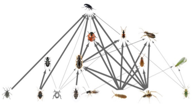
Latent Block Model for Overdispersed Count Data
Application in Microbial Ecology

J. Aubert

julie.aubert@agroparistech.fr



Journée NetBio
Saclay

Micro-biote/biome : définitions

Une communauté microbienne (ou microbiote) :

est un ensemble d'organismes issus de populations d'espèces distinctes qui cohabitent sur une même aire géographique à un moment donné et qui constituent un réseau d'interactions (= individus)

Le microbiome :

Le microbiote dans son environnement

Microbiomes d'intérêt :

Le phytobiome (microbiome d'un végétal) et notamment

- celui de la rhizosphère (dans le sol, autour du système racinaire)
- et de la phyllosphère (sur et sous les feuilles).

Le microbiome du sol
Le microbiome intestinal

Quelques questions d'intérêt

- Des questions de diversité : qui est là ? en quelle quantité ?
- Comment les communautés microbiennes se forment-elles, évoluent et interagissent-elles ?
- Quelle est l'influence des communautés microbiennes sur la santé (des plantes, des hommes) ? sur les performances de plante ? sur les fonctions de l'écosystème ?

A typical metagenomic experiment

Amplicon-based sampling. Consider

- n different (bacterial, fungal, ...) species / OTU and
- m different samples / patients / media / conditions.

NGS provides

$$\begin{aligned} Y_{ij} &= \text{number of reads from species } i \text{ in sample } j \\ &\propto \text{abundance of species } i \text{ in sample } j \end{aligned}$$

Question. Can we exhibit some patterns in the distribution of the species abundances across samples ?

Bi-clustering problem

Rephrased problem : Find

- groups of species having similar abundance profile across the samples and
- groups of samples histing the different species in similar proportions.

Bi-clustering problem

Rephrased problem : Find

- groups of species having similar abundance profile across the samples and
- groups of samples histing the different species in similar proportions.

Bi-clustering problem : Simultaneously determine

- row clusters and
- column clusters

in a $n \times m$ matrix of counts.

Bi-clustering problem

	S_1	S_2	S_3	$\dots$	S_j	$\dots$	S_m
OTU 1	0	0	0	$\dots$	y_{1j}	$\dots$	3
OTU 2	59	17	43	$\dots$	y_{2j}	$\dots$	3
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
OTU i	y_{i1}	y_{i2}	y_{i3}	$\dots$	y_{ij}	$\dots$	y_{id}
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
OTU n	90	1 20	123	$\dots$	y_{nj}	$\dots$	2
Seq. depth	4738	5157	6010	$\dots$	$\sum_{i=1}^{n} y_{ij}$	$\dots$	5916

y_{ij} = number of sequences from sample j assigned to Operational Taxonomic Unit (OTU) i.

Approach

Model-based clustering :

$\rightarrow$ LBM = Latent Block-Model

(Govaert and Nadif, 2005 ; Brault and Mariadassou, 2015)

Approach

Model-based clustering :

$\rightarrow$ LBM = Latent Block-Model

(Govaert and Nadif, 2005 ; Brault and Mariadassou, 2015)

Specificities of NGS data :

- count data,
- over dispersed (wrt Poisson),
- with heterogeneous sampling effort (= sequencing depth),
- with high variation among the species abundances,
- possibly with replicates.

Latent Block Model

Bi-clustering. K species groups, G sample groups

- $Z_i =$ group to which species i belongs to $(\in \{1, ...K\})$;
- $W_j =$ group to which sample j belongs to $(\in \{1, ...G\})$

both latent = hidden = unobserved.

→ Incomplete data model

Latent Block Model

Bi-clustering. K species groups, G sample groups

- $Z_i =$ group to which species i belongs to $(\in \{1, ...K\})$;
- $W_j =$ group to which sample j belongs to $(\in \{1, ...G\})$

both latent = hidden = unobserved.

→ Incomplete data model

Ex : Poisson LBM.

$$\begin{aligned} (Z_i) \text{ iid} &\sim \pi && \text{(species prop.)} \\ (W_j) \text{ iid} &\sim \rho && \text{(sample prop.)} \\ (Y_{ij}) \text{ indep } |(Z_i);(W_j) &\sim \mathcal{P}(\lambda_{Z_i W_j}) \end{aligned}$$

Does not accommodate for NGS data specificities.

Over-dispersion

Negative-binomial. Most popular distribution of NGS counts :

$$Y \sim \mathcal{NB}(\lambda, \phi) \qquad \mathbb{E}(Y) = \lambda, \qquad \mathbb{V}(Y) = \lambda(1 + \phi\lambda) \geq \lambda.$$

Over-dispersion

Negative-binomial. Most popular distribution of NGS counts :

$$Y \sim \mathcal{NB}(\lambda, \phi) \qquad \mathbb{E}(Y) = \lambda, \qquad \mathbb{V}(Y) = \lambda(1 + \phi\lambda) \geq \lambda.$$

Gamma-Poisson representation. Take $a = 1/\phi$ and draw

$$U \sim \mathcal{G}\text{am}(a, a), \quad Y \mid U \sim \mathcal{P}(\lambda U) \qquad \Rightarrow \qquad Y \sim \mathcal{NB}(\lambda, \phi).$$

Negative binomial = Poisson with latent Gamma

→ Incomplete data model (Y is observed, U is not).

LBM for metagenomic data

Hidden layer :

$$\begin{array}{rcll} (Z_i) \text{ iid} & \sim & \pi & \text{(species prop.)} \\ (W_j) \text{ iid} & \sim & \rho & \text{(sample prop.)} \\ (U_{ij}) \text{ iid} & \sim & \mathcal{G}\text{am}(a_{Z_i W_j}, a_{Z_i W_j}) & \end{array}$$

LBM for metagenomic data

Hidden layer :

$$\begin{aligned}
(Z_i) \text{ iid} &\sim \pi && \text{(species prop.)} \\
(W_j) \text{ iid} &\sim \rho && \text{(sample prop.)} \\
(U_{ij}) \text{ iid} &\sim \mathcal{G}\text{am}(a_{Z_i W_j}, a_{Z_i W_j})
\end{aligned}$$

Observed counts : (interest of model-based approaches)

$$Y_{ij} \mid Z, W, U \sim \mathcal{P}\left(\mu_i \, \nu_j \, \alpha_{Z_i W_j} \, U_{ij}\right)$$

where

- μ_i : mean abundance of species i
- ν_j : sequencing depth in sample j (fixed)
- α_{kg} : interaction term between group species k and sample group g.

LBM for metagenomic data

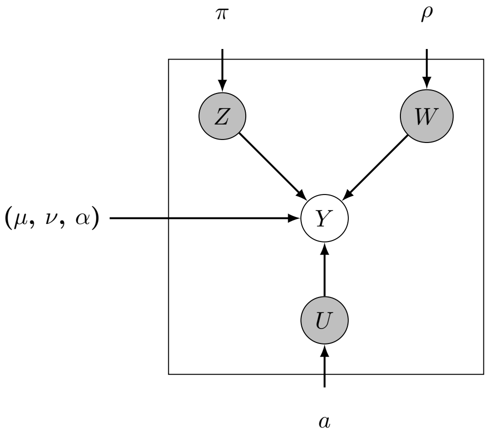


FIGURE – The proposed over-dispersed Poisson LBM presented as a directed graphical model. Legend : observed variables (filled white), latent variables (filled gray), parameters are outside the box.

Inference

Aim : Retrieve

- $Z_i =$ species group, or at least $P(i \in k|Y)$;
- $W_j =$ sample group, or at least $P(j \in g|Y)$;

and estimate the interaction parameter $\alpha = (\alpha_{kg})$.

Inference

Aim : Retrieve

- Z_i = species group, or at least $P(i \in k|Y)$;
- W_j = sample group, or at least $P(j \in g|Y)$;

and estimate the interaction parameter $\alpha = (\alpha_{kg})$.

Which means (maximum-likelihood approach)

- Compute $p(Z, W, U|Y)$;
- Maximize $\log p_\theta(Y)$, where $\theta = (\alpha, \mu)$.

Most popular algorithm : EM (Dempster et al., 1977).

Variational approximation

Species group Z_i and sample group W_j are not independent given Y_{ij}

$$\rightarrow p(Z, W, U \mid Y) \text{ intractable}$$

Variational approximation (Jordan, 1999). Find

$$\begin{array}{rrcl} & \widetilde{p}(Z, W, U) & \simeq & p(Z, W, U|Y) \\ \text{such that} & \widetilde{p}(Z, W, U) & = & \widetilde{p}_1(Z)\,\widetilde{p}_2(W)\,\widetilde{p}_3(U) \end{array}$$

(mean-field approximation).

→ Variational EM (VEM) algorithm provide a lower bound

$$J(Y, \widetilde{p}, \widehat{\theta}) \leq \log p_{\widehat{\theta}}(Y).$$

Penalized 'likelihood' criteria

Penalized criterion. $\log p_{\widehat{\theta}}(Y)$ intractable

$$\log p_{\widehat{\theta}}(Y) - \text{pen}(p_{\widehat{\theta}}) \qquad \rightarrow \qquad J(Y, \widetilde{p}, \widehat{\theta}) - \text{pen}(p_{\widehat{\theta}})$$

BIC & ICL. $\mathcal{H}$ = entropy

$$\text{pen}_{BIC} = \left[(K-1)\log n - (G-1)\log m - KG\log(nm)\right]/2$$

$$\text{pen}_{ICL_1} = \text{pen}_{BIC} + \mathcal{H}(\widetilde{p}_Z) + \mathcal{H}(\widetilde{p}_W) \qquad \text{(classification entropy)}$$

Model comparison

Likelihood ratio for nested models.

$\mathcal{M} \subset \mathcal{M}'$, the likelihood ratio is defined as

$$LR(\mathcal{M}, \mathcal{M}') = 2 \left[\log p(\mathbf{Y}; \widehat{\theta}_{\mathcal{M}'}) - \log p(\mathbf{Y}; \widehat{\theta}_{\mathcal{M}}) \right].$$

Interest of block structure.

$\mathcal{M}_{\text{min}} := \mathcal{M}_{1,1} \subset \mathcal{M}_{K,G} \subset \mathcal{M}_{\text{max}} := \mathcal{M}_{n,m}$

Lower bounds for likelihood ratios.

$$\begin{aligned} (a): \quad LR(\mathcal{M}_{\text{min}}, \mathcal{M}_{K,G}) &\geq 2 \left[\mathcal{J}(\mathbf{Y}, \widehat{q}_{K,G}, \widehat{\theta}_{K,G}) - \log p(\mathbf{Y}; \widehat{\theta}_{1,1}) \right], \\ (b): \quad LR(\mathcal{M}_{K,G}, \mathcal{M}_{\text{max}}) &\leq 2 \left[\log p(\mathbf{Y}; \widehat{\theta}_{n,p}) - \mathcal{J}(\mathbf{Y}, \widehat{q}_{K,G}, \widehat{\theta}_{K,G}) \right]. \end{aligned}$$

Three 16S or 18S rRNA amplicon-based datasets

- MetaRhizo : plants and bacteria communities living in their rhizosphere (collab. C. Mougel, INRA Rennes)
- Oak powdery mildew : bacteria and fungi including *Erysiphe alphitoides* living in the phyllosphere (collab. C. Vacher, INRA Bordeaux)
- Macaroni : microbial community assembly in soil (collab. L. Philippot, A. Spor, INRA Dijon)

Aim : to understand the structure of these relationships

Meta-rhizo

Dataset : Medicago truncatula rhizosphere.

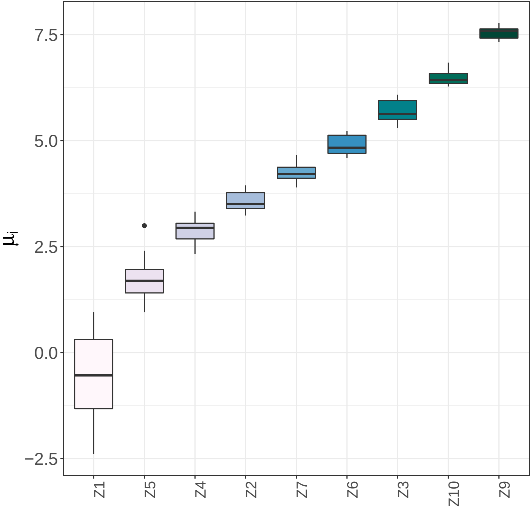
- $n = 288$ bacteria (genus)
- $m = 483$ samples = rhizosphere of different plants (genotypes)

The total counts per sample go from 29410 to 33840 number of sequences.
19.2% of data are null
Range from 0 to 5084 with a median = 9 and mean = 110

Results :

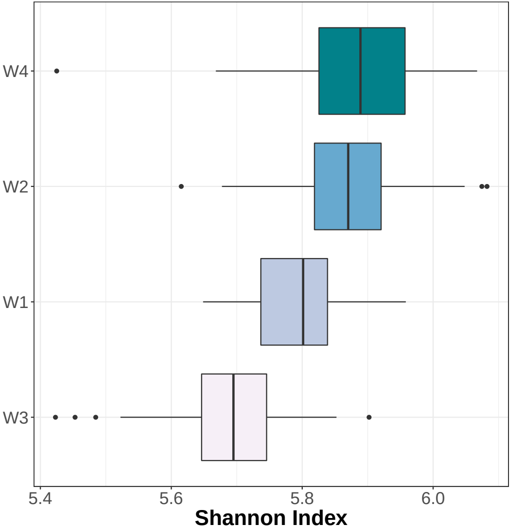
- $\widehat{K} = 10$ groups of bacteria
- $\widehat{G} = 4$ groups of samples
- $\hat{a} = 7.29$

MetaRhizo



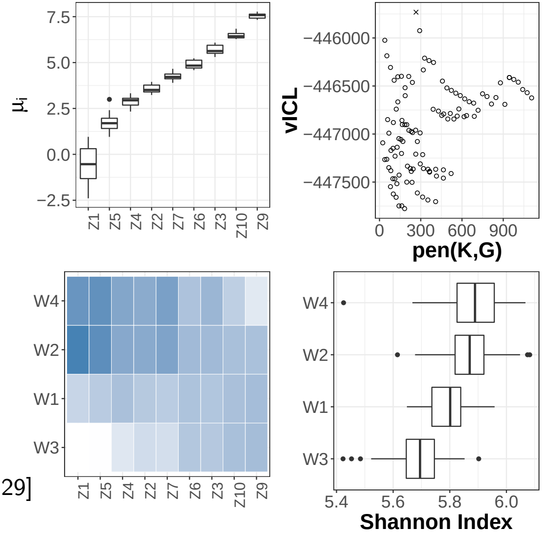
Despite ν_j, bacteria groups correspond to abundance groups.

MetaRhizo



Plant groups corresponds to diversity levels (Shannon index).

MetaRhizo



$\alpha_{kg} \in [0.6; 3.29]$

MetaRhizo

Goodness of fit.

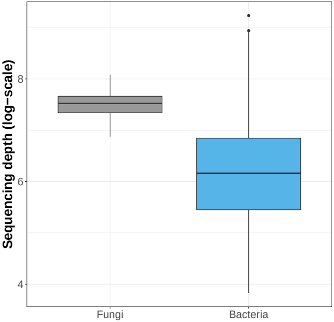
TABLE – MetaRhizo data. Goodness-of-fit. LR is the likelihood ratio statistic as defined in Section and df stands for difference in terms of free parameters.

$\mathcal{M}, \mathcal{M}'$	$LR(\mathcal{M}, \mathcal{M}')$	df	$LR(\mathcal{M}, \mathcal{M}')/df$
$\mathcal{M}_{\text{min}}, \mathcal{M}_{KG}$	37804.75	40	945.12
$\mathcal{M}_{KG}, \mathcal{M}_{\text{max}}$	143881	139064	1.03

Oak powdery mildew

Dataset : Pathobiome of the Erysiphe alphitoides (Jakuschkin et al. 2016).

- $n = 114 = \textit{E. alphitoides} + 47$ fungal $+66$ bacterial otus
- $m = 116$ leaves from 3 trees (resistant, intermediate, susceptible)
- 34% of data are null
- Range from 0 to 2228 (median $= 2$; mean $= 24.17$)

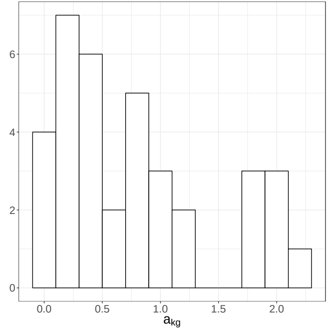


$$\rightarrow 2\nu_j = (\nu_j^{\text{bact}}, \nu_j^{\text{fung}})$$

Oak powdery mildew

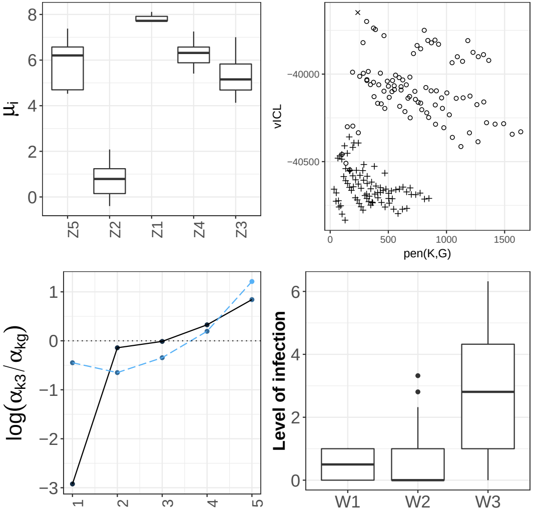
Results :

- Common a : $(\widehat{K} = 1,\ \widehat{G} = 1)$
- a_{kg} : $(\widehat{K} = 5,\ \widehat{G} = 3)$



$\alpha_{kg} \in [0.22; 2.14]$ (ratio from 1 to 9.6).

Oak powdery mildew



Oak powdery mildew

Comments :

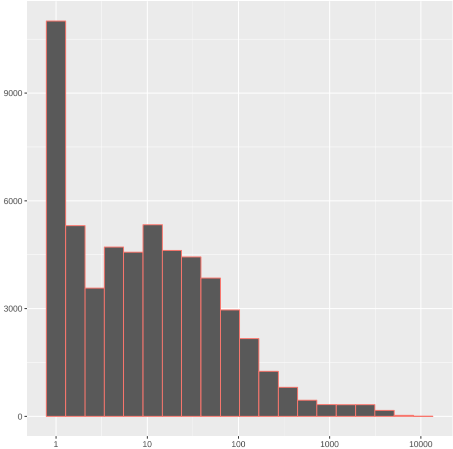
- Heterogeneous over-dispersion parameters (a_{kg}),
- Groups reveal the abundance of *E. alphitoides* (pathogene)

MicrobiAl Community Assembly Rules and functiONIng

Aim : Identify biotic interactions between microbial groups using a targeted subtractive approach by removal and enrichment of specific microbial groups

Data : After filtering steps, 353 OTUs and 347 biological samples (10 treatments)

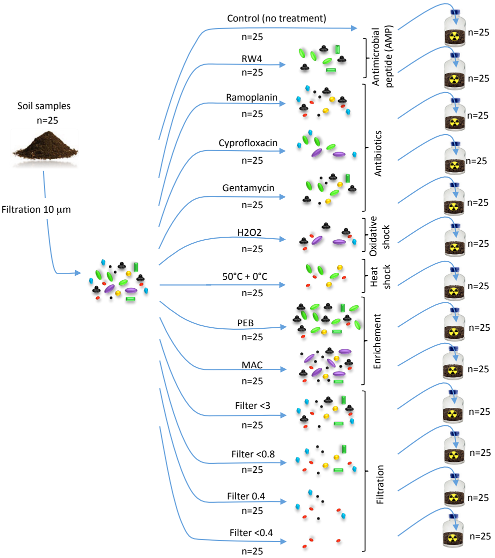
- 54% of data are null, Mean = 35.3, Max = 10598.



Approach and methods

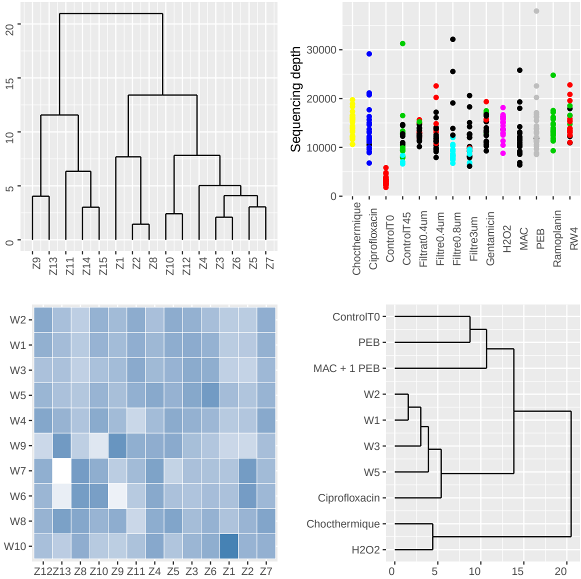
1. Ten time dilution of soil suspension filtered at 10 μm to focus on dominant bacterial groups.
2. **Removal, killing or preventing the growth of specific groups**
   - according to their cell size using filtration (4 size classes)
   - by incubating the soil suspension with (*i*) antibiotics targeting different groups and (*ii*) group specific antimicrobial peptides
   - according to the membrane properties by subjecting the soil suspension to osmotic and heat shocks
   - enrichment by incubating the soil suspension with inhibitors
3. For each treatment : inoculation into 25 microcosms containing sterilized soils.
4. Collect after 45 days for molecular and activity analyses.
5. Illumina Miseq sequencing
6. Bioinformatic annalysis with house pipeline (A. Spor)

Experimental Design



Selected latent block model

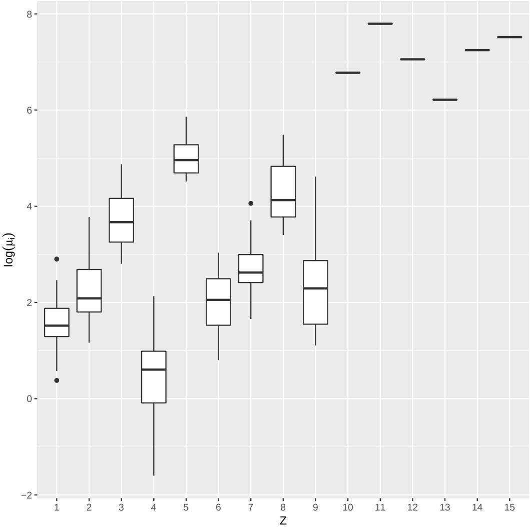
Common a ($\hat{a} = 0.32$) : ($\widehat{K} = 15$, $\widehat{G} = 10$)



Description of groups in columns

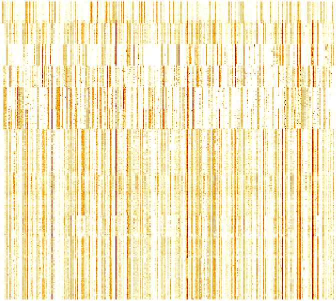
	W_1	W_2	W_3	W_4	W_5	W_6	W_7	W_8	W_9	W_{10}
Chocthermique	0	0	0	0	0	0	25	0	0	0
Ciprofloxacin	1	0	0	24	0	0	0	0	0	0
ControlT0	0	0	0	0	0	0	0	0	0	25
ControlT45	6	0	16	0	3	0	0	0	0	0
Filtrat0.4um	13	10	2	0	0	0	0	0	0	0
Filtre0.4um	17	8	0	0	0	0	0	0	0	0
Filtre0.8um	5	0	0	0	20	0	0	0	0	0
Filtre3um	19	1	0	0	5	0	0	0	0	0
Gentamicin	17	4	3	0	0	0	0	0	0	0
H2O2	1	0	0	0	0	23	0	0	0	0
MAC	0	0	0	0	0	0	0	0	25	0
PEB	0	0	0	0	0	0	0	24	1	0
Ramoplanin	0	0	24	0	0	0	0	0	0	0
RW4	4	18	3	0	0	0	0	0	0	0

Groups of bacteria

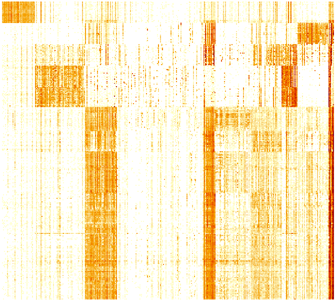


Heatmap

Before

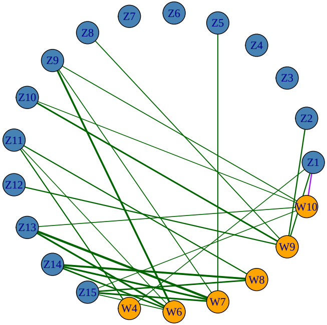


After



Network representation

- One vertex = one group of microorganism (Z_i in blue) or one group of soil (W_j in orange)
- Incidence matrix : use of α_{kg} matrix (abs. value > 1)
- Edge color : green for negative, purple for positive interactions



Discussions

Summary

- Parsimonious and complex model enables us to reduce data dimension
- ICL criteria to select number of groups
- Parameters biologically interpretable
- cobiclust R package

Possible extensions

Comments

- Dispersion parameter
- Normalization
- Zero-inflation

Acknowledgments

For experiments, datasets and biological expertise

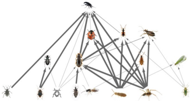
Agroécologie
Dijon
Unité de Recherche

C. Mougel (IGEPP)
A. Zancarini
C. Le Signor

L. Philippot (UMR AgroEcologie)
S. Rhodmane
A. Spor

For the statistical part

S. Robin S. Schbath S. Ouadah



ANR-17-CE32-0011-Next Generation Biomonitoring of change in ecosystem structure and function

References

[1] Govaert, G. and Nadif, M. (2010), **Latent Block Model for contingency table**, *Communications in Statistics - Theory and Methods*, 39(3), 416–425.

[2] Brault, V. and Mariadassou, M. (2015), **Co-clustering through latent bloc model : a review.**, *Journal de la Société Française de Statistique*, 156.

[3] Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977), **Maximum likelihood from incomplete data via the EM algorithm**, *Journal of the Royal Statistical Society, B* 39(1), 1-38.

[4] Jordan, M. I. et al. (1999), **Graphical models, exponential families, and variational inference. Found.**, *Trends Mach. Learn.*,1, 1–305.

[6] Jakushchkin, B. et al. (1999), **Deciphering the pathobiome : intra and interkingdom interactions involving the pathogem erysihe alphitoides**, *Microb. Ecol.*, 72(4) : 870-880.