Matrix Completion under Low-Rank Missing Mechanism

Xiaojun Mao*, Raymond K. W. Wong† and Song Xi Chen‡

December 20, 2018

Abstract

This paper investigates the problem of matrix completion from corrupted data, when a low-rank missing mechanism is considered. The better recovery of missing mechanism often helps completing the unobserved entries of the high-dimensional target matrix. Instead of the widely used uniform risk function, we weight the observations by inverse probabilities of observation, which are estimated through a specifically designed high-dimensional estimation procedure. Asymptotic convergence rates of the proposed estimators for both the observation probabilities and the target matrix are studied. The empirical performance of the proposed methodology is illustrated via both numerical experiments and a real data application.

Keywords: Low-rank missing mechanism; Missing data; Nuclear-norm regularization.

1 Introduction

The problem of recovering a high-dimensional matrix $\boldsymbol{A}_\star \in \mathbb{R}^{n_1 \times n_2}$ from very few (noisy) observations of its entries is commonly known as matrix completion, whose applications include, for examples, collaborative filtering, computer visions and positioning. From a statistical viewpoint, it is a high-dimensional missing data problem where a high percentage of matrix entries are missing. As in many missing data problems, the underlying missing (sampling/observation) mechanism

*Xiaojun Mao is Assistant Professor, School of Data Science, Fudan University, Shanghai 200433, China (Email: maoxj@fudan.edu.cn).

†Raymond K. W. Wong is Assistant Professor, Department of Statistics, Texas A&M University, College Station, TX 77843, USA (Email: raywong@stat.tamu.edu).

‡Song Xi Chen is Chair Professor, Department of Business Statistics and Econometrics, Guanghua School of Management and Center for Statistical Science, Peking University, Beijing 100651, China (Email: csx@gsm.pku.edu.cn).

plays an important role. Most existing work (e.g., Candès and Recht, 2009; Keshavan et al., 2009; Recht, 2011; Rohde and Tsybakov, 2011; Koltchinskii et al., 2011) adopt a uniform observation mechanism, where each entry has the same marginal probability of being observed. This leads to significant simplifications, and enables the domain to move forward rapidly with various theoretical breakthroughs in the last decade. However, the uniform mechanism is often unrealistic. Recent works (Negahban and Wainwright, 2012; Klopp, 2014; Cai and Zhou, 2016; Cai et al., 2016; Bi et al., 2017; Mao et al., 2018) have been devoted to relaxing such an restrictive assumption by adopting other missing structures. The usage of these settings hinges on strong prior knowledge of the underlying problems. At a high level, many of them utilize some special forms of low-rank structures, e.g., rank-1 structure. In this paper, we aim at recovering the target matrix $\boldsymbol{A}_\star$ under a flexible high-dimensional low-rank sampling structure. This is achieved by a weighted empirical risk minimization, with application of inverse probability weighting (IPW) (e.g., Schnabel et al., 2016; Mao et al., 2018) to adjust for the effect of non-uniform missingness.

Data arising in many applications of matrix completion, such as recommender systems, usually possesses complex "sampling" structure and its distribution are largely unknown. For example of a movie recommender system, some believe that users tend to rate movies that they prefer or dislike most, while often remain "silent" to other movies. Another example of the complex sampling regime is in the online merchandising, some users may purchase certain items regularly without often rating them, but evaluate products that they rarely buy with a higher chance. Similar to the widely adopted model that ratings are generated from a small number of hidden factors, it is reasonable to believe that the missingness is also governed by a small and possibly different set of hidden factors, which leads to a low-rank modeling of the missing structure.

Inspired by generalized linear models (GLM), we model the probabilities of observation $\boldsymbol{\Theta}_\star = (\theta_{\star,ij})_{i,j=1}^{n_1,n_2} \in (0,1)^{n_1 \times n_2}$ by a high-dimensional low-rank matrix $\boldsymbol{M}_\star = (m_{\star,ij})_{i,j=1}^{n_1,n_2} \in \mathbb{R}^{n_1 \times n_2}$ through a known function f. That means, on the entry-wise level, we have $\theta_{\star,ij} = f(m_{\star,ij})$. In GLM, the linear predictor $m_{\star,ij}$ is further modeled as a linear function of observed covariates.

However, to reflect difficulties to attain (appropriate and adequate) covariate information and the complexity in the modeling of $\boldsymbol{\Theta}_\star$ in some situations of the matrix completion, the predictor matrix $\boldsymbol{M}_\star$ is assumed completely hidden in this study. Despite $\boldsymbol{M}_\star$ being hidden, as demonstrated in this work, the low-rankness of $\boldsymbol{M}_\star$ together with the high dimensionality allows both identification and consistent estimation of $\boldsymbol{\Theta}_\star$, which facilitates IPW-based matrix completion. Motivated by the nature of matrix completions, we propose a novel parametrization $\boldsymbol{M}_\star = \mu_\star \mathbf{1}_{n_1}\mathbf{1}_{n_2}^\intercal + \boldsymbol{Z}_\star$ where $\boldsymbol{Z}_\star$ satisfies $\mathbf{1}_{n_1}^\intercal \boldsymbol{Z}_\star \mathbf{1}_{n_2} = 0$. Our proposal extends the work of Davenport et al. (2014), which aims to solve a binary matrix completion problem and pursues a different goal. Compared with Davenport et al. (2014), the proposed method does not regularize the estimation of $\mu_\star$, but only regularize the nuclear norm of the estimation of $\boldsymbol{Z}_\star$. This modification requires different algorithmic treatment. The underlying reason for such modification is to avoid bias caused by the nuclear-norm penalty.

There are three fundamental challenges that set our work aside from the existing works of matrix completion and the IPW-based estimator: (i) the high-dimensional nature of the sampling mechanism; (ii) the diminishing lower bound of the observation probabilities (as n_1, n_2 go to infinity) in common settings of matrix completion, and added issue to the instability of IPW; (iii) the effect of estimation error in IPW to the matrix completion procedure. Challenges (i) and (ii) are unique to our problem, and not found in the literature of missing data. The work related to Challenge (iii) is sparse in the literature of matrix completion. One notable example is Mao et al. (2018), which focuses on a low-dimensional parametric modeling of IPW with observable covariates.

We develop non-asymptotic upper bounds of the mean squared errors (MSE) for the proposed estimators of the observation probabilities and the target matrix. The theoretical analysis shows that the IPW-based matrix completion with the underlying probability $\theta_{\star,ij} = f(m_{\star,ij})$ offers a better upper bound than matrix completion that ignore the missing mechanism all together as in Klopp (2014). But, the IPW-based completed matrix using the aforementioned low-rank estimation of $\boldsymbol{M}_\star$ endures a slower convergence rate due to the high-dimensionality of $\boldsymbol{M}_\star$ and low levels of observation probabilities. Indeed, in many matrix completion problems, $\theta_L := \min_{i,j} \theta_{\star,i,j}$ often goes to zero when $n_1, n_2 \to \infty$. Not only does this inflate the estimation error of $\boldsymbol{\Theta}_\star$ but also leads to unstable inverse probability weights, which reduces the convergence rate. To circumvent this issue, we propose to re-estimate $\boldsymbol{Z}_\star$ by constraining the magnitude of its entries to a smaller threshold. Our result shows that the proposed constrained IPW estimator achieve the optimal rate (up to a logarithmic factor). We also compare the IPW-based completion based on the proposed constrained estimation, with the completion based on direct weight trimming (or winsorization), a known practice in the conventional missing value literature (e.g., Rubin, 2001; Kang and Schafer, 2007; Schafer and Kang, 2008) and show that the constrained estimation has both theoretical and empirical advantages.

The rest of the paper is organized as follows. The proposed model is constructed in Section 2. The corresponding estimation and computational algorithm for both the observation probabilities and the target matrix $\boldsymbol{A}_\star$ are developed in Section 3 and 4 separately, while the non-asymptotic upper bounds are given in Section 5. Numerical performances of the proposed method are illustrated in a simulation study in Section 6, and an application to a Yahoo! music rating dataset in Section 7. Concluding remarks are given in Section 8, while some technical details are delegated to a supplementary material.

2 Model and Method

2.1 General Setup

Let $\boldsymbol{A}_\star = (a_{\star,ij})_{i,j=1}^{n_1,n_2} \in \mathbb{R}^{n_1 \times n_2}$ be an unknown high-dimensional matrix of interest, and $\boldsymbol{Y} = (y_{ij})_{i,j=1}^{n_1,n_2}$ be a contaminated version of $\boldsymbol{A}_\star$ according to the following additive noise model:

$$y_{ij} = a_{\star,ij} + \epsilon_{ij}, \quad \text{for } i = 1, \ldots, n_1; j = 1, \ldots, n_2, \tag{2.1}$$

where $\{\epsilon_{ij}\}$ are independently distributed random errors with zero mean and finite variance. In the setting of matrix completion, only a portion of $\{y_{ij}\}$ is observed. For the (i,j)-th entry, define the sampling indicator $w_{ij} = 1$ if y_{ij} is observed, and 0 otherwise, and assume $\{\epsilon_{ij}\}$ are independent of

$\{w_{ij}\}$.

As for the sampling mechanism, we adopt a Bernoulli model where $\{w_{ij}\}$ are independent Bernoulli random variables with observation probabilities $\{\theta_{\star,ij}\}$, collectively denoted by a matrix $\boldsymbol{\Theta}_\star := (\theta_{\star,ij})_{i,j=1}^{n_1,n_2} \in (0,1)^{n_1\times n_2}$. Similar to generalized linear models (GLM), the observation probabilities can be expressed in terms of an unknown matrix $\boldsymbol{M}_\star = (m_{\star,ij})_{i,j=1}^{n_1,n_2} \in \mathbb{R}^{n_1\times n_2}$ and a pre-specified monotone and differentiable function $f : \mathbb{R} \to [0,1]$, i.e., $\theta_{\star,ij} = f(m_{\star,ij})$ for all i, j. The matrix $\boldsymbol{M}_\star$ plays the same role as a linear predictor in GLM, while the function f is an inverse link function. Two popular choices of f are inverse logit function $g(m) = e^m/(1+e^m)$ (logistic model) and the standard normal cumulative distribution function (probit model).

2.2 Low-rank Modeling of $\boldsymbol{A}_\star$ and $\boldsymbol{M}_\star$

The above setup is general. Without additional assumption, it is virtually impossible to recover the hidden feature matrix $\boldsymbol{M}_\star$ and also the target matrix $\boldsymbol{A}_\star$. A common and powerful assumption is that $\boldsymbol{A}_\star$ is a low-rank matrix, i.e., $\text{rank}(\boldsymbol{A}_\star) \ll \min\{n_1, n_2\}$. Take the Yahoo! Webscope data set (to be analyzed in Section 7) as an example. This data set contains a partially observed matrix of ratings from 15,400 users to 1000 songs, and the goal is to complete the rating matrix. The low-rank assumption reflects the belief that users' ratings are generated by a small number of factors, representing several standard preference profiles for songs. This viewpoint has been proven useful in the modeling of recommender systems (e.g., Candès and Plan, 2010; Cai et al., 2010).

The same idea could be adapted to the missing pattern, despite that the factors that induce the missingness may be different from those that generate the ratings. To this end, we assume $\boldsymbol{M}_\star$ is also low-rank, and in addition, it can be decomposed as

$$\boldsymbol{M}_\star = \mu_\star \boldsymbol{J} + \boldsymbol{Z}_\star \quad \text{where} \quad \mathbf{1}_{n_1}^\intercal \boldsymbol{Z}_\star \mathbf{1}_{n_2} = 0 \tag{2.2}$$

with $\mathbf{1}_n$ being a n-vector of ones, and $\boldsymbol{J} = \mathbf{1}_{n_1}\mathbf{1}_{n_2}^\intercal$. Here $\mu_\star$ is the mean of $\boldsymbol{M}_\star$, i.e, $\mu_\star = \mathbf{1}_{n_1}^\intercal \boldsymbol{M}_\star \mathbf{1}_{n_2}/(n_1 n_2)$. Although this parametrization holds for any matrix, this allows different treatments in the estimations of $\mu_\star$ and $\boldsymbol{Z}_\star$. See Section 3 for details. Further, the low-rank assumption of $\boldsymbol{M}_\star$ can be translated to the low-rank assumption of $\boldsymbol{Z}_\star$.

We note that the rank of $\boldsymbol{M}_\star$ is not the same as the rank of $\boldsymbol{\Theta}_\star$ due to the nonlinear transformation f. Generally, the low-rank structure of $\boldsymbol{M}_\star$ implies a specific low-dimensional nonlinear structure of $\boldsymbol{\Theta}_\star$. For a common high missingness scenario, most entries of $\boldsymbol{M}_\star$ are significantly negative, at where many common choices of the inverse link function can be well-approximated by a linear function. So our modeling can be regarded as a low-rank modeling of $\boldsymbol{\Theta}_\star$ in certain sense.

There are a few related but more specialized models. Srebro and Salakhutdinov (2010) and Negahban and Wainwright (2012) utilize an independent row and column sampling mechanism, leading to a rank-1 structure for $\boldsymbol{\Theta}_\star$. Cai et al. (2016) consider a block structure for $\boldsymbol{\Theta}_\star$ and hence $\boldsymbol{M}_\star$, which can be regarded as a special case of the low-rank modeling. Mao et al. (2018) considered the case when the missingness is dependent on observable covariates, and adopted a low-rank modeling with a known row space of $\boldsymbol{M}_\star$. The proposal in this paper is for the situation when the missingness is dependent on some hidden factors, which reflects situations when obvious covariates are unknown or not available.

2.3 IPW-based Matrix Completion: Motivations and Challenges

Write the Hadamard product as $\circ$ and the Frobenius norm as $\|\cdot\|_F$. To recover the target matrix $\boldsymbol{A}_\star$, many existing matrix completion techniques assume uniform missing structure and hence utilize an unweighted/uniform empirical risk function $\widehat{R}_{\mathrm{UNI}}(\boldsymbol{A}) = (n_1 n_2)^{-1}\|\boldsymbol{W}\circ(\boldsymbol{A}-\boldsymbol{Y})\|_F^2$ (e.g., Candès and Plan, 2010; Koltchinskii et al., 2011; Mazumder et al., 2010), which is an unbiased estimator of the risk $R(\boldsymbol{A}) := \mathsf{E}(\|\boldsymbol{A}-\boldsymbol{Y}\|_F^2)/(n_1 n_2)$ (up to a multiplicative constant) under uniform missingness. The work of Klopp (2014) is a notable exception that considers the use of $\widehat{R}_{\mathrm{UNI}}$ under non-uniform missingness.

For any matrix $\boldsymbol{B} = (b_{ij})_{i,j=1}^{n_1,n_2}$, we denote $\boldsymbol{B}^\dagger = (b_{ij}^{-1})_{i,j=1}^{n_1,n_2}$ and $\boldsymbol{B}^\ddagger = (b_{ij}^{-1/2})_{i,j=1}^{n_1,n_2}$. Under general missingness (uniform or non-uniform), one can show that, for any $\boldsymbol{A}\in\mathbb{R}^{n_1\times n_2}$,

$$R(\boldsymbol{A}) = \frac{1}{n_1 n_2}\mathsf{E}\left(\|\boldsymbol{A}-\boldsymbol{Y}\|_F^2\right) = \frac{1}{n_1 n_2}\mathsf{E}\left(\left\|\boldsymbol{W}\circ\boldsymbol{\Theta}_\star^\ddagger\circ(\boldsymbol{A}-\boldsymbol{Y})\right\|_F^2\right).$$

Clearly, $\boldsymbol{A}_\star$ uniquely minimizes the risk R. If $\boldsymbol{\Theta}$ were known, an unbiased estimator of R would be

$$\widehat{R}(\boldsymbol{A}) = \frac{1}{n_1 n_2} \left\| \boldsymbol{W} \circ \boldsymbol{\Theta}_\star^\ddagger \circ (\boldsymbol{A} - \boldsymbol{Y}) \right\|_F^2, \tag{2.3}$$

which involves IPW, and motivates the use of IPW in matrix completion as in Mao et al. (2018). In addition, our theoretical analysis shows that the nuclear-norm-regularized empirical risk estimator (to be defined in details later) based on $\widehat{R}$ (assuming the use of true observation probabilities) improves upon existing error upper bound of corresponding estimator based on $\widehat{R}_{\text{UNI}}$ achieved by Klopp (2014). See Section 5.3 for details. However, the inverse probability weights $\boldsymbol{\Theta}_\star^\ddagger$ are often unknown and have to be estimated in practice. The proposed estimation of $\boldsymbol{\Theta}_\star^\ddagger$ will be addressed carefully. Now we dicuss some challenges with the estimation particularly related to the matrix completion settings.

Despite the popularity of IPW in missing data literature, it is known to often produce unstable estimation due to occurrences of small probabilities (e.g., Rubin, 2001; Kang and Schafer, 2007; Schafer and Kang, 2008). This problematic scenario is indeed common for matrix completion problems where one intends to recover a target matrix from *very few* observations. Theoretically, a reasonable setup should allow some $\theta_{\star,ij}$ to go to zero as $n_1, n_2 \to \infty$, leading to diverging weights and a non-standard setup of IPW. Due to these observations, a careful construction of the estimation procedure is required.

For uniform sampling ($\theta_{\star,ij} \equiv \theta_0$ for some probability θ_0), one only has to worry about a small common probability θ_0 (or that θ_0 diminishes in an asymptotic sense.) Although small θ_0 increases the difficulty of estimation, R changes only up to a multiplicative constant. However, for non-uniform setting, it is not as straightforward due to the heterogeneity among $\{\theta_{\star,ij}\}$. To demonstrate the issue, we now briefly look at the Yahoo! Webscope dataset described in Section 7. One sign of the strong heterogeneity in $\{\theta_{\star,ij}\}$ is a large θ_U/θ_L, where $\theta_L := \min_{i,j} \theta_{\star,ij}$ and $\theta_U := \max_{i,j} \theta_{\star,ij}$. We found that the corresponding ratio of estimated probabilities $\widehat{\theta}_U/\widehat{\theta}_L$ based on the rank-1 structure of Negahban and Wainwright (2012) was 25656.2, and that based on our proposed method (without re-estimation, to be described below) was 23988.0. These huge ratios are signs of strong heterogeneity in the probability of the observation. From our analysis, strong heterogeneity could jeopardize the convergence rate of our estimator, and so will be addressed rigorously in our framework.

In the following, we propose an estimation of $\boldsymbol{\Theta}_\star$ in Section 3.1 and an appropriate modification in Section 3.3 which, when substituted into the empirical risk $\widehat{R}$, allows us to construct a stable estimator for $\boldsymbol{A}_\star$.

3 Estimation of $\boldsymbol{\Theta}_\star$

3.1 Regularized Maximum Likelihood Estimation

We develop the estimation of $\boldsymbol{\Theta}_\star$ based upon the framework of regularized maximum likelihood. Given the inverse of link function f, the log-likelihood function with respect to the indicator matrix $\boldsymbol{W} := (w_{ij}) \in \mathbb{R}^{n_1 \times n_2}$ is given by

$$\ell_{\boldsymbol{W}}\left(\boldsymbol{M}\right) := \sum_{i,j} \left\{ \mathbb{1}_{[w_{ij}=1]} \log\left(f\left(m_{ij}\right)\right) + \mathbb{1}_{[w_{ij}=0]} \log\left(1 - f\left(m_{ij}\right)\right) \right\},$$

for any $\boldsymbol{M} = (m_{ij})_{i,j=1}^{n_1,n_2} \in \mathbb{R}^{n_1 \times n_2}$, where $\mathbb{1}_{\mathcal{A}}$ is the indicator of an event $\mathcal{A}$. Due to the low-rank assumption of $\boldsymbol{M}_\star$, one natural candidate of estimators is the maximizer of the regularized log-likelihood $\ell_{\boldsymbol{W}}(\boldsymbol{M}) - \lambda\|\boldsymbol{M}\|_*$, where $\|\cdot\|_*$ represents the nuclear norm and $\lambda > 0$ is a tuning parameter. It is also common to enforce an additional max-norm constraint $\|\boldsymbol{M}\|_\infty \leq \alpha$ for some $\alpha > 0$ in the maximization (e.g., Davenport et al., 2014). Note that the nuclear norm penalty flavors $\boldsymbol{M} = \boldsymbol{0}$, corresponding to that $\Pr(w_{ij} = 1) = 0.5$ for all i, j. Nevertheless, this does not align well with common settings of matrix completion under which the average probability of observations is small, and hence results in a large bias. In view of this, we instead adopt the following estimator of $(\mu_\star, \boldsymbol{Z}_\star)$:

$$\left(\widehat{\mu}, \widehat{\boldsymbol{Z}}\right) = \underset{(\mu, \boldsymbol{Z}) \in \mathcal{C}_{n_1,n_2}(\alpha_1,\alpha_2)}{\arg\max} \ \ell_{\boldsymbol{W}}\left(\mu \boldsymbol{J} + \boldsymbol{Z}\right) - \lambda \left\|\boldsymbol{Z}\right\|_*, \text{where} \tag{3.1}$$

$$\mathcal{C}_{n_1,n_2}\left(\alpha_1, \alpha_2\right) := \{(\mu, \boldsymbol{Z}) \in \mathbb{R} \times \mathbb{R}^{n_1 \times n_2} : |\mu| \leq \alpha_1,\ \left\|\boldsymbol{Z}\right\|_\infty \leq \alpha_2,\ \boldsymbol{1}_{n_1}^\intercal \boldsymbol{Z} \boldsymbol{1}_{n_2} = 0\}.$$

Note that the mean μ of the linear predictor $\mu\boldsymbol{J}+\boldsymbol{Z}$ is not penalized. With $(\widehat{\mu}, \widehat{\boldsymbol{Z}})$, the corresponding estimator of $\boldsymbol{M}_\star$ is defined as $\widehat{\boldsymbol{M}} = \widehat{\mu}\boldsymbol{J}+\widehat{\boldsymbol{Z}}$. The constraint $\mathbf{1}_{n_1}^\intercal \boldsymbol{Z}\mathbf{1}_{n_2} = 0$ ensures the identifiability of μ and $\boldsymbol{Z}$. Apparently, the constraints in $\mathcal{C}_{n_1,n_2}(\alpha_1, \alpha_2)$ are analogous to $\|\boldsymbol{M}\|_\infty \le \alpha_0$, where $\alpha_0 = \alpha_1 + \alpha_2$, but on the parameters μ and $\boldsymbol{Z}$ respectively.

Davenport et al. (2014) considered a regularized maximum likelihood approach for a binary matrix completion problem. But their goal was different from ours, as they aimed to recover a binary rating matrix in lieu of the missing structure, they considered a regularization on $\boldsymbol{M}$ (instead of $\boldsymbol{Z}$) via $\|\boldsymbol{M}\|_* \le \alpha' \sqrt{\mathrm{rank}(\boldsymbol{M}_\star) n_1 n_2}$. In addition, this constraint required a prior knowledge of the true rank of $\boldsymbol{M}_\star$, which is not required in our proposed method (3.1). As for the scaling parameter α', Davenport et al. (2014) considered an α' independent of the dimensions n_1 and n_2 to restrict the "spikiness" of $\boldsymbol{M}$. As explained earlier, in our framework, θ_L should be allowed to go to zero as $n_1, n_2 \to \infty$. To this end, we allow α_1 and α_2 to depend on the dimensions n_1 and n_2. See more details in Section 5.

3.2 Computational algorithm and tuning parameter selection

To solve the optimization (3.1), we begin with the observation that $\ell_{\boldsymbol{W}}$ is a smooth concave function, which allows the usage of an iterative algorithm called accelerated proximal gradient algorithm (Beck and Teboulle, 2009). Given a pair $(\mu_{\text{old}}, \boldsymbol{Z}_{\text{old}})$ from a previous iteration, a quadratic approximation of the objective function $-\ell_{\boldsymbol{W}}(\mu\boldsymbol{J} + \boldsymbol{Z}) + \lambda\|\boldsymbol{Z}\|_*$ is formed:

$$\begin{aligned} P_L\left\{(\mu, \boldsymbol{Z}), (\mu_{\text{old}}, \boldsymbol{Z}_{\text{old}})\right\} := &- \ell_{\boldsymbol{W}}\left(\mu_{\text{old}}\boldsymbol{J} + \boldsymbol{Z}_{\text{old}}\right) \\ &+ (\mu - \mu_{\text{old}})\, \mathbf{1}_{n_1}^\intercal \left(-\nabla_\mu \ell_{\boldsymbol{W}}\left(\mu_{\text{old}}\boldsymbol{J} + \boldsymbol{Z}_{\text{old}}\right)\right) \mathbf{1}_{n_2} + \frac{L n_1 n_2}{2} (\mu - \mu_{\text{old}})^2 \\ &+ \langle \boldsymbol{Z} - \boldsymbol{Z}_{\text{old}}, -\nabla_{\boldsymbol{Z}} \ell_{\boldsymbol{W}}\left(\mu_{\text{old}}\boldsymbol{J} + \boldsymbol{Z}_{\text{old}}\right)\rangle + \frac{L}{2}\left\|\boldsymbol{Z} - \boldsymbol{Z}_{\text{old}}\right\|_F^2 + \lambda \left\|\boldsymbol{Z}\right\|_*, \end{aligned}$$

where $L > 0$ is an algorithmic parameter determining the step size of the proximal gradient algorithm, and is chosen by a backtracking method (Beck and Teboulle, 2009). Here $\langle \boldsymbol{B}, \boldsymbol{C}\rangle = \sum_{i,j} b_{ij} c_{ij}$ for any matrices $\boldsymbol{B} = (b_{ij})$ and $\boldsymbol{C} = (c_{ij})$ of same dimensions.

In this iterative algorithm, a successive update of $(\mu, \boldsymbol{Z})$ can be obtained by

$$\underset{(\mu,\boldsymbol{Z})\in\mathcal{C}_{n_1,n_2}(\alpha_1,\alpha_2)}{\arg\min} \quad P_L\left\{(\mu,\boldsymbol{Z}),(\mu_{\text{old}},\boldsymbol{Z}_{\text{old}})\right\},$$

where the optimization with respect to μ and $\boldsymbol{Z}$ can be performed separately. For μ, one can derive a closed-form update

$$\min\{\alpha_1, \max\{-\alpha_1, \mu_{\text{old}} + (Ln_1n_2)^{-1}\mathbf{1}_{n_1}^{\intercal}(-\nabla_{\mu}\ell_{\boldsymbol{W}}(\mu_{\text{old}}\boldsymbol{J}+\boldsymbol{Z}_{\text{old}}))\mathbf{1}_{n_2}\}\}.$$

As for $\boldsymbol{Z}$, we need to perform the minimization

$$\underset{\|\boldsymbol{Z}\|_\infty\le\alpha_2,\,\mathbf{1}_{n_1}^{\intercal}\boldsymbol{Z}\mathbf{1}_{n_2}=0}{\arg\min} \quad \langle \boldsymbol{Z}-\boldsymbol{Z}_{\text{old}}, -\nabla_{\boldsymbol{Z}}\ell_{\boldsymbol{W}}\left(\mu_{\text{old}}\boldsymbol{J}+\boldsymbol{Z}_{\text{old}}\right)\rangle + \frac{L}{2}\left\|\boldsymbol{Z}-\boldsymbol{Z}_{\text{old}}\right\|_F^2 + \lambda\left\|\boldsymbol{Z}\right\|_*,$$

which is equivalent to

$$\underset{\|\boldsymbol{Z}\|_\infty\le\alpha_2,\,\mathbf{1}_{n_1}^{\intercal}\boldsymbol{Z}\mathbf{1}_{n_2}=0}{\arg\min} \quad \frac{1}{2}\left\|\boldsymbol{Z}-\boldsymbol{Z}_{\text{old}}-\frac{1}{L}\nabla_{\boldsymbol{Z}}\ell_{\boldsymbol{W}}\left(\mu_{\text{old}}\boldsymbol{J}+\boldsymbol{Z}_{\text{old}}\right)\right\|_F^2 + \frac{\lambda}{L}\left\|\boldsymbol{Z}\right\|_*. \tag{3.2}$$

We apply a three-block extension of the alternative direction method of multipliers (ADMM) (Chen et al., 2016) to an equivalent form of (3.2):

$$\underset{\boldsymbol{Z}=\boldsymbol{G}_1=\boldsymbol{G}_2,\,\mathbf{1}_{n_1}^{\intercal}\boldsymbol{G}_1\mathbf{1}_{n_2}=0,\,\|\boldsymbol{G}_2\|_\infty\le\alpha_2}{\arg\min} \quad \frac{\lambda}{L}\left\|\boldsymbol{Z}\right\|_* + \frac{1}{2}\left\|\boldsymbol{G}_2-\boldsymbol{Z}_{\text{old}}-\frac{1}{L}\nabla_{\boldsymbol{Z}}\ell_{\boldsymbol{W}}\left(\mu_{\text{old}}\boldsymbol{J}+\boldsymbol{Z}_{\text{old}}\right)\right\|_F^2. \tag{3.3}$$

Write $\boldsymbol{H} = (\boldsymbol{H}_1, \boldsymbol{H}_2)$. The augmented Lagrangian for (3.3) is given by

$$\begin{aligned}\mathcal{L}_u\left(\boldsymbol{Z},\boldsymbol{G}_1,\boldsymbol{G}_2;\boldsymbol{H}\right) =& \frac{\lambda}{L}\left\|\boldsymbol{Z}\right\|_* + \frac{1}{2}\left\|\boldsymbol{G}_2-\boldsymbol{Z}_{\text{old}}-\frac{1}{L}\nabla_{\boldsymbol{Z}}\ell_{\boldsymbol{W}}\left(\mu_{\text{old}}\boldsymbol{J}+\boldsymbol{Z}_{\text{old}}\right)\right\|_F^2 \\ &- \langle\boldsymbol{H}_1,\boldsymbol{Z}-\boldsymbol{G}_1\rangle - \langle\boldsymbol{H}_2,\boldsymbol{Z}-\boldsymbol{G}_2\rangle + \frac{u}{2}\left\|\boldsymbol{Z}-\boldsymbol{G}_1\right\|_F^2 + \frac{u}{2}\left\|\boldsymbol{Z}-\boldsymbol{G}_2\right\|_F^2 \\ &+ \mathbb{I}_{[\{\mathbf{1}_{n_1}^{\intercal}\boldsymbol{G}_1\mathbf{1}_{n_2}=0]} + \mathbb{I}_{[\|\boldsymbol{G}_2\|_\infty\le\alpha_2]},\end{aligned}$$

where $u > 0$ is an algorithmic parameter and, $\mathbb{I}_{\mathcal{A}} = 0$ if the constraint $\mathcal{A}$ holds and ∞ otherwise. The detailed algorithm to solve (3.3) is summarized in Algorithm 1. It is noted that, in general, the multi-block ADMM may fail to converge for some $u > 0$ (Chen et al., 2016). In those cases, an appropriate selection of u is crucial. However, we are able to show that the form of our ADMM algorithm belongs to a special class (Chen et al., 2016) in which convergence is guaranteed for any $u > 0$. Therefore, we simply set $u = 1$. We summarize the corresponding convergence result in the following theorem whose proof is provided in the supplementary material.

Theorem 1. *The sequence* $\{\boldsymbol{Z}^{(k)}, \boldsymbol{G}_1^{(k)}, \boldsymbol{G}_2^{(k)}\}$*, generated by Algorithm 1, converges to the solution of* (3.3)*.*

Notice that the ADMM algorithm is nested within the proximal gradient algorithm. But, from our practical experiences, both the number of inner iterations (ADMM) and outer iterations (proximal gradient) are small, usually less than twenty in our numerical experiments.

We now discuss the choice of tuning parameters. For α_1 and α_2, they can be chosen according to prior knowledge of the problem setup, if available. In practice when prior knowledge is not available, one can choose large values for these parameters. Once these parameters are large enough, our method is not sensitive to their specific values. A more principled way to tune α_1 and α_2 is a challenging problem and beyond the scope of this work. As for λ, we adopt Akaike information criterion (AIC) where the degree of freedom is approximated by $r_{\widehat{\boldsymbol{M}}}(n_1 + n_2 - r_{\widehat{\boldsymbol{M}}})$.

3.3 Constrained estimation

To use $\widehat{R}$ of (2.3), a naive idea is to obtain $\widehat{\boldsymbol{\Theta}} = (\hat{\theta}_{ij})_{i,j=1}^{n_1,n_2} := \mathcal{F}(\widehat{\boldsymbol{M}})$, where $\mathcal{F}$ is an operator defined by $\mathcal{F}(\boldsymbol{M}) = (f(m_{ij}))_{i,j=1}^{n_1,n_2} \in \mathbb{R}^{n_1 \times n_2}$ for any $\boldsymbol{M} = (m_{ij})_{i,j=1}^{n_1,n_2} \in \mathbb{R}^{n_1 \times n_2}$, and then replace $\boldsymbol{\Theta}_\star^\ddagger$ by $\widehat{\boldsymbol{\Theta}}^\ddagger := (\hat{\theta}_{ij}^{-1/2})_{i,j=1}^{n_1,n_2}$. However, this direct implementation is not robust to extremely small probabilities of observation, and our theoretical analysis shows that this could lead to slower convergence rate of the estimator of $\boldsymbol{A}_\star$. In the literature of missing data, a simple solution to robustify is weight truncation (e.g., Potter, 1990; Scharfstein et al., 1999), i.e., winsorizing small probabilities.

In the estimation of $\widehat{\boldsymbol{\Theta}}$ defined in (3.1), assuming $\|\boldsymbol{Z}_\star\|_\infty \leq \alpha_2$, a large α_2 has an adverse effect on the estimation. In the setting of diverging α_2 (due to diminishing θ_L), the convergence rate of $\widehat{\boldsymbol{Z}}$ becomes slower and the estimator obtained after direct winsorization will also be affected. That is, even though the extreme probabilities could be controlled by winsorizing, the unchanged entries of $\widehat{\boldsymbol{Z}}$ (in the procedure of winsorizing) may already suffer from a slower rate of convergence. This results in a larger estimation error bound under certain settings of missingness, which will be discussed theoretically in Section 5.

Algorithm 1 The ADMM used to solve (3.3)

```
1: Initialize $k = 0$, and select $u$, $\boldsymbol{H}^{(k)}$, $\boldsymbol{Z}^{(k)}$, $\boldsymbol{G}_1^{(k)}$, $\boldsymbol{G}_2^{(k)}$ such that $\boldsymbol{Z}^{(k)}$ is a solution of (3.3) without constraints, $\mathbf{1}_{n_1}^\intercal \boldsymbol{G}_1^{(k)} \mathbf{1}_{n_2} = 0$ and $\|\boldsymbol{G}_2^{(k)}\|_\infty \le \alpha_2$.
2: Minimize $\mathcal{L}_u(\boldsymbol{Z}, \boldsymbol{G}_1^{(k)}, \boldsymbol{G}_2^{(k)}; \boldsymbol{H}^{(k)})$ with respect to $\boldsymbol{Z}$:
       $\boldsymbol{Z}^{(k+1)} = \mathcal{SVT}_{(uL)^{-1}\lambda}\{1/2(\boldsymbol{G}_1^{(k)} + \boldsymbol{G}_2^{(k)} + 1/u\boldsymbol{H}_1^{(k)} + 1/u\boldsymbol{H}_2^{(k)})\}.$
   Here $\mathcal{SVT}_c$ is the singular value soft-thresholding operator defined as
       $\mathcal{SVT}_c(\boldsymbol{D}) = \boldsymbol{U}\mathrm{diag}(\{(\sigma_i - c)_+\})\boldsymbol{V}^\intercal \quad \text{for any } c \ge 0,$
   where $x_+ = \max(x, 0)$, and $\boldsymbol{U\Sigma V}^\intercal$, with $\boldsymbol{\Sigma} = \mathrm{diag}(\{\sigma_i\})$, is the singular value decomposition (SVD) of a matrix $\boldsymbol{D}$.
3: Minimize $\mathcal{L}_u(\boldsymbol{Z}^{(k+1)}, \boldsymbol{G}_1, \boldsymbol{G}_2^{(k)}; \boldsymbol{H}^{(k)})$ with respect to $\boldsymbol{G}_1$:
       $\boldsymbol{G}_1^{(k+1)} = \mathop{\arg\min}_{\mathbf{1}_{n_1}^\intercal \boldsymbol{G}_1 \mathbf{1}_{n_2} = 0} \ \frac{1}{2}\left\|\boldsymbol{G}_1 - \left(\boldsymbol{Z}^{(k+1)} - 1/u\boldsymbol{H}_1^{(k)}\right)\right\|_F^2,$
   Let $\boldsymbol{B}_1 = \boldsymbol{Z}^{(k+1)} - 1/u\boldsymbol{H}_1^{(k)}$ and simplifies to $\boldsymbol{G}_1^{(k+1)} = \boldsymbol{B}_1 - (n_1 n_2)^{-1}\mathbf{1}_{n_1}^\intercal \boldsymbol{B}_1 \mathbf{1}_{n_2} \boldsymbol{J}$.
4: Minimize $\mathcal{L}_u(\boldsymbol{Z}^{(k+1)}, \boldsymbol{G}_1^{(k+1)}, \boldsymbol{G}_2; \boldsymbol{H}^{(k)})$ with respect to $\boldsymbol{G}_2$:
       $\boldsymbol{G}_2^{(k+1)} = \mathop{\arg\min}_{\|\boldsymbol{G}_2\|_\infty \le \alpha_2} \ \left\|\boldsymbol{G}_2 - \left\{\boldsymbol{Z}_{\mathrm{old}} + \frac{1}{L}\nabla_{\boldsymbol{Z}}\ell_{\boldsymbol{W}}\left(\mu_{\mathrm{old}}\boldsymbol{J} + \boldsymbol{Z}_{\mathrm{old}}\right) - \boldsymbol{H}_2^{(k)} + u\boldsymbol{Z}^{(k+1)}\right\}/(1+u)\right\|_F^2.$
   Let $\boldsymbol{B}_2 = \{\boldsymbol{Z}_{\mathrm{old}} + \frac{1}{L}\nabla_{\boldsymbol{Z}}\ell_{\boldsymbol{W}}(\mu_{\mathrm{old}}\boldsymbol{J} + \boldsymbol{Z}_{\mathrm{old}}) - \boldsymbol{H}_2^{(k)} + u\boldsymbol{Z}^{(k+1)}\}/(1+u)$ and simplifies to $\boldsymbol{G}_2^{(k+1)}(i,j) = \min\{\alpha_2, \max\{-\alpha_2, \boldsymbol{B}_2(i,j)\}\}$.
5: Update the dual variable $\boldsymbol{H}^{(k+1)} = (\boldsymbol{H}_1^{(k+1)\intercal}, \boldsymbol{H}_2^{(k+1)\intercal})^\intercal$ by
       $\boldsymbol{H}_1^{(k+1)} = \boldsymbol{H}_1^{(k)} - u(\boldsymbol{Z}^{(k+1)} - \boldsymbol{G}_1^{(k+1)}) \quad \text{and} \quad \boldsymbol{H}_2^{(k+1)} = \boldsymbol{H}_2^{(k)} - u(\boldsymbol{Z}^{(k+1)} - \boldsymbol{G}_2^{(k+1)}).$
6: Return $\boldsymbol{Z} = \boldsymbol{Z}^{(k+1)}$ if converged. Otherwise, increment $k$ and repeat Steps 2-6.
```

A seemingly better strategy is to impose a tighter constraint directly in the minimization problem (3.1). That is to adopt the constraint $\|\boldsymbol{Z}\|_\infty \leq \beta$ where $0 \leq \beta \leq \alpha_2$. Theoretically, one can better control the errors on those entries of magnitude smaller than β. However, the mean-zero constraint of $\boldsymbol{Z}$ no longer makes sense as the constraint $\|\boldsymbol{Z}\|_\infty \leq \beta$ may have shifted the mean.

We propose a re-estimation of $\boldsymbol{Z}_\star$ with a different constraint level β:

$$\widehat{\boldsymbol{Z}}_\beta = \underset{\boldsymbol{Z}\in\mathbb{R}^{n_1\times n_2}}{\arg\max}\, \ell_{\boldsymbol{W}}\left(\widehat{\mu}\boldsymbol{J} + \boldsymbol{Z}\right) - \lambda' \left\|\boldsymbol{Z}\right\|_* \quad \text{subject to} \quad \left\|\boldsymbol{Z}\right\|_\infty \leq \beta. \tag{3.4}$$

Note that we only re-compute $\boldsymbol{Z}$ but not μ, which allows us to drop the mean-zero constraint. Thus we have $\widehat{\boldsymbol{M}}_\beta = \widehat{\mu}\boldsymbol{J} + \widehat{\boldsymbol{Z}}_\beta$. The corresponding algorithm for optimization (3.4) can be derived similarly as in Davenport et al. (2014), and is provided in Section **??** of the supplementary material. Similar to many IPW method, the tuning of β is a challenging problem. In what follows, we write $\widehat{\boldsymbol{\Theta}} = \mathcal{F}(\widehat{\boldsymbol{M}})$ and $\widehat{\boldsymbol{\Theta}}_\beta = \mathcal{F}(\widehat{\boldsymbol{M}}_\beta)$.

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

Now, we come back to (2.3) and replace $\boldsymbol{\Theta}_\star^\ddagger$ by $\widehat{\boldsymbol{\Theta}}_\beta^\ddagger$ to obtain a modified empirical risk:

$$\widetilde{R}\left(\boldsymbol{A}\right) = \frac{1}{n_1 n_2}\left\|\boldsymbol{W} \circ \widehat{\boldsymbol{\Theta}}_\beta^\ddagger \circ \left(\boldsymbol{A} - \boldsymbol{Y}\right)\right\|_F^2, \tag{4.1}$$

where $\widehat{\boldsymbol{\Theta}}_\beta^\ddagger := (\widehat{\theta}_{ij,\beta}^{-1/2}) \in \mathbb{R}^{n_1\times n_2}$. Since $\boldsymbol{A}$ is a high-dimensional parameter, a direct minimization of $\hat{R}^*$ would often result in over-fitting. To circumvent this issue, we consider a regularized version:

$$\widetilde{R}\left(\boldsymbol{A}\right) + \tau \left\|\boldsymbol{A}\right\|_*, \tag{4.2}$$

where $\tau > 0$ is a regularization parameter. Again, the nuclear norm regularization encourages low-rank solution. Based on (4.2), our estimator of $\boldsymbol{A}_\star$ is defined as

$$\widehat{\boldsymbol{A}}_\beta = \underset{\|\boldsymbol{A}\|_\infty \leq a}{\arg\min}\left\{\frac{1}{n_1 n_2}\left\|\boldsymbol{W} \circ \widehat{\boldsymbol{\Theta}}_\beta^\ddagger \circ \left(\boldsymbol{A} - \boldsymbol{Y}\right)\right\|_F^2 + \tau \left\|\boldsymbol{A}\right\|_*\right\}, \tag{4.3}$$

where a is an upper bound on $\|\boldsymbol{A}_\star\|_\infty$. The above $\widehat{\boldsymbol{A}}_\beta$ contains as special cases (i) the matrix completion $\widehat{\boldsymbol{A}}_{\alpha_2}$, with unconstrained probability estimator $\widehat{\boldsymbol{\Theta}}$, by setting $\beta = \alpha_2$ and (ii) the estimator $\widehat{\boldsymbol{A}}_\beta$, with constrained probability estimator $\widehat{\boldsymbol{\Theta}}_\beta$, when $\beta < \alpha_2$.

We use an accelerated proximal gradient algorithm (Beck and Teboulle, 2009) to solve (4.3). For the choice of tuning parameter τ in (4.3), we adopt a 5-fold cross-validation to select the remaining tuning parameters. Due to the non-uniform missing mechanism, we use a weighted version of validation errors. The specific details are shown in Algorithm 2.

Algorithm 2 Estimation of target matrix $\widehat{\boldsymbol{A}}_\beta$.

```
Input covariate matrix $\boldsymbol{A}$, incomplete data matrix $\boldsymbol{Y}$, estimated probability matrices $\widehat{\boldsymbol{\Theta}}_\beta$ (or $\widehat{\boldsymbol{\Theta}}$), tuning parameter candidates $\tau^{(1)}, \ldots, \tau^{(k)}$, where $k$ is the grid length used for the search of parameter $\tau$ and a $k$ evaluation matrix $\boldsymbol{Q} = (Q_{ij})$ to be $\boldsymbol{Q} = \mathbf{0}$.
Randomly partition the observed entries of $\boldsymbol{Y}$ into 5 equal sized subsamples. These subsamples are used in turn as a test set. When subsample $l$ is used as test data, the remaining 4 subsamples are used as training data. Denote the corresponding indicator matrix of test data by $\boldsymbol{W}_*^{(l)}$ and that of training data by $\boldsymbol{W}^{(l)}$.
For each $i = 1, \ldots, k_1$ and $l = 1, \ldots, 5$, calculate $\widehat{\boldsymbol{A}}_\beta^{(l),\tau^{(i)}}$ by plugging $\boldsymbol{W}^{(l)}$ and $\tau^{(i)}$ in (4.3).
For $i = 1, \ldots, k$, $Q_i = \sum_{l=1}^5 \|\boldsymbol{W}_*^{(l)} \circ \boldsymbol{\Theta}_\beta^\ddagger \circ (\widehat{\boldsymbol{A}}_\beta^{(l),\tau^{(i)}} - \boldsymbol{Y})\|_F^2$.
Output the best parameters $\tau^{(j)}$ that minimize $Q_i$ among the entries of $\boldsymbol{Q}$.
Calculate $\widehat{\boldsymbol{A}}_\beta^{\tau^{(j)}}$ by plugging $\boldsymbol{W}$ and $\tau^{(j)}$ in (4.3).
```

5 Theoretical Properties

Let $\|\boldsymbol{B}\| = \sigma_{\max}(\boldsymbol{B})$, $\|\boldsymbol{B}\|_\infty = \max_{i,j} |b_{ij}|$ and $\|\boldsymbol{B}\|_{\infty,2} = \sqrt{\max_i \sum_j b_{ij}^2}$ be the spectral norm, the maximum norm and $l_{\infty,2}$-norm of a matrix $\boldsymbol{B}$ respectively. We use the symbol $\asymp$ to represent the asymptotic equivalence in order, i.e., $a_n \asymp b_n$ is equivalent to $a_n = O(b_n)$ and $b_n = O(a_n)$. We define the average squared distance between two matrices $\boldsymbol{B}, \boldsymbol{C} \in \mathbb{R}^{n_1 \times n_2}$ as $d^2(\boldsymbol{B}, \boldsymbol{C}) = \|\boldsymbol{B} - \boldsymbol{C}\|_F^2/(n_1 n_2)$. The average squared errors of $\widehat{\boldsymbol{M}}_\beta$ and $\widehat{\boldsymbol{\Theta}}_\beta^\dagger$ are then defined as $d^2(\widehat{\boldsymbol{M}}_\beta, \boldsymbol{M}_\star)$ and $d^2(\widehat{\boldsymbol{\Theta}}_\beta^\dagger, \boldsymbol{\Theta}_\star^\dagger)$ respectively. To measure the similarity between two probability matrices, we adopt the Hellinger distance d_H defined as follows. For any two matrices $\boldsymbol{S}, \boldsymbol{T} \in [0,1]^{n_1 \times n_2}$, $d_H^2(\boldsymbol{S}, \boldsymbol{T}) = (n_1 n_2)^{-1} \sum_{i,j} d_H^2(s_{ij}, t_{ij})$ where $d_H^2(s,t) = \left(\sqrt{s} - \sqrt{t}\right)^2 + \left(\sqrt{1-s} - \sqrt{1-t}\right)^2$ for $s, t \in [0,1]$. In the literature of matrix completion, most discussions related to optimal convergence rate are only accurate up to certain polynomial orders of $\log(n)$. For convenience, we use the notation polylog(n) to represent some polynomial of $\log(n)$.

5.1 Probabilities of observation

In this subsection, we investigate the asymptotic properties of $\widehat{\boldsymbol{M}}_\beta$ and $\widehat{\boldsymbol{\Theta}}_\beta^\dagger$ defined in Section 3. To this end, we introduce the following conditions on the missing structure.

C1. The indicators $\{w_{ij}\}_{i,j=1}^{n_1,n_2}$ are mutually independent, and independent of $\{\epsilon_{ij}\}_{i,j=1}^{n_1,n_2}$. For $i = 1, \ldots, n_1$ and $j = 1, \ldots, n_2$, w_{ij} follows a Bernoulli distribution with probability of success $\theta_{\star,ij} = f(m_{\star,ij}) \in (0, 1)$. Furthermore, f is monotonic increasing and differentiable.

C2. The hidden feature matrix $\boldsymbol{M}_\star = \mu_\star \boldsymbol{J} + \boldsymbol{Z}_\star$ where $\mathbf{1}_{n_1}^T \boldsymbol{Z}_\star \mathbf{1}_{n_2} = 0$, $|\mu_\star| \le \alpha_1 < \infty$ and $\|\boldsymbol{Z}_\star\|_\infty \le \alpha_2 < \infty$. Here α_1 and α_2 are allowed to depend on the dimensions n_1 and n_2. This also implies that there exists a lower bound $\theta_L \in (0, 1)$ (allowed to depend on n_1, n_2) such that $\min_{i,j}\{\theta_{ij}\} \ge \theta_L \ge f(-\alpha_1 - \alpha_2) > 0$.

For clear presentation, we assume $n_1 = n_2 = n$ and choose the logit function as the inverse link function f in the rest of Section 5, while corresponding results under general settings of n_1, n_2 and f are delegated to Section S1.2 in the supplementary material. We first establish the convergence results for $\widehat{\mu}$, $\widehat{\boldsymbol{Z}}$ and $\widehat{\boldsymbol{M}}$, respectively. To simplify notations, let $\alpha_0 = \alpha_1 + \alpha_2$ and $h_{\alpha_1,\beta} = (1 + e^{\alpha_1+\beta})^{-1}$.

Theorem 2. *Suppose Conditions C1-C2 hold, and $(\mu_\star, \boldsymbol{Z}_\star) \in \mathcal{C}_{n_1,n_2}(\alpha_1, \alpha_2)$. Consider $\widehat{\boldsymbol{M}} = \widehat{\mu}\boldsymbol{J} + \widehat{\boldsymbol{Z}}$ where $(\widehat{\mu}, \widehat{\boldsymbol{Z}})$ is the solution to* (3.1). *There exist positive constants C_1, C_2 such that for $\lambda \ge (8e + 1)\sqrt{n}$, we have with probability at least $1 - C_1/n$,*

$$(\mu_\star - \widehat{\mu})^2 \le C_2\left(\alpha_1^2 \wedge \Gamma_{n,\lambda}\right), \quad \frac{1}{n_1 n_2}\left\|\widehat{\boldsymbol{Z}} - \boldsymbol{Z}_\star\right\|_F^2 \le C_2\left(\alpha_2^2 \wedge \Gamma_{n,\lambda}\right)$$
$$\text{and} \quad \frac{1}{n_1 n_2}\left\|\widehat{\boldsymbol{M}} - \boldsymbol{M}_\star\right\|_F^2 \le C_2\left(\alpha_0^2 \wedge \Gamma_{n,\lambda}\right), \tag{5.1}$$

where

$$\Gamma_{n,\lambda} := \min\left\{\frac{e^{\alpha_0}\lambda}{n^2}\|\boldsymbol{Z}_\star\|_*, \frac{r_{\boldsymbol{Z}_\star} e^{2\alpha_0}\lambda^2}{n^2}\right\}.$$

The three upper bounds in (5.1) all consist of a trivial bound α_j^2 and a more dedicated bound $\Gamma_{n,\lambda}$. The trivial upper bounds α_1^2, α_2^2 and α_0^2 can be easily derived from the constraint set $\mathcal{C}_{n_1,n_2}(\alpha_1, \alpha_2)$. For extreme settings of increasing α_0, the more dedicated bound $\Gamma_{n,\lambda}$ is diverging and the trivial bounds may provide better control. For example, under the extreme scenario $\theta_L \asymp n^{-1}\text{polylog}(n)$ where the target matrix is still recoverable (Candès and Plan, 2010; Koltchinskii et al., 2011; Mao et al., 2018), we have $\alpha_1 + \alpha_2 = \alpha_0 \geq -\log \theta_L \asymp \text{polylog}(n)$. Then $\alpha_k = o(n^{-1/4}\|\boldsymbol{Z}_\star\|_*^{1/2})$ and $\alpha_k = o(r_{\boldsymbol{Z}_\star} n)$ for $k = 0, 1, 2$ which implies the trivial bounds are of the smallest order compared with $\Gamma_{n,\lambda}$. Thus it is necessary that we keep these trivial upper bounds α_k in the right hand sides of (5.1). The term $\Gamma_{n,\lambda}$ can be controlled by either the nuclear norm and the rank of $\boldsymbol{Z}_\star$. For a range of non-extreme scenarios, i.e., $\alpha_0 \leq 1/2 \log n$ or $\theta_L \geq n^{-1/2}$, the second term in $\Gamma_{n,\lambda}$ achieves the smallest order once $r_{\boldsymbol{Z}_\star} = O(e^{-\alpha_0} n^{-1/2}\|\boldsymbol{Z}_\star\|_*)$.

We now consider the constrained estimation for $\boldsymbol{Z}_\star$, $\boldsymbol{M}_\star$ and $\boldsymbol{\Theta}_\star^\dagger$. For any matrix $\boldsymbol{B} = (b_{ij})_{i,j=1}^{n_1,n_2}$, define the winsorizing operator $\mathcal{T}_\beta$ by $\mathcal{T}_\beta(\boldsymbol{B}) = (T_\beta(b_{ij}))$ where

$$T_\beta(b_{ij}) = b_{ij}\mathbb{1}_{[-\beta \leq b_{ij} \leq \beta]} + \beta\mathbb{1}_{[b_{ij} > \beta]} - \beta\mathbb{1}_{[b_{ij} < -\beta]} \quad \text{for any } \beta \geq 0. \tag{5.2}$$

Write $\boldsymbol{M}_{\star,\beta} = \mu_\star \boldsymbol{J} + \mathcal{T}_\beta(\boldsymbol{Z}_\star)$ and $\widehat{\boldsymbol{M}}_{\star,\beta} = \widehat{\mu}\boldsymbol{J} + \mathcal{T}_\beta(\boldsymbol{Z}_\star)$, **and** $\boldsymbol{\Theta}_{\star,\beta} = \mathcal{F}(\boldsymbol{M}_{\star,\beta})$ and $\widehat{\boldsymbol{\Theta}}_{\star,\beta} = \mathcal{F}(\widehat{\boldsymbol{M}}_{\star,\beta})$ respectively. It is noted that $\widehat{\boldsymbol{M}}_{\star,\beta}$ serves as a "bridge" between the underlying $\boldsymbol{M}_{\star,\beta}$ and the empirical $\widehat{\boldsymbol{M}}_\beta$. Write $N_\beta := \sum_{i,j}(\mathbb{1}_{[z_{\star,ij} > \beta]} + \mathbb{1}_{[z_{\star,ij} < -\beta]})$ as the number of extreme values in $\boldsymbol{Z}_\star$ at level β. The convergence rates of $d^2(\widehat{\boldsymbol{M}}_\beta, \boldsymbol{M}_\star)$ and $d^2(\widehat{\boldsymbol{\Theta}}_\beta^\dagger, \boldsymbol{\Theta}_\star^\dagger)$ are investigated in the next theorem.

Theorem 3. *Assume that Conditions C1-C2 hold, and $(\mu_\star, \boldsymbol{Z}_\star) \in \mathcal{C}_{n_1,n_2}(\alpha_1, \alpha_2)$. Consider $\widehat{\boldsymbol{M}}_\beta = \widehat{\mu}\boldsymbol{J} + \widehat{\boldsymbol{Z}}_\beta$ where $\widehat{\boldsymbol{Z}}_\beta$ is the solution to (3.4) and $\beta \geq 0$, there exist some positive constants C_1, C_2 and C_3 such that for $\lambda, \lambda' \geq (8e+1)\sqrt{n}$, we have with probability at least $1 - 2C_1/n$,*

$$d^2\left(\widehat{\boldsymbol{Z}}_\beta, \mathcal{T}_\beta\left(\boldsymbol{Z}_\star\right)\right) \leq C_3\Lambda_{n,\lambda'}, \quad d^2\left(\widehat{\boldsymbol{M}}_\beta, \boldsymbol{M}_\star\right) \leq C_2\left(\alpha_1^2 \wedge \Gamma_{n,\lambda}\right) + C_3\Lambda_{n,\lambda'} + \frac{2(\alpha_2 - \beta)_+^2 N_\beta}{n^2} \tag{5.3}$$

$$\textit{and} \quad d^2\left(\widehat{\boldsymbol{\Theta}}_\beta^\dagger, \boldsymbol{\Theta}_\star^\dagger\right) \leq \frac{C_2}{h_{\alpha_1,\beta}^2}\left(\alpha_1^2 \wedge \Gamma_{n,\lambda}\right) + \frac{C_3\Lambda_{n,\lambda'}}{h_{\alpha_1,\beta}^2} + \frac{8N_\beta}{n^2\theta_L^2}, \tag{5.4}$$

where

$$\Lambda_{n,\lambda'} := \min\left\{\beta^2, \tilde{\Gamma}_{n,\lambda'} + \frac{\beta\left(8N_\beta + \left(n^2 - N_\beta\right)|\mu_\star - \widehat{\mu}|\right)}{h_{\alpha_1,\beta}n^2}\right\}, \quad \textit{and}$$

$$\tilde{\Gamma}_{n,\lambda'} := \min\left\{\frac{\lambda'}{h_{\alpha_1,\beta}n^2}\left\|\mathcal{T}_\beta\left(\boldsymbol{Z}_\star\right)\right\|_*, \frac{r_{\mathcal{T}_\beta(\boldsymbol{Z}_\star)}\lambda'^2}{h_{\alpha_1,\beta}^2 n^2}\right\}.$$

In our proof, the second term in (5.3) provides an upper bound $C_3\Lambda_{n,\lambda'}$ for $d^2(\widehat{\boldsymbol{Z}}_\beta, \mathcal{T}_\beta(\boldsymbol{Z}_\star))$. Similarly, we can derive an upper bound $4\beta^2$ for $d^2(\widehat{\boldsymbol{Z}}_{\mathsf{Win},\beta}, \mathcal{T}_\beta(\boldsymbol{Z}_\star))$ from the second term in Theorem 2 where $\widehat{\boldsymbol{Z}}_{\mathsf{Win},\beta} = \mathcal{T}_\beta(\widehat{\boldsymbol{Z}})$ is directly winsorized from $\widehat{\boldsymbol{Z}}$. Obviously, the order of this upper bound is larger than or equal to $\Lambda_{n,\lambda'}$. Moreover, there are scenarios where $\Lambda_{n,\lambda'}$ is a smaller order of β^2. To illustrate, assume that both $\alpha_1 \asymp 1$ and $\beta \asymp 1$, we have $h_{\alpha_1,\beta} \asymp 1$. Once we have $N_\beta = o(n)$, $r_{\mathcal{T}_\beta(\boldsymbol{Z}_\star)} = o(n)$ and $|\widehat{\mu} - \mu_\star| = o(1)$, then $\Lambda_{n,\lambda'} = o(\beta^2)$. With a more dedicated investigation of (5.4), one can also derive an upper bound for $d^2(\widehat{\boldsymbol{\Theta}}^\dagger_\beta, \widehat{\boldsymbol{\Theta}}^\dagger_{\star,\beta})$, which will be used in Section 5.2. Due to the facts that $\|\boldsymbol{Z}_\star\|_* \le \alpha_2 r^{1/2}_{\boldsymbol{Z}_\star} n$ and $\|\mathcal{T}_\beta(\boldsymbol{Z}_\star)\|_* \le \min\{\beta r^{1/2}_{\mathcal{T}_\beta(\boldsymbol{Z}_\star)} n, \alpha_2 r^{1/2}_{\boldsymbol{Z}_\star} n + (\alpha_2 - \beta)_+ N_\beta n^{1/2}\}$, such an upper bound is of order $k_{\alpha_1,\alpha_2,\beta,n} h^{-2}_{\alpha_1,\beta}$ where

$$k_{\alpha_1,\alpha_2,\beta,n} \asymp \min\Big[\beta^2, h^{-1}_{\alpha_1,\beta} n^{-1/2} \min\Big\{\beta r^{1/2}_{\mathcal{T}_\beta(\boldsymbol{Z}_\star)}, \alpha_2 r^{1/2}_{\boldsymbol{Z}_\star} + (\alpha_2 - \beta)_+ \, N_\beta n^{-1/2}, h^{-1}_{\alpha_1,\beta} n^{-1/2} r_{\mathcal{T}_\beta(\boldsymbol{Z}_\star)}\Big\}$$
$$+\beta\left(8N_\beta + \left(n^2 - N_\beta\right) k'^{1/2}_{\alpha_1,\alpha_2,n}\right) n^{-2}\Big], \quad \text{and}$$

$$k'_{\alpha_1,\alpha_2,n} := \min\left\{\alpha_1^2, r_{\boldsymbol{Z}_\star} \alpha_2 e^{\alpha_0} n^{-1/2}, r_{\boldsymbol{Z}_\star} e^{2\alpha_0} n^{-1}\right\}.$$

In particular, the term $8N_\beta n^{-2}\theta_L^{-2}$ is due to $d^2(\boldsymbol{\Theta}^\dagger_{\star,\beta}, \boldsymbol{\Theta}^\dagger_\star)$.

5.2 Target matrix

To study the asymptotic convergence of $d^2(\widehat{\boldsymbol{A}}_\beta, \boldsymbol{A}_\star)$, we require the following conditions of the random errors $\boldsymbol{\epsilon}$ and the target matrix $\boldsymbol{A}_\star$. Recall that $\widehat{\boldsymbol{A}}_\beta$ includes both the estimations obtained with the unconstrained estimator $\widehat{\boldsymbol{\Theta}}$ and the constrained estimator $\widehat{\boldsymbol{\Theta}}_\beta$.

C3. (a) The random errors $\{\epsilon_{ij}\}$ in Model (2.1) are independently distributed random variables such that $\mathsf{E}(\epsilon_{ij}) = 0$ and $\mathsf{E}(\epsilon_{ij}^2) = \sigma_{ij}^2 < \infty$ for all i, j. (b) For some finite positive constants c_σ and η, $\max_{i,j} \mathsf{E}|\epsilon_{ij}|^l \le \frac{1}{2} l! c_\sigma^2 \eta^{l-2}$ for any positive integer $l \ge 2$.

C4. There exists a positive constant a such that $\|\boldsymbol{A}_\star\|_\infty \le a$.

Denote $\widehat{\boldsymbol{\Theta}}_{\star,\beta} = (\widehat{\theta}_{\star,ij,\beta})_{i,j=1}^{n_1,n_2}$, $h_{(1),\beta} := \max_{i,j}(\theta^{-1}_{\star,ij}\widehat{\theta}_{\star,ij,\beta})$ and

$$\Delta := \max\left\{\frac{(c_\sigma \vee a)\, e^{-\mu_\star/2 + \alpha_2 - \beta + |\alpha_2/2 - \beta|}\sqrt{n \log n}}{n^2}, \frac{\eta e^{\mu_\star/2 + \alpha_1 + |\alpha_2/2 - \beta|} k^{1/2}_{\alpha_1,\alpha_2,\beta,n} \log^{3/2} n}{h_{\alpha_1,\beta} n}\right\}. \quad (5.5)$$

The following theorem established a general upper bound for $d^2(\widehat{\boldsymbol{A}}_\beta, \boldsymbol{A}_\star)$.

Theorem 4. *Assume Conditions C1-C4 hold. For $\beta \geq 0$, there exist some positive constants C_4 and C_5, both independent of β, such that for $h_{(1),\beta}\tau \geq C_4\Delta$, we have with probability at least $1 - C_5/n$,*

$$d^2\left(\widehat{\boldsymbol{A}}_\beta, \boldsymbol{A}_\star\right) \leq \min\left\{2h_{(1),\beta}\tau \left\|\boldsymbol{A}_\star\right\|_*, 16n^2 r_{\boldsymbol{A}_\star} h^2_{(1),\beta}\tau^2\right\}. \tag{5.6}$$

As for the estimator of the target matrix based on direct winsorization $\widehat{\boldsymbol{\Theta}}_{\mathsf{Win},\beta} = \mathcal{F}(\widehat{\mu}\boldsymbol{J} + \widehat{\boldsymbol{Z}}_{\mathsf{Win},\beta})$ where $\widehat{\boldsymbol{Z}}_{\mathsf{Win},\beta} = \mathcal{T}_\beta(\widehat{\boldsymbol{Z}})$, an upper bound can be derived using Theorem 3. As noted in a remark after Theorem 3, $d^2(\widehat{\boldsymbol{Z}}_{\mathsf{Win},\beta}, \mathcal{T}_\beta(\boldsymbol{Z}_\star))$ converges at a slower rate which will cause a larger error bound for the target matrix.

Now, we discuss the rates of $d^2(\widehat{\boldsymbol{A}}_\beta, \boldsymbol{A}_\star)$ under various missing structures. For simplicity, the following discussion focuses on the low-rank linear predictor ($\boldsymbol{M}_\star$) setting, i.e., $r_{\boldsymbol{M}_\star} \asymp 1$. Under the uniform missingness, i.e., $\theta_{ij} \equiv \theta_0$, it has been shown in Koltchinskii et al. (2011) that $\theta_0^{-1}n^{-1}\text{polylog}(n)$ is the optimal rate for $d^2(\widehat{\boldsymbol{A}}_\beta, \boldsymbol{A}_\star)$. Therefore it is reasonable to require $\alpha_1 + \alpha_2 = \alpha_0 = O(\text{polylog}(n))$ for the convergence of $d^2(\widehat{\boldsymbol{A}}_\beta, \boldsymbol{A}_\star)$. Under the uniform missingness, we have $\alpha_2 = 0$, $\alpha_0 = \alpha_1$ and $e^{\mu_\star} \asymp \theta_0$. For $\beta = 0$, our estimator $\widehat{\boldsymbol{A}}_\beta$ degenerates to the estimator based on the unweighted empirical risk function. Theorem 4 shows that $\widehat{\boldsymbol{A}}_\beta$ achieves the optimal rate $\theta_0^{-1}n^{-1}\text{polylog}(n)$. As for $\beta > 0$, by taking $\beta \to 0$ such that $k_{\alpha_1,\alpha_2,\beta,n} = O(e^{\mu_\star - 2\alpha_1 - 2\beta}n^{-1}\log^{-2} n)$, the estimator can also reach the optimal rate. Of interest here is that β is allowed to be strictly positive to achieve the same rate.

Under the non-uniform missingness, suppose the lower and upper bounds of observation probability satisfy $\theta_L \asymp e^{\mu_\star - \alpha_2}$ and $\theta_U \asymp e^{\mu_\star + \alpha_2}$. For the non-constrained case of $\beta = \alpha_2$ and $h_{\alpha_1,\beta} \asymp e^{-\alpha_1 - \alpha_2}$, the second term of Δ in (5.5) dominates due to the fact that

$$e^{-\mu_\star/2 + \alpha_2/2}n^{-3/2}\log^{1/2} n = o(e^{\mu_\star/2 + 5\alpha_1/2 + 3\alpha_2/2}n^{-5/4}\log^{3/2} n).$$

Thus the convergence rate of $d^2(\widehat{\boldsymbol{A}}_\beta, \boldsymbol{A}_\star)$ is $e^{\mu_\star + 5\alpha_1 + 3\alpha_2}n^{-1/2}\log^3 n$. To guarantee convergence, as $e^{\mu_\star/2 + 5\alpha_1/2 + 3\alpha_2/2} \leq e^{3\alpha_1 + 3\alpha_2/2}$, it requires that $\alpha_1 + \alpha_2/2 < (1/12)\log n$ which implies that $\theta_L^{-1} = O(n^{1/6})$.

However, the above range of $\theta_L = O(n^{1/6})$ excludes $\theta_L \equiv (n^{-1}\text{polylog}(n))$, the case that results in the number of the observed matrix entries at the order of $n\,\mathrm{polylog}(n)$ which represents the most sparse case of observation where the matrix can still be recovered (Candès and Recht, 2009; Candès and Plan, 2010; Koltchinskii et al., 2011; Negahban and Wainwright, 2012). We will show in the following that with an appropriately chosen β, the constrained estimator $\widehat{\boldsymbol{\Theta}}_\beta$ can accommodate the case of $\theta_L^{-1} = O(n \log^{-1} n)$.

To demonstrate this, we start with the absolute constrained case, i.e., $\beta = 0$, which forces the estimated probabilities to be uniform and implies $e^{-\mu_\star/2+\alpha_2-\beta+|\alpha_2/2-\beta|} = e^{-\mu_\star/2+3\alpha_2/2} \asymp \theta_U^{1/2}\theta_L^{-1}$. Then, according to Theorem 4, $d^2(\widehat{\boldsymbol{A}}_\beta, \boldsymbol{A}_\star)$ attains the convergence rate $\theta_U\theta_L^{-2}n^{-1}\log(n)$, which converges to 0 provided $\theta_U\theta_L^{-2} = o(n\log^{-1} n)$. Obviously, the condition $\theta_U\theta_L^{-2} = o(n\log^{-1} n)$ includes the extreme case of $\theta_L^{-1} = O(n\log^{-1} n)$ and $n\,\mathrm{polylog}(n)$ observations.

For the more interesting setting $\beta > 0$, to simplify the discussion, we concentrate on the case when the first term in $k_{\alpha_1,\alpha_2,\beta,n}$ is of smallest order, which can be achieved by choosing $\beta = O(e^{-\mu_\star-2\alpha_1+\alpha_2}n^{-1/2}\log^{-1} n)$. Then, according to Theorem 4,

$$d^2(\widehat{\boldsymbol{A}}_\beta, \boldsymbol{A}_\star) = O_p(e^{-\mu_\star+2\alpha_2-2\beta+2|\alpha_2/2-\beta|}n^{-1}\log n) = O_p(e^{\alpha_1/2+3\alpha_2/2}n^{-1}\log n),$$

since $e^{-\mu_\star/2+\alpha_2-\beta+|\alpha_2/2-\beta|} \leq e^{\alpha_1/2+3\alpha_2/2}$. In the following we consider two cases: (i) $\alpha_2 = O((\log\log n)^{-1}\alpha_1)$ and (ii) $\alpha_1 = o(\alpha_2 \log\log n)$. Note that for either $\alpha_2 = O((\log\log n)^{-1}\alpha_1)$ or $\alpha_1 = o(\alpha_2\log\log n)$, we can simplify $e^{-\mu_\star+2\alpha_2-2\beta+2|\alpha_2/2-\beta|} \asymp \theta_U\theta_L^{-2}$ which leads to

$$d^2(\widehat{\boldsymbol{A}}_\beta, \boldsymbol{A}_\star) = O_p(\theta_U\theta_L^{-2}n^{-1}\log n).$$

If $\alpha_2 = O((\log\log n)^{-1}\alpha_1)$, we require $\alpha_1 < (1 + 3\log\log n)^{-1}(\log n - \log\log n)$ to guarantee convergence, which implies that $\theta_L = O(n^{-1})$. Thus, we only lose a $\mathrm{polylog}(n)$ factor when compared with the most extreme but feasible setting of $\theta_L^{-1} = O(n(\mathrm{polylog}(n))^{-1})$. Also $\beta = O(e^{-\mu_\star-2\alpha_1+\alpha_2}n^{-1/2}\log^{-1} n)$ implies that $\beta = O(n^{-1/2}\log^{-1} n)$. That is, we improve the rate of θ_L from $\theta_L^{-1} = O(n^{1/6})$ to $\theta_L^{-1} = O(n)$ under Case (i) while still able to attain consistency for the completed matrix $\hat{A}_\beta$. This implies that the proposed estimator can achieve the optimal rate (up to a $\mathrm{polylog}(n)$ order) with the above chosen $\beta = O(n^{-1/2}\log^{-1} n)$.

If $\alpha_1 = o((\log\log n)\alpha_2)$, we require that $\alpha_2 < (3 + (\log\log n)^{-1})^{-1}(\log n - \log\log n)$ which leads to $\theta_L^{-1} = O(n^{1/3})$. Also $\beta = O(e^{-\mu_\star - 2\alpha_1 + \alpha_2} n^{-1/2} \log^{-1} n)$ implies that $\beta = O(n^{-1/6} \log^{-1} n)$. However, to make $d^2(\widehat{\boldsymbol{A}}_\beta, \boldsymbol{A}_\star)$ convergent, there is still a gap, i.e., the attained rate for θ_L^{-1} has to be $O(n^{1/3})$, which does not cover the most extreme case of $\theta_L^{-1} = O(n(\text{polylog}(n))^{-1})$. The reason for not being able to attain the most extreme case of $\theta_L^{-1} = O(n(\text{polylog}(n))^{-1})$ is that the current Case (ii) allows more heterogeneity in $\boldsymbol{Z}_\star$ as reflected by having a larger α_2 than that prescribed under Case (i). As $\mu_\star$ is jointly estimated with $\boldsymbol{Z}_\star$ in the unconstrained estimation (Section 3.1), stronger heterogeneity slows down the convergence rate for the estimation of $\mu_\star$, which becomes one of the bottleneck for further improvement. If $\mu_\star$ was observable, the gap would not be as serious despite the adverse effect of stronger heterogeneity on the estimation of $\boldsymbol{Z}_\star$. Given our current result, for Case (ii), when the missingness is not extreme, i.e., $\theta_L^{-1} = O(n^{1/3})$, with an appropriately chosen $\beta > 0$, the proposed estimator can also achieve the optimal rate (up to polylog(n) order). Or otherwise, β should be set as 0 to achieve the optimal rate.

5.3 Comparison with Uniform Objective Function

Recall that the unweighted empirical risk function $\widehat{R}_{\text{UNI}}(\boldsymbol{A}) = (n_1 n_2)^{-1}\|\boldsymbol{W} \circ (\boldsymbol{A} - \boldsymbol{Y})\|_F^2$ is adopted by many existing matrix completion techniques (Klopp, 2014). An interesting question is whether there is any benefit in adopting the proposed weighted empirical risk function for matrix completion. In this subsection, we aim to shed some light on this aspect by comparing the non-asymptotic error bounds of the corresponding estimators. Due to the additional complication from the estimation error of the observation probability matrix, we only focus on the weighted empirical risk function with true inverse probability weighting in this section. We will demonstrate empirically in Sections 6 and 7 the benefits of the weighted objective function with estimated weights.

Most existing work with unweighted empirical risk function assume the true missingness is uniform (Candès and Plan, 2010; Koltchinskii et al., 2011). One notable exception is Klopp (2014), where unweighted empirical risk function is studied under possibly non-uniform missing structure. The estimator of Klopp (2014) is equivalent to our estimator when $\beta = 0$, which is denoted by

$\widehat{\boldsymbol{A}}^{\text{UNI}}$. Thus, according to Theorem 4, we have with probability at least $1 - C_5/n$,

$$d^2\left(\widehat{\boldsymbol{A}}^{\text{UNI}}, \boldsymbol{A}_\star\right) \leq \min\left\{2\theta_U^{1/2}\theta_L^{-1}n^{-3/2}\log^{-1/2} n \left\|\boldsymbol{A}_\star\right\|_*, 16r_{\boldsymbol{A}_\star}\theta_U\theta_L^{-2}n^{-1}\log^{-1} n\right\} := U^{\text{UNI}},$$

which is the same upper bound obtained in Klopp (2014). Define $\widehat{\boldsymbol{A}}^{\text{KNOWN}}$ as the estimator which minimizes the known weighted empirical risk function (2.3). Then,

$$d^2\left(\widehat{\boldsymbol{A}}^{\text{KNOWN}}, \boldsymbol{A}_\star\right) \leq \min\left\{2\theta_L^{-1/2}n^{-3/2}\log^{-1/2} n \left\|\boldsymbol{A}_\star\right\|_*, 16r_{\boldsymbol{A}_\star}\theta_L^{-1}n^{-1}\log^{-1} n\right\} := U^{\text{KNOWN}}.$$

The improvement in the upper bounds of the weighted objective function $\widehat{R}$ lies in that, under non-uniform missingness, $\theta_U\theta_L^{-1} > 1$ which implies that $U^{\text{KNOWN}} < U^{\text{UNI}}$ as summarized below.

Theorem 5. *Assume Conditions C1-C4 holds, and take $\tau_{KNOWN} = C_3\theta_L^{-1/2}n^{-3/2}\log^{-1/2} n$ and $\tau_{UNI} = C_3\theta_U^{1/2}f^{-1}(\mu_\star)n^{-3/2}\log^{-1/2} n$. The upper bound of $d^2(\widehat{\boldsymbol{A}}^{UNI}, \boldsymbol{A}_\star)$ is the same as U^{UNI} and the upper bound of $d^2(\widehat{\boldsymbol{A}}^{KNOWN}, \boldsymbol{A}_\star)$ is the same as U^{KNOWN}. In addition, $U^{KNOWN} \leq U^{UNI}$, and $U^{KNOWN} < U^{UNI}$ if $\theta_U > \theta_L$, i.e., the true missing mechanism is non-uniform.*

6 Simulation Study

This section reports results from simulation experiments which were designed to evaluate the numerical performance of the proposed methodologies. We first evaluate the estimation performances of the observation probabilities in Section 6.1 and then those of the target matrix in Section 6.2.

6.1 Missingness

In the simulation, the true observation probabilities $\boldsymbol{\Theta}_\star$ and the target matrix $\boldsymbol{A}_\star$ were randomly generated once and kept fixed for each simulation setting to be described below. To generate $\boldsymbol{\Theta}_\star$, we first generated $\boldsymbol{U}_{\boldsymbol{M}_\star} \in \mathbb{R}^{n_1 \times (r_{\boldsymbol{M}_\star}-1)}$ and $\boldsymbol{V}_{\boldsymbol{M}_\star} \in \mathbb{R}^{(r_{\boldsymbol{M}_\star}-1) \times n_2}$ as random Gaussian matrices with independent entries each following $\mathcal{N}(-0.4, 1)$. We then obtained $\boldsymbol{M}_\star = \boldsymbol{U}_{\boldsymbol{M}_\star}\boldsymbol{V}_{\boldsymbol{M}_\star}^\intercal - \bar{m}_{n_1,n_2,r_{\boldsymbol{M}_\star}}\boldsymbol{J}$ where $\bar{m}_{n_1,n_2,r_{\boldsymbol{M}_\star}}$ is a scalar chosen to ensure the average observation rate is 0.2 in each simulation setting. We finally set $\boldsymbol{\Theta}_\star = \mathcal{F}(\boldsymbol{M}_\star)$ where the inverse link function f is a logistic function.

In our study, we set $r_{\boldsymbol{M}_\star} = 11$, (or $r_{\boldsymbol{Z}_\star} = 10$) and chose $n_1 = n_2$ with four sizes: 600, 800, 1000 and 1200, and the number of simulation runs for each settings was 500.

For the purpose of benchmarking, we compared various estimators of the missingness:

1. the non-constrained estimator $\widehat{\boldsymbol{\Theta}}_\alpha$ defined in (3.1);

2. the constrained estimator $\widehat{\boldsymbol{\Theta}}_\beta$ defined in (3.4);

3. the directly winsorized estimator $\widehat{\boldsymbol{\Theta}}_{\mathsf{Win},\beta} = \mathcal{F}(\widehat{\mu}\boldsymbol{J} + \mathcal{T}_\beta(\widehat{\boldsymbol{Z}}))$;

4. the 1-bit estimator $\widehat{\boldsymbol{\Theta}}_{\mathsf{1\text{-}bit},\alpha}$ proposed in Davenport et al. (2014) and its corresponding constrained and winsorized versions $\widehat{\boldsymbol{\Theta}}_{\mathsf{1\text{-}bit},\beta}$ and $\widehat{\boldsymbol{\Theta}}_{\mathsf{1\text{-}bit,Win},\beta}$; (note that the 1-bit estimator $\widehat{\boldsymbol{\Theta}}_{\mathsf{1\text{-}bit},\alpha}$ imposes the nuclear-norm regularization on the whole $\boldsymbol{M}$ instead of $\boldsymbol{Z}$, when compared to $\widehat{\boldsymbol{\Theta}}_\alpha$)

5. the rank-1 probability estimator $\widehat{\boldsymbol{\Theta}}_{\mathsf{NW}}$ used in Negahban and Wainwright (2012) where $g_{i.} = n_2^{-1}\sum_{j=1}^{n_2} w_{ij}$, $g_{.j} = n_1^{-1}\sum_{i=1}^{n_1} w_{ij}$ and $\theta_{ij,\mathsf{NW}} = g_{i.}g_{.j}$;

6. the uniform estimator $\widehat{\boldsymbol{\Theta}}_{\mathsf{UNI}} = N/(n_1 n_2)\boldsymbol{J}$.

For the non-constrained estimator $\widehat{\boldsymbol{\Theta}}_\alpha$ and the 1-bit estimator $\widehat{\boldsymbol{\Theta}}_{\mathsf{1\text{-}bit},\alpha}$, the parameter α is set according to the knowledge of the true $\boldsymbol{M}_\star$. For the constrained estimators $\widehat{\boldsymbol{\Theta}}_\beta$ and $\widehat{\boldsymbol{\Theta}}_{\mathsf{Win},\beta}$, the constraint level β was chosen so that either 5% or 10% of the elements in $\widehat{\boldsymbol{Z}}_\alpha$ were winsorized. Similarly for $\widehat{\boldsymbol{\Theta}}_{\mathsf{1\text{-}bit},\beta}$ and $\widehat{\boldsymbol{\Theta}}_{\mathsf{1\text{-}bit,Win},\beta}$.

To quantify the estimation performance of linear predictor $\boldsymbol{M}_\star$ and observation probabilities $\boldsymbol{\Theta}_\star$, we considered the empirical root mean squared errors RMSE$(\boldsymbol{B}, \boldsymbol{C})$ with respect to any two matrices $\boldsymbol{B}$ and $\boldsymbol{C}$ of dimension $n_1 \times n_2$, and the Hellinger distance $d_H^2(\widehat{\boldsymbol{\Theta}}, \boldsymbol{\Theta}_\star)$ between $\widehat{\boldsymbol{\Theta}}$ and $\boldsymbol{\Theta}_\star$ defined as follows:

$$\mathrm{RMSE}\left(\boldsymbol{B}, \boldsymbol{C}\right) := \frac{\|\boldsymbol{B} - \boldsymbol{C}\|_F}{\sqrt{n_1 n_2}} \quad \text{and} \quad d_H^2\left(\widehat{\boldsymbol{\Theta}}, \boldsymbol{\Theta}_\star\right) := \frac{\sum_{i,j}^{n_1,n_2} d_H^2\left(\widehat{\theta}_{ij}, \theta_{\star,ij}\right)}{\sqrt{n_1 n_2}},$$

where $d_H^2(s,t) = (\sqrt{s} - \sqrt{t})^2 + (\sqrt{1-s} - \sqrt{1-t})^2$ as defined in Section 5. As the estimators $\mathcal{F}^{-1}(\widehat{\boldsymbol{\Theta}}_\alpha)$ and $\mathcal{F}^{-1}(\widehat{\boldsymbol{\Theta}}_{\mathsf{1\text{-}bit},\alpha})$ are both low-rank, we also report their corresponding ranks.

Table 1: The empirical root mean squared errors (RMSEs) RMSE$(\widehat{\boldsymbol{M}}, \boldsymbol{M}_\star)$, Hellinger distance $d_H^2(\widehat{\boldsymbol{\Theta}}, \boldsymbol{\Theta}_\star)$, rank of linear predictor $\widehat{\boldsymbol{M}}$ and estimated $\widehat{\boldsymbol{\Theta}}$ and their standard errors (in parentheses) under the low rank missing observation mechanism, with $(n_1, n_2) = (600, 600), (800, 800), (1000, 1000), (1200, 1200)$ and $r_{\boldsymbol{M}_\star} = 11$, for the proposed estimators $\widehat{\boldsymbol{\Theta}}_\alpha$, $\widehat{\boldsymbol{\Theta}}_{\textsf{1-bit},\alpha}$ and the two existing estimators ($\widehat{\boldsymbol{\Theta}}_{\textsf{NW}}$ and $\widehat{\boldsymbol{\Theta}}_{\textsf{UNI}}$).

600	$\widehat{\boldsymbol{\Theta}}_\alpha$	$\widehat{\boldsymbol{\Theta}}_{\textsf{1-bit},\alpha}$	$\widehat{\boldsymbol{\Theta}}_{\textsf{NW}}$	$\widehat{\boldsymbol{\Theta}}_{\textsf{UNI}}$
RMSE$(\widehat{\boldsymbol{M}}, \boldsymbol{M}_\star)$	2.6923 (0.0342)	2.9155 (0.0295)	-	-
$d_H^2(\widehat{\boldsymbol{\Theta}}, \boldsymbol{\Theta}_\star)$	0.0369 (0.0015)	0.0450 (0.0016)	0.1233 (1e-04)	0.1729 (1e-04)
$r_{\widehat{\boldsymbol{M}}}$	12.45 (0.50)	12.69 (0.46)	-	-
$r_{\widehat{\boldsymbol{\Theta}}}$	600.00 (0.00)	600.00 (0.00)	-	-
800	$\widehat{\boldsymbol{\Theta}}_\alpha$	$\widehat{\boldsymbol{\Theta}}_{\textsf{1-bit},\alpha}$	$\widehat{\boldsymbol{\Theta}}_{\textsf{NW}}$	$\widehat{\boldsymbol{\Theta}}_{\textsf{UNI}}$
RMSE$(\widehat{\boldsymbol{M}}, \boldsymbol{M}_\star)$	2.5739 (0.0116)	2.7796 (0.0033)	-	-
$d_H^2(\widehat{\boldsymbol{\Theta}}, \boldsymbol{\Theta}_\star)$	0.0317 (5e-04)	0.0379 (1e-04)	0.1219 (1e-04)	0.1767 (1e-04)
$r_{\widehat{\boldsymbol{M}}}$	12.04 (0.20)	12.03 (0.17)	-	-
$r_{\widehat{\boldsymbol{\Theta}}}$	800.00 (0.00)	800.00 (0.00)	-	-
1000	$\widehat{\boldsymbol{\Theta}}_\alpha$	$\widehat{\boldsymbol{\Theta}}_{\textsf{1-bit},\alpha}$	$\widehat{\boldsymbol{\Theta}}_{\textsf{NW}}$	$\widehat{\boldsymbol{\Theta}}_{\textsf{UNI}}$
RMSE$(\widehat{\boldsymbol{M}}, \boldsymbol{M}_\star)$	2.4870 (0.0212)	2.7731 (0.0015)	-	-
$d_H^2(\widehat{\boldsymbol{\Theta}}, \boldsymbol{\Theta}_\star)$	0.0266 (8e-04)	0.0351 (1e-04)	0.1246 (1e-04)	0.1767 (1e-04)
$r_{\widehat{\boldsymbol{M}}}$	12.68 (0.53)	12.00 (0.00)	-	-
$r_{\widehat{\boldsymbol{\Theta}}}$	1000.00 (0.00)	1000.00 (0.00)	-	-
1200	$\widehat{\boldsymbol{\Theta}}_\alpha$	$\widehat{\boldsymbol{\Theta}}_{\textsf{1-bit},\alpha}$	$\widehat{\boldsymbol{\Theta}}_{\textsf{NW}}$	$\widehat{\boldsymbol{\Theta}}_{\textsf{UNI}}$
RMSE$(\widehat{\boldsymbol{M}}, \boldsymbol{M}_\star)$	2.3809 (0.0018)	2.6470 (0.0012)	-	-
$d_H^2(\widehat{\boldsymbol{\Theta}}, \boldsymbol{\Theta}_\star)$	0.0242 (1e-04)	0.0314 (1e-04)	0.1211 (1e-04)	0.1761 (1e-04)
$r_{\widehat{\boldsymbol{M}}}$	12.00 (0.00)	12.00 (0.00)	-	-
$r_{\widehat{\boldsymbol{\Theta}}}$	1200.00 (0.00)	1200.00 (0.00)	-	-

Table 1 summarizes the simulation results for the missingness. The most visible aspect of the results is that the proposed estimators $\widehat{\boldsymbol{\Theta}}_\alpha$ and $\widehat{\boldsymbol{\Theta}}_{\textsf{1-bit},\alpha}$ both have superior performance than the two existing estimators $\widehat{\boldsymbol{\Theta}}_{\textsf{NW}}$ and $\widehat{\boldsymbol{\Theta}}_{\textsf{UNI}}$ by having smaller root mean square errors with respect to $\widehat{\boldsymbol{M}}$, Hellinger distances $d_H^2(\widehat{\boldsymbol{\Theta}}, \boldsymbol{\Theta}_\star)$ and more accuracy estimated rank of $\boldsymbol{M}_\star$. Without the separation of $\mu_\star$ from $\boldsymbol{M}_\star$, $\widehat{\boldsymbol{\Theta}}_{\textsf{1-bit},\alpha}$ has larger error and Hellinger distance than the proposed estimators. The performance of $\widehat{\boldsymbol{\Theta}}_{\textsf{NW}}$ is roughly between the proposed estimators and the uniform estimator $\widehat{\boldsymbol{\Theta}}_{\textsf{UNI}}$. Estimator $\widehat{\boldsymbol{\Theta}}_{\textsf{UNI}}$ is a benchmark which captures no variation of the observation probabilities.

6.2 Target matrix

To generate a target matrix $\boldsymbol{A}_\star$, we first generated $\boldsymbol{U}_{\boldsymbol{A}_\star} \in \mathbb{R}^{n_1 \times (r_{\boldsymbol{A}_\star}-1)}$ and $\boldsymbol{V}_{\boldsymbol{A}_\star} \in \mathbb{R}^{(r_{\boldsymbol{A}_\star}-1) \times n_2}$ as random matrices with independent Gaussian entries distributed as $\mathcal{N}(0, \sigma^2_{\boldsymbol{A}_\star})$ and obtained $\boldsymbol{A}_\star = 2.5\boldsymbol{J} + \boldsymbol{U}_{\boldsymbol{A}_\star}\boldsymbol{V}^\intercal_{\boldsymbol{A}_\star}$. Here we set the standard deviation of the entries in the matrix product $\boldsymbol{U}_{\boldsymbol{A}_\star}\boldsymbol{V}^\intercal_{\boldsymbol{A}_\star}$ to be 2.5 to mimic the Yahoo! Webscope data set described in Section 7. To achieve this, $\sigma_{\boldsymbol{A}_\star} = (2.5^2/(r_{\boldsymbol{A}_\star}-1))^{1/4}$. The contaminated version of $\boldsymbol{A}_\star$ was 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 $\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2_\epsilon)$. The σ^2_ϵ is chosen such that SNR $= \sqrt{\mathsf{E}\|\boldsymbol{A}_\star\|^2_F / \mathsf{E}\|\boldsymbol{\epsilon}\|^2_F} = 1$, where $\mathsf{E}\|\boldsymbol{A}_\star\|^2_F = n_1 n_2 (r_{\boldsymbol{A}_\star} - 1 + 2.5^2)$ implies $\sigma_\epsilon = 0.5\sqrt{r_{\boldsymbol{A}_\star} - 1 + 2.5^2}$.

For the estimation of the target matrix, we evaluated ten versions of the proposed estimators Proposed_$\widehat{\boldsymbol{\Theta}}_\beta$_$t$, Proposed_$\widehat{\boldsymbol{\Theta}}_{\mathsf{Win},\beta}$_$t$, Proposed_$\widehat{\boldsymbol{\Theta}}_\alpha$, Proposed_$\widehat{\boldsymbol{\Theta}}_{\mathsf{1\text{-}bit},\beta}$_$t$, Proposed_$\widehat{\boldsymbol{\Theta}}_{\mathsf{1\text{-}bit,Win},\beta}$_$t$ and Proposed_$\widehat{\boldsymbol{\Theta}}_{\mathsf{1\text{-}bit},\alpha}$. Here Proposed indicates the estimators are obtained by solving problem (4.3), while $\widehat{\boldsymbol{\Theta}}_\beta$, $\widehat{\boldsymbol{\Theta}}_{\mathsf{Win},\beta}$, $\widehat{\boldsymbol{\Theta}}_\alpha$, $\widehat{\boldsymbol{\Theta}}_{\mathsf{1\text{-}bit},\beta}$, $\widehat{\boldsymbol{\Theta}}_{\mathsf{1\text{-}bit,Win},\beta}$ and $\widehat{\boldsymbol{\Theta}}_{\mathsf{1\text{-}bit},\alpha}$ represents the probability estimators used in (4.3), as described in Section 6.1, and $t = 0.05$ or 0.1 denote the winsorized proportion for which β is chosen. In addition, same as Mao et al. (2018), we also compared them with three existing matrix completion techniques: the methods proposed in Negahban and Wainwright (2012) (NW), Koltchinskii et al. (2011) (KLT) and Mazumder et al. (2010) (MHT). Among these three methods, NW is the only one that adjusts for non-uniform missingness. All three methods require tuning parameter selection, for which cross-validation is adopted. See Mao et al. (2018) for more details.

To quantify the performance of the matrix completion, in addition to the empirical root mean squared errors with respect to $\widehat{\boldsymbol{A}}_\beta$ and $\boldsymbol{A}_\star$, we used one more measure:

$$\text{Test Error} := \frac{\left\| \boldsymbol{W}^\star \circ \left(\widehat{\boldsymbol{A}}_\beta - \boldsymbol{A}_\star \right) \right\|^2_F}{\|\boldsymbol{W}^\star \circ \boldsymbol{A}_\star\|^2_F},$$

where $\boldsymbol{W}^\star$ is the matrix of missing indicator with the (i,j)-th entry being $(1 - w_{ij})$. The test error measures the relative estimation error of the unobserved entries to their signal strength. The estimated ranks of $\widehat{\boldsymbol{A}}_\beta$ are also reported.

Tables 2-3 summarize the simulation results for different dimensions $n_1{=}n_2$ ranges from 600 to 1200 and two different settings of $r_{\boldsymbol{A}_\star} = 11$. The results of $r_{\boldsymbol{A}_\star} = 31$ are delegated to Tables S1-S2 of Section S1.4 in the supplementary material. From the tables, we notice that the ten versions of the proposed methods possess superior performance than the three existing methods by having smaller RMSEs and Test Errors. Among the first five proposed methods in the tables, Proposed_$\widehat{\boldsymbol{\Theta}}_\beta$ is better than Proposed_$\widehat{\boldsymbol{\Theta}}_\alpha$ for most of the time. It is because that the constrained estimator $\widehat{\boldsymbol{\Theta}}_\beta$ has much smaller ratio $\widehat{\theta}_U/\widehat{\theta}_L$ than $\widehat{\boldsymbol{\Theta}}_\alpha$ which improve the stability of prediction and the accuracy. Another observation is that Proposed_$\widehat{\boldsymbol{\Theta}}_\beta$_0.1 performs better than Proposed_$\widehat{\boldsymbol{\Theta}}_{\text{1-bit},\alpha}$ at most times.

7 Real data application

In this section we demonstrate the proposed methodology by analyzing the Yahoo! Webscope dataset (ydata-ymusic-user-artist-ratings-v1_0) available at `http://research.yahoo.com/Academic_Relations`. 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, each rates exactly 10 songs that are not rated in the training set. The missing rates are 0.9763 overall, 0.3520 to 0.9900 across users, and 0.6372 to 0.9957 across songs. The non-uniformity of the missingness is shown in Figure 1. In this experiment, we applied those methods as described in Section 6 to the training set and evaluated the test errors based on the corresponding test set. Here α was set as 100, so that $\hat{\boldsymbol{Z}}_\alpha$ was not sensitive to larger α.

Table 4 reports the root mean squared prediction errors (RMSPEs), where RMSPE $:= \|\boldsymbol{W}^{test} \circ (\widehat{\boldsymbol{A}}_\beta - \boldsymbol{Y})\|_F / \sqrt{\sum_{i=1}^{n_1} \sum_{j=1}^{n_2} w_{ij}^{test}}$ and $\boldsymbol{W}^{test}$ is the indicator matrix of test set with the (i,j)-th entry being w_{ij}^{test}. Note that Proposed_$\widehat{\boldsymbol{\Theta}}_\beta$_0.05 performs the best among all ten versions of proposed methods. Besides, Proposed_$\widehat{\boldsymbol{\Theta}}_\alpha$ also has much smaller RMSPE than the other eight versions of proposed methods. This may indicate that only slight constraint is required for the probabilities estimator for this dataset. Note that we cannot gaurantee the optimal convergence rate or even asymptotic convergence in certain setting of missingness for Proposed_$\widehat{\boldsymbol{\Theta}}_\alpha$, see Section 5.2 for details.

Table 2: RMSEs, test errors, estimated ranks $r_{\widehat{\boldsymbol{A}}_\beta}$ and their standard deviations (in parentheses) under the low rank missing observation mechanism, for three existing methods and ten versions of the proposed methods where Proposed indicates the estimators are obtained by solving problem (4.3), while $\widehat{\boldsymbol{\Theta}}_\beta$, $\widehat{\boldsymbol{\Theta}}_{\mathsf{Win},\beta}$, $\widehat{\boldsymbol{\Theta}}_\alpha$, $\widehat{\boldsymbol{\Theta}}_{\mathsf{1\text{-}bit},\beta}$,$\widehat{\boldsymbol{\Theta}}_{\mathsf{1\text{-}bit,Win},\beta}$ and $\widehat{\boldsymbol{\Theta}}_{\mathsf{1\text{-}bit},\alpha}$ represents the probability estimators used in (4.3), as described in Section 6.1, and $t = 0.05$ or 0.1 denote the winsorized proportion for which β is chosen.

(n_1, n_2) = (600, 600)	RMSE($\widehat{\boldsymbol{A}}_\beta, \boldsymbol{A}_\star$)	Test Error	$r_{\widehat{\boldsymbol{A}}_\beta}$
Proposed_$\widehat{\boldsymbol{\Theta}}_{\mathsf{Win},\beta}$_0.05	1.5615 (0.0147)	0.3005 (0.0062)	65.28 (5.72)
Proposed_$\widehat{\boldsymbol{\Theta}}_\beta$_0.05	1.5548 (0.0085)	0.2996 (0.0034)	54.98 (3.01)
Proposed_$\widehat{\boldsymbol{\Theta}}_{\mathsf{Win},\beta}$_0.1	1.5621 (0.0111)	0.3013 (0.0046)	63.68 (5.36)
Proposed_$\widehat{\boldsymbol{\Theta}}_\beta$_0.1	1.5509 (0.0085)	0.2983 (0.0034)	53.13 (2.72)
Proposed_$\widehat{\boldsymbol{\Theta}}_\alpha$	1.5637 (0.0147)	0.3010 (0.0061)	65.63 (5.89)
Proposed_$\widehat{\boldsymbol{\Theta}}_{\mathsf{1\text{-}bit,Win},\beta}$_0.05	1.5664 (0.0093)	0.3028 (0.0037)	62.76 (5.96)
Proposed_$\widehat{\boldsymbol{\Theta}}_{\mathsf{1\text{-}bit},\beta}$_0.05	1.5573 (0.0089)	0.2996 (0.0036)	61.80 (5.34)
Proposed_$\widehat{\boldsymbol{\Theta}}_{\mathsf{1\text{-}bit,Win},\beta}$_0.1	1.5669 (0.0092)	0.3032 (0.0037)	62.78 (2.68)
Proposed_$\widehat{\boldsymbol{\Theta}}_{\mathsf{1\text{-}bit},\beta}$_0.1	1.5540 (0.0089)	0.2987 (0.0036)	60.79 (3.01)
Proposed_$\widehat{\boldsymbol{\Theta}}_{\mathsf{1\text{-}bit},\alpha}$	1.5612 (0.0097)	0.3005 (0.0040)	62.12 (4.76)
NW	2.0362 (0.2681)	0.4873 (0.1315)	174.76 (53.20)
KLT	2.2867 (0.0073)	0.5951 (0.0026)	1.00 (0.00)
MHT	1.6543 (0.0097)	0.3432 (0.0041)	51.20 (2.61)
(n_1, n_2) = (800, 800)	RMSE($\widehat{\boldsymbol{A}}_\beta, \boldsymbol{A}_\star$)	Test Error	$r_{\widehat{\boldsymbol{A}}_\beta}$
Proposed_$\widehat{\boldsymbol{\Theta}}_{\mathsf{Win},\beta}$_0.05	1.4754 (0.0107)	0.2669 (0.0041)	88.58 (10.81)
Proposed_$\widehat{\boldsymbol{\Theta}}_\beta$_0.05	1.4797 (0.0080)	0.2714 (0.0030)	71.79 (4.12)
Proposed_$\widehat{\boldsymbol{\Theta}}_{\mathsf{Win},\beta}$_0.1	1.4724 (0.0108)	0.2664 (0.0042)	86.25 (10.34)
Proposed_$\widehat{\boldsymbol{\Theta}}_\beta$_0.1	1.4763 (0.0082)	0.2704 (0.0031)	67.08 (4.22)
Proposed_$\widehat{\boldsymbol{\Theta}}_\alpha$	1.4783 (0.0115)	0.2676 (0.0041)	88.92 (11.70)
Proposed_$\widehat{\boldsymbol{\Theta}}_{\mathsf{1\text{-}bit,Win},\beta}$_0.05	1.4917 (0.0078)	0.2743 (0.0030)	83.51 (1.45)
Proposed_$\widehat{\boldsymbol{\Theta}}_{\mathsf{1\text{-}bit},\beta}$_0.05	1.4804 (0.0080)	0.2705 (0.0031)	82.60 (3.47)
Proposed_$\widehat{\boldsymbol{\Theta}}_{\mathsf{1\text{-}bit,Win},\beta}$_0.1	1.4972 (0.0080)	0.2765 (0.0031)	81.64 (7.23)
Proposed_$\widehat{\boldsymbol{\Theta}}_{\mathsf{1\text{-}bit},\beta}$_0.1	1.4800 (0.0078)	0.2708 (0.0030)	74.89 (3.54)
Proposed_$\widehat{\boldsymbol{\Theta}}_{\mathsf{1\text{-}bit},\alpha}$	1.4790 (0.0099)	0.2685 (0.0039)	88.57 (9.56)
NW	2.1281 (0.2279)	0.5303 (0.1150)	249.51 (59.38)
KLT	2.3447 (0.0064)	0.6081 (0.0020)	1.00 (0.00)
MHT	1.6067 (0.0086)	0.3245 (0.0036)	63.68 (3.02)

[1] With $r_{\boldsymbol{M}_\star} = 11$, $r_{\boldsymbol{A}_\star} = 11$, (n_1, n_2) = (1000, 1000), (1200, 1200) and SNR = 1. The three existing methods are proposed respectively in Negahban and Wainwright (2012)(NW), Koltchinskii et al. (2011)(KLT) and Mazumder et al. (2010)(MHT)

Table 3: RMSEs, test errors, estimated ranks $r_{\widehat{\boldsymbol{A}}_\beta}$ and their standard deviations (in parentheses) under the low rank missing observation mechanism, for three existing methods and ten versions of the proposed methods where Proposed indicates the estimators are obtained by solving problem (4.3), while $\widehat{\boldsymbol{\Theta}}_\beta$, $\widehat{\boldsymbol{\Theta}}_{\mathsf{Win},\beta}$, $\widehat{\boldsymbol{\Theta}}_\alpha$, $\widehat{\boldsymbol{\Theta}}_{\mathsf{1\text{-}bit},\beta}$, $\widehat{\boldsymbol{\Theta}}_{\mathsf{1\text{-}bit,Win},\beta}$ and $\widehat{\boldsymbol{\Theta}}_{\mathsf{1\text{-}bit},\alpha}$ represents the probability estimators used in (4.3), as described in Section 6.1, and $t = 0.05$ or 0.1 denote the winsorized proportion for which β is chosen.

(n_1, n_2) = (1000, 1000)	RMSE($\widehat{\boldsymbol{A}}_\beta, \boldsymbol{A}_\star$)	Test Error	$r_{\widehat{\boldsymbol{A}}_\beta}$
Proposed_$\widehat{\boldsymbol{\Theta}}_{\mathsf{Win},\beta}$_0.05	1.3975 (0.0142)	0.2375 (0.0035)	114.67 (19.73)
Proposed_$\widehat{\boldsymbol{\Theta}}_\beta$_0.05	1.3909 (0.0064)	0.2391 (0.0023)	90.04 (6.51)
Proposed_$\widehat{\boldsymbol{\Theta}}_{\mathsf{Win},\beta}$_0.1	1.3878 (0.0078)	0.2354 (0.0023)	100.69 (16.20)
Proposed_$\widehat{\boldsymbol{\Theta}}_\beta$_0.1	1.3852 (0.0062)	0.2375 (0.0022)	81.79 (4.75)
Proposed_$\widehat{\boldsymbol{\Theta}}_\alpha$	1.4024 (0.0242)	0.2389 (0.0062)	115.40 (22.21)
Proposed_$\widehat{\boldsymbol{\Theta}}_{\mathsf{1\text{-}bit,Win},\beta}$_0.05	1.4068 (0.0062)	0.2430 (0.0022)	98.97 (2.55)
Proposed_$\widehat{\boldsymbol{\Theta}}_{\mathsf{1\text{-}bit},\beta}$_0.05	1.3920 (0.0072)	0.2383 (0.0027)	97.88 (6.06)
Proposed_$\widehat{\boldsymbol{\Theta}}_{\mathsf{1\text{-}bit,Win},\beta}$_0.1	1.4121 (0.0062)	0.2449 (0.0022)	105.50 (1.16)
Proposed_$\widehat{\boldsymbol{\Theta}}_{\mathsf{1\text{-}bit},\beta}$_0.1	1.3913 (0.0064)	0.2383 (0.0023)	100.94 (7.12)
Proposed_$\widehat{\boldsymbol{\Theta}}_{\mathsf{1\text{-}bit},\alpha}$	1.3894 (0.0084)	0.2353 (0.0029)	113.92 (11.35)
NW	1.9844 (0.2217)	0.4568 (0.1003)	279.45 (47.64)
KLT	2.3207 (0.0053)	0.5964 (0.0016)	1.00 (0.00)
MHT	1.5083 (0.0084)	0.2857 (0.0033)	77.47 (5.31)
(n_1, n_2) = (1200, 1200)	RMSE($\widehat{\boldsymbol{A}}_\beta, \boldsymbol{A}_\star$)	Test Error	$r_{\widehat{\boldsymbol{A}}_\beta}$
Proposed_$\widehat{\boldsymbol{\Theta}}_{\mathsf{Win},\beta}$_0.05	1.3389 (0.0168)	0.2171 (0.0040)	135.84 (25.41)
Proposed_$\widehat{\boldsymbol{\Theta}}_\beta$_0.05	1.3226 (0.0057)	0.2157 (0.0020)	106.13 (5.81)
Proposed_$\widehat{\boldsymbol{\Theta}}_{\mathsf{Win},\beta}$_0.1	1.3270 (0.0073)	0.2148 (0.0019)	112.28 (19.72)
Proposed_$\widehat{\boldsymbol{\Theta}}_\beta$_0.1	1.3144 (0.0054)	0.2135 (0.0018)	97.71 (5.49)
Proposed_$\widehat{\boldsymbol{\Theta}}_\alpha$	1.3453 (0.0287)	0.2187 (0.0071)	138.51 (29.08)
Proposed_$\widehat{\boldsymbol{\Theta}}_{\mathsf{1\text{-}bit,Win},\beta}$_0.05	1.3415 (0.0054)	0.2202 (0.0019)	115.63 (1.37)
Proposed_$\widehat{\boldsymbol{\Theta}}_{\mathsf{1\text{-}bit},\beta}$_0.05	1.3237 (0.0066)	0.2146 (0.0025)	115.07 (8.29)
Proposed_$\widehat{\boldsymbol{\Theta}}_{\mathsf{1\text{-}bit,Win},\beta}$_0.1	1.3489 (0.0054)	0.2226 (0.0019)	125.48 (1.04)
Proposed_$\widehat{\boldsymbol{\Theta}}_{\mathsf{1\text{-}bit},\beta}$_0.1	1.3259 (0.0058)	0.2157 (0.0019)	119.25 (10.60)
Proposed_$\widehat{\boldsymbol{\Theta}}_{\mathsf{1\text{-}bit},\alpha}$	1.3289 (0.0103)	0.2141 (0.0025)	137.05 (17.28)
NW	1.9273 (0.1956)	0.4399 (0.0877)	323.36 (48.91)
KLT	2.3494 (0.0044)	0.6041 (0.0013)	1.00 (0.00)
MHT	1.4649 (0.0062)	0.2706 (0.0024)	84.03 (4.49)

[2] With $r_{\boldsymbol{M}_\star} = 11$, $r_{\boldsymbol{A}_\star} = 11$, (n_1, n_2) = (1000, 1000), (1200, 1200) and SNR = 1. The three existing methods are proposed respectively in Negahban and Wainwright (2012)(NW), Koltchinskii et al. (2011)(KLT) and Mazumder et al. (2010)(MHT)

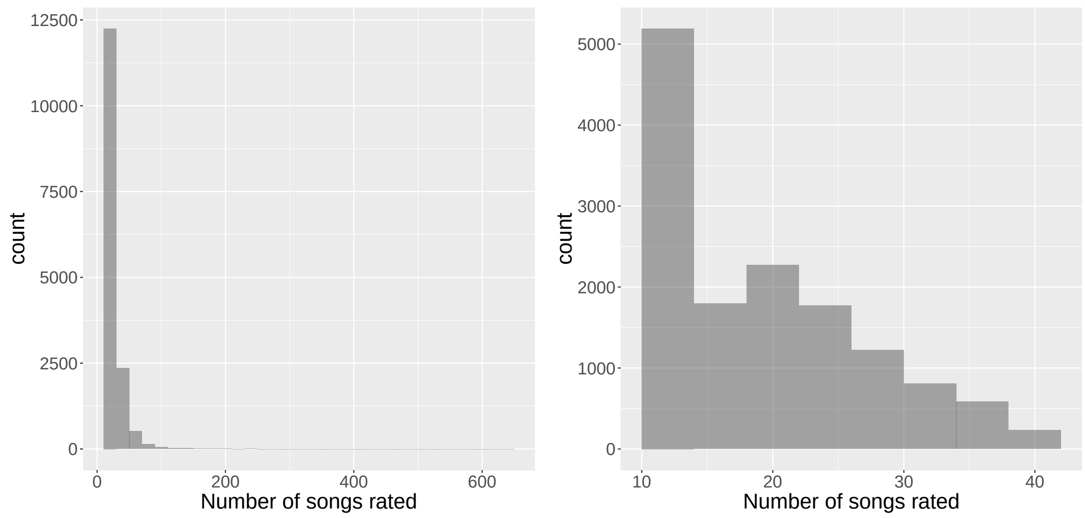

Figure 1: Left: The histogram of the number of songs rated per user in the Yahoo! Webscope dataset. Right: Similar to the left figure but restricted to no more than 40 songs rated per user.

With the separation of μ, Proposed_$\widehat{\boldsymbol{\Theta}}_{\alpha}$ is better than Proposed_$\widehat{\boldsymbol{\Theta}}_{\text{1-bit},\alpha}$; analogously, Proposed_$\widehat{\boldsymbol{\Theta}}_{\beta}$_$t$ is better than Proposed_$\widehat{\boldsymbol{\Theta}}_{\text{1-bit},\beta}$_$t$ with different constraint level t, same to Proposed_$\widehat{\boldsymbol{\Theta}}_{\text{Win},\beta}$_$s$ and Proposed_$\widehat{\boldsymbol{\Theta}}_{\text{1-bit,Win},\beta}$_$s$ with different winsorization level s.

As compared with the existing methods NW, KLT and MHT, our proposed methods perform significantly better in terms of RMSPEs, and achieve as much as 25% improvement when compared with MHT (the best among the three existing methods). This suggests that a more flexible modeling of missing structure improves the prediction power.

Table 4: Root mean squared prediction errors (RMSPEs) based on Yahoo! Webscope dataset for the ten versions of the proposed method and the three existing methods proposed respectively in Negahban and Wainwright (2012)(NW), Koltchinskii et al. (2011)(KLT) and Mazumder et al. (2010)(MHT).

	Proposed_$\widehat{\boldsymbol{\Theta}}_{\text{Win},\beta}$_0.05	Proposed_$\widehat{\boldsymbol{\Theta}}_{\beta}$_0.05	Proposed_$\widehat{\boldsymbol{\Theta}}_{\text{Win},\beta}$_0.1	
RMSPE	1.0396	1.0381	1.0476	
	Proposed_$\widehat{\boldsymbol{\Theta}}_{\beta}$_0.1	Proposed_$\widehat{\boldsymbol{\Theta}}_{\alpha}$	Proposed_$\widehat{\boldsymbol{\Theta}}_{\text{1-bit,Win},\beta}$_0.05	
RMSPE	1.0490	1.0383	1.0831	
	Proposed_$\widehat{\boldsymbol{\Theta}}_{\text{1-bit},\beta}$_0.05	Proposed_$\widehat{\boldsymbol{\Theta}}_{\text{1-bit,Win},\beta}$_0.1	Proposed_$\widehat{\boldsymbol{\Theta}}_{\text{1-bit},\beta}$_0.1	
RMSPE	1.1091	1.0760	1.0523	
	Proposed_$\widehat{\boldsymbol{\Theta}}_{\text{1-bit},\alpha}$	NW	KLT	MHT
RMSPE	1.1065	1.7068	3.6334	1.3821

8 Concluding Remarks

When the matrix entries are heterogeneously observed due to selection bias, this heterogeneity should be taken into account. This paper focuses on the problem of matrix completion under low-rank missing structure. In the recovery of probabilities of observation, we adopt a generalized linear model with a low-rank linear predictor matrix. To avoid unnecessary bias, we introduce a separation of the mean effect μ. As the extreme values of probabilities may lead to unstable estimation of target matrix, we propose an IPW-based method with constrained probability estimates and demonstrate the improvements in empirical perspectives. Our theoretical result shows that the estimator of the high dimensional probability matrix can be embedded into the IPW framework without compromising the rate of convergence of the target matrix (for an appropriately tuned $\beta > 0$), and reveals a possible regime change in the tuning of the constraint parameter ($\beta > 0$ vs. $\beta = 0$). In addition, corresponding computational algorithms are developed, and a related algorithmic convergence result is established. Empirical studies show the attractive performance of the proposed methods as compared with existing matrix completion methods.

References

Beck, A. and Teboulle, M. (2009), "A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems," *SIAM Journal on Imaging Sciences*, 2, 183–202.

Bi, X., Qu, A., Wang, J., and Shen, X. (2017), "A Group-Specific Recommender System," *Journal of the American Statistical Association*, 112, 1344–1353.

Cai, J.-F., Candès, E. J., and Shen, Z. (2010), "A Singular Value Thresholding Algorithm for Matrix Completion," *SIAM Journal on Optimization*, 20, 1956–1982.

Cai, T., Cai, T. T., and Zhang, A. (2016), "Structured Matrix Completion with Applications to Genomic Data Integration," *Journal of the American Statistical Association*, 111, 621–633.

Cai, T. T. and Zhou, W.-X. (2016), "Matrix Completion via Max-Norm Constrained Optimization," *Electronic Journal of Statistics*, 10, 1493–1525.

Candès, E. J. and Plan, Y. (2010), "Matrix Completion with Noise," *Proceedings of the IEEE*, 98, 925–936.

Candès, E. J. and Recht, B. (2009), "Exact Matrix Completion via Convex Optimization," *Foundations of Computational Mathematics*, 9, 717–772.

Chen, C., He, B., Ye, Y., and Yuan, X. (2016), "The Direct Extension of ADMM for Multi-block Convex Minimization Problems is Not Necessarily Convergent," *Mathematical Programming*, 155, 57–79.

Davenport, M. A., Plan, Y., van den Berg, E., and Wootters, M. (2014), "1-Bit Matrix Completion," *Information and Inference*, 3, 189–223.

Kang, J. D. and Schafer, J. L. (2007), "Demystifying Double Robustness: A Comparison of Alternative Strategies for Estimating a Population Mean from Incomplete Data," *Statistical science*, 22, 523–539.

Keshavan, R. H., Montanari, A., and Oh, S. (2009), "Matrix Completion from Noisy Entries," in *Advances in Neural Information Processing Systems*, pp. 952–960.

Klopp, O. (2014), "Noisy Low-Rank Matrix Completion with General Sampling Distribution," *Bernoulli*, 20, 282–303.

Koltchinskii, V., Lounici, K., and Tsybakov, A. B. (2011), "Nuclear-Norm Penalization and Optimal Rates for Noisy Low-Rank Matrix Completion," *The Annals of Statistics*, 39, 2302–2329.

Mao, X., Chen, S. X., and Wong, R. K. (2018), "Matrix Completion with Covariate Information," *Journal of the American Statistical Association*, in press.

Mazumder, R., Hastie, T., and Tibshirani, R. (2010), "Spectral Regularization Algorithms for Learning Large Incomplete Matrices," *Journal of Machine Learning Research*, 11, 2287–2322.

Negahban, S. and Wainwright, M. J. (2012), "Restricted Strong Convexity and Weighted Matrix Completion: Optimal Bounds with Noise," *Journal of Machine Learning Research*, 13, 1665–1697.

Potter, F. J. (1990), "A Study of Procedures to Identify and Trim Extreme Sampling Weights," in *Proceedings of the American Statistical Association, Section on Survey Research Methods*, vol. 225230.

Recht, B. (2011), "A Simpler Approach to Matrix Completion," *Journal of Machine Learning Research*, 12, 3413–3430.

Rohde, A. and Tsybakov, A. B. (2011), "Estimation of High-Dimensional Low-Rank Matrices," *The Annals of Statistics*, 39, 887–930.

Rubin, D. B. (2001), "Using Propensity Scores to Help Design Observational Studies: Application to the Tobacco Litigation," *Health Services and Outcomes Research Methodology*, 2, 169–188.

Schafer, J. L. and Kang, J. (2008), "Average Causal Effects from Nonrandomized Studies: a Practical Guide and Simulated Example." *Psychological methods*, 13, 279.

Scharfstein, D. O., Rotnitzky, A., and Robins, J. M. (1999), "Adjusting for Nonignorable Drop-Out Using Semiparametric Nonresponse Models," *Journal of the American Statistical Association*, 94, 1096–1120.

Schnabel, T., Swaminathan, A., Singh, A., Chandak, N., and Joachims, T. (2016), "Recommendations as Treatments: Debiasing Learning and Evaluation," *arXiv preprint arXiv:1602.05352*.

Srebro, N. and Salakhutdinov, R. R. (2010), "Collaborative Filtering in a Non-Uniform World: Learning with the Weighted Trace Norm," in *Advances in Neural Information Processing Systems*, pp. 2056–2064.