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 space model (LSM)

Adding covariates: $\text{logit}\left(P(X_{ij} \mid \theta)\right) = \alpha + \beta Y_{ij} - d(z_i, z_j)$

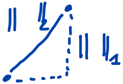
The covariate Y_{ij} can be used to provide extra information to the model on the pairs of nodes. For instance:

- Y_{ij} is the nb of years in common in a club/society between i and j.
- a type of relationship (categorial var) ~ $Y_{ij} \in \{1, \dots, K\} \Longrightarrow Y_{ij} = (0, 0, 1, 0, 0) \Rightarrow \beta$ is a vector
  ↑ $Y_{ij} = 3$

Choice of the distance:

Another way to extend this model is to play with the definition of the distance within the latent space

- $d(z_i, z_j) = \|z_i - z_j\|_2$ or $\|z_i - z_j\|_2^2$
- $d(z_i, z_j) = \|z_i - z_j\|_1$ (Manhattan distance)



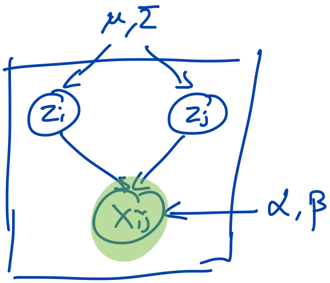
Modifying the model: A specific and interesting case is the situation of directed networks, in which there are the roles of sender and receiver. It is naturally interesting to model this. A way to do that:

$$\text{logit}\left(P(X_{ij}=1 \mid \theta)\right) = \alpha + \beta Y_{ij} - d(z_i, z_j) + \underbrace{\delta_i + \gamma_j}_{\text{a sender-receiver effect}}$$

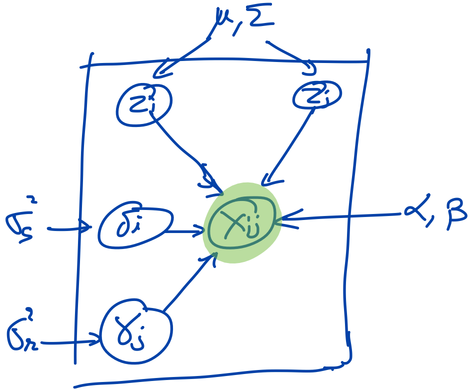
$$\text{where} \begin{cases} d_i \sim N(0, \sigma_s^2) & \leftarrow \text{the prior for the propensity to send messages} \\ \gamma_j \sim N(0, \sigma_r^2) & \leftarrow \text{the prior for receiving messages.} \end{cases}$$

Rmk: This model is highly parametrized: it has $(3n+2)$ parameters to estimate.

Exercise: draw the graphical model for this LSM version.



Bayesian LSM.



Outline

1. Introduction

2. Characterization and manipulation of networks

3. The visualization of networks

4. Clustering of networks

5. Texts

6. Images

The clustering of networks

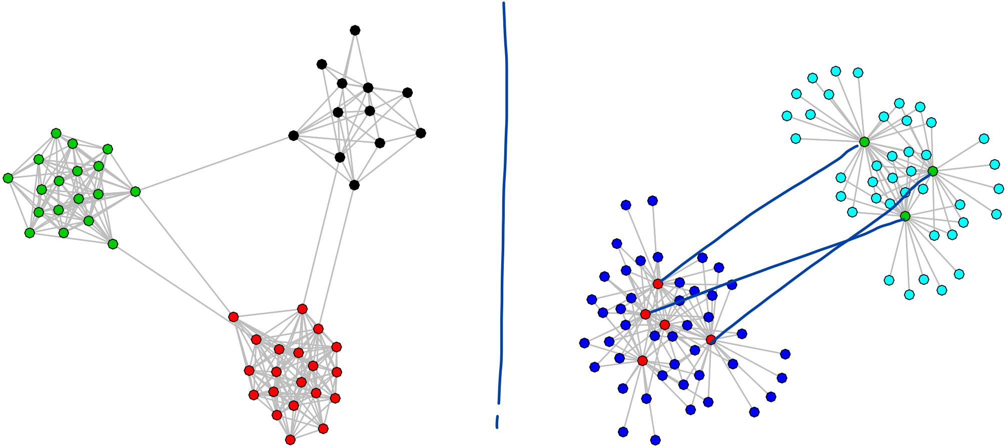
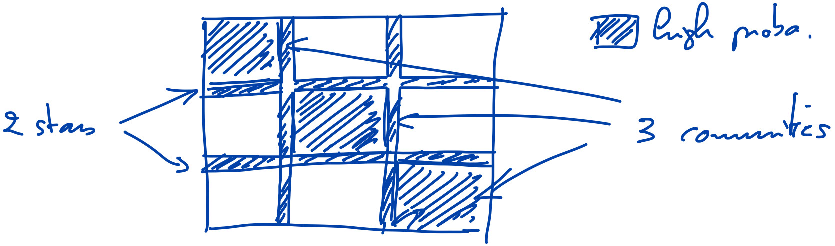


Figure: Clustering of communities vs. stars.

The clustering of networks

Difference between communities and stars:

- in communities, people have a higher probability connection within the community than with other communities
- stars are people that connect less within the group than outside the group.



The latent position cluster model (LPCM)

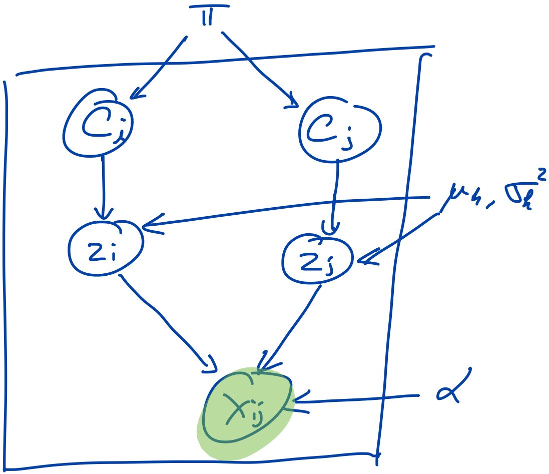
The LPCM extends LSM by adding a clustering structure:

- $\operatorname{logit}\left(P(X_{ij}=1 \mid \theta)\right) = \alpha - d(z_i, z_j)$
- $C_i \sim \mathcal{M}(1; \pi)$ where π_k is the prior probability for cluster k, $k \in \{1, \dots, K\}$
- $Z_i \mid C_{ik} = 1 \sim \mathcal{N}(\mu_k, \sigma_k^2 I)$

$$\Rightarrow \quad Z_i \sim \sum_{k=1}^{K} \pi_k \, \mathcal{N}(\mu_k, \sigma_k^2 I) .$$

The latent position cluster model (LPCM)

The model:



clustering

visualization...

⇒ the inference of this Bayesian model has to be done using MCMC or advanced inference strategies (VBEM)