Inference of species interaction networks with missing actors from abundance data

Raphaëlle Momal

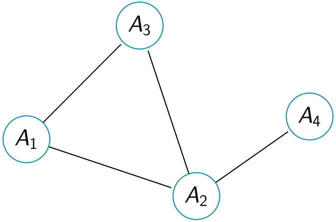
Supervision: S. Robin[1] and C. Ambroise[2]

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

March 10th, 2020

Statistical framework for conditional dependence

Graphical Models:



- Connected: all variables are dependant
- Markov property : G encodes the conditional independences

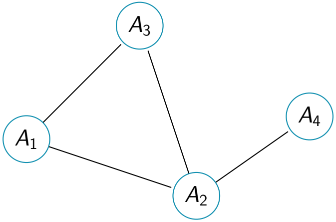
  e.g. $A_4 \perp\!\!\!\perp (A_1, A_3) \,|A_2$

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

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

Statistical framework for conditional dependence

Graphical Models:



- Connected: all variables are dependant

- Markov property : G encodes the conditional independences

  e.g. $A_4 \perp\!\!\!\perp (A_1, A_3) \,|A_2$

Here:

$$P(A) \propto \psi_1(A_1, A_2, A_3)\, \psi_2(A_2, A_4)$$

Gaussian Graphical Models (GGM)

$$Y = (Y_1, ..., Y_d) \sim \mathcal{N}_d(0, \Omega^{-1})$$

The factorization is straightforward:

$$p(y) \propto \prod_{j,k:\omega_{jk}\neq 0} exp(-y_j \omega_{jk} y_k / 2)$$

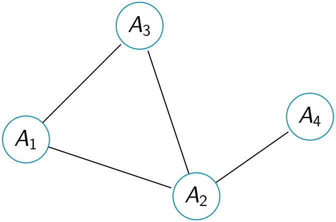
Gaussian Graphical Models (GGM)

$$Y = (Y_1, ..., Y_d) \sim \mathcal{N}_d(0, \Omega^{-1})$$

The factorization is straightforward:

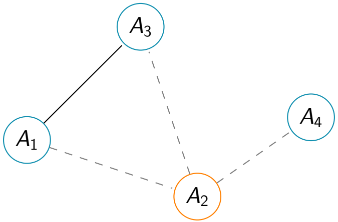
$$p(y) \propto \prod_{j,k:\omega_{jk} \neq 0} exp(-y_j \omega_{jk} y_k / 2)$$

$$\Omega = \begin{pmatrix} * & * & * & 0 \\ * & * & * & * \\ * & * & * & 0 \\ 0 & * & 0 & * \end{pmatrix} \Rightarrow$$



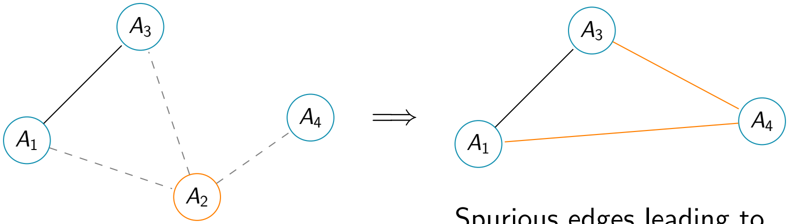
Missing actor

A_2 is not observed:



Missing actor

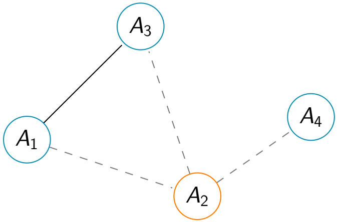
A_2 is not observed:



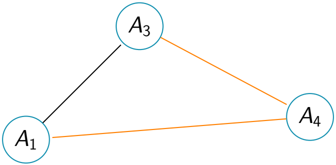
Spurious edges leading to wrong interpretation

Missing actor

A_2 is not observed:



$\Longrightarrow$

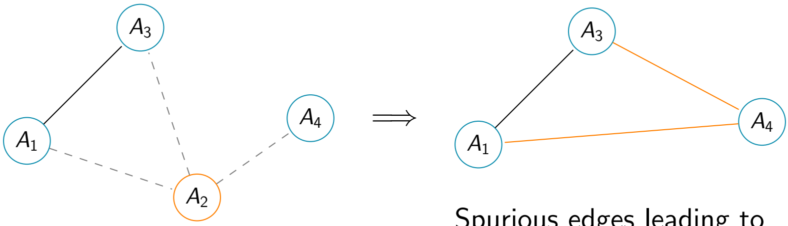


Spurious edges leading to wrong interpretation

How to infer a missing actor in a network ?

Missing actor

A_2 is not observed:



Spurious edges leading to wrong interpretation

How to infer (a missing actor in) a network from abundance data?

Graphical model for abundance data

$P\ell N$ model:

$$Y_{ij} \sim \mathcal{P}\Big(\exp(\underbrace{o_{ij} + x_i^\intercal \boldsymbol{\theta}_j}_{\text{fixed}} + \underbrace{Z_{ij}}_{\text{random}})\Big).$$

Graphical model for abundance data

$P\ell N$ model:

$$Y_{ij} \sim \mathcal{P}\Big(\exp(\underbrace{o_{ij} + x_i^\intercal \boldsymbol{\theta}_j}_{\text{fixed}} + \underbrace{Z_{ij}}_{\text{random}})\Big).$$

- Classically (Aitchison and Ho, 1989): $\mathbf{Z}_i \sim \mathcal{N}(0, \Omega^{-1})$ iid
- Easy handling of multi-variate data, offsets and covariates (Chiquet et al., 2018)

GGM: Ω encodes the conditional dependency structure.

Graphical model for abundance data

$P\ell N$ model:

$$Y_{ij} \sim \mathcal{P}\Big(\exp(\underbrace{o_{ij} + x_i^\intercal \boldsymbol{\theta}_j}_{\text{fixed}} + \underbrace{Z_{ij}}_{\text{random}})\Big).$$

- Classically (Aitchison and Ho, 1989): $\mathbf{Z}_i \sim \mathcal{N}(0, \Omega^{-1})$ iid
- Easy handling of multi-variate data, offsets and covariates (Chiquet et al., 2018)

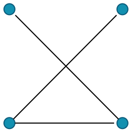
GGM: Ω encodes the conditional dependency structure.

Momal et al. (2020): foster sparsity with a random spanning tree:

$$Z|T \sim \mathcal{N}(0, {\Omega_T}^{-1})$$

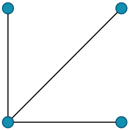
Inference with an efficient variational EM algorithm.

Explore the space with trees



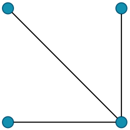
$p(T = t_1) = 0.12$

Explore the space with trees



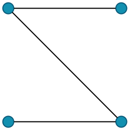
$p(T = t_2) = 0.51$

Explore the space with trees



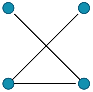
$p(T = t_3) = 0.02$

Explore the space with trees

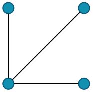


$p(T = t_4) = 0.3$

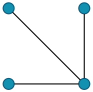
Explore the space with trees



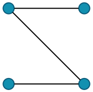
$p(t_1) = 0.12$



$p(t_2) = 0.51$



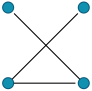
$p(t_3) = 0.02$



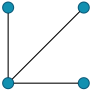
$p(t_4) = 0.3$

$\vdots$

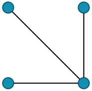
Explore the space with trees



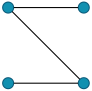
$p(t_1) = 0.12$



$p(t_2) = 0.51$



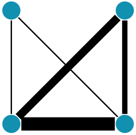
$p(t_3) = 0.02$



$p(t_4) = 0.3$

⋮

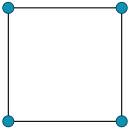
Edge probabilities[1]:



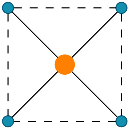
$$\mathbb{P}((j,k) \in T) = \sum_{\substack{t \in \mathcal{T} \\ (j,k) \in t}} p(t)$$

[1] https://github.com/Rmomal/EMtree

More dimensions ?



More dimensions ?



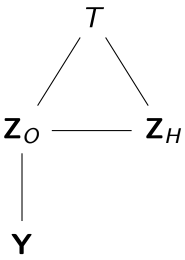
- Unobserved species
- Unobserved covariate ?

Gaussian case: Robin et al. (2019)

Graphical model with missing actors

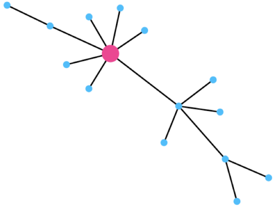
$$vBIC_0 = \mathcal{J}_0(\mathbf{Y}) - \frac{D}{2}\log(n)$$

Graphical model with missing actors

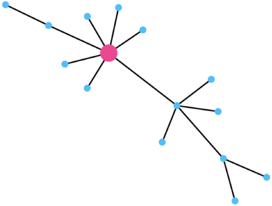


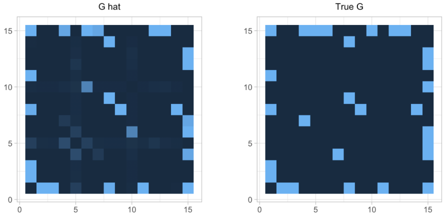
$$vBIC_0 = \mathcal{J}_0(\mathbf{Y}) - \frac{D}{2}\log(n) \qquad vBIC_q = \mathcal{J}_q(\mathbf{Y}) - \frac{D+2pq}{2}\log(n)$$

Results on a scale-free graph /!\Work in progress /!\

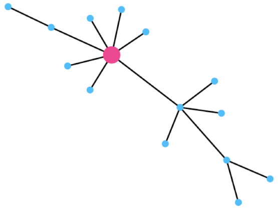


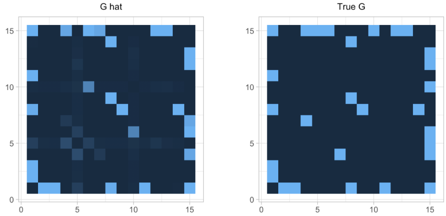
Results on a scale-free graph /!\Work in progress /!\

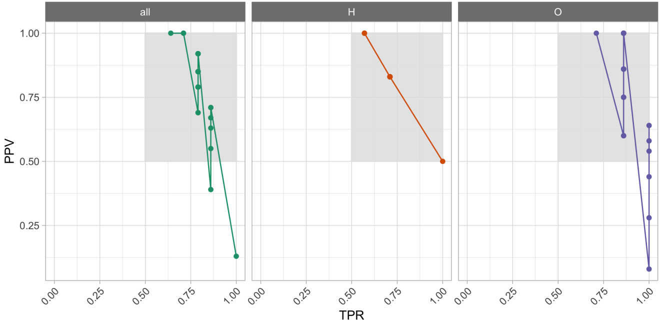




Results on a scale-free graph /!\Work in progress /!\







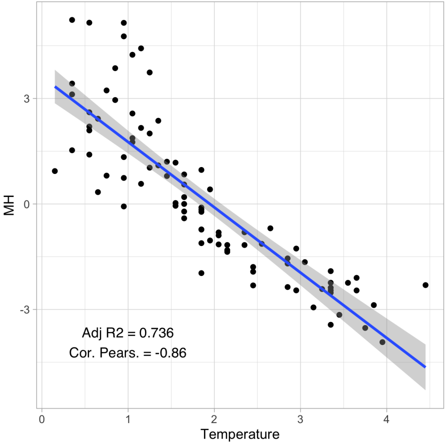
Barents fish data

- Abundances of 30 species in 89 sites of the Barents sea
- 4 available covariates (temperature, longitude, latitude, depth)

Selection model criteria (no covariates):

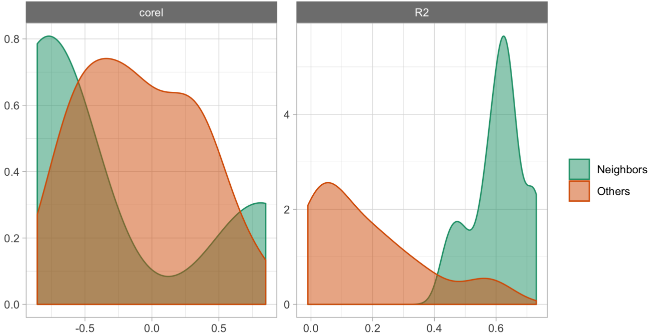
$$vBIC_0 = 4431 \quad \text{vs.} \quad vBIC_1 = 8501$$

Inference with no covariates



The inferred means of the missing actor are highly linked to the temperature

Inference with no covariates



Neighbors of the missing actor are also highly linked to the temperature

In a nutshell

Contributions

- Probabilistic model for detecting missing actor in species interaction network inference from abundance data
- Efficient variational inference, selection model criteria
- Missing actor characterization through its means and neighborhood

Perspectives

- Accounting for spatial effects of ecological datasets for network inference (violation of sites independence hypothesis)
- Spatial effects as missing actors in the network ?

Thank you!

Contact :

email raphaelle.momal@agroparistech.fr

Web Rmomal.github.io

Twitter @MomalRaphaelle

Currently looking for a job opportunity!

References I

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

Chiquet, J., Mariadassou, M., Robin, S., et al. (2018). Variational inference for probabilistic poisson pca. *The Annals of Applied Statistics*, 12(4):2674–2698.

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

Robin, G., Ambroise, C., and Robin, S. (2019). Incomplete graphical model inference via latent tree aggregation. *Statistical Modelling*, 19(5):545–568.