Master 2 – research internship

Comparing approximations in estimating generalized linear mixed models for categorical data

Supervisors: Jean-Baptiste Durand[(1)], Sandra Plancade[(2)] and Nathalie Peyrard[(2)]
[(1)] Jean-Baptiste.Durand@cirad.fr, Cirad, AMAP Montpellier and Inria, Laboratoire Jean Kuntzmann,Grenoble.
[(2)] Sandra.Plancade@inrae.fr,Nathalie.Peyrard@inrae.fr, MIAT, INRAE, Auzeville-Tolosane
Location of internship: AMAP, Montpellier

Allowance: about 550 euros per month
Duration: 5-6 month (starting in March 2022)

Context

The development of automated platforms form plant phenotyping generates high volumes of data. These are likely to be leveraged for improving crop management, provided adequate models are built to analyse them. Particularly in fruit trees, trunks produce lateral branches described by several categorical variables (length classes, etc.). Current models hypothesise trunk growth being subjected to successive, non-observable phases, with variable durations from a tree to another and impacting the observed variable distribution. These latent phases can be inferred from the observed variables. From a mathematical point of view, such assumptions lead to multivariate hidden semi-Markov models (MHSMM [1]).

In this framework, we would like to compare tree growth under different conditions (cultivars, nitrogen treatments, etc.). These are introduced in the model through fixed effects in parameters, completed by random effects, to account for dependencies induced by repeated measurements on a same individual. Such extensions to classical MHSMMs lead to resorting to generalised linear mixed models (GLMMs) for categorical data. Bayesian estimation of these GLMMs does not lead to closed form estimators for the posterior distribution and thus, numerical approximations are required.

Tasks

This internship aims at comparing two approximation families for Bayesian estimation of GLMMs, from both simulated and real data on juvenile apple trees. A first strategy based on Markov Chain Monte Carlo algorithms (MCMC [2]) was tried and highlighted extremely slow convergence, together with highly variable estimates between MCMC estimates. A promising alternative is offered by quadrature methods, particularly INLA [3].

Firstly, the intern will have to acquire knowledge on MCMC and INLA. Then, he or she will apply INLA to GLMM estimation from simulated data, to be compared with MCMC estimates, based on R libraries. The impact of estimation accuracy on selection criteria for fixed effects (for example, the effect of cultivar) will be assessed as well. The most accurate approximated estimator will be then applied to true data sets. Depending of the advances achieved by the intern, two extensions can be considered: either the transposition of the methodology to the analysis of willow trees, or more methodological extensions considering the application of INLA-like quadrature methods in maximum likelihood estimation of MHSMM with mixed effects.

Prerequisites

The intern should have some strong mathematical background with a special focus on statistics, together with knowledge in programming. Additional skills in statistical Bayesian modelling would be an asset. The intern should be motivated by applications to agronomy/plant growth modelling (although no experience is required in this field).

Remark

This work may be continued as a PhD thesis, depending on the ability of the candidate to obtain fellowships.

References

[1]Yu, S.-Z. Hidden semi-Markov models (2010). *Artificial intelligence,* **174**(2), 215—243.

[2] Hadfield, J. D. MCMCglmm: MCMC Methods for Multi-Response GLMMs in R (2010). *Journal of Statistical Software*, **33**(2), 1–22.

[3] Rue, H., Martino, S. and Chopin, N. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations (2009). *Journal of the Royal Statististical Society Series B*, **71** (Part 2), 319-392.