Matrix Completion with Model-free Weighting

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

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 & Srebro, 2005), computer vision (Weinberger & Saul, 2006), positioning (Montanari & 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 & 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 & Zhou, 2016; Kang et al., 2016; Zhu et al., 2016; Bi et al., 2017; Fithian & Mazumder, 2018; Dai et al., 2019; Robin et al., 2020; Chen et al., 2020). Although many statistical and computational breakthroughs (e.g., Candès & 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 & Srebro, 2011; Klopp, 2014; Cai & 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 & Salakhutdinov, 2010; Foygel et al., 2011; Negahban & 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 & 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

[1] Department of Statistics, Texas A&M University, College Station, TX 77843, USA [2] School of Data Science, Fudan University, Shanghai, 200433, China [3] Department of Biostatistics, University of Washington, Seattle, WA 98195, USA. Correspondence to: Xiaojun Mao <maoxj@fudan.edu.cn>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021. Copyright 2021 by the author(s).

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 & 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 $\mathrm{polylog}(n)$ to represent a polylogarithmic function of n, i.e., a polynomial in $\log n$. So $\mathcal{O}(\mathrm{polylog}(n))$ 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:

$$R(\boldsymbol{A}) = \frac{1}{n_1n_2}\mathsf{E}\left(\|\boldsymbol{Y} - \boldsymbol{A}\|_F^2\right), \quad \boldsymbol{A} \in \mathbb{R}^{n_1\times n_2}.$$

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

$$\widehat{R}_{\text{uni}}(\boldsymbol{A}) = \frac{1}{n_1n_2}\|\boldsymbol{T}\circ(\boldsymbol{Y} - \boldsymbol{A})\|_F^2, \quad \boldsymbol{A} \in \mathbb{R}^{n_1\times n_2},$$

which is unbiased for $\pi R(\boldsymbol{A})$ (e.g., Candès & Recht, 2009; Candès & 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., $\{\boldsymbol{A} : \operatorname{rank}(\boldsymbol{A}) \le r\}$), and a nuclear norm ball of radius ν (i.e., $\{\boldsymbol{A} : \|\boldsymbol{A}\|_* \le \nu\}$). 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 & 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 & 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 & Wainwright, 2012; Klopp, 2014; Lafond et al., 2014; Cai & 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 ($\ge$ 20000) in the Yahoo! Webscope dataset, under low-rank models of $\boldsymbol{\Pi}$ (e.g., Negahban & 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:

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

where $\boldsymbol{W} = (W_{ij})_{i,j=1}^{n_1,n_2}$ is a matrix composed of weights such that $W_{ij} \ge 1$ 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:

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

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 & Schafer, 2007). More problematically for matrix completion, the estimation quality of a high-dimensional probability matrix $\boldsymbol{\Pi} = (\pi_{ij})_{i,j=1}^{n_1,n_2}$ 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 $\boldsymbol{\Pi}$ is required. One could come up with a variety of ways to model $\boldsymbol{\Pi}$. But it is not obvious how to choose a good model for $\boldsymbol{\Pi}$. 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 $\boldsymbol{\Pi}$ 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 & Srebro, 2011; Cai & 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 A 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:

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

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

The weights $\{W_{ij}\}$ are restricted to be greater than or equal to 1, as their counterparts, inverse probabilities, satisfy $\pi_{ij}^{-1} \ge 1$. 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

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

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

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

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 $\|\boldsymbol{A}_\star\|_{\max} \le \beta$, it is natural to take $\beta' = 2\beta$, since $\|\boldsymbol{\Delta}\|_{\max} = \|\boldsymbol{A} - \boldsymbol{A}_\star\|_{\max} \le 2\beta$ for any $\boldsymbol{A}$ such that $\|\boldsymbol{A}\|_{\max} \le \beta$. 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 & 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 & Wainwright (2012); Klopp (2014); Lafond et al. (2014); Cai & Zhou (2016). Our bound is directly comparable to the work of Cai & 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 & 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*

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

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 & Shraibman, 2005):

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

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 & Zhou (2016) under asymptotically homogeneous settings, where their bound is derived via an additional assumption $\|\boldsymbol{A}_\star\|_\infty \leq \alpha$. 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

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

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.

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

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 & Zhou (2016) (`CZ`), Fang et al. (2018) (`FLT`), Koltchinskii et al. (2011) (`KLT`) and Negahban & 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 F 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

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)

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 & Zhou (2016) (`CZ`), Fang et al. (2018)(`FLT`), Negahban & 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

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 E.2 of the supplemental document. For both datasets, we separate half of the test data set 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 `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., Bayati, M., Doudchenko, N., Imbens, G., and Khosravi, K. Matrix completion methods for causal panel data models. Technical report, National Bureau of Economic Research, 2018.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Srebro, N. and Salakhutdinov, R. R. 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, 2010.

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

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

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

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

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

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

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