Asymptotic results for the stochastic block model

Jan van Waaij

In collaboration with

Bas Kleijn

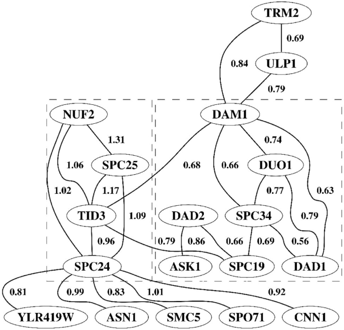
https://github.com/Jan-van-Waaij/Statistical_seminar_HU_Berlin

Community detection on networks

- More present than ever... social networks, protein-protein networks...
- A lot of attention... Abbe (2017), "Community detection and stochastic block models: recent developments"

Example

Protein-protein networks. Protein's interact, and find communities of interacting protein's, Chen and Yuan (2006).



The stochastic block model

Consider a graph of $2n$ nodes, partitioned in two sets of size n.

$$\Pr(\text{edge between i and j}) = \begin{cases} p_n & \text{i and j same partition,} \\ q_n & \text{i and j different partition.} \end{cases}$$

Parameter space Set of labels $\Theta \ni \theta = (\theta_1, \ldots, \theta_{2n})$,
$\theta_i \in \{0, 1\}$, $\sum_i \theta_i = n$.
θ and $\neg\theta = (1 - \theta_1, \ldots, 1 - \theta_{2n})$ give rise to the same partition.
Define $\boldsymbol{\Theta} = \Theta / \sim$, $\theta \sim \eta :\Leftrightarrow \theta = \eta$ or $\theta = \neg\eta$.

Statistical problem

- Given a graph of $2n$ nodes and its random edges, can we determine the communities?
- Of course impossible when $p_n = q_n$.
- When $p_n - q_n > c$ is constant, then it's easy.
- E.g. Facebook, giant graph > 2 billion, average number of friends $155 \approx 7.19 * \log(2'271'000'000)$,

$$\mathbb{P}(\text{probability of being friends}) \approx \frac{7.19 * \log n}{n},$$

where n is the number of users.

Theorem

A graph is connected with probability converging to one, if

$$p_n = \frac{a \log n}{n}, \quad q_n = \frac{b \log n}{n}, \quad , a, b > 0 \textit{ and } a + b > 1.$$

A graph has a connected component of $\mathcal{O}(n)$ with probability converging to one, if

$$p_n = \frac{a}{n}, \quad q_n = \frac{b}{n}, \quad , a, b > 0 \textit{ and } a + b > 1.$$

Different modes of estimation: exact recovery

$$\mathbb{P}_{\theta_0}(\hat{\theta} = \theta_0) \to 1$$

- Only possible if the graph is connected.
- When $0 < c < a_n, b_n < C$ and

$$p_n = \mathbb{P}(\text{edge between vertices of the same class}) = \frac{a_n \log n}{n}$$
$$q_n = \mathbb{P}(\text{edge between vertices of different class}) = \frac{b_n \log n}{n}$$

Then exact recovery is possible if and only if

$$(a_n + b_n - 2\sqrt{a_n b_n} - 1) \log n + \frac{1}{2} \log \log n \to \infty.$$

Different modes of estimation: detection

$$\mathbb{P}(\text{fraction of mismatched labels goes to zero}) \to 1.$$

- Only possible if there is a giant component.
-

$$p_n = \mathbb{P}(\text{edge between vertices of the same class}) = \frac{a_n}{n}$$
$$q_n = \mathbb{P}(\text{edge between vertices of different class}) = \frac{b_n}{n}$$

Then exact recovery is possible if and only if

$$\frac{n(a_n - b_n)^2}{a_n + b_n} \to \infty.$$

Bayesian setup

- Recall, parameter space is set of labels $\Theta \ni \theta = (\theta_1, \ldots, \theta_{2n})$, $\theta_i \in \{0,1\}$, $\sum_i \theta_i = n$. Identify θ and $\neg\theta = (1-\theta_1, \ldots, 1-\theta_{2n})$, $\boldsymbol{\Theta} = \Theta / \sim$.
- Θ has $\binom{2n}{n}$ elements, $\#\boldsymbol{\Theta} = \frac{1}{2}\binom{2n}{n}$.
- Uniform prior, $\pi(\boldsymbol{\theta}) = \frac{1}{\frac{1}{2}\binom{2n}{n}}$.
- Posterior is given by

$$\pi(\theta \mid X) = \frac{p_{\boldsymbol{\theta}}(X)\pi(\boldsymbol{\theta})}{\sum_{\boldsymbol{\eta} \in \boldsymbol{\Theta}} p_{\boldsymbol{\eta}}(X)\pi(\boldsymbol{\eta})} = \frac{p_{\boldsymbol{\theta}}(X)}{\sum_{\boldsymbol{\eta} \in \boldsymbol{\Theta}} p_{\boldsymbol{\eta}}(X)}$$

- Maximum a-posteriori (MAP) estimator same as maximum likelihood estimator (MLE).

Goals

- Establish posterior consistency.
- Confidence sets.

Posterior consistency

Theorem

If $(p_n = \frac{a_n \log n}{n})$ *and* $(q_n = \frac{b_n \log n}{n})$ *are such that*

$$(a_n + b_n - 2\sqrt{a_n b_n} - 2)\log n \to \infty, \ \textit{then } \pi(\theta_0 \mid X) \xrightarrow{P_{\theta_0}} 1.$$

Idea of the proof.

1. Choose $\theta_0 \in \boldsymbol{\theta}_0$. Let $Z_\theta = \{i : \theta_i = 0\}$. Let $V_k = \{\theta : \#(Z_{\theta_0} \setminus Z_\theta) = k\}$. V_k has $\binom{n}{k}^2$ elements.
2. For every $\theta \in V_k$ there is a test of power $a_{n,k}$,

$$\mathbb{P}_{\theta_0}\phi(X) + \mathbb{P}_\theta(1 - \phi(X)) \leq a_{n,k}.$$

3. $\mathbb{P}_{\theta_0}\Pi(\boldsymbol{\theta} \neq \boldsymbol{\theta}_0 \mid X) \leq \sum_{k=1}^{\lfloor n/2 \rfloor} \binom{n}{k}^2 a_{k,n} = \mathbf{o}(1).$

Credible sets turn out to be confidence sets

Definition

A $1-\alpha$ credible set, is any measurable set $D(X) \subset \boldsymbol{\Theta}$ with $\Pi(D(X) \mid X) \geq 1-\alpha$.

Definition

A $1-\alpha$ confidence set is any measurable set $C(X) \subset \Theta$ with $P_{\theta_0}(\theta_0 \in C(X)) \geq 1-\alpha$.

Theorem

When exact recovery holds, $1-\alpha$-credible sets, are $1-3\alpha$ confidence sets.

Proof.

See blackboard. $\square$