(C_1 Z - |Z-1|\right)^2 \le E\left(C_1 Z\right)^2 + E\left(Z-1\right)^2 = 3C_1^2 + 2 \tag{7.48}$$

$$E\left(C_2^2 Z^2 - (Z-1)^2\right) = 3C_2^2 - 2 \tag{7.49}$$

$$E\left(C_2^2 Z^2 - (Z-1)^2\right)^2 \le C_2^4 E\left(Z\right)^4 + E\left(Z-1\right)^4 = 105C_2^4 + 60 \tag{7.50}$$

Next we consider the random quadratic form of Σ. Suppose $\beta = (\beta_1, \cdots, \beta_p) \overset{iid}{\sim} N(0,1)$, $X_1, \cdots \overset{iid}{\sim} N(0,1)$, $\lambda_1(\Sigma), \cdots, \lambda_p(\Sigma)$ are the eigenvalues of Σ. Since Σ is positive definite, we have

$$E\beta^T\Sigma\beta = \mathrm{tr}(\Sigma) = \|\Sigma\|_* \tag{7.51}$$

$$\begin{aligned} E(\beta^T\Sigma\beta)^2 =& E(\sum_i \beta_i^2\Sigma_{ii} + 2\sum_{i<j}\Sigma_{ij}\beta_i\beta_j)^2 \\ =& \sum_i \Sigma_{ii}^2 E\beta_i^4 + 2\sum_{i<j}\Sigma_{ii}\Sigma_{jj}E\beta_i^2\beta_j^2 + \sum_{i<j}4\Sigma_{ij}^2 E\beta_i^2\beta_j^2 \\ =& 2(\sum_{i,j}\Sigma_{ij}^2) + (\sum_{ii}\Sigma_i)^2 = 2\|\Sigma\|_F^2 + \|\Sigma\|_*^2 \end{aligned}$$

Hence,

$$\|\Sigma\|_*^2 \le E(\beta^T\Sigma\beta)^2 \le 3\|\Sigma\|_*^2 \tag{7.52}$$

$$\begin{aligned} E(\beta^T\Sigma\beta)^4 &= E(\sum_{i=1}^{p}\lambda_i(\Sigma)X_i^2)^4 \\ &= \sum_{1\le i,j,s,t\le p}\lambda_i(\Sigma)\lambda_j(\Sigma)\lambda_s(\Sigma)\lambda_t(\Sigma)EX_i^2X_j^2X_s^2X_t^2 \\ &\le \sum_{1\le i,j,s,t\le p}\lambda_i(\Sigma)\lambda_j(\Sigma)\lambda_s(\Sigma)\lambda_t(\Sigma)7!! = 105\|\Sigma\|_*^4 \end{aligned} \tag{7.53}$$

Then we consider $C_1\xi_i^2 - |z_i|$ and $C_2^2\xi_i^4 - z_i^2$. By (7.45) and (7.46), we have

$$C_1\xi_i^2 - |z_i| = \beta^{(i)T}\Sigma\beta^{(i)} \cdot (C_1 Z_i - |Z_i - 1|),$$
$$C_2^2\xi_i^4 - z_i^2 = \left(\beta^{(i)T}\Sigma\beta^{(i)}\right)^2 \left(C_2^2 Z^2 - (Z-1)^2\right),$$

while $\beta^{(i)}$ and Z_i are independent in the equation above. By (7.47)-(7.53), we obtain an estimation of the first and second moment of these two quantities as

$$E\left(C_1\xi_i^2 - |z_i|\right) \geq (C_1 - \sqrt{2})\|\Sigma\|_* \tag{7.54}$$

$$\begin{aligned}\operatorname{Var}\left(C_1\xi_i^2 - |z_i|\right) &\leq E\left(C_1\xi_i^2 - |z_i|\right)^2 \\ =&E\left(C_1 Z_i - |Z_i - 1|\right)^2 E\left(\beta^{(i)T}\Sigma\beta^{(i)}\right)^2 \leq \left(9C_1^2 + 6\right)\|\Sigma\|_*^2\end{aligned} \tag{7.55}$$

$$E\left(C_2^2\xi_i^4 - z_i^2\right) \geq \left(3C_2^2 - 2\right) \cdot \|\Sigma\|_*^2 \tag{7.56}$$

$$\operatorname{Var}\left(C_2^2\xi_i^4 - z_i^2\right) \leq E\left(C_2^2\xi_i^4 - z_i^2\right)^2 \leq 105(105C_2^4 + 60)\|\Sigma\|_*^4 \tag{7.57}$$

We note that $Q_1 - \|z\|_1/n$ and $Q_2 - \|z\|_2^2/n$ are the average of n i.i.d. copy of $C_1\xi_i^2 - |z_i|$ and $C_2^2\xi_i^4 - z_i^2$. We can immediately get an estimation of the mean and variance of $Q_1 - \|z\|_1/n$ and $Q_2 - \|z\|_2^2/n$ based on (7.54)-(7.57). Finally, by Chebyshev's inequality,

$$P\left(Q_1 \leq \|z\|_1/n\right) \leq \frac{\operatorname{Var}(Q_1 - \|z\|_1/n)}{\left(E\left(Q_1 - \|z\|_1/n\right)\right)^2} \leq \frac{9C_1^2 + 6}{n(C_1 - \sqrt{2})^2}$$
$$P\left(Q_2 \leq \frac{\|z\|_2^2}{n}\right) \leq \frac{\operatorname{Var}\left(Q_2 - \|z\|_2^2/n\right)}{\left(E\left(Q_2 - \|z\|_2^2/n\right)\right)^2} \leq \frac{105\left(105C_2^4 + 60\right)}{n(3C_2^2 - 2)^2}$$

Since $C_3 > 1$ and $n \geq 3$, we know $C_3 \log n Z \geq Z - 1$ with probability 1. Suppose $i_0 = \arg\max_i \beta^{(i)T}\Sigma\beta^{(i)}$, then

$$\begin{aligned}
&P\left(Q_3 \leq \|z\|_\infty\right) \\
\leq&P\left(\max_i\left(C_3\log n(\beta^{(i)T}\Sigma\beta^{(i)})Z_i\right) \leq \max_i\left((\beta^{(i)T}\Sigma\beta^{(i)})(Z_i - 1)\right)\right) \\
&+ P\left(\max_i\left(C_3\log n(\beta^{(i)T}\Sigma\beta^{(i)})Z_i\right) \leq \max_i\left((\beta^{(i)T}\Sigma\beta^{(i)})(1 - Z_i)\right)\right) \\
\leq&0 + P\left(\max_i\left(C_3\log n(\beta^{(i)T}\Sigma\beta^{(i)})Z_i\right) \leq \max_i\left(\beta^{(i)T}\Sigma\beta^{(i)}\right)\right) \\
\leq&P\left(C_3\log n\beta^{(i_0)T}\Sigma\beta^{(i_0)}Z_{i_0} \leq \beta^{(i_0)T}\Sigma\beta^{(i_0)}\right) \\
\leq&P\left(Z_{i_0} \leq \frac{1}{C_3\log n}\right) = P\left(|N(0,1)| \leq \frac{1}{\sqrt{C_3\log n}}\right) \leq \frac{2}{\sqrt{2\pi C_3\log n}}
\end{aligned}$$

- Then we consider the latter part of (7.8)-(7.10). We can do similar calculations as the first part of the proof and get

$$EQ_1 = C_1 E\xi_i^2 = C_1 E(\beta^{(i)T}\Sigma\beta^{(i)}) = C_1\|\Sigma\|_*$$

$$\begin{aligned}\mathrm{Var}(Q_1) =& \frac{C_1^2}{n}\mathrm{Var}\xi_i^2 \le \frac{C_1^2}{n}E\xi_i^4 = \frac{C_1^2}{n}E\left(\beta^{(i)T}\Sigma\beta^{(i)}\right)^2 EZ^2 \le \frac{9C_1^2}{n}\|\Sigma\|_*^2\\ EQ_2 =& C_2^2 EZ^2\cdot E\left(\beta^{(i)T}\Sigma\beta^{(i)}\right)^2 \le 9C_2^2\|\Sigma\|_*^2\\ \mathrm{Var}(Q_2) =& \frac{1}{n}\mathrm{Var}\left(C_2^2\xi_i^4\right) \le \frac{C_2^4}{n}E\xi_i^8\\ =& \frac{C_2^4}{n}EZ^4\cdot E\left(\beta^{(i)T}\Sigma\beta^{(i)}\right)^4 \le \frac{105^2C_2^4}{n}\|\Sigma\|_*^4\end{aligned}$$

So by Chebyshev's inequality,

$$P\left(Q_1 \ge M_1C_1\|\Sigma\|_*\right) \le P\left(Q_1 - EQ_1 \ge (M_1-1)C_1\|\Sigma\|_*\right) \le \frac{9}{(M_1-1)^2 n}$$

$$\begin{aligned}&P\left(Q_2 \ge M_2C_2^2\|\Sigma\|_*^2\right) \le P\left(Q_2 - EQ_2 \ge (M_2-9)C_2^2\|\Sigma\|_*^2\right)\\ \le& \frac{\mathrm{Var}(Q_2)}{(M_2-9)^2C_2^4\|\Sigma\|_*^4} \le \frac{105^2}{n(M_1-9)^2}\end{aligned}$$

which provide the latter part of (7.8) and (7.9). Finally we note that $\xi_i^2 = (\beta^{(i)T}\Sigma\beta^{(i)})Z_i$. By Lemma 1 in [27] and the fact that $\|\Sigma\|_* = \sum_i \lambda_i(\Sigma)$, $\|\Sigma\| = \max_i \lambda_i(\Sigma)$, $\|\Sigma\|_F = \sqrt{\sum_i \lambda_i^2(\Sigma)}$, $\|\Sigma\|_F \le \sqrt{\|\Sigma\|_*\cdot\|\Sigma\|}$, we have

(7.58)

$$\begin{aligned}&P\left(\beta^T\Sigma\beta \ge \left(\sqrt{\|\Sigma\|_*} + \sqrt{2M_3\log n\|\Sigma\|}\right)^2\right)\\ =&P\left(\sum_{i=1}^n \lambda_i(\Sigma)X_i^2 \ge \|\Sigma\|_* + 2\sqrt{2M_3\log n\|\Sigma\|_*\|\Sigma\|} + 2M_3\log n\|\Sigma\|\right)\\ \le&P\left(\sum_{i=1}^n \lambda_i(\Sigma)(X_i^2-1) \ge 2\sqrt{2M_3\log n\sum_{i=1}^n\lambda_i^2(\Sigma)} + 2M_3\log n\max_i\lambda_i(\Sigma)\right)\\ \le& n^{-M_3}\end{aligned}$$

and

$$P\left(Z \ge 2M_3\log n\right) \le \exp(-2M_3\log n/2) = n^{-M_3}.$$

Hence,

$$P\left(C_3 \log n \xi_i^2 \geq 2C_3 M_3 \log^2 n \left(\sqrt{\|\Sigma\|_*} + \sqrt{2M_3 \log n \|\Sigma\|}\right)^2\right) \leq 2n^{-M_3},$$

and consequently

$$P\left(C_3 \log n \max_{1\leq i\leq n} \xi_i^2 \geq 2C_3 M_3 \log^2 n \left(\sqrt{\|\Sigma\|_*} + \sqrt{2M_3 \log n \|\Sigma\|}\right)^2\right) \leq 2n^{-M_3+1},$$

which gives the right side of (7.10). □

Department of Statistics
The Wharton School
University of Pennsylvania
Philadelphia, PA, 19104.
E-mail: tcai@wharton.upenn.edu
URL: http://www-stat.wharton.upenn.edu/~tcai/
E-mail: anrzhang@wharton.upenn.edu
URL: http://www-stat.wharton.upenn.edu/~anrzhang/