PLNmodels

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

J. Chiquet, M. Mariadassou, S. Robin + B. Batardière, J. Kwon

Paris-Saclay, AgroParisTech, INRAE

Last update 16 September, 2021

https://pln-team.github.io/PLNmodels

R/C++ Package PLNmodels

Last stable release on CRAN, development version available on GitHub.

```
install.packages("PLNmodels")
devtools::install_github("PLN-team/PLNmodels")
```

```
library(PLNmodels)
packageVersion("PLNmodels")
```

```
## [1] '0.11.4'
```

Python module

A Python + PyTorch implementation is coming

Advertisement (more, sorry about that)

https://computo.sfds.asso.fr, a new journal promoting reproducible research

Resources

Help and documentation

The PLNmodels website contains the standard package documentation and a set of comprehensive vignettes for the top-level functions

Publications

Chiquet, J., M. Mariadassou, and S. Robin (2018). "Variational inference for probabilistic Poisson PCA". In: *The Annals of Applied Statistics* 12, pp. 2674-2698. URL: http://dx.doi.org/10.1214/18-AOAS1177.

Chiquet, J., M. Mariadassou, and S. Robin (2019). "Variational inference for sparse network reconstruction from count data". In: *Proceedings of the 19th International Conference on Machine Learning (ICML 2019)*.

Chiquet, J., M. Mariadassou, and S. Robin (2021). "The Poisson-Lognormal Model as a Versatile Framework for the Joint Analysis of Species Abundances". In: *Frontiers in Ecology and Evolution* 9. DOI: 10.3389/fevo.2021.588292.

Facon, B., A. Hafsi, M. C. de la Masselière, et al. (2021). "Joint species distributions reveal the combined effects of host plants, abiotic factors and species competition as drivers of species abundances in fruit flies". In: *Ecological Letters*. DOI: 10.1111/ele.13825.

Lejal, E., J. Chiquet, J. Aubert, et al. (2021). "Temporal patterns in Ixodes ricinus microbial communities: an insight into tick-borne microbe interactions". In: *Microbiome* 9.153. DOI: 10.1186/s40168-021-01051-8.

Generic form of data sets

Routinely gathered in ecology/microbiology/genomics

Data tables

- Abundances: read counts of species/transcripts j in sample i
- Covariates: value of environmental variable k in sample i
- Offsets: sampling effort for species/transcripts j in sample i

Need a framework to model *dependencies between counts*

- understand environmental effects
  ⤳ explanatory models (multivariate regression, classification)
- exhibit patterns of diversity
  ⤳ summarize the information (clustering, dimension reduction)
- understand between-species interactions
  ⤳ 'network' inference (variable/covariance selection)
- correct for technical and confounding effects
  ⤳ account for covariables and sampling effort

Models for multivariate count data

If we were in a Gaussian world…

The general linear model [MKB79] would be appropriate! For each sample $i = 1, \ldots, n,$

$$\underbrace{\mathbf{Y}_i}_{\text{abundances}} = \underbrace{\mathbf{x}_i^\top \mathbf{\Theta}}_{\text{covariates}} + \underbrace{\mathbf{o}_i}_{\text{sampling effort}} + \boldsymbol{\varepsilon}_i, \quad \boldsymbol{\varepsilon}_i \sim \mathcal{N}(\mathbf{0}_p, \underbrace{\mathbf{\Sigma}}_{\text{between-species dependencies}})$$

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

$\rightsquigarrow$ This model gives birth to Principal Component Analysis, Discriminant Analysis, Gaussian Graphical Models, Gaussian Mixture models and many others . . .

With count data…

There is no generic model for multivariate counts

- Data transformation (log, $\sqrt{\,}$) : quick and dirty
- Non-Gaussian multivariate distributions [Ino+17]: do not scale to data dimension yet
- Latent variable models: interaction occur in a latent (unobserved) layer

The Poisson Lognormal model (PLN)

The PLN model [AH89] is a multivariate generalized linear model, where

- the counts $\mathbf{Y}_i$ are the response variables
- the main effect is due to a linear combination of the covariates $\mathbf{x}_i$
- a vector of offsets $\mathbf{o}_i$ can be specified for each sample.

$$\mathbf{Y}_i | \mathbf{Z}_i \sim \mathcal{P}\left(\exp \mathbf{Z}_i\right), \qquad \mathbf{Z}_i \sim \mathcal{N}(\mathbf{o}_i + \mathbf{x}_i^\top \mathbf{\Theta}, \mathbf{\Sigma}),$$

The unkwown parameters are

- $\mathbf{\Theta}$, the regression parameters
- $\mathbf{\Sigma}$, the variance-covariance matrix

Stacking all individuals together,

- $\mathbf{Y}$ is the $n \times p$ matrix of counts
- $\mathbf{X}$ is the $n \times d$ matrix of design
- $\mathbf{O}$ is the $n \times p$ matrix of offsets

Properties: over-dispersion, arbitrary-signed covariances

- mean: $\mathbb{E}(Y_{ij}) = \exp\left(o_{ij} + \mathbf{x}_i^\top \mathbf{\Theta}_{\cdot j} + \sigma_{jj}/2\right) > 0$
- variance: $\mathbb{V}(Y_{ij}) = \mathbb{E}(Y_{ij}) + \mathbb{E}(Y_{ij})^2 \left(e^{\sigma_{jj}} - 1\right) > \mathbb{E}(Y_{ij})$
- covariance: $\mathrm{Cov}(Y_{ij}, Y_{ik}) = \mathbb{E}(Y_{ij})\mathbb{E}(Y_{ik})\left(e^{\sigma_{jk}} - 1\right).$

Natural extensions

Various tasks of multivariate analysis

- **Dimension Reduction**: rank constraint matrix $\boldsymbol{\Sigma}$.

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

- **Classification**: maximize separation between groups with means

$$\mathbf{Z}_i \sim \mathcal{N}(\boldsymbol{\mu}_k \mathbf{1}_{\{i \in k\}}, \boldsymbol{\Sigma}), \quad \text{for known memberships.}$$

- **Clustering**: mixture model in the latent space

$$\mathbf{Z}_i \mid i \in k \sim \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \quad \text{for unknown memberships.}$$

- **Network inference**: sparsity constraint on inverse covariance.

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

- **Variable selection**: sparsity constraint on regression coefficients

$$\mathbf{Z}_i \sim \mathcal{N}(\mathbf{x}_i^\top \boldsymbol{\Theta}, \boldsymbol{\Sigma}), \quad \|\boldsymbol{\Theta}\|_1 < c.$$

Estimate $\theta = (\mathbf{\Theta}, \mathbf{\Sigma})$, predict the $\mathbf{Z}_i$, while the model marginal likelihood is

$$p_\theta(\mathbf{Y}_i) = \int_{\mathbb{R}_p} \prod_{j=1}^{p} p_\theta(Y_{ij}|Z_{ij})\, p_\theta(\mathbf{Z}_i)\mathrm{d}\mathbf{Z}_i$$

Maximum likelihood for incomplete data model: EM

With $\mathcal{H}(p) = -\mathbb{E}_p(\log(p))$ the entropy of p,

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

EM requires to evaluate (some moments of) $p_\theta(\mathbf{Z}\,|\,\mathbf{Y})$, but there is no close form!

Solutions

- [AH89] resort on numerical integration; [Kar05] Monte-Carlo integration
- Several heuristics, not always well motivated, found in the literature...
- Variational approach [WJ08]: use a proxy of $p_\theta(\mathbf{Z}\,|\,\mathbf{Y})$.

Variational approximation

Principle

- Find a proxy of the conditional distribution $p(\mathbf{Z} \mid \mathbf{Y})$:

$$q(\mathbf{Z}) \approx p_\theta(\mathbf{Z}|\mathbf{Y}).$$

- Choose a convenient class of distribution $\mathcal{Q}$ and minimize a divergence

$$q(\mathbf{Z})^\star \arg\min_{q \in \mathcal{Q}} D\left(q(\mathbf{Z}), p(\mathbf{Z}|\mathbf{Y})\right).$$

Popular choice

The Küllback-Leibler divergence (error averaged wrt the approximated distribution)

$$KL\left(q(\mathbf{Z}), p(\mathbf{Z}|\mathbf{Y})\right) = \mathbb{E}_q\left[\log\frac{q(z)}{p(z)}\right] = \int_{\mathcal{Z}} q(z) \log\frac{q(z)}{p(z)} \mathrm{d}z.$$

Variational EM & PLN

Class of distribution: diagonal multivariate Gaussian

$$\mathcal{Q} = \left\{ q : \quad q(\mathbf{Z}) = \prod_i q_i(\mathbf{Z}_i), \quad q_i(\mathbf{Z}_i) = \mathcal{N}\left(\mathbf{Z}_i; \mathbf{m}_i, \mathrm{diag}(\mathbf{s}_i \circ \mathbf{s}_i)\right), \mathbf{m}_i, \mathbf{s}_i \in \mathbb{R}_p \right\}$$

Maximize the ELBO (Evidence Lower BOund):

$$J(\theta, q) = \log p_\theta(\mathbf{Y}) - KL[q_\theta(\mathbf{Z})||p_\theta(\mathbf{Z}|\mathbf{Y})] = \mathbb{E}_q[\log p_\theta(\mathbf{Y}, \mathbf{Z})] + \mathcal{H}[q(\mathbf{Z})]$$

Variational EM

- VE step: find the optimal q (here, $\{(\mathbf{m}_i, \mathbf{s}_i)\}_{i=1,\ldots,n} = \{\mathbf{M}, \mathbf{S}\}$):

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

- M step: update $\hat{\theta}^h$

$$\theta^h = \arg\max J(\theta, q^h) = \arg\max_\theta \mathbb{E}_q[\log p_\theta(\mathbf{Y}, \mathbf{Z})]$$

ELBO and gradients for PLN

Let $\mathbf{A} = \mathbb{E}_q[\exp(\mathbf{Z})] = \exp\left(\mathbf{O} + \mathbf{M} + \frac{1}{2}\mathbf{S}^2\right)$

Variational bound

$$
\begin{aligned}
J(\mathbf{Y}) \quad &= \mathbf{1}_n^\mathsf{T} \left(\left[\mathbf{Y} \circ (\mathbf{O} + \mathbf{M}) - \mathbf{A} + \log(\mathbf{S})\right]\right) \mathbf{1}_p + \frac{n}{2}\log|\mathbf{\Omega}| \\
&\quad - \frac{1}{2}\mathrm{trace}\left(\mathbf{\Omega}\left[(\mathbf{M} - \mathbf{X}\mathbf{\Theta})^\mathsf{T}(\mathbf{M} - \mathbf{X}\mathbf{\Theta}) + \mathrm{diag}(\mathbf{1}_n^\mathsf{T}\mathbf{S}^2)\right]\right) + \mathrm{cst}.
\end{aligned}
$$

M-step (Analytical)

$$
\hat{\mathbf{\Theta}} = \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}\mathbf{M}, \quad \hat{\mathbf{\Sigma}} = \frac{1}{n}\left(\mathbf{M} - \mathbf{X}\hat{\mathbf{\Theta}}\right)^\top\left(\mathbf{M} - \mathbf{X}\hat{\mathbf{\Theta}}\right) + \frac{1}{n}\mathrm{diag}(\mathbf{1}^\mathsf{T}\mathbf{S}^2)
$$

Variational E-step (optimization)

$$
\frac{\partial J(q)}{\partial \mathbf{M}} = (\mathbf{Y} - \mathbf{A} - (\mathbf{M} - \mathbf{X}\mathbf{\Theta})\mathbf{\Omega}), \qquad \frac{\partial J(q)}{\partial \mathbf{S}} = \frac{1}{\mathbf{S}} - \mathbf{S} \circ \mathbf{A} - \mathbf{S}\mathbf{D}_{\mathbf{\Omega}}.
$$

Application to the optimization of PLN models

Property of PLN variational approximation

The ELBO $J(\theta, q)$ is bi-concave, i.e.

- concave wrt $q = (\mathbf{M}, \mathbf{S})$ for given θ
- convace wrt $\theta = (\boldsymbol{\Sigma}, \boldsymbol{\Theta})$ for given q but not jointly concave in general.

Optimization

Gradient ascent for the set of variaional parameters $(\mathbf{M}, \mathbf{S})$

Medium scale problems

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

⤳ Comfortable up to a thousand of sites ($n \approx 1000$), hundreds of species ($p \approx 100s$)

Large scale problems

Sophisticated Adaptive Stochastic Gradient Descent

- Rprop (1993) uses the gradient sign and update each variable independently:

- AdaGrad (2011) uses adaptive coordinate-wise step-sizes

- RMSProp (2012) adds momentum to the step-sizes

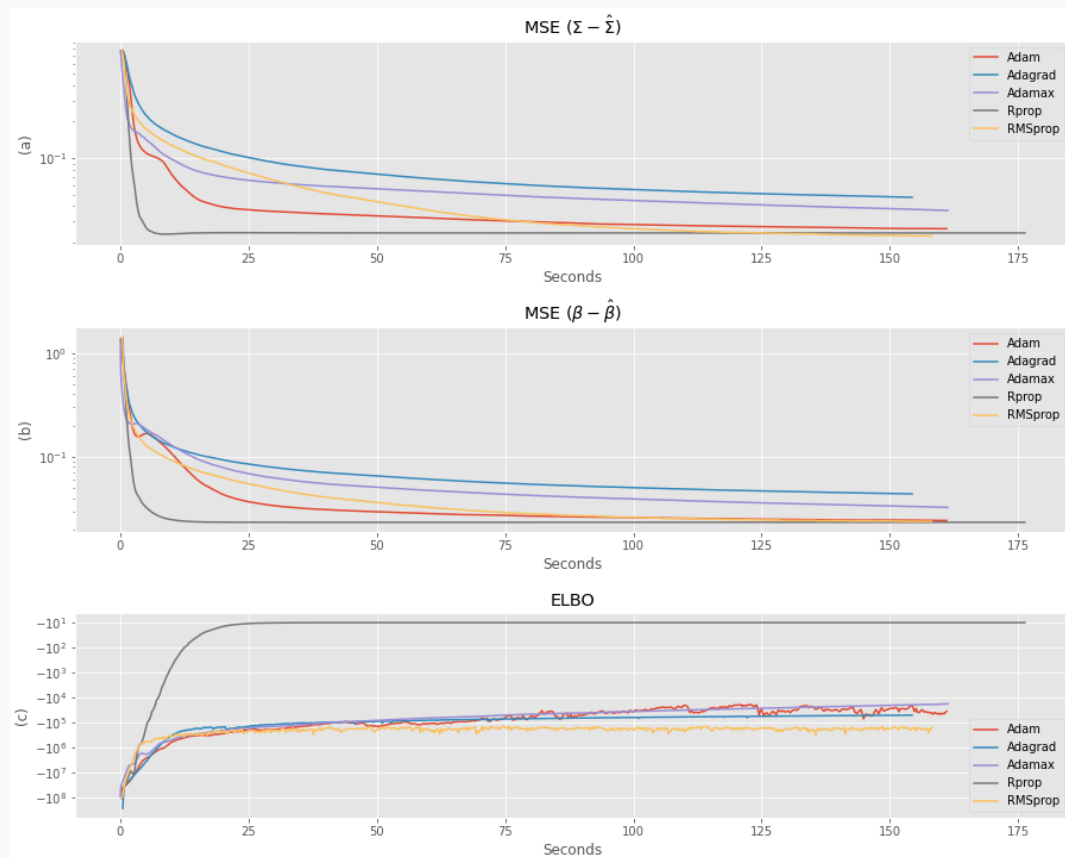- Adam (2015) also adds momentum to the gradients
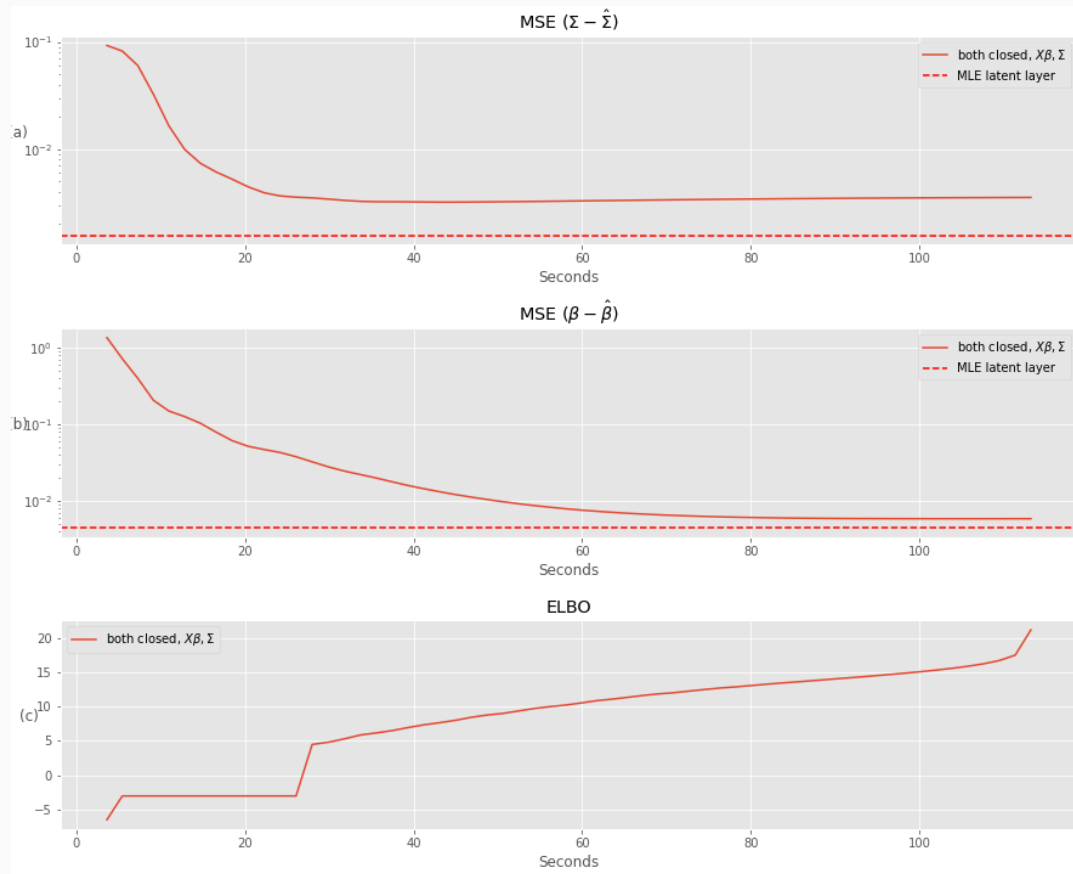

(a) SGD without momentum   (b) SGD with momentum
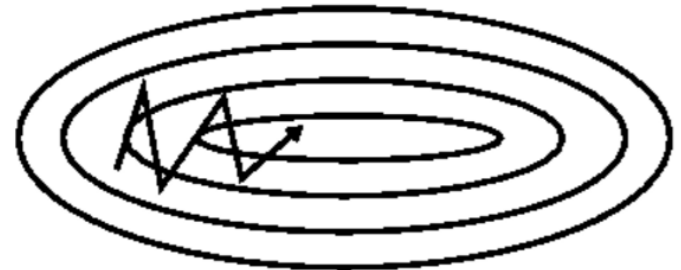
⤳ All available in **Pytorch** with auto-differentiation.

Optimizers comparison

n = 1,000, p = 200, d = 2. Rprop is much faster.

Performance

n = 10,000, p = 2,000, d = 2 (running time: 1 min 40s)

⤳ Work up to $n = 100,000, p = 10,000s$ (useful for microbiological data, single cell data, now found in ecology)

Statistical Inference

A first try: Wald test

Test $\mathcal{H}_0 : R\theta = r_0$ with the statistic

$$(R\hat{\theta} - r_0)^\top \left[nR\hat{\mathbb{V}}(\hat{\theta})R^\top\right]^{-1}(R\hat{\theta} - r_0) \sim \chi_k^2 \quad \text{where} \quad k = \text{rank}(R).$$

The Fisher Information matrix

$$I(\hat{\theta}) = -\mathbb{E}_\theta \left[\frac{\partial^2 \log \ell(\theta; x)}{\partial \theta^2}\right]$$

can be used as an approximation of $n\mathbb{V}(\hat{\theta})^{-1}$.

Application

Derive confidences interval for the inverse covariance $\boldsymbol{\Omega}$ and the regression parameters $\boldsymbol{\Theta}$.

Variational Wald-test

Variational Fisher Information

The Fisher information matrix is given by

$$I(\theta) = \begin{pmatrix} \frac{1}{n}(\mathbf{I}_p \otimes \mathbf{X}^\top)\text{diag}(\text{vec}(\mathbf{A}))(\mathbf{I}_p \otimes \mathbf{X}) & \mathbf{0} \\ \mathbf{0} & \frac{1}{2}\mathbf{\Omega}^{-1} \otimes \mathbf{\Omega}^{-1} \end{pmatrix}$$

and can be inverted blockwise to estimate $\mathbb{V}(\hat{\theta})$.

Wald test and coverage

- $\hat{\mathbb{V}}(\mathbf{\Theta}_{kj}) = [n(\mathbf{X}^\top \text{diag}(\text{vec}(\hat{A}_{\cdot j}))\mathbf{X})^{-1}]_{kk}$
- $\hat{\mathbb{V}}(\Omega_{kl}) = 2\Omega_{kk}\Omega_{ll}$

The confidence intervals at level α are given by

- $B_{kj} = \hat{B}_{kj} \pm \frac{q_{1-\alpha/2}}{\sqrt{n}}\sqrt{\hat{\mathbb{V}}(\mathbf{\Theta}_{kj})}$
- $\Omega_{kl} = \hat{\Omega}_{kl} \pm \frac{q_{1-\alpha/2}}{\sqrt{n}}\sqrt{\hat{\mathbb{V}}(\Omega_{kl})}.$

Numerical study

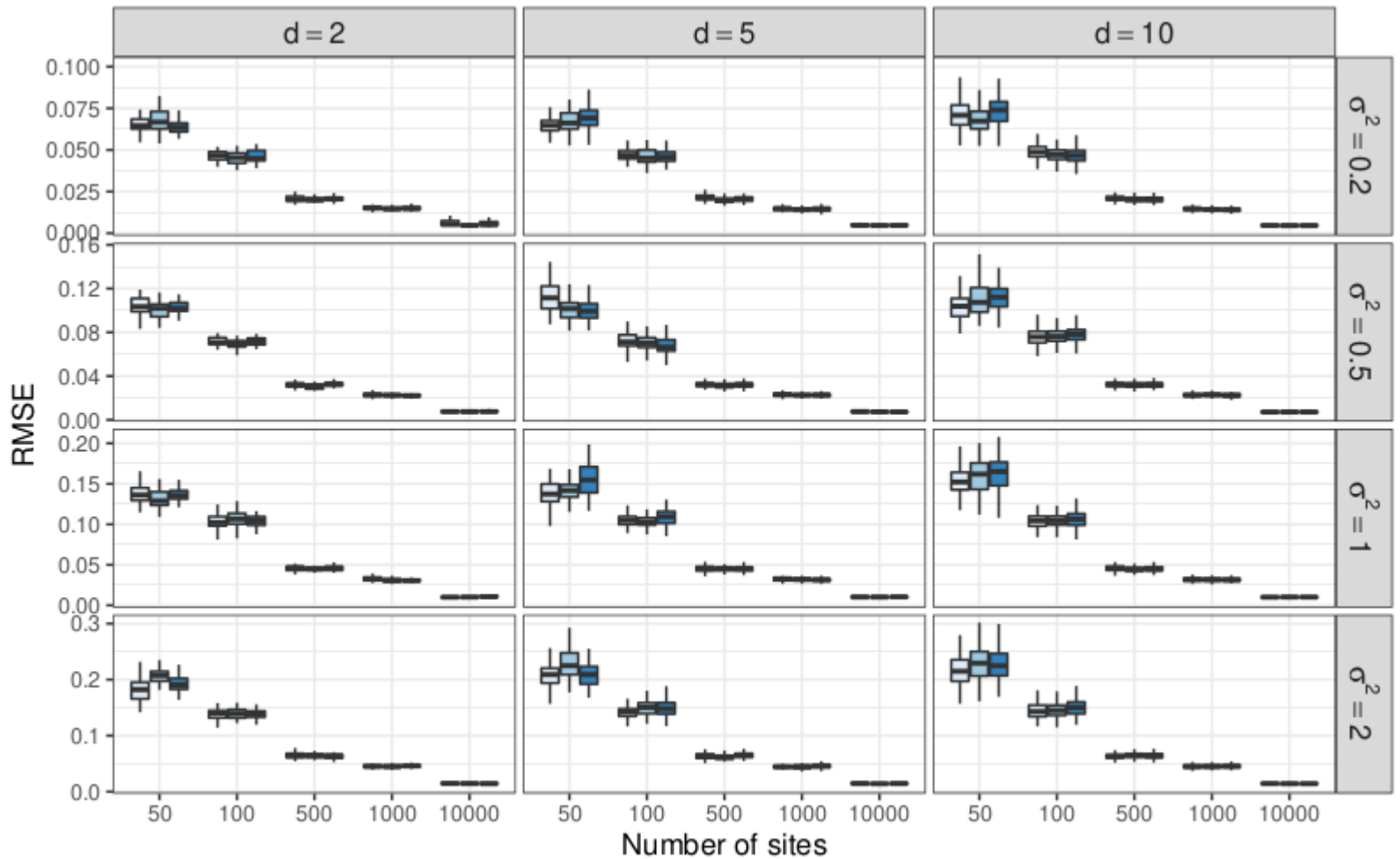## Study Bias and coverage of the estimator

- number of samples $n \in \{50, 100, 500, 1000, 10000\}$

- number of species/genes $p \in \{20, 200\}$

- number of covariates $d \in \{2, 5, 10\}$

- sampling effort $TSS \in \{\text{low}, \text{medium}, \text{high}\}$ (\$\approx 10^4\$, 10^5 and 10^6)

- noise level $\sigma^2 \in \{0.2, 0.5, 1, 2\}$

- $\boldsymbol{\Sigma}$ as $\sigma_{jk} = \sigma^2 \rho^{|j-k|}$, with $\rho = 0.2$

- $\boldsymbol{\Theta}$ with entries sampled from $\mathcal{N}(0, 1/d)$

Bias of $\hat{\Theta}$
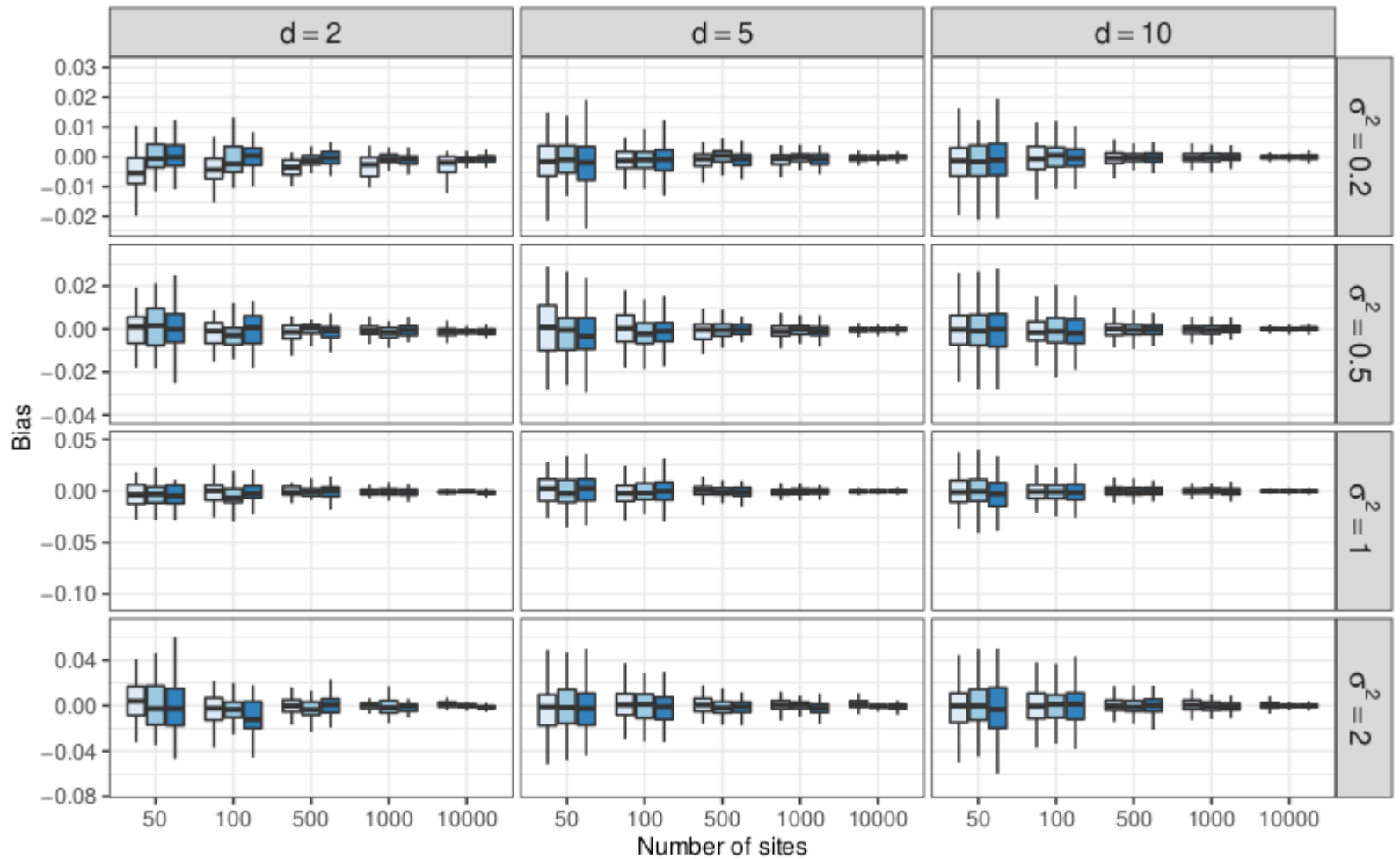
⤳ asymptotically unbiased

Root mean square error of $\hat{\Theta}$

⤳ asymptotically unbiased

95% confident interval - Coverage
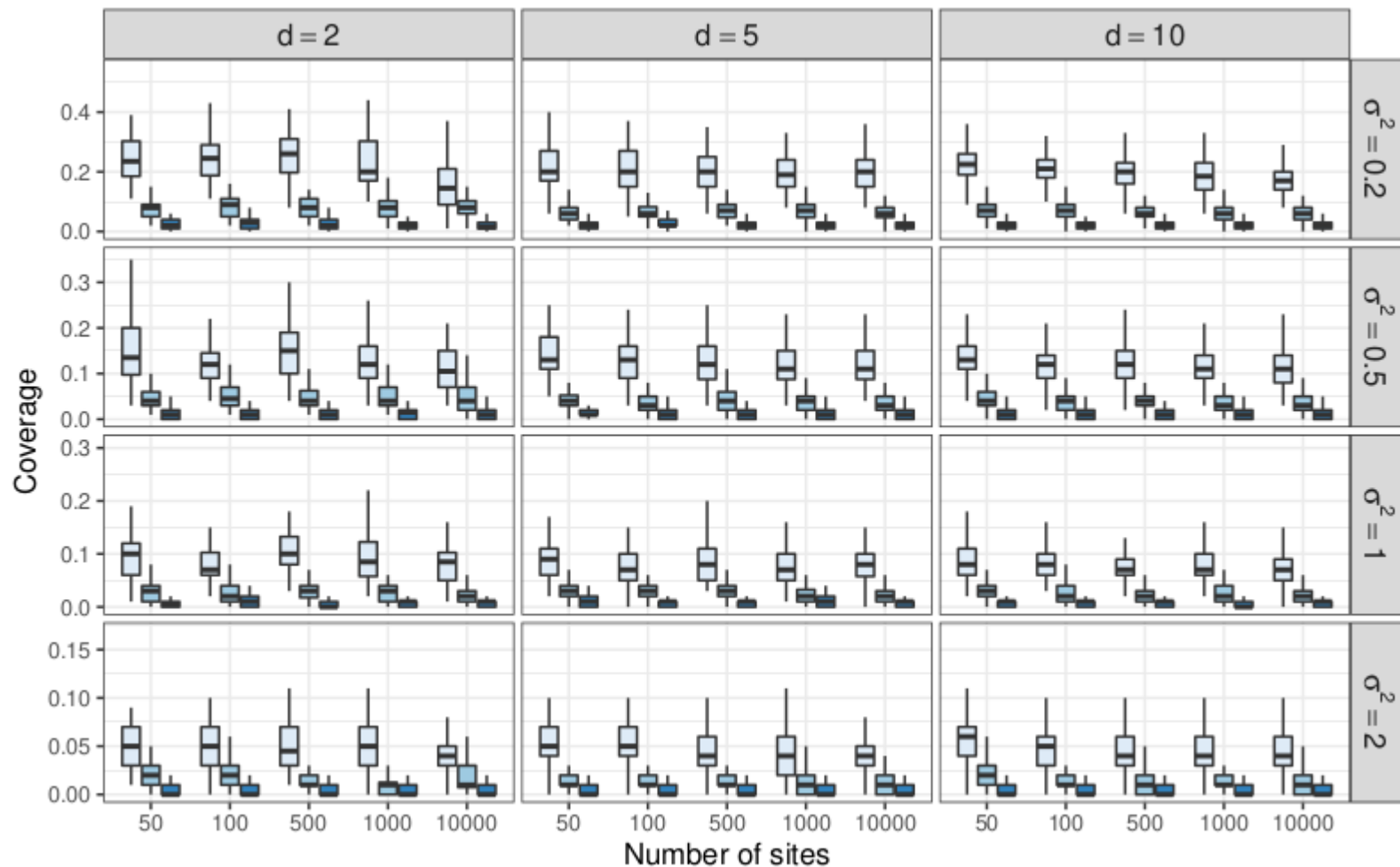
variance underestimated, no trusted confidence intervals can be derived out-of-the box

Other ideas

M-estimation (with M. Mariadassou)

We can derive asymptotic behavior but V-EM stationary point is <mark>not a log-likelihood stationary point</mark>, see [WM15] ⤳ Use sandwich estimator, correction.

Exact likelihood maximization (with B. Batardière, J. Kwon)

To compare and assess at least empirically the performance of the VE-M estimator

- Use importance sampling to estimate the likelihood:

$$p_\theta(Y_i) = \int \tilde{p}_\theta(Y_i|Z)p(Z)\mathrm{d}Z = \int \tilde{p}_\theta(Z)\mathrm{d}Z \approx \frac{1}{n_s}\sum_{k=1}^{n_s}\frac{\tilde{p}_\theta(V_k)}{g(V_k)}, \quad (V_k)_{1\le k\le n_s} \overset{iid}{\sim} g$$

- Estimate the gradients of the log-likelihood by plug-in:

$$\nabla_\theta \log p_\theta(Y_i) \approx \nabla_\theta \log\left(\frac{1}{n_s}\sum_{k=1}^{n_s}\frac{\tilde{p}_\theta^{(u)}(V_k)}{g(V_k)}\right)$$

Oaks powdery mildew data set overview

Study <mark>effects of the pathogen</mark> *E.Aphiltoïdes* (mildew) wrt bacterial and microbial communities

Species Abundances

- <mark>Microbial communities</mark> sampled on the surface of $n = 116$ oak leaves
- Communities sequenced and cleaned resulting in $p = 114$ OTUs (66 bacteria, 48 fungi).

Covariates and offsets

Characterize the samples and the sampling, most important being

- `tree` : Tree status with respect to the pathogen (susceptible, intermediate or resistant)
- `distTOground` : Distance of the sampled leaf to the base of the ground
- `orientation` : Orientation of the branch (South-West SW or North-East NE)
- `readsTOTfun` : Total number of ITS1 reads for that leaf
- `readsTOTbac` : Total number of 16S reads for that leaf

```
data(oaks)
oaks$Abundance %>% as_tibble() %>%
  dplyr::select(1:10) %>%
  head() %>% DT::datatable(fillContainer = FALSE)
```

Show 10 entries   Search:

	b_OTU_1045	b_OTU_109	b_OTU_1093	b_OTU_11	b_OTU_112	b_OTU_1191
1	0	0	0	6	146	1
2	0	0	0	0	0	1
3	0	0	0	2	0	0
4	0	0	0	1	1	1
5	0	0	0	4	1	1
6	0	0	0	77	2	20

Showing 1 to 6 of 6 entries   Previous   1   Next

Abundance table (II)

```r
log(1 + oaks$Abundance) %>%
  corrplot::corrplot(is.corr = FALSE,
    addgrid.col = NA,  tl.cex = .5,  cl.pos = "n")
```
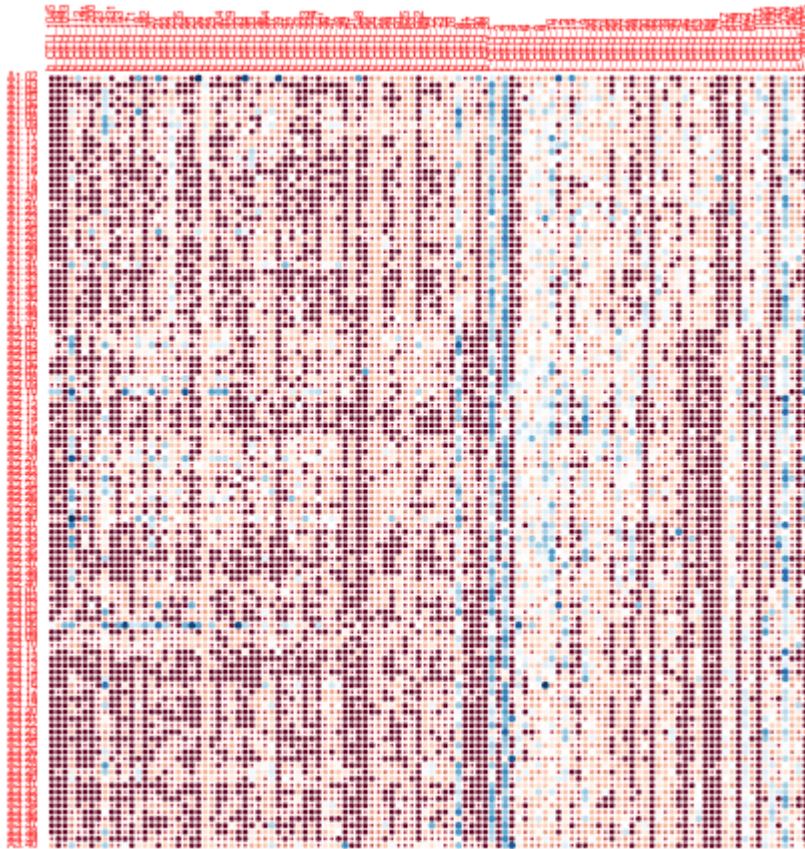
PLN with offsets and covariates (1)

Offset: modeling sampling effort

The predefined offset uses the total sum of reads, accounting for technologies specific to fungi and bacteria:

```
M01_oaks ← PLN(Abundance ~ 1 + offset(log(Offset)) , oaks)
```

Covariates: tree and orientation effects ('ANOVA'-like)

The `tree` status is a natural candidate for explaining a part of the variance.

- We chose to describe the tree effect in the regression coefficient (mean)
- A possibly spurious effect regarding the interactions between species (covariance).

```
M11_oaks ← PLN(Abundance ~ 0 + tree + offset(log(Offset)), oaks)
```

What about adding more covariates in the model, e.g. the orientation?

```
M21_oaks ← PLN(Abundance ~  0 + tree + orientation + offset(log(Offset)), oaks)
```

PLN with offsets and covariates (2)

There is a clear gain in introducing the tree covariate in the model:

```
rbind(M01 = M01_oaks$criteria,
      M11 = M11_oaks$criteria, M21 = M21_oaks$criteria) %>%
  knitr::kable(format = "html")
```

	nb_param	loglik	BIC	ICL
M01	6669	-32252.14	-48102.98	-52169.64
M11	6897	-31524.16	-47916.91	-51644.03
M21	7011	-31438.58	-48102.29	-51727.13

Looking at the coefficients Θ associated with `tree` bring additional insights:
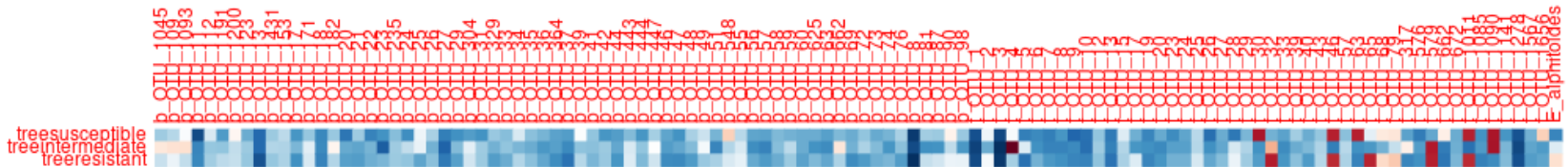
Discriminant Analysis

Use the `tree` variable for grouping (`grouping` is a factor of group to be considered)

```
myLDA_tree ←
    PLNLDA(Abundance ~ 1 + offset(log(Offset)), grouping = tree, data = oaks)
```
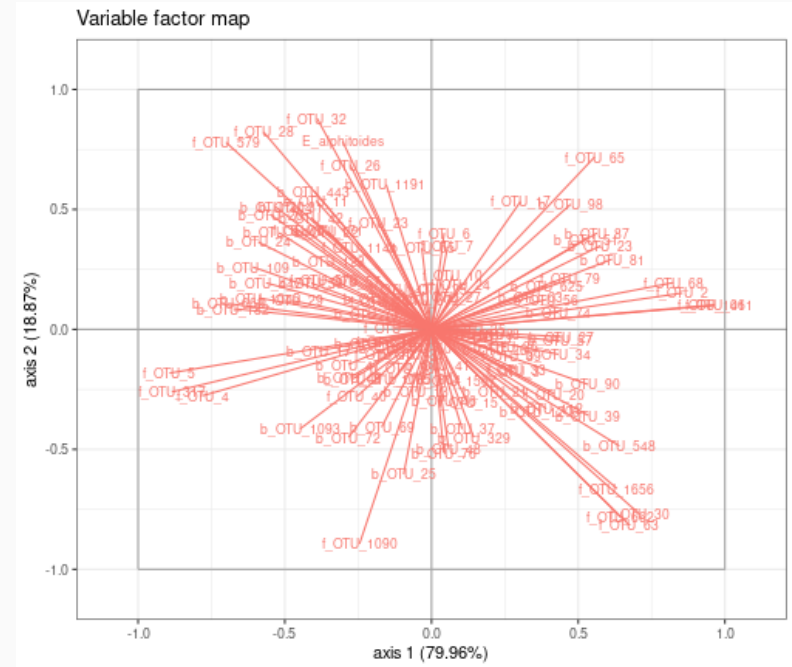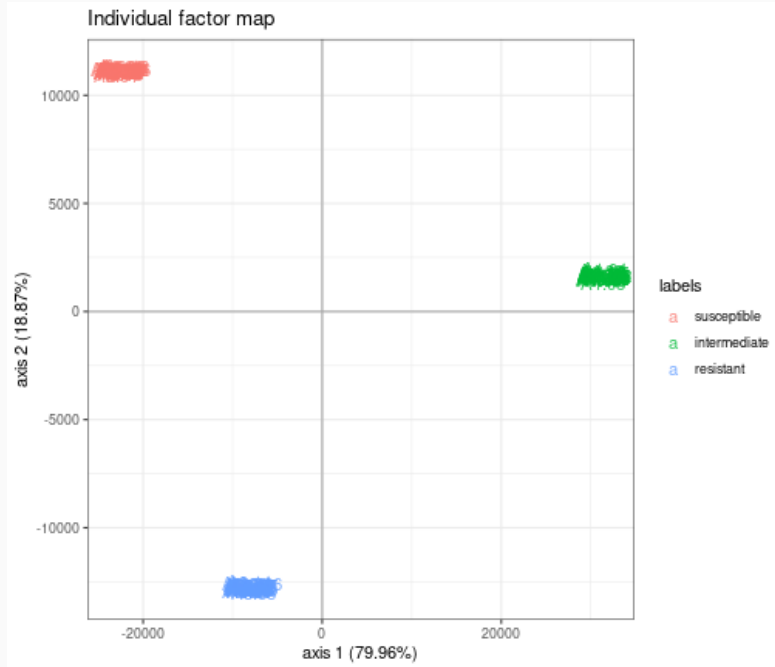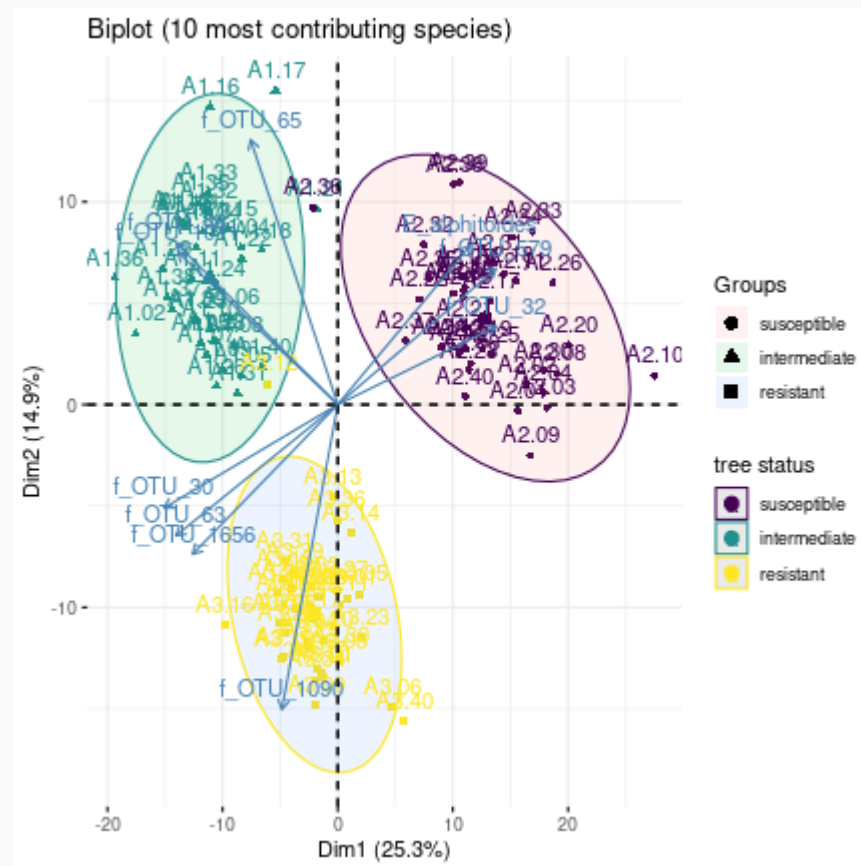
A PCA analysis of the oaks data set

```
PCA_offset ←
  PLNPCA(Abundance ~ 1 + offset(log(Offset)), data = oaks, ranks = 1:30)
PCA_offset_BIC ← getBestModel(PCA_offset, "BIC")
```


Biplot (10 most contributing species)

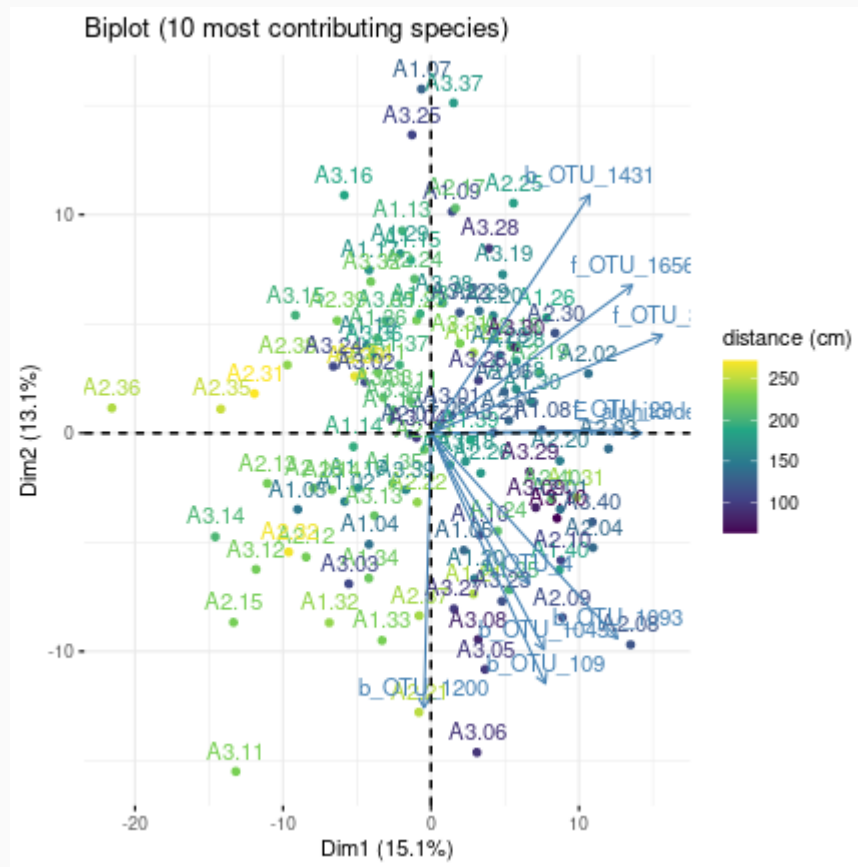# PCA: removing covariate effects

To hopefully find some hidden effects in the data, we can try to remove confounding ones:

```
PCA_tree ←
   PLNPCA(Abundance ~ 0 + tree + offset(log(Offset)), data = oaks, ranks = 1:30)
```


Biplot (10 most contributing species)
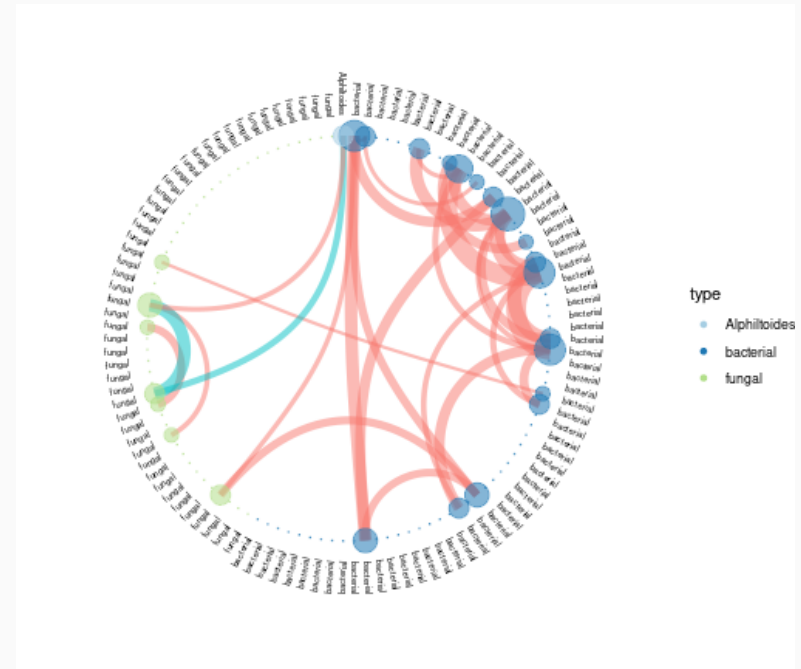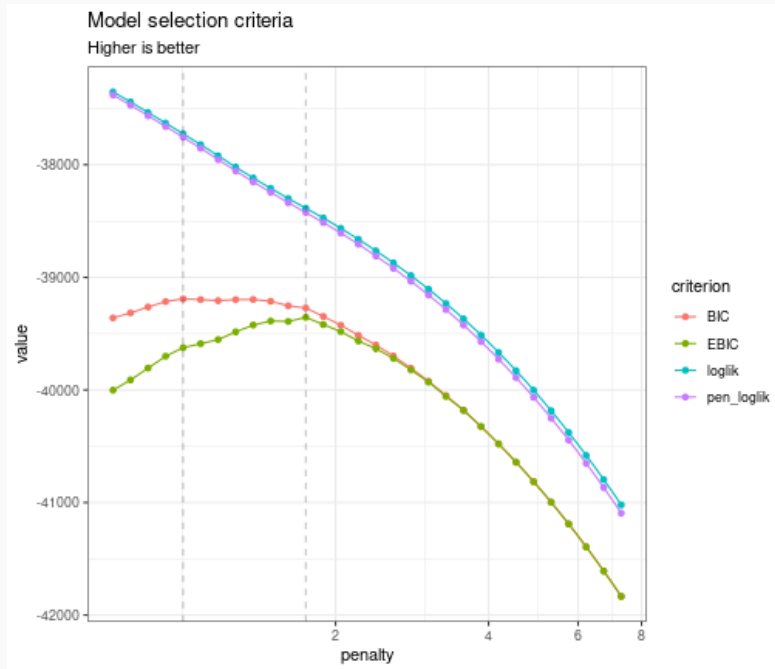
Clustering of the oaks samples

```
PLN_mixtures ←
    PLNmixture(Abundance ~ 1 + offset(log(Offset)), data = oaks, clusters = 1:3)
myPLN_mix ← getModel(PLN_mixtures, 3)
```

```
myPLN_mix$plot_clustering_pca()                    myPLN_mix$plot_clustering_data()
```

Network inference

```
networks ← PLNnetwork(Abundance ~ 0 + tree + offset(log(Offset)), data = oaks)
```

Conclusion

Summary

- PLN = generic model for multivariate count data analysis
- Flexible modeling of the covariance structure, allows for covariates
- Efficient V-EM algorithm

Extensions

- Other variants
  - zero inflation (data with *a lot* of zeros)
  - covariance structures (spatial, time series, …)
  - Variable selection (ℓ_1-penalty on the regression coefficients)
- Other models
  - Bernoulli/multinomial counterpart to PLN
  - functional data
  - multiple-data integration (e.g., Bernoulli + Poisson)

References

Aitchison, J. and C. Ho (1989). "The multivariate Poisson-log normal distribution". In: *Biometrika* 76.4, pp. 643-653.

Chiquet, J., M. Mariadassou, and S. Robin (2018). "Variational inference for probabilistic Poisson PCA". In: *The Annals of Applied Statistics* 12, pp. 2674-2698. URL: http://dx.doi.org/10.1214/18-AOAS1177.

Chiquet, J., M. Mariadassou, and S. Robin (2019). "Variational inference for sparse network reconstruction from count data". In: *Proceedings of the 19th International Conference on Machine Learning (ICML 2019)*.

Chiquet, J., M. Mariadassou, and S. Robin (2021). "The Poisson-Lognormal Model as a Versatile Framework for the Joint Analysis of Species Abundances". In: *Frontiers in Ecology and Evolution* 9. DOI: 10.3389/fevo.2021.588292.

Facon, B., A. Hafsi, M. C. de la Masselière, et al. (2021). "Joint species distributions reveal the combined effects of host plants, abiotic factors and species competition as drivers of species abundances in fruit flies". In: *Ecological Letters*. DOI: 10.1111/ele.13825.