Survival regression models with dependent Bayesian nonparametric priors

Alan Riva-Palacio
Departamento de Probabilidad y Estadística, Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas
Universidad Nacional Autónoma de México
and
Fabrizio Leisen *
School of Mathematical Sciences
University of Nottingham
and
Jim Griffin
Department of Statistical Science
University College London

November 17, 2020

Abstract

We present a novel Bayesian nonparametric model for regression in survival analysis. Our model builds on the classical *neutral to the right* model of Doksum (1974) and on the *Cox proportional hazards* model of Kim and Lee (2003). The use of a vector of dependent Bayesian nonparametric priors allows us to efficiently

*Fabrizio Leisen was supported by the European Community's Seventh Framework Programme [FP7/2007-2013] under grant agreement no: 630677.

model the hazard as a function of covariates whilst allowing nonproportionality. The model can be seen as having competing latent risks. We characterize the posterior of the underlying dependent vector of completely random measures and study the asymptotic behavior of the model. We show how an MCMC scheme can provide Bayesian inference for posterior means and credible intervals. The method is illustrated using simulated and real data.

Keywords: Bayesian nonparametrics, Survival Analysis, Dependent Completely Random Measures.

1 Introduction

The statistical analysis of the, potentially censored, survival time to an event has a long history. Often, estimates of the effects of observed covariates on the survival time distribution are key statistical quantities of interest. For example, information about white blood cells may be useful for the prediction of the time to death of leukaemia patients. There are several standard regression models. The accelerated failure time (AFT) model takes into account the effect of a covariate by accelerating or decelerating over time its effect on the survival time (Buckley and James, 1979). Alternatively, a parametric effect for the covariates can be combined with a nonparametric estimate of a baseline distribution of the survival time. The most popular example of this type of model is the semiparametric Cox (1972) model which has had a substantial impact in statistical and medical research, being introduced in one of the most cited statistical papers of all time (Ryan and Woodall, 2005).

The Cox regression model assumes proportional hazards (PH) and can be easily fitted with partial likelihood methods (Cox, 1975). The combination of this inference method with the counting process formulation of the model (Andersen and Gill, 1982) has led to extensions to stratified analysis, proportional intensity models, frailty models, and so on (Therneau and Grambsch, 2000). The model also leads medical researchers to focus on differences in instantaneous risk (hazard) rather than mean or median survival as in common regression models. Under the PH assumption, the survival curves for any combinations of covariate values must have hazard functions that are proportional over time, i.e. have constant hazard ratios. This is sometimes not realistic. For example, if a treatment effect is negative at the beginning of a study and positive by the end. Failing to account for this can lead to poor model fits, particularly in the tails of the

survival distribution. Such problems can be addressed by including interactions with time or stratifying according to the treatment (Kalbfleisch and Prentice, 2011). However, these approaches can lead to difficulties with interpretation of effects. Alternatively, the structure of the model can be changed. For example, the proportional odds (PO) model relaxes the PH assumption of a constant hazard ratio by assuming hazard functions such that this property holds only when the time goes to infinity (Cheng et al., 1995; Murphy et al., 1997; Yang and Prentice , 1999).

From the Bayesian perspective, the analysis of survival data was one of the first areas of application of Bayesian nonparametric techniques, see Doksum (1974) and Ferguson (1974), and Hjort et al. (2010) for a review. Popular priors include the beta process prior for the cumulative hazard function (Hjort, 1990), the extended gamma process (Dykstra and Laud, 1981), and the wide-class of neutral to the right (NTR) distributions (Doksum, 1974).

In this paper, we focus on the NTR model which assumes that the survival function of a survival time of interest Y, which is $S(t) = \mathbb{P}[Y > t]$, is given by

$$S(t) = e^{-\mu(0,t]}. \tag{1}$$

where μ is a completely random measure (CRM) (Kingman, 1967) for which $\mu(\mathbb{R}^+) \overset{\text{a.s.}}{=} \infty$ to ensure that the distribution of Y is supported in $\mathbb{R}^+$. As noted in Doksum (1974), such distribution is *neutral to the right* in the sense that if $F(t) = 1 - S(t)$ is the associated cumulative distribution function then

$$F(t_1),\ \frac{F(t_2) - F(t_1)}{1 - F(t_1)},\ \ldots,\ \frac{F(t_k) - F(t_{k-1})}{1 - F(t_{k-1})} \tag{2}$$

are independent for every $t_1 < \cdots < t_k$. The structure is very general and includes the Dirichlet process (Ferguson, 1973) and Beta-Stacy process (Muliere and Walker, 1997)

as special cases. The family of NTR distributions has desirable theoretical properties with survival data such as conjugacy for right-censored data (Ferguson and Phadia, 1979) and posterior consistency at an optimal rate (Kim and Lee, 2004). They are also a natural Bayesian nonparametric analogue of the widely-used frequentist Kaplan-Meier estimator. The approach was extended to multiple samples by Epifani and Lijoi (2010) and Riva-Palacio and Leisen (2018) and to Cox regression modelling by Kim and Lee (2003). Their model assumes that the survival function, $S_{\boldsymbol{X}}(t)$, for covariate value $\boldsymbol{X} = (X_1, \ldots, X_m) \in \mathbb{R}^m$ is modeled by

$$S_{\boldsymbol{X}}(t) = \mathbb{P}[Y > t \,|\boldsymbol{X}] = e^{-e^{\langle \boldsymbol{\beta}, \boldsymbol{X} \rangle} \mu(0,t]}.$$

Alternatively, Bayesian nonparametric regression survival models can be built by modelling the logarithm of the survival time using a dependent nonparametric prior. These allow for crossing survival and hazard curves and have straightforward interpretations. For example, the linear dependent Dirichlet process mixture (LDDP) (De Iorio et al., 2009) uses dependent Dirichlet processes (MacEachern, 1999) and the linear dependent tailfree process (Jara and Hanson, 2011) builds on Pólya tree priors. These approaches are reviewed by Hanson and Jara (2013). Other approaches to non-proportional hazards include Nieto-Barajas (2014) who introduces a semiparametric model based on increasing additive processes, Nipoti et al. (2018) who propose a partially proportional hazards model using cluster-dependent random hazards, and Fernández et al. (2016) who model the hazard as a logistic transform of a sum of Gaussian processes.

In this paper we build a tractable Bayesian nonparametric regression model for survival data based on the class of NTR distributions. The model assumes that

$$S_{\boldsymbol{X}}(t) = \mathbb{P}[Y > t \,|\boldsymbol{X}] = e^{-f_1(\boldsymbol{\beta}, \boldsymbol{X}) \mu_1(0,t] - \ldots - f_d(\boldsymbol{\beta}, \boldsymbol{X}) \mu_d(0,t]}$$

where $f_1, \ldots, f_d$ are known functions of the covariates $\boldsymbol{X}$, $\boldsymbol{\beta}$ are unknown parameters and $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_d)$ is a vector of completely random measures (VCRM). This extends the proportional hazards Cox regression model of Kim and Lee (2001) to allow for more general functions of the covariates and can be interpreted in a competing risks framework. Vectors of completely random measures have proven to be a useful tool for inducing dependence in Bayesian nonparametric priors; see for example Lijoi et al. (2014), Camerlenghi et al. (2019a) and Camerlenghi et al. (2019b). We term this model a *generalized additive neutral to the right regression* (GANTR) model.

GANTR provides a flexible regression modelling approach within an NTR model. It allows for non-proportional hazards and leads to clustering of observations into subpopulations (associated with different causes in the competing risks interpretation) according to covariate values. The model can be seen as a generalization of the multiple-sample model of Riva-Palacio and Leisen (2018) to allow for stratification into unknown covariate dependent clusters/sub-samples. The structure of the prior allows us to develop a posterior characterization and use it to construct an inference scheme that depends on the VCRM through its Laplace exponent. We concentrate on the class of compound random measures (Griffin and Leisen, 2017), and develop both an MCMC method and empirical Bayes method for estimating the hyperparameters. Unlike AFT models, the NTR approach models the survival function directly which eases the interpretation of the overall model and, particularly, the covariate effects. The tractability of the GANTR model allows us to derive the likelihood of the regression coefficients and the hyperparameters of the VCRM and so implement fast *maximum a posteriori* inference methods. We also find that the GANTR model leads to better fit of the data than the LDDP in both simulated and real data.

The outline of the paper is as follows. In the next section we consider CRMs and VCRMs in more detail. In Section 3 we formally present the GANTR model. Section 4 develops a posterior characterization of the model and results necessary for the ensuing inference procedures. We present simulation and real data studies for our model in Section 5. Conclusions for the work are presented in Section 6. Proofs, further properties regarding asymptotic behavior, details regarding inference and further simulation studies are included in the supplementary material. Code for our model is available in https://github.com/alan7riva/GANTR.

2 Preliminaries

VCRMs are a key building block of our proposed Bayesian nonparametric model. This section will introduce some basic ideas and representations through Laplace functional transforms. We will focus on the compound random measure (CoRM) class of VCRMs which were recently introduced by Griffin and Leisen (2017).

Let $\mathbb{Y}$ be a complete and separable metric space with corresponding Borel σ-algebra $\mathcal{Y}$ and probability space $(\Omega, \mathcal{F}, \mathbb{P})$. We denote by $\mathbb{M}_{\mathbb{Y}}$ the space of boundedly finite measures in the measure space $(\mathbb{Y}, \mathcal{Y})$ and the associated Borel σ-algebra by $\mathcal{M}_{\mathbb{Y}}$.

Definition 1. A random measure μ on $(\mathbb{Y}, \mathcal{Y})$ is called a completely random measure (CRM) if for any collection of disjoint sets $\{A_i\}_{i=1}^n \subset \mathcal{Y}$ the random variables $\{\mu(A_i)\}_{i=1}^n$ are mutually independent.

In this paper, we restrict attention to CRMs of the form

$$\mu = \sum_{i=1}^{\infty} w_i \delta_{Y_i}$$

where $\{w_i\}_{i=1}^{\infty}$ and $\{Y_i\}_{i=1}^{\infty}$ are collections of random variables taking values in $\mathbb{R}^+$ and $\mathbb{Y}$ respectively. We will refer to w_i as jump heights and Y_i as jump locations. Such CRMs can be characterized by their Laplace transform

$$\mathbb{E}\left[e^{-\lambda\mu(A)}\right] = e^{-\int_{\mathbb{R}^+\times A}(1-e^{-\lambda s})\nu(\mathrm{d}s,\mathrm{d}y)} \tag{3}$$

where $\lambda > 0$ and ν is a measure on $\mathbb{R}^+ \times \mathbb{Y}$ such that

$$\int_{\mathbb{R}^+\times A} \min\{s,1\}\nu(\mathrm{d}s,\mathrm{d}y) < \infty$$

for any bounded set $A \in \mathcal{Y}$. The measure ν is usually called the Lévy intensity of μ. We denote the Laplace exponent of a CRM as ψ_t where for $t, \lambda \in \mathbb{R}^+$

$$\psi_t(\lambda) = -\log\left(\mathbb{E}\left[e^{-\lambda\mu(0,t]}\right]\right).$$

See Kingman (1967) for a full review of CRMs. We say that a Lévy intensity ν is homogeneous if it can be written in the form

$$\nu(\mathrm{d}s,\mathrm{d}y) = \rho(\mathrm{d}s)\kappa(\mathrm{d}y)$$

where κ is a non-atomic measure on $\mathbb{Y}$ referring to the jump locations and ρ is a measure on $\mathbb{R}^+$ referring to the jump heights. For example, we will use the homogeneous CRM with Lévy intensity

$$\nu(\mathrm{d}s,\mathrm{d}y) = \frac{\gamma e^{-\alpha s}}{s}\mathrm{d}s\,\kappa(\mathrm{d}y), \tag{4}$$

where $\gamma > 0$ and $\alpha > 0$ which is a particular case of the Gamma process, see Phadia (2015). We refer to this process as the Gamma CRM which is denoted Gamma(α,γ). There is a natural generalization of CRMs to the multivariate setting.

Definition 2. A vector $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_d)$ of random measures on $(\mathbb{Y}, \mathcal{Y})$ is called a vector of completely random measures (VCRM) if, for any collection of disjoint sets $\{A_i\}_{i=1}^n \subset \mathcal{Y}$, the random vectors $\{(\mu_1(A_i), \ldots, \mu_d(A_i))\}_{i=1}^n$ are mutually independent.

The corresponding multivariate analogue of the Laplace transform (3) is

$$\mathbb{E}\left[e^{-\lambda_1\mu_1(A)-\ldots-\lambda_d\mu_d(A)}\right] = e^{-\int_{(\mathbb{R}^+)^d\times A}(1-e^{-\langle\boldsymbol{\lambda},\boldsymbol{s}\rangle})\nu(\mathrm{d}\boldsymbol{s},\mathrm{d}y)}$$

where $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_d) \in (\mathbb{R}^+)^d$ and ν is a measure on $(\mathbb{R}^+)^d \times \mathbb{Y}$ satisfying

$$\int_{(\mathbb{R}^+)^d\times A} \min\{\|\boldsymbol{s}\|, 1\}\nu(\mathrm{d}\boldsymbol{s}, \mathrm{d}y) < \infty \tag{5}$$

for any bounded $A \in \mathcal{Y}$. A d-dimensional VCRM with such Laplace transform can be represented as

$$\boldsymbol{\mu} = \left(\sum_{i=1}^{\infty} w_{1,i}\delta_{Y_i}, \ldots, \sum_{i=1}^{\infty} w_{d,i}\delta_{Y_i}\right)$$

for a random collection of vectors $\{(w_{1,i}, \ldots, w_{d,i})\}_{i=1}^{\infty}$ taking values in $(\mathbb{R}^+)^d$ and $\{Y_i\}_{i=1}^{\infty}$ taking values in $\mathbb{Y}$. The associated Laplace exponent for $t \in \mathbb{R}^+$ and $\boldsymbol{\lambda} = (\lambda_1, \cdots, \lambda_d) \in (\mathbb{R}^+)^d$ is now given by

$$\psi_t(\boldsymbol{\lambda}) = -\log\left(\mathbb{E}\left[e^{-\lambda_1\mu_1((0,t])-\ldots-\lambda_d\mu_d((0,t])}\right]\right).$$

The corresponding homogeneous case arises when

$$\nu(\mathrm{d}\boldsymbol{s}, \mathrm{d}y) = \rho(\mathrm{d}\boldsymbol{s})\kappa(\mathrm{d}y).$$

Griffin and Leisen (2017) introduced a flexible class of VCRMs called compound random measures where the dependence structure of the vector is modeled in a constructive way.

Definition 3. Given a d-variate probability density function h and an univariate Lévy intensity $\nu^\star$ we say that a d-variate VCRM $\boldsymbol{\mu}$ is a compound random measure (CoRM) with score distribution h and directing Lévy measure $\nu^\star$ if the d-variate Lévy intensity of $\boldsymbol{\mu}$ is given by

$$\nu(\mathrm{d}\boldsymbol{s}, \mathrm{d}y) = \int_{\mathbb{R}^+} z^{-d} h\left(\frac{s_1}{z}, \ldots, \frac{s_d}{z}\right) \nu^\star(\mathrm{d}z, \mathrm{d}y)\mathrm{d}\boldsymbol{s}.$$

In Riva-Palacio and Leisen (2019), the existence of marginal first moments for the score distribution in a CoRM is shown to be sufficient for the integrability condition (5) to be satisfied; in this work we only consider CoRMs with such score distributions. Furthermore Riva-Palacio and Leisen (2019) shows that CoRMs have an elegant interpretation in terms of discrete measures. Indeed, if μ is a homogeneous univariate CRM with Lévy intensity $\nu^\star(\mathrm{d}s, \mathrm{d}y)$ and series representation

$$\mu(\cdot) \overset{\text{a.s.}}{=} \sum_{j=1}^{\infty} w_j \delta_{u_j}(\cdot)$$

for random sequences $\{w_i\}_{i=1}^{\infty}$ in $\mathbb{R}^+$ and $\{u_i\}_{i=1}^{\infty}$ in $\mathbb{Y}$, and if $\{\boldsymbol{v}_j = (v_{1,j}, \ldots, v_{d,j})\}_{j=1}^{\infty}$ is an independent identically distributed (i.i.d.) sequence with common distribution h; then the associated CoRM $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_d)$ is such that

$$\mu_i(\cdot) \overset{\text{a.s.}}{=} \sum_{j=1}^{\infty} v_{i,j} w_j \delta_{u_j}(\cdot). \tag{6}$$

An interesting example of a CoRM is defined by a LogNormal$(\boldsymbol{m}, \boldsymbol{\Sigma})$, score distribution, with mean vector $\boldsymbol{m}$ and covariance matrix $\boldsymbol{\Sigma}$. Such choice allows us to distribute the mass of the directing Lévy intensity across the d-dimensional space of the CoRM intensity. The Lévy intensity of a CoRM with such score distribution and Gamma directing Lévy measure is presented in the supplementary material. In particular we will use the following construction.

Definition 4. We say that a d-dimensional random variable Z is given a δ-LogNormal distribution if its probability density function is

$$p(z) = \frac{1}{d}\sum_{i=1}^{d} \text{LogNormal}\left(z \,\middle|\, (1-\delta)\boldsymbol{e}_i + \delta\mathbf{1}\,,\, \sigma \mathrm{I}^{(d)}\right).$$

where $\text{LogNormal}(z|\boldsymbol{m},\boldsymbol{\Sigma})$ is the probability density function of a multivariate lognormal distribution using the parameterization discussed at the end of Section 2, $\delta \in (0,1]$, $\{\boldsymbol{e}_i\}_{i=1}^d$ is the canonical basis in $\mathbb{R}^d$, $\mathbf{1} = (1,\ldots,1) \in \mathbb{R}^d$, $\sigma > 0$ and $\mathrm{I}^{(d)}$ is the d-dimensional identity matrix.

We can use the above as the score distribution in a CoRM with Gamma directing Lévy measure. Observe that when using a mixture for the score distribution of a CoRM the Lévy intensity is a sum of the Lévy intensities corresponding to the mixture components with the directing Lévy measure fixed.

The δ-LogNormal is a d component mixture model. For small values of σ, the effect of the parameter δ is twofold. Firstly, when used as the score distribution in a CoRM, it controls the dependence between dimensions of the VCRM. If $\delta = 1$ the mass of the distribution is accumulated near the point $\mathbf{1}$ and the related CoRM has a Lévy intensity which accumulates mass near the identity axis. This CoRM is close to a completely dependent VCRM where all dimensions are almost surely equal. On the other hand, if $\delta \to 0$, the δ-LogNormal distribution accumulates mass near the points $\{\boldsymbol{e}_i\}_{i=1}^d$ and the related CoRM Lévy intensity accumulates mass near the axes in $(\mathbb{R}^+)^d$, which will be close to an independent entries VCRM. Values of $\delta \in (0,1)$ will modulate between these distributions and VCRMs. Secondly, in the multiple-sample information setting where the regression functions $f_i(\boldsymbol{\beta}, \boldsymbol{X}_k) = \mathbb{1}_{\{\boldsymbol{X}_{k,j}=1\}}$ for $i = 1,\ldots,d$, then as $\delta \to 0$ and $\sigma \to 0$ the GANTR model is equivalent to an NTR model for each sample. While if $\delta = 1$ and

$\sigma \to 0$ the GANTR model is equivalent to a NTR model. The parameter σ serves to diffuse the mass of the distribution. The modulating effect of δ decreases as σ increases so we chose a relatively small value of σ.

3 Survival regression model

The neutral to the right process (Doksum, 1974) for the survival function was defined in (1) as the exponential transform of a CRM μ for which $\mu(\mathbb{R}^+) \overset{\text{a.s.}}{=} \infty$. We say that a random variable Y with this survival function has a *neutral to the right* distribution, which is denoted $Y \sim \text{NTR}(\mu)$, where μ is a CRM.

Definition 5. Let $n, m, d, b \in \mathbb{N} \setminus \{0\}$, $\hat{Y} = \{Y_i\}_{i=1}^n$ with $Y_i \in \mathbb{R}^+$, and $\hat{\boldsymbol{X}} = \{\boldsymbol{X}_i\}_{i=1}^n$ with $\boldsymbol{X}_i = (X_{i,1}, \ldots, X_{i,m}) \in \mathbb{R}^m$ be a random sample. We say $\{\hat{Y}, \hat{\boldsymbol{X}}\}$ follow a generalized additive neutral to the right regression (GANTR) model if, for $\boldsymbol{t} = (t_1, \ldots, t_n) \in (\mathbb{R}^+)^n$, the joint survival function is

$$S_{\hat{\boldsymbol{X}}}(\boldsymbol{t}) = \mathbb{P}\left[Y_1 > t_1, \ldots, Y_n > t_n \,|\hat{\boldsymbol{X}}\right] = \prod_{i=1}^{n} e^{-f_1(\boldsymbol{\beta}, \boldsymbol{X}_i)\mu_1(0,t_i] - \ldots - f_d(\boldsymbol{\beta}, \boldsymbol{X}_i)\mu_d(0,t_i]} \tag{7}$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_b) \in \mathbb{R}^b$, $\boldsymbol{\mu} = \{\mu_1, \ldots, \mu_d\}$ are a VCRM with Lévy intensity $\nu_{\boldsymbol{c}}(d\boldsymbol{s}, dt)$, where $\boldsymbol{c}$ are parameters of the Lévy measure, and $f_i : \mathbb{R}^b \times \mathbb{R}^m \to \mathbb{R}^+$ for $i \in \{1, \ldots, d\}$.

The GANTR model for a single observation can be seen as an NTR distribution conditionally on the covariates $\boldsymbol{X}_i$

$$Y_i | \boldsymbol{X}_i \overset{\text{ind}}{\sim} \text{NTR}\left(\sum_{j=1}^{d} f_j(\boldsymbol{\beta}, \boldsymbol{X}_i)\mu_j\right) \tag{8}$$

which allows us to use results about NTR processes with our model. For example, neutrality to the right as in equation (2) is satisfied conditionally on the covariates, as in Proposition 3 of Riva-Palacio and Leisen (2018)

Some previously proposed models arise as special cases of the GANTR model. If $\boldsymbol{X}_i \in \{0,1\}^m$ such that $\boldsymbol{X}_{i,j} = 1$ if and only if the i-th observation belongs to the j-th sample, the multiple-sample NTR model of Riva-Palacio and Leisen (2018) for m samples can be recovered by choosing $f_j(\boldsymbol{\beta}, \boldsymbol{X}_i) = \mathbb{1}_{\{\boldsymbol{X}_{i,j}=1\}}$. The Cox NTR model of Kim and Lee (2003) arises when $d = 1$ and $f_1(\boldsymbol{\beta}, \boldsymbol{X}) = e^{\langle \boldsymbol{\beta}, \boldsymbol{X} \rangle}$. Unlike the Cox NTR model, the GANTR model allows for nonproportional hazards. Indeed, if $S_{\boldsymbol{X}_1}(t)$ and $S_{\boldsymbol{X}_2}(t)$ are the survival functions at time t of GANTR distributed random variables Y_1, Y_2 with respective covariates $\boldsymbol{X}_1 = (X_{1,1}, \ldots, X_{1,m})$, $\boldsymbol{X}_2 = (X_{2,1}, \ldots, X_{2,m})$, such that $\boldsymbol{X}_1 \neq \boldsymbol{X}_2$ then

$$S_{\boldsymbol{X}_1}(t) - S_{\boldsymbol{X}_2}(t) = S_{\boldsymbol{X}_1}(t) \left(1 - \prod_{i=1}^{d} e^{-r_i \mu_i(0,t]} \right)$$

where $r_i = f_i(\boldsymbol{\beta}, \boldsymbol{X}_2) - f_i(\boldsymbol{\beta}, \boldsymbol{X}_1)$, $i \in \{1, \ldots, d\}$, and, clearly, the survival functions for different covariates values can cross at any point $t \in \mathbb{R}^+$ if $d > 1$.

The GANTR model has been motivated as a flexible model of the effects of covariates on the survival function but it can also be viewed as a competing risks model (Prentice et al., 1978). We assume d independent latent causes for the event of interest and define the survival function for the j-th cause with covariates $\boldsymbol{X}$ to be

$$\tilde{S}_{j,\boldsymbol{X}}(t) = e^{-f_j(\boldsymbol{\beta}, \boldsymbol{X}) \mu_j(0,t]}.$$

The survival function for all risks across a sample is the GANTR model. The structure of models means that f_j and μ_j are not separately identified (although, their product is identified). This is not necessarily a problem for Bayesian inference if we are only

interested in functions of these products (such as the survival function) and a prior can be placed on β and μ_j. Alternatively, we can choose a parameterization which identifies each product (for example, by fixing the value of $f_j(\boldsymbol{\beta}, \boldsymbol{X}_0)$ for a specific covariate value $\boldsymbol{X}_0$). The survival function for all risks for a covariate $\boldsymbol{X}$ be re-expressed as

$$S_{\boldsymbol{X}}(t) = e^{-\sum_{j=1}^{d} f_j(\boldsymbol{\beta},\boldsymbol{X})\mu_j(0,t]} = e^{-f^\star(\boldsymbol{\beta},\boldsymbol{X})\sum_{j=1}^{d} w_j(\boldsymbol{\beta},\boldsymbol{X})\mu_j(0,t]}, \tag{9}$$

where $w_j(\boldsymbol{\beta}, \boldsymbol{X}) = f_j(\boldsymbol{\beta}, \boldsymbol{X}) / \sum_{k=1}^{d} f_k(\boldsymbol{\beta}, \boldsymbol{X})$ and $f^\star(\boldsymbol{\beta}, \boldsymbol{X}) = \sum_{k=1}^{d} f_k(\boldsymbol{\beta}, \boldsymbol{X})$. Under the latter parametrization, the w_j's are weightings on each latent cause (which depends on covariates) and allow departures from a Cox proportional hazards models which occurs if $w_j(\boldsymbol{\beta}, \boldsymbol{X})$'s do not depend on $\boldsymbol{X}$. Observations which have similar $w_j(\boldsymbol{\beta}, \boldsymbol{X})$ will have similar survival curves and this allows us to define subpopulations which tend to have similar survival outcomes. We illustrate this idea in Section 5 using data from melanoma patients. This parameterization can be identified by fixing the value of $f^\star(\boldsymbol{\beta}, \boldsymbol{X}_0)$ at covariate value $\boldsymbol{X}_0$. The competing risk interpretation also leads to a simple simulation scheme for our model. We sample $\tilde{Y}_i$ as survival times according to the survival function $\tilde{S}_{j,\boldsymbol{X}}(t)$ and set $Y = \min\{\tilde{Y}_1, \ldots, \tilde{Y}_d\}$.

The GANTR model with a CoRM chosen as the VCRM $\boldsymbol{\mu}$, can be represented as a conditional NTR model where, by substituting (6) into (8), $Y \sim \text{NTR}(\mu^\star)$ where

$$\mu^\star = \sum_{j=1}^{\infty} \left(\sum_{i=1}^{d} f_i(\boldsymbol{\beta}, \boldsymbol{X}) m_{i,j} \right) w_j \delta_{u_j}.$$

The measure $\mu^\star$ is a CoRM with the same directing Lévy measure as $\mu_1, \ldots, \mu_d$ and scores $\sum_{i=1}^{d} f_i(\boldsymbol{\beta}, \boldsymbol{X}) m_{i,j}$. In this form, the score distribution of the CoRM is marginally a random linear combination of the regression functions $f_1, \ldots, f_d$.

4 Posterior characterization

The form of the GANTR model allows us to derive an analytic expression for the posterior distribution given right-censored data. This result allows us to construct an inference scheme for the model as explained in Section 5. We assume that a sample of size n is observed and that there are (right) censoring times $C_1, \ldots, C_n$ which are i.i.d. and independent of the survival times $Y_1, \ldots, Y_n$. We observe the time $T_i = \min\{Y_i, C_i\}$ and the indicator variable $J_i = \mathbb{1}_{(0,C_i]}(Y_i)$ which is 0 if the i-th observation is censored. Let $\boldsymbol{\mathcal{D}} = \{(T_i, J_i, \boldsymbol{X}_i)\}_{i=1}^n$ be the survival data available for analysis. The $k \leq n$ order statistics (without repetition) of $T_1, \ldots, T_n$ are represented by $T_{(1)} < \cdots < T_{(k)}$ and define $T_{(0)} = 0$ and $T_{(k+1)} = \infty$. The number of right-censored observations (for which $J_i = 0$) and exact observations (for which $J_i = 1$) at time $T_{(j)}$ are n_j^c and n_j^e respectively. The indices of the right-censored and exact observations at time $T_{(j)}$ are

$$I_j^{(e)} = \{l : T_l = T_{(j)} \text{ and } J_l = 1\} \text{ and } I_j^{(c)} = \{l : T_l = T_{(j)} \text{ and } J_l = 0\}$$

respectively. The indices of all exact observations is $I^{(e)} = \cup_{j=1}^k I_j^{(e)}$. It is useful to define the pairs of functions, for $\boldsymbol{b} \in \mathbb{R}^b$ and $i \in \{1, \ldots, d\}$,

$$h_{i,j}^{(e)}(\boldsymbol{b}, \hat{\boldsymbol{X}}) = \sum_{l \in I_j^{(e)}} f_i(\boldsymbol{b}, \boldsymbol{X}_l), \qquad \bar{h}_{i,j}^{(e)}(\boldsymbol{b}, \hat{\boldsymbol{X}}) = \sum_{r=j}^{k} h_{i,r}^{(e)}(\boldsymbol{b}, \hat{\boldsymbol{X}})$$

and

$$h_{i,j}^{(c)}(\boldsymbol{b}, \hat{\boldsymbol{X}}) = \sum_{l \in I_j^{(c)}} f_i(\boldsymbol{b}, \boldsymbol{X}_l), \qquad \bar{h}_{i,j}^{(c)}(\boldsymbol{b}, \hat{\boldsymbol{X}}) = \sum_{r=j}^{k} h_{i,r}^{(c)}(\boldsymbol{b}, \hat{\boldsymbol{X}})$$

for $j \in \{1, \ldots, k\}$ and $\bar{h}_{i,k+1}^{(e)}(\boldsymbol{b}, \hat{\boldsymbol{X}}) = \bar{h}_{i,k+1}^{(c)}(\boldsymbol{b}, \hat{\boldsymbol{X}}) = 0$. We group these functions in the vectors $\bar{\boldsymbol{h}}_j^{(e)}(\boldsymbol{b}, \hat{\boldsymbol{X}}) = \left(\bar{h}_{1,j}^{(e)}(\boldsymbol{b}, \hat{\boldsymbol{X}}), \ldots, \bar{h}_{d,j}^{(e)}(\boldsymbol{b}, \hat{\boldsymbol{X}})\right)$ and $\bar{\boldsymbol{h}}_j^{(c)}(\boldsymbol{b}, \hat{\boldsymbol{X}}) = \left(\bar{h}_{1,j}^{(c)}(\boldsymbol{b}, \hat{\boldsymbol{X}}), \ldots, \bar{h}_{d,j}^{(c)}(\boldsymbol{b}, \hat{\boldsymbol{X}})\right)$.

Initially, we derive the likelihood of right-censored data $\boldsymbol{\mathcal{D}}$ in the GANTR model. We assume the following condition on the GANTR model.

Condition 1. the VCRM $\boldsymbol{\mu}$ has a Lévy intensity $\nu(\boldsymbol{s}, \mathrm{d}y)\mathrm{d}\boldsymbol{s}$ such that $\eta_t(\boldsymbol{s}) = \nu\left(\boldsymbol{s}, (0, t]\right)$ is differentiable in the sense that the partial derivative $\eta'_{t_0}(\boldsymbol{s}) = \partial\eta_t(\boldsymbol{s})/\partial t\big|_{t=t_0}$ exists and, as $s \to \infty$, $\eta'_t(s) = \mathcal{O}(\exp(ks))$ with $k < \min_{j\in I^{(e)}}\left\{\sum_{i=1}^{d}\left(\bar{h}^{(e)}_{j+1,i}(\boldsymbol{\beta}, \hat{\boldsymbol{X}}) + \bar{h}^{(c)}_{j,i}(\boldsymbol{\beta}, \hat{\boldsymbol{X}})\right)\right\}$ for any $t_0 > 0$.

This is a weak condition and is equivalent to requiring that the derivative of $\kappa(t)$ exists in the homogeneous case.

In the following result we provide a convenient expression for the likelihood of $\boldsymbol{\beta}$, the regression coefficients, and $\boldsymbol{c}$, the hyperparameters of the Lévy intensity, in the GANTR model. We want to emphasize the dependence of the Lévy intensity on $\boldsymbol{c}$ so in the following proposition we use the particular notation $\nu_{\boldsymbol{c}}$, $\eta'_{t,\boldsymbol{c}}$ and $\psi_{t,\boldsymbol{c}}$ for the Lévy intensity, partial derivative as above and Laplace exponent, respectively.

Proposition 1. *Let $\boldsymbol{\mathcal{D}}$ be survival data and assume a GANTR model with Condition 1. Let $\psi_{t,\boldsymbol{c}}$ be the Laplace exponent associated to $\nu_{\boldsymbol{c}}$, then the likelihood of $\boldsymbol{\beta}$ and $\boldsymbol{c}$ is given by*

$$
\begin{aligned}
l(\boldsymbol{\beta}, \boldsymbol{c}; \boldsymbol{\mathcal{D}}) = {} & e^{-\sum_{j=1}^{k}\left(\psi_{T_{(j)},\boldsymbol{c}}\left(\bar{\boldsymbol{h}}_j^{(c)}(\boldsymbol{\beta},\hat{\boldsymbol{X}})+\bar{\boldsymbol{h}}_j^{(e)}(\boldsymbol{\beta},\hat{\boldsymbol{X}})\right)-\psi_{T_{(j-1)},\boldsymbol{c}}\left(\bar{\boldsymbol{h}}_j^{(c)}(\boldsymbol{\beta},\hat{\boldsymbol{X}})+\bar{\boldsymbol{h}}_j^{(e)}(\boldsymbol{\beta},\hat{\boldsymbol{X}})\right)\right)} \\
& \times \prod_{j\in I^{(e)}}\left\{\int_{(\mathbb{R}^+)^d}\prod_{i=1}^{d}\left(e^{-\left(\bar{h}^{(e)}_{j+1,i}(\boldsymbol{\beta},\hat{\boldsymbol{X}})+\bar{h}^{(c)}_{j,i}(\boldsymbol{\beta},\hat{\boldsymbol{X}})\right)s_i}\right)\prod_{l\in I_j^{(e)}}\left(1-\prod_{i=1}^{d}e^{-f_i(\boldsymbol{\beta},\boldsymbol{X}_l)s_i}\right)\eta'_{T_{(j)},\boldsymbol{c}}(\boldsymbol{s})\mathrm{d}\boldsymbol{s}\right\}
\end{aligned}
\tag{10}
$$

The next theorem provides the posterior distribution of the model in (7) with a general VCRM and possibly right-censored data.

Theorem 1. *Let $\mathcal{D}$ be survival data and assume a GANTR model with Condition 1. If $f_i > 0$ for at least one $i \in \{1, \dots, d\}$, the posterior distribution of $\boldsymbol{\mu}$ is the distribution of the random measure*

$$(\mu_1^\circ, \dots, \mu_d^\circ) + \sum_{j \in I^{(e)}} (M_{1,j}\delta_{T_{(j)}}, \dots, M_{d,j}\delta_{T_{(j)}})$$

where

i) $\boldsymbol{\mu}^\circ = (\mu_1^\circ, \dots, \mu_d^\circ)$ *is a d-variate CRM with Levy intensity*

$$\nu^\circ(\mathrm{d}\boldsymbol{s}, \mathrm{d}y) = \sum_{j=1}^{k+1} e^{-\langle\, \bar{\boldsymbol{h}}_j^{(e)}(\boldsymbol{\beta}, \hat{\boldsymbol{X}}) + \bar{\boldsymbol{h}}_j^{(c)}(\boldsymbol{\beta}, \hat{\boldsymbol{X}})\, , \boldsymbol{s}\, \rangle} \nu(\mathrm{d}\boldsymbol{s}, \mathrm{d}y) \mathbb{1}_{\{\mathrm{d}y \in (T_{(j-1)}, T_{(j)})\}}$$

ii) The vectors of jumps $\{(M_{1,j}, \dots, M_{d,j})\}_{j \in I^{(e)}}$ are mutually independent and the vector of jumps corresponding to the exact observation $T_{(j)}$ has density

$$g_j(\boldsymbol{s}) \propto \prod_{i=1}^{d} \left(e^{-\left(\bar{h}_{i,j+1}^{(e)}(\boldsymbol{\beta}, \hat{\boldsymbol{X}}) + \bar{h}_{i,j}^{(c)}(\boldsymbol{\beta}, \hat{\boldsymbol{X}})\right) s_i} \right) \prod_{l \in I_j^{(e)}} \left(1 - \prod_{i=1}^{d} e^{-f_i(\boldsymbol{\beta}, \boldsymbol{X}_l) s_i} \right) \eta'_{T_{(j)}}(\boldsymbol{s})$$

iii) The random measure $\boldsymbol{\mu}^\circ$ is independent of $\{(M_{1,j}, \dots, M_{d,j})\}_{j \in I^{(e)}}$.

The above characterization can be seen as a conjugacy property where, similarly to NTR distributions (see for example in Ferguson and Phadia, 1979), the posterior is updated to be GANTR model with $(\boldsymbol{\mu}^\circ)$, furthermore this can be used to calculate the posterior mean of the survival function of a new time event $Y^\star$ with associated new covariate $\boldsymbol{X}^\star$, *i.e.* $\mathbb{E}_{\boldsymbol{\mu}|\mathcal{D}}\left[S_{\boldsymbol{X}^\star}(t)\right] = \mathbb{E}_{\boldsymbol{\mu}|\mathcal{D}}\left[\mathbb{P}[Y^\star > t \,|\boldsymbol{\mu}, \boldsymbol{\beta}, \boldsymbol{X}^\star]\right]$. Such a posterior mean, where we have integrated out the underlying VCRM $\boldsymbol{\mu}$, can be used for estimation purposes; its calculations is made explicit in the next corollary.

Corollary 1. *In the setting of Theorem 1, we denote*

$$\tilde{I}^{(e)}(t) = \{l \,:\, T_{(l)} \text{ is an exact observation } \} \cap \{l \,:\, T_{(l)} \leq t\}.$$

Let $S_{\boldsymbol{X}^\star}(t) = \mathbb{P}[Y^\star > t \,|\boldsymbol{\mu}, \boldsymbol{\beta}, \boldsymbol{X}^\star]$ *be the survival function of an GANTR distributed r.v.* $Y^\star$ *associated with a covariate vector* $\boldsymbol{X}^\star$*. Then*

$$\hat{S}_{\boldsymbol{X}^\star}(t) = \mathbb{E}[S_{\boldsymbol{X}^\star}(t)|\boldsymbol{\beta}, \boldsymbol{X}^\star, \boldsymbol{\mathcal{D}}] = e^{-\sum_{j=1}^{k+1}\left(\psi^\circ_{t\wedge T_{(j)}}(\boldsymbol{V}^\star)-\psi^\circ_{t\wedge T_{(j-1)}}(\boldsymbol{V}^\star)\right)\mathbb{1}_{\{T_{(j-1)}<t\}}}$$
$$\prod_{j\in\tilde{I}^{(e)}(t)} \frac{\int_{(\mathbb{R}^+)^d} e^{-\langle \boldsymbol{V}^\star+\bar{\boldsymbol{h}}^{(e)}_{j+1}(\boldsymbol{\beta},\hat{\boldsymbol{X}})+\bar{\boldsymbol{h}}^{(c)}_{j}(\boldsymbol{\beta},\hat{\boldsymbol{X}}),\boldsymbol{s}\rangle}\prod_{l\in I^{(e)}_j}\left(1-\prod_{i=1}^{d}e^{-f_i(\boldsymbol{\beta},\boldsymbol{X}_l)s_i}\right)\eta'_{T_{(j)}}(\boldsymbol{s})\mathrm{d}\boldsymbol{s}}{\int_{(\mathbb{R}^+)^d} e^{-\langle \bar{\boldsymbol{h}}^{(e)}_{j+1}(\boldsymbol{\beta},\hat{\boldsymbol{X}})+\bar{\boldsymbol{h}}^{(c)}_{j}(\boldsymbol{\beta},\hat{\boldsymbol{X}}),\boldsymbol{s}\rangle}\prod_{l\in I^{(e)}_j}\left(1-\prod_{i=1}^{d}e^{-f_i(\boldsymbol{\beta},\boldsymbol{X}_l)s_i}\right)\eta'_{T_{(j)}}(\boldsymbol{s})\mathrm{d}\boldsymbol{s}} \tag{11}$$

where $\boldsymbol{V}^\star = \left(V_1^\star \ldots, V_d^\star\right) = (f_1(\boldsymbol{\beta}, \boldsymbol{X}^\star), \ldots, f_d(\boldsymbol{\beta}, \boldsymbol{X}^\star))$ *and* ψ° *is the Laplace exponent of* μ° *in Theorem 1.*

The following lemma gives an analytic expression for integrals of the type that appear in both the posterior mean of Corollary 1 and likelihood function of Proposition 2.

Lemma 1. *Let* ν *be a Lévy intensity associated to a d-variate VCRM,* ψ_t *its Laplace exponent,* $\boldsymbol{q}_1, \ldots, \boldsymbol{q}_m \in (\mathbb{R}^+)^d \setminus \{0\}$*,* $\boldsymbol{q} \in (\mathbb{R}^+)^d$ *and* $\varnothing \neq I = \{i_1, \ldots, i_{|I|}\} \subset \{1, \ldots, m\}$*. We define* $I \setminus 1 = \{i_2, \ldots, i_{|I|}\}$*. Then*

$$\begin{aligned}&\int_{(\mathbb{R}^+)^d\times(0,t]} e^{-\langle \boldsymbol{q},\boldsymbol{s}\rangle}\prod_{l\in I}\left(1-e^{-\langle \boldsymbol{q}_l,\boldsymbol{s}\rangle}\right)\nu(\mathrm{d}\boldsymbol{s},\mathrm{d}x)\\ &= \sum_{S\subset I\setminus 1}(-1)^{\#(S)}\left(\psi_t\left(\boldsymbol{q}_1+\sum_{l\in S}\boldsymbol{q}_l+\boldsymbol{q}\right)-\psi_t\left(\sum_{l\in S}\boldsymbol{q}_l+\boldsymbol{q}\right)\right).\end{aligned}$$

The above lemma provides a readily available estimator $\hat{S}_{\boldsymbol{X}^\star}$, in Corollary 1, if the Laplace exponent associated to the underlying VCRM can be easily computed. Further theoretical properties of the GANTR model regarding asymptotic behavior are discussed in the supplementary material. The results in this section are valid for GANTR models based on general non-homogeneous VCRMs. In the rest of the paper, however, we will work with homogeneous VCRMs consisting of CoRMs with LogNormal score distributions. Epifani and Lijoi (2010) discuss the flexibility of NTR models built using homogeneous VCRMs which is illustrated by the asymptotic properties of this sub-class of the GANTR model, see supplementary material.

5 Simulation and real data studies

In this section we analyze a simulation study and two real survival datasets with the GANTR model and the Linear Dependent Dirichlet Process (LDDP) (De Iorio et al., 2009) using the implementation in the R library "DPpackage" (Jara et al., 2011). The first real dataset illustrates the identification of subpopulations with GANTR and the second dataset illustrates the performance of the GANTR model relative to the Cox regression model. The GANTR model uses a CoRM as the underlying VCRM with a δ-LogNormal score distribution with $\sigma = 0.1$ and a homogeneous Gamma directing Lévy process with parameters α and γ, whose intensity is given in (4), and $\kappa(\mathrm{d}y) = \mathrm{d}y$. There are also analysis of a further simulation study and a real survival dataset in the Supplementary material.

We consider two hybrid inference approaches. In both approaches, we first set (α, γ). We have found that an effective approach is to use the MAP estimate $\hat{\boldsymbol{c}}^{\mathrm{MAP}} = (\hat{\alpha}^{\mathrm{MAP}}, \hat{\gamma}^{\mathrm{MAP}})$

under the corresponding NTR model using the homogeneous Gamma CRM μ without considering covariates. In such NTR setting, a priori the mean survival is given by $\hat{S}(t) = \mathbb{E}\left[e^{-\mu(0,t]}\right] = \exp\left(-\gamma \log\left(1 + \frac{1}{\alpha}\right) t\right)$. So we can assign priors on γ and α which reflect the rate of survival times in an exponential model. In particular we used a log-normal prior centered in $n/\sum_{i=1}^{n} T_i$ and variance 0.001 for γ and a log-normal prior centered at one with variance 0.1 for the bone marrow data and 0.01 for the Kidney transplant data due to the rates in the different data sets. This centers the GANTR model around the data and captures the overall rate of the survival times while allowing small values of δ to indicate departure from the NTR model that does not consider covariate effects.

The posterior distribution of δ and $\boldsymbol{\beta}$, conditional on (α, γ) or $(\hat{\alpha}, \hat{\gamma})$, can be calculated using the likelihood $l(\boldsymbol{\beta}, (\delta, \hat{\alpha}, \hat{\gamma}); \boldsymbol{\mathcal{D}})$ in (10). A closed form expression for the Laplace exponent of the δ-LogNormal CoRM is not available but a Monte Carlo estimate can be easily calculated using draws from the score distribution and the Laplace exponent of the directing gamma CRM (this can also be used for the calculation of the posterior mean survival curve in Corollary 1). The two inference approaches differ in how δ and $\boldsymbol{\beta}$ are inferred. Firstly, a MCMC scheme (see Supplementary material) can be used to draw samples from the posterior distribution of δ and $\boldsymbol{\beta}$ allowing Monte Carlo estimates of the posterior mean survival and credible intervals to be calculated. Alternatively, a *maximum a posteriori* (MAP) estimate of δ and $\boldsymbol{\beta}$ can be found using numerical optimisation of the posterior distribution. We use the LFBGS routine of the Optim pacakge in Julia (Mogensen and Riseth, 2018). Details regarding evaluation of the likelihood gradient are given in the supplementary material. This only involves one evaluation of equation (11), in contrast to the MCMC approach, but at the expense of ignoring posterior uncer-

tainty in the parameters. We will denote the MAP estimate of a generic parameter θ by $\hat{\theta}^{\text{MAP}}$.

5.1 Simulated example

We consider a competing risks example. For $i = 1, \ldots, n$, there is a covariate Z_i which normal distribution with mean 1 and variance 0.75 truncated to $(0, 2)$ and there are three possible risks: $Y_i^{(0)} \overset{\text{i.i.d.}}{\sim} \text{We}\left(2.2, \frac{1.75}{l^{1/2.2}}\right)$, $Y_i^{(1)} \overset{\text{i.i.d.}}{\sim} \text{We}\left(1.2, \frac{2.2}{(1-l)^{1/1.2}}\right)$, $Y_i^{(2)} \overset{\text{i.i.d.}}{\sim} \text{We}\left(5.3, \frac{1.5}{(1-l)^{1/5.3}}\right)$ where $\text{We}(k, \lambda)$ represents a Weibull distributions with shape parameter k and scale λ and $l \in (0, 1)$. If $Z_i \leq 1$, we observe $Y_i = \min\left\{Y_i^{(1)}, Y_i^{(0)}\right\}$ and, otherwise, $Y_i = \min\left\{Y_i^{(2)}, Y_i^{(0)}\right\}$. This defines two sub-populations which are determined by whether (or not) the covariate is above the threshold value of 1. The parameter l controls the difference between the distributions for the two groups with the distributions becoming more similar as moves from 0 to 1. We generate two data sets of 200 observations with $l = 0.1$ and $l = 0.9$.

We fit the GANTR model with regression functions $f_1(z, \beta) = \mathbb{1}_{\{z \leq \beta\}}$ and $f_2(z, \beta) = \mathbb{1}_{\{z > \beta\}}$ where β controls the threshold between two sub-populations. For simplicity we took $(\alpha, \gamma) = (1, 1)$. We use both the MAP and MCMC methods for inference. We ran the MCMC algorithm with 10000 iterations, which essentially achieved convergence.

Results of fitting the GANTR and LDDP models to data generated with $l = 0.1$ are shown in Figure 1. Both models can clearly reconstruct the true survival curves with both inference methods for the GANTR model providing estimates that are closer to the true survival curves than the LDDP models. This visual impression is supported by the L_2 distance between a point estimate and the true curves in Table 1. In fact, the fitted survival curves are very similar for the full posterior inference and MAP estimation.

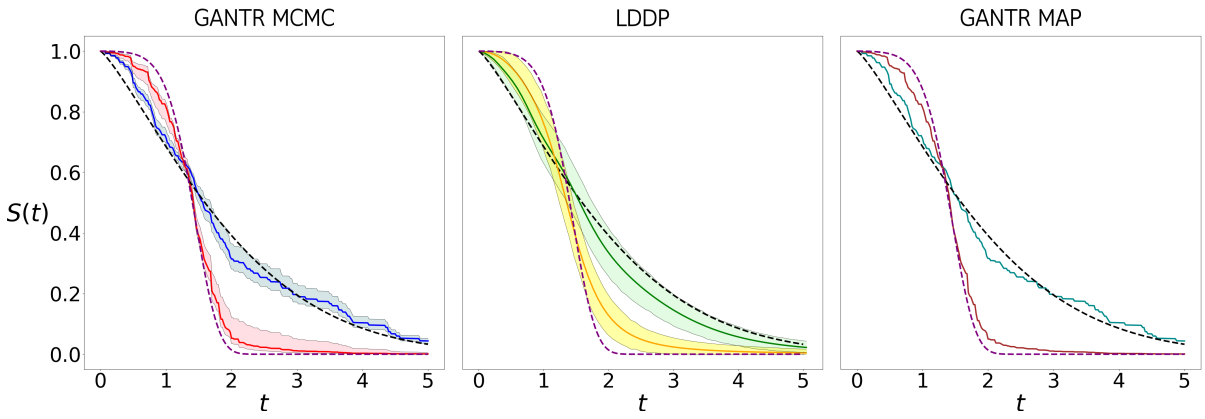

Figure 1: Competing risk simulation study ($l = 0.1$): posterior mean of the survival function (with 95% credible intervals) from the GANTR and LDDP models and GANTR MAP fits for $z = 0.66$ (true - - - - , GANTR —, LDDP —, GANTR MAP —), and $z = 1.44$ (true - - - - , GANTR —, LDDP —, GANTR MAP —).

Simulation study	$l = 0.1$		$l = 0.9$	
L_2 distance with true survival	$z = 0.66$	$z = 1.44$	$z = 0.66$	$z = 1.44$
GANTR mean survival	0.278	0.404	0.272	0.258
GANTR MAP survival	0.193	0.551	0.253	0.24
LDDP mean survival	0.33	0.619	0.38	0.824

Table 1: L_2 distance between estimated survival functions and true survival functions for $z = 0.66$ and $z = 1.44$ over evaluation meshes at every $1/60$ between 0 to 3.5 for $l = 0.1$ and at every $1/50$ between 0 and 5 for $l = 0.9$.

Other aspects of inference are also similar, the MAP and posterior mean estimates of the threshold parameter β are 1.01 and 1.02 respectively (which are very close to the true

value), and similarly, the corresponding estimates of δ are 0.18 and 0.19 respectively. The estimated value of δ is close to zero which indicates that the two groups are nearly independent. Table 1 also shows the L_2 distances between point estimates of the survival curve and its true value for data generated with $l = 0.9$. Again, both GANTR estimates have smaller L_2 distances than the LDDP model. Fitting the GANTR model using MCMC took about 2 hours for 200 observations and about 3 hours for 400 observations. On the other hand, MAP estimation took around 5 minutes for both 200 and 400 observations. The similarity of estimated survival curves and the shorter computation time leads us to only consider MAP estimation for the real data examples in the rest of this section. The supplementary material includes results from fitting the GANTR and LDDP models to simulated data sets with $l = 0.1$ with 400 observations and $l = 0.9$ with both 200 and 400 observations.

5.2 Real data studies

5.2.1 Melanoma survival data

Andersen et al. (2012) includes a study of 205 patients with melanoma who had a tumor removed by surgery. The thickness of the tumor was a covariate of interest as an increase in the tumor's thickness is thought to increase the chances of death. The data were right-censored for 72% of the patients. Again we use a two-dimensional GANTR model.

The lack of a straightforward way to stratify the patients into subpopulations according to tumor thickness motivated us to use a flexible regression model using two regression functions. There are many possible constructions for these regression functions such as univariate or multivariate splines (Denison et al., 2002) or Gaussian processes.

We choose to use a Lagrange interpolator polynomial (see Friedberg et al., 2013). Let $\boldsymbol{q} = (q_1, q_2, q_3, q_4, q_5)$ be the five number summary of the observed tumor thickness values and $L_{\boldsymbol{q},\boldsymbol{\beta}}$ be the Lagrange interpolator polynomial with knots $\{(q_i, \beta_i)\}_{i=1}^4 \cup \{(q_5, 0)\}$ where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_4) \in [0,1]^4$ are unknown parameters to define the regression functions: $f_1(z, \boldsymbol{\beta}) = \max\left\{\min\left\{1,\, L_{\boldsymbol{q},\boldsymbol{\beta}}(z)\right\},\, 0\right\}$ and $f_2(z, \boldsymbol{\beta}) = \max\left\{1 - \max\left\{L_{\boldsymbol{q},\boldsymbol{\beta}}(z),\, 0\right\},\, 0\right\}$. This leads to non-negative regression functions which are constrained so that $f_1(z, \boldsymbol{\beta}) + f_2(z, \boldsymbol{\beta}) = 1$. These functions can be interpreted as weights on two subpopulations, as in (9), which are determined by whether (or not) $f_1(z, \boldsymbol{\beta}) > f_2(z, \boldsymbol{\beta})$. The parameter δ controls the sharing between these subpopulations (*i.e.* there is very little sharing if δ is close to 0). The fixed value of the last knot, $(q_5, 0)$, constrains $f_2(z, \boldsymbol{\beta})$ to be close to one (and $f_1(z, \boldsymbol{\beta})$ to be close to zero) for large values of z and so we can interpret μ_2 as the competing risk of patients with large tumor thicknesses and μ_1 as the competing risk of patients with small tumor thicknesses.

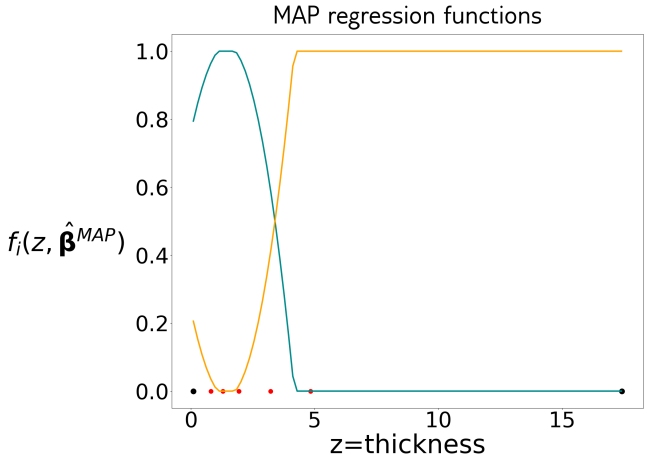

Figure 2: Melanoma study: plots of $f_1(z, \boldsymbol{\beta})$ (—) and $f_2(z, \boldsymbol{\beta})$ (—) for the MAP estimate of $\boldsymbol{\beta}$ with the minimum and maximum (•) and sixtiles (•) of the thickness values.

The MAP estimates of $\boldsymbol{\beta}$ imply that $f_1(z, \boldsymbol{\beta})$ crosses $f_2(z, \boldsymbol{\beta}) = 1$ at 3.3984 (shown in Figure 2). and the estimated $\hat{\delta}^{\text{MAP}} = 0.000972$ indicates little sharing of information. This suggests two subpopulations defined by the threshold tumor thickness of 3.3984 with substantially different survival curves in each subpopulation. This is illus-

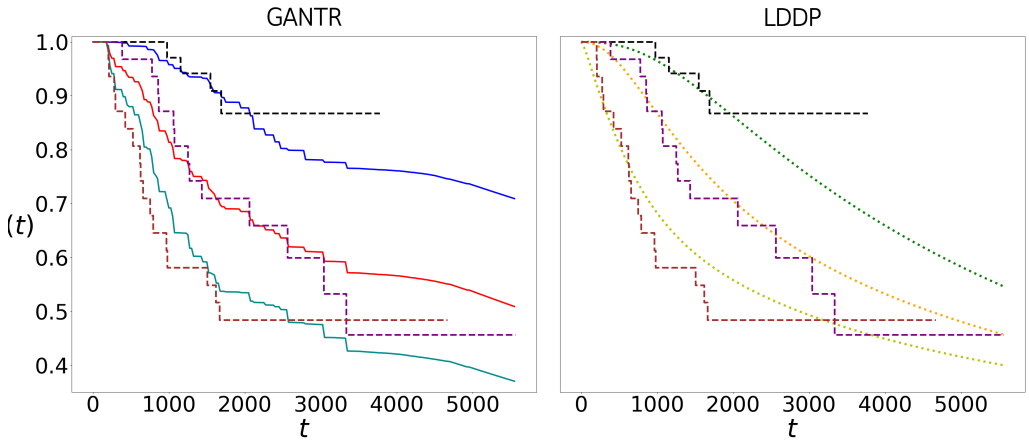

Figure 3: Melanoma study: fitted survival curves for thickness 1.5 (GANTR MAP ——, LDDP mean ·····), thickness 3.4 (GANTR MAP ——, LDDP mean ·····), and thickness 6.1 (GANTR MAP ——, LDDP mean ·····). The Kaplan-Meier fits of observations with thicknesses in the windows: $(1.255, 1.75)$ (- - - -), $(2.7, 4.1)$ (- - - -) and $(4.1, 8.1)$, (- - - -).

trated by the estimated survival curves with different tumor thicknesses in Figure 3. The estimated survival curves for each subpopulation are illustrated by the curves for thicknesses 1.5 and 6.1 which show clear differences with a better prognosis for smaller tumor thicknesses. The estimated survival curve for thickness 3.4 (close to the threshold value) shows the smoothing induced by the model between the subpopulations. The presence of the two heterogeneous subpopulations detected by the GANTR model are supported by the plotted Kaplan-Meier curves and the LDDP mean fits for the survival

curves, which are shown for comparison.

5.2.2 Kidney transplant data

We consider the Kidney transplant dataset from the survival analysis book of Klein and Moeschberger (2006) which is available in the R package "KMsurv" by Yan (2012). This dataset consists of 863 observations of which 723 were right-censored. There are two binary covariates: sex (male or female), and race (white or black), and age is treated as a continuous covariate. The combinations of race and sex can be used to divide the patients into four groups: 1) Male-White, 2) Male-Black, 3) Female-White and 4) Female-Black. We consider a GANTR model with $d = 4$ and regression functions

$$f_{\text{mw}}(\boldsymbol{z}) = e^{\beta_{0,\text{mw}}+\beta_{1,\text{mw}}z_{\text{age}}}\mathbb{1}_{\{\text{Male-White}\}}, \qquad f_{\text{mb}}(\boldsymbol{z}) = e^{\beta_{0,\text{mb}}+\beta_{1,\text{mb}}z_{\text{age}}}\mathbb{1}_{\{\text{Male-Black}\}},$$
$$f_{\text{fw}}(\boldsymbol{z}) = e^{\beta_{0,\text{fw}}+\beta_{1,\text{fw}}z_{\text{age}}}\mathbb{1}_{\{\text{Female-White}\}}, \qquad f_{\text{fb}}(\boldsymbol{z}) = e^{\beta_{0,\text{fb}}+\beta_{1,\text{fb}}z_{\text{age}}}\mathbb{1}_{\{\text{Female-Black}\}}.$$

The intercept coefficients $\beta_{0,\text{mw}}$, $\beta_{0,\text{mb}}$, $\beta_{0,\text{fw}}$, and $\beta_{0,\text{fb}}$ account for the heterogeneity in the populations. The coefficients of the interactions between group and age $\beta_{1,\text{mw}}$, $\beta_{1,\text{mb}}$, $\beta_{1,\text{fw}}$, and $\beta_{1,\text{fb}}$ account for differences in the effect of age between the groups. The White-Male subpopulation consists of 431 observations and the White-Female of 278 observations. The two other groups are much smaller with 92 observations in the Black-Male group and 59 observations in the Black-Female group; this restricts the ages for which Kaplan-Meier estimates can be provided. In the GANTR model, the estimated value of $\hat{\delta}^{\text{MAP}} = 0.3731$ which is indicative of the borrowing of information in the model's fit. The estimated values for the regression functions' parameters are presented in Table 2. We find that the Cox regression provides good fits for many ages. However, we find some discrepancies between the Cox regression model and both nonparametric

$\hat{\beta}^{\text{MAP}}_{0,\text{mw}}$	$\hat{\beta}^{\text{MAP}}_{1,\text{mw}}$	$\hat{\beta}^{\text{MAP}}_{0,\text{fw}}$	$\hat{\beta}^{\text{MAP}}_{1,\text{fw}}$	$\hat{\beta}^{\text{MAP}}_{0,\text{mb}}$	$\hat{\beta}^{\text{MAP}}_{1,\text{mb}}$	$\hat{\beta}^{\text{MAP}}_{0,\text{fb}}$	$\hat{\beta}^{\text{MAP}}_{1,\text{fb}}$
-5.2504	0.0521	-3.8825	0.0155	-4.6746	0.0331	-2.6801	0.00002

Table 2: MAP estimators for regression functions' parameters in kidney transfer real data study.

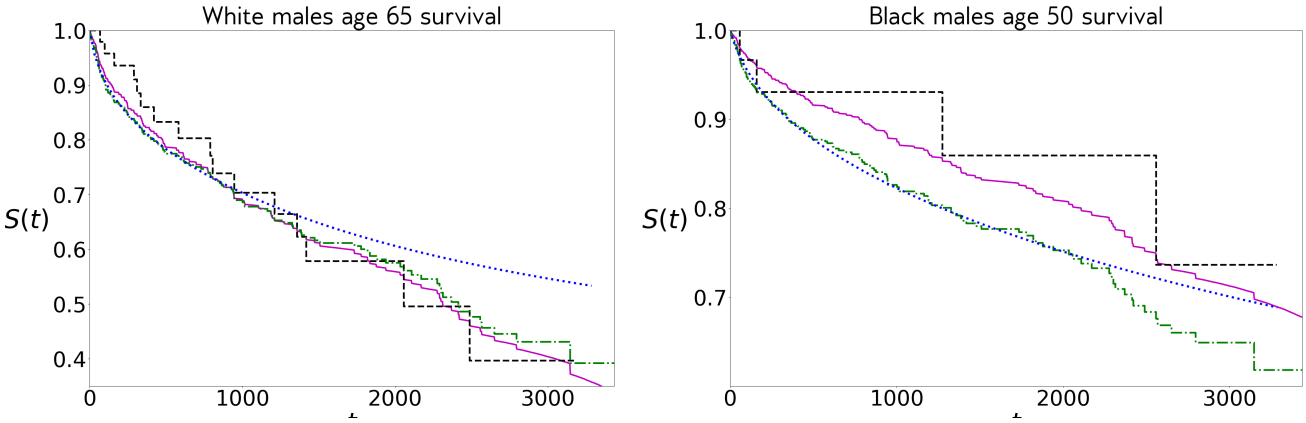

Figure 4: Kidney transplant data: Fits for white-male, ages 40 and 65, black-male and black-female, ages 50, sub-populations; GANTR MAP (——), Kaplan-Meier (- - - -), Cox regression (– – ·–) and LDDP mean (·····).

models (white, male, age 65) and between the two nonparametric models (black, male, age 50). In both cases, the GANTR fit is much closer to the Kaplan-Meier than the Cox regression model or LDDP for black, males aged 50. Further fits are presented in the supplementary material.

6 Conclusions

In a Bayesian nonparametric setting, we have introduced the GANTR model for possibly right-censored survival data. Our model generalizes the NTR models in a regression setting where non-proportional hazards are allowed. Our model can be interpreted in a competing risks framework. As a particular case of the GANTR, we can recover the multiple-sample models of Epifani and Lijoi (2010) and Riva-Palacio and Leisen (2018). The posterior characterization of the model was presented in Theorem 1 and asymptotic properties of the model are discussed in the supplementary material. We presented two approaches to draw posterior mean estimators for the survival curve, where the vector of completely random has been integrated out. The first relies on an MCMC sampler and the second in a maximum a posteriori procedure. Simulations studies provide evidence of the accuracy of our methodology and ease of interpretation. We also showed how these models can be used in in real data studies to allow for non-proportional hazard effects and crossing survival functions, and to discover subpopulations with different survival curves.

The GANTR model relies on the random weights of the underlying VCRM to allocate mass on latent competing risks. A generalized additive model is applied to the competing risks where interpretable covariate effects, e.g. Cox proportional hazards effects, are introduced in the regression functions. Thus the model in the NTR setting focuses on the random weight structure of the underlying CRM. However, time varying effects, such as accelerated failure times (e.g., Christensen & Johnson, 1988), can be considered in an interpretable manner by focusing on the location component of the underlying CRM.

Another approach that could be considered in survival analysis is the frailty model. It generalizes the proportional hazards model by introducing a multiplicative random

effect. Frailty models are often used to model clustered survival data, for example arising in multi-center clinical trials. However, such heterogeneities can be modeled directly with the GANTR model in a multiple-sample framework given by the different clinical centers. As such, although it is possible to include a frailty term in our model to introduce a mixed effect, we preferred to focus on the multiple-sample interpretation of our model. Future work will be devoted to explore this research line.

We have not considered inference for left-censored or interval-censored observations. Previous approaches to this problem in the Bayesian nonparametric setting include Doss (1994) based on mixtures of Dirichlet processes, Jara and Hanson (2011) based on a linear dependent Poisson-Dirichlet process and Kim and Lee (2003) focused on NTR distributions from the focus of cumulative hazards. We believe that our approach could be extended to using the approach of Kim and Lee (2003) but we leave this problem to future work.

The GANTR approach leads to an analytic form for the marginal likelihood (integrating over the vector of completely random measures). This is an attractive feature which allows us to calculate MAP estimates of hyperparameters. This can be seen as an empirical Bayes approach which approximates the fully Bayesian approach that we also consider. Petrone et al. (2014) provide a discussion on empirical Bayesian methods, including asymptotic results. The use of MAP estimates for hyperparameters is becoming increasingly popular in Bayesian nonparametrics, see *e.g.* Masoero et al. (2019) and Di Benedetto et al. (2017), where the number of hyperparameters is usually small and well-informed by the data. This contrasts with flexible Bayesian nonparametric regression approaches which model the logarithm of the survival time using a dependent Dirichlet process. We believe that this allows the GANTR to be applied more easily to problems

with many observations or covariates where MCMC samplers may mix slowly.

References

Andersen, P. K. and Gill, R. D. (1982). Cox's regression model for counting processes: A large sample study. The Annals of Statistics, 10, 1100–1120.

Andersen, P. K., Borgan, O., Gill, R. D., & Keiding, N. (2012). *Statistical models based on counting processes*. Springer Science & Business Media.

Buckley, J., and James, I. (1979). Linear regression with censored data. Biometrika, 66, 429–436.

Camerlenghi, F., Lijoi, A., Orbanz, P., and Prünster, I. (2019). Distribution theory for hierarchical processes. The Annals of Statistics, 47, 67–92.

Camerlenghi, F., Dunson, D. B., Lijoi, A., Prünster, I., and Rodríguez, A. (2019). Latent nested nonparametric priors (with discussion). Bayesian Analysis, 14, 1303–1356.

Cheng, S. C., Wei, L. J. and Ying, Z. (1995). Analysis of transformation models with censored data. Biometrika, 82, 835–845.

Christensen, R., and Johnson, W. (1988). Modelling accelerated failure time with a Dirichlet process. Biometrika, 75, 693–704.

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

Cox, D. R. (1975). Partial likelihood. Biometrika, 62, 269–276.

De Iorio, M., Johnson, W. O., Mueller, P., and Rosner, G. L. (2009). Bayesian nonparametric nonproportional hazards survival modeling. Biometrics, 65, 762–771.

Denison, D. G., Holmes, C. C., Mallick, B. K., and Smith, A. F. (2002). Bayesian methods for nonlinear classification and regression (Vol. 386). John Wiley & Sons.

Di Benedetto, G., Caron, F. and Teh, Y.-W. (2017). Non-exchangeable random partition models for microclustering. *arXiv:1711.07287*

Doksum, K. (1974). Tailfree and neutral random probabilities and their posterior distributions. The Annals of Probability, 2, 183–201.

Doss, H. (1994). Bayesian nonparametric estimation for incomplete data via successive substitution sampling. The Annals of Statistics, 22, 1763–1786.

Dykstra, R. L. and Laud, P. (1981). A Bayesian nonparametric approach to reliability. The Annals of Statistics, 9, 356–367.

Epifani, I. and Lijoi. A. (2010). Nonparametric priors for vectors of survival functions. Statistica Sinica, 20, 1455–1484.

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems, The Annals of Statistics, 1, 209–230.

Ferguson, T. S. (1974). Prior distributions on spaces of probability measures. The Annals of Statistics, 2, 615–629.

Ferguson, T. S., and Phadia, E. G. (1979). Bayesian nonparametric estimation based on censored data. The Annals of Statistics, 7, 163–186.

Fernández, T., Rivera, N. and Teh, Y. W. (2016). Gaussian processes for survival analysis. In *Advances in Neural Information Processing Systems 29*, eds. D. D. Lee and M. Sugiyama and U. V. Luxburg and I. Guyon and R. Garnett, 5021–5029.

Friedberg, S. H., Insel, A. J., and Spence, L. E. (2013). Linear Algebra: Pearson New International Edition. Pearson Higher Ed.

Griffin, J. and Leisen, F. (2017). Compound random measures and their use in Bayesian nonparametrics. Journal of the Royal Statistical Society, Series B, 79, 525–545.

Hjort, N. L. (1990). Nonparametric Bayes estimators based on beta processes in models for life history data. The Annals of Statistics, 18, 1259–1294.

Hjort, N. L., Holmes, C., Müller, P., & Walker, S. G. (Eds.) (2010). Bayesian nonparametrics. Cambridge University Press.

Hanson, T. E. and Jara A. (2013). Surviving fully Bayesian nonparametric regression models. In: Bayesian Theory and Applications. Oxford University Press.

Jara, A. and Hanson, T. E. (2011). A class of mixtures of dependent tail-free processes. Biometrika, 98, 553–566.

Jara, A., Hanson, T. E., Quintana, F. A., Müller, P. and Rosner, G. L. (2011). "DPpackage: Bayesian semi-and nonparametric modeling in R." Journal of Statistical Software, 40.

Kalbfleisch, J. D. and Prentice, R. L. (2011). *The Statistical Analysis of Failure Time Data*. John Wiley & Sons.

Kim, Y. and Lee, J. (2001). On posterior consistency of survival models. The Annals of Statistics, 29, 666–686.

Kim, Y. and Lee, J. (2003). Bayesian analysis of proportional hazard models. The Annals of Statistics, 31, 493–511.

Kim, Y., and Lee, J. (2004). A Bernstein-von Mises theorem in the nonparametric right-censoring model. The Annals of Statistics, 32, 1492–1512.

Kingman, J. (1967). Completely random measures. Pacific Journal of Mathematics, 21, 59–78.

Klein, J. P., and Moeschberger, M. L. (2006). *Survival Analysis: Techniques for Censored and Truncated Data.* Springer Science & Business Media.

Lijoi, A., Nipoti, B., and Prünster, I. (2014). Bayesian inference with dependent normalized completely random measures. Bernoulli, 20(3), 1260–1291.

MacEachern, S. N. (1999). Dependent nonparametric processes. In: *ASA Proceedings of the Section on Bayesian Statistical Science*, Alexandria, VA. American Statistical Association.

Masoero, L., Camerlenghi, F., Favaro, S. and Broderick, T. (2019). More for less: Predicting and maximizing genetic variant discovery via Bayesian nonparametrics. *arXiv:1912.05516*.

Mogensen, P. K., and Riseth, A. N. (2018). Optim: A mathematical optimization package for Julia. Journal of Open Source Software, 3.

Muliere, P. and Walker, S. G. (1997). Beta-Stacy processes and a generalization of the Pólya-urn scheme. The Annals of Statistics, 25, 1762–1780.

Murphy, S. A., Rossini, A. J. and Van der Vaart, A. W. (1997). Maximum likelihood estimation in the proportional odds model. Journal of the American Statistical Association, 92, 968–976.

Nieto-Barajas, L. E. (2014). Bayesian semiparametric analysis of short-and long-term hazard ratios with covariates. Computational Statistics & Data Analysis, 71, 477–490.

Nipoti, B., Jara, A., and Guindani, M. (2018). A Bayesian semiparametric partially PH model for clustered time-to-event data. Scandinavian Journal of Statistics, 45, 1016–1035.

Petrone, S., Rizzelli, S., Rousseau, J., and Scricciolo, C. (2014). Empirical Bayes methods in classical and Bayesian inference. Metron, 72, 201–215.

Phadia, E. G. (2015). *Prior Processes and Their Applications.* New York, NY: Springer.

Prentice, R. L., Kalbrleisch, J. D., Peterson, Jr, A. V., Flournoy, N., Farewell, V. T. and Breslow, N. E. (1978). The analysis of failure times in the presence of competing risks. Biometrics, 34, 541–554.

Riva-Palacio, A. and Leisen, F. (2018). Bayesian nonparametric estimation of survival functions with multiple-samples information. Electronic Journal of Statistics, 12, 1330–1357.

Riva-Palacio, A., and Leisen, F. (2019). Compound vectors of subordinators and their associated positive Lévy copulas. arXiv preprint arXiv:1909.12112.

Riva-Palacio, A., Leisen, F. and Griffin, J. Supplement to "Survival regression models with dependent Bayesian nonparametric priors".

Ryan, T. and Woodall, W. (2005). The most cited statistical papers. Journal of Applied Statistics, 32, 461–474.

Therneau, T. M. and Grambsch, P. M. (2000). Modeling survival data: Extending the Cox model. Springer-Verlag Inc.

Original by Klein, J. P. and Moeschberger, M. L. and modifications by Yan, J. (2012). KMsurv: Data sets from Klein and Moeschberger (1997), Survival Analysis. R package version 0.1-5. https://CRAN.R-project.org/package=KMsurv

Yang, S. and Prentice, R. L. (1999). Semiparametric inference in the proportional odds regression model. Journal of the American Statistical Association, 94, 125–136.