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: Y_{ij}

- a nb of years in common in a society, ...
- a type of relationship (categorical var)

$Y_{ij} \in \{1, \dots, K\} \longrightarrow \tilde{Y}_{ij} = (0,0,1,0,0) \Rightarrow \beta$ is now a vector.

$Y_{ij} = 3$.

Choice of the distance: $\| z_i - z_j \|_2^2$

It could be an Euclidean distance $\| \ \|_2$ (the most natural), but it could be any other distance if you prefe ($\| \ \|_1$, ...)

It is possible to compute afterward $P(Y_{ij}=1|\hat{\theta})$

$$\text{Logit}\left(P(Y_{ij}=1|\theta)\right) = \alpha - d(z_i, z_j)$$

$$\Leftrightarrow \log\left(\frac{p}{1-p}\right) = \alpha - d(z_i, z_j)$$

$$\Leftrightarrow p/(1-p) = \exp\left(\alpha - d(z_i, z_j)\right)$$

$$\Leftrightarrow p = \exp(-) - p\exp(-)$$

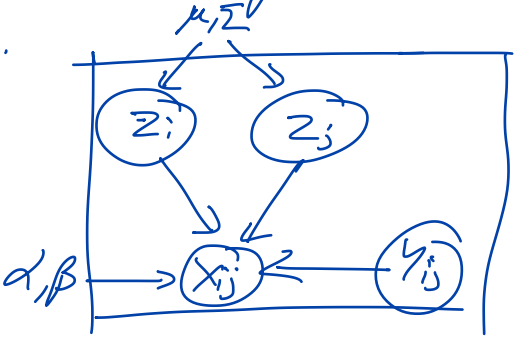
$$\Leftrightarrow p\left(1+\exp(-)\right) = \exp(-)$$

$$\Leftrightarrow P(Y_{ij}=1|\hat{\theta}) = \frac{\exp\left(\hat{\alpha} - d(\hat{z}_i, \hat{z}_j)\right)}{1+\exp\left(\hat{\alpha} - d(\hat{z}_i, \hat{z}_j)\right)}$$

The latent space model (LSM)

LSM in R via the *latentnet* package: which implements the original approach of Hoff et al (2001) ⇒ MCMC for a Bayesian version of LSM.

= LSM + $Z_i \sim N(\mu, \Sigma)$



and the VBLPCM package implements a VBEM algorithm for this model.

It is possible to extend the LSM model :

- by adding covariates on the nodes :

$$\text{Logit}\left(P(Y_{ij}=1 \mid \theta)\right) = \alpha + \beta X_{ij} - d(z_i, z_j)$$

where X_{ij} is a covariate about the pair of nodes

Ex: $X_{ij} = |age_i - age_j|$ or $X_{ij} = \mathbb{1}\{pos_i = pos_j\}$

- by "playing" with the notion of distance :

- $d(z_i, z_j) = \|z_i - z_j\|^2$ is ok for symmetric graphs.
- $d_i(z_i, z_j) = |z_i| \cos \theta_{ij}$ where $\theta_{ij} = \frac{z_i^t z_j}{|z_j|}$ is better for asymmetric networks

- by modifying the model :

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

where $\begin{cases} \delta_i \sim \mathcal{N}(0, \sigma_s^2) \\ \gamma_j \sim \mathcal{N}(0, \sigma_r^2) \end{cases}$

σ_s^2 ← propensity to send connections

σ_r^2 ← to receive connections.

⇒ this is of course for directed networks $\left(Y_{ij}=1 \not\Rightarrow Y_{ji}=1\right)$.

Rmk: this is a quite highly parametrized model

⇒ this model has $(3n+2)$ parameters to estimate and requires a lot of connections to estimated properly!