

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}

¹LMPB; ²Clermont Auvergne University; ³UMRF, INRAE, VetAgro Sup; ⁴LARIS, Angers University; ⁵ICES, 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

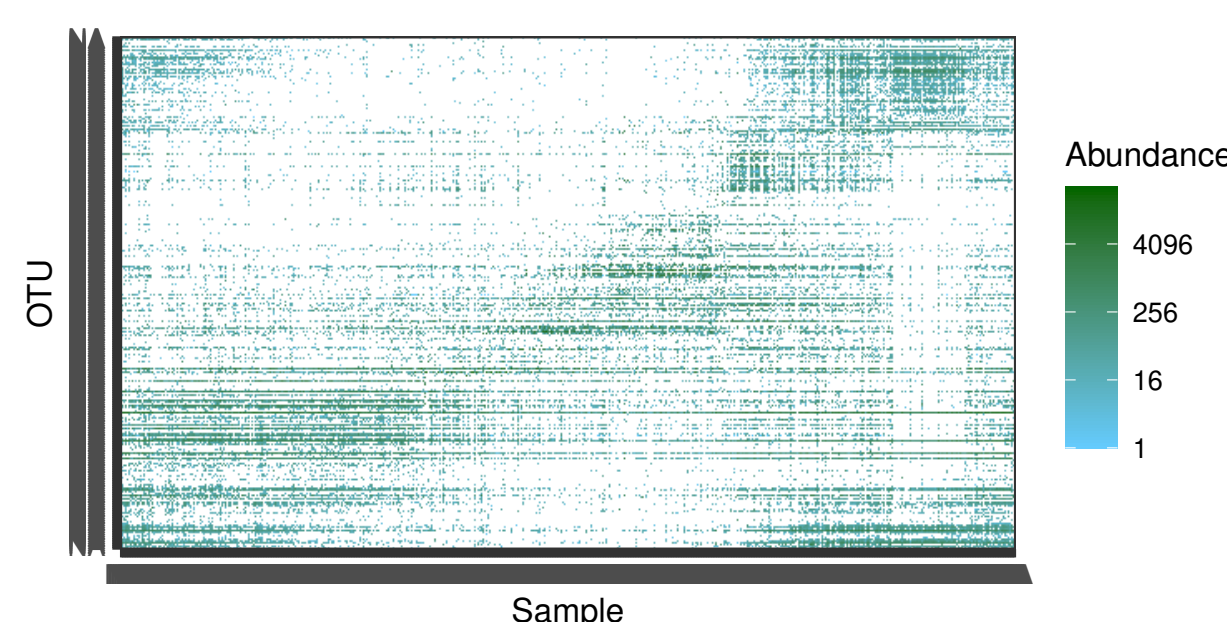
- **Microbial** composition
- **Physico-chemical** composition
- Presence of **pathogens**
- Type of **farming system**



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: $Y_i | Z_i \sim \mathcal{P}(\exp(Z_i))$

latent: $Z_i \sim \mathcal{N}_p(\mathbf{o}_i + \mathbf{x}_i^T \mathbf{B}, \Sigma)$

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

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

Inference of PLN

Marginal likelihood: $\log p_\theta(\mathbf{Y}) = \int_{\mathbb{R}^p} p_\theta(\mathbf{Y}, \mathbf{Z}) d\mathbf{Z}$

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

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

$$J(\mathbf{Y}, \theta, \psi) = \log p_\theta(\mathbf{Y}) - \text{KL}[q_\psi(\mathbf{Z}) || p_\theta(\mathbf{Z} | \mathbf{Y})]$$

$$= \mathbb{E}_{q_\psi}[\log p_\theta(\mathbf{Y}, \mathbf{Z})] - \mathbb{E}_{q_\psi}[\log q_\psi(\mathbf{Z})]$$

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

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).

SPARSE INFERENCE

Ideal variable selection strategy

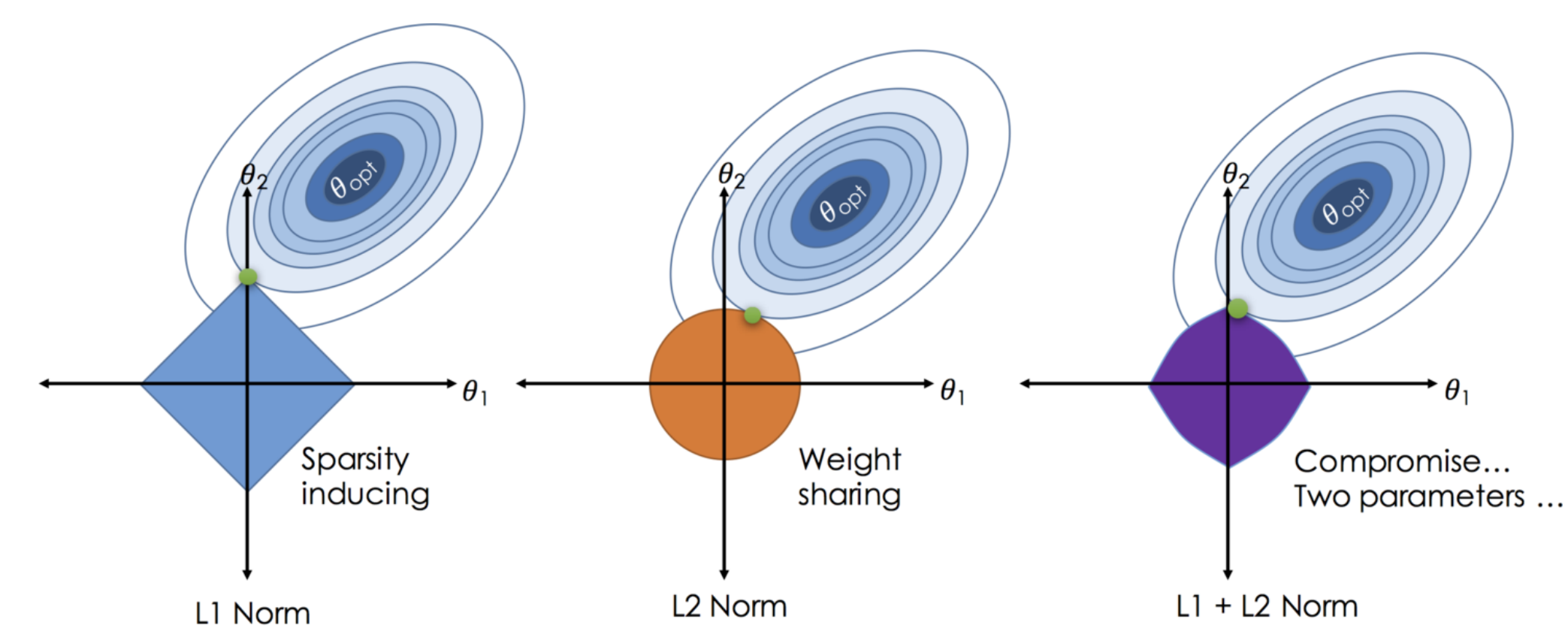
- Add an ℓ_0 penalty
- NP-hard problem
- Difficult to optimize
- ℓ_0 is non-convex

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

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



USING SMOOTH INFORMATION CRITERION (SIC) [3]

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

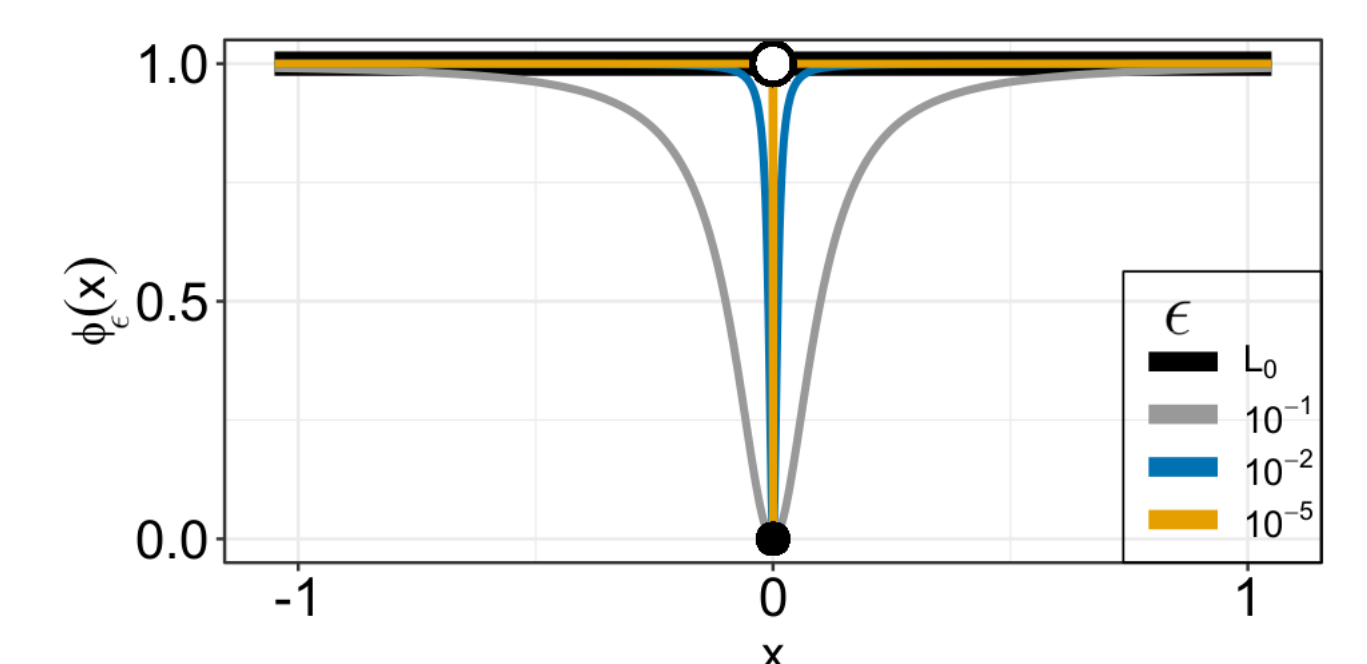
- $\|\mathbf{B}\|_{0,\varepsilon} = \sum_{j=1}^p \sum_{i=1}^d \phi_\varepsilon(\mathbf{B}_{i,j})$

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

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

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

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



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

Optimization algorithm: coupling epsilon-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 $\theta = (\mathbf{B}, \Sigma)$ for ψ fixed

APPLICATIONS

Numerical Study

- **Data**: $n = 10000$, $d = 6$, $p = 4$

► **Coefficients B with entries**

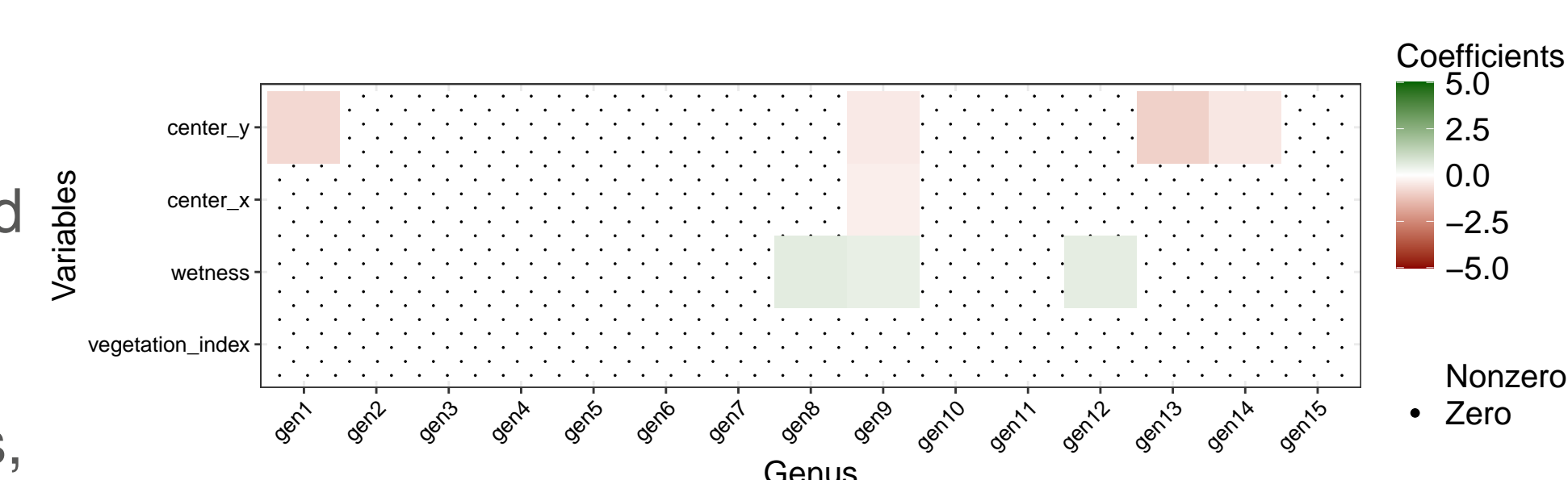
- 0: no effect
- 0.5: weak effect
- 1: strong effect

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
- Selection by maximising BIC and avoiding calibration of a tuning parameter