Nouvelles approches de sélection de variables en grande dimension pour les modèles non linéaires à effets mixtes

Application en amélioration des plantes

Proposition de stage niveau M2 (printemps 2026)

Pour postuler envoyer CV et dernier relevé de notes à
maud.delattre@inrae.fr et laure.sansonnet@sorbonne-universite.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. Plus largement, la sélection de variables en grande dimension dans les modèles non linéaires à effets mixtes constitue une thématique de recherche en plein développement, présentant un intérêt croissant dans de nombreux domaines appliqués [1, 2, 5].

Objectifs

Un enjeu majeur en amélioration des plantes est de tenir compte du fait que les effets génétiques peuvent évoluer au cours du temps : ils peuvent être plus marqués en début qu'en fin de croissance, ou n'apparaître que dans une phase précise du développement de la plante. Il est donc important d'intégrer cette information au modèle et de réaliser la sélection de variables en tenant compte de cette évolution temporelle. Des approches bayésiennes ont été proposées, notamment dans [3], basées sur une décomposition des coefficients des covariables et l'utilisation d'algorithmes MCMC. Toutefois, il n'existe à ce jour aucune implémentation efficace pour les modèles non linéaires à effets mixtes, pourtant essentiels en écophysiologie et dans de nombreux autres domaines. L'objectif du stage est de développer et d'évaluer une méthode de sélection de variables adaptée à ces modèles, tout en intégrant la problématique des effets génétiques évoluant au cours du temps. Le travail s'appuiera sur la méthode `SAEMVS`, proposée dans [5], qui combine un *prior* bayésien de type *spike and slab* gaussien et l'algorithme SAEM (Stochastic Approximation EM).

Le stage commencera par un travail bibliographique visant à se familiariser avec le formalisme des modèles non linéaires à effets mixtes et les approches de sélection de variables en grande dimension qui leur sont associées. La/le stagiaire proposera ensuite un cadre de modélisation adapté aux covariables dont les effets varient dans le temps, puis adaptera la procédure SAEMVS à ce cadre. Les performances de la méthode seront évaluées par des simulations, et, en fonction de l'avancement, la démarche pourra être testée sur des données réelles en collaboration avec nos partenaires biologistes.

Ce stage pourra constituer une première étape vers une thèse, visant à développer des méthodes de sélection de variables robustes aux colinéarités fréquentes entre covariables de grande dimension.

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. Il est à noter qu'aucune connaissance en sciences du vivant n'est exigée.

Conditions du stage

Laboratoires d'accueil

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

Laboratoire de Probabilités, Statistique et Modélisation (LPSM), Sorbonne Université, Université Paris Cité, CNRS, F-75005 Paris, France

Encadrantes

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

Laure Sansonnet : `laure.sansonnet@sorbonne-universite.fr`

Durée 4-6 mois

Gratification environ 550 euros nets par mois

Références

[1] Caillebotte, A., Kuhn, E., & Lemler, S. (2025). *Estimation and variable selection in nonlinear mixed-effects models.* arXiv :2503.20401.

[2] Gabaut, A., Thiébaut, R., Proust-Lima, C., & Prague, M. (2025). *A Stability-Enhanced Lasso Approach for Covariate Selection in Non-Linear Mixed Effect Model.* medrxiv :2025.05.15.25327667.

[3] Heuclin, B., Mortier, F., Trottier, C., & Denis, M. (2021). *Bayesian varying coefficient model with selection : An application to functional mapping.* Journal of the Royal Statistical Society Series C : Applied Statistics, 70(1), 24-50.

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

[5] Naveau, M., Kon Kam King, G., Rincent, R., Sansonnet, L., & Delattre, M. (2024). *Bayesian high-dimensional covariate selection in non-linear mixed-effects models using the SAEM algorithm.* Statistics and Computing, 34(1), 53.