Sujet de Stage de recherche M2

Analyse discriminante pour données de comptage multivariées : approche linéaire, quadratique et régularisée

Encadrants: Nicolas Jouvin, Guillem Rigaill, Julien Chiquet
Contact : nicolas.jouvin@inrae.fr

Contexte

L'analyse discriminante (voir [1]) est un outil classique de l'analyse multivariée qui permet de prédire l'appartenance à des groupes connus dans une population à l'aide d'un classifieur simple. L'avantage d'une telle approche, dite « model-based », est d'expliquer la structuration d'un jeu de données en populations et de fournir des éléments d'interprétation du phénomène d'étude, contrairement à des classifieurs issus de l'apprentissage automatique. Ce type d'approche est donc plébiscité dans les sciences du vivant, où le besoin en explicabilité est grand. Cependant, la formulation standard de l'analyse discriminante n'est pas adaptée à l'analyse de données de comptage multivariée, couramment rencontrées en écologie, génomique ou astrophysique.

Les modèles Poisson-lognormaux (PLN, voir [2]) fournissent un cadre générique pour la modélisation des données de comptage en s'appuyant sur une couche latente gaussienne. Une adaptation PLN de l'analyse discriminante dans sa version linéaire a été proposé dans [3] ainsi que les méthodes d'estimation associées, s'appuyant sur des approches variationnelles. Cependant, la version linéaire de l'analyse discriminante ne permet pas de répondre à un certain nombre de questions d'intérêt pour les applications, en particulier en biologie : est-ce que la structure de dépendances entre les sous-populations peut-être considérées comme identique ? Comment intégrer une structure de groupe sur les variables d'études dans les structures de dépendances ? Comment les comparer ?

Sujet

Le stage attaque ces questions en proposant une série de généralisations incrémentales de la version PLN de l'analyse discriminante linéaire (PLNLDA), en relachant l'hypothèse de covariance identique entre sous-populations : tout d'abord, une version quadratique de l'analyse discriminante, puis une version régularisée faisant l'hypothèse d'une structure de groupes entre les variables. Ces modèles feront l'objet de la mise en place d'un algorithme d'inférence de type variationnel-EM, et pourra s'appuyer sur les bibliothèques classiques d'optimisation ou d'autodifférentiation (`(py)torch`, `JAX`). Les algorithmes d'estimation développés pourront *in fine* intégrer le package `R/C++ PLNmodels` [4] ou sa version Python.

Les méthodes développées seront utilisées pour étudier un jeu de données issues du projet PEERSIM, coordonné par Guillem Rigaill et Étienne Delannoy, qui s'intéresse à l'étude des stress multiples chez les plantes dans un contexte de réchauffement climatique. Les modèles permettront d'analyser les changements d'expression des gènes (les variables), dont on observe des comptages dans différentes conditions de stress (les groupes) pour plusieurs plantes (les individus). Un des enjeux est de déterminer si les changements d'expression des gènes dans les différentes conditions de stress induisent des changements dans la moyenne de l'expression (décrits par les vecteurs de moyenne des différents groupes de

la population), mais également par des changements dans la structure de dépendance entre gènes (décrits pas les matrices de covariances des groupes de la population).

Profil du/de la candidat(e)

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

Conditions d'exercise

— Financement : MP Digit-Bio, projet PEERSIM
— Lieu : UMR MIA Paris-Saclay, Université Paris-Saclay, AgroParisTech INRAE, Campus Agro Paris Saclay
— Collaborateurs : Julien Chiquet, Guillem Rigaill, Nicolas Jouvin
— Début du stage : à déteminer avec le ou la candidat·e, possible dès mi-mars
— Durée du stage : 4 à 6 mois.

Comment candidater

Les candidat·e·s intéressé·e·s doivent envoyer un CV et une lettre de motivation à Nicolas Jouvin (nicolas.jouvin@inrae.fr).

Références

[1] Mardia, Kantilal Varichand, John T. Kent, and John M. Bibby. "Multivariate analysis." Probability and mathematical statistics (1979).

[2] Aitchison, John, and C. H. Ho. "The multivariate Poisson-log normal distribution." Biometrika 76.4 (1989) : 643-653.

[3] 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

[4] PLNmodels : A collection of Poisson lognormal models for multivariate count data analysis, https://github.com/jchiquet/PLNmodels