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[1,2], Estelle Kuhn [2], Sarah Lemler [3]

[1] *Université Paris-Saclay, INRAE, UMR GQE-Moulon, France, caillebotte.antoine@inrae.fr,*
[2] *Université Paris-Saclay, INRAE, UMR MaIAGE, France, estelle.kuhn@inrae.fr,*
[3] *Université Paris-Saclay, Laboratoire MICS, France, sarah.lemler@centralesupelec.fr*

Abstract. This paper considers 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 effects model, where the regression function serves as a link function incorporated into a Cox model as a covariate. In that way, the longitudinal data is related to the survival time at a given 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 maximization of a Lasso penalized likelihood. To tackle the optimization problem, we implement a pre-conditioned 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 parameters' estimation method in the joint modeling of a Cox and logistic model.

Keywords. Joint model, non-linear mixed effects model, Cox model, high dimension, preconditioned stochastic gradient, proximal operator

1 Introduction

A very current issue in many fields is better understanding 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 phenomena considered are often complex, both in terms of their modes of interaction and their temporal and spatial dynamics. Moreover, these phenomena are often 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 variabilities 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 covariates. It is also possible to include longitudinal data modeling as a covariate in the Cox model via a linking function. The objective is then to estimate the model parameters from the observations and to select relevant covariates. Several authors have proposed such approaches ([26], [19], [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], [20], [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 particularly easy 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 trick has then been proposed to face this restrictive assumption of the curved exponential family. It consists in 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 the 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 ([18], [21], [13]).

Besides, in many applications, current 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 which combines, through a link function, a nonlinear mixed effect 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 survival and mixed-effects model

We consider N individuals and study, for each individual i, the survival time $\mathbf{T}_i$, corresponding to the duration until the occurrence of an event of interest, and longitudinal data, more precisely 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 number of longitudinal observations for each individual of the population. The following describes the joint model we considered.

2.1 Survival model

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

$$h_i(t) := \lim_{dt \to 0} \frac{\mathbb{P}(t \leqslant \mathbf{T}_i < t + dt | \mathbf{T}_i \geqslant t)}{dt}; \forall t \geqslant 0. \tag{1}$$

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 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 a large number of covariates, such as p is very large with respect to N. The Cox model for individual i is written as follows:

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

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 $h_{\theta_{base}}$ where $\theta_{base} \in \mathbb{R}^b$ are its parameters. Therefore, the Cox model's unknown parameters are β and θ_{base}.

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:

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

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 survival and mixed-effects model

We assume that the hazard of the survival time is related to the longitudinal data dynamic through the link function m as follows:

$$h(t|\mathcal{M}(t, \mathbf{Z}_i), U_i) = h_{\theta_{base}}(t) \exp(\beta^T U_i + \alpha m(t, \mathbf{Z}_i)), \ \forall t \geqslant 0, \tag{4}$$

where $\mathcal{M}(t; \mathbf{Z}i) = \{m(s; \mathbf{Z}i) | \forall s, 0 \leqslant s < t\}$ describes the past values of the longitudinal dynamic up to time t. The parameter α represents the influence of the longitudinal dynamic on the survival data. The joint model can be written as follows:

$$\begin{cases} h(t|\mathcal{M}(t, \mathbf{Z}_i), U_i) & = h_{\theta_{base}}(t) \exp(\beta^T U_i + \alpha m(t, \mathbf{Z}_i)) \\ Y_{i,j} & = m(t_j; \mathbf{Z}_i) + \varepsilon_{i,j} \\ \mathbf{Z}_i \underset{i.i.d.}{\sim} \mathcal{N}(\mu, \Gamma) & ; \ \varepsilon_{i,j} \underset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2). \end{cases} \quad \forall 1 \leqslant i \leqslant N, 1 \leqslant j \leqslant J \tag{5}$$

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 $\theta = (\theta_{base}, \beta, \mu, \Gamma, \sigma^2, \alpha) \in \Theta$ the vector of unknown parameters with $\Theta \subset \mathbb{R}^d$ being the parameter space. In the following section, we propose an estimation method for these parameters.

3 Inference method

Note that there is often censoring in survival analysis, which leads to partially observed data: survival times are not directly observed. Available information is censored times and indicators, making the estimation task more complex. For the sake of simplicity, since we focus on the high-dimensional covariates selection task, we will not consider censoring in our approach for the moment. However, it will be part of further work.

3.1 Definition 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 $\mathcal{D} = (\mathbf{Y}, \mathbf{T})$ be the observed variables:

$$\mathcal{L}_{marg}(\theta; \mathcal{D}) = \int f_\theta(\mathcal{D}, \mathbf{Z}) d\mathbf{Z} = \int g_\theta(\mathcal{D}|\mathbf{Z}) p_\theta(\mathbf{Z}) d\mathbf{Z} \tag{6}$$

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

We introduce a penalty and consider a penalized maximum likelihood estimator to deal with the high dimension of the covariates. 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 proposed for linear regression models ([25]) and the Cox model ([24]). 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:

$$\text{pen}(\theta) = \|\beta\|_1 = \sum_{k=1}^{p} |\beta_k| ,$$

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:

$$\hat{\theta} = \underset{\theta \in \Theta}{\arg\max} \left(\log L_{marg}(\theta; \mathcal{D}) - \lambda \, \text{pen}(\theta) \right) , \tag{7}$$

where Θ denotes the parameter space and where λ is a positive parameter called the regularization 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 more free the components of β will be. It is customary to determine the value of λ using cross-validation ([25]).

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 algorithme ([16]). The inconvenience of these procedures is that it is well-adapted to models belonging to the curved exponential family, which is not the case for the joint model we consider. Recently [2] have proposed 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 proposed by [1] and [9]. Thus, we add a proximal gradient in the

procedure proposed in [2] and implement a preconditioned stochastic proximal gradient algorithm to calculate the estimator.

3.3 Implementation of the inference procedure

We deal simultaneously both with unobserved random effects of the mixed-effects model and the penalty term by implementing a preconditioned stochastic proximal gradient, called SPG-FIM in the sequel. A forward–backward splitting algorithm such that a Stochastic Proximal Gradient can compute the estimate 7. The latter has a stochastic approach to replace missing data with simulations. 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 proposed 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 proposed by [6]. The last step, called *Backward*, deals with the penalty term. We apply the classical proximal operator ([15, 22], defined below

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

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

$$(\mathrm{Prox}_{Lasso}(\beta))_i = \begin{cases} 0 & \text{si } |\beta_i| < \lambda \\ \beta_i - \lambda & \text{si } \beta_i \geqslant \lambda \\ \beta_i + \lambda & \text{si } \beta_i \leqslant -\lambda \end{cases} ; \ \forall i \in \{1, ..., p\}. \tag{9}$$

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 $(\gamma_k)_{k \geqslant 1}$ 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 preconditionning (SPG-FIM)

Require : Number of iterations $K \geqslant 1$; sequence of step-size $(\gamma_k)_{k\geqslant 1}$

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 $v_k = \dfrac{1}{N}\sum_{i=1}^{N} \nabla \log f_{\theta_k}(\mathcal{D}_i \mathbf{Z}_i^{(k)})$
- **FIM computation :**
  - Compute the stochastic approximation
  $\forall i \in \{1, ..., N\}, \Delta_i^{(k)} = (1-\gamma_k)\Delta_i^{(k-1)} + \gamma_k \nabla \log f_{\theta_k}(\mathcal{D}_i \mathbf{Z}_i^{(k)})$
  - 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 :
  $$\theta_{k+1} = \mathrm{Prox}_{\gamma_k \mathrm{pen}}(\omega_{k+1}) = \underset{\theta' \in \Theta}{\arg\min} \left\{ \gamma_k \mathrm{pen}(\theta') + \frac{1}{2} \left\| \omega_{k+1} - \theta' \right\|_2^2 \right\}$$

end

return $\hat{\theta} = \theta_K$

4 Simulation study

We generated data according to the joint model presented previously in Equation 5. We consider $N = 100$ individuals, each individual being observed $J = 20$ times. We use the classical logistic function in the nonlinear mixed effect model detailed in Equation 3:

$$m : t \mapsto \frac{Z_1}{1 + \exp\left(\dfrac{Z_2 - t}{Z_3}\right)}, \tag{10}$$

where Z_1 represents the asymptotical maximum value of the curve, Z_2 represents the value of the sigmoid's midpoint, and Z_3 represents the logistic growth rate. We model for each individual i the corresponding individual parameter $\mathbf{Z}_i \in \mathbb{R}^3$ through a Gaussian random variable with expectation $\mu \in \mathbb{R}^3$ and a diagonal variance $\Gamma = diag(\gamma_1^2, \gamma_2^2, 0)$ meaning that the third parameter μ_3 is modeled as a fixed effect. We consider a fixed Weibull baseline defined as $h_{a,b}(t) = ba^{-b}t^{b-1}$, where $a = 80$ and $b = 35$ are fixed (i.e. not estimated) in the simulation study.

We fix $p = 100$ and choose the vector β such that the first four components are equal to

$(-2, -1, 1, 2)$ 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 $U_{i,l} \sim \mathcal{U}([-1, 1])$; $\forall i \in 1, ..., N, l \in 1, ..., p$. We choose finally the link function parameter $\alpha = 11.11$. All the parameter values are detailed in Table 4.

Parameter	μ_1	μ_2	μ_3	γ_1^2	γ_2^2	σ^2	α
True value	0.3	90	7.5	$2.5\ 10^{-3}$	20	10^{-3}	11.11

Parameter	β_1	β_2	β_3	β_4	β_5	...	β_p
True value	-2	-1	1	2	0	...	0

Table 1: True parameter values used for the simulation

We focus in this simulation study on the selection of variables and on the inference of the parameters of the mixed-effect model as well as α the multiplicative parameter of the Cox model. 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 as it is usually the case 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 conduct the following inference methodology:

- Run the SPG-FIM algorithm in order to compute

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

  for different values of λ on a fixed grid.

- Select the best regularization parameter such that $\lambda_m = \arg\min_{\lambda} BIC(\lambda)$ according to the BIC criterion (see [23]):

$$BIC(\lambda) = -2 \log(\mathcal{L}_{marg}(\hat{\theta}_{\mathbf{Lasso}}(\lambda); \mathcal{D})) + k \, \log(N)$$

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

- Choose the reduced support of β according to the estimate $\hat{\theta}_{\mathbf{Lasso}}(\lambda_m)$ obtained previously from a run of the SPG-FIM (Algorithm 1).

- Compute $\hat{\theta}_{\mathbf{MLE}}$ the maximum likelihood estimate in the reduced model without the penalty term and therefore without bias with the SG-FIM (SPG-FIM without the *Backward* step).