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 01 December, 2021

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

Reproducibility

R/C++ Package PLNmodels

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

```r
install.packages("PLNmodels")
remotes::install_github("PLN-team/PLNmodels@dev")
```

```r
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.

PLNmodels: what is done[1]

1. Motivations
2. A family of models for multivariate analysis
3. Efficient variational inference
4. Illustration

[1] and published

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 frameworks to model *dependencies between counts*

- understand **environmental effects**
  $\rightsquigarrow$ explanatory models (multivariate regression, classification)
- exhibit **patterns of diversity**
  $\rightsquigarrow$ summarize the information (clustering, dimension reduction)
- 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 [MKB79] would be appropriate! For each sample $i = 1, \dots, n$,

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

null covariance ⇔ independence ⇝ uncorrelated species/transcripts do not interact

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

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: $\mathbb{E}(Y_{ij}) = \exp\big(o_{ij} + \mathbf{x}_i^\top \mathbf{\Theta}_{\cdot j} + \sigma_{jj}/2\big) > 0$
- 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 $\mathbf{\Sigma}$.

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

- Classification: maximize separation between groups with means

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

- Clustering: mixture model in the latent space

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

- Network inference: sparsity constraint on inverse covariance.

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

- Variable selection: sparsity constraint on regression coefficients

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

Inference: latent model but intractable EM

Estimate $\theta = (\boldsymbol{\Theta}, \boldsymbol{\Sigma})$, 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, $\{(\mathbf{m}_i, \mathbf{s}_i)\}_{i=1,\dots,n} = \{\mathbf{M}, \mathbf{S}\}$):

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

- 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

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

M-step (Analytical)

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

Variational E-step (optimization)

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

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 $\theta = (\mathbf{\Sigma}, \mathbf{\Theta})$ for given q but not jointly concave in general.

Optimization

Gradient ascent for the set of variational 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-transformation 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 ($p \approx 100s$)

Oaks powdery mildew data set overview

Jakuschkin, Fievet, Schwaller, Fort, Robin, and Vacher [Jak+16] Study effects of the pathogen *E.Aphiltoïdes* (mildew) wrt bacterial and microbial communities

Species Abundances

- Microbial communities 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

Abundance table (I)

```
data(oaks)
oaks$Abundance %>% as_tibble() %>%
  dplyr::select(1:10) %>%
  rmarkdown::paged_table()
```

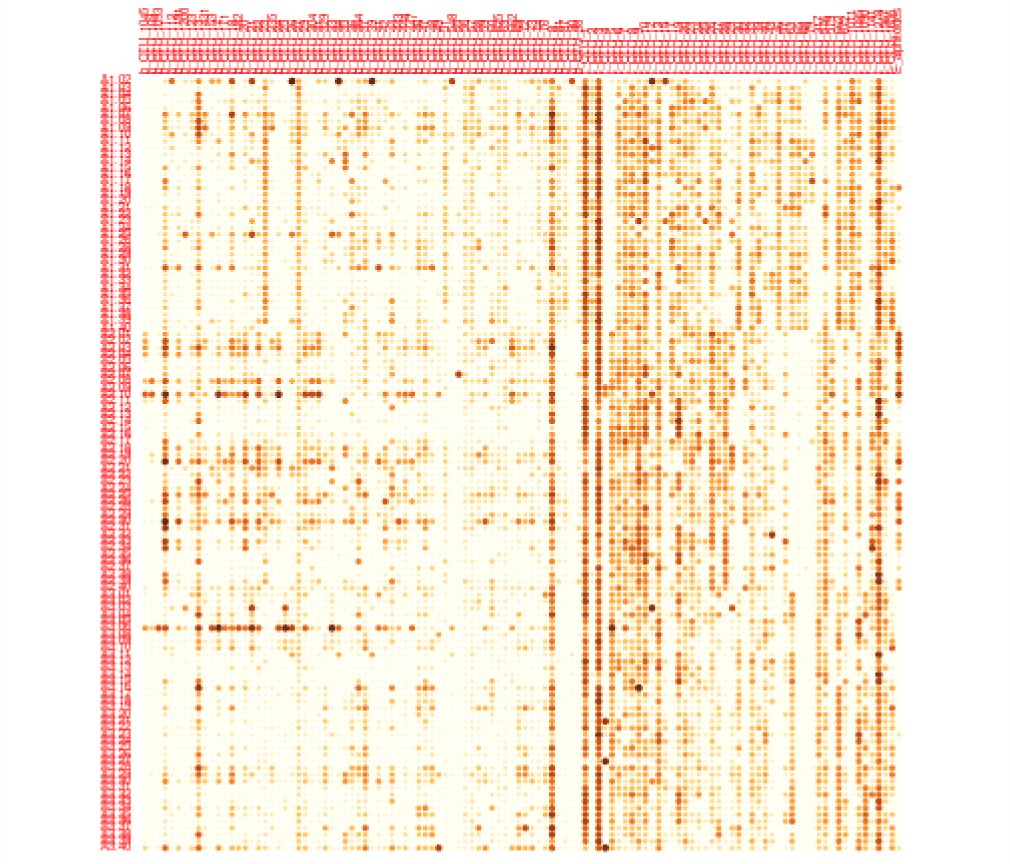
b_OTU_1045	b_OTU_109	b_OTU_1093	b_OTU_11	b_OTU_112
<int>	<int>	<int>	<int>	<int>
0	0	0	6	146
0	0	0	0	0
0	0	0	2	0
0	0	0	1	1
0	0	0	4	1
0	0	0	77	2
0	0	0	21	2
0	0	0	27	4
0	0	0	7	42
0	0	0	7	2

1-10 of 116 rows | 1-5 of 10 columns

Previous **1** 2 3 4 5 6 ... 12 Next

Abundance table (II)

```
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)
```

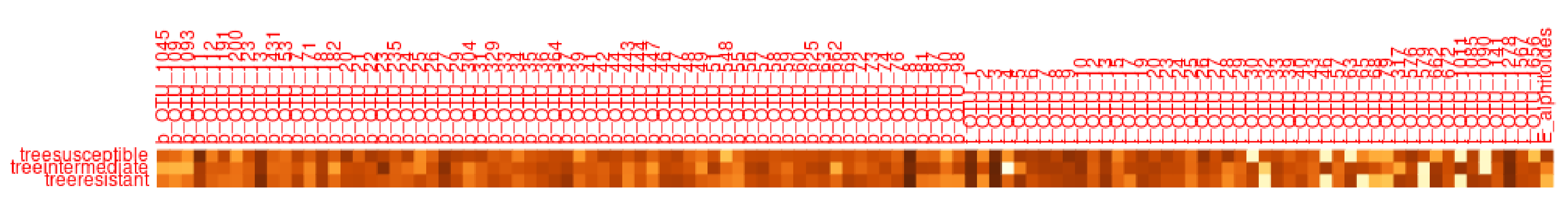
PLN with offsets and covariates (2)

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

```r
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

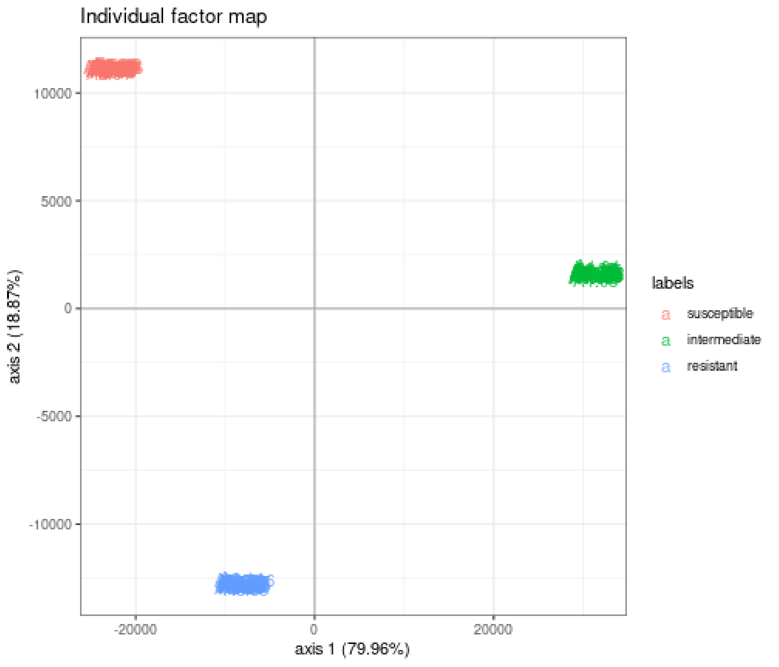
Looking at the coefficients $\mathbf{\Theta}$ associated with `tree` bring additional insights:

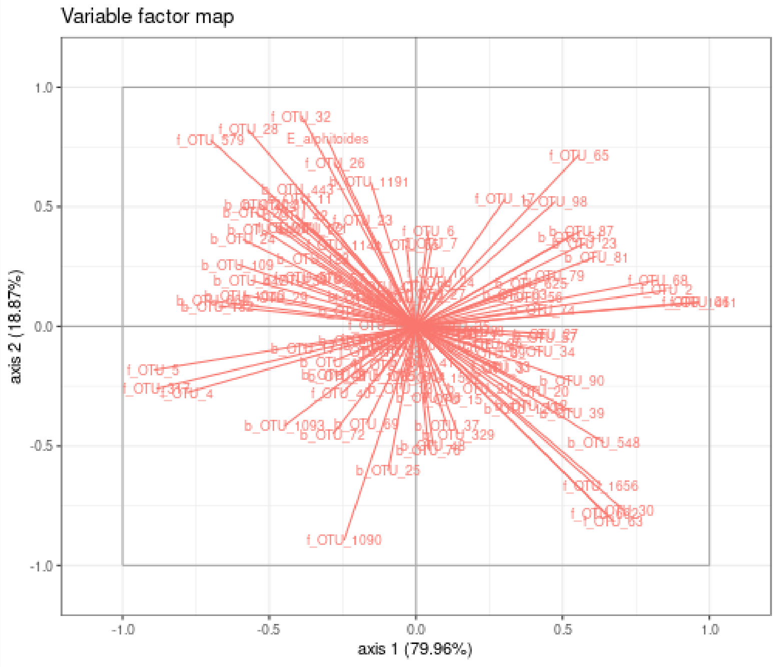


Discriminant Analysis

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

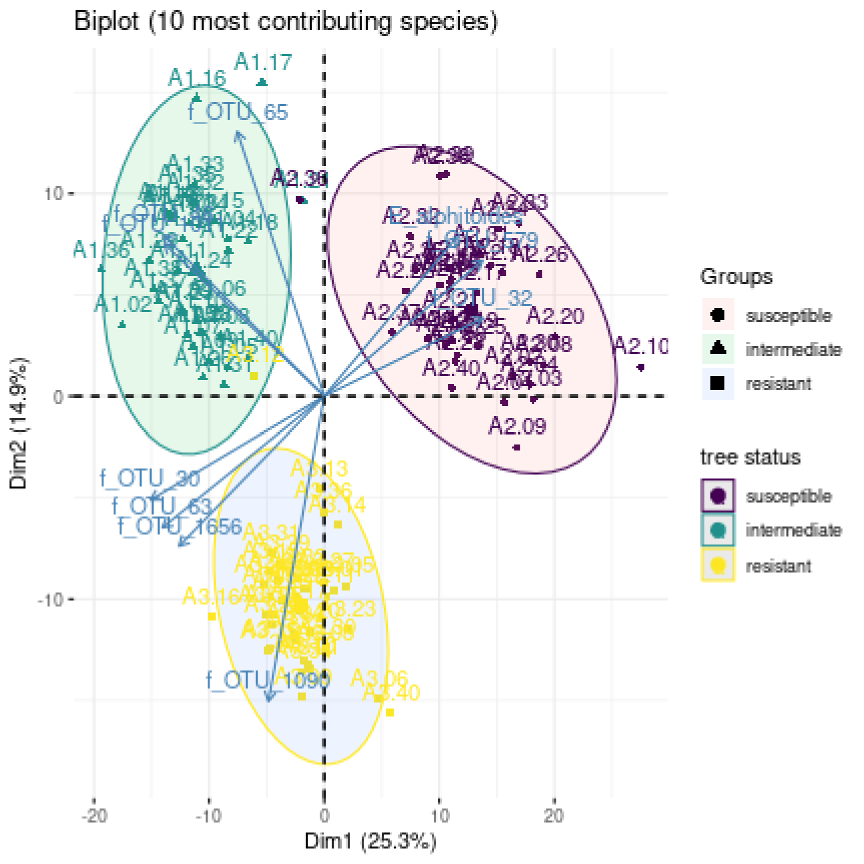
```
myLDA_tree ←
  PLNLDA(Abundance ~ 1 + offset(log(Offset)), grouping = oaks$tree, data = oaks)
```





A PCA analysis of the oaks data set

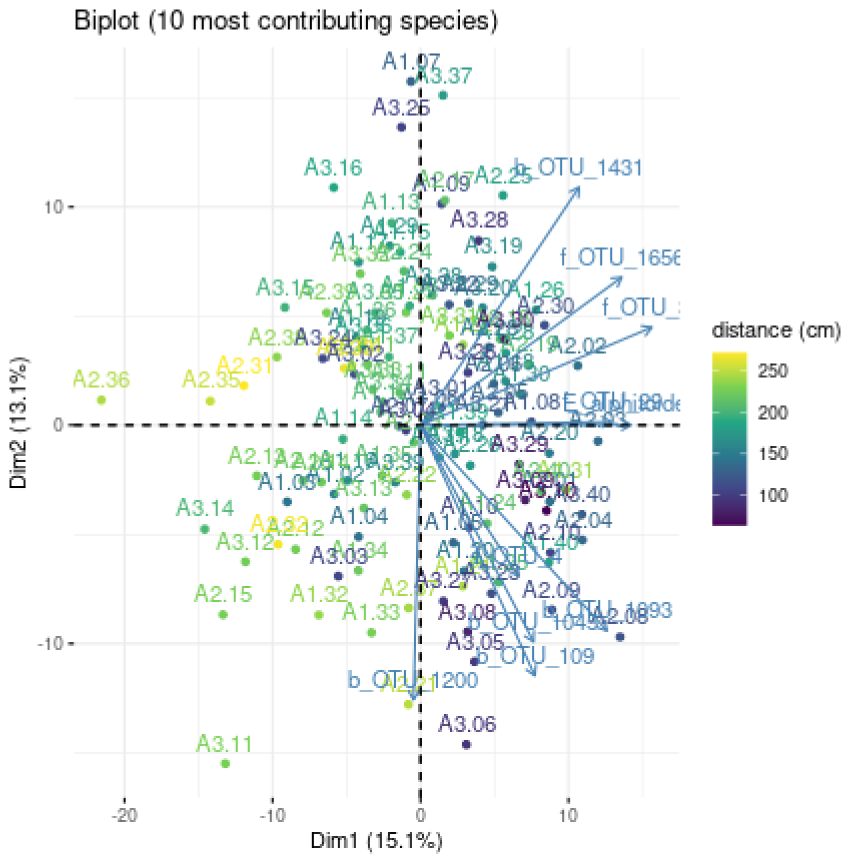
```
PCA_offset <-
  PLNPCA(Abundance ~ 1 + offset(log(Offset)), data = oaks, ranks = 1:30)
PCA_offset_BIC <- getBestModel(PCA_offset, "BIC")
```



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)
```



Clustering of the oaks samples

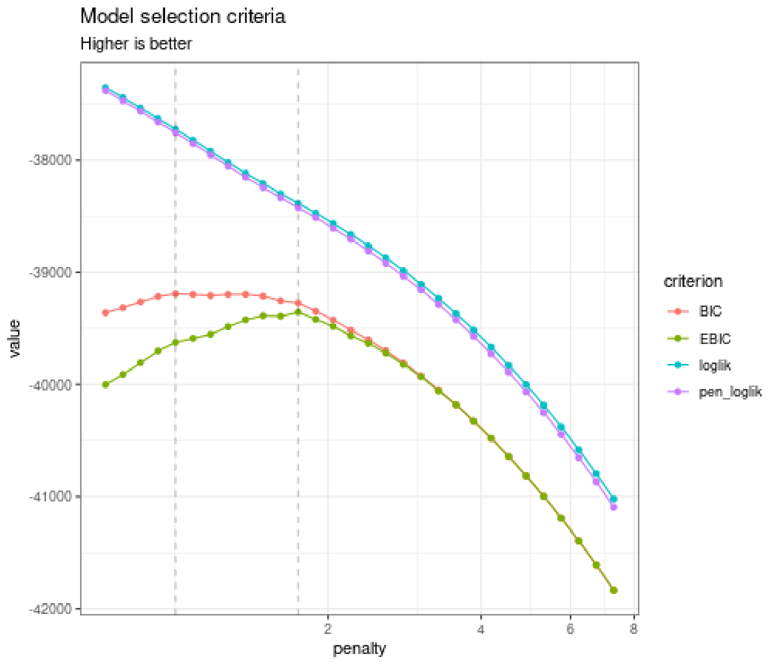
```
PLN_mixtures <-
   PLNmixture(Abundance ~ 1 + offset(log(Offset)), data = oaks, clusters = 1:3)
myPLN_mix <- getModel(PLN_mixtures, 3)
```

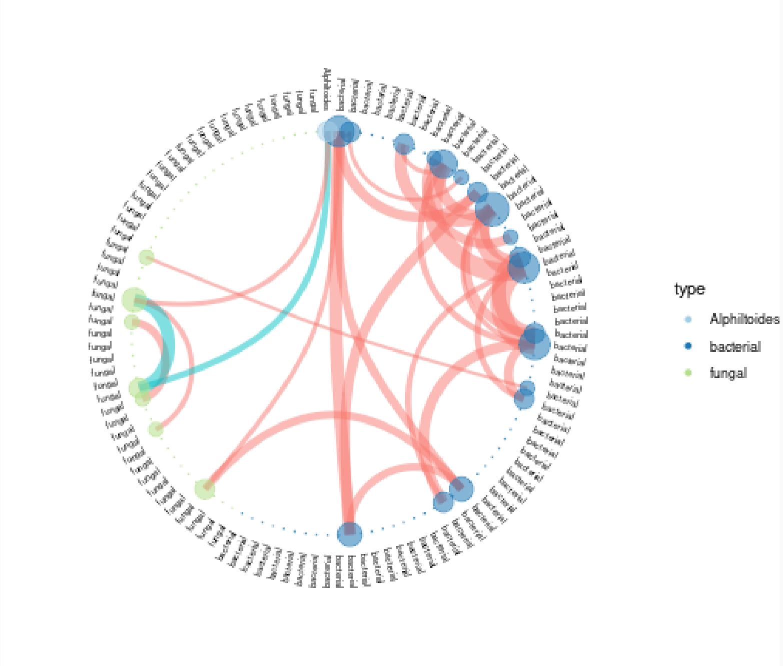
```
myPLN_mix$plot_clustering_pca()
```

```
myPLN_mix$plot_clustering_data()
```

Network inference

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





Some more recent research

Limitations

- No guarantee for the variational estimators
- No test on the inference parameters
- Environmental Genomics, Single-Cell: more zeros, more rows, more columns

Ongoing work

- Hypothesis testing
- Zero-inflated version
- Large scale and/or exact optimization

Statistical Inference

We only consider the standard PLN model.

A first try: Wald test

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

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

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 $\mathbf{\Omega}$ and the regression parameters $\mathbf{\Theta}$.

Variational Wald-test

Variational Fisher Information

The Fisher information matrix is given by

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

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

Wald test and coverage

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

The confidence intervals at level α are given by

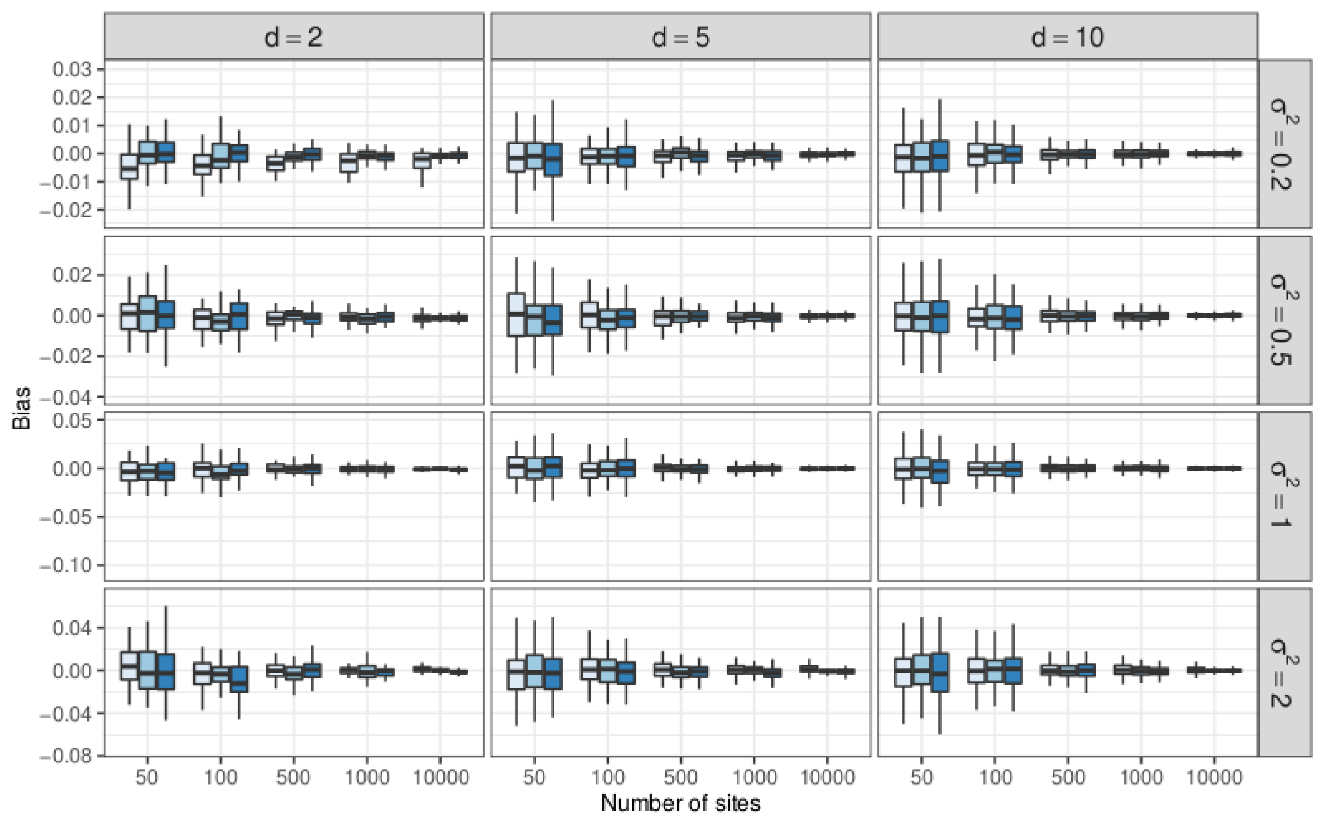
- $B_{kj} = \hat{B}_{kj} \pm \frac{q_{1-\alpha/2}}{\sqrt{n}}\sqrt{\hat{\mathbb{V}}(\boldsymbol{\Theta}_{kj})}$
- $\Omega_{kl} = \hat{\Omega}_{kl} \pm \frac{q_{1-\alpha/2}}{\sqrt{n}}\sqrt{\hat{\mathbb{V}}(\Omega_{kl})}$.

Numerical study

Study Bias and coverage of the estimator

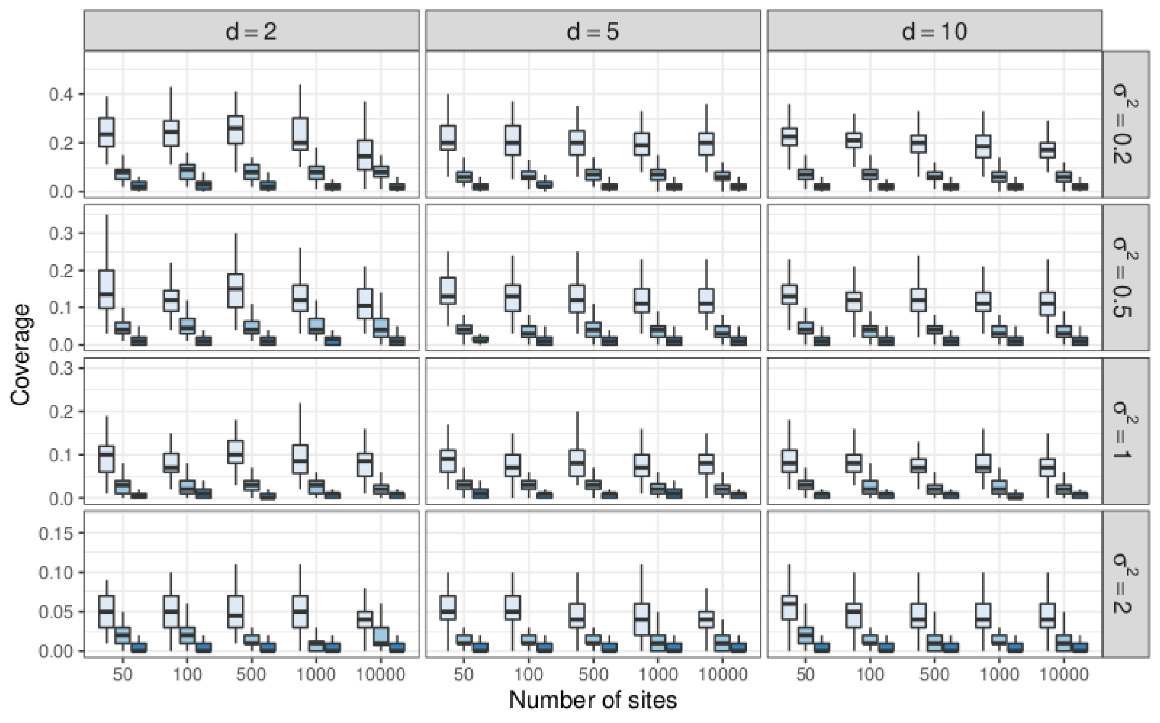
- number of samples $n \in \{50, 100, 500, 1000, 10000\}$
- number of species/genes $p \in \{20, 200\}$
- number of covariates $d \in \{2, 5, 10\}$
- sampling effort $TSS \in \{\text{low}, \text{medium}, \text{high}\}$ ($\approx 10^4$, 10^5 and 10^6)
- noise level $\sigma^2 \in \{0.2, 0.5, 1, 2\}$
- $\boldsymbol{\Sigma}$ 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 $\hat{\Theta}$



The variational estimator are asymptotically unbiased

95% confident interval - Coverage



Variance underestimated, no trusted confidence intervals can be derived out-of-the box

Idea: M-estimation (with M. Mariadassou, S. Robin)

- Asymptotics for the VEM stationary point (not a log-likelihood stationary point).
- Use sandwich estimator for correction as in [WM15]

A zero-inflated PLN with M. Mariadassoua and F. Gindraud

Motivations

- account for a large amount of zero, i.e. with single-cell data,
- try to separate "true" zeros from "technical"/dropouts

The Model

Use two latent vectors $\mathbf{W}_i$ and $\mathbf{Z}_i$ to model excess of zeroes and dependence structure

$$\begin{gathered}
\mathbf{Z}_i \sim \mathcal{N}(\mathbf{o}_i + \mathbf{x}_i^\top \boldsymbol{\Theta}, \boldsymbol{\Sigma}) \\
W_{ij} \sim \mathcal{B}(\mathrm{logit}^{-1}(\mathbf{x}_i^\top \boldsymbol{\Theta}_j^0)) \\
Y_{ij} \,|\, W_{ij}, Z_{ij} \sim W_{ij}\delta_0 + (1 - W_{ij})\mathcal{P}\left(\exp\{Z_{ij}\}\right),
\end{gathered}$$

The unkwown parameters are

- $\boldsymbol{\Theta}$, the regression parameters (from the PLN component)
- $\boldsymbol{\Theta}^0$, the regression parameters (from the Bernoulli component)
- $\boldsymbol{\Sigma}$, the variance-covariance matrix

$\rightsquigarrow$ ZI-PLN is a mixture of PLN and Bernoulli distribution with shared covariates.

ZI-PLN identifiability

Consider the standard ZIPLN model (*i.e.* not the ZIPLN-regression model) with 1 sample:

$$
\begin{aligned}
(W_j)_{j=1\dots p} &\sim \mathcal{B}^{\otimes}(\pi) = \mathcal{B}(\pi_1) \otimes \dots \mathcal{B}(\pi_p) \\
(Z_j)_{j=1\dots p} &\sim \mathcal{N}_p(\mu, \mathbf{\Sigma}) \\
Y_j | W_j, Z_j &\sim (1 - W_j)\mathcal{P}(e^{Z_j}) + W_j \delta_0
\end{aligned}
$$

Proposition

The standard ZIPLN model defined above with parameter $\theta = (\pi, \mu, \mathbf{\Sigma})$ and parameter space $(0,1)^p \times \mathbb{R}^p \times \mathbb{S}_p^{++}$ is identifiable.

Proof. We used the moments of $\mathbf{Y}$ to prove identifiability and rely on the following results for Gaussian and Poisson distributions:

- If $U \sim \mathcal{N}(\mu, \sigma^2)$, then $\mathbb{E}[e^U] = \exp(\mu + \sigma^2/2)$
- If $U \sim \mathcal{P}(\lambda)$ then $\mathbb{E}[U] = \lambda \quad \mathbb{E}[U^2] = \lambda(1 + \lambda) \quad \mathbb{E}[U^2] = \lambda(1 + 3\lambda + \lambda^2)$

Each coordinate of θ can be expressed as a simple functions of the (first three) moments of p_θ and thus $p_\theta = p_{\theta'} \Rightarrow \theta = \theta'$.

ZI-PLN Inference

Same routine...

Variational approximation

$$p(\mathbf{Z}_i, \mathbf{W}_i \mathbf{Y}_i) \approx q_\psi(\mathbf{Z}_i, \mathbf{W}_i) \approx q_{\psi_1}(\mathbf{Z}_i) q_{\psi_2}(\mathbf{W}_i)$$

with

$$q_{\psi_1}(\mathbf{Z}_i) = \mathcal{N}(\mathbf{Z}_i; \mathbf{m}_i, \mathrm{diag}(\mathbf{s}_i \circ \mathbf{s}_i)), \qquad q_{\psi_2}(\mathbf{W}_i) = \otimes_{j=1}^{p} \mathcal{B}(W_{ij}, \pi_{ij})$$

Variational lower bound

Let $\theta = (\mathbf{\Theta}, \mathbf{\Theta}^0, \mathbf{\Sigma})$ and $\psi = (\mathbf{M}, \mathbf{S}, \mathbf{\Pi})$, then

$$\begin{aligned} J(\theta, \psi) &= \log p_\theta(\mathbf{Y}) - KL(p_\theta(.\,|\mathbf{Y}) \| q_\psi(.\,)) \\ &= \mathbb{E}_{q_\psi} \log p_\theta(\mathbf{Z}, \mathbf{W}, \mathbf{Y}) - \mathbb{E}_{q_\psi} \log q_\psi(\mathbf{Z}, \mathbf{W}) \\ &= \mathbb{E}_{q_\psi} \log p_\theta(\mathbf{Y}|\mathbf{Z}, \mathbf{W}) + \mathbb{E}_{q_{\psi_1}} \log p_\theta(\mathbf{Z}) + \mathbb{E}_{q_{\psi_2}} \log p_\theta(\mathbf{W}) \\ &\qquad - \mathbb{E}_{q_{\psi_1}} \log q_{\psi_1}(\mathbf{Z}) - \mathbb{E}_{q_{\psi_2}} \log q_{\psi_2}(\mathbf{W}) \end{aligned}$$

Property: J is separately concave in θ, ψ_1 and ψ_2.

Optimizaton

A sparse criterion

Recall that $\theta = (\boldsymbol{\Theta}, \boldsymbol{\Theta}^0, \boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1})$. Sparsity allows to control the number of parameters:

$$\arg\min_{\theta,\psi} J(\theta, \psi) + \lambda_1 \|\boldsymbol{\Theta}\|_1 + \lambda_2 \|\boldsymbol{\Omega}\|_1 \left(+ \lambda_1 \|\boldsymbol{\Theta}^0\|_1 \right)$$

Alternate optimization

- (Stochastic) Gradient-descent on $\boldsymbol{\Theta}^0, \mathbf{M}, \mathbf{S}$
- Closed-form for posterior probabilities $\boldsymbol{\Pi}$
- Inverse covariance $\boldsymbol{\Omega}$
  - if $\lambda_2 = 0$, $\hat{\boldsymbol{\Sigma}} = n^{-1} \left[(\mathbf{M} - \mathbf{X}\boldsymbol{\Theta})^\top (\mathbf{M} - \mathbf{X}\boldsymbol{\Theta}) + \bar{\mathbf{S}}^2 \right]$
  - if $\lambda_2 > 0$, ℓ_1 penalized MLE ($\rightsquigarrow$ Graphical-Lasso with $\hat{\boldsymbol{\Sigma}}$ as input)
- PLN regression coefficient $\boldsymbol{\Theta}$
  - if $\lambda_1 = 0$, $\hat{\boldsymbol{\Theta}} = [\mathbf{X}^\top \mathbf{X}]^{-1} \mathbf{X}^\top \mathbf{M}$
  - if $\lambda_1 > 0$, vectorize and solve a ℓ_1 penalized least-squared problem

Initialize Θ^0 with logistic regression on $\delta_0(\mathbf{Y})$, $\boldsymbol{\Theta}$ with Poisson regression

A quick example in genomics (1)

scRNA data set

The dataset `scRNA` contains the counts of the 500 most varying transcripts in the mixtures of 5 cell lines in human liver (obtained with standard 10x scRNAseq Chromium protocol).

We subsample 500 random cells and the keep the 200 most varying genes

```r
library(ZIPLN)
data(scRNA)
scRNAsub         ← scRNA[sample.int(nrow(scRNA), 500), ]
scRNAsub$counts ← scRNAsub$counts[, 1:200]
scRNAsub$counts %>% as_tibble() %>% rmarkdown::paged_table()
```

KRT81	AKR1B10	LCN2	AKR1C2	ALDH1A1
<int>	<int>	<int>	<int>	<int>
211	123	1	60	36
3	2	2	1	0
1	4	285	10	0
2	1	1	0	2
7	2	190	2	3

A quick example in genomics (2)

Model fits

We adjust the standard PLN model and the ZI-PLN model with some sparsity on the precision matrix:

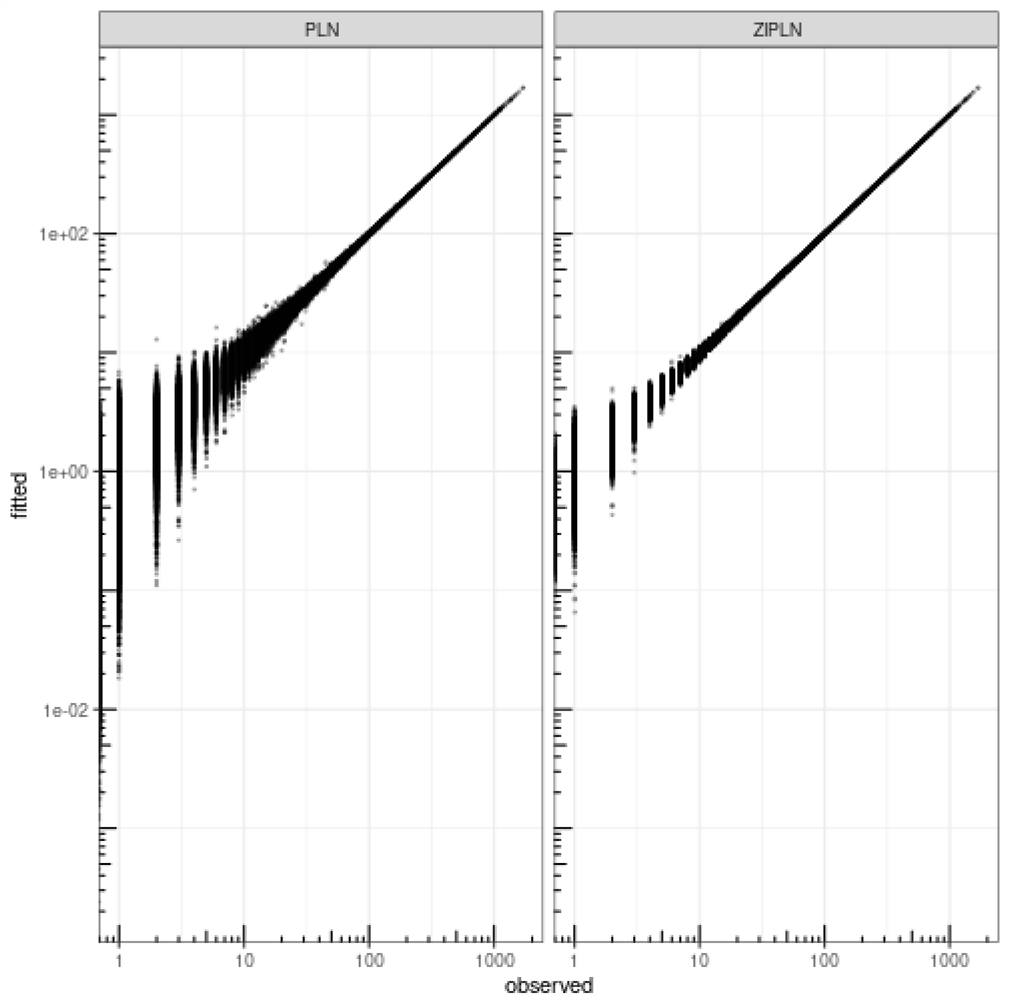
```
system.time(myPLN <-
    PLN(counts ~ 1 + offset(log(total_counts)),
        data = scRNAsub,control = list(trace = 0)))
```

```
##    user  system elapsed
##  23.090   0.032  23.122
```

```
system.time(myZIPLN <-
    ZIPLN(counts ~ 1 + offset(log(total_counts)), rho = .25,
          data = scRNAsub, control = list(trace = 0)))
```

```
##    user  system elapsed
##  50.279   0.052   7.572
```

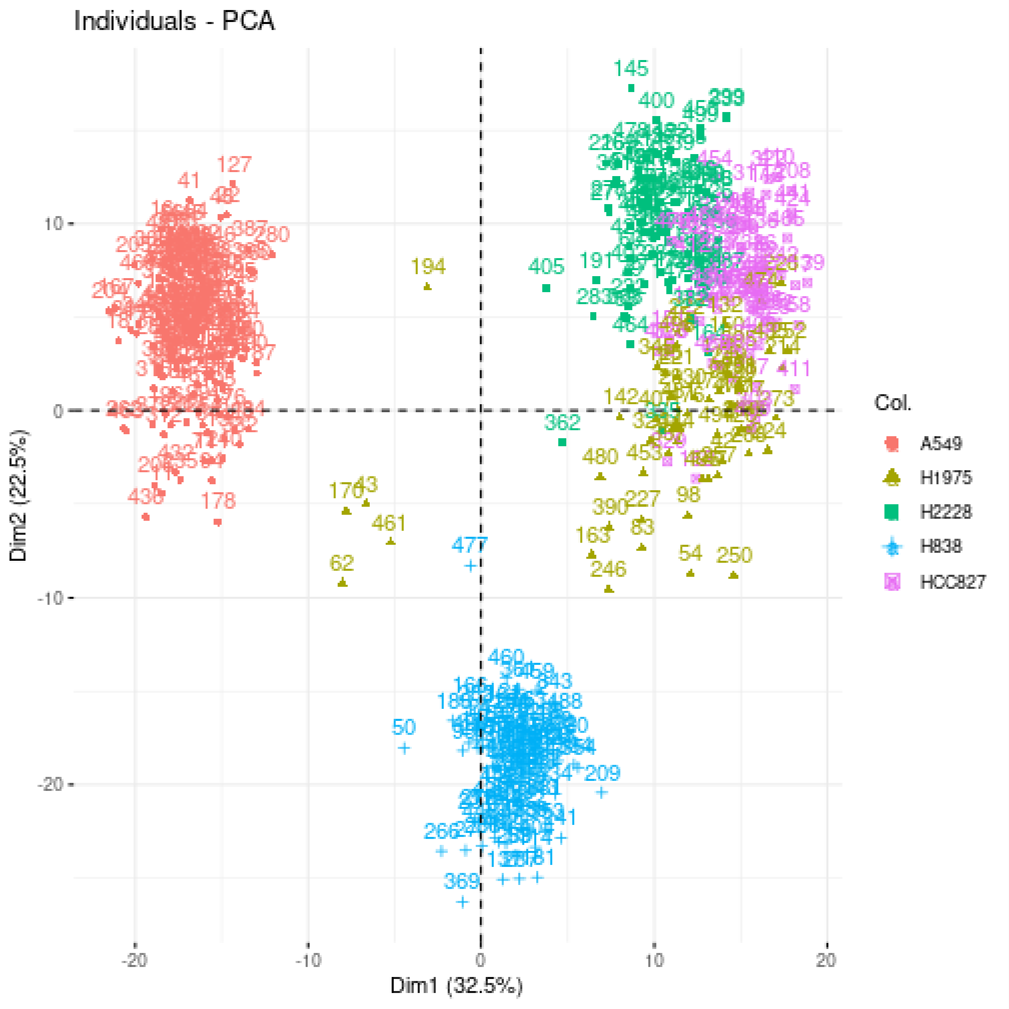
A quick example in genomics (3)



ZI-PLN seems to be less variant for predicting small counts

A quick example in genomics (4)

```
prcomp(myZIPLN$latent) %>% factoextra::fviz_pca_ind(col.ind = scRNAsub$cell_line)
```



A quick example in genomics (5)

```r
library(sbm)
A <- myZIPLN$model_par$Omega != 0; diag(A) <- 0
mySBM <- estimateSimpleSBM(A, estimOptions=list(plot=FALSE))
```

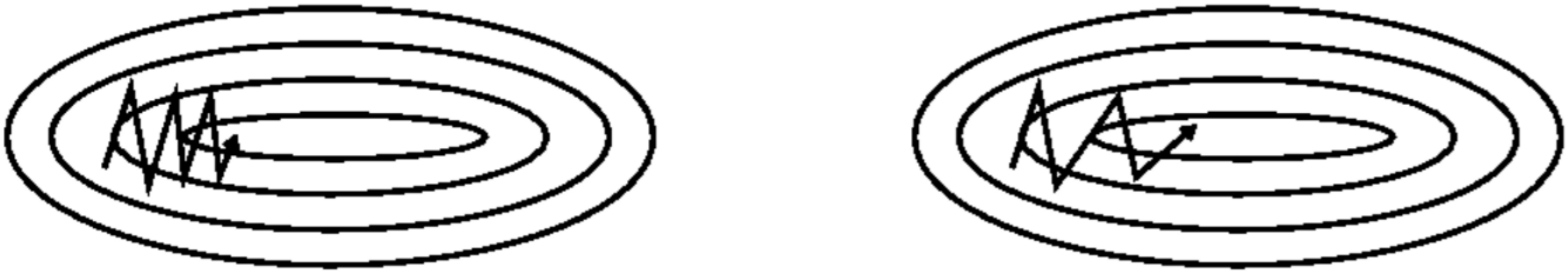
A quick example in genomics (6)

```r
library(igraph)
G <- igraph::graph.adjacency(A, mode = "undirected")
C <- igraph::cluster_fast_greedy(G)
plot(C, G)
```

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

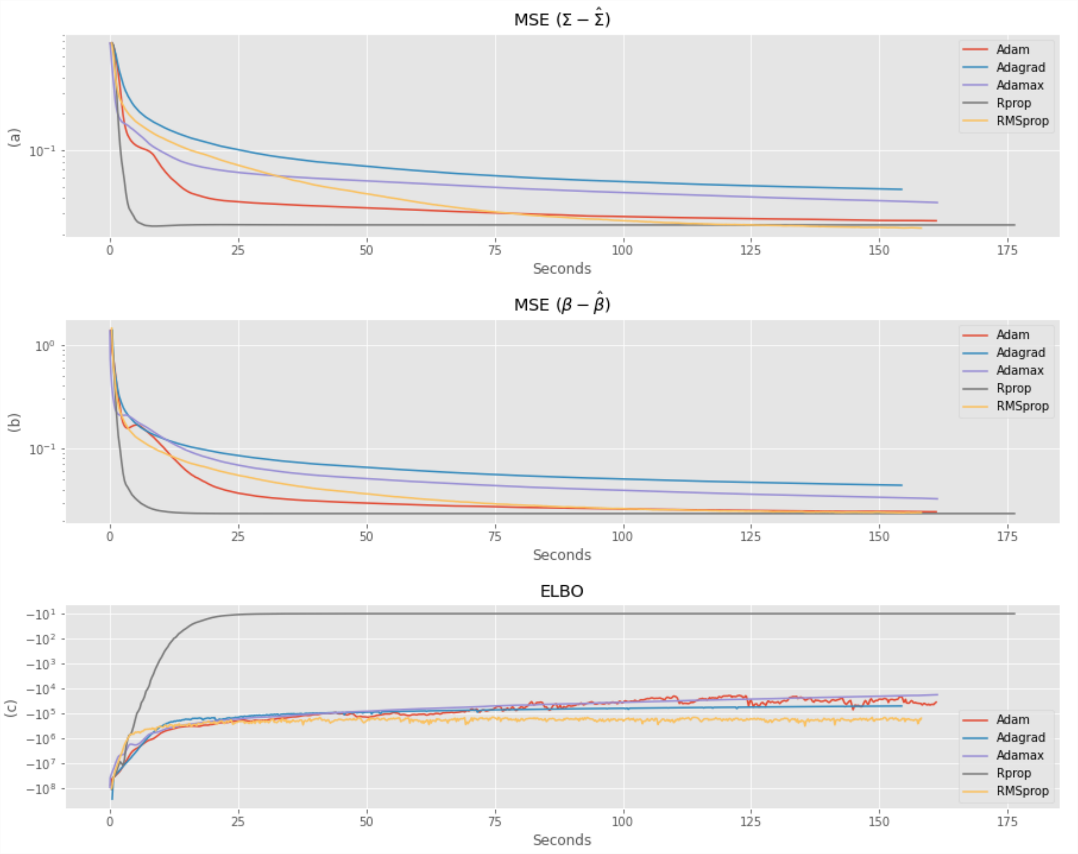


(a) SGD without momentum

(b) SGD with momentum

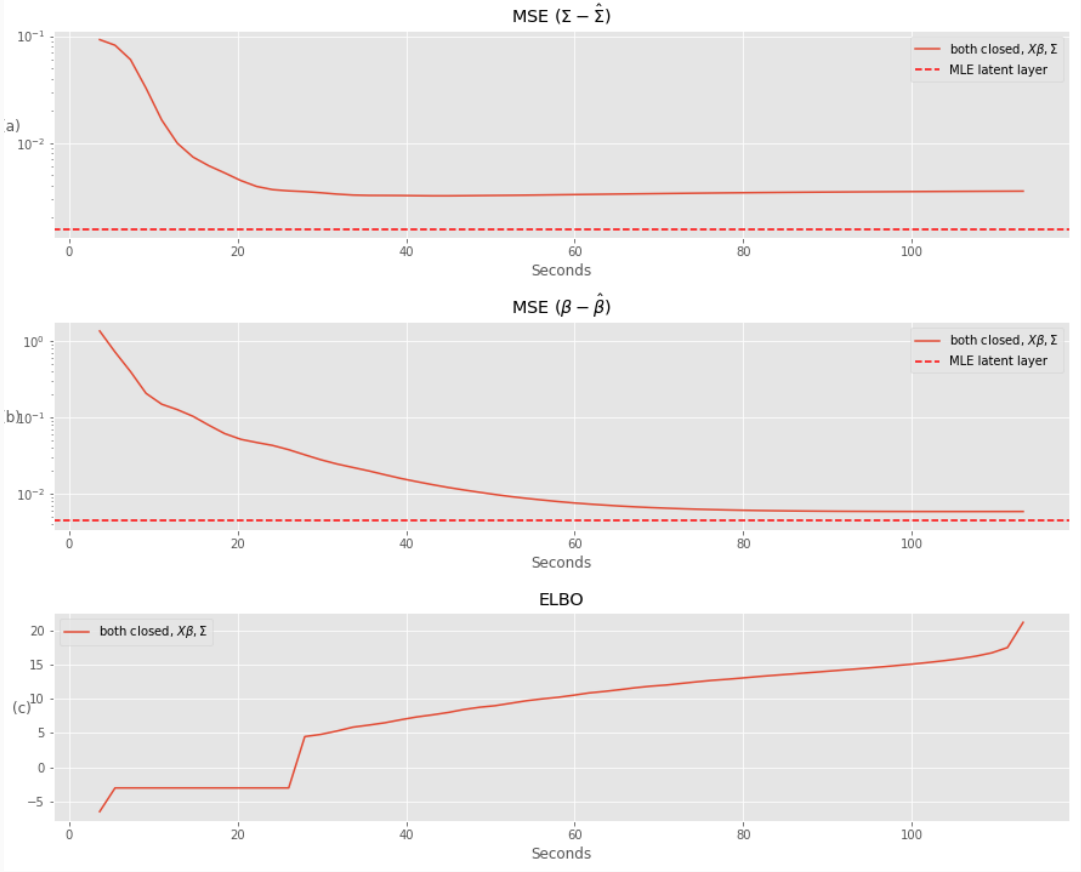
⇝ All available in **Pytorch** with auto-differentiation.

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)

$\rightsquigarrow$ Work up to $n = 100,000, p = 10,000s$

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:

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

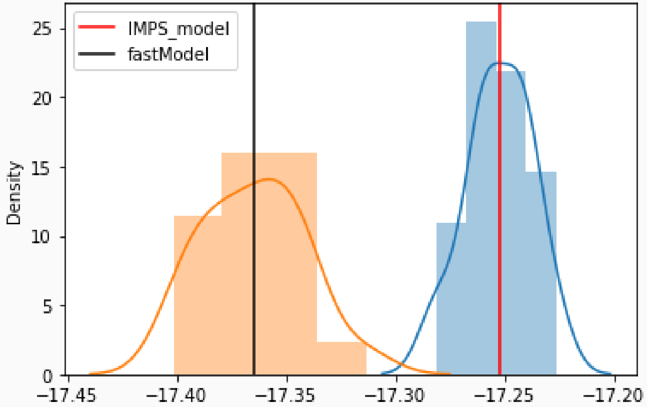
- Estimate the gradients of the log-likelihood by plug-in:

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

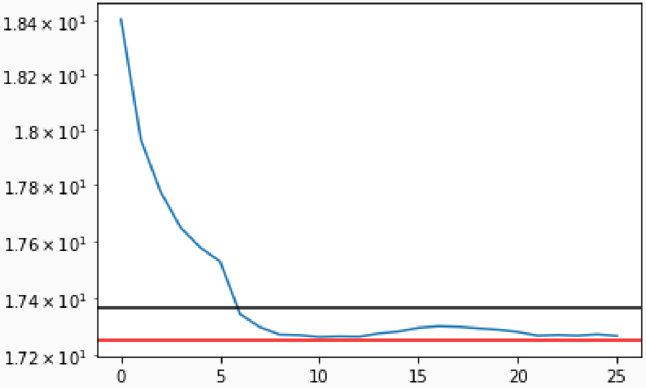
- Use SAGA (F. Bach), a smart incremental gradient descent approach

$\rightsquigarrow$ Work up to $n = 1,000, p = 30$, $n = 1,000, q = 30, p >> q$ for PCA variant

Comparing exact and variational estimators



Log-likelihood computed with importance sampling and variational estimators on 20 replicates



Same, on a single run (blue = IS, black = fastPLN, red = true log-likelihood

The true log-likelihood used for comparison is computed numerically.

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
  - covariance structures (spatial, time series, genetics...)
- 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.