Inference of species interaction networks from abundances

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

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

September 26th, 2019

In a few words

A project using mathematics, statistical modelling and machine learning techniques for applications in microbiology, metagenomics, or ecology.

- Direction:

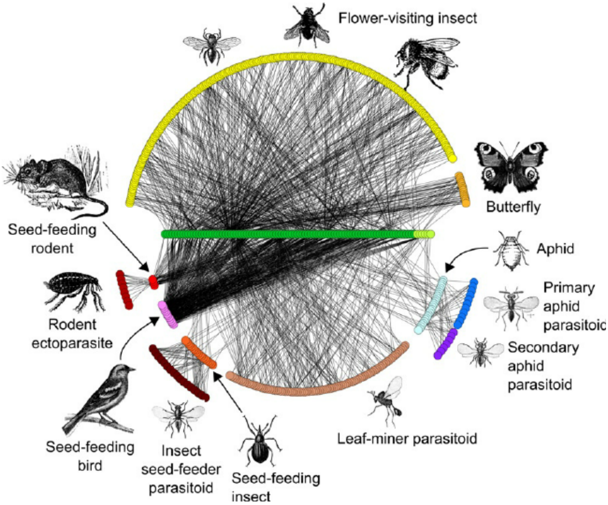
Stéphane Robin

Chistophe Ambroise

- Supports:

INRA SCIENCE & IMPACT — AgroParisTech — LABEX Mathématique Hadamard — Université Paris Sud / Université Paris-Saclay

Network example in ecology



Pocock et. al 2012

- Tool to better understand species interactions, eco-systems organizations
- Allows for resilience analyses, pathogens control, ecosystem comparison, response prediction...

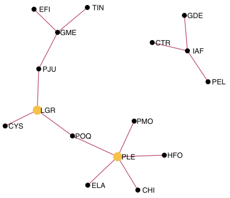
Aim of network inference from abundance data

date	site
apr93	km03
apr93	km03
apr93	km03
apr93	km03
apr93	km17
apr93	km17
⋮	⋮

(*a*) covariates **X**

EFI	ELA	GDE	GME	HFA
71	1	5	6	0
118	2	3	0	0
69	0	6	2	0
56	0	0	0	0
0	1	1	0	0
0	0	2	0	0
⋮	⋮	⋮	⋮	⋮

(*b*) species abundances **Y**



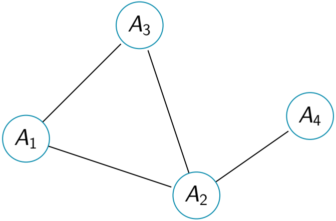
(*c*) inferred network

Data sample from the Fatala river dataset (ade4 R package).

- Unknown underlying structure
- Unobserved interaction data

Graphical models: a statistical framework for conditional dependence

Example:

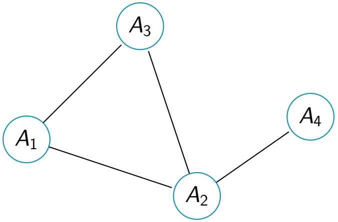


- Connected: all variables are dependant
- Direct dependence or conditional independence

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

Graphical models: a statistical framework for conditional dependence

Example:



- Connected: all variables are dependant
- Direct dependence or conditional independence

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

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

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

$P\ell N$ model

$$Y_{ij} \sim \mathcal{P}\left(\exp(o_{ij} + x_i^\intercal \boldsymbol{\theta}_j + Z_{ij})\right).$$

- A latent variable model
- easy handling of multi-variate data, offsets and covariates

Random effects Z add dependence among species. Classically (Aitchison and Ho, 1989):

$$Z \sim \mathcal{N}(0, \Sigma)$$

We foster sparsity with a mixture of tree structures:

$$Z \sim \sum p(T)\mathcal{N}(0, \Sigma_T), \qquad T \sim \prod_{jk} \beta_{jk}/B$$

Maximum likelihood with hidden data

$$\left.\begin{array}{c}\text{observations } Y \\ \text{hidden parameters } H\end{array}\right\} \Rightarrow \log p(Y) \text{ intractable.}$$

EM algorithm maximizes a surrogate for the log-likelihood :

$$Q = \mathbb{E}[\log p(Y,H)|Y] = \int \log p(Y,h)p(h|Y)dh$$

In most cases the conditional density $p(h|Y)$ is intractable.

Variational EM (VEM) resorts to a proxy $q(h) = \tilde{p}(h|Y)$.

Two hidden quantities

Our model includes two hidden layers of parameters. We need to compute conditional probabilities:

- $p(T|Y)$: computationally complex but tractable thanks to an algebraic mathematical tool (E: Kirshner (2008), M: Meilă and Jaakkola (2006)).
- $p(Z|Y)$: no close form, a VEM gives $\hat{\Sigma}$ and $\hat{\theta}$ (VEM: Chiquet et al. (2017)).

Mixture of trees: sparse and efficient

Sparse structures:

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

Mixture of trees: sparse and efficient

Sparse structures:

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

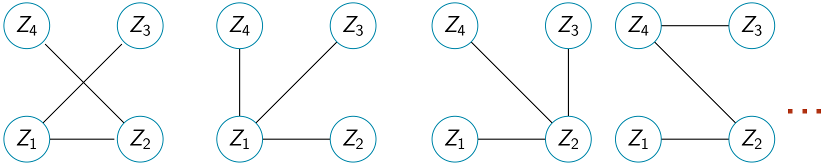
Suitable algebraic tool:

Matrix tree theorem (Chaiken and Kleitman, 1978)

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

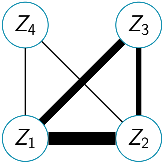
Approach: infer the network by averaging spanning trees

Concept of tree averaging



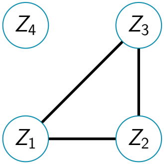
$P\{T = t_1|Y\}$ $P\{T = t_2|Y\}$ $P\{T = t_3|Y\}$ $P\{T = t_4|Y\}$

Compute edge probabilities:



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

Thresholding probabilities:



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.

$$Y_{ij} \sim \mathcal{P}\left(\exp(o_{ij} + x_i^\intercal \boldsymbol{\theta}_j + Z_{ij})\right).$$

$$Z \sim \sum p(T)\mathcal{N}(0, \Sigma_T), \qquad T \sim \prod_{jk} \beta_{jk}/B$$

EMtree algorithm

Input: Abundance data, covariates, offsets

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

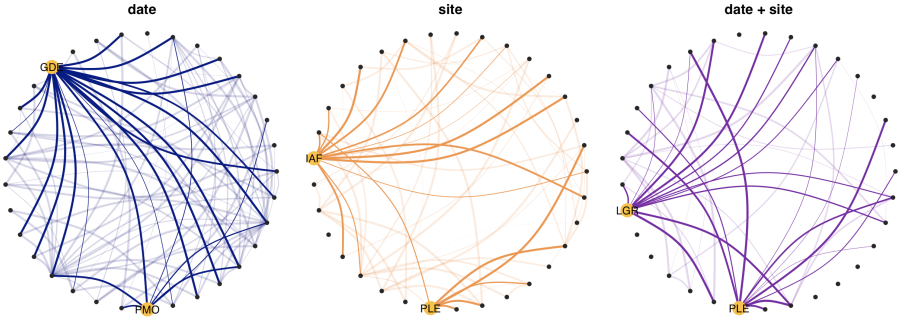
2nd step: EM algorithm to update the β_{jk} $\Rightarrow$ conditional probabilities for all edges.

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`

Inferred networks



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:

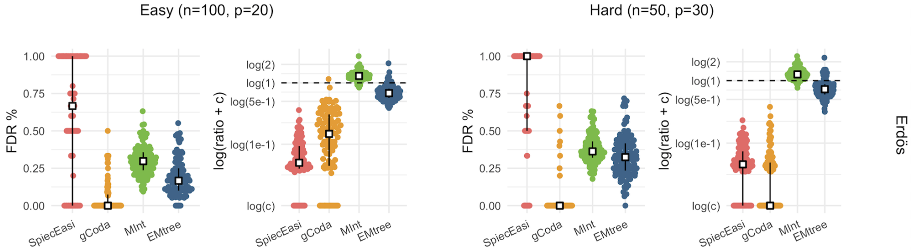
- MInt Biswas et al. (2016) (uses PLN model)

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

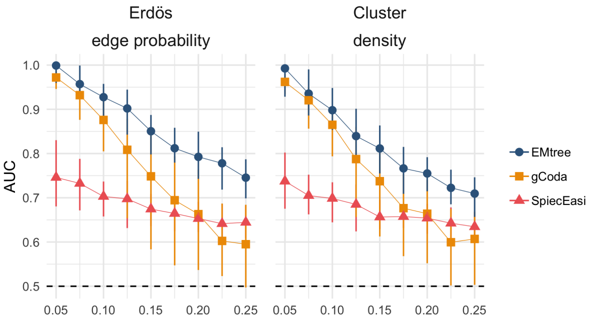
False Discovery Rate (FDR): how many false edges there is among what is detected ?
ratio: number of detections over the number of true edges



- EMtree is a sparser approach than MInt

Network density

Area under the (ROC) curve (AUC): "how good is a classifier to rank true positives higher"
100 observations, 20 species:



Effect of graph density on the evolution of AUC median and inter-quartile intervals in Erdös and Cluster structures.

To be published soon

Tree-based Reconstruction of Ecological Network from Abundance Data

Raphaëlle Momal[1]*, Stéphane Robin[1], Christophe Ambroise[2]

1: UMR MIA-Paris, AgroParisTech, INRA, Université Paris-Saclay, 75005 Paris, France
2: Laboratoire de Mathématiques et Modélisation d'Évry, 23 bvd de France, Évry, France

May 8, 2019

Summary

1. The behavior of ecological systems mainly relies on the interactions between the species it involves. In many situations, these interactions are not observed and have to be inferred from species abundance data. To be relevant, any reconstruction network methodology needs to handle count data and to account for possible environmental effects. It also needs to dis-

Conclusion

Contributions:

- Formal probabilistic model for network inference from count data
- R package: `https://github.com/Rmomal/EMtree`
- Preprint: *Tree-based Reconstruction of Ecological Network from Abundance Data.* `https://arxiv.org/pdf/1905.02452.pdf`

Perspectives:

- Sign and strength of interactions according to graphical models theory
- Missing major actor (species/covariates)
- More collaborations with experts in macro-ecology field

Thank you

Contact :

email	raphaelle.momal@agroparistech.fr
Web	Rmomal.github.io
Twitter	@MomalRaphaelle

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.

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.

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:

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

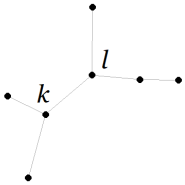
$$\mathbb{P}((k,l) \in T|Z) = \sum_{T \in \mathcal{T}:(k,l) \in T} \mathbb{P}(T|Z) = \frac{\sum_{(k,l)\in T} \mathbb{P}(T)\mathbb{P}(Z|T)}{\sum_T \mathbb{P}(T)\mathbb{P}(Z|T)}$$

$$= 1 - \frac{|Q^*_{uv}(\beta\Psi^{-kl})|}{|Q^*_{uv}(\beta\Psi)|}$$

$$= \tau_{kl}$$

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 :

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

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

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 :

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

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 :

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

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

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