

Joint modeling of longitudinal and survival data in high dimension

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

Antoine Caillebotte¹

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

¹Université Paris-Saclay, INRAE, MalAGE, ²CentraleSupelec

³INRAE Génétique Quantitative et Evolution Le Moulon

Internship presentations

4 juillet 2022

1 Introduction

- Flowering date
- Ostrina attack proportion
- Combining the two models

2 Methodology

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

3 Conclusion

- perspectives

biological context



FIGURE 1 – ostrinia larva



FIGURE 2 – ostrinia attack

Photo credit : Sacha Revillon

Flowering date

- flowering date of each corn plant observed,

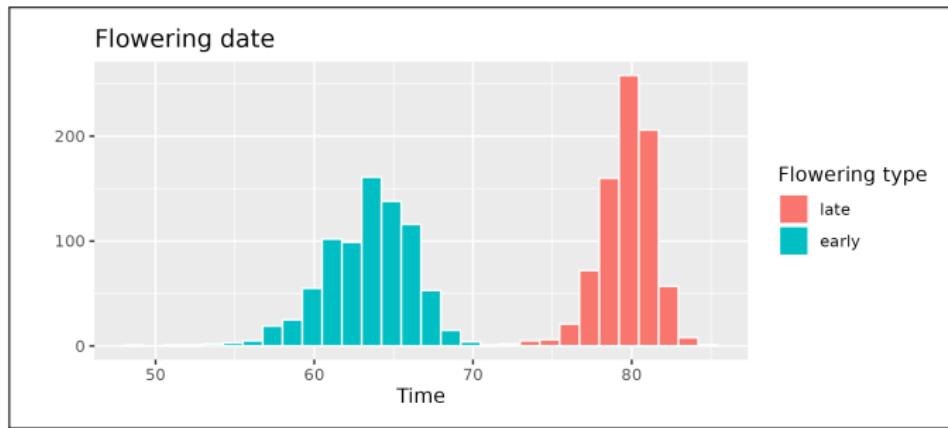


FIGURE 3 – Flowering date

- What impacts flowering : Genetic variability ? Insect attack ?

Survival analysis

We focus on the analysis of **time 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**.

T is a random variable positive.

Hazard function λ

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

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



Survival analysis : Cox Model

Definition : Cox model

The Cox model is a regression model that links the survival time of an individual to explanatory variables. And it allows to assume a common behavior to all individuals.

The Hazard function is given by :

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

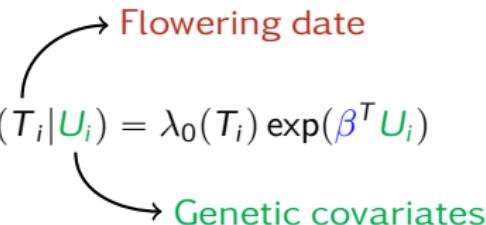
- λ_0 : baseline hazard which explain the common behvior
- U_i : covariate for individual i
- β regression parameter



Modeling the flowering date

- for any individual $i \in \{1, \dots, I\}$

Hazard that the flowering occurs at T_i :

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


- $T_i \in \mathbb{R}$: time event of interest (**observation**) ,
- λ_0 base line hazard **unknown** ,
- $\mathbf{U}_i \in \mathbb{R}^p$: covariate for individual i (**known**) ,
- $\boldsymbol{\beta} \in \mathbb{R}^p$: fixed effect **not observed** .

Current population parameters : $\theta = \beta$

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

Ostrinia attack proportion

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

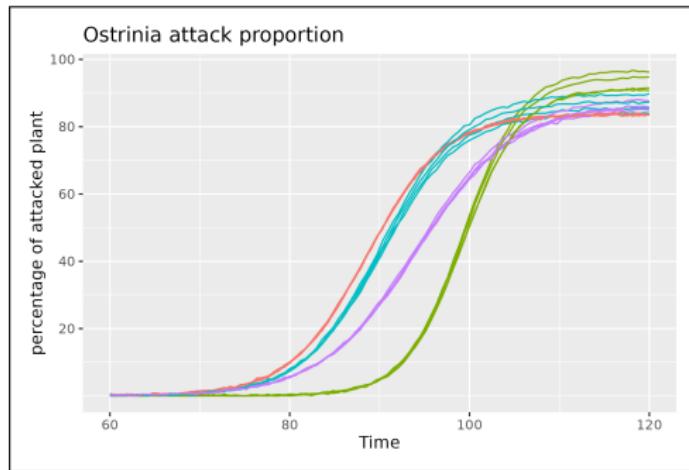


FIGURE 4 – Ostrinia attack proportion

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

Non-linear mixed-effects model (NLME)

- Longitudinal data modeling :

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

behavior based on genetics

- $Y_{i,j} \in \mathbb{R}$: i-th observation of the i-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 :

$$\varphi_i = \mu + \xi_i ; \xi_i \stackrel{i.i.d.}{\sim} \mathcal{N}(0, \Omega^2)$$

- $\mu \in \mathbb{R}^3$: intercept(c'est ça?) **unknown** ,

Current population parameters : $\theta = (\sigma^2, \mu, \Omega^2, \beta)$



Joint Model : NLME and Survival model

- Combining the two models using a link function m .

We assume that the proportion of attack at the time of flowering is explanatory of it.

- The joint model is :

$$\begin{cases} \lambda(T_{g,i}|U_{g,i}, \gamma_g) = \lambda_0(T_{g,i}) \exp(\beta^T U_{g,i} + \alpha m(T_{g,i}, \varphi_g, \eta_{g,i}) + \gamma_g) \\ Y_{g,i,j} = m(t_j; \varphi_g, \eta_{g,i}) + \epsilon_{g,i,j} \\ \dots \end{cases}$$

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

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



Hierachical model

- Observation : T, Y

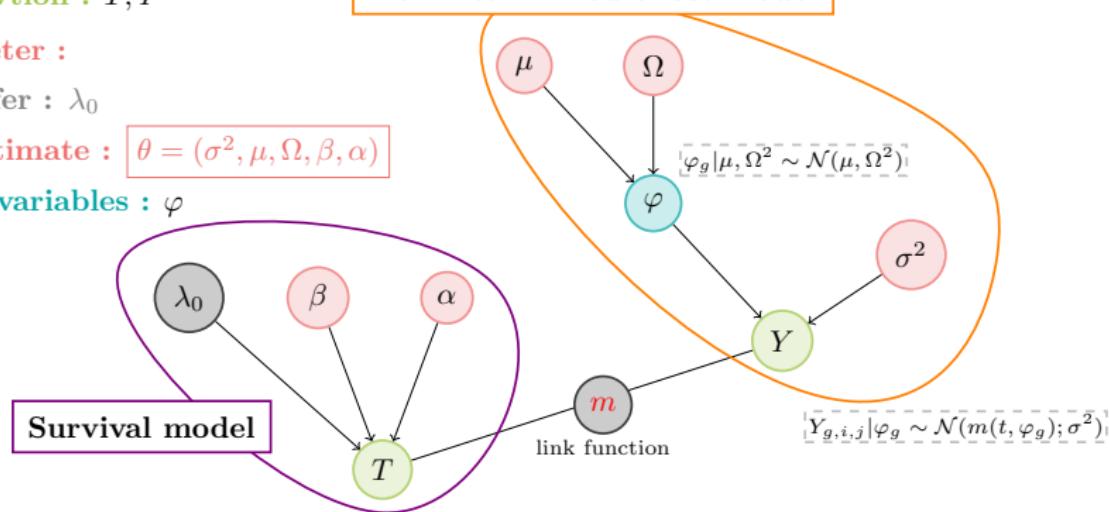
- Parameter :

- To infer : λ_0

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

- Latent variables : φ

Nonlinear mixed effect model



1 Introduction

- Flowering date
- Ostrina attack proportion
- Combining the two models

2 Methodology

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

3 Conclusion

- perspectives

Handle the high dimension of covariate

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



Expectation Maximisation algorithm

- **Require :** Starting point $\theta_0 \in \mathbb{R}^{100000}$ and a subset Θ of \mathbb{R}^{10000}
- At the iteration $k \geq 0$:

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

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

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

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

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

Problem : $Q(\theta|\theta_k)$ is not easy to compute due to the lack of observation of φ

Stochastic Approximation EM algorithm

- **Require :** Starting point $\theta_0 \in \mathbb{R}^{100000}$ and a subset Θ of \mathbb{R}^{10000}
- At the iteration $k \geq 0$:
 - ① **S-Step (Simulation)**, simulate $\varphi^{(k)}$ according to $f(\varphi|Y)$
 - ② **A-Step (stochastic Approximation)**, approximate the quantity :

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

- 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 $(u_k)_{k \in \mathbb{N}}$ is a sequence of positive step size such that $\sum_{k=1}^{\infty} u_k = \infty$ and $\sum_{k=1}^{\infty} u_k^2 < \infty$

1 Introduction

- Flowering date
- Ostrina attack proportion
- Combining the two models

2 Methodology

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

3 Conclusion

- perspectives

Conclusion and perspectives

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

