Statistical network inference

UMR MIA-Paris: S. Ouadah, S. Robin, ...

INRA / AgroParisTech

Learn Biocontrol, AgroParisTech, November 2016

Outline

Statistics for networks

Network inference

Gaussian graphical models

Tree-based Bayesian inference

Extensions

Concluding remarks & questions

Outline

Statistics for networks

Network inference

Gaussian graphical models

Tree-based Bayesian inference

Extensions

Concluding remarks & questions

Network: convenient way to describe

connexions / relations / interactions → links $i \sim j$

between

individuals / genes / species / entities → nodes i

Network: convenient way to describe

connexions / relations / interactions → links $i \sim j$

between

individuals / genes / species / entities → nodes i

'Network' = graph G:

$$G = (V, E)$$

$V = \{1, \dots p\}$ = set of nodes, E = set of edges.

Network: convenient way to describe

connexions / relations / interactions → links $i \sim j$

between

individuals / genes / species / entities → nodes i

'Network' = graph G:

$$G = (V, E)$$

$V = \{1, \dots p\}$ = set of nodes, E = set of edges.

Alternatively: Adjacency matrix

$$A = [A_{ij}] : \qquad A_{ij} = \begin{cases} 1 & \text{if } i \sim j, \\ 0 & \text{otherwise.} \end{cases}$$

Different questions

Understanding the network topology:

- Data = observed network
- Questions: central nodes? cluster structure? small-world property? ...

Different questions

Understanding the network topology:

- Data = observed network
- Questions: central nodes? cluster structure? small-world property? ...

Inferring the network:

- Data = repeated signal observed at each node
- Questions: which nodes are connected?

Different questions

Understanding the network topology:

- Data = observed network
- Questions: central nodes? cluster structure? small-world property? ...

Inferring the network:

- Data = repeated signal observed at each node
- Questions: which nodes are connected?

Using the network:

- Data = a given network + signal on nodes
- Questions: how the epidemic spreads along the network?

Different questions

Understanding the network topology:

- Data = observed network
- Questions: central nodes? cluster structure? small-world property? ...

Inferring the network:

- Data = repeated signal observed at each node
- Questions: which nodes are connected?

Using the network:

- Data = a given network + signal on nodes
- Questions: how the epidemic spreads along the network?

Each to be combined with covariates, time, missing data, ...

Outline

Statistics for networks

Network inference

Gaussian graphical models

Tree-based Bayesian inference

Extensions

Concluding remarks & questions

A brief review

Data. p species ($i = 1..p$), n replicates ($r = 1..n$)

$$\begin{aligned} Y_{ir} &= \text{abundance of species } i \text{ in replicate } r \\ Y_r = (Y_{ir})_{i=1..p} &= \text{vector of abundances in replicate } r \end{aligned}$$

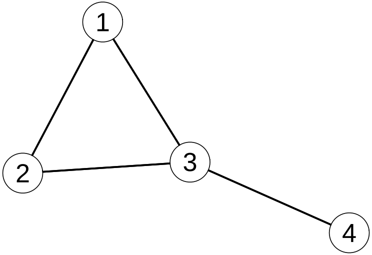
Goal. Infer the species interaction network from the set of Y_r? [15]

Remarks.

- Need to specify the type of interaction
- Need to distinguish between direct and indirect interactions

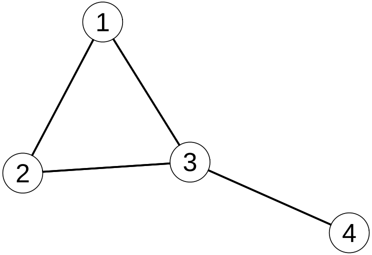
General framework: Graphical models [6]

Example: (undirected graph)



General framework: Graphical models [6]

Example: (undirected graph)

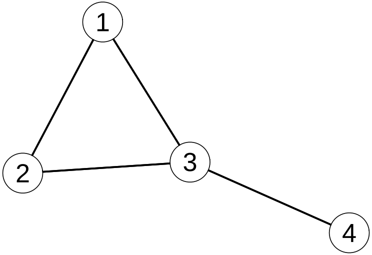


Properties:

All variables are dependent (connected graph).

General framework: Graphical models [6]

Example: (undirected graph)



Properties:

All variables are dependent (connected graph).

Some are conditionally independent, e.g.

$$Y_4 \perp (Y_1, Y_2) | Y_3$$

Graphical models

More precisely. $P =$ joint distribution faithful to G iff

$$P(Y_1, \dots, Y_p) \propto \prod_{C \in \mathcal{C}_G} \psi_C(Y_C)$$

where $\mathcal{C}_G =$ set of cliques of G. [4]

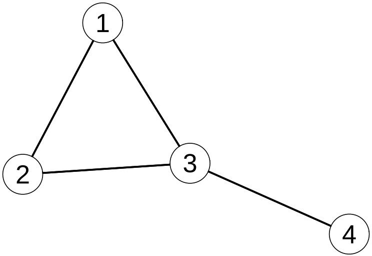
Graphical models

More precisely. $P =$ joint distribution faithful to G iff

$$P(Y_1, \dots, Y_p) \propto \prod_{C \in \mathcal{C}_G} \psi_C(Y_C)$$

where $\mathcal{C}_G =$ set of cliques of G. [4]

Example.



$$P(Y_1, Y_2, Y_3, Y_4) \propto$$

$$\psi_1(Y_1, Y_2, Y_3) \times \psi_2(Y_3, Y_4)$$

Network inference

Data. $Y_{ir} =$ (say) abundance of species i in replicate r.

Network inference

Data. Y_{ir} = (say) abundance of species i in replicate r.

Generic statistical model.

$$(Y_r)_{r=1..n} \text{ iid } \sim P, \qquad P \text{ is faithful to } G$$

Network inference

Data. Y_{ir} = (say) abundance of species i in replicate r.

Generic statistical model.

$$(Y_r)_{r=1..n} \text{ iid } \sim P, \qquad P \text{ is faithful to } G$$

Network inference problem.

Based on the dataset $Y = (Y_{ir})$, infer G.

Network inference

Data. Y_{ir} = (say) abundance of species i in replicate r.

Generic statistical model.

$$(Y_r)_{r=1..n} \text{ iid } \sim P, \qquad P \text{ is faithful to } G$$

Network inference problem.

Based on the dataset $Y = (Y_{ir})$, infer G.

Critical issue. There are $2^{p(p-1)/2}$ possible graphs

p	5	10	20	50
# graphs	10^3	10^{14}	10^{57}	10^{369}

Outline

Statistics for networks

Network inference

Gaussian graphical models

Tree-based Bayesian inference

Extensions

Concluding remarks & questions

GGM = Gaussian graphical models

Gaussian distribution.

$$Y_r \sim \mathcal{N}_p(\mu, \Sigma)$$

μ = vector of means, Σ = covariance matrix.

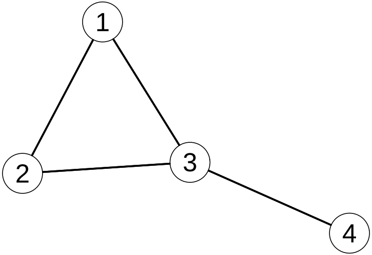
GGM = Gaussian graphical models

Gaussian distribution.

$$Y_r \sim \mathcal{N}_p(\mu, \Sigma)$$

μ = vector of means, Σ = covariance matrix.

A nice property.



Adjacency matrix

$$A = \left[\begin{array}{cccc} 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{array}\right]$$

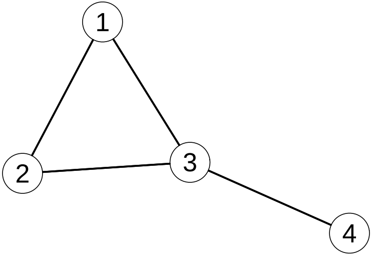
GGM = Gaussian graphical models

Gaussian distribution.

$$Y_r \sim \mathcal{N}_p(\mu, \Sigma)$$

μ = vector of means, Σ = covariance matrix.

A nice property.



Covariance matrix

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

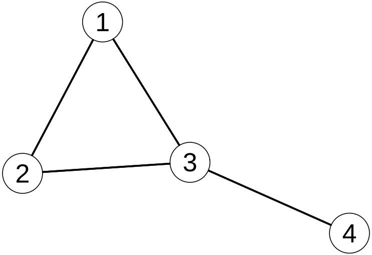
GGM = Gaussian graphical models

Gaussian distribution.

$$Y_r \sim \mathcal{N}_p(\mu, \Sigma)$$

μ = vector of means, Σ = covariance matrix.

A nice property.



Inverse covariance matrix

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

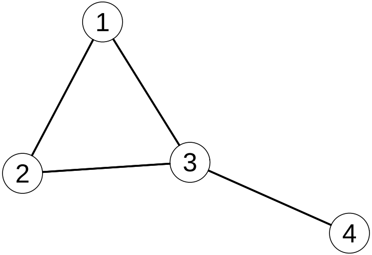
GGM = Gaussian graphical models

Gaussian distribution.

$$Y_r \sim \mathcal{N}_p(\mu, \Sigma)$$

$\mu =$ vector of means, $\Sigma =$ covariance matrix.

A nice property.



Estimated inverse covariance matrix

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

$(n = 100)$

Sparsity

Sparsity assumption:

$$\Omega = \Sigma^{-1} \text{ is sparse}$$

$=$ most entries of Ω are zeros.

Sparsity

Sparsity assumption:

$$\Omega = \Sigma^{-1} \text{ is sparse}$$

$=$ most entries of Ω are zeros.

A series of approaches for Gaussian data

- Sparse regression of each species on the others:

$$Y_i = a_j + \sum_{j \neq i} b_{ij} Y_j + E_i \quad \text{forcing most } b_{ij} = 0 \text{ [8]}$$

- Directly estimate Ω forcing most $\omega_{ij} = 0$ [3]

using, e.g, Lasso penalty [14,2] or more refined [1].

Outline

Statistics for networks

Network inference

Gaussian graphical models

Tree-based Bayesian inference

Extensions

Concluding remarks & questions

Inference of tree-shaped network [7,13]

Same problem. Infer G based on an iid sample $(Y_r) \sim P$ faithful to G.

Inference of tree-shaped network [7,13]

Same problem. Infer G based on an iid sample $(Y_r) \sim P$ faithful to G.

Tree assumption: The network G is a spanning tree, i.e.

$$\left.\begin{array}{l} G \text{ is connected} \\ G \text{ has no cycle} \end{array}\right\} \rightarrow G \text{ has } p-1 \text{ edges (sparse)}$$

Inference of tree-shaped network [7,13]

Same problem. Infer G based on an iid sample $(Y_r) \sim P$ faithful to G.

Tree assumption: The network G is a spanning tree, i.e.

$$\left.\begin{array}{l} G \text{ is connected} \\ G \text{ has no cycle} \end{array}\right\} \rightarrow G \text{ has } p-1 \text{ edges (sparse)}$$

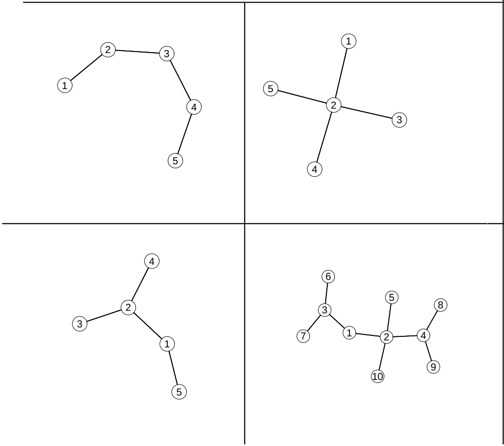
Bayesian inference: aim at providing

- the global posterior distribution of G given the data Y: $P(G|Y)$
- or edge probabilities:

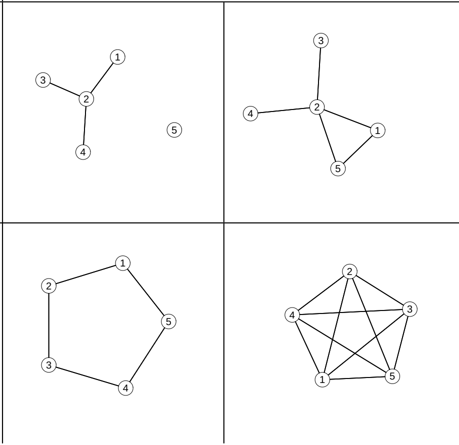
$$P(i \sim j|Y)$$

Tree-shaped network

Spanning trees



Non spanning trees



Exact Bayesian inference

Bayesian inference requires to sum over all possible graphs
$\rightarrow$ sum over all possible spanning trees ($\#\mathcal{T} = p^{p-2}$).

Exact Bayesian inference

Bayesian inference requires to sum over all possible graphs
$\rightarrow$ sum over all possible spanning trees ($\#\mathcal{T} = p^{p-2}$).

An algebraic tool. w_{ij} = weight of (i, j)

- Score of a tree = product of the weights of its branches

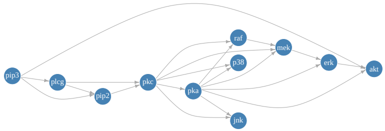
$$s(T) = \prod_{(i,j) \in T} w_{ij}$$

- Matrix-tree theorem:

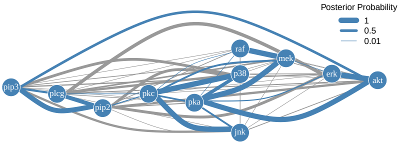
$$\sum_{T \in \mathcal{T}} s(T) \text{ is computable in } O(p^3)$$

Illustration: Raf pathway [13]

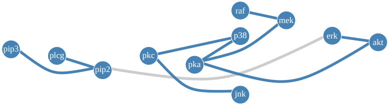
Flow cytometry data for $p = 11$ proteins from the Raf signaling pathway [11]



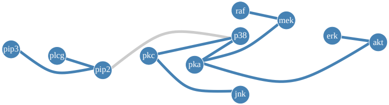
'ground truth'



posterior probabilities



most likely tree



second most likely tree

Outline

Statistics for networks

Network inference

Gaussian graphical models

Tree-based Bayesian inference

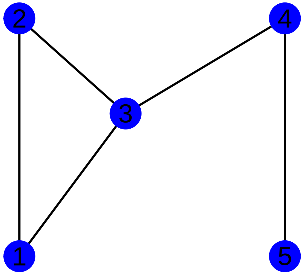
Extensions

Concluding remarks & questions

Removing nodes

Complete network (all nodes).

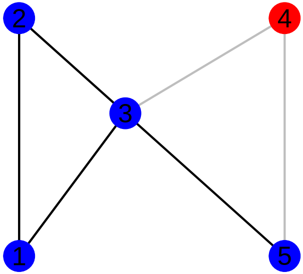
Graphical model



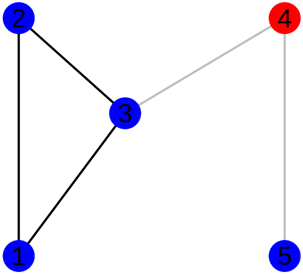
Removing nodes

Removing one node.

Marginal graph (missing node)



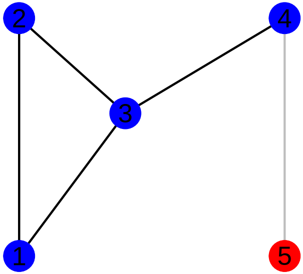
Conditional graph (observed node)



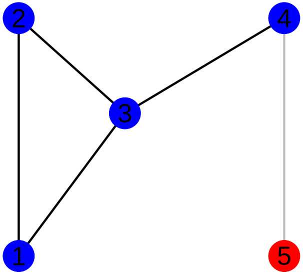
Removing nodes

Removing one node.

Marginal graph (missing node)



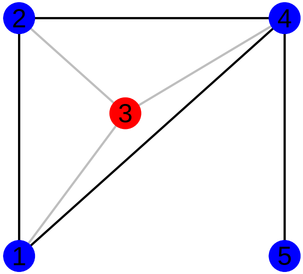
Conditional graph (observed node)



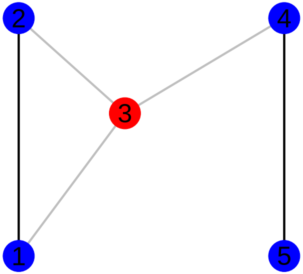
Removing nodes

Removing one node.

Marginal graph (missing node)



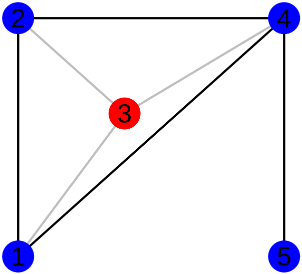
Conditional graph (observed node)



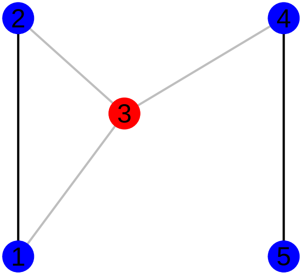
Removing nodes

Removing one node.

Marginal graph (missing node)



Conditional graph (observed node)



More complex patterns for directed graphical models.

Accounting for covariates (1/3)

Deciphering the pathobiome: Intra-and interkingdom interactions involving the pathogen *E. alphitoides*. [5]

Accounting for covariates (1/3)

Deciphering the pathobiome: Intra-and interkingdom interactions involving the pathogen *E. alphitoides*. [5]

Data.

- 3 trees × few tens of leaves per tree
- Abundance of few tens of (fungal and bacterial) species on each leaves via NGS
- Few covariates describing both trees and leaves

Accounting for covariates (1/3)

Deciphering the pathobiome: Intra-and interkingdom interactions involving the pathogen *E. alphitoides*. [5]

Data.

- 3 trees × few tens of leaves per tree
- Abundance of few tens of (fungal and bacterial) species on each leaves via NGS
- Few covariates describing both trees and leaves

Questions.

- Infer the ecological network
- Account for covariates (to avoid spurious edges)
- Deal with NGS counts as (Y_{ir}) measurements

Accounting for covariates (2/3)

Proposed strategy.

1. Perform regression on the covariates x for each species i:

$$Y_{ir} \sim \mathcal{P}(\mu_{ir}), \qquad \log \mu_{ir} = x_r \beta_i \qquad \rightarrow \qquad \widehat{\mu}_{ir} = e^{x_r \widehat{\beta}_i}$$

2. Compute the Pearson residuals $\widetilde{Y}_{ir} = (Y_{ir} - \widehat{\mu}_{ir}) / \sqrt{\widehat{\mu}_{ir}}$
3. Infer the network from the corrected abundances $(\widetilde{Y}_{ir})$ using [13].

Accounting for covariates (2/3)

Proposed strategy.

1. Perform regression on the covariates x for each species i:

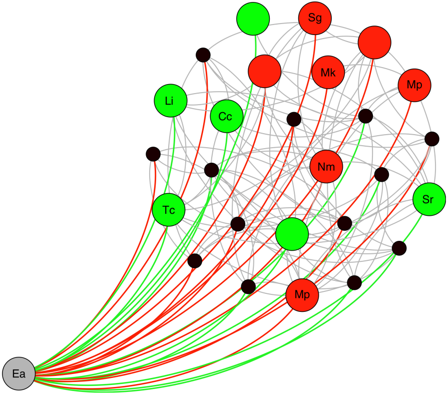
$$Y_{ir} \sim \mathcal{P}(\mu_{ir}), \qquad \log \mu_{ir} = x_r \beta_i \qquad \rightarrow \qquad \widehat{\mu}_{ir} = e^{x_r \widehat{\beta}_i}$$

2. Compute the Pearson residuals $\widetilde{Y}_{ir} = (Y_{ir} - \widehat{\mu}_{ir})/\sqrt{\widehat{\mu}_{ir}}$
3. Infer the network from the corrected abundances $(\widetilde{Y}_{ir})$ using [13].

Still unsolved problem.

- How to account for the uncertainty of the $\widehat{\beta}_i$ in the inference of G?

Accounting for covariates (3/3)



●: *E. alphitoides*

🔴: positively correlated fungal OTUs

🟢: negatively correlated fungal OTUs

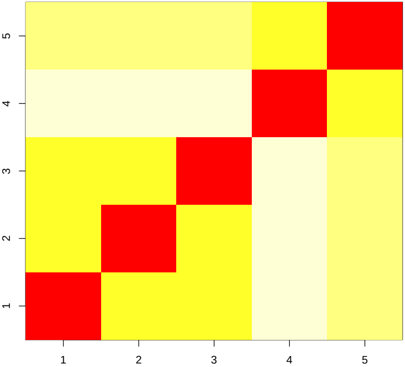
●: bacterial OTUs

Missing nodes (1/3)

Fact. Block-structured empirical correlation matrix.

→ Each block could be associated with an unobserved node.

→ Can we infer such missing nodes?



Missing nodes (2/3)

Problem statement. There exist a complete vector

$$(\underbrace{Y_1, \dots Y_p}_{O=\text{observed}}, \underbrace{Z_1, \dots Z_r}_{H=\text{hidden}})$$

the distribution P of which is faithful to a graph G.

Missing nodes (2/3)

Problem statement. There exist a complete vector

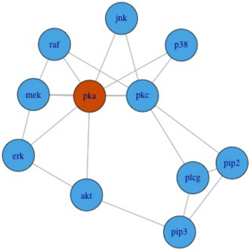
$$(\underbrace{Y_1, \dots Y_p}_{O=\text{observed}}, \underbrace{Z_1, \dots Z_r}_{H=\text{hidden}})$$

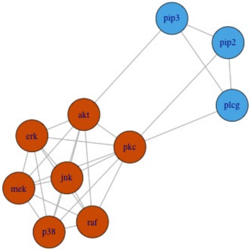
the distribution P of which is faithful to a graph G.

EM algorithm. [10]

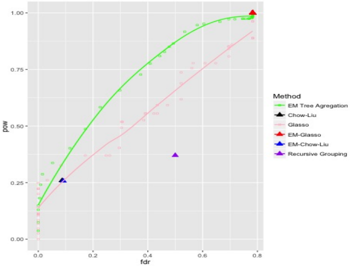
- E-step: infer H from O with current parameters $(\mu, \Sigma, G, ...)$
- M-step: update the parameters using O and $\widehat{H}$ (using [13] for G).

Missing nodes (3/3)





Accuracy for edge detection based on edge probability:



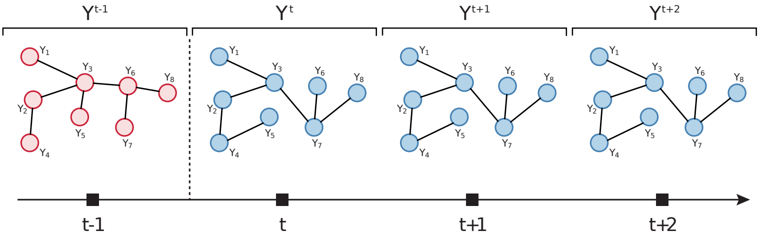
Change-point detection (1/2)

Data: $Y_t = (Y_{1t}, \dots Y_{pt})$ collected along time t.

Change-point detection (1/2)

Data: $Y_t = (Y_{1t}, \dots Y_{pt})$ collected along time t.

Problem: Suppose that Y_t is associated with graph G_t, that is affected by abrupt changes:



→ Infer both the change-points and the network associated with each period [9,12]

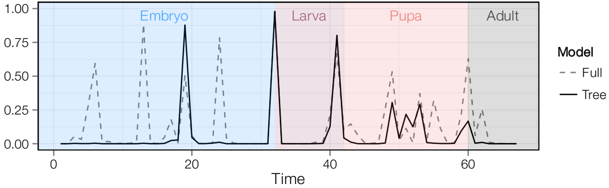
Change-point detection (2/2)

Data: $N = 67$ time points, $p = 11$ genes, four expected regions

Change-point detection (2/2)

Data: $N = 67$ time points, $p = 11$ genes, four expected regions

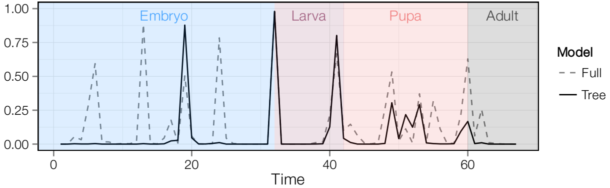
Posterior probability of change-points:



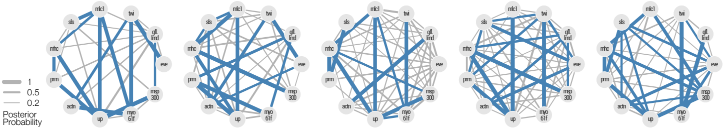
Change-point detection (2/2)

Data: $N = 67$ time points, $p = 11$ genes, four expected regions

Posterior probability of change-points:



Inferred networks:



1 - 18 | 19 - 31 | 32 - 40 | 41 - 52 | 53 - 67

Concluding remarks & questions

A generic framework for network inference, with not completely solved problems:

- Deal with NGS counts (non Gaussian)
- Accounting for covariates
- Incompletely observed species and/or variables

Concluding remarks & questions

A generic framework for network inference, with not completely solved problems:

- Deal with NGS counts (non Gaussian)
- Accounting for covariates
- Incompletely observed species and/or variables

Questions & remarks.

- Network = set of binary interactions
- What is an ecological network?

References I

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

B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani.
Least angle regression.
The Annals of statistics, 32(2):407–499, 2004.

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

J. M. Hammersley and P. Clifford.
Markov fields on finite graphs and lattices.
unpublished, 1971.

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

M. Meilă and T. Jaakkola.
Tractable Bayesian learning of tree belief networks.
March 2006.

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

References II

G. Rigaill, E. Lebarbier, and S. Robin.
Exact posterior distributions over the segmentation space and model selection for multiple change-point detection problem.
Stat. Comp., 22:917–29, 2011.
DOI: 10.1007/s11222-011-9258-8.

G. Robin.
Inférence de réseaux parcimonieux en présence de variables inobservées.
Master's thesis, univ. Paris-Saclay, 2016.

K. Sachs, O. Perez, D. Pe'er, D. A. Lauffenburger, and G. P. N.
Causal protein-signaling networks derived from multiparameter single-cell data.
Science (New York, N.Y.), 308:523–529, 2005.

L. Schwaller and S. Robin.
Exact bayesian inference for off-line change-point detection in tree-structured graphical models.
Statistics and Computing, pages 1–15, 2016.

L. Schwaller, S. Robin, and M. Stumpf.
Bayesian Inference of Graphical Model Structures Using Trees.
Technical report, ArXiv:1504.02723, April 2015.

R. Tibshirani.
Regression shrinkage and selection via the lasso.
Journal of the Royal Statistical Society. Series B (Methodological), pages 267–288, 1996.

C. Vacher, A. Tamaddoni-Nezhad, S. Kamenova, N. Peyrard, Y. Moalic, R. Sabbadin, L. Schwaller, J. Chiquet, M.A. Smith, J. Vallance, V. Fievet, B. Jakuschkin, and D. A. Bohan.
Chapter one - Learning ecological networks from next-generation sequencing data.
Advances in Ecological Research, 54:1–39, 2016.