Metamodeling for high-dimensional variable selection in complex nonlinear mixed-effects models.

Application in plant breeding.

M2 internship (spring 2024)

To apply send CV and latest school transcript to
maud.delattre@inrae.fr et marion.naveau@inrae.fr.

Application context

Mixed-effects models can be used to analyze observations collected repeatedly from several individuals. The variability intrinsic to the data is then attributable to different sources (intra-individual, inter-individual, residual), which must be taken into account to characterize the biological mechanisms behind the observations without bias. In a mixed-effects model, inter-individual variability is described by means of covariates and random effects. Covariates describe the differences between individuals due to observed characteristics, while random effects represent the part of the variability between individuals that is not attributable to measured covariates. In plant breeding, non-linear mixed-effects models are used to describe plant development as a function of genotype and environmental conditions. They help to understand the role of genotype-environment interactions in plant evolution, and are used to predict the performance of different varieties under specific environmental conditions. The covariates considered are generally high-dimensional, since varieties are characterized by thousands of genetic covariates (molecular markers, for example). In addition, some of the non-linear mixed-effects models used in plant breeding are based on ecophysiological models, which simulate the various stages of plant development very precisely, and are costly to evaluate. In this specific case, the computation times required by conventional inference algorithms are too long for them to be used on real data.

Objectives

This internship follows on from Marion Naveau's recent work in which covariate selection in nonlinear mixed-effects models is achieved by coupling the SAEM algorithm and a Bayesian spike-and-slab prior. This procedure is iterative, requiring several evaluations of the nonlinear regression function at each iteration. When the regression function is costly to evaluate, the execution time of the complete methodology becomes prohibitive. Metamodeling approaches, already successfully applied for parameter estimation in nonlinear mixed-effects models in [1], could reduce computation times.

The trainee will start with a bibliographical study aimed at understanding the formalism of nonlinear mixed-effects models and classical metamodeling approaches. He will then consider the use of metamodels to improve the methodology developed in [2]. He will then implement the proposed method and run simulations to validate its numerical behavior. Finally, he will apply it to real data. The application to real data will be carried out in collaboration with Renaud Rincent (UMR GQE - Le Moulon - Paris Saclay).

Profile

The candidate must be a M2 student (or equivalent) in statistics. An interest in statistical modeling, notions of statistical learning (possibly in high dimensions) and programming in R or Python are expected.

No knowledge of life sciences is required.

Internship conditions

Host laboratory

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

Supervisors

Maud Delattre : maud.delattre@inrae.fr

Marion Naveau : marion.naveau@inrae.fr

Duration 4-6 months

Gratification about 550 euros net per month

Références

[1] Barbillon, P., Barthélémy, C., & Samson, A. (2017) *Parameter estimation of complex mixed models based on meta-model approach.* Statistics and Computing, 27(4), 1111-1128.

[2] Naveau, M., Kon Kam King, G., Rincent, R., Sansonnet, L. & Delattre, M. (2022) *Bayesian high-dimensional covariate selection in non-linear mixed-effects models using the SAEM algorithm.* arXiv preprint arXiv :2206.01012.

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