Copula Link-Based Additive Models for Right-Censored Event Time Data

Journal:	*Journal of the American Statistical Association*
Manuscript ID	JASA-T&M-2018-0237.R3
Manuscript Type:	Article – Theory & Methods
Keywords:	additive predictor, bivariate survival data, copula, joint model, link function

Copula Link-Based Additive Models for Right-Censored Event Time Data

2019-03-03

Abstract

This article proposes an approach to estimate and make inference on the parameters of copula link-based survival models. The methodology allows for the margins to be specified using flexible parametric formulations for time-to-event data, the baseline survival functions to be modeled using monotonic splines, and each parameter of the assumed joint survival distribution to depend on an additive predictor incorporating several types of covariate effects. All the model's coefficients as well as the smoothing parameters associated with the relevant components in the additive predictors are estimated using a carefully structured efficient and stable penalized likelihood algorithm. Some theoretical properties are also discussed. The proposed modeling framework is evaluated in a simulation study and illustrated using a real data set. The relevant numerical computations can be easily carried out using the freely available `GJRM R` package.

Key Words: additive predictor, bivariate survival data, copula, joint model, link function, penalized log-likelihood, regression spline representation, simultaneous parameter estimation.

1 Introduction

Bivariate survival data consist of pairs of event times which may be right-censored and exhibit strong association, and are often encountered in biomedical studies. Applications utilizing such type of data include the study of Danish twin pairs (Wienke et al., 2003), the association of age at a marker event and age at menopause (Nan et al., 2006), and the dependence between time to myocardial infarction and time to stroke (Li et al., 2017). Copulae are well-suited to build bivariate models for survival outcomes since they can flexibly link marginal survival functions to form a joint survival distribution. Their use in survival analysis dates back to Clayton (1978), Oakes (1982) and Oakes (1986), and there have been a number of recent methodological developments in the area (e.g., Prenen et al., 2017; Romeo et al., 2018). While other frameworks can be adopted to analyze jointly event times (e.g., frailty and scale change models), the copula approach offers a good deal of flexibility in specifying the model and is usually computationally more tractable.

Clayton (1978) suggested that, when adjusting for covariates, the marginal survival functions as well as the copula dependence parameter can help uncover the presence of underlying factors influencing the probability of event times simultaneously. However, the majority of the articles published since then have mainly focused on controlling for covariates at the marginal level, hence neglecting the inclusion of covariate information in the association structure of the event times. The works by Bogaerts & Lesaffre (2008), Geerdens et al. (2018), Meyer & Romeo (2015) and Romeo et al. (2018) (see also the relevant references therein) have addressed this issue in copula models with several types of survival margins.

In this work, we contribute in this direction by developing an efficient and theoretically founded estimation and inferential likelihood-based framework for fitting flexible copula survival models for right-censored bivariate survival data. The proposed methodology allows for the simultaneous estimation of all the parameters of the assumed joint survival distribution. Moreover, each parameter can depend on an additive predictor incorporating a vast variety of covariate effects that are represented using the penalized regression spline approach (Wood, 2017). The margins are modeled via transformations of the survival functions, which, when combined with the use of additive predictors, give rise to marginal generalized additive survival or link-based models

(e.g., Liu et al., 2018; Royston & Parmar, 2002). These can essentially be regarded as flexible parametric model formulations for time-to-event data where transformations of the baseline survival functions can be flexibly modeled using for example B-splines, and the covariate effects are determined via additive predictors. It is important to note that working with transformations of the survival functions avoids the need for numerical integration (to evaluate, for instance, the cumulative hazard function), and that time-varying covariate effects can be easily accounted for (Royston & Parmar, 2002). Cox has encouraged the broader use of parametric survival models for empirical modeling (Reid, 1994). In fact, they facilitate model estimation and comparison, easily allow for the visualization of the estimated baseline hazard and survival functions, and allow us to calculate several quantities of interest and their variances, such as time-dependent hazard or odds ratios, which would otherwise be more difficult to obtain with a non-parametric approach (Hjort, 1992). The smoothing parameters associated with the spline components in the model's additive predictors are efficiently estimated from the data using a general and automatic approach.

The challenge with flexibly estimating transformations of the baseline survival functions is that they must be monotone in the time variables. In our view, this problem is best theoretically and computationally addressed using the monotonic P-spline approach introduced by Pya & Wood (2015). Alternative techniques make it difficult to efficiently and/or reliably estimate a vector of smoothing parameters in a shape constrained context. For instance, methods based on subjecting the spline coefficients to linear inequality constraints (e.g., Meyer, 2012; Zhang, 2004) make the derivatives of classic smoothness criteria with respect to multiple smoothing parameters change discontinuously. This is because constraints enter or leave the set of active constraints during the optimization. Preliminary experimentation with one such approach revealed that derivative based fitting methods often fails, hence hindering the possibility of developing an efficient scheme for automatic multiple smoothing parameter estimation for joint survival models.

It may be argued that using a two-stage estimation approach instead of a simultaneous one would, for example, make the fitting problem easier to deal with in exchange for some loss in efficiency. However, as argued and illustrated via simulation in the paper, the simultaneous method exhibits a superior performance in the context of the models developed in this paper.

To summarize, the proposed framework allows one to estimate joint survival models where

two flexible parametric survival models are linked by a copula function, all the model's parameters can be specified as functions of various types of covariate effects, and monotonic P-splines of transformations of the baseline survival functions are utilized to provide coherent marginal survival fits. The estimation approach is based on penalized maximum likelihood and consists of a carefully constructed optimization scheme that allows for the simultaneous penalized estimation of the model's parameters as well as for stable and efficient automatic multiple smoothing parameter selection. The construction of confidence intervals for linear and non-linear functions of the model's coefficients is discussed, whereas p-values for the model's smooth components (which may, for example, be useful to test for the null hypothesis of dependence parameter constancy but not only) are obtained by adapting to the current context some of the results available in the spline literature. The new modeling framework has been implemented in the `R` package `GJRM` (Marra & Radice, 2019), which has been created to facilitate the use of such models in industry and academia and to enhance reproducible research.

The proposed model, estimation and inferential methods, and some theoretical properties are discussed in Section 2. Section 3 revisits a case study on appendectomy, whereas Section 4 provides a discussion. Details on smooth function specifications, large sample properties, software implementation, model building, and a simulation study are collected in the on-line supplementary material for the sake of space.

2 Methodology

We consider the case of bivariate right censored data; the true event times are not always recorded, in which case lower times (the censoring times) are observed. For individual i, let (C_{1i}, C_{2i}) denote a vector of bivariate censoring times which is assumed to be independent of the pair of survival times (T_{1i}, T_{2i}) conditional on a generic $\mathbf{x}_i$ (the vector of baseline covariates), and non-informative. We observe $(Y_{1i}, Y_{2i}) = (\min\{T_{1i}, C_{1i}\} \in \mathbb{R}^+, \min\{T_{2i}, C_{2i}\} \in \mathbb{R}^+)$ and the corresponding vector of censoring indicators $(u_{1i}, u_{2i}) = (I\{T_{1i} \leq C_{1i}\}, I\{T_{2i} \leq C_{2i}\})$. Let also $\boldsymbol{\delta} \in \mathbb{R}^W$ be a generic vector of parameters of dimension W, and $i = 1, 2, \ldots, n$ where n represents the sample size.

2.1 Model formulation

In this section, we introduce copula link-based additive survival models by describing the components that make them up and the assumptions they are based on. Let T_{1i} and T_{2i} have conditional marginal survival functions generically defined as $S_v(t_{vi}|\mathbf{x}_{vi};\boldsymbol{\beta}_v) = P(T_{vi} > t_{vi}|\mathbf{x}_{vi};\boldsymbol{\beta}_v) \in (0,1)$ for $v = 1,2$, and conditional joint survival function expressed as $S(t_{1i},t_{2i}|\mathbf{x}_i;\boldsymbol{\delta}) = P(T_{1i} > t_{1i}, T_{2i} > t_{2i}|\mathbf{x}_i;\boldsymbol{\delta})$. In order to link T_{1i} and T_{2i} we assume the copula model

$$S(t_{1i},t_{2i}|\mathbf{x}_i;\boldsymbol{\delta}) = C\left(S_1(t_{1i}|\mathbf{x}_{1i};\boldsymbol{\beta}_1), S_2(t_{2i}|\mathbf{x}_{2i};\boldsymbol{\beta}_2); m\left\{\eta_{3i}(\mathbf{x}_{3i};\boldsymbol{\beta}_3)\right\}\right),$$

where $\boldsymbol{\delta}^\mathsf{T} = (\boldsymbol{\beta}_1^\mathsf{T}, \boldsymbol{\beta}_2^\mathsf{T}, \boldsymbol{\beta}_3^\mathsf{T})$, $\mathbf{x}_{1i}$, $\mathbf{x}_{2i}$ and $\mathbf{x}_{3i}$ are vectors of covariates (which can all be equal to $\mathbf{x}_i$ but have not to) with associated coefficient vectors $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$ and $\boldsymbol{\beta}_3$ of dimensions W_1, W_2 and W_3 such that $W = W_1 + W_2 + W_3$, $C : (0,1)^2 \to (0,1)$ is a uniquely defined 2-dimensional copula function with coefficient $\theta_i = m\left\{\eta_{3i}(\mathbf{x}_{3i};\boldsymbol{\beta}_3)\right\}$ capturing the (possibly varying) conditional dependence of (T_{1i}, T_{2i}) across observations (e.g., Marra & Radice, 2017; Patton, 2002; Sklar, 1973), $\eta_{3i}(\mathbf{x}_{3i};\boldsymbol{\beta}_3) \in \mathbb{R}$ is a predictor which includes generic additive covariate effects, and m is an inverse monotonic and differentiable link function which ensures that the dependence parameter lies in its range (see Table 1). The margins are modeled using generalized survival or link-based models (Liu et al., 2018; Royston & Parmar, 2002). That is, $S_v(t_{vi}|\mathbf{x}_{vi};\boldsymbol{\beta}_v)$ is defined as $G_v\left\{\eta_{vi}(t_{vi},\mathbf{x}_{vi};\boldsymbol{\beta}_v)\right\}$, where G_v is an inverse link function and the additive predictors $\eta_{vi}(t_{vi},\mathbf{x}_{vi};\boldsymbol{\beta}_v) \in \mathbb{R}$, for $v = 1,2$, must include baseline functions of time (or a stratified set of functions of time) as conveyed by the notation. For the sake of clarity, the set up of the additive predictors will be discussed in detail in the next section. Except for some cases, it may not be straightforward to understand the magnitude of the association between T_{1i} and T_{2i} from the knowledge of θ. In such situation, the well known Kendall's τ, which takes values in the customary range $[-1,1]$, can be employed. The above construction shows that the copula framework allows one to create a joint survival function from the knowledge of (arbitrary) marginal survival functions and a function C that binds them together.

The copulae considered in this work are reported in Table 1. Counter-clockwise rotated versions of copulae such as Clayton and Gumbel can be obtained using the following expressions:

$C_{90} = p_2 - C(1-p_1, p_2)$, $C_{180} = p_1 + p_2 - 1 + C(1-p_1, 1-p_2)$, $C_{270} = p_1 - C(p_1, 1-p_2)$, where the subscript indicates the degree of rotation, p_1 and p_2 are margins and θ has been suppressed for simplicity (e.g., Brechmann & Schepsmeier, 2013). More details on copulae and their theoretical properties can be found in Nelsen (2006). Function $G_v\left\{\eta_{vi}(t_{vi}, \mathbf{x}_{vi}; \boldsymbol{\beta}_v)\right\}$ can be specified as shown in Table 2.

The marginal cumulative hazard and hazard functions, H_v and h_v ($v = 1, 2$), are given by

$$H_v(t_{vi}|\mathbf{x}_{vi}; \boldsymbol{\beta}_v) = -\log\left[G_v\left\{\eta_{vi}(t_{vi}, \mathbf{x}_{vi}; \boldsymbol{\beta}_v)\right\}\right]$$

and

$$h_v(t_{vi}|\mathbf{x}_{vi}; \boldsymbol{\beta}_v) = -\frac{G_v'\left\{\eta_{vi}(t_{vi}, \mathbf{x}_{vi}; \boldsymbol{\beta}_v)\right\}}{G_v\left\{\eta_{vi}(t_{vi}, \mathbf{x}_{vi}; \boldsymbol{\beta}_v)\right\}} \frac{\partial \eta_{vi}(t_{vi}, \mathbf{x}_{vi}; \boldsymbol{\beta}_v)}{\partial t_{vi}}, \tag{1}$$

where $G_v'\left\{\eta_{vi}(t_{vi}, \mathbf{x}_{vi}; \boldsymbol{\beta}_v)\right\} = \partial G_v\left\{\eta_{vi}(t_{vi}, \mathbf{x}_{vi}; \boldsymbol{\beta}_v)\right\} / \partial \eta_{vi}(t_{vi}, \mathbf{x}_{vi}; \boldsymbol{\beta}_v)$. The joint functions can be defined in a similar way.

Remark 1. Let us consider a copula with asymmetric dependence (e.g., Clayton and Gumbel), and express the joint survival function of (T_1, T_2) as $S(t_1, t_2) = C\left(S_1(t_1), S_2(t_2)\right)$. While $S(t_1, t_2)$ assumes strong upper (lower) tail association, the same copula function but with margins defined using cumulative distribution functions $1 - S_1(t_1)$ and $1 - S_2(t_2)$ assumes strong lower (upper) tail dependence. Note also that $C_{180}\left(S_1(t_1), S_2(t_2)\right)$ models the same dependence structure as $C\left(1 - S_1(t_1), 1 - S_2(t_2)\right)$.

2.1.1 Predictor specification

This section provides some details on the set up of the three model's predictors. The main difference between $\eta_{vi}(t_{vi}, \mathbf{x}_{vi}; \boldsymbol{\beta}_v)$ for $v = 1, 2$ and $\eta_{3i}(\mathbf{x}_{3i}; \boldsymbol{\beta}_3)$ is that the former two must include smooth functions of time. Apart from that, the design matrix set up is the same across the three additive predictors since t_{vi} can be treated as a regressor. Therefore, let us consider a generic $\eta_{\nu i}$ ($\nu = 1, 2, 3$), where the dependence on the covariates and parameters is momentarily dropped, and an overall covariate vector $\mathbf{z}_{\nu i}$ made up of $\mathbf{x}_{\nu i}$ as well as $t_{\nu i}$ when $\nu = 1, 2$. For simplicity, the dimensions of $\mathbf{z}_{1i}$ and $\mathbf{z}_{2i}$ are assumed to be W_1 and W_2 since t_{1i} and t_{2i} can be treated as covariates.

Copula	$C(p_1, p_2; \theta)$	Range of θ	Link	Kendall's τ
AMH (`"AMH"`)	$\frac{p_1 p_2}{1-\theta(1-p_1)(1-p_2)}$	$\theta \in [-1, 1]$	$\tanh^{-1}(\theta)$	$-\frac{2}{3\theta^2}\{\theta + (1-\theta)^2 \log(1-\theta)\} + 1$
Clayton (`"C0"`)	$\left(p_1^{-\theta} + p_2^{-\theta} - 1\right)^{-1/\theta}$	$\theta \in (0, \infty)$	$\log(\theta)$	$\frac{\theta}{\theta+2}$
FGM (`"FGM"`)	$p_1 p_2 \{1 + \theta(1-p_1)(1-p_2)\}$	$\theta \in [-1, 1]$	$\tanh^{-1}(\theta)$	$\frac{2}{9}\theta$
Frank (`"F"`)	$-\theta^{-1} \log\{1 + (\exp\{-\theta p_1\} - 1)(\exp\{-\theta p_2\} - 1)/(\exp\{-\theta\} - 1)\}$	$\theta \in \mathbb{R} \setminus \{0\}$	–	$1 - \frac{4}{\theta}[1 - D_1(\theta)]$
Gaussian (`"N"`)	$\Phi_2\left(\Phi^{-1}(p_1), \Phi^{-1}(p_2); \theta\right)$	$\theta \in [-1, 1]$	$\tanh^{-1}(\theta)$	$\frac{2}{\pi}\arcsin(\theta)$
Gumbel (`"G0"`)	$\exp\left[-\left\{(-\log p_1)^\theta + (-\log p_2)^\theta\right\}^{1/\theta}\right]$	$\theta \in [1, \infty)$	$\log(\theta - 1)$	$1 - \frac{1}{\theta}$
Joe (`"J0"`)	$1 - \left\{(1-p_1)^\theta + (1-p_2)^\theta - (1-p_1)^\theta(1-p_2)^\theta\right\}^{1/\theta}$	$\theta \in (1, \infty)$	$\log(\theta - 1)$	$1 + \frac{4}{\theta^2} D_2(\theta)$
Plackett (`"PL"`)	$\left(Q - \sqrt{R}\right) / \{2(\theta - 1)\}$	$\theta \in (0, \infty)$	$\log(\theta)$	–
Student-t (`"T"`)	$t_{2,\zeta}\left(t_\zeta^{-1}(p_1), t_\zeta^{-1}(p_2); \zeta, \theta\right)$	$\theta \in [-1, 1]$	$\tanh^{-1}(\theta)$	$\frac{2}{\pi}\arcsin(\theta)$

Table 1: Definition of the copulae implemented in `GJRM`, with corresponding parameter range of association parameter θ, link function of θ, and relation between Kendall's τ and θ. $\Phi_2(\cdot, \cdot; \theta)$ denotes the cumulative distribution function (cdf) of a standard bivariate normal distribution with correlation coefficient θ, and $\Phi(\cdot)$ the cdf of a univariate standard normal distribution. $t_{2,\zeta}(\cdot, \cdot; \zeta, \theta)$ indicates the cdf of a standard bivariate Student-t distribution with correlation θ and fixed $\zeta \in (2, \infty)$ degrees of freedom, and $t_\zeta(\cdot)$ denotes the cdf of a univariate Student-t distribution with ζ degrees of freedom. $D_1(\theta) = \frac{1}{\theta}\int_0^\theta \frac{t}{\exp(t)-1} dt$ is the Debye function and $D_2(\theta) = \int_0^1 t \log(t)(1-t)^{\frac{2(1-\theta)}{\theta}} dt$. Quantities Q and R are given by $1 + (\theta - 1)(p_1 + p_2)$ and $Q^2 - 4\theta(\theta - 1)p_1 p_2$, respectively. The Kendall's τ for `"PL"` is computed numerically as no analytical expression is available. Argument `BivD` of `gjrm()` in `GJRM` allows the user to employ the desired copula function and can be set to any of the values within brackets next to the copula names in the first column; for example, `BivD = "J0"`. For Clayton, Gumbel and Joe, the number after the capital letter indicates the degree of rotation required: the possible values are `0`, `90`, `180` and `270`.

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 2: Link functions implemented in `GJRM`. Argument `margins` of `gjrm()` in `GJRM` allows the user to employ the desired marginal models and can be set to any of the values within brackets next to the models' names in the first column; for example, `margins = c("PH", "PO")`. 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.

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 (Hastie & Tibshirani, 1990; Ruppert et al., 2003; Wood, 2017). However, note that the additive assumption here means that not all the interaction terms among the covariates may be included in the predictor (e.g., Wood, 2017).

An additive predictor can be defined as

$$\eta_{\nu i} = \beta_{\nu 0} + \sum_{k_\nu=1}^{K_\nu} s_{\nu k_\nu}(\mathbf{z}_{\nu k_\nu i}),\ i = 1, \ldots, n, \tag{2}$$

where $\beta_{\nu 0} \in \mathbb{R}$ is an overall intercept, $\mathbf{z}_{\nu k_\nu i}$ denotes the k_ν^{th} sub-vector of the complete vector $\mathbf{z}_{\nu i}$ and the K_ν functions $s_{\nu k_\nu}(\mathbf{z}_{\nu k_\nu i})$ represent generic effects which are chosen according to the type of covariate(s) considered. Each $s_{\nu k_\nu}(\mathbf{z}_{\nu k_\nu i})$ can be represented as a linear combination of $J_{\nu k_\nu}$ basis functions $b_{\nu k_\nu j_{\nu k_\nu}}(\mathbf{z}_{\nu k_\nu i})$ and regression coefficients $\beta_{\nu k_\nu j_{\nu k_\nu}} \in \mathbb{R}$, that is (e.g., Wood, 2017)

$$\sum_{j_{\nu k_\nu}=1}^{J_{\nu k_\nu}} \beta_{\nu k_\nu j_{\nu k_\nu}} b_{\nu k_\nu j_{\nu k_\nu}}(\mathbf{z}_{\nu k_\nu i}). \tag{3}$$

The above formulation implies that the vector of evaluations $\{s_{\nu k_\nu}(\mathbf{z}_{\nu k_\nu 1}), \ldots, s_{\nu k_\nu}(\mathbf{z}_{\nu k_\nu n})\}^\mathsf{T}$ can be written as $\mathbf{Z}_{\nu k_\nu}\boldsymbol{\beta}_{\nu k_\nu}$ with $\boldsymbol{\beta}_{\nu k_\nu} = (\beta_{\nu k_\nu 1}, \ldots, \beta_{\nu k_\nu J_{\nu k_\nu}})^\mathsf{T}$ and design matrix $\mathbf{Z}_{\nu k_\nu}[i, j_{\nu k_\nu}] = b_{\nu k_\nu j_{\nu k_\nu}}(\mathbf{z}_{\nu k_\nu i})$. This allows the predictor in equation (2) to be written as

$$\boldsymbol{\eta}_\nu = \beta_{\nu 0}\mathbf{1}_n + \mathbf{Z}_{\nu 1}\boldsymbol{\beta}_{\nu 1} + \ldots + \mathbf{Z}_{\nu K_\nu}\boldsymbol{\beta}_{\nu K_\nu}, \tag{4}$$

where $\mathbf{1}_n$ is an n-dimensional vector made up of ones. Equation (4) can also be written in a more compact way as $\boldsymbol{\eta}_\nu = \mathbf{Z}_\nu\boldsymbol{\beta}_\nu$, where $\mathbf{Z}_\nu = (\mathbf{1}_n, \mathbf{Z}_{\nu 1}, \ldots, \mathbf{Z}_{\nu K_\nu})$ and $\boldsymbol{\beta}_\nu = (\beta_{\nu 0}, \boldsymbol{\beta}_{\nu 1}^\mathsf{T}, \ldots, \boldsymbol{\beta}_{\nu K_\nu}^\mathsf{T})^\mathsf{T}$.

Each $\boldsymbol{\beta}_{\nu k}$ has an associated quadratic penalty $\lambda_{\nu k_\nu}\boldsymbol{\beta}_{\nu k_\nu}^\mathsf{T}\mathbf{D}_{\nu k_\nu}\boldsymbol{\beta}_{\nu k_\nu}$, used in fitting, whose role is to enforce specific properties on the k_ν^{th} function, such as smoothness. The 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 the estimate smooth function $\hat{s}_{\nu k_\nu}(\mathbf{z}_{\nu k_\nu i})$. The overall penalty can be defined as $\boldsymbol{\beta}_\nu^\mathsf{T}\mathbf{D}_\nu\boldsymbol{\beta}_\nu$, where $\mathbf{D}_\nu = \mathrm{diag}(0, \lambda_{\nu 1}\mathbf{D}_{\nu 1}, \ldots, \lambda_{\nu K_\nu}\mathbf{D}_{\nu K\nu})$. Finally, smooth functions are typically

subject to centering (identifiability) constraints (see Wood (2017) for more details).

The above formulation allows one to employ a rich variety of covariate effects; the reader is referred to Supplementary Material A for some examples of penalty and basis function specifications.

Remark 2. In some cases, like smooth functions of continuous covariates, quantity $J_{\nu k_\nu}$ has to be fixed to some value to make the computation feasible. Hence, the unknown $s_{\nu k_\nu}(\mathbf{z}_{\nu k_\nu i})$ may not have an exact representation as given in (3). In practical situations, $J_{\nu k_\nu}$ is usually set to an arbitrary value that allows for enough flexibility in estimating the smooth term. The coefficients of the spline basis are then penalized in the estimation process to suppress that part of the smooth term's complexity which is not supported by the data and that would lead to over-fitting.

Remark 3. Let us write $\eta_{vi}(t_{vi}, \mathbf{x}_{vi}; \boldsymbol{\beta}_v)$ as $\mathbf{Z}_{vi}(t_{vi}, \mathbf{x}_{vi})^\mathsf{T}\boldsymbol{\beta}_v$. Then

$$\frac{\partial \eta_{vi}(t_{vi}, \mathbf{x}_{vi}; \boldsymbol{\beta}_v)}{\partial t_{vi}} = \lim_{\varepsilon \to 0} \left\{ \frac{\mathbf{Z}_{vi}(t_{vi} + \varepsilon, \mathbf{x}_{vi}) - \mathbf{Z}_{vi}(t_{vi} - \varepsilon, \mathbf{x}_{vi})}{2\varepsilon} \right\}^\mathsf{T} \boldsymbol{\beta}_v = \mathbf{Z}'^\mathsf{T}_{vi}\boldsymbol{\beta}_v,$$

which is needed in equation (1). Depending on the type of spline basis employed $\mathbf{Z}'_{vi}$ can be calculated either by a finite-difference method or analytically.

Remark 4. To make the link between the marginal model defined by additive predictor (2) with link function $g(S)$ (as defined in Table 2) and the known proportional hazards and odds models, let us write each of the link-based marginal models as (Royston & Parmar, 2002)

$$g_v\left\{S_v(t_{vi}|\mathbf{x}_{vi})\right\} = g_v\left\{S_{v0}(t_{vi})\right\} + \sum_{k_v=2}^{K_v} s_{vk_v}(\mathbf{x}_{vk_vi}), \tag{5}$$

where $g_v : (0,1) \to (-\infty, \infty)$ is a differentiable and invertible link function (see Table 2) and $S_{v0}(t_{vi})$ is a background survival function. If we replace $g_v\left\{S_{v0}(t_{vi})\right\}$ with $s_{v0}(t_{vi})$ then the RHS of (5) becomes notationally consistent with (2). This passage allows us to see that $s_{v0}(t_{vi})$ is effectively modeling a transformation of the respective baseline survival function, exactly as in Royston & Parmar (2002). Continuing the reasoning, model (5) yields the proportional hazards

model when choosing the log-log link. That is,

$$\log\{H_v(t_{vi}|\mathbf{x}_{vi})\} = \log\{H_{v0}(t_{vi})\} + \sum_{k_v=2}^{K_v} s_{vk_v}(\mathbf{x}_{vk_vi}), \tag{6}$$

where $H_v(t_{vi}|\mathbf{x}_{vi}) = -\log\{S_v(t_{vi}|\mathbf{x}_{vi})\}$ and $H_{v0}(t_{vi}) = -\log\{S_{v0}(t_{vi})\}$ is the background cumulative hazard function. Important advantages of modeling on the log-cumulative hazard scale are that the corresponding function is more stable than the log-hazard function (which is advantageous when estimating the model), that quantities such as $h_v(t_{vi}|\mathbf{x}_{vi})$ and $S_v(t_{vi}|\mathbf{x}_{vi})$ can be directly obtained without the need for numerical integration, and that time-dependent effects can be easily incorporated in the model via terms like $s_{vk_v}(t_{vi})\mathbf{x}_{vk_vi}$. Moreover, given the parametric but flexible nature of the link-based marginal models employed here, the presence of ties in the outcomes will not be problematic. Note that when the RHS of (6) contains time-dependent effects, the model loses the proportional hazards interpretation. Model (5) yields the proportional odds model when the -logit link is chosen.

2.2 Penalised log-likelihood

Let us assume that a random *i.i.d.* sample $\{(y_{1i}, y_{2i}, u_{1i}, u_{2i}, \mathbf{x}_i)\}_{i=1}^n$ is available, 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

$$\begin{aligned}\ell(\boldsymbol{\delta}) = \sum_{i=1}^{n} & u_{1i}u_{2i}\log\left[\frac{\partial C\{G_1(\eta_{1i}), G_2(\eta_{2i}); \theta_i\}}{\partial G_1(\eta_{1i})\partial G_2(\eta_{2i})} G_1'(\eta_{1i}) G_2'(\eta_{2i}) \frac{\partial \eta_{1i}}{\partial y_{1i}} \frac{\partial \eta_{2i}}{\partial y_{2i}}\right] \\ & + u_{1i}(1-u_{2i})\log\left[-\frac{\partial C\{G_1(\eta_{1i}), G_2(\eta_{2i}); \theta_i\}}{\partial G_1(\eta_{1i})} G_1'(\eta_{1i}) \frac{\partial \eta_{1i}}{\partial y_{1i}}\right] \\ & + (1-u_{1i})u_{2i}\log\left[-\frac{\partial C\{G_1(\eta_{1i}), G_2(\eta_{2i}); \theta_i\}}{\partial G_2(\eta_{2i})} G_2'(\eta_{2i}) \frac{\partial \eta_{2i}}{\partial y_{2i}}\right] \\ & + (1-u_{1i})(1-u_{2i})\log\left[C\{G_1(\eta_{1i}), G_2(\eta_{2i}); \theta_i\}\right],\end{aligned} \tag{7}$$

where η_{vi} is the shorthand notation for $\eta_{vi}(y_{vi}, \mathbf{x}_{vi}; \boldsymbol{\beta}_v)$.

The first three lines of (7) involve $\partial\eta_{vi}/\partial y_{vi}$ $(v = 1, 2)$ which can be calculated using $\mathbf{z}_{vi}'^{\mathsf{T}}\boldsymbol{\beta}_v$ (as per **Remark** 3) and *must* be positive to ensure that the hazard functions are positive. To this

end we propose modeling the time effects using B-splines with coefficients constrained such that the resulting smooth functions of time are monotonically increasing. Specifically, let $s_v(y_{vi}) = \sum_{j_v=1}^{J_v} \gamma_{vj_v} b_{vj_v}(y_{vi})$, where the b_{vj_v} are B-spline basis functions of at least second order built over the interval $[a, b]$, based on equally spaced knots, and γ_{vj_v} are spline coefficients. A sufficient condition for $s'_v(y_{vi}) \geq 0$ over $[a, b]$ is that $\gamma_{vj_v} \geq \gamma_{vj_v-1}, \forall j$ (e.g., Leitenstorfer & Tutz, 2007). Such condition can be imposed by re-parametrizing the spline coefficient vector so that $\boldsymbol{\gamma}_v = \boldsymbol{\Sigma}_v \tilde{\boldsymbol{\beta}}_v$, where $\boldsymbol{\beta}_v^\mathsf{T} = (\beta_{v1}, \beta_{v2}, \ldots, \beta_{vJ_v})$, $\tilde{\boldsymbol{\beta}}_v^\mathsf{T} = \{\beta_{v1}, \exp(\beta_{v2}), \ldots, \exp(\beta_{vJ_v})\}$ and $\boldsymbol{\Sigma}_v[\iota_{v1}, \iota_{v2}] = 0$ if $\iota_{v1} < \iota_{v2}$ and $\boldsymbol{\Sigma}_v[\iota_{v1}, \iota_{v2}] = 1$ if $\iota_{v1} \geq \iota_{v2}$, with ι_{v1} and ι_{v2} denoting the row and column entries of the respective matrix. When setting up the penalty term we penalize the squared differences between adjacent β_{vj_v}, starting from β_{v2}, using $\mathbf{D}_v = \mathbf{D}_v^{*\mathsf{T}} \mathbf{D}_v^*$ where $\mathbf{D}_v^*$ is a $(J_v - 2) \times J_v$ matrix made up of zeros except that $\mathbf{D}_v^*[\iota_v, \iota_v + 1] = -\mathbf{D}_v^*[\iota_v, \iota_v + 2] = 1$ for $\iota_v = 1, ..., J_v - 2$ (Pya & Wood, 2015). Matrix $\boldsymbol{\Sigma}_v$ can be absorbed into $\mathbf{Z}_v$.

Our model specification allows for a high degree of flexibility in modeling data (see also **Remark** 2). If an unpenalized approach is employed to estimate $\boldsymbol{\delta}$ then the resulting smooth function estimates are likely to be unduly wiggly (e.g., Ruppert et al., 2003). To prevent over-fitting, we maximize

$$\ell_p(\boldsymbol{\delta}) = \ell(\boldsymbol{\delta}) - \frac{1}{2}\boldsymbol{\delta}^\mathsf{T}\mathbf{S}\boldsymbol{\delta}, \tag{8}$$

where ℓ_p is the penalized log-likelihood, $\mathbf{S} = \text{diag}(\mathbf{D}_1, \mathbf{D}_2, \mathbf{D}_3)$, $\mathbf{D}_1$, $\mathbf{D}_2$ and $\mathbf{D}_3$ are overall penalties which contain $\boldsymbol{\lambda}_1$, $\boldsymbol{\lambda}_2$, $\boldsymbol{\lambda}_3$ and $\boldsymbol{\lambda}_\nu = (\lambda_{\nu 1}, \ldots, \lambda_{\nu K_\nu})^\mathsf{T}$. The smoothing parameter vectors can be collected in the overall vector $\boldsymbol{\lambda} = (\boldsymbol{\lambda}_1^\mathsf{T}, \boldsymbol{\lambda}_2^\mathsf{T}, \boldsymbol{\lambda}_3^\mathsf{T})^\mathsf{T}$.

2.3 Estimation of $\boldsymbol{\delta}$

As it can be seen from (7), because of right-censoring, the log-likelihood function is made up of four main components. This makes the structure of the score vector and Hessian matrix more involved as compared to the case of no censoring. Such structure is considerably further complicated by the non-linear dependence of $\boldsymbol{\gamma}_v$ on the coefficients contained in $\boldsymbol{\beta}_v$ that correspond to the B-spline bases of y_{vi}, which creates the need to account for terms like $\partial^2 \eta_{vi}(y_{vi}, \mathbf{x}_{vi}; \boldsymbol{\beta}_v)/\partial y_{vi} \partial \boldsymbol{\beta}_v = \mathbf{z}_{vi}'^\mathsf{T} \mathbf{E}_v$ and $\partial \eta_{vi}(y_{vi}, \mathbf{x}_{vi}; \boldsymbol{\beta}_v)/\partial \boldsymbol{\beta}_v = \mathbf{z}_{vi}^\mathsf{T} \mathbf{E}_v$, where $\mathbf{E}_v$ is a vector such that $\mathbf{E}_v[vk_v j_{vk_v}] = 1$ if

$\tilde{\beta}_{vk_vj_{vk_v}} = \beta_{vk_vj_{vk_v}}$ and $\exp(\beta_{vk_vj_{vk_v}})$ otherwise. Furthermore, the non-linear dependence of $\boldsymbol{\gamma}_v$ on $\boldsymbol{\beta}_v$ makes the optimization problem more difficult than in the case of unconstrained B-spline coefficients.

Preliminary experimentation revealed that the use of various optimization schemes, including those based on derivative free and quasi-Newton methods, is generally problematic, even when using not very complex model specifications. For instance, we found that several gradient and Hessian components are poorly approximated by numerical differentiation techniques. To make the fitting problem easier to deal with, we also experimented with a two-stage estimation approach as often seen in several copula contexts. In this case, the estimation of the marginal models and of the copula function is carried out in two separate steps; the use of a two-stage algorithm resulted in inefficient and (on occasion) unstable computations as compared to the joint approach. Eventually, we opted for a simultaneous estimation approach based on fully analytical first and second order derivatives. In practice, this was implemented using a trust region algorithm which was found to be efficient and well suited for the problem at hand. Supplementary Material C provides some simulation-based evidence. Specifically, compare Figures 3 and 6 (simultaneous estimation approach) with Figures 5 and 8 (two-stage approach).

Holding $\boldsymbol{\lambda}$ fixed at a vector of values and for a given $\boldsymbol{\delta}^{[a]}$, where a is an iteration index, we maximize equation (8) using

$$\boldsymbol{\delta}^{[a+1]} = \boldsymbol{\delta}^{[a]} + \underset{\mathbf{e}:\|\mathbf{e}\|\leq\Delta^{[a]}}{\arg\min}\ \breve{\ell}_p(\boldsymbol{\delta}^{[a]}), \tag{9}$$

where $\breve{\ell}_p(\boldsymbol{\delta}^{[a]}) = -\left\{\ell_p(\boldsymbol{\delta}^{[a]}) + \mathbf{e}^\mathsf{T}\mathbf{g}_p(\boldsymbol{\delta}^{[a]}) + \frac{1}{2}\mathbf{e}^\mathsf{T}\boldsymbol{H}_p(\boldsymbol{\delta}^{[a]})\mathbf{e}\right\}$, $\mathbf{g}_p(\boldsymbol{\delta}^{[a]}) = \mathbf{g}(\boldsymbol{\delta}^{[a]}) - \mathbf{S}\boldsymbol{\delta}^{[a]}$ and $\boldsymbol{H}_p(\boldsymbol{\delta}^{[a]}) = \boldsymbol{H}(\boldsymbol{\delta}^{[a]}) - \mathbf{S}$. Vector $\mathbf{g}(\boldsymbol{\delta}^{[a]})$ consists of $\mathbf{g}_1(\boldsymbol{\delta}^{[a]}) = \partial\ell(\boldsymbol{\delta})/\partial\boldsymbol{\beta}_1|_{\boldsymbol{\beta}_1=\boldsymbol{\beta}_1^{[a]}}, \ldots, \mathbf{g}_3(\boldsymbol{\delta}^{[a]}) = \partial\ell(\boldsymbol{\delta})/\partial\boldsymbol{\beta}_3|_{\boldsymbol{\beta}_3=\boldsymbol{\beta}_3^{[a]}}$, the Hessian matrix has elements $\boldsymbol{H}(\boldsymbol{\delta}^{[a]})_{o,h} = \partial^2\ell(\boldsymbol{\delta})/\partial\boldsymbol{\beta}_o\partial\boldsymbol{\beta}_h^\mathsf{T}|_{\boldsymbol{\beta}_o=\boldsymbol{\beta}_o^{[a]},\boldsymbol{\beta}_h=\boldsymbol{\beta}_h^{[a]}}$ where $o, h = 1, 2, 3$, $\|\cdot\|$ denotes the Euclidean norm, and $\Delta^{[a]}$ is the radius of the trust region which is adjusted through the iterations. The first line of (9) uses a quadratic approximation of $-\ell_p$ about $\boldsymbol{\delta}^{[a]}$ (the so-called model function) in order to choose the best $\mathbf{e}^{[a+1]}$ within the ball centered in $\boldsymbol{\delta}^{[a]}$ of radius $\Delta^{[a]}$, the trust-region. Note that, near the solution, the trust region method typically behaves as a classic Newton-Raphson unconstrained algorithm (e.g., Nocedal & Wright, 2006, Chapter 4).

The expressions for $\mathbf{g}(\boldsymbol{\delta})$ and $\boldsymbol{H}(\boldsymbol{\delta})$ are very tedious (due to right-censoring and the non-linear dependence of $\boldsymbol{\gamma}_v$ on $\boldsymbol{\beta}_v$) and have been analytically and modularity derived for all choices reported in Tables 1 and 2. Modularity here means that it will be easy to extend our algorithm to other parametric copulae and marginal link functions.

2.4 Estimation of $\boldsymbol{\lambda}$

As argued in Marra et al. (2017), automatic multiple smoothing parameter estimation in the context of complex joint models is more successfully achieved if the smoothing criterion is based on $\mathbf{g}(\boldsymbol{\delta})$ and $\boldsymbol{H}(\boldsymbol{\delta})$. Here, we re-iterate the main ideas and remark some useful results.

For notational convenience, let us denote with $\mathbf{g}_p^{[a]}$, $\mathbf{g}^{[a]}$, $\boldsymbol{H}_p^{[a]}$ and $\boldsymbol{H}^{[a]}$ the shorthand notations for $\mathbf{g}_p(\boldsymbol{\delta}^{[a]})$, $\mathbf{g}(\boldsymbol{\delta}^{[a]})$, $\boldsymbol{H}_p(\boldsymbol{\delta}^{[a]})$ and $\boldsymbol{H}(\boldsymbol{\delta}^{[a]})$ defined in the previous section. We first need to express the parameter estimator in terms of gradient and Hessian, which is achieved as follows. A first order Taylor expansion of $\mathbf{g}_p^{[a+1]}$ about $\boldsymbol{\delta}^{[a]}$ yields $\mathbf{0} = \mathbf{g}_p^{[a+1]} \approx \mathbf{g}_p^{[a]} + \left(\boldsymbol{\delta}^{[a+1]} - \boldsymbol{\delta}^{[a]}\right) \boldsymbol{H}_p^{[a]}$. We then have $\mathbf{0} = \mathbf{g}_p^{[a]} + \left(\boldsymbol{\delta}^{[a+1]} - \boldsymbol{\delta}^{[a]}\right) \left(\boldsymbol{H}^{[a]} - \mathbf{S}\right)$ which leads to $\boldsymbol{\delta}^{[a+1]} = \left(-\boldsymbol{H}^{[a]} + \mathbf{S}\right)^{-1} \sqrt{-\boldsymbol{H}^{[a]}}\mathbf{M}^{[a]}$, where $\mathbf{M}^{[a]} = \boldsymbol{\mu}_{\mathbf{M}}^{[a]} + \boldsymbol{\epsilon}^{[a]}$, $\boldsymbol{\mu}_{\mathbf{M}}^{[a]} = \sqrt{-\boldsymbol{H}^{[a]}}\boldsymbol{\delta}^{[a]}$ and $\boldsymbol{\epsilon}^{[a]} = \sqrt{-\boldsymbol{H}^{[a]}}^{-1}\mathbf{g}^{[a]}$. The square root of $-\boldsymbol{H}^{[a]}$ and its inverse are obtained by eigen-value decomposition. From likelihood theory, $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and $\mathbf{M} \sim \mathcal{N}(\boldsymbol{\mu}_{\mathbf{M}}, \mathbf{I})$, where $\mathbf{I}$ is an identity matrix, $\boldsymbol{\mu}_{\mathbf{M}} = \sqrt{-\boldsymbol{H}}\boldsymbol{\delta}^0$ and $\boldsymbol{\delta}^0$ is the true parameter vector. The predicted value vector for $\mathbf{M}$ is $\hat{\boldsymbol{\mu}}_{\mathbf{M}} = \sqrt{-\boldsymbol{H}}\hat{\boldsymbol{\delta}} = \mathbf{A}\mathbf{M}$, where $\mathbf{A} = \sqrt{-\boldsymbol{H}}\left(-\boldsymbol{H} + \mathbf{S}\right)^{-1}\sqrt{-\boldsymbol{H}}$. Our aim is to estimate $\boldsymbol{\lambda}$ so that the smooth terms' complexity which is not supported by the data is suppressed. Therefore, we use the following criterion

$$\mathbb{E}\left(\|\boldsymbol{\mu}_{\mathbf{M}} - \hat{\boldsymbol{\mu}}_{\mathbf{M}}\|^2\right) = \mathbb{E}\left(\|\mathbf{M} - \mathbf{A}\mathbf{M}\|^2\right) - \check{n} + 2\text{tr}(\mathbf{A}), \tag{10}$$

where $\check{n} = 3n$ and $\text{tr}(\mathbf{A})$ is the number of effective degrees of freedom (edf) of the penalized model. In practice, $\boldsymbol{\lambda}$ is estimated by minimizing an estimate of (10), i.e.

$$\widehat{\|\boldsymbol{\mu}_{\mathbf{M}} - \hat{\boldsymbol{\mu}}_{\mathbf{M}}\|^2} = \|\mathbf{M} - \mathbf{A}\mathbf{M}\|^2 - \check{n} + 2\text{tr}(\mathbf{A}). \tag{11}$$

The RHS of (11) depends on $\boldsymbol{\lambda}$ through $\mathbf{A}$ while $\mathbf{M}$ is associated with the un-penalized part of

the model. Note that (11) is approximately equivalent to the Akaike information criterion (AIC, Akaike, 1973), as shown at the end of this section. This means that $\boldsymbol{\lambda}$ is estimated by minimizing what is effectively the AIC with number of parameters given by $\text{tr}(\mathbf{A})$. Holding the model's parameter vector value fixed at $\boldsymbol{\delta}^{[a+1]}$, we solve problem

$$\boldsymbol{\lambda}^{[a+1]} = \arg\min_{\boldsymbol{\lambda}} \|\mathbf{M}^{[a+1]} - \mathbf{A}^{[a+1]}\mathbf{M}^{[a+1]}\|^2 - \check{n} + 2\text{tr}(\mathbf{A}^{[a+1]}), \tag{12}$$

using the automatic stable and efficient computational routine by Wood (2004). This approach is based on Newton's method and can evaluate in an efficient and stable way the components in (12) and their first and second derivatives with respect to $\log(\boldsymbol{\lambda})$ (since the smoothing parameters can only take positive values).

The methods for estimating $\boldsymbol{\delta}$ and $\boldsymbol{\lambda}$ are iterated until the algorithm satisfies the criterion $\frac{\left|\ell(\boldsymbol{\delta}^{[a+1]}) - \ell(\boldsymbol{\delta}^{[a]})\right|}{0.1 + \left|\ell(\boldsymbol{\delta}^{[a+1]})\right|} < 1e-07$. The selection of starting values plays an important role as it would in the majority of optimization problems. In this case, values for the marginal models are obtained by employing the `gamlss()` function within `GJRM`, which has been extended to fit univariate generalized survival models using the estimation approach proposed in this paper. This can be regarded as a contribution in itself as, to the best of our knowledge, the treatment of survival link-based models with flexible additive predictors and integrated automatic and stable multiple smoothing parameter selection has not been dealt with in the literature. An initial value for the copula parameter is obtained by using a transformation of the empirical Kendall's association between the responses.

Remark 5. The *edf* for a model containing only unpenalized terms is equal to ψ, the dimension of $\boldsymbol{\delta}$, since in this case $\text{tr}(\mathbf{A}) = \text{tr}(\mathbf{I})$. The *edf* for a penalized model is $\text{tr}(\mathbf{A})$ which can also be written as $\psi - \text{tr}\left\{(-\boldsymbol{H} + \mathbf{S})^{-1}\mathbf{S}\right\}$. The latter expression clearly shows the role of the smoothing parameter vector (contained in $\mathbf{S}$); if $\boldsymbol{\lambda} \to \mathbf{0}$ then $\text{tr}(\mathbf{A}) \to \psi$ and if $\boldsymbol{\lambda} \to \boldsymbol{\infty}$ then $\text{tr}(\mathbf{A}) \to \psi - \zeta$, where ζ is the total number of model's parameters subject to penalization. When $\mathbf{0} < \boldsymbol{\lambda} < \boldsymbol{\infty}$, the model's *edf* is equal to a value in the range $[\psi - \zeta, \psi]$. The *edf* of a single smooth or penalized component is given by the sum of the corresponding trace elements and has a value smaller than or equal to $J_{\nu k_\nu}$.

Remark 6. Equation (11) is approximately equivalent to $AIC = 2edf - 2\ell(\hat{\boldsymbol{\delta}})$, which can be shown as follows. A second order Taylor expansion of $-2\ell(\hat{\boldsymbol{\delta}})$ about $\boldsymbol{\delta}$ yields $-2\ell(\hat{\boldsymbol{\delta}}) \approx -2\ell(\boldsymbol{\delta}) - 2(\hat{\boldsymbol{\delta}} - \boldsymbol{\delta})^{\mathsf{T}}\mathbf{g} - (\hat{\boldsymbol{\delta}} - \boldsymbol{\delta})^{\mathsf{T}}\boldsymbol{H}(\hat{\boldsymbol{\delta}} - \boldsymbol{\delta})$. Recalling the definition of $\mathbf{M}$ and after some manipulation, $-(\hat{\boldsymbol{\delta}}-\boldsymbol{\delta})^{\mathsf{T}}\boldsymbol{H}(\hat{\boldsymbol{\delta}}-\boldsymbol{\delta})$ equals $\|\mathbf{M}-\sqrt{-\boldsymbol{H}}\hat{\boldsymbol{\delta}}\|^2 - 2\left\langle \mathbf{M} - \sqrt{-\boldsymbol{H}}\hat{\boldsymbol{\delta}}, \sqrt{-\boldsymbol{H}}^{-1}\mathbf{g}\right\rangle + \|\sqrt{-\boldsymbol{H}}^{-1}\mathbf{g}\|^2$, where $\langle\cdot,\cdot\rangle$ is the inner product. Similarly, $(\hat{\boldsymbol{\delta}}-\boldsymbol{\delta})^{\mathsf{T}}\mathbf{g}$ can be re-written as $-\left\langle \mathbf{M} - \sqrt{-\boldsymbol{H}}\hat{\boldsymbol{\delta}}, \sqrt{-\boldsymbol{H}}^{-1}\mathbf{g}\right\rangle + \|\sqrt{-\boldsymbol{H}}^{-1}\mathbf{g}\|^2$. These results lead to $2edf - 2\ell(\boldsymbol{\delta}) - \|\sqrt{-\boldsymbol{H}}^{-1}\mathbf{g}\|^2 + \|\mathbf{M} - \sqrt{-\boldsymbol{H}}\hat{\boldsymbol{\delta}}\|^2$. Dropping the terms that are not affected by $\boldsymbol{\lambda}$, we have that $2edf + \|\mathbf{M} - \sqrt{-\boldsymbol{H}}\hat{\boldsymbol{\delta}}\|^2$, where the latter quantity is a quadratic approximation of $-2\ell(\hat{\boldsymbol{\delta}})$.

2.5 Some theoretical results

In this section, we present the main asymptotic result related to the proposed estimator and then discuss the construction of confidence intervals. The large sample behavior of the penalized maximum likelihood estimator, $\hat{\boldsymbol{\delta}} = \arg\max_{\boldsymbol{\delta}} \ell_p(\boldsymbol{\delta})$, can be established under the relatively mild conditions of the consistency of the maximum likelihood estimator. Specifically,

Theorem 1. *Under the assumptions in Supplementary Material B and as $n \to \infty$, it follows that*

$$\hat{\boldsymbol{\delta}} - \boldsymbol{\delta}^0 = O_P(1/\sqrt{n}).$$

For the sake of space, we refer the reader to Supplementary Material B for more details and remarks.

As for the construction of intervals, it is more convenient to take a Bayesian view of the model and employ at convergence the result $\boldsymbol{\delta} \dot{\sim} \mathcal{N}(\hat{\boldsymbol{\delta}}, \mathbf{V}_{\boldsymbol{\delta}})$, where $\mathbf{V}_{\boldsymbol{\delta}} = -\boldsymbol{H}_p(\hat{\boldsymbol{\delta}})^{-1}$. As shown theoretically and via simulation by Marra & Wood (2012) for generalized additive models, 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. The above posterior can be justified using the distribution of $\mathbf{M}$ given in Section 2.4, making the large sample assumption that $\boldsymbol{H}(\boldsymbol{\delta})$ can be treated as fixed, and making the prior Bayesian assumption for smooth models $\boldsymbol{\delta} \sim \mathcal{N}(\mathbf{0}, \mathbf{S}^{-1})$, where $\mathbf{S}^{-1}$ is the Moore-Penrose pseudo-inverse of $\mathbf{S}$ (e.g., Silverman, 1985; Wood, 2017).

Remark 7. The covariance matrix of $\hat{\boldsymbol{\delta}}$ can be shown to be equal to $\mathbf{Cov}(\hat{\boldsymbol{\delta}}) \approx -\mathbf{F}^{-1}\mathbb{E}\left[\boldsymbol{H}(\boldsymbol{\delta}^0)\right]\mathbf{F}^{-1}$, where $\mathbf{F} = -\mathbb{E}\left[\boldsymbol{H}(\boldsymbol{\delta}^0)\right] + \mathbf{S}$. However, assumption $\mathbf{S} = o\left(\sqrt{n}\right)$ (in Supplementary Material B used in Theorem 1) implies that $\sqrt{n}\mathbf{Cov}(\hat{\boldsymbol{\delta}}) \approx \left\{\frac{1}{\sqrt{n}}\mathbb{E}\left[-\boldsymbol{H}(\delta^0)\right]\right\}^{-1}$ and $\sqrt{n}\mathbf{V}_{\boldsymbol{\delta}} \approx \left\{-\frac{1}{\sqrt{n}}\boldsymbol{H}(\delta^0)\right\}^{-1}$. Although the frequentist asymptotic approximation and the Bayesian result become equivalent as $n \to \infty$, as explained in the previous paragraph the latter will deliver better calibrated intervals in practical situations (see the last paragraph of Supplementary Material C for some simulation-based evidence).

Remark 8. Point-wise intervals for linear functions of the model's coefficients (such as smooth components) can be straightforwardly obtained using the Bayesian posterior distribution. As for intervals for non-linear functions of the model's coefficients (e.g., τ, hazard functions), these can be conveniently obtained by simulation, hence avoiding computationally expensive parametric bootstrap. That is,

1. Draw n_{sim} random vectors from $\mathcal{N}(\hat{\boldsymbol{\delta}}, \mathbf{V}_{\boldsymbol{\delta}})$.
2. Calculate n_{sim} simulated realizations of the quantity of interest. As an example, consider the Gaussian copula model where $\tau_i = \frac{2}{\pi}\arcsin\left[\tanh\left\{\eta_{3i}(\mathbf{x}_{3i};\boldsymbol{\beta}_3)\right\}\right]$. In this case, we would obtain $\boldsymbol{\tau}_i^{sim} = (\tau_i^{sim_1}, \tau_i^{sim_2}, \ldots, \tau_i^{sim_{n_{sim}}})^{\mathsf{T}}\ \forall i = 1, \ldots, n$ using $\boldsymbol{\beta}_3^{sim_j}\ \forall j = 1, \ldots, n_{sim}$.
3. For each $\boldsymbol{\tau}_i^{sim}$, calculate the lower, $\varsigma/2$, and upper, $1-\varsigma/2$, quantiles.

A small value for n_{sim}, say 100, typically gives accurate results, whereas ς is usually set to 0.05. Note that the distribution of non-linear functions of $\boldsymbol{\delta}$ need not be symmetric. To derive intervals for non-linear functions of the model's coefficients, we also considered using a frequentist approach based on the asymptotic covariance matrix shown in **Remark** 7 and the delta method. At finite sample sizes, the results were not satisfactory since the intervals were symmetric (which is typically not the case for non-linear functions of model's parameters) and did not take into account smoothing bias (see Figure 9 and the last paragraph of Supplementary Material C for more comments on this).

Remark 9. As pointed out by Pya & Wood (2015), interval estimates for the monotonic smooth terms in the model can be easily obtained using the distribution for $\tilde{\boldsymbol{\delta}}^{\mathsf{T}} = (\tilde{\boldsymbol{\beta}}_1^{\mathsf{T}}, \tilde{\boldsymbol{\beta}}_2^{\mathsf{T}}, \boldsymbol{\beta}_3^{\mathsf{T}})$, since such smooth components depend linearly on $\tilde{\boldsymbol{\beta}}_1$ and $\tilde{\boldsymbol{\beta}}_2$. The distribution of $\tilde{\boldsymbol{\delta}}$ is $\tilde{\boldsymbol{\delta}} \dot{\sim} \mathcal{N}(\hat{\tilde{\boldsymbol{\delta}}}, \mathbf{V}_{\tilde{\delta}})$,

where $\mathbf{V}_{\tilde{\boldsymbol{\delta}}} = \text{diag}(\mathbf{E})\,\mathbf{V}_{\boldsymbol{\delta}}\,\text{diag}(\mathbf{E})$, $\mathbf{E}^{\mathsf{T}} = (\mathbf{E}_1^{\mathsf{T}}, \mathbf{E}_2^{\mathsf{T}}, \mathbf{1}^{\mathsf{T}})$ and $\mathbf{1}$ has the same dimension of $\boldsymbol{\beta}_3$. This is obtained by considering a Taylor series expansion of $\tilde{\boldsymbol{\delta}}$ as a vector of functions of $\boldsymbol{\delta}$, i.e. $\tilde{\boldsymbol{\delta}} - \hat{\tilde{\boldsymbol{\delta}}} \approx \text{diag}(\mathbf{E})\left(\boldsymbol{\delta} - \hat{\boldsymbol{\delta}}\right)$. This shows that $\tilde{\boldsymbol{\delta}} - \hat{\tilde{\boldsymbol{\delta}}}$ is approximately a linear function of $\boldsymbol{\delta}$. Recalling that linear functions of normally distributed random variables follow normal distributions, the result in this Remark follows.

Remark 10. P-values for the smooth components in the model are obtained by adapting the results discussed in Wood (2013) to the current context. Note that $\mathbf{V}_{\tilde{\boldsymbol{\delta}}}$ is employed for p-value calculations, which is especially relevant for the monotonic terms in the model since it allows us to directly test these smooth functions for equality to zero.

Tools to aid the model building process are described in Supplementary Material E. The modeling and estimation framework discussed in this paper has been implemented in the `R` package `GJRM` (Marra & Radice, 2019) and we refer the reader to Supplementary Material D for a brief description of the software. Supplementary Material C provides the details and results of a simulation study.

3 Data analysis

In this section, we apply the proposed approach to data collected through a questionnaire survey of adult members of the Australian NH&MRC Twin Registry (Duffy et al., 1990) which have been recently analyzed by Romeo et al. (2018). One of the aims of this study was to investigate whether the magnitude of the dependence within adult twin pairs as to the risk of the onset of acute appendicitis is different for monozygotic (MZ) and dizygotic (DZ) twins. This would provide information on the role of heredity in the onset of appendicitis since the strength of the dependence within MZ and DZ twins is expected to be very similar and a difference in such strength would be indicative of a genetic effect on the risk of acute appendicitis. As in Romeo et al. (2018), we considered female twin pairs who had an appendectomy; the sample sizes were 1231 and 748 for MZ and DZ twins, respectively. The outcome variable was age at appendectomy (or censoring age), and the censoring rate was about 73% for each twin member in both zygotes. For more details and descriptive statistics see Romeo et al. (2018) and references therein. To facilitate the

	MZ twins				DZ twins			
Copula	AIC	BIC	τ	(95% CIs)	AIC	BIC	τ	(95% CIs)
N	7119.5	7182.0	0.31	(0.26, 0.36)	4303.1	4347.4	0.19	(0.11, 0.26)
C0	7133.3	7194.7	0.45	(0.39, 0.51)	4308.6	4353.1	0.27	(0.17, 0.38)
C180	7124.2	7187.3	0.20	(0.17, 0.25)	4302.9	4346.8	0.12	(0.08, 0.18)
J0	7122.3	7184.9	0.18	(0.15, 0.23)	4302.8	4346.3	0.10	(0.06, 0.16)
J180	7134.9	7196.2	0.46	(0.40, 0.53)	4308.9	4353.3	0.28	(0.19, 0.39)
G0	7115.5	7178.0	0.24	(0.19, 0.29)	4301.6	4345.2	0.13	(0.08, 0.20)
G180	7117.3	7179.1	0.38	(0.33, 0.44)	4303.7	4348.1	0.23	(0.16, 0.33)
F	7122.8	7184.7	0.33	(0.27, 0.38)	4306.4	4350.7	0.19	(0.11, 0.26)
AMH	7143.7	7205.3	0.33	(-0.18, 0.33)	4308.1	4352.4	0.22	(0.06, 0.30)
FGM	7144.6	7206.9	0.22	(-0.22, 0.22)	4307.8	4352.1	0.19	(-0.01, 0.22)
PL	7117.3	7179.4	0.33	(0.28, 0.38)	4305.4	4349.7	0.19	(0.12, 0.26)
T(3)	7108.2	7169.6	0.28	(0.22, 0.34)	4302.3	4345.7	0.14	(0.06, 0.22)
T(4)	7109.2	7170.7	0.29	(0.23, 0.34)	4301.5	4345.1	0.15	(0.06, 0.23)
T(5)	7110.3	7172.0	0.30	(0.24, 0.35)	4301.2	4345.1	0.16	(0.07, 0.23)
T(6)	7111.3	7173.1	0.30	(0.25, 0.36)	4301.2	4345.2	0.16	(0.09, 0.24)
T(7)	7112.1	7174.0	0.30	(0.24, 0.35)	4301.3	4345.3	0.16	(0.08, 0.24)
T(8)	7112.8	7174.8	0.30	(0.25, 0.35)	4301.3	4345.4	0.17	(0.09, 0.24)

Table 3: Values of model selection criteria for several copula models and estimates of Kendall's τ for MZ and DZ twins. The values in brackets next to the estimates for τ represent 95% intervals obtained using the approach described in Section 2.5. The values within brackets next to the Student-t copulae refer to ζ, the assumed degrees of freedom of the distribution.

comparison of results, we first followed the modeling strategy of Romeo et al. (2018) and then tried out a few more model specifications.

For the marginal equations, the smooth functions of the time variables were specified using monotonic penalized B-splines with penalty defined in Section 2.2 and 10 bases. Following the suggestion of Royston & Parmar (2002), smoothing was implemented on the log-time scale which usually yields very smooth fitted functions and hence it helps for example to reduce the chance of potential artifacts in the estimated hazard functions. All available link functions were considered in the modeling whereas, for the selection of the copula function, we started off with the Gaussian and then, based on the (negative or positive) sign of the dependence, we tried out the alternative specifications that were consistent with this initial finding. Using a 2.20-GHz Intel(R) Core(TM) computer running Windows 7, the average computing time was about 4 seconds and the total number of estimated parameters was 21.

Table 3 shows the values of the AIC and Bayesian information criterion (BIC), and the estimates of Kendall's τ (as well as 95% intervals) obtained when employing various copula models.

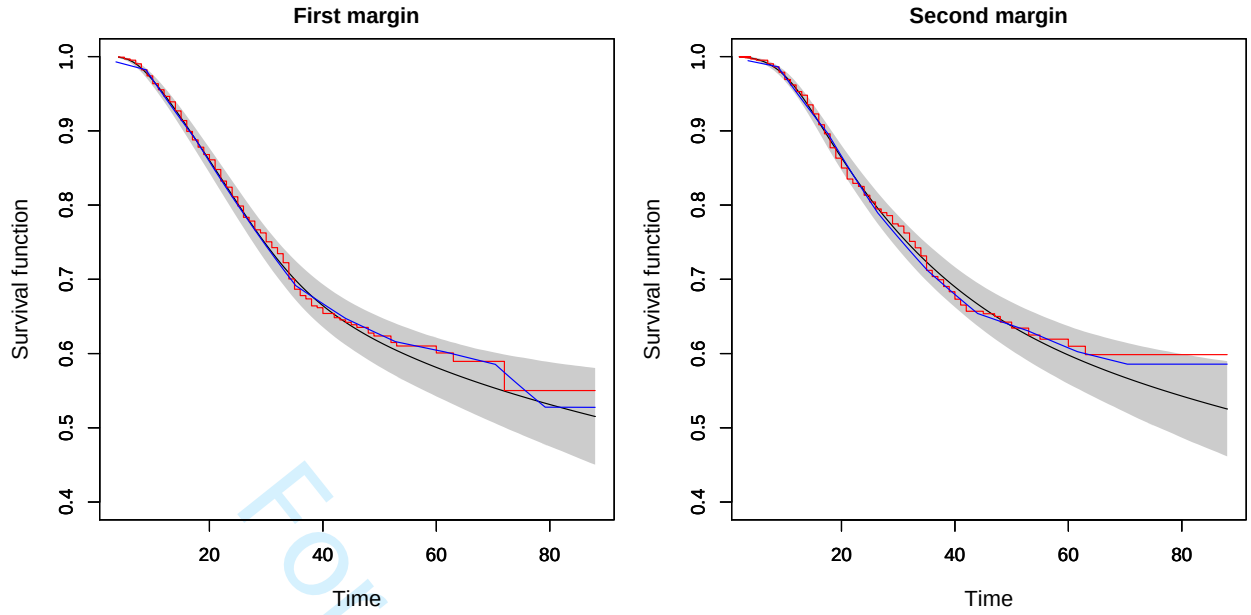

Figure 1: Survival function estimates obtained when applying to MZ twins data the proposed approach (black lines, with 95% intervals represented by the shaded areas), Kaplan-Meier estimator (red lines) and piecewise exponential model (blue lines) based on 10 intervals. The 95% intervals have been obtained using the approach described in Section 2.5.

For MZ twins, the Student-t with 3 degrees of freedom provides the best fit. For DZ twins, the situation is less clear cut in that several copulae look plausible, namely the Student-t and Gumbel. Looking at Kendall's τ, we see that the dependence between MZ pairs is stronger than that between DZ pairs, and that the confidence intervals either do not overlap or overlap slightly. This points to the presence of a genetic component to the disease as mentioned at the beginning of this section. These results are in line with those of Romeo et al. (2018) with the difference that we found marginally stronger dependencies albeit with slightly wider intervals. Our AIC and BIC values (when compared to those in Tables 5 and 6 of the above authors) suggest that the proposed approach yields slightly improved model fits. As for the marginals, we chose `PH` link functions although using `PO` and `probit` links led to very similar information criteria values as well as virtually identical results.

Figure 1 shows the survival function estimates produced when applying to MZ twins data the proposed approach, Kaplan-Meier estimator and piecewise exponential model (based on 10 intervals). The latter estimates were derived using the `R` packages `survival` (Therneau & Lumley, 2018) and `pssm` (Schoenfeld, 2017), and they fall overall within the 95% intervals obtained from the proposed method. Moreover, using a smaller number of intervals for the exponential model

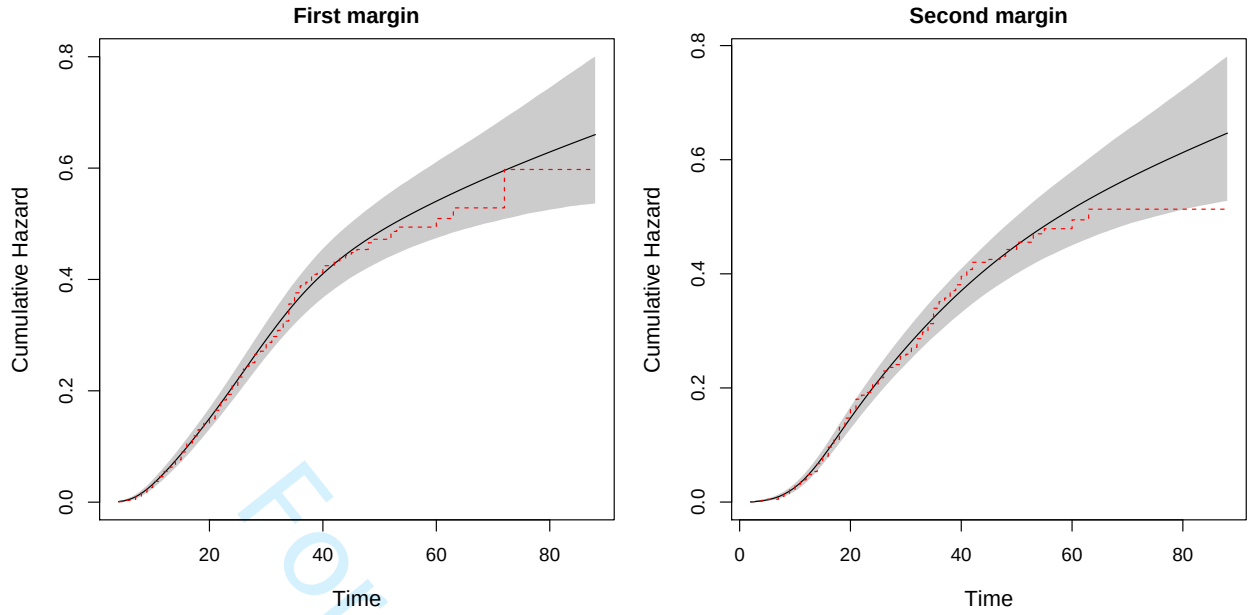

Figure 2: Cumulative hazards function estimates obtained when applying to MZ twins data the proposed approach (black lines, with 95% intervals represented by the shaded areas) and Kaplan-Meier estimator (red lines). The 95% intervals have been obtained using the approach described in Section 2.5.

(i.e., 7, 8 and 9) gave similar results. Towards the end of follow-up, when there are fewer events, there is some discrepancy in the fits produced by the three methods; this is related to a proportion of subjects that are not susceptible to the event of interest, in which case techniques developed in the area of cure rate models could be exploited to address this problem. It is worth pointing out that under the proposed copula link-based additive survival models it would be straightforward to predict, for instance, the survival probability of a new individual; this would be especially relevant when covariates are included in the model. Also, in a spline context estimating simultaneously all the smoothing parameters in a data-driven and automatic manner is crucial for practical purposes, and using a different approach to specify the marginals would not have allowed us to benefit from the efficient and stable multiple smoothing parameter selection technique presented in Section 2.4. The results for DZ twins were very similar. Figure 2 shows the plots of the cumulative hazards functions from the marginal proportional hazards equations of the proposed copula model and the Kaplan-Meier estimator; they are close and the Kaplan-Meier estimates fall overall within the intervals of the proposed method.

As in Romeo et al. (2018), we then merged the MZ and DZ data sets and specified the copula parameter of the Student-t with 3 degrees of freedom as function of type of zygosity (here used as a dichotomous covariate). This has the advantage of estimating the Kendall's τ for MZ and DZ twins

without splitting the data set. The results for τ were $0.29\ (0.23, 0.34)$ and $0.13\ (0.05, 0.21)$ for MZ and DZ twins, respectively, which are in line with those obtained from the separate analyzes.

4 Discussion

In this article, we have introduced copula link-based additive models for survival data and demonstrated their potential using simulated and real data. Important features of the proposed estimation and inferential framework are that: the marginal models can be specified using parametric but flexible formulations for time-to-event data which have several advantages including the easy post-estimation interpretation and calculation, hence visualization, of the flexibly estimated baseline hazard functions; monotonic splines are utilized to provide coherent marginal survival fits; each parameter of the assumed joint survival distribution is allowed to depend on an additive predictor incorporating several types of covariate effects; theoretically founded inferential results are employed for interval construction and hypothesis testing; all the model's parameters are estimated simultaneously using a carefully constructed efficient and stable algorithm that makes full use of the information contained in the data; the models can be easily employed using a freely available `R` package which allows for a number of modeling choices; the modularity of the implementation allows for easy inclusion of potentially any parametric link marginal function and copula.

It is worth noting that the methodology developed in this paper, although flexible, is fundamentally parametric and as such it may suffer from the usual potential drawbacks resulting from departures from the model assumptions. Developments where the margins and/or copula function are estimated using techniques that are more robust to model mis-specification were explored and based on Kauermann et al. (2013) and Segers et al. (2014). However, these were found to be limited with respect to the inclusion of flexible covariate effects and to require large sample sizes to produce reliable results in a regression context. We eventually elected to develop a flexible parametric modeling framework that would allow us to conveniently combine arbitrary marginal survival functions with various types of dependence structures linking them, and to allow for the possibility to specify all the model's parameters as functions of additive predictors which can be advantageous in the empirical applications.

Future research will focus on extending the models to the cases on left and interval censored responses, and will look into the extension to modeling more than two event times using, for instance, multivariate Archimedean copulae, mixtures of powers, pair-copulae constructions, the multivariate Gaussian and Student-t distributions, and the composite likelihood approach. We will also investigate the use of alternative model selection criteria such as cross-validation with score based on log-likelihood joint function evaluations.

Acknowledgments

We are indebted to four anonymous reviewers, the Associate Editor and Editor for their well thought out suggestions which helped us to improve and clarify several aspects of the article.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. *In: Petrov, B.N., Csaki, B.F. (eds.) Second International Symposium on Information Theory. Academiai Kiado, Budapest*.

Bogaerts, K. & Lesaffre, E. (2008). Modeling the association of bivariate interval-censored data using the copula approach. *Statistics in Medicine*, 27, 6379–6392.

Brechmann, E. C. & Schepsmeier, U. (2013). Modeling dependence with c- and d-vine copulas: The R package CDVine. *Journal of Statistical Software*, 52(3), 1–27.

Clayton, D. G. (1978). A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika*, 65(1), 141–151.

Collett, D. (2015). *Modelling survival data in medical research*. Third Edition, Chapman & Hall/CRC Press, London.

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

Duffy, D. L., Martin, N. G., & Mathews, J. D. (1990). Appendectomy in australian twins. *The American Journal of Human Genetics*, 47(3), 590–592.

Eilers, P. H. C. & Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science*, 11(2), 89–121.

Geerdens, C., Acar, E. F., & Janssen, P. (2018). Conditional copula models for right-censored clustered event time data. *Biostatistics*, 19(2), 247–262.

Gentle, J. E. (2003). *Random number generation and Monte Carlo methods*. Springer-Verlag, London.

Gray, R. J. (1992). Flexible methods for analyzing survival data using splines, with applications to breast cancer prognosis. *Journal of the American Statistical Association*, 87(420), 942–951.

Hastie, T. J. & Tibshirani, R. J. (1990). *Generalized Additive Models*. Chapman & Hall/CRC, London.

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

Kauermann, G. (2005). Penalized spline smoothing in multivariable survival models with varying coefficients. *Computational Statistics and Data Analysis*, 49(1), 169–186.

Kauermann, G., Krivobokova, T., & Fahrmeir, L. (2009). Some asymptotic results on generalized penalized spline smoothing. *Journal of the Royal Statistical Society Series B*, 71(2), 487–503.

Kauermann, G., Schellhase, C., & Ruppert, D. (2013). Flexible copula density estimation with penalized hierarchical b-splines. *Scandinavian Journal of Statistics*, 40(4), 685–705.

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, R., Cheng, Y., Chen, Q., & Jason, F. (2017). Quantile association for bivariate survival data. *Biometrics*, 73(2), 506–516.

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. (2017). Bivariate copula additive models for location, scale and shape. *Computational Statistics & Data Analysis*, 112, 99–113.

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

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

Marra, G. & Wood, S. (2012). Coverage properties of confidence intervals for generalized additive model components. *Scandinavian Journal of Statistics*, 39(1), 53–74.

Meyer, M. (2012). Constrained penalized splines. *Canadian Journal of Statistics*, 40(1), 190–206.

Meyer, R. & Romeo, J. (2015). Bayesian semi-parametric analysis of recurrent failure time data using copulas. *Biometrical Journal*, 57, 982–1001.

Nan, B., Lin, X., Lisabeth, L. D., & Harlow, S. D. (2006). Piecewise constant cross-ratio estimation for association of age at a marker event and age at menopause. *Journal of the American Statistical Association*, 101(473), 65–77.

Nelsen, R. (2006). *An Introduction to Copulas*. Second Edition, Springer, New York.

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

Oakes, D. (1982). A model for association in bivariate survival data. *Journal of the Royal Statistical Society Series B*, 44(3), 414–422.

Oakes, D. (1986). A class of multivariate failure time distributions. *Biometrika*, 73(3), 671–678.

Patton, A. J. (2002). Applications of copula theory in financial econometrics. *Ph.D. thesis, University of California, San Diego*.

Prenen, L., Braekers, R., & Duchateau, L. (2017). Extending the archimedean copula methodology to model multivariate survival data grouped in clusters of variable size. *Journal of the Royal Statistical Society Series B*, 79(2), 483–505.

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

R Development Core Team (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Reid, N. (1994). A conversation with Sir David Cox. *Statistical Science*, 9(3), 439–455.

Romeo, J., Meyer, R., & Gallardo, D. (2018). Bayesian bivariate survival analysis using the power variance function copula. *Lifetime Data Analysis*, 24(2), 355–383.

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.

Rue, H. & Held, L. (2005). *Gaussian Markov Random Fields: Theory and Applications*. Chapman & Hall/CRC, Florida.

Ruppert, D., Wand, M. P., & Carroll, R. J. (2003). *Semiparametric Regression*. Cambridge University Press, New York.

Schoenfeld, D. A. (2017). *pssm: Piecewise Exponential Model for Time to Progression and Time from Progression to Death*. R package version 1.1.

Segers, J., van den Akker, R., & Werker, B. J. M. (2014). Semiparametric gaussian copula models: Geometry and efficient rank-based estimation. *Annals of Statistics*, 42(5), 1911–1940.

Silverman, B. W. (1985). Some aspects of the spline smoothing approach to non-parametric regression curve fitting. *Journal of the Royal Statistical Society Series B*, 47(1), 1–52.

Sklar, A. (1973). Random variables, joint distributions, and copulas. *Kybernetica*, 9(6), 449–460.

Therneau, T. M. & Lumley, T. (2018). *survival: Survival Analysis*. R package version 2.42-6.

Trivedi, P. K. & Zimmer, D. M. (2007). Copula modeling: An introduction for practitioners. *Foundations and Trends (R) in Econometrics*.

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

Wienke, A., Lichtenstei, P., & Yashin, A. (2003). A bivariate frailty model with a cure fraction for modeling familial correlations in diseases. *Statistics and Computing*, 59(4), 1178–1183.

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. (2013). On p-values for smooth components of an extended generalized additive model. *Biometrika*, 100(1), 221–228.

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

Wood, S. N. (2018). *mgcv: Mixed GAM Computation Vehicle with GCV/AIC/REML Smoothness Estimation*. R package version 1.8-24.

Yoshida, T. & Naito, K. (2014). Asymptotics for penalized splines in generalized additive models. *Journal of Nonparametric Statistics*, 26(2), 269–289.

Zhang, J. (2004). A simple and efficient monotone smoother using smoothing splines. *Journal of Nonparametric Statistics*, 16(5), 779–796.

Supplementary Material: "Copula Link-Based Additive Models for Right-Censored Event Time Data"

Giampiero Marra[1] and Rosalba Radice [2]

Supplementary Material A

This section complements Section 2.1.1 by providing some examples of penalty and basis function specifications.

Linear and random effects For parametric, linear effects, equation (3) becomes $\mathbf{z}_{\nu k_\nu i}^{\mathsf{T}}\boldsymbol{\beta}_{\nu k_\nu}$, and the design matrix is obtained by stacking all covariate vectors $\mathbf{z}_{\nu k_\nu i}$ into $\mathbf{Z}_{\nu k_\nu}$. No penalty is typically assigned to linear effects ($\mathbf{D}_{\nu k_\nu} = \mathbf{0}$). This would be the case for binary and categorical variables. However, sometimes it is desirable to penalize parametric linear effects. For instance, the coefficients of some factor variables in the model may be weakly or not identified by the data. In this case, a ridge penalty could be employed to make the model's parameters estimable (here $\mathbf{D}_{\nu k_\nu} = \mathbf{I}$ where $\mathbf{I}$ is an identity matrix). This is equivalent to the assumption that the coefficients are $i.i.d.$ normal random effects with unknown variance (e.g., Wood, 2017).

Non-linear effects For continuous variables the smooth functions are represented using the regression spline approach (e.g., Wood, 2017). Specifically, for each continuous variable $z_{\nu k_\nu i}$, $s_{\nu k_\nu}(z_{\nu k_\nu i})$ is approximated by $\sum_{j\nu k_\nu=1}^{J_{\nu k_\nu}} \beta_{\nu k_\nu j_{\nu k_\nu}} b_{\nu k_\nu j_{\nu k_\nu}}(z_{\nu k_\nu i})$, where the $b_{\nu k_\nu j_{\nu k_\nu}}(z_{\nu k_\nu i})$ are known spline basis functions. The design matrix $\mathbf{Z}_{\nu k_\nu}$ comprises the basis function evaluations for each i, and hence describe $J_{\nu k_\nu}$ curves which have potentially varying degrees of complexity. We typically employ low rank thin plate regression splines which are numerically stable and have con-

[1]Department of Statistical Science, University College London, Gower Street, London WC1E 6BT, UK, giampiero.marra@ucl.ac.uk.

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

venient mathematical properties, although other spline definitions and corresponding penalties are supported in our implementation. Note that for one-dimensional smooth functions, the choice of spline definition does not play an important role in determining the shape of $\hat{s}_{\nu k_\nu}(z_{\nu k_\nu})$. To enforce smoothness, a conventional integrated square second derivative spline penalty is often employed (this is also the default option in the software). That is, $\mathbf{D}_{\nu k_\nu} = \int \mathbf{d}_{\nu k_\nu}(z_{\nu k_\nu})\mathbf{d}_{\nu k_\nu}(z_{\nu k_\nu})^\mathsf{T} dz_{\nu k_\nu}$, where the $j^{th}_{\nu k_\nu}$ element of $\mathbf{d}_{\nu k_\nu}(z_{\nu k_\nu})$ is given by $\partial^2 b_{\nu k_\nu j_{\nu k_\nu}}(z_{\nu k_\nu})/\partial z^2_{\nu k_\nu}$ and integration is over the range of $z_{\nu k_\nu}$. The formulas used to compute the basis functions and penalties for many spline definitions are provided in Wood (2017). This specification allows us to avoid arbitrary modeling decisions, such as choosing the appropriate degree of a polynomial or specifying cut-points, which could induce mis-specification bias. Many other types of spline bases and respective penalties can be employed, such as penalized cubic regression spline and P-splines.

Spatial effects When the geographic area (or country) of interest is split up into discrete contiguous geographic units (or regions) and such information is available, a Markov random field approach can be employed to exploit the information contained in neighboring observations which are located in the same country. In this case, equation (3) becomes $\mathbf{z}^\mathsf{T}_{\nu k_\nu i}\boldsymbol{\beta}_{\nu k_\nu}$ where $\boldsymbol{\beta}_{\nu k_\nu} = (\beta_{\nu k_\nu 1}, \ldots, \beta_{\nu k_\nu J_{\nu k_\nu}})^\mathsf{T}$ represents the vector of spatial effects, $J_{\nu k_\nu}$ denotes the total number of regions and $\mathbf{z}_{\nu k_\nu i}$ is made up of a set of area labels. The design matrix linking an observation i to the corresponding spatial effect is therefore defined as

$$\mathbf{Z}_{\nu k_\nu}[i, j_{\nu k_\nu}] = \begin{cases} 1 \text{ if the observation belongs to region } j_{\nu k_\nu} \\ 0 \text{ otherwise} \end{cases},$$

where $j_{\nu k_\nu} = 1, \ldots, J_{\nu k_\nu}$. The smoothing penalty is based on the neighborhood structure of the geographic units, so that spatially adjacent regions share similar effects. That is,

$$\mathbf{D}_{\nu k_\nu}[r, q] = \begin{cases} -1 \text{ if } r \neq q \wedge \ r \text{ and } q \text{ are adjacent neighbors} \\ 0 \quad \text{if } r \neq q \wedge \ r \text{ and } q \text{ are not adjacent neighbors} \\ N_r \text{ if } r = q \end{cases},$$

where N_r is the total number of neighbors for region r. In a stochastic interpretation, this penalty is equivalent to the assumption that $\boldsymbol{\beta}_{\nu k_\nu}$ follows a Gaussian Markov random field.

Several other specifications can be adopted. These include varying coefficient smooths obtained by multiplying one or more smooth components by some covariate(s), and smooth functions of two or more continuous covariates (Wood, 2017). The smoothers utilized here are obtained from the `R mgcv` package.

Supplementary Material B

This section provides some details on the assumptions required for Theorem 1 in Section 2.5 as well as some further results. Let us fix the $J_{\nu k_\nu}$ at a high value, and let L^t and L denote the likelihood functions for the true and employed models, with corresponding log-likelihoods ℓ^t and $\ell(\boldsymbol{\delta})$. Let also $\boldsymbol{\delta}^0$ be the minimizer of the Kullback-Leibler distance, that is $\boldsymbol{\delta}^0 = \arg\min_{\boldsymbol{\delta}} \mathrm{KL}\left[L^t|L\right]$, where $\mathrm{KL}\left[L^t|L\right] = \mathbb{E}\left[\ell^t - \ell(\boldsymbol{\delta})\right]$ with expectation taken with respect to the true model's distribution and $\boldsymbol{\delta}$. Theorem 1 holds under some usual customary assumptions which are listed in full in Vatter & Chavez-Demoulin (2015, Appendix A) and include: $\boldsymbol{\delta}$ is in a compact closed and bounded parametric space Θ, $\boldsymbol{\delta}^0$ for the model is in the interior point of Θ, $\ell_p(\boldsymbol{\delta})$ is continuous and differentiable, the employed link functions are monotonic and differentiable, $\mathbf{g}(\boldsymbol{\delta}^0) = O_P(\sqrt{n})$, $\mathbb{E}\left[\boldsymbol{H}(\boldsymbol{\delta}^0)\right] = O(n)$, $\boldsymbol{H}(\boldsymbol{\delta}^0) - \mathbb{E}\left[\boldsymbol{H}(\boldsymbol{\delta}^0)\right] = O_P(\sqrt{n})$, and $\mathbf{S} = o\left(\sqrt{n}\right)$. The last assumption can be equivalently formulated as $\lambda_{\nu k_\nu} = o\left(\sqrt{n}\right)$ for $k_\nu = 1, \ldots, K_\nu$, $\nu = 1, 2, 3$, assuming that the $\mathbf{D}_{\nu k_\nu}$ are asymptotically bounded. This assumption is rather weak as it allows the smoothing parameters to grow as the sample size increases, at a rate smaller than $\sqrt{n}$. In fact, the sequence $\hat{\boldsymbol{\lambda}}$ based on the mean squared error criterion described in Section 2.4 is bounded in probability.

Remark 11. The result in Theorem 1 assumes that the $J_{\nu k_\nu}$ values are fixed at a high value. This is a convenient assumption since the unknown $s_{\nu k_\nu}$ may not have an exact representation as linear combinations of the given bases and coefficients. However, in applied research the $J_{\nu k_\nu}$ values have to be fixed and assuming that these are high enough to assume that a good representation of the unknown functions can be obtained, it is possible to assume heuristically that the approx-

imation bias is negligible compared to estimation variability (e.g., Vatter & Chavez-Demoulin, 2015).

Remark 12. The asymptotic bias and covariance matrix of $\hat{\boldsymbol{\delta}}$ can be shown to be equal to $\mathbf{bias}(\hat{\boldsymbol{\delta}}) = \mathbb{E}\left(\hat{\boldsymbol{\delta}} - \boldsymbol{\delta}^0\right) \approx -\mathbf{F}^{-1}\mathbf{S}\boldsymbol{\delta}^0$ and $\mathbf{Cov}(\hat{\boldsymbol{\delta}}) \approx -\mathbf{F}^{-1}\mathbb{E}\left[\boldsymbol{H}(\boldsymbol{\delta}^0)\right]\mathbf{F}^{-1}$, where $\mathbf{F} = -\mathbb{E}\left[\boldsymbol{H}(\boldsymbol{\delta}^0)\right] + \mathbf{S}$. Furthermore, $\mathbf{bias}(\hat{\boldsymbol{\delta}}) = o(1/\sqrt{n})$ and $\mathbf{Cov}(\hat{\boldsymbol{\delta}}) = O(1/n)$.

Remark 13. Since $\mathbf{g}(\boldsymbol{\delta}^0)$ is a sum of *i.i.d.* components, it follows that $\left\{-\mathbb{E}\boldsymbol{H}(\boldsymbol{\delta}^0)\right\}^{-1/2}\mathbf{g}(\boldsymbol{\delta}^0) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbf{I})$. This implies asymptotic normality of the normalized estimator $\hat{\boldsymbol{\delta}}$.

These theoretical results are based on the assumptions described above and can essentially be proved from a Taylor expansion of the penalized log-likelihood.

Supplementary Material C

This section provides evidence on the empirical effectiveness of the proposed methodology in recovering true covariate effects and baseline functions.

Survival time T_{1i} was generated from a proportional hazards model defined, on the survival function scale, as $\log\left[-\log\left\{S_{10}(t_{1i})\right\}\right] + \beta_{11}z_{1i} + s_{11}(z_{2i})$ where $S_{10i}(t_{1i}) = 0.9\exp(-0.4t_{1i}^{2.5}) + 0.1\exp(-0.1t_{1i})$. Time T_{2i} was generated from a proportional odds model defined as $\log\left[\left\{1 - S_{20}(t_{2i})\right\}/S_{20}(t_{2i})\right] + \beta_{21}z_{1i} + s_{21}(z_{3i})$ where $S_{20}(t_{2i}) = S_{10}(t_{2i})$. The random censoring times C_{1i} and C_{2i} were obtained using uniform distributions with limits chosen so that censoring rates were about 42% and 33% for the first group of simulations and 75% and 50% for the second one. Observations were generated using the Brent's univariate root-finding method. The two survival times were joined using a Clayton copula where the predictor for the respective dependence parameter was specified as $\eta_{3i} = \beta_{31}z_{1i} + s_{31}(z_{2i})$. In practice, this was achieved using the conditional sampling approach. The set up of η_3 allowed dependence to vary across observations, with Kendall's τ values ranging approximately from 0.10 to 0.90. The smooth functions were $s_{11}(z_i) = \sin(2\pi z_i)$, $s_{21}(z_i) = -0.2\exp(3.2z_i)$, $s_{31}(z_i) = 3\sin(\pi z_i)$, whereas $\beta_{11} = -1.5$, $\beta_{21} = 1.2$ and $\beta_{31} = -1.5$. Finally, 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. Covariate z_{1i} was dichotomized by simply rounding it.

Sample sizes were set to 200, 500 and 1000, the number of replicates to 1000. The models were fitted using `gjrm()` in `GJRM` by employing all the marginal links and copulae listed in Tables 1 and 2. We also experimented with a two-stage approach where the `gamlss()` function from the same package was employed to obtain marginal fits and then the copula function estimated using a simplification of the code employed to fit the simultaneous model. The smooth components of the covariates were represented using penalized low rank thin plate splines with second order penalty (see Supplementary Material A) and 10 bases, and the smooths of times using monotonic penalized B-splines with penalty defined in Section 2.2 and 10 bases. For each replicate, curve estimates were constructed using 200 equally spaced fixed values in the $(0, 8)$ range for the monotonic functions and $(0, 1)$ otherwise.

We did not consider sample sizes smaller than 200 since the models involve three smooth functions (two for the margins and one for the copula parameter) and three parametric effects, hence imposing the estimation of 33 model's coefficients and 3 smoothing parameters; considering sample sizes smaller than 200 would produce meaningless results as it is known that the use of splines requires the availability of more information in the data. Preliminary experiments based on smaller samples confirmed this.

The main findings of the simulation study are organized in the bullet points below.

- *Parametric effects*: Figure 3 and Table 4 show that overall the mean estimates 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. At $n = 200$ the estimates for β_{31} (the effect of z_1 contained in the additive predictor of the copula parameter) are more variable and exhibit some bias as compared to those of the other parameters. However, the situation quickly improves as more observations are available for model fitting. We investigated this issue further and found that the profile log-likelihood of the relevant copula coefficient tends to be less sharp around the optimum than those related to the marginal equations, especially at low sample sizes. This suggests that the parameters related to the copula function may be more difficult to estimate when using a small data set. Therefore, in such a situation, more care is likely to be needed when deciding on the complexity of the

copula's additive predictor. These results are in line with those of Romeo et al. (2018) and references therein who found the same difficulty, particularly under a low level of dependence and small sample sizes.

- *Smooth effects*: Figure 6 and Table 4 show that overall the true functions are recovered well by the proposed estimation method and that the results improve in terms of bias and efficiency as the sample size increases. As for β_{31}, we see that at $n = 200$ estimation of $s_{31}(z_2)$ is more challenging. However, the performance improves dramatically as the sample grows large.

- *Impact of censoring rates*: Comparing Figures 3 and 6 (mild censoring rates) with Figures 4 and 7 (high censoring rates), and Table 4 (mild censoring) with Table 5 (high censoring), we see that the presence of high censoring deteriorates the estimation performance. Moreover, the most affected parameters are those belonging to the copula's additive predictor (for the same reason given in the first bullet point). These results do not come as a surprise given the loss of information caused by right-censoring. As the sample size increases the estimates improve considerably. Finally, high censoring caused the algorithm to fail to converge for a few simulation replicates which were discarded from the results.

- *Results from two-stage approach*: Comparing Figures 3 and 6 (simultaneous estimation approach) with Figures 5 and 8 (two-stage approach), we observe that, despite the two-stage method generally produces slightly more variable and biased estimates, the results are overall close. At $n = 200$, the differences are more tangible and the worse performance of the two-stage technique can be attributed to convergence failures (in around 20% of the replicates) at the copula step (the one involving the estimation of the copula's additive predictor). As elaborated in the first bullet point, the copula parameter is the most difficult to estimate and having a carefully constructed algorithm, that can exploit all the information available in the data, is advantageous in the context of the models developed in this paper; see also Marra & Radice (2017) who found similar results in a related model setting.

- *Model selection*: For each scenario considered in the simulation study, we fitted the correct model (based on proportional hazards and proportional odds margins for the first and second

equations respectively, and the Clayton copula) as well as several mis-specified models. The latter were first based on the correct margins (PH and PO for the two equations) and incorrect copulae (all those listed in Table 1 but the Clayton), and then based on the incorrect margins (here we swapped the marginal links by employing PO and PH for the two equations, respectively) and all copulae listed in Table 1. At $n = 500, 1000$, for each scenario and replicate, the correct model was always chosen by the AIC and BIC. At $n = 200$, the mis-specified model based on the correct margins and Joe copula rotated by 180 degrees was favored around 30% of times over the correct model. This result was not unexpected because the Clayton and rotated Joe copulae capture similar dependence structures, hence the differences between them may be hard to detect at small sample sizes.

Computing time for the proposed approach was on average 12 seconds for $n = 1000$ and around 7 seconds for $n = 200, 500$. Following a reviewer comment, we also calculated 95% average coverage probabilities for s_{11}, s_{21} and s_{31} using point-wise intervals based on the frequentist and Bayesian covariance matrices given in Section 2.5. For all smooth terms and scenarios considered, the coverages obtained using the Bayesian result ranged from 0.94 to 0.96, whereas those obtained with the frequentist approximation were lower by 0.02 on average. This confirmed that neglecting smoothing bias has a negative impact on the empirical performance of the intervals. We also considered a non-linear function of the model's coefficients, namely the Kendall's τ. The related simulated Bayesian intervals yielded close-to-nominal coverage probabilities as opposed to the frequentist approach (based on the asymptotic covariance matrix given in **Remark** 7 and the delta method) which produced intervals with severe under-coverage at times. To illustrate this point, Figure 9 shows the histogram and kernel density estimates of simulated Kendall's τ values obtained using the Bayesian posterior simulation approach. It is clear that the distribution of the values is asymmetric, a feature that the frequentist approach can not account for. However, we also found that the situation improves at bigger sample sizes. This was expected since the frequentist approximation and the Bayesian result are asymptotically equivalent and the asymmetry of the distribution becomes less marked.

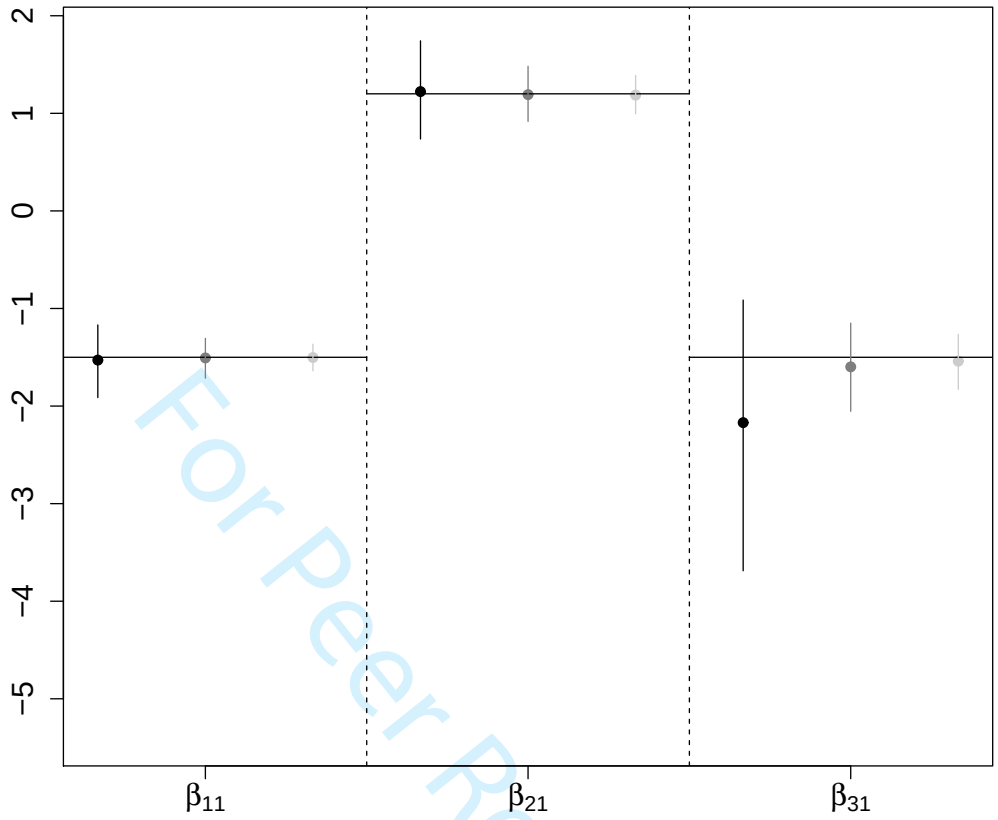

Figure 3: Linear coefficient estimates obtained by applying `gjrm()` to bivariate survival simulated data with mild censoring rates (about 42% and 33% for the two responses). Circles indicate mean estimates while bars represent the estimates' ranges resulting from 5% and 95% quantiles. True values are indicated by black solid horizontal lines. Black circles and vertical bars refer to the results obtained for $n = 200$, whereas those for $n = 500$ and $n = 1000$ are given in dark gray and light gray, respectively.

	Bias			RMSE		
	$n = 200$	$n = 500$	$n = 1000$	$n = 200$	$n = 500$	$n = 1000$
β_{11}	-0.029	-0.001	-0.003	0.230	0.125	0.086
β_{21}	0.023	-0.010	-0.011	0.306	0.174	0.121
β_{31}	-0.670	-0.098	-0.040	1.451	0.305	0.176
h_{10}	0.061	0.034	0.024	0.170	0.104	0.077
h_{20}	0.046	0.041	0.035	0.198	0.141	0.064
s_{11}	0.033	0.022	0.016	0.194	0.102	0.069
s_{21}	0.038	0.026	0.025	0.255	0.133	0.091
s_{31}	0.192	0.062	0.061	1.039	0.424	0.212

Table 4: Bias and root mean squared error (RMSE) obtained by applying the `gjrm()` to bivariate survival simulated data with mild censoring rates (about 42% and 33% for the two responses). 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 $n_s^{-1} \sum_{i=1}^{n_s} \sqrt{n_{rep}^{-1} \sum_{rep=1}^{n_{rep}} (\hat{s}_{rep,i} - s_i)^2}$, 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.

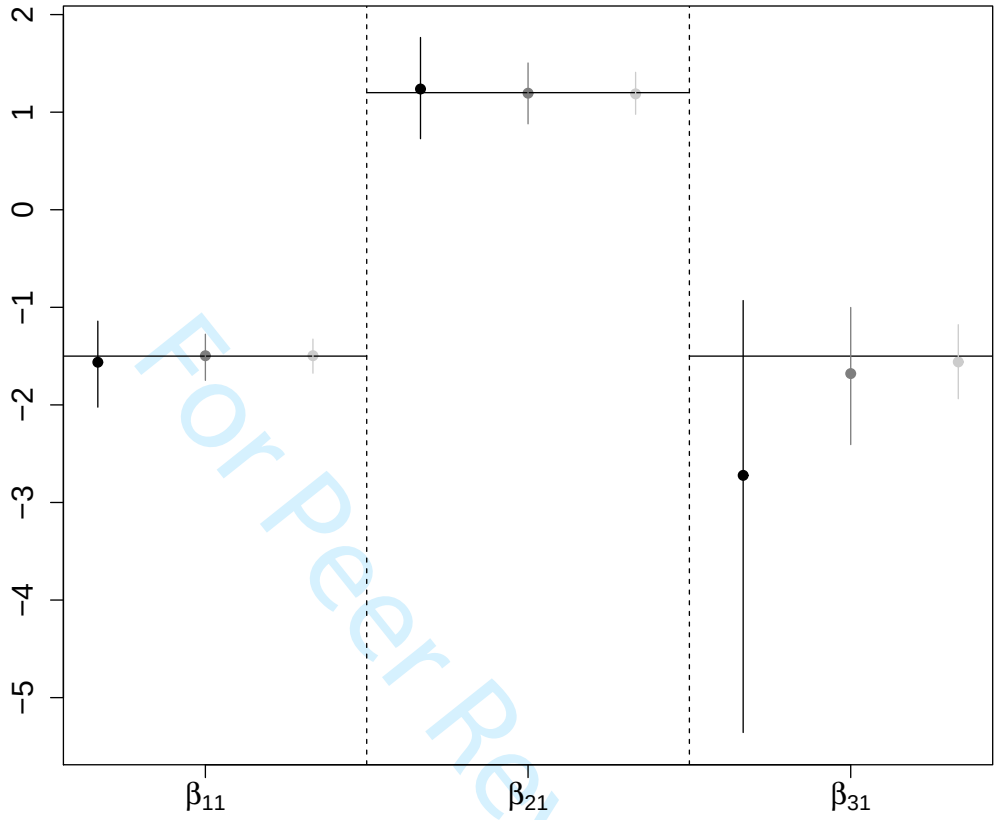

Figure 4: Linear coefficient estimates obtained by applying `gjrm()` to bivariate survival simulated data with high censoring rates (about 75% and 50% for the two responses). Further details are given in the caption of Figure 3.

	Bias			RMSE		
	$n = 200$	$n = 500$	$n = 1000$	$n = 200$	$n = 500$	$n = 1000$
β_{11}	-0.063	0.004	0.005	0.276	0.146	0.106
β_{21}	0.037	-0.005	-0.009	0.329	0.192	0.132
β_{31}	-1.222	-0.179	-0.059	2.581	0.474	0.245
h_{10}	0.246	0.138	0.083	0.378	0.247	0.158
h_{20}	0.181	0.088	0.070	0.420	0.205	0.173
s_{11}	0.047	0.036	0.023	0.251	0.129	0.084
s_{21}	0.039	0.034	0.032	0.331	0.164	0.109
s_{31}	0.336	0.265	0.080	1.089	0.764	0.275

Table 5: Bias and root mean squared error (RMSE) obtained by applying the `gjrm()` to bivariate survival simulated data with high censoring rates (about 75% and 50% for the two responses). Further details are given in the caption of Table 4.

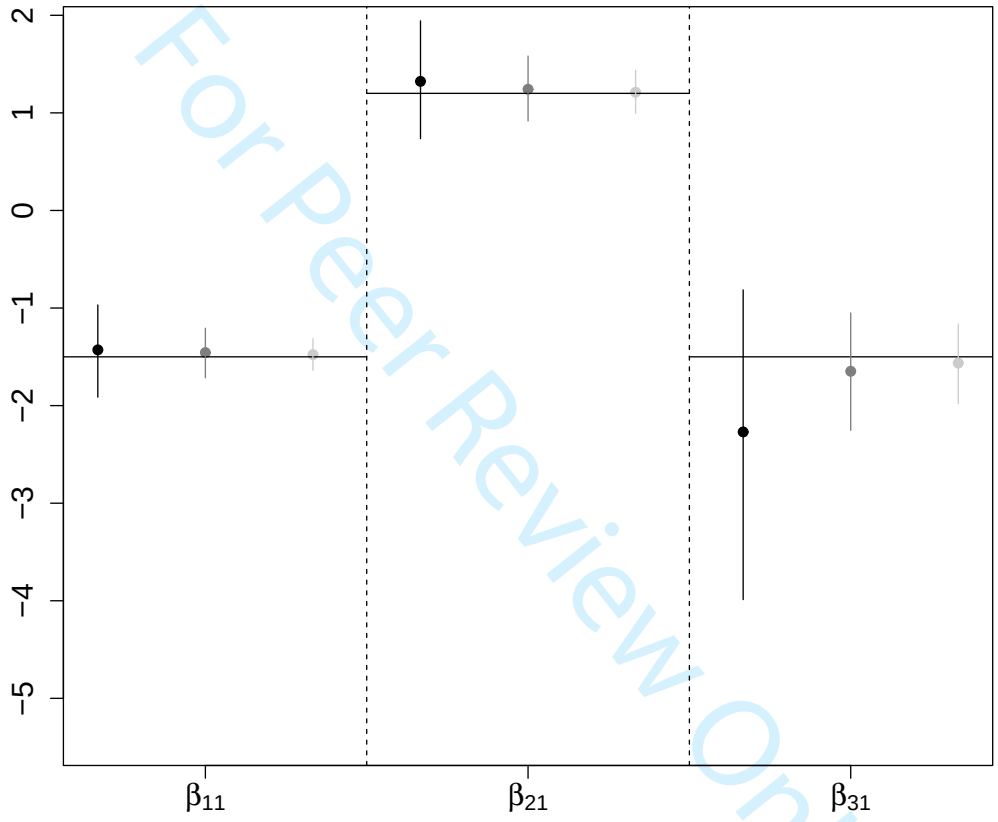

Figure 5: Linear coefficient estimates obtained by applying a two-stage estimation approach to bivariate survival simulated data with mild censoring rates (about 42% and 33% for the two responses). Further details are given in the caption of Figure 3.

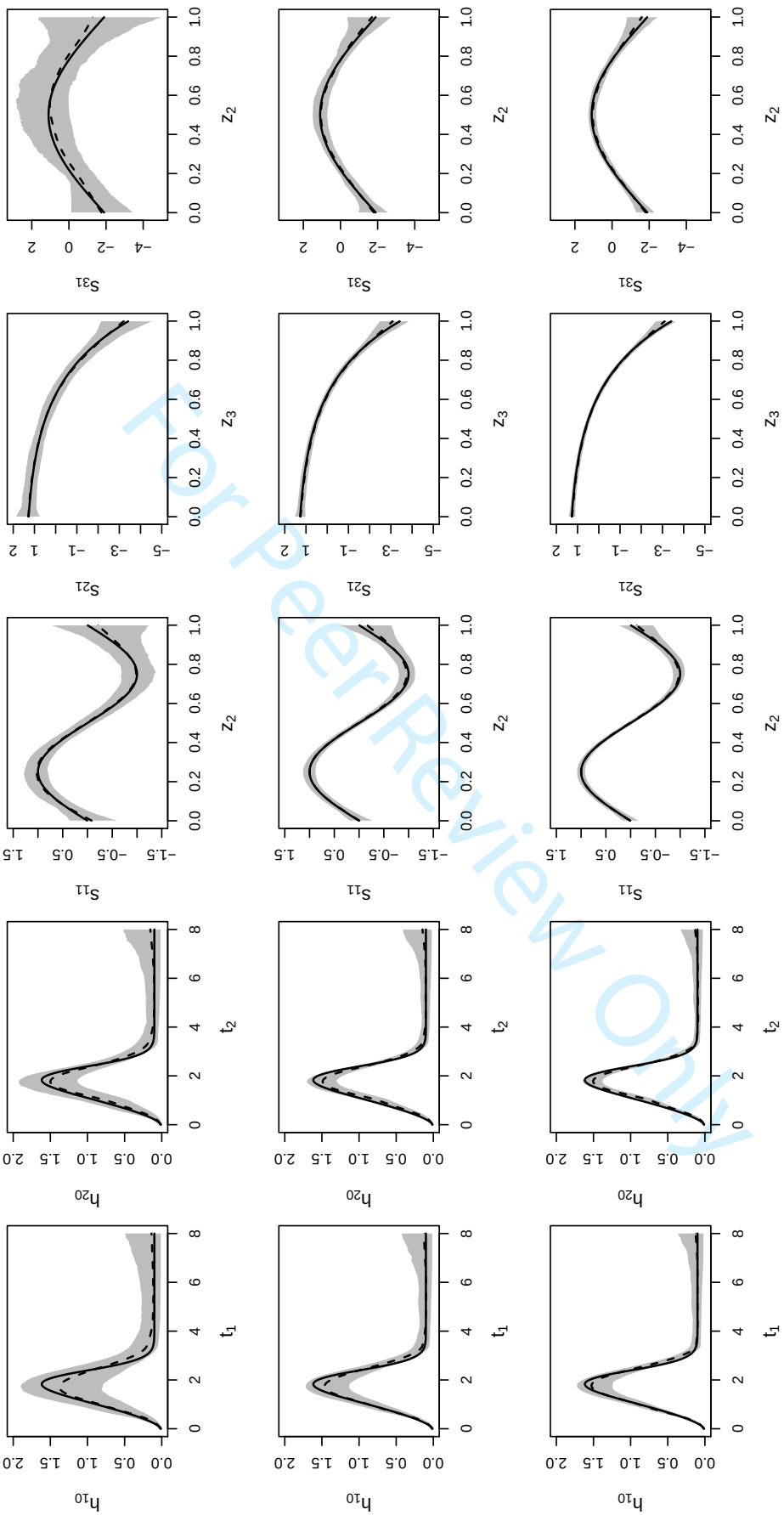

Figure 6: Smooth function estimates obtained by applying `gjrm()` to bivariate survival simulated data with mild censoring rates (about 42% and 33% for the two responses). True functions are represented by black solid lines, mean estimates by dashed lines and pointwise ranges resulting from 5% and 95% quantiles by shaded areas. The results in the first row refer to $n = 200$, whereas those in the second and third rows to $n = 500$ and $n = 1000$.

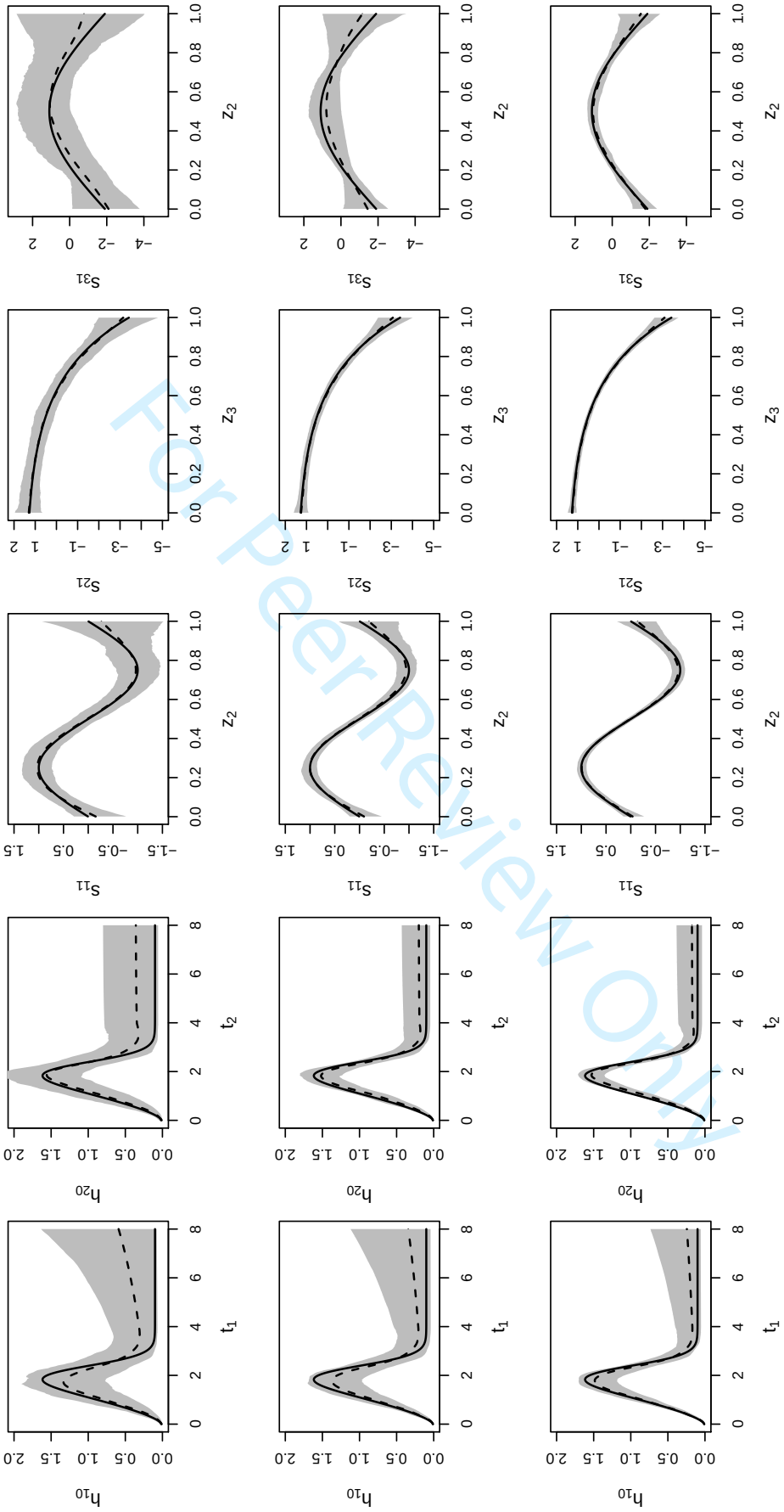

Figure 7: Smooth function estimates obtained by applying `gjrm()` to bivariate survival simulated data with high censoring rates (about 75% and 50% for the two responses). Further details are given in the caption of Figure 6.

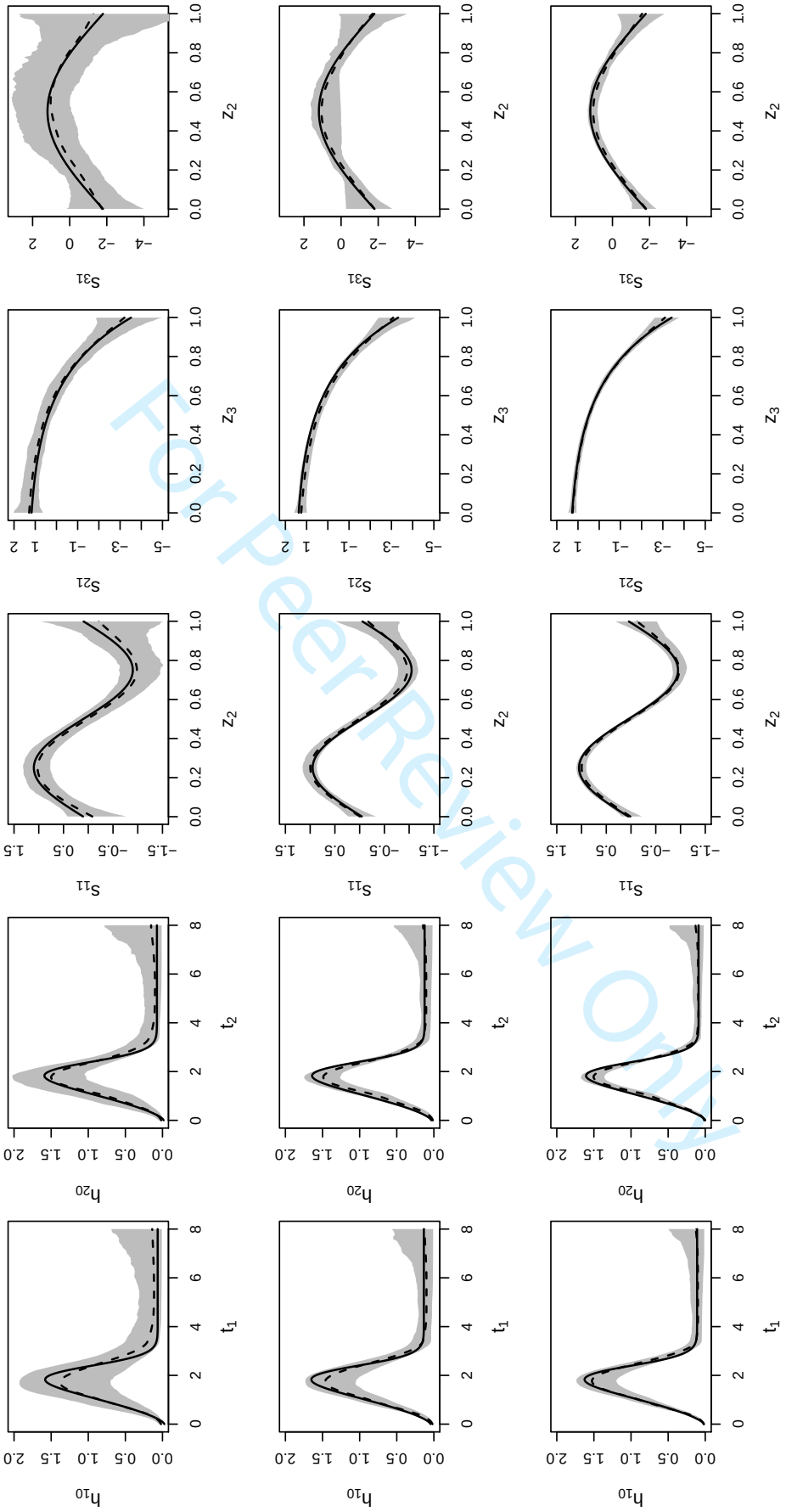

Figure 8: Smooth function estimates obtained by applying a two-stage estimation approach to bivariate survival simulated data with mild censoring rates (about 42% and 33% for the two responses). Further details are given in the caption of Figure 6.

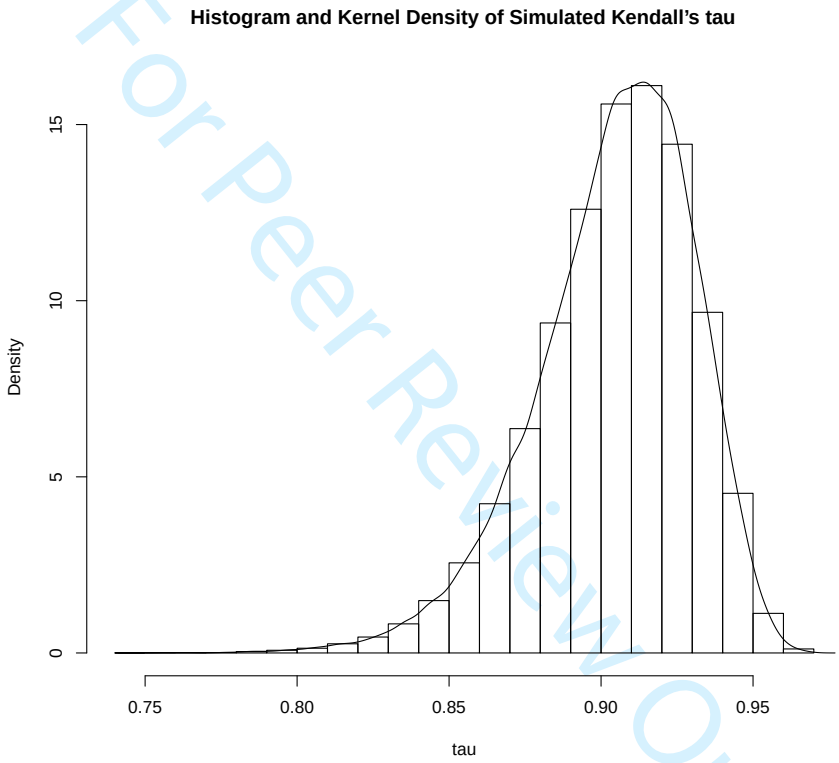

Figure 9: An example of histogram and kernel density estimates for 50000 simulated Kendall's τ values obtained using the Bayesian posterior simulation approach discussed in Section 2.5, after fitting the proposed model to bivariate survival simulated data with mild censoring rates and 200 observations.

Supplementary Material D

We have implemented the proposed models and estimation approach in `R` (R Development Core Team, 2019), by extending the `gjrm()` function within the package `GJRM` (Marra & Radice, 2019). This package has been created to enhance reproducible research as well as with transparent and straightforward dissemination of results in mind. The function is generally very easy to use, especially if the user is already familiar with the syntax of (generalized) linear and additive models in `R`. For instance, one of the calls used for the appendectomy analysis of this paper is

```
eq1 <- onset1 ~ s(log(onset1), bs = "mpi")
eq2 <- onset2 ~ s(log(onset2), bs = "mpi")
eq3 <-        ~ zyg
f.list <- list(eq1, eq2, eq3)
out <- gjrm(f.list, data = dat.fem, surv = TRUE,
            BivD = "T", margins = c("PH", "PH"),
            cens1 = app1, cens2 = app2, Model = "B")
```

where `onset1` is the age at appendectomy for twin 1 with censoring indicator `app1` (1 if the twin underwent appendectomy and 0 otherwise), and `zyg` is the type of zygosity (MZ or DZ). `onset2` and `app2` refer to twin 2. `dat.fem` is a data frame containing the variables in the model, `surv` must be set to `TRUE` in order to employ a joint bivariate survival model, `cens1` and `cens2` are the two censoring indicators, the possible choices for `BivD` and `margins` are given in Tables 1 and 2, `f.list` is a list of equations for the survival outcomes and the copula dependence parameter, and argument `bs` specifies the type of spline basis (e.g., `tp` for thin plate regression spline (the default) and `mpi` for monotonic P-spline). Monotonic P-splines must always be used for smooth terms of the responses, otherwise the program will produce an error message. Model `summary()` and `plot()` functions work in a similar fashion as those of generalized linear and additive models, and `AIC()` and `BIC()` can be used in the usual manner. `post.check()` produces plots of the Cox-Snell residuals for the two marginal models, and `hazsurv.plot()` produces hazard and survival plots. More details and options can be found in the documentation of the `GJRM R` package.

Supplementary Material E

Model building in our modeling framework involves the choice of the copula function, of the pair of link functions and the selection of relevant covariates in the model's additive predictors. To this end, we recommend using the AIC, BIC, Cox-Snell residuals and hypothesis testing. The AIC and BIC are given by $-2\ell(\hat{\boldsymbol{\delta}}) + 2edf$ and $-2\ell(\hat{\boldsymbol{\delta}}) + \log(n)edf$, where the log-likelihood is evaluated at the penalized parameter estimates and $edf = \text{tr}(\hat{\mathbf{A}})$. The residuals are defined as $r_{vi} = -\log\left\{S_v(y_{vi}|\mathbf{x}_{vi};\hat{\boldsymbol{\beta}}_v)\right\} \sim \text{Exp}(1)$, $v = 1, 2$, $i = 1, \ldots, n$ and can be used as follows. Let us denote the observed cumulative hazard as $\hat{H}_{r_v}(r_{vi})$ (derived from the Kaplan-Meier estimate). If the model is correct then the plot of the pairs $\left\{r_{vi}, \hat{H}_{r_v}(r_{vi})\right\}$ will have a 45° slope. This plot provides an overall assessment of the model's goodness of fit and can not suggest the type of mis-specification when the points do not follow the reference line. Note that the above definition of residuals is the same as that employed for more standard survival models. In fact, no special definition is required here since the proposed model is essentially parametric.

As a possible strategy, the researcher could use same set of covariates in all equations and choose the copula and link functions using the AIC, BIC and Cox-Snell residuals. The same tools can then be used to select the most relevant covariates in the model's predictors (using stepwise backward and/or forward selection). To favor more parsimonious models, small differences in the AIC and BIC values of competing models can be assisted by looking at the significance of the estimated effects; for example, a covariate could be excluded if the respective effect's p-value is larger than 5% or 10%. The model building process can be simplified if the researcher wishes to include variables in the model based on some prior belief or knowledge, or wishes to employ a particular set of link functions for the sake of interpretation.