Matrix Completion with Covariate Information

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

September 24, 2017

Abstract

This paper investigates the problem of matrix completion from corrupted data, when additional covariates are available. Despite being seldomly considered in the matrix completion literature, these covariates often provide valuable information for completing the unobserved entries of the high-dimensional target matrix $\boldsymbol{A}_0$. Given a covariate matrix $\boldsymbol{X}$ with its rows representing the row covariates of $\boldsymbol{A}_0$, we consider a column-space-decomposition model $\boldsymbol{A}_0 = \boldsymbol{X}\boldsymbol{\beta}_0 + \boldsymbol{B}_0$ where $\boldsymbol{\beta}_0$ is a coefficient matrix and $\boldsymbol{B}_0$ is a low-rank matrix orthogonal to $\boldsymbol{X}$ in terms of column space. This model facilitates a clear separation between the interpretable covariate effects ($\boldsymbol{X}\boldsymbol{\beta}_0$) and the flexible hidden factor effects ($\boldsymbol{B}_0$). Besides, our work allows the probabilities of observation to depend on the covariate matrix, and hence a missing-at-random mechanism is permitted. We propose a novel penalized estimator for $\boldsymbol{A}_0$ by utilizing both Frobenius-norm and nuclear-norm regularizations with an efficient and scalable algorithm. Asymptotic convergence rates of the proposed estimators are studied. The empirical performance of the proposed methodology is illustrated via both numerical experiments and a real data application.

Keywords: High-dimensional statistics; Low-rank estimation; Missing data; Nuclear-norm regularization.

*Xiaojun Mao is Ph.D. candidate, Department of Statistics, Iowa State University, Ames, IA 50011, USA (Email: mxjki@iastate.edu).

†Author of Correspondence. 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). His research is partially supported by Chinas National Key Research Special Program Grants 2016YFC0207701 and 2015CB856000, and National Natural Science Foundation of China grants 11131002, 71532001 and 71371016.

‡Raymond K. W. Wong is Assistant Professor, Department of Statistics, Texas A&M University, College Station, TX 77843, USA (Email: raywong@stat.tamu.edu). His research is partially supported by the National Science Foundation under Grants DMS-1612985 and DMS-1711952 (subcontract).

1 Introduction

In recent years the problem of recovering a low-rank data matrix from relatively few observed entries has drawn significant amount of attention. This problem arises from a variety of applications including collaborative filtering, computer visions and positioning. In these applications, the low-rank assumption is often used to reflect the belief that rows (or columns) are generated from a relatively few number of hidden factors. For instance, in the Netflix prize problem (Feuerverger et al., 2012), viewers' ratings are assumed to be adequately modeled by a few hidden profiles.

In the noiseless setting, earlier works (Candès and Recht, 2009; Recht, 2011) have established strong theoretical guarantees on perfect matrix recovery. A typical form of this remarkable result is stated as follows. An n_1-by-n_2 matrix $\boldsymbol{A}_0$ of rank $r_{\boldsymbol{A}_0}$, fulfilling certain incoherence conditions, can be recovered exactly with high probability from $c(n_1 + n_2)r_{\boldsymbol{A}_0} \log^2(n_1 + n_2)$ observed entries sampled uniformly at random via a convex and tractable constrained nuclear norm minimization for a positive constant c. As for the noisy setting where observed entries are corrupted by noise, extensive works on matrix completion (Candès and Plan, 2010; Koltchinskii et al., 2011; Rohde and Tsybakov, 2011) can be found under various forms of noise assumptions.

Some applications come with covariate information in the form of additional row and/or column information. For instance, the MovieLens 100K data set (Harper and Konstan, 2016) has both viewer demographics (age, gender, occupation and zip code) and movie features (release date and genre). These row and column covariates play similar roles as covariates in regression analysis and therefore can potentially lead to significant improvements in matrix recovery. Recent works (Abernethy et al., 2009; Natarajan and Dhillon, 2014) have shown such promises. In the noiseless setting, theoretical guarantees of perfect matrix recovery with covariates are available (Xu et al., 2013; Chiang et al., 2015). Yet, there have been limited attempts with theoretical results at the more realistic setting where observed entries are corrupted by noise. One notable study is the work by Zhu et al. (2016), which study a partial latent model for personalized prediction and its likelihood estimation.

Moreover, the probabilities of observation may vary with respect to the row and/or column attributes. As suggested by our real data analysis of the MovieLens data (Section 7), the sampling mechanism of the ratings varies across different viewer groups. The earlier literature of matrix completion (Candès and Recht, 2009; Abernethy et al., 2009; Keshavan et al., 2009; Recht, 2011; Rohde and Tsybakov, 2011; Koltchinskii et al., 2011) focused on uniform sampling mechanism, where each entry has the same marginal probability of being sampled. There are recent studies (Srebro and Salakhutdinov, 2010; Negahban and Wainwright, 2012; Klopp, 2014; Cai and Zhou, 2016; Cai et al., 2016; Bi et al., 2016) devoted to relaxing such restrictive assumption to the nonuniform case, where probabilities of observation are allowed to be different across rows and columns to some extent. However, the covariates are not taken into account in the modeling of the probabilities of observation. Driven by the aforementioned empirical observation, we model probabilities of observation with a missing-at-random (MAR) mechanism, where the probability of observation is independent of the matrix entry when conditional on the covariates.

In this paper we utilize the covariate information in both modelings of the observation probability and the completion of the target matrix. We focus on the use of only row (or equivalently column) covariates and leave the joint usage of both row and column covariates as a future work. More specifically, we consider a column-space-decomposition model of a target matrix $\boldsymbol{A}_0 \in \mathbb{R}^{n_1 \times n_2}$:

$$\boldsymbol{A}_0 = \boldsymbol{X}\boldsymbol{\beta}_0 + \boldsymbol{B}_0,$$

where $\boldsymbol{X} \in \mathbb{R}^{n_1 \times m}$ is a covariate matrix with its rows representing the row covariates of $\boldsymbol{A}_0$, $\boldsymbol{\beta}_0 \in \mathbb{R}^{m \times n_2}$ is a coefficient matrix, and $\boldsymbol{B}_0 \in \mathbb{R}^{n_1 \times n_2}$ is a low-rank matrix. To ensure identification, the column spaces of $\boldsymbol{X}$ and $\boldsymbol{B}_0$ are orthogonal. The above model shares some similarities with a recent work by Zhu, Shen, and Ye (2016), but differs in the aspect that they did not impose the orthogonality condition.

The purpose of considering covariate information is to improve the accuracy of the completion of $\boldsymbol{A}_0$ and $\boldsymbol{B}_0$. It is achieved by estimating $\boldsymbol{\beta}_0$ and $\boldsymbol{B}_0$ via minimizing a regularized empirical risk which allows separation with respect to $\boldsymbol{\beta}$ and $\boldsymbol{B}$. This means that the proposed estimators

$\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{B}}$ can be computed separately by two separate minimizations, which is scalable and non-iterative. Specifically, unlike many matrix completion algorithms that involve multiple singular value decompositions (SVD), our computation requires only one single SVD. This SVD can be re-used in computations of the proposed estimators with respect to different tuning parameters, which leads to significant computation reduction in tuning parameter selection. In addition, our algorithm can be coupled with the fast randomized singular value thresholding (FRSVT) procedure (Oh et al., 2015) for efficient computation in large matrix completion problems.

As for theoretical properties, we first provide a general asymptotic upper bounds for the mean squared error (MSE) achieved by the completed matrices under a general missing mechanism, followed by specific results for uniform missing and MAR satisfying the logistic regression. To demonstrate the benefits of including the covariate information, we show a faster convergence of the covariate part $\boldsymbol{X}\hat{\boldsymbol{\beta}}$ than the low-rank part $\hat{\boldsymbol{B}}$. In addition, we provide a non-asymptotic upper bound for the mean squared error (MSE) of the completed matrix $\hat{\boldsymbol{B}}$ and show it is no larger than the one by Koltchinskii et al. (2011) under the uniform missingness. Besides, the proposed matrix completion is shown to attain the minimax optimal rate (up to a logarithmic factor) in the estimation of both the entire matrix and its lower rank part $\boldsymbol{B}$ under the uniform missingness. Additional results for non-uniform missingness are also provided.

The rest of the paper is organized as follows. The proposed model is constructed in Section 2. The associated estimation, computation and tuning parameter selection are all developed in Section 3 while the asymptotic convergence rates are given in Section 4. In Section 5, we discuss the benefit of the covariate information with a set of theoretical results. Numerical performances of the proposed method are illustrated in a simulation study in Section 6 and an application to a MovieLens dataset in Section 7. Concluding remarks are given in Section 8, while all technical details are delegated to a supplementary material.

2 Proposed model

Let $\boldsymbol{A}_0 = (A_{0,ij}) \in \mathbb{R}^{n_1 \times n_2}$ be an unknown high dimensional matrix of interest, and $\boldsymbol{Y} = (Y_{ij})$ be a contaminated version of $\boldsymbol{A}_0$ where only a portion of $\{Y_{ij}\}$ is observed. For the (i, j)-th entry, consider the sampling indicator $\omega_{ij} = 1$ if Y_{ij} is observed, and 0 otherwise. The contamination follows the model:

$$Y_{ij} = A_{0,ij} + \epsilon_{ij}, \quad \text{for } i = 1, \ldots, n_1;\ j = 1, \ldots, n_2, \tag{2.1}$$

where $\{\epsilon_{ij}\}$ are independently distributed random errors with zero mean and finite variance. We assume that $\{\epsilon_{ij}\}$ are independent of $\{\omega_{ij}\}$.

In additional to the incomplete matrix $\boldsymbol{Y}$, we have an accompanying covariate matrix $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{n_1})^\intercal \in \mathbb{R}^{n_1 \times m}$, where $\boldsymbol{x}_i \in \mathbb{R}^{m \times 1}$ for $i = 1, \ldots, n_1$. Each row of $\boldsymbol{X}$, namely $\boldsymbol{x}_i^\intercal$, records m covariates associated with the corresponding row of $\boldsymbol{A}_0$. We assume that $\boldsymbol{A}_0$ is nonrandom given the covariates $\boldsymbol{X}$. For notational simplicity, $\boldsymbol{X}$ is assumed to be nonrandom. Compared with common settings of matrix completions, our setting has an additional covariate matrix $\boldsymbol{X}$, which is treated as an additional piece of information for the recovery of $\boldsymbol{A}_0$.

Regarding the sampling (or missingness) mechanism, we adopt the Bernoulli model $\omega_{ij} \sim \text{Bernoulli}(\theta_{ij}(\boldsymbol{x}_i))$ where the observation probabilities may depend on the covariate. For notational simplification, we denote $\theta_{ij} = \theta_{ij}(\boldsymbol{x}_i)$ in the rest of the paper. The detailed assumptions of $\{\epsilon_{ij}\}$ and $\{\theta_{ij}\}$ are specified in Conditions C1 and C4 in Section 4.

Prior to the discussion of our model, we briefly present two existing models of $\boldsymbol{A}_0$. The first one is a low-rank model of $\boldsymbol{A}_0$ which assumes each row (or column) of $\boldsymbol{A}_0$ is a linear combination of a small number of hidden factors. This assumption stems from the classical factor model. The second one assumes $\boldsymbol{A}_0$ is modeled as $\boldsymbol{X}\boldsymbol{\beta}_0$ with a coefficient matrix $\boldsymbol{\beta}_0 \in \mathbb{R}^{m \times n_2}$, where the problem of recovering $\boldsymbol{A}_0$ can be treated as a classical multivariate regression (Mardia et al., 1980; Freedman, 2009) (with missingness). This linear modeling affords easy interpretation of the covariate effect.

Our model is a combination of these two models, aiming to incorporate the covariate effect as well as to allow the hidden factor effect for accurate estimation of $\boldsymbol{A}_0$. To allow separation of these

two effects, we project $\boldsymbol{A}_0$ to the column space of $\boldsymbol{X}$ and its orthogonal complement such that $\boldsymbol{A}_0 = \boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{A}_0 + \boldsymbol{P}_{\boldsymbol{X}}^{\perp}\boldsymbol{A}_0$, where $\boldsymbol{P}_{\boldsymbol{X}} = \boldsymbol{X}(\boldsymbol{X}^\intercal\boldsymbol{X})^{-1}\boldsymbol{X}^\intercal$ and $\boldsymbol{P}_{\boldsymbol{X}}^{\perp} = \boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{X}}$.

By assuming that $\boldsymbol{B}_0 = \boldsymbol{P}_{\boldsymbol{X}}^{\perp}\boldsymbol{A}_0$ is of low rank, and $\boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{A}_0$ is linear in $\boldsymbol{X}$ such that $\boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{A}_0 = \boldsymbol{X}\boldsymbol{\beta}_0$, we have a specification of $\boldsymbol{A}_0$ in (2.1):

$$\boldsymbol{A}_0 = \boldsymbol{X}\boldsymbol{\beta}_0 + \boldsymbol{B}_0, \tag{2.2}$$

The low-rank assumption of $\boldsymbol{B}_0$ implies that $\boldsymbol{B}_0 = \boldsymbol{U}_0\boldsymbol{V}_0^\intercal$ where $\boldsymbol{U}_0 \in \mathbb{R}^{n_1 \times r_{\boldsymbol{B}_0}}$, $\boldsymbol{V}_0 \in \mathbb{R}^{n_2 \times r_{\boldsymbol{B}_0}}$ and $r_{\boldsymbol{B}_0}$ is the rank of $\boldsymbol{B}_0$ with $r_{\boldsymbol{B}_0} \ll \min\{n_1, n_2\}$.

Let $\tilde{\boldsymbol{U}}_0 = (\boldsymbol{X}, \boldsymbol{U}_0)$ and $\tilde{\boldsymbol{V}}_0 = (\boldsymbol{\beta}_0^\intercal, \boldsymbol{V}_0)$, then $\boldsymbol{A}_0 = \tilde{\boldsymbol{U}}_0\tilde{\boldsymbol{V}}_0^\intercal$. When compared with the typical matrix completion, model (2.2) has part of the column space of $\boldsymbol{A}_0$ being known due to $\boldsymbol{X}$. The coefficient matrix $\boldsymbol{\beta}_0$ signifies the strengths of the m covariate effects with respect to the n_2 columns of $\boldsymbol{A}_0$ and permits more interpretability in addition to the completion of $\boldsymbol{A}_0$. The goal of this paper is to recover the matrix $\boldsymbol{A}_0 = \boldsymbol{X}\boldsymbol{\beta}_0 + \boldsymbol{B}_0$, together with the coefficient matrix $\boldsymbol{\beta}_0$ and the low-rank matrix $\boldsymbol{B}_0$, in the presence of observation noise.

Our model shares some similarities with a recent work by Zhu, Shen, and Ye (2016), which allows the joint usage of row and column covariates. When only row covariates are used, the authors studied a model similar to (2.2) under the restriction that $\boldsymbol{\beta}_0 = (\boldsymbol{\alpha}, \ldots, \boldsymbol{\alpha})$, where $\boldsymbol{\alpha} \in \mathbb{R}^m$.

3 Estimation

3.1 Estimation of $\boldsymbol{\beta}_0$ and $\boldsymbol{B}_0$

We develop the estimators of $\boldsymbol{\beta}_0$ and $\boldsymbol{B}_0$ based on the framework of regularized empirical risk minimization. Define $\mathcal{C}(\boldsymbol{X})$ be the column space of a matrix $\boldsymbol{X}$, $\mathcal{N}(\boldsymbol{X}) = \{\boldsymbol{B} \in \mathbb{R}^{n_1 \times n_2} : \mathcal{C}(\boldsymbol{B}) \perp \mathcal{C}(\boldsymbol{X})\}$, $\boldsymbol{W} = (\omega_{ij})$ and $\boldsymbol{\Theta}^* = (\theta_{ij}^{-1})$. For any $\boldsymbol{\beta} \in \mathbb{R}^{m \times n_2}$ and $\boldsymbol{B} \in \mathcal{N}(\boldsymbol{X})$, we consider a population risk function

$$R(\boldsymbol{\beta}, \boldsymbol{B}) = \frac{1}{n_1 n_2}\mathsf{E}\left(\|\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{B} - \boldsymbol{W} \circ \boldsymbol{\Theta}^* \circ \boldsymbol{Y}\|_F^2\right),$$

where $\circ$ is the Hadamard product and $\|\cdot\|_F$ stands for the Frobenius norm. Our interest of this risk function originates from the following result established in Section S1 of the supplementary material.

Proposition 1. *Suppose that $\boldsymbol{X}^\intercal\boldsymbol{X}$ is invertible. Under Conditions C1(a) and C4 stated in Section 4, $(\boldsymbol{\beta}_0, \boldsymbol{B}_0)$ uniquely minimizes the risk function $R(\boldsymbol{\beta}, \boldsymbol{B})$.*

One nice feature of R is that $\boldsymbol{\beta}$ and $\boldsymbol{B}$ can be separated orthogonally. To appreciate this, we observe that the inner product $\langle \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{P}_{\boldsymbol{X}}(\boldsymbol{W}\circ\boldsymbol{\Theta}^*\circ\boldsymbol{Y}), \boldsymbol{B} - \boldsymbol{P}_{\boldsymbol{X}}^{\perp}(\boldsymbol{W}\circ\boldsymbol{\Theta}^*\circ\boldsymbol{Y})\rangle = 0$ for any $\boldsymbol{B} \in \mathcal{N}(\boldsymbol{X})$. Consequently,

$$R(\boldsymbol{\beta}, \boldsymbol{B}) = \frac{1}{n_1 n_2}\left[\mathsf{E}\left\{\|\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{P}_{\boldsymbol{X}}(\boldsymbol{W}\circ\boldsymbol{\Theta}^*\circ\boldsymbol{Y})\|_F^2\right\} + \mathsf{E}\left\{\left\|\boldsymbol{B} - \boldsymbol{P}_{\boldsymbol{X}}^{\perp}(\boldsymbol{W}\circ\boldsymbol{\Theta}^*\circ\boldsymbol{Y})\right\|_F^2\right\}\right].$$

This decomposition will facilitate the fast computation of the proposed estimators and simplify their theoretical analyses.

If $\{\theta_{ij}\}$ were known, a natural unbiased estimator of R would be

$$\hat{R}(\boldsymbol{\beta}, \boldsymbol{B}) = \frac{1}{n_1 n_2}\left\{\|\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{P}_{\boldsymbol{X}}(\boldsymbol{W}\circ\boldsymbol{\Theta}^*\circ\boldsymbol{Y})\|_F^2 + \left\|\boldsymbol{B} - \boldsymbol{P}_{\boldsymbol{X}}^{\perp}(\boldsymbol{W}\circ\boldsymbol{\Theta}^*\circ\boldsymbol{Y})\right\|_F^2\right\}. \tag{3.1}$$

As $\{\theta_{ij}\}$ are often unknown, we modify $\hat{R}$ by plugging in consistent estimators $\{\hat{\theta}_{ij}\}$ of $\{\theta_{ij}\}$. We note that our proposed matrix recovery method can accommodate a variety of models of $\{\theta_{ij}\}$. To achieve various theoretical guarantees, $\{\hat{\theta}_{ij}\}$ are only required to fulfill a mild condition (C5 in Section 4) under the chosen model of $\{\theta_{ij}\}$. In the following, instead of $\hat{R}$, we consider

$$\hat{R}^*(\boldsymbol{\beta}, \boldsymbol{B}) = \frac{1}{n_1 n_2}\left\{\left\|\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{P}_{\boldsymbol{X}}\left(\boldsymbol{W}\circ\hat{\boldsymbol{\Theta}}^*\circ\boldsymbol{Y}\right)\right\|_F^2 + \left\|\boldsymbol{B} - \boldsymbol{P}_{\boldsymbol{X}}^{\perp}\left(\boldsymbol{W}\circ\hat{\boldsymbol{\Theta}}^*\circ\boldsymbol{Y}\right)\right\|_F^2\right\}, \tag{3.2}$$

where $\hat{\boldsymbol{\Theta}}^* = (\hat{\theta}_{ij}^{-1}) \in \mathbb{R}^{n_1\times n_2}$ contains reciprocals of the estimated observed rates $\{\hat{\theta}_{ij}\}$.

Since $\boldsymbol{\beta}$ and $\boldsymbol{B}$ are high dimensional parameters, a direct minimization of $\hat{R}^*$ would often result in over-fitting. To avoid such an issue, we incorporate penalty terms as regularizations. Specifically, the estimators $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{B}})$ is defined as the minimizer of

$$f(\boldsymbol{\beta}, \boldsymbol{B}; \lambda_1, \lambda_2, \alpha) = \hat{R}^*(\boldsymbol{\beta}, \boldsymbol{B}) + \lambda_1\|\boldsymbol{\beta}\|_F^2 + \lambda_2\left(\alpha\|\boldsymbol{B}\|_* + (1-\alpha)\|\boldsymbol{B}\|_F^2\right) \tag{3.3}$$

with respect to $\boldsymbol{\beta} \in \mathbb{R}^{m \times n_2}$ and $\boldsymbol{B} \in \mathcal{N}(\boldsymbol{X})$, where $\|\cdot\|_*$ is the nuclear norm and, $\lambda_1, \lambda_2 > 0$ along with $0 \leq \alpha \leq 1$ are regularization parameters. The two Frobenius norm terms, $\lambda_1 \|\boldsymbol{\beta}\|_F^2$ and $\lambda_2(1-\alpha)\|\boldsymbol{B}\|_F^2$, are equivalent to the computationally efficient ℓ_2-shrinkage of $\mathrm{vec}(\boldsymbol{\beta})$ as well as $\mathrm{vec}(\boldsymbol{B})$, while the nuclear norm term, $\lambda_2\alpha\|\boldsymbol{B}\|_*$, corresponds to the sparsity-promoting ℓ_1-shrinkage of the singular values of $\boldsymbol{B}$. The combination of these regularizations allows efficient computation and encourages the low-rank solution. Here the parameter α strikes a balance between the ℓ_1 and ℓ_2-shrinkage of $\boldsymbol{B}$. In our theoretical analysis, either $\alpha = 1$ or $\alpha \to 1$ would lead to the convergence of the proposed estimators. However, it is known that an appropriate amount of ℓ_2-regularization often improves finite sample performance (Zou and Hastie, 2005; Sun and Zhang, 2012). Hence, instead of fixing $\alpha = 1$, we select α, together with λ_1 and λ_2, by the 5-fold cross-validation (Friedman et al., 2013).

Due to the orthogonal separation of $\boldsymbol{\beta}$ and $\boldsymbol{B}$ in (3.2), the minimization of (3.3) is equivalent to the following two separate minimizations:

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta} \in \mathbb{R}^{m \times n_2}}{\arg\min} \left\{ \frac{1}{n_1 n_2} \left\| \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{P}_{\boldsymbol{X}} \left(\boldsymbol{W} \circ \hat{\boldsymbol{\Theta}}^* \circ \boldsymbol{Y} \right) \right\|_F^2 + \lambda_1 \|\boldsymbol{\beta}\|_F^2 \right\} \text{ and} \tag{3.4}$$

$$\hat{\boldsymbol{B}} = \underset{\boldsymbol{B} \in \mathcal{N}(\boldsymbol{X})}{\arg\min} \left\{ \frac{1}{n_1 n_2} \left\| \boldsymbol{B} - \boldsymbol{P}_{\boldsymbol{X}}^{\perp} \left(\boldsymbol{W} \circ \hat{\boldsymbol{\Theta}}^* \circ \boldsymbol{Y} \right) \right\|_F^2 + \lambda_2 \left(\alpha \|\boldsymbol{B}\|_* + (1-\alpha) \|\boldsymbol{B}\|_F^2 \right) \right\}. \tag{3.5}$$

3.2 Closed-form expressions and fast computation

We discuss how to compute $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{B}}$ given in (3.4) and (3.5). As (3.4) is essentially a ridge regression problem, straightforward algebra gives

$$\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^\intercal \boldsymbol{X} + \lambda_1' \boldsymbol{I}_{m \times m} \right)^{-1} \boldsymbol{X}^\intercal \left(\boldsymbol{W} \circ \hat{\boldsymbol{\Theta}}^* \circ \boldsymbol{Y} \right), \tag{3.6}$$

where $\lambda_1' = n_1 n_2 \lambda_1$ and $\boldsymbol{I}_{m \times m}$ denotes the m-by-m identity matrix. We observe that the matrix inversion in (3.6) is performed to a m-by-m matrix, which does not scale with n_1 and n_2. So it can be computed quite efficiently despite the high dimensionality of $\boldsymbol{A}$. As for the solution $\hat{\boldsymbol{B}}$ in (3.5), the minimization over $\boldsymbol{B} \in \mathcal{N}(\boldsymbol{X})$ is not straightforward. The following proposition, whose proof is given in Section S1 of the supplementary material, shows that the minimization problem (3.5) can

be carried out by extending the domain from $\mathcal{N}(\boldsymbol{X})$ to $\mathbb{R}^{n_1 \times n_2}$. This domain enlargement reduces the complexity of the minimization.

Proposition 2. *Suppose that* $\boldsymbol{X}^\intercal \boldsymbol{X}$ *is invertible, the minimization problem* (3.5) *is equivalent to*

$$\underset{\boldsymbol{B} \in \mathbb{R}^{n_1 \times n_2}}{\arg\min} \left\{ \frac{1}{n_1 n_2} \left\| \boldsymbol{B} - \boldsymbol{P}_{\boldsymbol{X}}^{\perp} \left(\boldsymbol{W} \circ \hat{\boldsymbol{\Theta}}^* \circ \boldsymbol{Y} \right) \right\|_F^2 + \lambda_2 \left(\alpha \left\| \boldsymbol{B} \right\|_* + (1-\alpha) \left\| \boldsymbol{B} \right\|_F^2 \right) \right\}. \tag{3.7}$$

An advantage of (3.7), over (3.5), is the availability of a closed-form solution based on existing results on singular value shrinkage (Mazumder et al., 2010) described as follows. To express the solution, let $\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\intercal$ be the singular value decomposition (SVD) of a matrix $\boldsymbol{D}$ where $\boldsymbol{\Sigma} = \mathrm{diag}(\{\sigma_i\})$. Define the corresponding singular value soft-thresholding (SVT) operator $\mathcal{T}_c$ by

$$\mathcal{T}_c(\boldsymbol{D}) = \boldsymbol{U}\mathrm{diag}(\{(\sigma_i - c)_+\})\boldsymbol{V}^\intercal \quad \text{for any } c \geq 0, \tag{3.8}$$

where $x_+ = \max(x, 0)$. As suggested by its name, this operator soft-thresholds the singular values of the input matrix $\boldsymbol{D}$ at a specified threshold c. It can be shown that the solution of (3.7) possesses the following closed-form expression:

$$\hat{\boldsymbol{B}} = \frac{1}{1 + 2(1-\alpha)\lambda_2'} \left\{ \mathcal{T}_{\alpha\lambda_2'} \left(\boldsymbol{P}_{\boldsymbol{X}}^{\perp} \left(\boldsymbol{W} \circ \hat{\boldsymbol{\Theta}}^* \circ \boldsymbol{Y} \right) \right) \right\}, \tag{3.9}$$

where $\lambda_2' = n_1 n_2 \lambda_2 / 2$. The proof of this result follows from the proof of Theorem 1 in Mazumder et al. (2010), which utilizes simple sub-gradient arguments after re-parameterizing the variable $\boldsymbol{B}$ of (3.7) in terms of its singular values and singular vectors. The explicit solution (3.9) indicates that both the singular value soft-thresholding procedure ($\mathcal{T}_{\alpha\lambda_2'}$) and a scaling procedure ($1/\{1 + 2(1-\alpha)\lambda_2'\}$) are involved in $\hat{\boldsymbol{B}}$. Observe that these two procedures arise separately from the nuclear norm regularization and the Frobenius norm regularization. When $\alpha = 1$ (only nuclear norm regularization), (3.9) involves no scaling. As for $\alpha = 0$ (only Frobenius norm regularization), no soft-thresholding is administrated.

Among existing matrix completion algorithms, a set of them (Troyanskaya et al., 2001; Mazumder et al., 2010; Ma et al., 2011) require iterative applications of SVD to n_1-by-n_2 matrices. In contrast, the computation of $\hat{\boldsymbol{B}}$ in (3.9) requires only a single SVD of the matrix $\boldsymbol{P}_{\boldsymbol{X}}^{\perp}(\boldsymbol{W} \circ \hat{\boldsymbol{\Theta}}^* \circ \boldsymbol{Y})$ due to

the application of $\mathcal{T}_{\alpha\lambda_2'}$. Specifically, to obtain $\hat{\boldsymbol{B}}$ with respect to multiple choices of λ_2' (or λ_2) and α, the exact same SVD is needed. This is particularly favorable to tuning parameter selection, and allows us to perform the k-fold cross-validation procedure (Mazumder et al., 2010; Xu et al., 2013; Chiang et al., 2015) with much reduced computational burden. In all of our numerical evaluations, we choose $k = 5$. As for most alternative matrix completion algorithms, iterative applications of SVD need to be re-applied for every choice of tuning parameters, leading to a nested loop of SVDs and hence significant computational burden.

To further improve the computational efficiency of our method, we provide an approximate computational procedure for the low-rank solutions (3.7). This approximate procedure is particularly useful, when n_1 and n_2 are large, as the computation of a full SVD requires significant computational resources. The key component is the fast randomized singular value (soft-)thresholding (FRSVT) procedure (Oh et al., 2015), which utilizes random projections (Halko et al., 2011) to approximate the SVT operator. Recent work (Halko et al., 2011) has shown that random projections can explore the low-rank structure effectively, and are suitable for constructing efficient algorithms of approximate low-rank matrix factorizations. In FRSVT, random projections are obtained through the generation of Gaussian random matrix with independent entries. To approximate SVT with output rank at most L, the number of random projections $L + d$ is required to be higher than L. In the numerical illustrations of this paper, we set $L = 150$ and $d = 5$.

4 Asymptotic Convergence Rates

Let $\|\boldsymbol{A}\| = \sigma_{\max}(\boldsymbol{A})$ and $\|\boldsymbol{A}\|_\infty = \max_{i,j} |A_{ij}|$ be the spectral and the maximum norms of a matrix $\boldsymbol{A}$, 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 $b_n = O(a_n)$, and $n = n_1 + n_2$. The mean squared error of a generic estimator $\tilde{\boldsymbol{A}}$ is defined as $d^2(\tilde{\boldsymbol{A}}, \boldsymbol{A}_0) = \|\tilde{\boldsymbol{A}} - \boldsymbol{A}_0\|_F^2/(n_1 n_2)$.

In this section, we first establish a general convergence result on $d^2(\hat{\boldsymbol{A}}, \boldsymbol{A}_0)$ in Theorem 1, followed by more specific results on the convergence rates under the uniform probability of observation

model and the logistic regression model, respectively. Further, the convergence rate of $\|\hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_{0j}\|_F$ is established.

The technical conditions needed for our analysis are given as follows.

C1. (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, $\max_{i,j} \mathsf{E}|\epsilon_{ij}|^l \leq \frac{1}{2} l! c_\sigma^2 \eta^{l-2}$ for any positive integer $l \geq 2$.

C2. The design matrix $\boldsymbol{X}$ is of size $n_1 \times m$ such that $n_1 > m$. Moreover, there exists a positive constant a_x such that $\|\boldsymbol{X}\|_\infty < a_x$ and $\boldsymbol{X}^\intercal \boldsymbol{X}$ is invertible. Furthermore, there exists a finite symmetric matrix $\boldsymbol{S}_x$ with $0 < \sigma_{\min}(\boldsymbol{S}_x) \leq \|\boldsymbol{S}_x\| < \infty$ such that $n_1^{-1} \boldsymbol{X}^\intercal \boldsymbol{X} \to \boldsymbol{S}_x$ as $n_1 \to \infty$.

C3. There exist some positive constants a_1 and a_2 such that

$$\max\left\{\|\boldsymbol{X}\boldsymbol{\beta}_0\|_\infty, \|\boldsymbol{A}_0\|_\infty\right\} \leq \sqrt{\log(n)} a_1 \quad \text{and} \quad \max\left\{\|\boldsymbol{A}_0\|_{\infty,2}, \|\boldsymbol{A}_0^\intercal\|_{\infty,2}\right\} \leq \sqrt{n_1 \vee n_2} a_2.$$

C4. The indicators of observed entries $\{\omega_{ij}\}_{i,j=1}^{n_1,n_2}$ are mutually independent and $\omega_{ij} \sim \text{Bern}(\theta_{ij})$ for $\theta_{ij} \in (0,1)$, and are independent of $\{\epsilon_{ij}\}_{i,j=1}^{n_1,n_2}$. Furthermore, for $i = 1, \ldots, n_1$ and $j = 1, \ldots, n_2$, $\mathsf{P}(\omega_{ij} = 1 | \boldsymbol{x}_i, Y_{ij}) = \mathsf{P}(\omega_{ij} = 1 | \boldsymbol{x}_i) =: \theta_{ij}(\boldsymbol{x}_i) = \theta_{ij}$ where $\boldsymbol{x}_i^\intercal$ is the i-th row of the covariate matrix.

C5. (a) There exists a lower bound $\theta_L \in (0,1)$ such that $\min_{i,j}\{\theta_{ij}\} \geq \theta_L > 0$, where θ_L is allowed to depend on n_1 and n_2. (b) The estimators $\{\hat{\theta}_{ij}\}$ are consistent to $\{\theta_{ij}\}$, free of the tuning parameters λ_1', λ_2' and α, and are independent of $\{\epsilon_{ij}\}$. Moreover, there exists a positive constant t_0 such that for all $t > t_0$, $\mathsf{P}\{\sum_{ij}(1/\hat{\theta}_{ij} - 1/\theta_{ij})^2 \geq c_{n_1,n_2} t\} \leq g(t) + h_{n_1,n_2}$, where c_{n_1,n_2} and h_{n_1,n_2} are model specific nonrandom sequences depending on n_1 and n_2 and are independent of t such that $\lim_{n_1,n_2 \to \infty} h_{n_1,n_2} = 0$; and $g(t)$ is a function independent of n_1 and n_2 such that $\lim_{t \to \infty} g(t) \to 0$.

Condition C1(b) is the Bernstein condition which, together with C1(a), covers a variety of distributions for ϵ_{ij} including the Gaussian distribution $\epsilon_{ij} \sim \mathcal{N}(0, \sigma_{ij}^2)$ for positive constants σ_{ij}^2. In Condition C2, the requirement $n_1 > m$ is easily met as the number of covariates per subject is fixed. As the dimensions of $n_1^{-1} \boldsymbol{X}^\intercal \boldsymbol{X}$ are fixed at m-by-m, the rest of Condition C2 are quite standard. Condition C3 extends the conditions that $\|\boldsymbol{X}\boldsymbol{\beta}_0\|_\infty < \infty$ and $\|\boldsymbol{A}_0\|_\infty < \infty$ as assumed, for instance, by Keshavan et al. (2009), Koltchinskii et al. (2011), Sun and Zhang (2012) and Cai

and Zhou (2016), by allowing both $\boldsymbol{X}\boldsymbol{\beta}_0$ and $\boldsymbol{A}_0$ diverge at certain rates.

Condition C4 prescribes the independent Bernoulli model for the indicator of observing Y_{ij}, where the probability of observation θ_{ij} can depend on the covariate. This is analogous to the notion of the missing-at-random (MAR) commonly assumed in the missing value literature (Little and Rubin, 2014). A specific MAR model is the logistic regression model

$$\theta_{ij} = \theta_{ij}(\boldsymbol{x}_i) = \frac{\exp\{(1, \boldsymbol{x}_i^\intercal)\boldsymbol{\gamma}_{\cdot j}\}}{1 + \exp\{(1, \boldsymbol{x}_i^\intercal)\boldsymbol{\gamma}_{\cdot j}\}}, \tag{4.1}$$

where $\boldsymbol{\gamma}_{\cdot j} \in \mathbb{R}^{m+1}$ are the j-th column specific parameter vectors. Most of the existing studies in matrix completion (Keshavan et al., 2009; Gross, 2011; Recht, 2011; Rohde and Tsybakov, 2011; Koltchinskii et al., 2011; Sun and Zhang, 2012) focus on the so-called Uniform Sampling at Random (USR) scheme. Let $N = \sum_{i,j} w_{ij}$ be the total number of observations. Conditioning on N, the USR takes a random sample of N observed indices from the set $\{(i,j) : i \in \{1,\ldots,n_1\}, j \in \{1,\ldots,n_2\}\}$, independently with the uniform sampling probability $N/(n_1 n_2)$ with replacement. The "with replacement" means that a $A_{0,ij}$ can be observed more than once, which is not suitable for some matrix completion problems, for instance the Netflix prize problem (Feuerverger et al., 2012) as a viewer would not rate a movie more than once. There are studies (Srebro and Salakhutdinov, 2010; Negahban and Wainwright, 2012; Klopp, 2014; Cai and Zhou, 2016) which adopt heterogeneous sampling probability models without utilizing covariates, for instance heterogeneity with respect to the rows and columns while assuming the sampling of the row and the column are independent. Condition C4 introduces heterogeneity through covariates while including the aforementioned uniform and logistic regression models as special cases.

In Condition C5(a), imposing the lower bound θ_L in the probabilities of observation ensures each entry of the matrix has a minimum positive probability of observation. However, our condition does not impose the restriction that the number of observed entries is of the same order as $n_1 n_2$, since θ_L is allowed to go to 0 with n_1 and n_2 growing. For instance, one could take $\theta_L \asymp r_{\boldsymbol{B}_0} n \log^2(n)/n_1 n_2$ to mimic scenarios with $c r_{\boldsymbol{B}_0} n \log^2(n)$ observed entries as discussed in Section 1. The second part of Condition C5(b) is used to quantify the sum of squared errors in estimating $1/\theta_{ij}$ by the consistent

estimator $1/\hat{\theta}_{ij}$. The convergence rate c_{n_1,n_2} and the error bound functions $g(t)$ and h_{n_1,n_2} are given in a general setting, whose orders of magnitude are dependent of the model for θ_{ij}. We establish Condition C5(b) in Section S3 under the logistic regression model given in (4.1) via the uniform asymptotic normality of the maximum likelihood estimators (MLE) by applying Sweeting (1980)'s result. Condition C5(b) is also fulfilled under other sampling mechanisms including the uniform probability of observation model (i.e. $\theta_{ij} \equiv \theta_0$).

For any $\delta_\sigma > 0$, and $t \in (0, t_0)$, c_{n_1,n_2} specified in Condition C5(b), define

$$\Delta\left(\delta_\sigma, t\right) = \max\left\{\frac{\sqrt{\left(n_1 \vee n_2\right)\log\left(n\right)}}{\sqrt{\theta_L} n_1 n_2}, \left(n_1 n_2\right)^{-3/4}\left(c_{n_1,n_2} t\right)^{1/2}\log^{\delta_\sigma/4}\left(n\right)\right\} \tag{4.2}$$

and $\eta_{n_1,n_2}(g, \delta_\sigma, t) = 4g(t) + 4h_{n_1,n_2} + C\log^{-\delta_\sigma}(n)$ for a positive constant C. Here, $g(t)$ and h_{n_1,n_2} are specified in C5(b), and C5(b) implies that $\lim_{t\to\infty}\lim_{n_1,n_2\to\infty}\left\{\eta_{n_1,n_2}\left(g, \delta_\sigma, t\right)\right\} = 0$. The following Theorem 1 is proved in Section S5 of the supplementary material.

Theorem 1. *Assume Conditions C1-C5,* $0 < \alpha \leq 1$, $\lambda_1 = o(n_2^{-1})$ *and* $\lambda_2\alpha \geq (2+4m)C_0\Delta(\delta_\sigma, t)$, *for any* $t > t_0$ *and positive constants* δ_σ *and* C_0. *Then, for a positive constant* C',

$$\begin{aligned} d^2\left(\hat{\boldsymbol{A}}, \boldsymbol{A}_0\right) \leq & C' \max\left\{\min\left\{\lambda_2\alpha\left\|\boldsymbol{B}_0\right\|_*, n_1 n_2 r_{\boldsymbol{B}_0}\left(\lambda_2\alpha\right)^2\right\}, \right. \\ & \left. \lambda_2\left(1-\alpha\right)\left\|\boldsymbol{B}_0\right\|_F^2, n_1 n_2\Delta^2(\delta_\sigma, t), n_2^2\lambda_1^2\left\|\boldsymbol{X}\boldsymbol{\beta}_0\right\|_F^2\right\} \end{aligned} \tag{4.3}$$

with probability at least $1 - \eta_{n_1,n_2}(g, \delta_\sigma, t)$.

The diminishing $\eta_{n_1,n_2}(g, \delta_\sigma, t)$ means that $d^2(\hat{\boldsymbol{A}}, \boldsymbol{A}_0)$ is bounded by the right hand side of (4.3) with probability approaching 1 for n_1, n_2 and t large enough. We note that the order of the upper bound for $d^2(\hat{\boldsymbol{A}}, \boldsymbol{A}_0)$, as prescribed in (4.3), depends on the specific orders of $\Delta(\delta_\sigma, t)$, $\|\boldsymbol{B}_0\|_*$, $r_{\boldsymbol{B}_0}$, $\|\boldsymbol{X}\boldsymbol{\beta}_0\|_F$ and $\|\boldsymbol{B}_0\|_F$ and the choices of parameters λ_1, λ_2 and α. In the following, from (4.3), we derive specific convergence rates for $d^2(\hat{\boldsymbol{A}}, \boldsymbol{A}_0)$ under two models of θ_{ij}.

We first consider the uniform probability of observation model such that $\theta_{ij} \equiv \theta_0$. Under this model, the MLE for θ_0 is $\hat{\theta}_{ij} \equiv N/(n_1 n_2)$. It can be shown that we can choose $c_{n_1,n_2} = (1-\theta_0)/\theta_0$, for any $t_0 > 0$, $g(t) = \mathsf{P}\{\chi_1^2 > t\}$ and $h_{n_1,n_2} = \sup_t|\mathsf{P}\{\theta_0(1/\hat{\theta} - 1/\theta_0)^2/(1-\theta_0) \geq t\} - g(t)|$ in

Condition C5(b) so that C5(b) holds for any positive t. With the above choice of c_{n_1,n_2}, $0 < \delta_\sigma < 2$ and choosing t such that

$$t_0 < t < (n_1 n_2)^{-1/2} (n_1 \vee n_2) \log^{1-\delta_\sigma/2}(n), \tag{4.4}$$

then $\sup_t \Delta(\delta_\sigma, t) \asymp \Delta_1 =: \theta_0^{-1/2}(n_1 \vee n_2)^{1/2}(n_1 n_2)^{-1} \log^{1/2}(n)$.

Corollary 1. *Assume Conditions C1-C5, under the uniform probability of observation model, choose* $c_{n_1,n_2} = (1-\theta_0)/\theta_0$, $0 < \delta_\sigma < 2$ *and* t *as in* (4.4), $\lambda_1 = n_2^{-1}\log^{-1/2}(n)\Delta_1$, $1-\alpha \asymp 1/(n_1 n_2)$, $\lambda_2 \asymp \theta_0^{-1/2}(n_1 \wedge n_2)^{-1/2}(n_1 n_2)^{-1/2}\log^{1/2}(n)$ *in* (3.3). *Then, for a positive constant* C', *with probability at least* $1 - \eta_{n_1,n_2}(g, \delta_\sigma, t)$,

$$\textit{both} \quad d^2\left(\hat{\boldsymbol{A}}, \boldsymbol{A}_0\right) \quad \textit{and} \quad d^2\left(\hat{\boldsymbol{B}}, \boldsymbol{B}_0\right) \leq C' r_{\boldsymbol{B}_0} \theta_0^{-1} (n_1 \wedge n_2)^{-1} \log(n).$$

The corollary establishes that $d^2(\hat{\boldsymbol{A}}, \boldsymbol{A}_0)$ and $d^2(\hat{\boldsymbol{B}}, \boldsymbol{B}_0)$ are all $O_p\{r_{\boldsymbol{B}_0}\theta_0^{-1}(n_1 \wedge n_2)^{-1}\log(n)\}$. We note that the choice of parameter λ_2 actually depend on the magnitude of the noise $c_\sigma^2 = \max_{i,j}\{\sigma_{ij}^2\}$ as shown in Lemmas S4.1-S4.3 of Section S4 of the supplementary material. This means that $d^2(\hat{\boldsymbol{A}}, \boldsymbol{A}_0)$ depends implicitly on the level of the noise as well. Although the corollary assumes the uniform observation probability, its conclusions are valid for other missing models that accommodate the rate of $c_{n_1,n_2} = (1-\theta_0)/\theta_0$. In our analysis, the effect of the sample size N enters our results through the Binomial mean $n_1 n_2 \theta_0$ as it is of the same order of N. We note that Condition C5(a) allows $\theta_0 = \theta_L$ to depend on n_1 and n_2 and to diminish to zero as n_1 and n_2 diverge to infinity.

We note that the rate attained by Corollary 1 coincides with that of the other matrix completion methods, for instance Sun and Zhang (2012)'s calibrated elastic regularization estimator $\hat{\boldsymbol{A}}^{\mathrm{SZ}}$, Negahban and Wainwright (2012)'s row/column weighted regularization estimator $\hat{\boldsymbol{A}}^{\mathrm{NW}}$, Koltchinskii et al. (2011)'s prior mask distribution estimator $\hat{\boldsymbol{A}}^{\mathrm{KLT}}$ and Mazumder et al. (2010)'s matrix lasso estimator $\hat{\boldsymbol{A}}^{\mathrm{MHT}}$, under either the USR or the row and column product weight model of Negahban and Wainwright (2012). These methods also require the "incoherence conditions" (Candès and Recht, 2009), and/or the spikiness measure $\alpha(\boldsymbol{A}_0) = \sqrt{n_1 n_2}\|\boldsymbol{A}_0\|_\infty / \|\boldsymbol{A}_0\|_F$ of $\boldsymbol{A}_0$ to be bounded.

We now consider the scenario where the observation probability θ_{ij} follows the logistic regression model given in (4.1). As will be shown in the next corollary, this induces a different rate for c_{n_1,n_2} and a slower convergence rates for the estimators. For any $\delta_\sigma > 0$, it is shown in Section S3 of the supplementary material that for some constants η_g depending on θ_L and C_m, we can choose $c_{n_1,n_2} = \eta_g^{-1} n_2 \log(n_2)$, $t_0 = m+3$, $g(t) = C_m t \exp\{-t/2\}$, and $h_{n_1,n_2} = n_2 \max_j \sup_t |\mathsf{P}\{\sum_i (1/\hat{\theta}_{ij} - 1/\theta_{ij})^2 \geq t\} - \mathsf{P}(\chi^2_{m+1} \geq \eta_g t)|$ in Condition C5(b) so that C5(b) holds for any positive $t > t_0$ for the logistic model.

By choosing t such that

$$m + 3 < t < \log^{\delta_\sigma/6}(n)\,, \tag{4.5}$$

we have $\sup_t \Delta(\delta_\sigma, t) = \Delta_2(\delta_\sigma) \asymp \eta_g^{-1/2} n_1^{-3/4} n_2^{-1/4} \log^{1/2}(n_2) \log^{\delta_\sigma/3}(n)$. This implies that the convergence rate of $d^2(\hat{\boldsymbol{A}}, \boldsymbol{A}_0)$ given in (4.3) is $\eta_g^{-1} n_1^{-1/2} n_2^{1/2} \log(n_2) \log^{2\delta_\sigma/3}(n)$, as summarized in the following corollary.

Corollary 2. *Assume Conditions C1-C5, $n_1 n_2 \theta_L > (n_1 \vee n_2)\log(n)$ and the logistic model. Choose $c_{n_1,n_2} = \eta_g^{-1} n_2 \log(n_2)$, t as (4.5), $\lambda_1 = n_2^{-1} \log^{-1/2}(n) \Delta_2(\delta_\sigma)$ for any $\delta_\sigma > 0$, $1-\alpha \asymp 1/(n_1 n_2)$, $\lambda_2 \asymp \eta_g^{-1/2} n_1^{-3/4} n_2^{-1/4} \log^{1/2}(n_2) \log^{\delta_\sigma/3}(n)$ in (3.3). Then, for a positive constant C', with probability at least $1 - \eta_{n_1,n_2}(g, \delta_\sigma, t)$,*

$$\textit{both} \quad d^2\left(\hat{\boldsymbol{A}}, \boldsymbol{A}_0\right) \quad \textit{and} \quad d^2\left(\hat{\boldsymbol{B}}, \boldsymbol{B}_0\right) \leq C' r_{\boldsymbol{B}_0} \eta_g^{-1} n_1^{-1/2} n_2^{1/2} \log(n_2) \log^{2\delta_\sigma/3}(n)\,.$$

Corollary 2 implies that $d^2(\hat{\boldsymbol{A}}, \boldsymbol{A}_0)$ and $d^2(\hat{\boldsymbol{B}}, \boldsymbol{B}_0)$ are both $O_p\{r_{\boldsymbol{B}_0} \eta_g^{-1} n_1^{-1/2} n_2^{1/2} \log(n_2) \log^{2\delta_\sigma/3}(n)\}$. The assumption that $n_1 n_2 \theta_L > (n_1 \vee n_2)\log(n)$ is usually considered in existing matrix completion works. Using the proof of Corollary 2, it can be shown that the convergence rates for $d^2(\hat{\boldsymbol{A}}, \boldsymbol{A}_0)$ and $d^2(\hat{\boldsymbol{B}}, \boldsymbol{B}_0)$ can be simplified to $r_{\boldsymbol{B}_0} \log^{-2\delta_\sigma/3}(n_2)$ if $n_1 \asymp \eta_g^2 n_2 \log^{2+2\delta_\sigma}(n_2)$. In our results, we only specify the order of λ_2 although the choice of λ_2 depends on the magnitude of the noise $c_\sigma^2 = \max_{i,j}\{\sigma_{ij}^2\}$, as shown in Lemmas S4.1-S4.3 of Section S4 of the supplementary material.

Compared with the case of the uniform probability of observation considered in Corollary 1, the convergence rate of $r_{\boldsymbol{B}_0} \eta_g^{-1} n_1^{-1/2} n_2^{1/2} \log(n_2) \log^{2\delta_\sigma/3}(n)$ is much slower than $r_{\boldsymbol{B}_0} \theta_L^{-1}(n_1 \wedge$

$n_2)^{-1}\log(n)$. This is because of a much larger c_{n_1,n_2} due to the heterogeneity in the probability of observation as prescribed by the logistic model. This heterogeneity results in a larger amount of errors being accumulated in the estimation of $\{\theta_{ij}\}$, that slows down the convergence.

The coefficient matrix $\boldsymbol{\beta}_0$ helps to interpret the role of covariates in completing the target matrix through the parametric component $\boldsymbol{X}\boldsymbol{\beta}_0$. The following theorem provides the convergence rate of $\hat{\boldsymbol{\beta}}_j$ under a general setting.

Theorem 2. *Let $\hat{\boldsymbol{\beta}}_j$ and $\boldsymbol{\beta}_{0j}$ be the j-th column of $\hat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}_0$ respectively. Assume Conditions C1, C2, C4 and C5(a), and the estimators $\hat{\theta}_{ij}$ of θ_{ij} satisfy that for $|\hat{\theta}_{ij}-\theta_{ij}| = O_p(n_1^{-1/2})$. If $\|\boldsymbol{\beta}_0\|_F > 0$, $\|\boldsymbol{\beta}_0\|_\infty < \infty$ and $\lambda_1 = o(n_2^{-1})$, we have $\|\hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_{0j}\|_F = O_p(n_1^{-1/2})$ for each $j = 1,\ldots,n_2$.*

While the convergence of $\hat{\boldsymbol{\beta}}_j$ is of the standard rate, the theorem does not require any specification of c_{n_1,n_2} and any restriction on the regularization parameters λ_2 and α as in Theorem 1 and its two corollaries. Furthermore, Condition C5(b) is replaced by a mild convergence rate of the estimators $\{\hat{\theta}_{ij}\}$ which is more easily met. These are all due to the closed-form expression of $\hat{\boldsymbol{\beta}}$ given in (3.6). However, despite the $\sqrt{n_1}$-convergence rate of each $\hat{\boldsymbol{\beta}}_j$, we are unable to translate this rate for $\hat{\boldsymbol{\beta}}$. This is because the convergence rates for the whole matrix as stated in Theorem 1 as well as Corollaries 1 and 2 are slower than the $\sqrt{n_1}$-rate.

5 Benefits of Covariate Information

In this section, we outline some theoretical benefits of considering covariate information. More specifically, we compare the upper bounds of the mean squared errors of $\boldsymbol{A}_0$ achieved by our estimator and the one from Koltchinskii et al. (2011) under uniform missingness.

If $m \ll \min(n_1, n_2)$ and $\boldsymbol{B}_0$ is of low rank, our target matrix $\boldsymbol{A}_0 = \boldsymbol{X}\boldsymbol{\beta}_0 + \boldsymbol{B}_0$ is also a low-rank matrix. Without using the covariate $\boldsymbol{X}$, one can recover $\boldsymbol{A}_0$ by existing matrix completion techniques. A natural question is whether the utilization of the covariates improves the estimation. This question is addressed theoretically in this section by comparing non-asymptotic upper

bounds of mean squared errors. In addition, empirical evidences are shown in Sections 6 and 7 to demonstrate the benefits of using covariates.

To provide a simple and transparent comparison with existing results, we restrict our study to the uniform missingness while the target matrix follows $\boldsymbol{A}_0 = \boldsymbol{X}\boldsymbol{\beta}_0 + \boldsymbol{B}_0$.

Write $N = \sum_{i,j} \omega_{ij}$. Under the uniform missing mechanism, one can use N/n_1n_2 to estimate the common observation probability $\theta_{ij} \equiv \theta_0$ where $\theta_0 > 0$ is allowed to depend on n_1 and n_2 in our analysis; see Condition C5(a) in Section 4 for details. For clarity, we write the estimator $(\hat{\boldsymbol{\beta}}^{\mathrm{UNI}}, \hat{\boldsymbol{B}}^{\mathrm{UNI}})$ of the proposed methodology as

$$\hat{\boldsymbol{\beta}}^{\mathrm{UNI}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^{m\times n_2}}{\arg\min} \left\{ \frac{1}{n_1n_2} \left\| \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{P}_{\boldsymbol{X}} \left(\frac{n_1n_2}{N} \boldsymbol{W} \circ \boldsymbol{Y} \right) \right\|_F^2 + \lambda_1 \|\boldsymbol{\beta}\|_F^2 \right\} \text{ and} \tag{5.1}$$

$$\hat{\boldsymbol{B}}^{\mathrm{UNI}} = \underset{\boldsymbol{B}\in\mathbb{R}^{n_1\times n_2}}{\arg\min} \left\{ \frac{1}{n_1n_2} \left\| \boldsymbol{B} - \boldsymbol{P}_{\boldsymbol{X}}^{\perp} \left(\frac{n_1n_2}{N} \boldsymbol{W} \circ \boldsymbol{Y} \right) \right\|_F^2 + \lambda_2 \|\boldsymbol{B}\|_* \right\}, \tag{5.2}$$

when α in (3.5) is set to 1. By writing $\hat{\boldsymbol{A}}^{\mathrm{UNI}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}^{\mathrm{UNI}} + \hat{\boldsymbol{B}}^{\mathrm{UNI}}$, the mean squared error $d^2(\hat{\boldsymbol{A}}^{\mathrm{UNI}}, \boldsymbol{A}_0)$ can be decomposed as $d^2(\boldsymbol{X}\hat{\boldsymbol{\beta}}^{\mathrm{UNI}}, \boldsymbol{X}\boldsymbol{\beta}_0) + d^2(\hat{\boldsymbol{B}}^{\mathrm{UNI}}, \boldsymbol{B}_0)$. If the covariates are not utilized, (5.2) (without the projection $\boldsymbol{P}_{\boldsymbol{X}}^{\perp}$) alone leads to the estimator $\hat{\boldsymbol{A}}^{\mathrm{KLT}}$ of Koltchinskii et al. (2011):

$$\hat{\boldsymbol{A}}^{\mathrm{KLT}} = \underset{\boldsymbol{A}\in\mathbb{R}^{n_1\times n_2}}{\arg\min} \left\{ \frac{1}{n_1n_2} \left\| \boldsymbol{A} - \frac{n_1n_2}{N} \boldsymbol{W} \circ \boldsymbol{Y} \right\|_F^2 + \lambda_{\mathrm{KLT}} \|\boldsymbol{A}\|_* \right\}.$$

In the following, we compare $\hat{\boldsymbol{A}}^{\mathrm{UNI}}$ and $\hat{\boldsymbol{A}}^{\mathrm{KLT}}$ to reveal a benefit of the covariate.

It is shown in Theorem 3 of Koltchinskii et al. (2011) that if $\lambda_{\mathrm{KLT}} \geq 2\|\boldsymbol{M}\|$, then

$$d^2\left(\hat{\boldsymbol{A}}^{\mathrm{KLT}}, \boldsymbol{A}_0\right) \leq \lambda_{\mathrm{KLT}} \min \left\{ 2\|\boldsymbol{A}_0\|_*, \left(\frac{1+\sqrt{2}}{2} \right)^2 \lambda_{\mathrm{KLT}} n_1 n_2 r_{\boldsymbol{A}_0} \right\} =: U_{\mathrm{KLT}}, \tag{5.3}$$

say, where $\boldsymbol{M} = \boldsymbol{W} \circ \boldsymbol{Y}/N - \boldsymbol{A}_0/(n_1n_2)$. Similarly, for the proposed estimator, it can be shown that if $\lambda_2 \geq 2\|\boldsymbol{M}\|$,

$$d^2\left(\hat{\boldsymbol{B}}^{\mathrm{UNI}}, \boldsymbol{B}_0\right) \leq \lambda_2 \min \left\{ 2\|\boldsymbol{B}_0\|_*, \left(\frac{1+\sqrt{2}}{2} \right)^2 \lambda_2 n_1 n_2 r_{\boldsymbol{B}_0} \right\} =: U_{\mathrm{UNI}}. \tag{5.4}$$

Due to Lemmas S4.1-S4.3 of the supplementary material, there exist positive constants C and δ_σ such that $\|\boldsymbol{M}\| \leq C\theta_0^{-1/2}(n_1 \wedge n_2)^{-1/2}(n_1n_2)^{-1/2}\log^{1/2}(n)$ with probability at least $1 - 2/n -$

$4\log^{-\delta_\sigma}(n)$. We note that Koltchinskii et al. (2011) obtain the same rate for $\|\boldsymbol{M}\|$ in a similar fashion. Due to this theoretical guarantee, we pick $\lambda_2 = \lambda_{\mathrm{KLT}} = C\theta_0^{-1/2}(n_1 \wedge n_2)^{-1/2}(n_1 n_2)^{-1/2}\log^{1/2}(n)$.

The benefit of the covariate lies in the fast convergence of $\boldsymbol{X}\hat{\boldsymbol{\beta}}^{\mathrm{UNI}}$. As shown in Section S2.1 of the supplementary material, if $\lambda_1 = o\{n_1^{-1}n_2^{-3/2}\log^{-1}(n)\}$, then $d^2(\boldsymbol{X}\hat{\boldsymbol{\beta}}^{\mathrm{UNI}}, \boldsymbol{X}\boldsymbol{\beta}_0) = O_p(n_1^{-1})$ which is dominated by the bound U_{UNI} of $d^2(\hat{\boldsymbol{B}}^{\mathrm{UNI}}, \boldsymbol{B}_0)$ in (5.4). As $d^2(\hat{\boldsymbol{A}}^{\mathrm{UNI}}, \boldsymbol{A}_0) = d^2(\boldsymbol{X}\hat{\boldsymbol{\beta}}^{\mathrm{UNI}}, \boldsymbol{X}\boldsymbol{\beta}_0) + d^2(\hat{\boldsymbol{B}}^{\mathrm{UNI}}, \boldsymbol{B}_0)$, we only have to compare the bounds U_{KLT} and U_{UNI} in (5.3) and (5.4) when n_1 is large enough. Since these two bounds are of the same order, we have to analyze the corresponding constant factors. Since $r_{\boldsymbol{B}_0} \le r_{\boldsymbol{A}_0}$ and $\|\boldsymbol{B}_0\|_* \le \|\boldsymbol{A}_0\|_*$ (Proposition S2.1 of the supplementary material), we can conclude that $U_{\mathrm{UNI}} \le U_{\mathrm{KLT}}$. In addition, if $\boldsymbol{\beta}_0 \neq \boldsymbol{0}^{m\times n_2}$ and the rank of $\boldsymbol{A}_0$ is small, i.e., of order $o\{\theta_0^{1/2}(n_1 \wedge n_2)^{1/2}\}$, we have $U_{\mathrm{UNI}} < U_{\mathrm{KLT}}$, which implies a strictly better upper bound for $d^2(\hat{\boldsymbol{A}}^{\mathrm{UNI}}, \boldsymbol{A}_0)$ than $d^2(\hat{\boldsymbol{A}}^{\mathrm{KLT}}, \boldsymbol{A}_0)$. This illustrates the benefit of utilizing the covariates. The details are summarized in the following theorem whose proof is given in Section S2.1 of the supplementary material.

Theorem 3. *Assume Conditions C1-C3, and take* $\lambda_2 = \lambda_{\mathrm{KLT}} = C\theta_0^{-1/2}(n_1 \wedge n_2)^{-1/2}(n_1 n_2)^{-1/2}\log^{1/2}(n)$ *in both* (5.3) *and* (5.4). *Then* $U_{\mathrm{UNI}} \le U_{\mathrm{KLT}}$. *Furthermore,* $U_{\mathrm{UNI}} < U_{\mathrm{KLT}}$ *if* $\boldsymbol{\beta}_0 \neq \boldsymbol{0}^{m\times n_2}$ *and either one of the two following conditions holds: (i). (low-rank condition)* $r_{\boldsymbol{A}_0} = r_{\boldsymbol{B}_0} + m = o\{\theta_0^{1/2}(n_1 \wedge n_2)^{1/2}\}$, *or (ii). (row space condition)* $\mathcal{R}(\boldsymbol{\beta}_0) \nsubseteq \mathcal{R}(\boldsymbol{B}_0)$.

In the following, we provide a lower bound for $d^2(\hat{\boldsymbol{A}}^{\mathrm{UNI}}, \boldsymbol{A}_0)$. To this end, define two matrix classes

$$\beta(a_1) = \left\{\boldsymbol{\beta} \in \mathbb{R}^{m\times n_2} : \|\boldsymbol{X}\boldsymbol{\beta}\|_\infty \le a_1\right\}, \quad \mathcal{B}(r, a_1) = \left\{\boldsymbol{B} \in \mathbb{R}^{n_1\times n_2} : r_{\boldsymbol{B}} \le r, \|\boldsymbol{B}\|_\infty \le a_1\right\}.$$

Theorem 4. *Fix* $a_1 > 0$, *for* $r_{\boldsymbol{B}_0}$ *such that* $1 \le r_{\boldsymbol{B}_0} \le \min(n_1, n_2) - m$, $(n_1 \vee n_2)r_{\boldsymbol{B}_0} \le n_1 n_2\theta_0$. *Assume that* $\omega_{ij} \sim Bern(\theta_0)$ *for* $\theta_0 \in (0,1)$. *Let* $\{\epsilon_{ij}\}$ *be IID Gaussian* $\mathcal{N}(0, \sigma^2)$ *with* $\sigma^2 > 0$. *Then, there exist absolute constants* $\alpha \in (0,1)$, $c > 0$ *and* $0 \le l \le r_{\boldsymbol{B}_0}$ *such that*

$$\inf_{\hat{\boldsymbol{\beta}}^{\mathrm{UNI}}, \hat{\boldsymbol{B}}^{\mathrm{UNI}}} \sup_{\boldsymbol{\beta}_0 \in \beta(a_1), \boldsymbol{B}_0 \in \mathcal{B}(r_{\boldsymbol{B}_0}, a_1)} P\left(d^2(\hat{\boldsymbol{A}}^{\mathrm{UNI}}, \boldsymbol{A}_0) > c(\sigma \wedge a_1)^2 \frac{(n_1 \vee n_2)(r_{\boldsymbol{B}_0} + l)}{n_1 n_2 \theta_0}\right) \ge \alpha.$$

Theorem 4 establishes $c(\sigma \wedge a_1)^2(n_1 \vee n_2)(r_{\boldsymbol{B}_0} + l)/(n_1 n_2 \theta_0)$ as a lower bound for $d^2(\hat{\boldsymbol{A}}^{\mathrm{UNI}}, \boldsymbol{A}_0)$. This lower bound is of the same order as the one for $d^2(\hat{\boldsymbol{A}}^{\mathrm{KLT}}, \boldsymbol{A}_0)$ provided in Theorem 6 of Koltchinskii et al. (2011). Comparing Theorem 4 with Corollary 1 we see that, under the i.i.d Gaussian noise ϵ_{ij}, the rate of convergence of estimator $\hat{\boldsymbol{A}}^{\mathrm{UNI}}$ is optimal in a minimax sense on the class of matrices that $\boldsymbol{\beta}_0 \in \beta(a_1)$ and $\boldsymbol{B}_0 \in \mathcal{B}(r_{\boldsymbol{B}_0}, a_1)$ up to a logarithmic factor $\log(n)$.

As for the non-uniform missingness, we can derive similar upper bound for $d^2(\hat{\boldsymbol{B}}, \boldsymbol{B}_0)$ and lower bound for $d^2(\hat{\boldsymbol{A}}, \boldsymbol{A}_0)$ under the knowledge of the true missing probabilities $\boldsymbol{\Theta}$. In this case, the non-asymptotic upper bound for $d^2(\hat{\boldsymbol{B}}, \boldsymbol{B}_0)$ enjoys different constant factors due to the condition $\lambda_2 \geq 2\|\boldsymbol{W} \circ \boldsymbol{\Theta}^* \circ \boldsymbol{Y} - \boldsymbol{A}_0\|$, while the lower bound is different by replacing θ_0 by θ_L. The details can be found in Section S2.3 of the supplementary material. If we plug in the general estimator $\hat{\boldsymbol{\Theta}}$ of $\boldsymbol{\Theta}$ in the upper bound, it is complicated to trace the constant factors. Instead, we have investigated the corresponding rates of convergence in the asymptotic regime of n_1, n_2 in Section 4.

6 Simulation study

This section reports results from simulation experiments which were designed to evaluate the numerical performance of the proposed estimator $\hat{\boldsymbol{A}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{B}}$ where $\hat{\boldsymbol{\beta}}$ is given by (3.4) and $\hat{\boldsymbol{B}}$ is given by (3.5). We also carried out comparative evaluation with four existing matrix completion method.

In the simulation, the target matrix $\boldsymbol{A}_0 = \boldsymbol{X}\boldsymbol{\beta}_0 + \boldsymbol{B}_0$ was randomly generated once and kept as fixed for each setting of (n_1, n_2, m, r). We generate $\boldsymbol{X} \in \mathbb{R}^{n_1 \times m}$, $\boldsymbol{\beta}_0 \in \mathbb{R}^{m \times n_2}$, $\boldsymbol{U}_0 \in \mathbb{R}^{n_1 \times r}$ and $\boldsymbol{V}_0 \in \mathbb{R}^{n_2 \times r}$ as random matrices with independent standard Gaussian entries independently and obtain $\boldsymbol{B}_0 = \boldsymbol{P}_{\boldsymbol{X}}^{\perp}\boldsymbol{U}_0\boldsymbol{V}_0^{\intercal}$. This ensures $\boldsymbol{B}_0 \in \mathcal{N}(\boldsymbol{X})$. Although we do not explicitly enforce that $\boldsymbol{A}_0$, $\boldsymbol{X}$ and $\boldsymbol{\beta}_0$ are of full rank, this happens with probability 1. The contaminated version of $\boldsymbol{A}_0$ was then generated as $\boldsymbol{Y} = \boldsymbol{A}_0 + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \in \mathbb{R}^{n_1 \times n_2}$ has i.i.d. mean zero Gaussian entries $\epsilon_{ij} \sim \mathcal{N}(0, \sigma_\epsilon^2)$. The σ_ϵ^2 is chosen such that the signal-to-noise ratio (SNR) is 1, namely $\mathrm{SNR} = \sqrt{\mathrm{Signal}(\boldsymbol{A}_0)/\sigma_\epsilon^2} = 1$, where $\mathrm{Signal}(\boldsymbol{A}_0) = \sum_{i=1}^{n_1}\sum_{j=1}^{n_2}(A_{0ij} - \bar{A}_0)^2/(n_1 n_2 - 1)$ and $\bar{A}_0 = \sum_{i=1}^{n_1}\sum_{j=1}^{n_2} A_{0ij}/(n_1 n_2)$.

The simulation was conducted under two sampling mechanisms: *MAR: missing-at-random* and *UNI: uniform observation*. For MAR, we adopted the logistic model (4.1) with $\boldsymbol{\gamma}_{\cdot j} = (\gamma_{1j}, \gamma_{2j}, \gamma_{3j}, \gamma_{4j}, 0, \ldots, 0)^{\intercal}_{1\times(m+1)}$. The entries $\gamma_{1j}, \gamma_{2j}, \gamma_{3j}$ and γ_{4j} were drawn independently according to $\gamma_{1j} \sim \mathcal{N}(-1.5, 0.1^2)$ and $\gamma_{kj} \sim \mathcal{N}(0.3, 0.1^2)$ for $k = 2, 3, 4$. Once generated, they were kept fixed throughout all MAR settings. For UNI, we set $\theta_{ij} = 0.2$, which is close to the average θ_{ij} under MAR, for all i, j. Throughout the study, we set $m = 20$ and $r = 10$, and chose $n_1 = n_2$ with four sizes: 400, 600, 800 and 1000, and the number of simulation for each (n_1, n_2) combination was 500.

The binary likelihood is used to estimate $\{\theta_{ij}\}$ via estimating $\gamma_{\cdot j}$ first under the MAR. See Section S3 of the supplementary material for more details on the MLEs.

Under the MAR, we implemented four versions of the proposed matrix completion approach: (i) the full SVT (full SVD followed by the singular value soft-thresholding and scaling procedures) with the tuning parameter α chosen by the 5-fold cross-validation (SVT-$\hat{\alpha}$-LOG); (ii) the approximate SVT ($\widehat{\text{SVT}}$) as described in Section 3.2 with the tuning parameter α chosen by the 5-fold cross-validation ($\widehat{\text{SVT}}$-$\hat{\alpha}$-LOG); (iii) the full SVT with $\alpha = 1$ (SVT-1-LOG); (iv) the approximate SVT with $\alpha = 1$ ($\widehat{\text{SVT}}$-1-LOG). We also experimented these four variates of the proposed matrix completion estimators under the UNI and denote them as SVT-$\hat{\alpha}$-UNI, $\widehat{\text{SVT}}$-$\hat{\alpha}$-UNI, SVT-1-UNI and $\widehat{\text{SVT}}$-1-UNI.

For the purpose of benchmarking, we compared with four existing matrix completion techniques: the methods proposed in Sun and Zhang (2012) (SZ), Negahban and Wainwright (2012) (NW), Koltchinskii et al. (2011) (KLT) and Mazumder et al. (2010) (MHT). Note that these methods were not designed to incorporate the covariate information $\boldsymbol{X}$, and therefore they only provided an estimate for $\boldsymbol{A}_0$. For SZ, the tuning parameter α was given by a formula in Sun and Zhang (2012) and λ were chosen by the 5-fold cross-validation. For the other three methods as well as the proposed method, the 5-fold cross-validation was used to select the tuning parameters.

To quantify the performance of the matrix completion, we used two empirical measures

$$\text{Test Error} = \frac{\left\|\boldsymbol{W}^{\star} \circ \left(\hat{\boldsymbol{A}} - \boldsymbol{A}_0\right)\right\|_F^2}{\|\boldsymbol{W}^{\star} \circ \boldsymbol{A}_0\|_F^2} \quad \text{and} \quad \text{RMSE}\left(\boldsymbol{A}_0\right) = \frac{\left\|\hat{\boldsymbol{A}} - \boldsymbol{A}_0\right\|_F}{\sqrt{n_1 n_2}},$$

where $\boldsymbol{W}^{\star}$ is the matrix of missing indicator with the (i, j)-th entry being $(1 - \omega_{ij})$. The test error

measures the relative estimation error of the unobserved entries to their signal strength. Moreover, the RMSE measure can be similarly defined for the proposed estimators of $\boldsymbol{\beta}_0$ and $\boldsymbol{B}_0$.

Tables 1 and 2 summarize the simulation results, with Table 1 for the MAR and Table 2 for the UNI probability of observation. The most visible aspect of the simulation results was that the four versions of the proposed methods had superior performance than the four existing methods by having smaller RMSEs and Test Errors. The proposed estimators with $\alpha = 1$, namely SVT-1-LOG and $\widehat{\text{SVT}}$-1-LOG, had more accurate rank estimates than the four existing methods in all cases. The two estimators SVT-$\hat{\alpha}$-LOG and $\widehat{\text{SVT}}$-$\hat{\alpha}$-LOG over-estimated the rank (the true rank was 30) when the sample sizes were relatively small under the logistic model, which may be viewed as a price paid for having better RMSEs and Test Errors than their counterparts with $\alpha = 1$. We note that $\alpha = 1$ meant that the penalty on the low-rank matrix $\boldsymbol{B}$ was entirely based on the nuclear norm. By inspecting the empirical values of $\hat{\alpha}$ from the simulations for the logistic model, we found $\hat{\alpha}$ appeared to converge to 1 as the sample sizes got larger. This explained why the aforementioned over-estimation in the ranks by SVT-$\hat{\alpha}$-LOG and $\widehat{\text{SVT}}$-$\hat{\alpha}$-LOG were reduced for the sample sizes of 800 and 1000. Another feature exhibited from the tables was that as the size of the matrix n_1 and n_2 increased, both the RMSEs and Test Errors of the proposed methods got smaller. This was also the case for the four existing methods under the logistic model in Table 1. The latter was likely due to the reduction of the variance owing to having more "data" despite employing a misspecified model. In contrast, the reason for the proposed methods' having smaller RMSEs and Test Errors was due to their ability to reduce both the bias and the variance in the completed matrices as the methods are consistent as shown in the theoretical analyses in Section 4.

Comparing the results in Table 1 with those in Table 2, it was clear that the presence of the heterogeneity in the observation probability made the matrix completion more difficult as reflected by Table 1 having larger RMSEs and Test Errors. This comparison was fair as the overall observed rate under the logistic model was close to 0.2, the rate under the UNI. As the true rank in all settings was 30, It appeared that the estimated ranks were the most affected by the heterogeneity.

However, despite the heterogeneity, the proposed methods tended to produce more accurate (and smaller) ranks than the four existing methods.

The simulation results reported in Tables 1 and 2 consistently showed that the full SVT and the approximate SVT gave very close results, which confirmed that the approximate SVT can achieve computational reduction without sacrificing much accuracy. Under the MAR setting (Table 1), the proposed methods with the tuning parameter α chosen by the 5-fold cross validation produced completed matrices with larger ranks but smaller RMSEs than their counterparts with $\alpha = 1$, which confirmed an early remark made in Section 3 regarding the role of α in balancing between the nuclear and the Frobenius norms in the regularization of the low rank matrix $\boldsymbol{B}$. With the dimensions n_1 and n_2 growing, the chosen α approached 1 which led to more compatible rank estimates and the RMSEs between the two approaches of choosing α.

Furthermore, we conducted an additional simulation study where the covariates are not useful (i.e. $\boldsymbol{A}_0 = \boldsymbol{B}_0$). Table S1 in the supplementary material summarizes the corresponding simulation results under uniform probability of observation. The simulation results indicated that the two versions of the proposed methods had slightly inferior performance than the four existing methods by having larger RMSEs and test errors. This is expected since the existing methods assume no covariates, which matches with the underlying model. Although $\boldsymbol{\beta}_0 = \mathbf{0}$ is allowed in the model of the proposed methods, the proposed methods lose efficiency by considering a more general model.

7 Empirical study

We demonstrate the proposed methodology by analyzing the MovieLens 100K data set as described in Harper and Konstan (2016). This data set includes 100,000 movie ratings, ranging from 1 to 5, appraised by 943 viewers on 1682 movies, where each viewer had rated at least 20 movies. The data came with additional information on both viewers and movies. In this analysis, we adopted age and gender as the covariates for our proposed method. For evaluation purpose, the data provider split the 100,000 ratings into a training set with 90,570 ratings and a test set with 9,430 ratings,

Table 1: Empirical root mean square errors (RMSEs), test errors, estimated ranks and their standard errors (in parentheses) under model $\boldsymbol{A}_0 = \boldsymbol{X}\boldsymbol{\beta}_0 + \boldsymbol{B}_0$ and the logistic missing-at-random model (MAR), with (n_1, n_2)=(400,400), (600,600), (800,800), (1000,1000), $m = 20$, and $r = 10$, for four versions of the proposed methods, and the four existing methods (SZ, NW, KLT and MHT).

$n_1 = n_2 = 400$	RMSE($\boldsymbol{\beta}_0$)	RMSE($\boldsymbol{B}_0$)	RMSE($\boldsymbol{A}_0$)	Test error	Rank
SVT-$\hat{\alpha}$-LOG	0.6938 (0.0059)	3.1099 (0.0504)	4.4007 (0.0469)	0.6658 (0.0054)	117.27 (26.55)
SVT-1-LOG	0.6964 (0.0059)	3.1778 (0.1419)	4.4581 (0.1100)	0.6759 (0.0059)	24.55 (3.35)
$\widehat{\text{SVT}}$-$\hat{\alpha}$-LOG	0.6939 (0.0059)	3.1063 (0.0503)	4.3985 (0.0469)	0.6658 (0.0054)	111.96 (21.88)
$\widehat{\text{SVT}}$-1-LOG	0.6964 (0.0059)	3.1778 (0.1419)	4.4581 (0.1100)	0.6759 (0.0059)	24.55 (3.35)
SZ			4.8593 (0.0232)	0.8627 (0.0054)	49.76 (3.04)
NW			4.8340 (0.0221)	0.8565 (0.0056)	102.46 (5.34)
KLT			4.9789 (0.0214)	0.8869 (0.0055)	34.55 (2.12)
MHT			4.8507 (0.0234)	0.8595 (0.0056)	50.05 (2.72)
$n_1 = n_2 = 600$	RMSE($\boldsymbol{\beta}_0$)	RMSE($\boldsymbol{B}_0$)	RMSE($\boldsymbol{A}_0$)	Test error	Rank
SVT-$\hat{\alpha}$-LOG	0.6227 (0.0043)	3.1239 (0.0416)	4.1704 (0.0379)	0.5749 (0.0039)	124.97 (17.11)
SVT-1-LOG	0.6237 (0.0041)	3.2491 (0.1484)	4.2686 (0.1203)	0.5834 (0.0055)	50.15 (3.93)
$\widehat{\text{SVT}}$-$\hat{\alpha}$-LOG	0.6230 (0.0043)	3.1162 (0.0412)	4.1653 (0.0375)	0.5752 (0.0040)	113.57 (12.63)
$\widehat{\text{SVT}}$-1-LOG	0.6237 (0.0041)	3.2476 (0.1475)	4.2675 (0.1195)	0.5835 (0.0055)	49.67 (4.03)
SZ			4.5510 (0.0195)	0.7438 (0.0050)	80.71 (3.77)
NW			4.4681 (0.0182)	0.7186 (0.0051)	170.32 (6.03)
KLT			4.7097 (0.0143)	0.7821 (0.0041)	60.00 (1.59)
MHT			4.5201 (0.0191)	0.7341 (0.0051)	83.26 (3.29)
$n_1 = n_2 = 800$	RMSE($\boldsymbol{\beta}_0$)	RMSE($\boldsymbol{B}_0$)	RMSE($\boldsymbol{A}_0$)	Test error	Rank
SVT-$\hat{\alpha}$-LOG	0.5661 (0.0033)	3.0785 (0.0343)	3.9787 (0.0300)	0.5146 (0.0037)	101.03 (10.43)
SVT-1-LOG	0.5664 (0.0032)	3.1118 (0.0673)	4.0055 (0.0555)	0.5148 (0.0044)	69.41 (2.06)
$\widehat{\text{SVT}}$-$\hat{\alpha}$-LOG	0.5663 (0.0032)	3.0716 (0.0334)	3.9739 (0.0295)	0.5154 (0.0037)	93.00 (8.11)
$\widehat{\text{SVT}}$-1-LOG	0.5665 (0.0031)	3.1094 (0.0669)	4.0037 (0.0552)	0.5154 (0.0044)	66.94 (2.15)
SZ			4.3308 (0.0128)	0.6636 (0.0035)	103.45 (3.36)
NW			4.2144 (0.0142)	0.6284 (0.0039)	222.56 (7.28)
KLT			4.5276 (0.0111)	0.7132 (0.0031)	78.13 (1.55)
MHT			4.2855 (0.0147)	0.6498 (0.0038)	108.63 (4.71)
$n_1 = n_2 = 1000$	RMSE($\boldsymbol{\beta}_0$)	RMSE($\boldsymbol{B}_0$)	RMSE($\boldsymbol{A}_0$)	Test error	Rank
SVT-$\hat{\alpha}$-LOG	0.5109 (0.0027)	2.9337 (0.0461)	3.7107 (0.0388)	0.4601 (0.0037)	87.47 (2.03)
SVT-1-LOG	0.5109 (0.0027)	2.9336 (0.0459)	3.7106 (0.0387)	0.4601 (0.0037)	87.36 (1.88)
$\widehat{\text{SVT}}$-$\hat{\alpha}$-LOG	0.5112 (0.0026)	2.9272 (0.0458)	3.7062 (0.0385)	0.4613 (0.0037)	80.20 (1.65)
$\widehat{\text{SVT}}$-1-LOG	0.5111 (0.0026)	2.9281 (0.0460)	3.7068 (0.0387)	0.4611 (0.0037)	81.14 (2.30)
SZ			4.0069 (0.0151)	0.5897 (0.0036)	122.87 (7.36)
NW			3.8522 (0.0119)	0.5439 (0.0031)	270.96 (9.54)
KLT			4.2491 (0.0092)	0.6500 (0.0026)	91.56 (1.40)
MHT			3.9447 (0.0122)	0.5716 (0.0032)	136.57 (5.27)

Table 2: Empirical root mean square errors (RMSEs), test errors, estimated ranks and their standard errors (in parentheses) under model $\boldsymbol{A}_0 = \boldsymbol{X}\boldsymbol{\beta}_0 + \boldsymbol{B}_0$ and the uniform observation mechanism (UNI), with (n_1, n_2)=(400,400), (600,600), (800,800), (1000,1000) $m = 20$, and $r = 10$, for four versions of the proposed methods, and the four existing methods (SZ, NW, KLT and MHT).

$n_1 = n_2 = 400$	RMSE($\boldsymbol{\beta}_0$)	RMSE($\boldsymbol{B}_0$)	RMSE($\boldsymbol{A}_0$)	Test error	Rank
SVT-$\hat{\alpha}$-UNI	0.6343 (0.0050)	2.8815 (0.0181)	4.0473 (0.0200)	0.5898 (0.0053)	42.86 (3.47)
SVT-1-UNI	0.6344 (0.0051)	2.8804 (0.0177)	4.0466 (0.0201)	0.5896 (0.0053)	42.22 (2.13)
$\widehat{\text{SVT}}$-$\hat{\alpha}$-UNI	0.6343 (0.0050)	2.8816 (0.0181)	4.0474 (0.0200)	0.5898 (0.0054)	42.78 (3.45)
$\widehat{\text{SVT}}$-1-UNI	0.6344 (0.0051)	2.8805 (0.0177)	4.0467 (0.0202)	0.5896 (0.0053)	42.18 (2.13)
SZ			4.8318 (0.0251)	0.8528 (0.0060)	52.54 (3.12)
NW			4.8293 (0.0259)	0.8493 (0.0064)	97.47 (5.29)
KLT			4.8994 (0.0217)	0.8721 (0.0052)	45.42 (2.38)
MHT			4.8238 (0.0252)	0.8492 (0.0062)	51.27 (2.75)
$n_1 = n_2 = 600$	RMSE($\boldsymbol{\beta}_0$)	RMSE($\boldsymbol{B}_0$)	RMSE($\boldsymbol{A}_0$)	Test error	Rank
SVT-$\hat{\alpha}$-UNI	0.5711 (0.0037)	2.7570 (0.0136)	3.7423 (0.0145)	0.4893 (0.0035)	58.17 (1.75)
SVT-1-UNI	0.5711 (0.0037)	2.7571 (0.0136)	3.7424 (0.0145)	0.4893 (0.0035)	58.12 (1.75)
$\widehat{\text{SVT}}$-$\hat{\alpha}$-UNI	0.5711 (0.0037)	2.7566 (0.0138)	3.7420 (0.0146)	0.4892 (0.0035)	57.04 (1.64)
$\widehat{\text{SVT}}$-1-UNI	0.5711 (0.0037)	2.7568 (0.0137)	3.7421 (0.0146)	0.4892 (0.0035)	57.51 (1.72)
SZ			4.5228 (0.0176)	0.7322 (0.0047)	84.41 (3.07)
NW			4.4838 (0.0201)	0.7181 (0.0052)	160.25 (6.91)
KLT			4.6427 (0.0147)	0.7700 (0.0040)	74.71 (1.89)
MHT			4.4895 (0.0175)	0.7212 (0.0048)	84.30 (2.67)
$n_1 = n_2 = 800$	RMSE($\boldsymbol{\beta}_0$)	RMSE($\boldsymbol{B}_0$)	RMSE($\boldsymbol{A}_0$)	Test error	Rank
SVT-$\hat{\alpha}$-UNI	0.5155 (0.0028)	2.6277 (0.0117)	3.4884 (0.0119)	0.4188 (0.0027)	71.39 (1.53)
SVT-1-UNI	0.5155 (0.0028)	2.6278 (0.0117)	3.4884 (0.0119)	0.4188 (0.0027)	71.34 (1.51)
$\widehat{\text{SVT}}$-$\hat{\alpha}$-UNI	0.5155 (0.0028)	2.6240 (0.0120)	3.4856 (0.0120)	0.4180 (0.0027)	68.25 (1.35)
$\widehat{\text{SVT}}$-1-UNI	0.5155 (0.0028)	2.6247 (0.0119)	3.4861 (0.0120)	0.4181 (0.0027)	69.01 (1.62)
SZ			4.2348 (0.0128)	0.6329 (0.0036)	109.41 (2.41)
NW			4.1667 (0.0135)	0.6115 (0.0038)	214.47 (4.28)
KLT			4.4071 (0.0117)	0.6872 (0.0032)	98.45 (1.67)
MHT			4.1837 (0.0138)	0.6171 (0.0038)	111.15 (3.85)
$n_1 = n_2 = 1000$	RMSE($\boldsymbol{\beta}_0$)	RMSE($\boldsymbol{B}_0$)	RMSE($\boldsymbol{A}_0$)	Test error	Rank
SVT-$\hat{\alpha}$-UNI	0.4646 (0.0022)	2.4614 (0.0106)	3.2128 (0.0097)	0.3683 (0.0021)	82.59 (1.49)
SVT-1-UNI	0.4646 (0.0022)	2.4614 (0.0106)	3.2128 (0.0097)	0.3683 (0.0021)	82.59 (1.47)
$\widehat{\text{SVT}}$-$\hat{\alpha}$-UNI	0.4646 (0.0022)	2.4517 (0.0110)	3.2054 (0.0099)	0.3664 (0.0022)	77.11 (1.28)
$\widehat{\text{SVT}}$-1-UNI	0.4646 (0.0022)	2.4528 (0.0109)	3.2063 (0.0099)	0.3666 (0.0022)	77.94 (1.78)
SZ			3.8886 (0.0105)	0.5524 (0.0029)	129.51 (2.50)
NW			3.8064 (0.0109)	0.5278 (0.0029)	257.67 (5.05)
KLT			4.1026 (0.0099)	0.6189 (0.0027)	117.78 (1.63)
MHT			3.8277 (0.0111)	0.5342 (0.0030)	132.35 (3.62)

such that there were exactly 10 ratings per viewer in the test set. Two versions of such splitting are provided, which are referred to as Split1=(Training Set1, Test Set1) and Split2=(Training Set2, Test Set2), respectively. Further, we know that Test Set1 and Test Set2 are disjoint. In our experiment, we applied those methods as described in Section 6 to the training sets and evaluated the test errors based on the corresponding test sets. As common pre-processing steps, we removed the movies with no ratings in training sets, and applied the bi-scaling procedure (Mazumder et al., 2010) which standardizes a matrix to have row and column means zero and variances one, before applying any matrix completion methods.

To construct the covariate matrix $\boldsymbol{X}$, gender was encoded as "0" for male and "1" for female. Age was given as a numerical variable and used directly. Thus the covariate matrix $\boldsymbol{X}$ (viewers' demographic) was of dimension 943×2. As a standard procedure, every column of $\boldsymbol{X}$ was normalized to avoid any scaling issues in the penalties.

Next, we focus on the probabilities of observation $\{\theta_{ij}\}$. Our preliminary analysis suggested a non-monotone trend of observed rates with respect to age. To see this, we divide age into 7 categories: under 18, $18-24$, $25-34$, $35-44$, $45-49$, $50-55$ and $56+$, which are denoted by A1, A2, ..., A7, respectively. These age categories were suggested by the document accompanying with the data set (`http://files.grouplens.org/datasets/movielens/ml-1m-README.txt`). The non-monotonicity is demonstrated in Figure 1(a), which showed that the rate of observation peaked at the age group of $18-24$, continued to decline till the $45-49$ age group and then had a slight increase afterward. This indicated a strong age effect on the probability of observation. To gauge the gender effect, we split each age group into two sub-groups of male and female. This gave rise to 14 age and gender combinations which are denoted by MA1, FA1, ...,FA7. As shown in Figure 1(b), the sample observed rates varied across different viewer groups as determined by age and gender. Of interest was that female had higher rates of observation than their male counterparts for all age groups, which suggested the existence of the gender effect.

To reduce the number of parameters in the probability of observation, we explored the possibility

of merging some age-gender categories. However, it was computationally expensive to examine all possible merging combinations. In our analysis, a simple data-driven screening method was conducted. We took the uniform probability of observation model as the benchmark model, denoted as Benchmark, and considered 14 models for the observational probability that had exactly one of the 14 age-gender categories separated out to have its own individual rate of observation, once at a time, while the rest of the 13 categories was estimated by a common rate of observation. Then we applied our matrix completion procedure SVT-$\hat{\alpha}$-LOG and recorded the empirical validation error. For all the 14 models and the benchmark model, by applying similar procedure, we obtained the corresponding validation errors $Q_{\mathsf{MA1}}, \ldots, Q_{\mathsf{FA7}}, Q_{\mathsf{Benchmark}}$ shown in Table 3. If the validation error of a model was smaller than $Q_{\mathsf{Benchmark}}$, the corresponding group was marked as required individual modeling and should be separated out from the rest.

For Split1, seven groups (FA1, MA3, FA3, FA4, FA5, FA6, and FA7) were classified as that individual modeling was needed. For these seven groups, the corresponding sample proportions of observation were used as the estimates for their respective observation probabilities. The remaining seven groups were assumed to share a same observation probability, which was estimated by the pooled sample proportions of observation. Denote this final model for Split1 by Final1. As shown in Table 3, we note that the corresponding validation error $Q_{\mathsf{Final1}} = 4.4297$ was the smallest among all the evaluated models for Split1. This provided some validity of this final choice. For Split2, we identified seven groups (FA1, MA2, MA3, FA3, FA5, FA6 and MA7) and the corresponding final model Final2 also attained the smallest validation error $Q_{\mathsf{Final2}} = 4.4230$ among all the evaluated models. Since the proposed methods require only one SVD for each sampling probability model, we can perform this additional exploration of the sampling mechanism while keeping the computational costs significantly lower than most of the competitors.

Table 4 reports the root mean square prediction errors (RMSPEs) and estimated ranks of different estimators for both Split1 and Split2, where $\text{RMSPE} = \|\boldsymbol{W}^{test} \circ (\hat{\boldsymbol{A}} - \boldsymbol{Y})\|_F / \sqrt{\sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \omega_{ij}^{test}}$, where $\boldsymbol{W}^{test}$ is the indicator matrix of test set with the (i, j)-th entry being ω_{ij}^{test}. Since Test Set1

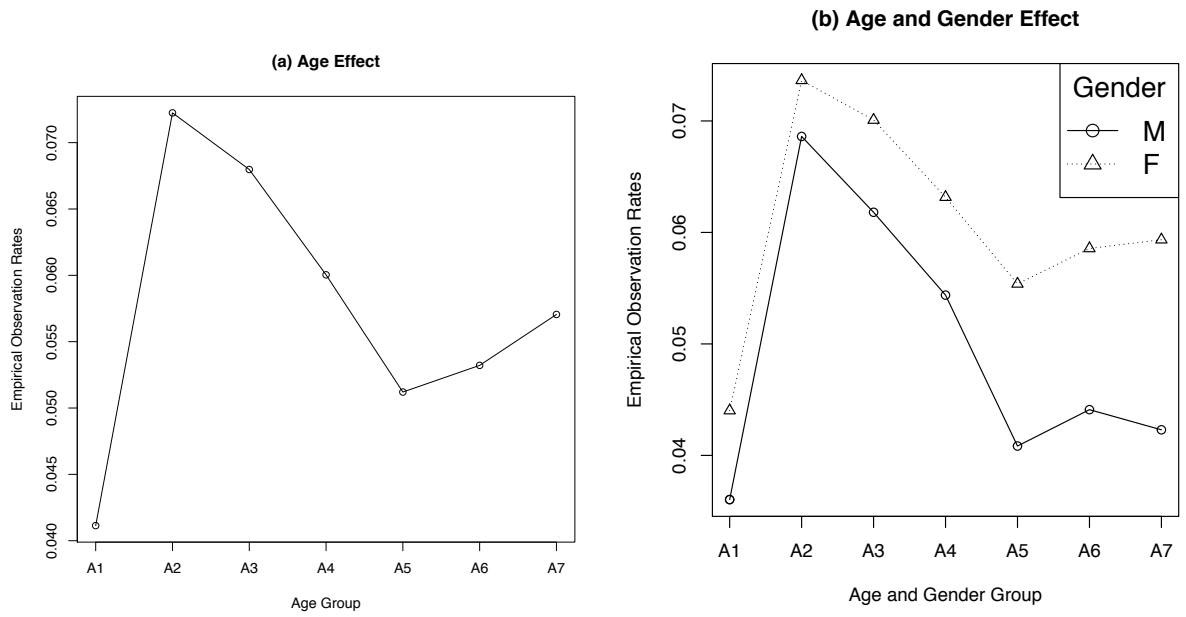

Figure 1: Empirical observation rates of the MovieLens 100K data. Panel (a): with respect to the seven age groups; Panel (b): with respect to the 14 combination groups of age and gender.

Table 3: Empirical validation errors Q under the 14 models,the Benchmark and the final selected models (Final), where $*$ and † denotes the age-gender combination that requires individual modeling for Split1 and Split2 respectively.

Model	MA1	FA1	MA2	FA2	MA3	FA3	MA4	FA4
Split1	4.4342	4.4310*	4.4319	4.4346	4.4317*	4.4307*	4.4322	4.4317*
Split2	4.4279	4.4235†	4.4239†	4.4269	4.4240†	4.4237†	4.4247	4.4240
Model	MA5	FA5	MA6	FA6	MA7	FA7	Benchmark	Final
Split1	4.4338	4.4317*	4.4335	4.4313*	4.4318	4.4317*	4.4317	4.4297
Split2	4.4263	4.4239†	4.4260	4.4236†	4.4239†	4.4240	4.4240	4.4230

and Test Set2, the corresponding test sets of Split1 and Split2, were disjoint and of the same size, it is fair to calculate the overall RMSPEs for evaluation of different methods. Similarly as the simulation results reported in the previous section, SVT-$\hat{\alpha}$-LOG and $\widehat{\text{SVT}}$-$\hat{\alpha}$-LOG produced highly comparable results, which indicated the applicability of $\widehat{\text{SVT}}$-$\hat{\alpha}$-LOG to larger data sets whenever computational resources are scarce. In both Split1 and Split2, the proposed methods outperformed NW, KLT and MHT in terms of smaller RMSPEs and either smaller or more reasonable rank estimation. Although the proposed methods were slightly inferior to SZ in Split1, they outperformed SZ significantly in Split2 by having smaller RMSPEs. Among the six matrix completion methods

considered, the two proposed methods and the KLT method offered the most consistent results between Split1 and Split2, while the other three methods exhibited much larger variations, especially in the estimated ranks. That KLT method gave rank 1 estimates was likely due to its ignoring the heterogeneity in the probability of observation, which amplified the difference between the largest and the rest of the eigenvalues. As a result, $(n_1 n_2/N)\sigma_1 \boldsymbol{u}_1 \boldsymbol{v}_1^\intercal$ explained most of the target matrix $\boldsymbol{A}_0$, leading to the rank-1 estimates in Table 4. Overall speaking, the two proposed methods were among the top two performers of the analysis reported in Table 4.

As suggested by an anonymous referee, we experimented treating the age as categorical variables with the number of categories ranging from three to seven. Corresponding details are given in Section S8 of the supplementary material. As reported, the prediction errors of using the four and five age categories were the best among the five categories. However, they were still inferior to the method of treating the age as a continues variable as shown in Table S2 of Section S8. This was likely due to an increase in the rank of $\boldsymbol{X}$ as a result of the age categorization. Nevertheless, we note that using the categorical age with four or five groups produced better results than the typical matrix completion without utilizing covariate information.

Table 4: Root mean square prediction errors (RMSPEs) and ranks of the completed matrix based on Split1 and Split2 for the two versions of the proposed method (SVT-$\hat{\alpha}$-LOG) and ($\widehat{\text{SVT}}$-$\hat{\alpha}$-LOG) and the four existing methods proposed respectively in Sun and Zhang (2012)(SZ), Negahban and Wainwright (2012)(NW), Koltchinskii et al. (2011)(KLT) and Mazumder et al. (2010)(MHT).

	Split1		Split2		Overall
	RMSPE	Rank	RMSPE	Rank	RMSPE
SVT-$\hat{\alpha}$-LOG	0.9415	47	0.9541	46	0.9478
$\widehat{\text{SVT}}$-$\hat{\alpha}$-LOG	0.9418	45	0.9542	43	0.9480
SZ	0.9412	39	0.9563	31	0.9488
NW	0.9421	269	0.9589	289	0.9506
KLT	0.9584	1	0.9688	1	0.9636
MHT	0.9414	56	0.9568	46	0.9491

8 Concluding remarks

This paper investigates the problem of matrix completion with covariate information. We have shown that utilizing such information can lead to more accurate completed matrix and more interpretable results. When the matrix entries are heterogeneously observed due to selection bias of covariates, this heterogeneity should be taken into account. Our real data analysis on the MovieLens 100K data revealed the existence of the heterogeneity by the age and the gender of the movie viewers. The heterogeneity, without proper treatment, can render the consistency of the existing matrix completion methods. Under a column-space-decomposition model, we propose a matrix completion procedure that adjusts for the heterogeneity in the observation mechanism by taking into account the covariate effect. The proposed matrix completion estimator can be coupled with the fast randomized singular value thresholding (FRSVT) procedure to achieve improved computational efficiency for high dimensional matrices. A general convergence of the matrix completion procedure is provided (Theorem 1), and specific convergence rates under two popular models for the probability of observation are also given. The column-space-decomposition model provides an interpretive coefficient matrix that can quantify the effect of the covariates. Empirical studies show the attractive performance of the proposed methods as compared with existing matrix completion methods in terms of the root mean square prediction errors and the ranks of completed matrices.

Acknowledgment

The authors are most grateful to the reviewers and the associate editor for their constructive comments which led to a much improved version of the paper.

References

Abernethy, J., Bach, F., Evgeniou, T., and Vert, J.-P. (2009), "A New Approach to Collaborative Filtering: Operator Estimation with Spectral Regularization," *Journal of Machine Learning Research*, 10, 803–826.

Bi, X., Qu, A., Wang, J., and Shen, X. (2016), "A Group-Specific Recommender System," *Journal of the American Statistical Association*.

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

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.

Chiang, K.-Y., Hsieh, C.-J., and Dhillon, I. S. (2015), "Matrix Completion with Noisy Side Information," in *Advances in Neural Information Processing Systems*, pp. 3429–3437.

Feuerverger, A., He, Y., and Khatri, S. (2012), "Statistical Significance of the Netflix Challenge," *Statistical Science*, 27, 202–231.

Freedman, D. A. (2009), *Statistical Models: Theory and Practice*, New York: Cambridge University Press.

Friedman, J., Hastie, T., and Tibshirani, R. (2013), *The Elements of Statistical Learning*, New York: Springer.

Gross, D. (2011), "Recovering Low-Rank Matrices from Few Coefficients in any Basis," *IEEE Transactions on Information Theory*, 57, 1548–1566.

Halko, N., Martinsson, P.-G., and Tropp, J. A. (2011), "Finding Structure with Randomness: Probabilistic Algorithms for Constructing Approximate Matrix Decompositions," *SIAM review*, 53, 217–288.

Harper, F. M. and Konstan, J. A. (2016), "The MovieLens Datasets: History and Context," *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 5, 19:1–19:19.

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.

Little, R. J. and Rubin, D. B. (2014), *Statistical Analysis with Missing Data*, New Jersey: John Wiley & Sons.

Ma, S., Goldfarb, D., and Chen, L. (2011), "Fixed Point and Bregman Iterative Methods for Matrix Rank Minimization," *Mathematical Programming*, 128, 321–353.

Mardia, K. V., Kent, J. T., and Bibby, J. M. (1980), *Multivariate Analysis*, London: Academic 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.

Natarajan, N. and Dhillon, I. S. (2014), "Inductive Matrix Completion for Predicting Gene–Disease Associations," *Bioinformatics*, 30, i60–i68.

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.

Oh, T.-H., Matsushita, Y., Tai, Y.-W., and Kweon, I. S. (2015), "Fast Randomized Singular Value Thresholding for Nuclear Norm Minimization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4484–4493.

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.

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.

Sun, T. and Zhang, C.-H. (2012), "Calibrated Elastic Regularization in Matrix Completion," in *Advances in Neural Information Processing Systems*, pp. 863–871.

Sweeting, T. (1980), "Uniform Asymptotic Normality of the Maximum Likelihood Estimator," *The Annals of Statistics*, 8, 1375–1381.

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., and Altman, R. B. (2001), "Missing Value Estimation Methods for DNA Microarrays," *Bioinformatics*, 17, 520–525.

Xu, M., Jin, R., and Zhou, Z.-H. (2013), "Speedup Matrix Completion with Side Information: Application to Multi-Label Learning," in *Advances in Neural Information Processing Systems*, pp. 2301–2309.

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

Zou, H. and Hastie, T. (2005), "Regularization and Variable Selection via the Elastic Net," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67, 301–320.

Supplemental Document for "Matrix Completion with Covariate Information"

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

September 24, 2017

Abstract

This document provides supplementary material to the article "Matrix Completion with Covariate Information" written by the same authors.

S1 Proof of Propositions

Proof of Proposition 1. We have

$$\begin{aligned}
&\mathsf{E}\,\|\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{B}-\boldsymbol{W}\circ\boldsymbol{\Theta}^*\circ\boldsymbol{Y}\|_F^2\\
=&\|\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{B}\|_F^2-2\left\langle\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{B},\mathsf{E}\left(\boldsymbol{W}\circ\boldsymbol{\Theta}^*\circ\boldsymbol{Y}\right)\right\rangle+\mathsf{E}\,\|\boldsymbol{W}\circ\boldsymbol{\Theta}^*\circ\boldsymbol{Y}\|_F^2\\
=&\|(\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{B})-(\boldsymbol{X}\boldsymbol{\beta}_0+\boldsymbol{B}_0)\|_F^2-\|\boldsymbol{X}\boldsymbol{\beta}_0+\boldsymbol{B}_0\|_F^2+\sum_{ij}\frac{\left((X\beta_0)_{ij}+B_{0ij}\right)^2+\sigma_{ij}^2}{\theta_{ij}},
\end{aligned}$$

due to Conditions C1(a) and C4. For any minimizer $(\boldsymbol{\beta}_s,\boldsymbol{B}_s)$ of R, we have $\boldsymbol{X}\boldsymbol{\beta}_0+\boldsymbol{B}_0=\boldsymbol{X}\boldsymbol{\beta}_s+\boldsymbol{B}_s$, which implies $\boldsymbol{X}(\boldsymbol{\beta}_0-\boldsymbol{\beta}_s)=\boldsymbol{B}_s-\boldsymbol{B}_0$. Since $\boldsymbol{B}_s-\boldsymbol{B}_0\in\mathcal{N}(\boldsymbol{X})$, we can conclude both $\boldsymbol{X}\boldsymbol{\beta}_s=\boldsymbol{X}\boldsymbol{\beta}_0$

*Xiaojun Mao is Ph.D. candidate, Department of Statistics, Iowa State University, Ames, IA 50011, USA (Email: mxjki@iastate.edu).

†Author of Correspondence. 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). His research is partially supported by Chinas National Key Research Special Program Grants 2016YFC0207701 and 2015CB856000, and National Natural Science Foundation of China grants 11131002, 71532001 and 71371016.

‡Raymond K. W. Wong is Assistant Professor, Department of Statistics, Texas A&M University, College Station, TX 77843, USA (Email: raywong@stat.tamu.edu). His research is partially supported by the National Science Foundation under Grants DMS-1612985 and DMS-1711952 (subcontract).

and $\boldsymbol{B}_s = \boldsymbol{B}_0$. As matrix $\boldsymbol{X}^\intercal \boldsymbol{X}$ is invertible, we know that $\boldsymbol{\beta}_s = \boldsymbol{\beta}_0$. This also implies that $(\boldsymbol{\beta}_0, \boldsymbol{B}_0)$ is the unique minimizer. □

Proof of Proposition 2. By operator inequality and matrix $\boldsymbol{X}^\intercal \boldsymbol{X}$ is invertible, we have $\|\boldsymbol{P}_{\boldsymbol{X}}^{\perp} \boldsymbol{B}\|_* \leq \|\boldsymbol{P}_{\boldsymbol{X}}^{\perp}\| \|\boldsymbol{B}\|_* \leq \|\boldsymbol{B}\|_*$. For any $\boldsymbol{B} \in \mathbb{R}^{n_1 \times n_2}$,

$$
\begin{aligned}
&\frac{1}{n_1 n_2} \left\| \boldsymbol{P}_{\boldsymbol{X}}^{\perp} \boldsymbol{B} - \boldsymbol{P}_{\boldsymbol{X}}^{\perp} \left(\boldsymbol{W} \circ \hat{\boldsymbol{\Theta}}^* \circ \boldsymbol{Y} \right) \right\|_F^2 + \lambda_2 \left(\alpha \left\| \boldsymbol{P}_{\boldsymbol{X}}^{\perp} \boldsymbol{B} \right\|_* + (1-\alpha) \left\| \boldsymbol{P}_{\boldsymbol{X}}^{\perp} \boldsymbol{B} \right\|_F^2 \right) \\
\leq & \frac{1}{n_1 n_2} \left\| \boldsymbol{P}_{\boldsymbol{X}}^{\perp} \boldsymbol{B} - \boldsymbol{P}_{\boldsymbol{X}}^{\perp} \left(\boldsymbol{W} \circ \hat{\boldsymbol{\Theta}}^* \circ \boldsymbol{Y} \right) \right\|_F^2 + \frac{1}{n_1 n_2} \|\boldsymbol{P}_{\boldsymbol{X}} \boldsymbol{B}\|_F^2 + \lambda_2 \left(\alpha \left\| \boldsymbol{P}_{\boldsymbol{X}}^{\perp} \boldsymbol{B} \right\|_* + (1-\alpha) \left\| \boldsymbol{P}_{\boldsymbol{X}}^{\perp} \boldsymbol{B} \right\|_F^2 \right) \\
&+ \lambda_2 (1-\alpha) \|\boldsymbol{P}_{\boldsymbol{X}} \boldsymbol{B}\|_F^2 \\
\leq & \frac{1}{n_1 n_2} \left\| \boldsymbol{B} - \boldsymbol{P}_{\boldsymbol{X}}^{\perp} \left(\boldsymbol{W} \circ \hat{\boldsymbol{\Theta}}^* \circ \boldsymbol{Y} \right) \right\|_F^2 + \lambda_2 \left(\alpha \|\boldsymbol{B}\|_* + (1-\alpha) \|\boldsymbol{B}\|_F^2 \right),
\end{aligned}
$$

where the first inequality is strict whenever $\boldsymbol{P}_{\boldsymbol{X}} \boldsymbol{B} \neq \boldsymbol{0}$. Therefore the solution of (3.7) belongs to $\mathcal{N}(\boldsymbol{X})$ and hence it is also a solution of (3.5). □

S2 Benefit of Covariate Information

Before discussing the benefit of using covariates, we need the following proposition which describes the relationship between $\|\boldsymbol{A}_0\|_*$ and $\|\boldsymbol{B}_0\|_*$.

Proposition S2.1. *Let $\boldsymbol{A}_0 = \boldsymbol{X}\boldsymbol{\beta}_0 + \boldsymbol{B}_0$, where $\boldsymbol{B}_0 \in \mathcal{N}(\boldsymbol{X})$, we have $\|\boldsymbol{B}_0\|_* \leq \|\boldsymbol{A}_0\|_*$. If $\mathcal{R}(\boldsymbol{\beta}_0) \nsubseteq \mathcal{R}(\boldsymbol{B}_0)$, once $\boldsymbol{\beta}_0 \neq \boldsymbol{0}^{m \times n_2}$, we have $\|\boldsymbol{B}_0\|_* < \|\boldsymbol{A}_0\|_*$. Here $\mathcal{R}(\boldsymbol{Y})$ is the row space of a matrix $\boldsymbol{Y}$.*

Proof. For any $\boldsymbol{Z} \in \partial \|\boldsymbol{B}_0\|_*$, we have $\|\boldsymbol{A}_0\|_* \geq \|\boldsymbol{B}_0\|_* + \langle \boldsymbol{Z}, \boldsymbol{X}\boldsymbol{\beta}_0 \rangle$. Write the SVD of $\boldsymbol{B}_0$ as $\sum_{i=1}^{r_{\boldsymbol{B}_0}} \sigma_i(\boldsymbol{B}_0) \boldsymbol{u}_{\boldsymbol{B}_0}^{(i)} \boldsymbol{v}_{\boldsymbol{B}_0}^{(i)T}$. Let $\mathcal{B}_{\boldsymbol{u}}$ be the linear span of $\boldsymbol{u}_{\boldsymbol{B}_0}^{(1)}, \ldots, \boldsymbol{u}_{\boldsymbol{B}_0}^{(r_{\boldsymbol{B}_0})}$ and $\mathcal{B}_{\boldsymbol{v}}$ be the linear span of $\boldsymbol{v}_{\boldsymbol{B}_0}^{(1)}, \ldots, \boldsymbol{v}_{\boldsymbol{B}_0}^{(r_{\boldsymbol{B}_0})}$. We have the fact that the sub-differential of the convex function $\boldsymbol{B}_0 \mapsto \|\boldsymbol{B}_0\|_*$ is the following set of matrices:

$$
\partial \|\boldsymbol{B}_0\|_* = \left\{ \sum_{i=1}^{r_{\boldsymbol{B}_0}} \boldsymbol{u}_{\boldsymbol{B}_0}^{(i)} \boldsymbol{v}_{\boldsymbol{B}_0}^{(i)T} + \boldsymbol{P}_{\mathcal{B}_{\boldsymbol{u}}^{\perp}} \boldsymbol{Z} \boldsymbol{P}_{\mathcal{B}_{\boldsymbol{v}}^{\perp}} : \left\| \boldsymbol{P}_{\mathcal{B}_{\boldsymbol{u}}^{\perp}} \boldsymbol{Z} \boldsymbol{P}_{\mathcal{B}_{\boldsymbol{v}}^{\perp}} \right\| \leq 1 \right\}.
$$

On the other hand, by Lemma 3.2 in Candès and Recht (2009), there exist matrix $\bar{\boldsymbol{Z}}$ with $\|\bar{\boldsymbol{Z}}\| = 1$ such that $\langle \bar{\boldsymbol{Z}}, \boldsymbol{X}\boldsymbol{\beta}_0\rangle = \|\bar{\boldsymbol{Z}}\|\|\boldsymbol{X}\boldsymbol{\beta}_0\|_* = \|\boldsymbol{X}\boldsymbol{\beta}_0\|_*$. Pick $\boldsymbol{Z} \in \partial\|\boldsymbol{B}_0\|_*$ such that $\boldsymbol{P}_{\mathcal{B}_{\boldsymbol{u}}^\perp}\boldsymbol{Z}\boldsymbol{P}_{\mathcal{B}_{\boldsymbol{v}}^\perp} = \boldsymbol{P}_{\mathcal{B}_{\boldsymbol{u}}^\perp}\bar{\boldsymbol{Z}}\boldsymbol{P}_{\mathcal{B}_{\boldsymbol{v}}^\perp}$, then we have

$$\begin{aligned}\langle \boldsymbol{Z}, \boldsymbol{X}\boldsymbol{\beta}_0\rangle &= \left\langle \sum_{i=1}^{r_{\boldsymbol{B}_0}} \boldsymbol{u}_{\boldsymbol{B}_0}^{(i)}\boldsymbol{v}_{\boldsymbol{B}_0}^{(i)T} + \boldsymbol{P}_{\mathcal{B}_{\boldsymbol{u}}^\perp}\bar{\boldsymbol{Z}}\boldsymbol{P}_{\mathcal{B}_{\boldsymbol{v}}^\perp}, \boldsymbol{X}\boldsymbol{\beta}_0 \right\rangle \\ &= 0 + \left\langle \bar{\boldsymbol{Z}}\boldsymbol{P}_{\mathcal{B}_{\boldsymbol{v}}^\perp}, \boldsymbol{X}\boldsymbol{\beta}_0 \right\rangle = \left\langle \bar{\boldsymbol{Z}}, \boldsymbol{X}\boldsymbol{\beta}_0\right\rangle - \left\langle \bar{\boldsymbol{Z}}\boldsymbol{P}_{\mathcal{B}_{\boldsymbol{v}}}, \boldsymbol{X}\boldsymbol{\beta}_0\right\rangle \\ &\geq \|\boldsymbol{X}\boldsymbol{\beta}_0\|_* - \left\|\bar{\boldsymbol{Z}}\boldsymbol{P}_{\mathcal{B}_{\boldsymbol{v}}}\right\| \|\boldsymbol{X}\boldsymbol{\beta}_0\|_* \geq \|\boldsymbol{X}\boldsymbol{\beta}_0\|_* - \|\boldsymbol{X}\boldsymbol{\beta}_0\|_* = 0.\end{aligned}$$

Thus we show that $\|\boldsymbol{B}_0\|_* \leq \|\boldsymbol{A}_0\|_*$.

If $\mathcal{R}(\boldsymbol{\beta}_0) \nsubseteq \mathcal{R}(\boldsymbol{B}_0)$, it implies that $\boldsymbol{\beta}_0\boldsymbol{P}_{\mathcal{B}_{\boldsymbol{v}}} \neq \boldsymbol{\beta}_0$. Thus for the inequality above, we always have $\langle \boldsymbol{Z}, \boldsymbol{X}\boldsymbol{\beta}_0\rangle > 0$ which implies $\|\boldsymbol{A}_0\|_* > \|\boldsymbol{B}_0\|_*$. □

S2.1 Compare the Upper Bounds

As for $d^2(\boldsymbol{X}\hat{\boldsymbol{\beta}}^{\mathrm{UNI}}, \boldsymbol{X}\boldsymbol{\beta}_0)$, it follows from the closed form of $\hat{\boldsymbol{\beta}}^{\mathrm{UNI}}$ that

$$\begin{aligned}\boldsymbol{X}\hat{\boldsymbol{\beta}}^{\mathrm{UNI}} - \boldsymbol{X}\boldsymbol{\beta}_0 =& \boldsymbol{X}(n_1^{-1}\boldsymbol{X}^\intercal\boldsymbol{X} + n_2\lambda_1\boldsymbol{I}_{m\times m})^{-1}n_1^{-1}\boldsymbol{X}^\intercal\left(\frac{n_1n_2}{N}\boldsymbol{W}\circ\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}_0\right) \\ &- \boldsymbol{X}(n_1^{-1}\boldsymbol{X}^\intercal\boldsymbol{X} + n_2\lambda_1\boldsymbol{I}_{m\times m})^{-1}n_2\lambda_1 n_1^{-1}\boldsymbol{X}\boldsymbol{\beta}_0.\end{aligned}$$

Take $\lambda_1 = o(n_2^{-1})$, $n_2\lambda_1 = o(1)$, we have $\boldsymbol{X}(n_1^{-1}\boldsymbol{X}^\intercal\boldsymbol{X} + n_2\lambda_1\boldsymbol{I}_{m\times m})^{-1}n_1^{-1}\boldsymbol{X}^\intercal = \boldsymbol{P}_{\boldsymbol{X}}(1+o(1))$. It implies that,

$$\frac{1}{n_1n_2}\left\|\boldsymbol{X}\hat{\boldsymbol{\beta}}^{\mathrm{UNI}} - \boldsymbol{X}\boldsymbol{\beta}_0\right\|_F^2 \leq \frac{1}{n_1n_2}\left\|\boldsymbol{P}_{\boldsymbol{X}}\left(\frac{n_1n_2}{N}\boldsymbol{W}\circ\boldsymbol{Y} - \boldsymbol{A}_0\right)\right\|_F^2(1+o(1)) + n_2^2\lambda_1^2\|\boldsymbol{X}\boldsymbol{\beta}_0\|_F^2(1+o(1)).$$

Let $\boldsymbol{P}_{\boldsymbol{X}} = (s_{ij})$, $\mathsf{E}\|\boldsymbol{P}_{\boldsymbol{X}}(\frac{n_1n_2}{N}\boldsymbol{W}\circ\boldsymbol{Y} - \boldsymbol{A}_0)\|_F^2 = \sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\mathsf{E}(\sum_{k=1}^{n_1}s_{ik}(n_1n_2\omega_{kj}Y_{kj}/N - A_{0kj}))^2 \leq 2\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}(\sum_{k=1}^{n_1}s_{ik}^2\mathsf{E}(n_1n_2\omega_{kj}A_{0kj}/N - A_{0kj})^2 + \sum_{k=1}^{n_1}s_{ik}^2\mathsf{E}(n_1n_2\omega_{kj}\epsilon_{kj}/N)^2)$. Due to Condition C1 and C4, we have $\max\mathsf{E}\epsilon_{ij}^2 \leq c_\sigma^2$ and $\|\boldsymbol{A}_0\|_\infty \leq \sqrt{\log(n)}a_1$. Since $\omega_{kj} \overset{\text{i.i.d}}{\sim} \mathrm{Bern}(\theta_0)$, we have

$$\begin{aligned}\mathsf{E}\left(\frac{\omega_{kj}}{N}\right) =& \mathsf{E}\left(\frac{\omega_{kj}}{\omega_{kj} + \sum_{(s,t)\neq(k,j)}\omega_{st}}\right) = \mathsf{E}\left\{\mathsf{E}\left(\frac{\omega_{kj}}{\omega_{kj}+c}\middle|\sum_{(s,t)\neq(k,j)}\omega_{st} = c\right)\right\} \\ =& \mathsf{E}\left\{\sum_{c=0}^{n_1n_2-1}\frac{\theta_0}{1+c}\right\} = \frac{1}{n_1n_2}(1-(1-\theta_0)^{n_1n_2}) \leq \frac{1}{n_1n_2},\end{aligned}$$

and similarly,

$$\mathsf{E}\left(\frac{\omega_{kj}}{N^2}\right) = \mathsf{E}\left\{\frac{\omega_{kj}}{\left(\omega_{kj}+\sum_{(s,t)\neq(k,j)}\omega_{st}\right)^2}\right\} = \mathsf{E}\left\{\sum_{c=0}^{n_1n_2-1}\frac{\theta_0}{(1+c)^2}\right\} \leq \frac{2}{n_1n_2\left(n_1n_2+1\right)\theta_0}.$$

Combine the above two results together, we have

$$\begin{aligned}\mathsf{E}\left\|\boldsymbol{P}_{\boldsymbol{X}}\left(\frac{n_1n_2}{N}\boldsymbol{W}\circ\boldsymbol{Y}-\boldsymbol{A}_0\right)\right\|_F^2 \leq & 2\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\left\{\left(\frac{2n_1n_2}{\left(n_1n_2+1\right)\theta_0}+2+1\right)\left\{\log\left(n\right)\right\}a_1^2\sum_{k=1}^{n_1}s_{ik}^2+\right.\\ & \left.\frac{2n_1n_2c_\sigma^2}{\left(n_1n_2+1\right)\theta_0}\sum_{k=1}^{n_1}s_{ik}^2\right\}\\ \leq & 2\left\{\left(\frac{2n_1n_2}{\left(n_1n_2+1\right)\theta_0}+3\right)\left\{\log\left(n\right)\right\}a_1^2+\frac{2n_1n_2c_\sigma^2}{\left(n_1n_2+1\right)\theta_0}\right\}n_2m.\end{aligned}$$

Take $\lambda_1 = o\{n_1^{-1}n_2^{-3/2}\log^{-1}(n)\}$, we have $d^2(\boldsymbol{X}\hat{\boldsymbol{\beta}}^{\mathrm{UNI}}, \boldsymbol{X}\boldsymbol{\beta}_0) = O_p(n_1^{-1})$.

Proofs of Theorem 3. Under Condition C3, we have $\|\boldsymbol{A}_0\|_* = O\{\sqrt{n_1n_2\log(n)}\}$ and $\|\boldsymbol{B}_0\|_* = O\{\sqrt{n_1n_2\log(n)}\}$. Under the low rank condition that $r_{\boldsymbol{A}_0} = r_{\boldsymbol{B}_0} + m = o\{\theta_0^{1/2}(n_1\wedge n_2)^{1/2}\}$, we have $\lambda_{\mathrm{KLT}}n_1n_2r_{\boldsymbol{A}_0} = o(\|\boldsymbol{A}_0\|_*)$ and $\lambda_2n_1n_2r_{\boldsymbol{B}_0} = o(\|\boldsymbol{B}_0\|_*)$ since $\lambda_2 = \lambda_{\mathrm{KLT}} \asymp \theta_0^{-1/2}(n_1\wedge n_2)^{-1/2}(n_1n_2)^{-1/2}\log^{1/2}(n)$. Namely, both the first terms in U_{KLT} and U_{UNI} dominate and we compare the second terms. As $r_{\boldsymbol{A}_0} = r_{\boldsymbol{B}_0} + m$, we can claim that $U_{\mathrm{UNI}} < U_{\mathrm{KLT}}$.

For the high rank case, i.e the second term dominates or of the same order as the first term, the first terms in U_{KLT} and U_{UNI} are the smaller order. If $\mathcal{R}(\boldsymbol{\beta}_0) \nsubseteq \mathcal{R}(\boldsymbol{B}_0)$, once there exists the covariate effect, i.e $\boldsymbol{\beta}_0 \neq \boldsymbol{0}^{m\times n_2}$, as given in Proposition S2.1, $\|\boldsymbol{B}_0\|_* < \|\boldsymbol{A}_0\|_*$ which implies $U_{\mathrm{UNI}} < U_{\mathrm{KLT}}$. For the remaining cases, we obtain the result $U_{\mathrm{UNI}} \leq U_{\mathrm{KLT}}$ by $\|\boldsymbol{B}_0\|_* \leq \|\boldsymbol{A}_0\|_*$.

□

S2.2 Proofs of Theorem 4

Proof. For some constant $0 \leq \gamma \leq 1$, if $n_1 \geq n_2$, define

$$\mathcal{C}_1 = \left\{\tilde{\boldsymbol{B}} = (B_{ij}) \in \mathbb{R}^{n_1\times r_{\boldsymbol{B}_0}} : B_{ij} \in \left\{0, \gamma\left(\sigma\wedge a_1\right)\left(\frac{r_{\boldsymbol{B}_0}}{\left(n_1\wedge n_2\right)\theta_0}\right)^{1/2}\right\}, \forall 1\leq i\leq n_1, 1\leq j\leq r_{\boldsymbol{B}_0}\right\},$$

and consider the associated set of block matrices

$$\mathcal{A}\left(\mathcal{C}_1\right) = \left\{\boldsymbol{A} = \boldsymbol{X}\tilde{\boldsymbol{\beta}} + \left(\tilde{\boldsymbol{B}}|\ldots|\tilde{\boldsymbol{B}}|\boldsymbol{0}\right) \in \mathbb{R}^{n_1\times n_2} : \tilde{\boldsymbol{\beta}} \in \beta\left(a_1\right), \tilde{\boldsymbol{B}} \in \mathcal{C}_1\right\},$$

where $\mathbf{0}$ denotes the $n_1 \times (n_2 - r_{\boldsymbol{B}_0}\lfloor n_2/r_{\boldsymbol{B}_0}\rfloor)$ zero matrix.

It is easy to see that any element of $\mathcal{B}(\mathcal{C}_1)$ and the difference of any two elements of $\mathcal{B}(\mathcal{C}_1)$ has rank at most $r_{\boldsymbol{B}_0}$. The entries of any matrix in $\mathcal{B}(\mathcal{C}_1)$ are within $[0, a_1]$. Due to Lemma 2.9 in Tsybakov (2009), there exists a subset $\mathcal{B}^0 \subset \mathcal{B}(\mathcal{C}_2)$ containing the zero $n_1 \times n_2$ matrix $\mathbf{0}$ where $\mathrm{Card}(\mathcal{B}^0) \geq 2^{r_{\boldsymbol{B}_0} n_1/2} + 1$ and for any two distinct elements $\boldsymbol{B}_1$ and $\boldsymbol{B}_2$ of $\mathcal{B}^0$,

$$\|\boldsymbol{B}_1 - \boldsymbol{B}_2\|_F^2 \geq \frac{n_1 r_{\boldsymbol{B}_0}}{8}\left(\gamma^2 (\sigma \wedge a_1)^2 \left(\frac{r_{\boldsymbol{B}_0}}{(n_1 \wedge n_2)\,\theta_0}\right)\left\lfloor \frac{n_2}{r_{\boldsymbol{B}_0}} \right\rfloor\right) \geq \frac{\gamma^2}{16}(\sigma \wedge a_1)^2 \left(\frac{n_1 n_2 r_{\boldsymbol{B}_0}}{(n_1 \wedge n_2)\,\theta_0}\right).$$

For $0 \leq l \leq r_{\boldsymbol{B}_0}$, take $\boldsymbol{\beta}^0 \subset \beta(a_1)$ such that

$$\boldsymbol{\beta}^0 = \left\{\tilde{\boldsymbol{\beta}} \in \mathbb{R}^{m \times n_2} : (X\beta)_{ij} = \gamma(\sigma \wedge a_1)\left(\frac{l}{(n_1 \wedge n_2)\,\theta_0}\right)^{1/2}, \forall 1 \leq i \leq n_1, 1 \leq j \leq n_2\right\}.$$

For any $\boldsymbol{A} \in \mathcal{A}^0 = \boldsymbol{\beta}^0 \cup \mathcal{B}^0$, the Kullback-Leibler divergence $K(\mathbb{P}_{\mathbf{0}}, \mathbb{P}_{\boldsymbol{A}})$ between $\mathbb{P}_{\mathbf{0}}$ and $\mathbb{P}_{\boldsymbol{A}}$ satisfies

$$K(\mathbb{P}_{\mathbf{0}}, \mathbb{P}_{\boldsymbol{A}}) = \mathsf{E}_{\mathbb{P}_{\mathbf{0}}}\left(\sum_{ij} \omega_{ij} \frac{A_{0ij}^2 - 2A_{0ij}Y_{0ij}}{2\sigma^2}\right) = \theta_0 \frac{\|\boldsymbol{A}\|_F^2}{2\sigma^2} \leq \frac{\gamma^2 (r_{\boldsymbol{B}_0} + l)\, n_1 n_2}{n_1 \wedge n_2}.$$

It is easy to know that $\mathrm{Card}(\mathcal{A}^0) = \mathrm{Card}(\mathcal{B}^0) \geq 2^{r_{\boldsymbol{B}_0} n_1/2} + 1$. From above we deduce the condition

$$\frac{1}{\mathrm{Card}(\mathcal{A}^0) - 1} \sum_{\boldsymbol{A} \in \mathcal{A}^0} K(\mathbb{P}_{\mathbf{0}}, \mathbb{P}_{\boldsymbol{A}}) \leq \alpha \log\left(\mathrm{Card}(\mathcal{A}^0) - 1\right) \tag{S2.1}$$

is satisfied for any $\alpha > 0$ if $0 < \gamma < \sqrt{\alpha}/2$ and $l \leq r_{\boldsymbol{B}_0}$. The result now follows by application of Theorem 2.5 in Tsybakov (2009).

For $n_1 \leq n_2$, similarly, define

$$\mathcal{C}_2 = \left\{\tilde{\boldsymbol{B}} = (B_{ij}) \in \mathbb{R}^{r_{\boldsymbol{B}_0} \times n_2} : B_{ij} \in \left\{0, \gamma(\sigma \wedge a_1)\left(\frac{r_{\boldsymbol{B}_0}}{(n_1 \wedge n_2)\,\theta_0}\right)^{1/2}\right\}, \forall 1 \leq i \leq r_{\boldsymbol{B}_0}, 1 \leq j \leq n_2\right\},$$

and consider the associated set of block matrices

$$\mathcal{A}(\mathcal{C}_2) = \left\{\boldsymbol{A} = \boldsymbol{X}\tilde{\boldsymbol{\beta}} + \left(\tilde{\boldsymbol{B}}|\ldots|\tilde{\boldsymbol{B}}|\mathbf{0}\right)^T \in \mathbb{R}^{n_1 \times n_2} : \tilde{\boldsymbol{\beta}} \in \beta(a_1), \tilde{\boldsymbol{B}} \in \mathcal{C}_2\right\},$$

where $\mathbf{0}$ denotes the $(n_1 - r_{\boldsymbol{B}_0}\lfloor n_1/r_{\boldsymbol{B}_0}\rfloor) \times n_2$ zero matrix here. Follow the same proof, we have the same result. □

S2.3 Non-Uniform Missing

For the non-uniform missing, we assume that the missing probability $\boldsymbol{\Theta} = (\theta_{ij})$ is known. Namely, we know $\boldsymbol{\Theta}^* = (1/\theta_{ij})$ in the risk function (3.1). Thus

$$\hat{\boldsymbol{B}}^{\text{NON-UNI}} = \underset{\boldsymbol{B}\in\mathbb{R}^{n_1\times n_2}}{\arg\min} \left\{ \frac{1}{n_1 n_2} \|\boldsymbol{B} - \boldsymbol{W}\circ\boldsymbol{\Theta}^*\circ\boldsymbol{Y}\|_F^2 + \lambda_2 \|\boldsymbol{B}\|_* \right\}. \tag{S2.2}$$

Follow the same proof of Theorem 3 of Koltchinskii et al. (2011), we have that

Theorem S2.1. *Assume Conditions C1-C4, if* $\lambda_2 \geq 2\|\boldsymbol{W}\circ\boldsymbol{\Theta}^*\circ\boldsymbol{Y} - \boldsymbol{A}_0\|$, *then*

$$d^2\left(\hat{\boldsymbol{B}}^{\text{NON-UNI}}, \boldsymbol{B}_0\right) \leq \lambda_2 \min\left\{ 2\|\boldsymbol{B}_0\|_*, \left(\frac{1+\sqrt{2}}{2}\right)^2 \lambda_2 n_1 n_2 r_{\boldsymbol{B}_0} \right\}.$$

As for $d^2(\boldsymbol{X}\hat{\boldsymbol{\beta}}^{\text{NON-UNI}}, \boldsymbol{X}\boldsymbol{\beta}_0)$, it follows from the closed form of $\hat{\boldsymbol{\beta}}$ that

$$\begin{aligned}\boldsymbol{X}\hat{\boldsymbol{\beta}}^{\text{NON-UNI}} - \boldsymbol{X}\boldsymbol{\beta}_0 =& \boldsymbol{X}(n_1^{-1}\boldsymbol{X}^\intercal\boldsymbol{X} + n_2\lambda_1\boldsymbol{I}_{m\times m})^{-1} n_1^{-1}\boldsymbol{X}^\intercal(\boldsymbol{W}\circ\boldsymbol{\Theta}^*\circ\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}_0) \\ & - \boldsymbol{X}(n_1^{-1}\boldsymbol{X}^\intercal\boldsymbol{X} + n_2\lambda_1\boldsymbol{I}_{m\times m})^{-1} n_2\lambda_1 n_1^{-1}\boldsymbol{X}\boldsymbol{\beta}_0.\end{aligned}$$

Take $\lambda_1 = o(n_2^{-1})$, $n_2\lambda_1 = o(1)$, we have $\boldsymbol{X}(n_1^{-1}\boldsymbol{X}^\intercal\boldsymbol{X} + n_2\lambda_1\boldsymbol{I}_{m\times m})^{-1} n_1^{-1}\boldsymbol{X}^\intercal = \boldsymbol{P}_{\boldsymbol{X}}(1+o(1))$. It implies that,

$$\frac{1}{n_1 n_2}\left\|\boldsymbol{X}\hat{\boldsymbol{\beta}}^{\text{NON-UNI}} - \boldsymbol{X}\boldsymbol{\beta}_0\right\|_F^2 \leq \frac{1}{n_1 n_2}\|\boldsymbol{P}_{\boldsymbol{X}}(\boldsymbol{W}\circ\boldsymbol{\Theta}^*\circ\boldsymbol{Y} - \boldsymbol{A}_0)\|_F^2 (1+o(1)) + n_2^2\lambda_1^2\|\boldsymbol{X}\boldsymbol{\beta}_0\|_F^2 (1+o(1)).$$

It is not hard to show that $\mathsf{E}\|\boldsymbol{P}_{\boldsymbol{X}}(\boldsymbol{W}\circ\boldsymbol{\Theta}^*\circ\boldsymbol{Y} - \boldsymbol{A}_0)\|_F^2 \leq \{(1/\theta_L - 1)\log(n)a_1^2 + c_\sigma^2/\theta_L\} n_2 m$. Then take $\lambda_1 = o(n_1^{-1}n_2^{-3/2}\log^{-1}(n))$, we have $d^2(\boldsymbol{X}\hat{\boldsymbol{\beta}}^{\text{NON-UNI}}, \boldsymbol{X}\boldsymbol{\beta}_0) = O_p(n_1^{-1})$.

The lower bound can be given in the following theorem.

Theorem S2.2. *Assume Condition C4, fix* $a_1 > 0$, *for* $r_{\boldsymbol{B}_0}$ *such that* $1 \leq r_{\boldsymbol{B}_0} \leq \min(n_1, n_2) - m$, $(n_1 \vee n_2) r_{\boldsymbol{B}_0} \leq n_1 n_2 \theta_L$. *Let the variables* ϵ_{ij} *be Gaussian* $\mathcal{N}(0, \sigma^2)$, $\sigma^2 > 0$ *for* $i = 1, \ldots, n_1$, $j = 1, \ldots, n_2$. *Then there exist absolute constants* $\alpha \in (0,1)$, $c > 0$ *and* $0 \leq l \leq r_{\boldsymbol{B}_0}$, *such that*

$$\inf_{\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{B}}\boldsymbol{\beta}_0 \in \beta(a_1), \boldsymbol{B}_0 \in \mathcal{B}\left(r_{\boldsymbol{B}_0}, a_1\right)} \sup P\left(d^2(\hat{\boldsymbol{A}}, \boldsymbol{A}_0) > c(\sigma \wedge a_1)^2 \frac{(n_1 \vee n_2)(r_{\boldsymbol{B}_0} + l)}{n_1 n_2 \theta_L} \right) \geq \alpha.$$

S3 Justification of Condition C5(b)

Sweeting (1980) presented a very general result concerning the uniform asymptotic normality of the MLEs. In this section, we want to verify Condition C5(b) under the logistic sampling model given in (4.1) by applying Sweeting's results. A natural estimator of $\boldsymbol{\gamma}_{\cdot j}$ is the conditional MLE $\hat{\boldsymbol{\gamma}}_{\cdot j}$, denoted as that maximizes the log-likelihood,

$$\ell_{n_1}\left(\boldsymbol{\gamma}_{\cdot j}\right)=\sum_{i=1}^{n_1}\left\{\omega_{ij}\log\theta_{ij}+\left(1-\omega_{ij}\right)\log\left(1-\theta_{ij}\right)\right\}.$$

We know that the MLE $\hat{\boldsymbol{\gamma}}_{\cdot j}$ of $\boldsymbol{\gamma}_{\cdot j}$ is a consistent estimator and the asymptotic normality of $\boldsymbol{\gamma}_{\cdot j}$ for each $j=1,\ldots,n_2$ under some regularity conditions. Then we apply Sweeting's result to show the uniform asymptotic normality of these MLEs.

The conditional Fisher information matrix is

$$I_{n_1}\left(\boldsymbol{\gamma}_{\cdot j}\right)=\mathsf{E}\left(-\frac{\partial^2\ell_{n_1}\left(\boldsymbol{\gamma}_{\cdot j}\right)}{\partial\boldsymbol{\gamma}_{\cdot j}^2}\right)=\sum_{i=1}^{n_1}\theta_{ij}\left(1-\theta_{ij}\right)\tilde{\boldsymbol{x}}_i\tilde{\boldsymbol{x}}_i^{\mathsf{T}}. \tag{S3.1}$$

Let $\bar{\boldsymbol{x}}_c=\lim_{n_1\to\infty}n_1^{-1}\sum_{i=1}^{n_1}\boldsymbol{x}_i$ and $\tilde{\boldsymbol{S}}_x=\begin{bmatrix}1 & \bar{\boldsymbol{x}}_c^{\mathsf{T}}\\ \bar{\boldsymbol{x}}_c & \boldsymbol{S}_x\end{bmatrix}$. To guarantee the sum of squared errors in Condition C5(b), we require the following conditions for the sampling model:

CA(a). (i) There exists a universal upper bound $\theta_U\in(0,1)$, where θ_U is allowed to depend on n_1 and n_2, such that $\max_{i,j}\{\theta_{ij}\}\le\theta_U<1$ uniformly. (ii) $0<\|\tilde{\boldsymbol{S}}_x\|<\infty$ and $\tilde{\boldsymbol{S}}_x>0$.

Condition CA(a) is a mild condition. The upper bound θ_U and the lower bound θ_L in C5(a) are considered together to ensure the invertibility of $I_{n_1}(\boldsymbol{\gamma}_{\cdot j})$.

Denote the parameter space $\boldsymbol{\Xi}$ is a bounded subset of $\mathbb{R}^{m+1}$ which covers the parameters $\boldsymbol{\gamma}_{\cdot j}$ for $j=1,\ldots,n_2$. Let $P_{\boldsymbol{\xi}}$, $P_{n_1,\boldsymbol{\xi}}$, $n_1\ge1$, be probability measures of random variables $\boldsymbol{A}(\boldsymbol{\xi})$, $\boldsymbol{A}_{n_1}(\boldsymbol{\xi})$, $n_1\ge1$ defined on the Borel subset of a metric space depending on a $\boldsymbol{\xi}\in\boldsymbol{\Xi}$, and let $C(\mathbb{R}^{m+1})$ be the space of real bounded uniformly continuous functions, $\boldsymbol{A}_{n_1}(\boldsymbol{\xi})\overset{u}{\Rightarrow}\boldsymbol{A}(\boldsymbol{\xi})$ in $\boldsymbol{\xi}\in\boldsymbol{\Xi}$ if and only if

$$\sup_{\boldsymbol{\xi}\in\boldsymbol{\Xi}}\left|P_{n_1,\boldsymbol{\xi}}\left(\boldsymbol{S}\right)-P_{\boldsymbol{\xi}}\left(\boldsymbol{S}\right)\right|\to0,\text{as }n_1\to\infty,$$

for any Borel set $\boldsymbol{S}$ with $P_{\boldsymbol{\xi}}\left(\partial\boldsymbol{S}\right)=0$.

In order to show the uniform weak convergence of MLEs, Sweeting proposed additional two regularity conditions in Sweeting (1980), which we present in a form that would connect well to the logistic regression model setting.

CA(b). There exist nonrandom square matrices $\boldsymbol{D}_{n_1}(\boldsymbol{\xi})$, continuous in $\boldsymbol{\xi}$, satisfying $\sup_{\boldsymbol{\xi}\in\boldsymbol{\Xi}}\|\boldsymbol{D}_{n_1}^{-1}(\boldsymbol{\xi})\|_F \to 0$, as $n_1 \to \infty$, such that

$$\boldsymbol{W}_{n_1}(\boldsymbol{\xi}) \equiv \boldsymbol{D}_{n_1}^{-1}(\boldsymbol{\xi})\, I_{n_1}(\boldsymbol{\xi}) \left\{\boldsymbol{D}_{n_1}^{-1}(\boldsymbol{\xi})\right\}^{\intercal} \overset{u}{\Rightarrow} \boldsymbol{W}(\boldsymbol{\xi}),$$

and $\mathsf{P}(\boldsymbol{W}(\boldsymbol{\xi}) > 0) = 1$.

CA(c). For all $\epsilon > 0$, (i) $\sup_{\boldsymbol{\xi}\in\boldsymbol{\Xi}}\sup_{\boldsymbol{\xi}'\in\mathcal{A}(\boldsymbol{\xi},\epsilon)} \|\boldsymbol{D}_{n_1}^{-1}(\boldsymbol{\xi})\boldsymbol{D}_{n_1}(\boldsymbol{\xi}') - \boldsymbol{I}_{m+1}\|_F \to 0$, where $\mathcal{A}(\boldsymbol{\xi},\epsilon) = \{\boldsymbol{\xi}' \in \boldsymbol{\Xi} : \|\boldsymbol{D}_{n_1}^{\intercal}(\boldsymbol{\xi})(\boldsymbol{\xi}' - \boldsymbol{\xi})\|_F \le \epsilon\}$, and

(ii)$\sup_{\boldsymbol{\xi}\in\boldsymbol{\Xi}}\sup_{\boldsymbol{\xi}^k\in\mathcal{A}(\boldsymbol{\xi},\epsilon),1\le k\le(m+1)} \|\boldsymbol{D}_{n_1}^{-1}(\boldsymbol{\xi})\{(I_{n_1}(\boldsymbol{\xi}^1)_{1.}^{\intercal},\ldots,I_{n_1}(\boldsymbol{\xi}^{m+1})_{(m+1).}^{\intercal}) - I_{n_1}(\boldsymbol{\xi})\}\{\boldsymbol{D}_{n_1}^{-1}(\boldsymbol{\xi})\}^{\intercal}\|_F \to 0$, where $I_{n_1}(\boldsymbol{\xi}^k)_{k.}$ is the $k-$th row of $I_{n_1}(\boldsymbol{\xi}^k)$ for $1 \le k \le m+1$.

Under growth and convergence Condition CA(b) and continuity Condition CA(c), Corollary 1 of Sweeting (1980) showed that the MLE of $\hat{\boldsymbol{\xi}}$ is asymptotic normal uniformly with respect to $\boldsymbol{\xi} \in \boldsymbol{\Xi}$,

$$\boldsymbol{W}_{n_1}^{1/2}(\boldsymbol{\xi})\, \boldsymbol{D}_{n_1}(\boldsymbol{\xi}) \left(\hat{\boldsymbol{\xi}} - \boldsymbol{\xi}\right) \overset{u}{\Rightarrow} \boldsymbol{Z}, \text{as } n_1 \to \infty,$$

where $\boldsymbol{Z}$ is the standard normal random vector in $\mathbb{R}^{m+1}$ and independent of $\boldsymbol{W}(\boldsymbol{\xi})$.

In the case of the logistic regression model, the parameter space $\boldsymbol{\Xi}$ is an open subset of $\mathbb{R}^{m+1}$ such that for any $\boldsymbol{\xi} \in \boldsymbol{\Xi}$ and $\theta_{i\boldsymbol{\xi}} = \exp(\boldsymbol{x}_i^{\intercal}\boldsymbol{\xi})/\{1+\exp(\boldsymbol{x}_i^{\intercal}\boldsymbol{\xi})\}$, $0 < \theta_L \le \min_{i,j}\{\theta_{i\boldsymbol{\xi}}\} \le \max_{i,j}\{\theta_{i\boldsymbol{\xi}}\} \le \theta_U < 1$. Let $\pi_{\boldsymbol{\xi}} = n_1^{-1}\sum_{i=1}^{n_1}\theta_{i\boldsymbol{\xi}}(1-\theta_{i\boldsymbol{\xi}})$, $\boldsymbol{D}_{n_1}(\boldsymbol{\xi}) = (n_1\pi_{\boldsymbol{\xi}})^{1/2}\boldsymbol{I}_{m+1}$ and $\boldsymbol{W}(\boldsymbol{\xi}) = \tilde{\boldsymbol{S}}_x$, thus $\boldsymbol{W}_{n_1}(\boldsymbol{\xi}) \equiv \boldsymbol{D}_{n_1}^{-1}(\boldsymbol{\xi})I_{n_1}(\boldsymbol{\xi})\{\boldsymbol{D}_{n_1}^{-1}(\boldsymbol{\xi})\}^{\intercal} = (n_1\pi_{\boldsymbol{\xi}})^{-1}I_{n_1}(\boldsymbol{\xi})$, where $I_{n_1}(\boldsymbol{\xi})$ is defined as the Fisher matrix in (S3.1). The justifications of Conditions CA(b) and CA(c) on any $\boldsymbol{\xi} \in \boldsymbol{\Xi}$ are given in the following.

Justification of Condition CA(b). For any $\boldsymbol{\xi} \in \boldsymbol{\Xi}$, since $\boldsymbol{\Xi}$ is a bounded subset of $\mathbb{R}^{m+1}$, then $\pi_{\boldsymbol{\xi}} = \sum_{i=1}^{n_1}\theta_{i\boldsymbol{\xi}}(1-\theta_{i\boldsymbol{\xi}})/n_1 \in \{\min\{\theta_U(1-\theta_U), \theta_L(1-\theta_L)\}, 1\}$. It is easy to see that $\sup_{\boldsymbol{\xi}\in\boldsymbol{\Xi}}\|\boldsymbol{D}_{n_1}^{-1}(\boldsymbol{\xi})\|_F = \sqrt{m+1}(n_1\pi_{\boldsymbol{\xi}})^{-1/2} \to 0$ under the case $\theta_L > (n_1 n_2)^{-1}(n_1 \vee n_2)\log(n)$ as $n_1 \to \infty$.

Under Condition C2, there exist a positive constant a_x such that $\|\boldsymbol{X}\|_\infty < a_x$, $\lim_{n_1\to\infty} n_1^{-1}\boldsymbol{X}^{\intercal}\boldsymbol{X} =$

$\lim_{n_1\to\infty} n_1^{-1}\sum_{i=1}^{n_1}\boldsymbol{x}_i\boldsymbol{x}_i^\mathsf{T} = \boldsymbol{S}_x$. Also we have $\bar{\boldsymbol{x}}_c = \lim_{n_1\to\infty} n_1^{-1}\sum_{i=1}^{n_1}\boldsymbol{x}_i$, thus

$$\begin{aligned} n_1^{-1}\sum_{i=1}^{n_1}\theta_{i\boldsymbol{\xi}}\left(1-\theta_{i\boldsymbol{\xi}}\right)\boldsymbol{x}_i - \pi_{\boldsymbol{\xi}}\bar{\boldsymbol{x}}_c =& n_1^{-1}\sum_{i=1}^{n_1}\theta_{i\boldsymbol{\xi}}\left(1-\theta_{i\boldsymbol{\xi}}\right)\boldsymbol{x}_i - n_1^{-1}\sum_{i=1}^{n_1}\theta_{i\boldsymbol{\xi}}\left(1-\theta_{i\boldsymbol{\xi}}\right)\bar{\boldsymbol{x}}_c \\ &\leq \left| n_1^{-1}\sum_{i=1}^{n_1}\theta_{i\boldsymbol{\xi}}\left(1-\theta_{i\boldsymbol{\xi}}\right)\left(\boldsymbol{x}_i-\bar{\boldsymbol{x}}_c\right)\right| \to 0. \end{aligned}$$

Similarly, we have $n_1^{-1}\sum_{i=1}^{n_1}\theta_{i\boldsymbol{\xi}}(1-\theta_{i\boldsymbol{\xi}})\boldsymbol{x}_i \to \pi_{\boldsymbol{\xi}}\bar{\boldsymbol{x}}_c$ and $n_1^{-1}\sum_{i=1}^{n_1}\theta_{i\boldsymbol{\xi}}(1-\theta_{i\boldsymbol{\xi}})\boldsymbol{x}_i\boldsymbol{x}_i^\mathsf{T} \to \pi_{\boldsymbol{\xi}}\boldsymbol{S}_x$. These imply, $(n_1\pi_{\boldsymbol{\xi}})^{-1}I_{n_1}(\boldsymbol{\xi}) = (n_1\pi_{\boldsymbol{\xi}})^{-1}\sum_{i=1}^{n_1}\theta_{i\boldsymbol{\xi}}(1-\theta_{i\boldsymbol{\xi}})\tilde{\boldsymbol{x}}_i\tilde{\boldsymbol{x}}_i^\mathsf{T} \to \tilde{\boldsymbol{S}}_x$.

Since $\boldsymbol{D}_{n_1}(\boldsymbol{\xi}) = (n_1\pi_{\boldsymbol{\xi}})^{1/2}\boldsymbol{I}_{m+1}$, $\boldsymbol{W}(\boldsymbol{\xi}) = \tilde{\boldsymbol{S}}_x$,

$$\boldsymbol{W}_{n_1}(\boldsymbol{\xi}) \equiv (n_1\pi_{\boldsymbol{\xi}})^{-1} I_{n_1}(\boldsymbol{\xi}) \overset{u}{\Rightarrow} \tilde{\boldsymbol{S}}_x,$$

Here $\boldsymbol{W}(\boldsymbol{\xi}) = \tilde{\boldsymbol{S}}_x$ and $\mathsf{P}(\tilde{\boldsymbol{S}}_x > 0) = 1$. □

Justification of Condition CA(c). For Condition CA(c)(i), for $\boldsymbol{\xi} \in \boldsymbol{\Xi}$, the set $\mathcal{A}(\boldsymbol{\xi},\epsilon) = \|\boldsymbol{D}_{n_1}^\mathsf{T}(\boldsymbol{\xi})(\boldsymbol{\xi}'-\boldsymbol{\xi})\|_F \leq \epsilon$ implies

$$\mathrm{tr}\left\{\left(\boldsymbol{\xi}'-\boldsymbol{\xi}\right)^\mathsf{T}\boldsymbol{D}_{n_1}(\boldsymbol{\xi})\boldsymbol{D}_{n_1}^\mathsf{T}(\boldsymbol{\xi})\left(\boldsymbol{\xi}'-\boldsymbol{\xi}\right)\right\} = (n_1\pi_{\boldsymbol{\xi}})\,\mathrm{tr}\left\{\left(\boldsymbol{\xi}'-\boldsymbol{\xi}\right)^\mathsf{T}\left(\boldsymbol{\xi}'-\boldsymbol{\xi}\right)\right\} \leq \epsilon^2.$$

Let $\theta_{i\boldsymbol{\xi}'} = \exp(\tilde{\boldsymbol{x}}_i^\mathsf{T}\boldsymbol{\xi}')/\{1+\exp(\tilde{\boldsymbol{x}}_i^\mathsf{T}\boldsymbol{\xi}')\}$ and $\pi_{\boldsymbol{\xi}'} = \sum_{i=1}^{n_1}\theta_{i\boldsymbol{\xi}'}(1-\theta_{i\boldsymbol{\xi}'})$. Since we have $\theta_{i\boldsymbol{\xi}'} - \theta_{i\boldsymbol{\xi}} = \theta_{i\boldsymbol{\xi}}(1-\theta_{i\boldsymbol{\xi}})\tilde{\boldsymbol{x}}_i^\mathsf{T}(\boldsymbol{\xi}'-\boldsymbol{\xi}) + (\boldsymbol{\xi}'-\boldsymbol{\xi})^\mathsf{T}\tilde{\boldsymbol{x}}_i(1-2\theta_{i\boldsymbol{\xi}^*})\theta_{i\boldsymbol{\xi}^*}(1-\theta_{i\boldsymbol{\xi}^*})\tilde{\boldsymbol{x}}_i^\mathsf{T}(\boldsymbol{\xi}'-\boldsymbol{\xi})$ for $\boldsymbol{\xi}^* \in \mathcal{B}^{m+1}(\boldsymbol{\xi}, d(\boldsymbol{\xi},\boldsymbol{\xi}'))$, where $\mathcal{B}^{m+1}(\boldsymbol{\xi}, d(\boldsymbol{\xi},\boldsymbol{\xi}'))$ is the ball belongs to $\mathbb{R}^{m+1}$ with center at $\boldsymbol{\xi}$ and radius $d(\boldsymbol{\xi},\boldsymbol{\xi}')$, $d(\boldsymbol{\xi},\boldsymbol{\xi}')$ is euclidean distance between the vector $\boldsymbol{\xi}$ and $\boldsymbol{\xi}'$. Since $\boldsymbol{\xi}^* \in \boldsymbol{\Xi}$, we have $|(1-2\theta_{i\boldsymbol{\xi}^*})\theta_{i\boldsymbol{\xi}^*}(1-\theta_{i\boldsymbol{\xi}^*})| < 2$. Combining the fact that there exist a positive constant a_x such that $\|\boldsymbol{X}\|_\infty < a_x$, $\|\tilde{\boldsymbol{x}}_i\tilde{\boldsymbol{x}}_i^\mathsf{T}\| < \infty$, we can say that $\theta_{i\boldsymbol{\xi}'} - \theta_{i\boldsymbol{\xi}} = \theta_{i\boldsymbol{\xi}}(1-\theta_{i\boldsymbol{\xi}})\tilde{\boldsymbol{x}}_i^\mathsf{T}(\boldsymbol{\xi}'-\boldsymbol{\xi}) + o((\boldsymbol{\xi}'-\boldsymbol{\xi}))$ and $(\theta_{i\boldsymbol{\xi}'}-\theta_{i\boldsymbol{\xi}})^2 = \theta_{i\boldsymbol{\xi}}^2(1-\theta_{i\boldsymbol{\xi}})^2\mathrm{tr}(\tilde{\boldsymbol{x}}_i\tilde{\boldsymbol{x}}_i^\mathsf{T}(\boldsymbol{\xi}'-$

$\boldsymbol{\xi})(\boldsymbol{\xi}'-\boldsymbol{\xi})^\intercal)+o((\boldsymbol{\xi}'-\boldsymbol{\xi})^\intercal(\boldsymbol{\xi}'-\boldsymbol{\xi}))$. It implies that

$$
\begin{aligned}
(n_1\pi_{\boldsymbol{\xi}})^{1/2}\left|\pi_{\boldsymbol{\xi}'}-\pi_{\boldsymbol{\xi}}\right| =& (n_1\pi_{\boldsymbol{\xi}})^{1/2}\left|n_1^{-1}\sum_{i=1}^{n_1}\left\{\theta_{i\boldsymbol{\xi}'}\left(1-\theta_{i\boldsymbol{\xi}'}\right)-\theta_{i\boldsymbol{\xi}}\left(1-\theta_{i\boldsymbol{\xi}}\right)\right\}\right| \leq n_1^{-1/2}\pi_{\boldsymbol{\xi}}^{1/2}\sum_{i=1}^{n_1}3\left|\theta_{i\boldsymbol{\xi}'}-\theta_{i\boldsymbol{\xi}}\right| \\
\leq& 3n_1^{-1/2}\pi_{\boldsymbol{\xi}}^{1/2}\sqrt{\frac{n_1}{n_1\pi_{\boldsymbol{\xi}}}\sum_{i=1}^{n_1}\left\{n_1\pi_{\boldsymbol{\xi}}\left(\theta_{i\boldsymbol{\xi}'}-\theta_{i\boldsymbol{\xi}}\right)^2\right\}} \\
\leq& 3n_1^{-1/2}\sqrt{\sum_{i=1}^{n_1}n_1\pi_{\boldsymbol{\xi}}\mathrm{tr}\left\{\tilde{\boldsymbol{x}}_i\tilde{\boldsymbol{x}}_i^\intercal\left(\boldsymbol{\xi}'-\boldsymbol{\xi}\right)\left(\boldsymbol{\xi}'-\boldsymbol{\xi}\right)^\intercal+o\left(\left(\boldsymbol{\xi}'-\boldsymbol{\xi}\right)^\intercal\left(\boldsymbol{\xi}'-\boldsymbol{\xi}\right)\right)\right\}} \\
\leq& 3n_1^{-1/2}\sqrt{2n_1\mathrm{tr}\left\{\left(\frac{1}{n_1}\sum_{i=1}^{n_1}\tilde{\boldsymbol{x}}_i\tilde{\boldsymbol{x}}_i^\intercal\right)n_1\pi_{\boldsymbol{\xi}}\left(\boldsymbol{\xi}'-\boldsymbol{\xi}\right)\left(\boldsymbol{\xi}'-\boldsymbol{\xi}\right)^\intercal\right\}} \leq 3\sqrt{2\left\|\tilde{\boldsymbol{S}}_x\right\|}\epsilon,
\end{aligned}
$$

which implies $\sup_{\boldsymbol{\xi}\in\boldsymbol{\Xi}}\sup_{\boldsymbol{\xi}'\in\mathcal{A}(\boldsymbol{\xi},\epsilon)}\|\boldsymbol{D}_{n_1}^{-1}(\boldsymbol{\xi})\boldsymbol{D}_{n_1}(\boldsymbol{\xi}')-\boldsymbol{I}_{m+1}\|_F=(m+1)|\pi_{\boldsymbol{\xi}'}/\pi_{\boldsymbol{\xi}}-1|=(m+1)(n_1\pi_{\boldsymbol{\xi}})^{-1/2}(n_1\pi_{\boldsymbol{\xi}})^{1/2}|\pi_{\boldsymbol{\xi}'}-\pi_{\boldsymbol{\xi}}|\to 0$ as $n_1\to\infty$.

For Condition CA(c)(ii), Let $\theta_{i\boldsymbol{\xi}^k}=\exp(\tilde{\boldsymbol{x}}_i^\intercal\boldsymbol{\xi}^k)/\{1+\exp(\tilde{\boldsymbol{x}}_i^\intercal\boldsymbol{\xi}^k)$ and $\pi_{\boldsymbol{\xi}^k}=\sum_{i=1}^{n_1}\theta_{i\boldsymbol{\xi}^k}(1-\theta_{i\boldsymbol{\xi}^k})/n_1$. For any $\boldsymbol{\xi}\in\boldsymbol{\Xi}$, since $(n_1\pi_{\boldsymbol{\xi}^k})^{-1}I_{n_1}(\boldsymbol{\xi}^k)\to\tilde{\boldsymbol{S}}_x$ and $(n_1\pi_{\boldsymbol{\xi}})^{-1}I_{n_1}(\boldsymbol{\xi})\to\tilde{\boldsymbol{S}}_x$ showed as before, we have over the sets $\|\boldsymbol{D}_{n_1}^\intercal(\boldsymbol{\xi})(\boldsymbol{\xi}^k-\boldsymbol{\xi})\|_F\leq\epsilon$, for $1\leq k\leq m+1$, as $n_1\to\infty$,

$$
\begin{aligned}
\left\|(n_1\pi_{\boldsymbol{\xi}})^{-1}\left\{I_{n_1}\left(\boldsymbol{\xi}^k\right)-I_{n_1}\left(\boldsymbol{\xi}\right)\right\}\right\|_F \leq& \left\|\left\{(n_1\pi_{\boldsymbol{\xi}})^{-1}-\left(n_1\pi_{\boldsymbol{\xi}^k}\right)^{-1}\right\}I_{n_1}\left(\boldsymbol{\xi}^k\right)\right\|_F \\
&+\left\|\left(n_1\pi_{\boldsymbol{\xi}^k}\right)^{-1}I_{n_1}\left(\boldsymbol{\xi}^k\right)-(n_1\pi_{\boldsymbol{\xi}})^{-1}I_{n_1}\left(\boldsymbol{\xi}\right)\right\|_F \\
\leq (n_1\pi_{\boldsymbol{\xi}})^{-3/2}(n_1\pi_{\boldsymbol{\xi}})^{1/2}|\pi_{\boldsymbol{\xi}^k}-\pi_{\boldsymbol{\xi}}|&\left\|\left(n_1\pi_{\boldsymbol{\xi}^k}\right)^{-1}I_{n_1}\left(\boldsymbol{\xi}^k\right)\right\|_F+\left\|\left(n_1\pi_{\boldsymbol{\xi}^k}\right)^{-1}I_{n_1}\left(\boldsymbol{\xi}^k\right)-(n_1\pi_{\boldsymbol{\xi}})^{-1}I_{n_1}\left(\boldsymbol{\xi}\right)\right\|_F
\end{aligned}
$$

By the inequalities $\|\tilde{\boldsymbol{S}}_x\|_F\leq\sqrt{m+1}\|\tilde{\boldsymbol{S}}_x\|<\infty$ and $\sqrt{n_1\pi_{\boldsymbol{\xi}}}|\pi_{\boldsymbol{\xi}^k}-\pi_{\boldsymbol{\xi}}|\leq 3\sqrt{2\|\tilde{\boldsymbol{S}}_x\|}\epsilon$, we have $\|(n_1\pi_{\boldsymbol{\xi}})^{-1}\{I_{n_1}(\boldsymbol{\xi}^k)-I_{n_1}(\boldsymbol{\xi})\}\|_F\to 0$.

Thus we have

$$
\begin{aligned}
&\sup_{\boldsymbol{\xi}\in\boldsymbol{\Xi}}\sup_{\boldsymbol{\xi}^k\in\mathcal{A}(\boldsymbol{\xi},\epsilon)}\left\|\boldsymbol{D}_{n_1}^{-1}\left(\boldsymbol{\xi}\right)\left\{\left(I_{n_1}\left(\boldsymbol{\xi}^1\right)_{1.}^\intercal,\ldots,I_{n_1}\left(\boldsymbol{\xi}^{m+1}\right)_{(m+1).}^\intercal\right)-I_{n_1}\left(\boldsymbol{\xi}\right)\right\}\left\{\boldsymbol{D}_{n_1}^{-1}\left(\boldsymbol{\xi}\right)\right\}^\intercal\right\|_F \\
\leq&\sup_{\boldsymbol{\xi}\in\boldsymbol{\Xi}}\sum_{k=1}^{m+1}\sup_{\boldsymbol{\xi}^k\in\mathcal{A}(\boldsymbol{\xi},\epsilon)}\left\|(n_1\pi_{\boldsymbol{\xi}})^{-1}\left\{I_{n_1}\left(\boldsymbol{\xi}^k\right)-I_{n_1}\left(\boldsymbol{\xi}\right)\right\}\right\|_F\to 0.
\end{aligned}
$$

□

Applying Corollary 1 in Sweeting (1980) we have that $I^{1/2}(\boldsymbol{\xi})(\hat{\boldsymbol{\xi}}-\boldsymbol{\xi})\overset{u}{\Rightarrow}\boldsymbol{Z}$ for all $\boldsymbol{\gamma}\in\boldsymbol{\Xi}$. Under Condition CA(a), we have the parameters $\boldsymbol{\gamma}_{\cdot j}\in\boldsymbol{\Xi}$ for $j=1,\ldots,n_2$. Namely, $I^{1/2}(\boldsymbol{\gamma}_{\cdot j})(\hat{\boldsymbol{\gamma}_{\cdot j}}-\boldsymbol{\gamma}_{\cdot j})\overset{u}{\Rightarrow}$

$\boldsymbol{Z}$ which implies $I^{1/2}(\boldsymbol{\gamma}_{\cdot j})(\hat{\boldsymbol{\gamma}}_{\cdot j} - \boldsymbol{\gamma}_{\cdot j}) \xrightarrow{d} \mathcal{N}(0,1)$ for all $j = 1, \ldots, n_2$. For $j = 1, \ldots, n_2$, define $\pi_j = \sum_{i=1}^{n_1} \theta_{ij}(1-\theta_{ij})/n_1$, and $\pi_j^* \in ((1-\theta_U)^2/\theta_L^2, (1-\theta_L)^2/\theta_U^2)$, then we have $\pi_j \in \{\min\{\theta_U(1-\theta_U), \theta_L(1-\theta_L)\}, 1\}$ and $\pi_j^* = \sum_{i=1}^{n_1}\{(1-\theta_{ij})^2/\theta_{ij}^2\}/n_1$. As shown in Justification of Condition CA(b), $(n_1\pi_j)^{-1/2} I^{1/2}(\boldsymbol{\gamma}_{\cdot j}) \to \tilde{\boldsymbol{S}}_x^{\,1/2}$. Thus we have $\sqrt{n_1\pi_j}(\hat{\boldsymbol{\gamma}}_{\cdot j} - \boldsymbol{\gamma}_{\cdot j}) \overset{u}{\Rightarrow} \boldsymbol{Z}_j$, where $\boldsymbol{Z}_j \sim \mathcal{N}(\boldsymbol{0}, \tilde{\boldsymbol{S}}_x^{\,-1})$ for all $1 \le j \le n_2$. For each $j = 1, \ldots, n_2$, $|\hat{\boldsymbol{\gamma}}_{\cdot j} - \boldsymbol{\gamma}_{\cdot j}| = O_p(\boldsymbol{1/\sqrt{n_1\pi_j}})$.

The estimator of θ_{ij} is given by $\hat{\theta}_{ij} = \exp(\tilde{\boldsymbol{x}}_i^{\intercal}\hat{\boldsymbol{\gamma}}_{\cdot j})/\{1+\exp(\tilde{\boldsymbol{x}}_i^{\intercal}\hat{\boldsymbol{\gamma}}_{\cdot j})\}$, thus, for each $j = 1, \ldots, n_2$, $\sup_i |\hat{\theta}_{ij} - \theta_{ij}| = O_p(1/\sqrt{n_1\pi_j})$. Also, we have that for specific $\boldsymbol{\gamma}_{\cdot j}^* \in \mathcal{B}^{m+1}(\boldsymbol{\gamma}_{\cdot j}, d(\boldsymbol{\gamma}_{\cdot j}, \hat{\boldsymbol{\gamma}}_{\cdot j}))$,

$$\frac{1}{\hat{\theta}_{ij}} - \frac{1}{\theta_{ij}} = -\frac{1}{\theta_{ij}^2}\left(\frac{\partial\theta_{ij}}{\partial\boldsymbol{\gamma}_{\cdot j}}\right)^{\intercal}(\hat{\boldsymbol{\gamma}}_{\cdot j} - \boldsymbol{\gamma}_{\cdot j}) + (\hat{\boldsymbol{\gamma}}_{\cdot j} - \boldsymbol{\gamma}_{\cdot j})^{\intercal}\frac{\partial^2(1/\theta_{ij})}{\partial\boldsymbol{\gamma}_{\cdot j}^2}\Big|_{\boldsymbol{\gamma}_{\cdot j}^*}(\hat{\boldsymbol{\gamma}}_{\cdot j} - \boldsymbol{\gamma}_{\cdot j}),$$

which can simplify to be

$$\frac{1}{\hat{\theta}_{ij}} - \frac{1}{\theta_{ij}} = -\frac{(1-\theta_{ij})}{\theta_{ij}}\tilde{\boldsymbol{x}}_i^{\intercal}(\hat{\boldsymbol{\gamma}}_{\cdot j} - \boldsymbol{\gamma}_{\cdot j}) + (\hat{\boldsymbol{\gamma}}_{\cdot j} - \boldsymbol{\gamma}_{\cdot j})^{\intercal}\frac{\left(1-\theta_{ij}^*\right)}{\theta_{ij}^*}\tilde{\boldsymbol{x}}_i\tilde{\boldsymbol{x}}_i^{\intercal}(\hat{\boldsymbol{\gamma}}_{\cdot j} - \boldsymbol{\gamma}_{\cdot j}).$$

Since there exist a positive constant a_x such that $\|\boldsymbol{X}\|_\infty < a_x$, we have $\|\tilde{\boldsymbol{x}}_i\tilde{\boldsymbol{x}}_i^{\intercal}\|_\infty < \infty$. Also $\boldsymbol{\gamma}_{\cdot j}^* \in B(\boldsymbol{\gamma}_{\cdot j}, d(\boldsymbol{\gamma}_{\cdot j}, \hat{\boldsymbol{\gamma}}_{\cdot j}))$ and $\|\boldsymbol{X}\|_\infty < a_x$ implies $\theta_{ij}^* \to \theta_{ij}$, as $n_1 \to \infty$. Namely, $(1-\theta_{ij}^*)/\theta_{ij}^* \to (1-\theta_{ij})/\theta_{ij}$, as $n_1 \to \infty$. Once $\theta_{ij} \neq 0$ and $\tilde{\boldsymbol{x}}_i \neq \boldsymbol{0}$, by Taylor expansion and continuous mapping theorem, we can see that:

$$\left(\frac{1}{\hat{\theta}_{ij}} - \frac{1}{\theta_{ij}}\right)^2 = \frac{(1-\theta_{ij})^2}{\theta_{ij}^2}(\hat{\boldsymbol{\gamma}}_{\cdot j} - \boldsymbol{\gamma}_{\cdot j})^{\intercal}\tilde{\boldsymbol{x}}_i\tilde{\boldsymbol{x}}_i^{\intercal}(\hat{\boldsymbol{\gamma}}_{\cdot j} - \boldsymbol{\gamma}_{\cdot j}) + o\left((\hat{\boldsymbol{\gamma}}_{\cdot j} - \boldsymbol{\gamma}_{\cdot j})^{\intercal}(\hat{\boldsymbol{\gamma}}_{\cdot j} - \boldsymbol{\gamma}_{\cdot j})\right),$$

for $i = 1, \ldots, n_1$, $j = 1, \ldots, n_2$. As $n_1 \to \infty$, we have

$$\sum_{i=1}^{n_1}\frac{(1-\theta_{ij})^2}{\theta_{ij}^2}\tilde{\boldsymbol{x}}_i\tilde{\boldsymbol{x}}_i^{\intercal}/\left(n_1\pi_j^*\right) \to \tilde{\boldsymbol{S}}_x.$$

By Slutsky theorem,

$$\frac{\pi_j}{\pi_j^*}\sum_{i=1}^{n_1}\left(\frac{1}{\hat{\theta}_{ij}} - \frac{1}{\theta_{ij}}\right)^2 \overset{u}{\Rightarrow} \boldsymbol{Z}_j^{\intercal}\left(\tilde{\boldsymbol{S}}_x^{\,-1}\right)^{-1}\boldsymbol{Z}_j^{\intercal},$$

which implies that $\pi_j\pi_j^{*-1}\sum_{i=1}^{n_1}(1/\hat{\theta}_{ij} - 1/\theta_{ij})^2 \overset{u}{\Rightarrow} \boldsymbol{U}_j$, where $\boldsymbol{U}_j \sim \chi_{m+1}^2$ for all $j = 1, \ldots, n_2$.

By using Polya's theorem, we have for any $t > \eta_g^{-1}(m+1)$, let $\eta_g = \min\{\pi_j/\pi_j^*\}$, $k_{n_1} = \max_j \sup_t |\mathsf{P}(\sum_i(1/\hat{\theta}_{ij} - 1/\theta_{ij})^2 \ge t) - \mathsf{P}(\chi_{m+1}^2 \ge \eta_g t)| \le 1/n_2^2$ there exists a positive integer N_{1/n_2^2},

for $n_1 \geq N_{1/n_2^2}$,

$$\sup_j \mathsf{P}\left\{\sum_{i=1}^{n_1}\left(\frac{1}{\hat{\theta}_{ij}}-\frac{1}{\theta_{ij}}\right)^2 \geq t\right\} \leq \sup_j \mathsf{P}\left\{\frac{\pi_j}{\pi_j^*}\sum_{i=1}^{n_1}\left(\frac{1}{\hat{\theta}_{ij}}-\frac{1}{\theta_{ij}}\right)^2 \geq \eta_g t\right\}$$

$$\leq \sup_j \left\{\mathsf{P}\left(\chi^2_{m+1} \geq \eta_g t\right)\right\} + k_{n_1} \leq \left[\frac{\eta_g t}{m+1}\exp\left\{1-\frac{\eta_g t}{m+1}\right\}\right]^{\frac{m+1}{2}} + k_{n_1}.$$

Take $c_{n_1,n_2} = n_2\log(n_2)/\eta_g$ and $t_0 = (m+3)$, for $t > t_0$, we have

$$\mathsf{P}\left\{\sum_{j=1}^{n_2}\sum_{i=1}^{n_1}\left(\frac{1}{\hat{\theta}_{ij}}-\frac{1}{\theta_{ij}}\right)^2 \geq c_{n_1,n_2}t\right\} \leq \sum_{j=1}^{n_2}\mathsf{P}\left\{\sum_{i=1}^{n_1}\left(\frac{1}{\hat{\theta}_{ij}}-\frac{1}{\theta_{ij}}\right)^2 \geq \frac{c_{n_1,n_2}t}{n_2}\right\}$$

$$= \sum_{j=1}^{n_2}\mathsf{P}\left\{\chi^2_{m+1} \geq \frac{\eta_g c_{n_1,n_2}t}{n_2}\right\} \leq \sum_{j=1}^{n_2}\left[\frac{\eta_g c_{n_1,n_2}t}{n_2(m+1)}\exp\left\{1-\frac{\eta_g c_{n_1,n_2}t}{n_2(m+1)}\right\}\right]^{\frac{m+1}{2}} + n_2 k_{n_1}$$

$$\leq (m+1)^{-(m+1)/2}\exp\left\{\frac{m+1}{2}-\frac{\eta_g c_{n_1,n_2}t}{2n_2}+\log(t)+\frac{m+1}{2}\log\left(\frac{\eta_g c_{n_1,n_2}}{n_2}\right)+\log(n_2)\right\} + n_2 k_{n_1}$$

$$\leq (m+1)^{-(m+1)/2}\exp\left\{\frac{m+1}{2}-\frac{t\log(n_2)}{2}+\log(t)+\frac{m+3}{2}\log(n_2)\right\} + n_2 k_{n_1}$$

$$\leq (m+1)^{-(m+1)/2}\exp\left\{m+2-\frac{t}{2}+\log(t)\right\} + n_2 k_{n_1}.$$

Let $g(t) = (m+1)^{-(m+1)/2}\exp\{m+2-t/2+\log(t)\}$ and $h_{n_1,n_2} = n_2 k_{n_1} \leq 1/n_2$ in Condition C5(b), we have $\lim_{t\to\infty} g(t) = 0$ and $\lim_{n_1,n_2\to\infty} h_{n_1,n_2} = 0$. It satisfies the requirements.

S4 Lemmas and Proofs

In this section, we provide various results required in the proofs of Theorems 1 and 2, as well as Corollaries 1 and 2. First, we review some basic facts about matrices which will be useful in the following development. For any $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{n_1\times n_2}$, we have

- Trace Duality Property:

$$|\mathrm{tr}(\boldsymbol{A}^\intercal\boldsymbol{B})| \leq \|\boldsymbol{B}\|\,\|\boldsymbol{A}\|_*. \tag{S4.1}$$

- Norm Inequalities:

$$\|\boldsymbol{A}\|_F \leq \|\boldsymbol{A}\|_* \leq \sqrt{r_{\boldsymbol{A}}}\,\|\boldsymbol{A}\|_F \quad \text{and} \quad \|\boldsymbol{A}\| \leq \|\boldsymbol{A}\|_F \leq \sqrt{r_{\boldsymbol{A}}}\,\|\boldsymbol{A}\|, \tag{S4.2}$$

where $r_{\boldsymbol{A}}$ is the rank of matrix $\boldsymbol{A}$.

Write $\boldsymbol{J}_{ij} = \boldsymbol{e}_i(n_1)\boldsymbol{e}_j^\intercal(n_2)$, where $\boldsymbol{e}_i(n) \in \mathbb{R}^n$ is the standard basis vector with the i-th element being 1 and the rest being 0. Now we present several lemmas.

Lemma S4.1. *Let $\Psi^{(1)} = \sum_{ij} \omega_{ij}\epsilon_{ij}\boldsymbol{J}_{ij}/(n_1 n_2 \hat{\theta}_{ij})$. Under Conditions C1, C4 and C5, for some positive constants c_σ, η, δ_σ and all $t > t_0$, there exists $\Delta^{(1)}(\delta_\sigma, t)$ such that*

$$\left\|\Psi^{(1)}\right\| \leq \Delta^{(1)}(\delta_\sigma, t) \asymp \max\left\{\frac{\sqrt{(n_1 \vee n_2)\log(n)}}{\sqrt{\theta_L} n_1 n_2}, (n_1 n_2)^{-3/4}\left(c_{n_1,n_2} t\right)^{1/2}\log^{\delta_\sigma/4}(n)\right\}$$

holds with probability at least $1 - 1/n - g(t) - h_{n_1,n_2} - 12c_\sigma^2\eta^2\log^{-\delta_\sigma}(n)$.

More specifically, for the uniform missingness, we have $\theta_{ij} \equiv \theta_0$ and $\hat{\theta}_{ij} \equiv N/(n_1 n_2)$ and for some positive constants δ_σ and C_1 such that

$$\left\|\Psi^{(1)}\right\| \leq C_1 \frac{\sqrt{(n_1 \vee n_2)\log(n)}}{\sqrt{\theta_0} n_1 n_2}$$

holds with probability at least $1 - 1/n - \log^{-\delta_\sigma}(n) - 2/(n_1 \vee n_2)$.

To prove Lemma S4.1, we apply Theorem 6.2 which is matrix Bernstein inequality for the sub-exponential case provided by Tropp (2012).

Proof of Lemma S4.1. For any rectangular matrix $\boldsymbol{M}$, let $\mathcal{L}(\boldsymbol{M})$ be the self-adjoint dilation of $\boldsymbol{M}$ defined as

$$\mathcal{L}(\boldsymbol{M}) := \begin{bmatrix} 0 & \boldsymbol{M} \\ \boldsymbol{M}^\intercal & 0 \end{bmatrix}.$$

In our case, for $i = 1, \ldots, n_1$, $j = 1, \ldots, n_2$, let

$$\boldsymbol{G}_{n_2(i-1)+j} = \mathcal{L}\left(\frac{\epsilon_{ij}\omega_{ij}}{\theta_{ij}}\boldsymbol{J}_{ij}\right) \quad \text{and} \quad \boldsymbol{H}_{n_2(i-1)+j} = \mathcal{L}\left(\frac{c_\sigma}{\sqrt{\theta_L}}\boldsymbol{J}_{ij}\right).$$

To apply Theorem 6.2 of Tropp (2012), we verify the conditions needed in the following.

Since ϵ_{ij} is independent of ω_{ij}, we have

$$\mathsf{E}\left(\frac{\epsilon_{ij}\omega_{ij}}{\theta_{ij}}\boldsymbol{J}_{ij}\right) = \mathsf{E}(\epsilon_{ij})\,\mathsf{E}\left(\frac{\omega_{ij}\boldsymbol{J}_{ij}}{\theta_{ij}}\right) = \boldsymbol{0},$$

which implies $\mathsf{E}\left(\boldsymbol{G}_{n_2(i-1)+j}\right) = \boldsymbol{0}$. Write $\eta_H = \eta/\theta_L$, where η is the constant in Condition C1. Now we want to show that

$$\mathsf{E}\left\{\mathcal{L}\left(\frac{\epsilon_{ij}\omega_{ij}}{\theta_{ij}}\boldsymbol{J}_{ij}\right)^l\right\} \leq \frac{l!}{2}\cdot\eta_H^{l-2}\mathcal{L}\left(\frac{c_\sigma}{\sqrt{\theta_L}}\boldsymbol{J}_{ij}\right)^2 \quad \text{for } l = 2, 3, 4, \ldots. \tag{S4.3}$$

In our case, under Condition C1 and C4, for a finite constant η, we have

$$\mathsf{E}\left|\frac{\epsilon_{ij}\omega_{ij}}{\theta_{ij}}\right|^l = \frac{\mathsf{E}\left|\epsilon_{ij}\right|^l \mathsf{E}\omega_{ij}}{\theta_{ij}^l} \leq \frac{\max \mathsf{E}\left|\epsilon_{ij}\right|^l}{\theta_L^{l-1}} \leq \frac{1}{2} l! \left(\frac{c_\sigma}{\sqrt{\theta_L}}\right)^2 \left(\frac{\eta}{\theta_L}\right)^{l-2}, \; l = 2, 3, \ldots.$$

Thus it suffices to show that $\mathcal{L}^l(\boldsymbol{J}_{ij}) \leq \mathcal{L}^2(\boldsymbol{J}_{ij})$ for $l = 2, 3, 4, \ldots$.

Let $\boldsymbol{K}_{n,i} = \boldsymbol{e}_i(n)\boldsymbol{e}_i^\intercal(n)$, where $\boldsymbol{e}_i(n) \in \mathbb{R}^n$ is the standard basis vector of $\mathbb{R}^n$ with the i-th element being 1 and the rest being 0. By the properties of $\boldsymbol{J}_{ij}$, it is not hard to show that for $l = 2s$ or $2s+1$, we have

$$\mathcal{L}^{2s}\left(\boldsymbol{J}_{ij}\right) = \begin{bmatrix} \boldsymbol{K}_{n_1,i} & 0 \\ 0 & \boldsymbol{K}_{n_2,j} \end{bmatrix} \quad \text{and} \quad \mathcal{L}^{2s+1}\left(\boldsymbol{J}_{ij}\right) = \begin{bmatrix} 0 & \boldsymbol{J}_{ij} \\ \boldsymbol{J}_{ij}^\intercal & 0 \end{bmatrix} = \mathcal{L}\left(\boldsymbol{J}_{ij}\right).$$

Hence (S4.3) is verified as $\begin{bmatrix} \boldsymbol{K}_{n_1,i} & -\boldsymbol{J}_{ij} \\ -\boldsymbol{J}_{ij}^\intercal & \boldsymbol{K}_{n_2,j} \end{bmatrix} \geq 0$.

Set the constant $\sigma_H^2 = \|\sum_{ij} \mathcal{L}(c_\sigma \boldsymbol{J}_{ij}/\sqrt{\theta_L})^2\| = c_\sigma^2 \|\sum_{ij} \mathcal{L}(\boldsymbol{J}_{ij})^2\|/\theta_L$. Since

$$\left\|\sum_{ij} \mathcal{L}\left(\boldsymbol{J}_{ij}\right)^2\right\| = \left\|\begin{bmatrix} \sum_{ij} \boldsymbol{K}_{n_1,i} & 0 \\ 0 & \sum_{ij} \boldsymbol{K}_{n_2,j} \end{bmatrix}\right\| = \max\left\{\left\|\sum_{ij} \boldsymbol{K}_{n_1,i}\right\|, \left\|\sum_{ij} \boldsymbol{K}_{n_2,j}\right\|\right\}$$

$$= \max\left\{\left\|n_2 \boldsymbol{I}_{n_1}\right\|, \left\|n_1 \boldsymbol{I}_{n_2}\right\|\right\} = n_1 \vee n_2,$$

we have $\sigma_H^2 = c_\sigma^2(n_1 \vee n_2)/\theta_L$. By the property of dilation (2.12) of Tropp (2012),

$$\mathsf{P}\left[\lambda_{\max}\left\{\sum_{ij} \mathcal{L}\left(\frac{\epsilon_{ij}\omega_{ij}}{\theta_{ij}}\boldsymbol{J}_{ij}\right)\right\} \geq t\right] = \mathsf{P}\left(\left\|\sum_{ij} \frac{\epsilon_{ij}\omega_{ij}}{\theta_{ij}}\boldsymbol{J}_{ij}\right\| \geq t\right).$$

By the Matrix Bernstein Inequality in Theorem 6.2 of Tropp (2012), we show that, for all $t_1 > 0$,

$$\mathsf{P}\left(\left\|\sum_{ij} \frac{\epsilon_{ij}\omega_{ij}}{\theta_{ij}}\boldsymbol{J}_{ij}\right\| \geq t_1\right) \leq n \cdot \exp\left\{\frac{-t_1^2/2}{c_\sigma^2\left(n_1 \vee n_2\right)/\theta_L + \eta_H t_1}\right\}$$

$$\leq \begin{cases} n \cdot \exp\left\{\frac{-t_1^2}{4c_\sigma^2(n_1 \vee n_2)/\theta_L}\right\} & \text{for } t_1 \leq c_\sigma^2\left(n_1 \vee n_2\right)/\left(\theta_L \eta_H\right) \\ n \cdot \exp\left\{\frac{-t_1}{4\eta_H}\right\} & \text{for } t_1 \geq c_\sigma^2\left(n_1 \vee n_2\right)/\left(\theta_L \eta_H\right) \end{cases}$$

In other words, for any $s_1 > 0$, with probability at least $1 - \exp\{-s_1\}$, we have

$$\left\|\sum_{ij} \frac{\epsilon_{ij}\omega_{ij}}{\theta_{ij}}\boldsymbol{J}_{ij}\right\| \leq \max\left\{2c_\sigma\sqrt{\frac{(n_1 \vee n_2)\left\{s_1 + \log\left(n\right)\right\}}{\theta_L}}, 4\eta_H\left\{s_1 + \log\left(n\right)\right\}\right\}$$

where $s_1' = s_1 + \log(n)$. Choose $s_1 = \log(n)$, i.e, $s_1' = 2\log(n)$. With probability at least $1 - 1/n$, we have

$$\frac{1}{n_1 n_2}\left\|\sum_{ij}\frac{\epsilon_{ij}\omega_{ij}}{\theta_{ij}}\boldsymbol{J}_{ij}\right\| \leq \frac{2c_\sigma\sqrt{2\left(n_1 \vee n_2\right)\log\left(n\right)}}{\sqrt{\theta_L}n_1 n_2} := \Delta^{(1)'}.$$

We also know that

$$\begin{aligned}\left\|\sum_{ij}\epsilon_{ij}\omega_{ij}\left(\frac{1}{\hat{\theta}_{ij}}-\frac{1}{\theta_{ij}}\right)\boldsymbol{J}_{ij}\right\|^2 \leq& \left\|\sum_{ij}\epsilon_{ij}\omega_{ij}\left(\frac{1}{\hat{\theta}_{ij}}-\frac{1}{\theta_{ij}}\right)\boldsymbol{J}_{ij}\right\|_F^2 = \sum_{ij}\epsilon_{ij}^2\omega_{ij}^2\left(\frac{1}{\hat{\theta}_{ij}}-\frac{1}{\theta_{ij}}\right)^2 \\ \leq& \sum_{ij}\epsilon_{ij}^2\left(\frac{1}{\hat{\theta}_{ij}}-\frac{1}{\theta_{ij}}\right)^2 \leq \max\epsilon_{ij}^2\sum_{ij}\left(\frac{1}{\hat{\theta}_{ij}}-\frac{1}{\theta_{ij}}\right)^2.\end{aligned}$$

Due to Markov inequality, under Condition C1, we have for any $a > 0$,

$$\mathsf{P}\left(\max\epsilon_{ij}^2 \geq a\right) = \mathsf{P}\left(\max\epsilon_{ij}^4 \geq a^2\right) \leq \frac{\sum_{ij}\mathsf{E}\epsilon_{ij}^4}{a^2} \leq \frac{12n_1n_2c_\sigma^2\eta^2}{a^2}.$$

Take $a = (n_1n_2)^{1/2}\log^{\delta_\sigma/2}(n)$ for a positive constant δ_σ, we have $\max\epsilon_{ij}^2 \leq (n_1n_2)^{1/2}\log^{\delta_\sigma/2}(n)$ with probability at least $1 - 12c_\sigma^2\eta^2\log^{-\delta_\sigma}(n)$.

Combining with Condition C5(b), we have for $t > t_0$, with probability at least $1-g(t)-h_{n_1,n_2}-12c_\sigma^2\eta^2\log^{-\delta_\sigma}(n)$, $\|\sum_{ij}\epsilon_{ij}\omega_{ij}(1/\hat{\theta}_{ij} - 1/\theta_{ij})\boldsymbol{J}_{ij}\| \leq (n_1n_2)^{1/4}(c_{n_1,n_2}t)^{1/2}\log^{\delta_\sigma/4}(n)$.

Then for $t > t_0$, with probability at least $1 - 1/n - g(t) - h_{n_1,n_2} - 12c_\sigma^2\eta^2\log^{-\delta_\sigma}(n)$, we have

$$\begin{aligned}&\frac{1}{n_1n_2}\left\|\sum_{ij}\frac{\epsilon_{ij}\omega_{ij}}{\hat{\theta}_{ij}}\boldsymbol{J}_{ij}\right\| \leq \frac{1}{n_1n_2}\left\|\sum_{ij}\frac{\epsilon_{ij}\omega_{ij}}{\theta_{ij}}\boldsymbol{J}_{ij}\right\| + \frac{1}{n_1n_2}\left\|\sum_{ij}\epsilon_{ij}\omega_{ij}\left(\frac{1}{\hat{\theta}_{ij}}-\frac{1}{\theta_{ij}}\right)\boldsymbol{J}_{ij}\right\| \\ \leq&\Delta^{(1)'} + \left(n_1n_2\right)^{-3/4}\left(c_{n_1,n_2}t\right)^{1/2}\log^{\delta_\sigma/4}\left(n\right) \\ :=&\Delta^{(1)}\left(\delta_\sigma, t\right) \asymp \max\left\{\frac{\sqrt{\left(n_1 \vee n_2\right)\log\left(n\right)}}{\sqrt{\theta_L}n_1n_2}, \left(n_1n_2\right)^{-3/4}\left(c_{n_1,n_2}t\right)^{1/2}\log^{\delta_\sigma/4}\left(n\right)\right\}.\end{aligned}$$

As for the uniform missingness, for the first term without the estimators $\hat{\theta}_{ij}$, we have the same upper bound. We also know that for the second term,

$$\begin{aligned}\mathsf{E}\left\|\sum_{ij}\epsilon_{ij}\omega_{ij}\left(\frac{1}{\hat{\theta}_{ij}}-\frac{1}{\theta_{ij}}\right)\boldsymbol{J}_{ij}\right\|^2 \leq& \mathsf{E}\left\{\sum_{ij}\epsilon_{ij}^2\omega_{ij}^2\left(\frac{1}{\hat{\theta}_{ij}}-\frac{1}{\theta_{ij}}\right)^2\right\} \leq c_\sigma^2\mathsf{E}\left\{\sum_{ij}\omega_{ij}\left(\frac{n_1n_2}{N}-\frac{1}{\theta_0}\right)^2\right\} \\ =&c_\sigma^2\mathsf{E}\left\{N\left(\frac{n_1n_2}{N}-\frac{1}{\theta_0}\right)^2\right\} = c_\sigma^2\left(n_1n_2\right)^2\mathsf{E}\left\{\frac{1}{N}-\frac{1}{n_1n_2\theta_0}\right\}.\end{aligned}$$

Also $\mathsf{E}(N) = n_1 n_2 \theta_0$ and Taylor expansions for the moments of functions of random variables implies that $\mathsf{E}(1/N) = 1/(\theta_0 n_1 n_2) + 1/(\theta_0 n_1 n_2)^3 \mathsf{Var}(N)(1+o(1)) = 1/(\theta_0 n_1 n_2) + (1-\theta_0)/(\theta_0 n_1 n_2)^2 (1+o(1))$ due to the fact that $\mathsf{E}(N - (n_1 n_2)\theta_0)^4 = o(\mathsf{Var}(N))$. We have $\mathsf{E}\|\epsilon_{ij}\omega_{ij}(n_1 n_2/N - 1/\theta_0)\boldsymbol{J}_{ij}\| \leq 2c_\sigma^2(1-\theta_0)/\theta_0^2$.

Due to Markov inequality, we have for $0 < \delta_\sigma < 2$, $\|\epsilon_{ij}\omega_{ij}(n_1 n_2/N - 1/\theta_0)\boldsymbol{J}_{ij}\| \leq c_\sigma^2(1-\theta_0)\log^{\delta_\sigma}(n)/\theta_0^2 \leq c_\sigma^2 \log^{\delta_\sigma}(n)/\theta_0^2$ with probability at least $1 - 2\log^{-\delta_\sigma}(n)$. Since $n_1 n_2 \theta_0 > (n_1 \vee n_2)\log(n)$, we have $\log^{\delta_\sigma}(n)/\theta_0^2 < (n_1 \vee n_2)\log(n)/\theta_0$. Then we have under the uniform missingness, for a positive constant C_1,

$$\left\|\Psi^{(1)}\right\| \leq C_1 \frac{\sqrt{(n_1 \vee n_2)\log(n)}}{\sqrt{\theta_0} n_1 n_2}$$

holds with probability at least $1 - 1/n - 2\log^{-\delta_\sigma}(n)$.

□

Lemma S4.2. *Let* $\Psi^{(2)} = \sum_{ij} A_{0ij}(\omega_{ij}/\theta_{ij} - 1)\boldsymbol{J}_{ij}/(n_1 n_2)$. *Under Conditions C3-C5, there exists* $\Delta^{(2)}$ *such that*

$$\left\|\Psi^{(2)}\right\| \leq \Delta^{(2)} \asymp \frac{\sqrt{|1/\theta_L - 1|\,(n_1 \vee n_2)\log(n)}}{n_1 n_2}$$

holds with probability at least $1 - 1/n$.

To prove Lemma S4.2, we utilize Proposition 1 given by Koltchinskii et al. (2011) as an immediate consequence of the Matrix Bernstein Inequality due to Ahlswede and Winter (2002) and Tropp (2012). For matrix $\boldsymbol{A}_0$, define that:

$$|\boldsymbol{A}_0|^* := \max\left\{ \sqrt{\frac{\max_{1\leq i\leq n_1} \sum_{j=1}^{n_2} |1/\theta_{ij} - 1| A_{0,ij}^2}{n_1 n_2}}, \sqrt{\frac{\max_{1\leq j\leq n_2} \sum_{i=1}^{n_1} |1/\theta_{ij} - 1| A_{0,ij}^2}{n_1 n_2}} \right\}. \tag{S4.4}$$

Proof of Lemma S4.2. Let $\boldsymbol{M}_{n_2(i-1)+j} = A_{0ij}(\omega_{ij}/\theta_{ij} - 1)\boldsymbol{J}_{ij}$. Under Conditions C4 and C5, it is easy to show that $\max_k \|\boldsymbol{M}_k\| \leq \max\{1/\theta_{ij} - 1, 1\}\|\boldsymbol{A}_0\|_\infty \leq \max\{1/\theta_L - 1, 1\}\|\boldsymbol{A}_0\|_\infty$ and

$$\sigma_M = \max\left\{ \frac{1}{n_1 n_2}\left\|\sum_k \mathsf{E}(\boldsymbol{M}_k \boldsymbol{M}_k^\mathsf{T})\right\|^{1/2}, \frac{1}{n_1 n_2}\left\|\sum_k \mathsf{E}(\boldsymbol{M}_k^\mathsf{T} \boldsymbol{M}_k)\right\|^{1/2} \right\} \leq |\boldsymbol{A}_0|^*.$$

Take $U_M = \max\{1/\theta_L - 1, 1\}\|\boldsymbol{A}_0\|_\infty$. By Proposition 1 of Koltchinskii et al. (2011), we have, for all $t > 0$,

$$\left\|\Psi^{(2)}\right\| \le 2\max\left\{|\boldsymbol{A}_0|^* \sqrt{\frac{t+\log(n)}{n_1 n_2}}, \max\left\{\frac{1}{\theta_L} - 1, 1\right\}\|\boldsymbol{A}_0\|_\infty \frac{t+\log(n)}{n_1 n_2}\right\}$$

with probability at least $1 - \exp\{-t\}$.

According to (S4.4), under Conditions C3 and C5, we have

$$|\boldsymbol{A}_0|^* \le \sqrt{\frac{|1/\theta_L - 1|}{n_1 n_2}} \max\left\{\|\boldsymbol{A}_0\|_{\infty,2}, \|\boldsymbol{A}_0^\intercal\|_{\infty,2}\right\} \le a_2 \sqrt{\frac{|1/\theta_L - 1|}{n_1 \wedge n_2}}.$$

Under additional Condition C3 and $t = \log(n)$, with probability at least $1 - 1/n$,

$$\left\|\Psi^{(2)}\right\| \le 2(a_1 \vee a_2)\max\left\{\sqrt{\frac{2|1/\theta_L - 1|\log(n)}{(n_1 \wedge n_2) n_1 n_2}}, 2\max\left\{\frac{1}{\theta_L} - 1, 1\right\}\frac{\log^{3/2}(n)}{n_1 n_2}\right\} := \Delta^{(2)},$$

for some positive constants a_1 and a_2 defined in Condition C3.

Since $(n_1 n_2)^{-1}\log^{3/2}(n) = o\{(n_1 \vee n_2)^{1/2}(n_1 n_2)^{-1}\log^{1/2}(n)\}$ and $\sqrt{|1/\theta_L - 1|} = o(\max\{1/\theta_L - 1, 1\})$ when $\theta_L = o(1)$, we have $\Delta^{(2)} \asymp \sqrt{|1/\theta_L - 1|}(n_1 \vee n_2)^{1/2}(n_1 n_2)^{-1}\log^{1/2}(n)$. □

Lemma S4.3. *Let $\Psi^{(3)} = \sum_{ij} A_{0ij}(\omega_{ij}/\hat{\theta}_{ij} - \omega_{ij}/\theta_{ij})\boldsymbol{J}_{ij}/(n_1 n_2)$. Under Conditions C3 and C5, for all $t > t_0$, there exists $\Delta^{(3)}(t)$ such that*

$$\left\|\Psi^{(3)}\right\| \le \Delta^{(3)}(t) \asymp \frac{\sqrt{c_{n_1,n_2} t\log(n)}}{n_1 n_2}$$

holds with probability at least $1 - g(t) - h_{n_1,n_2}$.

More specifically, for the uniform missingness, we have $\theta_{ij} \equiv \theta_0$ and $\hat{\theta}_{ij} \equiv N/(n_1 n_2)$ and for $0 < \delta_\sigma < 2$, such that

$$\left\|\Psi^{(3)}\right\| \le \frac{\sqrt{2(n_1 \vee n_2)\log(n)}a_1}{\sqrt{\theta_0} n_1 n_2}$$

holds with probability at least $1 - 2\log^{-\delta_\sigma}(n)$.

Proof of Lemma S4.3. By the inequality (S4.2), we have

$$\left\|\Psi^{(3)}\right\| \le \frac{1}{n_1 n_2}\left\|\boldsymbol{W} \circ \hat{\boldsymbol{\Theta}}^* \circ \boldsymbol{A}_0 - \boldsymbol{W} \circ \boldsymbol{\Theta}^* \circ \boldsymbol{A}_0\right\|_F$$

$$= \frac{1}{n_1 n_2}\sqrt{\sum_{ij} A_{0ij}^2 \omega_{ij}^2 \left(\frac{1}{\hat{\theta}_{ij}} - \frac{1}{\theta_{ij}}\right)^2} \le \frac{\|\boldsymbol{A}_0\|_\infty}{n_1 n_2}\sqrt{\sum_{ij}\left(\frac{1}{\hat{\theta}_{ij}} - \frac{1}{\theta_{ij}}\right)^2}.$$

Under Condition C5, $\sqrt{\sum_{ij}(1/\hat{\theta}_{ij}-1/\theta_{ij})^2} \le \sqrt{c_{n_1,n_2}t}$ with probability at least $1-g(t)-h_{n_1,n_2}$. It implies that under Condition C3, with probability at least $1-g(t)-h_{n_1,n_2}$,

$$\begin{aligned}\left\|\Psi^{(3)}\right\| \le & \frac{\sqrt{c_{n_1,n_2}t\log(n)}a_1}{n_1n_2} \le \frac{\sqrt{c_{n_1,n_2}t\log(n)}\,(a_1\vee a_2)}{n_1n_2} \\ :=& \Delta^{(3)}(t) \asymp \frac{\sqrt{c_{n_1,n_2}t\log(n)}}{n_1n_2}.\end{aligned}$$

Since $(n_1n_2)^{-1}\log^{1/2}(n) = o((n_1n_2)^{-3/4}\log^{\delta_\sigma/4}(n))$, we have $\Delta^{(3)}(t) = o(\Delta^{(1)}(\delta_\sigma,t))$.

As for the uniform missingness, similarly as the proof in Lemma S4.1, we have that $\mathsf{E}\{1/N - 1/(n_1n_2\theta_0)\} \le 2(1-\theta_0)/(\theta_0 n_1 n_2)^2$. Then for $0<\delta_\sigma<2$, with probability at least $1-2\log^{-\delta_\sigma}(n)$, $\|\omega_{ij}(n_1n_2/N - 1/\theta_0)\boldsymbol{J}_{ij}\| \le 2(1-\theta_0)\log^{\delta_\sigma}(n)/\theta_0^2 \le 2\log^{\delta_\sigma}(n)/\theta_0^2 \le 2(n_1\vee n_2)\log(n)/\theta_0$ for $n_1n_2\theta_0 > (n_1\vee n_2)\log(n)$. Thus it is not hard to conclude that, for $0<\delta_\sigma<2$, with probability at least $1-2\log^{-\delta_\sigma}(n)$,

$$\left\|\Psi^{(3)}\right\| \le \frac{\sqrt{2\,(n_1\vee n_2)\log(n)}a_1}{\sqrt{\theta_0}n_1n_2}.$$

□

S5 Proofs of Theorem 1 and Corollary 1

Proof of Theorem 1. Under Conditions C1 and C3-C5, Lemmas S4.1-S4.3 show that there exist constants $\Delta^{(1)}(\delta_\sigma,t)$, $\Delta^{(2)}$ and $\Delta^{(3)}(t)$ such that

$$\left\|\Psi^{(1)}\right\| \le \Delta^{(1)}(\delta_\sigma,t), \quad \left\|\Psi^{(2)}\right\| \le \Delta^{(2)}, \quad \left\|\Psi^{(3)}\right\| \le \Delta^{(3)}(t),$$

with probability at least $1-1/n-g(t)-h_{n_1,n_2}-12c_\sigma^2\eta^2\log^{-\delta_\sigma}(n)$, $1-1/n$ and $1-g(t)-h_{n_1,n_2}$ respectively. As defined in (4.2), $\Delta(\delta_\sigma,t) = \max\{\theta_L^{-1/2}(n_1\vee n_2)^{1/2}(n_1n_2)^{-1}\log^{1/2}(n), (n_1n_2)^{-3/4}(c_{n_1,n_2}t)^{1/2}\log^{\delta_\sigma/4}(n)\}$ We have for a positive constant C_0, $\Delta^{(1)}(\delta_\sigma,t)+\Delta^{(2)}+\Delta^{(3)}(t) \le C_0\Delta(\delta_\sigma,t)$.

It follows from the closed form of $\hat{\boldsymbol{\beta}}$ that

$$\begin{aligned}\boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{X}\boldsymbol{\beta}_0 =& \boldsymbol{X}(n_1^{-1}\boldsymbol{X}^\intercal\boldsymbol{X} + n_2\lambda_1\boldsymbol{I}_{m\times m})^{-1}n_1^{-1}\boldsymbol{X}^\intercal\left(\boldsymbol{W}\circ\hat{\boldsymbol{\Theta}}^*\circ\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}_0\right) \\ & - \boldsymbol{X}(n_1^{-1}\boldsymbol{X}^\intercal\boldsymbol{X} + n_2\lambda_1\boldsymbol{I}_{m\times m})^{-1}n_2\lambda_1n_1^{-1}\boldsymbol{X}\boldsymbol{\beta}_0.\end{aligned}$$

Take $\lambda_1 = o(n_2^{-1})$, $n_2\lambda_1 = o(1)$, we have $\boldsymbol{X}(n_1^{-1}\boldsymbol{X}^\intercal\boldsymbol{X} + n_2\lambda_1\boldsymbol{I}_{m\times m})^{-1}n_1^{-1}\boldsymbol{X}^\intercal = \boldsymbol{P}_{\boldsymbol{X}}(1+o(1))$. It implies that,

$$\begin{aligned}\frac{1}{n_1n_2}\left\|\boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{X}\boldsymbol{\beta}_0\right\|_F^2 \leq & \frac{1}{n_1n_2}\left\|\boldsymbol{P}_{\boldsymbol{X}}\left(\boldsymbol{W}\circ\hat{\boldsymbol{\Theta}}^*\circ\boldsymbol{Y} - \boldsymbol{A}_0\right)\right\|_F^2(1+o(1)) + n_2^2\lambda_1^2\left\|\boldsymbol{X}\boldsymbol{\beta}_0\right\|_F^2(1+o(1)) \\ \leq & \frac{m}{n_1n_2}\left\|\boldsymbol{P}_{\boldsymbol{X}}\left(\boldsymbol{W}\circ\hat{\boldsymbol{\Theta}}^*\circ\boldsymbol{Y} - \boldsymbol{A}_0\right)\right\|^2(1+o(1)) + n_2^2\lambda_1^2\left\|\boldsymbol{X}\boldsymbol{\beta}_0\right\|_F^2(1+o(1)) \\ \leq & 2mn_1n_2\left(C_0^2\Delta^2(\delta_\sigma,t) + a_1n_2^2\left\{\log(n)\right\}\lambda_1^2\right)\end{aligned}$$

with the probability at least $1 - 2/n - 2g(t) - 2h_{n_1,n_2} - 12c_\sigma^2\eta^2\log^{-\delta_\sigma}(n)$.

It follows from the definition of $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{B}}$ that

$$\begin{aligned}&\frac{1}{n_1n_2}\left\|\hat{\boldsymbol{A}} - \boldsymbol{W}\circ\hat{\boldsymbol{\Theta}}^*\circ\boldsymbol{Y}\right\|_F^2 + \lambda_1\left\|\hat{\boldsymbol{\beta}}\right\|_F^2 + \lambda_2\left(\alpha\left\|\hat{\boldsymbol{B}}\right\|_* + (1-\alpha)\left\|\hat{\boldsymbol{B}}\right\|_F^2\right) \\ \leq & \frac{1}{n_1n_2}\left\|\boldsymbol{X}\hat{\boldsymbol{\beta}} + \boldsymbol{B}_0 - \boldsymbol{W}\circ\hat{\boldsymbol{\Theta}}^*\circ\boldsymbol{Y}\right\|_F^2 + \lambda_1\left\|\hat{\boldsymbol{\beta}}\right\|_F^2 + \lambda_2\left(\alpha\left\|\boldsymbol{B}_0\right\|_* + (1-\alpha)\left\|\boldsymbol{B}_0\right\|_F^2\right).\end{aligned} \tag{S5.1}$$

Since we can rewrite the first term in the left hand side of (S5.1) as

$$\frac{1}{n_1n_2}\left\|\hat{\boldsymbol{A}} - \boldsymbol{W}\circ\hat{\boldsymbol{\Theta}}^*\circ\boldsymbol{Y}\right\|_F^2 = \frac{1}{n_1n_2}\left\|\boldsymbol{X}\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{B}} - \boldsymbol{B}_0 + \boldsymbol{B}_0 - \boldsymbol{W}\circ\hat{\boldsymbol{\Theta}}^*\circ\boldsymbol{Y}\right\|_F^2,$$

the inequality (S5.1) is equivalent to

$$\begin{aligned}\frac{1}{n_1n_2}\left\|\hat{\boldsymbol{B}} - \boldsymbol{B}_0\right\|_F^2 \leq & \frac{2}{n_1n_2}\left(\left\langle\hat{\boldsymbol{B}} - \boldsymbol{B}_0, \boldsymbol{W}\circ\hat{\boldsymbol{\Theta}}^*\circ\boldsymbol{\epsilon}\right\rangle + \left\langle\hat{\boldsymbol{B}} - \boldsymbol{B}_0, \boldsymbol{W}\circ\boldsymbol{\Theta}^*\circ\boldsymbol{A}_0 - \boldsymbol{A}_0\right\rangle\right. \\ & \left. + \left\langle\hat{\boldsymbol{B}} - \boldsymbol{B}_0, \boldsymbol{X}\boldsymbol{\beta}_0 - \boldsymbol{X}\hat{\boldsymbol{\beta}}\right\rangle + \left\langle\hat{\boldsymbol{B}} - \boldsymbol{B}_0, \boldsymbol{W}\circ\hat{\boldsymbol{\Theta}}^*\circ\boldsymbol{A}_0 - \boldsymbol{W}\circ\boldsymbol{\Theta}^*\circ\boldsymbol{A}_0\right\rangle\right) \\ & + \lambda_2\alpha\left(\left\|\boldsymbol{B}_0\right\|_* - \left\|\hat{\boldsymbol{B}}\right\|_*\right) + \lambda_2(1-\alpha)\left(\left\|\boldsymbol{B}_0\right\|_F^2 - \left\|\hat{\boldsymbol{B}}\right\|_F^2\right).\end{aligned}$$

We focus on the bound related to $\|\boldsymbol{B}_0\|_*$ in (4.3), namely,

$$d^2\left(\hat{\boldsymbol{B}}, \boldsymbol{B}_0\right) \leq C'\max\left\{\lambda_2\alpha\left\|\boldsymbol{B}_0\right\|_*, \lambda_2(1-\alpha)\left\|\boldsymbol{B}_0\right\|_F^2, n_1n_2\Delta^2(\delta_\sigma,t)\right\}, \tag{S5.2}$$

first. By the trace duality property given in (S4.1), with probability at least $1 - 2/n - 2g(t) - 2h_{n_1,n_2} - 12c_\sigma^2\eta^2\log^{-\delta_\sigma}(n)$, we have

$$\begin{aligned}\frac{1}{n_1n_2}\left\|\hat{\boldsymbol{B}} - \boldsymbol{B}_0\right\|_F^2 \leq & 2\left\|\hat{\boldsymbol{B}} - \boldsymbol{B}_0\right\|_*\left(\left\|\Psi^{(1)}\right\| + \left\|\Psi^{(2)}\right\| + \left\|\Psi^{(3)}\right\|\right) \\ & + \lambda_2\alpha\left(\left\|\boldsymbol{B}_0\right\|_* - \left\|\hat{\boldsymbol{B}}\right\|_*\right) + \lambda_2(1-\alpha)\left(\left\|\boldsymbol{B}_0\right\|_F^2 - \left\|\hat{\boldsymbol{B}}\right\|_F^2\right) \\ \leq & 2C_0\left(\left\|\hat{\boldsymbol{B}}\right\|_* + \left\|\boldsymbol{B}_0\right\|_*\right)\Delta(\delta_\sigma,t) + \lambda_2\alpha\left(\left\|\boldsymbol{B}_0\right\|_* - \left\|\hat{\boldsymbol{B}}\right\|_*\right) + \lambda_2(1-\alpha)\left(\left\|\boldsymbol{B}_0\right\|_F^2 - \left\|\hat{\boldsymbol{B}}\right\|_F^2\right).\end{aligned}$$

For $0 < \alpha \leq 1$ and $\lambda_2 \alpha \geq 2C_0 \Delta(\delta_\sigma, t)$, we can simplify the inequality to

$$\frac{1}{n_1 n_2} \left\| \hat{\boldsymbol{B}} - \boldsymbol{B}_0 \right\|_F^2 \leq \left(2C_0 \Delta \left(\delta_\sigma, t \right) + \lambda_2 \alpha \right) \left\| \boldsymbol{B}_0 \right\|_* + \lambda_2 \left(1 - \alpha \right) \left\| \boldsymbol{B}_0 \right\|_F^2,$$

with probability at least $1 - 2/n - 2g(t) - 2h_{n_1,n_2} - 12c_\sigma^2 \eta^2 \log^{-\delta_\sigma}(n)$.

Now we focus on the bound related to $r_{\boldsymbol{B}_0}$ in (4.3), namely,

$$d^2 \left(\hat{\boldsymbol{B}}, \boldsymbol{B}_0 \right) \leq C' \max \left\{ n_1 n_2 r_{\boldsymbol{B}_0} \left(\lambda_2 \alpha \right)^2, \lambda_2 \left(1 - \alpha \right) \left\| \boldsymbol{B}_0 \right\|_F^2 \right\}. \tag{S5.3}$$

To prove the remaining bounds, note that for any $\boldsymbol{Z} \in \partial \| \boldsymbol{B}_0 \|_*$, we have $\| \boldsymbol{B}_0 \|_* + \langle \boldsymbol{Z}, \hat{\boldsymbol{B}} - \boldsymbol{B}_0 \rangle \leq \| \hat{\boldsymbol{B}} \|_*$. The inequality (S5.1) implies, for any $\boldsymbol{Z} \in \partial \| \boldsymbol{B}_0 \|_*$

$$\begin{aligned} &\frac{1}{n_1 n_2} \left\| \hat{\boldsymbol{B}} - \boldsymbol{B}_0 \right\|_F^2 \\ \leq &\frac{2}{n_1 n_2} \left\langle \hat{\boldsymbol{B}} - \boldsymbol{B}_0, \boldsymbol{W} \circ \hat{\boldsymbol{\Theta}}^* \circ \boldsymbol{Y} - \boldsymbol{B}_0 - \boldsymbol{X} \hat{\boldsymbol{\beta}} \right\rangle + \lambda_2 \alpha \left\langle \boldsymbol{Z}, \boldsymbol{B}_0 - \hat{\boldsymbol{B}} \right\rangle + \lambda_2 \left(1 - \alpha \right) \left\| \boldsymbol{B}_0 \right\|_F^2. \end{aligned} \tag{S5.4}$$

On the other hand, by definition of $\partial \| \boldsymbol{B}_0 \|_*$, $\boldsymbol{Z} = \sum_{i=1}^{r_{\boldsymbol{B}_0}} \boldsymbol{u}_{\boldsymbol{B}_0}^{(i)} \boldsymbol{v}_{\boldsymbol{B}_0}^{(i)T} + \boldsymbol{P}_{\mathcal{B}_{\boldsymbol{u}}^\perp} \boldsymbol{W} \boldsymbol{P}_{\mathcal{B}_{\boldsymbol{v}}^\perp}$, where $\boldsymbol{W}$ is an arbitrary matrix with $\| \boldsymbol{W} \| \leq 1$. It follows from the trace duality (S4.1) that there exists $\boldsymbol{W}$ with $\| \boldsymbol{W} \| \leq 1$ such that

$$\left\langle \boldsymbol{P}_{\mathcal{B}_{\boldsymbol{u}}^\perp} \boldsymbol{W} \boldsymbol{P}_{\mathcal{B}_{\boldsymbol{v}}^\perp}, \boldsymbol{B}_0 - \hat{\boldsymbol{B}} \right\rangle = - \left\langle \boldsymbol{P}_{\mathcal{B}_{\boldsymbol{u}}^\perp} \boldsymbol{W} \boldsymbol{P}_{\mathcal{B}_{\boldsymbol{v}}^\perp}, \hat{\boldsymbol{B}} \right\rangle = \left\langle \boldsymbol{W}, \boldsymbol{P}_{\mathcal{B}_{\boldsymbol{u}}^\perp} \hat{\boldsymbol{B}} \boldsymbol{P}_{\mathcal{B}_{\boldsymbol{v}}^\perp} \right\rangle = \left\| \boldsymbol{P}_{\mathcal{B}_{\boldsymbol{u}}^\perp} \hat{\boldsymbol{B}} \boldsymbol{P}_{\mathcal{B}_{\boldsymbol{v}}^\perp} \right\|_*.$$

For this particular choice of $\boldsymbol{W}$, (S5.4) implies that

$$\begin{aligned} &\frac{1}{n_1 n_2} \left\| \hat{\boldsymbol{B}} - \boldsymbol{B}_0 \right\|_F^2 + \lambda_2 \alpha \left\| \boldsymbol{P}_{\mathcal{B}_{\boldsymbol{u}}^\perp} \hat{\boldsymbol{B}} \boldsymbol{P}_{\mathcal{B}_{\boldsymbol{v}}^\perp} \right\|_* \\ \leq &\frac{2}{n_1 n_2} \left\langle \hat{\boldsymbol{B}} - \boldsymbol{B}_0, \boldsymbol{W} \circ \hat{\boldsymbol{\Theta}}^* \circ \boldsymbol{Y} - \boldsymbol{B}_0 - \boldsymbol{X} \hat{\boldsymbol{\beta}} \right\rangle + \lambda_2 \alpha \left\langle \sum_{i=1}^{r_{\boldsymbol{B}_0}} \boldsymbol{u}_{\boldsymbol{B}_0}^{(i)} \boldsymbol{v}_{\boldsymbol{B}_0}^{(i)T}, \boldsymbol{B}_0 - \hat{\boldsymbol{B}} \right\rangle + \lambda_2 \left(1 - \alpha \right) \left\| \boldsymbol{B}_0 \right\|_F^2. \end{aligned} \tag{S5.5}$$

Using the facts that $\| \sum_{i=1}^{r_{\boldsymbol{B}_0}} \boldsymbol{u}_{\boldsymbol{B}_0}^{(i)} \boldsymbol{v}_{\boldsymbol{B}_0}^{(i)T} \| = 1$ and $\langle \sum_{i=1}^{r_{\boldsymbol{B}_0}} \boldsymbol{u}_{\boldsymbol{B}_0}^{(i)} \boldsymbol{v}_{\boldsymbol{B}_0}^{(i)T}, \boldsymbol{B}_0 - \hat{\boldsymbol{B}} \rangle = \langle \sum_{i=1}^{r_{\boldsymbol{B}_0}} \boldsymbol{u}_{\boldsymbol{B}_0}^{(i)} \boldsymbol{v}_{\boldsymbol{B}_0}^{(i)T}, \boldsymbol{P}_{\mathcal{B}_{\boldsymbol{u}}} (\boldsymbol{B}_0 - \hat{\boldsymbol{B}}) \boldsymbol{P}_{\mathcal{B}_{\boldsymbol{v}}} \rangle$, we deduce from (S5.5) that

$$\begin{aligned} &\frac{1}{n_1 n_2} \left\| \hat{\boldsymbol{B}} - \boldsymbol{B}_0 \right\|_F^2 + \lambda_2 \alpha \left\| \boldsymbol{P}_{\mathcal{B}_{\boldsymbol{u}}^\perp} \hat{\boldsymbol{B}} \boldsymbol{P}_{\mathcal{B}_{\boldsymbol{v}}^\perp} \right\|_* \\ \leq &2 \left\langle \hat{\boldsymbol{B}} - \boldsymbol{B}_0, \boldsymbol{M} \right\rangle + \lambda_2 \alpha \left\| \boldsymbol{P}_{\mathcal{B}_{\boldsymbol{u}}} \left(\boldsymbol{B}_0 - \hat{\boldsymbol{B}} \right) \boldsymbol{P}_{\mathcal{B}_{\boldsymbol{v}}} \right\|_* + \lambda_2 \left(1 - \alpha \right) \left\| \boldsymbol{B}_0 \right\|_F^2, \end{aligned} \tag{S5.6}$$

where $\boldsymbol{M} = (\boldsymbol{W} \circ \hat{\boldsymbol{\Theta}}^* \circ \boldsymbol{Y} - \boldsymbol{B}_0 - \boldsymbol{X} \hat{\boldsymbol{\beta}}) / (n_1 n_2)$.

To provide an upper bound on $2\langle\hat{\boldsymbol{B}}-\boldsymbol{B}_0,\boldsymbol{M}\rangle$ we use the following decomposition:

$$\begin{aligned}\left\langle\hat{\boldsymbol{B}}-\boldsymbol{B}_0,\boldsymbol{M}\right\rangle=&\left\langle\hat{\boldsymbol{B}}-\boldsymbol{B}_0,\mathcal{P}_{\boldsymbol{B}_0}\left(\boldsymbol{M}\right)\right\rangle+\left\langle\hat{\boldsymbol{B}}-\boldsymbol{B}_0,\boldsymbol{P}_{\mathcal{B}_{\boldsymbol{u}}^{\perp}}\boldsymbol{M}\boldsymbol{P}_{\mathcal{B}_{\boldsymbol{v}}^{\perp}}\right\rangle\\=&\left\langle\mathcal{P}_{\boldsymbol{B}_0}\left(\hat{\boldsymbol{B}}-\boldsymbol{B}_0\right),\mathcal{P}_{\boldsymbol{B}_0}\left(\boldsymbol{M}\right)\right\rangle+\left\langle\hat{\boldsymbol{B}},\boldsymbol{P}_{\mathcal{B}_{\boldsymbol{u}}^{\perp}}\boldsymbol{M}\boldsymbol{P}_{\mathcal{B}_{\boldsymbol{v}}^{\perp}}\right\rangle,\end{aligned}$$

where $\mathcal{P}_{\boldsymbol{B}_0}(\boldsymbol{M})=\boldsymbol{M}-\boldsymbol{P}_{\mathcal{B}_{\boldsymbol{u}}^{\perp}}\boldsymbol{M}\boldsymbol{P}_{\mathcal{B}_{\boldsymbol{v}}^{\perp}}$. Due to the trace duality (S4.1),

$$\begin{aligned}2\left|\left\langle\hat{\boldsymbol{B}}-\boldsymbol{B}_0,\boldsymbol{M}\right\rangle\right|\leq&\Lambda\left\|\mathcal{P}_{\boldsymbol{B}_0}\left(\hat{\boldsymbol{B}}-\boldsymbol{B}_0\right)\right\|_F+\Gamma\left\|\boldsymbol{P}_{\mathcal{B}_{\boldsymbol{u}}^{\perp}}\hat{\boldsymbol{B}}\boldsymbol{P}_{\mathcal{B}_{\boldsymbol{v}}^{\perp}}\right\|_*\\\leq&\Lambda\left\|\hat{\boldsymbol{B}}-\boldsymbol{B}_0\right\|_F+\Gamma\left\|\boldsymbol{P}_{\mathcal{B}_{\boldsymbol{u}}^{\perp}}\hat{\boldsymbol{B}}\boldsymbol{P}_{\mathcal{B}_{\boldsymbol{v}}^{\perp}}\right\|_*,\end{aligned}$$

where $\Lambda=2\|\mathcal{P}_{\boldsymbol{B}_0}(\boldsymbol{M})\|_F$ and $\Gamma=2\|\boldsymbol{P}_{\mathcal{B}_{\boldsymbol{u}}^{\perp}}(\boldsymbol{M})\boldsymbol{P}_{\mathcal{B}_{\boldsymbol{v}}^{\perp}}\|$. Note that $\Gamma\leq 2\|\boldsymbol{M}\|\leq 2C_0\Delta(\delta_\sigma,t):=\Gamma^\star$.

Since $\mathcal{P}_{\boldsymbol{B}_0}(\boldsymbol{M})=\boldsymbol{P}_{\mathcal{B}_{\boldsymbol{u}}^{\perp}}\boldsymbol{M}\boldsymbol{P}_{\mathcal{B}_{\boldsymbol{v}}}+\boldsymbol{P}_{\mathcal{B}_{\boldsymbol{u}}}\boldsymbol{M}$, $\operatorname{rank}(\boldsymbol{P}_{\mathcal{B}_{\boldsymbol{u}}})\leq r_{\boldsymbol{B}_0}$ and $\operatorname{rank}(\boldsymbol{P}_{\mathcal{B}_{\boldsymbol{v}}})\leq r_{\boldsymbol{B}_0}$, we have

$$\Lambda\leq 2\sqrt{\operatorname{rank}\left(\mathcal{P}_{\boldsymbol{B}_0}(\boldsymbol{M})\right)}\|\mathcal{P}_{\boldsymbol{B}_0}(\boldsymbol{M})\|\leq 2\sqrt{2r_{\boldsymbol{B}_0}}C_0\Delta\left(\delta_\sigma,t\right):=\Lambda^\star.$$

Due to the facts that

$$\left\|\boldsymbol{P}_{\mathcal{B}_{\boldsymbol{u}}}\left(\boldsymbol{B}_0-\hat{\boldsymbol{B}}\right)\boldsymbol{P}_{\mathcal{B}_{\boldsymbol{v}}}\right\|_*\leq\sqrt{r_{\boldsymbol{B}_0}}\left\|\boldsymbol{P}_{\mathcal{B}_{\boldsymbol{u}}}\left(\boldsymbol{B}_0-\hat{\boldsymbol{B}}\right)\boldsymbol{P}_{\mathcal{B}_{\boldsymbol{v}}}\right\|_F\leq\sqrt{r_{\boldsymbol{B}_0}}\left\|\boldsymbol{B}_0-\hat{\boldsymbol{B}}\right\|_F,$$

we have

$$\begin{aligned}&\frac{1}{n_1n_2}\left\|\hat{\boldsymbol{B}}-\boldsymbol{B}_0\right\|_F^2+\lambda_2\alpha\left\|\boldsymbol{P}_{\mathcal{B}_{\boldsymbol{u}}^{\perp}}\hat{\boldsymbol{B}}\boldsymbol{P}_{\mathcal{B}_{\boldsymbol{v}}^{\perp}}\right\|_*\\\leq&\left(\Lambda+\lambda_2\alpha\sqrt{r_{\boldsymbol{B}_0}}\right)\left\|\hat{\boldsymbol{B}}-\boldsymbol{B}_0\right\|_F+\Gamma\left\|\boldsymbol{P}_{\mathcal{B}_{\boldsymbol{u}}^{\perp}}\hat{\boldsymbol{B}}\boldsymbol{P}_{\mathcal{B}_{\boldsymbol{v}}^{\perp}}\right\|_*+\lambda_2\left(1-\alpha\right)\|\boldsymbol{B}_0\|_F^2,\end{aligned}$$

which implies

$$\begin{aligned}&\frac{1}{n_1n_2}\left\|\hat{\boldsymbol{B}}-\boldsymbol{B}_0\right\|_F^2+\left(\lambda_2\alpha-2C_0\Delta(\delta_\sigma,t)\right)\left\|\boldsymbol{P}_{\mathcal{B}_{\boldsymbol{u}}^{\perp}}\hat{\boldsymbol{B}}\boldsymbol{P}_{\mathcal{B}_{\boldsymbol{v}}^{\perp}}\right\|_*\\\leq&\left(\Lambda+\lambda_2\alpha\sqrt{r_{\boldsymbol{B}_0}}\right)\left\|\hat{\boldsymbol{B}}-\boldsymbol{B}_0\right\|_F+\lambda_2\left(1-\alpha\right)\|\boldsymbol{B}_0\|_F^2.\end{aligned}$$

Take $\lambda_2\alpha\geq 2C_0\Delta(\delta_\sigma,t)$, we have

$$\frac{1}{n_1n_2}\left\|\hat{\boldsymbol{B}}-\boldsymbol{B}_0\right\|_F^2\leq n_1n_2r_{\boldsymbol{B}_0}\left(2\sqrt{2}C_0\Delta\left(\delta_\sigma,t\right)+\lambda_2\alpha\right)^2+2\lambda_2\left(1-\alpha\right)\|\boldsymbol{B}_0\|_F^2.$$

Note that $2C_0\Delta(\delta_\sigma,t)\leq\lambda_2\alpha$, this means (S5.3) holds.

Finally, in Theorem 1, under the choice of parameters $0<\alpha\leq 1$ and $\lambda_2\alpha\geq(2+4m)C_0\Delta(\delta_\sigma,t)$, we have $n_1n_2C_0^2\Delta^2(\delta_\sigma,t)\leq n_1n_2r_{\boldsymbol{B}_0}(\lambda_2\alpha)^2$. Thus (4.3) follows from (S5.2) and (S5.3). □

Proof of Corollary 1 and Corollary 2. For Corollary 1, it is readily shown that $\sqrt{n_1 n_2 \theta_0/(1-\theta_0)}(1/\hat{\theta} - 1/\theta_0) \xrightarrow{d} \mathcal{N}(0,1)$. Since $\mathsf{P}\{(1/\hat{\theta} - 1/\theta_0)^2 \geq (1-\theta_0)t/\theta_0 \leq \mathsf{P}\{\chi_1^2 > t\} + \sup_t |\mathsf{P}\{\chi_1^2 > t\} - \mathsf{P}\{\theta_0(1/\hat{\theta} - 1/\theta_0)^2/(1-\theta_0) \geq t\}|$ where χ_1^2 is the chi-square random variable with one degree of freedom. Choose $c_{n_1,n_2} = (1-\theta_0)/\theta_0$, $t_0 > 0$, $g(t) = \mathsf{P}\{\chi_1^2 > t\}$ and $h_{n_1,n_2} = \sup_t |\mathsf{P}\{\theta_0(1/\hat{\theta} - 1/\theta_0)^2/(1-\theta_0) \geq t\} - g(t)|$ in Condition C5(b). While that $\lim_{t\to\infty} g(t) = 0$ is obvious, by Polya's theorem, $\lim_{n_1,n_2\to\infty} h_{n_1,n_2} = 0$. Thus Condition C5(b) holds for any positive t under the uniform probability of observation model. Under Condition C2 and C3, we have $\|\boldsymbol{B}_0\|_F = O\{\sqrt{n_1 n_2 \log(n)}\}$ and $\|\boldsymbol{X}\boldsymbol{\beta}_0\|_F = O\{\sqrt{n_1 n_2 \log(n)}\}$. Thus the dominate term in the right hand side is $n_1 n_2 r_{\boldsymbol{B}_0} \Delta_1^2$.

For Corollary 2, it is shown in Section S1.4 that by taking $c_{n_1,n_2} = \eta_g^{-1} n_2 \log(n_2)$ and $t_0 = (m+3)$, we have

$$\mathsf{P}\left\{\sum_{j=1}^{n_2}\sum_{i=1}^{n_1}\left(\frac{1}{\hat{\theta}_{ij}} - \frac{1}{\theta_{ij}}\right)^2 \geq c_{n_1,n_2} t\right\} \leq (m+1)^{-(m+1)/2} \exp\left\{m + 2 - \frac{t}{2} + \log(t)\right\} + n_2 k_{n_1}$$

where η_g is a constant depend on θ_L, χ_{m+1}^2 is the chi-square random variable with $m+1$ degrees of freedom, and $k_{n_1} = \max_j \sup_t |\mathsf{P}\{\sum_i (1/\hat{\theta}_{ij} - 1/\theta_{ij})^2 \geq t\} - \mathsf{P}(\chi_{m+1}^2 \geq t)|$.

Take $g(t) = (m+1)^{-(m+1)/2} \exp\{m+2-t/2+\log(t)\}$, and $h_{n_1,n_2} = n_2 k_{n_1}$. Then, $\lim_{t\to\infty} g(t) = 0$. By Polya's theorem, it is shown in Section S1.4 that there exists a positive integer N such that for $n_1 > N$ and $k_{n_1} < 1/n_2^2$, which implies that $\lim_{n_1,n_2\to\infty} h_{n_1,n_2} = 0$. Thus Condition C5(b) holds for any positive $t > t_0$ for the logistic model. Choose t as (4.5), we have $\sup_t \Delta(\delta_\sigma, t) = \Delta_2(\delta_\sigma) \asymp \eta_g^{-1/2} n_1^{-3/4} n_2^{-1/4} \log^{1/2}(n_2) \log^{\delta_\sigma/3}(n)$. This implies that the convergence rate for $d^2(\hat{\boldsymbol{A}}, \boldsymbol{A}_0)$ given in (4.3) is $\eta_g^{-1} n_1^{-1/2} n_2^{1/2} \log(n_2) \log^{2\delta_\sigma/3}(n)$. Under Condition C2 and C3, we have $\|\boldsymbol{B}_0\|_F = O\{\sqrt{n_1 n_2 \log(n)}\}$ and $\|\boldsymbol{X}\boldsymbol{\beta}_0\|_F = O\{\sqrt{n_1 n_2 \log(n)}\}$. Thus the dominate term in the right hand side is $n_1 n_2 r_{\boldsymbol{B}_0} \Delta_2^2(\delta_\sigma)$.

Assume that $n_1 \asymp \eta_g^2 n_2 \log^{2+2\delta_\sigma}(n_2)$, then right hand side becomes $r_{\boldsymbol{B}_0} \log^{-2\delta_\sigma/3}(n_2)$. □

S6 Proof of Theorem 2

Proof of Theorem 2. Since that $\lambda_1 = o(n_2^{-1})$, $n_2\lambda_1 = o(1)$, we have

$$\left(n_1^{-1}\boldsymbol{X}^\intercal\boldsymbol{X} + n_2\lambda_1\boldsymbol{I}_{m\times m}\right)^{-1} \to \boldsymbol{S}_x^{-1}.$$

We have the estimators $\hat{\theta}_{ij}$ of θ_{ij} satisfy that for $|\hat{\theta}_{ij} - \theta_{ij}| = O_p(n_1^{-1/2})$. Thus for the jth column of matrix $\boldsymbol{W}\circ\hat{\boldsymbol{\Theta}}^*\circ\boldsymbol{Y}$, we have

$$\left(\boldsymbol{W}\circ\hat{\boldsymbol{\Theta}}^*\circ\boldsymbol{Y}\right)_j = \left(\boldsymbol{W}\circ\left(1+O_p\left(n_1^{-1/2}\right)\right)\boldsymbol{\Theta}^*\circ\boldsymbol{Y}\right)_j.$$

Let $\boldsymbol{Z}_j = n_1^{-1}\boldsymbol{X}^\intercal(\boldsymbol{W}\circ\boldsymbol{\Theta}^*\circ\boldsymbol{Y})_j$. Then $Z_{kj} = n_1^{-1}\sum_{i=1}^{n_1} x_{ik}\omega_{ij}Y_{ij}/\theta_{ij}$ for each $k = 1,\ldots,m$. Since $\mathsf{E}(x_{ik}\omega_{ij}Y_{ij}/(n_1\theta_{ij})) = x_{ik}(X\beta_0 + B_0)_{ij}/n_1$, $\mathsf{Var}(x_{ik}\omega_{ij}Y_{ij}/(n_1\theta_{ij})) = x_{ik}^2(1-\theta_{ij})\{(X\beta_0 + B_0)_{ij}^2 + \sigma_{ij}^2\}/(n_1^2\theta_{ij})$, define $s_{n_1}^2 = \sum_{i=1}^{n_1} x_{ik}^2(1-\theta_{ij})\{(X\beta_0+B_0)_{ij}^2+\sigma_{ij}^2\}/(n_1^2\theta_{ij})$. Also we have

$$\begin{aligned}\mathsf{E}\left|x_{ik}\omega_{ij}Y_{ij}/\left(n_1\theta_{ij}\right) - x_{ik}\left(X\beta_0+B_0\right)_{ij}/n_1\right|^3 = \Big(&x_{ik}^3(1-\theta_{ij})\{(X\beta_0+B_0)_{ij}^3 + (X\beta_0+B_0)_{ij}\sigma_{ij}^2\}/\theta_{ij}^2\\ &-3x_{ik}^3\left(1-\theta_{ij}\right)\left\{\left(X\beta_0+B_0\right)_{ij}^3 + \left(X\beta_0+B_0\right)_{ij}\sigma_{ij}^2\right\}/\theta_{ij} + 2x_{ik}^3\left(X\beta_0+B_0\right)_{ij}^3\Big)/n_1^3,\end{aligned}$$

implies the Lyapunovs condition satisfied, namely,

$$\lim_{n_1\to\infty}\frac{1}{s_{n_1}^3}\sum_{i=1}^{n_1}\mathsf{E}\left|x_{ik}\omega_{ij}Y_{ij}/\theta_{ij} - x_{ik}\left(X\beta_0+B_0\right)_{ij}\right|^3 = 0.$$

By Lyapunov Central Limit Theorem, we have

$$\frac{1}{s_{n_1}}\sum_{i=1}^{n_1}\left(x_{ik}\omega_{ij}Y_{ij}/\left(n_1\theta_{ij}\right) - x_{ik}\left(X\beta_0+B_0\right)_{ij}/n_1\right) \xrightarrow{d} \mathcal{N}\left(0,1\right).$$

Combining with $n_1^{-1}\boldsymbol{X}^\intercal\boldsymbol{X}\to\boldsymbol{S}_x$, we have $\boldsymbol{Z}_j = n_1^{-1}\boldsymbol{X}^\intercal(\boldsymbol{W}\circ\boldsymbol{\Theta}^*\circ\boldsymbol{Y})_j = \boldsymbol{S}_x\boldsymbol{\beta}_{0j} + O_p(\mathbf{1}/\sqrt{\boldsymbol{n_1}})$.

For the estimator $\hat{\boldsymbol{\beta}}_j = (n_1^{-1}\boldsymbol{X}^\intercal\boldsymbol{X}+n_2\lambda_1\boldsymbol{I}_{m\times m})^{-1}n_1^{-1}\boldsymbol{X}^\intercal(\boldsymbol{W}\circ\hat{\boldsymbol{\Theta}}^*\circ\boldsymbol{Y})_j = (1+o(1))\boldsymbol{S}_x^{-1}n_1^{-1}\boldsymbol{X}^\intercal(\boldsymbol{W}\circ(1+O_p(n_1^{-1/2}))\boldsymbol{\Theta}^*\circ\boldsymbol{Y})_j$, we have $\hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_{0j}\xrightarrow{p}0$ and $\|\hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_{0j}\|_F^2 = O_p(m/n_1) = O_p(1/n_1)$. This completes the proof of Theorem 2. □

Table S1: Empirical root mean square errors (RMSEs), test errors, estimated ranks and their standard errors (in parentheses) under model $\boldsymbol{A}_0 = \boldsymbol{B}_0$ and uniform observation mechanism (UNI), with (n_1, n_2)=(400,400), (600,600), (800,800), (1000,1000) $m = 20$, and $r = 10$, for two versions of the proposed methods, and the four existing methods (SZ, NW, KLT and MHT).

$n_1 = n_2 = 400$	RMSE($\boldsymbol{\beta}_0$)	RMSE($\boldsymbol{B}_0$)	RMSE($\boldsymbol{A}_0$)	Test error	Rank
SVT-$\hat{\alpha}$-UNI	0.0121 (1e-04)	2.2346 (0.015)	2.2354 (0.015)	0.5723 (0.0071)	62.41 (1.59)
$\widehat{\text{SVT}}$-$\hat{\alpha}$-UNI	0.0121 (1e-04)	2.2342 (0.015)	2.2350 (0.015)	0.5721 (0.0071)	62.23 (1.58)
SZ			2.1082 (0.0167)	0.5059 (0.0076)	46.76 (2.74)
NW			2.0417 (0.0172)	0.4722 (0.0076)	94.48 (5.73)
KLT			2.2565 (0.0148)	0.5827 (0.007)	42.07 (1.58)
MHT			2.0550 (0.0171)	0.4796 (0.0076)	51.42 (2.57)
$n_1 = n_2 = 600$	RMSE($\boldsymbol{\beta}_0$)	RMSE($\boldsymbol{B}_0$)	RMSE($\boldsymbol{A}_0$)	Test error	Rank
SVT-$\hat{\alpha}$-UNI	0.0147 (1e-04)	2.0246 (0.0104)	2.0257 (0.0104)	0.4540 (0.0044)	75.82 (1.49)
$\widehat{\text{SVT}}$-$\hat{\alpha}$-UNI	0.0147 (1e-04)	2.0206 (0.0105)	2.0217 (0.0105)	0.4521 (0.0044)	74.51 (1.4)
SZ			1.8500 (0.0132)	0.3725 (0.0048)	58.17 (5.15)
NW			1.7794 (0.013)	0.3425 (0.0047)	120.92 (10.29)
KLT			2.0389 (0.0106)	0.4594 (0.0045)	55.49 (1.49)
MHT			1.7902 (0.011)	0.3476 (0.0042)	66.43 (2.46)
$n_1 = n_2 = 800$	RMSE($\boldsymbol{\beta}_0$)	RMSE($\boldsymbol{B}_0$)	RMSE($\boldsymbol{A}_0$)	Test error	Rank
SVT-$\hat{\alpha}$-UNI	0.0170 (1e-04)	1.8712 (0.0093)	1.8728 (0.0092)	0.3794 (0.0036)	85.54 (1.38)
$\widehat{\text{SVT}}$-$\hat{\alpha}$-UNI	0.0170 (1e-04)	1.8617 (0.0093)	1.8633 (0.0093)	0.3753 (0.0036)	82.46 (1.19)
SZ			1.6731 (0.0105)	0.2956 (0.0034)	60.91 (5.64)
NW			1.6055 (0.0085)	0.2707 (0.0029)	130.13 (6.05)
KLT			1.8817 (0.0092)	0.3824 (0.0036)	64.86 (1.36)
MHT			1.6107 (0.0099)	0.2734 (0.0032)	80.98 (6.26)
$n_1 = n_2 = 1000$	RMSE($\boldsymbol{\beta}_0$)	RMSE($\boldsymbol{B}_0$)	RMSE($\boldsymbol{A}_0$)	Test error	Rank
SVT-$\hat{\alpha}$-UNI	0.0185 (1e-04)	1.7238 (0.0073)	1.7258 (0.0073)	0.3275 (0.0027)	93.03 (1.36)
$\widehat{\text{SVT}}$-$\hat{\alpha}$-UNI	0.0185 (1e-04)	1.7090 (0.0073)	1.7111 (0.0073)	0.3216 (0.0026)	88.14 (1.12)
SZ			1.5076 (0.0069)	0.2435 (0.0023)	72.89 (2.72)
NW			1.4485 (0.0103)	0.2234 (0.0029)	157.62 (18.01)
KLT			1.7317 (0.0073)	0.3291 (0.0027)	72.37 (1.27)
MHT			1.4556 (0.0068)	0.2260 (0.0021)	85.43 (2.48)

S7 (Cont') Simulation study

S8 (Cont') Empirical Study

As suggested at `http://files.grouplens.org/datasets/movielens/ml-1m-README.txt`, we divide age into 7 categories: under 18, $18-24$, $25-34$, $35-44$, $45-49$, $50-55$ and 56+ in the modeling of probability estimator $\hat{\boldsymbol{\Theta}}^*$. However, it will cost much more ranks than keep it as numerical in the covariate $\boldsymbol{X}$ for prediction. To achieve a balance, we merge some age categories to form three to seven categories of the age variable. Specifically, the three categories layout is: under 24, $25-49$ and 50+; the four categories: under 24, $25-34$, $35-49$ and 50+; the five categories: under 24, $25-34$, $35-44$, $45-49$ and 50+; the six categories: under 18, $18-24$, $25-34$, $35-44$, $45-49$ and 50+; and the seven categories:under 18, $18-24$, $25-34$, $35-44$, $45-49$, $50-55$ and 56+. The predictions errors of using the four and five age categories are the best among the choices of three to seven categorization of the age.

Table S2: Root mean square prediction errors (RMSPEs) and ranks of the completed matrix based on Split1 and Split2 for the two versions of the proposed method (SVT-$\hat{\alpha}$-LOG) and ($\widehat{\text{SVT}}$-$\hat{\alpha}$-LOG) and the four existing methods proposed respectively in Sun and Zhang (2012)(SZ), Negahban and Wainwright (2012)(NW), Koltchinskii et al. (2011)(KLT) and Mazumder et al. (2010)(MHT).

		Split1		Split2		Overall
rank(X)		RMSPE	Rank	RMSPE	Rank	RMSPE
2	SVT-$\hat{\alpha}$-LOG	0.9415	47	0.9541	45	0.9478
	$\widehat{\text{SVT}}$-$\hat{\alpha}$-LOG	0.9416	45	0.9543	42	0.9480
4	SVT-$\hat{\alpha}$-LOG	0.9420	48	0.9540	42	0.9480
	$\widehat{\text{SVT}}$-$\hat{\alpha}$-LOG	0.9423	46	0.9540	42	0.9482
5	SVT-$\hat{\alpha}$-LOG	0.9420	49	0.9544	43	0.9483
	$\widehat{\text{SVT}}$-$\hat{\alpha}$-LOG	0.9422	47	0.9544	43	0.9483
	SZ	0.9412	39	0.9563	31	0.9488
	NW	0.9421	269	0.9589	289	0.9506
	KLT	0.9584	1	0.9688	1	0.9636
	MHT	0.9414	56	0.9568	46	0.9491

Table S2 reports the root mean square prediction errors (RMSPEs), estimated ranks and overall

RMSPEs of different estimators for both Split1 and Split2. The result with two categorical covariate $\boldsymbol{X}$ are included. Similarly as the simulation results reported in the Section 6, SVT-$\hat{\alpha}$-LOG and $\widehat{\text{SVT}}$-$\hat{\alpha}$-LOG produced highly comparable results, which indicated the applicability of $\widehat{\text{SVT}}$-$\hat{\alpha}$-LOG to larger data sets whenever computational resources are scarce. In Split2, the proposed methods outperformed SZ NW, KLT and MHT in terms of smaller RMSPEs and either smaller or more reasonable rank estimation. Although the proposed methods were slightly inferior to SZ and MHT in Split1, they outperformed SZ and MHT significantly in Split2 by having smaller RMSPEs. Among the ten matrix completion methods considered, the six proposed methods and the KLT method offered the most consistent results between Split1 and Split2, while the other three methods exhibited much larger variations, especially in the estimated ranks. Overall speaking, the two proposed methods were among the top two performers of the analysis reported in Table S2.

References

Ahlswede, R. and Winter, A. (2002), "Strong Converse for Identification via Quantum Channels," *IEEE Transactions on Information Theory*, 48, 569–579.

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

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.

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.

Sun, T. and Zhang, C.-H. (2012), "Calibrated Elastic Regularization in Matrix Completion," in *Advances in Neural Information Processing Systems*, pp. 863–871.

Sweeting, T. (1980), "Uniform Asymptotic Normality of the Maximum Likelihood Estimator," *The Annals of Statistics*, 8, 1375–1381.

Tropp, J. A. (2012), "User-Friendly Tail Bounds for Sums of Random Matrices," *Foundations of Computational Mathematics*, 12, 389–434.

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