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 Caillebotte[1]
E.Kuhn[1] S.Lemler[2] E.Marchadier [3] J.Legrand [3]

[1]Université Paris-Saclay, INRAE, MaIAGE, [2]CentraleSupelec

[3]INRAE Génétique Quantitative et Evolution Le Moulon

Internship presentations
4 juillet 2022

Stat4Plant

université PARIS-SACLAY

INRAE

biological context

FIGURE 1 – ostrinia larva

FIGURE 2 – ostrinia attack

Photo credit : Sacha Revillon

Flowering date

- Flowering date of each corn plant observed

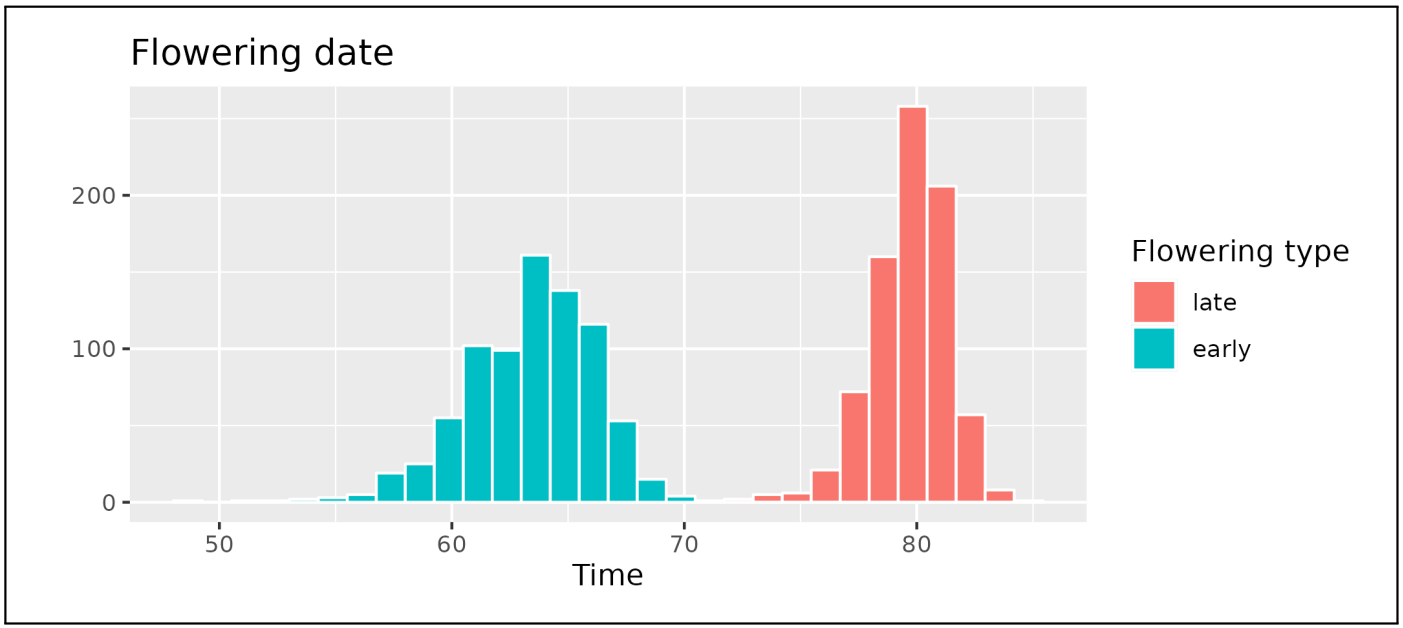


FIGURE 3 – Flowering date

- What impacts flowering : Genetic variability ? Insect attack ?

Survival analysis and 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 positive random variable.

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 :

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

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 behavior
- U_i : covariate for individual i
- β regression parameter

Modeling the flowering date

- For any plant $i \in \{1, ..., I\}$
  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 to be estimated .

Current population parameter : $\boxed{\theta = \beta}$

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

Ostrina attack proportion

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

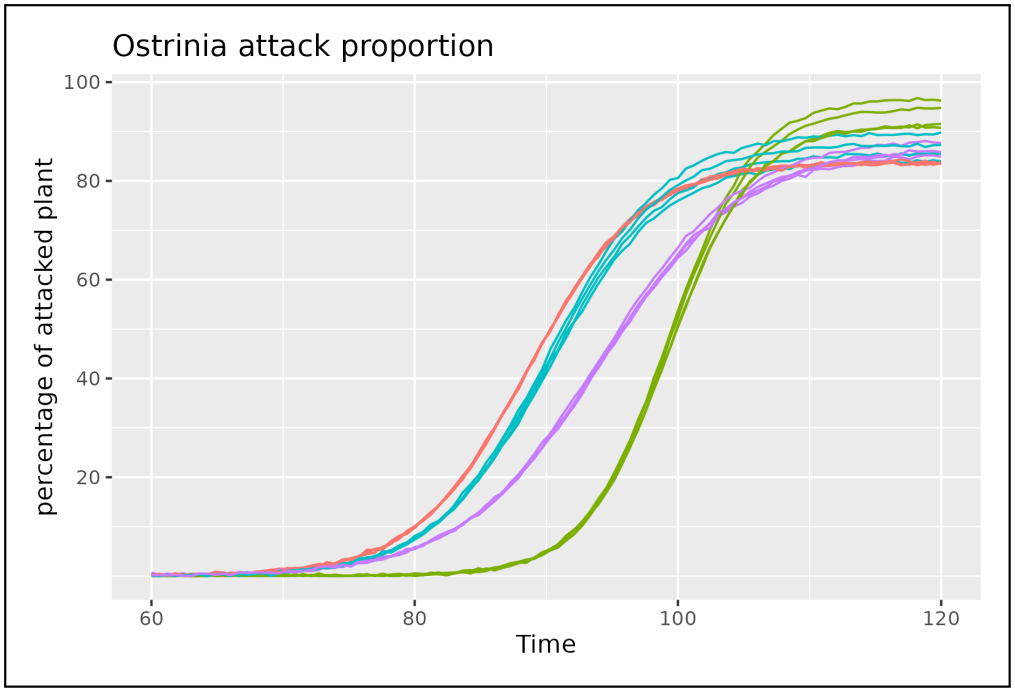


FIGURE 4 – Ostrina attack proportion

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

Non-linear mixed-effects model (NLME)

- Longitudinal data modeling :

behavior based on genetics

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

  - $Y_{i,j} \in \mathbb{R}$: 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 :

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

  - $\mu \in \mathbb{R}^3$: 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 :

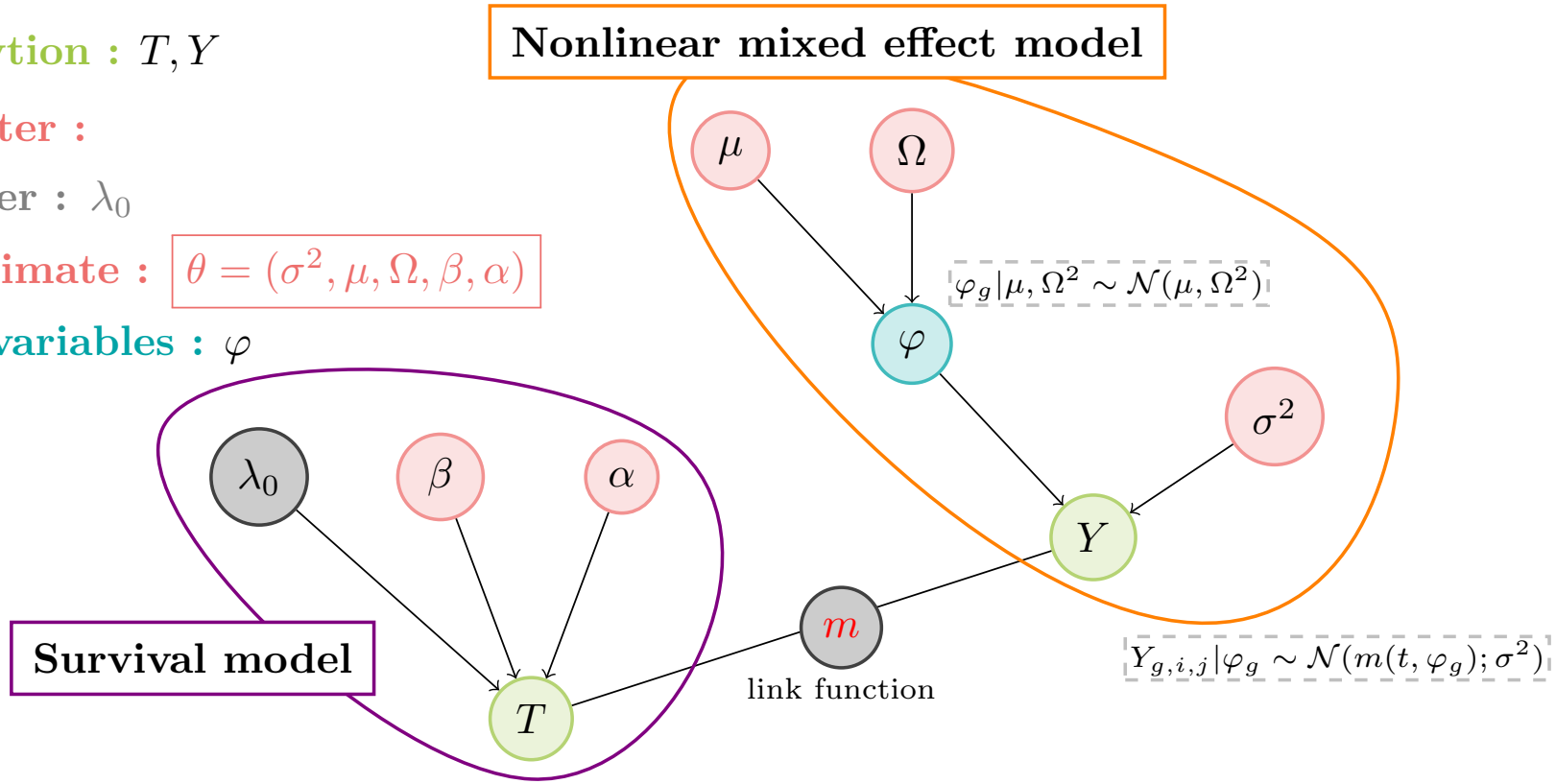
$$\begin{cases} \lambda(T_{g,i}|U_{g,i},\gamma_g) &= \lambda_0(T_{g,i})\exp(\beta^T U_{g,i} + \alpha m(T_{g,i},\varphi_g,\eta_{g,i}) + \gamma_g) \\ Y_{g,i,j} &= m(t_j;\varphi_g,\eta_{g,i}) + \epsilon_{g,i,j} \\ &\dots \end{cases}$$

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



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 $\theta_0 \in \mathbb{R}^{100000}$ and a subset Θ of $\mathbb{R}^{10000}$
- At the iteration $k \geqslant 0$:
  1. **E-Step (Expectation)**, evaluate the quantity :

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

  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 $k \geqslant 0$:
  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 $\sum_{k=1}^{\infty} u_k^2 < \infty$

Perspectives

Les algorithmes proximaux c'est cool !

Thank you for your attention !

Appendix 1 : full model

- **Obseravtion :** T, Y
- **Parameter :**
  - **Fixed hyperparameters :** $\omega_{\mathbf{a}}^{\mathbf{2}}, \omega_{\mathbf{b}}^{\mathbf{2}}$
  - **To estimate :** $\theta = (\sigma^2, \mu, \Omega, \omega_\eta^2, \omega_\gamma^2, \beta, \alpha)$
- **Latent variables :** $Z = (\varphi, \eta, \gamma, a, b)$