Sujet de Stage de recherche M2

Introduction de dépendance dans le modèle Poisson lognormal

Contacts: julien.chiquet@inrae.fr, mahendra.mariadassou@inrae.fr,
stephane.robin@sorbonne-universite.fr, bertrand.servin@inrae.fr

Contexte

Le modèle Poisson lognormal (PLN, voir [1]) décrit la distribution jointe d'un ensemble de p variables de comptages non indépendantes. L'espérance de chacun de ces comptages y dépend d'un ensemble de covariables via un terme de régression linéaire généralisée, tandis que la dépendance entre les comptages est modélisée au moyen d'un vecteur gaussien latent de dimension p . La structure de dépendance entre les p comptages y est donc décrite par la matrice de covariance Σ du vecteur gaussien latent. Ce modèle connaît de nombreuses applications en écologie [2], génétique ou en économie.

L'inférence des paramètres de ce modèle ne peut se faire au moyen d'un algorithme EM classique, mais des algorithmes efficaces fondés sur des approximations variationnelles sont disponibles [3].

Sujet

La version classique du modèle PLN suppose qu'on observe les n vecteurs de comptages (chacun de dimension p) indépendants. L'objectif de ce stage est d'étendre le modèle PLN afin de prendre en compte une dépendance entre les vecteurs de comptages. En fonction des applications, on pourra supposer que la structure de dépendance entre les observations est libre ou prend une forme spécifique. Il s'agira alors de développer l'algorithme d'inférence associé. Cet algorithme pourra *in fine* intégrer le package R/C++ `PLNmodels` [3].

Ce sujet est notamment motivé par deux applications. La première concerne la génétique des populations, dans laquelle les p comptages composant un vecteur sont issus de p individus éventuellement apparentés et où chaque vecteur de comptages est associé à un locus le long du génome. Il s'agit ici de prendre en compte la dépendance entre locus voisins (appelé déséquilibre de liaison). La deuxième concerne l'adaptation du modèle au cadre de l'*holobionte*, dans lequel on considère l'hôte et son microbiote comme une seule entité biologique. Il s'agit ici de prendre en compte l'apparentement entre hôtes (dépendances entre lignes) en plus des dépendances entre espèces microbiennes (déjà capturées par Σ).

Profil du/de la candidat(e)

Compétence en Statistique/Machine Learning, programmation (R et/ou Python).

Environnement

- Lieu : UMR MIA Paris-Saclay, Campus de Palaiseau
- Encadrants : Julien Chiquet, Stéphane Robin, Mahendra Mariadassou, Bertrand Servin
- Stage financé par le projet "*Behind the Count'her*" du programme de recherche Digit-Bio d'INRAE (*Biologie numérique pour explorer et prédire le vivant*)

Références

- [1] Aitchison, John, and C. H. Ho. "The multivariate Poisson-log normal distribution." *Biometrika* 76.4 (1989) : 643-653.
- [2] Chiquet, Julien, Mariadassou, Mahendra, and Robin, Stéphane. The Poisson-Lognormal Model as a Versatile Framework for the Joint Analysis of Species Abundances, *Frontiers in Ecology and Evolution*, 2021
- [3] `PLNmodels` : A collection of Poisson lognormal models for multivariate count data analysis, <https://github.com/jchiquet/PLNmodels>