Ecological network reconstruction from count data

Raphaëlle Momal
Supervision: S. Robin[1] and C. Ambroise[2]

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

February 12th, 2019

Context

Rising interest in jointly analysed species abundances:

- Metagenomics
- Microbiology
- Ecology

Ecological network

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

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

Example

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 $\Rightarrow$ $p \times p$ matrix O

Goal:

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

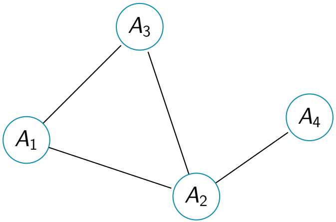
$$\widehat{G} = f(Y, X, O)$$

Challenges

- Statistical network inference
- Count data
- Offsets and covariates

Graphical models: a statistical framework for network inference

Example:

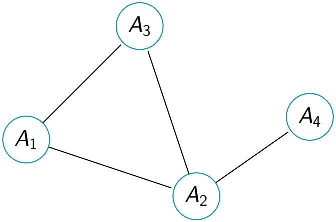


- Connected: all variables are dependant
- Some are conditionally independent (i.e. indirectly dependant)

  A_4 is independent from (A_1, A_3) conditionally on A_2

Graphical models: a statistical framework for network inference

Example:



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

  A_4 is independent from (A_1, A_3) conditionally on A_2

$$P(A_1, \dots, A_p) \propto \prod_{C \in \mathcal{C}_G} \psi_C(A_C)$$

PLN model

Poisson log-Normal distribution (Aitchison and Ho, 1989)

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

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

PLN model

Poisson log-Normal distribution (Aitchison and Ho, 1989)

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

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

PLN model + Graphical model

Poisson log-Normal distribution (Aitchison and Ho, 1989)

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

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

Proposed method: PLN + Spanning trees

Tree structure on PLN latent layer

EMtree model

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

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

Why Spanning trees

Sparse structures:

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

Why Spanning trees

Sparse structures:

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

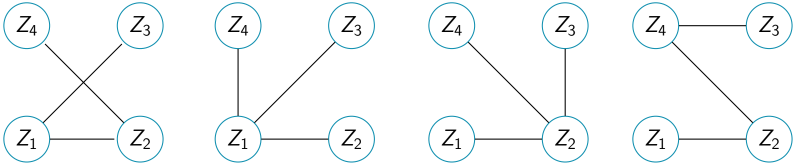
Suitable algebraic tool:

Matrix tree theorem (Chaiken and Kleitman, 1978)

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

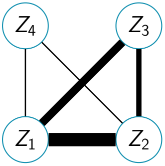
Approach: infer the network by averaging spanning trees

Tree averaging



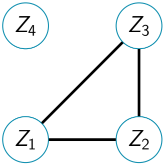
$P\{T = T_1|Z\}$ $P\{T = T_2|Z\}$ $P\{T = T_3|Z\}$ $P\{T = T_4|Z\}$ ...

Compute edge probabilities:



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

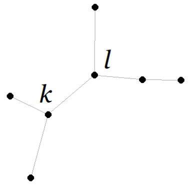
Thresholding probabilities:



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

Tree structured data

- Data dependency structure relies on a tree



- Likelihood factorizes on nodes and edges (Chow and Liu, 1968):

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

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 :

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

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

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:

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

$\Rightarrow$ **EM algorithm** (E: Kirshner (2008), M: Meilă and Jaakkola (2006))

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 β_{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 β_{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 Kurtz et al. (2015)
- gCoda Fang et al. (2017)

One taking raw counts and covariates:

- MInt Biswas et al. (2016) (uses PLN model)

Evaluation strategy

Alternatives:

Two methods on transformed counts, no covariates:

- SpiecEasi algorithm Kurtz et al. (2015)
- gCoda Fang et al. (2017)

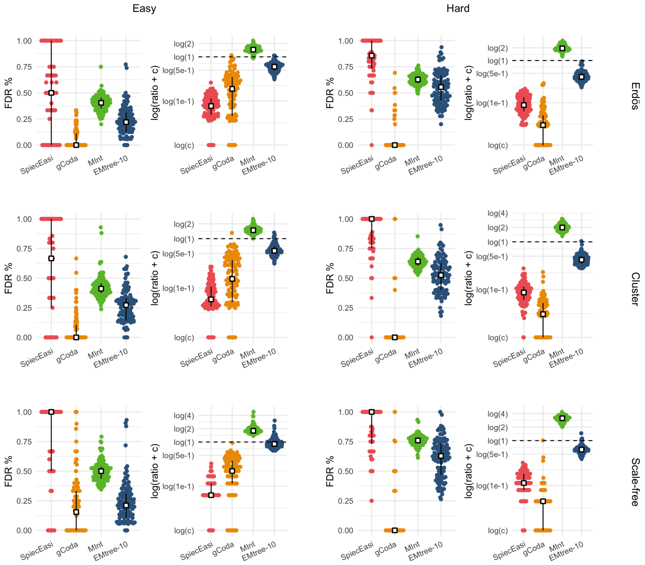
One taking raw counts and covariates:

- MInt Biswas et al. (2016) (uses PLN model)

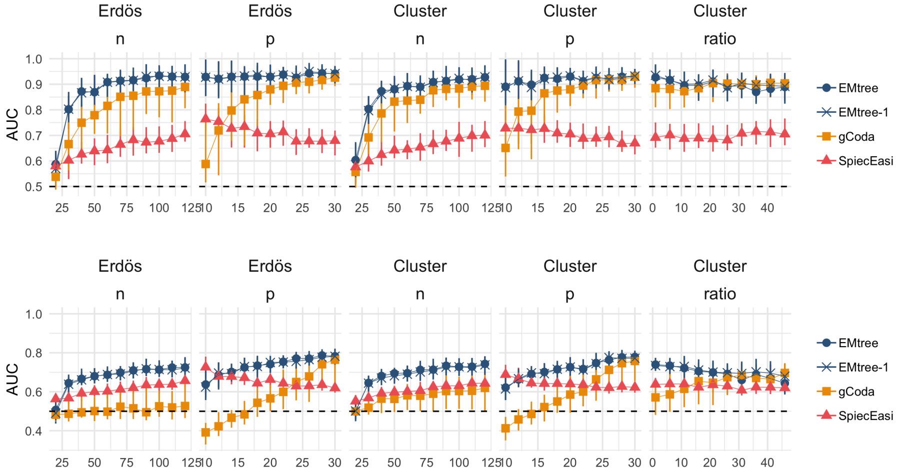
Simulation design:

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

Difficulty level

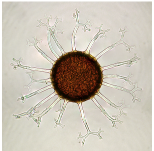


Network density



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 $5/p$.

Oak Mildew



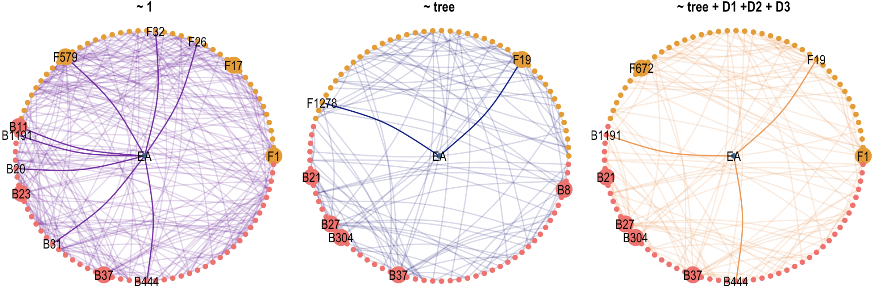
Pathogen Erysiphe alphitoides (EA).

Oak leaf with powdery mildew.

Metabarcoding of oak tree leaves microbiome (Jakuschkin et al., 2016).

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

Inferred networks



Conclusion

Contributions:

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

Perspectives:

- Network comparison
- Missing major actor (species/covariates)
- Model for the inference in the observed counts layer

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

Conditional probability computation

Kirchhoff's theorem (matrix tree, Aitchison and Ho (1989))

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 Meilă and Jordan (2000)

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 :

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

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

$$\hat{\beta}_{kl}^{h+1} = \frac{\tau_{kl}^h}{M_{kl}^h}$$

References I

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

Biswas, S., McDonald, M., Lundberg, D. S., Dangl, J. L., and Jojic, V. (2016). Learning microbial interaction networks from metagenomic count data. *Journal of Computational Biology*, 23(6):526–535.

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. to appear in *Annals of Applied Statistics*.

Chow, C. and Liu, C. (1968). Approximating discrete probability distributions with dependence trees. *IEEE Transactions on Information Theory*, 14(3):462–467.

Fang, H., Huang, C., Zhao, H., and Deng, M. (2017). gcoda: conditional dependence network inference for compositional data. *Journal of Computational Biology*, 24(7):699–708.

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.

Kirshner, S. (2008). Learning with tree-averaged densities and distributions. In *Advances in Neural Information Processing Systems*, pages 761–768.

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.

Meilă, M. and Jaakkola, T. (2006). Tractable bayesian learning of tree belief networks. *Statistics and Computing*, 16(1):77–92.

Meilă, M. and Jordan, M. I. (2000). Learning with mixtures of trees. *Journal of Machine Learning Research*, 1:1–48.