ARTICLE TYPE

Factor Analysis on Citations, Using a Combined Latent and Logistic Regression Model

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

[1]School of Industrial and Systems Engineering, Georgia Institute of Technology, 755 Ferst Dr, Atlanta, GA, USA

[2]Information System Division (AFRL/RIS), the Air Force Research Laboratory, Department of the Air Force, Air Force Materiel Command, AFRL/RIK - Rome, 26 Electronic Parkway, Rome, New York, USA

Correspondence
*Xiaoming Huo, Georgia Institute of Technology, Email: huo@gatech.edu

Present Address
755 Ferst Dr, Atlanta, GA, USA.

Summary

We propose a combined model, which integrates the latent factor model and the logistic regression model, for the citation network. It is noticed that neither a latent factor model nor a logistic regression model alone is sufficient to capture the structure of the data. The proposed model has a latent (i.e., factor analysis) model to represents the main technological trends (a.k.a., factors), and adds a sparse component that captures the remaining ad-hoc dependence. Parameter estimation are carried out through the construction of a joint-likelihood function of edges and properly chosen penalty terms. The convexity of the objective function allows us to develop an efficient algorithm, while the penalty terms push towards a low-dimensional latent component and a sparse graphical structure. Simulation results show that the proposed method works well in practical situations. The proposed method has been applied to a real application, which contains a citation network of statisticians (Ji and Jin, 2016 [11]). Some interesting findings are reported.

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

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 $1 \leq i, j \leq n$,

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

---

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

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

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

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 $(1 \le k \le K)$ for factor k, despite different nodes. We denote a matrix $F = \{f_1, f_2, \ldots, f_n\} \in \mathbb{R}^{K \times n}$. Each column i in matrix F contains the factor loadings associated with the node i $(1 \le i \le n)$. 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:

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

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

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

where we assume $X_{ii} = 0$ 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. The above delivers a factor analysis model. Various linear and nonlinear latent variable models have been studied extensively in the literature (e.g.,[12, 15, 14, 16, 10, 13]).

Our work is motivated from a recent work named *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. In our model, we also assume that there exist two dependent structures among citation edges in a 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 weight coefficients d_i's). There is still a small remainder due to the sparse graphical component.

Though it may seem similar to Chen et al [6], we work on a different model formulation in several aspects. We summarize the differences as follows.

1. 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 a random graph.

2. In FLaG model, a collection of binary responses for each question in the questionnaire follows a joint distribution, which is a combination of the Item Response Theory (IRT) model and the Ising model. We model the citation edges among papers as random variables, whose dependent structure is characterized by the combination of the Latent Factor Analysis model and the Sparse Graphical model.

3. FLaG approximates the original likelihood through constructing pseudo-likelihood function by taking advantage of conditional independence among the nodes. In our model, 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 ([1, 4, 5, 18]). Specifically, [5] studies statistical inference of a multivariate Gaussian model whose precision matrix admits the form of a low-rank matrix plus a sparse matrix. The inference and optimization of the current model are different from the aforementioned cases. We will construct a regularized-likelihood function, based on which estimator will be proposed for simultaneous model selection and parameter estimation. The objective function in the optimization problem for

the regularized estimator is 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 the latent dependent structure due to the common topics and the remaining sparse ad-hoc dependent structure. In Section 3, we will discuss the assumptions in our model and penalization on likelihood function that is constructed in Section 2. In Section 4, we provide a non-asymptotic error bound of the estimator. Section 5 gives the detailed procedure on how to compute the estimator of the optimization problem that is formulated in Section 3. In Section 6, 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 7 with several open questions 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):

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

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 the 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 as follows. 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. Note that the matrix S is introduced to capture the ad-hoc links in the graph. If we have $S_{ij} \le 0$, 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:

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

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:

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

3 | ESTIMATION

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

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

If we consider maximizing the above log-likelihood function, we will encounter several technical issues that 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 $F^T D F$. 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

$$F^T D F = F^T P^T P D P^T P F = (F')^T D' F',$$

   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 $F^T D F$ 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 an arbitrary matrix L, we assume that the eigen-vector of L is centered, that is,

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

   where $\mathbb{1}$ denotes a n-dimensional vector whose entries are all 1's. If we have $L = F^T D F$, 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 $F^T D F$ to L, where L is a low rank matrix. Consequently, the log-likelihood function in (8) can be rewritten as

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

We propose a penalized 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 $L = F^T D F$ 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 mimics 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 $\hat{L} = F^T D F$. This is the price we have to pay for an amenable computational approach. The rank of estimated $\hat{L}$ will be our

estimate of the number of factors (i.e., the number of underlying common topics). For the issue on assigning the community membership of each node i, we will discuss this later in Section 6.

4 | 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 number of papers in a network is explicitly tracked. We are interested in solving the following optimization problem :

$$\min_{\substack{\alpha \in R, S = S^T \\ L \geq 0}} -\frac{1}{n} \log \prod_{1 \leq i,j \leq n} \frac{\exp\left(X_{ij}\left(\alpha + L_{ij} + S_{ij}\right)\right)}{1 + \exp\left(\alpha + L_{ij} + S_{ij}\right)} + \delta\|L\|_* + \gamma\|S\|_1 \tag{11}$$

For the convenience of theoretical investigation, we slightly modify the first term in the objective function summing over all (i, j) pairs. After scaling, due to symmetry of $X, L,$ and S, the only difference between (10) and (11) is in the inclusion of terms in diagonal pairs $(i, i), \forall i = 1, \ldots, n$. This slight modification leads to neither theoretical consequence nor noticeable difference in practice. Let $(\widehat{\alpha}, \widehat{L}, \widehat{S})$ be the solution to (11), and (α^*, L^*, S^*) be the ground truth, which governs the data generating process. Let $\widehat{\Theta}$ and Θ^* be defined respectively as $\widehat{\Theta} = \widehat{\alpha} \mathbb{1}\mathbb{1}^T + \widehat{L} + \widehat{S}$ and $\Theta^* = \alpha^* \mathbb{1}\mathbb{1}^T + L^* + S^*$. And denote the error term for each parameter as $\widehat{\Delta}^\Theta = \widehat{\Theta} - \Theta^*, \widehat{\Delta}^\alpha = \widehat{\alpha} - \alpha^*, \widehat{\Delta}^L = \widehat{L} - L^*, \widehat{\Delta}^S = \widehat{S} - S^*$. Throughout the discussion, let $P^* = \left\{\frac{\exp(\Theta^*_{ij})}{1+\exp(\Theta^*_{ij})}\right\}_{1 \leq i,j \leq n} \in \mathbb{R}^{n \times n}$. We then 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 (11), $h(\Theta) = -\frac{1}{n}\sum_{i,j}\left\{X_{ij}\Theta_{ij} - \log(1 + \exp(\Theta_{ij}))\right\}$. We assume that $h(\Theta)$ is τ-strongly convex in a sense that lowest eigenvalue of Hessian matrix of the log-likelihood function is bounded away from zero ($\tau > 0$):

$$\nabla^2 h(\Theta) = \text{diag}\left(\text{vec}\left(\frac{1}{n}\frac{\exp(\Theta)}{(1+\exp(\Theta))^2}\right)\right) \geq \tau I_{n^2 \times n^2}$$

For any vector a, diag(a) is the diagonal matrix with elements of a on its diagonal. For any matrix $B = [b_1, \ldots, b_n] \in \mathbb{R}^{n \times n}$, vec($B$) $\in \mathbb{R}^{n^2}$ is obtained by stacking $b_1, \ldots, b_n$ in order. For any square matrix A and B, we have $A \geq B$ if and only if matrix $A - B$ 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 that the latent variables are centered, that is $JL = L$, where $J = I_n - \frac{1}{n}\mathbb{1}\mathbb{1}^T$, where $\mathbb{1}$ denotes all one vector in $\mathbb{R}^n$. We impose a spikiness condition $\|L\|_\infty \leq \frac{\kappa}{\sqrt{n \times n}}$ on L, to ensure the separation of L and S matrix [1]. We would also like to note that the constraint $|\alpha| \leq C\kappa$, 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 the following theorem. In our result, we measure error using squared Frobenius norm summed across three matrices:

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

Theorem 1. Under the assumptions 1 and 2, if we solve the convex problem (11) with a pair of regularization parameter (δ, γ) satisfying

$$\delta \geq 2\left\|\frac{1}{n}(X - P^*)\right\|_{op} \quad and \quad \gamma \geq 2\left\|\frac{1}{n}(X - P^*)\right\|_\infty + 4\kappa\tau\left(\frac{Cn+1}{n}\right), \tag{12}$$

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

$$e^2(\widehat{\alpha}\mathbb{1}\mathbb{1}^T, \widehat{L}, \widehat{S}) \leq \underbrace{c_1\frac{\delta^2}{\tau^2}}_{\mathcal{K}_{\alpha^*}} + \underbrace{c_2\frac{\delta^2}{\tau^2}\left\{k + \frac{\tau}{\delta}\sum_{j=k+1}^n \sigma_j(L^*)\right\}}_{\mathcal{K}_{L^*}} + \underbrace{c_3\frac{\gamma^2}{\tau^2}\left\{s + \frac{\tau}{\gamma}\sum_{(i,j)\notin M}|S^*_{ij}|\right\}}_{\mathcal{K}_{S^*}}, \tag{13}$$

where M is an arbitrary subset of matrix indices of cardinality at most s.

We would first like to note that the result presented in Theorem 1 can be thought of as an extension of Theorem 1 presented in paper [1] to a generalized linear model. Specifically, our work considers a logistic loss function whose parameter is characterized

by a sparse matrix plus a low rank matrix, whereas Agarwal, et al.[1] work on a general linear observation model whose parameter is also characterized by sum of a low rank matrix and a sparse matrix.

Astute readers might have noticed that the upper bound in 13 consists of three different terms, where we denote them as $\mathcal{K}_{\alpha*}$, $\mathcal{K}_{L*}$ and $\mathcal{K}_{S*}$. Each respective term is involved with estimating three model parameters, α, L and S. To be more specific, both $\mathcal{K}_{L*}$ and $\mathcal{K}_{S*}$ have two types of error: 1) The first one is called as an "estimation error". This error represents the statistical cost of estimating parameters which belong to the model subspace. 2) Another quantity is referred as "approximation error". This error occurs when we only focus on estimating parameters within the model subspace, and it shrinks as the model subspace becomes large.

The result of the Theorem 1 provides a family of upper-bounds, one for each indexed by a specific choice of model subspace M, and rank parameter k. In other words, this means that the subset M and the target rank k can be adaptively chosen so as to obtain the tightest upper bound. In ideal case where L^* is an exact low rank matrix with rank k $\left(\text{i.e., rank}(L^*) = k\right)$ and S^* is a sparse matrix, whose support lies within the model subspace M $\left(\text{i.e., supp}(S^*) \subset M\right)$, we can easily see "*approximation error*" terms in $\mathcal{K}_{L*}$ $\left(\text{i.e., } \delta \sum_{j=k+1}^{n} \sigma_j(L^*)\right)$ and in $\mathcal{K}_{S*}$ $\left(\text{i.e., } \gamma \sum_{(i,j)\notin M} |S_{ij}^*|\right)$ disappear, giving us Frobenius error bound as follows:

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

Here we use the notation $X \lesssim Y$ to denote that there exists universal absolute constant C such that $X \leq CY$.

5 | COMPUTATION

We propose a method that takes advantage of the special structure of the L_1 and the nuclear norm 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

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

are linear in α, L, and S. The term

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

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.

5.1 | 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

$$\min_{x \in \chi_f, z \in \chi_g} \{f(x) + g(z)\}, \tag{14}$$
$$\text{subject to} \quad x = z.$$

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 as in [3]. Let p^* be the minimal value in (14).

Theorem 2 (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 (14)

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

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

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

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

$$\chi_f = \{(\alpha, M, L, S) : \alpha \in \mathbb{R}, M, L, S \in \mathbb{R}^{n \times n}, L \text{ is positive semidefinite}, S \text{ is symmetricg}\},$$

$$f(x) = -\frac{\alpha}{n} \sum_{1 \leq i < j \leq n} X_{ij} - \frac{1}{2n} X \bullet M + \frac{1}{n} \sum_{1 \leq i \leq j \leq n} \log\left(1 + e^{\alpha + M_{ij}}\right) + \gamma \|S\|_1 + \delta \|L\|_*,$$

$$\chi_g = \{(\alpha, M, L, S) : \alpha \in \mathbb{R}, M, L, S \in \mathbb{R}^{n \times n}, M \text{ is symmetric and } M = L + S\}, \text{ and}$$

$$g(x) = 0, \text{ for } x \in \chi_g.$$

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

$$x^m = (x_\alpha^m, x_M^m, x_L^m, x_S^m), \quad z^m = (z_\alpha^m, z_M^m, z_L^m, z_S^m), \quad u^m = (u_\alpha^m, u_M^m, u_L^m, u_S^m).$$

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

$$x_\alpha^{m+1}, x_M^{m+1} = \arg\min_{\alpha, M} \quad -\frac{\alpha}{n} \sum_{1 \leq i < j \leq n} X_{ij} - \frac{1}{2n} X \bullet M + \frac{1}{n} \sum_{1 \leq i \leq j \leq n} \log\left(1 + e^{\alpha + M_{ij}}\right)$$

$$+ \frac{1}{2\lambda} \left[\alpha - (z_\alpha^m - u_\alpha^m)\right]^2 + \frac{1}{2\lambda} \|M - (z_M^m - u_M^m)\|_F^2, \tag{15}$$

$$x_L^{m+1} = \arg\min_L \quad \delta \|L\|_* + \frac{1}{2\lambda} \|L - (z_L^m - u_L^m)\|_F^2, \tag{16}$$

$$\text{subject to } L \text{ is positive semidefinite};$$

$$x_S^{m+1} = \arg\min_S \quad \gamma \|S\|_1 + \frac{1}{2\lambda} \|S - (z_S^m - u_S^m)\|_F^2, \tag{17}$$

$$\text{subject to } S \text{ is symmetric},$$

where $\|\cdot\|_F$ is the matrix Frobenius norm, defined as $\|M\|_F^2 = \sum_{i,j} m_{ij}^2$ for a matrix $M = \{(m_{ij})\}_{i,j=1}^n$. The problem in (15) may not have a closed-form solution. We use a simple gradient descent to solve in this step, setting the step size equal to 0.05 and stopping criteria as $\max\left(|x_{\alpha,m}^{(t+1)} - x_{\alpha,m}^{(t)}|, \|x_{M,m}^{(t+1)} - x_{M,m}^{(t)}\|_\infty\right) \le 10^{-9}$. Note that there are close-form solutions to (16) and (17), while (15) is a unconstrained convex optimization problem. More specifically, in (16), suppose the eigenvalue decomposition of the symmetric matrix $(z_L^m - u_L^m)$ can be written as

$$z_L^m - u_L^m = T \Lambda T^T,$$

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

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

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 (17), we have, for $i \ne j$,

$$S_{ij} = \begin{cases} 0, & \text{if } |(z_S^m - u_S^m)_{ij}| < \lambda\gamma, \\ (z_S^m - u_S^m)_{ij} - \lambda\gamma, & \text{if } (z_S^m - u_S^m)_{ij} > \lambda\gamma, \\ (z_S^m - u_S^m)_{ij} + \lambda\gamma, & \text{if } (z_S^m - u_S^m)_{ij} < -\lambda\gamma. \end{cases}$$

Step 2. We solve $z^{m+1} = \mathbf{P}_{\lambda,g}(x^{m+1} + u^m)$. A close-form solution exists here. Denote $\bar{\alpha} = x_\alpha^{m+1} + u_\alpha^m$, $\bar{M} = x_M^{m+1} + u_M^m$, $\bar{L} = x_L^{m+1} + u_L^m$, and $\bar{S} = x_S^{m+1} + u_S^m$, then evaluating $\mathbf{P}_{\lambda,g}(x^{m+1} + u^m)$ becomes

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

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

$$z_\alpha^{m+1} = \bar{\alpha},$$
$$z_M^{m+1} = \frac{1}{3}\bar{M} + \frac{1}{3}\bar{M}^T + \frac{1}{3}\bar{L} + \frac{1}{3}\bar{S},$$
$$z_L^{m+1} = \frac{1}{6}\bar{M} + \frac{1}{6}\bar{M}^T + \frac{2}{3}\bar{L} - \frac{1}{3}\bar{S}, \quad \text{and}$$
$$z_S^{m+1} = \frac{1}{6}\bar{M} + \frac{1}{6}\bar{M}^T - \frac{1}{3}\bar{L} + \frac{2}{3}\bar{S}.$$

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 the mth iteration, we have $\|x_M^m - x_L^m - x_S^m\|_F \le \delta$, with $\delta = 10^{-7}$.

6 | NUMERICAL ANALYSIS AND APPLICATIONS

Section 6 is divided into two parts. In Section 6.1, we conduct an empirical study of our proposed method with artificial graphical structures. In Section 6.2, we perform a real data analysis with citation network for statisticians.

6.1 | Numerical experiments with synthetic data

First, we introduce two synthetic scenarios that we want to explore. Then, we describe three model selection criteria and four evaluation metrics for the selected model. Subsequently, we elaborate experimental results from the synthetic networks and several interesting findings from those results.

6.1.1 | Synthetic Setting

Before specifying the network settings in two scenarios that we want to explore, let us first describe a set of procedures for setting the model parameters, α^*, F^*, D^* and S^* sequentially.

1. We draw an intercept term α^* in the logistic regression model from the uniform distribution that is supported on [-11,-10]. In this way, we can make α^* have the least effects in creating edges in the network.

2. Recall that the binary factor loading matrix F^* encodes the relation between topics and papers (i.e., if ith paper studies about kth topic, we denote $F_{ki}^* = 1$, otherwise $F_{ki}^* = 0$). First, we assume that there are n papers in the network, and K topics are embedded in it. Each of them consists of $\frac{n}{K}$ papers. This can be expressed in F^* as follows:

$$
F^* = \begin{bmatrix} \underbrace{1 \cdots 1}_{n/K} & & & \\ & \underbrace{1 \cdots 1}_{n/K} & \ddots & \\ & & & \underbrace{1 \cdots 1}_{n/K} \end{bmatrix} \in R^{K \times n},
$$

where each row of F^* has $\frac{n}{K}$ 1's and each column has only one 1. Then, we assume that n_l papers share l topics and n_m papers share m topics ($1 < l < m \le K$). This can be represented in the F^* in following steps: 1) Pick distinct n_l indices randomly from $\{1, 2, \ldots, n\}$. We will denote the set of the indices as Ω_{n_l}. 2) Choose n_m indices from the set $\{1, 2, \ldots, n\} \setminus \Omega_{n_l}$ and denote the set of those indices as Ω_{n_m}. 3) Make the columns of F^* with corresponding indices in set $\Omega_{n_l} \cup \Omega_{n_m}$ zeros. 4) We use a notation f_j^* to denote the jth column of the matrix F^*. Fill arbitrary l entries of f_j^* for $j \in \Omega_{n_l}$ with 1s, and also fill arbitrary m entries of f_j^* for $j \in \Omega_{n_m}$ with 1s. Lastly, we set $F^* = J F^*$ where $J = I_n - \frac{1}{n} \mathbb{1}\mathbb{1}^T$.

3. Generate the weight coefficients of the factors D_{ii}^* from the uniform distribution that is supported on [19, 20] $\forall 1 \le i \le K$. In this way, we can let the papers with same topic cluster together.

4. Remind that the positive entries of S^* can characterize the links in the network which cannot be accounted by the common topics. In the spirit of this notion, we construct ad-hoc links in a way that they connect the clusters of papers with one topic. Specifically, we assume that there are $\frac{|S^*|}{_KC_2}$ edges between two clusters, where $|S^*|$ denotes the number of non-zero entries of the upper-triangular part of the matrix S^*. This can be implemented via following steps: 1) We construct K sets $C_1^*, \ldots, C_K^*$ that are defined as follows:

$$
C_1^* \subseteq \left\{1, 2, \ldots, \frac{n}{K}\right\} \setminus \left\{\Omega_{n_l} \cup \Omega_{n_m}\right\}
$$
$$
C_2^* \subseteq \left\{\frac{n}{K} + 1, \ldots, \frac{2n}{K}\right\} \setminus \left\{\Omega_{n_l} \cup \Omega_{n_m}\right\}
$$
$$
\vdots
$$
$$
C_K^* \subseteq \left\{\frac{(K-1)n}{K} + 1, \ldots, n\right\} \setminus \left\{\Omega_{n_l} \cup \Omega_{n_m}\right\}
$$

where each of them has arbitrary $\frac{|S^*|}{_KC_2}$ elements. 2) Create a set I_{S^*} whose elements are pairs of indices such that

$$
I_{S^*} = \left\{ (i_r, j_r) : i_r \in C_p^*, j_r \in C_q^*, 1 \le p < q \le K, r = 1, 2, \ldots, \frac{|S^*|}{_KC_2} \right\},
$$

where we use i_r to indicate rth element i in the set. 3) Draw $S_{ij}^* \sim \text{Unif}[19, 20], \forall (i, j) \in I_{S^*}$. 4) Lastly, make it symmetric by setting $S_{ji}^* = S_{ij}^*, \forall 1 \le i < j \le n$.

5. Create an upper-triangular part of the adjacency matrix X whose each entry X_{ij} follows Bernoulli distribution. The distribution's parameter is parametrized by a probability, $P_{ij}^* = \frac{\exp(\alpha^* + f_i^{*T} D^* f_j^* + S_{ij}^*)}{1 + \exp(\alpha^* + f_i^{*T} D^* f_j^* + S_{ij}^*)}$. After drawing all the entries of X in the upper-triangular part, then make the matrix symmetric by setting $X_{ji} = X_{ij}, \forall 1 \le i < j \le n$.

With these settings in mind, we consider two scenarios, where each of them has three synthetic networks.

1. In the first scenario, we consider three networks, in which each of them consists of papers with only one topic. Specifically, following three networks $\{(n^{(i)}, n_1^{(i)}, K^{(i)}, |S^*|^{(i)})\}_{i=1}^3 = \{(30, 30, 3, 9), (80, 80, 4, 18), (120, 120, 5, 30)\}$ are considered. If we take an example, the notation $(n^{(1)}, n_1^{(1)}, K^{(1)}, |S^*|^{(1)}) = (30, 30, 3, 9)$ means that we generate a network with 30 papers. There are 3 topics embedded in the network, and 9 random ad-hoc links connect 3 clusters of papers, where each cluster represents a collection of papers with same topic.

2. In the second scenario, we consider three networks, in which each of them has some papers that have more than one topic. In particular, we consider $\{(n^{(i)}, n_2^{(i)}, n_3^{(i)}, K^{(i)}, |S^*|^{(i)})\}_{i=1}^3 = \{(120, 0, 10, 3, 18), (210, 50, 0, 3, 18), (210, 10, 10, 3, 18)\}$. For instance, in the third case, we have a network with 210 papers in total. There are 3 topics commonly shared across the network. Among 210 papers, 10 papers randomly share 2 topics out of 3, other 10 papers have 3 mixed topics, whereas remaining 190 papers only discuss 1 topic. Note that the 3 clusters from these 190 papers are connected through 18 random ad-hoc links.
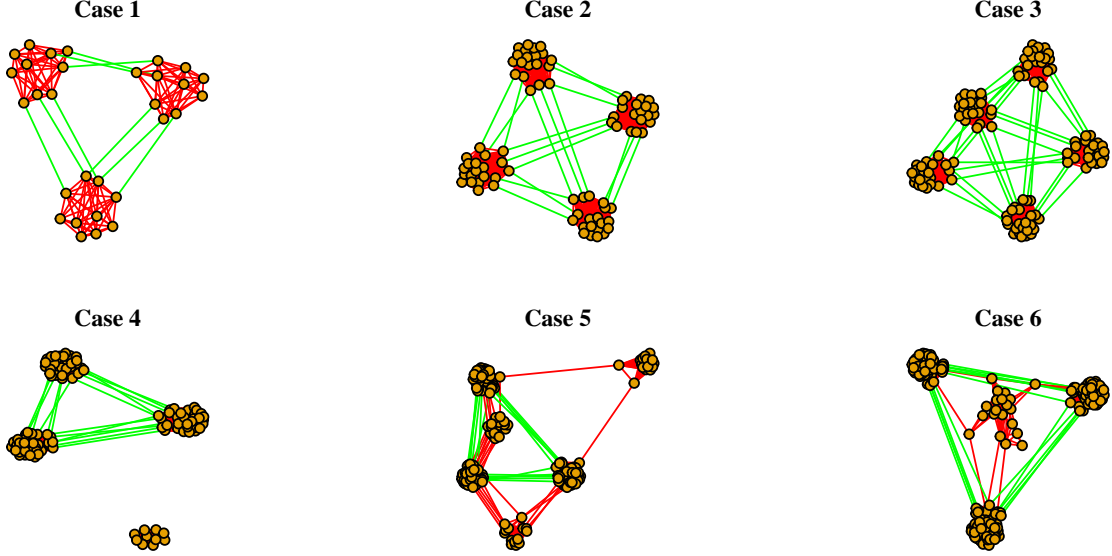

FIGURE 1 Graphical illustrations of six synthetic networks. Nodes that share the common factors are clustered. The cross cluster links are ad-hoc Citations. All the graphs are drawn via R package 'igraph'.

All six networks that are elaborated in scenarios 1 and 2 are visualized in Fig.1 . Notice that the nodes that share the common topics are clustered, and the cross clustered links are ad-hoc citations.

6.1.2 | Choosing the tuning parameters and evaluation criteria

Choosing a good pair of tuning parameters is an important yet challenging issue in our setting. Here we present a heuristic procedure for choosing a good pair of tuning parameters (γ, δ). Following the scree-plot approach in Ji and Jin [11], we plot the largest 15 eigenvalues of the adjacency matrix X, and find an "elbow" point where the eigenvalues seem to level off. An index of the point, which is to the left of this elbow point, is considered as the number of the topics embedded in the network. (We will denote this number as $\widehat{K}^{\text{Scree}}$.) We want to note the readers that the scree-plot analysis serves as a good guideline for determining the range of grids to search over. With the estimate of the number of topics in the network in mind, we record the rank$(\widehat{L}^{\gamma,\delta})$ and $|\widehat{S}^{\gamma,\delta}|$ for each tuning parameter pair on a given grid. We need to go through several iterations of this recording procedure to find a proper range of grid, in which we can get rank$(\widehat{L}^{\gamma,\delta}) = \widehat{K}^{\text{Scree}}$ and $10^{-4}|X| \leq |\widehat{S}^{\gamma,\delta}| \leq 10^{-1}|X|$, via adjusting the range of grid for γ and δ repeatedly. Once we find a grid, which satisfies above constraints, we choose a pair of tuning parameters:

$$(\gamma^{\text{Heu}}, \delta^{\text{Heu}}) := \left\{ (\gamma, \delta) : \text{rank}(\widehat{L}^{\gamma,\delta}) = \widehat{K}^{\text{Scree}}, \text{mode}|\widehat{S}^{\gamma,\delta}| : \text{subject to} 10^{-4}|X| \leq |\widehat{S}^{\gamma,\delta}| \leq 10^{-1}|X| \right\}$$

One might wonder how traditional model selection methods, such as the Bayes Information Criterion (BIC;[17]) and the AIC, work where we recall that BIC and AIC are defined as follows:

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

and

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

Here M indicates the current model, which is implicitly understood that the model is obtained from certain tuning parameter pair (γ, δ). We use $\mathbb{L}_n(\hat{\beta}(M))$ to denote 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 $\hat{S}^{\gamma,\delta}$ and the low-rank matrix $\hat{L}^{\gamma,\delta}$. In detail, if we have rank$(\hat{L}^{\gamma,\delta}) = K$, we can establish the following

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

since the number of free parameters in $\hat{L}^{\gamma,\delta}$ is K plus $nK - K(K+1)/2$, which is the number of free parameters in determining K orth-normal vectors. Additional 1 in the last term is due to $\hat{\alpha}$. We want to find a pair (γ, δ), which minimizes BIC(M) and AIC(M) as a function of (γ, δ) respectively, where we denote them as follows:

$$(\gamma^{BIC}, \delta^{BIC}) := \arg\min_{\gamma,\delta}\text{BIC}(M), \quad (\gamma^{AIC}, \delta^{AIC}) := \arg\min_{\gamma,\delta}\text{AIC}(M)$$

We evaluate the models that are selected via our heuristic approach, BIC, and AIC by using following four evaluation metrics.

$$
\begin{aligned}
M_1 &= \mathbb{1}\{\text{rank}(\hat{L}) = \text{rank}(L^*)\}, \\
M_2 &= \frac{\left|\{(i,j) : i < j : S_{i,j}^* \neq 0 \quad \& \quad \hat{S}_{i,j} \neq 0\}\right|}{\left|\{(i,j) : i < j : S_{i,j}^* \neq 0\}\right|}, \\
M_3 &= \frac{\left|\{(i,j) : i < j : S_{i,j}^* = 0 \quad \& \quad \hat{S}_{i,j} \neq 0\}\right|}{\left|\{(i,j) : i < j : S_{i,j}^* = 0\}\right|}, \\
M_4 &= \frac{\left|\{\text{Misclassified Nodes}\}\right|}{n},
\end{aligned}
$$

where M_1 is a metric on whether the selected model recovers the true low rank structure of network, M_2 evaluates the positive selection rate of the sparse ad-hoc structure in network, and M_3 evaluates the false discovery rate of ad-hoc edges. M_4 calculates the proportion of mis-classified nodes to the entire nodes in the network. With properly selected tuning parameter, M_1 will be 1, M_2 will be close to 1, and M_3 and M_4 will get close to 0. We present the evaluation results of the six cases via the four criteria, M_1, M_2, M_3, and M_4 in Table.1 .

6.1.3 | Several Observations

1. **Model Selection.** First and foremost, choosing a good pair of tuning parameters is critical when it comes to making a good statistical inference on data. As presented in Table.1 , both BIC and AIC, which are well known for their model selection consistency in asymptotic setting, appear to under-estimate either the number of topics or the number of ad-hoc links in the networks in our synthetic settings. This may be caused by the fact that these traditional methods take the sample size into account, and therefore penalizes the model complexity too harshly. In the heuristic approach, scree-plot plays an important role when it comes to recovering the number of topics, and this strategy leads us to good model selection results for all the six cases considered in two scenarios. (See Fig.2)

2. **Node Membership.** After fitting the model with a proper pair of tuning parameters, (γ, δ), we need to determine whether the ith paper belongs to the kth topic or not. We apply a simple k-means clustering algorithm on the $\hat{L}$ matrix's K eigenvectors where K denotes the rank of matrix $\hat{L}$. For three cases in Scenario 1, where each paper in the network only has one topic, we confirm that k-means algorithm performs well on classifying papers in the network. However, in Scenario 2 where we allow the papers in the network can have more than one topic, naive implementation of k-means algorithm entails a problem, since it is not able to afford the overlapped membership of nodes. Here, we create a matrix $\hat{E}_K \in \mathbb{R}^{n \times K}$, whose ith column corresponds to the ith eigenvector of the $\hat{L}$. In order to obtain a sense on how many clusters of papers exist in latent space, we project each row of the $\hat{E}_K$ on the first and second principal components of data matrix $\hat{E}_K$, and plot the projected points on 2-dimensional plane. We count the number of distinct clusters plotted

	Scenario 1								
	Case 1			Case 2			Case 3		
	Heuristic	AIC	BIC	Heuristic	AIC	BIC	Heuristic	AIC	BIC
M_1	1 (3)	0 (2)	0 (2)	1 (4)	0 (3)	0 (3)	1 (5)	0 (4)	0 (4)
M_2	1	0	0	1	1	1	1	1	1
M_3	0	0	0	0	0	0	0	0	0
M_4	0	$\frac{10}{30}$	$\frac{10}{30}$	0	$\frac{20}{80}$	$\frac{20}{80}$	0	$\frac{24}{120}$	$\frac{24}{120}$
	Scenario 2								
	Case 4			Case 5			Case 6		
	Heuristic	AIC	BIC	Heuristic	AIC	BIC	Heuristic	AIC	BIC
M_1	1 (3)	1 (3)	1 (3)	1 (3)	1 (3)	1 (3)	1 (3)	1 (3)	1 (3)
M_2	$\frac{17}{18}$	0	0	$\frac{17}{18}$	0	0	$\frac{16}{18}$	0	0
M_3	0	0	0	0	0	0	0	0	0
M_4	0	0	0	0	0	0	0	0	0

TABLE 1 For two scenarios, our heuristic method chooses the model with $\hat{L}$ with true rank, $\hat{S}$ whose M_2 value is close to 1, and M_3 value is close to 0. Also note that it chooses a model whose mis-classification rate is close to 0. A number in the parentheses represents the rank of $\hat{L}$ estimated from $(\gamma^{\text{Heu}}, \delta^{\text{Heu}})$, $(\gamma^{\text{AIC}}, \delta^{\text{AIC}})$ and $(\gamma^{\text{BIC}}, \delta^{\text{BIC}})$ for each case.
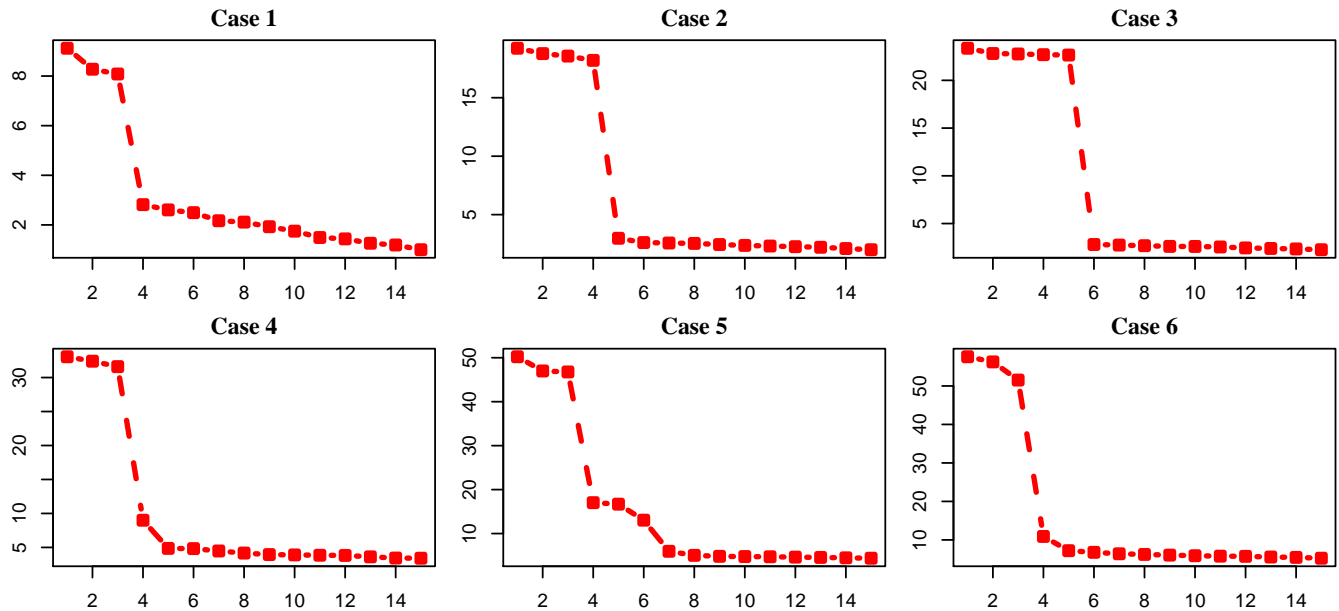

FIGURE 2 Scree plots for six synthetic Networks. $\hat{K}^{\text{Scree}}$ recovers the number of topics in the network correctly for all six cases.

on the plane. Subsequently, we run the k-means algorithm on the projected points. Refer Table.1 and Fig.3 for the references on how this procedure performs on six synthetic cases.

6.2 | Citation networks for statisticians

Recently, Ji and Jin [11] published an interesting dataset on citation network of papers from statistics journals. Specifically, this dataset is based upon all papers published from 2003 to the first half of 2012, from the four top statistical journals: Annals of Statistics, Biometrika, Journal of American Statistical Association and Journal of Royal Statistical Society (Series B). Citational relationships of 3248 papers are given in the form of adjacency matrix. In our analysis, we focus our attentions on the papers
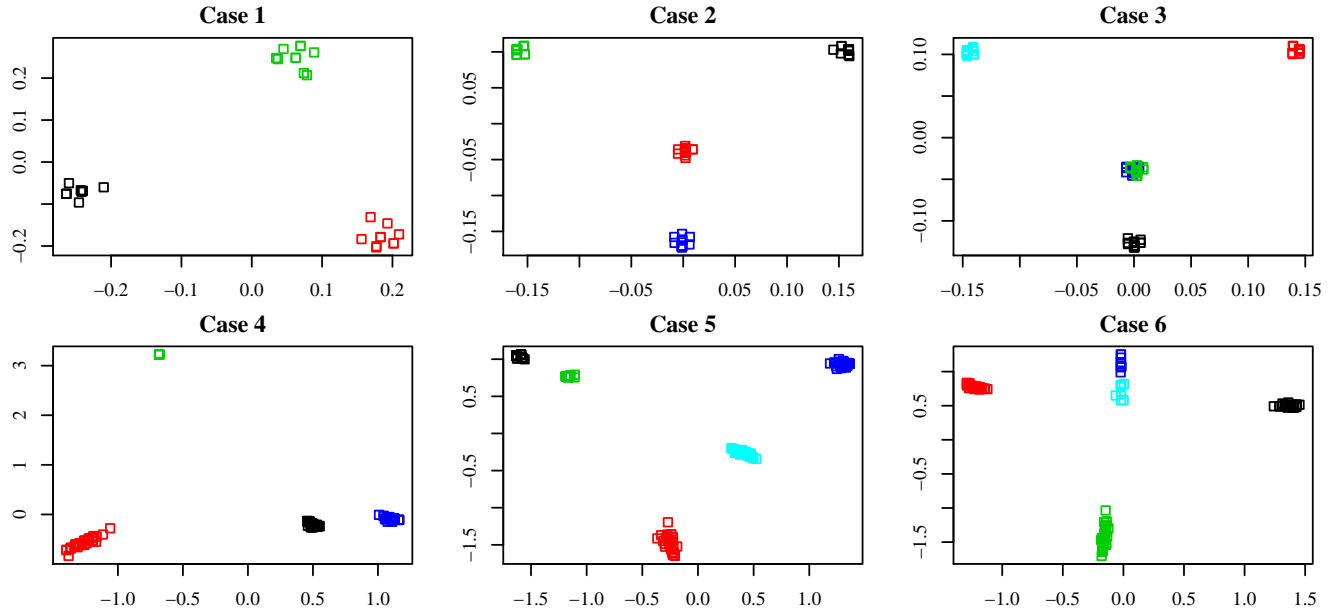
FIGURE 3 Case 1 ∼ 3 : Plots of rows from the first two eigenvectors of $\widehat{L}^{\gamma^{\text{Heu}},\delta^{\text{Heu}}}$. Case 4 ∼ 6 : Plots of the projected points on the first (X-axis) and second (Y-axis) principal component of the data matrix $\widehat{E}_K^{\gamma^{\text{Heu}},\delta^{\text{Heu}}}$. Different colors represent different clusters of papers that k-means algorithm assigns.
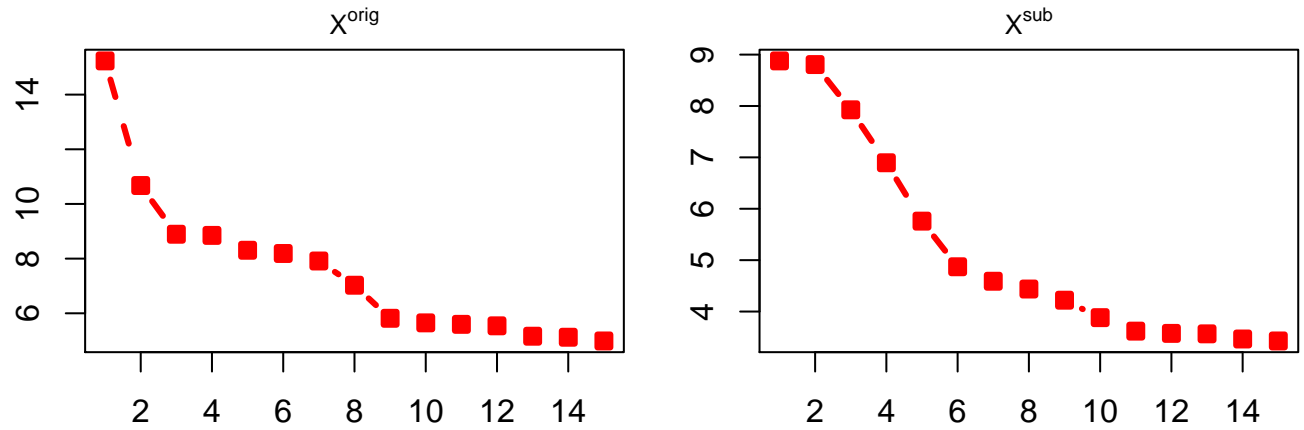

FIGURE 4 From left to right : Scree plots of the adjacency matrix X^{orig} and X^{sub}.

which have greater than or equal to 10 citational edges in the network of Ji and Jin [11]. After collecting papers with greater than or equal to 10 citational edges and eliminating those that have no connecting edges from the rest, we have 232 papers in total.

We denote the adjacency matrix of these 232 papers as X^{orig}. Elbow points of the scree plot from X^{orig} may be at the 3rd, 5th, or 9th largest eigenvalue, suggesting that there are from 2 to 8 embedded topics in the network. (See Fig. 4) In light of this, we conduct the analysis in two steps:

1. First, we assume that the network X^{orig} has 2 distinct topics and one giant mixed-component, which has a sub-network structure. Under this assumption, we set $\widehat{K}^{\text{scree}}$ as 3, and select a proper model via our heuristic method. Then, we perform a k-means algorithm on matrix $\widehat{E}_3$ treating each row of the matrix as one data point. Note that we set the total number of clusters in the network as 3 when we run the clustering algorithm.

2. Next, we restrict the network to the giant component ignoring all the edges to/from outside and obtain a subnetwork. We denote the adjacency matrix of this subnetwork as X^{sub}. We set $\widehat{K}^{\text{scree}}$ as 5, and also select a proper model through the heuristic method. Here, we run the k-means algorithm on $\widehat{E}_5$ setting the number of clusters as 5.

In the first step, a pair of parameters, $(\gamma^{\text{Heu}}, \delta^{\text{Heu}}) = (0.0021094, 0.01913)$, gives us $\widehat{L}$ with rank 3, and $\widehat{S}$ with $|\widehat{S}| = 51$. We list the first two topics discovered through our analysis.

- Variable selection, which includes 43 paper.

- Multiple Hypothesis Testing, which includes 31 papers.

The first topic studies on Variable Selection with high-dimensional data. The second topic discusses Controlling False Discovery Rate in various statistical settings. The third group, which consists of 158 papers, is hard to interpret and appears to have sub-network structures. For further investigation, we set this group as a giant component in the network, and denote the corresponding component's adjacency matrix as X^{sub}. We perform a model selection as described in aforementioned Step 2. A pair of tuning parameters, $(\gamma^{\text{Heu}}, \delta^{\text{Heu}}) = (0.00312, 0.0146)$, gives us the model with $\widehat{L}$ with rank 5, and $\widehat{S}$ with $|\widehat{S}| = 88$, and we can obtain five sub-communities as follows:

- Non-parametric Bayesian Statistics, which includes 15 papers.

- Functional / Longitudinal Data analysis, which includes 16 papers.

- Dimension Reduction, which includes 14 papers.

- High-dimensional covariance estimation, which includes 15 papers.

- Mixed topics, which includes 98 papers.

From the sub-network X^{sub}, we got four meaningful topics: Bayesian Statistics, Functional/Longitudinal Data Analysis, Dimension Reduction, and high-dimensional covariance estimation. Due to the small volume of each community, we could manually check that the false discovery for each community is all zero. (Full list of papers for each community is provided in $https : //sites.google.com/site/namjoonsuh/publications.$)

Sub-network structure has also a big collection of papers which we refer it as "Mixed Topics" cluster. 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 the five communities listed above, but also we could identify the papers with combinations of two or three topics. Papers, such as 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, the paper, 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 11 citation relationships it has with other papers, 6 of them came from the relationships with papers from Mixed topics.

Non-zero components of $\widehat{S}$ capture the citation relationships among papers that are not attributable to the common topics. The selected model has 51 sparse edges, and all of them are positive edges. In Table 6.2, we provide 10 pairs of papers that have the largest estimated $\widehat{S}_{ij}$. All the 10 edges come from the 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 appear 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. The second pair of papers comes from the Mixed topics community (denoted as *Mixed*) and the Variable Selection (*VarSel*) community. This case is interesting since both papers study about variable selection problem, but they are classified in different communities and connected via an ad-hoc link. Specifically, the paper from $Mixed$ community study the variable selection problem under the non-parametric Bayesian framework, and compare their method with the one that is proposed in *VarSel* paper, "Adaptive Lasso".

Pair	Community	Title
1	FuncAn	Properties of principal component methods for functional and longitudinal data analysis
	VarSel	Nonconcave penalized likelihood with a diverging number of parameters
2	VarSel	The adaptive lasso and its oracle properties
	Mixed	Nonparametric Bayes conditional distribution modeling with variable selection
3	DimRed	Contour projected dimension reduction
	VarSel	Factor profiled sure independence screening
4	VarSel	Factor profiled sure independence screening
	DimRed	Sliced regression for dimension reduction
5	CovEst	Two sample tests for high-dimensional covariance matrices
	VarSel	The sparsity and bias of the LASSO selection in high-dimensional linear regression
6	MulT	Innovated higher criticism for detecting sparse signals in correlated noise
	CovEst	Regularized estimation of large covariance matrices
7	DimRed	A constructive approach to the estimation of dimension reduction directions
	VarSel	Factor profiled sure independence screening
8	VarSel	A majorization-minimization approach to variable selection using spike and slab priors
	Mixed	Empirical Bayes selection of wavelet thresholds
9	Mixed	Nonparametric inferences for additive models
	VarSel	Nonparametric independence screening in sparse ultra-high-dimensional additive models
10	VarSel	Sure independence screening in generalized linear models with NP-dimensionality
	Mixed	Maximum likelihood estimation in semiparametric regression models with censored data

TABLE 2 All the 10 edges come from pairs of papers from different communities. For instance in the first pair, *FuncAn* paper cites *VarSel* paper for borrowing a mathematical representation to build a theorem. But they are not related in terms of topic.

7 | DISCUSSION

We propose a new combined latent factor and sparse graphical model. We consider the regularized likelihood by means of L_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 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.

Nonetheless, there are still several questions remaining to be answered. 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 well in general situation, 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 for citation network in real world, 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 directions of edges in our model. However, it would be interesting to consider incorporating 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 Fig.4 (left), we can see that bunch of Multiple Testing papers are assigned as Mixed cluster, and in Fig.4 (right) 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.

8 | APPENDIX

In this Section, first we briefly introduce several notations, a notion on the decomposability of regularizer, and a useful lemma proved in the work [1]. Then, we present Lemma2 and its proof. Finally, we present the proof of our Theorem1.

8.1 | Preliminary

Throughout the proof, we adopt the convenient short-hand notation on projection of matrix P on subspace M as P_M. We use $\langle A, B \rangle$ to denote the trace inner product of two matrices A and B $\left(\text{i.e.,} \langle A, B \rangle = \text{tr}\left(A^T B\right)\right)$. We use $\|A\|_\infty$ to denote the maximum absolute entry of matrix A, and use $\|B\|_{op}$ to denote the largest singular value of matrix B. And we will use the notion of decomposability of L_1 norm with respect to a pair of subspace $(M, M^\perp)$. Given an arbitrary subset $S \subseteq \{1, 2, \ldots, n\} \times \{1, 2, \ldots, n\}$ of matrix indices, M is defined as follows:

$$M(S) := \{U \in \mathbb{R}^{n \times n} | U_{ij} = 0, \forall (i, j) \in S\}$$

and $M^\perp(S) := (M(S))^\perp$. With this in mind, we recall the formal definition of the decomposability of L_1 norm as follows:

Definition 1. Given a subspace $M \subset \mathbb{R}^{n \times n}$ and its orthogonal complement $M^\perp$, an elementwise L_1 norm is decomposable with respect to $(M, M^\perp)$ if

$$\|A + B\|_1 = \|A\|_1 + \|B\|_1, \forall A \in M \text{ and } B \in M^\perp,$$

The notion of decomposability is used to penalize the perturbation from the model subspace M, and to obtain the tightest bound the L_1 norm can achieve. We will also use two results in our proof, which are presented and proved in the work [1]. For the convenience of readers, we present them here as follows:

Lemma 1. (Agarwal, et al [1]) For any $k = 1, 2, \ldots, n$, there is a decomposition $\widehat{\Delta}^L = \widehat{\Delta}_A^L + \widehat{\Delta}_B^L$ such that:

1. The decomposition satisfies

$$\text{rank}(\widehat{\Delta}_A^L) \leq 2k, \quad and \quad (\widehat{\Delta}_A^L)^T \widehat{\Delta}_B^L = (\widehat{\Delta}_B^L)^T \widehat{\Delta}_A^L = 0 \tag{18}$$

2. The difference $\mathbb{Q}(L^*, S^*) - \mathbb{Q}(\widehat{\Delta}^L + L^*, \widehat{\Delta}^S + S^*)$ is upper-bounded by

$$\mathbb{Q}(\widehat{\Delta}_A^L, \widehat{\Delta}_M^S) - \mathbb{Q}(\widehat{\Delta}_B^L, \widehat{\Delta}_{M^\perp}^S) + 2 \sum_{j=k+1}^n \sigma_j(L^*) + 2\frac{\gamma}{\delta} \|S_{M^\perp}^*\|_1 \tag{19}$$

where the notation $\mathbb{Q}(L, S)$ is defined as the weighted combination of the two regularizers for any pair of positive tuning parameters (γ, δ):

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

8.2 | Lemma 2

Lemma 2. If a pair of regularization parameters (δ, γ) satisfies condition (12), then for $\mathbb{Q}\left(\widehat{\Delta}_B^L, \widehat{\Delta}_{M^\perp}^S\right)$, we have

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

Proof. Through the application of basic inequality by using optimality of $\widehat{\Theta}$ and feasibility of Θ^* to convex program (11), we have

$$h(\widehat{\Theta}) - h(\Theta^*) \leq \delta \mathbb{Q}(L^*, S^*) - \delta \mathbb{Q}(\widehat{\Delta}^L + L^*, \widehat{\Delta}^S + S^*) \tag{20}$$

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

$$
\begin{aligned}
h(\widehat{\Theta}) - h(\Theta^*) &\geq \langle \nabla_\Theta h(\Theta^*), \widehat{\Theta} - \Theta^* \rangle \\
&= -\langle \frac{1}{n}(X - P^*), \widehat{\Delta}^\alpha \mathbb{1}\mathbb{1}^T + \widehat{\Delta}^L + \widehat{\Delta}^S \rangle \\
&\geq -\frac{1}{n}\|X - P^*\|_{op}(\|\widehat{\Delta}^\alpha \mathbb{1}\mathbb{1}^T\|_* + \|\widehat{\Delta}^L\|_*) + \frac{1}{n}\|X - P^*\|_\infty \|\widehat{\Delta}^S\|_1 \\
&\geq -\frac{\delta}{2}(\|\widehat{\Delta}^\alpha \mathbb{1}\mathbb{1}^T\|_F + \|\widehat{\Delta}_A^L\|_* + \|\widehat{\Delta}_B^L\|_*) - \frac{\gamma}{2}(\|\widehat{\Delta}_M^S\|_1 + \|\widehat{\Delta}_{M^\perp}^S\|_1)
\end{aligned}
\tag{21}
$$

An application of 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:

$$
\mathbb{Q}(\widehat{\Delta}_A^L, \widehat{\Delta}_M^S) - \mathbb{Q}(\widehat{\Delta}_B^L, \widehat{\Delta}_{M^\perp}^S) + 2\sum_{j=k+1}^n \sigma_j(L^*) + 2\frac{\gamma}{\delta}\|S_{M^\perp}^*\|_1
\tag{22}
$$

By combining relations (21) and (22), we can get the upper bound of $\mathbb{Q}(\widehat{\Delta}_B^L, \widehat{\Delta}_{M^\perp}^S)$:

$$
\mathbb{Q}(\widehat{\Delta}_B^L, \widehat{\Delta}_{M^\perp}^S) \leq \|\widehat{\Delta}^\alpha \mathbb{1}\mathbb{1}^T\|_F + 3\mathbb{Q}(\widehat{\Delta}_A^L, \widehat{\Delta}_M^S) + 4\sum_{j=k+1}^n \sigma_j(L^*) + 4\frac{\gamma}{\delta}\|S_{M^\perp}^*\|_1
\tag{23}
$$

$\square$

8.3 | Proof of Theorem 1

Proof. Since $\widehat{\Theta}$ and Θ^* are optimal minimizer and feasible solution respectively for the convex program (11), we have

$$
h(\widehat{\Theta}) + \delta\left\|\widehat{L}\right\|_* + \gamma\left\|\widehat{S}\right\|_1 \leq h(\Theta^*) + \delta\|L^*\|_* + \gamma\|S^*\|_1
\tag{23}
$$

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

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

By rearranging the term in (23) and plugging in above inequality relation, we get:

$$
\frac{\tau}{2}\|\widehat{\Delta}^\Theta\|_F^2 \leq -\langle \nabla_\Theta h(\Theta^*), \widehat{\Theta} - \Theta^* \rangle + \delta\|L^*\|_* + \gamma\|S^*\|_1 - \delta\|\widehat{L}\|_* - \gamma\|\widehat{S}\|_1
\tag{24}
$$

Through the definition of $\mathbb{Q}$, we can rewrite (24) as follows:

$$
\frac{\tau}{2}\|\widehat{\Delta}^\Theta\|_F^2 \leq -\langle \nabla_\Theta h(\Theta^*), \widehat{\Theta} - \Theta^* \rangle + \delta\mathbb{Q}(L^*, S^*) - \delta\mathbb{Q}\left(\widehat{\Delta}^L + L^*, \widehat{\Delta}^S + S^*\right)
\tag{25}
$$

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

$$
\mathbb{Q}\left(\widehat{\Delta}_A^L, \widehat{\Delta}_M^S\right) - \mathbb{Q}\left(\widehat{\Delta}_B^L, \widehat{\Delta}_{M^\perp}^S\right) + 2\sum_{j=k+1}^n \sigma_j(L^*) + 2\frac{\gamma}{\delta}\left\|S_{M^\perp}^*\right\|_1
\tag{26}
$$

First, we want to control upper bound of the term $-\langle \nabla_\Theta h(\Theta^*), \widehat{\Theta} - \Theta^* \rangle$ in (25).

$$
\begin{aligned}
-\langle \nabla_\Theta h(\Theta^*), \widehat{\Theta} - \Theta^* \rangle &= \langle \frac{1}{n}(X - P^*), \widehat{\Delta}^\alpha \mathbb{1}\mathbb{1}^T + \widehat{\Delta}^L + \widehat{\Delta}^S \rangle \\
&\leq \frac{1}{n}\|X - P^*\|_{op}\left(\left\|\widehat{\Delta}^\alpha \mathbb{1}\mathbb{1}^T\right\|_* + \left\|\widehat{\Delta}^L\right\|_*\right) + \frac{1}{n}\|X - P^*\|_\infty \left\|\widehat{\Delta}^S\right\|_1 \\
&\leq \frac{1}{n}\|X - P^*\|_{op}\left(\left\|\widehat{\Delta}^\alpha \mathbb{1}\mathbb{1}^T\right\|_F + \left\|\widehat{\Delta}_A^L\right\|_* + \left\|\widehat{\Delta}_B^L\right\|_*\right) + \frac{1}{n}\|X - P^*\|_\infty \left(\left\|\widehat{\Delta}_M^S\right\|_1 + \left\|\widehat{\Delta}_{M^\perp}^S\right\|_1\right) \\
&\leq \frac{\delta}{2}\left(\left\|\widehat{\Delta}^\alpha \mathbb{1}\mathbb{1}^T\right\|_F + \left\|\widehat{\Delta}_A^L\right\|_* + \left\|\widehat{\Delta}_B^L\right\|_*\right) + \frac{\gamma}{2}\left(\left\|\widehat{\Delta}_M^S\right\|_1 + \left\|\widehat{\Delta}_{M^\perp}^S\right\|_1\right)
\end{aligned}
\tag{27}
$$

Combining the inequalities (26) and (27), we can obtain the upper bound of RHS in (25) as follows:

$$
\frac{\tau}{2}\|\widehat{\Delta}^\Theta\|_F^2 \leq \frac{\delta}{2}\|\widehat{\Delta}^\alpha \mathbb{1}\mathbb{1}^T\|_F + \frac{3\delta}{2}\mathbb{Q}(\widehat{\Delta}_A^L, \widehat{\Delta}_M^S) + 2\delta\sum_{j=k+1}^n \sigma_j(L^*) + 2\gamma\|S_{M^\perp}^*\|_1
\tag{28}
$$

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

$$
\begin{aligned}
\|\widehat{\Delta}^{\Theta}\|_F^2 &= \|\widehat{\Theta} - \Theta^*\|_F^2 \\
&= \|\widehat{\Delta}^{\alpha}\mathbb{1}\mathbb{1}^T + \widehat{\Delta}^L + \widehat{\Delta}^S\|_F^2 \\
&= \|\widehat{\Delta}^{\alpha}\mathbb{1}\mathbb{1}^T\|_F^2 + \|\widehat{\Delta}^L + \widehat{\Delta}^S\|_F^2 + 2\langle \widehat{\Delta}^L + \widehat{\Delta}^S, \widehat{\Delta}^{\alpha}\mathbb{1}\mathbb{1}^T\rangle \\
&= \|\widehat{\Delta}^{\alpha}\mathbb{1}\mathbb{1}^T\|_F^2 + \|\widehat{\Delta}^L\|_F^2 + \|\widehat{\Delta}^S\|_F^2 + 2\langle \widehat{\Delta}^L + \widehat{\Delta}^S, \widehat{\Delta}^{\alpha}\mathbb{1}\mathbb{1}^T\rangle + 2\langle \widehat{\Delta}^L, \widehat{\Delta}^S\rangle
\end{aligned}
\tag{29}
$$

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, apply triangular inequality on $\widehat{\Delta}^{\alpha}$, and lastly we apply the constraint imposed on $|\alpha|$ stated in Assumption 2.

$$
\begin{aligned}
|\langle \widehat{\Delta}^L + \widehat{\Delta}^S, \widehat{\Delta}^{\alpha}\mathbb{1}\mathbb{1}^T\rangle| &= |\langle \widehat{\Delta}^S, \widehat{\Delta}^{\alpha}\mathbb{1}\mathbb{1}^T\rangle| \\
&\leq \|\widehat{\Delta}^{\alpha}\mathbb{1}\mathbb{1}^T\|_{\infty}\|\widehat{\Delta}^S\|_1 \\
&\leq (|\widehat{\alpha}| + |\alpha^*|)\|\widehat{\Delta}^S\|_1 \\
&\leq 2C\kappa\|\widehat{\Delta}^S\|_1
\end{aligned}
\tag{30}
$$

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.

$$
\begin{aligned}
|\langle \widehat{\Delta}^L, \widehat{\Delta}^S\rangle| &\leq \|\widehat{\Delta}^L\|_{\infty}\|\widehat{\Delta}^S\|_1 \\
&\leq (\|\widehat{L}\|_{\infty} + \|L^*\|_{\infty})\|\widehat{\Delta}^S\|_1 \\
&\leq \left(\frac{2\kappa}{n}\right)\|\widehat{\Delta}^S\|_1
\end{aligned}
\tag{31}
$$

We can combine the inequality (29), (30) and (31). Then applying the assumption on regularization parameter γ, and the fact $\|\widehat{\Delta}^L\|_* \geq 0$ sequentially, we can get,

$$
\begin{aligned}
\frac{\tau}{2}\|\widehat{\Delta}^{\Theta}\|_F^2 &\geq \frac{\tau}{2}\|\widehat{\Delta}^{\alpha}\mathbb{1}\mathbb{1}^T\|_F^2 + \frac{\tau}{2}\|\widehat{\Delta}^L\|_F^2 + \frac{\tau}{2}\|\widehat{\Delta}^S\|_F^2 - \kappa\tau\left(\frac{Cn+1}{n}\right)\|\widehat{\Delta}^S\|_1 \\
&\geq \frac{\tau}{2}\|\widehat{\Delta}^{\alpha}\mathbb{1}\mathbb{1}^T\|_F^2 + \frac{\tau}{2}\|\widehat{\Delta}^L\|_F^2 + \frac{\tau}{2}\|\widehat{\Delta}^S\|_F^2 - \frac{\gamma}{2}\|\widehat{\Delta}^S\|_1 \\
&\geq \frac{\tau}{2}\|\widehat{\Delta}^{\alpha}\mathbb{1}\mathbb{1}^T\|_F^2 + \frac{\tau}{2}\|\widehat{\Delta}^L\|_F^2 + \frac{\tau}{2}\|\widehat{\Delta}^S\|_F^2 - \frac{\delta}{2}\mathbb{Q}(\widehat{\Delta}^L, \widehat{\Delta}^S)
\end{aligned}
\tag{32}
$$

By combining the relations (28) and (32), applying triangular inequality, $\mathbb{Q}(\widehat{\Delta}^L, \widehat{\Delta}^S) \leq \mathbb{Q}(\widehat{\Delta}_A^L, \widehat{\Delta}_M^S) + \mathbb{Q}(\widehat{\Delta}_B^L, \widehat{\Delta}_{M^\perp}^S)$, and rearranging the term, we can get following inequality,

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

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

$$
\frac{\tau}{2}\|\widehat{\Delta}^{\alpha}\mathbb{1}\mathbb{1}^T\|_F^2 + \frac{\tau}{2}\|\widehat{\Delta}^L\|_F^2 + \frac{\tau}{2}\|\widehat{\Delta}^S\|_F^2 - \frac{\delta}{2}\|\widehat{\Delta}^{\alpha}\mathbb{1}\mathbb{1}^T\|_F \leq \frac{7\delta}{2}\mathbb{Q}(\widehat{\Delta}_A^L, \widehat{\Delta}_M^S) + 4\delta\sum_{j=k+1}^n \sigma_j(L^*) + 4\gamma\|S_{M^\perp}^*\|_1
\tag{33}
$$

Noting that $\widehat{\Delta}_A^L$ has rank at most $2k$ and that $\widehat{\Delta}_M^S$ lies in the model space M, we find that

$$
\begin{aligned}
\delta\mathbb{Q}(\widehat{\Delta}_A^L, \widehat{\Delta}_M^S) &\leq \sqrt{2k}\delta\|\widehat{\Delta}_A^L\|_F + \Psi(M)\gamma\|\widehat{\Delta}_M^S\|_F \\
&\leq \sqrt{2k}\delta\|\widehat{\Delta}^L\|_F + \Psi(M)\gamma\|\widehat{\Delta}^S\|_F
\end{aligned}
\tag{34}
$$

Here $\Psi(M)$ measures the compatibility between Frobenius norm and component-wise L_1 regularizer, where M is an arbitrary subset of matrix indices of cardinality at most s.

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

Using Cauchy-Schwarz inequality, we can easily check the quantity $\Psi(M)$ is bounded by at most $\sqrt{s}$. Plugging in the relation (34) into (33) and rearranging the term relevant with $e^2(\hat{\alpha}\mathbb{1}\mathbb{1}^T, \hat{L}, \hat{S})$ yield the claim. $\qquad\square$

Bibliography

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

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

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

[18] 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.