Estimation and variable selection in joint model of survival times and longitudinal outcomes with correlated random effects. Application to analyze the effects of European corn borer attacks on maize flowering date

PhD position to be filled in October 2021

Keywords.
statistics, frailty models, mixed effects models, joint modeling, maximum likelihood estimation, variable selection, penalized criterion, EM algorithm, stochastic approximation, quantitative genetics, molecular markers, biotic stress, maize flowering, pest development model, dynamical modeling, spatial dependence

Context.
Plant development typically proceeds through successive phenological stages, such as the duration from sowing to the germination or the vegetative, flowering, and grain-filling stages. Transitions between phenological stages are discrete events that depend on the variety-dependent plant growth dynamics and the biotic and abiotic environmental sequence experienced. The latter can include biotic stress, such as the intrinsic dynamics of all the plant biotic partners, including pests, and abiotic stress, such as climatic conditions.

The INRAE unit GQE-Le Moulon led a divergent selection experiment (DSE) for the flowering time of maize since 1997 in Saclay. The main objective of this experiment is to study the genetic basis of the response to selection and understand how genetic and environmental factors impact plant phenology. Therefore, the earliest and latest individuals from two initial maize lines F252 and MBS were selected every year and reproduced by selfing, leading to two late families and two early families of related genotypes per genetic background. Every year, ten descendants of each family and the ancestral lines are sown, one hundred individuals per genotype, evaluated and selected for their flowering time measured in thermal time (Durand, E. *et al.* [2010]). This biological material has been used to investigate the response to selection and the genetic factors affecting the flowering date. More recently, abiotic factors, such as climate variables, have been studied, particularly in the ongoing project WarmRules of DATAIA Institut. However, biotic factors such as the European Corn Borer (ECB) pressure could also impact plant phenology. Therefore GQE-Le Moulon also monitored the ECB dynamics on this experimental field during two consecutive years (Sanane, I. *et al.* [2018]). Other field experiments are ongoing to study the relation between plant phenology and pest succes (Phenofore project).

Objectives.
The applied objective is to investigate how the presence/absence of one pest can affect plant phenological stages characterized by different times of interest, such as the flowering time. Indeed, the presence/absence of a pest on a given plant depends on the pest dynamics, which itself depends on abiotic factors such as temperature and biotic factors, including plant variations such as development and genotypic variability leading to complex dependence structure in the dynamical phenomenon.

The mathematical objective will be first to build a joint model of the time of interest in plant growth and the dynamics of related traits (e.g. pest presence/absence dynamics), following the ideas of those initially developed for medical applications (Tsiatis, A.A. and Davidian, M. [2004]) and for epidemiological applications (Belay, D. B. *et al.* [2017]), and second to develop the corresponding methodology for statistical inference. This joint model will allow to model inter-and intra-individual variabilities through random effects and to include high-dimensional covariates such as -omic data. Statistical procedures for estimation, variables selection, and prediction will be proposed and analyzed from theoretical points of view. Corresponding efficient numerical tools based on stochastic algorithms adapted to high-dimensional setting will be developed and applied to analyze data of maize flowering times and European Corn Borer attacks.

Methodology.
In the case where the mixed model is linear, such joint model has been widely studied from a numerical point of view, with applications in the medical field (Tsiatis, A.A. and Davidian, M. [2004]) and theoretical properties have been established (Fu, R. and Gilbert, P.B. [2017]). However, less has been done in the non-linear setting, which is nevertheless of particular interest for modeling issues. Some papers (see e.g., Murawska, M. *et al.* [2012]) propose a two-step procedure to estimate the parameters of the mixed effect model and then those of the survival model. Other papers deal with the joint likelihood to simultaneously estimate all the joint model parameters (see, e.g., Mbogning, C. *et al.* [2014]). However, there is no theoretical guarantee.

Statistical methods dedicated to inference in the non-linear joint model will be developed using a penalized criterion to take into account the high-dimension of the covariates and at the computational level based on stochastic approximation algorithms. To establish theoretical properties, classical results on the M-estimators (Van Der Vaart, AW and Wellner, JA [1996]) will be adapted to the latent variable setting of the joint model. All these developments will also rely on previous works dedicated to survival models, including high-dimensional covariates (Guilloux, A. *et al.* [2016]), non-linear mixed-effects modeling (Kuhn, E. *et al.* [2005]) and stochastic optimization procedure in high-dimension for the computational part (Allassonniere, S. *et al.* [2015]).

The applied objective consists in understanding how pest presence may affect the phenological dynamics of maize plants in addition to other factors characterizing the environment (e.g climatic data) or plant traits (e.g NIRS data). Therefore the statistical developments will be

used to model the flowering date jointly in different genotypes of maize and the dynamics of European Corn Borer (ECB) on these plants. Such a model has been previously used to model, for example, mosquito abundance and malaria incidence and identify environmental factors affecting malaria incidence via the presence of mosquito (Belay, D. B. *et al.* [2017], Das, K. *et al.* [2012]). However, the dynamic of the mosquito was not taken into account within the modeling. An appropriate dynamical model for ECB predicting the probability of its presence on a plant depending on its phenology (e.g., its response to temperature), plant genotype and phenology and taking into account spatial autocorrelation due to female and larvae movements in the field will be built. The rich dataset produced by the divergent selection experiment (DSE) on flowering time contains flowering dates, height, and yield of the different genotypes available each year at the plant level. Climatic variables (such as temperature, wind, radiation, humidity, precipitations...) are also available. In addition, for the past two years, the ECB dynamics were monitored weekly for 2.5 months each year. Plants showing signs of ECB attacks were counted. As the DSE is an ongoing experiment, such data will be collected every year in the future. Other dataset related to this topic are currently collected at GQE-Le Moulon and could be analyzed with the joint modeling approach.

Description of the interdisciplinary collaboration.
This thesis is part of the Stat4Plant project dedicated to developing statistical methods to characterize the interactions between the plant and its environment. The ANR funds this project in the axis mathematics for applications in biology and health over the period 2021-2025. The consortium of the project gathers six teams mixing statisticians and biologists mainly located on Paris-Saclay campus. Therefore, the thesis will benefit from the rich scientific environment of the University of Paris Saclay within an interdisciplinary collaboration between the MaIAGE unit of the INRAE center of Jouy-en-Josas, the GQE-Le Moulon unit, and the MICS unit of CentraleSupélec in Gif-sur-Yvette.

Supervision and funding.
The PhD will be jointly supervised by Estelle Kuhn (INRAE, Unité MaIAGE, Jouy-en-Josas), Judith Legrand (GQE-Le Moulon, Université Paris Saclay), Sarah Lemler (MICS, CentraleSupélec). The PhD will be funded by the ANR PRC Stat4Plant. The PhD will be conducted in the MaIAGE unit in Jouy-en-Josas and in the GQE-Le Moulon unit in Gif-sur-Yvette in Saclay area.

Profile and skills required.
The applicant will have a strong background in mathematics, in particular in statistics, for example master degree or engineering school degree. He or she will also have taste in applications in life sciences. Experience in computer programming will be a benefit. If not, computer skills should be gained.

Application.
To apply, please send CV and motivation letter to :
estelle.kuhn@inrae.fr and sarah.lemler@centralesupelec.fr.

Références

Allassonniere, S. *et al.* Convergent stochastic expectation maximization algorithm with efficient sampling in high dimension. application to deformable template model estimation. *Comp. Stat. Data Anal.*, 91, 2015. doi : 10.1016/j.csda.2015.04.011.

Belay, D. B. *et al.* Joint bayesian modeling of time to malaria and mosquito abundance in ethiopia. *BMC infectious diseases*, 17(1), 2017. doi : 10.1186/s12879-017-2496-4.

Das, K. *et al.* A bayesian framework for functional mapping through joint modeling of longitudinal and time-to-event data. *Inter. jour. of plant genomics*, 2012, 2012. doi : 10.1155/2012/680634.

Durand, E. *et al.* Standing variation and new mutations both contribute to a fast response to selection for flowering time in maize inbreds. *BMC evolutionary biology*, 10(2), 2010. doi : 10.1186/1471-2148-10-2.

Fu, R. and Gilbert, P.B. Joint modeling of longitudinal and survival data with the cox model and two-phase sampling. *Lifetime data analysis*, 23(1) :136–159, 2017. doi : 10.1007/s10985-016-9364-1.

Guilloux, A. *et al.* Adaptive kernel estimation of the baseline function in the cox model with high-dimensional covariates. *Jour. Multivar. Anal.*, 148, 2016. doi : 10.1016/j.jmva.2016.03.002.

Kuhn, E. *et al.* Maximum likelihood estimation in nonlinear mixed effects models. *Comp. Stat. Data Anal.*, 49(4), 2005. ISSN 0167-9473. doi : https://doi.org/10.1016/j.csda.2004.07.002.

Mbogning, C. *et al.* Joint modeling of longitudinal and repeated time-to-event data with maximum likelihood estimation via the saem algorithm. *Jour. of Stat. Comp. and Sim.*, 2014. doi : 10.1080/00949655.2013.878938.

Murawska, M. *et al.* A two-stage joint model for nonlinear longitudinal response and a time-to-event with application in transplantation studies. *Jour. of Proba. and Stat.*, 2012, 2012. doi : 10.1155/2012/194194.

Sanane, I. *et al.* Plant versus herbivores : dynamics of interactions between corn and stem borers lepidoptera. *Maize genetics conference* , 2018.

Tsiatis, A.A. and Davidian, M. Joint modeling of longitudinal and time-to-event data : an overview. *Statistica Sinica*, 2004.

Van Der Vaart, AW and Wellner, JA. Weak convergence and empirical processes : With applications to statistics springer series in statistics. *Springer*, 1996. doi : 10.1007/978-1-4757-2545-2.