Étude comparative de lois a priori bayésiennes pour la sélection de variables dans les modèles non linéaires à effets mixtes

Proposition de stage niveau M2 à INRAE Jouy-en-Josas (printemps 2022)

Pour postuler, envoyez un CV et votre dernier relevé de notes à `maud.delattre@inrae.fr` *et* `guillaume.kon-kam-king@inrae.fr`.

Contexte applicatif

Les modèles à effets mixtes permettent d'analyser des observations collectées de façon répétée sur plusieurs individus, une situation typique dans des domaines comme la médecine, le marketing, les statistiques sportives ou la génétique. La variabilité intrinsèque aux données est alors attribuable à différentes sources (intra-individuelle, inter-individuelle, résiduelle) dont la prise en compte est essentielle pour caractériser sans biais les mécanismes biologiques à l'origine des observations. Dans un modèle à effets mixtes, la variabilité entre individus est décrite au moyen de covariables et d'effets aléatoires. Les covariables décrivent les différences entre individus dues à des caractéristiques observées tandis que les effets aléatoires représentent la part de la variabilité entre individus qui n'est pas attribuable aux covariables mesurées. Un exemple d'application envisagé concerne l'amélioration des plantes de culture (maïs, blé, etc.). Dans ce cas, les modèles non linéaires à effets mixtes peuvent être utilisés pour décrire le développement des plantes en fonction de leur génotype et des conditions environnementales. Ils permettent de comprendre le rôle des interactions entre le génotype et l'environnement dans l'évolution de la plante et sont utilisés pour prédire les performances de différentes variétés dans des conditions environnementales spécifiques. Les covariables considérées sont généralement nombreuses puisque les variétés sont caractérisées par des milliers de covariables génétiques (des marqueurs moléculaires par exemple) dont on sait que la plupart d'entre elles n'ont aucun effet sur certains traits phénotypiques. Il est donc intéressant d'envisager une sélection de variables à la fois pour identifier les régions du génome qui affectent effectivement le caractère d'intérêt et pour améliorer la capacité de prédiction du modèle. La grande dimension des données génomiques implique d'aborder la sélection de variables dans un cadre où le nombre de covariables est plus grand que le nombre d'individus. À notre connaissance, la question de la sélection de variables en grande dimension, pourtant populaire en Statistique et Machine Learning, a été peu étudiée dans le cadre spécifique des modèles non linéaires à effets mixtes.

Objectifs

Ce stage fait suite au travail récent de Marion Naveau dans lequel la sélection de covariables dans les modèles non linéaires à effets mixtes s'appuie sur l'utilisation d'un prior bayésien de type *spike and slab* [1]. L'approche bayésienne de la sélection de variable présente un certain nombre d'attraits: un degré d'interprétabilité, une stabilité numérique avantageuse et une flexibilité utile. Les résultats obtenus avec

le prior *spike and slab* sont très positifs et incitent à continuer l'exploration d'autres lois a priori pour construire un panorama de leurs mérites comparés. Le travail du stagiaire visera à étudier numériquement les différences de performances obtenues lorsque d'autres lois a priori - en particulier le prior Horseshoe - sont spécifiées sur les coefficients du modèle pour réaliser la sélection de variables : taux de fausses découvertes, robustesse à la collinéarité entre les covariables, complexité algorithmique (pour le passage au Big Data), etc.

Le stagiaire débutera par un travail bibliographique visant à comprendre le formalisme des modèles non linéaires à effets mixtes [2], les spécificités des différentes lois a priori utilisées en sélection de variables bayésienne [3, 4], et apprendra à maîtriser un des langages de programmation probabilistes pour l'inférence Bayésienne (Stan, NIMBLE, PyMC, Tensorflow probability, JAGS, Turing. . .). Il élaborera ensuite un plan de simulations permettant d'étudier différents scenarios d'intérêt pour la sélection de variables. Il mettra ensuite en œuvre les expériences numériques dans le langage choisi. Il se placera d'abord dans des situations où l'ensemble des covariables sujettes à sélection est de petite dimension devant le nombre d'observations avant de considérer des covariables de grande dimension si la durée du stage le permet.

Les compétences acquises par le stagiaire à l'issue du stage couvriront une variété de savoirs recherchés dans le monde académique et industriel : familiarité avec les modèles mixtes et l'inférence bayésienne, connaissance de l'état de l'art en sélection de variable, maîtrise d'un langage de programmation probabiliste et notions de statistique computationnelle.

Profil recherché

Le candidat doit être en formation de M2 (ou une formation équivalente) en statistique. Un intérêt pour la modélisation statistique, des notions d'apprentissage statistique et de programmation en `R` ou Python sont attendus.

Conditions du stage

Laboratoires d'accueil

UR 1404 Mathématiques et Informatique Appliquées du Génome à l'Environnement (MaIAGE), INRAE, 78352 Jouy-en-Josas

Encadrants

Maud Delattre : `maud.delattre@inrae.fr`

Guillaume Kon Kam King : `guillaume.kon-kam-king@inrae.fr`

Durée 4-6 mois

Gratification environ 550 euros nets par mois

References

[1] Delattre, M., Kon Kam King, G., Naveau, M. and Sansonnet, L. *Bayesian high-dimensional covariate selection in non-linear mixed-effects models using the SAEM algorithm.* (hal-03685060)

[2] Lavielle, M. (2014) *Mixed Effects Models for the Population Approach: Models, Tasks, Methods and Tools.* Chapman & Hall/CRC biostatistics series.

[3] Tadesse, M. G., & Vannucci, M. (2021). Handbook of Bayesian Variable Selection. CRC Press.

[4] Sutton, M. (2020). *Bayesian Variable Selection.* In K. L. Mengersen, P. Pudlo, & C. P. Robert (Eds.), Case Studies in Applied Bayesian Data Science: CIRM Jean-Morlet Chair, Fall 2018 (pp. 121–135). Springer International Publishing. https://doi.org/10.1007/978-3-030-42553-1_5