VARIABLE SELECTION BY AN APPROXIMATION OF THE ℓ_0 NORM IN PLN MODEL

Togo Jean Yves KIOYE[2,3], Paul-Marie GROLLEMUND[1,2,3], Jocelyn CHAUVET[4,5] and Christophe CHASSARD[2,3]

[1]LMPB; [2]Clermont Auvergne University; [3]UMRF, INRAE, VetAgro Sup; [4]LARIS, Angers University; [5]ICES, Vendée Catholic Institute

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

STUDY OF MICROBIAL COMMUNITIES

Several ecosystems concerned

- **Environment**: soil, grass, air
- **Farm**: barn, bedding, feed
- **Cow**: teats, faeces, rumen, milk
- **Milk storage**, **cheese**

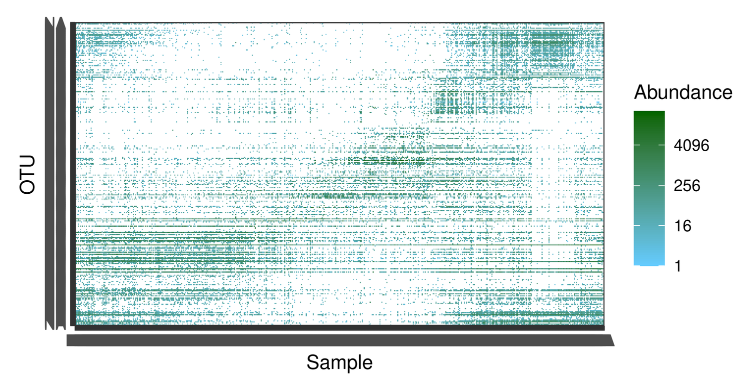
Data collected about each ecosystem

- **Microbial** composition
- **Physico-chemical** composition
- Presence of **pathogens**
- Type of **farming system**

TOOLS OF DATA COLLECTION

Measuring the presence of microbes in milk [1]

- Advanced genomics techniques: abundance of each species
- Tracking microbial species or strains over several years
- **Big datasets** (**biostatistics and bioinformatics**)



Database of projects

- **Amont Saint-Nectaire** project: impact of environmental variables
- **MINDS** project: differences in microbiota as a function of botanical diversity
- **TANDEM** project: difference in microbiota, agroecology / intensive agriculture, resilience to disturbance

OBJECTIVES

- Studying the **joint abundances of bacteria**
- Evaluating the influence of **environmental factors**
- Understanding the structural **interactions between bacteria**
- Taking account of **offsets**
- **Variable selection**

MODELISATION AND INFERENCE

Poisson Log-Normal (PLN) model [2]

observation: $Y_i \mid Z_i \sim \mathcal{P}\left(\exp(Z_i)\right)$

latent: $Z_i \sim \mathcal{N}_p\left(o_i + x_i^\top B, \Sigma\right)$

Model parameters: $\theta = (B, \Sigma)$

- $Y \in \mathbb{N}^{n \times p}$: responses
- $X \in \mathbb{R}^{n \times d}$: variables
- $O \in \mathbb{N}^{n \times p}$: offsets
- $B \in \mathbb{R}^{d \times p}$: regressors
- $\Sigma \in \mathbb{R}^{p \times p}$: covariance

Inference of PLN

Marginal likelihood: $\log p_\theta(Y) = \int_{\mathbb{R}_p} p_\theta(Y, Z)\, dZ$

EM algorithm: $\mathbb{E}_\theta[\log p_\theta(Y, Z) \mid Y]$ (**intractable**)

Variational EM [2]: Maximises the Evidence Lower Bound (ELBO)

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

Variational parameters: $\psi = (M, S)$

SPARSE INFERENCE

Ideal variable selection strategy

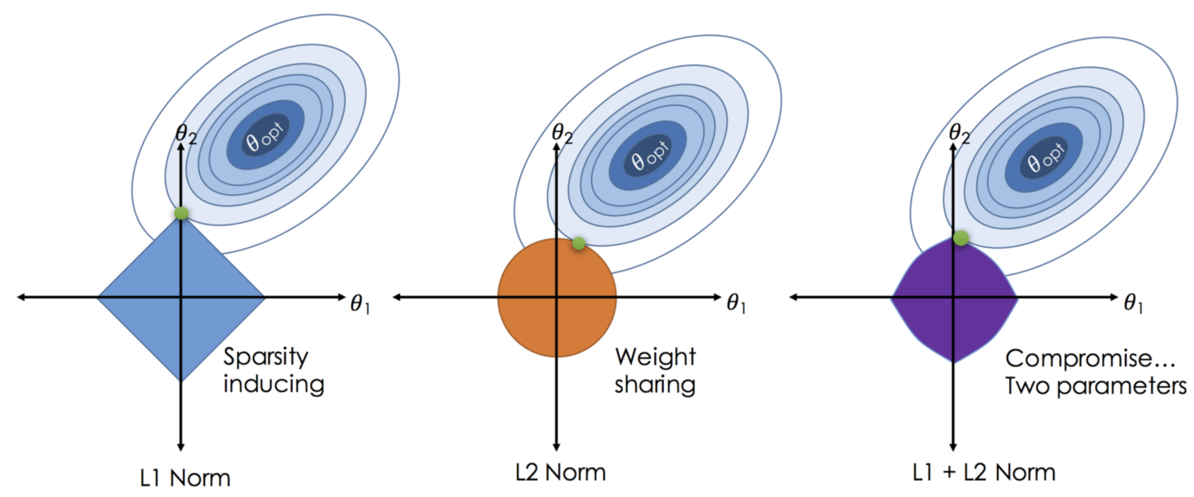
- Add an ℓ_0 penalty
- NP-hard problem
- Difficult to optimize
- ℓ_0 is non-convex

Some relaxing strategies

- Add an ℓ_q penalty to the lower bound of the likelihood (ELBO)
- Select an optimal tuning parameter λ
- Maximizing an information criterion: BIC, AIC

Penalization of the ELBO

$$J_{pen}(Y, B, \Sigma, \psi) = J(Y, B, \Sigma, \psi) - \lambda \|B\|_q$$



USING SMOOTH INFORMATION CRITERION (SIC) [3]

$$J_{pen}(Y, B, \Sigma, \psi) = J(Y, B, \Sigma, \psi) - \lambda \|B\|_{0,\varepsilon}$$

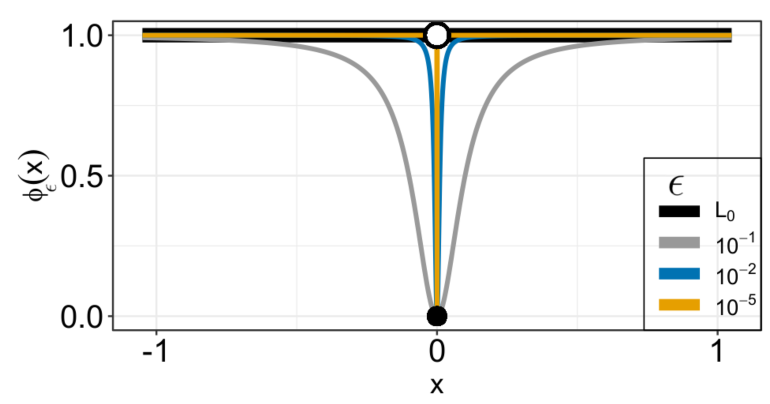
- $\|B\|_{0,\varepsilon} = \sum_{i=1}^{p} \sum_{j=1}^{d} \phi_\varepsilon(B_{i,j})$

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

- ϕ_ε is differentiable for $\varepsilon > 0$, and

$$\lim_{\varepsilon \to 0} \phi_\varepsilon(x) = \|x\|_0$$

- For **BIC**: $\lambda = \log(n)$; for **AIC**: $\lambda = 2$ (computationally advantageous)



- **ε-telescoping approach** to stabilize the optimization procedure

Optimization algorithm: coupling ε-telescoping and VEM

For each decreasing value of ε:

- **VE step**: Optimization of variational parameters ψ for θ fixed
- **VM step**: Optimization of model parameters $\theta = (B, \Sigma)$ for ψ fixed

APPLICATIONS

Numerical Study

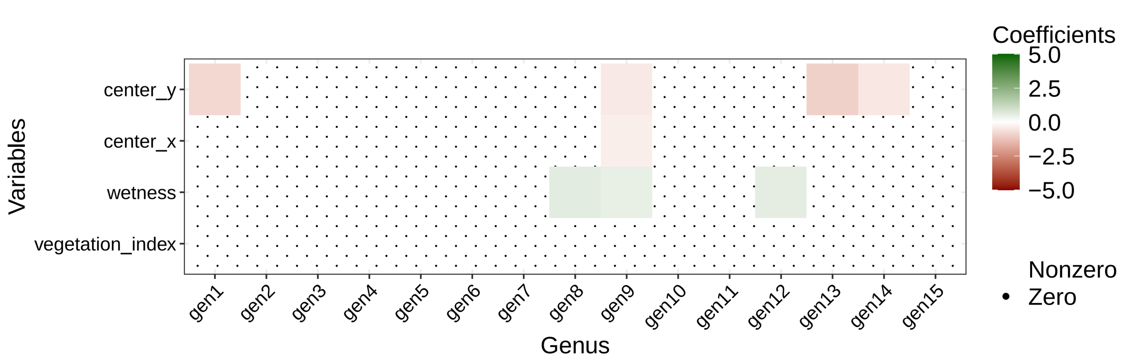
- **Data**: $n = 10000$, $d = 6$, $p = 4$
- **Coefficients B with entries**
  - **0**: no effect
  - **0.5**: weak effect
  - **1**: strong effect

Real (estimated) coefficients

	specie 1	specie 2	specie 3	specie 4
x_1	0 (0)	0.5 (0.47)	1 (1.02)	1 (1.03)
x_2	1 (0.98)	0 (0)	0.5 (0.50)	1 (0.92)
x_3	1 (0.97)	0 (0)	0.5 (0.50)	0 (0)
x_4	1 (1.02)	1 (0.90)	1 (0.97)	0 (0)
x_5	1 (1.01)	1 (0.92)	1 (1.03)	0.5 (0.42)
x_6	0 (0)	0 (0)	0 (0)	0 (0)

Genus data[4]

- **Sample size**: 1000
- **Number of genus**: 15
- **Abundance**: between 0 and 203
- **Variables**: center_y, center_x, wetness, vegetation_index



CONCLUSION

- Extension of SIC to the PLN model
- Identifies relevant variables by stepwise approximation of the ℓ_0 norm and decreases the coefficients of non-active variables to zero
- Application on UMRF data (Amont Saint-Nectaire, MINDS, TANDEM)
- Extension on zero-inflated PLN

REFERENCES

[1]Chassard, C. et al. "Lactic Starter Dose Shapes S. aureus and STEC O26: H11 Growth, and Bacterial Community Patterns in Raw Milk Uncooked Pressed Cheeses". In: *Microorganisms* 9.5 (2021).

[2]J. Chiquet, M. Mariadassou, and S. Robin. "The Poisson-lognormal model as a versatile framework for the joint analysis of species abundances". In: *Frontiers in Ecology and Evolution* 9 (2021).

[3]Meadhbh O'Neill and Kevin Burke. "Variable selection using a smooth information criterion for distributional regression models". In: *Statistics and Computing* 33.3 (2023), p. 71.

[4]J. Chauvet, C. Trottier, and X. Bry. "Component-Based Regularization of Multivariate Generalized Linear Mixed Models". In: *Journal of Computational and Graphical Statistics* 28.4 (2019).

Université Clermont Auvergne