

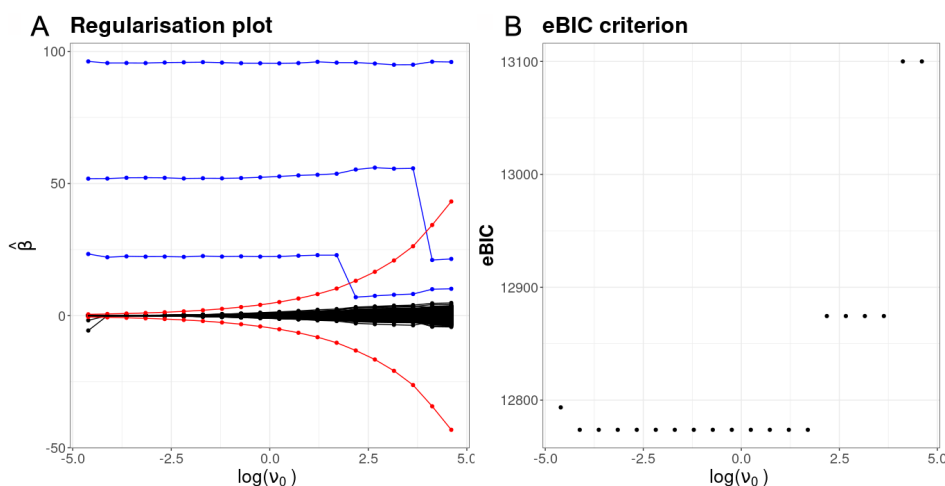
## ➤ Décrypter la variabilité génétique avec les modèles à effets mixtes : nouvelles perspectives pour l'amélioration des plantes.

Les modèles non linéaires à effets mixtes sont essentiels pour analyser des données issues de mesures répétées, notamment en amélioration des plantes, où ils permettent de modéliser le développement des variétés en fonction de leur génotype et de l'environnement. Ces modèles prennent en compte la variabilité intra- et inter-individuelle, ainsi que des covariables génétiques souvent nombreuses et redondantes. Une méthodologie novatrice combinant un prior bayésien spike-and-slab Gaussien et l'algorithme SAEM (Stochastic Approximation of EM) a été développée pour sélectionner efficacement les covariables pertinentes et améliorer les performances prédictives, tout en réduisant les temps de calcul par rapport aux approches MCMC classiques. Testée sur des données réelles de génétique végétale, cette méthode a permis d'identifier des marqueurs associés à la sénescence du blé, en accord avec des analyses antérieures.



### CONTEXTE ET ENJEUX

Les modèles à effets mixtes permettent d'analyser des observations répétées sur plusieurs individus en attribuant la variabilité des données à différentes sources (intra-individuelle, inter-individuelle, résiduelle). La variabilité inter-individuelle est décrite à l'aide de covariables et d'effets aléatoires : les covariables reflètent les différences attribuables à des caractéristiques observées, tandis que les effets aléatoires représentent la part de variabilité non expliquée par les covariables mesurées. En amélioration des plantes, les modèles non linéaires à effets mixtes (NLME) sont utilisés pour décrire le développement des plantes en fonction de leur génotype et des conditions environnementales. Ces modèles intègrent souvent des milliers de covariables génétiques (marqueurs moléculaires), dont la plupart n'ont aucun effet sur certains traits phénotypiques. Une sélection de variables est donc essentielle pour identifier les régions génomiques pertinentes et améliorer la capacité prédictive du modèle. Cependant, la grande dimension des données génomiques impose d'aborder ce problème dans un cadre où le nombre de covariables dépasse le nombre d'individus.



Graphe de régularisation obtenu sur un modèle de croissance simulé. La Figure (A) présente les effets fixes des covariables pour chaque valeur du paramètre de régularisation sur une grille. Les lignes bleues correspondent aux trois covariables pertinentes, tandis que les lignes noires représentent les effets fixes nuls. Les lignes rouges indiquent les seuils de sélection des covariables. Une covariable est sélectionnée si la valeur estimée de son coefficient est située en dehors des deux lignes rouges. La Figure (B) montre la valeur du critère eBIC pour chaque valeur du paramètre de régularisation sur la même grille. Les covariables sélectionnées correspondent au paramètre de régularisation associé à la plus petite valeur de eBIC. La procédure SAEMVS identifie exactement le bon modèle dans cet exemple simulé. ([https://github.com/Marion-Naveau/Supp\\_Information\\_SAEMVS](https://github.com/Marion-Naveau/Supp_Information_SAEMVS))



## RESULTATS

Des chercheurs d'INRAE et AgroParisTech ont développé une méthodologie novatrice, nommée SAEMVS, pour répondre à ces enjeux. Elle combine des approches bayésiennes et fréquentistes en s'appuyant sur un prior spike-and-slab Gaussien pour la régularisation et sur l'algorithme SAEM pour l'estimation des paramètres dans les NLME. Inspirée de la régression LASSO, cette méthode explore une grille de paramètres pour ajuster la régularisation et optimiser la sélection des variables. Par rapport aux approches MCMC classiques, l'utilisation du SAEM offre des avantages computationnels significatifs tout en maintenant des performances solides sur des données simulées. Appliquée à des données réelles, cette méthode a permis d'identifier des marqueurs génétiques associés à la sénescence du blé. Ce travail a été réalisé et financé au sein du projet ANR interdisciplinaire STAT4PLANT <https://stat4plant.mathnum.inrae.fr>.



## PERSPECTIVES

Certains NLME en amélioration des plantes reposent sur des modèles écophysiologiques qui simulent précisément les étapes du développement des plantes, mais leur évaluation est numériquement coûteuse. Ces contraintes rendent les algorithmes d'inférence classiques inadaptés aux données réelles. Une solution prometteuse est le recours à la métamodélisation, qui permet de réduire significativement les temps de calcul en approximant ces modèles complexes. Déjà utilisée pour l'estimation de paramètres dans d'autres NLME, la métamodélisation pourrait rendre la méthodologie SAEMVS applicable à ces modèles écophysiologiques, offrant ainsi aux biologistes une solution adaptée à leurs problématiques.



## RÉFÉRENCES

> Naveau M., Kon Kam King G., Rincet R. & al. - Bayesian high-dimensional covariate selection in non-linear mixed-effects models using the SAEM algorithm. Stat Comput 34, 53 (2024)  
<https://doi.org/10.1007/s11222-023-10367-4>



Contact : Maud Delattre  
Unité : MaAGE  
Département : MathNum  
Centre INRAE : Île de France Jouy-en-Josas  
INRAE 2030 Orientations scientifiques (OS) et Orientations de politique générale (OP) : OS 1.3  
Grands Objectifs Scientifiques (GOS) MathNum : GOS 1  
Mots-clés : Modèles à effets mixtes, sélection de variables, grande dimension, modèles écophysiologiques, sélection variétale