Latent variable models in biology and ecology

Chapter 4: Stochastic Block Models and Latent Block Models

Sophie Donnet. **INRAE**

Master 2 MathSV. February 8, 2022

Introduction

Descriptive statistics
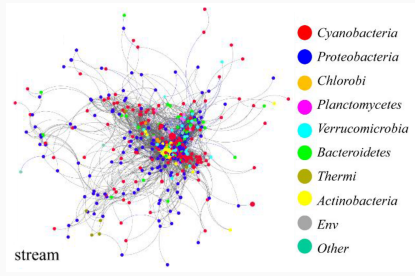
Probabilistic model

Inference

Networks can account for

- **Ecological networks** : Food web,
  Co-existence networks, Host-parasite
  interactions, Plant-pollinator
  interactions,
- **Social networks**
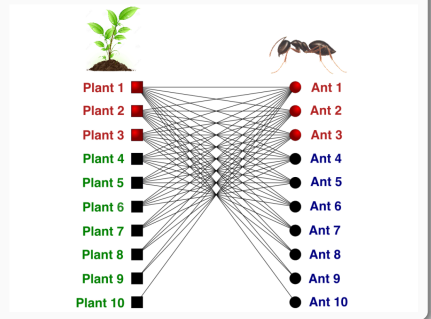- **Inventory datasets**
- ...

Networks may be or not bipartite: Interactions between nodes belonging to the same or to different functional group(s).

Simple network

Bipartite network

Terminology

A network consists in:

- nodes/vertices which represent individuals / species /ships which may interact or not,
- links/edges/connections which stand for an interaction between a pair of nodes / dyads.
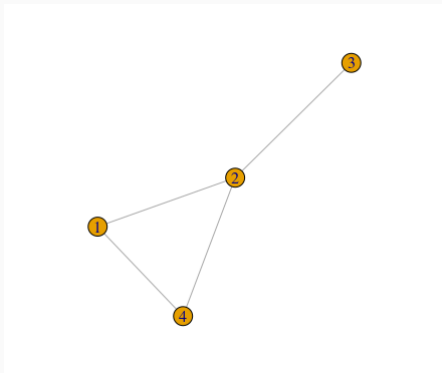
A network may be

- directed / oriented (e.g. food web...),
- symmetric / undirected (e.g. coexistence network),
- with or without loops.

This distinction only makes sense for simple networks (not bipartite).

Network representation and adjacency matrix

For a non-directed network

$$Y = \begin{pmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \end{pmatrix}$$
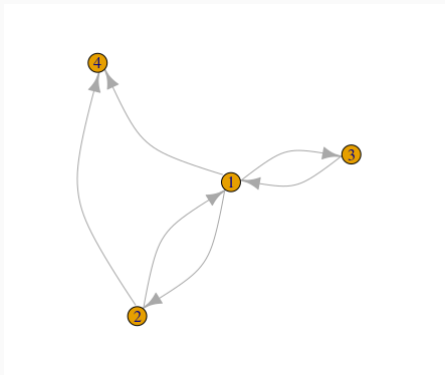

- n rows and n columns,
- symmetric matrix

Network representation and adjacency matrix

For a directed network

$$Y = \begin{pmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$
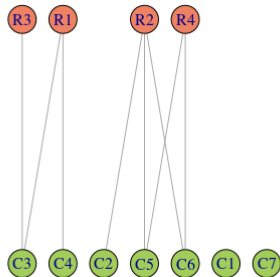

- n rows and n columns,
- non symmetric matrix

Bipartite network and incidence matrix

$$Y = \begin{pmatrix} 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 \end{pmatrix}$$

- n rows and m columns, rectangular matrix.

- the network provided as:
  - an adjacency matrix (for simple network) or an incidence matrix (for bipartite network),
  - a list of pair of nodes / dyads which are linked.
- some additional covariates on nodes, dyads which can account for sampling effort.

Goal

- Unraveling / describing / modeling the network topology.
- Discovering particular structure of interaction between some subsets of nodes.
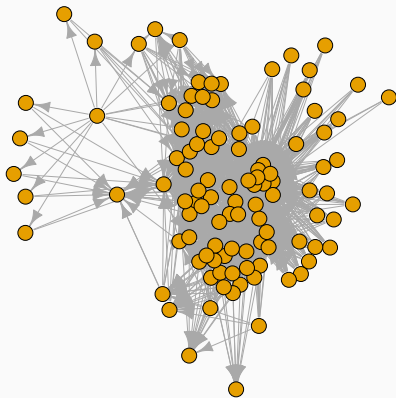- Understanding network heterogeneity.
- Not inferring the network !

Aim : give a short description of the network, give a hint about its structure, look for heterogeneity in the connections

- Many metrics supplied for simple networks
- Have been extended to bipartite networks

[Kéfi et al., 2016]

Degree

$$\begin{array}{rcl rcl}
\deg(u) &=& \sum_{v \in V}(u \leftrightarrow v), & \deg(v) &=& \sum_{u \in U}(u \leftrightarrow v) \\
\deg_i &=& \sum_{j=1}^{|V|} Y_{ij} & \deg_j &=& \sum_{i=1}^{|U|} Y_{ij}
\end{array}$$

- Nodes with high degree are hubs
- Nodes with null degree are isolated
- If edges are oriented : in- and out- degrees can be computed.

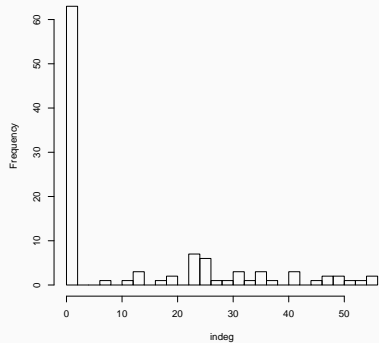## Out degree distribution

Histogram of outdeg
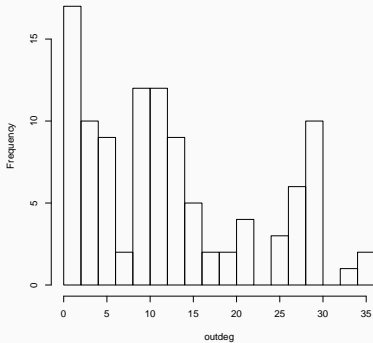
In degree distribution

Histogram of indeg

Closeness centrality

Property on a node

Definition

Determine whether a node can communicate with other nodes of the network directly or through the short paths.

$$C(u) = \frac{1}{\sum_{w \in U \cup V} d(u, w)}$$

where $d(u, w)$ is the length of the shortest path between u and w (through the network).

Note that, for bipartite networks

- A node $u \in U$ can have a minimum distance of 1 with $v \in V$.
- A node $u \in U$ can have a minimum distance of 2 with $u' \in U$.
- All paths between nodes of the same set are of even length.

Betweenness centrality

Property on a node

Definition

Betweenness centrality quantifies the number of times a node acts as a bridge along the shortest path between two other nodes.

The betweenness of a vertex v is computed as follows.

- For each pair of vertices (w, w'), compute the shortest paths between them. $\delta_{w,w'}$ is the number of shortest paths between (w, w')
- For each pair of vertices (w, w'), determine the fraction of shortest paths that pass through v : $\frac{\delta_{w,w'}(v)}{\delta_{w,w'}}$
- Sum this fraction over all pairs of vertices (w, w').

$$B(v) = \sum_{w \neq w' \neq v} \frac{\delta_{w,w'}(v)}{\delta_{w,w'}}$$

Betweenness centrality
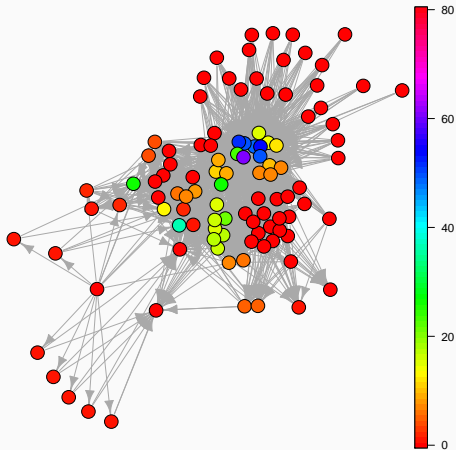
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.000	0.000	0.000	6.604	6.929	59.570

Property on the network

Definition

- Important property in ecology
- Defined as a pattern of interactions in which specialists (e.g. pollinators that visit few plant species) interact with plants that are visited by generalists.
- Mathematically, looking for a reordering of rows and columns such that Y is nested
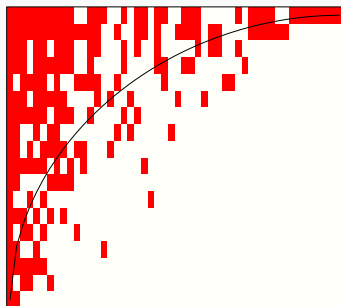
- more generally used on incidence matrices,
- significance of the nestedness index computed by random permutations of the matrix,
- this food web is found to be nested.

Property on the network

Definition

Existence of clusters (blocks, module, communities) where nodes are much more connected than with other clusters

Modularity

	1	2	3	4
	69	17	7	13

- very low modularity.

Probabilistic approach

- Context: our matrix Y is the realization of a stochastic process.
- Aim: Propose a stochastic process is able to mimic heterogeneity in the connections.
- Advantage: benefit from the statistical tools (tests, model selection, etc...)

A first random graph model for network: null model

[Erdös and Rényi, 1959] Model for n nodes

$$\forall 1 \leq i, j \leq n, \quad Y_{ij} \overset{i.i.d.}{\sim} b(p),$$

where b is the Bernoulli distribution and $p \in [0, 1]$ a probability for a link to exist.

degree distribution
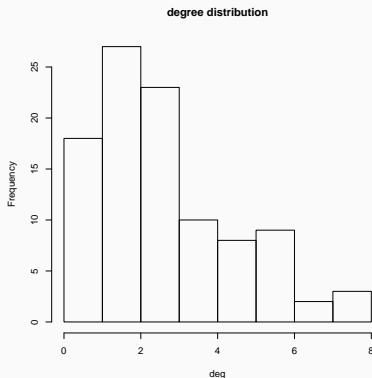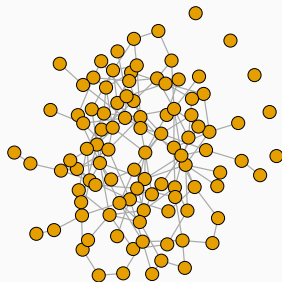
Limitations of an ER graph to describe real networks

- Homogeneity of the connections
- Degree distribution too concentrated, no high degree nodes,
- All nodes are equivalent (no nestedness...),
- No modularity, no hubs

Stochastic Block Model

[Nowicki and Snijders, 2001] Let (Y_{ij}) be an adjacency matrix

Latent variables

- The nodes $i = 1, \ldots, n$ are partitionned into K clusters
- $Z_i = k$ if node i belongs to cluster (block) k
- Z_i independant variables

$$\mathbb{P}(Z_i = k) = \pi_k$$

Conditionally to $(Z_i)_{i=1,\ldots,n}$**...**

(Y_{ij}) independant and

$$Y_{ij}|Z_i, Z_j \sim \mathcal{B}ern(\alpha_{Z_i, Z_j}) \quad \Leftrightarrow \quad P(Y_{ij} = 1|Z_i = k, Z_j = \ell) = \alpha_{k\ell}$$

Parameters

Let n nodes divided into 3 clusters

- $\mathcal{K} = \{\bullet, \bullet, \bullet\}$ clusters
- $\pi_\bullet = \mathbb{P}(i \in \bullet), \ \bullet \in \mathcal{K}, i = 1, \ldots, n$
- $\alpha_{\bullet\bullet} = \mathbb{P}(i \leftrightarrow j | i \in \bullet, j \in \bullet)$

$$Z_i = \mathbf{1}_{\{i \in \bullet\}} \ \sim^{\text{iid}} \mathcal{M}(1, \pi), \quad \forall \bullet \in \mathcal{K},$$
$$Y_{ij} \mid \{i \in \bullet, j \in \bullet\} \sim^{\text{ind}} \mathcal{B}(\alpha_{\bullet\bullet})$$
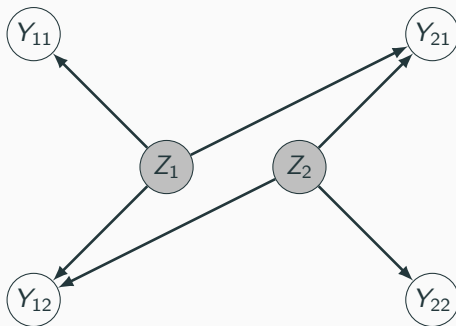
SBM : A great generative model

- Generative model : easy to simulate
- No a priori on the type of structure
- Combination of modularity, nestedness, etc...

Networks with hubs generated by SBM

Reordered adjacency matrix
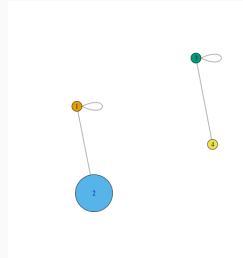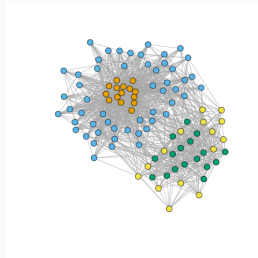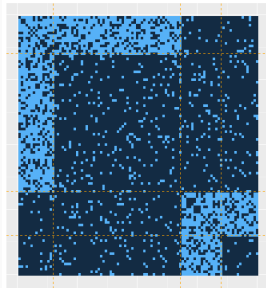
Histogram of degrees

Betweeness by block
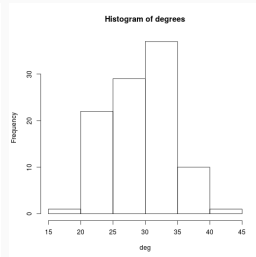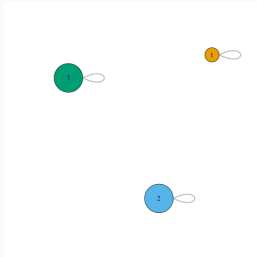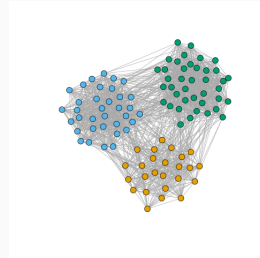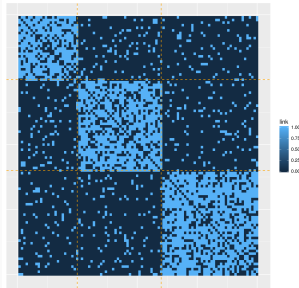
Nestedness generated by SBM
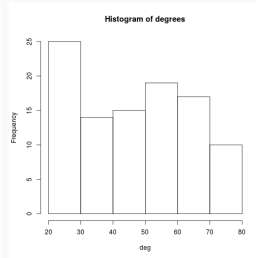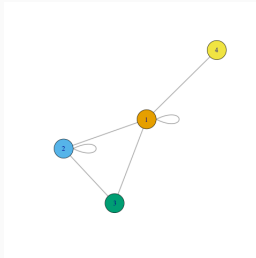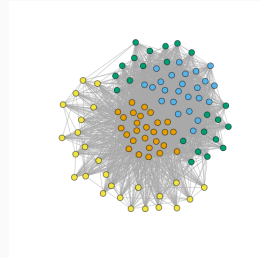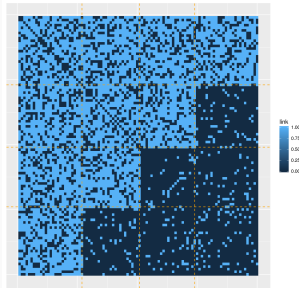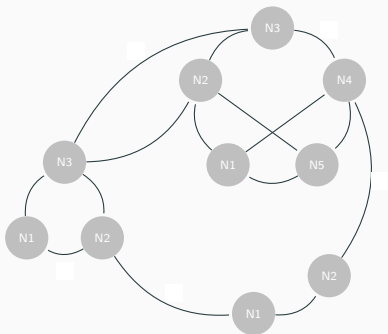
Stochastic Block Model

Let n nodes divided into

- $\mathcal{K} = \{\bullet, \bullet, \bullet\}$, card($\mathcal{K}$) known
- $\pi_{\bullet} =?,$
- $\alpha_{\bullet\bullet} =?$

[Nowicki and Snijders, 2001], [Daudin et al., 2008]

R package: blockmodels, sbm

From....
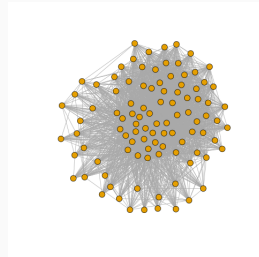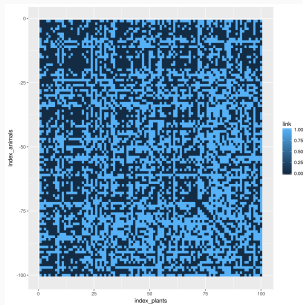
... to
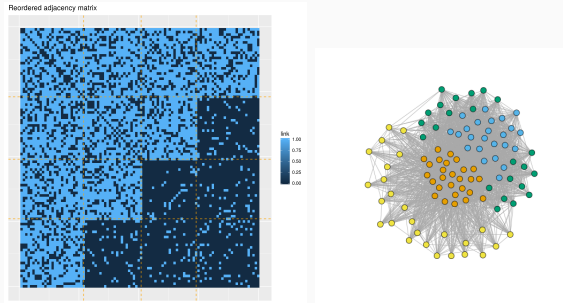
Tasks

- Find the clusters
- Find the number of clusters
- Practical implementation
- Theoretical results

Probabilistic model for binary bipartite networks

Let Y_{ij} be a bi-partite network. Individuals in row and cols are not the same.

Latent variables : bi-clustering

- Nodes $i = 1, \ldots, n_1$ partitionned into K_1 clusters, nodes $j = 1, \ldots, n_2$ partitionned into K_2 clusters

- $$Z_i^1 = k \quad \text{if node } i \text{ belongs to cluster (block) } k$$
  $$Z_j^2 = \ell \quad \text{if node } j \text{ belongs to cluster (block) } \ell$$

- Z_i^1, Z_j^2 independent variables

$$\mathbb{P}(Z_i^1 = k) = \pi_k^1, \quad \mathbb{P}(Z_j^2 = \ell) = \pi_\ell^2$$

Probabilistic model for binary bipartite networks

Conditionally to $(Z_i^1)_{i=1,\ldots,n_1}, (Z_j^2)_{j=1,\ldots,n_2}$ ...

(Y_{ij}) independent and

$$Y_{ij}|Z_i^1, Z_j^2 \sim \mathcal{B}ern(\alpha_{Z_i^1, Z_j^2}) \quad \Leftrightarrow \quad \mathbb{P}(Y_{ij} = 1|Z_i^1 = k, Z_j^2 = \ell) = \alpha_{k\ell}$$

[Govaert and Nadif, 2008]

Latent Block Model : illustration
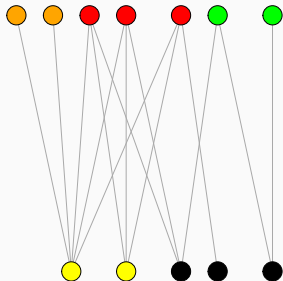
Latent Block Model

- n_1 row nodes $\mathcal{K}_1 = \{\bullet, \bullet, \bullet\}$ classes
- $\pi_\bullet^1 = \mathbb{P}(i \in \bullet), \bullet \in \mathcal{K}_1, i = 1, \ldots, n$
- n_2 column nodes $\mathcal{K}_2 = \{\bullet, \bullet\}$ classes
- $\pi_\bullet^2 = \mathbb{P}(j \in \bullet), \bullet \in \mathcal{K}_2, j = 1, \ldots, m$
- $\alpha_{\bullet\bullet} = \mathbb{P}(i \leftrightarrow j | i \in \bullet, j \in \bullet)$

$$Z_i^1 = \mathbf{1}_{\{i \in \bullet\}} \sim^{\text{iid}} \mathcal{M}(1, \pi^1), \quad \forall \bullet \in \mathcal{Q}_1,$$
$$Z_j^2 = \mathbf{1}_{\{j \in \bullet\}} \sim^{\text{iid}} \mathcal{M}(1, \pi^2), \quad \forall \bullet \in \mathcal{Q}_2,$$
$$Y_{ij} \mid \{i \in \bullet, j \in \bullet\} \sim^{\text{ind}} \mathcal{B}ern(\alpha_{\bullet\bullet})$$

Valued-edge networks

Values-edges networks

Information on edges can be something different from presence/absence. It can be:

1. a count of the number of observed interactions,
2. a quantity interpreted as the interaction strength,

Natural extensions of SBM and LBM

1. Poisson distribution: $Y_{ij} \mid \{i \in \bullet, j \in \bullet\} \sim^{\text{ind}} \mathcal{P}(\lambda_{\bullet\bullet})$,
2. Gaussian distribution: $Y_{ij} \mid \{i \in \bullet, j \in \bullet\} \sim^{\text{ind}} \mathcal{N}(\mu_{\bullet\bullet}, \sigma^2)$, [Mariadassou et al., 2010]
3. More generally,

$$Y_{ij} \mid \{i \in \bullet, j \in \bullet\} \sim^{\text{ind}} \mathcal{F}(\theta_{\bullet\bullet})$$

Multiplex networks

Several kind of interactions between nodes . For instance :

- Love and friendship
- Working relations and friendship
- In ecology : mutualistic and competition

Block model for multiplex networks

$Y_{ij} \in \{0,1\}^Q = (Y_{ij}^a, Y_{ij}^b)$, $\forall w \in \{0,1\}^2$

$$\mathbb{P}(Y_{ij}^a, Y_{ij}^b = w | Z_i = k, Z_j = \ell) = \alpha_{k\ell}^w$$

[Kéfi et al., 2016], [Barbillon et al., 2017]

In R package: blockmodels when two relations are at stake.

Remark: a particular case of multiplex network is dynamic network, [Matias and Miele, 2017].

Taking into account covariates

Sometimes covariates are available. They may be on:

- nodes,
- edges,
- both.

1. They can be used a posteriori to explain blocks inferred by SBM.
2. Extension of the SBM which takes into account covariates. Blocks are structure of interaction which is not explained by covariates !

If covariates are sampling conditions, case 2 be may more interesting.

SBM with covariates

- As before : (Y_{ij}) be an adjacency matrix
- Let $x^{ij} \in \mathbb{R}^p$ denote covariates describing the pair (i, j)

Latent variables : as before

- The nodes $i = 1, \ldots, n$ are partitioned into K clusters
- Z_i independent variables

$$\mathbb{P}(Z_i = k) = \pi_k$$

Conditionally to $(Z_i)_{i=1,\ldots,n}$...

(Y_{ij}) independent and

$$Y_{ij}|Z_i, Z_j \sim \mathcal{B}ern(\text{logit}(\alpha_{Z_i, Z_j} + \beta \cdot x_{ij})) \quad \text{if binary data}$$
$$Y_{ij}|Z_i, Z_j \sim \mathcal{P}(\exp(\alpha_{Z_i, Z_j} + \beta \cdot x_{ij})) \quad \text{if counting data}$$

If $K = 1$: all the connection heterogeneity is explained by the covariates.

Statistical Inference

- Selection of the number of clusters K for SBM or K_1, K_2 for LBM
- Estimation of the parameters π, θ for a given number of clusters
- Clustering $\hat{\mathbf{Z}}$

Complete likelihood ($\mathbf{Y}$) et ($\mathbf{Z}$)

$$
\begin{aligned}
\ell_c(\mathbf{Y}, \mathbf{Z}; \theta) &= p(\mathbf{Y}|\mathbf{Z}; \alpha)p(\mathbf{Z}; \pi) \\
&= \prod_{i,j} f_{\alpha_{Z_i, Z_j}}(Y_{ij}) \times \prod_i \pi_{Z_i} \\
&= \prod_{i,j} \alpha_{Z_i, Z_j}^{Y_{ij}} (1 - \alpha_{Z_i, Z_j})^{1 - Y_{ij}} \prod_i \pi_{Z_i}
\end{aligned}
$$

Marginal likelihood ($\mathbf{Y}$)

$$
\log \ell(\mathbf{Y}; \theta) = \log \sum_{\mathbf{Z} \in \mathcal{Z}} \ell_c(\mathbf{Y}, \mathbf{Z}; \theta). \tag{1}
$$

$$\log \ell(\mathbf{Y}; \theta) = \log \sum_{\mathbf{Z} \in \mathcal{Z}} \ell_c(\mathbf{Y}, \mathbf{Z}; \theta).$$

Remark

$\mathcal{Z} = \{1, \dots, K\}^n \Rightarrow$ when K and n increase, impossible to compute.

Standard tool to maximize the likelihood when latent variables involved : EM algorithm.

Standard EM

At iteration (t) :

- **Step E**: compute

$$Q(\theta|\theta^{(t-1)}) = \mathbb{E}_{\mathbf{Z}|\mathbf{Y},\theta^{(t-1)}}[\log \ell_c(\mathbf{Y}, \mathbf{Z}; \theta)]$$

- **Step M**:

$$\theta^{(t)} = \arg\max_\theta Q(\theta|\theta^{(t-1)})$$

Limitations of standard EM
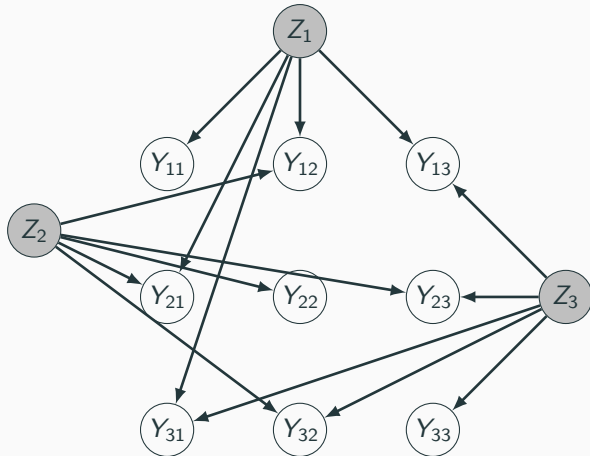
- Step E requires the computation of $\mathbb{E}_{\mathbf{Z}|\mathbf{Y},\theta^{(t-1)}}[\log \ell_c(\mathbf{Y}, \mathbf{Z}; \theta)]$
- However, once conditioned by par $\mathbf{Y}$, the $\mathbf{Z}$ are not independent anymore: complex distribution if K and n big.

Variational EM : maximization of a lower bound

Idea : replace the complicated distribution $p(\cdot|\mathbf{Y}; \theta) = [\mathbf{Z}|\mathbf{Y}, \theta]$ by a simpler one.

Let $\mathcal{R}_{\mathbf{Y}, \tau}$ be any distribution on $\mathbf{Z}$

Central identity

$$
\begin{aligned}
\mathcal{I}_\theta(\mathcal{R}_{\mathbf{Y}, \tau}) &= \log \ell(\mathbf{Y}; \theta) - \mathbf{KL}[\mathcal{R}_{\mathbf{Y}, \tau}, p(\cdot|\mathbf{Y}; \theta)] \quad \leq \log \ell(\mathbf{Y}; \theta) \\
&= \mathbb{E}_{\mathcal{R}_{\mathbf{Y}, \tau}} [\log \ell_c(\mathbf{Y}, \mathbf{Z}; \theta)] - \sum_{\mathbf{Z}} \mathcal{R}_{\mathbf{Y}, \tau}(\mathbf{Z}) \log \mathcal{R}_{\mathbf{Y}, \tau}(\mathbf{Z}) \\
&= \mathbb{E}_{\mathcal{R}_{\mathbf{Y}, \tau}} [\log \ell_c(\mathbf{Y}, \mathbf{Z}; \theta)] + \mathcal{H}(\mathcal{R}_{\mathbf{Y}, \tau}(\mathbf{Z}))
\end{aligned}
$$

Note that:

$$
\mathcal{I}_\theta(\mathcal{R}_{\mathbf{Y}, \tau}) = \log \ell(\mathbf{Y}; \theta) \Leftrightarrow \mathcal{R}_{\mathbf{Y}, \tau} = p(\cdot|\mathbf{Y}; \theta)
$$

Proof i

By Bayes

$$
\begin{aligned}
\log \ell_c(\mathbf{Y}, \mathbf{Z}; \theta) &= \log p(\mathbf{Z}|\mathbf{Y}; \theta) + \log \ell(\mathbf{Y}; \theta) \\
\log \ell(\mathbf{Y}; \theta) &= \log \ell_c(\mathbf{Y}, \mathbf{Z}; \theta) - \log p(\mathbf{Z}|\mathbf{Y}; \theta)
\end{aligned}
$$

By integration against $\mathcal{R}_{\mathbf{Y}, \tau}$:

$$
\begin{aligned}
\mathbb{E}_{\mathcal{R}_{\mathbf{Y}, \tau}}[\log \ell(\mathbf{Y}; \theta)] &= \mathbb{E}_{\mathcal{R}_{\mathbf{Y}, \tau}}[\log \ell_c(\mathbf{Y}, \mathbf{Z}; \theta)] - \mathbb{E}_{\mathcal{R}_{\mathbf{Y}, \tau}}[\log p(\mathbf{Z}|\mathbf{Y}; \theta)] \\
\log \ell(\mathbf{Y}; \theta) &= \mathbb{E}_{\mathcal{R}_{\mathbf{Y}, \tau}}[\log \ell_c(\mathbf{Y}, \mathbf{Z}; \theta)] - \mathbb{E}_{\mathcal{R}_{\mathbf{Y}, \tau}}[\log p(\cdot|\mathbf{Y}; \theta)]
\end{aligned}
$$

Proof ii

As a consequence:

$$
\begin{aligned}
\mathcal{I}_\theta(\mathcal{R}_{\mathbf{Y},\tau}) &= \log \ell(\mathbf{Y};\theta) - \mathbf{KL}[\mathcal{R}_{\mathbf{Y},\tau}, p(\cdot|\mathbf{Y};\theta)] \\
&= \mathbb{E}_{\mathcal{R}_{\mathbf{Y},\tau}}[\log \ell_c(\mathbf{Y},\mathbf{Z};\theta)] - \mathbb{E}_{\mathcal{R}_{\mathbf{Y},\tau}}[\log p(\mathbf{Z}|\mathbf{Y};\theta)] \\
&\quad - \mathbb{E}_{\mathcal{R}_{\mathbf{Y},\tau}}\left[\log \frac{\mathcal{R}_{\mathbf{Y},\tau}(\mathbf{Z})}{p(\mathbf{Z}|\mathbf{Y};\theta)}\right] \\
&= \mathbb{E}_{\mathcal{R}_{\mathbf{Y},\tau}}[\log \ell_c(\mathbf{Y},\mathbf{Z};\theta)] - \mathbb{E}_{\mathcal{R}_{\mathbf{Y},\tau}}[\log p(\mathbf{Z}|\mathbf{Y};\theta)] \\
&\quad - \underbrace{\mathbb{E}_{\mathcal{R}_{\mathbf{Y},\tau}}[\log \mathcal{R}_{\mathbf{Y},\tau}(\mathbf{Z})]}_{\mathcal{H}(\mathcal{R}_{\mathbf{Y},\tau}(\mathbf{Z}))} + \mathbb{E}_{\mathcal{R}_{\mathbf{Y},\tau}}[\log p(\mathbf{Z}|\mathbf{Y};\theta)]
\end{aligned}
$$

Variational EM

- Maximization of $\log \ell(\mathbf{Y}; \theta)$ w.r.t. θ replaced by maximization of the lower bound $\mathcal{I}_\theta(\mathcal{R}_{\mathbf{Y},\tau})$ w.r.t. τ and θ.
- **Benefit** : we choose $\mathcal{R}_{\mathbf{Y},\tau}$ such that the maximization calculus can be done explicitly
  - In our case: mean field approximation : neglect dependencies between the (Z_i)

$$P_{\mathcal{R}_{\mathbf{Y},\tau}}(Z_i = k) = \tau_{ik}$$

Variational EM

Algorithm

At iteration (t), given the current value $(\theta^{(t-1)}, \mathcal{R}_{\mathbf{Y}, \tau^{(t-1)}})$,

- **Step VE** Maximization w.r.t. τ

$$
\begin{aligned}
\tau^{(t)} &= \arg \max_{\tau \in \mathcal{T}} \mathcal{I}_{\theta^{(t-1)}} (\mathcal{R}_{\mathbf{Y}, \tau}) \\
&= \arg \max_{\tau \in \mathcal{T}} \mathbb{E}_{\mathcal{R}_{\mathbf{Y}, \tau}} \left[\log \ell_c(\mathbf{Y}, \mathbf{Z}; \theta^{(t-1)}) \right] + \mathcal{H}(\mathcal{R}_{\mathbf{Y}, \tau}(\mathbf{Z})) \\
&= \arg \max_{\tau \in \mathcal{T}} \log \ell(\mathbf{Y}; \theta^{(t-1)}) - \mathbf{KL}[\mathcal{R}_{\mathbf{Y}, \tau}, p(\cdot|\mathbf{Y}; \theta^{(t-1)})] \\
&= \arg \min_{\tau \in \mathcal{T}} \mathbf{KL}[\mathcal{R}_{\mathbf{Y}, \tau}, p(\cdot|\mathbf{Y}; \theta^{(t-1)})]
\end{aligned}
$$

Variational EM

Algorithm

- **Step M** Maximization w.r.t. θ

$$
\begin{aligned}
\theta^{(t)} &= \arg\max_{\theta} \mathcal{I}_{\theta}(\mathcal{R}_{\mathbf{Y},\tau^{(t)}}) \\
&= \arg\max_{\theta} \mathbb{E}_{\mathcal{R}_{\mathbf{Y},\tau^{(t)}}}\left[\log \ell_c(\mathbf{Y},\mathbf{Z};\theta)\right] + \mathcal{H}\left(\mathcal{R}_{\mathbf{Y},\tau^{(t)}}(\mathbf{Z})\right) \\
&= \arg\max_{\theta} \mathbb{E}_{\mathcal{R}_{\mathbf{Y},\tau^{(t)}}}\left[\log \ell_c(\mathbf{Y},\mathbf{Z};\theta)\right]
\end{aligned}
$$

VE-step for SBM i

$$\tau^{(t)} = \arg\min_\tau \mathbf{KL}[\mathcal{R}_{\mathbf{Y},\tau}, p(\cdot|\mathbf{Y}; \theta^{(t-1)})] = \arg\max_\tau \mathcal{I}_{\theta^{(t-1)}}(\mathcal{R}_{\mathbf{Y},\tau}).$$

(we drop out the index $^{(t-1)}$ on θ)

$$\mathcal{I}_\theta(\mathcal{R}_{\mathbf{Y},\tau}) \ = \ \sum_{\mathbf{Z}} \mathcal{R}_{\mathbf{Y},\tau}(\mathbf{Z}) \log \ell_c(\mathbf{Y}, \mathbf{Z}; \theta) - \sum_{\mathbf{Z}} \mathcal{R}_{\mathbf{Y},\tau}(\mathbf{Z}) \log \mathcal{R}_{\mathbf{Y},\tau}(\mathbf{Z}),$$

VE-step for SBM ii

with

$$
\begin{aligned}
\log \ell_c(\mathbf{Y}, \mathbf{Z}; \theta) &= \log p(\mathbf{Y}|\mathbf{Z}; \theta) + \log p(\mathbf{Z}; \theta), \\
&= \sum_{i,j=1, i \neq j}^{n} \log p(Y_{ij}|Z_i, Z_j; \theta) + \sum_{i=1}^{n} \log \pi_{Z_i}. \\
&= \sum_{i,j=1, i \neq j}^{n} \sum_{k,\ell=1}^{K} Z_{ik} Z_{j\ell} \log p(Y_{ij}|\alpha_{k\ell}) + \sum_{i=1}^{n} \sum_{k=1}^{K} Z_{ik} \log \pi_k
\end{aligned}
$$

Integration of the $\mathbf{Z}$ where $\mathbf{Z} \sim \mathcal{R}_{\mathbf{Y}, \tau}$

$$
\mathcal{I}_\theta(\mathcal{R}_{\mathbf{Y}, \tau}) = \sum_{i,j=1, i \neq j}^{n} \sum_{k,\ell=1}^{K} \tau_{iq} \tau_{j\ell} \log p(Y_{ij}|\alpha_{k\ell}) + \sum_{i=1}^{n} \sum_{k=1}^{K} \tau_{ik} \log \pi_k
$$

61

Maximization under the constraint: $\forall i = 1 \ldots n$, $\sum_{k=1}^{Q} \tau_{ik} = 1$.

- Derivatives of

$$\mathcal{I}_\theta(\mathcal{R}_{\mathbf{Y},\tau}) + \sum_{i=1}^{n} \lambda_i \left[\sum_{k=1}^{K} \tau_{ik} - 1 \right]$$

with respect to $(\lambda_i)_{i=1\ldots n}$ and $(\tau_{ik})_{i=1\ldots n, k=1\ldots K}$ where λ_i are the Lagrange multipliers,

- Leads to collection of equations: for $i = 1 \ldots n$ and $q = 1 \ldots Q$,

$$\sum_{\ell=1}^{K} \sum_{j=1, j\neq i}^{n} \log p(Y_{ij}|\alpha_{k\ell})\tau_{j\ell} + \log \pi_k - \log \tau_{ik} + 1 + \lambda_i = 0 \,,$$

VE-step for SBM iv

- Leads to the following fixed point problem:

$$\widehat{\tau}_{ik} = e^{1+\lambda_i} \alpha_k \prod_{j=1, j \neq i}^{n} \prod_{\ell=1}^{K} p(Y_{ij}|\alpha_{k\ell})^{\widehat{\tau}_{j\ell}}, \quad \forall i = 1 \ldots n, \forall k = 1 \ldots K,$$

which has to be solved under the constraints $\forall i = 1 \ldots n$, $\sum_{k=1}^{K} \tau_{ik} = 1$. This optimization problem is solved using a standard fixed point algorithm.

M-step for SBM i

$$\theta^{(t)} = \arg \max_{\theta} \mathcal{I}_{\theta^{(t)}}(\mathcal{R}_{\mathbf{Y}, \tau^{(t)}})$$

under the constraints: $\sum_{k=1}^{k} \pi_k = 1$.

Maximization with respect to $\boldsymbol{\pi}$ is quite direct:

$$\widehat{\pi}_q = \frac{1}{n} \sum_{i=1}^{n} \widehat{\tau}_{ik}$$

For the Bernoulli SBM:

$$\widehat{\alpha}_{k\ell} = \frac{\sum_{i,j=1,i\neq j}^{n} \widehat{\tau}_{ik} \widehat{\tau}_{j\ell} Y_{ij}}{\sum_{i,j=1,i\neq j}^{n} \widehat{\tau}_{ik} \widehat{\tau}_{j\ell}}$$

M-step for SBM ii

If the edge probabilities depend on covariates:

$$\text{logit}(p_{k\ell}) = \alpha_{k\ell} + \beta \cdot x_{ij},$$

then the optimization of $(\alpha_{k\ell})$ and (β) at step M of the VEM is not explicit anymore and one should resort to optimization algorithms such as Newton-Raphson algorithm.

In practice

- Really fast
- Strongly depend on the initial values

Penalized likelihood criterion

- Selection of the number of clusters K (or K_1, K_2 in the LBM)
- Integrated Classification Likelihood (ICL) [Biernacki et al., 2000]

$$ICL(\mathcal{M}_{\mathbf{K}}) = \log \ell_c(\mathbf{Y}, \hat{\mathbf{Z}}; \hat{\theta}_{\mathbf{K}}) - \mathrm{pen}(\mathcal{M}_{\mathbf{K}}) \qquad (2)$$

where

$$\hat{Z}_i = \underset{k \in \{1,\dots,K\}}{\arg\max} \ \hat{\tau}_{ik}. \qquad (3)$$

- Integrated Complete Likelihood (ICL)

$$ICL(\mathcal{M}_{\mathbf{K}}) = \mathbb{E}_{p(\cdot|\mathbf{Y}, \hat{\theta}_{\mathbf{K}})}[\log \ell_c(\mathbf{Y}, \hat{\mathbf{Z}}; \hat{\theta}_{\mathbf{K}}) - \mathrm{pen}(\mathcal{M}_{\mathbf{K}}) \qquad (4)$$

Expression of the penalization

For SBM

$$
pen_{\mathcal{M}} = \left\{
\begin{array}{ll}
-\frac{1}{2}\left\{(K-1)\log(n) + K^2\log\left(n^2-n\right)\right\} & \text{for directed network} \\[2ex]
-\frac{1}{2}\left\{\underbrace{(K-1)\log(n)}_{\text{Clust.}} + \frac{K(K+1)}{2}\log\left(\frac{n^2-n}{2}\right)\right\} & \text{for undirected network}
\end{array}
\right.
$$

For LBM

$$
pen_{\mathcal{M}} = -\frac{1}{2}\left\{ \underbrace{(K_1-1)\log(n_1) + (K_2-1)\log(n_2)}_{\text{Bi-Clust.}} \right.
$$

$$
\left. + \underbrace{(K_1 K_2)\log(n_1 n_2)}_{\text{Connection}} \right\}
$$

Advantages of ICL

- its capacity to outline the clustering structure in networks
- Involves a trade-off between goodness of fit and model complexity
- ICL values : goodness of fit AND clustering sharpness.

Comments on the ICL versus BIC

> **Conjecture**
>
> $$BIC(\mathcal{M}) = \log \ell(\mathbf{Y}; \hat{\theta}, \mathcal{M}) - \mathsf{pen}(\mathcal{M})$$
>
> with the same penalty

- Under this conjecture

$$
\begin{aligned}
ICL(\mathcal{M}) &= BIC(\mathcal{M}) + \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{Y}; \hat{\theta}_{\mathbf{K}}) \log p(\mathbf{Z}|\mathbf{Y}; \hat{\theta}_{\mathbf{K}}) \\
&= BIC(\mathcal{M}) - \mathcal{H}(p(\cdot|\mathbf{Y}; \theta))
\end{aligned}
$$

- As a consequence, because of the entropy, ICL will encourage clustering with well-separated groups

-

$$\widehat{ICL}(\mathcal{M}) = BIC(\mathcal{M}) + \sum_{\mathbf{Z}} \mathcal{R}_{\mathbf{Y}}(\mathbf{Z}, \hat{\tau}) \log \mathcal{R}_{\mathbf{Y}, \hat{\tau}}(\mathbf{Z}) - \mathbf{KL}[\mathcal{R}_{\mathbf{Y}, \hat{\tau}}, p(\cdot|\mathbf{Y}; \hat{\theta})].$$

Algorithm in practice

- Going trough the models and initiate VEM at the same time
- Bounds on K : $\{K_{min}, \ldots, K_{max}\}$

Stepwise procedure

Starting from K

- **Split** : if $K < K_{max}$
    - Maximize the likelihood (lower bound) of $\mathcal{M}_{K+1}$
    - K initializations of the VEM are proposed : split each cluster into 2 clusters
- **Merge** : If $K > K_{min}$
    - Maximize the likelihood (lower bound) of model $\mathcal{M}_{K-1}$
    - $\frac{K(K-1)}{2}$ initializations of the VEM are proposed : merging all the possible pairs of clusters

Theoretical properties for SBM

- Identifiability and a first consistency result by [Celisse et al., 2012]
- Consistency of the posterior distribution of the latent variables [Mariadassou and Matias, 2015]
- Consistency and properties of the variational estimators [Bickel et al., 2013]

Other extensions

- Time evolving networks Matias
- Multipartite, Multiplexe networks (R-package sbm, Bar-Hen, Barbillon, Donnet)
- Multilevel networks (individuals and organizations) (Chabbert-Liddell)
- Missing data in the network,

Probabilistic model for networks in a nutshell

SBM/LBM

- generative models,
- flexible,
- comprehensive models which can be linked to a lot of classical descriptors.

Barbillon, P., Donnet, S., Lazega, E., and Bar-Hen, A. (2017).

Stochastic block models for multiplex networks: an application to a multilevel network of researchers.

Journal of the Royal Statistical Society: Series A (Statistics in Society), 180(1):295–314.

Bickel, P., Choi, D., Chang, X., Zhang, H., et al. (2013).

Asymptotic normality of maximum likelihood and its variational approximation for stochastic blockmodels.

The Annals of Statistics, 41(4):1922–1943.

Biernacki, C., Celeux, G., and Govaert, G. (2000).

Assessing a mixture model for clustering with the integrated completed likelihood.

IEEE transactions on pattern analysis and machine intelligence, 22(7):719–725.

Celisse, A., Daudin, J.-J., and Pierre, L. (2012).

Consistency of maximum-likelihood and variational estimators in the stochastic block model.

Electronic Journal of Statistics, 6:1847–1899.

Daudin, J.-J., Picard, F., and Robin, S. (2008).

A mixture model for random graphs.

Statistics and computing, 18(2):173–183.

Erdös, P. and Rényi, A. (1959).

On random graphs i.

Publicationes Mathematicae Debrecen, 6:290.

References ii

Govaert, G. and Nadif, M. (2008).

Block clustering with bernoulli mixture models: Comparison of different approaches.

Computational. Statistics and Data Analysis, 52(6):3233–3245.

Kéfi, S., Miele, V., Wieters, E. A., Navarrete, S. A., and Berlow, E. L. (2016).

How structured is the entangled bank? the surprisingly simple organization of multiplex ecological networks leads to increased persistence and resilience.

PLOS Biology, 14(8):1–21.

Mariadassou, M. and Matias, C. (2015).

Convergence of the groups posterior distribution in latent or stochastic block models.

Bernoulli, 21(1):537–573.

Mariadassou, M., Robin, S., and Vacher, C. (2010).

Uncovering latent structure in valued graphs: a variational approach.

The Annals of Applied Statistics, 4(2):715–742.

Matias, C. and Miele, V. (2017).

Statistical clustering of temporal networks through a dynamic stochastic block model.

Journal of the Royal Statistical Society: Series B (Statistical Methodology), 79(4):1119–1141.

Nowicki, K. and Snijders, T. A. B. (2001).

Estimation and prediction for stochastic blockstructures.

Journal of the American Statistical Association, 96(455):1077–1087.