

Joint modeling of longitudinal and survival data in high dimension

Application to the analysis of the effects of ostrinia attacks on the flowering date of corn

Antoine Caillebotte¹

E.Kuhn¹ S.Lemler² E.Marchadier³ J.Legrand³

¹Université Paris-Saclay, INRAE, MalAGE, ²CentraleSupélec MICS

³INRAE Génétique Quantitative et Evolution Le Moulon

Internship presentations

4 juillet 2022

1 Introduction

- Flowering date
- ostrinia attack proportion and mixed effect model
- Joint modeling

2 Methodology

- Objectives
- High dimension
- Expectation Maximisation algorithm
- Stochastic Approximation EM algorithm

3 perspectives

biological context



FIGURE 1 – ostrinia larva



FIGURE 2 – ostrinia attack

Photo credit : Sacha Revillon

Divergent selection

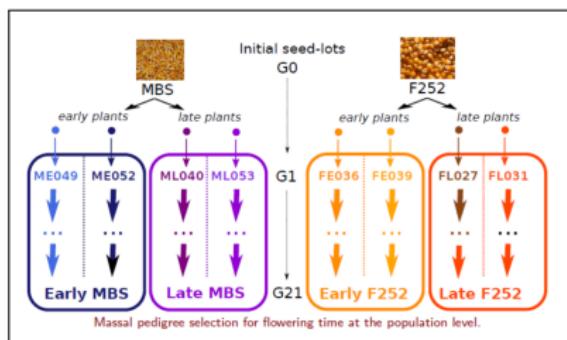


FIGURE 3 – *ostrinia larva*

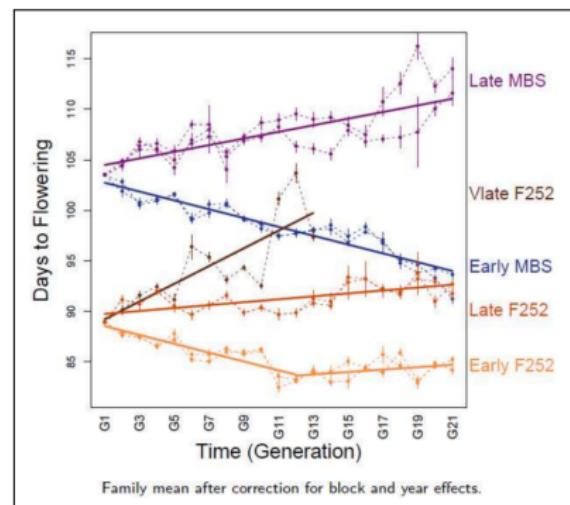


FIGURE 4 – *ostrinia attack*

Flowering date

- Flowering date of each corn plant observed

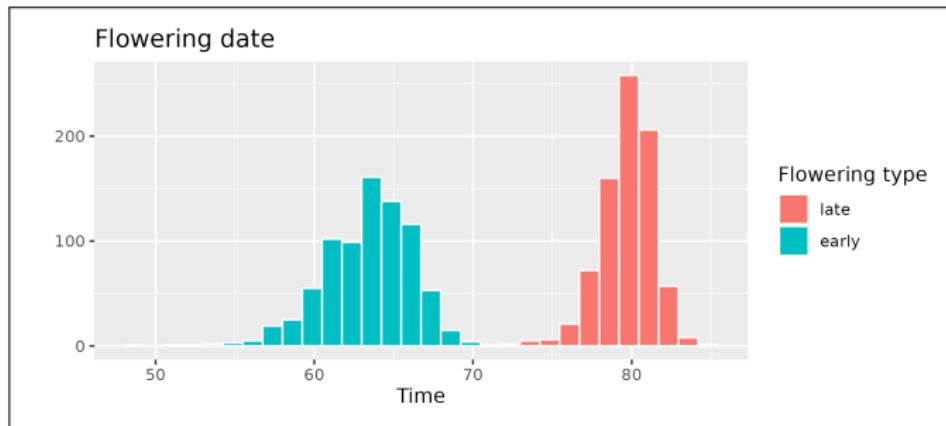


FIGURE 5 – Flowering date

- What impacts flowering : Genetic variability ? Insect attack ?

Survival analysis

Focus on **time to event of interest**

- In medicine : Time of remission or time to death
- In our case : **Date of flowering.**

Definition : Survival time

The survival time $T > 0$ is the time that elapses between an initial moment (start of the study) and the appearance of an event interest.

Hazard function λ

T is a positive random variable. $\lambda(t)$ probability that the event occurs in a small time interval after t , knowing that it did not occur until time t :

$$\lambda(t) = \lim_{h \rightarrow 0} \frac{\mathbb{P}(t < T \leq t + h | T > t)}{h}$$



Cox Model

Hazard function λ

$$\lambda(t) = \lim_{h \rightarrow 0} \frac{\mathbb{P}(t < T \leq t+h | T > t)}{h}$$

Regression model that links the survival time to explanatory variables and assume a common behavior.

The Hazard function is given by :

$$\lambda(T | \mathbf{U}_i) = \lambda_0(T) \exp(\boldsymbol{\beta}^T \mathbf{U}_i)$$

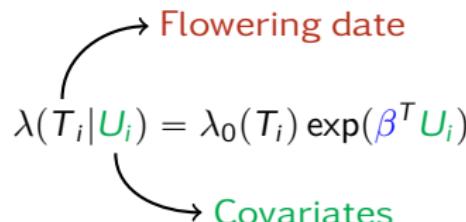
- T time event
- λ_0 : baseline hazard which explain the common behavior
- \mathbf{U}_i : covariate for individual i
- $\boldsymbol{\beta}$ regression parameter



Modeling the flowering date

- For any plant $i \in \{1, \dots, I\}$

Hazard that the flowering occurs at T_i :

$$\lambda(T_i | U_i) = \lambda_0(T_i) \exp(\beta^T U_i)$$


- $T_i \in \mathbb{R}$: time event of interest (**observation**) ,
- λ_0 base line hazard **unknown** ,
- $U_i \in \mathbb{R}^p$: covariats for individual i (**known**) ,
- $\beta \in \mathbb{R}^p$: fixed effect **unknown** .

Model parameter : $\theta = \beta$

- Objective** : model the proportion of attack and link it into this model!

ostrinia attack proportion

- Repeated observation of the number of attacks over time in a line of the field

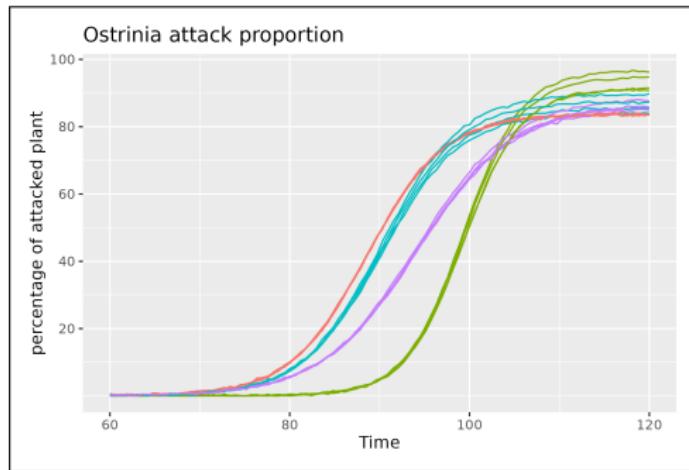


FIGURE 6 – ostrinia attack proportion

- intra (time-dependent behavior) and inter-individual variation (genetic background)

Non-linear mixed-effects model (NLME)

- Longitudinal data modeling :

$$Y_{g,j} = m(t_{g,j}; \varphi_g) + \epsilon_{g,j} ; \epsilon_{g,j} \underset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$$

behavior based on genetics

- $Y_{g,j} \in \mathbb{R}$: j-th responses of the g-th individual at time t_j (**observation**) ,
- $\varphi_g \in \mathbb{R}^3$: random group effects **not observed** ,
- m : a nonlinear function for φ .

- Inter-individual variation :

$$\varphi_g = \mu + \xi_g ; \xi_g \underset{i.i.d.}{\sim} \mathcal{N}(0, \Omega^2)$$

- $\mu \in \mathbb{R}^3, \Omega^2 \in \mathcal{M}_3(\mathbb{R})$: **unknown** ,

Model parameters : $\theta = (\sigma^2, \mu, \Omega^2)$



Joint Model : NLME and Survival model

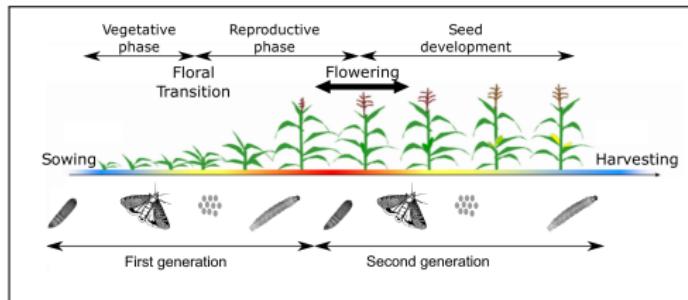


FIGURE 7 – Synchronization of the life cycles of the ostrinia and corn

- Combining the two models using a link function m .

$$\left\{ \begin{array}{l} \lambda(T_g | U_g, \gamma_g) = \lambda_0(T_g) \exp(\beta^T U_g + \alpha m(T_g, \varphi_g)) \\ Y_{g,j} = m(t_j; \varphi_g) + \epsilon_{g,j} \\ \dots \end{array} \right.$$

where α quantifies the association between the flowering and the attack proportion.

Model parameters : $\theta = (\sigma^2, \mu, \Omega^2, \beta, \alpha)$



Hierachical model

- Observation : T, Y

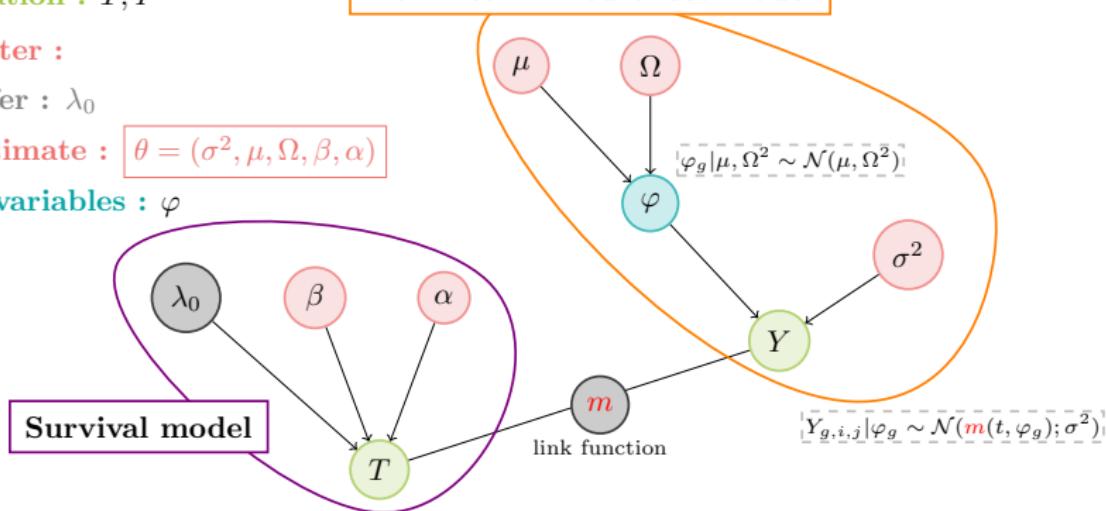
- Parameter :

- To infer : λ_0

- To estimate : $\theta = (\sigma^2, \mu, \Omega, \beta, \alpha)$

- Latent variables : φ

Nonlinear mixed effect model



1 Introduction

- Flowering date
- ostrinia attack proportion and mixed effect model
- Joint modeling

2 Methodology

- Objectives
- High dimension
- Expectation Maximisation algorithm
- Stochastic Approximation EM algorithm

3 perspectives



Objectives of the interships

mettre de la grande dimension dans ce jolie model



Handle the high dimension of covariate

$$\hat{\beta}_k = \arg \max_{\beta} \mathcal{L}(\beta) + pen(\beta)$$

Brouillon/remarque

sarah : parler du choix de la pénalité pour faire de la sélection de variables
antoine : on souhaite faire de la section de variable. seulement un petit nombre de variable sont pertinente et sont explicative de la date de floraison. Il faudrait trouver une pénalité qui permette de mettre à 0 les composante de β et garder les beta qui ont un sens dans la regression.



Expectation Maximisation algorithm

- **Require :** Starting point θ_0
- At the iteration $k \geq 0$:

① **E-Step (Expectation)**, evaluate the quantity :

$$Q(\theta|\theta_k) = \mathbb{E}_{\varphi|(\gamma, \theta_k)} [\log(\mathcal{L}_{comp}(\theta; \varphi, Y)) | Y, \theta_k]$$

② **M-Step (Maximisation)**, compute :

$$\theta_{k+1} = \arg \max_{\theta} Q(\theta|\theta_k)$$

• **Return :** $\hat{\theta} = \theta_K$ with K large enough

Problem : $Q(\theta|\theta_k)$ is not easy to compute



Stochastic Approximation EM algorithm

- **Require :** Starting point θ_0

- At the iteration $k \geq 0$:

- ① **S-Step (Simulation)**, simulate $\varphi^{(k)}$ according to $\mathcal{L}(\varphi|Y, \theta_k)$
- ② **A-Step (stochastic Approximation)**, evaluate :

$$Q_{k+1}(\theta) = (1 - u_k)Q_k(\theta) + u_k \log \mathcal{L}_{compt}(\theta; \varphi^{(k)}, Y)$$

- ③ **M-Step (Maximisation)**, compute :

$$\theta_{k+1} = \arg \max_{\theta \in \Theta} Q_{k+1}(\theta)$$

- **Return :** $\hat{\theta} = \theta_K$ with K large enough

where $(u_k)_{k \in \mathbb{N}}$ such that $\sum_{k=1}^{\infty} u_k = \infty$ and $\sum_{k=1}^{\infty} u_k^2 < \infty$

1 Introduction

- Flowering date
- ostrinia attack proportion and mixed effect model
- Joint modeling

2 Methodology

- Objectives
- High dimension
- Expectation Maximisation algorithm
- Stochastic Approximation EM algorithm

3 perspectives

Perspectives

Les algorithmes proximaux c'est cool!

- coding SAEM in the joint model
- integrating proximal algorithm to deal with high dimension of β
- analyse real data



Thank you for your attention!



Appendix 1 : full model

- Observation : T, Y

- Parameter :

- Fixed hyperparameters : ω_a^2, ω_b^2

- To estimate : $\theta = (\sigma^2, \mu, \Omega, \omega_\eta^2, \omega_\gamma^2, \beta, \alpha)$

- Latent variables : $Z = (\varphi, \eta, \gamma, a, b)$

