Estimation and variable selection in a joint model of survival times and longitudinal outcomes with random effects.

Antoine Caillebotte, Estelle Kuhn, Sarah Lemler

Estimation and variable selection in a joint model of survival times and longitudinal outcomes with random effects.

Antoine Caillebotte *email:caillebotte.antoine@inrae.fr*
Université Paris-Saclay, INRAE, UR MaIAGE, UMR GQE-Moulon, France

Estelle Kuhn *email:estelle.kuhn@inrae.fr*
Université Paris-Saclay, INRAE, UR MaIAGE, France,

Sarah Lemler *email:sarah.lemler@centralesupelec.fr*
Université Paris-Saclay, Laboratoire MICS, France,

May 23, 2024

Abstract: We consider a joint survival and mixed-effects model to explain the survival time from longitudinal data and high-dimensional covariates. The longitudinal data is modeled using a nonlinear mixed-effects model, where the regression function serves as a link function incorporated into a Cox model. In that way, the longitudinal data is related to the survival time. Additionally, the Cox model takes into account the inclusion of high-dimensional covariates. The main objectives of this research are two-fold: first, to identify the relevant covariates that contribute to explaining survival time, and second, to estimate all unknown parameters of the joint model. For that purpose, we consider the estimate defined through maximizing a Lasso penalized likelihood. To tackle the optimization problem, we implement a preconditioned stochastic gradient to handle the latent variables of the nonlinear mixed-effects model associated with a proximal operator to manage the non-differentiability of the penalty. We provide relevant simulations that showcase the performance of the proposed variable selection and the parameter estimation method in the joint modeling of the Cox and logistic model.

Key words: Joint model, nonlinear mixed-effects model, Cox model, high dimension covariates, penalized estimation, preconditioned stochastic gradient, proximal operator

1 Introduction

A very current issue in many fields is the need for a better understanding of the interactions between dependent dynamic phenomena. For example, in medicine, this may involve the dynamics of a patient's tumors in oncology and the effects of anti-cancer treatments administered to the patient. Another example in plant science is the dynamics of plant development in a plot and the spread of an epidemic disease or pests in that plot. The considered phenomena are often complex, both in terms of their modes of interaction and their temporal and spatial dynamics. Moreover, these phenomena are frequently observed in populations of heterogeneous or structured individuals, such as patients or plants.

Mathematical modeling has proven to be a powerful tool for understanding the interactions between multiple dynamic phenomena. It also allows for considering variability present in the observed population of individuals. Joint modeling of several phenomena has demonstrated its effectiveness in several fields, including medicine, pharmacology, and biology ([13]). A particular case of joint models concerns the simultaneous modeling of longitudinal data and survival data observed on the same individual. In this type of joint model, longitudinal data are often modeled by a mixed-effects model ([17, 4]), and survival data by a survival model such as the Cox model ([3]). The latter allows for modeling the instantaneous risk of the survival variable as a function of the covariates. It is also possible to include longitudinal data modeling as a covariate in the Cox model via a linking function. Several authors have studied this model ([26], [20], [14]). Due to the presence of latent variables in the mixed-effects model, inference by maximum likelihood can be made via Expectation Maximization (EM) like algorithms ([26], [11], [19], [7]). The EM-type algorithms, such as the classical Stochastic Approximation Expectation Maximization (SAEM), are the most classical approaches for inferring parameters in the presence of latent variables. They have been developed for estimation in general latent variable models. They are straightforward to implement in the context of a curved exponential family based on sufficient statistics of the model. Moreover, theoretical convergence results have been established in this context. However, when the model does not belong to the exponential family, which is the case in our context, the methodology is not generic in practice, and the theoretical results fail.

Some exponentialization tricks have then been proposed to face this restrictive assumption of the curved exponential family. For instance, one of them consists of considering some unknown parameters as random population variables. However, [5] have shown that, in general, the parameter returned by the SAEM on the modified model is not a maximum likelihood of the initial model, and they have suggested the use of this exponentialization trick with variances of the new random population variables that decrease as the iterations of the algorithm progress. This approach also has limitations in practice due to complex algorithmic settings and tuning. The gradient-based methods are another type of approach, often omitted for estimating parameters in latent models. Recently, [2] suggested using a preconditioned stochastic gradient algorithm to deal with parameter estimation in the presence of latent variables. This approach is particularly interesting when considering a model that does not belong to the exponential family, as is the case for our joint model. [2] showed that this algorithm performs well for the nonlinear logistic growth mixed-effects model, which can be used to represent some longitudinal data. Note that Bayesian numerical methods have also been proposed in parallel to estimate the parameters of a joint model. ([18], [21], [13]).

Besides, in many applications, today's technological means allow for collecting high-dimensional explanatory covariates. These may include, for example, genetic markers or omics data. In addition to the wealth of information provided by these covariates, they also generate difficulties in the statistical analysis of models as it is necessary to adapt statistical and numerical approaches to their high dimensionality. One possible approach is to consider a penalized estimator, such as the Lasso ([12], [27]), and adapted numerical methods, such as stochastic proximal gradient ([1], [9]).

In this paper, we consider a joint model that combines, through a link function, a nonlinear mixed-effects model for longitudinal data and a Cox model for the survival times, including covariates of high dimension. Our work aims to select the relevant variables among the high-dimensional covariates in the Cox model part of the joint model based on the whole dataset and then to estimate the model's unknown parameters. For that purpose, we propose an estimate for model parameters, which include a Lasso penalization for the regression parameter of the Cox model. To calculate this estimate in practice, we develop an algorithm combining a preconditioned stochastic gradient to deal with the latent variables in the joint model out of the exponential family and a proximal gradient to handle the non-differentiability of the Lasso penalty used for variable selection in the Cox model part. The proposed algorithm is easy to implement in general joint

models without assuming that model density belongs to the curved exponential family.

The paper is organized as follows. In Section 2, we detail the joint model constructed from a nonlinear mixed-effects model for longitudinal data and a Cox model for survival data, with high-dimensional covariates and a link function. In Section 3, we present the proposed inference method based on a Lasso penalized estimator and numerical procedure based on a stochastic proximal gradient algorithm. Finally, we illustrate the methodology in Section 4 through a simulation study. The paper ends with a conclusion.

2 Joint model for survival times and longitudinal outcomes

We consider N individuals, and for each individual i, we denote the survival time $\mathbf{T}_i^*$, corresponding to the duration until the occurrence of an event of interest, and longitudinal data, which are repeated observations J times denoted by $\mathbf{Y}_{i,j}$ with $i \in \{1, \dots, N\}$ and $j \in \{1, \dots, J\}$. Note that our work can easily be generalized to the case where there are different numbers of longitudinal observations for each individual in the population. The following describes the joint model we considered.

2.1 Survival Model

The survival time $\mathbf{T}_i^*$ of individual i is the time between a fixed initial moment and the occurrence of an event of interest. To characterize the distribution of $\mathbf{T}_i^*$, we use the hazard function defined by:

$$h_i(t) := \lim_{\Delta t \to 0} \frac{\mathbb{P}(t \leq \mathbf{T}_i^* < t + \Delta t | \mathbf{T}_i^* \geq t)}{\Delta t}; \forall t \geq 0. \tag{1}$$

The Cox model ([3]) is one of the most classical models in survival analysis. It allows us to relate the hazard function of the survival time $\mathbf{T}_i^*$ to the covariates $U_i \in \mathbb{R}^p$, with p being the number of covariates. In our approach, we will consider the high-dimensional setting with many covariates so that p is very large with respect to N. The hazard of the Cox model for individual i is written as follows:

$$h(t|U_i) = h_0(t) \exp(\beta^T U_i); \forall t \geq 0, \tag{2}$$

with $\beta \in \mathbb{R}^p$ a regression parameter and h_0 the baseline hazard function that characterizes a common behavior in the observed population. In the sequel, we will consider a parametric baseline function denoted by h_{θ_b} where θ_b are its parameters. Therefore, the Cox model's unknown parameters are β and θ_b. It is a positive random variable. In some settings the survival time is not observed directly due to censorship. Thus let's suppose that $\mathbf{T}_i^*$ is censored on the right, and let's denote by C_i the censoring time. We then observe $\mathbf{T}_i = \min(\mathbf{T}_i^*, C_i)$ and $\delta_i = \mathbb{1}_{\mathbf{T}_i^* \leq C_i}$ respectively the censored survival time and the censoring indicator.

In addition to the covariates, we consider explaining some of the survival time variability using the longitudinal data dynamic, which will be modeled using a nonlinear mixed-effects model. Let us present the mixed-effects model before explaining the integration of this new component into the Cox model.

2.2 Nonlinear Mixed-Effects Model

The longitudinal data are observed J times for each individual $i \in \{1, \dots, N\}$. Let us denoted by $\mathbf{Y}_{i,j}$ the $j-th$ observation of the $i-th$ individual for $j \in \{1, \dots, J\}$ and $i \in \{1, \dots, N\}$. We model this longitudinal observation using a nonlinear function m that depends on individual parameters represented by the latent variable $\mathbf{Z}_i$ as follows:

$$\begin{cases} \mathbf{Y}_{i,j} & = m(t_j; \mathbf{Z}_i) + \varepsilon_{i,j}, \forall 1 \leq i \leq N, 1 \leq j \leq J \\ \mathbf{Z}_i & \underset{i.i.d.}{\sim} \mathcal{N}(\mu, \Gamma) \ ; \ \varepsilon_{i,j} \underset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2), \end{cases} \tag{3}$$

where, t_j is the $j-th$ observation time, and $\varepsilon_{i,j}$ is an additive noise assumed centered Gaussian with unknown variance σ^2. The latent variable $\mathbf{Z}_i$ describes the inter-individual variability of the population. It is assumed that $\mathbf{Z}_i$ follows a Gaussian distribution with unknown expectation μ and variance Γ. The unknown parameters of the nonlinear mixed-effects model are therefore μ, Γ, and σ^2.

Let us introduce in the following the link function, which will combine the two previous models by modeling the influence of the dynamic of the longitudinal observation of the hazard function.

2.3 Joint model

We introduce the link function denoted by f, which is linearly parameterized with α and depends on the history of the true unobserved longitudinal data $\mathcal{M}(t;\mathbf{Z}_i) = \{m(s;\mathbf{Z}_i)|\forall s, 0 \leq s < t\}$. $\mathcal{M}(t;\mathbf{Z}_i)$ describes the past values of the longitudinal dynamic up to time t. Several forms for f can be taken, an example of which is given below. We assume that the hazard of the survival time of individual i is related to the longitudinal data dynamic through the link function f as follows, $\forall 1 \leq i \leq N, \forall t \geq 0$:

$$h(t|\mathcal{M}(t,\mathbf{Z}_i), U_i) = h_{\theta_b}(t)\exp\big(\beta^T U_i + f(\alpha, \mathcal{M}(t;\mathbf{Z}_i))\big), \tag{4}$$

The parameter α represents the influence of the longitudinal dynamic on the survival data.

The unknown parameters for the joint model include the parameters of the Cox model and those of the nonlinear mixed-effects model, as well as the link function parameter of the joint model. We note $\theta = (\theta_b, \beta, \alpha, \mu, \Gamma, \sigma^2) \in \Theta$ the vector of unknown parameters with $\Theta \subset \mathbb{R}^d$ being the parameter space.

3 Inference Method

In this section, we propose a method for estimating the model parameters presented above. We will define two estimators to deal first with the latent variable context and then with the high dimension.

3.1 Expression of the Marginal Likelihood

We consider the maximum likelihood estimator to infer the joint model parameters. In the context of latent variable models, the marginal likelihood, denoted by $\mathcal{L}_{marg}$, is obtained by integrating the complete likelihood over the latent variables, which are not observed. Let $\boldsymbol{\mathcal{D}}_i = (\mathbf{Y}_i, \mathbf{T}_i, \delta_i)$ be the observed variables for $1 \leq i \leq N$ and $\boldsymbol{\mathcal{D}} = (\boldsymbol{\mathcal{D}}_i)_{1\leq i\leq N}$:

$$\begin{aligned}\mathcal{L}_{marg}(\theta;\boldsymbol{\mathcal{D}}) &= \prod_{i=1}^{n}\int p_\theta(\boldsymbol{\mathcal{D}}_i, \mathbf{Z}_i)d\mathbf{Z}_i \\ &= \prod_{i=1}^{n}\int p_\theta(\boldsymbol{\mathcal{D}}_i|\mathbf{Z}_i)p_\theta(\mathbf{Z}_i)d\mathbf{Z}_i,\end{aligned} \tag{5}$$

where $p_\theta(\boldsymbol{\mathcal{D}}, \mathbf{Z}), p_\theta(\boldsymbol{\mathcal{D}}|\mathbf{Z}), p_\theta(\mathbf{Z})$ are respectively the density of the pair $(\boldsymbol{\mathcal{D}}, \mathbf{Z})$, the density of $\boldsymbol{\mathcal{D}}$ conditionally to $\mathbf{Z}$, and the density of $\mathbf{Z}$.

$$\begin{aligned}\mathcal{L}_{marg}(\theta;\boldsymbol{\mathcal{D}}) =& \prod_{i=1}^{n}\int\Bigg(\left\{\prod_{j=1}^{J} p_\theta(Y_{i,j}|\mathbf{Z}_i)\right\} \\ &\times p_\theta(\mathbf{T}_i, \delta_i|\mathcal{M}(t,\mathbf{Z}_i), U_i)p_\theta(\mathbf{Z}_i)\Bigg)d\mathbf{Z}_i \\ =& \prod_{i=1}^{n}\int\Bigg(\left\{\prod_{j=1}^{J}\Phi(Y_{i,j}; m(t_j,\mathbf{Z}_i), \sigma^2)\right\} \\ &\times \left(h(T_i|\mathcal{M}(T_j,\mathbf{Z}_i), U_i)\right)^{\delta_i} \\ &\times \exp\left(-\int_0^{T_i} h(s|\mathcal{M}(s,\mathbf{Z}_i), U_i)ds\right) \\ &\times \Phi(\mathbf{Z}_i;\mu,\Gamma)\Bigg)d\mathbf{Z}_i,\end{aligned}$$

where $\Phi(.; m, v)$ is the density of normal distribution centered in m and of variance v.

Due to the integral, it is difficult to directly compute the maximum of the marginal likelihood, which does not have an analytical form in this latent variable model. Therefore, we will use numerical methods to solve this maximization problem.

3.2 Definition of the Penalized Estimator for Variable Selection

One can usually estimate the model parameters by maximizing the likelihood using the maximum likelihood estimator written as follows :

$$\hat{\theta}_{\mathbf{MLE}}(\boldsymbol{\mathcal{D}}) = \underset{\theta \in \Theta}{\arg\max} \left(\log \mathcal{L}_{marg}(\theta; \boldsymbol{\mathcal{D}}) \right), \tag{6}$$

where Θ denotes the parameter space. However, in our context we must deal with the high dimension of the covariates, so we introduce a penalty and consider a penalized maximum likelihood estimator. We aim to select relevant variables among the covariates of the survival model. We use the Lasso (Least Absolute Shrinkage and Selection Operator) procedure, which was initially developed for linear regression models ([24]) and the Cox model ([25]). This method enables us to handle high-dimensional data and select a subset of explanatory covariates from a large collection. We consider a Lasso penalty, which only depends on the parameter β:

$$\mathrm{pen}_{\lambda}(\theta) = \lambda \left\| \beta \right\|_1 = \lambda \sum_{k=1}^{p} |\beta_k| ,$$

where λ is a positive real called regularization parameter. Our goal is then to maximize the logarithm of the marginal likelihood where the penalty is integrated as follows. Let us define the penalized maximum likelihood estimator by:

$$\hat{\theta}^{\lambda}_{\mathbf{Lasso}}(\boldsymbol{\mathcal{D}}) = \underset{\theta \in \Theta}{\arg\max} \left(\log \mathcal{L}_{marg}(\theta; \boldsymbol{\mathcal{D}}) - \mathrm{pen}_{\lambda}(\theta) \right), \tag{7}$$

where Θ denotes the parameter space and where λ is a positive parameter. The larger the value of λ, the more β will be constrained to have zero components. Conversely, the smaller the value of λ, the freer the components of β will be. It is customary to determine the value of λ using cross-validation ([24]). But here, without any predictive context, we study the BIC criterion to find the best regularization value ([23]).

Usually, when we deal with latent variables, since the marginal likelihood is non-analytic, classical methods used to infer the unknown parameters are Expectation Maximization like algorithms ([16]). The inconvenience of these procedures is that they are well adapted to models of the curved exponential family, which is not the case for the joint model we consider. The nonlinearity of the function m on the individual parameters φ prevents it from being in this exponential family. Recently [2] have presented a preconditioned stochastic gradient descent for estimation in a latent variable model adapted to general latent variables models. Moreover, due to the non-differentiability of the considered penalty, we will use a proximal algorithm as presented by [1] and [9]. Thus, we add a proximal gradient in the procedure presented in [2] and implement a preconditioned stochastic proximal gradient algorithm to calculate the estimator.

3.3 Implementation of the Inference Procedure

We deal simultaneously with unobserved random effects of the mixed-effects model and the penalty term by implementing a Preconditioned Stochastic Proximal Gradient, SPG-FIM, in the sequel. The latter is a forward-backward splitting algorithm that can compute the penalized estimate (7). The algorithm is divided into three steps; a realization of the latent variables is sampled with a first step called *Simulation*, which uses a Metropolis-Hastings sampler ([10]). The second step is the classical gradient descent on the approximate complete likelihood, the *Forward* step. Following the procedure presented in [2], we have chosen to use a preconditioning of the gradient with an estimate of the Fisher information matrix. The latter is updated during the iterations using the estimate presented by [6]. The last step, called *Backward*, deals with the penalty term. We apply the classical proximal operator ([15, 22], defined below :

$$\text{Prox}_{\text{pen}_\lambda}(\beta) = \underset{\beta' \in \mathbb{R}^p}{\arg\min} \left(\text{pen}_\lambda(\beta') + \frac{1}{2} \|\beta - \beta'\|_2^2 \right). \tag{8}$$

With the Lasso penalty, the proximal operator has an explicit form:

$$(\text{Prox}_{\text{pen}_\lambda}(\beta))_i = \begin{cases} 0 & \text{if } |\beta_i| < \lambda \\ \beta_i - \lambda & \text{if } \beta_i \geq \lambda \\ \beta_i + \lambda & \text{if } \beta_i \leq -\lambda \end{cases} ; \forall i \in \{1, ..., p\}. \tag{9}$$

The *Backward* step corresponds to the application of the proximal operator on the result of the *Forward* step.

As the penalty only depends on β, the proximal operator selects the β components that seem to be the most explanatory of the data. It computes a sparse solution for β but also applies shrinkage on the non-zero components so that the Lasso estimator is biased. Therefore, we detail a method to obtain an unbiased estimator in what follows.

Algorithm 1 provides the steps of the stochastic proximal gradient, where $(\gamma_k)_{k\geq 1}$ is a step size such that $\forall k \in \mathbb{N}, \gamma_k \in [0, 1], \sum_{k=1}^{\infty} \gamma_k = \infty$ and $\sum_{k=1}^{\infty} \gamma_k^2 < \infty$.

Algorithm 1: Stochastic proximal gradient with FIM preconditioning (SPG-FIM)

Require: Number of iterations $K \geq 1$; sequence of step-size $(\gamma_k)_{k\geq 1}$
Initialize Starting point $\theta_0 \in \mathbb{R}^d$, Δ_0
for $k = 1$ **to** K **do**
  • **Simulation step :**
    Draw $\mathbf{Z}^{(k)}$ **using a single step of a Hastings Metropolis procedure**
  • **Gradient computation :**
    Compute $v_k = \frac{1}{N} \sum_{i=1}^{N} \nabla \log f_{\theta_k}(\boldsymbol{\mathcal{D}}_i, \mathbf{Z}_i^{(k)})$
  • **FIM computation :**
    • **Compute the stochastic approximation**
    $\forall i \in \{1, ..., N\}$,
      $\Delta_i^{(k)} = (1 - \gamma_k)\Delta_i^{(k-1)} + \gamma_k \nabla \log f_{\theta_k}(\boldsymbol{\mathcal{D}}_i, \mathbf{Z}_i^{(k)})$
    • **Compute the FIM :**
      $FIM_k = \frac{1}{N} \sum_{i=1}^{N} \Delta_i^{(k)} \left(\Delta_i^{(k)}\right)^T$
  • **Gradient descent :**
    • **Forward step :** $\omega_{k+1} = \theta_k - \gamma_k FIM_k^{-1} v_k$
    • **Backward step :** $\theta_{k+1} = \text{Prox}_{\gamma_k \text{pen}_\lambda}(\omega_{k+1})$
end
return $\hat{\theta} = \theta_K$

Note that we can differentiate between the two step-size sequences involved in either stochastic approximation or gradient descent. However, for the sake of clarity, we have noted them in the same way here.

3.4 The Estimation Procedure

As explained, the proximal operator (9) has a shrinking effect on the estimator after its application, meaning that the values found for β are smaller than expected. Therefore, the estimator of β is biased; we thus divide the inference into two steps. An exploratory one that allows us to select the support of the vector β through a Lasso penalization estimation and a second step of inference without penalization, where we have restricted the number of covariates with respect to the selected support. We also need to select a well balance value for the regularization parameter, for that we use the BIC criterion. We thus apply these two steps for a grid of values for the regularization parameter λ and select the model and estimation that minimizes the BIC criterion. We conduct the following inference methodology :

- Run the SPG-FIM procedure (algorithm 1) in order to compute

$$\hat{\theta}^{\lambda}_{\mathbf{Lasso}}(\boldsymbol{\mathcal{D}}) = \arg\max_{\theta\in\Theta} \left\{\log \mathcal{L}_{marg}(\theta;\boldsymbol{\mathcal{D}}) - \text{pen}_{\lambda}(\theta)\right\},$$

  for different values of λ on a fixed grid.

- Choose the reduced support of β according to the estimate $\hat{\theta}^{\lambda}_{\mathbf{Lasso}}(\boldsymbol{\mathcal{D}})$. We got a reduced model. Compute $\hat{\theta}^{\lambda}_{\mathbf{MLE}}(\boldsymbol{\mathcal{D}})$ the maximum likelihood estimate in this reduced model without the penalty term and therefore without bias with the SG-FIM algorithm (SPG-FIM without the *Backward* step).

- We obtain a collection of estimates for each possible value of λ. For each regularization parameter, we compute the BIC criterion (see [23]):

$$BIC(\lambda) = -2\log(\mathcal{L}_{marg}(\hat{\theta}^{\lambda}_{\mathbf{MLE}}(\boldsymbol{\mathcal{D}});\boldsymbol{\mathcal{D}})) + k\ \log(N\times(J+1)),$$

  where k is the number of non-zeros components in β.
  Note that the quantity $\mathcal{L}_{marg}(\theta;\boldsymbol{\mathcal{D}}) = \int_{\mathbf{Z}} f_{\theta}(\boldsymbol{\mathcal{D}},\mathbf{Z})d\mathbf{Z}$ is computed by approximating the integral using a Monte Carlo procedure.

- Finally, we keep the estimate $\hat{\theta}^{\hat{\lambda}}_{\mathbf{MLE}}(\boldsymbol{\mathcal{D}})$ and the regularization parameter $\hat{\lambda}$ that minimizes the BIC, as :

$$\hat{\lambda} = \arg\min_{\lambda} BIC(\lambda).$$

So, with this method, we can finally estimate the relevant parameters from the joint modeling of survival and longitudinal data.

4 Simulation study

In this section, we propose to study the performance of the procedure we have just presented. First, we want to numerically show the consistency of the estimator (7) by studying four scenarios, varying the number N of individuals observed. Then, we will demonstrate the method's ability to select variables by varying the number p of covariates in the initial model. This simulation study is conducted using joint modeling for longitudinal logistic data. In each scenario, we generate an independent data set, and each time, we fit the model (4) using the routine described in subsection 3.4.

To compare the results of the different scenarios, we look at two metrics: the root mean square error (RMSE) to measure the estimation quality of the method and the sensitivity, specificity, and accuracy to study the selection capacity of the method. Sensitivity is the proportion of true positives correctly identified. Specificity is the proportion of the true negatives correctly identified. Finally, accuracy is the proportion of true results, either true positive or true negative.

4.1 Logistic Joint-Model Specification

We consider the joint model defined in section 2 with the classical logistic function for the nonlinear mixed-effects model, defined by :

$$m : t \mapsto \frac{\mathbf{Z}_1}{1+\exp\left(\dfrac{\mathbf{Z}_2 - t}{\tau}\right)}, \tag{10}$$

where $\mathbf{Z}_1$ represents the asymptotical maximum value of the curve, $\mathbf{Z}_2$ represents the value of the sigmoid's midpoint, and τ represents the logistic growth rate. We model for each individual i the corresponding individual parameter $\mathbf{Z}_i \in \mathbb{R}^2$ through a Gaussian random variable with expectation $\mu \in \mathbb{R}^2$ and diagonal variance-covariance matrix $\Gamma = diag(\gamma_1^2, \gamma_2^2)$. Regarding the survival model, we consider a fixed Weibull

baseline defined as $h_{a,b}(t) = ba^{-b}t^{b-1}$, where $a = 800$ and $b = 10$ are fixed (i.e. not estimated) in the simulation study. Moreover we assume that the time of interest $\mathbf{T}^*$ can be related to the longitudinal dynamic using the following linking function: $f(\alpha, \mathcal{M}(t; \mathbf{Z}_i)) = \alpha m(t, \mathbf{Z}_i)$.

We focus in this simulation study on the variable selection procedure and on the inference of the parameters of the mixed-effects model as well as α, the multiplicative parameter of the Cox model.

4.1.1 Simulation setup

We generated 100 data sets according to the joint model presented above. For each different value of p, we choose the vector β such that the first four components are equal to $(-3, -2, 2, 3)$ and the rest are equal to zero. Additionally, we generate the matrix of covariates U with N rows and p columns, following a uniform distribution $U_{i,l} \sim \mathcal{U}([-1, 1])$; $\forall i \in 1, ..., N, l \in 1, ..., p$. Afterward, we renormalize all the matrix columns to obtain centered data. All the parameter values are detailed in the table 4.1.1 below.

Table 1: Parameter values used in simulation

Parameter	μ_1	μ_2	τ	γ_1^2	γ_2^2	σ^2	α
True value	200	500	150	40	100	10	0.05
Parameter	β_1	β_2	β_3	β_4	β_5	...	β_p
True value	-3	-2	2	3	0	...	0

The initial parameters θ_0 were randomly drawn for each independent run. To make the algorithm numerically more robust, we add a regularization to the Fisher information matrix when estimating it. The Fisher information is thus calculated as follows: $FIM_k = (1-\lambda_k)Id + \lambda_k \frac{1}{N} \sum_{i=1}^{N} \Delta_i^{(k)} \left(\Delta_i^{(k)}\right)^T$. As explained before, we can set up several step-size sequences for stochastic gradient approximation, gradient descent, and fisher regularization. We define them as follows λ_{sto}, λ_{grad} and λ_{fim}. The gradient descent step size is chosen to be first increasing, then stable for a given time interval, and then decreasing. This provides a warm-up period to stabilize the Gibbs sampler. During this warm-up phase, step sizes for the stochastic approximation and the regularization of the Fisher information were also chosen to increase and then become stable until the end of the procedure. We choose the following sequences of steps according to $k_{preheating}$, $k_{heating}$ the start and end iterations of the preheating and heating phases respectively :

$$\gamma_{sto,k} = \begin{cases} \gamma_0^{\left(1-\frac{k}{k_{preheating}}\right)} & \text{if } k \leq k_{preheating} \\ 1 & \text{else} \end{cases}$$

$$\gamma_{grad,k} = \begin{cases} \gamma_0^{\left(1-\frac{k}{k_{preheating}}\right)} & \text{if } k \leq k_{preheating} \\ 1 & \text{else if } k \leq k_{heating} \\ (k - k_{heating})^{-0.65} & \text{else} \end{cases} \tag{11}$$

$$\gamma_{fim,k} = \begin{cases} \gamma_0^{\left(1-\frac{k}{k_{preheating}}\right)} & \text{if } k \leq k_{preheating}. \\ 0.9 & \text{else} \end{cases}$$

In this simulation study, we use the following configuration: $k_{preheating} = 1000$, $k_{heating} = 1400$ and $\gamma_0 = 10^{-9}$.

Table 2 summarizes the root mean square errors (RMSE) obtained when the number of covariates p is equal to 500, the number of observations J is fixed at 15, and the number of individuals increases $N \in \{200, 400\}$. We aim to show the numerical consistency of the proposed estimator (7). A very small error is obtained for a large proportion of the parameters. The results improve as the number of individuals increases. The variance of random effects is difficult to estimate, particularly that associated with the position of the midpoint in the logistic function γ_2^2. It can be seen that the parameters related to the explanatory covariates (i.e. $\beta_1, \beta_2, \beta_3, \beta_4$) are more sensitive than the other parameters to the number of individuals. Excluding the variances, the algorithm allows us to obtain estimates below an error of 5% for $N = 400$.

Table 2: Root mean squared error (RMSE) and Relative-RMSE (RMSE divided by the simulated true value), for a total of 100 repetitions. The simulated values are $\mu_1 = 200$, $\mu_2 = 500$, $\tau = 150$, $\gamma_1^2 = 40$, $\gamma_2^2 = 100$, $\sigma^2 = 10$, $\alpha = 0.05$, $\beta = [-3, -2, 2, 3]$

	P = 500			
	N = 200		N = 400	
	RMSE	RRMSE(%)	RMSE	RRMSE(%)
μ_1	0.467	0.233	0.303	0.151
μ_2	0.793	0.159	0.623	0.125
τ	0.371	0.248	0.276	0.184
γ_1^2	4.596	11.489	3.105	7.763
γ_2^2	15.501	15.501	9.542	9.542
σ^2	0.259	2.587	0.209	2.094
α	0.001	1.390	0.001	1.060
β_1	0.109	3.621	0.088	2.928
β_2	0.119	5.949	0.084	4.195
β_3	0.143	7.170	0.097	4.862
β_4	0.131	4.372	0.094	3.144

In addition to the estimation, we also want to demonstrate our method's ability to perform variable selection. To do this, we set the number of individual parameters to $N = 200$ and $J = 15$ and study the cases where $p \in \{200, 300, 500\}$. Table 3 shows the results for these scenarios.

Table 3: Average of sensitivity (Se), specificity (Sp), accuracy (Ac), and error using L_1-norm and L_2-norm over 100 repetitions, with their associated variance in brackets.

	N = 200				
P	Se	Sp	Ac	$\|.\|_1$	$\|.\|_2$
200	1.0	0.999	0.999	0.478	0.268
	(0.0)	(1.62e-05)	(1.57e-05)	(0.123)	(0.0255)
300	1.0	0.999	0.999	0.481	0.273
	(0.0)	(8.46e-06)	(8.29e-06)	(0.119	(0.0251)
500	1.0	0.999	0.999	0.548	0.307
	(0.0)	(2.23e-06)	(2.2e-06)	(0.131	(0.0305)

We can see that sensitivity, i.e., the proportion of true positives, is 100% with zero variance, i.e., over the 100 repetitions each time the four explanatory covariates were selected. The proportion of false positives, or specificity, is almost equally 100%. For the three values of p tested in less than 15% of cases, one to four additional variables are selected at worst. The error selection error for the L_1 or L_2-norms is constant for the scenarios we tested.

As an illustration, Figure 1 displays the estimated parameter as a function of iterations during the execution of the SPG-FIM algorithm. In addition to the algorithm's convergence to the values used in the simulation, we can also observe the preheating and heating phases before stabilizing the algorithm.

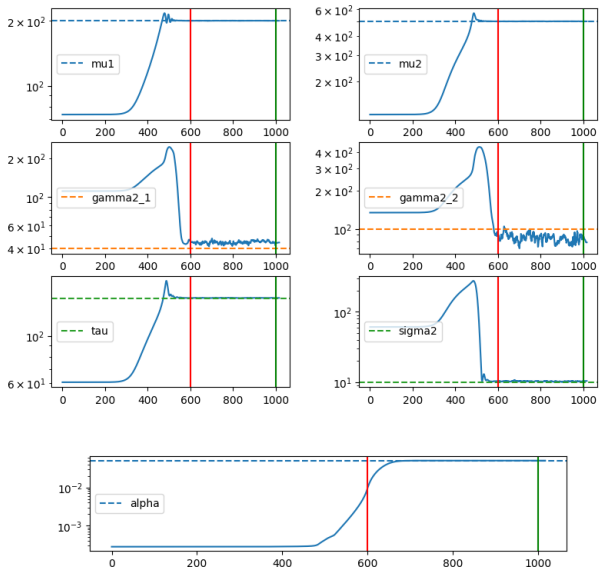

Figure 1: Parameter estimates across the SPG-FIM iterations. Dotted line: simulated value, red vertical line: end of the pre-heating, green vertical line: end of the heating

5 Conclusion and perspectives

In this work, we jointly addressed variable selection and parameter estimation in a joint model for survival data and longitudinal outcomes. We connected a survival model with a nonlinear mixed-effects model through a linking function and covariates eventually in high dimension. To estimate the unknown parameters of this joint model, we conduct a preconditioned proximal stochastic gradient to deal with the latent variables and the Lasso penalty. Our methodology has been thoroughly evaluated on simulated data to demonstrate its performance.

One interesting perspective of this work consists of addressing the prediction task carefully. It would be interesting to set up a method for predicting survival time from some start of longitudinal data observation. Previous studies, such as [8] and [13]), have explored this task within the context of joint models.

Furthermore, considering the partial likelihood for the survival part rather than the complete likelihood would be an intriguing avenue to explore. This approach would eliminate the need to make assumptions about the form of the baseline risk, which is especially relevant when dealing with the Cox model. Indeed, when considering the Cox model, the partial likelihood allows for estimating the regression parameter without knowing the baseline function. It would be interesting to see how our algorithm performs with this partial likelihood instead of the complete one.

Funding and Acknowledgements

This work was funded by the https://stat4plant.mathnum.inrae.fr/(Stat4Plant) project ANR-20-CE45-0012.

References

[1] Massil Achab. *Learning from Sequences with Point Processes*. Phd thesis, Université Paris-Saclay, 2017. Issue: 2017SACLX068.

[2] Baey, Delattre, Kuhn, Leger, and Lemler. Efficient precondtionned stochastic gradient descent fro estimation in latent variables models. *ICML*, 2023.

[3] D. R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):187–220, 1972. Publisher: [Royal Statistical Society, Wiley].

[4] M. Davidian and D.M. Giltinan. *Nonlinear Models for Repeated Measurement Data*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis, 1995.

[5] Vianney Debavelaere and Stéphanie Allassonnière. On the curved exponential family in the stochastic approximation expectation maximization algorithm. *ESAIM: Probability & Statistics*, 25, 2021.

[6] Maud Delattre and Estelle Kuhn. Computing an empirical Fisher information matrix estimate in latent variable models through stochastic approximation. *Computo*, 2023.

[7] Solène Desmée, France Mentré, Christine Veyrat-Follet, Bernard Sébastien, and Jeremie Guedj. Using the saem algorithm for mechanistic joint models characterizing the relationship between nonlinear psa kinetics and survival in prostate cancer patients. *Biometrics*, 73(1):305–312, 2017.

[8] Solène Desmée, France Mentré, Christine Veyrat-Follet, Bernard Sébastien, and Jérémie Guedj. Nonlinear joint models for individual dynamic prediction of risk of death using hamiltonian monte carlo: application to metastatic prostate cancer. *BMC Medical Research Methodology*, 17(1):105, 2017.

[9] Gersende Fort, Edouard Ollier, and Adeline Samson. Stochastic proximal gradient algorithms for penalized mixed models, 2017.

[10] Stuart Geman and Donald Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-6(6):721–741, 1984.

[11] Jeremie Guedj, Rodolphe Thiébaut, and Daniel Commenges. Joint modeling of the clinical progression and of the biomarkers' dynamics using a mechanistic model. *Biometrics*, 67(1):59–66, 2011.

[12] Zangdong He, Wanzhu Tu, Sijian Wang, Haoda Fu, and Zhangsheng Yu. Simultaneous variable selection for joint models of longitudinal and survival outcomes: Variable selection in joint models. *Biometrics*, 71(1):178–187, 2015.

[13] Marion Kerioui, Julie Bertrand, René Bruno, François Mercier, Jérémie Guedj, and Solène Desmée. Modelling the association between biomarkers and clinical outcome: An introduction to nonlinear joint models. *British Journal of Clinical Pharmacology*, 88(4):1452–1463, 2022.

[14] Agnieszka Król, Loïc Ferrer, Jean-Pierre Pignon, Cécile Proust-Lima, Michel Ducreux, Olivier Bouché, Stefan Michiels, and Virginie Rondeau. Joint model for left-censored longitudinal data, recurrent events and terminal event: Predictive abilities of tumor burden for cancer evolution with application to the ffcd 2000–05 trial. *Biometrics*, 72(3):907–916, 2016.

[15] Jean Jacques Moreau. Fonctions convexes duales et points proximaux dans un espace hilbertien. *Comptes rendus hebdomadaires des séances de l'Académie des sciences*, 255:2897–2899, 1962.

[16] Shu Kay Ng, Thriyambakam Krishnan, and Geoffrey J McLachlan. The em algorithm. *Handbook of computational statistics: concepts and methods*, pages 139–172, 2012.

[17] José C. Pinheiro and Douglas M. Bates. *Mixed-Effects Models in S and S-PLUS*. Statistics and Computing. Springer-Verlag, 2000.

[18] Elizabeth R. Brown and Joseph G. Ibrahim. A bayesian semiparametric joint hierarchical model for longitudinal and survival data. *Biometrics*, 59(2):221–228, 2003.

[19] Dimitris Rizopoulos. Jm: An r package for the joint modelling of longitudinal and time-to-event data. *Journal of statistical software*, 35:1–33, 2010.

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

[21] Dimitris Rizopoulos. The r package jmbayes for fitting joint models for longitudinal and time-to-event data using mcmc. *Journal of Statistical Software*, 72(7):1–46, 2016.

[22] R. Tyrrell Rockafellar. Monotone operators and the proximal point algorithm. *SIAM Journal on Control and Optimization*, 14(5):877–898, 1976.

[23] Gideon Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978. Publisher: Institute of Mathematical Statistics.

[24] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

[25] Robert Tibshirani. The lasso method for variable selection in the cox model. *Statistics in Medicine*, 16(4):385–395, 1997.

[26] Michael S. Wulfsohn and Anastasios A. Tsiatis. A joint model for survival and longitudinal data measured with error. *Biometrics*, 53(1):330–339, 1997.

[27] Fengting Yi, Niansheng Tang, and Jianguo Sun. Simultaneous variable selection and estimation for joint models of longitudinal and failure time data with interval censoring. *Biometrics*, 78(1):151–164, 2022.