Math. Program., Ser. A (2016) 159:289–338
DOI 10.1007/s10107-015-0961-7

FULL LENGTH PAPER

A rank-corrected procedure for matrix completion with fixed basis coefficients

Weimin Miao[1] · **Shaohua Pan**[2] · **Defeng Sun**[3]

Received: 10 April 2014 / Accepted: 13 October 2015 / Published online: 30 October 2015

Abstract For the problems of low-rank matrix completion, the efficiency of the widely-used nuclear norm technique may be challenged under many circumstances, especially when certain basis coefficients are fixed, for example, the low-rank correlation matrix completion in various fields such as the financial market and the low-rank density matrix completion from the quantum state tomography. To seek a solution of high recovery quality beyond the reach of the nuclear norm, in this paper, we propose a rank-corrected procedure using a nuclear semi-norm to generate a new estimator. For this new estimator, we establish a non-asymptotic recovery error bound. More importantly, we quantify the reduction of the recovery error bound for this rank-corrected procedure. Compared with the one obtained for the nuclear norm penalized least squares estimator, this reduction can be substantial (around 50 %). We also pro-

W. Miao author's research is supported in part by Willis Research Network.
D. Sun author's research is supported in part by Academic Research Fund under Grant R-146-000-149-112.
S. Pan author's research is supported in part by National Natural Science Foundation of China under project No. 11571120.

✉ Weimin Miao
miaoweimin@nus.edu.sg

Shaohua Pan
shhpan@scut.edu.cn

Defeng Sun
matsundf@nus.edu.sg

1 Risk Management Institute, National University of Singapore, 21 Heng Mui Keng Terrace, Singapore 119613, Singapore

2 Department of Mathematics, South China University of Technology, Tianhe District of Guangzhou City, China

3 Department of Mathematics and Risk Management Institute, National University of Singapore, 10 Lower Kent Ridge Road, Singapore 119076, Singapore

vide necessary and sufficient conditions for rank consistency in the sense of Bach (J Mach Learn Res 9:1019–1048, 2008). Very interestingly, these conditions are highly related to the concept of constraint nondegeneracy in matrix optimization. As a byproduct, our results provide a theoretical foundation for the majorized penalty method of Gao and Sun (A majorized penalty approach for calibrating rank constrained correlation matrix problems. http://www.math.nus.edu.sg/~matsundf/MajorPen_May5.pdf, 2010) and Gao (2010) for structured low-rank matrix optimization problems. Extensive numerical experiments demonstrate that our proposed rank-corrected procedure can simultaneously achieve a high recovery accuracy and capture the low-rank structure.

Keywords Matrix completion · Fixed basis coefficients · Low-rank · Convex optimization · Rank consistency · Constraint nondegeneracy

Mathematics Subject Classification 90C90

1 Introduction

The low-rank matrix completion is to recover an unknown low-rank matrix from the under-sampled observations with or without noises. This problem is of considerable interest in many application areas, from machine learning to quantum state tomography. A basic idea to address a low-rank matrix completion problem is to minimize the rank of a matrix subject to certain constraints from observations. Since the direct minimization of rank function is generally NP-hard, a widely-used convex relaxation approach is to replace the rank function with the nuclear norm—the convex envelope of the rank function over a unit ball of the spectral norm [19].

The nuclear norm technique has been observed to provide a low-rank solution in practice for a long time (see, e.g., [19,54,55]). The first remarkable theoretical characterization for the minimum rank solution via the nuclear norm minimization was given by Recht et al. [64], with the help of the concept of restricted isometric property (RIP). Recognizing that the matrix completion problem does not obey the RIP, Candès and Recht [8] introduced the concept of incoherence property and proved that most low-rank matrices can be exactly recovered from a surprisingly small number of noiseless observations of randomly sampled entries via the nuclear norm minimization. The bound of the number of sampled entries was later improved to be near-optimal by Candès and Tao [9] through a counting argument. Such a bound was also obtained by Keshavan et al. [37] for their proposed OptSpace algorithm. Later, Gross [30] sharpened the bound by employing a novel technique from quantum information theory developed in [31], in which noiseless observations were extended from entries to coefficients relative to an arbitrary basis. This technique was also adapted by Recht [63], leading to a short and intelligible analysis. Besides the above results for the noiseless case, matrix completion with noise was first addressed by Candès and Plan [7]. More recently, nuclear norm penalized estimators for matrix completion with noise have been well studied by Koltchinskii et al. [44], Negahban and Wainwright [58], and Klopp [40] under different settings. Besides the nuclear norm, estimators with other penalties for matrix completion have also been considered in terms of recoverability in the literature, e.g., [25,39,43,68,70].

The nuclear norm technique has been demonstrated to be a successful approach to encourage a low-rank solution for matrix completion. However, its efficiency may be challenged in some circumstances. For example, Salakhutdinov and Srebro [69] showed that when certain rows and/or columns are sampled with high probability, the nuclear norm minimization may fail in the sense that the number of observations required for recovery is much more than the setting of most matrix completion problems. It means that the efficiency of the nuclear norm techniques could be highly weakened under a general sampling scheme. Negahban and Wainwright [58] also pointed out the impact of such heavy sampling schemes on the recovery error bound. As a remedy for this, a weighted nuclear norm (trace norm), based on row- and column-marginals of the sampling distribution, was suggested in [24,58,69] if the prior information on sampling distribution is available. Moreover, the conditions characterized by Bach [3] for rank consistency of the nuclear norm penalized least squares estimator may not be satisfied, especially when certain constraints are involved.

A concrete example of interest is to recover a density matrix of a quantum system from Pauli measurements in quantum state tomography (see, e.g., [22,31,74]). A density matrix is a Hermitian positive semidefinite matrix of trace one. Clearly, if the constraints of positive semidefiniteness and trace one are simultaneously imposed on the nuclear norm minimization, the nuclear norm completely fails in promoting a low-rank solution. Thus, one of the two constraints has to be abandoned in the nuclear norm minimization and then be restored in the post-processing stage. In fact, this idea has been much explored in [22,31] and the numerical results there indicated its relative efficiency though it still has much room for improvement.

All the above examples motivate us to ask whether it is possible to go beyond the nuclear norm approach for practical use to seek for better performance in low-rank matrix completion. In this paper, we provide a positive answer to this question with both theoretical and empirical supports. We first establish a unified low-rank matrix completion model, which allows for the imposition of fixed basis coefficients so that the correlation and the density matrix completion are included as special cases. It means that in our setting, for any given basis of the matrix space, a few basis coefficients of the true matrix are assumed to be fixed due to a certain structure or some prior information, and the rest are allowed to be observed with noises under a general sampling scheme. To pursue a low-rank solution with a high recovery accuracy, we propose a rank-correction step to generate a new estimator. The rank-correction step solves a penalized least squares problem with its penalization being the nuclear norm minus a linear rank-correction term constructed on a reasonable initial estimator. A satisfactory choice of the initial estimator could be the nuclear norm penalized least squares estimator or one of its analogies. The resulting convex matrix optimization problem can be solved by the efficient algorithms recently developed in [21,34–36] even for large-scale cases.

The idea of using a two-stage or even multi-stage procedure is not brand new for dealing with sparse recovery in the statistical and machine learning literature. The l_1-norm penalized least squares method, also known as the Lasso [71], is very attractive and popular for variable selection in statistics, thanks to the invention of the fast and efficient LARS algorithm [12]. On the other hand, the l_1-norm penalty has long been known by statisticians to yield biased estimators and cannot achieve the best

estimation performance [14,18]. The issue of bias can be overcome by nonconvex penalization methods, see, e.g., [13,47,77]. A multi-stage procedure naturally occurs if the nonconvex problem obtained is solved by an iterative algorithm [45,81]. In particular, once a good initial estimator is used, a two-stage estimator is enough to achieve the desired asymptotic efficiency, e.g., the adaptive Lasso proposed by Zou [80]. There are also a number of important works along this line on variable selection, including [15,33,47,52,53,78,79], to name only a few. For a broad overview, the interested readers are referred to the recent survey papers [16,17]. It is natural to extend the ideas from the vector case to the matrix case. Fazel et al. [20] first proposed the reweighted trace minimization for minimizing the rank of a positive semidefinite matrix. In [3], Bach made an important step in extending the adaptive Lasso of Zou [80] to the matrix case for rank consistency. However, it is not clear how to apply Bach's idea to our matrix completion model with fixed basis coefficients since the required rate of convergence of the initial estimator for achieving asymptotic properties is no longer valid, as far as we can see. More critically, there are numerical difficulties in efficiently solving the resulting optimization problems. Numerical difficulties also occur in the reweighted nuclear norm approach proposed by Mohan and Fazel [56] as an extension of [20] for rectangular matrices. Iterative reweighted least squares minimization is an alternative extension of [20] independently proposed by Mohan and Fazel [57] and Fornasier et al. [23], taking advantage of the property that the rank of a matrix is equal to the rank of the product of this matrix and its transpose. However, the resulting smoothness of inner-iteration subproblems is weak in encouraging a low-rank solution so much more iterations are needed in general and thus the computational cost is high especially when hard constraints such as fixed basis coefficients are involved.

The rank-correction step to be proposed in this paper is for overcoming the above difficulties. This approach is inspired by the majorized penalty method proposed by Gao and Sun [27] for solving structured matrix optimization problems with a low-rank constraint. For our proposed rank-correction step, we establish a non-asymptotic recovery error bound in Frobenius norm, following a similar argument adopted by Klopp in [40]. We also discuss the impact of adding the rank-correction term on recovery error. More importantly, we provide an affirmative guarantee that under mild condition the rank-correction step highly improves the recoverability, compared with the nuclear norm penalized least squares estimator. As the estimator is expected to be of low-rank, we also study the asymptotic property—rank consistency in the sense of Bach [3], under the setting that the matrix size is assumed to be fixed. This setting may not be ideal for analyzing asymptotic properties for matrix completion, but it does allow us to take the crucial first step to gain insights into the limitation of the nuclear norm penalization. Among others, the concept of constraint nondegeneracy for conic optimization problem plays a key role in our analysis. Interestingly, our results of recovery error bound and rank consistency suggest a consistent criterion for constructing a suitable rank-correction function. In particular, for the correlation and the density matrix completion problems, we prove that rank consistency automatically holds for a broad selection of rank-correction functions. For most cases, a single rank-correction step is sufficient for a substantial improvement, unless the sample ratio is rather low so that the rank-correction step may be iteratively used for two or three times to achieve the limit of improvement. Owing to this property, the advantage of

our proposed method is more apparent in practical computations especially when fixed basis coefficients are involved. Finally, we remark that our results can also be used to provide a theoretical foundation in the statistical setting for the majorized penalty method of Gao and Sun [27] and Gao [26] for structured low-rank matrix optimization problems.

This paper is organized as follows. In Sect. 2, we introduce the observation model of matrix completion with fixed basis coefficients and formulate the rank-correction step. In Sect. 3, we establish a non-asymptotic recovery error bound for the estimator generated from the rank-correction step and provide a quantification of the improvement in recoverability. Section 4 provides necessary and sufficient conditions for rank consistency. Section 5 is devoted to the construction of the rank-correction function. In Sect. 6, we report numerical results to validate the efficiency of our proposed rank-corrected procedure. We conclude this paper in Sect. 7. All relevant material and all proofs of theorems are left in the appendices.

Notation Here we provide a brief summary of the notation used in this paper.

- Let $\mathbb{R}^{n_1\times n_2}$ and $\mathbb{C}^{n_1\times n_2}$ denote the space of all $n_1\times n_2$ real and complex matrices, respectively. Let $\mathcal{S}^n$ ($\mathcal{S}^n_+$, $\mathcal{S}^n_{++}$) denote the set of all $n\times n$ real symmetric (positive semidefinite, positive definite) matrices and $\mathcal{H}^n$ ($\mathcal{H}^n_+$, $\mathcal{H}^n_{++}$) denote the set of all $n\times n$ Hermitian (positive semidefinite, positive definite) matrices.
- Let $\mathbb{V}^{n_1\times n_2}$ represent $\mathbb{R}^{n_1\times n_2}$, $\mathbb{C}^{n_1\times n_2}$, $\mathcal{S}^n$ or $\mathcal{H}^n$. We define $n:=\min(n_1,n_2)$ for the previous two cases and stipulate $n_1=n_2=n$ for the latter two cases. Let $\mathbb{V}^{n_1\times n_2}$ be endowed with the trace inner product $\langle\cdot,\cdot\rangle$ and its induced norm $\|\cdot\|_F$, i.e., $\langle X,Y\rangle:=\mathrm{Re}\big(\mathrm{Tr}(X^{\mathbb{T}}Y)\big)$ for $X,Y\in\mathbb{V}^{n_1\times n_2}$, where "Tr" stands for the trace of a matrix and "Re" means the real part of a complex number.
- For the real case, i.e., $\mathbb{V}^{n_1\times n_2}=\mathbb{R}^{n_1\times n_2}$ or $\mathbb{V}^{n_1\times n_2}=\mathcal{S}^n$, let $\mathbb{S}^n$ ($\mathbb{S}^n_+$, $\mathbb{S}^n_{++}$) represent $\mathcal{S}^n$ ($\mathcal{S}^n_+$, $\mathcal{S}^n_{++}$); and for the complex case, i.e., $\mathbb{V}^{n_1\times n_2}=\mathbb{C}^{n_1\times n_2}$ or $\mathbb{V}^{n_1\times n_2}=\mathcal{H}^n$, let $\mathbb{S}^n$ ($\mathbb{S}^n_+$, $\mathbb{S}^n_{++}$) represent $\mathcal{H}^n$ ($\mathcal{H}^n_+$, $\mathcal{H}^n_{++}$).
- For the real case, $\mathbb{O}^{n\times k}$ denotes the set of all $n\times k$ real matrices with orthonormal columns, and for the complex case, $\mathbb{O}^{n\times k}$ denotes the set of all $n\times k$ complex matrices with orthonormal columns. When $k=n$, we write $\mathbb{O}^{n\times k}$ as $\mathbb{O}^n$ for short.
- The notation $^{\mathbb{T}}$ denotes the transpose for the real case and the conjugate transpose for the complex case. The notation * means the adjoint of a linear operator.
- For any index set π, let $|\pi|$ denote the cardinality of π, i.e., the number of elements in π. For any $x\in\mathbb{R}^n$, let $|x|$ denote the vector in $\mathbb{R}^n_+$ whose i-th component is $|x_i|$, let x_+ denote the vector in $\mathbb{R}^n_+$ whose i-th component is $\max(x_i,0)$ and let x_- denote the vector in $\mathbb{R}^n_+$ whose i-th component is $\min(-x_i,0)$.
- For any given vector x, $\mathrm{Diag}(x)$ denotes a rectangular diagonal matrix of suitable size with the i-th diagonal entry being x_i.
- For any $x\in\mathbb{R}^n$, let $\|x\|_2$ and $\|x\|_\infty$ denote the Euclidean norm and the maximum norm, respectively. For any $X\in\mathbb{V}^{n_1\times n_2}$, let $\|X\|$ and $\|X\|_*$ denote the spectral norm and the nuclear norm, respectively.
- The notations $\xrightarrow{a.s.}$, $\xrightarrow{p}$ and $\xrightarrow{d}$ mean almost sure convergence, convergence in probability and convergence in distribution, respectively. We write $x_m=O_p(1)$ if x_m is bounded in probability.

- For any set K, let $\delta_K(x)$ denote the indicator function of K, i.e., $\delta_K(x) = 0$ if $x \in K$, and $\delta_K(x) = +\infty$ otherwise. Let I_n denote the $n \times n$ identity matrix.

2 Problem formulation

In this section, we formulate the model of the matrix completion problem with fixed basis coefficients, and then propose an adaptive nuclear semi-norm penalized least squares estimator for solving this class of problems.

2.1 The observation model

Let $\{\Theta_1, \ldots, \Theta_d\}$ be a given orthonormal basis of the given real inner product space $\mathbb{V}^{n_1 \times n_2}$. Then, any matrix $X \in \mathbb{V}^{n_1 \times n_2}$ can be uniquely expressed in the form of $X = \sum_{k=1}^{d} \langle \Theta_k, X \rangle \Theta_k$, where $\langle \Theta_k, X \rangle$ is called the basis coefficient of X relative to Θ_k. Throughout this paper, let $\overline{X} \in \mathbb{V}^{n_1 \times n_2}$ be the unknown low-rank matrix to be recovered and let $\mathrm{rank}(\overline{X}) = r$. In some practical applications, for example, the correlation and density matrix completion, a few basis coefficients of the unknown matrix $\overline{X}$ are fixed (or assumed to be fixed) due to a certain structure or reliable prior information. We let $\alpha \subseteq \{1, 2, \ldots, d\}$ denote the set of the indices relative to which the basis coefficients are fixed, and β denote the complement of α in $\{1, 2, \ldots, d\}$, i.e., $\alpha \cap \beta = \emptyset$ and $\alpha \cup \beta = \{1, \ldots, d\}$. We define $d_1 := |\alpha|$ and $d_2 := |\beta|$.

When a few basis coefficients are fixed, one only needs to observe the rest for recovering the unknown matrix $\overline{X}$. Assume that we are given a collection of m noisy observations of the basis coefficients relative to $\{\Theta_k : k \in \beta\}$ in the following form

$$y_i = \left\langle \Theta_{\omega_i}, \overline{X} \right\rangle + \nu \xi_i, \quad i = 1, \ldots, m, \tag{1}$$

where ω_i are the indices randomly sampled from the index set β, ξ_i are the independent and identically distributed (i.i.d.) noises with $\mathbb{E}(\xi_i) = 0$ and $\mathbb{E}(\xi_i^2) = 1$, and $\nu > 0$ controls the magnitude of noise. Unless otherwise stated, we assume a general weighted sampling (with replacement) scheme with the sampling distributions of ω_i as follows.

Assumption 1 The indices $\omega_1, \ldots, \omega_m$ are i.i.d. copies of a random variable ω that has a probability distribution Π over $\{1, \ldots, d\}$ defined by

$$\Pr(\omega = k) = \begin{cases} 0 & \text{if } k \in \alpha, \\ p_k > 0 & \text{if } k \in \beta. \end{cases}$$

Note that each $\Theta_k, k \in \beta$ is assumed to be sampled with a positive probability in this sampling scheme. In particular, when the sampling probability of all $k \in \beta$ are equal, i.e., $p_k = 1/d_2 \; \forall\, k \in \beta$, we say that the observations are sampled uniformly at random.

For notational simplicity, let Ω be the multiset of all the sampled indices from the index set β, i.e., $\Omega = \{\omega_1, \ldots, \omega_m\}$. With a slight abuse on notation, we define the sampling operator $\mathcal{R}_\Omega \colon \mathbb{V}^{n_1 \times n_2} \to \mathbb{R}^m$ associated with Ω by

$$\mathcal{R}_\Omega(X) := \big(\langle \Theta_{\omega_1}, X\rangle, \ldots, \langle \Theta_{\omega_m}, X\rangle\big)^{\mathbb{T}}, \quad X \in \mathbb{V}^{n_1\times n_2}.$$

Then, the observation model (1) can be expressed in the following vector form

$$y = \mathcal{R}_\Omega(\overline{X}) + \nu\xi, \tag{2}$$

where $y = (y_1, \ldots, y_m)^{\mathbb{T}} \in \mathbb{R}^m$ and $\xi = (\xi_1, \ldots, \xi_m)^{\mathbb{T}} \in \mathbb{R}^m$ denote the observation vector and the noise vector, respectively.

Next, we present some examples of low-rank matrix completion problems in the above settings.

(1) *Correlation matrix completion* A correlation matrix is an $n \times n$ real symmetric or Hermitian positive semidefinite matrix with all diagonal entries being ones. Let e_i be the vector with the i-th entry being one and the others being zeros. Then, $\langle e_i e_i^{\mathbb{T}}, \overline{X}\rangle = \overline{X}_{ii} = 1 \; \forall\, 1 \le i \le n$. The recovery of a correlation matrix is based on the observations of entries. For the real case, $\mathbb{V}^{n_1\times n_2} = \mathcal{S}^n$, $d = n(n+1)/2$, $d_1 = n$,

$$\Theta_\alpha = \big\{e_i e_i^{\mathbb{T}} \mid 1 \le i \le n\big\} \quad \text{and} \quad \Theta_\beta = \left\{ \frac{1}{\sqrt{2}}(e_i e_j^{\mathbb{T}} + e_j e_i^{\mathbb{T}}) \;\middle|\; 1 \le i < j \le n \right\};$$

and for the complex case, $\mathbb{V}^{n_1\times n_2} = \mathcal{H}^n$, $d = n^2$, $d_1 = n$,

$$\Theta_\alpha = \big\{e_i e_i^{\mathbb{T}} \mid 1 \le i \le n\big\} \text{ and}$$
$$\Theta_\beta = \left\{ \frac{1}{\sqrt{2}}(e_i e_j^{\mathbb{T}} + e_j e_i^{\mathbb{T}}), \frac{\sqrt{-1}}{\sqrt{2}}(e_i e_j^{\mathbb{T}} - e_j e_i^{\mathbb{T}}) \;\middle|\; i < j \right\}.$$

Here, $\sqrt{-1}$ represents the imaginary unit. Of course, one may fix some off-diagonal entries in specific applications.

(2) *Density matrix completion* A density matrix of dimension $n = 2^l$ for some positive integer l is an $n \times n$ Hermitian positive semidefinite matrix with trace one. In quantum state tomography, one aims to recover a density matrix from Pauli measurements (observations of the coefficients relative to the Pauli basis) [22,31], given by

$$\Theta_\alpha = \left\{ \frac{1}{\sqrt{n}} I_n \right\} \text{ and } \Theta_\beta = \left\{ \frac{1}{\sqrt{n}}(\sigma_{s_1} \otimes \cdots \otimes \sigma_{s_l}) \;\middle|\; (s_1, \ldots, s_l) \in \{0,1,2,3\}^l \right\} \Big\backslash \Theta_\alpha,$$

where "$\otimes$" means the Kronecker product of two matrices and

$$\sigma_0 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad \sigma_1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad \sigma_2 = \begin{pmatrix} 0 & -\sqrt{-1} \\ \sqrt{-1} & 0 \end{pmatrix}, \quad \sigma_3 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$$

are the Pauli matrices. In this setting, $\mathbb{V}^{n_1\times n_2} = \mathcal{H}^n$, $\mathrm{Tr}(\overline{X}) = \langle I_n, \overline{X}\rangle = 1$, $d = n^2$, and $d_1 = 1$.

(3) *Rectangular matrix completion* Assume that a few entries of a rectangular matrix are known and let $\mathcal{I}$ be the index set of these entries. One aims to recover this rectangular matrix from the observations of the rest entries. For the real case, $\mathbb{V}^{n_1\times n_2} = \mathbb{R}^{n_1\times n_2}$, $d = n_1n_2$, $d_1 = |\mathcal{I}|$,

$$\Theta_\alpha = \big\{e_ie_j^{\mathbb{T}} \mid (i,j) \in \mathcal{I}\big\} \quad \text{and} \quad \Theta_\beta = \big\{e_ie_j^{\mathbb{T}} \mid (i,j) \notin \mathcal{I}\big\};$$

and for the complex case, $\mathbb{V}^{n_1\times n_2} = \mathbb{C}^{n_1\times n_2}$, $d = 2n_1n_2$, $d_1 = 2|\mathcal{I}|$,

$$\Theta_\alpha = \big\{e_ie_j^{\mathbb{T}}, \sqrt{-1}e_ie_j^{\mathbb{T}} \mid (i,j) \in \mathcal{I}\big\} \quad \text{and} \quad \Theta_\beta = \big\{e_ie_j^{\mathbb{T}}, \sqrt{-1}e_ie_j^{\mathbb{T}} \mid (i,j) \notin \mathcal{I}\big\}.$$

Now we introduce some linear operators that are frequently used in the subsequent sections. For any given index set $\pi \subseteq \{1,\ldots,d\}$, say α or β, we define the linear operators $\mathcal{R}_\pi\colon \mathbb{V}^{n_1\times n_2} \to \mathbb{R}^{|\pi|}$, $\mathcal{P}_\pi\colon \mathbb{V}^{n_1\times n_2} \to \mathbb{V}^{n_1\times n_2}$ and $\mathcal{Q}_\pi\colon \mathbb{V}^{n_1\times n_2} \to \mathbb{V}^{n_1\times n_2}$ respectively, by

$$\mathcal{R}_\pi(X) := \big(\langle\Theta_k, X\rangle\big)_{k\in\pi}^{\mathbb{T}}, \quad \mathcal{P}_\pi(X) := \sum_{k\in\pi}\langle\Theta_k, X\rangle\Theta_k \quad \text{and} \quad \mathcal{Q}_\pi(X) := \sum_{k\in\pi} p_k\langle\Theta_k, X\rangle\Theta_k.$$

For convenience of discussions, in the rest of this paper, for any given $X \in \mathbb{V}^{n_1\times n_2}$, we denote by $\sigma(X) = \big(\sigma_1(X),\ldots,\sigma_n(X)\big)^{\mathbb{T}}$ the singular value vector of X arranged in the nonincreasing order and define

$$\mathbb{O}^{n_1,n_2}(X) := \big\{(U,V) \in \mathbb{O}^{n_1}\times\mathbb{O}^{n_2} \mid X = U\mathrm{Diag}\big(\sigma(X)\big)V^{\mathbb{T}}\big\}.$$

In particular, when $\mathbb{V}^{n_1\times n_2} = \mathbb{S}^n$, we denote by $\lambda(X) = \big(\lambda_1(X),\ldots,\lambda_n(X)\big)^{\mathbb{T}}$ the eigenvalue vector of X with $|\lambda_1(X)| \ge \cdots \ge |\lambda_n(X)|$ and define

$$\mathbb{O}^n(X) := \big\{P \in \mathbb{O}^n \mid X = P\mathrm{Diag}(\lambda(X))P^{\mathbb{T}}\big\}.$$

For any $X \in \mathbb{V}^{n_1\times n_2}$ and any $(U,V) \in \mathbb{O}^{n_1,n_2}(X)$, we write $U = [U_1\ U_2]$ and $V = [V_1\ V_2]$ with $U_1 \in \mathbb{O}^{n_1\times r}$, $U_2 \in \mathbb{O}^{n_1\times(n_1-r)}$, $V_1 \in \mathbb{O}^{n_2\times r}$ and $V_2 \in \mathbb{O}^{n_2\times(n_2-r)}$. In particular, for any $X \in \mathbb{S}_+^n$ and any $P \in \mathbb{O}^n(X)$, we write $P = [P_1\ P_2]$ with $P_1 \in \mathbb{O}^{n\times r}$ and $P_2 \in \mathbb{O}^{n\times(n-r)}$.

2.2 The rank-correction step

In many situations, the nuclear norm penalization performs well for matrix recovery, but its efficiency may be challenged if the observations are sampled at random obeying a general distribution such as the one considered in [69]. The setting of fixed basis coefficients in our matrix completion model can also be regarded to be under an extreme sampling scheme. In particular, for the correlation and density matrix completion, the nuclear norm completely loses its efficiency since it reduces to a constant in these two cases. In order to overcome the shortcomings of the nuclear norm penalization, we

propose a rank-correction step to generate an estimator in pursuit of a better recovery performance.

Recall that $\overline{X}$ is the unknown true matrix of rank r. Given an initial estimator $\widetilde{X}_m$ of $\overline{X}$, say, the nuclear norm penalized least squares estimator or one of its analogies, our proposed rank-correction step is to solve the convex optimization problem

$$\begin{aligned} \widehat{X}_m \ \in \ \underset{X\in\mathbb{V}^{n_1\times n_2}}{\arg\min} \ \frac{1}{2m}\|y-\mathcal{R}_\Omega(X)\|_2^2+\rho_m\big(\|X\|_*-\langle F(\widetilde{X}_m),X\rangle\big) \\ \text{s.t. } \mathcal{R}_\alpha(X)=\mathcal{R}_\alpha(\overline{X}), \quad \|\mathcal{R}_\beta(X)\|_\infty\le b, \quad X\in\mathcal{C}, \end{aligned} \tag{3}$$

where $\rho_m > 0$ is the penalty parameter (depending on the number of observations), b is an upper bound of the magnitudes of basis coefficients of $\overline{X}$, $\mathcal{C}\subseteq\mathbb{V}^{n_1\times n_2}$ is a closed convex set that contains $\overline{X}$, and $F:\mathbb{V}^{n_1\times n_2}\to\mathbb{V}^{n_1\times n_2}$ is a spectral operator associated with a symmetric function $f:\mathbb{R}^n\to\mathbb{R}^n$. One may refer to "Appendix 1" for more information on the concept of spectral operators. (Indeed, based on the subsequent analysis for better recovery performance, the choice $f:\mathbb{R}^n\to[0,1]^n$ is much preferred, for which the penalization $\|X\|_*-\langle F(\widetilde{X}_m),X\rangle$ is indeed a nuclear semi-norm. But this choice criterion is not compulsory). The bound restriction is very mild since such a bound is often available in applications, for example, the correlation and the density matrix completion. This boundedness setting can also be found in previous works done by Negahban and Wainwright [58] and Klopp [40].

Hereafter, we call F the rank-correction function and $\langle F(\widetilde{X}_m),X\rangle$ the rank-correction term. Note that, when $F\equiv 0$, the rank-correction step (3) reduces to the nuclear norm penalized least squares estimator, which equally penalizes singular values to promote a low-rank solution for matrix completion. Certainly, for this purpose, penalizing more on small singular values or even directly penalizing the rank function could serve better, but only theoretically rather than practically, due to the lack of convexity. Also note that an initial estimation, if deviates not too much from the true matrix, could contain some information of the singular values and/or the rank of the true matrix to a certain extent. Therefore, provided such an initial estimator is available, it is achievable to construct a rank-correction term with a suitable F to substantially offset the penalization of large singular values from the nuclear norm penalty. Consequently, we can expect the rank-correction step (3) to have a better low-rank promoting ability and outperform the nuclear norm penalized least squares estimator.

The key issue is then how to construct a favored rank-correction function F. In the next two sections, we provide theoretical supports to our proposed rank-correction step, from which some important guidelines on the construction of F can be captured. In particular, if one chooses the nuclear norm penalized least squares estimator to be the initial estimator $\widetilde{X}_m$, and also suitably chooses the spectral operator F so that $\|X\|_*-\langle F(\widetilde{X}_m),X\rangle$ is a semi-norm, called nuclear semi-norm, then the estimator $\widehat{X}_m$ generated from this two-stage procedure is called the adaptive nuclear semi-norm penalized least squares estimator associated with F.

2.3 Relation with the majorized penalty approach

The rank-correction step above is inspired by the majorized penalty approach proposed by Gao and Sun [27] for solving the rank constrained matrix optimization problem:

$$\min_{X\in\mathcal{C}} \big\{h(X) : \ \mathrm{rank}(X) \le r\big\}, \tag{4}$$

where $r \ge 1$, $h : \mathbb{V}^{n_1\times n_2} \to \mathbb{R}$ is a given continuous function and $\mathcal{C} \in \mathbb{V}^{n_1\times n_2}$ is a closed convex set. Note that for any $X \in \mathbb{V}^{n_1\times n_2}$, the constraint $\mathrm{rank}(X) \le r$ is equivalent to

$$0 = \sigma_{r+1}(X) + \cdots + \sigma_n(X) = \|X\|_* - \|X\|_{(r)},$$

where $\|X\|_{(r)} := \sigma_1(X) + \cdots + \sigma_r(X)$ denotes the Ky Fan r-norm. The central idea of the majorized penalty approach is to solve the following penalized version of (4):

$$\min_{X\in\mathcal{C}} \ h(X) + \rho\big(\|X\|_* - \|X\|_{(r)}\big),$$

where $\rho > 0$ is the penalty parameter. With the current iterate X^k, the majorized penalty approach yields the next iterate X^{k+1} by solving the convex optimization problem

$$\min_{X\in\mathcal{C}} \ \widehat{h}^k(X) + \rho\big(\|X\|_* - \langle G^k, X\rangle\big), \tag{5}$$

where G^k is a subgradient of the convex function $\|X\|_{(r)}$ at X^k, and $\widehat{h}^k$ is a convex majorization function of h at X^k. By comparing with (3), one may notice that our proposed rank-correction step is close to a single step of the majorized penalty approach.

Note that the rank constrained least squares problem is of great consideration in matrix completion especially when the rank information is known. However, different from the noiseless case, for matrix completion with noise, the solution to the rank constrained least squares problem (assuming the uniqueness) is in general not the true matrix though quite close to it. Indeed, there may exist many candidate matrices surrounding the true matrix and having its rank. The rank constrained least squares solution is only one of them. It deviates the least from the noisy observations rather than the true matrix. Naturally, it is conceivable that some candidate matrices may deviate a bit more from the noisy observations but less from the true matrix. So, for the purpose of matrix completion, there is no need to aim precisely at the rank constrained least squares solution and find this solution accurately. An approach roughly towards it such as our proposed rank-correction step (3) is good enough to bring similar good recovery performance.

3 Error bounds

In this section, we aim to derive a recovery error bound in Frobenius norm for the estimator generated from the rank-correction step (3) and discuss the impact of the

rank-correction term on the resulting bound. The analysis mainly follows Klopp's arguments in [40], which is also in line with those used by Negahban and Wainwright [58].

We start the analysis by defining a quantity, which plays a key role in the subsequent analysis, as

$$a_m := \frac{1}{\sqrt{r}}\|F(\widetilde{X}_m) - \overline{U}_1\overline{V}_1^{\mathbb{T}}\|_F. \tag{6}$$

A basic relation between the true matrix $\overline{X}$ and its estimate $\widehat{X}_m$ can be obtained by using the optimality of $\widehat{X}_m$ to the problem (3) as follows.

Theorem 1 *For any $\kappa > 1$, if $\rho_m \geq \kappa\nu\left\|\frac{1}{m}\mathcal{R}_\Omega^*(\xi)\right\|$, then the following inequality holds:*

$$\frac{1}{2m}\left\|\mathcal{R}_\Omega(\widehat{X}_m - \overline{X})\right\|_2^2 \leq \left(\frac{\sqrt{2}}{\kappa} + a_m\right)\rho_m\sqrt{r}\|\widehat{X}_m - \overline{X}\|_F. \tag{7}$$

We emphasize that κ is not restricted to be a constant in Theorem 1 but could be set to depend on the size of matrix. This realization is important as can be seen in the sequel. According to Theorem 1, the choice of the penalty parameter ρ_m depends on the observation noises ξ_i and the sampling operator $\mathcal{R}_\Omega$. Therefore, we make the following assumption on the noises ξ_i as follows:

Assumption 2 The i.i.d. noise variables ξ_i are sub-exponential, i.e., there exist positive constants c_1, c_2 and c_3 such that for all $t > 0$, $\Pr(|\xi_i| \geq t) \leq c_1\exp(-c_2t^{c_3})$.

Moreover, based on Assumption 1, we further define quantities μ_1 and μ_2 that control the sampling probability for observations as

$$\mu_1 \geq \frac{1}{d_2}\cdot\max_{k\in\beta}\left\{\frac{1}{p_k}\right\} \quad \text{and} \quad \mu_2 \geq \sqrt{d_2}\cdot\max\left\{\left\|\sum_{k\in\beta}p_k\Theta_k\Theta_k^{\mathbb{T}}\right\|,\ \left\|\sum_{k\in\beta}p_k\Theta_k^{\mathbb{T}}\Theta_k\right\|\right\}. \tag{8}$$

It is easy to obtain that $\mu_1 \geq 1$ and $\mu_2 \geq 1$, according to the facts $\sum_{k\in\beta}p_k = 1$ and $\mathrm{Tr}\big(\sum_{k\in\beta}p_k\Theta_k\Theta_k^{\mathbb{T}}\big) = \mathrm{Tr}\big(\sum_{k\in\beta}p_k\Theta_k^{\mathbb{T}}\Theta_k\big) = 1$, respectively. In general, the values of μ_1 and μ_2 depend on the sampling distribution. The more extreme the sampling distribution is, the larger these two values have to be. Assume that there exist some positive constants γ_1 and γ_2 such that $\gamma_1/d_2 \leq p_k \leq \gamma_2/d_2$, $\forall k \in \beta$. Then we can easily set $\mu_1 := 1/\gamma_1$. The setting of μ_2 is not universal for different cases. For example, consider the cases described in Sect. 2. For correlation matrix completion, we can set $\mu_2 := \gamma_2/\sqrt{2}$ for the real case and $\mu_2 := \gamma_2$ for the complex case. For density matrix completion, we can set $\mu_2 := 1$ for any sampling distribution. For rectangular matrix completion, we can set $\mu_2 := \gamma_2$ for the real case and $\mu_2 := \sqrt{2}\gamma_2$ for the complex case. Note that $\gamma_1 = \gamma_2 = 1$ for uniform sampling.

Theorem 1 reveals the key to deriving a recovery error bound in Frobenius norm, that is, to establish the relation between $\frac{1}{m}\|\mathcal{R}_\Omega(\widehat{X}_m - \overline{X})\|_2^2$ and $\|\widehat{X}_m - \overline{X}\|_F^2$. This can be achieved by looking into some RIP-like property of the sampling operator $\mathcal{R}_\Omega$, as done previously in [40,44,49,58]. Following this idea, we obtain an explicit recovery error bound as follows:

Theorem 2 *Under Assumptions 1 and 2, there exist some positive absolute constants* c_0, c_1, c_2, c_3 *and some positive constants* C_0, C_1 *(only depending on the* ψ_1 *Orlicz norm of* ξ_k*) such that when* $m \geq c_3\sqrt{d_2}\log^3(n_1+n_2)/\mu_2$*, for any* $\kappa > 1$*, if* ρ_m *is chosen as*

$$\rho_m = C_1\kappa\nu\sqrt{\frac{\mu_2\log(n_1+n_2)}{\sqrt{d_2}m}}, \tag{9}$$

then with probability at least $1 - c_1(n_1+n_2)^{-c_2}$,

$$\frac{\|\widehat{X}_m - \overline{X}\|_F^2}{d_2} \leq C_0\left(c_0{}^2(\sqrt{2}+\kappa a_m)^2\nu^2 + \left(\frac{\kappa}{\kappa-1}\right)^2(\sqrt{2}+a_m)^2b^2\right)\mu_1^2\mu_2\frac{\sqrt{d_2}r\log(n_1+n_2)}{m}. \tag{10}$$

Theorem 2 shows that for any rank-correction function F, controlling the recovery error only needs the samples size m to be of roughly the degree of freedom of a rank r matrix up to a logarithmic factor in the matrix size. Besides the information on the order of magnitude, Theorem 2 also provides us more details on the constant part in the recovery error bound, which also plays an important role in practice. The impact of different choices of rank-correction functions on recovery error is fully embodied with the value of a_m. Note that the smaller a_m is, the smaller the error bound (10) is for a fixed κ, and thus the smaller value this error bound can achieve for the best κ (as well as the best ρ_m). Therefore, we aim to establish an explicit relationship between a_m and F in the next theorem.

Theorem 3 *For any given* $\widetilde{X}_m \in \mathbb{V}^{n_1\times n_2}$ *such that* $\|\widetilde{X}_m - \overline{X}\|_F/\sigma_r(\overline{X}) < 1/2$*, we have*

$$a_m \leq -\frac{1}{\sqrt{2r}}\log\left(1 - \sqrt{2}\,\frac{\|\widetilde{X}_m - \overline{X}\|_F}{\sigma_r(\overline{X})}\right) + \varepsilon_F(\widetilde{X}_m),$$

where $\varepsilon_F(\widetilde{X}_m) := \frac{1}{\sqrt{r}}\|F(\widetilde{X}_m) - \widetilde{U}_{m,1}\widetilde{V}_{m,1}^{\mathbb{T}}\|_F$.

It is immediate from Theorem 3 that

$$\frac{\|\widetilde{X}_m - \overline{X}\|_F}{\sigma_r(\overline{X})} < \frac{1}{\sqrt{2}}\left(1 - e^{-\sqrt{2r}(1-\varepsilon_F(\widetilde{X}_m))}\right) \Longrightarrow a_m < 1. \tag{11}$$

Recall that the nuclear norm penalized least squares estimator corresponds to the rank-correction step with $F \equiv 0$ so that $a_m = 1$. Therefore, Theorem 3 guarantees that if the initial estimator $\widetilde{X}_m$ does not deviate too much from $\overline{X}$, the rank-correction step outperforms the nuclear norm penalized least squares estimator in the sense of recovery error, provided that $F(\widetilde{X}_m)$ is close to $\widetilde{U}_{m,1}\widetilde{V}_{m,1}^{\mathbb{T}}$. For example, consider the case when the rank of the true matrix is known. One may simply choose $F(X) = U_1V_1^{\mathbb{T}}$ to take advantage of the rank information. In this case, the requirement in (11) ensuring $a_m < 1$ simply reduces to $\frac{\|\widetilde{X}_m - \overline{X}\|_F}{\sigma_r(\overline{X})} < 0.535 < \frac{1}{\sqrt{2}}(1 - e^{-\sqrt{2r}})$. Moreover, further

suppose that $\widetilde{X}_m$ is the nuclear norm penalized least squares estimator. Then, according to Theorems 2 and 3, one only needs samples with size

$$m = O\left(\sqrt{d_2} r^2 \log^{1+2\tau}(n_1+n_2) \cdot \frac{d_2}{\sigma_r^2(\overline{X})}\right) \Longrightarrow a_m = O(\log^{-\tau}(n_1+n_2)),$$

where $\tau > 0$. As can be seen, the larger the matrix size n is, the easier a_m becomes less than 1 or even close to 0. If the rank of the true matrix is unknown, one could construct the rank-correction function F on account of the tradeoff between optimality and robustness, to be discussed in Sect. 5. An experimental example of the relationship between a_m and F can be found in Table 1.

Next, we demonstrate the power of the rank-correction term with more details. It is interesting to notice that the value of κ (as well as ρ_m) has a substantial impact on the recovery error bound (10). The part related to the magnitude of noise ν increases as κ increases, while the part related to the upper bound b of entries slightly decreases to its limit as κ increases. Therefore, our first target is to find the smallest error bound in terms of (10) among all possible $\kappa > 1$. It is possible to work on the error bound (10) directly for its minimum in κ but the subsequent analysis is much more tedious. For simplicity of illustration, instead, we perform our analysis on a slightly relaxed version instead as

$$\frac{\|\widehat{X}_m - \overline{X}\|_F^2}{d_2} \le C_0\, \eta_m^2\, \mu_1^2\, \mu_2\, \frac{\sqrt{d_2} r \log(n_1+n_2)}{mn},$$

where

$$\eta_m := c_0\big(\sqrt{2} + \kappa a_m\big)\nu + \left(\frac{\kappa}{\kappa - 1}\right)\big(\sqrt{2} + a_m\big)b.$$

Direct calculation shows that over $\kappa > 1$, η_m attains its minimum

$$\overline{\eta}_m = \big(\sqrt{2} + a_m\big)(c_0\nu + b) + 2\sqrt{a_m\big(\sqrt{2} + a_m\big)c_0\nu b} \quad \text{at} \quad \overline{\kappa} = 1 + \sqrt{\left(1 + \frac{\sqrt{2}}{a_m}\right)\frac{b}{c_0\nu}}.$$

It is worthwhile to note that $\overline{\kappa} = O\big(1/\sqrt{a_m}\big)$ when $a_m \ll 1$, meaning that the optimal choice of κ is inversely proportional to $\sqrt{a_m}$ rather than a simple constant. (This observation is important for achieving the rank consistency in Sect. 4.) In other words, for achieving the best possible recovery error, the penalty parameter ρ_m chosen for the rank-correction step (3) with $a_m < 1$ should be larger than that for the nuclear norm penalized least squares estimator. In addition, consider two extreme cases with $a_m = 1$ and $a_m = 0$ respectively:

$$\overline{\eta}_m = \begin{cases} \overline{\eta}^0 := \sqrt{2}(c_0\nu + b) & \text{if } a_m = 0, \\ \overline{\eta}^1 := \big(\sqrt{2}+1\big)(c_0\nu + b) + 2\sqrt{\big(\sqrt{2}+1\big)c_0\nu b} & \text{if } a_m = 1. \end{cases}$$

By direct calculations, we obtain $\overline{\eta}^0/\overline{\eta}^1 \in (0.356, 0.586)$, where the lower bound is attained when $c_0\nu = b$ and the upper bound is approached when $c_0\nu/b \to 0$ or $c_0\nu/b \to \infty$. This finding motivates us to wonder whether the recovery error can be reduced by around half in practice. This inference is further validated by numerical experiments in Sect. 6.

4 Rank consistency

In this section we consider the asymptotic behavior of the estimator generated from the rank-correction step (3) in term of its rank. We expect that the resulting $\widehat{X}_m$ has the same rank as the true matrix $\overline{X}$. Theorem 2 only reveals a flavored parameter ρ_m in terms of the optimal order but rather its exact value. In practice, for a chosen parameter ρ_m, there is hardly any clue to know the recovery performance of the resulting solution since the true matrix is unknown. However, if the rank property holds as expected, the observable rank information may be used to infer the recovery quality of the resulting solution of a parameter and thus help in parameter searching. Numerical experiments in Sect. 6 demonstrate the practicability of this idea.

For the purpose above, we study the rank consistency in the sense of Bach [3] under the setting that the matrix size is fixed. An estimator X_m of the true matrix $\overline{X}$ is said to be rank consistent if

$$\lim_{m\to\infty} \Pr\big(\mathrm{rank}(X_m) = \mathrm{rank}(\overline{X})\big) = 1.$$

Throughout this section, we make the following assumptions:

Assumption 3 The spectral operator F is continuous at $\overline{X}$.

Assumption 4 The initial estimator $\widetilde{X}_m$ satisfies $\widetilde{X}_m \xrightarrow{p} \overline{X}$ as $m \to \infty$.

Epi-convergence in distribution gives us an elegant way in analyzing the asymptotic behavior of optimal solutions of a sequence of constrained optimization problems. Based on this technique, we obtain the following result.

Theorem 4 *If $\rho_m \to 0$, then $\widehat{X}_m \xrightarrow{p} \overline{X}$ as $m \to \infty$.*

We first focus on the characterization of necessary and sufficient conditions for rank consistency of $\widehat{X}_m$. Unlike in the analysis of recovery error bound, additional information represented by the set $\mathcal{C}$ could affect the path along which $\widehat{X}_m$ converges to $\overline{X}$ and thus may break the rank consistency. In the sequel, we only discuss two most common cases: the rectangular case $\mathcal{C} = \mathbb{V}^{n_1\times n_2}$ (recovering a rectangular matrix or a symmetric/Hermitian matrix) and the positive semidefinite case $\mathcal{C} = \mathbb{S}^n_+$ (recovering a symmetric/Hermitian positive semidefinite matrix).

For notational simplicity, we divide the index set β into three subsets as

$$\beta^+ := \{k \in \beta \mid \langle \Theta_k, \overline{X}\rangle = b\}, \quad \beta^- := \{k \in \beta \mid \langle \Theta_k, \overline{X}\rangle = -b\}, \quad \beta^\circ := \beta\backslash(\beta^+ \cup \beta^-). \tag{12}$$

Then, we define a linear operator $\mathcal{Q}^\dagger_\beta : \mathbb{V}^{n_1\times n_2} \to \mathbb{V}^{n_1\times n_2}$ as

$$\mathcal{Q}_\beta^\dagger(X) := \sum_{k\in\beta^\circ} \frac{1}{p_k}\langle \Theta_k, X\rangle \Theta_k + \sum_{k\in\beta^+} \frac{1}{p_k}(\langle \Theta_k, X\rangle)_- \Theta_k + \sum_{k\in\beta^-} \frac{1}{p_k}(\langle \Theta_k, X\rangle)_+ \Theta_k.$$

Here, we use the superscript "†" because of its inverse-like property in terms of

$$\mathcal{Q}_\beta(\mathcal{Q}_\beta^\dagger(Z)) = \mathcal{Q}_\beta^\dagger(\mathcal{Q}_\beta(Z)) = \mathcal{P}_\beta(Z) \quad \forall\, Z \in \{Z \in \mathbb{V}^{n_1\times n_2} \mid \mathcal{R}_{\beta^+}(Z) \le 0, \mathcal{R}_{\beta^-}(Z) \ge 0\}.$$

By extending the arguments of Bach [3] for the nuclear norm penalized least squares estimator from the unconstrained case to the constrained case, we obtain the following results.

Theorem 5 *For the rectangular case $\mathcal{C} = \mathbb{V}^{n_1\times n_2}$, consider the linear system*

$$\overline{U}_2^{\mathbb{T}} \mathcal{Q}_\beta^\dagger(\overline{U}_2 \Gamma \overline{V}_2^{\mathbb{T}})\overline{V}_2 = \overline{U}_2^{\mathbb{T}} \mathcal{Q}_\beta^\dagger\big(\overline{U}_1 \overline{V}_1^{\mathbb{T}} - F(\overline{X})\big)\overline{V}_2. \tag{13}$$

If $\rho_m \to 0$ and $\sqrt{m}\rho_m \to \infty$, then for the rank consistency of $\widehat{X}_m$,

(i) a necessary condition: (13) has a solution $\widehat{\Gamma} \in \mathbb{V}^{(n_1-r)\times(n_2-r)}$ with $\|\widehat{\Gamma}\| \le 1$;

(ii) a sufficient condition: (13) has a unique solution $\widehat{\Gamma} \in \mathbb{V}^{(n_1-r)\times(n_2-r)}$ with $\|\widehat{\Gamma}\| < 1$.

For the positive semidefinite case, the nuclear norm $\|X\|_*$ in (3) simply reduces to the trace $\langle I_n, X\rangle$. We assume that the Slater condition holds.

Assumption 5 For the positive semidefinite case $\mathcal{C} = \mathbb{S}_+^n$, the Slater condition holds, i.e., there exists some $X^0 \in \mathbb{S}_{++}^n$ such that $\mathcal{R}_\alpha(X^0) = \mathcal{R}_\alpha(\overline{X})$ and $\|\mathcal{R}_\beta(X^0)\|_\infty < b$.

Theorem 6 *For the positive semidefinite case $\mathcal{C} = \mathbb{S}_+^n$, consider the linear system*

$$\overline{P}_2^{\mathbb{T}} \mathcal{Q}_\beta^\dagger(\overline{P}_2 \Lambda \overline{P}_2^{\mathbb{T}})\overline{P}_2 = \overline{P}_2^{\mathbb{T}} \mathcal{Q}_\beta^\dagger\big(I_n - F(\overline{X})\big)\overline{P}_2. \tag{14}$$

Under Assumption 5, if $\rho_m \to 0$ and $\sqrt{m}\rho_m \to \infty$, then for the rank consistency of $\widehat{X}_m$,

(i) a necessary condition: (14) has a solution $\widehat{\Lambda} \in \mathbb{S}_+^{n-r}$;

(ii) a sufficient condition: (14) has a unique solution $\widehat{\Lambda} \in \mathbb{S}_{++}^{n-r}$.

Next, we provide a theoretical guarantee on the uniqueness of the solution to the linear systems (13) and (14) with the help of constraint nondegeneracy. The concept of constraint nondegeneracy was pioneered by Robinson [65] and later extensively developed by Bonnans and Shapiro [5]. We say that the constraint nondegeneracy holds at $\overline{X}$ to (3) with $\mathcal{C} = \mathbb{V}^{n_1\times n_2}$ if

$$\mathcal{R}_{\alpha\cup\beta^+\cup\beta^-}\big(\mathcal{T}(\overline{X})\big) = \mathbb{R}^{|\alpha\cup\beta^+\cup\beta^-|}, \tag{15}$$

where $\mathcal{T}(\overline{X}) = \left\{H \in \mathbb{V}^{n_1\times n_2} \mid \overline{U}_2^{\mathbb{T}} H \overline{V}_2 = 0\right\}$. Meanwhile, we say that the constraint nondegeneracy holds at $\overline{X}$ to (3) with $\mathcal{C} = \mathbb{S}^n_+$ if

$$\mathcal{R}_{\alpha\cup\beta^+\cup\beta^-}\left(\mathrm{lin}(\mathcal{T}_{\mathbb{S}^n_+}(\overline{X}))\right) = \mathbb{R}^{|\alpha\cup\beta^+\cup\beta^-|}, \tag{16}$$

where $\mathrm{lin}(\mathcal{T}_{\mathbb{S}^n_+}(\overline{X})) = \left\{H \in \mathbb{S}^n \mid \overline{P}_2^{\mathbb{T}} H \overline{P}_2 = 0\right\}$. One may refer to "Appendix 2" for more details of constraint nondegeneracy.

To take a closer look at the linear systems (13) and (14), we define linear operators $\mathcal{B}_1 : \mathbb{V}^{n_1\times n_2} \to \mathbb{V}^{(n_1-r)\times(n_2-r)}$ and $\mathcal{B}_2 : \mathbb{V}^{(n_1-r)\times(n_2-r)} \to \mathbb{V}^{(n_1-r)\times(n_2-r)}$ associated with $\overline{X}$, respectively, by

$$\mathcal{B}_1(Y) := \overline{U}_2^{\mathbb{T}} \mathcal{Q}^{\dagger}_{\beta}(Y)\overline{V}_2 \ \text{ and } \ \mathcal{B}_2(Z) := \overline{U}_2^{\mathbb{T}} \mathcal{Q}^{\dagger}_{\beta}(\overline{U}_2 Z \overline{V}_2^T)\overline{V}_2, \tag{17}$$

where $Y \in \mathbb{V}^{n_1\times n_2}$ and $Z \in \mathbb{V}^{(n_1-r)\times(n_2-r)}$. From the definition of $\mathcal{Q}^{\dagger}_{\beta}$, we know that the operator $\mathcal{B}_2$ is self-adjoint and positive semidefinite. Then, for the rectangular case $\mathcal{C} = \mathbb{V}^{n_1\times n_2}$, the linear system (13) can be rewritten as

$$\mathcal{B}_2(\Gamma) = \mathcal{B}_1(\overline{U}_1\overline{V}_1^{\mathbb{T}} - F(\overline{X})), \quad \Gamma \in \mathbb{V}^{(n_1-r)\times(n_1-r)}, \tag{18}$$

and for the positive semidefinite case $\mathcal{C} = \mathbb{S}^n_+$, the linear system (14) can be rewritten as

$$\mathcal{B}_2(\Lambda) = \mathcal{B}_2(I_{n-r}) + \mathcal{B}_1(\overline{P}_1\overline{P}_1^{\mathbb{T}} - F(\overline{X})), \quad \Lambda \in \mathbb{S}^{n-r}, \tag{19}$$

since both $\overline{U}_i$ and $\overline{V}_i$ reduce to $\overline{P}_i$ for $i = 1, 2$ for $\overline{X} \in \mathbb{S}^n_+$.

Clearly, the invertibility of $\mathcal{B}_2$ is equivalent to the uniqueness of the solution to the linear systems (13) and (14). The following result provides a link between the constraint nondegeneracy and the positive definiteness of $\mathcal{B}_2$.

Theorem 7 *For either the rectangular case $\mathcal{C} = \mathbb{V}^{n_1\times n_2}$ or the positive semidefinite case $\mathcal{C} = \mathbb{S}^n_+$, if the constraint nondegeneracy holds at $\overline{X}$ to the problem (3), then the self-adjoint linear operator $\mathcal{B}_2$ defined by (17) is positive definite.*

Combining Theorems 5, 6 and 7 together with (18) and (19), we immediately have the following result of rank consistency.

Theorem 8 *Suppose that $\rho_m \to 0$ and $\sqrt{m}\rho_m \to \infty$. If*

(i) for the rectangular case $\mathcal{C} = \mathbb{V}^{n_1\times n_2}$, the constraint nondegeneracy (15) holds at $\overline{X}$ to the problem (3) and

$$\left\|\mathcal{B}_2^{-1}\mathcal{B}_1(\overline{U}_1\overline{V}_1^{\mathbb{T}} - F(\overline{X}))\right\| < 1; \tag{20}$$

(ii) for the positive semidefinite case $\mathcal{C} = \mathbb{S}^n_+$, the constraint nondegeneracy (16) holds at $\overline{X}$ to the problem (3) and

$$I_{n-r} + \mathcal{B}_2^{-1}\mathcal{B}_1(\overline{P}_1\overline{P}_1^{\mathbb{T}} - F(\overline{X})) \in \mathbb{S}^{n-r}_{++}, \tag{21}$$

then the estimator $\widehat{X}_m$ generated from the rank-correction step (3) is rank consistent.

From Theorem 8, it is not difficult to see that when $F(\overline{X})$ is sufficiently close to $\overline{U}_1\overline{V}_1^{\mathbb{T}}$, the conditions (20) and (21) hold automatically and so does the rank consistency. Thus, Theorem 8 provides us a guideline to construct a suitable rank-correction function F to achieve the rank consistency. In particular, for the positive semidefinite matrix completion, we further consider two important classes as follows.

Class I The covariance matrix completion with partial positive diagonal entries fixed. Due to the positive semidefinite structure, the magnitudes of off-diagonal entries are fully controlled by the magnitudes of diagonal entries. Therefore, we remove all the bounded constraints corresponding to off-diagonal entries from the rank-correction step (3) as they are redundant. Thus, the constraints are reduced to

$$X_{ii} = \overline{X}_{ii} \quad \forall\, i \in \pi, \quad X_{ii} \le b \ \forall\, i \in \pi^c, \quad X \in \mathbb{S}^n_+,$$

where (π, π^c) is a partition of the index set $\{1, \ldots, n\}$. This class of problems includes the correlation matrix completion as a special case, in which all diagonal entries are fixed to be ones.

Class II The density matrix completion with its trace fixed to be one.
Due to the positive semidefinite structure, all the coefficients of Pauli basis are controlled because of the trace one constraint. Therefore, we remove all the bounded constraints from the rank-correction step (3) as they are redundant. Thus, in this case the constraints are reduced to

$$\frac{1}{\sqrt{n}}\mathrm{Tr}(X) = \frac{1}{\sqrt{n}}, \quad X \in \mathbb{S}^n_+.$$

Interestingly, for the matrix completion problems of Classes I and II, the constraint nondegeneracy automatically holds at $\overline{X}$. More importantly, if observations are sampled uniformly at random, the rank consistency can be guaranteed for a broad class of rank-correction functions F.

Theorem 9 *For the matrix completion problems of Classes I and II under uniform sampling, if $\rho_m \to 0$, $\sqrt{m}\rho_m \to \infty$ and F is a spectral operator associated with a symmetric function $f : \mathbb{R}^n \to \mathbb{R}^n$ such that for $i = 1, \ldots, n$,*

$$\begin{cases} f_i(x) > 0 & \text{if } x_i > 0, \\ f_i(x) = 0 & \text{if } x_i = 0, \end{cases} \qquad \forall\, x \in \mathbb{R}^n_+ \ \text{ and } \ \forall\, i = 1, \ldots, n, \tag{22}$$

then the estimator $\widehat{X}_m$ generated from the rank-correction step (3) is rank consistent.

5 Construction of the rank-correction function

In this section, we focus on the construction of a suitable rank-correction function F based on the results in Sects. 3 and 4. For achieving a smaller recovery error, according to Theorem 2, we desire a construction such that $F(\widetilde{X}_m)$ is close to $\overline{U}_1\overline{V}_1^{\mathbb{T}}$. Meanwhile,

for achieving the rank consistency, according to Theorem 8, we desire a construction such that $F(\overline{X})$ is close to $\overline{U}_1\overline{V}_1^{\mathbb{T}}$. Therefore, these two guidelines consistently suggest a natural idea, i.e., if possible, choosing

$$F(X) \approx U_1 V_1^{\mathbb{T}} \quad \text{near } \overline{X}.$$

Next, we proceed with the construction of the rank-correction function F for the rectangular case. For the positive semidefinite case, one only needs to replace the singular value decomposition with the eigenvalue decomposition and conduct exactly the same analysis.

5.1 The rank is known

If the rank of the true matrix $\overline{X}$ is known, it is clear that the best choice of F is

$$F(X) := U_1 V_1^{\mathbb{T}}, \tag{23}$$

where $(U, V) \in \mathbb{O}^{n_1,n_2}(X)$ and $X \in \mathbb{V}^{n_1\times n_2}$. Note that F defined by (23) is not a spectral operator over the whole space of $\mathbb{V}^{n_1\times n_2}$, but in a neighborhood of $\overline{X}$ it is indeed a spectral operator and is actually twice continuously differentiable (see, e.g., [11, Proposition 8]). With this rank-correction function, the rank-correction step is essentially the same as a single step of the majorized penalty method developed in [27].

5.2 The rank is unknown

If the rank of the true matrix $\overline{X}$ is unknown, we intend to construct a spectral operator F to imitate the case when the rank is known. Here, we propose F to be a spectral operator

$$F(X) := U\mathrm{Diag}\big(f(\sigma(X))\big)V^{\mathbb{T}} \tag{24}$$

associated with the symmetric function $f : \mathbb{R}^n \to \mathbb{R}^n$ defined by

$$f_i(x) = \begin{cases} \phi\left(\dfrac{x_i}{\|x\|_\infty}\right) & \text{if } x \in \mathbb{R}^n\backslash\{0\}, \\ 0 & \text{if } x = 0, \end{cases} \tag{25}$$

where $(U, V) \in \mathbb{O}^{n_1,n_2}(X)$, $X \in \mathbb{V}^{n_1\times n_2}$, and the scalar function $\phi : \mathbb{R} \to \mathbb{R}$ takes the form

$$\phi(t) := \mathrm{sgn}(t)(1+\varepsilon^\tau)\frac{|t|^\tau}{|t|^\tau + \varepsilon^\tau}, \quad t \in \mathbb{R}, \tag{26}$$

for some $\tau > 0$ and $\varepsilon > 0$.

Corollary 10 *Let F be a spectral operator defined by (24), (25) and (26).*

(i) If $\frac{\|\widetilde{X}_m-\overline{X}\|_F}{\sigma_r(\overline{X})} < \frac{1}{\sqrt{2}}\left(1-e^{-\sqrt{2r}}\right)$, *then for any* ε *satisfying* $\frac{\sigma_{r+1}(\widetilde{X}_m)}{\sigma_1(\widetilde{X}_m)} < \varepsilon < \frac{\sigma_r(\widetilde{X}_m)}{\sigma_1(\widetilde{X}_m)}$, *there exists some* $\overline{\tau}_1 > 0$ *such that* $a_m < 1$ *for any* F *with* $\tau \geq \overline{\tau}_1$.

(ii) Suppose that the constraint nondegeneracy holds at $\overline{X}$ *to the problem (3). If* $\rho_m \to 0$ *and* $\sqrt{m}\rho_m \to \infty$, *then for any* ε *satisfying* $0 < \varepsilon < \frac{\sigma_r(\overline{X})}{\sigma_1(\overline{X})}$, *there exists some* $\overline{\tau}_2 > 0$ *such that the rank consistency of* $\widehat{X}_m$ *holds for any* F *with* $\tau \geq \overline{\tau}_2$.

The proof of Corollary 10 is straightforward so we omit it. Corollary 10 suggests an ideal choice of ε for the recovery error reduction, i.e., $\varepsilon \in \left(\frac{\sigma_{r+1}(\widetilde{X}_m)}{\sigma_1(\widetilde{X}_m)}, \frac{\sigma_r(\widetilde{X}_m)}{\sigma_1(\widetilde{X}_m)}\right)$, provided that $\widetilde{X}_m$ does not deviate too much from $\overline{X}_m$, and also an ideal choice of ε for rank consistency, i.e., $\varepsilon \in \left(0, \frac{\sigma_r(\overline{X}_m)}{\sigma_1(\overline{X}_m)}\right)$. Note that these two intervals may not overlap each other, implying the theoretical possibility that the recovery error reduction and the rank consistency may not be achieved simultaneously if the initial estimator $\widetilde{X}_m$ is not close to $\overline{X}_m$.

The interval of ε for the recovery error reduction is disclosed if the true rank is accessible. Therefore, this ideal interval is an important insight that can be used to guide the choice of ε in practice since the initial $\widetilde{X}_m$ should contain some information of the true rank in general. Indeed, the value of ε can be regarded as a divide of confidence on whether $\sigma_i(\widetilde{X}_m)$ is believed to come from a nonzero singular values of $\overline{X}$ with perturbation—positive confidence if $\sigma_i(\widetilde{X}_m) > \varepsilon\sigma_1(\widetilde{X}_m)$ and negative confidence if $\sigma_i(\widetilde{X}_m) < \varepsilon\sigma_1(\widetilde{X}_m)$. Next we look for a suitable τ. It is observed from Fig. 1 that the parameter $\tau > 0$ mainly controls the shape of ϕ over $t \in [0, 1]$. The function ϕ is concave if $0 < \tau \leq 1$ and S-shaped with a single inflection point at $\varepsilon\left(\frac{\tau-1}{\tau+1}\right)^{1/\tau}$ if $\tau > 1$. It should be good to choose an S-shaped function ϕ. But one also needs to take account of the steepness of ϕ, which increases when τ increases. In particular for any ε satisfying $0 < \varepsilon < 1$, ϕ approaches to the step function taking the value 0 if $0 \leq t < \varepsilon$ and the value 1 if $\varepsilon < t \leq 1$ as $\tau \to \infty$. Since the rank of $\overline{X}$ is unknown and the singular values of $\widetilde{X}_m$ are unpredictable, choosing a large τ could be risky. Therefore, one needs to choose τ with certain conservation, sacrificing certain recovery quality in exchange for robustness strategically. Here, we provide a recommendation of the choices $\varepsilon \approx 0.05$ (or within $0.01 \sim 0.1$) and $\tau = 2$ (or within $1 \sim 3$) for most cases, particularly when the initial estimator is generated from the nuclear norm penalized least squares problem. These choices have performed very stably for plenty of problems, as validated in Sect. 6.

We also remark that for the positive semidefinite case, the rank-correction function defined by (24), (25) and (26) is related to the reweighted trace norm for the matrix rank minimization proposed by Fazel et al. [20,56]. The reweighted trace norm in [20,56] for the positive semidefinite case is $\langle (X^k + \varepsilon I_n)^{-1}, X\rangle$, which arises from the derivative of the surrogate function $\log\det(X + \varepsilon I_n)$ of the rank at an iterate X^k, where ε is a small positive constant. Meanwhile, in our proposed rank-correction step, if we choose $\tau = 1$, then $I_n - \frac{1}{1+\varepsilon}F(\widetilde{X}_m) = \varepsilon'(\widetilde{X}_m + \varepsilon' I_n)^{-1}$ with $\varepsilon' = \varepsilon\|\widetilde{X}_m\|$. Superficially, similarity occurs; however, it is notable that ε' depends on $\widetilde{X}_m$, which is different from the constant ε in [20,56]. More broadly speaking, the rank-correction function F defined by (24), (25) and (26) is not a gradient of any real-valued function.

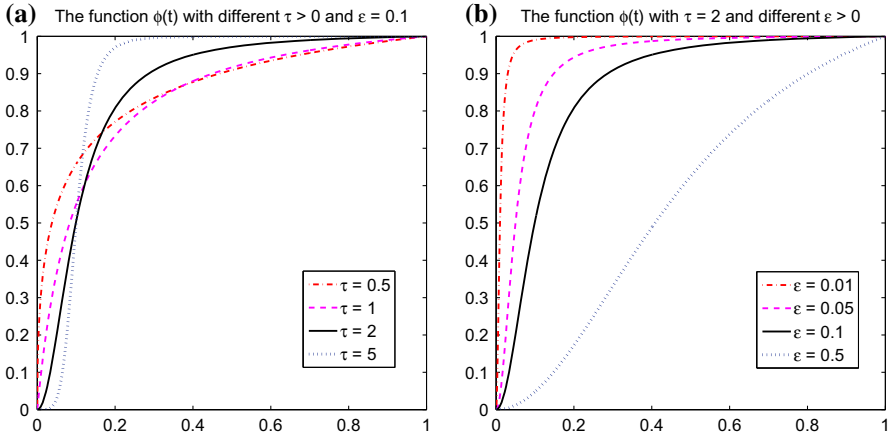

Fig. 1 Shapes of the function ϕ with different $\varepsilon > 0$ and $\tau > 0$. **a** $\varepsilon = 0.1$ with different $\tau > 0$. **b** $\tau = 2$ with different $\varepsilon > 0$

This distinguishes our proposed rank-correction step from the reweighted trace norm minimization in [20,56] even for the positive semidefinite case.

6 Numerical experiments

In this section, we validate the power of our proposed rank-correction step on the recovery by applying it to different matrix completion problems. We adopted the proximal alternating direction method of multipliers (proximal ADMM) to solve the optimization problem (3). For more details of the proximal ADMM, the readers may refer to Appendix B of [21]. For convenience, in the sequel, the NNPLS estimator and the RCS estimator, respectively, stand for the estimators from the nuclear norm penalized least squares problem (i.e., $F \equiv 0$) and the rank-correction step (3) with F specified in Sect. 5. Given an estimator X_m of $\overline{X}_m$, the **relative error** (**relerr** for short) is defined by

$$\text{relerr} = \frac{\|X_m - \overline{X}\|_F}{\max(10^{-8}, \|\overline{X}\|_F)}.$$

6.1 Influence of fixed basis coefficients on the recovery

In this subsection, we test the performance of the NNPLS estimator and the RCS estimator for different patterns of fixed basis coefficients. We randomly generated a correlation matrix by the following command:

```
M = randn(n,r)/sqrt(sqrt(n));  ML = weight*M(:,1:k);
 M(:,1:k) = ML;
Xtemp = M*M'; D = diag(1./sqrt(diag(Xtemp)));
 X_bar =  D*Xtemp*D.
```

We took the true matrix $\overline{X}$ = `X_bar` with dimension `n` = 500, rank `r` = 5, `weight` = 5 and `k` = 1. Here, the parameter `weight` is used to control the relative magnitude difference between the first k largest eigenvalues and the left $r - k$ nonzero eigenvalues. We randomly fixed partial diagonal and off-diagonal entries of $\overline{X}$ and then uniformly sampled the rest entries with i.i.d. Gaussian noise. The noise level, defined by $\|\nu\xi\|_2/\|y\|_2$ in (2) hereafter, was set to be 10 % and the upper bound of the non-fixed diagonal entries was set to be 1. We further assumed that the rank of the true matrix was known so that for RCS estimator we chose the rank-correction function (23).

In Fig. 2, we plot the curves of the relative recovery error and the rank of both the NNPLS estimator (the subfigures on the left) and the RCS estimator (the subfigures on the rigth) for different patterns of fixed entries. Note that both m and ρ_m in the rank-correction step (3) depend on the problem of consideration. Thus, we report $m\rho_m$ as a whole in the x-axis. (Note that for a specific problem, only ρ_m is adjustable.) In the captions of subfigures, **diag** means the number of fixed diagonal entries, and **off-diag** means the number of fixed off-diagonal entries. For each subfigure on the right side, the initial $\widetilde{X}_m$ for the RCS estimator is the point with the smallest recovery error from the corresponding subfigure on the left side.

Figure 2 fully manifests the advantage of the RCS estimator over the NNPLS estimator. It is shown that compared with the NNPLS estimator, the RCS estimator substantially reduces the recovery error and significantly improves the rank consistency. Moreover, the RCS estimator possesses a wide rage of the parameter ρ_m to achieve a desired small recovery error and the rank of the true matrix simultaneously. It indicates that whether the resulting solution of a parameter ρ_m achieves the true rank can be used to infer the recovery quality. Even if the true rank is unknown in advance, it is still possible to pick out a satisfied solution via monitoring the change of rank in parameter searching. Such advantages are far beyond the reach of the NNPLS estimator.

6.2 Performance of different rank-correction functions for recovery

In this subsection, we test the performance of different rank-correction functions for recovering a correlation matrix. We randomly generated the true matrix $\overline{X}$ by the command in Sect. 6.1 with `n` = 1000, `r` = 10, `weight` = 2 and `k` = 5. We fixed all the diagonal entries of $\overline{X}$ and then sampled partial off-diagonal entries uniformly at random with i.i.d. Gaussian noise. The noise level was set to be 10 %. We chose the (nuclear norm penalized) least squares estimator to be the initial estimator $\widetilde{X}_m$. In Fig. 3, we plot four curves corresponding to the rank-correction functions F defined by (24), (25) and (26) with different ε and τ, and additional two curves corresponding to the rank-correction functions F defined by (23) at $\widetilde{X}_m$ (i.e., $\widetilde{U}_1\widetilde{V}_1^{\mathbb{T}}$) and $\overline{X}$ (i.e., $\overline{U}_1\overline{V}_1^{\mathbb{T}}$), respectively. The values of a_m and the best recovery error are listed in Table 1.

For all the rank-correction functions plotted in Fig. 3, when ρ_m increases, the recovery error first decreases together with the rank and then increases after the rank of the true matrix is attained. The only exception is $\overline{U}_1\overline{V}_1^{\mathbb{T}}$. This exactly validates our discussion about the recovery error at the end of Sect. 3. It is worthwhile to point out

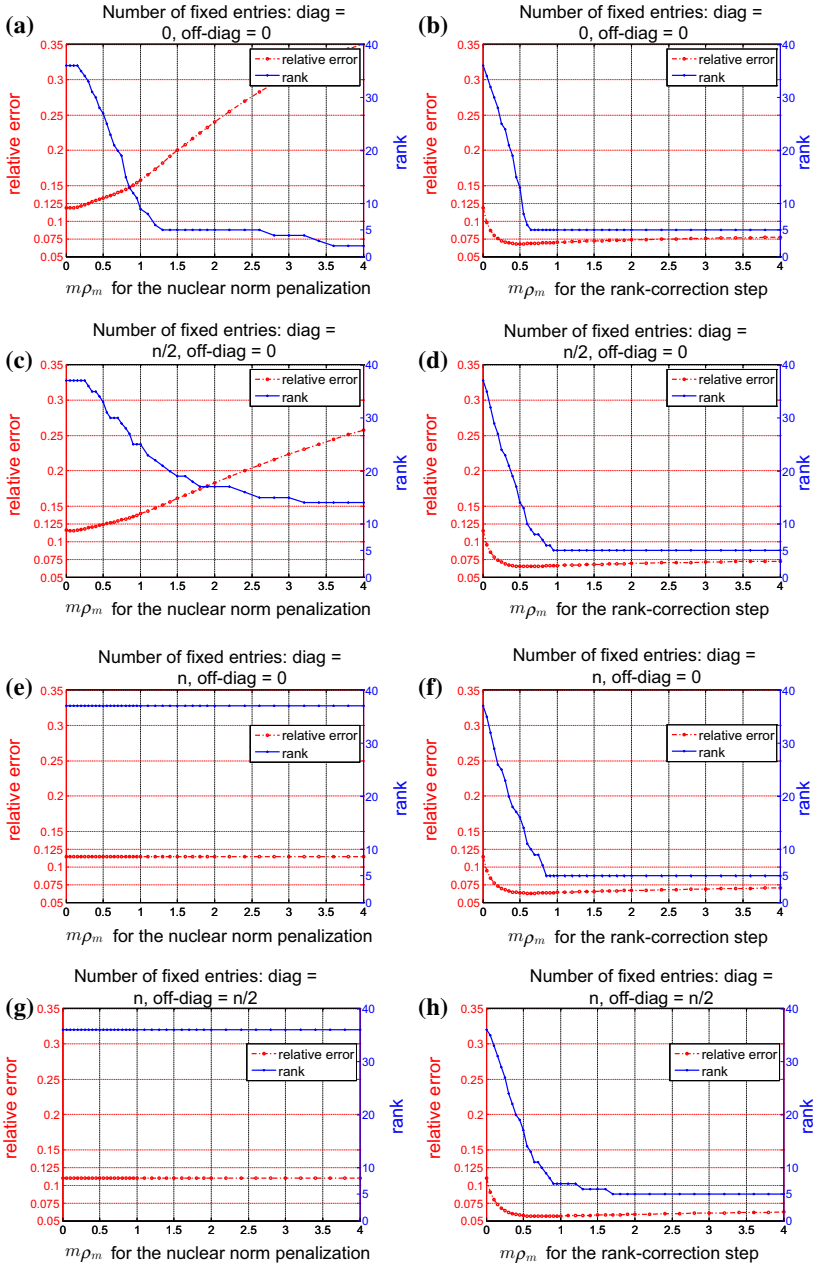

Fig. 2 Influence of fixed basis coefficients on recovery (sample ratio = 6.4 %). **a** Nuclear norm: diag = 0, off-diag = 0. **b** Rank-correction step: diag = 0, off-diag = 0. **c** Nuclear norm: diag = n/2, off-diag = 0. **d** Rank-correction step: diag = n/2, off-diag = 0. **e** Nuclear norm: diag = n, off-diag = 0. **f** Rank-correction step: diag = n, off-diag = 0. **g** Nuclear norm: diag = n, off-diag = n/2. **f** Rank-correction step: diag = n, off-diag = n/2

Table 1 Influence of the rank-correction term on the recovery error

F	Zero function	$\varepsilon = 0.1$ $\tau = 1$	$\varepsilon = 0.1$ $\tau = 2$	$\varepsilon = 0.1$ $\tau = 3$	$\varepsilon = 0.05$ $\tau = 2$	$\widetilde{U}_1\widetilde{V}_1^{\mathbb{T}}$	$\overline{U}_1\overline{V}_1^{\mathbb{T}}$
a_m	1	0.3126	0.1652	0.1402	0.1849	0.1355	0
Optimal relerr (%)	10.66	5.92	5.84	5.83	5.83	5.84	3.00

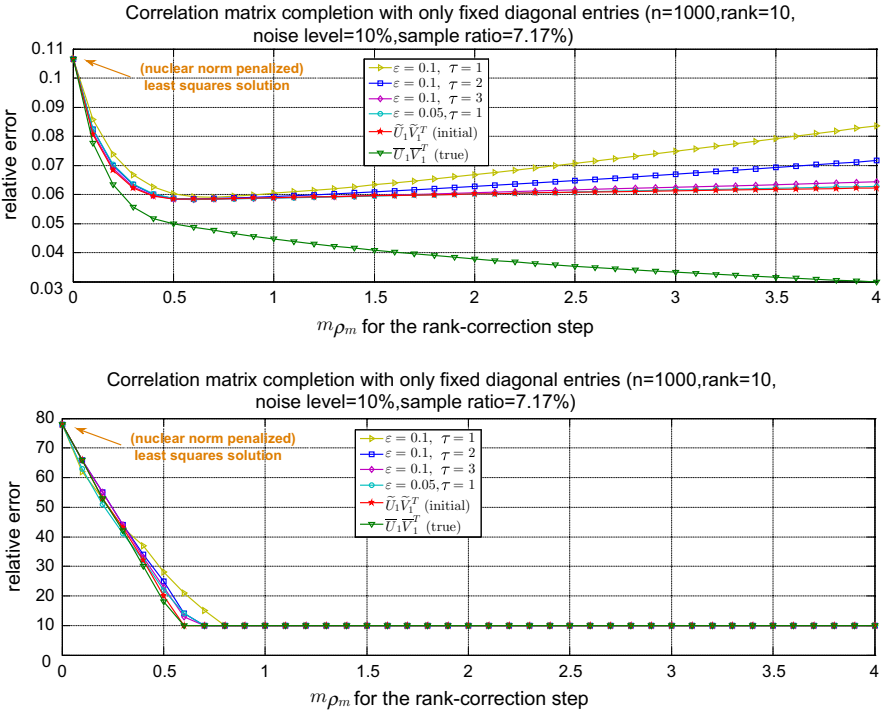

Fig. 3 Influence of the rank-correction term on the recovery

that, according to our observations of many tests, in practice, if a_m is larger than 1 but not too much, the recovery performance of the RCS estimator still has a high chance to be much better than that of the NNPLS estimator.

6.3 Performance of different initial NNPLS estimators for recovery

In this subsection, we take the covariance matrix completion for example to test the performance of the RCS estimator with different initial NNPLS estimators $\widetilde{X}_m$. We generated the true matrix $\overline{X}$ by the command in Sect. 6.1 with `n` $= 500$, `r` $= 5$, `weight` $= 3$ and `k` $= 1$ except that `D = eye(n)`. The upper bound of the non-fixed diagonal entries was set to be double of the largest absolute value among all the noisy

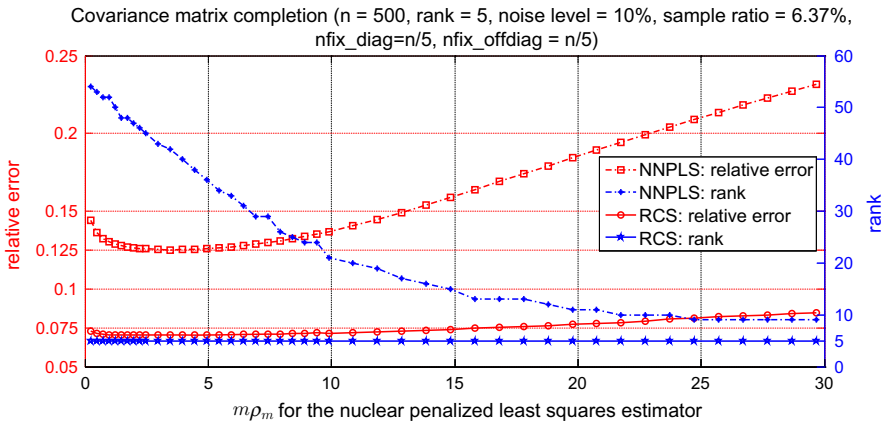

Fig. 4 Performance of the RCS estimator with different initial NNPLS estimators

observations of entries together with the fixed entries. We assumed that the rank of the true matrix was known so that we chose the rank-correction function (23).

For each ρ_m, we first produced the NNPLS estimator, and then use it as the initial point to produce a sequence of RCS estimators with different penalty parameters. Next we choose the RCS estimators that attains the correct rank with the smallest penalty parameter. As can be seen from Fig. 2, this choice of the RCS estimator results in the desired small recovery error. The test results are plotted in Fig. 4, where the dash curves represent for the NNPLS estimator and the solid curves represent for the chosen RCS estimator. We clearly observe from Fig. 4 that, no matter which NNPLS estimator is given to be the initial estimator, the RCS estimator can always substantially improve the recovery quality in terms of both the error and the rank.

6.4 Performance for different matrix completion problems

In this subsection, we test the performance of the RCS estimator for different matrix completion problems. Figure 2 has revealed that a good choice of the parameter ρ_m for the RCS estimator could be the smallest value that attains a stable rank. Therefore, the bisection search method can be used to find such a parameter ρ_m. This is actually what we benefit from rank consistency. In the following experiments, we apply this strategy to find a suitable ρ_m for the RCS estimator.

A natural question then arises: Will multiple rank-correction steps further improve the recovery quality? The answer can be found in Tables 2, 3 and 4 below, which report the experimental results for covariance matrix completion, rectangular matrix completion and density matrix completion, respectively. The reported NNPLS estimator is the one with the smallest recovery error among all different ρ_m presuming the true matrix is known. The initial estimator of the first RCS estimator is the NNPLS estimator with a single preset $\rho_m = 0.4\frac{\eta\|y\|_2}{\sqrt{m}}\sqrt{\frac{\log(n_1+n_2)}{mn}}$, where η is the noise level. This choice of ρ_m follows (9) with $C = 0.4$, $\kappa = 1$, $\mu = 1$ and ν taken its expected value based on observations. The second (third) RCS estimator takes the first (second)

Table 2 Performance for covariance matrix completion problems with $n = 1000$

r	Diag/off-diag	Sample ratio (%)	Relerr (rank)			
			NNPLS	1st RCS	2st RCS	3rd RCS
5	1000/0	2.40	1.94e−1 (47)	8.84e−2 (5)	8.03e−2 (5)	7.85e−2 (5)
	1000/0	7.99	6.08e−2 (50)	3.39e−2 (5)	3.38e−2 (5)	3.38e−2 (5)
	500/500	2.39	2.28e−1 (56)	1.07e−1 (5)	8.99e−2 (5)	8.48e−2 (5)
	500/500	7.98	1.16e−1 (56)	5.62e−2 (5)	5.42e−2 (5)	5.40e−2 (5)
10	1000/0	5.38	1.59e−1 (77)	7.42e−2 (10)	7.23e−2 (10)	7.22e−2 (10)
	1000/0	8.96	9.15e−2 (81)	5.06e−2 (10)	5.05e−2 (10)	5.05e−2 (10)
	500/500	5.38	1.65e−1 (82)	7.70e−2 (10)	7.29e−2 (10)	7.28e−2 (10)
	500/500	8.96	9.54e−2 (85)	5.16e−2 (10)	5.11e−2 (10)	5.11e−2 (10)

RCS estimator to be the initial estimator. The rank-correction function F is defined by (24), (25) and (26) with $\varepsilon = 0.05$ and $\tau = 2$.

For the covariance matrix completion problems, we generated the true matrix $\overline{X}$ by the command in Sect. 6.1 with `n` $= 1000$, `weight` $= 2$ and `k` $= 1$ except that `D = eye(n)`. The rank of $\overline{X}$ and the number of fixed diagonal and non-diagonal entries of $\overline{X}$ are reported in the first and the second columns of Table 2, respectively. We sampled partial off-diagonal entries uniformly at random with i.i.d. Gaussian noise at the noise level 10 %. The upper bound of the non-fixed diagonal entries was set to be double of the largest absolution value among all the noisy observations of entries together with the fixed entries. From Table 2, we see that when the sample ratio is reasonable, a single rank-correction step is fully capable to yield a desired result. However, when the sample ratio is very low, especially if some off-diagonal entries are fixed, one or two further rank-correction steps could still bring some improvement in recovery quality.

For the density matrix completion problems, we generated the true density matrix $\overline{X}$ by the following command:

```
M = randn(n,r)+i*randn(n,r);  ML = weight*M(:,1:k);
 M(:,1:k) = ML;
Xtemp = M*M';  X_bar = Xtemp/sum(diag((Xtemp))).
```

During the testing, we set `n` $= 1024$, `weight` $= 2$ and `k` $= 1$, and sampled partial Pauli measurements except the trace of $\overline{X}$ uniformly at random with 10 % i.i.d. Gaussian noise. Besides this statistical noise, we further added the depolarizing noise, which frequently appears in quantum systems. The strength of the depolarizing noise was set to be 0.01. This case is labeled as the mixed noise in the last four rows of Table 3. We remark here that the depolarizing noise differs from our assumption on noise since it does not have randomness. One may refer to [22,31] for details of the quantum depolarizing channel. In [22], Flammia et al. proposed a two-step method for seeking a feasible solution of low-rank—(1) evaluating an NNPLS estimator by dropping the trace one constraint; (2) normalizing the resulting solution to be of trace one. We tested this method in our experiments, with the NNPLS estimator without trace one constraint chosen to be the one with the smallest recovery error among all

Table 3 Performance for density matrix completion problems with $n = 1024$

Noise	r	Noise level (%)	Sample ratio (%)	NNPLS1			NNPLS2			RCS		
				Fidelity	Relerr	Rank	Fidelity	Relerr	Rank	Fidelity	Relerr	Rank
Statistical	3	10.0	1.5	0.716	2.49e−1	3	0.962	2.34e−1	3	0.992	8.47e−2	3
		10.0	4.0	0.915	8.14e−2	3	0.997	6.88e−2	3	0.998	4.13e−2	3
	5	10.0	2.5	0.696	2.56e−1	5	0.959	2.71e−1	5	0.992	8.28e−2	5
		10.0	5.0	0.886	1.04e−1	5	0.994	9.61e−2	5	0.997	4.81e−2	5
Mixed	3	12.5	1.5	0.657	2.95e−1	3	0.959	2.41e−1	3	0.990	9.89e−2	3
		12.4	4.0	0.842	1.42e−1	3	0.996	7.48e−2	3	0.997	6.20e−2	3
	5	12.4	2.5	0.631	3.05e−1	5	0.954	2.87e−1	5	0.990	9.81e−2	5
		12.5	5.0	0.814	1.62e−1	5	0.994	1.03e−1	5	0.996	6.94e−2	5

Table 4 Performance for rectangular matrix completion problems

Setting	Sample	Fixed	Sample ratio (%)	Relerr (rank)			
				NNPLS	1st RCS	2st RCS	3rd RCS
dim = 1000 × 1000, rank = 10	Uniform	0	5.97	1.98e−1 (119)	7.69e−2 (10)	7.31e−2 (10)	7.30e−2 (10)
		0	11.9	8.34e−2 (114)	4.49e−2 (10)	4.48e−2 (10)	4.48e−2 (10)
		1000	5.98	1.93e−1 (120)	7.45e−2 (10)	7.01e−2 (10)	7.00e−2 (10)
		1000	12.0	8.20e−2 (108)	4.35e−2 (10)	4.34e−2 (10)	4.34e−2 (10)
	Non-uniform	0	5.97	3.20e−1 (144)	1.22e−1 (10)	9.31e−2 (10)	8.77e−2 (10)
		0	11.9	1.27e−1 (171)	5.32e−2 (10)	5.12e−2 (10)	5.11e−2 (10)
		1000	5.98	3.07e−1 (146)	1.16e−1 (10)	8.78e−2 (10)	8.30e−2 (10)
		1000	12.0	1.24e−1 (173)	5.14e−2 (10)	4.93e−2 (10)	4.92e−2 (10)
dim = 500 × 1500, rank = 5	Uniform	0	3.99	2.31e−1 (73)	9.15e−2 (5)	8.10e−2 (5)	7.95e−2 (5)
		0	7.98	9.01e−2 (78)	4.60e−2 (5)	4.58e−2 (5)	4.58e−2 (5)
		1000	4.00	2.15e−1 (74)	8.77e−2 (5)	7.58e−2 (5)	7.36e−2 (5)
		1000	7.99	8.72e−2 (69)	4.34e−2 (5)	4.31e−2 (5)	4.31e−2 (5)
	Non-uniform	0	3.99	3.37e−1 (91)	1.53e−1 (5)	1.18e−1 (5)	1.07e−1 (5)
		0	7.98	1.37e−1 (128)	5.62e−2 (5)	5.33e−2 (5)	5.31e−2 (5)
		1000	4.00	3.11e−1 (93)	1.39e−1 (6)	1.06e−1 (5)	9.55e−2 (5)
		1000	7.99	1.29e−1 (104)	5.21e−2 (5)	4.91e−2 (5)	4.89e−2 (5)

that attain the true rank, presuming that the true matrix is known. The two-step results are reported as NNPLS1 and NNPLS2, respectively, in Table 3. Besides the relative recovery error (**relerr**), we also report the (squared) **fidelity**, which is a measure of the closeness of two quantum states defined by $\|\widehat{X}_m^{1/2}\overline{X}^{1/2}\|_*^2$. From Table 3, we can see that the RCS estimator is superior to the NNPLS2 estimator in terms of both the fidelity and the relative error.

For the rectangular matrix completion problems, we generated the true matrix $\overline{X}$ by the following command:

```
ML = randn(nr,r);   MR = randn(nc,r);
 MW = weight*ML (:,1:k);
ML(:,1:k) = MW;   X_bar = ML*MR'.
```

We set `weight` $= 2$, `k` $= 1$ and took $\overline{X} =$ `X_bar` with different dimensions and ranks. Both the uniform sampling scheme and the non-uniform sampling scheme were tested for comparison. For the non-uniform sampling scheme, the probability to sample the first $1/4$ rows and the first $1/4$ columns were 3 times as much as that of other rows and columns respectively. In other words, the density of sampled entries in the top-left part was 3 times as much as that in the bottom-left part and the top-right part respectively and 9 times as much as that in the bottom-right part. We added 10 % i.i.d. Gaussian noise to the sampled entries. We also fixed partial entries of $\overline{X}$ uniformly from the rest un-sampled entries. The upper bound of the non-fixed entries was set to be double of the largest absolution value among all the noisy observations of entries together with the fixed entries. What we observe from Table 4 for the rectangular matrix completion is similar to that for the covariance matric completion. Moreover, we can see that the non-uniform sampling scheme greatly weakens the recoverability of the NNPLS estimator in terms of both the recovery error and the rank, especially when the sample ratio is low. Meanwhile, the advantage of the RCS estimators in such cases becomes more remarkable.

7 Conclusions

In this paper, we proposed a rank-corrected procedure for low-rank matrix completion problems with fixed basis coefficients. This approach can substantially overcome the limitation of the nuclear norm technique for recovering a low-rank matrix. We confirmed the improvement of the rank-correction step in both the reduction of recovery error and the achievement of rank consistency (in the sense of Bach [3]). Due to the presence of fixed basis coefficients, constraint nondegeneracy plays an important role in our analysis. Extensive numerical experiments show that our approach can significantly improve the recovery performance compared with the nuclear norm penalized least square estimator. As a byproduct, our results also provide a theoretical foundation for the majorized penalty method of Gao and Sun [27] and Gao [26] for structured low-rank matrix optimization problems.

Our proposed rank-correction step also allows additional constraints according to other possible prior information. In order to better fit the under-sampling setting of matrix completion, in the future work, it would be of great interest to extend the

asymptotic rank consistency results to the case where the matrix size is allowed to grow. It would also be interesting to extend this approach to deal with other low-rank matrix problems.

Acknowledgments The authors would like to thank Professor Wotao Yin for his valuable comments on possibly choosing the optimal penalty parameter for recovery error bounds and Dr. Kaifeng Jiang for helpful discussions on efficiently solving the density matrix completion problem.

Appendix 1: Spectral operator

The concept of spectral operator is associated with a symmetric vector-valued function. A function $f : \mathbb{R}^n \to \mathbb{R}^n$ is said to be symmetric if

$$f(x) = Q^{\mathbb{T}} f(Qx) \quad \forall \text{ signed permutation matrix } Q \text{ and } x \in \mathbb{R}^n,$$

where a signed permutation matrix is a real matrix that contains exactly one nonzero entry 1 or -1 in each row and column and 0 elsewhere. From this definition, we see that

$$f_i(x) = 0 \quad \text{if } x_i = 0.$$

The spectral operator $F : \mathbb{V}^{n_1 \times n_2} \to \mathbb{V}^{n_1 \times n_2}$ associated with the function f is defined by

$$F(X) := U \mathrm{Diag}\big(f(\sigma(X))\big) V^{\mathbb{T}}, \tag{27}$$

where $(U, V) \in \mathbb{O}^{n_1, n_2}(X)$ and $X \in \mathbb{V}^{n_1 \times n_2}$. From [10, Theorems 3.1 & 3.6], the symmetry of f guarantees the well-definiteness of the spectral operator F, and the (continuous) differentiability of f implies the (continuous) differentiability of F. When $\mathbb{V}^{n_1 \times n_2} = \mathbb{S}^n$, we have that

$$F(X) = P \mathrm{Diag}\big(f(|\lambda(X)|)\big)\big(P \mathrm{Diag}(s(X))\big)^{\mathbb{T}},$$

where $P \in \mathbb{O}^n(X)$ and $s(X) \in \mathbb{R}^n$ with the i-th component $s_i(X) = -1$ if $\lambda_i(X) < 0$ and $s_i(X) = 1$ otherwise. In particular for the positive semidefinite case, both U and V in (27) reduce to P. For more details on spectral operators, the readers may refer to the PhD thesis [10].

Appendix 2: Constraint nondegeneracy

Consider the following constrained optimization problem

$$\min_{X \in \mathbb{V}^{n_1 \times n_2}} \Big\{ \Phi(X) + \Psi(X) : \ \mathcal{A}(X) - b \in K \Big\}, \tag{28}$$

where $\Phi : \mathbb{V}^{n_1\times n_2} \to \mathbb{R}$ is a continuously differentiable function, $\Psi : \mathbb{V}^{n_1\times n_2} \to \mathbb{R}$ is a convex function, $\mathcal{A} : \mathbb{V}^{n_1\times n_2} \to \mathbb{R}^l$ is a linear operator and $K \subseteq \mathbb{R}^l$ is a closed convex set. Let $\widehat{X}$ be a given feasible point of (28) and $\widehat{z} := \mathcal{A}(\widehat{X}) - b$. When Ψ is differentiable at $\widehat{X}$, we say that the constraint nondegeneracy holds at $\widehat{X}$ if

$$\mathcal{A}\,\mathbb{V}^{n_1\times n_2} + \mathrm{lin}\big(\mathcal{T}_K(\widehat{z})\big) = \mathbb{R}^l, \tag{29}$$

where $\mathcal{T}_K(\widehat{z})$ denotes the tangent cone of K at $\widehat{z}$ and $\mathrm{lin}(\mathcal{T}_K(\widehat{z}))$ denotes the largest linearity space contained in $\mathcal{T}_K(\widehat{z})$, i.e., $\mathrm{lin}(\mathcal{T}_K(\widehat{z})) = \mathcal{T}_K(\widehat{z}) \cap (-\mathcal{T}_K(\widehat{z}))$. When the function Ψ is nondifferentiable, we can rewrite the optimization problem (28) equivalently as

$$\min_{X\in\mathbb{V}^{n_1\times n_2},\, t\in\mathbb{R}} \Big\{\Phi(X) + t :\ \widetilde{\mathcal{A}}(X,t) \in K \times \mathrm{epi}\Psi\Big\},$$

where $\mathrm{epi}\Psi := \big\{(X,t) \in \mathbb{V}^{n_1\times n_2} \times \mathbb{R} \mid \Psi(X) \le t\big\}$ denotes the epigraph of Ψ and $\widetilde{\mathcal{A}} : \mathbb{V}^{n_1\times n_2} \times \mathbb{R} \to \mathbb{R}^l \times \mathbb{V}^{n_1\times n_2} \times \mathbb{R}$ is a linear operator defined by

$$\widetilde{\mathcal{A}}(X,t) := \begin{pmatrix} \mathcal{A}(X) - b \\ X \\ t \end{pmatrix}, \quad (X,t) \in \mathbb{V}^{n_1\times n_2} \times \mathbb{R}.$$

From (29) and [67, Theorem 6.41], the constraint nondegeneracy holds at $(\widehat{X},\widehat{t})$ with $\widehat{t} = \Psi(\widehat{X})$ if

$$\widetilde{\mathcal{A}}\begin{pmatrix} \mathbb{V}^{n_1\times n_2} \\ \mathbb{R} \end{pmatrix} + \begin{pmatrix} \mathrm{lin}\big(\mathcal{T}_K(\widehat{z})\big) \\ \mathrm{lin}\big(\mathcal{T}_{\mathrm{epi}\Psi}(\widehat{X},\widehat{t})\big) \end{pmatrix} = \begin{pmatrix} \mathbb{R}^l \\ \mathbb{V}^{n_1\times n_2} \\ \mathbb{R} \end{pmatrix}.$$

By the definition of $\widetilde{\mathcal{A}}$, it is not difficult to verify that this condition is equivalent to

$$[\mathcal{A}\ \ 0]\big(\mathrm{lin}(\mathcal{T}_{\mathrm{epi}\Psi}(\widehat{X},\widehat{t}))\big) + \mathrm{lin}\big(\mathcal{T}_K(\widehat{z})\big) = \mathbb{R}^l. \tag{30}$$

One can see that the problem (3) with $\mathcal{C} = \mathbb{V}^{n_1\times n_2}$ can be cast into (28) with $\Psi = \|\cdot\|_*$, $\mathcal{A} = [\mathcal{R}_\alpha\ \mathcal{R}_\beta]$, $K = \{0\}^{|\alpha|} \times [-b,b]^{|\beta|}$, and meanwhile the problem (3) with $\mathcal{C} = \mathbb{S}^n_+$ can be cast into (28) with $\Psi = \delta_{\mathbb{S}^n_+}$, $\mathcal{A} = \mathcal{R}_\alpha$, $K = \{0\}$. In the previous case, the condition (30) reduces to (15) according to the expression of $\mathcal{T}_{\mathrm{epi}\Psi}(\overline{X},\bar{t})$ with $\bar{t} = \|\overline{X}\|_*$ (e.g., see [34]). In the latter case, the condition (30) reduces to (16) according to Arnold's characterization of the tangent cone $\mathcal{T}_{\mathbb{S}^n_+}(\overline{X}) = \big\{H \in \mathbb{S}^n \mid \overline{P}_2^{\mathbb{T}} H \overline{P}_2 \in \mathbb{S}^{n-r}_+\big\}$ in [2].

Appendix 3: Proofs of Theorems

Proof of Theorem 1

Let $\Delta_m := \widehat{X}_m - \overline{X}$. Using the optimality of $\widehat{X}_m$ to the problem (3), we obtain that

$$\frac{1}{2m}\|\mathcal{R}_\Omega(\Delta_m)\|_2^2 \le \left\langle \frac{\nu}{m}\mathcal{R}_\Omega^*(\xi), \Delta_m \right\rangle - \rho_m\big(\|\widehat{X}_m\|_* - \|\overline{X}\|_* - \langle F(\widetilde{X}_m), \Delta_m\rangle\big). \quad (31)$$

Then, we introduce an orthogonal decomposition $\mathbb{V}^{n_1\times n_2} = T \oplus T^\perp$ with

$$\begin{cases} T := \big\{X \in \mathbb{V}^{n_1\times n_2} \mid X = X_1 + X_2 \text{ with } \mathrm{col}(X_1) \subseteq \mathrm{col}(\overline{X}) \text{ and } \mathrm{row}(X_2) \subseteq \mathrm{row}(\overline{X})\big\}, \\ T^\perp := \big\{X \in \mathbb{V}^{n_1\times n_2} \mid \mathrm{row}(X) \perp \mathrm{row}(\overline{X}) \text{ and } \mathrm{col}(X) \perp \mathrm{col}(\overline{X})\big\}, \end{cases}$$

where $\mathrm{row}(X)$ and $\mathrm{col}(X)$ denote the row space and column space of X, respectively. Let $\mathcal{P}_T$ and $\mathcal{P}_{T^\perp}$ be orthogonal projections onto T and $T^\perp$, respectively, given by

$$\mathcal{P}_T(X) = \overline{U}_1\overline{U}_1^{\mathbb{T}}X + X\overline{V}_1\overline{V}_1^{\mathbb{T}} - \overline{U}_1\overline{U}_1^{\mathbb{T}}X\overline{V}_1\overline{V}_1^{\mathbb{T}} \text{ and } \mathcal{P}_{T^\perp}(X) = \overline{U}_2\overline{U}_2^{\mathbb{T}}X\overline{V}_2\overline{V}_2^{\mathbb{T}} \quad (32)$$

for any $X \in \mathbb{V}^{n_1\times n_2}$ and $(\overline{U}, \overline{V}) \in \mathbb{O}^{n_1,n_2}(\overline{X})$. Then, it follows from the choice of ρ_m that

$$\left\langle \frac{\nu}{m}\mathcal{R}_\Omega^*(\xi), \Delta_m \right\rangle \le \left\| \frac{\nu}{m}\mathcal{R}_\Omega^*(\xi) \right\| \|\Delta_m\|_* \le \frac{\rho_m}{\kappa}\big(\|\mathcal{P}_T(\Delta_m)\|_* + \|\mathcal{P}_{T^\perp}(\Delta_m)\|_*\big). \quad (33)$$

Moreover, from the directional derivative of the nuclear norm at $\overline{X}$, (see [75, Theorem 1]), we have

$$\begin{aligned} \|\widehat{X}_m\|_* - \|\overline{X}\|_* - \langle F(\widetilde{X}_m), \Delta_m\rangle &\ge \langle \overline{U}_1\overline{V}_1^{\mathbb{T}}, \Delta_m\rangle + \|\overline{U}_2^{\mathbb{T}}\Delta_m\overline{V}_2\|_* - \langle F(\widetilde{X}_m), \Delta_m\rangle \\ &\ge \|\mathcal{P}_{T^\perp}(\Delta_m)\|_* - \|\overline{U}_1\overline{V}_1^{\mathbb{T}} - F(\widetilde{X}_m)\|_F\|\Delta_m\|_F \\ &= \|\mathcal{P}_{T^\perp}(\Delta_m)\|_* - a_m\sqrt{r}\|\Delta_m\|_F. \end{aligned} \quad (34)$$

Then, by substituting (33) and (34) into (31), we have

$$\frac{1}{2m}\|\mathcal{R}_\Omega(\Delta_m)\|_2^2 \le \rho_m\left(a_m\sqrt{r}\|\Delta_m\|_F + \frac{1}{\kappa}\|\mathcal{P}_T(\Delta_m)\|_* - \frac{\kappa-1}{\kappa}\|\mathcal{P}_{T^\perp}(\Delta_m)\|_*\right). \quad (35)$$

Note that $\mathrm{rank}(\mathcal{P}_T(\Delta_m)) \le 2r$. Hence, $\|\mathcal{P}_T(\Delta_m)\|_* \le \sqrt{2r}\|\mathcal{P}_T(\Delta_m)\|_F \le \sqrt{2r}\|\Delta_m\|_F$ and then the desired result (7) follows.

Proof of Theorem 2

We first show that the sampling operator $\mathcal{R}_\Omega$ satisfies some RIP-like property for matrices specified in a certain set with high probability. Similar results can also be found in [40,44,49,58].

For this purpose, define

$$\vartheta_m := \mathbb{E}\left\|\frac{1}{m}\mathcal{R}_\Omega^*(\epsilon)\right\| \quad \text{with } \epsilon = (\epsilon_1, \ldots, \epsilon_m)^{\mathbb{T}}, \tag{36}$$

where $\{\epsilon_1, \ldots, \epsilon_m\}$ is an i.i.d. Rademacher sequence, i.e., an i.i.d. sequence of Bernoulli random variables taking the values 1 and -1 with probability $1/2$.

Lemma 11 *Given any $s > 0$ and $t > 0$, define*

$$K(s,t) := \Big\{\Delta \in \mathbb{V}^{n_1\times n_2} \;\Big|\; \mathcal{R}_\alpha(\Delta) = 0,\ \|\mathcal{R}_\beta(\Delta)\|_\infty = 1,\ \|\Delta\|_* \le s\|\Delta\|_F,\ \langle \mathcal{Q}_\beta(\Delta), \Delta\rangle \ge t\Big\}.$$

Then, for any given $\gamma > 1$, $\tau_1 \in (0,1)$ and $\tau_2 \in (0, \tau_1/\gamma)$, with probability at least $1 - \frac{\exp(-(\tau_1-\gamma\tau_2)^2 mt^2/2)}{1-\exp(-(\gamma^2-1)(\tau_1-\gamma\tau_2)^2 mt^2/2)}$,

$$\frac{1}{m}\|\mathcal{R}_\Omega(\Delta)\|_2^2 \ge (1-\tau_1)\langle \mathcal{Q}_\beta(\Delta), \Delta\rangle - \frac{16}{\tau_2}s^2\mu_1 d_2 \vartheta_m^2 \quad \forall\, \Delta \in K(s,t). \tag{37}$$

Proof The proof is similar to that of [40, Lemma 12]. For any $s, t > 0$, $\gamma > 1$, $\tau_1 \in (0,1)$ and $\tau_2 \in (0, \tau_1/\gamma)$, we need to show that the event

$$E = \Big\{\exists\, \Delta \in K(s,t) \text{ such that } \Big|\frac{1}{m}\|\mathcal{R}_\Omega(\Delta)\|_2^2 - \langle \mathcal{Q}_\beta(\Delta), \Delta\rangle\Big| \ge \tau_1\langle \mathcal{Q}_\beta(\Delta), \Delta\rangle + \frac{16}{\tau_2}s^2\mu_1 d_2\vartheta_m^2\Big\}$$

occurs with probability less than $\frac{\exp(-(\tau_1-\gamma\tau_2)^2 mt^2/2)}{1-\exp(-(\gamma^2-1)(\tau_1-\gamma\tau_2)^2 mt^2/2)}$. We decompose $K(s,t)$ as

$$K(s,t) = \bigcup_{k=1}^{\infty}\Big\{\Delta \in K(s,t) \;\Big|\; \gamma^{k-1}t \le \langle \mathcal{Q}_\beta(\Delta), \Delta\rangle \le \gamma^k t\Big\}.$$

For any $a \ge t$, we further define $K(s,t,a) := \{\Delta \in K(s,t) \mid \langle \mathcal{Q}_\beta(\Delta), \Delta\rangle \le a\}$. Then we get $E \subseteq \bigcup_{k=1}^{\infty} E_k$ with

$$E_k = \Big\{\exists\, \Delta \in K(s,t,\gamma^k t) \text{ such that } \Big|\frac{1}{m}\|\mathcal{R}_\Omega(\Delta)\|_2^2 - \langle \mathcal{Q}_\beta(\Delta), \Delta\rangle\Big| \ge \gamma^{k-1}\tau_1 t + \frac{16}{\tau_2}s^2\mu_1 d_2\vartheta_m^2\Big\}.$$

Now we need to estimate the probability of each event E_k. Define

$$Z_a := \sup_{\Delta\in K(s,t,a)}\Big|\frac{1}{m}\|\mathcal{R}_\Omega(\Delta)\|_2^2 - \langle \mathcal{Q}_\beta(\Delta), \Delta\rangle\Big|.$$

Notice that for any $\Delta \in \mathbb{V}^{n_1\times n_2}$,

$$\frac{1}{m}\|\mathcal{R}_\Omega(\Delta)\|_2^2 = \frac{1}{m}\sum_{i=1}^m \langle \Gamma_{\omega_i}, \Delta\rangle^2 \overset{a.s.}{\to} \mathbb{E}(\langle \Gamma_{\omega_i}, \Delta\rangle^2) = \langle \mathcal{Q}_\beta(\Delta), \Delta\rangle.$$

Since $\|\mathcal{R}_\beta(\Delta)\|_\infty \le 1$ for all $\Delta \in K(s,t)$, from Massart's Hoeffding type concentration inequality [51, Theorem 1.4] for suprema of empirical processes, we have

$$\Pr\big(Z_a \ge \mathbb{E}(Z_a) + \varepsilon\big) \le \exp\left(-\frac{m\varepsilon^2}{2}\right) \quad \forall\, \varepsilon > 0. \tag{38}$$

Next, we use the standard Rademacher symmetrization in the theory of empirical processes to further derive an upper bound of $\mathbb{E}(Z_a)$. Let $\{\epsilon_1, \ldots, \epsilon_m\}$ be a Rademacher sequence. Then, we have

$$\begin{aligned}
\mathbb{E}(Z_a) &= \mathbb{E}\left(\sup_{\Delta\in K(s,t,a)} \left|\frac{1}{m}\sum_{i=1}^m \langle \Gamma_{\omega_i}, \Delta\rangle^2 - \mathbb{E}\big(\langle \Gamma_{\omega_i}, \Delta\rangle^2\big)\right|\right) \\
&\le 2\mathbb{E}\left(\sup_{\Delta\in K(s,t,a)} \left|\frac{1}{m}\sum_{i=1}^m \epsilon_i \langle \Gamma_{\omega_i}, \Delta\rangle^2\right|\right) \le 8\mathbb{E}\left(\sup_{\Delta\in K(s,t,a)} \left|\frac{1}{m}\sum_{i=1}^m \epsilon_i \langle \Gamma_{\omega_i}, \Delta\rangle\right|\right) \\
&= 8\mathbb{E}\left(\sup_{\Delta\in K(s,t,a)} \left|\frac{1}{m}\sum_{i=1}^m \langle \mathcal{R}^*_\Omega(\epsilon), \Delta\rangle\right|\right) \le 8\mathbb{E}\left\|\frac{1}{m}\mathcal{R}^*_\Omega(\epsilon)\right\|\left(\sup_{\Delta\in K(s,t,a)} \|\Delta\|_*\right),
\end{aligned} \tag{39}$$

where the first inequality follows from the symmetrization theorem (e.g., see [73, Lemma 2.3.1] and [6, Theorem 14.3]) and the second inequality follows from the contraction theorem (e.g., see [46, Theorem 4.12] and [6, Theorem 14.4]). Moreover, from (8), we have

$$\langle \mathcal{Q}_\beta(\Delta), \Delta\rangle \ge (\mu_1 d_2)^{-1}\|\Delta\|_F^2 \quad \forall\, \Delta \in \{\Delta \in \mathbb{V}^{n_1\times n_2} \mid \mathcal{R}_\alpha(\Delta) = 0\}. \tag{40}$$

This leads to

$$\|\Delta\|_* \le s\|\Delta\|_F \le s\sqrt{\mu_1 d_2 \langle \mathcal{Q}_\beta(\Delta), \Delta\rangle} \le s\sqrt{\mu_1 d_2 a} \quad \forall\, \Delta \in K(s,t,a). \tag{41}$$

Combining (39) and (41) with the definition of ϑ_m in (36), we obtain that

$$\mathbb{E}(Z_a) + \left(\frac{\tau_1}{\gamma} - \tau_2\right)a \le 8\vartheta_m s\sqrt{\mu_1 d_2 a} + \left(\frac{\tau_1}{\gamma} - \tau_2\right)a \le \frac{16}{\tau_2}s^2\mu_1 d_2 \vartheta_m^2 + \frac{\tau_1}{\gamma}a,$$

where the second inequality follows from the simple fact $x_1x_2 \le (x_1^2 + x_2^2)/2$ for any $x_1, x_2 \ge 0$. Then, it follows from (38) that

$$\Pr\left(Z_a \geq \frac{\tau_1}{\gamma}a + \frac{16}{\tau_2}s^2\mu_1 d_2 \vartheta_m^2\right) \leq \Pr\left(Z_a \geq \mathbb{E}(Z_a) + \left(\frac{\tau_1}{\gamma} - \tau_2\right)a\right)$$
$$\leq \exp\left(-\left(\frac{\tau_1}{\gamma} - \tau_2\right)^2 \frac{ma^2}{2}\right).$$

This implies that $\Pr(E_k) \leq \exp\left(-\frac{1}{2}\gamma^{2(k-1)}(\tau_1 - \gamma\tau_2)^2 mt^2\right)$. Then, since $\gamma > 1$, by using $\gamma^k \geq 1 + k(\gamma - 1)$ for any $k \geq 1$, we have

$$\begin{aligned}\Pr(E) &\leq \sum_{k=1}^{\infty}\Pr(E_k) \leq \sum_{k=1}^{\infty}\exp\left(-\frac{1}{2}\gamma^{2(k-1)}(\tau_1 - \gamma\tau_2)^2 mt^2\right)\\ &\leq \exp\left(-\frac{1}{2}(\tau_1 - \gamma\tau_2)^2 mt^2\right)\sum_{k=1}^{\infty}\exp\left(-\frac{1}{2}(k-1)(\gamma^2-1)(\tau_1 - \gamma\tau_2)^2 mt^2\right)\\ &\leq \frac{\exp(-(\tau_1 - \gamma\tau_2)^2 mt^2/2)}{1 - \exp(-(\gamma^2-1)(\tau_1 - \gamma\tau_2)^2 mt^2/2)}.\end{aligned}$$

Thus, we complete the proof of Lemma 11.

Now we proceed with the proof of Theorem 2. Let $\Delta_m := \widehat{X}_m - \overline{X}$. Notice that the equality (35) implies that

$$\|\mathcal{P}_{T^\perp}(\Delta_m)\|_* \leq \frac{1}{\kappa - 1}\|\mathcal{P}_T(\Delta_m)\|_* + \frac{\kappa}{\kappa - 1}a_m\sqrt{r}\|\Delta_m\|_F.$$

This, together with $\|\mathcal{P}_T(\Delta_m)\|_* \leq \sqrt{2r}\|\Delta_m\|_F$, leads to

$$\|\Delta_m\|_* \leq \|\mathcal{P}_T(\Delta_m)\|_* + \|\mathcal{P}_{T^\perp}(\Delta_m)\|_* \leq \frac{\kappa}{\kappa - 1}\left(\sqrt{2} + a_m\right)\sqrt{r}\|\Delta_m\|_F. \tag{42}$$

Let $b_m := \|\mathcal{R}_\beta(\Delta_m)\|_\infty \leq 2b$. For any fixed $c > 0$, $\gamma > 1$, $\tau_1 \in (0, 1)$ and $\tau_2 \in (0, \tau_2/\gamma)$, define $t_m := \sqrt{\frac{2c\log(n_1+n_2)}{(\tau_1 - \gamma\tau_2)^2 m}}$ so that direct calculation yields

$$\frac{\exp(-(\tau_1 - \gamma\tau_2)^2 mt_m^2/2)}{1 - \exp(-(\gamma^2-1)(\tau_1 - \gamma\tau_2)^2 mt_m^2/2)} = \frac{(n_1+n_2)^{-c}}{1 - (n_1+n_2)^{-(\gamma^2-1)c}} \leq \frac{(n_1+n_2)^{-c}}{1 - 2^{-(\gamma^2-1)c}}.$$

Then we separate the discussion into two cases:

Case 1: $\langle\mathcal{Q}_\beta(\Delta_m), \Delta_m\rangle \leq b_m^2 t_m$. It follows from (40) that $\|\Delta_m\|_F^2/d_2 \leq 4b^2\mu_1 t_m$.

Case 2: $\langle\mathcal{Q}_\beta(\Delta_m), \Delta_m\rangle > b_m^2 t_m$. It follows from (42) that $\Delta_m/b_m \in K(s_m, t_m)$ with $s_m := \frac{\kappa}{\kappa-1}\left(\sqrt{2} + a_m\right)\sqrt{r}$. Then for any given τ_3 satisfying $0 < \tau_3 < 1$, we obtain that with probability at least $1 - \frac{(n_1+n_2)^{-c}}{1-2^{-(\gamma^2-1)c}}$,

$$\begin{aligned}\frac{\|\Delta_m\|_F^2}{d_2} &\le \mu_1\langle\mathcal{Q}_\beta(\Delta_m),\Delta_m\rangle \le \frac{\mu_1}{1-\tau_1}\left(\frac{1}{m}\|\mathcal{R}_\Omega(\Delta_m)\|_2^2+\frac{16}{\tau_2}s_m^2\mu_1 d_2\vartheta_m^2 b_m^2\right)\\ &\le \frac{2}{1-\tau_1}\left(\frac{\sqrt{2}}{\kappa}+a_m\right)\mu_1\rho_m\sqrt{r}\|\Delta_m\|_F+\frac{16}{(1-\tau_1)\tau_2}s_m^2\mu_1^2 d_2\vartheta_m^2 b_m^2\\ &\le \tau_3\frac{\|\Delta_m\|_F^2}{d_2}+\frac{2}{(1-\tau_1)^2\tau_3}\left(\frac{\sqrt{2}}{\kappa}+a_m\right)^2\mu_1^2\rho_m^2 r d_2+\frac{16}{(1-\tau_1)\tau_2}s_m^2\mu_1^2 d_2\vartheta_m^2 b_m^2,\end{aligned}$$

where the first inequality follows from (40), the second inequality follows from Lemma 11 and the third inequality follows from Theorem 1. Plugging in s_m further leads to

$$\frac{\|\Delta_m\|_F^2}{d_2}\le\frac{\mu_1^2 d_2 r}{1-\tau_3}\left(\frac{2}{(1-\tau_1)^2\tau_3}\left(\frac{\sqrt{2}}{\kappa}+a_m\right)^2\rho_m^2+\frac{64}{(1-\tau_1)\tau_2}\left(\frac{\kappa}{\kappa-1}\right)^2(\sqrt{2}+a_m)^2\vartheta_m^2 b^2\right).$$

Combing the above two cases together, with γ, τ_1, τ_2 and τ_3 chosen to be absolute constants, we arrive at an intermediate result that there exist some positive absolute constants c_0', c_1', c_2' and C_0' such that for any $\kappa>1$, if ρ_m is chosen as in Theorem 1, then with probability at least $1-c_1'(n_1+n_2)^{-c_2'}$,

$$\begin{aligned}\frac{\|\widehat{X}_m-\overline{X}\|_F^2}{d_2}\le C_0'\max\Bigg\{&\mu_1^2 d_2 r\left(c_0'^2\left(\frac{\sqrt{2}}{\kappa}+a_m\right)^2\rho_m^2+\left(\frac{\kappa}{\kappa-1}\right)^2(\sqrt{2}+a_m)^2\vartheta_m^2 b^2\right),\\ &b^2\mu_1\sqrt{\frac{\log(n_1+n_2)}{m}}\Bigg\}.\end{aligned}\tag{43}$$

To further derive explicit estimations of ρ_m and ϑ_m, we introduce the noncommutative Bernstein inequality taken from [42, Corollary 2.1], which provides a probability control of the deviation of the sum of random matrices from its mean in the operator norm. The noncommutative Bernstein inequality introduced here is a recently-extended version, with the random matrices being controlled by the Orlicz norms (see [42–44]) rather than the operator norm (see, e.g., [30,63,72]). The Orlicz norms are used to characterize the tail behavior of random variables. Given any $s\ge 1$, the ψ_s Orlicz norm of a random variable z is defined by $\|z\|_{\psi_s}:=\inf\{t>0\,\big|\,\mathbb{E}\exp(|z|^s/t^s)\le 2\}$.

Lemma 12 (Koltchinskii [42]) *Let $Z_1,\ldots,Z_m\in\mathbb{V}^{n_1\times n_2}$ be independent random matrices with mean zero. Suppose that* $\max\left\{\left\|\|Z_i\|\right\|_{\psi_s},2\mathbb{E}^{\frac{1}{2}}(\|Z_i\|^2)\right\}<\varpi_s$ *for some constant ϖ_s. Define*

$$\sigma_Z:=\max\left\{\left\|\frac{1}{m}\sum_{i=1}^m\mathbb{E}(Z_iZ_i^{\mathbb{T}})\right\|^{1/2},\ \left\|\frac{1}{m}\sum_{i=1}^m\mathbb{E}(Z_i^{\mathbb{T}}Z_i)\right\|^{1/2}\right\}.$$

Then, there exists a constant C such that for all $t > 0$, with probability at least $1-\exp(-t)$,

$$\left\| \frac{1}{m} \sum_{i=1}^{m} Z_i \right\| \leq C \max \left\{ \sigma_Z \sqrt{\frac{t+\log(n_1+n_2)}{m}}, \varpi_s \left(\log \frac{\varpi_s}{\sigma_Z} \right)^{1/s} \frac{t+\log(n_1+n_2)}{m} \right\}.$$

With the help of Lemma 12, we obtain the following result, which is an extension of [44, Lemma 2] and [40, Lemmas 5 and 6] from the standard basis to an arbitrary orthonormal basis. A similar result can also be found in [58, Lemma 6].

Lemma 13 *Under Assumption 2, there exists a positive constant C' (only depending on the ψ_1 Orlicz norm of ξ_k) such that for all $t > 0$, with probability at least $1 - \exp(-t)$,*

$$\left\| \frac{1}{m} \mathcal{R}_{\Omega}^{*}(\xi) \right\| \leq C' \max \left\{ \sqrt{\frac{\mu_2(t+\log(n_1+n_2))}{\sqrt{d_2}m}}, \frac{\log(d_2)(t+\log(n_1+n_2))}{2m} \right\}. \tag{44}$$

In particular, when $m \geq \sqrt{d_2}\log^3(n_1+n_2)/\mu_2$, we also have

$$\mathbb{E} \left\| \frac{1}{m} \mathcal{R}_{\Omega}^{*}(\xi) \right\| \leq C' \sqrt{\frac{2e\mu_2 \log(n_1+n_2)}{\sqrt{d_2}m}}, \tag{45}$$

where e is the exponential constant.

Proof Recall that $\frac{1}{m}\mathcal{R}_{\Omega}^{*}(\xi) = \frac{1}{m}\sum_{i=1}^{m} \xi_i \Theta_{\omega_i}$. Let $Z_i := \xi_i \Theta_{\omega_i}$. Since $\mathbb{E}(\xi_i) = 0$, the independence of ξ_i and Θ_{ω_i} implies that $\mathbb{E}(Z_i) = 0$. Since $\|\Theta_{\omega_i}\|_F = 1$, we have that $\|Z_i\| \leq \|Z_i\|_F = |\xi_i|\|\Theta_{\omega_i}\|_F = |\xi_i|$. It follows that $\|\|Z_i\|\|_{\psi_1} \leq \|\xi_i\|_{\psi_1}$ and thus finite. (It is known that a random variable is sub-exponential if and only its ψ_1 Orlicz norm is finite [73]). Meanwhile, $\mathbb{E}^{\frac{1}{2}}(\|Z_i\|^2) \leq \mathbb{E}^{\frac{1}{2}}(\|Z_i\|_F^2) = \mathbb{E}^{\frac{1}{2}}(\xi_i^2) = 1$. Then direct calculation yields

$$\mathbb{E}\left(Z_i Z_i^{\mathbb{T}}\right) = \mathbb{E}\left(\xi_i^2 \Theta_{\omega_i} \Theta_{\omega_i}^{\mathbb{T}}\right) = \mathbb{E}\left(\Theta_{\omega_i} \Theta_{\omega_i}^{\mathbb{T}}\right) = \sum_{k\in\beta} p_k \Theta_k \Theta_k^{\mathbb{T}}.$$

The calculation for $\mathbb{E}\left(Z_i^{\mathbb{T}} Z_i\right)$ is similar. We obtain from (8) that $1/\sqrt{d_2} \leq \sigma_Z^2 \leq \mu_2/\sqrt{d_2}$. Then, applying this to Lemma 12 yields (44). The remaining proof of (45) follows the same as the proof of Lemma 6 in [40]. For simplicity, we omit it.

A good estimation of ρ_m can be achieved by choosing $t = c_2' \log(n_1+n_2)$ in Lemma 13 for an optimal order bound, where c_2' is the same as that in (43). With this choice, when $m \geq 4(1+c_2')\sqrt{d_2}\log^2(d_2)\log(n_1+n_2)/\mu_2$, the first term in the maximum of (44) dominates the second one. Thus, with probability at least $1-(n_1+n_2)^{-c_2'}$, one can choose

$$\rho_m = \kappa\nu \cdot C' \sqrt{\frac{(1+c_2')\mu_2 \log(n_1+n_2)}{\sqrt{d_2}m}}.$$

Moreover, since Bernoulli random variables are sub-exponential, Lemma 13 also provides an upper bound of ϑ_m in (45). It is worthwhile to note that after plugging the above estimations of ρ_m and ϑ_m, the second term in the maximum of (43) is negligible compared with the first term. Therefore, the second term is further dropped for simplicity and thus we complete the proof.

Proof of Theorem 3

For notational simplicity, we drop the subscript of $\widetilde{X}_m$ in this proof. With $(\widetilde{U}, \widetilde{V}) \in \mathbb{O}^{n_1,n_2}(\widetilde{X})$, one immediately obtains from the definition of a_m in (6) that

$$a_m \le \frac{1}{\sqrt{r}}\big(\|F(\widetilde{X}) - \widetilde{U}_1\widetilde{V}_1^{\mathbb{T}}\|_F + \|\widetilde{U}_1\widetilde{V}_1^{\mathbb{T}} - \overline{U}_1\overline{V}_1^{\mathbb{T}}\|_F\big) \le \varepsilon_F(\widetilde{X}) + \frac{1}{\sqrt{r}}\|\widetilde{U}_1\widetilde{V}_1^{\mathbb{T}} - \overline{U}_1\overline{V}_1^{\mathbb{T}}\|_F. \tag{46}$$

The left proof is to find an upper bound of $\|\widetilde{U}_1\widetilde{V}_1^{\mathbb{T}} - \overline{U}_1\overline{V}_1^{\mathbb{T}}\|_F$. Let $\delta := \|\widetilde{X} - \overline{X}\|_F$ and $\mathcal{N}_\delta(\overline{X}) := \{X \in \mathbb{V}^{n_1\times n_2} \mid \|X - \overline{X}\|_F \le \delta\}$.

Let $\widehat{F} : \mathbb{V}^{n_1\times n_2} \to \mathbb{V}^{n_1\times n_2}$ be a spectral operator associated with a symmetric function $\widehat{f} : \mathbb{R}^n \to \mathbb{R}^n$ given by $\widehat{f}_i(x) = \phi(x_i)$, $i = 1, \ldots, n$, where $\phi : \mathbb{R} \to \mathbb{R}$ is an odd scalar function with $\phi(t) = -\phi(-t)$ for $t < 0$, and $\phi(t)$ for $t \ge 0$ is defined as

$$\phi(t) = \begin{cases} 1 & \text{if } t \ge 2\sigma_r(\overline{X})/3 - \delta/3, \\ \frac{t-(\sigma_r(\overline{X})/3+\delta/3)}{\sigma_r(\overline{X})/3-2\sigma/3} & \text{if } \sigma_r(\overline{X})/3 + \delta/3 < t < 2\sigma_r(\overline{X})/3 - \delta/3, \\ 0 & \text{if } 0 \le t \le \sigma_r(\overline{X})/3 + \delta/3. \end{cases}$$

Note that for any $X \in \mathcal{N}_\delta(\overline{X})$,

$$|\sigma_i(X) - \sigma_i(\overline{X})| \le \sigma_1(X - \overline{X}) \le \|X - \overline{X}\|_F \le \delta, \quad i = 1, \ldots, n.$$

Since $\delta/\sigma_r(\overline{X}) < 1/2$, we further have $\sigma_r(X) \ge \sigma_r(\overline{X}) - \delta > \delta \ge \sigma_{r+1}(X)$. This means

$$\widehat{F}(X) = U_1V_1^{\mathbb{T}} \qquad \forall\, X \in \mathcal{N}_\delta(\overline{X}).$$

Moreover, $\widehat{F}$ is continuously differentiable over $\mathcal{N}_\delta(\overline{X})$. Hence, we can apply the Mean Value Theorem to obtain

$$\widetilde{U}_1\widetilde{V}_1^{\mathbb{T}} - \overline{U}_1\overline{V}_1^{\mathbb{T}} = \widehat{F}(\widetilde{X}) - \widehat{F}(\overline{X}) = \int_0^1 \widehat{F}'(\widetilde{X}_t)(\widetilde{X} - \overline{X})\,\mathrm{d}t, \tag{47}$$

where $\widetilde{X}_t := \overline{X} + t(\widetilde{X} - \overline{X})$. Clearly, $\widetilde{X}_t \in \mathcal{N}_\delta(\overline{X})$ when $t \in [0, 1]$.

Regarding (47), we need to look into the derivative of $\widehat{F}$ over $\mathcal{N}_\delta(\overline{X})$. Let $X \in \mathcal{N}_\delta(\overline{X})$ be arbitrary and $(U, V) \in \mathbb{O}^{n_1,n_2}(X)$. Without loss of generality, we assume $n_1 \le n_2$. Let $\chi_1 := \{1, \ldots, r\}$, $\chi_2 := \{r + 1, \ldots, n_1\}$ and $\chi_3 := \{n_1 + 1, \ldots, n_2\}$.

Then, according to [10, Theorem 3.6], we have that for any $H \in \mathbb{V}^{n_1\times n_2}$,

$$\widehat{F}'(X)(H) = U\left[\mathcal{E}_1(X)\circ\frac{\widetilde{H}_1+\widetilde{H}_1^{\mathbb{T}}}{2}+\mathcal{E}_2(X)\circ\frac{\widetilde{H}_1-\widetilde{H}_1^{\mathbb{T}}}{2}\quad \Upsilon(X)\circ\widetilde{H}_2\right]V^{\mathbb{T}}, \tag{48}$$

where $[\widetilde{H}_1\ \widetilde{H}_2] = \widetilde{H} := UHV^{\mathbb{T}}$ with $\widetilde{H}_1 \in \mathbb{V}^{n_1\times n_1}$, $\widetilde{H}_2 \in \mathbb{V}^{n_1\times(n_2-n_1)}$, and $\mathcal{E}_1(X) \in \mathbb{V}^{n_1\times n_1}$, $\mathcal{E}_2(X) \in \mathbb{V}^{n_1\times n_1}$, $\Upsilon(X) \in \mathbb{V}^{n_1\times(n_2-n_1)}$ take the form

$$\big(\mathcal{E}_1(X)\big)_{ij} = \begin{cases} \frac{1}{\sigma_i(X)-\sigma_j(X)} & \text{if } i\in\chi_1, j\in\chi_2 \text{ or } i\in\chi_2, j\in\chi_1,\\ 0 & \text{otherwise,}\end{cases}$$

$$\big(\mathcal{E}_2(X)\big)_{ij} = \begin{cases} \frac{2}{\sigma_i(X)+\sigma_j(X)} & \text{if } i\in\chi_1, j\in\chi_1,\\ \frac{1}{\sigma_i(X)-\sigma_j(X)} & \text{if } i\in\chi_1, j\in\chi_2 \text{ or } i\in\chi_2, j\in\chi_1,\\ 0 & \text{otherwise,}\end{cases}$$

$$\big(\Upsilon(X)\big)_{ij} = \begin{cases} \frac{1}{\sigma_i(X)} & \text{if } i\in\chi_1, j\in\chi_3,\\ 0 & \text{otherwise.}\end{cases}$$

Here, "$\circ$" stands for the Hadamard product of matrices. Let Δ denote the matrix in the bracket of (48). Moreover, let Δ_{χ_i,χ_j} and $\widetilde{H}_{\chi_i,\chi_j}$ denote the submatrices of Δ and $\widetilde{H}$ with row indices χ_i and column indices χ_j, respectively. Then, a direct calculation yields

$$\|\Delta_{\chi_1,\chi_1}\|_F^2 \le \frac{\|\widetilde{H}_{\chi_1,\chi_1}\|_F^2}{\sigma_r^2(X)}, \qquad \|\Delta_{\chi_1,\chi_2}\|_F^2+\|\Delta_{\chi_2,\chi_1}\|_F^2 \le \frac{\|\widetilde{H}_{\chi_1,\chi_2}\|_F^2+\|\widetilde{H}_{\chi_2,\chi_1}\|_F^2}{(\sigma_r(X)-\sigma_{r+1}(X))^2},$$

$$\|\Delta_{\chi_2,\chi_2}\|_F^2 = 0, \quad \|\Delta_{\chi_1,\chi_3}\|_F^2 \le \frac{\|\widetilde{H}_{\chi_1,\chi_3}\|_F^2}{\sigma_r^2(X)} \quad \text{and} \quad \|\Delta_{\chi_2,\chi_3}\|_F^2 = 0.$$

Note that $\|\widehat{F}'(X)(H)\|_F = \|\Delta\|_F$ and $\|\widetilde{H}\|_F = \|H\|_F$. By summing up the above inequalities together, we obtain that for any $X \in \mathcal{N}_\delta(\overline{X})$,

$$\big\|\widehat{F}'(X)(H)\big\|_F \le \frac{\sqrt{\|H_{\chi_1,\chi_1\cup\chi_2\cup\chi_3}\|_F^2+\|H_{\chi_2,\chi_1}\|_F^2}}{\sigma_r(X)-\sigma_{r+1}(X)} \le \frac{\|H\|_F}{\sigma_r(X)-\sigma_{r+1}(X)}. \tag{49}$$

Now, we proceed with the proof by applying (49) to (47). This leads to

$$\big\|\widetilde{U}_1\widetilde{V}_1^{\mathbb{T}}-\overline{U}_1\overline{V}_1^{\mathbb{T}}\big\|_F \le \int_0^1 \big\|\widehat{F}'(\widetilde{X}_t)(\widetilde{X}-\overline{X})\big\|_F\,\mathrm{d}t \le \int_0^1 \frac{\delta}{\sigma_r(\widetilde{X}_t)-\sigma_{r+1}(\widetilde{X}_t)}\,\mathrm{d}t. \tag{50}$$

Moreover, using [4, Theorems IV.3.4& II.3.1], we have

$$\begin{aligned}&\big(\sigma_r(\widetilde{X}_t)-\sigma_r(\overline{X})\big)^2+\sigma_{r+1}^2(\widetilde{X}_t)\\ &\quad\le \big\|\sigma(\widetilde{X}_t)-\sigma(\overline{X})\big\|_F^2 \le \big\|\sigma(\widetilde{X}_t-\overline{X})\big\|_F^2 = \big\|\widetilde{X}_t-\overline{X}\big\|_F^2 \le t^2\delta^2.\end{aligned}$$

This implies that $\sigma_r(\widetilde{X}_t)-\sigma_r(\overline{X})=\delta_t\cos\theta$ and $\sigma_{r+1}(\widetilde{X}_t)=\delta_t\sin\theta$ for some $\delta_t\le t\delta$ and $\theta\in[0,2\pi)$. Thus,

$$\sigma_r(\widetilde{X}_t)-\sigma_{r+1}(\widetilde{X}_t)=\sigma_r(\overline{X})+\delta_t\cos\theta-\delta_t\sin\theta\ge\sigma_r(\overline{X})-\sqrt{2}\delta_t\ge\sigma_r(\overline{X})-\sqrt{2}t\delta. \tag{51}$$

Substituting (51) into (50), we obtain that

$$\|\widetilde{U}_1\widetilde{V}_1^{\mathbb{T}}-\overline{U}_1\overline{V}_1^{\mathbb{T}}\|_F\le\int_0^1\frac{\delta}{\sigma_r(\overline{X})-\sqrt{2}t\delta}\,\mathrm{d}t=-\frac{1}{\sqrt{2}}\log\left(1-\frac{\sqrt{2}\,\delta}{\sigma_r(\overline{X})}\right).$$

This, together with (46), completes the proof.

Proof of Theorem 4

We first prove the following properties of the sample operator $\mathcal{R}_\Omega$ and its adjoint $\mathcal{R}_\Omega^*$.

Lemma 14 *(i) For any given $X\in\mathbb{V}^{n_1\times n_2}$, the random matrix $\frac{1}{m}\mathcal{R}_\Omega^*\mathcal{R}_\Omega(X)\xrightarrow{a.s.}\mathcal{Q}_\beta(X)$.*

(ii) The random vector $\frac{1}{\sqrt{m}}\mathcal{R}_{\alpha\cup\beta}\mathcal{R}_\Omega^*(\xi)\xrightarrow{d}N\big(0,\mathrm{Diag}(p)\big)$, where $p=(p_1,\ldots,p_d)^{\mathbb{T}}$.

Proof (i) It follows from the definitions of $\mathcal{R}_\Omega$ and its adjoint $\mathcal{R}_\Omega^*$ that $\frac{1}{m}\mathcal{R}_\Omega^*\mathcal{R}_\Omega(X)=\frac{1}{m}\sum_{i=1}^m\langle\Theta_{\omega_i},X\rangle\,\Theta_{\omega_i}$. This is an average value of m i.i.d. random matrices $\langle\Theta_{\omega_i},X\rangle\Theta_{\omega_i}$. Note that $\mathbb{E}\big(\langle\Theta_{\omega_i},X\rangle\Theta_{\omega_i}\big)=\mathcal{Q}_\beta(X)\ \forall\, i=1,\cdots,m$. Then the result follows directly from the strong law of large numbers.

(ii) It directly follows from the definitions of $\mathcal{R}_\Omega^*$ and $\mathcal{R}_{\alpha\cup\beta}$ that $\frac{1}{\sqrt{m}}\mathcal{R}_{\alpha\cup\beta}\mathcal{R}_\Omega^*(\xi)=\frac{1}{\sqrt{m}}\mathcal{R}_{\alpha\cup\beta}\big(\sum_{i=1}^m\xi_i\Theta_{\omega_i}\big)=\frac{1}{\sqrt{m}}\sum_{i=1}^m\xi_i\mathcal{R}_{\alpha\cup\beta}(\Theta_{\omega_i})$. Since $\mathbb{E}(\xi_i)=0$ and $\mathbb{E}(\xi_i^2)=1$, according to the independence of ξ_i and $\mathcal{R}_{\alpha\cup\beta}(\Theta_{\omega_i})$, we obtain $\mathbb{E}\big(\xi_i\mathcal{R}_{\alpha\cup\beta}(\Theta_{\omega_i})\big)=0$ and $\mathrm{cov}\big(\xi_i\mathcal{R}_{\alpha\cup\beta}(\Theta_{\omega_i})\big)=\mathrm{Diag}(p)$. Then, applying the vector-valued central limit theorem yields the result.

To prove the convergence in distribution of minimizers, the following theorem of Knight [41, Theorem 1] on epi-convergence in distribution is particularly useful in this regard (see also [32, Proposition 9]).

Lemma 15 (Knight [41]) *Let $\{\Phi_m\}$ be a sequence of random lower-semicontinuous functions that epi-converges in distribution to Φ. Assume that*

(i) $\widehat{x}_m$ is an ε_m-minimizer of Φ_m, i.e., $\Phi_m(\widehat{x}_m)\le\inf\Phi_m(x)+\varepsilon_m$, where $\varepsilon_m\xrightarrow{p}0$;
(ii) $\widehat{x}_m=O_p(1)$;
(iii) the function Φ has a unique minimizer $\overline{x}$.

Then, $\widehat{x}_m\xrightarrow{d}\overline{x}$. In addition, if Φ is a deterministic function, then $\widehat{x}_m\xrightarrow{p}\overline{x}$.

It is know from [29] that $\widehat{x}_m$ is guaranteed to be $O_p(1)$ when all Φ_m are convex functions and Φ has a unique minimizer. For more details on epi-convergence in distribution, one may refer to King and Wets [38], Geyer [28], Pflug [59,60] and Knight [41]. As Lemma 15 is only applicable to unconstrained optimization problems, constrained optimization problems need to be equivalently converted to unconstrained ones using the indicator function of feasible set. This leads to the issue of epi-convergence in distribution of the sum of two sequences of random functions; see, e.g., Pflug [60, Lemma 1].

Now we proceed with the proof of Theorem 4. Let Φ_m denote the objective function of (3) and $\mathcal{F}$ denote the feasible set. Then, the problem (3) can be concisely written as

$$\min_{X\in\mathbb{V}^{n_1\times n_2}} \big\{\Phi_m(X) + \delta_{\mathcal{F}}(X)\big\}.$$

By Assumptions 3 and 4 and Lemma 14, we have that the convex function Φ_m converges pointwise in probability to the convex function Φ, where $\Phi(X) := \frac{1}{2}\langle X - \overline{X}, \mathcal{Q}_\beta(X - \overline{X})\rangle$ for any $X \in \mathbb{V}^{n_1\times n_2}$. As a direct extension of Rockafellar [66, Theorem 10.8], Andersen and Gill [1, Theorem II.1] proved that the pointwise convergence in probability implies the convergence in probability (and thus in distribution) with respect to the topology of uniform convergence on compact subset. Then, according to Pflug [60, Lemma 1], we further obtain that $\Phi_m + \delta_{\mathcal{F}}$ epi-converges in distribution to $\Phi + \delta_{\mathcal{F}}$. Note that $\overline{X}$ is the unique minimizer of $\Phi(X) + \delta_{\mathcal{F}}(X)$ since $\Phi(X)$ is strongly convex over the feasible set $\mathcal{F}$. Thus, we complete the proof by applying Lemma 15 on epi-convergence in distribution.

Proof of Theorem 5

Theorem 4 actually implies that $\widehat{X}_m$ has a higher rank than $\overline{X}$ with probability converging to 1 if $\rho_m \to 0$, due to the straightforward result:

Lemma 16 *If* $X_m \xrightarrow{p} \overline{X}$, *then* $\lim_{m\to\infty} \Pr\big(\mathrm{rank}(X_m) \geq \mathrm{rank}(\overline{X})\big) = 1$.

Proof It follows from the Lipschitz continuity of singular values that

$$\sigma_k(X_m) \xrightarrow{p} \sigma_k(X) \quad \forall\, 1 \leq k \leq n.$$

Thus, for any $\varepsilon > 0$, we have

$$\mathbb{P}\big(\mathrm{rank}(X_m) \geq \mathrm{rank}(\overline{X})\big) \geq \mathbb{P}\big(|\sigma_r(X_m) - \sigma_r(\overline{X})| \leq \varepsilon\sigma_r(\overline{X})\big) \to 1 \quad \text{as} \quad m \to \infty.$$

Now we take a look at the local property for the rank function for the perturbation.

Lemma 17 *Let* $\overline{\Delta} \in \mathbb{V}^{n_1\times n_2}$ *satisfy* $\overline{U}_2^{\mathbb{T}}\overline{\Delta}\,\overline{V}_2 \neq 0$. *Then, for all* $\rho \neq 0$ *sufficiently small and* Δ *sufficiently close to* $\overline{\Delta}$, $\mathrm{rank}(\overline{X} + \rho\Delta) > \mathrm{rank}(\overline{X})$.

Proof Let $\sigma_i'(X;\cdot)$ denote the directional derivative function of the i-th largest singular value function $\sigma_i(\cdot)$ at X. Let $r := \mathrm{rank}(\overline{X})$. Note that $\sigma_{r+1}(\overline{X}) = 0$. Then, according to [48, Section 5.1] and [11, Proposition 6], for any $\Delta \in \mathbb{V}^{n_1\times n_2}$ and $\rho \to 0$, we have

$$\sigma_{r+1}(\overline{X}+\rho\Delta) - \sigma_{r+1}'(\overline{X};\rho\Delta) = O(\|\rho\Delta\|_F^2),$$

where $\sigma_{r+1}'(\overline{X};\rho\Delta) = \|\overline{U}_2^{\mathbb{T}}(\rho\Delta)\overline{V}_2\|$. Since $\overline{U}_2^{\mathbb{T}}\overline{\Delta}\,\overline{V}_2 \neq 0$, from the sign-preserving property of limits, for any $\rho \neq 0$ sufficiently small and Δ sufficiently close to $\overline{\Delta}$, we have

$$\frac{\sigma_{r+1}(\overline{X}+\rho\Delta)}{|\rho|} = \|\overline{U}_2^{\mathbb{T}}\Delta\overline{V}_2\| + O(|\rho|\|\Delta\|_F^2) > 0.$$

This implies that $\mathrm{rank}(\overline{X}+\rho\Delta) > \mathrm{rank}(\overline{X})$.

Define $\widehat{\Delta}_m := \rho_m^{-1}(\widehat{X}_m - \overline{X})$. To guarantee the efficiency of the nuclear semi-norm on encouraging a low-rank solution, the parameter ρ_m should not decay too fast. Then, for a slow decay on ρ_m, we can establish the following result.

Lemma 18 *If $\rho_m \to 0$ and $\sqrt{m}\rho_m \to \infty$, then $\widehat{\Delta}_m \xrightarrow{p} \widehat{\Delta}$, where $\widehat{\Delta}$ is the unique optimal solution to the following convex optimization problem*

$$\begin{array}{l}\min\limits_{\Delta\in\mathbb{V}^{n_1\times n_2}} \dfrac{1}{2}\langle\mathcal{Q}_\beta(\Delta),\Delta\rangle + \langle\overline{U}_1\overline{V}_1^{\mathbb{T}} - F(\overline{X}),\Delta\rangle + \|\overline{U}_2^{\mathbb{T}}\Delta\overline{V}_2\|_* \\ \quad \text{s.t. } \mathcal{R}_\alpha(\Delta) = 0,\ \ \mathcal{R}_{\beta^+}(\Delta) \le 0,\ \ \mathcal{R}_{\beta^-}(\Delta) \ge 0.\end{array} \tag{52}$$

Proof Take a variable transformation $\Delta := \rho_m^{-1}(X - \overline{X})$ in the optimization problem (3). Then one can easily see that $\widehat{\Delta}_m$ is the optimal solution to

$$\begin{array}{l}\min\limits_{\Delta\in\mathbb{V}^{n_1\times n_2}} \dfrac{1}{2m}\|\mathcal{R}_\Omega(\Delta)\|_2^2 - \dfrac{\nu}{m\rho_m}\langle\mathcal{R}_\Omega^*(\xi),\Delta\rangle + \dfrac{1}{\rho_m}\big(\|\overline{X}+\rho_m\Delta\|_* - \|\overline{X}\|_*\big) - \langle F(\widetilde{X}_m),\Delta\rangle \\ \quad \text{s.t.}\ \ \Delta \in \mathcal{F}_m := \rho_m^{-1}(\mathcal{K} - \overline{X}),\end{array} \tag{53}$$

where $\mathcal{K} := \big\{X \in \mathbb{S}^n \mid \mathcal{R}_\alpha(X) = \mathcal{R}_\alpha(\overline{X}),\ \|\mathcal{R}_\beta(X)\|_\infty \le b\big\}$. Let Φ_m and Φ denote the objective functions of (53) and (52), respectively. By the definition of directional derivative and [75, Theorem 1], we have

$$\lim_{\rho_m\to 0}\frac{1}{\rho_m}\big(\|\overline{X}+\rho_m\Delta\|_* - \|\overline{X}\|_*\big) = \langle\overline{U}_1\overline{V}_1^{\mathbb{T}},\Delta\rangle + \|\overline{U}_2^{\mathbb{T}}\Delta\overline{V}_2\|_*.$$

Then, under Assumptions 3 and 4, according to Lemma 14, we obtain that Φ_m converges pointwise in probability to Φ. Together with the convexity of $\mathcal{K}$, we know that $\mathcal{F}_m$ converges in the sense of Painlevé-Kuratowski to the tangent cone $\mathcal{T}_{\mathcal{K}}(\overline{X})$ (see [5,67]), taking the form

$$\mathcal{T}_{\mathcal{K}}(\overline{X}) = \big\{\Delta \in \mathbb{V}^{n_1\times n_2} \mid \mathcal{R}_\alpha(\Delta) = 0,\ \mathcal{R}_{\beta^+}(\Delta) \le 0,\ \mathcal{R}_{\beta^-}(\Delta) \ge 0\big\}. \tag{54}$$

Since epi-convergence of functions corresponds to set convergence of their epigraphs [67], we obtain that $\delta_{\mathcal{F}_m}$ epi-converges to $\delta_{\mathcal{T}_{\mathcal{K}}(\overline{X})}$. Then, by using the same argument as in the proof of Theorem 4, we obtain that $\Phi_m + \delta_{\mathcal{F}_m}$ epi-converges in distribution to $\Phi + \delta_{\mathcal{T}_{\mathcal{K}}(\overline{X})}$. In addition, the optimal solution to (52) is unique due to the strong convexity of Φ over the feasible set $\mathcal{K}$. Then, applying Lemma 15 on the epi-convergence in distribution leads to the desired result.

Note that $\widehat{X}_m = \overline{X} + \rho_m \widehat{\Delta}_m$. From Lemmas 16, 17 and 18, we can see that $\overline{U}_2^{\mathbb{T}} \widehat{\Delta} \overline{V}_2 = 0$ is a necessary condition for the rank consistency of $\widehat{X}_m$. Then, we look into an explicit characterization of this condition.

Lemma 19 *Let $\widehat{\Delta}$ be the optimal solution to the problem (52). Then $\overline{U}_2^{\mathbb{T}} \widehat{\Delta} \overline{V}_2 = 0$ if and only if the linear system (13) has a solution $\widehat{\Gamma} \in \mathbb{V}^{(n_1-r)\times(n_2-r)}$ with $\|\widehat{\Gamma}\| \leq 1$. Moreover, in this case, $\widehat{\Delta} = \mathcal{Q}_\beta^\dagger\big(\overline{U}_2 \widehat{\Gamma}\, \overline{V}_2^{\mathbb{T}} - \overline{U}_1 \overline{V}_1^{\mathbb{T}} + F(\overline{X})\big)$.*

Proof Assume that $\overline{U}_2^{\mathbb{T}} \widehat{\Delta} \overline{V}_2 = 0$. Since $\widehat{\Delta}$ is the optimal solution to (52), from the optimality condition, the subdifferential of $\|X\|_*$ at 0, and [66, Theorem 23.7], we obtain that there exist some $\widehat{\Gamma} \in \mathbb{V}^{(n_1-r)\times(n_2-r)}$ with $\|\widehat{\Gamma}\| \leq 1$ and $(\widehat{\eta}^0, \widehat{\eta}^1, \widehat{\eta}^2) \in \mathbb{R}^{|\alpha|} \times \mathbb{R}^{|\beta^+|} \times \mathbb{R}^{|\beta^-|}$ such that

$$\begin{cases} \mathcal{Q}_\beta(\widehat{\Delta}) + \overline{U}_1 \overline{V}_1^{\mathbb{T}} - F(\overline{X}) + \mathcal{R}_\alpha^*(\widehat{\eta}^0) + \mathcal{R}_{\beta^+}^*(\widehat{\eta}^1) + \mathcal{R}_{\beta^-}^*(\widehat{\eta}^2) - \overline{U}_2 \widehat{\Gamma}\, \overline{V}_2^{\mathbb{T}} = 0, \\ \mathcal{R}_\alpha(\widehat{\Delta}) = 0, \\ \mathcal{R}_{\beta^+}(\widehat{\Delta}) \leq 0, \quad \widehat{\eta}^1 \geq 0, \quad \langle \mathcal{R}_{\beta^+}(\widehat{\Delta}), \widehat{\eta}^1 \rangle = 0, \\ \mathcal{R}_{\beta^-}(\widehat{\Delta}) \geq 0, \quad \widehat{\eta}^2 \leq 0, \quad \langle \mathcal{R}_{\beta^-}(\widehat{\Delta}), \widehat{\eta}^2 \rangle = 0. \end{cases} \tag{55}$$

Note that $\mathcal{R}_{\beta^+}(\widehat{\Delta}) \leq 0$ and $\mathcal{R}_{\beta^-}(\widehat{\Delta}) \geq 0$ implies that $\mathcal{Q}_\beta^\dagger \mathcal{Q}_\beta(\widehat{\Delta}) = \mathcal{P}_\beta(\widehat{\Delta})$. Moreover, $\mathcal{Q}_\beta^\dagger \mathcal{R}_\alpha^*(\widehat{\eta}^0) = \mathcal{Q}_\beta^\dagger \mathcal{R}_{\beta^+}^*(\widehat{\eta}^1) = \mathcal{Q}_\beta^\dagger \mathcal{R}_{\beta^-}^*(\widehat{\eta}^2) = 0$. Then, we apply the operator $\mathcal{Q}_\beta^\dagger$ to the first equation of (55) and then obtain

$$\mathcal{P}_\beta(\widehat{\Delta}) + \mathcal{Q}_\beta^\dagger(\overline{U}_1 \overline{V}_1^{\mathbb{T}} - F(\overline{X})) - \overline{U}_2 \widehat{\Gamma}\, \overline{V}_2^{\mathbb{T}}) = 0. \tag{56}$$

Further note that $\mathcal{R}_\alpha(\widehat{\Delta}) = 0$ implies $\mathcal{P}_\alpha(\widehat{\Delta}) = 0$. This leads $\overline{U}_2^{\mathbb{T}} \mathcal{P}_\beta(\widehat{\Delta}) \overline{V}_2 = 0$ since $\overline{U}_2^{\mathbb{T}} \widehat{\Delta} \overline{V}_2 = 0$. Then, together with(56), we obtain that $\widehat{\Gamma}$ is a solution to (13).

Conversely, if the linear system (13) has a solution $\widehat{\Gamma}$ with $\|\widehat{\Gamma}\| \leq 1$, then it is easy to check that the KKT conditions (55) are satisfied with $\widehat{\Delta} = \mathcal{Q}_\beta^\dagger(\widehat{Z})$ and $\widehat{\eta}^0 = \mathcal{R}_\alpha(\widehat{Z})$, $\widehat{\eta}^1 = (\mathcal{R}_{\beta^+}(\widehat{Z}))_+$, $\widehat{\eta}^2 = (\mathcal{R}_{\beta^-}(\widehat{Z}))_-$, where $\widehat{Z} = \overline{U}_2 \widehat{\Gamma}\, \overline{V}_2^{\mathbb{T}} - \overline{U}_1 \overline{V}_1^{\mathbb{T}} + F(\overline{X})$. Then, $\overline{U}_2^{\mathbb{T}} \widehat{\Delta} \overline{V}_2 = 0$ directly follows from (13).

With Lemma 19, the necessary part of Theorem 5 is immediate due to the necessity of the condition $\overline{U}_2^{\mathbb{T}} \widehat{\Delta} \overline{V}_2 = 0$ for rank consistency. Now we proceed with the sufficient part.

Define β_m^+, β_m^-, β_m° similar to (12) with $\overline{X}$ replaced by $\widehat{X}_m$. From Theorem 4, we have $\widehat{X}_m \xrightarrow{p} \overline{X}$ as $m \to \infty$. The convergence implies that $\beta_m^+ \subseteq \beta^+$ and $\beta_m^- \subseteq \beta^-$ for

sufficiently large m. In this circumstance, the estimator $\widehat{X}_m$ is the optimal solution to (3) with $\mathcal{C} = \mathbb{V}^{n_1\times n_2}$ if and only if there exists a subgradient $\widehat{G}_m$ of the nuclear norm at $\widehat{X}_m$ and $(\widehat{\eta}_m^0, \widehat{\eta}_m^1, \widehat{\eta}_m^2) \in \mathbb{R}^{|\alpha|} \times \mathbb{R}^{|\beta_m^+|} \times \mathbb{R}^{|\beta_m^-|}$ such that $(\widehat{X}_m, \widehat{\eta}_m^0, \widehat{\eta}_m^1, \widehat{\eta}_m^2)$ satisfies the KKT conditions:

$$\begin{cases} \dfrac{1}{m}\mathcal{R}_\Omega^*\big(\mathcal{R}_\Omega(\widehat{X}_m)-y\big)+\rho_m\big(\widehat{G}_m-F(\widetilde{X}_m)\big)+\mathcal{R}_\alpha^*(\widehat{\eta}_m^0)+\mathcal{R}_{\beta_m^+}^*(\widehat{\eta}_m^1)+\mathcal{R}_{\beta_m^-}^*(\widehat{\eta}_m^2)=0, \\ \mathcal{R}_\alpha(\widehat{X}_m)=\mathcal{R}_\alpha(\overline{X}), \\ \mathcal{R}_{\beta_m^\circ}(\widehat{X}_m)<b,\ \mathcal{R}_{\beta_m^+}(\overline{X}_m)=b,\ \mathcal{R}_{\beta_m^-}(\overline{X}_m)=-b,\ \eta_m^1\geq 0,\ \eta_m^2\leq 0. \end{cases} \tag{57}$$

Let $(\widehat{U}_m, \widehat{V}_m) \in \mathbb{O}^{n_1,n_2}(\widehat{X}_m)$ with $\widehat{U}_{m,1} \in \mathbb{O}^{n_1\times r}$, $\widehat{U}_{m,2} \in \mathbb{O}^{n_1\times(n_1-r)}$, $\widehat{V}_{m,1} \in \mathbb{O}^{n_2\times r}$ and $\widehat{V}_{m,2} \in \mathbb{O}^{n_2\times(n_2-r)}$. From Theorem 4 and Lemma 16, we know that $\mathrm{rank}(\widehat{X}_m) \geq r$ with probability tending to one. When $\mathrm{rank}(\widehat{X}_m) \geq r$ holds, from the characterization of the subdifferential of the nuclear norm [75,76], we have that $\widehat{G}_m = \widehat{U}_{m,1}\widehat{V}_{m,1}^{\mathbb{T}} + \widehat{U}_{m,2}\widehat{\Gamma}_m\widehat{V}_{m,2}^{\mathbb{T}}$ for some $\widehat{\Gamma}_m \in \mathbb{V}^{(n_1-r)\times(n_2-r)}$ satisfying $\|\widehat{\Gamma}_m\| \leq 1$. Now we want to show $\|\widehat{\Gamma}_m\| < 1$ so that $\mathrm{rank}(\widehat{X}_m) = r$. Since $\widehat{X}_m \xrightarrow{p} \overline{X}$, by [11, Proposition 8] we have $\widehat{U}_{m,1}\widehat{V}_{m,1}^{\mathbb{T}} \xrightarrow{p} \overline{U}_1\overline{V}_1^{\mathbb{T}}$. As $\widehat{\Gamma}$ is the unique optimal solution to (13), applying Lemma 14 with the equation (2) leads to

$$\frac{1}{m\rho_m}\mathcal{R}_\Omega^*\big(\mathcal{R}_\Omega(\widehat{X}_m)-y\big)+\widehat{U}_{m,1}\widehat{V}_{m,1}^{\mathbb{T}}-F(\widetilde{X}_m)\xrightarrow{p}\mathcal{Q}_\beta(\widehat{\Delta})+\overline{U}_1\overline{V}_1^{\mathbb{T}}-F(\overline{X}),$$

Then, by further applying the operator $\mathcal{Q}_\beta^\dagger$ to the above equation, together with (56) in Lemma 19 and (57), we obtain that

$$\overline{U}_2^{\mathbb{T}}\mathcal{Q}_\beta^\dagger(\widehat{U}_{m,2}\widehat{\Gamma}_m\widehat{V}_{m,2}^{\mathbb{T}})\overline{V}_2 \xrightarrow{p} \overline{U}_2^{\mathbb{T}}\mathcal{Q}_\beta^\dagger(\overline{U}_2\widehat{\Gamma}\overline{V}_2^{\mathbb{T}})\overline{V}_2. \tag{58}$$

Since $\widehat{X}_m \xrightarrow{p} \overline{X}$, according to [11, Proposition 7], there exist two sequences of matrices $Q_{m,U} \in \mathbb{O}^{n_1-r}$ and $Q_{m,V} \in \mathbb{O}^{n_2-r}$ such that

$$\widehat{U}_{m,2}Q_{m,U} \xrightarrow{p} \overline{U}_2 \quad \text{and} \quad \widehat{V}_{m,2}Q_{m,V} \xrightarrow{p} \overline{V}_2. \tag{59}$$

Moreover, the uniqueness of the solution to the linear system (13) is equivalent to the non-singularity of its linear operator. By combining (58) and (59), we obtain that $Q_{m,U}^{\mathbb{T}}\widehat{\Gamma}_m Q_{m,V} \xrightarrow{p} \widehat{\Gamma}$. Hence, we obtain that $\|\widehat{\Gamma}_m\| < 1$ and thus $\mathrm{rank}(\widehat{X}_m) = r$ with probability tending to one since $\|\widehat{\Gamma}\| < 1$. Thus, we complete the proof of Theorem 5.

Proof of Theorem 6

The proof of Theorem 6 is similar to the proof of Theorem 5. Define $\widehat{\Delta}_m := \rho_m^{-1}(\widehat{X}_m - \overline{X})$.

Lemma 20 *If $\rho_m \to 0$ and $\sqrt{m}\rho_m \to \infty$, then $\widehat{\Delta}_m \xrightarrow{p} \widehat{\Delta}$, where $\widehat{\Delta}$ is the unique optimal solution to the following convex optimization problem*

$$\begin{array}{ll} \min\limits_{\Delta\in\mathbb{S}^n} & \frac{1}{2}\langle \mathcal{Q}_\beta(\Delta), \Delta\rangle + \langle I_n - F(\overline{X}), \Delta\rangle \\ \text{s.t.} & \mathcal{R}_\alpha(\Delta) = 0, \ \ \mathcal{R}_{\beta^+}(\Delta) \le 0, \ \ \mathcal{R}_{\beta^-}(\Delta) \ge 0, \ \ \overline{P}_2^{\mathbb{T}}\Delta\overline{P}_2 \in \mathbb{S}_+^{n-r}. \end{array} \tag{60}$$

Proof It is easy to verify that $\widehat{\Delta}_m$ is the optimal solution to

$$\begin{array}{ll} \min\limits_{\Delta\in\mathbb{S}^n} & \frac{1}{2m}\|\mathcal{R}_\Omega(\Delta)\|_2^2 - \frac{\nu}{m\rho_m}\langle \mathcal{R}_\Omega^*(\xi), \Delta\rangle + \langle I_n - F(\widetilde{X}_m), \Delta\rangle \\ \text{s.t.} & \Delta \in \mathcal{F}_m := \rho_m^{-1}(\mathcal{K}\cap\mathbb{S}_+^n - \overline{X}), \end{array} \tag{61}$$

where $\mathcal{K} := \{X \in \mathbb{S}^n \mid \mathcal{R}_\alpha(X) = \mathcal{R}_\alpha(\overline{X}), \ \|\mathcal{R}_\beta(X)\|_\infty \le b\}$. Then, $\mathcal{F}_m$ converges in the sense of Painlevé-Kuratowski to the tangent cone $\mathcal{T}_{\mathcal{K}\cap\mathbb{S}_+^n}(\overline{X})$ (see [5,67]). Note that the Slater condition in Assumption 5 implies that $\mathcal{K}$ and $\mathbb{S}_+^n$ cannot be separated. Then, from [67, Theorem 6.42], we have $\mathcal{T}_{\mathcal{K}\cap\mathbb{S}_+^n}(\overline{X}) = \mathcal{T}_{\mathcal{K}}(\overline{X}) \cap \mathcal{T}_{\mathbb{S}_+^n}(\overline{X})$ with $\mathcal{T}_{\mathcal{K}}(\overline{X})$ taking the form of (54) and $\mathcal{T}_{\mathbb{S}_+^n}(\overline{X}) = \{\Delta \in \mathbb{S}^n \mid \overline{P}_2^{\mathbb{T}}\Delta\overline{P}_2 \in \mathbb{S}_+^{n-r}\}$ according to Arnold [2]. Then, the proof can be completed by using the same argument as in the proof of Lemma 18.

For the case $\mathcal{C} = \mathbb{S}_+^n$, Lemmas 16, 17 and 20 imply that $\overline{P}_2^{\mathbb{T}}\widehat{\Delta}\overline{P}_2 = 0$ is a necessary condition for the rank consistency of $\widehat{X}_m$. Then we look into an explicit characterization of this condition.

Lemma 21 *Let $\widehat{\Delta}$ be the optimal solution to the problem (60). Then $\overline{P}_2^{\mathbb{T}}\widehat{\Delta}\overline{P}_2 = 0$ if and only if the linear system (14) has a solution $\widehat{\Lambda} \in \mathbb{S}_+^{n-r}$. Moreover, in this case, $\widehat{\Delta} = \mathcal{Q}_\beta^\dagger\big(\overline{P}_2\widehat{\Lambda}\,\overline{P}_2^{\mathbb{T}} - I_n + F(\overline{X})\big)$.*

Proof Note that the Slater condition also holds for the problem (60). (One may check the point $X^0 - \overline{X}$.) Hence, $\widehat{\Delta}$ is the optimal solution to (60) if and only if there exists $(\widehat{\zeta}^0, \widehat{\zeta}^1, \widehat{\zeta}^2, \widehat{\Lambda}) \in \mathbb{R}^{|\alpha|} \times \mathbb{R}^{|\beta^+|} \times \mathbb{R}^{|\beta^-|} \times \mathbb{S}^{n-r}$ such that

$$\left\{\begin{array}{l} \mathcal{Q}_\beta(\widehat{\Delta}) + I_n - F(\overline{X}) + \mathcal{R}_\alpha^*(\widehat{\zeta}^0) + \mathcal{R}_{\beta^+}^*(\widehat{\zeta}^1) + \mathcal{R}_{\beta^-}^*(\widehat{\zeta}^2) - \overline{P}_2\widehat{\Lambda}\overline{P}_2^{\mathbb{T}} = 0, \\ \mathcal{R}_\alpha(\widehat{\Delta}) = 0, \\ \mathcal{R}_{\beta^+}(\widehat{\Delta}) \le 0, \quad \widehat{\zeta}^1 \ge 0, \quad \langle \mathcal{R}_{\beta^+}(\widehat{\Delta}), \widehat{\zeta}^1\rangle = 0, \\ \mathcal{R}_{\beta^-}(\widehat{\Delta}) \ge 0, \quad \widehat{\zeta}^2 \le 0, \quad \langle \mathcal{R}_{\beta^-}(\widehat{\Delta}), \widehat{\zeta}^2\rangle = 0, \\ \overline{P}_2^{\mathbb{T}}\widehat{\Delta}\overline{P}_2 \in \mathbb{S}_+^{n-r}, \ \widehat{\Lambda} \in \mathbb{S}_+^{n-r}, \ \langle \overline{P}_2^{\mathbb{T}}\widehat{\Delta}\overline{P}_2, \widehat{\Lambda}\rangle = 0. \end{array}\right. \tag{62}$$

Then, applying the operator $\mathcal{Q}_\beta^\dagger$ to the first equation of (62) yields the desired expression of $\widehat{\Delta}$ if $\overline{P}_2^{\mathbb{T}}\widehat{\Delta}\overline{P}_2 = 0$. It immediately follows that $\widehat{\Lambda}$ is a solution to (14).

Conversely, if the linear system (14) has a solution $\widehat{\Lambda} \in \mathbb{S}^{n-r}_+$, it is easy to check that (62) is satisfied with $\widehat{\Delta} = \mathcal{Q}^{\dagger}_{\beta}(\widehat{Z})$ and $\widehat{\zeta}^0 = \mathcal{R}_{\alpha}(\widehat{Z})$, $\widehat{\zeta}^1 = (\mathcal{R}_{\beta^+}(\widehat{Z}))_+$, $\widehat{\zeta}^2 = (\mathcal{R}_{\beta^-}(\widehat{Z}))_-$, where $\widehat{Z} = \overline{P}_2\widehat{\Lambda}\,\overline{P}_2^{\mathbb{T}} - I_n + F(\overline{X})$. Then, $\overline{P}_2^{\mathbb{T}}\widehat{\Delta}\overline{P}_2 = 0$ directly follows from (14).

The necessary part of Theorem 6 is immediate from Lemma 21 due to the necessity of the condition $\overline{P}_2^{\mathbb{T}}\widehat{\Delta}\overline{P}_2 = 0$ for rank consistency. Now we proceed with the sufficient part.

Define β^+_m, β^-_m, β°_m by (12) with $\overline{X}$ replaced by $\widehat{X}_m$. From Theorem 4, we have $\widehat{X}_m \xrightarrow{p} \overline{X}$ as $m \to \infty$. The convergence implies that $\beta^+_m \subseteq \beta^+$ and $\beta^-_m \subseteq \beta^-$ for sufficiently large m. In this circumstance, the Slater condition implies that $\widehat{X}_m$ is the optimal solution to (3) if and only if there exists multipliers $(\widehat{\zeta}^0_m, \widehat{\zeta}^1_m, \widehat{\zeta}^2_m, \widehat{S}_m) \in \mathbb{R}^{|\alpha|} \times \mathbb{R}^{|\beta^+|} \times \mathbb{R}^{|\beta^-|} \times \mathbb{S}^n$ such that $(\widehat{X}_m, \widehat{\zeta}^0_m, \widehat{\zeta}^1_m, \widehat{\zeta}^2_m, \widehat{S}_m)$ satisfies the KKT conditions:

$$\begin{cases} \dfrac{1}{m}\mathcal{R}^*_{\Omega}\big(\mathcal{R}_{\Omega}(\widehat{X}_m) - y\big) + \rho_m\big(I_n - F(\widetilde{X}_m)\big) + \mathcal{R}^*_{\alpha}(\widehat{\zeta}^0_m) + \mathcal{R}^*_{\beta^+}(\widehat{\zeta}^1_m) + \mathcal{R}^*_{\beta^-}(\widehat{\zeta}^2_m) - \widehat{S}_m = 0, \\ \mathcal{R}_{\alpha}(\widehat{X}_m) = \mathcal{R}_{\alpha}(\overline{X}), \\ \mathcal{R}_{\beta^{\circ}_m}(\widehat{X}_m) < b,\ \mathcal{R}_{\beta^+_m}(\overline{X}_m) = b,\ \mathcal{R}_{\beta^-_m}(\overline{X}_m) = -b,\ \eta^1_m \ge 0,\ \eta^2_m \le 0, \\ \widehat{X}_m \in \mathbb{S}^n_+,\ \widehat{S}_m \in \mathbb{S}^n_+,\ \langle \widehat{X}_m, \widehat{S}_m\rangle = 0. \end{cases} \tag{63}$$

The last equation in (63) implies that $\widehat{X}_m$ and $\widehat{S}_m$ can have a simultaneous eigenvalue decomposition. Let $\widehat{P}_m \in \mathbb{O}^n(\widehat{X}_m)$ with $\widehat{P}_{m,1} \in \mathbb{O}^{n\times r}$ and $\widehat{P}_{m,2} \in \mathbb{O}^{n\times(n-r)}$. From Theorem 4 and Lemma 16, we know that $\mathrm{rank}(\widehat{X}_m) \ge r$ with probability tending to one. When $\mathrm{rank}(\widehat{X}_m) \ge r$ holds, we can write $\widehat{S}_m = \widehat{P}_{m,2}\widehat{\Lambda}_m\widehat{P}^{\mathbb{T}}_{m,2}$ for some diagonal matrix $\widehat{\Lambda}_m \in \mathbb{S}^{n-r}_+$. In addition, if $\widehat{\Lambda}_m \in \mathbb{S}^{n-r}_{++}$, then $\mathrm{rank}(\widehat{X}_m) = r$. Since $\widehat{X}_m \xrightarrow{p} \overline{X}$, according to [11, Proposition 1], there exist a sequence of matrices $Q_m \in \mathbb{O}^{n-r}$ such that $\widehat{P}_{m,2}Q_m \xrightarrow{p} \overline{P}_2$. Then, using the similar arguments to the proof of Theorem 5, we obtain that $Q_m^{\mathbb{T}}\widehat{\Lambda}_m Q_m \xrightarrow{p} \widehat{\Lambda}$. Since $\widehat{\Lambda} \in \mathbb{S}^n_{++}$, we have $\widehat{\Lambda}_m \in \mathbb{S}^n_{++}$ with probability tending to one. Thus, we complete the proof of Theorem 6.

Proof of Theorem 7

We first prove for the rectangular case $\mathcal{C} = \mathbb{V}^{n_1\times n_2}$ by contradiction. Assume that there exists some $\mathbb{V}^{(n_1-r)\times(n_2-r)} \ni \overline{\Gamma} \ne 0$ such that $\mathcal{B}_2(\overline{\Gamma}) = \overline{U}_2^{\mathbb{T}}\mathcal{Q}^{\dagger}_{\beta}(\overline{U}_2\overline{\Gamma}\,\overline{V}_2^{\mathbb{T}})\overline{V}_2 = 0$. Then $\langle \overline{\Gamma}, \overline{U}_2^{\mathbb{T}}\mathcal{Q}^{\dagger}_{\beta}(\overline{U}_2\overline{\Gamma}\,\overline{V}_2^{\mathbb{T}})\overline{V}_2\rangle = \langle \overline{U}_2\overline{\Gamma}\,\overline{V}_2^{\mathbb{T}}, \mathcal{Q}^{\dagger}_{\beta}(\overline{U}_2\overline{\Gamma}\,\overline{V}_2^{\mathbb{T}})\rangle = 0$. This immediately leads to $(\mathcal{Q}^{\dagger}_{\beta})^{1/2}(\overline{U}_2\overline{\Gamma}\,\overline{V}_2^{\mathbb{T}}) = 0$ since $\mathcal{Q}^{\dagger}_{\beta}$ is a self-adjoint and positive semidefinite operator. It then follows that $[\mathcal{R}_{\beta^{\circ}}; (\mathcal{R}_{\beta^+})_-; (\mathcal{R}_{\beta^-})_+](\overline{U}_2\overline{\Gamma}\,\overline{V}_2^{\mathbb{T}}) = 0$, where $(\mathcal{R}_{\pi})_{\pm}(\cdot) := (\mathcal{R}_{\pi}(\cdot))_{\pm}$ with $\pi = \beta^+$ or β^-. Then by using this equality, we have that for any $H \in \mathcal{T}(\overline{X})$,

$$\begin{aligned} 0 &= \langle \overline{\Gamma}, \overline{U}_2^{\mathbb{T}}H\overline{V}_2\rangle = \langle \overline{U}_2\overline{\Gamma}\,\overline{V}_2^{\mathbb{T}}, H\rangle = \langle \mathcal{R}_{\alpha\cup\beta}(\overline{U}_2\overline{\Gamma}\,\overline{V}_2^{\mathbb{T}}), \mathcal{R}_{\alpha\cup\beta}(H)\rangle \\ &= \langle [\mathcal{R}_{\alpha}; (\mathcal{R}_{\beta^+})_+; (\mathcal{R}_{\beta^-})_-](\overline{U}_2\overline{\Gamma}\,\overline{V}_2^{\mathbb{T}}), \mathcal{R}_{\alpha\cup\beta^+\cup\beta^-}(H)\rangle. \end{aligned}$$

By using the arbitrariness of $\mathcal{R}_{\alpha\cup\beta^+\cup\beta^-}(H)$ over $\mathbb{R}^{|\alpha\cup\beta^+\cup\beta^-|}$ implied by the constraint nondegeneracy (15), we further have $[\mathcal{R}_\alpha; (\mathcal{R}_{\beta^+})_+; (\mathcal{R}_{\beta^-})_-](\overline{U}_2\overline{\Gamma}\,\overline{V}_2^{\mathbb{T}}) = 0$. Therefore, we obtain $\overline{U}_2\overline{\Gamma}\,\overline{V}_2^{\mathbb{T}} = 0$ and thus $\overline{\Gamma} = 0$, which leads to a contradiction. Therefore, the linear operator $\mathcal{B}_2$ is positive definite. The proof for the positive semidefinite case is similar.

Proof of Theorem 9

We first prove for the constraint nondegeneracy.

Lemma 22 *For the matrix completion problems of Classes I and II, the constraint nondegeneracy (16) holds at* $\overline{X}$.

Proof For the real covariance matrix case, the proof is given in [61, Lemma 3.3] and [62, Proposition 2.1]. For the complex covariance matrix case, one can use the similar arguments to prove the result.

We next consider the density matrix case. Suppose that $\overline{X}$ satisfies the density constraint, i.e., $\mathcal{R}_\alpha(\overline{X}) = \frac{1}{\sqrt{n}}\mathrm{Tr}(\overline{X}) = \frac{1}{\sqrt{n}}$. Note that for any $t \in \mathbb{R}$, we have $t\overline{X} \in \mathrm{lin}(\mathcal{T}_{\mathcal{H}^n_+}(\overline{X}))$. This, along with $\mathrm{Tr}(\overline{X}) = 1$, implies that

$$\frac{1}{\sqrt{n}}\mathrm{Tr}\big(\mathrm{lin}(\mathcal{T}_{\mathcal{H}^n_+}(\overline{X}))\big) = \mathcal{R}_\alpha\big(\mathrm{lin}(\mathcal{T}_{\mathcal{H}^n_+}(\overline{X}))\big) = \mathbb{R}.$$

This means that the constraint nondegeneracy (16) holds.

From Theorem 7 and Lemma 22, for both Classes I an II, the linear system (14) has a unique solution $\widehat{\Lambda}$. Moreover, for both Classes I and II, uniform sampling yields $\mathcal{Q}^\dagger_\beta(Z) = \mathcal{P}_\beta(Z)/d_2$ for any $Z \in \mathbb{S}^n_+$. Thus, from (14), we have

$$\widehat{\Lambda} - \overline{P}_2^{\mathbb{T}}\mathcal{P}_\alpha(\overline{P}_2\widehat{\Lambda}\overline{P}_2^{\mathbb{T}})\overline{P}_2 = \overline{P}_2^{\mathbb{T}}\mathcal{P}_\beta(\overline{P}_2\widehat{\Lambda}\overline{P}_2^{\mathbb{T}})\overline{P}_2 = \overline{P}_2^{\mathbb{T}}\mathcal{P}_\beta(I_n - F(\overline{X}))\overline{P}_2. \quad (64)$$

Then we first prove for Class I by contradiction. For any $Z \in \mathbb{S}^n_+$, $\mathcal{P}_\alpha(Z)$ is the diagonal matrix whose i-th diagonal entries is X_{ii} for all $i \in \pi$ and the other entries are 0. Assume that $\widehat{\Lambda} \notin \mathbb{S}^{n-r}_{++}$, i.e., $\lambda_{\min}(\widehat{\Lambda}) \le 0$, where $\lambda_{\min}(\cdot)$ denotes the smallest eigenvalue. Then, we have

$$\lambda_{\min}(\widehat{\Lambda}) = \lambda_{\min}(\overline{P}_2\widehat{\Lambda}\overline{P}_2^{\mathbb{T}}) \le \lambda_{\min}\big(\mathcal{P}_\alpha(\overline{P}_2\widehat{\Lambda}\overline{P}_2^{\mathbb{T}})\big) \le \lambda_{\min}\big(\overline{P}_2^{\mathbb{T}}\mathcal{P}_\alpha(\overline{P}_2\widehat{\Lambda}\overline{P}_2^{\mathbb{T}})\overline{P}_2\big),$$

where the equality follows from the fact that $\widehat{\Lambda}$ and $\overline{P}_2\widehat{\Lambda}\overline{P}_2^{\mathbb{T}}$ have the same nonzero eigenvalues, the first inequality follows from the fact that the vector of eigenvalues is majorized by the vector of diagonal entries, (e.g., see [50, Theorem 9.B.1]), and the second inequality follows from the Courant-Fischer minmax theorem, (e.g., see [50, Theorem 20.A.1]). As a result, the left-hand side of (64) is not positive definite. Notice

that $\overline{P}_2^{\mathbb{T}} F(\overline{X})\overline{P}_2 = 0$. Thus, the right-hand side of (64) can be written as

$$\begin{aligned}\overline{P}_2^{\mathbb{T}}\mathcal{P}_\beta(I_n - F(\overline{X}))\overline{P}_2 &= \overline{P}_2^{\mathbb{T}}\mathcal{P}_\beta(I_n)\overline{P}_2 + \overline{P}_2^{\mathbb{T}}\mathcal{P}_\alpha(F(\overline{X}))\overline{P}_2 \\ &= \overline{P}_2^{\mathbb{T}}\big(\mathcal{P}_\beta(I_n) + \mathcal{P}_\alpha(F(\overline{X}))\big)\overline{P}_2.\end{aligned}$$

Since $\mathrm{rank}(\overline{X}) = r$, with the choice (22) of F, we have that for any $i \in \pi$,

$$\overline{X}_{ii} = \sum_{j=1}^{r} \lambda_j(\overline{X})|\overline{P}_{ij}|^2 > 0 \quad \text{implies} \quad \big(F(\overline{X})\big)_{ii} = \sum_{j=1}^{r} f_i\big(\lambda_j(\overline{X})\big)|\overline{P}_{ij}|^2 > 0.$$

Moreover, $\mathcal{P}_\beta(I_n)$ is the diagonal matrix with the last $n - r$ diagonal entries being ones and the other entries being zeros. Thus, $\mathcal{P}_\beta(I_n) + \mathcal{P}_\alpha(F(\overline{X}))$ is a diagonal matrix with all positive diagonal entries. It follows that the right-hand side of (64) is positive definite. Thus, we obtain a contradiction. Therefore, we should have $\widehat{\Lambda} \in \mathbb{S}_{++}^{n-r}$. Then, we can obtain the rank consistency according to Theorem 6.

Next, we prove for Class II. It is easy to see $\mathcal{P}_\alpha(\cdot) = \frac{1}{n}\mathrm{Tr}(\cdot)I_n$. By further using $\overline{P}_2^{\mathbb{T}} F(\overline{X})\overline{P}_2 = 0$ and $\mathcal{P}_\beta(I_n) = 0$, we can rewrite (64) as

$$\widehat{\Lambda} - \frac{1}{n}\mathrm{Tr}(\widehat{\Lambda})I_{n-r} = \frac{1}{n}\mathrm{Tr}(F(\overline{X}))I_{n-r}.$$

By taking the trace on both sides, we obtain that $\widehat{\Lambda} = \frac{1}{r}\mathrm{Tr}(F(\overline{X}))I_{n-r}$. Since $\overline{X}$ is a density matrix of rank r, with the choice (22) of F, we have that

$$\mathrm{Tr}(\overline{X}) = \sum_{i=1}^{n}\sum_{j=1}^{r} \lambda_j(\overline{X})|\overline{P}_{ij}|^2 = 1 \quad \text{implies} \quad \mathrm{Tr}\big(F(\overline{X})\big) = \sum_{i=1}^{n}\sum_{j=1}^{r} f_i\big(\lambda_j(\overline{X})\big)|\overline{P}_{ij}|^2 > 0.$$

It follows that $\widehat{\Lambda} \in \mathbb{S}_{++}^{n-r}$ and thus we obtain the rank consistency.

References

1. Andersen, P.K., Gill, R.D.: Cox's regression model for counting processes: a large sample study. Ann. Stat. **10**(4), 1100–1120 (1982)
2. Arnold, V.I.: On matrices depending on parameters. Russ. Math. Surv. **26**(2), 29–43 (1971)
3. Bach, F.R.: Consistency of trace norm minimization. J. Mach. Learn. Res. **9**, 1019–1048 (2008)
4. Bhatia, R.: Matrix Analysis, vol. 169. Springer, Berlin (1997)
5. Bonnans, J.F., Shapiro, A.: Perturbation analysis of optimization problems. Springer, Berlin (2000)
6. Bühlmann, P., Van De Geer, S.: Statistics for High-Dimensional Data: Methods, Theory and Applications. Springer, Berlin (2011)
7. Candès, E.J., Plan, Y.: Matrix completion with noise. Proc. IEEE **98**(6), 925–936 (2010)
8. Candès, E.J., Recht, B.: Exact matrix completion via convex optimization. Found. Comut. Math. **9**(6), 717–772 (2009)
9. Candès, E.J., Tao, T.: The power of convex relaxation: near-optimal matrix completion. IEEE Trans. Inf. Theory **56**(5), 2053–2080 (2010)
10. Ding, C.: An Introduction to a Class of Matrix Optimization Problems. PhD thesis, National University of Singapore (2012)

11. Ding, C., Sun, D.F., Toh, K.C.: An introduction to a class of matrix cone programming. Math. Program. **144**(1), 141–179 (2014)
12. Efron, B., Hastie, T., Johnstone, I., Tibshirani, R.: Least angle regression. Ann. Stat. **32**(2), 407–499 (2004)
13. Fan, J.: Comments on "Wavelets in statistics: a review" by A. Antoniadis. Stat. Methods Appl. **6**(2), 131–138 (1997)
14. Fan, J., Li, R.: Variable selection via nonconcave penalized likelihood and its oracle properties. J. Am. Stat. Assoc. **96**(456), 1348–1360 (2001)
15. Fan, J., Lv, J.: Sure independence screening for ultrahigh dimensional feature space. J. R. Stat. Soc. Ser B **70**(5), 849–911 (2008)
16. Fan, J., Lv, J.: A selective overview of variable selection in high dimensional feature space. Stat. Sin. **20**(1), 101 (2010)
17. Fan, J., Lv, J., Qi, L.: Sparse high dimensional models in economics. Annu. Rev. Econ. **3**, 291 (2011)
18. Fan, J., Peng, H.: Nonconcave penalized likelihood with a diverging number of parameters. Ann. Stat. **32**(3), 928–961 (2004)
19. Fazel, M.: Matrix Rank Minimization with Applications. PhD thesis, Stanford University (2002)
20. Fazel, M., Hindi, H., Boyd, S.P.: Log-det heuristic for matrix rank minimization with applications to hankel and euclidean distance matrices. In: Proceedings of the American Control Conference, 2003, vol. 3, pp. 2156–2162. Ieee (2003)
21. Fazel, M., Pong, T.K., Sun, D.F., Tseng, P.: Hankel matrix rank minimization with applications in system identification and realization. SIAM J. Matrix Anal. Appl. **34**(3), 946–977 (2013)
22. Flammia, S.T., Gross, D., Liu, Y.K., Eisert, J.: Quantum tomography via compressed sensing: error bounds, sample complexity and efficient estimators. New J. Phys. **14**(9), 095022 (2012)
23. Fornasier, Massimo, Rauhut, Holger, Ward, Rachel: Low-rank matrix recovery via iteratively reweighted least squares minimization. SIAM J. Optim. **21**(4), 1614–1640 (2011)
24. Foygel, R., Salakhutdinov, R., Shamir, O., Srebro, N.: Learning with the weighted trace-norm under arbitrary sampling distributions. In: Advances in Neural Information Processing Systems (NIPS) 24, vol. 24, pp. 2133–2141 (2011)
25. Foygel, R., Srebro, N.: Concentration-based guarantees for low-rank matrix reconstruction. In: 24nd Annual Conference on Learning Theory (COLT), (2011)
26. Gao, Y.: Structured Low Rank Matrix Optimization Problems: A Penalized Approach. PhD thesis, National University of Singapore (2010)
27. Gao, Y., Sun, D.F.: A majorized penalty approach for calibrating rank constrained correlation matrix problems. Preprint available at http://www.math.nus.edu.sg/~matsundf/MajorPen_May5.pdf (2010)
28. Geyer, C.J.: On the asymptotics of constrained M-estimation. The Ann. Stat. **22**(4), 1993–2010 (1994)
29. Geyer, C.J.: On the asymptotics of convex stochastic optimization. Unpublished manuscript (1996)
30. Gross, D.: Recovering low-rank matrices from few coefficients in any basis. IEEE Trans. Inf. Theory **57**(3), 1548–1566 (2011)
31. Gross, D., Liu, Y.K., Flammia, S.T., Becker, S., Eisert, J.: Quantum state tomography via compressed sensing. Phys. Rev. Lett. **105**(15), 150401 (2010)
32. Han, C., Phillips, P.C.B.: GMM with many moment conditions. Econometrica **74**(1), 147–192 (2006)
33. Huang, J., Ma, S., Zhang, C.H.: Adaptive lasso for sparse high-dimensional regression models. Stat. Sin. **18**(4), 1603 (2010)
34. Jiang, K., Sun, D., Toh, K.C.: An inexact accelerated proximal gradient method for large scale linearly constrained convex SDP. SIAM J. Optim. **22**(3), 1042–1064 (2012)
35. Jiang, K., Sun, D.F., Toh, K.C.: A partial proximal point algorithm for nuclear norm regularized matrix least squares problems. Math. Program. Comput. **6**(3), 281–325 (2014)
36. Jiang, K., Sun, D.F., Toh, K.C.: Solving nuclear normregularized and semidefinitematrix least squares problems with linear equality constraints. In: Bezdek, K., Deza, A., Ye, Y. (eds.) Discrete Geometry and Optimization, pp. 133–162. Springer, New York (2013)
37. Keshavan, R.H., Montanari, A., Oh, S.: Matrix completion from a few entries. IEEE Trans. Inf. Theory **56**(6), 2980–2998 (2010)
38. King, A.J., Wets, R.J.B.: Epi-consistency of convex stochastic programs. Stoch. Int. J. Probab. Stoch. Process. **34**(1–2), 83–92 (1991)
39. Klopp, O.: Rank penalized estimators for high-dimensional matrices. Electron. J. Stat. **5**, 1161–1183 (2011)

40. Klopp, O.: Noisy low-rank matrix completion with general sampling distribution. Bernoulli **20**(1), 282–303 (2014)
41. Knight, K.: Epi-convergence in distribution and stochastic equi-semicontinuity. Unpublished manuscript (1999)
42. Koltchinskii, V.: Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems: Ecole Dâeté de Probabilités de Saint-Flour XXXVIII-2008, vol. 2033. Springer, Berlin (2011)
43. Koltchinskii, V.: Von Neumann entropy penalization and low-rank matrix estimation. Ann. Stat. **39**(6), 2936–2973 (2012)
44. Koltchinskii, V., Lounici, K., Tsybakov, A.B.: Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. Ann. Stat. **39**(5), 2302–2329 (2011)
45. Lai, Ming-Jun, Yangyang, Xu, Yin, Wotao: Improved iteratively reweighted least squares for unconstrained smoothed \ell_q minimization. SIAM J. Numer Anal. **51**(2), 927–957 (2013)
46. Ledoux, M., Talagrand, M.: Probability in Banach Spaces: Isoperimetry and Processes, vol. 23. Springer, Berlin (1991)
47. Leng, C., Lin, Y., Wahba, G.: A note on the lasso and related procedures in model selection. Stat. Sin. **16**(4), 1273 (2006)
48. Lewis, A.S., Sendov, H.S.: Nonsmooth analysis of singular values. Part II: applications. Set-Valued Anal. **13**(3), 243–264 (2005)
49. Liu, Y.K.: Universal low-rank matrix recovery from Pauli measurements. In: Advances in Neural Information Processing Systems, pp 1638–1646 (2011)
50. Marshall, A.W., Olkin, I., Arnold, B.: Inequalities: Theory of Majorization and Its Applications. Springer, Berlin (2010)
51. Massart, P.: Optimal constants for hoeffding type inequalities. Technical report, Mathematiques Université de Paris-Sud, Report 98.86 (1998)
52. Meinshausen, N.: Relaxed lasso. Comput. Stat. Data Anal. **52**(1), 374–393 (2007)
53. Meinshausen, N., Bühlmann, P.: High-dimensional graphs and variable selection with the lasso. Ann. Stat. **34**(3), 1436–1462 (2006)
54. Mesbahi, M.: On the rank minimization problem and its control applications. Syst. Control Lett. **33**(1), 31–36 (1998)
55. Mesbahi, M., Papavassilopoulos, G.P.: On the rank minimization problem over a positive semidefinite linear matrix inequality. IEEE Trans. Autom. Control **42**(2), 239–243 (1997)
56. Mohan, K., Fazel, M.: Reweighted nuclear norm minimization with application to system identification. In: IEEE American Control Conference (ACC), 2010, pp. 2953–2959 (2010)
57. Mohan, Karthik, Fazel, Maryam: Iterative reweighted algorithms for matrix rank minimization. J. Mach. Learn. Res. **13**(1), 3441–3473 (2012)
58. Negahban, S., Wainwright, M.J.: Restricted strong convexity and weighted matrix completion: optimal bounds with noise. J. Mach. Learn. Res. **13**, 1665–1697 (2012)
59. Pflug, G.C.: Asymptotic dominance for solutions of stochastic programs. Czechoslov. J. Oper. Res. **1**(1), 21–30 (1992)
60. Pflug, G.C.: Asymptotic stochastic programs. Math. Oper. Res. **20**(4), 769–789 (1995)
61. Qi, H., Sun, D.F.: A quadratically convergent newton method for computing the nearest correlation matrix. SIAM J. Matrix Anal. Appl. **28**(2), 360 (2006)
62. Qi, H., Sun, D.F.: An augmented Lagrangian dual approach for the H-weighted nearest correlation matrix problem. IMA J. Numer. Anal. **31**(2), 491–511 (2011)
63. Recht, B.: A simpler approach to matrix completion. J. Mach. Learn. Res. **12**, 3413–3430 (2011)
64. Recht, B., Fazel, M., Parrilo, P.A.: Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. SIAM Rev. **52**(3), 471–501 (2010)
65. Robinson, S.M.: Local structure of feasible sets in nonlinear programming, Part II: nondegeneracy. Math. Program. Oberwolfach II 217–230 (1984)
66. Rockafellar, R.T.: Convex Analysis. Princeton University Press, Princeton (1970)
67. Rockafellar, R.T., Wets, R.J.B.: Variational Analysis. Springer, Berlin (1998)
68. Rohde, A., Tsybakov, A.B.: Estimation of high-dimensional low-rank matrices. Ann. Stat. **39**(2), 887–930 (2011)
69. Salakhutdinov, R., Srebro, N.: Collaborative filtering in a non-uniform world: learning with the weighted trace norm. Adv. Neural Inf. Process. Syst. **23**, 2056–2064 (2010)
70. Srebro, N., Rennie, J.D.M., Jaakkola, T.: Maximum-margin matrix factorization. Adv. Neural Inf. Process. Syst. **17**(5), 1329–1336 (2005)

71. Tibshirani, R.: Regression shrinkage and selection via the lasso. J. R. Stat. Soc. Ser. B **58**(1), 267–288 (1996)
72. Tropp, J.A.: User-friendly tail bounds for sums of random matrices. Found. Comput. Math. **12**(4), 389–432 (2012)
73. Van Der Vaart, A.W., Wellner, J.A.: Weak Convergence and Empirical Processes. Springer, Berlin (1996)
74. Wang, Y.: Asymptotic equivalence of quantum state tomography and noisy matrix completion. Ann. Stat. **41**(5), 2462–2504 (2013)
75. Watson, G.A.: Characterization of the subdifferential of some matrix norms. Linear Algebra Appl. **170**, 33–45 (1992)
76. Watson, G.A.: On matrix approximation problems with Ky Fan k norms. Numer. Algorithms **5**(5), 263–272 (1993)
77. Zhang, C.H.: Nearly unbiased variable selection under minimax concave penalty. Ann. Stat. **38**(2), 894–942 (2010)
78. Zhao, P., Yu, B.: On model selection consistency of lasso. J. Mach. Learn. Res. **7**(2), 2541 (2007)
79. Zhou, S., Van De Geer, S., Bühlmann, P.: Adaptive Lasso for high dimensional regression and Gaussian graphical modeling. Arxiv preprint. arXiv:0903.2515 (2009)
80. Zou, H.: The adaptive lasso and its oracle properties. J. Am. Stat. Assoc. **101**(476), 1418–1429 (2006)
81. Zou, H., Li, R.: One-step sparse estimates in nonconcave penalized likelihood models. Ann. Stat. **36**(4), 1509 (2008)