Bayesian Analysis (2006) **1**, Number 3, pp. 575–594

Modeling Censored Lifetime Data Using a Mixture of Gammas Baseline

Timothy E. Hanson*

Abstract. We propose a Bayesian semiparametric accelerated failure time (AFT) model in which the baseline survival distribution is modeled as a Dirichlet process mixture of gamma densities. The model is highly flexible and readily captures features such as multimodality in predictive survival densities. The approach can be used in a "black-box" manner in that the prior information needed to fit the model can be quite vague, and we recommend a particular prior in the absence of information on the baseline survival distribution. The resulting posterior baseline distribution has mass only on the positive reals, a desirable feature in a failure-time model. The formulae needed to fit the model are available in closed-form and the model is relatively easy to code and implement. We provide both simulated and real data examples, including data on the cosmetic effects of cancer therapy.

Keywords: Accelerated failure time, Dirichlet process mixture.

1 Introduction

Although fruitful, traditional parametric approaches to the accelerated failure time model suffer from a marked lack of flexibility. Most statistical packages on the market today require one to choose from a set of three or four basic baseline models quite similar in shape and character; for example, log-logistic and generalized gamma regression models restrict predictive survival densities to be right-skewed and unimodal. Parametric models may be enriched by considering mixtures of parametric densities for the baseline distribution.

In this paper, we consider a Dirichlet process mixture of gamma densities for the baseline in the accelerated failure time model. A mixture of gammas can provide a highly flexible baseline model, allowing for multiple modes. Wiper, Rios Insua, and Ruggeri (2001) note that any continuous density $f(x)$ on $\mathbb{R}^+$ such that $\lim_{x\to\infty} f(x) = 0$ can be approximated arbitrarily closely by a countable weighted sum of gamma densities.

Although Dirichlet process mixing in the context of density estimation has received much attention in the literature (Lo, 1984; Escobar and West, 1995), little research has been done regarding density estimation on the positive reals, suitable for failure time modeling. A notable exception is the work of Kuo and Mallick (1997), who propose two classes of Dirichlet process mixture models in the accelerated failure time (AFT) setting. In this paper, we consider their "MDPV" class based on gamma densities. In examples, Kuo and Mallick consider a location mixture of normal kernels with an exponential base-measure. This approach requires some fine-tuning in the selection of the kernel

[*] Division of Biostatistics, School of Public Health,University of Minnesota, MN, http://www.stat.umn.edu/~hanson

variance since the variance must be small enough to avoid placing significant mass on the negative reals. Thus, in order to achieve a parsimonious fit to data, the number of distinct components in the mixture needs to be relatively large. We compliment the work of Kuo and Mallick by considering a family of densities that have support only on the positive reals. The kernel is not used to smooth the Dirichlet process; rather, the baseline survival distribution is modeled as a weighted sum of possibly quite distinct densities. The prior on the gamma component shape and scale parameters is somewhat vague, and the expected number of components is a slowly growing function of the sample size when the Dirichlet process precision is set to unity, diminishing the possibility of overfitting.

Semiparametric approaches to the AFT model date at least to the initial work of Buckley and James (1979) in the frequentist realm and Christensen and Johnson (1988) in the Bayesian realm. More recent frequentist approaches include those by Ying, Jung, and Wei (1995) and Yang (1999). All four of these approaches are essentially fitting techniques focused on the estimation of regression effects. In fact, although the latter two papers include the analysis of failure time data, there are no predictive survival curves or densities nor mention of how one might obtain these very common loci of inference. Other recent Bayesian approaches include the work of Kuo and Mallick (1997), Walker and Mallick (1999), Kottas and Gelfand (2001), and Hanson and Johnson (2002). Walker and Mallick (1999) and Hanson and Johnson (2002) propose a Polya tree and a mixture of Polya trees, respectively, as priors on the log-baseline survival distribution in the AFT model. These approaches work quite well on data they present, although the baseline is centered around a single distribution or family of parametric distributions for the respective models. Kottas and Gelfand (2001) present Dirichlet mixtures of split, skewed unimodal densities. Although very useful for a skewed baseline, posterior predictive densities are also necessarily unimodal and prior specification requires special consideration.

Recently, Ghosh and Ghosal (2004) proposed a "proportional means" model where the baseline model is a Dirichlet process scale mixture of Weibull distributions. The model is fit using a truncated approximation to the Dirichlet process presented by Ishwaran and Zarepour (2002), and computation is carried out using WinBUGS software (Spiegelhalter et al., 2003). In the model we consider here, we mix over both the scale *and* the shape of the gamma distribution, and augment the model to allow learning about the Dirichlet process hyperparameters from the data. Mixing over both the shape and the scale provides a flexible model allowing for the possibility of quite different gamma component shapes. In related work, Kottas (2005) considers a Dirichlet process mixture of Weibull distributions for censored survival data without covariates. The Dirichlet process mixture that Kottas considers mixes over both the Weibull shape and scale, yielding a highly flexible model that performs well in simulations and on real data.

An alternative mixture of gamma components model is explored by Wiper, Rios Insua, and Ruggeri (2001). Wiper et al. place prior probabilities on the number of components in the mixture, and conditional on the number of components, place priors on the remaining parameters in each mixture model. The reversible jump approach of Green (1995) and Richardson and Green (1997) is used to obtain inference. The proposed prior on the gamma components includes a fixed exponential distribution on the shape parameter and a fixed inverse-gamma distribution on the mean of each gamma component. The parameters for the exponential scale and inverse-gamma mean priors are presumedly picked using some prior knowledge on the data generating mechanism. A problem with incorporating this fixed prior in the accelerated failure time setting is that often the baseline distribution is not interpretable, or prior information is simply not available. Thus, it is desirable to allow the baseline model to accommodate a very wide variety of shapes and spreads, as does the hierarchical baseline model we consider in this paper.

Predictive inference from the Dirichlet process mixture model can be quite similar to model averaged inference obtained through reversible jump. The reversible jump approach allows for very precise prior information on the number of components in the mixture. However, a Dirichlet process mixture does allow control over the expected number of distinct components (e.g. see Section 2.2 and Antoniak, 1974). Additionally, in practice many researchers simply take the mixture weights, conditional on the number of components d, to have a Dirichlet($\delta \mathbf{1}_d$) distribution, yielding a Dirichlet/multinomial allocation model. This model can be seen as an approximation to a Dirichlet process mixture as $\delta \rightarrow 0^+$, conditional on components being non-empty (Green and Richardson, 2001). We view the significant additional programming involved in setting up the reversible jump algorithm as unnecessary as we focus on models with a relatively small number of mixture components, ably fit via Dirichlet process mixtures. We thus also avoid potential problems with convergence assessment associated with reversible jump, which can be problematic as there are two types of mixing to consider: convergence of parameters within a component model in the product space, and mixing among the component models via reversible jump (Brooks and Giudici, 2000). In the model we present, the full conditional distributions involved in Gibbs sampling have closed form and coding is relatively painless.

Section 2 outlines a Dirichlet process mixture of gamma densities model and offers implementation guidelines that have worked well in practice. Section 3 provides several simulated and real examples using this model, and we summarize our conclusions and future research in Section 4.

2 The Model

Modeling a density as a Dirichlet process mixture of continuous kernel densities (Lo, 1984; Escobar and West, 1995) is easily extended to allow covariates in the AFT setting. We consider typical event-time data. The event-time T_i is to be stochastically modeled as a function of the p-dimensional covariate vector $\mathbf{x}_i$ and may be censored to lie in the interval $[a_i, b_i)$, where b_i is taken to be ∞ for a right-censored observation. In the exposition that follows we assume the data, $\mathbf{T} = (T_1, \ldots, T_n)'$, are uncensored, but we describe how to include censored data in Section 2.2.

2.1 The general model

Kuo and Mallick (1997) propose modeling the baseline survival distribution in the AFT model as a Dirichlet process mixture of smooth kernel densities. Specifically, their model is written hierarchically as follows:

$$\begin{aligned} T_i = \exp(-\mathbf{x}_i'\boldsymbol{\beta})V_i, \quad V_i|\boldsymbol{\theta}_i \overset{ind.}{\sim} k(v|\boldsymbol{\theta}_i) \quad \text{for } i = 1, \ldots, n, \\ \boldsymbol{\beta} \sim f(\boldsymbol{\beta}), \quad \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n|G \overset{iid}{\sim} G, \quad G|\alpha, \boldsymbol{\eta} \sim DP(\alpha G_{\boldsymbol{\eta}}), \end{aligned} \tag{B.1}$$

where $k(v|\boldsymbol{\theta})$ is a continuous density in v given $\boldsymbol{\theta}$ and $G_{\boldsymbol{\eta}}$ is a specified parametric probability measure having a density $g(\cdot|\boldsymbol{\eta})$. A complete treatment of the Dirichlet process may be found in Ferguson (1973). For k-dimensional $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n|G \overset{iid}{\sim} G$, the Dirichlet process prior on the distribution G – denoted $G|\alpha, \boldsymbol{\eta} \sim DP(\alpha G_{\boldsymbol{\eta}})$ – is a prior on the set of discrete probability distributions on $\mathbb{R}^k$. The distribution $G_{\boldsymbol{\eta}}$ centers the process in the sense that for any measurable A, $E\{G(A)\} = G_{\boldsymbol{\eta}}(A)$. The parameter $\alpha > 0$ controls how closely the process G is to $G_{\boldsymbol{\eta}}$: $Var\{G(A)\} = G_{\boldsymbol{\eta}}(A)[1 - G_{\boldsymbol{\eta}}(A)]/(\alpha + 1)$.

Kuo and Mallick (1997) essentially use this model for smoothing the Dirichlet process with a known, continuous kernel, and therefore take a kernel $k(v|\boldsymbol{\theta})$ with small variance in examples. In the sequel, we emphasize a baseline model that is a weighted sum mixture of possibly quite distinct densities on the positive reals.

Let $a_{\boldsymbol{\eta}}(v) \overset{def}{=} \int k(v|\boldsymbol{\theta})g(\boldsymbol{\theta}|\boldsymbol{\eta})d\boldsymbol{\theta}$ and note that given T_i and $\boldsymbol{\beta}$ we have $V_i = \exp(\mathbf{x}_i'\boldsymbol{\beta})T_i$. Using results from Escobar (1994), the full conditional distribution of the ith element of $\boldsymbol{\theta}$ is

$$\boldsymbol{\theta}_i|\boldsymbol{\theta}^{(i)}, \mathbf{V}, \boldsymbol{\eta} \left\{ \begin{array}{ll} \sim \frac{k(V_i|\boldsymbol{\theta})g(\boldsymbol{\theta}|\boldsymbol{\eta})}{a_{\boldsymbol{\eta}}(V_i)} & \text{w/ prob. } \frac{\alpha a_{\boldsymbol{\eta}}(V_i)}{\alpha a_{\boldsymbol{\eta}}(V_i) + \sum_{j \neq i} k(V_i|\boldsymbol{\theta}_j)} \\ = \boldsymbol{\theta}_j, \ j \neq i & \text{w/ prob. } \frac{k(V_i|\boldsymbol{\theta}_j)}{\alpha a_{\boldsymbol{\eta}}(V_i) + \sum_{j \neq i} k(V_i|\boldsymbol{\theta}_j)} \end{array} \right\}, \tag{B.2}$$

where for an arbitrary vector $\mathbf{y}$, the vector $\mathbf{y}^{(i)}$ denotes the vector $\mathbf{y}$ with the i^{th} component removed. For conjugate families, the parametric density $g(\cdot|\boldsymbol{\eta})$ can be chosen such that $a_{\boldsymbol{\eta}}(v) = \int k(v|\boldsymbol{\theta})g(\boldsymbol{\theta}|\boldsymbol{\eta})d\boldsymbol{\theta}$ is relatively easy to compute. Non-conjugate models, however, are possible to work with in this context; MacEachern and Müller (1998) and Neal (2000) describe how to accommodate them. Kuo and Mallick (1997) derive the density for $\boldsymbol{\beta}|\{\boldsymbol{\theta}_i\}, \mathbf{T}$, given by

$$f(\boldsymbol{\beta}|\{\boldsymbol{\theta}_i\}, \mathbf{T}) \propto f(\boldsymbol{\beta}) \prod_{i=1}^{n} \exp(\mathbf{x}_i'\boldsymbol{\beta})k(T_i \exp(\mathbf{x}_i'\boldsymbol{\beta})|\boldsymbol{\theta}_i), \tag{B.3}$$

which is readily sampled via one or more Metropolis-Hastings steps or through componentwise slice sampling (Neal, 2003).

The basic model (B.1) of Kuo and Mallick (1997) may be extended by further taking $\boldsymbol{\eta} \sim f(\boldsymbol{\eta})$, inducing a mixture of Dirichlet processes prior on G; this allows greater flexibility in modeling the mixing distribution of the baseline survival function. The baseline model then becomes:

$$V_1,\ldots,V_n|G \overset{iid}{\sim} \int k(v|\boldsymbol{\theta})G(d\boldsymbol{\theta}), \qquad G|\alpha,\boldsymbol{\eta} \sim DP(\alpha G_{\boldsymbol{\eta}}), \qquad \boldsymbol{\eta} \sim f(\boldsymbol{\eta}). \tag{B.4}$$

Sampling from the full conditional for the hyperparameters $\boldsymbol{\eta}$ is accomplished using Lemma 1 from Antoniak (1974):

$$f(\boldsymbol{\eta}|\{\boldsymbol{\theta}_j\}) \propto f(\boldsymbol{\eta}) \prod_{i=1}^{d\{\boldsymbol{\theta}_j\}} g(\boldsymbol{\theta}_i^*|\boldsymbol{\eta}), \tag{B.5}$$

where $\{\boldsymbol{\theta}_i^*\}_{i=1}^{d\{\boldsymbol{\theta}_j\}}$ is the set of $d\{\boldsymbol{\theta}_j\}$ distinct values of $\{\boldsymbol{\theta}_j\}$.

A Markov chain is set up and a dependent random sample from the posterior $\{(\boldsymbol{\beta}^j,\boldsymbol{\theta}^j,\boldsymbol{\eta}^j|\mathbf{T})\}_{j=1}^J$ is collected.

2.2 Mixture of gamma densities baseline

As a possible Dirichlet process mixture model, one may consider as a baseline the full mixture presented in Escobar and West (1995). However, as in the examples presented in Kuo and Mallick (1997), this model places mass on the negative reals. A kernel that places mass only on $\mathbb{R}^+$ may be better suited for modeling survival data and provide more parsimonious fit with fewer mixture components. In this paper, we examine the mixture of gammas model $\boldsymbol{\theta}_i = (\lambda_i,\gamma_i)$, $V_i|\{\lambda_i,\gamma_i\} \sim \Gamma(\lambda_i,\gamma_i)$. To allow for quite different gamma components in the mixture, the parametric base-measure of the Dirichlet process $G_{\boldsymbol{\eta}}$ should be somewhat dispersed. We take $\boldsymbol{\eta} = (a_\lambda, a_\gamma)$ and

$$g(\lambda,\gamma|\boldsymbol{\eta}) = a_\lambda \exp(-a_\lambda\lambda) a_\gamma \exp(-a_\gamma\gamma).$$

That is, under $G_{\boldsymbol{\eta}}$, λ and γ are distributed as independent exponential random variables. We further take flexible gamma priors for the hyperparameters; $a_\lambda \sim \Gamma(b_\lambda, c_\lambda)$ independent of $a_\gamma \sim \Gamma(b_\gamma, c_\gamma)$, where in examples we fix $b_\lambda = c_\lambda = b_\gamma = c_\gamma = 0.0001$.

Although assuming independent exponential distributions is computationally convenient, one might expect some dependence *a priori* between λ_i and γ_i. Certainly this dependence is expressed in the posterior of $E(G)$, but it may be useful to allow some prior dependence in G as well. A referee has suggested that the bivariate Dirichlet process of Walker and Muliere (2003) may prove useful in this regard.

Copsey and Webb (2003) consider a similar mixture but rather assume λ follows a Poisson distribution shifted by unity – $\lambda \sim$ Poisson(ϕ)+1, independent of $\gamma \sim \Gamma(\nu,\xi)$. In their example, $\phi = 2$, ξ is a function of training data, and $\nu = 1.001$, so γ is approximately exponential as in the model we consider here.

The parameter α, along with the sample size n, affects the expected number of gamma components $E(d\{\boldsymbol{\theta}_j\}) = E(d)$. Table 1 in Escobar (1994) gives $E(d)$ for various values of $\alpha = \alpha(n)$ and samples sizes n. In the absence of prior information on the number of gamma components needed to adequately model the baseline, a default value might be $\alpha = 1$. Using results from Antoniak (1974) and Escobar (1994), when $\alpha = 1$, the prior expected number of gamma components for a sample size of $n = 20$ is 4. The prior expected number of components for n equal to 100, 200, and 1000 is 5, 6, and 7 components, respectively. The prior expected number of components grows with n, but slowly, and the possibility of overfitting is diminished as compared to, for example, $\alpha = \sqrt{n}$. When $\alpha = \sqrt{n}$, $n = 20$ implies $E(d) \approx 8$ and $n = 100$ implies $E(d) \approx 24$. A finite mixture model for $n = 20$ with 8 gamma components would not be identifiable. A finite mixture model with 24 components seems excessive under most circumstances.

As an alternative to fixing α at a pre-specified value, Escobar and West (1995) describe a useful data augmentation trick allowing for the prior $\alpha \sim \Gamma(a_\alpha, b_\alpha)$ if desired. For every paper published using a Dirichlet process mixture at some level of model hierarchy there is a different gamma prior or set of priors considered for α; a sampling includes $\Gamma(1,1)$, $\Gamma(1,0.2)$, $\Gamma(1,0.005)$, $\Gamma(5,0.5)$, $\Gamma(2,4)$, $\Gamma(3,0.005)$, $\Gamma(2,0.1)$, and many more. The general argument is that the prior should "support large and small values" of α, possibly appealing to Escobar (1994), where a uniform prior on $\log_n \alpha$ over the range (n^{-1}, n^2) was suggested, yielding the prior $p(\alpha) \propto 1/\alpha$ on the interval (n^{-1}, n^2). Escobar (1994) actually considered a discrete prior placing mass 0.25 on the values $\{n^{-1}, 1, n, n^2\}$, allowing the prior $E(d)$ to range from one to n, with either extreme value equally likely. The priors $\Gamma(2,2)$ and $\Gamma(2,0.5)$ are considered in Section 3.2 along with fixing $\alpha = 0.1$ and $\alpha = 1$.

We derive the induced prior on the mean and variance of a gamma component. Where $V \sim \Gamma(\lambda, \gamma)$, let $\mu = \lambda/\gamma$ and $\sigma^2 = \lambda/\gamma^2$; then, given a_λ and a_γ, the induced prior on μ is $f(\mu|a_\lambda, a_\gamma) = a_\lambda a_\gamma (a_\gamma + a_\lambda \mu)^{-2}$ for $\mu > 0$. Given μ, a_λ, and a_γ, the precision σ^{-2} is distributed $\Gamma(2, a_\lambda \mu^2 + a_\gamma \mu)$. Note that the induced density for μ is monotone decreasing and can be very diffuse on the positive reals. The larger μ is, the smaller the precision (and hence the larger the variance) is expected to be. Under the prior, a component close to zero will typically have a small variance whereas a component far from zero will have a relatively large variance.

Under the model of Wiper et al. (2001), the prior on the mean is specified $\mu^{-1} \sim \Gamma(a, b)$ and the conditional precision can be shown to be distributed $\sigma^{-2}|\mu \sim \exp(\theta\mu^2)$, where a, b, and θ are fixed. Note then that, roughly similar to the prior we suggest, the conditional expected precision $E(\sigma^{-2}|\mu)$ decreases quadratically with increasing μ. Wiper et al. (2001) suggest that they might typically take $\theta = 0.01$ and $a = b = 1$, and further note that this implies the prior moments of μ do not exist. In simulations, Wiper et al. find that the model fits data admirably. Probably in practice, one would take a, b, and θ to somehow reflect prior belief in the location and spread of the gamma components. Alternatively, analogous to the normal component model of Richardson and Green (1997) and the model we consider here, perhaps one or more of these parameters might be assumed to arise *iid* from a hyperprior distribution. Regardless, we see that the priors behave somewhat similarly, with perhaps the major difference being that the values of θ, a, and b are fixed in the Wiper et al. prior, whereas the parameters a_λ and a_γ are random in the model described above, and can be learned about from the data.

Note that in the model we propose, the marginal priors for λ and γ are $p(\lambda|b_\lambda, c_\lambda) = b_\lambda c_\lambda^{b_\lambda}/(\lambda + c_\lambda)^{b_\lambda+1}$ independent of $p(\gamma|b_\gamma, c_\gamma) = b_\gamma c_\gamma^{b_\gamma}/(\gamma + c_\gamma)^{b_\gamma+1}$. These densities are similar to Pareto pdfs and quantities of interest are readily computed. For example $E(\lambda|b_\lambda, c_\lambda) = c_\lambda/(b_\lambda - 1)$ when $b_\lambda > 1$, the cdf is $F(t|b_\lambda, c_\lambda) = 1 - [1 + t/c_\lambda]^{-b_\lambda}$, the quantile function is $F^{-1}(p|b_\lambda, c_\lambda) = c_\lambda[(1-p)^{-1/b_\lambda} - 1]$, et cetera. Like Pareto pdfs, these marginal densities are strictly decreasing and convex, as are the conditional exponential densities. When b_λ, c_λ, b_γ, and c_γ are quite small, the (marginal) prior is approximately $p(\lambda, \gamma) \propto (\lambda\gamma)^{-1}$. However, an informative prior for the gamma components can be specified through appropriate choice of b_λ, c_λ, b_γ, and c_γ. An empirical Bayes approach would estimate $\tilde{b}_\lambda$, $\tilde{c}_\lambda$, $\tilde{b}_\gamma$, and $\tilde{c}_\gamma$ from an initial fit fixing $b_\lambda = c_\lambda = b_\gamma = c_\gamma = \epsilon$, where ϵ is some small value.

Kottas (2005) describes a related model for survival estimation without covariates. Briefly, the Weibull kernel $k(t|\lambda, \gamma) = \lambda\gamma t^{\lambda-1}\exp(-\gamma t^\lambda)$ is used in the Dirichlet process mixture, with the baseline model $\gamma|\tau \sim \Gamma(2, \tau)$ independent of $\lambda|\phi \sim \text{Uniform}(0, \phi)$. The shape and scale parameters have independent distributions and the hyperpriors are further specified $\tau \sim \exp(a_\tau)$ independent of $\phi \sim \text{Pareto}(2, b_\phi)$. Kottas suggests choosing a_τ and b_ϕ to match a prior marginal median and interquartile range for one Weibull component, obtained when $\alpha \to 0^+$.

It is straightforward, but tedious, to find the necessary formulae to implement the Gibbs sampler in the model proposed in this paper. Where $k(v|\lambda, \gamma)$ denotes a $\Gamma(\lambda, \gamma)$ density, the algorithm uses the following results:

Result 1 $$\int k(v|\lambda, \gamma)P(d\lambda, d\gamma|a_\lambda, a_\gamma) = \frac{a_\lambda a_\gamma}{v(v + a_\gamma)[a_\lambda - \log(v/(v + a_\gamma))]^2}.$$

Result 2 *The conditional distribution of the random vector (λ, γ) given (V, a_λ, a_γ), with density $k(V|\lambda, \gamma)g(\lambda, \gamma|\boldsymbol{\eta})/a_{\boldsymbol{\eta}}(V)$, can be sampled $\lambda|V, a_\lambda, a_\gamma \sim \Gamma(2, a_\lambda - \log[V/(V + a_\gamma)])$, then $\gamma|\lambda, V, a_\lambda, a_\gamma \sim \Gamma(\lambda + 1, V + a_\gamma)$.*

Result 3 *Let $(\lambda_i^*, \gamma_i^*), i = 1, \ldots, d\{\boldsymbol{\theta}_j\}$ denote the $d\{\boldsymbol{\theta}_j\}$ distinct values of $\{\boldsymbol{\theta}_j\}$. Then $a_\lambda|\boldsymbol{\theta} \sim \Gamma(b_\lambda + d\{\boldsymbol{\theta}_j\}, c_\lambda + \sum_{i=1}^{d\{\boldsymbol{\theta}_j\}} \lambda_i^*)$ independent of $a_\gamma|\boldsymbol{\theta} \sim \Gamma(b_\gamma + d\{\boldsymbol{\theta}_j\}, c_\gamma + \sum_{i=1}^{d\{\boldsymbol{\theta}_j\}} \gamma_i^*)$.*

Results 1 and 2 enable the use of an unusually simple MCMC sampling scheme for obtaining inference. Let $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_n)'$ and $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_n)'$. Using these results, the simplest algorithm, corresponding to Algorithm 1 in Neal (2000), for obtaining inference follows:

Algorithm 1 *We set up the Markov chain $\{\boldsymbol{X}^j\}_{j\geq1}$ with stationary distribution π, where $\pi(A) = P((\boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\gamma}, a_\lambda, a_\gamma) \in A|\mathbf{T})$. Initialize the initial state $\mathbf{X}^0 = (\boldsymbol{\beta}^0, \boldsymbol{\lambda}^0, \boldsymbol{\gamma}^0, a_\lambda^0, a_\gamma^0)$. Where $\boldsymbol{\theta}_i = (\lambda_i, \gamma_i)$, the j^{th} state is generated:*

1. *Let $\boldsymbol{\theta}_i = \boldsymbol{\theta}_i^{j-1}$ and $V_i^{j-1} = T_i \exp(\mathbf{x}_i'\boldsymbol{\beta}^{j-1})$, $i = 1, \ldots, n$. In turn, generate $\boldsymbol{\theta}_i|\boldsymbol{\theta}^{(i)}, \mathbf{V}^{j-1}, \boldsymbol{\eta}^{j-1}$, $i = 1, \ldots, n$, according to (B.2). Set $\boldsymbol{\theta}_i^j = \boldsymbol{\theta}_i$, $i = 1, \ldots, n$.*

Result 1 facilitates the computation of the probabilities in (2); Result 2 allows for sampling $k(V_i|\boldsymbol{\theta})g(\boldsymbol{\theta}|\boldsymbol{\eta})/a_{\boldsymbol{\eta}}(V_i)$ in (2).

2. *Repeat one or more times:*

   (a) *Sample $\boldsymbol{\beta}^* \sim q(\boldsymbol{\beta}^*|\boldsymbol{\beta}^{j-1})$ where q is a symmetric candidate generating kernel: $q(\mathbf{x}|\boldsymbol{y}) = q(\boldsymbol{y}|\mathbf{x})$ for all $\mathbf{x}, \boldsymbol{y} \in \mathbb{R}^p$. Set $V_i^* = T_i \exp(\mathbf{x}_i'\boldsymbol{\beta}^*)$, $i = 1, \ldots, n$.*

   (b) *Using (B.3), form $\rho(\boldsymbol{\beta}^*, \boldsymbol{\beta}^{j-1}) = \min\left\{1, \dfrac{f(\boldsymbol{\beta}^*|\lambda^j, \gamma^j, \mathbf{T})}{f(\boldsymbol{\beta}^{j-1}|\lambda^j, \gamma^j, \mathbf{T})}\right\}$.*

   (c) *Take $\boldsymbol{\beta}^j = \boldsymbol{\beta}^*$ with prob. $\rho(\boldsymbol{\beta}^*, \boldsymbol{\beta}^{j-1})$ or else take $\boldsymbol{\beta}^j = \boldsymbol{\beta}^{j-1}$.*

3. *Sample $\boldsymbol{\eta}^j$ using (B.5), which reduces to Result 3.*

In practice, Algorithm 2 in Neal (2000) can significantly improve the convergence properties of the Markov chain with some additional coding. Depending on one's tolerance for programming, the additional coding may be preferred over the simpler algorithm above. The examples in this paper use the above algorithm and convergence has not been an issue. However, the MCMC sample sizes used in Section 3, conservatively ranging from 5,000 to 1,000,000, could certainly be reduced, so the improved algorithm is outlined for completeness.

Essentially, the improved algorithm (Bush and MacEachern, 1996) employs step 1 in the above algorithm to sample a configuration of ties in $\{\boldsymbol{\theta}_j\}$. This configuration is represented by n variables c_i, $i = 1, \ldots, n$, which indicate which of the distinct $\{\boldsymbol{\theta}_j^*\}$ that $\boldsymbol{\theta}_i$ is equal to. Conditional on a configuration $\mathbf{c} = (c_1, \ldots, c_n)'$, $\boldsymbol{\theta}_j^* = (\lambda_j^*, \gamma_j^*)$ is drawn from the density

$$p_{\lambda_j^*, \gamma_j^*}(\lambda, \gamma|\mathbf{V}, \mathbf{c}, a_\lambda, a_\gamma) \propto \frac{\gamma^{N_j\lambda}}{\Gamma(\lambda)^{N_j}} \tilde{V}_j^{N_j\lambda} e^{-\gamma S_j} a_\lambda e^{-a_\lambda \lambda} a_\gamma e^{-a_\gamma \gamma},$$

where N_j is the number of $\mathbf{c}$ such that $c_i = j$, $\tilde{V}_j = \left[\prod_{i:c_i=j} V_i\right]^{N_j}$ is the geometric mean of those $\{V_i\}$ associated with cluster j and $S_j = \sum_{i:c_i=j} V_i$ is the sum. This is a non-standard density, but sampling can proceed by first sampling λ_j^* from the marginal density proportional to

$$\frac{\Gamma(N_j\lambda + 1)}{\Gamma(\lambda)^{N_j}} \left[\frac{\tilde{V}_j}{S_j + a_\gamma}\right]^{N_j\lambda} e^{-a_\lambda \lambda}$$

via a Metropolis step. Note that this density boils down to that described in Result 2 when only one of $\{V_i\}$ is associated with cluster j. If λ_j^* is accepted, then γ_j^* is sampled from $\Gamma(N_j\lambda_j^* + 1, S_j + a_\gamma)$; otherwise both are left at their previous values. Alternative algorithms for non-conjugate pairings of base-measure and kernel are provided in MacEachern and Müller (1998) and Neal (2000), but these are unnecessary for the model we propose.

A candidate generating kernel for sampling $\boldsymbol{\beta}^*$ can be obtained by a least squares fit of the data to the log-linear model $Y_i = \log(T_i) = \beta_0 - \mathbf{x}_i'\boldsymbol{\beta} + \epsilon_i$. Define the following:

$$X = \begin{bmatrix} 1 & -\mathbf{x}_1' \\ 1 & -\mathbf{x}_2' \\ \vdots & \vdots \\ 1 & -\mathbf{x}_n' \end{bmatrix}, \quad \hat{\boldsymbol{\beta}} = (X'X)^{-1}X'\mathbf{Y}, \text{ and } MSE = \sum_{i=1}^{n} \frac{(Y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}})^2}{n-p-1}.$$

We take the initial value $\boldsymbol{\beta}^0$ to be the last p components of $\hat{\beta}$ and use the scaled, lower right $p \times p$ submatrix of $(X'X)^{-1}MSE$ as the covariance matrix for a multivariate normal candidate generating kernel for $\boldsymbol{\beta}^*$. Define the "baseline residuals" $\hat{v}_i = T_i \exp(\mathbf{x}_i'\boldsymbol{\beta}^0 - \hat{\beta}_0)$. We start the Markov chain at one distinct gamma component using method of moments estimates based on $\hat{v}_1, \ldots, \hat{v}_n$. Where $\bar{v} = \sum_{i=1}^n \hat{v}_i/n$ and $s_v^2 = \sum_{i=1}^n (\hat{v}_i - \bar{v})^2/n$, take $d(\boldsymbol{\theta}^0) = 1$, $\lambda_i^0 = \bar{v}^2/s_v^2$, and $\gamma_i^0 = \bar{v}/s_v^2$. This approach is taken in Section 3.3.

Note that alternatively, the parametric gamma model can be fit to censored data and the resulting maximum likelihood estimates $\hat{\beta}$, $\hat{\lambda}$, and $\hat{\gamma}$ used as starting values instead of values described above. The estimated asymptotic covariance matrix $\widehat{\text{cov}}(\hat{\beta})$ provides a reasonable multivariate normal proposal for a random walk Metropolis step for sampling the full conditional for β in the Gibbs sampler. This approach is taken in Section 3.2.

Using Bayes' rule and Result 1, given $\boldsymbol{\lambda}, \boldsymbol{\gamma}, a_\lambda, a_\gamma$, we compute the posterior predictive baseline density for V_{n+1} as

$$f_{V_{n+1}}(v|\boldsymbol{\lambda}, \boldsymbol{\gamma}, a_\lambda, a_\gamma) = \frac{1}{\alpha+n}\left[\frac{\alpha a_\lambda a_\gamma}{v(v+a_\gamma)[a_\lambda - \log(v/(v+a_\gamma))]^2} + \sum_{i=1}^{n} k(v|\lambda_i, \gamma_i)\right].$$

Note that as the sample size increases, the predictive density essentially becomes a weighted sum of gamma densities. Given the vector $\boldsymbol{\beta}$ and covariates $\mathbf{x}$, the predictive survival density for T_{n+1} is given by

$$f_{T_{n+1}}(t|\boldsymbol{\beta}, \mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\gamma}, a_\lambda, a_\gamma) = f_{V_{n+1}}(t\exp(\mathbf{x}'\boldsymbol{\beta})|\boldsymbol{\lambda}, \boldsymbol{\gamma}, a_\lambda, a_\gamma)\exp(\mathbf{x}'\boldsymbol{\beta}).$$

Given realizations from the Markov chain $\{(\boldsymbol{\beta}^j, \boldsymbol{\lambda}^j, \boldsymbol{\gamma}^j, a_\lambda^j, a_\gamma^j)\}_{j=1}^J$, an estimate of the posterior predictive density for covariates $\mathbf{x}$ is given by

$$\hat{f}_{T_{n+1}}(t|\mathbf{T}, \mathbf{x}) = \frac{1}{J}\sum_{j=1}^{J} f_{V_{n+1}}(t\exp(\mathbf{x}'\boldsymbol{\beta}^j)|\boldsymbol{\lambda}^j, \boldsymbol{\gamma}^j, a_\lambda^j, a_\gamma^j)\exp(\mathbf{x}'\boldsymbol{\beta}^j).$$

Now assume that the first n_1 observations are known only up to the censoring intervals $T_i \in [a_i, b_i)$, $i = 1, \ldots, n_1$, where $b_i = \infty$ for right-censored data. Split $\mathbf{T}$ into $\mathbf{T} = (\mathbf{T}_1', \mathbf{T}_2')'$, where $\mathbf{T}_1 = (T_1, \ldots, T_{n_1})'$ and $\mathbf{T}_2 = (T_{n_1+1}, \ldots, T_n)'$. We denote the inclusion of the first n_1 observations in their respective censoring intervals as $\mathbf{T}_1 \in [\mathbf{a}, \mathbf{b})$. The sampling of the latent T_i's is worked into the Markov chain by considering the conditional distribution of $\{T_i|T_i \in [a_i, b_i), \boldsymbol{\beta}, \lambda_i, \gamma_i\}$. Note that T_i is conditionally independent of $\{\boldsymbol{\lambda}^{(i)}, \boldsymbol{\gamma}^{(i)}, \mathbf{T}^{(i)}, a_\lambda, a_\gamma\}$ given $\{T_i \in [a_i, b_i), \boldsymbol{\beta}, \lambda_i, \gamma_i\}$. Observing $T_i \in [a_i, b_i)$ implies $V_i \in [a_i \exp(\mathbf{x}_i'\boldsymbol{\beta}), b_i \exp(\mathbf{x}_i'\boldsymbol{\beta}))$, and $\{V_i|V_i \in [a_i \exp(\mathbf{x}_i'\boldsymbol{\beta}), b_i \exp(\mathbf{x}_i\boldsymbol{\beta})), \Lambda_i, \Gamma_i\}$ is simply distributed $\Gamma(\lambda_i, \gamma_i)$ restricted to the interval $[a_i \exp(\mathbf{x}_i'\boldsymbol{\beta}), b_i \exp(\mathbf{x}_i'\boldsymbol{\beta}))$. We thus sample $\{T_i|T_i \in [a_i, b_i), \mathbf{T}^{(i)}, \boldsymbol{\beta}, \lambda_i, \gamma_i\}$ using the inverse CDF method by sampling $U \sim U(K(a_i \exp(\mathbf{x}_i'\boldsymbol{\beta})|\lambda_i, \gamma_i), K(b_i \exp(\mathbf{x}_i'\boldsymbol{\beta})|\lambda_i, \gamma_i))$, taking $V_i = K^{-1}(U|\lambda_i, \gamma_i)$ where $K(v|\lambda, \gamma)$ is the c.d.f. of a $\Gamma(\lambda, \gamma)$ random variable, and then setting $T_i = V_i \exp(-\mathbf{x}_i'\boldsymbol{\beta})$. This approach, also used in Ghosh and Ghosal (2004) and Kottas (2005), requires little computation and does not use the rejection algorithm of Kuo and Mallick (1997), and thus is efficient even for very small censoring intervals.

The algorithm for obtaining inference yields an approximation to the posterior of all model parameters, having marginalized the Dirichlet process. A procedure for obtaining full semiparametric posterior inference of arbitrary functionals of the posterior G is given by Gelfand and Kottas (2002). This algorithm is useful for obtaining credible intervals for posterior survival quantiles or hazard functions, for example (Kottas, 2005). Note that the implicit modeling of baseline draws $\{V_i\}$ and the log-linear structure of the model allows the use of standard residual plots to check model adequacy. That is, $e_i = E\{\log(V_i)|\mathbf{T}_1 \in [\mathbf{a}, \mathbf{b}), \mathbf{T}_2\}$ can be easily computed from the MCMC output and plotted versus various predictors or log-predicted values. These plots should display homoscedascity and lack any overall pattern if the model fits.

3 Examples

3.1 Simulated data

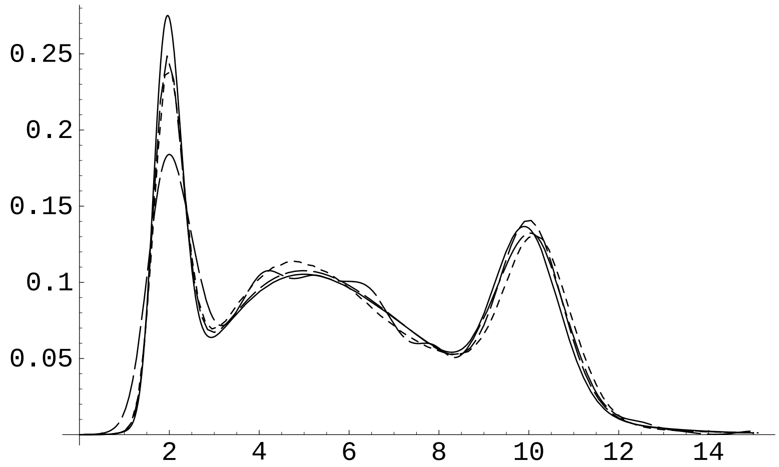

Figure 1: Density estimates. Solid is $f(x)$, long-dashed is kernel estimate, medium-dashed is reversible jump estimate, and short-dashed is Dirichlet process mixture estimate.

Wiper et al. (2001) consider the reversible jump algorithm of Green (1995) using gamma density mixtures for density estimation. The reversible jump algorithm allows transitions between parameter spaces of differing dimensions, which in this case is the number of components in the gamma mixture and their associated parameters. Within the Markov chain, the transitions are limited to "jumps" in the number of components in either direction (one less component or one more component) – hence the term "reversible jump." By considering a Dirichlet process mixture, we achieve the same result but bypass the need to restrict the order of the component locations and allow components of differing locations to have the same spread.

We examine one of the distributions considered in Wiper et al. – a mixture of three gamma densities – and compare three density estimation techniques. The density is

$$f(x) = 0.2k(x|40, 20) + 0.6k(x|6, 1) + 0.2k(x|200, 20).$$

We generated a random sample of size 1000 from $f(x)$ and computed the posterior density estimates from the proposed Dirichlet process mixture of gamma densities (the no-covariate AFT model) and from the reversible jump model of Wiper et al. (2001), using recommendations provided for both models in the absence of any real prior information. We were also interested in how a frequentist model might fare on these data and computed a Gaussian kernel-smoothed estimate using a bandwidth chosen by pseudo-likelihood cross validation (Habbema, Hermans, and van den Broak, 1974). Figure 1 displays the density estimates and the generating density. All three approaches yield valid density estimates, although the kernel smoothed estimate is more "wiggly" by comparison and less accurate at the first density mode. The two Bayesian mixture models yield remarkably similar estimates, which is typical in the author's experience.

To quantify the difference in estimating this density at small sample sizes, we examine the mean integrated squared error (MISE) of the Dirichlet process mixture estimate, the reversible jump estimate of Wiper et al., and the kernel smoothed estimate. The integrated square error (ISE) of an estimate $\hat{f}(x|\mathbf{X})$, based on a data sample $\mathbf{X} = (X_1, \ldots, X_n)'$, is defined to be $ISE = \int(\hat{f}(x|\mathbf{X}) - f(x))^2 dx$. The ISE is a random variable, but the "typical" error is given by $MISE = E_{\mathbf{X}}(ISE)$. We generated 1000 samples from the density $f(x)$ of sizes $n = 10$ and $n = 100$. We fixed $\alpha = 1$ in the Dirichlet process mixture model and used the pseudo-likelihood cross validation bandwidth to obtain kernel smoothed estimates. The $MISE$ estimates and standard errors for the estimates are presented in Table 1. The Dirichlet process mixture and the Wiper et al. approach perform better than the kernel-smoothed estimate for the chosen kernel and bandwidth selection procedure. This is not surprising for two reasons: first, kernel-smoothed estimates are known to be biased at the extrema of $f(x)$ and second, one would expect procedures based on a mixture of gammas to estimate a mixture of gammas well. The Dirichlet process mixture is inferior to Wiper et al. when $n = 10$, but superior when $n = 100$. This may be due in part to decreasing weight placed on the function $\frac{a_\lambda a_\gamma}{v(v+a_\gamma)[a_\lambda - \log(v/(v+a_\gamma))]^2}$ in the predictive density as the sample size increases, and in part due to increased flexibility from specifying a hyperprior for the gamma component parameters rather than fixed values.

Table 1: Comparison of MISE across models.

	DP mixture	Wiper et al.	Kernel Smoothed
$n = 10$	0.0396 (0.0016)	0.0349 (0.0012)	0.0466 (0.0017)
$n = 100$	0.0097 (0.0003)	0.0156 (0.0005)	0.0316 (0.0027)

3.2 Cosmetic effects of cancer therapy

Beadle et al. (1984) and Finkelstein and Wolfe (1985) consider data from a retrospective study designed to compare the cosmetic effects of radiotherapy versus radiotherapy *and* chemotherapy on women with early breast cancer; both treatments are alternatives to a mastectomy that preserve (and thus enhance the appearance of) the breast. It is postulated that chemotherapy in addition to radiotherapy reduces the cosmetic effect of the procedure by inducing breast retraction more quickly than radiotherapy alone.

A retrospective study of 46 radiation only and 48 radiation plus chemotherapy patients was made. Patients were observed typically every 4 to 6 months and at each observation a clinician recorded the level of breast retraction that had taken place since the last visit: none, moderate, or severe. The time-to-event considered was the time until moderate or severe breast retraction, and this time is interval-censored between patient visits or right censored if no breast retraction was detected over the study period of 48 months.

To test whether chemotherapy in addition to radiotherapy has an effect on time-to-breast-retraction, we fit the Dirichlet process mixture of gammas model to these data with $\alpha = 1$. The covariate of interest is $x_i = 0$ if the i^{th} patient had radiotherapy only, and $x_i = 1$ if the i^{th} patient had radiotherapy and chemotherapy; a flat prior was placed on the regression effect.

The posterior median for the regression effect is 0.59 and the 95% equal-tailed credible interval is $(0.22, 0.97)$, indicating that including chemotherapy significantly reduces the time to deterioration; the mean and median time to deterioration is reduced by a factor ranging from 0.4 to 0.8. A log-normal maximum likelihood analysis yields a point estimate and confidence interval of $0.21(0.01, 0.40)$, which is only marginally significant (p-value=0.04). Allowing a nonparametric baseline survival function changes the regression effects markedly. Figure 2 displays predictive survival curves for the two treatment groups across four α priors described below. For $\alpha = 1$ the posterior mode number of months until deterioration is 17, versus 31 when chemotherapy is added to radiotherapy; the posterior median number of months is 22 versus 40.

As a small sensitivity analysis on the choice of α, three other specifications were fit: $\alpha = 0.1$, $\alpha \sim \Gamma(2, 2)$, and $\alpha \sim \Gamma(2, 0.5)$. Fixing $\alpha = 0.1$ produces an essentially parametric analysis. The posterior median of β is reduced to $\hat{\beta} = 0.46$, closer to the log-normal result (Table 2), and predictive survival densities look like a single gamma component (Figure 2). Fixing $\alpha = 1$ yields a mixture model with a small but random number of components; predictive survival distributions have a bit of a "dip" to them relative to $\alpha = 0.1$. The prior $\alpha \sim \Gamma(2, 2)$ has mean $E(\alpha) = 1$ and produces predictive survival densities similar to $\alpha = 1$, but encourages a much greater spread on the number of distinct components $d = d\{\boldsymbol{\theta}_j\}$. The prior $\alpha \sim \Gamma(2, 0.5)$ has mean $E(\alpha) = 4$ and larger variance than $\alpha \sim \Gamma(2, 2)$; this prior has the effect of allowing for a much greater number of components in the mixture (Table 2) and also produces a second mode in the predictive survival densities (Figure 2). Except when fixing the precision to be quite small, $\alpha = 0.1$, posterior inferences, including predictive survival densities and posterior regression coefficient estimates, are fairly robust to the choice of α prior, fixed or continuous. This has been the experience of the author in general. Kottas (2005) also notes robustness of posterior estimates of functionals of the survival density to the choice of continuous α prior.

These data are also analyzed using a Bayesian semiparametric AFT model with a mixture of Dirichlet processes baseline (versus a Dirichlet process mixture) in Hanson and Johnson (2004). Inferences from the semiparametric models closely agree.

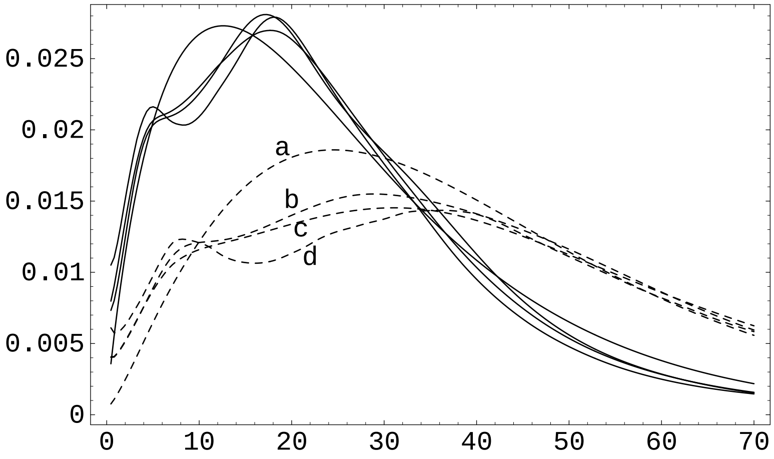

Figure 2: Predictive survival densities for $\alpha = 0.1$ (a), $\alpha = 1$ (b), $\alpha \sim \Gamma(2, 2)$ (c), and $\alpha \sim \Gamma(2, 0.5)$ (d). Dashed is radiotherapy only, solid is radiotherapy and chemotherapy.

3.3 Lung cancer data

We consider a data set presented in Maksymiuk et al. (1993) and subsequently analyzed by Ying, Jung, and Wei (1995), Walker and Mallick (1999), Yang (1999), and Kottas and Gelfand (2001), on the treatment of limited-stage small cell lung cancer in $n = 121$ patients. In the study, it was of interest to determine which sequencing of the drugs cisplaten and etoposide increased the lifetimes of those with limited-stage small cell lung cancer. Treatment A treated an individual with cisplaten followed by etoposide;

Table 2: Posterior inferences for various α priors. Estimates $\hat{\alpha}$, $\hat{\beta}$, and $\hat{d}$ are posterior medians. The posterior credible intervals (CI) are equal-tailed.

	$\alpha = 0.1$	$\alpha = 1$	$\alpha \sim \Gamma(2,2)$	$\alpha \sim \Gamma(2,0.5)$
$\hat{\beta}$	0.46	0.59	0.61	0.61
95% CI: β	(0.19,0.95)	(0.22,0.97)	(0.18,0.98)	(0.23,0.99)
$\hat{\alpha}$	0.1	1.0	1.05	3.16
95% CI: α	NA	NA	(0.20,3.36)	(0.51,10.95)
$\hat{d}$	1	4	4	10
95% CI: d	[1,3]	[1,9]	[1,13]	[2,26]
$P(d = 1)$	0.75	0.03	0.06	0.02
$P(d = 2)$	0.22	0.10	0.14	0.03
$P(d = 3)$	0.03	0.18	0.15	0.05
$P(d = 4)$	0	0.20	0.15	0.05
$P(d \geq 5)$	0	0.49	0.50	0.85

treatment B was etoposide followed by cisplaten. We indicate $x_{i,1} = 0, 1$ for treatments A and B respectively. The patient's age in years at entry into the study $x_{i,2}$ was included as a concomitant variable. Treatment A was administered to 62 patients, while treatment B was administered to 59 patients; 23 patients were administratively right-censored. The data set is thus $\{(T_i, \delta_i, \mathbf{x}_i)\}_{i=1}^{121}$ where T_i is either the observed lifetime in days ($\delta = 1$) or the censoring time ($\delta = 0$).

Ying et al. (1995) analyzed these data with a median-regression model; inference was based on estimating equations and asymptotic normality. Yang (1999) also proposed a median-regression model, but used weighted empirical survival and hazard functions in estimation. We compare the median-regression approaches of Ying et al. (1995) and Yang (1999), mean-regression approach of Buckley and James (1979), a mixture of Polya trees median regression model (Hanson and Johnson, 2002), a Dirichlet process mixture of gamma densities AFT model, and a parametric generalized gamma fit of these data. The model in each case is written

$$Y_i = \log(T_i) = \beta_1 x_{i,1} + \beta_2 x_{i,2} + W_i,$$

with different assumptions on the error W_i depending on the model.

The results of the parametric generalized gamma fit were obtained in SAS (SAS Institute, Inc.), the results of the Ying et al. (1995) and Yang (1999) models were obtained from their papers, and the results of the Buckley and James (1979) model were obtained using Frank Harrell's *Design* library for R, specifically the `bj()` funtion, obtainable from the Comprehensive R Archive Network (CRAN), for example at `http://cran.us.r-project.org/`. We fit the mixture of finite Polya trees regression model to the data as described in Hanson and Johnson (2002) based on their recommendations and using uninformative priors. Convergence was conservatively assessed through running quantile plots.

Table 3: Lung cancer data: regression effect estimates across models.

Coef.	Generalized Gamma	Ying, Jung, & Wei (1995)
β_1	-0.404 (-0.678,-0.130)	-0.375 (-0.983,-0.081)
β_2	-0.017 (-0.033,-0.001)	-0.009 (-0.037,0.007)
Coef.	Buckley & James (1979)	Mixture of Polya trees
β_1	-0.419 (-0.640,-0.196)	-0.363 (-0.682,-0.186)
β_2	-0.018 (-0.030,-0.005)	-0.005 (-0.033,0.007)
Coef.	Yang (1999)	DP mixture of gammas
β_1	-0.368 (-0.530,-0.230)	-0.398 (-0.635,-0.157)
β_2	-0.012 (-0.018,-0.002)	-0.010 (-0.026,0.004)

A Dirichlet process mixture of gamma densities model was fit to the data, fixing $\alpha = 1$ and taking a flat prior on $\boldsymbol{\beta}$: $f(\boldsymbol{\beta}) \propto 1$. The resulting predictive survival density for treatment A at the sample mean age of 62.11 is shown in Figure 3, along with the density from the generalized gamma fit of the data. The densities are somewhat close, which is what we would expect as the simpler parametric model provides fairly adequate fit based on various residual plots. However, the mixture model is slightly shifted and indicates that perhaps two or more gamma components are necessary to fit these data. In our experience with simulated data, the Dirichlet process mixture model tends to give similar results for a wide range of α's and typically does not "overfit" data when $\alpha = 1$. A plot of $e_i = E\{\log(V_i)|\mathbf{T}_1 \in [\mathbf{a}, \mathbf{b}), \mathbf{T}_2\}$ versus age $x_{i,2}$ shows no signs of curvature or heteroscedascity (Figure 4).

We present results on the regression coefficients for each model in Table 3. The four frequentist approaches yield a point estimate and 95% confidence interval for the parameters of interest; the two Bayesian models yield posterior median and equal-tailed 95% credible intervals. The results from all models agree somewhat, although the age effect is significant in neither Bayesian semiparametric model nor the model of Ying et al. (1995).

4 Conclusions

Mixture of gammas models provide a rich baseline for the accelerated failure time model. We have proposed a Dirichlet process mixture of gammas model that can be used with a minimal amount of prior specification and which does not tend to "overfit" data. The model can accommodate arbitrarily censored data (including interval censoring) and is relatively easy to code and implement.

Often, semiparametric and nonparametric techniques are overlooked by practitioners due to the availability of simpler methods in widely available packages. We are currently working on implementing this and related models in the popular R statistical package

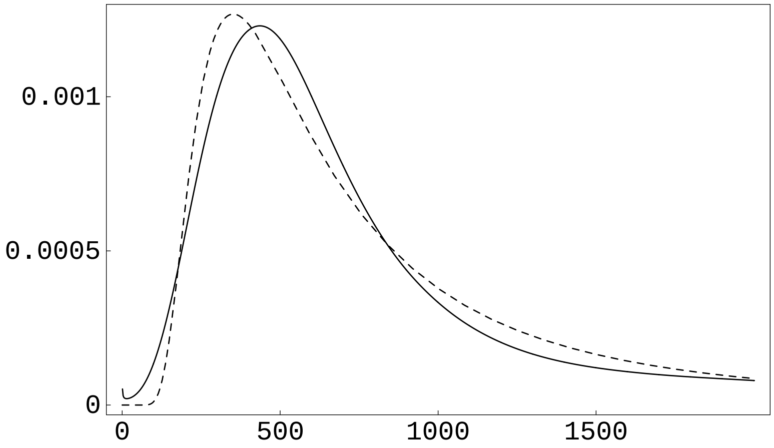

Figure 3: Lung cancer data: predictive survival densities in days for treatment A, age at entry 62. Solid is DP mixture of gammas, dashed is generalized gamma.

(Ihaka and Gentleman, 1996) for those who wish to explore these models, but are not fond of coding. Two successful examples of such novel R code for fitting AFT models include Arnost Komarek's functions `smoothSurvReg()`, which models the log-error distribution in the AFT model as a penalized G-spline and `bayessurvreg1()`, which models the log-error distribution as a mixture of Gaussian components via reversible jump. Both of these functions are available from the CRAN.

Clustered survival data are often modeled through the use of random effects termed *frailties*. It is straightforward to extend the model considered in this paper to the frailty model

$$T_{ij} = \exp(-\mathbf{x}_{ij}'\boldsymbol{\beta})W_i V_{ij},$$

where $i = 1, \ldots, n$ indexes clusters and $j = 1, \ldots, n_i$ indexes observations within a cluster. The V_{ij} are *iid* from the Dirichlet process mixture of gammas baseline, and the frailties W_i arise as *iid* from some distribution with mean or median one – often $\Gamma(\rho, \rho)$ or log-normal$(0, \rho)$ in the literature. Walker and Mallick (1997) consider a Polya tree prior on the distribution of $\{W_i\}$ whereas Sahu and Dey (2004) introduce a family of log-skew-t distributions. Another possibility is a mixture of Polya trees constrained to have median one (Hanson, 2004).

Ghosh and Ghosal (2004) consider a scale Dirichlet process mixture of Weibull components baseline in the AFT model, whereas Kottas (2005) considers Dirichlet process mixing over both the shape and the scale, but without covariates. A formal comparison of these approaches and the Dirichlet process mixtures of gammas model considered in this paper, with all models incorporating covariates, would be a welcome addition to the literature. A comprehensive comparison could also include the AFT models of Walker and Mallick (1999), Kottas and Gelfand (2001), and Hanson and Johnson (2002), as well

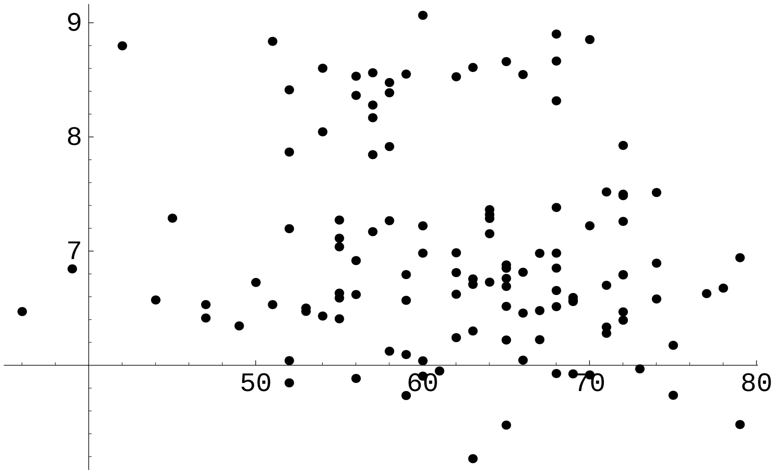

Figure 4: Lung cancer data. A plot of the posterior error estimates $e_i = E\{\log(V_i)|\mathbf{T}_1 \in [\mathbf{a}, \mathbf{b}), \mathbf{T}_2\}$ versus the predictor age $x_{i,2}$ shows no obvious lack of fit.

as the work of Brunner (1995), which considers log-baseline densities that are unimodal and symmetric.

References

Beadle, G. F., Come, S., Henderson, C., Silver, B., and Hellman, S. A. H. (1984). "The effects of adjuvant chemotherapy on the cosmetic results after primary radiation treatment for early stage breast cancer." *International Journal of Radiation Oncology, Biology, and Physics*, 10: 2131–2137.

Brooks, S. P. and Giudici, P. (2000). "MCMC convergence assessment via two-way ANOVA." *Journal of Computational and Graphical Statistics*, 9: 266–285.

Brunner, L. J. (1995). "Bayesian linear regression with error terms that have symmetric unimodal densities." *Journal of Nonparametric Statistics*, 4: 335–348.

Buckley, J. and James, I. (1979). "Linear regression with censored data." *Biometrics*, 8: 907–925.

Bush, C. A. and MacEachern, S. N. (1996). "A semiparametric Bayesian model for randomized block designs." *Biometrika*, 83: 275–285.

Christensen, R. and Johnson, W. O. (1988). "Modeling accelerated failure time with a Dirichlet process." *Biometrika*, 75: 693–704.

Copsey, K. and Webb, A. (2003). "Bayesian gamma mixture model approach to radar target recognition." *IEEE Transactions on Aerospace and Electronic Systems*, 39: 1201–1217.

Escobar, M. D. (1994). "Estimating normal means with a Dirichlet process prior." *Journal of the American Statistical Association*, 89: 268–277.

Escobar, M. D. and 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." *Annals of Statistics*, 1: 209–230.

Finkelstein, D. M. and Wolfe, R. A. (1985). "A semiparametric model for regression analysis of interval-censored failure time data." *Biometrics*, 41: 933–945.

Gelfand, A. E. and Kottas, A. (2002). "A computational approach for full nonparametric Bayesian inference under Dirichlet process mixture models." *Journal of Computational and Graphical Statistics*, 11: 289–305.

Ghosh, S. K. and Ghosal, S. (2004). "Proportional mean regression models for censored data." *Technical Report, North Carolina State University Department of Statistics*.

Green, P. (1995). "Reversible jump MCMC computation and Bayesian model determination." *Biometrika*, 82: 711–732.

Green, P. and Richardson, S. (2001). "Modelling heterogeneity with and without the Dirichlet process." *Scandinavian Journal of Statistics*, 28: 355–375.

Habbema, J. D. F., Hermans, J., and van den Broak, K. (1974). "A stepwise discrimination analysis program using density estimation." In *Compstat 1974: Proceedings in Computational Statistics*, 1974. Vienna: Physica Verlag.

Hanson, T. (2004). "Inference for mixtures of finite Polya tree models." *Technical Report, University of New Mexico Department of Mathematics and Statistics*.

Hanson, T. and Johnson, W. O. (2002). "Modeling regression error with a mixture of Polya trees." *Journal of the American Statistical Association*, 97: 1020–1033.

— (2004). "A Bayesian semiparametric AFT model for interval censored data." *Journal of Computational and Graphical Statistics*, 13: 341–361.

Ihaka, R. and Gentleman, R. (1996). "R: A language for data analysis and graphics." *Journal of Computational and Graphical Statistics*, 5: 299–314.

Ishwaran, H. and Zarepour, M. (2002). "Dirichlet prior sieves in finite normal mixtures." *Statistica Sinica*, 12: 941–963.

Kottas, A. (2005). "Nonparametric Bayesian survival analysis using mixtures of Weibull distributions." *Journal of Statistical Planning and Inference*.

Kottas, A. and Gelfand, A. E. (2001). "Bayesian semiparametric median regression modeling." *Journal of the American Statistical Association*, 95: 1458–1468.

Kuo, L. and Mallick, B. (1997). "Bayesian semiparametric inference for the accelerated failure-time model." *Canadian Journal of Statistics*, 25: 457–472.

Lo, A. Y. (1984). "On a class of Bayesian nonparametric estimates: I. Density estimates." *Annals of Statistics*, 12: 351–357.

MacEachern, S. N. and Müller, P. (1998). "Estimating mixture of Dirichlet process models." *Journal of Computational and Graphical Statistics*, 7: 223–238.

Maksymiuk, A. W., Jett, J. R., Earle, J. D., Su, J. Q., Diegert, F. A., Mailliard, J. A., Kardinal, C. G., Veeder, J. E. K. M. H., Wiesenfeld, M., Tschetter, L. K., and Levitt, R. (1994). "Sequencing and schedule effects of cisplatin plus etoposide in small cell lung cancer results of a north central cancer treatment group randomized clinical trial." *Journal of Clinical Oncology*, 12: 70–76.

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

— (2003). "Slice sampling (with discussion)." *Annals of Statistics*, 31: 705–767.

Richardson, S. and Green, P. (1997). "On Bayesian analysis of mixtures with an unknown number of components." *Journal of the Royal Statistical Society, Series B*, 59: 731–792.

Sahu, S. K. and Dey, D. K. (2004). *On multivariate survival models with a skewed frailty and a correlated baseline hazard process*, 321–338. Boca Raton, FL: CRC/Chapman & Hall.

Spiegelhalter, D., Thomas, A., Best, N., and Lunn, D. (2003). *WinBUGS 1.4 User Manual*. MRC Biostatistics Unit. Http://www.mrc-bsu.cam.ac.uk/bugs.

Tierney, L. (1994). "Markov chains for exploring posterior distributions." *Annals of Statistics*, 22: 1701–1762.

Walker, S. and Muliere, P. (2003). "A bivariate Dirichlet process." *Statistics and Probability Letters*, 64: 1–7.

Walker, S. G. and Mallick, B. K. (1997). "Hierarchical generalized linear models and frailty models with Bayesian nonparametric mixing." *Journal of the Royal Statistical Society, Series B*, 59: 845–860.

— (1999). "Semiparametric accelerated life time model." *Biometrics*, 55: 477–483.

Wiper, M., Insua, D. R., and Ruggeri, F. (2001). "Mixtures of gamma distributions with applications." *Journal of Computational and Graphical Statistics*, 10: 440–454.

Yang, S. (1999). "Censored median regression using weighted empirical survival and hazard functions." *Journal of the American Statistical Association*, 94: 137–145.

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