Réunion de rentrée

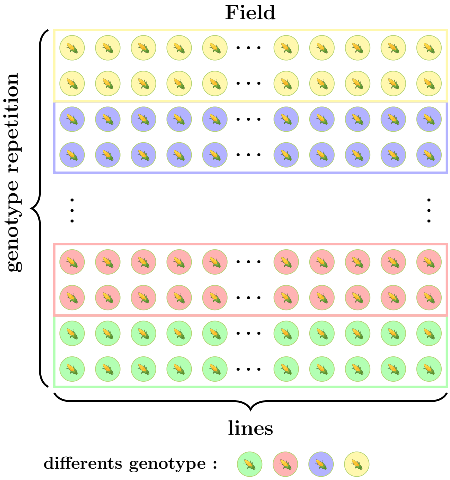
Antoine Caillebotte[1]
E.Kuhn[1] S.Lemler[2] E.Marchadier[3] J.Legrand[3]

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

[3]INRAE Génétique Quantitative et Evolution Le Moulon

14 octobre 2022

Notation



We denote by :

- G : number of genotype total,
- L : number of lines,
- J : number of repeated measures of the attacks.

1. Modeling
   - Cox Model
   - Non-linear mixed-effects model
   - Joint modeling
2. Inference in the joint Model
   - Expectation Maximisation algorithm
   - Penalized SAEM algorithm
   - Proximal Gradient Descent
3. Simulation study
   - Methodology
   - Results

Cox Model

Reference : Cox 1972

Hazard function h

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

Regression model that links the survival time to explanatory variables.
The Hazard function is given by :

$$h(T|U) = h_0(T) \exp(\beta^T U)$$

- T survival time
- h_0 : baseline hazard
- U : covariate for an individual
- β : regression parameter

Modeling the flowering date

- For any genotype $1 \leqslant g \leqslant G$ and line $1 \leqslant \ell \leqslant L$
  Hazard that the flowering occurs at $T_{g,\ell}$:

$$\underbrace{h(\overbrace{T_{g,\ell}}^{\text{Flowering date}}|U_{g\ell})}_{}\quad h(T_{g,\ell}|\underbrace{U_{g\ell}}_{\text{Covariates}}) = h_0(T_{g,\ell})\exp(\beta^T U_{g\ell})$$

  - $T_{g,\ell} \in \mathbb{R}$: time event of interest observed,
  - h_0 baseline hazard unknown,
  - $U_{g\ell} \in \mathbb{R}^p$: ℓ-th line of the g-th genotype's covariates known,
  - $\beta \in \mathbb{R}^p$: fixed effect unknown.

Model parameters : $\boxed{\theta = (h_0, \beta)}$

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

Non-linear mixed-effects model (NLME)

Reference : Davidian et Giltinan 1995

- Longitudinal data modeling : For any $1 \leqslant g \leqslant G$, $1 \leqslant \ell \leqslant L$ and $1 \leqslant j \leqslant J$

behavior based on genetics

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

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

- Inter-individual variation :

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

  - $\mu = (\mu_1, \mu_2, \mu_3) \in \mathbb{R}^3, \Omega = diag(\omega_1^2, \omega_2^2, \omega_3^2) \in \mathcal{M}_3(\mathbb{R})$: unknown,

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

Joint Model : NLME and Survival model

Reference : Rizopoulos 2012

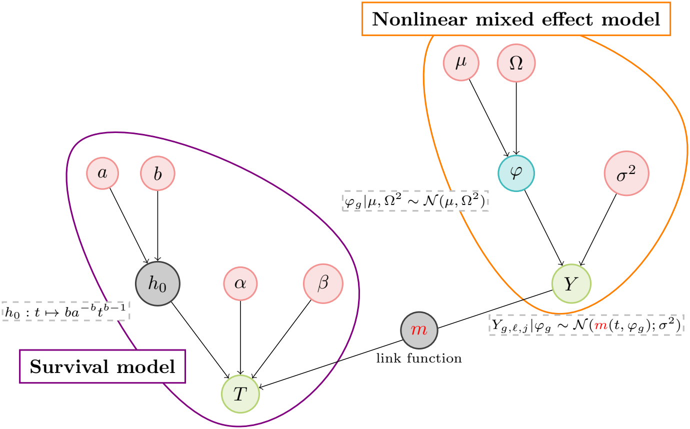
- Combining the two models using the link function m.
  For any $1 \leqslant g \leqslant G$, $1 \leqslant \ell \leqslant L$ and $1 \leqslant j \leqslant J$

$$\begin{cases} h_{g,\ell}(T_{g,\ell}|U_{g\ell}) = h_0(T_{g,\ell}) \exp(\beta^T U_{g\ell} + \alpha m(T_{g,\ell}; \varphi_g)) \\ Y_{g,\ell,j} = m(t_j; \varphi_g) + \epsilon_{g,\ell,j} \\ \varphi_g \underset{i.i.d.}{\sim} \mathcal{N}(\mu, \Omega) \quad ; \quad \epsilon_{g,\ell,j} \underset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2) \end{cases} \tag{1}$$

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

Model parameters : $\boxed{\theta = (\sigma^2, \mu, \Omega, h_0, \beta, \alpha)}$

Hierachical model



1. Modeling
   - Cox Model
   - Non-linear mixed-effects model
   - Joint modeling
2. Inference in the joint Model
   - Expectation Maximisation algorithm
   - Penalized SAEM algorithm
   - Proximal Gradient Descent
3. Simulation study
   - Methodology
   - Results

General estimate in latent variable joint model

$$\begin{cases} h_{g,\ell}(T_{g,\ell}|U_{g\ell}) = h_0(T_{g,\ell})\exp(\beta^T U_{g\ell} + \alpha m(T_{g,\ell};\varphi_g)) \\ Y_{g,\ell,j} = m(t_j;\varphi_g) + \epsilon_{g,\ell,j} \end{cases} \quad (2)$$

With : $\boxed{\theta = (\sigma^2, \mu, \Omega, h_0^{(a,b)}, \beta, \alpha)}$

Marginal likelihood written with complete likelihood

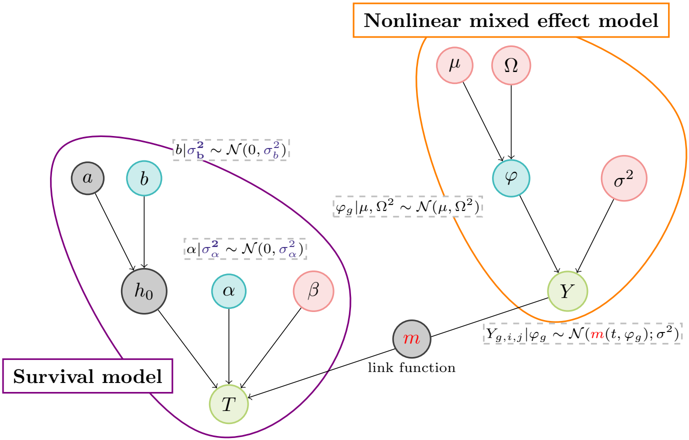
$$\mathcal{L}_{marg}(\theta|T,Y) = \int \mathcal{L}_{comp}(\theta|T,Y;\varphi)d\varphi$$

Recall : φ is not observed

Maximum likelihood Estimator (MLE)

$$\boxed{\hat{\theta} = \underset{\theta\in\Theta}{\arg\max}\, \mathcal{L}_{marg}(\theta|T,Y)} \quad (3)$$

Hierachical model adapted for the exponential family



SAEM - Exponential family

If we can write : $\log \mathcal{L}_{comp}(\theta; \varphi, b, \alpha; Y, T) = \langle \Phi(\theta); S(\varphi, b, \alpha)\rangle - \psi(\theta)$

- **Require :** Starting point θ_0
- At the iteration $0 \leqslant k \leqslant K$:
  1. **S-Step (Simulation)**, simulate $\varphi^{(k)}, b^{(k)}, \alpha^{(k)}$ according to $p(\varphi, b, \alpha | Y, T, \theta_k)$
  2. **A-Step (stochastic Approximation)**, evaluate :

$$S_{k+1} = (1 - u_k) S_k + u_k S(\varphi^{(k)}, b^{(k)}, \alpha^{(k)})$$

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

$$\theta_{k+1} = \arg\max_{\theta \in \Theta} \{\langle \Phi(\theta); S_{k+1}\rangle - \psi(\theta)\}$$

- **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 $\sum_{k=1}^{\infty} u_k^2 < \infty$

Estimate in joint model with covariates in high dimension

We separate the parameters in small dimension and those in large dimension :

$$\boxed{\theta = (\underbrace{\sigma^2, \mu, \Omega, b, \alpha}_{=\nu}, \beta) = (\nu, \beta)}$$

Variable selection $pen(\theta) = pen(\beta) = \lambda \|\beta\|_1 = \lambda \sum_{i=1}^{p} |\beta_i|$

Penalized Estimator

$$\boxed{(\hat{\nu}, \hat{\beta}) = \underset{\beta \in \mathbb{R}^p, \nu \in \mathbb{R}^d}{\arg\max} \{\mathcal{L}_{marg}(\nu, \beta | Y, T) - pen(\beta)\}}$$

Penalized SAEM algorithm

- **Require :** Starting point $\theta_0 = (\nu_0, \beta_0)$
- At the iteration $0 \leqslant k \leqslant K$:
  1. **S-Step (Simulation)**, simulate $\varphi^{(k)}, b^{(k)}, \alpha^{(k)}$ according to $p(\varphi, b, \alpha | Y, T, \nu_k, \beta_k)$
  2. **A-Step (stochastic Approximation)**, evaluate :

$$S_{k+1} = (1 - u_k) S_k + u_k S(\varphi^{(k)}, b^{(k)}, \alpha^{(k)})$$

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

$$\nu_{k+1} = \underset{\nu \in \mathbb{R}^d}{\arg\max} \{\langle \Phi(\nu, \beta_k); S_{k+1} \rangle - \psi(\nu, \beta_k)\}$$

$$\beta_{k+1} = \underset{\beta \in \mathbb{R}^p}{\arg\max} \{\langle \Phi(\nu_{k+1}, \beta); S_{k+1} \rangle - \psi(\nu_{k+1}, \beta) - pen(\beta)\}$$

- **Return :** $\hat{\theta} = (\nu_K, \beta_K)$ with K large enough

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

Proximal Gradient Descent

Gradient descent on the function $Q : \beta \mapsto \langle \Phi(\nu_{k+1}, \beta); S_{k+1} \rangle - \psi(\nu_{k+1}, \beta)$
Where ν_{k+1} is the current value in the SAEM algorithm
and S_{k+1} is the stochastic approximation of the sufficient statistic
Reference : Achab 2017

- **Require :** Starting point $\beta_0 \in \mathbb{R}^p$, the last compute of ν_{k+1} and S_{k+1}
- At the iteration $0 \leqslant k \leqslant K$:
  1. $\omega_k \leftarrow \beta_{k-1} - \gamma_k \dfrac{\nabla Q(\beta_{k-1})}{\|\nabla Q(\beta_{k-1})\|_2}$
  2. $\beta_k \leftarrow prox_{\gamma_k pen}(\omega_k)$
- **Return :** $\hat{\beta} = \beta_K$ with K large enough

where $(\gamma_k)_{k\in\mathbb{N}}$ is a sequence of steps and $\gamma_k > 0$

Proximal Operator

The proximal operator (Moreau 1962 ; Rockafellar 1976) defined below extends the gradient descents to non-differentiable functions.

Proximal operator

$$prox_{pen}(\beta) = \underset{\beta' \in \mathbb{R}^p}{\arg\min} \left(pen(\beta') + \frac{1}{2} \left\| \beta - \beta' \right\|_2^2 \right)$$

With Lasso penalization, $pen(\beta) = \|\beta\|_1$, we have the explicit form :

$$\boxed{(prox_{lasso}(\beta))_i = \begin{cases} 0 & \text{if } |\beta_i| < \lambda \\ \beta_i - \lambda & \text{if } \beta_i \geqslant \lambda \\ \beta_i + \lambda & \text{if } \beta_i \leqslant -\lambda \end{cases}} \tag{4}$$

Methodology

1. Simulate one single data set with $G = 40$, $L = 4$, $J = 10$ and $p = 1000$
2. Run SAEM resolution with 200 iterations

$$\boxed{\tilde{\theta}_{Lasso} = \arg\max_{\theta \in \Theta} \{\mathcal{L}_{marg}(\theta | T, Y) - pen_{Lasso}(\theta)\}}$$

3. Reduce the model with the variables selected by the Lasso, p « 1000
4. Run SAEM resolution with 200 iterations

$$\boxed{\hat{\theta}_{MLE} = \arg\max_{\theta \in \Theta} \mathcal{L}_{marg}(\theta | T, Y)}$$

Results

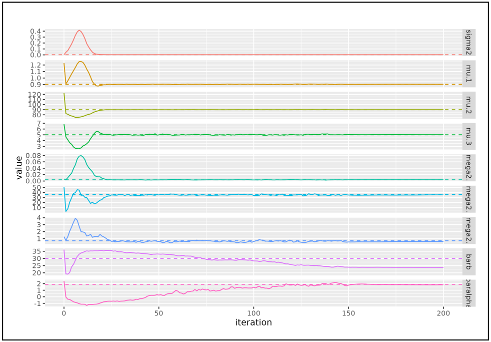


Figure 1 – $\tilde{\theta}_{Lasso}$ based on SAEM iterations on a single dataset ($G = 40, L = 4, J = 10$)

Estimate of the parameter with Lasso penalization

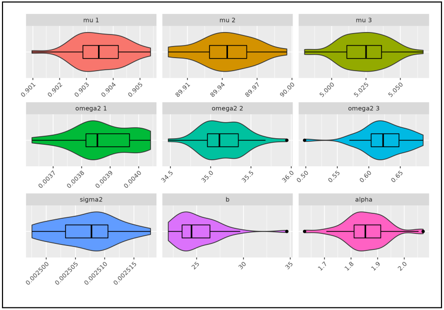


FIGURE 2 – $\tilde{\theta}_{Lasso}$ for 40 SAEM on a single dataset ($G = 40, L = 4, J = 10$) with $\lambda = \dfrac{1}{\sqrt{GL}}$

Variable selection procedure

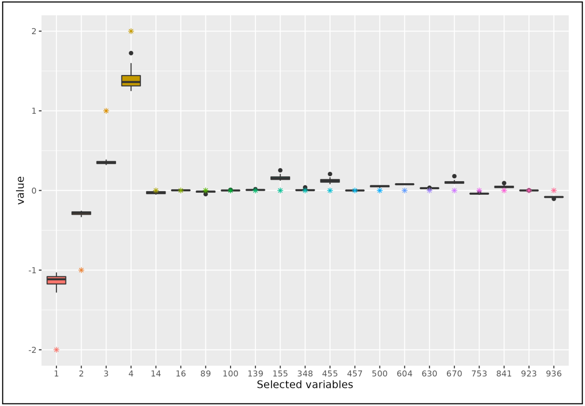


FIGURE 3 – $\tilde{\beta}_{Lasso} \in \mathbb{R}^{1000}$ for 40 SAEM on a single dataset ($G = 40, L = 4, J = 10$) with $\lambda = \dfrac{1}{\sqrt{GL}}$

Effect of the regularization choice

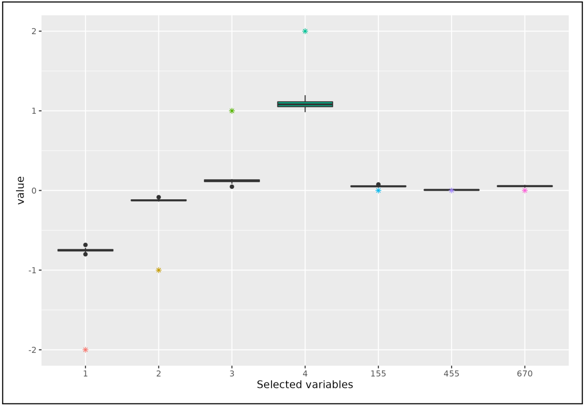


FIGURE 4 – $\tilde{\beta}_{Lasso} \in \mathbb{R}^{1000}$ for 40 SAEM on a single dataset ($G = 40, L = 4, J = 10$) with $\lambda = \dfrac{1.2}{\sqrt{GL}}$

MLE of β after the variable selection

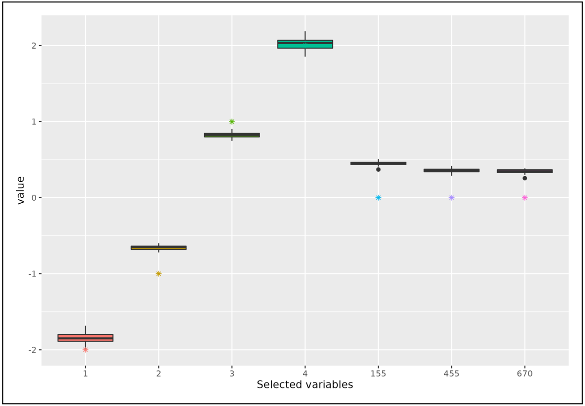


FIGURE 5 – $\hat{\beta}_{MLE} \in \mathbb{R}^7$ for 40 SAEM on a single dataset ($G = 40$, $L = 4$, $J = 10$)

Thank you for your attention!