Factor Analysis on Report Citations, Using a Combined Latent and Graphical Model

Namjoon Suh∗, Eric Heim†, Timothy Van Slyke†, Lee Seversky†, V. Xiaoming Huo∗

School of Industrial and Systems Engineering, Georgia Institute of Technology ∗, Air Force Research Laboratory†,

Model Formulation

- Motivated by the work [4], we study a citation network, where each node (i.e., item) can be a technical report or a publication. We denote a binary random variable X_{ij}, where $1 \le i, j \le n$ 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 $X_{ij} = 0$.
- **Latent Variable Model** 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) $k, 1 \le k \le K$. Here K is the total number of underlying topics (i.e., factors, or trends). We assume a logistic model for X_{ij}'s: for $1 \le i, j \le n$,

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

The justification of above model is that when both node i and node j are related with topic k, they have a higher chance to cite one another.

- **Conditional Graphical Model** 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

$$\mathbb{P}(X_{ij} = 1) = \frac{e^{\alpha' + S_{ij}}}{1 + e^{\alpha' + S_{ij}}}, \tag{2}$$

where $S_{ij} \in \mathbb{R}$, $S_{ij} \ge 0$, for $1 \le i, j \le n$, denotes the relation between nodes i and j.

- **Combined Model** By combining above two models, we can fully account the dependent structure of citation network. Under the assumption of independence of X_{ij}, $1 \le i, j \le n$, we can write joint probability function as follows

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

Assumptions and Penalization

- We want the matrix $S \in \mathbb{R}^{n \times n}$ to be as **sparse** as possible.
- 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**.
- Overall, the rank of matrix $F^T D F$ cannot be larger than $\min\{n, K\}$ ($k << n$)
- 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 identify matrix. Notice that matrix $F' = PF$ is also a factor loading matrix, and matrix $D' = PDP^T$ is still a diagonal matrix; we have

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

- Neither rank K of L nor the graphical structure is known.

Along with the line of these assumptions, we propose a penalized log-likelihood estimation approach as follows:

$$(\hat{\alpha}, \hat{L}, \hat{S}) = \arg\min_{\alpha, L, S} \left\{ -\frac{1}{n} \mathbb{L}_n(\alpha, L, S; X) + \gamma \|S\|_1 + \delta \|L\|_* \right\} \tag{4}$$

On the choice of Tuning parameters

We can choose γ and δ in (4) by minimizing the Bayes information criterion (BIC; [7]) that is known to yield consistent variable selection. BIC is defined as

$$\mathrm{BIC}(M) = -2\mathbb{L}_n(\hat{\beta}(M)) + |M| \log N,$$

where M is the current model, $\mathbb{L}_n(\hat{\beta}(M))$ is the maximal log-likelihood for a given model M. Note that $N = n(n-1)/2$, when n denotes the number of papers in network. If $\mathrm{rank}(L) = K$, we can establish the following

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

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.

Summary

- We propose a combined latent and graphical model for the citation network, where either a latent model or a graphical 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 graphical component** that captures the remaining **ad hoc dependence**.
- Model selection and parameter estimation are carried out simultaneously through construction of a pseudo-likelihood function and properly chosen penalty terms. The convexity of the **pseudo-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 which can demonstrate our new method works well in practical situations. The proposed method has been applied to a real application in **HEP-Ph (high energy physics phenomenology) citation data set**.

Synthetic and Real Data Analysis

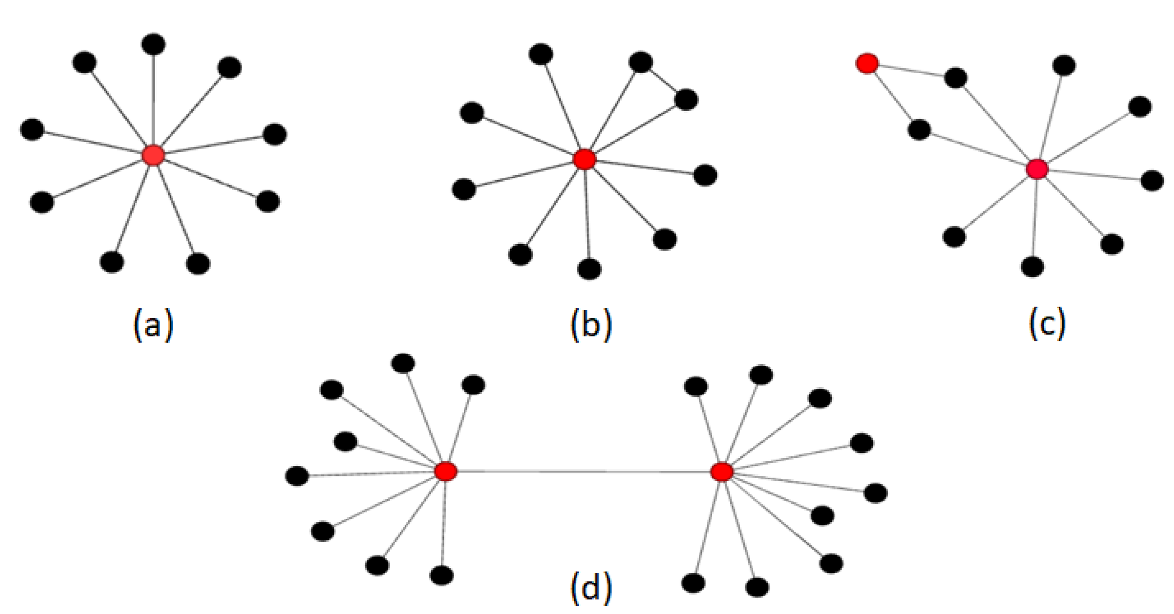

TABLE 1 Parameter estimation results of four synthetic cases with optimal tuning parameters.

	Case 1 (a)	Case 2 (b)	Case 3 (c)	Case 4 (d)
$\hat{\alpha}$	-4.625	-5.427	-1.914	-4.747
$\hat{S}_+$	∅	{(7,8)}	{(8,10), (9,10)}	∅
Rank($\hat{L}$)	1	1	1	2
$(\hat{\gamma}, \hat{\delta})$	(0.01, 0.01)	(0.01, 0.0146)	(0.0316, 0.0548)	(0.01, 0.01)

TABLE 2 We denote the true parameter as (S^*, L^*). And set the number of nonzero entries of upper triangular part of matrix S^* as $n^2 * 0.01$. (Each entry of S^* is non-negative, following the assumption in our paper). Set $K = n * 0.1$ as rank of matrix L^*, and L^* should be symmetric positive definite matrix. Generate binary random matrix $F \in \{0, 1\}^{K X n}$, non-negative diagonal matrix $D \in R^{K X K}$, and produce $L^* = F^T DF \in R^{n X n}$.

Dimension n	rank(L^*)	$\|S^*\|_0$	rank($\hat{L}$)	$\|\hat{S}\|_0$	$R_{S^* \cap \hat{S}}$
10	1	1	1	1	1/1
20	2	4	2	2	2/4
30	3	9	3	8	5/9
40	4	16	4	15	14/16

Figure 1: (a) is the case where all other nine papers are connected with one in the center. We can quickly realize that ten papers in (a) are connected by one commonly shared topic, and there is no ad-hoc dependent structure, which cannot be explained by this common topic. In second and third case, (b) and (c), we add one and two artificial edges to the first case, (a), which can be considered as ad-hoc dependent structures of the network system. Fourth case, (d), displays the network with two common latent topics and no ad-hoc dependency between 20 papers. We also perform additional experiment whether our proposed method decomposes the sparse and latent component well under the setting stated in Table2. [3]

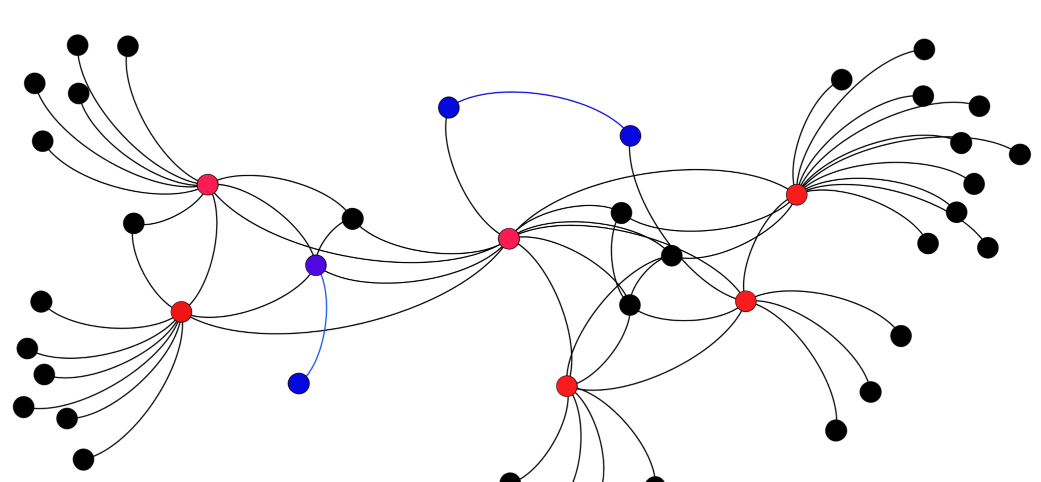

TABLE 3 10 most likely terms in each 6 topics obtained through LDA analysis. Numbers at the first row of table are IDs of 6 corresponding hub papers. Words colored in blue are topical words commonly shared by 6 topics. Words in red specifically characterize each embedded topic.

9907233	9704296	9606402	9702314	9611297	9903217
Gamma	Gamma	Tau	Gamma	Gamma	Meson
Phi	Mass	Meson	Bar	Chiral	Channel
Bar	Meson	Decay	Eta	Rho	Chiral
Meson	Phi	State	Phi	Meson	Lagrangian
Process	Theory	Final	Mass	Theory	Use
Channel	Chiral	Strange	Decay	Vector	Resonance
Branch	Rho	Vector	Resonance	Phi	Comment
Ratio	Perturbation	Page	Data	Comment	Coupling
Resonance	Electromagnetic	Comment	State	Page	GEV
Decay	Parameter	Right arrow	Meson	Right Arrow	Energy

Figure 2: Hub papers of each topics are represented with red circles, papers which form the ad-hoc dependencies with blue circles. Respective IDs of blue nodes on the above are 9311274 and 9506257 from left to right, those on below are 9803214 and 9706487 from up to down.

- We present the analysis of real citation graph provided as part of the 2003 KDD Cup [5]. The HEP-Ph (high energy physics phenomenology) citation graph from the e-print arXiv covers all the citations within a dataset of n=34,546 papers with e = 421,578 edges.
- We extract arbitrary 70 papers and the links between them. We did this so that the computational costs of running the combined latent and graphical model remain within reasonable time limits, given the 100 iterations of algorithm for grid search to find a minimized BIC value. Among 70, 43 papers turns out to have either incoming or outgoing edges with each other.
- We use a celebrated text-based topic model, LDA [1], to uncover the topics of each chunks of papers. We perform the analysis with the abstracts of papers, preprocess the text data by following the procedures introduced in paper [6] (Hornik et al), and use the standard VEM method for parameter estimation.

References

[1] Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

[2] Boyd, S., Parikh, N., Chu, E., Peleato, B., and Eckstein, J. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning*, 3(1):1–122.

[3] Candès, E. J., Li, X., Ma, Y., and Wright, J. (2011). Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3):11.

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

[5] Gehrke, J., Ginsparg, P., and Kleinberg, J. (2003). Overview of the 2003 kdd cup. *ACM SIGKDD Explorations Newsletter*, 5(2):149–151.

[6] Hornik, K. and Grün, B. (2011). topicmodels: An r package for fitting topic models. *Journal of Statistical Software*, 40(13):1–30.

[7] Schwarz, G. et al. (1978). Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464.

Computation of Estimator : ADMM

We present each of the three steps of the ADMM algorithm [2] here. 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. Due to the special structure of (4), $x_\alpha^{m+1}, x_M^{m+1}, x_L^{m+1}$, and x_S^{m+1} can be updated separately. More precisely, we have

$$\begin{aligned}
x_\alpha^{m+1}, x_M^{m+1} &= \arg\min_{\alpha, M} f(\alpha, M) = -\frac{\alpha}{n} \sum_{1 \le i < j \le n} X_{ij} - \frac{1}{2n} X \bullet M \\
&\quad + \frac{1}{n} \sum_{1 \le i < j \le n} \log\left(1 + e^{\alpha + M_{ij}}\right) + \frac{1}{2\lambda}[\alpha - (z_\alpha^m - u_\alpha^m)]^2 \\
&\quad + \frac{1}{2\lambda} \|M - (z_M^m - u_M^m)\|_F^2, \\
x_L^{m+1} &= \arg\min_{L \succ 0} \quad \delta \|L\|_* + \frac{1}{2\lambda} \|L - (z_L^m - u_L^m)\|_F^2, \\
x_S^{m+1} &= \arg\min_{S = S^\top} \quad \gamma \|S\|_1 + \frac{1}{2\lambda} \|S - (z_S^m - u_S^m)\|_F^2,
\end{aligned}$$

We can utilize a standard optimization algorithm to update x_M^{m+1}, x_α^{m+1} such as the **BFGS algorithm**. x_L^{m+1} and x_S^{m+1} can also be easily updated through **eigen** and **soft thresholding**, respectively.

Step 2. A closed-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 $\bar{S} = x_S^{m+1} + u_S^m$,

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

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

$$\begin{aligned}
z_\alpha^{m+1} &= \bar{\alpha}, \\
z_M^{m+1} &= \frac{1}{3}\bar{M} + \frac{1}{3}\bar{M}^T + \frac{1}{3}\bar{L} + \frac{1}{3}\bar{S} \\
z_L^{m+1} &= \frac{1}{6}\bar{M} + \frac{1}{6}\bar{M}^T + \frac{2}{3}\bar{L} - \frac{1}{3}\bar{S} \\
z_S^{m+1} &= \frac{1}{6}\bar{M} + \frac{1}{6}\bar{M}^T - \frac{1}{3}\bar{L} + \frac{2}{3}\bar{S}.
\end{aligned}$$

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

Scalability Issue

- The BFGS algorithm used at updating $x_\alpha^{m+1}, x_M^{m+1}$ in the first step of ADMM has a quadruple time complexity, $\mathcal{O}(n^4)$, when n denotes the number of paper.
- **Consensus Algorithm** We decompose a function $f(\alpha, M)$ in the first step of ADMM into n sub-functions, so that they can be solved in parallel fashion by introducing local variables $\alpha_j \in \mathbb{R}$ and a common global variable α, as follows:

$$\begin{aligned}
\underset{\alpha, \alpha_1, \ldots, \alpha_n, M_1, \ldots, M_n}{\text{minimize}} \quad & \sum_{j=1}^n f_j(\alpha_j, M_j) \\
\text{subject to} \quad & \alpha = \alpha_j, \; j = 1, \ldots, n.
\end{aligned}$$

- The consensus algorithm's time complexity of our case is $\mathcal{O}(kn^2)$, where k denotes the number of iterations for the algorithm to be converged. If we warm-start α_j-updates with α and dual variable obtained from previously converged ADMM, the k decreases fast as ADMM in section 6 iterates. [2]

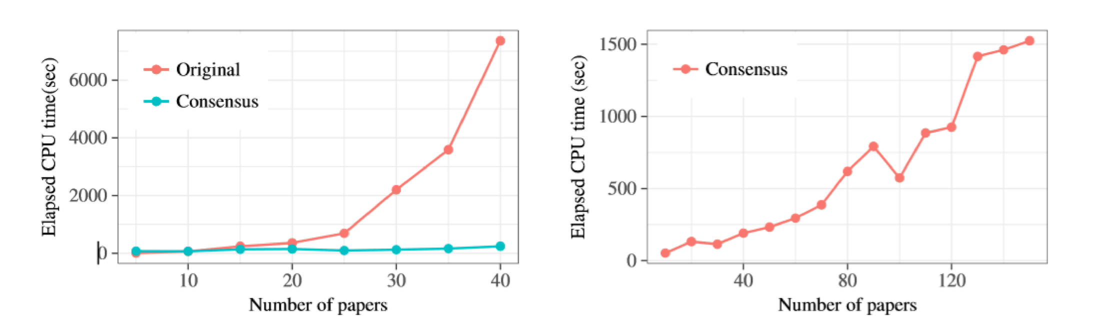

Figure 3: (Left) Comparison of CPU time taken to solve the Case 1 while increasing the total number of papers from 5 to 40 with 5 equal interval. (Right) CPU time taken to solve the Case 1 using consensus method varying the total number of papers from 10 to 150 with 10 equal interval.