Received 13 May 2019; | Revised 13 May 2019; | Accepted 13 May 2019
DOI: xxx/xxxx

ARTICLE TYPE

Factor Analysis on Report Citations, using a Combined Latent and logistic Model

Namjoon Suh*[1] | Xiaoming Huo[1] | Eric Heim[2] | Timothy Van Slyke[2] | Lee Seversky[2]

[1]Org Division, Org Name, State name, Country name
[2]Org Division, Org Name, State name, Country name

Correspondence
*Corresponding author name, This is sample corresponding address. Email: authorone@gmail.com

Present Address
This is sample for present address text this is sample for present address text

Summary

We propose a combined latent and logistic model for the citation network, where either a latent model or a logistic model alone is often insufficient to capture the structure of the data. The proposed model has a latent (i.e., factor analysis) model to represents the main technological trends (aka factors), and adds a sparse logistic component that captures the remaining ad hoc dependence. Parameter estimation are carried out through construction of a joint-likelihood function of edges and properly chosen penalty terms. The convexity of the likelihood function allows us to develop an efficient algorithm, while the penalty terms generate a low-dimensional latent component and a sparse graphical structure. Simulation results are reported that show the new method works well in practical situations. The proposed method has been applied to a real application in citation network of statistician data set (Ji and Jin, 2016 [11]).

KEYWORDS:
citation network, matrix decomposition, latent variable model, logistic model, convex optimization

1 | INTRODUCTION

We study a citation network, where each node (i.e., item) can be a technical report or a publication. A node may cite another node. Associated with a pair of nodes i and j, we denote a binary random variable X_{ij}, where $1 \leq i, j \leq n$ and n is the total number of nodes. We have $X_{ij} = 1$ if and only if either node i cites node j or vice versa; otherwise $X_{ij} = 0$. For each node i, we assume that there is an associated binary vector $f_i \in \mathbb{R}^K$, such that the kth entry of f_i, $f_{ik} = 1$, if and only if node i is related to topic (i.e., factor) k, $1 \leq k \leq K$. Here K is the total number of underlying topics (i.e., factors, or trends). We assume a logistic model for X_{ij}'s: for $1 \leq i, j \leq n$,

$$\mathbb{P}(X_{ij} = 1) = \frac{e^{\alpha + f_i^T D f_j}}{1 + e^{\alpha + f_i^T D f_j}}, \tag{1}$$

where $\alpha \in \mathbb{R}$ is a parameter and matrix $D \in \mathbb{R}^{K \times K}$ is a diagonal matrix: $D = \mathrm{diag}\{d_1, d_2, \ldots, d_K\}$. We assume $d_{ii} > 0$ for $1 \leq i \leq K$. Another way to put (1) is

$$\mathbb{P}(X_{ij} = 1) = \frac{\exp\left\{\alpha + \sum_{k=1}^K f_{ik} f_{jk} d_k\right\}}{1 + \exp\left\{\alpha + \sum_{k=1}^K f_{ik} f_{jk} d_k\right\}}. \tag{2}$$

[0]**Abbreviations:** fused factor analysis and logistic graphical models, ADMM, network model

A justification of the above model is that when both node i and node j are related to topic k, they have a higher chance to cite one way or the other. We have assumed a common strengthen coefficient d_k $(1 \leq k \leq K)$ for factor k, despite different nodes. We denote a matrix $F = \{f_1, f_2, \dots, f_n\} \in \mathbb{R}^{K\times n}$. Each column i in matrix F contains the factor loadings associated with the node i $(1 \leq i \leq n)$. Given the diagonal matrix D and the factor loading matrix F, we assume that X_{ij}'s are independent; therefore we have the total conditional probability function as follows:

$$\mathbb{P}(\{X_{ij}, 1 \leq i, j \leq n\}) = \prod_{1\leq i<j\leq n} \mathbb{P}(X_{ij}) = \prod_{1\leq i<j\leq n} \frac{e^{X_{ij}(\alpha + f_i^T D f_j)}}{1 + e^{\alpha + f_i^T D f_j}}, \tag{3}$$

where $\mathbb{P}(X_{ij})$ is given in (2). The last equation holds because X_{ij} only takes binary (i.e., 0 or 1) values. Recall that the dot product of two matrices with same dimensionality, $A, B \in \mathbb{R}^{a\times b}$, is defined as $A \bullet B = \text{trace}(A^T B) = \sum_{i=1}^{a} \sum_{j=1}^{b} a_{ij} b_{ij}$. The above (3) can be further rewritten as

$$\mathbb{P}(\{X_{ij}, 1 \leq i, j \leq n\}) = \frac{\exp(\alpha \sum_{1\leq i<j\leq n} X_{ij} + \frac{1}{2} X \bullet (F^T D F))}{\prod_{1\leq i<j\leq n} 1 + e^{f_i^T D f_j}}, \tag{4}$$

where we assume $X_{ii} = 0$ for all i $(1 \leq i \leq n)$ and $X_{ij} = X_{ji}$ for all i and j $(1 \leq i, j \leq n)$, i.e., the matrix X is symmetric. The above delivers a factor analysis model. Various linear and nonlinear latent variable models have been studied extensively in the literature (e.g.,[12, 16, 14, 17, 10, 13]).

Our work is motivated from a recent work *Fused Latent and Graphical (FLaG) model* (Chen et al, 2016, [6]). They assume that majority of variation of responses can be accounted by low dimensional latent vector, and remaining dependent structure of responses can be explained by sparse graphical structure. Thus, the resulting model contains a low-dimensional latent vector and a sparse conditional graph. Their key idea is to separate these two dependent structures so that they can facilitate the statistical inference. Our model also assumes that there exist two dependent structures among citation edges in network. A low-dimensional version of the aforementioned latent vector model is largely correct and majority of the citations among the nodes are induced by these common latent vectors f_i's (with weigh coefficients d_i's). There is still a small remainder due to the graphical component.

Though it may seem similar to Chen et al [6], we work on a different model formulation in several aspects. First, FLaG is built to analyze the Eysenck's Personality Questionnaire that consists of items designed to measure Psychoticism, Extraversion, and Neuroticism. So there are p questions that need to be answered, and each questions fall into above three categories. If there are n respondents to questions, we have n independent data generated from same distribution. In our case, the observed citation network can be thought of as one realization of random graph. Second, FLaG set a collection of binary responses for each individual questions follow a joint distribution which is a combination of Item Response Theory (IRT) model and Ising model, whereas we model the citation edges among papers as random variables, whose dependent structure is characterized by the combination of Latent Factor Analysis model and logistic model. Last but not least, FLaG approximate the original likelihood through constructing pseudo-likelihood function by taking advantage of conditional independence among the nodes, but our model's likelihood function is directly accessible due to the conditional independence among edges given parameters.

The proposed modeling framework is also related with the analysis of decomposing a matrix into low-rank and sparse components and the statistical inference of a multivariate Gaussian model whose precision matrix admits the form of a low-rank matrix plus a sparse matrix ([5, 4, 19]). However, the inference and optimization of the current model are different from those cases. We will construct a regularized-likelihood function, based on which estimator will be proposed for simultaneous model selection and parameter estimation. The optimization for the regularized estimator will be convex, for which we will develop an efficient algorithm through the alternating direction method of multiplier (ADMM : [3, 8, 9]).

The rest of the paper is organized as follows. In section 2 we will give a presentation on how to build a model which can encode both latent dependent structure due to the common topics and remaining ad-hoc dependent structure. In section 3, we will talk about assumptions in our model and penalization on likelihood function constructed in section 2. Section 4 gives the detailed procedure on how to compute the estimator of the optimization problem in Section 3. In Section 5, we will present simple numerical experiments on synthetic data, and application of our model on real citation network of statisticians. We finally conclude this work in section 6 with several open questions remain to be solved and some directions for future research.

2 | MODEL FORMULATION

Recall the following graphical model that was established in (4), which is essentially a factor model (latent variable model):

$$\mathbb{P}(\{X_{ij}, 1 \le i, j \le n\}) = \frac{\exp(\alpha \sum_{1\le i<j\le n} X_{ij} + \frac{1}{2} X \bullet (F^T DF))}{\prod_{1\le i<j\le n} 1 + e^{f_i^T Df_j}},$$

where $X_{ij}, 1 \le i, j \le n$, are binary random variables indicating either node i cites node j or vice versa, matrix $X = \{X_{ij}\} \in \mathbb{R}^{n\times n}$ is symmetric with diagonal entries all being equal to zero, factor loading matrix $F = [f_1, f_2, \dots, f_n] \in \mathbb{R}^{K\times n}$ records the relation between nodes and underlying topics, F^T is the transpose of F, and matrix $D \in \mathbb{R}^{K\times K}$ is diagonal with entries being the weight coefficients of factors.

The above specifies a latent model (or equivalently a factor model). We now describe a graphical model in the following. The graphical model will complement the latent model by characterizing links that are not interpretable via common factors. For the aforementioned binary random variable X_{ij}, $1 \le i, j \le n$, we define

$$\mathbb{P}(X_{ij} = 1) = \frac{e^{\alpha' + S_{ij}}}{1 + e^{\alpha' + S_{ij}}}, \tag{5}$$

where $S_{ij} \in \mathbb{R}$, for $1 \le i, j \le n$, denotes the relation between nodes i and j. If we have $S_{ij} = 0$, then it is less likely to have a citational relationship between nodes i and j. On the other hand, if $S_{ij} > 0$, then it is more likely to have a citation link between nodes i and j. Here parameter $\alpha' \in \mathbb{R}$ plays the same role as parameter α does in model (1). Denote the matrix $S = \{S_{ij}, 1 \le i, j \le n\} \in \mathbb{R}^{n\times n}$. Assume that given the matrix S, the binary random variables X_{ij}'s are independent; consequently, we have the total conditional probability function as follows:

$$\begin{aligned}
\mathbb{P}(\{X_{ij}, 1 \le i, j \le n\}) &= \prod_{1\le i<j\le n} \mathbb{P}(X_{ij}) \\
&= \prod_{1\le i<j\le n} \frac{e^{X_{ij}(\alpha' + S_{ij})}}{1 + e^{\alpha' + S_{ij}}} \\
&= \frac{\exp(\alpha' \sum_{1\le i<j\le n} X_{ij} + \frac{1}{2} X \bullet S)}{\prod_{1\le i<j\le n} 1 + e^{\alpha' + S_{ij}}}.
\end{aligned} \tag{6}$$

Recall that we have assumed that $X_{ii} = 0$ for all i ($1 \le i \le n$) and $X_{ij} = X_{ji}$ for all i and j ($1 \le i, j \le n$), i.e., the matrix X is symmetric. In the combined model, we integrate (4) and (6) to render the joint conditional probability function as follows:

$$\begin{aligned}
\mathbb{P}(X \mid \alpha, F, D, S) &= \prod_{1\le i<j\le n} \frac{e^{X_{ij}(\alpha + S_{ij} + f_i^T Df_j)}}{1 + e^{\alpha + S_{ij} + f_i^T Df_j}} \\
&= \frac{\exp\left(\alpha \sum_{1\le i<j\le n} X_{ij} + \frac{1}{2} X \bullet (F^T DF) + \frac{1}{2} X \bullet S\right)}{\prod_{1\le i<j\le n} \left(1 + e^{\alpha + f_i^T Df_j + S_{ij}}\right)}.
\end{aligned} \tag{7}$$

3 | ESTIMATION

Note that in the model (7), the log-likelihood function has the form as follows:

$$\begin{aligned}
\mathbb{L}(\alpha, F, D, S; X) = \alpha \sum_{1\le i<j\le n} X_{ij} + \frac{1}{2} X \bullet (F^T DF) + \frac{1}{2} X \bullet S \\
- \sum_{1\le i<j\le n} \log\left(1 + e^{\alpha + f_i^T Df_j + S_{ij}}\right).
\end{aligned} \tag{8}$$

If we consider maximizing the above log-likelihood function, we will encounter several technical issues which are described below.

1. We would like the matrix $S \in \mathbb{R}^{n\times n}$ to have as many zero entries as possible; i.e., matrix S is *sparse*.

2. There is an identifiability issue with the formation $F^T DF$. More specifically, let $P \in \mathbb{R}^{K\times K}$ be a signed permutation matrix, then we have $P^T P = I_n$, where $I_n \in \mathbb{R}^{K\times K}$ is the identity matrix. Notice that matrix $F' = PF$ is also a factor

loading matrix, and matrix $D' = PDP^T$ is still a diagonal matrix; we have

$$F^T DF = F^T P^T PDP^T PF = (F')^T D' F',$$

i.e., the choice of F and D is not unique.

3. We would like the number of nonzeros in each column of F to be small, reflecting that each node is associated with a small number of underlying topics.

4. Overall, the rank of matrix $F^T DF$ cannot be larger than $\min\{n, K\}$. With the application that we have in mind, in this paper, we assume that K is much smaller than n.

5. To ensure the separation of matrices $\alpha \mathbb{1}\mathbb{1}^T$ and L, we assume that the eigen-vector of L is centered, that is,

$$JLJ = L \quad \text{where} \quad J = I_n - \frac{1}{n}\mathbb{1}\mathbb{1}^T.$$

This condition uniquely identifies F up to a common orthogonal transformation of its columns.

Directly maximizing the objective function in (8) is not going to be an easy task. Following the approaches that were mentioned in Introduction, we propose to relax $F^T DF$ to L, where L is a low rank matrix. Consequently, the log-likelihood function in (8) can be rewritten as

$$\begin{aligned} \mathbb{L}_n(\alpha, L, S; X) = \alpha & \sum_{1 \le i < j \le n} X_{ij} + \frac{1}{2} X \bullet L + \frac{1}{2} X \bullet S \\ & - \sum_{1 \le i < j \le n} \log\left(1 + e^{\alpha + L_{ij} + S_{ij}}\right). \end{aligned} \tag{9}$$

We propose a penalize likelihood estimation approach as follows:

$$(\hat{\alpha}, \hat{L}, \hat{S}) = \arg\min_{\alpha, L, S} \left\{ -\frac{1}{n} \mathbb{L}_n(\alpha, L, S; X) + \gamma \|S\|_1 + \delta \|L\|_* \right\}, \tag{10}$$

where $\gamma > 0$ and $\delta > 0$ are algorithmic parameters whose values will be discussed later, the L_1 norm of matrix S is defined as $\|S\|_1 = \sum_{i \ne j} S_{ij}$ (Note that we do not penalize the diagonal entries of S), and nuclear norm of matrix L is defined as $\|L\|_* = \text{trace}\sqrt{(L^T L)}$. Recall that both S and L are symmetric matrices. The entries of matrix S can either be positive or negative. Note that we have imposed the diagonal entries of the matrix X to be zeros. Given that $L = F^T DF$ where matrix D is diagonal with nonnegative diagonal entries, it is easy to see that matrix L is positive semidefinite; which consequently leads to $\|L\|_* = \text{trace}(L)$, which is a linear functional to the matrix L. The nuclear norm of L penalizes the number of nonzero eigenvalues of L, which is the same as the rank of L. The regularization based on the nuclear norm was proposed in [7] and its statistical properties are studied in [2].

After we have obtained $\hat{S}$ in (10), we can uncover the graphical model by investigating non-zero entries in $\hat{S}$. On the other hand when we have calculated $\hat{L}$, we may not be able to find binary matrix F and nonnegative diagonal matrix D such that $\hat{L} = F^T DF$. This is the price we have to pay for an amenable computational problem. The rank of estimated $\hat{L}$ will be our estimate of the number of factors (i.e., underlying common topics). For the issue on assigning the community membership of each node i, we will discuss this later in Section 6.

4 | COMPUTATION

We propose a method that takes advantage of the special structure of the L_1 and nuclear norms by means of the alternating direction method of multiplier (ADMM), which is a method that has recently gained momentum. An examination of the objective function in (10) unvails that terms

$$\alpha \sum_{1 \le i < j \le n} X_{ij} + \frac{1}{2} X \bullet L + \frac{1}{2} X \bullet S$$

are linear in α, L, and S. The term

$$\sum_{1 \le i < j \le n} \log\left(1 + e^{\alpha + L_{ij} + S_{ij}}\right)$$

is convex with respect to α, L, and S. Functions $\|S\|_1$ and $\|L\|_*$ are known to be convex functions. Therefore, the objective function in (10) is convex. The above convex optimization problem can be solved via ADMM as follows.

4.1 | ADMM approach

We give a review of the alternating direction method of multiplier (ADMM). Consider two closed convex functions

$$f : \chi_f \to \mathbb{R} \text{ and } g : \chi_g \to \mathbb{R},$$

where the domain χ_f and χ_g of functions f and g are closed convex subsets of $\mathbb{R}^d$, and $\chi_f \bigcap \chi_g$ is nonempty. Both f and g are possibly non-differentiable. The alternating direction method of multiplier is an iterative algorithm that solves the following generic optimization problem:

$$\min_{x\in\chi_f \bigcap \chi_g} \{f(x) + g(x)\},$$

or equivalently

$$\begin{aligned} &\min_{x\in\chi_f, z\in\chi_g} \{f(x) + g(z)\}, \\ &\text{subject to} \qquad x = z. \end{aligned} \tag{11}$$

To describe the algorithm, we will need the following proximal operators

- $\mathbf{P}_{\lambda,f} : \mathbb{R}^d \to \chi_f$ as
$$\mathbf{P}_{\lambda,f}(v) = \arg\min_{x\in\chi_f} \left\{ f(x) + \frac{1}{2\lambda}\|x - v\|_2^2 \right\}$$
- and $\mathbf{P}_{\lambda,g} : \mathbb{R}^d \to \chi_g$ as
$$\mathbf{P}_{\lambda,g}(v) = \arg\min_{x\in\chi_g} \left\{ g(x) + \frac{1}{2\lambda}\|x - v\|_2^2 \right\},$$
where $\|\cdot\|_2$ is the usual Euclidean norm on $\mathbb{R}^d$ and λ is a scale parameter that is a fixed positive constant.

The algorithm starts with some initial values $x^0 \in \chi_f, z^0 \in \chi_g, u^0(= \lambda y^0) \in \mathbb{R}^d$. At the $(m + 1)$th iteration, (x^m, z^m, u^m) is updated according to the following steps until convergence

- Step 1: $x^{m+1} = \mathbf{P}_{\lambda,f}(z^m - u^m)$,
- Step 2: $z^{m+1} = \mathbf{P}_{\lambda,g}(x^{m+1} + u^m)$,
- Step 3: $u^{m+1} = u^m + x^{m+1} - z^{m+1}$.

The convergence properties of the algorithm are summarized in the following result in [3]. Let p^* be the minimal value in (11).

Theorem 1 (Boyd et al., 2011)**.** Assume functions $f : \chi_f \to \mathbb{R}$ and $g : \chi_g \to \mathbb{R}$ are closed convex functions, whose domains χ_f and χ_g are closed convex subsets of $\mathbb{R}^d$ and $\chi_f \bigcap \chi_g \neq \emptyset$. Assume the Lagrangian of (11)

$$L(x, z, y) = f(x) + g(z) + y^T(x - z)$$

has a saddle point, that is, there exists (x^*, z^*, y^*) (not necessarily unique) that $x^* \in \chi_f$ and $z^* \in \chi_g$, for which

$$L(x^*, z^*, y) \le L(x^*, z^*, y^*) \le L(x, z, y^*), \qquad \forall x, z, y \in \mathbb{R}^d.$$

Then the ADMM has the following convergence properties.

1. Residual convergence. $x^m - z^m \to 0$ as $m \to \infty$; i.e., the iterates approach feasibility.
2. Objective convergence. $f(x^m) + g(z^m) \to p^*$ as $m \to \infty$; i.e., the objective function of the iterates approaches the optimal value.
3. Dual variable convergence. $y^m \to y^*$ as $m \to \infty$, where y^* is a dual optimal point.

Now we see how ADMM can be adopted to solve for our penalized likelihood estimation (10). We reparameterize $M = L+S$ and let $x = (\alpha, M, L, S)$ (viewed as a vector). We define the following:

$$\begin{aligned} \chi_f &= \{(\alpha, M, L, S) : \alpha \in \mathbb{R}, M, L, S \in \mathbb{R}^{n\times n}, L \text{ is positive semidefinite}, S \text{ is symmetricg}\}, \\ f(x) &= -\frac{\alpha}{n}\sum_{1\le i<j\le n} X_{ij} - \frac{1}{2n} X \bullet M + \frac{1}{n}\sum_{1\le i<j\le n} \log\left(1 + e^{\alpha + M_{ij}}\right) + \gamma\|S\|_1 + \delta\|L\|_*, \\ \chi_g &= \{(\alpha, M, L, S) : \alpha \in \mathbb{R}, M, L, S \in \mathbb{R}^{n\times n}, M \text{ is symmetric and } M = L + S\}, \text{ and} \\ g(x) &= 0, \text{ for } x \in \chi_g. \end{aligned}$$

One can verify that (10) can be written as

$$\min_{x \in \chi_f \bigcap \chi_g} \{f(x) + g(x)\}.$$

We now present each of the three steps of the ADMM algorithm and show that the proximal operators $\mathbf{P}_{\lambda,f}$ and $\mathbf{P}_{\lambda,g}$ are easy to evaluate. Let

$$x^m = (x^m_\alpha, x^m_M, x^m_L, x^m_S), \quad z^m = (z^m_\alpha, z^m_M, z^m_L, z^m_S), \quad u^m = (u^m_\alpha, u^m_M, u^m_L, u^m_S).$$

Step 1. We solve $x^{m+1} = \mathbf{P}_{\lambda,f}(z^m - u^m)$. Due to the special structure of $f(\cdot)$, x^{m+1}_α, x^{m+1}_M, x^{m+1}_L, and x^{m+1}_S can be updated separately. More precisely, we have

$$\begin{aligned}
x^{m+1}_\alpha, x^{m+1}_M &= \arg\min_{\alpha,M} \quad -\frac{\alpha}{n}\sum_{1\le i<j\le n} X_{ij} - \frac{1}{2n} X \bullet M + \frac{1}{n}\sum_{1\le i<j\le n} \log\left(1 + e^{\alpha + M_{ij}}\right) \\
&\quad + \frac{1}{2\lambda}\left[\alpha - (z^m_\alpha - u^m_\alpha)\right]^2 + \frac{1}{2\lambda}\|M - (z^m_M - u^m_M)\|_F^2, && (12)\\
x^{m+1}_L &= \arg\min_L \quad \delta\|L\|_* + \frac{1}{2\lambda}\|L - (z^m_L - u^m_L)\|_F^2, && (13)\\
&\quad \text{subject to } L \text{ is positive semidefinite;} \\
x^{m+1}_S &= \arg\min_S \quad \gamma\|S\|_1 + \frac{1}{2\lambda}\|S - (z^m_S - u^m_S)\|_F^2, && (14)\\
&\quad \text{subject to } S \text{ is symmetric,}
\end{aligned}$$

where $\|\cdot\|_F$ is the matrix Frobenius norm, defined as $\|M\|_F^2 = \sum_{i,j} m_{ij}^2$ for a matrix $M = (m_{ij})$. The problem in (12) may not have a closed-form solution. We use a simple gradient descent to solve this step, setting the step size, $\alpha = 0.1$ and stopping criteria, $\max(|x^{(t+1)}_{\alpha,m} - x^{(t)}_{\alpha,m}|, \|x^{(t+1)}_{M,m} - x^{(t)}_{M,m}\|_\infty) \le 10^{-9}$. Note that there are close-form solutions to (13) and (14), while (12) is a unconstrained convex optimization problem. More specifically, in (13), suppose the eigenvalue decomposition of the symmetric matrix $(z^m_L - u^m_L)$ can be written as

$$z^m_L - u^m_L = T\Lambda T^T,$$

where T is orthogonal ($TT^T = I_n$), then, for $J = I_n - \frac{1}{n}\mathbb{1}\mathbb{1}^T$, we have

$$x^{m+1}_L = J(T\mathrm{diag}(\Lambda - \lambda\delta)_+ T^T)J^T,$$

and $\mathrm{diag}(\Lambda - \lambda\delta)_+$ is a diagonal matrix with the jth diagonal entry being

$$(\Lambda_{jj} - \lambda\delta)_+ = \begin{cases} 0, & \text{if } \Lambda_{jj} < \lambda\delta \\ \Lambda_{jj} - \lambda\delta, & \text{if } \Lambda_{jj} \ge \lambda\delta. \end{cases}$$

In (14), we have, for $i \neq j$,

$$S_{ij} = \begin{cases} 0, & \text{if } |(z^m_S - u^m_S)_{ij}| < \lambda\gamma \\ (z^m_S - u^m_S)_{ij} - \lambda\gamma, & \text{if } (z^m_S - u^m_S)_{ij} > \lambda\gamma \\ (z^m_S - u^m_S)_{ij} + \lambda\gamma, & \text{if } (z^m_S - u^m_S)_{ij} < -\lambda\gamma \end{cases}$$

Step 2. We solve $z^{m+1} = \mathbf{P}_{\lambda,g}(x^{m+1} + u^m)$. A close-form solution exists here. Denote $\bar\alpha = x^{m+1}_\alpha + u^m_\alpha$, $\bar M = x^{m+1}_M + u^m_M$, $\bar L = x^{m+1}_L + u^m_L$, and $\bar S = x^{m+1}_S + u^m_S$, then evaluating $\mathbf{P}_{\lambda,g}(x^{m+1} + u^m)$ becomes

$$\begin{aligned}
\min_{\alpha,M,L,S} \quad & \tfrac{1}{2}[\alpha - \bar\alpha]^2 + \tfrac{1}{2}\|M - \bar M\|_F^2 + \tfrac{1}{2}\|L - \bar L\|_F^2 + \tfrac{1}{2}\|S - \bar S\|_F^2 \\
\text{subject to} \quad & M \text{ is symmetric and } M = L + S.
\end{aligned}$$

The above optimization problem has a close-form solution, which is as follows:

$$\begin{aligned}
z^{m+1}_\alpha &= \bar\alpha, \\
z^{m+1}_M &= \frac{1}{3}\bar M + \frac{1}{3}\bar M^T + \frac{1}{3}\bar L + \frac{1}{3}\bar S \\
z^{m+1}_L &= \frac{1}{6}\bar M + \frac{1}{6}\bar M^T + \frac{2}{3}\bar L - \frac{1}{3}\bar S \\
z^{m+1}_S &= \frac{1}{6}\bar M + \frac{1}{6}\bar M^T - \frac{1}{3}\bar L + \frac{2}{3}\bar S.
\end{aligned}$$

Step 3. We solve $u^{m+1} = u^m + x^{m+1} - z^{m+1}$, which is a simple arithmetic.

The most important implementation details of this algorithm are the choice of λ and stopping criterion. In this work, we simply choose $\lambda = 0.5$. We terminate the algorithm when in mth iteration, $\|x^m_M - x^m_L - x^m_S\|_F \le \delta$, with $\delta = 10^{-7}$.

5 | NUMERICAL ANALYSIS AND APPLICATIONS

In this section, we have conducted a small empirical study of measuring the performance of our proposed method with artificially specified graphical structures. Also we perform a real data analysis with citation network for statisticians.

5.1 | Synthetic data

Data Generation. Following the notation of our paper, each edge X_{ij} follows Bernoulli distribution, whose parameter is parameterized by the probability, $P_{ij} = \frac{\exp(\alpha+F_i^T DF_j+S_{ij})}{1+\exp(\alpha+F_i^T DF_j+S_{ij})}$. For sparse component S, we generate random numbers uniformly distributed over the indices of upper triangular part (off diagonal) of matrix S, and make it symmetry. F^T binary matrix has $\lfloor \frac{n}{K} \rfloor$ ones in each columns, and has exactly one 1 in each rows. We randomly choose one of rows, fill that row with ones, and project the column space of F onto the orthogonal complement subspace of $\mathbb{1}$ vector by multiplying the aforementioned $J = I_n - \frac{1}{n}\mathbb{1}\mathbb{1}^T$ to F. Diagonal elements of D are generated from Unif$[7, 8]$ (uniform distribution on $[7, 8]$), and alpha is generated from Unif$[-3, -2]$. We set $K = 3, 4, 5$, $n = 30, 80, 120$, and $nnz = 10, 20, 40$, where nnz denotes number of non-zero component of the upper-triangular part of S.

Choosing the tuning parameters and evaluation criteria. We have to admit the fact that choosing a good pair of tuning parameter is an important but challenging issue in our problem. We here present a purely heuristic approach. Following the Ji and Jin [11], we plot the largest 20 eigenvalues of adjacency matrix X, which can tell us the numbers of communities embedded in network. Then, we record the rank of L and number of non-zero elements in S for each tuning parameter pair on grid, and choose a proper pair which can give us interesting interpretation of data.

One might wonder how traditional model selection methods, such as the Bayes Information Criterion (BIC;[18]) or the AIC, work where

$$\text{BIC}(M) = -2\mathbb{L}_n(\hat{\beta}(M)) + |M| \log\left(\frac{n(n-1)}{2}\right),$$

and

$$\text{AIC}(M) = -2\mathbb{L}_n(\hat{\beta}(M)) + 2|M|,$$

where M is the current model, $\mathbb{L}_n(\hat{\beta}(M))$ is the maximal log-likelihood for a given model M, and $|M|$ is the number of free parameters in M, which is determined by the number of non-zeros in S and the low-rank matrix L. If rank$(L) = K$, we can establish the following

$$|M| = \sum_{i<j} 1_{\{S_{ij}>0\}} + nK - \frac{K(K-1)}{2} + 1$$

Because the number of free parameters in L is K plus $nk - K(K+1)/2$, which is the number of free parameters in determining K orth-normal vectors. Additionally 1 in the last is due to α. We want to find an M which minimizes BIC(M) or AIC(M) as a function of M. However these traditional methods seem to entail a serious problem. The assumptions used to derive BIC or AIC are not satisfied here, since each edges in a graph are not identically distributed. Our experience shows that both BIC and AIC choose the most parsimonious model with the smallest K and smallest cardinality of ad-hoc dependency among all the estimated $(\widehat{L},\widehat{S})$ from a pair of tuning parameters over the given range of grid. We evaluate the models selected via our heuristic approach, BIC, and AIC by using following three evaluation metrics.

$$\begin{aligned}
M1 &= \mathbb{1}\left\{rank(\widehat{L}) = rank(L^*)\right\} \\
M2 &= \frac{|\left\{(i,j) : i<j : S^*_{i,j} \neq 0 \quad \& \quad \widehat{S}_{i,j} \neq 0\right\}|}{|\left\{(i,j) : i<j : S^*_{i,j} \neq 0\right\}|} \\
M3 &= \frac{|\left\{(i,j) : i<j : S^*_{i,j} = 0 \quad \& \quad \widehat{S}_{i,j} \neq 0\right\}|}{|\left\{(i,j) : i<j : S^*_{i,j} = 0\right\}|}
\end{aligned}$$

$M1$ is a metric on whether selected model recovers true low rank structure of network, $M2$ evaluates the positive selection rate of sparse network structure, and lastly $M3$ evaluates the false discovery rate. With properly selected tuning parameter, $M1$ will be 1, $M2$ will be close to 1, and $M3$ will get close to 0. Evaluated results of models selected via three criteria are presented in Table 1.

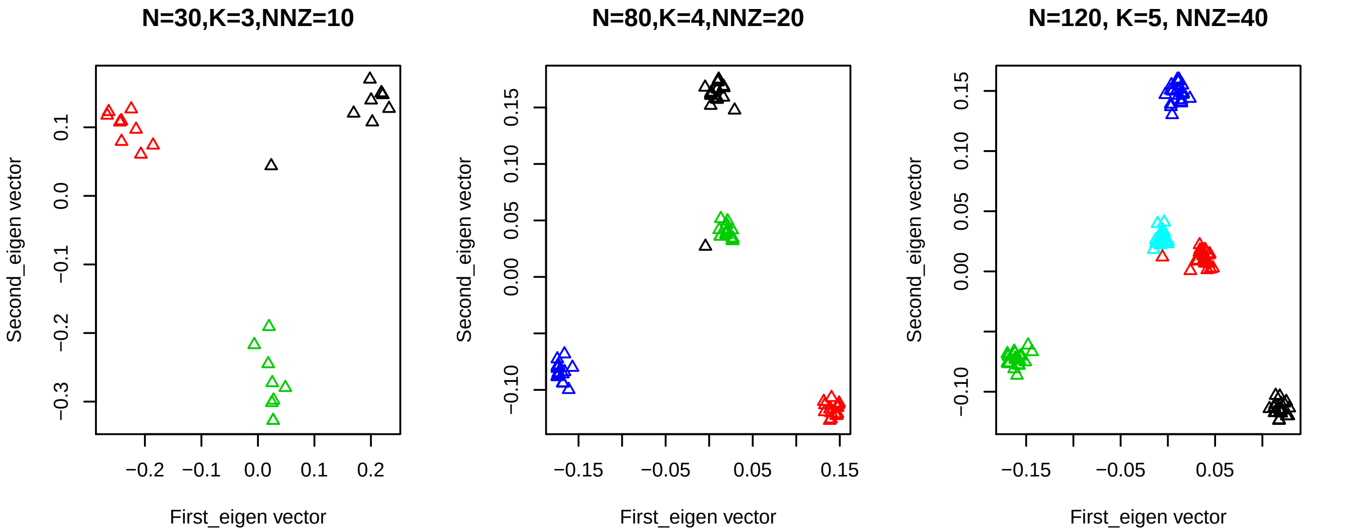

FIGURE 1 After fitting the model, result of application of k-means algorithm on first K eigenvectors of $\widehat{L}$ matrix. (Different color represents different clusters of nodes k-means assign for three synthetic settings.)

	Case 1			Case 2			Case 3		
	Heuristic	AIC	BIC	Heuristic	AIC	BIC	Heuristic	AIC	BIC
$M1$	1	0	0	1	0	0	1	0	0
$M2$	0.5	0	0	0.842	0	0	0.725	0	0
$M3$	0.007	0	0	0.008	0	0	0.0113	0	0

TABLE 1 Both AIC and BIC choose the model $\widehat{S} = 0$ with no ad-hoc structure, and $\widehat{L}$ with the lowest rank among all estimators via a pair of tuning parameter over the grid of search region. Our heuristic method correctly choose the model with $\widehat{L}$ with true rank and $\widehat{S} = 0$ catching all the ”across edges” of network. Note that

Community memberships of each nodes. Given a network data X, after fitting the model with a proper pair of tuning parameters, (γ, δ), we need to determine whether node i belongs to topic k. We applied a simple k-means clustering on the fitted L matrix's first K eigenvectors. Furthermore, we find one interesting phenomenon. Coordinates of each rows from first two eigenvectors of estimated $\widehat{L}$ matrix in our model characterize the clustering behavior of embedded topics reasonably well. Our experience shows that with only first two eigenvectors of $\widehat{L}$, we can also cluster the nodes comparably well.

Ad-hoc links. Ad-hoc links of synthetic network data can be thought of as “across edges” between clusters of nodes. Since indices of non-zero entries of S are randomly chosen, there might be some indices of non-zero S_{ij} entries where L_{ij} is also non-zero. In this case, this makes the edges indistinguishable if they come from S_{ij} or L_{ij}. Some of “across edges" are also attributed to α, and unfortunately, our algorithm cannot make a difference whether the edge comes from α or S_{ij}. So the total number of across edges might not be exactly same as we set in our data generation setting.

Simulation Results. Our goal in this simulation experiment is to check if our model can cluster each nodes into correct communities. Also to see if it can separate the ad-hoc links from edges which come from common topics. Tuning parameter pairs, $(\gamma, \delta) = (0.01, 0.03), (0.004, 0.014), (0.003, 0.013)$, respectively, give us the desirable results for the three cases. Note that different dynamics of network requires different choices of tuning parameters even the synthetic setting is same, since we are generating the random graph. However, we set the seeds in our simulation so that the results are reproducible. Fig. 1 shows us the result of clustered nodes using the k-means algorithm on the first $K(= \text{rank}(\widehat{L}))$ eigenvectors. We plot the rows of first two eigenvectors on the cartesian plane since it visualizes well the clustering behavior of the nodes. Our model chooses well the ad-hoc links between clusters of nodes as can be verified in Fig. 2 . We color the edges in blue whose corresponding elements of estimated $\widehat{S}$ are non-zero, and black whose corresponding entries of $\widehat{S}$ are zero.

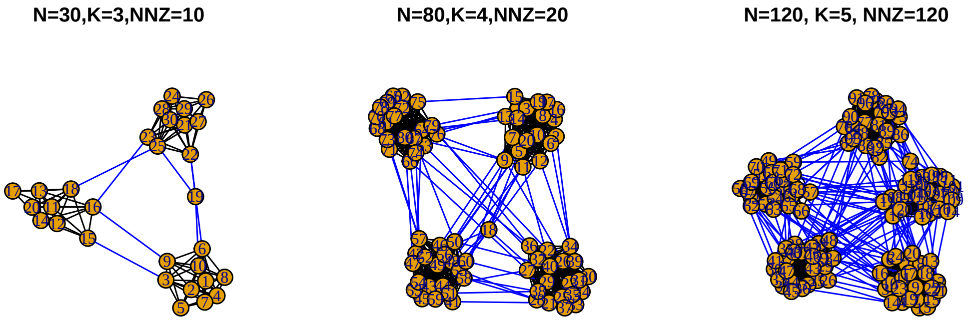

FIGURE 2 Edges colored with blue correspond to non-zero entries of estimated $\widehat{S}$ matrix. For all three cases, our algorithm correctly separates the "across edges" from edges coming from low-rank matrix structure.

5.2 | Citation networks for statisticians

Recently, Ji and Jin [11] published an interesting data set on citation and co-authorship networks of statisticians. Citation relationship over 3000 papers from 4 statistical journals, published between 2003 and 2012, were collected, and collaborative network between authors of papers is given as well. In this section, we analyze the citation network.

Citation Network This dataset is based on all published papers from 2003 to first half of 2012, in 4 of the top statistical journals: Annals of Statistics, Biometrika, Journal of American Statistical Association and Journal of Royal Statistical Society (Series B). Among 3248 papers, we restrict our attention on papers which have greater than or equal to 7 citations from collected papers in original dataset. We have 543 papers in our hands. The elbow points of the scree plot may be at the 3rd, 7th, or 9th largest eigenvalue, suggesting that there might be from 2 to 8 topics in network. In light of this, we performed our analysis under the assumption that there might exist 2 distinct topics and one giant component in which 3 or 4 topics mixed up.

$(\gamma, \delta) = (0.000912, 0.0097)$ gives us $\widehat{L}$ with rank 3, and $\widehat{S}$ with $|supp(\widehat{S})| = 197$. We apply k-means algorithm on 3 eigenvectors of estimated $\widehat{L}$ matrix, and visualize the result of clustering by plotting the rows of first two eigenvectors of $\widehat{L}$ on plane in $Fig4$. 404 papers colored in green are densely clustered nearby the origin with a short tail, ($Fig4$), and k-means classifies these papers as third topic. We list the first two topics discovered through our analysis. (Full lists of papers for each communities are provided in (*https://sites.google.com/site/namjoonsuh/publications*)

- Variable selection (85 papers)
- Multiple Hypothesis Testing (54 papers)

These two topics are classical research topics in field of statistics. First group talks about "Variable selection" in high-dimensional data. Majority of papers in second community study about "Controlling False Discovery Rate" in various statistical settings. As expected, third group is quite hard to interpret and seems to have substructures. For further investigation on this group, after obtaining the sub-network which is comprised only with papers in third group, we performed same analysis once again under the assumption that $K = 4$. We obtain four sub-communities as follows:

- Non-parametric Bayesian Statistics (17 papers)
- Functional Data analysis (35 papers)
- Dimension Reduction (23 papers)
- Mixed topics (329 papers)

Out of this one big chunk, we got three small, but meaningful topics: "Bayesian Statistics", "Functional/longitudinal data analysis", "Dimension Reduction". Due to small volume of each communities, we could check that false discoveries for each community are all zero.

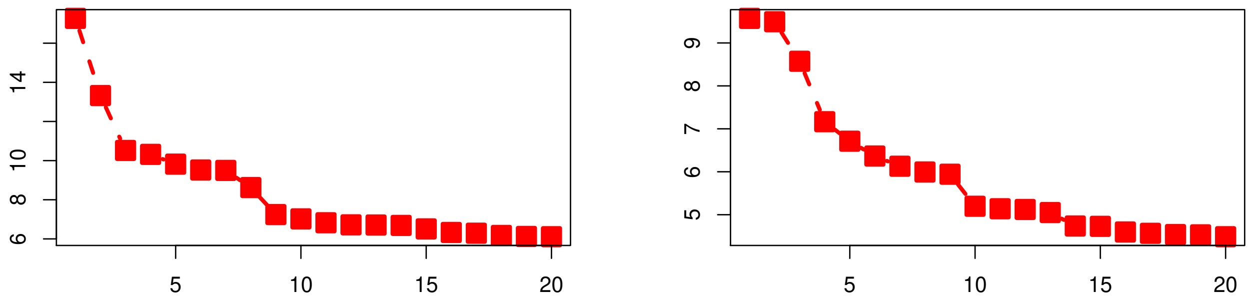

FIGURE 3 *Scree plots. From left to right: 543-Citation network, Mixed Topics-Citation network.*

Sub-network structure has also a big chunk of papers which we refer to as "Mixed Topics". Not only could we see the papers with topics on "Learning Theory", "Non-parametric/semi-parametric statistics", "Spatial Statistics", "Theoretical Machine Learning", which does not seem to belong to any of five communities listed above, but also we could identify the papers with combinations of two or three topics. Papers like "The Bayesian Lasso (T. Park, et al. 2008)", "Coordinate-independent sparse sufficient dimension reduction and variable selection (X. Chen, et al. 2010)" are the examples of these papers. It is also interesting to think about a reason on papers which seem to have obvious membership in one of 5 communities other than Mixed topic classified as Mixed topic. For instance, "On the "degrees of freedom" on the LASSO (H. Zou, et al. 2007)" is classified as Mixed Topic paper. We can simply guess model selection has lots of applications in other topics, so it might cite or have been cited by many papers in other communities. Actually, out of 19 citation relationships it has with other papers, 12 of them came from the relationships with papers from Mixed topics.

It is worth to note that our result is consistent with that from Ji and Jin in a sense that they also recover "Multiple Testing", "Variable selection" and "non-parametric Bayesian statistics". This is interesting since even though the dataset we consider is different from theirs, result is consistent. They focus their attentions on analyzing *"weakly connected giant component"* for a citation network of each paper's authors (i.e., each node is an author). We consider the citation network of papers whose node degree is greater than or equal to 7.

Ad-hoc links. Non-zero components of $\widehat{S}$ capture the citation relationships among papers that are not attributable to the common topics. The model we select has 197 sparse edges (9% of total edges), and all of them are positive edges. In Table 1, we provide 15 pairs of papers that have the most positive edges. All the 15 edges come from pairs of papers from different communities. For instance, the first pair of papers comes from the Functional Analysis community (denoted by *FuncAn*) and Variable selection (*VarSel*) community. *FuncAn* paper cites *VarSel* paper for borrowing a mathematical representation to build a theorem. Though it might seem to be a crucial step for building a theorem in their paper, we cannot say that two papers are closely related in terms of topic. Second pair comes from *MulT* and *VarSel* community. *MulT* paper briefly mention about *VarSel* paper in future work section, suggesting a possible way of combining their work and work in *VarSel* paper. And we also observe that as the weight of edges decreases, increasing number of pairs of papers both come from "Mixed topics" community.

6 | CONCLUSION

We propose a new combined latent factor analysis and logistic model. We consider the regularized likelihood by means of $\ell 1$ and nuclear norm penalties. The computation of the regularized estimator is facilitated by developing an algorithm based on the alternating direction method of multiplier to optimize a non-smooth and convex objective function. The proposed method is applied to citation and co-authorship network of statisticians, and the estimated model renders good interpretative power. Specifically, our analysis on statistician's citation network sheds the new light on the interpretation of dataset.

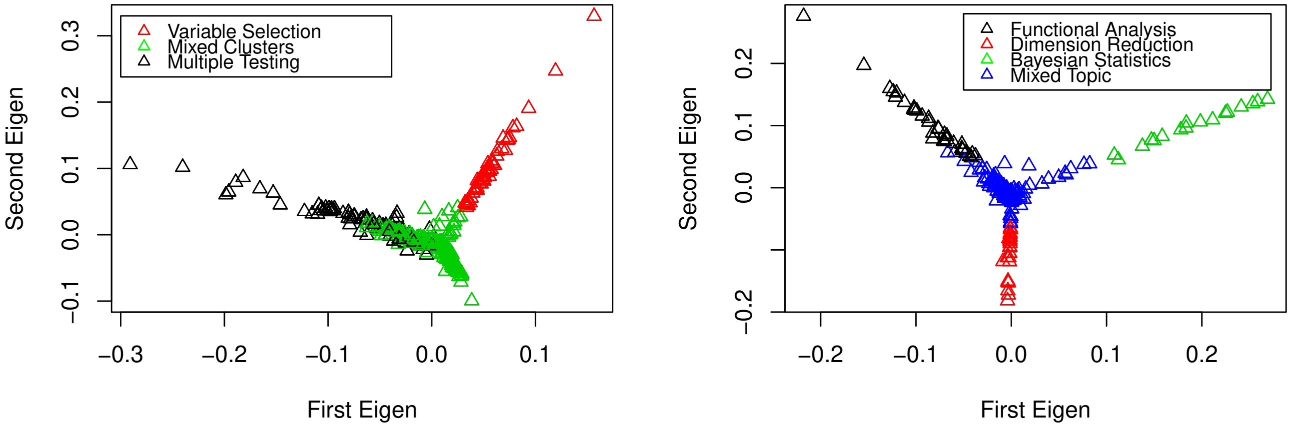

FIGURE 4 Illustration of first two eigenvectors of estimated $\widehat{L}$ matrix for original citation network with 543 papers(left) and sub-network graph(right). Presentation on clustering results of k-means algorithm with different colors.

However, there are still several open problems which require us to do more careful analysis. First of all, it remains unclear on how to choose the proper tuning parameters. Classical methods for choosing tuning parameters such as BIC or AIC did not work in our case, since they tend to choose the most parsimonious models. We also do not have systematic ways to do cross validation in network data. Not only because it is computationally expensive procedure, but also because if we partition the network data, we can loose fair amount of information on dependent structures among edges. This problem is also closely related with determining the number of communities in network. In lieu of using BIC or AIC, our analysis is heavily relying on heuristic approach when choosing the tuning parameter, and during this procedure, we use the screeplot for determining number of communities in network. Screeplot approach works usually well, but it does not necessarily always guarantee the correct estimate of number of communities. We need a more reliable and theoretically well understood way to determine K.

Secondly, we only consider the undirected case which is somewhat unrealistic assumption in real citation network, in a sense that it is not usual for two papers to cite each other at the same time. Since in our research, we were interested in separating the low rank structure of edges and ad-hoc links in network, we did not take into account the directionalities of edges in our model. However, it would be interesting to think about the way to incorporate the directed network into our matrix decomposition framework.

Last but not least, when we assign the memberships of each nodes, we use k-means clustering algorithm. However, k-means algorithm turns out that it tends to assign nodes conservatively to each communities. For example, in $Fig4(a)$, we can see that bunch of Multiple Testing papers are assigned as Mixed cluster, and in $Fig4(b)$ also, many papers which should have been classified to among three communities other than mixed topic, have been assigned as Mixed topic community. This is probably because k-means does not allow overlapping membership of each nodes. It would be interesting to see what happens if we apply some clustering methods which allow the overlapped memberships of nodes in network.

7 | AUXILIARY RESULT ON NON-ASYMPTOTIC ERROR BOUND OF THE ESTIMATOR

In this section, we focus on investigating the behaviour of non-asymptotic error bound of our estimator in the context where the sample size is fixed as 1, and the number of variables in model is explicitly tracked. We are interested in solving following optimization problem :

$$\min_{\substack{\alpha \in R, S=S^T \\ L \succ 0}} -\frac{1}{n} \log \prod_{1 \le i,j \le n} \frac{exp\left(X_{ij}\left(\alpha + L_{ij} + S_{ij}\right)\right)}{1 + exp\left(\alpha + L_{ij} + S_{ij}\right)} + \delta \|L\|_* + \gamma \|S\|_1 \tag{15}$$

Let $(\hat{\alpha}, \widehat{L}, \widehat{S})$ be the solution to (15), and (α^*, L^*, S^*) be the ground truth which governs the data generating process. Let $\widehat{\Theta}, \Theta^*$ be defined respectively as $\widehat{\Theta} = \hat{\alpha}\mathbb{1}\mathbb{1}^T + \widehat{L} + \widehat{S}$ and $\Theta^* = \alpha^*\mathbb{1}\mathbb{1}^T + L^* + S^*$. And denote the error term for each parameter as $\widehat{\Delta}^\Theta = \widehat{\Theta} - \Theta^*, \widehat{\Delta}^\alpha = \hat{\alpha} - \alpha^*, \widehat{\Delta}^L = \widehat{L} - L^*, \widehat{\Delta}^S = \widehat{S} - S^*$. Throughout the discussion, let $P^* = \left\{\frac{\exp(\Theta^*_{ij})}{1+\exp(\Theta^*_{ij})}\right\}_{1\le i,j\le n} \in \mathbb{R}^{n\times n}$. Astute readers might notice the slight difference of first term in objective function between (10) and (15). Justification of this modification is well explained in *remark 3.1* of the paper [15]. We further impose several assumptions for theoretical guarantees of our estimator.

Assumption 1. (**Strong convexity**) For any $\Theta \in \mathbb{R}^{n\times n}$, define the log-likelihood in (15), $h(\Theta) = -\frac{1}{n}\sum_{i,j}\left\{X_{ij}\Theta_{ij} - \log(1+\exp(\Theta_{ij}))\right\}$.We assume that $h(\Theta)$ is τ-strongly convex in a sense that lowest eigenvalue of Hessian matrix of likelihood function is bounded away from zero ($\tau > 0$):

$$\nabla^2 h(\Theta) = diag\Big(vec\Big(\frac{\exp(\Theta)}{n(1+\exp(\Theta))^2}\Big)\Big) \succcurlyeq \tau I_{n^2\times n^2}$$

For any vector a, diag(a) is the diagonal matrix with elements of a on its diagonal. For any square matrix A and B, $A \succcurlyeq B$ if and only if A-B is positive semi-definite.

Assumption 2. (**Identifiability of $\alpha\mathbb{1}\mathbb{1}^T$ and L, Spikiness of L**) To ensure the identifiability of $\alpha\mathbb{1}\mathbb{1}^T$ and L, we assume the latent variables are centered, that is $JL = L$, where $J = I_n - \frac{1}{n}\mathbb{1}\mathbb{1}^T$, $\mathbb{1}$ denotes all one vector in $\mathbb{R}^n$. We impose a spikiness condition $\|L\|_\infty \le \frac{\kappa}{\sqrt{n\times n}}$ on L, to ensure the separation of L and S matrix [1]. We would also like to note that the constraint $|\alpha| \le C\kappa$, for an absolute constant C, is included partially for obtaining theoretical guarantees.

With these assumptions we present the behavior of non-asymptotic error bound of our estimator through following theorem. In our result, we measure error using squared Frobenius norm summed across three matrices:

$$e^2(\hat{\alpha}\mathbb{1}\mathbb{1}^T, \widehat{L}, \widehat{S}) := \|\widehat{\Delta}^\alpha\mathbb{1}\mathbb{1}^T\|_F^2 + \|\widehat{\Delta}^L\|_F^2 + \|\widehat{\Delta}^S\|_F^2$$

Theorem 2. Under the assumptions 1 and 2, if we solve the convex problem (15) with a pair of regularization parameter (δ, γ) satisfying

$$\delta \ge 2\|\frac{1}{n}(X - P^*)\|_{op} \quad and \quad \gamma \ge 2\|\frac{1}{n}(X - P^*)\|_\infty + 4\kappa\tau\left(\frac{Cn+1}{n}\right) \tag{16}$$

There exist universal constants c_j, j = 1,2,3, for all integers $k = 1, 2, ..., n$ and $s = 1, 2, ..., n^2$, we have following upper bound on $e^2(\hat{\alpha}\mathbb{1}\mathbb{1}^T, \widehat{L}, \widehat{S})$

$$e^2(\hat{\alpha}\mathbb{1}\mathbb{1}^T, \widehat{L}, \widehat{S}) \le \underbrace{c_1\frac{\delta^2}{\tau^2}}_{\mathcal{K}_{\alpha^*}} \underbrace{+c_2\frac{\delta^2}{\tau^2}\left\{k + \frac{\tau}{\delta}\sum_{j=k+1}^{n}\sigma_j(L^*)\right\}}_{\mathcal{K}_{L^*}} \underbrace{+c_3\frac{\gamma^2}{\tau^2}\left\{s + \frac{\tau}{\gamma}\sum_{(i,j)\notin\mathbb{M}}|S^*_{ij}|\right\}}_{\mathcal{K}_{S^*}} \tag{17}$$

where $\mathbb{M}$ is an arbitrary subset of matrix indices of cardinality at most s.

Our result provides the upper-bound on non-asymptotic Frobenius norm between estimator $\hat{\alpha}\mathbb{1}\mathbb{1}^T$, $\widehat{L}$ and $\widehat{S}$, and ground truth $\alpha^*\mathbb{1}\mathbb{1}^T$, L^* and S^*, where matrices L^* that can be either approximately or exactly low rank, and matrices S^* that also can be either approximately or exactly sparse. A close observation on terms in upper-bound reveals that error bounds associated with $\|\widehat{\Delta}^\alpha\mathbb{1}\mathbb{1}^T\|_F^2$, $\|\widehat{\Delta}^L\|_F^2$, and $\|\widehat{\Delta}^S\|_F^2$ are represented in the form of $\mathcal{K}_{\alpha^*}$, $\mathcal{K}_{L^*}$, and $\mathcal{K}_{S^*}$ respectively. More specifically, terms $\mathcal{K}_{L^*}$ and $\mathcal{K}_{S^*}$ are comprised of two types of errors, *Estimation error* and *Approximation error*. "*Estimation error*" represents statistical cost of estimating a parameter belong to the model subspace $\mathbb{M}$

This result helps us to understand the behavior of our proposed estimator both in practice and in ideal cases. In real-world, large scale citation network analysis, it is difficult to expect exactly $k(\ll n)$ number of topics embedded in network. There must be

Specifically, when L^* is an exact low rank matrix with rank k (i.e., $rank(L^*) = k$) and S^* is a sparse matrix, which lies within model subspace $\mathbb{M}$ (i.e., $|supp(S^*)| = s$), then we have Frobenius error bound as follows:

$$e^2(\hat{\alpha}\mathbb{1}\mathbb{1}^T, \widehat{L}, \widehat{S}) \lesssim (\delta^2 + 1)k + \gamma^2 s$$

Comparison to Agarwal et al, [1] We would like to refer readers that result presented in **Theorem 2** is similar with that of Theorem 1 presented in paper [1]. However, it should be noted that the convex program we consider is distinct from [1]. [1] considers matrix decomposition problem, where the observations are noisy realization of linear transformation Ξ of the sum of (approximately) low rank and (approximately) sparse matrices. They used the notions of Restricted Strong Convexity of quadratic loss function and decomposable regularizer to prove their claims. In our work, we consider a convex problem, whose logistic loss For the completeness of paper, we provide the proof of theorem in appendix section.

8 | APPENDIX

Proof. Since $\widehat{\Theta}$ and Θ^* are optimal and feasible (respectively) for the convex program (15), we have

$$h(\widehat{\Theta}) + \delta\|\widehat{L}\|_* + \gamma\|\widehat{S}\|_1 \leq h(\Theta^*) + \delta\|L^*\|_* + \gamma\|S^*\|_1 \tag{18}$$

Through the assumption of strong convexity on $h(\Theta)$, and by the Taylor expansion, we can get a following lower bound on term $h(\widehat{\Theta}) - h(\Theta^*)$:

$$h(\widehat{\Theta}) - h(\Theta^*) \geq \langle \nabla_\Theta h(\Theta^*), \widehat{\Theta} - \Theta^* \rangle + \frac{\tau}{2}\|\widehat{\Delta}^\Theta\|_F^2$$

By rearranging the term in (18) and plugging in above inequality relation, we get:

$$\frac{\tau}{2}\|\widehat{\Delta}^\Theta\|_F^2 \leq -\langle \nabla_\Theta h(\Theta^*), \widehat{\Theta} - \Theta^* \rangle + \delta\|L^*\|_* + \gamma\|S^*\|_1 - \delta\|\widehat{L}\|_* - \gamma\|\widehat{S}\|_1 \tag{19}$$

Here, we introduce another notation, for any pair of positive tuning parameters (δ, γ) which is defined as the weighted combination of the two regularizers :

$$\mathbb{Q}(L, S) := \|L\|_* + \frac{\gamma}{\delta}\|S\|_1$$

Through the definition of $\mathbb{Q}$, we can rewrite (19) as follows:

$$\frac{\tau}{2}\|\widehat{\Delta}^\Theta\|_F^2 \leq -\langle \nabla_\Theta h(\Theta^*), \widehat{\Theta} - \Theta^* \rangle + \delta\mathbb{Q}(L^*, S^*) - \delta\mathbb{Q}(\widehat{\Delta}^L + L^*, \widehat{\Delta}^S + S^*) \tag{20}$$

According to Agarwal et al [1]'s second element of lemma 1, the difference $\mathbb{Q}(L^*, S^*) - \mathbb{Q}(\widehat{\Delta}^L + L^*, \widehat{\Delta}^S + S^*)$ is upper-bounded by

$$\mathbb{Q}(\widehat{\Delta}_A^L, \widehat{\Delta}_M^S) - \mathbb{Q}(\widehat{\Delta}_B^L, \widehat{\Delta}_{M^\perp}^S) + 2\sum_{j=k+1}^{n} \sigma_j(L^*) + 2\frac{\gamma}{\delta}\|S_{M^\perp}^*\|_1 \tag{21}$$

First, we want to control upper bound of the term $-\langle \nabla_\Theta h(\Theta^*), \widehat{\Theta} - \Theta^* \rangle$ in (20).

$$\begin{aligned}
-\langle \nabla_\Theta h(\Theta^*), \widehat{\Theta} - \Theta^* \rangle &= \langle \frac{1}{n}(X - P^*), \widehat{\Delta}^{\alpha\mathbb{1}\mathbb{1}^T} + \widehat{\Delta}^L + \widehat{\Delta}^S \rangle \\
&\leq \frac{1}{n}\|X - P^*\|_{op}(\|\widehat{\Delta}^\alpha\mathbb{1}\mathbb{1}^T\|_* + \|\widehat{\Delta}^L\|_*) + \frac{1}{n}\|X - P^*\|_\infty\|\widehat{\Delta}^S\|_1 \\
&\leq \frac{1}{n}\|X - P^*\|_{op}(\|\widehat{\Delta}^\alpha\mathbb{1}\mathbb{1}^T\|_F + \|\widehat{\Delta}_A^L\|_* + \|\widehat{\Delta}_B^L\|_*) + \frac{1}{n}\|X - P^*\|_\infty(\|\widehat{\Delta}_M^S\|_1 + \|\widehat{\Delta}_{M^\perp}^S\|_1) \\
&\leq \frac{\delta}{2}(\|\widehat{\Delta}^\alpha\mathbb{1}\mathbb{1}^T\|_F + \|\widehat{\Delta}_A^L\|_* + \|\widehat{\Delta}_B^L\|_*) + \frac{\gamma}{2}(\|\widehat{\Delta}_M^S\|_1 + \|\widehat{\Delta}_{M^\perp}^S\|_1)
\end{aligned} \tag{22}$$

Combining the inequalities (21) and (22), we can obtain the upper bound of RHS in (20) as follows:

$$\frac{\tau}{2}\|\widehat{\Delta}^\Theta\|_F^2 \leq \frac{\delta}{2}\|\widehat{\Delta}^\alpha\mathbb{1}\mathbb{1}^T\|_F + \frac{3\delta}{2}\mathbb{Q}(\widehat{\Delta}_A^L, \widehat{\Delta}_M^S) + 2\delta\sum_{j=k+1}^{n}\sigma_j(L^*) + 2\gamma\|S_{M^\perp}^*\|_1 \tag{23}$$

Second, we wish to control the lower bound of the term $\frac{\tau}{2}\|\widehat{\Delta}^\Theta\|_F^2$ with respect to $\hat{\Delta}^\alpha, \hat{\Delta}^L, \hat{\Delta}^S$.

$$\begin{aligned}
\|\widehat{\Delta}^\Theta\|_F^2 &= \|\widehat{\Theta} - \Theta^*\|_F^2 \\
&= \|\widehat{\Delta}^{\alpha\mathbb{1}\mathbb{1}^T} + \widehat{\Delta}^L + \widehat{\Delta}^S\|_F^2 \\
&= \|\widehat{\Delta}^\alpha\mathbb{1}\mathbb{1}^T\|_F^2 + \|\widehat{\Delta}^L + \widehat{\Delta}^S\|_F^2 + 2\langle \widehat{\Delta}^L + \widehat{\Delta}^S, \widehat{\Delta}^\alpha\mathbb{1}\mathbb{1}^T \rangle \\
&= \|\widehat{\Delta}^\alpha\mathbb{1}\mathbb{1}^T\|_F^2 + \|\widehat{\Delta}^L\|_F^2 + \|\widehat{\Delta}^S\|_F^2 + 2\langle \widehat{\Delta}^L + \widehat{\Delta}^S, \widehat{\Delta}^\alpha\mathbb{1}\mathbb{1}^T \rangle + 2\langle \widehat{\Delta}^L, \widehat{\Delta}^S \rangle
\end{aligned} \tag{24}$$

We want to get the further lower bound on trace inner product terms, $\langle \widehat{\Delta}^L + \widehat{\Delta}^S, \widehat{\Delta}^\alpha\mathbb{1}\mathbb{1}^T \rangle$, $\langle \widehat{\Delta}^L, \widehat{\Delta}^S \rangle$. To control the first trace inner product term, we use the relation $\widehat{\Delta}^L\mathbb{1} = 0$, apply the definition of dual norm on inner product term, triangular inequality on $\widehat{\Delta}^\alpha$, and lastly we apply the constraint imposed on $|\alpha|$ stated in Assumption 2.

$$\begin{aligned}
|\langle \widehat{\Delta}^L + \widehat{\Delta}^S, \widehat{\Delta}^\alpha \mathbb{1}\mathbb{1}^T \rangle| &= |\langle \widehat{\Delta}^S, \widehat{\Delta}^\alpha \mathbb{1}\mathbb{1}^T \rangle| \\
&\le \|\widehat{\Delta}^\alpha \mathbb{1}\mathbb{1}^T\|_\infty \|\widehat{\Delta}^S\|_1 \\
&\le (|\hat{\alpha}| + |\alpha^*|)\|\widehat{\Delta}^S\|_1 \\
&\le 2C\kappa \|\widehat{\Delta}^S\|_1 \qquad (25)
\end{aligned}$$

To control the term $\langle \widehat{\Delta}^L, \widehat{\Delta}^S \rangle$, we first apply the definition of dual norm on trace inner product term, then apply triangular inequality on $\widehat{\Delta}^L$ and spikiness condition.

$$\begin{aligned}
|\langle \widehat{\Delta}^L, \widehat{\Delta}^S \rangle| &\le \|\widehat{\Delta}^L\|_\infty \|\widehat{\Delta}^S\|_1 \\
&\le (\|\widehat{L}\|_\infty + \|L^*\|_\infty)\|\widehat{\Delta}^S\|_1 \\
&\le \left(\frac{2\kappa}{n}\right)\|\widehat{\Delta}^S\|_1 \qquad (26)
\end{aligned}$$

We can combine the inequality (24), (25) and (26). Then applying the assumption on regularization parameter γ, and the fact $\|\widehat{\Delta}^L\|_* \ge 0$ sequentially, we can get,

$$\begin{aligned}
\frac{\tau}{2}\|\widehat{\Delta}^\Theta\|_F^2 &\ge \frac{\tau}{2}\|\widehat{\Delta}^\alpha \mathbb{1}\mathbb{1}^T\|_F^2 + \frac{\tau}{2}\|\widehat{\Delta}^L\|_F^2 + \frac{\tau}{2}\|\widehat{\Delta}^S\|_F^2 - 2\kappa\tau\left(\frac{Cn+1}{n}\right)\|\widehat{\Delta}^S\|_1 \\
&\ge \frac{\tau}{2}\|\widehat{\Delta}^\alpha \mathbb{1}\mathbb{1}^T\|_F^2 + \frac{\tau}{2}\|\widehat{\Delta}^L\|_F^2 + \frac{\tau}{2}\|\widehat{\Delta}^S\|_F^2 - \frac{\gamma}{2}\|\widehat{\Delta}^S\|_1 \\
&\ge \frac{\tau}{2}\|\widehat{\Delta}^\alpha \mathbb{1}\mathbb{1}^T\|_F^2 + \frac{\tau}{2}\|\widehat{\Delta}^L\|_F^2 + \frac{\tau}{2}\|\widehat{\Delta}^S\|_F^2 - \frac{\delta}{2}\mathbb{Q}(\widehat{\Delta}^L, \widehat{\Delta}^S) \qquad (27)
\end{aligned}$$

By combining the relations (23) and (27), applying triangular inequality, $\mathbb{Q}(\widehat{\Delta}^L, \widehat{\Delta}^S) \le \mathbb{Q}(\widehat{\Delta}^L_A, \widehat{\Delta}^S_M) + \mathbb{Q}(\widehat{\Delta}^L_B, \widehat{\Delta}^S_{M^\perp})$, and rearranging the term, we can get following inequality,

$$\frac{\tau}{2}\|\widehat{\Delta}^\alpha \mathbb{1}\mathbb{1}^T\|_F^2 + \frac{\tau}{2}\|\widehat{\Delta}^L\|_F^2 + \frac{\tau}{2}\|\widehat{\Delta}^S\|_F^2 \le \frac{\delta}{2}\|\widehat{\Delta}^\alpha \mathbb{1}\mathbb{1}^T\|_F + 2\mathbb{Q}(\widehat{\Delta}^L_A, \widehat{\Delta}^S_M) + \frac{\delta}{2}\mathbb{Q}(\widehat{\Delta}^L_B, \widehat{\Delta}^S_{M^\perp}) + 2\delta\sum_{j=k+1}^{n}\sigma_j(L^*) + 2\gamma\|S^*_{M^\perp}\|_1$$

Further, by plugging in Lemma 1 to get an upper bound on $\mathbb{Q}(\widehat{\Delta}^L_B, \widehat{\Delta}^S_{M^\perp})$, we can rewrite the above inequality as follows:

$$\frac{\tau}{2}\|\widehat{\Delta}^\alpha \mathbb{1}\mathbb{1}^T\|_F^2 + \frac{\tau}{2}\|\widehat{\Delta}^L\|_F^2 + \frac{\tau}{2}\|\widehat{\Delta}^S\|_F^2 - \frac{\delta}{2}\|\widehat{\Delta}^\alpha \mathbb{1}\mathbb{1}^T\|_F \le \frac{7\delta}{2}\mathbb{Q}(\widehat{\Delta}^L_A, \widehat{\Delta}^S_M) + 4\delta\sum_{j=k+1}^{n}\sigma_j(L^*) + 4\gamma\|S^*_{M^\perp}\|_1 \qquad (28)$$

Noting that $\widehat{\Delta}^L_A$ has rank at most $2k$ and that $\widehat{\Delta}^S_M$ lies in the model space $\mathbb{M}$, we find that

$$\begin{aligned}
\delta\mathbb{Q}(\widehat{\Delta}^L_A, \widehat{\Delta}^S_M) &\le \sqrt{2k}\delta\|\widehat{\Delta}^L_A\|_F + \Psi(\mathbb{M})\gamma\|\widehat{\Delta}^S_{\mathbb{M}}\|_F \\
&\le \sqrt{2k}\delta\|\widehat{\Delta}^L\|_F + \Psi(\mathbb{M})\gamma\|\widehat{\Delta}^S\|_F \qquad (29)
\end{aligned}$$

Here $\Psi(\mathbb{M})$ measures the compatibility between Frobenius norm and component-wise ℓ_1 regularizer, where $\mathbb{M}$ is an arbitrary subset of matrix indices of cardinality at most s.

$$\Psi(\mathbb{M}) := \sup_{U\in\mathbb{M}, U\neq 0} \frac{\|U\|_1}{\|U\|_F}$$

Using Cauchy-Schwarz inequality, we can easily check the quantity $\Psi(\mathbb{M})$ is bounded by at most $\sqrt{s}$. Plugging in the relation (29) into (28) and rearranging the term relevant with $e^2(\hat{\alpha}\mathbb{1}\mathbb{1}^T, \widehat{L}, \widehat{S})$ yield the claim. □

8.1 | Proof of Lemma 1

Proof. Through the application of basic inequality by using optimality of $\widehat{\Theta}$ and feasibility of Θ^* to convex program (15), we have

$$h(\widehat{\Theta}) - h(\Theta^*) \le \delta\mathbb{Q}(L^*, S^*) - \delta\mathbb{Q}(\widehat{\Delta}^L + L^*, \widehat{\Delta}^S + S^*) \qquad (30)$$

By using convexity of $h(\Theta)$, we can write

$$\begin{aligned} h(\widehat{\Theta}) - h(\Theta^*) &\geq \langle \nabla_\Theta h(\Theta^*), \widehat{\Theta} - \Theta^* \rangle \\ &= -\langle \frac{1}{n}(X - P^*), \hat{\Delta}^\alpha \mathbb{1}\mathbb{1}^T + \widehat{\Delta}^L + \widehat{\Delta}^S \rangle \\ &\geq -\frac{1}{n}\|X - P^*\|_{op}(\|\hat{\Delta}^\alpha \mathbb{1}\mathbb{1}^T\|_* + \|\widehat{\Delta}^L\|_*) + \frac{1}{n}\|X - P^*\|_\infty \|\widehat{\Delta}^S\|_1 \\ &\geq -\frac{\delta}{2}(\|\hat{\Delta}^\alpha \mathbb{1}\mathbb{1}^T\|_F + \|\widehat{\Delta}^L_A\|_* + \|\widehat{\Delta}^L_B\|_*) - \frac{\delta}{2}(\|\widehat{\Delta}^S_{\mathbb{M}}\|_1 + \|\widehat{\Delta}^S_{\mathbb{M}^\perp}\|_1) \end{aligned} \tag{31}$$

One more round of application on Agarwal et al [1]'s second element of lemma 1, we can get an upper bound of difference $\mathbb{Q}(L^*, S^*) - \mathbb{Q}(\widehat{\Delta}^L + L^*, \widehat{\Delta}^S + S^*)$ as follows:

$$\mathbb{Q}(\widehat{\Delta}^L_A, \widehat{\Delta}^S_M) - \mathbb{Q}(\widehat{\Delta}^L_B, \widehat{\Delta}^S_{M^\perp}) + 2\sum_{j=k+1}^{n} \sigma_j(L^*) + 2\frac{\gamma}{\delta}\|S^*_{M^\perp}\|_1 \tag{32}$$

By combining relations (31) and (32), we can get the upper bound of $\mathbb{Q}(\widehat{\Delta}^L_B, \widehat{\Delta}^S_{M^\perp})$:

$$\mathbb{Q}(\widehat{\Delta}^L_B, \widehat{\Delta}^S_{M^\perp}) \leq \|\hat{\Delta}^\alpha \mathbb{1}\mathbb{1}^T\|_F + 3\mathbb{Q}(\widehat{\Delta}^L_A, \widehat{\Delta}^S_M) + 4\sum_{j=k+1}^{n} \sigma_j(L^*) + 4\frac{\gamma}{\delta}\|S^*_{M^\perp}\|_1$$

$\square$

Pair	Weight	Community	Title
1	6.03	FuncAn	Properties of principal component methods for functional and longitudinal data analysis
		VarSel	Nonconcave penalized likelihood with a diverging number of parameters
2	3.08	MulT	Innovated higher criticism for detecting sparse signals in correlated noise
		VarSel	Regularized estimation of large covariance matrices
3	2.47	DimRed	Contour projected dimension reduction
		VarSel	Factor profiled sure independence screening
4	2.15	VarSel	Factor profiled sure independence screening
		DimRed	Sliced regression for dimension reduction
5	1.86	VarSel	Covariance regularization by thresholding
		MulT	Adapting to unknown sparsity by controlling the false discovery rate
6	1.77	VarSel	Asymptotic properties of bridge estimators in sparse high-dimensional regression models
		Mixed	Marginal asymptotics for the ″large p small n″ paradigm: with applications to microarray data
7	1.71	VarSel	A majorization-minimization approach to Variable Selection using spike and slab priors
		Mixed	Empirical Bayes selection of wavelet thresholds
8	1.66	VarSel	Spades and mixture models
		MulT	Adapting to unknown sparsity by controlling the false discovery rate
9	1.61	DimRed	A constructive approach to the estimation of dimension reduction directions
		VarSel	Factor profiled sure independence screening
10	1.54	VarSel	A majorization-minimization approach to variable selection using spike and slab priors
		Mixed	Spike and slab variable selection: frequentist and Bayesian strategies
11	1.53	VarSel	variable selection in nonparametric additive models
		Mixed	Nonparametric estimation of an additive model with a link function
12	1.39	Mixed	Nonparametric inferences for additive models
		VarSel	Nonparametric independence screening in sparse ultra-high-dimensional additive models
13	1.29	Mixed	Bayesian variable selection in structured high-dimensional covariate spaces with applications in genomics
		VarSel	Model selection and estimation in regression with grouped variables
14	1.20	VarSel	The sparsity and bias of the LASSO selection in high-dimensional linear regression
		Mixed	Tests for high-dimensional covariance matrices
15	1.19	FuncAn	Empirical dynamics for longitudinal data
		Mixed	Variable selection in nonparametric varying-coefficient models for analysis of repeated measurements

References

[1] Agarwal, A., S. Negahban, M. J. Wainwright, et al., 2012: Noisy matrix decomposition via convex relaxation: Optimal rates in high dimensions. *The Annals of Statistics*, **40**, no. 2, 1171–1197.

[2] Bach, F. R., 2008: Consistency of trace norm minimization. *Journal of Machine Learning Research*, **9**, no. Jun, 1019–1048.

[3] Boyd, S., N. Parikh, E. Chu, B. Peleato, J. Eckstein, et al., 2011: Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning*, **3**, no. 1, 1–122.

[4] Candès, E. J., X. Li, Y. Ma, and J. Wright, 2011: Robust principal component analysis? *Journal of the ACM (JACM)*, **58**, no. 3, 11.

[5] Chandrasekaran, V., P. A. Parrilo, and A. S. Willsky, 2010: Latent variable graphical model selection via convex optimization. *2010 48th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, IEEE, 1610–1613.

[6] Chen, Y., X. Li, J. Liu, and Z. Ying, 2016: A fused latent and graphical model for multivariate binary data. *arXiv preprint arXiv:1606.08925*.

[7] Fazel, M., H. Hindi, S. P. Boyd, et al., 2001: A rank minimization heuristic with application to minimum order system approximation. *Proceedings of the American control conference*, Citeseer, volume 6, 4734–4739.

[8] Gabay, D. and B. Mercier, 1975: *A dual algorithm for the solution of non linear variational problems via finite element approximation*. Institut de recherche d'informatique et d'automatique.

[9] Glowinski, R. and A. Marrocco, 1975: On the solution of a class of nonlinear Dirichlet problems by a penalty-duality method and finite elements of order one. *Optimization Techniques IFIP Technical Conference*, Springer, 327–333.

[10] Harman, H. H., 1960: *Modern factor analysis*. Univ. of Chicago Press.

[11] Ji, P., J. Jin, et al., 2016: Coauthorship and citation networks for statisticians. *The Annals of Applied Statistics*, **10**, no. 4, 1779–1812.

[12] Jöreskog, K. G., 1969: A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika*, **34**, no. 2, 183–202.

[13] — 1970: A general method for estimating a linear structural equation system. *ETS Research Bulletin Series*, **1970**, no. 2, i–41.

[14] Lord, F. M. and M. R. Novick, 2008: *Statistical theories of mental test scores*. IAP.

[15] Ma, Z. and Z. Ma, 2017: Exploration of large networks with covariates via fast and universal latent space model fitting. *arXiv preprint arXiv:1705.02372*.

[16] McDonald, R. P., 2014: *Factor analysis and related methods*. Psychology Press.

[17] Rasch, G., 1980: Probabilistic models for some intelligence and attainment tests. 1960. *Copenhagen, Denmark: Danish Institute for Educational Research*.

[18] Schwarz, G. et al., 1978: Estimating the dimension of a model. *The annals of statistics*, **6**, no. 2, 461–464.

[19] Zhou, Z., X. Li, J. Wright, E. Candes, and Y. Ma, 2010: Stable principal component pursuit. *2010 IEEE international symposium on information theory*, IEEE, 1518–1522.

How to cite this article: Suh Namjoon., Xiaoming Huo, Eric Heim, Timothy Van Slyke, and Lee Seversky (2019), xxx, *journal., 2019;00:1–6*.