Variable selection by an approximation of the L_0 norm in Poisson log-normal (PLN) model

Application in the study of microbial communities in milk production processes

KIOYE Jean Yves[1]
joint work with GROLLEMUND P. M.[1,2]; CHASSARD C.[1] and CHAUVET J.[3]

UMRF[1], LMBP[2], LARIS[3]

July 6, 2023

Context and motivations

Understand what underlies milk quality

- Sensorial quality and biochemical composition
- Prairie biodiversity and livestock farming practices
- Relationship between different microbial communities

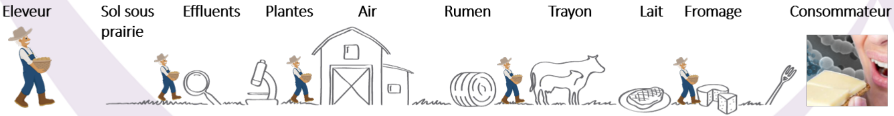
Context and motivations

Understand what underlies milk quality

- Sensorial quality and biochemical composition
- Prairie biodiversity and livestock farming practices
- Relationship between different microbial communities

Improving approaches at agri-food system level

- Impact of farming practices
- Upstream and downstream microbial flows
- Identification of determining factors



Modelisation

- Studying the joint abundances of bacteria
- Evaluating the influence of environmental factors
- Understanding the structural interactions between bacteria
- Take account sampling effort
- Variable selection

The Poisson log-normal (PLN) model

PLN model[1] : special case of generalized linear model

- $\boldsymbol{Y} \in \mathbb{N}^{n\times p}$: response matrix
- $\boldsymbol{X} \in \mathbb{R}^{n\times d}$: environmental variables matrix
- $\boldsymbol{O} \in \mathbb{N}^{n\times p}$: offsets matrix
- $\boldsymbol{B} \in \mathbb{R}^{d\times p}$: regressor matrix
- $\boldsymbol{\Sigma} \in \mathbb{R}^{p\times p}$: covariance matrix

PLN model is :

$$\boldsymbol{Y}_i \mid \boldsymbol{Z}_i \sim \mathcal{P}\big(\exp(\boldsymbol{Z}_i)\big) \qquad \text{(observation layer)}$$

$$\boldsymbol{Z}_i \sim \mathcal{N}_p\big(\boldsymbol{o}_i + \boldsymbol{x}_i^\top \boldsymbol{B}, \boldsymbol{\Sigma}\big) \qquad \text{(latent layer)}$$

1. John AITCHISON et CH HO. « The multivariate Poisson-log normal distribution ». In : *Biometrika* 76.4 (1989), p. 643-653.

PLN Inference

Estimate : $\theta = (\boldsymbol{B}, \boldsymbol{\Sigma})$

2. Dimitris KARLIS. « EM algorithm for mixed Poisson and other discrete distributions ». In : *ASTIN Bulletin : The Journal of the IAA* 35.1 (2005), p. 3-24

3. Julien CHIQUET, Mahendra MARIADASSOU et Stéphane ROBIN. « The Poisson-lognormal model as a versatile framework for the joint analysis of species abundances ». In : *Frontiers in Ecology and Evolution* 9 (2021), p. 588292

PLN Inference

Estimate : $\theta = (\boldsymbol{B}, \boldsymbol{\Sigma})$

Marginal likelihood

$$\log p_\theta(\boldsymbol{Y}) = \int_{\mathbb{R}_p} p_\theta(\boldsymbol{Y}, \boldsymbol{Z}) \, \mathrm{d}\boldsymbol{Z}$$

2. Dimitris Karlis. « EM algorithm for mixed Poisson and other discrete distributions ». In : *ASTIN Bulletin : The Journal of the IAA* 35.1 (2005), p. 3-24

3. Julien Chiquet, Mahendra Mariadassou et Stéphane Robin. « The Poisson-lognormal model as a versatile framework for the joint analysis of species abundances ». In : *Frontiers in Ecology and Evolution* 9 (2021), p. 588292

PLN Inference

Estimate : $\theta = (\boldsymbol{B}, \boldsymbol{\Sigma})$

Marginal likelihood

$$\log p_\theta(\boldsymbol{Y}) = \int_{\mathbb{R}_p} p_\theta(\boldsymbol{Y}, \boldsymbol{Z}) \,\mathrm{d}\boldsymbol{Z}$$

EM algorithm

$\mathbb{E}_\theta[\log p_\theta(\boldsymbol{Y}, \boldsymbol{Z}) | \boldsymbol{Y}]$, but $p_\theta(\boldsymbol{Z}|\boldsymbol{Y}) = \prod_{i=1}^n p_\theta(\boldsymbol{Z}_i|\boldsymbol{Y}_i)$ is intractable

To solve intractability :

- Numerical integration or Monte-Carlo integration [2]
- Variational approximations [3]

2. Dimitris Karlis. « EM algorithm for mixed Poisson and other discrete distributions ». In : *ASTIN Bulletin : The Journal of the IAA* 35.1 (2005), p. 3-24
3. Julien Chiquet, Mahendra Mariadassou et Stéphane Robin. « The Poisson-lognormal model as a versatile framework for the joint analysis of species abundances ». In : *Frontiers in Ecology and Evolution* 9 (2021), p. 588292

Variational inference of the PLN model

Variational approximation

- Approximate $p_\theta(\boldsymbol{Z}_i|\boldsymbol{Y}_i)$ with a multivariate gaussian distribution q_i with mean $\boldsymbol{m}_i$ and variance $\boldsymbol{s^2}_i$
- Replace $p_\theta(\boldsymbol{Z}|\boldsymbol{Y})$ with $\prod_i \mathcal{N}(\boldsymbol{Z}_i; \boldsymbol{m}_i, \text{diag}(\boldsymbol{s}_i^2))$

$\psi = (\boldsymbol{M}, \boldsymbol{S})$: variational parameters

Variational inference of the PLN model

Variational approximation

- Approximate $p_\theta(\boldsymbol{Z}_i|\boldsymbol{Y}_i)$ with a multivariate gaussian distribution q_i with mean $\boldsymbol{m}_i$ and variance $\boldsymbol{s^2}_i$
- Replace $p_\theta(\boldsymbol{Z}|\boldsymbol{Y})$ with $\prod_i \mathcal{N}(\boldsymbol{Z}_i; \boldsymbol{m}_i, \text{diag}(\boldsymbol{s}_i^2))$

$\psi = (\boldsymbol{M}, \boldsymbol{S})$: variational parameters

Evidence Lower Bound (ELBO) of PLN

$$
\begin{aligned}
J(\boldsymbol{Y}, \theta, \psi) &= \log p_\theta(\boldsymbol{Y}) - \text{KL}[q_\psi(\boldsymbol{Z}) || p_\theta(\boldsymbol{Z}|\boldsymbol{Y})] \\
&= \mathbb{E}_{q_\psi}[\log p_\theta(\boldsymbol{Y}, \boldsymbol{Z})] - \mathbb{E}_{q_\psi}[\log q_\psi(\boldsymbol{Z})]
\end{aligned}
$$

Variational EM

- VE step : Optimization of ψ for θ fixed
- VM step : Optimization of θ for ψ fixed

Variable selection

- Regularization of the regression coefficients matrix $\boldsymbol{B}$
- Methods used : Smooth Information Criterion (SIC) [4]
- θ : model parameters
- $\tilde{\theta}$: parameters to be regularized
- k : number of unregulated parameters
- $\ell(\theta)$: log-likelihood

$$\mathrm{SIC} = -2\ell(\boldsymbol{\theta}) + \lambda\left[\|\widetilde{\boldsymbol{\theta}}\|_{0,\varepsilon} + k\right]$$

where $\lambda = 2$ (respectively $\lambda = \log(n)$) for AIC (respectively BIC)

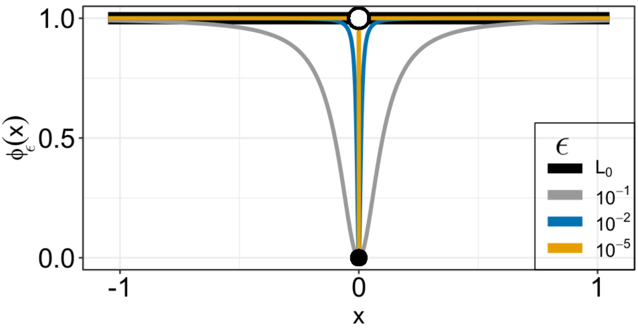
4. Meadhbh O'Neill et Kevin Burke. « Variable selection using a smooth information criterion for distributional regression models ». In : *Statistics and Computing* 33.3 (2023), p. 71.

Smooth Information Criterion (SIC)

- $\|\widetilde{\boldsymbol{\theta}}\|_{0,\varepsilon} = \sum_{j=1}^{d} \phi_\varepsilon(\theta_j)$ is an aproximation of the ≪ L_0 *norm* ≫, with

$$\phi_\varepsilon(x) = \frac{x^2}{x^2 + \varepsilon^2}$$

- ϕ_ε is differentiable for $\varepsilon > 0$, and $\lim_{\varepsilon \to 0} \phi_\varepsilon(x) = \|x\|_0$



Smooth Information Criterion (SIC) [5]

ε-telescoping approach for stable optimization procedure

- *ε-telescoping* : decreasing sequence of ε values
- Avoids the best tuning parameter selection problem
- No requirement to adjust many models
- Computationally advantageous

SIC Algorithm

1: Input : objective, parameters $\boldsymbol{\theta}$
2: decreasing sequence of ε values
3: **For** *each ε value in sequence*
Optimization
$$-2\ell(\boldsymbol{\theta}) + \log(n)\left[\|\boldsymbol{\theta}\|_{0,\varepsilon} + k\right]$$
4: Output : $\boldsymbol{\theta}$

5. Meadhbh O'Neill et Kevin Burke. « Variable selection using a smooth information criterion for distributional regression models ». In : *Statistics and Computing* 33.3 (2023), p. 71

Our contributions

How to adapt the SIC approach to the PLN model ?

- Complex model and multivariate responses
- Coupling ε*-telescoping* for each optimization step

PLN ELBO [6] :

$$J(\boldsymbol{Y},\boldsymbol{\theta},\boldsymbol{\psi}) = \boldsymbol{I}_n\Big[\boldsymbol{Y}\odot(\boldsymbol{O}+\boldsymbol{M}) - \boldsymbol{A} + \frac{1}{2}\log(\boldsymbol{S}^2)\Big]\boldsymbol{I}_p + \frac{n}{2}\log|\boldsymbol{\Omega}| \\ - \frac{n}{2}\mathrm{trace}\Big(\boldsymbol{\Omega}\Big[(\boldsymbol{M}-\boldsymbol{XB})^\top(\boldsymbol{M}-\boldsymbol{XB}) + \mathrm{diag}(\boldsymbol{I}_n^\top\boldsymbol{S}^2)\Big]\Big) \\ + \mathrm{const}$$

ELBO penalized with SIC :

$$J^{pen}(\boldsymbol{Y},\boldsymbol{\theta},\boldsymbol{\psi}) = J(\boldsymbol{Y},\boldsymbol{\theta},\boldsymbol{\psi}) - \lambda\,\|\boldsymbol{B}\|_{0,\varepsilon}$$

6. Julien Chiquet, Stephane Robin et Mahendra Mariadassou. « Variational inference for sparse network reconstruction from count data ». In : *International Conference on Machine Learning*. PMLR. 2019, p. 1162-1171

Proposed algorithm

Input : $\pi^0 = (\boldsymbol{B}^0, \boldsymbol{\Sigma}^0, \boldsymbol{M}^0, \boldsymbol{S}^0)$, $\boldsymbol{E} = (\varepsilon_1, \cdots, \varepsilon_T)$ with $\varepsilon_t = \varepsilon_1 r^{t-1}$ and $r \in]0, 1[$
Output : $\pi^T = (\boldsymbol{B}^T, \boldsymbol{\Sigma}^T, \boldsymbol{M}^T, \boldsymbol{S}^T)$

▷ Start ε-telescoping

For *t in* 1 *to* *T*

▷ Start VEM

Repeat

E step : Variational parameters optimization $\psi = (\boldsymbol{M}, \boldsymbol{S})$
M step : Parameters optimization $\theta = (\boldsymbol{B}, \boldsymbol{\Sigma})$

$$\frac{dJ^{pen}(\boldsymbol{Y}, \boldsymbol{\theta}, \boldsymbol{\psi})}{d\boldsymbol{B}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{M} - \frac{\log(n)}{2} \phi'_{\varepsilon_t}(\boldsymbol{B})$$

$$\frac{dJ^{pen}(\boldsymbol{Y}, \boldsymbol{\theta}, \boldsymbol{\psi})}{d\boldsymbol{\Sigma}} = \cdots$$

until convergence;

▷ End VEM

$\pi^t = (\boldsymbol{B}^t, \boldsymbol{\Sigma}^t, \boldsymbol{M}^t, \boldsymbol{S}^t)$

▷ End ε-telescoping

Simulation studies

Simulation process :

- Variables ($n = 10000$, $d = 6$) following $\boldsymbol{x}_i \sim \mathcal{U}_{[0.5,1.5]}$

Different regression parameter values $\boldsymbol{B}$

- No effect (0)
- weak effect (0.5)
- strong effect (1)

Diagonal covariance matrix

Count data according to PLN

Simulation studies

Simulation process :

- Variables ($n = 10000$, $d = 6$) following $\boldsymbol{x}_i \sim \mathcal{U}_{[0.5,1.5]}$

Different regression parameter values *B*

- No effect (0)
- weak effect (0.5)
- strong effect (1)

Diagonal covariance matrix
Count data according to PLN

Aims :

- Decreases the values of non-active variables to zero
- Minimise the errors in the estimated coefficients

Simulation Results

Table – Real coefficients (estimated coefficients with PLN)

	species 1	species 2	species 3	species 4
x_1	0 (0.159)	0.5 (0.546)	1 (1.120)	1 (1.048)
x_2	1 (1.107)	0 (0.161)	0.5 (0.559)	1 (1.007)
x_3	1 (1.143)	0 (0.089)	0.5 (0.649)	0 (0.026)
x_4	1 (1.148)	1 (1.037)	1 (1.111)	0 (0.098)
x_5	1 (1.136)	1 (1.034)	1 (1.127)	0.5 (0.571)
x_6	0 (0.098)	0 (0.096)	0 (0.090)	0 (0.095)

Simulation Results

Table – Real coefficients (estimated coefficients with PLN)

	species 1	species 2	species 3	species 4
x_1	0 (0.159)	0.5 (0.546)	1 (1.120)	1 (1.048)
x_2	1 (1.107)	0 (0.161)	0.5 (0.559)	1 (1.007)
x_3	1 (1.143)	0 (0.089)	0.5 (0.649)	0 (0.026)
x_4	1 (1.148)	1 (1.037)	1 (1.111)	0 (0.098)
x_5	1 (1.136)	1 (1.034)	1 (1.127)	0.5 (0.571)
x_6	0 (0.098)	0 (0.096)	0 (0.090)	0 (0.095)

Table – Real coefficients (estimated coefficients with PLN SIC PLN)

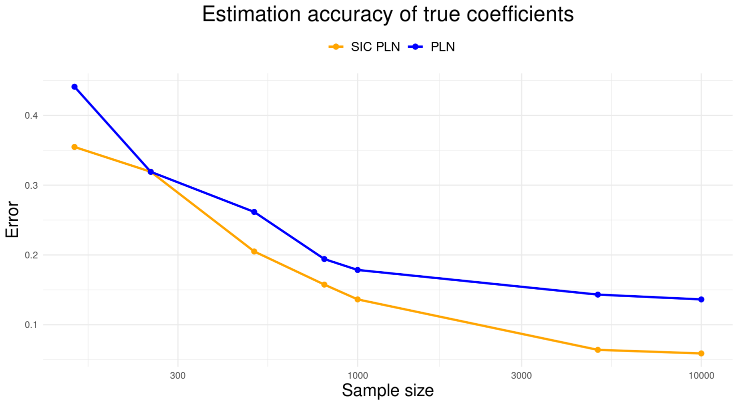
	species 1	species 2	species 3	species 4
x_1	0 (0.059)	0.5 (0.446)	1 (1.020)	1 (0.948)
x_2	1 (1.006)	0 (0.061)	0.5 (0.459)	1 (0.907)
x_3	1 (1.043)	0 (0)	0.5 (0.549)	0 (0)
x_4	1 (1.048)	1 (0.937)	1 (1.011)	0 (0)
x_5	1 (1.036)	1 (0.934)	1 (1.027)	0.5 (0.471)
x_6	0 (0)	0 (0)	0 (0)	0 (0)

- Estimation error for with PLN $\widehat{B}$

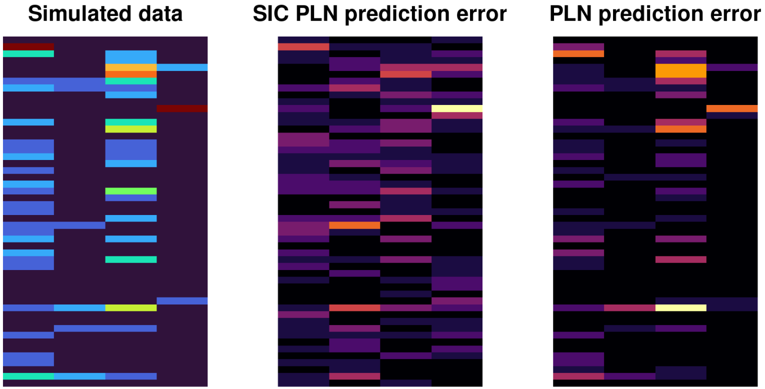
$$\frac{\|B - \widehat{B}\|_F}{\|B\|_F} = 0.136$$

- Estimation error with SIC PLN $\widehat{B}$

$$\frac{\|B - \widehat{B}\|_F}{\|B\|_F} = 0.058$$



Prediction accuracy



Microbiology data UMRF (In progress)

Holoflux metaprogram : 3 projects on microbial flows in agri-food systems

Project amont saint nectaire

Microbiology data UMRF (In progress)

Holoflux metaprogram : 3 projects on microbial flows in agri-food systems

Project amont saint nectaire

- Sample size : 536
- Bacteria 1458
- Abundance : between 0 and 39671

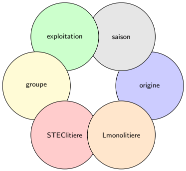
Microbiology data UMRF (In progress)

Holoflux metaprogram : 3 projects on microbial flows in agri-food systems

Project amont saint nectaire

- Sample size : 536
- Bacteria 1458
- Abundance : between 0 and 39671

- categorical environmental variables



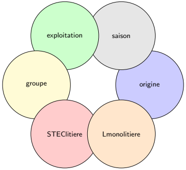
Microbiology data UMRF (In progress)

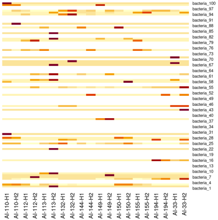
Holoflux metaprogram : 3 projects on microbial flows in agri-food systems

Project amont saint nectaire

- Sample size : 536
- Bacteria 1458
- Abundance : between 0 and 39671

- categorical environmental variables





Other data

- Project MINDS : botanical diversity
- Project TANDEM : agricultural practices

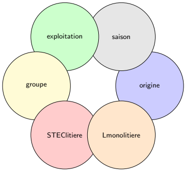
Microbiology data UMRF (In progress)

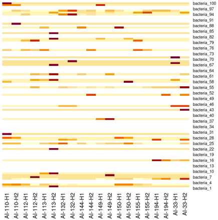
Holoflux metaprogram : 3 projects on microbial flows in agri-food systems

Project amont saint nectaire

- Sample size : 536
- Bacteria 1458
- Abundance : between 0 and 39671

- categorical environmental variables





Other data

- Project MINDS : botanical diversity
- Project TANDEM : agricultural practices

What environmental variables and farming practices explain microbial community abundances ?

Conclusion & perspectives

Conclusion

- Extension of SIC to the PLN model
- Identifies relevant continuous variables by stepwise approximation of the L_0 norm
- Selection by maximising an information criterion

Perspectives

- Penalized coefficients matrix and covariance matrix
- How SIC works on categorical data ?

Thank you for your attention ! ! !

"All models are wrong, but some are useful." *George E. P. Box*