Ecological network reconstruction from count data

Christophe Ambroise, Raphaëlle Momal, Stéphane Robin

[1]UMR AgroParisTech / INRA MIA-Paris
[2]LaMME, Evry

March15th, 2019

Abundance Data

Jointly analysed species abundances in Ecology

- Metagenomics
- Microbiology
- Ecology

Understanding abundance data

- often analysed using species distribution model (**?**), where species are traditionally considered as disconnected entities.
- biotic interactions are relevant descriptors of an ecosystem (**??**).
- interactions can be conveniently represented by ecological networks

- This work focuses on ecological network reconstruction based on observed species abundance data.

Interaction and statistics

Direct interactions (Causality ?)

- a useful network shall include only direct interactions between species.
- Indirect statistical association may be observed between two species either because they are both affected by the same environmental variations
- any approach aiming at reconstructing ecological networks needs to account for covariates

Graphical models: a statistical framework for network inference

Generic statistical model.

$$(Y_i)_{i=1..n} \text{ iid } \sim P,$$

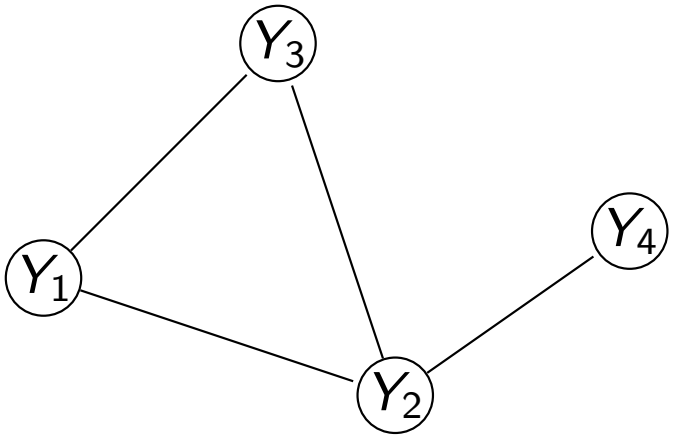
Faithfulness.

A distribution P is said to be Markov faithful to a graph G if, for any triple (A, B, S) of disjoint subsets of nodes, it holds that

$$(\text{S separates A from B in G}) \Leftrightarrow A \perp\!\!\!\perp B|S.$$

Graphical models: a statistical framework for network inference

Example:

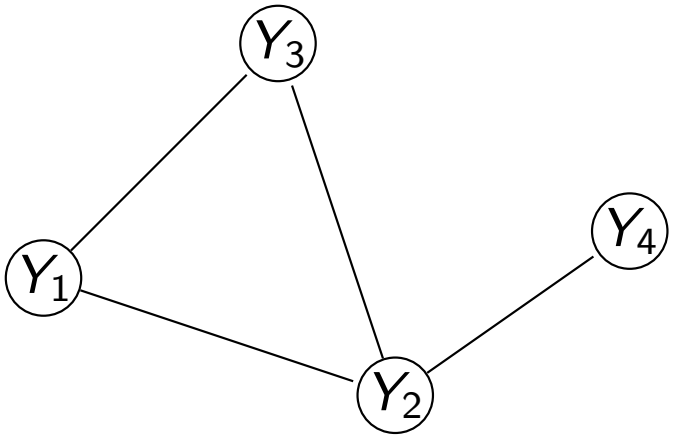


- Connected: all variables are dependant
- Some are conditionally independent (i.e. indirectly dependant)

  Y_4 is independent from (Y_1, Y_3) conditionally on Y_2

Graphical models: a statistical framework for network inference

Example:



- Connected: all variables are dependant
- Some are conditionally independent (i.e. indirectly dependant)

  Y_4 is independent from (Y_1, Y_3) conditionally on Y_2

$$P(Y_1, \ldots, Y_p) \propto \prod_{C \in \mathcal{C}_G} \psi_C(Y_C) \propto \psi_1(Y_1, Y_2, Y_3) \times \psi_2(Y_3, Y_4)$$

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

Network inference

Network inference problem.

Based on the dataset $Y = (Y_{ij})$, infer G.

Network inference

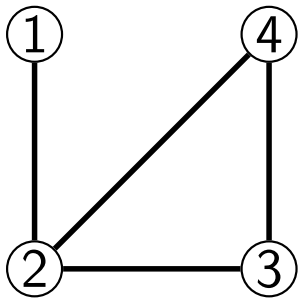
Network inference problem.

Based on the dataset $Y = (Y_{ij})$, infer G.

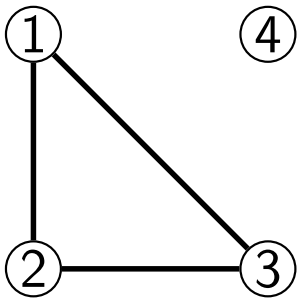
Critical issue. There are $2^{p(p-1)/2}$ possible graphs

p	5	10	20	50
# graphs	10^3	10^{14}	10^{57}	10^{369}

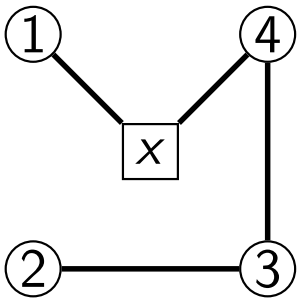
Practical influence of covariates



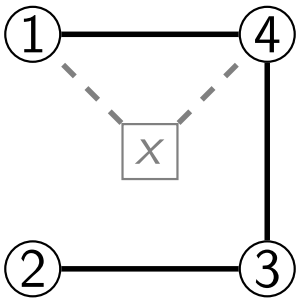
(*a*) connected



(*b*) disconnected



(*c*) with covariate



(*d*) missing covar

Examples of graphical models.

Our problem

Data:

- Species: bacteria, fungi...
- Abundances: read counts from Next-Generation Sequencing technologies (metabarcoding) $\Rightarrow$ $n \times p$ matrix Y
- Covariates: temperature, water depth... $\Rightarrow$ $n \times d$ matrix X
- Offsets: species-specific, sample-specific $\Rightarrow$ $n \times p$ matrix O

Goal:

Infer the species interaction network $\widehat{G}$ from count data Y, accounting for X and O :

$$\widehat{G} = f(Y, X, O)$$

Challenges

- Statistical network inference
- Count data
- Offsets and covariates

PLN model

Poisson log-Normal distribution (?)

$$\left.\begin{array}{ll} Z_i \ iid & \sim \mathcal{N}_d(0, \Sigma) \\ & (Y_{ij})_j \perp\!\!\!\perp | Z_i \\ Y_{ij}|Z_{ij} & \sim \mathcal{P}(e^{Z_{ij}}) \end{array}\right\} Y \sim \mathcal{PLN}(0, \Sigma)$$

- Dependency structure in the Gaussian latent layer
- Easy handling of multi-variate data

PLN model

Poisson log-Normal distribution (?)

$$\left.\begin{array}{ll} Z_i \ iid & \sim \mathcal{N}_d(0, \Sigma) \\ & (Y_{ij})_j \perp\!\!\!\perp | Z_i \\ Y_{ij}|Z_{ij} & \sim \mathcal{P}(e^{o_{ij} + x_i^\top \Theta_j + Z_{ij}}) \end{array}\right\} Y \sim \mathcal{PLN}(O + X^\top \Theta, \Sigma)$$

- Dependency structure in the Gaussian latent layer
- Easy handling of multi-variate data
- Allow adjustment for covariates and offsets
- Variational estimation algorithm (**?**)

PLN model + Graphical model

Poisson log-Normal distribution (?)

$$\left.\begin{array}{ll} Z_i \ iid & \sim \mathcal{N}_d(0, \Sigma_G) \\ & (Y_{ij})_j \perp\!\!\!\perp | Z_i \\ Y_{ij}|Z_{ij} & \sim \mathcal{P}(e^{o_{ij} + x_i^\top \Theta_j + Z_{ij}}) \end{array}\right\} Y \sim \mathcal{PLN}(O + X^\top \Theta, \Sigma_G)$$

- Dependency structure in the Gaussian latent layer
- Easy handling of multi-variate data
- Allow adjustment for covariates and offsets
- Variational estimation algorithm (**?**)

Sparsity and Gaussian dependency

General settings

let $\mathbf{K} = (K_{ij})_{(i,j)\in\mathcal{P}^2} := \mathbf{\Sigma}^{-1}$ be the concentration matrix.

let $\mathbf{K} = (K_{ij})_{(i,j)\in\mathcal{P}^2} := \mathbf{\Sigma}^{-1}$ be the concentration matrix.

The graphical interpretation

$$Z_i \perp\!\!\!\perp Z_j | Z_{\mathcal{P}\setminus\{i,j\}} \Leftrightarrow K_{ij} = 0 \Leftrightarrow \text{ edge } (i,j) \notin \text{ network},$$

since $r_{ij|\mathcal{P}\setminus\{i,j\}} = -K_{ij}/\sqrt{K_{ii}K_{jj}}$.

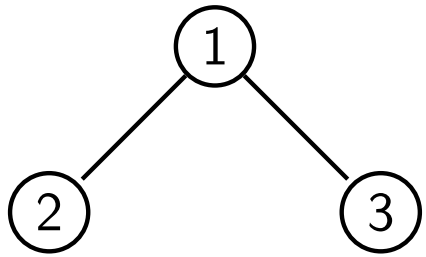
$\rightsquigarrow$ $\mathbf{K}$ describes the graph of conditional dependencies of the hidden structure.

Gaussian graphical models

Example

$$Z = \begin{pmatrix} Z_1 \\ Z_2 \\ Z_3 \end{pmatrix}, \qquad \Sigma = \begin{pmatrix} 2 & 1 & -1 \\ 1 & 1.5 & -0.5 \\ -1 & -0.5 & 1.5 \end{pmatrix}$$

$$K = \Sigma^{-1} = \begin{pmatrix} 1 & -0.5 & 0.5 \\ -0.5 & 1 & 0 \\ 0.5 & 0 & 1 \end{pmatrix}, \qquad \mathcal{G} =$$



- Underlying graph $\mathcal{G} = (V, E)$, $V = \{1, \ldots, p\}$
- The edge $\{i, j\}$ is in E if $K_{ij} \neq 0$

Inferring $\mathcal{G} \Leftrightarrow$ inferring the support of K.

Inference of K

Estimate K from data

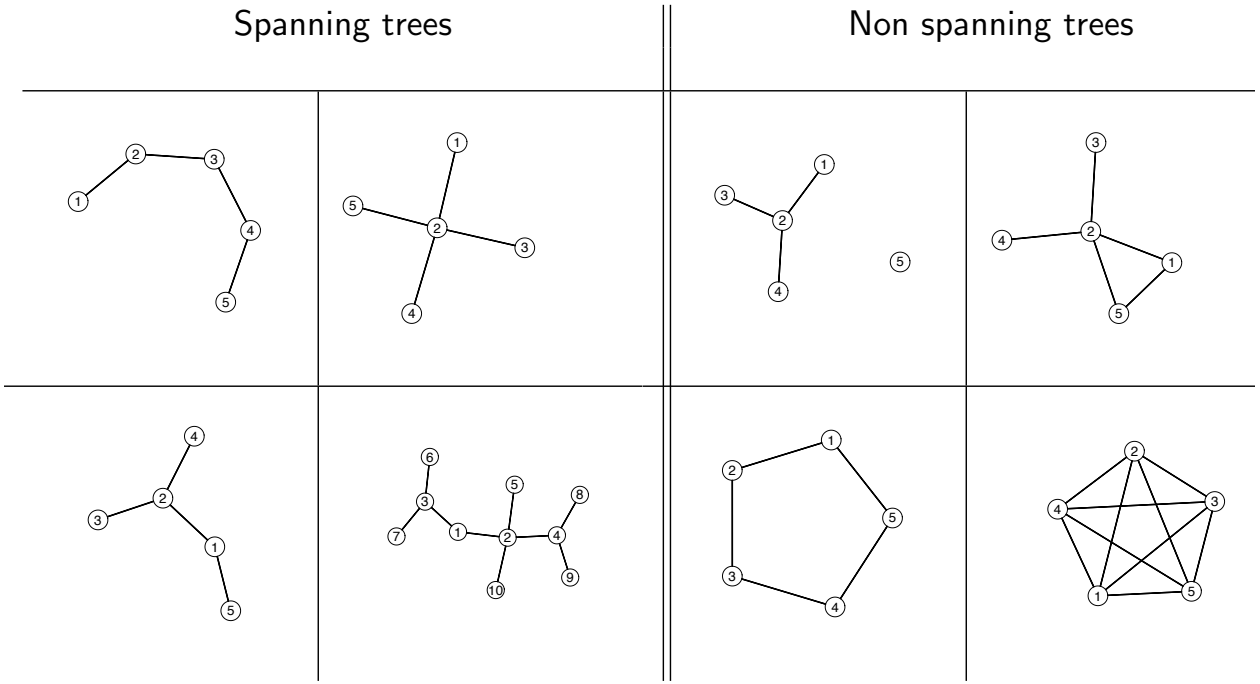
- Maximum likelihood estimator:

$$\begin{aligned} \hat{K}^{MLE} &= \arg\max_{K} \log\det(K) - \mathrm{tr}(K\Sigma_n) \\ &= \Sigma_n^{-1} \end{aligned} \tag{1}$$

Hypothesis on the structure of the support of K

- Penalized Log-likelihood
- Tree hypothesis

Tree-shaped network on PLN latent layer



PLN and spanning Tree

$$
\left.
\begin{aligned}
T &\sim \prod_{kl} \beta_{kl} / B \\
Z_i | T \ iid &\sim \mathcal{N}_d(0, \Sigma_T) \\
(Y_{ij})_j &\perp\!\!\!\perp | Z_i | T \\
Y_{ij} | Z_{ij}, T &\sim \mathcal{P}(e^{o_{ij} + x_i^\top \Theta_j + Z_{ij}})
\end{aligned}
\right\} Y \sim \mathcal{PLN}(O + X^\top \Theta, \Sigma_T)
$$

Tree-shaped network on PLN latent layer

PLN and spanning Tree

$$\left.\begin{aligned} T &\sim \prod_{kl} \beta_{kl}/B \\ Z_i|T \ iid &\sim \mathcal{N}_d(0, \Sigma_T) \\ &(Y_{ij})_j \perp\!\!\!\perp |Z_i|T \\ Y_{ij}|Z_{ij}, T &\sim \mathcal{P}(e^{o_{ij}+x_i^\top \Theta_j + Z_{ij}}) \end{aligned}\right\} Y \sim \mathcal{PLN}(O + X^\top \Theta, \Sigma_T)$$

Mixture of Trees

Z_i follows a mixture of centered Gaussian distributions with respective covariances matrices

$$Z_i \sim \sum_{T \in \mathcal{T}} P(T)\mathcal{N}(Z_i; 0, \Sigma_T)$$

- Both Z and T are latent variables

Why Spanning trees

Sparse structures:

$$\#\mathcal{G}_p = 2^{\frac{p(p-1)}{2}} \text{ reduced to } \#\mathcal{T}_p = p^{(p-2)}$$

p	5	10	20	50
# graphs	10^3	10^{14}	10^{57}	10^{369}
# spanning trees	10^2	10^8	10^{23}	10^{81}

Why Spanning trees

Sparse structures:

$\#\mathcal{G}_p = 2^{\frac{p(p-1)}{2}}$ reduced to $\#\mathcal{T}_p = p^{(p-2)}$

p	5	10	20	50
# graphs	10^3	10^{14}	10^{57}	10^{369}
# spanning trees	10^2	10^8	10^{23}	10^{81}

Suitable algebraic tool:

Matrix tree theorem (**?**)

$$\sum_{T \in \mathcal{T}} \prod_{(k,l) \in T} \psi_{k,l}(Y) = \det(L_{\psi(Y)}) \to \Theta(p^3)$$

Approach: infer the network by averaging spanning trees

Posterior probability for an edge

Prior on T: factorizes over the edges:

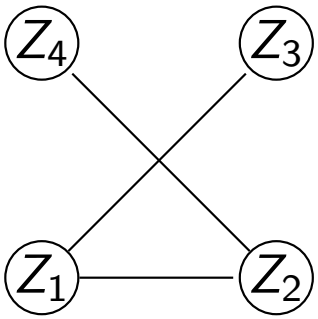
$$p(T) \propto \prod_{(j,k)\in T} \beta_{jk}$$

The existence of an edge between variables Z_j and Z_k can be assessed by

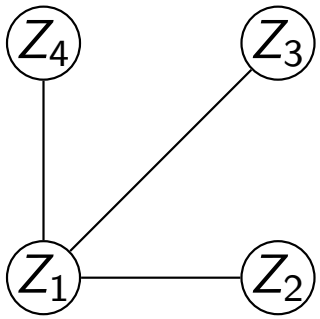
$$P((j,k) \in T \mid Z) \propto \sum_{T \ni (j,k)} p(T) p(Z \mid T)$$

which depends on the prior $p(T)$.

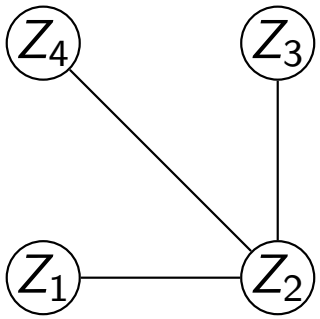
Tree averaging



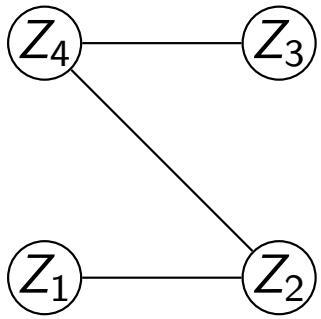
$P\{T = T_1|Z\}$



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



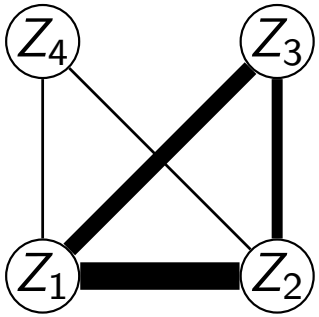
$P\{T = T_3|Z\}$



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

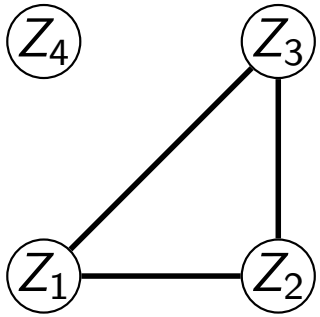
...

Compute edge probabilities:



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

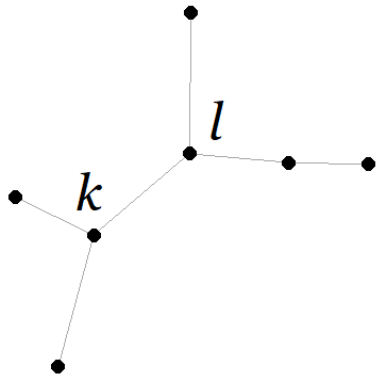
Thresholding probabilities:



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

Tree structured data

- Data dependency structure relies on a tree



- Likelihood factorizes on nodes and edges (**?**):

$$\mathbb{P}(Z|T) = \prod_{j=1}^{d} \mathbb{P}(Z_j) \prod_{k,l \in T} \psi_{kl}(Z) \quad ,$$

Where

$$\psi_{kl}(Z) = \frac{\mathbb{P}(Z_k, Z_l)}{\mathbb{P}(Z_k) \times \mathbb{P}(Z_l)}.$$

Rmq : with standardised gaussian data, $\hat{\Psi} = [\hat{\psi_{kl}}] \propto (1 - \hat{\rho_Z}^2)^{-1/2}$

Direct EM algorithm ?

- Complete likelihood :

$$\mathbb{P}(Y, Z, T) = \mathbb{P}(T) \times \mathbb{P}(Z|T) \times \mathbb{P}(Y|Z)$$

$$\log(\mathbb{P}(Y, Z, T)) = \sum_{k,l} \mathbb{1}_{\{(k,l)\in T\}} (\log(\beta_{kl}) + \log(\psi_{kl}(Z))) - \log(B)$$
$$+ \sum_{k} (\log(\mathbb{P}(Z_k)) + \log(\mathbb{P}(Y_k|Z_k)))$$

Direct EM algorithm ?

- Complete likelihood :

$$\mathbb{P}(Y, Z, T) = \mathbb{P}(T) \times \mathbb{P}(Z|T) \times \mathbb{P}(Y|Z)$$

$$\log(\mathbb{P}(Y, Z, T)) = \sum_{k,l} \mathbb{1}_{\{(k,l)\in T\}} (\log(\beta_{kl}) + \log(\psi_{kl}(Z))) - \log(B) + \sum_{k} (\log(\mathbb{P}(Z_k)) + \log(\mathbb{P}(Y_k|Z_k)))$$

- Conditional expectation :

$$\mathbb{E}_\theta[\log(\mathbb{P}(Y, Z, T))|Y] = \sum_{k,l\in V} \mathbb{P}((k,l) \in T|Y) \log(\beta_{kl}) + \mathbb{E}[\mathbb{1}_{\{(k,l)\in T\}} \log(\psi_{kl}(Z)|Y)] + \sum_{k} \mathbb{E}[\log(\mathbb{P}(Z_k))|Y] + \mathbb{E}[\log(\mathbb{P}(Y_k|Z_k))|Y] - \log(B)$$

Two steps solution

The `PLNmodels` package approximates the distribution parameters:

1. Approximate $\hat{\Sigma}_Z$
2. Apply EM mixture tree to $Z \sim \mathcal{N}(0, \hat{\Sigma}_Z)$

Simplified conditional expectation writing:

$$\mathbb{E}_\theta[\log(\mathbb{P}(Z, T))|Z] = \sum_{k,l} \mathbb{P}((k,l) \in T|Z) \times \log(\beta_{kl}\psi_{kl}) - \log(B) + \sum_k \log(\mathbb{P}(Z_k))$$

$$\Rightarrow \textbf{EM algorithm} \text{ (E: ?, M: ?)}$$

EMtree algorithm

Input: Abundance data, covariates, offsets

1rst step: VEM algorithm to fit PLN model $\Rightarrow \hat{\theta}$, $\hat{\Sigma}_Z$.

2nd step: EM algorithm to update the $\beta_{jk} \Rightarrow$ conditional probabilities for all edges.

EMtree algorithm

Input: Abundance data, covariates, offsets

1rst step: VEM algorithm to fit PLN model $\Rightarrow \hat{\theta}$, $\hat{\Sigma}_Z$.

2nd step: EM algorithm to update the $\beta_{jk} \Rightarrow$ conditional probabilities for all edges.

Thresholding: Select edges with probability above the probability of edges in a tree drawn uniformly ($2/p$)

Resampling: Strengthen the results: only edges selected in more than 80% of S sub-samples are kept.

Available for download at `https://github.com/Rmomal/EMtree`

Evaluation strategy

Alternatives:
Two methods on transformed counts, no covariates:

- SpiecEasi algorithm **?** : centered log-ratio (clr) transformation
- gCoda **?** : log transformation to the relative counts

Principles

Both methods assume that

- the transformed counts have a Gaussian distribution,
- that ecological networks are sparse.
- The network is then reconstructed using the graphical lasso (**?**).

One taking raw counts and covariates:

- MInt **?** (uses PLN model): same PLN model (without Trees) and Graphical Lasso

'Simulation design

1. Choose G and define Σ_G accordingly
2. Sample count data Y from $\mathcal{PLN}(X, \Sigma_G)$
3. Infer the network with EMtree, SpiecEasi, gCoda, and MInt
4. Compare results with presence/absence of edges (FDR, AUC)

Simulation settings

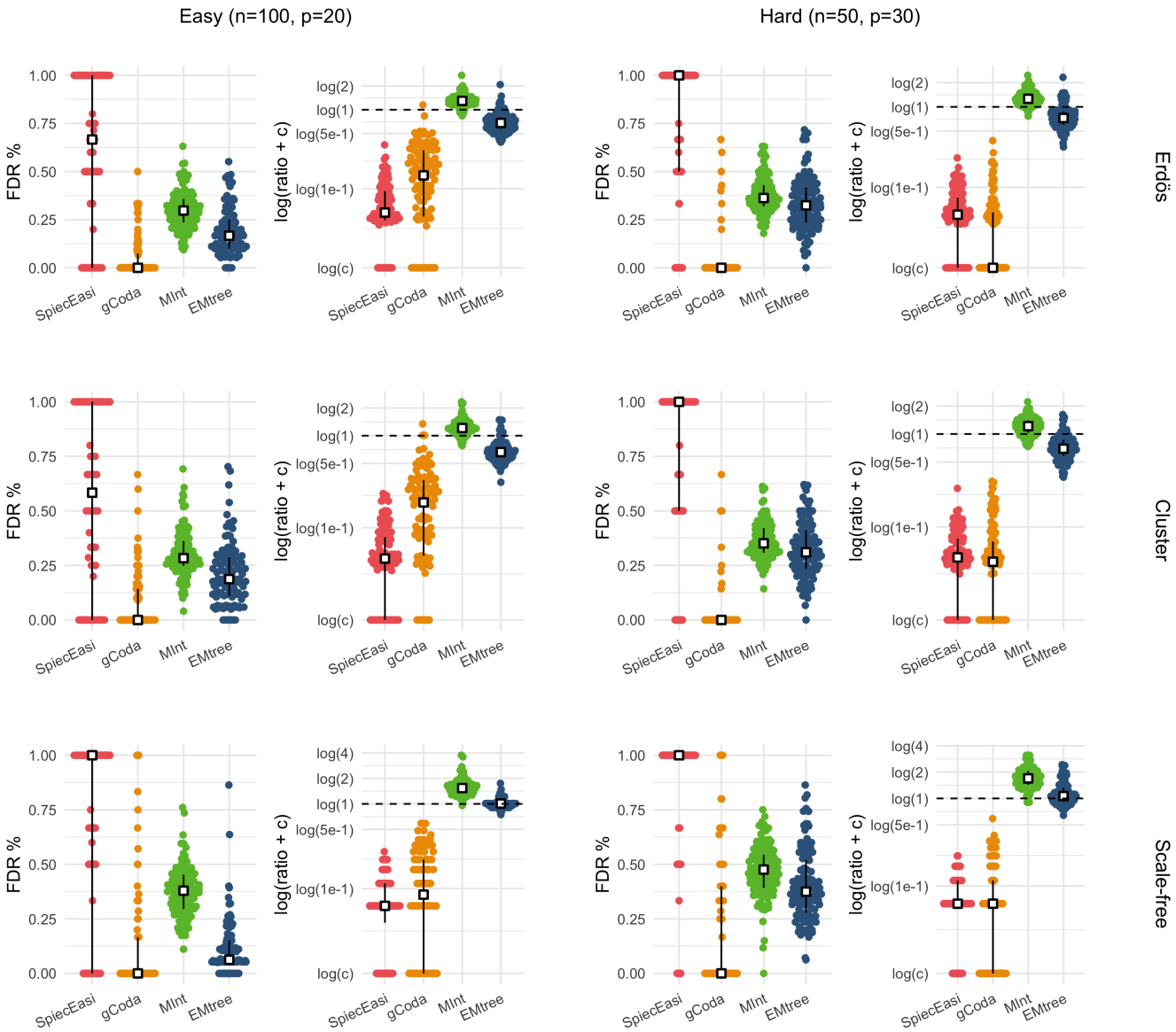
Difficulty

- easy setting ($n = 100$, $p = 20$)
- hard setting ($n = 50$, $p = 30$).

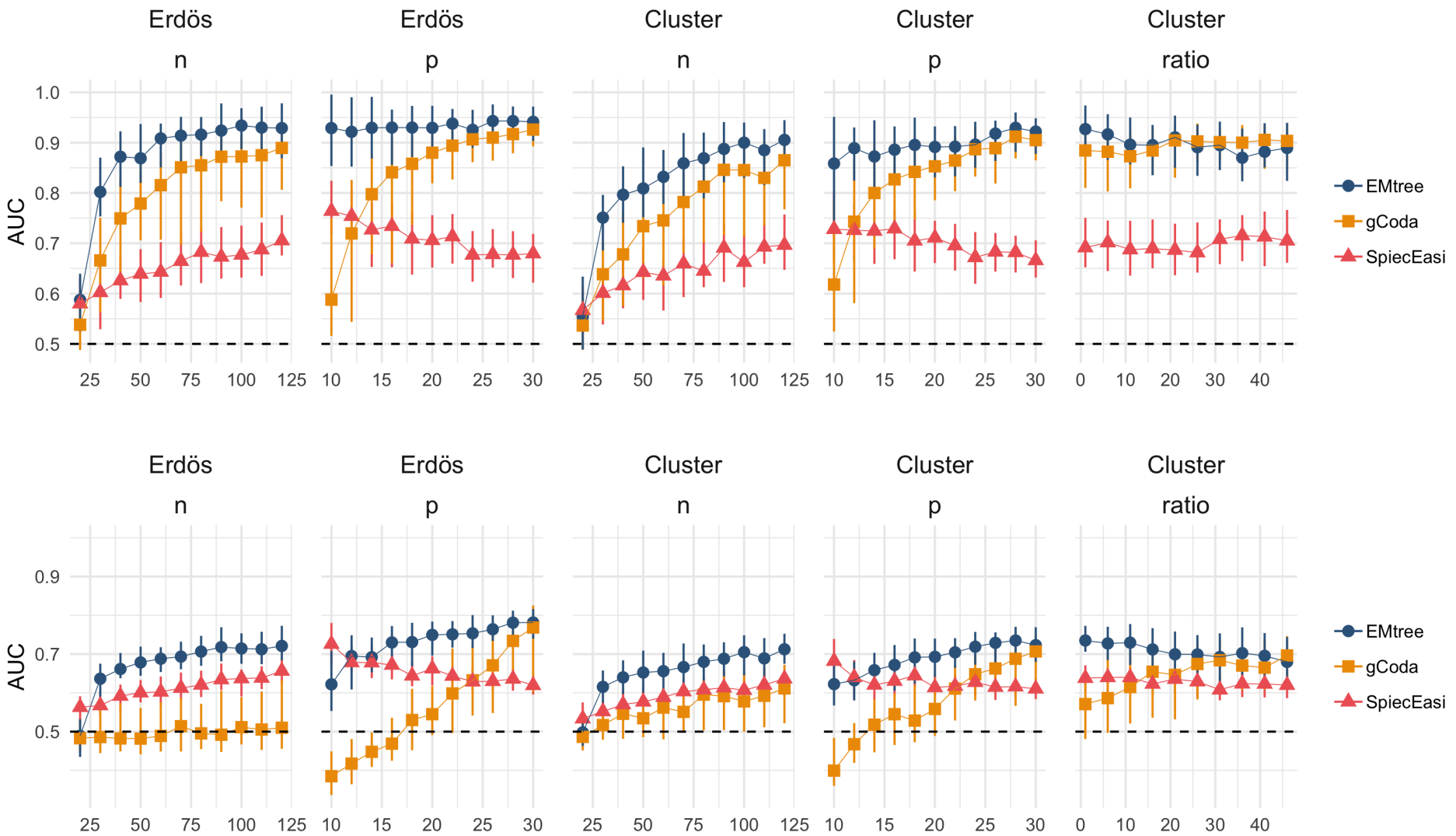
Network Structure

- Erdös
- Cluster
- Scale free

Difficulty level



Network density



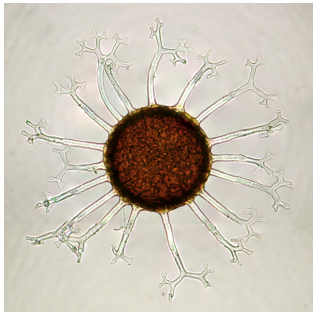
Effect of Erdös and Cluster structures on the evolutions of AUC median and inter-quartile intervals for parameters n, p and *ratio*. *Top*: densities set to $2/p$, *bottom*: densities set to $5/p$.

Running Times

	$n < 50$	$n \geq 50$	$p < 20$	$p \geq 20$
EMtree	0.41 (0.11)	0.6 (0.15)	0.38 (0.12)	0.71 (0.21)
gCoda	0.12 (0.47)	0.07 (0.03)	0.05 (0.03)	0.09 (0.06)
SpiecEasi	2.41 (0.25)	2.41 (0.25)	2.39 (0.25)	2.42 (0.25)

Median and standard-deviation of running times for each method in seconds, for n and p parameters. corresponding to Erdös and cluster structures with $5/p$ densities.

Oak Mildew



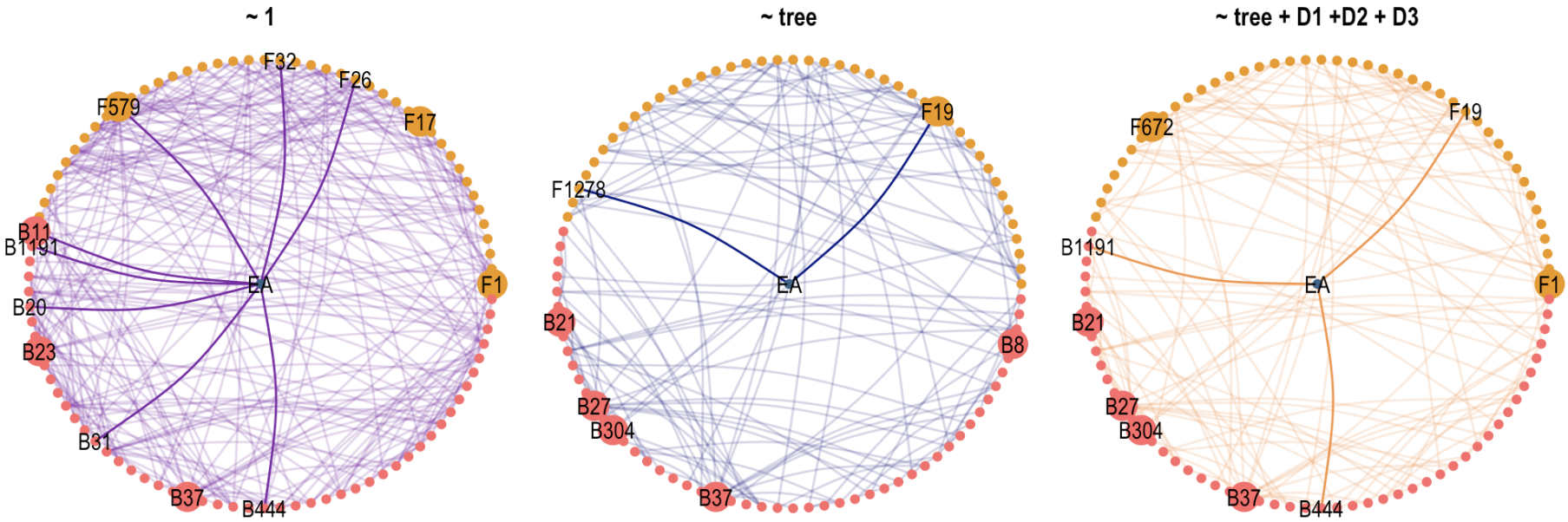
Pathogen Erysiphe alphitoides (EA).

Oak leaf with powdery mildew.

Metabarcoding of oak tree leaves microbiome (**?**).

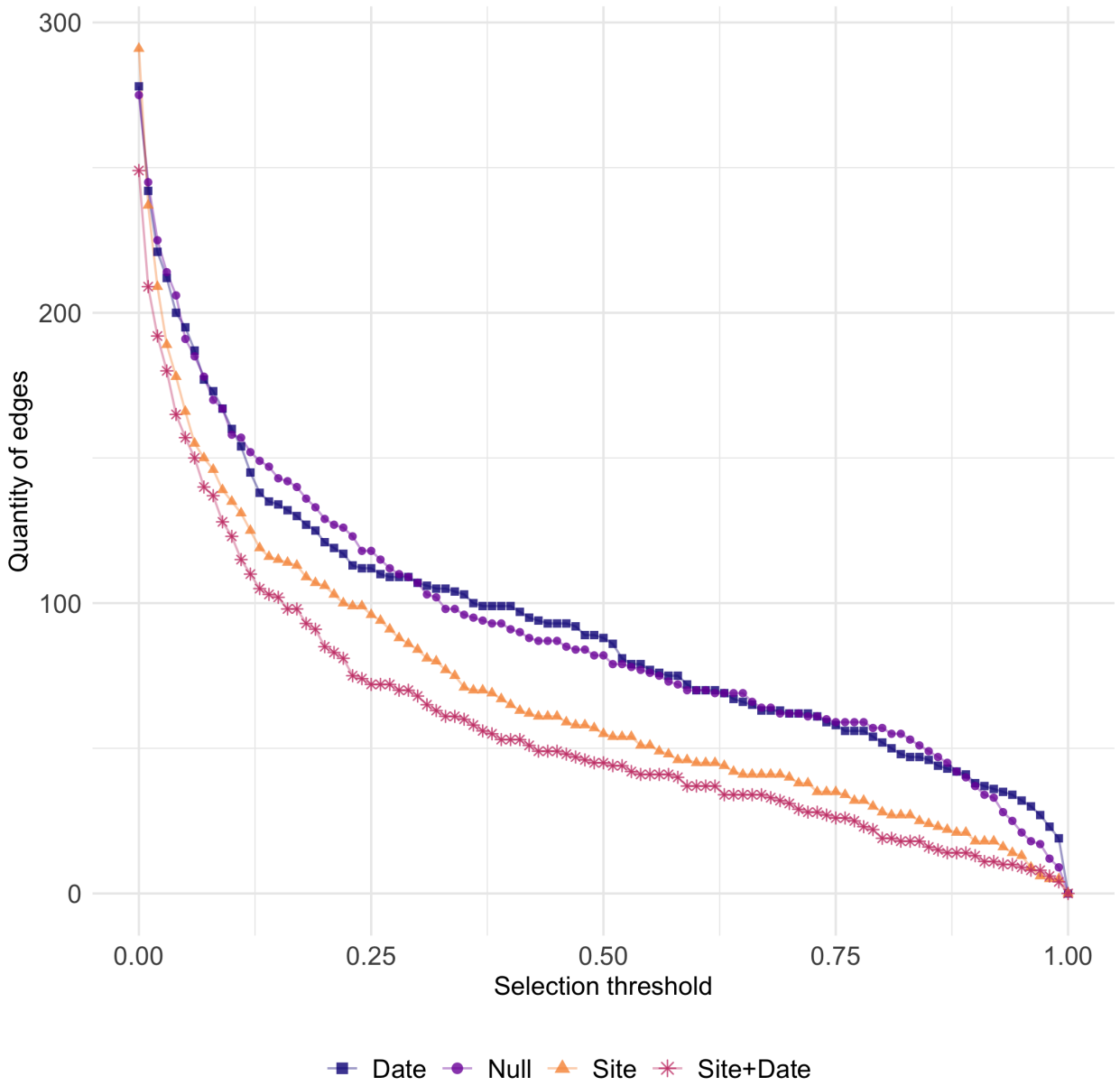
- 114 sample of 94 bacterial/fungal-OTUs (Operational Taxonomic Unit)
- Different read depth for bacteria and fungi
- covariates: tree status; distance to ground, to trunk and to base of the branch.

Inferred networks



- 10 neighbours for the pathogen Ea and 11 key-player OTUs.
- accounting for the covariates affects the network density, and the list of key-players.
- In addition, the connections of the pathogen Ea are greatly modified, highlighting changes in the microbiome of infected trees due to this agent.
- EMtree identifies three potential neighbors to the Ea and two key-players in these networks, which were not identified by **?**.

Selection Threshold



Conclusion

Contributions:

- Formal probabilistic model for network inference with count data
- Inclusion of offsets and covariates
- Variational estimation algorithm

Perspectives:

- Taking spatial position into account
- Improving precision of computation
- Network comparison
- Missing major actor (species/covariates)
- Model for the inference in the observed counts layer

Acknowledgments

Special thanks :

PLN team Julien Chiquet (MIA-Paris), Mahendra Mariadassou (INRA Jouy)

Data Corinne Vacher (INRA Bordeaux)

Conditional probability computation

Kirchhoff's theorem (matrix tree, ?)

For all $W = (a_{kl})_{k,l}$ a symmetric matrix, the corresponding Laplacian $Q(W)$ is defined as follows:

$$\mathcal{Q}_{uv}(W) = \begin{cases} -a_{uv} & 1 \leq u < v \leq n \\ \sum_{i=1}^{n} a_{vi} & 1 \leq u = v \leq n. \end{cases}$$

Then for all u et v:

$$|Q^*_{uv}(W)| = \sum_{T \in \mathcal{T}} \prod_{\{k,l\} \in E_T} a_{kl}$$

$$\begin{aligned} \mathbb{P}((k,l) \in T|Z) &= \sum_{T \in \mathcal{T}:(k,l) \in T} \mathbb{P}(T|Z) = \frac{\sum_{(k,l) \in T} \mathbb{P}(T)\mathbb{P}(Z|T)}{\sum_T \mathbb{P}(T)\mathbb{P}(Z|T)} \\ &= 1 - \frac{|Q^*_{uv}(\beta\Psi^{-kl})|}{|Q^*_{uv}(\beta\Psi)|} \\ &= \tau_{kl} \end{aligned}$$

M step

Goal : optimization of weights β_{kl}.

$$\underset{\beta_{kl}}{\operatorname{argmax}} \left\{ \sum_{k,l \in V} \tau_{kl}(\log(\beta_{kl}) + \log(\psi_{kl})) - \log(B) + \sum_{k} \log(\mathbb{P}(Z_k)) \right\}$$

With high combinatorial complexity of $B = \sum_{T \in \mathcal{T}} \prod_{k,l \in T} \beta_{kl}$

How to compute $\frac{\partial B}{\partial \beta_{kl}}$?

β_{kl} update

A result from Meilă ?

Inverting a minor of the laplacien Q, we define M :

$$\begin{cases} M_{uv} = [\mathcal{Q}^{*-1}]_{uu} + [\mathcal{Q}^{*-1}]_{vv} - 2[\mathcal{Q}^{*-1}]_{uv} & u, v < n \\ M_{nv} = M_{vn} = [\mathcal{Q}^{*-1}]_{vv} & v < n \\ M_{vv} = 0. \end{cases}$$

On peut montrer que :

$$\frac{\partial |Q^*_{uv}(W)|}{\partial \beta_{kl}} = M_{kl} \times |Q^*_{uv}(W)|$$

$$\frac{\partial \mathbb{E}_\theta[\log(\mathbb{P}(Z,T))|Z]}{\partial \beta_{kl}} = \frac{\tau_{kl}}{\beta_{kl}} - \frac{1}{B}\frac{\partial B}{\partial \beta_{kl}}$$

$$\hat{\beta}^{h+1}_{kl} = \frac{\tau^h_{kl}}{M^h_{kl}}$$

References I