Mixture tree model for network inference

Supervision: S. Robin[1] et C. Ambroise[12]

Raphaëlle Momal

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

July 7, 2018

Context

Rising interest in jointly analysed species abundances:

- Metagenomics
- Microbiology
- Ecology

Ecological network

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

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

Data example

- Species: bacteria, fungi...
- Abundances: read counts from Next-Generation Sequencing technologies (metabarcoding)
- Covariates: sequencing depth, temperature, water depth...

Repeated signal : n samples, p abundances.

Data table

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

- Y_{ij}: abundance of the jth species in the ith sample

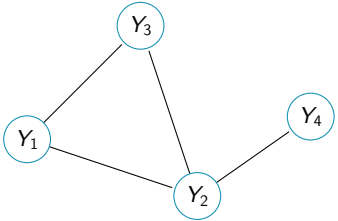
Infer the species interaction network from count data Y

Challenges

- Statistical network inference
- Count data
- Offsets and covariates

Graphical models: a statistical framework for network inference

Example:



- All variables are dependant
- Some are conditionally independent (i.e. indirectly dependeant)

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

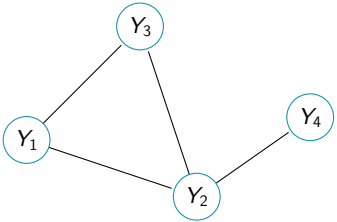
Graphical models

Definition [Lauritzen, 1996]

The joint distribution P is faithful to the graph G iff

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

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



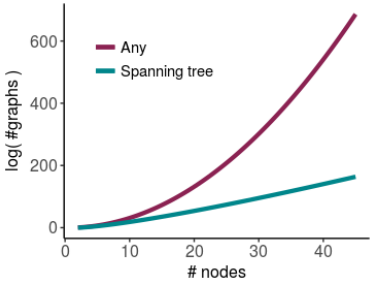
$$P(Y_1, Y_2, Y_3, Y_4) \propto \psi_1(Y_1, Y_2, Y_3) \times \psi_2(Y_3, Y_4)$$

Spanning trees

Unconstrained graph $\Rightarrow$ very large space to explore: $\#\mathcal{G}_p = 2^{\frac{p(p-1)}{2}}$

Spanning trees are a 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:

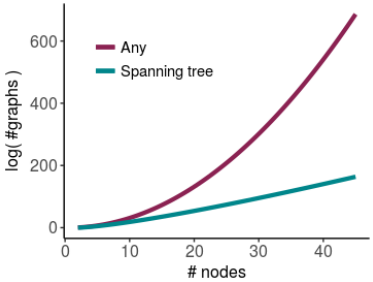
$$\#\mathcal{T}_p = p^{(p-2)}$$

Spanning trees

Unconstrained graph $\Rightarrow$ very large space to explore: $\#\mathcal{G}_p = 2^{\frac{p(p-1)}{2}}$

Spanning trees are a 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:

$$\#\mathcal{T}_p = p^{(p-2)}$$

Still a huge complexity...

Maximizing and summing over spanning trees

Maximum spanning tree Kruskall's algorithm

$$\hat{T} = \underset{T}{\operatorname{argmax}} \left\{ \prod_{(k,l)\in T} \psi_{k,l}(Y) \right\} \to \Theta(p^2)$$

Tree averaging Matrix tree theorem [Chaiken and Kleitman, 1978]

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

Maximizing and summing over spanning trees

Maximum spanning tree Kruskall's algorithm

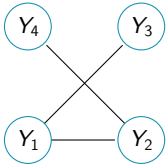
$$\hat{T} = \underset{T}{\operatorname{argmax}} \left\{ \prod_{(k,l)\in T} \psi_{k,l}(Y) \right\} \to \Theta(p^2)$$

Tree averaging Matrix tree theorem [Chaiken and Kleitman, 1978]

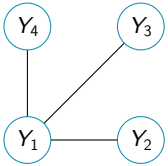
$$\sum_{T} \prod_{(k,l)\in T} \psi_{k,l}(Y) = \det(L(Y)) \to \Theta(p^3)$$

Approach: infer the network by averaging spanning trees

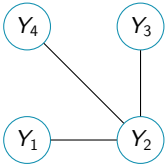
Tree averaging



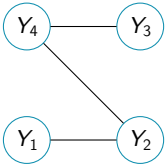
$P\{T = T_1|Y\}$



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



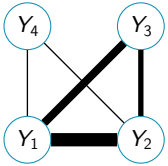
$P\{T = T_3|Y\}$



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

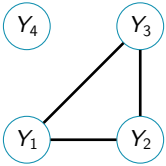
...

Compute edge probabilities:



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

Thresholding probabilities:



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

PLN model

Poisson log-Normal distribution [Aitchison and Ho, 1989]

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

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

PLN model

Poisson log-Normal distribution [Aitchison and Ho, 1989]

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

- Dependency structure in the Gaussian latent layer
- Easy handling of multi-variate data (contrary to Negative binomial distribution)
- Allow adjustment for covariates and offsets

PLN model

Poisson log-Normal distribution [Aitchison and Ho, 1989]

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

- Dependency structure in the Gaussian latent layer
- Easy handling of multi-variate data (contrary to Negative binomial distribution)
- Allow adjustment for covariates and offsets
- Variational estimation algorithm [Chiquet et al., 2017]

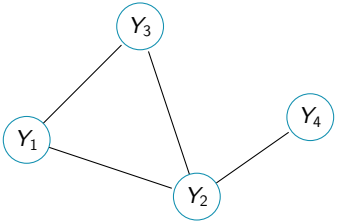
Approach: Infer the latent Gaussian network with an VEM algorithm.

Gaussian Graphical Models (GGM)

Gaussian distribution:

$$Y_i \sim \mathcal{N}_p(\mu, \Sigma),\ \mu = \text{vector of means},\ \Sigma = \text{covariance matrix}.$$

A nice property:



Inverse covariance matrix

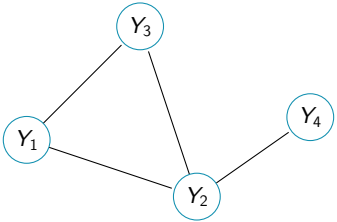
$$\Sigma^{-1} = \Omega \propto \begin{bmatrix} 1 & .5 & .5 & 0 \\ .5 & 1 & .5 & .5 \\ .5 & .5 & 1 & 0 \\ 0 & .5 & 0 & 1 \end{bmatrix}$$

Gaussian Graphical Models (GGM)

Gaussian distribution:

$$Y_i \sim \mathcal{N}_p(\mu, \Sigma),\ \mu = \text{vector of means},\ \Sigma = \text{covariance matrix}.$$

A nice property:



Inverse covariance matrix

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

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

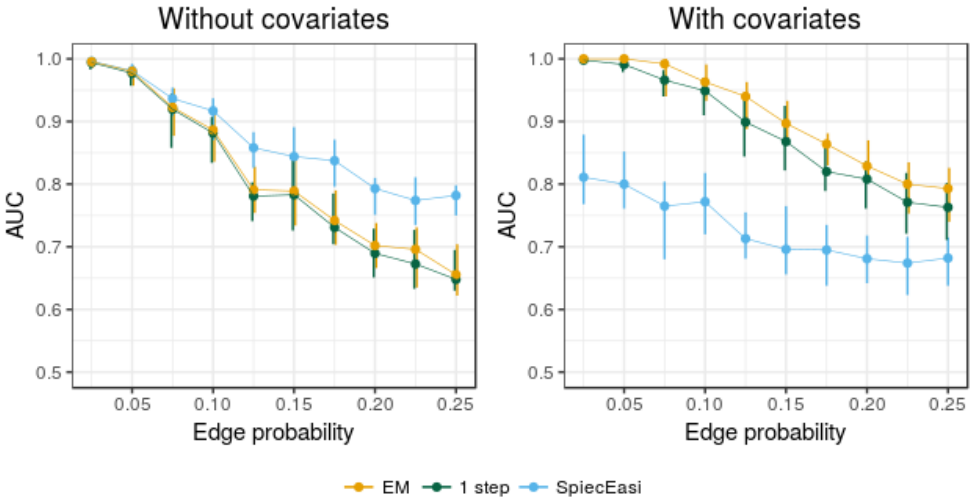
$\Rightarrow$ SpiecEasi method [Kurtz et al., 2015]: glasso on transformed counts

Simulation design

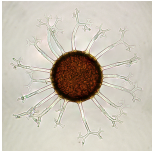
1. Choose G and define Ω accordingly
2. Sample count data Y from $\mathcal{PLN}(0, \Omega^{-1})$ with possible covariates
3. Infer the network with PLN + mixture tree VEM and SpiecEasi
4. Compare results with AUC (presence/absence of edges)

$\Rightarrow$ 40 replicates for each setting (p, n, edge probability)

Results: Erdös, 20 nodes



Oak Mildew



Pathogen Erysiphe alphitoides (EA).

Oak leaf with powdery mildew.

Metabarcoding of oak tree leaves microbiome [Jakuschkin et al., 2016].

- 114 sample of 94 microbial species counts (bacteria/fungi)
- Different read depth for bacteria and fungi: unsuited for normalization with SpiecEasi
- 3 quantitative covariates

We are interested in EA and F19, a second major fungi.

Model with covariates

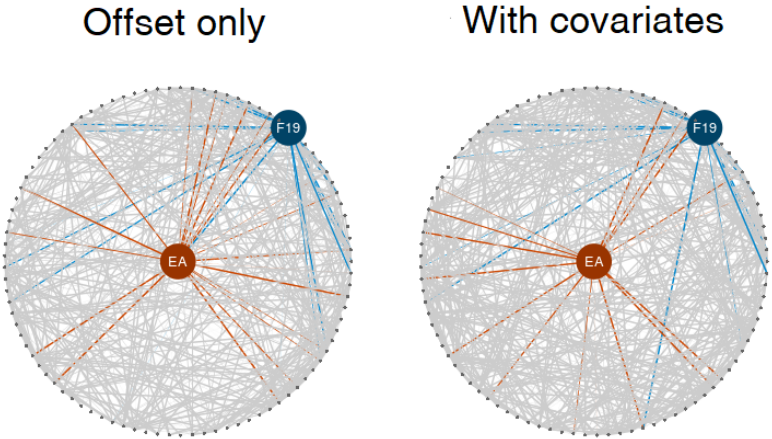
Regression coefficients

	Covariates ($\times 10^{-2}$)		
.	to base	to trunk	to ground
EA	-2.00	2.15	-2.51
F19	2.19	-1.72	1.43

Degree estimation

	Offset	Distances
EA	2.20	1.86
F19	3.03	2.80

Inferred networks



Conclusion

Contributions:

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

Perspectives:

- Method for determining the threshold
- Network comparison
- Model for the inference in the observed counts layer
- Missing major actor (species/covariable)

Acknowledgments

Special thanks :

Supervisors Stéphane Robin, Christophe Ambroise

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

Data Corinne Vacher (INRA Bordeaux)

Contact :

email raphaelle.momal@agroparistech.fr

Web Rmomal.github.io

References I

Aitchison, J. and Ho, C. (1989).
The multivariate Poisson-log normal distribution.
Biometrika, 76(4):643–653.

Chaiken, S. and Kleitman, D. J. (1978).
Matrix tree theorems.
Journal of combinatorial theory, Series A, 24(3):377–381.

Chiquet, J., Mariadassou, M., and Robin, S. (2017).
Variational inference for probabilistic Poisson PCA.
Technical report, arXiv:1703.06633.

References II

Jakuschkin, B., Fievet, V., Schwaller, L., Fort, T., Robin, C., and Vacher, C. (2016).
Deciphering the pathobiome: Intra- and interkingdom interactions involving the pathogen erysiphe alphitoides.
Microb Ecol, 72(4):870–880.
doi:10.1007/s00248-016-0777-x.

Kurtz, Z. D., Müller, C. L., Miraldi, E. R., Littman, D. R., Blaser, M. J., and Bonneau, R. A. (2015).
Sparse and compositionally robust inference of microbial ecological networks.
PLoS computational biology, 11(5):e1004226.

Lauritzen, S. L. (1996).
Graphical Models.