Matrix Completion with Model-free Weighting

Jiayi Wang[1], Raymond K. W. Wong[1], Xiaojun Mao[2], and Kwun Chuen Gary Chan[3]

[1]*Department of Statistics, Texas A&M University*
[2]*School of Data Science, Fudan University*
[3]*Department of Biostatistics, University of Washington*

Abstract

In this paper, we propose a novel method for matrix completion under general non-uniform missing structures. By controlling an upper bound of a novel balancing error, we construct weights that can actively adjust for the non-uniformity in the empirical risk without explicitly modeling the observation probabilities, and can be computed efficiently via convex optimization. The recovered matrix based on the proposed weighted empirical risk enjoys appealing theoretical guarantees. In particular, the proposed method achieves stronger guarantee than existing work in terms of the scaling with respect to the observation probabilities, under asymptotically heterogeneous missing settings (where entry-wise observation probabilities can be of different orders). These settings can be regarded as a better theoretical model of missing patterns with highly varying probabilities. We also provide a new minimax lower bound under a class of heterogeneous settings. Numerical experiments are also provided to demonstrate the effectiveness of the proposed method.

1 Introduction

Matrix completion is a modern missing data problem where the object of interest is a high-dimensional and often low-rank matrix. In its simplest form, a partial (noisy) observation of the target matrix is collected, and the goal is to impute the missing entries and sometimes also to de-noise the observed ones. There are various related applications in, e.g., bioinformatics Chi et al. (2013), causal inference Athey et al. (2018); Kallus et al. (2018), collaborative filtering Rennie and Srebro (2005), computer vision Weinberger and Saul (2006), positioning Montanari and Oh (2010), survey imputation Davenport et al. (2014); Zhang et al. (2020); Sengupta et al. (2021) and quantum state tomography Wang (2013); Cai et al. (2016). Matrix completion has been popularized by the famous Netflix prize problem Bennett and Lanning (2007), in which a large matrix of movie ratings is partially observed. Each row of this matrix consists of ratings from a particular customer while each column records the ratings to a particular movie.

Matrix completion has attracted significant interest from the machine learning and statistics communities (e.g., Koltchinskii et al., 2011; Hernández-Lobato et al., 2014; Klopp, 2014;

Lafond et al., 2014; Hastie et al., 2015; Klopp et al., 2015; Bhaskar, 2016; Cai and Zhou, 2016; Kang et al., 2016; Zhu et al., 2016; Bi et al., 2017; Fithian and Mazumder, 2018; Dai et al., 2019; Robin et al., 2020; Chen et al., 2020). Although many statistical and computational breakthroughs (e.g., Candès and Recht, 2009; Koltchinskii et al., 2011; Recht, 2011) have been made in this area in the last decade, most work (with theoretical guarantees) is developed under a uniform missing structure where every entry is assumed to be observed with the same probability. However, uniform missingness is unrealistic in many applications.

The work under non-uniform missingness is relatively sparse, and can be roughly divided into two major classes. The first class (e.g., Srebro et al., 2005; Foygel and Srebro, 2011; Klopp, 2014; Cai and Zhou, 2016) focuses on a form of robustness result, and shows that without actively adjusting for the non-uniform missing structure (e.g., simply applying a uniform empirical risk function $\hat{R}_{\text{uni}}$ defined below), nuclear-norm and max-norm regularized methods can still lead to consistent estimations. Since no direct adjustment is imposed, there is no need to model the non-uniform missing structure. The second class aims to improve the estimation by modeling the missing structure and actively adjusting for non-uniformity. Several works (e.g., Srebro and Salakhutdinov, 2010; Foygel et al., 2011; Negahban and Wainwright, 2012; Mao et al., 2019) fall into this class. However, many of the underlying models can be viewed as special low-rank (e.g., rank 1) missing structures. For instance, a common model is the product sampling model (Negahban and Wainwright, 2012) where row and column are chosen independently according to possibly non-uniform marginal distributions, leading to a rank-1 matrix of observation probability. The specific model choices of non-uniformity restrict the applicability and theoretical guarantees of these works. One notable exception is Foygel et al. (2011), which actively adjusts for a product sampling model via a variant of weighted trace-norm regularization, but still provides guarantee under general missing structure. Despite these efforts, the study of non-uniform missing mechanisms is still far from comprehensive.

In this work, we propose a novel method of *balancing* weighting to actively adjust for the non-uniform empirical risk due to general unbalanced (i.e., non-uniform) sampling, *without* explicitly modeling the probabilities of observation. This is especially attractive when such model is hard to choose or estimate. We summarize our major contributions as follows.

First, we propose a novel balancing idea to adjust for the non-uniformity in matrix completion problems. Unlike many existing works, this idea does not require specific modeling of the observation probabilities. Thanks to the proposed relaxation of the balancing error (Lemma 1), the balancing weights can then be obtained via a constrained spectral norm minimization, which is a convex optimization problem.

Second, we provide theoretical guarantees on the balancing performance of the proposed weights, as well as the matrix recovery via the corresponding *weighted* empirical risk estimator. We note that the estimation nature of the balancing weights introduces non-trivial dependence in the weighted empirical risk, as opposed to the typical unweighted empirical risk (often assumed to be a sum of independent quantities). This leads to a non-standard analysis of the proposed matrix estimator.

Third, we investigate a new type of asymptotic regime — asymptotically heterogeneous missing structures. This regime allows observation probabilities to be of different orders, a more reasonable asymptotic model for the scenarios with highly varying probabilities among entries. Under asymptotically heterogeneous settings, we show that our estimator achieves

a significantly better error upper bound than existing upper bounds in terms of the scaling with respect to the observation probabilities. Such scaling is shown to be optimal via a new minimax result based on a class of asymptotically heterogeneous settings. Note that we focus on the challenging uniform error d^2 as opposed to the weighted (non-uniform) error $\tilde{d}^2$ (see Section 5), so as to ensure entries with high missing rate would be given non-neglible emphasis in our error measure.

2 Background

2.1 Notation

Throughout the paper, we use several matrix norms: nuclear norm $\|\cdot\|_*$, Frobenius norm $\|\cdot\|_F$, spectral norm $\|\cdot\|$ entry-wise maximum norm $\|\cdot\|_\infty$ and max norm $\|\cdot\|_{\max}$. Specifically, the *entry-wise* maximum norm of a matrix $\boldsymbol{B} = (B_{ij})$ is defined as $\|\boldsymbol{B}\|_\infty = \max_{i,j}|B_{ij}|$, while the max norm is defined as

$$\|\boldsymbol{B}\|_{\max} = \inf\{\|\boldsymbol{U}\|_{2,\infty}\|\boldsymbol{V}\|_{2,\infty} : \boldsymbol{B} = \boldsymbol{U}\boldsymbol{V}^\intercal\},$$

where $\|\cdot\|_{2,\infty}$ denotes the maximum ℓ_2-row-norm of a matrix. See, e.g., Srebro and Shraibman (2005) for the properties of max norm. The Frobenius inner product and Hadamard product between two matrices $\boldsymbol{B}_1 = (B_{1,ij})$ and $\boldsymbol{B}_2 = (B_{2,ij})$ of the same dimensions are represented by $\langle \boldsymbol{B}_1, \boldsymbol{B}_2\rangle = \sum_{i,j} B_{1,ij}B_{2,ij}$ and $\boldsymbol{B}_1 \circ \boldsymbol{B}_2 = (B_{1,ij}B_{2,ij})$ respectively. For any $a \in \mathbb{R}$ and any matrix $\boldsymbol{B} = (B_{ij})$, we write $\boldsymbol{B}^{\circ(a)} = (B_{ij}^a)$.

We also adopt the following asymptotic notations. Let $(b_n)_{n\geq 1}$ and $(c_n)_{n\geq 1}$ be two sequences of nonnegative numbers. We write $b_n = \mathcal{O}(c_n)$ if $b_n \leq Kc_n$ for some constant $K > 0$; and $b_n \asymp c_n$ if $b_n = \mathcal{O}(c_n)$ and $c_n = \mathcal{O}(b_n)$. In addition, we use polylog(n) to represent a polylogarithmic function of n, i.e., a polynomial in $\log n$. So $\mathcal{O}(\text{polylog}(n))$ represents a polylogarithmic order in n.

2.2 Setup

We aim to recover an unknown target matrix $\boldsymbol{A}_\star = (A_{\star,ij})_{i,j=1}^{n_1,n_2} \in \mathbb{R}^{n_1\times n_2}$ from partial observation of its noisy realization $\boldsymbol{Y} = (Y_{ij})_{i,j=1}^{n_1,n_2} \in \mathbb{R}^{n_1\times n_2}$. Denote the observation indicator matrix $\boldsymbol{T} = (T_{ij})_{i,j=1}^{n_1,n_2} \in \mathbb{R}^{n_1\times n_2}$, where $T_{ij} = 1$ if Y_{ij} is observed and $T_{ij} = 0$ otherwise. We consider an additive noise model

$$Y_{ij} = A_{\star,ij} + \epsilon_{ij}, \quad i = 1,\ldots,n_1; j = 1,\ldots,n_2,$$

where $\{\epsilon_{ij}\}$ are independent errors with zero mean, and are independent of $\{T_{ij}\}$. Also, $\{T_{ij}\}$ are independent Bernoulli random variables with $\pi_{ij} = \Pr(T_{ij} = 1)$. We write $\boldsymbol{\Pi} = (\pi_{ij})_{i,j=1}^{n_1,n_2}$.

2.3 Uniformity Versus Non-uniformity

Due to complexity of data, it is often undesirable to posit an additional distributional model for $\{\varepsilon_{ij}\}$ (such as normality) in practice. To recover $\boldsymbol{A}_\star$, an empirical risk minimization

framework is commonly adopted with the risk function:

$$R(\boldsymbol{A}) = \frac{1}{n_1 n_2} \mathsf{E}\left(\|\boldsymbol{Y} - \boldsymbol{A}\|_F^2\right), \quad \boldsymbol{A} \in \mathbb{R}^{n_1 \times n_2}.$$

Under uniform sampling (i.e., $\pi_{ij} \equiv \pi$), this motivates the use of the popular empirical risk

$$\widehat{R}_{\text{uni}}(\boldsymbol{A}) = \frac{1}{n_1 n_2} \|\boldsymbol{T} \circ (\boldsymbol{Y} - \boldsymbol{A})\|_F^2, \quad \boldsymbol{A} \in \mathbb{R}^{n_1 \times n_2},$$

which is unbiased for $\pi R(\boldsymbol{A})$ (e.g., Candès and Recht, 2009; Candès and Plan, 2010; Koltchinskii et al., 2011; Klopp, 2014). To minimize $\widehat{R}_{\text{uni}}$, we can ignore the constant multiplier π. In such settings, a popular form of estimator is $\arg\min_{\boldsymbol{A} \in \mathcal{A}_{n_1,n_2}} \widehat{R}_{\text{uni}}(\boldsymbol{A})$, where examples of the hypothesis class $\mathcal{A}_{n_1,n_2}$ include a set of matrices with rank at most r (i.e., $\{\boldsymbol{A} : \text{rank}(\boldsymbol{A}) \leq r\}$), and a nuclear norm ball of radius ν (i.e., $\{\boldsymbol{A} : \|\boldsymbol{A}\|_* \leq \nu\}$). In the latter case, one can also adopt an equivalent minimization

$$\arg\min_{\boldsymbol{A}} \{\widehat{R}_{\text{uni}}(\boldsymbol{A}) + \lambda \|\boldsymbol{A}\|_*\},$$

obtained by the method of Lagrange multipliers.

However, uniform sampling is a strong assumption and often not satisfied (e.g., Srebro and Salakhutdinov, 2010; Foygel et al., 2011; Hernández-Lobato et al., 2014). In the empirical risk minimization framework, it is natural to adjust for such non-uniformity since $\widehat{R}_{\text{uni}}$ is no longer unbiased for R. Interestingly, such biasedness does not lead to an incorrect estimator in an asymptotic sense (Klopp, 2014), a form of robustness result (the first category of works under non-uniformity mentioned in Section 1). This is because $\boldsymbol{A}_\star$ still minimizes $\mathsf{E}\{\widehat{R}_{\text{uni}}(\boldsymbol{A})\}$ even when π_{ij}'s are heterogeneous, and, to achieve consistency, the theory requires that $\mathcal{A}_{n_1,n_2}$ grows asymptotically so that some appropriate "distance" between $\boldsymbol{A}_\star$ and the set $\mathcal{A}_{n_1,n_2}$ converges to zero. For finite sample, one often encounters some forms of misspecification ($\boldsymbol{A}_\star$ is not close to $\mathcal{A}_{n_1,n_2}$). In such settings, the estimator based on $\widehat{R}_{\text{uni}}(\boldsymbol{A})$ is inclined to favor entries with a higher chance of observation, which is often not desirable. For movie recommendation, it is generally not a good idea to neglect those people who rate less frequently, as they might be the customers who do not watch as frequently, and successful movie recommendation would help retain these customers from discontinuing movie subscription services. This is highly related to misspecification in low-dimensional models where misspecification requires weighting adjustments (Wooldridge, 2007). However, matrix completion problems involve a much more challenging high-dimensional setup with possibly diminishing observation probabilities (e.g., Candès and Recht, 2009; Koltchinskii, 2011). That is, $\pi_L := \min_{i,j} \pi_{ij} \to 0$ as $n_1, n_2 \to \infty$. In fact, the diminishing setting is of great interest and plays a central role in most analyses, since it mimics high missing situations such as in the Netflix prize problem ($< 1\%$ of observed ratings).

2.4 Extremely Varying Probabilities: Heterogeneity Meets Asymptotics

For non-uniform settings, one expects heterogeneity among the entries of $\boldsymbol{\Pi}$. We argue that there exist different levels of heterogeneity, and only the "simplest" level has been well-

studied. Define

$$\pi_U := \max_{i,j} \pi_{ij} \quad \text{and} \quad \pi_L := \min_{i,j} \pi_{ij}.$$

Existing work (e.g., Negahban and Wainwright, 2012; Klopp, 2014; Lafond et al., 2014; Cai and Zhou, 2016) is based on an assumption that $\pi_U \asymp \pi_L$, which enforces that all observation probabilities are of the same order. We call this asymptotically homogeneous missing structure. When the observation probabilities vary highly among different entries, this asymptotic framework may not reflect the empirical world. Highly varying probabilities are not rare. As demonstrated in Section 2.3 of Mao et al. (2020), the estimated ratio of π_U to π_L can be high (≥ 20000) in the Yahoo! Webscope dataset, under low-rank models of $\mathbf{\Pi}$ (e.g., Negahban and Wainwright, 2012). In our theoretical analysis (Section 5), we also look into the asymptotically heterogeneous settings where π_U and π_L are of different orders.

3 Empirical Risk Balancing

3.1 Propensity Approaches and their Drawbacks

To deal with non-uniformity, a natural idea is to utilize a weighted empirical risk:

$$\widehat{R}_{\boldsymbol{W}}(\boldsymbol{A}) = \frac{1}{n_1 n_2} \|\boldsymbol{T} \circ \boldsymbol{W}^{\circ(1/2)} \circ (\boldsymbol{Y} - \boldsymbol{A})\|_F^2, \tag{1}$$

where $\boldsymbol{W} = (W_{ij})_{i,j=1}^{n_1,n_2}$ is a matrix composed of weights such that $W_{ij} \geq 1$ for all i, j. A natural choice of $\boldsymbol{W}$ is $(\pi_{ij}^{-1})_{i,j=1}^{n_1,n_2}$, which leads to an unbiased risk estimator for $R(\boldsymbol{A})$, and such method is known as inverse probability weighting (IPW) in the missing data literature. As $\{\pi_{ij}\}$ are unknown in general, most methods with IPW insert the estimated probabilities based on certain models. These ideas have been studied in, e.g., Schnabel et al. (2016) under the form of a nuclear-norm regularized estimator:

$$\arg\min_{\boldsymbol{A}} \{\widehat{R}_{\boldsymbol{W}}(\boldsymbol{A}) + \lambda \|\boldsymbol{A}\|_*\}, \tag{2}$$

where $\lambda > 0$ is a tuning parameter. Despite its conceptual simplicity, it is well-known in the statistical literature that IPW estimators could produce unstable results due to extreme weights (Rubin, 2001; Kang and Schafer, 2007). More problematically for matrix completion, the estimation quality of a high-dimensional probability matrix $\mathbf{\Pi} = (\pi_{ij})_{i,j=1}^{n_1,n_2}$ could also be worsened significantly by diminishing probabilities of observation (as $n_1, n_2 \to \infty$) (Davenport et al., 2014). To solve this problem, Mao et al. (2020) imposed a constraint (effectively an upper bound) on the estimated inverse probabilities, where the constraint has to be aggressively chosen such that some true inverse probabilities do not necessarily satisfy in finite sample. However, there are still two general issues in this line of research. First, the estimation of $\mathbf{\Pi}$ is required. One could come up with a variety of ways to model $\mathbf{\Pi}$. But it is not obvious how to choose a good model for $\mathbf{\Pi}$. Second, the constraint level is tricky to select, and difficult to analyze theoretically. Indeed, the analysis of the effect of the constriant to matrix recovery forms the bulk of the analysis in Mao et al. (2020).

The goal of this work is to propose a method that does not require specific modeling and estimation of $\mathbf{\Pi}$ but still actively adjust for the non-uniformity in the sampling. This

method aims to directly find a stable weight matrix $\boldsymbol{W}$ that adjusts for non-uniformity, *without* enforcing $\boldsymbol{W}$ to be IPW derived from a specific model.

3.2 Balancing Weights

When $\varepsilon_{ij} = 0$ for all i, j (only for motivation purpose, not required for the proposed techniques), we aim to choose $\boldsymbol{W}$ such that $\widehat{R}_{\boldsymbol{W}}$ (left hand side) approximates the desirable "fully-observed" one (right hand side):

$$\frac{1}{n_1 n_2}\|\boldsymbol{T} \circ \boldsymbol{W}^{\circ(1/2)} \circ (\boldsymbol{A}_\star - \boldsymbol{A})\|_F^2 \approx \frac{1}{n_1 n_2}\|\boldsymbol{A}_\star - \boldsymbol{A}\|_F^2, \tag{3}$$

for a set of $\boldsymbol{A}$ (a hypothesis class of $\boldsymbol{A}_\star$ which grows with n_1, n_2) to be specified below. Indeed, we only need to determine those W_{ij} such that $T_{ij} = 1$, since the values of the remaining W_{ij} play no role in (3). Intuitively, the *weights* $\boldsymbol{W}$ are introduced to maintain *balance* between the left and right hand sides of (3). Therefore, we may work with $\widehat{R}_{\boldsymbol{W}}$ as if we were using the uniform empirical risk $\widehat{R}_{\text{uni}}$. The condition (3) can be written as

$$0 \approx \frac{1}{n_1 n_2}\left|\langle (\boldsymbol{T} \circ \boldsymbol{W} - \boldsymbol{J}) \circ \boldsymbol{\Delta}, \boldsymbol{\Delta}\rangle\right|, \tag{4}$$

where $\boldsymbol{\Delta} = \boldsymbol{A} - \boldsymbol{A}_\star$ and $\boldsymbol{J} \in \mathbb{R}^{n_1 \times n_2}$ is a matrix of ones. We call the right hand side the *balancing error* of $\boldsymbol{\Delta}$ with respect to $\boldsymbol{W}$, denoted by $S(\boldsymbol{W}, \boldsymbol{\Delta})$. Naturally, we want to find weights $\boldsymbol{W}$ that minimize the *uniform* balancing error

$$F(\boldsymbol{W}) := \sup_{\boldsymbol{\Delta} \in \mathcal{D}_{n_1,n_2}} S(\boldsymbol{W}, \boldsymbol{\Delta}),$$

for a (standardized) set $\mathcal{D}_{n_1,n_2}$, induced by the hypothesis class $\mathcal{A}_{n_1,n_2}$ of $\boldsymbol{A}_\star$.

A typical assumption is that $\boldsymbol{A}_\star$ is low-rank or approximately low-rank. Various classes are shown to be able to achieve such modeling. For instance, $\mathcal{A}_{n_1,n_2}$ can be chosen as a max-norm ball $\{\boldsymbol{A} : \|\boldsymbol{A}\|_{\max} \le \beta\}$ (e.g., Srebro et al., 2005; Foygel and Srebro, 2011; Cai and Zhou, 2013, 2016; Fang et al., 2018), and the induced choice of $\mathcal{D}_{n_1,n_2}$ would be $\{\boldsymbol{\Delta} : \|\boldsymbol{\Delta}\|_{\max} \le 2\beta\}$. However, the uniform balancing error does not have a closed form and so the computation of the weights would be significantly more difficult and expensive. Similar difficulty exists for nuclear-norm balls.

To solve this problem, we have developed the following novel lemma which allows us to focus on a relaxed version of balancing error that enjoys strong theoretical guarantees (see Section 5).

Lemma 1. *For any matrices $\boldsymbol{B}, \boldsymbol{C} \in \mathbb{R}^{n_1 \times n_2}$, we have*

$$|\langle \boldsymbol{C} \circ \boldsymbol{B}, \boldsymbol{B}\rangle| \le \|\boldsymbol{C}\|\|\boldsymbol{B}\|_{\max}\|\boldsymbol{B}\|_* \le \sqrt{n_1 n_2}\|\boldsymbol{C}\|\|\boldsymbol{B}\|_{\max}^2.$$

The proof of this lemma can be found in Section S1 of the supplemental document. The inequalities in Lemma 1 are tight in general: if $\boldsymbol{C}{=}a\boldsymbol{J}$ and $\boldsymbol{B}{=}b\boldsymbol{J}$ where $a, b \in \mathbb{R}$ and $\boldsymbol{J}$ is the matrix whose entries are all 1, the two equalities would hold simultaneously.

By Lemma 1, $S(\boldsymbol{W}, \boldsymbol{\Delta}) \le \sqrt{n_1 n_2}\|\boldsymbol{T} \circ \boldsymbol{W} - \boldsymbol{J}\|\|\boldsymbol{\Delta}\|_{\max}^2$ for any $\boldsymbol{\Delta} \in \mathbb{R}^{n_1 \times n_2}$, where the right hand side can be regarded as the relaxed balancing error. If we focus on the max-norm

ball (for $\mathcal{A}_{n_1,n_2}$ and hence $\mathcal{D}_{n_1,n_2}$) as discussed before, we are only required to control the spectral norm of $\|\boldsymbol{T} \circ \boldsymbol{W} - \boldsymbol{J}\|$, which is a convex function of $\boldsymbol{W}$. Therefore, we propose the following novel weights:

$$\widehat{\boldsymbol{W}} = \arg\min_{\boldsymbol{W}} \|\boldsymbol{T} \circ \boldsymbol{W} - \boldsymbol{J}\| \tag{5}$$
$$\text{subject to} \quad \|\boldsymbol{T} \circ \boldsymbol{W}\|_F \leq \kappa \quad \text{and} \quad W_{ij} \geq 1,$$

where the optimization is taken only over W_{ij} such that $T_{ij} = 1$. Here $\kappa \geq \sum_{i,j} T_{ij}$ is a tuning parameter.

The weights $\{W_{ij}\}$ are restricted to be greater than or equal to 1, as their counterparts, inverse probabilities, satisfy $\pi_{ij}^{-1} \geq 1$. The term $\|\boldsymbol{T} \circ \boldsymbol{W}\|_F$ regularizes $\boldsymbol{W}$ and is particularly important when ε_{ij}'s are not zero.

Let $h(\kappa) = \|\boldsymbol{T} \circ \widehat{\boldsymbol{W}} - \boldsymbol{J}\|$ where $\widehat{\boldsymbol{W}}$ is defined by (5) with the tuning parameter κ. It is proportional to the relaxed balancing error with respect to $\widehat{\boldsymbol{W}}$. As κ increases, a weaker constraint is imposed on $\boldsymbol{W}$. Therefore $h(\kappa)$ is non-increasing as κ increases. It can be shown that $h(\kappa)$ stays constant for all large enough κ, i.e., h achieves its smallest value. The percentage of (relaxed) balancing with respect to a specific κ is defined as $[M - h(\kappa)]/(M - m)$ where $M := \max_\kappa h(\kappa)$ and $m = \min_\kappa h(\kappa)$. One way to tune κ is to choose κ that achieves certain pre-specified percentage of balancing. We can also select κ from multiple values of κ with respect to certain balancing percentages, via a validation set. In Sections 6 and 7, we compare κ with respect to balancing percentages 100%, 75%, and 50%, and select the one with the smallest validation error.

3.3 Computation

The dual Lagrangian form of the constrained problem (5) is

$$\min_{W_{ij} \geq 1} \left\{ \|\boldsymbol{T} \circ \boldsymbol{W} - \boldsymbol{J}\| + \kappa' \|\boldsymbol{T} \circ \boldsymbol{W}\|_F^2 \right\}, \tag{6}$$

where κ' is the dual parameter. Denote $\boldsymbol{X} = \boldsymbol{T} \circ \boldsymbol{W} - \boldsymbol{J}$, we can obtain the analytic form of the subgradient of the largest singular value by $\partial \|\boldsymbol{X}\| = \boldsymbol{u}_1^\intercal (\partial \boldsymbol{X}) \boldsymbol{v}_1$ where $\boldsymbol{u}_1$ and $\boldsymbol{v}_1$ are the corresponding left and right singular vectors with respect to the largest singular value of matrix $\boldsymbol{X}$. Thus we have

$$\frac{\partial \|\boldsymbol{X}\|}{\partial W_{ij}} = \frac{\partial \|\boldsymbol{X}\|}{\partial \boldsymbol{X}} \frac{\partial \boldsymbol{X}}{\partial W_{ij}} = \boldsymbol{u}_1 \boldsymbol{v}_1^\intercal T_{ij},$$

and $\partial \|\boldsymbol{T} \circ \boldsymbol{W}\|_F^2 / \partial W_{ij} = 2 T_{ij} W_{ij}$. This allows us to efficiently adopt typical algorithms for smooth optimization with box-constraints such as "L-BFGS-B" algorithm.

4 Estimation of $\boldsymbol{A}_\star$

Given the weight estimator $\widehat{\boldsymbol{W}}$ defined by (5), we propose the following hybrid estimator that utilizes the advantages of both max-norm and nuclear-norm regularizations:

$$\widehat{\boldsymbol{A}} = \underset{\|\boldsymbol{A}\|_{\max}\le\beta}{\arg\min} \left\{ \widehat{R}_{\widehat{\boldsymbol{W}}}(\boldsymbol{A}) + \mu\|\boldsymbol{A}\|_* \right\}, \tag{7}$$

where $\|\cdot\|_*$ denotes the nuclear norm, and $\beta > 0$, $\mu \ge 0$ are tunning parameters. As explained in Section 3.2, the balancing weights $\widehat{\boldsymbol{W}}$ aims to make $\widehat{R}_{\widehat{\boldsymbol{W}}}$ behave like the uniform empirical risk $\widehat{R}_{\text{uni}}$ over a max-norm ball. Although not entirely necessary, the additional nuclear-norm penalty can sometimes produce tighter relaxation as shown in Lemma 1. As discussed in Fang et al. (2018), the additional nuclear norm bound shows its advantages under the uniform sampling scheme when the target matrix is exactly low-rank. We also find that using the hybrid of max-norm and nuclear-norm regularizations improve the estimation performance. If one enforces all the elements of $\widehat{\boldsymbol{W}}$ to be 1 (uniform weighting), then the estimator (7) degenerates to the estimator defined in Fang et al. (2018). The major novelty of our work is the stable weights.

We extend the algorithm proposed in Fang et al. (2018) to handle the weighted empirical risk function, so as to solve (7). Corresponding details can be found in Section S5.1 of the supplemental document.

5 Theoretical Properties

We provide a non-asymptotic analysis of the proposed estimator (7). One major challenge of our analysis is the estimation nature of the weights. As the same set of data is used to obtain the weights, the weighted empirical risk $\widehat{R}_{\widehat{\boldsymbol{W}}}(\boldsymbol{A})$ possesses complicated dependence structure, as opposed to the uniform empirical risk $\widehat{R}_{\text{uni}}(\boldsymbol{A})$ (which is assumed to be a sum of independent variables), even for a fixed $\boldsymbol{A}$. To study the convergence, we carefully decompose the errors into different components. We utilize the properties of true weights to control the balancing error term. Besides, we develop a novel lemma (Lemma S4) to study the concentration of the dual max-norm of the noise matrix with entry-wise multiplicative perturbation.

The following two assumptions will be used in our theoretical analysis. Recall that $\pi_U = \max_{i,j}\pi_{ij}$ and $\pi_L = \min_{i,j}\pi_{ij}$.

Assumption 1. *The observation indicators $\{T_{ij}\}$ are independent Bernoulli random variables with $\pi_{ij} = \Pr(T_{ij} = 1)$. The minimum observation probability π_L is positive, but it can depend on n_1, n_2. In particular, both π_U and π_L are allowed to diminish to zero when $n_1, n_2 \to \infty$.*

Assumption 2. *The random errors $\{\epsilon_{ij}\}$ are independent and centered sub-Gaussian random variables such that $E(\epsilon_{ij}) = 0$ and $\max_{i,j}\|\epsilon_{ij}\|_{\psi_2} \le \tau$ where $\|\epsilon_{ij}\|_{\psi_2} := \inf\{t > 0 : E[\exp(\epsilon_{ij}^2/t^2)] \le 2\}$ is the sub-Gaussian norm of ϵ_{ij}. Also, $\{\epsilon_{ij}\}$ are independent of $\{T_{ij}\}$.*

We start with an essential result that the estimated weights $\widehat{\boldsymbol{W}}$ possess the power to balance the non-uniform empirical risk. More specifically, in the following theorem, we derive

a non-asymptotic upper bound of the uniform balancing error evaluated at $\widehat{\boldsymbol{W}}$, where the balancing error can be written as

$$S(\boldsymbol{W}, \boldsymbol{\Delta}) = \frac{1}{n_1 n_2} \left| \|\boldsymbol{T} \circ \boldsymbol{W}^{\circ(1/2)} \circ \boldsymbol{\Delta}\|_F^2 - \|\boldsymbol{\Delta}\|_F^2 \right|.$$

Theorem 1. *Suppose Assumption 1 holds. Take* $\kappa \geq (2\sum_{i,j} \pi_{ij}^{-1})^{1/2}$. *There exists an absolute constant* $C_1 > 0$ *such that for any* $\beta' > 0$,

$$\sup_{\|\boldsymbol{\Delta}\|_{\max} \leq \beta'} S(\widehat{\boldsymbol{W}}, \boldsymbol{\Delta}) \leq C_1 \frac{\beta'^2}{\sqrt{\pi_L (n_1 \wedge n_2)}} \min\left\{ [\log(n_1 + n_2)]^{1/2}, \pi_L^{-1/2} \right\},$$

with probability at least $1 - \exp\{-2^{-1}(\log 2)\pi_L^2 \sum_{i,j} \pi_{ij}^{-1}\} - 1/(n_1 + n_2)$.

If $\|\boldsymbol{A}_\star\|_{\max} \leq \beta$, it is natural to take $\beta' = 2\beta$, since $\|\boldsymbol{\Delta}\|_{\max} = \|\boldsymbol{A} - \boldsymbol{A}_\star\|_{\max} \leq 2\beta$ for any $\boldsymbol{A}$ such that $\|\boldsymbol{A}\|_{\max} \leq \beta$. Therefore, we can take $\beta' = 2\beta$ in Theorem 1 to achieve uniform control over the balancing error associated with the estimation (7).

With the above balancing guarantee, we are now in a good position to study $\widehat{\boldsymbol{A}}$. Our guarantee for $\widehat{\boldsymbol{A}}$ is in terms of the uniform error $d^2(\widehat{\boldsymbol{A}}, \boldsymbol{A}_\star) := (n_1 n_2)^{-1} \|\widehat{\boldsymbol{A}} - \boldsymbol{A}_\star\|_F^2$, instead of the non-uniform error $\tilde{d}^2(\widehat{\boldsymbol{A}}, \boldsymbol{A}_\star) = \|\boldsymbol{\Pi}^{\circ(1/2)} \circ (\widehat{\boldsymbol{A}} - \boldsymbol{A}_\star)\|_F^2 / \|\boldsymbol{\Pi}^{\circ(1/2)}\|_F^2$ (e.g., Klopp, 2014; Cai and Zhou, 2016). Note that the non-uniform error $\tilde{d}^2(\widehat{\boldsymbol{A}}, \boldsymbol{A}_\star)$ places less emphases on entries that are less likely to be observed, although the guarantee in terms of the non-uniform error can be stronger and is easier to obtain. In asymptotically heterogeneous missing settings (i.e., π_U and π_L are of different orders), entries with probabilities of order smaller than π_U may be ignored within the non-uniform error in the asymptotic sense. Therefore it is not a good measure of performance if the guarantee over these entries are also important. In the following theorem, we provide a non-asymptotic error bound of our estimator (7) (based on the estimated weights).

Theorem 2. *Suppose Assumptions 1–2 hold. Assume* $\|\boldsymbol{A}_\star\|_{\max} \leq \beta$, *and* $\mu = \mathcal{O}(\min\{[\log(n_1 + n_2)]^{1/2}, \pi_L^{-1/2}\}/\sqrt{\pi_L(n_1 \wedge n_2)})$. *Then there exists an absolute constant* $C_2 > 0$ *such that for any* $\kappa \geq (2\sum_{i,j} \pi_{ij}^{-1})^{1/2}$,

$$d^2\left(\widehat{\boldsymbol{A}}, \boldsymbol{A}_\star\right) \leq C_2 \left[\frac{\beta^2}{\sqrt{\pi_L(n_1 \wedge n_2)}} \times \min\left\{ [\log(n_1 + n_2)]^{1/2}, \pi_L^{-1/2} \right\} + \frac{\beta \tau \kappa \sqrt{n_1 + n_2}}{n_1 n_2} \right]$$

with probability at least $1 - \exp\{-2^{-1}(\log 2)\pi_L^2 \sum_{i,j} \pi_{ij}^{-1}\} - 2\exp\{-(n_1 + n_2)\} - 1/(n_1 + n_2)$.

First, we consider the asymptotically homogeneous missing structures (i.e., $\pi_L \asymp \pi_U$) which most existing work assumes. Under $\pi_L \asymp \pi_U$, the two errors $d^2(\widehat{\boldsymbol{A}}, \boldsymbol{A}_\star)$ and $\tilde{d}^2(\widehat{\boldsymbol{A}}, \boldsymbol{A}_\star)$ are of the same order because

$$\frac{\pi_L}{\pi_U} d^2(\widehat{\boldsymbol{A}}, \boldsymbol{A}_\star) \leq \tilde{d}^2(\widehat{\boldsymbol{A}}, \boldsymbol{A}_\star) \leq \frac{\pi_U}{\pi_L} d^2(\widehat{\boldsymbol{A}}, \boldsymbol{A}_\star). \tag{8}$$

Therefore, the upper bound for $\tilde{d}^2(\widehat{\boldsymbol{A}}, \boldsymbol{A}_\star)$ that most existing work provides can be directly used to derive an upper bound for $d^2(\widehat{\boldsymbol{A}}, \boldsymbol{A}_\star)$, which shares the same order. Note that π_U and π_L are allowed to be different despite $\pi_U \asymp \pi_L$. So certain non-uniform missing structures

are still allowed under the setting of asymptotically homogeneous missingness. This setting has been studied in Negahban and Wainwright (2012); Klopp (2014); Lafond et al. (2014); Cai and Zhou (2016). Our bound is directly comparable to the work of Cai and Zhou (2016) which studies a max-norm constrained estimation. Their result assumes $\|\boldsymbol{A}_\star\|_\infty \leq \alpha$ for some α, which allows their bound to depend on $\alpha\beta$ instead of β^2 as in our bound. The comparision of error bounds between max-norm-constrained estimation and nuclear-norm-regularized estimation is given in Section 3.5 of Cai and Zhou (2016). As for exactly low-rank matrices, we can further show that our estimator achieves optimal error bound (up to a logarithmic order). Roughly speaking, if κ is small (so weights are close to constant), our estimator would behave like a standard nuclear-norm regularized estimator, and hence share the (near-)optimality of such estimator. We provide the error bound of our estimator under exactly low-rank setting and asymptotically homogeneous missingness, in Theorem S1 of the supplemental document.

For non-uniform missing structures, the orders of π_U and π_L do not necessarily match. When their orders are different, we call these missing structures asymptotically heterogeneous. We now focus on how the upper bound depends on π_U and π_L. As mentioned before, existing results are scarce. Recently, Mao et al. (2020) (their Section 5.3) provided an extension of existing upper bounds to possibly asymptotically heterogeneous settings, with a careful analysis. Corresponding upper bound scales with $\pi_L^{-1}\pi_U^{1/2}$. They also provided an additional result when one has access to the *true* probabilities $\boldsymbol{\Pi}$, and show that the upper bound of the estimator based on the empirical risk defined via the true probabilities can achieve the scaling $\pi_L^{-1/2}$, which is significantly better than $\pi_L^{-1}\pi_U^{1/2}$. However, until now, it remains unclear whether there exists an estimator with this scaling of π_U and π_L, without access to the true probabilities. Interestingly, Theorem 2 provides a positive result, and shows that the upper bound for the proposed estimator achieves this scaling $\pi_L^{-1/2}$ under very mild assumption that π_L is diminishing in at least a slow order, more specifically $\pi_L = \mathcal{O}(1/\log(n_1+n_2))$.

Next, we provide a theoretical result indicating that the scaling $\pi_L^{-1/2}$ cannot be improved under the asymptotically heterogeneous missing structures. In below, we give a minimax lower bound based on a class of asymptotically heterogeneous settings. To the best of the authors' knowledge, the minimax lower bounds under asymptotically heterogeneous regimes have never been studied.

The heterogeneous class that we consider posits

$$(n_1 n_2)^{-1} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \pi_{ij} \asymp \pi_L. \tag{9}$$

It is clear that (9) does not exclude asymptotically homogeneous settings. To demonstrate the heterogeneity, we provide an example as follows. Suppose there is only a fixed number of entries with observation probabilities in constant order, and the observation probabilities of the remaining entries are of the same order as π_L. Then $\pi_U \asymp 1$, and (9) is satisfied. Therefore, for any diminishing π_L, this setting is asymptotically heterogeneous.

Now, we provide the minimax result.

Theorem 3. *Let $\{\epsilon_{ij}\}$ be i.i.d. Gaussian $\mathcal{N}(0,\sigma^2)$ with $\sigma^2 > 0$. For any $\beta > 0$, assume (9) holds with $\pi_L^{-1} = \mathcal{O}(\beta^2(n_1 \wedge n_2)/(\sigma \wedge \beta)^2)$. Then, there exist constants $\delta \in (0,1)$ and $c > 0$*

such that

$$\inf_{\widehat{\boldsymbol{A}}} \sup_{\|\boldsymbol{A}_\star\|_{\max} \le \beta} \Pr\left(d^2\left(\widehat{\boldsymbol{A}}, \boldsymbol{A}_\star\right) > \frac{c(\sigma \wedge \beta)\beta}{\sqrt{\pi_L\left(n_1 \wedge n_2\right)}} \right) \ge \delta.$$

In the discussion below, we focus on $\sigma \asymp 1$, which, most notably, excludes asymptotically noiseless settings. Theorem 3 shows that the scaling $\pi_L^{-1/2}$ in our upper bound obtained in Theorem 2 is essential. Due to the general inequality (Srebro and Shraibman, 2005):

$$\|\boldsymbol{A}_\star\|_\infty \le \|\boldsymbol{A}_\star\|_{\max} \le \sqrt{\operatorname{rank}(\boldsymbol{A}_\star)}\|\boldsymbol{A}_\star\|_\infty, \tag{10}$$

β is not expected to grow fast for low-rank $\boldsymbol{A}_\star$ with bounded entries. For $\beta = \mathcal{O}(\operatorname{polylog}(n))$, our upper bound matches with the lower bound in Theorem 3 up to a logarithmic factor. For general β, our upper bound scales with β^2 instead of $(\sigma \wedge \beta)\beta$ despite its matching scaling with respect to π_L. Indeed, a mismatch between the upper bound and the lower bound also occurs in Cai and Zhou (2016) under asymptotically homogeneous settings, where their bound is derived via an additional assumption $\|\boldsymbol{A}_\star\|_\infty \le \alpha$. Their upper bound scales with $\alpha\beta$ instead of $(\sigma \wedge \alpha)\beta$ as in their minimax lower bound. We leave a more detailed study of the scaling with respect to β as a future direction.

6 Simulations

In this simulation study, we let the target matrix $\boldsymbol{A}_\star \in \mathbb{R}^{n_1 \times n_2}$ be generated by $\boldsymbol{A}_\star = \boldsymbol{U}\boldsymbol{V}^\intercal$, where $\boldsymbol{U} \in \mathbb{R}^{n_1 \times r}, \boldsymbol{V} \in \mathbb{R}^{n_2 \times r}$, and each entry of $\boldsymbol{U}$ and $\boldsymbol{V}$ is sampled uniformly and independently from $[0, 2]$. We set $n_1 = n_2 = 200$ and $r = 5$. Therefore, the rank of the target matrix is 5. The contaminated version of $\boldsymbol{A}_\star$ is then generated as $\boldsymbol{Y} = \boldsymbol{A}_\star + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \in \mathbb{R}^{n_1 \times n_2}$ has i.i.d. mean zero Gaussian entries $\epsilon_{ij} \sim N(0, \sigma_\epsilon^2)$. There are three settings of σ_ϵ, and they are chosen such that the signal-to-noise ratios (SNR:= $(\mathsf{E}\|\boldsymbol{A}_\star\|_F^2/\mathsf{E}\|\boldsymbol{\epsilon}\|_F^2)^{1/2}$) are 1, 5 and 10.

We consider three different missing mechanisms and generate observation indicator matrix $\boldsymbol{T}$ from $\boldsymbol{\Pi} = (\pi_{ij})_{i,j=1}^{n_1,n_2}$ that are specified as follows:

Setting 1: This setting is a uniform missing setting $\pi_{ij} = 0.25$ for all $i, j = 1, \ldots, 200$.

Setting 2: In this setting, we relate the missingness with the value of the target matrix. For entries that have high values, they are more likely to be observed. More specifically, we set

$$\pi_{ij} = \begin{cases} 1/16, & \text{if } A_{\star,ij} \le q_{0.25} \\ 0.25, & \text{if } q_{0.25} < A_{\star,ij} \le q_{0.75} \\ 7/16, & \text{if } A_{\star,ij} > q_{0.75} \end{cases}$$

where q_a is the a quantile of $A_{\star,ij}$, $i, j = 1, \ldots, 200$.

Setting 3: This setting is the contrary of Setting 2. For entries that have high values, they are less likely to be observed.

$$\pi_{ij} = \begin{cases} 7/16, & \text{if } A_{\star,ij} \le q_{0.25} \\ 0.25, & \text{if } q_{0.25} < A_{\star,ij} \le q_{0.75} \\ 1/16, & \text{if } A_{\star,ij} > q_{0.75} \end{cases}$$

where q_a is the a quantile of $A_{\star,ij}$, $i, j = 1, \ldots, 200$.

We generate 200 simulated data sets separately for each of the above settings to compare different matrix completion methods, including the proposed method (`BalWeights`) and five existing matrix completion methods: Mazumder et al. (2010) (`SoftImpute`), Cai and Zhou (2016) (`CZ`), Fang et al. (2018) (`FLT`), Koltchinskii et al. (2011) (`KLT`) and Negahban and Wainwright (2012) (`NW`). For all methods mentioned above, we randomly separate 20% of the observed entries in every simulated dataset and use it as the validation set to select tuning parameters.

In addition to the empirical root mean squared error (RMSE), we also include estimated rank and test error:

$$\text{TE} := \frac{\|(\boldsymbol{J} - \boldsymbol{T}) \circ (\widetilde{\boldsymbol{A}} - \boldsymbol{A}_\star)\|_F}{\sqrt{n_1 n_2 - N}},$$

where $\widetilde{\boldsymbol{A}}$ is a generic estimator of $\boldsymbol{A}_\star$; $\boldsymbol{T}$ is the matrix of observed indicator and N is the number of observed entries. The test error measures the relative estimation error of the unobserved entries. Due to the space limitation, we only present the results for SNR = 5. Results for SNR = 1 and SNR = 10 can be found in Section S6 of the supplemental document. Table 1 summarizes the average RMSE, average TE, and average estimated ranks for all three settings. In all three settings, `SoftImpute`, `CZ` and `KLT` do not provide competitive results as others. For Setting 1, `NW` achieves the smallest RMSE and TE, but `BalWeights` performs closely to it. When SNR = 1 (shown in supplemental document), `BalWeights` performs best — the average RMSE of `BalWeights` is 1.901 while the average RMSE of `NW` is 2.012. As for Settings 2 and 3, `BalWeights` outperforms other methods. Also, `NW` performs significantly worse than `BalWeights` in Setting 2. `FLT` has average RMSE and TE that are close to `BalWeights` in Setting 2 but does not perform well in Setting 3. As a result, we can see that `BalWeights` is quite robust across different missing structures.

7 Real Data Applications

We applied the above methods to two real datasets:

1. Coat Shopping Dataset, which is available at `http://www.cs.cornell.edu/~schnabts/mnar/`. As described in Schnabel et al. (2016), the dataset contains ratings from 290 Turkers on an inventory of 300 items. The self-selected ratings form the training set and the uniformly selected ratings form the test set. The training set consists of 6960 entries and test set consists of 4640 entries.

2. Yahoo! Webscope Dataset, which is available at `http://research.yahoo.com/AcademicRelations`. It contains (incomplete) ratings from 15,400 users on 1000 songs. The dataset consists of two subsets, a training set and a test set. The training set records approximately 300,000 ratings given by the aforementioned 15,400 users. Each song has at least 10 ratings. The test set was constructed by surveying 5,400 out of these 15,400 users, such that each selected user rates exactly 10 additional songs.

For the second dataset, due to its large size, we use a non-convex algorithm of Lee et al. (2010) to obtain `CZ`. Also, we modify this algorithm to incorporate another nuclear-norm regularization, to obtain `BalWeights` and `FLT`. Detailed algorithm can be found in Section S5.2 of the supplemental document. For both datasets, we separate half of the test data set

Table 1: Simulation results for three Settings when SNR=5. The average RMSE ($\overline{\text{RMSE}}$), average TE ($\overline{\text{TE}}$), and average estimated ranks ($\bar{\text{r}}$) with standard errors (SE) in parentheses are provided for six methods (`BalWeights`, `SoftImpute`, `CZ`, `FLT`, `NW` and `KLT`) in comparison. For the columns related $\overline{\text{RMSE}}$ and $\overline{\text{TE}}$, we bold results with the first two smallest errors.

	Setting 1		
Method	$\overline{\text{RMSE}}$	$\overline{\text{TE}}$	$\bar{\text{r}}$
`BalWeights`	**0.679(0.001)**	**0.700(0.001)**	25.150(0.128)
`SoftImpute`	0.699(0.001)	0.721(0.001)	45.005(0.161)
`CZ`	0.895(0.002)	0.899(0.002)	51.075(0.121)
`FLT`	0.682(0.001)	0.703(0.001)	26.705(0.131)
`NW`	**0.668(0.001)**	**0.688(0.001)**	28.04(0.187)
`KLT`	1.913(0.003)	1.976(0.003)	8.720(0.060)
	Setting 2		
Method	$\overline{\text{RMSE}}$	$\overline{\text{TE}}$	$\bar{\text{r}}$
`BalWeights`	**0.624(0.001)**	**0.635(0.001)**	24.980(0.136)
`SoftImpute`	0.648(0.001)	0.660(0.001)	41.240(0.104)
`CZ`	0.922(0.002)	0.945(0.002)	47.170(0.156)
`FLT`	**0.628(0.001)**	**0.640(0.001)**	26.045(0.145)
`NW`	0.665(0.002)	0.674(0.002)	22.030(0.806)
`KLT`	1.980(0.006)	1.880(0.004)	1.355(0.141)
	Setting 3		
Method	$\overline{\text{RMSE}}$	$\overline{\text{TE}}$	$\bar{\text{r}}$
`BalWeights`	**0.925(0.002)**	**1.002(0.002)**	24.090(0.138)
`SoftImpute`	1.143(0.003)	1.254(0.003)	47.240(0.144)
`CZ`	1.222(0.003)	1.324(0.003)	50.590(0.151)
`FLT`	1.026(0.002)	1.118(0.003)	32.440(0.131)
`NW`	**0.964(0.002)**	**1.043(0.002)**	18.350(0.319)
`KLT`	3.174(0.006)	3.477(0.006)	9.575(0.093)

as the validation set to select tuning parameters for all methods. And the remaining half test data set is used as the evaluation set.

Here, we include the test root mean squared error

$$\text{TRMSE} := \frac{\|\boldsymbol{T}_e \circ (\widetilde{\boldsymbol{A}} - \boldsymbol{A}_\star)\|_F}{\sqrt{N_e}},$$

where $\widetilde{\boldsymbol{A}}$ is a generic estimator of $\boldsymbol{A}_\star$; $\boldsymbol{T}_e$ is the indicator matrix for the evaluation set and N_e is the number of evaluation entries, and the test mean absolute error

$$\text{TMAE} := \frac{\sum_{\boldsymbol{T}_{e,ij}=1} |\widetilde{\boldsymbol{A}}_{ij} - \boldsymbol{A}_{\star,ij}|}{N_e},$$

to measure the performance of all the methods. Rank estimation is also provided.

Table 2 shows the TRMSE, TMAE and estimated ranks for the two datasets with all the methods mentioned above. For Coat Shopping Dataset, compared with the existing methods, the proposed method `BalWeights` achieves best TRMSE and TMAE. The errors of `FLT` are similar to that of `BalWeights`, but the estimated rank is larger than that of

Table 2: Test root mean squared errors (TRMSE), test mean absolute errors (TMAE) and estimated ranks (Rank) based on the evaluation set of Coat Shopping Dataset and Yahoo! Webscope Dataset for `BalWeights` and five existing methods proposed respectively in Mazumder et al. (2010) (`SoftImpute`), Cai and Zhou (2016) (`CZ`), Fang et al. (2018)(`FLT`), Negahban and Wainwright (2012) (`NW`) and Koltchinskii et al. (2011) (`KLT`). For the columns related TRMSE and TMAE, we bold results with the first two smallest errors.

	Coat Shopping Dataset		
Method	TRMSE	TMAE	Rank
`BalWeights`	**0.9888**	**0.7627**	26
`SoftImpute`	1.1401	0.8485	15
`CZ`	1.0354	0.8279	31
`FLT`	**0.9980**	**0.7723**	32
`NW`	1.0553	0.7972	25
`KLT`	2.0838	1.5733	2
	Yahoo! Webscope Dataset		
Method	TRMSE	TMAE	Rank
`BalWeights`	**1.0111**	**0.7739**	64
`SoftImpute`	1.2172	0.9230	31
`CZ`	1.0339	0.8156	29
`FLT`	1.0339	0.8156	29
`NW`	**1.0338**	**0.7954**	25
`KLT`	3.811	1.6589	1

`BalWeights`. In other words, `BalWeights` is significantly more efficient in capturing the signal. For Yahoo! Webscope Dataset, `BalWeights` also has the smallest errors among all the methods. However, compared with `CZ`, `FLT` and `NW` whose errors are relatively close to that of `BalWeights`, `BalWeights` has a higher estimated rank, though 64 is a reasonably small rank for a matrix with size 1000 by 15400. To confirm the fact that the higher errors of `CZ`, `FLT` and `NW` are not due to their smaller rank estimates, we look into the test error sequences obtained by varying the tuning parameters, for each of these three methods. We find that the change of test errors (based on the evaluation set) aligns well with the validation errors (based on the validation set), and the chosen tuning parameters indeed correspond to the almost smallest test errors they can achieve. This suggests that these three estimators are not able to capture additional useful information and hence produce a smaller rank estimates. But the proposed estimator is able to capitalize these additional signals to achieve reduction in errors.

Acknowledgements

The authors thank the reviewers for their helpful comments and suggestions. The work of Raymond K. W. Wong is partially supported by the US National Science Foundation (DMS-1711952, DMS-1806063 and CCF-1934904). The work of Xiaojun Mao is partially supported by NSFC Grant No. 12001109 and 92046021, Shanghai Sailing Program 19YF1402800, and the Science and Technology Commission of Shanghai Municipality grant 20dz1200600. The

work of K. C. G. Chan is partially supported by the US National Science Foundation (DMS-1711952). Portions of this research were conducted with high performance research computing resources provided by Texas A&M University (https://hprc.tamu.edu).

References

Athey, S., M. Bayati, N. Doudchenko, G. Imbens, and K. Khosravi (2018). Matrix completion methods for causal panel data models. Technical report, National Bureau of Economic Research.

Bennett, J. and S. Lanning (2007). The netflix prize. In *Proceedings of KDD cup and workshop*, Volume 2007, pp. 35.

Bhaskar, S. A. (2016). Probabilistic low-rank matrix completion from quantized measurements. *Journal of Machine Learning Research 17*(60), 1–34.

Bi, X., A. Qu, J. Wang, and X. Shen (2017). A group-specific recommender system. *Journal of the American Statistical Association 112*(519), 1344–1353.

Cai, T., D. Kim, Y. Wang, M. Yuan, and H. H. Zhou (2016). Optimal large-scale quantum state tomography with pauli measurements. *The Annals of Statistics 44*(2), 682–712.

Cai, T. T. and W.-X. Zhou (2013). A max-norm constrained minimization approach to 1-bit matrix completion. *The Journal of Machine Learning Research 14*(1), 3619–3647.

Cai, T. T. and W.-X. Zhou (2016). Matrix completion via max-norm constrained optimization. *Electronic Journal of Statistics 10*(1), 1493–1525.

Candès, E. J. and Y. Plan (2010). Matrix completion with noise. *Proceedings of the IEEE 98*(6), 925–936.

Candès, E. J. and B. Recht (2009). Exact matrix completion via convex optimization. *Foundations of Computational Mathematics 9*(6), 717–772.

Chen, Y., Y. Chi, J. Fan, C. Ma, and Y. Yan (2020). Noisy matrix completion: Understanding statistical guarantees for convex relaxation via nonconvex optimization. *SIAM Journal on Optimization 30*(4), 3098–3121.

Chi, E. C., H. Zhou, G. K. Chen, D. O. Del Vecchyo, and K. Lange (2013). Genotype imputation via matrix completion. *Genome research 23*(3), 509–518.

Dai, B., J. Wang, X. Shen, and A. Qu (2019). Smooth neighborhood recommender systems. *Journal of Machine Learning Research 20*(16), 1–24.

Davenport, M. A., Y. Plan, E. van den Berg, and M. Wootters (2014). 1-bit matrix completion. *Information and Inference 3*(3), 189–223.

Fang, E. X., H. Liu, K.-C. Toh, and W.-X. Zhou (2018). Max-norm optimization for robust matrix recovery. *Mathematical Programming 167*(1), 5–35.

Fithian, W. and R. Mazumder (2018). Flexible low-rank statistical modeling with missing data and side information. *Statistical Science 33*(2), 238–260.

Foygel, R., O. Shamir, N. Srebro, and R. R. Salakhutdinov (2011). Learning with the weighted trace-norm under arbitrary sampling distributions. In *Advances in Neural Information Processing Systems*, pp. 2133–2141.

Foygel, R. and N. Srebro (2011). Concentration-based guarantees for low-rank matrix reconstruction. In *Proceedings of the 24th Annual Conference on Learning Theory*, pp. 315–340.

Hastie, T., R. Mazumder, J. D. Lee, and R. Zadeh (2015). Matrix completion and low-rank svd via fast alternating least squares. *Journal of Machine Learning Research 16*(1), 3367–3402.

Hernández-Lobato, J. M., N. Houlsby, and Z. Ghahramani (2014). Probabilistic matrix factorization with non-random missing data. In *International Conference on Machine Learning*, pp. 1512–1520.

Kallus, N., X. Mao, and M. Udell (2018). Causal inference with noisy and missing covariates via matrix factorization. In *Advances in Neural Information Processing Systems*, pp. 6921–6932.

Kang, J. D. and J. L. Schafer (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science 22*(4), 523–539.

Kang, Z., C. Peng, and Q. Cheng (2016). Top-n recommender system via matrix completion. In *Thirtieth AAAI Conference on Artificial Intelligence*.

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

Klopp, O., J. Lafond, É. Moulines, J. Salmon, et al. (2015). Adaptive multinomial matrix completion. *Electronic Journal of Statistics 9*(2), 2950–2975.

Koltchinskii, V. (2011). *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems*. Springer.

Koltchinskii, V., K. Lounici, and A. B. Tsybakov (2011). Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *The Annals of Statistics 39*(5), 2302–2329.

Lafond, J., O. Klopp, E. Moulines, and J. Salmon (2014). Probabilistic low-rank matrix completion on finite alphabets. In *Advances in Neural Information Processing Systems*, pp. 1727–1735.

Lee, J. D., B. Recht, N. Srebro, J. Tropp, and R. R. Salakhutdinov (2010). Practical large-scale optimization for max-norm regularization. In *Advances in neural information processing systems*, pp. 1297–1305.

Mao, X., S. X. Chen, and R. K. Wong (2019). Matrix completion with covariate information. *Journal of the American Statistical Association 114*(525), 198–210.

Mao, X., R. K. Wong, and S. X. Chen (2020). Matrix completion under low-rank missing mechanism. *Statistica Sinica*.

Mazumder, R., T. Hastie, and R. Tibshirani (2010). Spectral regularization algorithms for learning large incomplete matrices. *Journal of Machine Learning Research 11*, 2287–2322.

Montanari, A. and S. Oh (2010). On positioning via distributed matrix completion. In *Sensor Array and Multichannel Signal Processing Workshop (SAM), 2010 IEEE*, pp. 197–200.

Negahban, S. and M. J. Wainwright (2012). Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. *Journal of Machine Learning Research 13*(1), 1665–1697.

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

Rennie, J. D. and N. Srebro (2005). Fast maximum margin matrix factorization for collaborative prediction. In *Proceedings of the 22nd international conference on Machine learning*, pp. 713–719.

Robin, G., O. Klopp, J. Josse, É. Moulines, and R. Tibshirani (2020). Main effects and interactions in mixed and incomplete data frames. *Journal of the American Statistical Association 115*(531), 1292–1303.

Rubin, D. B. (2001). Using propensity scores to help design observational studies: Application to the tobacco litigation. *Health Services and Outcomes Research Methodology 2*(3-4), 169–188.

Schnabel, T., A. Swaminathan, A. Singh, N. Chandak, and T. Joachims (2016). Recommendations as treatments: Debiasing learning and evaluation. Volume **48** of *Proceedings of Machine Learning Research*, New York, New York, USA, pp. 1670–1679. PMLR.

Sengupta, N., N. Srebro, and J. Evans (2021). Simple surveys: Response retrieval inspired by recommendation systems. *Social Science Computer Review 39*(1), 105–129.

Srebro, N., J. Rennie, and T. S. Jaakkola (2005). Maximum-margin matrix factorization. In *Advances in neural information processing systems*, pp. 1329–1336.

Srebro, N. and R. R. Salakhutdinov (2010). Collaborative filtering in a non-uniform world: Learning with the weighted trace norm. In *Advances in Neural Information Processing Systems*, Volume 23, pp. 2056–2064.

Srebro, N. and A. Shraibman (2005). Rank, trace-norm and max-norm. In *International Conference on Computational Learning Theory*, pp. 545–560. Springer.

Wang, Y. (2013). Asymptotic equivalence of quantum state tomography and noisy matrix completion. *The Annals of Statistics 41*(5), 2462–2504.

Weinberger, K. Q. and L. K. Saul (2006). Unsupervised learning of image manifolds by semidefinite programming. *International Journal of Computer Vision 70*(1), 77–90.

Wooldridge, J. M. (2007). Inverse probability weighted estimation for general missing data problems. *Journal of Econometrics 141*(2), 1281–1301.

Zhang, C., S. J. Taylor, C. Cobb, J. Sekhon, et al. (2020). Active matrix factorization for surveys. *Annals of Applied Statistics 14*(3), 1182–1206.

Zhu, Y., X. Shen, and C. Ye (2016). Personalized prediction and sparsity pursuit in latent factor models. *Journal of the American Statistical Association 111*(513), 241–252.

Supplementary Material for "Matrix Completion with Model-free Weighting"

Jiayi Wang[1], Raymond K. W. Wong[1], Xiaojun Mao[2], and Kwun Chuen Gary Chan[3]

[1]*Department of Statistics, Texas A&M University*
[2]*School of Data Science, Fudan University*
[3]*Department of Biostatistics, University of Washington*

S1 Proof of Lemma 1

Proof of Lemma 1. We first list two properties of max norm as follows. (i) As shown in Srebro and Shraibman (2005), $\|\boldsymbol{B}\|_* \leq \sqrt{n_1 n_2}\|\boldsymbol{B}\|_{\max}$. (ii) By an equivalent definition of max norm due to Lee et al. (2008) (also see equation (8) in Jalali and Srebro (2012)), we have $\|\boldsymbol{C}\circ\boldsymbol{B}\| \leq \|\boldsymbol{C}\|\|\boldsymbol{B}\|_{\max}$. Together with the duality of nuclear norm, we can show that

$$\begin{aligned} |\langle \boldsymbol{C}\circ\boldsymbol{B}, \boldsymbol{B}\rangle| &\leq \|\boldsymbol{C}\circ\boldsymbol{B}\|\|\boldsymbol{B}\|_* \leq \|\boldsymbol{C}\|\|\boldsymbol{B}\|_{\max}\|\boldsymbol{B}\|_* \\ &\leq \sqrt{n_1 n_2}\|\boldsymbol{C}\|\|\boldsymbol{B}\|_{\max}^2. \end{aligned} \tag{S1}$$

□

S2 Proofs of Theorems 1, 2 and 3

Let $\boldsymbol{e}_i(n) \in \mathbb{R}^n$ be the canonical basis vector, i.e., the i-th element of $\boldsymbol{e}_i(n)$ is 1 and the remaining elements are 0. We can define similar standard basis elements for n_1-by-n_2 matrices: $\boldsymbol{J}_{ij} = \boldsymbol{e}_i(n_1)\boldsymbol{e}_j^\intercal(n_2)$, which will be used in the applications of matrix Bernstein inequality in our proofs. For any $\beta \geq 0$, define the class of matrices $\mathcal{B}_{\max}(\beta)$ to be the max-norm ball with radius β, i.e.,

$$\mathcal{B}_{\max}(\beta) = \{\boldsymbol{A} \in \mathbb{R}^{n_1\times n_2} : \|\boldsymbol{A}\|_{\max} \leq \beta\}.$$

We also define

$$\mathcal{F} = \{\boldsymbol{u}\boldsymbol{v}^T : \boldsymbol{u} \in \{-1,+1\}^{n_1}, \boldsymbol{v} \in \{-1,+1\}^{n_2}\},$$

the set of rank-one sign matrices. Denote by $K_G \in (1.67, 1.79)$ the Grothendieck's constant. From Srebro and Shraibman (2005),

$$\text{conv}\mathcal{F} \subseteq \mathcal{B}_{\max}(1) \subseteq K_G \text{conv}\mathcal{F}. \tag{S2}$$

Moreover, the cardinality of $\mathcal{F}$ is $|\mathcal{F}| = 2^{n_1+n_2-1}$.

Lemma S1. *Suppose Assumption 1 hold. Let $\boldsymbol{W}_\diamond = (w_{\diamond,i,j}) \in \mathbb{R}^{n_1,n_2}$ where $W_{\diamond,i,j} = \pi_{i,j}^{-1}$. There exists a constant $C_1 \geq 0$ such that with probability at least $1 - 1/(n_1+n_2)$,*

$$\frac{1}{n_1 n_2} \|\boldsymbol{T} \circ \boldsymbol{W}_\diamond - \boldsymbol{J}\| \leq C_1 \min\left\{ \frac{\log^{1/2}(n_1+n_2)}{\sqrt{\pi_L (n_1 \wedge n_2) n_1 n_2}}, \frac{\sqrt{n_1+n_2}}{\pi_L n_1 n_2} \right\}.$$

Proof of Lemma S1. We use two different proof techniques to show the bounds. Depending on the rate of π_L, one of these two bounds is faster.

First, we show the proof for deriving the second bound. As $\{T_{ij}\}$ are Bernoulli random variables, each entry $T_{ij}W_{\diamond,i,j} - 1$ of matrix $\boldsymbol{T} \circ \boldsymbol{W}_\diamond - \boldsymbol{J}$ is sub-Gaussian random variable. Thus according to the definition of the ψ_2 norm, we have

$$\mathsf{E} \exp\{\log(2) \cdot (T_{ij}W_{\diamond,i,j} - 1)^2 / (\pi_{ij}^{-1} - 1)^2\} \leq 2,$$

which implies that $\|T_{ij}W_{\diamond,i,j} - 1\|_{\psi_2} \leq \log^{-1/2} 2 \cdot (\pi_{ij}^{-1} - 1) \leq 2(\pi_{ij}^{-1} - 1)$.

By Theorem S2 in Section S4 , taking $K = \max_{i,j} \|T_{ij}W_{\diamond,i,j} - 1\|_{\psi_2} \leq 2\pi_L^{-1}$ and $t = (n_1+n_2)^{1/2}$ in Theorem S2, there exists an absolute constant $C_1 > 0$ such that

$$\|\boldsymbol{T} \circ \boldsymbol{W}_\diamond - \boldsymbol{J}\| \leq \frac{C_1 \sqrt{n_1+n_2}}{\pi_L},$$

with probability at least $1 - 2 \cdot \exp(-(n_1+n_2))$.

Next, we consider applying the Matrix Bernstein inequality to derive the first bound. For $(n_1 n_2)^{-1} \|\boldsymbol{T} \circ \boldsymbol{W}_\diamond - \boldsymbol{J}\| = \|\sum_{i,j} (T_{ij}W_{\diamond,i,j} - 1)\boldsymbol{J}_{ij}/(n_1 n_2)\|$, where $\boldsymbol{J}_{ij}$ has 1 for $(i,j)-$th, but 0 for all the remaining entries, let $\boldsymbol{M}_{i,j} = (T_{ij}W_{\diamond,i,j} - 1)\boldsymbol{J}_{ij}$, $i = 1, \ldots, n_1$, $j = 1, \ldots, n_2$, then $(n_1 n_2)^{-1} \|\boldsymbol{T} \circ \boldsymbol{W}_\diamond - \boldsymbol{J}\| = \|(n_1 n_2)^{-1} \sum_{i,j} \boldsymbol{M}_{i,j}\|$. We can easily verify that $\mathsf{E}(\boldsymbol{M}_{i,j}) = \boldsymbol{0}$ and $\|\boldsymbol{M}_{i,j}\| \leq \max\{\pi_L^{-1} - 1, 1\}$ for each i, j by Assumption 1.

Since $\mathsf{E}(T_{ij}W_{\diamond,i,j} - 1)^2 = \pi_{ij}^{-1} - 1$, we can show that

$$\begin{aligned}
&\left\| \frac{1}{n_1 n_2} \sum_{i,j} \mathsf{E}\left(\boldsymbol{M}_{i,j} \boldsymbol{M}_{i,j}^\mathsf{T}\right) \right\| = \left\| \frac{1}{n_1 n_2} \sum_{i,j} \mathsf{E}\left(\boldsymbol{M}_{i,j}^\mathsf{T} \boldsymbol{M}_{i,j}\right) \right\| \\
\leq &\frac{1}{n_1 n_2} \max\left\{ \max_{1 \leq i \leq n_1} \sum_{j=1}^{n_2} |1/\pi_{ij} - 1|, \max_{1 \leq j \leq n_2} \sum_{i=1}^{n_1} |1/\pi_{ij} - 1| \right\} \\
\leq &\frac{1}{n_1 \wedge n_2} |1/\pi_L - 1|,
\end{aligned}$$

where the first inequality comes from Corollary 2.3.2 in Golub and Van Loan (1996).

By Theorem S3 in Section S4 , with probability at least $1 - 1/(n_1 + n_2)$, we have

$$\frac{1}{n_1 n_2} \|\boldsymbol{T} \circ \boldsymbol{W}_\diamond - \boldsymbol{J}\| \le 2 \max\left\{ \sqrt{\frac{2\,|1/\pi_L - 1| \log(n_1+n_2)}{(n_1 \wedge n_2)\, n_1 n_2}}, 2 \max\left\{\frac{1}{\pi_L} - 1, 1\right\} \frac{\log^{3/2}(n_1+n_2)}{n_1 n_2} \right\}.$$

Overall, the conclusion follows. □

Lemma S2. *Suppose Assumption 1 holds. With probability at least* $1 - \exp\{-2^{-1}(\log 2)\pi_L^2 \sum_{i,j} \pi_{ij}^{-1}\}$,

$$\|\boldsymbol{T} \circ \boldsymbol{W}_\diamond\|_F^2 \le 2 \sum_{i,j} \pi_{ij}^{-1}.$$

In particular, the probability is lower bounded by $1 - \exp\{-2^{-1}(\log 2) n_1 n_2 \pi_L^2 \pi_U^{-1}\}$.

Proof of Lemma S2. Note that $\|\boldsymbol{T} \circ \boldsymbol{W}_\diamond\|_F^2 = \sum_{i,j} T_{ij} \pi_{ij}^{-2}$. Let $\xi > 0$. By Markov inequality, for any $t \ge 0$,

$$\begin{aligned}\Pr\left(\|\boldsymbol{T} \circ \boldsymbol{W}_\diamond\|_F^2 \ge t\right) = \Pr\left\{\exp(\xi \|\boldsymbol{T} \circ \boldsymbol{W}_\diamond\|_F^2) \ge \exp(\xi t)\right\} &\le \exp(-\xi t) \mathsf{E} \exp\left(\xi \sum_{i,j} T_{ij} \pi_{ij}^{-2}\right) \\ &= \exp(-\xi t) \prod_{i,j} \mathsf{E} \exp(\xi T_{ij} \pi_{ij}^{-2}).\end{aligned}$$

For each (i, j), due to the inequality $1 + x \le \exp(x)$ for $x \ge 0$,

$$\mathsf{E} \exp(\xi T_{ij} \pi_{ij}^{-2}) = 1 + \{\exp(\xi \pi_{ij}^{-2}) - 1\}\pi_{ij} \le \exp[\{\exp(\xi \pi_{ij}^{-2}) - 1\}\pi_{ij}].$$

Combining with the above result and taking $t = 2 \sum_{i,j} \pi_{ij}^{-1}$,

$$\begin{aligned}\Pr\left(\|\boldsymbol{T} \circ \boldsymbol{W}_\diamond\|_F^2 \ge 2 \sum_{i,j} \pi_{ij}^{-1}\right) &\le \exp\left[-2\xi \sum_{i,j} \pi_{ij}^{-1} + \sum_{i,j} \{\exp(\xi \pi_{ij}^{-2}) - 1\}\pi_{ij}\right] \\ &= \exp\left[-\sum_{i,j} \pi_{ij} \left\{1 + 2\xi \pi_{ij}^{-2} - \exp(\xi \pi_{ij}^{-2})\right\}\right].\end{aligned}$$

Note the above inequality holds for any $\xi > 0$.

Next, we focus on the term $g(\xi \pi_{ij}^{-2})$ where $g(x) = 1 + 2x - \exp(x)$ for $x \ge 0$. It is easy to show that g attains its maximum at $x = \log 2$, and $g(\log 2) = 2 \log 2 - 1 > 0$. Also, $g(x)$ is increasing for $0 \le x \le \log 2$.

Take $\xi = (\log 2)\pi_L^2$. Then $0 \le \xi \pi_{ij}^{-2} \le \log 2$, and hence $g(\xi \pi_{ij}^{-2}) > 0$, for all i, j. The lower bound of $g(\xi \pi_{ij}^{-2})$ is crucial in determining the order of the probability bound. Since $g(x) \ge x/2$ for $0 \le x \le \log 2$,

$$g(\xi \pi_{ij}^{-2}) = g(\pi_L^2 \pi_{ij}^{-2} \log 2) \ge \frac{\log 2}{2} \pi_L^2 \pi_{ij}^{-2}, \quad \forall i, j$$

We conclude that

$$\sum_{i,j} \pi_{ij} \left\{ 1 + 2\xi \pi_{ij}^{-2} - \exp(\xi \pi_{ij}^{-2}) \right\} \geq \frac{\log 2}{2} \pi_L^2 \sum_{i,j} \pi_{ij}^{-1} \geq \frac{\log 2}{2} n_2 n_2 \pi_L^2 \pi_U^{-1},$$

which leads to the desired result. □

With these two lemmas, we are posed to prove Theorem 1.

Proof of Theorem 1. By Lemma S2, we can show that with probability at least $1-\exp\{-2^{-1}(\log 2)\pi_L^2 \sum_{i,j} \pi_{ij}^{-1}\}$, $\|\boldsymbol{T} \circ \boldsymbol{W}_\diamond\|_F \leq (2\sum_{i,j} \pi_{ij}^{-1})^{1/2}$ and hence $\boldsymbol{W}_\diamond$ is feasible for the constrained optimization (5).

Based on the definition of the proposed estimator $\widehat{\boldsymbol{W}}$, we have

$$\begin{aligned} S\left(\widehat{\boldsymbol{W}}, \boldsymbol{\Delta}\right) &= \frac{1}{n_1 n_2} \left| \left\langle \boldsymbol{\Delta}, \left(\boldsymbol{T} \circ \widehat{\boldsymbol{W}} - \boldsymbol{J} \right) \circ \boldsymbol{\Delta} \right\rangle \right| \\ &\leq \frac{1}{\sqrt{n_1 n_2}} \|\boldsymbol{T} \circ \boldsymbol{W}_\diamond - \boldsymbol{J}\| \, \|\boldsymbol{\Delta}\|_{\max}^2 \\ &\leq \frac{\beta'^2}{\sqrt{n_1 n_2}} \|\boldsymbol{T} \circ \boldsymbol{W}_\diamond - \boldsymbol{J}\| . \end{aligned}$$

The desired result then follows from Lemma S1. □

Our theoretical result of the final estimator $\widehat{\boldsymbol{A}}$ will be based on a key lemma (Lemma S4), which establishes the dual of max norm of random matrix $\boldsymbol{\epsilon}$ with general entry-wise scaling. Before we prove Lemma S4, we now show a comparison theorem between sub-Gaussian complexity and Gaussian complexity. This result (Lemma S3) extends Theorem 8 in Banerjee et al. (2014) to allow arbitrary entrywise scaling.

Define the Gaussian width and Gaussian complexity of the set $\mathcal{A}$ respectively as

$$w(\mathcal{A}) = \mathsf{E}_{\boldsymbol{G}} \left[\sup_{\boldsymbol{A} \in \mathcal{A}} \langle \boldsymbol{A}, \boldsymbol{G} \rangle \right] \quad \text{and} \quad \tilde{w}(\mathcal{A}) = \mathsf{E}_{\boldsymbol{G}} \left[\sup_{\boldsymbol{A} \in \mathcal{A}} |\, \langle \boldsymbol{A}, \boldsymbol{G} \rangle \,| \right],$$

where $\boldsymbol{G} = (G_{ij})$ and each $\{G_{ij}\}$ are independent standard Gaussian random variables. In our study, $\mathcal{A}$ is a max-norm ball, and so is symmetric. Therefore Gaussian width and Gaussian complexity are equivalent.

Lemma S3 (Extension of Theorem 8 in Banerjee et al. (2014))**.** *Suppose Assumption 2 holds. Let* $\boldsymbol{B} = (B_{ij}) \in \mathbb{R}^{n_1 \times n_2}$ *be a fixed matrix such that* $B_{ij} \geq 0$ *for each* i, j*. Then*

$$\mathsf{E}[\|\boldsymbol{B} \circ \boldsymbol{\epsilon}\|_{\max}^*] \leq \eta_0 \tau \mathsf{E}[\|\boldsymbol{B} \circ \boldsymbol{G}\|_{\max}^*],$$

where $\|\cdot\|_{\max}^*$ *is the dual norm of max norm,* $\boldsymbol{G} = (G_{ij})$ *has independent standard Gaussian entries which are also independent of the random errors* $\{\epsilon_{ij}\}$*, and* $\eta_0 > 0$ *is an absolute constant.*

Proof of Lemma S3. Since the desired result obviously holds if $\boldsymbol{B} = \mathbf{0}$, we assume $\boldsymbol{B} \neq \mathbf{0}$ in the rest of this proof. By definition, $\|\boldsymbol{C}\|_{\max}^* = \sup_{\|\boldsymbol{X}\|_{\max} \leq 1} \langle \boldsymbol{X}, \boldsymbol{C} \rangle$ for any $\boldsymbol{C} \in \mathbb{R}^{n_1 \times n_2}$. Therefore our goal is to bound a scaled sub-Gaussian complexity via the corresponding scaled Gaussian complexity. We now extend the proof of Theorem 8 of Banerjee et al. (2014) to allow an additional entrywise scaling parameter $\boldsymbol{B}$. We start with considering the sub-Gaussian process $Y_{\boldsymbol{X}} = \langle \boldsymbol{X}, \boldsymbol{B} \circ \boldsymbol{\epsilon} \rangle$ and the Gaussian process $Z_{\boldsymbol{X}} = \langle \boldsymbol{X}, \boldsymbol{B} \circ \boldsymbol{G} \rangle$, both indexed by $\boldsymbol{X} \in \mathcal{B}_{\max}(1)$. For any $\boldsymbol{X}_1, \boldsymbol{X}_2 \in \mathcal{B}_{\max}(1)$, by the general Hoeffding's inequality given in Theorem 2.6.3 of Vershynin (2018), we have

$$\Pr\left(|Y_{\boldsymbol{X}_1} - Y_{\boldsymbol{X}_2}| \geq t\right) \leq 2 \cdot \exp\left(-\frac{C_1 t^2}{\tau^2 \|\boldsymbol{B} \circ (\boldsymbol{X}_1 - \boldsymbol{X}_2)\|_F^2}\right), \quad t > 0, \tag{S3}$$

where $C_1 > 0$ is an absolute constant. One can show that $\mathsf{E}(Z_{\boldsymbol{X}_1} - Z_{\boldsymbol{X}_2})^2 = \|\boldsymbol{B} \circ (\boldsymbol{X}_1 - \boldsymbol{X}_2)\|_F^2$. According to Theorem 2.1.5 of Talagrand (2006), we can apply the generic chaining argument for upper bounds on the empirical processes $\sqrt{c} Y_{\boldsymbol{X}}/\tau$ and $Z_{\boldsymbol{X}}$. This yields

$$\mathsf{E}_{\boldsymbol{\epsilon}}\left[\sup_{\boldsymbol{X}_1, \boldsymbol{X}_2 \in \mathcal{B}_{\max}(1)} |Y_{\boldsymbol{X}_1} - Y_{\boldsymbol{X}_2}|\right] \leq \eta_1 \tau \mathsf{E}_{\boldsymbol{G}}\left[\sup_{\boldsymbol{X}_1 \in \mathcal{B}_{\max}(1)} Z_{\boldsymbol{X}_1}\right] = \eta_1 \tau w(\mathcal{B}_{\max}(1)), \tag{S4}$$

where η_1 is an absolute constant. Further, we can see that if $\boldsymbol{X} \in \mathcal{B}_{\max}(1)$, then $-\boldsymbol{X} \in \mathcal{B}_{\max}(1)$. Then we have

$$\begin{aligned}
\sup_{\boldsymbol{X}_1, \boldsymbol{X}_2 \in \mathcal{B}_{\max}(1)} |Y_{\boldsymbol{X}_1} - Y_{\boldsymbol{X}_2}| &= \sup_{\boldsymbol{X}_1, \boldsymbol{X}_2 \in \mathcal{B}_{\max}(1)} (Y_{\boldsymbol{X}_1} - Y_{\boldsymbol{X}_2}) = \sup_{\boldsymbol{X}_1 \in \mathcal{B}_{\max}(1)} Y_{\boldsymbol{X}_1} + \sup_{\boldsymbol{X}_2 \in \mathcal{B}_{\max}(1)} (-Y_{\boldsymbol{X}_2}) \\
&= \sup_{\boldsymbol{X}_1 \in \mathcal{B}_{\max}(1)} Y_{\boldsymbol{X}_1} + \sup_{-\boldsymbol{X}_2 \in \mathcal{B}_{\max}(1)} (\langle -\boldsymbol{X}_2, \boldsymbol{B} \circ \boldsymbol{\epsilon} \rangle) = 2 \sup_{\boldsymbol{X}_1 \in \mathcal{B}_{\max}(1)} Y_{\boldsymbol{X}_1}.
\end{aligned}$$

By taking the expectation on $\boldsymbol{\epsilon}$ on both side, we have

$$\mathsf{E}_{\boldsymbol{\epsilon}}\left[\sup_{\boldsymbol{X}_1, \boldsymbol{X}_2 \in \mathcal{B}_{\max}(1)} |Y_{\boldsymbol{X}_1} - Y_{\boldsymbol{X}_2}|\right] = 2\mathsf{E}_{\boldsymbol{\epsilon}}\left[\sup_{\boldsymbol{X}_1 \in \mathcal{B}_{\max}(1)} Y_{\boldsymbol{X}_1}\right]. \tag{S5}$$

As a result, with $\eta_0 = \eta_1/2$, we have

$$\mathsf{E}_{\boldsymbol{\epsilon}}\left[\sup_{\boldsymbol{X} \in \mathcal{B}_{\max}(1)} \langle \boldsymbol{B} \circ \boldsymbol{\epsilon}, \boldsymbol{X} \rangle\right] = \mathsf{E}_{\boldsymbol{\epsilon}}\left[\sup_{\boldsymbol{X} \in \mathcal{B}_{\max}(1)} Y_{\boldsymbol{X}}\right] \leq \eta_0 \tau w(\mathcal{B}_{\max}(1)). \tag{S6}$$

That completes the proof. □

Lemma S4. *Suppose Assumption 2 holds. Let $\boldsymbol{B} = (B_{ij}) \in \mathbb{R}^{n_1 \times n_2}$ be a fixed matrix such that $B_{ij} \geq 0$ for each i, j. There exists an absolute constant $C_2 > 0$ such that, with probability at least $1 - 2\exp\{-(n_1 + n_2)\}$,*

$$\|\boldsymbol{B} \circ \boldsymbol{\epsilon}\|_{\max}^* \leq C_2 \tau \|\boldsymbol{B}\|_F \sqrt{n_1 + n_2}.$$

Proof of Lemma S4. Define the set

$$\widetilde{\mathcal{B}}_{\max}(\beta) = \{\boldsymbol{B} \circ \boldsymbol{X} : \boldsymbol{X} \in \mathcal{B}_{\max}(\beta)\} \subset \mathbb{R}^{n_1 \times n_2}.$$

Note that we have

$$\mathsf{E}_{\boldsymbol{G}}\left[\sup_{\boldsymbol{X} \in \mathcal{B}_{\max}(1)} \langle \boldsymbol{B} \circ \boldsymbol{G}, \boldsymbol{X} \rangle\right] = \mathsf{E}_{\boldsymbol{G}}\left[\sup_{\boldsymbol{X} \in \mathcal{B}_{\max}(1)} \langle \boldsymbol{G}, \boldsymbol{B} \circ \boldsymbol{X} \rangle\right] = w(\widetilde{\mathcal{B}}_{\max}(1)).$$

Write $\widetilde{\mathcal{F}} = \{\boldsymbol{B} \circ \boldsymbol{X} : \boldsymbol{X} \in \mathcal{F}\}$. By the the relationship (S2), we have

$$\widetilde{\mathcal{F}} \subseteq \widetilde{\mathcal{B}}_{\max}(1) \subseteq \{\boldsymbol{B} \circ \boldsymbol{X} : \boldsymbol{X} \in K_G \operatorname{conv}(\mathcal{F})\} = K_G\{\boldsymbol{B} \circ \boldsymbol{X} : \boldsymbol{X} \in \operatorname{conv}(\mathcal{F})\} = K_G \operatorname{conv}(\widetilde{\mathcal{F}}).$$

Due to the properties of Gaussian width (see, e.g., Appendix A.1 of Banerjee et al., 2014), we have

$$w(\widetilde{\mathcal{B}}_{\max}(1)) \leq w(K_G \operatorname{conv}(\widetilde{\mathcal{F}})) = K_G w(\operatorname{conv}(\widetilde{\mathcal{F}})) = K_G w(\widetilde{\mathcal{F}}).$$

As for any $\boldsymbol{X} \in \mathcal{F}$, we have

$$\|\boldsymbol{B} \circ \boldsymbol{X}\|_F = \|\boldsymbol{B}\|_F,$$

and so $\langle \boldsymbol{G}, \boldsymbol{B} \circ \boldsymbol{X} \rangle \sim \mathcal{N}(0, \|\boldsymbol{B}\|_F^2)$. Recall that the $|\mathcal{F}| = 2^{n_1+n_2-1}$. By Proposition 3.1(ii) of Koltchinskii (2011), we have

$$w(\widetilde{\mathcal{F}}) \leq C_3 \|\boldsymbol{B}\|_F \sqrt{n_1 + n_2}.$$

where C_3 is a absolute constant. By Lemma S3, we conclude that

$$\begin{aligned}
\mathsf{E}_{\boldsymbol{\epsilon}}\left[\sup_{\boldsymbol{X} \in \mathcal{B}_{\max}(1)} \langle \boldsymbol{B} \circ \boldsymbol{\epsilon}, \boldsymbol{X} \rangle\right] &\leq \eta_0 \tau \mathsf{E}_{\boldsymbol{G}}\left[\sup_{\boldsymbol{X} \in \mathcal{B}_{\max}(1)} \langle \boldsymbol{B} \circ \boldsymbol{G}, \boldsymbol{X} \rangle\right] \\
&= \eta_0 \tau w(\widetilde{\mathcal{B}}_{\max}(1)) \leq K_G \eta_0 \tau w(\widetilde{\mathcal{F}}) \\
&\leq C_3 K_G \eta_0 \tau \|\boldsymbol{B}\|_F \sqrt{n_1 + n_2}. \qquad \text{(S7)}
\end{aligned}$$

Let $\varphi(\boldsymbol{Z}) = \sup_{\|\boldsymbol{X}\|_{\max} \leq 1} \langle \boldsymbol{B} \circ \boldsymbol{Z}, \boldsymbol{X} \rangle$ for any $\boldsymbol{Z} \in \mathbb{R}^{n_1 \times n_2}$. We aim to provide the concentration of $\varphi(\boldsymbol{\epsilon})$ to its expectation. For notational simplicity, we will focus on the setting with $B_{ij} > 0$ for all i, j; otherwise, one can reduce the support of φ to those entries corresponding to non-zero B_{ij}. Due to the possibly unbounded support of $\boldsymbol{\epsilon}$, we adopt an extension of McDiarmid's inequality Kontorovich (2014) with unbounded diameter. For any $\boldsymbol{Z}_1 = (Z_{1,ij}), \boldsymbol{Z}_2 = (Z_{2,ij}) \in \mathbb{R}^{n_1 \times n_2}$,

$$\begin{aligned}
|\varphi(\boldsymbol{Z}_1) - \varphi(\boldsymbol{Z}_2)| &\leq \sup_{\|\boldsymbol{X}\|_{\max} \leq 1} |\langle \boldsymbol{B} \circ \boldsymbol{Z}_1, \boldsymbol{X} \rangle - \langle \boldsymbol{B} \circ \boldsymbol{Z}_2, \boldsymbol{X} \rangle| \\
&\leq \sup_{\|\boldsymbol{X}\|_{\max} \leq 1} \sum_{i,j} B_{ij} |X_{ij}| |Z_{1,ij} - Z_{2,ij}| \\
&\leq \sup_{\boldsymbol{X} \in K_G \operatorname{conv}(\mathcal{F})} \sum_{i,j} B_{ij} |X_{ij}| |Z_{1,ij} - Z_{2,ij}| \\
&\leq q(\boldsymbol{Z}_1, \boldsymbol{Z}_2),
\end{aligned}$$

where $q(\boldsymbol{Z}_1, \boldsymbol{Z}_2) =: \sum_{i,j} q_{ij}(Z_{1,ij}, Z_{2,ij}) =: \sum_{i,j} K_G B_{ij}|Z_{1,ij} - Z_{2,ij}|$ is a metric. Therefore φ is 1-Lipschitz with respect to the metric q. Let ε'_{ij} be an independent copy of ε_{ij}, and γ_{ij} be an independent Rademacher random variable. We can show that the subgaussian norm of $\gamma_{ij} q_{ij}(\varepsilon_{ij}, \varepsilon'_{ij})$ is bounded by $C_4 \tau b_{ij}$ for some absolute constant $C_4 > 0$. By Theorem 1 of Kontorovich (2014), we conclude that

$$\mathsf{P}(|\varphi(\boldsymbol{\epsilon}) - \mathsf{E}\varphi(\boldsymbol{\epsilon})| > t) \leq 2\exp\left(-\frac{t^2}{2C_4\tau^2\|\boldsymbol{B}\|_F^2}\right), \quad t \geq 0.$$

Combining with (S7), we achieve the desired result.

□

Lemma S5. *Suppose Assumptions 1 and 2 hold. There exists an absolute constant $C_2 > 0$ such that with probability at least $1 - 2\exp\{-(n_1 + n_2)\}$,*

$$\left\|\boldsymbol{T} \circ \widehat{\boldsymbol{W}} \circ \boldsymbol{\epsilon}\right\|_{\max}^* \leq C_2 \tau\kappa\sqrt{n_1 + n_2}.$$

Proof of Lemma S5. Notice that $\boldsymbol{\epsilon}$ is independent of $\boldsymbol{T}$, and $\widehat{\boldsymbol{W}}$ is a function of $\boldsymbol{T}$. By Lemma S4, conditioned on $\boldsymbol{T}$, we have

$$\left\|\boldsymbol{T} \circ \widehat{\boldsymbol{W}} \circ \boldsymbol{\epsilon}\right\|_{\max}^* \leq C_2 \tau\|\boldsymbol{T} \circ \widehat{\boldsymbol{W}}\|_F\sqrt{n_1 + n_2}, \tag{S8}$$

with conditional probability at least $1 - 2\exp\{-(n_1 + n_2)\}$. Since the probability bound does not depend on $\boldsymbol{T}$, (S8) holds with the same probability bound unconditionally. By construction, $\|\boldsymbol{T} \circ \widehat{\boldsymbol{W}}\|_F \leq \kappa$, we have the desired result. □

Proof of Theorem 2. It follows from the definition of $\widehat{\boldsymbol{A}}$ that for $\boldsymbol{A}_\star \in \mathbb{R}^{n_1 \times n_2}$ with $\|\boldsymbol{A}_\star\|_{\max} \leq \beta$,

$$\frac{1}{n_1 n_2}\left\|\boldsymbol{T} \circ \widehat{\boldsymbol{W}}^{\circ 1/2} \circ \left(\widehat{\boldsymbol{A}} - \boldsymbol{Y}\right)\right\|_F^2 \leq \frac{1}{n_1 n_2}\left\|\boldsymbol{T} \circ \widehat{\boldsymbol{W}}^{\circ 1/2} \circ (\boldsymbol{A}_\star - \boldsymbol{Y})\right\|_F^2 + \mu(\|\boldsymbol{A}_\star\|_* - \|\widehat{\boldsymbol{A}}\|_*). \tag{S9}$$

Since we can rewrite the first term in the left hand side of (S9) as

$$\frac{1}{n_1 n_2}\left\|\boldsymbol{T} \circ \widehat{\boldsymbol{W}}^{\circ 1/2} \circ \left(\widehat{\boldsymbol{A}} - \boldsymbol{Y}\right)\right\|_F^2 = \frac{1}{n_1 n_2}\left\|\boldsymbol{T} \circ \widehat{\boldsymbol{W}}^{\circ 1/2} \circ \left(\widehat{\boldsymbol{A}} - \boldsymbol{A}_\star + \boldsymbol{A}_\star - \boldsymbol{Y}\right)\right\|_F^2,$$

the inequality (S9) leads to

$$\begin{aligned}\frac{1}{n_1 n_2}\left\|\boldsymbol{T} \circ \widehat{\boldsymbol{W}}^{\circ 1/2} \circ \left(\widehat{\boldsymbol{A}} - \boldsymbol{A}_\star\right)\right\|_F^2 \leq & \frac{2}{n_1 n_2}\left\langle \boldsymbol{T} \circ \widehat{\boldsymbol{W}}^{\circ 1/2} \circ \left(\widehat{\boldsymbol{A}} - \boldsymbol{A}_\star\right), \boldsymbol{T} \circ \widehat{\boldsymbol{W}}^{\circ 1/2} \circ (\boldsymbol{Y} - \boldsymbol{A}_\star)\right\rangle + \mu(\|\boldsymbol{A}_\star\|_* - \|\widehat{\boldsymbol{A}}\|_*) \\ = & \frac{2}{n_1 n_2}\left\langle \widehat{\boldsymbol{A}} - \boldsymbol{A}_\star, \boldsymbol{T} \circ \widehat{\boldsymbol{W}} \circ \boldsymbol{\epsilon}\right\rangle + \mu(\|\boldsymbol{A}_\star\|_* - \|\widehat{\boldsymbol{A}}\|_*).\end{aligned}$$

Therefore, due to Theorem 1, Lemma S5 and condition of μ, with the property that $\|\boldsymbol{A}_\star\|_* \leq \sqrt{n_1 n_2}\|\boldsymbol{A}_\star\|_{\max}$,

we have

$$\begin{aligned}
\frac{1}{n_1 n_2}\left\|\widehat{\boldsymbol{A}}-\boldsymbol{A}_\star\right\|_F^2 &\le \frac{1}{n_1 n_2}\left\langle \widehat{\boldsymbol{A}}-\boldsymbol{A}_\star, \left(\boldsymbol{T}\circ\widehat{\boldsymbol{W}}-\boldsymbol{J}\right)\circ\left(\boldsymbol{A}_\star-\widehat{\boldsymbol{A}}\right)\right\rangle + \frac{1}{n_1 n_2}\|\boldsymbol{T}\circ\widehat{\boldsymbol{W}}^{\circ(1/2)}\circ(\widehat{\boldsymbol{A}}-\boldsymbol{A}_\star)\|_F^2 \\
&\le S(\widehat{\boldsymbol{W}},\widehat{\boldsymbol{A}}-\boldsymbol{A}_\star) + \left|\frac{2}{n_1 n_2}\langle \widehat{\boldsymbol{A}}-\boldsymbol{A}_\star, \boldsymbol{T}\circ\widehat{\boldsymbol{W}}\circ\boldsymbol{\epsilon}\rangle\right| + \mu(\|\boldsymbol{A}_\star\|_* - \|\widehat{\boldsymbol{A}}\|_*) \\
&\le S(\widehat{\boldsymbol{W}},\widehat{\boldsymbol{A}}-\boldsymbol{A}_\star) + \frac{2}{n_1 n_2}\left\|\widehat{\boldsymbol{A}}-\boldsymbol{A}_\star\right\|_{\max}\left\|\boldsymbol{T}\circ\widehat{\boldsymbol{W}}\circ\boldsymbol{\epsilon}\right\|_{\max}^* + \mu\|\boldsymbol{A}_\star\|_* \\
&\le C_1(\beta^2+\beta)\min\left\{\frac{\log^{1/2}(n_1+n_2)}{\sqrt{\pi_L(n_1\wedge n_2)}}, \frac{\sqrt{n_1+n_2}}{\pi_L\sqrt{n_1 n_2}}\right\} + \frac{4C_2\beta\tau\kappa\sqrt{n_1+n_2}}{n_1 n_2} \qquad \text{(S10)} \\
&\le C_1(\beta^2)\min\left\{\frac{\log^{1/2}(n_1+n_2)}{\sqrt{\pi_L(n_1\wedge n_2)}}, \frac{\sqrt{n_1+n_2}}{\pi_L\sqrt{n_1 n_2}}\right\} + \frac{4C_2\beta\tau\kappa\sqrt{n_1+n_2}}{n_1 n_2}. \qquad \text{(S11)}
\end{aligned}$$

with probability at least $1-\exp\{-2^{-1}(\log 2)\pi_L^2\sum_{i,j}\pi_{ij}^{-1}\}-2\exp\{-(n_1+n_2)\}-1/(n_1+n_2)$. □

Proof of Theorem 3. Without loss of generality, we assume that $n_1 \ge n_2$. For some constant $0 \le \gamma \le 1$ such that $B = \sigma^{-2}(\sigma\wedge\beta)^2/(\gamma^2)$ is an integer and $B \le n_2$, define

$$\mathcal{C}_1 = \left\{\tilde{\boldsymbol{A}} = (A_{ij}) \in \mathbb{R}^{n_1\times B} : A_{ij} \in \{0, \gamma\beta\}, \forall 1 \le i \le n_1, 1 \le j \le B\right\},$$

and consider the associated set of block matrices

$$\mathcal{A}(\mathcal{C}_1) = \left\{\boldsymbol{A} = \left(\tilde{\boldsymbol{A}}|\ldots|\tilde{\boldsymbol{A}}|\boldsymbol{0}\right) \in \mathbb{R}^{n_1\times n_2} : \tilde{\boldsymbol{A}} \in \mathcal{C}_1\right\},$$

where $\boldsymbol{0}$ denotes the $n_1 \times (n_2 - B\lfloor n_2/B\rfloor)$ zero matrix.

It is easy to see that for any $\boldsymbol{A} \in \mathcal{A}(\mathcal{C}_1)$, we have that $\|\boldsymbol{A}\|_{\max} \le \sqrt{B}\|\boldsymbol{A}\|_\infty \le \beta$. Due to Lemma 2.9 in Tsybakov (2009), there exists a subset $\mathcal{A}^0 \subset \mathcal{A}(\mathcal{C}_1)$ containing the zero $n_1 \times n_2$ matrix $\boldsymbol{0}$ where $\mathrm{Card}(\mathcal{A}^0) \ge 2^{Bn_1/8}+1$ and for any two distinct elements $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$ of $\mathcal{A}^0$,

$$\|\boldsymbol{A}_1-\boldsymbol{A}_2\|_F^2 \ge \frac{n_1 B}{8}\left\{\gamma^2\beta^2\left\lfloor\frac{n_2}{B}\right\rfloor\right\} \ge \frac{n_1 n_2\gamma^2\beta^2}{16}. \qquad \text{(S12)}$$

For any $\boldsymbol{A} \in \mathcal{A}^0$, from the noisy observed model in section 2.2, the probability distribution $\mathbb{P}_{\boldsymbol{A}} = \Pi_{i,j}[(2\pi\sigma^2)^{-1/2}\exp\{-(Y_{ij}-A_{ij})^2/(2\sigma^2)\}]^{T_{ij}}$. Take $\mathbb{P}_{\boldsymbol{0}} = \Pi_{i,j}[(2\pi\sigma^2)^{-1/2}\exp\{-Y_{ij}^2/(2\sigma^2)\}]^{T_{ij}}$. Thus the Kullback-Leibler divergence $K(\mathbb{P}_{\boldsymbol{0}},\mathbb{P}_{\boldsymbol{A}}) = \mathsf{E}_{\mathbb{P}_{\boldsymbol{0}}}(\log(\mathbb{P}_{\boldsymbol{0}}/\mathbb{P}_{\boldsymbol{A}}))$ between $\mathbb{P}_{\boldsymbol{0}}$ and $\mathbb{P}_{\boldsymbol{A}}$ satisfies

$$K(\mathbb{P}_{\boldsymbol{0}},\mathbb{P}_{\boldsymbol{A}}) = \mathsf{E}_{\mathbb{P}_{\boldsymbol{0}}}\left(\sum_{ij} T_{ij}\frac{A_{ij}^2 - 2A_{ij}Y_{ij}}{2\sigma^2}\right) = \frac{\left\|\boldsymbol{\Pi}^{\circ 1/2}\circ\boldsymbol{A}\right\|_F^2}{2\sigma^2} \le \frac{\gamma^2\beta^2\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\pi_{ij}}{2\sigma^2} \le C_5\frac{\gamma^2\beta^2 n_1 n_2\pi_L}{2\sigma^2},$$

for some positive constant C_5. The last inequality is due to the condition that $n_1 n_2\pi_L \asymp \sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\pi_{ij}$.

From above we deduce the condition

$$\frac{1}{\mathrm{Card}(\mathcal{A}^0)-1}\sum_{\boldsymbol{A}\in\mathcal{A}^0} K\left(\mathbb{P}_{\mathbf{0}},\mathbb{P}_{\boldsymbol{A}}\right) \le \lambda \log\left(\mathrm{Card}(\mathcal{A}^0)-1\right), \tag{S13}$$

The above condition is valid when we take

$$\gamma^2 = C_6\left(\frac{(\sigma\wedge\beta)^2}{\beta^2 n_2 \pi_L}\right)^{1/2}$$

for some constant C_6 that depends on λ. Also, one can verify that under the conditions $\pi_L^{-1} = \mathcal{O}(\beta^2(n_1\wedge n_2)/(\sigma\wedge\beta)^2)$ and $\pi_L^{1/2} = \mathcal{O}((n_1\wedge n_2)^{1/2}\sigma^2/[\beta(\sigma\wedge\beta))]$, $\gamma\le 1$ and $B\le n_2$. Then we subsitute γ^2 in the bound of S12 and we achieve the final bound as the one showed in the Theorem.

Together with the similar argument when $n_2\ge n_1$, the result now follows by application of Theorem 2.5 in Tsybakov (2009). This completes the proof. □

Lemma S6. *Suppose Assumption 2 hold. For a fixed matrix $\boldsymbol{B}=(B_{ij})\in\mathbb{R}^{n_1\times n_2}$ where $B_{ij}\ge 0$, there exists an absolute constant $C_5>0$ such that, with probability at least $1-2\exp(-(n_1+n_2))$,*

$$\|\boldsymbol{B}\circ\boldsymbol{\epsilon}\| \le C_5\|\boldsymbol{B}\|_\infty\tau(\sqrt{n_1}+\sqrt{n_2}).$$

Proof. By the definition of $\|\cdot\|_{\psi_2}$,

$$\max_{i,j}\|B_{ij}\epsilon_{ij}\|_{\psi_2} \le \|\boldsymbol{B}\|_\infty\tau.$$

Apply Theorem S2 in Section S4 , and take $t=\sqrt{n_1}+\sqrt{n_2}$ in Theorem S2. Then conclusion follows.

□

S3 Non-asymptotic Error Bound under Low-rank Settings and Asymptotically Homogeneous Missingness

Theorem S1. *Suppose Assumption 2 hold and $\pi_L\asymp\pi_U\asymp\pi$. Assume $\|\boldsymbol{A}_\star\|_{\max}\le\beta$ and $\boldsymbol{A}_\star$ has rank R. If $\kappa'=\kappa-\|\boldsymbol{T}\|_F$ is bounded and $\mu\asymp\sqrt{\{\tau^2\pi\log(n_1+n_2)\}\{(n_1\wedge n_2)\,n_1n_2\}^{-1}}$, then there exists a constant $C_6>0$, such that with probability at least $1-3(n_1+n_2)^{-1}$,*

$$d^2(\widehat{\boldsymbol{A}},\boldsymbol{A}_\star) \le \frac{C_6R(\tau^2\vee\|A_\star\|_\infty^2)\log(n_1+n_2)}{[\pi(n_1\wedge n_2)]^{-1}}.$$

Proof outline. From the basic inequality, we have

$$\frac{1}{n_1n_2}\|\boldsymbol{T}\circ\widehat{\boldsymbol{W}}^{1/2}\circ(\widehat{\boldsymbol{A}}-\boldsymbol{A}_\star)\|_F^2 \le \frac{2}{n_1n_2}\langle\widehat{\boldsymbol{A}}-\boldsymbol{A}_\star,\boldsymbol{T}\circ\widehat{\boldsymbol{W}}\circ\boldsymbol{\epsilon}\rangle+\mu\|\boldsymbol{A}_\star\|_*-\mu\|\widehat{\boldsymbol{A}}\|_*.$$

Note that weights are restricted to be greater than 1. We then have

$$\frac{1}{n_1 n_2}\|\boldsymbol{T}\circ(\widehat{\boldsymbol{A}}-\boldsymbol{A}_\star)\|_F^2 \le \frac{1}{n_1 n_2}\|\boldsymbol{T}\circ\widehat{\boldsymbol{W}}^{1/2}\circ(\widehat{\boldsymbol{A}}-\boldsymbol{A}_\star)\|_F^2 \le \frac{2}{n_1 n_2}\|\widehat{\boldsymbol{A}}-\boldsymbol{A}_\star\|_*\|\boldsymbol{T}\circ\widehat{\boldsymbol{W}}\circ\boldsymbol{\epsilon}\| + \mu(\|\boldsymbol{A}_\star\|_* - \|\widehat{\boldsymbol{A}}\|_*).$$

Due to the constraint of κ', $\|\widehat{\boldsymbol{W}}\|_\infty$ is bounded, we can use Lemma S7 to derive the bound of $\|\boldsymbol{T}\circ\widehat{\boldsymbol{W}}\circ\boldsymbol{\epsilon}\|$. The remaining argument is rather standard and the same as the proof for No-weighted estimators with nuclear norm regularization (Klopp, 2014). □

Lemma S7. *Suppose assumptions in Corollary S1 hold. Then there exists a constant $C_7 > 0$ such that, with probability at least $1-(n_1+n_2)$,*

$$\frac{1}{n_1 n_2}\|\boldsymbol{T}\circ\widehat{\boldsymbol{W}}\circ\boldsymbol{\epsilon}\| \le C_7\sqrt{\frac{\tau^2\pi\log(n_1+n_2)}{(n_1\wedge n_2)\,n_1 n_2}}$$

Proof. The proof is very similar as the proof in Lemma S1.

We consider applying the Matrix Berinstein inequality for random matrices with bounded sub-exponential norm.

Due to the constraint of κ', there exists a constant C_8 such that $\|\hat{\boldsymbol{W}}\|_\infty \le C_8$.

For $(n_1n_2)^{-1}\|\boldsymbol{T}\circ\widehat{\boldsymbol{W}}\circ\boldsymbol{\epsilon}\| = \|\sum_{i,j}(T_{ij}\hat{W}_{i,j}\epsilon_{ij})\boldsymbol{J}_{ij}/(n_1n_2)\|$, where $\boldsymbol{J}_{ij}$ has 1 for $(i,j)-$th, but 0 for all the remaining entries, let $\boldsymbol{M}_{i,j} = (T_{ij}\hat{W}_{i,j}\epsilon_{ij})\boldsymbol{J}_{ij}$, then $(n_1n_2)^{-1}\|\boldsymbol{T}\circ\widehat{\boldsymbol{W}}\circ\boldsymbol{\epsilon}\| = \|(n_1n_2)^{-1}\sum_{i,j}\boldsymbol{M}_{i,j}\|$. We can easily verify that $\mathsf{E}(\boldsymbol{M}_{i,j}) = \boldsymbol{0}$. Note that ϵ_{ij} are sub-Gaussian random variables and therefore sub-exponential random variables. Then $\max_{i,j}\|\|\boldsymbol{M}_{i,j}\|\|_{\psi_1} \le \max_{i,j}\|T_{ij}\hat{W}_{i,j}\epsilon_{ij}\|_{\psi_1} \le C_9\tau$, where $\|\cdot\|_{\psi_1}$ is the sub-exponential norm of a random variable and C_9 is some constant depending on the C_8.

Since $\mathsf{E}(T_{ij}\hat{W}_{i,j}\epsilon_{ij})^2 \le cC_8^2\pi_{ij}\tau^2$ for some absolute constant c, we can show that

$$\begin{aligned}
&\left\|\frac{1}{n_1n_2}\sum_{i,j}\mathsf{E}\left(\boldsymbol{M}_{i,j}\boldsymbol{M}_{i,j}^\intercal\right)\right\| = \left\|\frac{1}{n_1n_2}\sum_{i,j}\mathsf{E}\left(\boldsymbol{M}_{i,j}^\intercal\boldsymbol{M}_{i,j}\right)\right\| \\
\le&\frac{1}{n_1n_2}\max\left\{\max_{1\le i\le n_1}\sum_{j=1}^{n_2}c_2\pi_{ij}\tau^2, \max_{1\le j\le n_2}\sum_{i=1}^{n_1}c_2\pi_{ij}\tau^2\right\} \\
\le&\frac{c_3\tau^2}{n_1\wedge n_2}\pi,
\end{aligned}$$

for some constant c_3.

By Proposition 11 in Klopp (2014) , there exsits a constant C_7, such that with probability at least $1-1/(n_1+n_2)$,

$$\frac{1}{n_1n_2}\left\|\boldsymbol{T}\circ\widehat{\boldsymbol{W}}\circ\boldsymbol{\epsilon}\right\| \le C_7\max\left\{\sqrt{\frac{\tau^2\pi\log(n_1+n_2)}{(n_1\wedge n_2)\,n_1n_2}}, \tau\log(1/\sqrt{\pi})\frac{\log^{3/2}(n_1+n_2)}{n_1n_2}\right\}.$$

Overall, the conclusion follows. □

S4 Useful Results

Theorem S2 (Theorem 4.4.5 of Vershynin (2018)). *Let $\boldsymbol{A}$ be an $n_1 \times n_2$ random matrix whose entries A_{ij} are independent, mean zero, sub-gaussian random variables. Then, for any $t > 0$ we have*

$$\|\boldsymbol{A}\| \leq CK(\sqrt{n_1} + \sqrt{n_2} + t)$$

with probability at least $1 - 2\exp(-t^2)$. Here $K = \max_{ij} \|A_{ij}\|_{\psi_2}$ and C is an absolute constant.

Proof. The proof can be found on Page 91 in Vershynin (2018). □

Theorem S3 (Proposition 1 of Koltchinskii et al. (2011)). *Let $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_N$ be independent random matrices with dimensions $n_1 \times n_2$ that satisfy $E\boldsymbol{Z}_i = 0$ and $\|\boldsymbol{Z}_i\| \leq U$ almost surely for some constant U and all $i = 1, \ldots, n$. Define*

$$\sigma_Z = \max\left\{ \left\| \frac{1}{N} \sum_{i=1}^{N} E(\boldsymbol{Z}_i \boldsymbol{Z}_i^{\intercal}) \right\|^{1/2}, \left\| \frac{1}{N} \sum_{i=1}^{N} E(\boldsymbol{Z}_i^{\intercal} \boldsymbol{Z}_i) \right\|^{1/2} \right\}.$$

Then, for all $t > 0$, with probability at least $1 - \exp(-t)$ we have

$$\left\| \frac{\boldsymbol{Z}_1 + \cdots + \boldsymbol{Z}_N}{N} \right\| \leq 2\max\left\{ \sigma_Z \sqrt{\frac{t + \log(n_1 + n_2)}{N}}, U\frac{t + \log(n_1 + n_2)}{N} \right\},$$

Proof. The proof can be found on Page 2325 in Koltchinskii et al. (2011). □

S5 Algorithm

S5.1 Convex Algorithm for Solving (7)

Follow Fang et al. (2018) and Cai and Zhou (2016), we consider an equivalent form objective function in (7) below.

$$\min_{\boldsymbol{X}, \boldsymbol{Z}} \frac{1}{n_1 n_2} \|\boldsymbol{T} \circ \widehat{\boldsymbol{W}}^{\circ(1/2)} \circ (\boldsymbol{Y} - \boldsymbol{Z}_{12})\|_F^2 + \mu\langle \boldsymbol{I}, \boldsymbol{X} \rangle,$$

$$\text{Subject to } \boldsymbol{X} \succcurlyeq 0,\ \boldsymbol{X} = \boldsymbol{Z},\ \boldsymbol{Z} \in \mathcal{P}_\beta$$

where $\boldsymbol{Z}, \boldsymbol{X} \in \mathbb{R}^{(n_1+n_2)\times(n_1+n_2)}$, $\mathcal{S}$ is the class of all symmetric matrices in $\mathbb{R}^{(n_1+n_2)\times(n_1+n_2)}$, $\mathcal{P}_\beta := \{\boldsymbol{C} \in \mathcal{S} : \text{diag}(\boldsymbol{C}) \geq 0, \|\boldsymbol{C}\|_\infty \leq \beta\}$, $\boldsymbol{I}$ is an identity matrix and

$$\boldsymbol{Z} = \begin{bmatrix} \boldsymbol{Z}_{11} & \boldsymbol{Z}_{12} \\ \boldsymbol{Z}_{12}^{\intercal} & \boldsymbol{Z}_{22} \end{bmatrix}, \boldsymbol{Z}_{11} \in \mathbb{R}^{n_1 \times n_1}, \boldsymbol{Z}_{22} \in \mathbb{R}^{n_2 \times n_2}$$

The derivation of above representation mainly comes from two facts: 1. The nuclear norm of $\boldsymbol{Z}_{12}$ is the the smallest possible sum of elements on the diagonal of $\boldsymbol{Z}$ given $\boldsymbol{Z} \succcurlyeq 0$ (Fazel et al., 2001); 2. The max norm of matrix $\boldsymbol{Z}_{12}$ is the smallest possible maximum element on the diagonal of $\boldsymbol{Z}$ given $\boldsymbol{Z} \succcurlyeq 0$ (Srebro et al., 2005).

The augmented Lagrangian function can be written as

$$\mathcal{L}(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{V}) = \frac{1}{n_1 n_2} \|\boldsymbol{T} \circ \widehat{\boldsymbol{W}}^{\circ(1/2)} \circ (\boldsymbol{Y} - \boldsymbol{Z}_{12})\|_F^2 + \mu \langle \boldsymbol{I}, \boldsymbol{X} \rangle + \langle \boldsymbol{V}, \boldsymbol{X} - \boldsymbol{Z} \rangle + \frac{\rho}{2} \|\boldsymbol{X} - \boldsymbol{Z}\|_F^2,$$

$$\text{Subject} \quad \boldsymbol{X} \succcurlyeq 0, \ \boldsymbol{Z} \in \mathcal{P}_\beta,$$

where $\boldsymbol{V} \in \mathbb{R}^{(n_1+n_2)\times(n_1+n_2)}$ is the dual variable and $\rho > 0$ is a hyper-parameter.

Then the alternating direction method of multipliers (ADMM) algorithm solves this optimization problem by minimizing the augmented Lagrangian with respect to different variables alternatingly. More explicitly, at the $(t+1)$-th iteration, the following updates are implemented:

$$\boldsymbol{X}^{t+1} = \Pi\{\boldsymbol{Z}^t + \rho^{-1}(\boldsymbol{V}^t + \mu \boldsymbol{I})\},$$

$$\boldsymbol{Z}^{(t+1)} = \arg\min_{\boldsymbol{Z} \in \mathcal{P}_\beta} \frac{1}{n_1 n_2} \|\boldsymbol{T} \circ \widehat{\boldsymbol{W}}^{\circ(1/2)} \circ (\boldsymbol{Y} - \boldsymbol{Z}_{12})\|_F^2 + \frac{\rho}{2} \|\boldsymbol{Z} - \boldsymbol{X}^{t+1} - \rho^{-1}\boldsymbol{V}^t\|_F^2 = \Phi_{\boldsymbol{T},\boldsymbol{Y},\widehat{\boldsymbol{W}},\beta}\{\boldsymbol{X}^{t+1} + \rho^{-1}\boldsymbol{V}^t\},$$

$$\boldsymbol{V}^{t+1} = \boldsymbol{V}^t + \tau\rho(\boldsymbol{X}^{t+1} - \boldsymbol{Z}^{t+1}),$$

where $\Pi(\cdot)$ is the projection to the space $\{\boldsymbol{C} \in \mathcal{S} : \boldsymbol{C} \succcurlyeq 0\}$, and $\Phi_{\boldsymbol{T},\boldsymbol{Y},\widehat{\boldsymbol{W}},\beta}$ is defined in Definition S1. Detailed derivation can be found in Fang et al. (2018) and Cai and Zhou (2016).

Definition S1. *We use $\boldsymbol{C}(i,j)$ to represent the element on the i-th row and j-th column of a matrix $\boldsymbol{C}$. For the matrix $\boldsymbol{C} \in \mathbb{R}^{(n_1+n_2)\times(n_1+n_2)}$, it can be partitioned into*

$$\boldsymbol{C} = \begin{bmatrix} \boldsymbol{C}_{11} & \boldsymbol{C}_{12} \\ \boldsymbol{C}_{12}^\intercal & \boldsymbol{C}_{22} \end{bmatrix}, \boldsymbol{C}_{11} \in \mathbb{R}^{n_1 \times n_1}, \boldsymbol{C}_{22} \in \mathbb{R}^{n_2 \times n_2}$$

Then

$$\Phi_{\boldsymbol{T},\boldsymbol{Y},\widehat{\boldsymbol{W}},\beta}(\boldsymbol{C}) = \begin{bmatrix} \Phi_{\boldsymbol{T},\boldsymbol{Y},\widehat{\boldsymbol{W}},\beta}(\boldsymbol{C})_{11} & \Phi_{\boldsymbol{T},\boldsymbol{Y},\widehat{\boldsymbol{W}},\beta}(\boldsymbol{C})_{12} \\ \Phi_{\boldsymbol{T},\boldsymbol{Y},\widehat{\boldsymbol{W}},\beta}(\boldsymbol{C})_{12}^\intercal & \Phi_{\boldsymbol{T},\boldsymbol{Y},\widehat{\boldsymbol{W}},\beta}(\boldsymbol{C})_{22} \end{bmatrix},$$

where

$$
\begin{aligned}
&\Phi_{\boldsymbol{T},\boldsymbol{Y},\widehat{\boldsymbol{W}},\beta}(\boldsymbol{C})_{11}(i,j) = \min\{\beta, \max\{\boldsymbol{C}_{11}(i,j), -\beta\}\} \qquad \text{if } i \neq j,\\
&\Phi_{\boldsymbol{T},\boldsymbol{Y},\widehat{\boldsymbol{W}},\beta}(\boldsymbol{C})_{11}(i,j) = \min\{\beta, \max\{\boldsymbol{C}_{11}(i,j), 0\}\} \qquad \text{if } i = j,\\
&\Phi_{\boldsymbol{T},\boldsymbol{Y},\widehat{\boldsymbol{W}},\beta}(\boldsymbol{C})_{22}(i,j) = \min\{\beta, \max\{\boldsymbol{C}_{22}(i,j), -\beta\}\}, \qquad \text{if } i \neq j,\\
&\Phi_{\boldsymbol{T},\boldsymbol{Y},\widehat{\boldsymbol{W}},\beta}(\boldsymbol{C})_{22}(i,j) = \min\{\beta, \max\{\boldsymbol{C}_{22}(i,j), 0\}\} \qquad \text{if } i = j,\\
&\Phi_{\boldsymbol{T},\boldsymbol{Y},\widehat{\boldsymbol{W}},\beta}(\boldsymbol{C})_{12}(i,j) = \min\left\{\beta, \max\left\{\frac{\boldsymbol{Y}(i,j)\widehat{\boldsymbol{W}}(i,j) + \rho\boldsymbol{C}(i,j)}{\widehat{\boldsymbol{W}}(i,j) + \rho}, -\beta\right\}\right\} \qquad \text{if } \boldsymbol{T}(i,j) = 1,\\
&\Phi_{\boldsymbol{T},\boldsymbol{Y},\widehat{\boldsymbol{W}},\beta}(\boldsymbol{C})_{12}(i,j) = \min\{\beta, \max\{\boldsymbol{C}_{12}(i,j), -\beta\}\} \text{ if } i \neq j \qquad \text{if } \boldsymbol{T}(i,j) = 0.
\end{aligned}
$$

We summarize the algorithm in Algorithm 1. Some piratical implementations to adaptively tune ρ and accelerate the computation can be found in Section 3.3 and 3.4 in Fang et al. (2018).

Algorithm 1: ADMM algorithm

Input: $\boldsymbol{Y}$, $\boldsymbol{T}$, β, μ, $\widehat{\boldsymbol{W}}$, $\rho = 0.1$, $\tau = 1.618$, K
Initialize $\boldsymbol{X}^0$, $\boldsymbol{Z}^0$, $\boldsymbol{V}^0$, R
for $t = 1$ **to** $K - 1$ **do**
  $\boldsymbol{X}^{t+1} \leftarrow \Pi\{\boldsymbol{Z}^t + \rho^{-1}(\boldsymbol{V}^t + \mu\boldsymbol{I})\}$
  $\boldsymbol{Z}^{(t+1)} \leftarrow \Phi_{\boldsymbol{T},\boldsymbol{Y},\widehat{\boldsymbol{W}},\beta}\{\boldsymbol{X}^{t+1} + \rho^{-1}\boldsymbol{V}^t\}$
  $\boldsymbol{V}^{t+1} \leftarrow \boldsymbol{V}^t + \tau\rho(\boldsymbol{X}^{t+1} - \boldsymbol{Z}^{t+1})$
  Stop if objective value changes less than tolerance
end for

S5.2 Nonconvex Algorithm for Solving (7)

The nonconvex algorithm for max-norm regularization developed in Lee et al. (2010) base on the equivalent definition of max-norm via matrix factorizations:

$$\|\boldsymbol{C}\|_{\max} := \inf\left\{\|\boldsymbol{U}\|_{2,\infty}\|\boldsymbol{V}\|_{2,\infty} : \boldsymbol{C} = \boldsymbol{U}\boldsymbol{V}^\intercal\right\},$$

where $\|\cdot\|_{2,\infty}$ denotes the maximum l_2 row norm of a matrix.

To incorporate the nuclear norm regularization, we also notice an equivalent definition of the nuclear

norm:

$$\|\boldsymbol{C}\|_* := \inf \frac{1}{2}\left\{\|\boldsymbol{U}\|_F^2 + \|\boldsymbol{V}\|_F^2 : \boldsymbol{C} = \boldsymbol{U}\boldsymbol{V}^\intercal\right\}.$$

Then we have the following relaxation of the objective function in (7). Take

$$f(\boldsymbol{L}, \boldsymbol{R}) = \frac{1}{n_1 n_2}\|\boldsymbol{T} \circ \widehat{\boldsymbol{W}}^{\circ(1/2)} \circ (\boldsymbol{Y} - \boldsymbol{L}\boldsymbol{R}^\intercal)\|_F^2 + \frac{\mu}{2}(\|\boldsymbol{L}\|_F^2 + \|\boldsymbol{R}\|_F^2),$$

and we obtain

$$\min_{\boldsymbol{L},\boldsymbol{R}} f(\boldsymbol{L}, \boldsymbol{R}),$$

$$\text{Subject to } \max\left\{\|\boldsymbol{L}\|_{2,\infty}, \|\boldsymbol{R}\|_{2,\infty}\right\} \leq \beta.$$

This optimization form is exactly the one in Lee et al. (2010) except that we add another nuclear penalty in the objective function f.

Like what Lee et al. (2010) considered, the projected gradient descent method can be applied to iteratively solve this problem. We define the project $\mathcal{P}_B$ as the Euclidean projection onto the set $\{\boldsymbol{M} : \|\boldsymbol{M}\|_{2,\infty} \leq B\}$. This projection can be computed by re-scaling the rows of current input matrix whose norms exceed B so their norms equal B. Rows with norms less than B are unchanged by the projection. We summarize the algorithm in Algorithm 2.

Algorithm 2: Projected gradient descent algorithm

Input: $\boldsymbol{Y}$, $\boldsymbol{T}$, β, μ, $\widehat{\boldsymbol{W}}$, step size τ, K
Initialize $\boldsymbol{L}^0$, $\boldsymbol{R}^0$,
for $t = 1$ **to** $K - 1$ **do**
  $\boldsymbol{L}^{t+1} \leftarrow \mathcal{P}_\beta\left(\boldsymbol{L} - \tau\frac{\partial f}{\partial \boldsymbol{L}}\right)$
  $\boldsymbol{R}^{t+1} \leftarrow \mathcal{P}_\beta\left(\boldsymbol{R} - \tau\frac{\partial f}{\partial \boldsymbol{R}}\right)$
  Stop if objective value changes less than tolerance
end for

S6 Additional Simulation Results

The simulation results for SNR = 1 and SNR = 10 are shown in Table S1 and S2 respectively.

Table S1: Similar to Table 1, but for SNR = 1.

Method	Setting 1 $\overline{\text{RMSE}}$	$\overline{\text{TE}}$	$\bar{\text{r}}$
`Proposed`	**1.901(0.004)**	**1.918(0.004)**	13.69(0.097)
`SoftImpute`	1.944(0.004)	1.961(0.004)	19.55(0.092
`CZ`	2.052(0.004)	2.044(0.004)	27.695(0.128)
`FLT`	**1.927(0.004)**	**1.946(0.004)**	15.265(0.105)
`NW`	2.012(0.004)	2.01(0.004)	25.61(0.069)
`KLT`	2.439(0.005)	2.492(0.005)	10.175(0.063)
Method	Setting 2 $\overline{\text{RMSE}}$	$\overline{\text{TE}}$	$\bar{\text{r}}$
`Proposed`	**1.716(0.004)**	**1.669(0.004)**	14.73(0.113)
`SoftImpute`	1.721(0.004)	1.685(0.004)	16.335(0.107)
`CZ`	1.86(0.004)	1.799(0.004)	25.965(0.115)
`FLT`	**1.711(0.004)**	**1.674(0.004)**	14.565(0.102)
`NW`	1.805(0.005)	1.747(0.005)	37.82(0.422)
`KLT`	2.16(0.005)	2.093(0.005)	2.065(0.110)
Method	Setting 3 $\overline{\text{RMSE}}$	$\overline{\text{TE}}$	$\bar{\text{r}}$
`Proposed`	**2.412(0.006)**	**2.586(0.007)**	12.495(0.098)
`SoftImpute`	2.923(0.007)	3.113(0.007)	29.15(0.112)
`CZ`	2.641(0.006)	2.812(0.006)	28.695(0.109)
`FLT`	2.878(0.007)	3.097(0.007)	20.105(0.105)
`NW`	**2.668(0.006)**	**2.779(0.007)**	33.115(0.066)
`KLT`	3.667(0.007)	3.969(0.007)	9.765(0.067)

References

Banerjee, A., S. Chen, F. Fazayeli, and V. Sivakumar (2014). Estimation with norm regularization. In *Advances in Neural Information Processing Systems*, pp. 1556–1564.

Cai, T. T. and W.-X. Zhou (2016). Matrix completion via max-norm constrained optimization. *Electronic Journal of Statistics 10*(1), 1493–1525.

Fang, E. X., H. Liu, K.-C. Toh, and W.-X. Zhou (2018). Max-norm optimization for robust matrix recovery. *Mathematical Programming 167*(1), 5–35.

Fazel, M., H. Hindi, and S. P. Boyd (2001). A rank minimization heuristic with application to minimum order system approximation. In *Proceedings of the 2001 American Control Conference.(Cat. No. 01CH37148)*, Volume 6, pp. 4734–4739. IEEE.

Table S2: Similar to Table 1, but for SNR = 10.

Method	Setting 1 $\overline{\text{RMSE}}$	$\overline{\text{TE}}$	$\bar{\text{r}}$
`Proposed`	**0.425(0.001)**	**0.443(0.001)**	30.565(0.143)
`SoftImpute`	0.483(0.001)	0.501(0.001)	91.615(1.845)
`CZ`	0.663(0.001)	0.688(0.001)	54.225(0.178)
`FLT`	0.427(0.001)	0.446(0.001)	32.105(0.129)
`NW`	**0.405(0.001)**	**0.422(0.001)**	30.180(0.196)
`KLT`	1.894(0.003)	1.958(0.003)	8.665(0.059)
Method	Setting 2 $\overline{\text{RMSE}}$	$\overline{\text{TE}}$	$\bar{\text{r}}$
`Proposed`	**0.401(0.001)**	**0.418(0.001)**	29.360(0.136)
`SoftImpute`	0.475(0.001)	0.484(0.001)	120.670(3.387)
`CZ`	0.684(0.002)	0.722(0.002)	59.300(0.485)
`FLT`	**0.409(0.001)**	**0.426(0.001)**	30.380(0.145)
`NW`	0.496(0.003)	0.510(0.003)	19.055(0.368)
`KLT`	1.975(0.006)	1.873(0.004)	1.375(0.144)
Method	Setting 3 $\overline{\text{RMSE}}$	$\overline{\text{TE}}$	$\bar{\text{r}}$
`Proposed`	**0.627(0.001)**	**0.688(0.002)**	32.675(0.144)
`SoftImpute`	0.868(0.002)	0.966(0.003)	77.115(1.343)
`CZ`	0.999(0.003)	1.105(0.004)	56.890(0.152)
`FLT`	**0.670(0.002)**	**0.736(0.002)**	39.740(0.820)
`NW`	0.703(0.003)	0.768(0.003)	21.860(0.614)
`KLT`	3.157(0.006)	3.460(0.006)	9.535(0.088)

Golub, G. H. and C. F. Van Loan (1996). Matrix computations. johns hopkins studies in the mathematical sciences.

Jalali, A. and N. Srebro (2012). Clustering using max-norm constrained optimization. In *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, pp. 481–488.

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

Koltchinskii, V. (2011). *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems*. Springer.

Koltchinskii, V., K. Lounici, and A. B. Tsybakov (2011). Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *The Annals of Statistics 39*(5), 2302–2329.

Kontorovich, A. (2014). Concentration in unbounded metric spaces and algorithmic stability. In *International Conference on Machine Learning*, pp. 28–36.

Lee, J. D., B. Recht, N. Srebro, J. Tropp, and R. R. Salakhutdinov (2010). Practical large-scale optimization for max-norm regularization. In *Advances in neural information processing systems*, pp. 1297–1305.

Lee, T., A. Shraibman, and R. Špalek (2008). A direct product theorem for discrepancy. In *2008 23rd Annual IEEE Conference on Computational Complexity*, pp. 71–80. IEEE.

Srebro, N., J. Rennie, and T. S. Jaakkola (2005). Maximum-margin matrix factorization. In *Advances in neural information processing systems*, pp. 1329–1336.

Srebro, N. and A. Shraibman (2005). Rank, trace-norm and max-norm. In *International Conference on Computational Learning Theory*, pp. 545–560. Springer.

Talagrand, M. (2006). *The generic chaining: upper and lower bounds of stochastic processes*. Springer Science & Business Media.

Tsybakov, A. B. (2009). *Introduction to Nonparametric Estimation*. New York: Springer-Verlag New York.

Vershynin, R. (2018). *High-dimensional probability: An introduction with applications in data science*, Volume 47. Cambridge university press.