

SBM for reconstructed network

November 20, 2019

Contents

1	Introduction	2
2	Model	3
3	Inference	4
3.1	Loss function	4
3.2	Estimation equation	4
4	Identifiability	6
4.1	Review of the literature	6
4.2	Proof in the parametric uni-dimensional context	6
5	Simulation study	8
5.1	Simulation design	8
6	Illustrations	9
A	Appendix	10
A.1	Non-parametric emission distributions	10

1 Introduction

2 Model

Data

- p nodes = species ($1 \leq i, j \leq n$)
- K clusters ($1 \leq k, \ell \leq K$)
- Z_i = cluster of node i , $Z_{ik} = \mathbb{I}\{Z_i = k\}$, $Z = (Z_i)$
- $G_{ij} = \mathbb{I}\{i \sim k\}$ = connection between nodes i and j , $G = (G_{ij})$ = unobserved network
- S_{ij} = score of edge between nodes i and j , $S = (S_{ij})$ = observed score matrix

Parameters

- $\pi = (\pi_k)$ = cluster proportions
- $\gamma = (\gamma_{k\ell})$ = between cluster connection probabilities
- ψ_0 = parameter of the score distribution for absent edge $p(S_{ij} \mid G_{ij} = 0)$ (idem ψ_1 for present edge), $\psi = (\psi_0, \psi_1)$
- $\theta = (\pi, \gamma, \psi)$

Model

- (Z_i) iid,

$$Z_i \sim \mathcal{M}(1, \pi)$$

- (G_{ij}) independent conditionally on Z ,

$$(G_{ij} \mid Z_i = k, Z_j = \ell) \sim \mathcal{B}(\gamma_{k\ell})$$

- (S_{ij}) independent conditionally on G ,

$$(S_{ij} \mid G_{ij} = u) \sim F(\cdot; \psi_u), \quad u = 0, 1$$

We further denote $F_u(\cdot) = F(\cdot; \psi_u)$ and $f_u(\cdot)$ the corresponding pdf.

Properties and definitions

- S and Z independent conditionally on G :

$$p(Z \mid G, S) = p(Z \mid G), \quad p(S \mid G, Z) = p(S \mid G)$$

- Distribution of G_{ij}

$$P(G_{ij} = 1 \mid S_{ij}, Z_i = k, Z_j = \ell) = \frac{\gamma_{k\ell} f_1(S_{ij})}{\gamma_{k\ell} f_1(S_{ij}) + (1 - \gamma_{k\ell}) f_0(S_{ij})} =: \eta_{ij}^{k\ell}$$

$$\tilde{P}(G_{ij} = 1 \mid S_{ij}) = \sum_{k, \ell} \tau_{ik} \tau_{j\ell} \eta_{ij}^{k\ell} =: \bar{\eta}_{ij}$$

- Kullback-Leibler divergence

$$\begin{aligned} KL(q(U); p(U)) &= \mathbb{E}_q(\log q(U) - \log p(U)) \\ KL(q(U, V); p(U, V)) &= \mathbb{E}_{q(U, V)}(\log q(U) + \log(q(V \mid U) - \log p(U) - \log p(V \mid U)) \\ &= KL(q(U); p(U)) + \mathbb{E}_{q(U)} KL(q(V \mid U), p(V \mid U)) \end{aligned}$$

3 Inference

3.1 Loss function

Log-likelihood

$$\begin{aligned}\log p(Z, G, S) &= \log p(Z; \pi) + \log p(G \mid Z; \gamma) + \log p(S \mid G; \psi) \\ &= \sum_{i,k} Z_{ik} \log \pi_k + \sum_{i < j} \sum_{k,\ell} Z_{ik} Z_{j\ell} (G_{ij} \log \gamma_{k\ell} + (1 - G_{ij}) \log(1 - \gamma_{k\ell})) \\ &\quad + \sum_{i < j} G_{ij} \log f_1(S_{ij}) + (1 - G_{ij}) \log f_0(S_{ij})\end{aligned}$$

Approximate distribution $q(Z, G) \approx p(Z, G \mid S)$

$$q(Z, G) = q(Z)q(G \mid Z) := q(Z)p(G \mid Z, S) \quad (1)$$

where

$$p(G \mid Z, S) = \prod_{i,j} p(G_{ij} \mid Z_i, Z_j, S_{ij})$$

and

$$q(Z) = \prod_i q_i(Z_i) = \prod_{i,k} \tau_{ik}^{Z_{ik}}.$$

Divergence $KL(q(Z, G); p(Z, G \mid S))$

$$\begin{aligned}KL(q(Z, G); p(Z, G \mid S)) &= KL(q(Z)p(G \mid Z, S); p(Z \mid S)p(G \mid Z, S)) \\ &= KL(q(Z); p(Z \mid S)) + \underbrace{\mathbb{E}_{q(Z)} KL(p(G \mid Z, S); p(G \mid Z, S))}_{=0}\end{aligned}$$

Still, the conditional entropy of $q(G \mid Z)$ contributes to the lower bound.

Lower bound $J(\theta, q)$

$$\begin{aligned}J(\theta, q) &= \log p_\theta(S) - KL(q(Z, G); p(Z, G \mid S)) \\ &= \mathbb{E}_q \log p_\theta(Z, G, S) + H(q(Z)) + \mathbb{E}_q H(q(G \mid Z))\end{aligned} \quad (2)$$

$$\begin{aligned}&= \sum_{i,k} \tau_{ik} \log \pi_k + \sum_{i < j} (\bar{\eta}_{ij} \log \gamma_{k\ell} + (1 - \bar{\eta}_{ij}) \log(1 - \gamma_{k\ell})) \\ &\quad + \sum_{i < j} \sum_{k,\ell} \tau_{ik} \tau_{j\ell} (\eta_{ij}^{k\ell} \log f_1(S_{ij}) + (1 - \eta_{ij}^{k\ell}) \log f_0(S_{ij})) \\ &\quad - \sum_{i,k} \tau_{ik} \log \tau_{ik} - \sum_{i < j} \sum_{k,\ell} \tau_{ik} \tau_{j\ell} (\eta_{ij}^{k\ell} \log \eta_{ij}^{k\ell} + (1 - \eta_{ij}^{k\ell}) \log(1 - \eta_{ij}^{k\ell}))\end{aligned} \quad (3)$$

3.2 Estimation equation

VE step Denoting

$$\log A_{ijk\ell} = \eta_{ij}^{k\ell} (\log \gamma_{k\ell} + \log f_1(S_{ij})) + (1 - \eta_{ij}^{k\ell}) (\log(1 - \gamma_{k\ell}) + \log f_0(S_{ij}))$$

setting the derivative wrt τ_{ik} to zero with the constraint $\sum_k \tau_{ik} = 0$ gives

$$\log \tau_{ik} = \log \pi_k + \sum_{j,\ell} \tau_{j\ell} \log A_{ijk\ell} + \text{cst} \quad \Leftrightarrow \quad \tau_{ik} \propto \pi_k \prod_{j,\ell} (A_{ijk\ell})^{\tau_{j\ell}}$$

M step Setting the derivative wrt to each parameter gives

$$\hat{\pi}_{ik} = \sum_i \tau_{ik} / n , \quad \hat{\gamma}_{k\ell} = \sum_{i < j} \sum_{k, \ell} \tau_{ik} \tau_{j\ell} \eta_{ij}^{k\ell} \Big/ \sum_{i < j} \sum_{k, \ell} \tau_{ik} \tau_{j\ell} .$$

Furthermore, if $f(\cdot, \psi_u) = \mathcal{N}(\cdot, \mu_u, \sigma_u^2)$ (i.e $\psi_u = (\mu_u, \sigma_u^2)$),

$$\begin{aligned} \hat{\mu}_0 &= \sum_{i < j} (1 - \bar{\eta}_{ij}) S_{ij} \Big/ \sum_{i < j} (1 - \bar{\eta}_{ij}) & \hat{\sigma}_0^2 &= \sum_{i < j} (1 - \bar{\eta}_{ij}) (S_{ij} - \hat{\mu}_0)^2 \Big/ \sum_{i < j} (1 - \bar{\eta}_{ij}) \\ \hat{\mu}_1 &= \sum_{i < j} \bar{\eta}_{ij} S_{ij} \Big/ \sum_{i < j} \bar{\eta}_{ij} S_{ij} & \hat{\sigma}_1^2 &= \sum_{i < j} \bar{\eta}_{ij} (S_{ij} - \hat{\mu}_0)^2 \Big/ \sum_{i < j} \bar{\eta}_{ij} S_{ij} \end{aligned}$$

The case of non-parametric version of f_0 and f_1 is considered in Appendix A.1

By-product The conditional probability for an edge to be part of G is denoted ψ_{ij}^1 :

$$\psi_{ij}^1 := \tilde{P}\{G_{ij} = 1\} = \sum_{k, \ell} \tau_{ik} \tau_{j\ell} \eta_{ij}^{k\ell}$$

and we denote $\psi_{ij}^0 = 1 - \psi_{ij}^1$.

4 Identifiability

4.1 Review of the literature

Notes on identifiability based on papers :

- [1] : "Allman, Elizabeth S. and Matias, Catherine and Rhodes, John A." : *Identifiability of parameters in latent structure models with many observed variables*
- [2] "Allman, Elizabeth S. and Matias, Catherine and Rhodes, John A." : *Parameter identifiability in a class of random graph mixture models*
- [5] "Teicher, Henry" : *Identifiability of Finite Mixtures*
- [6] "Teicher, Henry" : *Identifiability of Mixtures of product measures*

What is done in [2] : identifiability in weighted SBM

$$\begin{aligned} S_{ij}|Z_i = k, Z_j = \ell &\sim \mu_{k\ell} \\ \mu_{k\ell} &= (1 - \gamma_{k\ell})\delta_{\{0\}} + \gamma_{k\ell}F_{k\ell}(\cdot) \end{aligned}$$

for uni dimensional S and symmetric with

- $F_{k\ell}(\cdot)$ parametric (Theorem 12 of [2]) : $F(\cdot; \theta_{k\ell})$ under the following assumptions:
 - [A1] The $K(K + 1)/2$ parameter values $\theta_{k\ell}$ are distinct
 - [A2] The family of measures $\mathcal{M} = \{F(\cdot; \theta) | \theta \in \Theta\}$ is such that
 - [A2 (i)] all elements of \mathcal{M} have no point mass at 0
 - [A2 (ii)] the parameters of finite mixtures of measures of \mathcal{M} are identifiable (up to label switching) i.e.

$$\sum_{m=1}^M \alpha_m F(\cdots, \theta_m) = \sum_{m=1}^M \alpha'_m F(\cdots, \theta'_m) \Rightarrow \sum_{m=1}^M \alpha_m \delta_{\theta_m} = \sum_{m=1}^M \alpha'_m \delta_{\theta'_m}$$

In particular : true for Gaussian ([5]) and Laplace.

- $F_{k\ell}(\cdot)$ non-parametric (Theorem 14 of [2]) : if the $\mu_{k\ell}$ are *linearly independent* (to be detailed)

About the demonstrations

- *Parametric case* It is done from the distribution of a triplet (S_{ij}, S_{ik}, S_{jk}) and using [5]. How to adapt it to our case?
- *Nonparametric case* : only depends on the linear independancy of the $\mu_{k\ell}$. We have to precise it for our case?

4.2 Proof in the parametric uni-dimensional context

I tried to mimic/extend the proof of [2] but I don't think we are in the same scope.

Distribution of the S_{ij}

$$\begin{aligned}\mathbb{P}(S_{ij}) &= \sum_{q,\ell} \pi_q \pi_\ell [(1 - \gamma_{q\ell}) F_0(S_{ij}) + \gamma_{q\ell} F_1(S_{ij})] \\ &= \left[1 - \sum_{q,\ell} \pi_\ell \pi_q \gamma_{q,\ell} \right] F_0(S_{ij}) + \left[\sum_{q,\ell} \pi_q \pi_\ell \gamma_{q,\ell} \right] F_1(S_{ij})\end{aligned}$$

So assuming that F_0 and F_1 are such that any mixture of those two distributions is identifiable, we obtain the identifiability of θ_0, θ_1 and $\sum_{q,\ell} \pi_\ell \pi_q \gamma_{q,\ell}$.

So we have identifiability of $\pi' \gamma \pi$. It seems to me that once we have identified θ_0 and θ_1 we will be able to apply to proof of Clisse & al. [3], which is the one I know better. Which is the thing you said : meaning that once we have identified to high level, we are identifiable just like any binary SBM.

Distribution of the triplet (S_{ij}, S_{ik}, S_{jk})

$$\begin{aligned}\mathbb{P}(S_{ij}, S_{ik}, S_{jk}) &= \sum_{q,\ell,m} \pi_q \pi_\ell \pi_m [(1 - \gamma_{q\ell}) F_0(S_{ij}) + \gamma_{q\ell} F_1(S_{ij})] [(1 - \gamma_{qm}) F_0(S_{ik}) + \gamma_{qm} F_1(S_{ik})] \\ &\quad [(1 - \gamma_{\ell m}) F_0(S_{jk}) + \gamma_{\ell m} F_1(S_{jk})] \\ &= \sum_{q,\ell,m} \sum_{(u,v,w) \in \{0,1\}^3} \eta_{q,\ell,m,u,v} F_u(S_{ij}) F_v(S_{ik}) F_w(S_{jk}) \\ &= \sum_{(u,v,w) \in \{0,1\}^3} \left(\sum_{q,\ell,m} \eta_{q,\ell,m,u,v} \right) F_u(S_{ij}) F_v(S_{ik}) F_w(S_{jk}) \\ &= \sum_{(u,v,w) \in \{0,1\}^3} \left(\sum_{q,\ell,m} \eta_{q,\ell,m,u,v} \right) F_{u,v,w}(S_{ij}, S_{ik}, S_{jk})\end{aligned}$$

with

$$\eta_{q,\ell,m,u,v} = \pi_q \pi_\ell \pi_m (1 - \gamma_{q\ell})^{1-u} \gamma_{q\ell}^u (1 - \gamma_{q\ell})^{1-u} \gamma_{q\ell}^u (1 - \gamma_{qm})^{1-v} \gamma_{qm}^v (1 - \gamma_{\ell m})^{1-w} \gamma_{\ell m}^w.$$

The distribution of (S_{ij}, S_{ik}, S_{jk}) is a mixture (weights = $\sum_{q,\ell,m} \eta_{q,\ell,m,u,v}$) of the following distributions

$$F(s) = F_u(s_1, \theta_u) F_v(s_1, \theta_v) F_w(s_1, \theta_w)$$

where $F \in \mathcal{F}$ with

$$\mathcal{F} = \{F(s; \theta_0, \theta_1) : F(s; \theta_0, \theta_1) = F_u(s_1, \theta_u), F_v(s_2, \theta_v) F_w(s_3, \theta_w), (u, v, w) \in \{0, 1\}^3, \theta_0, \theta_1 \in \Theta_0, \theta_1 \in \Theta_1\}$$

Assumptions; [A1] we assume that any mixtures of elements of \mathcal{F} is identifiable. (to develop to get assumptions on F_0 and F_1).

5 Simulation study

5.1 Simulation design

Data simulation.

- $p = 20, 30, 50, 80$ nodes
- $n = 20, 50, 100, 200$ replicates
- $K = 3$ clusters
- $\pi = (1/6; 1/3, 1/2)$
- γ higher for smaller clusters, density $\bar{\gamma} = \pi^\top \gamma \pi = 1.5 \log(p)/p$
- $G \sim SBM(p, \pi, \gamma)$ conditional on G connected
- $\Omega = \text{Laplacian}(G)$ (+ increases the diagonal until positive-definite)
- $(Y_i)_{i=1\dots n}$ iid $\sim \mathcal{N}(0, \Omega^{-1})$

Inference methods.

oracle: SBM fit on (unobserved) G

vemGlasso: proposed VEM on glasso scores

vemMB: proposed VEM on M-B scores

vemTree: proposed VEM on tree-based edge probabilities

sbmGlasso: pipe-line = SBM on \hat{G}_{glasso} (with $eBIC$ selection)

vemMB: pipe-line = SBM on \hat{G}_{MB} (with ric selection)

vemTree: pipe-line = SBM on \hat{G}_{Tree} (with edge proba $> 2/p$ selection)

6 Illustrations

References

- [1] Elizabeth S. Allman, Catherine Matias, and John A. Rhodes. Identifiability of parameters in latent structure models with many observed variables. *Ann. Statist.*, 37(6A):3099–3132, 12 2009.
- [2] Elizabeth S. Allman, Catherine Matias, and John A. Rhodes. Parameter identifiability in a class of random graph mixture models. *arXiv e-prints*, page arXiv:1006.0826, Jun 2010.
- [3] Alain Celisse, Jean-Jacques Daudin, and Laurent Pierre. Consistency of maximum-likelihood and variational estimators in the stochastic block model. *Electron. J. Statist.*, 6:1847–1899, 2012.
- [4] E. Gassiat, A. Cleynen, and S. Robin. Inference in finite state space non parametric hidden Markov models and applications. *Statistics and Computing*, 26(1-2):61–71, 2016.
- [5] Henry Teicher. Identifiability of finite mixtures. *Ann. Math. Statist.*, 34(4):1265–1269, 12 1963.
- [6] Henry Teicher. Identifiability of mixtures of product measures. *The Annals of Mathematical Statistics*, 38(4):1300–1302, 1967.

A Appendix

A.1 Non-parametric emission distributions

Non-parametric estimates. Given a kernel function κ (s.t. $\int \kappa(x) dx = 1$), we propose to estimate the conditional score pdf f_u ($u = 0, 1$) as

$$\hat{f}_u(s) = \sum_{a < b} w_{ab}^u \kappa(s - S_{ab}), \quad \text{with } \sum_{a < b} w_{ab}^u = 1.$$

For each $u = 0, 1$, the maximisation of the lower bound (2) wrt $w^u = (w_{ab}^u)_{a < b}$ is equivalent to the maximization of

$$\sum_{i < j} h_{ij}^u \log \hat{f}_u(S_{ij}) - \lambda^u \sum_{a < b} w_{ab}^u$$

with $h_{ij}^1 = \sum_{k, \ell} \tau_{ik} \tau_{j\ell} \eta_{ij}^{k\ell}$ and $h_{ij}^0 = \sum_{k, \ell} \tau_{ik} \tau_{j\ell} (1 - \eta_{ij}^{k\ell})$. The derivative wrt w_{ab}^u is zero when

$$\sum_{i < j} h_{ij}^u \frac{\kappa(S_{ij} - S_{ab})}{\hat{f}_u(S_{ij})} - \lambda^u = 0,$$

which has no close form solution. However, close-form updates that increase the log-likelihood are provided in Propr. 3 of [4]. We may check if they still hold for the VEM lower bound.

Alternatively, a pragmatic, unjustified choice is to simply set $w_{ab}^u = \psi_{ab}^u$, that is to let each pair (a, b) contribute to the estimation f_1 (resp. f_0) proportionally to the probability for the edge G_{ab} to be equal to 1 (resp. 0).