Étude des communautés microbiennes intervenant dans le processus de production du lait

Sélection de variables dans un modèle de Poisson Log Normal (PLN)

KIOYE Togo Jean Yves[1]

GROLLEMUND P. M.[1,2] ; CHAUVET J.[3,4] ; CHASSARD C.[1]

UMRF[1], LMBP[2], LARIS[3], ICES[4]

Plan

Contexte et motivations

Modélisation de données de comptage multidimensionnelles

- Dépendance entre les comptages
- Comptages influencés par plusieurs variables explicatives (corrélation, redondance, pas d'effet)

Contexte et motivations

Modélisation de données de comptage multidimensionnelles

- Dépendance entre les comptages
- Comptages influencés par plusieurs variables explicatives (corrélation, redondance, pas d'effet)

Plusieurs domaines sont concernés

- Écologie : étude des abondances d'arbres, d'animaux, ...
- Sécurité routière : nombre de décès, d'accidents, ...
- Microbiologie : étude des communautés microbiennes, ...

Contexte et motivations

Microbiologie : comprendre ce qui sous-tend la qualité du lait

- Qualité sensorielle et composition biochimique
- Biodiversité des prairies et pratiques d'élevage
- Relation entre les différentes communautés microbiennes

Contexte et motivations

Microbiologie : comprendre ce qui sous-tend la qualité du lait

- Qualité sensorielle et composition biochimique
- Biodiversité des prairies et pratiques d'élevage
- Relation entre les différentes communautés microbiennes

Proposer des outils de compréhension pour aider à la prise de décision

- Impact des pratiques agricoles
- Étude des flux microbiens en amont et en aval
- Identification des facteurs déterminants

Exemple de données : Genus [1]

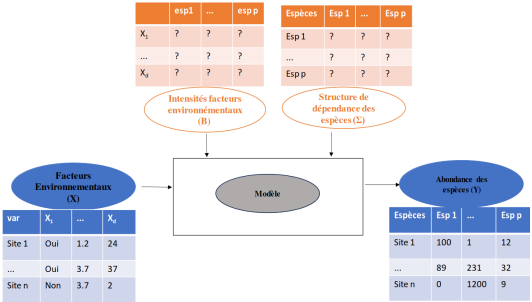
- Abondance : $\boldsymbol{Y}$ $(p = 15)$ d'arbres, $n = 1000$ parcelles (échantillons)

gen1	gen2	gen3	gen4	gen5	gen6	...	gen15
3	0	0	0	3	2	...	6
1	0	1	0	4	4	...	2
...	...	...	...	...	...	...	...

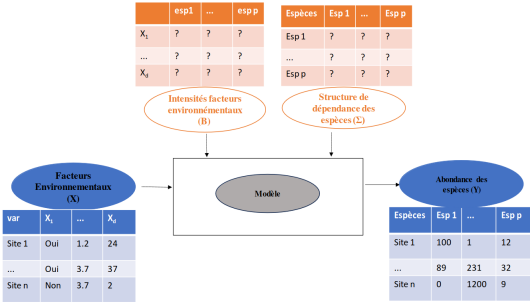
- Variables environnementales : $\boldsymbol{X}$ $(d = 5)$

surface	wetness	center_x	center_y	vegetation_index
15.00	13.00	13.43	3.65	10.15
20.00	13.35	13.48	3.56	10.05
...	...	...	...	...

Modélisation statistique



Modélisation statistique



- Étudier les abondances conjointes des espèces
- Évaluer l'intensité des facteurs environnementaux
- Considérer les interactions entre les espèces

Modèle de Poisson Log Normal (PLN) [2]

Le modèle PLN : cas particulier de modèle de régression

$$\begin{aligned} \boldsymbol{Y}_i \mid \boldsymbol{Z}_i &\sim \mathcal{P}\big(\exp(\boldsymbol{Z}_i)\big) && \text{(espace observé)} \\ \boldsymbol{Z}_i &\sim N_p\big(\boldsymbol{o}_i + \boldsymbol{x}_i^\top \boldsymbol{B}, \boldsymbol{\Sigma}\big) && \text{(espace latent)} \end{aligned} \tag{1}$$

Modèle de Poisson Log Normal (PLN) [2]

Le modèle PLN : cas particulier de modèle de régression

$$\begin{aligned} \boldsymbol{Y}_i \mid \boldsymbol{Z}_i &\sim \mathcal{P}\big(\exp(\boldsymbol{Z}_i)\big) && \text{(espace observé)} \\ \boldsymbol{Z}_i &\sim N_p\big(\boldsymbol{o}_i + \boldsymbol{x}_i^\top \boldsymbol{B}, \boldsymbol{\Sigma}\big) && \text{(espace latent)} \end{aligned} \tag{1}$$

Estimation complexe : approximation variationnelle

Modèle de Poisson Log Normal (PLN) [2]

Le modèle PLN : cas particulier de modèle de régression

$$\begin{aligned} \boldsymbol{Y}_i \mid \boldsymbol{Z}_i &\sim \mathcal{P}\big(\exp(\boldsymbol{Z}_i)\big) && \text{(espace observé)} \\ \boldsymbol{Z}_i &\sim N_p\big(\boldsymbol{o}_i + \boldsymbol{x}_i^\top \boldsymbol{B}, \boldsymbol{\Sigma}\big) && \text{(espace latent)} \end{aligned} \tag{1}$$

Estimation complexe : approximation variationnelle

Quelques besoins en analyse multidimensionnelle

- Résumer l'information de $\boldsymbol{Y}$ ou $\boldsymbol{X}$ (ACP, clustering, classification, ...)
- Identifier les interactions pertinentes : $\boldsymbol{\Sigma}$ (sélection de dépendances)
- Identifier les variables qui expliquent les abondances observées : $\boldsymbol{B}$

Sélection de variables

Plusieurs méthodes existent :

- Sélection du meilleur sous-ensemble : forward-backward, setpwise, etc.
- Coûteux sur le plan calculatoire
- Sélection de modèle : AIC, BIC, etc.

Sélection de variables

Plusieurs méthodes existent :

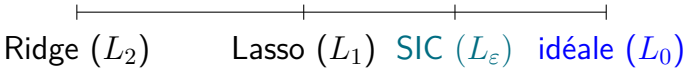
- Sélection du meilleur sous-ensemble : forward-backward, setpwise, etc.
- Coûteux sur le plan calculatoire
- Sélection de modèle : AIC, BIC, etc.

Question pas simple : approches modernes

- Méthode de régularisation : optimisation sous contrainte
- Contraintes relaxées : lasso, elastic-net, etc.
- Calibrage d'un paramètre de régularisation

Smooth Information Criterion (SIC)[3]

Récente contribution : **Smooth Information Criterion (SIC) [3]**



Smooth Information Criterion (SIC)[3]

Récente contribution : **Smooth Information Criterion (SIC) [3]**

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

- Approche du vrai problème :
  $\lim_{\varepsilon \to 0} \phi_\varepsilon(x) = \|x\|_0$ (L_0)
- Pas de calibrage d'un paramètre de régularisation

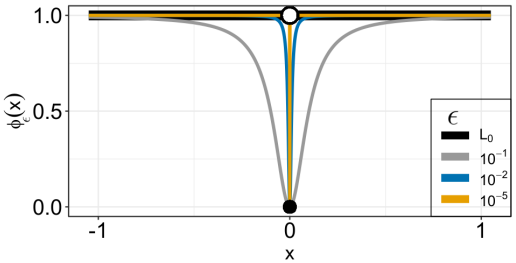
Smooth Information Criterion (SIC)[3]

Récente contribution : **Smooth Information Criterion (SIC) [3]**

Ridge (L_2) Lasso (L_1) SIC (L_ε) idéale (L_0)

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

- Approche du vrai problème : $\lim_{\varepsilon \to 0} \phi_\varepsilon(x) = \|x\|_0$ (L_0)
- Pas de calibrage d'un paramètre de régularisation



Article soumis au Journal of Multivariate Analysis (mars 2023)

SPARSE INFERENCE IN POISSON LOG-NORMAL MODEL BY APPROXIMATING THE L_0-NORM

Togo Jean Yves KIOYE,
Unité Mixte de Recherche sur le Fromage (UMRF)
Université Clermont Auvergne, France
togo_jean_yves.kioye@uca.fr

Paul-Marie GROLLEMUND
Laboratoire de Mathématiques Blaise Pascal (LMBP)
Unité Mixte de Recherche sur le Fromage (UMRF)
Université Clermont Auvergne, France
paul_marie.grollemund@uca.fr

Jocelyn CHAUVET
Laboratoire Angevin de Recherche en Ingénierie des Systèmes (LARIS)
Centre de recherche de l'ICES, France
jchauvet@ices.fr

Pierre DRUILHET, Erwan SAINT-LOUBERT-BIE
Laboratoire de Mathématiques Blaise Pascal (LMBP)
Université Clermont Auvergne, France
{pierre.druilhet,erwan.saint-loubert-bie}@uca.fr

Christophe CHASSARD
Unité Mixte de Recherche sur le Fromage (UMRF)
INRAE, France
christophe.chassard@inrae.fr

Nos contributions

Nos contributions

- Reformulations et interprétations

Nos contributions

- Reformulations et interprétations
- Travail théorique
- Nouvel algorithme SICPLN : Couplage VEM + ε-telescoping
- Illustration sur des données simulées et réelles

Données Genus [1]

Données Genus [1] : résultats d'estimation avec PLN

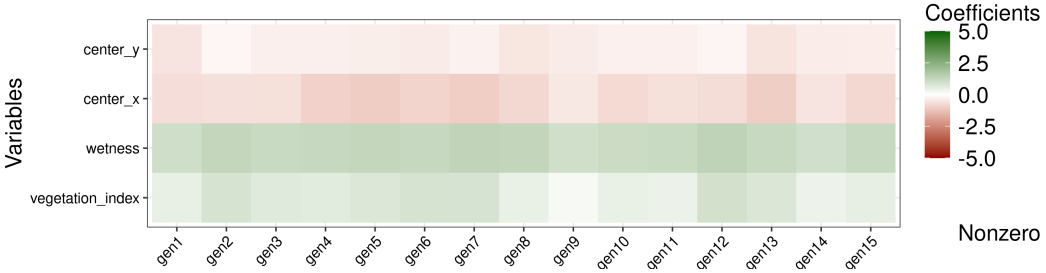


Figure 1 – Estimation des coefficients avec PLN.

Données Genus [1] : résultats d'estimation avec PLN

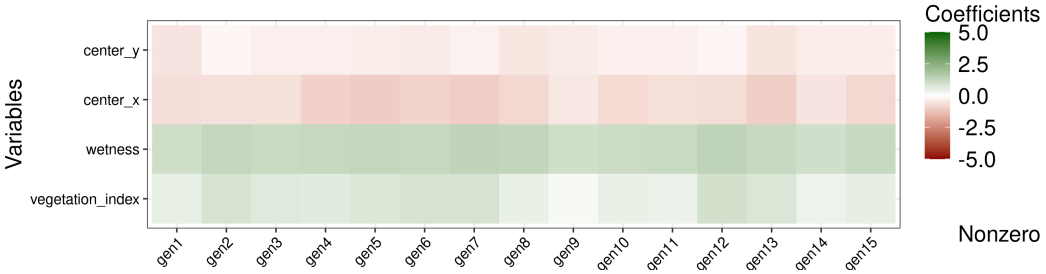


Figure 1 – Estimation des coefficients avec PLN.

- Les effets des variables ne sont pas les mêmes selon les genres

Données Genus [1] : résultats d'estimation avec PLN

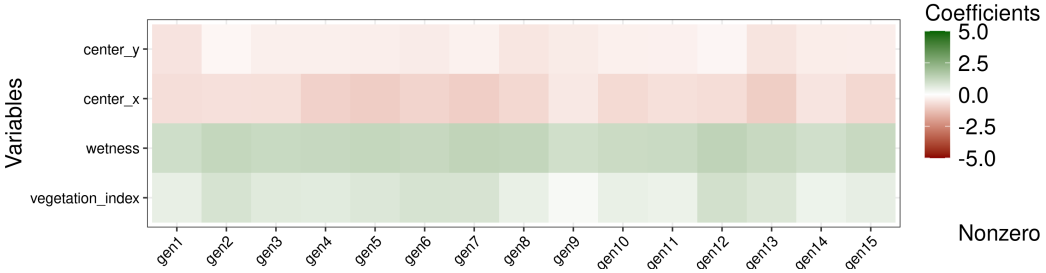


Figure 1 – Estimation des coefficients avec PLN.

- Les effets des variables ne sont pas les mêmes selon les genres
- Pas de sélection avec PLN : toutes les variables ont un effet non nul

Données Genus [1] : résultats d'estimation avec PLN

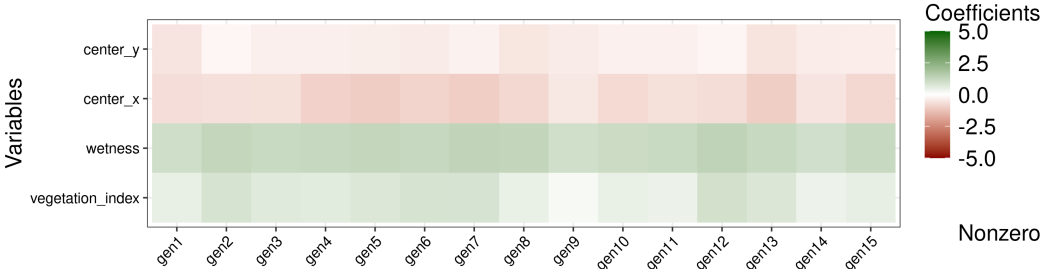


Figure 1 – Estimation des coefficients avec PLN.

- Les effets des variables ne sont pas les mêmes selon les genres
- Pas de sélection avec PLN : toutes les variables ont un effet non nul
- Possibilité de redondance d'informations et difficile à interpréter

Données Genus [1] : résultats d'estimation avec PLN

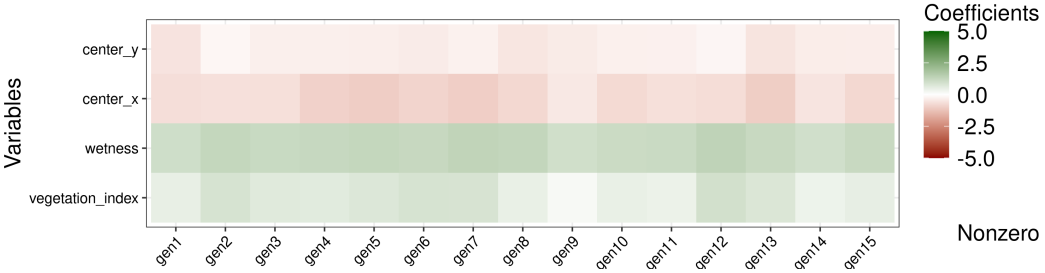


Figure 1 – Estimation des coefficients avec PLN.

- Les effets des variables ne sont pas les mêmes selon les genres
- Pas de sélection avec PLN : toutes les variables ont un effet non nul
- Possibilité de redondance d'informations et difficile à interpréter
- Recherche de parcimonie : identifier un sous ensemble de variables

Données Genus [1] : résultats d'estimation avec GLMNET

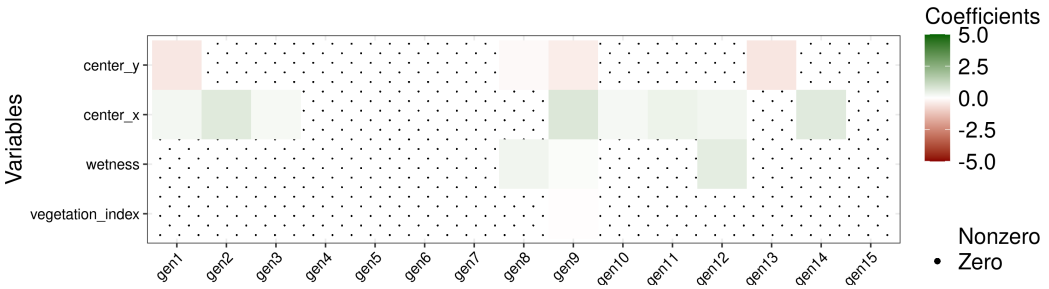


Figure 2 – Estimation des coefficients avec GLMNET.

Données Genus [1] : résultats d'estimation avec GLMNET

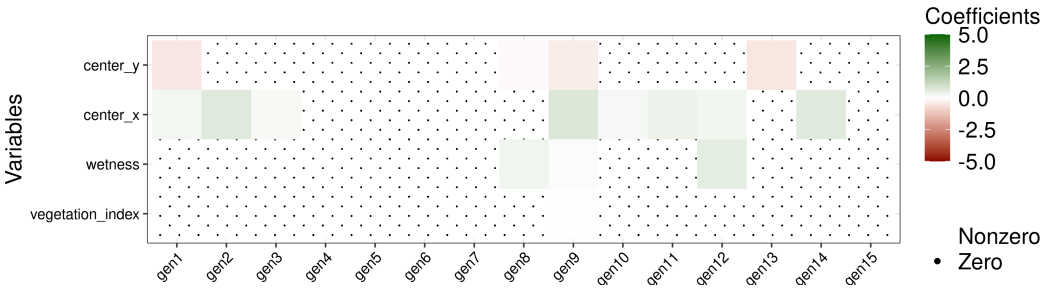


Figure 2 – Estimation des coefficients avec GLMNET.

- GLMNET ignore les relations de dépendance

Données Genus [1] : résultats d'estimation avec GLMNET

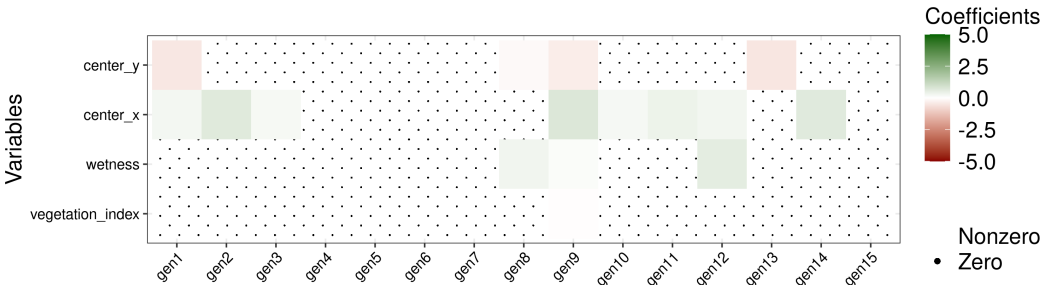


Figure 2 – Estimation des coefficients avec GLMNET.

- GLMNET ignore les relations de dépendance
- Permet d'obtenir de la parcimonie

Données Genus [1] : résultats d'estimation avec GLMNET

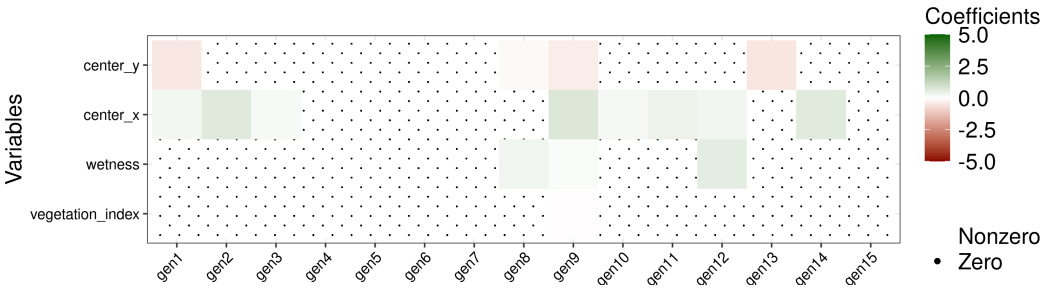


Figure 2 – Estimation des coefficients avec GLMNET.

- GLMNET ignore les relations de dépendance
- Permet d'obtenir de la parcimonie
- L'abondance de certaines espèces n'est pas en lien avec les variables

Données Genus [1] : résultats d'estimation avec SICPLN

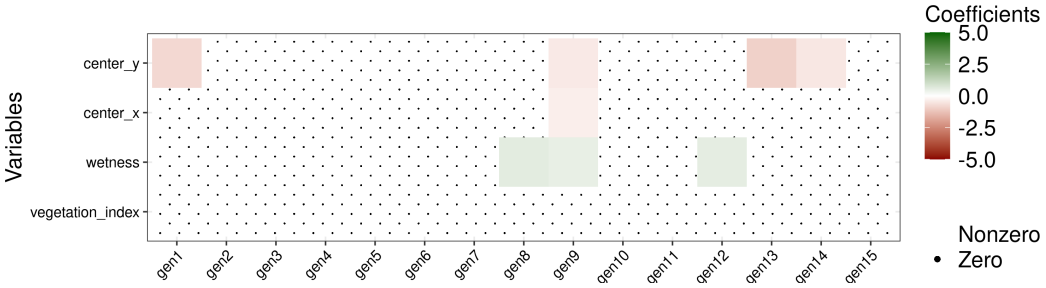


Figure 3 – Estimation des coefficients avec SICPLN.

Données Genus [1] : résultats d'estimation avec SICPLN

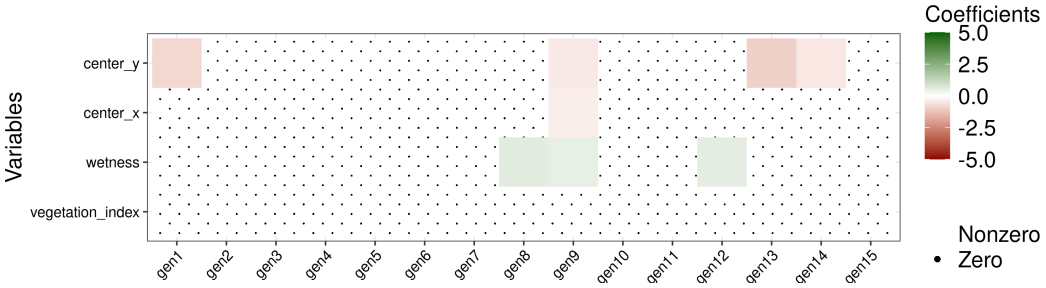


Figure 3 – Estimation des coefficients avec SICPLN.

- SICPLN prend en compte les relations de dépendance

Données Genus [1] : résultats d'estimation avec SICPLN

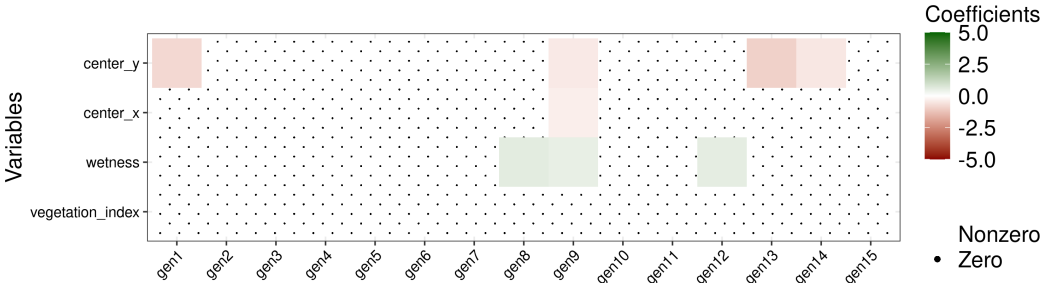


Figure 3 – Estimation des coefficients avec SICPLN.

- SICPLN prend en compte les relations de dépendance
- Plus parcimonieux que GLMNET

Données Genus [1] : résultats d'estimation avec SICPLN

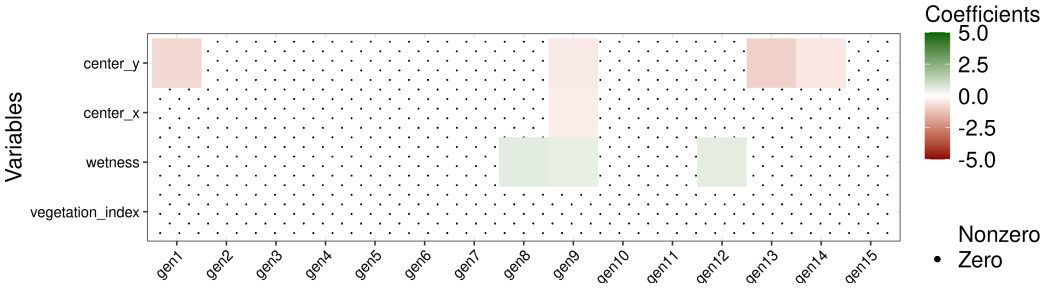


Figure 3 – Estimation des coefficients avec SICPLN.

- SICPLN prend en compte les relations de dépendance
- Plus parcimonieux que GLMNET
- Explication simplifiée de l'abondance

Données Hunting spider [4]

Données Hunting spider [4] : résultats d'estimation avec PLN

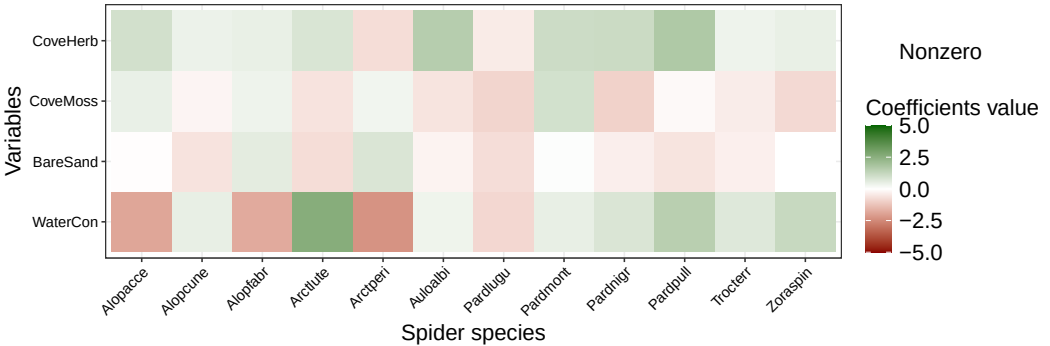


Figure 4 – Estimation des coefficients avec PLN.

- Pas de sélection de variables

Hunting spider [4] : résultats d'estimation avec GLMNET

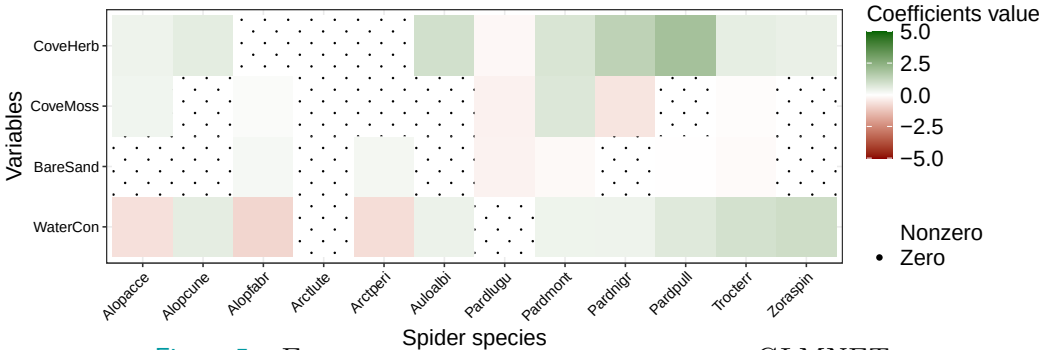


Figure 5 – Estimation des coefficients avec GLMNET.

- Plus de parcimonie que PLN (17 coefficients nuls)
- coveHerb, coveMoss et waterCon sont sélectionnées pour alopacce

Hunting spider [4] : résultats d'estimation avec SICPLN

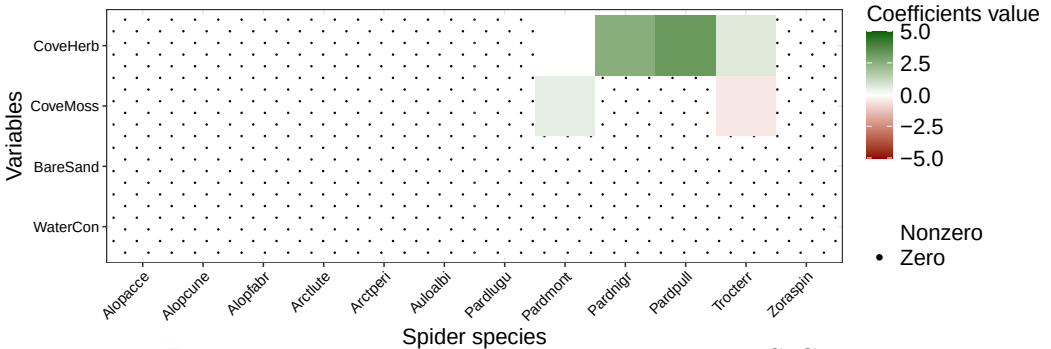


Figure 6 – Estimation des coefficients avec SICPLN.

- Plus de parcimonie que GLMNET (41 coefficients sont nuls)
- Espèce pardmont : coveHerb et coveMoss sont sélectionées
- Espèce alopcune : Aucune variable n'a été sélectionnée

Application en cours avec l'UMRF : étude des communautés microbiennes dans le processus de production du lait

Application sur les données de l'**UMRF**

Données Amont Saint-Nectaire

- Écosystèmes : air, eau, trayon, fecès, lait, filtre, litière
- Saison : hiver et été
- Lieu de collecte : 14 fermes
- Abondance : $n = 539$,$p = 1458$
- Covariables : $n = 539$, $d = 257$

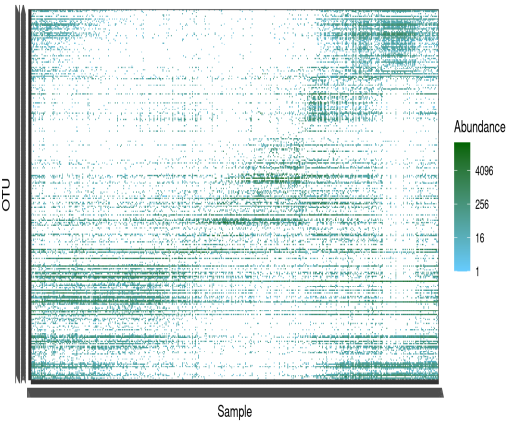


Figure 7 – Matrice d'abondance.

Communautés microbiennes provenant de l'air en hiver

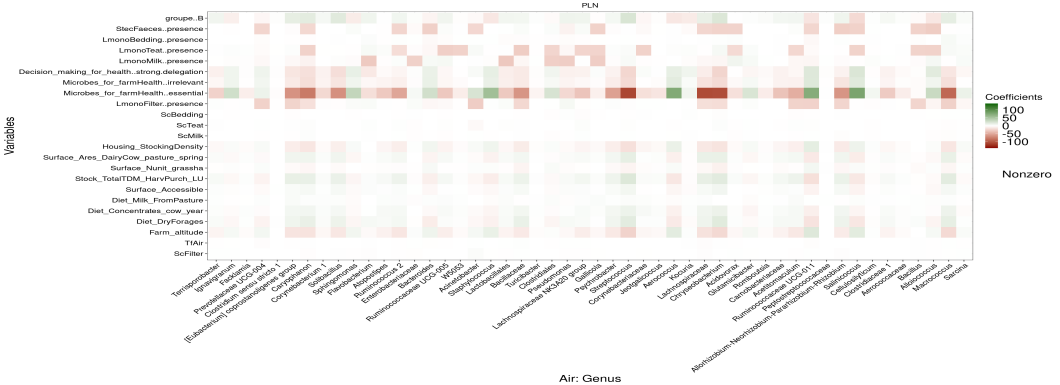


Figure 8 – Estimations avec PLN.

Communautés microbiennes provenant de l'air en hiver

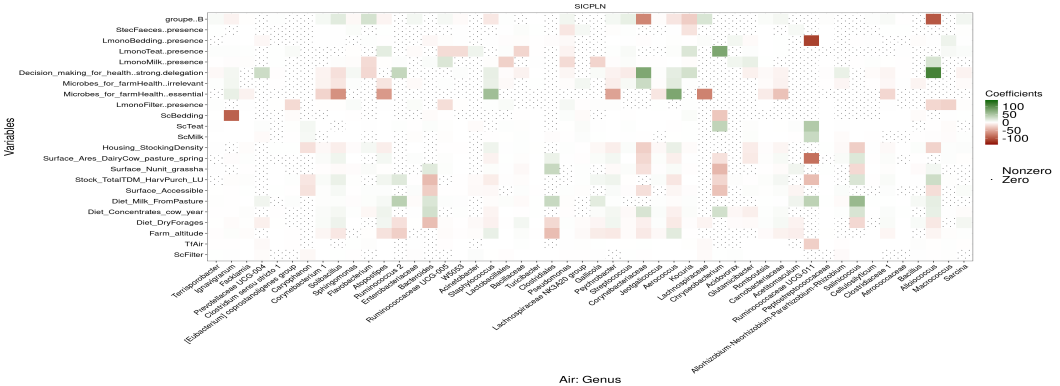


Figure 9 – Estimations avec SICPLN.

Données simulées : résultats de l'estimation

Table 1 – Estimations des coeffecients avec PLN, GLMNET et SICPLN.

Espèces	Méthode d'estimations	x_1	x_2	x_3	x_4	x_5	x_6
Espèce 1	Vrai coefficient	0	1	1	1	1	0
	PLN	0.08	1.04	1.09	1.07	1.10	0.09
	GLMNET	0	0.91	0.99	0.93	0.96	0
	SICPLN	0	0.95	1	0.98	0.92	0
Espèce 2	Vrai coefficient	0.5	0	0	1	1	0
	PLN	0.55	0.07	0.15	1.04	0.99	0.13
	GLMNET	0.43	0	0	0.98	0.87	0
	SICPLN	0.47	0	0	0.98	0.92	0
Espèce 3	Vrai coefficient	1	0.5	0.5	1	1	0
	PLN	1.10	0.58	0.54	1.05	1.06	0.10
	GLMNET	0.97	0.39	0.43	1.05	0.84	0
	SICPLN	1	0.48	0.44	0.96	0.97	0
Espèce 4	Vrai coefficient	1	1	0	0	0.5	0
	PLN	0.91	0.95	0.05	0.10	0.52	0.02
	GLMNET	0.64	1.14	0.10	-0.14	0.14	-0.21
	SICPLN	0.94	0.98	0	0	0.54	0

Conclusion

Résumé

- Formulations et interprétations du SIC
- Extension du SIC au modèle PLN
- Identification des variables pertinentes par approximation progressive de la norme L_0
- Sélection basée sur la maximisation du critère d'information de Bayes (BIC)

Travaux en cours et futurs

- Finalisation de l'application aux données UMRF
- Implémentation du SIC pour le modèle PLN zero-inflated
- Version non paramétrique

Merci de votre attention ! ! !

"Tous les modèles sont faux, mais certains sont utiles." George E. P. Box

Réferences I

- Jocelyn Chauvet, Catherine Trottier et Xavier Bry. « Component-Based Regularization of Multivariate Generalized Linear Mixed Models ». In : *Journal of Computational and Graphical Statistics* 28.4 (2019), p. 909-920.
- 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.
- 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.

Réferences II

 Nellie Smeenk-Enserink et PJM Van der Aart. « Correlations between distributions of hunting spiders (Lycosidae, Ctenidae) and environmental characteristics in a dune area ». In : *Netherlands Journal of Zoology* 25.1 (1974), p. 1-45.

Géométrie du SIC : biais induit

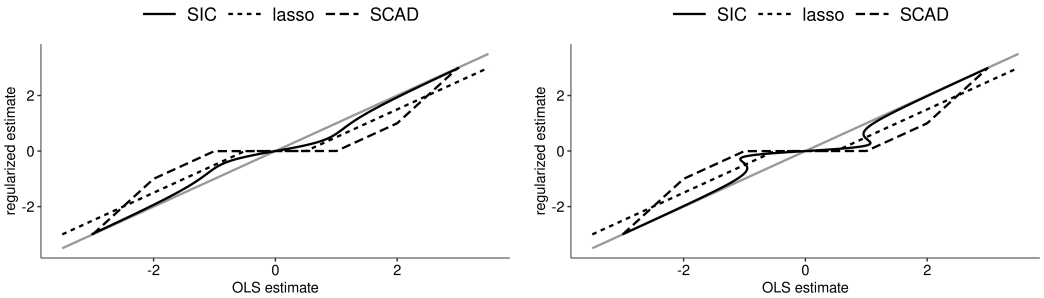


Figure 10 – Graphique des estimateurs de seuillage par rapport aux estimateurs des moindres carrés. La ligne grise est $y = x$. Le paramètre de réglage λ est fixé à 1 pour tous les estimateurs de seuils. La méthode SCAD implique un hyperparamètre de réglage supplémentaire, qui est fixé à $a = 3$. Le niveau d'approximation ε pour les graphiques sont 0.8 (gauche) et 0.4 (droite).

Données genus : matrice de précision

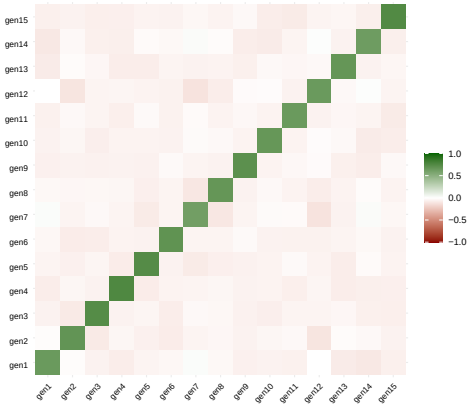


Figure 11 – Genus precision matrix

Données hunting spider : matrice de précision

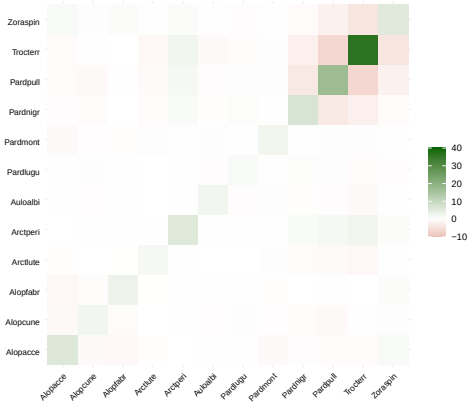


Figure 12 – Spider precision matrix