Statistica Sinica **31** (2021), 2005-2030
doi:https://doi.org/10.5705/ss.202019.0196

MATRIX COMPLETION UNDER LOW-RANK MISSING MECHANISM

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

Fudan University, Texas A&M University and Peking University

Abstract: Matrix completion is a modern missing-data problem in which both the missing structure and the underlying parameter are high dimensional. Despite the missing structure being a key component of any missing-data problem, existing matrix-completion methods often assume a simple uniform missing mechanism. In this work, we study matrix completion from corrupted data under a novel low-rank missing mechanism. The observation probability matrix is estimated using a high-dimensional, low-rank matrix-estimation procedure, and then used to complete the target matrix via inverse probability weighting. Owing to the high-dimensional and extreme (i.e., very small) nature of the true probability matrix, the effect of inverse probability weighting requires careful study. Lastly, we derive optimal asymptotic convergence rates of the proposed estimators for both the observation probabilities and the target matrix.

Key words and phrases: Low-rank, missing, 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, with applications including collaborative filtering, computer vision and positioning. From a statistical viewpoint, it is a high-dimensional, missing-data problem in which a high percentage of matrix entries are missing. As in many missing-data problems, the underlying missing mechanism plays an important role. Most existing works (e.g., Candès and Recht (2009); Keshavan, Montanari and Oh (2010); Recht (2011); Rohde and Tsybakov (2011); Koltchinskii, Lounici and Tsybakov (2011)) adopt a uniform observation mechanism, where each entry has the same marginal probability of being observed. This leads to significant simplifications, and has enabled the domain to move forward rapidly, with various theoretical breakthroughs occurring in the last decade. However, a uniform mechanism is often unrealistic. Recent works (Foygel et al. (2011); Negahban and Wainwright (2012); Klopp (2014); Cai and Zhou (2016);

Corresponding author: Song Xi Chen, Department of Business Statistics and Econometrics, Peking University, Beijing 100651, China. E-mail: csx@gsm.pku.edu.cn.

Cai, Cai and Zhang (2016); Bi et al. (2017); Mao, Chen and Wong (2019)) have tried to relaxing this restrictive assumption by adopting other missing structures. The use of these settings hinges on having strong prior knowledge of the underlying problem. At a high level, many of these works use a special form of low-rank structure for the missing mechanism. For instance, Foygel et al. (2011) and Negahban and Wainwright (2012) both adopt a rank-one structure based on the estimated marginal probabilities. In this study, we aim to recover the target matrix $\boldsymbol{A}_\star$ under a flexible, high-dimensional, low-rank sampling structure. This is achieved by using a weighted empirical risk minimization and by applying inverse probability weighting (e.g., Schnabel et al. (2016); Mao, Chen and Wong (2019)) to adjust for the effect of non-uniform missingness.

Data arising in many applications of matrix completion, such as recommender systems, usually possess a complex "sampling" structure that is largely unknown. In a movie recommender system, users tend to rate movies that they prefer or dislike the most, while often remaining "silent" about other movies. Another example of a complex sampling regime is that of online merchandising. Here some users may purchase certain items regularly without often rating them, but often evaluate products that they rarely buy. Similarly 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, leading to a low-rank model of the missing structure.

Inspired by generalized linear models, we model the probability of observation $\boldsymbol{\Theta}_\star = (\theta_{\star,ij})_{i,j=1}^{n_1,n_2} \in (0,1)^{n_1 \times n_2}$ using a high-dimensional, low-rank matrix $\boldsymbol{M}_\star = (m_{\star,ij})_{i,j=1}^{n_1,n_2} \in \mathbb{R}^{n_1 \times n_2}$ with a known function f. Therefore, at the entry level, we have $\theta_{\star,ij} = f(m_{\star,ij})$. In generalized linear models, the linear predictor $m_{\star,ij}$ is further modeled as a linear function of the observed covariates. However, to reflect the difficulty of attaining (appropriate and adequate) covariate information and the complexity of modeling $\boldsymbol{\Theta}_\star$ in some situations of the matrix completion, we assume the predictor matrix $\boldsymbol{M}_\star$ is completely hidden. Despite $\boldsymbol{M}_\star$ being hidden, the low rank and high dimensionality of $\boldsymbol{M}_\star$ allow both identification and consistent estimation of $\boldsymbol{\Theta}_\star$, which facilitates matrix completion based on inverse probability weighting. Motivated by the nature of matrix completion, we propose a novel parametrization $\boldsymbol{M}_\star = \mu_\star \mathbf{1}_{n_1} \mathbf{1}_{n_2}^\top + \boldsymbol{Z}_\star$, where $\boldsymbol{Z}_\star$ satisfies $\mathbf{1}_{n_1}^\top \boldsymbol{Z}_\star \mathbf{1}_{n_2} = 0$. Our proposal extends the work of Davenport et al. (2014), who solve a binary matrix-completion problem and pursue a different goal. In contrast to that of Davenport et al. (2014), the proposed method does not regularize the estimation of $\mu_\star$. Instead, it regularizes the nuclear norm of the estimation of $\boldsymbol{Z}_\star$, which requires a different algorithmic treatment to avoid bias caused by the

nuclear-norm penalty.

Three fundamental aspects that set our work apart from existing works on matrix completion: (i) the high-dimensional probability matrix $\boldsymbol{\Theta}$, the dimensions of which, n_1, n_2, go to infinity in our asymptotic setting; (ii) the diminishing lower bound of the observation probabilities (as n_1, n_2 go to infinity), and added issue to the instability of inverse probability weighting; (iii) the effects of the estimation error in the inverse probability weighting of the matrix completion procedure. Aspects (i) and (ii) are unique to our problem, and not found in the literature on missing data. Works related to aspect (iii) are sparse in the literature on matrix completion. Noted that Mao, Chen and Wong (2019) focused on the low-dimensional parametric modeling of inverse probability weighting with observable covariates.

We develop non-asymptotic upper bounds for the mean squared errors of the proposed estimators of the observation probabilities and the target matrix. To sustain the convergence rate of the target matrix under the high dimensionality of $\boldsymbol{M}_\star$ and the low levels of the observation probabilities, we propose re-estimating $\boldsymbol{Z}_\star$ by constraining the magnitude of its entries to a smaller threshold. Our analysis shows that the proposed constrained inverse probability weighting estimator achieves the optimal rate (up to a logarithmic factor in the estimation of the target matrix). We also compare the completion based on inverse probability weighting and the proposed constrained estimation with that 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)). As such, we show that the constrained estimation has both theoretical and empirical advantages.

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 context 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 zero otherwise, and assume $\{\epsilon_{ij}\}$ are independent of $\{w_{ij}\}$.

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 $\boldsymbol{\Theta}_\star = (\theta_{\star,ij})_{i,j=1}^{n_1,n_2} \in (0,1)^{n_1\times n_2}$. Similarly to generalized linear models, 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 prespecified monotone and differentiable function $f : \mathbb{R} \to [0,1]$; that is, $\theta_{\star,ij} = f(m_{\star,ij})$, for all i, j. The matrix $\boldsymbol{M}_\star$ plays the same role as a linear predictor in the generalized linear model. The function f is an inverse link function. Two popular choices of f are the 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 assumptions, it is virtually impossible to recover the hidden feature matrix $\boldsymbol{M}_\star$ or the target matrix $\boldsymbol{A}_\star$. A common and powerful assumption is that $\boldsymbol{A}_\star$ is a low-rank matrix; that is, $\operatorname{rank}(\boldsymbol{A}_\star) \ll \min\{n_1, n_2\}$. For example, consider the Yahoo! Webscope data set (see Section 7). This data set contains a partially observed matrix of ratings from 15,400 users fors 1,000 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 proven useful in modeling recommender systems (e.g., Candès and Plan (2010); Cai, Candès and Shen (2010)).

The same idea can be adapted to the missing pattern, despite the factors that induce the missingness possible differing from those that generate the ratings. To this end, we assume $\boldsymbol{M}_\star$ is of low rank. Next, we decompose $\boldsymbol{M}_\star$ as

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

where $\mathbf{1}_n$ is an n-vector of ones, and $\boldsymbol{J} = \mathbf{1}_{n_1}\mathbf{1}_{n_2}^{\mathrm{T}}$. Here, $\mu_\star$ is the mean of $\boldsymbol{M}_\star$; that is, $\mu_\star = \mathbf{1}_{n_1}^{\mathrm{T}} \boldsymbol{M}_\star \mathbf{1}_{n_2}/(n_1 n_2)$. Note that this decomposition holds for any matrix $\boldsymbol{M}$ by setting $\mu = (n_1 n_2)^{-1}\mathbf{1}_{n_1}^{\mathrm{T}} \boldsymbol{M} \mathbf{1}_{n_2}$ and $\boldsymbol{Z} = \boldsymbol{M} - \mu \boldsymbol{J}$. Moreover, the decomposition is unique, owing to the constraint that $\mathbf{1}_{n_1}^{\mathrm{T}} \boldsymbol{Z}_\star \mathbf{1}_{n_2} = 0$. The key here is to reparametrize $\boldsymbol{M}_\star$ in terms of $\mu_\star$ and $\boldsymbol{Z}_\star$, which require different treatments in their estimations. See Section 3 for details. Furthermore, the low-rankness of $\boldsymbol{M}_\star$ can be translated to the low-rankness of $\boldsymbol{Z}_\star$.

Note that the rank of $\boldsymbol{M}_\star$ is not the same as that of $\boldsymbol{\Theta}_\star$, owing to the nonlinear transformation f. In general, the low-rank structure of $\boldsymbol{M}_\star$ implies a specific low-

dimensional nonlinear structure of $\boldsymbol{\Theta}_\star$. In a common high-missingness scenario, most entries of $\boldsymbol{M}_\star$ are significant and negative, where many common choices of the inverse link function can be well approximated by a linear function. Thus our modeling can be regarded as a low-rank modeling of $\boldsymbol{\Theta}_\star$.

There are several related, but more specialized models. Srebro and Salakhutdinov (2010) and Negahban and Wainwright (2012) use an independent row and column sampling mechanism, leading to a rank-one structure for $\boldsymbol{\Theta}_\star$. Cai, Cai and Zhang (2016) consider a matrix block structure for $\boldsymbol{\Theta}_\star$ and hence $\boldsymbol{M}_\star$, that can be regarded as a special case of low-rank modeling. Mao, Chen and Wong (2019) examine the case when the missingness depends on observable covariates, and adopt a low-rank modeling with a known row space of $\boldsymbol{M}_\star$. In this study, we focus on the situation in which the missingness is dependent on some hidden factors, reflecting situations when obvious covariates are unknown or not available.

2.3. Inverse probability weighting-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 a uniform missing structure. Hence, they use 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, Lounici and Tsybakov (2011); Mazumder, Hastie and Tibshirani (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 R. If $\boldsymbol{\Theta}_\star$ 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 motivates the use of inverse probability weighting in matrix completion, as in Mao, Chen and Wong (2019). In addition, our theoretical analysis shows that

the nuclear-norm-regularized empirical risk estimator (defined later) based on $\widehat{R}$ (assuming the use of true observation probabilities) improves upon the existing error upper bound of the corresponding estimator based on $\widehat{R}_{\mathrm{UNI}}$ achieved by Klopp (2014), as shown in Section 5.3. However, the inverse probability weights $\boldsymbol{\Theta}_\star^\dagger$ are often unknown and have to be estimated, which has to be conducted carefully in the context of matrix completion.

Despite the popularity of inverse probability weighting in the missing-data literature, it is known to produce unstable estimations, owing to the occurrence of small probabilities (e.g., Rubin (2001); Kang and Schafer (2007); Schafer and Kang (2008)). This problematic scenario is common in matrix-completion problems in which we attempt 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 nonstandard setup of inverse probability weighting. Therefore, 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 a small θ_0 increases the difficulty of the estimation, $\widehat{R}(\boldsymbol{A})$ changes only up to a multiplicative constant. However, in a non-uniform setting, this is not as straightforward, owing to the heterogeneity among $\{\theta_{\star,ij}\}$. To demonstrate the issue, we examine the Yahoo! Webscope data set described in Section 7. A 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 $\theta_U = \max_{i,j} \theta_{\star,ij}$. The corresponding ratio of estimated probabilities $\widehat{\theta}_U/\widehat{\theta}_L$ based on the rank-one structure of Negahban and Wainwright (2012) is 25,656.2, and that based on our proposed method (without re-estimation, to be described below) is 23,988.0. This strong heterogeneity can jeopardize the convergence rate of our estimator, which was address properly in our framework.

In the following section, we propose an estimation approach for $\boldsymbol{\Theta}_\star$ 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 on the framework of a regularized maximum likelihood. Given the inverse of the 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

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

for any $\boldsymbol{M} = (m_{ij})_{i,j=1}^{n_1,n_2} \in \mathbb{R}^{n_1 \times n_2}$, where $\mathbb{I}_{\mathcal{A}}$ is the indicator of an event $\mathcal{A}$. Owing to the low-rank assumption of $\boldsymbol{M}_\star$, a natural candidate as an estimator for $\boldsymbol{M}_\star$ 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 $\|\boldsymbol{M}\|_\infty \le \alpha$, for some $\alpha > 0$, in the maximization (e.g., Davenport et al. (2014)). Note that the nuclear norm penalty favors $\boldsymbol{M} = 0$, corresponding to $\Pr(w_{ij} = 1) = 0.5$, for all i, j. Nevertheless, this does not align well with common settings of matrix completion, under which the average probability of observations is quite small, resulting in a large bias. In view of this, we instead adopt a parametrization $\boldsymbol{M}_\star = \mu_\star \boldsymbol{J} + \boldsymbol{Z}_\star$, and consider the following estimator of $(\mu_\star, \boldsymbol{Z}_\star)$:

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

where, $\mathcal{C}_{n_1,n_2}(\alpha_1, \alpha_2) = \{(\mu, \boldsymbol{Z}) \in \mathbb{R} \times \mathbb{R}^{n_1 \times n_2} : |\mu| \le \alpha_1,\ \|\boldsymbol{Z}\|_\infty \le \alpha_2,\ \mathbf{1}_{n_1}^{\mathrm{T}} \boldsymbol{Z} \mathbf{1}_{n_2} = 0\}$. Note that the mean μ of the linear predictor $\mu\boldsymbol{J} + \boldsymbol{Z}$ is not penalized. The constraint $\mathbf{1}_{n_1}^{\mathrm{T}} \boldsymbol{Z} \mathbf{1}_{n_2} = 0$ ensures the identifiability of μ and $\boldsymbol{Z}$. 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 $\boldsymbol{Z}$. With $(\widehat{\mu}, \widehat{\boldsymbol{Z}})$, the corresponding estimator of $\boldsymbol{M}_\star$ is $\widehat{\boldsymbol{M}} = \widehat{\mu}\boldsymbol{J} + \widehat{\boldsymbol{Z}}$.

Davenport et al. (2014) considered a regularized maximum likelihood approach for a binary matrix completion problem. Their goal was different, because they aimed at recovering a binary rating matrix in lieu of the missing structure, and considered a regularization on $\boldsymbol{M}$ (instead of $\boldsymbol{Z}$) via $\|\boldsymbol{M}\|_* \le \alpha'\{\mathrm{rank}(\boldsymbol{M}_\star) n_1 n_2\}^{1/2}$. 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 Section 5.

3.2. Computational algorithm and tuning parameter selection

To solve the optimization given in (3.1), we begin with the observation that $\ell_{\boldsymbol{W}}$ is a smooth concave function, which allows us to use an iterative algorithm

called the 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, as follows:

$$\begin{aligned} P_L\left\{(\mu,\boldsymbol{Z}),(\mu_{\text{old}},\boldsymbol{Z}_{\text{old}})\right\} = &-\ell_{\boldsymbol{W}}\left(\mu_{\text{old}}\boldsymbol{J}+\boldsymbol{Z}_{\text{old}}\right) \\ &+(\mu-\mu_{\text{old}})\,\mathbf{1}_{n_1}^{\mathrm{T}}\left\{-\nabla_\mu \ell_{\boldsymbol{W}}\left(\mu_{\text{old}}\boldsymbol{J}+\boldsymbol{Z}_{\text{old}}\right)\right\}\mathbf{1}_{n_2} \\ &+\frac{Ln_1n_2}{2}(\mu-\mu_{\text{old}})^2 \\ &+\left\langle \boldsymbol{Z}-\boldsymbol{Z}_{\text{old}}, -\nabla_{\boldsymbol{Z}}\ell_{\boldsymbol{W}}\left(\mu_{\text{old}}\boldsymbol{J}+\boldsymbol{Z}_{\text{old}}\right)\right\rangle \\ &+\frac{L}{2}\|\boldsymbol{Z}-\boldsymbol{Z}_{\text{old}}\|_F^2+\lambda\|\boldsymbol{Z}\|_*, \end{aligned}$$

where $L>0$ is an algorithmic parameter determining the step size of the proximal gradient algorithm, and is chosen using 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 the same dimensions.

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

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

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

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

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

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

which is equivalent to

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

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

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

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

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

where $u > 0$ is an algorithmic parameter, and $\mathbb{I}_{\mathcal{A}}$ is equal to zero if the constraint $\mathcal{A}$ holds, and ∞ otherwise. The detailed algorithm to solve (3.3) is provided as Algorithm 1 in the Supplementary Material. Noted that, in general, the multi-block alternative direction method of multipliers may fail to converge for some $u > 0$ (Chen et al. (2016)). In such cases, an appropriate selection of u is crucial. However, we are able to show that the form of our 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, the proof of which 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 *in the Supplementary Material, converges to the global optimum of* (3.3).

Note that the alternative direction method of multipliers algorithm is nested within the proximal gradient algorithm. However, our numerical experiments show that the numbers of inner iterations (alternative direction method of multipliers) and outer iterations (proximal gradient) are both small, usually less than 20. We summarize the corresponding convergence result of the overall proximal gradient algorithm in the following theorem, the proof of which is provided in the Supplementary Material.

Theorem 2. *The estimator* $(\widehat{\mu}, \widehat{\boldsymbol{Z}})$ *generated by the proximal gradient algorithm converges to the global optimum of* (3.1).

The tuning parameters α_1 and α_2 can be chosen based on prior knowledge of the problem setup, if available. When a prior knowledge is not available, one can choose large values for these parameters. Once these parameters are sufficiently large, 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. For λ, we adopt the Akaike information criterion (AIC), where we approximate the degrees of freedom 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 $\mathcal{F}(\boldsymbol{M}) = (f(m_{ij}))_{i,j=1}^{n_1,n_2} \in \mathbb{R}^{n_1 \times n_2}$, for any $\boldsymbol{M} = (m_{ij})_{i,j=1}^{n_1,n_2} \in \mathbb{R}^{n_1 \times n_2}$, and then to replace $\boldsymbol{\Theta}_\star^\ddagger$ with $\widehat{\boldsymbol{\Theta}}^\ddagger = (\hat{\theta}_{ij}^{-1/2})_{i,j=1}^{n_1,n_2}$. However, this direct implementation is not robust to extremely small probabilities of observation, and our theoretical analysis shows that it could lead to a slower convergence rate in the estimation of $\boldsymbol{A}_\star$. In the literature om missing data, a simple solution is to winsorize the small probabilities (Potter (1990); Scharfstein, Rotnitzky and Robins (1999)).

In the estimation of $\widehat{\boldsymbol{\Theta}}$ defined in (3.1) that assumes $\|\boldsymbol{Z}_\star\|_\infty \leq \alpha_2$, 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 is affected. That is, even though the extreme probabilities can be controlled by winsorizing, the unchanged entries of $\widehat{\boldsymbol{Z}}$ (in the winsorizing procedure) may already suffer from a slower rate of convergence. This results in a larger estimation error under certain settings of missingness, as revealed in Section 5.

A seemingly better strategy is to impose a tighter constraint directly in the minimization problem given in (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 entries of magnitude smaller than β. However, the mean-zero constraint of $\boldsymbol{Z}$ no longer makes sense, because 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:

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

Note that we only re-compute $\boldsymbol{Z}$, but not μ, which allows us to drop the mean-zero constraint. Thus, $\widehat{\boldsymbol{M}}_\beta = \widehat{\mu}\boldsymbol{J} + \widehat{\boldsymbol{Z}}_\beta$. The corresponding algorithm for optimization (3.4) can be derived similarly to that in Davenport et al. (2014), and is provided in the Supplementary Material. 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 $\boldsymbol{\Theta}_\star^\ddagger$ with $\widehat{\boldsymbol{\Theta}}_\beta^\ddagger$ to obtain a modified empirical risk:

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

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

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

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

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

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

We use an accelerated proximal gradient algorithm (Beck and Teboulle (2009)) to solve (4.3). For the choice of the tuning parameter τ in (4.3), we adopt a five-fold cross-validation to select the remaining tuning parameters. Owing to the non-uniform missing mechanism, we use a weighted version of the validation errors; see Algorithm 3 in the Supplementary Material.

5. Theoretical Properties

5.1. Probabilities of observation

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

To investigate the asymptotic properties of $\widehat{\boldsymbol{M}}_\beta$ and $\widehat{\boldsymbol{\Theta}}_\beta^\dagger$ defined in Section 3, 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 dis-

tribution 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 convenience in the theoretical analysis, we consider an equivalent estimator of $(\mu_\star, \boldsymbol{Z}_\star)$ defined by the constrained maximization problem (5.1) instead of the Lagrangian form (3.1). For $r_{\boldsymbol{Z}_\star} \le \min\{n_1, n_2\}$ and $\alpha_1, \alpha_2 \ge 0$,

$$\left(\widehat{\mu}, \widehat{\boldsymbol{Z}}\right) = \underset{(\mu,\boldsymbol{Z}) \in \widetilde{\mathcal{C}}_{n_1,n_2}(\alpha_1,\alpha_2)}{\operatorname{argmax}} \ell_{\boldsymbol{W}}\left(\mu \boldsymbol{J} + \boldsymbol{Z}\right), \text{ where} \tag{5.1}$$

$$\widetilde{\mathcal{C}}_{n_1,n_2}(\alpha_1,\alpha_2) = \big\{(\mu,\boldsymbol{Z}) \in \mathbb{R} \times \mathbb{R}^{n_1 \times n_2} : |\mu| \le \alpha_1, \|\boldsymbol{Z}\|_\infty \le \alpha_2, \|\boldsymbol{Z}\|_* \le \alpha_2 \sqrt{r_{\boldsymbol{Z}_\star} n_1 n_2}, \mathbf{1}_{n_1}^{\mathrm{T}} \boldsymbol{Z} \mathbf{1}_{n_2} = 0\big\}.$$

It is easy to see that we have $(\mu_\star, \boldsymbol{Z}_\star) \in \widetilde{\mathcal{C}}_{n_1,n_2}(\alpha_1,\alpha_2)$ once $(\mu_\star, \boldsymbol{Z}_\star) \in \mathcal{C}_{n_1,n_2}(\alpha_1,\alpha_2)$ holds. For ease of 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; the corresponding results under general settings of n_1, n_2, and f are delegated to Section S1.3 in the Supplementary Material. We first establish the convergence results for $\widehat{\mu}$, $\widehat{\boldsymbol{Z}}$, and $\widehat{\boldsymbol{M}}$. To simplify the notation, let $\alpha_0 = \alpha_1 + \alpha_2$, $h_{\alpha_1,\beta} = (1 + e^{\alpha_1+\beta})^{-1}$, and $\Gamma_n = e^{\alpha_0}(\alpha_1 + \alpha_2 r_{\boldsymbol{Z}_\star}^{1/2}) n^{-1/2}$.

Lemma 1. *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* (5.1). *There exist positive constants* C_1, C_2, *such that we have with probability at least* $1 - C_1/n$,

$$(\mu_\star - \widehat{\mu})^2 \le C_2\left(\alpha_1^2 \wedge \Gamma_n\right), \quad \frac{1}{n^2}\left\|\widehat{\boldsymbol{Z}} - \boldsymbol{Z}_\star\right\|_F^2 \le C_2\left(\alpha_2^2 \wedge \Gamma_n\right)$$

$$\text{and} \quad \frac{1}{n^2}\left\|\widehat{\boldsymbol{M}} - \boldsymbol{M}_\star\right\|_F^2 \le C_2\left(\alpha_0^2 \wedge \Gamma_n\right). \tag{5.2}$$

The upper bounds in (5.2) all consist of trivial bounds α_j^2 and a more dedicated bound Γ_n. The trivial upper bounds α_1^2, α_2^2, and α_0^2 can be derived easily 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 Γ_n is diverging and the trivial bounds may provide better control. The term Γ_n can be controlled by the rank of $\boldsymbol{Z}_\star$. For a range of non-extreme scenarios, that is, $\alpha_0 \le 1/2 \log n$ or $\theta_L \ge n^{-1/2}$, the second term in Γ_n

attains the convergence order once $r_{\boldsymbol{Z}_\star} = O(1)$.

Similarly, we study the theoretical results of the re-estimation of $\boldsymbol{Z}_\star$ in terms of the constrained optimization:

$$\widehat{\boldsymbol{Z}}_\beta = \underset{\boldsymbol{Z}\in\mathbb{R}^{n_1\times n_2}}{\operatorname{argmax}} \ell_{\boldsymbol{W}}\left(\widehat{\mu}\boldsymbol{J}+\boldsymbol{Z}\right) \text{ subject to } \|\boldsymbol{Z}\|_\infty \le \beta, \ \|\boldsymbol{Z}\|_* \le \beta\sqrt{r_{\mathcal{T}_\beta(\boldsymbol{Z}_\star)} n_1 n_2}. \tag{5.3}$$

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

$$T_\beta(b_{ij}) = b_{ij}\mathbb{I}_{[-\beta\le b_{ij}\le\beta]} + \beta\mathbb{I}_{[b_{ij}>\beta]} - \beta\mathbb{I}_{[b_{ij}<-\beta]}, \quad \text{for any } \beta\ge 0. \tag{5.4}$$

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 $\widehat{\boldsymbol{\Theta}}_{\star,\beta} = \mathcal{F}(\widehat{\boldsymbol{M}}_{\star,\beta})$. 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 $N_\beta = \sum_{i,j}(\mathbb{I}_{[z_{\star,ij}>\beta]} + \mathbb{I}_{[z_{\star,ij}<-\beta]})$ 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. Define $\Lambda_n = \min[\beta^2, \widetilde{\Gamma}_n + h_{\alpha_1,\beta}^{-1} n^{-2}\beta\{8N_\beta + (n^2 - N_\beta)|\mu_\star - \widehat{\mu}|\}]$, where $\widetilde{\Gamma}_n = h_{\alpha_1,\beta}^{-1}(\alpha_1 + \beta r_{\mathcal{T}_\beta(\boldsymbol{Z}_\star)}^{1/2}) n^{-1/2}$.

Theorem 3. *Assume that Conditions* C1−C2 *hold, and* $(\mu_\star, \boldsymbol{Z}_\star) \in \widetilde{\mathcal{C}}_{n_1,n_2}(\alpha_1,\alpha_2)$. *Consider* $\widehat{\boldsymbol{M}}_\beta = \widehat{\mu}\boldsymbol{J} + \widehat{\boldsymbol{Z}}_\beta$, *where* $\widehat{\boldsymbol{Z}}_\beta$ *is the solution to* (5.3) *and* $\beta \ge 0$. *Then there exist some positive constants* C_1, C_2, *and* C_3, *such that we have with probability at least* $1 - 2C_1/n$,

$$d^2\left\{\widehat{\boldsymbol{Z}}_\beta, \mathcal{T}_\beta\left(\boldsymbol{Z}_\star\right)\right\} \le C_3\Lambda_n, \tag{5.5}$$

$$d^2\left(\widehat{\boldsymbol{M}}_\beta, \boldsymbol{M}_\star\right) \le C_2\left(\alpha_1^2 \wedge \Gamma_n\right) + C_3\Lambda_n + \frac{2(\alpha_2-\beta)_+^2 N_\beta}{n^2}$$

$$\text{and} \quad d^2\left(\widehat{\boldsymbol{\Theta}}_\beta^\dagger, \boldsymbol{\Theta}_\star^\dagger\right) \le \frac{C_2}{h_{\alpha_1,\beta}^2}\left(\alpha_1^2 \wedge \Gamma_n\right) + \frac{C_3\Lambda_n}{h_{\alpha_1,\beta}^2} + \frac{8N_\beta}{n^2\theta_L^2}. \tag{5.6}$$

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 1, where $\widehat{\boldsymbol{Z}}_{\mathsf{Win},\beta} = \mathcal{T}_\beta(\widehat{\boldsymbol{Z}})$ is winsorized directly from $\widehat{\boldsymbol{Z}}$. Obviously, the order of this upper bound is larger than or equal to Λ_n. Moreover, there are scenarios in which Λ_n is a smaller order of β^2. To illustrate, assume that $\alpha_1 \asymp 1$ and $\beta \asymp 1$, and 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 $\Lambda_n = o(\beta^2)$.

With a more dedicated investigation of (5.5), one can derive an upper bound for $d^2(\widehat{\boldsymbol{\Theta}}_\beta^\dagger, \widehat{\boldsymbol{\Theta}}_{\star,\beta}^\dagger)$, which is used in Section 5.2. Denote $k'_{\alpha_1,\alpha_2,n} = \min\{\alpha_1^2, e^{\alpha_0}(\alpha_1 + \alpha_2 r_{\boldsymbol{Z}_\star}^{1/2}) n^{-1/2}\}$. Such an upper bound is of order $k_{\alpha_1,\alpha_2,\beta,n} h_{\alpha_1,\beta}^{-2}$, where

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

5.2. Target matrix

To study the convergence of $d^2(\widehat{\boldsymbol{A}}_\beta, \boldsymbol{A}_\star)$, we require the following conditions on the random errors $\boldsymbol{\epsilon}$ and the target matrix $\boldsymbol{A}_\star$. Recall that $\widehat{\boldsymbol{A}}_\beta$ includes the estimations obtained with the unconstrained estimator $\widehat{\boldsymbol{\Theta}}$ and those with the constrained estimator $\widehat{\boldsymbol{\Theta}}_\beta$, because $\widehat{\boldsymbol{A}}(\widehat{\boldsymbol{\Theta}}) = \widehat{\boldsymbol{A}}_{\alpha_2}$, with $\beta = \alpha_2$.

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, $\max_{i,j}\mathsf{E}|\epsilon_{ij}|^l \le (1/2)l!c_\sigma^2\eta^{l-2}$, for any positive integer $l \ge 2$.

C4. There exists a positive constant a_0 such that $\|\boldsymbol{A}_\star\|_\infty \le a_0$.

Denote $h_{(1),\beta} = \max_{i,j}(\theta_{\star,ij}^{-1}\theta_{\star,ij,\beta})$ and

$$\Delta = \max\left\{\frac{(c_\sigma \vee a)\, e^{-\mu_\star/2+\alpha_2-\beta+|\alpha_2/2-\beta|}\,(n\log n)^{1/2}}{n^2}, \frac{\eta e^{\mu_\star/2+\alpha_1+|\alpha_2/2-\beta|}k_{\alpha_1,\alpha_2,\beta,n}^{1/2}\log^{3/2} n}{h_{\alpha_1,\beta}n}\right\}. \tag{5.7}$$

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

Theorem 4. *Assume Conditions* C1−C4 *hold. For $\beta \ge 0$, there exist some positive constants C_4, C_5, C_6, and C_7, all independent of β, such that for $h_{(1),\beta}\tau \ge C_4\Delta$, we have with probability at least $1 - 3/(2n)$,*

$$d^2\left(\widehat{\boldsymbol{A}}_\beta, \boldsymbol{A}_\star\right) \le \max\left\{C_6 n^2 h_{(1),\beta}^2 r_{\boldsymbol{A}_\star}\tau^2 + C_7 h_{(1),\beta}^2 h_{(2),\beta}^2 r_{\boldsymbol{A}_\star} n^{-1}\log(n), C_5 h_{(1),\beta}h_{(3),\beta}n^{-1}\log^{1/2}(n)\right\}. \tag{5.8}$$

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 after Theorem 3, $d^2(\widehat{\boldsymbol{Z}}_{\mathsf{Win},\beta}, \mathcal{T}_\beta(\boldsymbol{Z}_\star))$ converges at a slower rate β^2, causing 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, such that $r_{\boldsymbol{M}_\star} \asymp 1$.

Uniform missingness. Under uniform missingness (i.e., $\theta_{ij} \equiv \theta_0$), Koltchinskii, Lounici and Tsybakov (2011) show 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 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 that based on the unweighted empirical risk function. Theorem 4 shows that this achieves the optimal rate $\theta_0^{-1}n^{-1}\text{polylog}(n)$. 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.

Non-uniform missingness. Under non-uniform missingness, suppose the lower and upper bounds of the observation probability satisfy $\theta_L \asymp e^{\mu_\star - \alpha_2}$ and $\theta_U \asymp e^{\mu_\star + \alpha_2}$, respectively. 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 (5.7) dominates because

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

Thus, the convergence rate of $d^2(\widehat{\boldsymbol{A}}_\beta, \boldsymbol{A}_\star)$ is $e^{\mu_\star+5\alpha_1+3\alpha_2}n^{-1/2}\log^3 n$. Because $e^{\mu_\star/2+5\alpha_1/2+3\alpha_2/2} \le e^{3\alpha_1+3\alpha_2/2}$, guaranteeing convergence 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 $\theta_L^{-1} = O(n^{1/6})$ excludes $\theta_L \equiv (n^{-1}\text{polylog}(n))$. This is the case that results in the number of observed matrix entries being of the order of $n\,\text{polylog}(n)$, which represents the most sparse case of observation where the matrix can still be recovered (Candès and Recht (2009); Candès and Plan (2010); Koltchinskii, Lounici and Tsybakov (2011); Negahban and Wainwright (2012)). We show in the following that with an appropriately chosen β, the constrained estimator $\widehat{\boldsymbol{\Theta}}_\beta$ can accommodate the case of $\theta_L^{-1} = O(n\log^{-1} n)$.

Case (I): $\beta = 0$. 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 zero provided that $\theta_U\theta_L^{-2} = o(n\log^{-1} n)$. 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 $n\,\text{polylog}(n)$ observations.

Case (II): $\beta > 0$. 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 a smaller 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,

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

because $e^{-\mu_\star/2 + \alpha_2 - \beta + |\alpha_2/2-\beta|} \leq e^{\alpha_1/2+3\alpha_2/2}$. In the following, we consider two further cases: (i) $\alpha_2 = O((\log\log n)^{-1}\alpha_1)$ and (ii) $\alpha_1 = o(\alpha_2 \log\log n)$. Note that for both cases, $e^{-\mu_\star + 2\alpha_2 - 2\beta + 2|\alpha_2/2-\beta|} \asymp \theta_U \theta_L^{-2}$, 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 polylog(n) factor compared with the most extreme, but feasible setting of $\theta_L^{-1} = O[n\{\text{polylog}(n)\}^{-1}]$. In addition, $\beta = O(e^{-\mu_\star - 2\alpha_1 + \alpha_2} n^{-1/2} \log^{-1} n)$ implies that $\beta = O(n^{-1/2} \log^{-1} n)$.

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 $\theta_L^{-1} = O(n^{1/3})$. Furthermore, $\beta = O(e^{-\mu_\star - 2\alpha_1 + \alpha_2} n^{-1/2} \log^{-1} n)$ implies that $\beta = O(n^{-1/6} \log^{-1} n)$. However, to make $d^2(\widehat{\boldsymbol{A}}_\beta, \boldsymbol{A}_\star)$ convergent, the attained rate for θ_L^{-1} has to be $O(n^{1/3})$, which excludes the most extreme heterogeneity case of $\theta_L^{-1} = O\{n(\text{polylog}(n))^{-1}\}$. The reason for not being able to cover 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). Because $\mu_\star$ is jointly estimated with $\boldsymbol{Z}_\star$ in the unconstrained estimation (Section 3.1), stronger heterogeneity slows down the convergence rate in the estimation of $\mu_\star$, which becomes a bottleneck for further improvement. If $\mu_\star$ were observable, the problem would not be as serious, despite the adverse effect of the stronger heterogeneity on the estimation of $\boldsymbol{Z}_\star$.

To summarize, under uniform missingness and Case (I) and (II)(i) under non-uniform missingness, we can achieve the optimal rate up to a polylog(n) order. For Case (II)(ii), when the missingness is not extreme, with an appropriately chosen $\beta > 0$, the proposed estimator can also attain the optimal rate up to the polylog(n) order.

5.3. Comparison with the uniform objective function

Recall that the unweighted empirical risk function $\widehat{R}_{\text{UNI}}(\boldsymbol{A}) = n^{-2}\|\boldsymbol{W} \circ (\boldsymbol{A} - \boldsymbol{Y})\|_F^2$ is adopted by many existing matrix-completion techniques (Klopp (2014)). An interesting question is whether there is any benefit to adopting the proposed weighted empirical risk function for matrix completion. In this subsection, we

examine this aspect by comparing the non-asymptotic error bounds of the corresponding estimators. Owing to the additional complication from the estimation error of the observation probability matrix, we focus only on the weighted empirical risk function with a true inverse probability weighting. We demonstrate empirically in Sections 6 and 7 the benefits of the weighted objective function with estimated weights.

Most existing works that use an unweighted empirical risk function assume the true missingness is uniform (Candès and Plan (2010); Koltchinskii, Lounici and Tsybakov (2011)). One notable exception is Klopp (2014), who studies an unweighted empirical risk function under a possibly non-uniform missing structure. The estimator of Klopp (2014) is equivalent to our estimator when $\beta = 0$, and is denoted by $\widehat{\boldsymbol{A}}^{\mathrm{UNI}}$. Thus, according to Theorem 4, we have with probability at least $1 - 3/(2n)$,

$$d^2\left(\widehat{\boldsymbol{A}}^{\mathrm{UNI}}, \boldsymbol{A}_\star\right) \le \min\left\{(C_6 + C_7) r_{\boldsymbol{A}_\star} \theta_U \theta_L^{-2} n^{-1} \log n, C_5 \theta_U^{1/2} \theta_L^{-1} n^{-1/2} \log^{1/2} n\right\} = U^{\mathrm{UNI}},$$

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

$$d^2\left(\widehat{\boldsymbol{A}}^{\mathrm{KNOWN}}, \boldsymbol{A}_\star\right) \le \min\left\{(C_6 + C_7) r_{\boldsymbol{A}_\star} \theta_L^{-1} n^{-1} \log n, C_5 \theta_L^{-1/2} n^{-1/2} \log^{1/2} n\right\} = U^{\mathrm{KNOWN}}.$$

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

Theorem 5. *Assume Conditions* C1−C4 *hold, 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,* $U^{KNOWN} \le U^{UNI}$ *and* $U^{KNOWN} < U^{UNI}$ *if* $\theta_U > \theta_L$; *that is, the true missing mechanism is non-uniform.*

Our approach draws inspiration from the missing-value literature. For instance, Chen, Leung and Qin (2008) show that using the estimated parameters in the inverse probability weighting can actually reduce the variance of the parameter of interest; see Theorem 1 of their paper. Given the results of Chen, Leung and Qin (2008), we expect that using the estimated parameters $\widehat{\boldsymbol{\Theta}}_\beta$ in the

weighting probability will not be inferior to the version with the true parameter $\widehat{\boldsymbol{\Theta}}_\star$. However, although this is verified using numerical studies, a theoretical proof cannot be achieved.

6. Simulation Study

6.1. Missingness

This section reports the results of our simulation experiments, which were designed to evaluate the numerical performance of the proposed methodologies. We first evaluate the estimation performance of the observation probabilities in Section 6.1, and then do so for the target matrix in Section 6.2. 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 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 $\boldsymbol{M}_\star = \boldsymbol{U}_{\boldsymbol{M}_\star}\boldsymbol{V}_{\boldsymbol{M}_\star}^{\mathrm{T}} - \bar{m}_{n_1,n_2,r_{\boldsymbol{M}_\star}}\boldsymbol{J}$, 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 choose $n_1 = n_2$, with four sizes: 600, 800, 1000, and 1200. The number of simulation runs for each setting is 500. For the purpose of benchmarking, we compare various missingness estimators:

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 one-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 $\widehat{\boldsymbol{\Theta}}_{\mathsf{1\text{-}bit,Win},\beta}$, respectively (note that the one-bit estimator $\widehat{\boldsymbol{\Theta}}_{\mathsf{1\text{-}bit},\alpha}$ imposes the nuclear-norm regularization on the whole $\boldsymbol{M}$ instead of $\boldsymbol{Z}$, in contrast to $\widehat{\boldsymbol{\Theta}}_\alpha$);
5. the rank-one probability estimator $\widehat{\boldsymbol{\Theta}}_{\mathsf{NW}}$ used in Negahban and Wainwright (2012), where $g_{i\cdot} = n_2^{-1}\sum_{j=1}^{n_2} w_{ij}$, $g_{\cdot j} = n_1^{-1}\sum_{i=1}^{n_1} w_{ij}$, and $\theta_{ij,\mathsf{NW}} = g_{i\cdot}g_{\cdot j}$;
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 one-bit estimator $\widehat{\boldsymbol{\Theta}}_{\mathsf{1\text{-}bit},\alpha}$, the parameter α is set based on our knowledge of the true $\boldsymbol{M}_\star$. For the constrained

Table 1. The root mean squared errors 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 $\widehat{\boldsymbol{\Theta}}_\alpha$, $\widehat{\boldsymbol{\Theta}}_{\mathsf{1\text{-}bit},\alpha}$ and the two existing estimators $\widehat{\boldsymbol{\Theta}}_{\mathsf{NW}}$ and $\widehat{\boldsymbol{\Theta}}_{\mathsf{UNI}}$.

600	$\widehat{\boldsymbol{\Theta}}_\alpha$	$\widehat{\boldsymbol{\Theta}}_{\mathsf{1\text{-}bit},\alpha}$	$\widehat{\boldsymbol{\Theta}}_{\mathsf{NW}}$	$\widehat{\boldsymbol{\Theta}}_{\mathsf{UNI}}$
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	$\widehat{\boldsymbol{\Theta}}_\alpha$	$\widehat{\boldsymbol{\Theta}}_{\mathsf{1\text{-}bit},\alpha}$	$\widehat{\boldsymbol{\Theta}}_{\mathsf{NW}}$	$\widehat{\boldsymbol{\Theta}}_{\mathsf{UNI}}$
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)	-	-
1,000	$\widehat{\boldsymbol{\Theta}}_\alpha$	$\widehat{\boldsymbol{\Theta}}_{\mathsf{1\text{-}bit},\alpha}$	$\widehat{\boldsymbol{\Theta}}_{\mathsf{NW}}$	$\widehat{\boldsymbol{\Theta}}_{\mathsf{UNI}}$
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)	-	-
$r_{\widehat{\boldsymbol{\Theta}}}$	1,000.00 (0.00)	1,000.00 (0.00)	-	-
1,200	$\widehat{\boldsymbol{\Theta}}_\alpha$	$\widehat{\boldsymbol{\Theta}}_{\mathsf{1\text{-}bit},\alpha}$	$\widehat{\boldsymbol{\Theta}}_{\mathsf{NW}}$	$\widehat{\boldsymbol{\Theta}}_{\mathsf{UNI}}$
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)	-	-
$r_{\widehat{\boldsymbol{\Theta}}}$	1,200.00 (0.00)	1,200.00 (0.00)	-	-

estimators $\widehat{\boldsymbol{\Theta}}_\beta$ and $\widehat{\boldsymbol{\Theta}}_{\mathsf{Win},\beta}$, the constraint level β is chosen so that either 5% or 10% of the elements in $\widehat{\boldsymbol{Z}}_\alpha$ are winsorized. The parameters for $\widehat{\boldsymbol{\Theta}}_{\mathsf{1\text{-}bit},\beta}$ and $\widehat{\boldsymbol{\Theta}}_{\mathsf{1\text{-}bit,Win},\beta}$ are set in a similar manner.

To quantify the estimation performance of the linear predictor $\boldsymbol{M}_\star$ and the observation probabilities $\boldsymbol{\Theta}_\star$, we consider 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:

$$\mathrm{RMSE}\left(\boldsymbol{B}, \boldsymbol{C}\right) = \frac{\|\boldsymbol{B} - \boldsymbol{C}\|_F}{(n_1 n_2)^{1/2}} \quad \text{and} \quad d_H^2\left(\widehat{\boldsymbol{\Theta}}, \boldsymbol{\Theta}_\star\right) = \frac{\sum_{i,j=1}^{n_1,n_2} d_H^2\left(\widehat{\theta}_{ij}, \theta_{\star,ij}\right)}{(n_1 n_2)^{1/2}}.$$

Because the estimators $\mathcal{F}^{-1}(\widehat{\boldsymbol{\Theta}}_\alpha)$ and $\mathcal{F}^{-1}(\widehat{\boldsymbol{\Theta}}_{\mathsf{1\text{-}bit},\alpha})$ are both of low rank, we also report their corresponding ranks.

Table 1 summarizes the simulation results. The most visible aspect of the results is that the proposed estimators $\widehat{\boldsymbol{\Theta}}_{\alpha}$ and $\widehat{\boldsymbol{\Theta}}_{\mathsf{1\text{-}bit},\alpha}$ both outperform the two existing estimators $\widehat{\boldsymbol{\Theta}}_{\mathsf{NW}}$ and $\widehat{\boldsymbol{\Theta}}_{\mathsf{UNI}}$. As such, the former have smaller root mean square errors with respect to $\widehat{\boldsymbol{M}}$, smaller Hellinger distances $d_H^2(\widehat{\boldsymbol{\Theta}}, \boldsymbol{\Theta}_\star)$, and more accurate estimated ranks of $\boldsymbol{M}_\star$. Without the separation of $\mu_\star$ from $\boldsymbol{M}_\star$, $\widehat{\boldsymbol{\Theta}}_{\mathsf{1\text{-}bit},\alpha}$ has a larger error and Hellinger distance than those of the proposed estimators. The performance of $\widehat{\boldsymbol{\Theta}}_{\mathsf{NW}}$ is roughly between that of the proposed estimators and that of the uniform estimator $\widehat{\boldsymbol{\Theta}}_{\mathsf{UNI}}$. The estimator $\widehat{\boldsymbol{\Theta}}_{\mathsf{UNI}}$ is a benchmark that 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 $\mathcal{N}(0, \sigma^2_{\boldsymbol{A}_\star})$, obtaining $\boldsymbol{A}_\star = 2.5\boldsymbol{J} + \boldsymbol{U}_{\boldsymbol{A}_\star}\boldsymbol{V}_{\boldsymbol{A}_\star}^{\mathrm{T}}$. Here, we set the standard deviation of the entries in the matrix product $\boldsymbol{U}_{\boldsymbol{A}_\star}\boldsymbol{V}_{\boldsymbol{A}_\star}^{\mathrm{T}}$ to 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 independent and identically distributed mean-zero Gaussian entries $\epsilon_{ij} \sim \mathcal{N}(0, \sigma_\epsilon^2)$. Here, σ_ϵ^2 is chosen such that $\mathrm{SNR} = (\mathsf{E}\|\boldsymbol{A}_\star\|_F^2/\mathsf{E}\|\boldsymbol{\epsilon}\|_F^2)^{1/2} = 1$, where $\mathsf{E}\|\boldsymbol{A}_\star\|_F^2 = n_1 n_2(r_{\boldsymbol{A}_\star} - 1 + 2.5^2)$ implies $\sigma_\epsilon = 0.5(r_{\boldsymbol{A}_\star} - 1 + 2.5^2)^{1/2}$.

For the estimation of the target matrix, we evaluated 10 versions of the proposed estimators $\mathsf{Prop}\_\widehat{\boldsymbol{\Theta}}_\beta\_t$, $\mathsf{Prop}\_\widehat{\boldsymbol{\Theta}}_{\mathsf{Win},\beta}\_t$, $\mathsf{Prop}\_\widehat{\boldsymbol{\Theta}}_\alpha$, $\mathsf{Prop}\_\widehat{\boldsymbol{\Theta}}_{\mathsf{1\text{-}bit},\beta}\_t$, $\mathsf{Prop}\_\widehat{\boldsymbol{\Theta}}_{\mathsf{1\text{-}bit,Win},\beta}\_t$, and $\mathsf{Prop}\_\widehat{\boldsymbol{\Theta}}_{\mathsf{1\text{-}bit},\alpha}$. Here, Prop indicates the estimators are obtained by solving problem (4.3), $\widehat{\boldsymbol{\Theta}}_\beta$, $\widehat{\boldsymbol{\Theta}}_{\mathsf{Win},\beta}$, $\widehat{\boldsymbol{\Theta}}_\alpha$, $\widehat{\boldsymbol{\Theta}}_{\mathsf{1\text{-}bit},\beta}$, $\widehat{\boldsymbol{\Theta}}_{\mathsf{1\text{-}bit,Win},\beta}$, and $\widehat{\boldsymbol{\Theta}}_{\mathsf{1\text{-}bit},\alpha}$ represent the probability estimators used in (4.3), as described in Section 6.1, and $t = 0.05$ or 0.1 denotes the winsorized proportion for which β is chosen. In addition, as in Mao, Chen and Wong (2019), we compare these with three existing matrix-completion techniques: the methods proposed in Negahban and Wainwright (2012) (NW), Koltchinskii, Lounici and Tsybakov (2011) (KLT), and Mazumder, Hastie and Tibshirani (2010) (MHT). Of 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, Chen and Wong (2019) 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 use one more measure: $\text{Test Error} = \|\boldsymbol{W}^\star \circ (\widehat{\boldsymbol{A}}_\beta - \boldsymbol{A}_\star)\|_F^2/\|\boldsymbol{W}^\star \circ \boldsymbol{A}_\star\|_F^2$, where $\boldsymbol{W}^\star$ is the matrix of the missing indicator, with the (i,j)th entry being $(1 - w_{ij})$. The test error

measures the estimation error of the unobserved entries relative to their signal strength. The estimated ranks of $\widehat{\boldsymbol{A}}_\beta$ are also reported.

Table 2 summarizes the simulation results for dimensions n_1=n_2 ranging from 600 to 800 and two different settings of $r_{\boldsymbol{A}_\star} = 11$. The results for $r_{\boldsymbol{A}_\star} = 11$ for dimensions n_1=n_2 ranging from 1,000 to 1,200 are relegated to Table S1, and the results for $r_{\boldsymbol{A}_\star} = 31$ are relegated to Tables S2−S3 of Section S1.5 in the Supplementary Material. The tables show that the 10 versions of the proposed methods outperform the three existing methods by having smaller root mean squared errors and Test Error values. Among the first five proposed methods, $\mathsf{Prop\_}\widehat{\boldsymbol{\Theta}}_\beta$ is better than $\mathsf{Prop\_}\widehat{\boldsymbol{\Theta}}_\alpha$ in most cases. This is because the constrained estimator $\widehat{\boldsymbol{\Theta}}_\beta$ has a much smaller ratio $\widehat{\theta}_U/\widehat{\theta}_L$ than $\widehat{\boldsymbol{\Theta}}_\alpha$, which improves the stability of the prediction and the accuracy. Furthermore, $\mathsf{Prop\_}\widehat{\boldsymbol{\Theta}}_\beta\mathsf{\_0.1}$ performs better than $\mathsf{Prop\_}\widehat{\boldsymbol{\Theta}}_{\mathsf{1\text{-}bit},\alpha}$ in most cases.

7. Real-Data Application

In this section, we demonstrate the proposed methodology by analyzing the Yahoo! Webscope data set (ydata-ymusic-user-artist-ratings-v1_0), available at `https://webscope.sandbox.yahoo.com/`. It contains (incomplete) ratings from 15,400 users on 1,000 songs. The data set 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 of whom rated exactly 10 songs that were not rated in the training set. The missing rates are 0.9763 overall, ranging from 0.3520 to 0.9900 across users, and from 0.6372 to 0.9957 across songs. The non-uniformity of the missingness is shown in Figure S1 of Section S1.6 in the Supplementary Material. In this experiment, we applied the methods described in Section 6 to the training set, and evaluated the test errors based on the corresponding test set. Because there is no prior knowledge about the true parameters α_1 and α_2, we suggest choosing α_1 and α_2 sufficiently large, say $\alpha_1 = 100$ and $\alpha_2 = 100$, to ensure that the range covers all missing probabilities. Noted that $\widehat{\boldsymbol{\Theta}}_\alpha$ is not sensitive to a larger α.

Table 3 reports the root mean squared prediction errors, where RMSPE = $\|\boldsymbol{W}^{test} \circ (\widehat{\boldsymbol{A}}_\beta - \boldsymbol{Y})\|_F/(\sum_{i=1}^{n_1}\sum_{j=1}^{n_2} w_{ij}^{test})^{1/2}$, and $\boldsymbol{W}^{test}$ is the indicator matrix of the test set with an (i,j)th entry of w_{ij}^{test}. Note that $\mathsf{Prop\_}\widehat{\boldsymbol{\Theta}}_\beta\mathsf{\_0.05}$ performs the best of the 10 versions of the proposed methods. In addition, $\mathsf{Prop\_}\widehat{\boldsymbol{\Theta}}_\alpha$ has a much smaller root mean squared prediction error than those of the other eight versions of the proposed methods. This may indicate that only a slight constraint

Table 2. Root mean squared errors, 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 10 versions of the proposed methods, where Prop indicates the estimators are obtained by solving problem (4.3), while $\widehat{\boldsymbol{\Theta}}_\beta$, $\widehat{\boldsymbol{\Theta}}_{\mathsf{Win},\beta}$, $\widehat{\boldsymbol{\Theta}}_\alpha$, $\widehat{\boldsymbol{\Theta}}_{\text{1-bit},\beta}$,$\widehat{\boldsymbol{\Theta}}_{\text{1-bit},\mathsf{Win},\beta}$, and $\widehat{\boldsymbol{\Theta}}_{\text{1-bit},\alpha}$ represent the probability estimators used in (4.3), as described in Section 6.1, and $t = 0.05$ or 0.1 denotes the winsorized proportion for which β is chosen.

$(n_1, n_2) = (600, 600)$	RMSE$(\widehat{\boldsymbol{A}}_\beta, \boldsymbol{A}_\star)$	Test Error	$r_{\widehat{\boldsymbol{A}}_\beta}$
Prop_$\widehat{\boldsymbol{\Theta}}_{\mathsf{Win},\beta}$_0.05	1.5615 (0.0147)	0.3005 (0.0062)	65.28 (5.72)
Prop_$\widehat{\boldsymbol{\Theta}}_{\beta}$_0.05	1.5548 (0.0085)	0.2996 (0.0034)	54.98 (3.01)
Prop_$\widehat{\boldsymbol{\Theta}}_{\mathsf{Win},\beta}$_0.1	1.5621 (0.0111)	0.3013 (0.0046)	63.68 (5.36)
Prop_$\widehat{\boldsymbol{\Theta}}_{\beta}$_0.1	1.5509 (0.0085)	0.2983 (0.0034)	53.13 (2.72)
Prop_$\widehat{\boldsymbol{\Theta}}_{\alpha}$	1.5637 (0.0147)	0.3010 (0.0061)	65.63 (5.89)
Prop_$\widehat{\boldsymbol{\Theta}}_{\text{1-bit},\mathsf{Win},\beta}$_0.05	1.5664 (0.0093)	0.3028 (0.0037)	62.76 (5.96)
Prop_$\widehat{\boldsymbol{\Theta}}_{\text{1-bit},\beta}$_0.05	1.5573 (0.0089)	0.2996 (0.0036)	61.80 (5.34)
Prop_$\widehat{\boldsymbol{\Theta}}_{\text{1-bit},\mathsf{Win},\beta}$_0.1	1.5669 (0.0092)	0.3032 (0.0037)	62.78 (2.68)
Prop_$\widehat{\boldsymbol{\Theta}}_{\text{1-bit},\beta}$_0.1	1.5540 (0.0089)	0.2987 (0.0036)	60.79 (3.01)
Prop_$\widehat{\boldsymbol{\Theta}}_{\text{1-bit},\alpha}$	1.5612 (0.0097)	0.3005 (0.0040)	62.12 (4.76)
NW	1.9896 (0.2814)	0.4676 (0.1341)	167.67 (54.78)
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	$r_{\widehat{\boldsymbol{A}}_\beta}$
Prop_$\widehat{\boldsymbol{\Theta}}_{\mathsf{Win},\beta}$_0.05	1.4754 (0.0107)	0.2669 (0.0041)	88.58 (10.81)
Prop_$\widehat{\boldsymbol{\Theta}}_{\beta}$_0.05	1.4797 (0.0080)	0.2714 (0.0030)	71.79 (4.12)
Prop_$\widehat{\boldsymbol{\Theta}}_{\mathsf{Win},\beta}$_0.1	1.4724 (0.0108)	0.2664 (0.0042)	86.25 (10.34)
Prop_$\widehat{\boldsymbol{\Theta}}_{\beta}$_0.1	1.4763 (0.0082)	0.2704 (0.0031)	67.08 (4.22)
Prop_$\widehat{\boldsymbol{\Theta}}_{\alpha}$	1.4783 (0.0115)	0.2676 (0.0041)	88.92 (11.70)
Prop_$\widehat{\boldsymbol{\Theta}}_{\text{1-bit},\mathsf{Win},\beta}$_0.05	1.4917 (0.0078)	0.2743 (0.0030)	83.51 (1.45)
Prop_$\widehat{\boldsymbol{\Theta}}_{\text{1-bit},\beta}$_0.05	1.4804 (0.0080)	0.2705 (0.0031)	82.60 (3.47)
Prop_$\widehat{\boldsymbol{\Theta}}_{\text{1-bit},\mathsf{Win},\beta}$_0.1	1.4972 (0.0080)	0.2765 (0.0031)	81.64 (7.23)
Prop_$\widehat{\boldsymbol{\Theta}}_{\text{1-bit},\beta}$_0.1	1.4800 (0.0078)	0.2708 (0.0030)	74.89 (3.54)
Prop_$\widehat{\boldsymbol{\Theta}}_{\text{1-bit},\alpha}$	1.4790 (0.0099)	0.2685 (0.0039)	88.57 (9.56)
NW	1.9515 (0.3625)	0.4585 (0.1593)	215.61 (82.24)
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)

With $r_{\boldsymbol{M}_\star} = 11$, $r_{\boldsymbol{A}_\star} = 11$, $(n_1, n_2) = (600, 600)$, $(800, 800)$, and SNR $= 1$. The three existing methods are proposed, respectively, in Negahban and Wainwright (2012)(NW), Koltchinskii, Lounici and Tsybakov (2011)(KLT), and Mazumder, Hastie and Tibshirani (2010)(MHT).

Table 3. Root mean squared prediction errors based on Yahoo! Webscope data set for the 10 versions of the proposed method and the three existing methods proposed, respectively, in Negahban and Wainwright (2012)(NW), Koltchinskii, Lounici and Tsybakov (2011)(KLT), and Mazumder, Hastie and Tibshirani (2010)(MHT).

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

is required for the probability estimator for this data set. Note that we cannot guarantee the optimal convergence rate or even asymptotic convergence in certain settings of missingness for $\mathsf{Prop\_}\widehat{\boldsymbol{\Theta}}_{\alpha}$; see Section 5.2 for details.

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

Compared with the existing methods NW, KLT, and MHT, our proposed methods perform significantly better in terms of the root mean squared prediction errors, achieving as much as a 25% improvement over the method of Mazumder, Hastie, and Tibshirani (the best of the three existing methods). This suggests that a more flexible model of a missing structure improves the model's prediction power.

8. Conclusion

When matrix entries are heterogeneously observed owing to a selection bias, this heterogeneity needs to be taken into account. This study focuses on the problem of matrix completion under a low-rank missing structure. To recover the observation probabilities, we adopted a generalized linear model with a low-rank linear predictor matrix. To avoid unnecessary bias, we introduced a separation of the mean effect μ. Because extreme values of the probabilities may lead to an unstable estimation of the target matrix, we propose an inverse probability weighting-based method with constrained probability estimates, and demonstrate its improvements empirically. Our theoretical result shows that the estimator

of the high-dimensional probability matrix can be embedded into the inverse probability weighting 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 compare the proposed methods with existing matrix-completion methods, demonstrating their appealing performance.

Supplementary Material

The online Supplementary Material contains useful lemmas, the proofs of the main theorems and some additional numerical studies.

Acknowledgments

The authors thank the editors and two reviewers for their constructive comments and suggestions. Xiaojun Mao's research is supported by NSFC Grants 12001109, 92046021, Shanghai Sailing Program 19YF1402800, and the Science and Technology Commission of Shanghai Municipality grant 20dz1200600. Raymond K.W. Wong's research was partially supported by the National Science Foundation under Grants DMS-1806063, DMS-1711952 (subcontract), and CCF-1934904. Song Xi Chen acknowledges support from the National Natural Science Foundation of China grants 92046021, 12071013 and 12026607, and LMEQF at Peking University.

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.

Chen, S. X., Leung, D. H. and Qin, J. (2008). Improving semiparametric estimation by using surrogate data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **70**, 803–823.

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

Foygel, R., Shamir, O., Srebro, N. and Salakhutdinov, R. R. (2011). Learning with the weighted trace-norm under arbitrary sampling distributions. In *Proceedings of the 24th International Conference on Neural Information Processing Systems*, 2133–2141.

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. (2010). Matrix completion from noisy entries. *The Journal of Machine Learning Research* **11**, 1532–4435.

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. (2019). Matrix completion with covariate information. *Journal of the American Statistical Association* **114**, 198–210.

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*. American Statistical Association Washington, DC.

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. In *Proceedings of the 33rd International Conference on Machine Learning* **48**, 1670–1679. PMLR, New York, USA.

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 **23**, 2056–2064.

Xiaojun Mao
School of Data Science, Fudan University, Shanghai 200433, China.
E-mail: maoxj@fudan.edu.cn
Raymond K. W. Wong
Department of Statistics, Texas A&M University, College Station, Texas 77843, USA.
E-mail: raywong@stat.tamu.edu
Song Xi Chen
Department of Business Statistics and Econometrics, Peking University, Beijing 100651, China.
E-mail: csx@gsm.pku.edu.cn

(Received May 2019; accepted March 2020)