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

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

[3]INRAE Génétique Quantitative et Evolution Le Moulon

Internship presentations
4 juillet 2022

Stat4Plant

université PARIS-SACLAY

INRAE

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

Synchronous development of ostrinia and corn

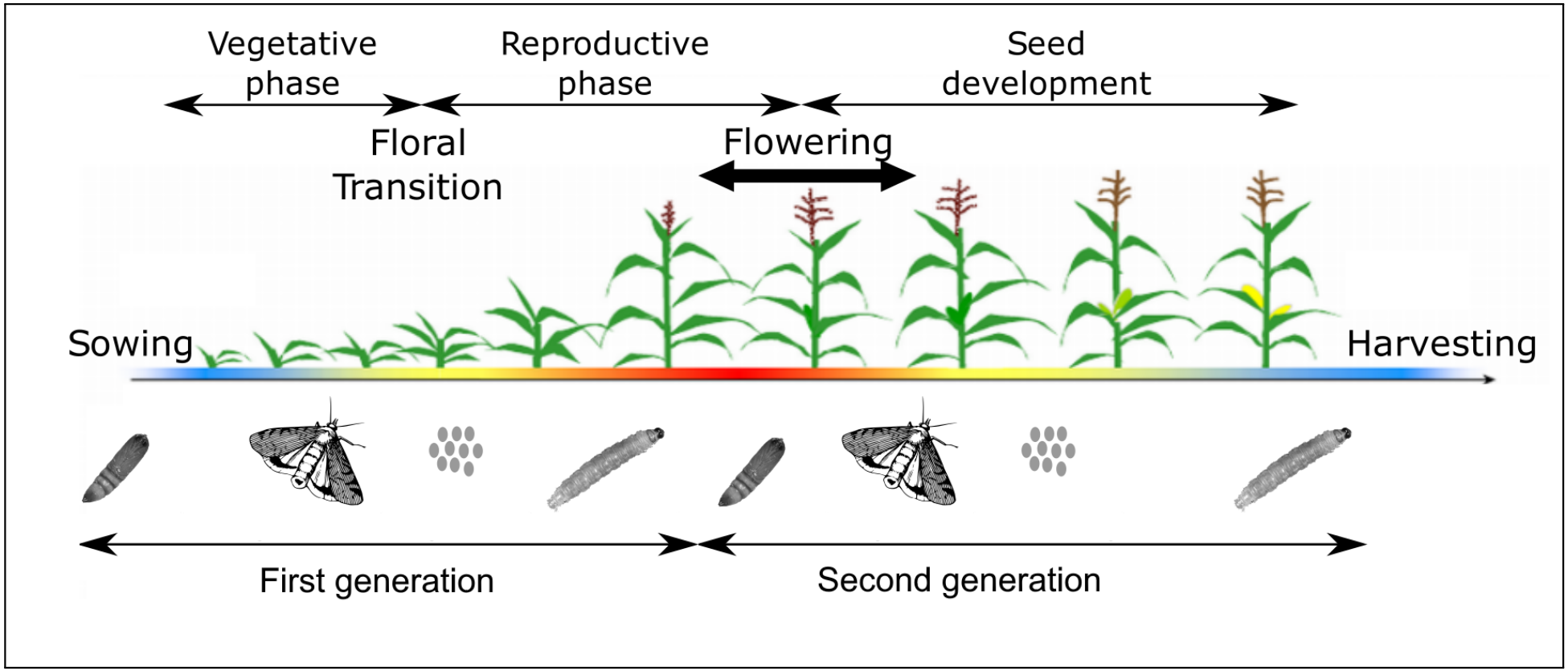


FIGURE 3 – Synchronization of the life cycles of the ostrinia and corn

Divergent selection

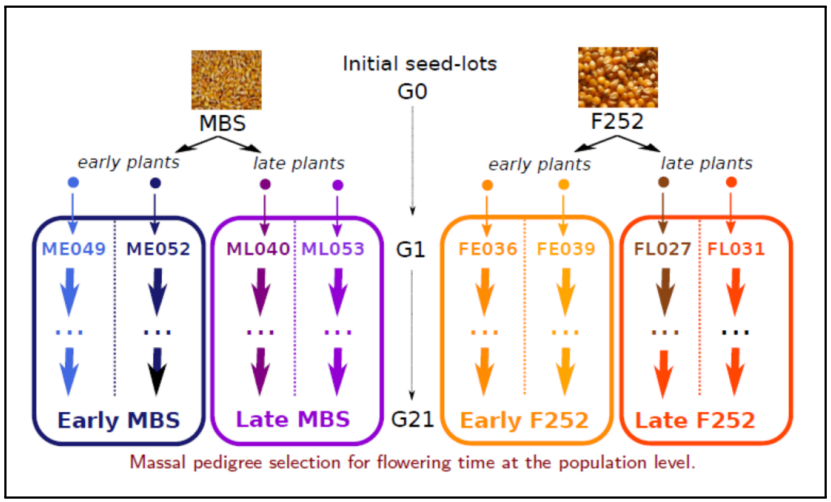


FIGURE 4 – Selection

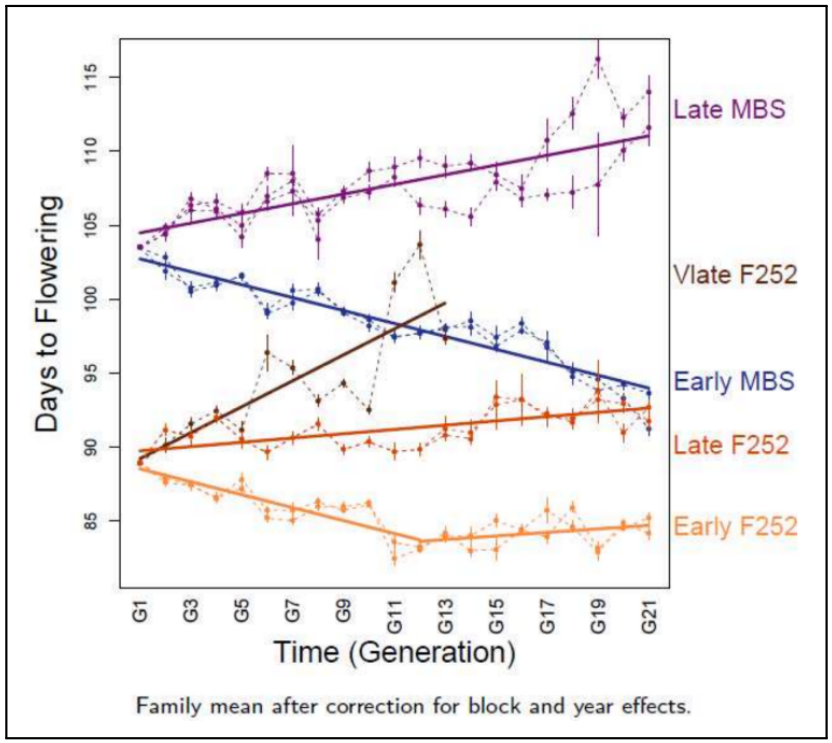


FIGURE 5 – Divergence

Flowering date

- Flowering date of each corn plant observed

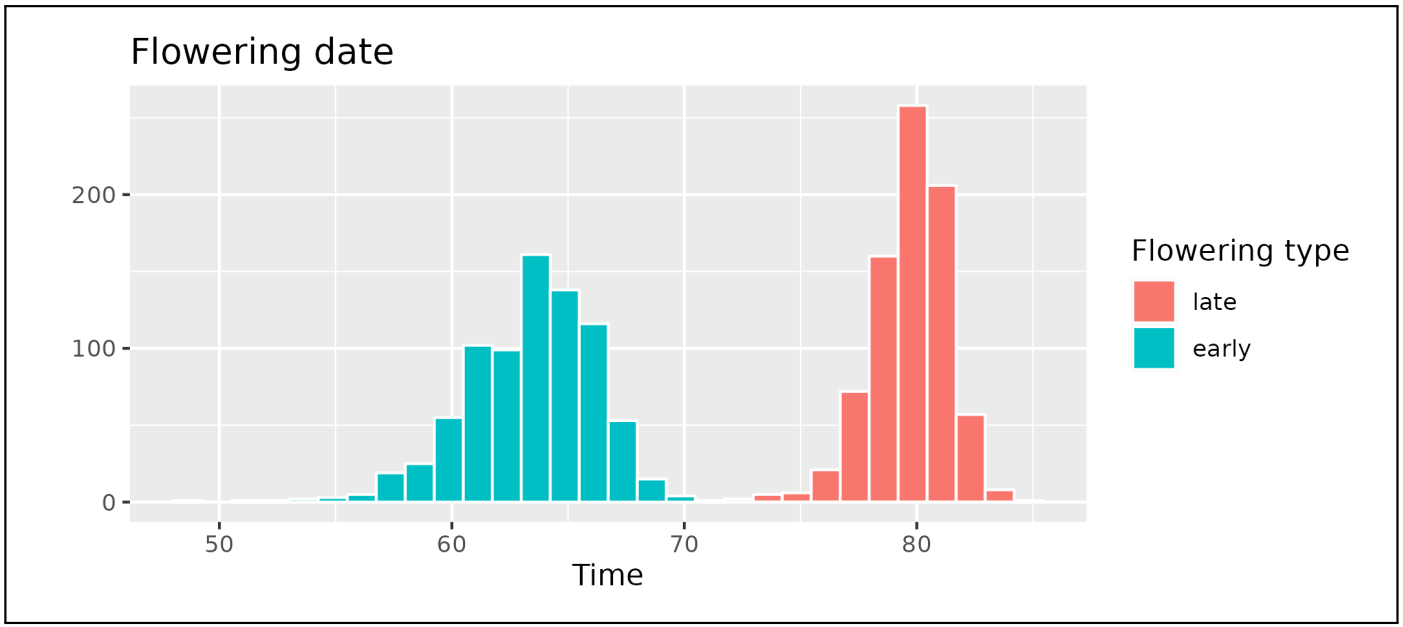


FIGURE 6 – Flowering date

- What impacts flowering : Genetic variability ? Insect attack ? Environmental conditions ?

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 :

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

Cox Model

Hazard function λ

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

Regression model that links the survival time to explanatory variables and assume a common behavior.
The Hazard function is given by :

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

- T time event
- 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 :

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

Flowering date

Covariates

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

Model parameter : $\boxed{\theta = \beta}$

- **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

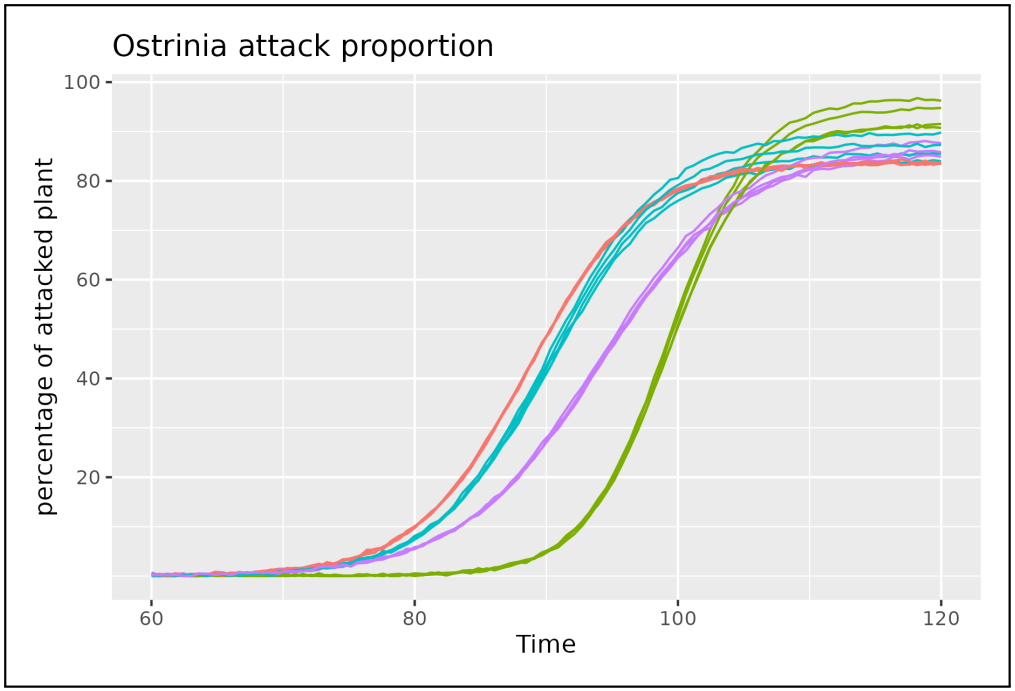


FIGURE 7 – ostrinia attack proportion

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

Non-linear mixed-effects model (NLME)

- Longitudinal data modeling :

behavior based on genetics

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

  - $Y_{g,j} \in \mathbb{R}$: j-th response of the g-th individual at time $t_{g,j}$ (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 : $\boxed{\theta = (\sigma^2, \mu, \Omega^2)}$

Joint Model : NLME and Survival model

- Combining the two models using a link function m.

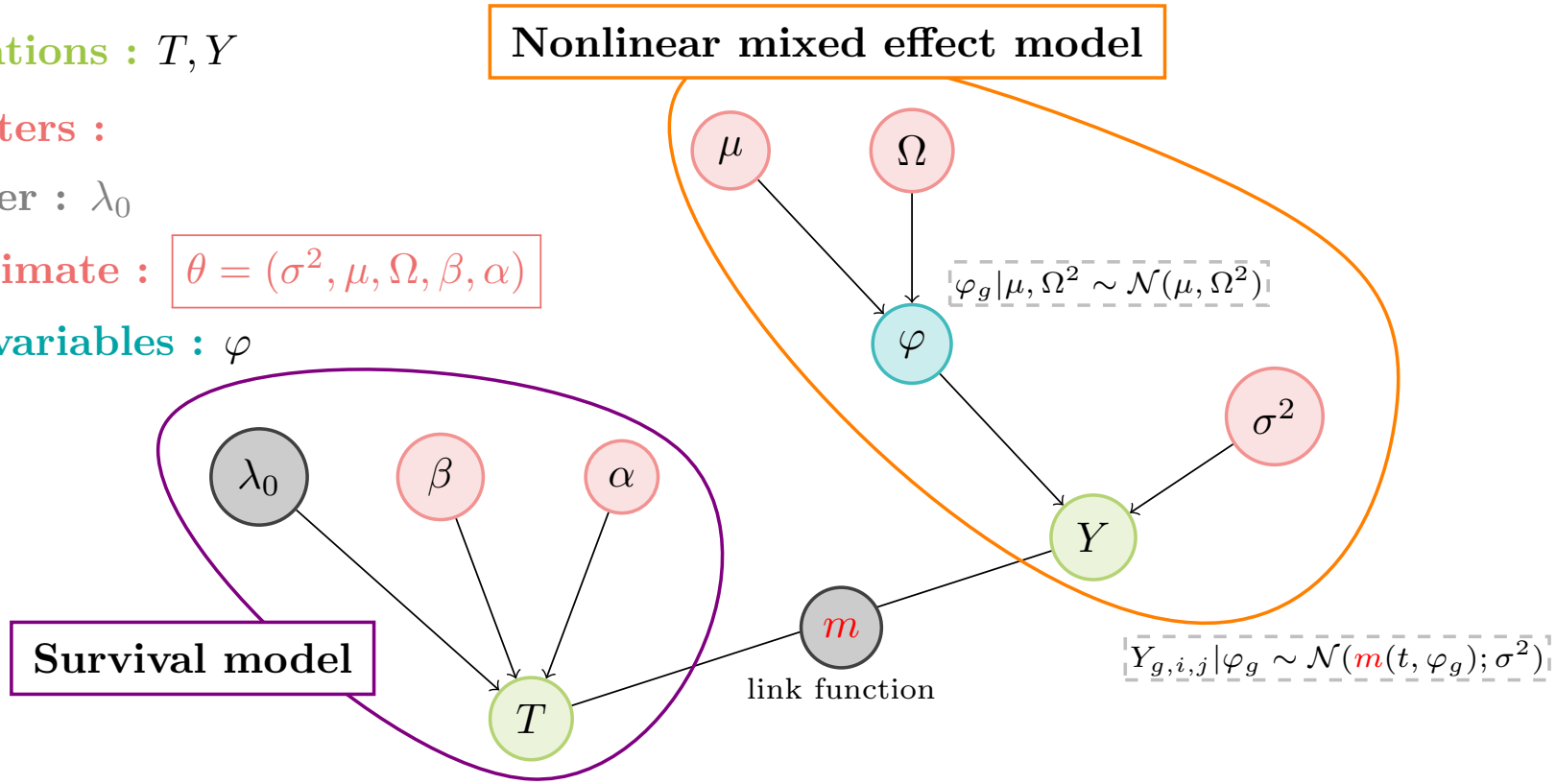
$$\begin{cases} \lambda(T_g|U_g,\gamma_g) &= \lambda_0(T_g)\exp(\beta^T U_g + \alpha m(T_g,\varphi_g)) \\ Y_{g,j} &= m(t_{g,j};\varphi_g) + \epsilon_{g,j} \\ \dots & \end{cases}$$

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

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

Hierachical model

- **Observations :** T, Y
- **Parameters :**
  - To infer : λ_0
  - To estimate : $\theta = (\sigma^2, \mu, \Omega, \beta, \alpha)$
- **Latent variables :** φ



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.=> lasso

Expectation Maximisation algorithm

- **Require :** Starting point θ_0
- At the iteration $k \geqslant 0$:
  1. **E-Step (Expectation)**, evaluate the quantity :

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

  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 $k \geqslant 0$:
  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
- intergrating proximal algorthm to deal with high dimension of β
- analyse real data

Thank you for your attention !

Appendix 1 : full model

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