Statistical Learning with Complex Data

Pr. Charles BOUVEYRON

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

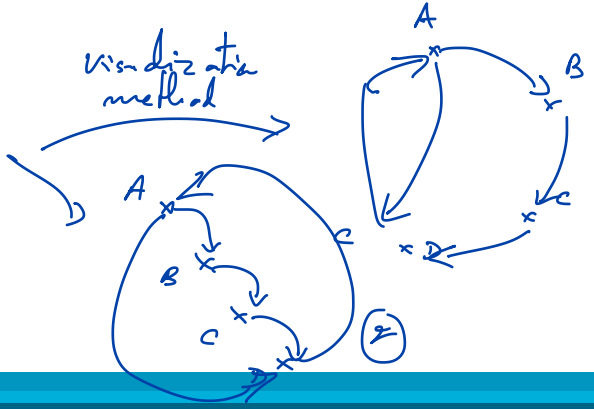
charles.bouveyron@univ-cotedazur.fr
@cbouveyron

Outline

The visualization of networks

First of all, it is important to understand that the visualization of a given network is not a trivial task. It is even a very difficult task if the network is dense.

A → B
B → C
C → D
D → A
A → D



From the previous example, it is obvious that positioning the nodes in a proper way (having a clear visualization of the relationships) is a difficult task.

For visualisation, we have different extensions of existing dimension reduction methods a dedicated statistical methods.

- MDS
- LSM

MDS for visualizing networks

Multidimensional Scaling (MDS) is a method used for the visualization of any kind of data (networks, quantitative, texts, images, ...) for which you are able to define a distance between the observations.

⇒ MDS has been quite popular for the dimension reduction of high-dimensional data.

MDS for visualizing networks

The goal of MDS is to find a low-dimensional representation of the data which is preserving the topology of the original data.

In practice, MDS looks for a positioning of the data points such that the distance between the points in the low-dimensional space are as close as possible than the distances in the original space.

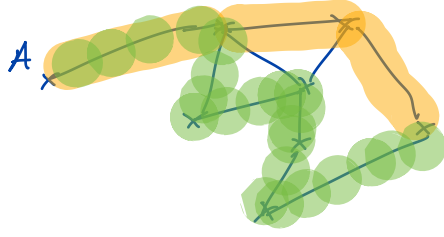
The translation of this problem in equations:

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

where d and δ are respectively the distances between the observations in the original and representation spaces, and $z_1, \dots, z_n$ are the latent representations of the data points $x_1, \dots, x_n$.

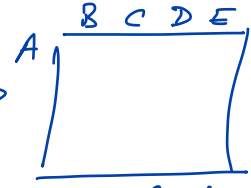
Applying this to a network requires to define:

- d: it could be the shortest-path distance on the graph.

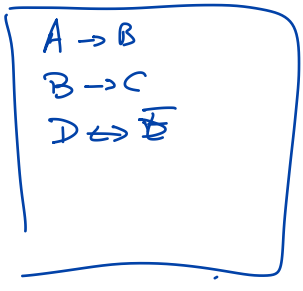


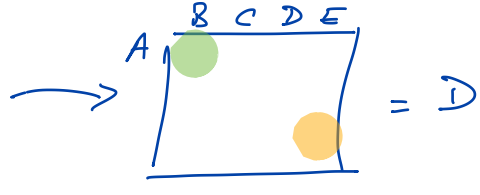
$d(A,B) = \min(3,6,5,4)$

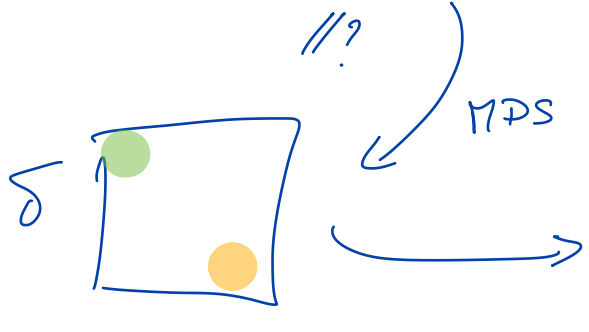
$= 3$



- δ: it could be simply the Euclidean distance in $\mathbb{R}^p$ ($p=2$?)


A → B
B → C
D ↔ E


B C D E
A
= D


MDS
A
B
C
D
E

Summary on MDS:

⊕ it is a generic method, working well for network but also other kinds of data.

⊕ the knowledge of the quality of the representation of the pairs is interesting in practice.

⊖ MDS could be limited for representing very complex networks.

⊖ MDS is not able to model a possible uncertainty on the observed edges.

Exercise: Use TDS to position the following nodes

	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

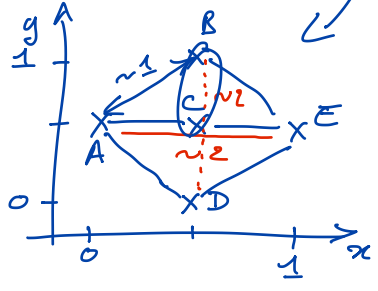
→

	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

= D

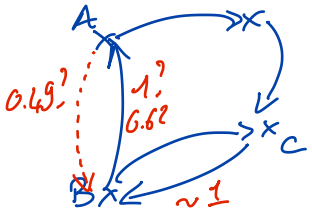
$\delta =$

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



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

LSM is the first statistical method ever proposed to visualize and model a network. This method in particular takes into account the possible uncertainty on the observed edges.



The latent space model (LSM)

The goal of the latent space model is two-fold: we would like to find a latent representation of the data points such that:

i) points that are close together should have a high probability to connect

ii) points that a far away should have a low probability to connect.

The latent space model (LSM)

Once again, translating this in equations leads to:

Let suppose that X_{ij} is a random variable such that:

$$\begin{cases} X_{ij} = 1 \text{ if } i \sim j \\ X_{ij} = 0 \text{ if } i \not\sim j \end{cases}$$

In this case, the LSM model assumes that:

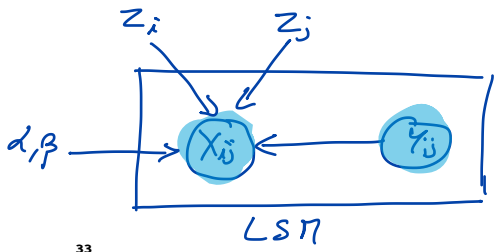
$$\text{logit}\left(P(X_{ij}=1|\theta)\right) = \log\left(\frac{P(X_{ij}=1|\theta)}{P(X_{ij}=0|\theta)}\right) = \alpha + \beta Y_{ij} - \|z_i - z_j\|^2$$

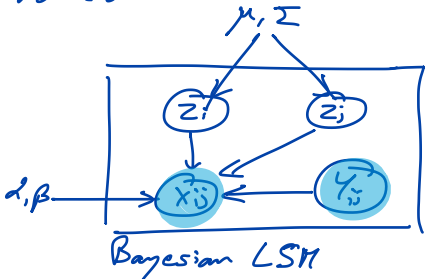
where α is a prior probability to connect, z_i is the latent position of the node i, Y_{ij} is a covariate information on the pair.

The latent space model (LSM)

In this model, the data are the pairs X_{ij}, $i = 1 \dots n$ and $j = 1 \dots n$, (the adjacency matrix) and the parameters of the model are $\alpha, \beta, z_1, \dots, z_n$.

Rmq: even though we have $n+2$ parameters to estimate we can use n^2 edge data to estimate them.





Inferring the LSM or the Bayesian LSM models requires to use either Maximum Likelihood or MCMC methods.

For the LSM model, the Log-Likelihood will have the following form:

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

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

Unfortunately, as for logistic regression, there is no closed-form solution and we have to rely on a optimization algorithm to maximize this function.

To summarize:

⊕ LSM both model the uncertainty of the edges while providing a visualization of the network.

⊕ LSM offered a first basic statistical model as a basis for a lot of extensions.

⊖ LSM is just able to model communities and not stars.