

PLN-TREE: UN MODÈLE LATENT D'INFÉRENCE DE RÉSEAUX SUR DES ARBRES D'ABONDANCE AU SEIN DU MICROBIOTE

Alexandre Chaussard¹ & Anna Bonnet¹ & Sylvain Le Corff¹ & Harry Sokol²

¹ *Laboratoire de Probabilités, Statistique et Modélisation, Sorbonne Université, France, chaussard@lpsm.paris*

² *CHU Saint-Antoine Assistance publique Hôpitaux de Paris (AP-HP), Sorbonne Université, France*

Résumé. Le microbiote intestinal est un écosystème complexe composé principalement de bactéries qui interagissent entre elles et avec leur environnement. Si la composition du microbiote s'est avérée être un biomarqueur pertinent pour plusieurs maladies, la structure spécifique du microbiote intestinal a été peu prise en compte à ce jour. En effet, la composition microbienne est décrite par des données discrètes et parcimonieuses qui s'organisent selon une structure d'arbre taxonomique. Plus précisément, chaque bactérie appartient à plusieurs groupes qui sont hiérarchiquement ordonnés du plus précis (espèce) au moins précis (domaine). Bien que cette information taxonomique soit connue, on ignore encore son rôle dans l'impact d'une bactérie sur son hôte. Des travaux récents (Chiquet et al. 2019) ont proposé un cadre permettant de modéliser des données de comptage parcimonieuses pour étudier des interactions entre espèces, mais dans lequel l'impact de la structure taxonomique n'a pas encore été étudié. Dans ce travail, nous visons donc à étendre les travaux existants afin d'établir un cadre et un modèle dédié aux données de comptage s'appuyant sur une structure taxonomique. Notre objectif est de proposer des méthodes interprétables pour expliquer l'interaction complexe entre les espèces bactériennes, avec une application dans le contexte des maladies inflammatoires de l'intestin.

L'approche proposée, basée sur le modèle Poisson log-normal (PLN), tient compte des dépendances markoviennes pour inclure la structure de l'arbre taxonomique reliant les bactéries tout en permettant aux taxons de différentes branches de s'influencer mutuellement. Cela permet non seulement d'étendre le champ d'application de la modélisation des interactions, mais aussi de s'aligner sur la complexité et la diversité inhérentes aux communautés microbiennes. En outre, nous présentons une nouvelle approche variationnelle qui incorpore la structure markovienne de la distribution *a posteriori* afin d'améliorer la précision de l'estimation variationnelle.

Nous avons ensuite appliqué notre méthode aux données issues d'une cohorte de patients diagnostiqués avec la maladie de Crohn. Le modèle est systématiquement comparé à des modèles alternatifs, ce qui permet une analyse complète de sa capacité à détecter des interactions microbiennes complexes associées à des conditions pathologiques.

Keywords. modèles latents, PLN, microbiote, données compositionnelles, arbre

Abstract. The gut microbiota is a complex ecosystem composed mostly of bacteria interacting with each other and their environment. If the composition of the microbiota has

proven to be a relevant biomarker for several diseases, the specific structure of gut microbiota has poorly been taken into account. Indeed, the microbial composition is described by discrete and sparse data, which also have a taxonomic structure. More precisely, each bacterium belongs to several groups that are hierarchically ordered from more precise (species) to less precise (domain). Although this taxonomic information is known, it remains unclear how it is related to the impact of a bacterium on its host. While some recent works (Chiquet et al. 2019) proposed a framework that accounts for the sparse and discrete nature of the microbiota, the impact of the taxonomic structure has not been investigated yet. In this work, we aim at introducing a new framework and dedicated algorithms that account for the sparse taxonomic abundance nature of the microbiota data. Our focus is to propose interpretable methods to unveil the complex interplay among bacterial species, with an application in the context of inflammatory bowel diseases.

The proposed approach, based on Poisson Log-Normal models, accounts for Markov dependencies to include the taxonomic tree structure linking the bacteria while enabling taxa of different branches to influence each other. This not only extends the scope of interaction modeling but also aligns with the inherent complexity and diversity of microbial communities. Additionally, we present a novel variational approach that incorporates the Markovian structure of the posterior distribution to enhance the precision of the variational estimation. Such variational families allow us to obtain theoretical guarantees on latent state estimation and data reconstruction.

To assess real-world applications, we conduct an empirical investigation on a cohort of patients diagnosed with Crohn’s disease. The model is systematically benchmarked against state-of-the-art algorithms, providing a comprehensive analysis of its efficiency in capturing complex microbial interactions associated with pathological conditions.

Keywords. latent models, PLN, microbiota, count data, tree

1 Introduction

Le microbiote intestinal humain est un écosystème complexe de 10^{13} micro-organismes résidant dans le tractus gastro-intestinal et qui joue un rôle clé dans la santé de l'hôte. Parmi les micro-organismes qui le composent, les bactéries constituent une part majoritaire, exerçant des effets profonds sur la physiologie, le métabolisme et les fonctions immunitaires de l'hôte. La compréhension des relations complexes au sein du microbiome intestinal est une étape cruciale vers la compréhension de son impact sur la santé humaine et la prédiction des résultats des traitements [Julien et al., 2022]. Dans cette optique, les progrès des technologies de séquençage à haut débit ont permis de caractériser les communautés microbiennes.

Parmi les différentes représentations du microbiote intestinal, nous nous concentrons sur les échantillons de taxa-abondance, qui permettent de représenter la composition d'un microbiome en combinant les données de comptage et les informations taxonomiques. Étant donné que la taxonomie fournit des informations sur les relations génétiques entre les espèces, cette structure d'arbre peut apporter des informations précieuses à la modélisation du nombre d'espèces. Cependant, les échantillons d'abondance taxonomique sont sujets à une grande variabilité inter-individuelle, ce qui complexifie le problème de la modélisation et de l'inférence. Outre la tâche de modélisation, nous visons à fournir un outil interprétable pour la recherche médicale en mettant l'accent sur les interactions entre espèces qui pourraient mieux expliquer le rôle des troubles du microbiome. Dans ce contexte, nous présentons un nouveau modèle pour les données de comptage associées à une taxonomie, PLNtree, conçu pour estimer les interactions microbiennes tout en modélisant les échantillons de taxa-abondance.

Le modèle PLNtree s'appuie sur le cadre Poisson Log-Normal (PLN), développé par [Aitchison and Ho, 1989] puis largement étudié par [Chiquet et al., 2021, Chiquet et al., 2019], qui modélise les données de comptage et les interactions entre espèces au moyen d'une variable latente gaussienne. Notre contribution consiste à étendre le modèle PLN et les méthodes d'inférence associées pour tenir compte de la structure arborescente inhérente à la taxonomie microbienne. En outre, pour garantir la flexibilité et l'adaptabilité, le modèle PLNtree incorpore une chaîne de Markov latente indépendante de la structure taxonomique qui capture la dynamique et les interactions sous-jacentes au sein de la communauté microbienne. Cette caractéristique permet d'explorer les relations cachées entre tous les taxons d'un niveau. Simultanément, les comptes observés suivent une chaîne de Markov contrainte qui adhère à la structure hiérarchique de la taxonomie, préservant ainsi l'intégrité de la composition des données de taxa-abondance.

De plus, comme l'apprentissage du modèle PLNtree nécessite l'utilisation d'un proxy variationnel, nous fournissons une approximation variationnelle "backward" amortie produisant des résultats affinés face à l'approche "mean-field".

Afin de valider l'efficacité du modèle PLNtree, nous présentons des résultats expérimentaux basés sur une cohorte de patients diagnostiqués avec la maladie de Crohn. Nous comparons le modèle PLNtree à des concurrents sur la tâche de génération en utilisant le critère de diversité alpha.

2 Cadre mathématique

2.1 Modèle Poisson Log-Normal model

La recherche d’interactions entre entités peut être formulée comme un problème d’inférence de réseau, pour lequel les méthodes canoniques sont basées sur les Modèles Graphiques Gaussiens [Altenbuchinger et al., 2020]. Toutefois, l’hypothèse de normalité ne s’applique pas aux données d’abondance. Des approches statistiques ont donc été proposées pour analyser les données d’abondance [Inouye et al., 2017]. En particulier, le modèle Poisson Log-Normal [Aitchison and Ho, 1989] a conduit à de nombreux développements théoriques et méthodologiques [Chiquet et al., 2019] dans l’inférence de réseau d’interactions pour les données de comptage. Le modèle PLN appartient à la famille des modèles à variables latentes, car il modélise l’abondance des entités en utilisant une distribution de Poisson paramétrée par une variable aléatoire gaussienne latente.

Soit $\mathbf{X} \in \mathbb{R}^{n \times d}$ une matrice d’abondance de d entités parmi n sites. Le modèle PLN repose sur les hypothèses suivantes.

- H1**
- $\mathbf{Z}_i \sim \mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$ où $\boldsymbol{\mu}_i \in \mathbb{R}^d, \boldsymbol{\Sigma} \in \mathbb{R}^{d \times d}$.
 - $(\mathbf{Z}_i, \mathbf{X}_i)_{1 \leq i \leq n}$ sont indépendants, et conditionnellement à \mathbf{Z} et $X_k, 1 \leq k \neq j \leq d$, X_j dépend de Z_j uniquement et $X_j \sim \mathcal{P}(e^{Z_j})$.

On note $p_{\mathbf{Z}, \boldsymbol{\theta}}$ la densité de \mathbf{Z} où $\boldsymbol{\theta} = \{\boldsymbol{\Sigma}, (\boldsymbol{\mu}_i)_{1 \leq i \leq n}\}$. Les modèles partiellement observés comme PLN font généralement usage de l’algorithme EM pour inférer le paramètre $\boldsymbol{\theta}$. Cependant, l’étape E nécessite le calcul d’une espérance conditionnelle sous la distribution $p_{\boldsymbol{\theta}}(\mathbf{Z}|\mathbf{X})$ qui n’est pas explicite dans le cadre de PLN. Les auteurs de [Chiquet et al., 2021] suggèrent une approximation variationnelle de l’étape E en optimisant une quantité sous-optimale pour la log-vraisemblance appelée ELBO [Kingma et al., 2019]:

$$\text{ELBO}(\boldsymbol{\theta}, \boldsymbol{\varphi}) = \mathbb{E}_{q_{\boldsymbol{\varphi}}}[\log p_{\boldsymbol{\theta}}(\mathbf{X}|\mathbf{Z})] - \text{KL}[q_{\boldsymbol{\varphi}}(\cdot|\mathbf{X}), p_{\mathbf{Z}, \boldsymbol{\theta}}(\cdot)] . \quad (1)$$

L’approximation variationnelle choisie par [Chiquet et al., 2021] s’associe à l’approche champ moyen :

$$q_{\boldsymbol{\varphi}}(\mathbf{Z}_i|\mathbf{X}_i) = \prod_{i=1}^n \mathcal{N}(\mathbf{Z}_i; \mathbf{m}_i, \mathbf{S}_i) ,$$

où $\mathcal{N}(\cdot; \mathbf{m}_i, \mathbf{S}_i)$ est la densité de la loi gaussienne de moyenne $\mathbf{m}_i \in \mathbb{R}^d$ et de variance $\mathbf{S}_i = \text{diag}(\mathbf{s}_i)$ avec $\mathbf{s}_i \in \mathbb{R}^d$ et où $\boldsymbol{\varphi}$ représente tous les paramètres variationnels. En conséquence, l’EM variationnel génère une séquence de paramètres $(\boldsymbol{\theta}^h, \boldsymbol{\varphi}^h)_{1 \leq h \leq N}$ obtenus en réalisant une optimisation alternée consistant d’abord à optimiser l’ELBO relativement à $\boldsymbol{\varphi}$ tel que

$$\begin{aligned} \boldsymbol{\varphi}^{h+1} &= \arg \max_{\boldsymbol{\varphi}} \text{ELBO}(\boldsymbol{\theta}^h, \boldsymbol{\varphi}) \\ \text{t.q. } &\forall i \in \{1, \dots, n\}, \mathbf{S}_i \succ 0 . \end{aligned}$$

L’approximation variationnelle choisie rend l’ELBO explicite, permettant ainsi une optimisation directe de $\boldsymbol{\varphi}$ [Chiquet et al., 2021, Chiquet et al., 2019]. Il devient ensuite possible

d'effectuer l'étape E en remplaçant la vraie loi *a posteriori* $p_{\theta}(\mathbf{Z}|\mathbf{X})$ par $q_{\varphi^{h+1}}(\mathbf{Z}|\mathbf{X})$, ce qui permet de calculer l'étape M pour obtenir θ^{h+1} , et ainsi de suite jusqu'à convergence.

Ce modèle s'est avéré très efficace pour adapter aux données de comptage un large spectre d'outils d'analyse multivariée tels que l'ACP, l'analyse discriminante et l'inférence de réseau. L'application d'intérêt pour notre problème est l'inférence de réseau, qui nous permettrait de découvrir les interactions fonctionnelles entre les bactéries du microbiome intestinal. Le problème d'inférence de réseau s'accompagne également d'une architecture informée par la parcimonie suggérée par [Chiquet et al., 2019], en introduisant une pénalisation de type LASSO sur la matrice de précision $\Omega = \Sigma^{-1}$.

Enfin, les auteurs soulignent le rôle clé des covariables et des offsets dans le problème d'inférence de graphe qui peuvent être intégrés dans la moyenne du modèle latent.

2.2 Top-Down PLNTree

Considérons maintenant que les données de comptage proviennent d'un arbre taxonomique connu, noté \mathcal{T} . Un exemple d'échantillon est fourni ci-dessous.

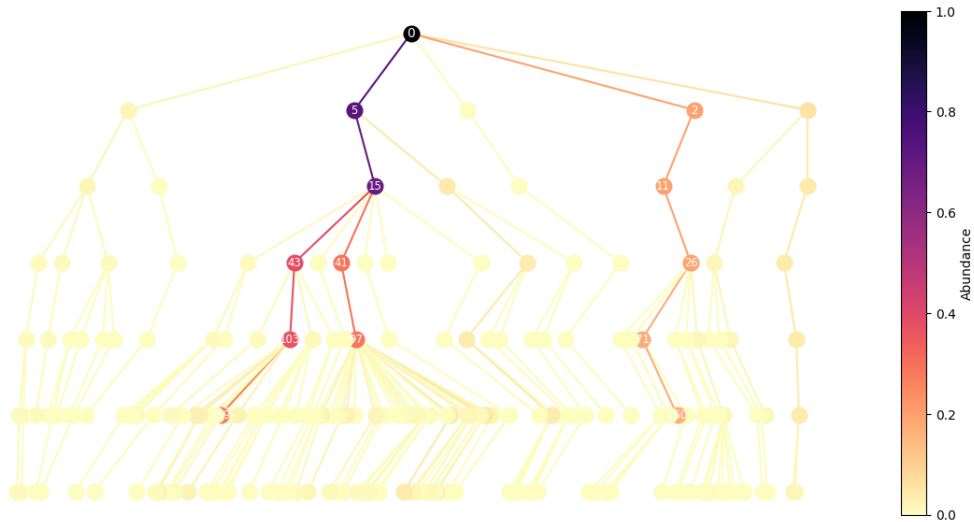


Figure 1: Exemple d'échantillon d'abondance de taxons (coloré selon l'abondance relative)

En supposant que \mathcal{T} comporte L couches, K_{ℓ} nœuds à la couche ℓ , nous désignons les échantillons d'abondance de taxons par

$$\mathbf{X} = (\mathbf{X}^1, \dots, \mathbf{X}^L),$$

où à la couche ℓ , $\mathbf{X}^{\ell} = (\mathbf{X}_1^{\ell}, \dots, \mathbf{X}_{K_{\ell}}^{\ell})$. Pour un nœud donné X_k^{ℓ} , nous introduisons la notation $\mathcal{C}(X_k^{\ell})$ qui est l'ensemble des enfants de ce nœud. La notation \mathcal{C}_{ℓ}^k renvoie à l'ensemble des indices des enfants du nœud X_k^{ℓ} .

Pour modéliser la structure séquentielle de l'arbre à travers ses couches, nous introduisons une dynamique de Markov dans l'espace latent et dans l'espace observé tout en restant dans le cadre du modèle PLN. Supposons que l'ensemble d'hypothèses suivant soit vérifié :

- H2** – Les $(\mathbf{Z}_i, \mathbf{X}_i)_{1 \leq i \leq n}$ sont indépendants, et conditionnellement à \mathbf{Z} et $\mathbf{X} \setminus \mathcal{C}(X_k^\ell)$, $1 \leq \ell \leq L, 1 \leq k \leq K_\ell$, l'ensemble des variables $\mathcal{C}(X_k^\ell)$ ne dépend que de $\mathcal{C}(Z_k^\ell)$ et de X_k^ℓ .
- Dynamique Markovienne latente (de haut en bas) :

$$\begin{aligned} \mathbf{Z}^1 &\sim \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) , \\ \mathbf{Z}^{\ell+1} &\sim \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{\theta}_{\ell+1}}(\mathbf{Z}^\ell), \boldsymbol{\Sigma}_{\boldsymbol{\theta}_{\ell+1}}(\mathbf{Z}^\ell)) , \end{aligned}$$

où $\boldsymbol{\mu}_{\boldsymbol{\theta}_{\ell+1}}(\cdot)$ et $\boldsymbol{\Sigma}_{\boldsymbol{\theta}_{\ell+1}}(\cdot)$ désignent des fonctions arbitraires paramétrées par $\boldsymbol{\theta}_{\ell+1}$.

- Dynamique observée contrainte :

$$\begin{aligned} \mathbf{X}^1 &\sim \mathcal{P}(e^{\mathbf{Z}^1}) , \\ \mathcal{C}(X_k^\ell) &\sim \mathcal{P}\left(e^{\mathcal{C}(Z_k^\ell)} \middle| \sum_{j \in \mathcal{C}_k^\ell} X_j^{\ell+1} = X_k^\ell\right) . \end{aligned}$$

Approximation Markov Backward. Comme dans le cas du modèle PLN, nous n'avons pas accès à un estimateur du maximum de vraisemblance. Au lieu de cela, comme suggéré par [Chiquet et al., 2021] dans le modèle PLN, nous effectuons une approximation variationnelle pour estimer la distribution a posteriori non-explicite. Sous H2, la distribution de lissage visée $p_\theta(\mathbf{Z}|\mathbf{X})$ est une chaîne de Markov "backward" :

$$p_\theta(\mathbf{Z}|\mathbf{X}) = p_\theta(\mathbf{Z}^L|\mathbf{X}^{1:L}) \prod_{\ell=1}^{L-1} p_\theta(\mathbf{Z}^\ell|\mathbf{Z}^{\ell+1}, \mathbf{X}^{1:\ell}) .$$

Puisque nous approchons la quantité ci-dessus à l'aide d'une approximation variationnelle, nous suggérons une famille variationnelle qui tient compte de la structure "backward" de la vraie distribution de lissage :

- H3** La loi variationnelle "backward" s'écrit :

$$q_\varphi(\mathbf{Z}|\mathbf{X}) = q_{\varphi^L}(\mathbf{Z}^L|\mathbf{X}^{1:L}) \prod_{\ell=1}^{L-1} q_{\varphi^\ell}(\mathbf{Z}^\ell|\mathbf{Z}^{\ell+1}, \mathbf{X}^{1:\ell}) ,$$

où

- $q_{\varphi^L}(\mathbf{Z}^L|\mathbf{X}^{1:L}) = \mathcal{N}(\mathbf{Z}^L; \mathbf{m}_{\varphi^L}(\mathbf{X}^{1:L}), \mathbf{S}_{\varphi^L}(\mathbf{X}^{1:L}))$
- $q_{\varphi^\ell}(\mathbf{Z}^\ell|\mathbf{Z}^{\ell+1}, \mathbf{X}^{1:\ell}) = \mathcal{N}(\mathbf{Z}^\ell; \mathbf{m}_{\varphi^\ell}(\mathbf{Z}^{\ell+1}, \mathbf{X}^{1:\ell}), \mathbf{S}_{\varphi^\ell}(\mathbf{Z}^{\ell+1}, \mathbf{X}^{1:\ell}))$

- Pour tout ℓ , $\mathbf{S}_{\varphi^\ell}$ produit une matrice diagonale qui peut être représentée par sa diagonale.

Manipuler les entrées des réseaux employés en pratique est une tâche difficile, sans formulation évidente. Nous suggérons une configuration expérimentale amortie utilisant un réseau de neurones récurrent, inspirée de récents travaux [Chagneux et al., 2022], de sorte que $\mathbf{X}^{1:\ell}$ soit incorporé sous une forme contractée pour neutraliser la nature expansive de la chaîne.

3 Résultats numériques

Dans cette section, nous visons à montrer les performances empiriques du modèle taxa-PLN, que nous souhaitons comparer à celles de PLN afin de mesurer l'intérêt d'intégrer la taxonomie au modèle. Considérons une étude pratique d'une cohorte de 732 patients atteints de la maladie de Crohn.

Approximation champ moyen. L'approximation variationnelle régulièrement utilisée dans la littérature est l'approximation champ moyen (mean-field) [Kingma et al., 2019], qui s'exprime comme suit.

H4 L'approximation fonctionnelle du champ moyen Gaussien est donnée par

$$q_\varphi(\mathbf{Z}|\mathbf{X}) = \prod_{\ell=1}^L q_{\varphi^\ell}(\mathbf{Z}^\ell|\mathbf{X}^\ell),$$

où $q_{\varphi^\ell}(\mathbf{Z}^\ell|\mathbf{X}^\ell) \sim \mathcal{N}(\mathbf{m}_{\psi_\ell}(\mathbf{X}^\ell), \mathbf{S}_{\psi_\ell}(\mathbf{X}^\ell))$, tel que $\mathbf{m}_{\psi_\ell}(\mathbf{X}^\ell) \in \mathbb{R}^{K_\ell}$, et $\mathbf{S}_{\psi_\ell}(\mathbf{X}^\ell) = \text{diag}(\mathbf{s}_{\psi_\ell}(\mathbf{X}^\ell))$ avec $\mathbf{s}_{\psi_\ell}(\mathbf{X}^\ell) \in \mathbb{R}_+^{K_\ell}$.

En utilisant l'approximation champ moyen, nous obtenons un premier candidat de comparaison avec l'approximation "backward" de PLNTree. Bien qu'il soit plus facile d'entraîner un modèle champ moyen, cette approximation peut cependant restreindre trop fortement le choix de la famille variationnelle par rapport à la vraie distribution *a posteriori*.

Benchmarks génératifs. Nous commençons par évaluer la capacité du modèle à générer des échantillons de qualité, dans le sens où ils sont proches de l'échantillon initial qui a servi à apprendre les paramètres du modèle. Pour ce faire, nous apprenons le modèle PLNTree sur l'ensemble des données et nous générons des données artificielles de taxa-abondance (MF : Champ moyen, Amortized : Chaîne de Markov backward amortie). Nous procédons à une procédure similaire en utilisant le modèle PLN à chaque couche. Toutefois, empiler les comptages obtenus via PLN ne fournit pas des échantillons plausibles, car ils ne respectent pas les contraintes de taxa-abondance. Par conséquent, nous fournissons également une variante PLN (fill) qui permet de générer des échantillons de taxa-abondance valides en

partant uniquement de la dernière couche, puis en remontant l’arbre par les contraintes déterministes de la taxonomie.

Les critères de comparaison des générations correspondent aux mesures de diversité alpha couramment utilisées dans l’analyse du microbiote, à savoir, l’entropie de Shannon, Inverse Simpson ou encore l’indice Chao1 [Chao, 1984].

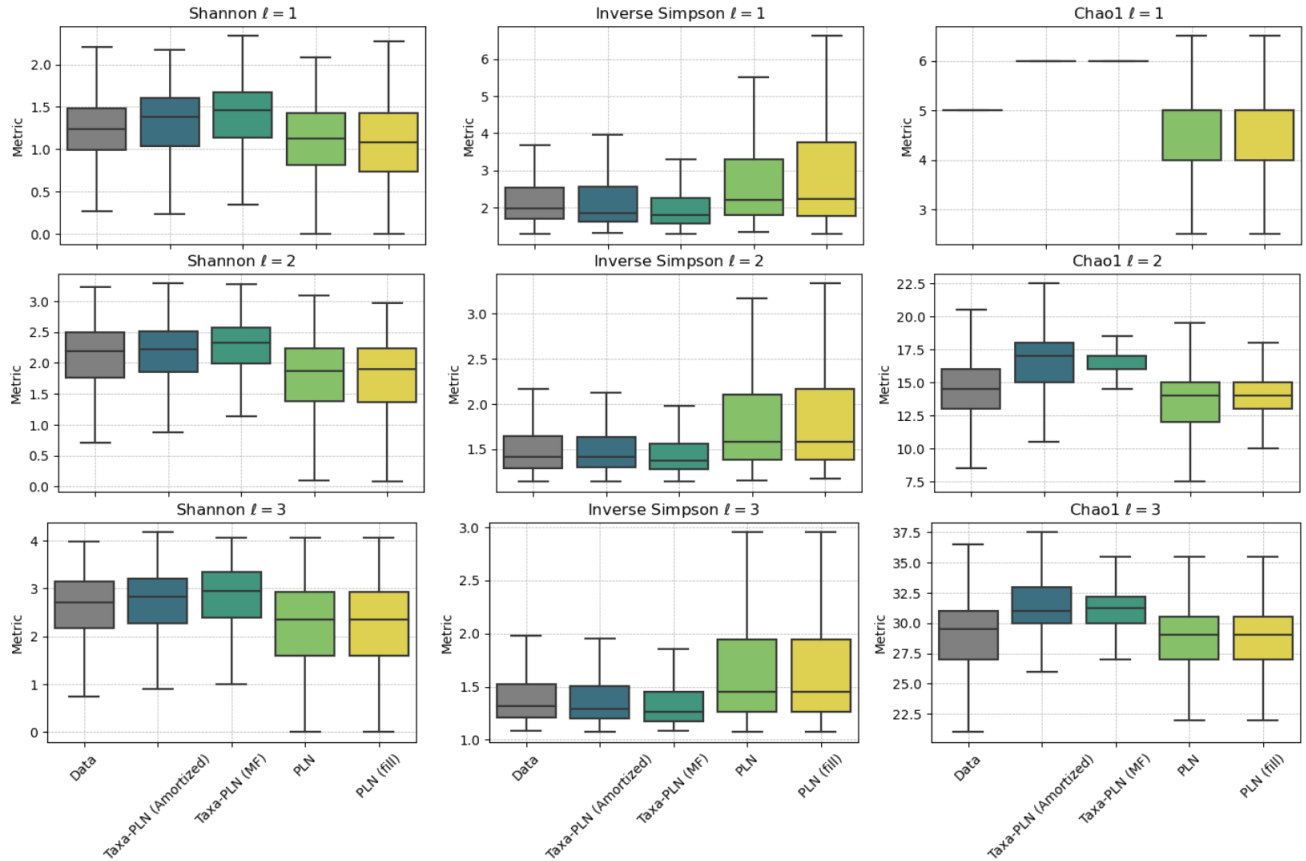


Figure 2: Benchmarks de génération de diversité alpha sur Suivitheque pour chaque couche.

En regardant la figure 2, il semble que l’exploitation de l’information taxonomique ait amélioré la qualité générative du modèle. En regardant l’indice Chao1, on peut supposer que les échantillons générés souffrent de comptes spurieux (comptes proches de 0) qui augmentent la diversité Chao1, ouvrant la voie à d’autres améliorations. Les figures 2 et 3 ainsi que le tableau 1 montrent également les améliorations apportées par l’inférence backward amortie par rapport à l’approche champ moyen.

	Taxa-PLN (MF)	Taxa-PLN (Amortized)	PLN	PLN (fill)
$\bar{\rho}$	0.96 ± 0.13	0.99 ± 0.08	0.99 ± 0.04	0.99 ± 0.04

Table 1: Corrélation moyenne entre les échantillons originaux et après encodage puis décodage.

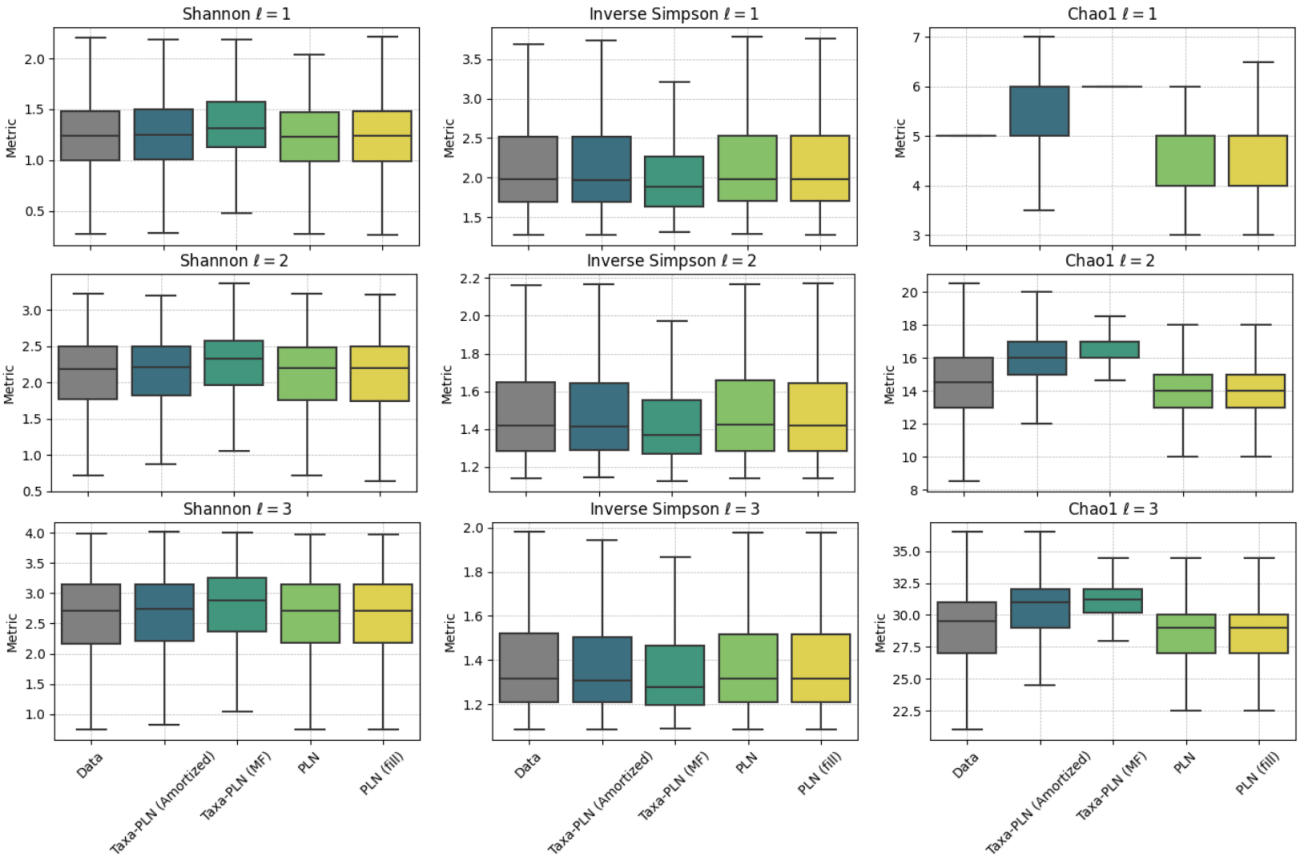


Figure 3: Benchmarks d'encodage de la diversité alpha sur Suivitheque pour chaque couche.

4 Conclusion

Le modèle PLNtree apparaît comme un modèle pertinent pour inférer les interactions microbiennes au sein du microbiome intestinal. Les gains de performance en alpha diversité proposés par PLNtree sur la tâche de génération montrent l'intérêt d'inclure l'information taxonomique dans l'inférence des interactions microbiennes. De plus, l'introduction d'une chaîne de Markov "backward" pour l'approximation variationnelle semble améliorer la précision du modèle, en montrant des résultats supérieurs au cas champ moyen. La perspective d'obtenir des garanties statistiques grâce à cette approximation ajoute une dimension prometteuse aux contributions potentielles du modèle PLNtree dans la compréhension des écosystèmes microbiens complexes.

Bibliography

References

- [Aitchison and Ho, 1989] Aitchison, J. and Ho, C. (1989). The multivariate poisson-log normal distribution. *Biometrika*, 76(4):643–653.
- [Altenbuchinger et al., 2020] Altenbuchinger, M., Weihs, A., Quackenbush, J., Grabe, H. J., and Zacharias, H. U. (2020). Gaussian and mixed graphical models as (multi-) omics data analysis tools. *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms*, 1863(6):194418.
- [Chagneux et al., 2022] Chagneux, M., Gassiat, É., Gloaguen, P., and Corff, S. L. (2022). Amortized backward variational inference in nonlinear state-space models. *arXiv preprint arXiv:2206.00319*.
- [Chao, 1984] Chao, A. (1984). Nonparametric estimation of the number of classes in a population. *Scandinavian Journal of statistics*, pages 265–270.
- [Chiquet et al., 2021] Chiquet, J., Mariadassou, M., and Robin, S. (2021). The poisson-lognormal model as a versatile framework for the joint analysis of species abundances. *Frontiers in Ecology and Evolution*, 9:588292.
- [Chiquet et al., 2019] Chiquet, J., Robin, S., and Mariadassou, M. (2019). Variational inference for sparse network reconstruction from count data. In *International Conference on Machine Learning*, pages 1162–1171. PMLR.
- [Inouye et al., 2017] Inouye, D. I., Yang, E., Allen, G. I., and Ravikumar, P. (2017). A review of multivariate distributions for count data derived from the poisson distribution. *Wiley Interdisciplinary Reviews: Computational Statistics*, 9(3):e1398.
- [Julien et al., 2022] Julien, C., Anakok, E., Treton, X., Nachury, M., Nancey, S., Buisson, A., Fumery, M., Filippi, J., Maggiori, L., Panis, Y., et al. (2022). Impact of the ileal microbiota on surgical site infections in crohn’s disease: a nationwide prospective cohort. *Journal of Crohn’s and Colitis*, 16(8):1211–1221.
- [Kingma et al., 2019] Kingma, D. P., Welling, M., et al. (2019). An introduction to variational autoencoders. *Foundations and Trends® in Machine Learning*, 12(4):307–392.