Journal of Machine Learning Research 20 (2019) 1-43 Submitted 9/17; Revised 3/19; Published 6/19

Low Permutation-rank Matrices: Structural Properties and Noisy Completion

Nihar B. Shah NIHARS@CS.CMU.EDU
Machine Learning Department and Computer Science Department
Carnegie Mellon University

Sivaraman Balakrishnan SIVA@STAT.CMU.EDU
Department of Statistics
Carnegie Mellon University

Martin J. Wainwright WAINWRIG@STAT.BERKELEY.EDU
EECS Department and Department of Statistics
University of California, Berkeley

Editor: Tong Zhang

Abstract

We consider the problem of noisy matrix completion, in which the goal is to reconstruct a structured matrix whose entries are partially observed in noise. Standard approaches to this underdetermined inverse problem are based on assuming that the underlying matrix has low rank, or is well-approximated by a low rank matrix. In this paper, we propose a richer model based on what we term the "permutation-rank" of a matrix. We first describe how the classical non-negative rank model enforces restrictions that may be undesirable in practice, and how and these restrictions can be avoided by using the richer permutation-rank model. Second, we establish the minimax rates of estimation under the new permutation-based model, and prove that surprisingly, the minimax rates are equivalent up to logarithmic factors to those for estimation under the typical low rank model. Third, we analyze a computationally efficient singular-value-thresholding algorithm, known to be optimal for the low-rank setting, and show that it also simultaneously yields a consistent estimator for the low-permutation rank setting. Finally, we present various structural results characterizing the uniqueness of the permutation-rank decomposition, and characterizing convex approximations of the permutation-rank polytope.

Keywords: Non-negative matrix completion, recommender systems, permutation-based model, minimax theory, oracle inequalities.

1. Introduction

In the problem of matrix completion, the goal is to reconstruct a matrix based on observations of a subset of its entries (Laurent, 2001). Matrix completion has a variety of applications, including recommender systems (Koren et al., 2009), image understanding (Lee and Seung, 1999), credit risk monitoring (Vandendorpe et al., 2008), fluorescence spectroscopy (Gobinet et al., 2004), and modeling signal-adaptive audio effects (Sarver and Klapuri, 2011). We refer the reader to the surveys by Gillis (2014); Davenport and Romberg (2016) for an overview of the vast literature on this topic. Throughout this paper, in order to provide a running example for our modeling, it will be convenient to refer back to a particular variant of a recommender system application. More concretely, suppose that there are $n \geq 2$ users and $d \geq 2$ items, as well as an unknown matrix $M^* \in [0,1]^{n \times d}$ that captures the users' preferences for the items. Specifically, the $(i,j)^{th}$ entry of M^* represents the probability that user i likes item j. The problem is to estimate this preference matrix $M^* \in [0,1]^{n \times d}$ from observing users' likes or dislikes for some subset of the items.

Following a long line of past work in this area (e.g., Chen et al., 2013; Gross, 2011; Srebro et al., 2005; Candes and Recht, 2009; Candès and Tao, 2010; Keshavan et al., 2010; Recht, 2011; Chatterjee, 2014), we consider the following form of random design observation model. For a given parameter $p_{\text{obs}} \in (0,1]$ and for any user-item pair (i,j), we observe user i's rating for item j with probability p_{obs}. We assume that when an entry (i,j) is observed, we observe a binary value—for instance, {like, dislike} or $\{0,1\}$—which arises as a Bernoulli realization of the true preference M^*_{ij}.[1] More formally, we observe a matrix $Y \in \{0, \frac{1}{2}, 1\}^{n \times d}$, where

$$Y_{ij} = \begin{cases} 1 & \text{with probability } p_{\text{obs}} M^*_{ij} & \text{(user } i \text{ likes item } j) \\ 0 & \text{with probability } p_{\text{obs}}(1 - M^*_{ij}) & \text{(user } i \text{ dislikes item } j) \\ \frac{1}{2} & \text{with probability } 1 - p_{\text{obs}} & \text{(no data available),} \end{cases} \tag{1}$$

for every $(i,j) \in [n] \times [d]$. The goal is to estimate the underlying matrix M^* based on the observed matrix Y.

It is clear that, if no structural conditions are imposed on the underlying matrix M^*, then this problem is ill-posed. A classical approach is to impose on a bound on either the rank or the non-negative rank of the matrix. We begin by describing the approach based on the non-negative rank, before turning to the alternative approach based on permutation rank that is the focus of this paper.

Non-negative rank: In the problem of non-negative low-rank matrix completion, the matrix M^* is assumed to have a factorization of the form

$$M^* = UV^T,$$

for some matrices $U \in \mathbb{R}_+^{n \times r}$ and $V \in \mathbb{R}_+^{d \times r}$. Here the integer $r \in \{1, \ldots, \min\{d,n\}\}$ is known as the *non-negative rank* of the matrix. (As a corner case, we also have that the zero matrix is the only matrix with a non-negative rank of $r = 0$.) It is often assumed that the non-negative rank r is a known quantity, but in this paper, we make no such assumptions. For any value of $r \in \{1, \ldots, \min\{d,n\}\}$, we let $\mathbb{C}_{\text{NR}}(r)$ denote the set of all matrices with a non-negative factorization of rank at most r—that is

$$\mathbb{C}_{\text{NR}}(r) := \left\{ M \in [0,1]^{n \times d} \mid M = UV^T,\ U \in \mathbb{R}_+^{n \times r},\ V \in \mathbb{R}_+^{d \times r} \right\}.$$

For any matrix M, the smallest value of r such that $M \in \mathbb{C}_{\text{NR}}(r)$ is termed its *non-negative rank*, and is denoted by $\overline{r}(M)$.

1. Our results readily extend to any rating scheme with bounded values, such as five-star ratings. We focus on the binary case for purposes of brevity.

In order to gain some intuition for the meaning of the non-negative rank, note that any matrix $M \in \mathbb{C}_{\text{NR}}(r)$ can be written as a sum of the form

$$M = \sum_{\ell=1}^{r} u^{\ell} (v^{\ell})^T.$$

Here $u^{\ell} \in \mathbb{R}_+^n$ and $v^{\ell} \in \mathbb{R}_+^d$ are vectors such that $u^{\ell}(v^{\ell})^T \in [0,1]^{n \times d}$ for every $\ell \in [r]$. Such a decomposition can be interpreted as the existence of r features, indexed by $\ell \in [r]$. The d entries of vector v^{ℓ} represent the contribution of feature ℓ to the d respective items, and the n entries of vector u^{ℓ} represent the amounts by which the n respective users are influenced by feature ℓ. The popular overview article by Koren et al. (2009) provides an explanation for this assumption:

> *"Latent factor models are an alternative approach that tries to explain the ratings by characterizing both items and users on, say, 20 to 100 factors inferred from the ratings patterns. ... For movies, the discovered factors might measure obvious dimensions such as comedy versus drama, amount of action, or orientation to children; less well-defined dimensions such as depth of character development or quirkiness; or completely uninterpretable dimensions. For users, each factor measures how much the user likes movies that score high on the corresponding movie factor."*

Concerns with non-negative rank: Let us now delve deeper into this model[2], continuing in the context of movie recommendations for the sake of concreteness. Suppose there are r features that govern the movie watching experience; examples of such features include the amount of comedy content or the depth of character development. For any user $i \in [n]$ and any feature $\ell \in [r]$, we let $u_i^{\ell} \in \mathbb{R}_+$ denote the "affinity" of user i towards feature ℓ, and for any movie $j \in [d]$, we let $v_j^{\ell} \in \mathbb{R}_+$ denote the amount of content associated to feature ℓ in movie j. The conventional low non-negative rank model then assumes that the affinity of user i towards movie j conditioned on feature ℓ is given by $u_i^{\ell} v_j^{\ell}$. Consequently, for given feature ℓ, the entire behavior of each user and movie is governed by a pair of parameters, namely u_i^{ℓ} and v_j^{ℓ} for user i and item j respectively. Such an assumption has some unnatural implications. For instance, consider any two movies, say A and B, and any two users, say X and Y. Then conditioned on any feature ℓ, we have the implication

$$\frac{\text{Preference of user } X \text{ for movie } A}{\text{Preference of user } X \text{ for movie } B} = \frac{\text{Preference of user } Y \text{ for movie } A}{\text{Preference of user } Y \text{ for movie } B}.$$

In words, the low non-negative rank model inherently imposes a condition that is potentially unrealistic—namely, that for any given feature, the ratio of preferences for any pair of movies is identical for all users. Likewise, for any given feature, the ratio of preferences of any pair of users is identical for all movies. With the goal of circumventing this possibly troublesome condition, let us now describe a generalization that we call permutation rank, which is the main focus of our paper.

2. A slightly different, alternative interpretation is discussed in Appendix A.

Permutation rank: As with the ordinary rank, the permutation rank of the all-zeros matrix is zero. Otherwise, for any non-zero matrix, the permutation rank ρ takes values in the set $\{1, \ldots, \min\{n, d\}\}$. We begin by describing the set $\mathbb{C}_{\text{PR}}(1)$ of matrices with permutation rank one:

$$\mathbb{C}_{\text{PR}}(1) := \{M \in [0,1]^{n\times d} \mid \exists \text{ permutations } \pi_1 : [n] \to [n] \text{ and } \pi_2 : [d] \to [d] \text{ such that}$$
$$M_{ij} \geq M_{i'j'} \text{ for every quadruple } (i,j,i',j') \text{ such that } \pi_1(i) \geq \pi_1(i') \text{ and } \pi_2(j) \geq \pi_2(j')\ \}.$$

In words, a non-zero matrix is said to have a permutation rank of 1 if there exists a permutation of its rows and columns such that the entries of the resulting matrix are non-decreasing down any column and to the right along any row. Observe that any matrix with the conventional (non-negative) rank equal to 1 also belongs to the set $\mathbb{C}_{\text{PR}}(1)$. However, a matrix in $\mathbb{C}_{\text{PR}}(1)$ can have any non-negative rank, meaning the set of matrices with a permutation-rank of 1 also includes some matrices with a full non-negative rank.

We now extend the definition of the permutation rank to any integer $\rho \in \{1, \ldots, \min\{n, d\}\}$. In particular, the set of matrices with permutation rank at most ρ is given by

$$\mathbb{C}_{\text{PR}}(\rho) := \Big\{M \in [0,1]^{n\times d} \;\Big|\; M = \textstyle\sum_{\ell=1}^{\rho} Q^{\ell} \text{ for some matrices } Q^1, \ldots, Q^{\rho} \in \mathbb{C}_{\text{PR}}(1)\Big\},$$

Note that this definition reduces to $\mathbb{C}_{\text{PR}}(1)$ in the special case $\rho = 1$. Otherwise, for $\rho > 1$, the permutations defining membership of each constituent matrix Q^{ℓ} in $\mathbb{C}_{\text{PR}}(\rho)$ are allowed to be different. For any matrix M, the smallest value of ρ such that $M \in \mathbb{C}_{\text{PR}}(\rho)$ is termed its *permutation rank*, and is denoted by $\overline{\rho}(M)$.

Revisiting the example of movie recommendations, the interpretation of this more general permutation-rank model is that conditioned on any feature $\ell \in [r]$, the preference ordering across movies continues to be consistent for different users, but the values of these preferences *need not* be identical scalings of each other. Observe that the conventional non-negative matrix-completion setting $\mathbb{C}_{\text{NR}}(r)$ is a special case of the permutation-rank matrix-completion setting where each matrix Q^{ℓ} is restricted to be of rank one. Whenever $r < \min\{d, n\}$, we have the strict inclusion $\mathbb{C}_{\text{NR}}(r) \subset \mathbb{C}_{\text{PR}}(r)$.

Outline and main contributions: Having discussed the limitations of the non-negative rank model and introduced the permutation rank, we now outline the remainder of the paper. In Section 2, we present our main results on the problem of estimating the matrix M^* (in the Frobenius norm) from partial and noisy observations. Specifically, we present a certain regularized least squares estimator, and prove that it achieves (nearly) minimax-optimal rates for estimation over the permutation-rank model. We also show that surprisingly, even if one considers the more restrictive non-negative rank model, and even if the rank is known, no estimator can achieve lower estimation error up to logarithmic factors. We also analyze the computationally efficient Singular Value Thresholding (SVT) algorithm, and show that it yields consistent estimates over the permutation-rank model, in addition to yielding the optimal estimate under the non-negative rank model. In Section 3, we establish some interesting properties of the permutation-rank model, and also derive certain relationships of this model with the non-negative rank model. In Section 4 we present the proofs of our results. We conclude the paper with a discussion in Section 5.

The paper also contains two appendices. Appendix A describes an alternative interpretation of the non-negative rank model. Appendix B is devoted to negative results, where we show that certain intuitive algorithms provably fail.

2. Main results on estimating M^*

We begin by considering the problem of estimating a low permutation rank matrix M^* based on noisy and partial observations. We first analyze a computationally expensive estimator, based on regularizing the least-squares cost with a multiple of the permutation rank, and show that it achieves minimax-optimal rates up to logarithmic factors. We then turn to a polynomial-time algorithm based on nuclear norm regularization, which is equivalent singular value thresholding in the current set-up.

2.1. Optimal oracle inequalities for estimation

Suppose that we collect an observation matrix Y of the form (1), where the unknown matrix M^* belongs $\mathbb{C}_{\text{PR}}$. In this section, we analyze a regularized form of least-squares estimation, as applied to the recentered matrix

$$Y' := \frac{1}{p_{\text{obs}}} Y - \frac{1 - p_{\text{obs}}}{2 p_{\text{obs}}} 11^T. \tag{2a}$$

We perform this recentering in order to obtain an unbiased estimate Y' of the true matrix M^* in the presence of missing observations, which is used in the least-squares estimator described below. As a sanity check, observe that when $p_{\text{obs}} = 1$, we have the direct relation $Y' = Y$.

Letting $\overline{\rho}(M)$ denote the permutation-rank of any matrix M, we then consider the estimator

$$\widehat{M}_{\text{LS}} \in \underset{M \in [0,1]^{n \times d}}{\arg\min} \Big(|\!|\!| Y' - M |\!|\!|_{\text{F}}^2 + \frac{\overline{\rho}(M) \max\{n, d\} \log^{2.01} d}{p_{\text{obs}}} \Big). \tag{2b}$$

Observe that importantly, the estimator $\widehat{M}_{\text{LS}}$ does *not* need to know the value of the true permutation-rank of the underlying matrix. Moreover, while the estimator (as stated) is based on a known value of p_{obs}, this assumption is not critical, since p_{obs} can be estimated accurately from the observed matrix Y.

We now turn to some theoretical guarantees on the performance of this estimator. Rather than assuming that, for a given rank ρ, the target matrix M^* has permutation rank exactly equal to ρ, we instead provide bounds that depend on distances to the set of all matrices with a given permutation rank. More precisely, for any given tolerance $\epsilon \geq 0$, define the set

$$B^{\text{P}}(\rho, \epsilon) := \big\{ M \in [0,1]^{n \times d} \mid \exists M' \in [0,1]^{n \times d} \text{ s.t. } \overline{\rho}(M') \leq \rho \text{ and } |\!|\!| M - M' |\!|\!|_{\text{F}} \leq \epsilon \big\},$$

corresponding to the set of all matrices that are at most ϵ distant from the set of matrices with permutation rank ρ. Similarly, we define the set

$$B^{\text{N}}(r, \epsilon) := \big\{ M \in [0,1]^{n \times d} \mid \exists M' \in [0,1]^{n \times d} \text{ s.t. } \overline{r}(M') \leq r \text{ and } |\!|\!| M - M' |\!|\!|_{\text{F}} \leq \epsilon \big\},$$

corresponding to matrices that are at most ϵ away from some matrix with non-negative-rank r.

In stating the following theorem, as well as throughout the remainder of the paper, we use c, c', c_1 etc. to denote positive universal constants. The values of these constants may differ from line to line.

Theorem 1 *(a) For any matrix $M^* \in [0,1]^{n\times d}$ and any integer $\rho \in [\min\{n,d\}]$, the regularized least squares estimator $\widehat{M}_{LS}$ satisfies the upper bound*

$$\frac{1}{nd}\|\widehat{M}_{LS} - M^*\|_F^2 \leq c_1 \min\Big\{1, \quad \min_{M\in\mathbb{C}_{PR}(\rho)} \frac{\|M - M^*\|_F^2}{nd} + \frac{\rho \ \log^{2.01}(nd)}{\min\{n,d\}p_{\mathrm{obs}}}\Big\}, \tag{3a}$$

with probability at least $1 - e^{-c_0 \max\{n,d\}\log(nd)}$.
(b) Conversely, for any integer $\rho \in [\min\{n,d\}]$, any scalar $\epsilon \geq 0$, and any estimator $\widehat{M}$, there exists a matrix $M^* \in B^{\mathrm{N}}(\rho,\epsilon)$ such that

$$\mathbb{E}\big[\frac{1}{nd}\|\widehat{M} - M^*\|_F^2\big] \geq c_2 \min\Big\{1, \quad \frac{\epsilon^2}{nd} + \frac{\rho}{\min\{n,d\}p_{\mathrm{obs}}}\Big\}. \tag{3b}$$

See Section 4.1 for the proof of this claim.

Interpretation as oracle inequality: The upper bound (3a) is an instance of an *oracle inequality*: it provides a family of upper bounds, one for each choice of the integer $\rho \in [\min\{n,d\}]$, on the estimation error associated with an arbitrary matrix $M^* \in [0,1]^{n\times d}$. For each choice of ρ, the upper bound (3a) consists of two terms. The first term, involving the minimum over $M \in \mathbb{C}_{\mathrm{PR}}(\rho)$, is a form of approximation error: it measures how well the unknown matrix M^* can be approximated with a matrix M of permutation rank at most ρ. The second term is a form of estimation error, measuring the difficulty of estimating a matrix that has permutation rank at most ρ. Since one such an upper bound holds for each choice of ρ, the bound (3a) shows that the estimator mimicks the behavior of an "oracle", which is allowed to choose ρ so as to optimize the trade-off between the approximation and estimation error.

Sandwiching of the risk: The upper bound (3a) of Theorem 1 can equivalently be stated in the following manner. For any integer $\rho \in [\max\{n,d\}]$ and any scalar $\epsilon \geq 0$ such that $M^* \in B^{\mathrm{P}}(\rho,\epsilon)$, the regularized least squares estimator $\widehat{M}_{\mathrm{LS}}$ satisfies the upper bound

$$\frac{1}{nd}\|\widehat{M}_{\mathrm{LS}} - M^*\|_{\mathrm{F}}^2 \leq c_1 \min\Big\{\frac{\epsilon^2}{nd} + \frac{\rho \ \log^{2.01}(nd)}{\min\{n,d\}p_{\mathrm{obs}}}, 1\Big\}, \tag{4a}$$

with probability at least $1 - e^{-c_0 \max\{n,d\}\log(\max\{nd\})}$. On the other hand, since $B^{\mathrm{N}}(\rho,\epsilon) \subseteq B^{\mathrm{P}}(\rho,\epsilon)$ for every value of ρ and ϵ, the lower bound (3b) of Theorem 1 implies the following result. For any integer $\rho \in [\min\{n,d\}]$, any scalar $\epsilon \geq 0$, and any estimator $\widehat{M}$, there exists a matrix $M^* \in B^{\mathrm{N}}(\rho,\epsilon)$ such that

$$\mathbb{E}\big[\frac{1}{nd}\|\widehat{M} - M^*\|_{\mathrm{F}}^2\big] \geq c_2 \min\Big\{1, \quad \frac{\epsilon^2}{nd} + \frac{\rho}{\min\{n,d\}p_{\mathrm{obs}}}\Big\}. \tag{4b}$$

Comparing the bounds (4a) and (4b), we see that our results are sharp up to logarithmic factors.

Specialization to minimax risk: When suitably specialized to matrices that have some fixed permutation (or non-negative) rank, Theorem 1 leads to sharp upper and lower bounds on the minimax risks for the problems of matrix completion over the sets $\mathbb{C}_{\mathrm{NR}}$ and $\mathbb{C}_{\mathrm{PR}}$. In order for a clear comparison between the two bounds, let us index both the non-negative rank and the permutation-rank using a generic notation k—the meaning of the notation will be clear from the context.

Part (a) of Theorem 1 implies that for any value $k \in [\min\{n, d\}]$ and any matrix $M^* \in \mathbb{C}_{\mathrm{PR}}(k)$, the regularized least squares estimator $\widehat{M}_{\mathrm{LS}}$ satisfies the bound

$$\frac{1}{dn}|\!|\!|\widehat{M}_{\mathrm{LS}} - M^*|\!|\!|_{\mathrm{F}}^2 \leq c_1 \min\Big\{\frac{k \ \log^{2.01}(nd)}{\min\{n, d\} p_{\mathrm{obs}}}, 1\Big\},$$

with probability at least $1 - e^{-c_0 \max\{n,d\} \log(nd)}$. Within the set of matrices $[0, 1]^{n\times d}$ under consideration, we have the deterministic upper bound $\frac{1}{nd}|\!|\!|\widehat{M}_{\mathrm{LS}} - M^*|\!|\!|_{\mathrm{F}}^2 \leq 1$. Consequently, our high probability upper bound also implies a uniform bound on the mean-squared error over the set $\mathbb{C}_{\mathrm{PR}}(k)$—that is

$$\sup_{M^* \in \mathbb{C}_{\mathrm{PR}}(k)} \frac{1}{dn}\mathbb{E}[|\!|\!|\widehat{M}_{\mathrm{LS}} - M^*|\!|\!|_{\mathrm{F}}^2] \leq c_1' \min\Big\{\frac{k \ \log^{2.01}(nd)}{\min\{n, d\} p_{\mathrm{obs}}}, 1\Big\}. \tag{5a}$$

Since $\mathbb{C}_{\mathrm{PR}}(k)$ is a superset of $\mathbb{C}_{\mathrm{NR}}(k)$, the same upper bound holds for the minimax risk over $\mathbb{C}_{\mathrm{NR}}(k)$:

$$\sup_{M^* \in \mathbb{C}_{\mathrm{NR}}(k)} \frac{1}{dn}\mathbb{E}[|\!|\!|\widehat{M}_{\mathrm{LS}} - M^*|\!|\!|_{\mathrm{F}}^2] \leq c_1' \min\Big\{\frac{\rho \ \log^{2.01}(nd)}{\min\{n, d\} p_{\mathrm{obs}}}, 1\Big\}. \tag{5b}$$

Conversely, part (b) of Theorem 1 implies that for any $k \in [\max\{n, d\}]$, the error incurred by any estimator $\widehat{M}$ over the set $\mathbb{C}_{\mathrm{NR}}(k)$ is error lower bounded as

$$\sup_{M^* \in \mathbb{C}_{\mathrm{NR}}(k)} \frac{1}{dn}\mathbb{E}[|\!|\!|\widehat{M} - M^*|\!|\!|_{\mathrm{F}}^2] \geq c_2 \min\Big\{\frac{k}{\min\{n, d\} p_{\mathrm{obs}}}, 1\Big\}. \tag{5c}$$

Since $\mathbb{C}_{\mathrm{NR}}(k) \subseteq \mathbb{C}_{\mathrm{PR}}(k)$, the error incurred by any estimator $\widehat{M}$ over the set $\mathbb{C}_{\mathrm{PR}}(k)$ is also lower bounded as

$$\sup_{M^* \in \mathbb{C}_{\mathrm{PR}}(k)} \frac{1}{dn}\mathbb{E}[|\!|\!|\widehat{M} - M^*|\!|\!|_{\mathrm{F}}^2] \geq c_2 \min\Big\{\frac{k}{\min\{n, d\} p_{\mathrm{obs}}}, 1\Big\}. \tag{5d}$$

We have thus characterized the minimax risk over both the families $\mathbb{C}_{\mathrm{PR}}$ or $\mathbb{C}_{\mathrm{NR}}$, with bounds (5) that are matching up to logarithmic factors.

A win-win for permutation-based models: An important consequence of our oracle and minimax results is the multi-fold benefit of moving from the restrictive non-negative-rank assumptions to the more general permutation-rank assumptions. Fitting a permutation-rank k model when the true matrix actually has a non-negative rank of k leads to relatively little additional (overfitting) error. On the other hand, we show later in the paper that fitting a non-negative rank k model when the true matrix actually has a permutation-rank of k can lead to very high error, due to model mismatch.

Resolving an open problem in estimation from pairwise comparisons: The aforementioned results resolve an important open problem in the area of estimation from pairwise comparisons. The pairwise-comparison setting involves a number of items (such as a number of cars or football teams), and the observed data comprises noisy comparisons between various pairs of these items. The classical literature in this area assumes that the probability of any item beating another in a pairwise comparison between them follows a restrictive parameter-based model. Recently, the paper Shah et al. (2017) established minimax rates of estimating these pairwise comparison probabilities under a strictly and significantly more general model called strong stochastic transitivity (SST). Understanding minimax rates when the pairwise comparison probabilities follow mixtures of SST models remained open.

The pairwise-comparison setting under the SST model is a special case of our present problem and corresponds to the setting where the value of ρ is known and equal to 1, the matrix M^* is square with $n = d$, and all entries of M^* satisfy the shifted-skew-symmetry condition $M^*_{ij} + M^*_{ji} = 1$. Theorem 1 provides guarantees on the estimation of mixtures of different SST models, thereby resolving an important open problem on this topic. In addition, the results of Theorem 1 are obtained in the framework of oracle inequalities, which is more general than the minimax framework of Shah et al. (2017), and sheds light on situations when the true matrix does not follow the assumed model.

2.2. Computationally efficient estimator

At this point, we do not know how to compute the regularized least squares estimator (2b) in an efficient manner, and we suspect that it may be computationally intractable to do so. Consequently, in this section, we turn to analyzing a different method based on singular value thresholding (SVT). Singular value thresholding has been used either directly or as a subroutine in several past papers on the conventional low-rank matrix completion problem (see, for example, the papers Cai et al., 2010; Donoho et al., 2014; Chatterjee, 2014). This approach is appealing due to its computational simplicity, involving only computation of the singular value decomposition, followed by a single pointwise non-linearity; see Cai and Osher (2010) for a fast algorithm. In the context of the permutation rank completion problem, we show here that the SVT estimator is consistent for estimation under the permutation-rank model, albeit with a rate that is suboptimal by a factor of $\sqrt{\min\{n,d\}p_{\text{obs}}}$. Note that these guarantees hold without the estimator needing to know that the matrix is drawn from a permutation-rank model, nor the value p_{obs} of the permutation rank.

The SVT estimator is straightforward to describe. From the observation matrix $Y \in \{0, \frac{1}{2}, 1\}^{n \times d}$, we first obtain the transformed observation matrix Y' as in equation (2a). Applying the singular value decomposition yields the representation $Y' = UDV^T$, where the $(n \times d)$ matrix D is diagonal, whereas the $(n \times n)$ and $(d \times d)$ matrices U and V respectively are both orthonormal. For a threshold $\lambda > 0$ to be specified, define another diagonal matrix $\widetilde{D}_\lambda$ with entries

$$[\widetilde{D}_\lambda]_{jj} = \begin{cases} 0 & \text{if } D_{jj} < \lambda \\ D_{jj} - \lambda & \text{if } D_{jj} \geq \lambda \end{cases} \qquad \text{for each } j \in [\min\{n, d\}]. \tag{6}$$

Finally, the SVT estimator is given by

$$\widehat{M}_{\text{SVT}} = U\widetilde{D}_{\lambda}V^{T}.$$

The following theorem now establishes guarantees for the singular value thresholding estimator.

Theorem 2 *Suppose that $p_{\text{obs}} \geq \frac{1}{\min\{n,d\}} \log^7(nd)$. Then for any matrix $M^* \in [0,1]^{n \times d}$, the SVT estimator $\widehat{M}_{\scriptscriptstyle SVT}$ with threshold $\lambda = 2.1\sqrt{\frac{n+d}{p_{\text{obs}}}}$ satisfies the bound*

$$\frac{1}{nd}|\!|\!|\widehat{M}_{\scriptscriptstyle SVT} - M^*|\!|\!|_{\scriptscriptstyle F}^2 \leq c_1 \min_{M \in [0,1]^{n \times d}} \Big(\min\Big\{ \frac{\overline{\rho}(M)}{\sqrt{\min\{n,d\}p_{\text{obs}}}}, \frac{\overline{r}(M)}{\min\{n,d\}p_{\text{obs}}} \Big\} + \frac{1}{nd}|\!|\!|M^* - M|\!|\!|_{\scriptscriptstyle F}^2 \Big), \tag{7}$$

with probability at least $1 - e^{-c_0 \max\{n,d\}}$.

Observe that the bound (7) on the risk of the SVT estimator has the term $\sqrt{\min\{n,d\}}$ in the denominator of the first expression in the minimum, as opposed to the $\min\{n,d\}$ in the upper bound (3a) from Theorem 1. This form of "$\sqrt{n}$-suboptimality" arises in several permutation-based problems of this type studied in recent papers (e.g., Shah et al., 2017, 2019; Chatterjee and Mukherjee, 2019; Shah et al., 2016; Flammarion et al., 2016; Pananjady et al., 2016). In some cases, this gap—between the performance of any polynomial-time algorithm and the best algorithm—is known to be unavoidable (Shah et al., 2019) conditional on the planted clique conjecture. It is interesting to speculate whether such a computational complexity gap exists in the context of the permutation-rank model.

The proof techniques underlying Theorem 2 can also be used to establish previously known guarantees (Koltchinskii et al., 2011; Chatterjee, 2014) for the non-negative rank model. In order to contrast with the permutation-rank model, let us state one such guarantee here. It is known from previous results (Koltchinskii et al., 2011; Chatterjee, 2014) for the non-negative rank model that for any matrix $M^* \in [0,1]^{n \times d}$, the SVT estimator incurs an error upper bounded by

$$\frac{1}{nd}|\!|\!|\widehat{M}_{\text{SVT}} - M^*|\!|\!|_{\text{F}}^2 \leq c_1' \frac{\overline{r}(M^*)}{\min\{n,d\}p_{\text{obs}}}, \tag{8}$$

with high probability. On the other hand, our permutation-based modeling approach yields a stronger guarantee for the classical SVT estimator—namely, setting $M = M^*$ in our upper bound (8) of Theorem 2 guarantees that the SVT estimator $\widehat{M}_{\text{SVT}}$ with threshold $\lambda = 2.1\sqrt{\frac{n+d}{p_{\text{obs}}}}$ satisfies the upper bound

$$\frac{1}{nd}|\!|\!|\widehat{M}_{\text{SVT}} - M^*|\!|\!|_{\text{F}}^2 \leq c_1 \min\left\{ \frac{\overline{r}(M^*)}{\min\{n,d\}p_{\text{obs}}}, \frac{\overline{\rho}(M^*)}{\sqrt{\min\{n,d\}p_{\text{obs}}}} \right\}, \tag{9}$$

with high probability. The bound (9) can yield results that are much sharper as compared to what may be obtained from the previously known guarantees (8) for the SVT estimator.

For example, suppose $n = d$ and $p_{\text{obs}} = 1$. Consider the matrix $M^* \in [0,1]^{n \times d}$ given by

$$M^*_{ij} = \begin{cases} 1 & \text{if } i > j \\ 0 & \text{if } i < j \\ \frac{1}{2} & \text{if } i = j. \end{cases}$$

Then we have $\overline{r}(M^*) = n$ and $\overline{\rho}(M^*) = 1$. Consequently, the bound (8) from past literature yields an upper bound of

$$\frac{1}{nd} \| \widehat{M}_{\text{SVT}} - M^* \|_{\text{F}}^2 \leq c_1',$$

whereas in contrast, our analysis (9) yields the sharper bound

$$\frac{1}{nd} \| \widehat{M}_{\text{SVT}} - M^* \|_{\text{F}}^2 \leq \frac{c_1}{\sqrt{n}}, \tag{10}$$

with high probability. Moreover, in our earlier work (Shah et al., 2017), we have shown that for this choice of M^*, the bound (10) is the best possible up to logarithmic factors for the SVT estimator with any fixed threshold λ.

3. Properties of permutation-rank models

In the previous section, we established some motivating properties of permutation-based models from the perspective of statistical estimation, in this section, we derive some more insights on the permutation-rank model.

3.1. Comparing permutation-rank and non-negative-rank

We begin by comparing the permutation-rank model with the conventional non-negative rank model. To this end, first observe that the definitions of the two models immediately imply that the permutation-rank of any matrix is always upper bounded by its non-negative rank, that is, for any matrix M, we have $\overline{\rho}(M) \leq \overline{r}(M)$. A natural question that now arises is whether there is any additional general condition beyond this simple relation that constrains the two notions of the matrix rank. The following proposition shows that there is no other guaranteed relation between the two notions of matrix rank.

Proposition 3 *For any values $0 < \rho \leq r \leq \min\{n, d\}$, there exist matrices whose permutation-rank is ρ and non-negative rank is r.*

A particular instance that underlies part of the proof of Proposition 3, associated to any pair of values (ρ, r), is the following block matrix $M_{\rho,r}$ of size $(n \times d)$:

$$M_{\rho,r} := \begin{bmatrix} J_{r-\rho+1} & 0 & 0 \\ 0 & I_{\rho-1} & 0 \\ 0 & 0 & 0 \end{bmatrix},$$

where for any value k, J_k denotes an upper triangular matrix of size $(k \times k)$ with all entries on and above the diagonal set as 1, and let I_k denote the identity matrix of size $(k \times k)$.

By construction, the matrix $M_{\rho,r}$ has a non-negative rank $\overline{r}(M_{\rho,r}) = r$ and a permutation rank $\overline{\rho}(M_{\rho,r}) = \rho$.

We now investigate a second relation between the two models. Recall from our discussion earlier that the assumptions of the permutation-rank model are much less restrictive than the assumptions of the non-negative rank model. With this context, a natural question that arises is to quantify the bias of an estimator that fits a matrix of non-negative rank of k when the true underlying matrix instead has a permutation rank of k. We answer this question using the notion of the Hausdorff distance: For any two sets $\mathbb{S}_1, \mathbb{S}_2 \in \mathbb{R}^{n \times d}$, the Hausdorff distance $\mathcal{H}(\mathbb{S}_1, \mathbb{S}_2)$ between the two sets in the squared Frobenius norm is defined as

$$\mathcal{H}(\mathbb{S}_1, \mathbb{S}_2) := \max\Big\{ \sup_{M \in \mathbb{S}_1} \inf_{M' \in \mathbb{S}_2} |\!|\!| M - M' |\!|\!|_{\mathrm{F}}^2 \;,\; \sup_{M' \in \mathbb{S}_2} \inf_{M \in \mathbb{S}_1} |\!|\!| M - M' |\!|\!|_{\mathrm{F}}^2 \Big\} \tag{11}$$

The following proposition quantifies the Hausdorff distance between non-negative-rank and permutation-rank models.

Proposition 4 *For any positive integer $k \leq \frac{1}{2}\min\{d, n\}$, the Hausdorff distance between the sets $\mathbb{C}_{NR}(k)$ and $\mathbb{C}_{PR}(k)$ is lower bounded as*

$$\mathcal{H}(\mathbb{C}_{NR}(k), \mathbb{C}_{PR}(k)) \geq c_3 \frac{nd}{k}. \tag{12}$$

Proposition 4 helps quantify the bias on fitting a non-negative rank model when the true matrix follows the permutation-rank model as follows. Consider any positive integer $k \leq \frac{1}{2}\min\{d, n\}$, and any estimator $\widetilde{M}_k$ that outputs a matrix in $\mathbb{C}_{\mathrm{NR}}(k)$. Then since $\mathbb{C}_{\mathrm{NR}}(k) \subseteq \rho(k)$, the error incurred by this estimator when the true matrix lies in the set $\mathbb{C}_{\mathrm{PR}}(k)$ is lower bounded as

$$\sup_{M^* \in \mathbb{C}_{\mathrm{PR}}(k)} \frac{1}{nd} |\!|\!| M - \widetilde{M}_k |\!|\!|_{\mathrm{F}}^2 \geq \frac{1}{nd} \mathcal{H}(\mathbb{C}_{\mathrm{NR}}(k), \mathbb{C}_{\mathrm{PR}}(k)) \geq c_3 \frac{1}{k}, \tag{13}$$

with probability 1.

Observe that when k is a constant (but n and d are allowed to grow), the right hand side of the bound (12) becomes a constant, and this is the largest possible order-wise gap between any pair of matrices in $[0,1]^{n \times d}$. Likewise, the right hand side of (13) becomes a constant, and this is the largest possible order-wise error for any estimator that outputs matrices in $[0,1]^{n \times d}$.

3.2. No "good" convex approximation

In this section, we investigate a question about an important property of the permutation-based set, and in particular, its primitive $\mathbb{C}_{\mathrm{PR}}(1)$. There are various estimators including our regularized least squares estimator (2b) as well as those studied in the literature (Shah et al., 2017, 2019, 2016) which require solving a an optimization problem over the set $\mathbb{C}_{\mathrm{PR}}(1)$. With this goal in mind, a natural question that arises is: Is the set $\mathbb{C}_{\mathrm{PR}}(1)$ convex? If not, then does it at least have a "good" convex approximation? The following proposition answers these questions in the negative using the notion of the Hausdorff distance between sets (11).

Proposition 5 *The Hausdorff distance* (11) *between the set of matrices with permutation-rank one and any arbitrary convex set* $\mathbb{C} \subseteq \mathbb{R}^{n \times n}$ *is lower bounded as*

$$\frac{1}{nd}\mathcal{H}(\mathbb{C}_{PR}(1), \mathbb{C}) \geq c,$$

where $c > 0$ *is a universal constant.*

A specific example of a convex set $\mathbb{C}$ is the convex hull of $\mathbb{C}_{\mathrm{PR}}(1)$. Then by definition we have the relation $\sup\limits_{M_1 \in \mathbb{C}_{\mathrm{PR}}(1)} \inf\limits_{M_2 \in \mathbb{C}} |||M_1 - M_2|||_{\mathrm{F}}^2 = 0$. Consequently, Proposition 5 implies that

$$\sup_{M_2 \in \mathbb{C}} \inf_{M_1 \in \mathbb{C}_{\mathrm{PR}}(1)} |||M_1 - M_2|||_{\mathrm{F}}^2 = \Theta(nd),$$

thus showing that the convex hull of $\mathbb{C}_{\mathrm{PR}}(1)$ is a much larger set than $\mathbb{C}_{\mathrm{PR}}(1)$ itself.

The proof of Proposition 5 relies on a more general result that we derive, one which relates a certain notion of inherent (lack of) convexity of a set to the Hausdorff distance between that set and any convex approximation. Note that this result does not preclude the possibility that an optimization procedure over a convex approximation to $\mathbb{C}_{\mathrm{PR}}(1)$ converges close enough to some element of $\mathbb{C}_{\mathrm{PR}}(1)$ itself. We leave the investigation of this possibility to future work.

3.3. On the uniqueness of decomposition

In this section, we investigate conditions for the uniqueness of the decomposition of any matrix into its constituent components that have a permutation-rank of one. In the conventional setting of low non-negative rank matrix completion, several past works (Donoho and Stodden, 2003; Theis et al., 2005; Laurberg et al., 2008; Gillis, 2012; Arora et al., 2012) investigate the conditions required for uniqueness of the decomposition of matrices into their constituent non-negative rank-one matrices. Here we consider an analogous question in the setting of permutation rank. More precisely, consider any matrix $M \in [0, 1]^{n \times d}$ with a permutation-rank decomposition of the form

$$M = \sum_{\ell=1}^{\overline{\rho}(M)} M^{(\ell)}, \tag{14}$$

where $M^{(\ell)} \in \mathbb{C}_{\mathrm{PR}}(1)$ for every $\ell \in [\overline{\rho}(M)]$. Under what conditions on the matrix M is the set $\{M^{(1)}, \ldots, M^{(\overline{\rho}(M))}\}$ of constituent matrices unique? The following result provides a necessary condition for uniqueness. In order to state the result, we use the notation $\mathbf{1}$ to denote the indicator function, that is, $\mathbf{1}\{x\} = 1$ if x is true and $\mathbf{1}\{x\} = 0$ if x is false.

Proposition 6 *A necessary condition for the uniqueness of a permutation-rank decomposition* (14) *for any matrix* M *is that for every coordinate* $(i, j) \in [n] \times [d]$*, there is at most one* $\ell \in [\overline{\rho}(M)]$ *such that* $M_{ij}^{(\ell)}$ *is non-zero and distinct from all other entries of* $M^{(\ell)}$*, that is,*

$$\sum_{\ell \in [\overline{\rho}(M)]} \mathbf{1}\Big\{M_{ij}^{\ell} \notin \{0\} \cup \{M_{i'j'}^{\ell}\}_{\substack{i' \in [n], j' \in [d], \\ (i',j') \neq (i,j)}}\Big\} \leq 1 \qquad \textit{for every } (i, j) \in [n] \times [d].$$

We note that the necessary condition continues to hold even if we restrict attention to only symmetric matrices. The necessary condition provided by Proposition 6 indicates that *any* sufficient condition for uniqueness of the decomposition must be quite strong. Moreover, we believe that the conditions for sufficiency may be significantly stronger than those necessitated by Proposition 6. The reason for such drastic requirements for uniqueness is the high-degree of flexibility offered by the permutation-rank model.

Let us illustrate the necessary condition from Proposition 6 with a simple example. Consider the following matrix M with $n = d = 2$ and $\overline{\rho}(M) = 2$ and decomposition into $M^{(1)}, M^{(2)} \in \mathbb{C}_{\text{PR}}(1)$:

$$M := \begin{bmatrix} 1 & .6 \\ .6 & 1 \end{bmatrix} = \begin{bmatrix} 0 & .3 \\ .3 & .9 \end{bmatrix} + \begin{bmatrix} 1 & .3 \\ .3 & .1 \end{bmatrix}$$

Observe that the necessary condition obtained in Proposition 6 is required to hold for every coordinate of the matrix. Let us first evaluate this condition for the coordinate $(1, 1)$ of the matrices. Since $M^{(1)}_{11} = 0$, there is at most one $\ell \in \{1, 2\}$ such that $M^{(\ell)}_{11}$ is non-zero, and hence the coordinate $(1, 1)$ satisfies the necessary condition. Moving on to coordinate $(1, 2)$, we have $M^{(1)}_{12} = M^{(1)}_{21}$ and $M^{(2)}_{12} = M^{(2)}_{21}$; hence the coordinate $(1, 2)$ also passes the necessary condition. The argument for coordinate $(1, 2)$ also applies to coordinate $(2, 1)$ since the matrices involved are symmetric. We finally test coordinate $(2, 2)$. Observe that $M^{(1)}_{22} \notin \{0, M^{(1)}_{11}, M^{(1)}_{12}, M^{(1)}_{21}\}$ and $M^{(2)}_{22} \notin \{0, M^{(2)}_{11}, M^{(2)}_{12}, M^{(2)}_{21}\}$. As a consequence, for both $\ell = 1$ and $\ell = 2$, we have that $M^{(\ell)}_{22}$ is non-zero and distinct from all other entries of $M^{(\ell)}$. The condition necessary for uniqueness is thus violated. Indeed, as guaranteed by Proposition 6, there exist other decompositions of M with permutation-rank 2—for instance, the decomposition

$$M = \begin{bmatrix} 1 & .6 \\ .6 & 1 \end{bmatrix} = \begin{bmatrix} 0 & .4 \\ .4 & .9 \end{bmatrix} + \begin{bmatrix} 1 & .2 \\ .2 & .1 \end{bmatrix}$$

is another example.

Finally, we put the negative result on the decomposition into some practical perspective with an analogy to tensor decompositions. The canonical polyadic (CP) decomposition of a tensor (Hitchcock, 1927) is not unique unless strong non-degeneracy conditions are imposed (Kruskal, 1977). From a theoretical perspective in many applications (for instance, in estimating latent variable models (Hsu and Kakade, 2013)) the CP decomposition is most useful or interpretable when it is unique. Furthermore, even when the decomposition is unique, computing it is NP-hard in the worst-case (Håstad, 1990). However, in practice the CP decomposition is often computed via ad-hoc methods that generate useful results (Kolda and Bader, 2009).

4. Proofs

We now turn to the proofs of our main results. In all our proofs, we assume that the values of n and d are larger than certain universal constants, so as to avoid subcases having to do with small values of (n, d). We will also ignore floors and ceilings wherever they are not critical. These assumptions entail no ultimate loss of generality, since our results continue to hold for all values with different constant prefactors. Throughout these and other proofs, we use the notation $\{c, c', c_0, c_1, C, C'\}$ and so on to denote positive constants whose values may change from line to line.

4.1. Proof of Theorem 1(a)

The proof of this theorem involves generalizing an argument used in our past work (Shah et al., 2017, Theorem 1). In particular, the problem setting of our past work (Shah et al., 2017, Theorem 1) is a special case of the present problem, restricted to the case of square matrices ($n = d$) and permutation rank $\rho = 1$. Here we develop a number of additional techniques in order to handle the generalization to non-square matrices and arbitrary permutation rank.

Throughout the proof, we assume without loss of generality that $n \leq d$; otherwise, we can apply the same argument to the matrix transposes. Now observe that we must have $\frac{1}{nd}|||\widehat{M}_{\text{LS}} - M^*|||_{\text{F}}^2 \leq 1$ deterministically, and hence it suffices to show that

$$\frac{1}{nd}|||\widehat{M}_{\text{LS}} - M^*|||_{\text{F}}^2 \leq c_1 \frac{1}{nd} \min_{M \in [0,1]^{n \times d}} \Big(|||M - M^*|||_{\text{F}}^2 + \frac{\overline{\rho}(M) d \log^{2.01}(nd)}{p_{\text{obs}}} \Big), \tag{15}$$

with high probability.

In order to succinctly capture the right hand side of (15), we define a matrix M_0 as

$$M_0 \in \operatorname*{arg\,min}_{M \in [0,1]^{n \times d}} \Big(|||M - M^*|||_{\text{F}}^2 + \frac{\overline{\rho}(M) d \log^{2.01} d}{p_{\text{obs}}} \Big).$$

One can alternatively interpret the matrix M_0 as the estimate that $\widehat{M}_{\text{LS}}$ would produce if the observation matrix Y' was replaced by the true matrix M^* in the optimization problem specified in (2b).

A key challenge in this analysis is to handle the complex structure of the permutation rank set. Specifically, any matrix in the set $\mathbb{C}_{\text{PR}}(k)$, for some $k \in [n]$, is defined to be the sum of k matrices each in the set $\mathbb{C}_{\text{PR}}(1)$. Each of these k constituent matrices can have *different and arbitrary* permutations of their rows/columns. In order to handle this complexity, we decompose this set into "simpler" sets of matrices, where each set of matrices is associated to a fixed set of permutations. We introduce some additional notation to this end. For any pair of permutations $\pi : [n] \to [n]$ and $\sigma : [d] \to [d]$, we first define the set

$$\mathbb{C}_{\text{PR}}(1; \pi, \sigma) := \{M \in \mathbb{C}_{\text{PR}}(1) \mid \text{rows and columns of } M \text{ are ordered according to } \pi \text{ and } \sigma \text{ respectively}\}.$$

In words, $\mathbb{C}_{\text{PR}}(1; \pi, \sigma)$ is the set of all matrices with permutation rank 1 and where the rows are ordered according to the permutation π and the columns are ordered according to the permutation σ.

Next we extend the notation $\mathbb{C}_{\text{PR}}(1; \pi, \sigma)$ to matrices with a more general permutation rank. Let Π denote the set of all possible permutations of d items, and let Σ denote the set of all possible permutations of the n users. Consider any value $k \in [n]$, any sequence

$\widetilde{\Pi}^{(k)} := (\pi_1, \ldots, \pi_k) \in \Pi^k$ and any sequence $\widetilde{\Sigma}^{(k)} := (\sigma_1, \ldots, \sigma_k) \in \Sigma^k$. Define the set

$$\mathbb{C}_{\text{PR}}(k; \widetilde{\Pi}^{(k)}, \widetilde{\Sigma}^{(k)}) := \Big\{ M = \sum_{\ell=1}^{k} M^{(\ell)} \Big| \ M^{(\ell)} \in \mathbb{C}_{\text{PR}}(1; \pi_\ell, \sigma_\ell) \text{ for every } \ell \in [k] \Big\}.$$

The set $\mathbb{C}_{\text{PR}}(k; \widetilde{\Pi}^{(k)}, \widetilde{\Sigma}^{(k)})$ is the set of all matrices with permutation rank k, and which can be expressed as the constituent permutation-rank-1 matrices following permutations $(\pi_1, \sigma_1), \ldots, (\pi_k, \sigma_k)$ respectively.

Having decomposed the permutation-rank set, we now apply a similar treatment to the estimator $\widehat{M}_{\text{LS}}$. The estimator $\widehat{M}_{\text{LS}}$ as defined in (2b) minimizes a certain objective over all matrices in $[0,1]^{n \times d}$. We will decompose this estimator in a certain manner in terms of multiple estimators, each of which optimize over a particular subset of $[0,1]^{n \times d}$. Specifically, for each set $\mathbb{C}_{\text{PR}}(k; \widetilde{\Pi}^{(k)}, \widetilde{\Sigma}^{(k)})$ defined above, let us define an associated estimator

$$M_{\widetilde{\Pi}^{(k)}, \widetilde{\Sigma}^{(k)}} \in \underset{M \in \mathbb{C}_{\text{PR}}(k; \widetilde{\Pi}^{(k)}, \widetilde{\Sigma}^{(k)})}{\arg\min} \|\!|Y' - M|\!\|_{\text{F}}^2. \tag{16}$$

Note that in defining the objective of minimization (16) for $M_{\widetilde{\Pi}^{(k)}, \widetilde{\Sigma}^{(k)}}$, we have omitted the regularization term $\frac{\overline{\rho}(M) d \log^{2.01} d}{p_{\text{obs}}}$ from (2b) because the value of the rank $\overline{\rho}(M)$ is the same for all matrices in the set $\mathbb{C}_{\text{PR}}(k; \widetilde{\Pi}^{(k)}, \widetilde{\Sigma}^{(k)})$. Observe that the regularized least squares estimate $\widehat{M}_{\text{LS}}$ can now be thought of as choosing the matrix $M_{\widetilde{\Pi}^{(k)}, \widetilde{\Sigma}^{(k)}}$ across all possible values of $(k; \widetilde{\Pi}^{(k)}, \widetilde{\Sigma}^{(k)})$ that minimizes the regularized least squares objective in (2b).

Finally, we define a set $\widetilde{\Gamma}$ as

$$\begin{aligned} \widetilde{\Gamma} := \Big\{ (k, \widetilde{\Pi}^{(k)}, \widetilde{\Sigma}^{(k)}) \in [n] \times \Pi^k \times \Sigma^k \Big| \|\!|Y' - M_{\widetilde{\Pi}^{(k)}, \widetilde{\Sigma}^{(k)}}|\!\|_{\text{F}}^2 + \frac{k d \log^{2.01} d}{p_{\text{obs}}} \\ \leq \|\!|Y' - M_0|\!\|_{\text{F}}^2 + \frac{\overline{\rho}(M_0) d \log^{2.01} d}{p_{\text{obs}}} \Big\}. \end{aligned}$$

The set $\widetilde{\Gamma}$ is the set of all tuples of k and sets of permutations $(\widetilde{\Pi}^{(k)}, \widetilde{\Sigma}^{(k)})$ such that the value of the objective in (2b) under the estimator $M_{\widetilde{\Pi}^{(k)}, \widetilde{\Sigma}^{(k)}}$ is no more than that obtained if one sets the estimate to M_0. The proof will use these quantities in the following manner. First, the matrix M_0 can intuitively be thought of as a "reasonable" regularized least squares estimate (although impractical since computing M_0 requires knowledge of M^*). Treating the objective value in the optimization problem (2b) attained by M_0 as a baseline, the proof can primarily focus on only those estimates (16) whose objective value in (2b) is at most that of M_0. The set of tuples $(k, \widetilde{\Pi}^{(k)}, \widetilde{\Sigma}^{(k)})$ satisfying this condition precisely comprises the set $\widetilde{\Gamma}$.

Note that the set $\widetilde{\Gamma}$ is guaranteed to be non-empty since the parameter and permutations corresponding to M_0 always lie in $\widetilde{\Gamma}$. We now state a lemma which forms a key component of this proof.

Lemma 7 *For any $(k, \widetilde{\Pi}^{(k)}, \widetilde{\Sigma}^{(k)}) \in \widetilde{\Gamma}$, it must be that*

$$\mathbb{P}\Big(\|\!|M_{\widetilde{\Pi}^{(k)}, \widetilde{\Sigma}^{(k)}} - M_0|\!\|_F^2 \leq c_1 \frac{\overline{\rho}(M_0) d \log^{2.01} d}{p_{\text{obs}}} + 4 \|\!|M_0 - M^*|\!\|_F^2 \Big) \geq 1 - e^{-4kd \log d}, \tag{17}$$

for some positive universal constant c_1.

Under our assumption of $d \geq n$, for any value of k the cardinality of the set $\widetilde{\Gamma}$ restricted to any k is at most $e^{2kd \log d}$. Hence a union bound over all $k \in [n]$ and all permutations—applied to equation (17)—yields

$$\mathbb{P}\Big(\max_{(k, \widetilde{\Pi}^{(k)}, \widetilde{\Sigma}^{(k)}) \in \widetilde{\Gamma}} \|M_{\widetilde{\Pi}^{(k)}, \widetilde{\Sigma}^{(k)}} - M_0\|_{\mathrm{F}}^2 \leq c_1 \frac{\overline{\rho}(M_0) d \log^{2.01} d}{p_{\mathrm{obs}}} + 4\|M_0 - M^*\|_{\mathrm{F}}^2 \Big) \geq 1 - e^{-d \log d}. \tag{18}$$

From the definition of the regularized least squares estimator $\widehat{M}_{\mathrm{LS}}$ in equation (2b) and the definition of set $\widetilde{\Gamma}$ above, we have that $\widehat{M}_{\mathrm{LS}}$ must equal $M_{\widetilde{\Pi}^{(k)}, \widetilde{\Sigma}^{(k)}}$ for some $(k, \widetilde{\Pi}^{(k)}, \widetilde{\Sigma}^{(k)}) \in \widetilde{\Gamma}$. As a consequence, the tail bound (18) ensures that

$$\mathbb{P}\Big(\|\widehat{M}_{\mathrm{LS}} - M_0\|_{\mathrm{F}}^2 \leq c_1 \frac{\overline{\rho}(M_0) d \log^{2.01} d}{p_{\mathrm{obs}}} + 4\|M_0 - M^*\|_{\mathrm{F}}^2 \Big) \geq 1 - e^{-d \log d}.$$

Finally, applying the triangle inequality yields the claimed result

$$\mathbb{P}\Big(\|\widehat{M}_{\mathrm{LS}} - M^*\|_{\mathrm{F}}^2 \leq 10\|M^* - M_0\|_{\mathrm{F}}^2 + 2c_1 \frac{\overline{\rho}(M_0) d \log^{2.01} d}{p_{\mathrm{obs}}} \Big) \geq 1 - e^{-d \log d}.$$

Proof of Lemma 7 We begin by writing an equivalent form of the (random) observation and noise matrices that will aid in our analysis. It is straightforward to verify that the observation matrix Y' can equivalently be written in the linearized form

$$Y' = M^* + \frac{1}{p_{\mathrm{obs}}} W', \tag{19a}$$

where W' has entries that are independent, and are distributed as

$$[W']_{ij} = \begin{cases} p_{\mathrm{obs}}(\frac{1}{2} - [M^*]_{ij}) + \frac{1}{2} & \text{with probability } p_{\mathrm{obs}}[M^*]_{ij} \\ p_{\mathrm{obs}}(\frac{1}{2} - [M^*]_{ij}) - \frac{1}{2} & \text{with probability } p_{\mathrm{obs}}(1 - [M^*]_{ij}) \\ p_{\mathrm{obs}}(\frac{1}{2} - [M^*]_{ij}) & \text{with probability } 1 - p_{\mathrm{obs}}. \end{cases} \tag{19b}$$

By definition, any $(k, \widetilde{\Pi}^{(k)}, \widetilde{\Sigma}^{(k)}) \in \widetilde{\Gamma}$ must satisfy the inequality

$$\|Y' - M_{\widetilde{\Pi}^{(k)}, \widetilde{\Sigma}^{(k)}}\|_{\mathrm{F}}^2 + \frac{kd \log^{2.01} d}{p_{\mathrm{obs}}} \leq \|Y' - M_0\|_{\mathrm{F}}^2 + \frac{\overline{\rho}(M_0) d \log^{2.01} d}{p_{\mathrm{obs}}}.$$

Using the linearized form (19a), some algebraic manipulations yield the inequality

$$\begin{aligned} &\|M^* - M_{\widetilde{\Pi}^{(k)}, \widetilde{\Sigma}^{(k)}}\|_{\mathrm{F}}^2 + \frac{2}{p_{\mathrm{obs}}} \langle\langle W', \ M^* - M_{\widetilde{\Pi}^{(k)}, \widetilde{\Sigma}^{(k)}} \rangle\rangle + \frac{kd \log^{2.01} d}{p_{\mathrm{obs}}} \\ &\qquad \leq \|M^* - M_0\|_{\mathrm{F}}^2 + \frac{2}{p_{\mathrm{obs}}} \langle\langle W', \ M^* - M_0 \rangle\rangle + \frac{\overline{\rho}(M_0) d \log^{2.01} d}{p_{\mathrm{obs}}}. \end{aligned} \tag{20}$$

Applying the inequality $\|M_{\widetilde{\Pi}^{(k)}, \widetilde{\Sigma}^{(k)}} - M_0\|_{\mathrm{F}}^2 \leq 2\|M^* - M_{\widetilde{\Pi}^{(k)}, \widetilde{\Sigma}^{(k)}}\|_{\mathrm{F}}^2 + 2\|M^* - M_0\|_{\mathrm{F}}^2$ and denoting $\widehat{\Delta}_{\widetilde{\Pi}^{(k)}, \widetilde{\Sigma}^{(k)}} := (M_{\widetilde{\Pi}^{(k)}, \widetilde{\Sigma}^{(k)}} - M_0)$, some algebraic manipulations yield the inequality

$$\frac{1}{2} \|\widehat{\Delta}_{\widetilde{\Pi}^{(k)}, \widetilde{\Sigma}^{(k)}}\|_{\mathrm{F}}^2 \leq \frac{2}{p_{\mathrm{obs}}} \langle\langle W', \ \widehat{\Delta}_{\widetilde{\Pi}^{(k)}, \widetilde{\Sigma}^{(k)}} \rangle\rangle + \frac{(\overline{\rho}(M_0) - k) d \log^{2.01} d}{p_{\mathrm{obs}}} + 2\|M_0 - M^*\|_{\mathrm{F}}^2. \tag{21}$$

Now consider the set of matrices

$$\mathbb{C}_{\text{DIFF}}(\widetilde{\Pi}^{(k)}, \widetilde{\Sigma}^{(k)}; M_0) := \Big\{\alpha(M - M_0) \mid M \in \mathbb{C}_{\text{PR}}(k; \widetilde{\Pi}^{(k)}, \widetilde{\Sigma}^{(k)}),\ \alpha \in [0,1]\Big\}, \tag{22}$$

and note that $\mathbb{C}_{\text{DIFF}}(\widetilde{\Pi}^{(k)}, \widetilde{\Sigma}^{(k)}; M_0) \subseteq [-1,1]^{n\times d}$. For each choice of radius $t > 0$, define the random variable

$$Z_{\widetilde{\Pi}^{(k)},\widetilde{\Sigma}^{(k)}}(t) := \sup_{\substack{M_{\text{DIFF}} \in \mathbb{C}_{\text{DIFF}}(\widetilde{\Pi}^{(k)},\widetilde{\Sigma}^{(k)};M_0), \\ |\!|\!| M_{\text{DIFF}} |\!|\!|_{\text{F}} \leq t}} \frac{2}{p_{\text{obs}}} \langle\!\langle W',\ M_{\text{DIFF}} \rangle\!\rangle. \tag{23}$$

Using the inequality (21), the Frobenius norm error $|\!|\!|\widehat{\Delta}_{\widetilde{\Pi}^{(k)},\widetilde{\Sigma}^{(k)}}|\!|\!|_{\text{F}}$ then satisfies the bound

$$\frac{1}{2}|\!|\!|\widehat{\Delta}_{\widetilde{\Pi}^{(k)},\widetilde{\Sigma}^{(k)}}|\!|\!|_{\text{F}}^2 \leq\ Z_{\widetilde{\Pi},\widetilde{\Sigma}}\big(|\!|\!|\widehat{\Delta}_{\widetilde{\Pi}^{(k)},\widetilde{\Sigma}^{(k)}}|\!|\!|_{\text{F}}\big) + \frac{(\overline{\rho}(M_0) - k) d \log^{2.01} d}{p_{\text{obs}}} + 2|\!|\!| M_0 - M^* |\!|\!|_{\text{F}}^2. \tag{24}$$

Thus, in order to obtain our desired bound, we need to understand the behavior of the random quantity $Z_{\widetilde{\Pi}^{(k)},\widetilde{\Sigma}^{(k)}}(t)$.

Dy definition, the set $\mathbb{C}_{\text{DIFF}}(\widetilde{\Pi}^{(k)}, \widetilde{\Sigma}^{(k)}; M_0)$ is "star-shaped", meaning that $\alpha M_{\text{DIFF}} \in \mathbb{C}_{\text{DIFF}}(\widetilde{\Pi}^{(k)}, \widetilde{\Sigma}^{(k)})$ for every $\alpha \in [0,1]$ and every $M_{\text{DIFF}} \in \mathbb{C}_{\text{DIFF}}(\widetilde{\Pi}^{(k)}, \widetilde{\Sigma}^{(k)}; M_0)$. Using this star-shaped property, we are guaranteed that $\mathbb{E}[Z_{\widetilde{\Pi}^{(k)},\widetilde{\Sigma}^{(k)}}(\delta)]$ grows at most linearly with δ. We are then in turn guaranteed the existence of some scalar $\delta_c > 0$ satisfying the critical inequality

$$\mathbb{E}[Z_{\widetilde{\Pi}^{(k)},\widetilde{\Sigma}^{(k)}}(\delta_c)] \leq \frac{\delta_c^2}{2}. \tag{25}$$

Our interest is in an upper bound to the smallest (strictly) positive solution δ_c to the critical inequality (25), and moreover, our goal is to show that for every $t \geq \delta_c$, we have $|\!|\!|\widehat{\Delta}|\!|\!|_{\text{F}} \leq c\sqrt{t\delta_c}$ with high probability. To this end, define a "bad" event $\mathcal{A}_t$ as

$$\mathcal{A}_t = \big\{\exists \Delta \in \mathbb{C}_{\text{DIFF}}(\widetilde{\Pi}^{(k)}, \widetilde{\Sigma}^{(k)}; M_0) \mid |\!|\!|\Delta|\!|\!|_{\text{F}} \geq \sqrt{t\delta_c} \quad \text{and} \quad \frac{2}{p_{\text{obs}}}\langle\!\langle W',\ \Delta\rangle\!\rangle \geq 2|\!|\!|\Delta|\!|\!|_{\text{F}}\sqrt{t\delta_c}\big\}. \tag{26}$$

Using the star-shaped property of $\mathbb{C}_{\text{DIFF}}(\widetilde{\Pi}^{(k)}, \widetilde{\Sigma}^{(k)}; M_0)$, it follows by a rescaling argument that

$$\mathbb{P}[\mathcal{A}_t] \leq \mathbb{P}[Z_{\widetilde{\Pi}^{(k)},\widetilde{\Sigma}^{(k)}}(\delta_c) \geq 2\delta_c\sqrt{t\delta_c}] \qquad \text{for all } t \geq \delta_c.$$

The following lemma helps control the behavior of the random variable $Z_{\widetilde{\Pi}^{(k)},\widetilde{\Sigma}^{(k)}}(\delta_c)$.

Lemma 8 *For any $\delta > 0$, the mean of $Z_{\widetilde{\Pi}^{(k)},\widetilde{\Sigma}^{(k)}}(\delta)$ is bounded as*

$$\mathbb{E}[Z_{\widetilde{\Pi}^{(k)},\widetilde{\Sigma}^{(k)}}(\delta)] \leq c_1 \frac{\max\{k, \overline{\rho}(M_0)\} d}{p_{\text{obs}}} \log^2 d,$$

and for every $u > 0$, its tail probability is bounded as

$$\mathbb{P}\Big(Z_{\widetilde{\Pi}^{(k)},\widetilde{\Sigma}^{(k)}}(\delta) > \mathbb{E}[Z_{\widetilde{\Pi}^{(k)},\widetilde{\Sigma}^{(k)}}(\delta)] + u\Big) \leq \exp\Big(\frac{-c_2 u^2 p_{\text{obs}}}{\delta^2 + \mathbb{E}[Z_{\widetilde{\Pi}^{(k)},\widetilde{\Sigma}^{(k)}}(\delta)] + u}\Big),$$

where c_1 and c_2 are positive universal constants.

From this lemma, we have the tail bound

$$\mathbb{P}\Big(Z_{\widetilde{\Pi}^{(k)},\widetilde{\Sigma}^{(k)}}(\delta_c) > \mathbb{E}[Z_{\widetilde{\Pi}^{(k)},\widetilde{\Sigma}^{(k)}}(\delta_c)] + \delta_c\sqrt{t\delta_c}\Big) \leq \exp\Big(\frac{-c_2(\delta_c\sqrt{t\delta_c})^2 p_{\mathrm{obs}}}{\delta_c^2 + \mathbb{E}[Z_{\widetilde{\Pi}^{(k)},\widetilde{\Sigma}^{(k)}}(\delta_c)] + (\delta_c\sqrt{t\delta_c})}\Big),$$

for all $t > 0$. By the definition of δ_c in (25), we have $\mathbb{E}[Z_{\widetilde{\Pi}^{(k)},\widetilde{\Sigma}^{(k)}}(\delta_c)] \leq \delta_c^2 \leq \delta_c\sqrt{t\delta_c}$ for all $t \geq \delta_c$, and consequently

$$\mathbb{P}[\mathcal{A}_t] \leq \mathbb{P}[Z_{\widetilde{\Pi}^{(k)},\widetilde{\Sigma}^{(k)}}(\delta_c) \geq 2\delta_c\sqrt{t\delta_c}] \;\leq\; \exp\Big(\frac{-c_2(\delta_c\sqrt{t\delta_c})^2 p_{\mathrm{obs}}}{3\delta_c\sqrt{t\delta_c}}\Big), \quad \text{for all } t \geq \delta_c. \tag{27}$$

Now we must have either $|\!|\!|\widehat{\Delta}_{\widetilde{\Pi}^{(k)},\widetilde{\Sigma}^{(k)}}|\!|\!|_{\mathrm{F}} \leq \sqrt{t\delta_c}$, or we have $|\!|\!|\widehat{\Delta}_{\widetilde{\Pi}^{(k)},\widetilde{\Sigma}^{(k)}}|\!|\!|_{\mathrm{F}} > \sqrt{t\delta_c}$. In the latter case, conditioning on the complement $\mathcal{A}_t^c$, our basic inequality (21) implies that

$$\frac{1}{2}|\!|\!|\widehat{\Delta}_{\widetilde{\Pi}^{(k)},\widetilde{\Sigma}^{(k)}}|\!|\!|_{\mathrm{F}}^2 \leq 2|\!|\!|\widehat{\Delta}_{\widetilde{\Pi}^{(k)},\widetilde{\Sigma}^{(k)}}|\!|\!|_{\mathrm{F}}\sqrt{t\delta_c} + \frac{(\overline{\rho}(M_0) - k)d\log^{2.01} d}{p_{\mathrm{obs}}} + 2|\!|\!|M_0 - M^*|\!|\!|_{\mathrm{F}}^2.$$

Using the identity $x^2 \leq ax + b \Rightarrow x^2 \leq 4a^2 + 2b$ for every non-negative triplet of scalars x, a and b, we obtain

$$|\!|\!|\widehat{\Delta}_{\widetilde{\Pi}^{(k)},\widetilde{\Sigma}^{(k)}}|\!|\!|_{\mathrm{F}}^2 \leq 16t\delta_c + 2\frac{(\overline{\rho}(M_0) - k)d\log^{2.01} d}{p_{\mathrm{obs}}} + 4|\!|\!|M_0 - M^*|\!|\!|_{\mathrm{F}}^2, \tag{28}$$

with probability at least $\mathbb{P}(\mathcal{A}_t^c) \geq 1 - \exp\Big(\frac{-c_2(\delta_c\sqrt{t\delta_c})^2 p_{\mathrm{obs}}}{3\delta_c\sqrt{t\delta_c}}\Big)$ for all $t \geq \delta_c$ (from (27)). Finally, from the bound on the expected value of $Z_{\widetilde{\Pi}^{(k)},\widetilde{\Sigma}^{(k)}}(t)$ in Lemma 8, we see that the critical inequality (25) is satisfied for

$$\delta_c = \sqrt{\frac{c_1 \max\{\overline{\rho}(M_0), k\}d}{p_{\mathrm{obs}}}\log d}.$$

Setting $t = c'\delta_c$ in (28) for a large enough constant c' yields

$$\mathbb{P}\Big(|\!|\!|\widehat{\Delta}_{\widetilde{\Pi}^{(k)},\widetilde{\Sigma}^{(k)}}|\!|\!|_{\mathrm{F}}^2 \leq \frac{c_1'\overline{\rho}(M_0)d}{p_{\mathrm{obs}}}\log^{2.01} d + 4|\!|\!|M_0 - M^*|\!|\!|_{\mathrm{F}}^2\Big) \geq 1 - \exp\Big(-4\max\{\overline{\rho}(M_0), k\}d\log d\Big),$$

for some constant $c_1' > 0$, thus yielding the claim of the lemma.

Proof of Lemma 8 We break our proof into two parts, corresponding to bounds on the mean of the random variable $Z_{\widetilde{\Pi}^{(k)},\widetilde{\Sigma}^{(k)}}(\delta)$ followed by control of its tail behavior.

Bounding the mean: We begin by establishing an upper bound on the mean $\mathbb{E}[Z_{\widetilde{\Pi}^{(k)},\widetilde{\Sigma}^{(k)}}(\delta)]$. In order to obtain the desired upper bound, in the proof we carefully account for the rank k and the different permutations associated to any matrix in $\mathbb{C}_{\mathrm{DIFF}}(\widetilde{\Pi}^{(k)}, \widetilde{\Sigma}^{(k)}; M_0)$.

For convenience of analysis, we introduce a new random variable

$$\widetilde{Z}_{\widetilde{\Pi}^{(k)},\widetilde{\Sigma}^{(k)}} := \sup_{M_{\mathrm{DIFF}} \in \mathbb{C}_{\mathrm{DIFF}}(\widetilde{\Pi}^{(k)},\widetilde{\Sigma}^{(k)};M_0)} \langle\!\langle W', \; M_{\mathrm{DIFF}}\rangle\!\rangle.$$

Then by definition, we have $\mathbb{E}[Z_{\widetilde{\Pi}^{(k)},\widetilde{\Sigma}^{(k)}}(\delta)] \leq \frac{1}{p_{\text{obs}}}\mathbb{E}[\widetilde{Z}_{\widetilde{\Pi}^{(k)},\widetilde{\Sigma}^{(k)}}]$ for every $\delta > 0$. In addition, since $M_0 \in \mathbb{C}_{\text{PR}}(k)$, it can be decomposed as $M_0 = \sum_{\ell=1}^{k} M_0^{(\ell)}$, for some matrices $M_0^{(1)}, \ldots, M_0^{(k)} \in \mathbb{C}_{\text{PR}}(1)$.

We introduce some additional notation for ease of exposition. If $\overline{\rho}(M_0) < k$, then let $M_0^{(\overline{\rho}(M_0)+1)}, \ldots, M_0^{k}$ denote all-zero matrices. Hence we can write $M_0 = \sum_{\ell=1}^{\max\{\overline{\rho}(M_0),k\}} M_0^{(\ell)}$. If $\overline{\rho}(M_0) > k$ then let $\pi_{k+1}, \ldots, \pi_{\overline{\rho}(M_0)}$ be arbitrary (but fixed) permutations of n items and $\sigma_{k+1}, \ldots, \sigma_{\overline{\rho}(M_0)}$ be arbitrary (but fixed) permutations of d items. With this notation in place, we have the following deterministic upper bound on the value of the random variable $\widetilde{Z}_{\widetilde{\Pi}^{(k)},\widetilde{\Sigma}^{(k)}}$:

$$\widetilde{Z}_{\widetilde{\Pi}^{(k)},\widetilde{\Sigma}^{(k)}} \leq \sum_{\ell=1}^{\max\{\overline{\rho}(M_0),k\}} \sup_{[M_{\text{DIFF}}]_\ell \in \mathbb{C}_{\text{DIFF}}(\{\pi_\ell\},\{\sigma_\ell\};M_0^{(\ell)})} \langle\!\langle W',\ [M_{\text{DIFF}}]_\ell \rangle\!\rangle.$$

We also recall our assumption that $d \geq n$ without loss of generality. Now let $\log N(\epsilon, \mathbb{C}, |\!|\!|\cdot|\!|\!|_{\text{F}})$ denote the ϵ metric entropy of class $\mathbb{C} \subset \mathbb{R}^{n\times d}$ in the Frobenius norm metric $|\!|\!|\cdot|\!|\!|_{\text{F}}$. Then the truncated form of Dudley's entropy integral inequality yields[3]

$$\mathbb{E}[\widetilde{Z}_{\widetilde{\Pi}^{(k)},\widetilde{\Sigma}^{(k)}}] \leq \sum_{\ell=1}^{\max\{\overline{\rho}(M_0),k\}} c\left\{ d^{-8} + \int_{\frac{1}{2}d^{-9}}^{2d} \sqrt{\log N(\epsilon, \mathbb{C}_{\text{DIFF}}(\{\pi_\ell\},\{\sigma_\ell\};M_0^{(\ell)}), |\!|\!|.|\!|\!|_{\text{F}})}(\Delta\epsilon)\right\}, \tag{29}$$

where we have used the fact that the diameter of the set $\mathbb{C}_{\text{DIFF}}(\{\pi_\ell\},\{\sigma_\ell\};M_0^{(\ell)})$ is at most $2d$ in the Frobenius norm.

In our past work (Shah et al., 2017, Lemma 2), we derived a bound on the metric entropy of the set $\mathbb{C}_{\text{DIFF}}(\{\pi_\ell\},\{\sigma_\ell\};M_0^{(\ell)})$ as:

$$\log N\Big(\epsilon, \mathbb{C}_{\text{DIFF}}(\{\pi_\ell\},\{\sigma_\ell\};M_0^{(\ell)}), |\!|\!|\cdot|\!|\!|_{\text{F}}\Big) \leq 16\frac{d^2}{\epsilon^2}\big(\log\frac{d}{\epsilon}\big)^2,$$

for any $\epsilon > 0$ and $\ell \in [k]$. Substituting this bound on the metric entropy into the Dudley bound (29) yields

$$\mathbb{E}[\widetilde{Z}_{\widetilde{\Pi}^{(k)},\widetilde{\Sigma}^{(k)}}] \leq c' \max\{\overline{\rho}(M_0),k\} d \log^2 d.$$

The inequality $\mathbb{E}[Z_{\widetilde{\Pi}^{(k)},\widetilde{\Sigma}^{(k)}}(\delta)] \leq \frac{1}{p_{\text{obs}}}\mathbb{E}[\widetilde{Z}_{\widetilde{\Pi}^{(k)},\widetilde{\Sigma}^{(k)}}]$ then yields the claimed result.

Bounding the tail: In order to establish the claimed tail bound on the deviations of $Z_{\widetilde{\Pi}^{(k)},\widetilde{\Sigma}^{(k)}}(\delta)$ above its mean, we use a Bernstein-type bound on the supremum of empirical processes due to Klein and Rio (2005, Theorem 1.1c), which we state in a simplified form here.

Lemma 9 (Klein and Rio) *Let $X := (X_1, \ldots, X_m)$ be any sequence of zero-mean, independent random variables, each taking values in $[-1,1]$. Let $\mathcal{V} \subset [-1,1]^m$ be any measurable*

3. Here we use $(\Delta\epsilon)$ to denote the differential of ϵ, so as to avoid confusion with the number of columns d.

set of m-length vectors. Then for any $u > 0$, the supremum $X^\dagger = \sup_{v \in \mathcal{V}} \langle X, v \rangle$ satisfies the upper tail bound

$$\mathbb{P}\big(X^\dagger > \mathbb{E}[X^\dagger] + u\big) \leq \exp\Big(\frac{-u^2}{2 \sup_{v \in \mathcal{V}} \mathbb{E}[\langle v,\, X\rangle^2] + 4\mathbb{E}[X^\dagger] + 3u}\Big).$$

We now call upon Lemma 9 setting $\mathcal{V} = \{M_{\text{DIFF}} \in \mathbb{C}_{\text{DIFF}}(\widetilde{\Pi}^{(k)}, \widetilde{\Sigma}^{(k)}; M_0) \mid |\!|\!| M_{\text{DIFF}} |\!|\!|_{\text{F}} \leq \delta\}$, $X = W'$, and $X^\dagger = p_{\text{obs}} Z_{\widetilde{\Pi}^{(k)}, \widetilde{\Sigma}^{(k)}}(\delta)$. The entries of the matrix W' are mutually independent, have a mean of zero, and are bounded by 1 in absolute value. Then we have $\mathbb{E}[X^\dagger] = p_{\text{obs}} \mathbb{E}[Z_{\widetilde{\Pi}^{(k)}, \widetilde{\Sigma}^{(k)}}(\delta)]$ and $\mathbb{E}[\langle\!\langle M_{\text{DIFF}},\ W' \rangle\!\rangle^2] \leq 4 p_{\text{obs}} |\!|\!| M_{\text{DIFF}} |\!|\!|_{\text{F}}^2 \leq 4 p_{\text{obs}} \delta^2$ for every $M_{\text{DIFF}} \in \mathcal{V}$. With these assignments, and some algebraic manipulations, we obtain that for every $u > 0$,

$$\mathbb{P}\big(Z_{\widetilde{\Pi}^{(k)}, \widetilde{\Sigma}^{(k)}}(\delta) > \mathbb{E}[Z_{\widetilde{\Pi}^{(k)}, \widetilde{\Sigma}^{(k)}}(\delta)] + u\big) \leq \exp\Big(\frac{-u^2 p_{\text{obs}}}{8\delta^2 + 4\mathbb{E}[Z_{\widetilde{\Pi}^{(k)}, \widetilde{\Sigma}^{(k)}}(\delta)] + 3u}\Big),$$

as claimed.

4.2. Proof of Theorem 1(b)

Assume without loss of generality that $d \geq n$. Throughout the proof, we ignore floor and ceiling conditions as these are not critical to the proof and affect the lower bound by only a constant factor.

Our proof is based on the framework of Fano's inequality which is a standard method in minimax analysis (see Cover and Thomas, 2012 or Tsybakov, 2008). Here we state the inequality in the context of our problem:

Lemma 10 (Fano's inequality) *Consider any integer $\eta \geq 2$ and any set of η matrices $G^1, \ldots, G^\eta \in B^N(r, \epsilon)$. For each $\ell \in [\eta]$, let $\mathbb{P}^\ell$ denote the probability distribution of the matrix Y obtained by setting $M^* = G^\ell$. Setting ξ_1 and ξ_2 as*

$$\xi_1 = \max_{\ell \neq \ell' \in [\eta]} D_{\text{KL}}(\mathbb{P}^\ell \| \mathbb{P}^{\ell'}) \qquad \textit{and} \qquad \xi_2 = \min_{\ell \neq \ell' \in [\eta]} |\!|\!| G^\ell - G^{\ell'} |\!|\!|_F^2,$$

the minimax risk of estimating M^* from Y is lower bounded as

$$\inf_{\widehat{M}} \sup_{M^* \in B^N(r,\epsilon)} \mathbb{E}\big[|\!|\!| M^* - \widehat{M} |\!|\!|_2^2\big] \geq \frac{\xi_2}{2}\Big(1 - \frac{\xi_1 + \log 2}{\log \eta}\Big). \tag{30}$$

In order to obtain the desired bound using Fano's inequality, we must construct a set of matrices with suitable values of η, ξ_1 and ξ_2. Our first step towards this goal is to call upon the Gilbert-Varshamov bound (Gilbert, 1952; Varshamov, 1957) from coding theory. Here we state this bound in the form of a special case that is required for our problem. As done throughout this manuscript, the notation c represents a positive universal constant.

Lemma 11 (Gilbert-Varshamov bound) *For any integer $\tau \geq 2$, there exists a set of at least $\exp(c\tau)$ binary vectors, each of length τ, such that the Hamming distance between any pair vectors in this set is at least $\tau/10$.*

We now use the framework of Fano and Gilbert-Varshamov to establish our claimed result.

Designing candidate set of matrices for Fano's inequality: In what follows, we use the Gilbert-Varshamov bound to construct a set of binary vectors with certain desired properties, and then use it to in turn obtain a set of matrices for use in Fano's inequality. Let $\tau = dr + p_{\text{obs}}\epsilon^2$. Then the Gilbert-Varshamov bound guarantees the existence of

$$\eta := \exp\Big(c(dr + p_{\text{obs}}\epsilon^2)\Big)$$

binary vectors $g^1, \ldots, g^\eta$, each of length $(dr + p_{\text{obs}}\epsilon^2)$, such that the Hamming distance between any pair of vectors in this set is lower bounded by $\frac{dr+p_{\text{obs}}\epsilon^2}{10}$.

For some $\delta \in (0, \frac{1}{4})$ whose value is specified later, define a related set of vectors $\widetilde{g}^1, \ldots, \widetilde{g}^\eta$ as

$$\widetilde{g}^\ell_j = \begin{cases} \frac{1}{2} + \delta & \text{if } g^\ell_j = 1 \\ \frac{1}{2} - \delta & \text{if } g^\ell_j = 0, \end{cases}$$

for every $\ell \in [\eta]$ and $j \in [dr + p_{\text{obs}}\epsilon^2]$. Next define a set of "low rank" matrices $G^1, \ldots, G^\eta \in [0,1]^{n\times d}$ where the matrix G^ℓ is obtained as follows. For each $\ell \in [\eta]$, arrange the first rd entries of vector $\widetilde{g}^\ell$ as the entries of an $(r \times d)$ matrix—this arrangement may be done in an arbitrary manner as long as it is consistent across every $\ell \in [\eta]$. Now append a $(\frac{p_{\text{obs}}\epsilon^2}{d} \times d)$ matrix at the bottom, whose entries comprise the last $p_{\text{obs}}\epsilon^2$ entries of the vector $\widetilde{g}^\ell$—again, this arrangement may be done in an arbitrary manner as long as it is consistent across every $\ell \in [\eta]$. Now stack $\frac{1}{p_{\text{obs}}}$ copies of the resulting $\big((r + \frac{p_{\text{obs}}\epsilon^2}{d}) \times d\big)$ matrix on top of each other to form a $\big((\frac{r}{p_{\text{obs}}} + \frac{\epsilon^2}{d}) \times d\big)$ matrix. Note that our assumption $\epsilon^2 + \frac{r\max\{n,d\}}{p_{\text{obs}}} \leq nd$, along with the assumption $d \geq n$, implies that $n \geq \frac{r}{p_{\text{obs}}} + \frac{\epsilon^2}{d}$. Append $(n - (\frac{r}{p_{\text{obs}}} + \frac{\epsilon^2}{d}))$ rows of all zeros at the bottom of this matrix, and denote the resultant $(n \times d)$ matrix as G^ℓ. The set of matrices $\{G^1, \ldots, G^\eta\}$ is our candidate set for use in Fano's inequality.

Applying Fano's inequality: In order to apply Fano's inequality, we first establish certain properties of the candidate set of matrices constructed above.

Lemma 12 *The set of matrices $\{G^1, \ldots, G^\eta\}$ satisfies the following three properties:*
(a) $G^\ell \in B^{\mathrm{N}}(r, \epsilon)$ for every $\ell \in [\eta]$.
(b) For every pair $\ell_1 \neq \ell_2 \in [\eta]$, it must be that

$$|||G^{\ell_1} - G^{\ell_2}|||_F^2 \geq \frac{\delta^2}{10}\big(\frac{dr}{p_{\text{obs}}} + \epsilon^2\big).$$

(c) For every $\ell \in [\eta]$, let $\mathbb{P}^\ell$ denotes the probability distribution of the matrix Y obtained by setting $M^* = G^\ell$. Then for every pair $\ell_1 \neq \ell_2 \in [\eta]$, it must be that

$$D_{\text{KL}}(\mathbb{P}^\ell_1 \| \mathbb{P}^{\ell_2}) \leq c'\delta^2 p_{\text{obs}}\big(\frac{dr}{p_{\text{obs}}} + \epsilon^2\big).$$

Part (a) of Lemma 12 ensures that the condition of membership in $B^{\mathrm{N}}(r, \epsilon)$ in the statement of Lemma 10 is satisfied. Then parts (b) and (c) of Lemma 12 allow us to set $\xi_1 =$

$c'\delta^2 p_{\text{obs}}\big(\frac{dr}{p_{\text{obs}}} + \epsilon^2\big)$ and $\xi_2 = \frac{\delta^2}{10}\big(\frac{dr}{p_{\text{obs}}} + \epsilon^2\big)$ in Lemma 10. Substituting these relations in Lemma 10 yields a lower bound on any estimator $\widehat{M}$ for M^* as

$$\begin{aligned}\inf_{\widehat{M}} \sup_{M^* \in B^{\text{N}}(r,\epsilon)} \mathbb{E}[\|\!|\widehat{M} - M^*\|\!|_{\text{F}}^2] &\geq \frac{\delta^2}{20}\Big(\frac{dr}{p_{\text{obs}}} + \epsilon^2\Big)\Big(1 - \frac{c'\delta^2 p_{\text{obs}}\big(\frac{dr}{p_{\text{obs}}} + \epsilon^2\big) + \log 2}{c(dr + p_{\text{obs}}\epsilon^2)}\Big) \\ &\overset{(i)}{\geq} c''\big(\frac{dr}{p_{\text{obs}}} + \epsilon^2\big),\end{aligned}$$

where inequality (i) is obtained by choosing δ^2 as a small enough constant (which depends only on c and c'). Recalling our assumption $d \geq n$, and consequently replacing d by $\max\{n, d\}$ in the bound yields the claimed result.

Proof of Lemma 12 Part (a): We show that the matrix $G^\ell \in [0,1]^{n \times d}$ can be decomposed into a sum of a low-rank matrix (of non-negative rank at most r) and a sparse matrix (number of non-zero entries at most ϵ^2). To this end, we first show that $G^\ell \in B^{\text{N}}(r,\epsilon)$ for every $\ell \in [\eta]$, that is, we show that the matrix $G^\ell \in [0,1]^{n \times d}$ can be decomposed into a sum of a low-rank matrix (of non-negative rank at most r) and a sparse matrix (number of non-zero entries at most ϵ^2). First we set to zero the entries in G^ℓ which correspond to the last $p_{\text{obs}}\epsilon^2$ entries of the vector $\widetilde{g}^\ell$. Let us denote the resulting matrix as $\widetilde{G}^\ell$. Each row of the matrix $\widetilde{G}^\ell$ is either all zero or is identical to one among the first r rows of G^ℓ. Consequently we have $\overline{r}(\widetilde{G}^\ell) \leq r$. Also observe that in the matrix $(G^\ell - \widetilde{G}^\ell)$, the number of non-zero entries is at most $\frac{1}{p_{\text{obs}}} \times p_{\text{obs}}\epsilon^2 = \epsilon^2$, and furthermore, each of these entries lie in the interval $[0,1]$. Hence we have $\|\!|G^\ell - \widetilde{G}^\ell\|\!|_{\text{F}}^2 \leq \epsilon^2$. The matrix G^ℓ thus satisfies all the requirements for membership in the set $B^{\text{N}}(r,\epsilon)$.

Part (b): The claim pertaining to the lower bound on the Frobenius norm difference is easily obtained via simple algebra:

$$\|\!|G^{\ell_1} - G^{\ell_2}\|\!|_{\text{F}}^2 = \frac{1}{p_{\text{obs}}}\|\!|\widetilde{g}^{\ell_1} - \widetilde{g}^{\ell_2}\|\!|_2^2 = \frac{4\delta^2}{p_{\text{obs}}} D_H(g^{\ell_1}, g^{\ell_2}),$$

where $D_H(g^{\ell_1}, g^{\ell_2})$ denotes the Hamming distance between the two binary vectors g^{ℓ_1} and g^{ℓ_2}. From our construction, the Hamming distance between any pair of vectors in the set $\{g^1, \ldots, g^\eta\}$ is at least $\frac{dr + p_{\text{obs}}\epsilon^2}{10}$, thereby yielding the claimed result.

Part (c): Since the entries of Y are mutually independent, we have the decomposition

$$D_{\text{KL}}(\mathbb{P}^{\ell_1} \| \mathbb{P}^{\ell_2}) = \sum_{i \in [n], j \in [d]} D_{\text{KL}}(\mathbb{P}^{\ell_1}_{ij} \| \mathbb{P}^{\ell_2}_{ij}),$$

where we have slightly abused notation to let $\mathbb{P}^\ell_{ij}$ denote the (marginal) probability distribution induced on Y_{ij} under $\mathbb{P}^\ell$. Now, the structure of the matrices $G^1, \ldots, G^\eta$ helps simplify this bound to

$$\begin{aligned} D_{\mathrm{KL}}(\mathbb{P}^{\ell_1} \| \mathbb{P}^{\ell_2}) &= \frac{1}{p_{\mathrm{obs}}} \sum_{i \in [r + \frac{p_{\mathrm{obs}}\epsilon^2}{d}], j \in [d]} D_{\mathrm{KL}}(\mathbb{P}^{\ell_1}_{ij} \| \mathbb{P}^{\ell_2}_{ij}) \\ &\overset{(i)}{\leq} \frac{1}{p_{\mathrm{obs}}} \Big(\frac{p_{\mathrm{obs}} \delta^2 (dr + p_{\mathrm{obs}} \epsilon^2)}{(1/4 - \delta^2)} \Big) \\ &\overset{(ii)}{\leq} c' \delta^2 p_{\mathrm{obs}} \big(\frac{dr}{p_{\mathrm{obs}}} + \epsilon^2 \big), \end{aligned}$$

where inequality (i) follows from the elementary identity $a \log \frac{a}{b} \leq (a-b)\frac{a}{b}$ for all $a, b \in (0,1)$ along with some algebraic simplifications, and inequality (ii) is due to the restriction $\delta \in (0, \frac{1}{4})$. We thus obtain the claimed result.

4.3. Proof of Theorem 2

We now turn to analysis of the singular-value thresholding (SVT) estimator. This proof is based on the framework of a proof from our earlier work (Shah et al., 2017, Theorem 2), which can be seen as a particular case with $n = d$ and $\rho = 1$. We introduce certain additional tricks in order to generalize the proof for general values of ρ and to obtain a sharp dependence on ρ. As in our previous proofs, we may assume without loss of generality that $n \leq d$.

The high-level idea behind the proof is as follows. The goal is to bound the term $\|\widehat{M}_{\text{SVT}} - M^*\|_{\text{F}}^2$. Observe that $\widehat{M}_{\text{SVT}}$ depends on the true matrix M^*, the realization of the noise Y and the threshold λ. The first step is to bound the target term in a manner that explicitly constitutes these components of $\widehat{M}_{\text{SVT}}$. Such a decomposition will then allow for a simplified analysis where we can separately analyze the individual constituents. The terms corresponding to M^* are then bounded in an algebraic manner, and the noise term is bounded using tools from random matrix theory. These individual results are then put together by means of some careful algebra in order to obtain the claimed result.

Let us now present the complete proof. Recall from equation (19a) that we can write our observation model as $Y' = M^* + \frac{1}{p_{\mathrm{obs}}} W'$, where $W' \in [-1,1]^{n \times d}$ is a zero-mean matrix with mutually independent entries. Also recall that these entries follow the distribution

$$[W']_{ij} = \begin{cases} p_{\mathrm{obs}}(\frac{1}{2} - [M^*]_{ij}) + \frac{1}{2} & \text{with probability } p_{\mathrm{obs}}[M^*]_{ij} \\ p_{\mathrm{obs}}(\frac{1}{2} - [M^*]_{ij}) - \frac{1}{2} & \text{with probability } p_{\mathrm{obs}}(1 - [M^*]_{ij}) \\ p_{\mathrm{obs}}(\frac{1}{2} - [M^*]_{ij}) & \text{with probability } 1 - p_{\mathrm{obs}}. \end{cases} \tag{31}$$

For any matrix A, let $\sigma_1(A), \sigma_2(A), \ldots$ denote its singular values in descending order.

Our proof of the upper bound is based on three lemmas. The first lemma is a result from our earlier work (Shah et al., 2017). The key idea behind this lemma is to bound the term $\|\widehat{M}_{\text{SVT}} - M^*\|_{\text{F}}^2$ in a manner that explicitly captures the terms constituting $\widehat{M}_{\text{SVT}}$.

Lemma 13 *(Shah et al., 2017, Lemma 3) If $\lambda \geq \frac{1.01 \|W'\|_{op}}{p_{\mathrm{obs}}}$, then*

$$\|\widehat{M}_{SVT} - M^*\|_F^2 \leq c \sum_{j=1}^{n} \min\big\{\lambda^2, \sigma_j^2(M^*)\big\} \tag{32}$$

with probability at least $1 - c_1 e^{-c'n}$*, where* c*,* c_1 *and* c' *are positive universal constants.*

Our second lemma now helps to bound the right hand side of the inequality (32). The lemma presents an approximation-theoretic result that bounds the tail of the singular values of any matrix with a given permutation-rank or non-negative rank. The proof of this lemma builds on a construction due to Chatterjee (2014).

Lemma 14 *(a) For any matrix* $M \in \mathbb{C}_{\scriptscriptstyle PR}(\rho)$ *and any* $s \in \{1, 2, \ldots, n-1\}$*, we have*

$$\sum_{j=s+1}^{n} \sigma_j^2(M) \leq \frac{nd\rho^2}{s}.$$

(b) For any matrix $M \in \mathbb{C}_{\scriptscriptstyle NR}(r)$ *and any* $s \in \{1, 2, \ldots, n-1\}$*, we have*

$$\sum_{j=s+1}^{n} \sigma_j^2(M) \leq nd \max\big\{\frac{r-s}{r}, 0\big\}.$$

Observe that the result (32) applies under the condition on the noise matrix W' stated in Lemma 13. Our third lemma hence controls the noise term W' in a manner that will allow us to apply Lemma 13.

Lemma 15 *Suppose that* $p_{\text{obs}} \geq \frac{1}{\min\{n,d\}} \log^7(nd)$*. Then given a random matrix* W' *with entries distributed according to the distribution* (31)*, we have*

$$\mathbb{P}\big(|\!|\!|W'|\!|\!|_{\scriptscriptstyle op} > 2.01\sqrt{p_{\text{obs}}(n+d)}\big) \leq e^{-c' \max\{n,d\}}.$$

Based on these three lemmas, we now complete the proof of the theorem. From Lemma 15 we see that the choice $\lambda = 2.1\sqrt{\frac{n+d}{p_{\text{obs}}}}$ guarantees that $\lambda \geq \frac{1.01|\!|\!|W'|\!|\!|_{\text{op}}}{p_{\text{obs}}}$ with probability at least $1 - e^{-c' \max\{n,d\}}$. Consequently, the condition required for an application of Lemma 13 is satisfied, and applying this lemma then yields the upper bound

$$|\!|\!|\widehat{M}_{\text{SVT}} - M^*|\!|\!|_{\text{F}}^2 \leq c \sum_{j=1}^{n} \min\big\{\frac{d}{p_{\text{obs}}}, \sigma_j^2(M^*)\big\} \tag{33}$$

with probability at least $1 - e^{-c'd}$, where we have also used our assumption that $n \leq d$.

Now consider any matrix $M_0 \in \mathbb{R}^{n \times d}$. In what follows, we convert the bound (33) into a bound that depends on the properties of M_0, namely $\overline{\rho}(M_0)$, $\overline{r}(M_0)$ and $|\!|\!|M^* - M_0|\!|\!|_{\text{F}}$. Let $\rho_0 = \overline{\rho}(M_0)$ and $r_0 = \overline{r}(M_0)$.

We first have the following deterministic upper bound

$$\begin{aligned}
\sum_{j=1}^{n} \min\big\{\frac{d}{p_{\text{obs}}}, \sigma_j^2(M^*)\big\} &\leq 2\sum_{j=1}^{n} \min\big\{\frac{d}{p_{\text{obs}}}, \sigma_j^2(M_0)\big\} \\
&\quad + 2\sum_{j=1}^{n} \Big(\min\big\{\sqrt{\frac{d}{p_{\text{obs}}}}, \sigma_j(M^*)\big\} - \min\big\{\sqrt{\frac{d}{p_{\text{obs}}}}, \sigma_j(M_0)\big\}\Big)^2 \\
&\leq 2\sum_{j=1}^{n} \min\big\{\frac{d}{p_{\text{obs}}}, \sigma_j^2(M_0)\big\} + 2\sum_{j=1}^{n} \big(\sigma_j(M^*) - \sigma_j(M_0)\big)^2, \quad (34)
\end{aligned}$$

where the inequality (34) is a consequence of the more general result that $(\min\{a, b_1\} - \min\{a, b_2\})^2 \leq (b_1 - b_2)^2$ for any three real numbers a, b_1, and b_2. In what follows, we bound the two terms on the right hand side of (34) separately.

Bounding the second term on the right hand side of (34): In order to bound the second term, we call upon a more general relation pertaining to matrices established in Schönemann (1968).

Lemma 16 (Schönemann, 1968) *For any pair of matrices $A, B \in \mathbb{R}^{n \times d}$ with singular value decompositions $A = U_1 D_1 V_1^T$ and $B = U_2 D_2 V_2^T$, it must be that*

$$(U_1 U_2^T,\ V_1 V_2^T) \in \underset{U \in \mathbb{R}^{n\times n},\ V \in \mathbb{R}^{d \times d}}{\arg\min} \ |||A - UBV^T|||_F^2 \qquad (35)$$
$$\text{such that } U^T = U^{-1},\ V^T = V^{-1}.$$

To bound the second term on the right hand side of (34), we use Lemma 16 with the choices $A = M^*$ and $B = M_0$. With this choice, some simple algebra yields that the minimum value of the objective $|||A - UBV^T|||_{\text{F}}^2$ in (35) equals $\sum_{j=1}^n (\sigma_j(M^*) - \sigma_j(M_0))^2$. On the other hand, the choice of U and V as identity matrices is feasible for (35), and the associated value of the objective equals $|||M^* - M_0|||_{\text{F}}^2$. Consequently, we have the inequality

$$\sum_{j=1}^n (\sigma_j(M^*) - \sigma_j(M_0))^2 \leq |||M^* - M_0|||_{\text{F}}^2. \qquad (36)$$

Bounding the first term on the right hand side of (34): As for the first term of (34), an application of Lemma 14(a) to the matrix M_0 yields the bound

$$\begin{aligned}\sum_{j=1}^n \min\big\{\frac{d}{p_{\text{obs}}}, \sigma_j^2(M_0)\big\} &\leq \min_{s \in [n]} \big(\frac{sd}{p_{\text{obs}}} + \frac{(\rho_0)^2 nd}{s}\big) \\ &\leq 3\rho_0 d \sqrt{\frac{n}{p_{\text{obs}}}}, \qquad (37)\end{aligned}$$

where inequality (37) is obtained with the choice $s = \lceil \rho_0 \sqrt{p_{\text{obs}} n}\rceil$. Separately, an application of Lemma 14(b) to the matrix M_0 yields

$$\begin{aligned}\sum_{j=1}^n \min\big\{\frac{d}{p_{\text{obs}}}, \sigma_j^2(M_0)\big\} &\leq \min_{s \in [n]} \big(\frac{sd}{p_{\text{obs}}} + nd \max\big\{1 - \frac{s}{r_0}, 0\big\}\big) \\ &\leq \frac{r_0 d}{p_{\text{obs}}}, \qquad (38)\end{aligned}$$

where the inequality (38) is obtained with the choice $s = r_0$.

Putting it all together: Combining the bounds (33), (34), (36), (37) and (38), we obtain the result that the inequality

$$|||\widehat{M}_{\text{SVT}} - M^*|||_{\text{F}}^2 \leq 2 \min\Big\{\frac{3\rho_0 d\sqrt{n}}{\sqrt{p_{\text{obs}}}}, \frac{r_0 d}{p_{\text{obs}}}\Big\} + 2|||M^* - M_0|||_{\text{F}}^2,$$

must hold with probability at least $1 - e^{c'd}$. Finally, recalling our assumption that $d \geq n$ and substituting $n = \min\{n, d\}$ and $d = \max\{n, d\}$ yields the claimed result.

Proof of Lemma 14 Part (a): Without loss of generality, assume that $d \geq n$.

We begin with an upper bound on the tail of the singular values of any matrix in $\mathbb{C}_{\text{PR}}(1)$, that is, of any matrix that has a permutation-rank of 1. The proof of this bound uses a construction due to Chatterjee (2014) for a rank $\widetilde{s}$ approximation of any matrix in $\mathbb{C}_{\text{PR}}(1)$, for any value $\widetilde{s} \in [n]$. We first reproduce Chatterjee's construction.

For a given matrix $M \in \mathbb{C}_{\text{PR}}(1)$, define the vector $\tau \in \mathbb{R}^d$ of column sums—namely, with entries $\tau_j = \sum_{i=1}^n [M]_{ij}$ for $j \in [d]$. Using this vector, define a rank $\widetilde{s}$ approximation $\widetilde{M}$ to M by grouping the columns according to the vector τ according to the following procedure:

- Observing that each $\tau_j \in [0, n]$, divide the full interval $[0, n]$ into $\widetilde{s}$ groups—say of the form
$$[0, n/\widetilde{s}), [n/\widetilde{s}, 2n/\widetilde{s}), \ldots [(\widetilde{s}-1)n/\widetilde{s}, n].$$
If τ_j falls into the interval α for some $\alpha \in [\widetilde{s}]$, then map column j to the group G_α of indices.
- For each $\alpha \in [\widetilde{s}]$ such that group G_α is non-empty, choose a particular column index $j' \in G_\alpha$ in an *arbitrary* fashion. For every other column index $j \in G_\alpha$, set $\widetilde{M}_{ij} = M_{ij'}$ for all $i \in [n]$.

By construction, the matrix $\widetilde{M}$ has at most $\widetilde{s}$ distinct rows, and hence rank at most $\widetilde{s}$. Now consider any column $j \in [d]$ and suppose that $j \in G_\alpha$ for some $\alpha \in [\widetilde{s}]$. Let j' denote the column chosen for the group G_α in the second step of the construction. Since $M \in \mathbb{C}_{\text{PR}}(1)$, we must either have $M_{ij} \geq M_{ij'} = \widetilde{M}_{ij}$ for every $i \in [n]$, or $M_{ij} \leq M_{ij'} = \widetilde{M}_{ij}$ for every $i \in [n]$. Then we are guaranteed that

$$\sum_{i=1}^{n} |\widetilde{M}_{ij} - M_{ij}| = | \sum_{i=1}^{n} (\widetilde{M}_{ij} - M_{ij}) | = |\tau_{j'} - \tau_j| \leq \frac{n}{\widetilde{s}}, \tag{39}$$

where we have used the fact the pair $(\tau_j, \tau_{j'})$ must lie in an interval of length at most $n/\widetilde{s}$. This completes the description of Chatterjee's construction.

In what follows, we use Chatterjee's result in order to obtain our claimed bound on the tail of the spectrum of any matrix $M \in \mathbb{C}_{\text{PR}}(\rho)$. We modify the result in a careful manner that allows us to obtain the desired dependence on the parameter ρ. Recall that any matrix $M \in \mathbb{C}_{\text{PR}}(\rho)$ can be decomposed as

$$M = \sum_{\ell=1}^{\rho} M^{(\ell)},$$

for some matrices $M^{(1)}, \ldots, M^{(\rho)} \in \mathbb{C}_{\text{PR}}(1)$. Let $\widetilde{s} = \frac{s}{\rho}$. For every $\ell \in [r]$, let $\widetilde{M}^{(\ell)}$ be a rank $\widetilde{s} = \frac{s}{\rho}$ approximation of $M^{(\ell)}$ obtained from Chatterjee's construction above, but with the following additional detail. Observe that in Chatterjee's construction, the choice of column j' from group G_α is arbitrary. For our construction, we will make a specific choice of this column: we choose the column whose entries have the smallest values among all columns in the group G_α. With this choice, we have the property

$$\widetilde{M}^{(\ell)}_{ij} \leq M^{(\ell)}_{ij} \qquad \text{for every } \ell \in [\rho],\ i \in [n], j \in [d]. \tag{40}$$

Now let $\widetilde{M} := \sum_{\ell=1}^{\rho} \widetilde{M}^{(\ell)}$. Since every entry of every matrix $\widetilde{M}^{(\ell)}$ is non-negative, we have that every entry of $\widetilde{M}$ is also non-negative. We also claim that

$$\widetilde{M}_{ij} = \sum_{\ell=1}^{\rho} \widetilde{M}^{(\ell)}_{ij} \overset{(i)}{\leq} \sum_{\ell=1}^{\rho} M^{(\ell)}_{ij} = M_{ij} \leq 1,$$

where the inequality (i) is a consequence of the set of inequalities (40). Thus we have that $\widetilde{M} \in [0,1]^{n \times d}$, $M \in [0,1]^{n \times d}$, and that the rank of $\widetilde{M}$ is at most $\rho\widetilde{s}$. This result then yields the bound

$$\begin{aligned}\sum_{j=\rho\widetilde{s}+1}^{n} \sigma_j^2(M) \leq \|M - \widetilde{M}\|_{\mathrm{F}}^2 \leq \sum_{i=1}^{n}\sum_{j=1}^{d} |M_{ij} - \widetilde{M}_{ij}| &= \sum_{i=1}^{n}\sum_{j=1}^{d} \Big|\sum_{\ell=1}^{\rho} (M^{(\ell)}_{ij} - \widetilde{M}^{(\ell)}_{ij})\Big| \\ &\overset{(i)}{\leq} \sum_{i=1}^{n}\sum_{j=1}^{d}\sum_{\ell=1}^{\rho} |M^{(\ell)}_{ij} - \widetilde{M}^{(\ell)}_{ij}|,\end{aligned}$$

where inequality (i) follows from the triangle inequality. Now recall that every matrix $\widetilde{M}^{(\ell)}$ is obtained from Chatterjee's construction, and rewriting Chatterjee's result (39) for the matrices $\widetilde{M}^{(\ell)}$ presently under consideration, we obtain

$$\sum_{i=1}^{n} |\widetilde{M}^{(\ell)}_{ij} - M^{(\ell)}_{ij}| \leq \frac{n}{\widetilde{s}},$$

for every $\ell \in [\rho]$. As a consequence, we have

$$\sum_{j=\rho\widetilde{s}+1}^{n} \sigma_j^2(M) \leq \frac{\rho n d}{\widetilde{s}} = \frac{\rho^2 n d}{s},$$

where we have substituted the relation $\widetilde{s} = \frac{s}{\rho}$ to obtain the final result.

Part (b): This result follows directly from the facts that the rank of M is at most r, and the square of its Frobenius norm is at most nd.

Proof of Lemma 15 Define an $((n+d) \times (n+d))$ matrix W'' as

$$W'' = \frac{1}{\sqrt{p_{\text{obs}}}} \begin{bmatrix} 0 & W' \\ (W')^T & 0 \end{bmatrix}.$$

From (31) and the construction above, we have that the matrix W'' is symmetric, and has mutually independent entries above the diagonal that have a mean of zero and a variance upper bounded by 1. Consequently, known results in random matrix theory (e.g., see Chatterjee, 2014, Theorem 3.4 or Tao, 2012, Theorem 2.3.21) yield the bound $\|W''\|_{\text{op}} \leq 2.01\sqrt{n+d}$ with probability at least $1 - e^{-c\max\{n,d\}}$, under the assumption $p_{\text{obs}} \geq \frac{1}{\min\{n,d\}} \log^7(nd)$. One can also verify that $\|W''\|_{\text{op}} = \frac{1}{\sqrt{p_{\text{obs}}}} \|W'\|_{\text{op}}$, yielding the claimed result.

4.4. Proof of Proposition 3

We recall that for any integer $k \geq 0$, the notation J_k denotes an upper triangular matrix of size $(k \times k)$ with all entries on and above the diagonal set as 1, and I_k denotes the identity matrix of size $(k \times k)$. Consider an $(n \times d)$ matrix M with the following block structure:

$$M := \begin{bmatrix} J_{r-\rho+1} & 0 & 0 \\ 0 & I_{\rho-1} & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

In the remainder of the proof, we show that $\overline{r}(M) = r$ and $\overline{\rho}(M) = \rho$. Using the ideas in the construction of M and the associated proof to follow, one can construct many other matrices that have a non-negative rank of r and a permutation-rank of ρ, for any given value $1 \leq \rho \leq r \leq \min\{n, d\}$.

We partition the proof into four parts.

Proof of $\overline{r}(M) \leq r$: One can write M as a sum of r matrices, each having a non-negative rank of one: for each non-zero row, consider a component matrix comprising that row and zeros elsewhere. Consequently, we have $\overline{r}(M) \leq r$.

Proof of $\overline{r}(M) \geq r$: Observe that the (conventional) rank of M^* equals r. Since the rank of any matrix is a lower bound on its non-negative rank, we have that $\overline{r}(M) \geq r$. We have thus established that the non-negative rank of this matrix equals exactly r.

Proof of $\overline{\rho}(M) \leq \rho$: Observe that the $(n \times d)$ matrix with $J_{r-\rho+1}$ as its top-left submatrix and zeros elsewhere has a permutation-rank of 1. Moreover, any $(n \times d)$ matrix with exactly one entry as 1 and the remaining entries 0 also has a permutation-rank of 1, and hence a $(n \times d)$ matrix with $I_{\rho-1}$ as its submatrix and zeros elsewhere has a permutation-rank of at most $(\rho - 1)$. Putting these arguments together, we obtain the bound $\overline{\rho}(M) \leq \rho$.

Proof that $\overline{\rho}(M) \geq \rho$: First observe that the matrix

$$I_{2\times 2} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

does not belong to $\mathbb{C}_{\text{PR}}(1)$. It follows that any matrix containing $I_{2\times 2}$ as a submatrix cannot belong to the set $\mathbb{C}_{\text{PR}}(1)$. It further follows that for any positive integer k, the matrix $I_{k\times k}$ must have a permutation rank of at least k. Finally, observe that the matrix M contains $I_{\rho\times\rho}$ as its submatrix (given by the intersection of rows $\{r - \rho, \ldots, r\}$ with the columns $\{r - \rho, \ldots, r\}$). It follows that M must have a permutation rank of at least ρ, thereby proving the claim.

4.5. Proof of Proposition 4

We assume for ease of exposition that n and d are divisible by k. Otherwise, since $k \leq \frac{1}{2}\min\{n, d\}$, one may take floors or ceilings which will change the result only by a constant factors. Since $\mathbb{C}_{\text{NR}}(k) \subseteq \mathbb{C}_{\text{PR}}(k)$, we have $\sup\limits_{M \in \mathbb{C}_{\text{NR}}(k)} \inf\limits_{M' \in \mathbb{C}_{\text{PR}}(k)} |||M - M'|||_{\text{F}}^2 = 0$. In what follows, we show that $\sup\limits_{M \in \mathbb{C}_{\text{PR}}(k)} \inf\limits_{M' \in \mathbb{C}_{\text{NR}}(k)} |||M - M'|||_{\text{F}}^2 \geq \frac{cnd}{k}$.

Consider the block matrix $\widetilde{M} \in [0,1]^{\frac{n}{k} \times \frac{d}{k}}$:

$$\widetilde{M} = \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix}, \tag{41}$$

where each of the four blocks is of size $(\frac{n}{2k} \times \frac{d}{2k})$. The following lemma shows that the best rank-1 approximation to $\widetilde{M}$ has a large approximation error:

Lemma 17 *For the matrix $\widetilde{M}$ defined in* (41)*, for any vectors $u \in \mathbb{R}^n$ and $v \in \mathbb{R}^d$, it must be that*

$$|||\widetilde{M} - uv^T|||_{\scriptscriptstyle F}^2 \geq c\frac{nd}{k^2},$$

where $c > 0$ is a universal constant.

We now use the matrix $\widetilde{M}$ defined in (41) to build the following block matrix $M \in \mathbb{C}_{\text{PR}}(k)$:

$$M := \begin{bmatrix} \widetilde{M} & 0 & \cdots & 0 \\ 0 & \widetilde{M} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \widetilde{M} \end{bmatrix}.$$

In words, the matrix M is a block-diagonal matrix where the diagonal has k copies of $\widetilde{M}$.

Due to the block diagonal structure of M, the singular values of M are simply k copies of the singular values of its constituent matrix $\widetilde{M}$. Consequently, we have that for any matrix $M' \in \mathbb{C}_{\text{NR}}(k)$:

$$|||M - M'|||_{\text{F}}^2 \geq k(|||\widetilde{M}|||_{\text{F}}^2 - |||\widetilde{M}|||_{\text{op}}^2) \stackrel{(i)}{\geq} c\frac{nd}{k},$$

as claimed, where the inequality (i) is a consequence of Lemma 17.

Proof of Lemma 17 Consider any value $i \in [\frac{n}{2k}]$ and $j \in [\frac{d}{2k}]$. Then we claim that

$$\begin{aligned}(\widetilde{M}_{i,j} - [uv^T]_{i,j})^2 + (\widetilde{M}_{i+\frac{n}{2k},j} - [uv^T]_{i+\frac{n}{2k},j})^2 + (\widetilde{M}_{i,j+\frac{d}{2k}} - [uv^T]_{i,j+\frac{d}{2k}})^2 \\ + (\widetilde{M}_{i+\frac{n}{2k},j+\frac{d}{2k}} - [uv^T]_{i+\frac{n}{2k},j+\frac{d}{2k}})^2 \geq 0.01.\end{aligned} \tag{42}$$

If not, then for the choice of $\widetilde{M}$ in (41), we must have $[uv^T]_{i,j} \in (0.9, 1.1)$, $[uv^T]_{i+\frac{n}{2k},j} \in (0.9, 1.1)$, $[uv^T]_{i,j+\frac{d}{2k}} \in (0.9, 1.1)$ and $[uv^T]_{i+\frac{n}{2k},j+\frac{d}{2k}} < 0.1$. However, since $[uv^T]_{i',j'} = u_{i'}v_{j'}$ for every coordinate (i', j'), we also have

$$[uv^T]_{i,j} \times [uv^T]_{i+\frac{n}{2k},j+\frac{d}{2k}} = [uv^T]_{i+\frac{n}{2k},j} \times [uv^T]_{i,j+\frac{d}{2k}},$$

which contradicts the required ranges of the individual coordinates. Summing the bound (42) over all values of $i \in [\frac{n}{2k}]$ and $j \in [\frac{d}{2k}]$ yields the claimed result.

4.6. Proof of Proposition 5

Consider any set $\mathbb{S}$ and any convex set $\mathbb{C}$. We begin with a key lemma that establishes a relation between $\mathcal{H}(\mathbb{S}, \mathbb{C})$ and a proposed notion of the inherent convexity of $\mathbb{S}$.

Lemma 18 *For any set $\mathbb{S} \subseteq [0,1]^{n\times d}$ and any convex set $\mathbb{C} \subseteq [0,1]^{n\times d}$, it must be that*

$$\mathcal{H}(\mathbb{S}, \mathbb{C}) \geq \frac{2}{9} \sup_{M_1 \in \mathbb{S},\ M_2 \in \mathbb{S}} \inf_{M_0 \in \mathbb{S}} \|\frac{1}{2}(M_1 + M_2) - M_0\|_F^2. \tag{43}$$

The left hand side of inequality (43) is the Hausdorff distance between the sets $\mathbb{S}$ and $\mathbb{C}$ in terms of the squared Frobenius norm. The right hand side of the inequality represents a notion of the inherent convexity of the set $\mathbb{S}$.

With this lemma in place, we now complete the remainder of the proof. To this end, we set $\mathbb{S} = \mathbb{C}_{\mathrm{PR}}(1)$, and let $\mathbb{C}$ be any convex set of $[0,1]$-valued $(n \times d)$ matrices.

We now construct a pair of matrices $M_1 \in \mathbb{C}_{\mathrm{PR}}(1)$ and $M_2 \in \mathbb{C}_{\mathrm{PR}}(1)$ that we use to lower bound the right hand side of (43). Define matrices $M_1 \in \mathbb{C}_{\mathrm{PR}}(1)$ and $M_2 \in \mathbb{C}_{\mathrm{PR}}(1)$ as

$$[M_1]_{ij} = \begin{cases} 1 & \text{if } i \leq \frac{n}{2},\ j \leq \frac{d}{2} \\ 0 & \text{otherwise,} \end{cases} \qquad \text{and} \qquad [M_2]_{ij} = \begin{cases} 1 & \text{if } i > \frac{n}{2},\ j > \frac{d}{2} \\ 0 & \text{otherwise.} \end{cases}$$

It follows that the entries of the matrix $\frac{1}{2}(M_1 + M_2)$ are given by:

$$[\frac{1}{2}(M_1 + M_2)]_{ij} = \begin{cases} \frac{1}{2} & \text{if } (i \leq \frac{n}{2},\ j \leq \frac{d}{2}) \text{ or } (i > \frac{n}{2},\ j > \frac{d}{2}) \\ 0 & \text{otherwise.} \end{cases}$$

Now consider any pair of integers $(i,j) \in [\lfloor n/2 \rfloor] \times [\lfloor d/2 \rfloor]$. Then the (2×2) submatrix of $\frac{1}{2}(M_1 + M_2)$ formed by its entries (i,j), $(i + \lceil n/2 \rceil, j)$, $(i, j + \lceil d/2 \rceil)$ and $(i + \lceil n/2 \rceil, j + \lceil d/2 \rceil)$ equals

$$\begin{bmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{2} \end{bmatrix}.$$

It is easy to verify that there is a constant $c > 0$ such that the squared Frobenius norm distance between this rescaled identity matrix and any (2×2) matrix in $\mathbb{C}_{\mathrm{PR}}(1)$ is at least c. Since this argument holds for any choice of $(i,j) \in [\lfloor n/2 \rfloor] \times [\lfloor d/2 \rfloor]$, summing up the errors across each of these sets of entries yields

$$\|\frac{1}{2}(M_1 + M_2) - M\|_{\mathrm{F}}^2 \geq c' nd, \qquad \text{for every matrix } M \in \mathbb{C}_{\mathrm{PR}}(1),$$

where $c' > 0$ is a universal constant. Finally, substituting this bound in Lemma 18 yields the claimed result.

It remains to prove Lemma 18.

Proof of Lemma 18. Consider any matrices $M_1 \in \mathbb{S}$ and $M_2 \in \mathbb{S}$. From the definition of the Hausdorff distance $\mathcal{H}$, we know that there exist matrices $\widetilde{M}_1 \in \mathbb{C}$ and $\widetilde{M}_2 \in \mathbb{C}$ such that

$$\|M_i - \widetilde{M}_i\|_{\mathrm{F}}^2 \leq \mathcal{H}(\mathbb{S}, \mathbb{C}), \quad \text{for } i \in \{1, 2\}. \tag{44}$$

Since $\mathbb{C}$ is a convex set, we also have $\frac{1}{2}(\widetilde{M}_1 + \widetilde{M}_2) \in \mathbb{C}$. Then from the definition of $\mathcal{H}$, we also know that there exists a matrix $M_0 \in \mathbb{S}$ such that

$$\|\frac{1}{2}(\widetilde{M}_1 + \widetilde{M}_2) - M_0\|_{\mathrm{F}}^2 \leq \mathcal{H}(\mathbb{S}, \mathbb{C}). \tag{45}$$

Finally, applying the triangle inequality to the bounds (44) and (45) yields

$$\begin{aligned} \|\frac{1}{2}(M_1 + M_2) - M_0\|_{\mathrm{F}}^2 &\leq 3\|\frac{1}{2}(\widetilde{M}_1 + \widetilde{M}_2) - M_0\|_{\mathrm{F}}^2 + \frac{3}{4}\|M_1 - \widetilde{M}_1\|_{\mathrm{F}}^2 + \frac{3}{4}\|M_2 - \widetilde{M}_2\|_{\mathrm{F}}^2 \\ &\leq \frac{9}{2}\mathcal{H}(\mathbb{S}, \mathbb{C}). \end{aligned}$$

4.7. Proof of Proposition 6

Suppose there exists a coordinate pair (i, j) such the stated condition is violated. Then there must exist two distinct values $\ell_1 \in [\overline{\rho}(M)]$ and $\ell_2 \in [\overline{\rho}(M)]$ that satisfy the following three conditions:

(a) $M_{ij}^{(\ell_1)} > 0$ and $M_{ij}^{(\ell_2)} > 0$,

(b) The value of $M_{ij}^{(\ell_1)}$ is different from all other entries in $M^{(\ell_1)}$, and

(c) The value of $M_{ij}^{(\ell_2)}$ is different from all other entries in $M^{(\ell_2)}$.

In addition, the fact that $M_{ij}^{(\ell_1)} + M_{ij}^{(\ell_2)} \in (0, 1)$ for every coordinate (i, j), along with condition (a) above, imply a fourth condition:

(d) $M_{ij}^{(\ell_1)} < 1$ and $M_{ij}^{(\ell_2)} < 1$.

Now for any $\epsilon > 0$, define $\widetilde{M}_\epsilon^{(\ell_1)}$ and $\widetilde{M}_\epsilon^{(\ell_2)}$ to be matrices obtained by replacing the $(i, j)^{th}$ entries of the matrices $M^{(\ell_1)}$ and $M^{(\ell_2)}$ with $(M_{ij}^{(\ell_1)} + \epsilon)$ and $(M_{ij}^{(\ell_2)} - \epsilon)$ respectively. Now, conditions (b)–(d) in tandem imply that there exists some value $\epsilon > 0$ such that *all* of the following properties hold:

(i) $[M_\epsilon^{(\ell_1)}]_{ij} \in [0, 1]$,

(ii) $[M_\epsilon^{(\ell_2)}]_{ij} \in [0, 1]$, and

(iii) The relative ordering of the entries of $M^{(\ell_1)}$ is identical to the relative ordering of the entries of $M_\epsilon^{(\ell_1)}$; the relative ordering of the entries of $M^{(\ell_1)}$ is identical to the relative ordering of the entries of $M_\epsilon^{(\ell_1)}$.

Properties (i) and (ii) imply that $\widetilde{M}_\epsilon^{(\ell_1)} \in [0, 1]^{n \times d}$ and $\widetilde{M}_\epsilon^{(\ell_2)} \in [0, 1]^{n \times d}$. Combined with property (iii), we also have $\widetilde{M}_\epsilon^{(\ell_1)} \in \mathbb{C}_{\mathrm{PR}}(1)$ and $\widetilde{M}_\epsilon^{(\ell_2)} \in \mathbb{C}_{\mathrm{PR}}(1)$. Finally, from the construction of the matrices $\widetilde{M}_\epsilon^{(\ell_1)}$ and $\widetilde{M}_\epsilon^{(\ell_2)}$, it is easy to see the relation

$$M = \widetilde{M}_\epsilon^{(\ell_1)} + \widetilde{M}_\epsilon^{(\ell_2)} + \sum_{i \in [\overline{\rho}(M)] \setminus \{\ell_1, \ell_2\}} M^{(i)}.$$

This decomposition of M is a different, valid permutation-rank decomposition of M.

5. Discussion and future work

We posit that the conventional low-rank models for matrix completion and denoising are equivalent to "parametric" assumptions with undesirable implications. We propose a new permutation-rank approach and argue, by means of a conceptual discussion as well as theoretical guarantees, that this approach offers significant benefits at little additional cost. Our work also contributes to a growing body of literature (Shah, 2017, Part 1), (Shah et al., 2017; Shah and Wainwright, 2018; Shah et al., 2016, 2019; Heckel et al., 2016; Chatterjee and Mukherjee, 2019; Flammarion et al., 2016; Chen et al., 2017) on moving towards more flexible models based on permutations that provide robustness to model mismatches.

In this paper we considered a binary-valued observation matrix. More generally, one may have observations that are real valued and bounded, in which case, our results continue to hold. Specifically, if we have every entry of Y and M^* is bounded by a constant, and if the noise is assumed to be of zero mean, then the statistical results presented here continue to hold (with possibly different constant factors which may depend on the bounding constant). We note that the bounded-entries assumption holds in most practical rating schemes deployed today.

We studied a low non-negative rank framework inspired by its interpretation discussed in Appendix A. It will be of interest to extend this framework to a regular low-rank matrix completion problem setting. The permutation-rank version of this setting would be identical to that defined in this paper, but where the constituent rank one components are not restricted to have entries $[0, 1]$. In terms of computational aspects of this model, we are optimistic for the following reason. Presently, given any matrix, we do not know how to compute its permutation rank (both non-negative and conventional) in a computationally efficient manner. However, the (classical) rank of a matrix is easily computable efficiently, whereas the non-negative rank is known to be computationally hard. We hope that the relative ease of computing the classical rank may carry over to the permutation-rank setting. In terms of statistical aspects, we expect some of our present results to carry over to the permutation-rank version of regular low-rank matrix completion, including Theorem 1(b) and Theorem 2 with some modifications. However, at this point we do not know if the proof of the information-theoretic upper bound of Theorem 1(a) can be readily translated (with or without any additional assumptions), and we leave this question open for future work.

Our work gives rise to some useful open problems that we hope to address in future work. In this paper, we established benefits of the permutation-based approach for the matrix completion problem under the random design observation setting. In the literature, the classical low (non-negative) rank matrix completion problem is more recently also studied under other observation models such as weighted random sampling (Negahban and Wainwright, 2012), fixed design (Jain et al., 2013; Klopp, 2014), streaming/active learning (Yun et al., 2015; Jin et al., 2016; Balcan and Zhang, 2016), or biased observation models (Hsieh et al., 2015), which are also of interest in the context of permutation-rank matrix completion. A second open problem is to close the gap between the statistically optimal minimax rate of estimation and the best known rate for polynomial-time computable algorithms for the permutation-rank model. Any solution to this problem may also contribute to the understanding of some other open problems in the literature (e.g., see Shah et al., 2017; Flammarion et al., 2016; Shah et al., 2016) on the gap between the statistical and computational aspects of estimation under an unknown permutation. Finally, we hope that the insights and techniques for flexible models we developed can be useful in problems of broader interest in the domain of "learning from people", to handle issues like miscalibration (Wang and Shah, 2019), subjectivity (Noothigattu et al., 2018), biases (Stelmakh et al., 2019), noise (Stelmakh et al., 2018), and strategic behavior (Xu et al., 2019), where models need to be flexible in order to accommodate complex human behavior.

Acknowledgments

The work of NBS was partially supported by NSF grants 1755656 and 1763734. The work of MJW and NBS was partially supported by DOD Advanced Research Projects Agency W911NF-16-1-0552 and National Science Foundation grant NSF-DMS-1612948. The work of SB was supported by NSF grants DMS-1713003 and CIF-1763734.

References

Sanjeev Arora, Rong Ge, Ravindran Kannan, and Ankur Moitra. Computing a nonnegative matrix factorization–provably. In *Proceedings of the forty-fourth annual ACM symposium on Theory of computing*, pages 145–162. ACM, 2012.

Maria-Florina F Balcan and Hongyang Zhang. Noise-tolerant life-long matrix completion via adaptive sampling. In *Advances In Neural Information Processing Systems*, 2016.

Jian-Feng Cai and Stanley Osher. Fast singular value thresholding without singular value decomposition. *UCLA CAM Report*, 5, 2010.

Jian-Feng Cai, Emmanuel J Candès, and Zuowei Shen. A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization*, 20(4):1956–1982, 2010.

Emmanuel J. Candes and Benjamin Recht. Exact matrix completion via convex optimization. *Found. Comput. Math.*, 9(6):717–772, December 2009.

Emmanuel J. Candès and Terence Tao. The power of convex relaxation: Near-optimal matrix completion. *IEEE Trans. Inf. Theor.*, 56(5):2053–2080, May 2010.

Sabyasachi Chatterjee and Sumit Mukherjee. Estimation in tournaments and graphs under monotonicity constraints. *IEEE Transactions on Information Theory*, 2019.

Sourav Chatterjee. Matrix estimation by universal singular value thresholding. *The Annals of Statistics*, 43(1):177–214, 2014.

Xi Chen, Sivakanth Gopi, Jieming Mao, and Jon Schneider. Competitive analysis of the top-k ranking problem. In *ACM-SIAM Symposium on Discrete Algorithms*, 2017.

Yudong Chen, Ali Jalali, Sujay Sanghavi, and Constantine Caramanis. Low-rank matrix recovery from errors and erasures. *IEEE Trans. Information Theory*, 59(7):4324–4337, 2013.

Thomas Cover and Joy Thomas. *Elements of information theory*. John Wiley & Sons, 2012.

Mark A Davenport and Justin Romberg. An overview of low-rank matrix recovery from incomplete observations. *IEEE Journal of Selected Topics in Signal Processing*, 10(4): 608–622, 2016.

David Donoho and Victoria Stodden. When does non-negative matrix factorization give a correct decomposition into parts? In *Advances in neural information processing systems*, 2003.

David Donoho, Matan Gavish, et al. Minimax risk of matrix denoising by singular value thresholding. *The Annals of Statistics*, 42(6):2413–2440, 2014.

Nicolas Flammarion, Cheng Mao, and Philippe Rigollet. Optimal rates of statistical seriation. *arxiv:1607.02435*, 2016.

Edgar N Gilbert. A comparison of signalling alphabets. *Bell System Technical Journal*, 31 (3):504–522, 1952.

Nicolas Gillis. Sparse and unique nonnegative matrix factorization through data preprocessing. *Journal of Machine Learning Research*, 13(Nov):3349–3386, 2012.

Nicolas Gillis. The why and how of nonnegative matrix factorization. *Regularization, Optimization, Kernels, and Support Vector Machines*, 12(257), 2014.

Cyril Gobinet, Eric Perrin, and Régis Huez. Application of non-negative matrix factorization to fluorescence spectroscopy. In *European Signal Processing Conference*, 2004.

David Gross. Recovering low-rank matrices from few coefficients in any basis. *IEEE Trans. Information Theory*, 57(3):1548–1566, 2011.

Johan Håstad. Tensor rank is np-complete. *Journal of Algorithms*, 11(4):644–654, 1990.

Reinhard Heckel, Nihar B Shah, Kannan Ramchandran, and Martin J Wainwright. Active ranking from pairwise comparisons and when parametric assumptions don't help. *arxiv:1606.08842*, 2016.

Frank L Hitchcock. The expression of a tensor or a polyadic as a sum of products. *Studies in Applied Mathematics*, 6(1-4):164–189, 1927.

Cho-Jui Hsieh, Nagarajan Natarajan, and Inderjit Dhillon. PU learning for matrix completion. In *International Conference on Machine Learning*, pages 2445–2453, 2015.

Daniel Hsu and Sham M Kakade. Learning mixtures of spherical gaussians: moment methods and spectral decompositions. In *ACM Innovations in Theoretical Computer Science*, 2013.

Prateek Jain, Praneeth Netrapalli, and Sujay Sanghavi. Low-rank matrix completion using alternating minimization. In *ACM symposium on Theory of computing*, 2013.

Chi Jin, Sham M Kakade, and Praneeth Netrapalli. Provable efficient online matrix completion via non-convex stochastic gradient descent. In *Advances in Neural Information Processing Systems*, 2016.

Raghunandan H Keshavan, Andrea Montanari, and Sewoong Oh. Matrix completion from noisy entries. *Journal of Machine Learning Research*, 11(Jul):2057–2078, 2010.

Thierry Klein and Emmanuel Rio. Concentration around the mean for maxima of empirical processes. *The Annals of Probability*, 33(3):1060–1077, 2005.

Olga Klopp. Noisy low-rank matrix completion with general sampling distribution. *Bernoulli*, 20(1):282–303, 2014.

Tamara G Kolda and Brett W Bader. Tensor decompositions and applications. *SIAM review*, 51(3):455–500, 2009.

Vladimir Koltchinskii, Karim Lounici, and Alexandre B Tsybakov. Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *The Annals of Statistics*, pages 2302–2329, 2011.

Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8), 2009.

Joseph B Kruskal. Three-way arrays: rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics. *Linear algebra and its applications*, 18(2):95–138, 1977.

Hans Laurberg, Mads Græsbøll Christensen, Mark D Plumbley, Lars Kai Hansen, and Søren Holdt Jensen. Theorems on positive data: On the uniqueness of NMF. *Computational intelligence and neuroscience*, 2008, 2008.

M. Laurent. Matrix completion problems. In *The Encyclopedia of Optimization*, pages 221—229. Kluwer Academic, 2001.

Daniel D Lee and H Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

Sahand Negahban and Martin J Wainwright. Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. *Journal of Machine Learning Research*, 13(May):1665–1697, 2012.

Ritesh Noothigattu, Nihar Shah, and Ariel Procaccia. Choosing how to choose papers. *arXiv preprint arxiv:1808.09057*, 2018.

Ashwin Pananjady, Martin J Wainwright, and Thomas A Courtade. Linear regression with an unknown permutation: Statistical and computational limits. *arxiv:1608.02902*, 2016.

Benjamin Recht. A simpler approach to matrix completion. *Journal of Machine Learning Research*, 12(Dec):3413–3430, 2011.

Ryan Sarver and Anssi Klapuri. Application of nonnegative matrix factorization to signal-adaptive audio effects. In *Proc. DAFx*, pages 249–252, 2011.

Peter H Schönemann. On two-sided orthogonal procrustes problems. *Psychometrika*, 33(1): 19–33, 1968.

Nihar Shah. *Learning From People*. PhD thesis, EECS Department, University of California, Berkeley, Jul 2017.

Nihar B Shah and Martin J Wainwright. Simple, robust and optimal ranking from pairwise comparisons. *Journal of machine learning research*, 2018.

Nihar B Shah, Sivaraman Balakrishnan, and Martin J Wainwright. A permutation-based model for crowd labeling: Optimal estimation and robustness. *arxiv:1606.09632*, 2016.

Nihar B Shah, Sivaraman Balakrishnan, Adityanand Guntuboyina, and Martin J Wainwright. Stochastically transitive models for pairwise comparisons: Statistical and computational issues. *IEEE Transactions on Information Theory*, 2017.

Nihar B Shah, Sivaraman Balakrishnan, and Martin J Wainwright. Feeling the Bern: Adaptive estimators for Bernoulli probabilities of pairwise comparisons. *IEEE Transactions on Information Theory*, 2019.

Nathan Srebro, Noga Alon, and Tommi S Jaakkola. Generalization error bounds for collaborative prediction with low-rank matrices. In *Advances In Neural Information Processing Systems*, 2005.

Ivan Stelmakh, Nihar Shah, and Aarti Singh. PeerReview4All: Fair and accurate reviewer assignment in peer review. *arXiv preprint arxiv:1806.06237*, 2018.

Ivan Stelmakh, Nihar Shah, and Aarti Singh. On testing for biases in peer review. In *ACM EC Workshop on Mechanism Design for Social Good*, 2019.

Terence Tao. *Topics in random matrix theory*, volume 132. American Mathematical Society Providence, RI, 2012.

Fabian J Theis, Kurt Stadlthanner, and Toshihisa Tanaka. First results on uniqueness of sparse non-negative matrix factorization. In *European Signal Processing Conference*, 2005.

A.B. Tsybakov. *Introduction to Nonparametric Estimation.* Springer Series in Statistics. 2008.

Antoine Vandendorpe, Ngoc-Diep Ho, Steven Vanduffel, and Paul Van Dooren. On the parameterization of the creditrisk+ model for estimating credit portfolio risk. *Insurance: Mathematics and Economics*, 42(2):736–745, 2008.

RR Varshamov. Estimate of the number of signals in error correcting codes. In *Dokl. Akad. Nauk SSSR*, 1957.

Jingyan Wang and Nihar B Shah. Your 2 is my 1, your 3 is my 9: Handling arbitrary miscalibrations in ratings. In *AAMAS*, 2019.

Yichong Xu, Han Zhao, Xiaofei Shi, and Nihar Shah. On strategyproof conference review. In *IJCAI*, 2019.

Se-Young Yun, Marc Lelarge, and Alexandre Proutiere. Streaming, memory limited matrix completion with noise. *arxiv:1504.03156*, 2015.

Appendix

Appendix A. Alternative interpretation of the non-negative rank model

In the non-negative rank model described in the introduction, one may wonder why the affinity of a user to a movie conditioned on a feature must be modeled as the product $u_i^\ell v_j^\ell$ of the separate connections of the user and movie to the feature. Secondly, one may also wonder why the net affinity of a user to a movie is the sum of the affinities across the features $\sum_{\ell=1}^r u_i^\ell v_j^\ell$. These two modeling assumptions may sometimes be confusing, and hence in what follows, we present an alternative interpretation of the low non-negative rank model for the recommender systems application.

Consider any feature $\ell \in [r]$. The affinities of users towards movies conditioned on this feature is a matrix, say $X^{(\ell)} \in [0,1]^{n \times d}$. The matrix $X^{(\ell)}$ is assumed to have a (non-negative) rank of 1. Hence the probability that user i likes movie j, when asked to judge only based on feature ℓ, equals $X_{ij}^{(\ell)}$.

Now, every user is assumed to have their own way of weighing features to decide which movies they like. Specifically, any user $i \in [n]$ is associated to values $\alpha_i^{(1)}, \ldots, \alpha_i^{(r)}$ such that $\alpha_i^{(\ell)} \geq 0$ for every $\ell \in [r]$ and $\sum_{\ell=1}^r \alpha_i^{(\ell)} = 1$. The probability that user i likes any movie j is assumed to be the convex combination

$$\sum_{\ell=1}^{r} \alpha_i^{(\ell)} X_{ij}^{(\ell)}.$$

This completes the description of the model.

Let us verify that the resulting user-movie matrix has a non-negative rank of r. Recall the assumption that $X^{(\ell)}$ has a non-negative rank of 1, and let $X^{(\ell)} = u^\ell (v^\ell)^T$ for some vectors u^ℓ and v^ℓ. Then the i^{th} row of the overall user-movie matrix equals $\sum_{\ell=1}^r \alpha_i^{(\ell)} u_i^\ell (v^\ell)^T$, and hence the overall user-movie matrix equals

$$\sum_{\ell=1}^{r} \widetilde{u}^\ell (v^\ell)^T, \qquad \text{where} \qquad \widetilde{u}^\ell = \begin{bmatrix} \alpha_1^{(\ell)} u_1^\ell \\ \vdots \\ \alpha_n^{(\ell)} u_n^\ell \end{bmatrix}.$$

This completes the alternative description of the non-negative rank model.

One can observe that the restriction $\sum_{\ell=1}^r \alpha_i^{(\ell)} = 1$ makes this model slightly more restrictive than the non-negative rank model described earlier in the main text. However, all of our results on estimation for the non-negative rank model described in Section 2.1 continue to apply to this model as well.

Appendix B. Intuitive algorithms that provably fail

In this section, we present two intuitive polynomial-time computable algorithms for the permutation-rank setting—one for estimating M^* from Y, and one for decomposing M^* into its constituent permutation-rank-one matrices—and show that these algorithms provably fail. Our goal in describing these negative results is as a complement to the positive results provided in the main text, and with the hope that the points of failure of these algorithms may form starting points for subsequent research.

B.1. An intuitive polynomial-time estimator for M^* from Y

In this section, we consider the problem of estimating the matrix M^* from noisy and partial observations Y as defined earlier in equation (1). For simplicity, we assume that $p_{\text{obs}} = 1$. For any vector $z \in \mathbb{R}^m$, we let vector $z_+ \in \mathbb{R}^m$ with entries $[z_+]_i = \max\{z_i, 0\}$ represent the positive component of z, and vector $z_- \in \mathbb{R}^m$ with entries $[z_-]_i = \min\{z_i, 0\}$ represent the negative component of z.

We first provide some intuition and background to motivate the estimator we study in this section, and then present a formal definition. Denote the permutation-rank of matrix M^* as $\rho^* := \overline{\rho}(M^*)$, and assume that the value of ρ^* is known. The goal is to obtain an estimate $\widehat{M} \in \mathbb{C}_{\text{PR}}(\rho^*)$ from the observed matrix Y such that the error $\frac{1}{nd}|\!|\!|M^* - \widehat{M}|\!|\!|_{\text{F}}^2$ is as small as possible. For any such matrix $\widehat{M}$, let us use the following notation for its permutation-rank decomposition: $\widehat{M} = \sum_{\ell=1}^{\rho^*} \widehat{M}^{(\ell)}$ where $\widehat{M}^{(\ell)} \in \mathbb{C}_{\text{PR}}(1)$ for every $\ell \in [\rho^*]$. Further, we let $\widehat{\pi}^{(\ell)}$ and $\widehat{\sigma}^{(\ell)}$ respectively denote the permutation of the rows and columns of $\widehat{M}^{(\ell)}$.

Past literature on computationally-efficient estimation for such problems provides us with estimators with the following two distinct goals: (i) to estimate the permutations of the constituent matrices in the permutation-rank decomposition, and (ii) estimators to compute the entries of the constituent matrices given the permutations. For each of these two goals, we describe a natural estimator below from past works.

Estimating the permutations via singular value decomposition: Compute the singular value decomposition $Y = \sum_{\ell=1}^{\min\{n,d\}} a^{(\ell)}[b^{(\ell)}]^T$ such that the vectors $\{a^{(1)}, \ldots, a^{(\min\{n,d\})}\}$ are mutually orthogonal, the vectors $\{b^{(1)}, \ldots, b^{(\min\{n,d\})}\}$ are mutually orthogonal, and $[a^{(1)}]^T b^{(1)} \geq \ldots \geq [a^{(\min\{n,d\})}]^T b^{(\min\{n,d\})}$. In order to resolve a global sign ambiguity, we also mandate the condition $\|a_+^{(\ell)}\|_2 \geq \|a_-^{(\ell)}\|_2$ for every $\ell \in [\min\{n,d\}]$. Finally, for each $\ell \in [\rho^*]$, set $\widehat{\pi}^{(\ell)}$ and $\widehat{\sigma}^{(\ell)}$ as the ordering of the entries of $a^{(\ell)}$ and $b^{(\ell)}$ respectively.

From past works (Shah et al., 2016), this estimator for the permutations is known to possess appealing properties for the case when $\rho^* = 1$. For instance, it is not hard to see that in a noiseless setting where $Y = M^*$, the estimator will yield exactly the row and column permutations of $M^* \in \mathbb{C}_{\text{PR}}(1)$. This fact is employed in Shah et al. (2016) to obtain consistent estimates of the permutations associated to an unknown matrix in $\mathbb{C}_{\text{PR}}(1)$ in the context of a "crowd labeling" problem. It is also not hard to verify that the estimator is can be computed in a computationally-efficient manner.

Estimating the entries via least squares, when given the permutations: Given some estimate $\widehat{\pi}^{(1)}, \widehat{\sigma}^{(1)}, \ldots, \widehat{\pi}^{(\rho^*)}, \widehat{\sigma}^{(\rho^*)}$ of the permutations associated to the permutation-rank decomposition of M^*, the following estimator $\widehat{M}$ provides an estimate of the matrix M^* as well as the matrices in its permutation-rank decomposition.

$$\widehat{M} \in \underset{M \in [0,1]^{n \times d}}{\arg\min} \; \|\!|Y - M|\!\|_{\mathrm{F}}^2 \tag{46}$$

$$\text{such that} \;\; M = \sum_{\ell=1}^{\rho^*} \widetilde{M}^{(\ell)}, \; \text{and}$$

$$\widetilde{M}^{(\ell)} \in \mathbb{C}_{\mathrm{PR}}(1) \text{ with rows and columns ordered by } (\widehat{\pi}^{(\ell)}, \widehat{\sigma}^{(\ell)}), \text{ for every } \ell \in [\rho^*].$$

The aforementioned estimator $\widehat{M}$ is a natural extension of the least-squares estimators studied in past works (Shah et al., 2017, 2019) for the case when $M^* \in \mathbb{C}_{\mathrm{PR}}(1)$. The estimator is known to have appealing properties from both the statistical and computational perspectives. From a computational standpoint, all of the constraints in the optimization program (46) can be expressed as a set of (polynomial number of) linear inequalities, thereby making the optimization problem computationally tractable. From a statistical standpoint, if the given permutations $(\widehat{\pi}^{(1)}, \widehat{\sigma}^{(1)}, \ldots, \widehat{\pi}^{(\rho^*)}, \widehat{\sigma}^{(\rho^*)})$ are exactly (or approximately) equal to the permutations associated to a permutation-rank decomposition of M^*, the proofs of the results in Shah et al. (2017, 2019) as well as Theorem 1 in the present paper imply that the estimator $\widehat{M}$ is minimax optimal for estimating M^* from Y. The estimator $\widehat{M}$ continues to remain statistically efficient if the permutations are known up to a reasonable approximation.

Given the two intuitive estimators discussed above, a natural means to estimate M^* from Y is to concatenate these two estimators to obtain the following two-step estimator:
Step 1: From Y, obtain an estimate $(\widehat{\pi}^{(1)}, \widehat{\sigma}^{(1)}, \ldots, \widehat{\pi}^{(\rho^*)}, \widehat{\sigma}^{(\rho^*)})$ of the permutations of the decomposition of M^* via the singular value decomposition-based estimator described above.
Step 2: Using the estimates of the permutations, obtain an estimate $\widehat{M}$ of M^* via the least squares projection (46).

We believe that when $\rho^* = 1$, this estimator is not only consistent, but it has an expected error decaying at the rate $\mathcal{O}(\min\{n, d\}^{-1/2})$. We now show that in fact as soon as one moves to the setting of $\rho^* > 1$, this estimator is no longer even consistent—even if there is no noise.

Proposition 19 *There exists a matrix $M^* \in \mathbb{C}_{PR}(2)$ such that when $Y = M^*$, the two-step estimator $\widehat{M}$ has an error lower bounded as*

$$\frac{1}{nd} \|\!|M^* - \widehat{M}|\!\|_F^2 \geq c_2,$$

with probability 1.

The proof of this result is provided in Appendix B.3.1. The proof also demonstrates an identical negative result for the following modified estimation algorithm: In computing

$\widehat{M}$ as above, instead of taking only the permutations of the top ρ^* singular vectors, collect permutations from singular vectors until you obtain ρ^* *distinct* permutations; then apply the least squares projection step to these ρ^* distinct permutations.

B.2. An intuitive greedy algorithm for permutation-rank decomposition

Consider any matrix any matrix $M \in [0,1]^{n \times d}$. The singular value decomposition of M into components having a (conventional) rank of one can be performed with the following greedy algorithm:

- Let $\widehat{k} = 1$
- While $M \neq \sum_{\ell=1}^{\widehat{k}-1} \widehat{M}^{(\ell)}$:
  - Let $\widehat{M}^{(\widehat{k})} \in \underset{\substack{M'=ab^T \\ (a,b)\in\mathbb{R}^n\times\mathbb{R}^d}}{\arg\min} \; |\!|\!| M - \sum_{\ell=1}^{\widehat{k}-1} \widehat{M}^{(\ell)} - M' |\!|\!|_F$
  - $\widehat{k} = \widehat{k} + 1$
- Output $\widehat{k}$ as the rank of M and $\{\widehat{M}^{(1)}, \ldots, \widehat{M}^{(\widehat{k})}\}$ as its singular value decomposition.

An obvious question that arises is whether a similar greedy algorithm works to obtain a permutation-rank decomposition.

To this end, consider any value $q \geq 1$, and for any matrix M, let $|\!|\!| M |\!|\!|_q$ denote its entry-wise norm $|\!|\!| M |\!|\!|_q := \big(\sum_{i,j} (M_{ij})^q \big)^{\frac{1}{q}}$. Then the natural analogue of the aforedescribed algorithm in the context of permutation-rank decomposition is as follows:

- Let $\widehat{\rho} = 1$
- While $M \neq \sum_{\ell=1}^{\widehat{\rho}-1} \widehat{M}^{(\ell)}$:
  - Let $\widehat{M}^{(\widehat{\rho})} \in \underset{M' \in \mathbb{C}_{\mathrm{PR}}(1)}{\arg\min} \; |\!|\!| M - \sum_{\ell=1}^{\widehat{\rho}-1} \widehat{M}^{(\ell)} - M' |\!|\!|_q$
  - $\widehat{\rho} = \widehat{\rho} + 1$
- Output $\widehat{\rho}$ as the permutation-rank of M and $\{\widehat{M}^{(1)}, \ldots, \widehat{M}^{(\widehat{\rho})}\}$ as its permutation-decomposition

The following proposition investigates whether such an algorithm will work.

Proposition 20 *For any values of n, d and $\rho \geq 2$, there exists an $(n \times d)$ matrix $M \in \mathbb{C}_{PR}(\rho)$ such that the above algorithm outputs a decomposition of permutation-rank at least $(\rho + 1)$.*

The guaranteed incorrectness of the permutation rank of the output of the algorithm also directly implies that the decomposition is also incorrect.

B.3. Proofs

We now present the proofs of the negative results introduced in this section.

B.3.1. Proof of Proposition 19

In what follows, for clarity of exposition, we ignore issues pertaining to floors and ceilings of numbers, as they affect the results only by a constant factor.

We begin by defining a matrix $M^* \in \mathbb{C}_{\text{PR}}(2)$ as

$$M^* = M^{(1)} + M^{(2)},$$

with

$$M^{(1)} = a^{(1)}(b^{(1)})^T + a^{(2)}(b^{(2)})^T \quad \text{and} \quad M^{(2)} = a^{(3)}(b^{(3)})^T.$$

Set

$$\begin{array}{lcllll} a^{(1)} & = & [1 & .9 \cdots .9 & .8 \cdots .8 & 0 \cdots 0 \quad]^T \\ a^{(2)} & = & [0 & .2 \cdots .2 & -.1 \cdots -.1 & 0 \cdots 0 \quad]^T \\ a^{(3)} & = & [0 & \underbrace{0 \cdots 0}_{\nu_1(n-1)} & \underbrace{0 \cdots 0}_{\nu_2(n-1)} & \underbrace{1 \cdots 1}_{\nu_3(n-1)} \quad]^T, \end{array}$$

and

$$\begin{array}{lcllll} b^{(1)} & = & [1 & .9 \cdots .9 & .8 \cdots .8 & 0 \cdots 0 \quad]^T \\ b^{(2)} & = & [0 & .2 \cdots .2 & -.1 \cdots -.1 & 0 \cdots 0 \quad]^T \\ b^{(3)} & = & [0 & \underbrace{0 \cdots 0}_{\nu_1(d-1)} & \underbrace{0 \cdots 0}_{\nu_2(d-1)} & \underbrace{1 \cdots 1}_{\nu_3(d-1)} \quad]^T, \end{array}$$

where $\nu_1 = .684$, $\nu_2 = .304$, and $\nu_3 = .012$. It is easy to verify that all entries of the matrices $M^*, M^{(1)}, M^{(2)}$ lie in the interval $[0,1]$ and that $M^{(1)} \in \mathbb{C}_{\text{PR}}(1)$ and $M^{(2)} \in \mathbb{C}_{\text{PR}}(1)$ and $M^* \in \mathbb{C}_{\text{PR}}(2) \backslash \mathbb{C}_{\text{PR}}(1)$.

One can further verify the following properties of this construction:

1. $\langle a^{(\ell)}, a^{(\ell')} \rangle = 0$ and $\langle b^{(\ell)}, b^{(\ell')} \rangle = 0$ for every $\ell \neq \ell' \in \{1,2,3\}$.
2. $\|a^{(1)}\|_2 > \|a^{(2)}\|_2 > \|a^{(3)}\|_2$ and $\|b^{(1)}\|_2 > \|b^{(2)}\|_2 > \|b^{(3)}\|_2$.
3. The (conventional) rank of M^* is 3.
4. $a^{(1)}$, $a^{(2)}$ and $a^{(3)}$ have different permutations of their entries; likewise, $b^{(1)}$, $b^{(2)}$ and $b^{(3)}$ have different permutations of their entries.
5. $\|a_+^{(\ell)}\|_2 \geq \|a_-^{(\ell)}\|_2$ and $\|b_+^{(\ell)}\|_2 \geq \|b_-^{(\ell)}\|_2$ for every $\ell \in [3]$.

The five properties listed above imply that the following decomposition of M^*,

$$M^* = \sum_{\ell=1}^{3} a^{(\ell)}(b^{(\ell)})^T,$$

is a valid singular value decomposition with the global signs of the constituent vectors satisfying the conditions of Step 1 of the algorithm. Consequently, the $\rho^* = 2$ estimated permutations in Step 1 of the algorithm are those given by the respective orderings of the entries of the vectors $\{a^{(1)}, b^{(1)}\}$ and $\{a^{(2)}, b^{(2)}\}$.

Observe that the entries 2 to $(1+\nu_1(n-1))$ of both $a^{(1)}$ and $a^{(2)}$ have values higher than the last $\nu_3(n-1)$ entries of these vectors, and hence this ordering is reflected in the respective permutations derived from these vectors. Likewise, the entries 2 through $(1+\nu_1(d-1))$ are ranked higher than the last $\nu_3(d-1)$ entries in the permutation derived from the vectors $b^{(1)}$ and $b^{(2)}$. Due to this collection of inequalities, the least squares program in Step 2 of the algorithm must mandate that

$$M_{ij} \geq M_{i'j'} \geq M_{i''j''}, \tag{47}$$

whenever $2 \leq i, i' \leq 1 + \nu_1(n-1)$; $n - \nu_3(n-1) < i'' \leq n$; $2 \leq j \leq 1 + \nu_1(d-1)$; and $d - \nu_3(d-1) < j', j'' \leq d$. However, for each coordinate in this range, we also have $M^*_{ij} = .85, M^*_{i'j'} = 0, M^*_{i''j''} = 1$. Consequently, any triplet of values $(M_{ij}, M_{i'j'}, M_{i''j''})$ that follows the ordering (47) must necessarily incur an error lower bounded as

$$(M_{ij} - M^*_{ij})^2 + (M_{i'j'} - M^*_{i'j'})^2 + (M_{i''j''} - M^*_{i''j''})^2 \geq c,$$

for some universal constant $c > 0$. Summing up the errors over all the entries of the matrix in the aforementioned coordinate set yields that any matrix M satisfying the constraints of the least squares problem must have squared Frobenius error at least $|\!|\!|M - M^*|\!|\!|_{\text{F}}^2 \geq c'nd$, for some universal constant $c' > 0$.

B.3.2. Proof of Proposition 20

First let $n = d = \rho = 2$. Consider the (2×2) matrix M defined as

$$M := \begin{bmatrix} 0 & .6 \\ .6 & .4 \end{bmatrix}.$$

It is easy to verify that $M \in \mathbb{C}_{\text{PR}}(2) \backslash \mathbb{C}_{\text{PR}}(1)$.

Let us now investigate the operation of the proposed algorithm on this matrix M. The following lemma controls the first step of the algorithm.

Lemma 21 *When the input matrix M is as defined above, the algorithm will select*

$$\widehat{M}^{(1)} = \begin{bmatrix} 0 & .4 \\ .4 & .4 \end{bmatrix}$$

in the first iteration.

As a consequence of this lemma, we have the following residual that is used for the subsequent iterations of the algorithm:

$$M - \widehat{M}^{(1)} = \begin{bmatrix} 0 & .2 \\ .2 & 0 \end{bmatrix}.$$

It is easy to see the that the residual matrix $(M - \widehat{M}^{(1)}) \in \mathbb{C}_{\text{PR}}(2) \backslash \mathbb{C}_{\text{PR}}(1)$. Also observe that in each iteration, the algorithm subtracts out a matrix in $\mathbb{C}_{\text{PR}}(1)$ from the residual.

Consequently, the algorithm will require at least two more iterations to terminate. The algorithm thus necessarily outputs a decomposition with $\widehat{\rho} \geq 3$, as claimed.

Next we extend these arguments to any arbitrary values of $n \geq 2, d \geq 2, \rho \geq 2$. Consider matrix M with entries:

- $M_{11} = 0$, $M_{12} = M_{21} = .6$, $M_{22} = .4$
- $M_{ii} = 1$ for every $i \in \{3, \ldots, \rho\}$
- $M_{ij} = 0$ for every other coordinate (i, j).

The matrix M has a block-diagonal structure with the top-left (2×2) block as one non-zero component and $(\rho - 2)$ other entries on the diagonal as $(\rho - 2)$ additional non-zero components.

The rest of the proof is partitioned into two cases:
Case I: Suppose that at some step $\widehat{\rho}$ of the algorithm, some entry of the residual matrix $(M - \sum_{\ell=1}^{\widehat{\rho}} \widehat{M}^{(\ell)}$ is strictly negative. Then the algorithm will never terminate because every subsequent candidate matrix in the minimization step of the algorithm must lie in $\mathbb{C}_{\text{PR}}(1)$ and hence must have non-negative entries. The algorithm will thus output $\widehat{\rho} = \infty$.

Case II: Now suppose that the residual matrices always have non-negative entires. Then given the block-diagonal structure of M, any matrix $\widehat{M}^{(1)}, \widehat{M}^{(2)}, \ldots$ in the iterations of the algorithm can be non-zero in exactly one of these diagonal components. As a result, the overall decomposition yielded by the algorithm decouples into ρ individual decompositions of the ρ respective blocks, each of which will contribute a permutation-rank of at least 1. Moreover, from the arguments for the case of $n = d = 2$ above, we also have that the top-left (2×2) block will induce a decomposition of permutation-rank 3 in the algorithm. Putting the pieces together, we see that the matrix M will induce the proposed algorithm to output a decomposition of permutation-rank at least $(\rho + 1)$, whereas $\overline{\rho}(M) = \rho$.

Proof of Lemma 21 Since $M_{11} = 0$, we must have $\widehat{M}^{(1)}_{11} = 0$. Now suppose the column ordering of $\widehat{M}^{(1)}$ is such that the first column is greater than the second column. Then we must have $\widehat{M}^{(1)}_{12} = 0$. Since $M^{(1)}$ is the minimizer of the optimization program in the algorithm we must then have $\widehat{M}^{(1)}_{21} = M_{21}$ and $\widehat{M}^{(1)}_{22} = M_{22}$, and consequently $\|M - \widehat{M}^{(1)}\|_q = \widehat{M}^{(1)}_{12} = .6$. An analogous argument holds if the first row is greater than the second row in the permutation of $\widehat{M}^{(1)}$. Finally, suppose that in the permutations of $M^{(1)}$, the second column is greater than the first column and the second row is greater than the first row. Then we must have $.4 \geq \widehat{M}^{(1)}_{22} \geq \max\{\widehat{M}^{(1)}_{12}, \widehat{M}^{(1)}_{21}\}$. With this condition, one can see that the minimizer of the optimization program is $\widehat{M}^{(1)}_{12} = \widehat{M}^{(1)}_{21} = \widehat{M}^{(1)}_{22} = .4$. Consequently, we have $\|M - \widehat{M}^{(1)}\|_q = .2 \times 2^{\frac{1}{q}} < .6$ for any $q \geq 1$. Thus the algorithm chooses

$$\widehat{M}^{(1)} = \begin{bmatrix} 0 & .4 \\ .4 & .4 \end{bmatrix}.$$