Variational Inference in the Poisson lognormal model

Application to multivariate analysis of count data

Julien Chiquet, MIA Paris

joint work with M. Mariadasou, S. Robin

PASADENA Workshop, Jussieu, February, 15 2019

J.C., Mahendra Mariadassou, Stéphane Robin,
Variational inference for probabilistic Poisson PCA
http://dx.doi.org/10.1214/18-AOAS1177 Ann Appl Statist 12: 2674–2698, 2018

J.C., Mahendra Mariadassou, Stéphane Robin,
Variational inference for sparse network reconstruction from count data
https://arxiv.org/abs/1806.03120 (submitted)

PLNmodels package, development version on github

```
devtools::install_github("jchiquet/PLNmodels@v0.7.0.2")
```

https://jchiquet.github.io/PLNmodels/
http://julien.cremeriefamily.info/slides/mollusk/mollusk.html

Motivations: oak powdery mildew pathobiome

Metabarcoding data from [JFS+16]

- $n = 116$ leaves, $p = 114$ species (66 bacteria, 47 fungies + *E. alphitoides*)

```
counts[1:3, c(1:4, 48:51)]

##       f_1 f_2 f_3 f_4 E_alphitoides b_1045 b_109 b_1093
## A1.02  72   5 131   0             0      0     0      0
## A1.03 516  14 362   0             0      0     0      0
## A1.04 305  24 238   0             0      0     0      0
```

- $d = 8$ covariates (tree susceptibility, distance to trunk, orientation, ...)

```
covariates[1:3, ]

##               tree distTOtrunk distTOground pmInfection orientation
## A1.02 intermediate         202        155.5           1          SW
## A1.03 intermediate         175        144.5           0          SW
## A1.04 intermediate         168        141.5           0          SW
```

- Sampling effort in each sample (bacteria $\neq$ fungi)

```
offsets[1:3, 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
## [3,] 2122 2122 2122 2122          2122    480   480    480
```

Problematic & Basic formalism

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

- 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 pathobiome

- 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 dependences between count variables

Models for multivariate count data

If we were in a Gaussian world, the general linear model would be appropriate

For each sample $i = 1, \dots, n$, it explains

- the abundances of the p species ($\mathbf{Y}_i$)
- by the values of the d covariates $\mathbf{X}_i$ and the p offsets $\mathbf{O}_i$

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

+ null covariance $\Leftrightarrow$ independence $\rightsquigarrow$ uncorrelated species do not interact

But we are not, and there is no generic model for multivariate counts

- Data transformation ($\log, \sqrt{}$) : quick and dirty
- Non-Gaussian multivariate distributions: do not scale to data dimension yet
- Latent variable models: interaction occur in a latent (unobserved) layer

Models for multivariate count data

If we were in a Gaussian world, the general linear model would be appropriate

For each sample $i = 1, \dots, n$, it explains

- the abundances of the p species $(\mathbf{Y}_i)$
- by the values of the d covariates $\mathbf{X}_i$ and the p offsets $\mathbf{O}_i$

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

+ ~~null covariance $\Leftrightarrow$ independence $\rightsquigarrow$ uncorrelated species do not interact~~

But we are not, and there is no generic model for multivariate counts

- Data transformation $(\log, \sqrt{})$: quick and dirty
- Non-Gaussian multivariate distributions: do not scale to data dimension yet
- Latent variable models: interaction occur in a latent (unobserved) layer

Poisson-log normal (PLN) distribution

A latent Gaussian model

Originally proposed by Atchisson [AH89]

$$\mathbf{Z}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$$

$$\mathbf{Y}_i \,|\, \mathbf{Z}_i \sim \mathcal{P}(\exp\{\mathbf{O}_i + \mathbf{X}_i^\intercal \mathbf{\Theta} + \mathbf{Z}_i\})$$

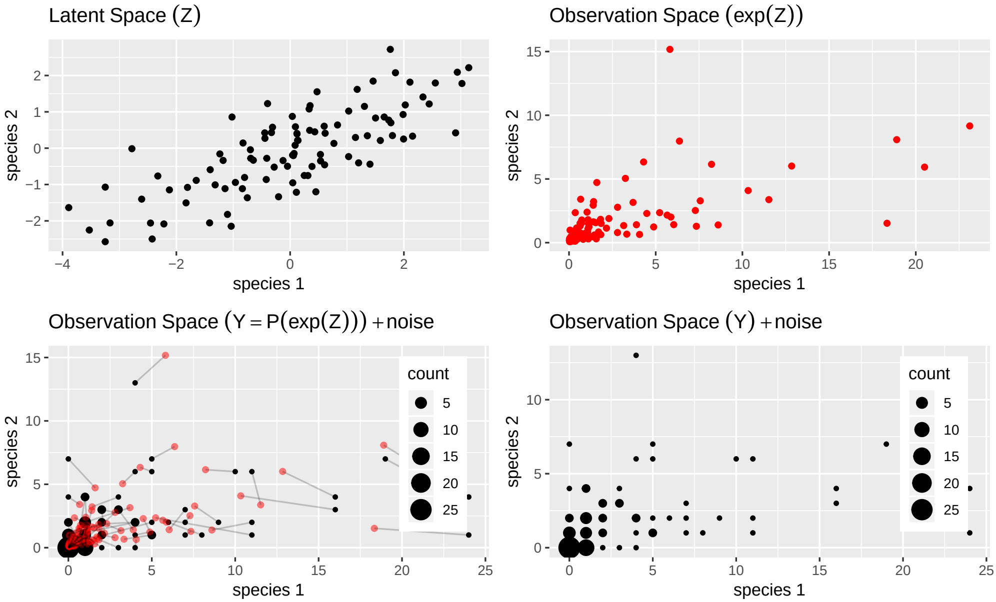
Interpretation

- Dependency structure encoded in the latent space (i.e. in $\mathbf{\Sigma}$)
- Additional effects are fixed
- Conditional Poisson distribution = noise model

Properties

+ over-dispersion
+ covariance with arbitrary signs
- maximum likelihood via EM algorithm is limited to a couple of variables

Geometrical view



Outline

Outline

Intractable EM

Aim of the inference:

- estimate $\boldsymbol{\beta} = (\boldsymbol{\Theta}, \boldsymbol{\Sigma})$
- predict the $\mathbf{Z}_i$

Maximum likelihood

PLN is an incomplete data model: try EM

$$\log p_{\boldsymbol{\beta}}(\mathbf{Y}) = \mathbb{E}[\log p_{\boldsymbol{\beta}}(\mathbf{Y}, \mathbf{Z}) \,|\, \mathbf{Y}] + \mathcal{H}[p_{\boldsymbol{\beta}}(\mathbf{Z} \,|\, \mathbf{Y})]$$

EM requires to evaluate (some moments of)

$$p(\mathbf{Z} \,|\, \mathbf{Y}) = \prod_i p(\mathbf{Z}_i \,|\, \mathbf{Y}_i)$$

but no close form for $p(\mathbf{Z}_i \,|\, \mathbf{Y}_i)$.

- [Kar05] resorts to numerical or Monte-Carlo integration.
- Variational approach [WJ08]: use a proxy of $p(\mathbf{Z} \,|\, \mathbf{Y})$.

Variational EM

Variational approximation: choose a class of distribution $\mathcal{Q}$

$$\mathcal{Q} = \left\{ \tilde{p} : \quad \tilde{p}(\mathbf{Z}) = \prod_i \tilde{p}_i(\mathbf{Z}_i), \quad \tilde{p}_i(\mathbf{Z}_i) = \mathcal{N}(\mathbf{Z}_i; \tilde{\mathbf{m}}_i, \tilde{\mathbf{s}}_i) \right\}$$

and maximize the lower bound ($\tilde{\mathbb{E}}$ = expectation under $\tilde{p}$)

$$J(\theta, \tilde{p}) = \log p_{\boldsymbol{\beta}}(\mathbf{Y}) - KL[\tilde{p}(\mathbf{Z}) \,||\, p_{\boldsymbol{\beta}}(\mathbf{Z} \,|\, \mathbf{Y})] = \tilde{\mathbb{E}}[\log p_{\boldsymbol{\beta}}(\mathbf{Y}, \mathbf{Z})] + \mathcal{H}[\tilde{p}(\mathbf{Z})]$$

Variational EM.

- VE step: find the optimal $\tilde{p}$:

$$\tilde{p}^h = \arg\max J(\boldsymbol{\beta}^h, \tilde{p}) = \arg\min_{\tilde{p} \in \mathcal{Q}} KL[\tilde{p}(\mathbf{Z}) \,||\, p_{\boldsymbol{\beta}^h}(Z \,|\, Y)]$$

- M step: update $\hat{\boldsymbol{\beta}}$

$$\hat{\boldsymbol{\beta}}^h = \arg\max J(\boldsymbol{\beta}, \tilde{p}^h) = \arg\max_{\boldsymbol{\beta}} \tilde{\mathbb{E}}[\log p_{\boldsymbol{\beta}}(\mathbf{Y}, \mathbf{Z})]$$

Variational EM

Variational approximation: choose a class of distribution $\mathcal{Q}$

$$\mathcal{Q} = \left\{ \tilde{p} : \quad \tilde{p}(\mathbf{Z}) = \prod_i \tilde{p}_i(\mathbf{Z}_i), \quad \tilde{p}_i(\mathbf{Z}_i) = \mathcal{N}(\mathbf{Z}_i; \tilde{\mathbf{m}}_i, \tilde{\mathbf{s}}_i) \right\}$$

and maximize the lower bound ($\tilde{\mathbb{E}}$ = expectation under $\tilde{p}$)

$$J(\theta, \tilde{p}) = \log p_{\boldsymbol{\beta}}(\mathbf{Y}) - KL[\tilde{p}(\mathbf{Z}) \,||\, p_{\boldsymbol{\beta}}(\mathbf{Z} \,|\, \mathbf{Y})] = \tilde{\mathbb{E}}[\log p_{\boldsymbol{\beta}}(\mathbf{Y}, \mathbf{Z})] + \mathcal{H}[\tilde{p}(\mathbf{Z})]$$

Variational EM.

- VE step: find the optimal $\tilde{p}$:

$$\tilde{p}^h = \arg\max J(\boldsymbol{\beta}^h, \tilde{p}) = \arg\min_{\tilde{p} \in \mathcal{Q}} KL[\tilde{p}(\mathbf{Z}) \,||\, p_{\boldsymbol{\beta}^h}(Z \mid Y)]$$

- M step: update $\hat{\boldsymbol{\beta}}$

$$\hat{\boldsymbol{\beta}}^h = \arg\max J(\boldsymbol{\beta}, \tilde{p}^h) = \arg\max_{\boldsymbol{\beta}} \tilde{\mathbb{E}}[\log p_{\boldsymbol{\beta}}(\mathbf{Y}, \mathbf{Z})]$$

Variational EM

Property: The lower $J(\boldsymbol{\beta}, \tilde{p})$ is bi-concave, i.e.

- wrt $\tilde{p} = (\tilde{\mathbf{M}}, \tilde{\mathbf{S}})$ for given $\boldsymbol{\beta}$
- wrt $\boldsymbol{\beta} = (\boldsymbol{\Sigma}, \boldsymbol{\Theta})$ for given $\tilde{p}$

but not jointly concave in general.

Optimization: projected gradient ascent for the complete parameter $(\tilde{\mathbf{m}}, \tilde{\mathbf{s}}, \boldsymbol{\beta})$

- algorithm: conservative convex separable approximations [Sva02]
- implementation: `NLopt` nonlinear-optimization package [Joh11]
- initialization: LM after log-trasnformation applied independently on each variables + concatenation of the regression coefficients + Pearson residuals

`PLNmodels` R-package:

https://github.com/jchiquet/PLNmodels

Outline

Model with offsets and covariates

Prepare the data set

Load the package, prepare the data, compute a basic offset

```
library(PLNmodels)
oaks <- prepare_data(counts, covariates)
str(oaks)

## 'data.frame': 116 obs. of  7 variables:
##  $ Abundance   : int [1:116, 1:114] 72 516 305 152 529 317 200 158 88 297 ...
##   ..- attr(*, "dimnames")=List of 2
##   .. ..$ : chr  "A1.02" "A1.03" "A1.04" "A1.05" ...
##   .. ..$ : chr  "f_1" "f_2" "f_3" "f_4" ...
##  $ tree        : Factor w/ 3 levels "intermediate",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ distTOtrunk : num  202 175 168 148 138 136 115 88 79 198 ...
##  $ distTOground: num  156 144 142 134 130 ...
##  $ pmInfection : int  1 0 0 1 1 1 1 5 1 0 ...
##  $ orientation : Factor w/ 2 levels "NE","SW": 2 2 2 2 2 2 2 2 2 2 ...
##  $ Offset      : num  9944 2104 1967 2498 2560 ...
```

Model with offsets and covariates

Fit the PLN model

Fit the model with offsets

```
system.time(PLN_offset <- PLN(Abundance ~ 1 + offset(log(Offset)), oaks))

##
##  Initialization...
##  Adjusting a PLN model with full covariance model
##  Post-treatments...
##  DONE!
##    user  system elapsed
##  10.220   0.017   2.850
```

Now the model with offsets and the 'tree' covariate

```
PLN_tree <- PLN(Abundance ~ 0 + tree + offset(log(Offset)), oaks)

##
##  Initialization...
##  Adjusting a PLN model with full covariance model
##  Post-treatments...
##  DONE!
```

Model with offsets and covariates

Accessing the model parameters

```
coef(PLN_offset) %>% exp() %>% head() %>% t() %>% knitr::kable(format = "latex")
```

	f_1	f_2	f_3	f_4	f_5	f_6
(Intercept)	0.0659714	0.0003894	0.1011897	0.0003132	0.0050419	0.005493

```
coef(PLN_tree) %>% exp() %>% head() %>% t() %>% knitr::kable(format = "latex")
```

	f_1	f_2	f_3	f_4	f_5	f_6
treeintermediate	0.0854752	0.0035076	0.1428223	0.0000063	0.0002539	0.0074122
treeresistant	0.0628723	0.0001377	0.1094356	0.0010107	0.0194141	0.0026414
treesusceptible	0.0524241	0.0001785	0.0650146	0.0001073	0.0122173	0.0081335

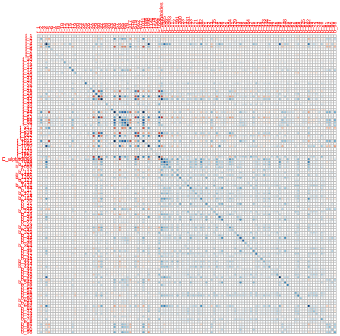
```
rbind(
  M00 = PLN_offset$criteria,
  M01 = PLN_tree$criteria) %>% knitr::kable(format = "latex")
```

	degrees_freedom	loglik	BIC	ICL	R_squared
M00	6669	-32376.95	-48227.80	-52270.95	0.9924747
M01	6897	-32503.54	-48896.29	-52808.74	0.9885133

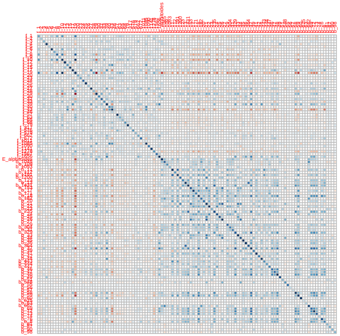
Model with offsets and covariates

A large part of the variance is explained by the covariates

```r
corrplot(
  vcov(PLN_offset),
  is.corr = FALSE,
  tl.cex = .5, cl.pos = "n"
)
```



```r
corrplot(
  vcov(PLN_tree),
  is.corr = FALSE,
  tl.cex = .5, cl.pos = "n"
)
```



PLN: natural extensions for multivariate analysis

Idea(s)

Put some additional constraint on the residual variance.

PCA: constraint the rank of $\mathbf{\Sigma}$.

LDA: a 'supervised' version of PCA

Network: put sparsity constraint on $\mathbf{\Omega} = \mathbf{\Sigma}^{-1}$.

Challenges

$\rightsquigarrow$ a variant of the variational algorithm is required for each model
$\rightsquigarrow$ interpretation is not exactly like in the "usual" Gaussian world

Outline

Probabilistic PCA

Dimension reduction. Typical task in multivariate analysis

Model: Probabilistic PCA (pPCA):

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

Recall that: $\text{rank}(\mathbf{\Sigma}) = q \quad \Leftrightarrow \quad \exists \mathbf{B}(p \times q) : \Sigma = \mathbf{B}\mathbf{B}^\intercal$.

pPCA in the PLN model. Variational inference:

$$\text{maximize } J(\boldsymbol{\beta}, \tilde{p})$$

$\rightsquigarrow$ Still bi-concave in $\boldsymbol{\beta} = (\mathbf{B}, \mathbf{\Theta})$ and $(\tilde{\mathbf{M}}, \tilde{\mathbf{S}})$

Model selection

Number of components q needs to be chosen.

Penalized 'likelihood'.

- $\log p_{\hat{\boldsymbol{\beta}}}(\mathbf{Y})$ intractable: replaced with $J(\hat{\boldsymbol{\beta}}, \tilde{p})$
- BIC $\rightsquigarrow$ $\tilde{\text{BIC}}_q = J(\hat{\boldsymbol{\beta}}, \tilde{p}) - \frac{1}{2} p(d+q)\log(n)$
- ICL $\rightsquigarrow$ $\tilde{\text{ICL}}_q = \tilde{\text{BIC}}_q - \mathcal{H}(\tilde{p})$

Chosen rank:

$$\hat{q} = \arg\max_q \tilde{\text{BIC}}_q \qquad \text{or} \qquad \hat{q} = \arg\max_q \tilde{\text{ICL}}_q$$

Visualization

PCA: Optimal subspaces nested when q increases.

PLN-pPCA: Non-nested subspaces.

$\rightsquigarrow$ For the selected dimension dimension $\hat{q}$:

- Compute the estimated latent positions $\mathbb{E}_{\tilde{p}}(\mathbf{Z}_i) = \tilde{\mathbf{M}}\hat{\mathbf{B}}^\top$
- Perform PCA on the $\tilde{\mathbf{M}}\hat{\mathbf{B}}^\top$
- Display result in any dimension $q \leq \hat{q}$

Goodness of fit

pPCA: Cumulated sum of the eigenvalues = % of variance preserved on the first q components.

PLN-pPCA: Deviance based criterion.

- Compute $\tilde{\mathbf{Z}}^{(q)} = \mathbf{O} + \mathbf{X}\hat{\mathbf{\Theta}}^\top + \tilde{\mathbf{M}}^{(q)} \left(\hat{\mathbf{B}}^{(q)}\right)^\top$
- Take $\lambda_{ij}^{(q)} = \exp\left(\tilde{Z}_{ij}^{(q)}\right)$
- Define $\lambda_{ij}^{\min} = \exp(\tilde{Z}_{ij}^{0})$ and $\lambda_{ij}^{\max} = Y_{ij}$
- Compute the Poisson log-likelihood $\ell_q = \log \mathbb{P}(\mathbf{Y}; \lambda^{(q)})$

Pseudo-R^2:

$$R_q^2 = \frac{\ell_q - \ell_{\min}}{\ell_{\max} - \ell_{\min}}$$

Outline

Fit the PLNPCA models

Fit the model with offsets, and various covariates

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

## Model with offset
PLN_offset <- PLNPCA(Abundance ~ 1 + offset(log(Offset)), data = oaks, ranks=Q)

## Models with offset and covariates (tree + orientation)
formula <- Abundance ~ 0 + tree + orientation + offset(log(Offset))
PLN_tree_orientation <- PLNPCA(formula, data = oaks, ranks = Q)

## model at initialization: log of count + LM
logLM_tree_orientation <-
  PLNPCA(
    formula, ranks = Q,
    control.main=list(inception="LM", maxeval=1),
    control.init=list(inception="LM", maxeval=1)
  )
```

Models selection criteria

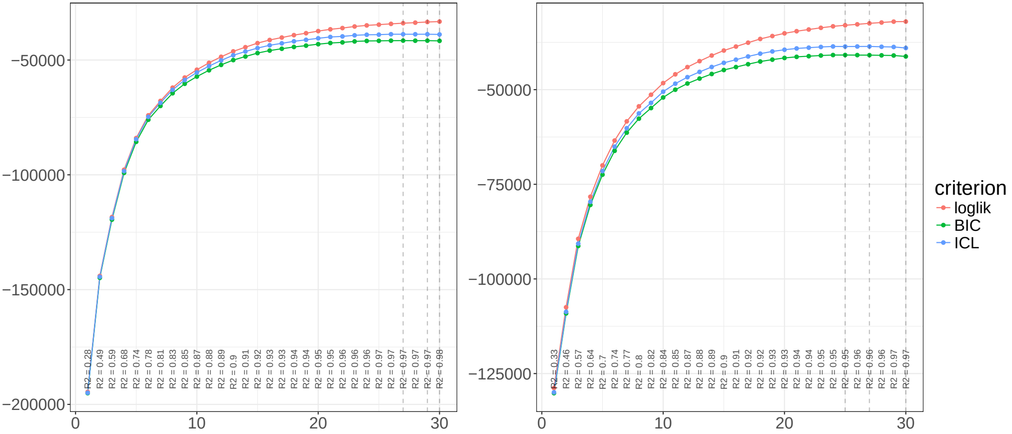


Figure: offset only: $\hat{q} = 24$ offset + covariates: $\hat{q} = 21$

PCA: vizualization I

basic transformation + LM fails at exhibiting basic structure in the data

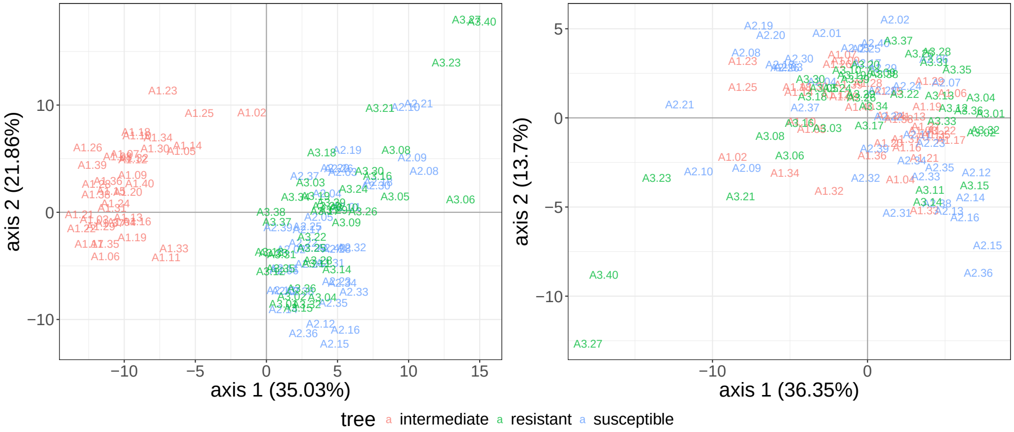


Figure: offset only offset + covariates

PCA: vizualization II

PLN PCA separates well the kind of tree

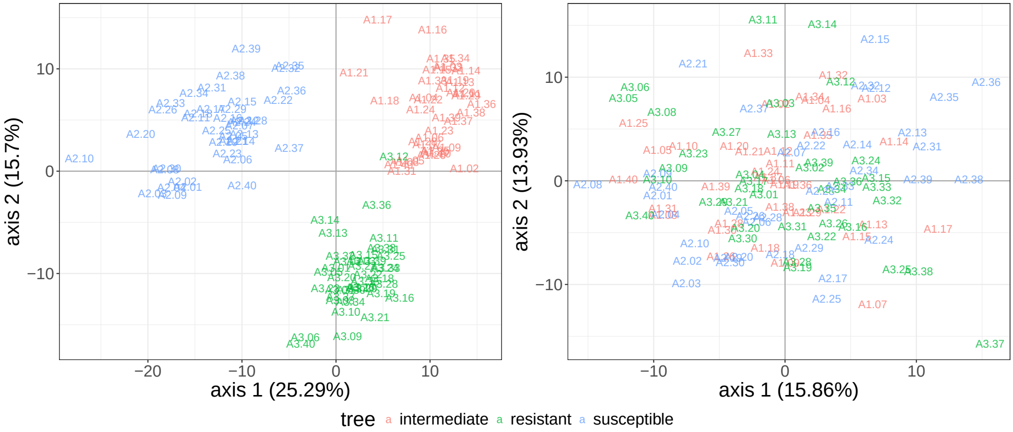


Figure: offset only offset + covariates

PCA: vizualization III

Introduction of covariates unravel hidden patterns

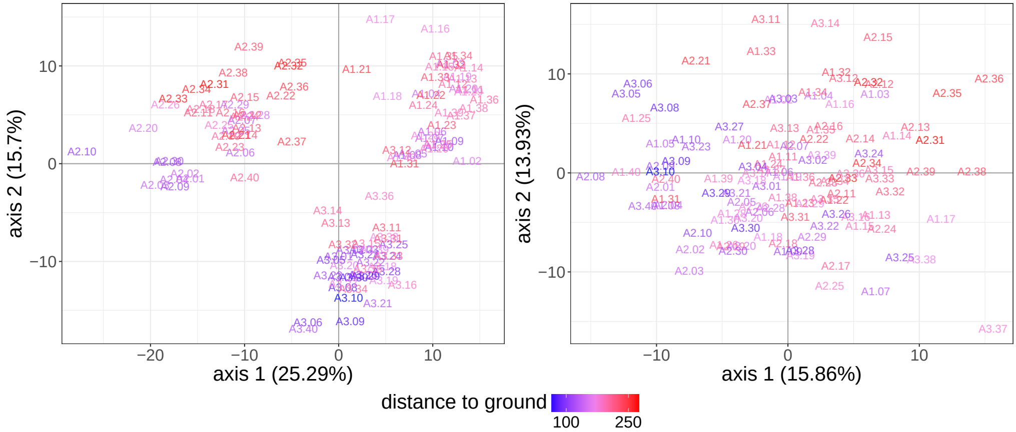


Figure: offset only offset + covariates

PCA: vizualization IV

Introduction of covariates unravel different groups of species

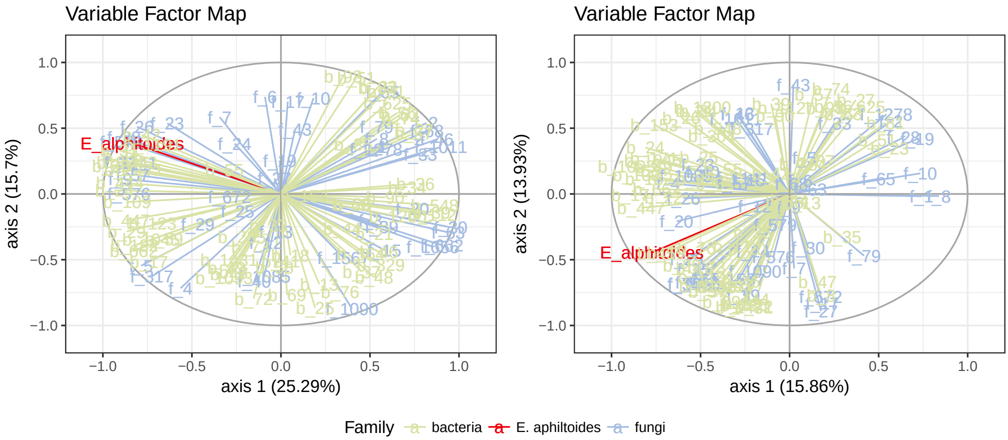


Figure: offset only offset + covariates

Outline

Gaussian LDA

Model

Assume the samples are distributed in K groups and note

- $\mathbf{G}$ the group membership matrix
- $\mathbf{U} = [\boldsymbol{\mu}_1^\top, \dots, \boldsymbol{\mu}_K^\top]^\top$ the matrix of group-specific means

The model is

$$\mathbf{Z}_i = \mathcal{N}(\underbrace{\mathbf{g}_i^\top \mathbf{U}}_{\boldsymbol{\mu}_i}, \boldsymbol{\Sigma})$$

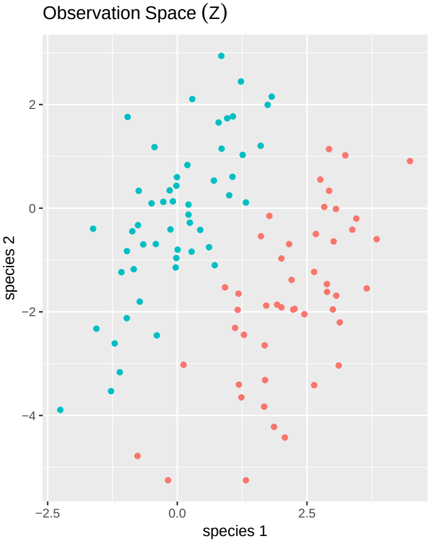
Aim of LDA. Find the linear combination(s) $\mathbf{Z}u$ ($u \in \mathbb{R}^p$) maximizing separation between groups

Solution

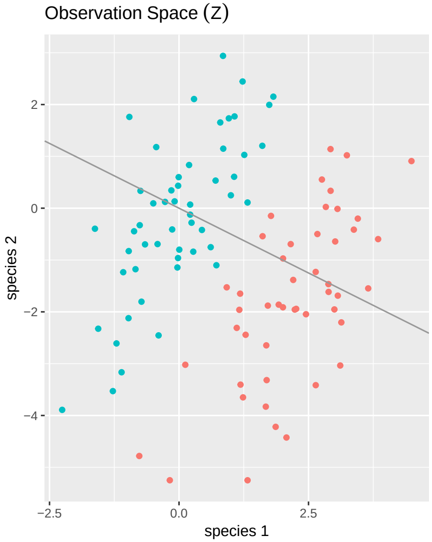
Find the first eigenvectors of $\mathbf{W}^{-1}\mathbf{B}$ where

- $\hat{\boldsymbol{\Sigma}}$ is the within-group variance matrix, i.e. the unbiased estimated of $\boldsymbol{\Sigma}$:
- $\hat{\mathbf{B}}$ is the between-group variance matrix, estimated from $\mathbf{U}$.

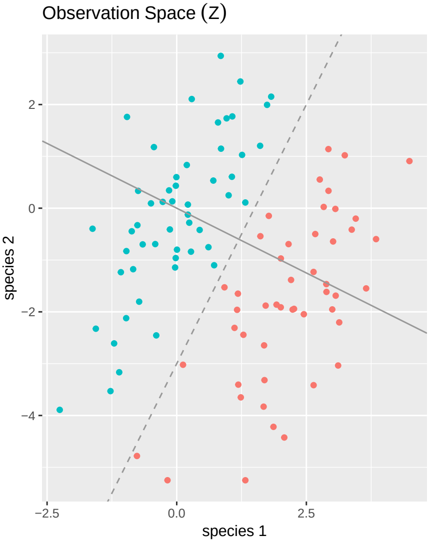
Geometric view: LDA



Geometric view: LDA



Geometric view: LDA

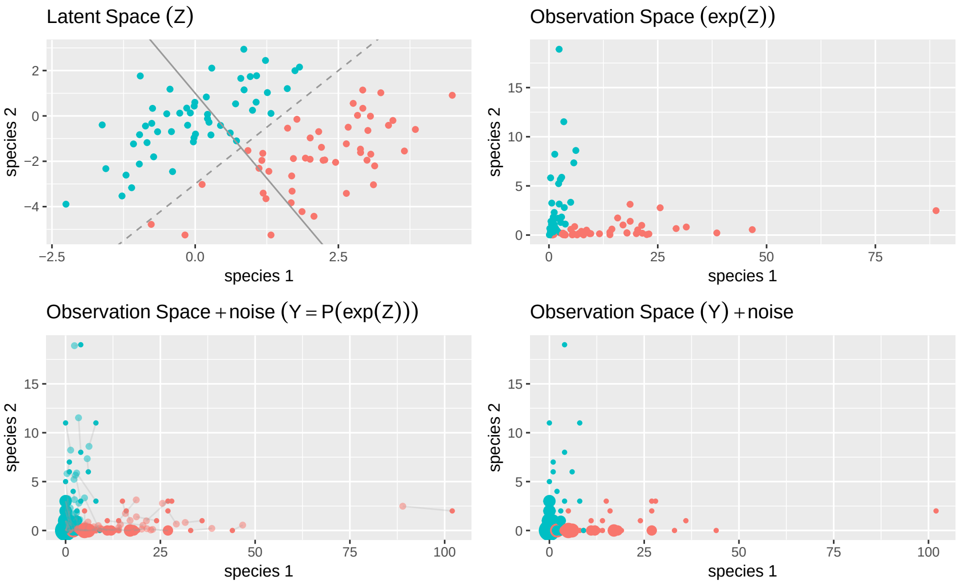


PLN LDA (I)

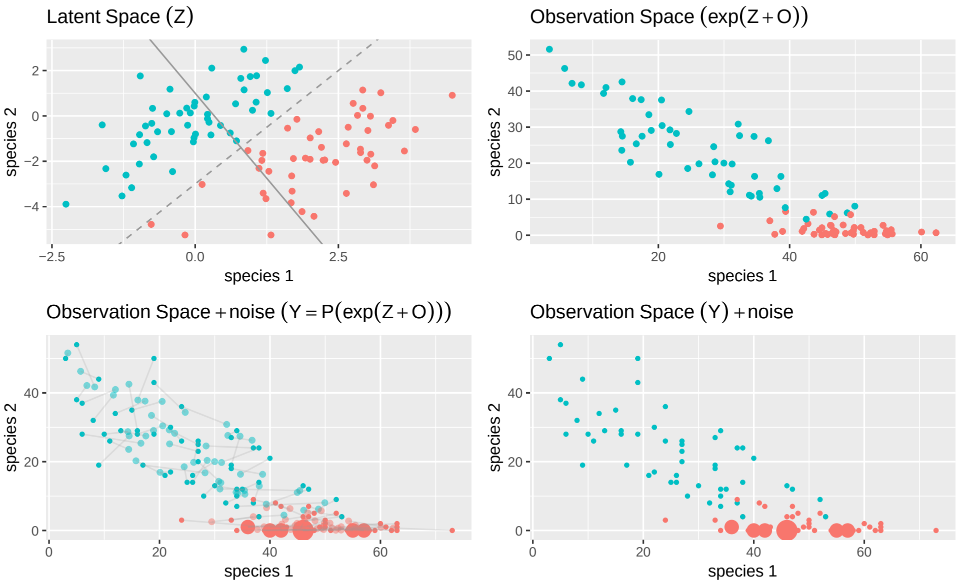
Model: LDA + Poisson layer

$$\mathbf{Z}_i \text{ indep.}, \quad \mathbf{Z}_i \sim \mathcal{N}(\mathbf{g}_i^\top \mathbf{U}, \mathbf{\Sigma})$$
$$\mathbf{Y}_i \text{ indep.} | \mathbf{Z}, \quad \mathbf{Y}_i \sim \mathcal{P}(\exp(\mathbf{O}_i + \mathbf{X}_i^\top \mathbf{\Theta} + \mathbf{Z}_i))$$

PLN LDA (I)


Latent Space (Z)
Observation Space (exp(Z))
Observation Space + noise (Y = P(exp(Z)))
Observation Space (Y) + noise
species 1
species 2

PLN LDA (I)



Latent Space (Z)
Observation Space (exp(Z + O))
Observation Space + noise (Y = P(exp(Z + O)))
Observation Space (Y) + noise
species 1
species 2

PLN LDA (II)

Similar to normal PLN with

- $\mathbf{X} \to (\mathbf{X}, \mathbf{G})$
- $\boldsymbol{\Theta} \to (\boldsymbol{\Theta}, \mathbf{U})$

Inference:

- Use variational inference to estimate $\boldsymbol{\Sigma}, \mathbf{U}$ and $\mathbf{B}$ and $\tilde{\mathbf{Z}}_i$
- Compute $\hat{\boldsymbol{\Theta}}$ as

$$\hat{\boldsymbol{\Theta}} = \frac{1}{K-1} \sum_k n_k (\hat{\boldsymbol{\mu}}_k - \hat{\boldsymbol{\mu}}_\bullet)(\hat{\boldsymbol{\mu}}_k - \hat{\boldsymbol{\mu}}_\bullet)^\intercal$$

- Compute first $K-1$ eigenvectors of $\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\Theta}} = \mathbf{P}^\top \Lambda \mathbf{P}$

PLN LDA (III)

Graphical representation

Mimick gaussian LDA:

- Center the estimated latent positions $\tilde{\mathbf{Z}}$
- Compute the estimated coordinates along the discriminant axes

$$\tilde{\mathbf{Z}}^{LDA} = \tilde{\mathbf{Z}}\mathbf{P}\Lambda^{1/2}$$

Prediction

For each group k

- Assume that the new sample $\mathbf{Y}_{\text{new}}$ comes from group k
- Compute (variational) likelihood $p_k = P(\mathbf{Y}_{\text{new}} | \hat{\Sigma}, \hat{\Theta}, \hat{\boldsymbol{\mu}}_k)$
- Compute posterior probability $\pi_k \propto \frac{n_k p_k}{n}$

Assign to group with highest π_k

PLN LDA (III)

Graphical representation

Mimick gaussian LDA:

- Center the estimated latent positions $\tilde{\mathbf{Z}}$
- Compute the estimated coordinates along the discriminant axes

$$\tilde{\mathbf{Z}}^{LDA} = \tilde{\mathbf{Z}}\mathbf{P}\Lambda^{1/2}$$

Prediction

For each group k

- Assume that the new sample $\mathbf{Y}_{\text{new}}$ comes from group k
- Compute (variational) likelihood $p_k = P(\mathbf{Y}_{\text{new}}|\hat{\mathbf{\Sigma}}, \hat{\mathbf{\Theta}}, \hat{\boldsymbol{\mu}}_k)$
- Compute posterior probability $\pi_k \propto \frac{n_k p_k}{n}$

Assign to group with highest π_k

Outline

Fit the PLNLDA models

find the linear combinaison that separates the grouping

Fit the model with offsets, and the 'tree' covariates

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

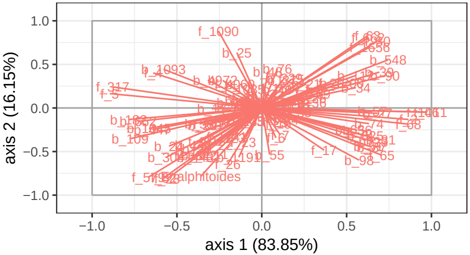
```
##
##  Performing discriminant Analysis...
##  DONE!
```

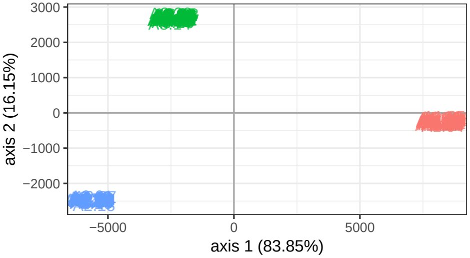
```
plot(myLDA_tree)
```

LDA on tree status

Axes contribution

axis 1 : 83.85%
axis 2 : 16.15%





classification
- a intermediate
- a resistant
- a susceptible

Outline

Background on Gaussian Graphical Models

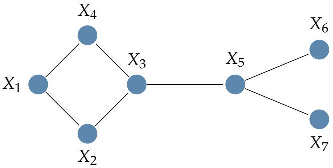
Suppose $\mathbf{Y}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{\Omega}^{-1} = \mathbf{\Sigma})$

Conditional independence structure

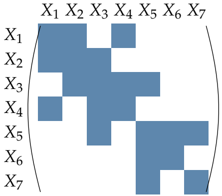
$$(i,j) \notin \mathcal{E} \Leftrightarrow Y_i \perp\!\!\!\perp Y_j \,|\, Y_{\backslash\{i,j\}} \Leftrightarrow \mathbf{\Omega}_{ij} = 0.$$

Graphical interpretation

$\mathcal{G} = (\mathcal{P}, \mathcal{E})$



$\mathbf{\Omega}$



Graphical-Lasso [BDE08,YL08,FHT07]

Network reconstruction is (roughly) a variable selection problem

$$\hat{\mathbf{\Omega}}_\lambda = \arg\max_{\mathbf{\Theta} \in \mathbb{S}_+} \ell(\mathbf{\Omega}; \mathbf{Y}) - \lambda \|\mathbf{\Theta}\|_1$$

PLN network model

Model:

$$\mathbf{Z}_i \text{ iid} \sim \mathcal{N}_p(\mathbf{0}_p, \mathbf{\Omega}^{-1}),$$
$$\mathbf{Y}_i \mid \mathbf{Z}_i \sim \mathcal{P}(\exp\{\mathbf{O}_i + \mathbf{X}_i^\top \mathbf{\Theta} + \mathbf{Z}_i\})$$

$\mathbf{\Omega}$ sparse

Interest: Similar to Gaussian graphical model (GGM) inference

Sparsity-inducing regularization: graphical lasso

$$\log p_{\boldsymbol{\beta}}(\mathbf{Y}) - \lambda \left\|\mathbf{\Omega}\right\|_{1,\text{off}}$$

Cheat: Use the PLN model and infer the graphical model of Z

Graphical model of $Z \neq$ *Graphical model of* Y

PLN network graphical model: examples I

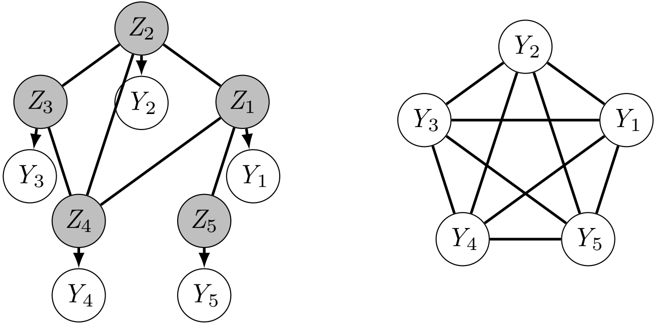


Figure: Left: joint distribution of $p(Z_i, Y_i)$. Right: marginal distribution $p(Y_i)$. The graph on the right is a clique because the graph of the Z_i's on the left is connected.

PLN network graphical model: examples II

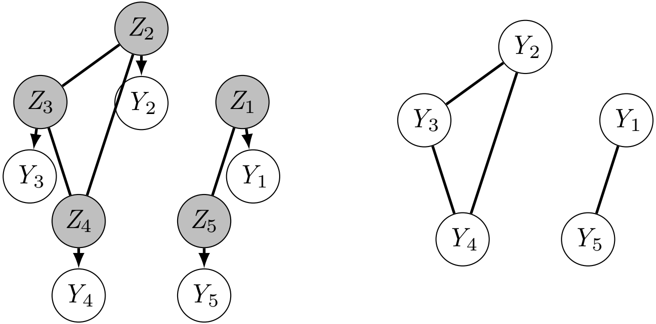


Figure: Left: joint distribution of $p(Z_i, Y_i)$. Right: marginal distribution $p(Y_i)$.

Variational inference

Same problem: $\log p_{\boldsymbol{\beta}}(\mathbf{Y})$ is intractable

Variational approximation: maximize

$$J(\boldsymbol{\beta}, \tilde{p}) - \lambda \, \|\boldsymbol{\Omega}\|_{1,\text{off}} = \tilde{\mathbb{E}}[\log p_{\boldsymbol{\beta}}(\mathbf{Y}, \mathbf{Z})] + \mathcal{H}[\tilde{p}(\mathbf{Z})] - \lambda \, \|\boldsymbol{\Omega}\|_{1,\text{off}}$$

taking $\tilde{p} \in \mathcal{Q}$.

$\rightsquigarrow$ Still bi-concave in $\boldsymbol{\beta} = (\boldsymbol{\Omega}, \boldsymbol{\Theta})$ and $\tilde{p} = (\tilde{\mathbf{M}}, \tilde{\mathbf{S}})$. Ex:

$$\hat{\boldsymbol{\Omega}} = \arg\max_{\boldsymbol{\Omega}} \frac{n}{2}\left(\log |\boldsymbol{\Omega}| - \operatorname{tr}(\hat{\boldsymbol{\Sigma}}\boldsymbol{\Omega})\right) - \lambda \|\boldsymbol{\Omega}\|_{1,\text{off}} : \quad \text{gLasso problem}$$

Sparsity level λ needs to be chosen.

Stability-based approach for Network by resampling: StARS

1. Infers B networks $\mathbf{\Omega}^{(b,\lambda)}$ on subsamples of size m for varying λ.
2. Frequency of inclusion of each edges $e = i \sim j$ is estimated by

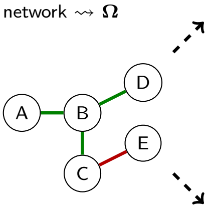
$$p_e^\lambda = \#\{b : \Omega_{ij}^{(b,\lambda)} \neq 0\}/B$$

3. Variance of inclusion of edge e is $v_e^\lambda = p_e^\lambda(1 - p_e^\lambda)$.
4. Network stability is $\mathrm{stab}(\lambda) = 1 - 2\overline{v}^\lambda$ where $\overline{v}^\lambda$ is the average of the v_e^λ.

$\rightsquigarrow$ StARS[1] selects the smallest λ (densest network) for which $\mathrm{stab}(\lambda) \geq 1 - 2\beta$

[1] [LRW10] suggest using $2\beta = 0.05$ and $m = \lfloor 10\sqrt{n} \rfloor$ based on theoretical results.

Simulation study



abundances $\log \mathbf{a}_i$

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

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

proportions $\boldsymbol{\pi}_i$ $\longrightarrow$ counts

	$i=1$	$i=2$
A	3	8
B	3	6
C	5	2
D	8	30
E	1	4
N_i	20	50

Figure: Compositional model used for data generation

i) Draw (unreachable) *abundances* $\mathbf{a}_i$: $\log(\mathbf{a}_i) \sim \mathcal{N}(\mathbf{X}\boldsymbol{\Theta}, \boldsymbol{\Omega}^{-1})$
  - $\mathbf{X}$ accounts for some covariates
  - $\boldsymbol{\Omega}$ is the latent network between species

ii) Transform abundances $\mathbf{a}_i$ to *proportions* $\boldsymbol{\pi}_i$ with logistic-transform

iii) Draw observed *counts* $Y_i \sim \mathcal{M}(N_i, \boldsymbol{\pi}_i)$ with random N_i – the sampling effort

Simulation results

Non-compositional methods fail

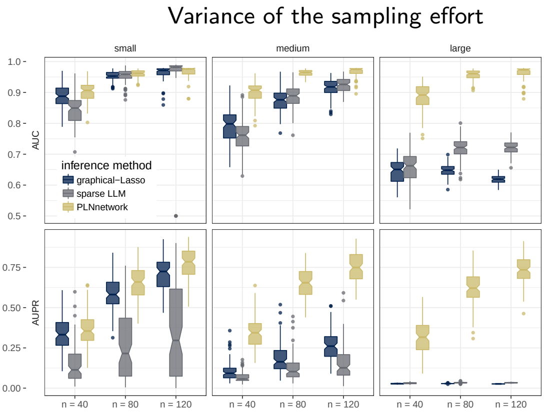


Figure: Effect of the variability of the sampling effort on the quality of the reconstruction of 50-node random networks (100 simulations.)

Simulation results

Accounting for covariates effect does matter

		area under the ROC			area under the PR		
covar.	**method**	$\mathbf{n = p/2}$	$\mathbf{n = p}$	$\mathbf{n = 2p}$	$\mathbf{n = p/2}$	$\mathbf{n = p}$	$\mathbf{n = 2p}$
		scale-free network					
small	`PLNnetwork`	.66 (0.05)	**.78** (0.05)	**.91** (0.03)	**.11** (0.04)	**.25** (0.07)	**.49** (0.08)
	`sparCC`	.66 (0.05)	.73 (0.05)	.79 (0.05)	.09 (0.03)	.16 (0.05)	.24 (0.07)
	`SPiEC-Easi`	**.67** (0.04)	.77 (0.05)	.85 (0.04)	.10 (0.03)	.17 (0.05)	.27 (0.07)
medium	`PLNnetwork`	**.62** (0.05)	**.73** (0.05)	**.85** (0.05)	**.09** (0.03)	**.18** (0.06)	**.34** (0.08)
	`sparCC`	.55 (0.05)	.57 (0.05)	.58 (0.05)	.05 (0.01)	.05 (0.01)	.06 (0.01)
	`SPiEC-Easi`	.61 (0.04)	.66 (0.04)	.71 (0.03)	.06 (0.01)	.06 (0.01)	.07 (0.01)
large	`PLNnetwork`	**.58** (0.05)	**.67** (0.05)	**.78** (0.05)	**.07** (0.03)	**.12** (0.04)	**.23** (0.07)
	`sparCC`	.52 (0.04)	.53 (0.04)	.53 (0.05)	.04 (0.01)	.04 (0.01)	.04 (0.01)
	`SPiEC-Easi`	.57 (0.04)	.60 (0.03)	.65 (0.03)	.05 (0.01)	.05 (0.01)	.05 (0.01)

Table: Areas under the ROC curve and Areas under the Precision-Recall curve of the compositional methods (`PLNnetwork`, `sparCC` and `SPiEC-Easi`) in various settings, averaged over 100 simulations, with standard errors.

Simulation results

Accounting for covariates effect does matter

		area under the ROC			area under the PR		
covar.	**method**	$\mathbf{n = p/2}$	$\mathbf{n = p}$	$\mathbf{n = 2p}$	$\mathbf{n = p/2}$	$\mathbf{n = p}$	$\mathbf{n = 2p}$
		random network					
small	PLNnetwork	.77 (0.07)	**.90** (0.04)	**.96** (0.01)	**.14** (0.07)	**.36** (0.11)	**.64** (0.09)
	sparCC	.76 (0.06)	.83 (0.06)	.89 (0.04)	.11 (0.05)	.23 (0.09)	.36 (0.11)
	SPiEC-Easi	**.78** (0.05)	.87 (0.04)	.92 (0.03)	.11 (0.05)	.23 (0.09)	.36 (0.11)
medium	PLNnetwork	**.72** (0.06)	**.85** (0.05)	**.94** (0.02)	**.09** (0.04)	**.24** (0.09)	**.49** (0.10)
	sparCC	.59 (0.06)	.61 (0.07)	.62 (0.06)	.03 (0.01)	.04 (0.02)	.04 (0.02)
	SPiEC-Easi	.67 (0.05)	.74 (0.05)	.77 (0.03)	.04 (0.01)	.05 (0.02)	.05 (0.01)
large	PLNnetwork	**.64** (0.07)	**.78** (0.06)	**.88** (0.04)	**.06** (0.03)	**.14** (0.07)	**.29** (0.09)
	sparCC	.54 (0.05)	.53 (0.06)	.54 (0.06)	.02 (0.01)	.02 (0.01)	.03 (0.01)
	SPiEC-Easi	.61 (0.05)	.65 (0.04)	.68 (0.03)	.03 (0.00)	.03 (0.00)	.03 (0.01)

Table: Areas under the ROC curve and Areas under the Precision-Recall curve of the compositional methods (PLNnetwork, sparCC and SPiEC-Easi) in various settings, averaged over 100 simulations, with standard errors.

Simulation results

Accounting for covariates effect does matter

		area under the ROC			area under the PR		
covar.	**method**	$\mathbf{n = p/2}$	$\mathbf{n = p}$	$\mathbf{n = 2p}$	$\mathbf{n = p/2}$	$\mathbf{n = p}$	$\mathbf{n = 2p}$
		community network					
small	PLNnetwork	.60 (0.04)	.69 (0.04)	**.78** (0.05)	**.17** (0.03)	**.26** (0.04)	**.38** (0.05)
	sparCC	**.62** (0.04)	.66 (0.04)	.70 (0.04)	.16 (0.02)	.21 (0.04)	.26 (0.04)
	SPiEC-Easi	**.62** (0.04)	**.70** (0.04)	.77 (0.04)	**.17** (0.02)	.24 (0.04)	.31 (0.04)
medium	PLNnetwork	.57 (0.03)	**.65** (0.04)	**.73** (0.05)	**.15** (0.02)	**.22** (0.03)	**.31** (0.05)
	sparCC	.55 (0.03)	.56 (0.04)	.56 (0.03)	.11 (0.02)	.12 (0.02)	.12 (0.02)
	SPiEC-Easi	**.58** (0.03)	.63 (0.03)	.67 (0.03)	.13 (0.02)	.14 (0.02)	.15 (0.02)
large	PLNnetwork	**.55** (0.03)	**.60** (0.04)	**.67** (0.04)	**.13** (0.02)	**.17** (0.03)	**.24** (0.04)
	sparCC	.52 (0.03)	.52 (0.03)	.52 (0.03)	.10 (0.02)	.10 (0.02)	.10 (0.02)
	SPiEC-Easi	**.55** (0.03)	.58 (0.03)	.62 (0.03)	.11 (0.01)	.11 (0.02)	.12 (0.01)

Table: Areas under the ROC curve and Areas under the Precision-Recall curve of the compositional methods (PLNnetwork, sparCC and SPiEC-Easi) in various settings, averaged over 100 simulations, with standard errors.

Outline

PLNnetwork models: consensus or tree-specific networks?

We consider 3 setups

1. resistant samples, with covariates
2. susceptible samples, with covariates
3. both samples samples, with covariates + tree effect and interactions

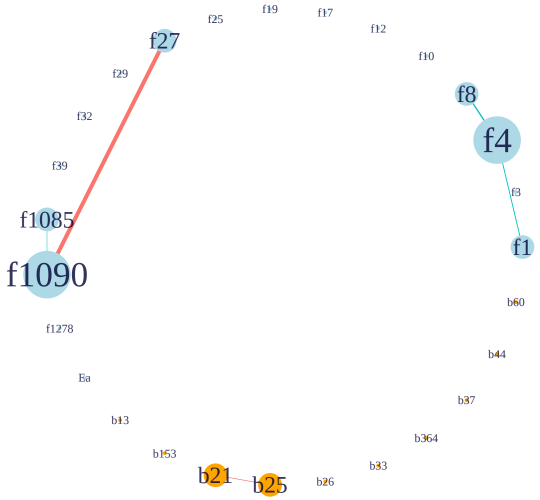
Network inference

`PLNnetwork` + 'StARS' for model selection

- 100 resamplings
- high level of stability (edges frequencies > 0.995)

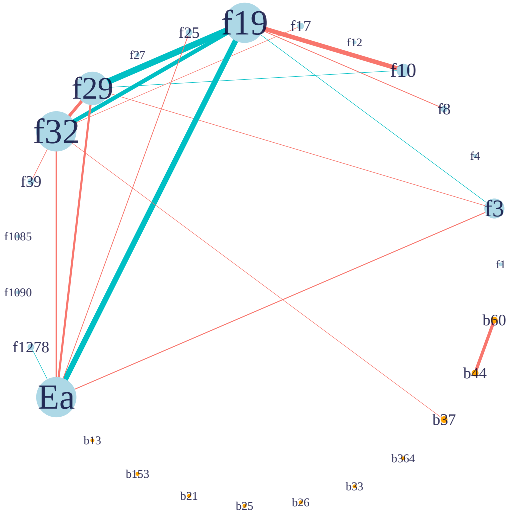
PLNnetwork models: resistant

Trees resistant to mildew (*E. Alphitoïdes*)



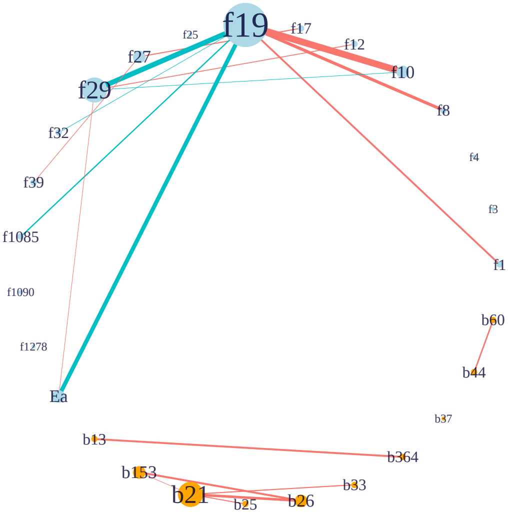
PLNnetwork models: susceptible

Trees susceptibles to mildew (*E. Alphitoïdes*)



PLNnetwork models: consensus

Both Trees



Discussion

Summary

- PLN = generic model for multivariate count data analysis
- Allows for covariates
- Flexible modeling of the covariance structure
- Efficient VEM algorithm
- `PLNmodels` package: `https://github.com/jchiquet/PLNmodels`

Extensions

- Model selection criterion for network inference
- Other covariance structures (spatial, time series, ...)
- Mixture model in the latent space
- Confidence interval and tests for the regular PLN

Statistical properties of variational estimates

General properties of VEM inference.

- VEM stationary point $\neq$ log-likelihood stationary point
- Some consistency results, typically when $p(Z \mid Y)$ asymptotically belongs to $\mathcal{Q}$ (SBM, Bayesian logistic regression).

Using VEM output as a starting point for regular inference:

- Get maximum-(composite-)likelihood estimates starting from $\tilde{\boldsymbol{\beta}}_{VEM}$ (proposed internship)

$\rightsquigarrow$ Hopefully: few iterations are needed

Thanks to you for your patience and to my co-workers

References

John Aitchison and CH Ho.
The multivariate poisson-log normal distribution.
Biometrika, 76(4):643–653, 1989.

B. Jakuschkin, V. Fievet, L. Schwaller, T. Fort, C. Robin, and C. Vacher.
Deciphering the pathobiome: Intra-and interkingdom interactions involving the pathogen Erysiphe alphitoides.
Microbial ecology, pages 1–11, 2016.

Steven G Johnson.
The NLopt nonlinear-optimization package, 2011.

D. Karlis.
EM algorithm for mixed Poisson and other discrete distributions.
Astin bulletin, 35(01):3–24, 2005.

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.

Krister Svanberg.
A class of globally convergent optimization methods based on conservative convex separable approximations.
SIAM journal on optimization, 12(2):555–573, 2002.

M. J. Wainwright and M. I. Jordan.
Graphical models, exponential families, and variational inference.
Found. Trends Mach. Learn., 1(1–2):1–305, 2008.