Link-Based Survival Additive Models under Mixed Censoring to Assess Risks of Hospital-Acquired Infections

Giampiero Marra* Alessio Farcomeni† Rosalba Radice‡

2020-09-16

Abstract

The majority of methods available to model survival data only deal with right censoring. However, there are many applications where left, right and/or interval censoring simultaneously occur. A methodology that is capable of handling all types of censoring as well as flexibly estimating several types of covariate effects is presented. The baseline hazard is modelled through monotonic P-splines. The model's parameters are estimated using an efficient and stable penalised likelihood algorithm. The proposed framework is evaluated in simulation, and illustrated using an original data example on time to first hospital infection or in-hospital death in cirrhotic patients. A peak of risk in the first week since hospitalisation is identified, together with a non-linear effect of Model for End-Stage Liver Disease (MELD) score. The `GJRM R` package, with an implementation of our approach, is freely available on CRAN.

Key Words: additive predictor; link function; mixed censoring; penalised log-likelihood; regression splines; survival data.

*Corresponding author. Department of Statistical Science, University College London, Gower Street, London WC1E 6BT, UK, giampiero.marra@ucl.ac.uk.

†Department of Economics and Finance, University of Rome "Tor Vergata", Via Columbia 2, 00133 Roma, Italy alessio.farcomeni@uniroma2.it.

‡Business School, City, University of London, 106 Bunhill Row, EC1Y 8TZ London, UK, rosalba.radice@city.ac.uk.

1 Introduction

Survival data are encountered in many applications and since the pioneering work of Cox (1972) a great deal of attention has been devoted to developing survival models, and related estimation techniques, for right-censored event times. However, there are many situations where the data are simultaneously affected by different types of censoring mechanisms. For example, AIDS trials are often concerned with determining the incubation period of the HIV virus, defined as the time elapsed from HIV infection to the onset of AIDS. Since the diagnosis of the disease is usually based on blood testing, which can only be carried out on a periodic basis, it is impossible to know exactly what the incubation period is, hence giving rise to interval-censoring (Odell et al., 1992). Other examples are carcinogenesis studies, such as the Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial (Wang et al., 2016). For more examples of interval-censored data, in various fields, we refer the reader to Sun (2006) and Zhang & Sun (2010). The presence of interval-censored observations does not rule out other types of censoring. In fact, it is perfectly possible for some patients to have experienced the event of interest before the first screening or, alternatively, to reach the end of the trial without ever experiencing it, thus generating left- and right-censored observations, respectively. In many cases, furthermore, it might be additionally possible to precisely measure the time to event for some subjects, therefore having additionally uncensored observations. We refer to this situation as *mixed* censoring (Schick & Yu, 2000), also referred to as partly interval-censoring, which naturally arises with composite endpoint definitions which are widespread, especially in the fields of cardiology, internal medicine and oncology. For instance, event-free survival in cancer studies is defined as the time between primary treatment and the occurrence of any of a series of cancer-specific complications of events. These might include events that can be measured precisely (e.g., death) and others that can only be guaranteed to have occurred in a time interval between two screenings (e.g., leukopenia). A common approach in the presence of truly uncensored events is to treat the interval censored ones as uncensored at the upper limit of the interval, that is, at the moment of diagnosis. This is well acknowledged to be a possible source of bias (e.g., Odell et al., 1992; Fleming et al., 2009).

Our work is motivated by an original application to the evaluation of risk of in-hospital adverse

events (death or new onset of infection) in cirrhotic patients. Italy, and southern Europe in general, is a high risk country for multi-drug resistant pathogens, which occur much more commonly in the form of hospital infections than in community-acquired ones (Merli et al., 2015; Bartoletti et al., 2018; Piano et al., 2019). Cirrhotic patients, due to compromised liver functionality or as a side effect of treatment, are additionally oftentimes immunodepressed and hence at higher risk of infections. Our data consist of $n = 678$ cirrhotic patients who were admitted to Policlinico Umberto I hospital in Rome, Italy, between 2009 and 2017. Of these, none was infected at admission, none was taking antibiotics, and none was scheduled for (nor had) major surgery during the hospital stay. The endpoint is a composite one, where an event is defined as the occurrence of an infection or death before hospital discharge. Times were recorded from admission. The main scientific questions with the data at hand revolve around the possibility of an increased risk of infection or death due to the use of catheterism, paracentesis, and overcrowding of the ward. We would like to model the effect of these binary predictors after non-parametrically adjusting for the effect of MELD, a score summarising the progression of liver failure. Indeed, a clearly non-linear effect of MELD will be discovered, indicating that a simple polynomial effect of this predictor would lead to misleading inference. Clearly, these data provide uncensored time-to-event in case death (before infection) is observed, and right-censored data if no event occurs before hospital discharge. Furthermore, in case an infection is observed, the event time is only known to have occurred between the last and current assessments (usually within a time span of 12 to 48 hours), therefore having also interval-censored event times.

At present, survival models with the simultaneous presence of different types of censoring can be easily handled through accelerated failure time (parametric) models. Cox regression with mixed censoring is computationally cumbersome (Satten, 1996; Goggins et al., 1998), although an efficient implementation can be found in Anderson-Bergman (2017). There are works which proposed estimating flexible survival models under mixed censoring and in the following we mention the perhaps most relevant to this paper. Recent articles include Liu et al. (2018) who proposed generalised survival models to estimate covariate effects flexibly while accounting for the monotonicity constraint on the survival function via a penalty term. Li & Ma (2019) employed a primal-dual

interior point algorithm to estimate additive hazards models with parametric covariate effects and non-negative constraints on the hazards via M-splines. Szabo et al. (2020) proposed a sieve maximum likelihood two-step estimation procedure based on polynomial splines for the accelerated hazards model. Wang et al. (2016) introduced an EM algorithm to estimate proportional hazards (PH) models that estimate covariate effects parametrically, and use monotone splines to approximate the cumulative baseline hazard function. The literature on survival modelling is vast and some interesting developments and `R` implementations are discussed in Fauvernier et al. (2019) and Komarek et al. (2005), and many models incorporated in the `survival` package. These, however, do not allow for either mixed censoring or flexible baseline and covariate effects via penalised regression splines.

Building on Marra & Radice (2020a), we present in this work a flexible parametric methodology that is capable of handling simultaneously all types of censoring, estimating covariate effects via additive predictors, and modelling the baseline hazard by means of monotonic P-splines. The proposed link-based survival additive model yields the widely used PH and proportional odds (PO) models as special cases. Importantly, the modelling framework avoids numerical integration, which may lead to unstable and slow computations. The resulting additive model is very flexible. Modelling the baseline hazard by means of monotonic P-splines is more efficient and parsimonious than using a non-parametric hazard as in Cox models, and at the same time much more flexible than making strong parametric assumptions as in Accelerated Failure Time (AFT) models. Parameter estimation is based on a penalised maximum likelihood approach with automatic multiple smoothing parameter selection, which allows for stable and efficient computations. Note that the closest approach to ours is that by Liu et al. (2017, 2018), however, as opposed to our proposal, these authors impose monotonicity via a penalty term and, as they point out, their algorithm requires improvements when it comes to multidimensional smoothing parameter estimation. In order to facilitate the use of the developments in this article in industry and academia, as well as enhance reproducible research, our methods are available within the `GJRM` package (Marra & Radice, 2020b) for the `R` (R Development Core Team, 2020) software.

The rest of the paper is organized as follows. Model formulation and parameter estimation are

discussed in Sections 2 and 3, with further details provided in Section 4. A simulation study is presented in Section 5, and the results obtained by applying the proposed modelling framework to real data are discussed in Section 6. The final section concludes the paper with some directions of future research.

2 Model formulation

The link-based additive survival models discussed below are essentially based on three ingredients: the survival function, a link function and an additive predictor. Although the key intuition behind this construction is provided by Younes & Lachin (1997), the model presented here is more flexible and is based on Marra & Radice (2020a). Let T_i have a conditional survival function generically denoted by $S(t_i|\mathbf{x}_i;\boldsymbol{\beta}) = P(T_i > t_i|\mathbf{x}_i;\boldsymbol{\beta}) \in (0,1)$, where $\mathbf{x}_i$ is a generic vector of patient characteristics that has an associated coefficient vector $\boldsymbol{\beta} \in \mathbb{R}^w$ with w given by the length of $\boldsymbol{\beta}$. Then a link-based additive survival model can be written as

$$g\left[S(t_i|\mathbf{x}_i;\boldsymbol{\beta})\right] = \eta_i(t_i, \mathbf{x}_i; \mathbf{f}(\boldsymbol{\beta})), \tag{1}$$

where $g : (0,1) \to \mathbb{R}$ is a monotone and twice continuously differentiable link function with bounded derivatives and hence invertible, and $\eta_i(t_i, \mathbf{x}_i; \mathbf{f}(\boldsymbol{\beta})) \in \mathbb{R}$ is an additive predictor, defined in more detail in the next paragraph, which includes a baseline function of time (or a stratified set of functions of time) to model the baseline hazard, and several types of covariate effects. $\mathbf{f}(\boldsymbol{\beta})$ is a vector function of $\boldsymbol{\beta}$ whose main role is to impose a monotonicity constraint when evaluating the baseline function of time contained in the additive predictor; this is discussed in detail in Section 3. A simple rearrangement of (1) yields $S(t_i|\mathbf{x}_i;\boldsymbol{\beta}) = G\left\{\eta_i(t_i, \mathbf{x}_i; \mathbf{f}(\boldsymbol{\beta}))\right\}$, where G is the inverse link function. The cumulative hazard and hazard functions, H and h, are defined as $H(t_i|\mathbf{x}_i;\boldsymbol{\beta}) = -\log\left[G\left\{\eta_i(t_i, \mathbf{x}_i; \mathbf{f}(\boldsymbol{\beta}))\right\}\right]$ and

$$h(t_i|\mathbf{x}_i;\boldsymbol{\beta}) = -\frac{G'\left\{\eta_i(t_i, \mathbf{x}_i; \mathbf{f}(\boldsymbol{\beta}))\right\}}{G\left\{\eta_i(t_i, \mathbf{x}_i; \mathbf{f}(\boldsymbol{\beta}))\right\}} \frac{\partial \eta_i(t_i, \mathbf{x}_i; \mathbf{f}(\boldsymbol{\beta}))}{\partial t_i}, \tag{2}$$

where $G'\left\{\eta_i(t_i, \mathbf{x}_i; \mathbf{f}(\boldsymbol{\beta}))\right\} = \partial G\left\{\eta_i(t_i, \mathbf{x}_i; \mathbf{f}(\boldsymbol{\beta}))\right\} / \partial \eta_i(t_i, \mathbf{x}_i; \mathbf{f}(\boldsymbol{\beta}))$. Table 1 displays the functions g, G and G' implemented for this work.

Model	Link $g(S)$	Inverse link $g^{-1}(\eta) = G(\eta)$	$G'(\eta)$
Prop. hazards (`"PH"`)	$\log\{-\log(S)\}$	$\exp\{-\exp(\eta)\}$	$-G(\eta)\exp(\eta)$
Prop. odds (`"PO"`)	$-\log\left(\frac{S}{1-S}\right)$	$\frac{\exp(-\eta)}{1+\exp(-\eta)}$	$-G^2(\eta)\exp(-\eta)$
probit (`"probit"`)	$-\Phi^{-1}(S)$	$\Phi(-\eta)$	$-\phi(-\eta)$

Table 1: Link functions implemented in `GJRM`. Φ and ϕ are the cumulative distribution and density functions of a univariate standard normal distribution. The first two functions are typically known as log-log and -logit links, respectively. These are the same as those in Liu et al. (2018).

Let us now consider the construction of η_i where in this paragraph, for the sake of simplicity, the dependence on covariates and parameters has been dropped. Since t_i can be treated as a regressor, we define an overall covariate vector $\mathbf{z}_i$ made up of $\mathbf{x}_i$ and t_i. The main benefits of using an additive predictor 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 about their functional forms. However, additivity here implies that not all the interaction terms among the covariates may be included in η_i. There are many textbooks on the subject and we refer the reader to Wood (2017) for a thorough discussion. An additive predictor can be defined as

$$\eta_i = \beta_0 + \sum_{k=1}^{K} s_k(\mathbf{z}_{ki}),\ i = 1, \ldots, n, \tag{3}$$

where $\beta_0 \in \mathbb{R}$ is an overall intercept, $\mathbf{z}_{ki}$ denotes the k^{th} sub-vector of the complete vector $\mathbf{z}_i$ and the K functions $s_k(\mathbf{z}_{ki})$ denote effects which are chosen according to the type of covariate(s) considered. Each $s_k(\mathbf{z}_{ki})$ can be represented as a linear combination of J_k basis functions $b_{kj_k}(\mathbf{z}_{ki})$ and coefficients $f_{kj_k}(\beta_{kj_k})$, that is

$$\sum_{j_k=1}^{J_k} f_{kj_k}(\beta_{kj_k}) b_{kj_k}(\mathbf{z}_{ki}). \tag{4}$$

The above formulation implies that the vector of evaluations $\{s_k(\mathbf{z}_{k1}), \ldots, s_k(\mathbf{z}_{kn})\}^\mathsf{T}$ can be written as $\mathbf{Z}_k \mathbf{f}_k(\boldsymbol{\beta}_k)$ with $\mathbf{f}_k(\boldsymbol{\beta}_k) = (f_{k1}(\beta_{k1}), \ldots, f_{kJ_k}(\beta_{kJ_k}))^\mathsf{T}$ and design matrix $\mathbf{Z}_k[i, j_k] = b_{kj_k}(\mathbf{z}_{ki})$.

This allows the predictor in equation (3) to be written as

$$\boldsymbol{\eta} = \beta_0 \mathbf{1}_n + \mathbf{Z}_1 \mathbf{f}_1(\boldsymbol{\beta}_1) + \ldots + \mathbf{Z}_K \mathbf{f}_K(\boldsymbol{\beta}_K), \tag{5}$$

where $\mathbf{1}_n$ is an n-dimensional vector made up of ones. Equation (5) can also be written in a more compact way as $\boldsymbol{\eta} = \mathbf{Z}\mathbf{f}(\boldsymbol{\beta})$, where $\mathbf{Z} = (\mathbf{1}_n, \mathbf{Z}_1, \ldots, \mathbf{Z}_K)$ and $\mathbf{f}(\boldsymbol{\beta}) = (\beta_0, \mathbf{f}_1(\boldsymbol{\beta}_1)^\mathsf{T}, \ldots, \mathbf{f}_K(\boldsymbol{\beta}_K)^\mathsf{T})^\mathsf{T}$. Recall that $\mathbf{f}(\boldsymbol{\beta})$ serves to impose a monotonicity constraint when evaluating the baseline smooth function of time. In fact, the $\mathbf{f}_k$ vector functions will all be set to the identity function except for the one related to the baseline which is specified in Section 3. Each $\boldsymbol{\beta}_k$ has an associated quadratic penalty $\lambda_k \boldsymbol{\beta}_k^\mathsf{T} \mathbf{D}_k \boldsymbol{\beta}_k$, used in fitting, whose role is to enforce specific properties on the k^{th} function, such as smoothness. Note that each matrix $\mathbf{D}_k$ only depends on the choice of the basis functions. Smoothing parameter $\lambda_k \in [0, \infty)$ controls the trade-off between fit and smoothness, and plays a crucial role in determining the shape of $\hat{s}_k(\mathbf{z}_{ki})$. The overall penalty can be defined as $\boldsymbol{\beta}^\mathsf{T}\mathbf{S}\boldsymbol{\beta}$, where $\mathbf{S} = \mathrm{diag}(0, \lambda_1 \mathbf{D}_1, \ldots, \lambda_K \mathbf{D}_K)$. Recall that smooth functions are typically subject to centering (identifiability) constraints. Depending on the types of covariate effects one wishes to model (e.g., non-linear, random, spatial), several definitions of basis functions and penalty terms are possible and we refer the reader to Wood (2017) for all the options available. The spline definition and penalty employed for the baseline smooth function of time are discussed in the second paragraph of Section 3.

In equation (2), quantity $\partial \eta_i(t_i, \mathbf{x}_i; \mathbf{f}(\boldsymbol{\beta}))/\partial t_i$ is required. The results of the previous paragraph allow us to re-write $\eta_i(t_i, \mathbf{x}_i; \mathbf{f}(\boldsymbol{\beta}))$ as $\mathbf{Z}_i(t_i, \mathbf{x}_i)^\mathsf{T} \mathbf{f}(\boldsymbol{\beta})$, where $\mathbf{Z}_i(t_i, \mathbf{x}_i)$ denotes the i^{th} row of the $\mathbf{Z}$ matrix (that is based on covariates and the time variable, as pointed out earlier). The derivative of interest can then be obtained as $\lim_{\varepsilon \to 0} \left\{ \frac{\mathbf{Z}_i(t_i+\varepsilon, \mathbf{x}_i) - \mathbf{Z}_i(t_i-\varepsilon, \mathbf{x}_i)}{2\varepsilon} \right\}^\mathsf{T} \mathbf{f}(\boldsymbol{\beta}) = \mathbf{Z}_i'^\mathsf{T} \mathbf{f}(\boldsymbol{\beta})$, where, depending on the type of spline basis employed, $\mathbf{Z}_i'$ can be calculated either by a finite-difference method or analytically.

Following, e.g., Royston & Parmar (2002), the link-based additive survival model can be written as

$$g\left\{S(t_i|\mathbf{x}_i)\right\} = g\left\{S_0(t_i)\right\} + \sum_{k=2}^{K} s_k(\mathbf{x}_{ki}), \tag{6}$$

where $S_0(t_i)$ is a baseline survival function. If we replace $g\{S_0(t_i)\}$ with $s_0(t_i)$ then the RHS of (6) becomes notationally consistent with (3), which also shows that $s_0(t_i)$ is effectively modelling a transformation of the baseline survival function. The use of $s_0(t_i)$ as a predictor leads, as stated above, to a semi-parametric baseline hazard.

The choice for g determines the scale of the analysis (e.g., Liu et al., 2018). For instance, model (6) yields the proportional hazards model when choosing the log-log link, i.e.

$$\log\{H(t_i|\mathbf{x}_i)\} = \log\{H_0(t_i)\} + \sum_{k=2}^{K} s_k(\mathbf{x}_{ki}), \tag{7}$$

where $H(t_i|\mathbf{x}_i) = -\log\{S(t_i|\mathbf{x}_i)\}$, and $H_0(t_i) = -\log\{S_0(t_i)\}$ is the baseline cumulative hazard function. Important benefits of modelling on the log-cumulative hazard scale are that the corresponding function is computationally more stable than the log-hazard function, that quantities such as $h(t_i|\mathbf{x}_i)$ and $S(t_i|\mathbf{x}_i)$ can be directly obtained without the need for numerical integration, and that time-dependent effects can be easily incorporated in the model using terms like $s_k(t_i)\mathbf{x}_{ki}$. When the RHS of (7) contains time-dependent effects, the model loses the proportional hazards interpretation. Model (6) yields the proportional odds model when the -logit link is chosen. Finally, note that time-varying covariates can be incorporated as usual by representing data in time intervals. For example, a subject with three measurements for a covariate will contribute with three time intervals, the first two of which will be associated with right-censoring.

Remark 1. For certain smooth functions, such as those modelling the effects of continuous covariates, quantity J_k in (4) has to be set to some value to make the computation feasible. This implies the well known fact that the unknown $s_k(\mathbf{z}_{ki})$ may not have an exact representation as given in (4). In practice, J_k is set to a typically large arbitrary value that allows for "enough" flexibility in estimating the smooth term. This is not problematic since the coefficients of the spline basis are penalised in the estimation process such that the smooth term's complexity that is not supported by the data is suppressed (e.g., Wood, 2017).

Remark 2. For left- and right-censored observations (call them l_i and r_i), as well as exact observations, the additive predictor is uniquely defined since there is only one relevant time-to-event

datum t_i for each individual. Specifically, $t_i = l_i$ for left-censored observations, $t_i = r_i$ for right-censored observations and $t_i = l_i = r_i$ for exact observations, where l_i and r_i are realisations of the respective random variables L_i and R_i. However, in the case of interval-censoring, the model formulation is slightly more involved due to the need to account for the information contained in both the lower and upper bounds of the censoring interval. This means that interval-censored observations require the set-up of two distinct design matrices, hence additive predictors, based on l_i and r_i. Ultimately, all the baseline covariates are the same and only one vector of parameters $\boldsymbol{\beta}$ will be estimated, but the predictor itself is a function of time and, as such, will take on different values depending on whether it is evaluated at l_i or r_i. If the ith observation is interval-censored then we need to define $\eta_i(l_i, \mathbf{x}_i; \mathbf{f}(\boldsymbol{\beta}))$ and $\eta_i(r_i, \mathbf{x}_i; \mathbf{f}(\boldsymbol{\beta}))$ which can be expressed as $\mathbf{Z}_{1i}(l_i, \mathbf{x}_i)^\mathsf{T}\mathbf{f}(\boldsymbol{\beta})$ and $\mathbf{Z}_{2i}(r_i, \mathbf{x}_i)^\mathsf{T}\mathbf{f}(\boldsymbol{\beta})$, where $\mathbf{Z}_{1i}(l_i, \mathbf{x}_i)$ and $\mathbf{Z}_{2i}(r_i, \mathbf{x}_i)$ are identical except for the time variables.

3 Parameter estimation

For each individual i, let T_i denote the true event time. Due to censoring, T_i may not be recorded exactly, in which case the random variable is only known to lie within the interval (L_i, R_i), where L_i and R_i are left and right censoring times. If $L_i = 0$ then the observation is defined as "left-censored", if $R_i = \infty$ then the observation is classified as "right-censored", and if L_i and R_i take on finite distinct non-zero values then the observation is classified as “interval-censored”. Exact observations correspond to the case $L_i = R_i$. The censoring type for the ith observation is represented by the indicator functions δ_{Li}, δ_{Ri}, δ_{Ii} and δ_{Ui}.

Let us assume that a random *i.i.d.* sample $\{(l_i, r_i, \delta_{Ui}, \delta_{Li}, \delta_{Ri}, \delta_{Ii}, \mathbf{x}_i)\}_{i=1}^n$ is available, where n represents the sample size, that there are no competing risks and that censoring is independent and non-informative conditional on $\mathbf{x}_i$. The log-likelihood function can be written as

$$\begin{aligned}\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} \delta_{Ui} \log\left[-\frac{\partial G\left\{\eta_i(l_i)\right\}}{\partial \eta_i(l_i)} \frac{\partial \eta_i(l_i)}{\partial l_i}\right] + \delta_{Li} \log\left[1 - S\left\{\eta_i(l_i)\right\}\right] \\ + \delta_{Ri} \log\left[S\left\{\eta_i(r_i)\right\}\right] + \delta_{Ii} \log\left[S\left\{\eta_i(l_i)\right\} - S\left\{\eta_i(r_i)\right\}\right],\end{aligned} \tag{8}$$

Note that, in the case of exact observations, l_i and r_i are interchangeable. The proposed model

allows for a high degree of flexibility, which is why penalised estimation of $\boldsymbol{\beta}$ is advisable. In order to prevent over-fitting we maximise the penalised log-likelihood

$$\ell_p(\boldsymbol{\beta}) = \ell(\boldsymbol{\beta}) - \frac{1}{2}\boldsymbol{\beta}^{\mathsf{T}}\mathbf{S}\boldsymbol{\beta}. \tag{9}$$

To ensure that the estimated survival function is monotonically decreasing or equivalently that the hazard function is positive (achieved if $\partial\eta_i(l_i)/\partial l_i$ is positive), we model the time effects using the monotonic P-spline approach, which is explained using a simplified notation for the sake of simplicity. Let $s(t_i) = \sum_{j=1}^{J} f_j(\beta_j)b_j(t_i)$, where the b_j are B-spline basis functions of at least second order built over the interval $[a, b]$, based on equally spaced knots, and the $f_j(\beta_j)$ are spline coefficients. A sufficient condition for $s'(t_i) \geq 0$ over $[a, b]$ is that $f_j(\beta_j) \geq f_j(\beta_{j-1})\ \forall j$ (e.g., Leitenstorfer & Tutz, 2007). Such condition can be imposed by defining $\mathbf{f}(\boldsymbol{\beta}) = \boldsymbol{\Sigma}\left\{\beta_1, \exp(\beta_2), \ldots, \exp(\beta_J)\right\}^{\mathsf{T}}$, where $\boldsymbol{\Sigma}[\iota_1, \iota_2] = 0$ if $\iota_1 < \iota_2$ and $\boldsymbol{\Sigma}[\iota_1, \iota_2] = 1$ if $\iota_1 \geq \iota_2$, with ι_1 and ι_2 denoting the row and column entries of $\boldsymbol{\Sigma}$. (Note that, in practice, $\boldsymbol{\Sigma}$ is absorbed into the design matrix containing the B-spline basis functions.) When setting up the penalty term we penalise the squared differences between adjacent β_j, starting from β_2, using $\mathbf{D} = \mathbf{D}^{*\mathsf{T}}\mathbf{D}^*$ where $\mathbf{D}^*$ is a $(J-2) \times J$ matrix made up of zeros except that $\mathbf{D}^*[\iota, \iota+1] = -\mathbf{D}^*[\iota, \iota+2] = 1$ for $\iota = 1, ..., J-2$ (Pya & Wood, 2015).

Following Marra & Radice (2020a), estimation of $\boldsymbol{\beta}$ and $\boldsymbol{\lambda}$ is achieved using a two-stage algorithm whose main ingredients are the analytical score vector and Hessian matrix (see Appendix A). Given the structure of (8), deriving such quantities has been somewhat tedious, especially because of the non-linear dependence of $\mathbf{f}(\boldsymbol{\beta})$ on $\boldsymbol{\beta}$ which gave rise to terms like $\partial^2\eta_i(t_i, \mathbf{x}_i; \mathbf{f}(\boldsymbol{\beta}))/\partial t_i\partial\boldsymbol{\beta} = \mathbf{z}_i'^{\mathsf{T}}\mathbf{E}$ and $\partial\eta_i(t_i, \mathbf{x}_i; \mathbf{f}(\boldsymbol{\beta}))/\partial\boldsymbol{\beta} = \mathbf{z}_i^{\mathsf{T}}\mathbf{E}$, where $\mathbf{E}$ is a vector such that $\mathbf{E}[kj_k] = 1$ if $f_{kj_k}(\beta_{kj_k}) = \beta_{kj_k}$ and $\exp(\beta_{kj_k})$ otherwise. However, the computational benefits of avoiding approximations justified the effort. The algorithm can be summarised as follows:

- *Trust region step*: holding $\boldsymbol{\lambda}$ fixed at a vector of values and for a given $\boldsymbol{\beta}^{[a]}$, where a is an

iteration index, maximise equation (9) using

$$\boldsymbol{\beta}^{[a+1]} = \boldsymbol{\beta}^{[a]} + \underset{\mathbf{e}:\|\mathbf{e}\|\leq\Delta^{[a]}}{\arg\min}\ \breve{\ell}_p(\boldsymbol{\beta}^{[a]}), \tag{10}$$

where $\breve{\ell}_p(\boldsymbol{\beta}^{[a]}) = -\left\{\ell_p(\boldsymbol{\beta}^{[a]}) + \mathbf{e}^\mathsf{T}\mathbf{g}_p(\boldsymbol{\beta}^{[a]}) + \frac{1}{2}\mathbf{e}^\mathsf{T}\boldsymbol{H}_p(\boldsymbol{\beta}^{[a]})\mathbf{e}\right\}$, $\mathbf{g}_p(\boldsymbol{\beta}^{[a]}) = \mathbf{g}(\boldsymbol{\beta}^{[a]}) - \mathbf{S}\boldsymbol{\beta}^{[a]}$ and $\boldsymbol{H}_p(\boldsymbol{\beta}^{[a]}) = \boldsymbol{H}(\boldsymbol{\beta}^{[a]}) - \mathbf{S}$. Vector $\mathbf{g}(\boldsymbol{\beta}^{[a]})$ is given by $\partial\ell(\boldsymbol{\beta})/\partial\boldsymbol{\beta}|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{[a]}}$, matrix $\boldsymbol{H}(\boldsymbol{\beta}^{[a]})$ by

$$\partial^2\ell(\boldsymbol{\beta})/\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^\mathsf{T}|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{[a]}},$$

$\|\cdot\|$ denotes the Euclidean norm, and $\Delta^{[a]}$ is the radius of the trust region which is adjusted through the iterations. Equation (10) uses a quadratic approximation of $-\ell_p$ about $\boldsymbol{\beta}^{[a]}$ (the so-called model function) in order to choose the best $\mathbf{e}^{[a+1]}$ within the ball centered in $\boldsymbol{\beta}^{[a]}$ of radius $\Delta^{[a]}$, the trust-region. Throughout the iterations, a proposed solution is accepted or rejected and the trust region adjusted (i.e., expanded or shrunken) based on the ratio between the improvement in the objective function when going from $\boldsymbol{\beta}^{[a]}$ to $\boldsymbol{\beta}^{[a+1]}$ and that predicted by the approximation. Note that, near the solution, the trust region method typically behaves as a classic Newton-Raphson unconstrained algorithm. For more details see (e.g., Nocedal & Wright, 2006, Chapter 4).

- *Smoothing step*: holding the model's parameter vector value fixed at $\boldsymbol{\beta}^{[a+1]}$, solve problem

$$\boldsymbol{\lambda}^{[a+1]} = \underset{\boldsymbol{\lambda}}{\arg\min}\ \|\mathbf{M}^{[a+1]} - \mathbf{A}^{[a+1]}\mathbf{M}^{[a+1]}\|^2 - n + 2\text{tr}(\mathbf{A}^{[a+1]}), \tag{11}$$

where, dropping the iteration index for simplicity, $\mathbf{M} = \boldsymbol{\mu}_\mathbf{M} + \boldsymbol{\epsilon}$, $\boldsymbol{\mu}_\mathbf{M} = \sqrt{-\boldsymbol{H}}\boldsymbol{\beta}$, $\boldsymbol{\epsilon} = \sqrt{-\boldsymbol{H}}^{-1}\mathbf{g}$ and $\mathbf{A} = \sqrt{-\boldsymbol{H}}\left(-\boldsymbol{H} + \mathbf{S}\right)^{-1}\sqrt{-\boldsymbol{H}}$. It can be proved that (11) is approximately equivalent to the Akaike information criterion (AIC). This means that $\boldsymbol{\lambda}$ is estimated by minimising what is effectively the AIC with number of parameters given by $\text{tr}(\mathbf{A})$. The above step is implemented adapting to the current context the routine by Wood (2004), which is based on Newton's method and can evaluate in an efficient and stable way the

components in (11) and their first and second derivatives with respect to $\log(\boldsymbol{\lambda})$ (since the smoothing parameters can only take positive values).

The two steps are iterated until the algorithm satisfies the criterion $\frac{\left|\ell(\boldsymbol{\beta}^{[a+1]})-\ell(\boldsymbol{\beta}^{[a]})\right|}{0.1+\left|\ell(\boldsymbol{\beta}^{[a+1]})\right|} < 1e-07$, and convergence is assessed by checking that the maximum of the absolute value of the score vector is numerically equivalent to 0 and that the observed Information matrix is positive definite. Reliable starting values are obtained by combining the use of the stable and efficient shape constrained smoothing approach of Pya & Wood (2015), implemented through the `scam` `R` package, with the procedure detailed in Section 2.3.1 of Liu et al. (2018).

We would like to point out that preliminary experimentation using, for instance, classical quasi-Newton and Newton methods revealed that estimation performance and convergence are not often satisfactory. As an example, we found that the Hessian is poorly approximated by numerical differentiation techniques, which was not surprising given the definition of (8). Trust region algorithms are generally more stable and faster compared to in-line search methods. The latter use the quadratic model of the objective function to find a search direction and suitable step lengths along such direction, whereas the former search the step that minimises the objective function within a previously defined region around the current iterate. If a function exhibits long plateaus and the current iterate is in that region, line search methods may search the next step far away from the current iterate in which case it may be possible that the evaluation of the log likelihood will not be finite. Instead, trust region methods define a maximum distance based on the trust region before evaluating the objective function. This is convenient because the new iterate will not lie too far away from the current one, and in the case of a non-definite evaluation, the proposed step will not be accepted. If a candidate which minimises the quadratic model and that also lies in the trust region does not improve the function sufficiently or gives a non-definite evaluation, the trust region will shrink and the algorithm will move back to the previous step. If the improvement is large enough, the trust region will expand in the next iteration.

As mentioned earlier, the non-linear dependence of $\mathbf{f}(\boldsymbol{\beta})$ on $\boldsymbol{\beta}$ makes the estimation problem more challenging which could in turn lead to numerical instabilities. Another potential issue is that the smoothing step neglects the dependence of the Hessian on the smoothing parameter vec-

tor; this will vanish asymptotically, but at finite sample sizes it may not be non-negligible. A smoothing approach addressing the latter issue has been proposed by Wood et al. (2016), however it requires computing the third and fourth order derivatives; this is daunting in the current context and generally such derivatives have to be implemented very carefully to avoid numerical instabilities. We found the estimation framework detailed in this section to work well in our simulation and case studies. Adaptations of the same framework have also been successfully utilised in different survival contexts (e.g., Dettoni et al., 2020; Marra & Radice, 2020a).

4 Further details

The number of effective degrees of freedom (edf) for a model containing only unpenalised terms is equal to say w, the dimension of $\boldsymbol{\beta}$, since in this case $\text{tr}(\mathbf{A}) = \text{tr}(\mathbf{I})$. The edf for a penalised model is $\text{tr}(\mathbf{A})$ which can also be written as $w - \text{tr}\left\{(-\boldsymbol{H} + \mathbf{S})^{-1}\mathbf{S}\right\}$. The latter shows the role of $\boldsymbol{\lambda}$ contained in $\mathbf{S}$; if $\boldsymbol{\lambda} \to \mathbf{0}$ then $\text{tr}(\mathbf{A}) \to w$ and if $\boldsymbol{\lambda} \to \boldsymbol{\infty}$ then $\text{tr}(\mathbf{A}) \to w - \zeta$, where ζ is the total number of model's parameters subject to penalisation. When $\mathbf{0} < \boldsymbol{\lambda} < \boldsymbol{\infty}$, the model's edf is equal to a value in the range $[w - \zeta, w]$. The edf of a single smooth or penalised term is given by the sum of the corresponding trace elements and has a value smaller than or equal to J_k.

As for the construction of intervals, it is convenient to take a Bayesian view of the model and employ at convergence the result $\boldsymbol{\beta} \dot{\sim} \mathcal{N}(\hat{\boldsymbol{\beta}}, \mathbf{V}_{\boldsymbol{\beta}})$, where $\mathbf{V}_{\boldsymbol{\beta}} = -\boldsymbol{H}_p(\hat{\boldsymbol{\beta}})^{-1}$ (Wood et al., 2016). Intervals constructed using this approach exhibit close-to-nominal frequentist coverage probabilities since they account for both sampling variability and smoothing bias, an aspect that is particularly relevant at finite sample sizes. Since the evaluation of the additive predictor (as defined, for instance, by equation (5)) and the quantities that rely on it (e.g., equation (2)) depend on $\mathbf{f}(\boldsymbol{\beta})$, it makes sense to obtain the relevant distribution, which, following Pya & Wood (2015), is $\mathbf{f}(\boldsymbol{\beta}) \dot{\sim} \mathcal{N}(\mathbf{f}(\hat{\boldsymbol{\beta}}), \mathbf{V}_{\mathbf{f}(\boldsymbol{\beta})})$, where $\mathbf{V}_{\mathbf{f}(\boldsymbol{\beta})} = \text{diag}\,(\mathbf{E})\,\mathbf{V}_{\boldsymbol{\beta}}\,\text{diag}\,(\mathbf{E})$. This is worked out by using a Taylor series expansion of $\mathbf{f}(\boldsymbol{\beta})$, i.e. $\mathbf{f}(\boldsymbol{\beta}) - \mathbf{f}(\hat{\boldsymbol{\beta}}) \approx \text{diag}\,(\mathbf{E})\left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\right)$, which shows that $\mathbf{f}(\boldsymbol{\beta}) - \mathbf{f}(\hat{\boldsymbol{\beta}})$ is approximately a linear function of $\boldsymbol{\beta}$. Recalling that linear functions of normally distributed random variables follow normal distributions, the result follows. P-values for the smooth components in the model are derived by adapting the results discussed in Wood (2017) and using $\mathbf{V}_{\mathbf{f}(\boldsymbol{\beta})}$

as covariance matrix.

Intervals for linear functions of the model's coefficients (such as smooth components) are obtained using the above result for $\mathbf{f}(\boldsymbol{\beta})$. Intervals for non-linear functions of the model's coefficients can instead be conveniently obtained by posterior simulation, hence avoiding computationally expensive parametric bootstrap or frequentist approximations. As an example, if we are interested in obtaining intervals for (2) then we need to obtain a number of simulated vectors for $\mathbf{f}(\boldsymbol{\beta})$ and for each of them evaluate (2). These evaluations are then used to construct intervals.

5 Simulation study

This section provides evidence on the empirical effectiveness of the proposed methodology in recovering true covariate effects and baseline functions, in the presence of all types of censoring and of linear and non-linear effects. The data generating process (DGP) described below has been designed to mimic some of the features of the results of the case study discussed in the next section. For instance, the chosen baseline function, values for β_1 and β_2, and shape and magnitude of one of the two smooth functions are in line with the empirical findings. In the DGP, however, we included an extra smooth function to make the estimation problem more challenging. We have not considered potential competitors in our study because, to the best of our knowledge, there are no alternative implementations capable of handling mixed censoring, of flexibly estimating the (linear or non-linear) shapes of the baseline function and covariate effects, and based on a fast and stable automatic multiple smoothing parameter selection approach.

The exact survival time T_i was generated from a proportional hazard model defined, on the survival function scale, as $\log\left[-\log S_0(t_i)\right] + \beta_1 z_{1i} + \beta_2 z_{2i} + s_1(z_{3i}) + s_2(z_{4i})$, where $S_0(t_i) = 0.7\exp(-0.03t_i^{1.8}) + 0.3\exp(-0.3t_i^{2.5})$, $\beta_1 = 1.3$, $\beta_2 = 0.5$, $s_1(x) = -0.075\exp(3.2x)$ and $s_2(x) = \sin(2\pi x)$. A very similar definition of $S_0(t_i)$ has previously been adopted by Liu et al. (2018) on the basis of biological plausibility of the underlying distribution. Correlated covariates were generated using a multivariate standard Gaussian with correlation parameters set at 0.5 and then transformed using the distribution function of a standard Gaussian (e.g., Marra & Radice, 2020a). Covariates z_{1i} and z_{2i} were dichotomised by simply rounding them. Observations were

generated using the Brent's univariate root-finding method. The values of L_i and R_i were determined through a visit process. Let U denote the uniform distribution. Two visits were simulated such that $V_1 \sim U[0, 2]$ and $V_2 = V_1 + U[0, 6]$. Then observations for which $T_i < V_1$ were left-censored (with $L_i = 0$ and $R_i = V_1$), observations for which $T_i > V_2$ were right-censored (with $L_i = V_2$ and $R_i = \infty$), and observations for which $V_1 < T_i < V_2$ were interval-censored (with $L_i = V_1$ and $R_i = V_2$). Uncensored observations were generated by randomly assigning (with probability equal to 0.2) to left and interval censored observations the respective observed survival times. The sample size was set to 700 (in line with the size of the data-set used in the case study) and the number of replicates to 1000. To assess the effect that several proportions of censoring types have on the estimation results, we tried different simulation settings; the performance of the estimation method was very similar to that discussed in this section.

The models were fitted using function `gamlss()` in `GJRM` described in Appendix B. The smooth components of the continuous covariates were represented using the default penalised low rank thin plate splines with second order penalty and 10 bases (Wood, 2017). Note that we could have employed different spline definitions and related penalties (e.g., cubic regression splines and P-splines which are available in the package). As explained in Wood (2017), for uni-dimensional smooths of continuous covariates, the specific choice of spline definition will not have an impact on the estimated curve(s) as long as a reliable smoothing method is available for model fitting. As for the number of basis functions, the chosen value of 10 is arbitrary and based on the fact that it generally offers enough modelling flexibility in applications. However, a sensitivity analysis using more bases was attempted; there was no virtual change in the results but, as expected, the computing time increased. Regarding the smooth function of the time variable, we employed the monotonic penalised B-spline approach detailed in Section 3. For each replicate, curve estimates were constructed using 200 equally spaced fixed values in the $(0, 6)$ range for the monotonic function and $(0, 1)$ otherwise.

Figure 1 summarises the results. Considering the small sample size and complexity of the model, the true functions and linear effects are overall recovered well by the proposed estimation method. As the sample size increases (results not shown here) the estimates improve and their

variability decreases. Computing time for fitting the model was on average 5 seconds.

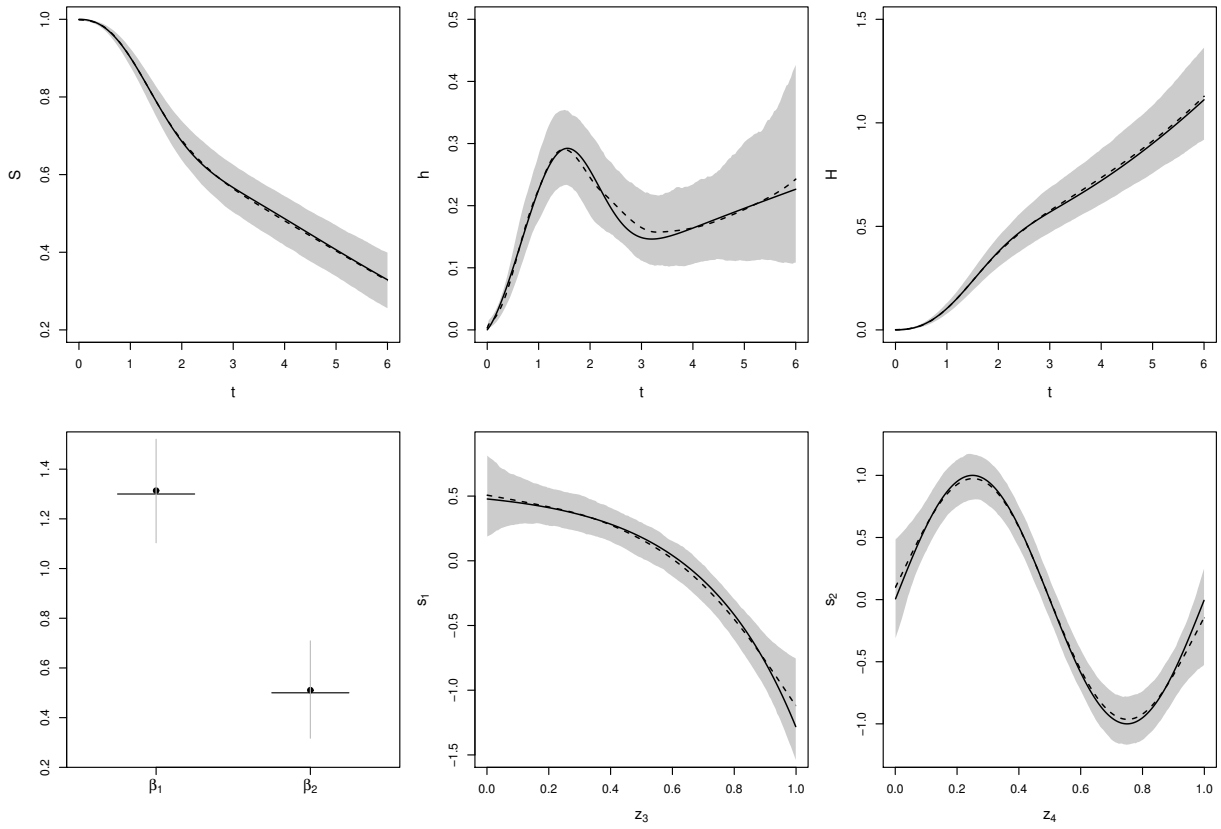

Figure 1: Smooth function and linear coefficient estimates obtained by applying `gamlss()` in `GJRM` to survival simulated data in the presence of all types of censoring. True functions are represented by black solid lines, mean estimates by dashed lines and point-wise ranges resulting from 5% and 95% quantiles by shaded areas. In the lower left plot, circles indicate mean estimates while gray bars represent the estimates' ranges resulting from 5% and 95% quantiles. True values are indicated by dashed horizontal lines. The first three (top) plots refer to the survival, hazard and cumulative hazard functions.

We also considered an alternative definition for s_2, namely $s_2(x) = 0.2x^{11}(10(1-x))^6 + 10(10x)^3(1-x)^{10}$. As far as the baseline smooth function of time, s_1 and the parametric components are concerned, we virtually obtained the same results as those in Figure 1. As for s_2, the result is reported in Figure 4, in Appendix C, which shows that overall the true function is recovered well by the estimation method.

6 Hospital-related risk assessment in cirrhotic patients

Our data example is about risk of in-hospital infection or death for cirrhotic patients not undergoing major surgery or at clear risk at admission. A retrospective study was conducted at the

Policlinico Umberto I hospital in Rome, reviewing data about patients admitted between January 2009 and March 2017. A total of $n = 678$ patients satisfy our inclusion criteria of (i) having a diagnosis of cirrhosis prior to hospitalisation, (ii) not having an infection or taking antibiotics at admission, (iii) not being hospitalised for major surgery (including, clearly, liver transplantation).

The endpoint of interest is composite; an event is defined as the occurrence of an infection or death before hospital discharge. We have 573 patients who were safely admitted, treated and discharged, therefore giving rise to right censored times. We also have 96 interval censored times for patients that have developed an infection during the hospital stay (whose precise onset is clearly impossible to measure), and 9 in-hospital deaths which give rise to uncensored events.

Follow-up times (before event or hospital discharge) range between 1 and 89 days, with a median of 7 days. Patients are on average 60.8$\pm$11.8 years old, 76% are males, 84 are staying in an extra bed, 138 recovered from an infection within the month prior to admission, and 361 have a history of alcohol abuse. The extra bed patients are those who are admitted without the availability of a bed in the ward, and hence hospitalised in the emergency room, or in a temporary bed set up in the corridors of the ward. During the hospital stay (for most patients within 48 hours of admission) 206 patients received a paracentesis procedure, 133 catheterisation of some sort, and 61 both procedures. Paracentesis is a procedure in which a needle is inserted into the peritoneal cavity to obtain ascitic fluid for diagnostic or therapeutic purposes, while catheterisation (usually at the level of the hepatic vein for these patients) involves the insertion of a catheter into a blood vessel. Both are routine procedures, whose associated risks in this patients' population should be carefully assessed. Finally, the patient status is summasized through the MELD score, ranging from 1 to 40 in our data, with a mean of 13.34 and a standard deviation of 5.21. MELD evaluates the severity of chronic liver disease and is also used to prioritize waiting lists for transplantation (where usually a MELD larger than fifteen points is an indication for entry into the waiting list).

Given that risk associated with MELD is well established in the literature, we estimate a model with two additive predictors for baseline assessment: $s(MELD)$ and $s(t_i)$. For the smooth function of time we also consider the option $s(\log(t_i))$. This typically helps producing a smoother fitted function which in turn reduces the chance of potential artifacts in the estimated hazard func-

tion, which may be especially relevant at low sample sizes (e.g., Royston & Parmar, 2002). Importantly, for outcomes characterised by marked changes in the values close to zero, employing the $\log$ transformation will considerably help modelling such patterns. Generally, we found that the $\log$ transform is often preferred in empirical applications. The chosen link function is PH. The log-transformation is preferred with a BIC of 1034.2 versus a BIC of 1059.9 for the identity transformation (using the AIC led to the same conclusion). The results of the chosen model are summarised in the upper panel of Figure 2, which clearly shows a non-linear effect of MELD (this will commented in more detail below). The number of basis functions for the smooth components was set to the default value of 10; increasing this value did not change the results. Overall we have p-values smaller than 5% for both additive terms.

We then build models with three predictors: the two additive components (one for MELD, one for time) from the base model, and each of the additional variables considered in the study. The estimated coefficient, standard error and p-value for each extra regressor in the model are reported in the left panel of Table 2.

Var	Coef	SE	p-value	Coef	SE	p-value
Age	0.002	0.008	0.771			
Gender	-0.202	0.226	0.372			
Paracentesis	0.578	0.206	0.005	0.415	0.209	0.047
Extra bed	1.408	0.229	< 0.001	1.288	0.229	< 0.001
Catheter	0.701	0.209	0.001	0.439	0.214	0.040
History	0.114	0.244	0.640			
Alcohol	-0.167	0.204	0.413			

Table 2: Cirrhotic patients data. Left panel: Estimated coefficient, standard error and p-value for the each regressor included in the base model (that includes a smoothed baseline risk and smoothed MELD). Right panel: Estimated coefficients, standard errors and p-values for linear effects included in the final multivariate model.

Finally, we select a multivariate model in a forward stepwise fashion. The final model includes additive components for the baseline risk and MELD, and linear effects of the indicator variables for paracentesis, overcrowding (extra bed) and catheterisation. (The same model is actually obtained through a different backward or stepwise selection algorithm.) Results are reported in the right panel of Table 2. In the lower panel of Figure 2, we report the estimated effects for the baseline hazard and MELD. The baseline hazard is slightly non-linear even on a log scale, increasing steeply at the beginning of follow-up but with a small but noticeable leveling up at around

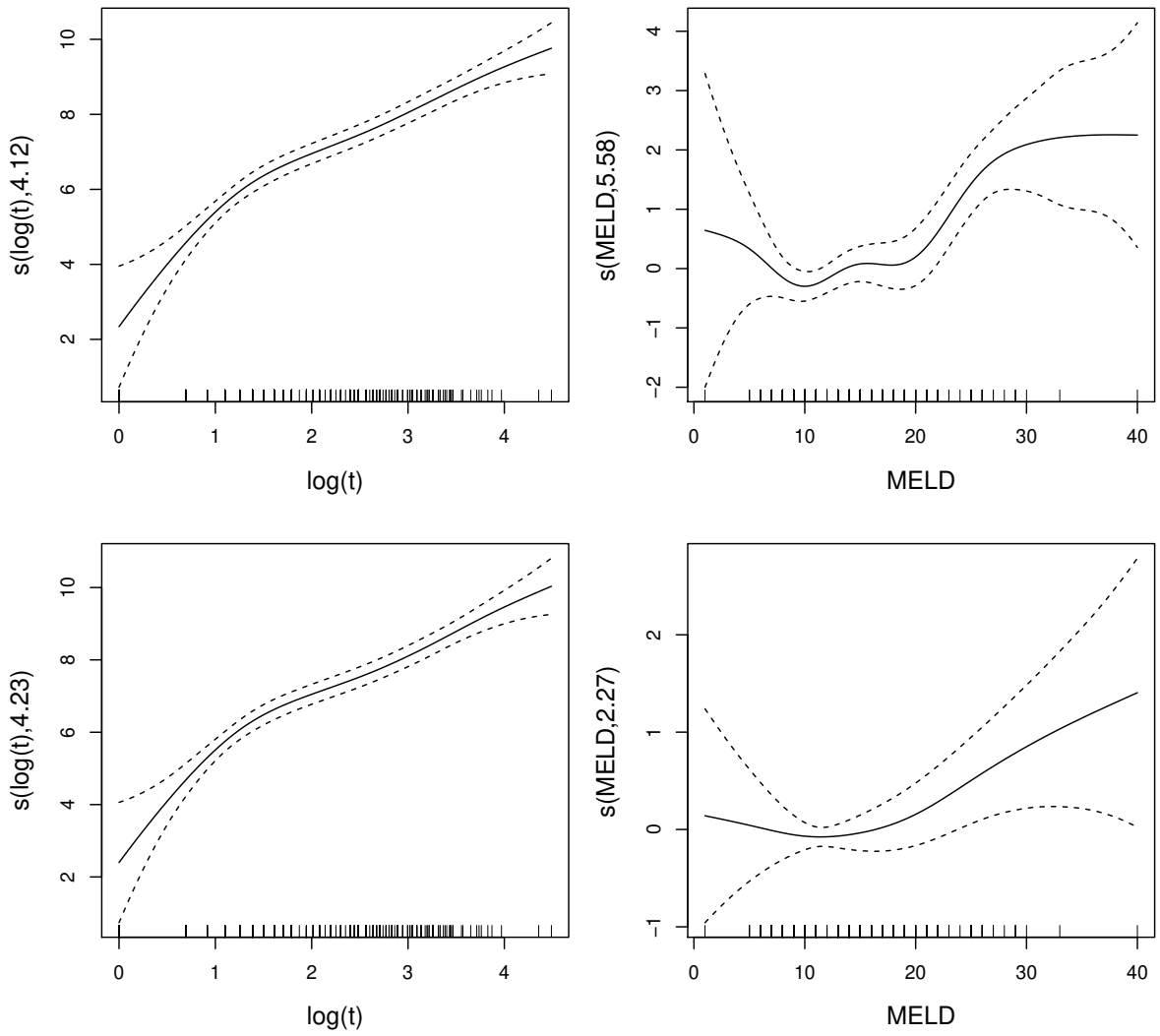

Figure 2: Cirrhotic patients data. Baseline risk and smoothed effect of MELD when used as the only predictors (upper panel), and in a multivariate model with additional linear effects for overcrowding, catheterisation, and paracentesis (lower panel). The 95% point-wise intervals are obtained as described in Section 4. The rug plot, at the bottom of each graph, shows the variables' values. The number in brackets in the y-axis caption represents the *edf* of the respective smooth curve.

$\exp(2) \cong 7$ days. This holds both when conditioning only on MELD (upper panel) and when additionally adjusting for important linear effects (lower panel). We can, therefore, conclude that the risk of events increases sharply during the hospital stay (as expected) but that the first week seems to be the most critical. The effect of MELD is also non-linear, but much smoother when the model adjusts for overcrowding and risky procedures. The sample size is comparatively high given our inclusion criteria, but still too small to draw strong conclusions. Nevertheless, it can be seen that for a MELD of up to around 15/20 points there is no risk differential, while after a certain threshold a sharp increase in risk is observed. This supports the use of MELD thresholds

for inclusion in waiting lists for liver transplantation, where 15 is clearly a good choice given that MELD will likely increase while in the waiting list. Finally, after non-parametrically adjusting for baseline risk and patient status (summarised by MELD), the variables overcrowding, paracentesis and catheterisation are found to be risk factors for the new onset of infections or death. The rationale for paracentesis and catheterisation is clear and might be also connected to a not perfect implementation of the procedure, while the high effect on risk of overcrowding is most likely linked to increased contacts among patients, and between patients and visitors.

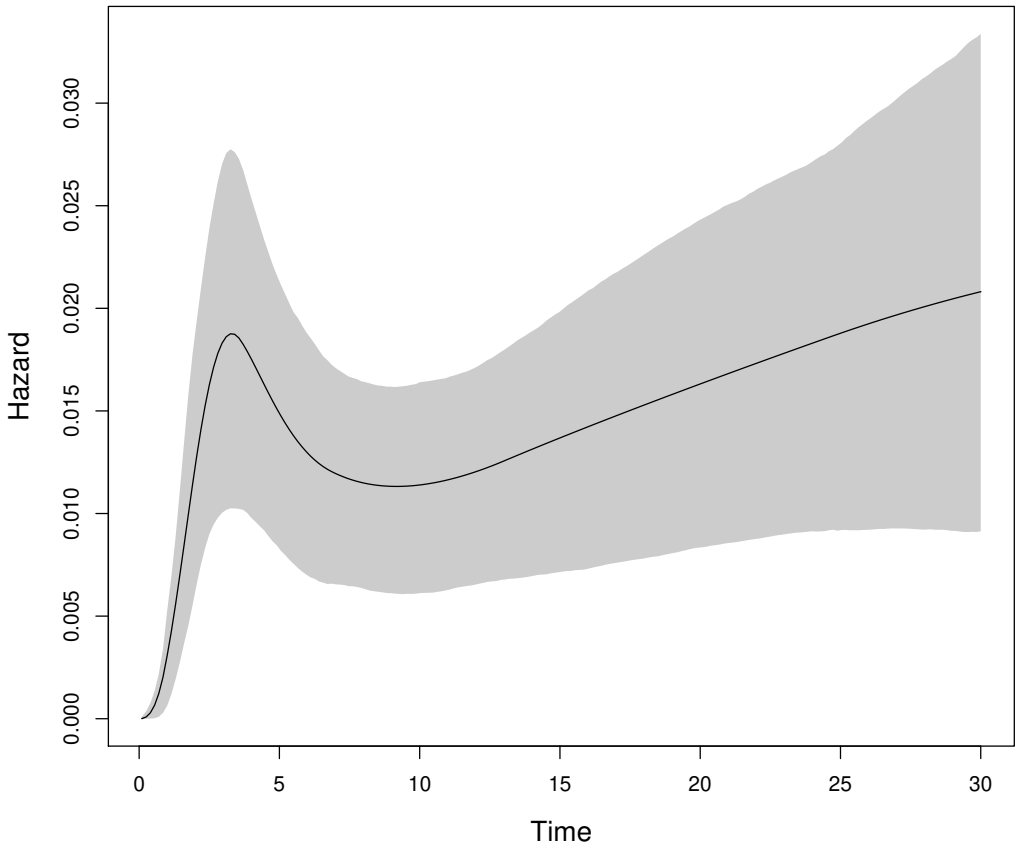

Figure 3: Cirrhotic patients data. Estimated hazard for a patient that underwent catheterisation, with a MELD equal to fifteen at admission. The 95% intervals have been obtained via posterior simulation using the approach described in Section 4.

To further illustrate the capabilities of our method, Figure 3 shows the estimated hazard for a patient having undergone catheterisation, with a MELD equal to fifteen at admission. As expected, the risk peaks within the first few days, it then flattens out, and subsequently slowly increases. Similar predictions can be carried out for different profiles.

Model	BIC
AFT-Weibull	1340.25
AFT-log-Normal	1336.01
Biased Cox	1091.04
Biased Cox-GAM	1099.31
PO	1013.05
Probit	1012.03
PH	995.74

Table 3: Cirrhotic patients data. BIC values obtained using the final multivariate specification when employing different modelling approaches: AFT (based on Weibull and log-Normal), a biased Cox regression, a biased Cox-GAM model, and the proposed approach based on PO, probit and PH links.

Finally, we compare the goodness of fit of our proposed model with that obtained using alternative and more classical approaches. Specifically, we considered parametric AFT models based on both the Weibull and log-Normal assumptions and that can handle mixed censoring, a (biased) Cox regression model that treats interval censored data as uncensored at R_i, and a similar biased Cox-GAM model which includes a non-parameteric additive effect of MELD. We also considered using different link function specifications for our proposed model. A further alternative is a (formally correct) Cox regression models in which interval censored data are treated as right censored at L_i. These are anyway not feasible with our data due to the presence of only nine deaths (uncensored events). Even after using a Firth's penalised likelihood, results are not credible for this specification. Table 3 reports the BIC for each approach, and our proposal with PH specification is clearly preferred. As a further consideration, treating the baseline risk semi-parametrically seems to be particularly important.

7 Concluding remarks

In many survival studies. mixed censoring (a situation where uncensored and left, right and interval censored observations mix together) may arise. There is, therefore, a strong need for theoretically founded, flexible and computationally efficient statistical methods for fitting survival models for this type of data. In this paper, we contributed in this direction by introducing link-based survival additive models under mixed censoring that can be fitted using a stable and efficient estimation approach. A clearly added value is the availability of analytic score and Hessian functions,

to estimate the model's coefficients and smoothing parameters, which make our implementation very convenient from a computational point of view. The inferential procedure is implemented in the accompanying `GJRM R` package.

The proposed approach performed well in simulation and has also been applied to an original data example on time to first hospital infection or in-hospital death in cirrhotic patients, showing that in real data scenarios ignoring the mixed nature of censoring or smooth non-linear effects might lead to lack of fit and bias. Code and data can be found at https://github.com/afarcome/GJRM.

Future research will focus on extending the modelling framework to the cases of: left truncation, excess hazard, multivariate response variables, complex survival outcomes including competing risks, multiple-events per subjects, and ultimately multi-state models. Joint modelling of survival and longitudinal outcomes should be a possible useful extension of our method. It would also be interesting to compare the performance of the proposed method versus the approach of Fauvernier et al. (2019) in the presence of time-dependent effects. In a similar vein as Liu et al. (2018), we also plan on extending the plotting function of the `GJRM R` package to include more conditional post-estimators based, for instance, on contrasts and various types of standardisations.

Acknowledgements

We would like to thank three reviewers, the associate editor, and the editor for their feedback which did help to clarify many aspects of the paper and strengthen its message, and Davide Lazzaro for contributing to the writing of the first paragraph of the Introduction and Remark 2. We also would like to thank Alessia Eletti for help with the writing of the analytical expressions of the score and Hessian reported in Appendix A.

Appendix A: Score and Hessian

This section contains the analytical expressions of the score and Hessian of the model's log-likelihood. Recall that the structure of equation (8) implies the presence of four main components,

that is

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} \delta_{Ui}\ell_{U_i} + \delta_{Li}\ell_{L_i} + \delta_{Ri}\ell_{R_i} + \delta_{Ii}\ell_{I_i}.$$

Exploiting this fact, the gradient and Hessian are reported according to the type of censoring considered to ease their readability. To simplify the notation, we use $S'_N\left\{\eta_i(t_i)\right\} = \partial S_N\left\{\eta_i(t_i)\right\}/\partial\eta_i(t_i)$, $S''_N\left\{\eta_i(t_i)\right\} = \partial^2 S_N\left\{\eta_i(t_i)\right\}/\partial\eta_i(t_i)^2$ and $S'''_N\left\{\eta_i(t_i)\right\} = \partial^3 S_N\left\{\eta_i(t_i)\right\}/\partial\eta_i(t_i)^3$, where t_i is adopted whenever the equality holds both for r_i as for l_i. To simplify the notation further, we also present the results for a single i^{th} observation.

Score

- Uncensored:

$$\frac{\partial}{\partial\boldsymbol{\beta}}\ell_{U_i}(\boldsymbol{\beta}) = \delta_{U_i}\left[S'_N\left\{\eta_i(l_i)\right\}\frac{\partial\eta_i(l_i)}{\partial l_i}\right]^{-1}\left[S''_N\left\{\eta_i(l_i)\right\}\frac{\partial\eta_i(l_i)}{\partial\boldsymbol{\beta}}\frac{\partial\eta_i(l_i)}{\partial l_i} + S'_N\left\{\eta_i(l_i)\right\}\frac{\partial^2\eta_i(l_i)}{\partial\boldsymbol{\beta}\partial l_i}\right].$$

- Left censoring:

$$\frac{\partial}{\partial\boldsymbol{\beta}}\ell_{L_i}(\boldsymbol{\beta}) = -\delta_{L_i}\left[1 - S_N\left\{\eta_i(l_i)\right\}\right]^{-1} S'_N\left\{\eta_i(l_i)\right\}\frac{\partial\eta_i(l_i)}{\partial\boldsymbol{\beta}}.$$

- Right censoring:

$$\frac{\partial}{\partial\boldsymbol{\beta}}\ell_{R_i}(\boldsymbol{\beta}) = \delta_{R_i}\left[S_N\left\{\eta_i(l_i)\right\}\right]^{-1} S'_N\left\{\eta_i(l_i)\right\}\frac{\partial\eta_i(l_i)}{\partial\boldsymbol{\beta}}.$$

- Interval censoring:

$$\frac{\partial}{\partial\boldsymbol{\beta}}\ell_{I_i}(\boldsymbol{\beta}) \quad = \delta_{I_i}\left[S_N\left\{\eta_i(l_i)\right\} - S_N\left\{\eta_i(r_i)\right\}\right]^{-1}\left[S'_N\left\{\eta_i(l_i)\right\}\frac{\partial\eta_i(l_i)}{\partial\boldsymbol{\beta}} - S'_N\left\{\eta_i(r_i)\right\}\frac{\partial\eta_i(r_i)}{\partial\boldsymbol{\beta}}\right].$$

Hessian

- Uncensored:

$$\frac{\partial^2}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T}\ell(\boldsymbol{\beta})_{U_i} =$$
$$\delta_{U_i}\left[-S'_N\left\{\eta_i(l_i)\right\}\right]^{-2}\left[S''_N\left\{\eta_i(l_i)\right\}^2\frac{\partial\eta_i(l_i)}{\partial\boldsymbol{\beta}}\left\{\frac{\partial\eta_i(l_i)}{\partial\boldsymbol{\beta}}\right\}^T - S'_N\left\{\eta_i(l_i)\right\}S'''_N\left\{\eta_i(l_i)\right\}\frac{\partial\eta_i(l_i)}{\partial\boldsymbol{\beta}}\left\{\frac{\partial\eta_i(l_i)}{\partial\boldsymbol{\beta}}\right\}^T\right]$$
$$+\delta_{U_i}\left[S'_N\left\{\eta_i(l_i)\right\}\right]^{-1}S''_N\left\{\eta_i(l_i)\right\}\frac{\partial^2\eta_i(l_i)}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T}$$
$$+\delta_{U_i}\left[-\frac{\partial\eta_i(l_i)}{\partial l_i}\right]^{-2}\frac{\partial^2\eta_i(l_i)}{\partial\boldsymbol{\beta}\partial l_i}\left\{\frac{\partial^2\eta_i(l_i)}{\partial\boldsymbol{\beta}\partial l_i}\right\}^T+\delta_{U_i}\left[\frac{\partial\eta_i(l_i)}{\partial l_i}\right]^{-1}\frac{\partial^3\eta_i(l_i)}{\partial\boldsymbol{\beta}^2\partial l_i}.$$

- Left censoring:

$$\frac{\partial^2}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T}\ell(\boldsymbol{\beta})_{L_i} = -\delta_{L_i}\left[1-S_N\left\{\eta_i(l_i)\right\}\right]^{-2}\left[S'_N\left\{\eta_i(l_i)\right\}\right]^2\frac{\partial\eta_i(l_i)}{\partial\boldsymbol{\beta}}\left\{\frac{\partial\eta_i(l_i)}{\partial\boldsymbol{\beta}}\right\}^T$$
$$-\delta_{L_i}\left[1-S_N\left\{\eta_i(l_i)\right\}\right]^{-1}S''_N\left\{\eta_i(l_i)\right\}\frac{\partial\eta_i(l_i)}{\partial\boldsymbol{\beta}}\left\{\frac{\partial\eta_i(l_i)}{\partial\boldsymbol{\beta}}\right\}^T$$
$$+\delta_{L_i}\left[1-S_N\left\{\eta_i(l_i)\right\}\right]^{-1}S'_N\left\{\eta_i(l_i)\right\}\left\{\frac{\partial^2\eta_i(l_i)}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T}\right\}^T.$$

- Right censoring:

$$\frac{\partial^2}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T}\ell(\boldsymbol{\beta})_{R_i} = -\delta_{R_i}\left[S_N\left\{\eta_i(l_i)\right\}\right]^{-2}\left[S'_N\left\{\eta_i(l_i)\right\}\right]^2\frac{\partial\eta_i(l_i)}{\partial\boldsymbol{\beta}}\left\{\frac{\partial\eta_i(l_i)}{\partial\boldsymbol{\beta}}\right\}^T$$
$$+\delta_{R_i}\left[S_N\left\{\eta_i(l_i)\right\}\right]^{-1}S''_N\left\{\eta_i(l_i)\right\}\frac{\partial\eta_i(l_i)}{\partial\boldsymbol{\beta}}\left\{\frac{\partial\eta_i(l_i)}{\partial\boldsymbol{\beta}}\right\}^T+\delta_{R_i}\left[S_N\left\{\eta_i(l_i)\right\}\right]^{-1}S'_N\left\{\eta_i(l_i)\right\}\left\{\frac{\partial^2\eta_i(l_i)}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T}\right\}^T.$$

- Interval censoring:

$$
\begin{aligned}
&\frac{\partial^2}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T}\ell(\boldsymbol{\beta})_{I_i} = -\delta_{I_i}\left[S_N\left\{\eta_i(l_i)\right\} - S_N\left\{\eta_i(r_i)\right\}\right]^{-2} \\
&\cdot\left[\left[S'_N\left\{\eta_i(l_i)\right\}\right]^2\frac{\partial\eta_i(l_i)}{\partial\boldsymbol{\beta}}\left\{\frac{\partial\eta_i(l_i)}{\partial\boldsymbol{\beta}}\right\}^T + \left[S'_N\left\{\eta_i(r_i)\right\}\right]^2\frac{\partial\eta_i(r_i)}{\partial\boldsymbol{\beta}}\left\{\frac{\partial\eta_i(r_i)}{\partial\boldsymbol{\beta}}\right\}^T\right. \\
&\left.- S'_N\left\{\eta_i(l_i)\right\}S'_N\left\{\eta_i(r_i)\right\}\left[\frac{\partial\eta_i(l_i)}{\partial\boldsymbol{\beta}}\left\{\frac{\partial\eta_i(r_i)}{\partial\boldsymbol{\beta}}\right\}^T + \frac{\partial\eta_i(r_i)}{\partial\boldsymbol{\beta}}\left\{\frac{\partial\eta_i(l_i)}{\partial\boldsymbol{\beta}}\right\}^T\right]\right] \\
&+ \delta_{I_i}\left[S_N\left\{\eta_i(l_i)\right\} - S_N\left\{\eta_i(r_i)\right\}\right]^{-1} \\
&\cdot\left[S''_N\left\{\eta_i(l_i)\right\}\frac{\partial\eta_i(l_i)}{\partial\boldsymbol{\beta}}\left\{\frac{\partial\eta_i(l_i)}{\partial\boldsymbol{\beta}}\right\}^T + S'_N\left\{\eta_i(l_i)\right\}\left\{\frac{\partial^2\eta_i(l_i)}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T}\right\}^T\right. \\
&\left.- S''_N\left\{\eta_i(r_i)\right\}\frac{\partial\eta_i(r_i)}{\partial\boldsymbol{\beta}}\left\{\frac{\partial\eta_i(r_i)}{\partial\boldsymbol{\beta}}\right\}^T - S'_N\left\{\eta_i(r_i)\right\}\left\{\frac{\partial^2\eta_i(r_i)}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T}\right\}^T\right].
\end{aligned}
$$

Appendix B: The R GJRM package

Link-based additive survival models with mixed censoring can be fitted using function `gamlss()` in the `GJRM` R package (Marra & Radice, 2020b). The `gamlss()` function is generally very easy to use, especially if the user is already familiar with the syntax of (generalised) linear and additive models in `R`. An example of call is

```
eq  <- list(t ~ s(log(t), bs = "mpi") + z1 + s(z2))
out <- gamlss(eq, data = dataset, surv = TRUE, margin = "PH",
              cens = cens, type.cens = "mixed", upperB = "t2")
```

where `t` is a survival variable with censoring mixed indicator `cens` (made up of four possible categories: `I` which stands for interval, `L` for left, `R` for right, and `U` for uncensored), and `z1` and `z2` are (e.g., binary and continuous) covariates. Variable `t2` is only used when interval censored observations are present in the `dataset`; in this case the intervals' upper bound values are required and the variable name of the upper bound has to be provided via argument `upperB`. Argument `surv` must be set to `TRUE` in order to employ a survival model. Argument `margin` of `gamlss()` in `GJRM` allows the user to employ the desired link function and the possible choices are given in Table 1; for example, `margin = "PH"` returns a proportional hazards model. Given

the modularity of our implementation, other link function specifications, such as those belonging to the Aranda-Ordaz family as described by Royston & Parmar (2002), can be considered. `eq` contains the equation of interest. Symbol `s()` stands for smooth function. As in `mgcv`, the default spline basis is `bs = "tp"` (penalised low rank thin plate spline) with `k = 10` (number of basis functions) and `m = 2` (order of derivatives). However, argument `bs` can also be set to, for example, `cr` (penalised cubic regression spline), `ps` (P-spline) and `mrf` (Markov random field), to name but a few. It is important to note that `bs` must be set to `mpi` (monotonic P-spline) for the baseline smooth of time. Model `summary()` and `plot()` functions work in a similar fashion as those of generalised linear and additive models, and `AIC()` and `BIC()` can be used in the usual manner. Function `hazsurv.plot()` allows the user to produce, post-estimation, hazard and survival plots. More details and options can be found in the documentation of the `GJRM R` package.

Appendix C: Further simulation results

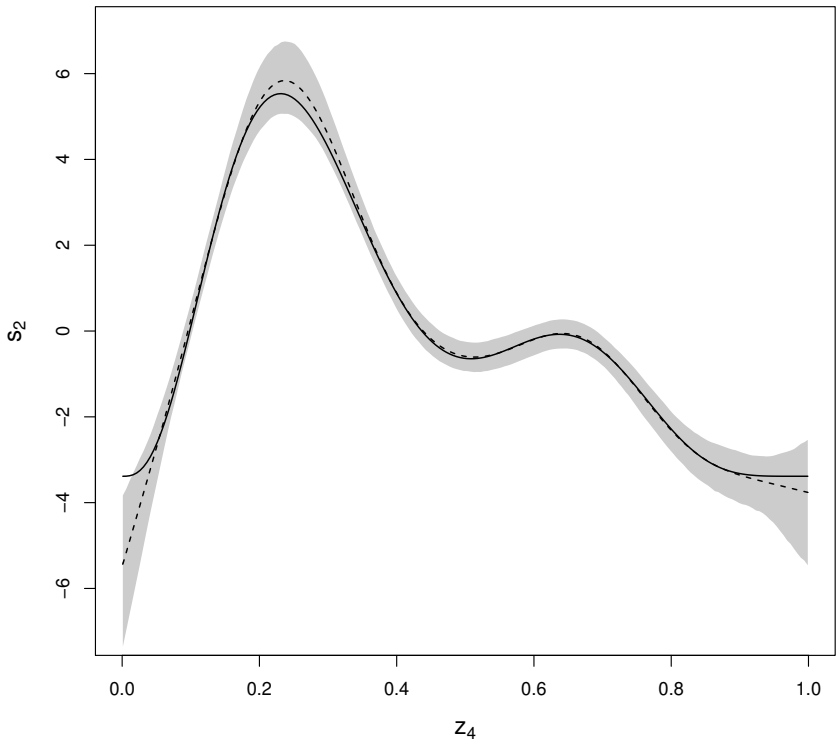

Figure 4: Results from the experiment that uses an alternative definition for s_2. The true function is represented by the black solid line, the mean estimate by the dashed line and poin-twise ranges resulting from 5% and 95% quantiles by the shaded area.

References

Anderson-Bergman, C. (2017). `icenReg`: Regression models for interval censored data in `r`. *Journal of Statistical Software*, 81(12).

Bartoletti, M., Giannella, M., Lewis, R., Caraceni, P., Tedeschi, S., Paul, M., Schramm, C., Bruns, T., Merli, M., et al. (2018). A prospective multicentre study of the epidemiology and outcomes of bloodstream infection in cirrhotic patients. *Clinical Microbiology and Infection*, 24(5), 546.e1 – 546.e8.

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

Dettoni, R., Marra, G., & Radice, R. (2020). Generalized link-based additive survival models with informative censoring. *Journal of Computational and Graphical Statistics*.

Fauvernier, M., Roche, L., Uhry, Z., Tron, L., Bossard, N., Remontet, L., & Challenges in the Estimation of Net Survival Working Survival Group (2019). Multi-dimensional penalized hazard model with continuous covariates: applications for studying trends and social inequalities in cancer survival. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 68(5), 1233–1257.

Fleming, T. R., Rothmann, M. D., & Hong, L. L. (2009). Issues in using progression-free survival when evaluating oncology products. *Journal of Clinical Oncology*, 27(17), 2874–2880.

Goggins, W., Finkelstein, D., Schoenfeld, D., & Zaslavksy, A. (1998). A Markov Chain Monte Carlo EM algorithm for analyzing interval-censored data under the Cox proportional hazards model. *Biometrics*, 54(4), 1498–1507.

Komarek, A., Lesaffre, E., & Hilton, J. (2005). Accelerated failure time model for arbitrarily censored data with smoothed error distribution. *Journal of Computational and Graphical Statistics*, 14(3), 726–745.

Leitenstorfer, F. & Tutz, G. (2007). Generalized monotonic regression based on B-splines with an application to air pollution data. *Biostatistics*, 8(3), 654–673.

Li, J. & Ma, J. (2019). Maximum penalized likelihood estimation of additive hazards models with partly interval censoring. *Computational Statistics & Data Analysis*, 137, 170–180.

Liu, X.-R., Pawitan, Y., & Clements, M. (2017). Generalized survival models for correlated time-to-event data. *Statistics in Medicine*, 36(29), 4743–4762.

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

Marra, G. & Radice, R. (2020a). Copula link-based additive models for right-censored event time data. *Journal of the American Statistical Association*, 115(530), 886–895.

Marra, G. & Radice, R. (2020b). *GJRM: Generalised Joint Regression Modelling*. R package version 0.2-2.

Merli, M., Lucidi, C., Di Gregorio, V., Falcone, M., Giannelli, V., Lattanzi, B., Giusto, M., Ceccarelli, G., Farcomeni, A., Riggio, O., & Venditti, M. (2015). The spread of multi drug resistant infections is leading to an increase in the empirical antibiotic treatment failure in cirrhosis: A prospective survey. *PLoS ONE*, 10(5), e0127448.

Nocedal, J. & Wright, S. J. (2006). *Numerical Optimization*. Springer-Verlag, New York.

Odell, P. M., Anderson, K. M., & D'Agostino, R. B. (1992). Maximum likelihood estimation for interval-censored data using a Weibull-based accelerated failure time model. *Biometrics*, 48(3), 951–959.

Piano, S., Singh, V., Caraceni, P., Maiwall, R., et al. (2019). Epidemiology and effects of bacterial infections in patients with cirrhosis worldwide. *Gastroenterology*, 156(5), 1368–1380.

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

R Development Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Royston, P. & Parmar, M. (2002). Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. *Statistics in Medicine*, 21(15), 2175–2197.

Satten, G. (1996). Rank-based inference in the proportional hazards model for interval censored data. *Biometrika*, 83(2), 355–370.

Schick, A. & Yu, Q. (2000). Consistency of the GMLE with mixed case interval-censored data. *Scandinavian Journal of Statistics*, 27(1), 45–55.

Sun, J. (2006). *The Statistical Analysis of Interval-Censored Data*. Berlin: Springer.

Szabo, Z., Liu, X., & Xiang, L. (2020). Semiparametric sieve maximum likelihood estimation for accelerated hazards model with interval-censored data. *Journal of Statistical Planning and Inference*, 205, 175–192.

Wang, L., McMahan, C., Hudgens, M., & Qureshi, Z. (2016). A flexible, computationally efficient method for fitting the proportional hazards model to interval-censored data. *Biometrics*, 72(1), 222–231.

Wood, S. N. (2004). Stable and efficient multiple smoothing parameter estimation for generalized additive models. *Journal of the American Statistical Association*, 99(467), 673–686.

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

Wood, S. N., Pya, N., & Säfken, B. (2016). Smoothing parameter and model selection for general smooth models. *Journal of the American Statistical Association*, 111(516), 1548–1563.

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

Zhang, Z. & Sun, J. (2010). Interval censoring. *Statistical Methods in Medical Research*, 9(1), 53–70.