Statistical Learning with Complex Data

Pr. Charles BOUVEYRON

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

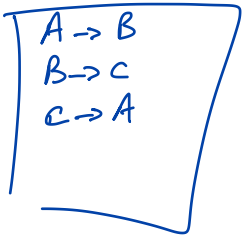
charles.bouveyron@univ-cotedazur.fr
@cbouveyron

Outline

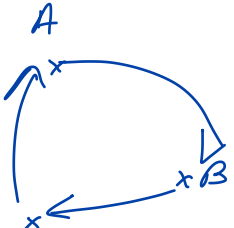
The visualization of networks

First, it is not an easy task to visualize a network and this is even a difficult task in general!

The visualization of networks is at least as difficult than the visualization of quantitative data.



Visualization = how to position the nodes in a quantitative space such that we can visualize easily all interactions



The High-school network of Moreno: this is a visualization of a dataset he collected in the 30's.

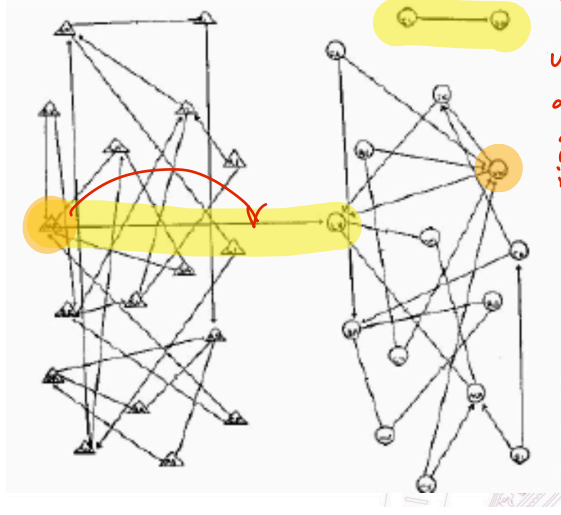
MANY MISFITS REVEALED

Dr. J. L. Moreno Calculates There Are 10 to 15 Million Isolated Individuals in Nation.

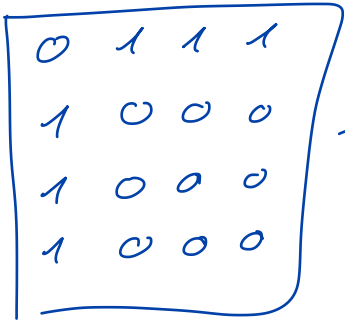
A new science, named psychological geography, which aims to chart the emotional currents, cross-currents and under-currents of human relationships in a community, was introduced here yesterday at the

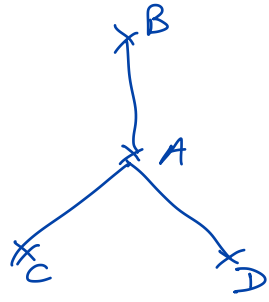
EMOTIONS MAPPED BY NEW GEOGRAPHY

Charts Seek to Portray the Psychological Currents of Human Relationships.



This visualization was hand made and it is quite good even though it could be slightly improved.





The question is clearly about finding a good position of the nodes in a "latent space"

MDS for visualizing networks

Multidimensional Scaling (MDS) is a method used for the visualization of any kind of data (networks, quantitative, categorical, texts ...) for which we can compute a distance.

MDS can be in particular used for dimension reduction of quantitative data.

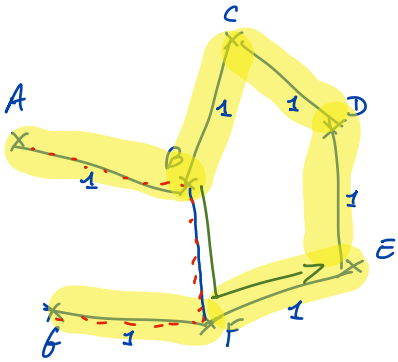
MDS for visualizing networks

The idea of MDS is to find a representation of the data in a low-dimensional space (p) which preserves the topology of the input data.

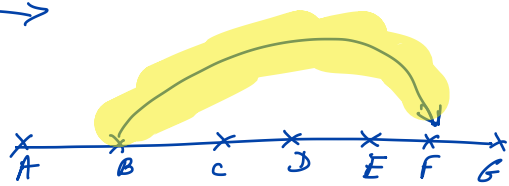
In practice, the distance between two data points in the latent space should be as close as possible to the distance of these two points in the original space:

$$\min_{z} \sum_{i=1}^{n} \sum_{j=1, j \neq i}^{n} \| d(x_i, x_j) - \delta(z_i, z_j) \|^2$$

where x_i is the original data and z_i are the projection in the visualization space.



MDS



$d(A,G) = \cancel{6}\ 3$ $\quad d(B,E) = 2$

The distance of the shortest path

$\delta(A,G) \simeq 6.$

For this new configuration, the straight line is not working anymore and DDS will have to find a compromise!

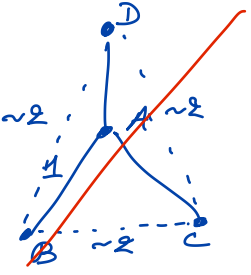
Exercise: "use" NDS to position the nodes with the following network structure:

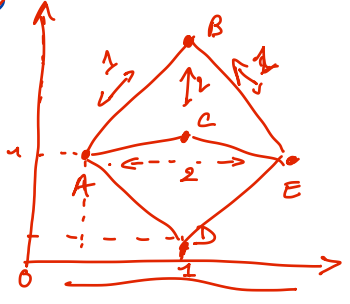
	A	B	C	D	E
A	0	1	1	1	0
B	1	0	0	0	1
C	1	0	0	0	1
D	1	0	0	0	1
E	0	1	1	1	0

Adj matrix

	A	B	C	D	E
A	0	1	1	1	2
B		0	2	2	1
C			0	2	1
D				0	1
E					0

Distance of the shortest path





MDS offers a single solution to visualize networks:

⊕ easy and allows to quantify the deformations.

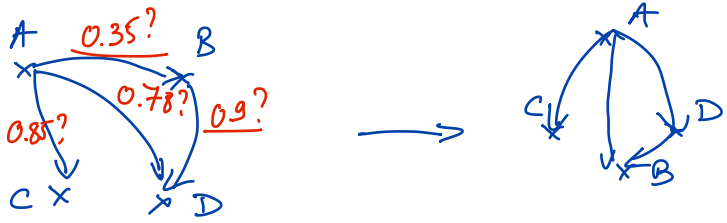
⊖ no way to choose the appropriate dimensionality for the visualization space ($p = 2$ or 3)

⊖ this is a deterministic approach that does not take into account the possible uncertainty on the data (on the edges).

The latent space model (LSM) (Hoff, Handcock and Raftery, 2001)

LSM is one of the first statistical model proposed to model and visualize networks.

One of the key features of LSM is its ability to model the uncertainty on the observed edges.



The latent space model (LSM)

The goal of the latent space model is to provide a latent representation of the data such that:

- two points that are close in the space are very likely to connect
- two points that are far away in the latent space will have a low probability to connect.

The latent space model (LSM)

Reminder: $\text{logit}(p) = \log\left(\frac{p}{1-p}\right)$

Let first consider the random variable X:

$$\begin{cases} X_{ij} = 1 & \text{if } i \text{ is connected to } j \\ X_{ij} = 0 & \text{if they are not connected} \end{cases}$$

The LSM model assumes that the probability of $X_{ij}=1$ is:

$$\text{logit}\left(P(X_{ij}=1 \mid \theta)\right) = \alpha + \beta Y_{ij} - \|z_i - z_j\|^2$$

where z_i, z_j are the coordinates of the nodes i and j in the latent space, and Y_{ij} is some covariates about the pair (i,j) and $\theta = \{\alpha, \beta, z_1, z_2, \dots, z_n\}$.

The latent space model (LSM)

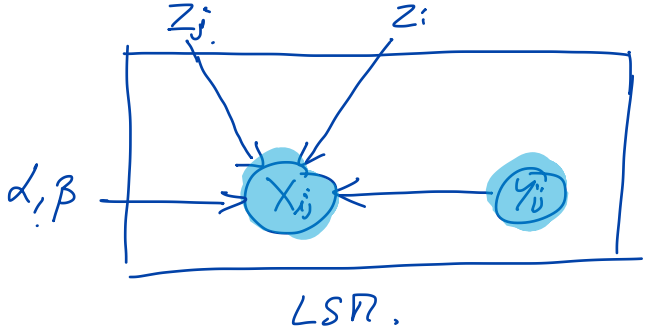
Thanks to this modeling, the LSM model will put close together nodes that have a high probability to connect.

The covariate Y_{ij} may be for instance some (possibly multivariate) information about the pair:

Y_{ij} = nb of years they know each other.

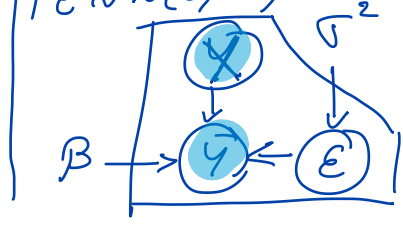
In this model the data are $(X_{ij})_{i,j=1,\dots,n}$ (and eventually the (Y_{ij})) and the model parameters are $\theta = \{\alpha, \beta, z_1, \dots, z_n\}$.

The graphical model for the LSM is



Linear regression

$$Y = \beta^t X + \varepsilon$$
$$\varepsilon \sim N(0, \sigma^2)$$



How to estimate model parameters?

It is possible to use the Maximum Likelihood approach to estimate $\{\alpha, \beta, z_1, \ldots, z_n\}$ from the data:

$$\log\left(\mathcal{L}(X,\Theta)\right) = \sum_{\substack{i=1 \\ j=1 \\ i \neq j}}^{n} \left[x_{ij}\left(\alpha + \beta y_{ij} - d_{ij}^2\right) - \log\left(1 + \exp\left(\alpha + \beta y_{ij} - d_{ij}^2\right)\right) \right]$$

where $d_{ij}^2 = \|z_i - z_j\|^2$.

Unfortunately, there is no closed-form solution for $\hat{\Theta}_{ML}$ and we have to numerically optimize this function.

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.

$Y_{ij} = 3$.

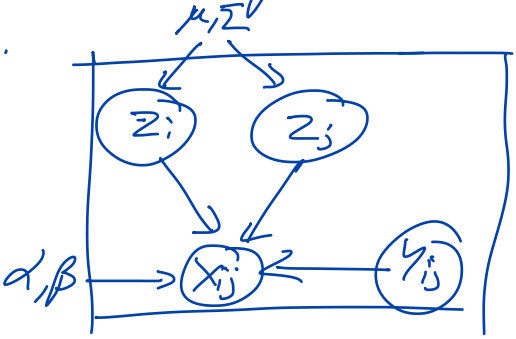
Choice of the distance: $\| z_i - z_j \|_2^2$

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

The latent space model (LSM)

LSM in R via the *latentnet* package: with 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.