Mixture tree model for network inference

Supervision: S. Robin\textsuperscript{1} et C. Ambroise\textsuperscript{12}

Raphaëlle Momal

\textsuperscript{1}UMR AgroParisTech / INRA MIA-Paris
\textsuperscript{2}LaMME, Evry

July 7, 2018
Motivation

Context

Rising interest in **jointly analysed** species abundances:

- Metagenomics
- Microbiology
- Ecology

Ecological network

Tool to better understand species interactions (direct/indirect), eco-systems organizations (clusters ?)

Allows for resilience analyses, pathogens control, ecosystem comparison, response prediction...
Data example

- **Species**: bacteria, fungi...
- **Abundances**: read counts from Next-Generation Sequencing technologies (metabarcoding)
- **Covariates**: sequencing depth, temperature, water depth...

Repeated signal : \( n \) samples, \( p \) abundances.

**Data table**

\[
Y = [Y_{ij}]_{(i,j)\in\{1,...,n\} \times \{1,...,p\}}
\]

- \( Y_{ij} \): abundance of the \( j^{th} \) species in the \( i^{th} \) sample

Infer the **species interaction network** from count data \( Y \)
Challenges

- Statistical network inference
- Count data
- Offsets and covariates
Graphical models: a statistical framework for network inference

Example:

- All variables are dependant
- Some are conditionally independent (i.e. indirectly dependant)

\[ Y_4 \text{ is independent from } (Y_1, Y_3) \text{ conditionally on } Y_2 \]
Definition [Lauritzen, 1996]

The joint distribution $P$ is faithful to the graph $G$ iff

$$P(Y_1, \ldots, Y_p) \propto \prod_{C \in C_G} \psi_C(Y_C)$$

where $C_G = \text{set of maximal cliques of } G$. 

\[ P(Y_1, Y_2, Y_3, Y_4) \propto \psi_1(Y_1, Y_2, Y_3) \times \psi_2(Y_3, Y_4) \]
Spanning trees

Unconstrained graph ⇒ very large space to explore: \( \#G_p = 2^{\frac{p(p-1)}{2}} \)

Spanning trees are a **sparse** solution :

\[
\begin{align*}
G & \text{ is connected} \\
G & \text{ has no cycle}
\end{align*}
\implies G \text{ has } (p - 1) \text{ edges}
\]

Much **smaller space** to explore:

\( \#T_p = p^{(p-2)} \)
Spanning trees

Unconstrained graph ⇒ very large space to explore: \( \#G_p = 2^{\frac{p(p-1)}{2}} \)

Spanning trees are a \textit{sparse} solution:

\[
\begin{align*}
G \text{ is connected} & \quad \text{G has no cycle} \\
G \text{ has } (p-1) \text{ edges}
\end{align*}
\]

Much \textit{smaller space} to explore:

\( \#T_p = p^{(p-2)} \)

Still a huge complexity...
Maximizing and summing over spanning trees

**Maximum spanning tree**  Kruskall’s algorithm

\[
\hat{T} = \arg\max_T \left\{ \prod_{(k,l) \in T} \psi_{k,l}(Y) \right\} \to \Theta(p^2)
\]

**Tree averaging**  Matrix tree theorem [Chaiken and Kleitman, 1978]

\[
\sum_T \prod_{(k,l) \in T} \psi_{k,l}(Y) = \det(L(Y)) \to \Theta(p^3)
\]
Maximizing and summing over spanning trees

**Maximum spanning tree** Kruskall’s algorithm

$$\hat{T} = \arg\max_T \left\{ \prod_{(k,l) \in T} \psi_{k,l}(Y) \right\} \to \Theta(p^2)$$

**Tree averaging** Matrix tree theorem [Chaiken and Kleitman, 1978]

$$\sum_T \prod_{(k,l) \in T} \psi_{k,l}(Y) = \det(L(Y)) \to \Theta(p^3)$$

**Approach:** infer the network by averaging spanning trees
Tree averaging

\[ P\{T = T_1|Y\} \]

\[ P\{T = T_2|Y\} \]

\[ P\{T = T_3|Y\} \]

\[ P\{T = T_4|Y\} \]

Compute edge probabilities:

\[ P\{(j, k) \in T|Y\} \]

Thresholding probabilities:

\[ P\{(j, k) \in T|Y\} \]
PLN model

Poisson log-Normal distribution [Aitchison and Ho, 1989]

\[
\begin{aligned}
Z_i \sim N_d(0, \Sigma) \\
(Y_{ij})_j \perp \perp |Z_i \\
Y_i | Z_i \sim P \mathcal{L} \mathcal{N}(0, \Sigma)
\end{aligned}
\]

- Dependency structure in the Gaussian latent layer
- Easy handling of multi-variate data (contrary to Negative binomial distribution)
PLN model

Poisson log-Normal distribution [Aitchison and Ho, 1989]

\[ Z_i \text{ iid } \sim \mathcal{N}_d(0, \Sigma) \]
\[ (Y_{ij})_j \perp \perp |Z_i \]
\[ Y_{ij}|Z_{ij} \sim \mathcal{P}(\exp^{o_{ij} + x_i^T \Theta_j + Z_{ij}}) \]
\[ Y \sim \mathcal{PLN}(O + x^T \Theta, \Sigma) \]

- Dependency structure in the Gaussian latent layer
- Easy handling of multi-variate data (contrary to Negative binomial distribution)
- Allow adjustment for covariates and offsets
PLN model

Poisson log-Normal distribution [Aitchison and Ho, 1989]

\[
Z_i \text{ iid } \sim \mathcal{N}_d(0, \Sigma) \\
(Y_{ij})_j \perp \perp |Z_i \\
Y_{ij}|Z_{ij} \sim \mathcal{P}(e^{o_{ij} + x_i^T \Theta_j + Z_{ij}}) \\
Y \sim \mathcal{PLN}(O + x^T \Theta, \Sigma)
\]

- Dependency structure in the Gaussian latent layer
- Easy handling of multi-variate data (contrary to Negative binomial distribution)
- Allow adjustment for covariates and offsets
- Variational estimation algorithm [Chiquet et al., 2017]

**Approach:** Infer the latent Gaussian network with an VEM algorithm.
Gaussian Graphical Models (GGM)

Gaussian distribution:

\[ Y_i \sim \mathcal{N}_p(\mu, \Sigma), \quad \mu = \text{vector of means}, \quad \Sigma = \text{covariance matrix}. \]

A nice property:

Inverse covariance matrix

\[
\Sigma^{-1} = \Omega \propto \begin{bmatrix}
1 & .5 & .5 & 0 \\
.5 & 1 & .5 & .5 \\
.5 & .5 & 1 & 0 \\
0 & .5 & 0 & 1 
\end{bmatrix}
\]
Gaussian Graphical Models (GGM)

Gaussian distribution:

\[ Y_i \sim \mathcal{N}_p(\mu, \Sigma), \quad \mu = \text{vector of means}, \quad \Sigma = \text{covariance matrix}. \]

A nice property:

\[
\begin{pmatrix}
Y_1 \\
Y_2 \\
Y_3 \\
Y_4
\end{pmatrix}
\]

Inverse covariance matrix

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

Glasso on gaussian data:

\[
\hat{\Omega}_\lambda = \arg \min_{\Omega \in S^+_d} \left\{ L(Y, \Omega) + \lambda \sum_{i \neq j} |\omega_{ij}| \right\}
\]

⇒ SpiecEasi method [Kurtz et al., 2015]: glasso on transformed counts
Simulation design

1. Choose $G$ and define $\Omega$ accordingly

2. Sample count data $Y$ from $\mathcal{PLN}(0, \Omega^{-1})$ with possible covariates

3. Infer the network with PLN + mixture tree VEM and SpiecEasi

4. Compare results with AUC (presence/absence of edges)

$\Rightarrow$ 40 replicates for each setting ($p, n, \text{edge probability}$)
Results: Erdös, 20 nodes
Oak Mildew

Pathogen *Erysiphe alphitoides* (EA).

Oak leaf with powdery mildew.

Metabarcoding of oak tree leaves microbiome [Jakuschkin et al., 2016].

- 114 sample of 94 microbial species counts (bacteria/fungi)
- Different read depth for bacteria and fungi: unsuited for normalization with SpiecEasi
- 3 quantitative covariates

We are interested in EA and F19, a second major fungi.
Model with covariates

Regression coefficients

<table>
<thead>
<tr>
<th></th>
<th>Covariates $\times 10^{-2}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>to base</td>
<td>to trunk</td>
</tr>
<tr>
<td>EA</td>
<td>-2.00</td>
</tr>
<tr>
<td>F19</td>
<td>2.19</td>
</tr>
</tbody>
</table>

Degree estimation

<table>
<thead>
<tr>
<th></th>
<th>Offset</th>
<th>Distances</th>
</tr>
</thead>
<tbody>
<tr>
<td>EA</td>
<td>2.20</td>
<td>1.86</td>
</tr>
<tr>
<td>F19</td>
<td>3.03</td>
<td>2.80</td>
</tr>
</tbody>
</table>
Inferred networks

Offset only

With covariates
Conclusion

Contributions:

- Formal probabilistic model for network inference with count data
- EM Estimation algorithm
- Inclusion of offsets and covariates

Perspectives:

- Method for determining the threshold
- Network comparison
- Model for the inference in the observed counts layer
- Missing major actor (species/covariable)
Acknowledgments

Special thanks:

**Supervisors** Stéphane Robin, Christophe Ambroise

**PLN team** Julien Chiquet (MIA-Paris), Mahendra Mariadassou (INRA Jouy)

**Data** Corinne Vacher (INRA Bordeaux)

Contact:

* email raphaelle.momal@agroparistech.fr
* Web Rmomal.github.io


