Sélection de variables en grande dimension dans les modèles non linéaires à effets mixtes.

Application en amélioration des plantes.

Proposition de stage niveau M2 (printemps 2021)

Pour postuler envoyer CV et dernier relevé de notes à
maud.delattre@inrae.fr et laure.sansonnet@agroparistech.fr.

Contexte applicatif

Les modèles à effets mixtes permettent d'analyser des observations collectées de façon répétée sur plusieurs individus. 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. En amélioration des plantes, les modèles non linéaires à effets mixtes sont utilisés pour décrire le développement des plantes en fonction de leurs génotypes 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. A notre connaissance, la question de la sélection de variables en grande dimension n'a jamais été étudiée dans les modèles non linéaires à effets mixtes.

Objectifs

Après s'être approprié le formalisme des modèles non linéaires à effets mixtes [**?**], le stagiaire s'intéressera à la mise en place d'une méthode de sélection de variables en grande dimension dans ces modèles (par exemple un critère de type "Lasso" ou encore une méthode de type "spike and slab"). Les objectifs du stage seront i) d'implémenter la méthode proposée, ii) de réaliser des simulations pour en valider le comportement numérique, iii) d'en étudier les propriétés théoriques, et iv) de l'appliquer à des données réelles. L'application sur données réelles se fera en collaboration avec Renaud Rincent (UMR GQE - Le Moulon - Paris Saclay).

Le stage pourra déboucher sur une thèse.

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 (éventuellement en grande dimension) et de programmation en R sont attendus.

Il est à noter qu'aucune connaissance en sciences du vivant n'est exigée et que selon le profil et les intérêts du candidat, le stage pourra se concentrer sur les aspects théoriques ou numériques.

Conditions du stage

Laboratoires d'accueil

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

UMR MIA-Paris, AgroParisTech, INRAE, 75005 Paris

Encadrantes

Maud Delattre : maud.delattre@inrae.fr

Laure Sansonnet : laure.sansonnet@agroparistech.fr

Durée 4-6 mois

Gratification environ 550 euros nets par mois

Références

[1] Delattre, M., Lavielle, M. and Poursat, M.A. (2014) *A note on BIC in mixed-effects models.* Electronic Journal of Statistics.

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

[3] Rokova, V. and George, E. I. (2014) *EMVS : The EM Approach to Bayesian Variable Selection.* Journal of the American Statistical Association.

[4] Schelldorfer, J., Bühlmann, P. and Van de Geer, S. (2011) *Estimation for High-Dimensional Linear Mixed-Effects Models Using ℓ_1-Penalization.* Scandinavian Journal of Statistics.

[5] Zhao, P. and Yu, B. (2006) *On Model Selection Consistency of Lasso.* Journal of Machine Learning Research