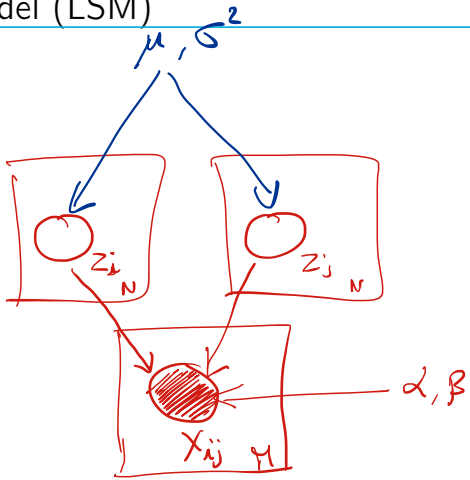
Statistical Learning with Complex Data

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)



N: nb of nodes

M: —— edges

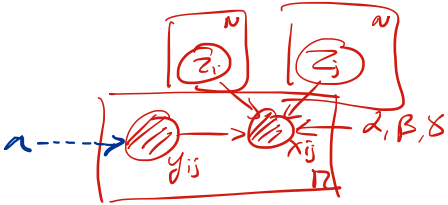
$\Rightarrow$ More efficient inference techniques relies on Bayesian version of the model $\Rightarrow$ $z_i \sim N(\mu, \sigma^2 I_p)$

The latent space model (LSM)

Adding covariates: $\text{logit}(P(X_{ij}=1|G)) = \alpha - \beta\|z_i - z_j\|^2 + \gamma y_{ij}$

where y_{ij} may indicate a difference between the nodes i and j

Ex: $y_{ij} = |Age_i - Age_j|$



Choice of the distance:

- Euclidean distance
- directional distance

$$d(z_i, z_j) = |z_i| \cos(\widehat{z_i z_j}) = \frac{z_i^t z_j}{|z_j|}$$

More adapted to directed network with a strong asymmetry

The latent space model (LSM)

A more tenable inference approach: will rely on a Bayesian modeling of the LSM $\Rightarrow$ MCMC procedure may be used to infer the model.

$\Rightarrow$ "latentnet" package implements a MCMC algorithm for the LSM model.

Outline

The clustering of networks

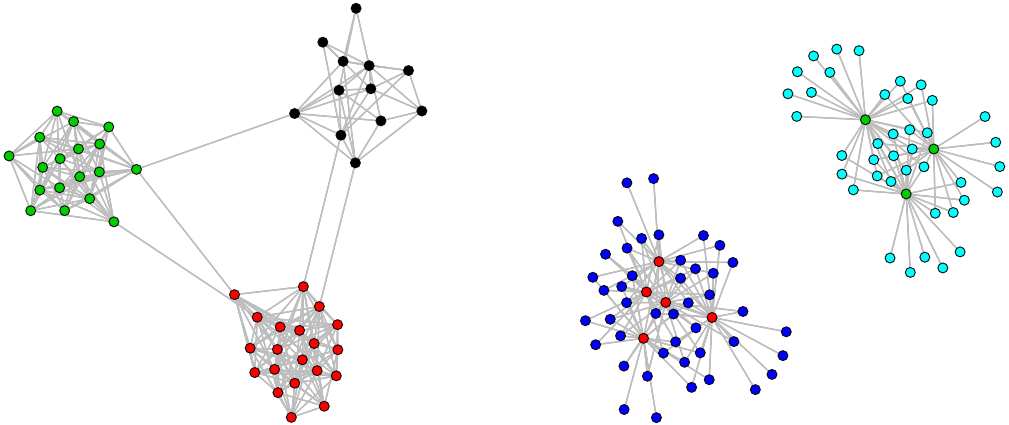


Figure: Clustering of communities vs. stars.

The clustering of networks

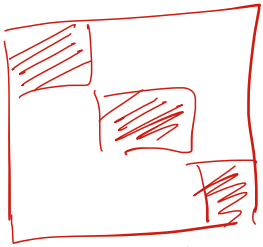
Difference between communities and stars:

- Community : $P(X_{ij}=1 \mid c_i = c_j) \geq P(X_{ij}=1 \mid c_i \neq c_j)$

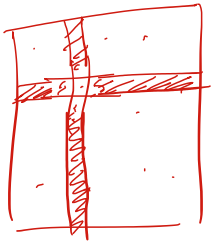
  ↳ assortative mixing

- Stars : $P(X_{ij}=1 \mid c_i \neq c_j) \geq P(X_{ij}=1 \mid c_i = c_j)$

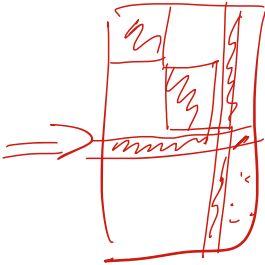
  ↳ dissortative mixing



Communities



Stars



The latent position cluster model (LPCM)

The LPCM extends LSM by adding a clustering structure: (Hoff, Raftery & Handcock, 2002)

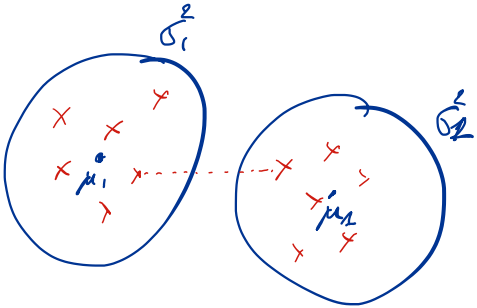
- $\text{logit}\left(P(X_{ij}=1 \mid \theta)\right) = \alpha - d(z_i, z_j) + \gamma y_{ij}$
  - $c_i \sim \mathcal{M}(\tau)$, where c_i indicates the group membership of i.

    $c_i = (0,0,1,0,0) \Leftrightarrow i$ belongs to cluster #3.
  - $z_i \mid c_{ik}=1 \sim N(\mu_k, \sigma_k^2 I_p)$

$$\Leftrightarrow z_i \sim \sum_{k=1}^{K} \tau_k N(\mu_k, \sigma_k^2 I_p)$$

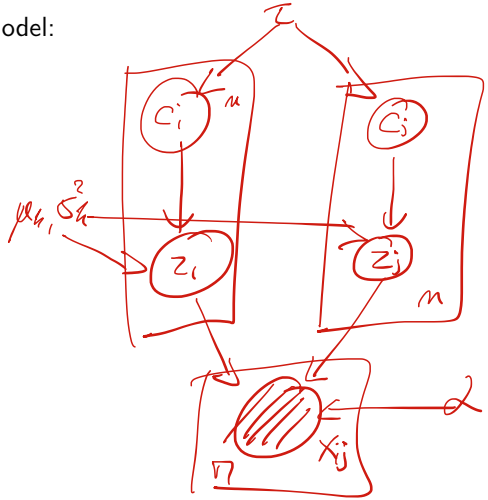
The latent position cluster model (LPCM)

The model:



The latent position cluster model (LPCM)

The graphical model:



The latent position cluster model (LPCM)

Inference:

- MCMC procedure → latentnet package
- VBEM procedure → VBLPCM ——

⚠ MCMC is not adapted to large networks (< 500 nodes)

⇒ VBEM may be used for networks smaller than 20 000 nods.

The latent position cluster model (LPCM)

In R, the *latentnet* and *VBLPCM* packages allow to use it:

The latent position cluster model (LPCM)

Extension #1: adding a sender/receiver effect

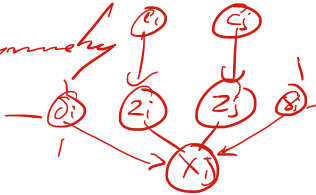
receiver effect

$$\text{logit}\left(P(X_{ij}=1 \mid \theta)\right) = \alpha - d(z_i, z_j) + \gamma_i + \delta_j$$

sender effect $\simeq$ prior probability for i to send a message to anyone.

with $\begin{cases} \gamma_i \sim N(\mu_\gamma, \sigma_\gamma^2) \\ \delta_j \sim N(\mu_\delta, \sigma_\delta^2) \end{cases}$

$\Rightarrow$ This model allows to break the symmetry of the LPCM / LSM model

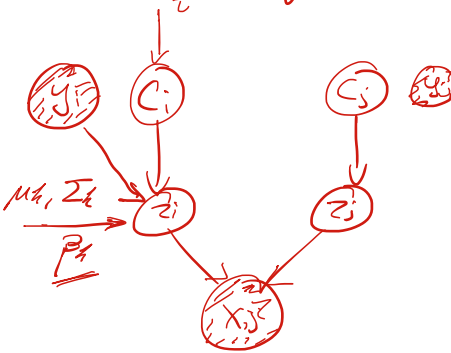


The latent position cluster model (LPCM)

Extension #2: mixture of experts LPCM incorporates covariate information within the distance model.

$$z_i \sim \sum_{k=1}^{K} \pi_k(y_i)\, \phi(z; \mu_k, \Sigma_k)$$

where $$\pi_k(y_i) = \frac{\exp(\beta_k^t y_i)}{\sum_{\ell=1}^{K} \exp(\beta_\ell^t y_i)}$$



$\Rightarrow$ MCMC / VBEM procedures to solve this problem.

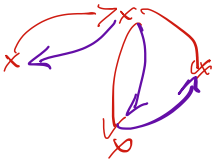
The latent position cluster model (LPCM)

Extension #3: taking into account a dynamic

$$\left(X_{ij}^{(t)}\right)_{t=1,\dots,T} \longrightarrow \pi_k^{(t)}$$

$$\longrightarrow \pi_k^{(w_s)}$$ where w_s is a cluster of times.

Extension #4: dealing with multi-networks



→ LVM 2 net package which implements an extension of the LPCM model to this situation.

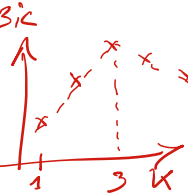
The latent position cluster model (LPCM)

Extension #1: adding a sender/receiver effect

The latent position cluster model (LPCM)

Choosing the number of clusters: Model selection is a theory which is based on penalized-likelihood criteria.

The most famous criterion is BIC:



$$\mathrm{BIC}(\mathcal{M}) = \log \mathcal{L}(\hat{\theta}) - \frac{\gamma(\mathcal{M})}{2} \log(n)$$

in the network context, the nb of edges

where $\gamma(\mathcal{M})$ is the nb of free parameters in the model.

For instance: $\mathrm{LPCM} = 1\alpha + K\mu + K\sigma^2 + (K-1)\pi \Rightarrow 3K$

The stochastic block model (SBM)

The SBM model ~~assumes~~: allows to deal with:

- networks which are directed or not,
- ——— that contain communities but also stars.

It assumes:

- C_i indicates the group membership of node i

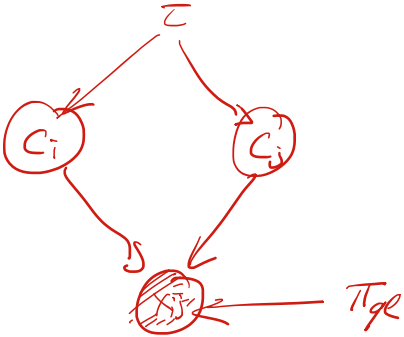
$$C_i \sim \mathcal{M}(\pi) \rightarrow C_i = (0,0,1,0)$$

- the connection between i and j:

$$X_{ij} \mid C_{iq} C_{jl} = 1 \sim \mathcal{B}(\pi_{ql})$$

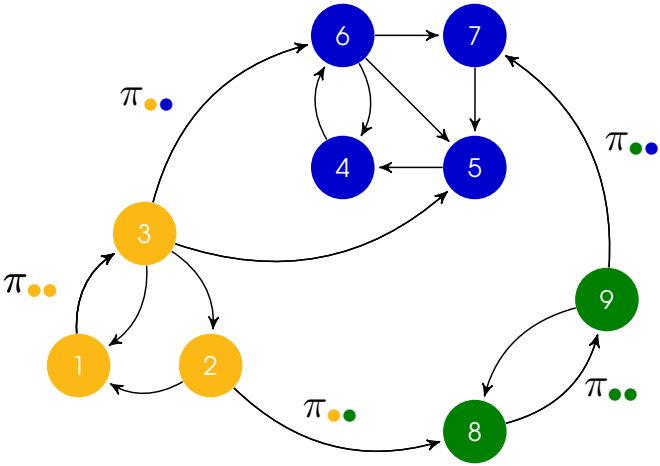
The stochastic block model (SBM)

The graphical model:



The stochastic block model (SBM)

A simple example:



The stochastic block model (SBM)

Inference: MCMC or V(B)EM procedures are used for fitting a SBM model from data.

Rmk:

if cluster #q is a hub, what are the values π_{ql}, $\forall l = 1 \dots K$?

$$\pi_{ql} \geq \pi_{qq} \quad \forall l \neq q$$

(and π_{lq})

The stochastic block model (SBM)

Choosing the number of clusters:

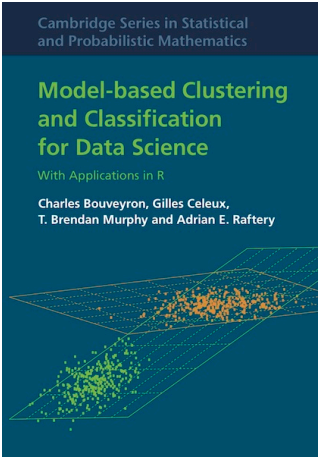
→ BIC

→ ICL

References

M. Salter-Townshend, A. White, I. Gollini and T. B. Murphy, *Review of Statistical Network Analysis: Models, Algorithms, and Software*, Statistical Analysis and Data Mining, Vol. 5(4), pp. 243–264, 2012.

References (more seriously ;-)



(Chapter 10 is devoted to network analysis!)