Estimation and variable selection in joint model of survival times and longitudinal data

Antoine Caillebotte[1]
E.Kuhn[1] S.Lemler[2]

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

[3]INRAE Génétique Quantitative et Evolution Le Moulon

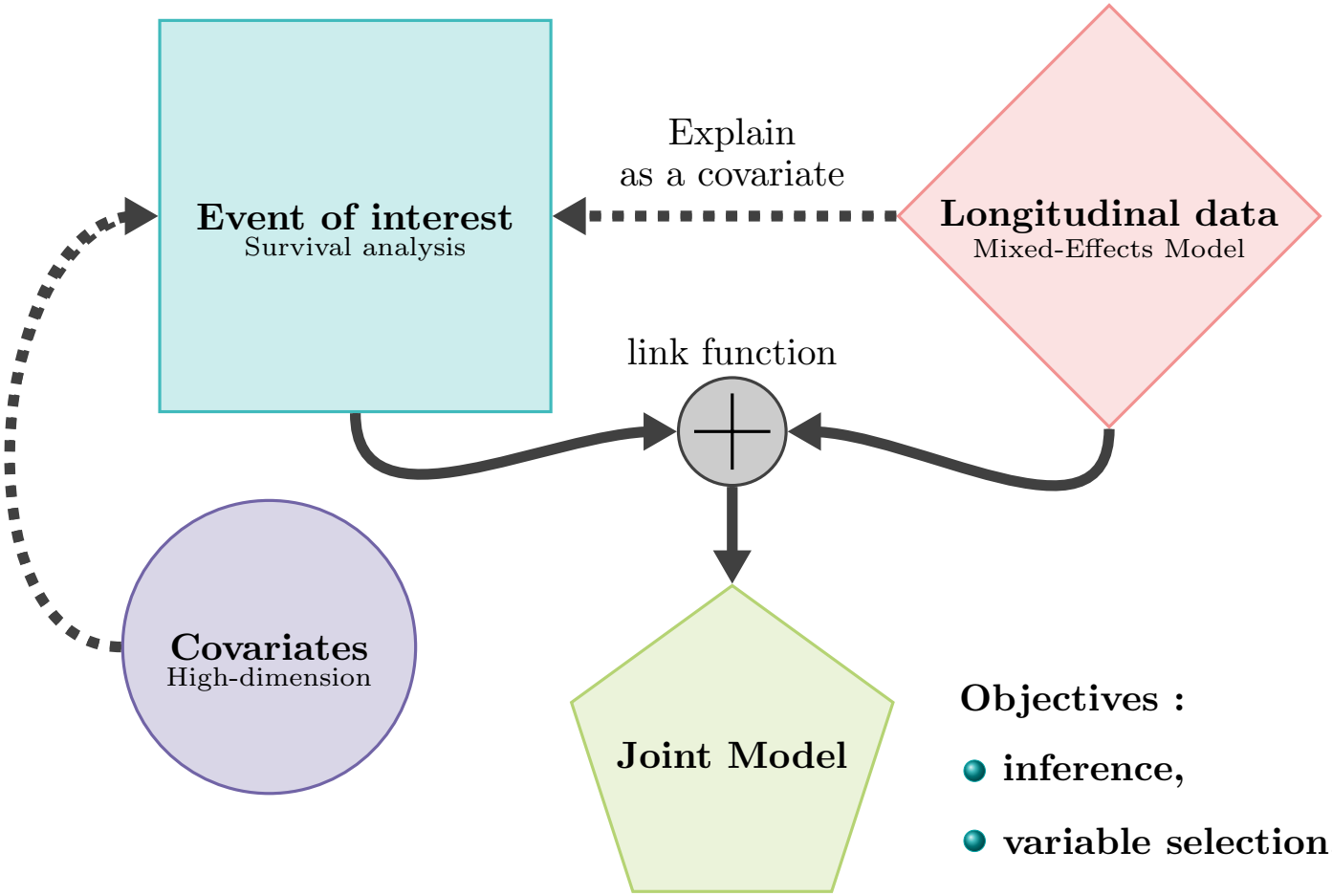
July 3rd, 2023 - EMS presentation

Stat4Plant

université PARIS-SACLAY

INRAE

Introduction



Objectives :

- **inference,**
- **variable selection.**

Survival analysis

Focus on **time to event of interest**

- In medicine : Time of remission / time to death

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 of interest.

Hazard function

Let T **be a positive random variable.** For $t \geqslant 0$, $h(t)$ probability that the event occurs in a small time interval after t, knowing that it did not occur until time t :

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

Cox model

Reference : Cox 1972

- For any individual $1 \leqslant i \leqslant N$
  Regression model that links the survival time to explanatory variables.
  The Hazard function at T_i is given by :

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

(Event of interest → T_i; Covariates → U_i)

  - $T_i \in \mathbb{R}$: time event of interest observed,
  - h_0 baseline hazard unknown,
  - $U_i \in \mathbb{R}^p$: the i-th individual's covariates known,
  - $\beta \in \mathbb{R}^p$: regression parameter vector unknown.

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

- **Objective :** model the longitudinal data and link it to this model !

Non-linear mixed-effects model (NLME)

Reference : Davidian et Giltinan 1995

- Longitudinal data modeling : For any $1 \leqslant i \leqslant N$ and $1 \leqslant j \leqslant J$

behavior based on individuals variation

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

  - $Y_{i,j} \in \mathbb{R}$: j-th response of the i-th individual at time t_j observation,
  - m : nonlinear function for Z,
  - $Z_i \in \mathbb{R}^d$: random effects (parameters of m) not observed.

- Inter-individual variation :

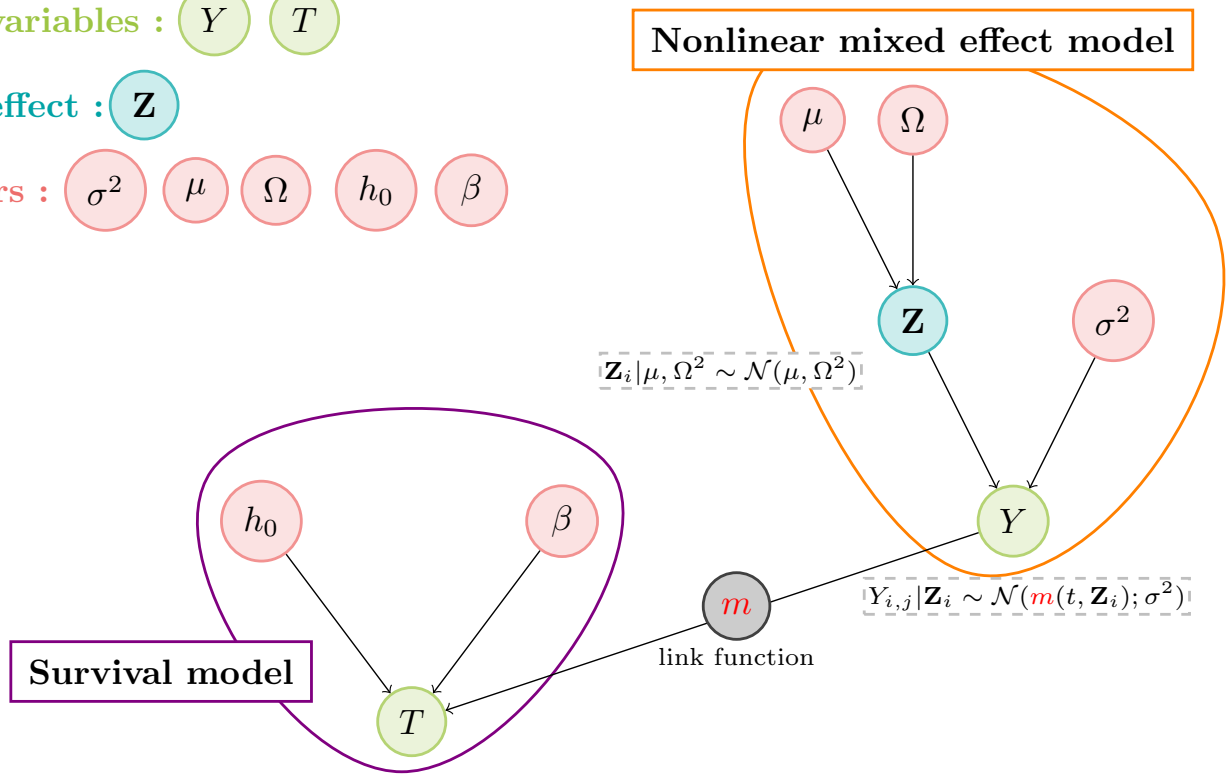
$$Z_i = \mu + \xi_i \; ; \; \xi_i \underset{i.i.d.}{\sim} \mathcal{N}(0, \Omega)$$

  - $\mu = (\mu_1, \ldots, \mu_d) \in \mathbb{R}^d, \Omega = diag(\omega_1^2, \ldots, \omega_d^2) \in \mathcal{M}_d(\mathbb{R})$: unknown,

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

Hierachical model

- **Random variables :** Y T
- **Random effect :** $\mathbf{Z}$
- **Parameters :** σ^2 μ Ω h_0 β



Joint Model : NLME and Survival model

Reference : Rizopoulos 2012

- Combining the two models using the link function m.
  For any $1 \leqslant i \leqslant N$ and $1 \leqslant j \leqslant J$

$$\begin{cases} h(t|U_i) = h_0(t)\exp(\beta^T U_i + \alpha m(T_i; Z_i)) \ ; \ \forall t \geqslant 0 \\ Y_{i,j} = m(t_j; Z_i) + \epsilon_{i,j} \\ Z_i \underset{i.i.d.}{\sim} \mathcal{N}(\mu, \Omega) \quad ; \quad \epsilon_{i,j} \underset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2) \end{cases} \tag{1}$$

  where α quantifies the association between survival and longitudinal data.

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

General estimate in latent variable joint model

$$\begin{cases} h(t|U_i) = h_0(t)\exp(\beta^T U_i + \alpha m(t;Z_i)) \ ; \ \forall t \geqslant 0 \\ Y_{i,j} = m(t_j;Z_i) + \epsilon_{i,j} \end{cases} \tag{2}$$

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

Marginal likelihood written with complete likelihood

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

Recall : Z is not observed

Maximum likelihood Estimator (MLE)

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

Stochastic Gradient with FIM preconditionning

Reference : Baey, Delattre, Kuhn, Leger et Lemler 2023

- **Require :** Starting point θ_0, Δ^0
- At the iteration $0 \leqslant k < K$:

Stochastic Gradient with FIM preconditionning

Reference : Baey, Delattre, Kuhn, Leger et Lemler 2023

- **Require :** Starting point θ_0, Δ^0
- At the iteration $0 \leqslant k < K$:
  1. **Simulation step**, Draw $Z^{(k)} \sim p(.|T, Y, \theta_k)$ using a Gibbs sampler algorithm

Stochastic Gradient with FIM preconditionning

Reference : Baey, Delattre, Kuhn, Leger et Lemler 2023

- **Require :** Starting point θ_0, Δ^0
- At the iteration $0 \leqslant k < K$:
  1. **Simulation step**, Draw $Z^{(k)} \sim p(.|T, Y, \theta_k)$ using a Gibbs sampler algorithm
  2. **Gradient computation**, $v_k = \frac{1}{N} \sum_{i=1}^{N} \nabla \log \mathcal{L}_{comp}(\theta_k; T_i, Y_i, Z_i^{(k)})$

Stochastic Gradient with FIM preconditionning

Reference : Baey, Delattre, Kuhn, Leger et Lemler 2023

- **Require :** Starting point θ_0, Δ^0
- At the iteration $0 \leqslant k < K$:
  1. **Simulation step**, Draw $Z^{(k)} \sim p(.|T, Y, \theta_k)$ using a Gibbs sampler algorithm
  2. **Gradient computation**, $v_k = \frac{1}{N} \sum_{i=1}^{N} \nabla \log \mathcal{L}_{comp}(\theta_k; T_i, Y_i, Z_i^{(k)})$
  3. **Stochastic Approximation**, Compute $\Delta_i^{(k)} = (1 - \gamma_k)\Delta_i^{(k-1)} + \gamma_k \nabla \log \mathcal{L}_{comp}(\theta_{k-1}; T_i, Y_i, Z_i^{(k)})$
  4. **FIM computation**, $FIM_k = \frac{1}{N} \sum_{i=1}^{N} \Delta_i^{(k)} \left(\Delta_i^{(k)}\right)^T$

FIM estimation

Stochastic Gradient with FIM preconditionning

Reference : Baey, Delattre, Kuhn, Leger et Lemler 2023

- **Require :** Starting point θ_0, Δ^0
- At the iteration $0 \leqslant k < K$:
  1. **Simulation step**, Draw $Z^{(k)} \sim p(.|T, Y, \theta_k)$ using a Gibbs sampler algorithm
  2. **Gradient computation**, $v_k = \frac{1}{N} \sum_{i=1}^{N} \nabla \log \mathcal{L}_{comp}(\theta_k; T_i, Y_i, Z_i^{(k)})$
  3. **Stochastic Approximation**, Compute $\Delta_i^{(k)} = (1 - \gamma_k)\Delta_i^{(k-1)} + \gamma_k \nabla \log \mathcal{L}_{comp}(\theta_{k-1}; T_i, Y_i, Z_i^{(k)})$
  4. **FIM computation**, $FIM_k = \frac{1}{N} \sum_{i=1}^{N} \Delta_i^{(k)} \left(\Delta_i^{(k)}\right)^T$

  FIM estimation
  5. **Forward, gradient** $\theta_{k+1} = \theta_k + \gamma_k FIM_k^{-1} v_k$
- **Return :** $\hat{\theta} = \theta_K$ with K large enough

Estimate in joint model with covariates in high dimension

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

$$\theta = (\underbrace{\sigma^2, \mu, \Omega, \theta_{h_0}, \alpha}_{=\nu \in \mathbb{R}^d}, \beta) = (\nu, \beta)$$

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

Penalized Estimator :

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

Problem : $\mathcal{L}_{comp}(\nu, \beta | T, Y) - pen(\beta)$ is not differentiable.

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 :

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

Stochastic Proximal Gradient with FIM preconditionning

- **Require :** Starting point θ_0, $\Delta^{(0)}$
- At the iteration $0 \leqslant k < K$:
  1. **Simulation step**, Draw $Z^{(k)} \sim p(.|T, Y, \theta_k)$ using a Gibbs sampler algorithm
  2. **Gradient computation**, $v_k = \frac{1}{N} \sum_{i=1}^{N} \nabla \log \mathcal{L}_{comp}(\theta_k; T_i, Y_i, Z_i^{(k)})$
  3. **Stochastic Approximation** , Compute $\Delta_i^{(k)} = (1 - \gamma_k)\Delta_i^{(k-1)} + \gamma_k \nabla \log \mathcal{L}_{comp}(\theta_{k-1}; T_i, Y_i, Z_i^{(k)})$
  4. **FIM computation**, $FIM_k = \frac{1}{N} \sum_{i=1}^{N} \Delta_i^{(k)} \left(\Delta_i^{(k)}\right)^T$

Stochastic Proximal Gradient with FIM preconditionning

- **Require :** Starting point θ_0, $\Delta^{(0)}$
- At the iteration $0 \leqslant k < K$:
  1. **Simulation step**, Draw $Z^{(k)} \sim p(.|T, Y, \theta_k)$ using a Gibbs sampler algorithm
  2. **Gradient computation**, $v_k = \frac{1}{N} \sum_{i=1}^{N} \nabla \log \mathcal{L}_{comp}(\theta_k; T_i, Y_i, Z_i^{(k)})$
  3. **Stochastic Approximation** , Compute $\Delta_i^{(k)} = (1 - \gamma_k)\Delta_i^{(k-1)} + \gamma_k \nabla \log \mathcal{L}_{comp}(\theta_{k-1}; T_i, Y_i, Z_i^{(k)})$
  4. **FIM computation**, $FIM_k = \frac{1}{N} \sum_{i=1}^{N} \Delta_i^{(k)} \left(\Delta_i^{(k)}\right)^T$
  5. **Forward, gradient** $(\nu_{k+1}, \omega_{k+1})^T = (\nu_k, \beta_k)^T + \gamma_k FIM_k^{-1} v_k$
  6. **Backward, penalization**
     $\beta_{k+1} = \text{Prox}_{\gamma_k pen}(\omega_{k+1})$
     $\theta_{k+1} = (\nu_{k+1}, \beta_{k+1})$
- **Return :** $\hat{\theta} = \theta_K$ with K large enough

Methodology

- Simulate one data set with $G = 100$, $J = 20$ and $p = 100$
- Choose $\lambda = \underset{\lambda \in \Lambda}{\arg\min}\, BIC(\lambda)$

Methodology

- Simulate one data set with $G = 100$, $J = 20$ and $p = 100$
- Choose $\lambda = \underset{\lambda \in \Lambda}{\arg\min}\, BIC(\lambda)$

1. Run SPG resolution

$$\hat{\theta}_{\text{Lasso}} = \underset{\theta \in \Theta}{\arg\max} \left\{ \mathcal{L}_{marg}(\theta | T, Y) - pen_{Lasso}(\theta) \right\}$$

   - Reduce the model with the variables selected by the Lasso, $s \ll 100$

Methodology

- Simulate one data set with $G = 100$, $J = 20$ and $p = 100$
- Choose $\lambda = \underset{\lambda \in \Lambda}{\arg\min}\, BIC(\lambda)$

1. Run SPG resolution

$$\hat{\theta}_{\text{Lasso}} = \underset{\theta \in \Theta}{\arg\max} \{\mathcal{L}_{marg}(\theta | T, Y) - pen_{Lasso}(\theta)\}$$

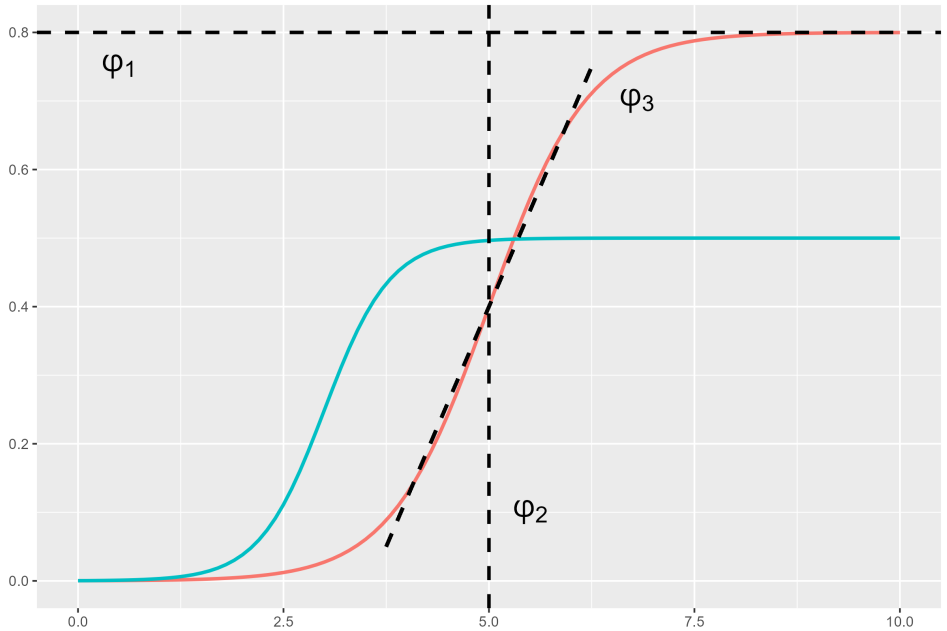
   - Reduce the model with the variables selected by the Lasso, $s \ll 100$

2. Run SG resolution

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

Non-linear function : Logistic

$$\text{Let } \varphi = (\varphi_1, \varphi_2, \varphi_3) \in \mathbb{R}^3 \quad m : t, \varphi \mapsto \frac{\varphi_1}{1 + \exp\left(\frac{\varphi_2 - t}{\varphi_3}\right)}$$



Where :
- φ_1 is the curve's asymptotic value,
- φ_2 is the value of the sigmoid midpoint,
- $\varphi_3 > 0$ is the logistic growth rate.

Results

	μ_1	μ_2	μ_3	ω_1^2	ω_2^2	σ^2	α
ν^*	0.3	90	7.5	2.5×10^{-3}	20	10^{-3}	11.11

	β_1	β_2	β_3	β_4	β_5	$\dots$	β_p
β^*	-2	-1	1	2	0	$\dots$	0

TABLE 1 – Simulation parameters

Results example

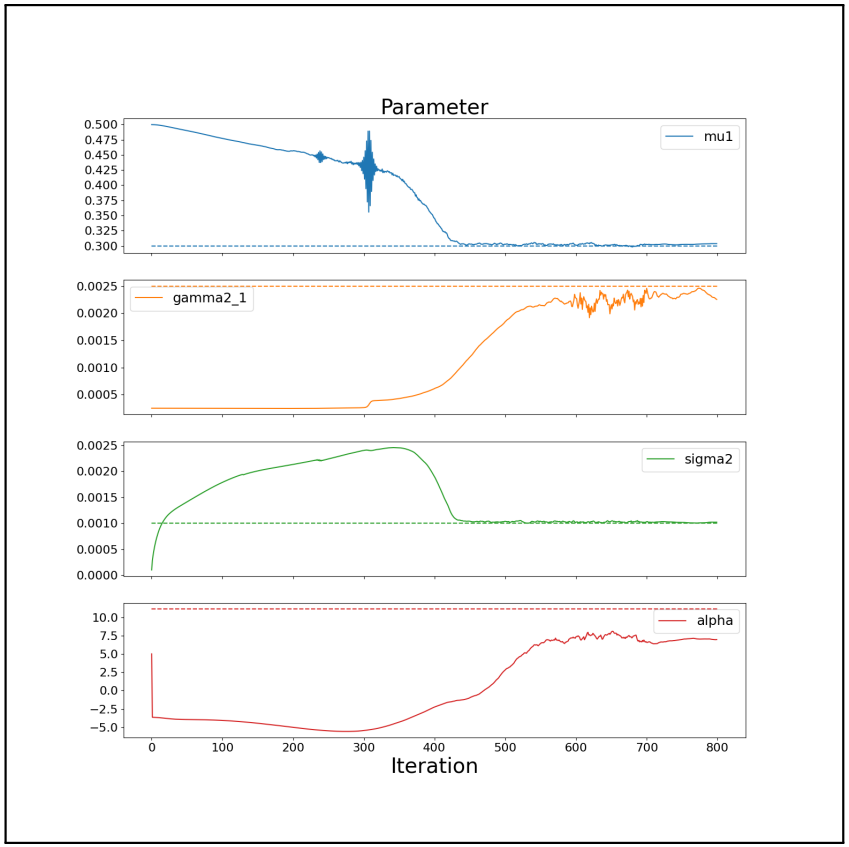


FIGURE 1 – $\hat{\theta}_{\text{Lasso}}$ during SPG-FIM iterations

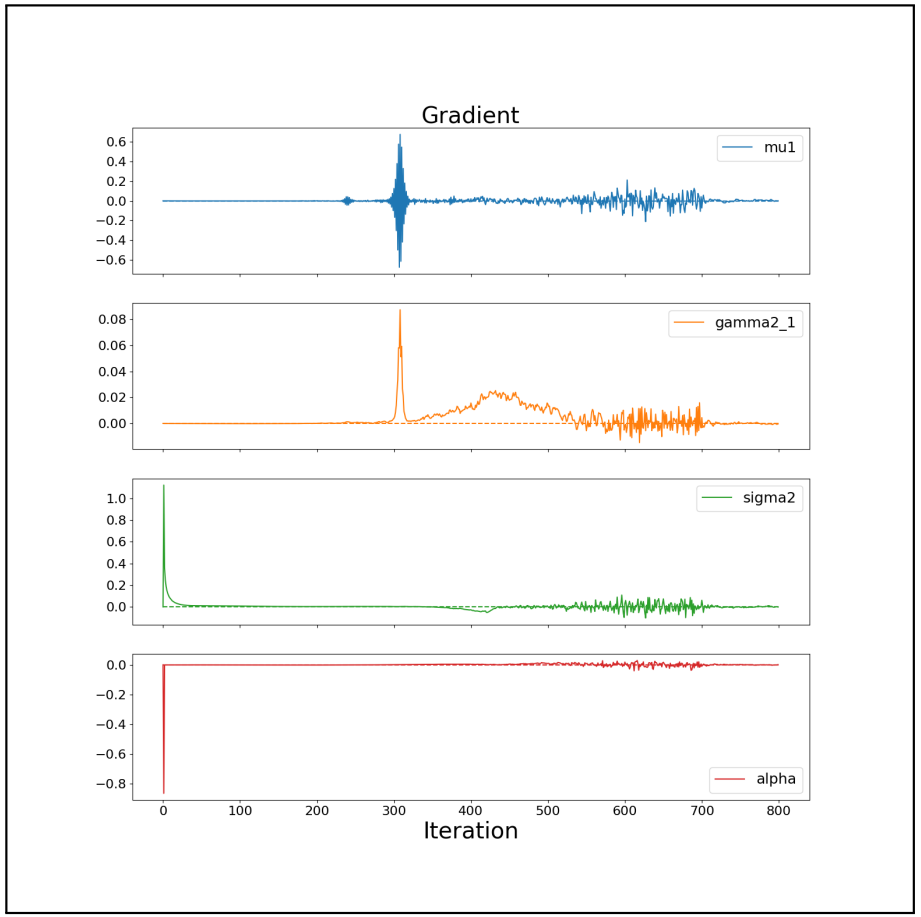


FIGURE 2 – Gradient value during SPG-FIM iterations

Results example

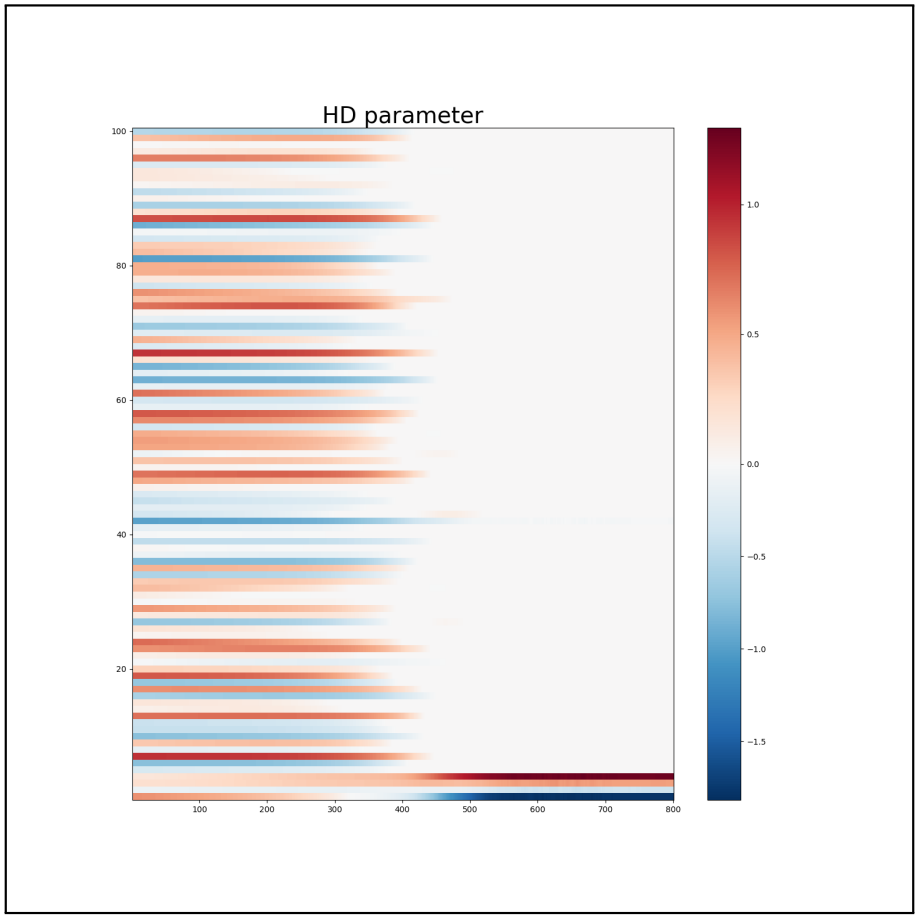


FIGURE 3 – Tile representation of the selected variables $\hat{\beta}_{\text{Lasso}}$ during an SPG-FIM

Variable selection procedure

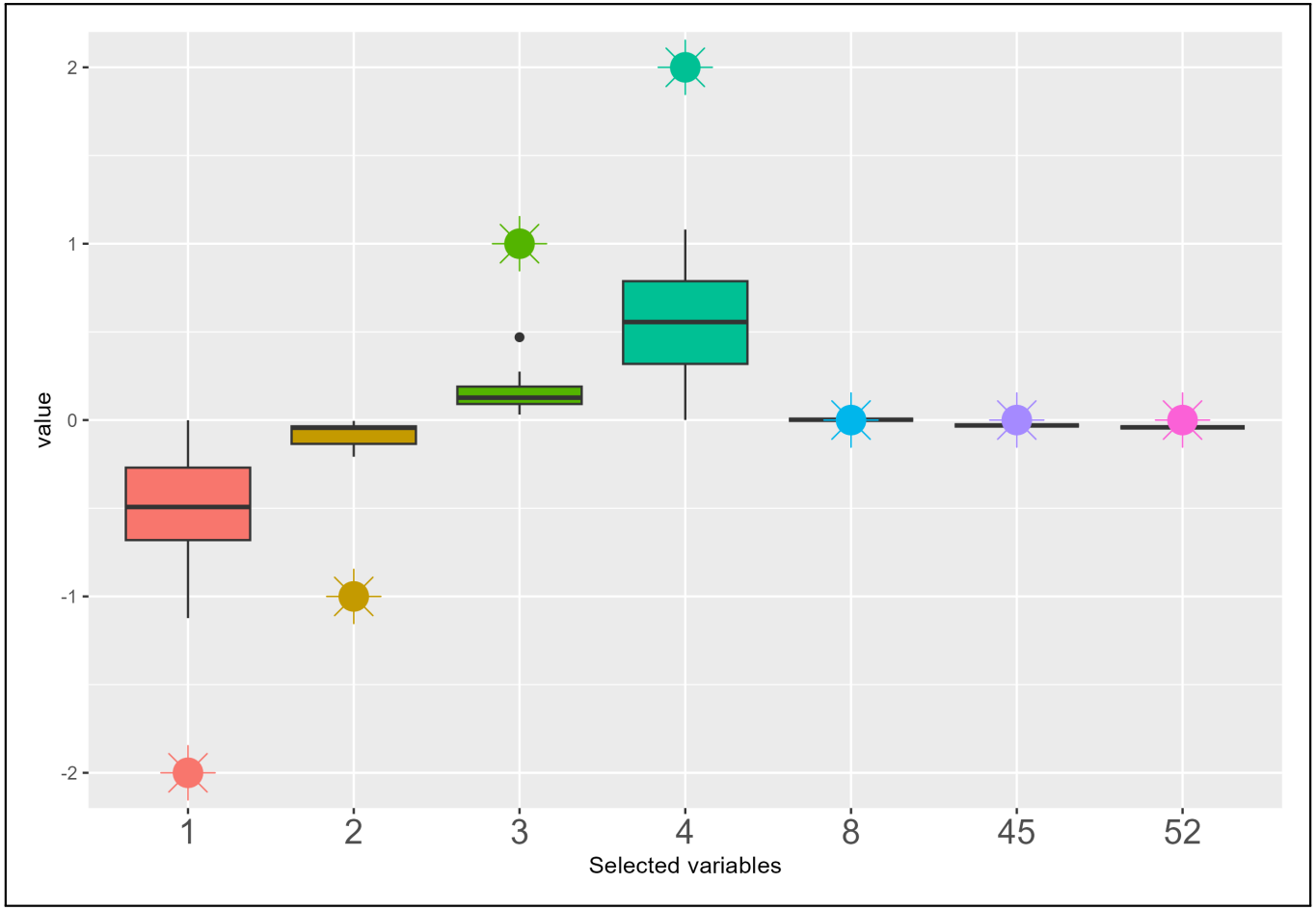


FIGURE 4 – $\hat{\beta}_{\text{Lasso}} \in \mathbb{R}^7$ for 50 SPG ($G = 100$, $J = 20$)
The colored stars indicate the value used during the data simulation

MLE of β after the variable selection

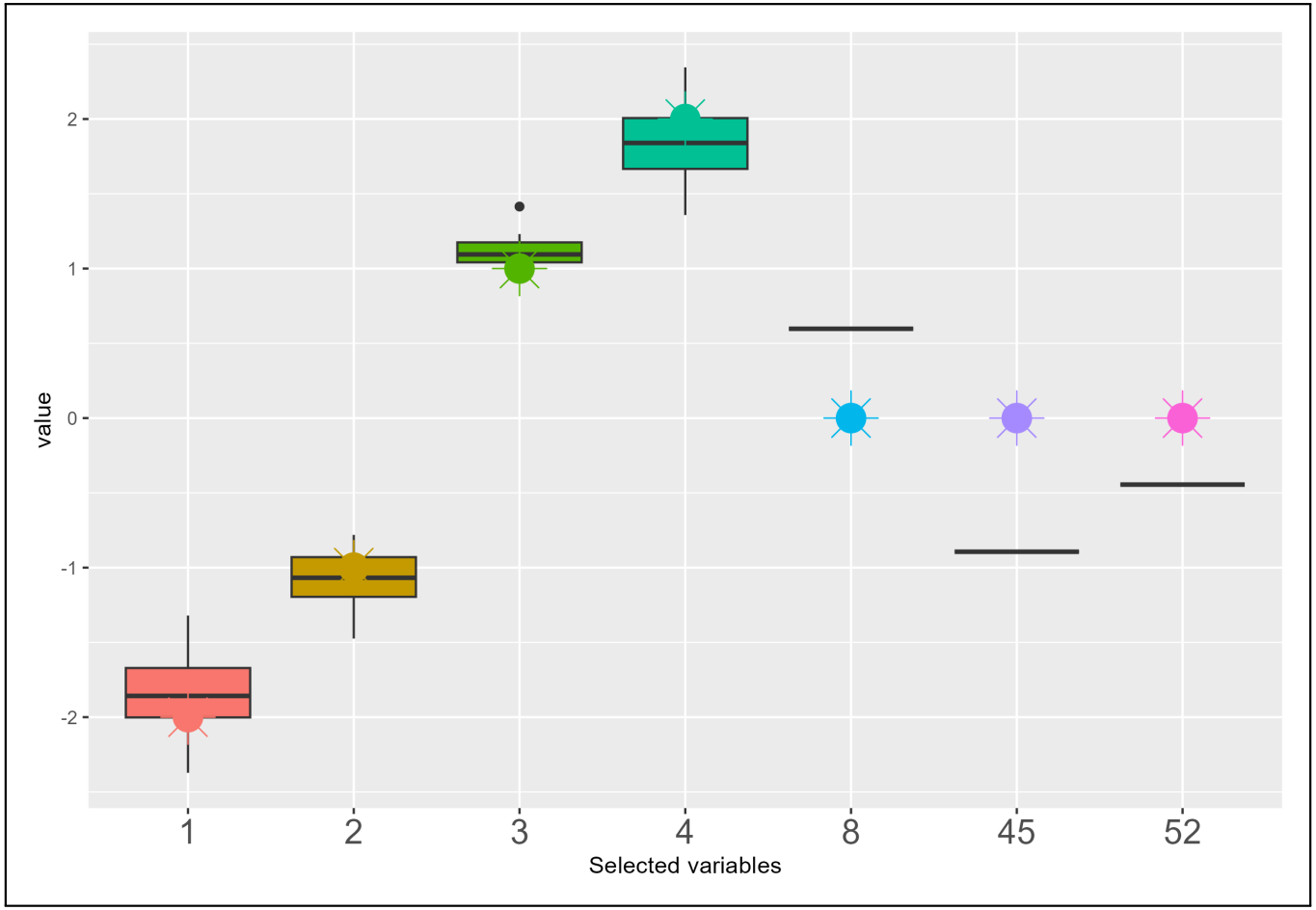


FIGURE 5 – $\hat{\theta}_{\text{MLE}} \in \mathbb{R}^7$ for 50 SPG ($G = 100$, $J = 20$)
The colored stars indicate the value used during the data simulation

Estimate of the parameter with Lasso penalization

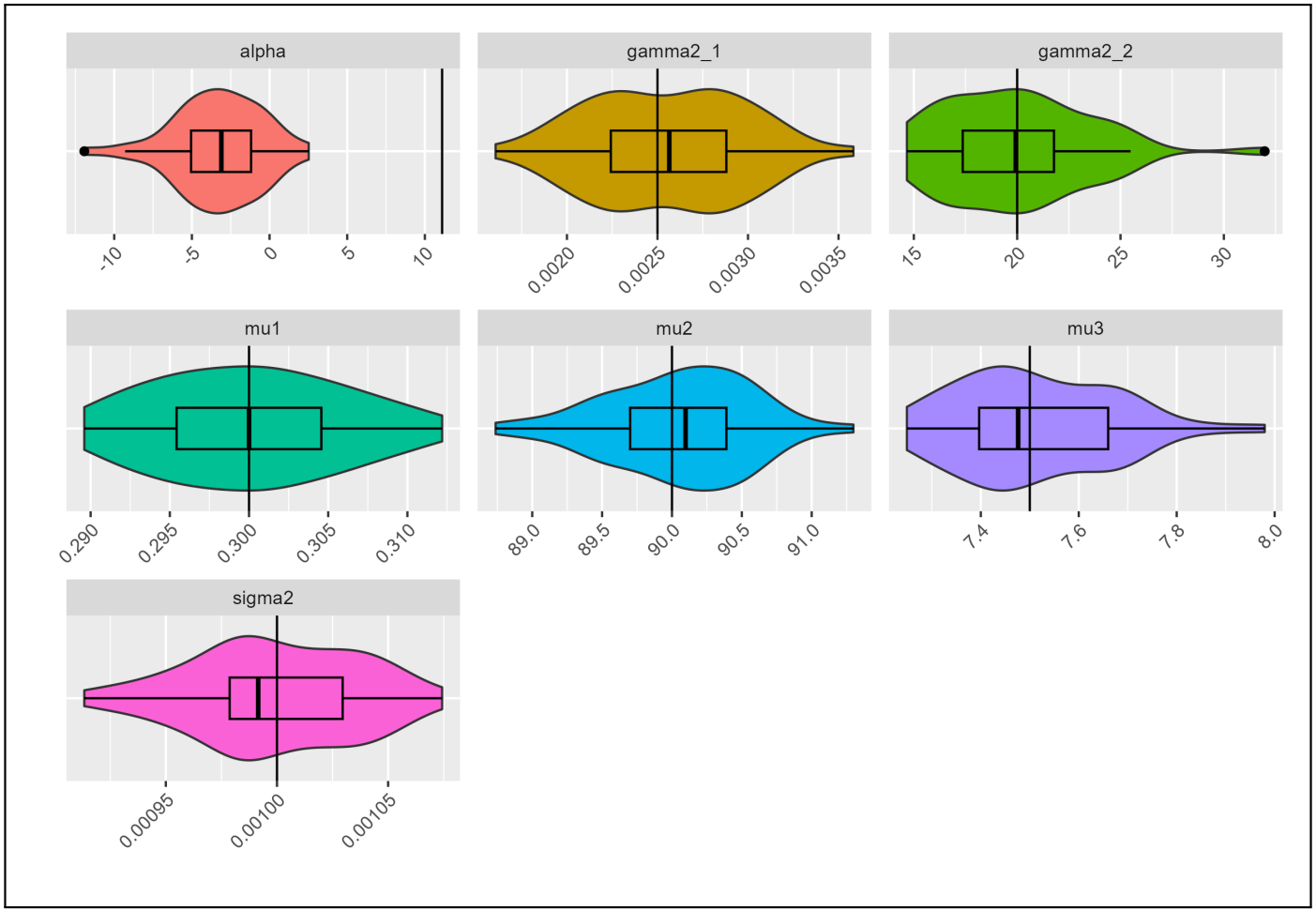


FIGURE 6 – $\hat{\theta}_{\text{Lasso}}$ for 50 SPG ($G = 100$, $J = 20$)

MLE Estimate

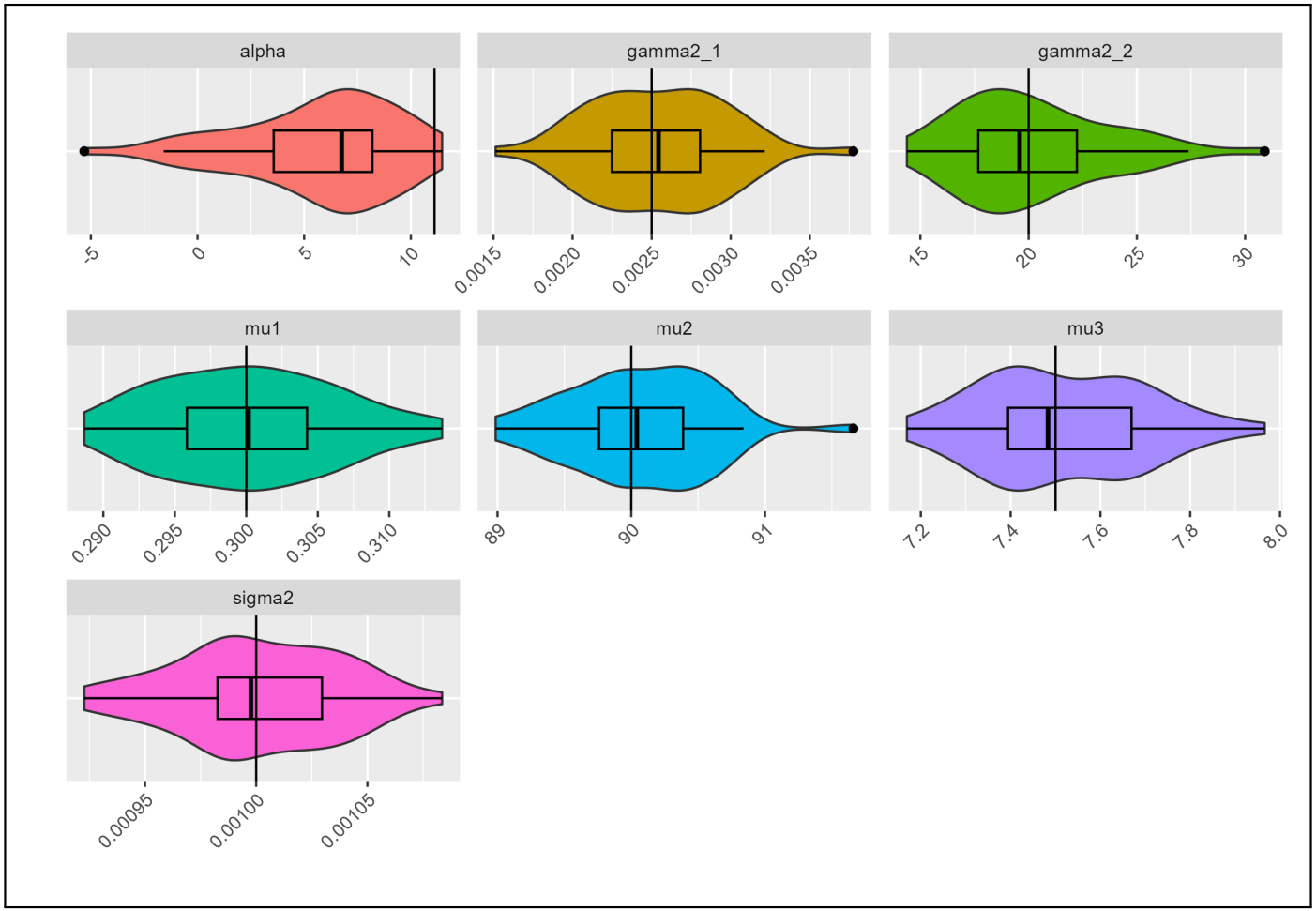


FIGURE 7 – $\hat{\theta}_{\text{MLE}}$ for 50 SG ($G = 100$, $J = 20$)

Conclusion

- Joint modeling of survival and longitudinal data,
- Estimation and variable selection procedure with $N = 100$ and $p = 100$
- Pre-conditionned stochastic gradient implementation with proximal step.

Conclusion

- Joint modeling of survival and longitudinal data,
- Estimation and variable selection procedure with $N = 100$ and $p = 100$
- Pre-conditionned stochastic gradient implementation with proximal step.

Perspectives :

- Apply this method to real data,
- Add high dimension to the mixed effects model,
- Do prediction ou Prediction of the survival time,
- Asymptotic properties of the penalized estimates

Thank you for your attention!

This work has led to a pre-publication on Arxiv, which can be found here :

https://arxiv.org/abs/2306.16765

Bibliographie

Baey, Delattre, Kuhn, Leger et Lemler (2023). « Efficient precondtionned stochastic gradient descent for estimation in latent variables models ». In : *ICML*.

Cox, D. R. (1972). « Regression Models and Life-Tables ». In : 34.2.

Davidian, M. et D.M. Giltinan (1995). *Nonlinear Models for Repeated Measurement Data*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis. ISBN : 978-0-412-98341-2. URL : https://books.google.fr/books?id=0eSIBPAL4qsC.

Moreau, Jean Jacques (1962). « Fonctions convexes duales et points proximaux dans un espace hilbertien ». In : *Comptes rendus hebdomadaires des séances de l'Académie des sciences* 255, p. 2897-2899.

Rizopoulos, Dimitris (2012). *Joint models for longitudinal and time-to-event data : With applications in R*. CRC press.

Rockafellar, R. Tyrrell (août 1976). « Monotone Operators and the Proximal Point Algorithm ». In : *SIAM Journal on Control and Optimization* 14.5, p. 877-898. ISSN : 0363-0129, 1095-7138. DOI : 10.1137/0314056. URL : http://epubs.siam.org/doi/10.1137/0314056 (visité le 28/08/2022).

Baseline hazard definition

A previous internship has highlighted the well-fitting of the data with a Weibull law for the baseline hazard : $h_0 : t \mapsto ba^{-b}t^{b-1}$

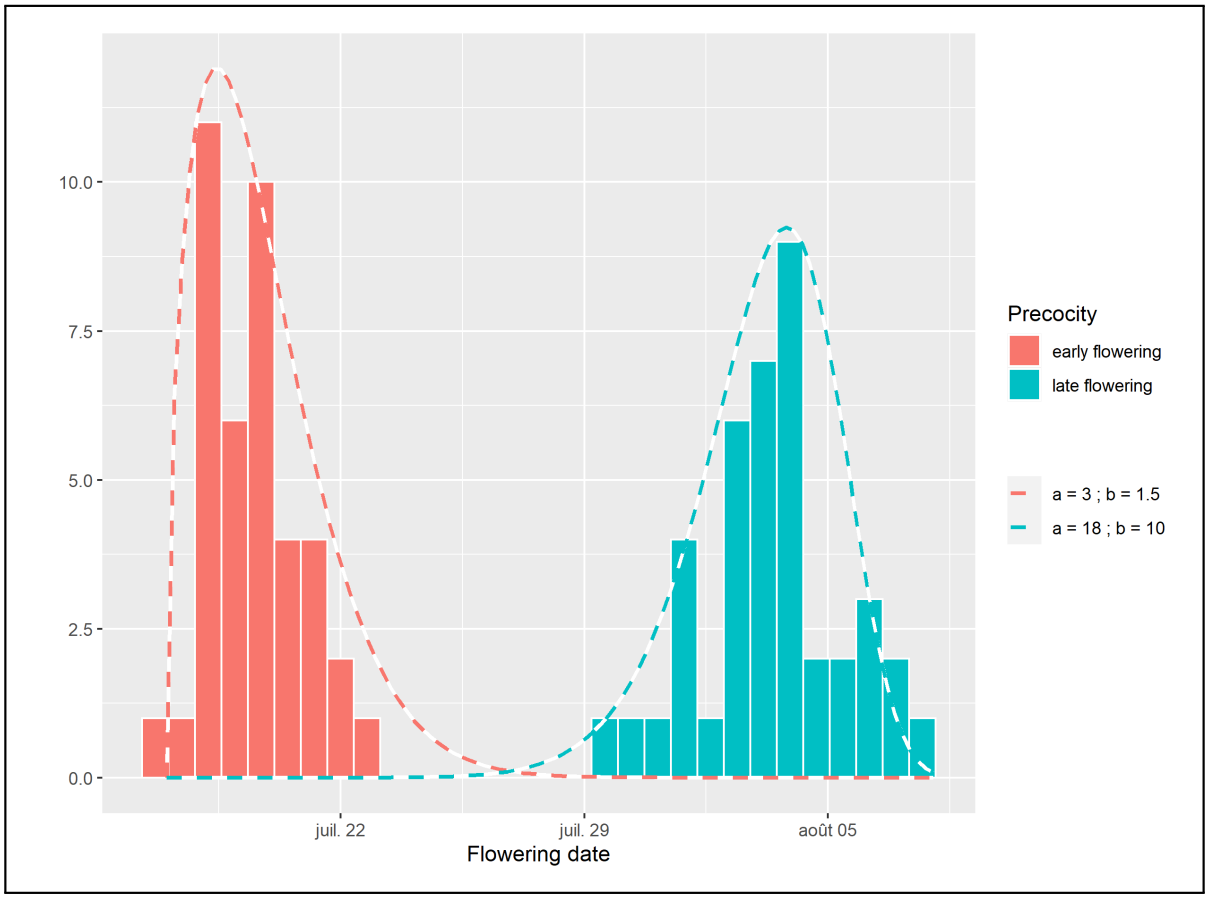


FIGURE 8 – Weibull density example