Lights: a generalized joint model for high-dimensional multivariate longitudinal data and censored durations

Simon Bussy*[1], Antoine Barbieri[2], Sarah Zohar[1], and Anne-Sophie Jannot[1,3]

[1] *INSERM, UMRS 1138, Centre de Recherche des Cordeliers, Paris, France*
[2] *INSERM, UMR 1219, Bordeaux Population Health Research Center, Univ. Bordeaux, France*
[3] *Biomedical Informatics and Public Health Department, EGPH, APHP, Paris, France*

Abstract

This paper introduces a prognostic method called *lights* to deal with the problem of joint modeling of longitudinal data and censored durations, where a large number of longitudinal features are available. Yet there is no standard model so far to learn from such high-dimensional multivariate longitudinal data in a survival analysis setting. Features are extracted from the longitudinal processes and included as potential risk factor in a group-specific Cox model with high-dimensional shared associations. Appropriate penalties are then used during inference to allow flexibility in modeling the dependency between the longitudinal features and the event time. The statistical performance of the method is examined on an extensive Monte Carlo simulation study, and finally illustrated on publicly available datasets. On these high-dimensional datasets, our proposed method significantly outperforms the state-of-the-art survival models regarding risk prediction in terms of C-index, with a computing time orders of magnitude faster. In addition, it provides powerful interpretability by automatically pinpointing significant features being relevant from a clinical perspective. Thus, we propose a powerfull tool for personalized medicine, with the ability of automatically determining significant prognostic longitudinal biomarkers, which is of increasing importance in many areas of medicine.

Keywords. High-dimensional estimation; Joint modeling; Multivariate longitudinal data; Survival analysis

1 Introduction

In many clinical studies, it has become increasingly common to record the values of longitudinal features (e.g., biomarkers) until the occurrence of an event of interest for a subject. The "joint modeling" approaches, namely modeling the longitudinal and survival outcomes through a joint likelihood model rather than separately, has received considerable attention during the past two decades [Tsiatis and Davidian, 2004]. Numerical studies suggest that these approaches are among the most satisfactory to combine information [Yu et al., 2004]. They have the advantage of making more efficient use of the data since information about survival also goes into modeling the longitudinal features. In addition, they produce unbiased estimates and do not rely on approximations for incorporating conplex longitudinal trajectories. Most developments have either focused on shared random-effect models (SREMs) [Wulfsohn and Tsiatis, 1997], in which characteristics of the longitudinal processes (for instance functions of the random effects) are included as features in the

*Corresponding author: simon.bussy@gmail.com

survival model ; or on joint latent class models (JLCMs) [Vermunt and Magidson, 2003], in which the population is considered as heterogeneous, with the assumption that there exist homogeneous latent subgroups sharing the same marker trajectories and the same prognostic.

The high-dimensional longitudinal data context. With the development of electronic health records, high-dimensional settings are becoming increasingly frequent in various contexts where the number of available features to consider as potential risk factors is tremendous. Moreover, with an increased focus on personalised medicine, the need to implement multivariate models that account for a large number of longitudinal outcomes is critical. Despite this, joint models have predominantly focused on univariate data, with attempts to fit multiple univariate joint models separately [Wang et al., 2012], which is inefficient [Lin et al., 2002a]. Despite many multivariate models being presented in full generality, questions arrising from the high-dimensional context – e.g., computational power, limits in numerical estimation, or sample size – are never considered in analyses (to the best of our knowledge), and the number of longitudinal outcomes considered in numerical studies are often very low (see Hickey et al. [2016] for a complete review). For instance, Jaffa et al. [2014] only considers 3 longitudinal outcomes in the simulation study while mentioning a "high-dimensional multivariate longitudinal data" context.

General framework. The setting of this paper is such that we want to incorporate high-dimensional time-dependent (longitudinal) features measured with error in a survival model. Let us consider the usual survival analysis framework. Following Andersen et al. [2012], let non-negative random variables $T^\star$ and C stand for the times of the event of interest and censoring times respectively. The event of interest could be for instance survival time, re-hospitalization, relapse or disease progression. We then denote T the right-censored time and Δ the censoring indicator, defined as

$$T = T^\star \wedge C \quad \text{and} \quad \Delta = \mathbb{1}_{\{T^\star \leq C\}}$$

respectively, where $a \wedge b$ denotes the minimum between two numbers a and b, and $\mathbb{1}_{\{\cdot\}}$ the indicator function taking the value 1 if the condition in $\{\cdot\}$ is satisfied and 0 otherwise.

Let X denotes the p-dimensional vector of time-independent features (e.g., patients characteristics, therapeutic strategy, or omics features recorded at the begining of a study), and let $Y(t) = \left(Y^1(t), \ldots, Y^L(t)\right)^\top \in \mathbb{R}^L$ denote the value of the L-dimensional longitudinal outcome at time point $t \geq 0$, with $L \in \mathbb{N}_+$.

Heterogeneity of the population. An assumption of heterogeneity within the patient population is frequently relevant in medical research where several differing profiles of subjects are expected [Bussy et al., 2019]. To take account of this, we introduce a latent variable $G \in \{0, \ldots, K-1\}$ modeling the $K \geq 1$ subgroups of different risk, which is a classical modeling assumption in JLCMs [Lin et al., 2002b, Proust-Lima et al., 2014]. Let us denote

$$\pi_{\xi_k}(x) = \mathbb{P}[G = k | X = x] \tag{1}$$

the latent class membership probability given time-independent features $x \in \mathbb{R}^p$, and consider a softmax link function given by

$$\pi_{\xi_k}(x) = \frac{e^{x^\top \xi_k}}{\sum_{k=0}^{K-1} e^{x^\top \xi_k}}$$

where $\xi_k \in \mathbb{R}^p$ denotes a vector of coefficients that quantifies the impact of each time-independent features on the probability that a subject belongs to the k-th group, with $\xi_0 = \mathbf{0}_p$ for overparameterization purpose, where $\mathbf{0}_p$ stands for the vector of $\mathbb{R}^p$ having all coordinates equal to zero. The intercept term is here omitted without loss of generality. From now on, all computations are done conditionally on features x.

Main contribution. In this paper, we propose a method called *lights* (generaLized joInt hiGH-dimensional longiTudinal Survival) which is from both JLCMs and SREMs, since we also include features extracted from the longitudinal processes as potential risk factor in the survival model, which is a group-specific Cox model [Cox, 1972] with high-dimensional shared associations. To allow flexibility in modeling the dependency between the longitudinal features and the event time, we use appropriate penalties : elastic net [Zou and Hastie, 2005] for feature selection in the latent class membership, and sparse group lasso [Simon et al., 2013] in the survival model, as well as for the fixed effect (allowing flexible representations of time). Indeed, this penalty acts like the lasso at the trajectory level, namely an entire trajectory may drop out of the model on one side. On the other hand, it yields sparsity for a given trajectory, namely feature selection. Inference is achieved using an efficient and novel Quasi-Newton Monte Carlo Expectation Maximization algorithm. Hence, the method provides interpretations of the high-dimensional longitudinal features, thus offering a powerful tool for clinical decision making in patient monitoring.

Organization of the paper. A precise description of the model is given in Section 2. Section 3 focuses on a regularized version of the model to exploit dimension reduction and prevent overfitting. Inference is presented under this framework, as well as the developed algorithm. Section 4 introduces the C-index metric, as well as a novel evaluation strategy to assess diagnostic prediction performances while mimicking a real-time use of the model in clinical care, and finally the considered competing methods. Section 5 presents the simulation procedure used to evaluate the performance of our method in a high-dimensional context and compares it with state-of-the-art ones. In Section 6, we apply our method to high-dimensional publicly available datasets. Finally, we discuss the obtained results in Section 7.

Notations. Throughout the paper, for every $q > 0$, we denote by $\|v\|_q$ the usual ℓ_q-quasi norm of a vector $v \in \mathbb{R}^m$, namely $\|v\|_q = (\sum_{k=1}^m |v_k|^q)^{1/q}$. We also denote $\|v\|_0 = |\{k : v_k \neq 0\}|$, where $|A|$ stands for the cardinality of a finite set A. For $u, v \in \mathbb{R}^m$, we denote by $u \odot v$ the Hadamard product $u \odot v = (u_1 v_1, \ldots, u_m v_m)^\top$. For a squared matrix M, vech(M) stacks columns of M one under another in a single vector, starting each column at its diagonal element. We write I_m for the identity matrix of $\mathbb{R}^{m \times m}$. Finally, we write, for short, $\mathbf{1}_m$ (resp. $\mathbf{0}_m$) for the vector of $\mathbb{R}^m$ having all coordinates equal to one (resp. zero).

2 Method

In this section, we descibe the longitudinal and time-to-event submodels, as well as the required hypothesis in order to write a likelihood and draw inference for the lights model.

2.1 Group-specific marker trajectories

We suppose a group-specific marker trajectory and a generalized linear mixed model for each longitudinal marker given subgroup G, so that for the l-th outcome at time $t \geq 0$ one has

$$h_l\big(\mathbb{E}[Y^l(t)|b^l, G = k]\big) = m_k^l(t) \tag{2}$$

where h_l denotes a known one-to-one link function, and m_k^l the linear predictor such that

$$m_k^l(t) = u^l(t)^\top \beta_k^l + v^l(t)^\top b^l$$

where $u^l(t) \in \mathbb{R}^{q_l}$ is a row vector of (possibly) time-varying features with corresponding unknown fixed effect parameters β_k^l, and $v^l(t) \in \mathbb{R}^{r_l}$ is a row vector of (possibly) time-varying features with $r_l \leq q_l$ and where the corresponding subject-and-longitudinal outcome specific random effects b^l that does not depend on the group membership, which is not a strong modeling assumption.

Assumption 1. *We suppose that the random effects are independent of the group membership, and that the latter remain independent conditional on the observed data (namely T, Δ and Y).*

A suitable distributional assumption for the random effects component is a zero-mean multivariate normal distribution [Hickey et al., 2016], that is

$$b^l \sim \mathcal{N}(0, D_{ll})$$

with $D_{ll} \in \mathbb{R}^{r_l \times r_l}$ the unstructured variance-covariance matrix. To account for dependence between the different longitudinal outcome types, we let $\mathrm{Cov}[b^l, b^{l'}] = D_{ll'}$ for $l \neq l'$ and we denote

$$D = \begin{bmatrix} D_{11} & \cdots & D_{1L} \\ \vdots & \ddots & \vdots \\ D_{1L}^\top & \cdots & D_{LL} \end{bmatrix}$$

the global variance-covariance matrix. In the sequel of the paper, our aim is to associate the true and unobserved value $m_k^l(t)$ of the l-th longitudinal outcome at time t with the event outcome $T^\star$.

2.2 Group-specific risk of event

To quantify the effect of the longitudinal outcomes on the risk for an event, we use a Cox [Cox, 1972] relative risk model of the form

$$\lambda(t|\mathcal{M}_k(t), G = k) = \lambda_0(t) \exp\left\{x^\top \gamma_{k,0} + \sum_{l=1}^{L} \sum_{a=1}^{\mathcal{A}} {\gamma_{k,a}^l}^\top \varphi_a(t, \beta_k^l, b^l)\right\}, \tag{3}$$

where $\lambda_0(\cdot)$ is an unspecified baseline hazard function and given that $G = k$, we denote

$$\mathcal{M}_k(t) = \{m_k^l(u), 0 \leq u < t\}$$

the history of the true unobserved longitudinal process up to time t. We choose to incorporate x, also used in (1), with no *a priori* on the choice of time-independant features involved in the definition of π_{ξ_k} nor λ, since independant regularizations are used during inference, see (8). Corresponding fixed effects are denoted $\gamma_{k,0} \in \mathbb{R}^p$, and for each l-th

longitudinal outcome, we consider $\mathcal{A} \in \mathbb{N}_+$ known functionals φ_a defining a shared association with $\gamma_{k,a}^l \in \mathbb{R}^{\rho_a}$ the corresponding joint association parameters, and $\rho_a \in \mathbb{N}_+$ the dimension of the corresponding $\mathrm{Im}(\varphi_a)$. This can be viewed as a generalization of SREMs [Rizopoulos, 2010]. Let us finally denote

$$\gamma_k = (\gamma_{k,0}^\top, \gamma_{k,1}^1, \ldots, \gamma_{k,\mathcal{A}}^1, \ldots, \gamma_{k,1}^L, \ldots, \gamma_{k,\mathcal{A}}^L)^\top \in \mathbb{R}^{p+L\mathcal{A}}.$$

Specification of the functionals $(\varphi_a)_{a\in\mathcal{A}}$. The association structure between the longitudinal and the time-to-event submodels is key to the joint modeling framework. In spite of that, rationale for selecting shared associations has received little attention. We then propose to include some of the most common parameterization with no *a priori*, set out in Table 1, and let the model select the relevant ones through the regularization strategy described in Section 3.1.

Description	$\varphi_a(t, \beta_k^l, b^l)$	$\frac{\partial \varphi_a(t,\beta_k^l,b^l)}{\partial \beta_k^l}$	ρ_a	Reference
Linear predictor	$m_k^l(t)$	$u^l(t)$	1	Chi and Ibrahim [2006]
Random effects	b^l	$\mathbf{0}_{q_l}$	r_l	Hatfield et al. [2011]
Time-dependent slope	$\frac{\mathrm{d}}{\mathrm{d}t} m_k^l(t)$	$\frac{\mathrm{d}}{\mathrm{d}t} u^l(t)$	1	Rizopoulos and Ghosh [2011]
Cumulative effect	$\int_0^t m_k^l(s)\mathrm{d}s$	$\int_0^t u^l(s)\mathrm{d}s$	1	Andrinopoulou et al. [2017]

Table 1: Description of the shared associations included in the group-specific risk of event submodel (3). The gradient with respect to β_k^l is also given, which will be useful for inference in Section 3.2.

2.3 Likelihood

Consider an independent and identically distributed (i.i.d.) cohort of n subjects

$$\mathcal{D}_n = \{(x_1, y_1^1, \ldots, y_1^L, t_1, \delta_1), \ldots, (x_n, y_n^1, \ldots, y_n^L, t_n, \delta_n)\}$$

where for each subject $i = 1, \ldots, n$, process Y_i^l is measured n_i^l times at $t_{i1}^l, \ldots, t_{in_i^l}^l$ (which can differ between subjects and outcomes) with $t_{ij}^l \leq t_{ij+1}^l$ for all $j = 1, \ldots, n_i^l - 1$ and such that

$$y_i^l = (y_{i1}^l, \ldots, y_{in_i^l}^l)^\top \in \mathbb{R}^{n_i^l} \quad \text{with} \quad y_{ij}^l = Y_i^l(t_{ij}^l)$$

for all $l = 1, \ldots, L$. For the i-th subject, let us denote

$$\begin{cases} y_i &= ({y_i^1}^\top \cdots {y_i^L}^\top)^\top \in \mathbb{R}^{n_i}, \\ b_i &= ({b_i^1}^\top \cdots {b_i^L}^\top)^\top \in \mathbb{R}^{r}, \end{cases}$$

with $n_i = \sum_{l=1}^L n_i^l$ and $r = \sum_{l=1}^L r_l$ the total number of longitudinal measurements (for subject i) and the total dimension of the random effects respectively, as well as the following design matrices

$$U_i = \begin{bmatrix} U_{i1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & U_{iL} \end{bmatrix} \in \mathbb{R}^{n_i \times q} \quad \text{and} \quad V_i = \begin{bmatrix} V_{i1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & V_{iL} \end{bmatrix} \in \mathbb{R}^{n_i \times r}$$

with $q = \sum_{l=1}^{L} q_l$ and where for all $l = 1, \ldots, L$, one writes

$$\begin{cases} U_{il} &= \left(u_i^l(t_{i1}^l)^\top \cdots u_i^l(t_{in_i^l}^l)^\top\right)^\top \in \mathbb{R}^{n_i^l \times q_l}, \\ V_{il} &= \left(v_i^l(t_{i1}^l)^\top \cdots v_i^l(t_{in_i^l}^l)^\top\right)^\top \in \mathbb{R}^{n_i^l \times r_l}. \end{cases}$$

From now on, all computations are done conditionally on the design matrices $(U_i)_{i=1,\ldots,n}$ and $(V_i)_{i=1,\ldots,n}$.

Assumption 2. *Suppose that conditional on the random effects, each Y_i^l is independent, and that the censoring times T_i are independent of b_i.*

Assumption 2 is a standard modeling assumption, see for instance Tsiatis and Davidian [2004]. For all $k = 0, \ldots, K-1$, we denote

$$\beta_k = ({\beta_k^1}^\top \cdots {\beta_k^L}^\top)^\top \in \mathbb{R}^q$$

and

$$M_{ik} = U_i\beta_k + V_i b_i \in \mathbb{R}^{n_i}.$$

Given (2), each y_i^l is assumed to be from a one-parameter exponential family with respect to a reference measure which is either the Lebesgue measure (e.g., in the Gaussian case) or the counting measure (e.g., in the logistic cases). The conditional distribution of $y_i|b_i, G_i = k$ is then assumed to be from a distribution with a density of the form

$$f(y_i|b_i, G_i = k) = \exp\left\{(y_i \odot \Phi_i)^\top M_{ik} - c_\phi(M_{ik}) + d_\phi(y_i)\right\}, \tag{4}$$

with

$$\Phi_i = (\phi_1^{-1}{\mathbf{1}_{n_i^1}}^\top \cdots \phi_L^{-1}{\mathbf{1}_{n_i^L}}^\top)^\top \in \mathbb{R}^{n_i}$$

and $\phi = (\phi_1, \ldots, \phi_L)^\top \in \mathbb{R}^L$. The density described in (4) encompasses several distributions, see Table 2. The functions $c_\phi(\cdot)$ and $d_\phi(\cdot)$ are known as well as the dispersion parameters ϕ_l, while parameters β_k have to be estimated. Note that $d_\phi(\cdot)$ is related to the normalizing constant.

Model	Support	Use cases	ϕ_l	$h_l(\cdot)$ in (2)	$c_\phi(\cdot)$
Gaussian	$\mathbb{R}$	Continuous response data	σ_l^2	$z \mapsto z$	$z \mapsto z^2/2\sigma_l^2$
Categorical	$\{0,1\}^C$	Outcome with C modalities	1	$z \mapsto \log\left(\frac{z}{1-z}\right)$	$z \mapsto \log(1+e^z)$
Poisson	$\mathbb{N}$	Count of occurrences	1	$z \mapsto \log(z)$	$z \mapsto e^z$

Table 2: Examples of standard distributions that fit in the considered setting, given in the univariate case for simplicity of the notations.

Let us denote

$$\theta = \left(\xi_0^\top \cdots \xi_{K-1}^\top, \beta_0^\top \cdots \beta_{K-1}^\top, \phi^\top, \mathrm{vech}(D), \lambda_0(t), \gamma_0^\top \cdots \gamma_{K-1}^\top\right)^\top \in \mathbb{R}^\vartheta$$

the collection of the $\vartheta \in \mathbb{N}_+$ unknown parameters to estimate. To write the log-likelihood $\ell_n(\theta)$ (rescaled by n^{-1}) for samples in $\mathcal{D}_n$, corresponding to the joint distribution of the time-to-event and longitudinal outcomes, let us make the following hypothesis.

Assumption 3. *Assume that both the random effects vector b_i and the group membership account for the association between the longitudinal and event outcomes, that is*

$$f(t_i, \delta_i, y_i | b_i, G_i = k; \theta) = f(t_i, \delta_i | b_i, G_i = k; \theta) f(y_i | b_i, G_i = k; \theta) \tag{5}$$

for all $i = 1, \ldots, n$.

Assumption 3 is a generalization of classical hypothesis used in SREMs and JLCMs [Hickey et al., 2016]. Then, one has

$$\begin{aligned} \ell_n(\theta) &= \ell_n(\theta; \mathcal{D}_n) \\ &= n^{-1} \sum_{i=1}^{n} \log \int_{\mathbb{R}^r} \sum_{k=0}^{K-1} \pi_{\xi_k}(x_i) f(t_i, \delta_i | b_i, G_i = k; \theta) f(y_i | b_i, G_i = k; \theta) f(b_i; \theta) \mathrm{d}b_i, \end{aligned} \tag{6}$$

where

$$f(t_i, \delta_i | b_i, G_i = k; \theta) = \big[\lambda(t_i | \mathcal{M}_k(t_i), G_i = k)\big]^{\delta_i} \exp\Big\{ - \int_0^{t_i} \lambda(s | \mathcal{M}_k(s), G_i = k) \mathrm{d}s \Big\}$$

and

$$f(b_i; \theta) = (2\pi)^{-\frac{r}{2}} |D|^{-\frac{1}{2}} \exp\Big\{ - \frac{1}{2} b_i^\top D^{-1} b_i \Big\}. \tag{7}$$

Specification of the design matrices. In many practical applications, subjects show highly nonlinear longitudinal trajectories. In (6), the complete longitudinal history is required for the computation of the survival function. Hence, in order to produce a good estimate of $\mathcal{M}_k(t)$, we consider a flexible representations for $u^l(t)$ using a high-dimensional vector of time monomials, namely

$$u^l(t) = (1, t, t^2, \ldots, t^\alpha)^\top$$

with $\alpha \in \mathbb{N}_+$. The idea here is to allow a wide range of polynomial orders for the representation so that a suitable one can be automatically chosen for each trajectory – depending on its inherent complexity – thanks to the regularization strategy proposed in (8). We then let

$$v^l(t) = (1, t)^\top$$

so that each trajectory of each subject gets an affine random effect. Hence with this choice in practice, one has $q_l = \alpha + 1$ and $r_l = 2$ for all $l = 1, \ldots, L$.

3 Inference

In this section, we describe the procedure for estimating the parameters of the lights model. Let us first present the considered penalized objective, and then focus on the algorithm proposed for inference.

3.1 Penalized objective

In order to avoid overfitting and improve the prediction power of our method, we propose to minimize the penalized objective

$$\ell_n^{\mathrm{pen}}(\theta) = -\ell_n(\theta) + \sum_{k=0}^{K-1} \zeta_{1,k} \|\xi_k\|_{\mathrm{en},\eta} + \zeta_{2,k} \|\gamma_k\|_{\mathrm{sgl}_1,\tilde{\eta}} + \zeta_{3,k} \|\beta_k\|_{\mathrm{sgl}_1,\tilde{\eta}} \tag{8}$$

where for all $k = 0, \dots, K-1$, we add an elastic net regularization [Zou and Hastie, 2005] of the vector ξ_k and a sparse group lasso regularization [Simon et al., 2013] of the vectors γ_k and β_k, for tuning hyper-parameters $(\zeta_{1,k}, \zeta_{2,k}, \zeta_{3,k})^\top \in \mathbb{R}^3_+$. Here, $(\eta, \tilde{\eta}) \in [0,1]^2$ are fixed and we denote

$$\|z\|_{\text{en},\eta} = (1-\eta)\|z\|_1 + \frac{\eta}{2}\|z\|_2^2$$

for any vector z, that is a linear combination of the lasso (ℓ_1) and ridge (squared ℓ_2) penalties, and

$$\|z\|_{\text{sgl}_1,\tilde{\eta}} = (1-\tilde{\eta}) \sum_{l=1}^{L} \|z^l\|_2 + \tilde{\eta}\|z\|_1$$

for the sparse group lasso penalty, for which we do not have to account for group sizes since they all have the same one. Note that in the γ_k penalty, l actually starts at 0 (and not 1), but for simplicity we do not change the notation. Hence, the resulting optimization problem is written

$$\hat{\theta} \in \text{argmin}_{\theta \in \mathbb{R}^\vartheta} \ell_n^{\text{pen}}(\theta). \tag{9}$$

One advantage of the considered regularization method is its ability to perform feature selection (the lasso part) and pinpoint the most important features relatively to the prediction objective : the support of ξ_k thus informs on the time-independant features involved in the k-th group membership. The ridge part allows to handle potential correlation between features. On the other hand, the sparse group lasso allows to perform feature selection for each trajectory : the support of γ_k^l informs on the features involved in the k-th group risk of event for the l-th longitudinal outcome. Finally, the sparse group lasso of β_k allows to consider a flexible representation of time for the design matrices $(U_i)_{i=1,\dots,n}$ and lets the model automatically fit each trajectory l with the right complexity. Note that in practice, the intercept is not regularized.

3.2 A Quasi-Newton Monte Carlo EM

In order to derive an algorithm for this objective, we introduce a so-called QNMCEM algorithm, being a combination between an EM algorithm [Dempster et al., 1977] with Monte Carlo approximations [Levine and Casella, 2001], and multiple L-BFGS-B algorithms [Zhu et al., 1997]. EM algorithm has already been used for multivariate data joint modeling (see Lin et al. [2002a] for instance), but here we face different original problems: for each subject i, the latent variables are the pairs (G_i, b_i) (not only the random effects); and then, we want to minimize the penalized objective ℓ_n^{pen} (not "only" the negative log-likelihood).

We first need to compute the negative completed log-likelihood (here scaled by n^{-1}), namely the negative joint distribution of $\mathcal{D}_n$, $\boldsymbol{b} = (b_1, \dots, b_n)$ and $\boldsymbol{G} = (G_1, \dots, G_n)$. It

can be written

$$\begin{aligned}\ell_n^{\text{comp}}(\theta) &= \ell_n^{\text{comp}}(\theta; \mathcal{D}_n, \boldsymbol{b}, \boldsymbol{G}) \\ &= -n^{-1} \sum_{i=1}^{n} -\frac{1}{2}(r \log 2\pi + \log |D| + b_i^\top D^{-1} b_i) + \sum_{k=0}^{K-1} \mathbb{1}_{\{G_i=k\}} \Big[\log \pi_{\xi_k}(x_i) \\ &\quad + \delta_i \Big(\log \lambda_0(t_i) + x_i^\top \gamma_{k,0} + \sum_{l=1}^{L} \sum_{a=1}^{\mathcal{A}} {\gamma_{k,a}^l}^\top \varphi_a(t_i, \beta_k^l, b_i^l) \Big) \\ &\quad - \int_0^{t_i} \lambda_0(s) \exp \Big\{ x_i^\top \gamma_{k,0} + \sum_{l=1}^{L} \sum_{a=1}^{\mathcal{A}} {\gamma_{k,a}^l}^\top \varphi_a(s, \beta_k^l, b_i^l) \Big\} \mathrm{d}s \\ &\quad + (y_i \odot \Phi_i)^\top M_{ik} - c_\phi(M_{ik}) + d_\phi(y_i) \Big].\end{aligned}$$

Suppose that we are at step $w+1$ of the algorithm, with current iterate denoted $\theta^{(w)}$.

Monte Carlo E-step. We need to compute the expected negative log-likelihood of the complete data conditional on the observed data and the current estimate of the parameters given by

$$\mathcal{Q}_n(\theta, \theta^{(w)}) = \mathbb{E}_{\theta^{(w)}}[\ell_n^{\text{comp}}(\theta) | \mathcal{D}_n].$$

Given Assumption 1, the previous expression requires to compute expectations of the form

$$\mathbb{E}_{\theta^{(w)}}[g(b_i, G_i) | t_i, \delta_i, y_i] = \sum_{k=0}^{K-1} \pi_{ik}^{\theta^{(w)}} \int_{\mathbb{R}^r} g(b_i, G_i) f(b_i | t_i, \delta_i, y_i; \theta^{(w)}) \mathrm{d}b_i$$

for different functions g, where we denote

$$\pi_{ik}^{\theta^{(w)}} = \mathbb{P}_{\theta^{(w)}}[G_i = k | t_i, \delta_i, y_i] \tag{10}$$

the posterior probability of the latent class membership using parameters $\theta^{(w)}$. Then, either $g(b_i, G_i) = \tilde{g}(b_i)$ (for instance $\tilde{g} : b_i \mapsto b_i^\top D^{-1} b_i$), or $g(b_i, G_i) = g_k(G_i)$ with $g_k : G_i \mapsto \mathbb{1}_{\{G_i=k\}}$.

In the case $g(b_i, G_i) = \tilde{g}(b_i)$, one has

$$\mathbb{E}_{\theta^{(w)}}[\tilde{g}(b_i) | t_i, \delta_i, y_i] = \int_{\mathbb{R}^r} \tilde{g}(b_i) f(b_i | t_i, \delta_i, y_i; \theta^{(w)}) \mathrm{d}b_i = \frac{\sum_{k=0}^{K-1} \pi_{\xi_k^{(w)}}(x_i) \Lambda_{ik,\tilde{g}}^{\theta^{(w)}}}{\sum_{k=0}^{K-1} \pi_{\xi_k^{(w)}}(x_i) \Lambda_{ik,1}^{\theta^{(w)}}} \tag{11}$$

with

$$\Lambda_{ik,\tilde{g}}^{\theta^{(w)}} = \int_{\mathbb{R}^r} \tilde{g}(b_i) f(t_i, \delta_i, y_i | b_i, G_i = k; \theta^{(w)}) f(b_i; \theta^{(w)}) \mathrm{d}b_i, \tag{12}$$

and $\Lambda_{ik,1}^{\theta^{(w)}}$ obtained from $\Lambda_{ik,\tilde{g}}^{\theta^{(w)}}$ taking $\tilde{g} : b_i \mapsto 1$. The integral in (12) is not tractable analytically, and some form of approximation must be used in practice. Standard numerical integration techniques (such as Gaussian quadrature) are not well suited in our high-dimensional context. We then propose to use Monte Carlo approximation, which has already been applied for generalized linear mixed models [Booth and Hobert, 1999], with antithetic normal variates method [Hammersley and Morton, 1956] to reduce variance of the simulation result. Algorithm 1 describes how to construct the set $S^{(w)}$ used for the approximation.

Algorithm 1: Construction of the samples in $S^{(w)}$

Input: $S^{(w)} = \varnothing$

Output: $S^{(w)}$ filled with $2N$ samples

Compute $C^{(w)} \in \mathbb{R}^{r\times r}$ such that $C^{(w)}{C^{(w)}}^\top = D^{(w)}$ // Cholesky decomposition

for $\breve{n} = 1, \dots, N$ **do**

  Sample $\Omega_{\breve{n}} \sim \mathcal{N}(0, I_r)$

  Compute $\breve{b}_{\breve{n}} = C^{(w)}\Omega_{\breve{n}} \in \mathbb{R}^r$

  Update $S^{(w)} \leftarrow S^{(w)} \cup \{\breve{b}_{\breve{n}}, -\breve{b}_{\breve{n}}\}$

Return: $S^{(w)}$

The approximation of (11) is finally obtained by

$$\hat{\mathbb{E}}_{\theta^{(w)}}[\tilde{g}(b_i)|t_i, \delta_i, y_i] = \frac{\sum_{k=0}^{K-1} \pi_{\xi_k^{(w)}}(x_i)\hat{\Lambda}_{ik,\tilde{g}}^{\theta^{(w)}}}{\sum_{k=0}^{K-1} \pi_{\xi_k^{(w)}}(x_i)\hat{\Lambda}_{ik,1}^{\theta^{(w)}}} \tag{13}$$

with

$$\hat{\Lambda}_{ik,\tilde{g}}^{\theta^{(w)}} = \frac{1}{2N}\sum_{\breve{b}\in S^{(w)}} \tilde{g}(\breve{b}) f(t_i, \delta_i, y_i|\breve{b}, G_i = k; \theta^{(w)}). \tag{14}$$

It is common in Monte Carlo EM to increase N with the iterations w [Wei and Tanner, 1990] since it is computationally inefficient to use a large N in the early steps of the algorithm, when $\theta^{(w)}$ is far from $\hat{\theta}$. Multiple techniques have been proposed (see for instance Law et al. [2002] where a subjectively rule is followed, or Booth and Hobert [1999] that uses a rule based on confidence intervals for $\theta^{(w)}$ which requires additional variance estimation). We opt for a simple automated approach using relative differences of the objective function defined in (8) (see Algorithm 2).

Now in the case $g = g_k$, one has $\mathbb{E}_{\theta^{(w)}}[g_k(Gi)|t_i, \delta_i, y_i] = \pi_{ik}^{\theta^{(w)}}$ and we compute its approximation given

$$\hat{\pi}_{ik}^{\theta^{(w)}} = \frac{\pi_{\xi_k^{(w)}}(x_i)\hat{\Lambda}_{ik,1}^{\theta^{(w)}}}{\sum_{k=0}^{K-1} \pi_{\xi_k^{(w)}}(x_i)\hat{\Lambda}_{ik,1}^{\theta^{(w)}}}. \tag{15}$$

Quasi-Newton M-step. Here, we need to compute

$$\theta^{(w+1)} \in \operatorname{argmin}_{\theta\in\mathbb{R}^\vartheta} \mathcal{Q}_n(\theta, \theta^{(w)}) + \sum_{k=0}^{K-1} \zeta_{1,k}\|\xi_k\|_{\mathrm{en},\eta} + \zeta_{2,k}\|\gamma_k\|_{\mathrm{sgl}_1,\tilde{\eta}} + \zeta_{3,k}\|\beta_k\|_{\mathrm{sgl}_1,\tilde{\eta}}.$$

Let us introduce the following useful functions

$$\begin{cases} \tilde{g}^1_{s,\gamma_k,\beta_k} & : b_i \mapsto \exp\left\{x_i^\top\gamma_{k,0} + \tilde{g}^2_{s,\gamma_k,\beta_k}(b_i)\right\}, \\ \tilde{g}^2_{s,\gamma_k,\beta_k} & : b_i \mapsto \sum_{l=1}^{L}\sum_{a=1}^{\mathcal{A}} {\gamma_{k,a}^l}^\top \varphi_a(s, \beta_k^l, b_i^l). \end{cases} \tag{16}$$

We then present the parameters updates in the order given in Algorithm 2. First, the update of $D^{(w)}$ is naturally given in closed-form by

$$D^{(w+1)} = n^{-1}\sum_{i=1}^{n} \hat{\mathbb{E}}_{\theta^{(w)}}[b_i b_i^\top|t_i, \delta_i, y_i]. \tag{17}$$

Then, let us focus on the update of $\xi_k^{(w)}$ for $k = 0, \ldots, K-1$. We denote

$$P_{n,k}^{(w)}(\xi_k) = -n^{-1} \sum_{i=1}^{n} \hat{\pi}_{ik}^{\theta^{(w)}} \log \pi_{\xi_k}(x_i)$$

based on the quantities involved in $\mathcal{Q}_n(\theta, \theta^{(w)})$ that depend on ξ_k. The update for $\xi_k^{(w)}$ therefore requires to solve the following convex minimization problem

$$\xi_k^{(w+1)} \in \operatorname{argmin}_{\xi_k \in \mathbb{R}^p} P_{n,k}^{(w)}(\xi_k) + \zeta_{1,k} \|\xi_k\|_{\text{en},\eta}. \tag{18}$$

It looks like the logistic regression objective, where labels are not fixed but softly encoded by the expectation step (computation of $\hat{\pi}_{ik}^{\theta^{(w)}}$ in (15)). We then choose to solve (18) using the L-BFGS-B algorithm [Zhu et al., 1997] which belongs to the class of quasi-Newton optimization routines and solves the given minimization problem by computing approximations of the inverse Hessian matrix of the objective function. It can deal with differentiable convex objectives with box constraints. In order to use it with ℓ_1 penalization, which is not differentiable, we use the trick borrowed from Andrew and Gao [2007]: for $a \in \mathbb{R}$, write $|a| = a^+ + a^-$, where a^+ and a^- are respectively the positive and negative part of a, and add the constraints $a^+ \geq 0$ and $a^- \geq 0$. Namely, we rewrite the minimization problem (18) as the following differentiable problem with box constraints

$$\begin{aligned} \text{minimize} \quad & P_{n,k}^{(w)}(\xi_k^+ - \xi_k^-) + \zeta_{1,k}\big((1-\eta)\sum_{j=1}^{p}(\xi_{k,j}^+ + \xi_{k,j}^-) + \frac{\eta}{2}\|{\xi_k}^+ - {\xi_k}^-\|_2^2\big) \\ \text{subject to} \quad & \xi_{k,j}^+ \geq 0 \text{ and } \xi_{k,j}^- \geq 0 \text{ for } j = 1, \ldots, p \end{aligned} \tag{19}$$

where $\xi_k^{\pm} = (\xi_{k,1}^{\pm}, \ldots, \xi_{k,p}^{\pm})^\top$. The L-BFGS-B solver requires the exact value of the gradient, which is easily given by

$$\frac{\partial P_{n,k}^{(w)}(\xi_k)}{\partial \xi_k} = -n^{-1} \sum_{i=1}^{n} \hat{\pi}_{ik}^{\theta^{(w)}} \big(1 - \pi_{\xi_k}(x_i)\big) x_i. \tag{20}$$

Similarly to the strategy of increasing N with the iterations w for computational reasons, it is not clever to use a too small solver tolerance in the early steps of the algorithm, where we call tolerance the value for which iterations stop when the stopping criterion is below it. We then decrease the L-BFGS-B tolerance (which we denote *tol* in Algorithm 2) with every increase of N.

Now, for the update of $\beta_k^{(w)}$ for $k = 0, \ldots, K-1$, we similarly denote

$$\begin{aligned} R_{n,k}^{(w)}(\beta_k) = -n^{-1} \sum_{i=1}^{n} \hat{\pi}_{ik}^{\theta^{(w)}} \Big[& \delta_i \hat{\mathbb{E}}_{\theta^{(w)}}[\tilde{g}^2_{t_i, \gamma_k^{(w)}, \beta_k}(b_i) | t_i, \delta_i, y_i] \\ & - \sum_{j=1}^{J} \lambda_0^{(w)}(t_j) \hat{\mathbb{E}}_{\theta^{(w)}}[\tilde{g}^1_{t_j, \gamma_k^{(w)}, \beta_k}(b_i) | t_i, \delta_i, y_i] \mathbb{1}_{\{t_j \leq t_i\}} \\ & + (y_i \odot \Phi_i^{(w)})^\top \hat{\mathbb{E}}_{\theta^{(w)}}[M_{ik} | t_i, \delta_i, y_i] \\ & - \hat{\mathbb{E}}_{\theta^{(w)}}[c_{\phi^{(w)}}(M_{ik}) | t_i, \delta_i, y_i] \Big] \end{aligned}$$

based on the quantities involved in $\mathcal{Q}_n(\theta, \theta^{(w)})$ depending on β_k, with $\tilde{g}^1_{t_j, \gamma_k^{(w)}, \beta_k}$ and $\tilde{g}^2_{t_i, \gamma_k^{(w)}, \beta_k}$ defined in (16). The integration over the survival process has been replaced

with a finite sum over the process evaluated at the distinct event times $(t_1, \ldots, t_J)$ with $J \in \mathbb{N}_+$, since $\lambda_0^{(w)}$ defined in (27) is always zero except at observed event times. Hence, we update $\beta_k^{(w)}$ by solving the following problem

$$\begin{aligned} &\text{minimize} \quad R_{n,k}^{(w)}(\beta_k^+ - \beta_k^-) + \zeta_{3,k}\big((1-\tilde{\eta})\sum_{l=1}^{L}\|{\beta_k^l}^+ - {\beta_k^l}^-\|_2 + \tilde{\eta}\sum_{j=1}^{q}(\beta_{k,j}^+ + \beta_{k,j}^-)\big) \\ &\text{subject to} \quad \beta_{k,j}^+ \geq 0 \text{ and } \beta_{k,j}^- \geq 0 \text{ for } j = 1, \ldots, q. \end{aligned} \tag{21}$$

The value of the gradient for the Gaussian case is here given by

$$\begin{aligned} \frac{\partial R_{n,k}^{(w)}(\beta_k)}{\partial \beta_k^l} = -n^{-1}\sum_{i=1}^{n}\hat{\pi}_{ik}^{\theta^{(w)}}\Big[\sum_{a=1}^{\mathcal{A}}{\gamma_{k,a}^l}^{(w)}\Big(\delta_i\hat{\mathbb{E}}_{\theta^{(w)}}[\frac{\partial\varphi_a(t_i,\beta_k^l,b_i^l)}{\partial\beta_k^l}|t_i,\delta_i,y_i] \\ - \sum_{j=1}^{J}\lambda_0^{(w)}(t_j)\hat{\mathbb{E}}_{\theta^{(w)}}[\frac{\partial\varphi_a(t_j,\beta_k^l,b_i^l)}{\partial\beta_k^l}\tilde{g}^1_{t_j,\gamma_k^{(w)},\beta_k}(b_i)|t_i,\delta_i,y_i]\mathbb{1}_{\{t_j\leq t_i\}}\Big) \\ + U_{il}^\top {\phi_l^{(w)}}^{-2} I_{n_{il}}(y_i^l - U_{il}\beta_k^l - V_{il}\hat{\mathbb{E}}_{\theta^{(w)}}[b_i|t_i,\delta_i,y_i])\Big] \end{aligned} \tag{22}$$

for all $l = 1, \ldots, L$. Note that the gradients of the considered φ_a functions with respect to β_k^l are given in Table 1.

Concerning the update of $\gamma_k^{(w)}$ for $k = 0, \ldots, K-1$, let us follow the same strategy by denoting

$$\begin{aligned} Q_{n,k}^{(w)}(\gamma_k) = -n^{-1}\sum_{i=1}^{n}\hat{\pi}_{ik}^{\theta^{(w)}}\Big[\delta_i\hat{\mathbb{E}}_{\theta^{(w)}}[\log\circ\tilde{g}^1_{t_i,\gamma_k,\beta_k^{(w+1)}}(b_i)|t_i,\delta_i,y_i] \\ - \sum_{j=1}^{J}\lambda_0^{(w)}(t_j)\hat{\mathbb{E}}_{\theta^{(w)}}[\tilde{g}^1_{t_j,\gamma_k,\beta_k^{(w+1)}}(b_i)|t_i,\delta_i,y_i]\mathbb{1}_{\{t_j\leq t_i\}}\Big] \end{aligned}$$

based on the quantities involved in $\mathcal{Q}_n(\theta, \theta^{(w)})$ that depend on γ_k, with $\tilde{g}^1_{t_j,\gamma_k,\beta_k^{(w+1)}}$ defined in (16). The update for $\gamma_k^{(w)}$ therefore requires to solve the following minimization problem

$$\gamma_k^{(w+1)} \in \operatorname{argmin}_{\gamma_k\in\mathbb{R}^{p+L\mathcal{A}}} Q_{n,k}^{(w)}(\gamma_k) + \zeta_{2,k}\|\gamma_k\|_{\text{sgl}_1,\tilde{\eta}}. \tag{23}$$

We then rewrite problem (23) as the following differentiable problem with box constraints

$$\begin{aligned} &\text{minimize} \quad Q_{n,k}^{(w)}(\gamma_k^+ - \gamma_k^-) + \zeta_{2,k}\big((1-\tilde{\eta})\sum_{l=1}^{L}\|{\gamma_k^l}^+ - {\gamma_k^l}^-\|_2 + \tilde{\eta}\sum_{j=1}^{p+L\mathcal{A}}(\gamma_{k,j}^+ + \gamma_{k,j}^-)\big) \\ &\text{subject to} \quad \gamma_{k,j}^+ \geq 0 \text{ and } \gamma_{k,j}^- \geq 0 \text{ for } j = 1, \ldots, p+L\mathcal{A}. \end{aligned} \tag{24}$$

One can also compute the gradient since one has

$$\frac{\partial Q_{n,k}^{(w)}(\gamma_k)}{\partial\gamma_{k,0}} = -n^{-1}\sum_{i=1}^{n}\hat{\pi}_{ik}^{\theta^{(w)}}\Big[\delta_i - \sum_{j=1}^{J}\lambda_0^{(w)}(t_j)\hat{\mathbb{E}}_{\theta^{(w)}}[\tilde{g}^1_{t_j,\gamma_k,\beta_k^{(w+1)}}(b_i)|t_i,\delta_i,y_i]\mathbb{1}_{\{t_j\leq t_i\}}\Big]x_i \tag{25}$$

and

$$\frac{\partial Q_{n,k}^{(w)}(\gamma_k)}{\partial \gamma_{k,a}^l} = -n^{-1} \sum_{i=1}^{n} \hat{\pi}_{ik}^{\theta^{(w)}} \Big[\delta_i \hat{\mathbb{E}}_{\theta^{(w)}}[\varphi_a(t_i, \beta_k^{l\,(w+1)}, b_i^l) | t_i, \delta_i, y_i] - \sum_{j=1}^{J} \lambda_0^{(w)}(t_j) \hat{\mathbb{E}}_{\theta^{(w)}}[\varphi_a(t_j, \beta_k^{l\,(w+1)}, b_i^l) \tilde{g}^1_{t_j, \gamma_k, \beta_k^{(w+1)}}(b_i) | t_i, \delta_i, y_i] \mathbb{1}_{\{t_j \leq t_i\}} \Big] \tag{26}$$

for all $a = 1, \ldots, \mathcal{A}$ and $l = 1, \ldots, L$.

For the update of $\lambda_0^{(w)}$, we treat the jump size at the observed event times as parameters to be estimated [Klein, 1992]. The closed-form update is then given by

$$\lambda_0^{(w+1)}(t) = \frac{\sum_{i=1}^{n} \delta_i \mathbb{1}_{\{t=t_i\}}}{\sum_{i=1}^{n} \sum_{k=0}^{K-1} \hat{\pi}_{ik}^{\theta^{(w)}} \hat{\Lambda}^{\theta^{(w)}}_{ik, \tilde{g}^1_{t_i, \gamma_k^{(w+1)}, \beta_k^{(w+1)}}} \mathbb{1}_{\{t_i \geq t\}}}, \tag{27}$$

which is a Breslow like estimator [Breslow, 1972] adapted to our model.

Finally, for the update of $\phi^{(w)}$ and regarding Table 2, we focus on the Gaussian case being particularly useful in practice. It is obtained in closed-form by

$$\phi_l^{(w+1)} = \Big(\sum_{i=1}^{n} n_i^l \Big)^{-1} \sum_{i=1}^{n} \sum_{k=0}^{K-1} \hat{\pi}_{ik}^{\theta^{(w)}} \Big[\big(y_i^l - U_{il} \beta_k^{l\,(w+1)} \big)^\top \Big(y_i^l - U_{il} \beta_k^{l\,(w+1)} - 2 V_{il} \hat{\mathbb{E}}_{\theta^{(w)}}[b_i^l | t_i, \delta_i, y_i] \Big) + \mathrm{Tr}\left(V_{il}^\top V_{il} \hat{\mathbb{E}}_{\theta^{(w)}}[b_i^l b_i^{l\,\top} | t_i, \delta_i, y_i] \right) \Big]. \tag{28}$$

Initialization. Because our algorithm gives a local minimum, it is clever to choose an initial value $\theta^{(0)}$ relatively close to the final solution $\hat{\theta}$. Then, let us give some details about the starting point of Algorithm 2. For all $k = 0, \ldots, K-1$, we first use $\xi_k^{(0)}$ as the zero vector.

The QNMCEM algorithm. Algorithm 2 describes the main steps of the resulting QNMCEM algorithm to solve the optimization problem (9).

3.3 The case $K = 2$

In many practical applications – including those addressed in Section 6 – we are interested in identifying one subgroup of the population with a high risk of adverse event compared to the others. Then, in the following, we consider $G \in \{0, 1\}$ where $G = 1$ means high-risk of early adverse event and $G = 0$ means low risk. To simplify notations, let us set $\xi = \xi_1$, $\pi_\xi(x)$ the conditional probability that a patient belongs to the group with high risk of adverse event given its time-independent features x, and $\zeta_1 = \zeta_{1,1}$. In this context, which is the one we consider in practice in Sections 5 and 6, we suppose that the "right" penalty strength for ξ_k and γ_k does not depend on k, so that we denote $\zeta_2 = \zeta_{2,0} = \zeta_{2,1}$ and $\zeta_3 = \zeta_{3,0} = \zeta_{3,1}$.

Algorithm 2: QNMCEM algorithm for the lights model inference

Input: Training data $\mathcal{D}_n$; initialize $N = 50L$; tuning hyper-parameters $(\zeta_{1,k}, \zeta_{2,k}, \zeta_{3,k})_{k=0,\dots,K-1}$

Output: Last parameters $\hat{\theta} \in \mathbb{R}^{\vartheta}$

/* Initialization */

Compute the starting parameters $\theta^{(0)} \in \mathbb{R}^{\vartheta}$

for $w = 1, \dots,$ *until convergence* **do**

  /* Monte Carlo E-step */

  Compute $\hat{\Lambda}_{ik,\tilde{g}}^{\theta^{(w)}}$ using (14) for the required functions $\tilde{g}$

  Compute $\hat{\pi}_{ik}^{\theta^{(w)}}$ using (15)

  if *Condition* **then**

    $N \leftarrow N+$

    $tol \leftarrow tol-$

  /* Quasi-Newton M-step */

  Update $D^{(w+1)}$ using (17)

  Update $(\xi_k^{(w+1)})_{k=0,\dots,K-1}$ by solving (19)

  Update $(\beta_k^{(w+1)})_{k=0,\dots,K-1}$ by solving (21)

  Update $\lambda_0^{(w+1)}$ using (27)

  Update $(\gamma_k^{(w+1)})_{k=0,\dots,K-1}$ by solving (24)

  Update $\phi^{(w+1)}$ using (28)

Return: $\hat{\theta}$

Cross-validation procedure. The hyper-parameters $(\zeta_1, \zeta_2, \zeta_3)^\top \in \mathbb{R}_+^3$ are then tuned during a 10-fold randomized search cross-validation procedure [Bergstra and Bengio, 2012] using predictive marker defined in Section 4.1 and the C-index score defined in Section 4.2. Random search is indeed more efficient in terms of computing times than classical grid search for hyper-parameters optimization and finds better models by effectively searching in a larger configuration space, which is appropriate to our high-dimensional problem. Moreover, the search interval for hyper-parameter ζ_1 is automatically computed in a data driven way described in Appendix A, while the search intervals for ζ_2 and ζ_3 are determined empirically.

4 Performance evaluation

In this section, we first present the proposed evaluation strategy to assess real-time prognostic prediction performances. We then introduce the appropriate metric used to this end: the C-index ; and finally, the considered models for performance comparisons.

4.1 Evaluation strategy: the real-time prediction paradigm

When validating predictions or comparing performances of competing models in a survival context, one can be interested either in the discriminative power of the predictive rule and use concordance measures such as the C-index [Heagerty and Zheng, 2005] (defined in Section 4.2), or in the predictive accuracy of the rule [Schemper and Henderson, 2000]. Developments in the field of joint modeling have primarily focused on modeling and estimation, and most studies do not consider goodness-of-fit nor generalization power

in a prognostic prediction perspective [Hickey et al., 2016]. However, with the prospect of using prognostic prediction in real-time on a daily basis, practitioners will naturally require predictive prognostic tools to evaluate models fit and compare models in terms of discriminative power. So we choose to put ourselves in this paradigm.

Predictive marker. The question now is to choose a discriminative marker rule for risk prediction to be used in the cross-validation procedure for selecting the best regularization hyper-parameters, and of course to be used as final risk prediction after hyper-parameters fine-tuning. The lights model has the ability to produce prognostic predictions that can be dynamically updated according to the observed trajectory of the features. Once the model is trained (so that one obtains $\hat{\theta}$ from (9) using our QNMCEM algorithm introduced in Section 3.2), posterior risk classification for subject i and group k can be made through $\hat{\pi}_{ik}^{\hat{\theta}}$ (see (15)). Similar posterior probabilities have been considered for goodness-of-fit evaluation in other JLCMs models, see Proust-Lima et al. [2014] for instance.

But in the real-time prediction paradigm, it makes no sense from a practical point of view to use this marker rule for risk prediction, since the latter requires to know the survival labels (t_i, δ_i), being intrinsically unknow in a context where we want to perform real-time risk prediction. Denoting t_i^{max} the time for subject i when one wants to perform the risk prediction – so in practice, the time up to which one has data measurements for Y_i, say the "present" time for prediction – we define

$$\mathcal{R}_i = \sum_{k=0}^{K-1} \pi_{\hat{\xi}_k}(x_i) \exp\left\{\tilde{g}^1_{t_i^{max},\hat{\gamma}_k,\hat{\beta}_k}(b_i)\right\} \tag{29}$$

as predictive marker rule of the lights model for subject i.

4.2 The C-index metric

We detail in this section the metric considered to evaluate risk prediction performances. Let us denote by M the marker under study (e.g., the posterior group membership probabilities $\hat{\pi}_{ik}^{\hat{\theta}}$ (see (15)) for the lights model), and assume that it is measured once at $t = 0$. A common concordance measure that does not depend on time is the C-index [Harrell et al., 1996] defined by

$$\mathcal{C} = \mathbb{P}[M_i > M_j | T_i^\star < T_j^\star],$$

with $i \neq j$ two independent subjects (which does not depend on i, j under the i.i.d. sample hypothesis). In our case, $T^\star$ is subject to right censoring, so one would typically consider the modified $\mathcal{C}_\tau$ defined by

$$\mathcal{C}_\tau = \mathbb{P}[M_i > M_j | T_i < T_j, T_i < \tau],$$

with τ corresponding to the fixed and prespecified follow-up period duration [Heagerty and Zheng, 2005]. A Kaplan-Meier estimator for the censoring distribution leads to a nonparametric and consistent estimator of $\mathcal{C}_\tau$ [Uno et al., 2011], already implemented in the `Python` package `lifelines`. Hence in the following, we consider the C-index metric to assess performances.

4.3 Competing models

In this section, we briefly introduce the models we consider for performance comparisons in the simulation study as well as in the applications on real datasets in Section 6.

Penalized Cox model with time-independant features. The first model we consider as a baseline is the well known Cox PH model with time-independant features. In this model introduced in Cox [1972], a parameter vector β is estimated by minimizing the partial log-likelihood given by

$$\ell_n^{\text{cox}}(\beta) = n^{-1} \sum_{i=1}^{n} \delta_i \big(x_i^\top \beta - \log \sum_{i':t_{i'} \geq t_i} \exp(x_{i'}^\top \beta) \big).$$

We use respectively the `R` packages `survival` and `glmnet` [Simon et al., 2011] for the partial log-likelihood and the minimization of the following quantity

$$-\ell_n^{\text{cox}}(\beta) + \xi \big((1-\eta)\|\beta\|_1 + \frac{\eta}{2}\|\beta\|_2^2 \big),$$

where ξ is chosen by the a 10-fold cross-validation procedure, for a given $\eta \in [0,1]$. Ties are handled via the Breslow approximation of the partial likelihood [Breslow, 1972]. We also choose to include basic time-independant features extracted from longitudinal processes, namely the average value for each time-dependant feature.

The time-dependent Cox model. A classical extension of the Cox model supposes that features depend on time [Sueyoshi, 1992] (and then loosing the risk proportionality property), that is with our notations

$$\lambda_i(t) = \lambda_0(t) \exp\left(y_i(t)^\top \beta \right).$$

Using the `survival` `R` package [Zhang et al., 2018], blabla description...

Multivariate joint modeling. `joineRML` package : in this `R` package [Hickey et al., 2018], blabla description...

The `JM` package : in this `R` package [Rizopoulos, 2010], blabla description...

The `JMbayes` package : n this `R` package [Rizopoulos, 2014], blabla description...

The `joineR` package : in this `R` package [Philipson et al., 2018], blabla description...

5 High-dimensional simulation study

In this section, we provide details regarding data generation, followed by the results of the extensive Monte Carlo simulation study to examine our method and compare it with state-of-the-art.

5.1 Simulation design

In order to assess the proposed method, we perform an extensive Monte Carlo simulation study.

simulation of time-varying features in a Cox model: see Therneau et al. [2017]

5.2 Results of simulation

Let us present now the simulation results.

6 Applications

In this section, we apply our lights method on two publicly available datasets and compare its performance with state-of-the-art methods.

6.1 The PBCseq dataset

This dataset is a follow-up to the original dataset [Fleming and Harrington, 2011, Murtaugh et al., 1994]from the Mayo Clinic trial in primary biliary cirrhosis (PBC) of the liver conducted between 1974 and 1984. A total of 424 PBC patients, referred to Mayo Clinic during that ten-year interval, met eligibility criteria for the randomized placebo controlled trial of the drug D-penicillamine. The first 312 cases in the data set participated in the randomized trial and contain largely complete data. The additional 112 cases did not participate in the clinical trial, but consented to have basic measurements recorded and to be followed for survival. Six of those cases were lost to follow-up shortly after diagnosis, so the data here are on an additional 106 cases as well as the 312 randomized participants. The dataset contains only baseline measurements of the laboratory paramters and contains multiple laboratory results, but only on the first 312 patients.

6.2 The MIMIC III dataset

The MIMIC III (Medical Information Mart for Intensive Care III) database is a large, freely-available hospital dataset containing de-identified data from over 40,000 patients. This data comes from patients who were admitted in critical care units to Beth Israel Deaconess Medical Center in Boston, Massachusetts from 2001 to 2012 [Johnson et al., 2016]. The dataset was populated with data that had been acquired during routine hospital care, so there was no associated burden on caregivers and no interference with their workflow.

7 Conclusion

In this paper, a generalized joint model for high-dimensional multivariate longitudinal data and censored durations (lights) has been introduced, and a new efficient estimation algorithm (QNMCEM) has been derived, that considers a penalization of the likelihood in order to perform covariate selection and to prevent overfitting.

Software

All the methodology discussed in this paper is implemented in `Python`. The code is available from https://github.com/SimonBussy/lights in the form of annotated programs, together with a notebook tutorial.

Acknowledgements

Conflict of Interest: None declared.

Appendices

A Numerical details

Concerning the cross-validation procedure for tuning

$$(\zeta_{1,0}, \zeta_{2,0}, \zeta_{3,0}, \ldots, \zeta_{1,K-1}, \zeta_{2,K-1}, \zeta_{3,K-1})^\top \in \mathbb{R}_+^{3K},$$

we use a randomized search with the C-index metric (see Section 4.1) where for all $k = 0, \ldots, K-1$ and $j \in \{1, 2, 3\}$, each $\zeta_{j,k}$ finds its candidates in the interval

$$[\zeta_{j,k}^{\max} \times 10^{-4}, \zeta_{j,k}^{\max}] \subset \mathbb{R},$$

with $\zeta_{j,k}^{\max}$ the interval upper bound computed hereafter.

Considering the convex minimization problem (18) at a given step w, let us denote $\zeta_{1,k}^1 \leq \zeta_{1,k}^2 \leq \cdots \leq \zeta_{1,k}^{\max}$ the randomly chosen candidate values for $\zeta_{1,k}$, such that at $\zeta_{1,k}^{\max}$, all coefficients $\hat{\xi}_{k,j}$ for all $j = 1, \ldots, p$ are exactly zero. The KKT conditions [Boyd and Vandenberghe, 2004] claim that

$$\begin{cases} \dfrac{\partial P_{n,k}^{(w)}(\hat{\xi}_k)}{\partial \xi_{k,j}} = \zeta_{1,k}\big((\eta - 1)\ \mathrm{sgn}(\hat{\xi}_{k,j}) - \eta \hat{\xi}_{k,j}\big) & \forall j \in \hat{\mathcal{A}}_k \\ \left| \dfrac{\partial P_{n,k}^{(w)}(\hat{\xi}_k)}{\partial \xi_{k,j}} \right| < \zeta_{1,k}(1-\eta) & \forall j \notin \hat{\mathcal{A}}_k \end{cases},$$

where $\hat{\mathcal{A}}_k = \{j = 1, \ldots, p : \hat{\xi}_{k,j} \neq 0\}$ is the active set of the $\hat{\xi}_k$ estimator, and for all $x \in \mathbb{R} \setminus \{0\}$, $\mathrm{sgn}(x) = \mathbb{1}_{\{x>0\}} - \mathbb{1}_{\{x<0\}}$. Then, using (20), one obtains

$$\hat{\xi}_{k,j} = 0 \ \Rightarrow \ \left| n^{-1} \sum_{i=1}^{n} \hat{\pi}_{ik}^{\theta^{(w)}} \big(1 - \pi_{\xi_k}(x_i)\big) x_{ij} \right| < \zeta_{1,k}(1-\eta)$$

for all $j = 1, \ldots, p$. Hence, we choose the following upper bound for the randomly chosen candidate interval during the cross-validation procedure

$$\zeta_{1,k}^{\max} = \frac{1}{n(1-\eta)} \max_{j=1,\ldots,p} \sum_{i=1}^{n} |x_{ij}|.$$

Similar strategy to automatically compute $\zeta_{2,k}^{\max}$ or $\zeta_{3,k}^{\max}$ is not easy since without additionnal hypothesis, the gradients 22, 25 and 26 are not upper bounded by a quantity that depends only on the data and not on the current iteration parameters estimates. We then choose those interval upper bounds empirically.

References

Per K Andersen, Ornulf Borgan, Richard D Gill, and Niels Keiding. *Statistical models based on counting processes*. Springer Science & Business Media, 2012.

Galen Andrew and Jianfeng Gao. Scalable training of l1-regularized log-linear models. In *International Conference on Machine Learning*, pages 33–40. ACM, 2007.

Eleni-Rosalina Andrinopoulou, Dimitris Rizopoulos, Johanna JM Takkenberg, and Emmanuel Lesaffre. Combined dynamic predictions using joint models of two longitudinal outcomes and competing risk data. *Statistical methods in medical research*, 26(4):1787–1801, 2017.

James Bergstra and Yoshua Bengio. Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13(2):281–305, 2012.

James G Booth and James P Hobert. Maximizing generalized linear mixed model likelihoods with an automated monte carlo em algorithm. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(1):265–285, 1999.

Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

Norman E Breslow. Contribution to discussion of paper by dr cox. *JJournal of the Royal Statistical Society: Series B (Statistical Methodology)*, 34:216–217, 1972.

Simon Bussy, Agathe Guilloux, Stéphane Gaïffas, and Anne-Sophie Jannot. C-mix: A high-dimensional mixture model for censored durations, with applications to genetic data. *Statistical methods in medical research*, 28(5):1523–1539, 2019.

Yueh-Yun Chi and Joseph G Ibrahim. Joint models for multivariate longitudinal and multivariate survival data. *Biometrics*, 62(2):432–445, 2006.

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

Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977.

Thomas R Fleming and David P Harrington. *Counting processes and survival analysis*, volume 169. John Wiley & Sons, 2011.

JM Hammersley and KW Morton. A new monte carlo technique: antithetic variates. In *Mathematical proceedings of the Cambridge philosophical society*, volume 52, pages 449–475. Cambridge University Press, 1956.

Frank E Harrell, Kerry L Lee, and Daniel B Mark. Tutorial in biostatistics multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in medicine*, 15:361–387, 1996.

Laura A Hatfield, Mark E Boye, and Bradley P Carlin. Joint modeling of multiple longitudinal patient-reported outcomes and survival. *Journal of Biopharmaceutical Statistics*, 21(5):971–991, 2011.

Patrick J Heagerty and Yingye Zheng. Survival model predictive accuracy and roc curves. *Biometrics*, 61(1):92–105, 2005.

Graeme L Hickey, Pete Philipson, Andrea Jorgensen, and Ruwanthi Kolamunnage-Dona. Joint modelling of time-to-event and multivariate longitudinal outcomes: recent developments and issues. *BMC medical research methodology*, 16(1):117, 2016.

Graeme L Hickey, Pete Philipson, Andrea Jorgensen, and Ruwanthi Kolamunnage-Dona. joinerml: a joint model and software package for time-to-event and multivariate longitudinal outcomes. *BMC medical research methodology*, 18(1):50, 2018.

Miran A Jaffa, Mulugeta Gebregziabher, and Ayad A Jaffa. A joint modeling approach for right censored high dimensional multivariate longitudinal data. *Journal of biometrics & biostatistics*, 5(4), 2014.

Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3:160035, 2016.

John P Klein. Semiparametric estimation of random effects using the cox model based on the em algorithm. *Biometrics*, pages 795–806, 1992.

Ngayee J Law, Jeremy MG Taylor, and Howard Sandler. The joint modeling of a longitudinal disease progression marker and the failure time process in the presence of cure. *Biostatistics*, 3(4):547–563, 2002.

Richard A Levine and George Casella. Implementations of the monte carlo em algorithm. *Journal of Computational and Graphical Statistics*, 10(3):422–439, 2001.

Haiqun Lin, Charles E McCulloch, and Susan T Mayne. Maximum likelihood estimation in the joint analysis of time-to-event and multiple longitudinal variables. *Statistics in Medicine*, 21(16):2369–2382, 2002a.

Haiqun Lin, Bruce W Turnbull, Charles E McCulloch, and Elizabeth H Slate. Latent class models for joint analysis of longitudinal biomarker and event process data: application to longitudinal prostate-specific antigen readings and prostate cancer. *Journal of the American Statistical Association*, 97(457):53–65, 2002b.

Paul A Murtaugh, E Rolland Dickson, Gooitzen M Van Dam, Michael Malinchoc, Patricia M Grambsch, Alice L Langworthy, and Chris H Gips. Primary biliary cirrhosis: prediction of short-term survival based on repeated patient visits. *Hepatology*, 20(1): 126–134, 1994.

Pete Philipson, Ines Sousa, Peter J Diggle, and Paula Williamson. Package 'joiner'. 2018.

Cécile Proust-Lima, Mbéry Séne, Jeremy MG Taylor, and Hélène Jacqmin-Gadda. Joint latent class models for longitudinal and time-to-event data: A review. *Statistical methods in medical research*, 23(1):74–90, 2014.

Dimitris Rizopoulos. The r package jmbayes for fitting joint models for longitudinal and time-to-event data using mcmc. *arXiv preprint arXiv:1404.7625*, 2014.

Dimitris Rizopoulos and Pulak Ghosh. A bayesian semiparametric multivariate joint model for multiple longitudinal outcomes and a time-to-event. *Statistics in medicine*, 30(12): 1366–1380, 2011.

Dimitris D Rizopoulos. Jm: An r package for the joint modelling of longitudinal and time-to-event data. *Journal of Statistical Software (Online)*, 35(9):1–33, 2010.

Michael Schemper and Robin Henderson. Predictive accuracy and explained variation in cox regression. *Biometrics*, 56(1):249–255, 2000.

Noah Simon, Jerome Friedman, Trevor Hastie, Rob Tibshirani, et al. Regularization paths for cox's proportional hazards model via coordinate descent. *Journal of statistical software*, 39(5):1–13, 2011.

Noah Simon, Jerome Friedman, Trevor Hastie, and Robert Tibshirani. A sparse-group lasso. *Journal of computational and graphical statistics*, 22(2):231–245, 2013.

Glenn T Sueyoshi. Semiparametric proportional hazards estimation of competing risks models with time-varying covariates. *Journal of econometrics*, 51(1-2):25–58, 1992.

Terry Therneau, Cindy Crowson, and Elizabeth Atkinson. Using time dependent covariates and time dependent coefficients in the cox model. *Survival Vignettes*, 2017.

Anastasios A Tsiatis and Marie Davidian. Joint modeling of longitudinal and time-to-event data: an overview. *Statistica Sinica*, pages 809–834, 2004.

Hajime Uno, Tianxi Cai, Michael J Pencina, Ralph B D'Agostino, and LJ Wei. On the c-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Statistics in medicine*, 30(10):1105–1117, 2011.

Jeroen K Vermunt and Jay Magidson. Latent class models for classification. *Computational Statistics & Data Analysis*, 41(3-4):531–537, 2003.

Ping Wang, Wei Shen, and Mark Ernest Boye. Joint modeling of longitudinal outcomes and survival using latent growth modeling approach in a mesothelioma trial. *Health Services and Outcomes Research Methodology*, 12(2-3):182–199, 2012.

Greg CG Wei and Martin A Tanner. A monte carlo implementation of the em algorithm and the poor man's data augmentation algorithms. *Journal of the American statistical Association*, 85(411):699–704, 1990.

Michael S Wulfsohn and Anastasios A Tsiatis. A joint model for survival and longitudinal data measured with error. *Biometrics*, pages 330–339, 1997.

Menggang Yu, Ngayee J Law, Jeremy MG Taylor, and Howard M Sandler. Joint longitudinal-survival-cure models and their application to prostate cancer. *Statistica Sinica*, pages 835–862, 2004.

Zhongheng Zhang, Jaakko Reinikainen, Kazeem Adedayo Adeleke, Marcel E Pieterse, and Catharina GM Groothuis-Oudshoorn. Time-varying covariates and coefficients in cox regression models. *Annals of translational medicine*, 6(7), 2018.

Ciyou Zhu, Richard H Byrd, Peihuang Lu, and Jorge Nocedal. Algorithm 778: L-bfgs-b: Fortran subroutines for large-scale bound-constrained optimization. *ACM Transactions on Mathematical Software (TOMS)*, 23(4):550–560, 1997.

Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.