Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: https://www.tandfonline.com/loi/uasa20

Matrix Completion With Covariate Information

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

To cite this article: Xiaojun Mao, Song Xi Chen & Raymond K. W. Wong (2019) Matrix Completion With Covariate Information, Journal of the American Statistical Association, 114:525, 198-210, DOI: 10.1080/01621459.2017.1389740

To link to this article: https://doi.org/10.1080/01621459.2017.1389740

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2019, VOL. 114, NO. 525, 198–210, Theory and Methods
https://doi.org/10.1080/01621459.2017.1389740

Matrix Completion With Covariate Information

Xiaojun Mao[a], Song Xi Chen[b], and Raymond K. W. Wong[c]

[a]Department of Statistics, Iowa State University, Ames, IA; [b]Department of Business Statistics and Econometrics, Guanghua School of Management and Center for Statistical Science, Peking University, Beijing, China; [c]Department of Statistics, Texas A&M University, College Station, TX

ABSTRACT

This article investigates the problem of matrix completion from the corrupted data, when the 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.

ARTICLE HISTORY
Received October 2016
Revised September 2017

KEYWORDS
High-dimensional statistics; Low-rank estimation; Missing data; Nuclear-norm regularization

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, He, and Khatri 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, Lounici, and Tsybakov 2011; Rohde and Tsybakov 2011) can be found under various forms of noise assumptions.

Some applications come with the covariate information in the form of additional row and/or column information. For instance, the MovieLens 100 K dataset (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 the perfect matrix recovery with covariates are available (Xu, Jin, and Zhou 2013; Chiang, Hsieh, and Dhillon 2015). Yet, there have been limited attempts with theoretical results at the more realistic setting where the observed entries are corrupted by noise. One notable study is the work by Zhu, Shen, and Ye (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 the 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 (Abernethy et al. 2009; Candès and Recht 2009; Keshavan, Montanari, and Oh 2009; Recht 2011; Rohde and Tsybakov 2011; Koltchinskii, Lounici, and Tsybakov 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, Cai, and Zhang 2016; Bi et al. 2016) devoted to relaxing such restrictive assumption to the nonuniform case, where the probabilities of observation are allowed to be different across the rows and the columns to some extent. However, the covariates are not taken into account in the modeling of the probabilities of observation.

CONTACT Song Chen ✉ songchen@iastate.edu Department of Business Statistics and Econometrics, Guanghua School of Management and Center for Statistical Science, Peking University, Beijing 100651, China.

Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA.

Driven by the aforementioned empirical observation, we model the 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 article, we use 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 the row and the 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 the 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 the 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 noniterative. Specifically, unlike many matrix completion algorithms that involve multiple singular value decompositions (SVD), our computation requires only one single SVD. This SVD can be reused in computations of the proposed estimators with respect to different tuning parameters, which leads to the 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 the 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 the 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 nonasymptotic upper bound for the MSE of the completed matrix $\hat{\boldsymbol{B}}$ and show it is no larger than the one by Koltchinskii, Lounici, and Tsybakov (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 nonuniform missingness are also provided.

The rest of the article 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

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

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 addition to the incomplete matrix $\boldsymbol{Y}$, we have an accompanying covariate matrix $\boldsymbol{X} = (\boldsymbol{x}_1, \dots, \boldsymbol{x}_{n_1})^\intercal \in \mathbb{R}^{n_1 \times m}$, where $\boldsymbol{x}_i \in \mathbb{R}^{m \times 1}$ for $i = 1, \dots, n_1$. 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 the 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, Kent, and Bibby 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 Equation (1)

$$\boldsymbol{A}_0 = \boldsymbol{X}\boldsymbol{\beta}_0 + \boldsymbol{B}_0. \tag{2}$$

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) 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 Equation (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 β_0 and 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

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

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,

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

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

$$\hat{R}(\boldsymbol{\beta}, \boldsymbol{B}) = \frac{1}{n_1 n_2}\left\{\left\|\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{P}_{\boldsymbol{X}}(\boldsymbol{W}\circ\boldsymbol{\Theta}^*\circ\boldsymbol{Y})\right\|_F^2 + \left\|\boldsymbol{B} - \boldsymbol{P}_{\boldsymbol{X}}^{\perp}(\boldsymbol{W}\circ\boldsymbol{\Theta}^*\circ\boldsymbol{Y})\right\|_F^2\right\}. \tag{3}$$

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

$$\hat{R}^*(\boldsymbol{\beta}, \boldsymbol{B}) = \frac{1}{n_1 n_2}\left\{\left\|\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{P}_{\boldsymbol{X}} big(\boldsymbol{W}\circ\hat{\boldsymbol{\Theta}}^*\circ\boldsymbol{Y}\right)\right\|_F^2 + \left\|\boldsymbol{B} - \boldsymbol{P}_{\boldsymbol{X}}^{\perp}\left(\boldsymbol{W}\circ\hat{\boldsymbol{\Theta}}^*\circ\boldsymbol{Y}\right)\right\|_F^2\right\}, \tag{4}$$

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

$$f(\boldsymbol{\beta}, \boldsymbol{B}; \lambda_1, \lambda_2, \alpha) = \hat{R}^*\left(\boldsymbol{\beta}, \boldsymbol{B}\right) + \lambda_1\left\|\boldsymbol{\beta}\right\|_F^2 + \lambda_2\left(\alpha\left\|\boldsymbol{B}\right\|_* + (1-\alpha)\left\|\boldsymbol{B}\right\|_F^2\right) \tag{5}$$

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 $0 \le \alpha \le 1$ 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, Hastie, and Tibshirani 2013).

Due to the orthogonal separation of $\boldsymbol{\beta}$ and $\boldsymbol{B}$ in Equation (4), the minimization of Equation (5) is equivalent to the following two separate minimizations

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^{m\times n_2}}{\arg\min}\left\{\frac{1}{n_1 n_2}\left\|\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{P}_{\boldsymbol{X}}\left(\boldsymbol{W}\circ\hat{\boldsymbol{\Theta}}^*\circ\boldsymbol{Y}\right)\right\|_F^2 + \lambda_1\left\|\boldsymbol{\beta}\right\|_F^2\right\} \tag{6}$$

and

$$\hat{\boldsymbol{B}} = \underset{\boldsymbol{B}\in\mathcal{N}(\boldsymbol{X})}{\arg\min}\left\{\frac{1}{n_1 n_2}\left\|\boldsymbol{B} - \boldsymbol{P}_{\boldsymbol{X}}^{\perp}\left(\boldsymbol{W}\circ\hat{\boldsymbol{\Theta}}^*\circ\boldsymbol{Y}\right)\right\|_F^2 + \lambda_2\left(\alpha\left\|\boldsymbol{B}\right\|_* + (1-\alpha)\left\|\boldsymbol{B}\right\|_F^2\right)\right\}. \tag{7}$$

3.2. Closed-Form Expressions and Fast Computation

We discuss how to compute $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{B}}$ given in Equations (6) and (7). As Equation (6) is essentially a ridge regression problem, straightforward algebra gives

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\intercal \mathbf{X} + \lambda_1' \mathbf{I}_{m\times m})^{-1}\mathbf{X}^\intercal\big(\mathbf{W}\circ\hat{\boldsymbol{\Theta}}^{*}\circ\mathbf{Y}\big), \tag{8}$$

where $\lambda_1' = n_1 n_2 \lambda_1$ and $\mathbf{I}_{m\times m}$ denotes the m-by-m identity matrix. We observe that the matrix inversion in Equation (8) 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 $\mathbf{A}$. As for the solution $\hat{\mathbf{B}}$ in Equation (7), the minimization over $\mathbf{B}\in\mathcal{N}(\mathbf{X})$ is not straightforward. The following proposition, whose proof is given in Section S1 of the supplementary material, shows that the minimization problem (7) can be carried out by extending the domain from $\mathcal{N}(\mathbf{X})$ to $\mathbb{R}^{n_1\times n_2}$. This domain enlargement reduces the complexity of the minimization.

Proposition 2. Suppose that $\mathbf{X}^\intercal\mathbf{X}$ is invertible, the minimization problem (7) is equivalent to

$$\operatorname*{arg\,min}_{\mathbf{B}\in\mathbb{R}^{n_1\times n_2}} \left\{ \frac{1}{n_1 n_2}\left\| \mathbf{B} - \mathbf{P}_{\mathbf{X}}^{\perp}\big(\mathbf{W}\circ\hat{\boldsymbol{\Theta}}^{*}\circ\mathbf{Y}\big)\right\|_F^2 + \lambda_2\,(\alpha\,\|\mathbf{B}\|_* \right.$$
$$\left. +\ (1-\alpha)\,\|\mathbf{B}\|_F^2)\right\}. \tag{9}$$

An advantage of Equation (9), over Equation (7), is the availability of a closed-form solution based on the existing results on singular value shrinkage (Mazumder, Hastie, and Tibshirani 2010) described as follows. To express the solution, let $\mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\intercal$ be the SVD of a matrix $\mathbf{D}$ where $\boldsymbol{\Sigma} = \operatorname{diag}(\{\sigma_i\})$. Define the corresponding singular value soft-thresholding (SVT) operator $\mathcal{T}_c$ by

$$\mathcal{T}_c(\mathbf{D}) = \mathbf{U}\operatorname{diag}(\{(\sigma_i - c)_+\})\mathbf{V}^\intercal \quad \text{for any } c \ge 0, \tag{10}$$

where $x_+ = \max(x, 0)$. As suggested by its name, this operator soft-thresholds the singular values of the input matrix $\mathbf{D}$ at a specified threshold c. It can be shown that the solution of Equation (9) possesses the following closed-form expression

$$\hat{\mathbf{B}} = \frac{1}{1 + 2\,(1-\alpha)\,\lambda_2'}\left\{\mathcal{T}_{\alpha\lambda_2'}\left(\mathbf{P}_{\mathbf{X}}^{\perp}\big(\mathbf{W}\circ\hat{\boldsymbol{\Theta}}^{*}\circ\mathbf{Y}\big)\right)\right\}, \tag{11}$$

where $\lambda_2' = n_1 n_2 \lambda_2/2$. The proof of this result follows from the proof of Theorem 1 in Mazumder, Hastie, and Tibshirani (2010), which uses simple subgradient arguments after reparameterizing the variable $\mathbf{B}$ of Equation (9) in terms of its singular values and singular vectors. The explicit solution (11) indicates that both the SVT procedure ($\mathcal{T}_{\alpha\lambda_2'}$) and a scaling procedure ($1/\{1+2(1-\alpha)\lambda_2'\}$) are involved in $\hat{\mathbf{B}}$. Observe that these two procedures arise separately from the nuclear norm regularization and the Frobenius norm regularization. When $\alpha = 1$ (only nuclear norm regularization), Equation (11) 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, Hastie, and Tibshirani 2010; Ma, Goldfarb, and Chen 2011) require iterative applications of SVD to n_1-by-n_2 matrices. In contrast, the computation of $\hat{\mathbf{B}}$ in Equation (11) requires only a single SVD of the matrix $\mathbf{P}_{\mathbf{X}}^{\perp}(\mathbf{W}\circ\hat{\boldsymbol{\Theta}}^{*}\circ\mathbf{Y})$ due to the application of $\mathcal{T}_{\alpha\lambda_2'}$. Specifically, to obtain $\hat{\mathbf{B}}$ with respect to the multiple choices of λ_2' (or λ_2) and α, the exact same SVD is needed. This is particularly favorable to the tuning parameter selection, and allows us to perform the k-fold cross-validation procedure (Mazumder, Hastie, and Tibshirani 2010; Xu, Jin, and Zhou 2013; Chiang, Hsieh, and Dhillon 2015) with much reduced computational burden. In all of our numerical evaluations, we choose $k = 5$. As for the most alternative matrix completion algorithms, iterative applications of SVD need to be reapplied 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 (9). 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 FRSVT procedure (Oh et al. 2015), which utilizes random projections (Halko, Martinsson, and Tropp 2011) to approximate the SVT operator. Recent work (Halko, Martinsson, and Tropp 2011) has shown that random projections can explore the low-rank structure effectively, and are suitable for constructing efficient algorithms of the approximate low-rank matrix factorizations. In FRSVT, random projections are obtained through the generation of Gaussian random matrix with the 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 article, we set $L = 150$ and $d = 5$.

4. Asymptotic Convergence Rates

Let $\|\mathbf{A}\| = \sigma_{\max}(\mathbf{A})$ and $\|\mathbf{A}\|_\infty = \max_{i,j}|A_{ij}|$ be the spectral and the maximum norms of a matrix $\mathbf{A}$, respectively. We use the symbol $\asymp$ to represent the asymptotic equivalence in order, that is, $a_n \asymp b_n$ is equivalent to $a_n = O(b_n)$ and $b_n = O(a_n)$, and $n = n_1 + n_2$. The MSE of a generic estimator $\tilde{\mathbf{A}}$ is defined as $d^2(\tilde{\mathbf{A}}, \mathbf{A}_0) = \|\tilde{\mathbf{A}} - \mathbf{A}_0\|_F^2/(n_1 n_2)$.

In this section, we first establish a general convergence result on $d^2(\hat{\mathbf{A}}, \mathbf{A}_0)$ 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 (1) are independently distributed random variables such that $\mathrm{E}\,(\epsilon_{ij}) = 0$ and $\mathrm{E}\,(\epsilon_{ij}^2) = \sigma_{ij}^2 < \infty$ for all i, j. (b) For some finite positive constants c_σ and η, $\max_{i,j}\mathrm{E}|\epsilon_{ij}|^l \le \frac{1}{2} l!\, c_\sigma^2 \eta^{l-2}$ for any positive integer $l \ge 2$.

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

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

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

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

C5. (a) There exists a lower bound $\theta_L \in (0,1)$ such that $\min_{i,j}\{\theta_{ij}\} \ge \theta_L > 0$, 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 $t > t_0$, $\mathsf{P}\{\sum_{ij}(1/\hat{\theta}_{ij} - 1/\theta_{ij})^2 \ge c_{n_1,n_2}t\} \le g(t) + h_{n_1,n_2}$, 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, Montanari, and Oh (2009), Koltchinskii, Lounici, and Tsybakov (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 MAR commonly assumed in the missing value literature (Little and Rubin 2014). A specific MAR model is the logistic regression model

$$\theta_{ij} = \theta_{ij}(\boldsymbol{x}_i) = \frac{\exp\left\{\left(1, \boldsymbol{x}_i^\intercal\right)\boldsymbol{\gamma}_{\cdot j}\right\}}{1 + \exp\left\{\left(1, \boldsymbol{x}_i^\intercal\right)\boldsymbol{\gamma}_{\cdot j}\right\}}, \tag{12}$$

where $\boldsymbol{\gamma}_{\cdot j} \in \mathbb{R}^{m+1}$ are the jth column-specific parameter vectors. Most of the existing studies in the matrix completion (Keshavan, Montanari, and Oh 2009; Gross 2011; Recht 2011; Rohde and Tsybakov 2011; Koltchinskii, Lounici, and Tsybakov 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 $N/(n_1n_2)$ 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, He, and Khatri 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 the 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_1n_2, since θ_L is allowed to go to 0 with n_1 and n_2 growing. For instance, one could take $\theta_L \asymp r_{\boldsymbol{B}_0}n\log^2(n)/n_1n_2$ to mimic scenarios with $cr_{\boldsymbol{B}_0}n\log^2(n)$ 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 Equation (12) 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

$$\Delta(\delta_\sigma, t) = \max\left\{\frac{\sqrt{(n_1 \vee n_2)\log(n)}}{\sqrt{\theta_L}n_1n_2}, (n_1n_2)^{-3/4}(c_{n_1,n_2}t)^{1/2}\log^{\delta_\sigma/4}(n)\right\} \tag{13}$$

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 $\lim_{t\to\infty}\lim_{n_1,n_2\to\infty}\{\eta_{n_1,n_2}(g, \delta_\sigma, t)\} = 0$. The following Theorem 1 is proved in Section S5 of the supplementary material.

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

$$d^2(\hat{\boldsymbol{A}}, \boldsymbol{A}_0) \le C'\max\{\min\left\{\lambda_2\alpha\|\boldsymbol{B}_0\|_*, n_1n_2r_{\boldsymbol{B}_0}(\lambda_2\alpha)^2\right\}, \lambda_2(1-\alpha)\|\boldsymbol{B}_0\|_F^2, n_1n_2\Delta^2(\delta_\sigma, t), n_2^2\lambda_1^2\|\boldsymbol{X}\boldsymbol{\beta}_0\|_F^2\} \tag{14}$$

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 Equation (14) with the 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 Equation (14), 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 Equation (14), we derive the 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 the 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 $t_0 > 0$, $g(t) = \mathrm{P}\{\chi_1^2 > t\}$, and $h_{n_1,n_2} = \sup_t |\mathrm{P}\{\theta_0(1/\hat{\theta} - 1/\theta_0)^2/(1-\theta_0) \geq t\} - g(t)|$ 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

$$t_0 < t < (n_1 n_2)^{-1/2}(n_1 \vee n_2)\log^{1-\delta_\sigma/2}(n), \qquad (15)$$

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 Equation (15), $\lambda_1 = n_2^{-1}\log^{-1/2}(n)\Delta_1$, $1-\alpha \asymp 1/(n_1 n_2)$, $\lambda_2 \asymp \theta_0^{-1/2}(n_1 \wedge n_2)^{-1/2}(n_1 n_2)^{-1/2}\log^{1/2}(n)$ in Equation (5). Then, for a positive constant C', with probability at least $1 - \eta_{n_1,n_2}(g, \delta_\sigma, t)$,

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

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 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. 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's (2012) 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, Lounici, and Tsybakov's (2011) prior mask distribution estimator $\hat{\boldsymbol{A}}^{\mathrm{KLT}}$, and Mazumder, Hastie, and Tibshirani (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 Equation (12). 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 $h_{n_1,n_2} = n_2 \max_j \sup_t |\mathrm{P}\{\sum_i (1/\hat{\theta}_{ij} - 1/\theta_{ij})^2 \geq t\} - \mathrm{P}(\chi_{m+1}^2 \geq \eta_g t)|$ in Condition C5(b) so that C5(b) holds for any positive $t > t_0$ for the logistic model.

By choosing t such that

$$m + 3 < t < \log^{\delta_\sigma/6}(n), \qquad (16)$$

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 Equation (14) 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 Equation (16), $\lambda_1 = n_2^{-1}\log^{-1/2}(n)\Delta_2(\delta_\sigma)$ 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 Equation (5). Then, for a positive constant C', with probability at least $1 - \eta_{n_1,n_2}(g, \delta_\sigma, t)$,

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

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 jth 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 Equation (8). 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 MSE of $\boldsymbol{A}_0$ achieved by our estimator and the one from Koltchinskii, Lounici, and Tsybakov (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 use of the covariates improves the estimation. This question is addressed theoretically in this section by comparing nonasymptotic upper bounds of MSE. 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

$$\hat{\boldsymbol{\beta}}^{\mathrm{UNI}} = \underset{\boldsymbol{\beta} \in \mathbb{R}^{m \times n_2}}{\arg\min} \left\{ \frac{1}{n_1 n_2} \left\| \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{P}_{\boldsymbol{X}} \left(\frac{n_1 n_2}{N} \boldsymbol{W} \circ \boldsymbol{Y} \right) \right\|_F^2 + \lambda_1 \|\boldsymbol{\beta}\|_F^2 \right\} \tag{17}$$

and

$$\hat{\boldsymbol{B}}^{\mathrm{UNI}} = \underset{\boldsymbol{B} \in \mathbb{R}^{n_1 \times n_2}}{\arg\min} \left\{ \frac{1}{n_1 n_2} \left\| \boldsymbol{B} - \boldsymbol{P}_{\boldsymbol{X}}^{\perp} \left(\frac{n_1 n_2}{N} \boldsymbol{W} \circ \boldsymbol{Y} \right) \right\|_F^2 + \lambda_2 \|\boldsymbol{B}\|_* \right\}, \tag{18}$$

when α in Equation (7) is set to 1. By writing $\hat{\boldsymbol{A}}^{\mathrm{UNI}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}^{\mathrm{UNI}} + \hat{\boldsymbol{B}}^{\mathrm{UNI}}$, the MSE $d^2(\hat{\boldsymbol{A}}^{\mathrm{UNI}}, \boldsymbol{A}_0)$ 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, Equation (18) (without the projection $\boldsymbol{P}_{\boldsymbol{X}}^{\perp}$) alone leads to the estimator $\hat{\boldsymbol{A}}^{\mathrm{KLT}}$ of Koltchinskii, Lounici, and Tsybakov (2011)

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

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, Lounici, and Tsybakov (2011) that if $\lambda_{\mathrm{KLT}} \geq 2\|\boldsymbol{M}\|$, then

$$d^2(\hat{\boldsymbol{A}}^{\mathrm{KLT}}, \boldsymbol{A}_0) \leq \lambda_{\mathrm{KLT}} \min \left\{ 2\|\boldsymbol{A}_0\|_*, \left(\frac{1+\sqrt{2}}{2} \right)^2 \lambda_{\mathrm{KLT}} n_1 n_2 r_{\boldsymbol{A}_0} \right\} =: U_{\mathrm{KLT}}, \tag{19}$$

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 $\lambda_2 \geq 2\|\boldsymbol{M}\|$,

$$d^2(\hat{\boldsymbol{B}}^{\mathrm{UNI}}, \boldsymbol{B}_0) \leq \lambda_2 \min \left\{ 2\|\boldsymbol{B}_0\|_*, \left(\frac{1+\sqrt{2}}{2} \right)^2 \lambda_2 n_1 n_2 r_{\boldsymbol{B}_0} \right\} =: U_{\mathrm{UNI}}. \tag{20}$$

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, Lounici, and Tsybakov (2011) obtained the same rate for $\|\boldsymbol{M}\|$ in a similar fashion. Due to this theoretical guarantee, we pick $\lambda_2 = \lambda_{\mathrm{KLT}} = C\theta_0^{-1/2}(n_1 \wedge n_2)^{-1/2}(n_1n_2)^{-1/2}\log^{1/2}(n)$.

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 Equation (20). 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 Equations (19) and (20) when n_1 is large enough. Since these two bounds are of the same order, we have to analyze the corresponding constant factors. Since $r_{\boldsymbol{B}_0} \leq r_{\boldsymbol{A}_0}$ and $\|\boldsymbol{B}_0\|_* \leq \|\boldsymbol{A}_0\|_*$ (Proposition S2.1 of the supplementary material), we can conclude that $U_{\mathrm{UNI}} \leq U_{\mathrm{KLT}}$. In addition, if $\boldsymbol{\beta}_0 \neq \boldsymbol{0}^{m \times n_2}$ and the rank of $\boldsymbol{A}_0$ is small, that is, 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 $\lambda_2 = \lambda_{\mathrm{KLT}} = C\theta_0^{-1/2}(n_1 \wedge n_2)^{-1/2}(n_1n_2)^{-1/2}\log^{1/2}(n)$ in both Equations (19) and (20). Then, $U_{\mathrm{UNI}} \leq U_{\mathrm{KLT}}$. 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

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

Theorem 4. Fix $a_1 > 0$, for $r_{\boldsymbol{B}_0}$ such that $1 \leq r_{\boldsymbol{B}_0} \leq \min(n_1, n_2) - m$, $(n_1 \vee n_2) r_{\boldsymbol{B}_0} \leq n_1 n_2 \theta_0$. Assume that $\omega_{ij} \sim \mathrm{Bern}(\theta_0)$ 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 $0 \leq l \leq r_{\boldsymbol{B}_0}$ such that

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

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, Lounici, and Tsybakov (2011). Comparing Theorem 4 with Corollary 1, we see that, under the iid 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 nonuniform 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 nonasymptotic 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 and n_2 in Section 4.

6. Simulation Study

This section reports results from the 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 Equation (6) and $\hat{\boldsymbol{B}}$ is given by Equation (7). 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 iid 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: *missing-at-random* and *UNI: uniform observation*. For MAR, we adopted the logistic model (12) 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 γ_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 SVT 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, Lounici, and Tsybakov (2011) (KLT), and Mazumder, Hastie, and Tibshirani (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

$$\text{test error} = \frac{\left\|\boldsymbol{W}^\star \circ (\hat{\boldsymbol{A}} - \boldsymbol{A}_0)\right\|_F^2}{\|\boldsymbol{W}^\star \circ \boldsymbol{A}_0\|_F^2} \quad \text{and}$$

$$\mathrm{RMSE}\ (\boldsymbol{A}_0) = \frac{\|\hat{\boldsymbol{A}} - \boldsymbol{A}_0\|_F}{\sqrt{n_1 n_2}},$$

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

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)

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 the 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

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)

proposed methods lose efficiency by considering a more general model.

7. Empirical Study

We demonstrate the proposed methodology by analyzing the MovieLens 100 K dataset as described in Harper and Konstan (2016). This dataset 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 9430 ratings, 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, Hastie, and Tibshirani 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 nonmonotone trend of observed rates with respect to age. To see this, we divide age into seven 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 dataset (*http://files.grouplens.org/datasets/movielens/ml-1m-README.txt*). The nonmonotonicity is demonstrated in Figure 1(a), which showed that the rate of observation peaked

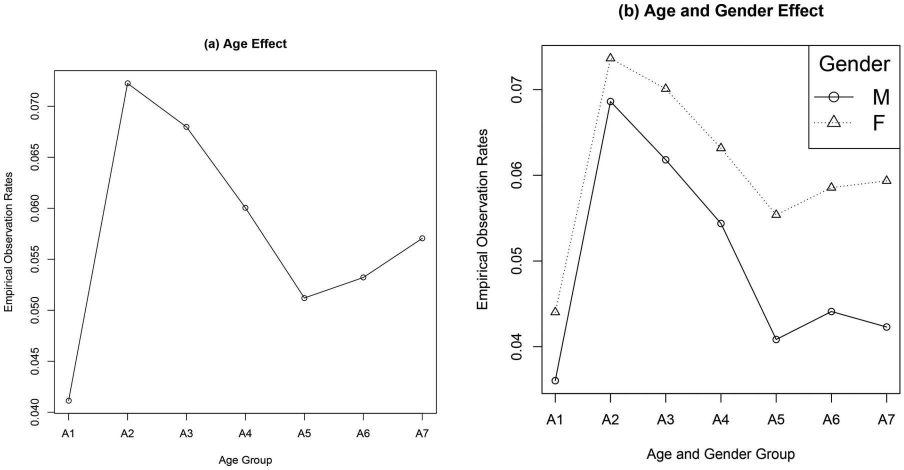

Figure 1. Empirical observation rates of the MovieLens 100 K data. Panel (a): with respect to the seven age groups; Panel (b): with respect to the 14 combination groups of age and gender.

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 subgroups 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 the age and the 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 the 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 $Q_{\text{MA1}}, \dots, Q_{\text{FA7}}, Q_{\text{Benchmark}}$ shown in Table 3. If the validation error of a model was smaller than $Q_{\text{Benchmark}}$, 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 the 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 the observation. Denote this final model for Split1 by Final1. As shown in Table 3, we note that the corresponding validation error $Q_{\text{Final1}} = 4.4297$ 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 $Q_{\text{Final2}} = 4.4230$ 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 RMSPE $= \|\boldsymbol{W}^{\text{test}} \circ (\hat{\boldsymbol{A}} - \boldsymbol{Y})\|_F / \sqrt{\sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \omega_{ij}^{\text{test}}}$, where $\boldsymbol{W}^{\text{test}}$ is the indicator matrix of test set with the (i, j)-th entry being $\omega_{ij}^{\text{test}}$. Since Test Set1 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

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

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, Lounici, and Tsybakov (2011)(KLT), and Mazumder, Hastie, and Tibshirani (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

indicated the applicability of $\widehat{\text{SVT}}$-$\hat{\alpha}$-LOG to larger datasets 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, $(n_1n_2/N)\sigma_1\boldsymbol{u}_1\boldsymbol{v}_1^{\intercal}$ 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 continuous 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 the covariate information.

8. Concluding Remarks

This paper investigates the problem of matrix completion with the 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 100 K 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 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 the existing matrix completion methods in terms of the RMSPE and the ranks of completed matrices.

Supplementary Materials

The supplementary materials contain technical details including the proofs of Propositions 1–2 and Theorems 1–4; and additional results of simulation and empirical study.

Acknowledgment

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

Funding

National Key Research Special Program of China [2016YFC0207701 and 2015CB856000] National Natural Science Foundation of China [11131002, 71371016 and 71532001] National Science Foundation [DMS-1612985 and DMS-1711952].

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. [198]

Bi, X., Qu, A., Wang, J., and Shen, X. (2016), "A Group-Specific Recommender System," *Journal of the American Statistical Association*, 112, 1344–1353. [198]

Cai, T., Cai, T. T., and Zhang, A. (2016), "Structured Matrix Completion with Applications to Genomic Data Integration," *Journal of the American Statistical Association*, 111, 621–633. [198]

Cai, T. T., and Zhou, W.-X. (2016), "Matrix Completion via Max-Norm Constrained Optimization," *Electronic Journal of Statistics*, 10, 1493–1525. [198,202]

Candès, E. J., and Plan, Y. (2010), "Matrix Completion with Noise," *Proceedings of the IEEE*, 98, 925–936. [198]

Candès, E. J., and Recht, B. (2009), "Exact Matrix Completion via Convex Optimization," *Foundations of Computational Mathematics*, 9, 717–772. [198,203]

Chiang, K.-Y., Hsieh, C.-J., and Dhillon, I. S. (2015), "Matrix Completion with Noisy Side Information," *Advances in Neural Information Processing Systems*, 28, 3447–3455. [198,201]

Feuerverger, A., He, Y., and Khatri, S. (2012), "Statistical Significance of the Netflix Challenge," *Statistical Science*, 27, 202–231. [198,202]

Freedman, D. A. (2009), *Statistical Models: Theory and Practice*, New York: Cambridge University Press. [199]

Friedman, J., Hastie, T., and Tibshirani, R. (2013), *The Elements of Statistical Learning*, New York: Springer. [200]

Gross, D. (2011), "Recovering Low-Rank Matrices from Few Coefficients in any Basis," *IEEE Transactions on Information Theory*, 57, 1548–1566. [202]

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. [201]

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. [198,207]

Keshavan, R. H., Montanari, A., and Oh, S. (2009), "Matrix Completion from Noisy Entries," *Advances in Neural Information Processing Systems*, 22, 952–960. [198,202]

Klopp, O. (2014), "Noisy Low-Rank Matrix Completion with General Sampling Distribution," *Bernoulli*, 20, 282–303. [198,202]

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. [198,199,202,203,204,205]

Little, R. J., and Rubin, D. B. (2014), *Statistical Analysis With Missing Data*, Hoboken NJ: Wiley. [202]

Ma, S., Goldfarb, D., and Chen, L. (2011), "Fixed Point and Bregman Iterative Methods for Matrix Rank Minimization," *Mathematical Programming*, 128, 321–353. [201]

Mardia, K. V., Kent, J. T., and Bibby, J. M. (1980), *Multivariate Analysis*, London: Academic Press. [199]

Mazumder, R., Hastie, T., and Tibshirani, R. (2010), "Spectral Regularization Algorithms for Learning Large Incomplete Matrices," *Journal of Machine Learning Research*, 11, 2287–2322. [201,203,205,207]

Natarajan, N., and Dhillon, I. S. (2014), "Inductive Matrix Completion for Predicting Gene–Disease Associations," *Bioinformatics*, 30, i60–i68. [198]

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. [198,202,203,205]

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. [199,201]

Recht, B. (2011), "A Simpler Approach to Matrix Completion," *Journal of Machine Learning Research*, 12, 3413–3430. [198,202]

Rohde, A., and Tsybakov, A. B. (2011), "Estimation of High-Dimensional Low-Rank Matrices," *The Annals of Statistics*, 39, 887–930. [198,202]

Srebro, N., and Salakhutdinov, R. R. (2010), "Collaborative Filtering in a Non-Uniform World: Learning with the Weighted Trace Norm," *Advances in Neural Information Processing Systems*, 23, 2056–2064. [198,202]

Sun, T., and Zhang, C.-H. (2012), "Calibrated Elastic Regularization in Matrix Completion," *Advances in Neural Information Processing Systems*, 25, 863–871. [200,202,203,205]

Sweeting, T. (1980), "Uniform Asymptotic Normality of the Maximum Likelihood Estimator," *The Annals of Statistics*, 8, 1375–1381. [202]

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. [201]

Xu, M., Jin, R., and Zhou, Z.-H. (2013), "Speedup Matrix Completion with Side Information: Application to Multi-Label Learning," *Advances in Neural Information Processing Systems*, 26, 2301–2309. [198,201]

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. [198,199,200]

Zou, H., and Hastie, T. (2005), "Regularization and Variable Selection via the Elastic Net," *Journal of the Royal Statistical Society*, Series B, 67, 301–320. [200]