
Journal of Computational and Graphical Statistics

ISSN: 1061-8600 (Print) 1537-2715 (Online) Journal homepage: https://www.tandfonline.com/loi/ucgs20

# Generalized Link-Based Additive Survival Models with Informative Censoring

Robinson Dettoni, Giampiero Marra & Rosalba Radice

# Generalized Link-Based Additive Survival Models with Informative Censoring

Robinson Dettoni [1,2,*], Giampiero Marra[3], Rosalba Radice[4]

[1]Department of Economics, Universidad de Santiago de Chile, Santiago 9170 022, Chile

[2]Department of Statistical Science, University College London, Gower Street, London WC1E 6BT, UK

[3]Department of Statistical Science, University College London, Gower Street, London WC1E 6BT, UK

[4]Cass Business School, City, University of London, 106 Bunhill Row, EC1Y 8TZ London, UK

*Corresponding author: r.dettoni@ucl.ac.uk



### Abstract

Time to event data differ from other types of data because they are censored. Most of the related estimation techniques assume that the censoring mechanism is non-informative while in many applications it can actually be informative. The aim of this work is to introduce a class of flexible survival models which account for the information provided by the censoring times. The baseline functions are estimated non-parametrically by monotonic P-splines, whereas covariate effects are flexibly determined using additive predictors. Parameter estimation is reliably carried out within a penalised maximum likelihood framework with integrated automatic multiple smoothing parameter selection. We derive the $\sqrt{n}$-consistency and asymptotic normality of the non-informative and informative estimators, and shed light on the efficiency gains produced by the newly introduced informative estimator when compared to its non-informative counterpart. The finite sample properties of the estimators are investigated via a Monte Carlo simulation study which highlights the good empirical performance of the proposal. The modelling framework is illustrated on data about infants hospitalised for pneumonia. The models and methods discussed in the paper have been implemented in the R package GJRM to allow for transparent and reproducible research.

Key Words: additive predictor, informative censoring, link-based survival model, penalised maximum likelihood, smoothing.

# 1 Introduction

Time to event data are different from other types of data because of censoring. This means that the response of interest, the time until a particular event occurs, can not be totally observed. As a result, models must be used to relate the observed and unobserved parts of the data since the recorded observations alone can not provide direct information on the event of interest. Most of the related estimation techniques assume that the censoring scheme is independent and non-informative conditional on covariates (e.g., Cox, 1972; Ma et al., 2014; Scheike & Zhang, 2003; Xue et al., 2018; Younes & Lachin, 1997). In many applications, however, these assumptions can at least be questioned (e.g., Chen, 2010; Huang & Zhang, 2008; Li & Peng, 2015; Lu & Zhang, 2012; Slud & Rubinstein, 1983; Wang et al., 2015; Xu et al., 2017, 2018; Zheng & Klein, 1995; Zeng et al., 2004).

Censoring is independent when the hazard rate of the event of interest for the censored observations is equal to the hazard rate for the uncensored ones, otherwise it is called dependent (Kalbfleisch & Prentice, 2002). If the event and censoring times are assumed to be dependent, then survival models accounting for this feature of the data face a problem of identification. In general, without additional assumptions, it is not possible to identify the survival distribution from the censored data alone or testing whether the censoring and survival mechanisms are independent (Cox, 1959; Tsiatis, 1975).

Censoring is informative when the censoring times, say $T_2$, contain information on the parameters of the distribution of the event variable, say $T_1$ (Lagakos, 1979; Kalbfleisch & Prentice, 2002). In particular, let us write the hazard functions for the event and censored times as $\mathrm{h}_{T_1}(t \mid \mathbf{x}_{T_1}; \boldsymbol{\theta}_{T_1})$ and $\mathrm{h}_{T_2}(t \mid \mathbf{x}_{T_2}; \boldsymbol{\theta}_{T_2})$. If the vector of parameters $\boldsymbol{\theta}_{T_1}$ and $\boldsymbol{\theta}_{T_2}$ have components in common then censoring is informative. In this case, the observable data $(Y, \delta) = \{\min(T_1, T_2), I(T_1 < T_2)\}$, where $I$ is the usual indicator function, provide sufficient information to identify the marginal survival functions of $T_1$ and $T_2$ (Kalbfleisch & Prentice, 2002).

Although dependent censoring is a well studied problem in the survival analysis and competing risk literature (e.g., Crowder, 2012; Emura & Chen, 2018), the specific literature analysing the problem of informative censoring is scarce, even though ignoring it may have detrimental consequences on inferential conclusions (e.g., Siannis et al., 2005; Lu & Zhang, 2012). In a seminal work, Koziol & Green (1976) proposed an informative survival model where the hazard functions of $T_1$ and $T_2$ satisfy $\mathrm{h}_{T_2}(t) = p\mathrm{h}_{T_1}(t)$, for some constant $0 < p < 1$. Since this model did not incorporate covariates, it was further extended. For instance, Yuan (2005) introduced a semiparametric Cox model estimated via profile likelihood in which, for a given vector of covariates $\mathbf{x}$, $\mathrm{h}_{T_2}(t \mid \mathbf{x}) = \varrho(t, \mathbf{x}; \theta)\mathrm{h}_{T_1}(t \mid \mathbf{x})$, where $\varrho$ is a function known up to a finite-dimensional parameter, $\theta$. The purpose of $\varrho$ was to capture the possible information contained in the censoring times. Lu & Zhang (2012) proposed a semi-parametric informative survival model where the baseline hazards are estimated non-parametrically and the covariate effects parametrically. In their approach, the hazard functions of $T_1$ and $T_2$ conditional on $\mathbf{x}$ are modelled using $\mathrm{h}_{T_\upsilon}(t \mid \mathbf{x}) = \mathrm{h}_{0,T_\upsilon}(t)\exp(\mathbf{x}^\top \boldsymbol{\varphi}_\upsilon)$, where $\mathbf{x}^\top \boldsymbol{\varphi}_\upsilon = \mathbf{x}_1^\top \vartheta_0 + \mathbf{x}_2^\top \vartheta_\upsilon$, for $\upsilon = 1, 2$.

In this article we deal with informative censoring. In particular, we develop a flexible, general and tractable survival modelling framework where the baseline functions are estimated non-parametrically via means of monotonic P-splines, covariate effects are flexibly determined using additive predictors, and informative censoring is accounted for. Model fitting is based on an optimization scheme that allows for the reliable simultaneous penalized estimation of all model's parameters as well as for stable and fast automatic multiple smoothing parameter selection. We provide the $\sqrt{n}$ -consistency and asymptotic normality of the non-informative and informative estimators, and show that the newly introduced informative estimator is more efficient than its non-informative counterpart. A Monte Carlo simulation study highlights the merits of the proposal, and the modelling framework is illustrated on data about infants hospitalised for pneumonia. The models and methods introduced in the article have been implemented in the `R` package `GJRM` (Marra & Radice, 2019) to allow for transparent and reproducible research. To the best of our knowledge, there are no alternative flexible survival models with informative censoring, nor respective software implementations, of the type proposed here. Given that the assumption of absence of informative censoring is often made for convenience, the proposed methodology is likely to appeal the wider audience wishing to estimate possibly more realistic survival models or at least assess whether allowing for informative censoring can produce more plausible results.

The article is organized as follows. In the next section, the proposed model and its theoretical properties are discussed. In Section 3, the effectiveness of the proposed methodology is explored by means of a simulation study. In Section 4, the framework is illustrated on data about infants hospitalised for pneumonia. Section 5 concludes the paper with a discussion.

## 2 Methodology

In this paper, the case of right censored data is considered; the true event time is not always observed, in which case censoring (lower) times are observed. For individual *i*, where $i = 1, \ldots, n$ and *n* represents the sample size, let $T_{1i}$ and $T_{2i}$ denote the true event and censoring times. Let also $\mathbf{z}_{\nu i}^{\top} = (z_{\nu 1 i}, \ldots, z_{\nu K_{\nu} i})$ be a vector of baseline covariates of dimension $K_{\nu}$, where $\mathbf{z}^{\top}$ stands for the transpose of a vector **z**, $\nu = 1, 2$ and $\mathbf{z}_i^{\top} = (\mathbf{z}_{1i}, \mathbf{z}_{2i})$. It is assumed that the $(T_{1i}, \mathbf{z}_i)$, for $i = 1, ..., n$, are independently and identically distributed (*i.i.d.*). The censoring times, $T_{2i}$, are also assumed to be *i.i.d*. The distribution of $T_2$ depends on **z**. In addition, we assume that $T_{1i}$ and $T_{i2}$ are conditionally independent given $\mathbf{z}_i$, and that $T_{i1}$ is informatively right censored by $T_{i2}$ through some covariates (Andersen & Keiding, 2006). We observe $(Y_i, \mathbf{z}_i, \delta_{1i})$, where $Y_i = \min\{T_{1i}, T_{2i}\}$ and $\delta_{1i} = I(T_{1i} \leq T_{2i})$. We also define $\delta_{2i} = [1 - \delta_{1i}]$. Finally, $\boldsymbol{\theta}$ is a generic vector of parameters.

## 2.1 Survival functions

The survival function of $T_{\nu i}$ taking values in (0, 1), conditional on $\mathbf{z}_{\nu i}$ and $\boldsymbol{\theta}_{\nu}$, can be expressed as

$$P(T_{\nu i} > t_{\nu i} \mid \mathbf{z}_{\nu i}; \boldsymbol{\theta}_{\nu}) = S_{\nu}(t_{\nu i} \mid \mathbf{z}_{\nu i}; \boldsymbol{\theta}_{\nu}) = \mathcal{G}_{\nu}[\xi_{\nu i}(t_{\nu i}, \mathbf{z}_{\nu i}; \boldsymbol{\theta}_{\nu})], \qquad (1)$$

where, for $\nu = 1, 2, \boldsymbol{\theta}_{\nu}$ and $\mathbf{z}_{\nu i}$ represent generic vectors of coefficients and covariates, respectively. The survival functions are modelled using generalised survival or link-based functions models (Younes & Lachin, 1997; Liu et al., 2018). That is, $S_{\nu}(t_{\nu i} \mid \mathbf{z}_{\nu i}; \boldsymbol{\theta}_{\nu})$ is defined as $\mathcal{G}_{\nu}[\xi_{\nu i}(t_{\nu i}, \mathbf{z}_{\nu i}; \boldsymbol{\theta}_{\nu})]$, where $\mathcal{G}_{\nu}$ is an inverse link function. The set up of the two $\xi$ predictors is discussed in the detail in the next section. As conveyed by the notation, $\xi_{1i}$ and $\xi_{2i}$ must include baseline functions of time. Different choices for function $\mathcal{G}_{\nu}[\xi_{\nu i}(t_{\nu i}, \mathbf{z}_{\nu i}; \boldsymbol{\theta}_{\nu})]$ can be specified; some common examples are shown in Table 1 reported in Supplementary Material A. The cumulative hazard function, $\mathrm{H}_{\nu}$, and the hazard function, $\mathrm{h}_{\nu}$, are given by

$$\mathrm{H}_{\nu}(t_{\nu i} \mid \mathbf{z}_{\nu i}; \boldsymbol{\theta}_{\nu}) = -\log \mathcal{G}_{\nu}[\xi_{\nu i}(t_{\nu i}, \mathbf{z}_{\nu i}; \boldsymbol{\theta}_{\nu})],$$
$$\mathrm{h}_{\nu}(t_{\nu i} \mid \mathbf{z}_{\nu i}; \boldsymbol{\theta}_{\nu}) = -\frac{\mathcal{G}'_{\nu}[\xi_{\nu i}(t_{\nu i}, \mathbf{z}_{\nu i}; \boldsymbol{\theta}_{\nu})]}{\mathcal{G}_{\nu}[\xi_{\nu i}(t_{\nu i}, \mathbf{z}_{\nu i}; \boldsymbol{\theta}_{\nu})]} \frac{\partial \xi_{\nu i}(t_{\nu i}, \mathbf{z}_{\nu i}; \boldsymbol{\theta}_{\nu})}{\partial t_{\nu i}}, \quad (2)$$

where $\mathcal{G}'_{\nu}[\xi_{\nu i}(t_{\nu i}, \mathbf{z}_{\nu i}; \boldsymbol{\theta}_{\nu})] = \partial \mathcal{G}_{\nu}[\xi_{\nu i}(t_{\nu i}, \mathbf{z}_{\nu i}; \boldsymbol{\theta}_{\nu})] / \partial \xi_{\nu i}(t_{\nu i}, \mathbf{z}_{\nu i}; \boldsymbol{\theta}_{\nu})$.

## 2.2 Additive predictors

This section provides some details on the set up of the two model's predictors for the cases of informative and non-informative censoring. Note that these must include baseline functions of time. To make the presentation simpler, the same design matrix is set up for the two additive predictors. Also, $t_{\nu i}$ can be treated like a covariate. The main advantages of using additive predictors are that various types of covariate effects can be dealt with and that such effects can be flexibly determined without making strong parametric a priori assumptions regarding their forms (e.g., Wood, 2017).

Let us consider a generic predictor $\xi_{\nu i} \in \mathbb{R}$ (where the dependence on the covariates and parameters is momentarily dropped), and the overall baseline covariate vector $\mathbf{x}_{\nu i}$, which contains $\mathbf{z}_{\nu i}$ and $t_{\nu i}$. The additive predictors for the censoring and event times can be defined generically as

$$\xi_{\nu i} = \alpha_{\nu 0} + \sum_{k_{\nu}=0}^{K_{\nu}} s_{\nu k_{\nu}}(\mathbf{x}_{\nu k_{\nu} i}), i = 1, \ldots, n. \quad (3)$$

In (3), $\alpha_{\nu 0} \in \mathbb{R}$ is an overall intercept, $\mathbf{x}_{\nu k_{\nu} i}$ denotes the $k_{\nu}^{th}$ sub-vector of the complete vector $\mathbf{x}_{\nu i}$ and the $K_{\nu}$ functions $s_{\nu k_{\nu}}(\mathbf{x}_{\nu k_{\nu} i})$ represent generic effects which are chosen according to the type of covariate(s) considered. Note that, in (3), $k_{\nu}$ starts from 0 since the summation also includes a smooth function of time.

If censoring is informative, some covariates in $\mathbf{x}_{1i}$ must also appear in $\mathbf{x}_{2i}$. In particular, let us define the vectors of informative and non-informative covariates of dimensions $Q$ and $Q_\nu$ as $\mathbf{x}_i^{0\top} = (x_{1i}^0, \ldots, x_{Qi}^0)$ and $\mathbf{x}_{\nu i}^{1\top} = (x_{\nu 1i}^1, \ldots, x_{\nu Q_\nu i}^1)$, where $K_\nu = Q + Q_\nu$. Informative censoring implies that some components of $\sum_{k_1=1}^{K_1} s_{1k_1}(\mathbf{x}_{1k_1 i})$ must appear in $\sum_{k_2=1}^{K_2} s_{2k_2}(\mathbf{x}_{2k_2 i})$. Without loss of generality, we assume that the first $Q$ components in $\sum_{k_1=1}^{K_1} s_{1k_1}(\mathbf{x}_{1k_1 i})$ appear in $\sum_{k_2=1}^{K_2} s_{2k_2}(\mathbf{x}_{2k_2 i})$. That is,

$$\begin{aligned}\sum_{k_1=1}^{K_1} s_{1k_1}(\mathbf{x}_{1k_1 i}) &= \sum_{q_0=1}^{Q} s_q(\mathbf{x}_{qi}^0) + \sum_{q_1=1}^{Q_1} s_{1q_1}(\mathbf{x}_{1q_1 i}^1) \\ \sum_{k_2=1}^{K_2} s_{2k_2}(\mathbf{x}_{2k_2 i}) &= \sum_{q=1}^{Q} s_q(\mathbf{x}_{qi}^0) + \sum_{q_2=1}^{Q_2} s_{2q_2}(\mathbf{x}_{2q_2 i}^1)\end{aligned}. \quad (4)$$

Therefore, using (4), equation (3) becomes

$$\xi_{\nu i} = \alpha_{\nu 0} + \sum_{q=1}^{Q} s_q(\mathbf{x}_{qi}^0) + \sum_{q_\nu=0}^{Q_\nu} s_{\nu q_\nu}(\mathbf{x}_{\nu q_\nu i}^1), \quad (5)$$

where $\mathbf{x}_{qi}^0$ and $\mathbf{x}_{\nu q_\nu i}^1$ denote the informative and non-informative sub-vectors of the complete vectors $\mathbf{x}_i^0$ and $\mathbf{x}_{\nu i}^1$ respectively, and $s_{\nu q_\nu}(\mathbf{x}_{\nu q_\nu i}^1) = s_{\nu 0}(t_{\nu i})$ when $q_\nu = 0$
.

In (5), the smooth functions are represented using the penalised regression spline approach (e.g., Wood, 2017). Specifically, each $s_{\nu q_\nu}(\mathbf{x}_{\nu q_\nu i}^1)$ can be approximated as a linear combination of $J_{\nu q_\nu}$ non-informative basis functions $\mathcal{Q}_{\nu q_\nu j_{\nu q_\nu}}(\mathbf{x}_{\nu q_\nu i}^1)$ and regression coefficients $\alpha_{\nu q_\nu j_{\nu q_\nu}} \in \mathbb{R}$. In a similar manner, each $s_q(\mathbf{x}_{qi}^0)$ can be approximated as a linear combination of $J_q$ informative basis functions $\mathcal{Q}_{qj_q}(\mathbf{x}_{qi}^0)$ and regression coefficients $\alpha_{0qj_q} \in \mathbb{R}$. More specifically, $s_q(\mathbf{x}_{qi}^0)$

and $s_{\nu q_\nu}(\mathbf{x}^1_{\nu q_\nu i})$ are given by $s_q(\mathbf{x}^0_{qi}) = \sum_{j_q=1}^{J_q} \alpha_{0qj_q} \mathcal{Q}_{qj_q}(\mathbf{x}^0_{qi})$ and $s_{\nu q_\nu}(\mathbf{x}^1_{\nu q_\nu i}) = \sum_{j_{\nu q_\nu}=1}^{J_{\nu q_\nu}} \alpha_{\nu q_\nu j_{\nu q_\nu}} \mathcal{Q}_{\nu q_\nu j_{\nu q_\nu}}(\mathbf{x}^1_{\nu q_\nu i})$, and therefore (5) can be written as

$$\xi_{\nu i} = \alpha_{\nu 0} + \sum_{q=1}^{Q} \boldsymbol{\mathcal{Q}}_q(\mathbf{x}^0_{qi})^\top \boldsymbol{\alpha}_{0q} + \sum_{q_\nu=0}^{Q_\nu} \boldsymbol{\mathcal{Q}}_{\nu q_\nu}(\mathbf{x}^1_{\nu q_\nu i})^\top \boldsymbol{\alpha}_{\nu q_\nu}, \qquad (6)$$

where $\boldsymbol{\mathcal{Q}}_q(\mathbf{x}^0_{qi})^\top \boldsymbol{\alpha}_{0q} = \sum_{j_q=1}^{J_q} \alpha_{0qj_q} \mathcal{Q}_{qj_q}(\mathbf{x}^0_{qi})$, $\boldsymbol{\mathcal{Q}}_{\nu q_\nu}(\mathbf{x}^1_{\nu q_\nu i})^\top \boldsymbol{\alpha}_{\nu q_\nu} = \sum_{j_{\nu q_\nu}=1}^{J_{\nu q_\nu}} \alpha_{\nu q_\nu j_{\nu q_\nu}} \mathcal{Q}_{\nu q_\nu j_{\nu q_\nu}}(\mathbf{x}^1_{\nu q_\nu i})$, $\boldsymbol{\mathcal{Q}}_q(\mathbf{x}^0_{qi}) = \{\mathcal{Q}_{q1}(\mathbf{x}^0_{qi}), \dots, \mathcal{Q}_{qJ_q}(\mathbf{x}^0_{qi})\}^\top$, $\boldsymbol{\mathcal{Q}}_{\nu q_\nu}(\mathbf{x}^1_{\nu q_\nu i}) = \{\mathcal{Q}_{\nu q_\nu 1}(\mathbf{x}^1_{\nu q_\nu i}), \dots, \mathcal{Q}_\nu$ and $\boldsymbol{\alpha}_{\nu q_\nu} = (\alpha_{\nu q_\nu 1}, \dots, \alpha_{\nu q_\nu J_{\nu q_\nu}})^\top$. To write equation (6) in a more compact way, we define $\boldsymbol{\mathcal{Q}}_i^{0\top} \boldsymbol{\alpha}_0 = \sum_{q=1}^{Q} \boldsymbol{\mathcal{Q}}_q(\mathbf{x}^0_{qi})^\top \boldsymbol{\alpha}_{0q}$ and $\boldsymbol{\mathcal{Q}}_{\nu i}^{1\top} \boldsymbol{\alpha}_\nu = \sum_{q_\nu=0}^{Q_\nu} \boldsymbol{\mathcal{Q}}_{\nu q_\nu}(\mathbf{x}^1_{\nu q_\nu i})^\top \boldsymbol{\alpha}_{\nu q_\nu}$, where $\boldsymbol{\alpha}_0 = (\boldsymbol{\alpha}_{01}, \dots, \boldsymbol{\alpha}_{0Q})^\top$, $\boldsymbol{\alpha}_\nu = (\alpha_{\nu 0}, \boldsymbol{\alpha}_{\nu 0}, \dots, \boldsymbol{\alpha}_{\nu Q_\nu})^\top$, $\boldsymbol{\mathcal{Q}}_i^0 = \{\boldsymbol{\mathcal{Q}}_1(\mathbf{x}^0_{1i})^\top, \dots, \boldsymbol{\mathcal{Q}}_Q(\mathbf{x}^0_{Qi})^\top\}^\top$ and $\boldsymbol{\mathcal{Q}}_{\nu i}^1 = \{1, \boldsymbol{\mathcal{Q}}_{\nu 0}(\mathbf{x}^1_{\nu 0 i})^\top, \dots, \boldsymbol{\mathcal{Q}}_{\nu Q_\nu}(\mathbf{x}^1_{\nu Q_\nu})^\top\}^\top$. Therefore,

$$\xi_{\nu i} = \boldsymbol{\mathcal{Q}}_i^{0\top} \boldsymbol{\alpha}_0 + \boldsymbol{\mathcal{Q}}_{\nu i}^{1\top} \boldsymbol{\alpha}_\nu. \quad (7)$$

If $Q > 0$ then censoring is informative and $\sum_{q=1}^{Q} s_q(\mathbf{x}^0_{qi})$ can be estimated using the information from both the censoring and event times. If $Q = 0$ (i.e., the components in $\sum_{k_1=1}^{K_1} s_{1k_1}(\mathbf{x}_{1k_1 i})$ and $\sum_{k_2=1}^{K_2} s_{2k_2}(\mathbf{x}_{2k_2 i})$ are assumed all distinct) then (6) reduces to the model with non-informative censoring and hence we would have

$$\xi_{\nu i} = \gamma_{\nu 0} + \sum_{k_\nu=0}^{K_\nu} \boldsymbol{\mathcal{Q}}_{\nu k_\nu}(\mathbf{x}_{\nu k_\nu i})^\top \boldsymbol{\gamma}_{\nu k_\nu},$$

where $\mathcal{Q}_{\nu k_\nu}(\mathbf{x}_{\nu k_\nu i}) = \{\mathcal{Q}_{\nu k_\nu 1}(\mathbf{x}_{\nu k_\nu i}), \ldots, \mathcal{Q}_{\nu k_\nu J_{\nu k_\nu}}(\mathbf{x}_{\nu k_\nu i})\}^\top$ and $\boldsymbol{\gamma}_{\nu k_\nu} = (\gamma_{\nu k_\nu 1}, \ldots, \gamma_{\nu k_\nu J_{\nu k_\nu}})^\top$. Furthermore, if $\mathcal{Q}_{\nu i}^\top \boldsymbol{\gamma}_\nu = \sum_{k_\nu=0}^{K_\nu} \mathcal{Q}_{\nu k_\nu}(\mathbf{x}_{\nu k_\nu i})^\top \boldsymbol{\gamma}_{\nu k_\nu}$, $\boldsymbol{\gamma}_\nu = (\gamma_{\nu 0}, \boldsymbol{\gamma}_{\nu 0}, \ldots, \boldsymbol{\gamma}_{\nu K_\nu})^\top$ and $\mathcal{Q}_{\nu i} = \{1, \mathcal{Q}_{\nu 0}(\mathbf{x}_{\nu 0 i})^\top, \ldots, \mathcal{Q}_{\nu K_\nu}(\mathbf{x}_{\nu K_\nu})^\top\}^\top$, we obtain

$$\xi_{\nu i} = \mathcal{Q}_{\nu i}^\top \boldsymbol{\gamma}_\nu. \quad (8)$$

Note that, for the case in which $Q = 0$, we have introduced the new parameter vector $\boldsymbol{\gamma}_\nu$ to stress the difference between the parameters of the informative and non-informative models. Some methods for determining the value of $Q$ are discussed in Supplementary Material F.

The vectors of parameters $\boldsymbol{\alpha}_{0q}$ and $\boldsymbol{\alpha}_{\nu q_\nu}$ have associated quadratic penalties $\lambda_q \boldsymbol{\alpha}_{0q}^\top \mathcal{D}_q^0 \boldsymbol{\alpha}_{0q}$ and $\lambda_{\nu q_\nu} \boldsymbol{\alpha}_{\nu q_\nu}^\top \mathcal{D}_{\nu q_\nu}^1 \boldsymbol{\alpha}_{\nu q_\nu}$ used in fitting, whose role is to enforce specific properties on the respective functions, such as smoothness. It is important to note that $\mathcal{D}_q^0$ and $\mathcal{D}_{\nu q_\nu}^1$ only depend on the choice of the basis functions. Smoothing parameter $\lambda_{\nu k_\nu} \in [0, \infty)$ controls the trade-off between fit and smoothness, and plays a crucial role in determining the shape of $\hat{s}_{\nu k_\nu}(\mathbf{x}_{\nu k_\nu i})$. The overall penalty can be defined as $\boldsymbol{\alpha}_\nu^\top \mathcal{D}_\nu \boldsymbol{\alpha}_\nu$, where $\mathcal{D}_\nu = \mathrm{diag}(\lambda_1 \mathcal{D}_1^0, \ldots, \lambda_Q \mathcal{D}_Q^0, 0, \lambda_{\nu 0} \mathcal{D}_{\nu 0}^1, \ldots, \lambda_{\nu Q_\nu} \mathcal{D}_{\nu Q_\nu}^1)$. Moreover, smooth functions are typically subject to centering (identifiability) constraints. The set up described above allows for several types of covariate effects such as linear, non-linear, spatial, random and functional effects, to name but a few. We refer the reader to Wood (2017) for the exact definitions of the spline bases and penalties of the above mentioned cases.

To give a concrete example, consider the informative additive model

$$g_\nu\{S_\nu(t_{\nu i} \mid \mathbf{z}_i^0, \mathbf{z}_{\nu i}^1)\} = g_\nu\{S_{\nu 0}(t_{\nu i})\} + \sum_{q=1}^{Q} \mathcal{Q}_q(\mathbf{z}_{qi}^0)^\top \boldsymbol{\alpha}_{0q} + \sum_{q_\nu=1}^{Q_\nu} \mathcal{Q}_{\nu q_\nu}(\mathbf{z}_{\nu q_\nu i}^1)^\top \boldsymbol{\alpha}_{\nu q_\nu}, \quad (9)$$

where $g_\nu : (0,1) \to (-\infty, \infty)$ is a differentiable and invertible link function (see Table 1 in Supplementary Material A), $S_{\nu 0}(t_{\nu i})$ is a baseline survival function, and $g_\nu\{S_{\nu 0}(t_{\nu i})\}$ is represented using a smooth function of time, $s_{\nu 0}(t_{\nu i})$. When the log-log link is chosen, equation (9) yields the proportional hazards model

$$\log\{\mathrm{H}_\nu(t_{\nu i} \mid \mathbf{z}_i^0, \mathbf{z}_{\nu i}^1)\} = \log\{\mathrm{H}_{\nu 0}(t_{\nu i})\} + \sum_{q=1}^{Q} \mathcal{Q}_q(\mathbf{z}_{qi}^0)^\top \boldsymbol{\alpha}_{0q} + \sum_{q_\nu=1}^{Q_\nu} \mathcal{Q}_{\nu q_\nu}(\mathbf{z}_{\nu q_\nu i}^1)^\top \boldsymbol{\alpha}_{\nu q_\nu},$$

where $\mathrm{H}_\nu(t_{\nu i} \mid \mathbf{z}_i^0, \mathbf{z}_{\nu i}^1) = -\log\{S_\nu(t_{\nu i} \mid \mathbf{z}_i^0, \mathbf{z}_{\nu i}^1)\}$ and $\log\{\mathrm{H}_{\nu 0}(t_{\nu i})\} = -\log\{S_{\nu 0}(t_{\nu i})\}$ is the cumulative baseline hazard function. Analogously, equation (9) yields the proportional odds model when the -logit link is chosen.

The models considered in this paper are fundamentally parametric but flexible. It is worth noting that the more extensive use of parametric survival models in applications has been encouraged by Cox; see the discussion in Reid (1994). Moreover, as pointed out for instance by Hjort (1992), parametric approaches simplify somewhat model estimation and comparison, easily allow for the visualization of the estimated baseline hazard and survival functions, and allow to calculate several quantities of interest and their variances which would otherwise be difficult to obtain with a non-parametric approach. Another important advantage is that there is no necessity to use numerical integration methods to estimate the cumulative hazard function.

## 2.3 Estimation framework

The data consist of $\{Y_i, \delta_{1i}, \mathbf{z}_i\}$, where $Y_i = \min\{T_{1i}, T_{2i}\}$ and $\delta_{1i} = I(T_{1i} \le T_{2i})$, for $i = 1, \ldots, n$. Let $f(t_1, t_2 \mid \mathbf{z})$ be the conditional joint distribution of $(T_1, T_2)$ given **z**. We can write $P(Y_i, \delta_{1i} = 1 \mid \mathbf{z}_i) = \int_{y_i}^{\infty} f(y_i, t_2 \mid \mathbf{z}_i) dt_2$ and $P(Y_i, \delta_{1i} = 0 \mid \mathbf{z}_i) = \int_{y_i}^{\infty} f(t_1, y_i \mid \mathbf{z}_i) dt_1$. Therefore, the conditional likelihood function of $(Y_i, \delta_{1i})$ given $\mathbf{z}_i$, for all $i = 1, \ldots, n$, is

$$\mathcal{L}=\prod_{i=1}^{n}\left\{\int_{y_i}^{\infty} f(y_i,t_2\,|\,\mathbf{z}_i)dt_2\right\}^{\delta_{1i}}\left\{\int_{y_i}^{\infty} f(t_1,y_i\,|\,\mathbf{z}_i)dt_1\right\}^{\delta_{2i}}.$$

Below we provide the relevant details for the cases of informative and non-informative censoring, which highlight the differences between the two estimators and that are also required for the theoretical derivations in Section 2.4.

If it is assumed that $T_{1i}$ and $T_{i2}$ are conditionally independent given $\mathbf{z}_i$, then $\int_{y_i}^{\infty} f(y_i,t_2\,|\,\mathbf{z}_i)dt_2 = f_1(y_i\,|\,\mathbf{z}_{1i};\gamma_1)S_2(y_i\,|\,\mathbf{z}_{2i};\gamma_2)$ and $\int_{y_i}^{\infty} f(t_1,y_i\,|\,\mathbf{z}_i)dt_1 = f_2(y_i\,|\,\mathbf{z}_{2i};\gamma_2)S_1(y_i\,|\,\mathbf{z}_{1i};\gamma_1)$ when censoring is non-informative. However, if censoring is informative $\gamma_1$ and $\gamma_2$ would have some components in common. Since it was assumed that the first $Q$ components of $\gamma_1$ are the same as the first $Q$ components of $\gamma_2$, we have $\mathcal{Q}_{\nu i}^{\top}\boldsymbol{\gamma}_{\nu}=\mathcal{Q}_{i}^{0\top}\boldsymbol{\alpha}_0+\mathcal{Q}_{\nu i}^{1\top}\boldsymbol{\alpha}_{\nu}$. Using (1), (2), (7), (8), and $\xi_{\nu i}(\boldsymbol{\gamma}_{\nu})$ and $\xi_{\nu i}(\boldsymbol{\alpha}_0,\boldsymbol{\alpha}_{\nu})$ as the shorthand notation for $\xi_{\nu i}(y_i,\mathbf{z}_{\nu i};\boldsymbol{\gamma}_{\nu})$ and $\xi_{\nu i}(y_i,\mathbf{z}_{\nu i};\boldsymbol{\alpha}_0,\boldsymbol{\alpha}_{\nu})$, the non-informative and informative log-likelihood functions can be written, respectively, as

$$\begin{aligned}
\ell(\boldsymbol{\gamma}) &= \sum_{i=1}^{n}\left\{\log\mathcal{G}_1\left[\xi_{1i}(\boldsymbol{\gamma}_1)\right]+\delta_{1i}\log\left\{-\frac{\mathcal{G}'_1\left[\xi_{1i}(\boldsymbol{\gamma}_1)\right]}{\mathcal{G}_1\left[\xi_{1i}(\boldsymbol{\gamma}_1)\right]}\frac{\partial\xi_{1i}(\boldsymbol{\gamma}_1)}{\partial y_i}\right\}\right\}\\
&+\sum_{i=1}^{n}\left\{\log\mathcal{G}_2\left[\xi_{2i}(\boldsymbol{\gamma}_2)\right]+\delta_{2i}\log\left\{-\frac{\mathcal{G}'_2\left[\xi_{2i}(\boldsymbol{\gamma}_2)\right]}{\mathcal{G}_2\left[\xi_{2i}(\boldsymbol{\gamma}_2)\right]}\frac{\partial\xi_{2i}(\boldsymbol{\gamma}_2)}{\partial y_i}\right\}\right\},\\
\ell(\boldsymbol{\alpha}) &= \sum_{i=1}^{n}\left\{\log\mathcal{G}_1\left[\xi_{1i}(\boldsymbol{\alpha}_0,\boldsymbol{\alpha}_1)\right]+\delta_{1i}\log\left\{-\frac{\mathcal{G}'_1\left[\xi_{1i}(\boldsymbol{\alpha}_0,\boldsymbol{\alpha}_1)\right]}{\mathcal{G}_1\left[\xi_{1i}(\boldsymbol{\alpha}_0,\boldsymbol{\alpha}_1)\right]}\frac{\partial\xi_{1i}(\boldsymbol{\alpha}_0,\boldsymbol{\alpha}_1)}{\partial y_i}\right\}\right\}\\
&+\sum_{i=1}^{n}\left\{\log\mathcal{G}_2\left[\xi_{2i}(\boldsymbol{\alpha}_0,\boldsymbol{\alpha}_2)\right]+\delta_{2i}\log\left\{-\frac{\mathcal{G}'_2\left[\xi_{2i}(\boldsymbol{\alpha}_0,\boldsymbol{\alpha}_2)\right]}{\mathcal{G}_2\left[\xi_{2i}(\boldsymbol{\alpha}_0,\boldsymbol{\alpha}_2)\right]}\frac{\partial\xi_{2i}(\boldsymbol{\alpha}_0,\boldsymbol{\alpha}_2)}{\partial y_i}\right\}\right\}.
\end{aligned}\qquad(10)$$

To ensure that the hazard functions in (10) are positive, $[\partial\xi_{\nu i}(\boldsymbol{\theta}_{\nu})/\partial y_i]$, for $\nu=(1,2)$, must be positive. To this end, we model the time effects using B-splines with coefficients constrained such that the resulting smooth functions of time are monotonically increasing. In particular, we define $s_{\nu 0}(y_i)=\sum_{j_\nu=1}^{J_\nu}\vartheta_{\nu 0 j_{\nu 0}}\mathcal{M}_{\nu 0 j_{\nu 0}}(y_i)$, where the $\mathcal{M}_{\nu 0 j_{\nu 0}}(y_i)$ are B-spline basis functions of at least second order built over the interval $[a,b]$, based on equally spaced knots, and $\vartheta_{\nu 0 j_{\nu 0}}$ are spline coefficients. A sufficient condition for $[\partial s_{\nu 0}(y_i)/\partial y_i]\geq 0$ over $[a,b]$ is that $\vartheta_{\nu 0 j_{\nu 0}}\geq\vartheta_{\nu 0 j-1_{\nu 0}}, \forall j$ (Leitenstorfer & Tutz, 2006). Such condition can be imposed by re-parametrising the spline coefficient vector so that $\vartheta_{\nu 0}=\mathbf{\Gamma}_{\nu 0}\boldsymbol{\beta}_{\nu 0}$, where $\boldsymbol{\beta}_{\nu 0}^{\top}=(\beta_{\nu 01},\beta_{\nu 02},\ldots,\beta_{\nu 0 J_{\nu 0}})$, $\boldsymbol{\beta}_{\nu}^{\top}=(\beta_{\nu 01},\exp(\beta_{\nu 02}),\ldots,\exp(\beta_{\nu 0 J_{\nu 0}}))$ and $\mathbf{\Gamma}_{\nu 0}[\kappa_{\nu 01},\kappa_{\nu 02}]=0$ if $\kappa_{\nu 01}<\kappa_{\nu 02}$, and $\mathbf{\Gamma}_{\nu 0}[\kappa_{\nu 01},\kappa_{\nu 02}]=1$ if $\kappa_{\nu 01}\geq\kappa_{\nu 02}$. Following Pya & Wood (2015, Section 2.2.1), the penalty term is set up to penalise the squared differences between adjacent $\beta_{\nu 0 j_{\nu 0}}$, starting from $\beta_{\nu 02}$, using $\boldsymbol{\mathcal{D}}_{\nu 0}=\boldsymbol{\mathcal{D}}_{\nu 0}^{\odot\top}\boldsymbol{\mathcal{D}}_{\nu 0}^{\odot}$, where $\boldsymbol{\mathcal{D}}_{\nu 0}^{\odot}$ is a $(J_{\nu 0}-2)\times J_{\nu 0}$ matrix made up of zeros except that $\boldsymbol{\mathcal{D}}_{\nu 0}^{\odot}[\kappa_{\nu 0},\kappa_{\nu 0}+1]=-\boldsymbol{\mathcal{D}}_{\nu 0}^{\odot}[\kappa_{\nu 0},\kappa_{\nu 0}+2]=1$ for $\kappa_{\nu 0}=1,\ldots,J_{\nu 0-2}$. Therefore, the non-informative and informative additive predictors, that ensure positive hazard functions in (10), are

$$\begin{aligned}\xi_{\nu i}&=\gamma_{\nu 0}+\boldsymbol{\mathcal{Q}}_{\nu 0}(y_i)^{\top}\mathbf{\Gamma}_{\nu 0}\tilde{\boldsymbol{\gamma}}_{\nu 0}+\sum_{k_\nu=1}^{K_\nu}\boldsymbol{\mathcal{Q}}_{\nu k_\nu}(\mathbf{x}_{\nu k_\nu i})^{\top}\boldsymbol{\gamma}_{\nu k_\nu},\\ \xi_{\nu i}&=\alpha_{\nu 0}+\boldsymbol{\mathcal{Q}}_{\nu 0}(y_i)^{\top}\mathbf{\Gamma}_{\nu 0}\boldsymbol{\alpha}_{\nu 0}+\sum_{q=1}^{Q}\boldsymbol{\mathcal{Q}}_{q}(\mathbf{x}_{qi}^{0})^{\top}\boldsymbol{\alpha}_{0q}+\sum_{q_\nu=1}^{Q_\nu}\boldsymbol{\mathcal{Q}}_{\nu q_\nu}(\mathbf{x}_{\nu q_\nu i}^{1})^{\top}\boldsymbol{\alpha}_{\nu q_\nu}.\end{aligned}\qquad(11)$$

Our model specification allows for a high degree of flexibility in modelling survival data. If an unpenalised estimation approach is employed to estimate $\boldsymbol{\gamma}=(\boldsymbol{\gamma}_1^{\top},\boldsymbol{\gamma}_2^{\top})^{\top}$ and $\boldsymbol{\alpha}=(\boldsymbol{\alpha}_0^{\top},\boldsymbol{\alpha}_1^{\top},\boldsymbol{\alpha}_2^{\top})^{\top}$, then the resulting smooth function estimates are likely to be unduly wiggly (e.g., Wood, 2017). Therefore, to prevent over-fitting, the following functions are maximized

$$\ell_p(\boldsymbol{\gamma})=\ell(\boldsymbol{\gamma})-\frac{1}{2}\boldsymbol{\gamma}^{\top}\boldsymbol{\mathcal{S}}\boldsymbol{\gamma},\qquad(12)$$

$$\ell_p(\boldsymbol{\alpha}) = \ell(\boldsymbol{\alpha}) - \frac{1}{2}\boldsymbol{\alpha}^\top \boldsymbol{\mathcal{S}}\boldsymbol{\alpha}, \qquad (13)$$

where $\ell_p(\boldsymbol{\gamma})$ and $\ell_p(\boldsymbol{\alpha})$ are the non-informative and informative penalized log-likelihoods. Moreover, $\boldsymbol{\mathcal{S}} = \mathrm{diag}(\boldsymbol{\mathcal{D}}_1, \boldsymbol{\mathcal{D}}_2)$, and $\boldsymbol{\mathcal{D}}_1$ and $\boldsymbol{\mathcal{D}}_2$ are overall penalties which contain $\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2$. The smoothing parameter vectors can be collected in the overall vector $\boldsymbol{\lambda} = (\boldsymbol{\lambda}_1^\top, \boldsymbol{\lambda}_2^\top)^\top$. Estimation of the models' parameters and smoothing coefficients is achieved by using a stable and efficient trust region algorithm with integrated automatic multiple smoothing parameter selection (see Supplementary Material C for details). This required working with first and second order analytical derivatives which have been tediously derived as well as verified using numerical derivatives. Their structures are shown below. Note that these results were also required for the theoretical proofs presented in Section 2.4.

When censoring is non-informative, the gradient of (12) can be obtained as

$$\nabla_{\boldsymbol{\gamma}} \ell_p(\boldsymbol{\gamma}) = \nabla_{\boldsymbol{\gamma}} \ell(\boldsymbol{\gamma}) - \boldsymbol{\gamma}\boldsymbol{\mathcal{S}},$$

where $\nabla_{\boldsymbol{\gamma}} \ell(\boldsymbol{\gamma}) = \left(\nabla_{\boldsymbol{\gamma}_1} \ell(\boldsymbol{\gamma})^\top, \nabla_{\boldsymbol{\gamma}_2} \ell(\boldsymbol{\gamma})^\top\right)^\top$. The components of $\nabla_{\boldsymbol{\gamma}_\nu} \ell(\boldsymbol{\gamma})$ can generically be calculated using the following expression

$$\nabla_{\boldsymbol{\gamma}_{\nu k_\nu}} \ell(\boldsymbol{\gamma}) = \begin{cases} \sum_{i=1}^{n} \left[\Delta_{\nu i} \boldsymbol{\mathcal{Q}}_{\nu 0}^{\vartriangle}(y_i) + \Omega_{\nu i} \boldsymbol{\mathcal{Q}}_{\nu 0}^{\vartriangle'}(y_i)\right] & \text{if } \boldsymbol{\gamma}_{\nu k_\nu} = \boldsymbol{\gamma}_{\nu 0}, \\ \sum_{i=1}^{n} \left[\Delta_{\nu i} \boldsymbol{\mathcal{Q}}_{\nu k_\nu}(\mathbf{x}_{\nu k_\nu i})\right] & \text{otherwise.} \end{cases} \qquad (14)$$

In (14), $\boldsymbol{\mathcal{Q}}_{\nu 0}^{\vartriangle}(y_i)$ and $\boldsymbol{\mathcal{Q}}_{\nu 0}^{\vartriangle'}(y_i)$ are design vectors. Furthermore, $\Omega_{\nu i} = \delta_{\nu i}\left(\frac{\partial \xi_{\nu i}}{\partial y_i}\right)^{-1}$ and $\Delta_{\nu i} = \left[\frac{\mathcal{G}'_\nu}{\mathcal{G}_\nu} + \delta_{\nu i}\left(\frac{\mathcal{G}''_\nu}{\mathcal{G}'_\nu} - \frac{\mathcal{G}'_\nu}{\mathcal{G}_\nu}\right)\right]$, for all $\nu = 1, 2$. The non-informative penalized Hessian can be calculated as

$$\nabla_{\boldsymbol{\gamma}\boldsymbol{\gamma}} \ell_p(\boldsymbol{\gamma}) = \nabla_{\boldsymbol{\gamma}\boldsymbol{\gamma}} \ell(\boldsymbol{\gamma}) - \boldsymbol{\mathcal{S}},$$

where

$$\nabla_{\gamma\gamma}\ell(\gamma) = \begin{bmatrix} \nabla_{\gamma_1\gamma_1}\ell(\gamma) & \mathbf{0} \\ \mathbf{0} & \nabla_{\gamma_2\gamma_2}\ell(\gamma) \end{bmatrix}.$$

Further, the elements of $\nabla_{\gamma_\nu\gamma_\nu}\ell(\gamma)$ are calculated using

$$\begin{aligned}
\nabla_{\gamma_{\nu k_\nu}\gamma_{\nu 0}}\ell(\gamma) &= \sum_{i=1}^{n}\left[\mathcal{Q}_{\nu k_\nu}(\mathbf{x}_{\nu k_\nu i})\Phi_{\nu i}\mathcal{Q}_{\nu 0}^{\vartriangle}(y_i)^\top\right], \\
\nabla_{\gamma_{\nu 0}\gamma_{\nu s_\nu}}\ell(\gamma) &= \sum_{i=1}^{n}\left[\mathcal{Q}_{\nu 0}^{\vartriangle}(y_i)\Phi_{\nu i}\mathcal{Q}_{\nu s_\nu}(\mathbf{x}_{\nu s_\nu i})^\top\right], \\
\nabla_{\gamma_{\nu k_\nu}\gamma_{\nu s_\nu}}\ell(\gamma) &= \sum_{i=1}^{n}\left[\mathcal{Q}_{\nu k_\nu}(\mathbf{x}_{\nu k_\nu i})\Phi_{\nu i}\mathcal{Q}_{\nu s_\nu}(\mathbf{x}_{\nu s_\nu i})^\top\right], \\
\nabla_{\gamma_{\nu 0}\gamma_{\nu 0}}\ell(\gamma) &= \sum_{i=1}^{n}\left[\mathcal{Q}_{\nu 0}^{\vartriangle}(y_i)\Phi_{\nu i}\mathcal{Q}_{\nu 0}^{\vartriangle}(y_i)^\top + \Delta_{\nu i}\mathcal{Q}_{\nu 0}^{\vartriangle\vartriangle}(y_i) - \mathcal{Q}_{\nu 0}^{\vartriangle'}(y_i)\Psi_{\nu i}\mathcal{Q}_{\nu 0}^{\vartriangle'}(y_i)^\top + \Omega_{\nu i}\mathcal{Q}_{\nu 0}^{\vartriangle\vartriangle'}(y_i)\right].
\end{aligned} \tag{15}$$

In these sub-matrices $\Phi_{\nu i} = \delta_{\nu i}\left(\frac{\mathcal{G}'''_\nu}{\mathcal{G}_\nu} - \frac{\mathcal{G}''^2_\nu}{\mathcal{G}'^2_\nu} - \frac{\mathcal{G}''_\nu}{\mathcal{G}_\nu} + \frac{\mathcal{G}'^2_\nu}{\mathcal{G}^2_\nu}\right)$ and $\Psi_{\nu i} = \left[\delta_{\nu i}\left(\frac{\partial \xi_{\nu i}}{\partial y_i}\right)^{-2}\right]$. In addition, $\mathcal{Q}_{\nu 0}^{\vartriangle\vartriangle}(y_i)$ and $\mathcal{Q}_{\nu 0}^{\vartriangle\vartriangle'}(y_i)$ are design diagonal matrices.

If the censoring is informative, the gradient of (13) can be calculated as

$$\nabla_{\alpha}\ell_p(\alpha) = \nabla_{\alpha}\ell(\alpha) - \alpha\mathcal{S},$$

where $\nabla_{\alpha}\ell(\alpha) = \left(\nabla_{\alpha_0}\ell(\alpha)^\top, \nabla_{\alpha_1}\ell(\alpha)^\top, \nabla_{\alpha_2}\ell(\alpha)^\top\right)^\top$. To obtain $\nabla_{\alpha_0}\ell(\alpha)$ and $\nabla_{\alpha_\nu}\ell(\alpha)$, we use

$$\begin{aligned}
\nabla_{\alpha_0}\ell(\alpha) &= \sum_{i=1}^{n}\left[\mathcal{Q}_i^0\left(\Delta_{1i} + \Delta_{2i}\right)\right], \\
\nabla_{\alpha_{\nu k_\nu}}\ell(\alpha) &= \begin{cases} \sum_{i=1}^{n}\left[\Delta_{\nu i}\mathcal{Q}_{\nu 0}^{\iota\vartriangle}(y_i) + \Omega_{\nu i}\mathcal{Q}_{\nu 0}^{\iota\vartriangle'}(y_i)\right] & \text{if } \alpha_{\nu k_\nu} = \alpha_{\nu 0}, \\ \sum_{i=1}^{n}\left[\Delta_{\nu i}\mathcal{Q}_{\nu k_\nu}(\mathbf{x}_{\nu k_\nu i})\right] & \text{otherwise,} \end{cases}
\end{aligned} \tag{16}$$

where $\mathcal{Q}_{\nu 0}^{\iota\triangle}(y_i)$ and $\mathcal{Q}_{\nu 0}^{\iota\triangle\prime}(y_i)$ are design vectors. The informative penalized Hessian can be obtained as follow

$$\nabla_{\boldsymbol{\alpha}\boldsymbol{\alpha}}\ell_p(\boldsymbol{\alpha}) = \nabla_{\boldsymbol{\alpha}\boldsymbol{\alpha}}\ell(\boldsymbol{\alpha}) - \boldsymbol{\mathcal{S}},$$

where

$$\nabla_{\boldsymbol{\alpha}\boldsymbol{\alpha}}\ell(\boldsymbol{\alpha}) = \begin{bmatrix} \nabla_{\boldsymbol{\alpha}_0\boldsymbol{\alpha}_0}\ell(\boldsymbol{\alpha}) & \nabla_{\boldsymbol{\alpha}_0\boldsymbol{\alpha}_1}\ell(\boldsymbol{\alpha}) & \nabla_{\boldsymbol{\alpha}_0\boldsymbol{\alpha}_2}\ell(\boldsymbol{\alpha}) \\ \nabla_{\boldsymbol{\alpha}_1\boldsymbol{\alpha}_0}\ell(\boldsymbol{\alpha}) & \nabla_{\boldsymbol{\alpha}_1\boldsymbol{\alpha}_1}\ell(\boldsymbol{\alpha}) & \mathbf{0} \\ \nabla_{\boldsymbol{\alpha}_2\boldsymbol{\alpha}_0}\ell(\boldsymbol{\alpha}) & \mathbf{0} & \nabla_{\boldsymbol{\alpha}_2\boldsymbol{\alpha}_2}\ell(\boldsymbol{\alpha}) \end{bmatrix}.$$

Furthermore, $\nabla_{\boldsymbol{\alpha}_0\boldsymbol{\alpha}_0}\ell(\boldsymbol{\alpha})$ and the components of $\nabla_{\boldsymbol{\alpha}_\nu\boldsymbol{\alpha}_0}\ell(\boldsymbol{\alpha})$ and $\nabla_{\boldsymbol{\alpha}_0\boldsymbol{\alpha}_\nu}\ell(\boldsymbol{\alpha})$ are obtained using

$$\begin{aligned} \nabla_{\boldsymbol{\alpha}_0\boldsymbol{\alpha}_0}\ell(\boldsymbol{\alpha}) &= \sum_{i=1}^{n}\left[\mathcal{Q}_i^0(\Phi_{1i}+\Phi_{2i})\mathcal{Q}_i^{0\top}\right], \\ \nabla_{\boldsymbol{\alpha}_0\boldsymbol{\alpha}_{\nu q_\nu}}\ell(\boldsymbol{\alpha}) &= \begin{cases} \sum_{i=1}^{n}\left[\mathcal{Q}_i^0\Phi_{\nu i}\mathcal{Q}_{\nu 0}^{\iota\triangle}(y_i)^\top\right] & \text{if } \boldsymbol{\alpha}_{\nu q_\nu} = \boldsymbol{\alpha}_{\nu 0}, \\ \sum_{i=1}^{n}\left[\mathcal{Q}_i^0\Phi_{\nu i}\mathcal{Q}_{\nu q_\nu}(\mathbf{x}_{\nu q_\nu i}^1)^\top\right] & \text{otherwise,} \end{cases} \\ \nabla_{\boldsymbol{\alpha}_{\nu q_\nu}\boldsymbol{\alpha}_0}\ell(\boldsymbol{\alpha}) &= \begin{cases} \sum_{i=1}^{n}\left[\mathcal{Q}_{\nu 0}^{\iota\triangle}(y_i)\Phi_{\nu i}\mathcal{Q}_i^{0\top}\right] & \text{if } \boldsymbol{\alpha}_{\nu q_\nu} = \boldsymbol{\alpha}_{\nu 0}, \\ \sum_{i=1}^{n}\left[\mathcal{Q}_{\nu q_\nu}(\mathbf{x}_{\nu q_\nu i}^1)\Phi_{\nu i}\mathcal{Q}_i^{0\top}\right] & \text{otherwise.} \end{cases} \end{aligned} \qquad (17)$$

Finally, the elements of $\nabla_{\boldsymbol{\alpha}_\nu\boldsymbol{\alpha}_\nu}\ell(\boldsymbol{\alpha})$ are calculated using

$$
\begin{aligned}
\nabla_{\boldsymbol{\alpha}_{\nu q_\nu}\boldsymbol{\alpha}_{\nu 0}}\ell(\boldsymbol{\alpha}) &= \sum_{i=1}^{n}\left[\mathcal{Q}_{\nu q_\nu}(\mathbf{x}^1_{\nu q_\nu i})\Phi_{\nu i}\mathcal{Q}^{\iota\triangle}_{\nu 0}(y_i)^\top\right],\\
\nabla_{\boldsymbol{\alpha}_{\nu 0}\boldsymbol{\alpha}_{\nu q_\nu}}\ell(\boldsymbol{\alpha}) &= \sum_{i=1}^{n}\left[\mathcal{Q}^{\iota\triangle}_{\nu 0}(y_i)\Phi_{\nu i}\mathcal{Q}_{\nu q_\nu}(\mathbf{x}^1_{\nu q_\nu i})^\top\right],\\
\nabla_{\boldsymbol{\alpha}_{\nu q_\nu}\boldsymbol{\alpha}_{\nu r_\nu}}\ell(\boldsymbol{\alpha}) &= \sum_{i=1}^{n}\left[\mathcal{Q}_{\nu q_\nu}(\mathbf{x}^1_{\nu q_\nu i})\Phi_{\nu i}\mathcal{Q}_{\nu r_\nu}(\mathbf{x}^1_{\nu r_\nu i})^\top\right],\\
\nabla_{\boldsymbol{\alpha}_{\nu 0}\boldsymbol{\alpha}_{\nu 0}}\ell(\boldsymbol{\alpha}) &= \sum_{i=1}^{n}\left[\mathcal{Q}^{\iota\triangle}_{\nu 0}(y_i)\Phi_{\nu i}\mathcal{Q}^{\iota\triangle}_{\nu 0}(y_i)^\top + \Delta_{\nu i}\mathcal{Q}^{\iota\triangle\triangle}_{\nu 0}(y_i) - \mathcal{Q}^{\iota\triangle\prime}_{\nu 0}(y_i)\Psi_{\nu i}\mathcal{Q}^{\iota\triangle\prime}_{\nu 0}(y_i)^\top + \Omega_{\nu i}\mathcal{Q}^{\iota\triangle\triangle\prime}_{\nu 0}(y_i)\right].
\end{aligned}
\tag{18}
$$

As before, $\mathcal{Q}^{\iota\triangle\triangle}_{\nu 0}(y_i)$ and $\mathcal{Q}^{\iota\triangle\triangle\prime}_{\nu 0}(y_i)$ represent design diagonal matrices.

The derivations of the results reported here as well as some algorithmic details are given in Supplementary Materials B and C.

Remark 1. The scores and Hessian components described in this section have been implemented in a modular way, hence no substantial programming work will be required to incorporate link functions not considered in this article. Furthermore, quantities such as those defined in (14), (15), (16), (17) and (18), are needed for the theoretical proofs of the next section.

## 2.4 Theoretical properties

In this section, we derive the $\sqrt{n}$ consistency and asymptotic normality of the non-informative and informative estimators, and shed light on the efficiency gains produced by the newly introduced informative estimator when compared to its non-informative counterpart. As far as the number of basis functions is concerned, we use the fixed-knot asymptotic framework since it is closer to practical statistical modelling (e.g., Vatter & Chavez-Demoulin, 2015, and references therein). In what follows, we define $\hat{S}_{\nu 0}(\boldsymbol{\theta}_{\nu 0}) = \mathcal{G}_{\nu 0}[s(\boldsymbol{\theta}_{\nu 0})]$ as the short notation for $\hat{S}_{\nu 0}(y_i, \boldsymbol{\theta}_{\nu 0}) = \mathcal{G}_{\nu 0}[s(y_i, \boldsymbol{\theta}_{\nu 0})]$ and $\boldsymbol{\theta}^*$ as the true vector of parameters.

The informative penalized maximum log-likelihood estimator (IPMLE) can be defined as

$$\boldsymbol{\alpha} = \underset{\boldsymbol{\alpha}\in\Theta}{\operatorname{argmax}}\, \ell_p(\boldsymbol{\alpha}),$$

and the non-informative counterpart (NPMLE) as

$$\hat{\boldsymbol{\gamma}} = \underset{\gamma\in\Theta}{\operatorname{argmax}}\, \ell_p(\boldsymbol{\gamma}).$$

Theorem 1 (Asymptotic properties of the IPMLE estimator). If the set of Assumptions **1** and **2** in Supplementary Material D hold then

(i) the informative penalized maximum log-likelihood estimator $\boldsymbol{\alpha}$ exists, is $\sqrt{n}$-consistent and

$$\sqrt{n}(\boldsymbol{\alpha}-\boldsymbol{\alpha}^{*}) \xrightarrow{d} \mathcal{N}\left\{\mathbf{0},[\mathcal{I}(\boldsymbol{\alpha}^{*})]^{-1}\right\},$$

where $\mathcal{I}(\boldsymbol{\alpha}^{*}) = \mathbb{E}[-\nabla_{\boldsymbol{\alpha}\boldsymbol{\alpha}}\ell(\mathbf{w};\boldsymbol{\alpha}^{*})]$.

(ii) $\hat{S}_{10}(\boldsymbol{\alpha}_{10})$ is asymptotically independent of $\hat{S}_{20}(\boldsymbol{\alpha}_{20})$ and

$$\sqrt{n}[\hat{S}_{\nu 0}(\boldsymbol{\alpha}_{\nu 0}) - S_{\nu 0}(\boldsymbol{\alpha}^{*}_{\nu 0})] \xrightarrow{d} \mathcal{N}\left\{\mathbf{0},\boldsymbol{\Sigma}_{\boldsymbol{\alpha}^{*}_{\nu 0}}\right\},\ \nu=1,2,$$

where $\boldsymbol{\Sigma}_{\boldsymbol{\alpha}^{*}_{\nu 0}} = \mathcal{G}'_{\nu 0}[s(\boldsymbol{\alpha}^{*}_{\nu 0})]\nabla_{\boldsymbol{\alpha}_{\nu 0}} s(\boldsymbol{\alpha}^{*}_{\nu 0})[\mathcal{I}(\boldsymbol{\alpha}^{*}_{\nu 0})]^{-1}\nabla_{\boldsymbol{\alpha}_{\nu 0}} s(\boldsymbol{\alpha}^{*}_{\nu 0})^{\top}\mathcal{G}'_{\nu 0}[s(\boldsymbol{\alpha}^{*}_{\nu 0})]$ and $\mathcal{I}(\boldsymbol{\alpha}^{*}_{\nu 0}) = \mathbb{E}[-\nabla_{\boldsymbol{\alpha}_{\nu 0}\boldsymbol{\alpha}_{\nu 0}}\ell(\mathbf{w};\boldsymbol{\alpha}^{*}_{\nu 0})]$.

Theorem 2 (Asymptotic properties of the NPMLE estimator). If the set of Assumptions **1** and **2** in Supplementary Material D hold then

(i) the non-informative penalized maximum log-likelihood estimator $\hat{\gamma}$ exists, is $\sqrt{n}$-consistent and

$$\sqrt{n}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*) \xrightarrow{d} \mathcal{N}\left\{\mathbf{0}, [\mathcal{I}(\boldsymbol{\gamma}^*)]^{-1}\right\},$$

where $\mathcal{I}(\boldsymbol{\gamma}^*) = \mathbb{E}[-\nabla_{\gamma\gamma}\ell(\mathbf{w};\boldsymbol{\gamma}^*)]$.

(ii) $\hat{S}_{10}(\hat{\gamma}_{10})$ is asymptotically independent of $\hat{S}_{20}(\hat{\gamma}_{20})$ and

$$\sqrt{n}[\hat{S}_{\nu 0}(\hat{\gamma}_{\nu 0}) - S_{\nu 0}(\boldsymbol{\gamma}^*_{\nu 0})] \xrightarrow{d} \mathcal{N}\left\{\mathbf{0}, \boldsymbol{\Sigma}_{\gamma^*_{\nu 0}}\right\}, \ \nu = 1, 2,$$

where $\boldsymbol{\Sigma}_{\gamma^*_{\nu 0}} = \mathcal{G}'_{\nu 0}[s(\boldsymbol{\gamma}^*_{\nu 0})]\nabla_{\gamma_{\nu 0}} s(\boldsymbol{\gamma}^*_{\nu 0})[\mathcal{I}(\boldsymbol{\gamma}^*_{\nu 0})]^{-1}\nabla_{\gamma_{\nu 0}} s(\boldsymbol{\gamma}^*_{\nu 0})^\top \mathcal{G}'_{\nu 0}[s(\boldsymbol{\gamma}^*_{\nu 0})]$ and $\mathcal{I}(\boldsymbol{\gamma}^*_{\nu 0}) = \mathbb{E}[-\nabla_{\gamma_{\nu 0}\gamma_{\nu 0}}\ell(\mathbf{w};\boldsymbol{\gamma}^*_{\nu 0})]$.

Theorem 3 (Efficiency of the IPMLE estimator). For $\nu = 1, 2$, let $\boldsymbol{\gamma}_\nu = (\boldsymbol{\gamma}^{\iota}_\nu, \boldsymbol{\gamma}^{n\iota}_\nu)^\top$ be the informative and non-informative parameters of the non-informative model, respectively. Under the set of Assumptions **1** and **2** in Supplementary Material D, and if we further assume that $\boldsymbol{\gamma}^{n\iota}_{\nu 0} = \boldsymbol{\alpha}_{\nu 0}$, then

$$\mathcal{A}\mathbf{Cov}(\boldsymbol{\alpha}_0) < \mathcal{A}\mathbf{Cov}(\hat{\boldsymbol{\gamma}}^{\iota}_\nu),$$
$$\mathcal{A}\mathbf{Cov}(\boldsymbol{\alpha}_\nu) < \mathcal{A}\mathbf{Cov}(\hat{\boldsymbol{\gamma}}^{n\iota}_\nu),$$

where $\mathcal{A}\mathbf{Cov}(\boldsymbol{\alpha}_0) = \Sigma_{\alpha^*_0}$, $\mathcal{A}\mathbf{Cov}(\boldsymbol{\alpha}_\nu) = \Sigma_{\alpha^*_\nu}$, $\mathcal{A}\mathbf{Cov}(\hat{\boldsymbol{\gamma}}^{\iota}_\nu) = \Sigma_{\gamma^{*\iota}_\nu}$, and $\mathcal{A}\mathbf{Cov}(\hat{\boldsymbol{\gamma}}^{n\iota}_\nu) = \Sigma_{\gamma^{*n\iota}_\nu}$ represent the asymptotic covariance matrices of $\boldsymbol{\alpha}_0, \boldsymbol{\alpha}_\nu, \hat{\boldsymbol{\gamma}}^{\iota}_\nu$ and $\hat{\boldsymbol{\gamma}}^{n\iota}_\nu$ respectively.

The proofs of Theorems 1, 2 and 3 are given in Supplementary Material D.

Remark 2. The fact that the informative and non-informative survival functions are orthogonal (part (ii) of Theorems 1 and 2) suggests that the estimation algorithm will yield more accurate parameter vector updates throughout the iterations (e.g., Nocedal & Wright, 2006). Moreover, Theorem 3 shows that under informative censoring it is possible to estimate the model's coefficients more efficiently since more information is exploited by the informative model.

Remark 3. As far as the construction of confidence intervals and p-values are concerned, for practical purposes it is convenient to adapt to the current context the results discussed in Marra et al. (2017). Supplementary Material E provides more details on this.

# 3 Simulation study

This section provides evidence on the empirical effectiveness of the proposed methodology in recovering true linear effects, non linear effects and baseline functions under informative censoring for three Data Generating Processes (DGPs). The performance of the informative penalized maximum log-likelihood estimator against that of its non-informative counterpart was also examined.

(i) DGP1 ($z_{1i}$ non-informative, $z_{2i}$ informative and censoring rate of about 78%). Event times, $T_{1i}$, were generated from a proportional hazard model, while censored times, $T_{2i}$, were generated from a proportional odd model. These, defined on the survival function scale, are given by

$$\begin{aligned} &\log\left[-\log\left\{S_{10}(t_{1i})\right\}\right]+\alpha_{01}+\alpha_{11}z_{1i}+s_{11}(z_{2i}),\\ &\log\left[\frac{\left\{1-S_{20}(t_{2i})\right\}}{S_{20}(t_{2i})}\right]+\alpha_{02}+\alpha_{12}z_{1i}+s_{12}(z_{2i}),\end{aligned} \qquad (19)$$

where $S_{10}(t_{1i})=0.72\exp(-0.4t_{1i}^{2.4})+0.28\exp(-0.1t_{1i}^{1.0})$ and $S_{20}(t_{2i})=0.99\exp(-0.1t_{2i}^{2.2})+0.01\exp(-0.4t_{2i}^{1.1})$ (Crowther & Lambert, 2013). Covariate $z_{1i}$ was generated using a binomial distribution and $z_{2i}$ using a uniform distribution. As for the smooth functions, we used $s_{11}(z_{2i})=s_{12}(z_{2i})=-0.2\exp(3.2z_i)$, whereas the parametric coefficients were: $\alpha_{01}=0.25, \alpha_{02}=0.85, \alpha_{11}=-2.0$ and $\alpha_{12}=1.8$.

Sample sizes were set to 500, 1000 and 4000, and the number of replicates to 1000. Replicates in which the models did not converge were discarded and replaced with additional ones. The models were fitted using `gamlss()` in `GJRM` by employing the proportional hazard link (`"PH"`) for the event times and the proportional odd link (`"PO"`) for the censoring times (see Supplementary Material A for some software details). The smooth components of $z_2$ were represented using penalized low rank thin plate splines with second order penalty and 10 bases (the default in `GJRM`), and the smooths of times using monotonic penalized B-splines with penalty defined in Section 2.3 and 10 bases. Note that smooth terms of explanatory variables can also be represented using different spline definitions (see Supplementary Material A). In the case of one-dimensional smooth functions, all definitions lead to virtually the same result as long as the amount of smoothing is selected in a data-driven manner (e.g., Wood, 2017). For each replicate, curve estimates were constructed using 200 equally spaced fixed values in the (0, 8) range for the baseline functions and (0, 1) otherwise.

**Results**: Regarding the estimates for $\alpha_{11}$ (the parameter of the non-informative covariate), Figure 4 (in Supplementary Material G) and Table 1 show that overall the mean estimates for the IPMLE and NPMLE are very close to the respective true values and improve as the sample size increases, and that the variability of the estimates decreases as the sample size grows large.

As for the smooth effect of the informative covariate, Figures 6 and 7 (in Supplementary Material G), and Table 1 show that overall the true functions are recovered well by the proposed estimation methods and that the results improve in terms of bias and efficiency as the sample size increases. However, the IPMLE is more efficient than the NPMLE for all sample sizes examined in the simulation study; for example, for $n = 500, 1000$ the RMSE for the NPMLE is more than twice as large as the

IPMLE. Some gains in efficiency are also observed for the baseline functions.

(ii) DGP2 ($z_{1i}$ informative, $z_{2i}$ informative and censoring rate of about 74%). As for DGP1, $T_{1i}$ and $T_{2i}$ were generated using the model defined in (19). However, in this case, the baseline survival functions were defined as $S_{10}(t_{1i}) = 0.75\exp(-0.4t_{1i}^{2.4}) + 0.25\exp(-0.1t_{1i}^{1.0})$ and $S_{20}(t_{2i}) = 0.99\exp(-0.1t_{2i}^{2.2}) + 0.01\exp(-0.4t_{2i}^{1.1})$. The informative covariates, $z_{1i}$ and $z_{2i}$, were generated using binomial and uniform distributions, respectively. Finally, $s_{11}(z_{2i}) = s_{12}(z_{2i}) = -0.2\exp(3.2z_i), \alpha_{01} = 0.25, \alpha_{02} = 0.85$ and $\alpha_{11} = \alpha_{12} = -1.5$.

**Results**: Similarly to DGP1, Figures 5, 8 and 9 and Table 3 (in Supplementary Material G) show that overall the mean estimates for the two estimators are very close to the respective true values and improve as the sample size increases. The variability of the estimates also decreases as the sample size grows large. However, the IPMLE is significantly more efficient than the NPMLE for all cases considered.

Computing times for the proposed approach were on average 8 seconds for *n* = 4000 and around 5 seconds for *n* = 1000, 500. A third DGP with a different smooth function for $z_{2i}$ and with a censoring rate of about 47% was explored (see Supplementary Material G). This DGP suggested the perhaps expected result that the gain in efficiency of the IPMLE tends not to be too significant when a mild censoring rate is considered. Finally, for the above DGPs, we explored the ability of information criteria such as the Akaike information criterion (AIC) and the Bayesian information criterion (BIC), defined in Supplementary Material F, to select the correct model. When doing this, we also considered the informative estimator with incorrectly chosen set of informative covariates (e.g., for DGP1, in estimation, $z_1$ was assumed to be informative instead of $z_2$). For all sample sizes and cases considered both AIC and BIC always chose the correct model.

# 4 Empirical illustration

The modelling framework is illustrated using the data employed by Lu & Zhang (2012), where the aim was to assess how several factors affect the contraction of pneumonia in infants in the presence of informative censoring. According to the World Health Organization (WHO), pneumonia accounted for 16% of all deaths of children under five years old in 2015. The data set consists of 3470 annual personal interviews conducted for the National Longitudinal Survey of Youth from 1979 through 1986 (NLSY, 1995). The response variable, $Y_i$, is the age, in months, at which the infant was hospitalised for pneumonia, and 97.9% of this variable is right censored.

The covariates considered in the modeling were age of the mother in years (`mthage`), `urban` environment (urban = 1, rural = 0), `region` (1 = north-east, 2 = north central, 3 = south, 4 = west), `poverty` (1 = yes, 0 = no), whether the infant had a normal birth weight as defined by weighting at least 5.5 pounds (`wmonth` = 1 if yes and 0 otherwise), `race` (1 = white, 2 = black, 3 = other), `education` (years of school of mother), month the child started to be on solid food (`sfmonth`), average number of cigarettes smoked per week during pregnancy (`smoke` = 0, 1 or 2) and `alcohol` used by mother during pregnancy (0, 1, 2), where the higher the number the higher the frequency of alcohol consumption. To capture the effect of housing crowding (since pneumonia is a communicable disease), number of siblings of the child (`nsibs`) was considered and grouped in three categories (0 for infants without siblings, 1 for infants with one to three siblings, and 2 for more than three siblings.

To assess whether the censoring mechanism was informative, we employed the AIC, BIC, and $K$-Fold Cross validation ($\Upsilon^{\mathrm{KCV}}$) with $K$ = 20 (decreasing or increasing this value did not alter the conclusions); see Supplementary Material F for their definitions. Since several combinations of covariates and link functions had to be considered, a number of models were tried out and the final models selected using the above mentioned criteria. Table 2 in Supplementary Material F shows the results for the chosen models and supports the presence of informative censoring through the `alcohol` and `region` variables (Model 3). Table 2 and Figure 1 present the results for Model 3 and Model 1 (the latter neglects informative censoring).

**Main findings**: From a quick overall look at Table 2, the results exhibit a smaller estimation uncertainty for the informative model. Analysing the table in more detail, the coefficients of `wmonth`, `nsibs1`, `nsibs2` are statistically significant for both models. For instance, the expected hazard for infants with one to three siblings is 2 times that for infants without siblings. Similarly, the expected hazard is 6.4 times higher in infants with more than 3 siblings as compared to infants with no siblings. The parameters of categories `alcohol1` and `region4` of the respective variables are statistically significant at the 10% level for the informative model and are not significant for the non-informative model. The implication of this result is that using the non-informative model the variables `alcohol` and `region` would most likely be removed from the model, hence missing out on some potentially important behavioral and geographical patterns. The table also shows that the smooth functions estimates for `s(u)` and `s(mthage)` are statistically significant for both models, whereas Figure 1 displays their estimated functional forms along with the survival and hazard curves. The plots show, for instance, that, after a certain point, the hazard to contract pneumonia decreases with mother's age. The survival and hazard curves are every similar across the two models with the main difference that the informative approach yields considerably less variable estimates. Our results are consistent with those of Lu & Zhang (2012) who found that the censoring mechanism is informative in this dataset, and that the informative model provides a better fit as compared to its non informative counterpart.

# 5 Discussion

In this article, we have introduced generalized link-based additive survival models with informative censoring and their potential illustrated using simulated and real data. The proofs of the $\sqrt{n}$-consistency and asymptotic normality of the non-informative and informative estimators have been provided. Further, we showed that the newly introduced informative estimator is more efficient than its non-informative counterpart.

Important features of the modelling framework are that: the survival models can account for informative censoring; the baseline functions are estimated non-parametrically via means of monotonic P-splines, which allows one to obtain coherent estimated survival functions; covariate effects are flexibly determined using additive predictors; the optimization scheme allows for the reliable simultaneous penalized estimation of all model's parameters as well as for stable and fast automatic multiple smoothing parameter selection; the models can be easily utilized using the freely available `GJRM` package which allows for several modelling choice as well as for transparent and reproducible research. Given that the assumption of absence of informative censoring is often made for mathematical convenience as well as lack of software, the proposed methodology is likely to appeal researchers in various fields wishing to estimate possibly more realistic survival models.

Future research will focus on extending the proposed informative model to include time varying covariates along with the incorporation of left and interval censored responses, and on the construction of efficient schemes for selecting automatically the set of informative covariates. We will also look into the case of dependent censoring as well as alternative estimation approaches such as sieve maximum likelihood.

# References

Andersen, P. K. & Keiding, N. (2006). *Survival and event history analysis*. Wiley Chichester.

Chen, Y.-H. (2010). Semiparametric marginal regression analysis for dependent competing risks under an assumed copula. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(2), 235–251.

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

Cox, D. R. (1959). The analysis of exponentially distributed life-times with two types of failure. *Journal of the Royal Statistical Society: Series B (Methodological)*, 21(2), 411–421.

Crowder, M. J. (2012). *Multivariate survival analysis and competing risks*. Chapman and Hall/CRC.

Crowther, M. J. & Lambert, P. C. (2013). Simulating biologically plausible complex survival data. *Statistics in medicine*, 32(23), 4118–4134.

Emura, T. & Chen, Y.-H. (2018). *Analysis of Survival Data with Dependent Censoring: Copula-Based Approaches*. Springer.

Hjort, N. L. (1992). On inference in parametric survival data models. *International Statistical Review/Revue Internationale de Statistique*, 60(3), 355–387.

Huang, X. & Zhang, N. (2008). Regression survival analysis with an assumed copula for dependent censoring: a sensitivity analysis approach. *Biometrics*, 64(4), 1090–1099.

Kalbfleisch, J. D. & Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data*. 2nd Ed. Hoboken, Wiley.

Koziol, J. A. & Green, S. B. (1976). A cramér-von mises statistic for randomly censored data. *Biometrika*, 63(3), 465–474.

Lagakos, S. (1979). General right censoring and its impact on the analysis of survival data. *Biometrics*, 35(1), 139–156.

Leitenstorfer, F. & Tutz, G. (2006). Generalized monotonic regression based on b-splines with an application to air pollution data. *Biostatistics*, 8(3), 654–673.

Li, R. & Peng, L. (2015). Quantile regression adjusting for dependent censoring from semicompeting risks. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 77(1), 107–130.

Liu, X.-R., Pawitan, Y., & Clements, M. (2018). Parametric and penalized generalized survival models. *Statistical methods in medical research*, 27(5), 1531–1546.

Lu, Z. & Zhang, W. (2012). Semiparametric likelihood estimation in survival models with informative censoring. *Journal of Multivariate Analysis*, 106, 187–211.

Ma, J., Heritier, S., & Lô, S. N. (2014). On the maximum penalized likelihood approach for proportional hazard models with right censored survival data. *Computational Statistics & Data Analysis*, 74, 142–156.

Marra, G. & Radice, R. (2019). *GJRM: Generalised Joint Regression Modelling*. R package version 0.2-1.

Marra, G., Radice, R., Bärnighausen, T., Wood, S. N., & McGovern, M. E. (2017). A simultaneous equation approach to estimating hiv prevalence with nonignorable missing responses. *Journal of the American Statistical Association*, 112(518), 484–496.

NLSY (1995). National Longitudinal Survey of Youth Handbook. The Ohio State University.

Nocedal, J. & Wright, S. (2006). *Numerical optimization, series in operations research and financial engineering*. Springer, New York, USA, 2006.

Pya, N. & Wood, S. N. (2015). Shape constrained additive models. *Statistics and Computing*, 25(3), 543–559.

Reid, N. (1994). A conversation with sir david cox. *Statistical Science*, 9(3), 439–455.

Scheike, T. H. & Zhang, M.-J. (2003). Extensions and applications of the cox-aalen survival model. *Biometrics*, 59(4), 1036–1045.

Siannis, F., Copas, J., & Lu, G. (2005). Sensitivity analysis for informative censoring in parametric survival models. *Biostatistics*, 6(1), 77–91.

Slud, E. V. & Rubinstein, L. V. (1983). Dependent competing risks and summary survival curves. *Biometrika*, 70(3), 643–649.

Tsiatis, A. (1975). A nonidentifiability aspect of the problem of competing risks. *Proceedings of the National Academy of Sciences*, 72(1), 20–22.

Vatter, T. & Chavez-Demoulin, V. (2015). Generalized additive models for conditional dependence structures. *Journal of Multivariate Analysis*, 141, 147–167.

Wang, A., Chandra, K., Xu, R., & Sun, J. (2015). The identifiability of dependent competing risks models induced by bivariate frailty models. *Scandinavian Journal of Statistics*, 42(2), 427–437.

Wood, S. N. (2017). *Generalized Additive Models: An Introduction With R*. Second Edition, Chapman & Hall/CRC, London.

Xu, J., Ma, J., Connors, M. H., & Brodaty, H. (2018). Proportional hazard model estimation under dependent censoring using copulas and penalized likelihood. *Statistics in medicine*, 37(14), 2238–2251.

Xu, J., Ma, J., & Prvan, T. (2017). Non parametric hazard estimation with dependent censoring using penalized likelihood and an assumed copula. *Communications in Statistics-Theory and Methods*, 46(22), 11383–11403.

Xue, X., Xie, X., & Strickler, H. D. (2018). A censored quantile regression approach for the analysis of time to event data. *Statistical methods in medical research*, 27(3), 955–965.

Younes, N. & Lachin, J. (1997). Link-based models for survival data with interval and continuous time censoring. *Biometrics*, 53(4), 1199–1211.

Yuan, M. (2005). Semiparametric censorship model with covariates. *Test*, 14(2), 489–514.

Zeng, D. et al. (2004). Estimating marginal survival function by adjusting for dependent censoring using many covariates. *The Annals of Statistics*, 32(4), 1533–1555.

Zheng, M. & Klein, J. P. (1995). Estimates of marginal survival for dependent competing risks based on an assumed copula. *Biometrika*, 82(1), 127–138.

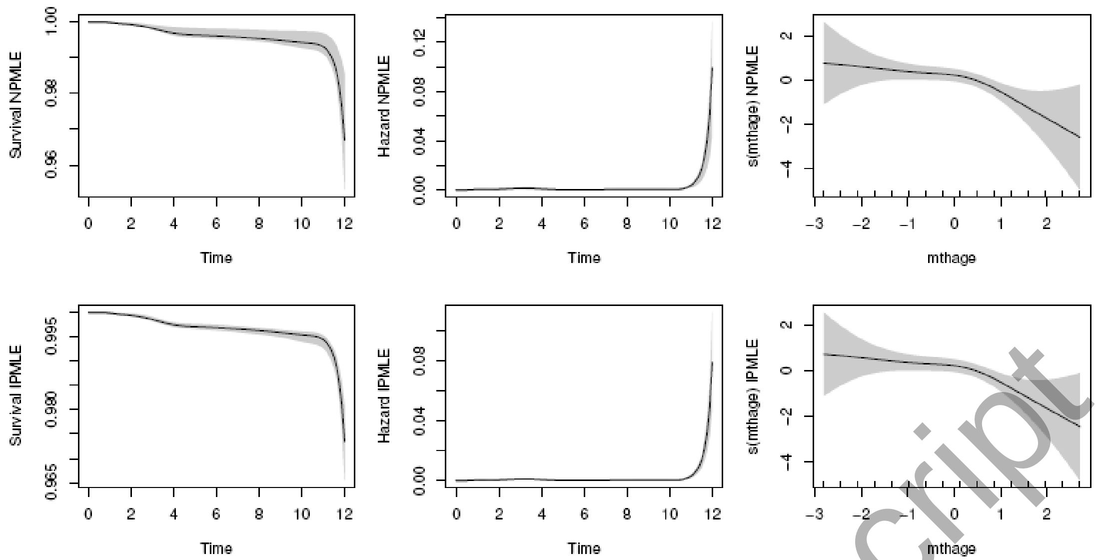


**Fig. 1** Smooth function estimates and their corresponding 95% intervals for Model 1 (non-informative model) and Model 3 (informative model) obtained by applying `gamlss()` in `GJRM` to pneumonia data. The intervals have been obtained using the approach described in Supplementary Material E.

**Table 1** Bias and root mean squared error (RMSE) for the IPMLE and NPMLE obtained by applying the `gamlss()` to informative survival data simulated according to DGP1 characterised by a censoring rate of about 78%. Bias and RMSE for the smooth terms are calculated, respectively, as $n_s^{-1}\sum_{i=1}^{n_s}|\bar{\hat{s}}_i - s_i|$ and $n_s^{-1}\sum_{i=1}^{n_s}\sqrt{n_{rep}^{-1}\sum_{rep=1}^{n_{rep}}\left(\hat{s}_{rep,i} - s_i\right)^2}$, where $\bar{\hat{s}}_i = n_{rep}^{-1}\sum_{rep=1}^{n_{rep}}\hat{s}_{rep,i}$, $n_s$ is the number of equally spaced fixed values in the (0, 8) or $(0,1)$ range, and $n_{rep}$ is the number of simulation replicates. In this case, $n_s$ = 200 and $n_{rep}$ = 1000. The bias for the smooth terms is based on absolute differences in order to avoid compensating effects when taking the sum.

| (a) Informative Penalized Maximum Log-likelihood Estimator (IPMLE) | | | | | | |
|---|---|---|---|---|---|---|
| | Bias | | | RMSE | | |
| | n = 500 | n = 1000 | n = 4000 | n = 500 | n = 1000 | n = 4000 |
| $\alpha_{11}$ | -0.047 | -0.013 | -0.001 | 0.369 | 0.239 | 0.118 |
| $s_{11}$ | 0.036 | 0.028 | 0.013 | 0.161 | 0.114 | 0.061 |
| $h_{10}$ | 0.095 | 0.069 | 0.034 | 0.336 | 0.245 | 0.104 |
| $S_{10}$ | 0.027 | 0.024 | 0.018 | 0.071 | 0.054 | 0.033 |
| (b) Non-informative Penalized Maximum Log-likelihood Estimator (NPMLE) | | | | | | |
| | Bias | | | RMSE | | |
| | n = 500 | n = 1000 | n = 4000 | n = 500 | n = 1000 | n = 4000 |
| $\alpha_{11}$ | -0.079 | -0.015 | -0.005 | 0.360 | 0.245 | 0.116 |
| $s_{11}$ | 0.085 | 0.069 | 0.046 | 0.383 | 0.206 | 0.118 |
| $h_{10}$ | 0.120 | 0.070 | 0.034 | 0.427 | 0.292 | 0.121 |
| $S_{10}$ | 0.034 | 0.025 | 0.017 | 0.086 | 0.068 | 0.039 |

**Table 2** Estimation results of the non-informative and informative models (Models 1 and 3, respectively, in Table 5 in Supplementary Material F) applied to pneumonia data. The models were fitted using `gamlss()` in `GJRM` by employing the "`PH-PH`" link functions combination. Furthermore, $\mathrm{EDF}$ and $\mathrm{Ref.DF}$ refer to the effective degrees of freedom and reference degrees of freedom of the smooths. More details can be founded in Supplementary Materials C and E.

| (a) Model 1 (NPMLE) | | | | |
|---|---|---|---|---|
| Linear Covariates | Estimate | Standart Error | Z-value | P-value |
| intercept | -71.28 | 45.52 | -1.566 | 0.117 |
| alcohol1 | 0.364 | 0.310 | 1.174 | 0.240 |
| alcohol2 | -0.130 | 0.336 | -0.386 | 0.700 |
| nsibs1 | 0.696 | 0.258 | 2.695 | 0.007 ** |
| nsibs2 | 1.833 | 0.761 | 2.408 | 0.016 * |
| region2 | -0.004 | 0.343 | -0.012 | 0.991 |
| region3 | -0.489 | 0.343 | -1.426 | 0.154 |
| region4 | -0.698 | 0.438 | -1.595 | 0.111 |
| wmonth | -0.767 | 0.293 | -2.617 | 0.009 ** |
| Smooth Variables | EDF | Ref.DF | Chi-square | P-value |
| s(u) | 7.776 | 8.640 | 101.94 | <2e-16 *** |
| s(mthage) | 2.503 | 3.171 | 10.41 | 0.019 * |
| (b) Model 3 (IPMLE) | | | | |
| Linear Covariates | Estimate | Standart Error | Z-value | P-value |
| intercept | -71.50 | 45.51 | -1.571 | 0.116 |
| alcohol1 | 0.086 | 0.046 | 1.859 | 0.063 . |
| alcohol2 | 0.022 | 0.046 | 0.472 | 0.637 |
| nsibs1 | 0.687 | 0.257 | 2.670 | 0.008 ** |
| nsibs2 | 1.860 | 0.760 | 2.448 | 0.014 * |
| region2 | -0.063 | 0.056 | -1.135 | 0.256 |
| region3 | -0.017 | 0.052 | -0.325 | 0.745 |
| region4 | -0.107 | 0.059 | -1.814 | 0.070 . |

| (a) Model 1 (NPMLE) | | | | |
|---|---|---|---|---|
| wmonth | -0.761 | 0.291 | -2.616 | 0.009 ** |
| Smooth Variables | EDF | Ref.DF | Chi-square | P-value |
| s(u) | 7.776 | 8.640 | 101.59 | <2e-16 *** |
| s(mthage) | 2.466 | 3.127 | 9.501 | 0.026 * |