Statistical Learning with Complex Data

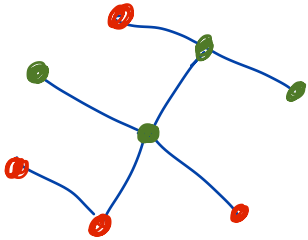
Pr. Charles BOUVEYRON

Professor of Statistics
Chair of the Institut 3IA Côte d'Azur
Université Côte d'Azur & Inria

charles.bouveyron@univ-cotedazur.fr
@cbouveyron

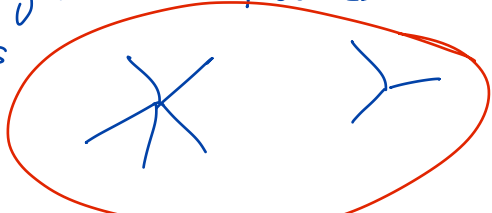
The idea of clustering of networks could be:

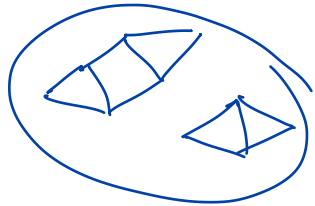
Task 1) group together the nodes of a network which have a similar role in the network



⇒ we will focus on this problem

Task 2) group together networks that have similar structures

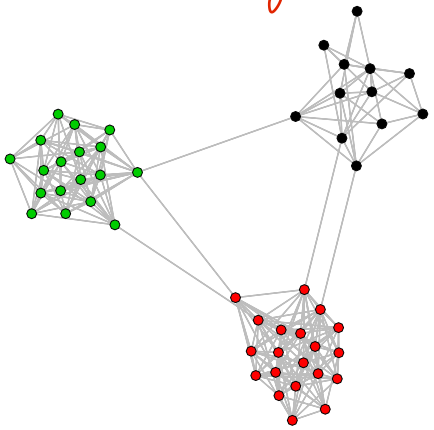


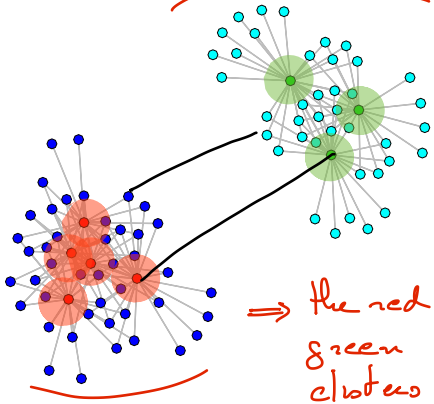


The clustering of networks

a Network made of communities

a network with stars.





⇒ the red and green clusters contain hubs

Figure: Clustering of communities vs. stars.

⇒ the two blue clusters contain nodes that are "govern" by the hubs.

The latent position cluster model (LPCM)

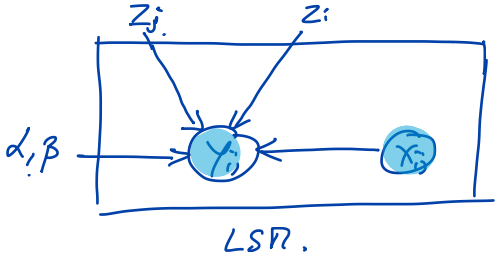
The LPCM extends LSM by adding a clustering structure:

- $\text{logit}\left(P(Y_{ij}=1 \mid \theta)\right) = \alpha + \beta X_{ij} - d(z_i, z_j)$
- $C_i \sim \mathcal{M}(1; \rho_k) \quad k = 1, \dots, K$
- $Z_i \mid C_{ik} = 1 \sim \mathcal{N}(\mu_k, \sigma_k^2 I)$

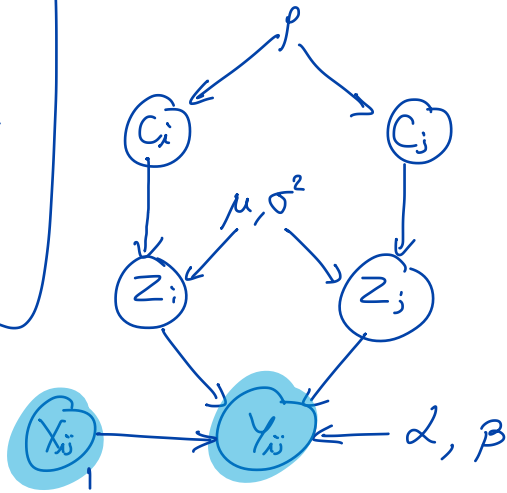
} This a Bayesian extension of LPM.

$$\Rightarrow p(z_i, \theta) \sim \sum_{k=1}^{K} \rho_k \, \mathcal{N}(\mu_k, \sigma_k^2 I)$$

Here the parameter $\rho = (\rho_1, \dots, \rho_K)$ is the vector of cluster proportions.



The inference of LPCR should be done with a MCMC procedure or a VBEM algorithm.



The graphical model for LPCR

Rmk: the VBEM algo is working for large networks (VBLPCM package)

The latent position cluster model (LPCM)

Extension #1: adding a sender/receiver effect

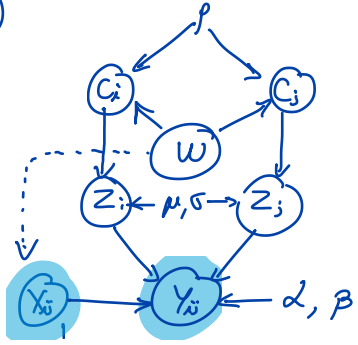
Naturally, it is possible to add to LPCM the sender receiver effect

- $\text{logit}\left(P(Y_{ij}=1|\theta)\right) = \alpha + \beta X_{ij} - d(z_i, z_j) + \delta_i + \gamma_j$
- $c_i \sim \mathcal{M}(1; \rho_k) \quad k = 1, \dots, K$
- $z_i \mid c_{ik} = 1 \sim \mathcal{N}(\mu_k, \sigma_k^2 I)$
- $\delta_i \sim \mathcal{N}(0, \sigma_S^2)$ and $\gamma_j \sim \mathcal{N}(0, \sigma_R^2)$

The latent position cluster model (LPCM)

Extension #2: mixture of experts LPCM : this model assumes that the covariates W may have an effect on the clustering.

$$p(z_i) = \sum_{k=1}^{K} \rho_k(w_i)\, \mathcal{N}(z_i; \mu_k, \sigma_k^2 I)$$



Rmk : this model may be useful in very specific situations and of course comes with difficulties for inference.

The stochastic block model (SBM)

$\Leftrightarrow$ cluster #3.

$C_i = (0,0,1,0,0)$

The SBM model assumes:

$$\begin{cases} C_i \sim \mathcal{M}(1, \rho) \\ Y_{ij} \mid C_{iq} C_{je} = 1 \sim \mathcal{B}(\pi_{qe}), \quad q, e \in \{1, \dots, K\} \end{cases}$$

which indicates the cluster membership of the node i

$C_{iq} = 1$ means that node i belongs to cluster q

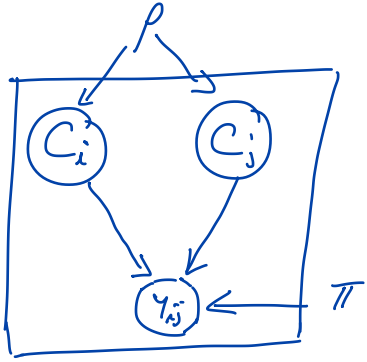
$\pi_{qe} \in [0,1]$ indicates the probability to connect for nodes coming from cluster q and e respectively.

$Y_{ij} = 1$ if $i \to j$ (in the directed case) or if $i \leftrightarrow j$ (in the undirected case)

$Y_{ij} = 0$ otherwise

The stochastic block model (SBM)

The graphical model:

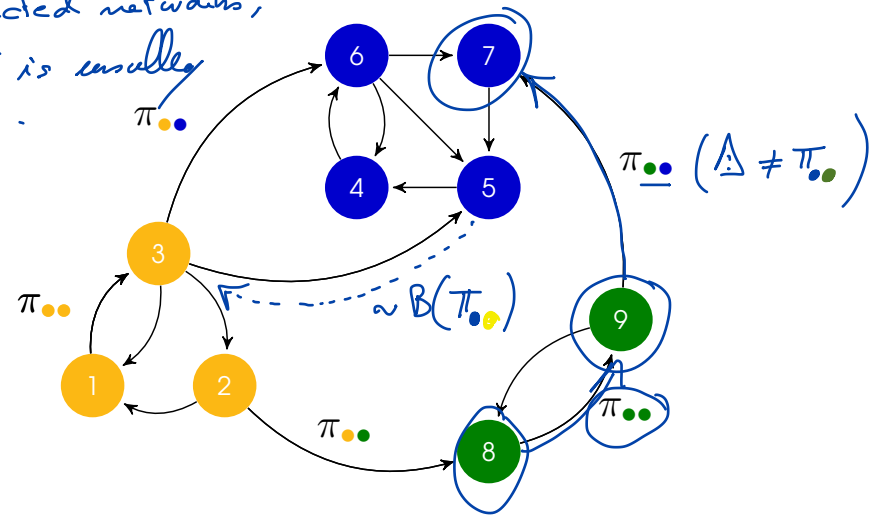


So, the inference of the SBM model will have to estimate from the data y, the parameters ρ and π, plus the cluster memberships C.

The stochastic block model (SBM)

A simple example:

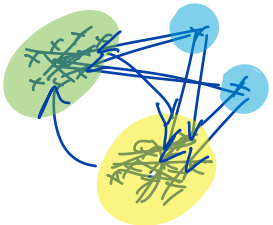
Rmk: for directed networks, the matrix π is usually not symmetric.



One of the most interesting feature of the SBM is that it can model:

- communities : $\pi_{qq} \geq \pi_{ql}$ for $l \neq q$
- stars / hubs : $\pi_{qq} \leq \pi_{ql}$ ——

and a mixture of stars and communities:



$$\pi = \begin{array}{c} \\ 1 \\ 2 \\ 3 \end{array} \begin{pmatrix} 0.01 & 0.4 & 0.35 \\ 0.2 & 0.25 & 0.05 \\ 0.3 & 0.02 & 0.5 \end{pmatrix}$$

→ grp 1 is a group of hubs / influencers

→ grp 2 and 3 are communities.

The stochastic block model (SBM)

Inference: SBM can be infer with :

- MCMC algorithm
- VEM algorithm

and other advanced technique.

⇒ VEM is implemented in the mixer package.

The stochastic block model (SBM)

Choosing the number of clusters: it can be choose thanks to model selection criteria (AIC, BIC, ...)

↳ in practice for this kind of application, the most popular technique is ICL (integrated Classification Likelihood)

$$ICL = BIC - \sum_i \sum_k \hat{t}_{ik} \log(\hat{t}_{ik})$$