

# Mixture tree model for network inference

## JOBIM

Raphaëlle Momal

UMR518 AgroParis Tech/INRA

June 21, 2018

## 1 Motivation

## 2 Network inference

- General Framework
- Using trees

## 3 With count data

- Model
- Illustration

# Context

Rising interest in **jointly analysed** species abundances:

- Metagenomics
- Microbiologie
- Ecology

## Ecological network

Tool to better understand species interactions (direct/indirect, nature),  
eco-systems organizations (clusters ?)

Allows for resilience analyses, pathogens control, ecosystem comparison,  
reaction prediction...

# Example

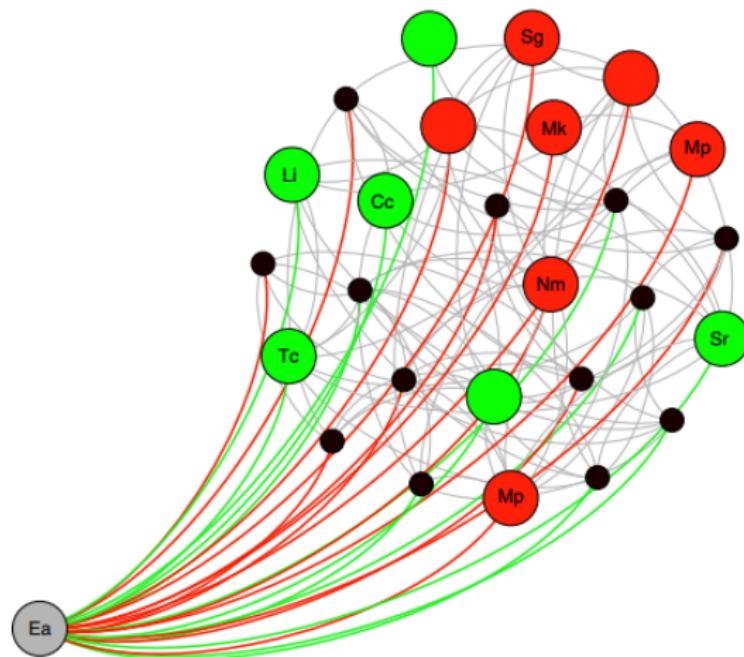


Figure: *Erysiphe alphitoides* pathobiom on oak tree leaves.

# Data

- **Species** : animal, bacteria, gene ...
- **Abundances** : ecological counts, Next-Generation Sequencing technologies...
- **Covariates** : coverage, temperature, water depth ...

Repeated signal :  $n$  samples of  $p$  abundances.

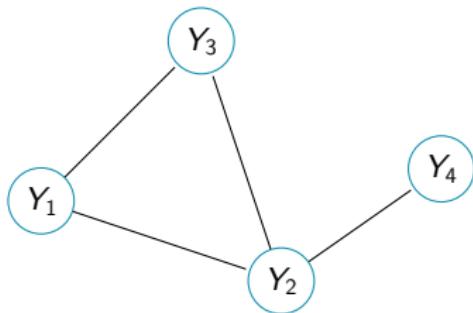
## Notations

$$Y = [Y_{ij}]_{(i,j) \in \{1, \dots, n\} \times \{1, \dots, p\}}$$

- $Y_{ij}$  : abundance of the  $j^{th}$  species in the  $i^{th}$  sample

Infer the species interaction network from  $Y$

# Graphical models



- All variables are dependant
- Some are **conditionnally independant**  
(i.e. indirectly dependant)

$$Y_4 \perp\!\!\!\perp (Y_1, Y_3) | Y_2$$

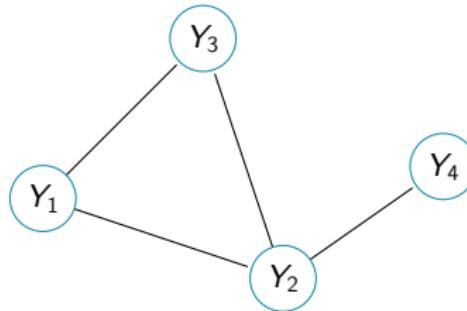
# Graphical models

## Factorization property

The joint distribution  $P$  is faithful to the graph  $G$  iff

$$P(Y_1, \dots, Y_p) \propto \prod_{C \in \mathcal{C}_G} \psi_C(Y_C)$$

where  $\mathcal{C}_G = \text{set of cliques of } G$ .



$$P(Y_1, Y_2, Y_3, Y_4) \propto \\ \psi_1(Y_1, Y_2, Y_3) \times \psi_2(Y_3, Y_4)$$

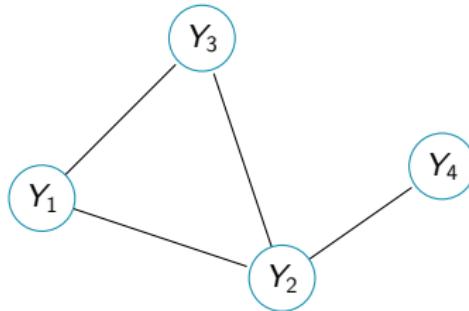
# Gaussian Graphical Models (GGM)

Gaussian distribution.

$$Y_r \sim \mathcal{N}_p(\mu, \Sigma)$$

$\mu$  = vector of means,  $\Sigma$  = covariance matrix.

A nice property.



# Gaussian Graphical Models (GGM)

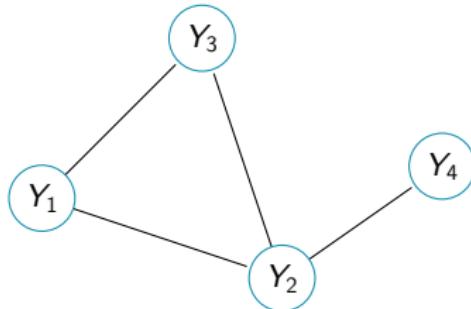
Gaussian distribution.

$$Y_r \sim \mathcal{N}_p(\mu, \Sigma)$$

$\mu$  = vector of means,  $\Sigma$  = covariance matrix.

A nice property.

Adjacency matrix



$$A = \begin{bmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix}$$

# Gaussian Graphical Models (GGM)

Gaussian distribution.

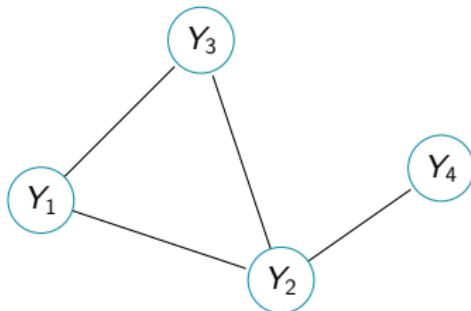
$$Y_r \sim \mathcal{N}_p(\mu, \Sigma)$$

$\mu$  = vector of means,  $\Sigma$  = covariance matrix.

A nice property.

Covariance matrix

$$\Sigma \propto \begin{bmatrix} 1 & -.25 & -.41 & .25 \\ -.25 & 1 & -.41 & .25 \\ -.41 & -.41 & 1 & -.61 \\ .25 & .25 & -.61 & 1 \end{bmatrix}$$



# Gaussian Graphical Models (GGM)

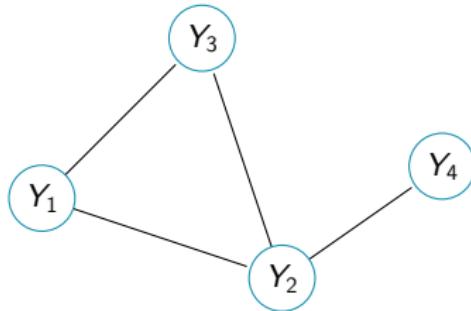
Gaussian distribution.

$$Y_r \sim \mathcal{N}_p(\mu, \Sigma)$$

$\mu$  = vector of means,  $\Sigma$  = covariance matrix.

A nice property.

Inverse covariance matrix



$$\Sigma^{-1} \propto \begin{bmatrix} 1 & .5 & .5 & 0 \\ .5 & 1 & .5 & 0 \\ .5 & .5 & 1 & .5 \\ 0 & 0 & .5 & 1 \end{bmatrix}$$

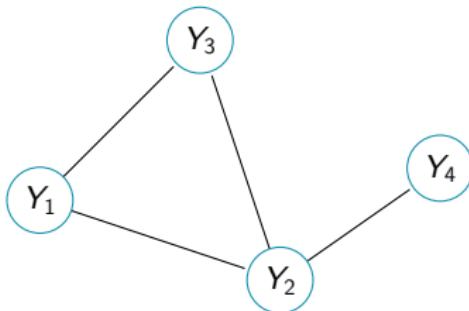
# Gaussian Graphical Models (GGM)

Gaussian distribution.

$$Y_r \sim \mathcal{N}_p(\mu, \Sigma)$$

$\mu$  = vector of means,  $\Sigma$  = covariance matrix.

A nice property.



Estimated inverse covariance matrix

$$\hat{\Sigma}^{-1} \propto \begin{bmatrix} 1 & .48 & .61 & \textcolor{red}{.09} \\ .48 & 1 & .67 & \textcolor{red}{.06} \\ .61 & .67 & 1 & .46 \\ \textcolor{red}{.09} & \textcolor{red}{.06} & .46 & 1 \end{bmatrix}$$

( $n = 100$ )

# Sparse with Glasso

Sparsity assumption  $\Rightarrow$  find a **sparse estimate** for  $\Omega = \Sigma^{-1}$

Negative log-likelihood :

$$L(Y, \Omega) \propto \frac{n}{2} \log(\det(\Omega)) - \frac{n}{2} Y^T \Omega Y$$

Graphical LASSO (glasso) :

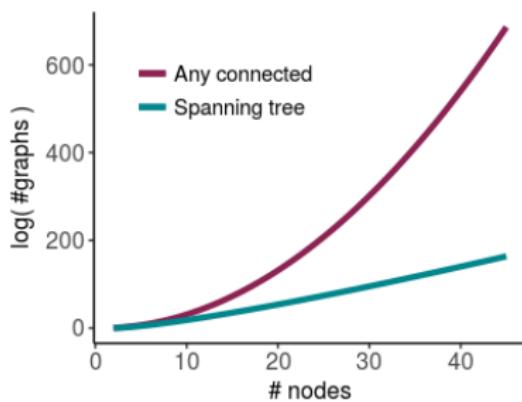
Glasso penalises the  $L_1$  norm of the concentration matrix:

$$\widehat{\Omega}_\lambda = \arg \min_{\Omega \in \mathcal{S}_d^+} \left\{ L(Y, \Omega) + \lambda \sum_{i \neq j} |w_{ij}| \right\}$$

# Spanning trees

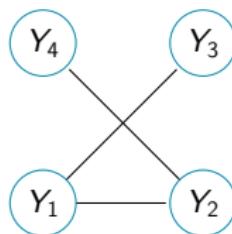
Spanning trees are another **sparse** solution :

$$\left. \begin{array}{l} G \text{ is connected} \\ G \text{ has no cycle} \end{array} \right\} G \text{ has } (p - 1) \text{ edges}$$

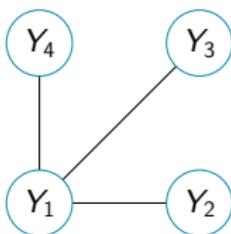


Much **smaller space** to explore:  
 $\#\mathcal{G} = 2^{\frac{p(p-1)}{2}}$  vs.  $\#\mathcal{T} = p^{(p-2)}$

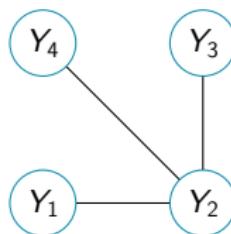
# Tree averaging



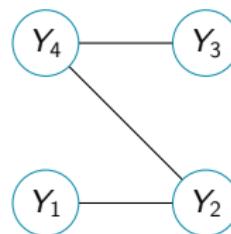
$$P\{T = T_1|Y\}$$



$$P\{T = T_2|Y\}$$



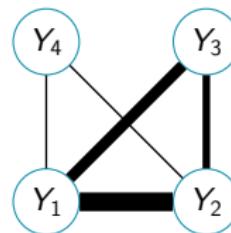
$$P\{T = T_3|Y\}$$



$$P\{T = T_4|Y\}$$

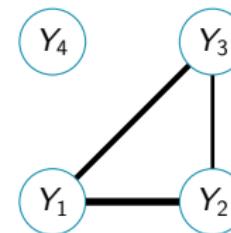
...

Compute edge probabilities:



$$P\{(j, k) \in T|Y\}$$

Thresholding probabilities:



$$P\{(j, k) \in T|Y\}$$

# PLN model

## Poisson log-Normal distribution

$$\left. \begin{array}{l} Z_i \text{ iid } \sim \mathcal{N}_d(\mu, \Sigma) \\ (Y_{ij})_j \perp\!\!\!\perp |Z_i \\ Y_{ij}|Z_{ij} \sim \mathcal{P}(e^{Z_{ij}}) \end{array} \right\} Y \sim \mathcal{PLN}(\mu, \Sigma)$$

- Gaussian latent layer
- Easy handling of multi-variate data (contrary to Negative binomial distribution)

# PLN model

## Poisson log-Normal distribution

$$\left. \begin{array}{ll} Z_i \text{ iid} & \sim \mathcal{N}_d(\mu, \Sigma) \\ (Y_{ij})_j \perp\!\!\!\perp |Z_i & \\ Y_{ij}|Z_{ij} & \sim \mathcal{P}(e^{x_i^T \Theta_j + Z_{ij}}) \end{array} \right\} Y \sim \mathcal{PLN}(x^T \Theta + \mu, \Sigma)$$

- Gaussian latent layer
- Easy handling of multi-variate data (contrary to Negative binomial distribution)
- Allow adjustment for covariates

# PLN model

## Poisson log-Normal distribution

$$\left. \begin{array}{l} Z_i \text{ iid } \sim \mathcal{N}_d(\mu, \Sigma) \\ (Y_{ij})_j \perp\!\!\!\perp | Z_i \\ Y_{ij}|Z_{ij} \sim \mathcal{P}(e^{x_i^T \Theta_j + Z_{ij}}) \end{array} \right\} Y \sim \mathcal{PLN}(x^T \Theta + \mu, \Sigma)$$

- Gaussian latent layer
- Easy handling of multi-variate data (contrary to Negative binomial distribution)
- Allow adjustment for covariates

**Idea:** Infer the **latent Gaussian network**.

# Oak Mildew

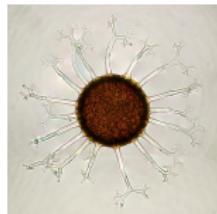


Figure: *Erysiphe alphitoides*.



Figure: Oak Leaf with powdery mildew.

Data : metabarcoding of microbiome on oak tree leaves

- 114 sample of 94 microbial species counts (bacteria/fungi)
- 8 covariates (leaf distance to the ground / tree / branch ...)