Network inference from incomplete abundance data

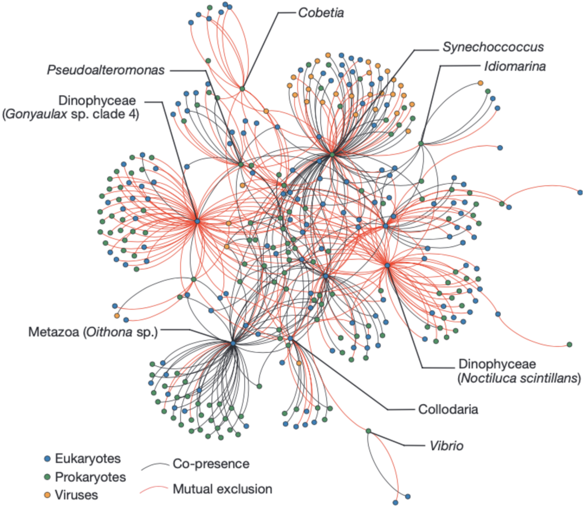
Raphaëlle Momal

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

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

February 18th, 2021

Species co-occurrence network



Integrated plankton community network related to carbon export at 150m (Guidi et. al, 2016)

Reasons for species co-occurrence

Two species can co-occur due to:

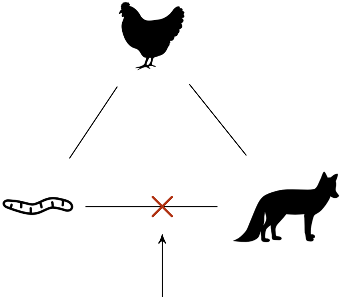
1. a similar response to the same environmental variable,
2. their response to a third species prensence/abundance (mediator species), even if they do not directly depend on one another,
3. their direct association.

Taking environmental effects into account is paramount, yet not enough to separate (2) from (3).

Simple dependencies

After adjusting for environmental covariates, we obtain (residual) correlations between species.

$$\text{correlation} \neq 0 \iff \text{dependence}$$
(Gaussian framework)



Spurious dependence

Dependencies can be direct, or indirect/spurious and due to a mediator species (or unaccounted environmental factor).

⇒ Conditional dependencies are always direct links.

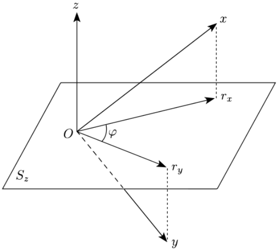
Interpretation of conditional dependencies

Measure of the dependence link between two species after having controlled for the effect of all others.

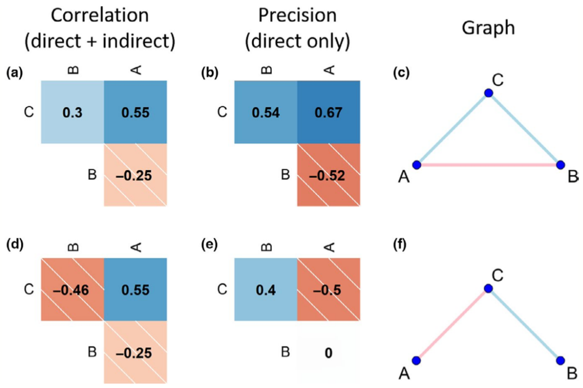
Regression: $Y = \beta_X X + \beta_Z Z + \varepsilon$.

- Y and X are dependent conditionnally on Z $\iff$ $\beta_X \neq 0$.
- Partial correlations quantify this dependence: correlation between the residuals of the regressions of X with Z and of Y with Z ($cos(\varphi)$).

Graphically: are the projections of X and Y on the hyperplan of Z orthogonal?



Two scenarios



Toy-example with Gaussian data (Popovic et al., 2019)

- 1^{rst} line: $A \sim B$,
  2^{nd} line: $A \nsim B$.
- Same $Cor(A, B)$ in both scenarios.
- Only conditional dependences can separate scenarios.

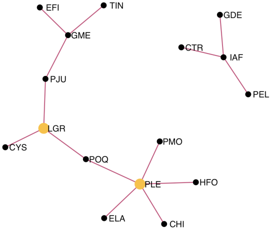
Aim of network inference from abundance data

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

(*a*) species abundances **Y**

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

(*b*) covariates **X**



(*c*) **G**

Data sample from the Fatala river dataset (Baran 1995).

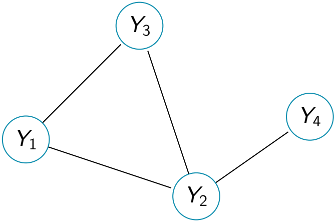
Mathematical framework

i Graphical Models

ii Graph exploration with trees

iii Poisson log-Normal model

Graphical Models



Global Markov:

Y_2 separates Y_3 from $Y_4 \Rightarrow Y_3 \perp\!\!\!\perp Y_4 \mid Y_2$.

Hammersley-Clifford:

Strictly positive and continuous density f:

f global Markov $\iff f(\boldsymbol{Y}) = \prod_{c \in \mathcal{C}} \psi(Y_c)$.

Here $\mathcal{C} = \big\{\{1,2,3\},\{2,4\}\big\}$:

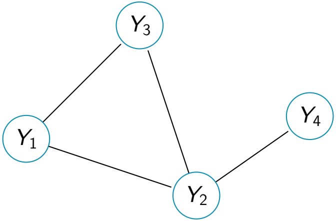
$$f(\boldsymbol{Y}) = \psi(Y_1, Y_2, Y_3) \times \psi(Y_2, Y_4)$$

Gaussian Graphical Models (GGM)

Let $\boldsymbol{Y} \sim \mathcal{N}(\mu, \boldsymbol{\Sigma})$ with precision matrix $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1} = (\omega_{jk})_{jk}$:

$$f(\boldsymbol{Y}) \propto \prod_{j,k,,\omega_{jk}\neq 0} \exp(-Y_k \omega_{jk} Y_j/2).$$

Faithful Markov property:



$$\iff \qquad \boldsymbol{\Omega} = \begin{pmatrix} * & * & * & 0 \\ * & * & * & * \\ * & * & * & 0 \\ 0 & * & 0 & * \end{pmatrix}$$

Gaussian precision terms and conditional dependence

Regression : $X \sim \mathcal{N}(\mu, \mathbf{\Omega}^{-1})$. In the regression $X_j = \sum_{k \neq j} \theta_{jk} X_k + \varepsilon_j$, it holds that $\varepsilon_j \sim \mathcal{N}(0, \omega_{jj}^{-1})$ and $\theta_{jk} = -\omega_{jk}/\omega_{jj}$. Thus $\omega_{jk} \propto \theta_{jk}$

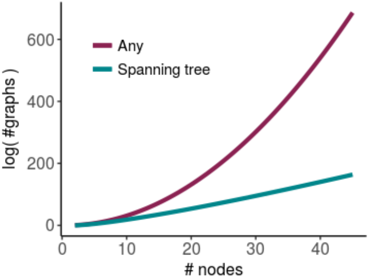
$$\begin{tikzcd}
\text{Covariance/Correlation matrix} \arrow[d, "\text{Inverse}"] \\
\text{Precision matrix } (\omega_{jk})_{jk} \arrow[d, "\text{-Normalized}"] \\
\text{Partial correlations } (\rho_{jk} = -\omega_{jk}/\sqrt{\omega_{jj}\omega_{kk}})
\end{tikzcd}$$

partial correlation/precision $\neq 0 \iff$ conditional dependence
(Gaussian framework)

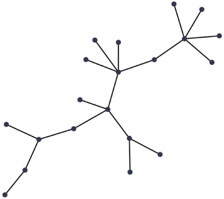
Exploring the graph space

Aim: infer **G**.
Very large space to explore: $\#\mathcal{G}_p = 2^{\frac{p(p-1)}{2}}$

Spanning trees are sparse and simple structures:



- no loops
- $(p-1)$ edges



Much smaller space to explore:

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

Summing over spanning trees

Let $\mathbf{W} = (w_{jk})_{jk}$ be a matrix with null diagonal and positive entries, and $\mathbf{Q}$ its Laplacian:

$$[\mathbf{Q}]_{jk} = \begin{cases} \sum_k w_{jk} & \text{if } j = k \\ -w_{jk} & \text{otherwise} \end{cases}$$

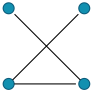
Matrix-tree Theorem (Chaiken and Kleitman, 1978)

All minors of $\mathbf{Q}$ are equal, and for any $1 \leq u, v, \leq p$:

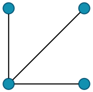
$$|\mathbf{Q}^{uv}| = \sum_{T \in \mathcal{T}} \prod_{jk \in T} w_{jk}$$

Allows to sum over $p^{(p-2)}$ trees in $\mathcal{O}(p^3)$ operations.

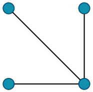
Exploring $\mathcal{T}$ with tree averaging



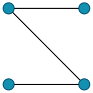
$p(t_1) = 0.12$



$p(t_2) = 0.51$



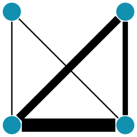
$p(t_3) = 0.02$



$p(t_4) = 0.3$

$\vdots$

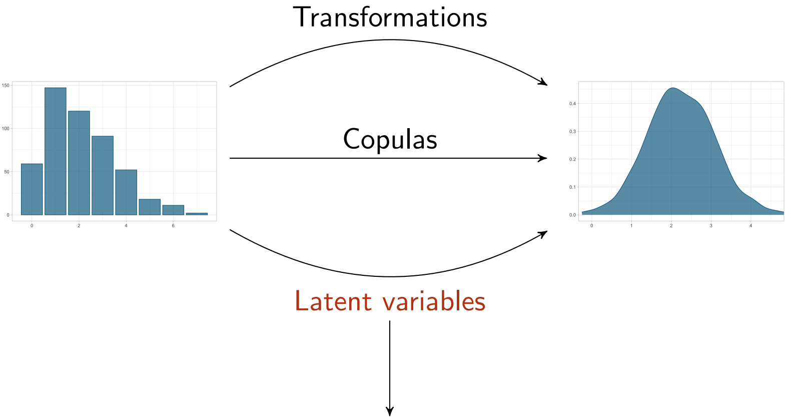
Network inference
= edge probabilities:



$$\mathbb{P}\{k\ell \in T\} = \sum_{\substack{T \in \mathcal{T} \\ k\ell \in T}} p(T)$$

$$p(T) \propto \prod_{kl \in T} w_{kl}$$

Getting back to Gaussian data



Transformations

Copulas

Latent variables

Modeling counts with Gaussian latent parameters

Poisson log-normal model

PℓN model (Aitchison and Ho, 1989) for sample *i* and species *j*:

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

$$Y_{ij} \mid \boldsymbol{Z}_i \sim \mathcal{P}(\exp(\underbrace{o_{ij} + \boldsymbol{x}_i^\intercal \boldsymbol{\theta}_j}_{\text{fixed}} + Z_{ij})).$$

- Latent variables are iid, observed data are independent conditionally on the $\boldsymbol{Z}_i$.
- A generalized multivariate linear mixed model : fixed abiotic and random biotic effects.
- Variational estimation algorithm (PLNmodels, Chiquet et al. (2018))

Network inference from counts

i Model

ii Inference

General model

- Assume a random tree dependency structure T
- Dependence structure in Gaussian layer $\boldsymbol{Z}$
- Distribution for counts $\boldsymbol{Y}$ accounting for covariates/offsets

$$\begin{tikzcd}
T \arrow[d] \\
\boldsymbol{Z} \arrow[d] \\
\boldsymbol{Y}
\end{tikzcd}$$

- Matrix Tree Theorem
- Gaussian Graphical Model
- Poisson log-normal model

$P\ell N$ model with tree-shaped Gaussian parameters

$$\begin{cases} T \sim \prod_{kl\in T} \beta_{kl}/B, \\ \boldsymbol{Z}_i \mid T \sim \mathcal{N}(0, \boldsymbol{\Omega}_T) \\ Y_{ij} \mid \boldsymbol{Z}_i \sim \mathcal{P}(\exp(o_{ij} + \boldsymbol{x}_i^\intercal \boldsymbol{\theta}_j + Z_{ij})). \end{cases}$$

Gaussian mixture with p^{p-2} components:

$$p(\boldsymbol{Z}) = \sum_{T\in\mathcal{T}} p(T)\mathcal{N}(\boldsymbol{Z} \mid T; 0, \boldsymbol{\Omega}_T).$$

Decomposition of the likelihood:

$$p(\boldsymbol{Y}, \boldsymbol{Z}, T) = p_{\boldsymbol{\beta}}(T)\; p_{\Omega_T}(\boldsymbol{Z} \mid T)\; p_{\boldsymbol{\theta}}(\boldsymbol{Y} \mid \boldsymbol{Z}).$$

Two-step procedure

EM algorithm (Dempster et al., 1977)

Maximizes the likelihood in presence of latent variables:

E step: Compute $\mathbb{E}[\log p_{\Theta^t}(\boldsymbol{Y}, \boldsymbol{Z}, T) \mid \boldsymbol{Y}]$

M step: $\Theta^{t+1} = \mathrm{argmax}_{\Theta} \left\{\mathbb{E}[\log p_{\Theta^t}(\boldsymbol{Y}, \boldsymbol{Z}, T) \mid \boldsymbol{Y}]\right\}$

Two-step procedure

EM algorithm (Dempster et al., 1977)

Maximizes the likelihood in presence of latent variables:

E step: Compute $\mathbb{E}[\log p_{\Theta^t}(\boldsymbol{Y}, \boldsymbol{Z}, T) \mid \boldsymbol{Y}]$

M step: $\Theta^{t+1} = \operatorname{argmax}_{\Theta} \left\{ \mathbb{E}[\log p_{\Theta^t}(\boldsymbol{Y}, \boldsymbol{Z}, T) \mid \boldsymbol{Y}] \right\}$

1. `PLNmodels` (Chiquet et al., 2018) gives $\widehat{\boldsymbol{\theta}}$ and approximates of $\boldsymbol{Z} \mid \boldsymbol{Y}$ sufficient statistics.
2. EM algorithm to get $\widehat{\boldsymbol{\beta}}$.

Actually: $\tilde{\mathbb{E}}[\log p_{\beta}(\boldsymbol{Y}, \boldsymbol{Z}, T) \mid \boldsymbol{Z}] = \tilde{\mathbb{E}}[\log p_{\beta}(\boldsymbol{Z}, T) \mid \boldsymbol{Z}] + cst.$

Factorization on the edges

Tree structure factorization:

$$p_{\mathbf{\Omega}_T}(\boldsymbol{Z} \mid T) = \prod_k p(\boldsymbol{Z}_k) \prod_{kl \in T} \frac{p(\boldsymbol{Z}_k, \boldsymbol{Z}_l)}{p(\boldsymbol{Z}_k)\, p(\boldsymbol{Z}_l)}$$

Only the 1^{rst} and 2^{nd} order moments of $\boldsymbol{Z} \mid \boldsymbol{Y}$ are required, replaced by their variational approximation from step 1.

Expression of the surrogate

$$\tilde{\mathbb{E}}[\log p_\beta(\boldsymbol{Z}, T) \mid \boldsymbol{Z}] = \sum_{j<k} P_{jk} \log\left(\beta_{jk}\widehat{\psi}_{jk}\right) - \log B + cst,$$

where $\widehat{\psi}_{jk} = (1 - \widehat{\rho}_{jk}^2)^{-n/2}$ and $P_{jk} = \mathbb{P}\{jk \in T \mid \boldsymbol{Z}\}$.

Proposed EM algorithm

The M matrix is built from the inverse of a Laplacian matrix (Meilă and Jaakkola, 2006).

E step: $p(T \mid \boldsymbol{Z})$ factorizes on the edges.
Using the weight matrix $\mathbf{W} = \boldsymbol{\beta} \odot \widehat{\psi}$, all probabilities can be computed at once:

$$P_{jk} = w_{jk} M(\mathbf{W})_{jk} \text{ (Kirshner, 2008)}$$

M step: Requires the computation of $\partial_{\beta_{jk}}(\sum_{T\in\mathcal{T}} \prod_{jk\in T} \beta_{jk})$.
Update formula:

$$\beta_{jk} = \frac{P_{jk}}{M(\boldsymbol{\beta})_{jk}}$$

This fixed-point problem is solved using optimization, with a gradient ascent procedure.

The M matrix

Lemma (Meilă and Jaakkola, 2006)

$\mathbf{Q}^{pp}$ is the Laplacian matrix $\mathbf{Q}$ of $\mathbf{W}$ to which the last column and row were removed. M is then defined as follows:

$$M(\mathbf{W})_{jk} = \begin{cases} [(\mathbf{Q}^{pp})^{-1}]_{jj} + [(\mathbf{Q}^{pp})^{-1}]_{kk} - 2[(\mathbf{Q}^{pp})^{-1}]_{jk} & 1 \leq j, k < p \\ [(\mathbf{Q}^{pp})^{-1}]_{jj} & k = p, 1 \leq j < p \\ 0 & k = j \end{cases}$$

With $B = \sum_{T \in \mathcal{T}} \prod_{jk \in T} \beta_{jk}$, we then have:

$$\partial_{\beta_{jk}} B = M(\boldsymbol{\beta})_{jk} \times B$$

Contributions

Article Momal R., Robin S., and Ambroise C. . *"Tree-based inference of species interaction networks from abundance data."* Methods in Ecology and Evolution 11.5 (2020): 621-632.

R package `EMtree`: `https://rmomal.github.io/EMtree/`.

The article provides with illustrations and comparison to alternative approaches (SpiecEasi, gCoda, ecoCopula, MInt, and MRFcov) on simulated data with different types of dependency structures and 20 to 30 variables.

Practical developments

i Larger networks

ii Threshold selection

iii Partial correlations

Numerical stability and the Matrix Tree Theorem

- The MTT operator $\sum_{T\in\mathcal{T}}\prod_{jk\in T} x = p^{(p-2)}x^{(p-1)}$ quickly reaches the machine precision (ex: $x=1$ and $p=200$ gives numerical infinity).

  Upper and lower bounds for $\boldsymbol{\beta}$, which depend on p and the machine precision limits Δ_{min} and Δ_{max}:

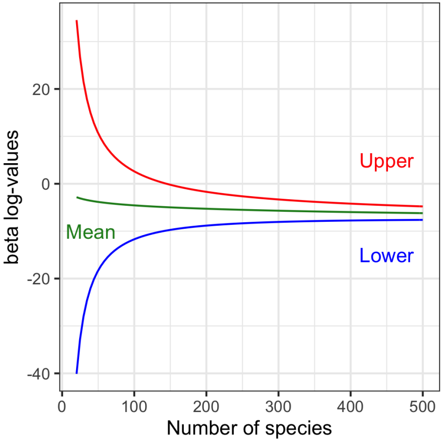
  $$\left(\Delta_{min}p^{-(p-2)}\right)^{1/(p-1)} < \beta_{jk} < \left(\Delta_{max}p^{-(p-2)}\right)^{1/(p-1)}$$

- If $\boldsymbol{\beta}$ has too many high values, $\mathbf{Q}(\boldsymbol{\beta})^{11}$ can become numerically non positive-definite (conditioning$< 1e-16$).

  Optimization under mean constraint:

  $$\bar{\boldsymbol{\beta}} = p^{-(p-2)/(p-1)}.$$

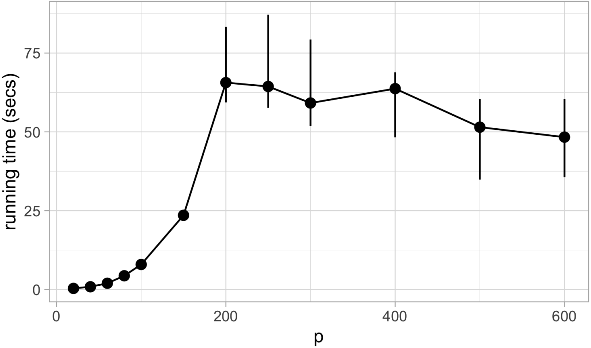
Numerical stability and the Matrix Tree Theorem



- These constraints are implemented using an L-BFGS-B optimization algorithm during the M step.
- This fosters numerical stability and allows for larger networks.

Evolution of running time

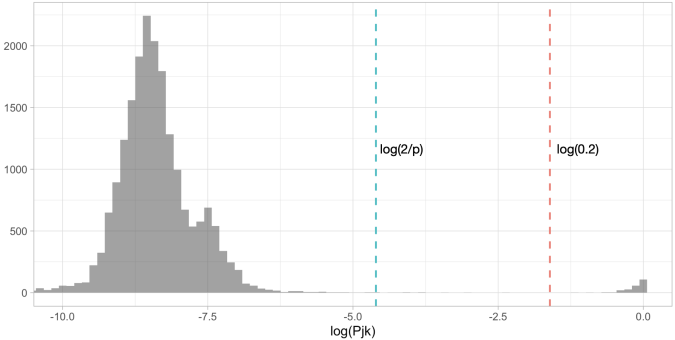
- Varying number of nodes p from 20 to 600.
- Erdös random graphs with edge probability of $3/p$.
- 20 graphs at each point.



Quality assessment

The AUC would give misleading results due to the growing amount of negatives. Setting a threshold, we can assess the quality of the selected set of edges.

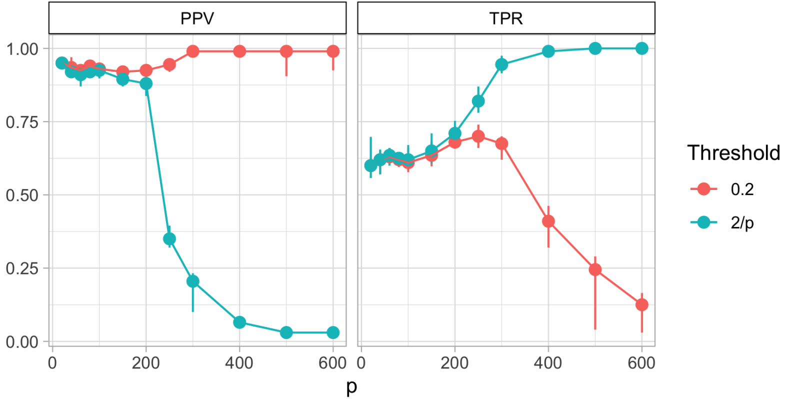
We use a fixed probability threshold of 0.2, and the average value $2/p$.



Example of a distribution of the P_{jk} probabilities with $p = 200$ nodes.

Inference quality for determined thresholds

- PPV= TP/(TP+FP): amount of truth among detection (precision).
- TPR=TP/(TP+FN): amount of truth detected (recall).



Inference of large networks with EMtree

- Is numerically possible
- Demands reasonable running time.
- Thresholds performance on simulated data: $2/p$ becomes too small and 0.2 too big.

$\Rightarrow$ Need for a threshold selection strategy.

Stability selection concept

Seminal paper: StARS (Stability approach to regularization selection, Liu et al., 2010)

- Developed in a regularization context to select the optimal penalty.
- Standard procedure for penalty selection in the inference with graphical LASSO.
- Measures, for each penalty, the average variability of edges selection across resamples.

Here we adapt this to a stability approach to threshold selection.

Edge selection frequencies

1. Create B random sub-samples using 80% of input data

2.

b	edges scores				
1	2e-04	0.0024	0.0414	0.2507	
2	1e-04	0.0013	0.0004	0.0574	
3	2e-04	0.0013	0.0008	0.0127	...
⋮	⋮	⋮	⋮	⋮	

3. Apply threshold α on all resampled scores

4. $f_{jk}^{\alpha} = \sum_{b=1}^{B} \mathbb{1}\{P_{jk}^{s} \geq \alpha\}/B$

q edges selection frequencies: 0.000 0.0381 0.0190 0.7048 ...

Stability of frequencies

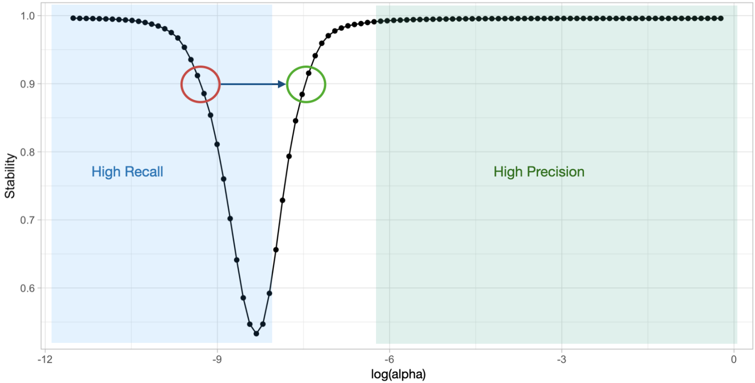
The stability s_α varies between 0 and 1 and is defined as:

$$s_\alpha = 1 - 4 \quad \underbrace{\frac{1}{q}\sum_{j<k} f_{jk}^{\alpha}(1 - f_{jk}^{\alpha})}_{\text{Mean of bernoulli variances}} \quad .$$

- Stability selection requires to set a desired stability value s^* (stability threshold).
- The optimal threshold α^* is then

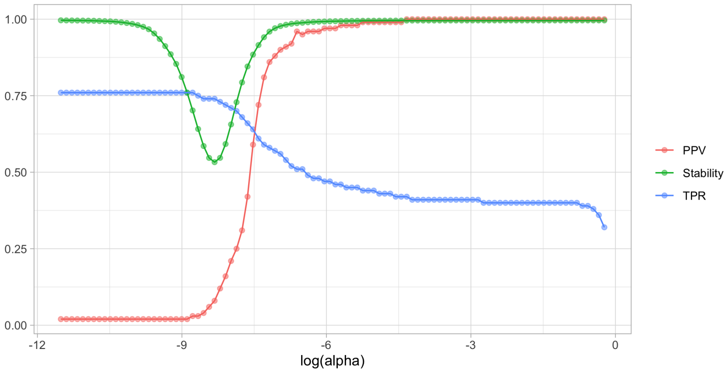
$$\alpha^* = \underset{\alpha}{\operatorname{argmin}}\{s_\alpha - s^*\}$$

Stability and quality



Stability is 1 if α is too big (empty selection) or too small (complete selection). For any s^*, the larger value for α should by chosen.

Example on a 200 nodes Erdös graph



Simulations

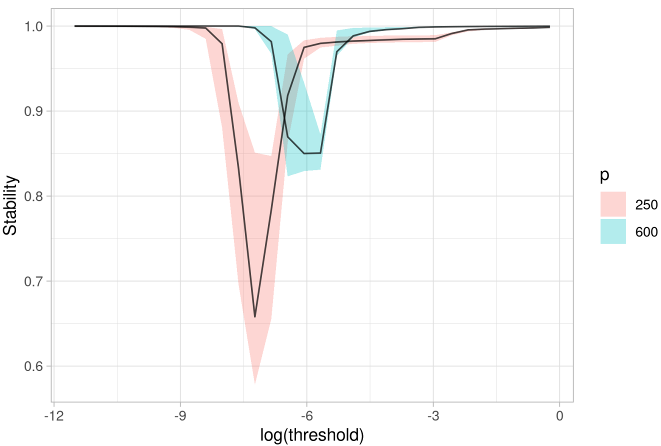
Design:

- 30 graphs for each $p \in \{250, 600\}$.
- Count data is simulated under the PLN model.
- Inference with EMtree and a resampling of size 30.

Thresholding: Keep the edges with a score higher than α^* in more than 90% of the resamples.

Performance: Comparison with thresholds 0.2 and $2/p$.

Stability profiles



Performance

Medians and standard deviations over the 30 inferences:

$p = 250$:

	2/p	0.2	$\alpha^*(90)$	$\alpha^*(95)$	$\alpha^*(98)$
PPV	0.67 (0.14)	1.00 (0.01)	0.60 (0.15)	0.64 (0.13)	0.70 (0.16)
TPR	0.66 (0.08)	0.59 (0.07)	0.82 (0.08)	0.79 (0.09)	0.71 (0.10)

$p = 600$:

	2/p	0.2	$\alpha^*(90)$	$\alpha^*(95)$	$\alpha^*(98)$
PPV	0.09 (0.015)	1.00 (0.05)	0.53 (0.16)	0.53 (0.09)	0.66 (0.04)
TPR	1.00 (0.00)	0.17 (0.09)	0.97 (0.05)	0.97 (0.05)	0.94 (0.04)

Computing partial correlations

$$\rho_{jk} = \frac{-\omega_{jk}}{\sqrt{\omega_{kk}\omega_{jj}}}$$

Partial correlations are paramount in the study of sign and strength of species interactions. They can be computed from estimates of $\boldsymbol{\Sigma}$ or $\boldsymbol{\Omega}$, which EMtree does not provide.

However, the R package `ggm` (Marchetti et al., 2006) implements an iterative procedure to fit the model by maximum likelihood (Speed and Kiiveri, 1986).

Input data:

- Empirical covariance matrix (S_{PLN})
- Estimate of the adjacency matrix ($\hat{G}$, output from EMtree)

Graphical LASSO

The glasso (Friedman et al., 2008) estimates the precision matrix with an ℓ_1 penalized regularization:

$$\underset{\mathbf{\Omega}\geq 0}{\operatorname{argmax}} \left\{ \log|\mathbf{\Omega}| + tr\mathbf{Z}^\intercal\mathbf{Z}\mathbf{\Omega} - \lambda||\mathbf{\Omega}||_1 \right\}, \qquad ||\mathbf{\Omega}||_1 = \sum_{j\neq k} |\omega_{jk}|.$$

The inference is conducted on a grid of λ. Here we choose the penalty giving the $\hat{\mathbf{\Omega}}$ which minimizes the error on the partial correlations.

Simulations

Partial computations are computed from:

- ggm oracle: MLE fit of $\boldsymbol{\Sigma}$ with G
- ggm $\hat{G}$: MLE fit of $\boldsymbol{\Sigma}$ with $\hat{G}$
- min_glasso: the glasso estimate of $\boldsymbol{\Omega}$ minimizing the MSE
- naive: S_{PLN}

	p=50	p=200
ggm oracle	2.6e-4 (1.5e-4)	7.0e-5 (3.6e-5)
ggm $\hat{G}$	1.5e-3 (3.3e-4)	2.7e-4 (7.1e-5)
min_glasso	2.1e-3 (2.9e-4)	6.2e-4 (6.4e-5)
naive	6.0e-3 (6.2e-4)	2.9e-3 (1.2e-4)

Median and standard deviation of mean square errors of the partial correlations, on 30 Erdös graphs.

Conclusion

Model

- A probabilistic model for the inference of conditional dependency networks from abundance data.
- Uses a latent mixture of trees-shaped Gaussian variables to cast the problem in the GGM framework.

Inference

- An EM algorithm which combines the GGM framework flexibility and spanning trees algebraic properties.
- Outputs edges probabilities of membership to the latent tree.

Conclusion

Developments

- Constraints on tree parameters makes it numerically possible to manage large datasets
- Stability approach gives promising results for threshold selection
- The combination of PLN and EMtree outputs allows to get partial correlation estimates.

Perspectives

- Robustness assessment.
- PostDoc at MetaGenoPolis: network comparison through microbial guilds in human gut microbiota.

Thank you!

raphaelle.momal@inrae.fr

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. (2018). Variational inference for probabilistic poisson pca. *The Annals of Applied Statistics*, 12(4):2674–2698.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Royal Statist. Soc., series B*, 39:1–38.

Friedman, J., Hastie, T., and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441.

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

Liu, H., Roeder, K., and Wasserman, L. (2010). Stability approach to regularization selection (stars) for high dimensional graphical models. In *Advances in neural information processing systems*, pages 1432–1440.

Marchetti, G. M. et al. (2006). Independencies induced from a graphical markov model after marginalization and conditioning: the r package ggm. *Journal of Statistical Software*, 15(6):1–15.

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

Popovic, G. C., Warton, D. I., Thomson, F. J., Hui, F. K. C., and Moles, A. T. (2019). Untangling direct species associations from indirect mediator species effects with graphical models. *Methods in Ecology and Evolution*, 10(9):1571–1583.

Speed, T. P. and Kiiveri, H. T. (1986). Gaussian markov distributions over finite graphs. *The Annals of Statistics*, pages 138–150.