Inference de réseaux écologiques à partir de données de comptage

J. Chiquet[2], J. Lao[1], M. Mariadassou[1], S. Robin[2], S. Schbath[1]

[1]MAIAGE, INRA, Université Paris-Saclay, 78350 Jouy-en-Josas, France
[2]MIA 518, INRA, AgroParisTech, 16 Rue Claude Bernard, 75005 Paris, France

Journées NETBIO
10 Novembre 2017

Plan

Métagénomique

Principe

Caractérisation de la **globalité** d'un écosystème microbien,
sans isolation au préalable de microorganismes qui le compose

Métagénomique

Principe

Caractérisation de la **globalité** d'un écosystème microbien,
sans isolation au préalable de microorganismes qui le compose

Pourquoi ?

- $< 1\%$ des microorganismes de la biosphère sont cultivables *in vitro*

Métagénomique

Principe

Caractérisation de la **globalité** d'un écosystème microbien,
sans isolation au préalable de microorganismes qui le compose

Pourquoi ?

- $< 1\%$ des microorganismes de la biosphère sont cultivables *in vitro*
- Perte des interactions écologiques avec les analyses par isolation

Métagénomique

Principe

Caractérisation de la **globalité** d'un écosystème microbien,
sans isolation au préalable de microorganismes qui le compose

Pourquoi ?

- $< 1\%$ des microorganismes de la biosphère sont cultivables *in vitro*
- Perte des interactions écologiques avec les analyses par isolation

Fort intérêt

Microbiote intestinal humain (MetaHit, HMP), écosystèmes océaniques (Tara Expeditions), et de la terre (Earth Microbiome Project), Produits Fermentés (FoodMicrobiome, MétaPDOCheese)

Amplicon

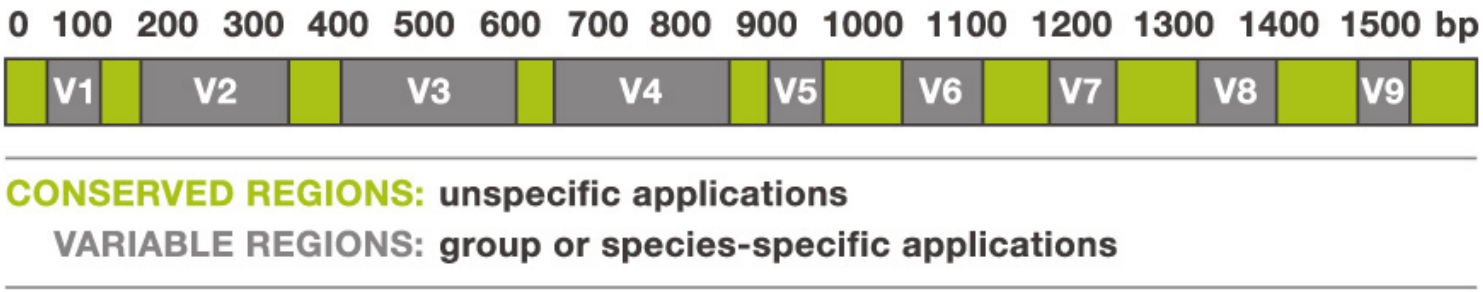
- Identification d'un mélange de bactéries via leurs séquences génétiques;

Amplicon

- Identification d'un mélange de bactéries via leurs séquences génétiques;
- Utilisation d'un marqueur génétique **universel** : ARNr 16S pour les prokaryotes

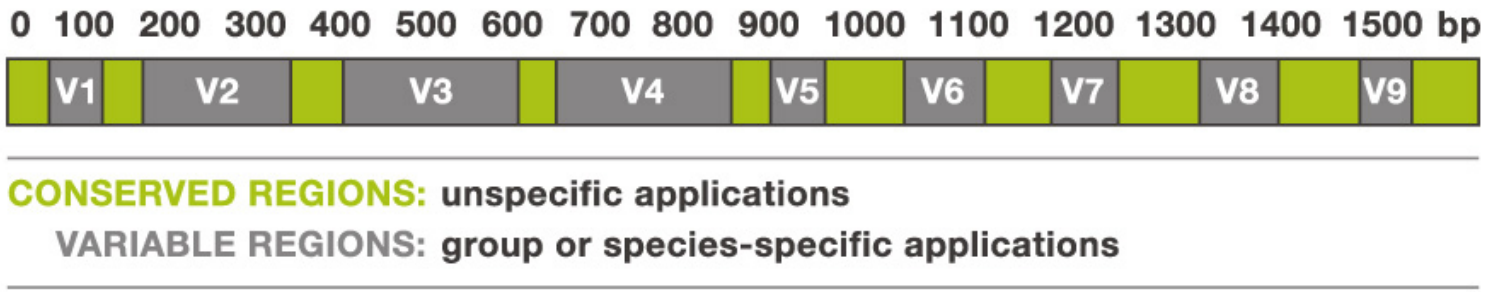
Amplicon

- Identification d'un mélange de bactéries via leurs séquences génétiques;
- Utilisation d'un marqueur génétique **universel** : ARNr 16S pour les prokaryotes
- Caractéristiques importantes :
  - Succession de régions variables et conservées



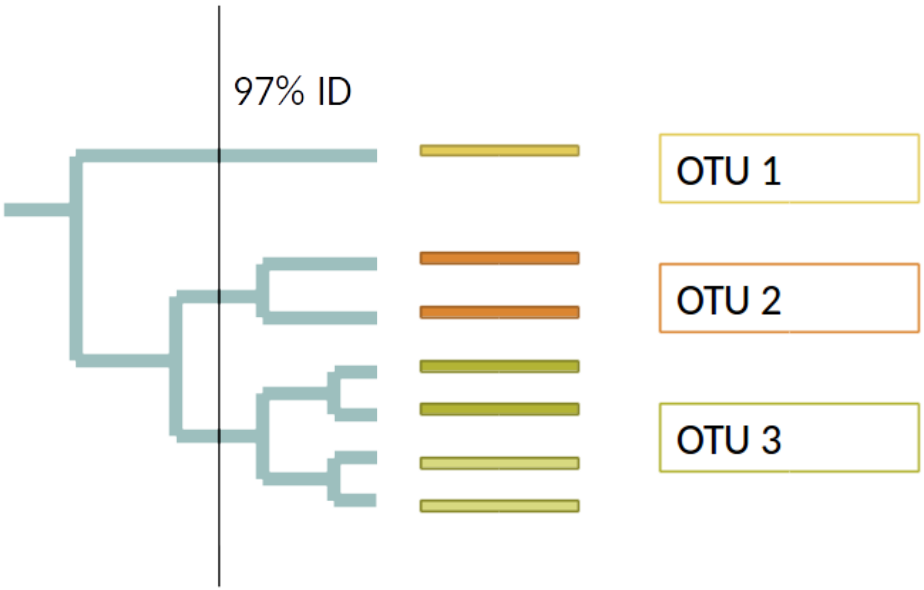
Amplicon

- Identification d'un mélange de bactéries via leurs séquences génétiques;
- Utilisation d'un marqueur génétique **universel** : ARNr 16S pour les prokaryotes
- Caractéristiques importantes :
  - Succession de régions variables et conservées

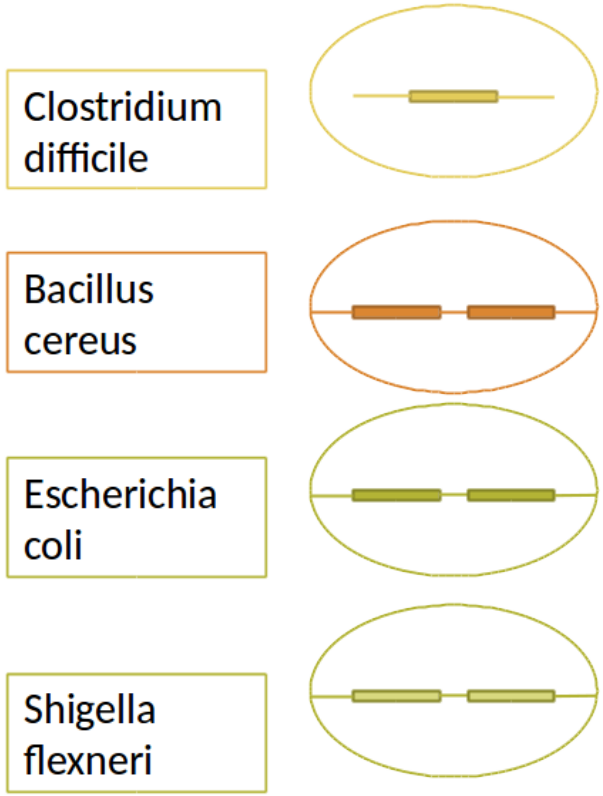


  - Amorces PCR possibles sur les régions flanquantes car conservées
  - Régions variables sont de longueur "séquençable"
  - Bases de données exhaustives et (à peu près) fiables

Principe des OTUs



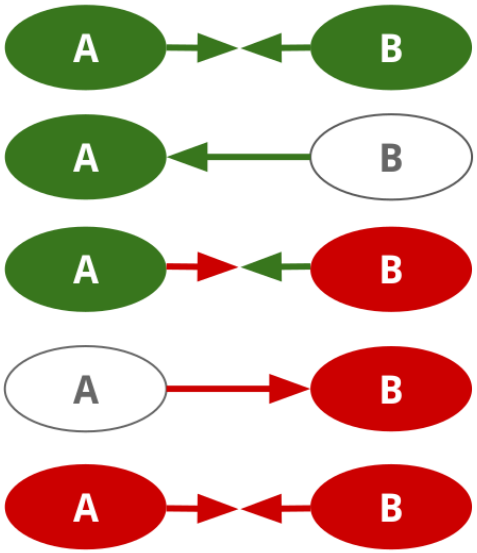
Phylogenetic tree of 16S rRNA

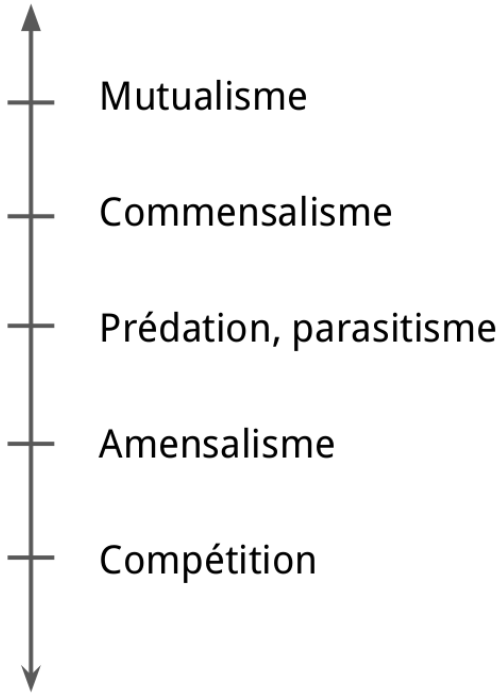


Données Types

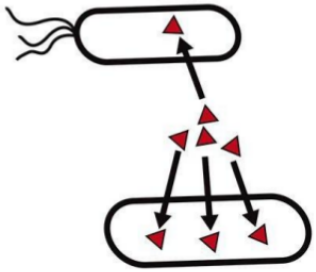
```
##        Sample_a Sample_b Sample_c Sample_d Sample_e
## OTU_1         3        3        3        3        6
## OTU_2         5        2        3        3        1
## OTU_3         2        0        2        4        3
## OTU_4         4        3        0        4        3
## OTU_5         0        2        2        1        4
## OTU_6         0        2        2        1        3
```

Interactions écologiques

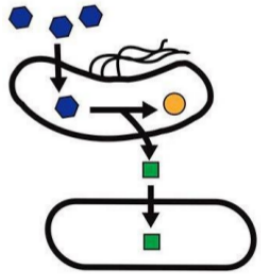




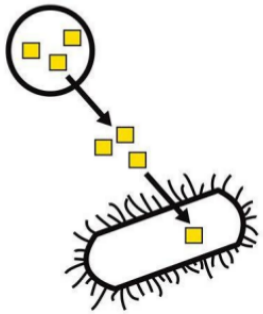
Compétition



Syntrophie

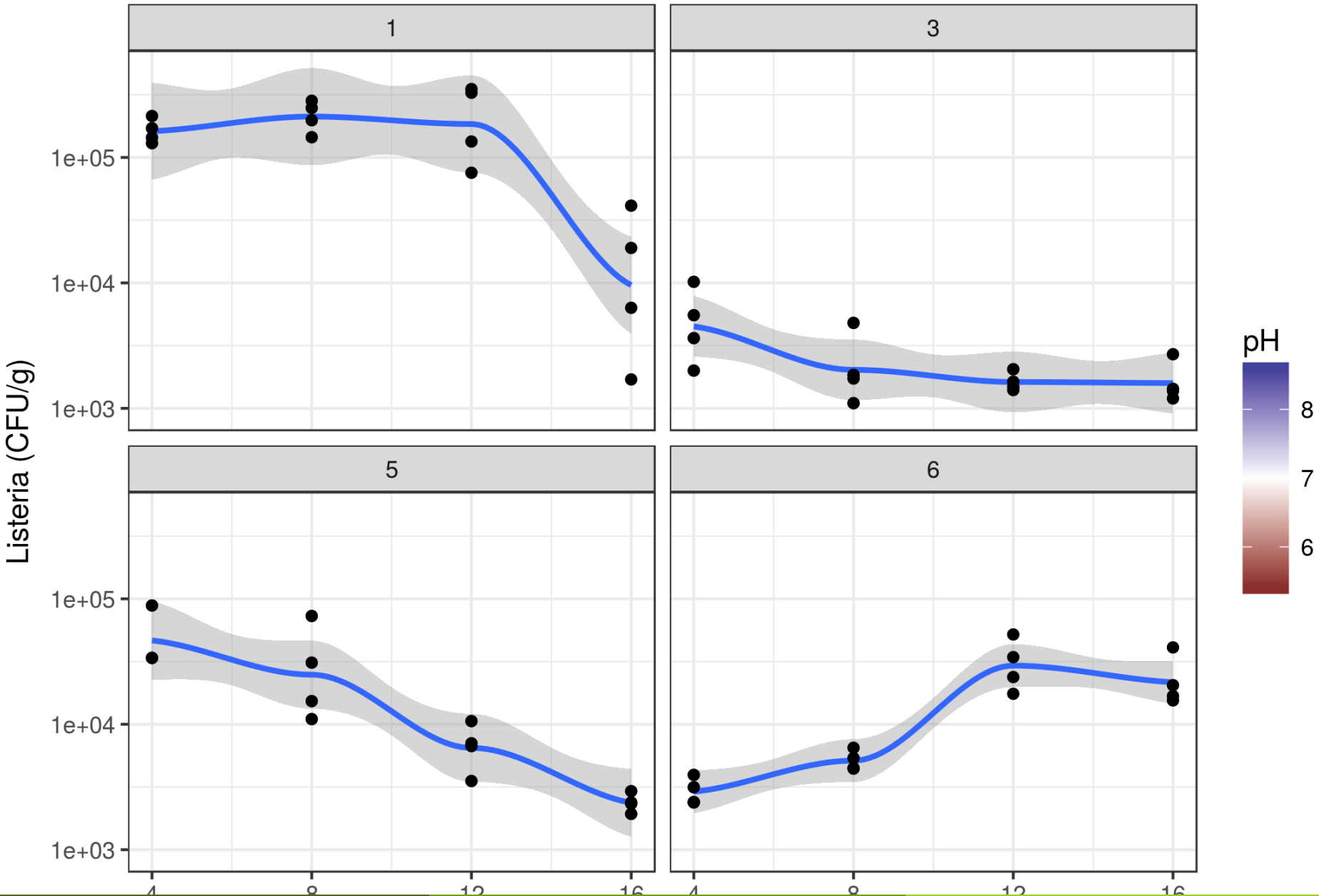


Alimentation croisée

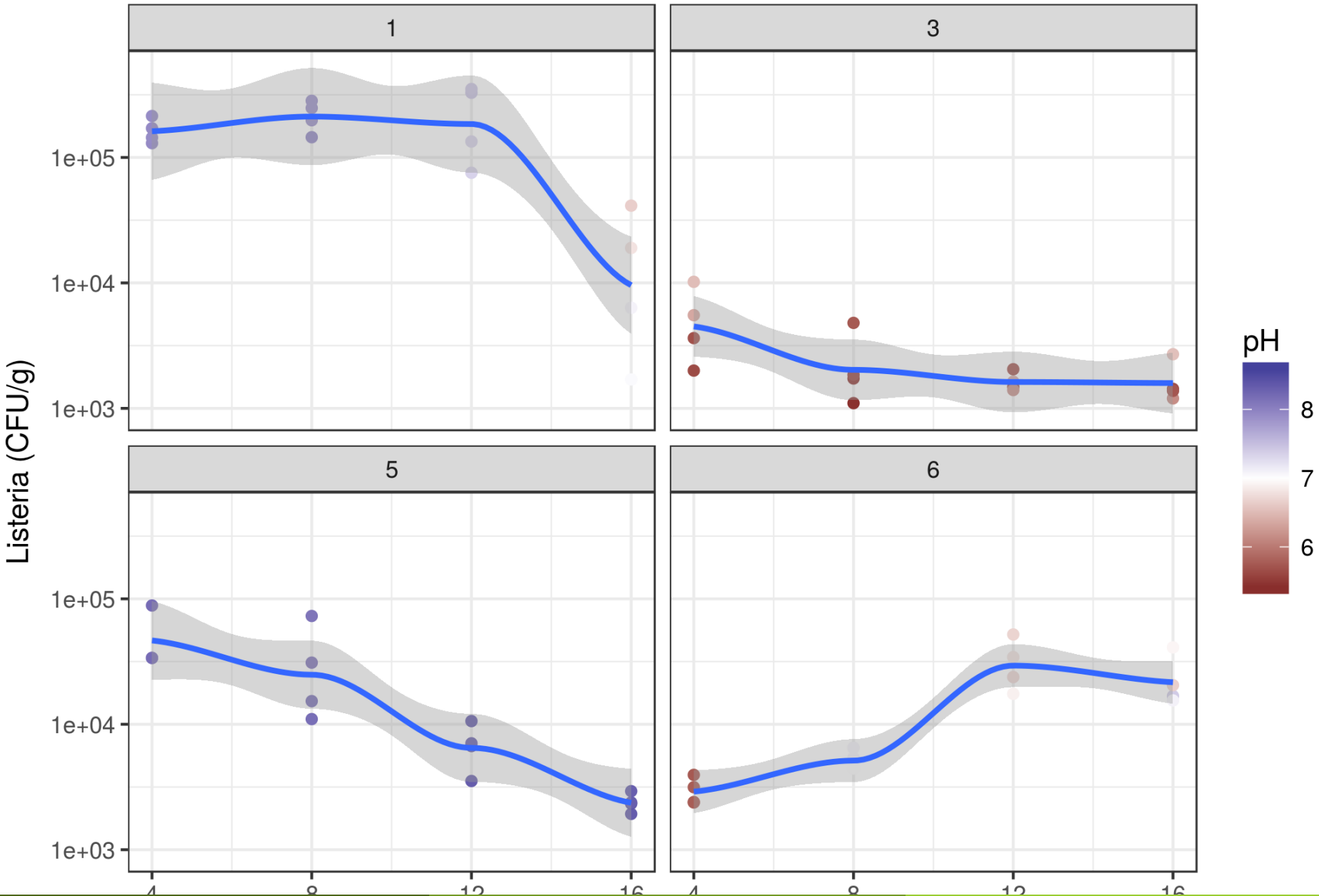


Seth EC, Taga ME (2014). Nutrient cross-feeding in the microbial world. *Front. Microbiol.* **5**(350). doi:10.3389/fmicb.2014.00350

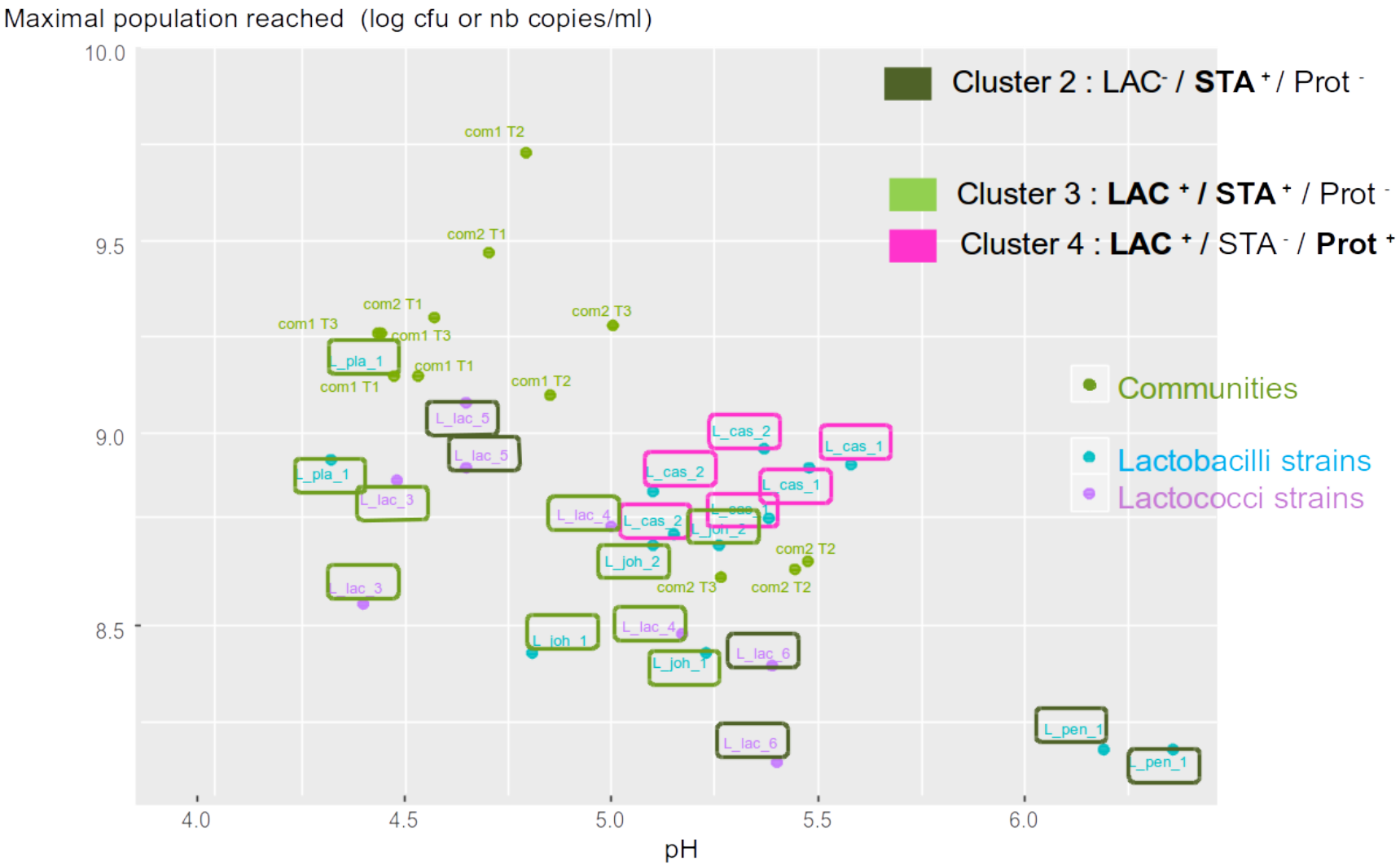
Exemple: Fortress (E. Dugat-Bony)



Exemple: Fortress (E. Dugat-Bony)



Exemple: Millup (V. Gagnaire/A. Thierry)



Interaction

- On se limite à des interactions entre deux taxa;
- On s'intéresse surtout à des interactions de type
  - *commensales*,
  - *mutualistes*,
  - *compétitives*

  Le reste étant trop difficile à détecter

Interaction $\simeq$ co-occurrence

- On se limite à des interactions entre deux taxa;
- On s'intéresse surtout à des interactions de type
  - *commensales*,
  - *mutualistes*,
  - *compétitives*

  Le reste étant trop difficile à détecter
- Ces 3 trois types induisent par de la co-présence (commensalisme, mutualisme) ou de la co-exclusion (compétition)
- On passe par la co-occurrence pour identifier des interactions

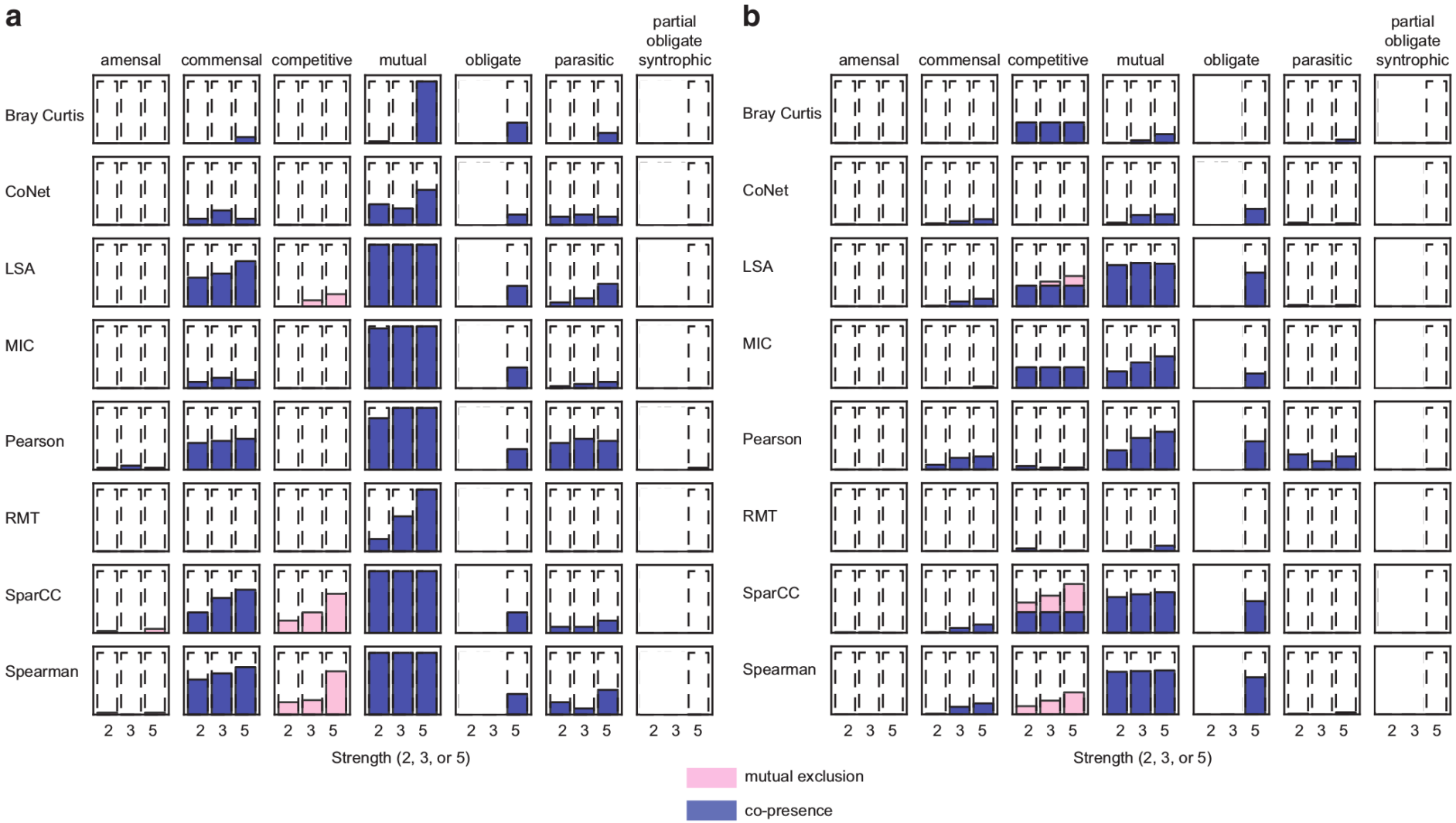
Interaction $\simeq$ co-occurrence

- On se limite à des interactions entre deux taxa;
- On s'intéresse surtout à des interactions de type
  - *commensales*,
  - *mutualistes*,
  - *compétitives*

  Le reste étant trop difficile à détecter
- Ces 3 trois types induisent par de la co-présence (commensalisme, mutualisme) ou de la co-exclusion (compétition)
- On passe par la co-occurrence pour identifier des interactions

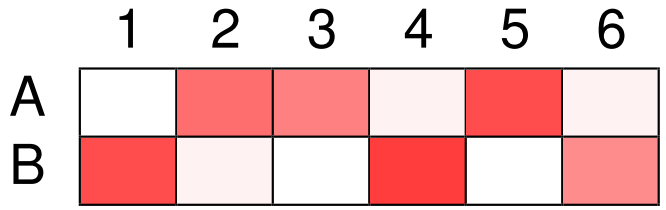
Interaction $\rightleftharpoons$ Co-occurrence

Autres interactions (d'après Weiss et al., 2016)



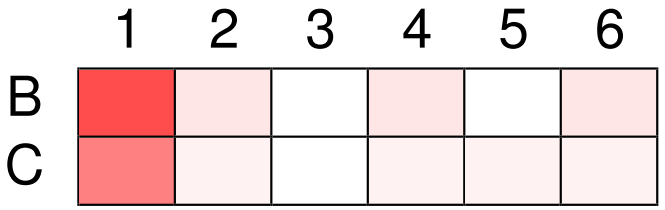
Réseaux de co-occurrences

Observation des abondances



Exclusion mutuelle de A et B

→ **Anti-corrélation**

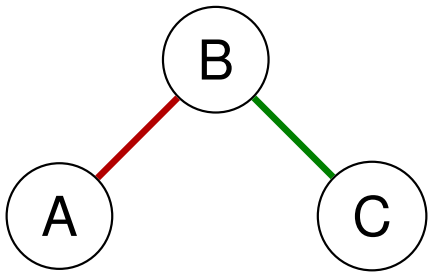


Co-occurrence de B et C

→ **Corrélation**

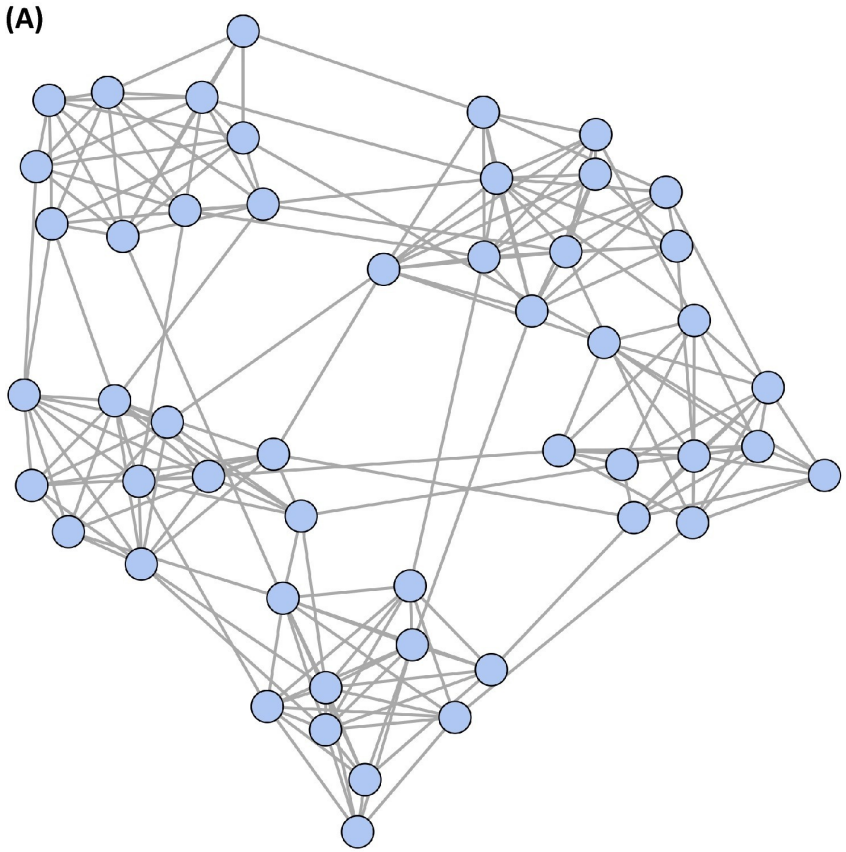
⇓

Inférence du réseau



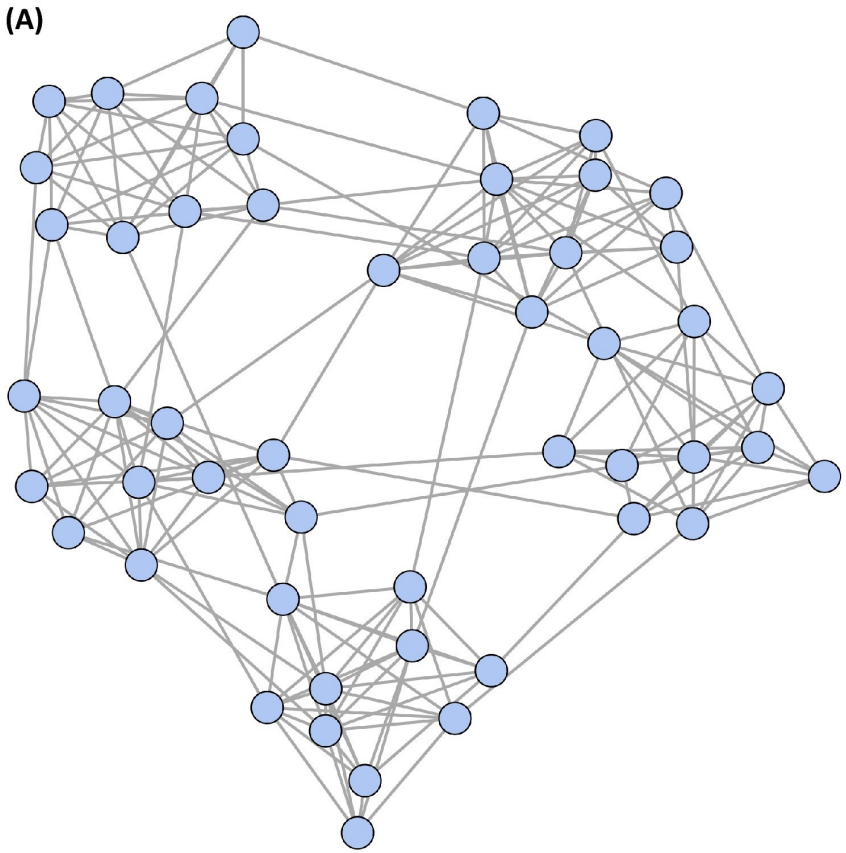
Des réseaux pour quoi faire? (Layeghifard et al. 2017)

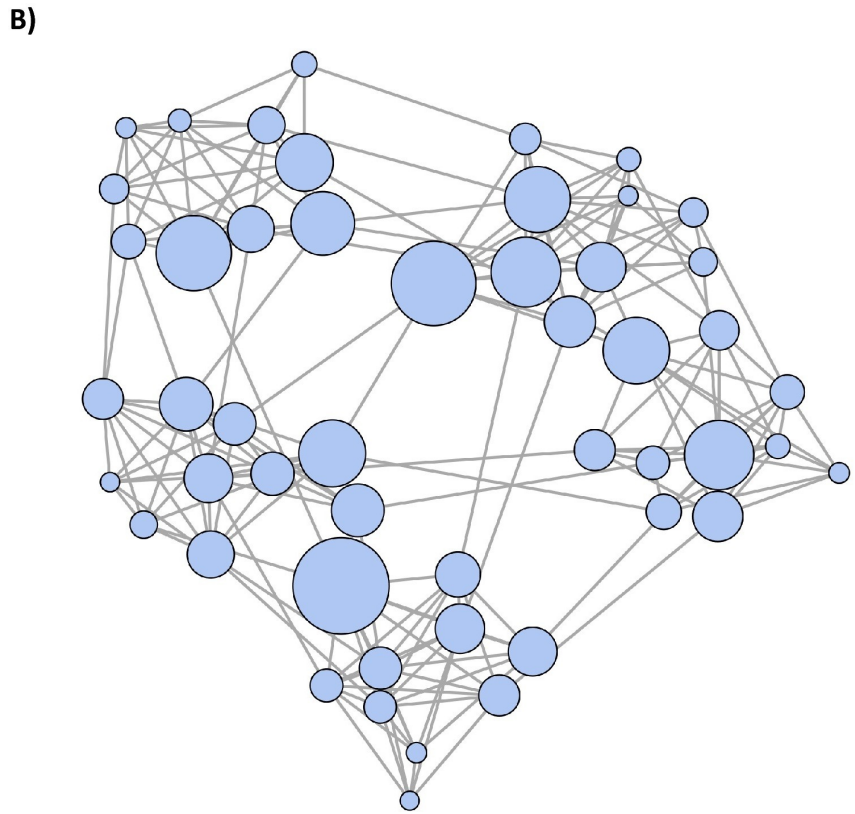
1. Avoir une jolie figure



Des réseaux pour quoi faire? (Layeghifard et al. 2017)

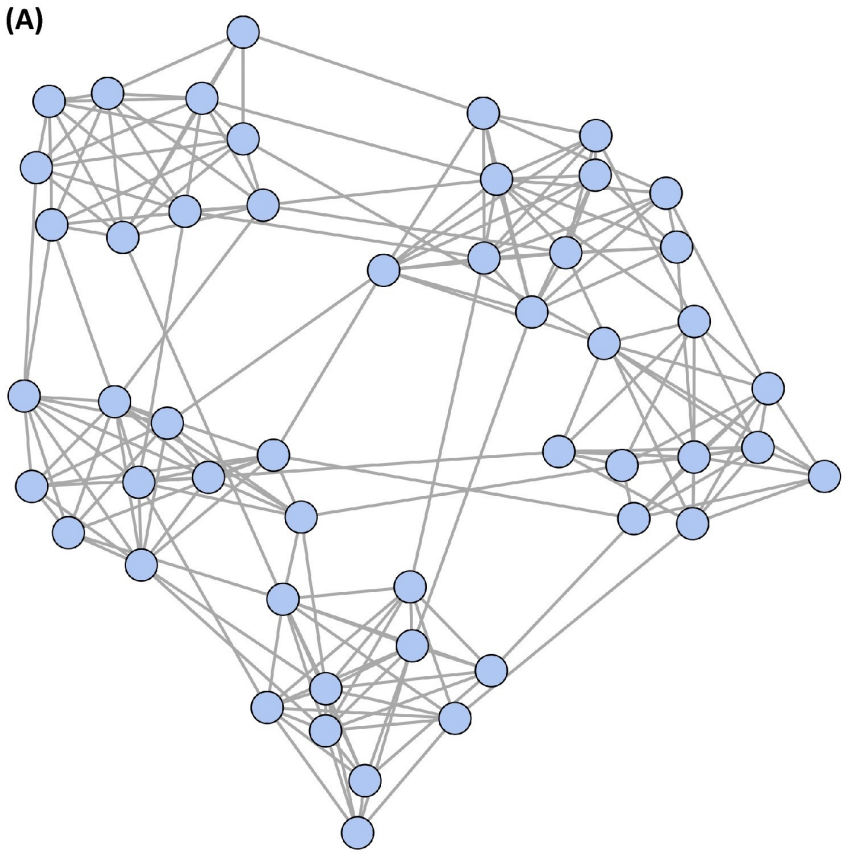
1. Avoir une jolie figure
2. Trouver des taxa systémiques

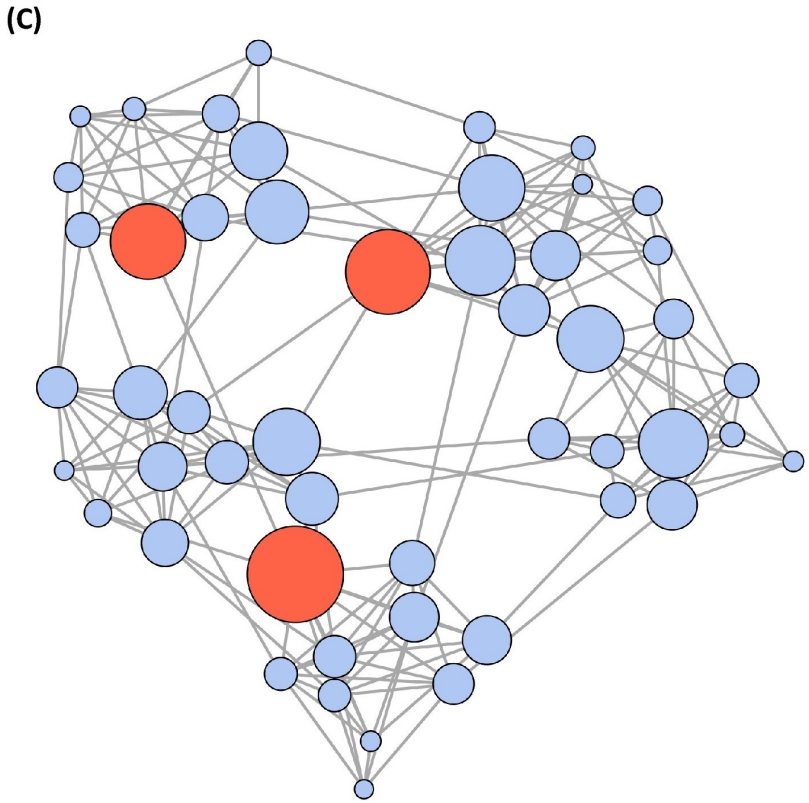




Des réseaux pour quoi faire? (Layeghifard et al. 2017)

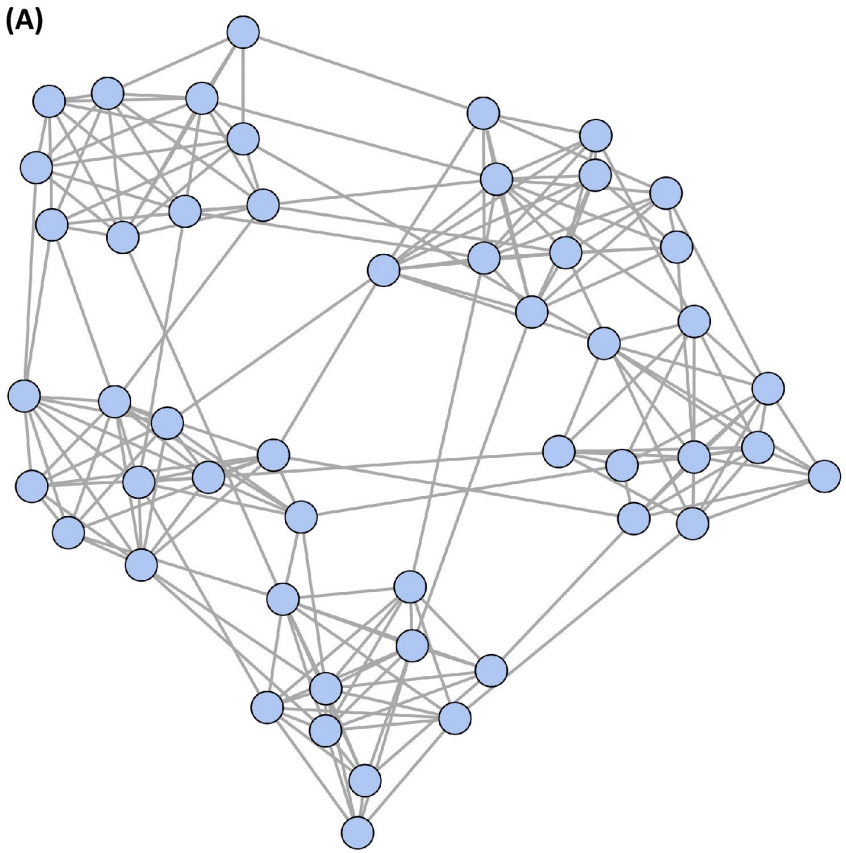
1. Avoir une jolie figure
2. Trouver des taxa systémiques
3. Trouver des taxa clé de voute (*keystone species*)

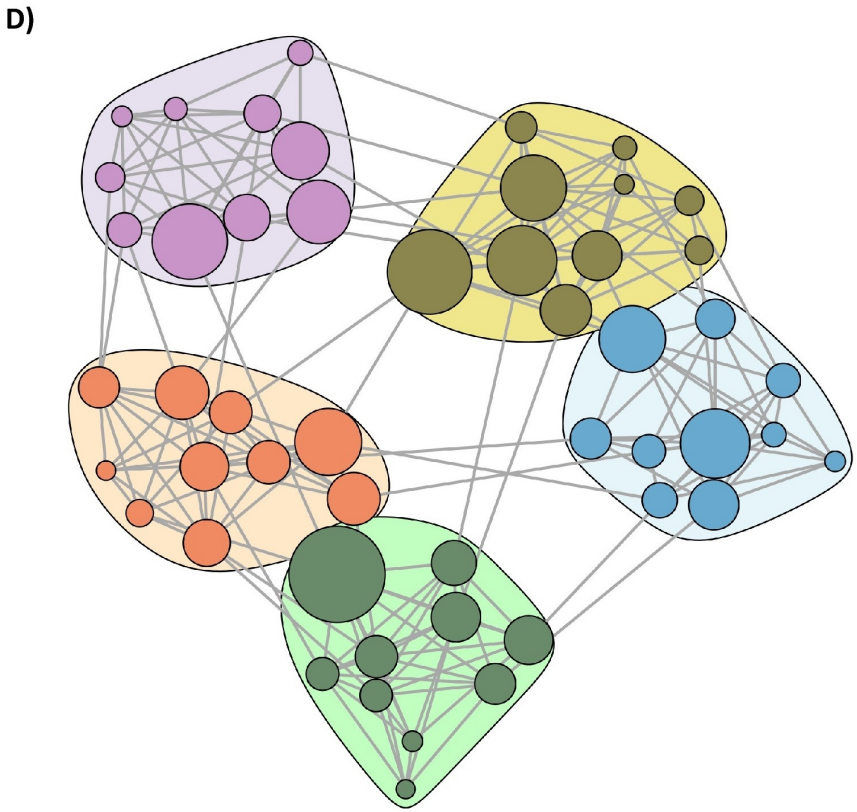




Des réseaux pour quoi faire? (Layeghifard et al. 2017)

1. Avoir une jolie figure
2. Trouver des taxa systémiques
3. Trouver des taxa clé de voute (*keystone species*)
4. Trouver des communautés de cluster de taxa en interaction





Généralités

Problématique

Évaluer la présence/absence d'une arête pour chaque **paire** de taxa possible

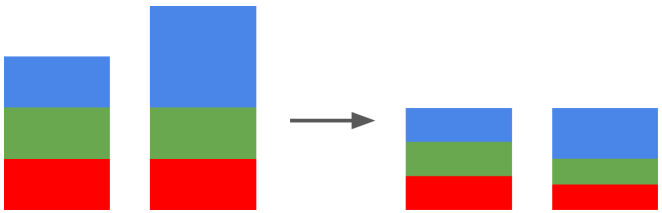
Problématique

Évaluer la présence/absence d'une arête pour chaque **paire** de taxa possible

Sparsité Beaucoup de 0 dans les tables d'abondances
$\rightarrow$ Détection de *co-absences* au détriment de taxa rares

Dimensionnalité Nombre d'OTUs $>>$ Nombre d'échantillons
$\rightarrow$ Manque de puissance

Compositionnalité Les abondances peuvent être biaisées : profondeurs de séquençage différentes, effet de la normalisation, ...



Lorsque les abondances de ces 2 communautés sont représenté de façon relative, les OTUs vert et rouge ont des abondances différentes

Plan

Une grande variété de méthodes...

Par ordre (approximatif) de complexité

1. Corrélation (Pearson, Spearman)

Une grande variété de méthodes...

Par ordre (approximatif) de complexité

1. Corrélation (Pearson, Spearman)
2. (Dis)similarité des profils d'occurence (Bray-Curtis, Kullback-Leibler, etc)

Une grande variété de méthodes...

Par ordre (approximatif) de complexité

1. Corrélation (Pearson, Spearman)
2. (Dis)similarité des profils d'occurence (Bray-Curtis, Kullback-Leibler, etc)
3. Modèles (Graphiques) Gaussiens
   - sur données *transformées* ($\log$, $\arcsin$, etc)
   - avec correction *compositionnelle* (sparCC, Rebacca, SPIEC-EASI)
   - pour données de *comptages* (Mint, PLN-network)

Une grande variété de méthodes...

Par ordre (approximatif) de complexité

3. Modèles (Graphiques) Gaussiens
   - sur données *transformées* ($\log$, $\arcsin$, etc)
   - avec correction *compositionnelle* (sparCC, Rebacca, SPIEC-EASI)
   - pour données de *comptages* (Mint, PLN-network)

...mais au final des calculs de corrélations

Interaction $\simeq$ Co-occurrence $\simeq$ Corrélation (directe/**partielle**)

...mais au final des calculs de corrélations

Interaction $\simeq$ Co-occurrence $\simeq$ Corrélation (directe/**partielle**)

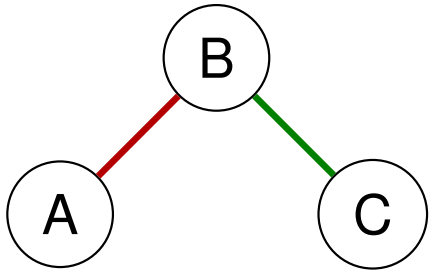
Réseau d'interaction $G \simeq$ Matrice de covariance directe (Σ) ou partielle (Σ^{-1})

...mais au final des calculs de corrélations

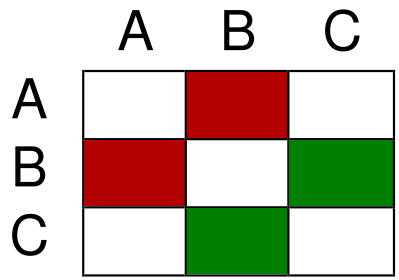
Interaction $\simeq$ Co-occurrence $\simeq$ Corrélation (directe/**partielle**)

Réseau d'interaction $G \simeq$ Matrice de covariance directe (Σ) ou partielle (Σ^{-1})

Réseau G



Matrice Σ ou Σ^{-1}



Modèles (graphiques) pour données de comptages

Approche compositionnelle

- Modéliser les proportions (à la Aitchinson)
- Ajouter une couche d'échantillonnage (Multinomiale)

Approche directe

- Modéliser directement les comptages (avec un paramètre pour la profondeur)
- Modèle graphique Poisson (Inouye et al., 2017)
- Espace latent et émission poissonnienne

Plan

Données compositionnelles

Definition

1. Données positives (ou nulles) sommant à 1

$$\sum_{j=1}^{p} x_j = 1; \quad \mathbf{x} = (x_1, \ldots, x_p) \in \mathbb{S}^{p-1}$$

2. Description des parties d'un tout

Données compositionnelles

Definition

1. Données positives (ou nulles) sommant à 1

$$\sum_{j=1}^{p} x_j = 1; \quad \mathbf{x} = (x_1, \ldots, x_p) \in \mathbb{S}^{p-1}$$

2. Description des parties d'un tout

- information relative entre les parties
- Si les x_j sont des proportions correspondants à des abondances absolues b_j, on a

$$\log(x_j/x_{j'}) = \log(b_j/b_{j'})$$

Transformation *clr*

La transformation centered log-ratio (clr) permet de se ramener à $\mathbb{H}^{p-1} = \{\boldsymbol{x}^\intercal \in \mathbb{R}^p : \boldsymbol{x}\mathbf{1} = 0\}$

Definition

$$clr : \begin{cases} \mathbb{S}^{p-1} & \to \mathbb{H}^{p-1} \subset \mathbb{R}^p \\ \mathbf{x} & \mapsto \boldsymbol{y} = clr(\boldsymbol{x}) = (x_1/m_g(\boldsymbol{x}), \ldots, x_p/m_g(\boldsymbol{x})) \end{cases}$$

où $m_g(\boldsymbol{x})$ est la moyenne géométrique de $\boldsymbol{x}$.

Si les x_j sont des proportions correspondants à des abondances absolues b_j, on a

$$clr(\boldsymbol{x}) = clr(\boldsymbol{b})$$

Modèle Multinomial lognormal

Abondances	$\mathbb{R}^p$	$\boldsymbol{b}_i$ i.i.d.	$\log(\boldsymbol{b}_i) \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
Proportions	$\mathbb{S}^{p-1}$	$\boldsymbol{x}_i = clr^{-1}(clr(\boldsymbol{b}_i))$	$clr(\boldsymbol{p}_i) \sim \mathcal{N}_p(clr(\boldsymbol{\mu}), \boldsymbol{\Gamma} = \boldsymbol{G}\boldsymbol{\Sigma}\boldsymbol{G})$
Comptages	$\mathbb{N}^p$	$\boldsymbol{n}_i$	$\boldsymbol{n}_i \sim \mathcal{M}(N_i, \boldsymbol{x}_i)$

Où $\boldsymbol{G} = \boldsymbol{I}_p - \frac{1}{p}\mathbf{1}^\intercal\mathbf{1}$ est la matrice de centrage.

Modèle Multinomial lognormal

Abondances	$\mathbb{R}^p$	$\boldsymbol{b}_i$ i.i.d.	$\log(\boldsymbol{b}_i) \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
Proportions	$\mathbb{S}^{p-1}$	$\boldsymbol{x}_i = clr^{-1}(clr(\boldsymbol{b}_i))$	$clr(\boldsymbol{p}_i) \sim \mathcal{N}_p(clr(\boldsymbol{\mu}), \boldsymbol{\Gamma} = \boldsymbol{G}\boldsymbol{\Sigma}\boldsymbol{G})$
Comptages	$\mathbb{N}^p$	$\boldsymbol{n}_i$	$\boldsymbol{n}_i \sim \mathcal{M}(N_i, \boldsymbol{x}_i)$

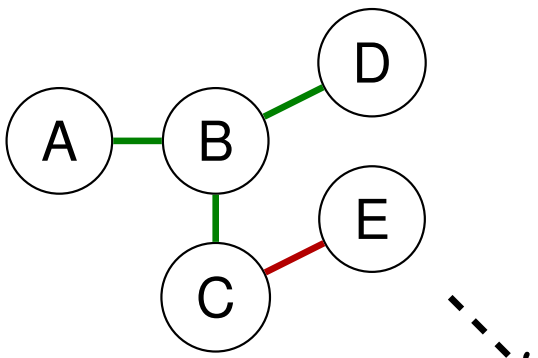
Où $\boldsymbol{G} = \boldsymbol{I}_p - \frac{1}{p}\mathbf{1}^\intercal\mathbf{1}$ est la matrice de centrage.

Le réseau est donnée par $\boldsymbol{\Sigma}$ / $\boldsymbol{\Sigma}^{-1}$

Modèle Multinomial lognormal (II)

Abondances
de bases $\boldsymbol{b}$

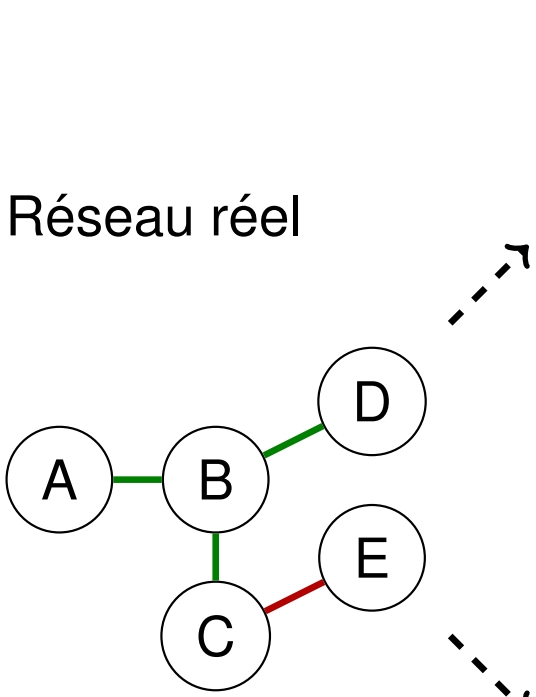
Réseau réel



$$\begin{pmatrix} 8 \\ 8 \\ 8 \\ 9 \\ 2 \end{pmatrix} \qquad \begin{pmatrix} 6 \\ 6 \\ 6 \\ 8 \\ 1 \end{pmatrix}$$

- Les abondances dans la communauté i sont gouvernées par des abondances de bases $\boldsymbol{b}_i$ log-normales $\log(\boldsymbol{b}_i) \sim \mathcal{N}(\mu, \boldsymbol{\Sigma})$

Modèle Multinomial lognormal (II)



Réseau réel

Abondances de bases $\boldsymbol{b}$

Abondances relatives $\boldsymbol{x}$

$$\begin{pmatrix} 8 \\ 8 \\ 8 \\ 9 \\ 2 \end{pmatrix} \longrightarrow \begin{pmatrix} 0.17 \\ 0.17 \\ 0.17 \\ 0.47 \\ 0.02 \end{pmatrix}$$

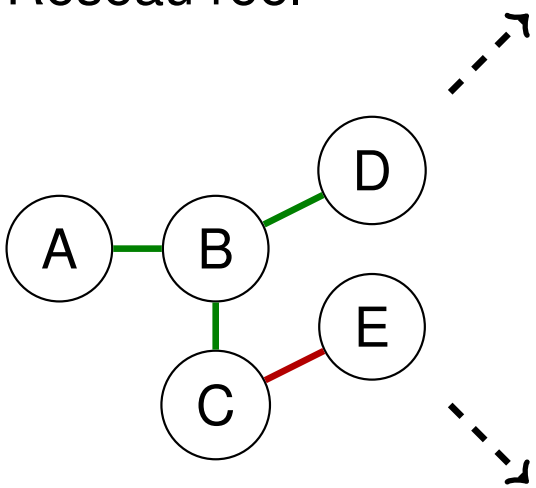
$$\begin{pmatrix} 6 \\ 6 \\ 6 \\ 8 \\ 1 \end{pmatrix} \longrightarrow \begin{pmatrix} 0.09 \\ 0.09 \\ 0.09 \\ 0.71 \\ 0.02 \end{pmatrix}$$

- L'abondance relative (ou proportion) x_{ij} du taxon j dans l'échantillon i se déduit de $\boldsymbol{b}_i$ par une transformation logistique:

$$x_{ij} = \frac{e^{b_{ij}}}{\sum_j e^{b_{ij}}}$$

Modèle Multinomial lognormal (II)

Réseau réel



Abondances de bases $\boldsymbol{b}$

$$\begin{pmatrix} 8 \\ 8 \\ 8 \\ 9 \\ 2 \end{pmatrix} \qquad \begin{pmatrix} 6 \\ 6 \\ 6 \\ 8 \\ 1 \end{pmatrix}$$

Abondances relatives $\boldsymbol{x}$

$$\begin{pmatrix} 0.17 \\ 0.17 \\ 0.17 \\ 0.47 \\ 0.02 \end{pmatrix} \qquad \begin{pmatrix} 0.09 \\ 0.09 \\ 0.09 \\ 0.71 \\ 0.02 \end{pmatrix}$$

Tableau de comptage

	1	2
A	3	8
B	3	6
C	5	2
D	8	30
E	1	4
N	20	50

- Les comptages $\mathbf{n}_i = (n_{i1}, \ldots, n_{ip})$ des différents taxa dans l'échantillon i suivent alors une loi multinomiale de paramètres N_i (profondeur de séquençage) et $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})$.

$$\mathbf{n}_i \sim \mathcal{M}(N_i, \boldsymbol{x}_i)$$

Matrice de variation

Definition

Matrice de variation On considère la *matrice de variation* $\boldsymbol{T}$ définie par

$$T_{jj'} = \frac{1}{2} Var(\log(x_j) - \log(x_{j'})) = \frac{1}{2} Var(\log(b_j) - \log(b_{j'}))$$

Remarques

- $\boldsymbol{T}$ peut-être estimée par sa version empirique $\hat{\boldsymbol{T}}$.
- $T_{ij} = \frac{1}{2}(\sigma_{ii} + \sigma_{jj}) - \sigma_{ij}$
- Si on connaissait les σ_{ii}, on aurait un estimateur de σ_{ij}.

Inférence: REBACCA, SparCC

Description

- **Objectif:** Reconstruire Σ
- **Méthode:** Méthode des moments glorifiée

Inférence: REBACCA, SparCC

Description

- **Objectif:** Reconstruire Σ
- **Méthode:** Méthode des moments glorifiée

1. Transformation des $\boldsymbol{n}_i$ en $\boldsymbol{x}_i$ via clr (avec pseudocomptage éventuel)

Inférence: REBACCA, SparCC

Description

- **Objectif:** Reconstruire Σ
- **Méthode:** Méthode des moments glorifiée

1. Transformation des $\boldsymbol{n}_i$ en $\boldsymbol{x}_i$ via clr (avec pseudocomptage éventuel)
2. Calcul de $\hat{\boldsymbol{T}}$ à partir des $\boldsymbol{x}_i$

Inférence: REBACCA, SparCC

Description

- **Objectif:** Reconstruire Σ
- **Méthode:** Méthode des moments glorifiée

1. Transformation des $\boldsymbol{n}_i$ en $\boldsymbol{x}_i$ via clr (avec pseudocomptage éventuel)
2. Calcul de $\hat{\boldsymbol{T}}$ à partir des $\boldsymbol{x}_i$
3. Écriture d'un système de $p(p-1)/2$ équations linéaires à $p(p-1)/2$ inconnues mais de rang $p(p-1)/2 - p$ reliant les σ_{ij} et les $\hat{T}_{ij}$:

$$\boldsymbol{A}\text{vech}(\boldsymbol{\Sigma}) = \text{vech}(\boldsymbol{T})$$

Inférence: REBACCA, SparCC

Description

- **Objectif:** Reconstruire Σ
- **Méthode:** Méthode des moments glorifiée

1. Transformation des $\boldsymbol{n}_i$ en $\boldsymbol{x}_i$ via clr (avec pseudocomptage éventuel)
2. Calcul de $\hat{\boldsymbol{T}}$ à partir des $\boldsymbol{x}_i$
3. Écriture d'un système de $p(p-1)/2$ équations linéaires à $p(p-1)/2$ inconnues mais de rang $p(p-1)/2 - p$ reliant les σ_{ij} et les $\hat{T}_{ij}$:

$$\boldsymbol{A}\text{vech}(\boldsymbol{\Sigma}) = \text{vech}(\boldsymbol{T})$$

4. Résolution du système sous l'hypothèse que

Inférence: REBACCA, SparCC

Description

- **Objectif:** Reconstruire Σ
- **Méthode:** Méthode des moments glorifiée

1. Transformation des $\boldsymbol{n}_i$ en $\boldsymbol{x}_i$ via clr (avec pseudocomptage éventuel)
2. Calcul de $\hat{\boldsymbol{T}}$ à partir des $\boldsymbol{x}_i$
3. Écriture d'un système de $p(p-1)/2$ équations linéaires à $p(p-1)/2$ inconnues mais de rang $p(p-1)/2 - p$ reliant les σ_{ij} et les $\hat{T}_{ij}$:

$$\boldsymbol{A}\text{vech}(\boldsymbol{\Sigma}) = \text{vech}(\boldsymbol{T})$$

4. Résolution du système sous l'hypothèse que

SparCC $\sum_{j \neq i} \sigma_{ij} \ll (p-1)\sigma_{ii} + \sum_{j \neq i} \sigma_{jj}$ (faibles corrélations)
$\Longrightarrow$ estimateur facile des σ_{ii} et plug-in pour les σ_{ij}

Inférence: REBACCA, SparCC

Description

- **Objectif:** Reconstruire Σ
- **Méthode:** Méthode des moments glorifiée

1. Transformation des $\boldsymbol{n}_i$ en $\boldsymbol{x}_i$ via clr (avec pseudocomptage éventuel)
2. Calcul de $\hat{\boldsymbol{T}}$ à partir des $\boldsymbol{x}_i$
3. Écriture d'un système de $p(p-1)/2$ équations linéaires à $p(p-1)/2$ inconnues mais de rang $p(p-1)/2 - p$ reliant les σ_{ij} et les $\hat{T}_{ij}$:

$$\boldsymbol{A}\text{vech}(\boldsymbol{\Sigma}) = \text{vech}(\boldsymbol{T})$$

4. Résolution du système sous l'hypothèse que

SparCC $\sum_{j \neq i} \sigma_{ij} \ll (p-1)\sigma_{ii} + \sum_{j \neq i} \sigma_{jj}$ (faibles corrélations)
$\Longrightarrow$ estimateur facile des σ_{ii} et plug-in pour les σ_{ij}

Rebacca $\boldsymbol{\Sigma}$ est sparse (moins de $p/4$ coefficients non-nuls par ligne), via une régularisation ℓ_1 du système:

$$\| \arg\min_{\boldsymbol{\Sigma}} \boldsymbol{A}\text{vech}(\boldsymbol{\Sigma}) - \text{vech}(\boldsymbol{T}) \| + \lambda \|\text{vech}(\boldsymbol{\Sigma})\|_1$$

Avantages, inconvénients

Avantages

+ Méthode très simple et très rapide
+ Simple résolution d'un système linéaire

Inconvénients

- Focus sur Σ et non Σ^{-1}
- Oubli de la couche multinomiale
- Complexité en espace $O(p^4)$ (pour REBACCA)
- Choix de λ (REBACCA) ou du nombre d'itérations (SparCC)

Inférence: SPIEC-EASI

Description

- **Objectif:** Reconstruire Σ^{-1}
- **Méthode:** Se ramener à un lasso graphique

1

Inférence: SPIEC-EASI

Description

- **Objectif:** Reconstruire Σ^{-1}
- **Méthode:** Se ramener à un lasso graphique

1. Transformation des $\boldsymbol{n}_i$ en $\boldsymbol{x}_i$ via clr (avec pseudocomptage éventuel)

1

Inférence: SPIEC-EASI

Description

- **Objectif:** Reconstruire Σ^{-1}
- **Méthode:** Se ramener à un lasso graphique

1. Transformation des $\boldsymbol{n}_i$ en $\boldsymbol{x}_i$ via clr (avec pseudocomptage éventuel)
2. Supposer que p est grand donc que $\boldsymbol{G} \simeq \boldsymbol{I}_p$ et $\boldsymbol{\Gamma} = \boldsymbol{G\Sigma G} \simeq \boldsymbol{\Sigma}$

1

Inférence: SPIEC-EASI

Description

- **Objectif:** Reconstruire Σ^{-1}
- **Méthode:** Se ramener à un lasso graphique

1. Transformation des $\boldsymbol{n}_i$ en $\boldsymbol{x}_i$ via clr (avec pseudocomptage éventuel)
2. Supposer que p est grand donc que $\boldsymbol{G} \simeq \boldsymbol{I}_p$ et $\boldsymbol{\Gamma} = \boldsymbol{G\Sigma G} \simeq \boldsymbol{\Sigma}$
3. Utiliser $\hat{\boldsymbol{\Gamma}}$ en lieu et place de $\hat{\boldsymbol{\Sigma}}$

1

Inférence: SPIEC-EASI

Description

- **Objectif:** Reconstruire Σ^{-1}
- **Méthode:** Se ramener à un lasso graphique

1. Transformation des $\boldsymbol{n}_i$ en $\boldsymbol{x}_i$ via clr (avec pseudocomptage éventuel)
2. Supposer que p est grand donc que $\boldsymbol{G} \simeq \boldsymbol{I}_p$ et $\boldsymbol{\Gamma} = \boldsymbol{G}\boldsymbol{\Sigma}\boldsymbol{G} \simeq \boldsymbol{\Sigma}$
3. Utiliser $\hat{\boldsymbol{\Gamma}}$ en lieu et place de $\hat{\boldsymbol{\Sigma}}$
4. Insérer dans un lasso graphique[1] pour estimer $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$

$$\hat{\boldsymbol{\Omega}} = \underset{\boldsymbol{\Omega} \in \mathbb{S}_{++}}{\arg\min} - \log\det \boldsymbol{\Omega} + \mathsf{Tr}(\boldsymbol{\Omega}\hat{\boldsymbol{\Gamma}}) + \lambda \|\boldsymbol{\Omega}\|_1$$

[1] Variante possible avec l'estimation de voisinage à la Meinshausen et Bühlmann

Avantages, inconvénients

Avantages

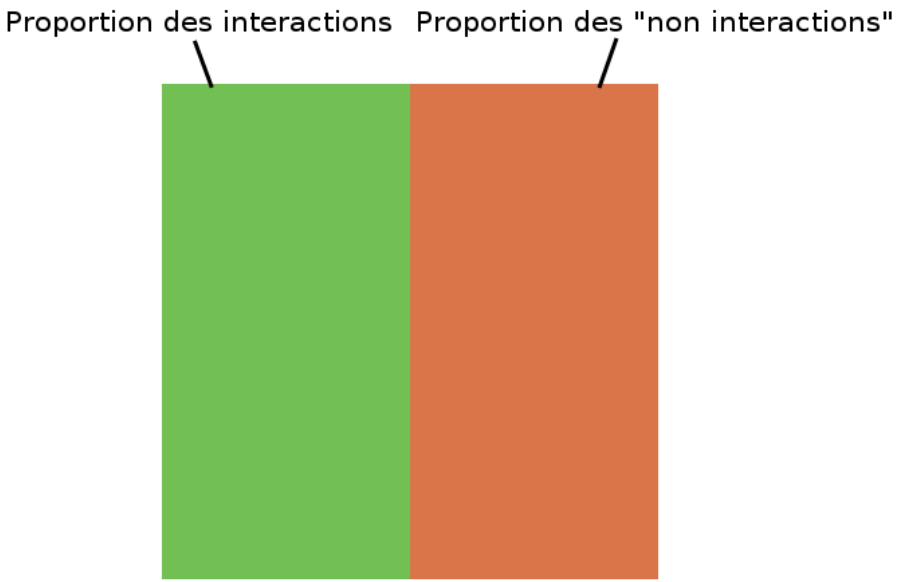
+ Focus sur $\mathbf{\Omega} = \mathbf{\Sigma}^{-1}$
+ Basée sur une méthode de vraisemblance pénalisée

Inconvénients

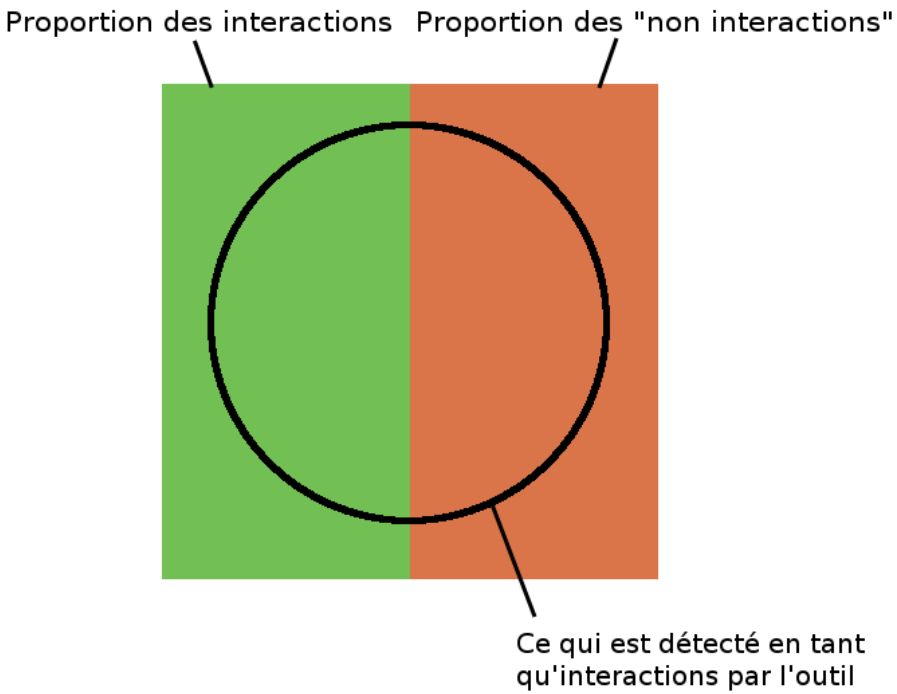
- Oubli de la couche multinomiale
- Choix de λ (REBACCA)
- Relativement long

Mesures de la performance

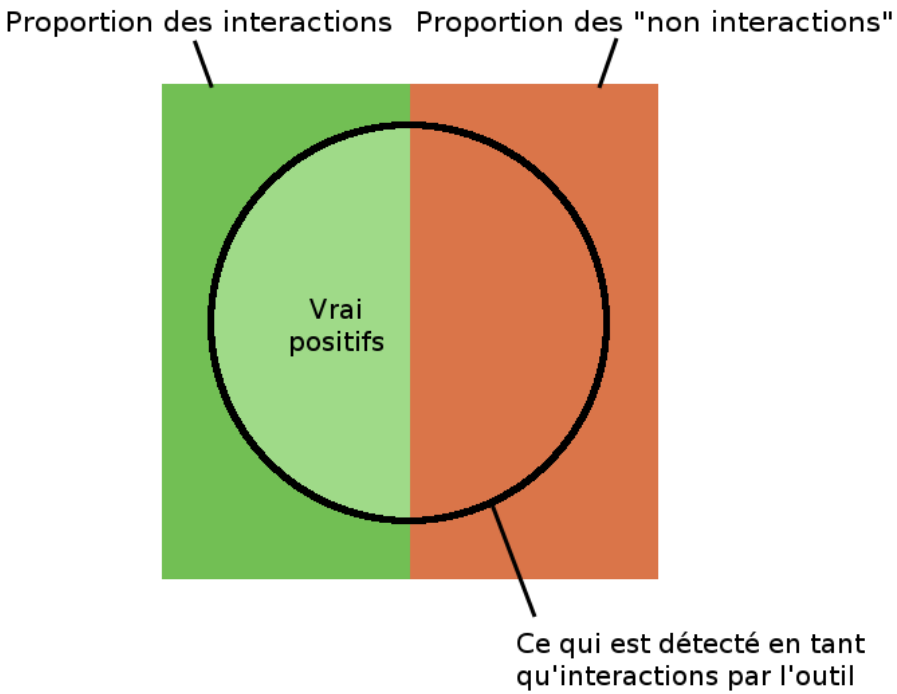
Mesures de la performance



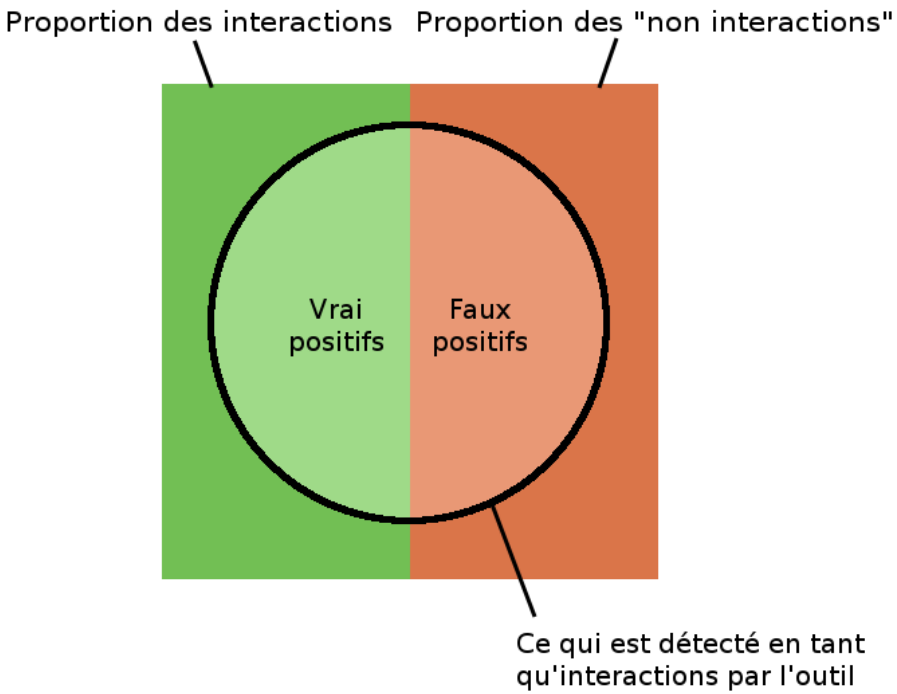
Mesures de la performance



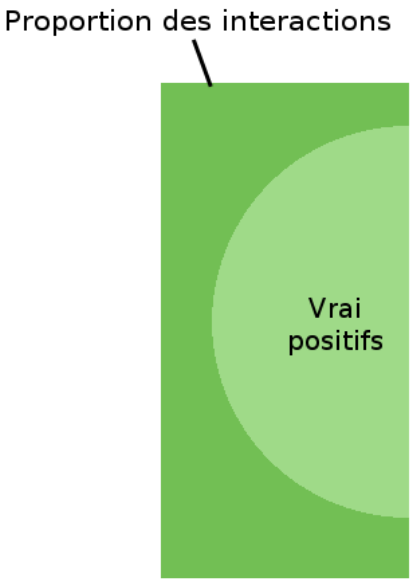
Mesures de la performance



Mesures de la performance



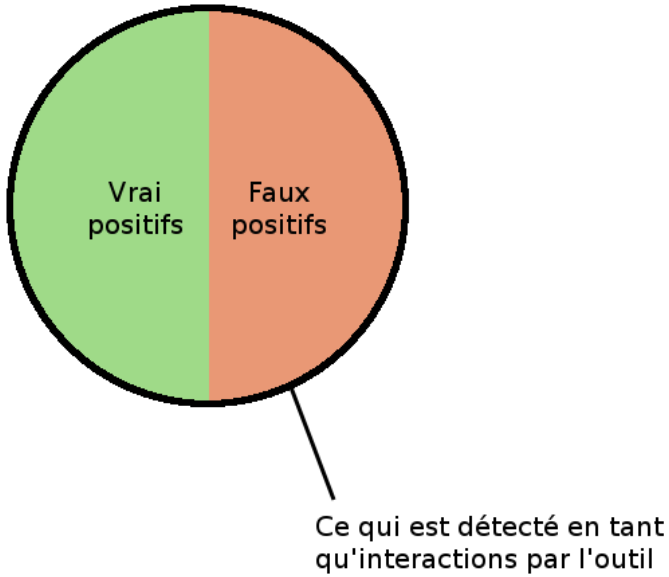
Mesures de la performance



$$\frac{Vrai\ positifs}{Positifs}$$

Sensibilité/Rappel Proportion des interactions du réseau ayant été détectées par l'outil

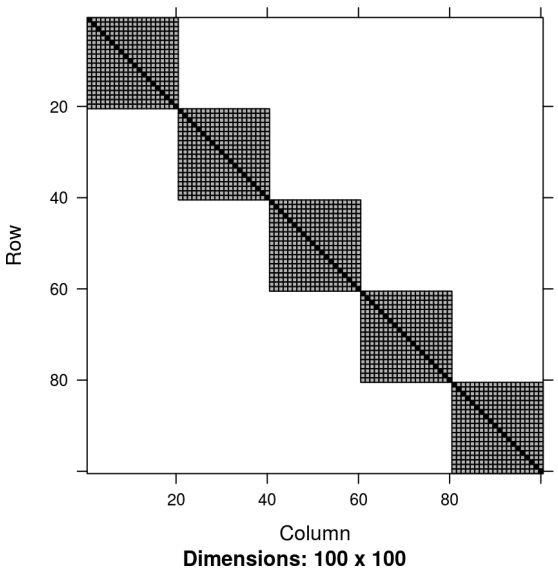
Mesures de la performance



$$\frac{Vrai\ positifs}{Vrai\ positifs + Faux\ positifs}$$

Précision Proportion des interactions ayant été correctement détectées parmi ce que l'outil a détecté

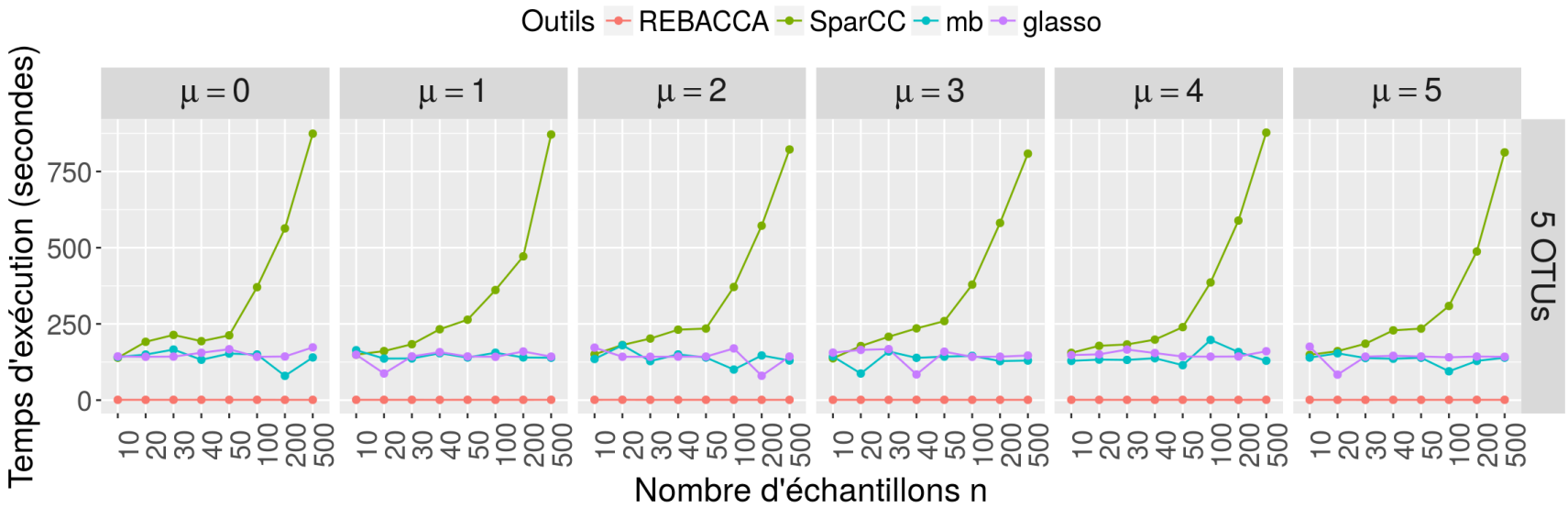
Schéma de simulation



- p Nombre d'OTUs
  $\rightarrow$ Réseaux de taille variable, passage à l'échelle
- μ Gamme des abondances des OTUs
  $\rightarrow$ Détection des taxa rares
- ρ Force moyenne des corrélations
  $\rightarrow$ Détection d'interactions faibles (puissance de l'outil)
- k Matrice diagonale par blocs de k blocs : Sparsité du réseau
  $\rightarrow$ Robustesse aux données non sparses

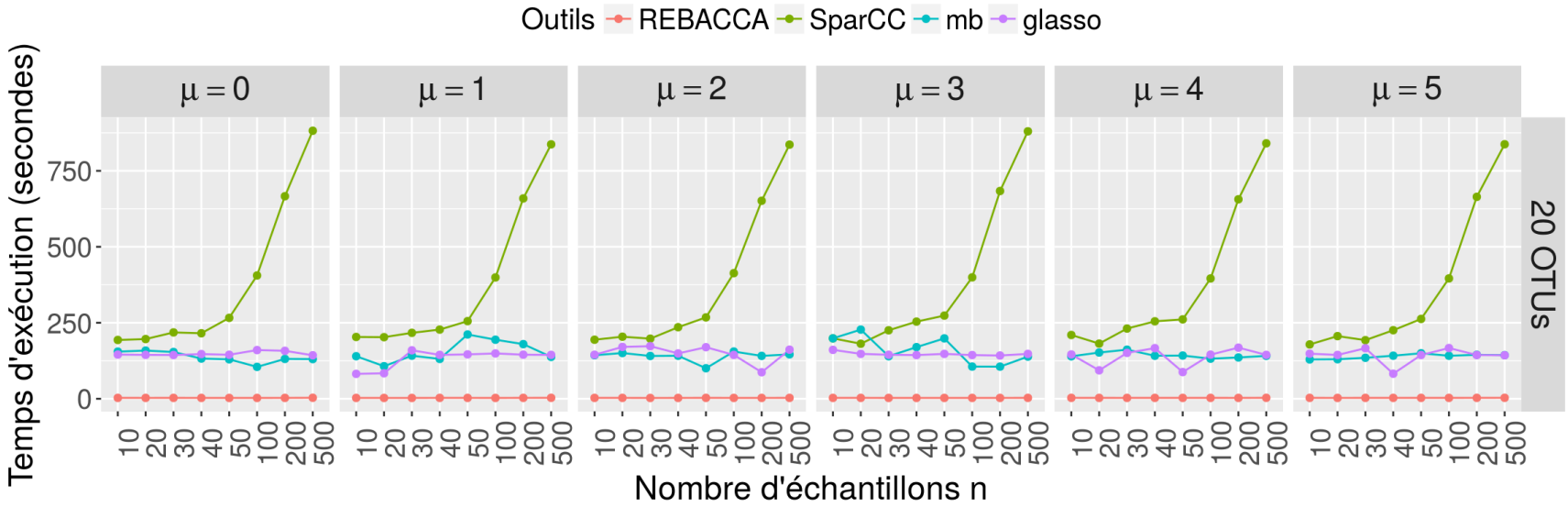
Temps d'exécution

SparCC est le plus lent pour des petits réseaux (12 minutes contre 4 minutes)



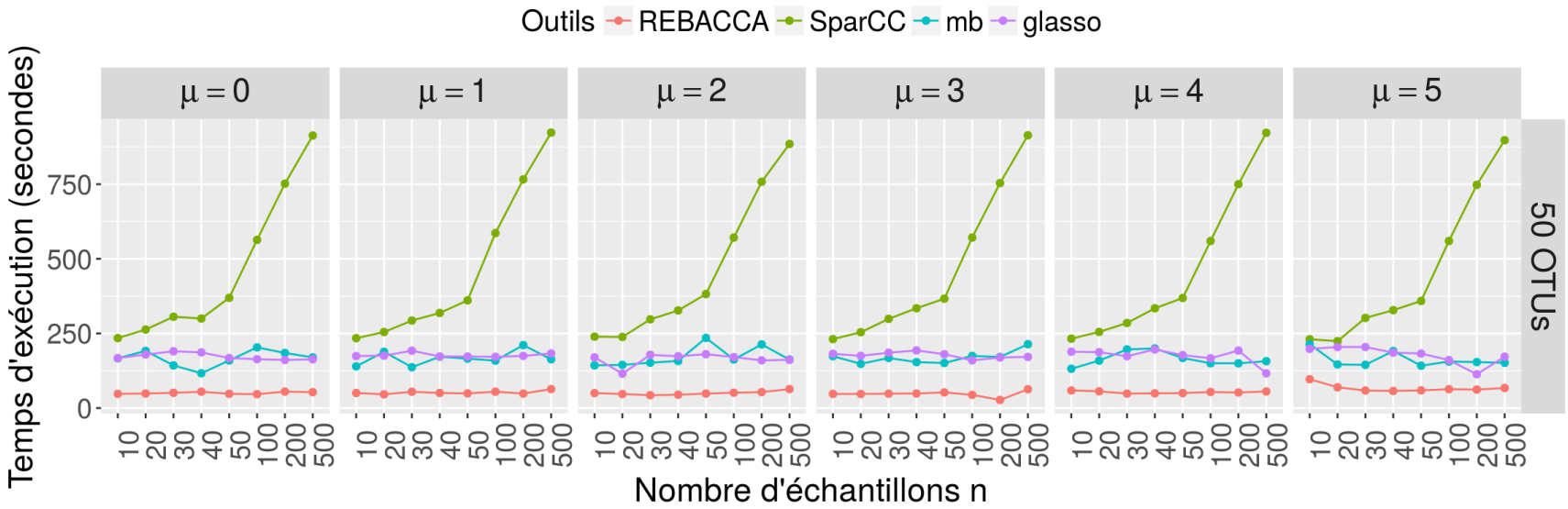
Temps d'exécution

SparCC est le plus lent pour des petits réseaux (12 minutes contre 4 minutes)



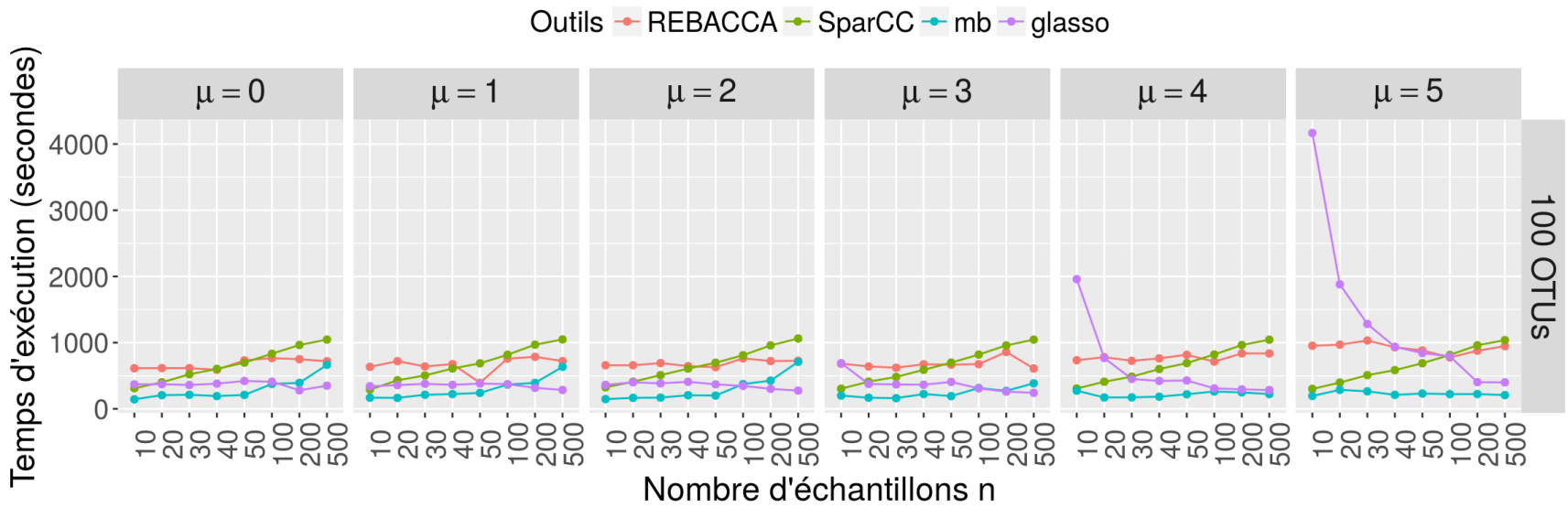
Temps d'exécution

SparCC est le plus lent pour des petits réseaux (12 minutes contre 4 minutes)



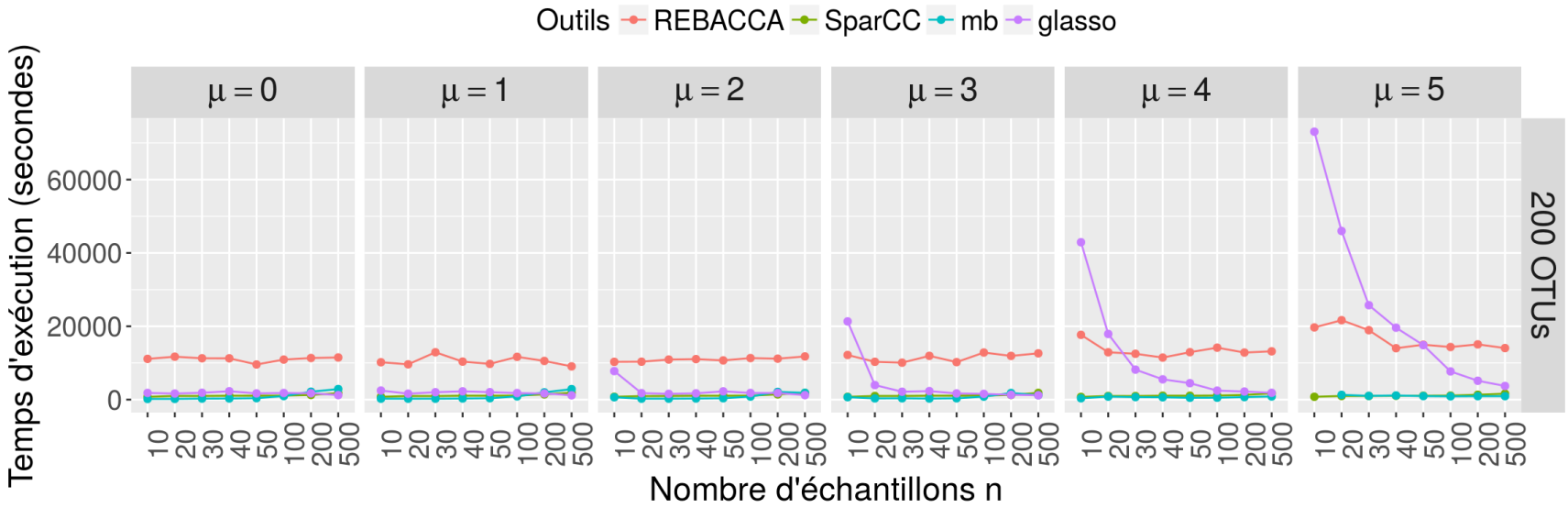
Temps d'exécution

SparCC possède les meilleurs temps sur les larges réseaux
glasso manque de puissance lorsque n est petit $(10 \dots 50)$:
2,5x plus lent que les autres méthodes



Temps d'exécution

SparCC possède les meilleurs temps sur les larges réseaux
glasso manque de puissance lorsque n est petit $(10 \dots 50)$:
Presque **5x plus lent** que les autres méthodes (19 heures contre 4 heures)



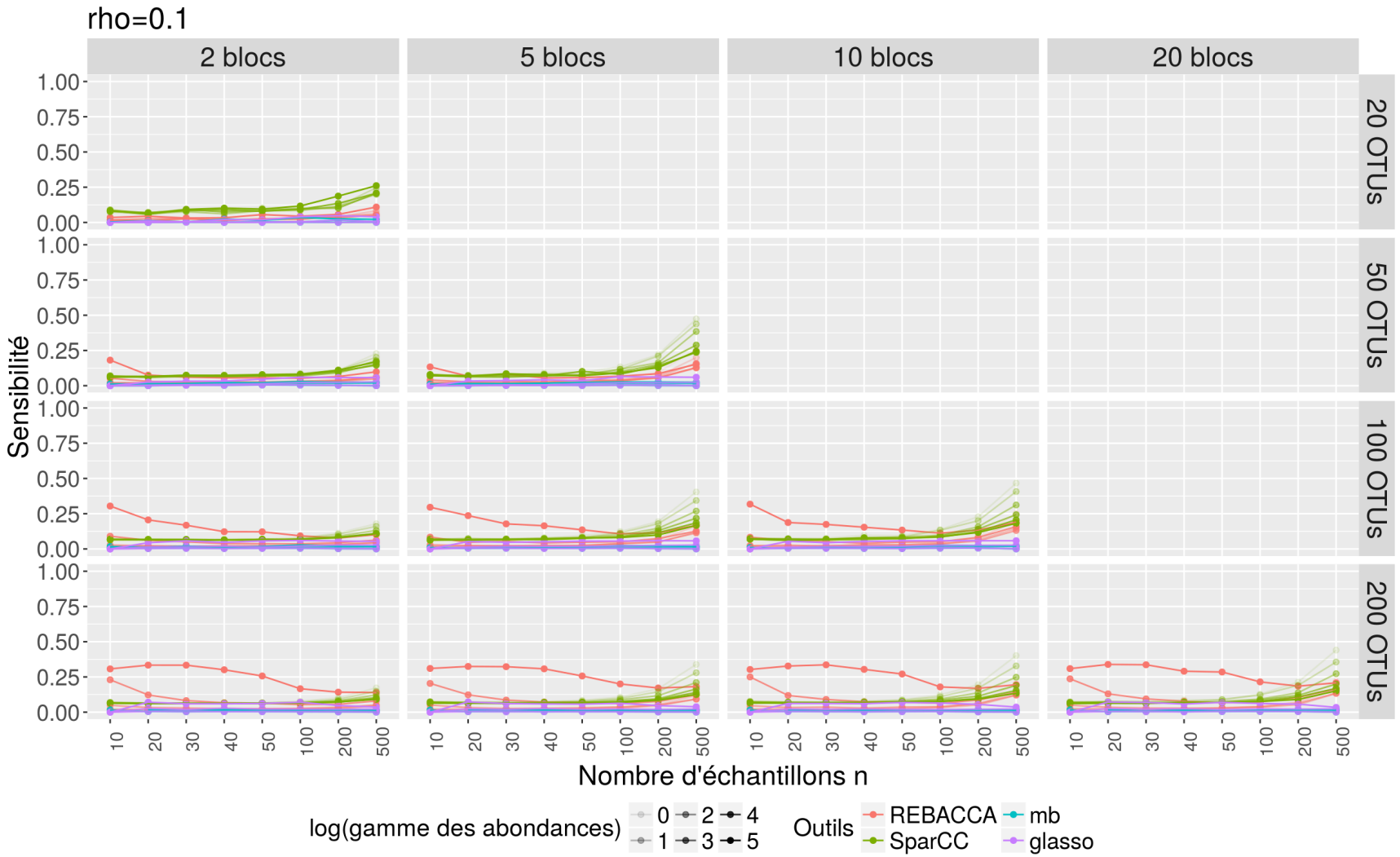
Limites en mémoire

- REBACCA est limité à 250 OTUs
  $\rightarrow$ Complexité en mémoire en $O(D^4)$

- Les simulations très difficiles n'aboutissent pas avec la méthode SPIEC-EASI (MB)
  $\rightarrow$ Compensation du manque de puissance pour détecter les taxa rares

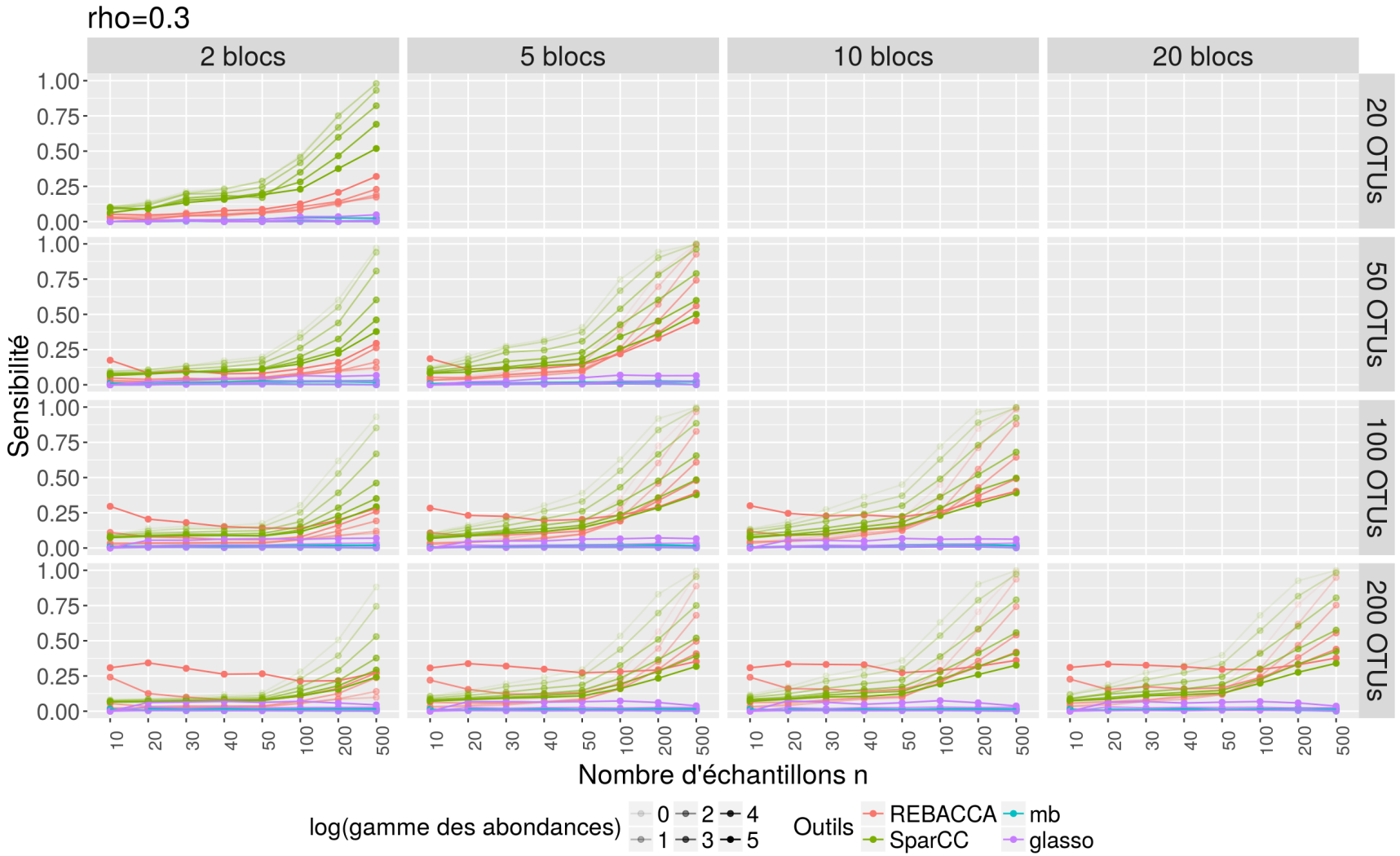
Réseaux de densité variable : Sensibilité

Mauvaise sensibilité pour de **faibles** interactions ($\rho = 0.1$)



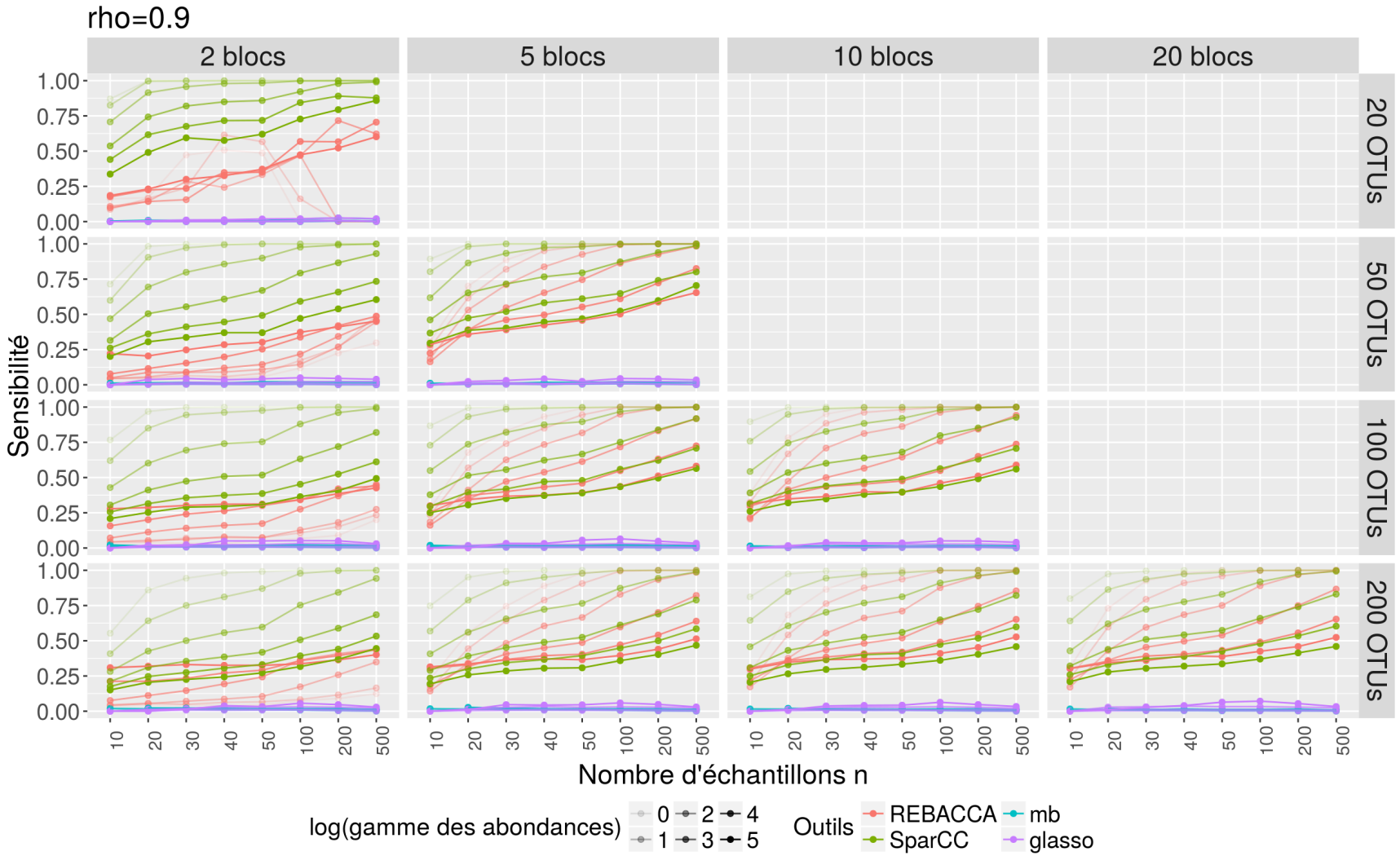
Réseaux de densité variable : Sensibilité

Bonne sensibilité lorsqu'il y a beaucoup d'échantillons à partir de $\rho = 0.3$



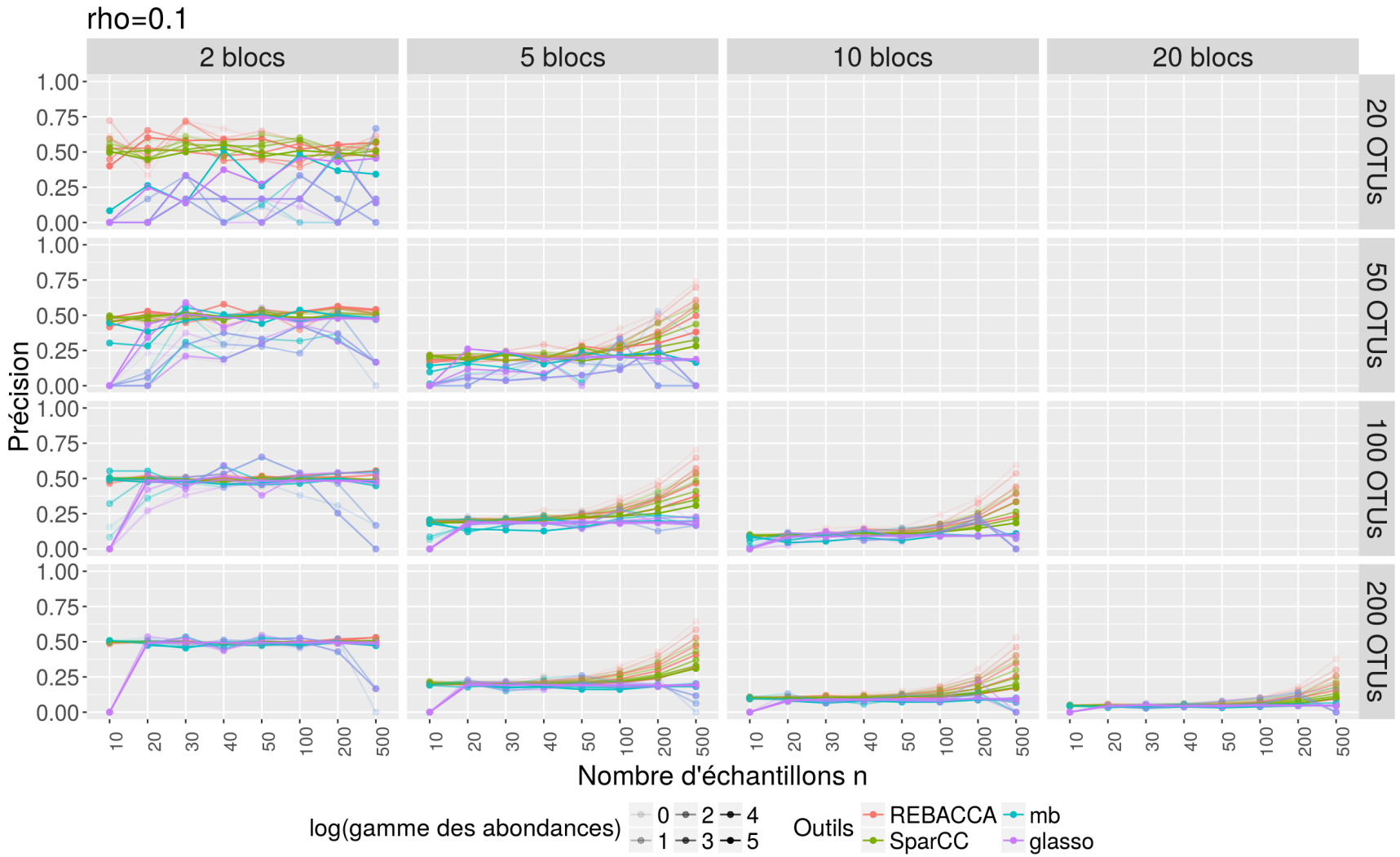
Réseaux de densité variable : Sensibilité

SparCC meilleur que REBACCA pour des réseaux denses



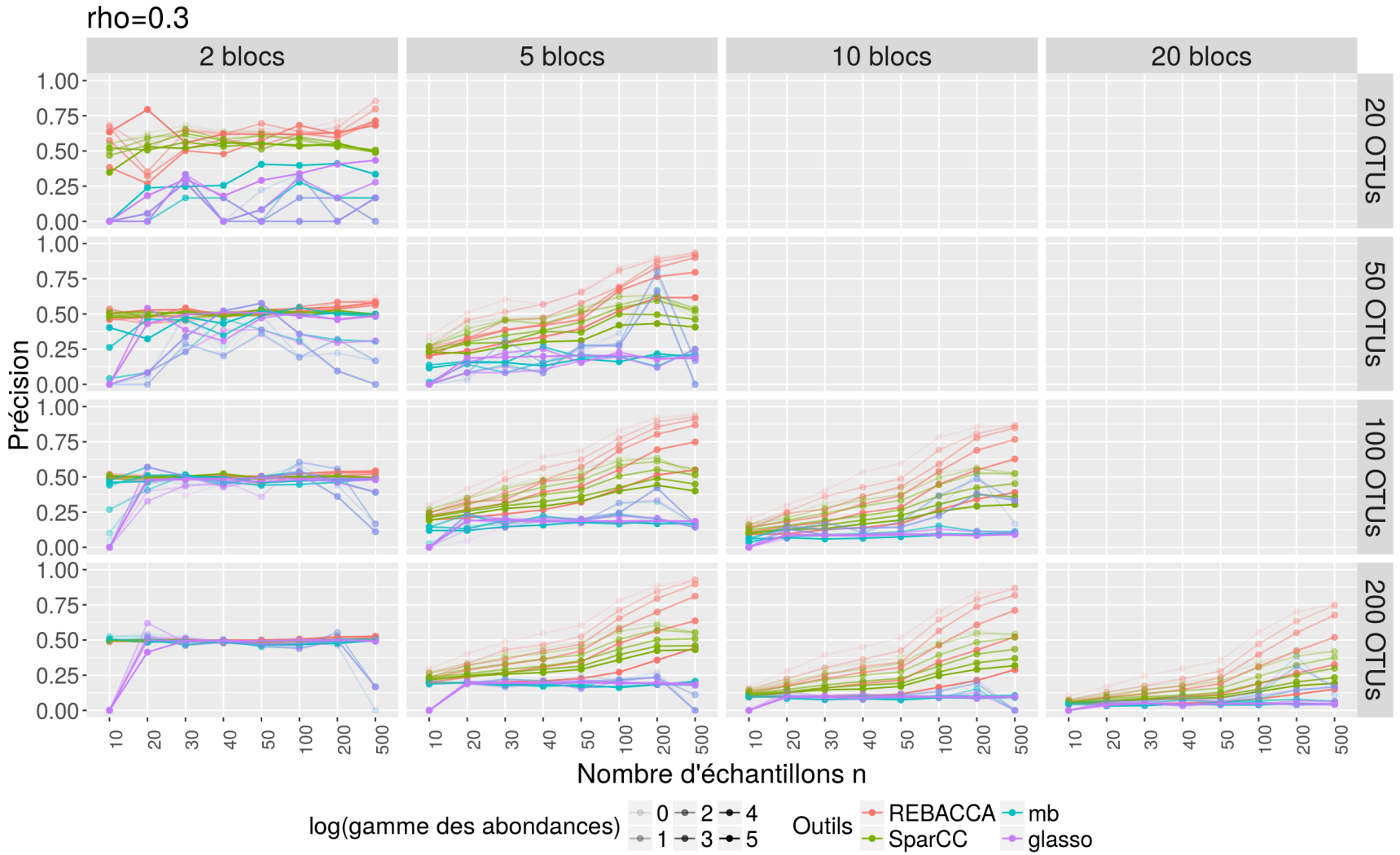
Réseaux de densité variable : Précision

Mauvaise précision pour de **faibles** interactions ($\rho = 0.1$)



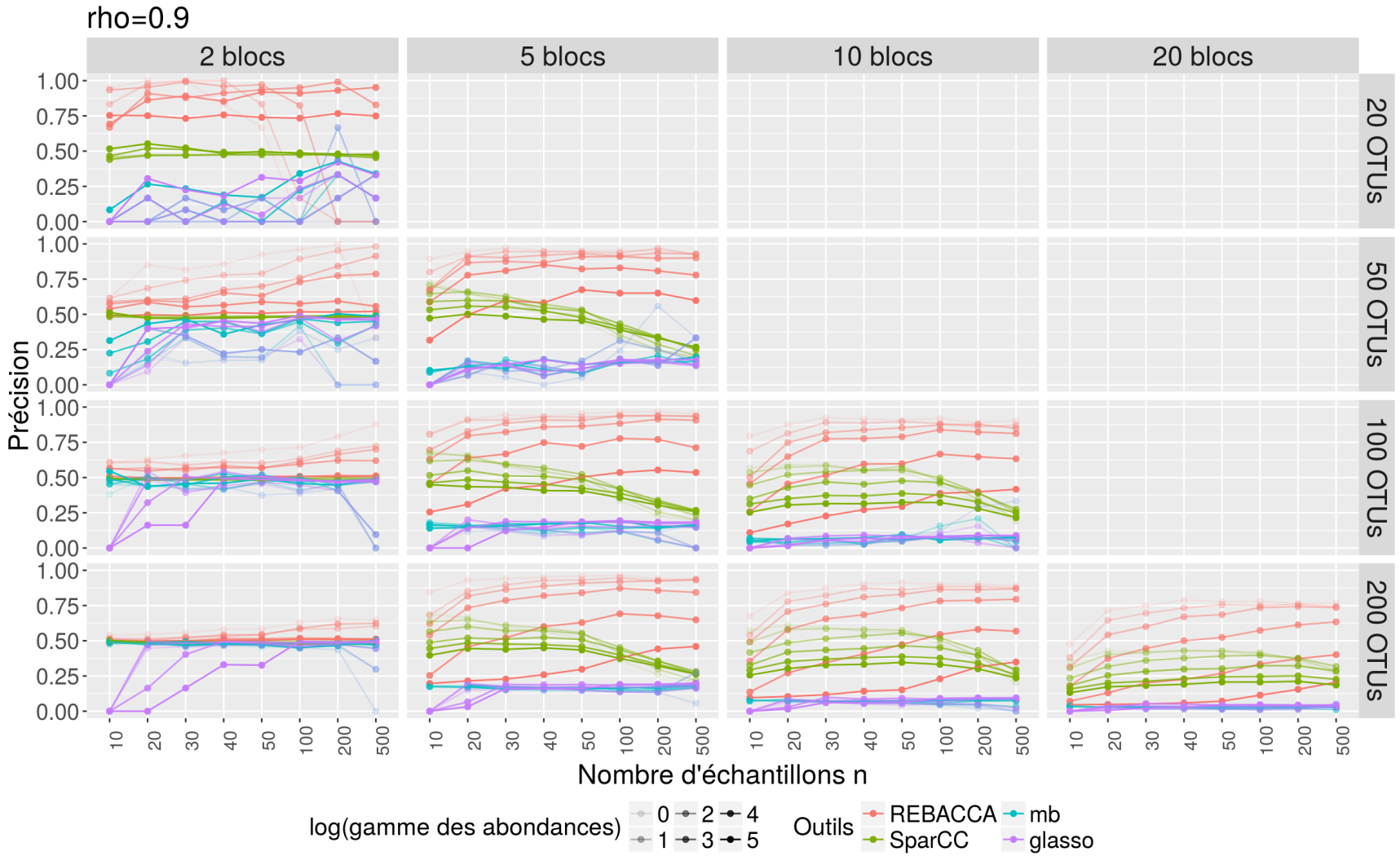
Réseaux de densité variable : Précision

Bonne précision lorsqu'il y a beaucoup d'échantillons à partir de $\rho = 0.3$



Réseaux de densité variable : Précision

REBACCA a une meilleure précision (0.75) que SparCC (0.5)



Conclusions Partielles

- REBACCA: (Assez) **peu** d'interactions mais plutôt **fiables**
  sensibilité moyenne, bonne précision
  - $\rightarrow$ Bonnes performances, mais limité à 250 OTUs.
  - $\rightarrow$ Solutions possibles : Agglomération d'OTUs en méta-OTUs selon leur assignation taxonomique.

Conclusions Partielles

- REBACCA: (Assez) **peu** d'interactions mais plutôt **fiables**
  sensibilité moyenne, bonne précision
  - $\rightarrow$ Bonnes performances, mais limité à 250 OTUs.
  - $\rightarrow$ Solutions possibles : Agglomération d'OTUs en méta-OTUs selon leur assignation taxonomique.

- SparCC: **Beaucoup** d'interactions mais **moins** fiables
  bonne sensibilité, précision moyenne

Conclusions Partielles

- REBACCA: (Assez) **peu** d'interactions mais plutôt **fiables**
  sensibilité moyenne, bonne précision
  - → Bonnes performances, mais limité à 250 OTUs.
  - → Solutions possibles : Agglomération d'OTUs en méta-OTUs selon leur assignation taxonomique.

- SparCC: **Beaucoup** d'interactions mais **moins** fiables
  bonne sensibilité, précision moyenne

- SPIEC-EASI: Mauvais résultats
  mauvaise sensibilité, mauvaise précision
  - → Résultats très inférieurs à ceux de la publi.
  - → Modèle de simulation (dans la publi) non compositionel.

Conclusions Partielles

- REBACCA: (Assez) **peu** d'interactions mais plutôt **fiables**
  sensibilité moyenne, bonne précision
  - → Bonnes performances, mais limité à 250 OTUs.
  - → Solutions possibles : Agglomération d'OTUs en méta-OTUs selon leur assignation taxonomique.
- SparCC: **Beaucoup** d'interactions mais **moins** fiables
  bonne sensibilité, précision moyenne
- SPIEC-EASI: Mauvais résultats
  mauvaise sensibilité, mauvaise précision
  - → Résultats très inférieurs à ceux de la publi.
  - → Modèle de simulation (dans la publi) non compositionel.
- Plus **difficile** pour des gammes d'abondances **variables**
- Aucun des 3 outils ne gère les co-variables (confusion entre préférence d'habitat et interaction).

Plan

Modèle de Poisson multivarié

- Peut-on modéliser directement la loi jointe de n?

Modèle de Poisson multivarié

- Peut-on modéliser directement la loi jointe de $\boldsymbol{n}$?
- Par analogie au cas Gaussien, on aimerait avoir une loi de la forme

$$P(\boldsymbol{n}) \propto \exp\left(\boldsymbol{n}^\intercal \boldsymbol{\Omega} \boldsymbol{n} + \boldsymbol{\eta}^\intercal \boldsymbol{n} - \sum_j \log(n_j!) - B(\boldsymbol{\eta}, \boldsymbol{\Omega})\right)$$

Modèle de Poisson multivarié

- Peut-on modéliser directement la loi jointe de $\boldsymbol{n}$?
- Par analogie au cas Gaussien, on aimerait avoir une loi de la forme

$$P(\boldsymbol{n}) \propto \exp\left(\boldsymbol{n}^\intercal \boldsymbol{\Omega} \boldsymbol{n} + \boldsymbol{\eta}^\intercal \boldsymbol{n} - \sum_j \log(n_j!) - B(\boldsymbol{\eta}, \boldsymbol{\Omega})\right)$$

qui garantit que

- les lois conditionnelles de chaque comptage sont Poisson
- $\Omega_{ij} = 0$ sont conditionnellement indépendantes

Modèle de Poisson multivarié

- Peut-on modéliser directement la loi jointe de $\boldsymbol{n}$?
- Par analogie au cas Gaussien, on aimerait avoir une loi de la forme

$$P(\boldsymbol{n}) \propto \exp\left(\boldsymbol{n}^\intercal \boldsymbol{\Omega} \boldsymbol{n} + \boldsymbol{\eta}^\intercal \boldsymbol{n} - \sum_j \log(n_j!) - B(\boldsymbol{\eta}, \boldsymbol{\Omega})\right)$$

  qui garantit que
  - les lois conditionnelles de chaque comptage sont Poisson
  - $\Omega_{ij} = 0$ sont conditionnellement indépendantes
- Possible uniquement si
  - $\Omega_{jj} = 0$ pour tout j
  - $\Omega_{ij} \leq 0$ pour tout $i \neq j$

Modèle de Poisson multivarié

- Peut-on modéliser directement la loi jointe de $\boldsymbol{n}$?
- Par analogie au cas Gaussien, on aimerait avoir une loi de la forme

$$P(\boldsymbol{n}) \propto \exp\left(\boldsymbol{n}^\intercal \boldsymbol{\Omega} \boldsymbol{n} + \boldsymbol{\eta}^\intercal \boldsymbol{n} - \sum_j \log(n_j!) - B(\boldsymbol{\eta}, \boldsymbol{\Omega})\right)$$

  qui garantit que
  - les lois conditionnelles de chaque comptage sont Poisson
  - $\Omega_{ij} = 0$ sont conditionnellement indépendantes
- Possible uniquement si
  - $\Omega_{jj} = 0$ pour tout j
  - $\Omega_{ij} \leq 0$ pour tout $i \neq j$
- $\Longrightarrow$ Uniquement des interactions négatives ☹

Modèle Poisson Log-Normal (PLN)

Modèle

(log)-abondances	$\mathbb{R}^p$	$\boldsymbol{b}_i$	i.i.d.	$\boldsymbol{b}_i \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
Comptages	$\mathbb{N}^p$	$n_{ij}\|b_{ij}$	indep.	$n_{ij}\|b_{ij} \sim \mathcal{P}(e^{b_{ij}})$

Modèle Poisson Log-Normal (PLN)

Modèle

$$\begin{array}{rcccc} \text{(log)-abondances} & \mathbb{R}^p & \boldsymbol{b}_i & \text{i.i.d.} & \boldsymbol{b}_i \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \\ \text{Comptages} & \mathbb{N}^p & n_{ij}|b_{ij} & \text{indep.} & n_{ij}|b_{ij} \sim \mathcal{P}(e^{b_{ij}}) \end{array}$$

Properties:

$$\begin{aligned} \mathbb{E}(n_{ij}) &= e^{\mu_j + \Sigma_{jj}/2} =: \lambda_j \\ \mathbb{V}(n_{ij}) &= \lambda_j + \lambda_j^2 \left(e^{\Sigma_{jj}} - 1\right) && \text{(surdispersion)} \\ \mathbb{C}\text{ov}(n_{ij}, n_{ik}) &= \lambda_j \lambda_k \left(e^{\Sigma_{jk}} - 1\right) && \text{(signe arbitraire)} \end{aligned}$$

PLN Covariables et Offset

Modèle

(log)-abondances	$\mathbb{R}^p$	$\boldsymbol{b}_i$	i.i.d.	$\boldsymbol{b}_i \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
Comptages	$\mathbb{N}^p$	$n_{ij}\|b_{ij}$	indep.	$n_{ij}\|b_{ij} \sim \mathcal{P}(e^{b_{ij}})$

PLN Covariables et Offset

Modèle

$$\begin{array}{rcccl} \text{(log)-abondances} & \mathbb{R}^p & \boldsymbol{b}_i & \text{i.i.d.} & \boldsymbol{b}_i \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \\ \text{Comptages} & \mathbb{N}^p & n_{ij}|b_{ij} & \text{indep.} & n_{ij}|b_{ij} \sim \mathcal{P}(e^{b_{ij}}) \end{array}$$

Propriétés:

$$\begin{aligned} \mathbb{E}(n_{ij}) &= e^{\mu_j + \Sigma_{jj}/2} =: \lambda_j \\ \mathbb{V}(n_{ij}) &= \lambda_j + \lambda_j^2 \left(e^{\Sigma_{jj}} - 1\right) && \text{(surdispersion)} \\ \mathbb{C}\text{ov}(n_{ij}, n_{ik}) &= \lambda_j \lambda_k \left(e^{\Sigma_{jk}} - 1\right) && \text{(signe arbitraire)} \end{aligned}$$

PLN Covariables et Offset

Modèle

$$\begin{array}{rccccc} \text{(log)-abondances} & \mathbb{R}^p & \boldsymbol{b}_i & \text{i.i.d.} & \boldsymbol{b}_i \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \\ \text{Comptages} & \mathbb{N}^p & n_{ij}|b_{ij} & \text{indep.} & n_{ij}|b_{ij} \sim \mathcal{P}(e^{b_{ij}}) \end{array}$$

Propriétés:

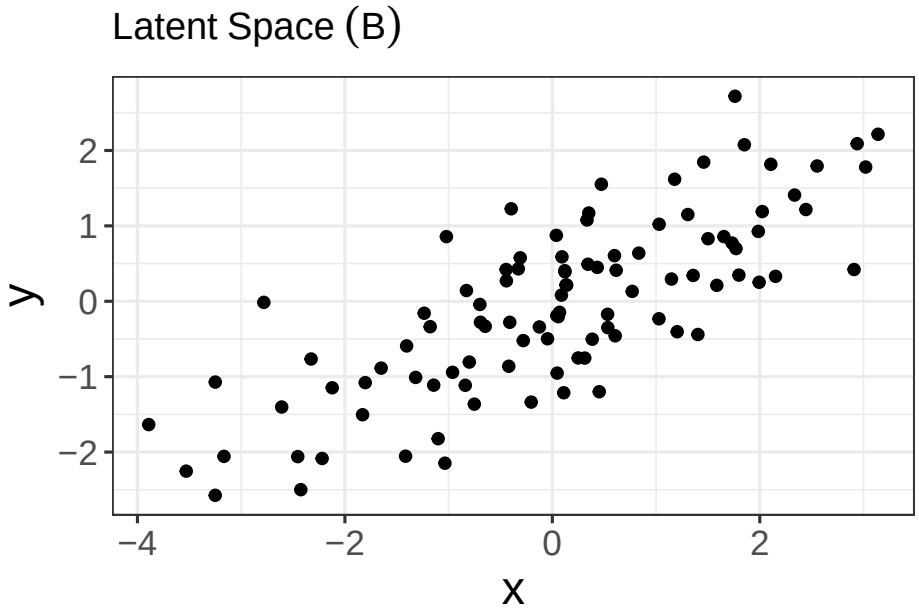
$$\begin{aligned} \mathbb{E}(n_{ij}) &= e^{\mu_j + \Sigma_{jj}/2} =: \lambda_j \\ \mathbb{V}(n_{ij}) &= \lambda_j + \lambda_j^2 \left(e^{\Sigma_{jj}} - 1\right) && \text{(surdispersion)} \\ \mathbb{C}\text{ov}(n_{ij}, n_{ik}) &= \lambda_j \lambda_k \left(e^{\Sigma_{jk}} - 1\right) && \text{(signe arbitraire)} \end{aligned}$$

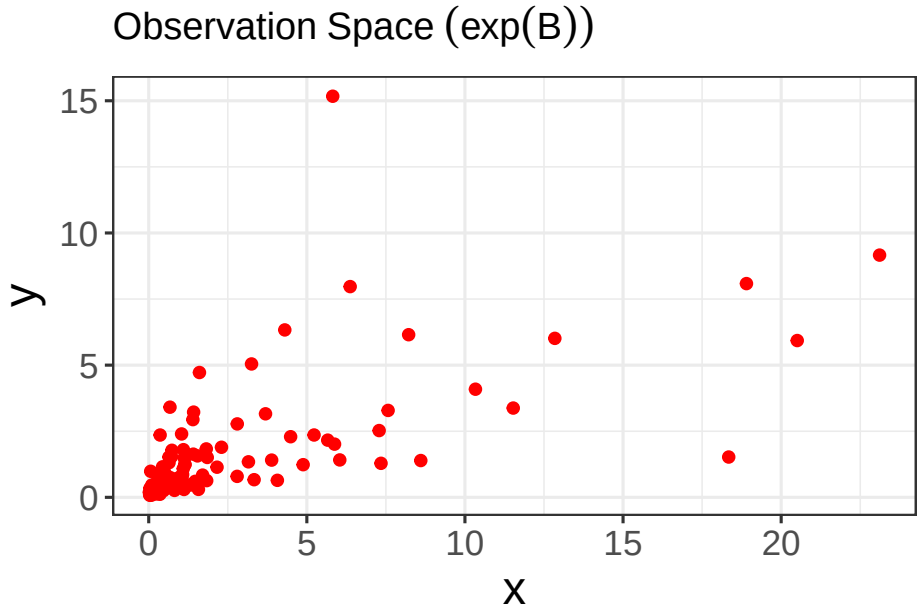
Covariables et Offset

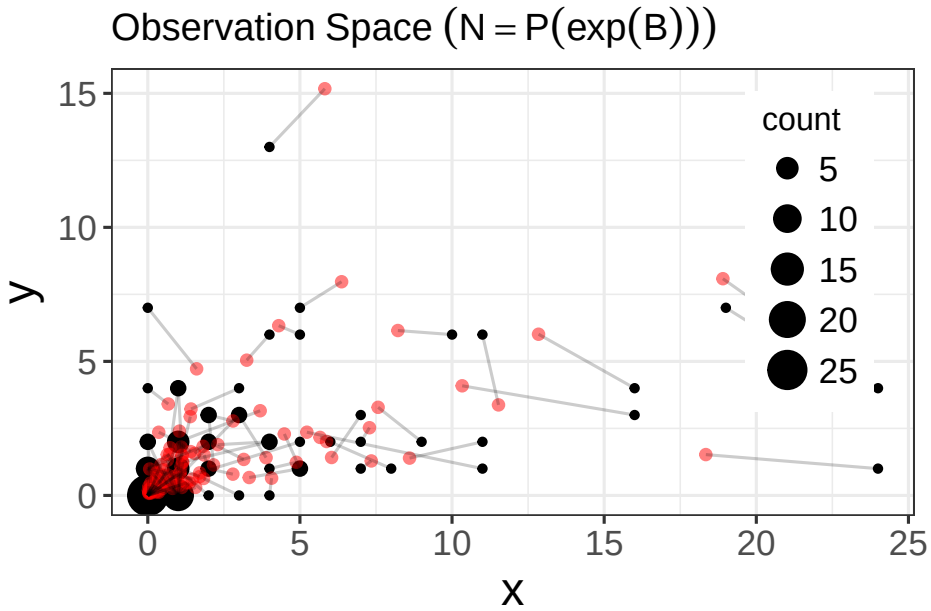
En présence de covariables $\boldsymbol{X}_i$ et d'offset $\boldsymbol{O}_i$, on peut modéliser $\boldsymbol{\mu}_i$ par

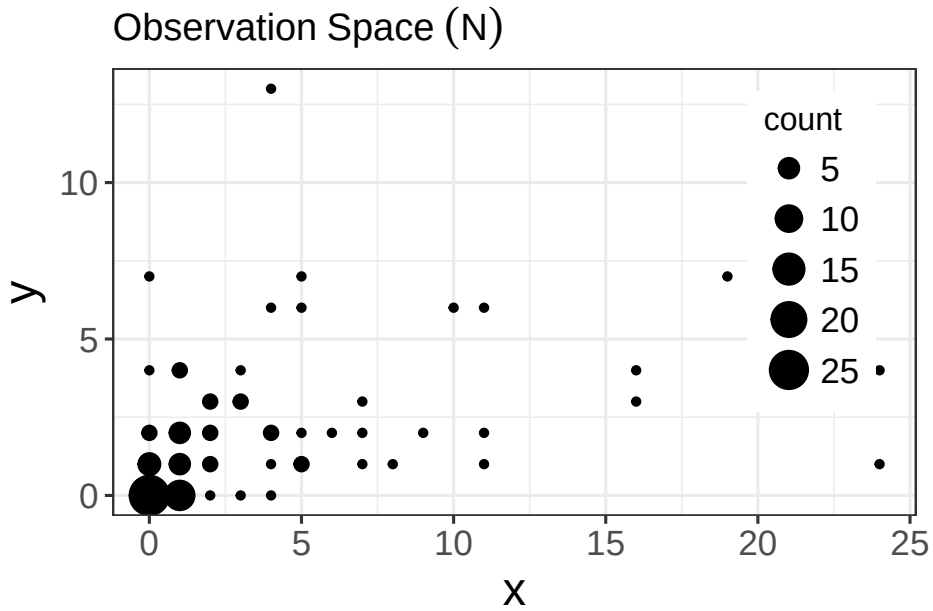
$$\boldsymbol{\mu}_i = \boldsymbol{O}_i + \boldsymbol{X}_i \Theta$$

Vue géométrique









PLN: Vraisemblance

$$
\begin{aligned}
\log p(\boldsymbol{N}, \boldsymbol{B}; \boldsymbol{\Theta}, \boldsymbol{\Sigma}) &= \sum_{i=1}^{n} \log p(\boldsymbol{n}_i | \boldsymbol{b}_i; \boldsymbol{\Theta}, \boldsymbol{\Sigma}) + \log p(\boldsymbol{b}_i) \\
&= \mathbf{1}_n^\intercal \left[\boldsymbol{N} \odot (\boldsymbol{X}\boldsymbol{\Theta} + \boldsymbol{B}) - \exp(\boldsymbol{X}\boldsymbol{\Theta} + \boldsymbol{B}) \right] \mathbf{1}_p \\
&\quad - \frac{\|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{B}\|_F^2}{2} - \frac{nq}{2} \log(2\pi) - K(\boldsymbol{N})
\end{aligned}
$$

où $\exp$ est appliquée par composante et $\odot$ le produit de Hadamard.

PLN: Vraisemblance

$$\begin{aligned}\log p(\boldsymbol{N}, \boldsymbol{B}; \boldsymbol{\Theta}, \boldsymbol{\Sigma}) &= \sum_{i=1}^{n} \log p(\boldsymbol{n}_i | \boldsymbol{b}_i; \boldsymbol{\Theta}, \boldsymbol{\Sigma}) + \log p(\boldsymbol{b}_i) \\ &= \mathbf{1}_n^\intercal \left[\boldsymbol{N} \odot (\boldsymbol{X}\boldsymbol{\Theta} + \boldsymbol{B}) - \exp(\boldsymbol{X}\boldsymbol{\Theta} + \boldsymbol{B})\right] \mathbf{1}_p \\ &\quad - \frac{\|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{B}\|_F^2}{2} - \frac{nq}{2}\log(2\pi) - K(\boldsymbol{N})\end{aligned}$$

où $\exp$ est appliquée par composante et $\odot$ le produit de Hadamard.

- Pas de forme simple pour $\log p(\boldsymbol{N}; \boldsymbol{\Theta}, \boldsymbol{\Sigma}) \Rightarrow$ Pas de solution exacte pour $\boldsymbol{\Theta}$ and $\boldsymbol{\Sigma}$
- Pas de forme simple pour $p(\boldsymbol{n}_i | \boldsymbol{b}_i; \boldsymbol{\Theta}, \boldsymbol{\Sigma}) \Rightarrow$ Pas d'algorithme EM standard

Approximation variationnelle 101

On veut calculer la vraisemblance intractable $\log p(\boldsymbol{N}; \boldsymbol{\Theta}, \boldsymbol{\Sigma})$.

Approximation variationnelle 101

On veut calculer la vraisemblance intractable $\log p(\boldsymbol{N}; \boldsymbol{\Theta}, \boldsymbol{\Sigma})$.

Approximation variationnelle 101

On veut calculer la vraisemblance intractable $\log p(\boldsymbol{N}; \boldsymbol{\Theta}, \boldsymbol{\Sigma})$.

On peut le développer en

$$\log p(\boldsymbol{N}; \boldsymbol{\Theta}, \boldsymbol{\Sigma}) = \mathbb{E}_{\boldsymbol{B}|\boldsymbol{N}}[\log p(\boldsymbol{N}, \boldsymbol{N}; \boldsymbol{\Theta}, \boldsymbol{\Sigma}) - \log p(\boldsymbol{B}|\boldsymbol{N})]$$

Approximation variationnelle 101

On veut calculer la vraisemblance intractable $\log p(\boldsymbol{N}; \boldsymbol{\Theta}, \boldsymbol{\Sigma})$.

On peut le développer en

$$\log p(\boldsymbol{N}; \boldsymbol{\Theta}, \boldsymbol{\Sigma}) = \mathbb{E}_{\boldsymbol{B}|\boldsymbol{N}}[\log p(\boldsymbol{N}, \boldsymbol{N}; \boldsymbol{\Theta}, \boldsymbol{\Sigma}) - \log p(\boldsymbol{B}|\boldsymbol{N})]$$

- $\log p(\boldsymbol{N}, \boldsymbol{B}; \boldsymbol{\Theta}$ est simple ☺;
- $\log p(\boldsymbol{B}|\boldsymbol{N})$ est complexe ☹.

Approximation variationnelle 101

On veut calculer la vraisemblance intractable $\log p(\boldsymbol{N}; \boldsymbol{\Theta}, \boldsymbol{\Sigma})$.

On peut le développer en

$$\log p(\boldsymbol{N}; \boldsymbol{\Theta}, \boldsymbol{\Sigma}) = \mathbb{E}_{\boldsymbol{B}|\boldsymbol{N}}[\log p(\boldsymbol{N}, \boldsymbol{N}; \boldsymbol{\Theta}, \boldsymbol{\Sigma}) - \log p(\boldsymbol{B}|\boldsymbol{N})]$$

- $\log p(\boldsymbol{N}, \boldsymbol{B}; \boldsymbol{\Theta}$ est simple ☺;
- $\log p(\boldsymbol{B}|\boldsymbol{N})$ est complexe ☹.

Idée: Remplacer $\log p(\boldsymbol{B}|\boldsymbol{N})$ par une approximation agréable $\tilde{p}(\boldsymbol{B})$.

Approximation variationnelle

Soit $\tilde{p} = \tilde{p}_{\boldsymbol{M},\boldsymbol{S}} = \otimes_i \mathcal{N}_p(\boldsymbol{m}_i, Diag(\boldsymbol{s_i})^2)$ une approximation de $p(\boldsymbol{B}|\boldsymbol{N})$.

On a la borne inférieure variationnelle de la vraisemblance

$$\log p(\boldsymbol{N}; \boldsymbol{\Theta}, \boldsymbol{\Sigma}) \geq J(\boldsymbol{M}, \boldsymbol{S}, \boldsymbol{\Theta}, \boldsymbol{\Sigma}) := J(\tilde{p}, \boldsymbol{\Theta}, \boldsymbol{\Sigma})$$

avec

$$J(\tilde{p}, \boldsymbol{\Theta}, \boldsymbol{\Sigma}) := \log p(\boldsymbol{N}; \boldsymbol{\Theta}, \boldsymbol{\Sigma}) - KL(\tilde{p}(\boldsymbol{B})||p(\boldsymbol{B}|\boldsymbol{N}; \boldsymbol{\Theta}, \boldsymbol{\Sigma}))$$

Approximation variationnelle

Soit $\tilde{p} = \tilde{p}_{\boldsymbol{M},\boldsymbol{S}} = \otimes_i \mathcal{N}_p(\boldsymbol{m}_i, Diag(\boldsymbol{s_i})^2)$ une approximation de $p(\boldsymbol{B}|\boldsymbol{N})$.

On a la borne inférieure variationnelle de la vraisemblance

$$\log p(\boldsymbol{N}; \boldsymbol{\Theta}, \boldsymbol{\Sigma}) \geq J(\boldsymbol{M}, \boldsymbol{S}, \boldsymbol{\Theta}, \boldsymbol{\Sigma}) := J(\tilde{p}, \boldsymbol{\Theta}, \boldsymbol{\Sigma})$$

avec

$$\begin{aligned} J(\tilde{p}, \boldsymbol{\Theta}, \boldsymbol{\Sigma}) &:= \log p(\boldsymbol{N}; \boldsymbol{\Theta}, \boldsymbol{\Sigma}) - KL(\tilde{p}(\boldsymbol{B})||p(\boldsymbol{B}|\boldsymbol{N}; \boldsymbol{\Theta}, \boldsymbol{\Sigma})) \\ &= \mathbb{E}_{\tilde{p}}[\log p(\boldsymbol{N}, \boldsymbol{B}; \boldsymbol{\Theta}, \boldsymbol{\Sigma}) - \log \tilde{p}(\boldsymbol{B})] \end{aligned}$$

Approximation variationnelle

Soit $\tilde{p} = \tilde{p}_{\boldsymbol{M},\boldsymbol{S}} = \otimes_i \mathcal{N}_p(\boldsymbol{m}_i, Diag(\boldsymbol{s_i})^2)$ une approximation de $p(\boldsymbol{B}|\boldsymbol{N})$.

On a la borne inférieure variationnelle de la vraisemblance

$$\log p(\boldsymbol{N}; \boldsymbol{\Theta}, \boldsymbol{\Sigma}) \geq J(\boldsymbol{M}, \boldsymbol{S}, \boldsymbol{\Theta}, \boldsymbol{\Sigma}) := J(\tilde{p}, \boldsymbol{\Theta}, \boldsymbol{\Sigma})$$

avec

$$\begin{aligned} J(\tilde{p}, \boldsymbol{\Theta}, \boldsymbol{\Sigma}) &:= \log p(\boldsymbol{N}; \boldsymbol{\Theta}, \boldsymbol{\Sigma}) - KL(\tilde{p}(\boldsymbol{B})||p(\boldsymbol{B}|\boldsymbol{N}; \boldsymbol{\Theta}, \boldsymbol{\Sigma})) \\ &= \mathbb{E}_{\tilde{p}}[\log p(\boldsymbol{N}, \boldsymbol{B}; \boldsymbol{\Theta}, \boldsymbol{\Sigma}) - \log \tilde{p}(\boldsymbol{B})] \\ &= \mathbf{1}_n^\intercal \left[\boldsymbol{N} \odot (\boldsymbol{X}\boldsymbol{\Theta} + \boldsymbol{M}) - \boldsymbol{A}\right] \mathbf{1}_p \\ &\quad - \frac{1}{2} \mathbf{1}_n^\intercal \left[\boldsymbol{M} \odot \boldsymbol{M} + \boldsymbol{S} \odot \boldsymbol{S} - 2\log(\boldsymbol{S}) - \mathbf{1}_{n,p}\right] \mathbf{1}_p - K(\boldsymbol{N}). \end{aligned}$$

où

$$\boldsymbol{A} = \exp\left(\boldsymbol{X}\boldsymbol{\Theta} + \boldsymbol{M} + \frac{1}{2}(\boldsymbol{S} \odot \boldsymbol{S})\right)$$

Inférence variationnelle

Pour reconstuire le réseau, on ajoute de la structure sur $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$ via une contrainte de sparsité $\|\boldsymbol{\Omega}\|_1$

Variational EM On maximise la borne J avec la structure sur $\|\boldsymbol{\Omega}\|_1$

$$\underset{\boldsymbol{M},\boldsymbol{S},\boldsymbol{\Theta},\boldsymbol{\Omega}}{\arg\max}\ J(\boldsymbol{M},\boldsymbol{S},\boldsymbol{\Theta},\boldsymbol{\Omega}) - \lambda\|\boldsymbol{\Omega}\|_1$$

Inférence variationnelle

Pour reconstuire le réseau, on ajoute de la structure sur $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$ via une contrainte de sparsité $\|\boldsymbol{\Omega}\|_1$

Variational EM On maximise la borne J avec la structure sur $\|\boldsymbol{\Omega}\|_1$

$$\underset{\boldsymbol{M},\boldsymbol{S},\boldsymbol{\Theta},\boldsymbol{\Omega}}{\arg\max}\, J(\boldsymbol{M},\boldsymbol{S},\boldsymbol{\Theta},\boldsymbol{\Omega}) - \lambda\|\boldsymbol{\Omega}\|_1$$

On alterne entre les étapes suivantes

- Mise à jour de $(\boldsymbol{M},\boldsymbol{S},\boldsymbol{\Theta}) = \arg\max_{\boldsymbol{M},\boldsymbol{S},\boldsymbol{\Theta}} J(\boldsymbol{M},\boldsymbol{S},\boldsymbol{\Theta},\boldsymbol{\Omega})$
- Mise à jour de $\boldsymbol{\Omega} = \arg\max_{\boldsymbol{\Omega}} J(\boldsymbol{M},\boldsymbol{S},\boldsymbol{\Theta},\boldsymbol{\Omega}) - \lambda\|\boldsymbol{\Omega}\|_1$

Algorithme

Propriétés de J_q

- J n'est pas globalement concave;
- $J(.,.,.,\mathbf{\Omega}) - \lambda\|\mathbf{\Omega}\|_1$ concave en $\mathbf{\Omega}$: résolution via un glasso
- J concave in $(\boldsymbol{M}, \boldsymbol{S}, \mathbf{\Theta})$: résolution via une desccente de gradient

Algorithme

Propriétés de J_q

- J n'est pas globalement concave;
- $J(.,.,.,\mathbf{\Omega}) - \lambda\|\mathbf{\Omega}\|_1$ concave en $\mathbf{\Omega}$: résolution via un glasso
- J concave in $(\boldsymbol{M}, \boldsymbol{S}, \mathbf{\Theta})$: résolution via une desccente de gradient

En pratique

Tests de plusieurs procédures:

- Procédure itérative / Optimisation de J
- Remplacer glasso par Meinshausen et Bülhman

Avantages, inconvénients

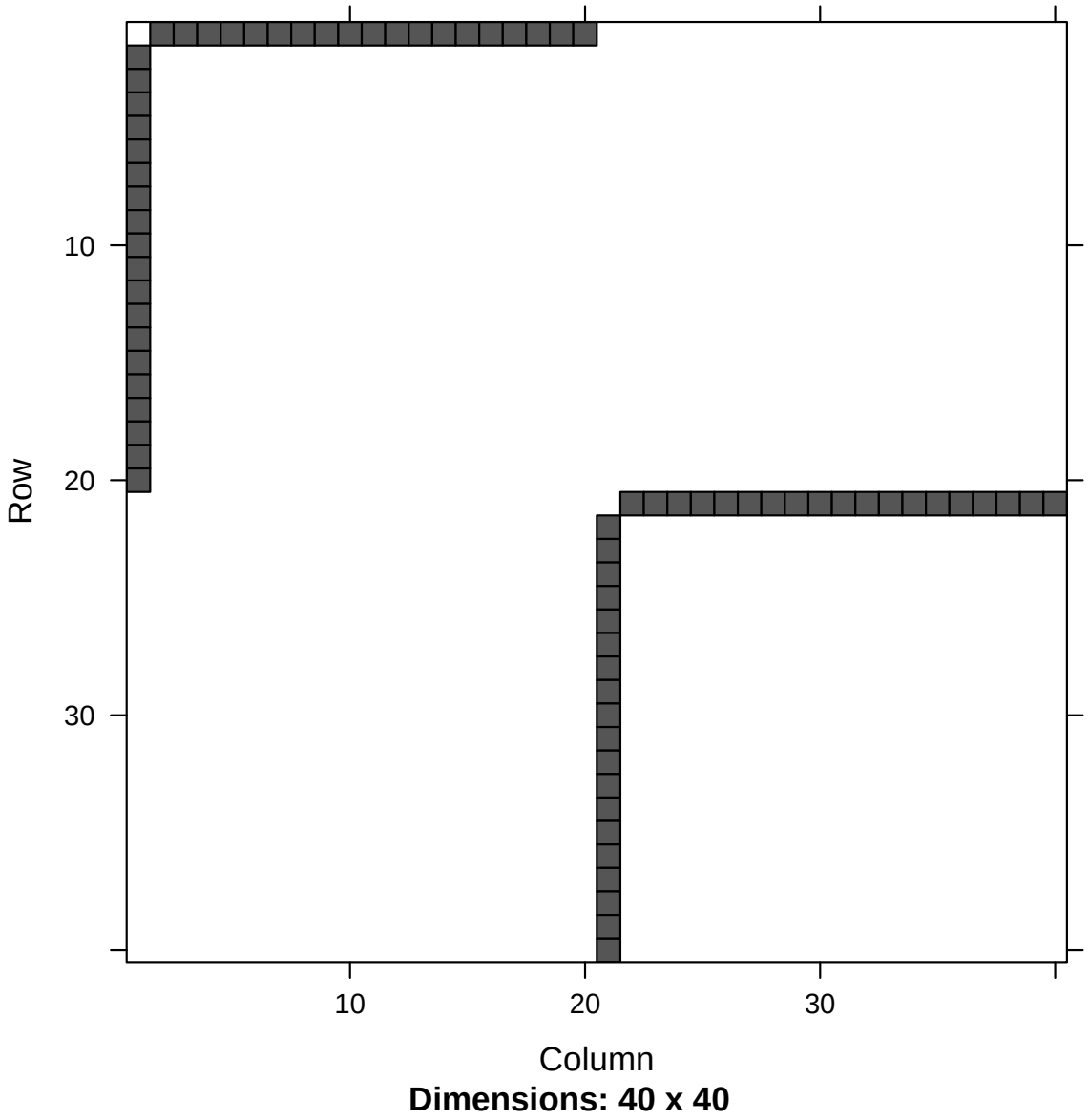
Avantages

+ Focus sur $\mathbf{\Omega} = \mathbf{\Sigma}^{-1}$
+ Basée sur une méthode de vraisemblance pénalisée
+ Modélisation des comptages et pas des fréquences

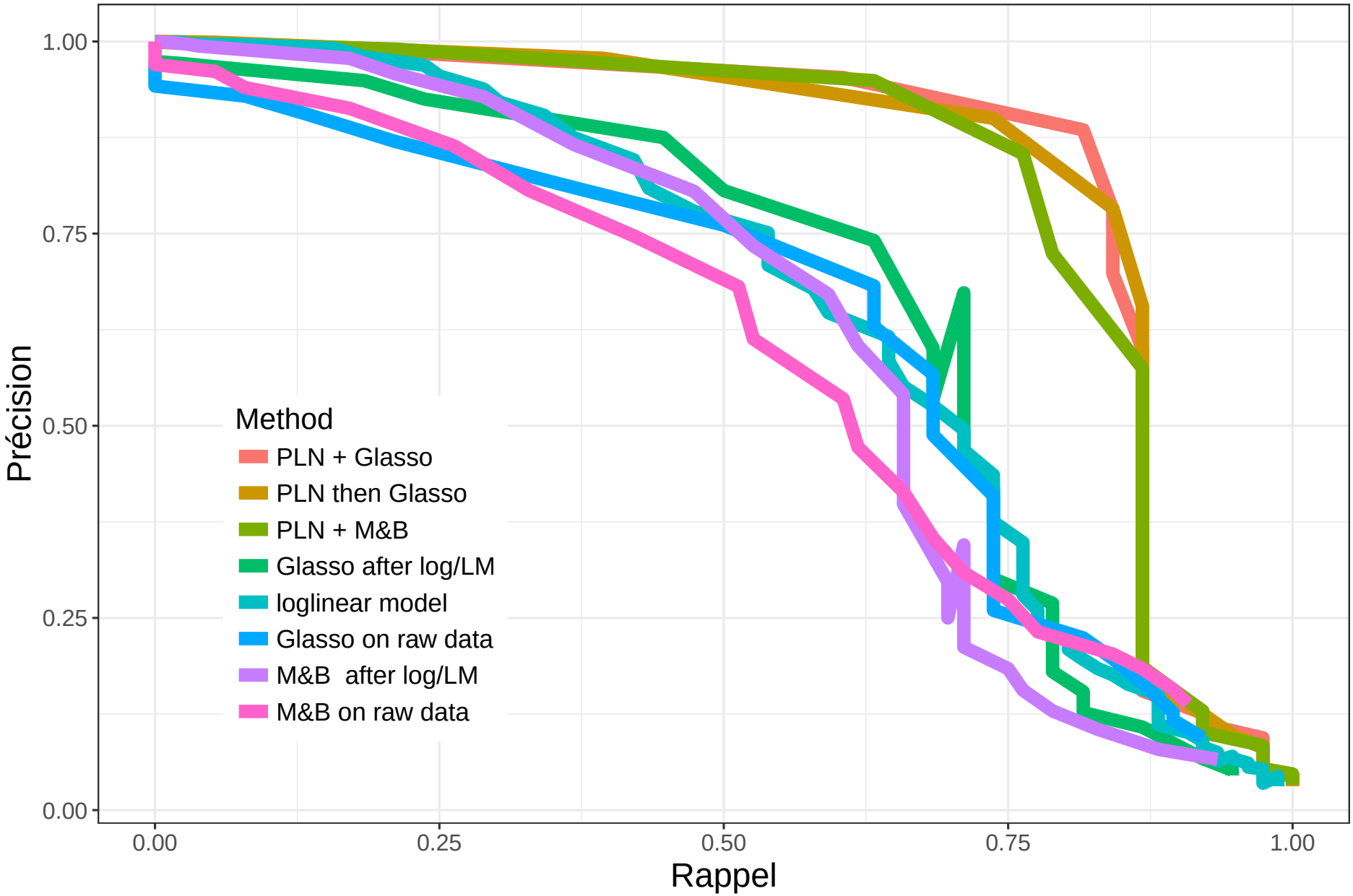
Inconvénients

- Inférence variationnelle (approchée)
- Choix de λ

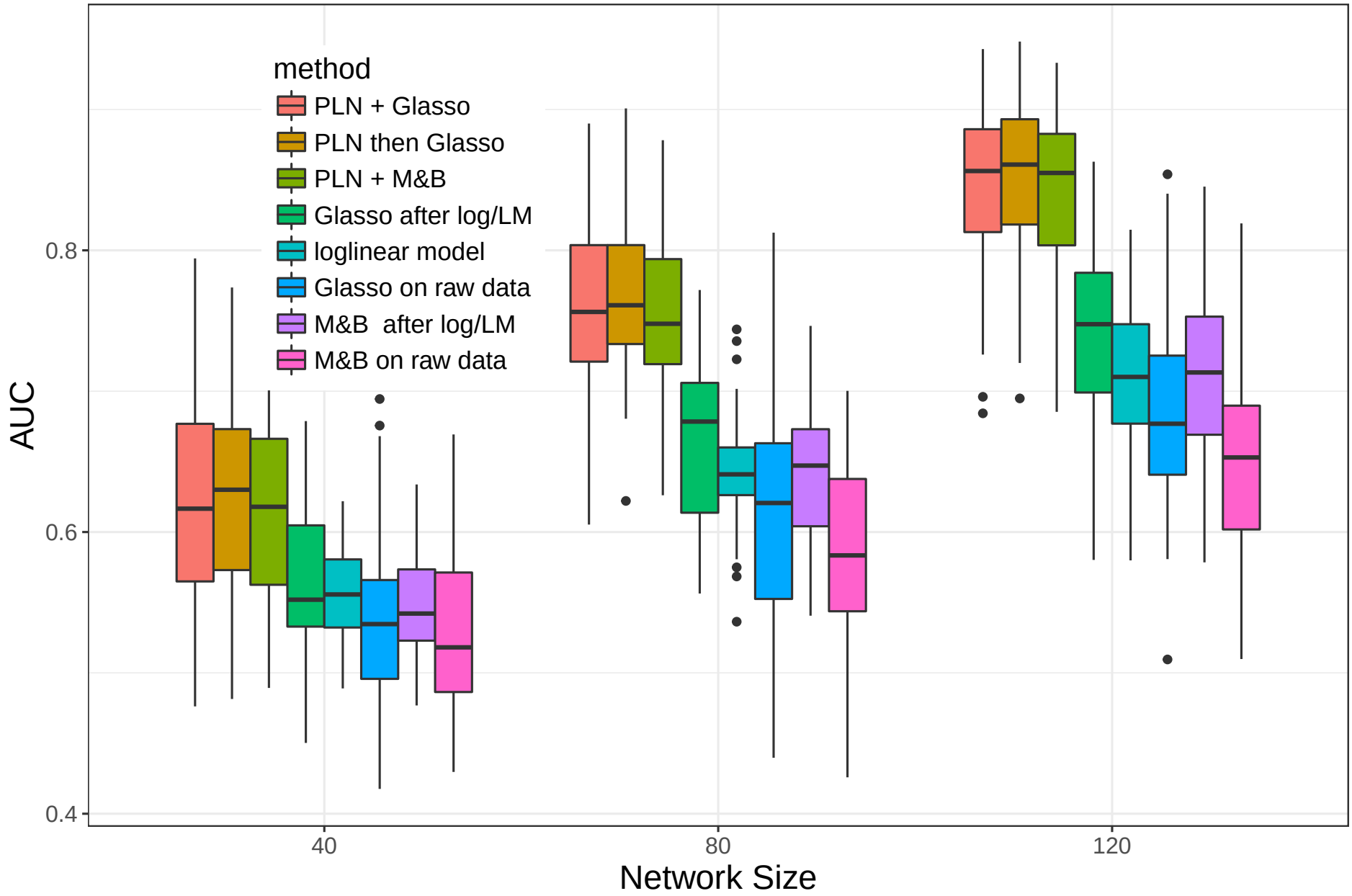
Un graphe simple (Hub)



Comparaison de différentes méthodes



(Petite) Étude de simulations



Plan

En résumé

Interaction en écologie

- Définition assez floue
- Beaucoup de méthodes (SparCC, REBACCA, SPIEC-EASI, etc)
- Mais chacune avec ses limitations

En résumé

Interaction en écologie

- Définition assez floue
- Beaucoup de méthodes (SparCC, REBACCA, SPIEC-EASI, etc)
- Mais chacune avec ses limitations

Modèle PLN

- Modèle générique pour données de comptages
- Adaptation *facile* aux familles exponentielles
- Algorithme variationnel pour l'inférence

En résumé

Interaction en écologie

- Définition assez floue
- Beaucoup de méthodes (SparCC, REBACCA, SPIEC-EASI, etc)
- Mais chacune avec ses limitations

Modèle PLN

- Modèle générique pour données de comptages
- Adaptation *facile* aux familles exponentielles
- Algorithme variationnel pour l'inférence

Extensions

- Ajouter du *Zero Inflated* pour distinguer zéros structuraux / de sous-échantillonnage
- Rajouter une structuration spatiale / temporelle dans l'espace latent.

Code

- Code **R/C++** disponible sur https://github.com/jchiquet/PLNmodels;
- Programmation OO avec des classes R6;

```
Y <- read.csv("myCounts.csv")
meta <- read.csv("myCovariates.csv")
## offset
O <- log(matrix(rep(rowSums(Y), ncol(Y)), ncol = ncol(Y)))
## Fit many models
PLNnets.models <- PLNnetwork(dat$Y)
## Choose best model
best.model <- PLNnets.models$getBestModel()
## Reconstruct and plot network
best.model$latentNetwork()
best.model$plot_network()
```

Références

1. J. T. Staley and a. A. Konopka, “Measurement of in Situ Activities of Nonphotosynthetic Microorganisms in Aquatic and Terrestrial Habitats,” Annual Review of Microbiology, vol. 39, no. 1, pp. 321–346, 1985.
2. W. Wade, “Unculturable bacteria—the uncharacterized organisms that cause oral infections,” Journal of the Royal Society of Medicine, vol. 95, pp. 81–83, Feb. 2002.
3. G. Lima-Mendez, K. Faust, et al., “Determinants of community structure in the global plankton interactome,” Science, vol. 348, p. 1262073, May 2015.
4. K. Faust and J. Raes, “Microbial interactions : from networks to models,” Nature Reviews Microbiology, vol. 10, pp. 538–550, Aug. 2012.
5. K. Faust, J. F. Sathirapongsasuti, J. Izard, N. Segata, D. Gevers, J. Raes, and C. Huttenhower, “Mi- crobial Co-occurrence Relationships in the Human Microbiome,” PLOS Computational Biology, vol. 8, p. e1002606, July 2012.
6. J. Friedman and E. J. Alm, “Inferring Correlation Networks from Genomic Survey Data,” PLOS Computational Biology, vol. 8, p. e1002687, Sept. 2012.
7. Y. Ban, L. An, and H. Jiang, “Investigating microbial co-occurrence patterns based on metagenomic compositional data,” Bioinformatics, vol. 31, pp. 3322–3329, Oct. 2015.
8. J. Aitchison, “A new approach to null correlations of proportions,” Journal of the International Association for Mathematical Geology, vol. 13, pp. 175–189, Apr. 1981.
9. L. C. Xia, D. Ai, J. Cram, J. A. Fuhrman, and F. Sun, “Efficient statistical significance approximation for local similarity analysis of high-throughput time series data,” Bioinformatics, vol. 29, pp. 230–237, Jan. 2013.
10. Y. Deng, Y.-H. Jiang, Y. Yang, Z. He, F. Luo, and J. Zhou, “Molecular ecological network analyses,” BMC Bioinformatics, vol. 13, p. 113, 2012.
11. 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, vol. 11, p. e1004226, May 2015.
12. S. Weiss, W. Van Treuren, et al., “Correlation detection strategies in microbial data sets vary widely in sensitivity and precision,” The ISME Journal, vol. 10, no. 7, pp. 1669–1681, 2016.
13. M. Layeghifard, D. M. Hwang, and D. S. Guttman, “Disentangling Interactions in the Microbiome: A Network Perspective” Trends in Microbiology, 2017, 25, 217-228