Journal of the American Statistical Association

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/uasa20

Statistical Inference For Noisy Matrix Completion Incorporating Auxiliary Information

Shujie Ma, Po-Yao Niu, Yichong Zhang & Yinchu Zhu

To cite this article: Shujie Ma, Po-Yao Niu, Yichong Zhang & Yinchu Zhu (24 Apr 2024): Statistical Inference For Noisy Matrix Completion Incorporating Auxiliary Information, Journal of the American Statistical Association, DOI: 10.1080/01621459.2024.2335591

To link to this article: https://doi.org/10.1080/01621459.2024.2335591

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2024, VOL. 00, NO. 0, 1–13: Theory and Methods
https://doi.org/10.1080/01621459.2024.2335591

Statistical Inference For Noisy Matrix Completion Incorporating Auxiliary Information

Shujie Ma[a], Po-Yao Niu[a], Yichong Zhang[b], and Yinchu Zhu[c]

[a]Department of Statistics, University of California at Riverside, Riverside, CA; [b]School of Economics, Singapore Management University, Singapore; [c]Department of Economics, Brandeis University, Waltham, MA

ABSTRACT

This article investigates statistical inference for noisy matrix completion in a semi-supervised model when auxiliary covariates are available. The model consists of two parts. One part is a low-rank matrix induced by unobserved latent factors; the other part models the effects of the observed covariates through a coefficient matrix which is composed of high-dimensional column vectors. We model the observational pattern of the responses through a logistic regression of the covariates, and allow its probability to go to zero as the sample size increases. We apply an iterative least squares (LS) estimation approach in our considered context. The iterative LS methods in general enjoy a low computational cost, but deriving the statistical properties of the resulting estimators is a challenging task. We show that our method only needs a few iterations, and the resulting entry-wise estimators of the low-rank matrix and the coefficient matrix are guaranteed to have asymptotic normal distributions. As a result, individual inference can be conducted for each entry of the unknown matrices. We also propose a simultaneous testing procedure with multiplier bootstrap for the high-dimensional coefficient matrix. This simultaneous inferential tool can help us further investigate the effects of covariates for the prediction of missing entries. Supplementary materials for this article are available online, including a standardized description of the materials available for reproducing the work.

ARTICLE HISTORY
Received August 2022
Accepted March 2024

KEYWORDS
Auxiliary covariates; Factor models; Missing data; Multiplier bootstrap; Simultaneous inference

1. Introduction

Advances in modern technology have facilitated us to collect large-scale data that are naturally presented in the form of a matrix with both dimensions increasing vastly. Recovering an intact matrix from partial observations, known as the matrix completion problem, has received considerable attention in different fields. Most existing methods for estimating missing entries of a matrix only use information from its partial observations. In many real applications, auxiliary information is often available in addition to the observed entries. For example, in a recommender system that aims to predict ratings of users based on the observed ratings from others, the data often contain additional information such as user demographical profiles, apart from the observed ratings by users. Indeed, such auxiliary information can be exploited to enrich the basic model and improve prediction accuracy, especially when only a few entries are observed. Because of the increased availability of auxiliary covariates in real-world datasets, there is a pressing need to develop matrix completion techniques that can make good use of the auxiliary information. As a result, a few computational algorithms have been recently proposed to tackle this problem, see, for example, Xu, Jin, and Zhou (2013), Chiang, Hsieh, and Dhillon (2015), Zhu, Shen, and Ye (2016), Alaya and Klopp (2019), and Jin, Ma, and Jiang (2022), and see Ibriga and Sun (2023) for tensor completion with covariate information.

In this article, we consider a semi-supervised model for the matrix completion problem with row-feature information, in which a target matrix $\Theta \in \mathbb{R}^{n\times m}$ can be written as

$$\Theta = X\beta' + \Gamma, \tag{1}$$

where $X \in \mathbb{R}^{n\times d}$ is an observable row-feature matrix, $\beta \in \mathbb{R}^{m\times d}$ is an unknown coefficient matrix for X, and $\Gamma \in \mathbb{R}^{n\times m}$ is an unknown low-rank matrix driven by unobserved latent factors, so that it can be decomposed as $\Gamma = LF'$ with $L \in \mathbb{R}^{n\times r}$ and $F \in \mathbb{R}^{m\times r}$. As a result, we have $\Theta = X\beta' + \Gamma = [X, L][\beta, F]'$. Therefore, the target matrix Θ can be learned from both observed covariates in X and unobserved latent variables in L of the subjects, while β and F can be considered unknown coefficients for X and L, respectively. For example, in a recommendation system, we use the observed baseline characteristics in X and the unobserved variables in L of users to predict their ratings. This model was also mentioned in Fithian and Mazumder (2018) and Mao, Chen, and Wong (2019), and they proposed different penalized methods for estimating the parameters. Moreover, Mao, Chen, and Wong (2019) has investigated the convergence rates of their proposed penalized estimators.

Unlike Fithian and Mazumder (2018) and Mao, Chen, and Wong (2019) that focus on the estimation of the target matrix, we aim to perform statistical inferences for the unknown matrix Θ and the high-dimensional coefficient matrix β in model (1). Our goal is to provide an interval estimator associated with a given confidence level rather than a point estimator for each entry in Θ and to test the significance of the covariates for the prediction of the missing entries. For the matrix completion problem with an incomplete and noise-corrupted data matrix, estimation error bounds in terms of entry-wise, Euclidean and spectral norm losses have been established for the estimators of the unknown

CONTACT Shujie Ma shujie.ma@ucr.edu Department of Statistics, University of California at Riverside, Riverside, CA.
Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA.

low-rank matrix, obtained from various convex and nonconvex optimization algorithms (e.g., Candes and Plan 2010; Koltchinskii, Lounici, and Tsybakov 2011; Negahban and Wainwright 2012; Chen et al. 2020; Athey et al. 2021). However, confidence intervals derived directly from such bounds are expected to be too conservative, which is mainly caused by the presence of a nonnegligible bias.

To quantify the uncertainty associated with a parameter estimator, one needs to characterize the (asymptotic) distribution of the estimator. In general, this is a challenging task to accomplish when fitting a high-dimensional statistical model, as it involves nonlinear and non-explicit parametric estimation procedures (Javanmard and Montanari 2014). It can be even more difficult to derive the asymptotic distribution when the data matrix has a large number of missing entries. Thus, the literature on inference for matrix completion is still scarce. There is some recent development in statistical inference for matrix completion without the observed auxiliary covariates based on either a de-biased strategy or singular value decomposition (SVD) estimation. Carpentier et al. (2018), Chen et al. (2019), and Xia and Yuan (2021) proposed de-biased estimators to construct confidence intervals of the unknown underlying matrix with a low-rank structure. The de-biased estimators are built upon initial estimates that can be obtained from nuclear norm penalization. A sample splitting step is needed in the approach considered in Carpentier et al. (2018) and Xia and Yuan (2021). Jin, Miao, and Su (2021) proposed an iterative SVD method with the missing entries replaced by the SVD estimates from the previous step. Their estimator requires that the number of iterations diverges with the sample size to have asymptotic normality. Under a block structure assumption for the observed entries, Bai and Ng (2021) and Cahan, Bai, and Ng (2022) proposed to impute the missing values using the estimated factors and loadings obtained from applying SVD on fully observed sub-matrices. Moreover, Xiong and Pelger (2023) applied SVD to an adjusted covariance matrix computed from observed data.

Unlike the aforementioned works, we consider an iterative least squares (LS) estimation procedure and provide an inferential analysis for the parameters of model (1) with auxiliary information. The iterative LS method has become a popular approach for matrix completion due to its computational advantages (Zhou et al. 2008; Hastie et al. 2015; Sun and Luo 2016). However, the literature on the asymptotic distributions of iterative LS estimators is still scarce. Our algorithm starts from the initial estimates of β and Γ, which are obtained from ordinary LS regression and SVD of the residual matrix, respectively. Based on these initial estimates, we show that we only need to iterate the LS estimation a finite number of times, and the resulting entry-wise estimators of β, Θ and Γ are guaranteed to have asymptotic normality. As a result, a pointwise confidence interval and individual inference can be conducted for each entry of the unknown matrices. The iterative LS method enjoys low computational cost compared to the iterative SVD approach (Jin, Miao, and Su 2021), but the development of its statistical properties is quite challenging. We show that without including the covariate matrix in the model, our iterative LS estimator of the unknown low-rank matrix Γ has the same asymptotic distribution as the iterative SVD estimator proposed in Jin, Miao, and Su (2021). Because our method only requires finite iterations of LS estimation, it is computationally more efficient and much faster than their method which needs to iterate the SVD procedure a diverging number of times. This computational advantage becomes more significant as the data matrices are larger. Moreover, we allow that the observational pattern of the responses depends on the baseline covariates and its probability goes to zero as the sample size increases, whereas the existing works on inference for matrix completion require the observational probability of the responses to be independent of the baseline covariates and/or be bounded below by a constant.

It is worth noting that each column of the coefficient matrix β is a high-dimensional vector when m is large. It is of practical interest to conduct simultaneous inference for these high-dimensional column-vectors in β, which correspond to the effects of the covariates for the prediction of all missing entries jointly. To achieve this goal, we develop a Gaussian multiplier bootstrap inferential procedure, and provide theoretical justification for our bootstrap-based simultaneous inference in this high-dimensional setting. Gaussian multiplier bootstrap that involves empirical processes is considered a powerful tool for conducting tests in classical statistical problems, and has recently been successfully applied to high-dimensional regression settings (Chernozhukov, Chetverikov, and Kato 2013, 2017). Our work is the first to apply this technique to the matrix completion problem with a thorough theoretical investigation. The proposed multiplier bootstrap inferential method can help us identify the important auxiliary covariates for the prediction of all missing entries.

In model (1), the rank of matrix Γ, which is r, is unknown a priori. We propose a new information criterion (eIC) method for estimating r based on our iterative LS method, and show that the proposed eIC approach can consistently estimate r with a high probability. This method has better finite sample performance than the commonly used singular-value-based approaches for rank selection in matrix completion, and its advantage becomes more significant when the data have more missing entries.

The rest of this article is organized as follows. The proposed estimators and the theoretical results are given in Section 2 and Section 3. Section 4 provides the information criterion method for rank estimation. The simultaneous inference for β is given in Section 5. Sections 6 and 7 provide simulation studies and analysis of the MovieLens 1M dataset using the proposed method, respectively. A conclusion is given in Section 8. All technical proofs and additional numerical results are provided in the supplementary materials.

Notations. Throughout the article, $\|\cdot\|$ denotes the spectral norm, $\|\cdot\|_*$ the nuclear norm, $\|\cdot\|_F$ the Frobenius norm, and $\|\cdot\|_\infty$ the maximum absolute value of the entries of a matrix. Let $A \circ B$ be the Hadamard product of two matrices A, B of the same dimensions. Let $n \wedge m$ ($n \vee m$) denote the minimum (maximum) of n and m. For two sequences of positive numbers a_n and b_n, $a_n \ll b_n$ means $a_n = o(b_n)$ and $a_n \lesssim b_n$ means that $a_n = O(b_n)$.

2. Model and Estimation

2.1. The Model

We consider the following model:

$$Y = \Theta + \varepsilon = X\beta' + \Gamma + \varepsilon, \tag{2}$$

where $Y, \Gamma, \varepsilon \in \mathbb{R}^{n\times m}$, $X = (X_1, \ldots, X_n)' \in \mathbb{R}^{n\times d}$ in which $X_i = (1, \tilde{X}_i')'$, and $\tilde{X}_i \in \mathbb{R}^{(d-1)\times 1}$ is the vector of baseline covariates for the ith subject. Moreover, $\beta = (\beta_1, \ldots, \beta_m)' \in \mathbb{R}^{m\times d}$ with $\beta_j \in \mathbb{R}^d$, so model (2) allows the unknown coefficients of the covariates to be different across j. We assume that $\Gamma = \{\Gamma_{ij}\} = LF'$ with $L = (L_1, \ldots, L_n)' \in \mathbb{R}^{n\times r}$ and $F = (F_1, \ldots, F_m)' \in \mathbb{R}^{m\times r}$. We let r and d be fixed. To identify β, we assume that $E(L_i|\tilde{X}_i) = 0$, and F_j are independent of $\tilde{X}_i$ and L_i. We do not observe all entries in $Y = (Y_{i,j})$, so let $\Xi = (\xi_{i,j}) \in \mathbb{R}^{n\times m}$ with each entry $\xi_{i,j} \in \{0, 1\}$ denoting the status of $Y_{i,j}$: $\xi_{i,j} = 1$ if and only if $Y_{i,j}$ is observed.

We assume that $P(\xi_{i,j} = 1|\tilde{X}_i) = \eta(\gamma_{0,n} + \tilde{X}_i'\gamma_1) = \pi_i$, where $\gamma_{0,n} = \log(\alpha_n) + \gamma_0$ and $\eta(\cdot)$ is the logit link function for logistic regression, so the probability of the observed rate depends on the baseline characteristics of each subject. We allow $\alpha_n \to 0$ as $n \to \infty$, so $\pi_i \to 0$ as $n \to \infty$. The probability of the observed responses π_i can be written as $\pi_i = \alpha_n e^{(\gamma_0 + \tilde{X}_i'\gamma_1)}/\{1 + e^{(\gamma_{0,n} + \tilde{X}_i'\gamma_1)}\}$, so the rate of α_n determines how fast π_i can go to zero, which will be discussed in Section 3.

2.2. The Estimation Procedure

2.2.1. Initial Estimators

To obtain an initial estimator of β, we compute the ordinary LS estimator $\hat{\beta} = (\hat{\beta}_1, \ldots, \hat{\beta}_m)'$ without considering the latent matrix Γ, so each $\hat{\beta}_j$ is obtained by

$$\hat{\beta}_j = \left(\sum_{i=1}^n X_i X_i' \xi_{i,j}\right)^{-1} \left(\sum_{i=1}^n X_i Y_{i,j} \xi_{i,j}\right). \tag{3}$$

Next, we obtain an SVD estimate of Γ as follows. Define

$$W_{i,j} = \hat{\pi}_i^{-1} \xi_{i,j} (Y_{i,j} - X_i' \hat{\beta}_j). \tag{4}$$

where $\hat{\pi}_i = \eta(\hat{\gamma}_{0,n} + \tilde{X}_i'\hat{\gamma}_1)$ is the estimated observation rate for the ith subject, in which

$$\hat{\gamma} = (\hat{\gamma}_{0,n}, \hat{\gamma}_1) = \arg\min_{r_0, r_1} \sum_{i=1}^n \sum_{j=1}^m Y_{ij}(r_0 + \tilde{X}_i' r_1) - \log(1 + \exp(r_0 + \tilde{X}_i' r_1)).$$

We perform SVD on W such that $W = UDV' = \sum_{s=1}^{m\wedge n} d_s u_s v_s'$ where d_s's are the singular values in D in decreasing order and u_s's, v_s's are the corresponding left and right singular vectors in U and V. Then for a given rank r, the SVD estimator of Γ is $\hat{\Gamma} = (\hat{\Gamma}_{i,j}) = \hat{L}\hat{F} = \sum_{s=1}^r d_s u_s v_s'$ where $\hat{L} = \sqrt{n}\,(u_1, \ldots, u_r)$ and $\hat{F} = 1/\sqrt{n}\,(d_1 \cdot v_1, \ldots, d_r \cdot v_r)$.

2.2.2. The Iterative LS Estimators

The initial estimator $\hat{\Gamma} = \hat{L}\hat{F}'$ is actually the minimizer of the following function:

$$f(\hat{\beta}, L, F) = \left\| LF' - \mathrm{diag}\,(\hat{\pi})^{-1}(Y - X\hat{\beta}') \circ \Xi \right\|_F^2, \tag{5}$$

where $\hat{\pi} = (\hat{\pi}_1, \ldots, \hat{\pi}_n)'$, and $\mathrm{diag}\,(\hat{\pi})$ is an $n \times n$ diagonal matrix with the diagonals being $\hat{\pi}_1, \ldots, \hat{\pi}_n$ and the off-diagonal entries equal to zeros.

In the above function, the missing values are treated as zeros and contribute to the residuals while fitting, so the resulting estimates may not be optimal as they ignore the information about the missing positions. To solve this problem, we consider another objective function in which the missing entries do not contribute to the residuals, and propose an updating procedure (algorithm) that iteratively updates the estimates using the estimates given in Section 2.2.1 as the initial values.

We define the following objective function where the missing values do not contribute to the residuals:

$$f^*(\beta, L, F) = \left\| \Xi \circ [\, LF' - (Y - X\beta')\,] \right\|_F^2. \tag{6}$$

Although it is hard to find out the joint minimizers of (6) explicitly, we can easily obtain its minimizer of each β, L, and F if the other two are fixed at their current values by solving an LS problem. Therefore, we can consider the following updating procedure:

$$\begin{aligned} \tilde{\beta}^{(g)} &= \arg\min_\beta f^*\left(\beta, \tilde{L}^{(g-1)}, \tilde{F}^{(g-1)}\right); \\ \tilde{F}^{(g)} &= \arg\min_F f^*\left(\tilde{\beta}^{(g)}, \tilde{L}^{(g-1)}, F\right); \\ \tilde{L}^{(g)} &= \arg\min_L f^*\left(\tilde{\beta}^{(g)}, L, \tilde{F}^{(g)}\right) \end{aligned} \tag{7}$$

for any given $g \geq 1$, where g is the step index in the iterative algorithm. This algorithm requires initial values for $\tilde{L}$ and $\tilde{F}$ to start with. An obvious option is to use $\tilde{L}^{(0)} = \hat{L}$, and $\tilde{F}^{(0)} = \hat{F}$ given in Section 2.2.1. Then the resulting estimator of Γ at the gth step is

$$\tilde{\Gamma}^{(g)} = \tilde{L}^{(g)} \tilde{F}^{(g)'}. \tag{8}$$

We call the corresponding estimators $\tilde{\beta}^{(g)}$ and $\tilde{\Gamma}^{(g)}$ the *iterative LS estimators*.

3. Asymptotic Theory

3.1. Assumptions

We make the following assumptions to investigate the asymptotic theories about the proposed iterative LS estimator.

Assumption 1.

(i) $Y_{i,j} \perp \xi_{i,j} | (\tilde{X}_i, L_i, F_j)$.

(ii) Assume $\xi_{i,j} = 1\{\gamma_{0,n} + \tilde{X}_i'\gamma_1 \geq \nu_{i,j}\}$, where $\{\nu_{i,j}\}_{1\leq i\leq n, 1\leq j\leq m}$ is a sequence of iid logistic random variables independent of (ϵ, X, L, F). Denote $P(\xi_{ij} = 1|\tilde{X}_i) = \eta(\gamma_{0,n} + \tilde{X}_i'\gamma_1) = \pi_i$, where $\gamma_{0,n} = \log(\alpha_n) + \gamma_0$, $\eta(\cdot)$ is the standard logistic cdf, and $\alpha_n \leq 1$ is some deterministic sequence.

(iii) $n^{1/2}\log(m + n) \lesssim m \lesssim n^2$ and $(n \vee m)^\varrho \ll (n \wedge m)\alpha_n^2$ for an arbitrarily small constant $0 < \varrho < 1$.

(iv) There exists a constant $C > 0$ such that $\sup_{||u||\leq C} \left\| \frac{1}{n\alpha_n} \sum_{i=1}^n \Lambda_i(u)(1 - \Lambda_i(u)) X_i X_i' - H_0 \right\| = o_P(1)$, where $\Lambda_i(u) = \Lambda_i(u_0, u_1) = \eta((\gamma_{0,n} + u_0(nm\alpha_n)^{-1/2}) + \tilde{X}_i'(\gamma_1 + u_1(nm\alpha_n)^{-1/2}))$, and $H_0 = E\{\exp(\gamma_0 + \tilde{X}_i'\gamma_1) X_i X_i'\}$ is positive definite.

(v) Entries of $\tilde{X}_i$ have sub-Gaussian norms bounded by a constant, $\Sigma_X = E\left(X_i X_i'\right)$ and $\alpha_n^{-1} E\left(\pi_i X_i X_i'\right)$ have eigenvalues bounded away from zero and infinity for $i = 1, \ldots, n$.

(vi) Entries of L_i, F_j, and $\varepsilon_{i,j}$ have sub-Gaussian norms bounded by a constant.
(vii) For some constant $c > 0$,

$$P\left(\sigma_r(L'L/n) > c\right) \to 1, \quad P\left(\sigma_r(F'F/m) > c\right) \to 1,$$
$$P\left(\sigma_r\left(\sum_{i=1}^{n} L_i L_i' \pi_i/(\alpha_n n)\right) > c\right) \to 1,$$
$$P\left(\sigma_r\left(\sum_{i=1}^{n} X_i X_i' \pi_i/(\alpha_n n)\right) > c\right) \to 1,$$

where $\sigma_r(A)$ is the rth largest singular value of A.
(viii) Conditional on (X, L, F), $\varepsilon_{i,j}$ is independent across (i, j) with $E(\varepsilon_{i,j} \mid L, F, X) = 0$. $\{\tilde{X}_i\}_{i=1}^n$, $\{L_i\}_{i=1}^n$ and $\{F_j\}_{j=1}^m$ are sequences of iid random variables, respectively.

With the above assumptions and the model identification assumption that $E(\Gamma_{ij}|X_i) = 0$, we can first show that $\hat{\beta}$ and $\hat{\Gamma}$ are consistent estimators of β and Γ (see supplement B).

Remark (Comments about Assumption 1). Assumption (i) assumes that the response $Y_{i,j}$ and the variable for missingness $\xi_{i,j}$ are independent conditional on the observed covariates and the latent variables. Assumption (ii) assumes that the missingness of each response depends on the observed baseline covariates of each individual, and the probability of the missing pattern is modeled through a logistic regression model, which is called the propensity score function (Rosenbaum and Rubin 1983). This assumption is more relaxed and practical than the "missing uniformly at random" condition imposed in Carpentier et al. (2018), Chen et al. (2019), and Jin, Miao, and Su (2021) for statistical inference. For example, in the MovieLens data in Section 7, whether users rate a movie or not often depends on their baseline characteristics, including gender, age, etc. It is worth noting that when the baseline characteristics for movies, denoted by Z_j for the jth movie, are observed, it is possible to include both $\tilde{X}_i$ and Z_j in the logistic model such that $\xi_{i,j} = 1\{\gamma_{0,n} + \tilde{X}_i'\gamma_1 + Z_j'\gamma_2 \geq \nu_{i,j}\}$. Our proposed estimation procedure and its statistical properties can be extended to this model. We can also consider a logistic model for missing probabilities by including the entries of a latent low-rank matrix, denoted by $A_{i,j}$, so $\xi_{i,j} = 1\{\gamma_{0,n} + \tilde{X}_i'\gamma_1 + A_{i,j} \geq \nu_{i,j}\}$. The estimation of the unobserved $A_{i,j}$ in this nonlinear model and the development of the associated statistical properties are nontrivial. We leave the study of extending our method to these two models for future work.

Assumption (iii) provides the order requirement for n and m. It is typically assumed to ensure the asymptotic properties of the estimators of L and F in factor models; see Bai and Ng (2002) and Jin, Miao, and Su (2021). Moreover, we allow α_n to decay to zero in polynomial order of $n \vee m$. Given the sub-Gaussianity of $\tilde{X}_i$ in Assumption (v), one has $\pi := E(\pi_i) = O(\alpha_n)$, so α_n is the observation rate that controls how fast the probability of the observed responses goes to zero. If n and m are of the same order, then the main restriction in Assumption (iii) is that α_n cannot decay to zero faster than $n^{-1/2}$. If only concerning the estimation in matrix completion models, α_n can decay to zero faster than the order given in Assumption (iii). For example, Klopp (2014) provided a Frobenius norm-based estimation error bound for the nuclear-norm penalized estimators under the condition that the observation rate is polylog$(n)/n$. To establish the distributional theory and uniform convergence rate of our iterative LS estimators in matrix completion, we require a higher observation rate, as the higher-order terms in our estimator involve the term $O(\alpha_n^{-2})$. To make the higher-order terms negligible so the resulting estimator can have an asymptotic linear expansion, α_n needs to satisfy the condition given in Assumption (iii).

Assumptions (iv)–(vii) are the moment and distribution conditions on the covariates, latent variables, and error terms. These are typical conditions for convergence rates and asymptotic analysis; see similar assumptions in Bai and Ng (2002), Chen et al. (2019), and Jin, Miao, and Su (2021). Specifically, Assumption (iv) can be directly verified by the uniform law of large numbers. One sufficient condition for H_0 being a positive definite matrix is that X_i has compact support and EX_iX_i' is of full rank. This condition is common for sparse logistic regressions; see, for example, Graham (2020, Assumption 3). Under the first condition in Assumption (viii), model (1) is correctly specified for the conditional mean of the responses. The second condition can be relaxed to that $\{\tilde{X}_i\}_{i=1}^n$, $\{L_i\}_{i=1}^n$ and $\{F_j\}_{j=1}^m$ are sequences of independent random variables. Our theoretical results still hold under this relaxed condition.

The following two theorems provide the asymptotic representations of $\tilde{\beta}^{(g)}$, $\tilde{\Gamma}^{(g)}$ and their proofs are left in supplement B–D.

Theorem 1. Let Model (2) and Assumption 1 hold and $\kappa_n = (1/n+1/m)\alpha_n^{-3/2}n^{1/q}$ for a constant $q > 0$ that can be arbitrarily large. The estimator $\tilde{\beta}^{(g)}$ obtained from the updating procedure (7) has the following asymptotic representation for any finite $g \geq 1$,

$$\left\|\tilde{\beta}^{(g)} - \beta - n^{-1}\left(\Xi \circ \varepsilon + \operatorname{diag}(\pi)LF'\right)' X(E\pi_i X_i X_i')^{-1}\right\|_{2,\infty} = O_P(\kappa_n),$$

where $\pi = (\pi_1, \ldots, \pi_n)'$, $\operatorname{diag}(\pi)$ is a diagonal matrix with π as the diagonal, and $\|\cdot\|_{2,\infty}$ is the maximum row 2-norm of a matrix.

Theorem 2. Let Model (2) and Assumption 1 hold. The estimator $\tilde{\Gamma}^{(g)}$ defined in (8) has the following asymptotic representation, for any finite $g \geq 1$, $\left\|\tilde{\Gamma}^{(g)} - \Gamma - \Delta\right\|_\infty = O_P(\kappa_n)$, where Δ is a $n \times m$ matrix with its (i, j)th entry being

$$\Delta_{i,j} = \frac{1}{n}L_i'(EL_iL_i'\pi_i)^{-1}\sum_{k=1}^{n} L_k\xi_{k,j}\varepsilon_{k,j} + \frac{1}{m\pi_i}\sum_{t=1}^{m} F_t'\xi_{i,t}\varepsilon_{i,t}\Sigma_F^{-1}F_j - \frac{1}{n}X_i'E(X_iX_i'\pi_i)^{-1}\sum_{k=1}^{n}\pi_k X_k L_k' F_j, \text{ where } \Sigma_F = E\left(F_jF_j'\right).$$

Remark. Point-wise confidence intervals or inference for each component in β_j and Γ can be constructed based on the asymptotic representations given in Theorems 1 and 2. In addition, we propose a multiplier bootstrap statistic in Section 5 for conducting simultaneous inference on the high-dimensional matrix β.

Remark. The iterative estimation algorithm starts from the initial ordinary LS estimator $\hat{\beta}$ given in (3). Under Assumption 1 (i), according to the derivation given in Lemma 1 of the supplement, we can obtain the asymptotic variance of $\hat{\beta}_j$, denoted by $\widetilde{\text{var}}\left(\hat{\beta}_j\right)$, as $\widetilde{\text{var}}\left(\hat{\beta}_j\right) = n^{-1}(E\pi_i X_i X_i')^{-1}\{\text{var}(\xi_{i,j}X_i\varepsilon_{i,j}) + \text{var}(\xi_{i,j}X_i L_i' F_j)\}(E\pi_i X_i X_i')^{-1}$. Moreover, from Theorem 1, we obtain the asymptotic variance of the iterative estimator $\tilde{\beta}_j^{(g)}$, for $g \geq 1$, denoted by $\widetilde{\text{var}}\left(\tilde{\beta}_j^{(g)}\right)$, as $\widetilde{\text{var}}\left(\tilde{\beta}_j^{(g)}\right) = n^{-1}(E\pi_i X_i X_i')^{-1}\{\text{var}(\xi_{i,j}X_i\varepsilon_{i,j}) + \text{var}(\pi_i X_i L_i' F_j)\}(E\pi_i X_i X_i')^{-1}$. Given that $E(\xi_{i,j}X_i L_i' F_j | X_i, L_i, F_j) = \pi_i X_i L_i' F_j$, one has $\text{var}(\xi_{i,j}X_i L_i' F_j) = E\{\pi_i(1-\pi_i)\Gamma_{ij}^2 X_i X_i'\} + \text{var}(\pi_i X_i L_i' F_j)$. Thus,

$$\widetilde{\text{var}}\left(\hat{\beta}_j\right) - \widetilde{\text{var}}\left(\tilde{\beta}_j^{(g)}\right) = n^{-1}(E\pi_i X_i X_i')^{-1} E\{\pi_i(1-\pi_i)\Gamma_{ij}^2 X_i X_i'\}(E\pi_i X_i X_i')^{-1} \geq 0.$$

This means that $\tilde{\beta}_j^{(g)}$ always has a smaller asymptotic variance than $\hat{\beta}_j$. In fact, when the observation rate $\alpha_n = o(1)$, $\text{var}(\pi_i X_i L_i' F_j)$ in $\widetilde{\text{var}}\left(\tilde{\beta}_j^{(g)}\right)$ is asymptotically negligible compared to $\text{var}(\xi_{i,j}X_i L_i' F_j)$ in $\widetilde{\text{var}}\left(\hat{\beta}_j\right)$, and thus the difference between the asymptotic variances of the initial and iterative estimators is larger when more observations are missing.

Remark. Without the existence of the covariate matrix X, model (2) becomes $Y = \Gamma + \varepsilon$. When $\pi_i = \pi$ such that missingness does not depend on covariates as considered in Jin, Miao, and Su (2021), our iterative LS estimator $\tilde{\Gamma}^{(g)} - \Gamma$ at any finite $g \geq 1$ has asymptotic representation: $\pi^{-1}\left(n^{-1}L\Sigma_L^{-1}L'(\varepsilon \circ \Xi) + m^{-1}(\varepsilon \circ \Xi)F\Sigma_F^{-1}F'\right)$. Therefore, it achieves the same efficiency as the iterative PCA estimator given in Jin, Miao, and Su (2021), and it has been shown in Jin, Miao, and Su (2021) that the iterative estimator has smaller asymptotic variance than the initial estimator $\hat{\Gamma}$ obtained from one-step PCA when $\pi < 1$. We also note that to achieve such efficiency improvement, Jin, Miao, and Su (2021) need the number of iterations to go to infinity, while our iterative LS estimator only requires a few iterations.

Remark. Based on the asymptotic linear expansions given in Theorems 1 and 2, one can immediately obtain the error bounds of our iterative LS estimators in Frobenius norm:

$$||\tilde{\beta}^g - \beta||_F^2/m = O_P(((n\wedge m)\alpha_n)^{-1}\log n); ||\tilde{\Gamma}^g - \Gamma||_F^2/(nm) = O_P(((n\wedge m)\alpha_n)^{-1}\log n). \quad (9)$$

Since our iterative LS estimators are asymptotically unbiased, the rate in (9) comes from the asymptotic variance. Under a similar model as ours, Mao, Chen, and Wong (2019) proposed a regularized estimation method penalizing the nuclear and Frobenius norms, and derived the convergence rate of the estimators for β and Γ. Without incorporating covariates, regularization methods based on different norms have been studied in the matrix completion problems; see, for example, Klopp (2014) and Cai and Zhou (2016). In general, the regularized estimators have an inherent bias term from the penalties that can go into the convergence rate in addition to the rate from the asymptotic variance. To conduct inference, a debiasing procedure is often needed for the regularized estimation, which is nontrivial in matrix completion problems. Our iterative estimators are asymptotically unbiased and have an asymptotic linear representation based on which we can conduct inference. Moreover, the iterative LS estimation enjoys computational convenience, which is important for modern large-scale data analysis.

4. Rank Estimation

In practice, $r = \text{rank}(\Gamma)$ is often unknown and needs to be estimated. In this section, we introduce a mean-square-error (MSE)-based approach to estimating the rank. This method fully takes advantage of the proposed iterative LS estimates, and it is described as follows. We compute $\hat{\beta}$ following (3). Recall W defined in (4) and its SVD $\sum_{s=1}^{m\wedge n} d_s u_s v_s'$. We then define $\hat{L}^k\hat{F}^{k\prime} = \sum_{s=1}^{k} d_s u_s v_s'$ as the analogues of $\hat{L}, \hat{F}$ in Section 2.2.1 with a superscript k denoting the rank used. Note that the true rank is unknown, and thus k could vary and is not necessarily equal to r. We then consider an estimation procedure similar to (7) but without updating β:

$$\begin{aligned}\tilde{F}^{k,(g+1)} &= \arg\min_F f^*\left(\hat{\beta}, \tilde{L}^{k,(g)}, F\right);\\ \tilde{L}^{k,(g+1)} &= \arg\min_L f^*\left(\hat{\beta}, L, \tilde{F}^{k,(g)}\right),\end{aligned} \quad (10)$$

where $f^*(\cdot)$ is defined in (6). The initial value $\tilde{L}^{k,(0)}$ is set as $\hat{L}^k$.

Given a fixed positive integer g and for any $k \ll n \wedge m$, we define the following function

$$\text{mse}(k,g) = \frac{1}{nm}\left\|\Xi \circ \left(Y - X\hat{\beta} - \tilde{\Gamma}^{k,(g)}\right)\right\|_F^2,$$

where $\tilde{\Gamma}^{k,(g)} = \tilde{L}^{k,(g)}\tilde{F}^{k,(g)'}$ is the rank k iterative LS estimator of Γ at step $g \geq 1$.

We define the MSE-based rank estimating criterion and the resulting estimator of the rank given as follows.

$$\begin{aligned}\text{eIC}(k\,|\,g) &= \log\text{mse}(k,g) + k\cdot h(n,m),\\ \hat{r}^{\text{eIC}(g)} &= \arg\min_{1\leq k\leq \bar{r}} \text{eIC}(k\,|\,g),\end{aligned} \quad (11)$$

for $k \geq 1$ and a predetermined upper bound $\bar{r}$, where $k\cdot h(n,m)$ is a penalty function that depends on n, m. The theorem for the statistical guarantee of $\hat{r}^{\text{eIC}(g)}$ is stated below, and its proof is in supplement E.

Theorem 3. Let Model (2) and Assumption 1 hold. Assume that $\bar{r}$ is fixed and satisfies $\bar{r} \geq r$. The rank estimator $\hat{r}^{\text{eIC}(g)}$ defined in (11) satisfies $P\left(\hat{r}^{\text{eIC}(g)} = r\right) \to 1$ if $h(n,m) = o(1)$ and $\sqrt{\frac{mn\alpha_n}{(m+n)}}h(n,m) \to \infty$ in a polynomial rate in $n \vee m$.

Remark. Theorem 3 shows that the MSE-based rank estimator $\hat{r}^{\text{eIC}(g)}$ can consistently estimate the true rank r when $h(n,m)$ satisfies certain conditions. Section 6.2 provides a formula for calculating $h(n,m)$ in our numerical analysis.

Remark. We have an interesting finding that the MSE-based method for rank selection cannot be constructed based on the initial estimates $\hat{\beta}$, $\hat{L}^k$ and $\hat{F}^k$, where $\hat{L}^k$ and $\hat{F}^k$ are rank k SVD estimates of L and F, because the MSE value may not be decreasing as k increases when the observation rate is small. A heuristic argument and the numerical illustration are given in Section G of the supplementary materials.

5. Bootstrap Inference of β

In this section, we provide a testing procedure for the null hypothesis:

$$H_0 : A_j\beta_j = a_j^0 \qquad \forall j \in \mathcal{G} \tag{12}$$

where each $A_j \neq 0$ is a given matrix with dimension $q \times k$, and $q \leq k$, each a_j^0 is a q-dimensional vector, and $\mathcal{G}$ is a subset of $\{1, \ldots, m\}$. By Theorem 1,

$$\tilde{\beta}_j^{(g)} - \beta_j = n^{-1}\sum_{i=1}^n \omega_{i,j} + \text{smaller terms},$$

where $\omega_{i,j} = E(\pi_i X_i X_i')^{-1} X_i \left(\xi_{i,j}\varepsilon_{i,j} + \pi_i \Gamma_{i,j}\right)$. Therefore, a simple test statistic is

$$T = \max_{j\in\mathcal{G}} \|A_j\tilde{\beta}_j^{(g)} - a_j^0\|_\infty. \tag{13}$$

We can use a simple multiplier bootstrap procedure to compute the p-value. Define

$$\hat{\omega}_{i,j} = \left(n^{-1}\sum_{i=1}^n \hat{\pi}_i X_i X_i'\right)^{-1} X_i \left(\xi_{i,j}\hat{\varepsilon}_{i,j} + \hat{\pi}_i\tilde{\Gamma}_{i,j}\right),$$

where $\hat{\varepsilon}_{i,j} = Y_{i,j} - X_i'\tilde{\beta}_j - \tilde{\Gamma}_{i,j}$, in which $\tilde{\beta}_j$ and $\tilde{\Gamma}_{i,j}$ are the iterative LS estimates of β_j and $\Gamma_{i,j}$ at the last step.

Let $\{\iota_i\}_{i=1}^n$ be random variables generated from $N(0,1)$ that are independent of the data. The bootstrapped test statistic is

$$T^* = \max_{j\in\mathcal{G}} \left\| n^{-1}\sum_{i=1}^n \iota_i A_j \hat{\omega}_{i,j} \right\|_\infty. \tag{14}$$

Conditional on the data, the randomness of T^* comes from the generated variables $\{\iota_i\}_{i=1}^n$. By generating many realizations of T^*, we can compute the $(1-\alpha)$ quantile of T^* conditional on the data, that is, $\mathcal{Q}(T^*, 1-\alpha)$ satisfies $P(T^* \leq \mathcal{Q}(T^*, 1-\alpha) \mid \text{data}) = 1-\alpha$.

Assumption 2. Suppose that the following conditions hold:

i) There exists a constant $M_1 > 0$ such that $\min_{i,j} E(\varepsilon_{i,j}^2 \mid X, L, F) \geq M_1$ almost surely.
ii) $(\log m)^{5/2} \ll \min\{nm^{-1/2}, mn^{-1/2}\}$.

Theorem 4. Let Model (2) and Assumptions 1 and 2 hold. Under the null hypothesis (12), if $|\mathcal{G}| \leq m$, then $P\left(T > \mathcal{Q}(T^*, 1-\alpha)\right) = \alpha + o(1)$, where T and T^* are defined in (13) and (14), and $\mathcal{Q}$ is the quantile function.

6. Simulation Studies

In this section, we conduct simulation studies to illustrate the finite sample performance of our proposed iterative LS method. We generate the responses by model (2): $Y = X\beta' + \Gamma + \epsilon$, where $\Gamma = LF'$, in which $L \in \mathbb{R}^{n\times r}$ and $F \in \mathbb{R}^{m\times r}$. We then generate the covariates, the coefficients, and the latent matrices as follows. For $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, m$, we independently generate the covariates by $X_i \sim N(0, \Sigma_X)$, the hidden matrix by $L_i \sim N(0, \Sigma_L)$, $F_j \sim N(0, 4\Sigma_F)$, and the noise by $\varepsilon_{i,j} \sim N(0,1)$. The covariance matrices are $(\Sigma_X)_{k,k'} = \text{cov}(X_{i,k}, X_{i,k'}) = 0.5^{|k-k'|}$, $(\Sigma_L)_{k,k'} = \text{cov}(L_{i,k}, L_{i,k'}) = 0.5^{|k-k'|}$ and $(\Sigma_F)_{k,k'} = \text{cov}(F_{j,k}, F_{j,k'}) = 0.2^{|k-k'|}$. We generate the coefficients by $\beta_j \sim N(0, 4I_d)$. We regenerate (X, L, F, ε) for each simulation replicate while β remains fixed. The dimension of X_i and β_j is $d = 3$ while the rank, $r = 3$, is considered for the latent factor matrix.

Next, we generate the observed entries of the responses according to the two data-generating processes (DGPs) with constant and covariate-dependent observation rates, respectively. For each type of DGP, we consider $n = m = 200, 500, 1000$ to see how the estimators and their asymptotic properties behave in different sample sizes.

DGP 1 (Constant observation rate π). In this design, we consider constant observed rates and run simulations for $\pi = 1, 0.8, 0.5, 0.2$ to see how the observed rate would affect the performance (the data is fully observed when $\pi = 1$).

DGP 2 (Covaraiate-dependent observation rate π_i). In this design, we let the observational rates of the response variables depend on the observed covariates of each individual, so we generate π_i from the logistic model: $P(\xi_{i,j} = 1 | X_i) = \pi_i = \eta(\gamma_{0,n} + X_i'\gamma_1)$, where $\gamma_{0,n} = \log(\alpha_n)$ with $\alpha_n = Cn^{-1/2}\log n$ and $\gamma_1 = (0.2, \ldots, 0.2)'$. We see that α_n controls the sparseness of the observed values of each response, and we allow that $\alpha_n \to 0$ as $n \to \infty$, so that $\pi_i \to 0$ as $n \to \infty$. We let $C = 1.0, 1.5, 2.0$. When the C value is larger, it corresponds to larger observation rates of the responses.

When data are generated from DGP 1, we compare the performance of our proposed iterative LS method with that of the iterative PCA method given in Jin, Miao, and Su (2021). In Jin, Miao, and Su (2021), they assume that the observed rate is a constant π which is bounded below by a constant. As a result, their setting only satisfies the condition on the observation rate in DGP1, not the one in DGP2.

Without the presence of the covariates X, Jin, Miao, and Su (2021) proposed to estimate Γ using an iterative PCA method with the missing values of Y replaced by the PCA estimate of Γ from the previous step. To make the iterative PCA method in Jin, Miao, and Su (2021) be accommodated to our model (2), once we obtain the estimate of Γ by PCA, we use the same LS method to obtain the estimate for β. To distinguish the estimators from our proposed method and the one from Jin, Miao, and Su (2021), we denote our gth step iterative LS estimator by $\tilde{\Gamma}_{ls}^{(g)}$, and their iterative PCA estimator by $\tilde{\Gamma}_{pca}^{(g)}$. The estimator $\tilde{\Gamma}_{ls}^{(g)}$ is obtained as described in Section 2.2.2. To adapt the iterative PCA method given in Jin, Miao, and Su (2021) for our model, at the gth step, $g \geq 1$, we replace the missing values in W by the corresponding values of the estimates obtained from the previous step, and then $\tilde{\Gamma}_{pca}^{(g)}$ is the rank r SVD of the updated W. Once the estimate of Γ is obtained, the estimate of β is obtained by the same LS method. The same initial estimator $\tilde{\Gamma}_{pca}^{(0)} = \hat{\Gamma}$ is used. In each simulation, we obtain $\tilde{\Gamma}_{pca}^{(g)}$ and $\tilde{\Gamma}_{ls}^{(g)}$ at the steps $g = 1, 2, 3$ and $g \to \infty$. The estimate at convergence denoted by $\tilde{\Gamma}^{(c)}$ is obtained by iterating the algorithm until convergence, that is, the maximum difference between the estimates from two consecutive steps, $\|X\tilde{\beta}^{(g)} + \tilde{\Gamma}^{(g)} - X\tilde{\beta}^{(g-1)} - \tilde{\Gamma}^{(g-1)}\|_\infty^2$, is smaller than the small threshold 10^{-6}.

The iterative PCA method in Jin, Miao, and Su (2021) requires that the number of iterations go to infinity to have the desired convergence rate and the asymptotic distribution of the estimator for Γ. We will show that our iterative LS estimator for Γ only needs a finite number of iterations to achieve the same asymptotic distribution, so our method enjoys great computational advantage, especially in the large dimensional

Table 1. The MSE of different estimators in DGP 1.

DGP 1		Initial		Iterative PCA				Iterative LS			
n, m	π	$\hat{\beta}$	$\hat{\Gamma}$	$\tilde{\beta}^{(3)}_{pca}$	$\tilde{\beta}^{(c)}_{pca}$	$\tilde{\Gamma}^{(3)}_{pca}$	$\tilde{\Gamma}^{(c)}_{pca}$	$\tilde{\beta}^{(3)}_{ls}$	$\tilde{\beta}^{(c)}_{ls}$	$\tilde{\Gamma}^{(3)}_{ls}$	$\tilde{\Gamma}^{(c)}_{ls}$
200	0.2	0.614	7.469	0.258	0.157	3.075	0.419	0.197	0.176	0.631	0.457
	0.5	0.230	1.121	0.126	0.125	0.321	0.283	0.125	0.125	0.285	0.285
	0.8	0.145	0.430	0.119	0.119	0.258	0.258	0.119	0.119	0.258	0.258
	1	0.117	0.250	–	–	–	–	–	–	–	–
500	0.2	0.224	2.354	0.068	0.057	0.835	0.150	0.058	0.058	0.152	0.152
	0.5	0.087	0.399	0.047	0.047	0.116	0.108	0.047	0.047	0.108	0.108
	0.8	0.054	0.167	0.044	0.044	0.099	0.099	0.044	0.044	0.099	0.099
	1	0.043	0.096	–	–	–	–	–	–	–	–
1000	0.2	0.110	0.726	0.030	0.029	0.246	0.074	0.029	0.029	0.074	0.074
	0.5	0.044	0.195	0.024	0.024	0.057	0.055	0.024	0.024	0.055	0.055
	0.8	0.028	0.084	0.023	0.023	0.050	0.050	0.023	0.023	0.050	0.050
	1	0.023	0.048	–	–	–	–	–	–	–	–

Table 2. Computing time in seconds* in each setting.

DGP 1		Time in sec to get estimators			Number of iterations		Ave. time for 1 iteration	
n, m	π	$\hat{\Gamma}$	$\tilde{\Gamma}^{(c)}_{ls}$	$\tilde{\Gamma}^{(c)}_{pca}$	$\tilde{\Gamma}^{(c)}_{ls}$	$\tilde{\Gamma}^{(c)}_{pca}$	*ls*	*pca*
200	0.8	0.04	0.16	0.38	4.0	9.3	0.041	0.041
	0.4	0.05	0.28	1.37	7.0	33.6	0.040	0.041
	0.2	0.04	0.60	4.01	15.7	99.9	0.038	0.040
500	0.8	0.41	0.44	3.18	3.6	7.5	0.124	0.423
	0.4	0.42	0.58	10.06	5.0	23.5	0.115	0.429
	0.2	0.41	0.80	27.54	7.4	64.5	0.108	0.427
1000	0.8	3.43	1.05	24.24	3.0	6.9	0.351	3.514
	0.4	3.03	1.16	62.22	4.0	19.9	0.287	3.130
	0.2	2.96	1.39	149.12	5.5	48.8	0.252	3.056

*The values are calculated based on 100 simulation replicates.

setting. Moreover, we will illustrate the performance of our proposed multiplier bootstrap inferential method for testing the high-dimensional coefficient matrix and the rank estimation methods.

In the following sections, we show partial simulation results due to the space limit. For the complete numerical results, we refer to Section H of the supplementary materials.

6.1. Performance of The Estimators

To evaluate the performance, we repeat the simulation under each setting 500 times and, for any estimator $\tilde{\theta}$ for a parameter θ_0, we calculate the average mean-square-error: $\text{MSE}(\tilde{\theta}) = \frac{1}{500|\theta_0|}\sum_{s=1}^{500}\left\|\tilde{\theta}_s - \theta_{0,s}\right\|_F^2$, where $\tilde{\theta}_s$, $\theta_{0,s}$ are the estimator and the true parameter in sth repetition, and $|\theta_0|$ is the number of elements in θ_0.

We first compare the performance of our iterative LS estimator with that of the iterative PCA estimator using DGP 1. Table 1 shows the MSE of different estimators obtained with the true rank based on the 500 simulation replicates in each setting of DGP 1 for $g = 3$ and $g = c$ (at convergence), and $r = 3$. Results for other cases are similar, and are provided in the supplementary materials.

For larger sample sizes $n, m = 500, 1000$, we see that $\tilde{\Gamma}^{(3)}_{ls}$ has much smaller MSE than the initial estimator $\hat{\Gamma}$, and it has the same MSE as $\tilde{\Gamma}^{(c)}_{ls}$ at all values of π. It indicates that our LS estimate of Γ at a finite step performs better than the initiate estimate, and it has a similar performance as the LS estimate at convergence. Moreover, $\tilde{\Gamma}^{(3)}_{ls}$ and $\tilde{\Gamma}^{(c)}_{pca}$ have similar MSE values, both of which are significantly smaller than the MSE obtained from $\tilde{\Gamma}^{(3)}_{pca}$. The difference between the MSE values of $\tilde{\Gamma}^{(3)}_{ls}$ and $\tilde{\Gamma}^{(3)}_{pca}$ becomes more dramatic as the observation rate π is smaller. This result corroborates our theoretical finding that the proposed iterative LS estimator at a finite step $g \geq 1$ achieves the same convergence rate and asymptotic property as the iterative PCA estimator at $g \to \infty$. For small sample size $n, m = 200$, we can observe the same pattern for $\tilde{\Gamma}^{(3)}_{ls}$ at $\pi = 0.5, 0.8$. The MSE of $\tilde{\Gamma}^{(3)}_{ls}$ is almost the same as that of $\tilde{\Gamma}^{(c)}_{ls}$ and $\tilde{\Gamma}^{(c)}_{pca}$ at $\pi = 0.5, 0.8$, but it is slightly worse at $\pi = 0.2$. However, the MSE of $\tilde{\Gamma}^{(3)}_{pca}$ is much larger than that of the other three estimates. This result further shows that the iterative PCA method needs a diverging number of iterations to achieve the desired convergence rate as proven in Jin, Miao, and Su (2021). The performance of the estimators of β is similar for both methods. Only in the case $n = m = 200$ and $\pi = 0.2$, $\tilde{\beta}^{(3)}_{pca}$ is slightly worse than $\tilde{\beta}^{(3)}_{ls}$.

Next, we compare the computing time of the iterative LS and the iterative PCA methods. When missing values exist, our proposed iterative LS method has a great computational advantage over the iterative PCA method in two aspects. First, for one complete iteration, the computational complexity of PCA on the updated matrix W is $O(mn^2 + m^3)$, and it is only $O(r^2\pi mn)$ for solving the two LS systems defined in (7) for L and F. Since we have the low-rank assumption, r is fixed and $r \ll \min(m, n)$, we see that our LS method is much more computationally efficient than the PCA method for one complete update. Second, our estimator only needs a finite number of iterations, while the iterative PCA estimator requires a diverging number of iterations to have the same asymptotic properties. This result was already demonstrated by the performance comparison in Table 1.

To test the actual computing time of the two estimators, we run simulations using the data generated from DGP 1 when the true rank $r = 3$, and the sample sizes, $n = m = 200, 500, 1000$, with observation rate $\pi = 0.8, 0.5, 0.2$, respectively. Based on 100 simulation replications of each setting, Table 2 reports the average computing time and the number of iterations needed to obtain the converged estimate for each method, and the average computing time of one iteration (one complete update). For a fair comparison, all simulations are run on a regular laptop with specs: Intel(R) Core(TM) i7-8750H CPU, 2667MHz 16 GB RAM without the help of GPU or CPU parallel computing.

The last two columns in Table 2 show the average time for one update by both methods at different sample sizes n, m. We see that the iterative PCA method has a more dramatic increase (from 0.04 for $n, m = 200$ to 3 sec for $n, m = 1000$) than our proposed iterative LS method (from 0.04 to 0.3 instead) when the sample size increases. We can also see that the number of iterations needed to converge increases as the observation rate π decreases. From the "Number of iterations" columns, we observe that the iterative PCA method in general needs more iterations to converge, and the difference between the PCA and the LS methods becomes more prominent as the π value becomes smaller even in the settings with large sample sizes. For instance, in the case with $n, m = 1000$, the iterative LS method needs around 3 iterations at $\pi = 0.8$ and 5 iterations at $\pi = 0.2$,

Table 3. The MSE of different estimators in DGP 2.

DGP 2		Initial		Iterative LS				
n,m	C	$\hat{\beta}$	$\hat{\Gamma}$	$\tilde{\beta}_{ls}^{(3)}$	$\tilde{\beta}_{ls}^{(c)}$	$\tilde{\Gamma}_{ls}^{(3)}$	$\tilde{\Gamma}_{ls}^{(c)}$	$N_{ls}^{(c)}$[2]
200	1	0.420	5.993	0.183	0.178	0.516	0.456	12.7
	1.5	0.318	3.612	0.155	0.154	0.379	0.375	9.7
	2	0.268	2.315	0.142	0.141	0.337	0.334	8.4
500	1	0.192	3.271	0.067	0.066	0.185	0.182	8.9
	1.5	0.143	1.472	0.057	0.057	0.147	0.147	7.2
	2	0.119	0.888	0.053	0.053	0.132	0.132	6.3
1000	1	0.115	1.535	0.034	0.034	0.094	0.094	7.3
	1.5	0.085	0.721	0.030	0.030	0.077	0.077	6.1
	2	0.070	0.482	0.028	0.028	0.071	0.071	5.6

The average number of complete iterations to get converged results.

Table 4. Average bias and 95% CI coverage rate for some estimators*.

DGP 2		Bias (10^{-2})						95% CI coverage rate					
n,m	C	$\tilde{\Gamma}_{11}$	$\tilde{\Gamma}_{23}$	$\tilde{\Gamma}_{35}$	$\tilde{Y}_{11}$	$\tilde{Y}_{23}$	$\tilde{Y}_{35}$	$\tilde{\Gamma}_{11}$	$\tilde{\Gamma}_{23}$	$\tilde{\Gamma}_{35}$	$\tilde{Y}_{11}$	$\tilde{Y}_{23}$	$\tilde{Y}_{35}$
200	1	0.3	−1.9	−1.9	−0.5	−2.6	−1.7	0.93	0.90	0.92	0.92	0.92	0.92
	1.5	−1.0	−0.1	−1.1	−0.5	−0.4	0.5	0.92	0.92	0.94	0.95	0.95	0.96
	2	−0.1	−0.0	−2.0	−0.0	0.2	−0.4	0.92	0.92	0.94	0.93	0.95	0.95
500	1	0.7	−0.4	−0.4	0.3	−0.3	2.0	0.92	0.93	0.93	0.94	0.95	0.93
	1.5	0.5	−0.4	0.6	0.2	−0.4	1.9	0.94	0.94	0.93	0.95	0.96	0.95
	2	0.1	−0.8	0.5	−0.1	−0.4	1.4	0.95	0.95	0.93	0.95	0.96	0.95
1000	1	0.3	−2.1	1.3	−0.9	−1.0	−1.1	0.94	0.92	0.94	0.94	0.94	0.96
	1.5	0.7	−2.4	1.8	−0.3	−0.7	0.0	0.95	0.92	0.96	0.93	0.96	0.95
	2	0.9	−1.9	1.3	0.0	−0.4	0.0	0.96	0.93	0.95	0.93	0.95	0.95

*$\tilde{\Gamma}_{i,j} = \tilde{\Gamma}_{i,j}^{(3)}$ and $\tilde{Y}_{i,j} = X_i'\tilde{\beta}_j^{(3)} + \tilde{\Gamma}_{i,j}^{(3)}$.

whereas the number of iterations for the iterative PCA method grows from 7 to 49.

Next, we show in Table 3 the MSE of our iterative LS estimators based on the 500 simulation replicates in each setting of DGP 2 for $g = 3$ and $r = 3$. We can observe similar patterns as shown in DGP 1; the estimators at $g = 3$ have almost the same MSE as the converged estimators in every case when $n, m = 500$ or 1000. Even for $C = 1$, $n, m = 200$, the estimator at $g = 3$ performs quite well. When the C value is larger, the response matrix is more densely observed, so the estimators are expected to have better performance. The last column shows the average number of iterations to obtain the converged estimator, and we can see that the numbers are all small. With the low computational complexity, it is possible to use the converged solution in practice, or use the estimate at $g = 3$ if the algorithm is implemented on large datasets and we need a faster computational speed.

In the last of this section, we construct pointwise confidence intervals for $\Gamma_{i,j}$ and $\mu_{i,j} = E(Y_{i,j} \mid L_i, X_i, F_j)$ for some given i, j based on the asymptotic representations in Theorem 1 and Theorem 2. For $g \geq 1$, let $\tilde{Y}_{i,j} = X_i'\tilde{\beta}_j^{(g)} + \tilde{\Gamma}_{i,j}^{(g)}$ and $\sigma^2 = E(\varepsilon_{i,j}^2)$, then $(\tilde{Y}_{i,j} - \mu_{i,j})/\sigma_{n,m}(\tilde{Y}_{i,j})$ and $(\tilde{\Gamma}_{i,j}^{(g)} - \Gamma_{i,j})/\sigma_{n,m}(\tilde{\Gamma}_{i,j}^{(g)})$ asymptotically follow $N(0,1)$, where

$$\begin{aligned}\sigma_{n,m}^2(\tilde{\Gamma}_{i,j}^{(g)}) &= \sigma^2\left[n^{-1}L_i'E(\pi_i L_i L_i')^{-1}L_i + (m\pi_i)^{-1}F_j'\Sigma_F^{-1}F_j\right] \\ &\quad + n^{-1}\zeta_{i,j}^2 \\ \sigma_{n,m}^2(\tilde{Y}_{i,j}) &= \sigma^2\left[n^{-1}\left(L_i'E(\pi_i L_i L_i')^{-1}L_i + X_i'E(\pi_i X_i X_i')^{-1}X_i\right)\right. \\ &\quad \left.+(m\pi_i)^{-1}F_j'\Sigma_F^{-1}F_j\right], \end{aligned} \tag{15}$$

and $\zeta_{i,j}^2 = X_i'E(\pi_i X_i X_i')^{-1}E(\pi_k^2 X_k L_k' F_j F_j' L_k X_k' \mid F_j)E(\pi_i X_i X_i')^{-1}X_i$. The unknown terms in the above representations can be replaced by empirical estimators so that we can get the estimated standard error $\hat{\sigma}_{n,m}(\cdot)$.

In the literature, the latent factor model is often considered for matrix completion without considering the covariates. In this model, the matrix Θ is directly decomposed as $\Theta = L^*F^{*\prime}$, where L^* and F^* are latent factors and their loadings, both of which are unknown and need to be estimated. Compared to model (1), we can write $L^* = [X, L]$ and $F^* = [\beta, F]$, but both X and β are treated as unknown variables in the latent factor model. When $E(X_i L_i') = 0$ and $E(\beta_j F_j') = 0$, it can be shown that the asymptotic variance of the estimator $\tilde{Y}_{i,j} := \tilde{\Theta}_{i,j}$ based on the latent factor model with Θ having rank $r + d$ is $\sigma_{n,m}^2(\tilde{Y}_{i,j}) = \sigma^2\left[n^{-1}\left(L_i'E(\pi_i L_i L_i')^{-1}L_i + X_i'E(\pi_i X_i X_i')^{-1}X_i\right) + (m\pi_i)^{-1}(F_j'\Sigma_F^{-1}F_j + \beta_j'\Sigma_\beta^{-1}\beta_j)\right]$, where $\Sigma_\beta = E\left(\beta_j\beta_j'\right)$, and $\tilde{\Theta}$ is the iterative LS estimator of Θ and has the rank of $d + r$. We can see that this asymptotic variance has one additional term $(m\pi_i)^{-1}\beta_j'\Sigma_\beta^{-1}\beta_j$ compared to the one given in (15), so it is larger than the asymptotic variance of the estimator for model (1) that incorporates the observed covariates. When the estimator $\tilde{\Theta}$ has a rank smaller than $d + r$, it has an asymptotically nonnegligible bias.

Table 4 shows the biases and the empirical coverage rates of 95% CI of three arbitrarily chosen estimators obtained at $g = 3$. Each value is calculated based on 500 simulation replicates. We observe that all the biases are very small, and the coverage rates are close to the nominal value except for the cases with a very small effective size ($n = m = 200$, $C = 1$). The empirical distributions of the Z-statistics of $\tilde{Y}_{i,j}$ for $(i,j) = (2,3)$ are shown in Figure 1. We can see that the distributions are close to the standard normal (shaded area) in those cases with larger effective sample sizes. Results for settings in DGP 1 are similar and provided in the supplementary materials.

6.2. Rank Estimation

The number of factors r is often unknown in practice. Section 4 introduces the estimator $\hat{r}^{\text{eIC}}$ to estimate the unknown rank r. With large n, m, Theorem 3 shows that this estimator can find the correct rank with a high probability if the penalty $h(n,m)$ satisfies the stated condition. We let

$$h(n,m) = C_h n^{\delta_h}\sqrt{(m+n)/(mn\hat{\alpha}_n)}, \tag{16}$$

with $C_h = 0.9$ and $\delta_h = 0.1$, where $\hat{\alpha}_n = \overline{\pi} = n^{-1}\sum_{i=1}^n \hat{\pi}_i$ for DGP1 and $\hat{\alpha}_n = e^{\hat{\gamma}_{0,n}}$ for DGP2, so that $h(n,m)$ satisfies the condition given in Theorem 3. To see the performance of the eIC method for rank estimation, we use the 500 simulation replicates in each setting of DGP 1 and DGP 2, and estimate the rank using the eIC criterion given in (11). We set $g = 3$ for the eIC method because our iterative LS estimates perform well with three iterations as shown in Section 6.1.

We report the accuracy (the percentage of obtaining the true rank) along with the average of the rank estimates based on 500 simulation replicates for all settings in Table 5. Note that cases with $\pi = 1$ in DGP 1 are omitted since we aim to find a method

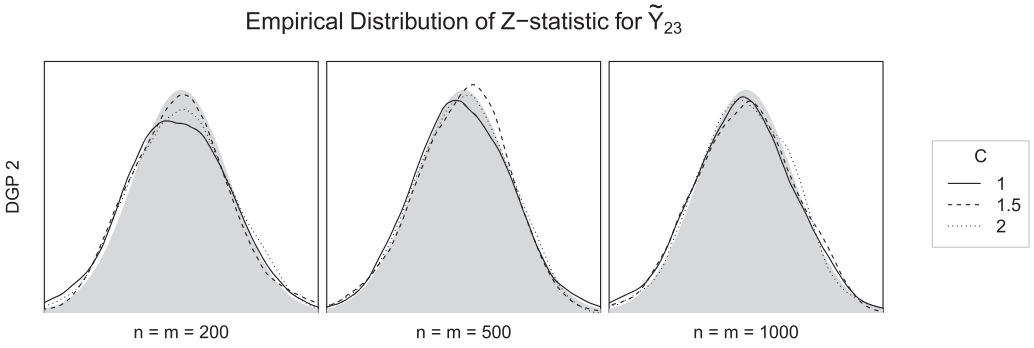

Figure 1. The empirical distribution of $\frac{\tilde{Y}_{23}-\mu_{23}}{\hat{\sigma}_{n,m}(\tilde{Y}_{23})}$ in different simulation settings. The shaded area is the density of standard normal distribution.

Table 5. The rank estimation results of $\hat{r}^{\text{eIC}(3)}$ based on 500 simulations in each setting with the true rank $r = 3$.

	DGP 1			DGP 2		
n, m	π	Acc.*	Ave.[2]	C	Acc.*	Ave.[2]
200	0.2	93.8	2.95	1.0	97.8	3.02
	0.5	100.0	3.00	1.5	99.8	3.00
	0.8	100.0	3.00	2.0	99.8	3.00
500	0.2	100.0	3.00	1.0	100.0	3.00
	0.5	100.0	3.00	1.5	100.0	3.00
	0.8	100.0	3.00	2.0	100.0	3.00
1000	0.2	100.0	3.00	1.0	100.0	3.00
	0.5	100.0	3.00	1.5	100.0	3.00
	0.8	100.0	3.00	2.0	100.0	3.00

The percentage of $\hat{r}^{\text{eIC}(3)} = r$.
The average of $\hat{r}^{\text{eIC}(3)}$.

that can accurately estimate the rank when the data have missing entries. We see that our proposed eIC method performs well in all settings for both DGP 1 and DGP 2. Even with a relatively small sample size ($n = m = 200$) and low observation rate ($\pi = 0.2$ or $C = 1.0$), the eIC method can correctly estimate the true rank with high probability. Its performance further improves as the sample size becomes larger.

6.3. Simultaneous Inference for The Coefficients

In this section, we conduct hypothesis tests on $H_0 : A_j \cdot \beta_j = a_j^0$ for $j \in \mathcal{G}$ at the significant level $\alpha = 0.05$ by the multiplier bootstrap method given in Section 5. We consider the following hypotheses: (i) $H_0 : \beta_{j,p} = 0 \quad \forall\, j, p$; (ii) $H_0 : \beta_{j,p_0} = 0 \quad \forall\, j$. Note that in (ii), p_0 is a fixed value (could be 1, 2 or 3 in our DGPs).

Remark. To follow the notation in Section 5, $A_j = I$ in (i), and $A_j = (1,0,0), (0,1,0), (0,0,1)$ in (ii) for $p_0 = 1, 2, 3$ respectively, and $\mathcal{G} = \{1, \ldots, m\}$ in all the tests.

To see the performance of the testing procedure under null and different alternative hypotheses, we generate our β from $N(0, 4\rho^2 I)$ in DGPs 1 and 2. We run 500 simulation replications in each setting with $\rho = 0$, e^{-3}, $e^{-2.5}$, e^{-2}, $e^{-1.5}$, and e^{-1}, respectively. Note that the null hypothesis H_0 is true when $\rho = 0$. For each setting and ρ value, we compute the empirical rejection rate of each test based on the 500 simulation replicates.

The results for DGP 2, $r = 3$ are presented in Figure 2. The numerical results of all scenarios are relegated to the supplementary materials. We observe that except for the case with a small sample size $n = m = 200$ and $\pi = 0.2$, the rejection rate is very close to the significant level 0.05 under the null hypothesis ($\rho = 0$). The power approaches 1 quickly as the ρ value becomes larger or the sample size increases. This corroborates our theoretical results for the proposed simultaneous testing method given in Section 5. The rejection rate for the case with $n = m = 200$ and $\pi = 0.2$ is slightly larger due to the small effective sample size.

7. Application

In this section, we apply the proposed method to the MovieLens 1M dataset.[1] MovieLens is a website where people can sign up and rate movies in their database, and it is run by a lab at the University of Minnesota called GroupLens. They provide movie recommendations to the users based on their rating history. The 1M dataset contains 1,000,209 ratings on 3952 movies from 6040 users. Some demographic information of users is provided using an assigned ID, including the user's gender, age, occupation, and zip code. Each rating is a number between 0.5 and 5 with 0.5 gaps between two ratings, linked to a user and a movie. The timestamp at which a rating was given was also recorded. In the dataset, each user has rated at least 20 movies. To provide appropriate recommendations to users, our goal is to : (i) estimate the ratings based on the proposed low-rank model with covariates given in Model (2) through our iterative LS procedure, (ii) conduct pointwise inference for each rating based on the established asymptotic distribution, and (iii) conduct simultaneous inference for the coefficients of the covariates based on our bootstrap procedure.

7.1. Application of The Proposed Method to MovieLens 1M

To apply our method to the MovieLens 1M dataset, we use Y to represent the rating matrix in which the ith row and jth column correspond to the ith user and jth movie, respectively. As a result, the dimension of Y should be 6040×3952. We consider gender and age as the covariates in Model (2); both of them may have effects on the movie ratings and the missingness of the ratings. Then, in the covariate matrix, the gender is encoded as "0" (female) and "1" (male); the age is factorized into four

[1] website: *https://grouplens.org/datasets/movielens/*

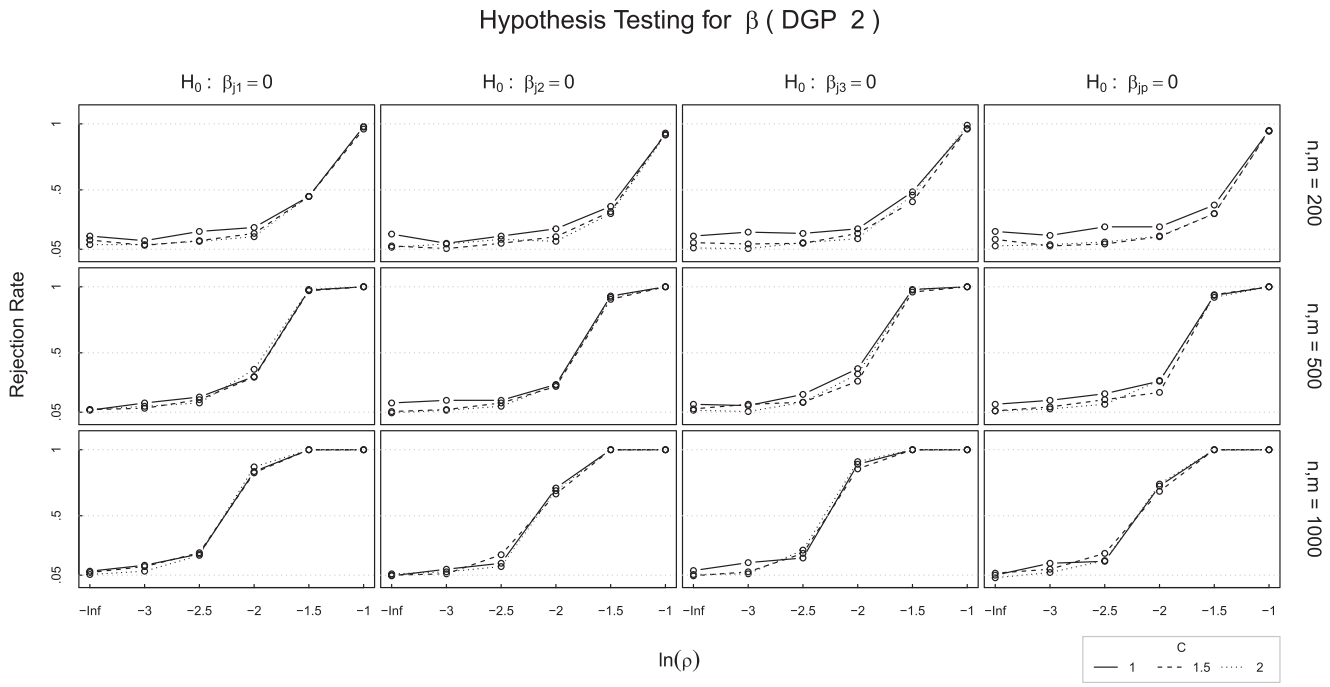

Figure 2. Empirical rejection rates at level $\alpha = 0.05$. Each column represents a hypothesis, and each row represents a sample size. When $x = \ln(\rho) = -\text{Inf}$, the null hypothesis is true.

groups: "0-24", "25-34", "35-49", "50+", and it is represented by three dummy variables. We also include the interactive terms between gender and age groups in the covariate matrix X, and then the dimension of X including the intercept is $m = 6040$ by $d = 8$. The dataset is split into a training set and a test set, and the test set contains 60, 400 ratings with 10 ratings from each user.

Let G_i be the indicator of the gender for the ith user, and $A_{i,k}$, $1 \le k \le 3$, be the indicator of age groups "25-34", "35-49", "50+", respectively. Both gender and age may affect the missingness of movie ratings, so we fit a logistic model for $\pi_i = P(\xi_{i,j} = 1|X_i)$:

$$\begin{aligned}\text{logit}(\pi_i) = \gamma_0 + G_i\gamma_1 + A_{i,1}\gamma_2 + A_{i,2}\gamma_3 + A_{i,3}\gamma_4 \\ +(G_i \cdot A_{i,1})\gamma_5 + (G_i \cdot A_{i,2})\gamma_6 + (G_i \cdot A_{i,3})\gamma_7,\end{aligned} \tag{17}$$

where $\text{logit}(x) = \log(x/(1-x))$. Moreover, we fit the following model for the responses:

$$\begin{aligned}Y_{i,j} = \beta_{j,0} + G_i\beta_{j,1} + A_{i,1}\beta_{j,2} + A_{i,2}\beta_{j,3} + A_{i,3}\beta_{j,4} + (G_i \cdot A_{i,1}) \\ \beta_{j,5} + (G_i \cdot A_{i,2})\beta_{j,6} + (G_i \cdot A_{i,3})\beta_{j,7} + L_i'F_j + \varepsilon_{i,j}.\end{aligned} \tag{18}$$

We also consider the sub-model with only the main effects of $G_\cdot$ and $A_{\cdot,k}$ as well as other sub-models which include only partial interactions between gender and age groups to see which model has the best prediction. We use the eIC method with the penalty given in (16) with $C_h = 0.2$, $\delta_h = 0.1$ to obtain the estimated rank $\hat{r} = \hat{r}^{\text{eIC}(c)}$. Tables 2 and 3 in Section 6.1 show that the iterative LS algorithm in general only needs a few iterations to converge. Then, in the real data analysis, we estimate β_j and $\Gamma_{i,j}$ by running the iterative LS algorithm until convergence or stopped at step=30. With the estimated rank, we then obtain the estimated ratings in the training set and the predicted ratings in the test set by $\tilde{Y}_{i,j} = X_i'\tilde{\beta}_j + \tilde{\Gamma}_{i,j}$ where $\tilde{\beta} = \tilde{\beta}^{(c)}$ and $\tilde{\Gamma}_{i,j} = \tilde{L}_i^{(c)'}\tilde{F}_j^{(c)}$. In addition, since the rating is limited to be between 0.5 and 5, we define the adjusted estimated rating as $\tilde{Y}_{i,j}^{\text{adj}} = (\tilde{Y}_{i,j} \vee 0.5) \wedge 5$, so as to enable the estimated ratings to have values between (0.5, 5).

Table 6. The fitting result for the logistic model for π_i.

	(Intercept)	G	A_1	A_2	A_3	$G \cdot A_1$	$G \cdot A_2$	$G \cdot A_3$
Estimate	−3.347	0.148	0.140	−0.037	−0.304	0.061	0.083	0.072
Std. error (10^{-3})	4.53	5.25	5.71	5.97	7.87	6.58	6.93	9.07
p-value	*	*	*	10^{-9}	*	*	*	10^{-14}

*Value $< 10^{-15}$.

7.2. The Fitting Results

The fitting result of the logistic model for π_i given in (17) is shown in Table 6. It shows the estimate and the standard error of each coefficient, and the p-values for testing whether each coefficient is zero or not. Table 6 shows that the p-values are all close to zero, indicating that all the coefficients for both the main and interaction effects are significantly different from zero. This result further demonstrates that the two baseline covariates, gender and age, and their interactions should have significant effects on the missing pattern of the movie rates. We will use the full model (17) for π_i in the follow-up analysis.

Next, we calculate the root mean square error (RMSE) of the estimated ratings, where $\text{RMSE} = \left[(n_S)^{-1}\sum_{(i,j)\in S}(Y_{i,j} - \tilde{Y}_{i,j})^2\right]^{1/2}$, respectively, for the training and test datasets, to check the prediction performance of each model. In the above formula, the S is the set of observed indices in the training set or the indices in the test set, and $n_S = |S|$ is the number of elements in S. The RMSEs for the training and

Table 7. RMSE of different models in training and test set.

Model			RMSE		adj. RMSE[2]	
Covariate(s)	p	$\hat{r}$	Training	Tested	Training	Tested
$1+G+A_1+A_2+A_3+G\cdot A_1$ $+G\cdot A_2+G\cdot A_3$	8	2	0.8381	0.9067	0.8379	0.8993
$1+G+A_1+A_2+A_3+G\cdot A_1+G\cdot A_2$	7	2	0.8395	0.9046	0.8393	0.8982
$1+G+A_1+A_2+A_3+G\cdot A_1+G\cdot A_3$	7	2	0.8398	0.9057	0.8396	0.8980
$1+G+A_1+A_2+A_3+G\cdot A_2+G\cdot A_3$	7	2	0.8398	0.9064	0.8396	0.8982
$1+G+A_1+A_2+A_3+G\cdot A_1$	6	2	0.8411	0.9032	0.8409	0.8962
$1+G+A_1+A_2+A_3+G\cdot A_2$	6	2	0.8411	0.9043	0.8409	0.8966
$1+G+A_1+A_2+A_3+G\cdot A_3$	6	2	0.8415	0.9044	0.8413	0.8968
$1+G+A_1+A_2+A_3$	5	2	0.8428	0.9020	0.8426	0.8949
1	1	2	0.8582	0.9813	0.8581	0.9010

adj. RMSE uses $\tilde{Y}_{i,j}^{adj}$ in stead of $\tilde{Y}_{i,j}$ in the RMSE formula.

Table 8. The hypothesis testing results for all contrasts.

H_0	Meaning	Test statistic	p-value*
$\beta_{j1}=0 \quad \forall j$	No difference in gender	5.25	<0.001
$\beta_{j2}=0 \quad \forall j$	No difference in age group (–24) and (25–34)	5.50	<0.001
$\beta_{j3}=0 \quad \forall j$	No difference in age group (–24) and (35–49)	10.75	<0.001
$\beta_{j4}=0 \quad \forall j$	No difference in age group (–24) and (50+)	6.15	<0.001
$\beta_{j2}-\beta_{j3}=0$	$\forall j$No difference in age group (25–34) and (35–49)	10.75	<0.001
$\beta_{j2}-\beta_{j4}=0$	$\forall j$ No difference in age group (25–34) and (50+)	7.81	<0.001
$\beta_{j3}-\beta_{j4}=0$	$\forall j$ No difference in age group (35–49) and (50+)	10.75	<0.001

*Each p-value is calculated through 1000 bootstrap results.

test datasets are provided in Table 7. Note that in this table, we also calculate the RMSE using the adjusted rating $\tilde{Y}_{i,j}^{adj}$. According to Table 7, we can see that all the estimated rank $\hat{r}$ by the eIC method is 2 for all cases, and the best prediction which gives the lowest RMSE in the test set is the model with only the main effects, no matter we use the original estimators or the adjusted ones. As a result, we will use the model with only the main effects in the follow-up analysis, and it is formulated as

$$Y_{i,j} = \beta_{j,0}+G_i\beta_{j,1}+A_{i,1}\beta_{j,2}+A_{i,2}\beta_{j,3}+A_{i,3}\beta_{j,4}+L_i'F_j+\varepsilon_{i,j}. \quad (19)$$

7.3. Insight into MovieLens 1M

With the selected model (19), we can run contrast tests on the coefficient matrix β to see if any category in the covariates is unnecessary or if any two (or more) of them can be combined. Since the tests concern the high dimensional coefficient matrix β, we use the multiplier bootstrap method provided in Section 5 to conduct simultaneous inference, and the results are presented in Table 8. All the p-values in Table 8 are very small, indicating that the different age and gender groups have significant effects on the prediction of movie ratings. For further illustration, Figure 3 shows the boxplots of the estimated movie ratings in different gender and age groups for some movies. We can see that for the movie "Antonia's Line," the ratings are very different between genders, but they are similar among different age groups of the same gender. However, for some other movies such as "The Brain That Wouldn't Die," in which both age and gender significantly affect the movie ratings; see more boxplot examples in the supplementary materials.

While the overall effects of different covariates on the ratings can be investigated through Table 8, examples in Figure 3 motivate us to perform individual tests on each movie, so that we can understand the effect of covariate on each movie. A z-test is then conducted for each movie based on the asymptotic result in Theorem 1. Table S5 in Section H.1 of the supplementary materials shows the top 10 movies with the smallest p-values in each test. All the p-values shown in Table S5 are significant after a Bonferroni adjustment. In Figure 4, we select two movies in which either gender or age has a significant effect and draw a quantile plot with 90% point-wise confidence intervals (CI) to further illustrate the effects. The movie " Set It Off" is the one in which the ratings are significantly different in gender (nonoverlapping CI bands), but not at all in age, and "Boys and Girls" shows the other way. Note that only the most significant pair of age groups are shown in this figure. We refer to the supplementary materials for the numerical results of more movie examples.

8. Conclusion

This article studies statistical inference for noisy matrix completion with auxiliary information when the missing pattern of the responses depends on baseline covariates and the observed rates can go to zero as the sample size increases. We show that the iterative LS method has a computational advantage over the iterative PCA method, and it is supported by reliable statistical properties for inference. With only a finite number of iterations, the resulting estimators of the latent low-rank matrix and the coefficient matrix for the observed covariates are asymptotically unbiased and guaranteed to have asymptotic normality under mild conditions. A new information criterion eIC method based on the iterative LS estimation is proposed for rank estimation. It

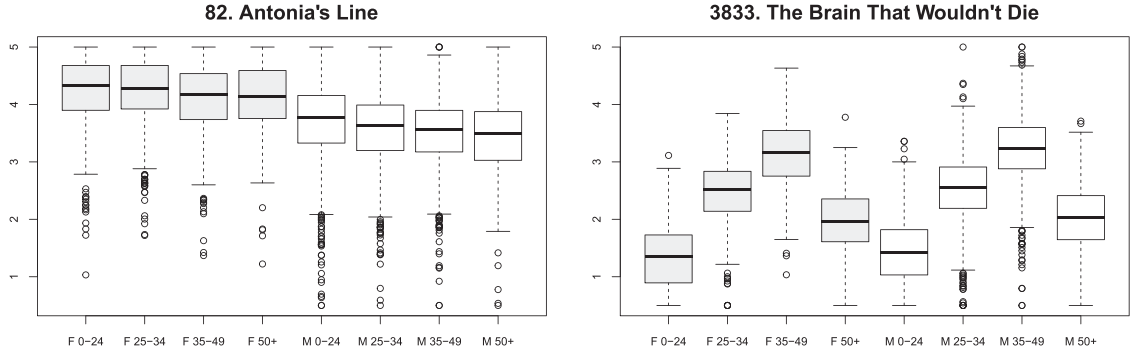

Figure 3. Boxplots of the estimated ratings in different gender and age groups for some movies.

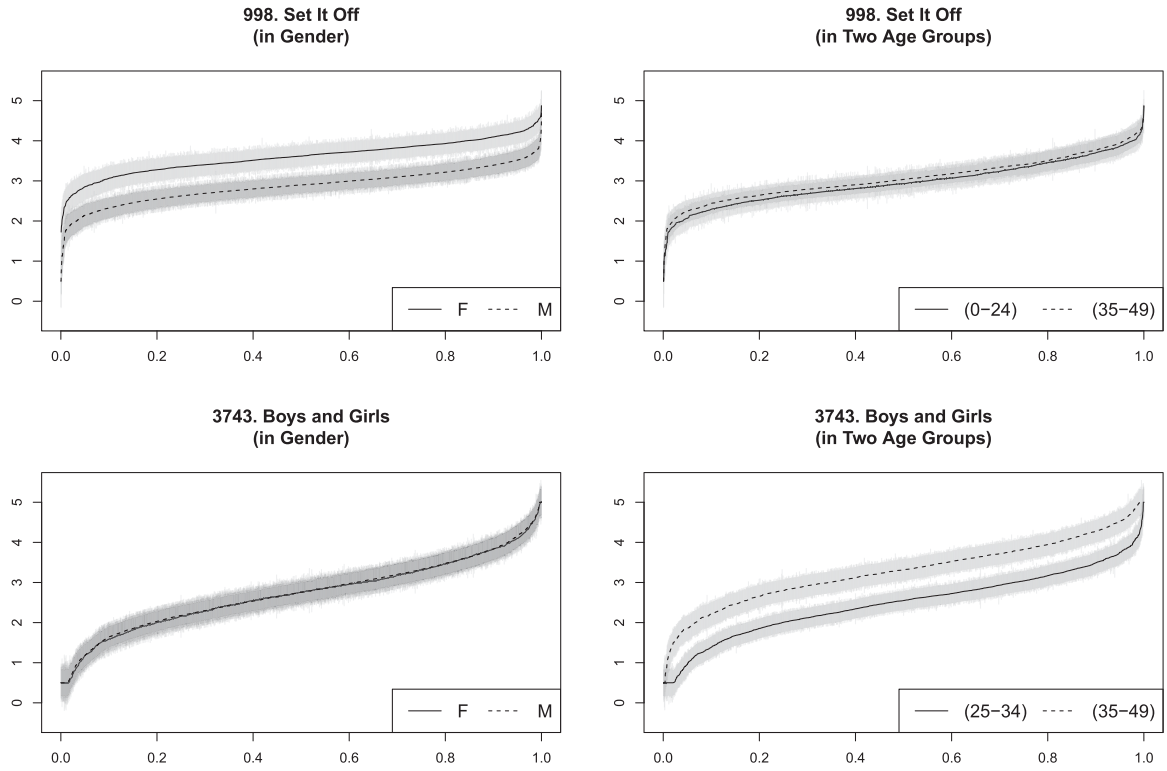

Figure 4. Estimated ratings and 90% point-wise confidence intervals in different groups. The *y*-axis is the rating and the *x*-axis is the percentile. Ratings are grouped by gender or age.

is supported by the consistency property and is demonstrated to have better numerical performance than the widely used IC criterion method based on the singular value estimation.

Moreover, we propose a simultaneous testing method for the high dimensional coefficient matrix β via a Gaussian multiplier bootstrap procedure. This inferential procedure can help us investigate the effects (or contrast effects) of the auxiliary covariates for the prediction of the missing entries. We have discussed in Section 6.1 and have shown in the real data application Section 7 that the use of the observed covariates in matrix completion does help the prediction and improves the prediction accuracy. Our proposed method has immediate applications in collaborative filtering, biological and social network recovery, recommender systems, and so forth. The semi-supervised model considered in our article makes use of row-feature information such as the user's demographic information to help the prediction of movie ratings. It is worth noting that Zhu, Shen, and Ye (2016) have considered a different model that incorporates user-specific and content-specific predictors by letting their coefficients be the same across all j and i, respectively. As an extended work, we can also consider incorporating the column-feature information into our proposed framework. Moreover, the development of the asymptotic distributions of the iterative LS estimators in the setting with the growing number of factors (Mao, Chen, and Wong 2019) or Γ with high rank is also an interesting future research topic to explore.

Supplementary Materials

All the technical proofs and additional numerical results are provided in the online supplementary materials.

Acknowledgments

The authors thank the Editor, the Associate Editor, and anonymous reviewers for their constructive comments that have helped us improve the article substantially.

Disclosure Statement

The authors report there are no competing interests to declare.

Funding

Niu and Ma's research was supported by NSF grant DMS-2014221, DMS-2310288, and UCR Academic Senate CoR grant. Yichong Zhang acknowledges the financial support from the NSFC under grant no. 72133002.

References

Alaya, M. Z., and Klopp, O. (2019), "Collective Matrix Completion," *Journal of Machine Learning*, 20, 1–43. [1]

Athey, S., Bayati, M., Doudchenko, N., Imbens, G., Khosravi, K. (2021), "Matrix Completion Methods for Causal Panel Data Models," *Journal of the American Statistical Association*, 116, 1716–1730. [2]

Bai, J., and Ng, S. (2002), "Determining the Number of Factors in Approximate Factor Models," *Econometrica*, 70, 191–221. [4]

——— (2021), "Matrix Completion, Counterfactuals, and Factor Analysis of Missing Data," *Journal of the American Statistical Association*, 116, 1746–1763. [2]

Cahan, E., Bai, J., and Ng, S. (2022), "Factor-based Imputation of Missing Values and Covariances in Panel Data of Large Dimensions," *Journal of Econometrics*, 233, 113–131. [2]

Cai, T. T., and Zhou, W.-X. (2016), "Matrix Completion via Max-Norm Constrained Optimization," *Electronic Journal of Statistics*, 10, 1493–1525. [5]

Candes, E. J., and Plan, Y. (2010), "Matrix Completion with Noise," *Proceedings of the IEEE*, 98, 925–936. [2]

Carpentier, A., Klopp, O., Löffler, M., and Nickl, R. (2018), "Adaptive Confidence Sets for Matrix Completion," *Bernoulli*, 24, 2429–2460. [2,4]

Chen, Y., Chi, Y., Fan, J., Ma, C., and Yan, Y. (2020), "Noisy Matrix Completion: Understanding Statistical Guarantees for Convex Relaxation via Nonconvex Optimization," *SIAM Journal on Optimization*, 30, 3098–3121. [2]

Chen, Y., Fan, J., Ma, C., and Yan, Y. (2019), "Inference and Uncertainty Quantification for Noisy Matrix Completion," *Proceedings of the National Academy of Sciences*, 116, 22931–22937. [2,4]

Chernozhukov, V., Chetverikov, D., and Kato, K. (2013), "Gaussian Approximations and Multiplier Bootstrap for Maxima of Sums of High-Dimensional Random Vectors," *The Annals of Statistics*, 41, 2786–2819. [2]

——— (2017), "Central Limit Theorems and Bootstrap in High Dimensions," *Annals of Probability*, 45, 2309–2352. [2]

Chiang, K., Hsieh, C., and Dhillon, I. S. (2015), "Matrix Completion with Noisy Side Information," in *Proceedings of the 28th International Conference on Neural Information Processing Systems* (Vol. 2), pp. 3447–3455. [1]

Fithian, W., and Mazumder, R. (2018), "Flexible Low-Rank Statistical Modeling with Side Information," *Statistical Science*, 33, 238–260. [1]

Graham, B. S. (2020), "Sparse Network Asymptotics for Logistic Regression," Technical Report, National Bureau of Economic Research. [4]

Hastie, T., Mazumder, R., Lee, J. D., and Zadeh, R. (2015), "Matrix Completion and Low-Rank SVD via Fast Alternating Least Squares," *Journal of Machine Learning Research*, 16, 3367–3402. [2]

Ibriga, H. S., and Sun, W. W. (2023), "Covariate-Assisted Sparse Tensor Completion," *Journal of the American Statistical Association*, 118, 2605–2619. [1]

Javanmard, A., and Montanari, A. (2014), "Confidence Intervals and Hypothesis Testing for High-Dimensional Regression," *Journal of Machine Learning Research*, 15, 2869–2909. [2]

Jin, H., Ma, Y., and Jiang, F. (2022), "Matrix Completion with Covariate Information and Informative Missingness," *Journal of Machine Learning Research*, 23, 1–62. [1]

Jin, S., Miao, K., and Su, L. (2021), "On Factor Models with Random Missing: Em Estimation, Inference, and Cross Validation," *Journal of Econometrics*, 222, 745–777. [2,4,5,6,7]

Klopp, O. (2014), "Noisy Low-Rank Matrix Completion with General Sampling Distribution," *Bernoulli*, 20, 282–303. [4,5]

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. [2]

Mao, X., Chen, S. X., and Wong, R. K. W. (2019), "Matrix Completion with Covariate Information," *Journal of the American Statistical Association*, 114, 198–210. [1,5,12]

Negahban, S., and Wainwright, M. J. (2012), "Restricted Strong Convexity and Weighted Matrix Completion: Optimal Bounds with Noise," *The Journal of Machine Learning Research*, 13, 1665–1697. [2]

Rosenbaum, P. R., and Rubin, D. B. (1983), "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika*, 79, 41–55. [4]

Sun, R., and Luo, Z. (2016), "Guaranteed Matrix Completion via Nonconvex Factorization," *IEEE Transactions on Information and Theory*, 62, 6535–6579. [2]

Xia, D., and Yuan, M. (2021), "Statistical Inferences of Linear Forms for Noisy Matrix Completion," *Journal of the Royal Statistical Society*, Series B, 83, 58–77. [2]

Xiong, R., and Pelger, M. (2023), "Large Dimensional Latent Factor Modeling with Missing Observations and Applications to Causal Inference," *Journal of Econometrics*, 1, 271–301. [2]

Xu, M., Jin, R., and Zhou, Z. (2013), "Speedup Matrix Completion with Side Information: Application to Multi-Label Learning," in *Proceedings of the 26th International Conference on Neural Information Processing Systems* (Vol. 2), pp. 2301–2309. [1]

Zhou, Y., Wilkinson, D., Schreiber, R., and Pan, R. (2008), *Large-Scale Parallel Collaborative Filtering for the Netflix Prize in Algorithmic Aspects in Information and Management*, Berlin: Springer. [2]

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. [1,12]