Uniform estimation of some random graph parameters

Ismaël Castillo

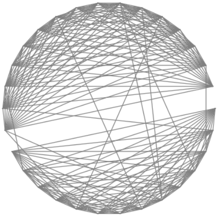
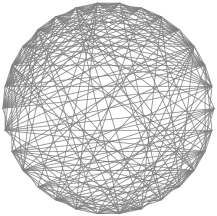
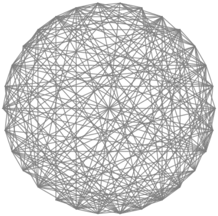
LPMA – Université Paris VI

joint work with Peter Orbanz (Columbia)

Mathematical Methods of Statistics
July 10th, 2017

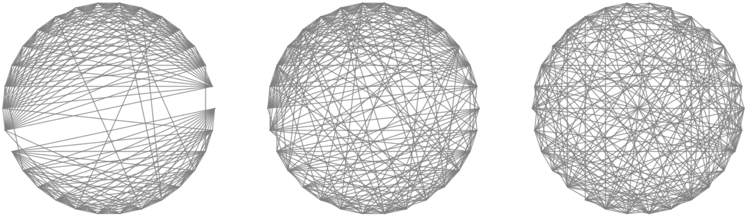
Random graph models

Graph samples with $n = 30$ nodes

Random graph models

Graph samples with $n = 30$ nodes



How does the *information* grow with n?

Random graph models

- Given a probability model for the random graph [next slides]

$$\mathcal{P} = \{P_\eta^{(n)},\ \eta \in H\}$$

  Collection of possible distributions for $(X_{ij})_{1 \le i < j \le n}$, where

  - $X_{ij} \in \{0, 1\}$ tells whether an edge is present or not between nodes i and j

- One may be interested in estimation of
  - the 'parameters' η
  - and/or of functionals $\psi(P_\eta)$ of those

- Example: edge density $E_{P_\eta}[X_{ij}]$
  - Possible estimator

$$\frac{1}{\binom{n}{2}} \sum_{i<j} X_{ij}$$

Random graph models: SBM

The *stochastic block model* SBM with K classes

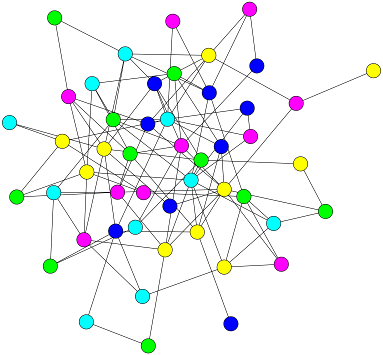
- Parameters
  - $p = (p_1, \ldots, p_K)$ prob. of classes
  - $Q = (Q_{kl})$ symmetric $K \times K$ matrix prob. of connection between classes
- Notation: labels
  - $\varphi : \{1, \ldots, n\} \to \{1, \ldots, K\}$ assigns a class to each vertex
- Observations: $(X_{ij})_{i<j}$ [the labelling map φ is not observed]

Let $\pi = p_1\delta_1 + \ldots + p_K\delta_K$. The data distribution is

$$(\varphi(1), \ldots, \varphi(n)) \sim \pi^{\otimes n},$$

$$(X_{ij})_{i<j} \mid \varphi \sim \bigotimes_{i<j} \mathrm{Be}(Q_{\varphi(i)\varphi(j)})$$

Random graph models: SBM



- $K = 5$
- $p = (p_1, \ldots, p_5)$ probabilities of classes
- Q a 5×5 matrix of connectivities

Random graph models: graphon model

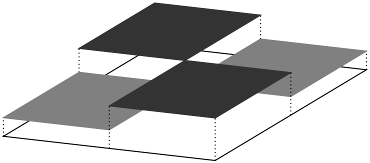
The *graphon model*

- Parameter
  - $f : [0,1]^2 \to [0,1]$ measurable symmetric
- Observations: $(X_{ij})_{i<j}$ [the design variables U_i are not observed]

$$
\begin{aligned}
(U_i)_i &\sim \mathrm{Unif}[0,1]^{\otimes n} \\
(X_{ij})_{i<j} \mid (U_i)_i &\sim \bigotimes_{i<j} \mathrm{Be}(f(U_i, U_j))
\end{aligned}
$$

Random graph models

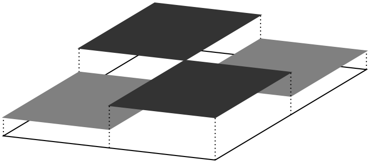
- The SBM model as a graphon model

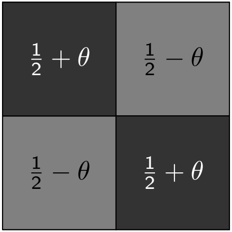


$\frac{1}{2}+\theta$	$\frac{1}{2}-\theta$
$\frac{1}{2}-\theta$	$\frac{1}{2}+\theta$

Random graph models

- The SBM model as a graphon model





- Matrix $Q = (Q_{ij})$ of SBM model can be read as heights of histogram with partition on $[0, 1]$ given by proportions vector p

SBM: estimation of parameters

[Bickel et al 2013] [here in case $\mathbb{P}[X_{ij} = 1] =: \rho \sim$ cst.]

SBM: estimation of parameters

[Bickel et al 2013] [here in case $\mathbb{P}[X_{ij} = 1] =: \rho \ \sim$ cst.]

- Assume that
  - $K = K_0$ is known and fixed
  - no two lines of matrix Q_0 are the same and $\min_k p_{0,k} \neq 0$

- First, what happens if labels $\varphi(\cdot)$ would be observed?

SBM: estimation of parameters

[Bickel et al 2013] [here in case $\mathbb{P}[X_{ij} = 1] =: \rho \sim$ cst.]

- Assume that
  - $K = K_0$ is known and fixed
  - no two lines of matrix Q_0 are the same and $\min_k p_{0,k} \neq 0$

- First, what happens if labels $\varphi(\cdot)$ would be observed?
  - the MLE $(\tilde{p}^{ML}, \tilde{Q}^{ML})$ is asymptotically normal

$$\sqrt{n}(\tilde{p}^{ML} - p_0) \mid \varphi \to \mathcal{N}(0, T_1)$$
$$n(\tilde{Q}^{ML} - Q_0) \mid \varphi \to \mathcal{N}(0, T_2)$$

$$\tilde{p}^{ML}_a = \frac{1}{n}\sum_{i=1}^n 1\!\!1_{\varphi(i)=a}, \qquad \tilde{Q}^{ML}_{ab} = \frac{\sum_{i<j} X_{ij} 1\!\!1_{\varphi(i)=a\,,\,\varphi(j)=b}}{\sum_{i<j} 1\!\!1_{\varphi(i)=a\,,\,\varphi(j)=b}}$$

SBM: estimation of parameters

[Bickel et al 2013]

- Theorem. If labels are unknown, under the previous assumptions
  - the MLE $(\hat{p}^{ML}, \hat{Q}^{ML})$ is asymptotically normal

$$\begin{aligned} \sqrt{n}(\hat{p}^{ML*} - p_0) &\to \mathcal{N}(0, T_1) \\ n(\hat{Q}^{ML*} - Q_0) &\to \mathcal{N}(0, T_2) \end{aligned}$$

where (p^*, Q^*) denotes a label switched-version of (p, Q)

SBM: estimation of parameters

[Bickel et al 2013]

- Theorem. If labels are unknown, under the previous assumptions
  - the MLE $(\hat{p}^{ML}, \hat{Q}^{ML})$ is asymptotically normal

$$\begin{aligned} \sqrt{n}(\hat{p}^{ML*} - p_0) &\to \mathcal{N}(0, T_1) \\ n(\hat{Q}^{ML*} - Q_0) &\to \mathcal{N}(0, T_2) \end{aligned}$$

  where (p^*, Q^*) denotes a label switched-version of (p, Q)

- *Pointwise* inference is *asymptotically* equivalent to given φ case
  - Idea : ML 'profiles out' the unknown φ
    $\to$ proof based on $\hat{\varphi}$ that consistently estimates φ asymptotically

SBM: estimation of parameters

[Bickel et al 2013]

- Theorem. If labels are unknown, under the previous assumptions
  - the MLE $(\hat{p}^{ML}, \hat{Q}^{ML})$ is asymptotically normal

$$\sqrt{n}(\hat{p}^{ML*} - p_0) \rightarrow \mathcal{N}(0, T_1)$$
$$n(\hat{Q}^{ML*} - Q_0) \rightarrow \mathcal{N}(0, T_2)$$

  where (p^*, Q^*) denotes a label switched-version of (p, Q)

- *Pointwise* inference is *asymptotically* equivalent to given φ case
  - Idea : ML 'profiles out' the unknown φ
    $\rightarrow$ proof based on $\hat{\varphi}$ that consistently estimates φ asymptotically

- Note that
  - p_0 estimated at 'slow' rate $\frac{1}{\sqrt{n}}$ $\rightarrow$ $\frac{1}{n}$ in terms of quadratic risk
  - Q_0 estimated at 'fast' rate $\frac{1}{n}$ $\rightarrow$ $\frac{1}{n^2}$ in terms of quadratic risk

Some questions

The previous results are *asymptotic* and *pointwise* at (p_0, Q_0)

Some questions

- *uniform* estimation of parameters?
  $\rightarrow$ say of the connectivity parameters = the matrix Q

- in practice n and K are free
  $\rightarrow$ non-asymptotic results where K is possibly 'large'?

Framework : SBM (to start with)

Related topics

- Testing and 'community' detection

  [Arias-Castro, Candès and Durand 2011]

  [Butucea and Ingster 2013]

  [Arias-Castro and Verzelen 2014-15]

- Estimation of the graphon function f [wrt squared L^2-type risk]

  [Olhede and Wolfe 2013]

  [Gao, Lu and Zhou 2015]

  [Klopp, Tsybakov and Verzelen 2015]

- Other random graph models
  'sparsified' graphon model, preferential attachment, graphex model, ...

A first result $K = 2$

For simplicity assume equiproportions $p_1 = p_2 = \frac{1}{2}$

- Consider the SBM submodel $p = [\frac{1}{2}, \frac{1}{2}]$, $Q = Q^\theta$, $K = 2$, with

$$Q^\theta = \begin{bmatrix} \frac{1}{2} + \theta & \frac{1}{2} - \theta \\ \frac{1}{2} - \theta & \frac{1}{2} + \theta \end{bmatrix}$$

A first result $K = 2$

For simplicity assume equiproportions $p_1 = p_2 = \frac{1}{2}$

- Consider the SBM submodel $p = [\frac{1}{2}\,,\,\frac{1}{2}]$, $Q = Q^\theta$, $K = 2$, with

$$Q^\theta = \begin{bmatrix} \frac{1}{2} + \theta & \frac{1}{2} - \theta \\ \frac{1}{2} - \theta & \frac{1}{2} + \theta \end{bmatrix}$$

- Theorem 1 [*minimax lower bound*]. For some $c_0 > 0$,

$$\inf_{\hat\theta} \sup_{|\theta| \le \frac{1}{2}} E_\theta \left[(\hat\theta - \theta)^2 \right] \ge \frac{c_0}{n}.$$

A first result $K = 2$

For simplicity assume equiproportions $p_1 = p_2 = \frac{1}{2}$

- Consider the SBM submodel $p = [\frac{1}{2}\,,\,\frac{1}{2}]$, $Q = Q^\theta$, $K = 2$, with

$$Q^\theta = \begin{bmatrix} \frac{1}{2}+\theta & \frac{1}{2}-\theta \\ \frac{1}{2}-\theta & \frac{1}{2}+\theta \end{bmatrix}$$

- Theorem 1 [*minimax lower bound*]. For some $c_0 > 0$,

$$\inf_{\hat\theta} \sup_{|\theta|<\frac{\delta_0}{n}} E_\theta\left[(\hat\theta-\theta)^2\right] \geq \frac{c_0}{n}.$$

A first result $K = 2$

For simplicity assume equiproportions $p_1 = p_2 = \frac{1}{2}$

- Consider the SBM submodel $p = [\frac{1}{2}\,,\,\frac{1}{2}]$, $Q = Q^\theta$, $K = 2$, with

$$Q^\theta = \begin{bmatrix} \frac{1}{2} + \theta & \frac{1}{2} - \theta \\ \frac{1}{2} - \theta & \frac{1}{2} + \theta \end{bmatrix}$$

- Theorem 1 *[minimax lower bound]*. For some $c_0 > 0$,

$$\inf_{\hat\theta} \sup_{|\theta| < \frac{\delta_0}{n}} E_\theta\left[(\hat\theta - \theta)^2\right] \geq \frac{c_0}{n}.$$

  - minimax local lower bound, relevant not only at $\theta = 0$, but also locally

    can show that if $t = \dfrac{\delta_0}{2n}$, $\quad \inf_{\hat\theta} \sup_{|\theta - t| < \frac{\delta_0}{2n}} E_\theta\left[(\hat\theta - \theta)^2\right] \geq \frac{c_0}{n}.$

A first result K=2

- Consider the submodel

$$Q^{\theta} = \begin{bmatrix} \frac{1}{2} + \theta & \frac{1}{2} - \theta \\ \frac{1}{2} - \theta & \frac{1}{2} + \theta \end{bmatrix}$$

  One observes SBM-data with $p = [\frac{1}{2}, \frac{1}{2}]$, $Q = Q^{\theta}$, $K = 2$

- Let $\hat{\theta}^{ML}$ be the MLE in the above submodel

- Theorem 1 (bis) [*minimax upper bound*]. For some $d_0 > 0$,

$$\sup_{|\theta| \leq \frac{1}{2}} E_{\theta}\left[(\hat{\theta}^{ML} - \theta)^2 \right] \leq \frac{d_0}{n}$$

  Proof: profile (pseudo)-maximum likelihood

Main result for $K = k$ classes, motivation

An example with $n = 30$, $K = 5$ and a 'difficult' matrix Q

$$\begin{bmatrix} .55 & .45 & .4 & .1 & .7 \\ .45 & .55 & .4 & .1 & .7 \\ .4 & .4 & .6 & .2 & .1 \\ .1 & .1 & .2 & .1 & .4 \\ .7 & .7 & .4 & .4 & .2 \end{bmatrix}$$

Main result for $K = k$ classes, motivation

An example with $n = 30$, $K = 5$ and a 'difficult' matrix Q

$$\begin{bmatrix} {\color{red}.55} & {\color{red}.45} & .4 & .1 & .7 \\ {\color{red}.45} & {\color{red}.55} & .4 & .1 & .7 \\ .4 & .4 & .6 & .2 & .1 \\ .1 & .1 & .2 & .1 & .4 \\ .7 & .7 & .4 & .4 & .2 \end{bmatrix}$$

Local uniform estimation rate of elements of this submatrix will be 'slow'

SBM result $K = k$ classes

Suppose equiproportions for simplicity [balanced proportions would work]

Let us consider estimation along the submodel

$$\begin{bmatrix} a_0 & a_1 & \dots & a_{k-2} \\ a_1 & & & \\ \vdots & & B & \\ a_{k-2} & & & \end{bmatrix}$$

SBM result $K = k$ classes

Suppose equiproportions for simplicity   [balanced proportions would work]

Let us consider estimation along the submodel

$$\begin{bmatrix} a_0 & a_1 & \dots & a_{k-2} \\ a_1 & & & \\ \vdots & & B & \\ a_{k-2} & & & \end{bmatrix} \to \begin{bmatrix} a_0 + \theta & a_0 - \theta & a_1 & \dots & a_{k-2} \\ a_0 - \theta & a_0 + \theta & a_1 & \dots & a_{k-2} \\ a_1 & a_1 & & & \\ \vdots & \vdots & & B & \\ a_{k-2} & a_{k-2} & & & \end{bmatrix} = Q^\theta$$

Set $A = [a_0 \ a_1 \ \cdots \ a_{k-2}]$ and $B = (b_{ij})$

SBM result $K = k$ classes

Suppose proportions are equidistributed [extends to 'balanced proportions']

- main message of Theorem 2 below

Around *any* point in the interior of the $K = k - 1$ classes model,

there is a direction coming from the $K = k$ classes model such that

rate of estimation of matrix parameter along this direction no better than $\frac{k}{n}$

SBM result $K = k$ classes

Suppose proportions are equidistributed [extends to 'balanced proportions']

- main message of Theorem 2 below

  Around *any* point in the interior of the $K = k - 1$ classes model,

  there is a direction coming from the $K = k$ classes model such that

  rate of estimation of matrix parameter along this direction no better than $\frac{k}{n}$

- One observes SBM-data with law E_θ specified by

$$K = k,\ p = [\frac{1}{K}, \ldots \frac{1}{K}],\ Q = Q^\theta \text{ the previous } K \times K \text{ matrix}$$

- Theorem 2 [*minimax lower bound]*. For some $c_1 > 0$, for *any* A and B,

$$\inf_{\hat{\theta}} \sup_{|\theta| \le \frac{1}{2}} E_\theta \left[(\hat{\theta} - \theta)^2 \right] \ge c_1 \frac{k}{n}.$$

Comments on the lower bound

- The bound is minimax *local*
- Idea of proof of lower bound

  a 'mixture vs mixture' lower bound argument

  more involved than before, as 'null hypothesis' is a mixture

SBM result $K = k$ classes

Implications

- The 'boundary' of SBM model with $K = k$ moves with k and n
- If one is interested in estimating (some of the) heights = elements of the Q matrix, one should take this moving boundary into account to determine precisely the accuracy of estimation
- The rate along constructed submodel deteriorates with k.

  The lower bound is non-asymptotic.

  For many k, n, the boundary area, of size at least $\frac{k}{n}$, can be 'large'

Upper-bound result

Recall
$$Q^\theta = \begin{bmatrix} a_0+\theta & a_0-\theta & a_1 & \dots & a_{k-2} \\ a_0-\theta & a_0+\theta & a_1 & \dots & a_{k-2} \\ a_1 & a_1 & & & \\ \vdots & \vdots & & B=(b_{ij}) & \\ a_{k-2} & a_{k-2} & & & \end{bmatrix}$$

- Require, for $\mathcal{C} := \{a_i, b_{ij},\ 1 \le i,j \le k-2\}$,

$$\min_{c\in\mathcal{C}} \{|c-a_0|\} \ge \kappa > 0 \qquad (C)$$

$$k^3 \log k \lesssim \kappa^4 n \qquad (D)$$

Upper-bound result

Recall
$$Q^\theta = \begin{bmatrix} a_0+\theta & a_0-\theta & a_1 & \dots & a_{k-2} \\ a_0-\theta & a_0+\theta & a_1 & \dots & a_{k-2} \\ a_1 & a_1 & & & \\ \vdots & \vdots & & B=(b_{ij}) & \\ a_{k-2} & a_{k-2} & & & \end{bmatrix}$$

- Require, for $\mathcal{C} := \{a_i, b_{ij},\ 1 \le i,j \le k-2\}$,

$$\min_{c\in\mathcal{C}} \{|c - a_0|\} \ge \kappa > 0 \qquad (C)$$

$$k^3 \log k \lesssim \kappa^4 n \qquad (D)$$

- Theorem 3 *[upper bound]*. For some $C_1 > 0$, for *any* A, B such that conditions (C)–(D) hold, for $\hat{\theta}$ a profile-MLE estimate,

$$\sup_{|\theta|\le\kappa} E_\theta\left[(\hat{\theta}-\theta)^2\right] \le C_1 \frac{k}{n}.$$

Upper-bound result – checking (C) and (D)

Conditions (C) and (D) are satisfied in the following settings

- *Example 1 (well-separated block)*

  If κ is a given positive constant e.g. $\kappa = 1/4$, then (C) means that the coefficients of A, B are fairly different from a_0

- *Example 2 (randomly sampled matrices A, B)*

  If the vector defining A and the upper half of the matrix B are sampled iid $\mathcal{U}[0,1]$, then (C)-(D) is satisfied with high probability if

$$k \lesssim n^{1/7}$$

Dependence on the 'environment' A and B

Consider a 'least-favorable case"

$$Z^\theta = \begin{bmatrix} \frac{1}{2}+\theta & \frac{1}{2}-\theta & \frac{1}{2} & \cdots & \frac{1}{2} \\ \frac{1}{2}-\theta & \frac{1}{2}+\theta & \frac{1}{2} & \cdots & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & \cdots & \frac{1}{2} \\ \vdots & \vdots & \vdots & & \vdots \\ \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & \cdots & \frac{1}{2} \end{bmatrix}.$$

[or a perturbation thereof]

Observe data from the corresponding SBM model with equiproportions

- Theorem 4. There exists $c_2 > 0$ such that

$$\inf_{\hat{\theta}} \sup_{\theta \in (-1/2, 1/2)} E_\theta(\hat{\theta} - \theta)^2 \geq c_2 \frac{k^2}{n}$$

Functionals of graphon model

In the more general graphon model, consider the functional

$$\psi(\langle f\rangle) = \left[\int_{[0,1]^2} \left(f(x,y) - \int_{[0,1]^2} f\right)^2 dxdy\right]^{1/2}$$

[Continuous 'graphon-analogue' of previous parameter θ]

- Theorem 5. For $\mathcal{P}$ a $\mathcal{C}^1(M)$-class of graphons, for some $c_3, c_4 > 0$,

$$\frac{c_3}{n} \leq \inf_{\hat\psi} \sup_{f\in\mathcal{P}} E_f\left[(\hat\psi - \psi(f))^2\right] \leq \frac{c_4}{n}.$$

Functionals of graphon model

- Lower bound also holds for other functionals such as

$$\psi(P_w) = \int_{[0,1]^2} \left| f(x,y) - \int_{[0,1]^2} |f(x,y)| dxdy \right| dxdy$$

- The quadratic 'slow rate' $1/n$ [or $1/\sqrt{n}$ non-quadratic] appears to be quite universal for uniform estimation of many functionals

Conclusion

Previously known results for SBMs (K, p, Q): given fixed $K = k$,

$\to$ proportions p estimated at 'slow' rate $\frac{1}{n}$ asymptotically

$\to$ connectivities Q estimated at 'fast' rate $\frac{1}{n^2}$ asymptotically, *pointwise*, in the *interior* of set of SBMs with k classes

Conclusions

- Fast rate for connectivities Q not achievable uniformly

  Uniform rates are at most k/n [=generic lower bound if k not too large] and can be as slow as k^2/n, from a non-asymptotic perspective in n and k

- This phenomenom happens close to *any* $k - 1$ classes model, not only around a 'least favorable one' $\to$ local lower bound

Open questions

- SBM: 'Continuum' of rates inbetween k/n and k^2/n depending on A, B?
- More general estimation theory of graphon functionals?