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

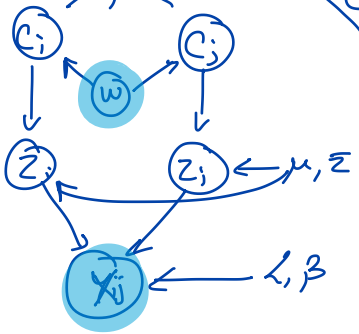
The latent position cluster model (LPCM)

Extension #2: mixture of experts LPCM

This model assumes that some covariates w may have an effect on the clustering

$$p(z_i) = \sum_{k=1}^{K} \pi_k(w_i)\, \mathcal{N}(z_i; \mu_k, \Sigma_k)$$

↑ the prior probability of the groups depends on the individual covariate.



Rmk: This model of course comes with some complications regarding inference

The latent position cluster model (LPCM)

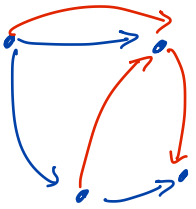
Extension #3: taking into account a dynamic

In order to model real-world networks where interactions may evolve along the time, it is interesting to model this. A way to do that is to assume that the cluster proportions π evolve:

$$\text{logit}(\pi_k(t)) = \alpha_k(t) \sim \mathcal{N}(\alpha_k(t-1), \sigma)$$

this modeling is known as the State Space Model (SSM).

Extension #4: dealing with multi-networks



LSM / LPCM to this situation

$$z_i \sim \sum \pi_k^c \mathcal{N}(\dots)$$

$$\text{logit}(P(X_{ij}^c | \theta)) = \alpha_c - \beta_c |z_i - z_j|$$

The Stochastic Block Model (SBM):

SBM is, at the moment, the most popular and efficient clustering model for networks. SBM has two main interests:

1) it is able to recover both communities and stars at the same time
2) the output of the model can be seen as a network summary (meta-network).

The stochastic block model (SBM)

The SBM model assumes:

- $C_i \sim \mathcal{M}(1; \rho)$
- $X_{ij} \mid C_{ik}=1, C_{jl}=1 \sim B(\pi_{kl})$

where $C_i = (c_{i1}, \dots, c_{iQ})$ with $c_{ik} \in \{0,1\}$ and $\rho = (\rho_1, \dots, \rho_Q)$ are the prior probabilities of the groups.

$C_{jl}=1$ means that j belongs to cluster l.

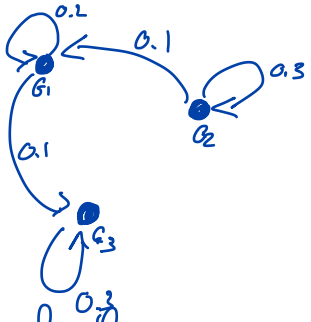
$C_i = (0,0,1,0)$ $\Rightarrow$ i belongs to cluster 3.

π_{kl} is the probability that people from cluster k connect with people from cluster l.

Two remarks:

(i) The matrix Π can be seen as a network between the groups (it adjacency matrix, weighted)

$$\Pi = \begin{pmatrix} 0.2 & 0 & 0.1 \\ 0.1 & 0.3 & 0 \\ 0 & 0 & 0.3 \end{pmatrix}$$

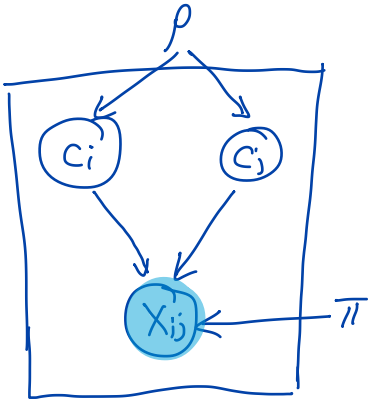


(ii) Π can also indicate if the groups are communities or stars

$$\Pi = \begin{pmatrix} 0.01 & 0. & 0 \\ 0.2 & 0.4 & 0 \\ 0.2 & 0 & 0.5 \end{pmatrix}$$

The stochastic block model (SBM)

The graphical model:



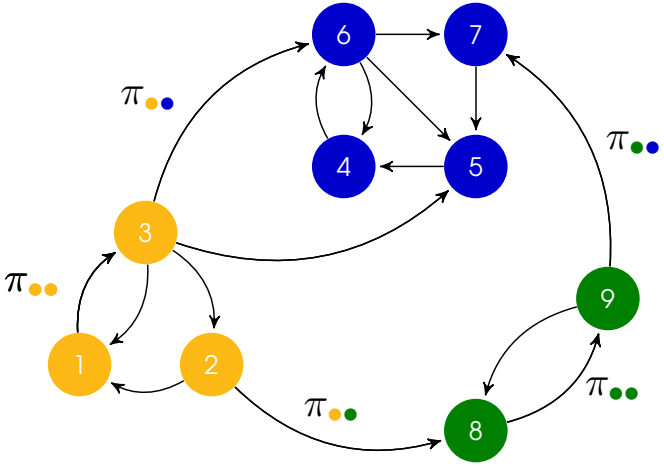
The inference of this model will have to estimate the model parameters, ρ and π, and the latent variables C.

→ Variational EM algorithm.

→ MCMC with a Bayesian version of the model.

The stochastic block model (SBM)

A simple example:



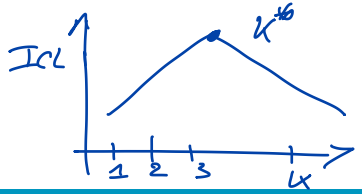
The stochastic block model (SBM)

Choosing the number of clusters:

As for other statistical models, we can rely here on model selection tools:

$$BIC(\mathcal{M}) = \log\left(\mathcal{L}(\hat{\theta})\right) - \frac{\nu(\mathcal{M})}{2}\log(n)$$

$$ICL(\mathcal{M}) = BIC - \sum_i \sum_k c_{ik} \log(c_{ik}).$$



The mixed membership SBM (MMSBM)

The MMSBM extends the SBM as follows; in order to allow people to have different clusters depending on their roles in the network.

- $C_{i\to j} \sim \mathcal{M}(1; \rho_i)$ and $C_{i\leftarrow j} \sim \mathcal{M}(1; \rho_j)$ and $\rho_i \sim \mathrm{Dir}(\alpha)$
- $X_{ij} \mid C_{i\to jk} C_{i\leftarrow jl} \sim \mathcal{B}(\pi_{kl})$