Training school INRAE BOTTOMS-UP

Analysing multispecies distribution with graphical models

Marc Ohlmann, Alpine Ecology Lab, Grenoble

05/10/22

Outline

Introduction

- Context
- Multispecies distribution models
- Conditional reasoning

Graphical models

- Bayesian networks
- Markov networks

Introduction

Introduction
Context

- Single Species Distribution Model: model the statistical link between a given species and its abiotic environment.
- Modelling of the realised niche

What about other species ?

- Each species has specific response to environment
- Species have interrelated distributions due to biotic interactions

Introduction
Context

Joint Species Distribution models: correlative models

- Lot of literature around the multivariate probit model (Pollock et al. 2014, Ovaskainen and Abrego 2020)
- Regression on environment while taking into account correlations between species distributions
- These correlations do not necessarily reflect biotic interactions !
- In gaussian JSDM, same estimation of the niche than SDMs (see Poggiato et al. 2021)

Introduction
Context

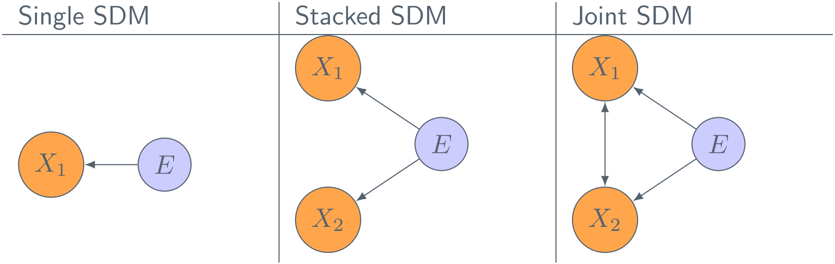
Joint Species Distribution models: correlative models

- Lot of literature around the multivariate probit model (Pollock et al. 2014, Ovaskainen and Abrego 2020)
- Regression on environment while taking into account correlations between species distributions
- These correlations do not necessarily reflect biotic interactions !
- In gaussian JSDM, same estimation of the niche than SDMs (see Poggiato et al. 2021)

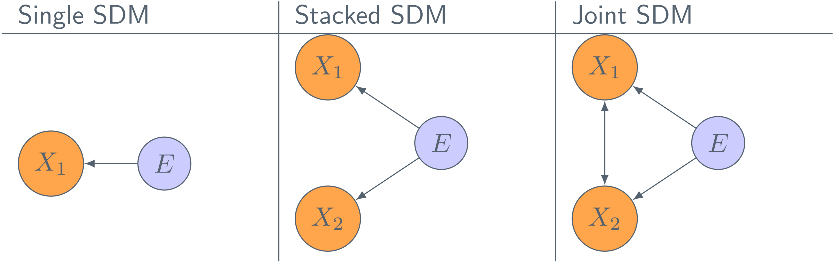
Here, focus on **graph based** approaches.

- Useful for interpretation of statistical dependencies structure

Introduction
Context



Introduction
Context



- In Joint SDM, we consider statistical links between species distribution (X_1 and $X2$) and environment (E) but also statistical links between species distribution

Introduction
Conditional independence

Let X,Y and Z three random variables (rvs)

Independence

$X \perp\!\!\!\perp Y$ if :

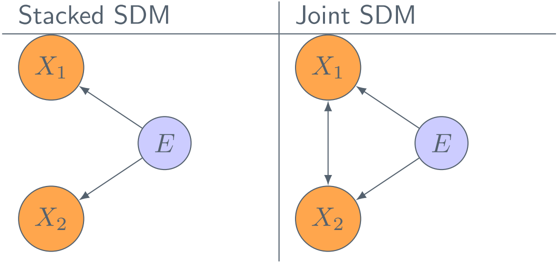
$$\mathbb{P}(X,Y) = \mathbb{P}(X)\mathbb{P}(Y)$$

Conditional independence

$X \perp\!\!\!\perp Y|Z$ if :

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

Introduction
Conditional independence and SDMs



- In stacked SDM: $X1 \perp\!\!\!\perp X2|E$ BUT $X_1 \not\perp\!\!\!\perp X_2$ due to the environment
- In Joint SDM, we do not assume $X1 \perp\!\!\!\perp X2|E$ and consider that species distributions can influence each others conditionally to the environment

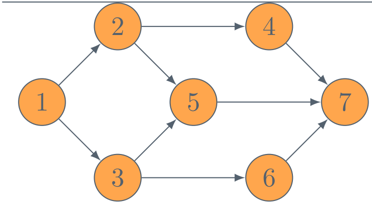
Introduction
Graphical models

Graphical models is a class of statistical models that aim mapping conditional independence statements through graphs.

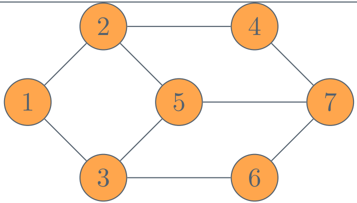
Graphs

A graph G is a pair $G = (V, E)$ where V is the set of nodes and E is the set of edges

Directed graph



Undirected graph



Introduction
Graphical models

Generally speaking, in graphical models (see Schwaller 2015 for an introduction):

- Nodes are random variables (species distribution, abiotic environment)
- Edges represent conditional independence statements
- Edges are inferred by the model

Introduction
Graphical models

Generally speaking, in graphical models (see Schwaller 2015 for an introduction):

- Nodes are random variables (species distribution, abiotic environment)
- Edges represent conditional independence statements
- Edges are inferred by the model

Graphical models and biotic interactions

Is conditional dependence a good representation of biotic interactions ?

- "other things being equal" thinking
- No causality in the definition: association networks

Several ecological processes can lead to correlated species distributions, do not over-interpret the inferred networks in terms of biotic interactions (see Blanchet, Cazelles, and Gravel 2020)

Graphical models

Two families of graphical models:

- Directed graphical models (Bayesian networks)
  Describe statistical dependencies using directed graph
- Undirected graphical models (Markov networks)
  Describe statistical dependencies using undirected graph

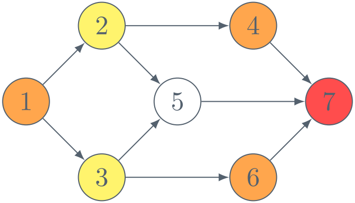
Graphical models

Bayesian networks
DAGs

Bayesian networks describe conditional independence statements using Directed Acyclic Graphs

Directed Acyclic Graph

A Directed Acyclic Graph (DAG) is a directed graph with no cycles



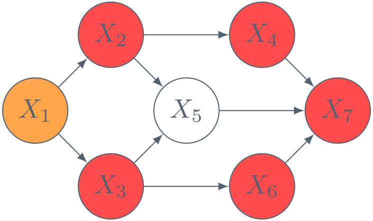
- 7 is a child of 5
- 2 and 3 are the parents of 5

Bayesian networks
Conditional independence

Markov blanket

The markov blanket of a node X, $MB(X)$ is the set of nodes that makes X conditionally independant of all the other nodes

- In Bayesian networks, the Markov blanket is made of children, parents and arents of the children



$$MB(X_5) = \{X_2, X_3, X_4, X_6, X_7\}$$

$$\mathbb{P}(X_5|X_1, X_2, X_3, X_4, X_6, X_7) = \mathbb{P}(X_5|X_2, X_3, X_4, X_6, X_7)$$

Bayesian networks
Factorisation

Factorisation

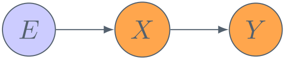
The joint probability factorises over the graph G as :

$$\mathbb{P}(X_1, ..., X_n) = \prod_i \mathbb{P}(X_i | \mathbf{X}_{\mathbf{Pa_i(G)}})$$

- The structure of G gives the dependencies
- Hierarchical perspective

Bayesian networks
Example

Let E be the rv associated to the environment, X plant distribution and Y herbivore distribution.
We assume the following Bayesian network to describe this system:



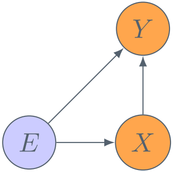
$$\mathbb{P}(E, X, Y) = \mathbb{P}(X|E)\mathbb{P}(Y|X)$$

We have: $E \perp\!\!\!\perp Y|X$, meaning that knowing plant distribution, environment and herbivore distribution are conditionally independent.

In ecological terms, environment does not affect directly herbivores, only trough plants

Bayesian networks
Example 2

We keep the same variables.

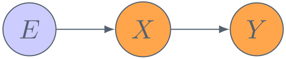


$$\mathbb{P}(E,X,Y) = \mathbb{P}(X|E)\mathbb{P}(Y|X,E)$$

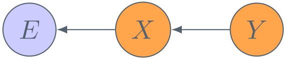
We have: $E \not\perp\!\!\!\perp Y|X$, meaning environment directly affects herbivore distribution, even when knowing plan distribution

Bayesian networks
Example 3

Do not interpret the direction of arrows in terms of causal relationships !



and



Represent the same conditional dependencies.

- Imagine successive linear regressions
- You can explain plant by environment or the opposite !

Bayesian networks
On data

The setting:

- You have species abundances and environmental covariates in several locations
- You want to infer dependencies structure using a Bayesian network

What tools ?

- Several R packages for structure learning are available (*bnlearn*,*bnstruct*), non-parametric or parametric
- Representation and analysis of the inferred network

Once the structure fixes, inference of the parameters (think of response curves between variables)

Bayesian networks
Miscellaneous

- Links between Bayesian networks and structural equation models (SEMs)
- In SEMs, structure is fixed according to expert-knowledge. Links are then tested

Bayesian networks
SEMs

Example from Martinez-Almoyna et al. 2019

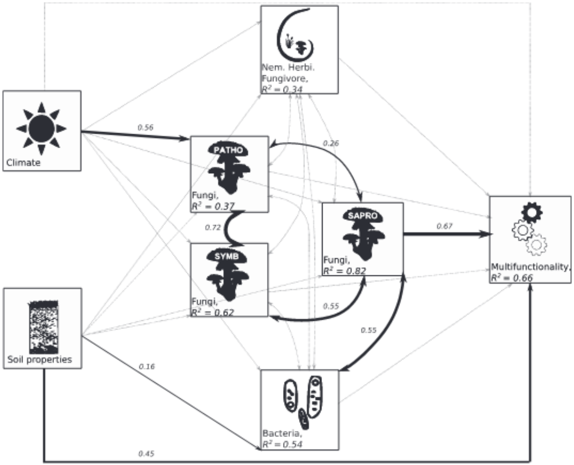
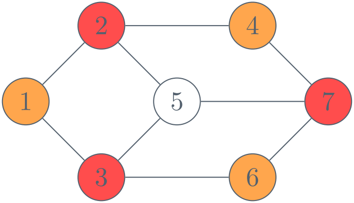


FIGURE 6 Integrated path model highlighting the direct and indirect effects of climate and soil properties turnover on turnover of ecosystem multifunctionality. The size of the arrows is proportional to the size of the associated standardized path coefficients (only for significant paths). Dotted grey lines represent non-significant paths. Paths with double arrows represent correlations

Markov networks

Markov networks describe conditional Independence statement between random variables using undirected graphs.



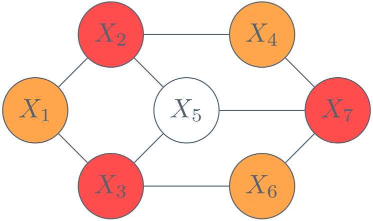
- 2,3,7 are neighbors of 5

Markov networks
Conditional independence

Markov blanket

The markov blanket of a node X, $MB(X)$ is the set of nodes that makes X conditionally independant of all the other nodes

- In Markov networks, the Markov blanket is made of neighbors
- Two disconnected nodes are conditionally independent



$$MB(X_5) = \{X_2, X_3, X_7\}$$

$$\mathbb{P}(X_5|X_1, X_2, X_3, X_4, X_6, X_7) = \mathbb{P}(X_5|X_2, X_3, X_7)$$

Markov networks
Gaussian graphical models

- We assume overall species distribution $\mathbf{X}$ is a multivariate normal distribution
- $X \sim \mathcal{N}(\mu, \mathbf{\Sigma})$
- $\mathbf{\Sigma}$ is the marginal variance-covariance matrix
- For normal distribution, $\mathbf{\Sigma}^{-1}$ is the precision matrix, that is linked to partial correlations

Markov networks

Gaussian graphical models

- We assume overall species distribution $\mathbf{X}$ is a multivariate normal distribution
- $X \sim \mathcal{N}(\mu, \mathbf{\Sigma})$
- $\mathbf{\Sigma}$ is the marginal variance-covariance matrix
- For normal distribution, $\mathbf{\Sigma}^{-1}$ is the precision matrix, that is linked to partial correlations

Partial correlation

What is a partial correlation ?

- Regression coefficient (partial) in multivariate linear model (r partial, R marginal)
- Correlation between two variables once removed the influence of the others

Markov networks
Gaussian graphical models

Partial correlation

What is partial correlation ?

- Regression coefficient (partial) in multivariate linear model (r partial, R marginal)
- Correlation between two variables once removed the influence of the others

What is the link between $\mathbf{\Omega} := \mathbf{\Sigma}^{-1}$ and partial correlations ?

$$\rho_{X_i,X_j\cdot\mathbf{V}\setminus\{X_i,X_j\}} = -\frac{\Omega_{i,j}}{\sqrt{\Omega_{i,i}\Omega_{j,j}}}$$

Markov networks

Gaussian graphical models

Density of $\mathbf{X}$:

$$\mathbb{P}(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^k|\mathbf{\Sigma}|}}\exp\left(-\frac{1}{2}(\mathbf{x}-\mu)^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\mu)\right)$$

- The density of the multivariate normal distribution involves the partial variance-covariance matrix

Markov networks
Gaussian graphical models: inference

A typical problem:

- You have samples from a multivariate normal (species distributions)
- You want to infer $\mathbf{\Theta} = \mathbf{\Sigma}^{-1}$, the precision matrix
- Adjacency matrix of the partial correlation network

Markov networks

Gaussian graphical models: inference

A typical problem:

- You have samples from a multivariate normal (species distributions)
- You want to infer $\mathbf{\Theta} = \mathbf{\Sigma}^{-1}$, the precision matrix
- Adjacency matrix of the partial correlation network

Graphical Lasso method Friedman, Hastie, and Tibshirani 2008

From the empirical correlation matrix $\mathbf{S} = \mathbf{X}^{\mathbf{T}}\mathbf{X}$,

$$\hat{\mathbf{\Theta}} = \text{argmin}\left(\text{tr}(\mathbf{S}\mathbf{\Theta}) - \log\det(\mathbf{\Theta}) + \lambda \sum_{j \neq k} \Theta_{j,k}\right)$$

where λ is the penalising parameter controlling number of edges

- Increasing λ decreases the number of edges of the partial correlation network

Markov networks

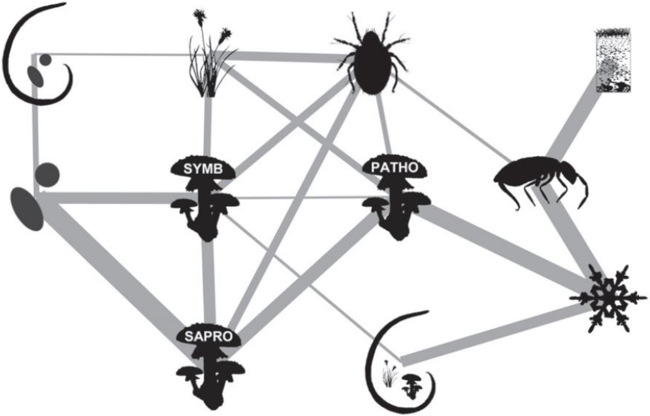
Gaussian graphical models: application on data

- Several R packages to apply graphical lasso on data: *glasso*, *GLASSOO*
- In *glasso* package, the user provides the penalty term λ
- In *GLASSOO*, λ is estimated using cross-validation

Markov networks
Gaussian graphical models: application on data

In Ohlmann et al. 2018

- eDNA soil data along an environmental gradient in the French Alps
- graphical lasso on β-diversity of several soil trophic groups and environmental variables



Markov networks
Adding the environment

How to deal with the abiotic environment in multivariate normal models.

- Either model it as a node of the graphical model (as the previous slide)
- Or do a regression (mean) and analyse the correlation structure of the residuals (JSDMs way)

$$\mathbf{Y} \sim \mathcal{N}(\mathbf{B}^{\mathbf{T}}\mathbf{X}, \mathbf{\Sigma})$$

$\mathbf{B}^{\mathbf{T}}\mathbf{X}$ is the regression on environmental covariables

Markov networks
For other distributions

Markov networks for several distribution (see Yang et al. 2014):

- binary graphical model
- Poisson graphical model (constraints on the association parameters)
- Graphical model mixing different types of distribution (see R package *XMRF*)

Markov networks
For other distributions

Markov networks for several distribution (see Yang et al. 2014):

- binary graphical model
- Poisson graphical model (constraints on the association parameters)
- Graphical model mixing different types of distribution (see R package *XMRF*)

Markov networks
Multivariate Poisson-Lognormal model

- Developped in Chiquet, Mariadassou, and Robin 2018
- Latent layer that is a Gaussian graphical model
- Poisson emission distribution
- Suitable for ecological count data (e.g. eDNA data)

PLN model

- $\mathbf{Z}$ latent layer, $\mathbf{Y}$ observation layer
- $\mathbf{Z} \sim \mathcal{N}(0, \mathbf{\Sigma})$
- $Y_j | Z_j \sim \mathcal{P}(\exp(\mu_j + Z_j)$

Markov networks
Multivariate Poisson-Lognormal model

- Developped in Chiquet, Mariadassou, and Robin 2018
- Latent layer that is a Gaussian graphical model
- Poisson emission distribution
- Suitable for ecological count data (e.g. eDNA data)

PLN model

- $\mathbf{Z}$ latent layer, $\mathbf{Y}$ observation layer
- $\mathbf{Z} \sim \mathcal{N}(0, \mathbf{\Sigma})$
- $Y_j | Z_j \sim \mathcal{P}(\exp(\mu_j + Z_j))$

Properties:

- $\mathbb{E}(Y_j) = \exp(\mu_j + \frac{\sigma_j^2}{2}) := \lambda_j$
- $\mathbb{V}(Y_j) = \lambda_j + \lambda_j^2(\exp(\sigma_j^2) - 1)$ (over-dispersion, variance is greater than mean)

Thank you for your attention,
Let’s do a R practical session now !

References I

Blanchet, F Guillaume, Kevin Cazelles, and Dominique Gravel (2020). "Co-occurrence is not evidence of ecological interactions". In: *Ecology Letters*.

Chiquet, Julien, Mahendra Mariadassou, and Stéphane Robin (2018). "Variational inference for sparse network reconstruction from count data". In: *arXiv preprint arXiv:1806.03120*.

Friedman, Jerome, Trevor Hastie, and Robert Tibshirani (2008). "Sparse inverse covariance estimation with the graphical lasso". In: *Biostatistics* 9.3, pp. 432–441.

Martinez-Almoyna, Camille et al. (2019). "Multi-trophic β-diversity mediates the effect of environmental gradients on the turnover of multiple ecosystem functions". In: *Functional Ecology* 33.10, pp. 2053–2064.

References II

Ohlmann, Marc et al. (2018). "Mapping the imprint of biotic interactions on β-diversity". In: *Ecology Letters* 21.11, pp. 1660–1669. DOI: 10.1111/ele.13143. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/ele.13143. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/ele.13143.

Ovaskainen, Otso and Nerea Abrego (2020). *Joint species distribution modelling: with applications in R*. Cambridge University Press.

Poggiato, Giovanni et al. (2021). "On the interpretations of joint modeling in community ecology". In: *Trends in Ecology & Evolution* 36.5, pp. 391–401.

Pollock, Laura J et al. (2014). "Understanding co-occurrence by modelling species simultaneously with a Joint Species Distribution Model (JSDM)". In: *Methods in Ecology and Evolution* 5.5, pp. 397–406.

Schwaller, Loıc (2015). "An introduction to graphical models". In.

References III

Yang, Eunho et al. (2014). “Mixed graphical models via exponential families”. In: *Artificial intelligence and statistics*. PMLR, pp. 1042–1050.