

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

Antoine Caillebotte¹
Estelle Kuhn¹ Sarah Lemler²

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

³INRAE Génétique Quantitative et Evolution Le Moulon

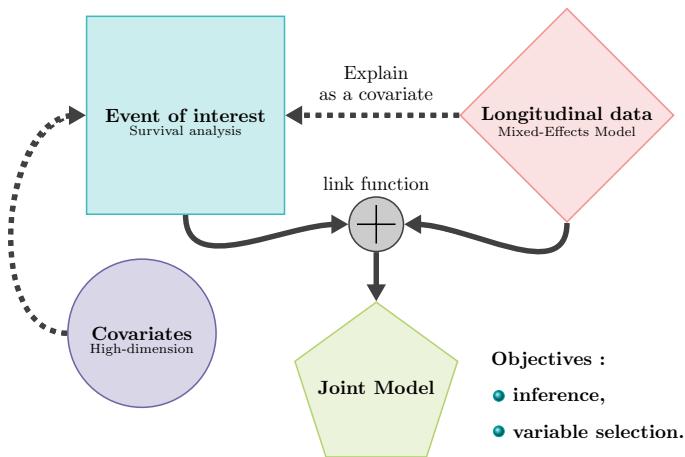
October 2nd, 2023 - JPS - 2023 presentation



- 1 Introduction
 - Survival data
 - Longitudinal data
 - Joint modeling
- 2 Inference in the joint model
 - Stochastic Gradient
 - Proximal Operator
 - Stochastic Proximal Gradient
- 3 Simulation study
 - Methodology
 - Results
- 4 Conclusion and perspectives



Introduction



Survival analysis

Focus on **time to event of interest**

- In medicine : Time to 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 observation of an event of interest**.

Hazard function

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

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



Cox model

Reference : Cox 1972

- For any individual $1 \leq i \leq N$

Regression model that links the survival time to explanatory variables.

The **Hazard function** at T_i is given by :

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

- $T_i \in \mathbb{R}$: 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 : $\theta = (h_0, \beta)$

- **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 $1 \leq i \leq N$ and $1 \leq j \leq J$

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

behavior based on individual variation

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



Non-linear mixed-effects model (NLME)

Reference : DAVIDIAN et GILTINAN 1995

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

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

behavior based on individual variation

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

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

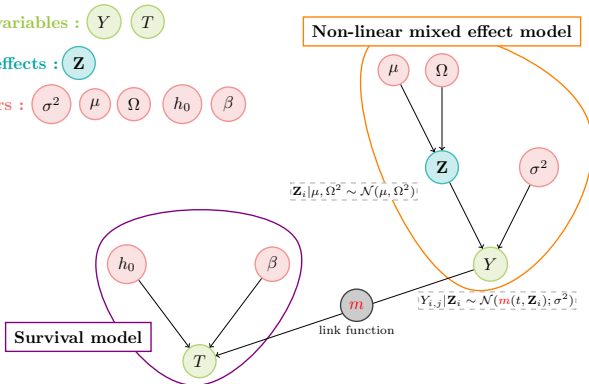
- $\boldsymbol{\mu} = (\mu_1, \dots, \mu_d) \in \mathbb{R}^d$, $\Omega = \text{diag}(\omega_1^2, \dots, \omega_d^2) \in \mathcal{M}_d(\mathbb{R})$: **unknown**,

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



Hierarchical model

- **Random variables** : Y T
- **Random effects** : 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 $1 \leq i \leq N$ and $1 \leq j \leq J$

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

where α quantifies the association between survival and longitudinal data.

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

- 1 Introduction
 - Survival data
 - Longitudinal data
 - Joint modeling
- 2 Inference in the joint model
 - Stochastic Gradient
 - Proximal Operator
 - Stochastic Proximal Gradient
- 3 Simulation study
 - Methodology
 - Results
- 4 Conclusion and perspectives



General estimate in latent variable joint model

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

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

Marginal likelihood written with complete likelihood

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

Recall : \mathbf{Z} is not observed

Maximum likelihood Estimator (MLE)

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



Stochastic Gradient with FIM preconditionning

Reference : BAEY, DELATTRE, KUHN, LEGER et LEMLER 2023



Stochastic Gradient with FIM preconditionning

Reference : BAEY, DELATTRE, KUHN, LEGER et LEMLER 2023

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



Stochastic Gradient with FIM preconditionning

Reference : BAEY, DELATTRE, KUHN, LEGER et LEMLER 2023

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

Stochastic Gradient with FIM preconditionning

Reference : BAEY, DELATTRE, KUHN, LEGER et LEMLER 2023

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

- ② **Gradient computation**, $\mathbf{v}_k = \frac{1}{N} \sum_{i=1}^N \nabla \log \mathcal{L}_{comp}(\theta_k; T_i, Y_i, \mathbf{Z}_i^{(k)})$

FIM estimation

- ③ **Stochastic Approximation**, compute $\Delta_i^{(k)} = (1 - \gamma_k) \Delta_i^{(k-1)} + \gamma_k \nabla \log \mathcal{L}_{comp}(\theta_{k-1}; T_i, Y_i, \mathbf{Z}_i^{(k)})$

- ④ **FIM computation**, $FIM_k = \frac{1}{N} \sum_{i=1}^N \Delta_i^{(k)} \left(\Delta_i^{(k)} \right)^T$

Stochastic Gradient with FIM preconditionning

Reference : BAEY, DELATTRE, KUHN, LEGER et LEMLER 2023

- **Require** : Starting point θ_0, Δ^0
- At the iteration $0 \leq k < K$:
 - ① **Simulation step**, Draw $\mathbf{Z}^{(k)} \sim p(\cdot | T, Y, \theta_k)$ using a Gibbs sampler algorithm
 - ② **Gradient computation**, $\mathbf{v}_k = \frac{1}{N} \sum_{i=1}^N \nabla \log \mathcal{L}_{comp}(\theta_k; T_i, Y_i, \mathbf{Z}_i^{(k)})$
 - ③ **Stochastic Approximation**, compute

$$\Delta_i^{(k)} = (1 - \gamma_k) \Delta_i^{(k-1)} + \gamma_k \nabla \log \mathcal{L}_{comp}(\theta_{k-1}; T_i, Y_i, \mathbf{Z}_i^{(k)})$$
 - ④ **FIM computation**, $FIM_k = \frac{1}{N} \sum_{i=1}^N \Delta_i^{(k)} \left(\Delta_i^{(k)} \right)^T$
 - ⑤ **Forward, gradient** $\theta_{k+1} = \theta_k + \gamma_k FIM_k^{-1} \mathbf{v}_k$
- **Return** : $\hat{\theta} = \theta_K$ with K large enough

FIM estimation

Estimate in joint model with covariates in high dimension

We separate the small-dimensional from the high-dimensional parameters :

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

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

Penalized Estimator : $(\hat{\nu}, \hat{\beta}) = \arg \max_{\beta \in \mathbb{R}^p, \nu \in \mathbb{R}^d} \{ \mathcal{L}_{margin}(\nu, \beta | T, Y) - pen(\beta) \}$

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 :

$$\text{prox}_{\text{pen}}(\beta) = \arg \min_{\beta' \in \mathbb{R}^p} \left(\text{pen}(\beta') + \frac{1}{2} \|\beta - \beta'\|_2^2 \right)$$

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

$$(\text{prox}_{\text{lasso}}(\beta))_i = \begin{cases} 0 & \text{if } |\beta_i| < \lambda \\ \beta_i - \lambda & \text{if } \beta_i \geq \lambda \\ \beta_i + \lambda & \text{if } \beta_i \leq -\lambda \end{cases} \quad (4)$$



Stochastic Proximal Gradient with FIM preconditioning

- **Require** : Starting point $\theta_0, \Delta^{(0)}$
- At the iteration $0 \leq k < K$:
 - ① Simulation step, Draw $\mathbf{Z}^{(k)} \sim p(\cdot | T, Y, \theta_k)$ using a Gibbs sampler algorithm
 - ② Gradient computation, $\mathbf{v}_k = \frac{1}{N} \sum_{i=1}^N \nabla \log \mathcal{L}_{comp}(\theta_k; T_i, Y_i, \mathbf{Z}_i^{(k)})$
 - ③ Stochastic Approximation , Compute
$$\Delta_i^{(k)} = (1 - \gamma_k) \Delta_i^{(k-1)} + \gamma_k \nabla \log \mathcal{L}_{comp}(\theta_{k-1}; T_i, Y_i, \mathbf{Z}_i^{(k)})$$
 - ④ FIM computation, $FIM_k = \frac{1}{N} \sum_{i=1}^N \Delta_i^{(k)} (\Delta_i^{(k)})^T$

Stochastic Proximal Gradient with FIM preconditionning

- **Require** : Starting point $\theta_0, \Delta^{(0)}$
- At the iteration $0 \leq k < K$:
 - ① Simulation step, Draw $\mathbf{Z}^{(k)} \sim p(\cdot | T, Y, \theta_k)$ using a Gibbs sampler algorithm
 - ② Gradient computation, $\mathbf{v}_k = \frac{1}{N} \sum_{i=1}^N \nabla \log \mathcal{L}_{comp}(\theta_k; T_i, Y_i, \mathbf{Z}_i^{(k)})$
 - ③ Stochastic Approximation, Compute $\Delta_i^{(k)} = (1 - \gamma_k) \Delta_i^{(k-1)} + \gamma_k \nabla \log \mathcal{L}_{comp}(\theta_{k-1}; T_i, Y_i, \mathbf{Z}_i^{(k)})$
 - ④ FIM computation, $FIM_k = \frac{1}{N} \sum_{i=1}^N \Delta_i^{(k)} (\Delta_i^{(k)})^T$
 - ⑤ **Forward, gradient** $(\nu_{k+1}, \omega_{k+1})^T = (\nu_k, \beta_k)^T + \gamma_k FIM_k^{-1} \mathbf{v}_k$
 - ⑥ **Backward, penalization**

$$\beta_{k+1} = \text{Prox}_{\gamma_k \text{pen}}(\omega_{k+1})$$

$$\theta_{k+1} = (\nu_{k+1}, \beta_{k+1})$$
- **Return** : $\hat{\theta} = \theta_K$ with K large enough



- 1 Introduction
 - Survival data
 - Longitudinal data
 - Joint modeling
- 2 Inference in the joint model
 - Stochastic Gradient
 - Proximal Operator
 - Stochastic Proximal Gradient
- 3 Simulation study
 - Methodology
 - Results
- 4 Conclusion and perspectives



Methodology

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

Methodology

- Simulate one data set with $N = 100$, $J = 20$ and $p = 100$
- Choose $\lambda = \arg \min_{\lambda \in \Lambda} BIC(\lambda)$
- ① Run SPG resolution

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

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

Methodology

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

1 Run SPG resolution

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

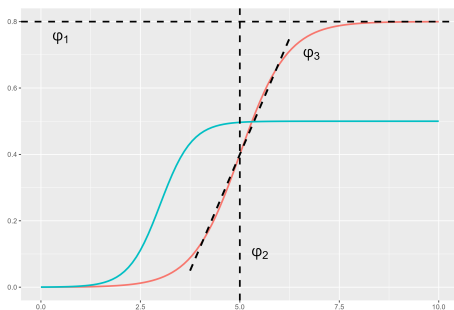
- Reduce the model with the variables selected by the Lasso, $s \ll 100$

2 Run SG resolution

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

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

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

β^*	β_1	β_2	β_3	β_4	β_5	...	β_p
	-2	-1	1	2	0	...	0

Table 1 – Simulation parameters

Results example

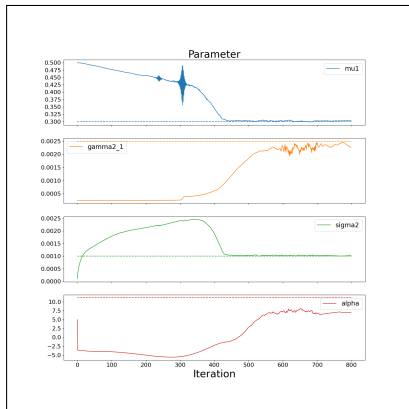


Figure 1 – $\hat{\theta}_{\text{Lasso}}$ during SPG-FIM iterations

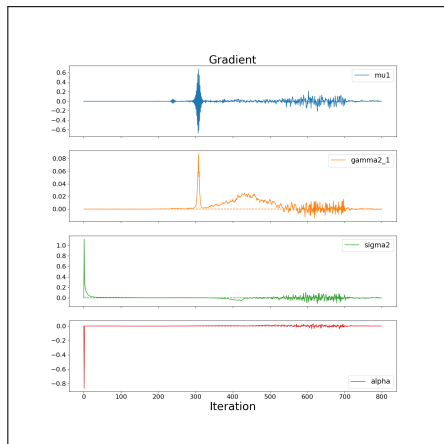


Figure 2 – Gradient value during SPG-FIM iterations

Results example

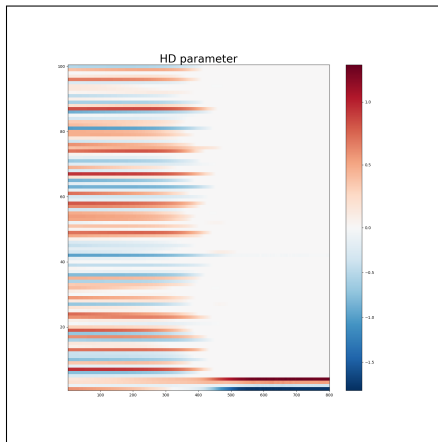


Figure 3 – Tile representation of the selected variables $\hat{\beta}_{Lasso}$ during an SPG-FIM

Variable selection procedure

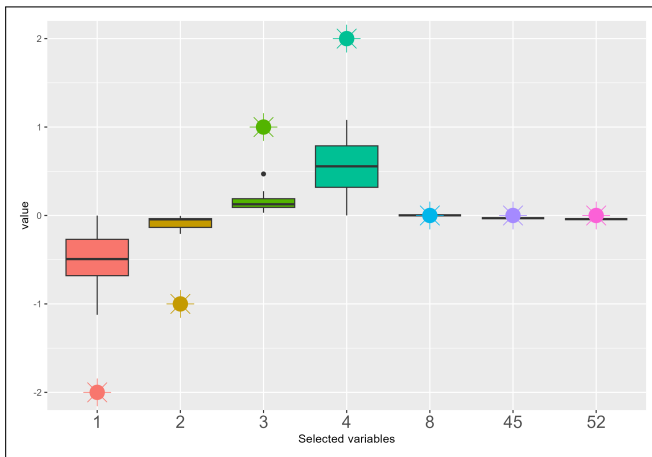


Figure 4 – $\hat{\beta}_{\text{Lasso}} \in \mathbb{R}^7$ for 50 SPG ($G = 100$, $J = 20$)

The colored stars indicate the value used during the data simulation



MLE of β after the variable selection

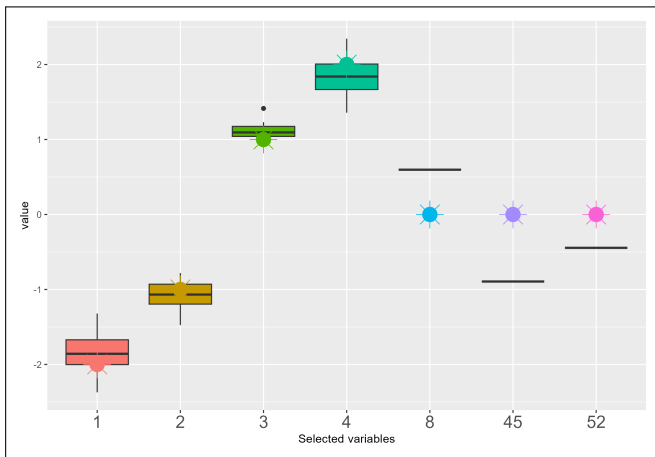


Figure 5 – $\hat{\theta}_{MLE} \in \mathbb{R}^7$ for 50 SPG ($G = 100$, $J = 20$)

The colored stars indicate the value used during the data simulation



Estimate of the parameter with Lasso penalization

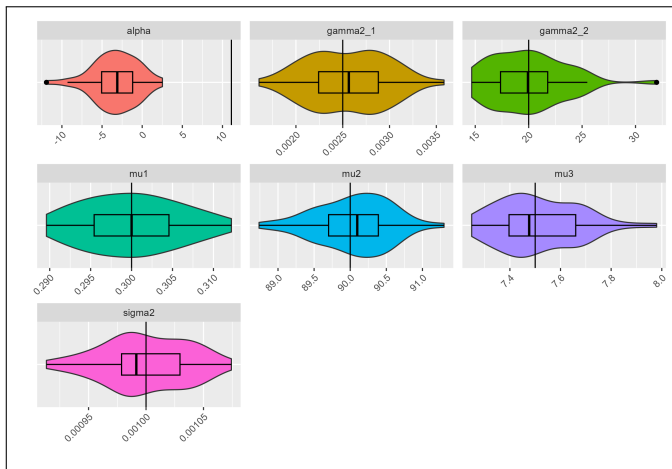


Figure 6 – $\hat{\theta}_{\text{Lasso}}$ for 50 SPG ($G = 100$, $J = 20$)

MLE Estimate

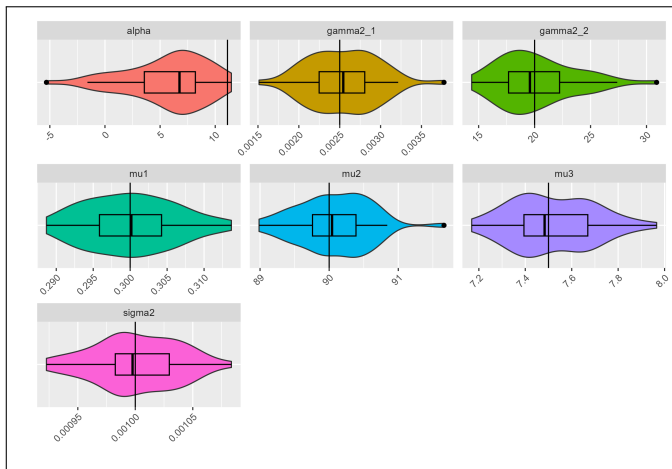


Figure 7 – $\hat{\theta}_{MLE}$ for 50 SG ($G = 100$, $J = 20$)

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,
- Asymptotic properties of the penalized estimates,
- Do prediction of the survival time.



Thank you for your attention!

This work has led to a pre-publication on Arxiv, by :








Caillebotte, Kuhn, Lemler,

which can be found here :

<https://arxiv.org/abs/2306.16765>



Bibliographie

-  **BAEY, DELATTRE, KUHN, LEGER et LEMLER (2023).** « Efficient preconditioned 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 : $h_0 : t \mapsto ba^{-b}t^{b-1}$

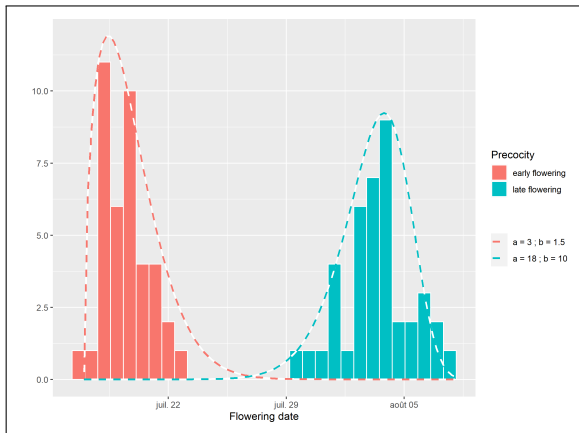


Figure 8 – Weibull density example

Regularization path

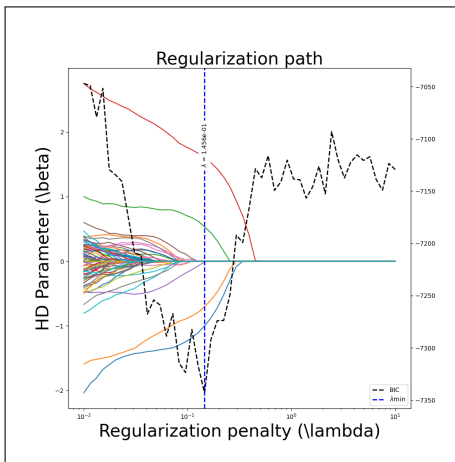


Figure 9 – Regularization path, BIC criterion displayed in dotted lines, while β component values are shown in solid colored lines. The vertical blue line indicates the value of λ for which the BIC is minimal.