Network inference and Graphical models

S. Robin

Sorbonne université

CESAB & EcoNet, Apr. 2026

Network analysis: Two different situations

Interactions are observed.

- Interaction networks, contact networks
- Trophic networks, plant-pollinators networks

→ Understand the organization/functioning of the network

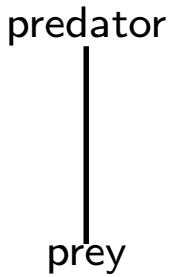
Interactions are not-observed.

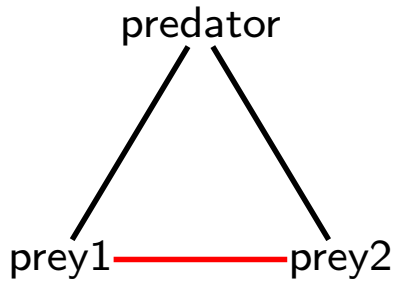
- Species abundance or presence/absence data (co-occurrence 'networks')
- Deep sequencing, metagenomics

→ Reconstruct the 'interaction' network

Species interaction network

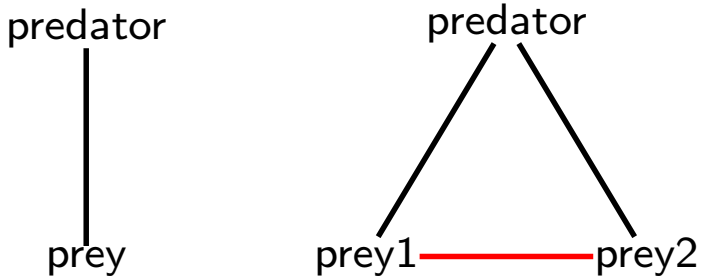
- Aim: Understand how species from a same community interact
- Network representation = draw an edge between interacting pairs of species





Species interaction network

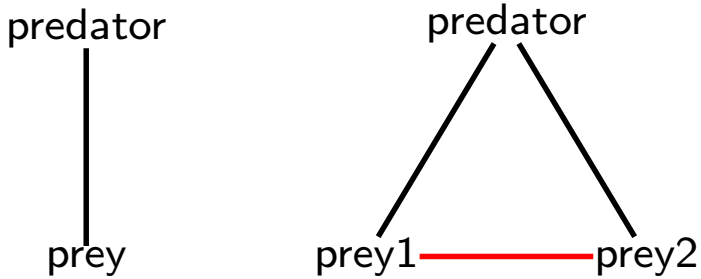
- Aim: Understand how species from a same community interact
- Network representation = draw an edge between interacting pairs of species



- Main issue: Distinguish direct interactions (predator-prey) from simple associations (= indirect interactions: two preys of a same predator)
- Co-occurences or correlations cannot distinguish between the two [PWT+19]

Species interaction network

- Aim: Understand how species from a same community interact
- Network representation = draw an edge between interacting pairs of species



- Main issue: Distinguish direct interactions (predator-prey) from simple associations (= indirect interactions: two preys of a same predator)
- Co-occurences or correlations cannot distinguish between the two [PWT+19]

Probabilistic translation.

$$\text{'network'} = \text{graph}$$
$$\text{'association'} = \text{marginal dependance}$$
$$\text{'direct interaction'} = \text{conditional dependance}$$

Network inference

Typical dataset at hand. n sites, p species,

- x_i = environmental description of site i, t_j = traits of species j
- Y_{ij} = abundance (or presence) of species j in site i

Network inference

Typical dataset at hand. n sites, p species,

- x_i = environmental description of site i, t_j = traits of species j
- Y_{ij} = abundance (or presence) of species j in site i

Assumption.

- Species 'direct' interactions are encoded in a network (= a graph) G.
- G is the same in all sites.

Network inference

Typical dataset at hand. n sites, p species,

- x_i = environmental description of site i, t_j = traits of species j
- Y_{ij} = abundance (or presence) of species j in site i

Assumption.

- Species 'direct' interactions are encoded in a network (= a graph) G.
- G is the same in all sites.

Aim. Retrieve the graph G from the species abundances Y :

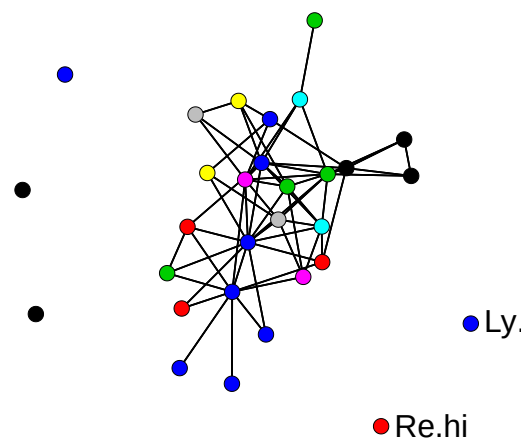
Abundances Y

Hi.pl	*An.lu*	*Me.ae*
31	0	108
4	0	110
27	0	788
13	0	295
23	0	13
20	0	97
⋮	⋮	⋮

Covariates X

Lat.	Long.	Depth
71.10	22.43	349
71.32	23.68	382
71.60	24.90	294
71.27	25.88	304
71.52	28.12	384
71.48	29.10	344
⋮	⋮	⋮

Inferred network G



Outline

Outline

Statistical model

Observed data. The adundance vector in site i:

$$y_i = [y_{i1} \dots y_{ij} \dots y_{im}]$$

is seen as a realisation of a random vector

$$Y = [Y_1 \dots Y_j \dots Y_m]$$

(For the time being, forget about the covariates: same x_i in all sites.)

Statistical model

Observed data. The adundance vector in site i:

$$y_i = [y_{i1} \dots y_{ij} \dots y_{im}]$$

is seen as a realisation of a random vector

$$Y = [Y_1 \dots Y_j \dots Y_m]$$

(For the time being, forget about the covariates: same x_i in all sites.)

Statistical model.

$$Y \sim p \quad = \text{joint distribution: } p(Y_1, \dots, Y_m)$$

Statistical model

Observed data. The adundance vector in site i:

$$y_i = [y_{i1} \dots y_{ij} \dots y_{im}]$$

is seen as a realisation of a random vector

$$Y = [Y_1 \dots Y_j \dots Y_m]$$

(For the time being, forget about the covariates: same x_i in all sites.)

Statistical model.

$$Y \sim p \quad = \text{joint distribution: } p(Y_1, \dots, Y_m)$$

Where is the network (graph)?

- The graph G has to be encoded in the distribution p in some way
- Graphical models establish a formal connexion between p and G

Graphical models

A generic framework [Lau96,WJ08]

Joint distribution $p = p(A, B, C, D, \dots)$

Graphical model $G =$ dependency structure between A, B, C, D, $\dots$

Graphical models

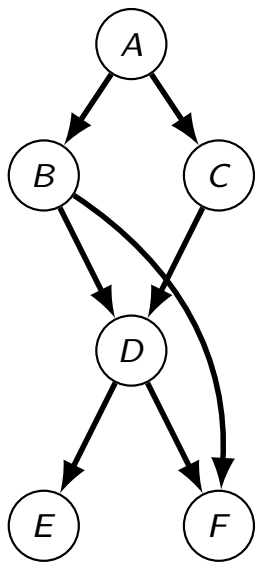
A generic framework [Lau96,WJ08]

Joint distribution $p = p(A, B, C, D, \dots)$

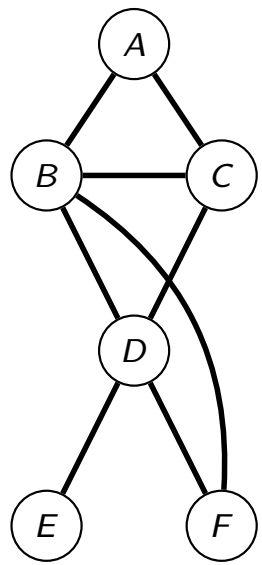
Graphical model $G =$ dependency structure between A, B, C, D, ...

Two kinds of graphical models.

Directed



Undirected



Outline

A reminder in probability

Two random variables A and B.

Joint distribution: $$p(a, b) = \Pr\{A = a, B = b\}$$

Marginal distribution: $$p(a) = \Pr\{A = a\} = \sum_{b \in \mathcal{B}} p(a, b)$$

Conditional distribution: $$p(b \mid a) = \Pr\{B = b \mid A = a\} = \frac{p(a, b)}{p(a)}$$

A reminder in probability

Two random variables A and B.

Joint distribution: $p(a, b) = \Pr\{A = a, B = b\}$

Marginal distribution: $p(a) = \Pr\{A = a\} = \sum_{b \in \mathcal{B}} p(a, b)$

Conditional distribution: $p(b \mid a) = \Pr\{B = b \mid A = a\} = \dfrac{p(a, b)}{p(a)}$

Factorisation. As a consequence:

$p(a, b) = p(a)\, p(b \mid a)$ 'A, then B given A' $A \to B$

$= p(b)\, p(a \mid b)$ 'B, then A given B' $B \to A$

Directed graphical models = Bayesian networks

Definition. Let D be a directed acyclic graph (DAG), the distribution p is said to factorize in D iff

$$p(a_1, \dots a_m) = \prod_{j=1}^{m} p(a_j \mid a_{pa_D(j)})$$

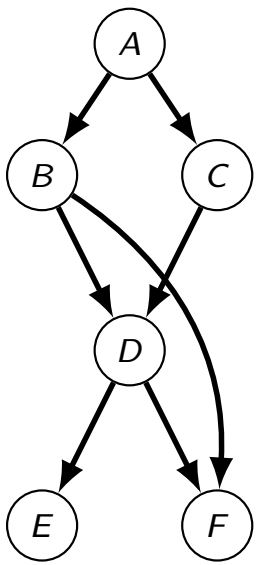
where $pa_D(i)$ stands for the set of parents of i in D.

Directed graphical models = Bayesian networks

Definition. Let D be a directed acyclic graph (DAG), the distribution p is said to factorize in D iff

$$p(a_1, \dots a_m) = \prod_{j=1}^{m} p(a_j \mid a_{pa_D(j)})$$

where $pa_D(i)$ stands for the set of parents of i in D.



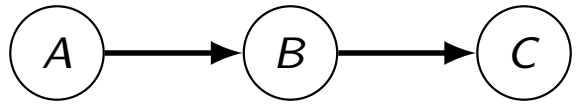
$$pa_D(A) = \emptyset, \qquad pa_D(D) = \{B, C\}$$

$$\begin{aligned} p(a, \dots f) \quad = \quad & p(a)\ p(b \mid a)\ p(c \mid a) \\ & p(d \mid b, c)\ p(e \mid d) \\ & p(f \mid b, d) \end{aligned}$$

See [SWA17] for an introduction in ecology

A simple (interesting) example

Consider $D =$

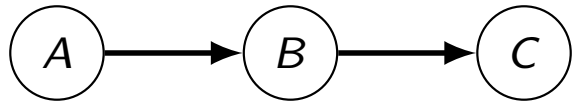


$p(a, b, c)$ is faithful to D iff

$$p(a, b, c) = p(a)\ p(b \mid a)\ p(c \mid a)$$

A simple (interesting) example

Consider $D =$



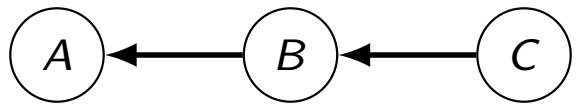
$p(a, b, c)$ is faithful to D iff

$$p(a, b, c) = p(a)\ p(b \mid a)\ p(c \mid a)$$

But

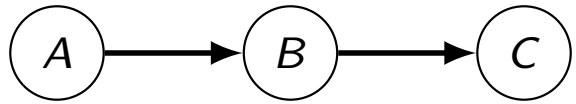
$$p(a)\ p(b \mid a)\ p(c \mid b) = p(a \mid b)\ p(b \mid c)\ p(c)$$

so p is also faithful to $D' =$



A simple (interesting) example

Consider $D =$



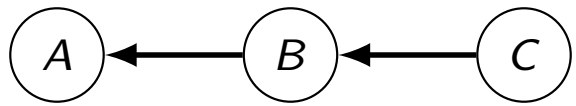
$p(a, b, c)$ is faithful to D iff

$$p(a, b, c) = p(a)\ p(b \mid a)\ p(c \mid a)$$

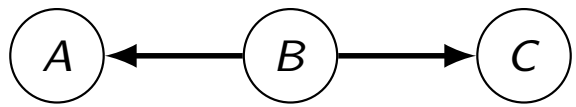
But

$$p(a)\ p(b \mid a)\ p(c \mid b) = p(a \mid b)\ p(b \mid c)\ p(c)$$

so p is also faithful to $D' =$

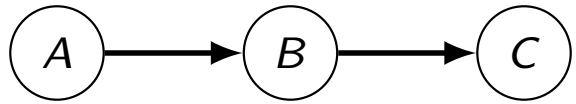


and to $D'' =$



A simple (interesting) example

Consider $D =$



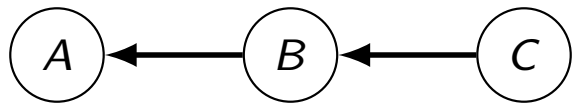
$p(a, b, c)$ is faithful to D iff

$$p(a, b, c) = p(a)\ p(b \mid a)\ p(c \mid a)$$

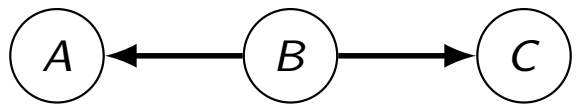
But

$$p(a)\ p(b \mid a)\ p(c \mid b) = p(a \mid b)\ p(b \mid c)\ p(c)$$

so p is also faithful to $D' =$



and to $D'' =$



Conclusion: The directed graphical model is most often not unique.

- $p(a, b, c, \dots)$ is not enough to retrieve the edge orientations (see appendix)
- No causal interpretation (causality not addressed here, see [Pea09a,Pea09b])

Outline

Undirected graphical models = Markov random fields

Definition. Let G be an *undirected graph*, the distribution p is said to factorize in G iff

$$p(a_1, \dots a_m) \propto \prod_{C \in \mathcal{C}(G)} \psi_C(a_C).$$

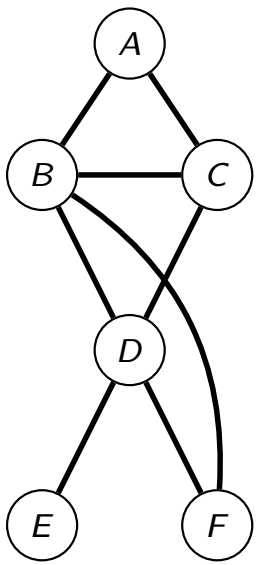
where $\mathcal{C}(G)$ is the set of maximal cliques of G

Undirected graphical models = Markov random fields

Definition. Let G be an *undirected graph*, the distribution p is said to factorize in G iff

$$p(a_1, \dots a_m) \propto \prod_{C \in \mathcal{C}(G)} \psi_C(a_C).$$

where $\mathcal{C}(G)$ is the set of maximal cliques of G



$$\begin{aligned} p(a, \dots f) \quad \propto \quad & \psi_1(a, b, c) \\ & \times \psi_2(b, c, d) \\ & \times \psi_3(b, d, f) \\ & \times \psi_4(d, e) \end{aligned}$$

Conditional independence

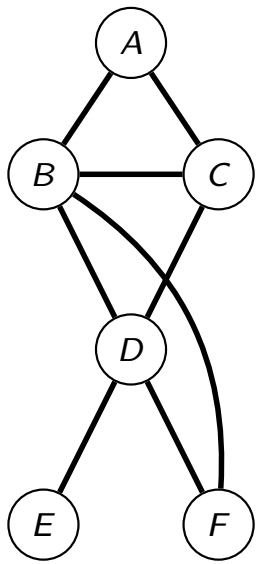
Property of undirected graphical models. If $p(a, b, c, \dots) > 0$,

$$\text{separation} \quad \Leftrightarrow \quad \text{conditional independence}$$

Conditional independence

Property of undirected graphical models. If $p(a, b, c, \dots) > 0$,

$$\text{separation} \quad \Leftrightarrow \quad \text{conditional independence}$$



- A and F are dependent
- A and F are dependent given B
- A and F are independent given $\{B, C\}$
- $\{A, B, C\}$ is independent from E given D
- A and C are dependent given all other variables
  A and C are 'conditionally dependent'

- Inferring G = Inferring conditional dependencies
- Task: retrieve G from the distribution p in an automatic way

Outline

Gaussian graphical models (GGM)

Gaussian setting. Suppose that the 'abundance' vector $Y = (Y_1 \dots Y_m)$ has a multivariate Gaussian distribution

$$Y \sim \mathcal{N}_m(\mu, \Sigma)$$

- $\mu = (m \times 1)$ vector of mean 'abundances'
- $\Sigma = (m \times m)$ covariance matrix:

$$\sigma_{jj} = \sigma_j^2 = \mathbb{V}(Y_j), \qquad \sigma_{jk} = \mathbb{C}\text{ov}(Y_j, Y_k),$$

- Correlation:

$$\rho_{jk} = \sigma_{jk} \big/ (\sigma_j \, \sigma_k)$$

Gaussian graphical models (GGM)

Gaussian setting. Suppose that the 'abundance' vector $Y = (Y_1 \dots Y_m)$ has a multivariate Gaussian distribution

$$Y \sim \mathcal{N}_m(\mu, \Sigma)$$

- $\mu = (m \times 1)$ vector of mean 'abundances'
- $\Sigma = (m \times m)$ covariance matrix:

$$\sigma_{jj} = \sigma_j^2 = \mathbb{V}(Y_j), \qquad \sigma_{jk} = \mathbb{C}\text{ov}(Y_j, Y_k),$$

- Correlation:

$$\rho_{jk} = \sigma_{jk} \big/ (\sigma_j \, \sigma_k)$$

A first nice property. For species j and k,

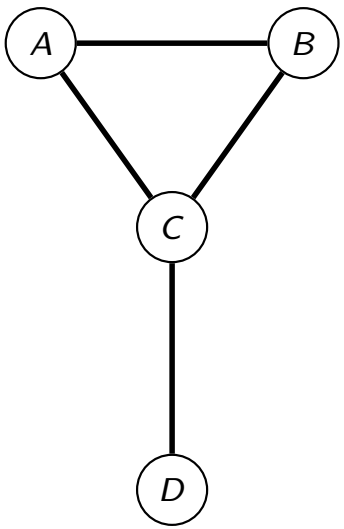
$$\text{null correlation} \quad \Leftrightarrow \quad \text{null covariance} \quad \Leftrightarrow \quad \text{independence},$$

$$\text{i.e.} \qquad \rho_{jk} = 0 \qquad \Leftrightarrow \qquad \sigma_{jk} = 0 \qquad \Leftrightarrow \qquad Y_j \perp\!\!\!\perp Y_k.$$

But we are interested in conditional dependency.

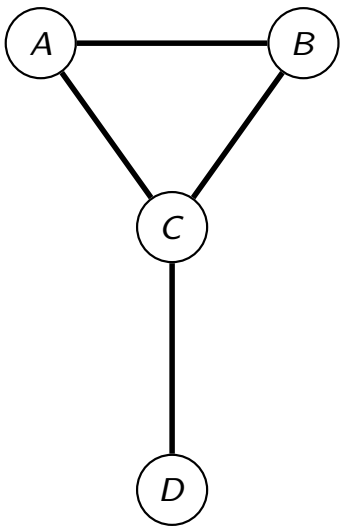
A nice property of GGM's

Graphical model.



A nice property of GGM's

Graphical model.



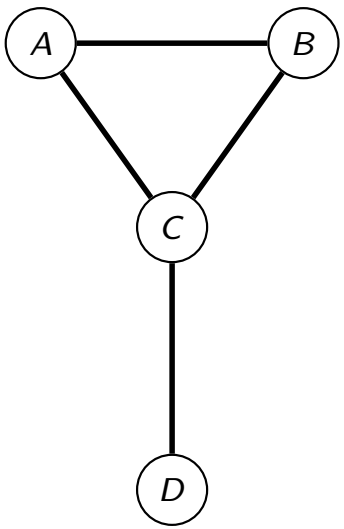
Adjacency matrix.

$$G = \begin{bmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix}$$

- Connected
- C separates D from $\{A, B\}$

A nice property of GGM's

Graphical model.



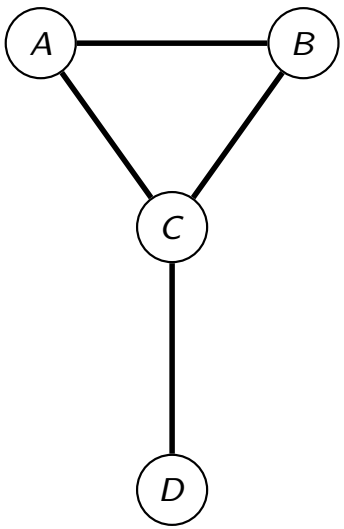
Covariance matrix.

$$\Sigma \propto \begin{bmatrix} 1 & -.25 & -.41 & .25 \\ -.25 & 1 & -.41 & .25 \\ -.41 & -.41 & 1 & -.61 \\ .25 & .25 & -.61 & 1 \end{bmatrix}$$

- No zero because G is Connected

A nice property of GGM's

Graphical model.



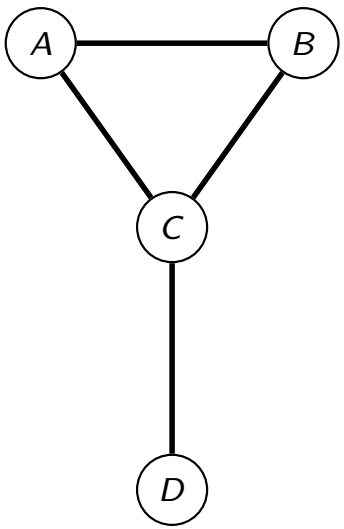
Inverse covariance matrix.

$$\Omega = \Sigma^{-1} \propto \begin{bmatrix} 1 & .5 & .5 & 0 \\ .5 & 1 & .5 & 0 \\ .5 & .5 & 1 & .5 \\ 0 & 0 & .5 & 1 \end{bmatrix}$$

- 0's at (1, 4) and (2, 4)
- Conditional dependencies are encoded in Ω = precision matrix
- (see appendix for a formal explanation)

A nice property of GGM's

Graphical model.



Estimated inverse covariance matrix.

$$\widehat{\Sigma}^{-1} \propto \begin{bmatrix} 1 & .48 & .61 & .09 \\ .48 & 1 & .67 & .06 \\ .61 & .67 & 1 & .46 \\ .09 & .06 & .46 & 1 \end{bmatrix}$$

$(n = 100)$

- No 'true zero' in the estimate
- Need to 'force' zeros to appear

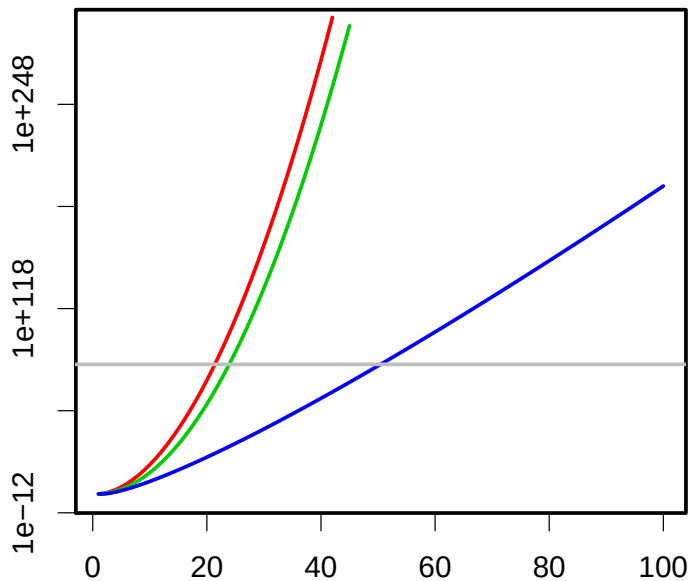
Network inference

Aim. Given a series of iid 'abundance' vectors $Y_1, \dots Y_n$, find the graph G that best fits the distribution of the Y_i's.

Network inference

Aim. Given a series of iid 'abundance' vectors $Y_1, \dots Y_n$, find the graph G that best fits the distribution of the Y_i's.

A huge search space:



Number of directed acyclic graphs (no close form), undirected graphs ($2^{m(m-1)/2}$), spanning trees (m^{m-2}). x-axis: m = number of species.

- Need for efficient algorithms / heuristics

Making Ω sparse

Graphical lasso [FHT08]. Penalized likelihood:

$$\max_{\mu,\Omega} \quad \underbrace{\log p(Y; \mu, \Omega)}_{\text{log-likelihood}} - \underbrace{\lambda \sum_{j<k} |\omega_{jk}|}_{\text{penalization}}$$

the ℓ_1 penalty ('lasso') forces some ω_{jk} to be zero.

- The higher λ, the sparser the network
- Turns out to be a convex problem, so fast solutions exists (`huge` R package)

Making Ω sparse

Graphical lasso [FHT08]. Penalized likelihood:

$$\max_{\mu,\Omega} \quad \underbrace{\log p(Y;\mu,\Omega)}_{\text{log-likelihood}} - \underbrace{\lambda \sum_{j<k} |\omega_{jk}|}_{\text{penalization}}$$

the ℓ_1 penalty ('lasso') forces some ω_{jk} to be zero.

- The higher λ, the sparser the network
- Turns out to be a convex problem, so fast solutions exists (`huge` R package)

Alternatives.

- Regression viewpoint (see appendix)
- Mixture of tree-shaped distributions: [MJ06,Kir07,JFS+16,SR17,MRA19]
- Other forms of penalties (ℓ_1, ℓ_2, combinations) induces other topologies
  Ex.: [ACM09] induce SBM structured networks.

Outline

Back to ecology

Specificities.

- Mostly count or presence/absence data, which do not fit a Gaussian distribution
- Environmental effects need to be accounted for

Back to ecology

Specificities.

- Mostly count or presence/absence data, which do not fit a Gaussian distribution
- Environmental effects need to be accounted for

Joint species distribution models (JSDM).

- n sites, m species
- $Y_i = [Y_{i1}, \dots Y_{im}]$ abundance vector observed in site i
- x_i = vector of environmental covariates for site i + species traits t_j
- JSDM: multivariate distribution

$$p(Y_i \mid x_i)$$

- See [WBO+15] for a general introduction to JSDM or [OA20] for some R tools

Remark. Univariate of JSDM = SDM $\simeq$ Generalized linear models (glm / glmm).

Network inference with JSDM

Problem.

- Not many flexible *multivariate* distributions for count or binary data do exist [IYAR17]

Network inference with JSDM

Problem.

- Not many flexible *multivariate* distributions for count or binary data do exist [IYAR17]

A common trick: latent variable models

- Most popular structure = latent GGM:
  Spiec-Easi [KMM+15], gCODA [FHZD17], MiNT [BML+16], PLNnetwork [CMR18b], tree-based PLN [MRA19]

- PLN = Poisson log-normal model (see next)

Network inference with JSDM

Problem.

- Not many flexible *multivariate* distributions for count or binary data do exist [IYAR17]

A common trick: latent variable models

- Most popular structure = latent GGM:
  Spiec-Easi [KMM+15], gCODA [FHZD17], MiNT [BML+16], PLNnetwork [CMR18b], tree-based PLN [MRA19]

- PLN = Poisson log-normal model (see next)

But also:

- Copula-based approaches [AdVPM19]

Outline

Poisson log-normal model

Data: n independent sites (no spatial structure), m species,

$$Y_{ij} = \text{ abundance of species } j \text{ in site } i$$

Poisson log-normal model

Data: n independent sites (no spatial structure), m species,

$$Y_{ij} = \text{ abundance of species } j \text{ in site } i$$

Poisson log-normal (PLN) model [AH89,CMR21]:

- A Gaussian vector is associated with each site i

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

Poisson log-normal model

Data: n independent sites (no spatial structure), m species,

$$Y_{ij} = \text{ abundance of species } j \text{ in site } i$$

Poisson log-normal (PLN) model [AH89,CMR21]:

- A Gaussian vector is associated with each site i

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

- Species abundances in site i are independent given Z_i

$$Y_{ij} \sim \mathcal{P}(\exp(x_i^\top \beta_j + Z_{ij}))$$

Poisson log-normal model

Data: n independent sites (no spatial structure), m species,

$$Y_{ij} = \text{ abundance of species } j \text{ in site } i$$

Poisson log-normal (PLN) model [AH89,CMR21]:

- A Gaussian vector is associated with each site i

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

- Species abundances in site i are independent given Z_i

$$Y_{ij} \sim \mathcal{P}(\exp(x_i^\top \beta_j + Z_{ij}))$$

summarized as $\{Y_i\}_{1 \leqslant i \leqslant n}$ independent

$$Y_i \sim PLN(x_i, \beta, \Sigma),$$

$\rightarrow$ multivariate mixed generalized linear model (glmm).

Poisson log-normal model

Data: n independent sites (no spatial structure), m species,

$$Y_{ij} = \text{ abundance of species } j \text{ in site } i$$

Poisson log-normal (PLN) model [AH89,CMR21]:

- A Gaussian vector is associated with each site i

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

- Species abundances in site i are independent given Z_i

$$Y_{ij} \sim \mathcal{P}(\exp(x_i^\top \beta_j + Z_{ij}))$$

summarized as $\{Y_i\}_{1 \leqslant i \leqslant n}$ independent

$$Y_i \sim PLN(x_i, \beta, \Sigma),$$

$\rightarrow$ multivariate mixed generalized linear model (glmm).

Interpretation.

- $x_i^\top \beta_j$ mean (log-)abundance of species j: abiotic effects
- Σ = dependency structure (encoded in the latent layer): biotic interactions

Network inference

PLN network model. Same model as PLN + sparsity assumption:

$$\{Y_i\} \text{ independent}, \qquad Y_i \sim PLN(x_i, \beta, \Sigma), \qquad +\Omega = \Sigma^{-1} \text{ sparse}$$

Network inference

PLN network model. Same model as PLN + sparsity assumption:

$$\{Y_i\} \text{ independent}, \qquad Y_i \sim PLN(x_i, \beta, \Sigma), \qquad +\Omega = \Sigma^{-1} \text{ sparse}$$

Inference algorithm. Variational EM + lasso penalty [CMR18a,CMR19]:

$$\max_{\beta,\Omega} \widetilde{\log p}(Y; x, \beta, \Sigma) - \lambda \sum_{j<k} |\omega_{jk}|$$

- Alternate convex problems
  $\rightarrow$ Fast solution (`PLNmodels` R package)
- 'Atutomatic' choice of λ: BIC, EBIC, cross-validation, ...
- Resampling (StARS [LRW10]) can be used to assess robustness

Outline

Illustration

Barents fish dataset:

- $n = 89$ stations,
- $m = 30$ fish species,
- $d = 4$ covariates (latitude, longitude, depth, temperature)

Application

see Rmarkdown

Outline

Missing actors

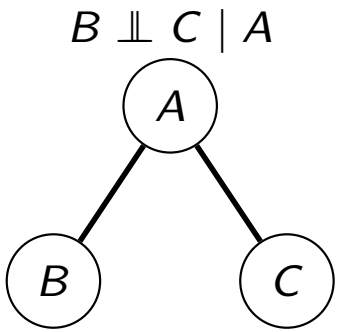
Incomplete observations. Most of the time, not all 'actors' (species, environmental covariates, ...) are observed

Missing actors

Incomplete observations. Most of the time, not all 'actors' (species, environmental covariates, ...) are observed

Missing variable = marginalisation.

- Remove the missing node and all edges connected to it.
- Connect all its neighbors with each other.

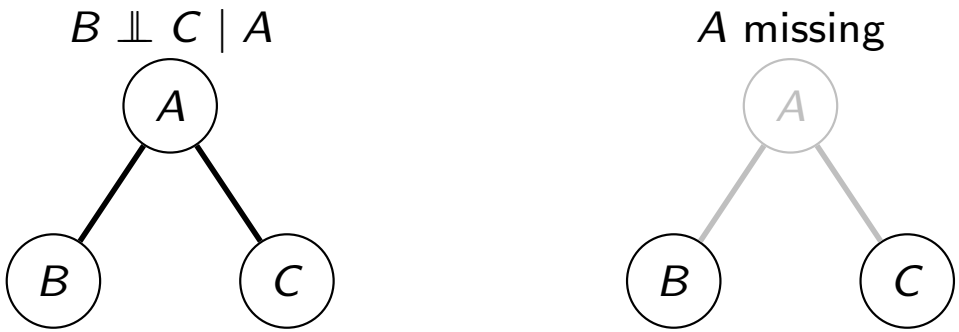


Missing actors

Incomplete observations. Most of the time, not all 'actors' (species, environmental covariates, ...) are observed

Missing variable = marginalisation.

- Remove the missing node and all edges connected to it.
- Connect all its neighbors with each other.

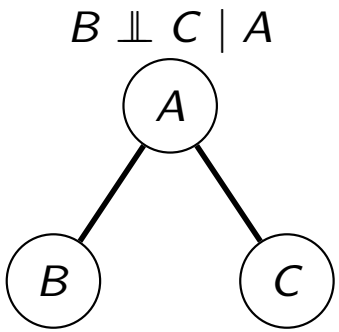


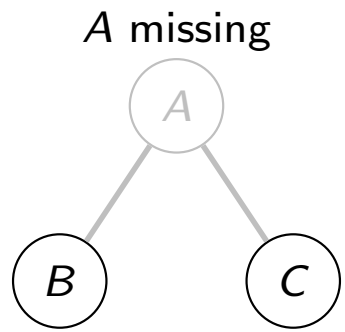
Missing actors

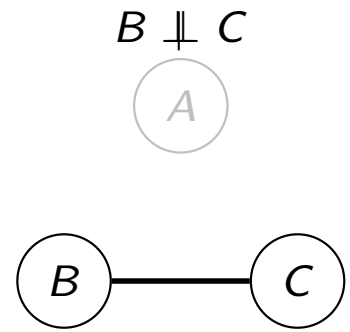
Incomplete observations. Most of the time, not all 'actors' (species, environmental covariates, ...) are observed

Missing variable = marginalisation.

- Remove the missing node and all edges connected to it.
- Connect all its neighbors with each other.





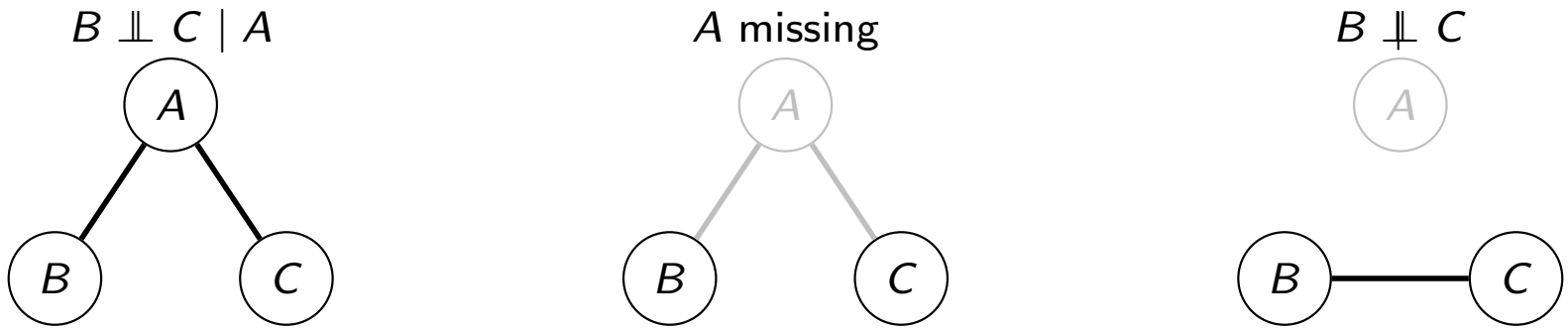


Missing actors

Incomplete observations. Most of the time, not all 'actors' (species, environmental covariates, ...) are observed

Missing variable = marginalisation.

- Remove the missing node and all edges connected to it.
- Connect all its neighbors with each other.



Indeed:

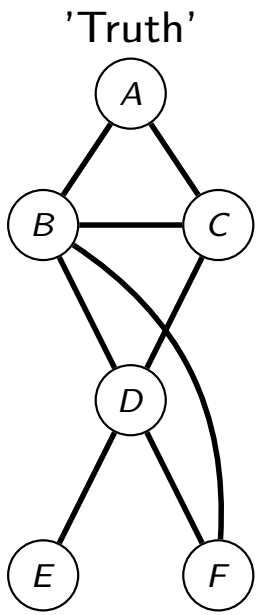
$$p(a, b, c) = p(a)p(b \mid a)p(c \mid a)$$

but A is not observed, so only $p(b, c)$ can be considered:

$$p(b, c) = \sum_a p(a, b, c) = \sum_a p(a)p(b \mid a)p(c \mid a) \neq p(b)p(c)$$

'Spurious' edges

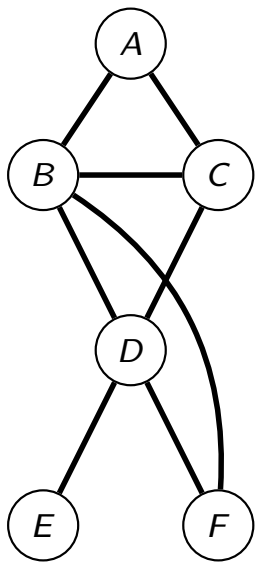
Possibly dramatic effect on the observable dependency structure:



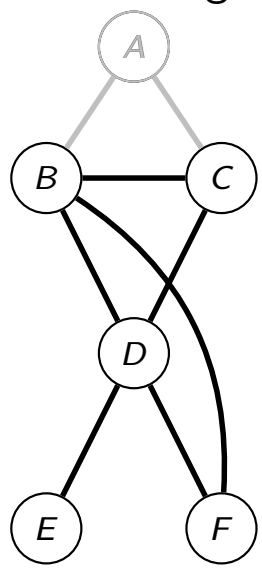
'Spurious' edges

Possibly dramatic effect on the observable dependency structure:

'Truth'



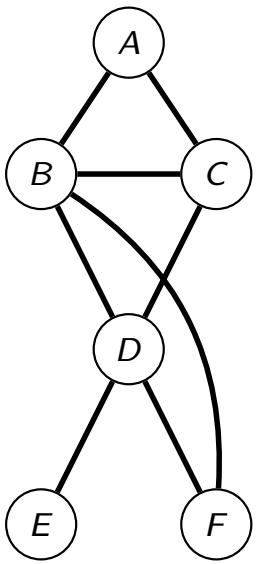
A missing



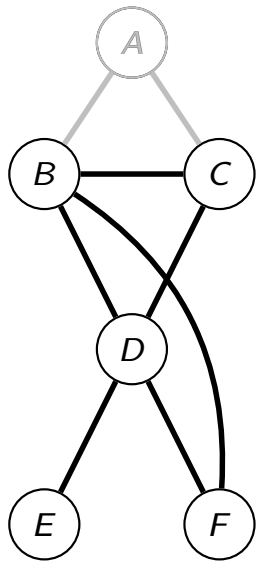
'Spurious' edges

Possibly dramatic effect on the observable dependency structure:

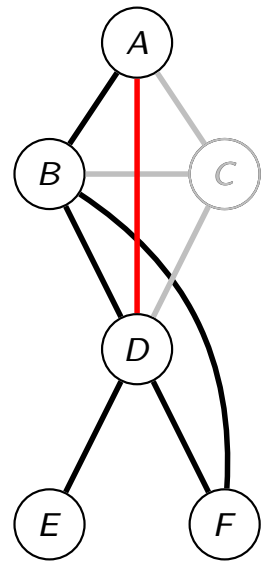
'Truth'



A missing

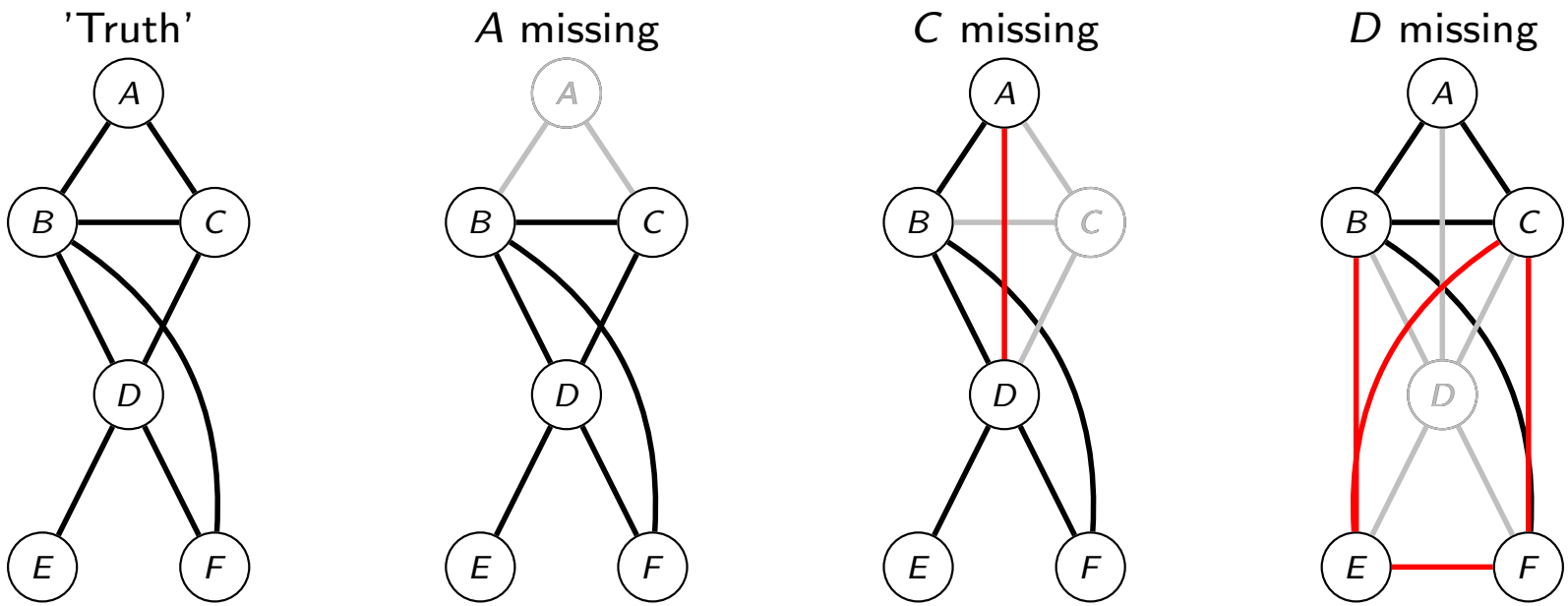


C missing



'Spurious' edges

Possibly dramatic effect on the observable dependency structure:



→ Need to account for 'all' available information to avoid 'spurious' edges

Illustration

Barents fish dataset:

- $n = 89$ stations,
- $m = 30$ fish species,
- not accounting for the environmental covariates

Application

see Rmarkdown

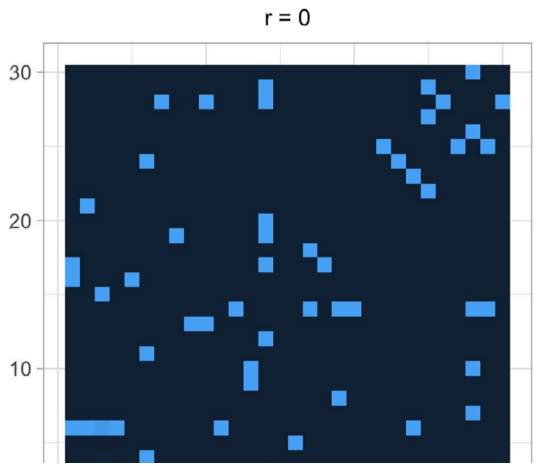
An attempt to retrieve missing actors

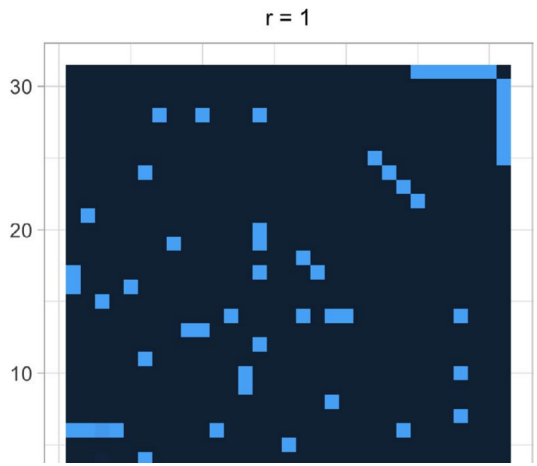
Barents' fishes. Based on tree-shaped network inference

Infered network [MRA20]

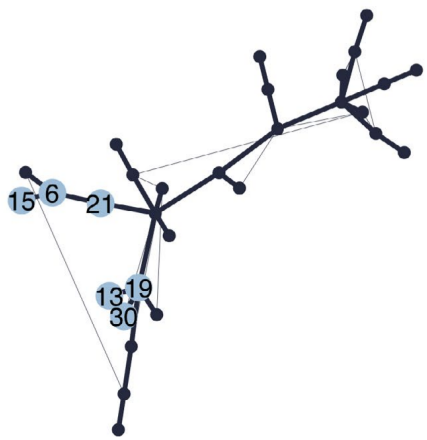
Infered missing actor [MRA21]

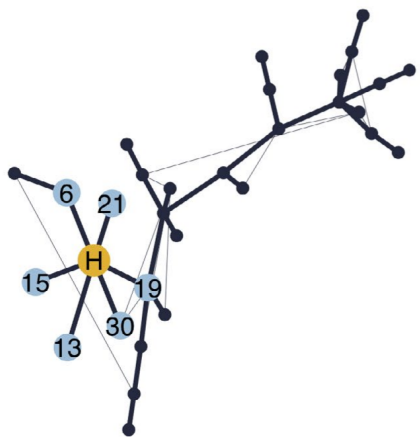
Adjacency matrix





Network





Outline

Temporal data

Data. Y_t = abundance vector at time t

$$Y_t = (A_t, B_t, C_t, D_t) = \text{ abundance vector at time } t$$

$\rightarrow$ Same population observed along time

Temporal data

Data. Y_t = abundance vector at time t

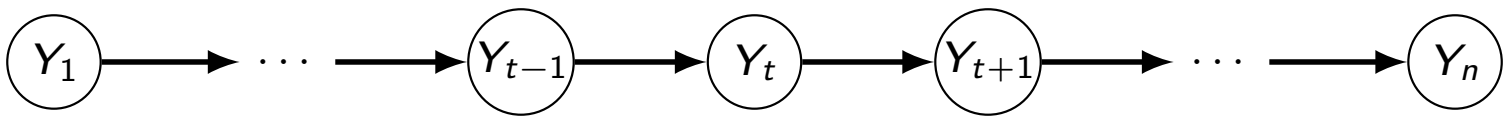
$$Y_t = (A_t, B_t, C_t, D_t) = \text{ abundance vector at time } t$$

→ Same population observed along time

A general model. Markov assumption:

$$p(Y) = p(Y_1)p(Y_2 \mid Y_1) \dots p(Y_n \mid Y_{n-1})$$

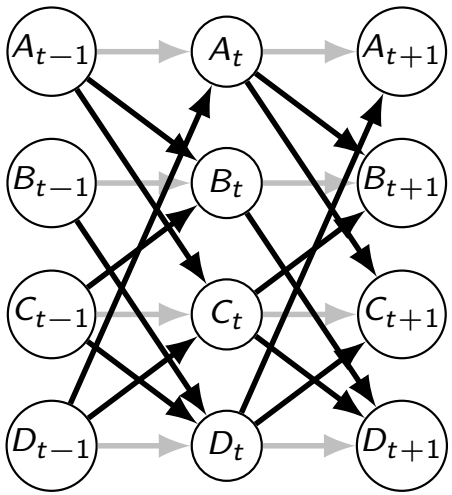
that is



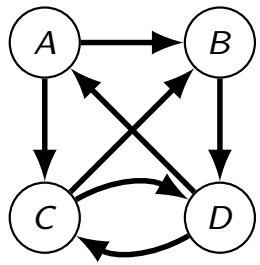
- Obviously oriented

A nice case: Dynamic 'Bayesian' networks (DBN)

Genuine graphical model: a DAG:



'Dynamic' network: not a DAG



Inference.

- Reconstruction problem: Find the parents of each species $p(A_{t+1} \mid A_t, B_t, C_t, \dots)$, independently
- *Sparsity* assumption: Each species has only few parents
  → Variable selection for regression

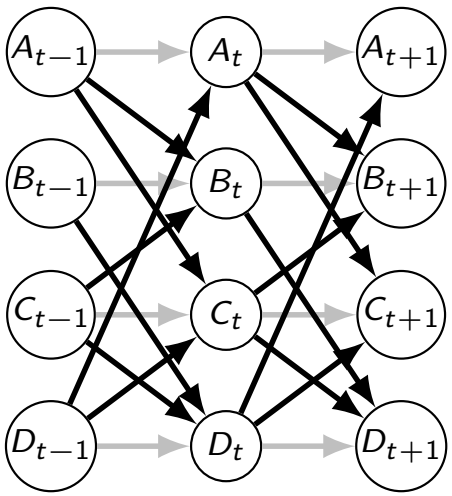
A common misunderstanding

A common setting. Dynamic model (e.g. Lotka-Volterra) but static data

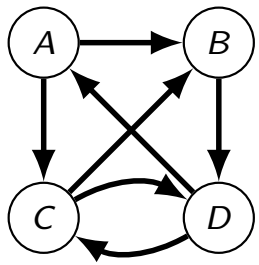
A common misunderstanding

A common setting. Dynamic model (e.g. Lotka-Volterra) but static data

Dynamic model



'Dynamic' network



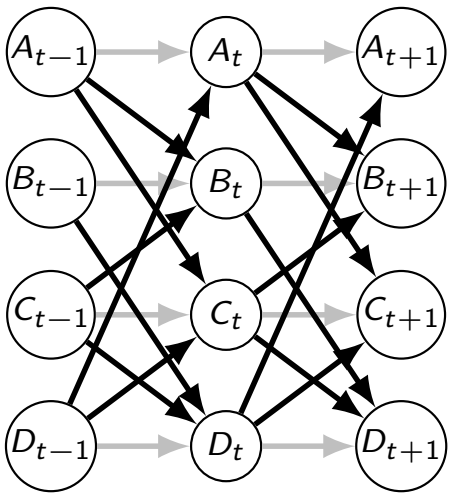
Data at hand

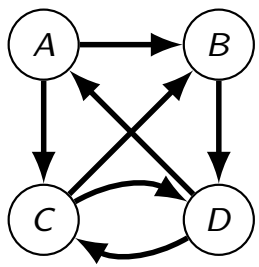
- n similar sites
- Each assumed to have reach a stationary regime (T large)
- n abundance vectors Y_T^i.

A common misunderstanding

A common setting. Dynamic model (e.g. Lotka-Volterra) but static data

Dynamic model 'Dynamic' network Data at hand





- n similar sites
- Each assumed to have reach a stationary regime (T large)
- n abundance vectors Y_T^i.

- The graphical model of Y_T is not the dynamic network
  $\rightarrow$ network inference applied to the 'stationary' abundance vectors (Y_T^i) will not retrieve the dynamic network
- Not clear if the dynamic network can be recovered from stationary (static) observations [Liu25].

Outline

Edge prediction

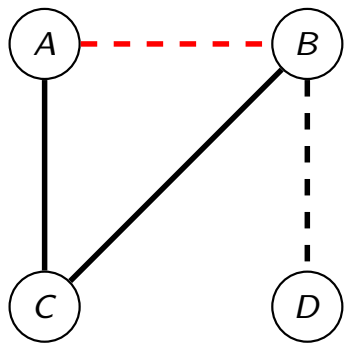
Context. Network are often only partially observed

Edge prediction

Context. Network are often only partially observed

Problem. Based on

- Traits t_a and t_b for species A and B,
- Other known interactions : $A \leftrightarrow C$, $A \nleftrightarrow D$, $B \leftrightarrow C$, $C \nleftrightarrow D$
- Similarities (phylogeny) between species
- ...



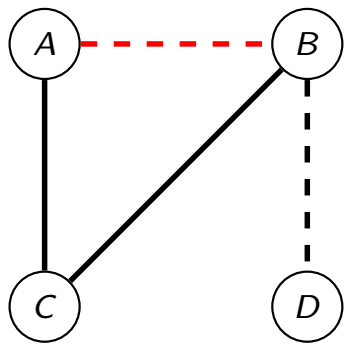
Can we predict if $A \leftrightarrow B$?

Edge prediction

Context. Network are often only partially observed

Problem. Based on

- Traits t_a and t_b for species A and B,
- Other known interactions : $A \leftrightarrow C$, $A \nleftrightarrow D$, $B \leftrightarrow C$, $C \nleftrightarrow D$
- Similarities (phylogeny) between species
- ...



Can we predict if $A \leftrightarrow B$?

A 0/1 prediction problem.

- Hundreds of (ML-like) prediction methods
- But heterogeneous structured predictors

Conclusion (?)

Summary.

- Statistical methods for infering species interaction networks do exist.
- They all rely on the notion of graphical model, most on Gaussian graphical models.
- The problem is not an easy task.
- The stability of the results needs to be assessed.

Conclusion (?)

Summary.

- Statistical methods for infering species interaction networks do exist.
- They all rely on the notion of graphical model, most on Gaussian graphical models.
- The problem is not an easy task.
- The stability of the results needs to be assessed.

What's next.

- Still many very partially solved problems (see 'Extensions').
- Experimental validation of the inferred edges is not an easy task.
- The inferred network is not an observed network
- Uncertainty of the inferred interactions needs to be accounted in downstream analyses

References I

C. Ambroise, J. Chiquet, and C. Matias. Inferring sparse gaussian graphical models with latent structure. *Electron. J. Statist.*, 3:205–38, 2009.

M. J Anderson, P. de Valpine, A. Punnett, and A. E Miller. A pathway for multivariate analysis of ecological communities using copulas. *Ecology and evolution*, 9(6):3276–3294, 2019.

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

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

J. Chiquet, M. Mariadassou, and S. Robin. Variational inference for probabilistic Poisson PCA. *The Annals of Applied Statistics*, 12(4):2674–2698, 2018.

J. Chiquet, M. Mariadassou, and S. Robin. Variational inference for sparse network reconstruction from count data. Technical Report 1806.03120, arXiv, 2018. *accepted in ICML 2019*.

J. Chiquet, M. Mariadassou, and S. Robin. Variational inference for sparse network reconstruction from count data. In *International Conference on Machine Learning*, pages 1162–1171, 2019.

J. Chiquet, M. Mariadassou, and S. Robin. The Poisson-lognormal model as a versatile framework for the joint analysis of species abundances. *Frontiers in Ecology and Evolution*, 9:188, 2021.

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

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

D. Inouye, E. Yang, G. I Allen, and P. Ravikumar. A review of multivariate distributions for count data derived from the Poisson distribution. *Computational Statistics*, 9(3), 2017.

References II

B. Jakuschkin, V. Fievet, L. Schwaller, T. Fort, C. Robin, and C. Vacher. Deciphering the pathobiome: Intra-and interkingdom interactions involving the pathogen Erysiphe alphitoides. *Microbial ecology*, pages 1–11, 2016.

S. Kirshner. Learning with tree-averaged densities and distributions. In *NIPS*, pages 761–768, 2007.

Z. D. Kurtz, C. L. Müller, E. R. Miraldi, D. R. Littman, M. J. Blaser, and R. A. Bonneau. Sparse and compositionally robust inference of microbial ecological networks. *PLoS computational biology*, 11(5):e1004226, 2015.

S. L. Lauritzen. *Graphical Models*. Oxford Statistical Science Series. Clarendon Press, 1996.

Bixuan Liu. Identifiability of var (1) model in a stationary setting. Technical Report 2504.03466, arXiv, 2025.

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

N. Meinshausen and P. Bühlmann. High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, pages 1436–1462, 2006.

M. Meilă and T. Jaakkola. Tractable Bayesian learning of tree belief networks. *Statistics and Computing*, March 2006.

R. Momal, S. Robin, and A. Ambroise. Tree-based reconstruction of ecological network from abundance data. Technical Report 1905.02452, arXiv, 2019.

R. Momal, S. Robin, and C. Ambroise. Tree-based inference of species interaction networks from abundance data. *Methods in Ecology and Evolution*, 11(5):621–632, 2020.

R. Momal, S. Robin, and C. Ambroise. Accounting for missing actors in interaction network inference from abundance data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 2021.

O. Ovaskainen and N. Abrego. *Joint species distribution modelling: With applications in R*. Cambridge University Press, 2020.

J. Pearl. Causal inference in statistics: An overview. *Statistics surveys*, 3:96–146, 2009.

References III

J. Pearl. *Causality*. Cambridge university press, 2009.

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

L. Schwaller. *Exact Bayesian Inference in Graphical Models : Tree-structured Network Inference and Segmentation*. Theses, Université Paris-Saclay, September 2016.

L. Schwaller and S. Robin. Exact Bayesian inference for off-line change-point detection in tree-structured graphical models. *Statistics and Computing*, 27(5):1331–1345, 2017.

E.L. Sander, J.T. Wootton, and S. Allesina. Ecological network inference from long-term presence-absence data. *Scientific reports*, 7(1):7154, 2017.

N. Verzelen. Minimax risks for sparse regressions: Ultra-high-dimensional phenomenons. *Electron. J. Stat.*, 6:38–90, 2012.

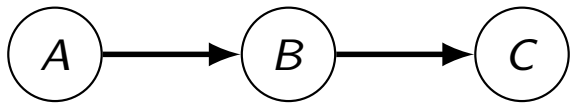
N. Verzelen and F. Villers. Tests for gaussian graphical models. *Computational Statistics & Data Analysis*, 53(5):1894 – 1905, 2009.

D. I. Warton, F. G. Blanchet, R. B. O'Hara, O. Ovaskainen, S. Taskinen, S. C Walker, and F. KC. Hui. So many variables: joint modeling in community ecology. *Trends in Ecology & Evolution*, 30(12):766–779, 2015.

M. J. Wainwright and M. I. Jordan. Graphical models, exponential families, and variational inference. *Found. Trends Mach. Learn.*, 1(1–2):1–305, 2008.

A simple (interesting) example

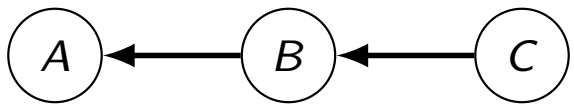
p faithful to $D =$



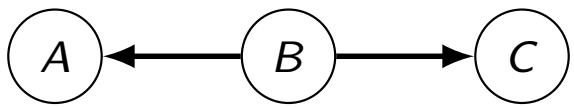
means that

$$p(a, b, c) = p(a)\ p(b \mid a)\ p(c \mid a)$$

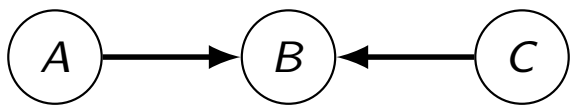
p is also faithful to $D' =$



and to $D'' =$



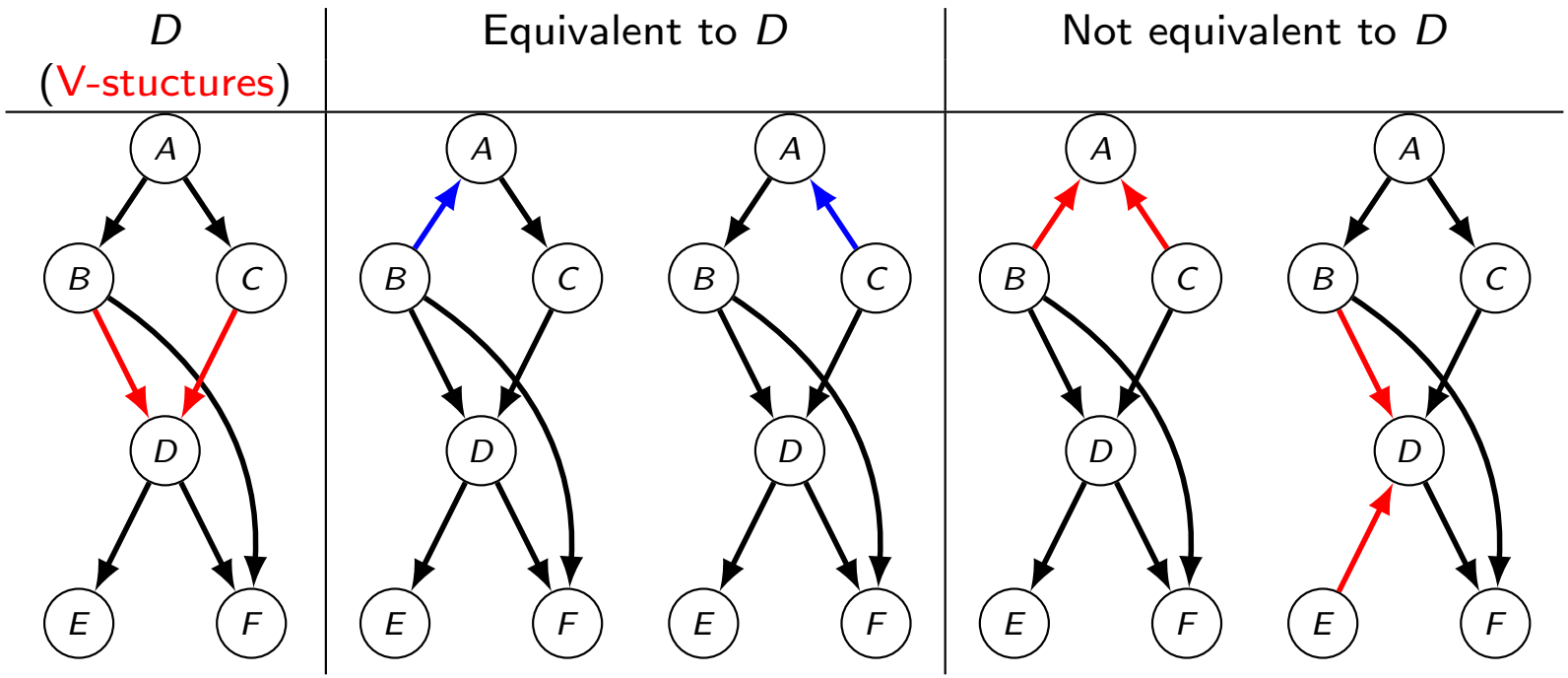
but not to $D''' =$



Interpretability?

Theorem [Pea09b]. Two DAGs are Markov equivalent (i.e. induce the same conditional dependences and independences) if they have the same skeleton (i.e. the same undirected edges) and the same V-structures.

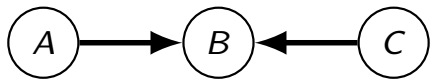
Example.



(Not speaking of possibly missing nodes)

V-structure

In the V-structured (or 'head to head') DAG:



A and C are conditionally dependent ($A \not\perp\!\!\!\perp C \mid B$):

$$p(a, b, c) = p(a)p(c)p(b \mid a, b)$$

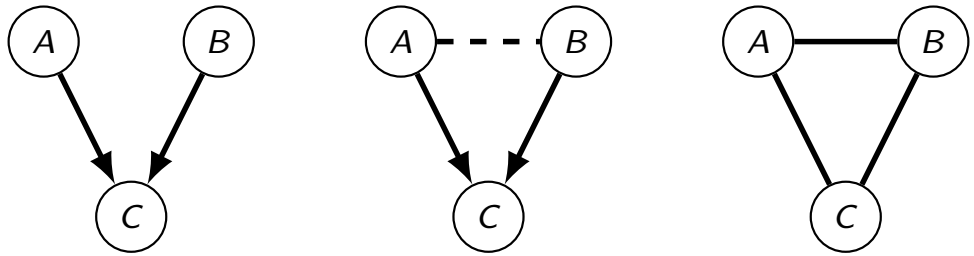
$$\Rightarrow \quad p(a, b \mid b) = \frac{p(a, b, c)}{p(b)} = \frac{p(a)\ p(c)\ p(b \mid a, c)}{p(b)}$$

Remark. A and C are marginally independent:

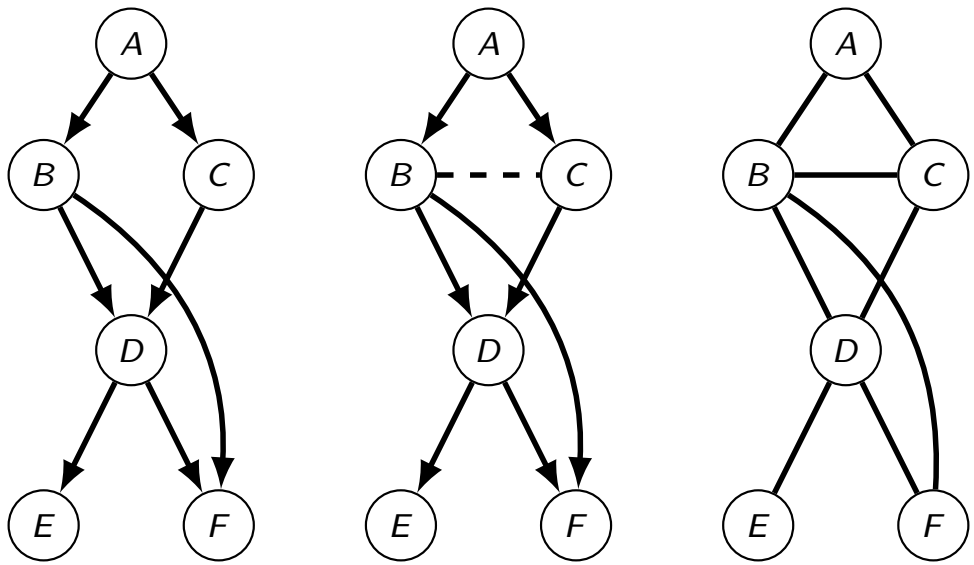
$$p(a, c) = \sum_b p(a)\ p(c)\ p(b \mid a, c) = p(a)\ p(c) \underbrace{\sum_b p(b \mid a, c)}_{=1}$$

From directed to undirected graphical models

Resolve *immoralities*. Graph *moralization* ('parents must be married'):



Example.



Regression point of view

Similarly to temporal networks, finding the neighbors of a node is equivalent to find its 'parents' in a regression:

$$Y_{ik} = \sum_{j \neq k} \beta_{jk} Y_{ij} + E_{ik}.$$

→ Requires a testing procedure [VV09] or a penalization [MB06] to determine non-zero coefficients

Remarks.

- Regression needs reconciliation when $\widehat{\beta}_{jk} = 0$ and $\widehat{\beta}_{jk} \neq 0$
- [Ver12] provides bounds for the recovery of the list of neighbors: let k = degree of a given node

$k \log(p/k) \leqslant n$	possible recovery
$k \log(p/k) > n \log n$	impossible recovery

GGM's nice property: More formally

If $Y \sim \mathcal{N}(0, \Sigma)$, then

$$
\begin{aligned}
p(Y) &\propto \exp\left(-\frac{1}{2}\|Y\|^2_{\Sigma^{-1}}\right) \\
&= \exp\left(-\frac{1}{2}\sum_{j,k}\omega_{jk}\, Y_j\, Y_k\right) \qquad \text{where } \Omega = [\omega_{jk}] = \Sigma^{-1} \\
&= \prod_{j,k} \underbrace{\exp\left(-\frac{1}{2}\omega_{jk}\, Y_j\, Y_k\right)}_{\phi_{jk}(Y_j, Y_k)}
\end{aligned}
$$

- The non-zeros of Ω correspond to the edges of G
- Furthermore:

$$-\omega_{jk} \propto \rho\left(Y_j, Y_k \mid Y_{\{j,k\}}\right) = \text{ 'partial' correlation}$$

Regression point of view

Similarly to temporal networks, finding the neighbors of a node is equivalent to find its 'parents' in a regression:

$$Y_{ik} = \sum_{j \neq k} \beta_{jk} Y_{ij} + E_{ik}.$$

→ Requires a testing procedure [VV09] or a penalization [MB06] to determine non-zero coefficients

Remarks.

- Regression needs reconciliation when $\widehat{\beta}_{jk} = 0$ and $\widehat{\beta}_{jk} \neq 0$
- [Ver12] provides bounds for the recovery of the list of neighbors: let k = degree of a given node

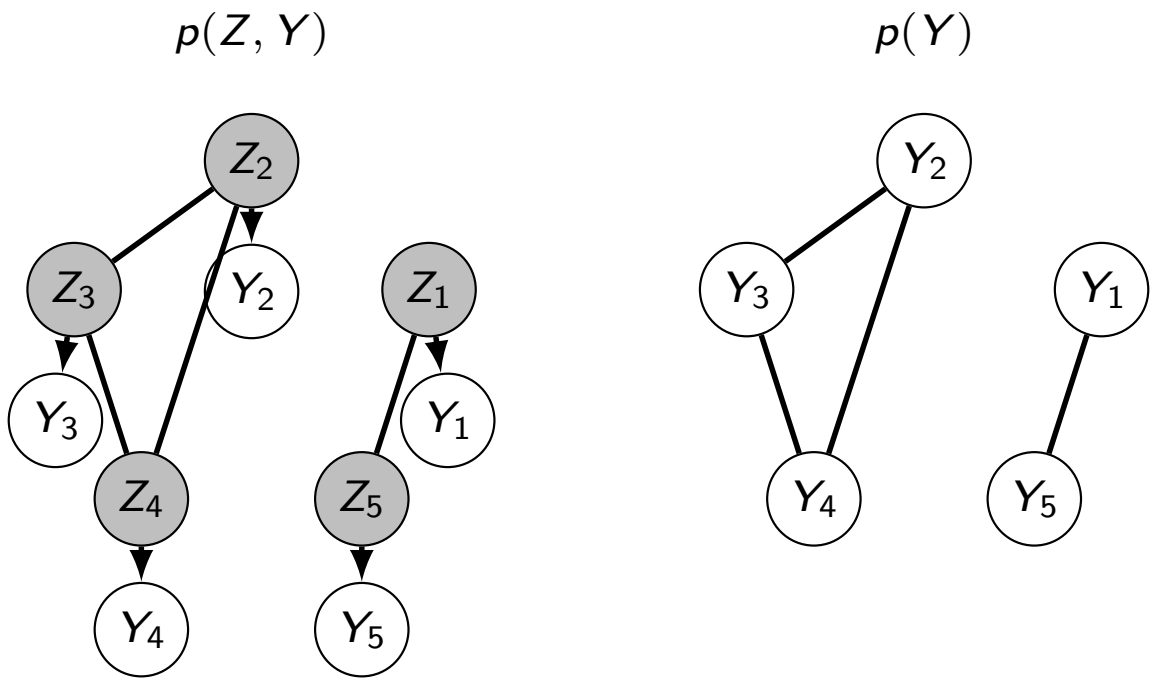
$k \log(p/k) \leqslant n$	possible recovery
$k \log(p/k) > n \log n$	impossible recovery

Network inference under the PLN model

All latent (GGM) models infer the dependency struture of the latent Z, not of the observed abundances Y

Network inference under the PLN model

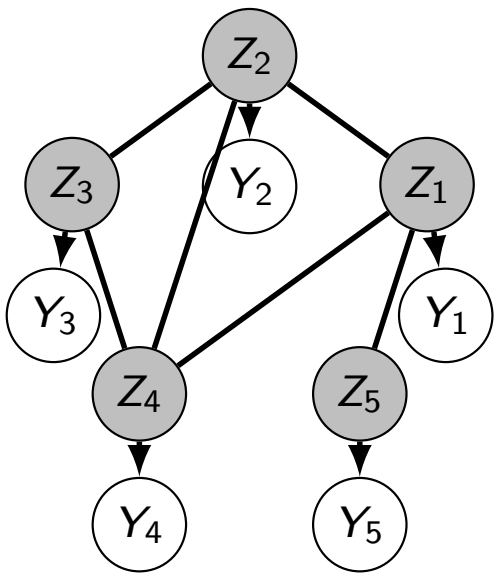
All latent (GGM) models infer the dependency struture of the latent Z, not of the observed abundances Y



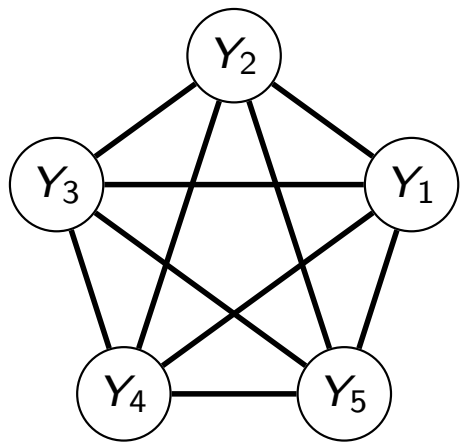
Network inference under the PLN model

All latent (GGM) models infer the dependency struture of the latent Z, not of the observed abundances Y

$p(Z, Y)$



$p(Y)$



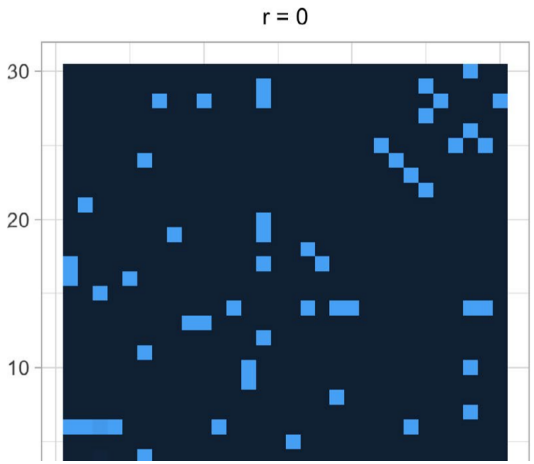
Missing actors

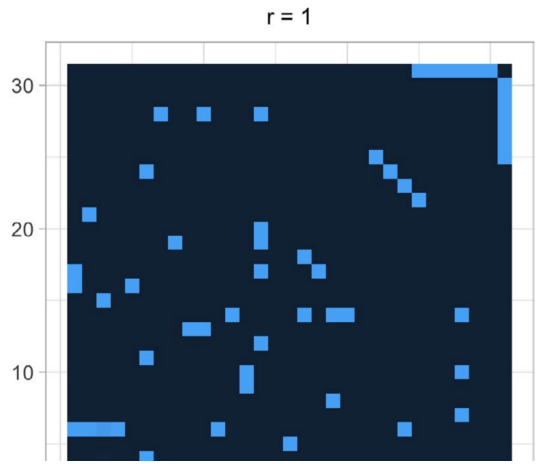
Barents'sea dataset. Based on tree-shaped network inference

Infered network [MRA20]

Infered missing actor [MRA21]

Adjacency matrix





Network