Estimation and Clustering in the Dynamic Stochastic Block Model

Marianna Pensky
Department of Mathematics
University of Central Florida, USA

Meeting in Mathematical Statistics
Lumini, December 2017

Recipe for success (Lepski and Tsybakov)

Recipe for success (Lepski and Tsybakov)

If your estimator sucks,
You need to do the minimax.
First, you chhose unknown theta
Using **Oleg Lepski method**
If your upper bounds are tight
You have a reason for delight.
For lower bounds
You need to look
With no doubts
Into **Sasha's book**.
Matching the bounds more or less
Will ensure your success.

Happy birthday!!

Stochastic Block Model (SBM)

Network $=$ undirected graph with n nodes

Observations:
$\mathbf{B}_{i,j} \sim$ Bernoulli$(\mathbf{\Theta}_{i,j})$, $1 \leq i < j \leq n$
$\mathbf{B}_{i,j} = \mathbf{B}_{j,i}$, $\mathbf{B}_{i,i} = 0$, $\mathbf{B}_{i,j}$ are independent for $1 \leq i < j \leq n$

Nodes are grouped into m classes $\Omega_1, \cdots, \Omega_m$

Probability of a connection $\mathbf{\Theta}_{i,j}$ is entirely determined to which groups the nodes i and j belong:
$\mathbf{\Theta}_{i,j} = \mathbf{G}_{k,k'} \quad \text{if} \quad i \in \Omega_k,\ j \in \Omega_{k'}$

Problems:
a) estimate matrix $\mathbf{\Theta}$
b) cluster the nodes into classes $\Omega_1, \cdots, \Omega_m$

Vast literature in the last 10-15 years

Stochastic Block Model

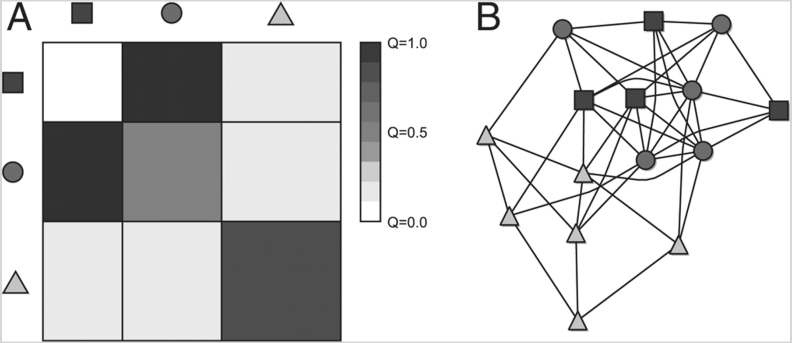

Figure: Stochastic Block Model with 3 blocks. Left panel: matrix **G**. Right panel: undirected graph with 15 nodes

Related problem: graphon estimation

Consider a collection of latent random variables $\zeta_1, \cdots, \zeta_n \in [0,1]^n$

Consider a symmetric function $f : [0,1]^2 \to [0,1]$

$f(x,y) = f(y,x)$, $f(x,x) = 0$

Model: $\mathbf{\Theta}_{i,j} = f(\zeta_i, \zeta_j)$, $i,j = 1, \cdots, n$

$\mathbf{\Theta}_{i,j}$ is **identifiable** up to re-labeling of the variables
f is **invariable** with respect to any Lebesgue measure-preserving bijection
$\mu : [0,1] \to [0,1]$, so that $f(\mu(x), \mu(y)) = f(x,y)$
Consider **equivalence classes** of graphons

Assumption: f is smooth

Objective: estimate generating function f and matrix $\mathbf{\Theta}$

However, everything in the world exists in time

Dynamic Stochastic Block Model (DSBM)

Network $=$ undirected graph with n nodes

Network is observed at L time instances $t_1, t_2, \cdots, t_L \in [0, T]$
For simplicity: $T = 1,\ t_l = l/L,\ l = 1, \cdots, L$

Observations:
$\mathbf{A}_{i,j,l} \sim$ Bernoulli$(\mathbf{\Theta}_{i,j,l})$, $1 \le i < j \le n,\ l = 1, \cdots, L$
$\mathbf{A}_{i,j,l} = \mathbf{A}_{j,i,l},\ \mathbf{A}_{i,i,l} = 0$
$\mathbf{A}_{i,j,l}$ are independent for $1 \le i < j \le n,\ l = 1, \cdots, L$

Nodes are grouped into m classes $\Omega_1, \cdots, \Omega_m$

$\mathbf{G}$ is the connectivity tensor:
$\mathbf{G}_{k,k',l} = \mathbf{G}_{k',k,l} = \mathbf{\Theta}_{i,j,l}$ if $i \in \Omega_k$ and $j \in \Omega_{k'}$

Estimation problem

Objective: estimate the tensor $\mathbf{\Theta} \in \mathbb{R}^{n \times m \times L}$

Assume some continuity:

1. probabilities $\mathbf{G}_{k,k',l}$ do not change drastically from one time instant to another
2. only few nodes change their memberships from one time point to another

Do not assume: knowledge of the number of classes m

Dynamic graphon estimation

Consider a collection of latent random variables $\zeta_1, \cdots, \zeta_n \in [0,1]^n$

Consider a symmetric function $f : [0,1]^3 \to [0,1]$

$f(x,y,t) = f(y,x,t)$, $f(x,x,t) = 0$

Model: $\mathbf{\Theta}_{i,j,l} = f(\zeta_i, \zeta_j, l/L)$, $i,j = 1, \cdots, n$, $l = 1, \cdots, L$

Assumptions:

- f is smooth in t and piecewise smooth in x and y
- enumeration of the nodes does not change in time

Objective: estimate generating function f and tensor of probabilities $\mathbf{\Theta}$

Existing results: estimation in the static SBM model

- If m is known, and $\boldsymbol{\Theta}$ is not sparse, the minimax rate for estimating $\boldsymbol{\Theta}$ is

$$\frac{1}{n^2}\,\|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}\|_F^2 \asymp \frac{\log m}{n} + \frac{m^2}{n^2}$$

(Gao, Lu and Zhou (AOS, 2015))

- m^2/n^2 is the parametric error: $M = m(m+1)/2$ parameters, $N = n(n-1)/2$ independent observations

  $n^{-1}\log m$ is the clustering error: $n^{-1}\log m \asymp N^{-1}\log(m^n)$ where m^n is the cardinality of the set of clustering matrices

- If $\boldsymbol{\Theta}$ is sparse, so that $\|\boldsymbol{\Theta}\|_\infty \le \rho_n$

$$\frac{1}{n^2}\,\|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}\|_F^2 \asymp \rho_n\left(\frac{\log m}{n} + \frac{m^2}{n^2}\right)$$

(Klopp, Tsybakov, Verzelen (2016))

Existing results: clustering in in the static SBM model

- Under the condition that the lowest eigenvalue of matrix **G** is separated from zero, Lei and Rinaldo (2015) derived clustering errors for an SBM with an arbitrary number of classes (sparse and non-sparse cases)
- Under assumptions that the SBM is balanced: $n_{\max} \asymp n_{\min} \asymp n/m$ and there is a community structure, Gao, Ma, Zhang and Zhou (2017) derived optimal minimax lower and upper bounds the misclassification proportion
- Under similar assumptions, Gao, Ma, Zhang and Zhou (2017) extended their results to the degree-corrected stochastic block model

Existing results: static graphon estimation

Let matrix $\mathbf{\Theta}$ be generated by the graphon f

- If f is in Holder class with a smoothness parameter α and α is known, then

$$\frac{1}{n^2}\,\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}\|_F^2 \asymp \frac{\log n}{n} + n^{-\frac{2\alpha}{\alpha+1}}$$

  (Gao, Lu and Zhou (AOS, 2015))

- Extension to the case where $\mathbf{\Theta}$ is sparse: $\|\mathbf{\Theta}\|_\infty \leq \rho_n$
  f is in Holder class with a smoothness parameter α, α is known
  (Klopp, Tsybakov, Verzelen (AOS, 2016))

- Non-combinatorial distance-based estimation of the graphon with $\alpha = 1$

$$\frac{1}{n^2}\,\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}\|_F^2 = O_P\left(\frac{\log n}{n}\right)^{1/2}$$

  (Zhang, Levina, Zhu (JNPS, 2016))

The model

Let $\mathcal{M}(m,n)$ be the **collection of clustering matrices** $\mathbf{X} \in \{0,1\}^{n\times m}$

$\mathbf{X} \in \mathcal{M}(m,n)$ have exactly one 1 per row
$\mathbf{X}_{ik} = 1$ iff node i belongs to the class Ω_k, $\mathbf{X}_{ik} = 0$ zero otherwise

Data: $\mathbf{A}_{i,j,l} \sim$ Bernoulli$(\mathbf{\Theta}_{i,j,l})$, $1 \le i < j \le n$, $l = 1, \cdots, L$
$\mathbf{A}_{i,j,l} = \mathbf{A}_{j,i,l}$, $\mathbf{A}_{i,i,l} = 0$
$\mathbf{A}_{i,j,l}$ are independent for $1 \le i < j \le n$, $l = 1, \cdots, L$

Model: $\mathbf{\Theta}_{*,*,l} = \mathbf{Z}^{(l)} \mathbf{G}_{*,*,l} (\mathbf{Z}^{(l)})^T$, $l = 1, \cdots, L$
$\mathbf{Z}^{(l)} \in \mathcal{M}(m,n)$ is a clustering matrix at the moment t_l
$\mathbf{G}_{*,*,l}$ is a matrix of block connection probabilities at the moment t_l

Assumptions: smoothness of block probabilities

- Assume that for each (k_1, k_2), $k_1, k_2 = 1, \cdots, m$, vector $\mathbf{G}_{k_1,k_2,l} = (\mathbf{G}_{k_1,k_2,1}, \cdots, \mathbf{G}_{k_1,k_2,L})$ represents values of a **smooth function**, so that vectors $\mathbf{G}_{k_1,k_2,*}$ have **sparse representation** in some orthogonal basis $\mathbf{H} \in \mathbb{R}^{L\times L}$ with $\mathbf{H}^T\mathbf{H} = \mathbf{H}\mathbf{H}^T = \mathbf{I}_L$

  For example, $\mathbf{H}$ is a matrix of the **Fourier** or a **wavelet transform**

- Assume that vectors $\mathbf{HG}_{k_1,k_2,*}$ have **only few large elements**, so that $\mathbf{HG}_{k_1,k_2,*}$ can be approximated using only few of its elements

  Let J be a set of all nonzero elements in $\mathbf{HG}_{k_1,k_2,*}$, $k_1, k_2 = 1, \cdots, m$ necessary for its representation

  Assumption: Cardinality $|J|$ of set J is small

Collection of clustering matrices

Let $\mathcal{C}_{n,m,L}$ be a **set of clustering matrices** such that

$$\mathcal{C}(m, n, L) \subseteq (\mathcal{M}(m, n), \cdots, \mathcal{M}(m, n)).$$

Assume that $\mathbf{C} = (\mathbf{X}^{(1)}, \cdots, \mathbf{X}^{(L)}) \in \mathcal{C}(m, n, L)$ for some m

No specific assumptions on the set of clustering matrices so far

The data summary

- Data is represented by the order 3 tensor $\mathbf{A}_{i,j,l} \sim$ Bernoulli$(\mathbf{\Theta}_{i,j,l})$, $1 \leq i < j \leq n$, $l = 1, \cdots, L$
- The tensor has an underlying structure $\mathbf{\Theta}_{*,*,l} = \mathbf{Z}^{(l)} \mathbf{G}_{*,*,l} (\mathbf{Z}^{(l)})^T$, $l = 1, \cdots, L$ with $\mathbf{G}_{*,*,l}$ representing values of smooth functions
- Data $\mathbf{A}$ and tensor $\mathbf{\Theta}$ are redundant: $\mathbf{A}_{i,j,l} = \mathbf{A}_{j,i,l}$, $\mathbf{A}_{i,i,l} = *$ and $\mathbf{G}_{k,k',l} = \mathbf{G}_{k',k,l}$

Solution: use the **Kronecker product** and **vectorization** to reduce the problem to a **structured regression problem**

The model

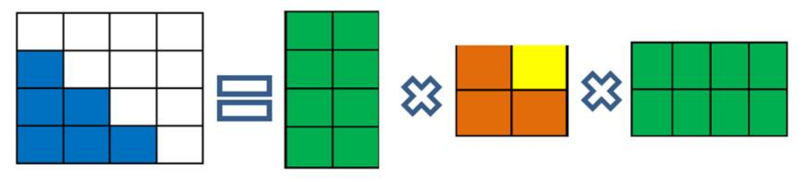

Figure: The model: $\mathbf{\Theta}_{*,*,1} = \mathbf{Z}^{(1)} \mathbf{G}_{*,*,1} (\mathbf{Z}^{(1)})^T$, $n = 4$, $m = 2$

The model: accounting for the symmetry

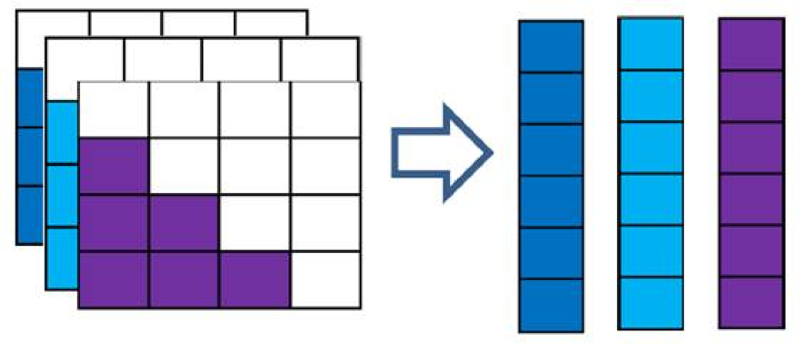

Figure: DSBM model with $n = 4$, $L = 3$

The model: accounting for the redundancy in G

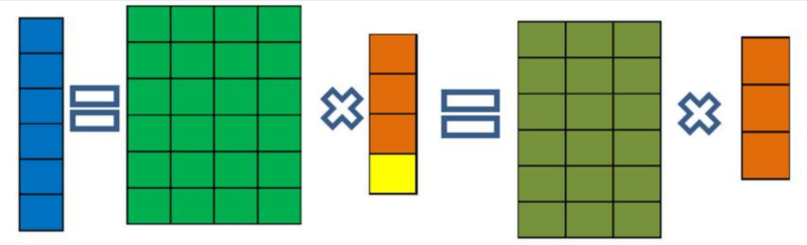

Figure: Removing redundancy in **G**: $n = 4$, $N = n(n-1)/2 = 6$, $m = 2$, $M = m(m+1)/2 = 3$.The dark green matrix is a clustering matrix for N elements and M classes

The model: vectorization summary

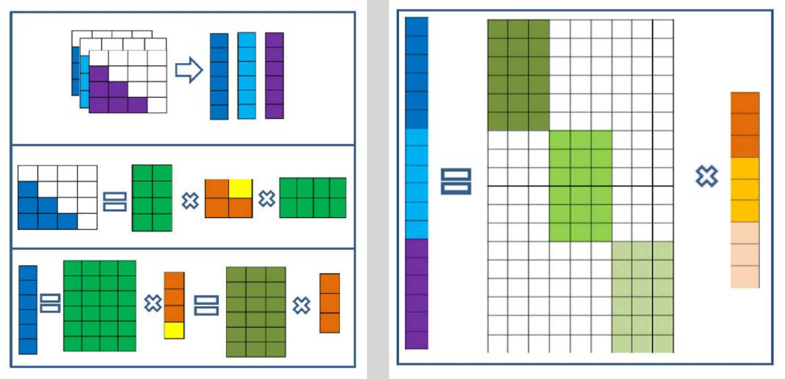

Figure: Vectorization of tensor $\mathbf{\Theta}$ with $n = 4,\ m = 2,\ N = 6,\ M = 3,\ L = 3$. **Left panel, top:** accounting for redundancy in $\mathbf{\Theta}$. **Left panel, middle:** the structure $\mathbf{\Theta}$. **Left panel, bottom:** accounting for redundancy in $\mathbf{G}$. **Right panel:** the final model $\boldsymbol{\theta} = \mathbf{C}\mathbf{q}$.

The final model

Let $N = n(n-1)/2, M = m(m+1)/2$

Let $\boldsymbol{\theta} \in [0,1]^N$ be the vectorized version of the tensor $\boldsymbol{\Theta}$ with the redundancy removed

Let $\mathbf{a} \in [0,1]^N$ be the vectorized version of the tensor $\mathbf{A}$ with the redundancy removed and, hence, **independent elements**

Let $\mathbf{q} \in [0,1]^M$ be the vectorized version of the tensor $\mathbf{G}$ with the redundancy removed

Define a block diagonal matrix $\mathbf{C} \in \{0,1\}^{NL \times ML}$ with blocks $\mathbf{C}^{(l)}$, $l = 1, \cdots, L$, on the diagonal. Each block $\mathbf{C}^{(l)} \in \{0,1\}^{N \times M}$ is a clustering matrix at time t_l

Final model:

$$\mathbf{a} = \boldsymbol{\theta} + \boldsymbol{\xi} \quad \text{with} \quad \boldsymbol{\theta} = \mathbf{C}\mathbf{q} \tag{1}$$

$$\mathbf{a}_i \quad \text{are independent Bernoulli}(\boldsymbol{\theta}_i), \quad i = 1, \cdots, NL$$

If the matrix $\mathbf{C}$ **were known**, then (1) would represent the **standard regression model with independent Bernoulli errors**

The final model with a structure

Recall that vectors $\mathbf{G}_{k_1,k_2,*}$ have **sparse representation** in the orthogonal basis $\mathbf{H} \in \mathbb{R}^{L \times L}$ (e.g., Fourier or wavelet)

Cardinality $|J|$ of set J of all nonzero elements in $\mathbf{HG}_{k_1,k_2,*},\ k_1, k_2 = 1, \cdots, m$ that are necessary for its representation, is small

Denote $\mathbf{W} = (\mathbf{H} \otimes \mathbf{I}_M)$ and observe that $\mathbf{W}$ is an orthogonal matrix: $\mathbf{W}^T\mathbf{W} = \mathbf{W}\mathbf{W}^T = \mathbf{I}_{ML}$

Let $\mathbf{d} = \mathbf{Wq}, \quad \mathbf{d} \in \mathbb{R}^{ML} \Longrightarrow \mathbf{a} = \mathbf{CW}^T\mathbf{d} + \xi$

Let $J = \{j :\ \mathbf{d}_j \neq 0\},\ \mathbf{d}_{J^c} = \mathbf{0}$ be the set necessary for representing $\mathbf{q}:\ \mathbf{q} = \mathbf{W}^T\mathbf{d}$

Cardinality $|J|$ of set J **is small**

Recall that $\mathcal{C}_{n,m,L}$ is a **set of clustering matrices** such that

$$\mathcal{C}(m, n, L) \subseteq (\mathcal{M}(m, n), \cdots, \mathcal{M}(m, n)).$$

Optimization problem

Find $m, J, \mathbf{d}$ and $\mathbf{C}$ as a solution of the penalized least squares optimization problem

$$(\widehat{m}, \widehat{J}, \widehat{\mathbf{d}}, \widehat{\mathbf{C}}) = \operatorname*{argmin}_{m,J,\mathbf{d},\mathbf{C}} \left\{ \|\mathbf{a} - \mathbf{C}\mathbf{W}^T\mathbf{d}\|^2 + \text{Pen}(|J|, m) \right\} \qquad (2)$$

s.t. $1 \le m \le n$, $J \equiv J_M$, $\mathbf{d}_{J^c} = 0$, $\mathbf{C} \in \mathcal{C}(m, n, L)$ with

$$\text{Pen}(|J|, m) = 11 \log(|\mathcal{C}(m, n, L)|) + \frac{11}{2} |J| \log\left(\frac{5\, m^2 L}{|J|}\right).$$

Here $\mathbf{d} \in \mathbb{R}^{ML}$, $\mathbf{W} \in \mathbb{R}^{ML \times ML}$, $M = m(m+1)/2$

Algorithm:

1. Solve the optimization problem separately for every m. Obtain $\widehat{\mathbf{d}}_M, \widehat{\mathbf{C}}_M$ and $\widehat{J}_M$.
2. Select the value $\widehat{M} = \widehat{m}(\widehat{m}+1)/2$ that delivers the minimum in (2), so that

$$\widehat{\mathbf{d}} = \widehat{\mathbf{d}}_{\widehat{M}}, \quad \widehat{\mathbf{C}} = \widehat{\mathbf{C}}_{\widehat{M}}, \quad \widehat{J} = \widehat{J}_{\widehat{M}}. \qquad (3)$$

3. Set $\widehat{\mathbf{W}} = (\mathbf{H} \otimes \mathbf{I}_{\widehat{M}})$ and calculate $\widehat{\mathbf{q}} = \widehat{\mathbf{W}}^T\widehat{\mathbf{d}}$, $\quad \widehat{\boldsymbol{\theta}} = \widehat{\mathbf{C}}\widehat{\mathbf{q}}$
4. Obtain $\widehat{\boldsymbol{\Theta}}$ by packing vector $\widehat{\boldsymbol{\theta}}$ into the tensor and taking the symmetries into account.

An oracle inequality

Consider a **DSBM** with a true matrix of probabilities $\boldsymbol{\Theta}^*$ and estimator $\widehat{\boldsymbol{\Theta}}$ obtained as the tensor version of the vector $\mathbf{C}\mathbf{W}^T\widehat{\mathbf{d}}$. Then, for any $t > 0$, with probability at least $1 - 9e^{-t}$

$$\frac{\|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*\|_F^2}{n^2\, L} \leq \min_{\substack{m,J,\mathbf{q} \\ \mathbf{C} \in \mathcal{C}(m,n,L)}} \left[\frac{6\, \|\mathbf{C}\mathbf{W}^T(\mathbf{W}\mathbf{q})_J - \boldsymbol{\theta}^*\|^2}{n^2\, L} + \frac{4\, \mathrm{Pen}(|J|, m)}{n^2\, L} \right] + \frac{38}{n^2\, L}\, t,$$

where

$$\mathrm{Pen}(|J|, m) = 11\, \log(|\mathcal{C}(m,n,L)|) + \frac{11}{2}\, |J| \log\left(\frac{5\, m^2 L}{|J|}\right).$$

and a **similar result holds for the expectation.**

- $(n^2 L)^{-1}\ \log(|\mathcal{C}(m,n,L)|)$ is **the clustering error**
- $(n^2 L)^{-1}\ |J| \log\left(\frac{5\, m^2 L}{|J|}\right)$ is **the nonparametric estimation error**

So far, we placed no restrictions on the collection of clustering matrices

Stability of clustering in time

Consider a collection $\mathcal{F}(m, n, n_0, L)$ that corresponds to the situation where **at most n_0 nodes can change their memberships from one time instant to another**

$$|\mathcal{F}(m,n,n_0,L)| = m^n \left[\binom{n}{n_0} m^{n_0} \right]^{L-1} \qquad |\mathcal{F}(m,n,0,L)| = m^n$$

If $n_0 = 0$, then the group memberships of the nodes do not change in time.

Then, for any $t > 0$, with probability at least $1 - 9e^{-t}$

$$\frac{\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*\|_F^2}{n^2\, L} \leq Const \min_{\substack{m,J,\mathbf{q} \\ \mathbf{C} \in \mathcal{F}(m,n,n_0,L)}} \left[\frac{\|\mathbf{C}\mathbf{W}^T(\mathbf{W}\mathbf{q})_J - \theta^*\|^2}{n^2\, L} + \frac{|J|}{n^2 L} \log\left(\frac{25 m^2 L}{|J|}\right) + \frac{\log m}{n\, L} + \frac{n_0\, \log(mne/n_0)}{n^2} \right] + \frac{38}{n^2\, L}\, t$$

The lower bounds for the error

$$\inf_{\widehat{\mathbf{\Theta}}} \sup_{\mathbf{\Theta} \in \mathcal{B}} \mathbb{P}_{\mathbf{\Theta}} \left\{ \frac{\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}\|_F^2}{n^2 L} \geq C_\gamma \left(\frac{\log m}{nL} + \frac{n_0 \log\left(\frac{mne}{n_0}\right)}{n^2} + \frac{|J| \log\left(L \min(1, m^2/2|J|)\right)}{n^2 L} \right) \right\} \geq \frac{1}{4}$$

The lower bound coincides with the upper bound up to a constant

- $\frac{\log m}{nL} = \frac{n \log m}{n^2 L}$ is the error due to the **initial clustering**
- $\frac{n_0 \log(mne/n_0)}{n^2}$ is the error due to **changes in the memberships**
- $\frac{|J| \log\left(L \min(1, m^2/2|J|)\right)}{n^2 L}$ is the error of **non-parametric estimation** due to the need of selecting and estimating $|J|$ unknown parameters

Discussion of the estimation results for the DSBM

- The estimator is **adaptive** with respect to the **unknown number of classes**. It is **not adaptive** with respect to the **complexity of the set of possible clustering matrices** which is due to changes in the memberships
- **No probabilistic assumptions on the mechanism generating changes in the memberships.** However, various mechanisms **can be accommodated** by identifying the **cardinality** of a collection of clustering matrices that is likely to occur with high probability.
- If $L = 1$ and $\mathbf{H} = 1$, we **automatically obtain an adaptive estimator of a sparse SBM** where **probabilities of interactions for some classes are equal to zero** (different kind of sparsity than in Klopp *et al.* (2016))
- Can handle the situation where the **number of nodes changes in time**. In this case, denote by n the **maximum number of nodes** that have been in the network over time and create a class with no interaction with all other classes. Place all inactive nodes into this class

Assumptions for clustering

- Assume, as before, that for each $(k_1, k_2),\ k_1, k_2 = 1, \cdots, m$, the vector $\{\mathbf{G}_{k_1,k_2,l}\}_{l=1,\cdots,L}$ represents values of a **smooth function**: $\mathbf{G}_{k_1,k_2,l} = G(l/L; k_1, k_2)$ where $G(t; k_1, k_2)$ **belongs to a Hölder class** $\Sigma(\beta, C_\Sigma)$
- Assume that $\mathbf{C} \in \mathcal{F}(m, n, n_0, L)$, so that **at most n_0 nodes can change their memberships from one time instant to another**
- Assume that there exists an absolute constant C_λ, $1 \le C_\lambda < \infty$, independent of n, l and L such that

$$C_\lambda^{-1} \le \lambda_{\min}(\mathbf{G}_{*,*,l}) \le \lambda_{\max}(\mathbf{G}_{*,*,l}) \le C_\lambda.$$

Spectral clustering algorithm

Input: Adjacency matrices $\mathbf{A}_l$, $l = 1, \ldots, L$; number of communities m ; approximation parameter ϵ .

Output: Membership matrices $\mathbf{Z}_l$ for any $l = 1, \ldots, L$.

Steps:

1: For each l, $l = 1, \cdots, L$:

2: Estimate $\mathbf{\Theta}_l$ by the discrete kernel estimator $\widehat{\mathbf{\Theta}}_{l,r}$ described later.

3: Let $\mathbf{U}_l \in \mathbb{R}^{n \times m}$ be a matrix representing the first m eigenvectors of $\widehat{\mathbf{\Theta}}_{l,r}$.

4: Apply the $(1+\epsilon)$-approximate k-means algorithm to the row vectors of $\mathbf{U}_l$

5: Obtain the solution $\widehat{\mathbf{Z}}_l$.

Discrete kernel estimator

Let $r < L/2$ be **the kernel width parameter**
Consider three pairs of sets of integers

$$\begin{aligned} \mathcal{T}_{r,1} &= \{-r, \cdots, r\}, \quad \mathcal{D}_{r,1} = \{1+r, \cdots, L-r\}; \\ \mathcal{T}_{r,2} &= \{0, \cdots, r\}, \quad \mathcal{D}_{r,2} = \{1, \cdots, r\}; \\ \mathcal{T}_{r,3} &= \{-r, \cdots, 0\}, \quad \mathcal{D}_{r,3} = \{L-r+1, \cdots, L\}. \end{aligned}$$

Introduce **discrete kernel functions** $W_{r,l}^{(j)}(i)$ of an integer argument i such that $|W_{r,l}^{(j)}(i)| \leq W_{\max}$, where $W_{\max}$ is independent of l, j and i, and for $j = 1, 2, 3$,

$$\frac{1}{|\mathcal{T}_{r,j}|} \sum_{i \in \mathcal{T}_{r,j}} i^k \, W_{r,l}^{(j)}(i) = \begin{cases} 1, & \text{if } k = 0, \\ 0, & \text{if } k = 1, \ldots, l. \end{cases}$$

If $l \in \mathcal{D}_{r,j}$, we construct an estimator of $\boldsymbol{\Theta}_l$ on the basis of $\mathbf{A}_{l+i}$, $i \in \mathcal{T}_{r,j}$, $j = 1, 2, 3.$ Here $|\mathcal{T}_{r,j}|$ is the cardinality of the set $\mathcal{T}_{r,j}$.

Adaptive discrete kernel estimator

The **optimal value** r^* of r depends on **the unknown values** of **the number of membership changes** n_0, **the size of the largest block** $n_{\max}$ and **the parameter of Hölder space** β

Therefore, in practice, the value r^* is **unavailable.**

Use **Lepskii method for construction of an adaptive estimator**
For any l, set

$$\widehat{r} \equiv \widehat{r}_l = \max \left\{ 0 \leq l \leq L/2 : \ \|\widehat{\boldsymbol{\Theta}}_{l,r} - \widehat{\boldsymbol{\Theta}}_{l,r'}\| \leq 4\, C_0 \sqrt{n/(r' \vee 1)} \quad \text{for any} \quad r' < r \right\}$$

Here C_0 is a constant specified in the paper

The clustering error measures

Clustering of the nodes can be recovered only **up to column permutations.** Denote the set of $m \times m$ permutation matrices by $\mathcal{E}_m$. Assume that the node's labels are fixed and do not depend on l.

1. Two **local measures of clustering accuracy** at time $t_l = l/L$:

a) **the highest relative clustering error** over the communities

$$\tilde{R}_l(\widehat{\mathbf{Z}}_l, \mathbf{Z}_l) = \min_{\mathbf{J} \in \mathcal{E}_m} \max_{1 \le k \le m} n_k^{-1} \, \|(\widehat{\mathbf{Z}}_l \mathbf{J} - \mathbf{Z}_l)_{\Omega_{l,k},*}\|_0,$$

where $\Omega_{l,k}$ is the k-th community at time t_l.

b) **the overall relative clustering error**

$$R_l(\widehat{\mathbf{Z}}_l, \mathbf{Z}_l) = n^{-1} \min_{\mathbf{J} \in \mathcal{E}_m} \|\widehat{\mathbf{Z}}_l \mathbf{J} - \mathbf{Z}_l\|_0$$

2. Two **global measures of clustering accuracy:**

a) **the highest relative error** over the communities and b) **the overall highest relative error**

$$\tilde{R}_{\max} = \max_{1 \le l \le L} \tilde{R}_l(\widehat{\mathbf{Z}}_l, \mathbf{Z}_l), \quad R_{\max} = \max_{1 \le l \le L} R_l(\widehat{\mathbf{Z}}_l, \mathbf{Z}_l)$$

The clustering accuracy

Let $L \leq n^{\tau_1}$ for some $\tau_1 < \infty$. Then, for any $\tau > 0$, **with probability at least** $1 - 4\, n^{-\tau}$, one has

$$\tilde{R}_l(\widehat{\mathbf{Z}}_l, \mathbf{Z}_l) \leq C_1(2+\epsilon)\, \frac{m\, n}{n_{\min}^2}\, \min\left(1; \left(\frac{n}{L^{2\beta}}\right)^{\frac{1}{2\beta+1}} + \sqrt{\frac{n_{\max}\, n_0}{n}}\right).$$

In addition, **with probability at least** $1 - 4\, n^{-(\tau-\tau_1)}$, one has

$$\tilde{R}_{\max} \leq C_2(2+\epsilon)\, \frac{m\, n}{n_{\min}^2}\, \min\left(1; \left(\frac{n}{L^{2\beta}}\right)^{\frac{1}{2\beta+1}} + \sqrt{\frac{n_{\max}\, n_0}{n}}\right).$$

Moreover,

$$R_l(\widehat{\mathbf{Z}}_l, \mathbf{Z}_l) \leq \tilde{R}_l(\widehat{\mathbf{Z}}_l, \mathbf{Z}_l)\, \frac{n_{\max}}{n}, \quad R_{\max} \leq \tilde{R}_{\max}\, \frac{n_{\max}}{n}$$

Here, constants C_1 and C_2 depend on C_λ, the parameters of the kernel W and the parameters of the Hölder class $\Sigma(\beta, C_\Sigma)$

Discussion of the clustering results for the DSBM

- The estimator is **adaptive** to the Hölder class parameter β and to the number of membership changes between two consecutive time points n_0
- The case when connection probabilities are uniformly small **(sparse case)** is studied in the paper (not in the talk).
- Although the procedure **requires knowledge of the number of clusters** m, one can run the procedure for a set of values of m and then choose m by using penalty in the estimation model

Time-dependent graphon

Consider a collection of latent random variables $\zeta_1, \cdots, \zeta_n$ and a symmetric function $f : [0,1]^3 \to [0,1]$ such that $f(x,y,t) = f(y,x,t)$

The model: $\boldsymbol{\Theta}_{i,j,l} = f(\zeta_i, \zeta_j, l/L)$, $i,j = 1, \cdots, n$, $l = 1, \cdots, L$
$f(x,y,t)$ is the **dynamic graphon**

The identifiability: for any f and any measure-preserving bijection $\mu : [0,1] \to [0,1]$, the functions $f(x,y,t)$ and $f(\mu(x), \mu(y), t)$ define the **same probability distribution** on random graphs

1. consider **equivalence classes of graphons**
2. **no label switching in time:** μ is independent of t

Let $\mathbf{H}$ be an orthogonal transform. Denote by $\mathbf{v}(\zeta_i, \zeta_j, \cdot) = \mathbf{H}\boldsymbol{\Theta}_{i,j,*}$ the transformation of vector $\boldsymbol{\Theta}_{i,j,*}$ along the time dimension

Since $\zeta_1, \cdots, \zeta_n$ are independent of l, **the estimation algorithm can be simplified**

Assumptions

Assumption: f is smooth in t and piecewise smooth in x and y

Assumptions

There exist an integer κ, some $\nu_1, \nu_2, C_{01}, C_{02} > 0$, and $0 = x_0 < x_1 < \cdots < x_\kappa = 1$ such that for any $x, x' \in [0,1]$ and $y, y' \in [0,1]$ and any $i, j = 1, \cdots \kappa$

$$\begin{aligned} [\mathbf{v}(x,y,l) - \mathbf{v}(x',y',l)]^2 &\leq C_{01}[|x - x'| + |y - y'|]^{2\nu_1} \\ \sum_{l=1}^{L} l^{2\nu_2}\, \mathbf{v}^2(x,y,l) &\leq C_{02} \end{aligned}$$

Two regimes

1. $\kappa = \kappa_n$ **is arbitrary**, $\nu_1 = \infty$. Corresponds to the **piecewise constant graphon** that forms a **dynamic SBM** with κ_n blocks
2. κ **is a fixed finite constant independent of** n. Corresponds to the **piecewise smooth graphon**

Piecewise constant graphon

Regime 1

Denote a class of functions f satisfying assumptions on the previous slide with $\kappa = \kappa_n$ and $\nu_1 = \infty$ by $\mathcal{W} \equiv \mathcal{W}(\kappa, \infty, \nu_2)$

Then **the upper and the lower bounds for the mean squared risk coincide up to a constant**

$$\sup_{f \in \mathcal{W}} \mathbb{E}\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}\|_F^2 \asymp \min\left[\frac{1}{L}\left(\frac{\kappa_n^2 \log(n/\kappa)}{n^2}\right)^{\frac{2\nu_2}{2\nu_2+1}}, \frac{\kappa_n^2}{n^2}\right] + \frac{\log \kappa_n}{nL}$$

Similar results hold in probability

Smooth graphon

Regime 2

Denote a class of functions f satisfying assumptions on the previous slide with a constant κ independent of n and $\nu_1 < \infty$ by $\mathcal{W} \equiv \mathcal{W}(\kappa, \nu_1, \nu_2)$

Then **the upper and the lower bounds for the mean squared risk coincide up to a logarithmic factor of L**

$$\sup_{f \in \mathcal{W}} \mathbb{E}\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}\|_F^2 \le C \left\{ \min \left[\frac{1}{L} \left(\frac{\log L}{n^2} \right)^{\frac{2\nu_1\nu_2}{(2\nu_2+1)(\nu_1+1)}}, \left(\frac{\log L}{n^2} \right)^{\frac{\nu_1}{\nu_1+1}} \right] + \frac{\log n}{nL} \right\}$$

$$\inf_{\widehat{\mathbf{\Theta}}} \sup_{f \in \mathcal{W}} \mathbb{E}\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}\|_F^2 \ge C \left\{ \min \left[\frac{1}{L} \left(\frac{1}{n^2} \right)^{\frac{2\nu_1\nu_2}{(2\nu_2+1)(\nu_1+1)}}, \left(\frac{1}{n^2} \right)^{\frac{\nu_1}{\nu_1+1}} \right] + \frac{\log n}{nL} \right\}$$

Similar results hold in probability

Discussion

Our contributions:

- **Penalized least squares** optimization algorithm
- **No assumptions on the mechanism** generating changes in the memberships of the nodes
- **No knowledge of the number of classes**
- **Oracle inequalities** for the mean squared risk of an estimator of $\mathbf{\Theta}$ (for the dynamic SBM and graphon)
- **Non-asymptotic minimax study** of the DSBM model and the dynamic graphon
- Allows a variety of **extensions**
- Vectorization of the model leads to more intuitive **much simpler mathematics** and **faster computations**

Discussion

The shortcomings:

- The maximum likelihood estimation is **NP-hard** and is not **computationally viable**

  $\Rightarrow$ in practice one has to use a **convex relation of the algorithm**

- The dynamic graphon defines $\mathbf{\Theta}_{i,j,l} = f(\zeta_i, \zeta_j, l/L)$, $i, j = 1, \cdots, n$, $l = 1, \cdots, L$, where $\zeta_1, \cdots, \zeta_n$ are random variables **independent of time**

  It corresponds to a **DSBM** where **nodes do not change membershops in time**

  Need to generalize the model to the case when $\zeta_1(t), \cdots, \zeta_n(t)$ are stochastic processes

  The notion of a dynamic graphon has not been developed yet by a probability community

References

- Pensky, M. Dynamic network models and graphon estimation. ArXiv1607.00673
- Pensky, M., Zhang, T. Spectral clustering in the dynamic stochastic block model. ArXiv1705.01204