Statistica Sinica Preprint No: SS-2020-0243	
Title	Robust Recommendation via Social Network Enhanced Matrix Completion
Manuscript ID	SS-2020-0243
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202020.0243
Complete List of Authors	Jingxuan Wang, Haipeng Shen and Fei Jiang
Corresponding Author	Fei Jiang
E-mail	fei.jiang@ucsf.edu

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0243

ROBUST RECOMMENDATION VIA SOCIAL NETWORK ENHANCED MATRIX COMPLETION

Jingxuan Wang[1], Haipeng Shen[2] and Fei Jiang[1]

[1] *University of California, San Francisco and* [2] *University of Hong Kong*

Abstract: Robust product recommendation enables internet platforms to boost their business. However, in practice, the user–product rating matrix often has many missing entries. Social network information generates new insights about user behaviors. To fully use such information, we develop a novel approach, called MCNet, that combines the random dot product graph model and low-rank matrix completion to recover missing entries in a user–product rating matrix. Our algorithm improves the accuracy and the efficiency of recovering the incomplete matrices. We study the asymptotic properties of the estimator. Furthermore, we perform extensive simulations and show that MCNet outperforms existing approaches, especially when the data have small signals. Moreover, MCNet yields robust estimation under misspecified models. We apply MCNet and its competitors to predict the missing entries in the user–product rating matrices of the Yelp and Douban movie platforms. The results show that, in general, MCNet gives the smallest testing errors among the comparative methods.

Key words and phrases: Low-rank estimation, matrix completion, missing data, random dot product graph, social network.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0243

1. Introduction

Uncovering true user ratings on products is critical for internet platforms such as Yelp, Facebook, and Amazon, because they help to promote their business. These platforms use estimated ratings to recommend products to users with the highest willingness to pay, thus maximizing their revenue. Such data sets are often arranged in matrix form, where the rows and columns correspond to users and products, respectively. However, typically, many entries are missing, because not every product has been exposed to all users. The ratings of these missing entries are crucial to the recommendation strategies of the products on the platforms.

Many matrix completion algorithms have been developed to recover missing entries in a user–product rating matrix, often adopting low-rank estimation frameworks. Srebro, Rennie, and Jaakkola (2005) developed a matrix factorization algorithm by assuming the underlying rating matrix has a finite known rank and the data are missing completely at random. Recht, Fazel, and Parrilo (2010) cast the matrix completion as a constrained convex optimization problem, and recovered the matrix by minimizing its nuclear norm. Similar convex relaxations are adopted by Candes and Plan (2010) and Koltchinskii, Lounici, and Tsybakov (2011) under different noise settings, and an imputation method, SoftImpute, was developed to accom-

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0243

modate matrices with growing ranks (Mazumder, Hastie, and Tibshirani (2010)). Recently, Negahban and Wainwright (2011), Koltchinskii, Lounici, and Tsybakov (2011), and Fan, Wang, and Zhu (2017) rigorously studied the statistical properties of low-rank matrix estimation under the linear regression setting. Klopp (2014) and Elsener and van de Geer (2018) considered using the Huber loss for the robust estimation of the low-rank matrix. Fan, Gong, and Zhu (2019) further extended the low-rank matrix estimation under the nonlinear model framework. In addition, a nonconvex relaxation method, called TopN, was proposed by Kang, Peng, and Cheng (2016). Furthermore, deep learning methods have been proposed in the recommendation system literature to recover missing matrix entries (Liu and Wu (2017); Zhang et al. (2019)). In particular, Wang, Wang, and Yeung (2015) developed a deep learning framework incorporating content information to address the matrix completion problem.

In addition to the user–product rating matrix, auxiliary covariates, such as users' demographics and products' attributes, are often collected on internet platforms. These covariates provide additional information that is not explained by the ratings matrix itself (Feuerverger, He, and Khatri (2012)). Incorporating this information has been shown to improve both the accuracy and the precision of estimations (Abernethy et al. (2009); Shi,

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0243

Larson, and Hanjalic (2014)). Chiang, Hsieh, and Dhillon (2015) and Xu, Jin, and Zhou (2013) illustrated the theoretical guarantees on matrix recovery with covariate information under noiseless settings. Furthermore, Mao, Chen, and Wong (2019) showed that considering users' features reduces the matrix estimation error under missing-at-random settings, where the probability of observation is independent of the unobserved target matrix, given the covariates (Rubin (1976)). Zhu, Shen, and Ye (2016) proposed a partial latent model that combines user and item covariates in a linear regression setting with l_1 and l_2 penalties. Moreover, by allowing high-dimensional features, Robin et al. (2018) introduced a sparse low-rank estimation to recover the rating matrix and the feature effects simultaneously.

In addition to the explicit features collected on the platforms, the social network contains rich information about the associations among users. The network implicitly tells how much information can be taken from the other users. This information has been translated into penalty terms in matrix completion (Ma et al. (2011); Yu, Pan, and Li (2011); Liu and Aberer (2013)) to reflect the intuition that the closer two users are, the more information must be taken from each other to recover their missing ratings. Rao et al. (2015) developed a graph Laplacian method that uses network information to assist the matrix completion. Jing et al. (2019) introduced a

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0243

penalized collaborative filtering method, called NetRec, that allows users to share information with their connections in the network. In addition, Dai et al. (2019) developed a smooth recommendation system based on the latent factor model that jointly incorporates social network, product network, and user–product-specific covariates using kernel weighting.

When the covariate information is not given, the missingness of the matrix entries is considered to be missing completely at random (Srebro, Rennie, and Jaakkola (2005); Candes and Plan (2010); Koltchinskii, Lounici, and Tsybakov (2011)). This assumption is not appropriate in our setting, because users may vary significantly in terms of their willingness for rating. The additional network information generates user-specific features, allowing us to estimate the missing probability tailored to the individual features. To estimate the user-specific missing probability, Mao, Wong, and Chen (2018) introduced a two-step inverse probability weighting–based matrix completion framework, where the observation probabilities are estimated using a generalized linear model with a low-rank predictor matrix. Furthermore, Bi et al. (2017) proposed a singular value decomposition–based group-specific model to use the between-subject dependency information from users and items that share similar missingness characteristics.

In this article, we propose a matrix completion social network (MC-

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0243

Net) algorithm that uses social network information to improve the matrix completion. Specifically, we embed the high-dimensional network structure into a low-dimensional space using adjacency spectral embedding (Sussman et al. (2012); Lyzinski et al. (2014); Athreya et al. (2018)), and generate a set of latent positions that best summarize the distances between the users. We then incorporate this embedding into the matrix completion model to improve the accuracy of recovering missing entries in the rating matrix. The theoretical guarantee of the adjacency spectral embedding has been studied by Oliveira (2009), Lu and Peng (2013), and Lei and Rinaldo (2015). Sussman, Tang, and Priebe (2013) and Lyzinski et al. (2016) formally established the estimation consistency, while Athreya et al. (2016) and Tang and Priebe (2018) provided the asymptotic normality of the resulting latent positions.

The contributions of this study are five-fold. (1) We develop a flexible model incorporating network information to improve the accuracy of matrix recovery. (2) We provide an efficient estimation procedure that yields smaller estimation errors than those of the competing methods proposed by Mazumder, Hastie, and Tibshirani (2010), Kang, Peng, and Cheng (2016), and Jing et al. (2019). (3) Our method is flexible in terms of considering different missing mechanisms, including covariate-independent and covariate-

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0243

dependent missingness. (4) We provide asymptotic upper bounds of the estimation errors. (5) We use simulation studies and two real-data analyses to show that our algorithm improves both the accuracy and the efficiency of recovering incomplete matrices.

The rest of the paper is organized as follows. In Section 2, we describe the MCNet model and the estimation procedure. In Section 3, we provide the theoretical results. We evaluate the MCNet method and compare it with existing methods using extensive simulations in Section 4. We apply MCNet to analyze data from a Douban Movie data set and the Yelp Dataset Challenge in Section 5. Section 6 concludes the paper.

2. Methodology

2.1 Notation

Before presenting the model, we define the notation used throughout the paper. For an $n_1 \times n_2$ matrix $\mathbf{H}$, we denote H_{ik} as the i,kth entry, $\mathbf{H}_{i\cdot}$ as the ith row, and $\mathbf{H}_{\cdot k}$ as the kth column. We define $\lambda_i(\mathbf{H})$ as the ith largest eigenvalue of $\mathbf{H}$, and $\sigma_i(\mathbf{H}) = \sqrt{\lambda_i(\mathbf{H}^{\mathrm{T}}\mathbf{H})}$ as the ith largest singular value of $\mathbf{H}$. In addition, $\|\mathbf{H}\|_2$, $\|\mathbf{H}\|_{\max}$, $\|\mathbf{H}\|_F$, and $\|\mathbf{H}\|_*$ denote the operator norm, element-wise maximum norm, Frobenius norm, and nuclear norm, respectively, of the matrix $\mathbf{H}$. Furthermore, we denote $\|\mathbf{H}\|_{2\to\infty}$ as the

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0243

maximum of the Euclidean norms for the rows of $\mathbf{H}$; that is, $\|\mathbf{H}\|_{2\to\infty} = \max_i\{\|\mathbf{H}_{i\cdot}^{\mathrm{T}}\|_2\}$. Then, $\langle \mathbf{H}_1, \mathbf{H}_2\rangle$ denotes the trace inner product between $\mathbf{H}_1$ and $\mathbf{H}_2$. Finally, we define the signal in the matrix $\mathbf{H}$ as $\sum_{ij}\{H_{ij} - \sum_{ij} H_{ij}/(n_1n_2)\}^2/(n_1n_2-1)$.

2.2 The MCNet Model

Let M_{ik} be the connection indicator such that $M_{ik} = M_{ki} = 1$ if the ith and the kth users are connected in the social network. Furthermore, we define $\mathbf{M}$ as the adjacency matrix, with M_{ik} being its i,kth entry, and assume

$$\Pr(\mathbf{M}|\mathbf{X}) = \prod_{i<k}(\mathbf{X}_{i\cdot}\mathbf{X}_{k\cdot}^{\mathrm{T}})^{M_{ik}}(1-\mathbf{X}_{i\cdot}\mathbf{X}_{k\cdot}^{\mathrm{T}})^{1-M_{ik}}, \tag{2.1}$$

where $\mathbf{X}_{i\cdot} \in \mathbb{R}^{1\times d}$ and $\mathbf{X} = (\mathbf{X}_{1\cdot}^{\mathrm{T}}, \ldots, \mathbf{X}_{n_1\cdot}^{\mathrm{T}})^{\mathrm{T}} \in \mathbb{R}^{n_1\times d}$ is the latent position matrix. We say that $\mathbf{M}$ follows a random dot product graph (RDPG) distribution with the latent position $\mathbf{X}$, denoted by $\mathbf{M} \sim \mathrm{RDPG}(\mathbf{X})$.

In addition, let Y_{ik} be the rating for product k by user i, and we assume

$$Y_{ik} = A_{0ik} + \epsilon_{ik}, \tag{2.2}$$

where $\{\epsilon_{ik}, i = 1, \ldots, n_1, k = 1, \ldots, n_2\}$ are independent mean zero random errors. We assume that $\mathbf{Y}$ and $\mathbf{M}$ are independent when $\mathbf{X}$ is given. Let $\mathbf{A}_0$ be the matrix with the i,kth entry being A_{0ik}, and we further decompose

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0243

the information in $\mathbf{A}_0$ as

$$\mathbf{A}_0 = \mathbf{X}\boldsymbol{\beta}_0 + \mathbf{B}_0. \tag{2.3}$$

Here, $\boldsymbol{\beta}_0 \in \mathbb{R}^{d\times n_2}$ is the unknown parameter matrix of interest, and $\mathbf{B}_0 \in \mathbb{R}^{n_1\times n_2}$ is an unknown low-rank matrix with columns that are orthogonal to the column space of $\mathbf{X}$ so that $\mathbf{P}_{\mathbf{X}}\mathbf{B}_0 = \mathbf{0}$, where $\mathbf{P}_{\mathbf{X}} = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}$. The low-rank assumption is commonly used in the matrix factorization literature (Srebro, Rennie, and Jaakkola (2005); Recht, Fazel, and Parrilo (2010); Candes and Plan (2010); Koltchinskii, Lounici, and Tsybakov (2011)), and assumes there are a few latent factors explaining most of the data. This assumption allows information to be borrowed across all observed entries in $\mathbf{Y}$. Furthermore, we assume $\mathbf{B}_0$ is orthogonal to the column space of $\mathbf{X}$ to ensure the model is identifiable. To see that, suppose $\mathbf{P}_{\mathbf{X}}\mathbf{B} \neq \mathbf{0}$. It is easy to see that $\mathbf{X}\boldsymbol{\beta}_0 + \mathbf{B}_0 = \mathbf{X}\{\boldsymbol{\beta}_0 + (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{B}_0\} + \mathbf{P}_{\mathbf{X}}^{\perp}\mathbf{B}_0$, where $\mathbf{P}_{\mathbf{X}}^{\perp} = \mathbf{I} - \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}$. Now, $\{\boldsymbol{\beta}_0 + (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{B}_0\}$ is an unknown parameter, because $\boldsymbol{\beta}_0$ and $\mathbf{B}_0$ are both unknown. Furthermore, $\mathbf{P}_{\mathbf{X}}^{\perp}\mathbf{B}_0$ is orthogonal to $\mathbf{X}$. Therefore, for any $\mathbf{B}_0$, $\mathbf{X}\boldsymbol{\beta}_0 + \mathbf{B}_0$ can be decomposed into a linear combination of $\mathbf{X}$ and a matrix orthogonal to the column space of $\mathbf{X}$. This assumption is also used in Mao, Chen, and Wong (2019).

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0243

2.3 MCNet Estimation

When $\mathbf{Y}$ and $\mathbf{X}$ are fully observed, $\boldsymbol{\beta}_0$ and $\mathbf{B}_0$ are the minimizers of

$$E\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{B}\|_F^2.$$

Because each user only rates a subset of the products, $\mathbf{Y}$ contains missing entries. Let $W_{ik} = 1$ if the product k has a rating from user i, and θ_{ik} be the true probability of the i,kth entry being observed. Let $\mathbf{W}, \boldsymbol{\Omega}_0$ be matrices with the i,kth entry being W_{ik} and θ_{ik}^{-1}, respectively. Then, we consider the risk function

$$L(\boldsymbol{\beta}, \mathbf{B}) = E\|\mathbf{Y} \circ \mathbf{W} \circ \boldsymbol{\Omega}_0 - \mathbf{X}\boldsymbol{\beta} - \mathbf{B}\|_F^2,$$

where $\circ$ is the Hadamard product. Under Conditions (C1) and (C3) in the next section, $\boldsymbol{\beta}_0$ and $\mathbf{B}_0$ are the minimizers of $L(\boldsymbol{\beta}, \mathbf{B})$ because $E(\mathbf{Y} \circ \mathbf{W} \circ \boldsymbol{\Omega}_0|\mathbf{X}) = \mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_0 + \mathbf{B}_0$ and the partial derivative of $E\|\mathbf{Y} \circ \mathbf{W} \circ \boldsymbol{\Omega}_0 - \mathbf{X}\boldsymbol{\beta} - \mathbf{B}\|_F^2$ with respect to $\boldsymbol{\beta}$ and $\mathbf{B}$ are zeros when $\boldsymbol{\beta} = \boldsymbol{\beta}_0$ and $\mathbf{B} = \mathbf{B}_0$. In addition, because $\mathbf{X}$ and $\mathbf{B}_0$ are orthogonal, it is easy to see that

$$\begin{aligned} L(\boldsymbol{\beta}_0, \mathbf{B}_0) &= E\|\mathbf{X}\boldsymbol{\beta}_0 - \mathbf{P}_{\mathbf{X}}(\mathbf{W} \circ \boldsymbol{\Omega}_0 \circ \mathbf{Y})\|_F^2 \\ &\quad + E\|\mathbf{B}_0 - \mathbf{P}_{\mathbf{X}}^{\perp}(\mathbf{W} \circ \boldsymbol{\Omega}_0 \circ \mathbf{Y})\|_F^2. \end{aligned}$$

In the above equation, $\boldsymbol{\beta}_0$ and $\mathbf{B}_0$ are in two separate loss functions, which allows us to estimate $\boldsymbol{\beta}_0$ and $\mathbf{B}_0$ separately.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0243

However, in general, $\mathbf{X}$ is unobservable. We estimate it using the relation in (2.1), as follows. Because $\mathbf{M}$ is a square matrix, we can write $\mathbf{M} = \sum_{i=1}^{n_1} \lambda_i(\mathbf{M})\mathbf{u}_i\mathbf{u}_i^{\mathrm{T}}$, where $\lambda_i(\mathbf{M})$ are the eigenvalues ordered by absolute magnitude, and $\mathbf{u}_1, \ldots, \mathbf{u}_{n_1}$ are the corresponding eigenvectors. Let $\mathbf{U_M} = (\mathbf{u}_1, \ldots, \mathbf{u}_d)$ and $\mathbf{S_M} = \mathrm{diag}\{|\lambda_1(\mathbf{M})|, \ldots, |\lambda_d(\mathbf{M})|\}$, where the hyperparameter d can be prescribed by the eigenvalue ratio test (Ahn and Horenstein (2013)). That is,

$$d = \underset{q}{\mathrm{argmax}} \frac{\log(V_{q-1}/V_q)}{\log(V_q/V_{q+1})},$$

where V_q is the sum of the first q largest absolute eigenvalues of $\mathbf{M}$. We then obtain the estimator $\widehat{\mathbf{X}} = \mathbf{U_M}\mathbf{S_M}^{1/2}$ using an adjacency spectral embedding that gives a continuous representation of the nodes in a social network as vectors in a low-dimensional space.

We replace $\mathbf{X}$ in $L(\boldsymbol{\beta}, \mathbf{B})$ by the adjacency spectral embedding $\widehat{\mathbf{X}}$, and obtain the estimators as

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^{d\times n_2}}{\mathrm{argmin}} \left\{ \frac{1}{n_1 n_2} \|\widehat{\mathbf{X}}\boldsymbol{\beta} - \mathbf{P}_{\widehat{\mathbf{X}}}(\mathbf{W} \circ \widehat{\boldsymbol{\Omega}} \circ \mathbf{Y})\|_F^2 + \lambda_1 \|\boldsymbol{\beta}\|_F^2 \right\} \tag{2.4}$$

and

$$\begin{aligned} \widehat{\mathbf{B}} \;=\; & \underset{\mathbf{B}\in\mathcal{N}(\widehat{\mathbf{X}})}{\mathrm{argmin}} \left\{ \frac{1}{n_1 n_2} \|\mathbf{B} - \mathbf{P}^{\perp}_{\widehat{\mathbf{X}}}(\mathbf{W} \circ \widehat{\boldsymbol{\Omega}} \circ \mathbf{Y})\|_F^2 \right. \\ & \left. + \lambda_2\{\alpha\|\mathbf{B}\|_* + (1-\alpha)\|\mathbf{B}\|_F^2\} \right\}, \end{aligned} \tag{2.5}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0243

where $\mathcal{N}(\widehat{\mathbf{X}})$ is the space orthogonal to the one spanned by $\widehat{\mathbf{X}}$, and $\widehat{\mathbf{\Omega}}$ is an element-wise consistent estimator for $\mathbf{\Omega}_0$. In (2.4), the term $\|\boldsymbol{\beta}\|_F^2$ leads to a ridge regression problem. In (2.5), the nuclear term $\|\mathbf{B}\|_*$ has a thresholding effect over singular values because it is equivalent to the l_1-norm penalty of the vector of singular values. The Frobenius term $\|\mathbf{B}\|_F^2 = \text{trace}(\mathbf{B}^{\mathrm{T}}\mathbf{B}) = \sum_i^{n_1 \wedge n_2} \sigma_i^2$ corresponds to the l_2-norm penalty of the singular values. The convex regularization leads to an elastic-net regularization of the singular values (Zou and Hastie (2005); Sun and Zhang (2012)). The parameter α controls the balance of the two penalties. The strong convex Frobenius penalty is introduced to improve the finite-sample performance (Sun and Zhang (2012); Mao, Chen, and Wong (2019)) and stability under highly corrupted data (Li, Chen, and Li (2012); Kim, Lee, and Oh (2015)).

The estimation in (2.4) leads to

$$\widehat{\boldsymbol{\beta}} = (\widehat{\mathbf{X}}^{\mathrm{T}}\widehat{\mathbf{X}} + n_1 n_2 \lambda_1 \mathbf{I})^{-1}\widehat{\mathbf{X}}^{\mathrm{T}}(\mathbf{W} \circ \widehat{\mathbf{\Omega}} \circ \mathbf{Y}). \tag{2.6}$$

From (2.5), $\widehat{\mathbf{B}}$ can be obtained using the one-step first-order scaled singular-value-thresholding operator based on the algorithm introduced in Cai, Candès, and Shen (2010). We show that $\widehat{\mathbf{B}}$ is guaranteed to be orthogonal to the column space of $\widehat{\mathbf{X}}$ in Section S1.2 of the Supplementary Material. Now, because $\widehat{\mathbf{X}}$ converges to $\mathbf{X}$ in the Frobenius norm asymptotically (shown in the next section), $\widehat{\mathbf{B}}$ is orthogonal to the column space of $\mathbf{X}$ asymptotically.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0243

Finally, we estimate $\mathbf{A}_0$ by $\widehat{\mathbf{A}} = \widehat{\mathbf{X}}\widehat{\boldsymbol{\beta}} + \widehat{\mathbf{B}}$. We illustrate the estimation procedure in Algorithm 1.

Algorithm 1 MCNet algorithm.

1. Estimate $\mathbf{X}$ by adjacency spectral embedding $\widehat{\mathbf{X}} = \mathbf{U}_\mathbf{M}\mathbf{S}_\mathbf{M}^{1/2}$
2. Estimate $\boldsymbol{\beta}_0$ by $\widehat{\boldsymbol{\beta}} = (\widehat{\mathbf{X}}^\mathrm{T}\widehat{\mathbf{X}} + n_1 n_2 \lambda_1 \mathbf{I})^{-1}\widehat{\mathbf{X}}^\mathrm{T}(\mathbf{W} \circ \widehat{\boldsymbol{\Omega}} \circ \mathbf{Y})$
3. Estimate $\mathbf{B}_0$ by $\widehat{\mathbf{B}} = \mathbf{U}\mathbf{D}_{n_1,n_2,\lambda_2,\alpha}\mathbf{V}^\mathrm{T}$, where $\mathbf{U}\mathbf{D}\mathbf{V}^\mathrm{T}$ is the singular value decomposition of $\mathbf{P}^{\perp}_{\widehat{\mathbf{X}}}(\mathbf{W} \circ \widehat{\boldsymbol{\Omega}} \circ \mathbf{Y})$, $\mathbf{D}_{n_1,n_2,\lambda_2,\alpha} = \mathrm{diag}\left\{\frac{(D_{1,1}-n_1 n_2 \lambda_2 \alpha/2)_+}{1+n_1 n_2 \lambda_2(1-\alpha)}, \ldots, \frac{(D_{n_1\wedge n_2, n_1 \wedge n_2} - n_1 n_2 \lambda_2 \alpha/2)_+}{1+n_1 n_2 \lambda_2(1-\alpha)}\right\}$, and $t_+ = t \vee 0$
4. Return $\widehat{\mathbf{A}} = \widehat{\mathbf{X}}\widehat{\boldsymbol{\beta}} + \widehat{\mathbf{B}}$

3. Main Results

We define $d(\mathbf{H}_1, \mathbf{H}_2) = \|\mathbf{H}_1 - \mathbf{H}_2\|_F/\sqrt{d_1 d_2}$ for $d_1 \times d_2$ matrices $\mathbf{H}_1, \mathbf{H}_2$. Because $\mathbf{X}\mathbf{X}^\mathrm{T} = \mathbf{X}\mathbf{O}\mathbf{O}^\mathrm{T}\mathbf{X}^\mathrm{T}$ for an orthogonal matrix $\mathbf{O}$, $\mathbf{X}$ and, in turn, $\boldsymbol{\beta}$ are only identifiable up to an orthogonal transformation. We show the convergences of $\widehat{\boldsymbol{\beta}}$ to $\mathbf{O}^\mathrm{T}\boldsymbol{\beta}_0$, $\widehat{\mathbf{B}}$ to $\mathbf{B}_0$, and $\widehat{\mathbf{A}}$ to $\mathbf{A}_0$ by establishing the upper bounds of $\|\widehat{\boldsymbol{\beta}}_{\cdot j} - \mathbf{O}^\mathrm{T}\boldsymbol{\beta}_{0,\cdot j}\|_2$, $d(\widehat{\mathbf{B}}, \mathbf{B}_0)$, and $d(\widehat{\mathbf{A}}, \mathbf{A}_0)$, respectively. First, we present the technical conditions needed for our analysis.

(C1) For all i and j, ϵ_{ij} is an independent random error with zero mean. There exist positive constants c_σ and η such that $\max_{i,j} E|\epsilon_{ij}|^l \leq l! c_\sigma^2 \eta^{l-2}/2$ holds, for any integer $l \geq 2$.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0243

(C2) There exist positive constants a_1, a_2, and a_3 such that

$$\begin{aligned} \|\mathbf{A}_0\|_{\max} &\leq \sqrt{\log(n_1+n_2)}a_1 \\ \|\mathbf{A}_0\|_{2\to\infty} &\leq a_2\sqrt{n_2} \\ \|\mathbf{A}_0^{\mathrm{T}}\|_{2\to\infty} &\leq a_2\sqrt{n_1} \\ \|\boldsymbol{\beta}_0\|_{\max} &\leq a_3. \end{aligned}$$

(C3) For all i and j, the observation indicator W_{ij} independently follows the Bernoulli distribution with parameter $\theta_{ij} := \Pr(W_{ij} = 1|\mathbf{X}_{i,.}, Y_{ij}) = \Pr(W_{ij} = 1|\mathbf{X}_{i,.}) \in (0, 1)$. In addition, W_{ij} is independent of ϵ_{ij}.

(C4) (a) There exists a lower bound $\theta_L \in (0, 1)$ such that $\theta_{ij} \geq \theta_L$, for all i, j; θ_L can vanish when $n_1, n_2 \to \infty$. (b) For all i and j, $|\widehat{\theta}_{ij} - \theta_{ij}| = O_p(n_1^{-1/2})$ and $\widehat{\theta}_{ij}$ is independent of $\{\epsilon_{ij}\}$.

(C5) (a) The true latent position matrix $\mathbf{X}$ has rank d. (b) $n_1^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{X} \to \mathbf{S}_x$ as $n_1 \to \infty$, $\|\mathbf{S}_x\|_2 < \infty$, and $\lambda_d(\mathbf{S}_x) \geq c_0$, where c_0 is a positive constant. (c) $\max_{i\leq n_1} \sum_{j=1}^{n_1} \Pi_{ij} > log^{4+a}(n_1)$, for some positive constant a, where $\mathbf{\Pi} = \mathbf{X}\mathbf{X}^{\mathrm{T}}$.

Conditions (C1)–(C4) are standard conditions discussed in Mao, Chen, and Wong (2019) to achieve estimation consistency when incorporating covariate information in matrix completion. Condition (C3) assumes that

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0243

the missingness follows either covariate-independent or covariate-dependent missingness mechanisms. Under covariate-independent settings (Koltchinskii, Lounici, and Tsybakov (2011); Recht (2011)), we estimate the observed probability by $\sum_{ij} W_{ij}/(n_1 n_2)$. Under the covariate-dependent missingness, if θ_L in Condition (C4) is bounded from below, we can use the standard logistic regression

$$\text{logit}(\theta_{ij}) = \tau_{0j} + \mathbf{X}_{i\cdot}\boldsymbol{\tau}_j \tag{3.1}$$

to generate root-n consistent estimators for the observed probabilities. If θ_L vanishes as n_1 and n_2 go to infinity, we can use a zero-inflated binomial regression to generate root-n consistent estimators (Diallo, Diop, and Dupuy (2017); Hall (2000)). Condition (C5) is a global convex condition that guarantees the consistency of recovering the latent position in the random dot product model.

Lemma 3.1. *Assume Condition (C5) holds. Then, there is an orthogonal matrix* $\mathbf{O} \in \mathbb{R}^{d\times d}$ *such that* $\|\widehat{\mathbf{X}}\mathbf{O}^{\mathrm{T}} - \mathbf{X}\|_F = O_p(n_1^{1/4})$.

In Lemma 3.1, we employ the Hoeffding Concentration Theorem and assume the probability of $M_{ij} = 1$ is strictly greater than zero for all i, j. The lemma is based on Lemma S1.3 in the Supplementary Material, which is equivalent to Lemma 50 in Athreya et al. (2018) and Theorem A.5 in Tang

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0243

et al. (2017) under Condition (C5) that the eigenvalues of the probability matrix $\mathbf{\Pi}$ grow linearly with n_1. This bound allows us to achieve the root-n convergence of $\widehat{\boldsymbol{\beta}}$ to the truth.

Theorem 3.1. *Assume that Conditions (C1)–(C5) hold and* $\lambda_1 = o(n_2^{-1})$. *Then, there exist an orthogonal matrix* $\mathbf{O}$ *and a constant* c *such that* $\|\widehat{\boldsymbol{\beta}}_{\cdot j} - \mathbf{O}^{\mathrm{T}}\boldsymbol{\beta}_{0,\cdot j}\|_2 = O_p[\max\{n_1^{-1/2} log^c(n_1)\theta_L^{-1}, n_1^{-1}\delta^{1/2}(\mathbf{\Pi})\}]$, *for each* $j = 1, \ldots, n_2$.

Theorem 3.1 gives the parametric convergence rate for $\widehat{\boldsymbol{\beta}}_{\cdot j}$ to the truth up to an orthogonal transformation. The fact that $\widehat{\mathbf{X}}$ instead of the true $\mathbf{X}$ was used in the loss function does not inflate the asymptotic errors. Furthermore, it is easy to see that $d(\widehat{\boldsymbol{\beta}}, \mathbf{O}^{\mathrm{T}}\boldsymbol{\beta}_0) = O_p[\max\{n_1^{-1/2}\log^c(n_1)\theta_L^{-1}, n_1^{-1}\delta^{1/2}(\mathbf{\Pi})\}]$ because $\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0$ has a finite rank. The proof is provided in the Supplementary Material.

Lemma 3.2. *Assume Conditions (C1)–(C5) hold and* $\lambda_1 = o(n_2^{-1})$. *Then,*

$$d^2(\widehat{\mathbf{X}}\widehat{\boldsymbol{\beta}}, \mathbf{X}\boldsymbol{\beta}_0) = O_p\left\{\max\left(\frac{log^{2c}(n_1)\delta(\mathbf{\Pi})}{n_1^2\theta_L^2}, \frac{\delta^2(\mathbf{\Pi})}{n_1^3}, \frac{\|\boldsymbol{\beta}_0\|_F^2}{n_1^{1/2}n_2}\right)\right\}.$$

Lemma 3.2 is a direct consequence of Lemma 3.1 and Theorem 3.1, and is necessary to establish the convergence of $\widehat{\mathbf{A}}$.

Lemma 3.3. *Assume Conditions (C1)–(C5) hold,* $\alpha \in (0, 1]$, *and* $\lambda_2 \geq \{2c_Y log^c(n_1)\theta_L^{-1} + a_2\}/\sqrt{n_1 n_2 \alpha^2} + 2\|\mathbf{B}_0\|_2/(n_1 n_2 \alpha)$ *for any positive constant*

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0243

c_Y. *Then,*

$$d^2(\widehat{\mathbf{B}}, \mathbf{B}_0) = O_p\left[\max\left\{\lambda_2\alpha\|\mathbf{B}_0\|_*, \lambda_2(1-\alpha)\|\mathbf{B}_0\|_F^2\right\}\right].$$

Lemma 3.3, together with Lemma 3.2, leads to the following.

Theorem 3.2. *Assume Conditions (C1)–(C5) hold and* $\lambda_1 = o(n_2^{-1})$. *Furthermore, for any given* $c_Y > 0$, *we assume* $\lambda_2 \geq \{2c_Y log^c(n_1)\theta_L^{-1} + a_2\}/\sqrt{n_1 n_2 \alpha^2} + 2\|\mathbf{B}_0\|_2/(n_1 n_2 \alpha)$. *Then,*

$$d^2(\widehat{\mathbf{A}}, \mathbf{A}_0) = O_p\left[\max\left\{\frac{log^{2c}(n_1)\delta(\mathbf{\Pi})}{n_1^2\theta_L^2}, \frac{\delta^2(\mathbf{\Pi})}{n_1^3}, \frac{\|\boldsymbol{\beta}_0\|_F^2}{n_1^{1/2}n_2}, \lambda_2\alpha\|\mathbf{B}_0\|_*, \lambda_2(1-\alpha)\|\mathbf{B}_0\|_F^2\right\}\right].$$

Remark 1. Compared with the state-of-the-art SoftImpute algorithm (Mazumder, Hastie, and Tibshirani, 2010), $d^2(\widehat{\mathbf{A}}, \mathbf{A}_0)$ of MCNet grows with $\|\boldsymbol{\beta}_0\|_F$, $\|\mathbf{B}_0\|_*$, and $\|\mathbf{B}_0\|_F$, while that of SoftImpute increases with $\|\mathbf{A}_0\|_*$ and $\|\mathbf{A}_0\|_F$, as shown in Corollary 1 of Koltchinskii, Lounici, and Tsybakov (2011). In general, $\|\mathbf{A}_0\|_* \geq \|\mathbf{B}_0\|_*$ and $\|\mathbf{A}_0\|_F$ is greater than both $\|\boldsymbol{\beta}_0\|_F$ and $\|\mathbf{B}_0\|_F$. Hence, MCNet improves the estimation accuracy by incorporating the social network information.

Compared with the most recent social network collaborative filtering method NetRec Jing et al. (2019), Theorem 3.2 shows that the estimation error of the social network structure, that is, $\widehat{\mathbf{X}}$, does not contribute to $d^2(\widehat{\mathbf{A}}, \mathbf{A}_0)$. On the other hand, for NetRec, its $d^2(\widehat{\mathbf{A}}, \mathbf{A}_0)$ increases with the additional bias induced by the social network penalty, as shown in Theorem

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0243

3.1 of Jing et al. (2019). This implies that, when the signal in $\mathbf{A}_0$ is relatively small compared with this social network bias, NetRec will perform poorly, whereas MCNet will still provide satisfactory results. We demonstrate this point using simulations in Section 4.2.

4. Simulations

We evaluate the convergence of MCNet, and compare it with that of Soft-Impute by Mazumder, Hastie, and Tibshirani (2010), TopN proposed by Kang, Peng, and Cheng (2016), and the NetRec method proposed by Jing et al. (2019).

4.1 Performance as the Signal-to-Noise Ratio Varies

We fix $d = 10$ and choose $n_1 = n_2 = 500$ to 2000 with a step size of 250. Each entry of $\mathbf{X}$ is generated from a beta distribution with parameters $(1.5, 1)$. We then scale each entry by a chosen constant to ensure $\max_{i,j} \mathbf{X}_{i\cdot}\mathbf{X}_{j\cdot}^{\mathrm{T}} = 0.9$. Furthermore, we generate $M_{ij} = M_{ji}$, for $i \neq j$, from a Bernoulli distribution with success rate $\mathbf{X}_{i\cdot}\mathbf{X}_{j\cdot}^{\mathrm{T}}$, and generate each entry in $\boldsymbol{\beta}_0$ from a mean zero normal distribution with variance d/n_2. Moreover, we define $\mathbf{B}_0 = \mathbf{P}_{\mathbf{X}}^{\perp}\mathbf{U}_0\mathbf{V}_0^{\mathrm{T}}d/\sqrt{n_1 n_2}$, where $\mathbf{U}_0 \in \mathbb{R}^{n_1\times 10}$ and $\mathbf{V}_0 \in \mathbb{R}^{n_2\times 10}$ are matrices with standard normal entries. Let $\mathbf{A}_0 = \mathbf{X}\boldsymbol{\beta}_0 + \mathbf{B}_0$ and $\mathbf{Y} = \mathbf{A}_0 + \boldsymbol{\epsilon}$,

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0243

where $\boldsymbol{\epsilon}$ is an error matrix with independent mean zero normal entries. The standard deviations of the errors are chosen to achieve 1, 0.5, 0.2, 0.1, and 0.05 signal-to-noise ratios.

We adopt two types of missing mechanisms. In Model I, we consider covariate-independent missingness with $\theta_{ij} = 0.2$ uniformly across all i, j. In Model II, the missing probability follows the logistic model in (3.1), where the regression parameters τ_{0j} and $\boldsymbol{\tau}_j$ are selected to achieve 20% missing rates in $\mathbf{Y}$, on average.

We select the tuning parameters λ_1, λ_2, and α using the error perturbation method introduced in Jing et al. (2019) in the simulations. More specifically, given $\mathbf{A}_0$, we generate $\boldsymbol{\epsilon}$ using a normal distribution and calculate $d(\widehat{\mathbf{A}}, \mathbf{A}_0)$ after the estimation. We repeat the procedure K times and obtain the average of $d(\widehat{\mathbf{A}}, \mathbf{A}_0)$ as the residual mean squared error (RMSE). Then, we select the tuning parameters that yield the smallest RMSE, on average. The same procedure is adopted to select the tuning parameters in the SoftImpute, TopN, and NetRec procedures.

We evaluate the convergence of MCNet, and compare the RMSEs from MCNet, SoftImpute, TopN, and NetRec. Figure 1 and Figure 2 show the results when the entries are missing according to Model I and Model II, respectively. They indicate that the convergence rates of $d(\widehat{\boldsymbol{\beta}}, \mathbf{O}^{\mathrm{T}}\boldsymbol{\beta}_0)$, $d(\widehat{\mathbf{X}}\mathbf{O}^{\mathrm{T}}, \mathbf{X})$,

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0243

and $d(\widehat{\mathbf{B}}, \mathbf{B}_0)$ are consistent with the theoretical rates derived in Theorem 3.1, Lemma 3.1, and Lemma 3.3, which are shown as the dashed curves. Here, the specific form of the orthogonal matrix $\mathbf{O}$ is described in the online Supplementary Material. The results also suggest that MCNet has smaller $d(\widehat{\mathbf{A}}, \mathbf{A}_0)$ compared with those from SoftImpute, TopN, and NetRec. The advantages are more obvious when the signal-to-noise ratio becomes smaller. Moreover, we summarize the square root of the area under the $d^2(\widehat{\mathbf{A}}, \mathbf{A}_0)$ curves under different settings in Table 1, which shows that MCNet outperforms SoftImpute, TopN, and NetRec in all settings with a smaller area under the curve, on average.

4.2 Performance as the Noise Level Varies

We compare MCNet, SoftImpute, TopN, and NetRec under the models with various noise levels σ_{ij}, and provide guidelines for choosing among the four methods in practice. The parameters $d, \boldsymbol{\beta}_0, \mathbf{B}_0$ and $\mathbf{X}$ are assumed to be the same as those in Section 4.1. We choose $n_1 = n_2 = 500$ to 2000 with a step size of 500. We generate the observation matrix using $\mathbf{Y} = \mathbf{A}_0 + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon}$ is an error matrix with entries from an independent zero mean normal distribution, with standard deviations varying from 0.2 to 1.6 with a step size of 0.2. Figure 3 shows that MCNet outperforms SoftImpute, TopN, and

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0243

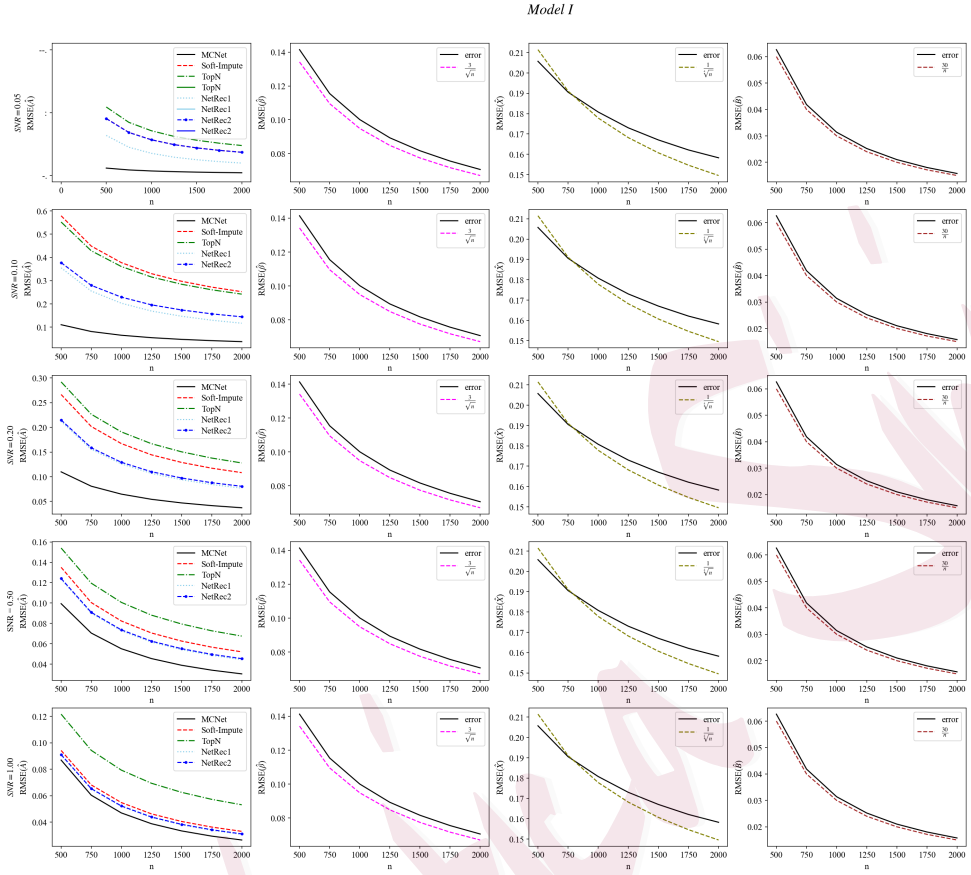

Figure 1: The estimation convergence under model I from 100 simulations. Column 1 shows the comparisons of $d(\widehat{\mathbf{A}}, \mathbf{A}_0)$ between five comparative methods. Columns 2–4 show $d(\widehat{\boldsymbol{\beta}}, \mathbf{O}^{\mathrm{T}}\boldsymbol{\beta}_0)$, $d(\widehat{\mathbf{X}}\mathbf{O}^{\mathrm{T}}, \mathbf{X})$, and $d(\widehat{\mathbf{B}}, \mathbf{B}_0)$, respectively.

NetRec with a smaller $d(\widehat{\mathbf{A}}, \mathbf{A}_0)$. The results are consistent with our finding in Remark 1 of Section 3 that the performance of the NetRec estimator in Jing et al. (2019) deteriorates, while our estimator still exhibits satisfactory

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0243

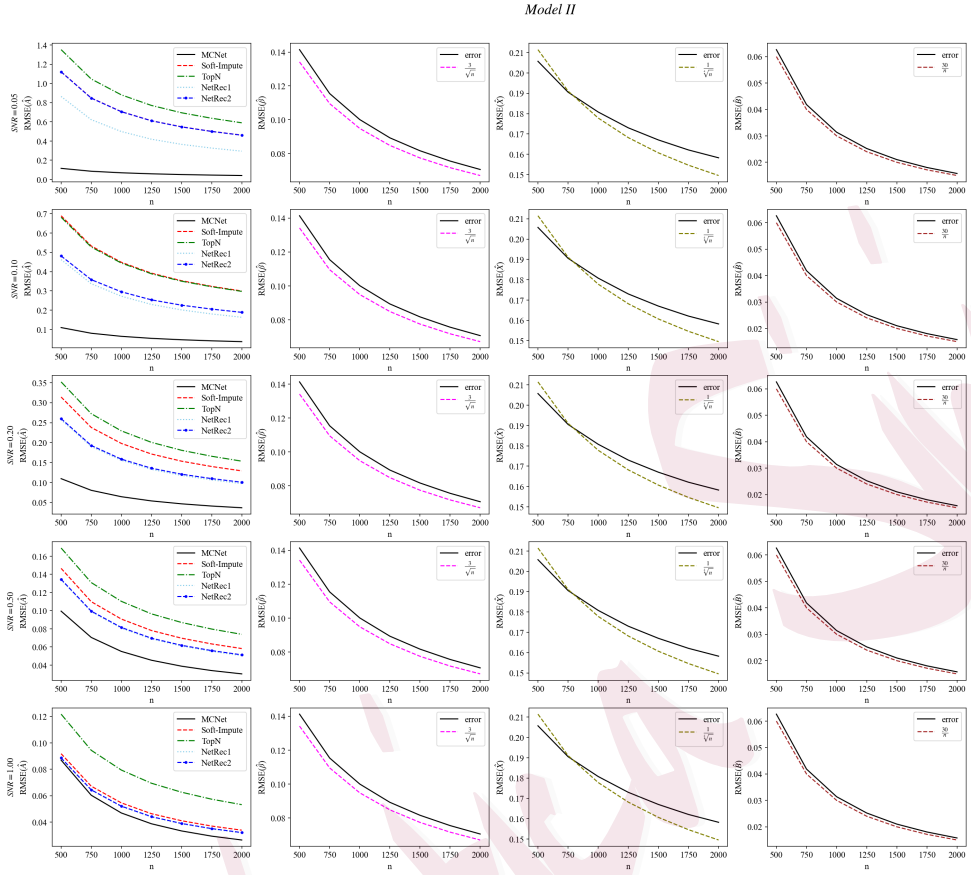

Figure 2: The estimation convergence under model II from 100 simulations. Column 1 shows the comparisons of $d(\widehat{\mathbf{A}}, \mathbf{A}_0)$ between five comparative methods. Columns 2–4 show $d(\widehat{\boldsymbol{\beta}}, \mathbf{O}^{\mathrm{T}}\boldsymbol{\beta}_0)$, $d(\widehat{\mathbf{X}}\mathbf{O}^{\mathrm{T}}, \mathbf{X})$, and $d(\widehat{\mathbf{B}}, \mathbf{B}_0)$, respectively.

results when the signal in $\mathbf{A}_0$ is small.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0243

	Model I					Model II				
	SNR					SNR				
	0.05	0.1	0.2	0.5	1.0	0.05	0.1	0.2	0.5	1.0
MCNet	**2.70**	**2.44**	**2.43**	**2.09**	**1.80**	**2.59**	**2.41**	**2.43**	**2.09**	**1.80**
	(0.04)	(0.03)	(0.02)	(0.02)	(0.01)	(0.03)	(0.02)	(0.02)	(0.02)	(0.01)
SoftImpute	21.36	14.14	6.28	3.10	2.08	26.53	16.79	7.44	3.41	2.06
	(0.39)	(0.17)	(0.10)	(0.04)	(0.03)	(0.44)	(0.20)	(0.11)	(0.05)	(0.03)
TopN	26.63	13.52	7.14	3.77	2.97	33.00	16.64	8.60	4.13	2.98
	(0.32)	(0.16)	(0.08)	(0.04)	(0.04)	(0.38)	(0.19)	(0.10)	(0.05)	(0.04)
NetRec1	13.43	7.68	4.76	2.75	1.97	18.88	10.27	5.87	3.04	1.96
	(0.30)	(0.16)	(0.09)	(0.04)	(0.03)	(0.36)	(0.19)	(0.11)	(0.04)	(0.03)
NetRec2	21.33	8.64	4.88	2.78	1.98	26.50	11.12	5.98	3.07	1.97
	(0.39)	(0.20)	(0.10)	(0.04)	(0.03)	(0.44)	(0.22)	(0.11)	(0.05)	(0.03)

Table 1: Square root of the area under the $d^2(\widehat{\mathbf{A}}, \mathbf{A})$ curves with standard errors (in parentheses). The best results among the algorithms are indicated by the bold font.

4.3 Performance as the Network Density Varies

We compare MCNet, SoftImpute, TopN, and NetRec under the models with various network densities. We fix $d = 10$ and $n_1 = n_2 = 500$. To vary the density of the social network, we generate entries of $\mathbf{X}$ from a beta distribution with parameters $(\alpha, 1)$, where α varies from 0.3 to 2.1 with a step size of 0.3. We then scale the entries of $\mathbf{X}$ by a chosen constant to

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0243

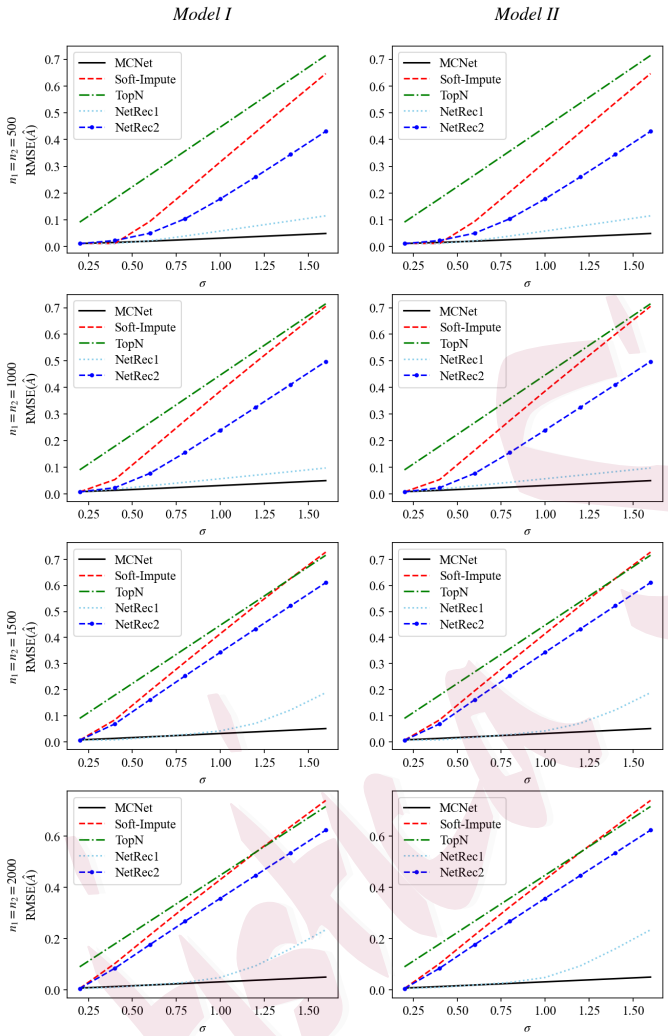

Figure 3: Performance of MCNet method and other methods under models I and II for 100 repetitions.

make sure $\max_{i,j} \mathbf{X}_{i\cdot}\mathbf{X}_{j\cdot}^{\mathrm{T}} = 0.9$. The adjacency matrix is generated from a Bernoulli distribution with success rate $\mathbf{X}\mathbf{X}^{\mathrm{T}}$. The parameters $\boldsymbol{\beta}_0$ and $\mathbf{B}_0$ are assumed to be the same as those in Section 4.1. Let $\mathbf{A}_0 = c(\mathbf{X}\boldsymbol{\beta}_0 + \mathbf{B}_0)$,

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0243

where c is a constant chosen to achieve $\|\mathbf{A}_0\|_F = 1000$, and $\mathbf{Y} = \mathbf{A}_0 + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon}$ is an error matrix with independent mean zero normal entries. The standard deviation of the errors is chosen to achieve a 0.5 signal-to-noise ratio. We only adopt model I for the missing mechanism. That is, each entry of the observation matrix $\mathbf{W}$ is generated from a Bernoulli distribution with mean 0.2.

Figure 4 shows the resulting $d(\widehat{\mathbf{A}}, \mathbf{A}_0)$ with 95% confidence intervals versus the mean density of the social network (mean of $\mathbf{\Pi} = \mathbf{X}\mathbf{X}^{\mathrm{T}}$). It indicates that our estimator is better than other comparative methods, with smaller RMSEs, on average. Furthermore, the advantages over other estimators are move obvious when the connective rates are higher.

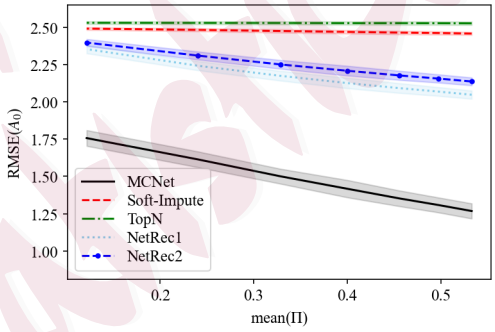

Figure 4: Connection probability. Performance of MCNet and other methods under model I for 100 repetitions.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0243

4.4 Robustness of MCNet

We further evaluate MCNet when model (2.1) is violated. We select $n_1 = 2000$ and $n_2 = 500$, and generate $\mathbf{Y}$ from $\mathbf{Y} = \mathbf{A}_0 + \boldsymbol{\epsilon}$, where $\mathbf{A}_0 = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^{\mathrm{T}}\mathrm{std}(\mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^{\mathrm{T}})^{-1}$, with $\mathbf{U} \in \mathbb{R}^{n_1 \times 10}$, $\mathbf{V} \in \mathbb{R}^{n_2 \times 10}$ being unitary matrices, $\boldsymbol{\Sigma}$ being the diagonal matrix with entries from the beta distribution with shape parameters 1 and 5, and $\mathrm{std}(\mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^{\mathrm{T}})$ being the element-wise standard deviation of the matrix $\mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^{\mathrm{T}}$. The errors are generated from a mean zero normal distribution, with standard deviations varying from 0.2 to 1.6. The entries are missing with probability 0.95. Furthermore, we generate M_{ij} and M_{ji} from a Bernoulli distribution with the success rate as the i, jth entry of $\exp(-\beta\mathbf{T})$, where $\mathbf{T}$ is a matrix with the i, jth entry equal to $\|\mathbf{A}_{0,i\cdot} - \mathbf{A}_{0,j\cdot}\|_2$ and β is selected to allow the mean number of edges on each node of the network to be eight.

Figure 5 shows the resulting $d(\widehat{\mathbf{A}}, \mathbf{A}_0)$ from MCNet and NetRec with their 95% confidence intervals. The results show that when the true model deviates from the one assumed in MCNet, MCNet still provides satisfactory results Furthermore, it outperforms NetRec (with the correct model assumption) with a smaller $d(\widehat{\mathbf{A}}, \mathbf{A}_0)$ when $\sigma_{ij} > 2.5$.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0243

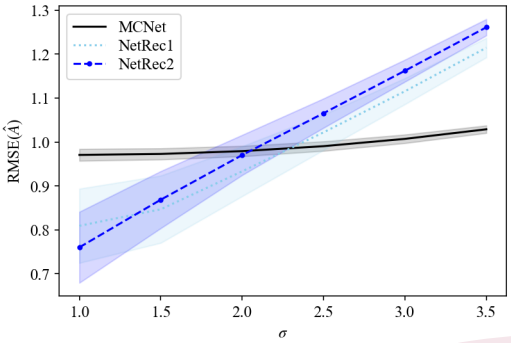

Figure 5: Robustness to assumptions. Performance of MCNet and NetRec methods under model I from 100 simulations.

5. Empirical Study

In this section, we evaluate MCNet on two real data sets: the Douban Movie data set from Douban.com and processed by Zheng et al. (2017), and the Yelp data set from the Yelp Dataset Challenge (Yelp (2019)).

5.1 Douban Movie Data Analysis

The data set contains 195,493 ratings from 3,022 users on 6,971 movies. We select a subset of movies that have at least 10 ratings. Furthermore, two users are considered as connected if they are friends. The i, jth entry of $\mathbf{Y}$ is the rating from user i on the jth movie, scaled by the total number of ratings received by the movie. Finally, the data set comprises 3,022 users, 3,810 movies, and 176,656 ratings. There are 1,366 connections in

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0243

the social network. The sample standard deviation of observed entries of $\mathbf{Y}$ after scaling is 0.08.

We select d using Ahn and Horenstein's method (Ahn and Horenstein (2013)). We assume that the missing mechanism follows model I. We randomly split the users' ratings and the corresponding missing indicators into training (80%) and testing (20%) data sets. Within the training steps, we implement five-fold cross-validation to select the tuning parameters for MCNet, while keeping the social network connections fixed. We report the RMSE for the observed entries on the testing set and compare our method with SoftImpute, TopN, and NetRec. We repeat the training–testing procedure 50 times and report the testing RMSEs in Table 2 and Figure 6. The results show that MCNet has a significantly smaller testing RMSE than those of the competing methods, where its 95% confidence interval does not overlap with the mean RMSEs from the other methods.

	MCNet	SoftImpute	TopN	NetRec1	NetRec2
Mean of RMSEs	**0.01187**	0.01299	0.01258	0.01225	0.01251
95% CI	0.01162-	0.01273-	0.01232-	0.01190-	0.01223-
	0.01211	0.01324	0.01283	0.01260	0.01279

Table 2: The mean out-of-sample RMSEs over 50 repetitions for the Douban data set. The minimum value is indicated by the bold font.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0243

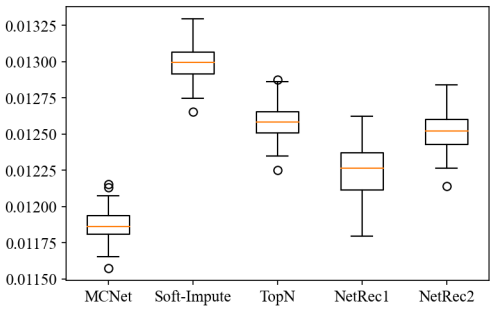

Figure 6: The testing RMSEs of MCNet and other methods over 50 repetitions.

5.2 Yelp Data Analysis

The data set contains six million reviews from one million users on 192 thousand businesses. We extract the users' social network connections and their ratings on restaurants from five cities: Mentor OH, Gastonia NC, Matthews NC, Laval QC (Canada), and Brampton ON (Canada). Two users are connected if they are friends. The i, jth entry of $\mathbf{Y}$ is the rating from user i on the jth restaurant, scaled by the total number of reviews received by the restaurant. We provide some descriptive statistics for the five cities in Table 3.

We use Ahn and Horenstein's method (Ahn and Horenstein (2013)) to select d and apply the same training–testing procedures in Section 5.1 to evaluate the methods. For each city, we repeat the training-testing proce-

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0243

City	No. of users (n_1)	No. of rstrnts. (n_2)	No. of ratings ($\sum_{ij} W_{ij}$)	No. of conns. ($\sum_{ij} M_{ij}$)	Standard deviation
Mentor	2,611	181	4,689	2,334	0.18
Gastonia	3,050	184	5,000	3,422	0.15
Matthews	5,505	207	9,600	13,808	0.13
Laval	1,380	282	2,843	3,064	0.35
Brampton	4,329	546	9,283	12,392	0.25

Table 3: Description of the data from five cities. Abbreviations: rstrnts, restaurants; conns, connections. Standard derivation is the sample standard deviation of observed entries of $\mathbf{Y}$ after scaling.

dure 50 times. The results are reported in Table 4 and Figure 7, which show that all methods perform equivalently because the 95% confidence intervals overlap with each other, while MCNet has a consistently smaller mean RMSE across all cities. We further calculate the sample standard deviation of the entries of $\mathbf{Y}$ in Table 3, which suggests that the observed rating in Yelp has larger variation than that in the Douban data set (sample standard deviation is 0.08). This phenomenon suggests that when the data variation is large, the methods perform equally, possibly because the signal is weak and there is not much room for improvement. When the variation of the observed rating is small, our method performs significantly better

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0243

than the other methods.

	Mentor		Gastonia		Matthews	
	Mean	95% CI	Mean	95% CI	Mean	95% CI
MCNet	**0.0479**	0.0370-0.0588	**0.0334**	0.0256-0.0411	**0.0202**	0.0161-0.0243
SoftImpute	0.0507	0.0397-0.0617	0.0354	0.0273-0.0434	0.0217	0.0176-0.0258
TopN	0.0507	0.0397-0.0617	0.0349	0.0269-0.0429	0.0214	0.0173-0.0255
NetRec1	0.0496	0.0388-0.0604	0.0349	0.0269-0.0430	0.0215	0.0175-0.0256
NetRec2	0.0496	0.0388-0.0604	0.0350	0.0269-0.0430	0.0215	0.0175-0.0256
	Laval		Brampton			
	Mean	95% CI	Mean	95% CI		
MCNet	**0.2060**	0.1729-0.2390	**0.0958**	0.0831-0.1084		
SoftImpute	0.2366	0.2033-0.2700	0.1020	0.0897-0.1144		
TopN	0.2351	0.2018-0.2684	0.1004	0.0881-0.1128		
NetRec1	0.2235	0.1904-0.2566	0.1005	0.0882-0.1128		
NetRec2	0.2213	0.1881-0.2545	0.0996	0.0873-0.1119		

Table 4: The mean testing RMSEs over 50 repetitions for the five cities. The minimum value of each column is indicated by the bold font.

6. Conclusion

We propose the MCNet method to incorporate social network information for matrix completion. MCNet generates latent features from a social network using the random dot product graph model, and the features are

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0243

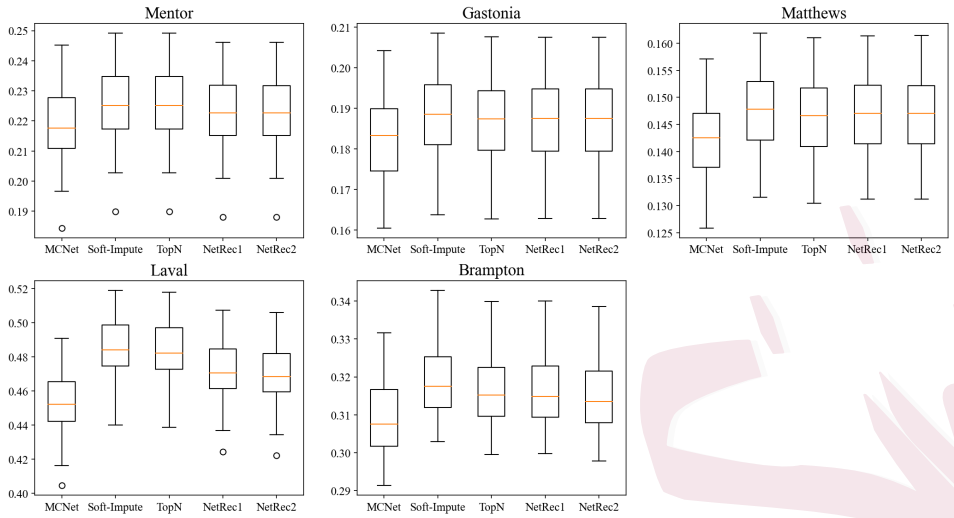

Figure 7: The testing RMSEs of MCNet and other methods over 50 repetitions.

used in the matrix completion to reduce the estimation errors. The algorithm is evaluated using extensive simulations and two real data analyses of Douban movie and Yelp data sets. The results show that MCNet outperforms the SoftImpute, TopN, and NecRec algorithms and provides robust performance when the signal-to-noise ratios are small.

In general, incorporating information in $\mathbf{Y}$ to estimate $\mathbf{X}$ will improve the estimation accuracy (Yu, Rao, and Dhillon (2016)). To achieve this goal, we could add an additional penalty term $\|\mathbf{M} - \mathbf{X}\mathbf{X}^{\mathrm{T}}\|_F$ to the loss function in (2.4) and (2.5). This enriched loss function will allow us to obtain $\widehat{\mathbf{X}}$, $\widehat{\boldsymbol{\beta}}$, and $\widehat{\mathbf{B}}$ in a single framework. However, this penalty function

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0243

is nonconvex in $\mathbf{X}$, which brings difficulties in parameter estimation. This topic is left to future research.

Supplementary Material

The online Supplementary Material contains detailed proofs for Lemmas 3.1–3.3 and Theorems 3.1–3.2.

Acknowledgments

The authors thank the reviewers and the associate editor for their constructive comments.

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.

Ahn, S. C. and Horenstein, A. R. (2013). Eigenvalue ratio test for the number of factors. *Econometrica* **81**, 1203-1227.

Athreya, A., Priebe, C. E., Tang, M., Lyzinski, V., Marchette, D. J. and Sussman, D. L. (2018). Statistical inference on random dot product graphs: a survey. *The Journal of Machine Learning Research* **18**, 1-92.

Athreya, A., Priebe, C. E., Tang, M., Lyzinski, V., Marchette, D. J. and Sussman, D. L. (2016).

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0243

A limit theorem for scaled eigenvectors of random dot product graphs. *Sankhya A* **78**, 1-18.

Bi, X., Qu, A., Wang, J. and Shen, X.(2017). A group-specific recommender system. *Journal of the American Statistical Association* **112**, 1344-1353.

Cai, J. F., Candès, E. J. and Shen, Z. (2010). A singular value thresholding algorithm for matrix completion. *SIAM Journal on optimization* **20**, 1956-1982.

Cai, T. T. and Zhou, W. X. (2018). Matrix completion via max-norm constrained optimization. *Electronic Journal of Statistics* **10**, 1493-1525.

Candes, E. J. and Plan, Y. (2010). Matrix completion with noise. *Proceedings of the IEEE* **98**, 925-936.

Chiang, K.-Y., Hsieh, C.-J. and Dhillon, I. S. (2015). Matrix completion with noisy side information. In *Advances in Neural Information Processing Systems*, 3447-3455.

Dai, B., Wang, J., Shen, X. and Qu, A. (2019). Smooth neighborhood recommender systems. *Journal of machine learning research* **20**, 1-24.

Diallo, A. O., Diop, A. and Dupuy, J. F. (2017). Asymptotic properties of the maximum-likelihood estimator in zero-inflated binomial regression. *Communications in Statistics-Theory and Methods* **46**, 9930-9948.

Elsener, A. and van de Geer, S. (2018). Robust low-rank matrix estimation. *The Annals of Statistics* **46**, 3481-3509.

Fan, J., Gong, W. and Zhu, Z. (2019). Generalized high-dimensional trace regression via nuclear

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0243

norm regularization. *Journal of Econometrics* **212**, 177-202.

Fan, J., Wang, W. and Zhu, Z. (2017). A shrinkage principle for heavy-tailed data: high-dimensional robust low-rank matrix recovery. *arXiv preprint arXiv:1603.08315.*

Feuerverger, A., He, Y. and Khatri, S. (2012). Statistical significance of the Netflix challenge. *Statistical Science* **27**, 202-231.

Foygel, R., Salakhutdinov, R., Shamir, O. and Srebro, N (2011). Learning with the weighted tracenorm under arbitrary sampling distributions. In *Advances in Neural Information Processing Systems*, 24.

Hall, D. B. (2000). Zero-inflated Poisson and binomial regression with random effects: a case study. *Biometrics* **56**, 1030-1039.

Jing, B., Li, T., Ying, N. and Yu, X. (2019). Collaborative filtering with awareness of social network. *Preprint.*

Kang, Z., Peng, C. and Cheng, Q. (2016). Top-n recommender system via matrix completion. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, 179-185.

Kim, E., Lee, M. and Oh, S. (2015). Elastic-net regularization of singular values for robust subspace learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 915-923.

Klopp, O. (2014). Noisy low-rank matrix completion with general sampling distribution. *Bernoulli* **20**, 282-303.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0243

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.

Lei, J. and Rinaldo, A. (2015). Consistency of spectral clustering in stochastic block models. *The Annals of Statistics* **43**, 215-237.

Li, H., Chen, N. and Li, L. (2012). Error analysis for matrix elastic-net regularization algorithms. *IEEE transactions on neural networks and learning systems* **23**, 737-748.

Liu, X. and Aberer, K. (2013). Soco: a social network aided context-aware recommender system. *Proceedings of the 22nd international conference on World Wide Web*, 781-802.

Liu, J. and Wu, C. (2017). Deep learning based recommendation: A survey. *International Conference on Information Science and Applications*, 451-458.

Lu, L. and Peng, X. (2013). Spectra of edge-independent random graphs. *Electronic Journal of Combinatorics* **20**, 27.

Lyzinski, V., Sussman, D. L., Tang, M., Athreya, A. and Priebe, C. E. (2014) Perfect clustering for stochastic blockmodel graphs via adjacency spectral embedding. *Electronic journal of statistics* **8**, 2905-2922.

Lyzinski, V., Tang, M., Athreya, A., Park, Y. and Priebe, C. E. (2016). Community detection and classification in hierarchical stochastic blockmodels. *IEEE Transactions on Network Science and Engineering* **4**, 13-26.

Ma, H., Zhou, D., Liu, C., Lyu, M. R. and King, I. (2011). Recommender systems with social

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0243

regularization. In *Proceedings of the fourth ACM international conference on Web search and data mining*, 287-296.

Mao, X., Chen, S. X. and Wong, R. K. (2019). Matrix completion with covariate information. *Journal of the American Statistical Association* **114**, 198-210.

Mao, X., Wong, R. K. and Chen, S. X. (2018). Matrix Completion under Low-Rank Missing Mechanism. *arXiv preprint arXiv:1812.07813.*

Mazumder, R., Hastie, T. and Tibshirani, R. (2010). Spectral regularization algorithms for learning large incomplete matrices. *Journal of machine learning research* **11**, 2287-2322.

Negahban, S. and Wainwright, M. J. (2011). Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *The Annals of Statistics* **39**, 1069-1097.

Oliveira, R. I. (2009). Concentration of the adjacency matrix and of the laplacian in random graphs with independent edges. *arXiv preprint arXiv:0911.0600.*

Rao, N., Yu, H. F., Ravikumar, P. and Dhillon, I. S. (2015). Collaborative filtering with graph information: Consistency and scalable methods. In *Advances in Neural Information Processing Systems*, 7.

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

Recht, B., Fazel, M. and Parrilo, P. A. (2010). Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review* **52**, 471-501.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0243

Robin, G., Wai, H.-T., Josse, J., Klopp, O. and Moulines, É. (2018). Low-rank interaction with sparse additive effects model for large data frames. In *Advances in Neural Information Processing Systems*, 5496-5506.

Rubin, D. B. (1976). Inference and missing data. *Biometrika* **63**, 581-592.

Shi, Y., Larson, M. and Hanjalic, A. (2014). Collaborative filtering beyond the user-item matrix: a survey of the state of the art and future challenges. *ACM Computing Surveys (CSUR)* **47**, 1-45.

Srebro, N., Rennie, J. and Jaakkola, T. S. (2005). Maximum-margin matrix factorization. In *Advances in Neural Information Processing Systems*, 1329-1336.

Sun, T and Zhang, C. H. (2012). Calibrated elastic regularization in matrix completion. In *Advances in Neural Information Processing Systems* 863-871.

Sussman, D. L., Tang, M., Fishkind, D. E. and Priebe, C. E. (2012). A consistent adjacency spectral embedding for stochastic blockmodel graphs. *Journal of the American Statistical Association* **107**, 1119-1128.

Sussman, D. L., Tang, M. and Priebe, C. E. (2013). Consistent latent position estimation and vertex classification for random dot product graphs. *IEEE transactions on pattern analysis and machine intelligence* **36**, 48-57.

Tang, M., Athreya, A., Sussman, D. L., Lyzinski, V., Park, Y. and Priebe, C. E. (2017). A semiparametric two-sample hypothesis testing problem for random graphs. *Journal of Computational and Graphical Statistics* **26**, 344-354.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0243

Tang, M. and Priebe, C. E. (2018). Limit theorems for eigenvectors of the normalized laplacian for random graphs. *The Annals of Statistics* **46**, 2360-2415.

Wang, H., Wang, N. and Yeung, D. Y. (2015). Collaborative deep learning for recommender systems. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, 1235-1244.

Xu, M., Jin, R. and Zhou, Z.-H. (2013). Speedup matrix completion with side information: application to multi-label learning. In *Advances in Neural Information Processing Systems*, 2301-2309.

Yelp (2019). Yelp dataset challenge @MISC. `http://www.yelp.com/dataset_challenge`.

Young, S. J. and Scheinerman, E. R. (2007). Random dot product graph models for social networks. In *International Workshop on Algorithms and Models for the Web-Graph*, 138-149. Springer.

Yu, L., Pan, R. and Li, Z. (2011). Adaptive social similarities for recommender systems. In *Proceedings of the fifth ACM conference on Recommender systems*, 257-260.

Yu, H. F., Rao, N. and Dhillon, I. S. (2016). Temporal Regularized Matrix Factorization for High-dimensional Time Series Prediction. In *Advances in Neural Information Processing Systems*, 847-855.

Zhang, S., Yao, L., Sun, A. and Tay, Y. (2019). Deep learning based recommender system: A survey and new perspectives. *ACM Computing Surveys (CSUR)* **52**, 1-38.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0243

REFERENCES

Zheng, J., Liu, J., Shi, C., Zhuang, F., Li, J. and Wu, B. (2017). Recommendation in heterogeneous information network via dual similarity regularization. *International Journal of Data Science and Analytics* **3**, 35-48.

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

Zou, H. and Hastie, T.(2005). Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)* **67**, 301-320.

Department of Epidemiology & Biostatistics, University of California, San Francisco, 550 16th St., San Francisco, CA 94158, USA.

E-mail: jingxuan.wang@ucsf.edu, fei.jiang@ucsf.edu

Innovation and Information Management, Faculty of Business and Economics, University of Hong Kong, Pokfulam Road, Hong Kong

E-mail: haipeng@hku.hk