Bayesian Nonparametric Erlang Mixture Modeling for Survival Analysis

Yunzhe Li, Juhee Lee, and Athanasios Kottas*
Department of Statistics, University of California, Santa Cruz

November 15, 2022

Abstract

We develop a flexible Erlang mixture model for survival analysis. The model for the survival density is built from a structured mixture of Erlang densities, mixing on the integer shape parameter with a common scale parameter. The mixture weights are constructed through increments of a distribution function on the positive real line, which is assigned a Dirichlet process prior. The model has a relatively simple structure, balancing flexibility with efficient posterior computation. Moreover, it implies a mixture representation for the hazard function that involves time-dependent mixture weights, thus offering a general approach to hazard estimation. We extend the model to handle survival responses corresponding to multiple experimental groups, using a dependent Dirichlet process prior for the group-specific distributions that define the mixture weights. Model properties, prior specification, and posterior simulation are discussed, and the methodology is illustrated with synthetic and real data examples.

Keywords: Bayesian nonparametrics; Dependent Dirichlet process; Dirichlet process; Erlang distribution; Hazard function; Markov chain Monte Carlo; Survival analysis.

*Yunzhe Li (yli566@ucsc.edu) is Ph.D. student, Juhee Lee (juheelee@soe.ucsc.edu) is Associate Professor, and Athanasios Kottas (thanos@soe.ucsc.edu) is Professor, Department of Statistics, University of California, Santa Cruz. This research was supported in part by the National Science Foundation under award DMS 2015428.

1 Introduction

The Erlang mixture model is defined as a weighted combination of gamma densities, $\sum_{m=1}^{M} \omega_m \, \text{Ga}(t \mid m, \theta)$, with each gamma density $\text{Ga}(t \mid m, \theta)$ indexed by its integer shape parameter, m, and with all densities sharing scale parameter θ. Hence, in contrast to traditional mixture models, Erlang mixtures comprise identifiable mixture components and a parsimonious model formulation built from kernels that involve a single parameter that needs to be estimated. Indeed, it is more natural to view the model as a basis representation for densities on $\mathbb{R}^+$, where the $\text{Ga}(t \mid m, \theta)$ play the role of the basis densities and the ω_m provide the corresponding weights. The key result for Erlang mixtures stems from the construction of the weights as increments of a distribution function G on $\mathbb{R}^+$, in particular, $\omega_m = G(m\theta) - G((m-1)\theta)$ (with the last weight adjusted such that the ω_m form a probability vector). Then, as $M \to \infty$ and $\theta \to 0$, the Erlang mixture density converges pointwise to the density of G (e.g., Butzer 1954, Lee & Lin 2010).

The Erlang mixture structure, in conjunction with the theoretical support from the convergence result, provide an appealing setting for nonparametric Bayesian modeling and inference. The key ingredient for such modeling is a nonparametric prior for distribution G, which, along with priors for θ and M, yields the full Bayesian model. Regarding relevant existing approaches, we are only aware of Xiao et al. (2021) where the Erlang mixture is used as a prior model for inter-arrival densities of homogeneous renewal processes. Also related is the prior model for Poisson process intensities in Kim & Kottas (2022), although for that model the weights are defined as increments of a cumulative intensity function. Finally, we note that the Erlang mixture model was used for density estimation in Venturini et al. (2008), albeit with fixed M and with a Dirichlet prior distribution for the vector of weights, i.e., without exploiting the construction through distribution G.

To our knowledge, Erlang mixtures have not been explored as a general methodological tool for nonparametric Bayesian survival analysis, and this is our motivation for the work in this article. The nonparametric Bayesian model is built from a Dirichlet process (DP) prior (Ferguson 1973) for distribution G, which defines the mixture weights, and from parametric priors for θ and M, which control the effective support and smoothness in the shape of the Erlang mixture density. The modeling approach is sufficiently flexible to handle non-standard shapes for important functionals of the survival distribution, including the survival function and the hazard function. We discuss prior specification for the model hyperparameters, and design an efficient posterior simulation method that draws from well-established techniques for DP mixture models. The model is extended to incorporate survival responses from multiple experimental groups, using a dependent Dirichlet process prior (MacEachern 2000, Quintana et al. 2022) for the group-specific distributions that define the mixture weights. The model extension retains the flexibility in the group-specific survival densities, and it also allows for general relationships between groups that bypass restrictive assumptions, such as proportional hazards.

Survival analysis is among the earliest application areas of Bayesian nonparametrics. The literature includes modeling and inference methods based on priors on the space of survival functions, survival densities, cumulative hazard functions, or hazard functions. Reviews can be found, for instance, in Ibrahim et al. (2001), Phadia (2013), Müller et al. (2015), and Mitra & Müller (2015). The part of this literature that is more closely related to our proposed methodology involves DP mixture models for the survival density. Such mixture models have been developed using kernels that include the Weibull distribution (e.g., Kottas 2006), log-normal distribution (e.g., De Iorio et al. 2009), and gamma distribution (e.g., Hanson 2006, Poynor & Kottas 2019). The convergence property for Erlang

mixtures is the only mathematical result we are aware of that supports the choice of a particular parametric kernel in mixture modeling for densities on $\mathbb{R}^+$.

Our main objective is to add a new practical tool to the collection of nonparametric Bayesian survival analysis methods. The DP-based Erlang mixture model may be attractive for its modeling perspective that involves a basis densities representation, its parsimonious mixture structure, and efficient posterior simulation algorithms (comparable to the ones for standard DP mixtures).

The rest of the article is organized as follows. Section 2 introduces the methodology, including approaches to prior specification and posterior simulation (with details for the latter given in the Appendixes). Sections 3 and Section 4 present results from synthetic and real data examples, respectively. Finally, Section 5 concludes with a summary.

2 Methodology

2.1 The modeling approach

Erlang Mixture Model. We propose a structured mixture model of Erlang densities for the density, $f(t)$, of the survival distribution, aiming at more general inference for survival functionals than what specific parametric distributions can provide. Specifically, let

$$f(t) \equiv f(t \mid M, \theta, \boldsymbol{\omega}) = \sum_{m=1}^{M} \omega_m \, \mathrm{Ga}(t \mid m, \theta), \quad t \in \mathbb{R}^+, \tag{1}$$

where $\boldsymbol{\omega} = \{\omega_m : m = 1, \ldots, M\}$ denotes the vector of mixture weights, and $\mathrm{Ga}(\cdot \mid m, \theta)$ the density of the Erlang distribution, that is, the gamma distribution with integer shape parameter m and scale parameter θ, such that the mean is $m\theta$ and the variance $m\theta^2$. Given the number of the Erlang mixture components, M, the kernel densities in (1) are fully specified up to the common scale parameter θ. Hence, compared with standard mixture

models, for which the number of unknown parameters increases with M, the model in (1) offers a parsimonious mixture representation.

A key component of the model specification revolves around the mixture weights. These are defined through increments of a distribution function G with support on $\mathbb{R}^+$, such that $\omega_m = G(m\theta) - G((m-1)\theta)$, for $m = 1, \ldots, M-1$, and $\omega_M = 1 - G((M-1)\theta)$. This formulation for the mixture weights provides appealing theoretical results for the Erlang mixture model in (1). In particular, as $M \to \infty$ and $\theta \to 0$, $f(t \mid M, \theta, \boldsymbol{\omega})$ converges pointwise to the density function of G. The convergence property for the density can be derived from more general probabilistic results (e.g., Butzer 1954); a proof of the convergence of the distribution function of $f(t \mid M, \theta, \boldsymbol{\omega})$ to G can be found in Lee & Lin (2010). This result highlights that using a prior with wide support for G is crucial to achieve the generality of the model in (1) required to capture non-standard shapes of a survival distribution. We provide details below on the nonparametric prior for G, as well as on the priors for parameters θ and M.

The model in (1) also offers a flexible, albeit parsimonious mixture representation for the survival function, $S(t \mid M, \theta, G)$, and the hazard function, $h(t \mid M, \theta, G)$. Note that, having defined the mixture weights $\boldsymbol{\omega}$ through distribution G, we use the latter in the notation for model parameters. Denote by $S_{\text{Ga}}(\cdot \mid m, \theta)$ and $h_{\text{Ga}}(\cdot \mid m, \theta)$ the survival and hazard function, respectively, of the Erlang distribution with parameters m and θ. Then, the survival function associated with the model in (1) is given by

$$S(t \mid M, \theta, G) \;=\; \sum_{m=1}^{M} \omega_m \, S_{\text{Ga}}(t \mid m, \theta), \tag{2}$$

that is, it has the same weighted combination representation as the density, replacing the Erlang basis densities by the corresponding survival functions. Moreover, the hazard

function under the Erlang mixture model can be expressed as

$$h(t \mid M, \theta, G) \;=\; \sum_{m=1}^{M} \omega_m^\star(t)\, h_{\mathrm{Ga}}(t \mid m, \theta), \tag{3}$$

where $\omega_m^\star(t) = \omega_m S_{\mathrm{Ga}}(t \mid m, \theta)/\{\sum_{m'=1}^{M} \omega_{m'}\, S_{\mathrm{Ga}}(t \mid m', \theta)\}$. The hazard function is a weighted combination of the hazard functions associated with the Erlang basis densities, and, importantly, the mixture weights in (3) vary with t. Such time-dependent weights allow for local adjustment, and thus $h(t \mid M, \theta, G)$ can achieve general shapes, despite the fact that the basis hazard functions, $h_{\mathrm{Ga}}(t \mid m, \theta)$, are non-decreasing in t (constant for $m = 1$, and increasing for $m \geq 2$).

Dirichlet Process Prior for G. As previously discussed, the key model component is distribution G as it defines the mixture weights ω_m through discretization of its distribution function on intervals $B_m = ((m-1)\theta, m\theta]$, for $m = 1, \ldots, M-1$, and $B_M = ((M-1)\theta, \infty)$. We place a DP prior on G, i.e., $G \mid \alpha, G_0 \sim \mathrm{DP}(\alpha, G_0)$, where $\alpha > 0$ is the total mass parameter and G_0 the centering distribution (Ferguson 1973). We work with an exponential distribution, $\mathrm{Exp}(\zeta)$, for G_0, with random mean ζ assigned an inverse-gamma hyperprior, $\zeta \sim \text{inv-Ga}(a_\zeta, b_\zeta)$. We further assume a gamma hyperprior for the total mass parameter, $\alpha \sim \mathrm{Ga}(a_\alpha, b_\alpha)$. Given M, the DP prior for G implies a Dirichlet prior distribution for the vector of mixture weights, $\boldsymbol{\omega} \mid M, \alpha, \zeta \sim \mathrm{Dir}(\alpha G_0(B_1), \ldots, \alpha G_0(B_M))$.

The nonparametric prior for G is of primary importance. The DP prior allows the corresponding distribution function realizations to admit general shapes that can concentrate probability mass on different time intervals B_m, thus favoring different Erlang basis densities through the associated ω_m. The key parameter in this respect is α, as it controls the extent of discreteness for realizations of G and the variability of such realizations around G_0. As an illustration, Figure 1 plots prior realizations for the mixture weights and the

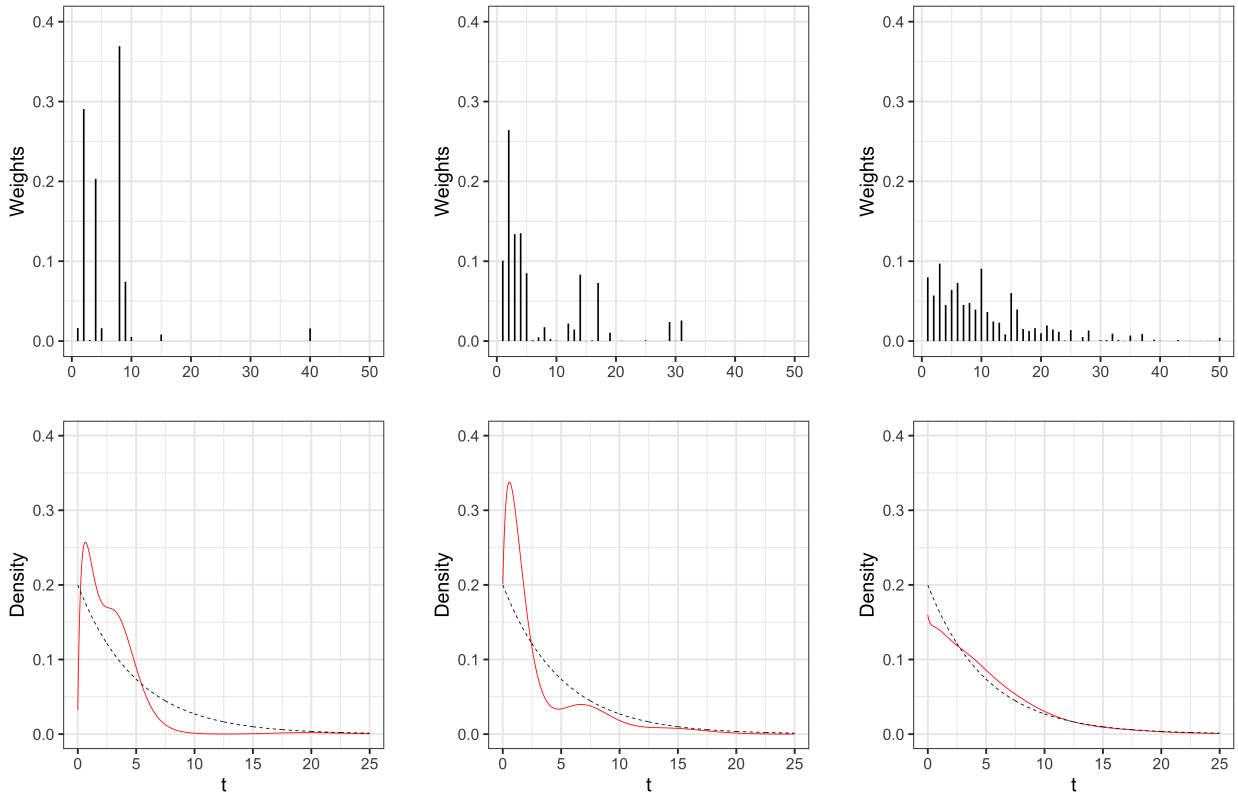
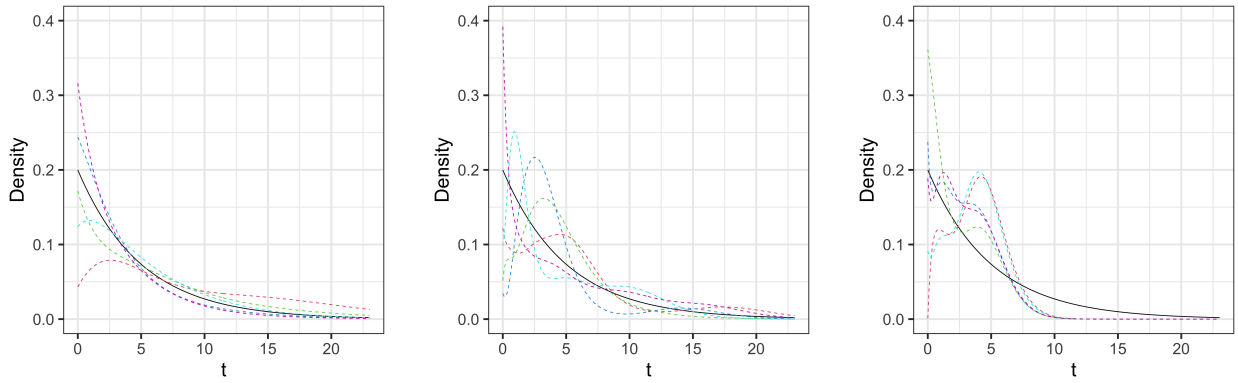


Figure 1: Prior realizations of the mixture weights $\boldsymbol{\omega}$ (top row) and the corresponding densities $f(t \mid M, \theta, G)$ given by the red solid lines (bottom row), under $\alpha = 1, 10, 100$ (left, middle, right columns). In all cases, $M = 50$, $\theta = 0.5$, and $G_0 = \text{Exp}(5)$. The black dotted line in the bottom row panels is the density of G_0.

corresponding Erlang mixture density under three values of α ($\alpha = 1, 10$ or 100), using in all cases $M = 50$, $\theta = 0.5$, and an Exp(5) distribution for G_0. The smaller α gets, the smaller the number of effective mixture weights becomes. Also, for larger α the Erlang mixture density becomes similar to the density of G_0, which is to be expected from the pointwise convergence result and the fact that larger α values imply smaller variability of G around G_0.

Priors for θ and M. Under the model construction for the mixture weights, θ controls the step size of the increments and thus how fine the discretization of G is. Moreover, θ



(a) $(M, \theta) = (10, 2)$ (b) $(M, \theta) = (40, 0.5)$ (c) $(M, \theta) = (10, 0.5)$

Figure 2: Prior realizations for $f(t \mid M, \theta, G)$, under $(M, \theta) = (10, 2)$, $(40, 0.5)$, and $(10, 0.5)$. In al cases, $\alpha = 10$ and $G_0 = \text{Exp}(5)$. The black dashed line denotes the density of G_0.

controls the location and dispersion of the Erlang basis densities in (1). With smaller θ, the Erlang densities are more concentrated around their mean $m\theta$, and the discretization of G becomes finer. Hence, and as the pointwise convergence result suggests, smaller θ values may be needed to accommodate non-standard density shapes. Also, the last component in (1) has mean $M\theta$ (with variance $M\theta^2$), and thus the effective support of $f(t \mid M, \theta, G)$ is jointly determined by M and θ; with smaller θ, a greater value of M is needed to achieve the same effective support. To illustrate, Figure 2 plots five prior realizations of the Erlang mixture density for each of three combinations of (M, θ), using in all cases $\alpha = 10$, and an Exp(5) distribution for G_0. For panels (a) and (b), $M\theta = 20$. The resulting density realizations have similar effective support, although the ones in panel (b) involve more variable shapes, as expected since the value of θ is smaller than that in panel (a). For panel (c), $M\theta = 5$, resulting in noticeably smaller effective support for the realized densities relative to panels (a) and (b).

We work with a joint prior for θ and M, $p(\theta, M) = p(\theta)p(M \mid \theta)$. We assume $\theta \sim \text{Ga}(a_\theta, b_\theta)$, and conditional on θ, assign to M a discrete uniform distribution, $M \mid \theta \sim$

Unif($\lceil M_1/\theta \rceil, \ldots, \lceil M_2/\theta \rceil$), where $\lceil a \rceil$ is the smallest integer that is larger or equal to a. To specify the hyperparameters a_θ, b_θ, M_1 and M_2, we use a relatively conservative approach, based on the range of the data. For M_1, we choose a value greater than the largest value in the data, and set $M_2 = cM_1$ for a relatively small integer c. The motivation for this choice is to ensure that the effective support of the Erlang mixture model is sufficiently large for the particular data application. To specify the prior hyperparameters for θ, we notice that $M_1/\theta \sim$ inv-Ga$(a_\theta, M_1/b_\theta)$, based on which we recommend selecting values for a_θ and b_θ such that E(M_1/θ) is between 10 and 50.

Posterior Simulation. The data point for the i^{th} subject is recorded as $y_i = \min(t_i, c_i)$, where t_i is the survival time and c_i the (independent) administrative censoring time, for $i = 1, \ldots, n$. The data set can be represented through $\mathcal{D} = \{(y_i, \nu_i) : i = 1, \ldots, n\}$, where the ν_i are binary censoring indicators such that $\nu_i = 1$ if t_i is observed, and $\nu_i = 0$ otherwise. Then, the likelihood function can be written as

$$L(M, \theta, G; \mathcal{D}) = \prod_{i=1}^{n} \left\{ f(y_i \mid M, \theta, G) \right\}^{\nu_i} \left\{ S(y_i \mid M, \theta, G) \right\}^{1-\nu_i}, \tag{4}$$

where $f(\cdot \mid M, \theta, G) \equiv f(\cdot \mid M, \theta, \boldsymbol{\omega})$ and $S(\cdot \mid M, \theta, G)$ are given in (1) and (2), respectively.

We implement posterior inference via Markov chain Monte Carlo (MCMC) simulation, using standard posterior simulation methods for DP mixture models (e.g., Escobar & West 1995, Neal 2000). The Erlang mixture density in (1) can be expressed as a DP mixture by exploiting the definition of the weights ω_m through distribution G, resulting in the following alternative mixture representation:

$$f(t \mid M, \theta, G) \;=\; \sum_{m=1}^{M} \omega_m \, \text{Ga}(t \mid m, \theta) \;=\; \int_0^\infty \left\{ \sum_{m=1}^{M} \mathbb{1}_{B_m}(\phi) \, \text{Ga}(t \mid m, \theta) \right\} \text{d}G(\phi).$$

Here, $\mathbb{1}_B(\cdot)$ is the indicator function for set B, and, as before, $B_m = ((m-1)\theta, m\theta]$, for $m = 1, \ldots, M-1$, and $B_M = ((M-1)\theta, \infty)$.

For posterior simulation, we augment the likelihood in (4) with subject-specific latent variables, $\phi_i \mid G \overset{i.i.d.}{\sim} G$, which indicate the mixture component for the associated observations. In particular, if ϕ_i falls into interval B_m, the i^{th} observation corresponds to the m^{th} Erlang basis density. The posterior distribution involves G, M, θ, the set of latent variables $\boldsymbol{\phi} = \{\phi_i : i = 1, ..., n\}$, and the DP hyperparameters (α, ζ). We marginalize G over its DP prior and work with the prior full conditionals for the ϕ_i, implied by the DP Pólya urn representation (Blackwell & MacQueen 1973), to sample from the marginal posterior distribution for all model parameters except G. To this end, we employ the MCMC method in Escobar & West (1995); the details are given in Appendix A.

Although we do not sample the mixture weights $\boldsymbol{\omega}$ during the MCMC simulation, it is straightforward to obtain posterior samples for $\boldsymbol{\omega}$, using their definition in terms of distribution G. The conditional posterior distribution for G, given (α, ζ) and $\boldsymbol{\phi}$, is characterized by a DP with updated total mass parameter $\alpha^\star = \alpha + n$, and centering distribution $G_0^\star = \alpha(\alpha+n)^{-1}\text{Exp}(\zeta)+(\alpha+n)^{-1}\sum_{i=1}^{n}\delta_{\phi_i}$. Hence, using the DP definition, the conditional posterior distribution for $\boldsymbol{\omega}$, given M, (α, ζ), and $\boldsymbol{\phi}$, is a Dirichlet distribution with parameter vector $(\alpha^\star G_0^\star(B_1), \ldots, \alpha^\star G_0^\star(B_M))$.

Two points about the posterior simulation method are worth making. First, note that the model parameters do not explicitly contain the vector of mixture weights. The mixture weights are estimated through the posterior distribution of G, which plays the role of the relevant parameter. This is practically important in that the dimension of the parameter space does not change with M, and we thus do not need to resort to more complex trans-dimensional MCMC algorithms. Second, the DP-based Erlang mixture model offers an interesting example where full posterior inference can be obtained from a DP mixture model without the need to truncate or approximate the DP prior. This is a result of the

use of a marginal MCMC method, as well as of the fact that distribution G enters the model only through increments of its distribution function, which define the mixture weights.

2.2 Model extension for control-treatment studies

A practically relevant scenario in studies where survival responses are collected involves data from multiple experimental groups, typically associated with different treatments. Evidently, it is of interest in these settings to compare survival distributions across different groups. We develop an extension of the Erlang mixture model in this direction, focusing on the case of two groups for, say, a generic control-treatment study. Our objective is to retain the flexible modeling approach for the survival distributions, avoiding restrictions to specific parametric shapes or rigid relationships, such as proportional hazards. We also seek a prior probability model that allows for dependence, and thus borrowing of information, between the two distributions.

We use the dependent DP (DDP) prior structure (MacEachern 2000) that extends the DP prior for distribution G to a prior model for a collection of covariate-dependent distributions, G_x, where x indexes the distributions in terms of values in the covariate space. Our context involves a binary covariate $x \in \mathcal{X} = \{C, T\}$, where C and T represent control and treatment groups, respectively. The DDP prior builds from the DP stick-breaking representation (Sethuraman 1994) by utilizing covariate-dependent weights and/or atoms. We work with a common-weights DDP prior model:

$$G_x = \sum_{\ell=1}^{\infty} p_\ell \, \delta_{\varphi^*_{x\ell}}, \qquad \text{for } x \in \mathcal{X}, \tag{5}$$

with $p_1 = v_1$, $p_\ell = v_\ell \prod_{r=1}^{\ell-1}(1 - v_r)$, for $\ell \geq 2$, where the v_ℓ are i.i.d. from a Beta$(1, \alpha)$ distribution, and the atoms $\boldsymbol{\varphi}^\star_\ell = (\varphi^\star_{\text{C}\ell}, \varphi^\star_{\text{T}\ell})$ arise i.i.d. from a bivariate distribution G_0. Note that, under this construction, G_x follows marginally a DP(α, G_{0x}) prior, where G_{0x},

for $x \in \mathcal{X}$, are the marginals of G_0 associated with the control and treatment groups. For G_0, we consider a bivariate log-normal distribution, such that $\boldsymbol{\varphi}_\ell^\star \mid \boldsymbol{\mu} \overset{i.i.d.}{\sim} \mathrm{LN}_2(\boldsymbol{\mu}, \Sigma)$, with Σ specified. We place a bivariate normal, $\mathrm{N}_2(\bar{\boldsymbol{\mu}}, \Sigma_0)$, hyperprior on $\boldsymbol{\mu}$, with $\bar{\boldsymbol{\mu}}$ and Σ_0 fixed, and a gamma hyperprior on the total mass parameter α.

Allowing also for group-specific number of Erlang basis densities, M_x, as well as group-specific Erlang scale parameter, θ_x, the extension of the Erlang mixture model in (1) can be expressed as

$$f_x(t) \equiv f(t \mid M_x, \theta_x, G_x) = \sum_{m=1}^{M_x} \omega_{xm} \, \mathrm{Ga}(t \mid m, \theta_x), \quad t \in \mathbb{R}^+, \tag{6}$$

where $\omega_{xm} = G_x(m\theta_x) - G_x((m-1)\theta_x)$, $m = 1, \ldots, M_x - 1$, and $\omega_{xM_x} = 1 - G_x((M_x - 1)\theta_x)$. Similar to the model in (1), the group-specific Erlang basis densities are fully specified given M_x and θ_x. Furthermore, the survival functions, $S_x(t)$, and hazard functions, $h_x(t)$, under the extended model have a mixture representation similar to (2) and (3),

$$S_x(t) = \sum_{m=1}^{M_x} \omega_{xm} \, S_{\mathrm{Ga}}(t \mid m, \theta_x) \quad \text{and} \quad h_x(t) = \sum_{m=1}^{M_x} \omega_{xm}^\star(t) \, h_{\mathrm{Ga}}(t \mid m, \theta_x), \tag{7}$$

where $\omega_{xm}^\star(t) = \omega_{xm} \, S_{\mathrm{Ga}}(t \mid m, \theta_x) / \{\sum_{m'=1}^{M_x} \omega_{xm'} \, S_{\mathrm{Ga}}(t \mid m', \theta_x)\}$. Note that both the mixture components and weights are indexed by x. Again, the time-dependent weights in the hazard mixture form allow for local adjustment, and thus for flexible group-specific hazard rate shapes. Importantly, the prior model allows for general relationships between the control and treatment group hazard functions. In particular, inference is not restricted by the proportional hazards assumption, implied by several commonly used parametric or semiparametric survival regression models.

To complete the full Bayesian model, we place priors on θ_x and M_x, using again the role of these parameters (discussed in Section 2.1). More specifically, for each x, the joint prior, $p(\theta_x, M_x) = p(\theta_x)p(M_x \mid \theta_x)$. We further assume $\theta_x \overset{ind}{\sim} \mathrm{Ga}(a_{x\theta}, b_{x\theta})$, and $M_x \mid \theta_x \overset{ind.}{\sim}$

Unif$(\lceil M_{x1}/\theta_x \rceil, \ldots, \lceil M_{x2}/\theta_x \rceil)$. We use an approach similar to the one described in Section 2.1 to specify M_{x1} and M_{x2}, and the hyperparameters for θ_x.

Posterior simulation for the DDP-based Erlang mixture model proceeds with a relatively straightforward extension of the MCMC simulation method in Section 2.1. The details are provided in Appendix B.

The primary focus of this paper is on the DP-based Erlang mixture model for survival analysis and its extension for the control-treatment setting. We note however that the DDP-based Erlang mixture model can be further extended to accommodate a general p-variate covariate vector $\boldsymbol{x}$. For example, we may consider a linear-DDP structure (De Iorio et al. 2009) to extend G_x in (5) to $G_{\boldsymbol{x}} = \sum_{\ell=1}^{\infty} p_\ell \, \delta_{\psi^\star_\ell(\boldsymbol{x})}$, where $\psi^\star_\ell(\boldsymbol{x}) = \exp((1, \boldsymbol{x}')\boldsymbol{\beta}_\ell)$ with the $\boldsymbol{\beta}_\ell$ i.i.d. from a baseline distribution. The structured DDP prior for $G_{\boldsymbol{x}}$ yields covariate-dependent mixture weights, and thus a nonparametric prior model for covariate-dependent survival densities and hazard functions. A regression model may also be used for M and/or θ. Different from the linear-DDP mixture of log-normal distributions in De Iorio et al. (2009), the extended model retains the parsimonious Erlang mixture structure.

3 Simulation Study

We use three simulation scenarios to illustrate the models developed in Section 2. For the Erlang mixture model for a single distribution, we consider simulated data from: a two-component log-normal mixture to demonstrate the model's capacity to estimate non-standard density and hazard function shapes (Section 3.1); and, a log-normal distribution sampled with different levels of censoring (Section 3.2). The DDP-based extension of the model is illustrated in Section 3.3 with a synthetic data example based on a log-normal control distribution and a two-component log-normal mixture treatment distribution, specified

such that the corresponding hazard functions cross each other.

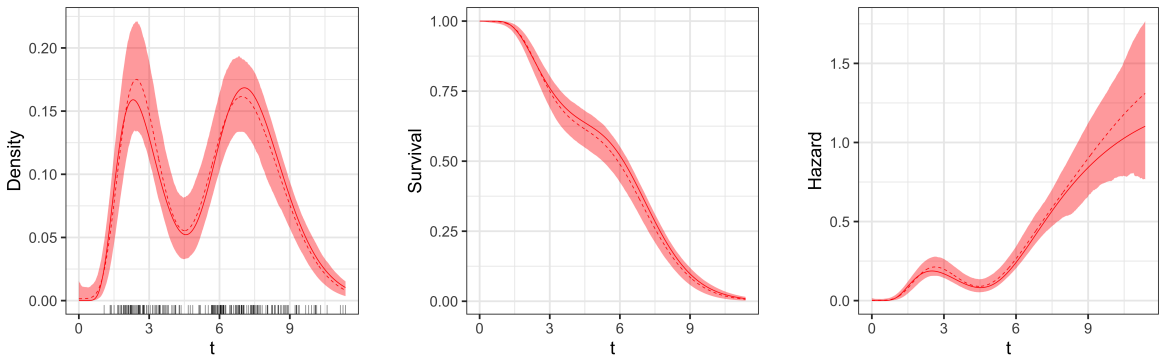
For all data examples considered here and in Section 4, we used the approach discussed in Section 2 to specify the prior hyperparameters. Consistent with inference results obtained from DP mixture models, we have observed some sensitivity to the prior choice for α. The effect on the posterior distribution for α is more noticeable for the small cell lung cancer data of Section 4.2 (involving the smallest sample size among our data examples). However, posterior inference results for survival functionals are largely unaffected even under fairly different priors for α. When the sample size is relatively small for each group, we recommend applying the DDP-based Erlang mixture model with a prior for α that supports small to moderate values, such as the $\text{Ga}(5, 1)$ prior used in Section 3.3 and 4.2.

We examined convergence and mixing of the MCMC algorithms using standard diagnostic techniques. In our experiments, we observed that parameters θ and M are highly correlated, and moderate thinning was used to improve efficiency. A general approach we take is to run the MCMC chain for 100,000 iterations, then discard the first 25% posterior samples and keep every 38th iteration for posterior inference.

3.1 Example 1: Bimodal density

We simulate $n = 200$ survival times from a mixture of two log-normal distributions, $0.4\,\text{LN}(1, 0.4) + 0.6\,\text{LN}(2, 0.2)$, which yields a bimodal density and a non-monotonic hazard function. The true underlying functions $f(t)$, $S(t)$ and $h(t)$ are plotted in Figure 3. Regarding prior specification, we used: $\alpha \sim \text{Ga}(2, 1)$; $\zeta \sim \text{inv-Ga}(3, 4)$; $\theta \sim \text{Ga}(1, 1)$; and, $M \mid \theta \sim \text{Unif}\,(\lceil M_1/\theta \rceil, \ldots, \lceil M_2/\theta \rceil)$, with $M_1 = 13$ and $M_2 = 3 \times M_1$.

Posterior inference is summarized in Figure 3. The complex features of the underlying survival functionals are captured well by the model. In particular, the inference results for

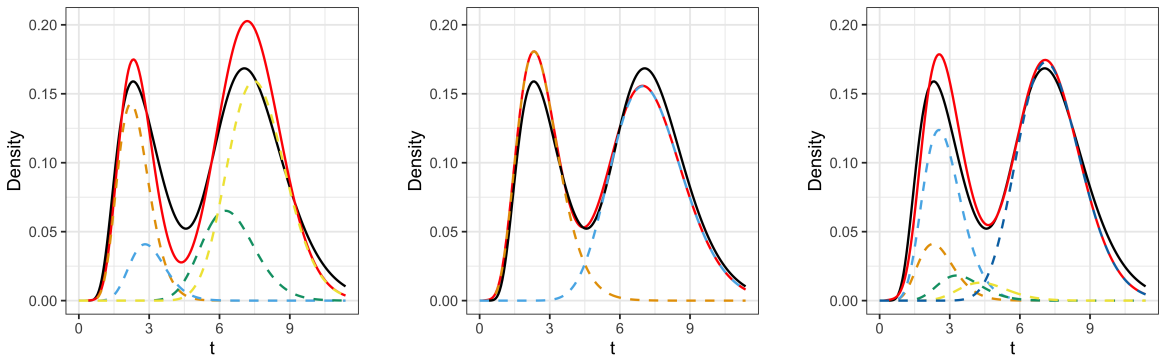


(a) Density function (b) Survival function (c) Hazard function

Figure 3: Simulation Example 1. Posterior mean (dashed lines) and 95% interval estimates (shaded regions) for the density function (left panel), survival function (middle panel) and hazard function (right panel). The red solid line in each panel corresponds to the true underlying function. The black marks on the x-axis in the left panel show the observed survival times.

the hazard function demonstrate the effectiveness of the model structure in (3) with the time-dependent weights allowing for local adjustment and estimation of a non-standard hazard function shape.

The posterior distribution for the common scale parameter θ is substantially concentrated on smaller values relative to its prior, in particular, the posterior mean and 95% credible interval estimates for θ are 0.28 and $(0.13, 0.39)$. Recalling the definition of the mixture weights, this indicates the level of partitioning needed to accommodate the non-standard, bimodal shape of the underlying density. The posterior mean and 95% credible interval estimates of the number M of mixture components are 101 and $(44, 223)$. However, the number of effective mixture components (i.e., effective basis densities) is considerably smaller than M. As an informal rule, we identify an effective Erlang basis density through its corresponding mixture weight taking value greater than a threshold of 0.01. Then, the number of effective mixture components is about 4 (on average across posterior samples).



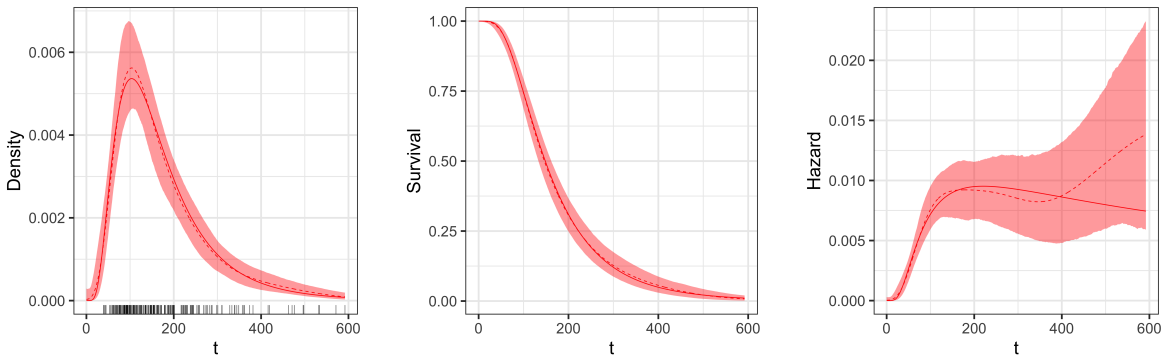
(a) $\theta = 0.20$, $M = 153$ (b) $\theta = 0.33$, $M = 78$ (c) $\theta = 0.25$, $M = 54$

Figure 4: Simulation Example 1. Plots (a)-(c) show the posterior realization of $f(t \mid M, \theta, G)$ (red solid line), based on three randomly chosen posterior samples. Each dashed line represents the Erlang basis density $\text{Ga}(t \mid m, \theta)$ for components with $\omega_m > 0.01$, multiplied by its corresponding weight. The black solid line is the true underlying density.

For a graphical illustration, Figure 4 plots three randomly selected posterior realizations of $f(t \mid M, \theta, G)$. The associated posterior draws for (θ, M) are $(0.2, 153)$, $(0.33, 78)$, and $(0.25, 54)$, whereas the number of effective Erlang basis densities is only 4, 2, and 5, respectively. The weighted effective basis densities (i.e., $\omega_m \times \text{Ga}(t \mid m, \theta)$ for m such that $\omega_m > 0.01$) are also plotted in Figure 4. This example highlights the critical importance of the nonparametric prior for distribution G that defines the weights for the Erlang mixture model.

3.2 Example 2: Unimodal density with censoring

For the second synthetic data example, we generate survival times from a log-normal distribution, $t_i \overset{i.i.d.}{\sim} \text{LN}(5, 0.6)$, $i = 1, \ldots, n$ with $n = 200$. The priors for the model parameters are: $\alpha \sim \text{Ga}(2, 1)$; $\zeta \sim \text{inv-Ga}(3, 1000)$; $\theta \sim \text{Ga}(2, 25)$; and, $M \mid \theta \sim \text{Unif}\left(\lceil 1000/\theta \rceil, \ldots, \lceil 3000/\theta \rceil\right)$.

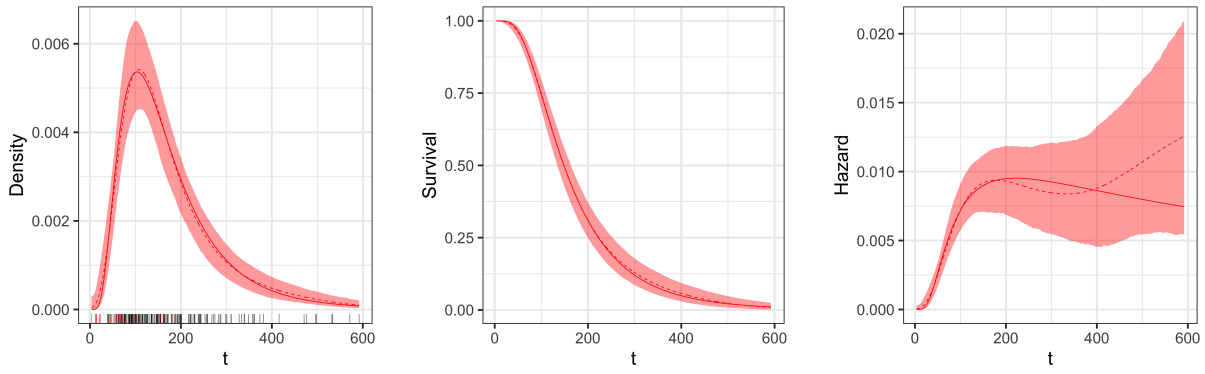


(a) Density function (b) Survival function (c) Hazard function

Figure 5: Simulation Example 2 (data without censoring). Posterior mean (dashed lines) and 95% interval estimates (shaded regions) for the density function (left panel), survival function (middle panel) and hazard function (right panel). The red solid line in each panel corresponds to the true underlying function, and the black rugs in the left panel show the survival times.

As shown in Figure 5, the model estimates well the density, survival and hazard function. The point estimate for the hazard function is less accurate beyond $t = 400$, which is to be expected given the very few observations that are greater than that time point, although the interval estimate contains the true function throughout the observation time window.

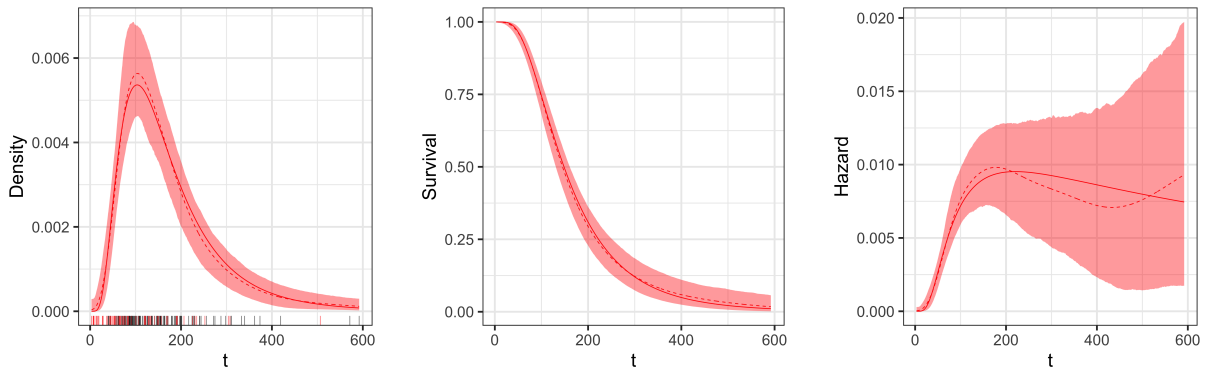
In addition, we examine the model's performance for data with censored observations. We simulate censoring times c_i from an exponential distribution with mean parameter κ, and define the observed times as $y_i = \min(t_i, c_i)$, with binary censoring indicators $\nu_i = 1(y_i \leq c_i)$. We generate the c_i under two different values of κ, resulting in two datasets with different proportions of censored observations, $g = 12\%$ and 33.5%. Figure 6 plots posterior mean and 95% interval estimates for the density, survival and hazard functionals. We note that censoring does not substantially affect the quality of the inference results, with the true function contained in all cases within the posterior interval estimates. The width of the posterior uncertainty bands increases with the larger censoring proportion,



(a) Density, $g = 12\%$

(b) Survival $g = 12\%$

(c) Hazard, $g = 12\%$



(d) Density, $g = 33.5\%$

(e) Survival, $g = 33.5\%$

(f) Hazard, $g = 33.5\%$

Figure 6: Simulation Example 2 (censored data). Posterior mean (dashed lines) and 95% interval estimates (shaded regions) for the density function (left column), survival function (middle column) and hazard function (right column). The top and bottom row corresponds to the data with censoring proportion $g = 12\%$ and 33.5%, respectively. The red solid line in each panel denotes the true underlying function. The rug plots in the left column panels indicate the data points, where the black and red marks correspond to observed and censored survival times, respectively.

with the increase more noticeable for the hazard function estimates.

3.3 Example 3: A control-treatment synthetic data set

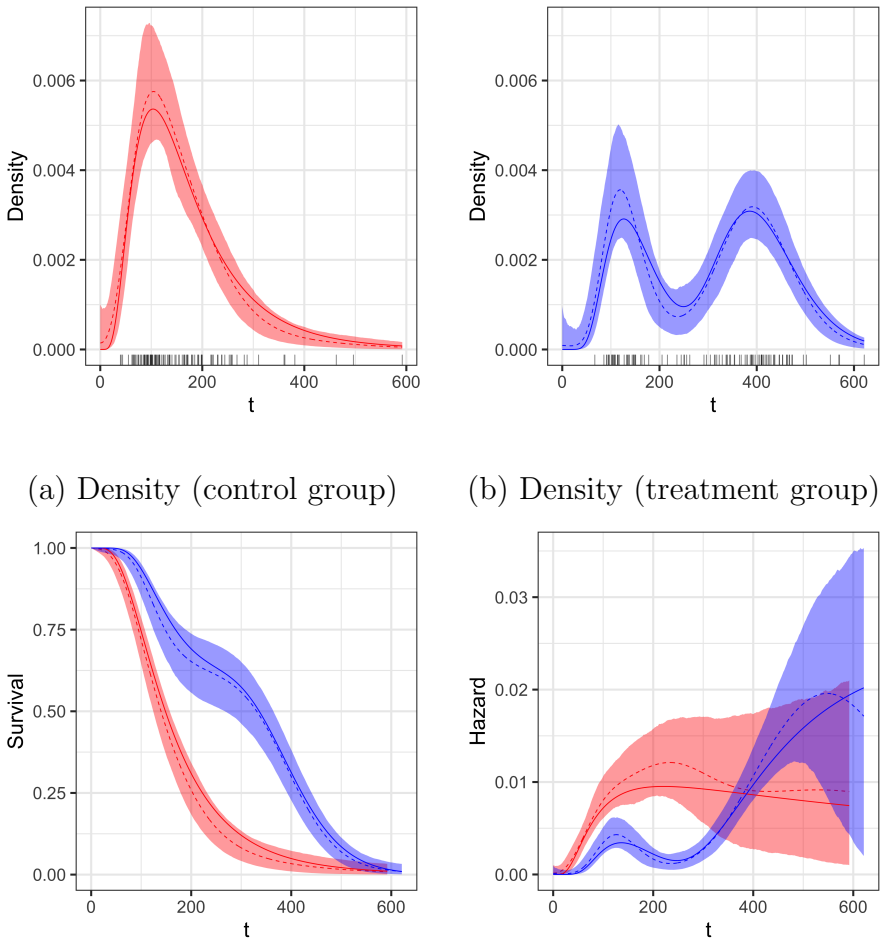
Here, we examine the performance of the DDP-based Erlang mixture model of Section 2.2. We consider a binary covariate, $x_i = C$ or T, with 100 responses in each group, such that $n = 200$. We generate $t_i \overset{i.i.d.}{\sim} \text{LN}(5, 0.6)$ for subjects with $x_i = C$, and $t_i \overset{i.i.d.}{\sim} 0.4\,\text{LN}(5, 0.4) + 0.6\,\text{LN}(6, 0.2)$ for subjects with $x_i = T$. The true density, survival and hazard functions are shown in Figure 7. The control group density is unimodal, whereas the treatment group has a bimodal density and a non-standard, non-monotonic hazard function. The truth is specified such that we have crossing hazard functions for the two groups, a scenario that traditional proportional hazards models can not accommodate.

Regarding the prior hyperparameters, we set: $\alpha \sim \text{Ga}(5, 1)$; $\boldsymbol{\mu} \sim \text{N}_2((5, 5.5)', 10\,\text{I}_2)$; $\Sigma = 3\,\text{I}_2$; $\theta_x \overset{ind.}{\sim} \text{Ga}(2, 50)$; and, $M_x \mid \theta_x \overset{ind.}{\sim} \text{Unif}\,(\lceil 1000/\theta_x \rceil, \ldots, \lceil 4000/\theta_x \rceil)$. As shown in Figure 7, the model captures effectively the shape of the survival functionals, despite the fact that the functions vary greatly across the two groups, and it successfully recovers the non-proportional hazards relationship between the groups. Again, with respect to hazard estimation, the point estimates are generally less accurate and the interval bands are wider for larger time points where data is scarce.

4 Data Examples

4.1 Liver metastases data

We consider data on survival times (in months) from 622 patients with liver metastases from a colorectal primary tumor without other distant metastases, available from the R



(a) Density (control group)

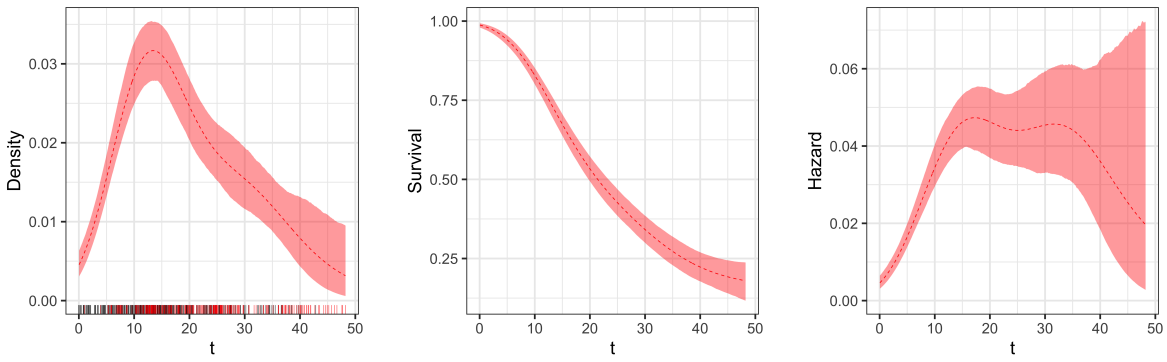
(b) Density (treatment group)

(c) Survival function

(d) Hazard function

Figure 7: Simulation Example 3. Panels (a) and (b) plot the estimates for the control and treatment group density, respectively (the rug plots show the corresponding survival times). Panels (c) and (d) compare the estimates for the survival and hazard function, respectively. In each panel, the dashed lines denote the posterior mean estimates, the solid line the true underlying function, and the shaded regions indicate the 95% credible intervals. Red and blue color is used for the control and treatment group, respectively.

package "locfit". The censoring proportion is high, with 259 censored responses. The data set has been used in earlier work to illustrate classical and Bayesian nonparametric methods for density and hazard estimation; see, e.g., Antoniadis et al. (1999) and Kottas (2006).



(a) Density function (b) Survival function (c) Hazard function

Figure 8: Liver metastases data. Panels (a), (b) and (c) plot posterior mean (dashed lines) and 95% interval estimates (shaded regions) for the density, survival and hazard function, respectively. The rug plot in panel (a) shows observed (black) and censored (red) survival times.

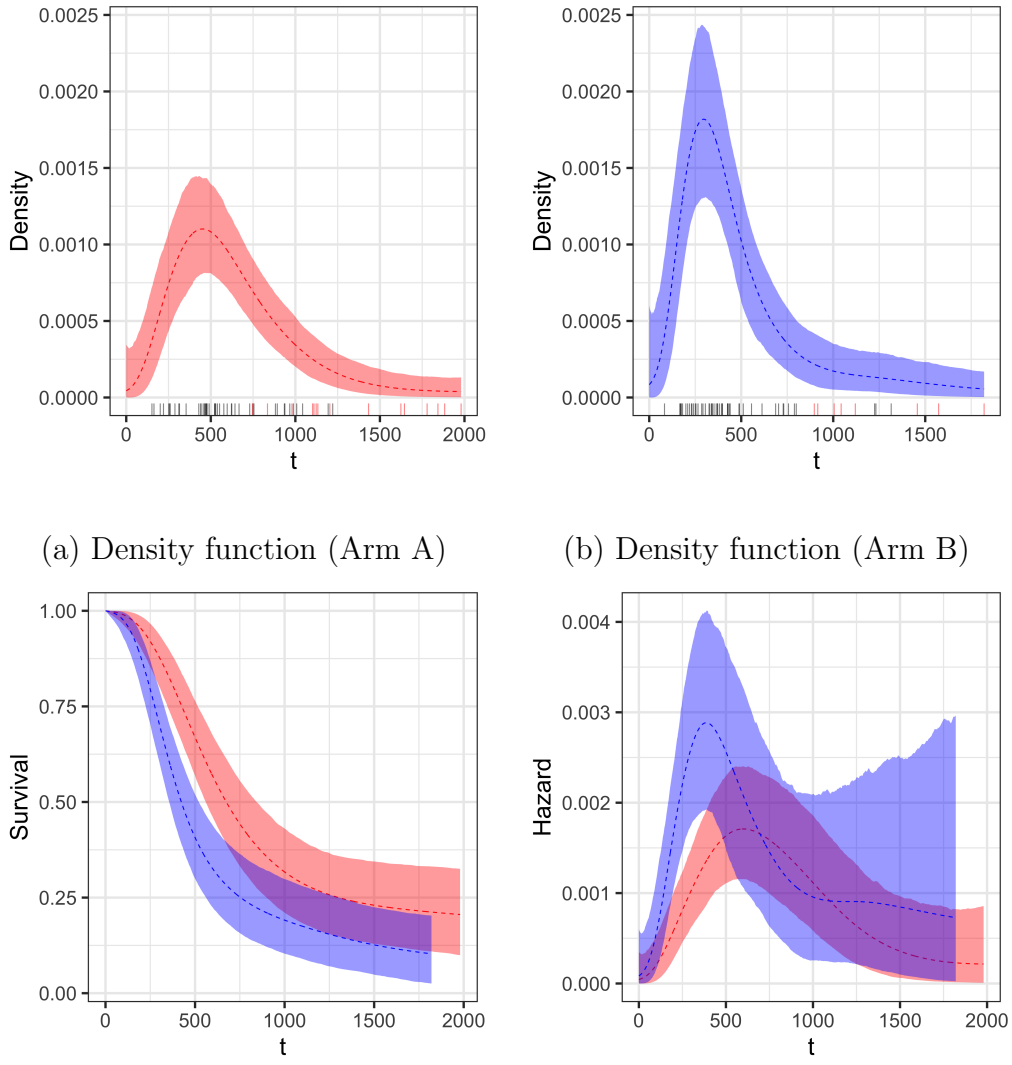
To apply the DP-based Erlang mixture model, we set the priors as follows: $\alpha \sim \text{Ga}(5, 1)$; $\zeta \sim \text{inv-Ga}(3, 80)$; $\theta \sim \text{Ga}(2, 2)$; and, $M \mid \theta \sim \text{Unif}(\lceil 100/\theta \rceil, \ldots, \lceil 300/\theta \rceil)$. Inference results for the density, survival, and hazard function are reported in Figure 8. The model estimates a unimodal survival density (with mode at about 13 months), with a non-standard, skewed right tail. The hazard rate estimate increases up to about 17 months, stays roughly constant between 17 to 35 months, and then decreases. The width of the posterior uncertainty bands for the hazard function increases considerably beyond 40 months, which is consistent with the fact that there are very few responses beyond that time point, and almost all of them are censored. Density and hazard rate estimates with similar shapes were obtained from the previous analyses in Antoniadis et al. (1999) and Kottas (2006). Overall, this example supports the findings from the simulation study regarding the Erlang mixture model's capacity to effectively estimate non-standard density and hazard function shapes.

4.2 Small cell lung cancer data

To illustrate the DDP-based Erlang mixture model with real data, we consider the data set from Ying & Wei (1995) on survival times (in days) of patients with small cell lung cancer. The data correspond to a study designed to evaluate two treatment regimens of drugs, etoposide (E) and cisplatin (P), given with a different sequence, with Arm A denoting the regimen where P is followed by E, and Arm B the regimen where E is followed by P. A total of 121 patients were randomly assigned to one of the treatment arms, resulting in 62 patients in Arm A, and 59 in Arm B. The survival times of 23 patients (15 in Arm A and 8 in Arm B) are administratively right censored.

The DDP-based Erlang mixture model is applied with $x \in \mathcal{X} = \{A, B\}$. The priors are set as follows: $\alpha \sim \text{Ga}(5, 1)$; $\theta_x \overset{ind.}{\sim} \text{Ga}(2, 50)$; $M_x \mid \theta_x \overset{ind.}{\sim} \text{Unif}\left(\lceil 2500/\theta_x \rceil, \ldots, \lceil 10000/\theta_x \rceil\right)$; $\boldsymbol{\mu} \sim \text{N}_2((6.7, 6.3)', 10\,\text{I}_2)$; and, $\Sigma = 3\,\text{I}_2$. Here, $(6.7, 6.3)'$ are the averages of the observed survival times for each treatment after logarithmic transformation.

Posterior mean and interval estimates for the density, survival, and hazard function are compared across the two treatments in Figure 9. The Arm B density estimate is more peaked, and the mode under Arm B is estimated to be smaller than that under Arm A. The posterior mean estimates for the survival functions indicate that survival time under Arm B is stochastically smaller than that under Arm A. However, we note the overlap in the interval estimates for the two treatment survival functions for smaller time points and, more emphatically, for time points beyond about $t = 700$ days. Based on the hazard function posterior mean estimates, the hazard rate under arm B is larger than that under arm A, with the exception of the time interval from about 700 to 1100 days that corresponds to a crossing of the estimated hazard functions. In this case, there is even more substantial overlap of the interval estimates, driven by the large posterior uncertainty for the arm B



(a) Density function (Arm A)

(b) Density function (Arm B)

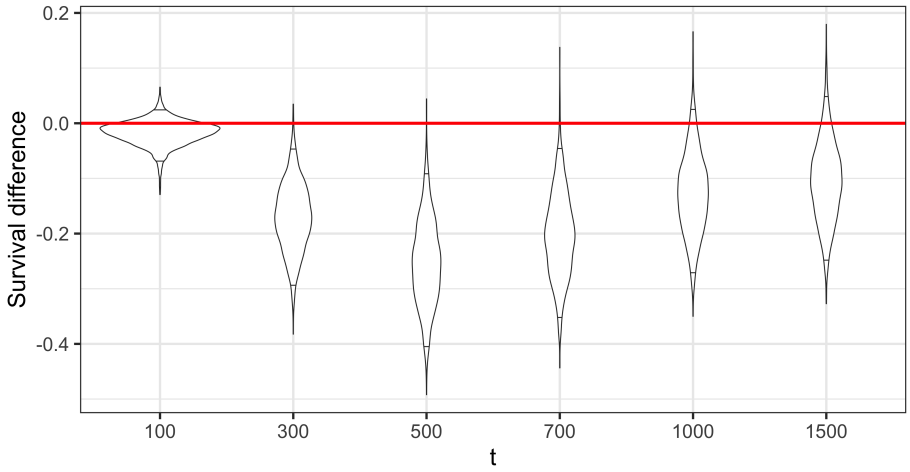
(c) Survival functions

(d) Hazard functions

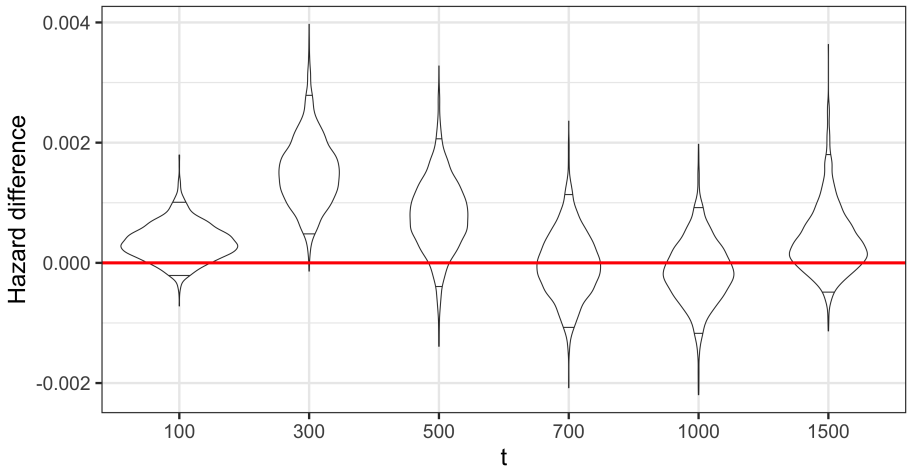
Figure 9: Small cell lung cancer data. Panels (a) and (b) plot estimates for the Arm A and Arm B density; the rug plots show observed (black) and censored (red) survival times. Panels (c) and (d) compare the estimates for the survival and hazard function. In each panel, the dashed lines denote the posterior mean estimates, and the shaded regions indicate the 95% credible intervals. Red and blue color is used for the Arm A and Arm B group, respectively.

hazard rate estimate. Nonetheless, the estimates strongly suggest that the proportional hazards assumption is not suitable for this study.

For a more focused comparison of the two treatments, Figure 10 plots the entire posterior



(a) $S_B(t) - S_A(t)$



(b) $h_B(t) - h_A(t)$

Figure 10: Small cell lung cancer data. Panels (a) and (b) show, through violin plots, the posterior distributions of the difference between the two treatment survival and hazard functions at six specific time points, $t = 100$, 300, 500, 700, 1000, and 1500 days. The short black solid lines within each violin plot indicate the 95% posterior credible interval.

distribution for the difference between the survival and hazard functions at six specific time points, $t = 100$, 300, 500, 700, 1000, and 1500 days. The lines within each violin plot indicate the 95% posterior credible interval for $S_B(t) - S_A(t)$ and $h_B(t) - h_A(t)$, and can thus be contrasted with the horizontal reference line at 0. Based on the 95% interval

estimate, treatment A outperforms treatment B at $t = 300$, 500 and 700 days with respect to survival probability, and at $t = 300$ days according to hazard rate.

5 Summary

We have developed a parsimonious Erlang mixture model as a general methodological tool for nonparametric Bayesian survival analysis. The model is built from a basis representation for the survival density, using Erlang basis densities with a common scale parameter. The weights are defined through increments of a random distribution function, which is flexibly modeled with a Dirichlet process prior. Utilizing a common-weights dependent Dirichlet process prior, the model has been extended to accommodate a categorical covariate associated with a generic control-treatment setting. The proposed methodology provides a useful balance between model flexibility and computational efficiency. The models were illustrated with synthetic and real data examples.

Appendix A: MCMC algorithm for the DP-based Erlang mixture model

In this section, we provide details of posterior simulation for the DP-based Erlang mixture model in Section 2.1. Recall that we have the augmented model using latent variables ϕ_i,

$$\begin{aligned}
t_i \mid \phi_i, \theta, M &\overset{ind.}{\sim} \sum_{m=1}^{M} \mathbb{1}_{B_m}(\phi_i)\mathrm{Ga}(t \mid m, \theta), \\
(\phi_1, \ldots, \phi_n) \mid \alpha, \zeta &\sim \mathrm{Exp}(\phi_1 \mid \zeta) \prod_{i=2}^{n} \left\{ \frac{\alpha}{\alpha + i - 1} \mathrm{Exp}(\phi_i \mid \zeta) + \frac{1}{\alpha + i - 1} \sum_{j=1}^{i-1} \delta_{\phi_j}(\phi_i) \right\}, \\
\zeta &\sim \text{inv-Ga}(a_\zeta, b_\zeta), \\
\theta &\sim \mathrm{Ga}(a_\theta, b_\theta), \\
M \mid \theta &\sim \mathrm{Unif}(\lceil M_1/\theta \rceil, \ldots, \lceil M_2/\theta \rceil), \\
\alpha &\sim \mathrm{Ga}(a_\alpha, b_\alpha),
\end{aligned}$$

where $B_m = ((m-1)\theta, m\theta]$ for $m = 1, \ldots, M-1$, and $B_M = ((M-1)\theta, \infty)$. Here, $\mathrm{Ga}(t \mid a, b)$ denotes the density of the gamma distribution with shape parameter a and scale parameter b evaluated at t, and $\mathrm{Exp}(\phi \mid a)$ the density of the exponential distribution with mean parameter a evaluated at ϕ. The likelihood function under the augmented model can be written as

$$L(M, \theta, \boldsymbol{\phi}; \mathcal{D}) = \prod_{i=1}^{n} \sum_{m=1}^{M} \left\{ \mathbb{1}_{B_m}(\phi_i)\mathrm{Ga}(y_i \mid m, \theta) \right\}^{\nu_i} \left\{ \mathbb{1}_{B_m}(\phi_i) S_{\mathrm{Ga}}(y_i \mid m, \theta) \right\}^{1-\nu_i}, \tag{8}$$

where $S_{\mathrm{Ga}}(y_i \mid m, \theta) = \int_{y_i}^{\infty} \mathrm{Ga}(u \mid m, \theta) du$, $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_n)$, and $\mathcal{D} = \{(y_i, \nu_i), i = 1, \ldots, n\}$. The joint posterior distribution of the random parameters, $\boldsymbol{\phi}, \theta, M, \zeta$, and α is

$$\begin{aligned}
p(\boldsymbol{\phi}, \theta, M, \zeta, \alpha \mid \mathcal{D}) \propto & \prod_{i=1}^{n} \sum_{m=1}^{M} \left\{ \mathbb{1}_{B_m}(\phi_i)\mathrm{Ga}(y_i \mid m, \theta) \right\}^{\nu_i} \left\{ \mathbb{1}_{B_m}(\phi_i) S_{\mathrm{Ga}}(y_i \mid m, \theta) \right\}^{1-\nu_i} \\
& \times p(\boldsymbol{\phi} \mid \alpha, \zeta) p(\zeta) p(\theta) p(M \mid \theta) p(\alpha).
\end{aligned}$$

We use a Metropolis-within-Gibbs algorithm for posterior simulation if direct sampling is not available. The parameters in the proposal distributions for Metropolis-Hastings update are automatically tuned by adaptive Metropolis-Hastings algorithms in Roberts & Rosenthal (2009) for fast convergence and improved mixing. We checked mixing and convergence of the Markov chain and did not find any evidence of converging to a wrong distribution. The full conditionals are given below.

1. M and θ

   - Sample M from the following categorical distribution,

   $$p(M = j_M \mid -) = \frac{L(M = j_M, \theta, \boldsymbol{\phi}; \mathcal{D})}{\sum_{i_M=\lceil \frac{M_1}{\theta} \rceil}^{\lceil \frac{M_2}{\theta} \rceil} L(M = i_M, \theta, \boldsymbol{\phi}; \mathcal{D})}, \qquad j_M = \left\lceil \frac{M_1}{\theta} \right\rceil, \ldots, \left\lceil \frac{M_2}{\theta} \right\rceil,$$

   where $L(j_M, \theta, \boldsymbol{\phi}; \mathcal{D})$ is the likelihood function of the augmented model in (8) evaluated with $M = j_M$ and the current values of $\boldsymbol{\phi}$ and θ.

   - The full conditional of θ is

   $$p(\theta \mid -) \propto \mathrm{Ga}(\theta \mid a_\theta, b_\theta) L(M, \theta, \boldsymbol{\phi}; \mathcal{D}).$$

   We update θ using a random walk Metropolis-Hasting algorithm.

   - We also jointly update (M, θ) via a Metropolis-Hasting algorithm. Given the current values $(M^{(t-1)}, \theta^{(t-1)})$ at iteration t, we first generate a proposal, $\theta^\star$ of θ; $\log(\theta^*) \sim \mathrm{N}(\log(\theta^{(t-1)}), \epsilon)$, where ϵ is an adaptive step size, and generate $M^\star$ from

   $$q(M^\star = j_M \mid M^{(t-1)}, \theta^\star) = \frac{\{(j_M - M^{(t-1)})^2 + 1\}^{-1}}{\sum_{i_M=\lceil \frac{M_1}{\theta} \rceil}^{\lceil \frac{M_2}{\theta} \rceil} \{(i_M - M^{(t-1)})^2 + 1\}^{-1}}, \quad j_M = \left\lceil \frac{M_1}{\theta^*} \right\rceil, \ldots, \left\lceil \frac{M_2}{\theta^*} \right\rceil.$$

   We then accept $(\theta^\star, M^\star)$ with probability $\min(1, r^\star)$, where

   $$r^\star = \frac{\theta^\star p(\theta^\star) p(M^\star \mid \theta^\star) L(M^\star, \theta^\star, \boldsymbol{\phi}; \mathcal{D}) q(M^{(t-1)} \mid \theta^{(t-1)}, M^\star)}{\theta^{(t-1)} p(\theta^{(t-1)}) p(M^{(t-1)} \mid \theta^{(t-1)}) L(M^{(t-1)}, \theta^{(t-1)}, \boldsymbol{\phi}; \mathcal{D}) q(M^\star \mid \theta^\star, M^{(t-1)})}.$$

2. ζ

Let $\boldsymbol{\phi}^\star = (\phi_1^\star, \ldots, \phi_{n^\star}^\star)$ the set of all distinct values in $(\phi_1, \ldots, \phi_n)$ and $n^\star$ the number of elements in $\boldsymbol{\phi}^\star$. The full conditional of ζ is

$$\text{inv-Ga}\left(a_\zeta + n^\star, b_\zeta + \sum_{j=1}^{n^\star} \phi_j^\star\right).$$

3. α

We use the augmentation method in (Escobar & West 1995) to update α. We first introduce an auxiliary variable η, $\eta \mid \alpha, n \sim \text{Be}(\alpha + 1, n)$, and sample α from a mixture of two gamma distributions;

$$\begin{aligned}\alpha \mid - \;\sim\; & \frac{a_\alpha + n^\star - 1}{n(b_\alpha^{-1} - \log(\eta)) + a_\alpha + n^\star - 1}\text{Ga}(a_\alpha + n^\star, (b_\alpha^{-1} - \log(\eta))^{-1}) \\ & + \frac{n(b_\alpha^{-1} - \log(\eta))}{n(b_\alpha^{-1} - \log(\eta)) + a_\alpha + n^\star - 1}\text{Ga}(a_\alpha + n^\star - 1, (b_\alpha^{-1} - \log(\eta))^{-1}).\end{aligned}$$

4. $\boldsymbol{\phi}$

Let $\boldsymbol{\phi}_i^{\star -} = (\phi_1^{\star -}, \ldots, \phi_{n^{\star -}}^{\star -})$ be the set of distinct values in $\boldsymbol{\phi}_{-i}$, where $\boldsymbol{\phi}_{-i} = (\phi_1, \ldots, \phi_{i-1}, \phi_{i+1}, \ldots, \phi_n)$ and $n^{\star -}$ is the number of elements in $\boldsymbol{\phi}_i^{\star -}$. Let n_j^- be the number of elements in $\boldsymbol{\phi}_{-i}$ that equal $\phi_j^{\star -}$. The full conditional of ϕ_i is

$$\begin{aligned}\phi_i \mid \boldsymbol{\phi}_{-i}, y_i, \alpha, \zeta, \theta, M \;\sim\; & \frac{\alpha q_0}{\alpha q_0 + \sum_{j=1}^{n^{\star -}} n_j^- q_j} h(\phi_i \mid y_i, \theta, M, \zeta) \\ & + \sum_{j=1}^{n^{\star -}} \frac{n_j^- q_j}{\alpha q_0 + \sum_{k=1}^{n^{\star -}} n_k^- q_k} \delta_{\phi_j^{\star -}}(\phi_i),\end{aligned}$$

where

$$\begin{aligned}q_0 \;=\; & \sum_{m=1}^{M-1} \{G_{\text{Exp}}(m\theta \mid \zeta) - G_{\text{Exp}}((m-1)\theta \mid \zeta)\} \{\text{Ga}(y_i \mid m, \theta)\}^{\nu_i} \{S_{\text{Ga}}(y_i \mid m, \theta)\}^{1-\nu_i} \\ & + \{1 - G_{\text{Exp}}((M-1)\theta \mid \zeta)\} \{\text{Ga}(y_i \mid m, \theta)\}^{\nu_i} \{S_{\text{Ga}}(y_i \mid m, \theta)\}^{1-\nu_i}, \\ q_j \;=\; & \sum_{m=1}^{M-1} \mathbb{1}_{((m-1)\theta, m\theta]}(\phi_j^{*-}) \{\text{Ga}(y_i \mid m, \theta)\}^{\nu_i} \{S_{\text{Ga}}(y_i \mid m, \theta)\}^{1-\nu_i} \\ & + \mathbb{1}_{((M-1)\theta, \infty)}(\phi_j^{*-}) \{\text{Ga}(y_i \mid m, \theta)\}^{\nu_i} \{S_{\text{Ga}}(y_i \mid m, \theta)\}^{1-\nu_i},\end{aligned}$$

with $G_{\text{Exp}}(\cdot \mid \zeta)$ denoting the exponential distribution function with mean ζ, and

$$h(\phi_i \mid y_i, \theta, M, \zeta) = \sum_{m=1}^{M} \Omega_m \text{T-Exp}_m(\phi_i \mid \zeta),$$

with

$$\begin{aligned}
\Omega_m &= \{\text{Ga}(y_i \mid m, \theta)\}^{\nu_i} \{S_{\text{Ga}}(y_i \mid m, \theta)\}^{1-\nu_i} \\
&\quad \times (G_{\text{Exp}}(m\theta \mid \zeta) - G_{\text{Exp}}((m-1)\theta \mid \zeta)) q_0^{-1}, m = 1, \ldots, M-1, \\
\Omega_M &= \{\text{Ga}(y_i \mid M, \theta)\}^{\nu_i} \{S_{\text{Ga}}(y_i \mid M, \theta)\}^{1-\nu_i} \\
&\quad \times (1 - G_{\text{Exp}}((M-1)\theta \mid \zeta)) q_0^{-1}.
\end{aligned}$$

Here, $h(\phi_i \mid y_i, \theta, M, \zeta)$ is a mixture of truncated exponential distributions, and $\text{T-Exp}_m(\phi \mid \zeta)$ is the density function of the truncated exponential distribution with mean parameter ζ with the support $((m-1)\theta, m\theta]$. ϕ_i is equal to $\phi_j^{\star -}$ with probability $n_j^- q_j / A$, where $A = \alpha q_0 + \sum_{h=1}^{n^{\star -}} n_h^- q_h$; or it is drawn from $h(\phi_i \mid t_i, \theta, M, \zeta)$. The inverse-cdf sampling method can be used to draw a sample from $h(\phi_i \mid t_i, \theta, M, \zeta)$.

Appendix B: MCMC algorithm for the DDP mixture model

We present here the posterior simulation details for the model developed in Section 2.2. The augmented model using latent variables $\boldsymbol{\varphi}_i = (\varphi_{Ci}, \varphi_{Ti})$ is written as

$$\begin{aligned}
t_i \mid M_{x_i}, \theta_{x_i}, \varphi_{x_i} &\overset{ind.}{\sim} \sum_{m=1}^{M_{x_i}} \mathbb{1}_{B_{x_i m}}(\varphi_{x_i,i}) \mathrm{Ga}(t \mid m, \theta_{x_i}),\ i = 1, \ldots, n, \text{ and } x_i \in \{C, T\}, \\
(\boldsymbol{\varphi}_1, \ldots, \boldsymbol{\varphi}_n) \mid \alpha, \boldsymbol{\mu} &\sim \mathrm{LN}_2(\boldsymbol{\varphi}_1 \mid \boldsymbol{\mu}, \Sigma) \prod_{i=2}^{n} \left\{ \frac{\alpha}{\alpha + i - 1} \mathrm{LN}_2(\boldsymbol{\varphi}_i \mid \boldsymbol{\mu}, \Sigma) + \frac{1}{\alpha + i - 1} \sum_{j=1}^{i-1} \delta_{\boldsymbol{\varphi}_j}(\boldsymbol{\varphi}_i) \right\}, \\
\theta_x &\overset{ind.}{\sim} \mathrm{Ga}(a_{x\theta}, b_{x\theta}), \\
M_x \mid \theta_x &\overset{ind.}{\sim} \mathrm{Unif}(\lceil M_{x1}/\theta_x \rceil, \ldots, \lceil M_{x2}/\theta_x \rceil), \\
\alpha &\sim \mathrm{Ga}(a_\alpha, b_\alpha), \\
\boldsymbol{\mu} &\sim \mathrm{N}_2(\bar{\boldsymbol{\mu}}, \Sigma_0),
\end{aligned}$$

where $B_{x_i m} = ((m-1)\theta_{x_i}, m\theta_{x_i}]$ for $m = 1, \ldots, M_{x_i} - 1$, and $B_{x_i M_{x_i}} = ((M_{x_i} - 1)\theta_{x_i}, \infty)$. The likelihood function for the augmented model for observation i is

$$L_i(M_{x_i}, \theta_{x_i}, \varphi_{x_i,i}; \mathcal{D}) = \left[\sum_{m=1}^{M_{x_i}} \mathbb{1}_{B_{x_i m}}(\varphi_{x_i,i}) \mathrm{Ga}(y_i \mid m, \theta_{x_i}) \right]^{\nu_i} \left[\sum_{m=1}^{M_{x_i}} \mathbb{1}_{B_{x_i m}}(\varphi_{x_i,i}) S_{\mathrm{Ga}}(y_i \mid m, \theta_{x_i}) \right]^{1-\nu_i}.$$

where $\mathcal{D} = \{(\dagger_\rangle, \nu_\rangle, \S_\rangle), \rangle = \infty, \ldots, \backslash\}$ denotes data. Similar to the algorithm in Appendix A, we use an adaptive Metropolis-within-Gibbs algorithm in Roberts & Rosenthal (2009) for the Metropolis-Hastings updates. Mixing and convergence of Markov chain are checked and no evidence is found of converging to a wrong distribution. The full conditionals are given below.

1. $\boldsymbol{M} = (M_C, M_T)$

Sample M_C from the following categorical distribution,

$$p(M_C = j_M \mid -) = \frac{L_C(j_M, \theta_C, \boldsymbol{\varphi}; \mathcal{D})}{\sum_{i_M=\left\lceil \frac{M_{C1}}{\theta_C} \right\rceil}^{\left\lceil \frac{M_{C2}}{\theta_C} \right\rceil} L_C(i_M, \theta_C, \boldsymbol{\varphi}; \mathcal{D})}, \qquad j_M = \left\lceil \frac{M_{C1}}{\theta_C} \right\rceil, \ldots, \left\lceil \frac{M_{C2}}{\theta_C} \right\rceil,$$

where $L_C(j_M, \theta_C, \boldsymbol{\varphi}; \mathcal{D}) = \prod_{i:x_i=C} L_i(j_M, \theta_C, \varphi_{Ci}; \mathcal{D})$. We then draw M_T in a similar way.

2. $\boldsymbol{\theta} = (\theta_C, \theta_T)$

   The full conditional of $\boldsymbol{\theta}$ is

$$p(\boldsymbol{\theta} \mid -) \propto \mathrm{Ga}(\theta_C \mid a_{C\theta}, b_{C\theta}) \mathrm{Ga}(\theta_T \mid a_{T\theta}, b_{T\theta}) \prod_{i=1}^{n} L_i(M_{x_i}, \theta_{x_i}, \varphi_{x_i,i}; \mathcal{D}).$$

   We use the algorithm in Roberts & Rosenthal (2009) to sample $\boldsymbol{\theta}$. Let $\boldsymbol{\theta}^{(t-1)} = (\theta_C^{(t-1}, \theta^{(t01)})_T)$ the current values of $\boldsymbol{\theta}$. A proposal of $\boldsymbol{\theta}$ is generated from

$$\log(\boldsymbol{\theta}^\star) \sim 0.95\mathrm{N}(\log(\boldsymbol{\theta}^{(t-1)}), 2.38^2/2\Sigma_n) + 0.05\mathrm{N}(\log(\boldsymbol{\theta}^{(t-1)}), 0.01/2I_2),$$

   where Σ_n is the empirical covariance matrix of $\log(\boldsymbol{\theta})$ based on the run so far. Then we accept $\boldsymbol{\theta}^\star$ with probability $\min(1, r^\star)$, where

$$r^\star = \frac{\theta_C^\star \theta_T^\star \mathrm{Ga}(\theta_C^\star | a_{C\theta}, b_{C\theta}) \mathrm{Ga}(\theta_T^\star \mid a_{T\theta}, b_{T\theta}) \prod_{i=1}^n L_i(M_{x_i}, \theta_{x_i}^\star, \varphi_{x_i,i}; \mathcal{D})}{\theta_C^{(t-1)} \theta_T^{(t-1)} \mathrm{Ga}(\theta_C^{(t-1)} \mid a_{C\theta}, b_{C\theta}) \mathrm{Ga}(\theta_T^{(t-1)} \mid a_{T\theta}, b_{T\theta}) \prod_{i=1}^n L_i(M_{x_i}, \theta_{x_i}^{(t-1)}, \varphi_{x_i,i}; \mathcal{D})}.$$

3. $\boldsymbol{\mu}$

   Let $\boldsymbol{\varphi}^\star = (\boldsymbol{\varphi}_1^\star, \ldots, \boldsymbol{\varphi}_{n^\star}^\star)$ be the set of distinct values in $\boldsymbol{\varphi}$, where $n^\star$ is the number of elements in $\boldsymbol{\varphi}^\star$. The full conditional of $\boldsymbol{\mu}$ is

$$\mathrm{N}_2(\boldsymbol{\mu}_1, \Sigma_1),$$

   where

$$\Sigma_1 = \left[\Sigma_0^{-1} + n^\star \Sigma^{-1}\right]^{-1} \quad \text{and} \quad \boldsymbol{\mu}_1 = \Sigma_1 \left[\Sigma_0^{-1} \bar{\boldsymbol{\mu}} + \Sigma^{-1} \sum_{i=1}^{n^\star} \log(\boldsymbol{\varphi}_i^\star)\right].$$

4. α

We use the augmentation method in Escobar & West (1995) to update α. We first introduce an auxiliary variable η, $\eta \mid \alpha, n \sim \text{Be}(\alpha+1, n)$, and sample α from a mixture of two gamma distributions;

$$\begin{aligned}\alpha \mid - \;\sim\; & \frac{a_\alpha + n^\star - 1}{n(b_\alpha^{-1} - \log(\eta)) + a_\alpha + n^\star - 1}\text{Ga}(a_\alpha + n^\star, (b_\alpha^{-1} - \log(\eta))^{-1}) \\ & + \frac{n(b_\alpha^{-1} - \log(\eta))}{n(b_\alpha^{-1} - \log(\eta)) + a_\alpha + n^\star - 1}\text{Ga}(a_\alpha + n^\star - 1, (b_\alpha^{-1} - \log(\eta))^{-1}).\end{aligned}$$

5. $\boldsymbol{\varphi}$

Let $\boldsymbol{\varphi}_i^{\star -} = (\boldsymbol{\varphi}_1^{\star -}, \ldots, \boldsymbol{\varphi}_{n^{\star -}}^{\star -})$ be the set of distinct values in $\boldsymbol{\varphi}_{-i} = (\boldsymbol{\varphi}_1, \ldots, \boldsymbol{\varphi}_{i-1}, \boldsymbol{\varphi}_{i+1}, \ldots, \boldsymbol{\varphi}_n)$, where $n^{\star -}$ is the number of elements in $\boldsymbol{\varphi}_i^{\star -}$. Let n_j^- be number of elements in $\boldsymbol{\varphi_{-i}}$ that is equal to $\boldsymbol{\varphi}_j^{\star -}$. The full conditional of $\boldsymbol{\varphi}_i$ is

$$\begin{aligned}\boldsymbol{\varphi}_i \mid \boldsymbol{\varphi}_{-i}, \boldsymbol{\mu}, \Sigma, \boldsymbol{\theta}, \boldsymbol{M}, \mathcal{D} \;\sim\; & \frac{\alpha q_0}{\alpha q_0 + \sum_{j=1}^{n^{*-}} n_j^- q_j} h(\boldsymbol{\varphi}_i \mid y_i, \boldsymbol{\mu}, \Sigma, \boldsymbol{\theta}, \boldsymbol{M}) \\ & + \sum_{j=1}^{n^{\star -}} \frac{n_j^- q_j}{\alpha q_0 + \sum_{j=1}^{n^{*-}} n_j^- q_j} \delta_{\boldsymbol{\varphi}_j^{*-}}(\boldsymbol{\varphi}_i),\end{aligned}$$

where, for $x_i = C$,

$$\begin{aligned}q_0 \;=\; & \sum_{m=1}^{M_C - 1} \{\text{Ga}(y_i \mid m, \theta_C)\}^{\nu_i} \{S_{\text{Ga}}(y_i \mid m, \theta_C)\}^{1-\nu_i} \\ & \quad \times \{G_{\text{LN}}(m\theta_C \mid \mu_{C|T}, \Sigma_{C|T}) - (G_{\text{LN}}((m-1)\theta_C \mid \mu_{C|T}, \Sigma_{C|T})\} \\ & \quad + \{\text{Ga}(y_i \mid M_C, \theta_C)\}^{\nu_i} \{S_{\text{Ga}}(y_i \mid M_C, \theta_C)\}^{1-\nu_i} \{1 - G_{\text{LN}}((M_C - 1)\theta_C \mid \mu_{C|T}, \Sigma_{C|T})\}, \\ q_j \;=\; & \sum_{m=1}^{M_C - 1} \mathbb{1}_{((m-1)\theta_C, m\theta_C]}(\varphi_{Ci}^{\star -}) \{\text{Ga}(y_i | m, \theta_C)\}^{\nu_i} \{S_{\text{Ga}}(y_i | m, \theta_C)\}^{1-\nu_i} \\ & \quad + \mathbb{1}_{((M_C - 1)\theta_C, \infty)}(\varphi_{Ci}^{\star -}) \{\text{Ga}(y_i | M_C, \theta_C)\}^{\nu_i} \{S_{\text{Ga}}(y_i | M_C, \theta_C)\}^{1-\nu_i},\end{aligned}$$

$$\mu_{C|T} \;=\; \mu_1 + \Sigma_{12}/\Sigma_{22}(\varphi_{Ti} - \mu_2),$$

$$\Sigma_{C|T} \;=\; \Sigma_{11} - \Sigma_{12}\Sigma_{21}/\Sigma_{22}$$

with $G_{\text{LN}}(\cdot \mid \mu_{C|T}, \Sigma_{C|T})$ denoting a lognormal distribution function with mean $\mu_{C|T}$

and variance $\Sigma_{C|T}$, and

$$h(\boldsymbol{\varphi}_i \mid y_i, \boldsymbol{\mu}, \Sigma, \boldsymbol{\theta}, \boldsymbol{M}) \;=\; \mathrm{LN}(\varphi_{Ti} \mid \mu_1, \Sigma_{11}) \times \sum_{m=1}^{M_C} \Omega_m \text{T-LN}_m(\varphi_{Ci} \mid \mu_{C|T}, \Sigma_{C|T})$$

with

$$\begin{aligned}
\Omega_m &= \{\mathrm{Ga}(y_i \mid m, \theta_C)\}^{\nu_i} \{S_{\mathrm{Ga}}(y_i \mid m, \theta_C)\}^{1-\nu_i} \\
&\qquad \times \{G_{\mathrm{LN}}(m\theta_C \mid \mu_{C|T}, \Sigma_{C|T}) - G_{\mathrm{LN}}((m-1)\theta_C \mid \mu_{C|T}, \Sigma_{C|T})\} q_0^{-1}, \\
&\qquad\qquad \text{for } m = 1, \ldots, M_C - 1, \\
\Omega_{M_C} &= \{\mathrm{Ga}(y_i \mid M_C, \theta_C)\}^{\nu_i} \{S_{\mathrm{Ga}}(y_i \mid M_C, \theta_C)\}^{1-\nu_i} \\
&\qquad \times \{1 - G_{\mathrm{LN}}((M_C - 1)\theta_C \mid \mu_{C|T}, \Sigma_{C|T})\} q_0^{-1}.
\end{aligned}$$

Similar to the algorithm of updating φ_i for the DP-based Erlang mixture model, we let $\boldsymbol{\varphi}_i = \boldsymbol{\varphi}_j^{\star -}$ with probability $n_j^- q_j / A$, where $A = \alpha q_0 + \sum_{h=1}^{n^{\star -}} n_h^- q_h$, or draw a new $\boldsymbol{\varphi}_i$ from $h(\boldsymbol{\varphi}_i \mid y_i, \boldsymbol{\mu}, \Sigma, \boldsymbol{\theta}, \boldsymbol{M})$ with probability $\alpha q_0 / A$. To draw a sample from $h(\boldsymbol{\varphi}_i \mid y_i, \boldsymbol{\mu}, \Sigma, \boldsymbol{\theta}, \boldsymbol{M})$, we first draw φ_{Ti} from $\mathrm{LN}(\mu_1, \Sigma_{11})$ and then, conditional on φ_{Ti}, draw φ_{Ci} from a mixture of truncated lognormal distributions using an inverse-cdf sampling method, where each component, T-LN_m is a lognormal distribution with support of $((m-1)\theta_C, m\theta_C]$. The same method is applied for the observations with $x_i = T$ by simply switching C with T.

References

Antoniadis, A., Grégoire, G. & Nason, G. (1999), 'Density and hazard rate estimation for right-censored data by using wavelet methods', *Journal of the Royal Statistical Society: Series B (Methodological)* **61**, 63–84.

Blackwell, D. & MacQueen, J. B. (1973), 'Ferguson distributions via Pólya urn schemes', *Annals of Statistics* **1**, 353–355.

Butzer, P. (1954), 'On the extensions of Bernstein polynomials to the infinite interval', *Proceedings of the American Mathematical Society* **5**, 547–553.

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

Escobar, M. D. & West, M. (1995), 'Bayesian density estimation and inference using mixtures', *Journal of the American Statistical Association* **90**, 577–588.

Ferguson, T. S. (1973), 'A Bayesian analysis of some nonparametric problems', *The Annals of Statistics* **1**, 209–230.

Hanson, T. E. (2006), 'Modeling censored lifetime data using a mixture of gammas baseline', *Bayesian Analysis* **1**, 575–594.

Ibrahim, J. G., Chen, M. & Sinha, D. (2001), *Bayesian Survival Analysis*, Springer, New York, NY.

Kim, H. & Kottas, A. (2022), 'Erlang mixture modeling for Poisson process intensities', *Statistics and Computing* **32**, 3.

Kottas, A. (2006), 'Nonparametric Bayesian survival analysis using mixtures of Weibull distributions', *Journal of Statistical Planning and Inference* **136**, 578–596.

Lee, S. C. K. & Lin, X. S. (2010), 'Modeling and evaluating insurance losses via mixtures of Erlang distributions', *North American Actuarial Journal* **14**, 107–130.

MacEachern, S. N. (2000), 'Dependent Dirichlet processes', *Technical Report, Ohio State University* .

Mitra, R. & Müller, P., eds (2015), *Nonparametric Bayesian Inference in Biostatistics*, Springer, Cham, Switzerland.

Müller, P., Quintana, F. A., Jara, A. & Hanson, T. (2015), *Bayesian Nonparametric Data Analysis*, Springer, Cham, Switzerland.

Neal, M. (2000), 'Markov Chain sampling methods for Dirichlet process mixture models', *Journal of Computational and Graphical Statistics* **9**, 249–265.

Phadia, E. G. (2013), *Prior Processes and Their Applications*, Springer, Berlin Heidelberg.

Poynor, V. & Kottas, A. (2019), 'Nonparametric Bayesian inference for mean residual life functions in survival analysis.', *Biostatistics* **20**, 240–255.

Quintana, F. A., Müller, P., Jara, A. & MacEachern, S. N. (2022), 'The dependent Dirichlet process and related models', *Statistical Science* **37**, 24–41.

Roberts, G. O. & Rosenthal, J. S. (2009), 'Examples of adaptive MCMC', *Journal of Computational and Graphical Statistics* **18**, 349–367.

Sethuraman, J. (1994), 'A constructive definition of Dirichlet priors', *Statistica Sinica* **4**, 639–650.

Venturini, S., Dominici, F. & Parmigiani, G. (2008), 'Gamma shape mixtures for heavy-tailed distributions', *The Annals of Applied Statistics* **2**, 756–776.

Xiao, S., Kottas, A., Sansó, B. & Kim, H. (2021), 'Nonparametric Bayesian modeling and estimation for renewal processes', *Technometrics* **63**, 100–115.

Ying, Z., J. S. & Wei, L. (1995), 'Survival analysis with median regression models', *Journal of the American Statistical Association* **90**, 178–184.