Degree-corrected block model

Jiaxin Hu

July 6, 2021

1 Main results

Models

Consider adjacency matrix $A \in \{0,1\}^{n\times n}$, where $A = A^T$ and the n nodes belongs to k communities. Let $z = (z(1), ..., z(n))^T \in [k]^n$ denote the label vector, $\theta = (\theta_1, ..., \theta_n) \in \mathbb{R}^n$ denote the degree correction parameters, and $B \in [0,1]^{k\times k}$ denotes the connection between k communities. Then we assume

$$A_{ij} = A_{ji} \sim_{ind} Ber(\theta_i\theta_j B_{z(i)z(j)}), \quad A_{ii} = 0, \quad \text{for all} \quad i \neq j \in [n].$$

Let $n_u = \sum_{i=1}^n \mathbf{1}\{z(i) = u\}$ denote the size of u-th community, and $P = \mathbb{E}[A] \in [0,1]^{n\times n}$. Consider the parameter space

$$\begin{aligned}\mathcal{P}_n(\theta, p, q, k, \beta, \delta) = \Big\{ & P \in [0,1]^{n\times n} : \text{ there exists a } z \in [k]^n \text{ and } B = B^T \in \mathbb{R}^{k\times k} \\ & s.t. \quad P_{ii} = 0, P_{ij} = \theta_i\theta_j B_{z(i),z(j)}, \quad i \neq j \in [n], \\ & \frac{1}{n_u}\sum_{z(i)=u} \theta_i \in [1-\delta, 1+\delta], \quad u \in [k], \\ & \max_{u\neq v} B_{uv} \leq q < p \leq \min_u B_{uu}, \\ & \frac{n}{\beta k} - 1 \leq n_u \leq \frac{\beta n}{k} + 1, \quad u \in [k] \Big\}.\end{aligned}$$

Consider the loss

$$\ell(\hat{z}, z) = \frac{1}{n}\min_{\pi\in\prod_k} H(\hat{z}, \pi(z)),$$

where $H(z_1, z_2) = \sum_{i\in[n]} \mathbf{1}\{z_1(i) \neq z_2(i)\}$.

Remark 1. There are three constrains in the parameter space need to be noticed: 1) the average of degree-corrected parameters θ_i is around 1; 2) gap between in-group and between-group connection is strictly positive; 3) size of each community is of order $\mathcal{O}(n)$. Correspondingly, in precision model, 1) the average of degree-corrected parameters u_k is around 0; 2) gap between groups is defined by the angle between factor matrices, i.e., $\cos(\Theta_a, \Theta_{a'}) < \delta$; 3) the size of each community $|I_r|$ is of

order $\mathcal{O}(K)$. Note that we do not assume the third condition in previous proof, and we may add this constraint in future. In addition, we need the bounded singular value in precision model since the community model bounds the entries in B.

Theory

Consider the estimator

$$\hat{z} = \arg\max_{z' \in \mathcal{P}_n} \mathcal{L}(z', B_0), \tag{1}$$

where $\mathcal{L}(z, B)$ is the likelihood with given degree-corrected parameters θ, B, and $B_{0,uu} = p, B_{0,uv} = q, u \neq v$.

Remark 2. In the model there are three blocks of parameters: z, θ, B. Obviously, $\hat{z}$ is not necessary the MLE since $\hat{z}$ is the optimal membership vector with given (which can be the hardest) degree-corrected parameters and the hardest community connection, B_0 whose gap between in-group and between-group is $p - q$. Note that the MLE of z' outperforms than $\hat{z}$ because $\hat{z}$ is the estimator under the worst θ and B. In precision model, we may also consider the membership $\hat{M}$ under the hardest U and Θ_r, e.g, $\frac{K}{\beta R} \leq |I_r| \leq \frac{\beta K}{R}$.

Theorem 1.1. *Consider the quantity*

$$\exp(-I) = \frac{1}{n} \sum_{i=1}^{n} \exp\left(-\theta_i \frac{n}{\beta k} (\sqrt{p} - \sqrt{q})^2 \right).$$

Consider any sequence $\{\mathcal{P}_n\}_{n=1}^{\infty}$ *such that* $n \to \infty, I \to \infty, p > q, \|\theta\|_\infty = o(n/k), \min_{i \in [n]} \theta_i \geq c, \log k = o(\min\{I, \log n\})$, *and* $\beta \in [1, \sqrt{5/3}]$. *Then, the estimator* (1) *satisfies*

$$\sup_{\mathcal{P}_n} \mathbb{E}[\ell(\hat{z}, z)] \leq \exp(-I).$$

as $n \to \infty$.

Remark 3. The theorem add additional assumption on the upper and lower bound of θ_i. In precision model, we also assume similar condition, i.e., $m < |u_k| < \sqrt{Mn}$, where the upper bound follows by the assumption that $\sum_{k \in I_r} \|u_k\|^2 = M$. For the convergence, the ideal rate corresponding to $\exp(-I)$ looks like

$$\ell(\hat{M}, M) \leq \exp(-\frac{K}{R} c(m, \tau_1, \tau_2, \beta, \delta^2)) g(n),$$

where $c(m, \tau_1, \tau_2, \beta, \delta^2)$ denote some constant related to the parameters $m, \tau_1, \tau_2, \beta, \delta^2$, and $g(n)$ is a decreasing function of n.

Proof Sketch

The exponential rate mainly relies on the following key step

$$\begin{aligned}
\mathbb{P}(n\ell(\hat{z},z)=m) &\le \sum_{\tilde{z}:|\Gamma|=m} \mathbb{P}(L(\tilde{z})>L(z)) \\
&\le \sum_{\tilde{z}:|\Gamma|=m} \prod_{i\neq j,\tilde{Y}_{ij}\neq Y_{ij}} \exp\left(-\frac{1}{4}\theta_i\theta_j(\sqrt{p}-\sqrt{q})^2\right) \\
&= \sum_{\tilde{z}:|\Gamma|=m} \exp\left(-\frac{1}{2}\sum_{i\in\Gamma}\theta_i\left[\sum_{j:\ i\neq j,\ \tilde{Y}_{ij}\neq Y_{ij}}\theta_j\right](\sqrt{p}-\sqrt{q})^2\right) \\
&= \sum_{\tilde{z}:|\Gamma|=m} \exp\left(-\frac{1}{2}\sum_{i\in\Gamma}\theta_i\left[\sum_{j\neq i}\theta_j-\sum_{j\in\Gamma}\theta_j\right](\sqrt{p}-\sqrt{q})^2\right) \\
&= \sum_{\tilde{z}:|\Gamma|=m} \prod_{i\in\Gamma} \exp\left(-\frac{1}{2}\theta_i\left[(1-\delta)2n_{\min}-\sum_{j\in\Gamma}\theta_j\right](\sqrt{p}-\sqrt{q})^2\right),
\end{aligned}$$

where $\Gamma=\bigcup_{u\in[k]}\{i\in[n],z(i)=u,\hat{z}(i)\neq u\}$, $Y_{ij}=\mathbf{1}\{z(i)=z(j)\}$, $n_{\min}=\frac{n}{\beta k}$ and $\sum_{j\in\Gamma}\theta_j=\mathcal{O}(n_{\min})$ by the assumption.

Remark 4. The key step implies the probability $\mathbb{P}(n\ell(\hat{z},z)=m)=\mathcal{O}\left(\left[\exp(-\frac{n}{\beta k})\right]^m\right)$ and the n inside $\exp(\cdot)$ comes from the sum of degree-corrected parameters $\theta_i\theta_j$ of the misclassified pair (i,j). I think we may meet two difficulties in precision model: 1) the sum of degree-corrected parameter is around 0 due to the intercept, while the community detection is more similar to the case without intercept; 2) community detection uses the connections of $n(n-1)/2$ pairs (i,j) to find the clustering of n nodes while precision model uses K precision matrices find the clustering of K categories. In this sense, biclustering has better performance than marginal clustering. So, if the average of u_k in is not 0, we still may not have K inside the $\exp(\cdot)$.

References