Detecting Overlapping and Correlated Communities without Pure Nodes: Identifiability and Algorithm

Kejun Huang
University of Florida

Xiao Fu
Oregon State University

International Conference on Machine Learning 2019

Mixed-membership Stochastic Blockmodel

MMSB [Airoldi et al., 2008]

- Given a graph adjacency matrix $\boldsymbol{A}$
- An edge is present/absent follows Bernoulli

$$\Pr(A_{ij} = \{0, 1\}) = P_{ij}^{A_{ij}} (1 - P_{ij})^{1 - A_{ij}}$$

- $\boldsymbol{P} = \boldsymbol{M}^\top \boldsymbol{B} \boldsymbol{M}$: $\quad \boldsymbol{B} \in [0, 1]^{k \times k}$ community interaction
  $\boldsymbol{m}_i \in \Delta = \{\boldsymbol{x} : \boldsymbol{x} \geq 0, \boldsymbol{1}^\top \boldsymbol{x} = 1\}$ mixed-membership of node i

★ Task: Uniquely identify (part of) $\boldsymbol{M}$ from data $\boldsymbol{A}$

★ Challenges: **identifiability & scalability**

2nd-order Graph Moment

inspired by Anandkumar et al. [2014]

- Divide the network into three sets of nodes $\mathcal{S}_0$, $\mathcal{S}_1$, and $\mathcal{S}_2$
  - $\mathcal{S}_2$: n nodes interested in finding their memberships
  - $\mathcal{S}_1$: $k-1$ nodes
  - $\mathcal{S}_0$: all the other nodes to act as 2-star samples

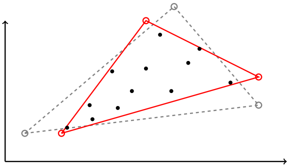
- $\widehat{Y}_{i_1 i_2} = \frac{1}{|\mathcal{S}_0|} \sum_{i_0 \in \mathcal{S}_0} A_{i_0 i_1} A_{i_0 i_2} \quad i_1 \in \mathcal{S}_1 \;\; i_2 \in \mathcal{S}_2$
- $Y_{i_1 i_2} = \mathrm{E}[\widehat{Y}_{i_1 i_2}] = \boldsymbol{m}_{i_1}^\top \boldsymbol{B}^\top \left(\frac{1}{|\mathcal{S}_0|} \sum_{i_0 \in \mathcal{S}_0} \boldsymbol{m}_{i_0} \boldsymbol{m}_{i_0}^\top \right) \boldsymbol{B} \boldsymbol{m}_{i_2}$
- Let $\boldsymbol{\Sigma} = \mathrm{E}[\boldsymbol{m}_{i_0} \boldsymbol{m}_{i_0}^\top]$ and $|\mathcal{S}_0| \to \infty$, then $\widehat{\boldsymbol{Y}} \to \boldsymbol{M}_1^\top \boldsymbol{B}^\top \boldsymbol{\Sigma} \boldsymbol{B} \boldsymbol{M}_2$

$$\boldsymbol{Y} = \boldsymbol{\Xi} \boldsymbol{M}_2$$

★ Can we uniquely recover $\boldsymbol{M}_2 \in \Delta^n$ from $\boldsymbol{Y} \in \mathbb{R}^{(k-1)\times n}$?

Geometric Interpretation

$$\boldsymbol{y}_{i_2} = \boldsymbol{\Xi}\boldsymbol{m}_{i_2} = \sum_{j=1}^{k} \boldsymbol{\xi}_j m_{ji_2} \qquad \boldsymbol{m}_{i_2} \in \Delta$$



- $\boldsymbol{y}_{i_2}$ is a **convex combination** of $\boldsymbol{\xi}_1, ..., \boldsymbol{\xi}_k$
- $\boldsymbol{y}_{i_2}$ belongs to the **convex hull** of $\boldsymbol{\xi}_1, ..., \boldsymbol{\xi}_k$
- There are infinitely many enclosing simplexes

★ Intuition: Find the one with **minimum volume**

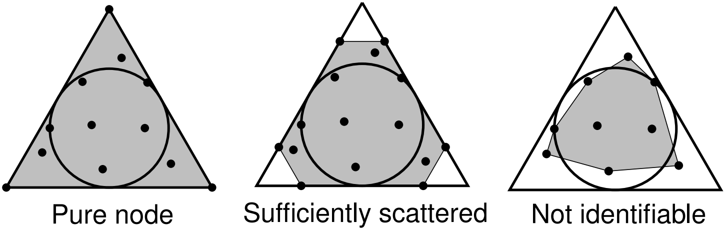
$$\begin{aligned} &\underset{\boldsymbol{\Xi}, \boldsymbol{M}_2}{\text{minimize}} \quad \frac{1}{(k-1)!}\left| \det \left[\boldsymbol{\xi}_1 - \boldsymbol{\xi}_k \quad \cdots \quad \boldsymbol{\xi}_{k-1} - \boldsymbol{\xi}_k \right] \right| \\ &\text{subject to} \;\; \boldsymbol{Y} = \boldsymbol{\Xi}\boldsymbol{M}_2, \;\; \boldsymbol{M}_2 \geq 0, \;\; \boldsymbol{1}^\top \boldsymbol{M}_2 = \boldsymbol{1}. \end{aligned}$$

Identifiability

Definition: Sufficiently Scattered (informal)

Let $\mathcal{D}$ be a "hyper-disc" on the hyperplane $\boldsymbol{1}^\top \boldsymbol{x} = 1$ defined as $\mathcal{D} = \{\boldsymbol{x} \in \mathbb{R}^k : \|\boldsymbol{x}\|^2 \leq \frac{1}{k-1}, \boldsymbol{1}^\top \boldsymbol{x} = 1\}$. A matrix $\boldsymbol{M}$, with all its columns in Δ, is called **sufficiently scattered** if $\mathcal{D} \subseteq \operatorname{conv}(\boldsymbol{M})$.

[Huang et al., 2014, 2016, 2018]



Identifiability

- Equivalently, define $\widetilde{\boldsymbol{Y}} = \begin{bmatrix} \boldsymbol{Y} \\ \boldsymbol{1}^\top \end{bmatrix}, \quad \widetilde{\boldsymbol{\Xi}} = \begin{bmatrix} \boldsymbol{\Xi} \\ \boldsymbol{1}^\top \end{bmatrix},$

$$\begin{aligned} &\underset{\widetilde{\boldsymbol{\Xi}}, \boldsymbol{M}_2}{\text{minimize}} \quad \left| \det \widetilde{\boldsymbol{\Xi}} \right| \\ &\text{subject to} \quad \widetilde{\boldsymbol{Y}} = \widetilde{\boldsymbol{\Xi}} \boldsymbol{M}_2, \boldsymbol{M}_2 \geq 0, \boldsymbol{e}_k^\top \widetilde{\boldsymbol{\Xi}} = \boldsymbol{1}^\top. \end{aligned} \tag{\$}$$

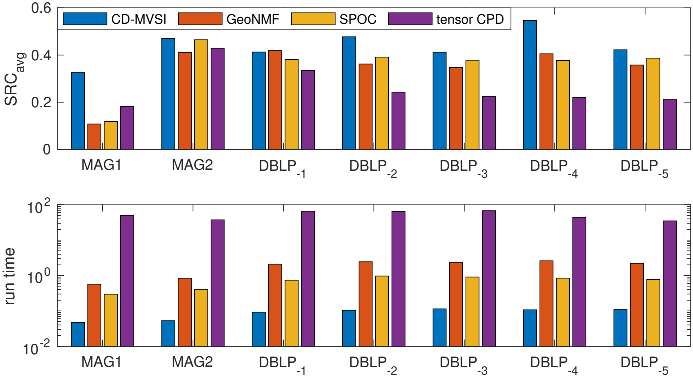
Theorem [Fu et al., 2015, Lin et al., 2015]

Suppose $\boldsymbol{Y} = \boldsymbol{\Xi}^\natural \boldsymbol{M}_2^\natural$, where $\mathrm{rank}(\widetilde{\boldsymbol{\Xi}}^\natural) = k$ and $\boldsymbol{M}_2^\natural \in \Delta^n$ is **sufficiently scattered**. Let $(\boldsymbol{M}_\star, \boldsymbol{\Xi}_\star)$ be an optimal solution for (\$), then there exists a permutation matrix $\boldsymbol{\Pi} \in \mathbb{R}^{k \times k}$ such that

$$\boldsymbol{M}_2^\natural = \boldsymbol{\Pi} \boldsymbol{M}_\star, \qquad \widetilde{\boldsymbol{\Xi}}^\natural = \boldsymbol{\Xi}_\star \boldsymbol{\Pi}^\top.$$

Experiment

- Data sets:
  - Coauthorship data from Microsoft Academic Graph (MAG) and DBLP [Mao et al., 2017]
  - Groundtruth community: “field of study” in MAG and venues in DBLP



References I

Edoardo M Airoldi, David M Blei, Stephen E Fienberg, and Eric P Xing. Mixed membership stochastic blockmodels. *Journal of Machine Learning Research*, 9: 1981–2014, 2008.

Animashree Anandkumar, Rong Ge, Daniel Hsu, and Sham M Kakade. A tensor approach to learning mixed membership community models. *Journal of Machine Learning Research*, 15(1):2239–2312, 2014.

Xiao Fu, Wing-Kin Ma, Kejun Huang, and Nicholas D Sidiropoulos. Blind separation of quasi-stationary sources: Exploiting convex geometry in covariance domain. *IEEE Transactions on Signal Processing*, 63(9), 2015.

Kejun Huang, Nicholas D Sidiropoulos, and Ananthram Swami. Non-negative matrix factorization revisited: Uniqueness and algorithm for symmetric decomposition. *IEEE Transactions on Signal Processing*, 62(1):211–224, 2014.

Kejun Huang, Xiao Fu, and Nikolaos D Sidiropoulos. Anchor-free correlated topic modeling: Identifiability and algorithm. In *Advances in Neural Information Processing Systems*, pages 1786–1794, 2016.

Kejun Huang, Xiao Fu, and Nicholas Sidiropoulos. Learning hidden Markov models from pairwise co-occurrences with application to topic modeling. In *International Conference on Machine Learning*, pages 2068–2077. PMLR, 2018.

References II

Chia-Hsiang Lin, Wing-Kin Ma, Wei-Chiang Li, Chong-Yung Chi, and ArulMurugan Ambikapathi. Identifiability of the simplex volume minimization criterion for blind hyperspectral unmixing: The no-pure-pixel case. *IEEE Transactions on Geoscience and Remote Sensing*, 53(10):5530–5546, 2015.

Xueyu Mao, Purnamrita Sarkar, and Deepayan Chakrabarti. On mixed memberships and symmetric nonnegative matrix factorizations. In *International Conference on Machine Learning*, pages 2324–2333, 2017.

Krzysztof Nowicki and Tom A B Snijders. Estimation and prediction for stochastic blockstructures. *Journal of the American Statistical Association*, 96(455): 1077–1087, 2001.

Tom AB Snijders and Krzysztof Nowicki. Estimation and prediction for stochastic blockmodels for graphs with latent block structure. *Journal of Classification*, 14(1): 75–100, 1997.