PLNmodels

A collection of Poisson lognormal models for multivariate analysis of count data

Julien Chiquet, MIA Paris

joint work with M. Mariadasou, S. Robin

AG MIA, Jouy-en-Josas, May, 22 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
In Proceedings of the 19th International Conference on Machine Learning (ICML'19)

PLNmodels package, development version on github
`install.packages("PLNmodels")`
https://jchiquet.github.io/PLNmodels/

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 a generic framework to model dependences between count variables

- account for peculiarities of count data
  $\rightsquigarrow$ vary over many orders of magnitude
  $\rightsquigarrow$ are overdispersed
- 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

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 $\mathbf{O}_i$

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

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
- 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 $\mathbf{O}_i$

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

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
- Latent variable models: interaction occur in a latent (unobserved) layer

Poisson-log normal (PLN) distribution

A latent Gaussian model

Originally proposed by Atchisson [AH89]

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

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

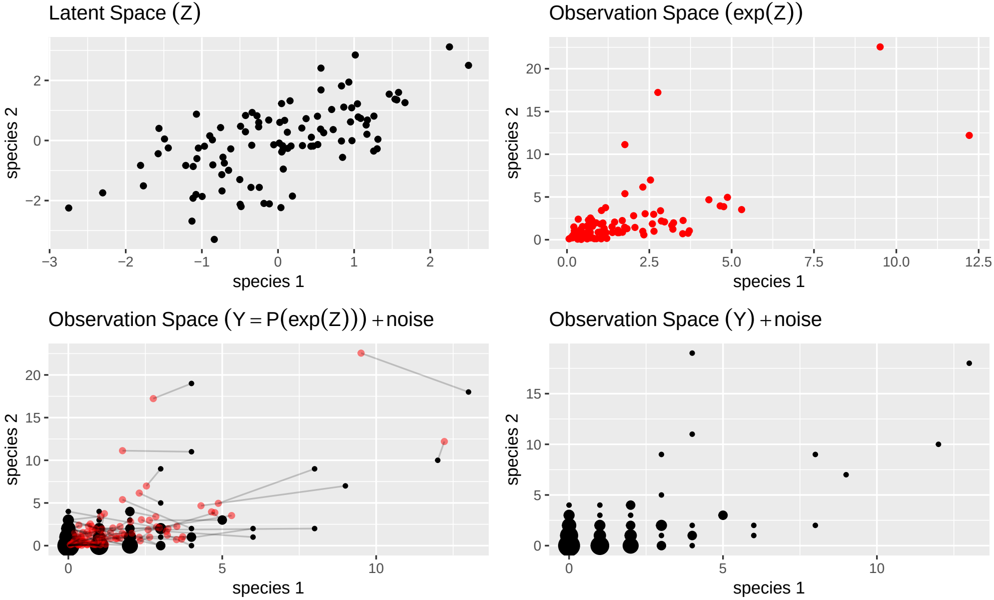
Interpretation

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

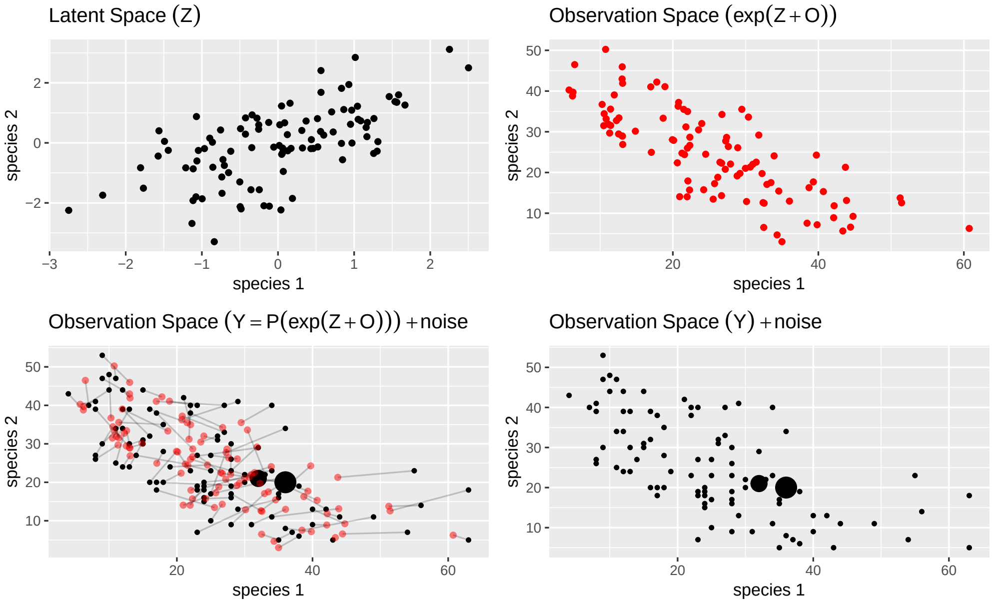
Properties

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

Geometrical view



Geometrical view (with offset)



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

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

EM requires to evaluate (some moments of)

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

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

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

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 $\tilde{p}$)

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

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 $\hat{\boldsymbol{\beta}}$

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

Optimization & Implementation

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/C++-package: `https://jchiquet.github.io/PLNmodels`

PLN: natural extensions towards multivariate analysis

- PCA: rank constraint on $\mathbf{\Sigma}$.

$$\mathbf{Z}_i \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{\Sigma} = \mathbf{B}\mathbf{B}^\top), \quad \mathbf{B} \in \mathcal{M}_{pk} \text{ with orthogonal columns.}$$

- Network: sparsity constraint on inverse covariance.

$$\mathbf{Z}_i \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{\Sigma} = \mathbf{\Omega}^{-1}), \quad \|\mathbf{\Omega}\|_1 < c.$$

- LDA: maximize separation between groups with means $\mathbf{M} = [\boldsymbol{\mu}_1^\top, \dots, \boldsymbol{\mu}_K^\top]^\top$

$$\mathbf{Z}_i \sim \mathcal{N}(\boldsymbol{\mu}_i = \mathbf{g}_i^\top \mathbf{M}, \mathbf{\Sigma}), \quad \mathbf{g}_i \text{ a group indicator vector.}$$

- Clustering: mixture model in the latent space

$$\mathbf{Z}_i \sim \prod_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{\mu}_k, \mathbf{\Sigma}_k), \quad \text{with, e.g., } \mathbf{\Sigma}_k \text{ diagonal matrices}$$

Challenge: a variant of the variational algorithm is required for each model

PLN network model

Model:

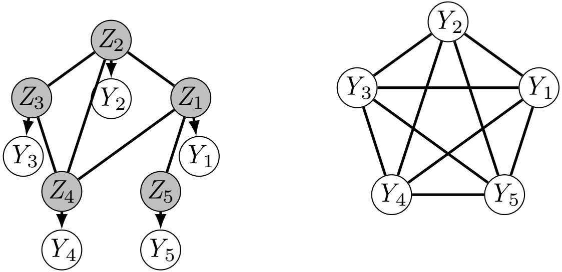
$$\mathbf{Z}_i \text{ iid} \sim \mathcal{N}_p(\mathbf{0}_p, \boldsymbol{\Omega}^{-1}), \qquad \boldsymbol{\Omega} \text{ sparse}, \quad \|\boldsymbol{\Omega}\|_{1,\text{offdiagonal}} < c$$

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

Cheat: Use the PLN model and infer the graphical model of $\mathbf{Z}$

$$(i,j) \notin \mathcal{E} \Leftrightarrow Z_i \perp\!\!\!\perp Z_j | Z_{\backslash\{i,j\}} \Leftrightarrow \boldsymbol{\Omega}_{ij} = 0.$$

Graphical interpretation: $p(\mathbf{Z}_i, \mathbf{Y}_i)$ vs $p(\mathbf{Y}_i)$



PLN network model

Model:

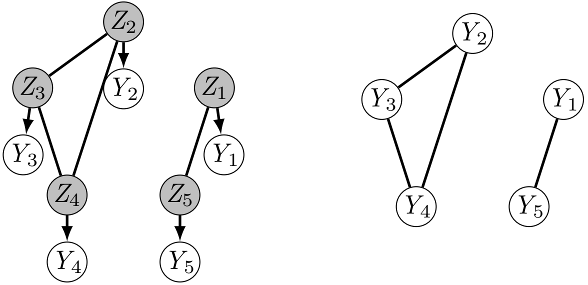
$$\mathbf{Z}_i \text{ iid} \sim \mathcal{N}_p(\mathbf{0}_p, \boldsymbol{\Omega}^{-1}), \qquad \boldsymbol{\Omega} \text{ sparse}, \quad \|\boldsymbol{\Omega}\|_{1,\text{offdiagonal}} < c$$

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

Cheat: Use the PLN model and infer the graphical model of $\mathbf{Z}$

$$(i,j) \notin \mathcal{E} \Leftrightarrow Z_i \perp\!\!\!\perp Z_j | Z_{\setminus\{i,j\}} \Leftrightarrow \boldsymbol{\Omega}_{ij} = 0.$$

Graphical interpretation: $p(\mathbf{Z}_i, \mathbf{Y}_i)$ vs $p(\mathbf{Y}_i)$



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:

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

Model selection

Alternative to model selection criteria

Sparsity level λ needs to be chosen.
Stability-based approach for Network by resampling: StARS

1. Infers B networks $\boldsymbol{\Omega}^{(b,\lambda)}$ 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 $\text{stab}(\lambda) = 1 - 2\bar{v}^\lambda$ where $\bar{v}^\lambda$ is the average of the v_e^λ.

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

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

An example in connection with the news

Data: first round of the French presidential election of 2017 (source: `https://data.gouv.fr`)

- votes cast for each of the 11 candidates in the more than 63, 000 polling stations
- voting population varied wildly
  From 10 to 105,891 , with a median at 736 and 99.5% of the stations with less than 1,700 voters.
- patterns depend on geography

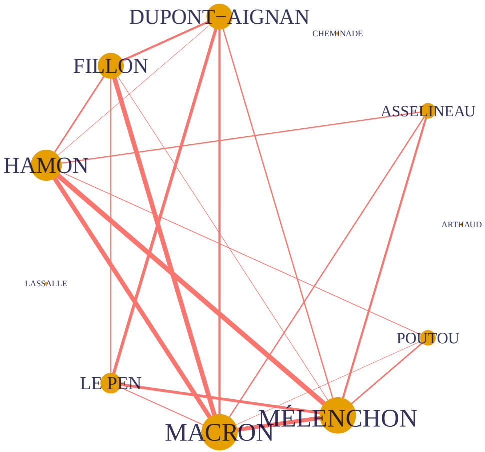
Models

- no offset
- offset: log-registered population of voters to account for different station sizes
- covariate: department as a proxy for geography.

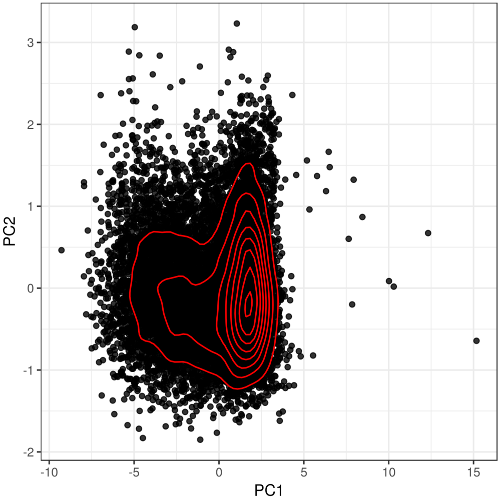
Question: find *competing* candidates, who appeal to different voters, and *compatible* candidates

French Presidential: no offset

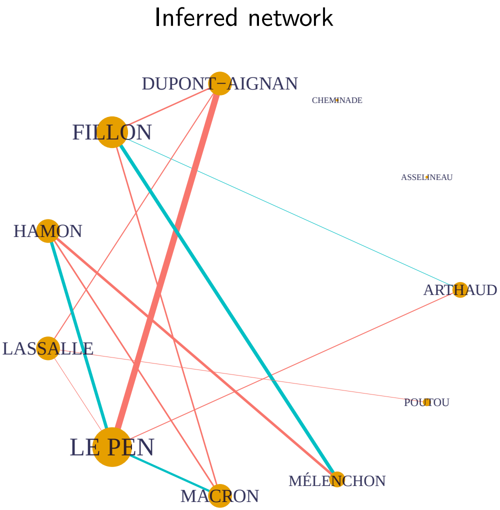
Inferred network

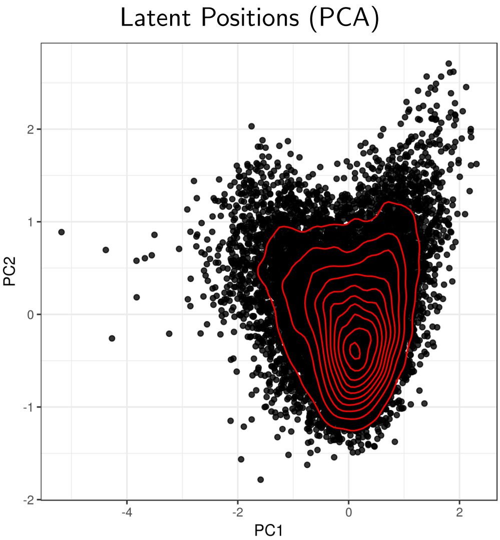


Latent Positions (PCA)



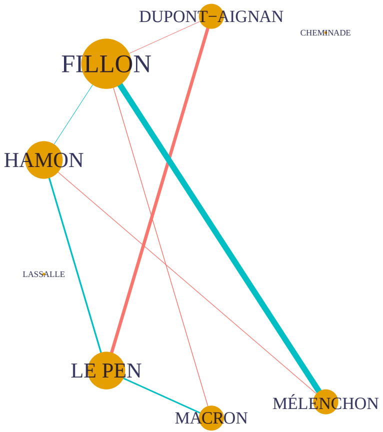
French Presidential: offset



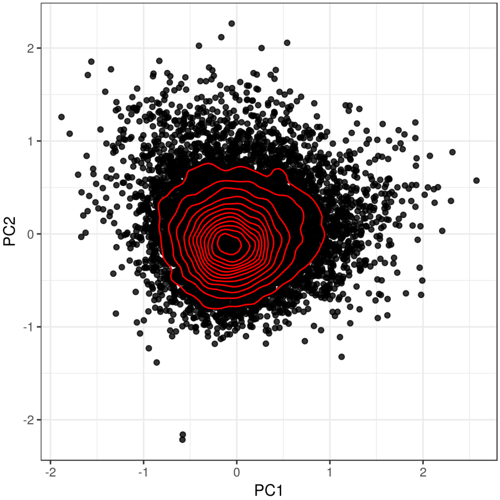


French Presidential: departments

Inferred network



Latent Positions (PCA)



More "conventional" example: Oak powdery mildew data set

Three setups

1. $n_r = 39$ resistant samples, with covariates (orientation, distance to ground)
2. $n_s = 39$ susceptible samples, with covariates (orientation, distance to ground)
3. both samples samples, with covariates + tree effect and interactions

Network inference

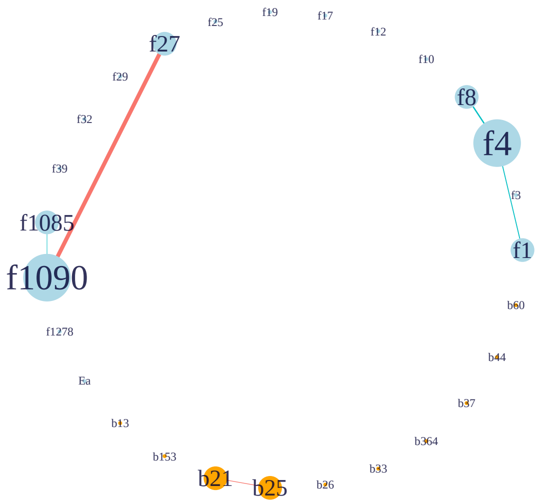
`PLNnetwork` + 'StARS' for model selection

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

Question: consensus or tree-specific networks?

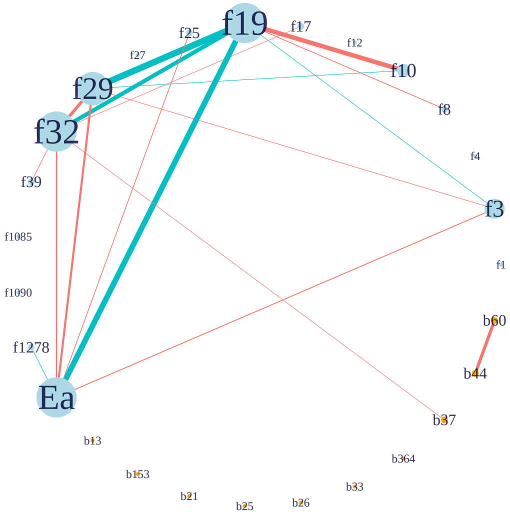
PLNnetwork models: resistant

Trees resistant to mildew (*E. Alphitoïdes*)



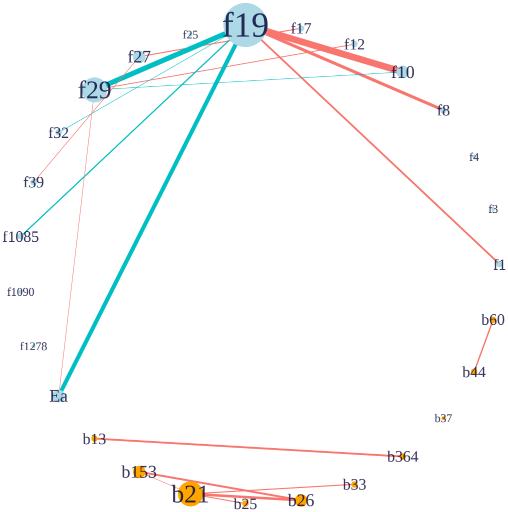
PLNnetwork models: susceptible

Trees susceptibles to mildew (*E. Alphitoïdes*)



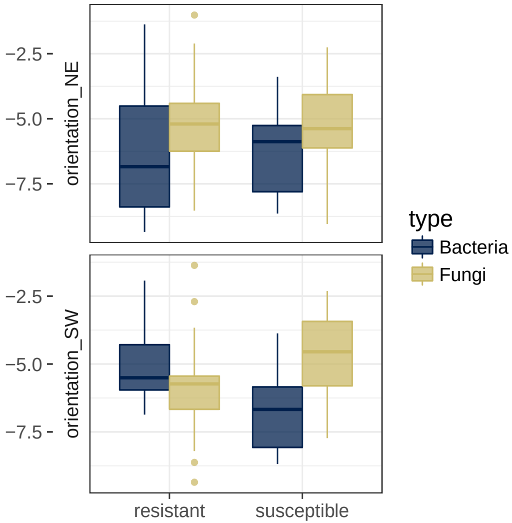
PLNnetwork models: consensus

Both Trees



PLNnetwork models: covariate effect

coefficients associated to orientation



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`

Ongoing extension...

- Confidence interval and tests for the regular PLN
- Other covariance structures (spatial, time series, ...), mixture models, ...
- Zero-Inflation

Following PLN Network Raphaëlle Momal's PhD (supervized by S. Robin and C. Ambroise)

- Tree-based decomposition of the underlying graphical model
- Other Model selection criterion for network inference

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.