Nonparametric Bayesian Survival Analysis using Mixtures of Weibull Distributions

ATHANASIOS KOTTAS
University of California, Santa Cruz

ABSTRACT. Bayesian nonparametric methods have been applied to survival analysis problems since the emergence of the area of Bayesian nonparametrics. However, the use of the flexible class of Dirichlet process mixture models has been rather limited in this context. This is, arguably, to a large extent, due to the standard way of fitting such models that precludes full posterior inference for many functionals of interest in survival analysis applications. To overcome this difficulty, we provide a computational approach to obtain the posterior distribution of general functionals of a Dirichlet process mixture. We model the survival distribution employing a flexible Dirichlet process mixture, with a Weibull kernel, that yields rich inference for several important functionals. In the process, a method for hazard function estimation emerges. Methods for simulation-based model fitting, in the presence of censoring, and for prior specification are provided. We illustrate the modeling approach with simulated and real data.

Key words: censored observations, Dirichlet process mixture models, hazard function, survival function

Running headline: Nonparametric Bayesian survival analysis

1. Introduction

The combination of the Bayesian paradigm and nonparametric methodology requires the construction of priors on function spaces. The area of Bayesian nonparametrics has grown rapidly following the work of Ferguson (1973) on the Dirichlet process (DP), a random probability measure on spaces of distribution functions. Bayesian nonparametric methods
are very well suited for survival data analysis, enabling flexible modeling for the unknown survival function, cumulative hazard function or hazard function, providing techniques to handle censoring and truncation, allowing incorporation of prior information and yielding rich inference that does not rely on restrictive parametric specifications.

We consider fully nonparametric modeling for survival analysis problems that do not involve a regression component. (See Ibrahim et al., 2001, chapters 3 and 10, for a review of Bayesian semiparametric regression modeling for survival data.) In this context, most of the existing approaches concentrate on a specific functional of the survival distribution and develop priors for the associated space of random functions.

Regarding prior models over the space of distribution functions (equivalently survival functions), early inferential work, typically restricted to point estimation, involved the DP, as in, e.g., Susarla & Van Ryzin (1976), and the class of neutral to the right processes (Doksum, 1974), as in Ferguson & Phadia (1979). More recently, simulation-based model fitting enabled the full Bayesian analyses of Doss (1994), Muliere & Walker (1997) and Walker & Damien (1998) based on mixtures of Dirichlet processes (Antoniak, 1974), Polya tree priors (Ferguson, 1974, Lavine, 1992) and beta-Stacy process priors (Walker & Muliere, 1997), respectively. Modeling for the cumulative hazard function employs Gamma process priors, as in Kalbfleisch (1978), or beta process priors (Hjort, 1990), as in Damien et al. (1996). In the context of hazard function estimation, Dykstra & Laud (1981) proposed the extended Gamma process, a model for monotone hazard functions. Other approaches for modeling hazard rates include Ammann (1984), Arjas & Gasbarra (1994) and Nieto-Barajas & Walker (2002).

DP mixture models form a very rich class of Bayesian nonparametric models. They emerge from the assumption that the mixing distribution, in a mixture of a parametric family of distributions, arises from a DP. DP mixtures have dominated the Bayesian nonparametric literature after the machinery for their fitting, using Markov chain Monte Carlo (MCMC) methods, was developed following the work of Escobar (1994). Being essentially countable mixtures of parametric distributions, they provide the attractive features and flexibility of mixture modeling. Given their popularity in the Bayesian nonparametric literature, it is somewhat surprising that very little work exists on DP mixture modeling.
for distributions supported by $R^+$, with applications in survival data analysis. In fact, even in this work, the problem is tackled with normal mixtures through the use of a logarithmic transformation of the data (as in Kuo & Mallick, 1997). No approach appears to exist that employs a DP mixture model, with a kernel having support on $R^+$, to provide inference for general functionals of survival populations. This can be attributed to the common feature of the MCMC algorithms that have been developed for fitting DP mixture models, being the marginalization over the mixing distribution. Hence the posterior of the mixing distribution is not obtained resulting in limited inference for many functionals of the mixture distribution. In particular, the posterior of non-linear functionals cannot be obtained. Thus most functionals of interest for survival populations, including the cumulative hazard function, hazard function and percentile life functionals, cannot be studied.

In this paper, we model the unknown survival distribution with a Weibull DP mixture model, mixing on both the shape and scale parameters of the Weibull kernel. We develop an efficient MCMC algorithm to fit the model to uncensored and right censored data. Moreover, extending the work of Gelfand & Kottas (2002), we show how the output of the MCMC algorithm can be used to obtain draws from the posterior of general functionals, linear and non-linear, of the mixture model. Hence, modeling the distribution function of the survival population with a flexible nonparametric mixture, full posterior inference is enabled for essentially any survival population functional that might be of interest. In particular, the model yields smooth data-driven estimates for the density function, survival function, cumulative hazard function and hazard function and quantifies the associated posterior uncertainty. We note here that obtaining the posterior of certain functionals of the survival distribution is not straightforward under some of the existing Bayesian nonparametric models, e.g., when modeling the hazard function to begin with. Finally, we demonstrate the utility of the model in comparisons of survival populations performed without forcing any specific relation, e.g., location-scale shift models or proportional hazards models, instead letting the data determine the form of differences for the functionals of interest.

The paper is organized as follows. Section 2 briefly reviews DP mixture models. Sec-
tion 3 provides a computational approach to obtain the posterior of functionals of DP mixtures. Section 4 presents the Weibull DP mixture model including methods for posterior inference (with the details given in the Appendix) and prior specification. Section 5 considers the analyses of three datasets to illustrate the model. Finally, section 6 offers a summary and discussion of related future research.

2. Dirichlet process mixture models

A DP mixture model is a mixture with a parametric kernel and a random mixing distribution modeled with a DP prior. The definition of the DP involves a parametric distribution function $G_0$, the center or base distribution of the process, and a positive scalar precision parameter $\nu$. The larger the value of $\nu$ the closer a realization of the process is to $G_0$. See Ferguson (1973, 1974) for the formal development. We write $G \sim DP(\nu G_0)$ to denote that a DP prior is placed on the distribution function $G$.

More explicitly, a DP mixture model is given by

$$F(\cdot;G) = \int K(\cdot | \theta) G(d\theta), \quad (1)$$

where $K(\cdot | \theta)$ is the distribution function of the parametric kernel of the mixture and $G_0 \sim DP(\nu G_0)$. If $k(\cdot | \theta)$ is the density corresponding to $K(\cdot | \theta)$, the density of the random mixture in (1) is $f(\cdot;G) = \int k(\cdot | \theta)G(d\theta)$. We refer to Antoniak (1974) as well as Ferguson (1983), Lo (1984) and Brunner & Lo (1989) for details on the theoretical aspects of such random mixture models. Equivalently, a DP mixture model can be viewed as a hierarchical model, where associated with each observation $Y_i$ of the data vector $D = \{Y_i, i = 1, \ldots, n\}$ is a latent $\theta_i$. Conditionally on the $\theta_i$, the $Y_i$ are assumed independent from $K(\cdot | \theta_i)$. Next, the $\theta_i$ given $G$ are independent and identically distributed (i.i.d.) from $G$ and finally $G \sim DP(\nu G_0)$. This is the simplest fully nonparametric version of the model. Further stages in the hierarchy can be added by assuming that $\nu$ and/or the parameters of $G_0$ are random. Moreover, semiparametric specifications emerge by writing $\theta = (\theta_1, \theta_2)$ and DP mixing on $\theta_1$ with a parametric prior for $\theta_2$.

Simulation-based model fitting for DP mixture models is well developed by now. The key idea is the marginalization over $G$ (Antoniak, 1974) which enables the construction of a Gibbs sampler to draw from the resulting finite dimensional posterior $[\theta_1, \ldots, \theta_n | D]$. (Here-
after we use the bracket notation for conditional and marginal distributions.) This Gibbs sampler is straightforward to implement provided it is easy to evaluate \[ \int k(\cdot \mid \theta) G_0(d\theta), \]
either in closed form or numerically, and it is feasible to draw from the distribution with density proportional to \[ k(\cdot \mid \theta) g_0(\theta), \]
where \( g_0 \) is the density of \( G_0 \) (Escobar, 1994, Escobar & West, 1995, West et al., 1994, and Bush & MacEachern, 1996). Other approaches (see, e.g., MacEachern & Müller, 1998, and Neal, 2000) have been proposed for DP mixtures for which it is difficult or inefficient to perform these operations.

However, regardless of the MCMC method used to fit a DP mixture model, inference for functionals of \( F(\cdot; G) \) is limited to moments of linear functionals (Gelfand & Mukhopadhyay, 1995). Gelfand & Kottas (2002) proposed an approach that enables full inference for general functionals by sampling (approximately) from the DP after fitting the model with one of the existing MCMC algorithms. The method was presented in the setting where \( \nu \) and the parameters of \( G_0 \) are fixed. In the next section we provide an extension that allows these parameters to be random and forms the basis of our approach for functionals of survival distributions.

3. Inference for functionals of Dirichlet process mixtures

Consider a generic DP mixture model, as described in section 2, with independent priors \([\nu]\) and \([\psi]\) placed on \( \nu \) and the parameters \( \psi \) of \( G_0 = G_0(\cdot \mid \psi) \). Hence the full hierarchical model becomes

\[
\begin{align*}
Y_i \mid \theta_i & \overset{ind.}{\sim} K(\cdot \mid \theta_i), i = 1, \ldots, n \\
\theta_i \mid G & \overset{i.i.d.}{\sim} G, i = 1, \ldots, n \\
G \mid \nu, \psi & \sim DP(\nu G_0); G_0 = G_0(\cdot \mid \psi) \\
\nu, \psi & \sim [\nu][\psi].
\end{align*}
\]

Let \( H(F(\cdot; G)) \) denote a functional of the random mixture in (1) with posterior \([H(F(\cdot; G)) \mid D]. \) If \( H \) is a linear functional, Fubini’s theorem yields

\[
H(F(\cdot; G)) = \int H(K(\cdot \mid \theta_0)) G(d\theta_0).
\]

This formula suggests a Monte Carlo integration for a realization from \([H(F(\cdot; G)) \mid D] \)
provided we are able to draw from \([G \mid D]\). To this end, note that

\[
[\theta_0, \theta, G, \nu, \psi \mid D] \propto [\theta_0 \mid G][G \mid \theta, \nu, \psi][\theta, \nu, \psi \mid D],
\]

where \(\theta_0\) and \(\theta = (\theta_1, \ldots, \theta_n)\) are conditionally independent given \(G\). Here \([G \mid \theta, \nu, \psi]\) is a DP with updated precision parameter \(\nu + n\) and base distribution \((\nu + n)^{-1} (\nu G_0(\cdot \mid \psi) + \sum_{i=1}^{n} \delta_{\theta_i}(\cdot))\), where \(\delta_a\) denotes the degenerate distribution at \(a\) (Ferguson, 1973, Antoniak, 1974). Therefore to sample from \([H(F(\cdot; G)) \mid D]\), we first fit model (2), employing one of the available MCMC methods, to obtain \(B\) draws \(\theta_b = (\theta_{b1}, \ldots, \theta_{bn}), \nu_b, \psi_b, b = 1, \ldots, B\), from \([\theta, \nu, \psi \mid D]\). Then for each \(b = 1, \ldots, B\):

(i) draw \(G_b \sim [G \mid \theta_b, \nu_b, \psi_b]\)

(ii) draw \(\theta_{bl} \sim G_b\) for \(l = 1, \ldots, L\)

(iii) compute \(H_b = L^{-1} \sum_{l=1}^{L} H(K(\cdot \mid \theta_{bl}))\).

Finally, \((H_b, b = 1, \ldots, B)\) are draws from the posterior of the linear functional \(H(F(\cdot; G))\).

In general, the Monte Carlo approximation to the integral in (3) stabilizes for \(L = 1,000\). (A conservative value \(L = 2,500\) was used for the examples in section 5).

Implementing step (i) requires sampling from a DP for which we use its constructive definition given by Sethuraman & Tiwari (1982) and Sethuraman (1994). According to this construction, a realization from \([G \mid \theta, \nu, \psi]\) is almost surely of the form \(\sum_{j=1}^{\infty} \omega_j \delta_{\theta_j}\), where \(\omega_1 = z_1, \omega_j = z_j \prod_{s=1}^{j-1} (1 - z_s), j = 2,3,\ldots, \) with \(z_s\) \(\sim\) Beta(1, \(\nu_b + n\)) and independently \(\theta_j \mid \theta, \nu, \psi\) i.i.d. from \((\nu_b + n)^{-1} (\nu_b G_0(\cdot \mid \psi) + \sum_{i=1}^{n} \delta_{\theta_i}(\cdot))\). We work with an approximate realization \(G_J = \sum_{j=1}^{J} w_j \delta_{\theta_j}\), where \(w_j = \omega_j, j = 1,\ldots, J-1\) and \(w_J = 1 - \sum_{j=1}^{J-1} w_j = \prod_{s=1}^{J-1} (1 - z_s)\). Noting that \(E(\sum_{j=1}^{J} \omega_j \mid \nu_b) = 1 - \{(\nu_b + n)/(\nu_b + n + 1)\}^J\), we specify \(J\) so that \(\{(\nu_b + n)/(\nu_b + n + 1)\}^J = \epsilon\), for small \(\epsilon\). (\(\epsilon = 0.0001\) was used for the examples of section 5.) Results on sensitivity analysis for the value of \(J\) suggest that this is a reliable choice and, in fact, somewhat conservative since typically smaller values of \(J\) produce essentially identical inference.

The approach yields the posterior of any linear functional as well as the posterior of any function of one or more linear functionals. Hence full inference is available for many non-linear functionals that can be expressed as functions of linear functionals, e.g., the cumulative hazard function functional \(\Lambda(t_0; G) = -\log(1 - F(t_0; G))\) and the hazard function functional \(\lambda(t_0; G) = f(t_0; G)/(1 - F(t_0; G))\), for any fixed \(t_0\). An important
non-linear functional, that cannot be handled in this fashion, is the quantile functional, denoted by $\eta_p(F(\cdot; G))$, $p \in (0, 1)$. However, its posterior can be obtained by drawing samples (of size $B$) from the posterior of the distribution function functional $F(t_m; G)$, for a grid of values $t_m$, $m = 1, \ldots, M$, over the support of $F(\cdot; G)$. The columns of the resulting $M \times B$ matrix yield random realizations from $[F(\cdot; G) | D]$ that can be inverted (with interpolation) to provide draws from $[\eta_p(F(\cdot; G)) | D]$.

The method to choose $J$, suggested above, provides a practical way to implement the algorithm. However, formal justification for the approach involves the study of convergence properties, as $J \to \infty$, of sequences of random variables $H(F(\cdot; G_J))$, defined by functionals arising under the partial sum approximation. Note that the approximation is used at each iteration $b$ given the draw $\theta_b$, $\nu_b$, $\psi_b$ from $[\theta, \nu, \psi | D]$. Hence the random variables of interest are $H(F(\cdot; G_J))$, $G_J$ being the partial sum approximation to a realization from $[G | \theta, \nu, \psi]$, where $\theta$, $\nu$, $\psi$ follow $[\theta, \nu, \psi | D]$. Gelfand & Kottas (2002) studied the limiting behavior, as $J \to \infty$, of $H(F(\cdot; G_J)) - H(F(\cdot; G))$, under certain conditions on $H$, $K(\cdot | \theta)$ and $G_0$, when $\nu$ and $\psi$ are fixed. The fact that the approximation is applied conditionally on $\theta$, $\nu$, $\psi$ makes the theorems in that paper applicable in this more general setting with minor modifications required in the proofs. Therefore, here, we only state the results indicating their applications in survival analysis problems.

**Lemma 1:** For any bounded linear functional $H$, $H(F(\cdot; G_J))$ converges to $H(F(\cdot; G))$ almost surely as $J \to \infty$.

**Lemma 2:** For any linear functional $H$ that satisfies $\int (H(K(\cdot | \theta)))^2 G_0(d\theta) < \infty$, $H(F(\cdot; G_J))$ converges to $H(F(\cdot; G))$ in mean of order 2 as $J \to \infty$.

**Lemma 3:** If $K(\cdot | \theta)$ has continuous support then for any fixed $p \in (0, 1)$, outside a set of Lebesgue measure 0, $\eta_p(F(\cdot; G_J))$ converges in probability to $\eta_p(F(\cdot; G))$ as $J \to \infty$.

For survival populations, Lemma 1 yields almost sure convergence for the survival function functional and convergence in probability for the cumulative hazard function functional, regardless of the kernel of the mixture and the base distribution of the DP. Provided the condition of Lemma 2 is satisfied for $H(K(\cdot | \theta)) = k(t_0 | \theta)$, for fixed $t_0$, we obtain convergence in quadratic mean for the density function functional and, combining this result with Lemma 1, convergence in probability for the hazard function functional.
Finally, for continuous survival populations, we have convergence in probability for the median survival time functional as well as for general percentile life functionals.

Finally, a similar method provides draws from the prior distribution of $H(F(\cdot; G))$. Again, the approach is motivated by formula (3) for linear functionals. Now $[\theta_0, G, \nu, \psi] \propto [\theta_0 | G][G | \nu, \psi][\nu][\psi]$. Hence steps (ii) and (iii) remain the same and, instead of step (i), we draw $\nu_b \sim [\nu]$, $\psi_b \sim [\psi]$ and then $G_b \sim [G | \nu_b, \psi_b] = DP(\nu_b G_0(\cdot | \psi_b))$. Here, for the number of terms $J$ in the partial sum approximation of the DP we take the value that satisfies $\{\max_b \nu_b/(1 + \max_b \nu_b)\}^J = \epsilon$, for small $\epsilon$. All theoretical results are readily extended. Prior distributions of functionals are useful indicating prior to posterior learning and the implications of the prior hyperparameters for $\nu$ and $\psi$.

4. A nonparametric mixture model for survival distributions

Section 4.1 motivates the modeling approach and presents the mixture model. Section 4.2 provides a strategy for posterior inference with the details given in the Appendix. Prior specification is discussed in section 4.3.

4.1. The model

Modeling with DP mixtures for distributions with support on $R$ (or $R^d$) typically employs normal (or multivariate normal) kernels (see, e.g., Ferguson, 1983, Escobar & West, 1995, and Müller et al., 1996). Besides the convenient form of normal densities, there is theoretical support for this choice (Ferguson, 1983, Lo, 1984). Other representation results (Choquet-type theorems) suggest the choice of the kernel of the DP mixture when certain distributional shapes or properties are desired (see, e.g., Brunner & Lo, 1989, Brunner, 1992, and Kottas & Gelfand, 2001).

In the context of survival analysis, the choice of the kernel is more delicate than for mixtures with support on $R$. Apart from flexible density shapes, e.g., allowing for skewness and multimodality, here other functionals are also important. In particular, we seek mixtures that can provide rich inference for the hazard function, being able to capture the shape of monotone and non-monotone hazards. In this regard, it is well known that the hazard function of a mixture with kernel with decreasing hazard is de-
creasing and, in fact, the same holds true for mixtures of exponential distributions that have constant hazard rates. It is also possible for mixtures with increasing hazard kernels to have ultimately decreasing hazard functions. See Gurland & Sethuraman (1995) for specific results on the reversal of increasing hazard rates through mixing. Even though analytic results are possible only for certain classes of mixtures, such results suggest that we need kernels that allow for increasing, including rapidly increasing, hazard functions. In a fully nonparametric setting, this points out the main drawback of the model \( f \Phi((\cdot - \mu)/\sigma)G(d\mu, d\sigma^2) \), where \( \Phi(\cdot) \) is the standard normal distribution function, for survival data on the (natural) logarithmic scale. This model, the natural choice given the existing work for distributions on \( R \), is equivalent to a mixture of lognormal distributions, for the data on the original scale, and will therefore produce ultimately decreasing hazard functions, because the hazard rate of a lognormal distribution is either decreasing or increases to a maximum and then decreases to 0 as time approaches infinity. Similar considerations exclude loglogistic or inverse Gaussian distributions.

Weibull, or possibly Gamma, kernels emerge as promising choices when fully nonparametric inference is sought for a range of functionals of the survival distribution. In terms of hazard function estimation, the Weibull distribution is preferable because it possesses hazards that increase more rapidly than those of the Gamma distribution. Moreover, its survival function is available in closed form making it computationally more attractive, especially for samples with censored observations, as our MCMC algorithm illustrates (see section 4.2).

Denote by \( K_W(t | \alpha, \lambda) = 1 - \exp(-\lambda^{-1}t^\alpha) \) and by \( k_W(t | \alpha, \lambda) = \lambda^{-1}t^{\alpha-1} \exp(-\lambda^{-1}t^\alpha) \) the distribution function and density function, respectively, of the Weibull distribution with shape parameter \( \alpha > 0 \) and scale parameter \( \lambda > 0 \). We model the distribution function of the survival population with the mixture

\[
F(\cdot; G) = \int K_W(\cdot | \alpha, \lambda) G(d\alpha, d\lambda),
\]

where \( G \sim DP(\nu G_0) \). Mixing on both the shape and scale parameters of the Weibull kernel results in a flexible mixture that can model a wide range of distributional shapes, in fact, approximate arbitrarily well any density on \( R^+ \). To see this, recall, from section 3, the almost sure representation for realizations from a DP, that yields the (almost sure)
representation $\sum_{j=1}^{\infty} \omega_j k_W(\cdot \mid \alpha_j, \lambda_j)$ for the density $f(\cdot \mid G)$ of the mixture. Next, note that, for any $t_0 \in R^+$, we can find $\alpha_0$ and $\lambda_0$ such that the density $k_W(\cdot \mid \alpha_0, \lambda_0)$ is centered at $t_0$ with arbitrarily small dispersion, e.g., we can set the median equal to $t_0$ and the interquartile range equal to $\varepsilon$, for (arbitrarily) small $\varepsilon$, and solve for $\alpha_0$ and $\lambda_0$. (In fact, under this choice, unless $t_0$ is close to 0, e.g., $t_0 < 0.0063$ for $\varepsilon = 0.01$, $k_W(\cdot \mid \alpha_0, \lambda_0)$ is unimodal with mode that is essentially identical to the median for small enough $\varepsilon$.) Finally, the argument is completed noting that mixtures of point masses are dense in the weak star topology (see, e.g., Diaconis & Ylvisaker, 1985).

The base distribution $G_0$ of a DP mixture model is typically chosen so that prior to posterior analysis is efficient, the model is flexible and prior information can be incorporated through the parameters of $G_0$. Although a base distribution $G_0$ yielding a closed form expression for the integral $\int k_W(\cdot \mid \alpha, \lambda)G_0(d\alpha,d\lambda)$ is not available, the choice

$$G_0(\alpha, \lambda \mid \phi, \gamma) = \text{Uniform}(\alpha \mid 0, \phi)\text{IGamma}(\lambda \mid d, \gamma)$$

(4)

achieves essentially all the aforementioned goals. (Here, $\text{IGamma}(\cdot \mid a, b)$ denotes the inverse Gamma distribution with mean $b/(a-1)$, provided $a > 1$.) Under this choice, the integral above reduces to that of a smooth function over a bounded interval and is easy to compute using numerical integration. Hence the, generally more efficient and easier to implement, standard Gibbs sampler (West et al., 1994, Bush & MacEachern, 1996) can be employed to fit the model. Of course, fixing $\phi$ would be restrictive and choosing its value awkward. Assuming $\phi$ random overcomes these difficulties. We set $d = 2$ yielding an inverse Gamma distribution with infinite variance. Flexibility is added by taking $\gamma$ random. In section 4.3 we discuss how prior information, in the form of prior percentiles, regarding the survival population can be used to specify the parameters of the priors for $\gamma$ and $\phi$. These are taken to be Gamma and Pareto distributions, respectively, leading to convenient updates in the Gibbs sampler. Finally, we place a Gamma prior on $\nu$ which also facilitates the implementation of the Gibbs sampler (Escobar & West, 1995).

If $t_i, i = 1, \ldots, n$ are the survival times, the full Bayesian model can be written in the
following hierarchical form

\[
\begin{align*}
  t_i & \mid \alpha_i, \lambda_i \quad \text{i.d.} \quad \sim K_W(t_i \mid \alpha_i, \lambda_i), i = 1, \ldots, n \\
  (\alpha_i, \lambda_i) & \mid G \quad \text{i.i.d.} \quad \sim \quad G_i, i = 1, \ldots, n \\
  G & \mid \nu, \gamma, \phi \quad \sim \quad DP(\nu G_0) \\
  \nu & \quad \sim \quad \text{Gamma}(\nu \mid a_\nu, b_\nu) \\
  \gamma & \quad \sim \quad \text{Gamma}(\gamma \mid a_\gamma, b_\gamma) \\
  \phi & \quad \sim \quad \text{Pareto}(\phi \mid a_\phi, b_\phi),
\end{align*}
\]

with \( G_0 \) defined in (4) and \( d \) and all the parameters of the priors for \( \nu, \gamma \) and \( \phi \) fixed. (Here, \( \text{Gamma}(\cdot \mid a, b) \) denotes the Gamma distribution with mean \( a/b \).) This expression of the model is generic with no assumption made regarding the status of survival times, uncensored or censored. The distinction is considered in the next section where we discuss posterior inference for model (5).

### 4.2. Posterior inference

Assume that the sample, of size \( n = n_o + n_c \), from the survival population consists of \( n_o \) uncensored survival times, \( t_{i_o}, i_o = 1, \ldots, n_o \) and \( n_c \) right censored survival times, \( z_{i_c}, i_c = 1, \ldots, n_c \). As discussed in section 2, simulation-based model fitting for models of the form in (5) proceeds by integrating out the random distribution \( G \). In our case, this leads to the posterior \( [(\alpha_1, \lambda_1), \ldots, (\alpha_n, \lambda_n), \nu, \gamma, \phi \mid D] \), where \( D = \{t_1, \ldots, t_{n_o}, z_1, \ldots, z_{n_c}\} \), that can be obtained using Gibbs sampling, as detailed in the Appendix, following West et al. (1994) and Bush & MacEachern (1996). We incorporate censoring by exploiting the closed form for the survival function of the Weibull distribution. This version of the Gibbs sampler for fitting DP mixture models to censored data does not appear to exist in the literature. Censoring is typically handled with a data augmentation technique (see, e.g., Kuo & Mallick, 1997, and Kottas & Gelfand, 2001).

Implementing the Gibbs sampler, we obtain draws from \( [(\alpha_1, \lambda_1), \ldots, (\alpha_n, \lambda_n), \nu, \gamma, \phi \mid D] \) which are then used with the approach of section 3 to sample from posteriors of survival population functionals. It is straightforward to verify that the condition of Lemma 2 holds for the functional \( H(K_W(\cdot \mid \alpha, \lambda)) = k_W(t_0 \mid \alpha, \lambda) \), with fixed \( t_0 \), making the convergence results for survival functionals, discussed in section 3, applicable. In section 5 we illustrate
with the posteriors of $f(t_0; G)$, $1 - F(t_0; G)$, $\lambda(t_0; G)$, for fixed time point $t_0$, and the posterior of median survival time $\eta_{0.5}(F(\cdot; G))$. Obtaining the former posteriors over a grid of $t_0$ values and connecting the corresponding point estimates (posterior means or medians) and interval estimates (based on posterior percentiles), we can provide posterior point estimates for the density function, survival function and hazard function functionals along with the associated uncertainty bands. Using the analogous approach outlined at the end of section 3, similar inference summaries are available for prior distributions of functionals, induced by the prior choices in model (5).

Model (5) yields a new method for flexible data-driven hazard function estimation. It recovers successfully many hazard shapes, as indicated by our practical experience with simulated data from various distributions with nonstandard shapes for their hazard functions. (Section 5.1 provides an illustration.) Of course, (5) models the distribution function of the survival population and hence cannot, in general, be used to incorporate hazard specific prior information, e.g., model specific hazard shapes. One exception is for decreasing hazard rates that can be forced by model (5) if we restrict the shape parameter of the Weibull kernel to lie in (0,1), with the corresponding adjustment for $G_0$ in (4).

4.3. Prior specification

To apply model (5), values for the parameters of the priors for $\nu$, $\gamma$ and $\phi$ must be chosen. Here, we provide a simple recommendation for choice of these values. In section 5 we present some results on prior sensitivity analysis.

Recall from section 2 that the parameter $\nu$ of the DP prior $DP(\nu G_0)$ controls how close a realization of the process is to the base distribution $G_0$. In the DP mixture model (5), $\nu$ controls the distribution of the number of distinct elements of the vector $((\alpha_1, \lambda_1), \ldots, (\alpha_n, \lambda_n))$ and hence the number of distinct components of the mixture. (See Antoniak, 1974, and Escobar & West, 1995, for more details.) Therefore, prior information about the number of components can be incorporated through the prior for $\nu$. In the absence of strong prior information in this direction, it appears natural to choose values for $a_\nu$ and $b_\nu$, yielding Gamma priors for $\nu$ that place mass both on small and large values. Practical experience with model (5), based on several real and simulated datasets, suggests
that there is posterior learning for $\nu$ when sample sizes are moderate to large (e.g., $n > 50$). However, with small sample sizes, it appears to be difficult for the data to inform about $\nu$. (See section 5.2 on the effect of this sensitivity to the prior for $\nu$ on posterior inference for functionals.)

Regarding the choice of prior hyperparameters for $\gamma$ and $\phi$, we simplify by setting $a_{\gamma} = 1$, resulting in an exponential prior for $\gamma$, and $a_{\phi} = 2$, resulting in infinite prior variance for $\phi$. To center the priors for $\gamma$ and $\phi$, i.e., choose $b_{\gamma}$ and $b_{\phi}$, we propose the following method, based on the parametric version of model (5), that replaces the second stage of (5) with $(\alpha_i, \lambda_i) \mid \gamma, \phi \overset{i.i.d.}{\sim} G_0$, with $G_0$ given in (4). Marginalizing over $\gamma$ and $\phi$ with respect to their priors Gamma(1, $b_{\gamma}$) and Pareto(2, $b_{\phi}$), respectively, we get the induced marginal priors $[\lambda] = 2b_{\gamma}/\{(1 + \lambda b_{\gamma})^3\}$, $\lambda > 0$, and $[\alpha] = 2b_{\phi}^2/\{3\max\{\alpha, b_{\phi}\}\}^3\}$, $\alpha > 0$. Now, using prior guesses for the median and the interquartile range of the survival population, we obtain values $\tilde{\alpha}$ and $\tilde{\lambda}$ corresponding to the Weibull distribution that best matches these prior guesses. Of course, if the prior information is in the form of two prior moments we can use the corresponding Weibull moments and solve for $\tilde{\alpha}$ and $\tilde{\lambda}$. Finally, we specify $b_{\gamma}$ and $b_{\phi}$ by setting the medians of $[\lambda]$ and $[\alpha]$ equal to $\tilde{\lambda}$ and $\tilde{\alpha}$, respectively.

5. Data illustrations

We consider three examples to illustrate model (5). The first example is based on a simulated dataset. The other two consider the analyses of real datasets with censoring.

For all three examples, we follow the approach of section 4.3 to specify the prior hyperparameters for $\gamma$ and $\phi$. (In the absence of actual prior information for these illustrative examples, we used values roughly equal to the sample median and interquartile range as prior guesses for the population median and interquartile range, respectively.) Prior distributions of functionals indicate that the approach results in a rather noninformative specification (see section 5.1 for an illustration.) We study the effect of the prior choice for $\nu$ considering Gamma priors with varying dispersion.

(Figure 1 here)

Convergence of the MCMC algorithm, assessed through multiple chains, was fast. For
all examples, a burn-in period of at most 5,000 iterations was adequate. Mixing of the
chains was also satisfactory considering the large number of latent variables involved. For
instance, thinning of at most 150 iterations was enough to eliminate autocorrelations for
the dataset of section 5.1. All the results are based on posterior samples of size $B = 10,000$
which are used, as discussed in section 4.2, to obtain the posteriors of functionals reported
in the following sections.

5.1. A simulated dataset

We test the performance of model (5) using data generated from a mixture of lognormal
distributions, $p\text{LN}(\mu_1, \sigma_1^2) + (1-p)\text{LN}(\mu_2, \sigma_2^2)$, with $\mu_1 = 0$, $\mu_2 = 1.2$, $\sigma_1^2 = 0.25$, $\sigma_2^2 = 0.02$
and $p = 0.8$. The density function of this mixture is bimodal and the hazard function
is non-monotone with three change points in the interval $(0, 5)$ where essentially all the
probability mass lies. (See Figures 2 and 3 for the actual curves.) Here we ignore censoring
and take $n = 200$, large enough to provide a representative sample from the mixture.

(Figure 2 here)

We consider three priors for $\nu$, Gamma(2, 0.9), Gamma(2, 0.1) and Gamma(3, 0.05)
distributions, yielding increasing values for the prior mean and variance of $\nu$. The third
choice is rather extreme, postulating $a priori$ a large number of distinct components (clus-
ters) $n^*$, relative to $n$, for the mixture model. Note, for instance, that for moderately
large $n$, the expected value of $n^*$, given $\nu$ and $n$, can be approximated by $\nu \log(1 + (n/\nu))$
(Escobar & West, 1995). Figure 1 provides posteriors for $\nu$ and $n^*$. In all cases, there
is learning for $\nu$ from the data, although, under the more dispersed priors, the tail of
the posterior is affected by the prior. However, the posterior changes are certainly less
dramatic than the changes in the prior. The associated posteriors for $n^*$ indicate that
larger posterior values for $\nu$ result in higher posterior probabilities for larger values of $n^*$.
Regardless, the data support roughly 3 to 15 distinct components in the mixture model.

(Figure 3 here)

Of more interest is the effect of these prior choices on posterior inference for function-
als. Figure 2 provides results for the survival function and density function functionals.
The point estimates are based on prior and posterior means. The uncertainty bands correspond to 95% pointwise interval estimates. The shapes of the true survival and density functions are recovered quite well and, in fact, based on rather vague prior specifications. Especially noteworthy, in this regard, are the prior uncertainty bands for the survival function. Posterior inference, under the three priors for $\nu$, is very similar, the only noticeable difference being that the posterior density estimate under the Gamma(3,0.05) prior captures the second mode of the true density more successfully than the other two. In Figure 3 we plot posterior point (posterior means) and 95% pointwise interval estimates for the hazard function functional, obtained under the Gamma(2,0.1) prior for $\nu$. We also compare posterior point estimates under the three priors for $\nu$. The shape and change points of the true hazard are captured successfully. Note that there are only four observations greater than 4, the largest being equal to 4.574. This explains the considerable increase, beyond time point 4, in the width of posterior uncertainty bands.

5.2. Remission times for leukemia patients

To demonstrate comparisons of survival functionals across populations, we consider data on remission times, in weeks, for leukemia patients taken from Lawless (1982, p. 351). The study involves two treatments, A and B, each with 20 patients. The actual observations are 1, 3, 3, 6, 7, 7, 10, 12, 14, 15, 18, 19, 22, 26, 28+, 29, 34, 40, 48+, 49+ for treatment A and 1, 1, 2, 2, 3, 4, 5, 8, 8, 9, 11, 12, 14, 16, 18, 21, 27+, 31, 38+, 44 for treatment B. (A + denotes a censored observation.) Assuming proportional hazard functions, i.e., $\lambda_B(t) = \delta \lambda_A(t)$, $\delta > 0$, Lawless (1982) tests equality of the associated survival functions (i.e., $\delta = 1$), based on classical test procedures that rely on approximate normality, concluding that “there is no evidence of a difference in distributions.” Damien & Walker (2002) also use this dataset to illustrate a Bayesian nonparametric approach for comparison of two treatments. Their approach does not assume any functional relationship between the distribution functions associated with the treatments and yields a result they regard “far from conclusive of no difference”.

(Figure 4 here)
We employ model (5) for each of the underlying population distributions forcing no particular relation between the corresponding distribution functions. We have again experimented with several priors for \( \nu \). In this case, with the small sample sizes, there is almost no learning about \( \nu \) from the data. However, the effect on posterior inference for functionals is relatively minor. As an illustration, the posterior means and 95% pointwise interval estimates of the survival function for treatment A, under three priors for \( \nu \), are shown in Figure 4(a). The following results, for both treatments, correspond to the Gamma(2,0.9) prior for \( \nu \).

(Figure 5 here)

In Figure 4(b), we compare posterior means and 95% pointwise interval estimates of the survival functions for treatments A and B. Figures 4(c) and 4(d) compare posterior means for the density functions and hazard functions, respectively. Note the relatively large posterior uncertainty, depicted in the posterior interval estimates in Figure 4(b), consistent with the small sample sizes. Even under this uncertainty, there is indication for differences in the distributions of remission times, under the two treatments, e.g., with regard to center and dispersion. The point estimates for the hazard functions provide meaningful results in the light of the data, e.g., for treatment B, there is an increase in the hazard rate during roughly the first 15 weeks, followed by a decrease up to roughly 25 weeks at which point the hazard rate stabilizes. Moreover, Figure 4(d), even though it shows only point estimates, indicates that the proportional hazards assumption is suspect. To study more carefully the validity of this assumption, we obtain \( [\lambda_B(t_0)/\lambda_A(t_0) \mid D] \), the posterior of the ratio of hazard functions for treatments B and A, over a grid of \( t_0 \) values. Figure 5(a), containing the resulting posterior means and 95% interval estimates, provides evidence against the proportional hazards assumption, e.g., note the nonoverlapping posteriors for \( \lambda_B(t_1)/\lambda_A(t_1) \) and \( \lambda_B(t_2)/\lambda_A(t_2) \), for any \( t_1 \in (0,10) \) and \( t_2 \in (20,30) \). Figure 5(b) summarizes, by posterior means, 80% and 95% interval estimates, for a range of \( t_0 \) values, \( [F_B(t_0) - F_A(t_0) \mid D] \), the posterior of the difference of survival functions for treatments A and B. Finally, Figure 5(c) shows \( [\eta_A - \eta_B \mid D] \), the posterior of the difference of median remission times for treatments A and B, that yields \( P(\eta_A > \eta_B \mid D) = 0.9135 \).
Figures 5(a) and 5(b) allow assessment of local and global differences for two important functionals of the survival populations and, along with the other results, do not justify a statement as strong as that of no difference between distributions of remission times.

5.3. Survival times for patients with liver metastases

Our final example is concerned with a large dataset, given in Haupt & Mansmann (1995), involving survival times, in months, for patients with liver metastases from a colorectal primary tumor without other distant metastases. Censoring is fairly heavy, with 259 censored observations among the 622 observations. Antoniadis et al. (1999) analyzed the data providing wavelet based point estimates for the density and hazard function and also comparing with other classical curve fitting techniques.

We have again tried a few different priors for \( \nu \), in fitting model (5) to this dataset. The resulting posteriors, under two prior choices, are shown in Figure 6(a). As expected, based on the previous analyses, posterior inference for functionals is robust with respect to the choice of prior on \( \nu \). The following results correspond to the Gamma(3,0.05) prior.

(Figure 6 here)

In Figures 6(b) and 6(c), we plot posterior means and 95% pointwise interval estimates for the density function and hazard function, respectively. In general, based on point estimates, results are similar to these obtained in Antoniadis et al. (1999). There is an increasing hazard up to about 15 months whereas the hazard rate remains roughly constant between 15 and 35 months. The decrease in the hazard point estimate for time points beyond 40 months might be due to the fact that all but 2 uncensored survival times are below 40 months (26 censored observations have values greater than 40 months). Regardless, our estimates avoid the boundary effects that some of the estimates reported in Antoniadis et al. (1999) exhibit. More importantly, the uncertainty associated with the point estimates can be assessed through posterior interval estimates. Finally, Figure 6(d) displays how the posterior of the survival function functional, \( 1 - F(t_0; G) \), evolves with time, showing the posteriors at time points \( t_0 = 15, 30, 40 \) and 45 months.
6. Summary and future work

We have developed a nonparametric mixture model for survival populations based on the Weibull distribution for the kernel and the DP prior for the mixing distribution. We have shown how full posterior inference for essentially any functional of interest in survival analysis problems can be obtained. In particular, the model yields rich data-driven inference for the density function, survival function and hazard function. We have investigated the effect of prior choices on posterior inference and demonstrated that a relatively small amount of prior input suffices for the implementation of the methodology.

The MCMC algorithm, designed to fit the model, can be extended to handle left or interval censored data. Future research will study applications of the methodology under these alternative censoring schemes as well as for truncated data. Of most interest is the extension of model (5) to regression settings. Current work studies two alternative formulations, involving DP mixtures of parametric regression models. They both result in flexible semiparametric regression models that allow deviations from specific parametric assumptions, e.g., proportional hazards, when they are not supported by the data.

Acknowledgements

The author wishes to thank Alan Gelfand and Peter Müller for helpful discussions.

References


Walker, S. & Muliere, P. (1997). Beta-Stacy processes and a generalization of the Pólya-

West, M., Müller, P. & Escobar, M.D. (1994). Hierarchical priors and mixture models,
with application in regression and density estimation. In *Aspects of uncertainty: A

Athanasios Kottas, Department of Applied Mathematics and Statistics, Baskin School of
Engineering, University of California at Santa Cruz, Santa Cruz, CA 95064, USA.
E-mail: thanos@ams.ucsc.edu

**Appendix: Details for simulation-based model fitting**

Here we present the Gibbs sampler designed to fit model (5). To simplify some of the
full conditionals, we assume that all survival times are greater than 1. In fact, the trans-
formation $1 + (t_i / \max t_i)$ facilitates the implementation of the algorithm. Of course, the
posteriors of all functionals of interest can be obtained on the original scale by applying the
transformation when computing the survival function and density function functionals.

A key property is the (almost sure) discreteness of the random distribution $G$, inducing
a clustering of the $(\alpha_i, \lambda_i)$ (and thus of the $t_i$). Let $n^*$ be the number of clusters in the
vector $((\alpha_1, \lambda_1),..., (\alpha_n, \lambda_n))$ and denote by $(\alpha_j^*, \lambda_j^*)$, $j = 1,...,n^*$, the distinct $(\alpha_i, \lambda_i)$'s.
The vector of configuration indicators $s = (s_1, ..., s_n)$, defined by $s_i = j$ if and only if
$(\alpha_i, \lambda_i) = (\alpha_j^*, \lambda_j^*)$, $i = 1,...,n$, determines the clusters. Let $n_j$ be the number of mem-
ers of cluster $j$, i.e., $n_j = |\{i : s_i = j\}|$, $j = 1,...,n^*$. Gibbs sampling to draw from
$[(\alpha_1, \lambda_1),..., (\alpha_n, \lambda_n), \nu, \gamma, \phi ] \mid D$] is based on the following full conditionals:

(a) $[(\alpha_i, \lambda_i, s_i ) \mid \{(\alpha_{i'}, \lambda_{i'}, s_{i'}), i' \neq i\}, \nu, \gamma, \phi ] \mid D$, for $i = 1,...,n$

(b) $[(\alpha_j^*, \lambda_j^* ) \mid s, n^*, \gamma, \phi ] \mid D$, for $j = 1,...,n^*$
(c) \( [\nu \mid n^*, D], [\phi \mid \{(\alpha_i^*, \lambda_j^*), j = 1, \ldots, n^*\}, n^*] \) and \([\gamma \mid \{(\alpha_i^*, \lambda_j^*), j = 1, \ldots, n^*\}, n^*]\)

In these expressions we condition only on the relevant variables exploiting the conditional independence structure of the model and properties of the DP.

The full conditionals in (a) are obtained using the Pólya urn representation of the DP (Blackwell & MacQueen, 1973). We use the superscript “-” to denote all relevant quantities when the \( i^{th} \) element \((\alpha_i, \lambda_i)\) is removed from the vector \((\alpha_1, \lambda_1), \ldots, (\alpha_n, \lambda_n)\).

Hence \(n^*\) is the number of clusters in \((\alpha_{i'}, \lambda_{i'}), i' \neq i\) and \(n^*_j\) is the number of elements in cluster \( j, j = 1, \ldots, n^*\), with \((\alpha_i, \lambda_i)\) removed. Again, \((\alpha_j^*, \lambda_j^*), j = 1, \ldots, n^*\), are the distinct cluster values. Then for each \(i_o = 1, \ldots, n_o\), corresponding to an uncensored survival time \(t_{i_o}\), the full conditional in (a) is the mixed distribution

\[
q_0^0 h^0(\alpha_{i_o}, \lambda_{i_o} \mid \gamma, \phi, t_{i_o}) + \sum_{j=1}^{n^*_o} n_j^* q_j^0 \delta(\alpha_j^*, \lambda_j^*)(\alpha_{i_o}, \lambda_{i_o}) \overline{q_0^0 + \sum_{j=1}^{n^*_o} n_j^* q_j^0}
\]

(A.1)

where \(q_j^0 = kW(t_{i_o} \mid \alpha_j^*, \lambda_j^*)\) and

\[
q_0^0 = \nu \int kW(t_{i_o} \mid \alpha, \lambda) G_0(d\alpha, d\lambda) = \frac{d\nu \gamma^d}{\phi t_{i_o}} \int_0^\phi \frac{\alpha i_o^\alpha}{(\gamma + t_{i_o}^\alpha)^{d+1}} d\alpha,
\]

(A.2)

that is easy to compute using numerical integration. Moreover, \(h^0(\alpha_{i_o}, \lambda_{i_o} \mid \gamma, \phi, t_{i_o}) \propto kW(t_{i_o} \mid \alpha_{i_o}, \lambda_{i_o}) g_0(\alpha_{i_o}, \lambda_{i_o} \mid \phi, \gamma)\), where \(g_0\) is the density of \(G_0\). Simulating from the mixed distribution (A.1) is straightforward if we can draw from its continuous piece. To this end, note that the density \(h^0\) can be written as \([\alpha_{i_o} \mid \gamma, \phi, t_{i_o}] \mid \lambda_{i_o} \mid \alpha_{i_o}, \gamma, \phi, t_{i_o}\) where \([\alpha_{i_o} \mid \gamma, \phi, t_{i_o}] \propto \alpha_{i_o} t_{i_o}^{\alpha_{i_o}} 1(0 \leq \alpha_{i_o} \leq \phi) / \{(\gamma + t_{i_o}^{\alpha_{i_o}})^{d+1}\}\) and \([\lambda_{i_o} \mid \alpha_{i_o}, \gamma, \phi, t_{i_o}]\) is the density of an IGamma\(\cdot \mid d + 1, \gamma + t_{i_o}^{\alpha_{i_o}}\). Hence we can simulate from the latter distribution given the draw from the former that can be obtained by discretizing \([\alpha_{i_o} \mid \gamma, \phi, t_{i_o}]\). Note that we already have the required values of the function from the evaluation of integral (A.2).

The full conditional in (a) corresponding to a right censored survival time \(z_{i_c}\) can be developed in a similar fashion. The difference is that for \(z_{i_c}\) the contribution from the first stage of model (5) is \(1 - kW(z_{i_c} \mid \alpha_{i_c}, \lambda_{i_c})\), instead of \(kW(t_{i_c} \mid \alpha_{i_c}, \lambda_{i_c})\) for an uncensored survival time \(t_{i_c}\). Therefore the full conditional of \((\alpha_{i_c}, \lambda_{i_c}), i_c = 1, \ldots, n_c\), is of the form (A.1) with \((\alpha_{i_o}, \lambda_{i_o}), t_{i_o}, q_0^0, h^0\) and \(q_j^0\) replaced by \((\alpha_{i_c}, \lambda_{i_c}), z_{i_c}, q_0^c, h^c\) and \(q_j^c\), respectively.
Here $q_j^o = 1 - K_W(z_{i_c} \mid \alpha_j^*, \lambda_j^*)$,

$$
q_0^c = \nu \int (1 - K_W(z_{i_c} \mid \alpha, \lambda)) G_0(d\alpha, d\lambda) = \frac{\nu \gamma^d}{\phi} \int_0^\phi \frac{1}{(\gamma + z_{i_c}^{\alpha_c})^d} d\alpha, \quad (A.3)
$$

and

$$
h^c(\alpha_{i_c}, \lambda_{i_c} \mid \gamma, \phi, z_{i_c}) \propto (1 - K_W(z_{i_c} \mid \alpha_{i_c}, \lambda_{i_c})) g_0(\alpha_{i_c}, \lambda_{i_c} \mid \phi, \gamma) = [\alpha_{i_c} \mid \gamma, \phi, z_{i_c}][\lambda_{i_c} \mid \alpha_{i_c}, \gamma, \phi, z_{i_c}], \quad (A.4)
$$

where $[\alpha_{i_c} \mid \gamma, \phi, z_{i_c}] \propto 1_{(0 \leq \alpha_{i_c} \leq \phi)} / \{(\gamma + z_{i_c}^{\alpha_c})^d\}$ and $[\lambda_{i_c} \mid \alpha_{i_c}, \gamma, \phi, z_{i_c}]$ is the density of an IGamma$(\cdot \mid d, \gamma + z_{i_c}^{\alpha_c})$. We follow the same guidelines with the previous paragraph to compute the integral in (A.3) and to draw from (A.4).

Note that updating $(\alpha_i, \lambda_i)$ implicitly updates $s_i$, $i = 1, \ldots, n$. Before proceeding to update $(\alpha_{i+1}, \lambda_{i+1})$, we redefine $n^*$, $(\alpha_j^*, \lambda_j^*)$, $j = 1, \ldots, n^*$, $s_i$, $i = 1, \ldots, n$ and $n_j$, $j = 1, \ldots, n^*$, which in turn define $n^{*-}$ and $n_j^-$ after removing $(\alpha_{i+1}, \lambda_{i+1})$.

Once step (a) is completed, we have a specific configuration $s = (s_1, \ldots, s_n)$ and the associated cluster locations $(\alpha_j^*, \lambda_j^*)$, $j = 1, \ldots, n^*$. Step (b) improves the mixing of the chain by moving these cluster locations (Bush & MacEachern, 1996). Note that the group of observations corresponding to cluster $j$ might consist of both uncensored and censored survival times. In this most general case, for each $j = 1, \ldots, n^*$, $[(\alpha_j^*, \lambda_j^*) \mid s, n^*, \gamma, \phi, D]$ is proportional to

$$
g_0(\alpha_j^*, \lambda_j^* \mid \phi, \gamma) \prod_{\{i_o: s_{i_o} = j\}} k_W(t_{i_o} \mid \alpha_j^*, \lambda_j^*) \prod_{\{i_c: s_{i_c} = j\}} (1 - K_W(z_{i_c} \mid \alpha_j^*, \lambda_j^*)), \quad (A.5)
$$

resulting in the unnormalized density

$$
\alpha_j^{n_j^o} 1_{(0 \leq \alpha_j \leq \phi)} \lambda_j^{-(d+n_j^o+1)} C(\alpha_j^*, D) \exp \left\{ -\lambda_j^{n_j^o-1} \left( \gamma + \sum_{\{i_o: s_{i_o} = j\}} t_{i_o}^{\alpha_j^*} + \sum_{\{i_c: s_{i_c} = j\}} z_{i_c}^{\alpha_j^*} \right) \right\},
$$

where $n_j^o = |\{i_o: s_{i_o} = j\}|$ and $C(\alpha_j^*, D) = \left( \prod_{\{i_o: s_{i_o} = j\}} t_{i_o}^{\alpha_j^*} \right)$. To generate from this distribution, we extend the Gibbs sampler to draw from the full conditional densities $[\lambda_j^* \mid \alpha_j^*, s, n^*, \gamma, \phi, D]$ and $[\alpha_j^* \mid \lambda_j^*, s, n^*, \gamma, \phi, D]$. The former is the density of an IGamma$(\cdot \mid d + n_j^o, \gamma + \sum_{\{i_o: s_{i_o} = j\}} t_{i_o}^{\alpha_j^*} + \sum_{\{i_c: s_{i_c} = j\}} z_{i_c}^{\alpha_j^*})$ whereas the latter has an awkward form rendering random generations difficult. We overcome this difficulty employing slice sampling (see, e.g., Damien et al., 1999). Specifically, consider auxiliary variables $v, w^o$
= \{w_{i_o}^o, i_o : s_{i_o} = j\}, \ w^c = \{w_{i_c}^c, i_c : s_{i_c} = j\}, \text{ taking positive values, such that the joint distribution } [\alpha_j^*, v, w^o, w^c | \lambda_j^*, s, n^*, \gamma, \phi, D] \text{ is proportional to }

\alpha_j^{*j} 1(0<\alpha_j^* \leq \phi) 1(0<v < C(\alpha_j^*, D)) \prod_{\{i_o: s_{i_o} = j\}} 1(0<w_{i_o}^o < \exp(-\lambda_j^{-1} t_{i_o}^o)) \prod_{\{i_c: s_{i_c} = j\}} 1(0<w_{i_c}^c < \exp(-\lambda_j^{-1} z_{i_c}^j))

(A.6)

yielding \([\alpha_j^* | \lambda_j^*, s, n^*, \gamma, \phi, D]\) upon marginalization over the auxiliary variables. Now the full conditionals, corresponding to (A.6), are all standard. In particular, they are uniform on \((0, C(\alpha_j^*, D))\), \((0, \exp(-\lambda_j^{-1} t_{i_o}^o))\) and \((0, \exp(-\lambda_j^{-1} z_{i_c}^j))\) for \(v, w_{i_o}^o, i_o \in \{i_o: s_{i_o} = j\}\) and \(w_{i_c}^c, i_c \in \{i_c: s_{i_c} = j\}\), respectively. Moreover, the full conditional for \(\alpha_j^*\) is proportional to \(\alpha_j^{*j} 1(\alpha_j^* < \gamma < \infty)\), where \(\overline{\alpha}_j^* = \max \left\{0, \left(\sum_{\{i_o: s_{i_o} = j\}} \log t_{i_o}\right)^{-1} \log v\right\}\) and

\[\overline{\alpha}_j^* = \min \left\{\phi, \min_{\{i_o: s_{i_o} = j\}} \left(\frac{\log(-\lambda_j^* \log w_{i_o}^o)}{\log t_{i_o}}\right), \min_{\{i_c: s_{i_c} = j\}} \left(\frac{\log(-\lambda_j^* \log w_{i_c}^c)}{\log z_{i_c}}\right)\right\}.

Drawing from this distribution is straightforward by using the inverse c.d.f. method.

For clusters that relate only to uncensored or only to censored survival times the third or second term, respectively, in (A.5) is absent. The full conditional for \(\lambda_j^*\) is, again, an inverse Gamma distribution with appropriately adjusted parameters. Sampling from the full conditional of \(\alpha_j^*\) is, again, facilitated by the introduction of auxiliary variables.

Turning to step (c), we update \(\nu\) using the augmentation method given in Escobar & West (1995). Briefly, an auxiliary variable \(u\) is introduced such that the joint density of \(\nu\) and \(u\) has full conditionals \([u | \nu, D] = \text{Beta}(\nu + 1, n)\) and \([\nu | u, n^*, D] = p\text{Gamma}(a_\nu + n^*, b_\nu - \log(u)) + (1 - p)\text{Gamma}(a_\nu + n^* - 1, b_\nu - \log(u))\), where \(p = (a_\nu + n^* - 1) / \{n(b_\nu - \log(u)) + a_\nu + n^* - 1\}\).

Finally, the full conditionals for \(\phi\) and \(\gamma\) are proportional to \([\phi] \prod_{j=1}^{n^*} \phi^{-1} 1(\phi \geq \alpha_j^*)\) and \([\gamma] \prod_{j=1}^{n^*} \gamma^d \exp(-\lambda_j^{-1} \gamma), \text{ respectively, where } [\phi] \text{ and } [\gamma] \text{ are their prior densities from model (5). Hence the last two densities in (c) correspond to a Pareto(\cdot | a_\phi + n^*, \max\{1 \leq j \leq n^*\} \alpha_j^*)\) distribution for \(\phi\) and a Gamma(\cdot | a_\gamma + dn^*, b_\gamma + \sum_{j=1}^{n^*} \lambda_j^{-1})\) distribution for \(\gamma\).
Figure 1: For the simulated data, histograms of posterior draws for $\nu$ and $n^*$, under three prior choices for $\nu$. The prior densities for $\nu$ are denoted by the solid lines.
Figure 2: Inference, under three prior choices for $\nu$, for the simulated data. The upper panels provide prior (dotted lines) and posterior (dashed lines) point and interval estimates for the survival function functional. The lower panels include the histogram of the data along with the posterior point estimate (dashed line) for the density function functional. In each graph, the solid line denotes the true curve.
Figure 3: For the simulated data, posterior inference for the hazard function functional. Under the $\text{Gamma}(2,0.1)$ prior for $\nu$, the left panel provides point and interval estimates (dashed lines). The right panel compares point estimates under the three priors for $\nu$, $\text{Gamma}(2,0.9)$ (smaller dashed line), $\text{Gamma}(2,0.1)$ (dashed line) and $\text{Gamma}(3,0.05)$ (dotted line). In each graph, the solid line denotes the true hazard function.
Figure 4: Data on remission times for leukemia patients. (a) Posterior point and interval estimates of the survival function for treatment A, under a Gamma(3,0.3) (dotted lines), a Gamma(2,0.9) (solid lines) and a Gamma(2,5) (dashed lines) prior for $\nu$. Under the Gamma(2,0.9) prior for $\nu$, Figures 4(b), 4(c) and 4(d) compare the survival functions (point and interval estimates), density functions and hazard functions (point estimates), respectively, for treatments A (solid lines) and B (dashed lines).
Figure 5: Data on remission times for leukemia patients. (a) Posterior point estimate (solid line) and 95% interval estimates (dashed lines) for $[\lambda_B(t_0)/\lambda_A(t_0) \mid D]$. (b) Posterior point estimate (solid line), 80% interval estimates (dotted lines) and 95% interval estimates (dashed lines) for $[F_B(t_0) - F_A(t_0) \mid D]$. (c) Histogram of draws from $[\eta_A - \eta_B \mid D]$. 
Figure 6: Liver metastases dataset. (a) Posteriors (dashed and solid line, respectively) for $\nu$ under Gamma(2,0.1) (dotted line) and Gamma(3,0.05) (smaller dashed line) priors. (b) Posterior point estimate (solid line) and interval estimates (dashed lines) for the density function. (c) Posterior point estimate (solid line) and interval estimates (dashed lines) for the hazard function. (d) Posteriors for the survival function functional at 15, 30, 40 and 45 months (dashed, dotted, solid and smaller dashed lines, respectively).