Une promenade parmi les modèles pour comptages multivariées

M. Mariadassou, INRAE-MaIAGE

Travail avec Julien Chiquet and Stéphane Robin

CNRS Villa Clythia, Fréjus, 6 octobre 2022

Julien Chiquet, M.M., Stéphane Robin,
Variational inference for probabilistic Poisson PCA
http://doi.org/10.1214/18-AOAS1177 (*Annals of Applied Statistics*, 2019)

Julien Chiquet, M.M., Stéphane Robin,
Variational inference for network inference with count data
http://proceedings.mlr.press/v97/chiquet19a/chiquet19a.pdf (*ICML19*)

Julien Chiquet, M.M., Stéphane Robin,
The Poisson-lognormal model as a versatile framework for the joint analysis of species abundances
http://doi.org/10.3389/fevo.2021.588292 (*Frontiers in Ecol. and Evol.*, 2021)

PLNmodels package, development version on github
`devtools::install_github("pln-team/PLNmodels", build_vignettes=TRUE)`
https://pln-team.github.io/PLNmodels/

MaIAGE

INRAE

Outline

Microbiome: Impact du sevrage sur le microbiote des porcelets

Données issues de [MBE+15].

- $n = 155$ échantillons ($= 31$ porcelets à 5 points de temps)
- $p = 1038$ espèces bactériennes (OTUs) avec prévalence ≥ 0.05
- Certaines covariables (sexe, portée, etc)
- Offsets: $o_i =$ offset de l'échantillon i (profondeur de séquençage)

Microbiome: Impact du sevrage sur le microbiote des porcelets

Données issues de [MBE+15].

- $n = 155$ échantillons ($= 31$ porcelets à 5 points de temps)
- $p = 1038$ espèces bactériennes (OTUs) avec prévalence ≥ 0.05
- Certaines covariables (sexe, portée, etc)
- Offsets: $o_i =$ offset de l'échantillon i (profondeur de séquençage)

But: Étudier l'impact du sevrage sur le microbiote intestinal

Un regard sur les données

Données de métagénétique de [MBE+15]

- count matrice de taille $n = 155$ échantillons, $p = 1038$ espèces

```
mach_counts[1:2, c(3, 9, 12, 15)]
```

```
##        5982 347 349 5854
## SF0901    0  23   3    0
## SF0902    8   0   4    0
```

- $d = 8$ covariables (sexe, portée, statut de sevrage, . . .)

```
mach_covariates[1:2, ]
```

```
##        Run Project Time Bande sex        mere Weaned
## SF0901   3 Kinetic  D14  1105   1 17MAG101814   TRUE
## SF0902   3 Kinetic  D36  1105   1 17MAG101814  FALSE
```

- Effort d'observation pour chaque échantillon

```
mach_offsets[1:2, c(1:4, 48:51)]
```

```
##        16342  164 5982 5980 10413 6307 8949  346
## SF0901  3084 3084 3084 3084  3084 3084 3084 3084
## SF0902  2182 2182 2182 2182  2182 2182 2182 2182
```

Pathobiome: Mildiou du bouleau

Données issues de [JFS+16].

- $n = 116$ feuilles de bouleau = échantillons
- $p = 114$ espèces microbiennes
  - $p_1 = 66$ espèces microbiennes (OTUs, issues du marqueur 16S)
  - $p_2 = 48$ espèces fongiques (OTUs, issues du marqueur ITS)
- covariables: arbre (résistant, intermédiaire, susceptible), hauteur, distance au tronc, ...
- offsets: $o_{i1} \neq o_{i2} =$ offset pour les bactéries, espèces fongiques

Pathobiome: Mildiou du bouleau

Données issues de [JFS+16].

- $n = 116$ feuilles de bouleau = échantillons
- $p = 114$ espèces microbiennes
  - $p_1 = 66$ espèces microbiennes (OTUs, issues du marqueur 16S)
  - $p_2 = 48$ espèces fongiques (OTUs, issues du marqueur ITS)
- covariables: arbre (résistant, intermédiaire, susceptible), hauteur, distance au tronc, ...
- offsets: $o_{i1} \neq o_{i2} =$ offset pour les bactéries, espèces fongiques

```
offsets[1:2, c(1:4, 48:51)]
```

```
##       f_1  f_2  f_3  f_4 E_alphitoides b_1045 b_109 b_1093
## [1,] 2488 2488 2488 2488          2488   8315  8315   8315
## [2,] 2054 2054 2054 2054          2054    662   662    662
```

Pathobiome: Mildiou du bouleau

Données issues de [JFS+16].

- $n = 116$ feuilles de bouleau = échantillons
- $p = 114$ espèces microbiennes
  - $p_1 = 66$ espèces microbiennes (OTUs, issues du marqueur 16S)
  - $p_2 = 48$ espèces fongiques (OTUs, issues du marqueur ITS)
- covariables: arbre (résistant, intermédiaire, susceptible), hauteur, distance au tronc, ...
- offsets: $o_{i1} \neq o_{i2}$ = offset pour les bactéries, espèces fongiques

```
offsets[1:2, c(1:4, 48:51)]
```

```
##       f_1  f_2  f_3  f_4 E_alphitoides b_1045 b_109 b_1093
## [1,] 2488 2488 2488 2488          2488   8315  8315   8315
## [2,] 2054 2054 2054 2054          2054    662   662    662
```

But. Comprendre les interactions entre les espèces, y compris le pathogène *E. alphitoides*.

Problématique & Formalism

Tableau de données: $\mathbf{Y} = (Y_{ij}), n \times p$; $\mathbf{X} = (X_{ik}), n \times d$; $\mathbf{O} = (O_{ij}), n \times p$

- $Y_{ij} =$ abondance (nombre de lectures) de l'espèce j dans l'échantillon i
- $X_{ik} =$ valeur de la covariable k dans l'échantillon i
- $O_{ij} =$ offset (effort d'observation) pour l'espèce j dans l'échantillon i

Problématique & Formalism

Tableau de données: $\mathbf{Y} = (Y_{ij}), n \times p$; $\mathbf{X} = (X_{ik}), n \times d$; $\mathbf{O} = (O_{ij}), n \times p$

- Y_{ij} = abondance (nombre de lectures) de l'espèce j dans l'échantillon i
- X_{ik} = valeur de la covariable k dans l'échantillon i
- O_{ij} = offset (effort d'observation) pour l'espèce j dans l'échantillon i

Besoin de méthodes d'analyses multivariées pour comprendre les écosytèmes microbiens

- Mettre en évidence des patrons de diversité
  $\rightsquigarrow$ Résumer l'information de $\mathbf{Y}$ (ACP, classification, ...)
- Comprendre les interactions entre espèces
  $\rightsquigarrow$ Inférence de réseau (sélection de covariance)
- Corriger les effets techniques et confondants
  $\rightsquigarrow$ Prise en compte de covariables et d'offsets.

Problématique & Formalism

Tableau de données: $\mathbf{Y} = (Y_{ij}), n \times p$; $\mathbf{X} = (X_{ik}), n \times d$; $\mathbf{O} = (O_{ij}), n \times p$

- Y_{ij} = abondance (nombre de lectures) de l'espèce j dans l'échantillon i
- X_{ik} = valeur de la covariable k dans l'échantillon i
- O_{ij} = offset (effort d'observation) pour l'espèce j dans l'échantillon i

Besoin de méthodes d'analyses multivariées pour comprendre les écosytèmes microbiens

- Mettre en évidence des patrons de diversité
  $\rightsquigarrow$ Résumer l'information de $\mathbf{Y}$ (ACP, classification, ...)
- Comprendre les interactions entre espèces
  $\rightsquigarrow$ Inférence de réseau (sélection de covariance)
- Corriger les effets techniques et confondants
  $\rightsquigarrow$ Prise en compte de covariables et d'offsets.

$\rightsquigarrow$ Besoin d'un cadre générique pour modéliser les données de comptages multivariées

Microbial Ecology 101

1. Appliquer votre distance préférée (Jaccard, Bray-Curtis, UniFrac, weighted UniFrac, Aitchison)

Microbial Ecology 101

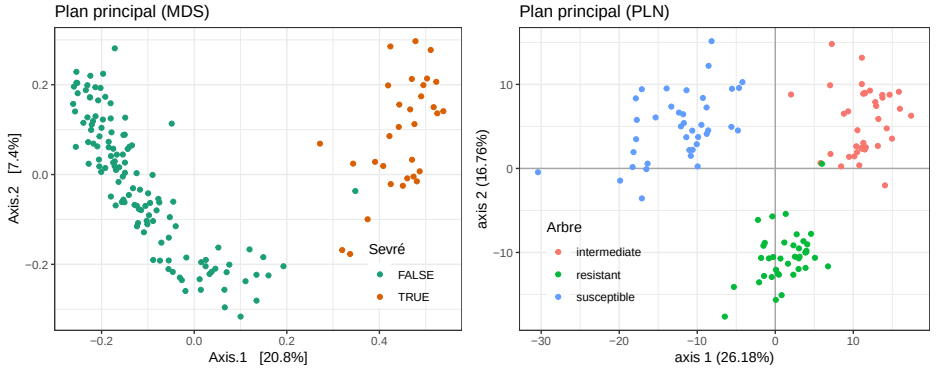
1. Appliquer votre distance préférée (Jaccard, Bray-Curtis, UniFrac, weighted UniFrac, Aitchison)
2. Appliquer votre méthode de réduction de dimension préférée (ACP, MDS/PCoA, NMDS, RDA, PLN, etc)

Microbial Ecology 101

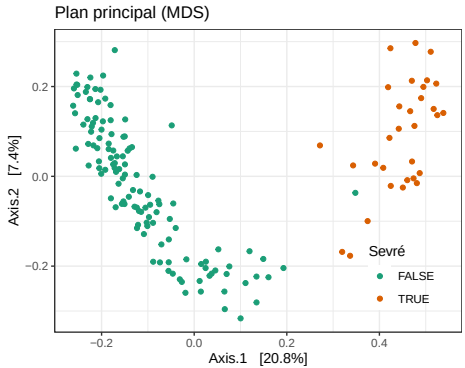
1. Appliquer votre distance préférée (Jaccard, Bray-Curtis, UniFrac, weighted UniFrac, Aitchison)
2. Appliquer votre méthode de réduction de dimension préférée (ACP, MDS/PCoA, NMDS, RDA, PLN, etc)
3. Tracer le résultat

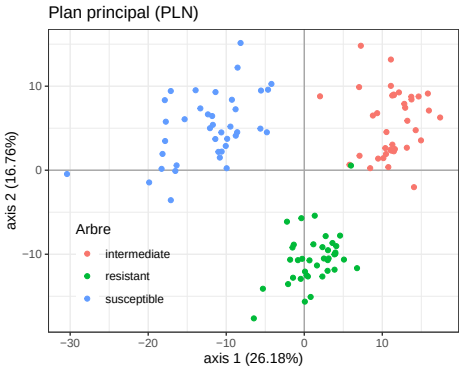
Microbial Ecology 101

1. Appliquer votre distance préférée (Jaccard, Bray-Curtis, UniFrac, weighted UniFrac, Aitchison)
2. Appliquer votre méthode de réduction de dimension préférée (ACP, MDS/PCoA, NMDS, RDA, PLN, etc)
3. Tracer le résultat
4. *Et voilà!*



Microbial Ecology 101



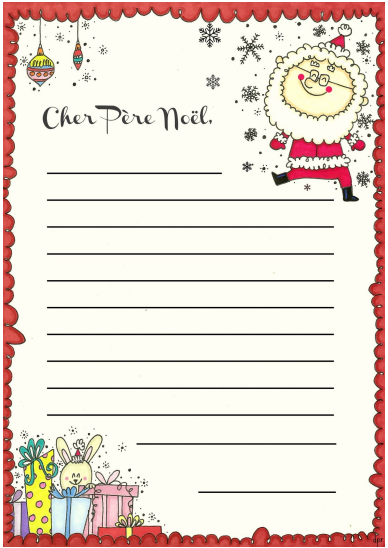


1. Parfait pour trouver des structures...
2. Mais peu idéal pour la modéliser

Quel type de modèle générique?

Quel cadre générique pour les données de comptages multivariées?

Quel type de modèle générique?



Cher Papa Noël

Pour Noël, je voudrais une famille de modèles génératifs qui sont:

Cher Papa Noël

Pour Noël, je voudrais une famille de modèles génératifs qui sont:

- Suffisamment flexibles pour:
  - modéliser les compositions moyennes;
  - modéliser la dispersion (variabilité biologique);
  - modéliser les interactions entre espèces (réseaux écologiques);
  - prendre en compte l'hétérogénéité des communautés;
  - intégrer les données issues de sources différentes (bactéries et champignons)

Cher Papa Noël

Pour Noël, je voudrais une famille de modèles génératifs qui sont:

- Suffisamment flexibles pour:
  - modéliser les compositions moyennes;
  - modéliser la dispersion (variabilité biologique);
  - modéliser les interactions entre espèces (réseaux écologiques);
  - prendre en compte l'hétérogénéité des communautés;
  - intégrer les données issues de sources différentes (bactéries et champignons)
- aussi parcimonieux que possible;

Cher Papa Noël

Pour Noël, je voudrais une famille de modèles génératifs qui sont:

- Suffisamment flexibles pour:
  - modéliser les compositions moyennes;
  - modéliser la dispersion (variabilité biologique);
  - modéliser les interactions entre espèces (réseaux écologiques);
  - prendre en compte l'hétérogénéité des communautés;
  - intégrer les données issues de sources différentes (bactéries et champignons)
- aussi parcimonieux que possible;
- interprétables;

Cher Papa Noël

Pour Noël, je voudrais une famille de modèles génératifs qui sont:

- Suffisamment flexibles pour:
  - modéliser les compositions moyennes;
  - modéliser la dispersion (variabilité biologique);
  - modéliser les interactions entre espèces (réseaux écologiques);
  - prendre en compte l'hétérogénéité des communautés;
  - intégrer les données issues de sources différentes (bactéries et champignons)
- aussi parcimonieux que possible;
- interprétables;
- rapides et faciles à estimer à partir de données;

Cher Papa Noël

Pour Noël, je voudrais une famille de modèles génératifs qui sont:

- Suffisamment flexibles pour:
  - modéliser les compositions moyennes;
  - modéliser la dispersion (variabilité biologique);
  - modéliser les interactions entre espèces (réseaux écologiques);
  - prendre en compte l'hétérogénéité des communautés;
  - intégrer les données issues de sources différentes (bactéries et champignons)
- aussi parcimonieux que possible;
- interprétables;
- rapides et faciles à estimer à partir de données;
- réalistes (*e.g.* simulent des échantillons réalistes).

Outline

Outline

Modèles Multinomiaux

Intuition

- Il y a p espèces, présentes en proportions $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_p)$ et en quantité infinie dans la population
- On échantillonne N (profondeur de séquençage) bactéries dans la population

Modèles Multinomiaux

Intuition

- Il y a p espèces, présentes en proportions $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_p)$ et en quantité infinie dans la population
- On échantillonne N (profondeur de séquençage) bactéries dans la population

Modèle mathématiques

$$\mathbf{Y} \sim \mathcal{M}(N, \boldsymbol{\pi})$$

Modèles Multinomiaux

Intuition

- Il y a p espèces, présentes en proportions $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_p)$ et en quantité infinie dans la population
- On échantillonne N (profondeur de séquençage) bactéries dans la population

Modèle mathématiques

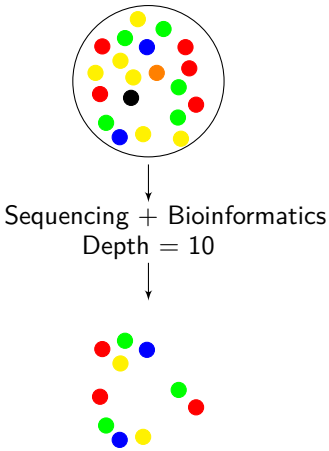
$$\mathbf{Y} \sim \mathcal{M}(N, \boldsymbol{\pi})$$

L'inférence est facile

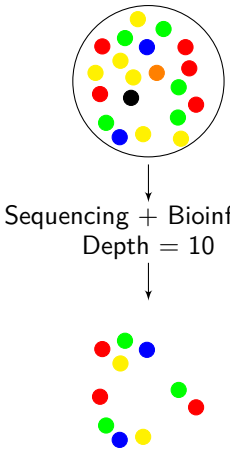
$$\hat{\pi}_j = \frac{\sum_{i=1}^{n} Y_{ij}}{\sum_{i=1}^{n} N_i}$$

avec Y_{ij} l'abundance de l'espèce j dans l'échantillon i et N_i la profondeur de l'échantillon i.

Distribution multinomiale: tirage dans une urne (avec remise)

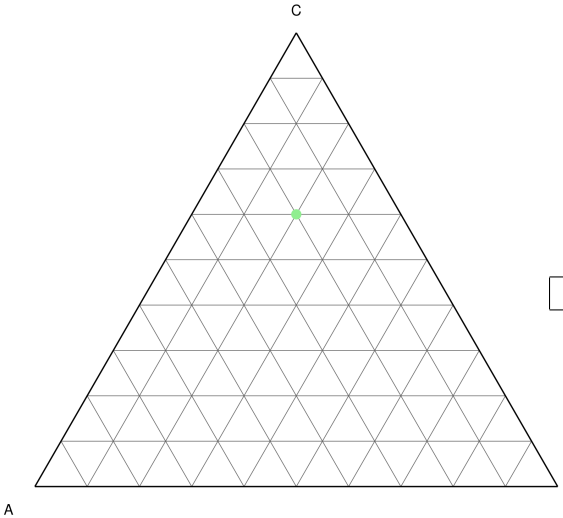


Distribution multinomiale: tirage dans une urne (avec remise)



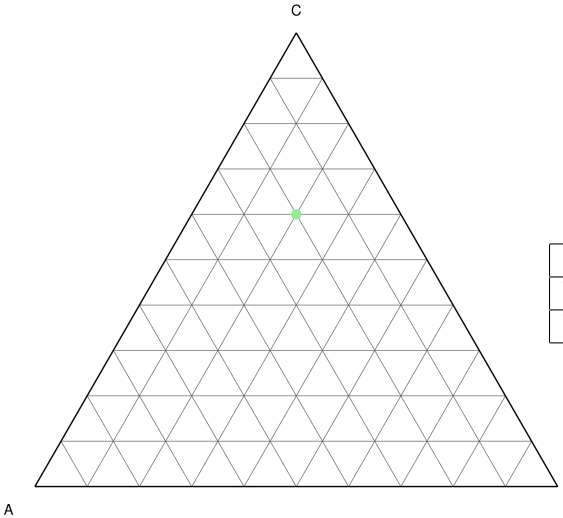
	●	●	●	●	●	●
Prop.	0.25	0.30	0.25	0.05	0.10	0.05
Counts	3	2	3	0	2	0
Obs. Prop.	0.3	0.2	0.3	0	0.2	0

Modèle multinomial



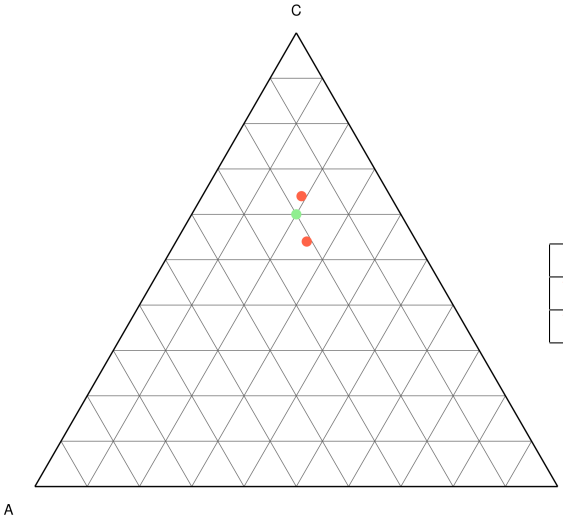
	A	B	C
π	0.2	0.2	0.6

Modèle multinomial



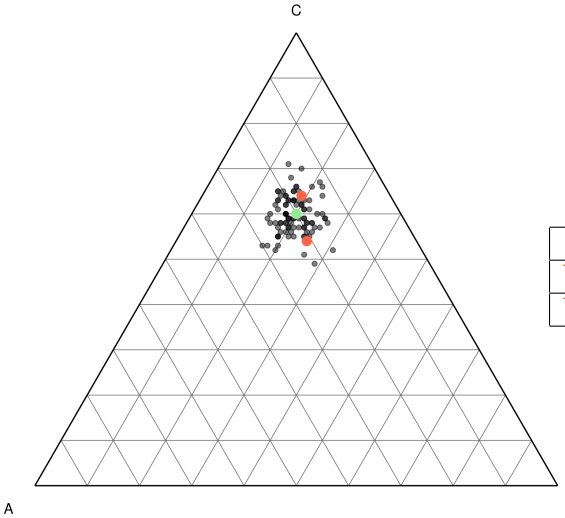
	A	B	C
π	0.2	0.2	0.6
$\mathbf{Y}_1$	17	19	64
$\mathbf{Y}_2$	21	25	54

Modèle multinomial



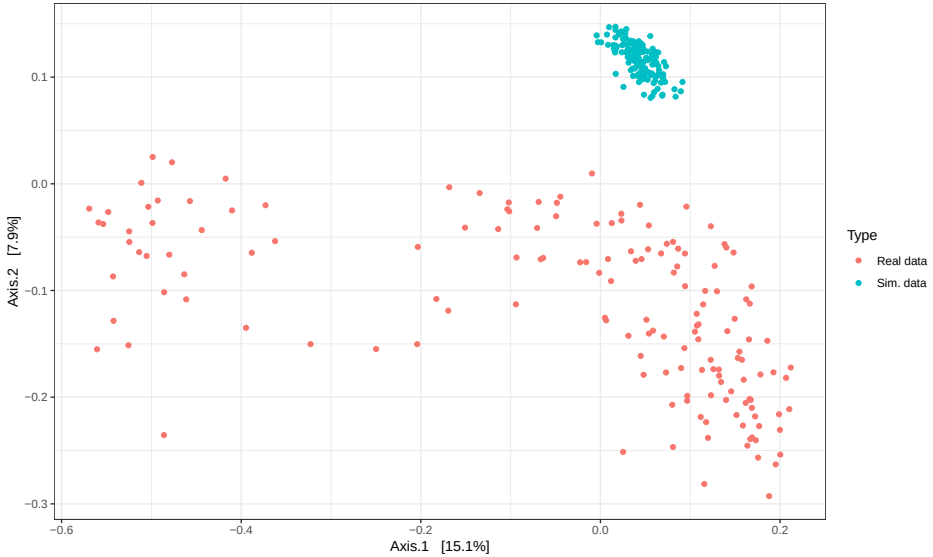
	A	B	C
π	0.2	0.2	0.6
$\mathbf{Y}_1$	17	19	64
$\mathbf{Y}_2$	21	25	54

Modèle multinomial



	A	B	C
π	0.2	0.2	0.6
$\mathbf{Y}_1$	17	19	64
$\mathbf{Y}_2$	21	25	54

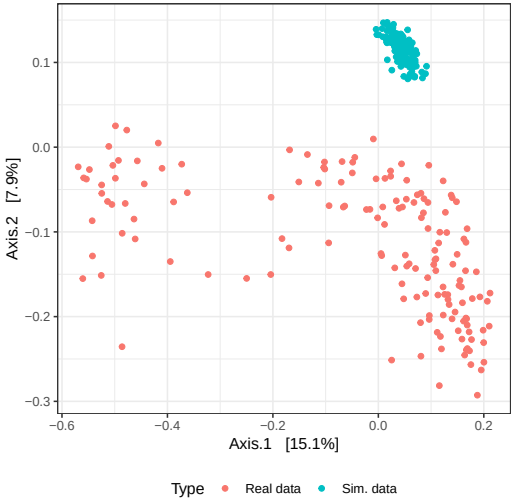
Exemple de modèle multinomial



Défauts

Hétérogénéité

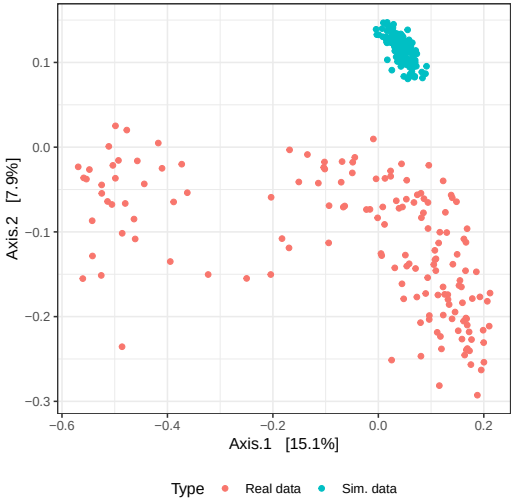
- Pas assez d'hétérogénéité



Défauts

Hétérogénéité

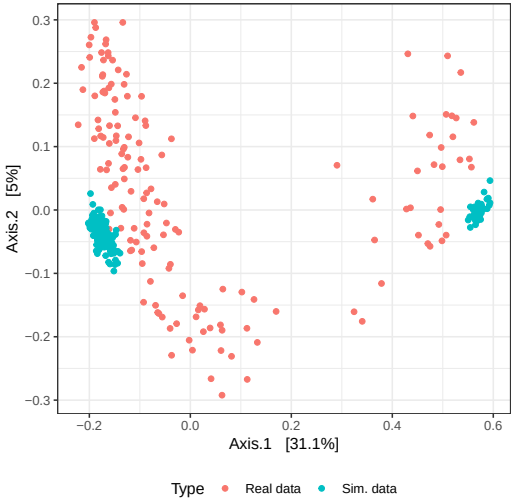
- Pas assez d'hétérogénéité
  ⇝ Ne s'ajuste qu'à une partie des données



Défauts

Hétérogénéité

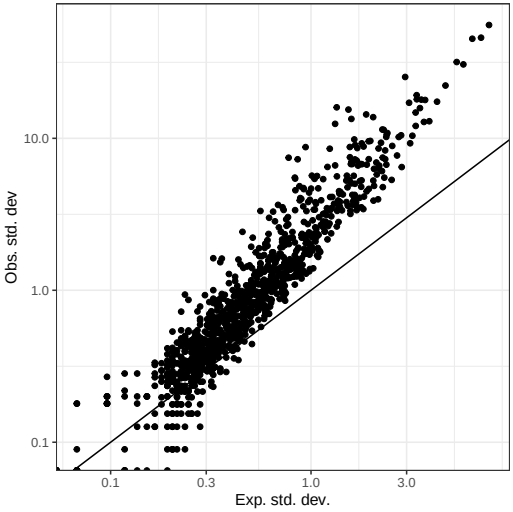
- Pas assez d'hétérogénéité
  ⇝ Ne s'ajuste qu'à une partie des données
- Variance faible



Défauts

Hétérogénéité

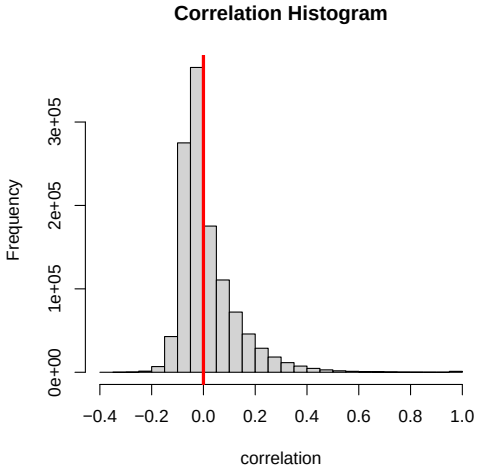
- Pas assez d'hétérogénéité
  $\rightsquigarrow$ Ne s'ajuste qu'à une partie des données
- Variance faible
- Dispersion faible



Défauts

Hétérogénéité

- Pas assez d'hétérogénéité
  $\rightsquigarrow$ Ne s'ajuste qu'à une partie des données
- Variance faible
- Dispersion faible
- Corrélations du mauvais signe



Avantages et Inconvénients

Avantages

+ Parcimonie: $p-1$ paramètres pour p abondances
+ Facile à estimer
+ Interprétable

Avantages et Inconvénients

Avantages

+ Parcimonie: $p-1$ paramètres pour p abondances
+ Facile à estimer
+ Interprétable

Inconvénients

- Mauvais pour hétérogénéité
- Mauvais pour la dispersion autour de la composition moyenne ($\simeq$ variabilité biologique)
- Mauvais pour les corrélations entre espèces

Outline

©Manfred Heyde

Modèles de mélange

Intuition

- Chaque échantillon appartient à un parmi K groupes
- Le groupe k est caractérisé par sa composition $\boldsymbol{\pi}_k$
- Un échantillon du groupe k a pour composition $\boldsymbol{\pi}_k$
- Les comptages sont échantillonnés suivant une loi multinomiale

Modèles de mélange

Intuition

- Chaque échantillon appartient à un parmi K groupes
- Le groupe k est caractérisé par sa composition $\boldsymbol{\pi}_k$
- Un échantillon du groupe k a pour composition $\boldsymbol{\pi}_k$
- Les comptages sont échantillonnés suivant une loi multinomiale

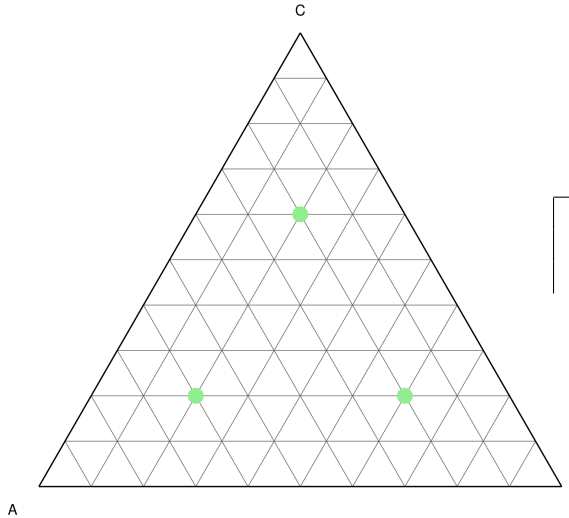
Modèle hiérarchique

$$Z \sim \mathcal{M}(1, \boldsymbol{\alpha})$$
$$Y|Z = k \sim \mathcal{M}(N, \boldsymbol{\pi}_k)$$

où

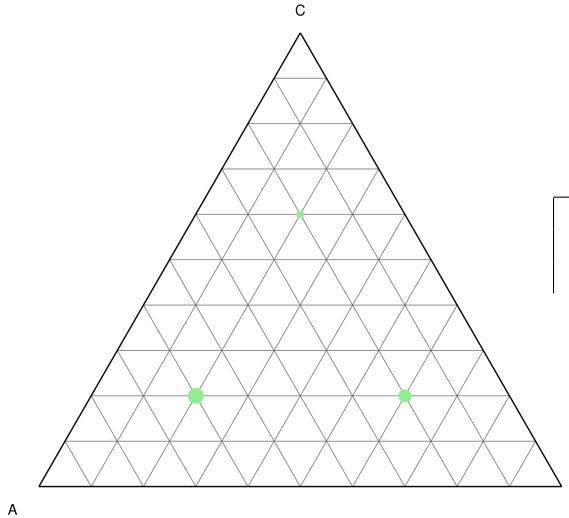
- $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)$ sont les proportions des K groupes,

Mélange de multinomiales



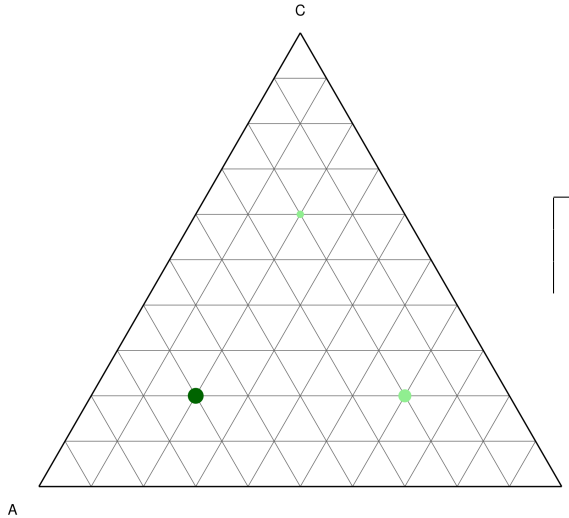
	A	B	C	α
π_1	0.6	0.2	0.2	
π_2	0.2	0.6	0.2	
π_3	0.2	0.2	0.6	

Mélange de multinomiales



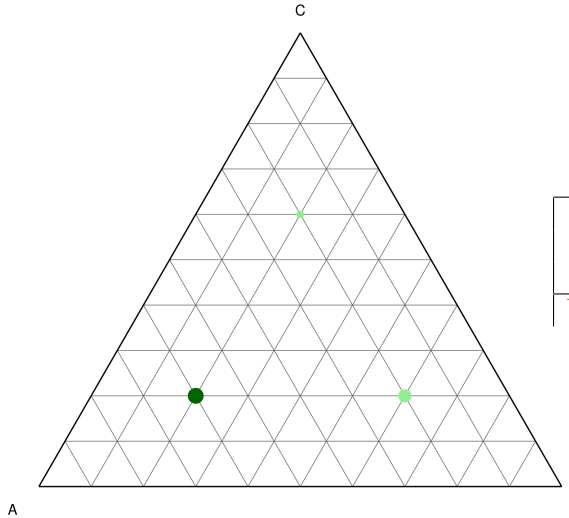
	A	B	C	α
π_1	0.6	0.2	0.2	0.5
π_2	0.2	0.6	0.2	0.4
π_3	0.2	0.2	0.6	0.1

Mélange de multinomiales



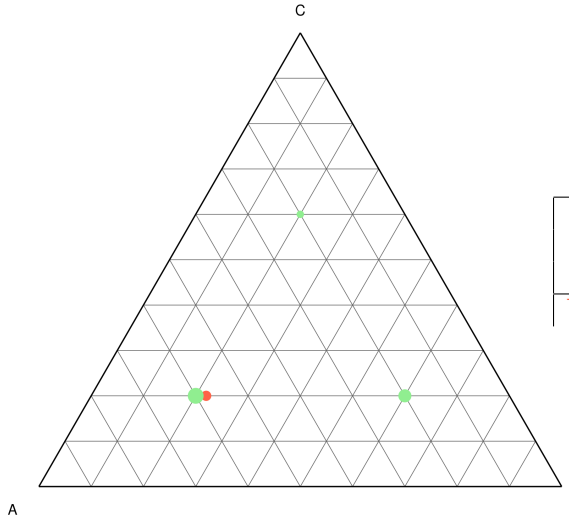
	A	B	C	α
$\boldsymbol{\pi}_1$	0.6	0.2	0.2	0.5
π_2	0.2	0.6	0.2	0.4
π_3	0.2	0.2	0.6	0.1

Mélange de multinomiales



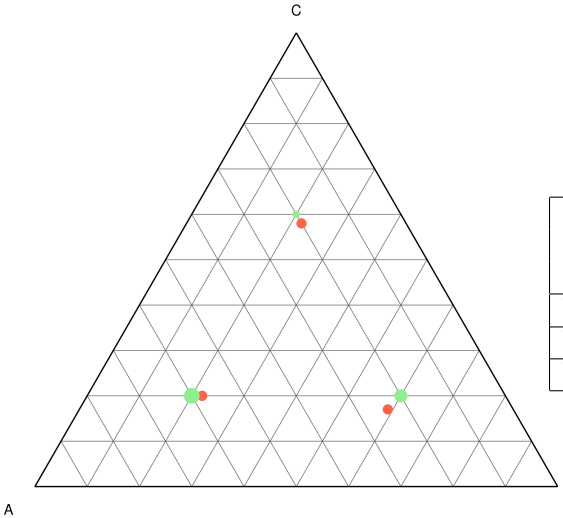
	A	B	C	α
$\boldsymbol{\pi}_1$	0.6	0.2	0.2	0.5
π_2	0.2	0.6	0.2	0.4
π_3	0.2	0.2	0.6	0.1
$\mathbf{Y}\|Z=1$	58	22	20	

Mélange de multinomiales



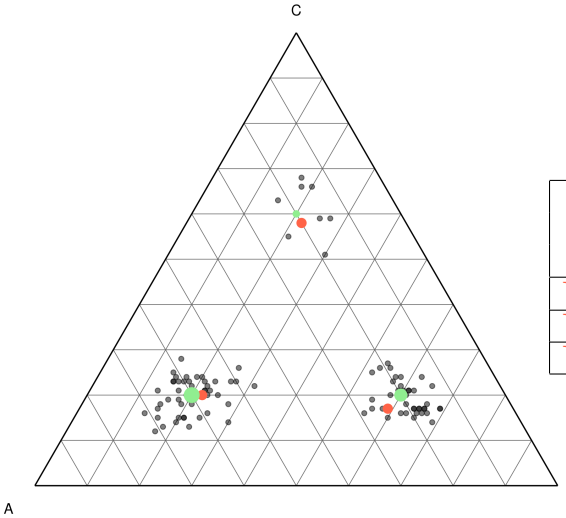
	A	B	C	α
$\boldsymbol{\pi}_1$	0.6	0.2	0.2	0.5
$\boldsymbol{\pi}_2$	0.2	0.6	0.2	0.4
$\boldsymbol{\pi}_3$	0.2	0.2	0.6	0.1
$\mathbf{Y}\|Z=1$	58	22	20	

Mélange de multinomiales



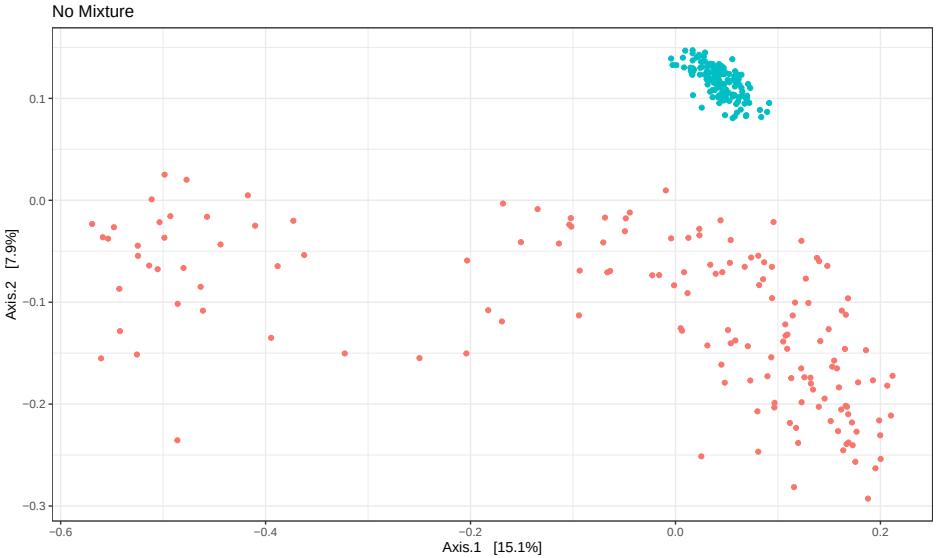
	A	B	C	α
$\boldsymbol{\pi}_1$	0.6	0.2	0.2	0.5
$\boldsymbol{\pi}_2$	0.2	0.6	0.2	0.4
$\boldsymbol{\pi}_3$	0.2	0.2	0.6	0.1
$\mathbf{Y}\|Z=1$	58	22	20	
$\mathbf{Y}\|Z=2$	24	59	17	
$\mathbf{Y}\|Z=3$	20	22	58	

Mélange de multinomiales

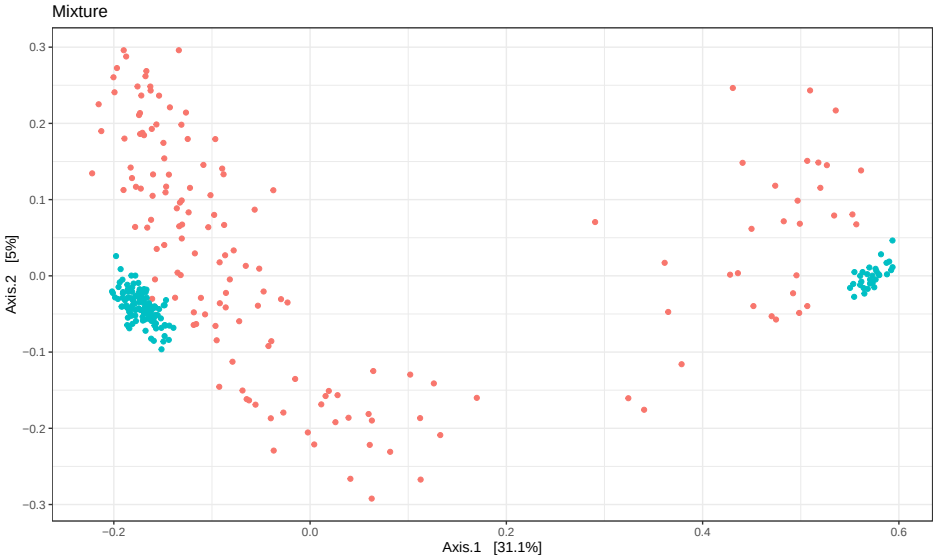


	A	B	C	α
$\boldsymbol{\pi}_1$	0.6	0.2	0.2	0.5
$\boldsymbol{\pi}_2$	0.2	0.6	0.2	0.4
$\boldsymbol{\pi}_3$	0.2	0.2	0.6	0.1
$\mathbf{Y}\|Z=1$	58	22	20	
$\mathbf{Y}\|Z=2$	24	59	17	
$\mathbf{Y}\|Z=3$	20	22	58	

Exemple de mélange de multinomial



Exemple de mélange de multinomial



Avantages et inconvénients

Avantages

+ Bon pour hétérogénéité
+ Parcimonieux: $Kp - 1$ paramètres pour K groups
+ L'inférence est facile quand les groupes sont connus $\rightsquigarrow$ simples moyennes

Avantages et inconvénients

Avantages

+ Bon pour hétérogénéité
+ Parcimonieux: $Kp - 1$ paramètres pour K groups
+ L'inférence est facile quand les groupes sont connus $\rightsquigarrow$ simples moyennes

Inconvénients

- Inférence moins facile quand les groupes sont inconnus
  $\rightsquigarrow$ algorithme itératif EM
- Mauvais pour la dispersion
- Mauvais pour les corrélations entre OTUs

Outline

Dirichlet - Multinomiale

Intuition

- π est la composition moyenne au niveau de l'écosystème

Dirichlet - Multinomiale

Intuition

- $\boldsymbol{\pi}$ est la composition moyenne au niveau de l'écosystème
- Échantillon i a sa **propre** composition $\boldsymbol{\pi}_i$ (version bruitée de $\boldsymbol{\pi}$) $\rightsquigarrow$ variabilité biologique

Dirichlet - Multinomiale

Intuition

- $\boldsymbol{\pi}$ est la composition moyenne au niveau de l'écosystème
- Échantillon i a sa **propre** composition $\boldsymbol{\pi}_i$ (version bruitée de $\boldsymbol{\pi}$) $\rightsquigarrow$ variabilité biologique
- N_i microbes tirés suivant $\mathcal{M}(1, \boldsymbol{\pi}_i)$ $\rightsquigarrow$ variabilité technique / échantillonnage

Dirichlet - Multinomiale

Intuition

- $\boldsymbol{\pi}$ est la composition moyenne au niveau de l'écosystème
- Échantillon i a sa **propre** composition $\boldsymbol{\pi}_i$ (version bruitée de $\boldsymbol{\pi}$) $\rightsquigarrow$ variabilité biologique
- N_i microbes tirés suivant $\mathcal{M}(1, \boldsymbol{\pi}_i)$ $\rightsquigarrow$ variabilité technique / échantillonnage

Modèle Hiérarchique

$$\begin{aligned} \boldsymbol{\pi} & \quad \text{Composition de l'écosystème} \\ \boldsymbol{\pi}_i \sim \mathcal{D}(\kappa\boldsymbol{\pi}) & \quad \text{Composition de l'échantillon} \\ \mathbf{Y}_i \sim \mathcal{M}(N_i, \boldsymbol{\pi}_i) & \quad \text{Comptages observés} \end{aligned}$$

où $1/\kappa$ modélise le niveau de variabilité (petit κ $\rightsquigarrow$ grande dispersion)

Dirichlet - Multinomiale

Intuition

- $\boldsymbol{\pi}$ est la composition moyenne au niveau de l'écosystème
- Échantillon i a sa **propre** composition $\boldsymbol{\pi}_i$ (version bruitée de $\boldsymbol{\pi}$) $\rightsquigarrow$ variabilité biologique
- N_i microbes tirés suivant $\mathcal{M}(1, \boldsymbol{\pi}_i)$ $\rightsquigarrow$ variabilité technique / échantillonnage

Modèle Hiérarchique

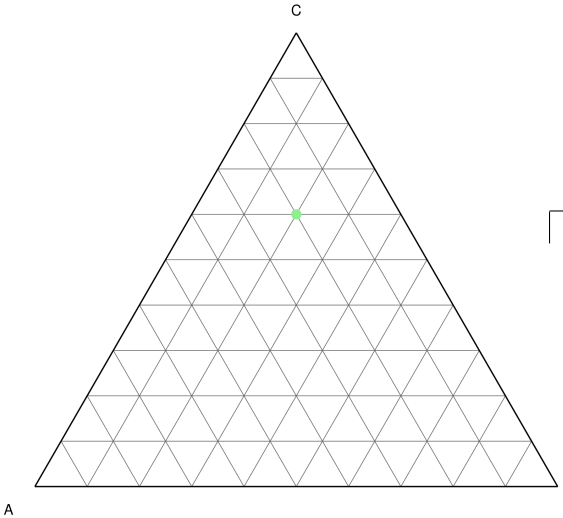
$$\begin{aligned} \boldsymbol{\pi} & \quad \text{Composition de l'écosystème} \\ \boldsymbol{\pi}_i \sim \mathcal{D}(\kappa\boldsymbol{\pi}) & \quad \text{Composition de l'échantillon} \\ \mathbf{Y}_i \sim \mathcal{M}(N_i, \boldsymbol{\pi}_i) & \quad \text{Comptages observés} \end{aligned}$$

où $1/\kappa$ modélise le niveau de variabilité (petit κ $\rightsquigarrow$ grande dispersion)

Couche de mélange

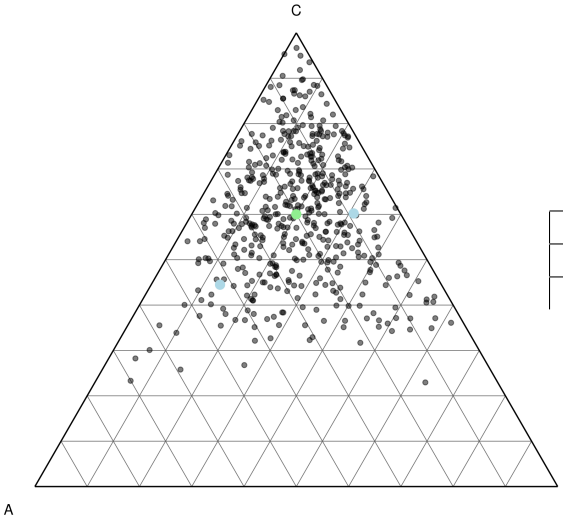
Ce modèle peut se combiner avec une couche de mélange

Distribution Dirichlet-Multinomiale



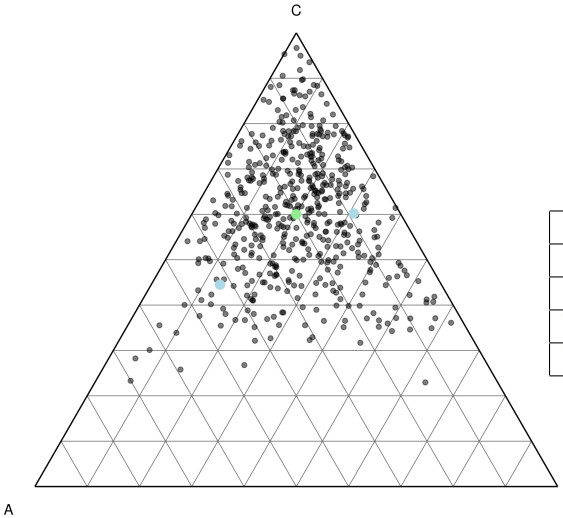
	A	B	C
π	0.2	0.2	0.6

Distribution Dirichlet-Multinomiale



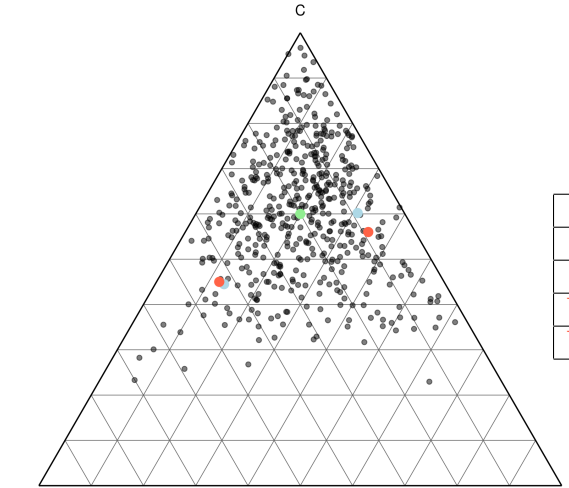
	A	B	C
π	0.2	0.2	0.6
π_1	0.089	0.309	0.602
π_2	0.423	0.132	0.445

Distribution Dirichlet-Multinomiale



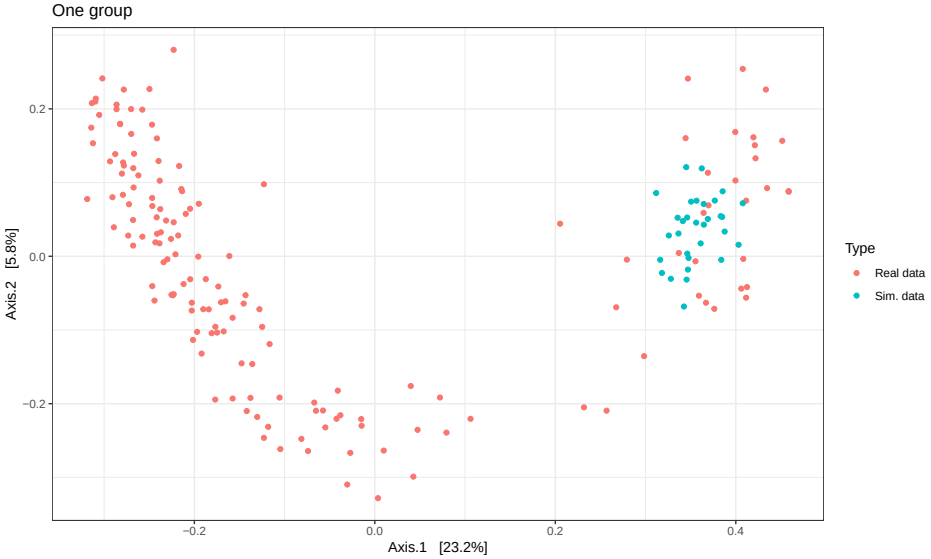
	A	B	C
$\boldsymbol{\pi}$	0.2	0.2	0.6
$\boldsymbol{\pi}_1$	0.089	0.309	0.602
$\boldsymbol{\pi}_2$	0.423	0.132	0.445
$\mathbf{Y}_1$	9	35	56
$\mathbf{Y}_2$	43	12	45

Distribution Dirichlet-Multinomiale

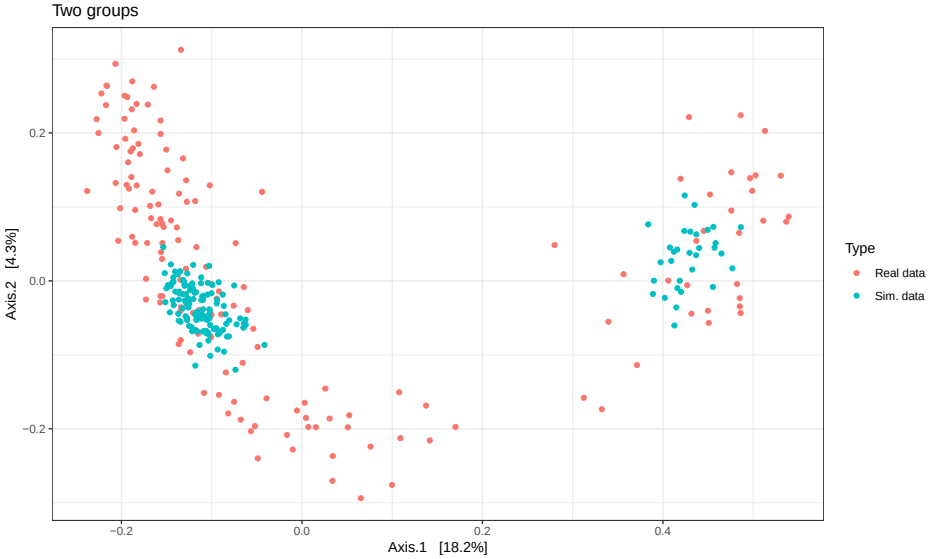


	A	B	C
$\boldsymbol{\pi}$	0.2	0.2	0.6
$\boldsymbol{\pi}_1$	0.089	0.309	0.602
$\boldsymbol{\pi}_2$	0.423	0.132	0.445
$\mathbf{Y}_1$	9	35	56
$\mathbf{Y}_2$	43	12	45

Exemple de Dirichlet-Multinomiale



Exemple de Dirichlet-Multinomiale (II)



Avantages et inconvénients

Avantages

+ Bon pour l'hétérogénéité
+ Moyen pour la dispersion
+ Parcimonieux: $K(p+1)-1$ paramètres pour K groupes

Avantages et inconvénients

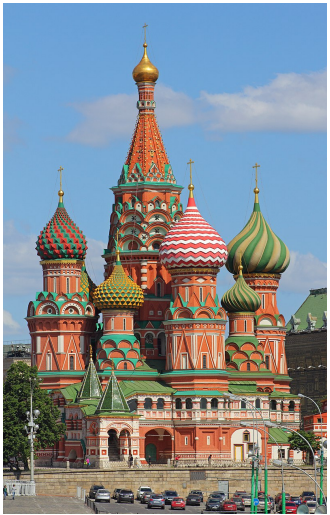
Avantages

+ Bon pour l'hétérogénéité
+ Moyen pour la dispersion
+ Parcimonieux: $K(p+1)-1$ paramètres pour K groupes

Inconvénients

- Inférence non triviale
  Groupes connus $\rightsquigarrow$ descente de gradient
  Groupes inconnus $\rightsquigarrow$ Algorithme itérative EM + descente de gradient
- Mauvais pour les corrélations entre OTUs

Outline



©A. Savin

Latent Dirichlet Allocation

Intuition

- Il y a K archétypes d'écosystemes $1, \ldots, K$

Latent Dirichlet Allocation

Intuition

- Il y a K archétypes d'écosystemes $1, \ldots, K$
- Chaque archétype a sa propre composition $\boldsymbol{\pi}_k$

Latent Dirichlet Allocation

Intuition

- Il y a K archétypes d'écosystemes $1, \ldots, K$
- Chaque archétype a sa propre composition $\boldsymbol{\pi}_k$
- Chaque échantillon $\mathbf{Y}$ est le mélange de plusieurs archétypes en proportions $(\theta_1, \ldots, \theta_K)$

Latent Dirichlet Allocation

Intuition

- Il y a K archétypes d'écosystemes $1, \ldots, K$
- Chaque archétype a sa propre composition $\boldsymbol{\pi}_k$
- Chaque échantillon $\mathbf{Y}$ est le mélange de plusieurs archétypes en proportions $(\theta_1, \ldots, \theta_K)$
- $\theta_k N$ microbes sont échantillonnés à partir d'une version bruitée de $\boldsymbol{\pi}_k$

Latent Dirichlet Allocation

Intuition

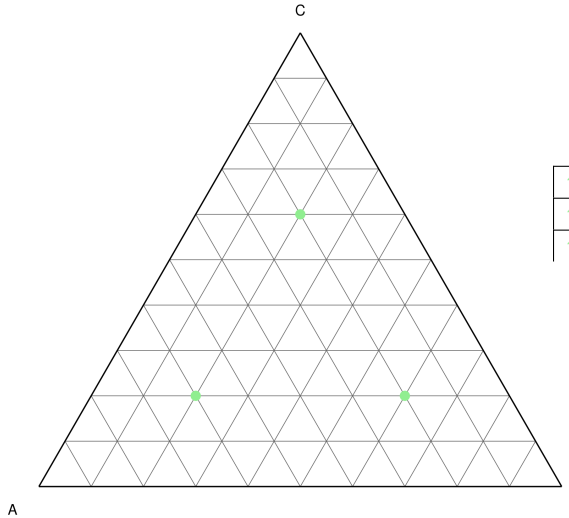
- Il y a K archétypes d'écosystemes $1, \dots, K$
- Chaque archétype a sa propre composition $\boldsymbol{\pi}_k$
- Chaque échantillon $\mathbf{Y}$ est le mélange de plusieurs archétypes en proportions $(\theta_1, \dots, \theta_K)$
- $\theta_k N$ microbes sont échantillonnés à partir d'une version bruitée de $\boldsymbol{\pi}_k$

Modèle Hiérarchique

$$
\begin{aligned}
\boldsymbol{\pi}_1, \dots, \boldsymbol{\pi}_K & \quad \text{Composition des archétypes} \\
\boldsymbol{\theta} \sim \mathcal{D}(\kappa \boldsymbol{\alpha}) & \quad \text{Proportion des archétypes dans un échantillon} \\
\tilde{\boldsymbol{\pi}}_k \sim \mathcal{D}(\kappa_k \boldsymbol{\pi}_k) & \quad \text{Version bruitée de } \boldsymbol{\pi}_k \\
z_i \sim \mathcal{M}(1, \boldsymbol{\theta}) & \quad \text{Archétype d'origine du comptage } i \\
w_i | z_i = k \sim \mathcal{M}(1, \tilde{\boldsymbol{\pi}}_k) & \quad \text{espèce du comptage} i
\end{aligned}
$$

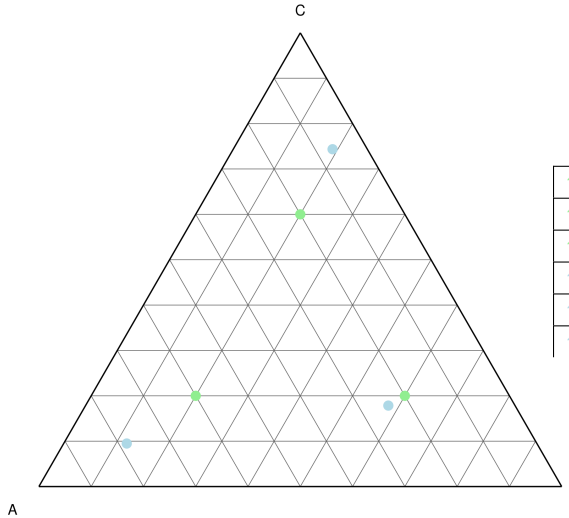
où κ et les κ_k contrôlent le niveau de variabilité.

Latent Dirichlet Allocation



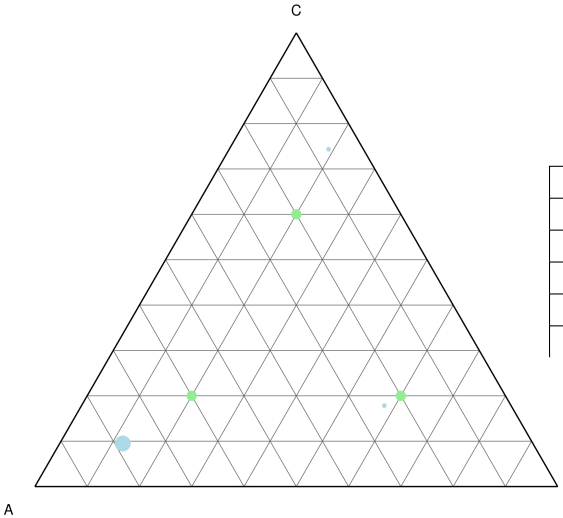
	A	B	C	θ
π_1	0.6	0.2	0.2	
π_2	0.2	0.6	0.2	
π_3	0.2	0.2	0.6	

Latent Dirichlet Allocation



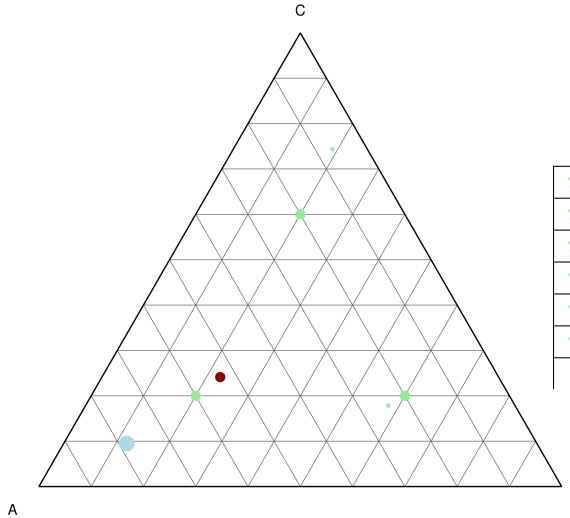
	A	B	C	θ
π_1	0.6	0.2	0.2	
π_2	0.2	0.6	0.2	
π_3	0.2	0.2	0.6	
$\tilde{\pi}_1$	0.784	0.121	0.095	
$\tilde{\pi}_2$	0.242	0.579	0.179	
$\tilde{\pi}_3$	0.423	0.132	0.445	

Latent Dirichlet Allocation



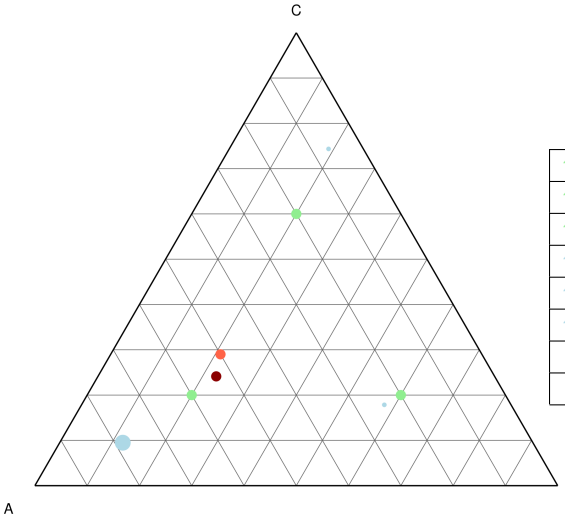
	A	B	C	θ
$\boldsymbol{\pi}_1$	0.6	0.2	0.2	
$\boldsymbol{\pi}_2$	0.2	0.6	0.2	
$\boldsymbol{\pi}_3$	0.2	0.2	0.6	
$\tilde{\boldsymbol{\pi}}_1$	0.784	0.121	0.095	0.6
$\tilde{\boldsymbol{\pi}}_2$	0.242	0.579	0.179	0.2
$\tilde{\boldsymbol{\pi}}_3$	0.423	0.132	0.445	0.2

Latent Dirichlet Allocation



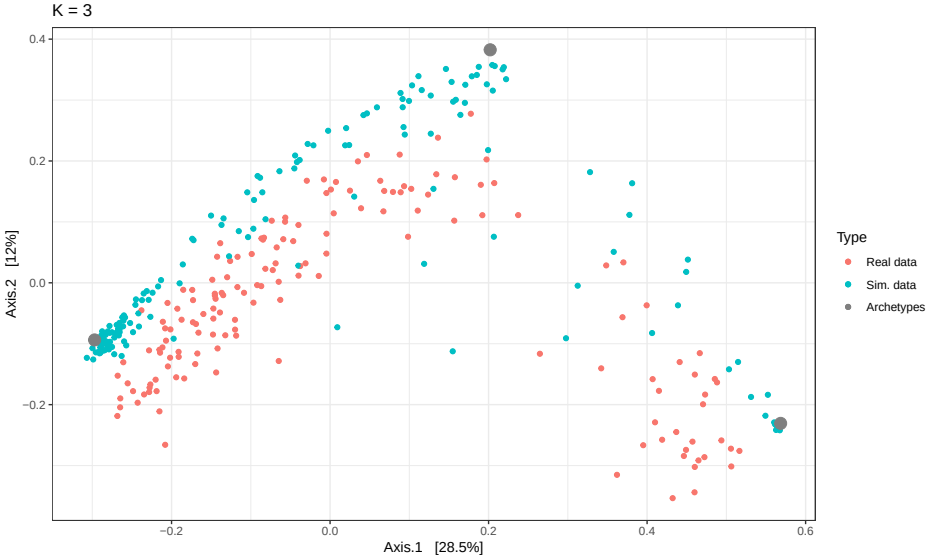
	A	B	C	θ
π_1	0.6	0.2	0.2	
π_2	0.2	0.6	0.2	
π_3	0.2	0.2	0.6	
$\tilde{\pi}_1$	0.784	0.121	0.095	0.6
$\tilde{\pi}_2$	0.242	0.579	0.179	0.2
$\tilde{\pi}_3$	0.423	0.132	0.445	0.2
π	0.532	0.226	0.241	

Latent Dirichlet Allocation

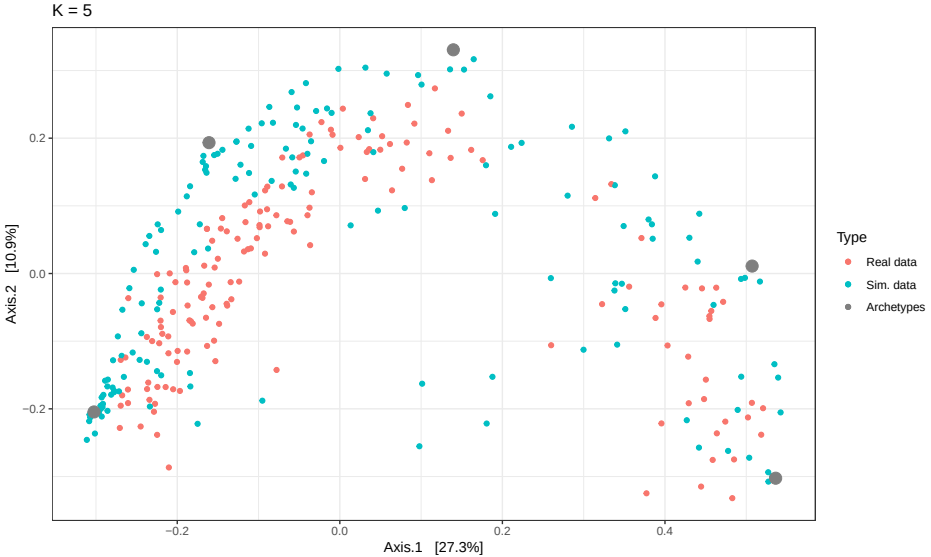


	A	B	C	θ
π_1	0.6	0.2	0.2	
π_2	0.2	0.6	0.2	
π_3	0.2	0.2	0.6	
$\tilde{\pi}_1$	0.784	0.121	0.095	0.6
$\tilde{\pi}_2$	0.242	0.579	0.179	0.2
$\tilde{\pi}_3$	0.423	0.132	0.445	0.2
$\boldsymbol{\pi}$	0.532	0.226	0.241	
$\mathbf{Y}$	54	18	28	

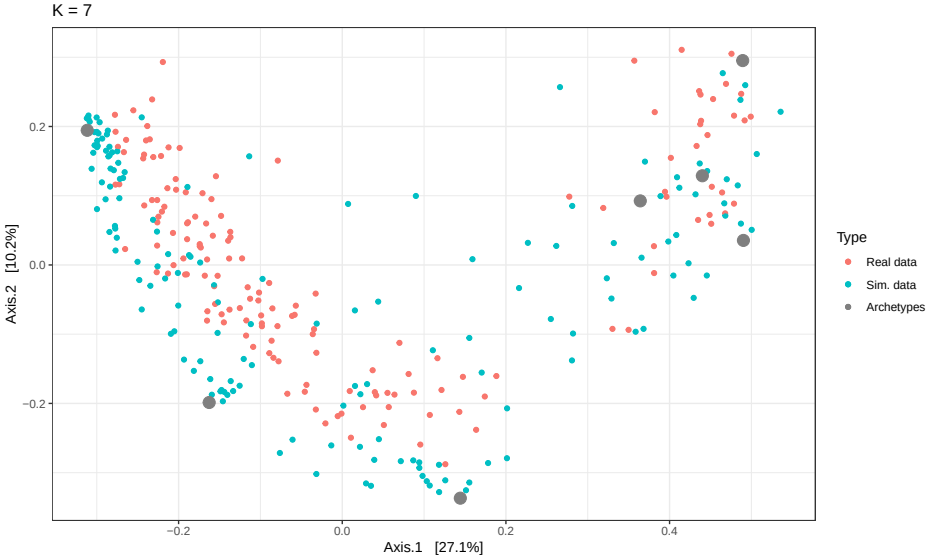
Exemple de Latent Dirichlet Allocation



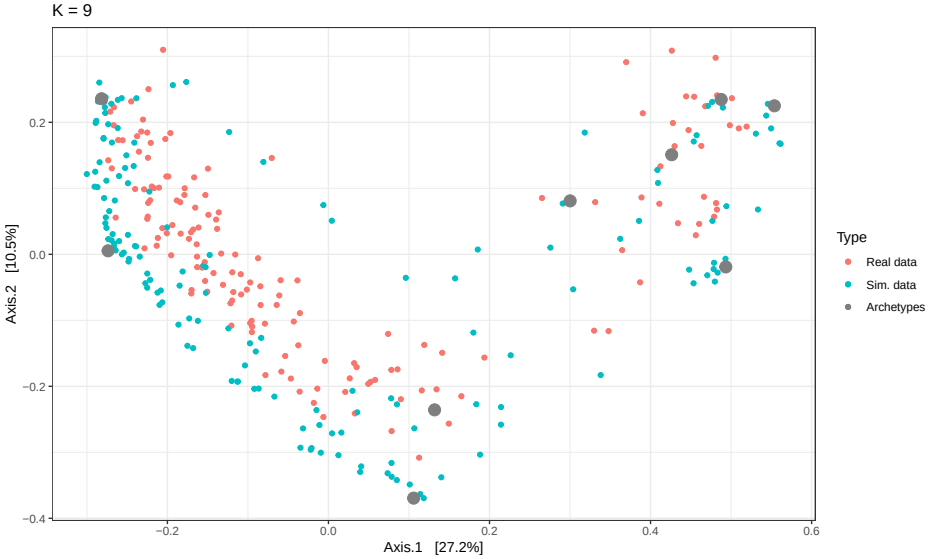
Exemple de Latent Dirichlet Allocation



Exemple de Latent Dirichlet Allocation



Exemple de Latent Dirichlet Allocation



Avantages et inconvénients

Avantages

+ Bon pour l'hétérogénéity
+ Bon pour la dispersion
+ Parcimonieux: $K(p+1)$ paramètres pour K archétypes

Avantages et inconvénients

Avantages

+ Bon pour l'hétérogénéity
+ Bon pour la dispersion
+ Parcimonieux: $K(p+1)$ paramètres pour K archétypes

Inconvénients

- Inférence moyennement complexe
  $\rightsquigarrow$ algorithme EM + descente de gradient / Échantillonnage de Gibbs
- Interprétation complexe $\rightsquigarrow$ les archétypes ne sont pas des groupes
- Mauvaises corrélations entre espèces

Résumé intermédiaire

Les modèles multinomiaux sont bons pour

- modéliser les compositions moyennes;
- modéliser la dispersion autour de ces moyennes;
- modéliser l'hétérogénéité;
- en utilisant (relativement) peu de paramètres

Résumé intermédiaire

Les modèles multinomiaux sont bons pour

- modéliser les compositions moyennes;
- modéliser la dispersion autour de ces moyennes;
- modéliser l'hétérogénéité;
- en utilisant (relativement) peu de paramètres

Les modèles multinomiaux sont mauvais pour

- modéliser les interactions entre espèces;
- prendre en compte les covariables;
- intégrer les données issues de différentes sources (*e.g.* 16S, ITS)

Outline

Modéliser les corrélations

Les modéles gaussiens multivariés sont le standard *de facto* pour modéliser les corrélations.

Modéliser les corrélations

Les modéles gaussiens multivariés sont le standard *de facto* pour modéliser les corrélations.

Pour des variables continues

- Les p variables $\mathbf{Y}_i$ (*e.g.* abondances d'espèces) sont expliquées
- par les valeurs de d covariables $\mathbf{X}_i$ et de p offsets $\mathbf{O}_i$

$$\mathbf{Y}_i = \underbrace{\mathbf{X}_i\mathbf{B}}_{\text{capture les covariables}} + \underbrace{\mathbf{O}_i}_{\text{captures les efforts d'observation}} + \varepsilon_i,\ \varepsilon_i \sim \mathcal{N}(\mathbf{0}_p, \underbrace{\mathbf{\Sigma}}_{\text{dépendances entre espèces}})$$

+ covariance nulle $\Leftrightarrow$ indépendance $\rightsquigarrow$ les espèces non-corrélées n'interagissent pas.

Modéliser les corrélations

Les modéles gaussiens multivariés sont le standard *de facto* pour modéliser les corrélations.

Pour des variables continues

- Les p variables $\mathbf{Y}_i$ (*e.g.* abondances d'espèces) sont expliquées
- par les valeurs de d covariables $\mathbf{X}_i$ et de p offsets $\mathbf{O}_i$

$$\mathbf{Y}_i = \underbrace{\mathbf{X}_i\mathbf{B}}_{\text{capture les covariables}} + \underbrace{\mathbf{O}_i}_{\text{captures les efforts d'observation}} + \varepsilon_i,\ \varepsilon_i \sim \mathcal{N}(\mathbf{0}_p, \underbrace{\mathbf{\Sigma}}_{\text{dépendances entre espèces}})$$

+ ~~covariance nulle $\Leftrightarrow$ indépendance $\rightsquigarrow$ les espèces non-corrélées n'interagissent pas.~~

Mais les abondances ne sont pas gaussiennes...

Modéliser les corrélations

Les modéles gaussiens multivariés sont le standard *de facto* pour modéliser les corrélations.

Pour des variables continues

- Les p variables $\mathbf{Y}_i$ (*e.g.* abondances d'espèces) sont expliquées
- par les valeurs de d covariables $\mathbf{X}_i$ et de p offsets $\mathbf{O}_i$

$$\mathbf{Y}_i = \underbrace{\mathbf{X}_i\mathbf{B}}_{\text{capture les covariables}} + \underbrace{\mathbf{O}_i}_{\text{captures les efforts d'observation}} + \varepsilon_i,\ \varepsilon_i \sim \mathcal{N}(\mathbf{0}_p, \underbrace{\boldsymbol{\Sigma}}_{\text{dépendances entre espèces}})$$

+ ~~covariance nulle $\Leftrightarrow$ indépendance $\rightsquigarrow$ les espèces non-corrélées n'interagissent pas.~~

Mais les abondances ne sont pas gaussiennes...

Modèle à variable latente avec une couche *latente* gaussienne et une couche *observée* comptage

Outline

©cbrettre

Multinomiale Log-Normale

Intuition

- La couche latente modélise des abondances de bases $\mathbf{z}$

Multinomiale Log-Normale

Intuition

- La couche latente modélise des abondances de bases $\mathbf{z}$
- Ces abondances sont transformées en une composition moyenne $\boldsymbol{\pi}$

Multinomiale Log-Normale

Intuition

- La couche latente modélise des abondances de bases $\mathbf{z}$
- Ces abondances sont transformées en une composition moyenne $\boldsymbol{\pi}$
- N microbes sont tirés suivant une multinomiale de paramètre $\boldsymbol{\pi}$

Multinomiale Log-Normale

Intuition

- La couche latente modélise des abondances de bases $\mathbf{z}$
- Ces abondances sont transformées en une composition moyenne $\boldsymbol{\pi}$
- N microbes sont tirés suivant une multinomiale de paramètre $\boldsymbol{\pi}$

Modèle hiérarchique

$$\mathbf{z} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \qquad \text{abondances de base}$$

$$\boldsymbol{\pi}|\mathbf{z} = \left(\frac{e^{z_j}}{\sum_{j'} e^{z_{j'}}} \right)_j \qquad \text{composition moyenne}$$

$$\mathbf{Y} \sim \mathcal{M}(N, \boldsymbol{\pi}) \qquad \text{comptages observés}$$

Multinomiale Log-Normale

Intuition

- La couche latente modélise des abondances de bases $\mathbf{z}$
- Ces abondances sont transformées en une composition moyenne $\boldsymbol{\pi}$
- N microbes sont tirés suivant une multinomiale de paramètre $\boldsymbol{\pi}$

Modèle hiérarchique

$$\mathbf{z} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \qquad \text{abondances de base}$$

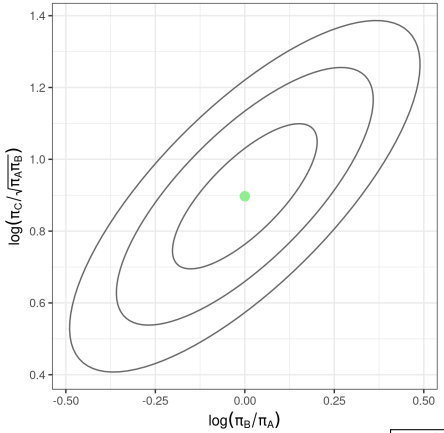
$$\boldsymbol{\pi}|\mathbf{z} = \left(\frac{e^{z_j}}{\sum_{j'} e^{z_{j'}}} \right)_j \qquad \text{composition moyenne}$$

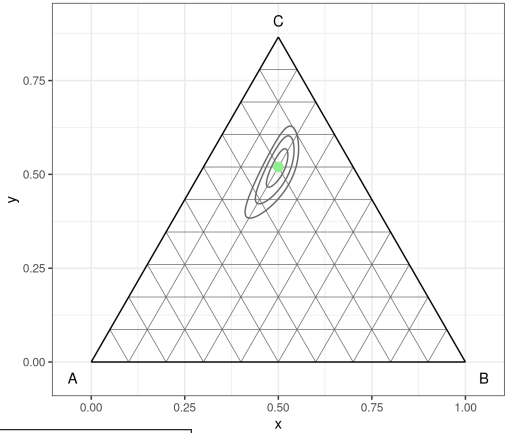
$$\mathbf{Y} \sim \mathcal{M}(N, \boldsymbol{\pi}) \qquad \text{comptages observés}$$

Couche de mélange

On peut rajouter une couche de mélange au modèle pour l'hétérogénéité.

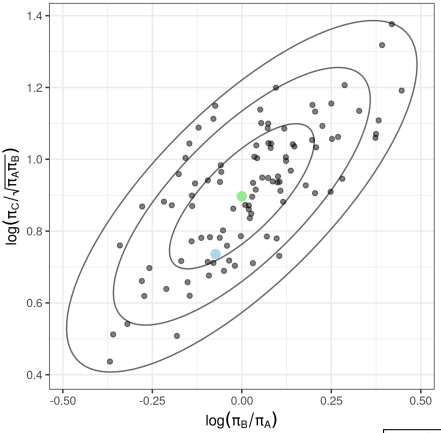
Multinomiale Log-Normale

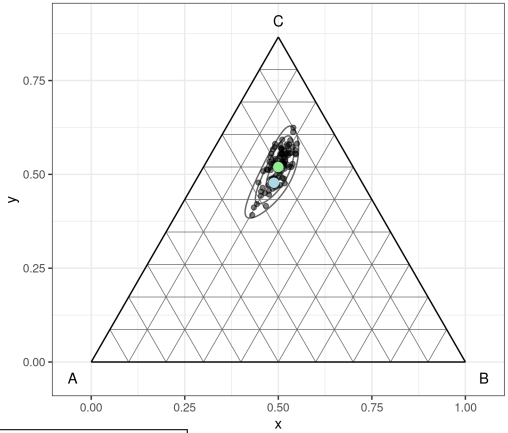




	A	B	C
π	0.2	0.2	0.6

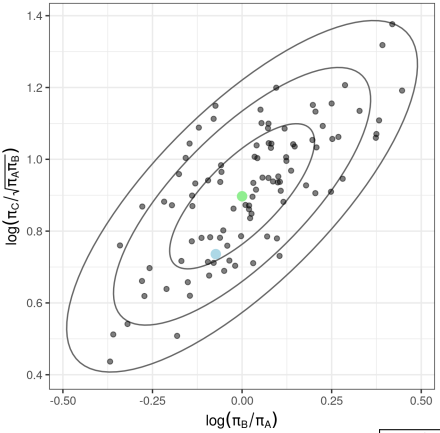
Multinomiale Log-Normale

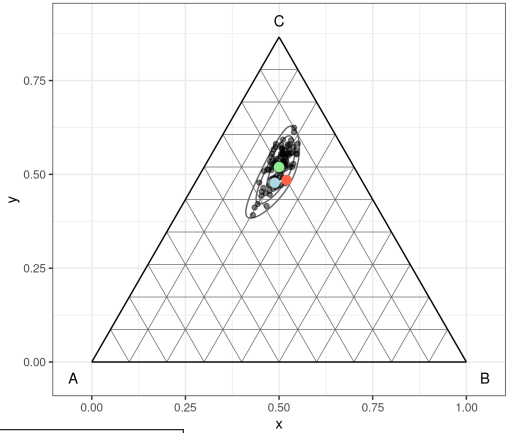




	A	B	C
π	0.2	0.2	0.6
π_1	0.235	0.213	0.552

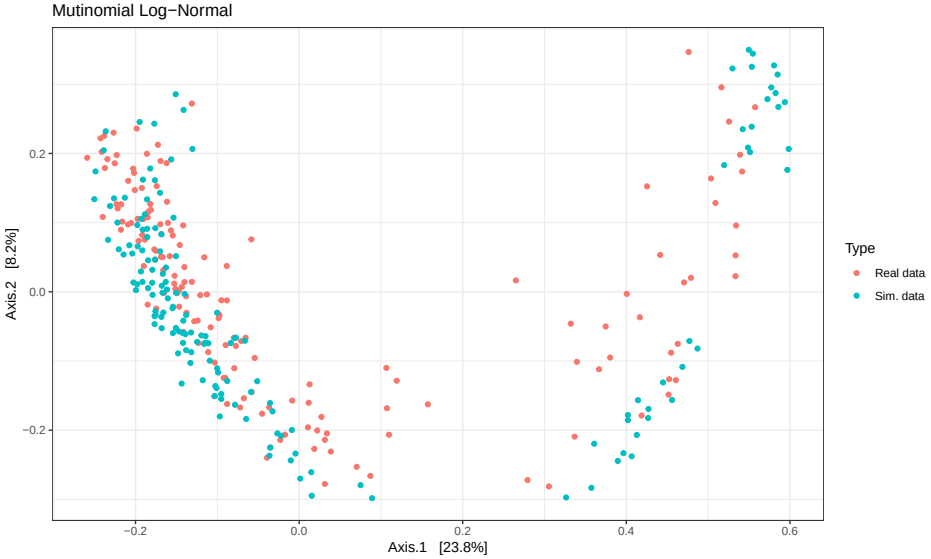
Multinomiale Log-Normale





	A	B	C
π	0.2	0.2	0.6
π_1	0.235	0.213	0.552
Y	20	24	56

Exemple de Multinomiale Log-Normale



Avantages et inconvénients

Avantages

+ Bon pour l'hétérogénéité
+ Bon pour la dispersion
+ Bon pour les corrélations entre espèces

Avantages et inconvénients

Avantages

+ Bon pour l'hétérogénéité
+ Bon pour la dispersion
+ Bon pour les corrélations entre espèces

Inconvénients

- Modèle non-parcimonieux: $p(p+3)/2$ paramètres
- Inférence complexe
  $\rightsquigarrow$ Algorithme EM itératif / Échantillonnage de Gibbs
- Modélisation faite au niveau des compositions et pas des comptages.

Outline

Distribution Poisson-log normale [AH89]

Intuition

- La couche latente modélise des bases $\mathbf{z}$

Distribution Poisson-log normale [AH89]

Intuition

- La couche latente modélise des bases $\mathbf{z}$
- Les bases sont *transformées* en comptages moyens

Distribution Poisson-log normale [AH89]

Intuition

- La couche latente modélise des bases $\mathbf{z}$
- Les bases sont *transformées* en comptages moyens
- Les comptages des microbes sont *tirés* suivant une distribution de Poisson

Distribution Poisson-log normale [AH89]

Intuition

- La couche latente modélise des bases $\mathbf{z}$
- Les bases sont *transformées* en comptages moyens
- Les comptages des microbes sont *tirés* suivant une distribution de Poisson

Modèle hiérarchique

$$\mathbf{z} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \quad \text{Bases}$$

$$\lambda_j | \mathbf{z} = e^{z_j} \quad \text{Comptage moyen de l'espèce } j$$

$$Y_j | z_j \sim \mathcal{P}(e^{z_j}) \quad \text{Comptage observé de l'espèce } j$$

Distribution Poisson-log normale [AH89]

Intuition

- La couche latente modélise des bases $\mathbf{z}$
- Les bases sont *transformées* en comptages moyens
- Les comptages des microbes sont *tirés* suivant une distribution de Poisson

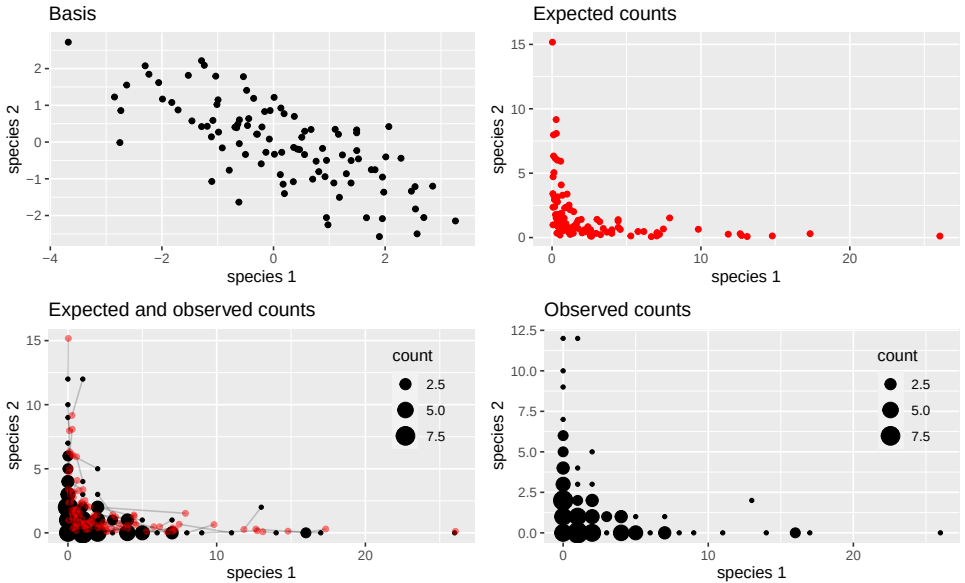
Modèle hiérarchique

$\mathbf{z} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$	Bases
$\lambda_j \vert \mathbf{z} = e^{z_j}$	Comptage moyen de l'espèce j
$Y_j \vert z_j \sim \mathcal{P}(e^{z_j})$	Comptage observé de l'espèce j

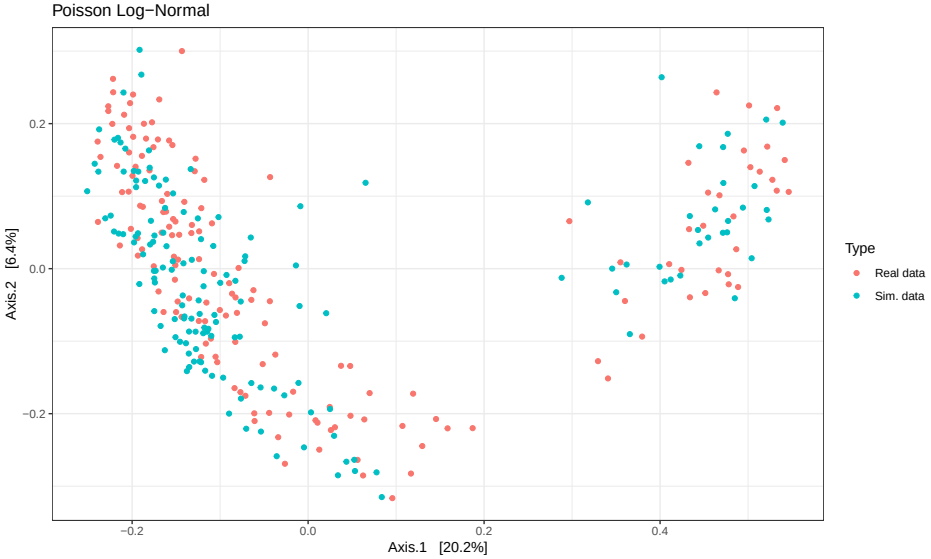
Couche mélange

On peut rajouter une couche de mélange au modèle pour l'hétérogénéité.

Interprétation géométrique



Exemple de Poisson Log-Normale



Avantages et inconvénients

Avantages

+ Bon pour l'hétérogénéité
+ Bon pour la dispersion
+ Bon pour les corrélations entre espèces
+ Modélisation faite au niveau des comptages
  $\rightsquigarrow$ les comptages peuvent être à des échelles différentes et provenir de différentes sources

Avantages et inconvénients

Avantages

+ Bon pour l'hétérogénéité
+ Bon pour la dispersion
+ Bon pour les corrélations entre espèces
+ Modélisation faite au niveau des comptages
  $\rightsquigarrow$ les comptages peuvent être à des échelles différentes et provenir de différentes sources

Inconvénients

- Modèle non-parcimonieux: $p(p+3)/2$ paramètres
- Inférence complexe
  $\rightsquigarrow$ Algorithme EM itératif + descente de gradient
- Comptage total seulement controlé en moyenne

Résumé intermédiaire

Les modèles Log-Normaux sont bons pour

- modéliser les compositions moyennes;
- modéliser la dispersion autour de ces moyennes;
- modéliser l'hétérogénéité;
- modéliser les interactions entre espèces;
- corriger les effets de covariables à l'aide d'un modèle linéaire

Résumé intermédiaire

Les modèles Log-Normaux sont bons pour

- modéliser les compositions moyennes;
- modéliser la dispersion autour de ces moyennes;
- modéliser l'hétérogénéité;
- modéliser les interactions entre espèces;
- corriger les effets de covariables à l'aide d'un modèle linéaire

Les modèles log-normaux sont mauvais pour

- leur grand nombre de paramètres
- la complexité de leurs méthodes d'inférence

Résumé intermédiaire

Les modèles Log-Normaux sont bons pour

- modéliser les compositions moyennes;
- modéliser la dispersion autour de ces moyennes;
- modéliser l'hétérogénéité;
- modéliser les interactions entre espèces;
- corriger les effets de covariables à l'aide d'un modèle linéaire

Les modèles log-normaux sont mauvais pour

- leur grand nombre de paramètres
- la complexité de leurs méthodes d'inférence

- Les modèles MLN sont plus faciles à interpréter (compositions)
- Les modèles PLN permettent de mixer les données de différentes sources (16S, ITS, etc.)

Outline

Modèle PLN en écologie microbienne

PLN: un modèle flexible pour prendre en compte:

- l'hétérogénéité et les comptages moyens ($\simeq$ moment de premier ordre)
- la dispersion et les corrélations entre espèces ($\simeq$ moment de second ordre)
- des covariables et des facteurs de confusion
- des comptages issus de différentes sources

Modèle PLN en écologie microbienne

PLN: un modèle flexible pour prendre en compte:

- l'hétérogénéité et les comptages moyens ($\simeq$ moment de premier ordre)
- la dispersion et les corrélations entre espèces ($\simeq$ moment de second ordre)
- des covariables et des facteurs de confusion
- des comptages issus de différentes sources

Permet de se ramener à des analyses multivariées *classiques*:

Idée: Mettre des contraintes sur le modèle

- ACP $\rightsquigarrow$ faible rang sur $\boldsymbol{\Sigma}$
- Analyse discriminante $\rightsquigarrow$ structure (connue) de groupe sur $\boldsymbol{\mu}$
- Inférence de réseaux $\rightsquigarrow$ $\boldsymbol{\Sigma}^{-1}$ parcimonieuse
- Modèles de mélange $\rightsquigarrow$ structure (inconnue) de groupe sur $\boldsymbol{\mu}$
- *etc.*

Outline

PLN-PCA: résumer l'information

Réduction de dimension et visualisation. Tâches typiques en analyse multivariée

$$\mathbf{Z}_i \text{ iid} \sim \mathcal{N}_p(\mathbf{0}_p, \boldsymbol{\Sigma}), \qquad \text{rank}(\boldsymbol{\Sigma}) = q \ll p$$
$$\mathbf{Y}_i \,|\, \mathbf{Z}_i \sim \mathcal{P}(\exp\{\mathbf{O}_i + \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\})$$

$\rightsquigarrow$ Trouver une base de faible dimension (axes de l'ACP) pour représenter la covariance latente

Ajustement d'un modèle PLNPCA avec offset et covariables.

```
Qmax = 30; Q <- 1:Qmax;

## Model with offset
models.offset <- PLNPCA(counts ~ 1 + offset(log(offsets)), ranks=Q)

## Models with offset and covariates (tree + orientation)
formula <- counts ~ 1 + covariates$tree + covariates$orientation + offset(log(offsets))
models.tree.orientation <- PLNPCA(formula, ranks=Q) # approx 10 mn
```

PCA: visualisation

PLN-PCA séparent les échantillons suivant l'arbre d'origine

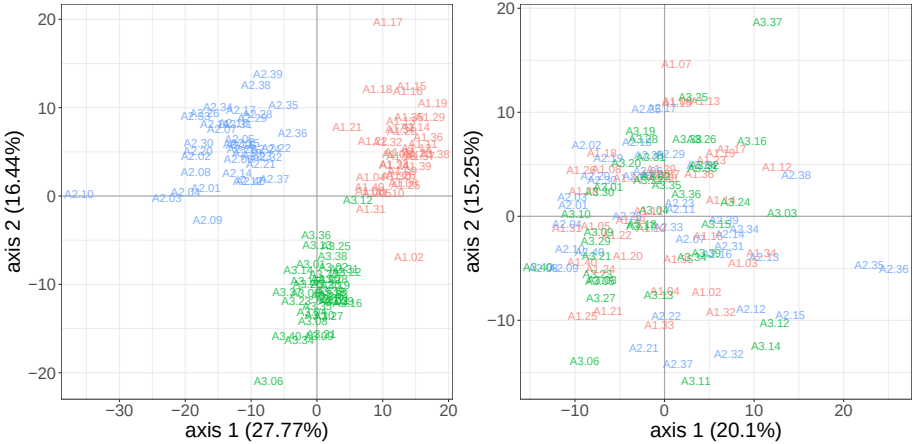


Figure: Offset seul Offset + covariables

PCA: visualisation II

L'introduction de covariables fait émerger des structures

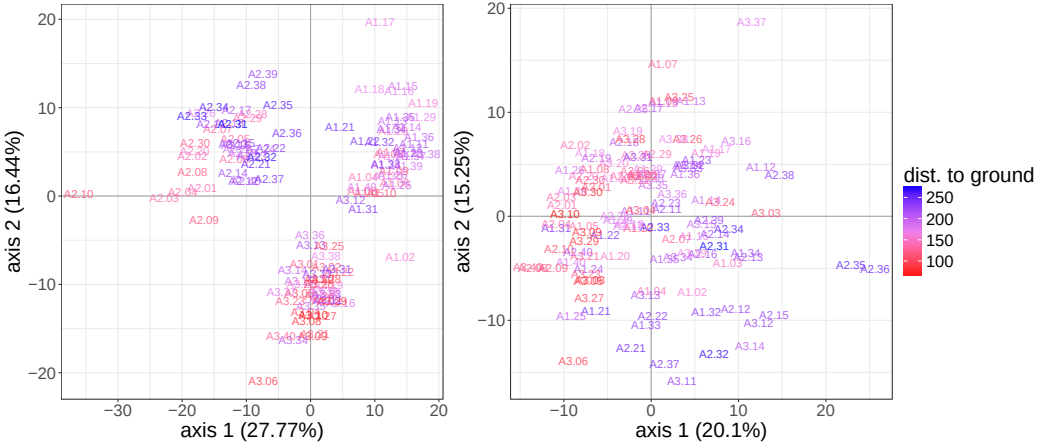


Figure: Offset seul Offset + covariables

Outline

Estimation des modèles PLN-LDA

Trouver la combinaison linéaire qui sépare les groupes

Estimation du modèle, avec offsets et covariables.

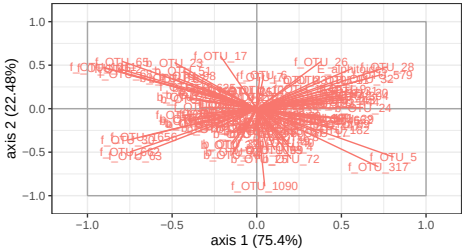
```
myLDA_tree   <- PLNLDA(Abundance ~ offset(log(Offset)), grouping = tree, data = oaks)

##
##  Performing discriminant Analysis...
##  DONE!
```

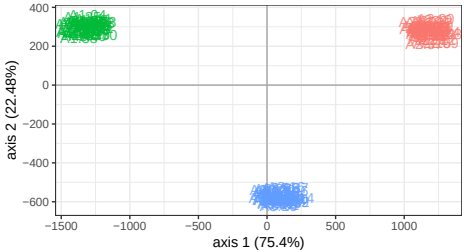
```
myLDA_tree$plot_LDA()
```

Analyse discriminante sur l'arbre d'origine

Axes contribution

axis 1 : 75.4%
axis 2 : 22.48%





classification

a susceptible

a intermediate

a resistant

Erreur de prédiction (10-fold validation croisée)

	susceptible	intermediate	resistant
intermediate	0	38	0
resistant	0	0	39
susceptible	39	0	0

Outline

Réseau pour données de comptages

Même cadre:

$$\begin{array}{rccc} \text{Espace des paramètres } \mathbb{R}^p & \mathbb{E}[\mathbf{Z}_i] = \mu_i & & \mathbf{Z}_i \sim \mathcal{N}(\mu_i, \mathbf{\Sigma} = \mathbf{B}^\intercal\mathbf{B}) \\ \text{Espace des observations } \mathbb{N}^p & Y_{ij}|Z_{ij} & \text{indep.} & Y_{ij}|Z_{ij} \sim \mathcal{P}(e^{Z_{ij}}) \end{array}$$

où $\mathbf{B}$ est de (faible-)rang $q \ll p$.

Réseau pour données de comptages

Même cadre, contraintes différentes:

$$\begin{array}{rccc} \text{Espace des paramètres } \mathbb{R}^p & \mathbb{E}[\mathbf{Z}_i] = \mu_i & & \mathbf{Z}_i \sim \mathcal{N}(\mu_i, \mathbf{\Sigma} = \mathbf{\Omega}^{-1}) \\ \text{Espace des observations } \mathbb{N}^p & Y_{ij}|Z_{ij} & \text{indep.} & Y_{ij}|Z_{ij} \sim \mathcal{P}(e^{Z_{ij}}) \end{array}$$

où $\mathbf{\Omega}$ est creuse et reflète la topologie du réseau.

Inférence variationnelle

EM variationnel Maximiser une borne inf J de la vraisemblance sous contrainte de parcimonie sur $\mathbf{\Omega}$:

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

Inférence variationnelle

EM variationnel Maximiser une borne inf J de la vraisemblance sous contrainte de parcimonie sur $\mathbf{\Omega}$:

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

Itérer jusqu'à convergence entre

- step 1: Mise à jour des paramètres $(\mathbf{M},\mathbf{S},\mathbf{\Theta})$ par descente de gradient

Inférence variationnelle

EM variationnel Maximiser une borne inf J de la vraisemblance sous contrainte de parcimonie sur $\mathbf{\Omega}$:

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

Itérer jusqu'à convergence entre

- step 1: Mise à jour des paramètres $(\mathbf{M},\mathbf{S},\mathbf{\Theta})$ par descente de gradient
- step 2: Mise à jour du réseau $\mathbf{\Omega}$ par lasso graphique (*glasso*)

Inférence variationnelle

EM variationnel Maximiser une borne inf J de la vraisemblance sous contrainte de parcimonie sur $\mathbf{\Omega}$:

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

Itérer jusqu'à convergence entre

- step 1: Mise à jour des paramètres $(\mathbf{M},\mathbf{S},\mathbf{\Theta})$ par descente de gradient
- step 2: Mise à jour du réseau $\mathbf{\Omega}$ par lasso graphique (*glasso*)

λ sélectionné en utilisant la méthode de rééchantillonnage StARS [LRW10].

Un example miroir de celui de Thibaud

Données

Suffrages de 63 242 bureaux de vote lors des élections présidentielles de 2017 (premier tour)

Objectif

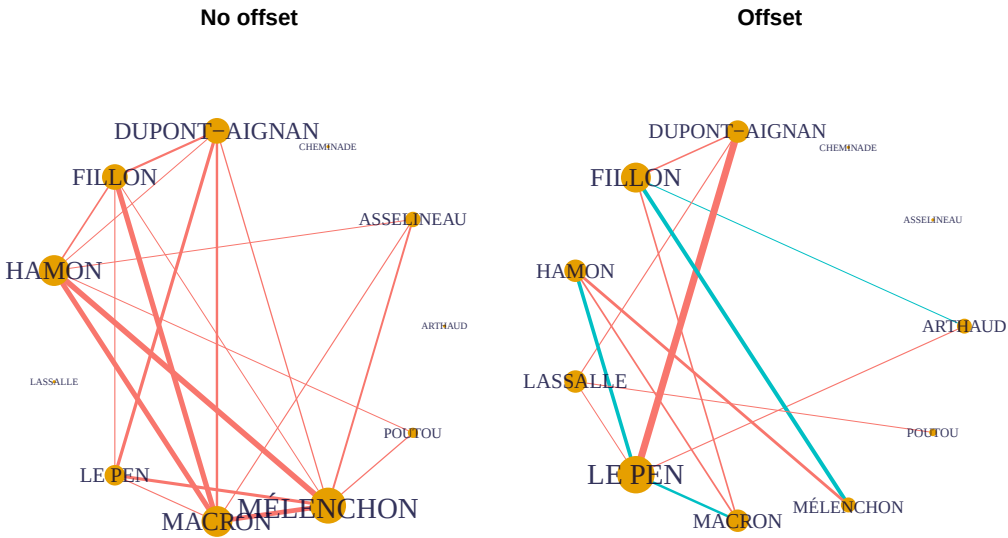
Reconstruire les corrélations partielles entre candidats.

Méthode

On ajuste deux modèles:

- Un modèle naïf sans offset (pas de prise en compte de la compositionalité)
- Un modèle moins naïf avec la taille du bureau de vote comme offset.

La compositionalité compte



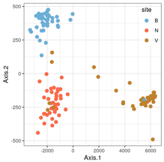
Conclusion

Résumé PLN = modèle générique pour les données de comptages

- Prise en compte des effets compositionels (via les *offset*)
- Prise en compte des covariables
- Modélisation statistique flexible
- `PLNmodels` R-package

Extension

- Ajout des zéros essentiels ("zéro-inflation")
- Extensions: ACP parcimonieuse
- Intervalle de confiance et tests
- Données manquantes...



Classification accuracy: 94.3%
(work with S. Even)

PLN-LDA: compare sites

Goal: find **systematic differences** between sites in different **classes.**

Constraint: **$\mu = \mu_k$** if site in **known** class k

PLN-mixture: find groups

Goal: cluster sites into **homogeneous groups**

Constraint: **$\mu = \mu_k$** if site in **unknown** group k

Constrain species abundances μ

Input

Counts Y

	sp 1	sp j	sp p
site 1	14	0	78
site j	36	34	56
site n	80	0	12

Covariates X

	cov 1	cov h	cov d
site 1	A	1.2	hi
site j	A	2.6	lo
site n	B	3.8	lo

Standard PLN

Latent layer
$Z \sim N(\mu, \Sigma)$

$\mu = X\Theta$: environmental effects

PLNmodels

Obs. counts
$Y|Z \sim P(e^Z)$

Σ: species dependencies

Output

Fixed effects Θ

	sp 1	sp j	sp p
cov 1	0.1	0.6	1.2
cov h	0.3	2.3	1.1
cov d	0.4	-1.2	1.4

Effect of cov. h on sp. j

Variance Σ

	sp 1	sp j	sp p
sp 1	1	-0.2	0.3
sp j	-0.2	2	1.1
sp p	0.3	1.1	3

Covariation between Species j and p

Constrain species dependencies Σ

PLN-PCA: find structure

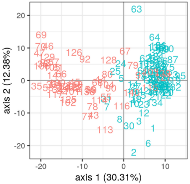
Goal: find **few structuring factors** governing species dependencies

Model: force **Σ** to have **low rank**

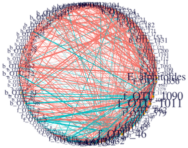
PLN-network: find interactions

Goal: find pairs of species in **direct interaction**

Model: force **$\Omega^{-1} = \Sigma$** to be **sparse**



Work with N. Peyrard and M.-J. Cros



Work with C. Vacher

References I

John Aitchison and CH Ho.

The multivariate poisson-log normal distribution.
Biometrika, 76(4):643–653, 1989.

Boris Jakuschkin, Virgil Fievet, Loïc Schwaller, Thomas Fort, Cécile Robin, and Corinne Vacher.

Deciphering the pathobiome: Intra- and interkingdom interactions involving the pathogen erysiphe alphitoides.
Microbial Ecology, 72(4):870–880, Nov 2016.

Han Liu, Kathryn Roeder, and Larry Wasserman.

Stability approach to regularization selection (stars) for high dimensional graphical models.
In *Proceedings of the 23rd International Conference on Neural Information Processing Systems - Volume 2*, NIPS'10, pages 1432–1440, USA, 2010. Curran Associates Inc.

Núria Mach, Mustapha Berri, Jordi Estellé, Florence Levenez, Gaëtan Lemonnier, Catherine Denis, Jean-Jacques Leplat, Claire Chevaleyre, Yvon Billon, Joël Doré, and et al.

Early-life establishment of the swine gut microbiome and impact on host phenotypes.
Environmental Microbiology Reports, 7(3):554–569, May 2015.