Inférence de réseaux d'associations à l'échelle de groupes à partir de données d'abondance avec le modèle PLN-Block

Jeanne Tous [1] & Julien Chiquet [2]

[1] *Université Paris-Saclay, AgroParisTech, INRAE, UMR MIA Paris-Saclay, 91120, Palaiseau, France, julien.chiquet@inrae.fr*
[2] *Université Paris-Saclay, AgroParisTech, INRAE, UMR MIA Paris-Saclay, 91120, Palaiseau, France, jeanne.tous@inrae.fr*

Résumé. Les réseaux d'association constituent un outil utile en écologie pour identifier des relations entre espèces qui ne peuvent être expliquées par les variables environnementales observées, et peuvent donc aider à la compréhension du fonctionnement de systèmes complexes. De tels réseaux peuvent être inférés à partir de données d'abondance, comme des comptages d'espèces en écologie, grâce à PLN-network [Chiquet, Mariadassou et Robin, 2019], une méthode d'inférence de réseaux fondée sur un modèle Poisson-log-normal couplée à une procédure d'estimation de type GLASSO [Friedman et al., 2007]. Cependant, lorsque le volume de données et le nombre d'espèces étudiés augmentent, les réseaux obtenus sont complexes à étudier puisque les associations entre espèces sont identifiées avec des degrés de confiance variés et leur interprétation individuelle est sujette à caution. Il existe des métriques pour agréger l'information contenue dans ces réseaux (nombre total d'associations, intensité moyenne de celles-ci, nombre de cliques dans le réseaux...). Cependant, ces métriques résument la complexité des réseaux à un faible nombre d'informations, nécessairement réductrices. Un compromis entre l'échelle très fine de chaque association et l'échelle globale de ces métriques peut être obtenu par un clustering sur les sommets du graphe et l'inférence d'un réseau à l'échelle des clusters plutôt que celle des espèces. De tels groupes rassembleraient des sommets dont les positionnements vis-à-vis du reste du réseau semblent similaires. Nous proposons le modèle PLN-Block, une méthode d'inférence de réseaux dérivée de PLN-network, qui vise à simultanément effectuer un clustering des espèces étudiées et inférer une structure de réseau entre les clusters, à partir de données d'abondance. Nous introduisons ici ce modèle, discutons les résultats du clustering et les améliorations qui peuvent lui être apportées, notamment pour mieux tenir compte de la nature spécifique des données écologiques.

Mots-clés. Modèles poisson-log-normaux ; clustering ; inférence variationnelle ; biodiversité ; réseaux d'associations.

Abstract. Association networks are a useful tool in fields such as ecology to identify relationships that cannot be explained by observed environmental factors and thus help understand the functioning of complex systems. Such networks can be inferred from abundance data, such as measured presence in ecology thanks to PLN-network [Chiquet, Mariadassou and Robin, 2019], a Poisson-log-normal based network inference method associated with an estimation procedure like GLASSO [Friedman et al., 2007]. However, as observed data and the number of studied species get more numerous, the resulting network becomes very tedious to study since pairwise associations can be identified with various degree of confidence and edges' individual interpretation can be questionable. Measures exist to summarize them thanks to a few metrics, such as the overall number of associations, their intensity, the number of cliques... but these metrics, on the other hand, reduce a very complex structure to a few simple figures, necessarily wanting. A compromise between the micro-scale of pairwise associations and the macro-scale of such measures could be to cluster graph vertices and infer a network at the scale of these clusters. Such groups would gather vertices that seem to behave similarly in regard to their positioning and associations profiles in the network. We introduce PLN-block, a network inference method derived from PLN-network that aims at simultaneously clustering studied species and derive a network structure between these clusters from abundance data. We introduce the model and the associated inference method, discuss the clustering results and how the model could be improved to better adapt to the specificity of ecological data.

Keywords. Poisson log-normal models, clustering, variational inference, biodiversity, association networks.

1 Introduction

Les réseaux d'association sont des objets mathématiques utiles pour décrire les relations entre espèces en écologie. Ils peuvent être inférés à partir de données d'abondance d'espèces : les relations de dépendance entre ces abondances sont mesurées par des outils de type corrélation et les dépendances qui ne peuvent être expliquées par les seules covariables sont identifiées comme des associations. Dans

certains cas, ces associations peuvent être vues comme des interactions, mais souvent elles n'en sont que des approximations, et l'interprétation de ces réseaux est plus limitée puisque les associations identifiées peuvent aussi être expliquées par des réactions similaires à des facteurs environnementaux non mesurés [Poggiato et. al, 2021].

L'inférence de réseaux menée dans notre cadre de travail se fonde sur des modèles graphiques non orientés [Lauritzen, 1996] ou champs aléatoires de Markov [Harris, 2016]. Dans ce cadre, les espèces i et j sont considérées comme liées dans le réseau si et seulement si leurs abondances sont conditionnellement dépendantes, au sens du modèle statistique défini, étant données les covariables et les abondances des autres espèces. Inférer un réseau d'association revient alors à calculer des corrélations partielles entre des variables liées aux abondances de chaque espèce, comme nous le détaillerons dans la description de PLN-block.

Du fait de la nature des données d'abondance, à savoir des données de comptage, le modèle statistique considéré ici doit être adapté aux données discrètes, ce qui n'est pas le cas des modèles graphiques gaussiens classiques. PLN-network [Chiquet, Mariadassou, Robin, 2019] est un modèle conçu pour tenir compte de la spécificité des données de comptage pour inférer des réseaux peu denses. Les données y sont modélisées par une distribution Poisson log-normale [Aitchison et Ho, 1989], tenant compte des covariables potentielles et intégrant une variable latente pour modéliser les dépendances résiduelles qui définissent le réseau. L'ajout d'une contrainte de sparsité sur le réseau reconstruit permet de sélectionner seulement les dépendances les plus fortes et de limiter la sélection de celles qui semblent les moins certaines.

Cependant, le réseau obtenu peut être complexe à interpréter, particulièrement lorsqu'un grand nombre d'espèces est étudié. En effet, il s'agit d'un objet complexe, informatiquement lourd à manipuler, et riche en informations. De plus, les associations résultantes doivent être considérées avec précaution dans la mesure où elles sont le résultat d'une procédure statistique et sont identifiées avec des degrés de confiance divers qui dépendent du modèle et de la sparsité du réseau. Le choix de l'hyper-paramètre qui définit le niveau de sparsité imposé dans le réseau permet ainsi de modifier le niveau de certitude avec lequel les associations sont retenues. Il existe des métriques pour agréger les informations contenues dans le réseau (nombre total d'associations, intensité moyenne de celles-ci...) mais elles semblent en revanche très réductrices au vu de la complexité des objets étudiés.

Une autre méthode consisterait à effectuer un clustering des nœuds du réseau, fondé sur leur profil d'associations avec les autres nœuds. Un tel résultat peut être

obtenu avec des Stochastic Block Models (SBM), introduits par Holland et al. en 1983 [Holland et al., 1983]. Dans leur formulation la plus simple, il s'agit de modéliser des associations binaires par un modèle de mélange : les associations suivent une loi de Bernouilli dont le paramètre est entièrement défini par les groupes auxquels appartiennent chacun des nœuds considérés. De nombreuses variations de ce modèle existent, par exemple pour tenir compte des distributions d'interactions possibles (gaussienne, Poisson...) de l'évolution de la composition des groupes [Yang et al., 2011], [Matias et Miele, 2017] ou des incertitudes sur les données observées [Rebafka et al., 2019]. Cependant, les SBM sont appliqués à des réseaux déjà reconstruits, de sorte que les résultats du clustering ne peuvent pas fournir d'information pour nourrir la construction initiale du réseau à l'échelle des clusters.

Avec PLN-Block, nous proposons d'effectuer le clustering des espèces observées et l'inférence d'un réseau d'associations à l'échelle de ces clusters simultanément grâce à une procédure d'optimisation alternée.PLN-block est directement dérivé de PLN-network par l'ajout d'une couche latente supplémentaire pour décrire les groupes auxquels les espèces appartiennent, dont découlent leurs probabilités d'association, dans la logique de modèle de mélange des SBM. Pour l'inférence des paramètres du modèle, nous utilisons une méthode d'inférence variationnelle, fondée sur le calcul d'une approximation de la fonction de vraisemblance.

Nous présentons le modèle PLN-block dans la section 2. La section 3 décrit la méthode d'inférence utilisée, tandis que la 4 présente les résultats de premières simulations. Enfin nous discutons des possibilités d'amélioration du modèle dans la section 5.

2 Modèle

Dans ce modèle,

- on considère p espèces observées dans n sites, et l'on suppose qu'elles se divisent en Q groupes (Q est un hyper-paramètre) ;
- la variable $Y_{i,j}$ décrit l'abondance (le nombre d'individus observés) de l'espèce j dans le site i ;
- $X_i \in \mathbb{R}^m$ est le vecteur des covariables du site i et B_j le vecteur de paramètres de régression correspondant pour l'espèce j, supposé indépendant du site con-

sidéré. On note $X \in \mathcal{M}_{n,m}(\mathbb{R})$ la matrice dont la i-ème ligne est X_i, et $B \in \mathcal{M}_{m,p}(\mathbb{R})$ la matrice dont la j-ème colonne est B_j ;

- la variable latente $Z_i \in \mathbb{R}^Q$ décrit la structure de dépendance entre les groupes d'espèces sur le site i : pour tout $i, Z_i \sim \mathcal{N}(0, \Sigma)$;
- la variable latente C_j indique le groupe auquel l'espèce j appartient. Pour tout $j, C_j \sim \mathcal{M}(1, \alpha)$ avec $\alpha = (\alpha_q)_{1 \leq q \leq Q}$ et $\sum_{q=1}^{Q} \alpha_k = 1$; α_q est donc la probabilité pour une espèce d'appartenir au groupe q ; on note $\alpha = (\alpha_q)_{1 \leq q \leq Q}$ et $C_{j,q} = \mathbb{1}_{C_j = q}$, $C_{j,q}$ vaut 1 si l'espèce j appartient au groupe q, 0 sinon ;
- $o_{i,j}$ décrit un offset pour l'espèce j sur le site i, pour permettre, le cas échéant, de tenir compte des différences d'efforts d'échantillonnage entre sites / espèces.

On suppose que chaque $Y_{i,j}$ suit, conditionnellement à Z_i et C_j une distribution Poisson log-normale dont le paramètre est déterminé par les covariables et par le $Z_{i,q}$ correspondant au groupe q auquel appartient l'espèce j :

$$Y_{i,j} | Z_i, C_j \sim \mathcal{P}(\exp(x_i^T B_j + o_{i,j} + \sum_{q=1}^{Q} Z_{i,q} C_{j,q})).$$

Le réseau entre les groupes est donc traduit par la variable latente Z ou, plus exactement, par sa matrice de précision $\Omega = \Sigma^{-1}$. En effet, la corrélation partielle entre Z_{q_1} et Z_{q_2} est donnée par $\rho_{q_1,q_2} = \frac{-\Omega_{q_1,q_2}}{\sqrt{\Omega_{q_1,q_1} \Omega_{q_2,q_2}}}$. Enfin, on peut ajouter une contrainte de sparsité sur Ω afin de contrôler la densité du réseau. Pour ce faire, on ajoute une pénalité ℓ_1 sur les termes non diagonaux de Ω, multipliée par un hyper-paramètres λ qui *in fine* contrôle le nombre de corrélations partielles non nulles, donc le nombre d'arêtes du réseau inféré.

3 Méthode d'inférence

On note $\theta = (B, \Omega, \alpha)$ les paramètres du modèle. Il n'existe pas de forme explicite de la vraisemblance complète du modèle, $\sum_C \int_{-\infty}^{+\infty} p_\theta(Y, Z, C) dZ$ et il n'est donc pas possible de calculer directement θ par maximum de vraisemblance. Une simple stratégie EM [Dempster et al., 1977] ne résout pas non plus le problème car il n'existe pas de formule explicite pour calculer les moments de $p_\theta(Z, C|Y)$.

On opte donc pour une stratégie d'inférence variationnelle. Pour approcher la distribution conditionnelle $p_\theta(Z, C|Y)$ on fait une approximation de champ moyen

[Blei et al., 2017]. On cherche donc π_ψ dans $\Pi = \{\pi_{\psi_1}(Z)\pi_{\psi_2}(C)\}$ ($\psi = \{\psi_1, \psi_2\}$ désigne l'ensemble des paramètres variationnels) pour approcher $p_\theta(Z, C|Y)$, avec :

- π_{ψ_1} : pour tout $i, \pi_{\psi_1}(Z_i) \sim \mathcal{N}(M_i, S_i)$, avec $S_i = diag(s^2_{i,q})_{q\in[\![1;Q]\!]}$ diagonale pour tout i. On note $S \in \mathcal{M}_{n,K}(\mathbb{R})$ définie par $S_{i,q} = s^2_{i,q}$ et $M \in \mathcal{M}_{n,K}(\mathbb{R})$ définie par $M_{i,q} = M_{i_q}$ de sorte que $\psi_1 = (M, S)$.

- π_{ψ_2} : pour tout $j, C_j \sim \mathcal{M}(1, (\tau_{q,j})_{1\leq q\leq Q})$ avec pour tout $j, \sum_{q=1}^{Q} \tau_{q,j} = 1$. On note $\tau \in \mathcal{M}_{Q,p}(\mathbb{R})$ la matrice définie par $(\tau_{q,j})_{j\in[\![1;p]\!],q\in[\![1;Q]\!]}$ de sorte que $\psi_2 = (\tau)$.

On fait également l'hypothèse que, sous π_ψ les Z_i sont independants entre eux, de même que les C_j. La tâche de clustering revient donc à trouver $\max((\tau_{q,j})_{1\leq q\leq Q})$ pour tout $1 \leq j \leq p$.

Dans le cadre de l'approximation variationnelle, on peut définir une vraisemblance approximative (ELBO) à maximiser :

$$\begin{aligned} J(Y;\theta,\psi) &= \log(p_\theta(Y)) - \mathrm{KL}(\pi_\psi(Z,C)||p_\theta(Z,C|Y)) \\ &= \mathbb{E}_\pi(p_\theta(Y,Z,C)) - \mathbb{E}_\pi(\log(\pi_{\psi_1}(Z))) - \mathbb{E}_\pi(\log(\pi_{\psi_2}(C))) \end{aligned}$$

Les calculs permettent d'aboutir à une expression matricielle explicite pour J.

Dans le cas où on ajoute une contrainte de sparsité sur Ω, on peut définir une nouvelle fonction objectif : $J_2 = J - \lambda\|\Omega\|_{\ell_1,\text{off}}$ où $\lambda\|\Omega\|_{\ell_1,\text{off}}$ désigne la norme de Ω considérée sans ses termes diagonaux (qui traduisent les variances au sein des groupes mais ne donnent pas d'informations sur les corrélations entre ces derniers), et λ est un hyper-paramètre.

La procédure d'inférence consiste ensuite à mettre alternativement à jour B, M, S, Ω et τ, α pour maximiser J ou $J2$, dans un cadre d'EM variationnel que l'on ne détaille pas ici.

4 Simulations

Les simulations effectuées jusqu'à présent visent d'abord à tester les capacités de clustering de l'algorithme d'inférence. Nous avons donc simulé des données sous le modèle décrit en section 2 et évaluer les capacités de l'algorithme à identifier

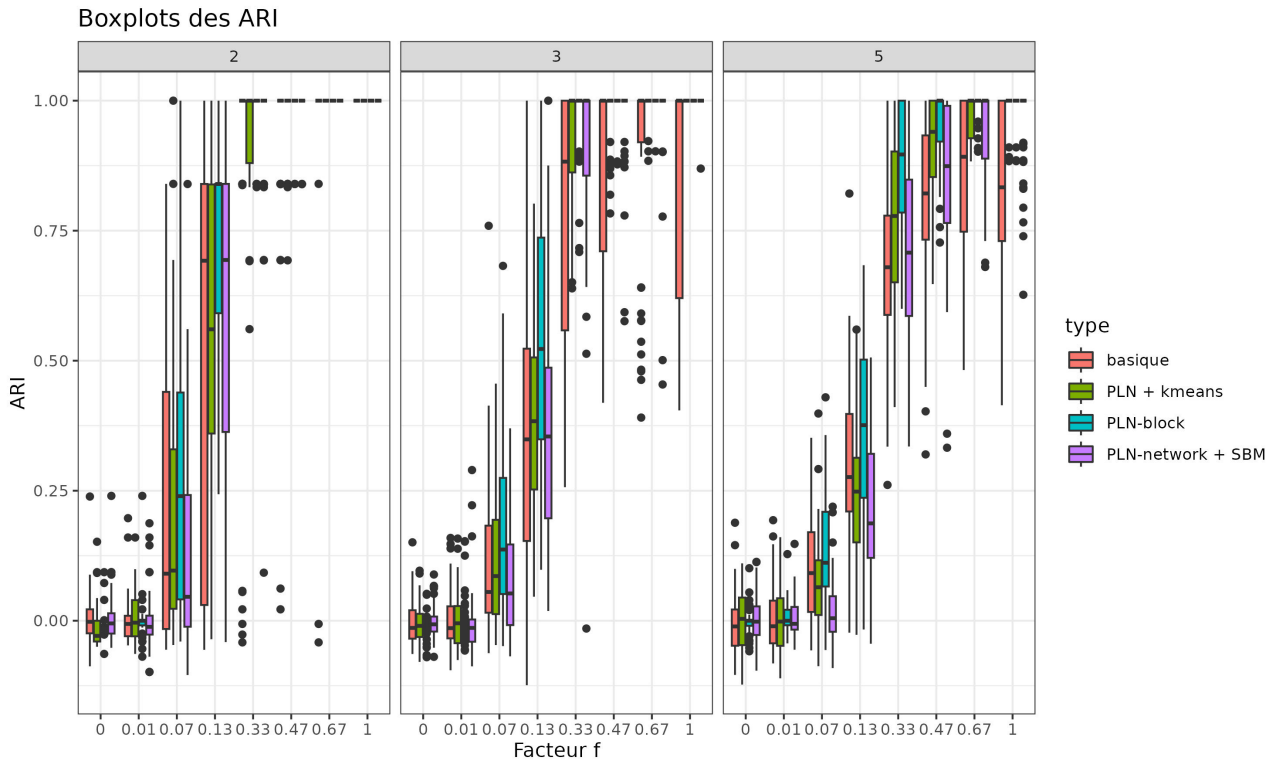

Figure 1: ARI des clusterings effectués sur des données simulées sous PLN-block avec $p = 25$, $n = 50$, $q = 2$, 3 ou 5. Abscisse : facteur par lequel est multipliée la matrice Σ réputée "facile".

correctement ces clusters par l'average rand index (ARI) entre les vrais clusters et les clusters inférés.

Nous avons comparé les ARI obtenus avec notre algorithme à ceux obtenus par différentes méthodes :

1. Une méthode "basique" qui consiste à effectuer une régression de Poisson sur les données suivi d'un clustering k-means sur les colonnes de résidus.

2. Une méthode "PLN + k-means" qui consiste à appliquer un modèle PLN simple aux données suivi d'un clustering k-means sur les résidus, obtenus grâce aux moyennes variationnelles de la variable latente Z.

3. Une méthode "PLN-network + SBM" : on reconstruit un réseau à l'échelle des

espèces avec PLN-network puis on lui applique un modèle SBM pour en tirer des clusters.

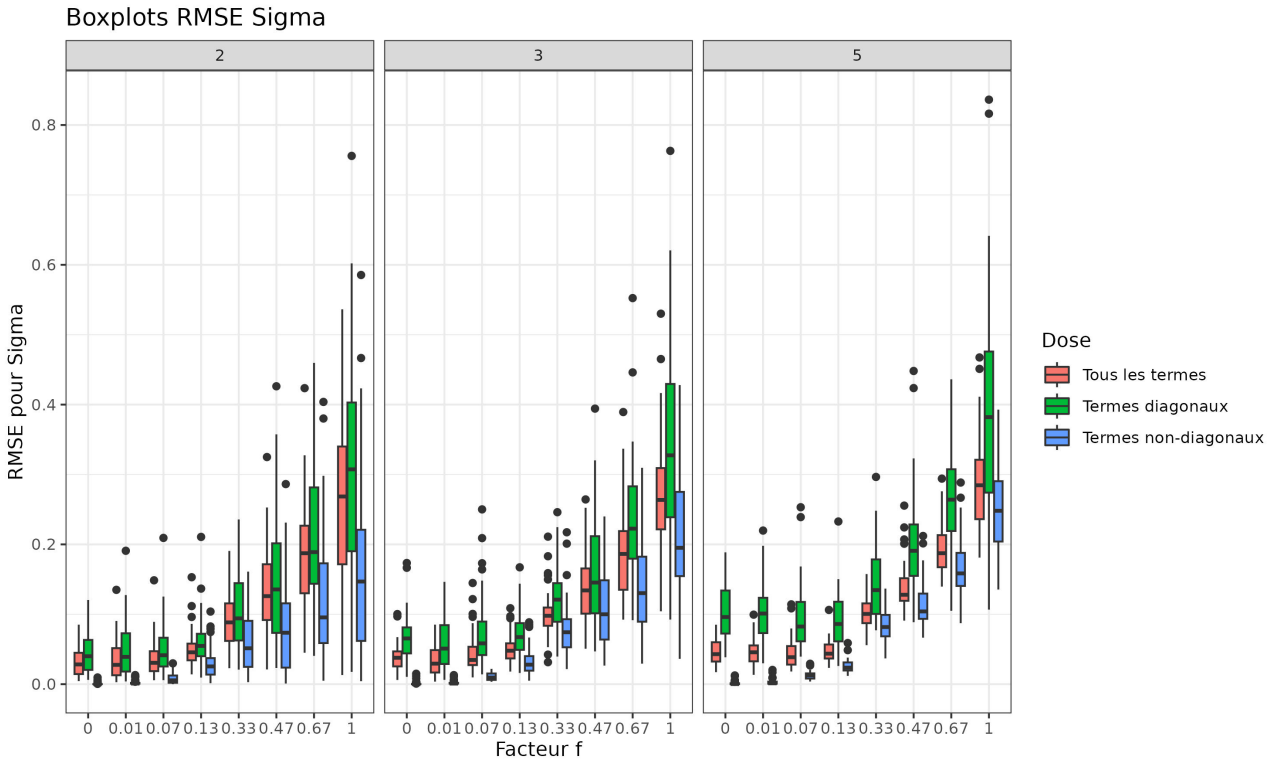

Figure 2: RMSE de Σ sur des données simulées sous PLN-block avec $p = 25$, $n = 50$, $q = 2$, 3 ou 5. Abscisse : facteur par lequel est multipliée la matrice Σ initiale.

Nous proposons ici de premières simulations avec $p = 25$ espèces, $n = 50$ sites. On tire aléatoirement les covariables environnementales (en deux dimensions) contenues dans X entre 0 et 1, et les paramètres de régression associés, contenus dans B, entre -1 et 1. Nous effectuons les tests pour $q = 2, 3$ ou 5 groupes. Pour chaque q, on a créé à la main une matrice définie positive Σ par laquelle on a cherché à mettre des contrastes importants entre les corrélations afin de rendre le clustering plus facile. Ensuite, nous avons simulé avec Σ multipliée par un facteur f entre 0 et 1. L'objectif est de rendre la tâche de clustering plus difficile lorsque f se rapproche de 0 car l'effet des groupes sur la variable Y observée est alors moins fort par rapport à celui des covariables environnementales. Les données sont simulées sous le modèle PLN-block.

La figure 1 montre les résultats de cette simulation. On constate que la tâche de clustering se complexifie effectivement beaucoup lorsque f se rapproche de 0, et ce quelle que soit la méthode employée. La méthode PLN-block semble plus robuste que les autres dès que f dépasse 0.07, constat plus marqué pour $q = 5$. L'algorithme d'optimisation semble donc aboutir aux résultats souhaités pour le clustering, dès lors que l'effet des groupes est suffisamment important. La figure 2 montre en revanche une augmentation de la RMSE pour l'estimation de Σ à mesure que f augmente, ce qui pourrait s'expliquer par une plus grande variance dans l'estimation de plus grandes valeurs de Σ.

Nous présenterons oralement des résultats concernant les réseaux inférés par PLN-block et les effets de l'hyper-paramètre de sparsité.

5 Discussion

PLN-block propose un modèle d'inférence de réseaux à l'échelle de groupes identifiés par ce même modèle. Les premiers résultats sur la tâche de clustering montrent la capacité du modèle à inférer correctement les clusters dès lors que leur effet sur les observations dépasse un certain seuil, qui semble dépendre du nombre de clusters. Cependant la robustesse du modèle dans des cas plus réalistes, pour lesquels des effets spécifiques des points du graphes s'additionnent à ceux des groupes n'est pas assurée. En particulier, il semblerait judicieux d'intégrer dans le modèle un effet spécifique des points (en plus de l'effet des clusters) sur la variance. Sans cela, , il se peut que des points soient regroupés dans le même cluster en raison de variances similaires (diagonale de Ω) et non en raison de corrélations similaires, ce qui est la visée de PLN-block. Nous travaillons actuellement à cette généralisation.

Dans un second temps, d'autres développements de PLN-block pourraient intégrer une dimension temporelle pour tenir compte de l'évolution des groupes ou des associations entre eux, ou encore d'y ajouter des *a priori* écologiques pour guider les tâches de clustering et d'inférence de réseau.

Bibliographie

Aitchison, J., et Ho, C. H. (1989), The multivariate Poisson-log normal distribution, *Biometrika*, 76, pp 643-653.

Blei, D. M., Kucukelbir, A., et McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518), 859-877.

Chiquet, J. Mariadassou, M. et Robin, S. (2019) Variational Inference of Sparse Network from Count Data, *Proceedings of Machine Learning Research*, 97, pp 1162-1171.

Chiquet, J. Mariadassou, M. et Robin, S. (2021) The Poisson-lognormal model as a versatile framework for the joint analysis of species abundances., *Frontiers in ecology and evolution*, 9, p 588292.

Dempster, A. P., Laird, N. M., et Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society: series B (methodological)*, 39(1), 1-22.

Friedman, J., Hastie, T. et Tibshirani, R. (2007) Sparse inverse covariance estimation with the graphical lasso, *Biostatistics*, 9(3), pp 432-441.

Harris, D.J. (2016) Inferring species interactions from co-occurrence data with Markov networks, *Ecology*, 97, pp 3308-3314.

Holland, P. W., Laskey, K. B. et Leinhardt, S.(1983) Stochastic block models: First steps, *Social networks*, 5, pp 109-137.

Lau, M. K., Borrett, S. R., Baiser, B., Gotelli, N. J. et Ellison, A. M.(2017) Ecological network metrics: opportunities for synthesis, *Ecosphere*, 8.

Lauritzen,S.L. (1996) *Graphical Models*, 17.

Matias, C. et Miele, V.(2017) Statistical clustering of temporal networks through a dynamic stochastic block model, *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 79(4), pp 1119-1141.

Poggiato, G., Münkemüller, T., Bystrova, D., Arbel, J., Clark, J.S. et Thuiller, W. (2021) On the Interpretations of Joint Modeling in Community Ecology, *Trends in Ecology and Evolution*, 2806.

Rebafka, T., Roquain, E. et Villers, F. (2019) On the Graph inference with clustering and false discovery rate control *rXiv preprint arXiv:1907.10176*.

Yang, T., Chi, Y., Zhu, S., Gong, Y. et Jin, R.(2011) Detecting communities and their evolutions in dynamic social networks—a Bayesian approach, *Machine learning*, 82, pp 157-189.