Statistical issues in the stochastic block model

Catherine Matias

CNRS - Université Pierre et Marie Curie, Paris
catherine.matias@math.cnrs.fr
http://cmatias.perso.math.cnrs.fr/

INI - Graph Limits and Statistics

Outline

Contact networks

Animal (or human) contact networks, built from

- Field observations of associations between animals;
- sensor based measurements;
- trapping data; ...

Some questions

- Is there a social structure?
  Understanding if there is a peculiar non-random mixing of individuals that would be a sign for a social organisation
- How does it vary with other factors?
  Breeding season, seasonal changes, response to stress, arrival/departure of a peculiar individual, ...
- How can we predict how infectious diseases can spread?
- What is the dynamics of the structure?

Modeling heterogeneity I

- Random graphs are heterogeneous (simple Erdös-Rényi model fails to capture this heterogeneity);
- A simple way to capture this heterogeneity is to assume group structure on the nodes;
- Classification of nodes also induces a summary of the information contained in the graph;

Motivation/Justification: Szemerédi's regularity Lemma [Szemerédi(1978)]

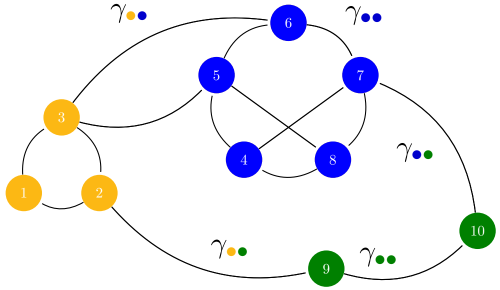
Every large enough graph can be divided into subsets of about the same size so that the edges between different subsets behave almost randomly.

Modeling heterogeneity II

Nodes classification: when and why?

- Practical methods discussed here are suited for graphs with up to a few thousands of nodes;
- With (very) large datasets, is node classification adapted anyway?

Stochastic block model (binary graphs)



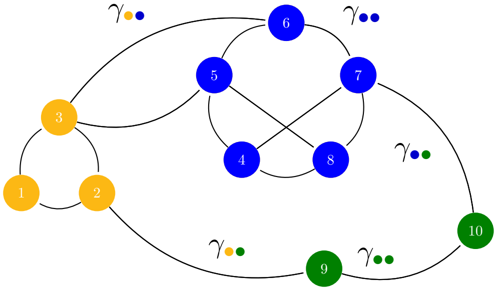
$n = 10, Z_{5\bullet} = 1$

$A_{12} = 1, A_{15} = 0$

Binary case (parametric model with $\theta = (\boldsymbol{\pi}, \boldsymbol{\gamma})$)

- K groups (=colors $\bullet\bullet\bullet$).
- $\{Z_i\}_{1 \leq i \leq n}$ i.i.d. vectors $Z_i = (Z_{i1}, \ldots, Z_{iK}) \sim \mathcal{M}(1, \boldsymbol{\pi})$, with $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)$ groups proportions. Z_i not observed (latent).
- Observations: presence/absence of an edge $\{A_{ij}\}_{1 \leq i < j \leq n}$,
- Conditional on $\{Z_i\}$'s, the r.v. A_{ij} are independent $\mathcal{B}(\gamma_{Z_i Z_j})$.

Stochastic block model (weighted graphs)



$n = 10, Z_{5\bullet} = 1$

$A_{12} \in \mathbb{R}, A_{15} = 0$

Weighted case (parametric model with $\theta = (\boldsymbol{\pi}, \boldsymbol{\gamma}^{(1)}, \boldsymbol{\gamma}^{(2)})$)

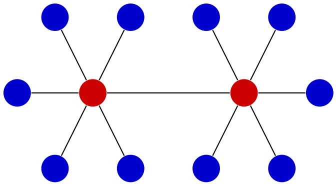
- Latent variables: *idem*
- Observations: 'weights' A_{ij} , where $A_{ij} = 0$ or $A_{ij} \in \mathbb{R}^s \setminus \{0\}$,
- Conditional on the $\{Z_i\}$'s, the random variables A_{ij} are independent with distribution

$$\mu_{Z_i Z_j}(\cdot) = \gamma^{(1)}_{Z_i Z_j} f(\cdot, \gamma^{(2)}_{Z_i Z_j}) + (1 - \gamma^{(1)}_{Z_i Z_j})\delta_0(\cdot)$$

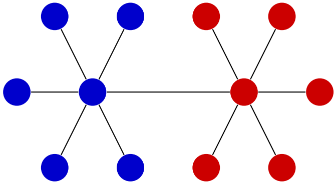
SBM classification vs community detection

SBM classification

- Nodes classification induced by the model reflects a common connectivity behaviour;
- Many clustering methods try to group nodes that belong to the same quasi-clique (ex: community detection)
- Toy example



SBM clusters



Community detection or SBM

Particular cases and generalisations

Particular case: Affiliation model (planted partition)

$$\boldsymbol{\gamma} = \begin{pmatrix} \alpha & \dots & \beta \\ \vdots & \ddots & \vdots \\ \beta & \dots & \alpha \end{pmatrix} \qquad (\alpha \gg \beta \implies \text{community detection})$$

Some generalisations

- Overlapping groups for binary graphs; SBM with covariates; Degree corrected SBM;...
- Latent block models (LBM), for array data or bipartite graphs [Govaert and Nadif(2003)];
- Nonparametric SBM (graphon);
- Dynamic SBM (come back in December!).

Outline

Outline

Parameter's identifiability

Problem

- Obviously, the model may only be identifiable up to a permutation on the group's labels.
  Parameter $\theta = (\boldsymbol{\pi}, \boldsymbol{\gamma})$ is identifiable whenever
  $\mathbb{P}_\theta = \mathbb{P}_{\theta'} \iff \exists \sigma \in \mathfrak{S}_K, \boldsymbol{\pi}_k = \boldsymbol{\pi}'_{\sigma(k)}$ and $\boldsymbol{\gamma}_{kl} = \boldsymbol{\gamma}'_{\sigma(k)\sigma(l)}$.
- But whether one may uniquely recover the parameter up to a permtation is a delicate task.

Necessary condition

Any 2 rows (or columns) of connectivity matrix γ must be different.

Is this sufficient? No general proof yet!

Parameter's identifiability

Problem

- Obviously, the model may only be identifiable up to a permutation on the group's labels.
  Parameter $\theta = (\boldsymbol{\pi}, \boldsymbol{\gamma})$ is identifiable whenever
  $\mathbb{P}_\theta = \mathbb{P}_{\theta'} \iff \exists \sigma \in \mathfrak{S}_K, \boldsymbol{\pi}_k = \boldsymbol{\pi}'_{\sigma(k)}$ and $\boldsymbol{\gamma}_{kl} = \boldsymbol{\gamma}'_{\sigma(k)\sigma(l)}$.
- But whether one may uniquely recover the parameter up to a permtation is a delicate task.

Necessary condition

Any 2 rows (or columns) of connectivity matrix $\boldsymbol{\gamma}$ must be different.

Is this sufficient? No general proof yet!

Some existing identifiability results

- For binary SBM with $K = 2$ groups, [Allman et al.(2009)] prove that $\gamma_{11}, \gamma_{12}, \gamma_{22}$ distinct is a sufficient condition ($n \geq 16$);
- For binary SBM with $K \geq 3$, [Allman et al.(2011)] prove a generic identifiability (condition is not explicit);
- For directed binary SBM, [Celisse et al.(2012)] prove that $\gamma\pi$ (or $\boldsymbol{\pi}^\intercal\gamma$) has distinct coordinates is sufficient ($n \geq 2K$).
- For *weighted* SBM, the only known result [Allman et al.(2011)] requires all parameters in γ to be distinct!

Open question: Do we really need these extra conditions?

Some existing identifiability results

- For binary SBM with $K = 2$ groups, [Allman et al.(2009)] prove that $\gamma_{11}, \gamma_{12}, \gamma_{22}$ distinct is a sufficient condition ($n \geq 16$);
- For binary SBM with $K \geq 3$, [Allman et al.(2011)] prove a generic identifiability (condition is not explicit);
- For directed binary SBM, [Celisse et al.(2012)] prove that $\gamma\pi$ (or $\boldsymbol{\pi}^\intercal\gamma$) has distinct coordinates is sufficient ($n \geq 2K$).
- For *weighted* SBM, the only known result [Allman et al.(2011)] requires all parameters in γ to be distinct!

Open question: Do we really need these extra conditions?

Outline

The stochastic block model

Statistical issues

- Identifiability
- Parameter estimation
- Some convergence results

Towards dynamic networks

Overview of algorithms

Goal is MLE. Likelihood computation is untractable for n not small.

Parameter estimation

- `em` algorithm not feasible because latent variables are not independent conditional on observed ones: $\mathbb{P}(\{Z_i\}_i|\{A_{ij}\}_{i,j}) \neq \prod_i \mathbb{P}(Z_i|\{A_{ij}\}_{i,j})$
- Alternatives:
  - Gibbs sampling
  - Variational approximation to `em`.
  - Ad-hoc methods: Composite likelihood or Moment methods [Ambroise and Matias(2012), Bickel et al.(2011)]; Degrees [Channarond et al.(2012)];

Variational approximation principle I

Log-likelihood decomposition

$\mathcal{L}_{\mathbf{A}}(\boldsymbol{\theta}) := \log \mathbb{P}(\mathbf{A}; \boldsymbol{\theta}) = \log \mathbb{P}(\mathbf{A}, \mathbf{Z}; \boldsymbol{\theta}) - \log \mathbb{P}(\mathbf{Z}|\mathbf{A}; \boldsymbol{\theta})$ and for any distribution $\mathbb{Q}$ on $\mathbf{Z}$,

$$\mathcal{L}_{\mathbf{A}}(\boldsymbol{\theta}) = \mathbb{E}_{\mathbb{Q}}(\log \mathbb{P}(\mathbf{A}, \mathbf{Z}; \boldsymbol{\theta})) + \mathcal{H}(\mathbb{Q}) + \mathcal{KL}(\mathbb{Q} \| \mathbb{P}(\mathbf{Z}|\mathbf{A}; \boldsymbol{\theta}))$$

`em` principle

- `e`-step: maximise the quantity $\mathbb{E}_{\mathbb{Q}}(\log \mathbb{P}(\mathbf{A}, \mathbf{Z}; \boldsymbol{\theta}^{(t)})) + \mathcal{H}(\mathbb{Q})$ with respect to $\mathbb{Q}$. This is equivalent to minimizing $\mathcal{KL}(\mathbb{Q} \| \mathbb{P}(\mathbf{Z}|\mathbf{A}; \boldsymbol{\theta}^{(t)}))$ with respect to $\mathbb{Q}$.
- `m`-step: keeping now $\mathbb{Q}$ fixed, maximize the quantity $\mathbb{E}_{\mathbb{Q}}(\log \mathbb{P}(\mathbf{A}, \mathbf{Z}; \boldsymbol{\theta})) + \mathcal{H}(\mathbb{Q})$ with respect to $\boldsymbol{\theta}$ and update the parameter value $\boldsymbol{\theta}^{(t+1)}$ to this maximiser. This is equivalent to maximizing the conditional expectation $\mathbb{E}_{\mathbb{Q}}(\log \mathbb{P}(\mathbf{A}, \mathbf{Z}; \boldsymbol{\theta}))$ w.r.t. $\boldsymbol{\theta}$.

Variational approximation principle II

Variational `em`

- `e`-step: search for an optimal $\mathbb{Q}$ within a restricted class $\mathcal{Q}$, e.g. class of factorized distr.

$$\mathbb{Q}(\mathbf{Z}) = \prod_{i=1}^{n} \mathbb{Q}(Z_i), \quad \mathbb{Q}^\star = \arg\min_{\mathbb{Q}\in\mathcal{Q}} \mathcal{KL}(\mathbb{Q} \| \mathbb{P}(\mathbf{Z}|\mathbf{A};\boldsymbol{\theta}^{(t)}))$$

- `m`-step: unchanged, *i.e.* $\theta^{(t+1)} = \arg\max_\theta \mathbb{E}_{\mathbb{Q}^\star}(\log \mathbb{P}(\mathbf{A},\mathbf{Z};\boldsymbol{\theta}))$
- A consequence of $\mathcal{KL} \geq 0$ is the lower bound

$$\mathcal{L}_{\mathbf{A}}(\boldsymbol{\theta}) \geq \mathbb{E}_{\mathbb{Q}}(\log \mathbb{P}(\mathbf{A},\mathbf{Z};\boldsymbol{\theta})) + \mathcal{H}(\mathbb{Q})$$

So that the variational approximation consists in maximizing a lower bound on the log-likelihood. Why does it make sense ?

Model selection

How do we choose the number of groups K?

Frequentist setting

- Maximal likelihood is not available (thus neither AIC or BIC),
- ICL criterion is used [Daudin et al.(2008)] (no consistency result on that).

Bayesian setting

- MCMC approach to select number of LBM groups [Wyse and Friel(2012)].
- Exact ICL requires greedy search optimization [Côme and Latouche(2015)]

Outline

Convergence results

Why does the variational approximation work?

- The variational approximation is empirically accurate for SBM.
- Variational approximation does not converge to MLE unless the true posterior $p(\mathbf{Z}|\mathbf{Y};\gamma)$ is degenerate [Gunawardana and Byrne(2005)].
- This is the case for SBM [Mariadassou and Matias(2015)].

Results from [Celisse et al.(2012), Bickel et al.(2013)] (parametric setting)

- Variational `em` estimators are asymptotically equivalent to MLE for SBM.
- MLE is convergent and asymptotically normal.

See Harry's talk about minimax rates in the graphon setting.

Outline

The stochastic block model

Statistical issues

- Identifiability
- Parameter estimation
- Some convergence results

Towards dynamic networks

Aggregation vs stream links I

Aggregation

- Most of the dynamic networks are constructed by aggregating data over pre-specified time intervals.
- How should we choose the window size?
- [Sulo-Caceres et al.(2011), Sulo-Caceres et al.(2010)] propose to select the time window optimizing some criteria (variance vs compression ratio of a statistic). Computer intensive.

Questions

- Other approaches to select window size?
- Can we quantify what we are loosing?

Aggregation vs stream links II

Dynamic SBM with aggregation

- Many proposals among which those by [Yang et al.(2011), Xu and Hero(2014)].
- Main issue is to decide whether the groups can change with time, or the connectivity parameters change with time, or both.
- In [Matias and Miele(2016)], we have proposed a model where both may change. Adding a restriction on the parameters, we can identify them.

Aggregation vs stream links III

Point processes for stream links

- Stream links record interactions between 2 individuals (i, j) at some time point t.
- Data has the form $\{(i_m, j_m, t_m); 1 \leq m \leq M; i_m, , j_m \in V, 0 \leq t_1 < \cdots < t_m\}$.
- Example: phone calls, emails, O/D data, contact nets ...

- $N_{ij}(t)$ =counting process of the interactions between i, j up to time t
- Take into account that $N_{ij}(t)$ and $N_{ik}(t)$ are related.
- See for e.g. [Matias et al.(2015)] for a dynamic point process extension of SBM.

Aggregation vs stream links IV

Questions with those point processes approaches

- We don't have graphs anymore, are we loosing something? (Also: visual representation issue)
- Point processes are not suited for interactions that last over some time interval.
- ...

Thank you for your attention !

References I

E.S. Allman, C. Matias, and J.A. Rhodes.
Identifiability of parameters in latent structure models with many observed variables.
Ann. Statist., 37(6A):3099–3132, 2009.

E.S. Allman, C. Matias, and J.A. Rhodes.
Parameters identifiability in random graph mixture models.
J. Statist. Plann. Inf., 141(5):1719–1736, 2011.

C. Ambroise and C. Matias.
New consistent and asymptotically normal parameter estimates for random graph mixture models.
JRSSB, 74(1):3–35, 2012.

P. Bickel, A. Chen and E. Levina
The method of moments and degree distributions for network models
Ann. Statist., 39(5):2280—2301, 2011.

References II

P. Bickel, D. Choi, X. Chang, H. Zhang
Asymptotic normality of maximum likelihood and its variational approximation for stochastic blockmodels
Ann. Statist., 41(4):1922—1943, 2013.

A. Celisse, J.-J. Daudin, and L. Pierre.
Consistency of maximum-likelihood and variational estimators in the Stochastic Block Model.
Electron. J. Statist., 6:1847–1899, 2012.

A. Channarond, J.-J. Daudin, and S. Robin.
Classification and estimation in the Stochastic Blockmodel based on the empirical degrees.
Electron. J. Statist., 6:2574—2601, 2012.

References III

E. Côme and P. Latouche.
Model selection and clustering in stochastic block models based on the exact integrated complete data likelihood.
Statistical Modelling, 2015.

J.-J. Daudin, F. Picard, and S. Robin.
A mixture model for random graphs.
Stat. Comput., 18(2):173–183, 2008.

G. Govaert and M. Nadif.
Clustering with block mixture models.
Pattern Recognition, 36(2):463–473, 2003.

A. Gunawardana and W. Byrne.
Convergence theorems for generalized alternating minimization procedures.
J. Mach. Learn. Res., 6:2049–2073, 2005.

References IV

M. Mariadassou and C. Matias.
Convergence of the groups posterior distribution in latent or stochastic block models.
Bernoulli, 21(1):537–573, 2015.

C. Matias and V. Miele.
Statistical clustering of temporal networks through a dynamic stochastic block model.
JRSSB, To appear.

C. Matias, T. Rebafka, and F. Villers.
A semiparametric extension of the stochastic block model for longitudinal networks.
arXiv:1512.07075, 2015.

References V

RS Caceres, T Berger-Wolf, R Grossman.
Temporal scale of processes in dynamic networks
IEEE 11th International Conference on Data Mining Workshops, 925-932, 2011.

RS Sulo, T Berger-Wolf, R Grossman.
Meaningful selection of temporal resolution for dynamic networks
Proceedings of the Eighth Workshop on Mining and Learning with Graphs, 127-136, 2010.

E. Szemerédi.
Regular partitions of graphs.
In *Problèmes combinatoires et théorie des graphes*, volume 260 of *Colloq. Internat. CNRS, Univ. Orsay, Orsay, 1976*, pages 399–401, 1978.

References VI

J. Wyse and N. Friel.
Block clustering with collapsed latent block models.
Statistics and Computing, 22(2):415–428, 2012.

K. Xu and A. Hero
Dynamic stochastic blockmodels for time-evolving social networks.
Selected Topics in Signal Processing, IEEE Journal of, 8(4), 552-–562, 2014.

T. Yang, Y. Chi, S. Zhu, Y. Gong, and R. Jin
Detecting communities and their evolutions in dynamic social networks—a Bayesian approach.
Machine learning, 82(2), 157—189, 2011.