Variational inference of the Poisson log-normal model
Some applications in ecology

S. Robin

Joint work with J. Chiquet & M. Mariadassou

AIGM, Dec. 2017, Toulouse

Community ecology

Abundance data. $Y = [Y_{ij}] : n \times p$:

$$\begin{aligned} Y_{ij} &= \text{abundance of species } j \text{ in sample } i \text{ (old)} \\ &= \text{number of reads associated with species } j \text{ in sample } i \text{ (new)} \end{aligned}$$

Need for multivariate analysis:

- to summarize the information from Y
- to exhibit patterns of diversity
- to understand between-species interactions

Community ecology

Abundance data. $Y = [Y_{ij}] : n \times p$:

$$\begin{aligned} Y_{ij} &= \text{abundance of species } j \text{ in sample } i \text{ (old)} \\ &= \text{number of reads associated with species } j \text{ in sample } i \text{ (new)} \end{aligned}$$

Need for multivariate analysis:

- to summarize the information from Y
- to exhibit patterns of diversity
- to understand between-species interactions

More generally, to model dependences between count variables

$\rightarrow$ Need for a generic (probabilistic) framework

Models for multivariate count data.

Abundance vector: $Y_i = (Y_{i1}, \dots Y_{ip})$, $Y_{ij} =$ counts $\in \mathbb{N}$

Models for multivariate count data.

Abundance vector: $Y_i = (Y_{i1}, \dots Y_{ip})$, $Y_{ij} =$ counts $\in \mathbb{N}$

No generic model for multivariate counts.

- Data transformation ($\widetilde{Y}_{ij} = \log(1 + Y_{ij}), \sqrt{Y_{ij}}$)
  $\rightarrow$ Pb when many counts are zero.

- Poisson multivariate distributions
  $\rightarrow$ Constraints of the form of the dependency [IYAR16]

- Latent variable models
  $\rightarrow$ Poisson-Gamma (= negative binomial): positive dependency
  $\rightarrow$ Poisson-log normal [AH89]

Poisson-log normal (PLN) distribution

Latent Gaussian model:

- $(Z_i)_i$: iid latent vectors $\sim \mathcal{N}_p(0, \Sigma)$
- $Y_i = (Y_{ij})_j$: counts independent conditional on Z_i

$$Y_{ij} \mid Z_{ij} \sim \mathcal{P}\left(e^{\mu_j + Z_{ij}}\right)$$

Poisson-log normal (PLN) distribution

Latent Gaussian model:

- $(Z_i)_i$: iid latent vectors $\sim \mathcal{N}_p(0, \Sigma)$
- $Y_i = (Y_{ij})_j$: counts independent conditional on Z_i

$$Y_{ij} \mid Z_{ij} \sim \mathcal{P}\left(e^{\mu_j + Z_{ij}}\right)$$

Properties:

$$\mathbb{E}(Y_{ij}) = e^{\mu_j + \sigma_j^2/2} =: \lambda_j > 0$$

$$\mathbb{V}(Y_{ij}) = \lambda_j + \lambda_j^2 \left(e^{\sigma_j^2} - 1\right) \qquad \text{(over-dispersion)}$$

$$\mathbb{C}\text{ov}(Y_{ij}, Y_{ik}) = \lambda_j \lambda_k \left(e^{\sigma_{jk}} - 1\right) \qquad \text{(same sign as } \sigma_{jk}\text{)}$$

Poisson-log normal (PLN) distribution

Extensions.

- x_i = vector of covariates for observation i;
- o_{ij} = known 'offset'.

$$Y_{ij} \mid Z_{ij} \sim \mathcal{P}(e^{o_{ij} + x_i^\intercal \beta_j + Z_{ij}})$$

Poisson-log normal (PLN) distribution

Extensions.

- x_i = vector of covariates for observation i;
- o_{ij} = known 'offset'.

$$Y_{ij} \mid Z_{ij} \sim \mathcal{P}(e^{o_{ij} + x_i^\intercal \beta_j + Z_{ij}})$$

Interpretation.

- Dependency structure encoded in the latent space (i.e. in Σ)
- Additional effects are fixed
- Conditional Poisson = noise model

Multivariate analysis of abundance data

Variational inference of PLN

Probabilistic PCA for counts

Network inference

Discussion

Intractable EM

Aim of the inference:

- estimate $\theta = (\beta, \Sigma)$
- predict the Z_i's

Intractable EM

Aim of the inference:

- estimate $\theta = (\beta, \Sigma)$
- predict the Z_i's

Maximum likelihood. EM requires to evaluate (some moments of)

$$p(Z \mid Y) = \prod_i p(Z_i \mid Y_i)$$

but no close form for $p(Z_i \mid Y_i)$.

- [Kar05] resorts to numerical or Monte-Carlo integration.

Variational EM

Variational approximation: replace $p(Z \mid Y)$ with

$$\widetilde{p}(Z) = \prod_i \mathcal{N}(Z_i; \widetilde{m}_i, \widetilde{S}_i)$$

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

$$\begin{aligned} J(\theta, \widetilde{p}) &= \log p_\theta(Y) - KL[\widetilde{p}(Z)||p(Z|Y)] \\ &= \widetilde{\mathbb{E}}[\log p_\theta(Y, Z)] + \mathcal{H}[\widetilde{p}(Z)] \end{aligned}$$

Variational EM

Variational approximation: replace $p(Z \mid Y)$ with

$$\widetilde{p}(Z) = \prod_i \mathcal{N}(Z_i; \widetilde{m}_i, \widetilde{S}_i)$$

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

$$\begin{aligned} J(\theta, \widetilde{p}) &= \log p_\theta(Y) - KL[\widetilde{p}(Z) || p(Z|Y)] \\ &= \widetilde{\mathbb{E}}[\log p_\theta(Y, Z)] + \mathcal{H}[\widetilde{p}(Z)] \end{aligned}$$

Variational EM.

- VE step: find the optimal $\widetilde{p}$ (i.e. $\widetilde{m}_i$'s and diagonal $\widetilde{S}_i$'s)
- M step: update $\widehat{\theta}$.

Variational EM

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

- wrt $\widetilde{p} = (\widetilde{M}, \widetilde{S})$ for fixed θ
- wrt $\theta = (\Sigma, \beta)$ for fixed $\widetilde{p}$ (close form for $\widehat{\Sigma} = n^{-1}(\widetilde{M}^\intercal \widetilde{M} + \widetilde{S}_+))$

but not jointly concave in general.

Implementation: Gradient ascent for the complete parameter $(\widetilde{M}, \widetilde{S}, \theta)$

- No formal VEM algorithm.

Variational EM

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

- wrt $\widetilde{p} = (\widetilde{M}, \widetilde{S})$ for fixed θ
- wrt $\theta = (\Sigma, \beta)$ for fixed $\widetilde{p}$ (close form for $\widehat{\Sigma} = n^{-1}(\widetilde{M}^\intercal \widetilde{M} + \widetilde{S}_+)$)

but not jointly concave in general.

Implementation: Gradient ascent for the complete parameter $(\widetilde{M}, \widetilde{S}, \theta)$

- No formal VEM algorithm.

`PLNmodels` package:

`https://github.com/jchiquet/PLNmodels`

Probabilistic PCA

Dimension reduction. Typical task in multivariate analysis

Probabilistic PCA

Dimension reduction. Typical task in multivariate analysis

Model: Probabilistic PCA (pPCA):

$$\begin{aligned} (Z_i)_i \text{ iid} &\sim \mathcal{N}_p(0, \Sigma), \qquad \text{rank}(\Sigma) = q \ll p \\ Y_{ij} | Z_{ij} &\sim \mathcal{P}(e^{o_{ij} + x_i^\intercal \beta_j + Z_{ij}}) \end{aligned}$$

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

Probabilistic PCA

Dimension reduction. Typical task in multivariate analysis

Model: Probabilistic PCA (pPCA):

$$\begin{aligned} (Z_i)_i \text{ iid} &\sim \mathcal{N}_p(0, \Sigma), \qquad \text{rank}(\Sigma) = q \ll p \\ Y_{ij} | Z_{ij} &\sim \mathcal{P}(e^{o_{ij} + x_i^\intercal \beta_j + Z_{ij}}) \end{aligned}$$

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

pPCA in the PLN model. Variational inference:

$$\text{maximize } J(\theta, \widetilde{p})$$

$\rightarrow$ Still bi-concave in $\theta = (B, \beta)$ and $(\widetilde{M}, \widetilde{S})$

Model selection

Number of components q: needs to be chosen.

Penalized 'likelihood'.

- $\log p_{\widehat{\theta}}(Y)$ intractable: replaced with $J(\widehat{\theta}, \widetilde{p})$
- BIC [Sch78] $\rightarrow$ $vBIC_q = J(\widehat{\theta}, \widetilde{p}) - pq \log(n)/2$
- ICL [BCG00] $\rightarrow$ $vICL_q = vBIC_q - \mathcal{H}(\widetilde{p})$

Chosen rank:

$$\widehat{q} = \arg\max_q vBIC_q \qquad \text{or} \qquad \widehat{q} = \arg\max_q vICL_q$$

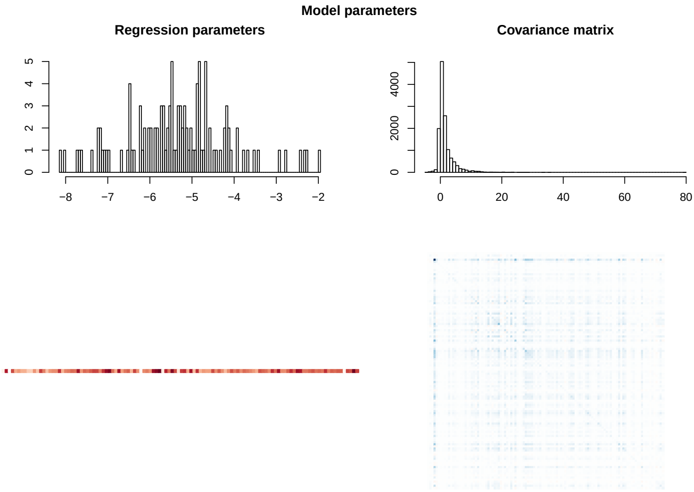
Pathobiome: Oak powdery mildew

Data from [JFS+16].

- $n = 116$ oak leaves = samples
- $p_1 = 66$ bacterial species (OTU)
- $p_2 = 48$ fungal species ($p = 114$)
- covariates: tree (resistant, intermediate, susceptible), branch height, distance to trunk, ...
- offsets: o_{i1}, o_{i2} = offset for bacteria, fungi

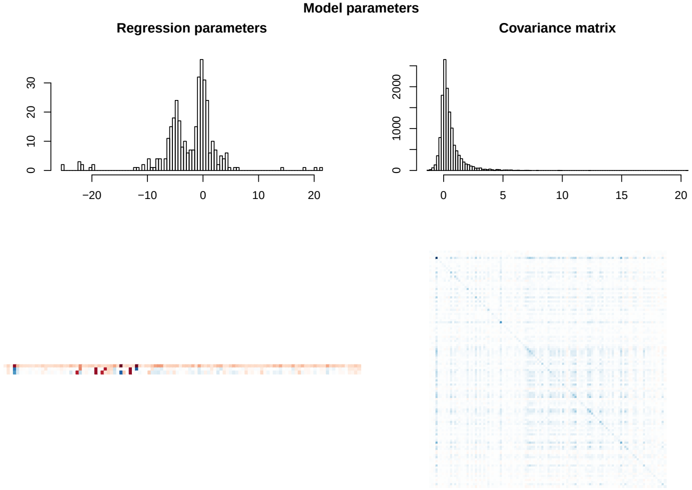
Pathobiome: PLN model ($q = p$)

Without covariates: offset only

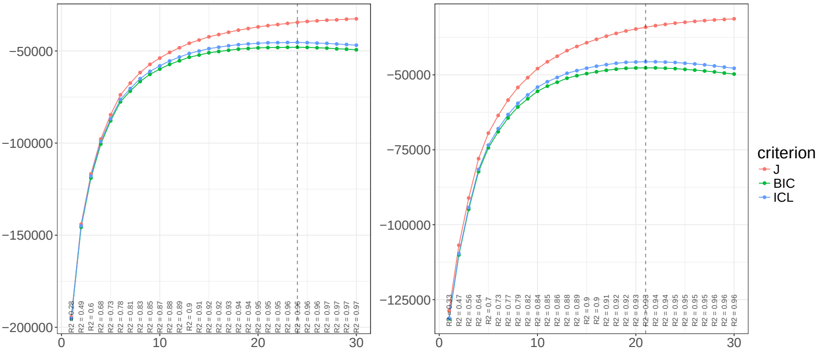


Pathobiome: PLN model ($q = p$)

With covariates: offset, tree (suscept., interm, resist.), orientation



Pathobiome: PCA rank selection



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

Visualization

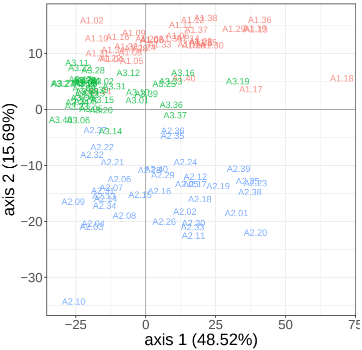
PCA: Optimal subspaces nested when q increases.

PLN-pPCA: Non-nested subspaces.

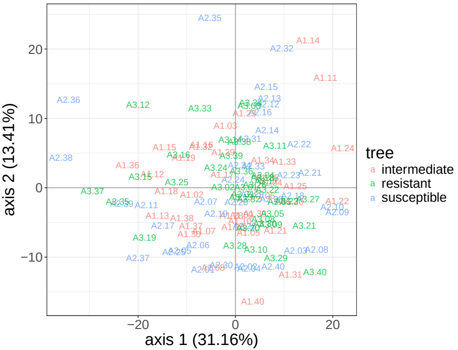
$\rightarrow$ For a the selected dimension $\widehat{q}$:

- Compute the estimated latent positions $\widetilde{M}$
- Perform PCA on the $\widetilde{M}$
- Display results in any dimension $q \leq \widehat{q}$

Pathobiome: First 2 PCs

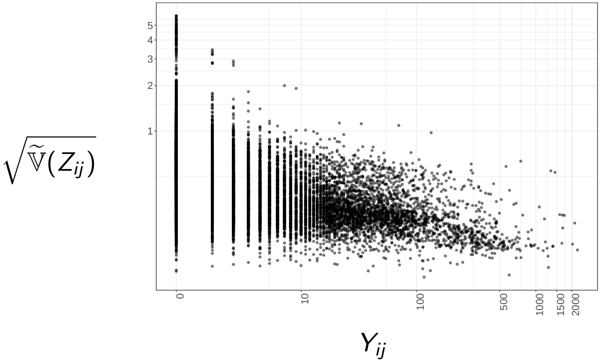


offset only



offset + covariates

Pathobiome: Precision of $\widehat{Z}_{ij}$



Due to the link function (log), $\widetilde{\mathbb{V}}(Z_{ij})$ is higher when Y_{ij} is close to 0.

Problem

Aim: 'infer the ecological network'

Statistical interpretation: infer the graphical model of the $Y_i = (Y_{i1}, \dots Y_{ip})$, i.e. the graph G such that

$$p(Y_i) \propto \prod_{C \in \mathcal{C}(G)} \psi_C(Y_i^C)$$

where $\mathcal{C}(G)$ = set of cliques of G

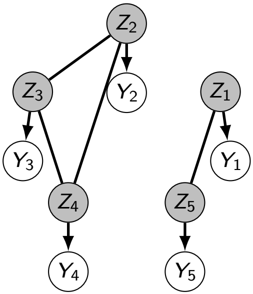
Count data: No generic framework (see Intro)

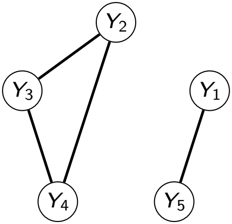
PLN network inference

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

PLN network inference

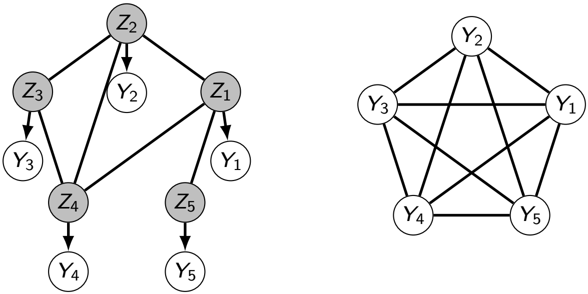
Cheat: Use the PLN model and infer the graphical model of Z





PLN network inference

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



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

PLN network model

Model:

$$\begin{aligned} (Z_i)_i \text{ iid} & \sim \mathcal{N}_p(0, \Omega^{-1}), \qquad \Omega \text{ sparse} \\ Y_{ij} | Z_{ij} & \sim \mathcal{P}(e^{o_{ij} + x_i^\intercal \beta_j + Z_{ij}}) \end{aligned}$$

PLN network model

Model:

$$\begin{aligned} (Z_i)_i \text{ iid} &\sim \mathcal{N}_p(0, \Omega^{-1}), \qquad \Omega \text{ sparse} \\ Y_{ij} | Z_{ij} &\sim \mathcal{P}(e^{o_{ij} + x_i^\intercal \beta_j + Z_{ij}}) \end{aligned}$$

Interest: Similar to Gaussian graphical model (GGM) inference

PLN network model

Model:

$$\begin{array}{rcl}(Z_i)_i \text{ iid} & \sim & \mathcal{N}_p(0, \Omega^{-1}), \qquad \Omega \text{ sparse} \\ Y_{ij} | Z_{ij} & \sim & \mathcal{P}(e^{o_{ij} + x_i^\intercal \beta_j + Z_{ij}})\end{array}$$

Interest: Similar to Gaussian graphical model (GGM) inference

Sparsity-inducing regularization: graphical lasso (gLasso, [FHT08])

$$\log p_\theta(Y) - \lambda \|\Omega\|_{1,\text{off}}$$

Variational inference

Same problem: $\log p_\theta(Y)$ is intractable

Variational inference

Same problem: $\log p_\theta(Y)$ is intractable

Variational approximation: maximize

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

with

$$\widetilde{p}(Z) = \prod \mathcal{N}(Z_i; \widetilde{m}_i, \widetilde{S}_i)$$

Variational inference

Same problem: $\log p_\theta(Y)$ is intractable

Variational approximation: maximize

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

with

$$\widetilde{p}(Z) = \prod \mathcal{N}(Z_i; \widetilde{m}_i, \widetilde{S}_i)$$

→ Still bi-concave in $\theta = (\Omega, \beta)$ and $\widetilde{p} = (\widetilde{M}, \widetilde{S})$. Ex:

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

Model selection

Network density: controlled by λ

Model selection

Network density: controlled by λ

Penalized 'likelihood'.

- $vBIC(\lambda) = J(\widehat{\theta}, \widetilde{p}) - \frac{\log n}{2} \left(pq + |\text{Support}(\widehat{\Omega}_\lambda)| \right)$
- $EBIC(\lambda)$: Extended BIC [FD10]

Model selection

Network density: controlled by λ

Penalized 'likelihood'.

- $vBIC(\lambda) = J(\widehat{\theta}, \widetilde{p}) - \frac{\log n}{2} \left(pq + |\text{Support}(\widehat{\Omega}_\lambda)|\right)$
- $EBIC(\lambda)$: Extended BIC [FD10]

Stability selection.

- Get B subsamples
- Get $\widehat{\Omega}_\lambda^b$ for an intermediate λ and $b = 1 \dots B$
- Count the selection frequency of each edge

Oak powdery mildew: PLNmodels package

Syntax:

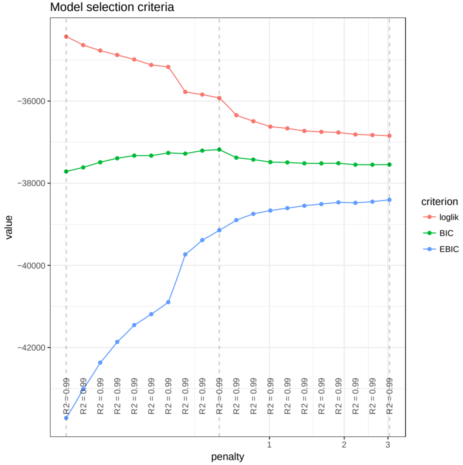
```
formula.offset <- Count ~ 1 + offset(log(Offset))

models.offset <- PLNnetwork(formula.offset)

best.offset <- models.offset$getBestModel("BIC")
```

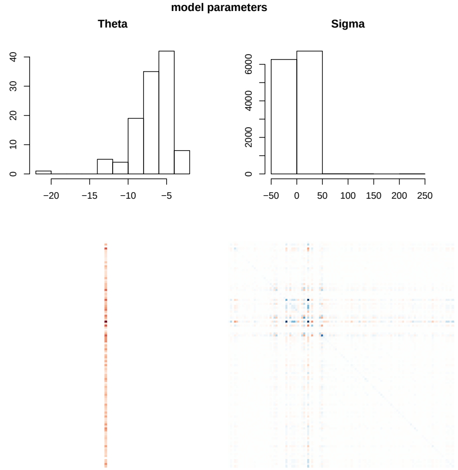
Oak powdery mildew: no covariates

```
models.offset$plot()
```



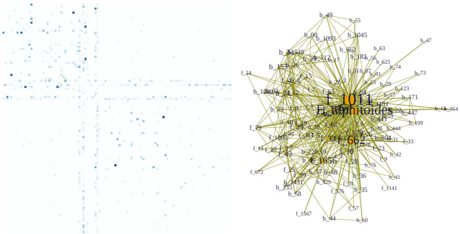
Oak powdery mildew: no covariates

```
best.offset$plot()
```



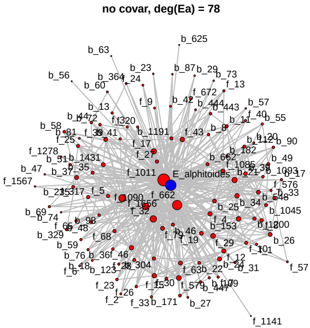
Oak powdery mildew: no covariates

```
best.offset$plot_network()
```

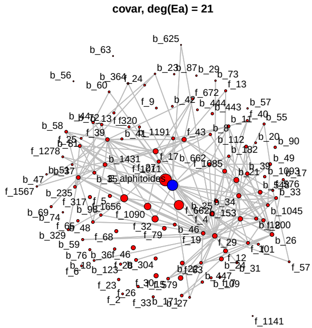


Oak powdery mildew: effect of the covariates

no covariates



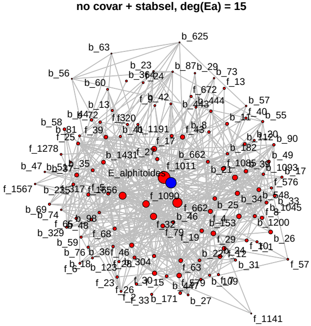
covariate = tree + orientation



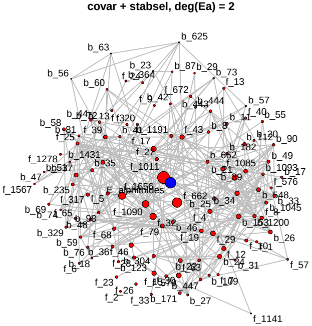
Ea = *Erysiphe alphitoides* = pathogene responsible for oak mildew

Oak powdery mildew: stability selection

no covariates



covariate = tree + orientation



Multivariate analysis of abundance data

Variational inference of PLN

Probabilistic PCA for counts

Network inference

Discussion

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`

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`

To do list

- Model selection criterion for network inference
- Tree-based network inference (R. Momal's PhD)
- Other covariance structures (spatial, time series, ...)
- Statistical properties of the variational estimates (for regular PLN)

References

J. Aitchison and C.H Ho. The multivariate Poisson-log normal distribution. *Biometrika*, 76(4):643–653, 1989.

C. Biernacki, G. Celeux, and G. Govaert. Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Trans. Pattern Anal. Machine Intel.*, 22(7):719–25, 2000.

R. Foygel and M. Drton. Extended Bayesian information criteria for gaussian graphical models. In *Advances in neural information processing systems*, pages 604–612, 2010.

J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

D. I. Inouye, E. Yang, G. I. Allen, and P. Ravikumar. A review of multivariate distributions for count data derived from the Poisson distribution. Technical Report 1609.00066, arXiv, 2016.

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.

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

G. Schwarz. Estimating the dimension of a model. *Ann. Statist.*, 6:461–4, 1978.