A Walk Along Models for Count Data in Microbial Ecology

M. Mariadassou, INRAE-MaIAGE

joint work with Julien Chiquet and Stéphane Robin

Shandong University Summer School, Weihai, 2021, July 20-23

📄 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/

Outline

1. Tentative Syllabi
2. Motivation
3. Multinomial Models
4. Log-Normal Models
5. Applications

Syllabi

Multivariate models

L1: Overview of the concepts

L2: Review of multivariate models based on multinomial distributions

L3: Review of other multivariate count models

L4: Log-normal models: MLN and PLN

PLN models

L5: Estimation in the PLN model

L6: PLN-PCA for dimension reduction

L7: PLN-LDA and PLNmixture for classification and clustering

L8: PLNnetwork for network inference

Outline

1. Tentative Syllabi
2. Motivation
3. Multinomial Models
4. Log-Normal Models
5. Applications

Microbiome: Impact of Weaning on Piglets' Microbiome

Data from [MBE$^+$15].

- $n = 155$ samples ($= 31$ piglets at 5 times)
- $p = 1038$ bacterial species (OTUs) with prevalence ≥ 0.05
- Some covariates (sex, sire, etc)
- Offsets: $o_i =$ offset for sample i (sequencing depth)

Microbiome: Impact of Weaning on Piglets' Microbiome

Data from [MBE+15].

- $n = 155$ samples ($= 31$ piglets at 5 times)
- $p = 1038$ bacterial species (OTUs) with prevalence ≥ 0.05
- Some covariates (sex, sire, etc)
- Offsets: $o_i =$ offset for sample i (sequencing depth)

Aim: Study impact of weaning on gut microbiota

A look at the data

Metabarcoding data from [MBE^{+}15]

- count matrix with $n = 155$ piglets, $p = 1038$ species

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

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

- $d = 8$ covariates (sex, mother, weaning status, ...)

```
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
```

- Sampling effort in each sample

```
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: Oak powdery mildew

Data from [JFS+16].

- $n = 116$ oak leaves = samples
- $p = 114$ microbial species
  - $p_1 = 66$ bacterial species (OTUs, based on the 16S)
  - $p_2 = 48$ fungal species (OTUs, based on the ITS)
- covariates: tree (resistant, intermediate, susceptible), height, distance to trunk, ...
- offsets: $o_{i1} \neq o_{i2}$ = offset for bacteria, fungi

Pathobiome: Oak powdery mildew

Data from [JFS+16].

- $n = 116$ oak leaves = samples
- $p = 114$ microbial species
  - $p_1 = 66$ bacterial species (OTUs, based on the 16S)
  - $p_2 = 48$ fungal species (OTUs, based on the ITS)
- covariates: tree (resistant, intermediate, susceptible), height, distance to trunk, ...
- offsets: $o_{i1} \neq o_{i2}$ = offset for bacteria, fungi

```
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: Oak powdery mildew

Data from [JFS+16].

- $n = 116$ oak leaves = samples
- $p = 114$ microbial species
  - $p_1 = 66$ bacterial species (OTUs, based on the 16S)
  - $p_2 = 48$ fungal species (OTUs, based on the ITS)
- covariates: tree (resistant, intermediate, susceptible), height, distance to trunk, ...
- offsets: $o_{i1} \neq o_{i2}$ = offset for bacteria, fungi

```
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
```

Aim. Understand the interaction between the species, including the oak mildew pathogene *E. alphitoides*.

Problematic & Basic formalism

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

- $Y_{ij} =$ abundance (read counts) of species j in sample i
- $X_{ik} =$ value of covariate k in sample i
- $O_{ij} =$ offset (sampling effort) for species j in sample i

Problematic & Basic formalism

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

- Y_{ij} = abundance (read counts) of species j in sample i
- X_{ik} = value of covariate k in sample i
- O_{ij} = offset (sampling effort) for species j in sample i

Need for multivariate analysis to help deciphering the ecosystem

- exhibit patterns of diversity
  $\rightsquigarrow$ summarize the information from $\mathbf{Y}$ (PCA, clustering, . . .)
- understand between-species interactions
  $\rightsquigarrow$ 'Network' inference (variable/covariance selection)
- correct for technical and confounding effects
  $\rightsquigarrow$ account for covariables and sampling effort

Problematic & Basic formalism

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

- Y_{ij} = abundance (read counts) of species j in sample i
- X_{ik} = value of covariate k in sample i
- O_{ij} = offset (sampling effort) for species j in sample i

Need for multivariate analysis to help deciphering the ecosystem

- exhibit patterns of diversity
  $\rightsquigarrow$ summarize the information from $\mathbf{Y}$ (PCA, clustering, ...)
- understand between-species interactions
  $\rightsquigarrow$ 'Network' inference (variable/covariance selection)
- correct for technical and confounding effects
  $\rightsquigarrow$ account for covariables and sampling effort

$\rightsquigarrow$ need a generic framework to model dependencies between count variables

Microbial Ecology 101

1. Apply your favorite distance (Jaccard, Bray-Curtis, UniFrac, weighted UniFrac, etc)

Microbial Ecology 101

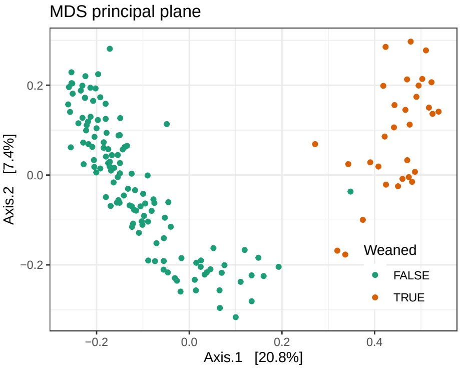
1. Apply your favorite distance (Jaccard, Bray-Curtis, UniFrac, weighted UniFrac, etc)
2. Apply your favorite dimension reduction technique (PCA, MDS/PCoA, NMDS, RDA, PLN, etc)

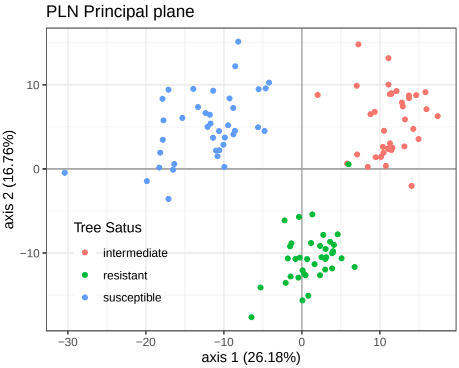
Microbial Ecology 101

1. Apply your favorite distance (Jaccard, Bray-Curtis, UniFrac, weighted UniFrac, etc)
2. Apply your favorite dimension reduction technique (PCA, MDS/PCoA, NMDS, RDA, PLN, etc)
3. Plot resulting *graph*

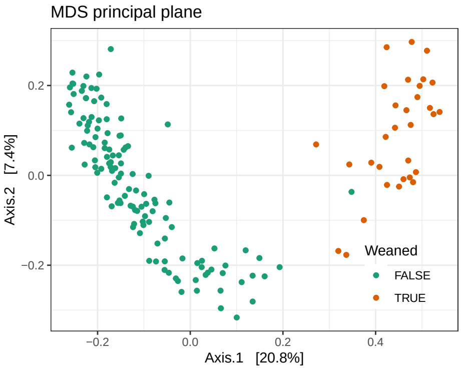
Microbial Ecology 101

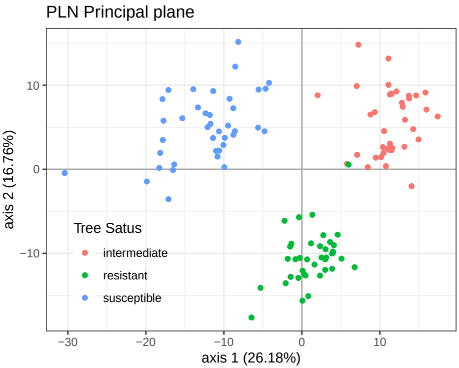
1. Apply your favorite distance (Jaccard, Bray-Curtis, UniFrac, weighted UniFrac, etc)
2. Apply your favorite dimension reduction technique (PCA, MDS/PCoA, NMDS, RDA, PLN, etc)
3. Plot resulting *graph*
4. *Et voilà!*


MDS principal plane


PLN Principal plane

Microbial Ecology 101





1. Perfect for finding structure...
2. But not for modeling it

What kind of generic models?

What kind of generic framework for multivariate count data?

What kind of generic models?

Cher Père Noël,

My Wish List to Santa

We want a family of generative models that are:

My Wish List to Santa

We want a family of generative models that are:

- Flexible enough to:
    - model average communities;
    - model dispersion (biological variability);
    - model interaction between OTUs (ecological networks);
    - accomodate heterogeneous communities;
    - integrate data from different sources (bacterial and fractions)

My Wish List to Santa

We want a family of generative models that are:

- Flexible enough to:
  - model average communities;
  - model dispersion (biological variability);
  - model interaction between OTUs (ecological networks);
  - accomodate heterogeneous communities;
  - integrate data from different sources (bacterial and fractions)
- yet as parcimonious as possible;

My Wish List to Santa

We want a family of generative models that are:

- Flexible enough to:
    - model average communities;
    - model dispersion (biological variability);
    - model interaction between OTUs (ecological networks);
    - accomodate heterogeneous communities;
    - integrate data from different sources (bacterial and fractions)
- yet as parcimonious as possible;
- interpretable;

My Wish List to Santa

We want a family of generative models that are:

- Flexible enough to:
  - model average communities;
  - model dispersion (biological variability);
  - model interaction between OTUs (ecological networks);
  - accomodate heterogeneous communities;
  - integrate data from different sources (bacterial and fractions)
- yet as parcimonious as possible;
- interpretable;
- fast and easy to fit to data;

My Wish List to Santa

We want a family of generative models that are:

- Flexible enough to:
  - model average communities;
  - model dispersion (biological variability);
  - model interaction between OTUs (ecological networks);
  - accomodate heterogeneous communities;
  - integrate data from different sources (bacterial and fractions)
- yet as parcimonious as possible;
- interpretable;
- fast and easy to fit to data;
- good fits to data (*e.g.* simulate realistic samples).

Outline

1. Tentative Syllabi
2. Motivation
3. Multinomial Models
   - Multinomial
   - Mixture of Multinomials
   - (Mixture of) Dirichlet-Multinomial
   - Latent Dirichlet Allocation
4. Log-Normal Models
5. Applications

Outline

1. Tentative Syllabi
2. Motivation
3. Multinomial Models
   - Multinomial
   - Mixture of Multinomials
   - (Mixture of) Dirichlet-Multinomial
   - Latent Dirichlet Allocation
4. Log-Normal Models
5. Applications

Multinomial Models

Intuition

- There are p species with proportions $\boldsymbol{\pi} = (\pi_1, \dots, \pi_p)$ in the species
- You pick N (sequencing depths) individuals with replacement

Multinomial Models

Intuition

- There are p species with proportions $\boldsymbol{\pi} = (\pi_1, \dots, \pi_p)$ in the species
- You pick N (sequencing depths) individuals with replacement

Mathematical Model

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

Multinomial Models

Intuition

- There are p species with proportions $\boldsymbol{\pi} = (\pi_1, \dots, \pi_p)$ in the species
- You pick N (sequencing depths) individuals with replacement

Mathematical Model

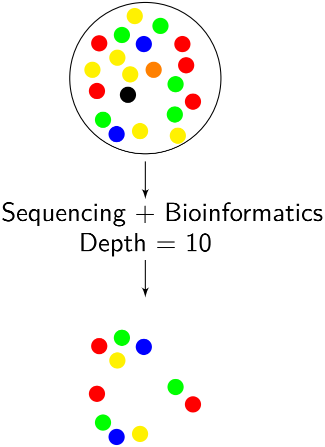
$$\mathbf{Y} \sim \mathcal{M}(N, \boldsymbol{\pi})$$

Inference is easy

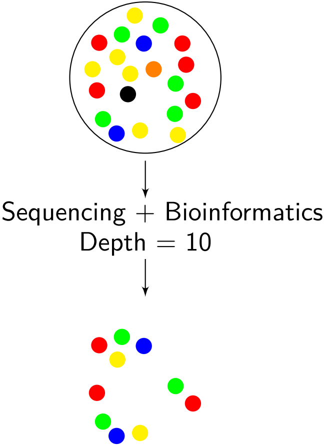
$$\hat{\pi}_j = \frac{\sum_{i=1}^n Y_{ij}}{\sum_{i=1}^n N_i}$$

with Y_{ij} the abundance of species j in sample i and N_i the depth of sample i.

Multinomial distribution: draw balls (with replacement) from a box



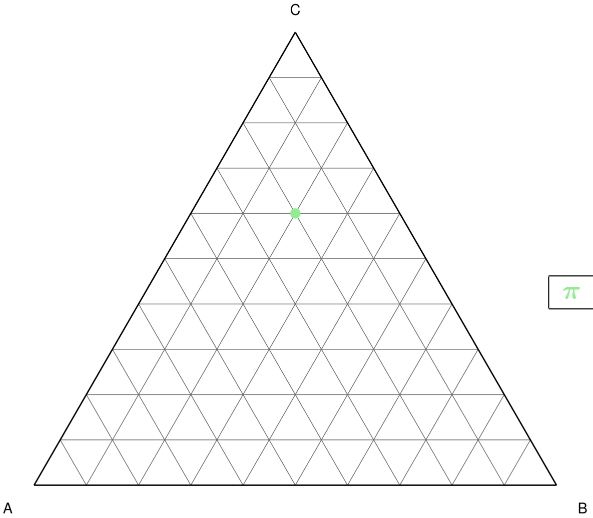
Multinomial distribution: draw balls (with replacement) from a box



Sequencing + Bioinformatics
Depth = 10

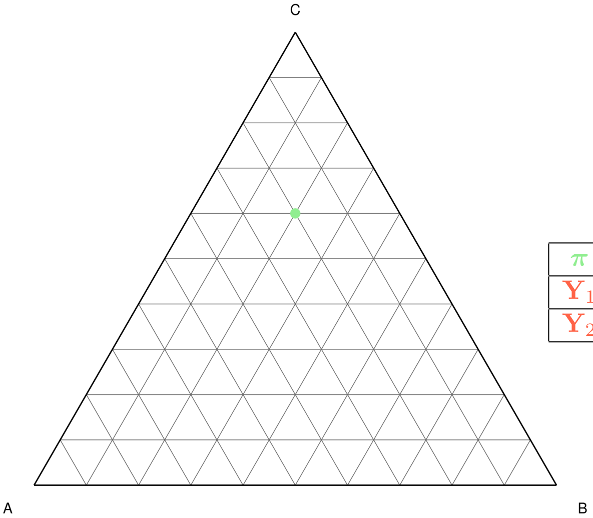
	🔴	🟡	🟢	⚫	🔵	🟠
Prop.	0.25	0.30	0.25	0.05	0.10	0.05
Counts	3	2	3	0	2	0
Obs. Prop.	0.3	0.2	0.3	0	0.2	0

Multinomial Model



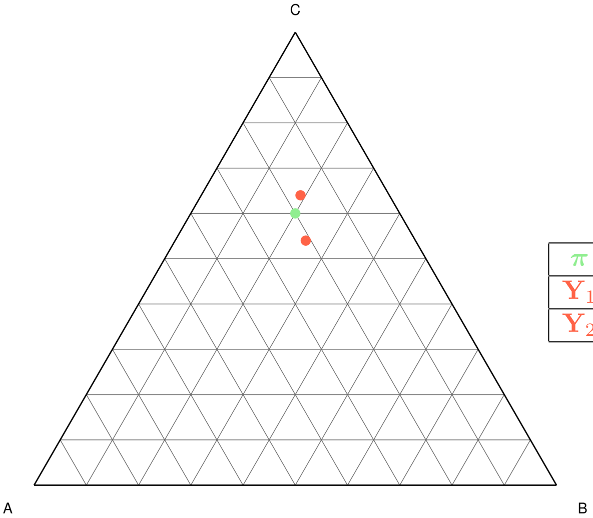
	A	B	C
π	0.2	0.2	0.6

Multinomial Model



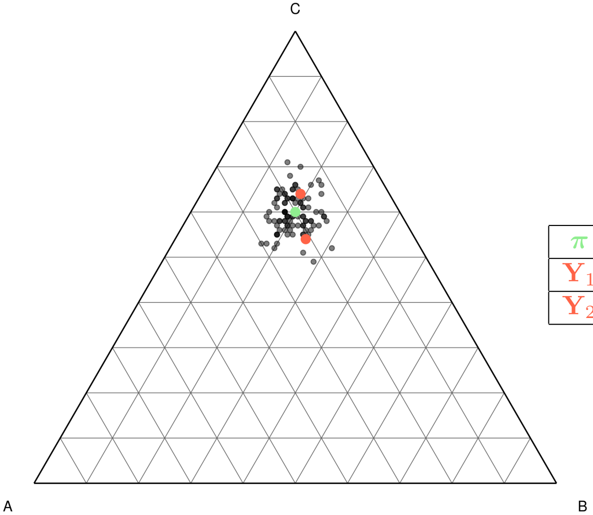
	A	B	C
π	0.2	0.2	0.6
$\mathbf{Y}_1$	17	19	64
$\mathbf{Y}_2$	21	25	54

Multinomial Model



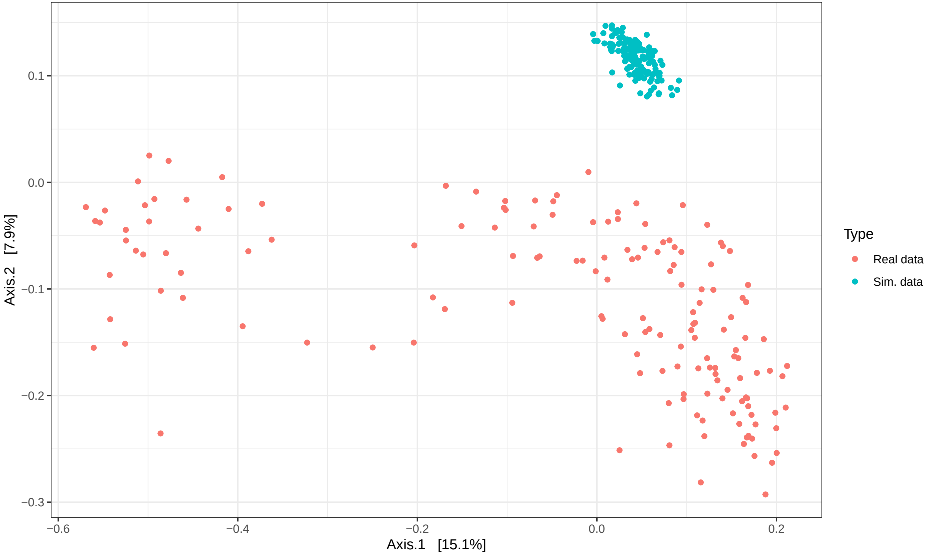
	A	B	C
π	0.2	0.2	0.6
$\mathbf{Y}_1$	17	19	64
$\mathbf{Y}_2$	21	25	54

Multinomial Model



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

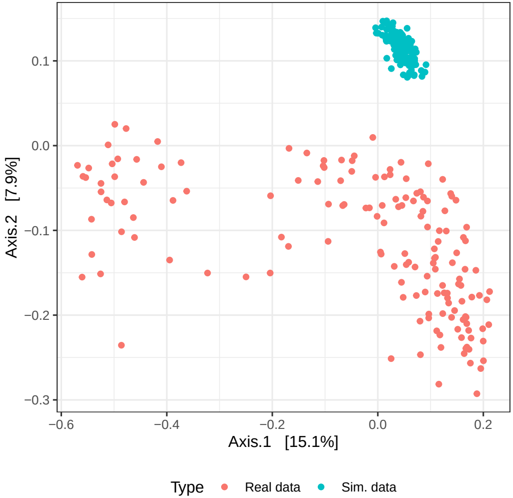
Example of Multinomial Model



Drawbacks

Heterogeneity

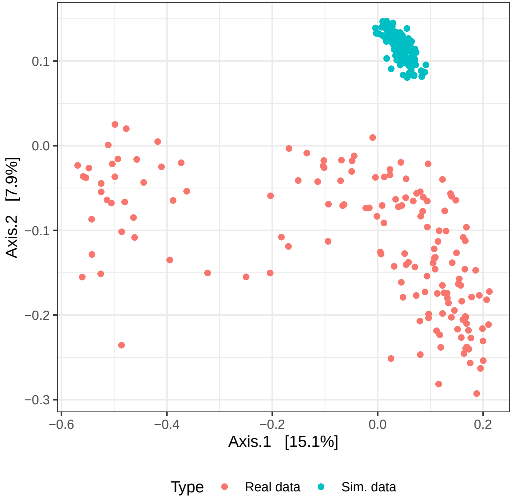
- Lack of heterogeneity



Drawbacks

Heterogeneity

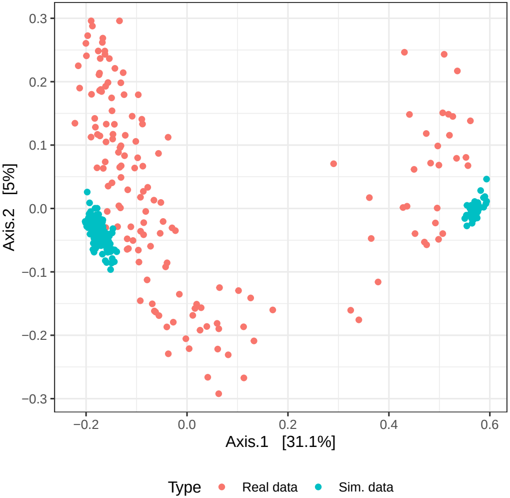
- Lack of heterogeneity
  ⇝ Fit only part of the data



Drawbacks

Heterogeneity

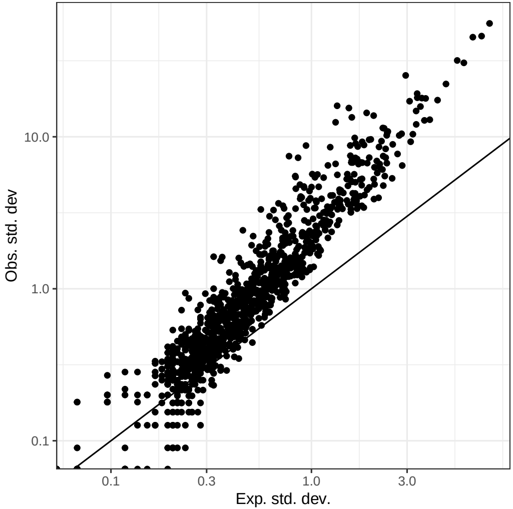
- Lack of heterogeneity
  ⇝ Fit only part of the data
- Lack of variance



Drawbacks

Heterogeneity

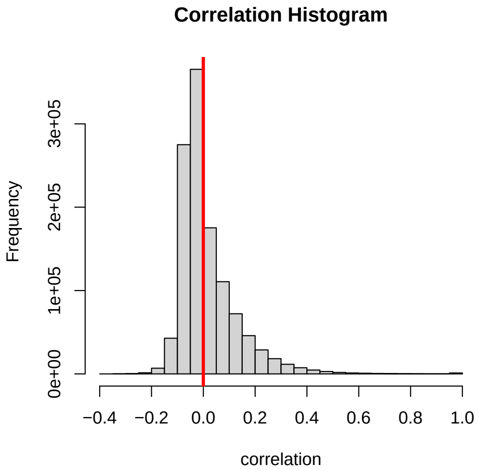
- Lack of heterogeneity
  ⇝ Fit only part of the data
- Lack of variance
- Small dispersion



Drawbacks

Heterogeneity

- Lack of heterogeneity
  ⇝ Fit only part of the data
- Lack of variance
- Small dispersion
- Wrong correlations



Pros and Cons

Pros

- + Parsimonious model: $p-1$ parameters to model p abundances
- + Easy to estimate
- + interpretable parameter

Pros and Cons

Pros

- + Parsimonious model: $p-1$ parameters to model p abundances
- + Easy to estimate
- + interpretable parameter

Cons

- - Bad for heterogeneity
- - Bad for dispersion around average composition ($\simeq$ biological variability)
- - Bad for correlations between OTUs

Outline

1. Tentative Syllabi

2. Motivation

3. Multinomial Models
   - Multinomial
   - Mixture of Multinomials
   - (Mixture of) Dirichlet-Multinomial
   - Latent Dirichlet Allocation

4. Log-Normal Models

5. Applications

©Manfred Heyde

Mixture Models

Intuition

- Each sample belongs to one of K groups
- Group k is characterized by its composition $\boldsymbol{\pi}_k$
- A sample from group k has composition $\boldsymbol{\pi}_k$
- Reads are sampled according to a multinomial process

Mixture Models

Intuition

- Each sample belongs to one of K groups
- Group k is characterized by its composition $\boldsymbol{\pi}_k$
- A sample from group k has composition $\boldsymbol{\pi}_k$
- Reads are sampled according to a multinomial process

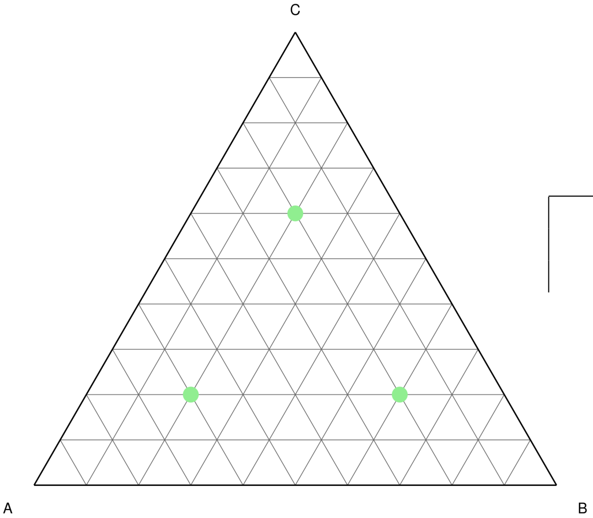
Hierarchical Model

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

where

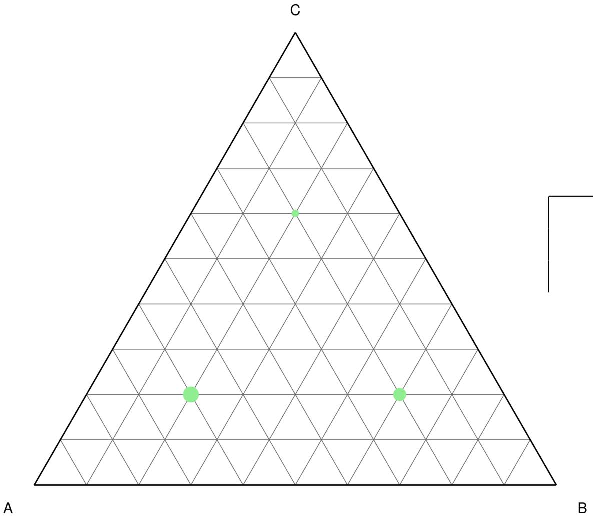
- $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)$ are the proportions of the K groups,

Mixture of Multinomial



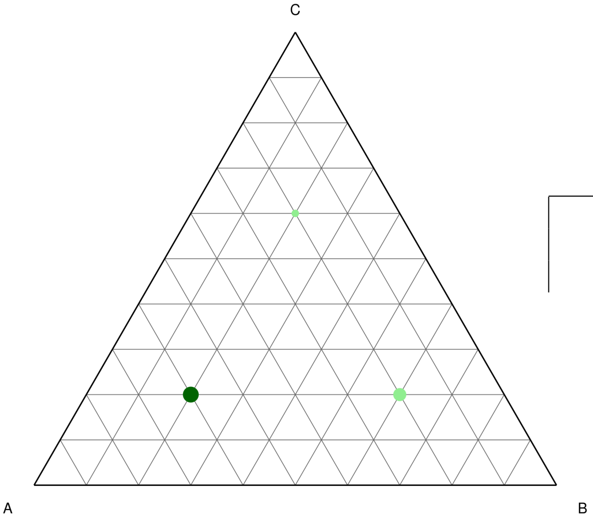
	A	B	C	α
π_1	0.6	0.2	0.2	
π_2	0.2	0.6	0.2	
π_3	0.2	0.2	0.6	

Mixture of Multinomial



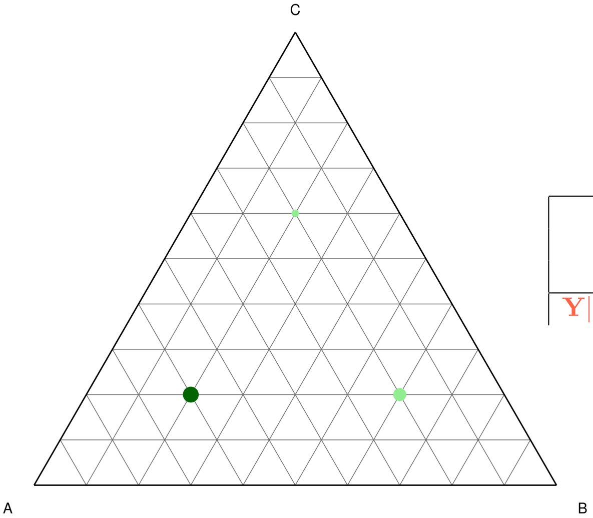
	A	B	C	α
π_1	0.6	0.2	0.2	0.5
π_2	0.2	0.6	0.2	0.4
π_3	0.2	0.2	0.6	0.1

Mixture of Multinomial



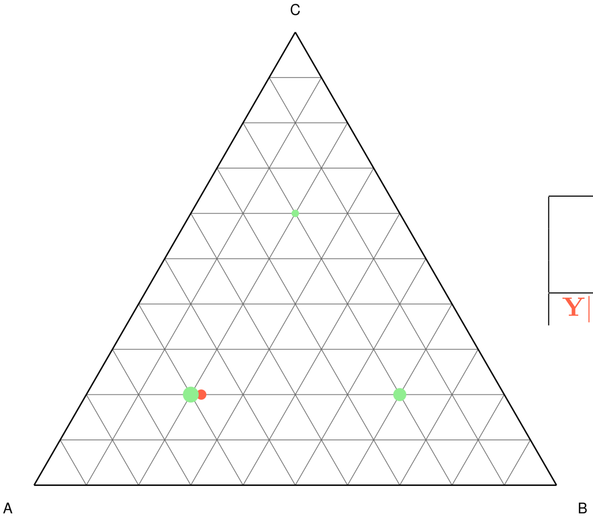
	A	B	C	α
π_1	0.6	0.2	0.2	0.5
π_2	0.2	0.6	0.2	0.4
π_3	0.2	0.2	0.6	0.1

Mixture of Multinomial



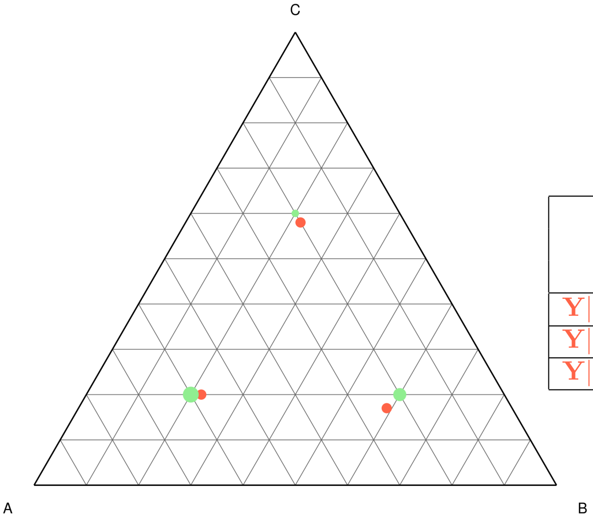
	A	B	C	α
$\boldsymbol{\pi}_1$	0.6	0.2	0.2	0.5
π_2	0.2	0.6	0.2	0.4
π_3	0.2	0.2	0.6	0.1
$\mathbf{Y} \vert Z = 1$	58	22	20	

Mixture of Multinomial



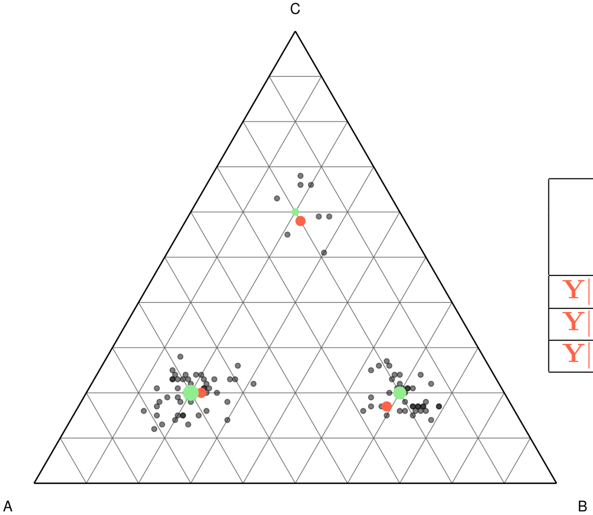
	A	B	C	α
π_1	0.6	0.2	0.2	0.5
π_2	0.2	0.6	0.2	0.4
π_3	0.2	0.2	0.6	0.1
$\mathbf{Y} \mid Z = 1$	58	22	20	

Mixture of Multinomial



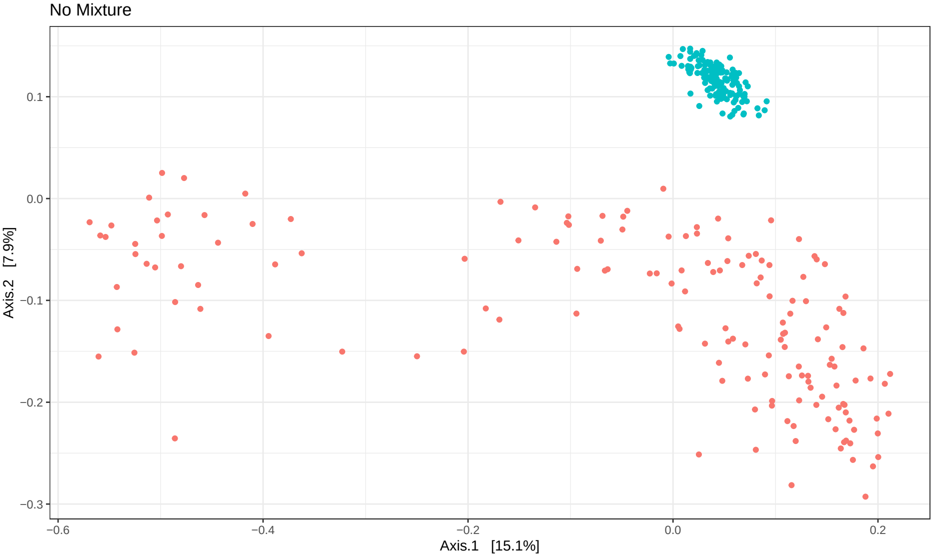
	A	B	C	α
π_1	0.6	0.2	0.2	0.5
π_2	0.2	0.6	0.2	0.4
π_3	0.2	0.2	0.6	0.1
$\mathbf{Y}\mid Z=1$	58	22	20	
$\mathbf{Y}\mid Z=2$	24	59	17	
$\mathbf{Y}\mid Z=3$	20	22	58	

Mixture of Multinomial

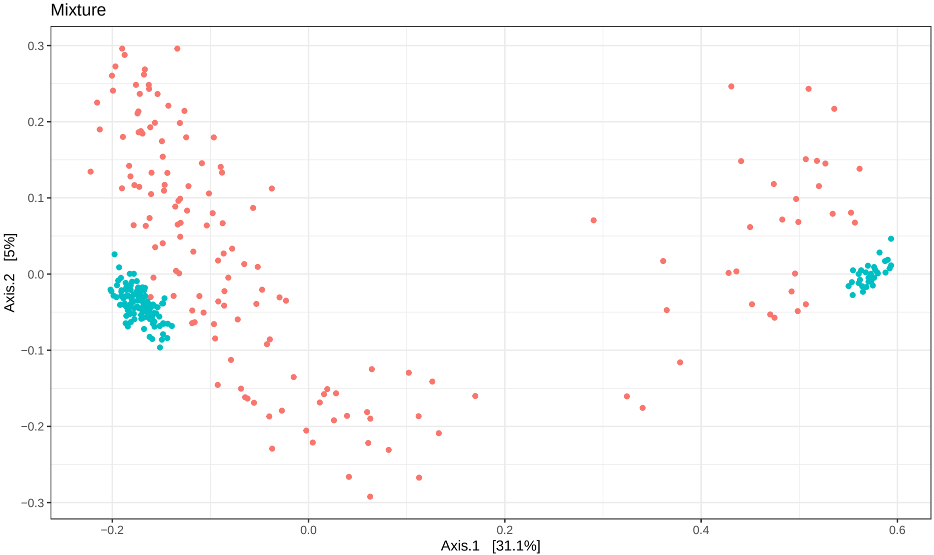


	A	B	C	α
π_1	0.6	0.2	0.2	0.5
π_2	0.2	0.6	0.2	0.4
π_3	0.2	0.2	0.6	0.1
$\mathbf{Y}\|Z=1$	58	22	20	
$\mathbf{Y}\|Z=2$	24	59	17	
$\mathbf{Y}\|Z=3$	20	22	58	

Example of Mixture Models



Example of Mixture Models



Pros and Cons

Pros

- + Good for heterogeneity
- + Parcimonious: $Kp - 1$ parameters for K groups
- + Inference is easy when groups are known $\rightsquigarrow$ simple averages

Pros and Cons

Pros

- + Good for heterogeneity
- + Parcimonious: $Kp - 1$ parameters for K groups
- + Inference is easy when groups are known $\rightsquigarrow$ simple averages

Cons

- - Inference is more involved when groups are unknown
  $\rightsquigarrow$ iterative EM algorithm
- - Bad for dispersion
- - Bad for correlations between OTUs

Outline

1. Tentative Syllabi

2. Motivation

3. Multinomial Models
   - Multinomial
   - Mixture of Multinomials
   - (Mixture of) Dirichlet-Multinomial
   - Latent Dirichlet Allocation

4. Log-Normal Models

5. Applications

Dirichlet - Multinomial

Intuition

- $\boldsymbol{\pi}$ is the ecosystem-level average composition

Dirichlet - Multinomial

Intuition

- $\boldsymbol{\pi}$ is the ecosystem-level average composition
- Sample i has **own** composition $\boldsymbol{\pi}_i$ (noisy version of $\boldsymbol{\pi}$) ⇝ Biological variability

Dirichlet - Multinomial

Intuition

- $\boldsymbol{\pi}$ is the ecosystem-level average composition
- Sample i has **own** composition $\boldsymbol{\pi}_i$ (noisy version of $\boldsymbol{\pi}$) $\rightsquigarrow$ Biological variability
- N_i reads are sampled from $\boldsymbol{\pi}_i$ according to a multinomial $\rightsquigarrow$ Technical / Sampling variability

Dirichlet - Multinomial

Intuition

- $\boldsymbol{\pi}$ is the ecosystem-level average composition
- Sample i has **own** composition $\boldsymbol{\pi}_i$ (noisy version of $\boldsymbol{\pi}$) ⇝ Biological variability
- N_i reads are sampled from $\boldsymbol{\pi}_i$ according to a multinomial ⇝ Technical / Sampling variability

Hierarchical Model

$$\begin{aligned} &\boldsymbol{\pi} && \text{Ecosystem average composition} \\ \boldsymbol{\pi}_i &\sim \mathcal{D}(\kappa\boldsymbol{\pi}) && \text{Sample average composition} \\ \mathbf{Y}_i &\sim \mathcal{M}(N_i, \boldsymbol{\pi}_i) && \text{Observed counts} \end{aligned}$$

where $1/\kappa$ models the level of variability (large $1/\kappa$ ⇝ large variablity)

Dirichlet - Multinomial

Intuition

- $\boldsymbol{\pi}$ is the ecosystem-level average composition
- Sample i has **own** composition $\boldsymbol{\pi}_i$ (noisy version of $\boldsymbol{\pi}$) $\rightsquigarrow$ Biological variability
- N_i reads are sampled from $\boldsymbol{\pi}_i$ according to a multinomial $\rightsquigarrow$ Technical / Sampling variability

Hierarchical Model

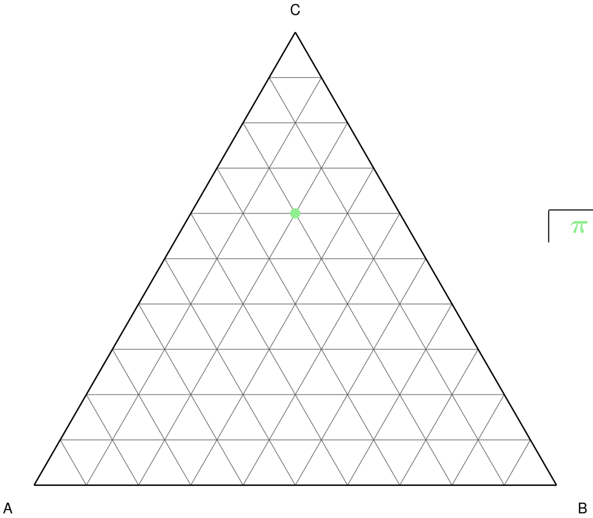
$$
\begin{aligned}
& \boldsymbol{\pi} && \text{Ecosystem average composition} \\
\boldsymbol{\pi}_i &\sim \mathcal{D}(\kappa \boldsymbol{\pi}) && \text{Sample average composition} \\
\mathbf{Y}_i &\sim \mathcal{M}(N_i, \boldsymbol{\pi}_i) && \text{Observed counts}
\end{aligned}
$$

where $1/\kappa$ models the level of variability (large $1/\kappa \rightsquigarrow$ large variablity)

Mixture Layer

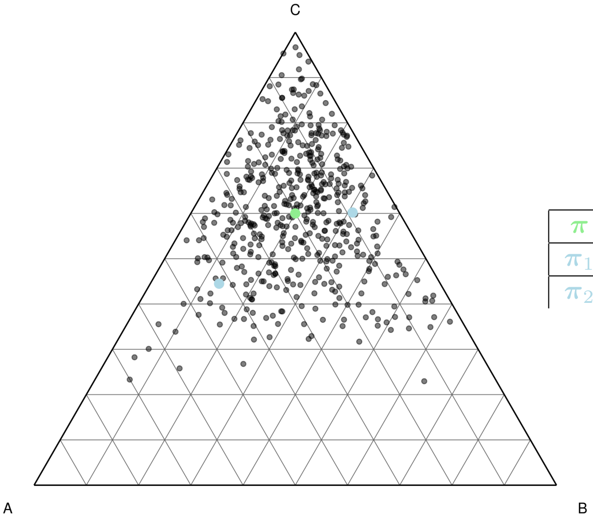
Can be combined with a mixture model

Dirichlet-Multinomial distribution



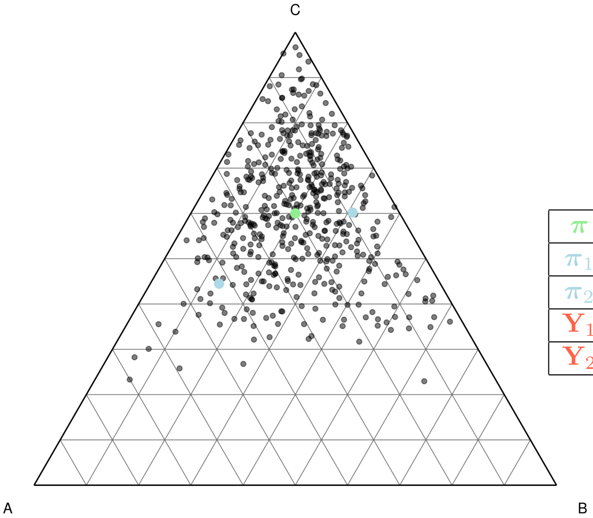
	A	B	C
π	0.2	0.2	0.6

Dirichlet-Multinomial distribution



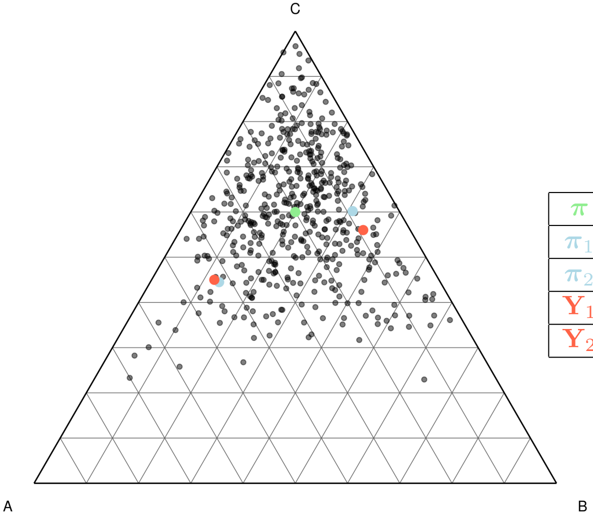
	A	B	C
π	0.2	0.2	0.6
π_1	0.089	0.309	0.602
π_2	0.423	0.132	0.445

Dirichlet-Multinomial distribution



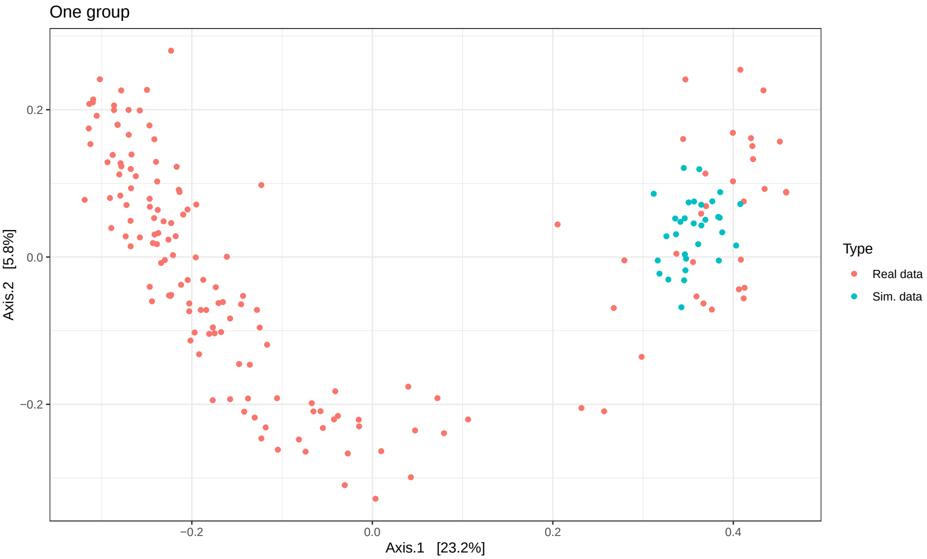
	A	B	C
π	0.2	0.2	0.6
π_1	0.089	0.309	0.602
π_2	0.423	0.132	0.445
$\mathbf{Y}_1$	9	35	56
$\mathbf{Y}_2$	43	12	45

Dirichlet-Multinomial distribution

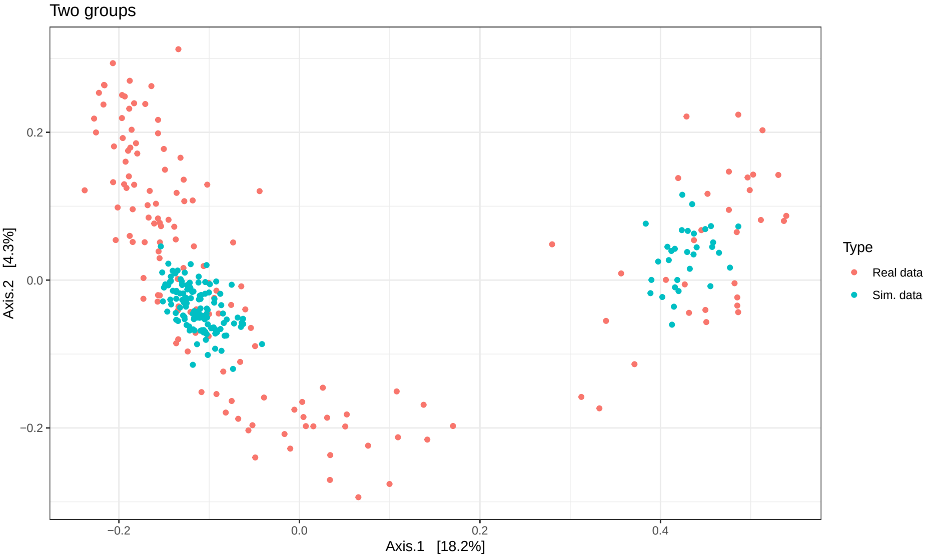


	A	B	C
π	0.2	0.2	0.6
π_1	0.089	0.309	0.602
π_2	0.423	0.132	0.445
$\mathbf{Y}_1$	9	35	56
$\mathbf{Y}_2$	43	12	45

Example of Dirichlet-Multinomial



Example of Dirichlet-Multinomial (Cont'd)



Pros and Cons

Pros

- Good for heterogeneity
- So-so of OK for dispersion
- Parcimonious: $K(p+1)-1$ parameters for K groups

Pros and Cons

Pros

- Good for heterogeneity
- So-so of OK for dispersion
- Parcimonious: $K(p+1)-1$ parameters for K groups

Cons

- Inference is more involved
  Known groups $\rightsquigarrow$ gradient descent
  Unknown groups $\rightsquigarrow$ Iterative EM algorithm + gradient descent
- Bad for correlations between OTUs

Outline

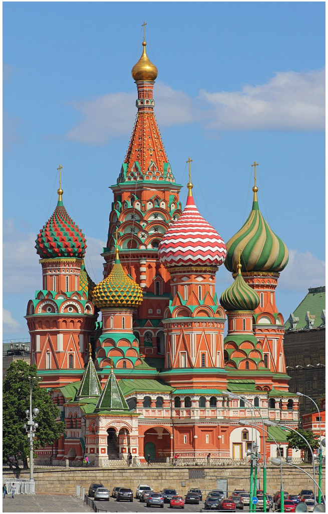
1. Tentative Syllabi

2. Motivation

3. Multinomial Models
    - Multinomial
    - Mixture of Multinomials
    - (Mixture of) Dirichlet-Multinomial
    - Latent Dirichlet Allocation

4. Log-Normal Models

5. Applications



©A. Savin

Latent Dirichlet Allocation

Intuition

- There are K archetype ecosystems $1, \dots, K$

Latent Dirichlet Allocation

Intuition

- There are K archetype ecosystems $1, \dots, K$
- Each archetype has its own composition $\boldsymbol{\pi}_k$

Latent Dirichlet Allocation

Intuition

- There are K archetype ecosystems $1, \dots, K$
- Each archetype has its own composition $\boldsymbol{\pi}_k$
- Each sample $\mathbf{Y}$ is made-up of several archetypes in proportions $(\theta_1, \dots, \theta_K)$

Latent Dirichlet Allocation

Intuition

- There are K archetype ecosystems $1, \dots, K$
- Each archetype has its own composition $\boldsymbol{\pi}_k$
- Each sample $\mathbf{Y}$ is made-up of several archetypes in proportions $(\theta_1, \dots, \theta_K)$
- $\theta_k N$ reads are sampled from a noisy version of $\boldsymbol{\pi}_k$

Latent Dirichlet Allocation

Intuition

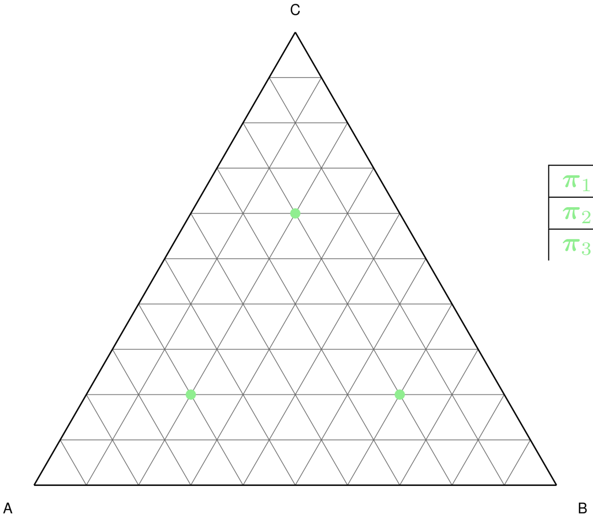
- There are K archetype ecosystems $1, \dots, K$
- Each archetype has its own composition $\boldsymbol{\pi}_k$
- Each sample $\mathbf{Y}$ is made-up of several archetypes in proportions $(\theta_1, \dots, \theta_K)$
- $\theta_k N$ reads are sampled from a noisy version of $\boldsymbol{\pi}_k$

Hierarchical Model

$\boldsymbol{\pi}_1, \dots, \boldsymbol{\pi}_K$	Archetypes average compositions
$\boldsymbol{\theta} \sim \mathcal{D}(\kappa \boldsymbol{\alpha})$	Proportion of archetypes in sample
$\tilde{\boldsymbol{\pi}}_k \sim \mathcal{D}(\kappa_k \boldsymbol{\pi}_k)$	Noisy version of $\boldsymbol{\pi}_k$
$z_i \sim \mathcal{M}(1, \boldsymbol{\theta})$	Archetype of origin of read i
$w_i \mid z_i = k \sim \mathcal{M}(1, \tilde{\boldsymbol{\pi}}_k)$	OTU of read i

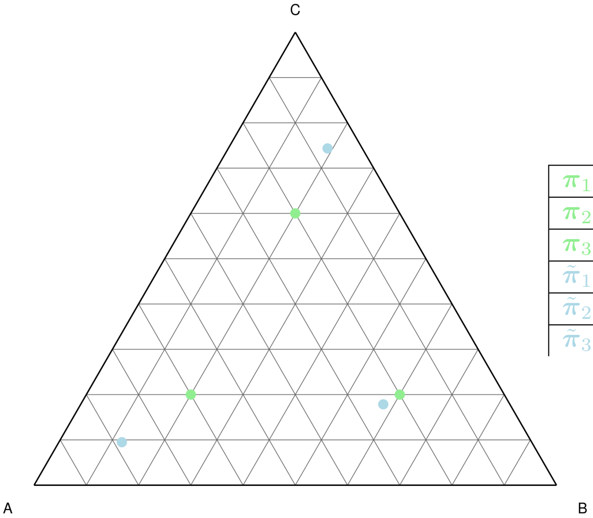
where κ and the κ_k control noise levels.

Latent Dirichlet Allocation



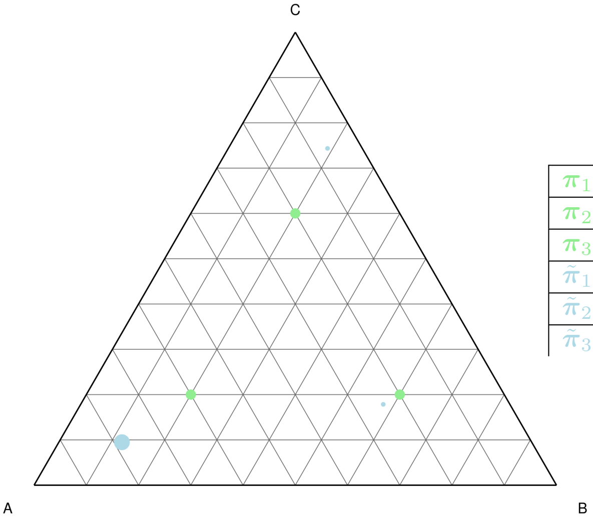
	A	B	C	θ
π_1	0.6	0.2	0.2	
π_2	0.2	0.6	0.2	
π_3	0.2	0.2	0.6	

Latent Dirichlet Allocation



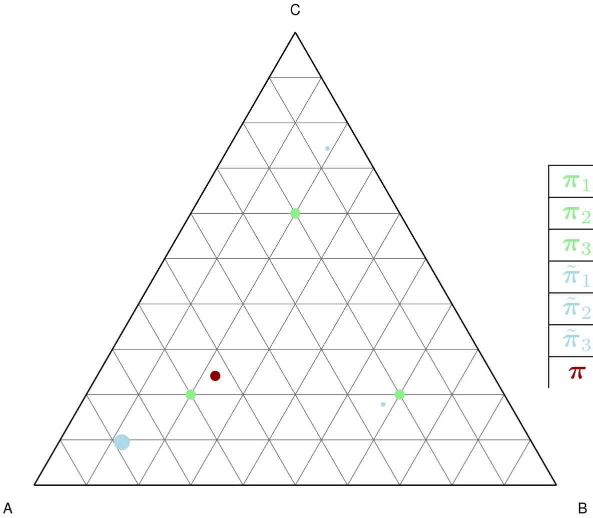
	A	B	C	θ
π_1	0.6	0.2	0.2	
π_2	0.2	0.6	0.2	
π_3	0.2	0.2	0.6	
$\tilde{\pi}_1$	0.784	0.121	0.095	
$\tilde{\pi}_2$	0.242	0.579	0.179	
$\tilde{\pi}_3$	0.423	0.132	0.445	

Latent Dirichlet Allocation



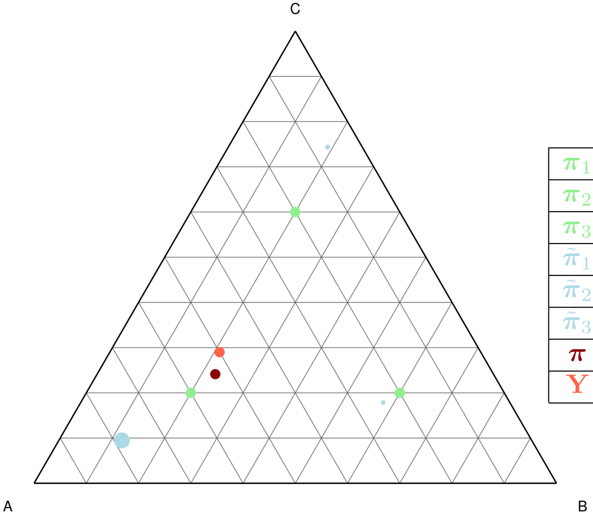
	A	B	C	θ
π_1	0.6	0.2	0.2	
π_2	0.2	0.6	0.2	
π_3	0.2	0.2	0.6	
$\tilde{\pi}_1$	0.784	0.121	0.095	0.6
$\tilde{\pi}_2$	0.242	0.579	0.179	0.2
$\tilde{\pi}_3$	0.423	0.132	0.445	0.2

Latent Dirichlet Allocation



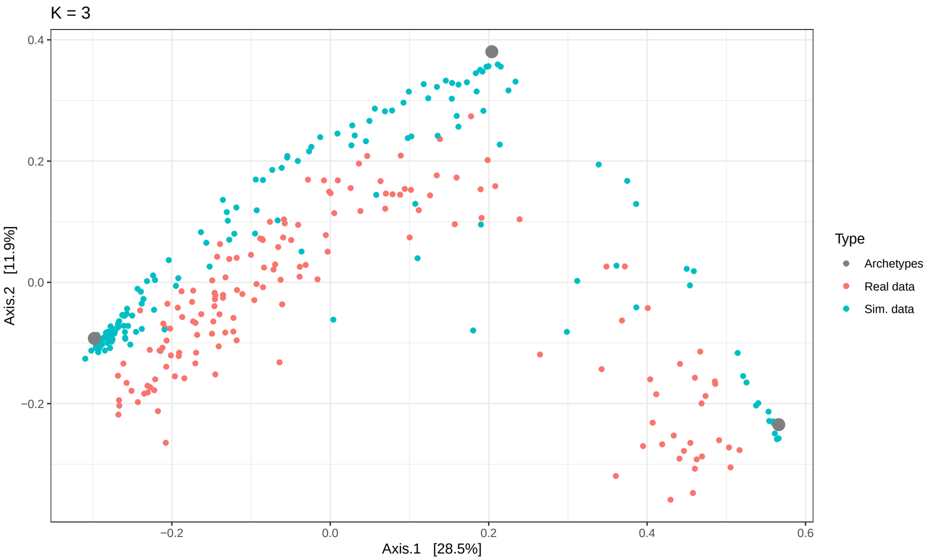
	A	B	C	θ
π_1	0.6	0.2	0.2	
π_2	0.2	0.6	0.2	
π_3	0.2	0.2	0.6	
$\tilde{\pi}_1$	0.784	0.121	0.095	0.6
$\tilde{\pi}_2$	0.242	0.579	0.179	0.2
$\tilde{\pi}_3$	0.423	0.132	0.445	0.2
π	0.532	0.226	0.241	

Latent Dirichlet Allocation

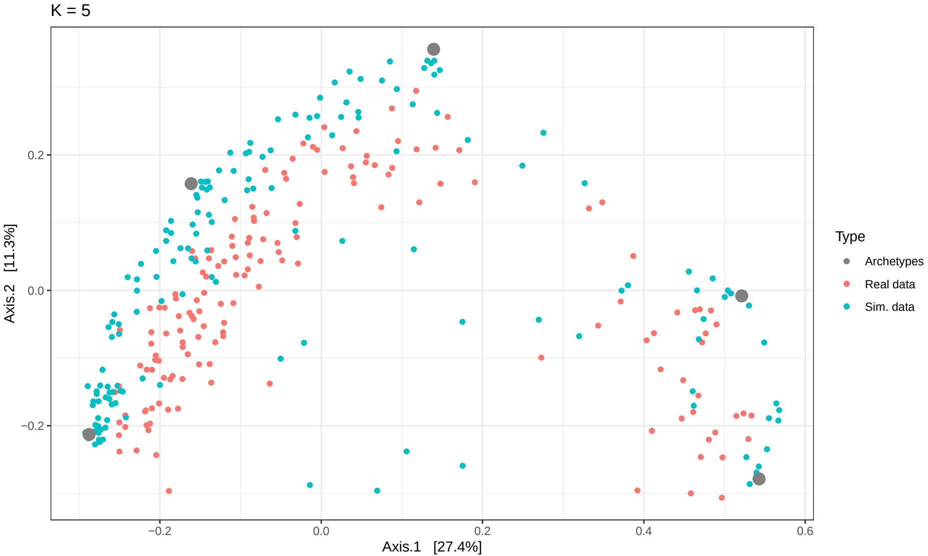


	A	B	C	θ
π_1	0.6	0.2	0.2	
π_2	0.2	0.6	0.2	
π_3	0.2	0.2	0.6	
$\tilde{\pi}_1$	0.784	0.121	0.095	0.6
$\tilde{\pi}_2$	0.242	0.579	0.179	0.2
$\tilde{\pi}_3$	0.423	0.132	0.445	0.2
π	0.532	0.226	0.241	
$\mathbf{Y}$	54	18	28	

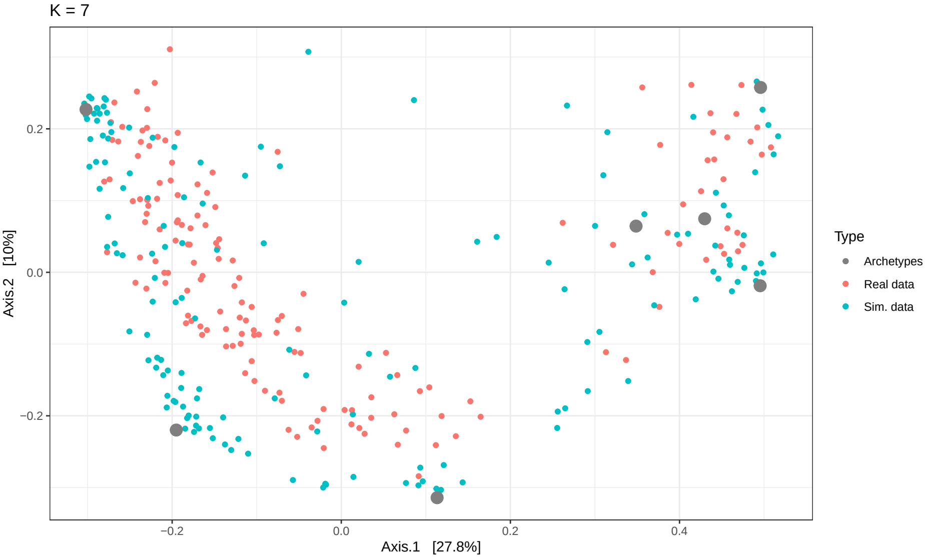
Example of Latent Dirichlet Allocation



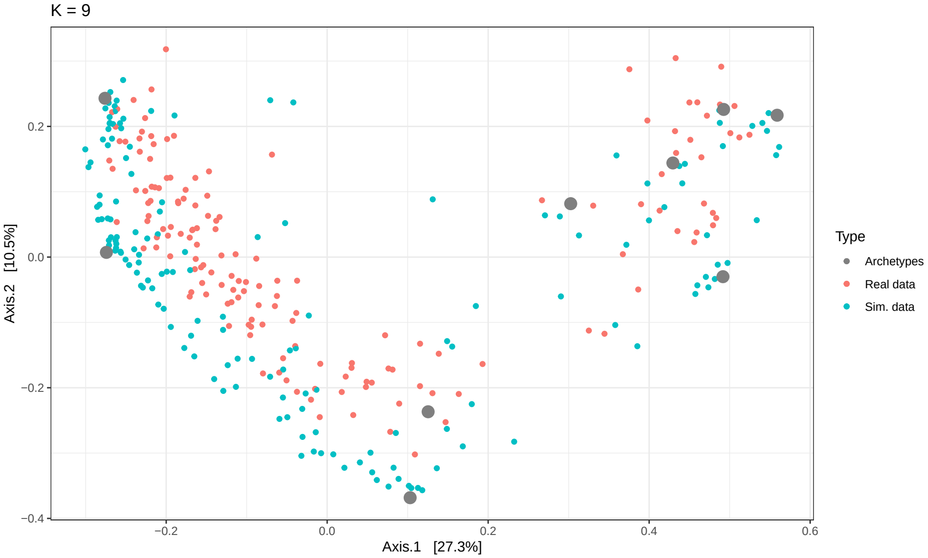
Example of Latent Dirichlet Allocation



Example of Latent Dirichlet Allocation



Example of Latent Dirichlet Allocation



Pros and Cons

Pros

- Good for heterogeneity
- Good for dispersion
- Parcimonious: $K(p+1)$ parameters for K archetypes

Pros and Cons

Pros

- + Good for heterogeneity
- + Good for dispersion
- + Parcimonious: $K(p+1)$ parameters for K archetypes

Cons

- - Inference is very involved
  $\rightsquigarrow$ gradient descent + EM algorithm / Gibbs sampling
- - Interpretation is complex $\rightsquigarrow$ archetypes are not groups
- - Bad for correlations between OTUs

Partial Summary

Multinomial-based models are good at

- modeling compositions;
- modeling dispersion around average compositions;
- modeling heterogeneity;
- using (relatively) few parameters

Partial Summary

Multinomial-based models are good at

- modeling compositions;
- modeling dispersion around average compositions;
- modeling heterogeneity;
- using (relatively) few parameters

Multinomial models are bad at

- modeling interactions between covariates;
- accounting for covariates;
- Integrating datasets from different sources (*e.g.* 16S, ITS)

Outline

1. Tentative Syllabi

2. Motivation

3. Multinomial Models

4. Log-Normal Models
   - Multinomial Log-Normal
   - Poisson Log-Normal

5. Applications

Modeling Correlations

Multivariate Gaussian models are the *de facto* distribution to model correlations.

Modeling Correlations

Multivariate Gaussian models are the *de facto* distribution to model correlations.

For continuous variables

- The p variables $\mathbf{Y}_i$ (*e.g.* species abundances) are explained
- by the values of the d covariates $\mathbf{X}_i$ and the p offsets $\mathbf{O}_i$

$$\mathbf{Y}_i = \underbrace{\mathbf{X}_i\mathbf{B}}_{\text{accounts for covariates}} + \underbrace{\mathbf{O}_i}_{\text{accounts for sampling effort}} + \varepsilon_i, \ \varepsilon_i \sim \mathcal{N}(\mathbf{0}_p, \underbrace{\mathbf{\Sigma}}_{\text{dependencies between species}})$$

+ null covariance ⇔ independence ⇝ uncorrelated species do not interact

Modeling Correlations

Multivariate Gaussian models are the *de facto* distribution to model correlations.

For continuous variables

- The p variables $\mathbf{Y}_i$ (*e.g.* species abundances) are explained
- by the values of the d covariates $\mathbf{X}_i$ and the p offsets $\mathbf{O}_i$

$$\mathbf{Y}_i = \underbrace{\mathbf{X}_i\mathbf{B}}_{\substack{\text{accounts for} \\ \text{covariates}}} + \underbrace{\mathbf{O}_i}_{\substack{\text{accounts for} \\ \text{sampling effort}}} + \boldsymbol{\varepsilon}_i, \ \boldsymbol{\varepsilon}_i \sim \mathcal{N}(\mathbf{0}_p, \underbrace{\boldsymbol{\Sigma}}_{\substack{\text{dependencies} \\ \text{between species}}})$$

+ ~~null covariance ⇔ independence ⇝ uncorrelated species do not interact~~

But abundances are not gaussian...

Modeling Correlations

Multivariate Gaussian models are the *de facto* distribution to model correlations.

For continuous variables

- The p variables $\mathbf{Y}_i$ (*e.g.* species abundances) are explained
- by the values of the d covariates $\mathbf{X}_i$ and the p offsets $\mathbf{O}_i$

$$\mathbf{Y}_i = \underbrace{\mathbf{X}_i\mathbf{B}}_{\substack{\text{accounts for}\\ \text{covariates}}} + \underbrace{\mathbf{O}_i}_{\substack{\text{accounts for}\\ \text{sampling effort}}} + \boldsymbol{\varepsilon}_i, \ \boldsymbol{\varepsilon}_i \sim \mathcal{N}(\mathbf{0}_p, \underbrace{\boldsymbol{\Sigma}}_{\substack{\text{dependencies}\\ \text{between species}}})$$

+ ~~null covariance ⇔ independence ⇝ uncorrelated species do not interact~~

But abundances are not gaussian...

Use a latent variable models with a gaussian latent layer and a count observed layer

Outline

1. Tentative Syllabi

2. Motivation

3. Multinomial Models

4. Log-Normal Models
   - Multinomial Log-Normal
   - Poisson Log-Normal

5. Applications

©cbrettre

Multinomial Log-Normal

Intuition

- The latent layer models so-called basis abundances $\mathbf{z}$

Multinomial Log-Normal

Intuition

- The latent layer models so-called basis abundances $\mathbf{z}$
- Basis are transformed to an average composition $\boldsymbol{\pi}$

Multinomial Log-Normal

Intuition

- The latent layer models so-called basis abundances $\mathbf{z}$
- Basis are transformed to an average composition $\boldsymbol{\pi}$
- N reads are sampled from $\boldsymbol{\pi}$ according to a multinomial distribution

Multinomial Log-Normal

Intuition

- The latent layer models so-called basis abundances $\mathbf{z}$
- Basis are transformed to an average composition $\boldsymbol{\pi}$
- N reads are sampled from $\boldsymbol{\pi}$ according to a multinomial distribution

Hierarchical Model

$\mathbf{z} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$	Abundance basis
$\boldsymbol{\pi} \mid \mathbf{z} = \left(\frac{e^{z_j}}{\sum_{j'} e^{z_{j'}}} \right)_j$	Average composition
$\mathbf{Y} \sim \mathcal{M}(N, \boldsymbol{\pi})$	Observed composition

Multinomial Log-Normal

Intuition

- The latent layer models so-called basis abundances $\mathbf{z}$
- Basis are transformed to an average composition $\boldsymbol{\pi}$
- N reads are sampled from $\boldsymbol{\pi}$ according to a multinomial distribution

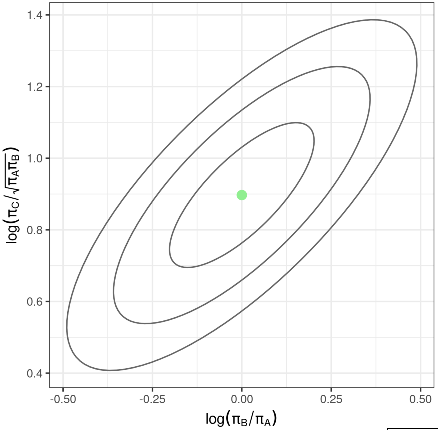
Hierarchical Model

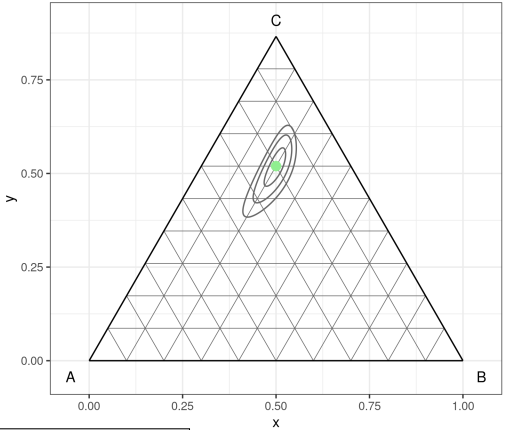
$$\begin{aligned}
\mathbf{z} &\sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) && \text{Abundance basis} \\
\boldsymbol{\pi} | \mathbf{z} &= \left(\frac{e^{z_j}}{\sum_{j'} e^{z_{j'}}} \right)_j && \text{Average composition} \\
\mathbf{Y} &\sim \mathcal{M}(N, \boldsymbol{\pi}) && \text{Observed composition}
\end{aligned}$$

Mixture Layer

Can be combined with a mixture model

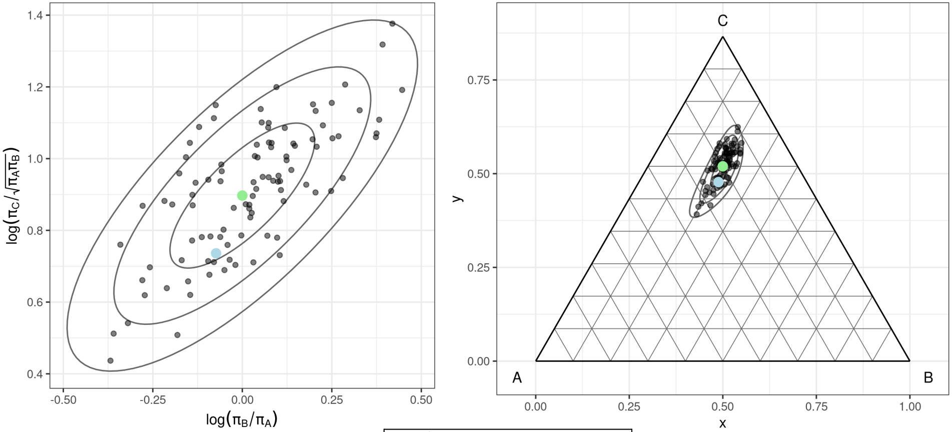
Multinomial Log-Normal





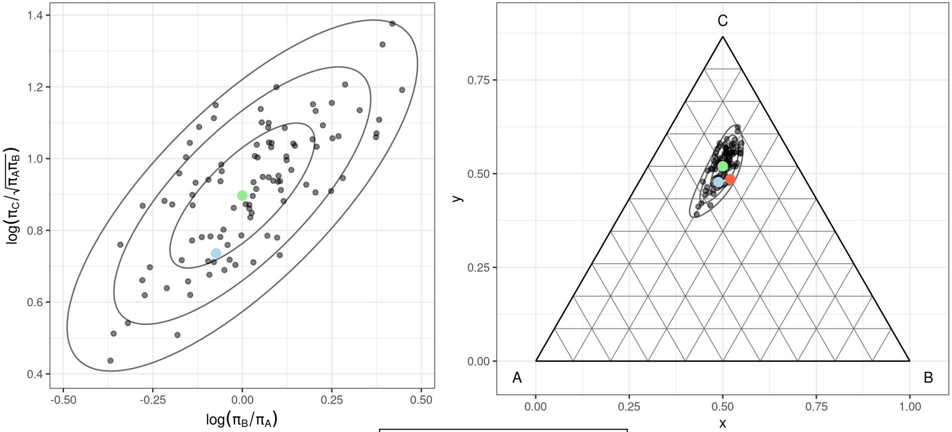
	A	B	C
π	0.2	0.2	0.6

Multinomial Log-Normal



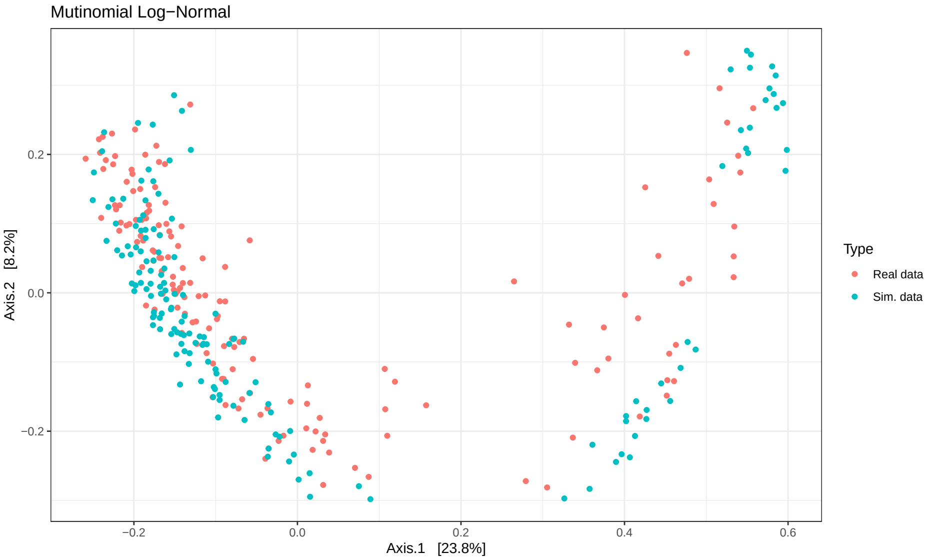
	A	B	C
π	0.2	0.2	0.6
π_1	0.235	0.213	0.552

Multinomial Log-Normal



	A	B	C
π	0.2	0.2	0.6
π_1	0.235	0.213	0.552
$\mathbf{Y}$	20	24	56

Example of Multinomial Log-Normal



Pros and Cons

Pros

- Good for heterogeneity
- Good for dispersion
- Good for correlations between OTUs

Pros and Cons

Pros

+ Good for heterogeneity
+ Good for dispersion
+ Good for correlations between OTUs

Cons

- The model is not parsimonious: $p(p+3)/2$ parameters
- Inference is involved
  $\rightsquigarrow$ iterative EM algorithm
- Modeling is done at the proportion level

Outline

1. Tentative Syllabi

2. Motivation

3. Multinomial Models

4. Log-Normal Models
   - Multinomial Log-Normal
   - Poisson Log-Normal

5. Applications

Poisson-log normal (PLN) distribution [AH89]

Intuition

- The latent layer models basis $\mathbf{z}$

Poisson-log normal (PLN) distribution [AH89]

Intuition

- The latent layer models basis $\mathbf{z}$
- Basis are *transformed* to average counts

Poisson-log normal (PLN) distribution [AH89]

Intuition

- The latent layer models basis $\mathbf{z}$
- Basis are *transformed* to average counts
- Reads are *sampled* according to Poisson distribution

Poisson-log normal (PLN) distribution [AH89]

Intuition

- The latent layer models basis $\mathbf{z}$
- Basis are *transformed* to average counts
- Reads are *sampled* according to Poisson distribution

Hierarchical Model

$$\begin{aligned} \mathbf{z} &\sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) && \text{Basis} \\ \lambda_j | \mathbf{z} &= e^{z_j} && \text{Average count of species } j \\ \mathbf{Y}_j | \mathbf{z} &\sim \mathcal{P}(e^{z_j}) && \text{Observed count of species } j \end{aligned}$$

Poisson-log normal (PLN) distribution [AH89]

Intuition

- The latent layer models basis $\mathbf{z}$
- Basis are *transformed* to average counts
- Reads are *sampled* according to Poisson distribution

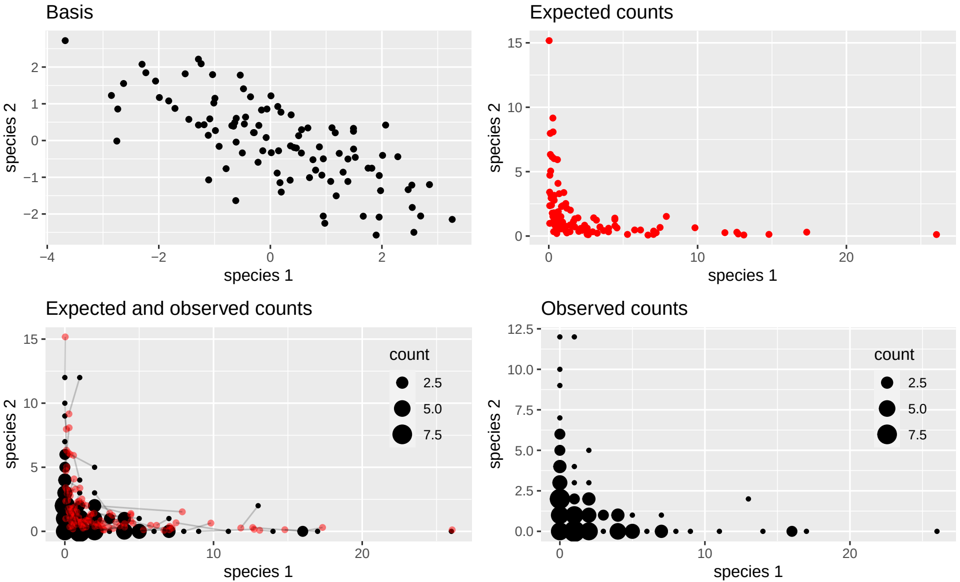
Hierarchical Model

$$\begin{aligned} \mathbf{z} &\sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) && \text{Basis} \\ \lambda_j | \mathbf{z} &= e^{z_j} && \text{Average count of species } j \\ \mathbf{Y}_j | \mathbf{z} &\sim \mathcal{P}(e^{z_j}) && \text{Observed count of species } j \end{aligned}$$

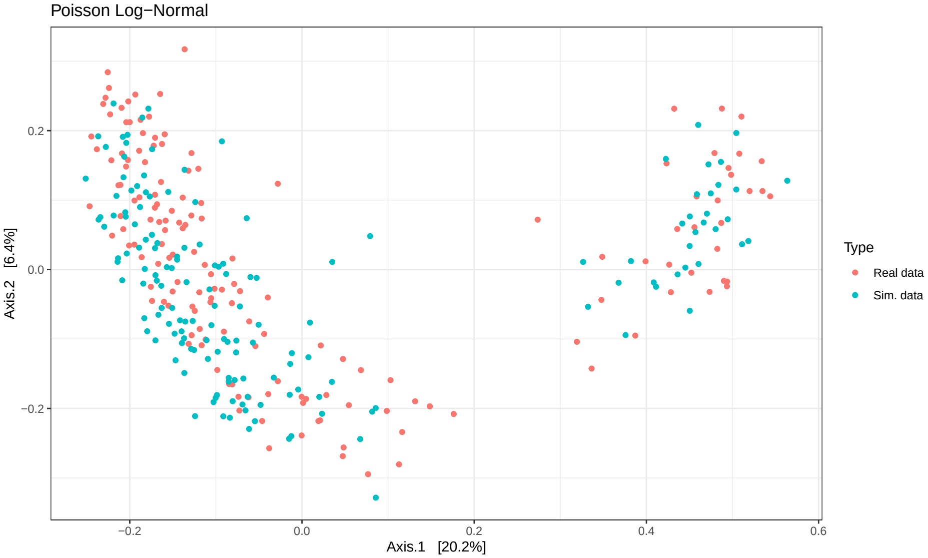
Mixture Layer

Can be combined with a mixture model

Geometrical view



Example of Poisson Log-Normal



Pros and Cons

Pros

- + Good for heterogeneity
- + Good for dispersion
- + Good for correlations between OTUs
- + Modeling done at the count level
  $\rightsquigarrow$ counts can be on different scales and come from different sources

Pros and Cons

Pros

- + Good for heterogeneity
- + Good for dispersion
- + Good for correlations between OTUs
- + Modeling done at the count level
  $\rightsquigarrow$ counts can be on different scales and come from different sources

Cons

- - The model is not parsimonious: $p(p+3)/2$ parameters
- - Inference is quite involved
  $\rightsquigarrow$ iterative EM algorithm + gradient descent
- - Sequencing depths are only controlled on average

Partial Summary

Log-Normal models are good at

- modeling compositions;
- modeling dispersion around average compositions;
- modeling heterogeneity;
- modeling interactions between OTUs;
- accounting for covariates through the linear model.

Partial Summary

Log-Normal models are good at

- modeling compositions;
- modeling dispersion around average compositions;
- modeling heterogeneity;
- modeling interactions between OTUs;
- accounting for covariates through the linear model.

Log-Normal models are bad at

- being parsimonious

Partial Summary

Log-Normal models are good at

- modeling compositions;
- modeling dispersion around average compositions;
- modeling heterogeneity;
- modeling interactions between OTUs;
- accounting for covariates through the linear model.

Log-Normal models are bad at

- being parsimonious

- MLN results are easier to interpret (proportions)
- PLN allows to mix data from different sources (16S, ITS, etc.)

Outline

1. Tentative Syllabi
2. Motivation
3. Multinomial Models
4. Log-Normal Models
5. Applications
    - PCA
    - Linear Discriminant Analysis

PLN model in Microbial Ecology

PLN: a flexible models accounting for:

- Heterogeneity and average compositions ($\simeq$ first order moments)
- Dispersion and correlation between OTUs ($\simeq$ second order moments)
- Structuring covariates
- Counts coming from dfferent data sources

PLN model in Microbial Ecology

PLN: a flexible models accounting for:

- Heterogeneity and average compositions ($\simeq$ first order moments)
- Dispersion and correlation between OTUs ($\simeq$ second order moments)
- Structuring covariates
- Counts coming from dfferent data sources

Allows for *traditional* multivariate analysis:

Idea: put additional constraints in the model

- PCA $\rightsquigarrow$ small rank $\boldsymbol{\Sigma}$
- Linear Discriminant Analysis $\rightsquigarrow$ known group structure on $\boldsymbol{\mu}$
- Network Inference $\rightsquigarrow$ sparse/tree-like $\boldsymbol{\Sigma}^{-1}$
- Mixture Models $\rightsquigarrow$ unknown group structure on $\boldsymbol{\mu}$
- *etc.*

Outline

1 Tentative Syllabi

2 Motivation

3 Multinomial Models

4 Log-Normal Models

5 Applications
- PCA
- Linear Discriminant Analysis

PLN-PCA: summarize information

Dimension reduction and vizualization. Typical task in multivariate analysis

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

$\rightsquigarrow$ Find a low-dimensional base (PCA axes) to represent the latent covariance

Fit the PLNPCA models with offsets and various covariates.

```
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: vizualization

PLN PCA separates well the kind of tree

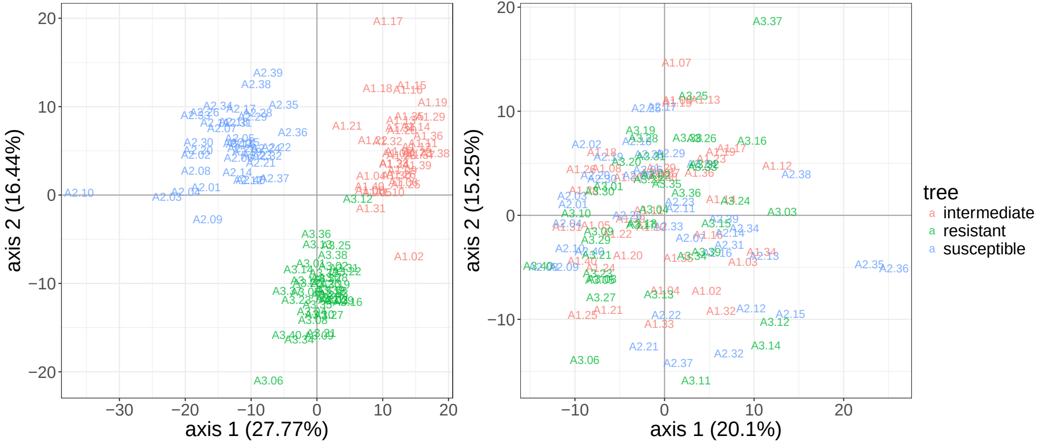


Figure: offset only offset + covariates

PCA: vizualization II

Introduction of covariates unravels hidden patterns

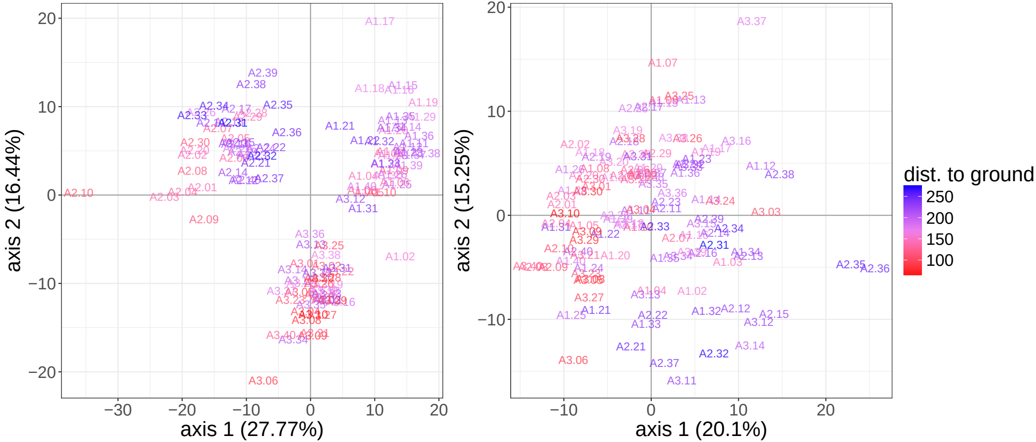


Figure: offset only offset + covariates

Outline

1. Tentative Syllabi

2. Motivation

3. Multinomial Models

4. Log-Normal Models

5. Applications
    - PCA
    - Linear Discriminant Analysis

Fit the PLNLDA models

find the linear combinaison that separates the grouping

Fit the model with offsets, and various covariates

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

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

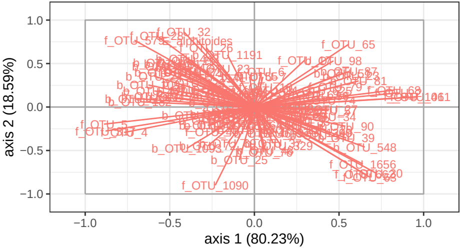
```
myLDA_tree$plot_LDA()
```

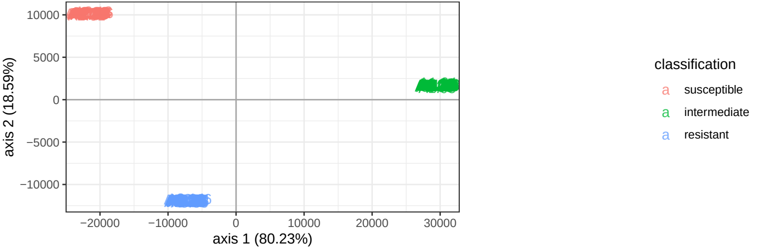
LDA on tree status

Axes contribution

axis 1 : 80.23%

axis 2 : 18.59%





Prediction error (10 fold cross-validation)

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

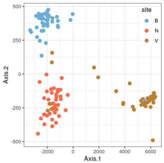
Conclusion

Summary PLN = generic model for multivariate counts

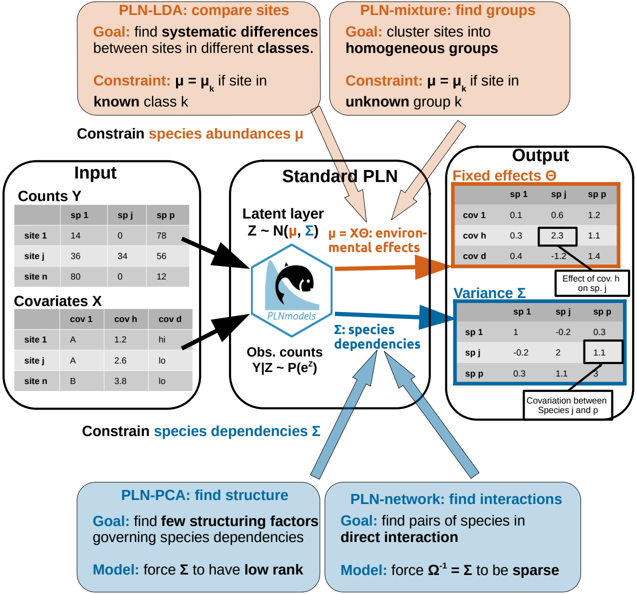
- Corrects for covariates and offset ($\simeq$ sequencing depths)
- Flexible statistical modeling
- `PLNmodels` R-package

Additional extensions

- Add technical/biological "zeros" (zero-inflation)
- Extensions: sparse PCA, mixture models
- Confidence interval and tests
- Missing data...



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

$\mu = X\Theta$: environ-mental 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

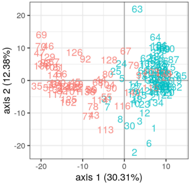
Goal: find **few structuring factors** governing species dependencies

Model: force Σ to have **low rank**

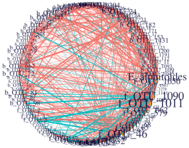
PLN-network: find interactions

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.

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.