UNIVERSITY OF CALIFORNIA
SANTA CRUZ

BAYESIAN NONPARAMETRIC GAMMA MIXTURES FOR MEAN RESIDUAL LIFE INFERENCE

A dissertation submitted in partial satisfaction of the
requirements for the degree of

DOCTOR OF PHILOSOPHY

in

STATISTICS AND APPLIED MATHEMATICS

by

Valerie Poynor

September 2013

The Dissertation of Valerie Poynor
is approved:

Professor Athanasios Kottas, Chair

Professor David Draper

Professor Marc Mangel

Professor Abel Rodriguez

Tyrus Miller
Vice Provost and Dean of Graduate Studies

Table of Contents

List of Figures

List of Tables

Abstract

Bayesian Nonparametric Gamma Mixtures for Mean Residual Life Inference

by

Valerie Poynor

In survival analysis interest lies in modeling data that describe the time to a particular event. Informative functions, namely the hazard function and mean residual life function, can be obtained from the model's distribution function. We focus on the mean residual life function which provides the expected remaining life given that the subject has survived (i.e., is event-free) up to a particular time. This function is of direct interest in reliability, medical, and actuarial fields. In addition to its practical interpretation, the mean residual life function characterizes the survival distribution. In terms of mean residual life function inference, there are two shortcomings present in the current literature. First off, the shape of the functional is often restricted, which forces the researcher to make an assumption that may not be appropriate. Secondly, in cases where the shape of the functional is not parametrically specified, full inference is not obtained. The aim of our research is to provide a modeling approach that yields full inference for the mean residual life function, and is not restrictive on the shape of the functional. In particular, we develop general Bayesian nonparametric modeling methods for inference for mean residual life functions built from a mixture model for the associated survival distribution. Although the prior model is not placed on the mean residual life function directly, our methods offer rich inference for the desired functional. We place a Dirichlet process

mixture model on the survival function, and discuss the importance of careful kernel selection to ensure desirable properties for the mean residual life function. We advocate for a mixture model with a gamma kernel and dependent baseline distribution for the Dirichlet process prior. We extend our model to the regression setting by modeling the joint distribution for the survival response and random covariates. This approach provides a flexible method for obtaining inference for the regression functionals when the number of random covariates is small to moderate. We further extend our methods to the scenario where interest lies in comparison of survival between two experimental groups. Typically, we expect the range of survival in the two groups to be the same, but exhibiting different characteristics over that range. Here, we develop a dependent Dirichlet process prior for the mixing distributions having shared locations across the two groups and varying weights to incorporate dependency between populations and achieve richer inferential results. The final scenario we consider is the case in which the researcher believes two populations have ordered mean residual life functions. For such applications, a prior model that incorporates an ordering constraint on the mean residual life functions is attractive. We introduce a mixture of Erlang distributions with weights constructed using Dirichlet process priors that provides the mean residual life ordering result. We demonstrate the utility of our modeling methods through simulation and real data examples. In addition, we draw comparisons with both parametric and semiparametric models.

To my parents, who have loved and supported me every step of the way.

Acknowledgments

I cherish all whom have played a part, big or small, in the continuous series of events that have resulted in where I am today. I am grateful for the listening ear of my family and friends. Phone conversations, visits, and even facebook messages have lessened the distance of loved ones, lifted my spirits, and made me laugh.

My colleagues and professors have played a critical role in my academic success. Between study groups, a great learning atmosphere in the classroom, office hours, and late night emails, I was able to comprehend complex material at a faster rate and on a deeper level. Without the personal and academic support of my peers and teachers, I would not be where I am today.

Finally, I would like to acknowledge my advisor, Thanasis, who has taught me never to give up, to take my time so that I produce my best work (but there is a such a thing as TOO much time), consuming a piece of steak solves all illnesses, and that Bayesian nonparametric statistics is just plain awesome. To Thanasis, I would like to say thank you for your guidance, your patience, for pushing me to be best I could be, and for believing in me the whole way through.

Chapter 1

Introduction

1.1 Background

Survival data describe the time to a particular event. This event is typically referred to as the failure of some machine or death of a person. However, survival data can also represent duration of unemployment, life expectancy of a product, the time until a patient relapses, etc. Continuous-time survival data is more prevalent, however, there are situations in which survival time lives on a discrete space. Situations in which discrete-time survival data is more appropriate include time to graduation of students in terms of semesters or quarters, the number of years of teachers serve before retirement, and the number of menstrual cycles occurs until a couple conceives. Often discrete-time survival data is the product of logistical and financial restrictions, such that the data can only be collected at certain time periods. In our research, we will consider only continuous survival data. For examples and methodologies involving discrete survival

data see, e.g., Singer & Willet (1993); Willet & Singer (1993); Scheike & Jensen (1997); and Cox & Oakes (1984).

The survival function of a continuous positive random variable T defines the probability of survival beyond time t, $S(t) = Pr(T > t) = 1 - F(t)$, where $F(t)$ is the distribution function. The hazard rate function computes the probability of a failure in the next instant given survival up to time t,

$$h(t) \quad = \quad \lim_{\Delta t \to 0} \frac{Pr[t < T \leq t + \Delta t | T > t]}{\Delta t} = \frac{f(t)}{S(t)}$$

where $f(t)$ is the probability density function. The mean residual life (mrl) function computes the expected remaining survival time of a subject given survival up to time t. Suppose that $F(0) = 0$ and $\mu \equiv E(T) = \int_0^\infty S(t)dt < \infty$. Then the mrl function for continuous T is defined as:

$$m(t) \quad = \quad E(T - t | T > t) = \frac{\int_t^\infty (u - t) f(u) du}{S(t)} = \frac{\int_t^\infty S(u) du}{S(t)} \tag{1.1}$$

and $m(t) \equiv 0$ whenever $S(t) = 0$. The mrl function is of particular interest because of its easy interpretability and large area of application (Guess & Proschan, 1985). Moreover, it characterizes the survival distribution via the Inversion Formula (1.2). Again for continuous T with finite mean, the survival function is defined through the mrl function:

$$S(t) = \frac{m(0)}{m(t)} \exp\left[- \int_0^t \frac{1}{m(u)} du \right]. \tag{1.2}$$

The Inversion Formula is one of several defining properties of the mrl function. The mrl function is also nonnegative and right continuous. More interesting is that the

mrl function plus its argument, $m(t) + t$, is nondecreasing. These properties, in addition to a couple more formalities that are formally stated in Section 2.1.1, provide the characterization theorem for the mrl function (Hall & Wellner, 1981).

As with the hazard function, it is possible to characterize the form of the mrl function for standard parametric distributions. In Section 2.1.3, we explore mrl functions for a number of commonly used parametric models for survival data. While the shapes of these mrl functions are transparent there is a downfall. Unfortunately, the shape of the mrl function is often limited to be monotonically increasing (INC) or decreasing (DCR), which may be appropriate for some situations, but not suitable for other populations. For instance, biological lifetime data tend to support lower mrl during infancy and elderly age while there is a higher mrl during the middle ages. The shape of this mrl function is unimodal and commonly referred to as upside-down bathtub (UBT) shape. Modifications of the Weibull model have been explored in order to develop more flexible parametric distributions with regard to the shapes of the hazard and mrl functions; see, e.g., Pham & Lai (2007).

Several papers have investigated the shape of the mrl function in relation to the hazard function. For instance, a well-known relationship for monotonically increasing (decreasing) hazard functions is that the corresponding mrl function will be monotonically decreasing (increasing); see Finkelstein (2002) for a review. Gupta & Akman (1995) establish sufficient conditions for the mrl function to be decreasing (increasing) or UBT (BT) given that the hazard is BT (UBT). Xie et al. (2004) look at the specific change points of mrl function and hazard function.

Another point of interest lies in inference for the mrl function. The classical survival analysis literature includes several estimation techniques for mrl functions. The most basic estimator, being the empirical estimate, was first studied in Yang (1978). The empirical estimate is defined by $\hat{m}_n(t) = ((\int_t^\infty S_n(u)du)/S_n(t))\mathbb{1}_{[0,T_{(n)}]}(t)$ where $S_n(t)$ is the empirical survival function and $T_{(n)}$ is the maximum observed survival time. It is shown that under this fixed finite interval, the estimator is asymptotically unbiased, is uniformly strong consistent, and as n goes to infinity it converges in distribution to a Gaussian process. Hall & Wellner (1979) extend the empirical estimator by defining it for values on the positive real line. Furthermore, they provide nonparametric confidence bands for the estimate via transformations of the limiting process of the estimator into Brownian motion. Abdous & Berred (2005) use a local linear fitting technique to find a smooth estimate assuming only that the smoothing kernel is symmetric. A nonparametric hypothesis testing procedure for comparing mrl functions from two independent groups was introduced by Berger et al. (1988). A practical benefit of this procedure is that mrl estimates of the two groups were allowed to cross, a pattern that is likely to arise in applications.

Classical estimation for the mrl function began to have a semiparametric regression flavor when Oakes & Dasu (1990) extended the class of distribution having linear mrl functions (Hall & Wellner, 1981), to a family having proportional mrl functions, $m_1(t) = \psi m_2(t)$ for $\psi > 0$. Maguluri & Zhang (1994) further extended the proportional mrl model to a regression setting, $m(t|z) = \exp(\psi z)m_0(t)$, where z is a vector of covariates, ψ is of vector of regression coefficients, and $m_0(t)$ is a baseline

mrl function. Chen & Cheng (2005) also extend the proportional mrl model to include inference for the regression parameters with censored data.

In contrast to the classical literature, there has been very little work on modeling and inference for mrl functions under the Bayesian framework. Lahiri & Park (1991) present nonparametric Bayes and empirical Bayes estimators under a Dirichlet process (Ferguson, 1973) prior for the probability distribution. They show that the Bayes estimator becomes a weighted average of the prior guess for the mrl function and the empirical mrl function of the data. Johnson (1999) discusses a Bayesian method for estimation of the mrl function under interval and right censored data, also using a Dirichlet process prior for the corresponding survival function.

1.2 Objectives and contributions

The wide range of application for the mrl function illustrates the practical importance of this functional. Ideally, we would like to develop flexible Bayesian prior models for the mrl function directly. Based on our literature search, such model structures do not exist. The reason for the lack of development on this topic may be due to the difficulty in obtaining a likelihood. Obtaining the likelihood requires the use of the inversion formula (1.2), which involves integration over the reciprocal of the mrl function. Note that classical nonparametric estimation techniques do not face this issue because they are not built from probabilistic modeling of the survival distribution of mrl function. Forms of the mrl functions that lead to convenient integration are re-

strictive with respect to the shape of the mrl function. One possibility is to look at a mixture of mrl functions for which the corresponding distribution function is easily obtainable. It is easy to show that a linear combination of mrl functions produces a valid mrl function. We have explored models for the mrl function by mixing over a class of parametric mrl functions. We developed these classes by creating functions that satisfy the characterization theorem for mrl functions. We were able to capture unique skewness characteristics as well as some basic shapes for the mrl function, however, we could not develop a class of mrl functions that allowed general (multimodal) shapes. More critically, implementation of posterior inference suffers from the complicated way the mrl function enters the likelihood.

The obstacles in modeling the mrl function directly have directed us to looking at the inference for the mrl function that is implied by Bayesian nonparametric mixture models for the density. The choice of kernel is important when the interest lies in mrl inference. We investigate and report the implications of the mrl function under various kernel mixture distributions, providing a basic set of criteria for kernel selection. In particular, we lend a sufficiency condition that ensures finiteness of the mean of the mixture distribution, a requirement for a well-defined mrl function. We also discuss the limitations that some kernels imply for the tail behavior of the mrl function. Taking these criteria and properties into consideration, we choose to work with a gamma kernel distribution. The model we propose is a model for the density function, however, we are still able to show that the model is dense, in the pointwise sense, on the space of mrl functions. This result indicates that under flexible prior framework, we will be able to

closely estimate any mrl function. Details of the model formulation and kernel selection can be found in Chapter 2.

In the case where predictor variables for the survival response distribution are present, we provide a framework for obtaining inference for mrl functions. In this context, fully nonparametric regression modeling is appealing as it can capture different mrl function shapes for different parts of the covariate space. We convey the details for joint modeling of survival times and random covariates, and provide a simulation example with a single continuous covariate demonstrating a variety of mrl forms over the space of covariate values. When the data consists of two experimental groups, the dependent Dirichlet process (DDP) prior for the mixing distribution allows dependency across the two populations to be incorporated in the model. In particular, we propose a DDP in which the populations share the same locations, but the weights vary for each group. This structure has not been developed in the literature as much as the more commonly used structure of having the groups sharing the same weights and varying locations. However, in the context of survival data, the former structure is more natural. Specifically, we may expect the range of the survival times of the two experimental groups to be the same, but exhibiting different prevalence across survival time. Further discussion of our model development and implementation for Bayesian nonparametric survival regression is divulged in Chapter 3.

When a researcher knows that the mrl of one population is higher than that of another population across the entire support, then a Bayesian model that has this ordering property in the prior is an attractive model for data obtained from these popu-

lations. Models that assume ordering between two distribution functions (i.e. stochastic order) have been studied and formulated from both a Bayesian and frequentist perspective. On the other hand, there is not substantial literature on study for mrl ordering, and to our knowledge, no work on model formulation from a Bayesian nonparametric perspective. In Chapter 4, we develop and implement a Bayesian nonparametric model for this setting.

Chapter 2

Nonparametric Bayesian inference for Mean Residual Life functions

The focus of this chapter is the development of a Bayesian nonparametric mixture model that achieves flexible inference for the mrl function. We begin by reviewing important properties of the mrl function, and looking at the shapes of a number of mrl functions that are associated with commonly used parametric models in survival analysis (Section 2.1). These commonly used parametric model motivate the need for a more flexible modeling approach to mrl inference. In Section 2.2, we present the gamma DDP mixture model for mrl inference. In Section 2.3, we demonstrate the ability of the model to capture unique features in the mrl function through data examples. We close the chapter with our concluding remarks in Senction 2.4.

2.1 Theory and properties of MRL functions

In this section, we review some important properties and characteristics of the mrl function and provide the form of the mrl function for several common distributions. We begin with some elementary properties that are well-established in the literature that either lead to the development of the Inversion Formula, presented in Equation (1.2), or become of interest once the Inversion Formula is provided. We state the characterization theorem for the mrl function (Hall & Wellner, 1981). Finally, we provide alternative forms of the mrl function that aid in studying various shapes of the mrl function for a number of commonly used distributions as well as the exponentiated Weibull distribution.

2.1.1 Properties of mrl functions

We start out by recalling an elementary relationship between the survival function and the moments of the distribution. If the r^{th} moment exists for a continuous random variable T, we have:

$$E(T^r) \quad = \quad r \int_0^\infty t^{r-1} S(t) dt \tag{2.1}$$

This expression is of interest, because once we establish the Inversion Formula (1.2), we have a way of obtaining the moments (when they exist) from the mrl function. Additionally, we have an expression for the variance in terms of the survival function: $Var(T) = 2\int_0^\infty tS(t)dt - \left[\int_0^\infty S(t)dt\right]^2$.

We have already defined the mrl as the expectation of the remaining survival

time given survival up to time t. Here we derive the expression for the mrl function through the survival function as stated in (1.1),

$$\begin{aligned} m(t) &= \int_t^\infty (u-t)dP(T \leq u|T > t) = \int_t^\infty (u-t)\left(\frac{-S'(u)du}{S(t)}\right) \\ &= \frac{\lim_{u\to\infty}(u-t)S(u) - (t-t)S(t) + \int_t^\infty S(u)du}{S(t)} = \frac{\int_t^\infty S(u)du}{S(t)} \end{aligned}$$

where the first limit in the last step tends to 0 since we assume that the first moment exists, and the second limit tends to 0 since $F(\infty) = 1$. It is now easily seen that the first moment is equivalent to the mrl function at $t = 0$.

$$m(0) = \frac{\int_0^\infty (u-0)f(u)du}{S(0)} = \frac{\int_0^\infty uf(u)du}{1} = \mu. \tag{2.2}$$

The following properties are also provided in Hall & Wellner (1981), and are essential for the development of the characterization theorem for mrl functions: (a) $m(t)$ is a nonnegative and right-continuous, and $m(0) = \mu > 0$; (b) $v(t) \equiv m(t) + t$ is non-decreasing; (c) $m(t^-) > 0$ for $t \in (0, X)$, where $X \equiv \inf\{t : F(t) = 1\} \leq \infty$. If $X < \infty$, $m(X^-) = 0$, and m is continuous at X, $(m(t^-) \equiv \lim_{t\to t^-} m(t))$; (d) $S(t) = m(0)/m(t)exp\left[-\int_0^t 1/m(u)du\right]$, for all $t < X$ (Inversion Formula); (e) $\int_0^t 1/m(u)du \to \infty$ as $t \to X$. Property (d) is known as the Inversion Formula (1.2). See Appendix A for proofs.

We conclude the review of properties for mrl functions with a key result that provides necessary and sufficient conditions such that a function is the mrl function for a survival distribution, and thus it characterizes mrl functions.

Characterization Theorem: Suppose a function $m(t)$ which maps $\mathbb{R}^+ \to \mathbb{R}^+$ satisfies (a) $m(t)$ is right-continuous and $m(0) > 0$; (b) $v(t) \equiv m(t) + t$ is non-decreasing; (c) if $m(t^-) = 0$ for some $t = t_0$, then $m(t) = 0$ for $t \in [t_0, \infty)$; (d) if $m(t^-) > 0$ for all t, then $\int_0^\infty 1/m(u)du = \infty$. Let $X \equiv inf\{t : m(t^-) = 0\} \leq \infty$, and define $S(t)$ by (1.2) for $t < X$ and $S(t) \equiv 0$ for $t \geq X$. Then $F(t) \equiv 1 - S(t)$ is a distribution function on $\mathbb{R}^+$ with $F(0) = 0$, $X_F = X$, finite mean $\mu_F = m(0)$, and mrl function $m_F(t) = m(t)$.

2.1.2 Linear mrl function

Oakes & Dasu (1990) focus on linear mrl functions discussed in Hall & Wellner (1981). The key result is that if the mrl function is linear, $m(t) = At + B$ $(A > -1, B > 0)$, then by use of the Inversion Formula (1.2), the survival function has the form:

$$S(t) = \left[\frac{B}{At+B}\right]_+^{\frac{1}{A}+1} \tag{2.3}$$

When $A = 0$, the survival distribution is an exponential with mean B, however, for $A \neq 0$:

$$S(t) = \left(\tfrac{B}{At+B}\right) exp\left[-\int_0^t \tfrac{1}{Au+B}du\right] = \left(\tfrac{B}{At+B}\right) exp\left[-\tfrac{1}{A}ln(Au+B)\right]_0^t$$
$$= \left(\tfrac{B}{At+B}\right) \frac{exp\left[ln(At+B)^{-\frac{1}{A}}\right]}{exp\left[ln(B)^{-\frac{1}{A}}\right]} = \left(\tfrac{B}{At+B}\right)\left(\tfrac{B}{At+B}\right)^{\frac{1}{A}} = \left(\tfrac{B}{At+B}\right)_+^{\frac{1}{A}+1}$$

where the positive part is necessary to satisfy the nonnegative property of the survival function.

For $A > 0$ the survival function is a Pareto distribution. The form of the survival function of the Pareto distribution for random variable Z is, $S(z) = (\beta/z)^\alpha$ for $\beta > 0$ (scale), $\alpha > 0$ (shape), and $z \in \beta, +\infty)$. If we consider the transformation

$Z = AT + B$ where $B = \beta$ and $1/A + 1 = \alpha$, then we have $Z \sim Pareto(\alpha, \beta)$. Note that the first moment only exists for the Pareto distribution when $\alpha > 1$ therefore, since $1/A + 1 > 1$, the mean of the survival distribution exists for linear mrl with $A, B > 0$. Finally, since $Z \geq \beta > 0 \Rightarrow \beta/z > 0$, the survival function is always positive, therefore, no precautions need be made with regard to taking only the positive part of the function.

For $-1 < A < 0$ the survival function is a rescaled beta distribution. The pdf of a rescaled beta distribution is $f(z|a, b, p, q) = ((z-a)^{p-1}(b-z)^{q-1})/(\mathbf{B}(p, q)(b-a)^{p+q+1})$ where $a \leq z \leq b, \quad p, q > 0$, and $\mathbf{B}(.,.)$ is the beta function defined as $\mathbf{B}(p, q) = \int_0^1 t^{p-1}(1-t)^{q-1}\, dt$. Start with the form of the survival function from the linear mrl, $S(t) = [B/(At + B)]_+^{1/A+1}$, note: that the positive part is obtained when $-At \leq B \rightarrow t \leq -B/A$. Then $F(t) = 1 - [B/(At + B)]^{1/A+1}$. Thus we have,

$$f(t) = -(1/A + 1)\left[\frac{B}{At+B}\right]^{\frac{1}{A}}\left[\frac{AB}{(At+B)^2}\right] = -\frac{\left(\frac{1}{A}+1\right)AB^{\frac{1}{A}+1}}{(At+B)^{\left(\frac{1}{A}+1\right)+1}} = -\frac{A(At+B)^{-\left(\frac{1}{A}+1\right)-1}}{\left(\frac{1}{A}+1\right)^{-1}B^{-\left(\frac{1}{A}+1\right)}}.$$

$$\text{Let } Z = -AT \Rightarrow \frac{dt}{dz} = -\frac{1}{A} \Rightarrow f(z) = \frac{+\frac{A}{A}(B-z)^{-\left(\frac{1}{A}+1\right)-1}}{\left(\frac{1}{A}+1\right)^{-1}B^{-\left(\frac{1}{A}+1\right)}} \overset{q=-\left(\frac{1}{A}+1\right)}{=} \frac{(B-t)^{q-1}}{\left(-\frac{1}{q}\right)B^{-q}}$$

Now we can see that $B = b, a = 0, p = 1$. When $p = 1 \Rightarrow \mathbf{B}(p = 1, q) = \int_0^1 (1-t)^{q-1}dt = -(1/q)$ we have, $f(z) = ((z-0)^{1-1}(b-z)^{q-1})/(\mathbf{B}(1, q)(b-0)^{q+1-1}) \quad 0 \leq z \leq b$. The survival function is given by,

$$S(z) = 1 - \int_0^z \frac{(u-0)^{1-1}(b-u)^{q-1}}{\mathbf{B}(1, q)(b-0)^{q+1-1}}du = 1 - \frac{\int_0^z (b-u)^{q-1}\, du}{\mathbf{B}(1, q)b^q} = \left(\frac{b-z}{b}\right)^q$$

which is precisely the transformed survival function.

2.1.3 The form of the mrl function for some common distributions

In this section, we summarize our investigation of the forms of mrl functions for a number of common distributions. In the previous section, we discussed the distributions having a linear mean residual life function namely the exponential, Pareto, and rescaled beta. These distributions share the convenient feature that they yield a closed form for the mrl function. On the other hand, the linearity of the mrl is too limiting to be of much practical use. There are a number of distributions having more flexible mrl functions, such as increasing and decreasing curvatures as well as BT or UBT shapes. The difficulty for these distributions lies in obtaining a closed form of the mrl. Recall from (1.1) that the mrl is defined as $\int_t^\infty S(u)du/S(t)$. Alternatively, the mrl can be written as,

$$m(t) = \frac{\int_t^\infty (u-t)f(u)du}{S(t)} = \frac{\int_t^\infty uf(u)du}{S(t)} - \frac{t\int_t^\infty f(u)du}{S(t)} = \frac{\int_t^\infty uf(u)du}{S(t)} - t \qquad (2.4)$$

Govilt & Aggarwal (1983) derive (2.4) by starting with $\int_t^\infty f(u)du$ and applying integration by parts and solving for $\int_t^\infty S(u)du$ to obtain $\int_t^\infty S(u)du = \int_t^\infty uf(u)du - tS(t)$, then dividing both sides by S(t). This derivation requires that $tS(t) \to 0$ as $t \to \infty$. This limit converges to 0 as long as the distribution function is right continuous and has finite mean. The distributions that we discuss meet these requirements. The mrl can also be obtained, perhaps more directly, from the equality stated in Hall & Wellner (1981) by subtracting t from both sides.

The distributions discussed here have no known closed form for their associated mrl making them difficult to explore. However, through the use of (2.4) and/or simple transformations of T, we are able to obtain forms of the mrl functions comprised

of well-known integrals. Although these forms are far from an ideal closed form, they are easy to evaluate with most statistical programming software.

Gamma Distribution

Govilt & Aggarwal (1983) use (2.4) to obtain a more convenient form of the mrl under the gamma distribution. Using shape parameter α and scale parameter λ, the numerator in (2.4), $\int_t^\infty uf(u)du$, simplifies to $1/\Gamma(\alpha)\int_t^\infty(w/\lambda)^\alpha exp[-w/\lambda]dw$. Under the integration by parts with $u = (w/\lambda)^\alpha$ and $dv = exp[-w/\lambda]dw$, the numerator is becomes, $(1/\Gamma(\alpha))(t/\lambda)^{\alpha-1}t^\alpha exp[-t\lambda] + \lambda\alpha\int_t^\infty f_{\mathbf{T}}(w)dw$. Substituting this expression to the numerator in (2.4), the mrl function is given by,

$$m(t) \quad = \quad \frac{t^\alpha exp\left[-\frac{t}{\lambda}\right]}{\lambda^{\alpha-1}\Gamma(\alpha)S_{\mathbf{T}}(t)} + \lambda\alpha - t \tag{2.5}$$

Gompertz Distribution

The Gompertz distribution with shape and scale parameters $\alpha, \lambda > 0$ respectively has survival function $S(t) = exp[(\lambda/\alpha)(1 - e^{\alpha t})]$. The numerator in (1.1) is written as: $\int_t^\infty S(u)du = \int_t^\infty exp[(\lambda/\alpha)(1-e^{\alpha u})]du = e^{(\lambda/\alpha)}\int_t^\infty exp[-(\lambda/\alpha)e^{\alpha u}]du$. If we let $z(u) = z = (\lambda/\alpha)e^{\alpha u}$, then $u = (1/\alpha)ln[(\lambda/\alpha)z] \Rightarrow du = (1/\alpha)(1/z)dz$. Denote the incomplete gamma function as $\Gamma_{inc}(a,t) = \int_t^\infty u^{a-1}e^{-u}du$ for $t, a \geq 0$, and define $z(t) = (\lambda/\alpha)e^{\alpha t}$. Substituting back into the survival function provides, $S(t) = e^{\lambda/\alpha}(1/\alpha)\int_{z(t)}^\infty z^{-1}e^{-z}dz = e^{\lambda/\alpha}(1/\alpha)\Gamma_{inc}(0, z(t))$.

$$m(t) = \frac{e^{\lambda/\alpha}\left(\frac{1}{\alpha}\right)\Gamma_{inc}(0,z(t))}{exp\left[\frac{\lambda}{\alpha}\left(1-e^{\alpha t}\right)\right]} = e^{z(t)}\left(\frac{1}{\alpha}\right)\Gamma_{inc}(0,z(t)) \tag{2.6}$$

Loglogistic Distribution

Consider the survival function for the loglogistic distribution with shape and scale parameters $\alpha, \lambda > 0$. The mean of the log-logistic distribution is only finite when the shape parameter is greater than 1, thus the mrl is only defined when $\alpha > 1$. The mrl for the log-logistic distribution is easily obtained from by simplifying (1.1) as is done by Gupta et al. (1999). The numerator in (1.1) is defined as $\int_t^\infty [1 + (u/\lambda)^\alpha]^{-1}$. Let $z(u) = z = ((u/\lambda)^\alpha)/(1 + (u/\lambda)^\alpha)$. Then $u = \lambda(z/(1-z))^{1/\alpha}$ and $du = (\lambda/\alpha)(z/(1-z))^{(1/\alpha)-1}(1/((1-z)^2))dz$. Applying the transformation, the integral becomes $(\lambda/\alpha)\Gamma(1-(1/\alpha))\Gamma(1/\alpha)\int_{z(t)}^1 \Gamma(1-(1/\alpha)+(1/\alpha))/(\Gamma(1-(1/\alpha))\Gamma(1\alpha)(1-z)^{(1-(1/\alpha))-1}z^{(1/\alpha)-1}dz$. Note that the integral is over a Beta kernel, thus the mrl function is given by,

$$m(t) = \left(\frac{\lambda}{\alpha}\right)\Gamma\left(1-\frac{1}{\alpha}\right)\Gamma\left(\frac{1}{\alpha}\right) S_{\mathbf{Z}}\left(z(t); 1-\frac{1}{\alpha}, \frac{1}{\alpha}\right)\left(1+\left(\frac{t}{\lambda}\right)^\alpha\right) \qquad (2.7)$$

Lognormal Distribution

The lognormal distribution has no closed form for the survival function, so (2.4) will be used to obtain the mrl function.

Using location μ and scale σ^2,the numerator in (2.4) is $(1/\sqrt{2\pi})\int_t^\infty (1/\sigma)$ $exp[-(1/2)((ln(u)-\mu)/\sigma)^2]du$. Let $z(u) = z = (ln(u)-\mu)/\sigma$, then $u = exp[z\sigma+\mu]$ and $du = \sigma exp[z\sigma+\mu]dz$. The numerator becomes, $(1/\sqrt{2\pi})\int_{z(t)}^\infty exp[-(1/2)z^2+z\sigma+\mu]dz =$ $e^{(\mu+(\sigma^2/2))}[1-\Phi((ln(t)-(\mu+\sigma^2))/\sigma)]$.

$$m(t) = \frac{e^{\left(\mu+\frac{\sigma^2}{2}\right)}\left[1-\Phi\left(\frac{ln(t)-(\mu+\sigma^2)}{\sigma}\right)\right]}{1-\Phi\left(\frac{ln(t)-\mu}{\sigma}\right)} - t \qquad (2.8)$$

Weibull Distribution

The Weibull distribution is closely related to the gamma distribution, so it is no

surprise that their mrl functions also behave similarly. Consider the Weibull distribution with shape parameter $\alpha > 0$ and scale parameter $\lambda > 0$. The numerator in (1.1) becomes $\int_t^\infty exp\left[-\left(\frac{u}{\lambda}\right)^\alpha\right] du$. Let $z(u) = z = u^\alpha$, then $u = z^{1/\alpha}$ and $du = \frac{1}{\alpha} z^{\frac{1}{\alpha}-1} dz$. Applying the transformation, we obtain, $\alpha^{-1} \int_{z(t)}^\infty z^{\alpha^{-1}-1} e^{-z/(\lambda^\alpha)} dz = \alpha^{-1}(\lambda^\alpha)^{\alpha^{-1}} \Gamma(\alpha^{-1}) \int_{z(t)}^\infty (z^{\alpha^{-1}-1} e^{-z/(\lambda^\alpha)})/((\lambda^\alpha)^{\alpha^{-1}} \Gamma(\alpha^{-1}))$. The last integral is exactly the survival function $S_{\mathbf{Z}}(z(t))$ with $Z \sim \Gamma\left(\frac{1}{\alpha}, \lambda^\alpha\right)$. Thus the mrl is given by,

$$m(t) = \frac{\left(\frac{\lambda}{\alpha}\right) \Gamma\left(\frac{1}{\alpha}\right) S_{\mathbf{Z}}(z(t))}{S_{\mathbf{T}}(t)} \tag{2.9}$$

Table 2.1, provides a summary of the possible shapes of the hazard rate and mrl functions for the distributions discussed in this section. The table shows how restricted these commonly used distribution are in modeling the mrl function. The gamma and Weibull are more versatile as they offer three potential shapes for the mrl function, but none of these shapes consider change points in the mrl function.

Exponentiated Weibull Distribution

Modifications of the Weibull model have been explored in order to develop a more flexible parametric model with regard to the shapes of the hazard and mrl functions; see Pham & Lai (2007) for an extensive list. We chose to focus of the exponentiated Weibull distribution which has closed form survival function and can take on a number of various forms for the mrl, namely monotone INC, monotone DCR, constant, UBT, or BT (Mudholkar & Strivasta, 1993). The distribution and mrl functions for the

Distribution	Density Function	Hazard Rate	Mean Residual Life
Gamma(γ, β) shape $\gamma > 0$ rate $\beta > 0$	$\frac{\beta^\gamma}{\Gamma(\gamma)} t^{\gamma-1} e^{-\beta t}$	$\gamma < 1$ DCR $\gamma = 1$ constant β $\gamma > 1$ INC	$\gamma < 1$ INC $\gamma = 1$ constant $1/\beta$ $\gamma > 1$ DCR
Gompertz(γ, λ) shape $\gamma > 0$ scale $\lambda > 0$	$\lambda\gamma e^{\lambda t} e^{\gamma} e^{\left(-\gamma e^{\lambda t}\right)}$	$\forall\gamma$ INC	$\forall\gamma$ DCR
Loglogistic(γ, λ) shape $\gamma > 0$ scale $\lambda > 0$	$\frac{(\gamma/\lambda)(t/\lambda)^{\gamma-1}}{[1+(t/\lambda)^\gamma]^2}$	$\gamma \leq 1$ DCR $\gamma > 1$ UBT	$\gamma \leq 1$ undefined $\gamma > 1$ BT
Lognormal(μ, σ) location $\mu \in \mathbb{R}$ scale $\sigma^2 > 0$	$\frac{e^{-\frac{(ln(t)-\mu)^2}{2\sigma^2}}}{t\sqrt{2\pi\sigma^2}}$	UBT	BT
Weibull(γ, λ) shape $\gamma > 0$ scale $\lambda > 0$	$\frac{\gamma}{\lambda}\left(\frac{t}{\lambda}\right)^{\gamma-1} e^{-(t/\lambda)^\gamma}$	$\gamma < 1$ DCR $\gamma = 1$ constant $1/\lambda$ $\gamma > 1$ INC	$\gamma < 1$ INC $\gamma = 1$ constant λ $\gamma > 1$ DCR

Table 2.1: Shapes of the mean residual life function for common parametric distributions. Shapes are described as being increasing (INC), decreasing (DCR), upside down bathtub (UBT), bathtub (BT), constant, or undefined.

exponentiated Weibull model are given by the following expressions:

$$\begin{aligned} F(t|\alpha,\theta,\sigma) &= \left[1-\exp\left(-\left(\frac{t}{\sigma}\right)^\alpha\right)\right]^\theta, \quad t>0, \alpha,\theta,\sigma>0 \qquad (2.10) \\ m(t|\alpha,\theta,\sigma) &= \frac{\int_t^\infty \left[1-\left[1-\exp\left(-\left(\frac{u}{\sigma}\right)^\alpha\right)\right]^\theta\right] du}{1-\left[1-\exp\left(-\left(\frac{t}{\sigma}\right)^\alpha\right)\right]^\theta} \end{aligned}$$

where α and θ are shape parameters and σ is a scale parameter. Note that σ, being a scale, will not play a role in determining the form of the hazard and mrl functions. Table 2.2 provides the parameter sets that result in each distinct shape for the mrl function.

Mudholkar & Strivasta (1993) provide a table similar to Table 2.2 for the hazard rate function for specific domains of α and θ. Xie et al. (2004) look at the role of the product of the shape parameters on the form of the hazard rate. Gupta & Akman (1995) prove that if the hazard rate function is BT and $h(0) > 1/m(0)$, then the mrl is

α	θ	$\alpha\theta$	form of mrl function
1	1	1	exponential distribution $\rightarrow$ constant mrl
–	1	–	weibull distribution $\rightarrow$ monotone (inc, dcr or constant) mrl
< 1	$\neq 1$	< 1	increasing
> 1	$\neq 1$	> 1	decreasing
> 1	< 1	< 1	UBT
< 1	> 1	> 1	BT

Table 2.2: Forms of MRL for Exponentiated Weibull Distribution

UBT, while $h(0) \leq 1/m(0)$ implies decreasing mrl function. Similarly if the hazard rate function is UBT and $h(0) > 1/m(0)$, then the mrl is BT, while $h(0) \geq 1/m(0)$ implies increasing function. Combining the aforementioned results, we are able to improve the table in Mudholkar & Strivasta (1993) to specify the exact shape of the mrl function for particular values of α and θ in conjunction with the value of the product of the parameters, yielding Table 2.2.

2.2 Nonparametric mixture model for MRL inference

In this section, we discuss our modeling methods for obtaining inference for the mrl function. Section 2.2.1 motivates the use of a nonparametric Dirichlet process mixture model (DPMM). We provide the model structure, and discuss the choice of kernel distribution. In Section 2.2.2, we discuss prior specification. Section 2.2.3 provides the techniques used to obtain posterior inference for the mixture distribution and functionals thereof.

2.2.1 Model formulation

When the data exhibits unusual distributional features such as multi-modality or skewness, parametric models tend to fail to capture these important features. A way to go about this issue is to use a mixture model that combines a number of distributions that we will refer to as components of the model. The question then becomes how many components should be used and how should they be combined together? These concerns can be addressed by bringing in a nonparametric aspect to the model, in particular, to the weights of each component and to the number of components.

We use a Dirichlet process (DP) prior for the mixing distribution resulting in a DP nonparametric mixture model, $f(t|G) = \int k(t|\boldsymbol{\theta})dG(\boldsymbol{\theta})$, for the density of the survival distribution. In practice, an appropriately supported kernel distribution, $k(t|\boldsymbol{\theta})$, is selected, and a $DP(\alpha, G_0)$ prior is assigned to G. The DP is a stochastic process with random sample paths that are distributions (Ferguson, 1973). Thus a realization from the DP provides a random cdf sample path. The G_0 parameter is the baseline or centering distribution, while α is a precision parameter; the larger the value of α the closer the DP sample path is to the centering distribution. We use the stick-breaking (SB) constructive definition of the DP defined by Sethuraman Sethuraman (1994), which states that a sample $G(\cdot)$ from $DP(\alpha, G_0)$ is almost surely of the form $\sum_{l=1}^{\infty} w_l \delta_{\boldsymbol{\theta}_l}(\cdot)$ where $\delta_{\boldsymbol{\theta}_l}(\cdot)$ is a point mass at $\boldsymbol{\theta}_l$. The $\boldsymbol{\theta}_l$, for all $l \in \{1, 2, ...\}$, are i.i.d. samples from the baseline distribution, G_0, and the w_l are the corresponding weights constructed sampling i.i.d. latent variables $v_r \sim \text{Beta}(1, \alpha)$, for all $r \in \{1, 2, ...\}$, then

$w_1 = v_1$ and $w_l = v_l \prod_{r=1}^{l-1}(1 - v_r)$, for $l \in \{2, 3, ...\}$.

We use the truncated version of the SB constructive definition of the DP, $G_L(\cdot) = \sum_{l=1}^{L} p_l \delta_{\boldsymbol{\theta}_l}(\cdot)$, where $\boldsymbol{\theta}_l \overset{iid}{\sim} G_0$ for $l = 1, ..., L$, and $p_1 = v_1$, $p_l = v_l \prod_{r=1}^{l-1}(1 - v_r)$, where $v_r \overset{iid}{\sim} Beta(1, \alpha)$ for $r = 1, ..., L - 1$, and $p_L = 1 - \sum_{l=1}^{L-1} p_l$. The model is given by:

$$f(t|G) \quad \sim \quad \int k(t|\boldsymbol{\theta}) dG(\boldsymbol{\theta}) = \sum_{l=1}^{L} p_l k(t|\boldsymbol{\theta}_l) \tag{2.11}$$

where p_l for $l = 1, ..., L$ are the weights obtained via the SB construction, described above, corresponding to the component $\boldsymbol{\theta}_l$ and L is the total number of components in the mixture model. Technically, since the number of components is predetermined there is no nonparametric element to the number of components. However, L is generally chosen to overestimate the true number of components, so that the number of components suggested by the data is captured by the model. In fact, many of the components will just be assigned a probability that is virtually zero. The number of components for the finite sum DP approximation can be found using $E(\sum_{l=1}^{L} p_l) = 1 - (\alpha/(\alpha + 1))^L$, in particular, solving for L in $(\alpha/(\alpha + 1))^L = \epsilon$ for small $\epsilon > 0$.

Our primary aim in this paper is to present a Bayesian model that provides both flexible and practical inference for the mean residual life function. The mrl function is defined by the distribution function and vice versa, thus we advocate for the nonparametric Dirichlet process mixture which provides flexible modeling on the distribution function. We obtain inference for the mrl function via fitting a DPMM on the distribution function. Since our interest is inference for the mrl function, it is necessary

that the mrl function of the DPMM exists and is finite. A sufficient condition for the finiteness of the mrl function for a given kernel distribution is provided later in this section. Although we do not place a prior directly on the mrl function, from the lemma stated at the end of this section, we can use prior knowledge of the tail behavior to select an agreeable kernel distribution. Essentially, we can induce a prior for the mrl function through the tail behavior. We complete the model formulation by addressing the aspect of dependency within $\boldsymbol{\theta}$. We consider modeling the dependence between the kernel parameters by using a joint baseline distribution, G_0, in the DPMM.

Care is needed in selecting a kernel distribution to ensure the mean of the DPMM is finite, $E(T|G) < \infty$ where $T \sim f(t|G) = \int_{\Theta} k(t|\boldsymbol{\theta}) dG(\boldsymbol{\theta})$. We provide sufficient conditions to ensure finiteness of the mean by following the argument in Theorem 3 of Ferguson (1973). Let $Z = E(T|G)$, where T is a $\mathbb{R}^+$ random valued. Recall that if $E(Z) < \infty$, then $Z < \infty$ almost surely. Hence we need to show $E(Z) < \infty$. Observe that $Z = \int_T t \int_{\Theta} k(t|\boldsymbol{\theta}) dG(\boldsymbol{\theta}) dt = \int_{\Theta} E(T|\boldsymbol{\theta}) dG(\boldsymbol{\theta}) = \sum_{j=1}^{\infty} w_j W(\boldsymbol{\theta}_j)$ where the w_j are the weights arising from the stick-breaking process and $W(\boldsymbol{\theta}_j) = E(T|\boldsymbol{\theta}_j)$. Define the sequence of $\mathbb{R}^+$ valued random variables $Z_n = \sum_{j=1}^{n} w_j W(\boldsymbol{\theta}_j)$, for $n \in \{1, 2, ...\}$. Note that Z_n is an almost surely increasing sequence and $Z_n \overset{a.s.}{\to} Z$. Thus by the monotone convergence theorem, $E(Z_n) \overset{a.s.}{\to} E(Z)$. Now, we can write $E(Z) = E\left[\sum_{j=1}^{\infty} w_j W(\boldsymbol{\theta}_j)\right]$. Using the independence of w_j and $W(\boldsymbol{\theta}_j)$, the expectation becomes $\sum_{j=1}^{\infty} E(w_j) E(W(\boldsymbol{\theta}_j)) = \sum_{j=1}^{\infty} E(w_j)(\int_{\boldsymbol{\Theta}} W(\boldsymbol{\theta}_j) dG_0(\boldsymbol{\theta}_j))$. Upon integration over $\boldsymbol{\theta}_j$ in the last expression, the resulting expression is free of the subscript j and is a function of the parameters of the baseline distribution, G_0, with parameters $\boldsymbol{\psi}$. Define $A(\boldsymbol{\psi}) = \int_{\boldsymbol{\Theta}} W(\boldsymbol{\theta}_j) dG_0(\boldsymbol{\theta}_j)$,

then $E(Z)$ becomes $A(\boldsymbol{\psi})\sum_{j=1}^{\infty} E(w_j) = A(\boldsymbol{\psi})$, since $\sum_{j=1}^{\infty} E(w_j) = 1$. Therefore, if $A(\boldsymbol{\psi}) = \int_{\boldsymbol{\Theta}} W(\boldsymbol{\theta}) dG_0(\boldsymbol{\theta}) < \infty$, then $E(T|G) < \infty$ almost surely. In words, the finiteness of the expected value of the mean of the kernel distribution with respect to the baseline distribution guarantees finiteness of the first moment of the DPMM.

Common kernel distributions in modeling survival data include the lognormal, Weibull, and gamma distributions. First, consider the lognormal kernel with $W(\boldsymbol{\theta}) \equiv E(T|\mu, \sigma) = exp(\mu + \sigma^2/2)$ and $G_0 = N(\mu|\lambda, \tau^2)\Gamma^{-1}(\sigma^2|a, \rho)$ where ρ denotes the scale parameter. Here, we have $A(\boldsymbol{\psi}) = \int_{-\infty}^{\infty} exp(\mu) N(\mu|\lambda, \tau^2) d\mu \int_0^{\infty} exp(\sigma^2/2)\Gamma^{-1}(\sigma^2|a, \rho) d\sigma^2$. The first integral is clearly finite, but the second integral would require a bound on σ^2 that would depend on ρ in order to be finite. We can get around the restriction by using a gamma baseline distribution, but the rate parameter of the gamma distribution would have to be truncated below at $1/2$. In either case, we will not have conjugacy.

If we use a Weibull kernel with $W(\boldsymbol{\theta}) \equiv E(T|\gamma, \sigma) = \sigma^{1/\gamma}\Gamma(1 + 1/\gamma)$ and $G_0 = \Gamma(\gamma|a, \rho \text{ (rate)})\Gamma^{-1}(\sigma|c, \lambda)$, $A(\boldsymbol{\psi})$ is given by $\int_0^{\infty}\int_0^{\infty} \sigma^{1/\gamma}\Gamma(\gamma|a, \rho)\Gamma^{-1}(\sigma|c, \lambda) d\sigma d\gamma$. We can integrate out σ without difficulty by recognizing another gamma distribution, however, the finiteness of the first integral requires $\gamma > 1/c$. This is not an unreasonable restriction for $c > 1$, allowing for decreasing and/or unimodal components in the mixture, however, the second integral yields more restrictions. We can obtain finiteness by constructing a function, $g(\gamma)$ that is greater than $\Gamma(1 + 1/\gamma)\Gamma(c - 1/\gamma)\lambda^{1/\gamma}$ for $\gamma > 1/c$, where the second part of the expression is a result of the first integral. Note that $\Gamma(c - 1/\gamma)$ goes to ∞ as $\gamma \to 1/c$, which causes problems in convergence. We can get around this by making γ bounded below by a value just slightly larger than $1/c$.

By showing $E(g(\gamma)) < \infty$ with respect to the distribution $\Gamma^{-1}(\gamma|a,\rho)$, then we know that $E(\Gamma(1+1/\gamma)\Gamma(c-1/\gamma)\lambda^{1/\gamma}) < \infty$. The idea is that $E(g(\gamma))$ is easily computed, so it is convenient to use a function of the form $g(\gamma) = \lambda^v(1/\gamma^w + \Gamma(c))$ for $v, w > 0$. Using this function form, will result in a restriction on the shape parameter, a, that will depend on w. Since w and a are both fixed parameters, this is not an unreasonable restriction, and slice sampling may be used in the MCMC for γ, but the sampling from the posterior conditional of ρ will require a Metropolis-Hastings step.

Consider a gamma kernel distribution with $W(\boldsymbol{\theta}) \equiv E(T|\alpha,\beta) = \alpha/\beta$ and $G_0 = f(\alpha|\boldsymbol{\omega})\Gamma(\beta|c,\lambda)$. We can separate the integrals in $A(\boldsymbol{\psi})$ to be $\int_0^\infty \alpha f(\alpha|\boldsymbol{\omega})d\alpha$ $\int_0^\infty \beta^{-1}\Gamma^{-1}\beta^{-1}\Gamma(\beta|c,\lambda)d\beta$, where the first integral is simply $E(\alpha)$ with respect to $f(\alpha|\boldsymbol{\omega})$ and the second integral is $E(\beta^{-1})$ where $\beta^{-1} \sim \Gamma^{-1}(\beta|c,\lambda)$. Therefore as long as we choose $f(\alpha|\boldsymbol{\omega})$ to have finite mean and set $c > 1$, then $A(\boldsymbol{\psi}) < \infty$. We do not have conjugacy in the MCMC for α, but our parameter restriction is minimal. The mean and variance of the gamma distribution are not independent, so we might consider a joint G_0. A convenient option would be to model using $\boldsymbol{\theta} = (\theta = log(\alpha), \phi = log(\beta))'$ and place a bivariate normal distribution on $G_0 = N_2(\boldsymbol{\theta}|\boldsymbol{\mu},\boldsymbol{\Sigma})$. Now, we have $A(\Psi) = \int\int e^{\theta-\phi}N_2((\theta,\phi)'|\boldsymbol{\mu},\boldsymbol{\Sigma})d\theta d\phi = E(e^{\theta-\phi})$, which can be obtained from the moment generating function of the bivariate normal, $E(e^{\boldsymbol{\theta}'t})$, with $t = (1,-1)'$. Hence $A(\Psi) = e^{(1,-1)\boldsymbol{\mu}+(1/2)(1,-1)\boldsymbol{\Sigma}(1,-1)'}$, which is finite for any $\boldsymbol{\mu} \in \mathbb{R}^2$ and any non-negative definite $\boldsymbol{\Sigma} \in \mathbb{R}^{2x2}$.

Another important consideration in the choice of the mixture kernel is the shape of the mixture mrl function relative to the mrl function of the kernel distribution.

The following lemma, whose proof can be found in Appendix B.1, provides a result on the tail behavior of the mrl function for the mixture distribution.

Lemma 1. Let $m(t|\boldsymbol{\theta})$ be the parametric mrl function of the corresponding to the DPMM kernel and $m(t|G_L)$ be the mrl function of the mixture, where G_L is the truncated approximation to the mixing distribution. Then,

1. If $\lim_{t\to\infty} m(t|\boldsymbol{\theta}) = \infty \ \forall \boldsymbol{\theta} \in \boldsymbol{\Theta}$, then $\lim_{t\to\infty} m(t|G_L) = \infty$.

2. If $\lim_{t\to\infty} m(t|\boldsymbol{\theta}) = 0 \ \forall \boldsymbol{\theta} \in \boldsymbol{\Theta}$, then $\lim_{t\to\infty} m(t|G_L) = 0$.

Taking into account the condition for $E(T|G) < \infty$ and the lemma above, the gamma distribution emerges as the more suitable choice for the kernel distribution. Referring back to Table 2.1, we can see that a lognormal kernel will always result in a mrl that goes to infinity in tail. A Gompertz kernel would result in a mrl that tends to zero in the tail. If there is prior knowledge regarding the tail behavior of the mrl, then it would make sense to choose a kernel that has a corresponding mrl with agreeable tail behavior. However, in the case that prior knowledge of the mrl tail behavior is not known, the gamma or Weibull kernel would be appropriate choices. Per our discussion regarding the sufficient condition for existent and finite mrl, the Weibull requires more restriction on the support of the model parameters.

Another important model property to investigate is the matter of denseness. Let $\mathscr{F}$ represent the space of absolutely continuous distribution functions on $\mathbb{R}^+$ with finite mean. Formally, a class of distributions, $\mathcal{C}$, is said to be dense in $\mathscr{F}$, if for any

distribution function, F, there exists a sequence of distribution functions, $\{F_n\} \subseteq \mathcal{C}$, that converges to F. The type of convergence implies a measure of distance between the limiting sequence and F. Johnson & Taaffe (1998) show denseness of infinite and finite mixtures of Erlang distributions on the space of cumulative distribution functions having support $[0, \infty)$. They provide details for weak convergence and make an argument for uniform convergence. A mixture of Erlang distributions is in the class of gamma mixtures, so the result holds true for gamma mixtures as well. More interesting from our prospective, however, is the denseness of the resulting mrl function. In appendix B.2, we show that for any continuous mrl function, $m(t)$, there exists a corresponding sequence of mrl functions for a mixture of gamma distributions, $\{m_n(t)\}$ such that for any $t_0 \geq 0$, $\lim_{n\to\infty} m_n(t_0) = m(t_0)$, converges pointwise, providing the denseness result in Lemma 2.

Lemma 2. The set of mrl functions corresponding with the class of gamma mixture distributions is dense, in the pointwise sense, in the space of continuous mrl functions.

Finally we turn to the choice of G_0. We seek to be more general in our modeling by using a dependent G_0 for the parameters of the gamma kernel. This allows the model to capture correlations between the kernel parameters. Note that, once one leaves the setting of normal mixtures, the kernel parameters are not naturally separated as location and scale parameters, making the assumption of an independent G_0 more restrictive than in mixing with Gaussian kernels. Recall that modeling the shape and

rate parameters of the gamma kernel on the log-scale allow us to use a bivariate normal for G_0, and we only need a non-negative definite 2×2 covariance matrix in G_0 to satisfy the sufficient condition. In the remainder of this paper, we will refer to this model as the gamma DPMM and assume a bivariate normal G_0 on the log-scale of the gamma kernel parameters.

2.2.2 Prior specification

When it comes to prior specification often there is not much prior knowledge on the behavior of the population of interest, but typically the researcher will have at least somewhat of an idea of the range and midpoint/midrange of the population. We would want to set our priors to have a prior predictive distribution that encompasses this range. One way to favor a prior predictive distribution that covers the range of the data is to imagine one relatively dispersed kernel component that is centered at the midrange with 2 standard deviations either way representing the prior range. In the data illustrations in Section 2.3, we set the range to about 2 times the data range. We can then divide the range by 4 and square that value to get the prior variance of the data. Specifically, $(range(T)/4)^2 \approx Var(T)$. This method can be implemented when fitting a gamma DPMM. We place the following distributions on the hyperparameters: $\boldsymbol{\mu} \sim N_2(a_\mu, B_\mu)$ and $\boldsymbol{\Sigma} \sim IWish(a_\Sigma, B_\Sigma)$. Making use of the moment generating function of the bivariate normal distribution, the independence property of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, and the first order Taylor expansion for $\exp((1/2)t_\cdot'\boldsymbol{\Sigma}t_\cdot)$ centered around $E((1/2)t_\cdot'\boldsymbol{\Sigma}t_\cdot)$,

we approximate $Var(T)$ as follows:

$$\begin{aligned} Var(T) &= Var(E(T|e^{\theta}, e^{\phi})) + E(Var(T|e^{\theta}, e^{\phi})) \qquad (2.12) \\ &= E(Var(e^{\theta-\phi}|\boldsymbol{\mu}, \boldsymbol{\Sigma})) + Var(E(e^{\theta-\phi}|\boldsymbol{\mu}, \boldsymbol{\Sigma})) + E(E(e^{\theta-2\phi}|\boldsymbol{\mu}, \boldsymbol{\Sigma})) \\ &= E(e^{t_1'\boldsymbol{\mu}})E(e^{t_1'\boldsymbol{\Sigma}t_1}) + E(e^{t_2'\boldsymbol{\mu}})E(e^{t_2'\boldsymbol{\Sigma}t_2}) - E^2(e^{t_3'\boldsymbol{\mu}})E^2(e^{t_3'\boldsymbol{\Sigma}t_3}) \\ &\approx e^{t_1'a_\mu + (1/2)t_1'B_\mu t_1} e^{\frac{(1/2)t_1'B_\Sigma t_1}{a_\Sigma - d - 1}} + e^{t_2'a_\mu + (1/2)t_2'B_\mu t_2} e^{\frac{(1/2)t_2'B_\Sigma t_2}{a_\Sigma - d - 1}} \\ &\quad - e^{2(t_3'a_\mu + (1/2)t_3'B_\mu t_3)} e^{2\frac{(1/2)t_3'B_\Sigma t_3}{a_\Sigma - d - 1}} \end{aligned}$$

where $t_1 = (1, -2)'$, $t_2 = (2, -2)'$, $t_3 = (1, -1)'$, and $d \times d$ is the dimension of $\boldsymbol{\Sigma}$, specifically, $d = 2$. We set $a_\Sigma = 4$, which is the smallest degrees of freedom for the inverse Wishart distribution that has finite mean. If we place priors of the form $B_\mu = ((b_\mu', 0)', (0, b_\mu'))$ and $B_\Sigma = ((b_\Sigma', 0)', (0, b_\Sigma'))$ for $b_\mu', b_\Sigma' > 0$, the expression is simplified, however, we still have four parameters to specify. One solution would be to incorporate the marginal expectation:

$$E(T) \approx e^{t_3'a_\mu + \frac{1}{2}t_3'B_\mu t_3 + \frac{1}{2}t_3'B_\Sigma t_3} = e^{(a_{\mu_1} - a_{\mu_2}) + b_\mu' + b_\Sigma'} \qquad (2.13)$$

where upon applying our earlier assumptions, we get the last expression in (2.13) with $a_\mu = (a_{\mu_1}, a_{\mu_2})'$. We can further simplify by setting $b_\mu' = b_\Sigma'$ resulting in two equations with three unknowns. Next, we can allocate a percentage of the marginal expectation (2.13) to $exp(a_{\mu_1} - a_{\mu_2})$ and a percentage of the marginal variance (2.12) to $exp(a_{\mu_1} - 2a_{\mu_2})$, solving for a_{μ_1} and a_{μ_2}. Finally, we can return to (2.12) and solve for b_μ' and b_Σ'.

Regarding the prior for α, we consider the relationship between the number of distinct components, n^*, and the value of α. In general, the number of distinct

components is large for large α and small for small α. If the data set is moderately large, $E(n^*|\alpha) \approx \alpha log\left(\frac{\alpha+n}{\alpha}\right)$ can be used to suggest an appropriate range of α values.

This approach to prior specification is based on a small amount of prior information regarding the survival distribution. In general, we recommend studying the implied prior distribution for important survival functionals, including prior point and interval estimates for the mrl function.

2.2.3 Posterior inference

Posterior simulation is simplified by truncating the mixing distribution $G_L(\cdot) \approx G(\cdot)$. Before we introduce $\boldsymbol{\theta}_l$, the first two levels of the model are,

$$\begin{aligned} t_i|\zeta_i &\overset{ind}{\sim} K(t_i|\zeta_i),\ i = 1, ..., n \\ \zeta_i|\boldsymbol{p}, \boldsymbol{\theta} &\overset{iid}{\sim} G_L,\ i = 1, ..., n \end{aligned}$$

where $\boldsymbol{p} = (p_1, ..., p_L)$ are the weights corresponding to the weights, $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_L)$. By marginalizing over the ζ_i we obtain the finite mixture model in (2.11). Now we can augment the model with configuration variables $\mathbf{w} = (\mathrm{w}_1, ..., \mathrm{w}_n)$ such that $\mathrm{w}_i = l$ iff $\zeta_i = \boldsymbol{\theta}_l$. Using a gamma kernel ($K$) and bivariate normal baseline (G_0) with $\boldsymbol{\theta}_l = (\theta_l, \phi_l)'$, the hierarchical model is given by,

$$\begin{aligned} t_i|\boldsymbol{\theta}, \mathrm{w}_i &\overset{ind}{\sim} \Gamma(t_i|e^{\theta_{\mathrm{w}_i}}, e^{\phi_{\mathrm{w}_i}}),\ i = 1, ..., n \\ \mathrm{w}_i|\mathbf{p} &\overset{iid}{\sim} \sum_{l=1}^{L} p_l \delta_l(\mathrm{w}_i),\ i = 1, ..., n \\ \mathbf{p}|\alpha &\sim f(\mathbf{p}|\alpha) \\ (\theta_l, \phi_l)'|\boldsymbol{\mu}, \boldsymbol{\Sigma} &\overset{iid}{\sim} N_2((\theta_l, \phi_l)'|\boldsymbol{\mu}, \boldsymbol{\Sigma}),\ l = 1, ..., L \end{aligned} \tag{2.14}$$

where $f(\boldsymbol{p}|\alpha) = \alpha^{L-1}p_L^{\alpha-1}(1-p_1)^{-1}(1-(p_1+p_2))^{-1} \times ... \times (1-\sum_{l=1}^{L-2} p_l)^{-1}$ is a special case of the generalized Dirichlet distribution (Conner & Mosemann, 1969). Here, we use conjugate priors, $\alpha \sim \Gamma(\alpha|a_\alpha, b_\alpha(\text{rate}))$, $\boldsymbol{\mu} \sim N_2(\boldsymbol{\mu}|a_\mu, B_\mu)$, and $\boldsymbol{\Sigma} \sim IWish(\boldsymbol{\Sigma}|a_\Sigma, B_\Sigma)$.

Now, we can utilize a blocked Gibbs sampler (Ishwaran & James, 2001) to obtain samples from the posterior distribution $p(\boldsymbol{\theta}, \mathbf{w}, \boldsymbol{p}, \boldsymbol{\psi}, \alpha|data)$ where $\boldsymbol{\psi} = (a_\mu, B_\mu, B_\Sigma)$. We have Gibbs steps for all parameters except $\boldsymbol{\theta}$, for which we use a Metropolis-Hastings step. The specifics of the posterior sampling method for the gamma DPMM are provided in Appendix C.

The posterior samples for $G_L \equiv (\boldsymbol{p}, \boldsymbol{\theta})$ can be used to obtain inference for the density, survival, and hazard functions at any time point t, by directly evaluating the expressions for these functions under the gamma DPMM. Obtaining the mrl function must be done by numerical integration approximation for the integral over the survival function. From (2.1) we know that the mrl function at 0 returns the expected survival time, $m(0) = \mu$. Hence, the mrl function can be written alternatively as follows:

$$m(t) \quad = \quad \frac{\int_t^\infty S(u)du}{S(t)} = \frac{\int_0^\infty S(u)du - \int_0^t S(u)du}{S(t)} = \frac{\mu - \int_0^t S(u)du}{S(t)} \qquad (2.15)$$

We can avoid having to truncate the upper bound of the integration in the numerator in (1.1) by using the form of the mrl function as described in (2.15). We obtain posterior point and interval estimates for the mrl function by evaluating expression (2.15) at the posterior samples from the MCMC. We do this over a grid of survival times, $t_{0,j}$ for $j = 1, ..., m$. The survival function is monotone decreasing so the trapezoid technique is an appropriate method of approximating the integral in the mrl. We evaluate the mrl

at the first grid point by $m(t_{0,1}|G_L) = [E(T|G_L) - 0.5(t_{0,1}(1 + S(t_{0,1}|G_L))]/S(t_{0,1}|G_L)$ and use the following expression for $j = 2, ..., m$:

$$m(t_{0,j}|G_L) = \frac{E(T|G_L) - \frac{1}{2}\left(t_{0,1}(1+S(t_{0,1}|G_L)) + \sum_{i=2}^{j}(t_{0,j} - t_{0,j-1})(S(t_{0,j}|G_L) + S(t_{0,j-1}|G_L))\right)}{S(t_{0,j}|G_L)}$$

We save a lower and upper quantile along with the median at each grid point for each mixture functional to obtain (point-wise) posterior point and interval estimates.

2.3 Data examples

In Section 2.3.1, we use simulated data to illustrate the ability of the gamma DPMM to capture non-standard mrl function shapes as well as the correlation between kernel parameters θ and ϕ. In Section 2.3.2, we fit a gamma DPMM as well as an exponentiated Weibull model to a data set involving survival times for subjects from two groups, including formal model comparison between the two models. In Section 2.3.3, we provide results of fitting the gamma DPMM to a data set of two groups both containing right censored data values.

2.3.1 Simulation examples

In this section, we will work with two simulated data sets. The first data set consists of 200 simulations from a mixture distribution of four gamma components in which the shape and scale parameters are positively associated: $T_1 \sim 0.35\Gamma(10, 0.5) + 0.4\Gamma(20, 1) + 0.15\Gamma(30, 5) + 0.1\Gamma(40, 8)$. We fit a gamma DPMM with priors $a_\mu =$

$(1.6, 0.4)$, $B_\mu = B_\sigma = ((0.39, 0)', (0, 0.39)')$, $a_\alpha = 2$, $b_\alpha = 1$, and $L = 40$. The effective posterior sample size is 2000. Results are shown in Figure 2.1.

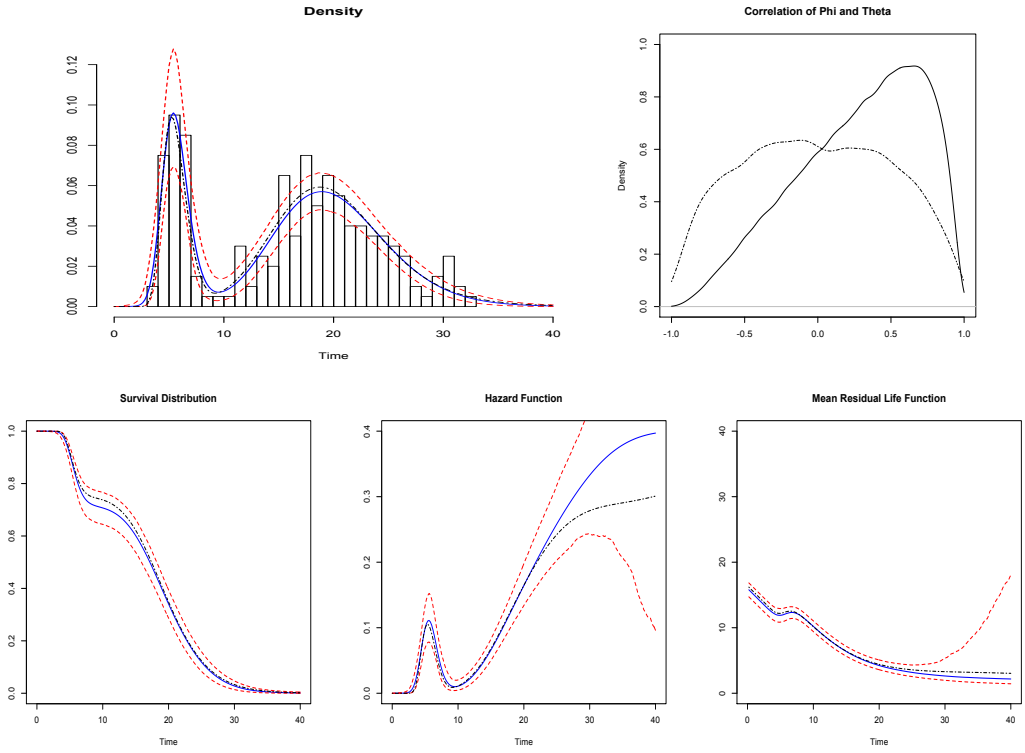

Figure 2.1: Simulation example 1. Point (solid) and interval (dashed) estimates of lifetime for the density (top left) overlaying the sample histogram and actual population density (dot-dashed), posterior (solid) and prior (dot-dashed) distribution of the correlation between θ and ϕ (top right), survival (lower left), hazard rate (lower middle), and mrl (lower right) functions of the two experimental groups under the gamma DPMM.

The red dashed line represents the 95% interval estimates, the blue solid line is the point estimate, and the black dot-dashed line is the truth of the appropriate functional. We can see that truth is well within the interval estimate, moreover the point estimate is close to the truth. The correlation shows that the model (black solid) is able to capture the positive relationship between the parameters, even though the

prior (black dot-dashed line) is evenly dispersed about the situation of zero correlation.

The second data set consists of 100 simulations from a distribution with negative correlation between the shape and rate parameters: $T_2 \sim 0.3\Gamma(15, 0.2)+0.25\Gamma(12, 0.5)+ 0.35\Gamma(8, 2) + 0.1\Gamma(3, 6)$. This population was chosen to test the model's ability to separate modes that are close together, as well as model a distribution with a long tail. A gamma DPMM was fit to the data with priors $a_\mu = (2.4, -1)$, $B_\mu = B_\sigma = ((0.18, 0)', (0, 0.18)')$, $a_\alpha = 2$, $b_\alpha = 1$, and $L = 40$. The effective posterior sample size is 2000. Results are shown in Figure 2.2. Once again the point estimates are close to the truth even in the tail where less data is available. The uncertainty band in the mrl plot has a large upper bound, which is likely an effect of the sparse data in the tail. Numerical instability in the computation of the mrl function is also likely to contribute to the large upper bound. The model picks up a strong negative correlation between the parameters.

2.3.2 Analysis of survival times of rats (ad libitum vs restricted eating)

This data set, considered earlier in Berger et al. (1988), is used to illustrate posterior inference under both an exponentiated Weibull model and the gamma DPMM. The data consists of survival times of rats in two experimental groups. The first group (Ad libitum group) is comprised of 90 rats who were allowed to eat freely as they desired. The second group (Restricted group) is comprised of 106 rats that were placed on a restricted diet.

Under the exponentiated Weibull model, we used the $P_1 = 10\%$, $P_2 = 50\%$,

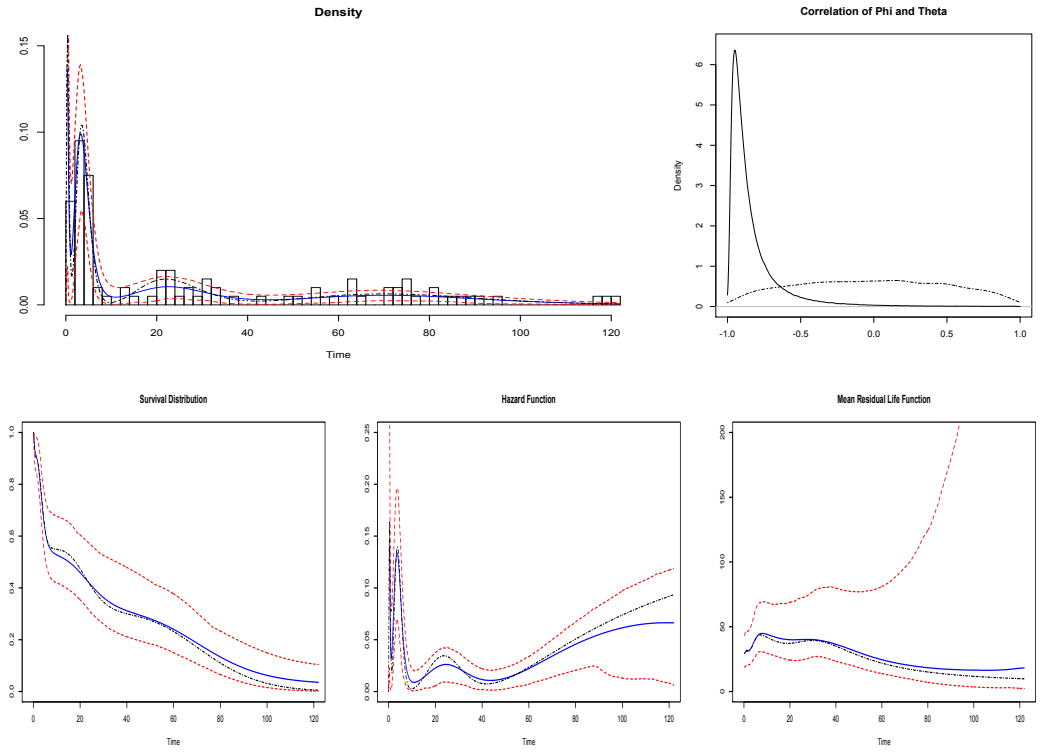

Figure 2.2: Simulation example 2. Point (solid) and interval (dashed) estimates of lifetime for the density (top left) overlaying the sample histogram and actual population density (dot-dashed), posterior (solid) and prior (dot-dashed) distribution of the correlation between θ and ϕ (top right), survival (lower left), hazard rate (lower middle), and mrl (lower right) functions of the two experimental groups under the gamma DPMM.

and $P_3 = 90\%$ quantiles of the data to obtain a system of three equations from the distribution function: $P = [1 - exp(-(Q/\sigma)^{\alpha})]^{\theta}$ where P is the percentile and Q is the survival time representing that quantile. The system of equations is solved to obtain prior means for α, σ and θ. For simplicity, exponential priors were placed on these parameters. The restricted group had respective quantile values of ($Q_1 = 1.55$, $Q_2 = 2.84$, $Q_3 = 3.34$). If we set $\alpha = 2$, $\theta = 5$, and $\sigma = 2$, then the corresponding quantiles are given as $Q_1' = 1.99$, $Q_2' = 2.85$, and $Q_3' = 4.07$ which we considered to

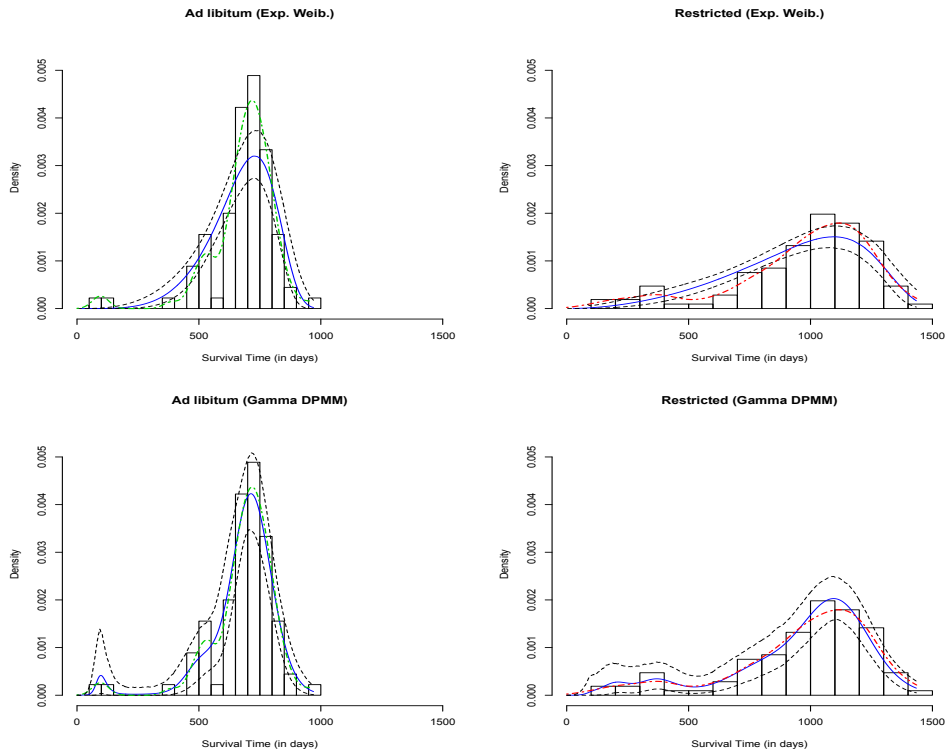

Figure 2.3: Relative frequency histogram and densities of lifetime (in days) of the two experimental groups (Ad libitum is left and Restricted is right) along with posterior mean and 95% interval estimates for the density functions under the exponentiated Weibull model (top) and the gamma DPMM (bottom).

be reasonably close to the observed quantiles. Therefore, we set hyper-parameters as $a_\alpha = 2$, $a_\theta = 5$, and $a_\sigma = 2$. Following the same methodology for the ad libitum group, we set the hyper-parameters as $a_\alpha = 4$, $a_\theta = 1$, and $a_\sigma = 2$. Posterior results were obtained using a Metropolis-Hastings algorithm in the MCMC with a trivariate normal proposal distribution on the log-scale. Point and interval estimates of the density function are plotted in the top row of Figure 2.3.

Prior specification for the gamma DPMM were determined using methods described in section 3.2 by allocating 60% of the marginal mean to $exp(a_{\mu_1} - a_{\mu_2})$ and

2.5% of the marginal variance to $exp(a_{\mu_1} - 2a_{\mu_2})$. For the restricted group, we use $a_\mu = (4.1, 3.6)$, $B_\mu = B_\sigma = ((0.1, 0)', (0, 0.1)')$, $a_\alpha = 2$, $b_\alpha = 1$, and $L = 40$. For the ad libitum group, we use $a_\mu = (4.16, 3.8)$, $B_\mu = B_\sigma = ((0.095, 0)', (0, 0.095)')$, $a_\alpha = 2$, $b_\alpha = 1$, and $L = 40$. The effective posterior sample size under both models are 2000 for each group. The bottom row in Figure 2.3 depicts the posterior estimates for the densities for the two groups under the DP mixture model.

In Figure 2.3, we note that the parametric model has some trouble capturing some of the characteristics of the data. In the ad libitum group (upper left) a minor mode is suggested just below the 200^{th} day. The unimodality of the exponentiated Weibull distribution makes it impossible for the parametric model to capture this shape. We note that the model tries to by reaching the tail of the estimated density out to these values, but this is at a cost of underestimating the density where most of the data exist, and overestimating the density where there is no data at all. There are many regions where the data and the density of the data (green dot-dashed) do not even fall within the interval estimates (black dashed). If we compare to how well the nonparametric model (lower left) performs we see quite a bit of improvement. The extra structure at the lower survival times is now being captured without the consequences of modeling poorly in other regions of the data. The data density remains within the interval estimates over the entire range of the data. We see similar results for the restricted group, which has a large left skew with a slight mode in the far tail. The exponentiated Weibull model (upper right) is able to model some of the skewness, but again runs into trouble by smoothing over obvious peaks and valleys. Again there are a number of regions in

which the density of the data (red dot dashed) is not contained in the interval estimates of the model. The gamma DPMM (lower right) is able to capture the peaks and valleys that the exponentiated Weibull model could not. There is a slight discrepancy from the point estimate (blue solid) and the density of the data around 1250 days. Nonetheless, the data density remains within the interval estimates of the model.

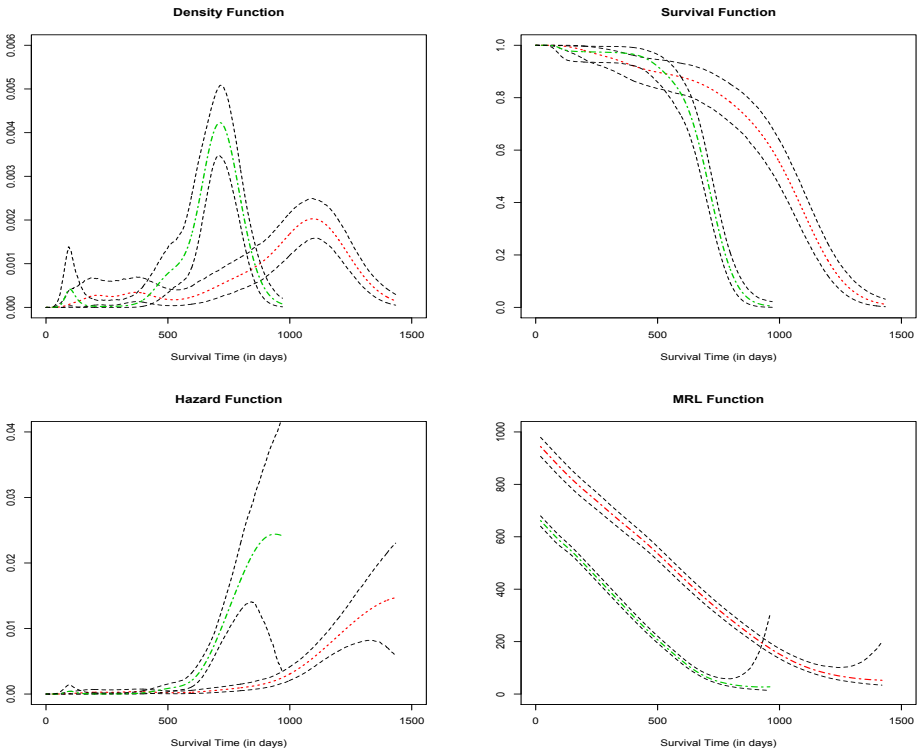

Figure 2.4: Point and 95% interval estimates of lifetime (in years) for the density (top left), survival (top right), and hazard rate (lower left), and point and 80% interval estimates for the mrl (lower right) functions of the two experimental groups under the gamma DPMM.

By comparing the densities under the two models, there is clear evidence that the nonparametric gamma DPMM is superior to the exponentiated Weibull model. Therefore, we will use the results under the nonparametric gamma DPMM to compare

the mrl functions under the two groups. In Figure , we plot point and interval estimates of the posterior density functions (upper left), survival functions (upper right), hazard functions (lower left), and mrl functions (lower right) for both the ad libitum (green) and restricted (red) groups. Note that the interval estimates for the mrl function are 80% probability bands as opposed to the other interval estimates, which are 95% probability bands. The reason for this is to reduce the steepness for which the upper bound shoots upward towards the tail of the data. This is likely due to the lower number of observations towards the end of the range of the data and also the numerical instability in computing the mrl function. Looking at the estimated densities survival functions we can see that the majority of the ad libitum group have lower survival times compared to the restricted group. The mrl functions are monotonically decreasing and do not cross with regard to the point estimates. Moreover the interval estimates do not cross until the we reach about 800 days. This leads us to conclude that the remaining life expectancy of a rat in the restricted group is higher than the remaining life expectancy of a rat in the ad libitum group until we reach about 800 days.

We use the minimum posterior predictive loss approach (Gelfand & Ghosh, 1998) to compare the exponentiated Weibull model to the nonparametric gamma DPMM. Under this criterion the goal is to minimize, within the collection of models under consideration, the expectation of a specified loss function under the posterior predictive distribution of replicate responses $\boldsymbol{t}_{rep}$ given the observed data $\boldsymbol{t}_{obs}$. Here, we use the square error loss function so that the general criterion is given by $D_k(m) = \sum_{i=1}^{n} var(t_{i,rep}|\boldsymbol{t}_{obs}, m) + \frac{k}{k+1}\sum_{i=1}^{n}(E(t_{i,rep}|\boldsymbol{t}_{obs}, m) - t_{i,obs})^2$, where $t_{i,rep}$ is a replicate

of the i^{th} observation, $t_{i,obs}$, under the posterior predictive distribution of the m^{th} model. The first term is representative of a penalty measure $P(m)$, and the second term is a goodness-of-fit measure $G(m)$. The value of k is specified as the relative regret for departure from $t_{i,rep}$. Note that as k tends to infinity, the criterion becomes the sum of the penalty $P(m)$ and goodness-of-fit $G(m)$ measures.

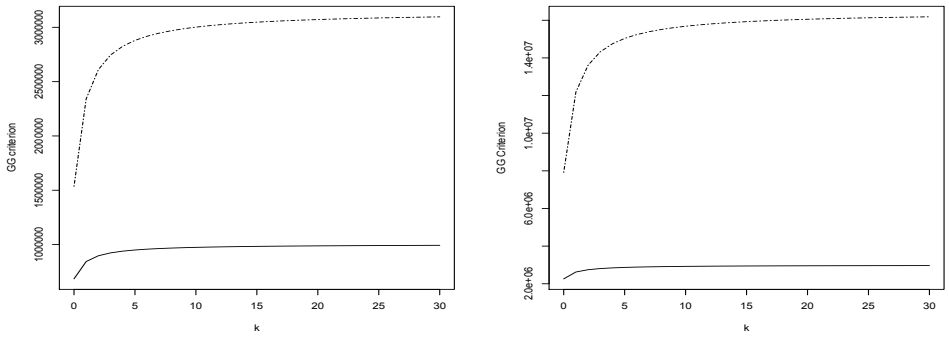

Figure 2.5: Values of the posterior predictive loss criterion for comparison between the parametric exponentiated Weibull model (dot dashed lines) and nonparametric gamma DPMM (solid lines).

For the exponentiated Weibull model (m_1), obtaining $E(t_{i,rep}|\boldsymbol{t}_{obs}, m)$ and $var(t_{i,rep}|\boldsymbol{t}_{obs}, m)$ is straightforward. The posterior predictive distribution is given by $p(t_{i,rep}|\boldsymbol{t}_{obs}) = \int EW(t_{i,rep}|\alpha, \theta, \sigma) \times p(\alpha, \theta, \sigma|data) d\alpha d\theta d\sigma$ and can thus be sampled by taking the posterior samples $(\alpha_b, \theta_b, \sigma_b)$, for $b = 1, ..., B$, and drawing $t_{i,rep,b}$ from the exponentiated Weibull distribution given each posterior parameter vector. Next, we compute the mean and variance of the B replicates. Important to note is that the mean and variance for one experimental group is going to be the same for each observation in that group. We find the $E(t_{i,rep}|\boldsymbol{t}_{obs}, m_1)$ and $var(t_{i,rep}|\boldsymbol{t}_{obs}, m_1)$ for the ad libitum group to be 671.2 and 17433.0, respectively, and for the the re-

stricted group to be 949.5 and 74691.7, respectively. Thus the ad libitum group has $G(m_1)^a = \sum_{i=1}^{90}(671.2 - t_{i,obs})^2 = 1615787$ and $P(m_1)^a = 90 * (17433.0) = 1568967$. The restricted group has $G(m_1)^r = \sum_{i=1}^{106}(949.5 - t_{i,obs})^2 = 8542725$ and $P(m_1)^r = 106 * (74691.7) = 7917319$.

Evaluating $D_k(m)$ under the nonparametric gamma DPMM (m_2) takes a little more care. Recall that $t_i|G \overset{ind}{\sim} \int \Gamma(t_i|exp(\theta), exp(\phi))dG(\theta, \phi)$ for $i = 1, ..., n$. In order to obtain replicates for each t_i, we need to know the l^{th} component from which the observed t_i came from according to the model, $t_{i,rep}|\boldsymbol{t}_{obs}, m_2 \sim \int \Gamma(t_{i,rep}|exp(\theta_{l_i}), exp(\phi_{l_i}))$ $p(exp(\theta_{l_i}), exp(\phi_{l_i})|data)d\theta_{l_i}d\phi_{l_i}$, for $i = 1, ..., n$, where the subscript l_i is the i^{th} value of the posterior sample of $\mathbf{w}$ and θ_{l_i} and ϕ_{l_i} are the l_i^{th} posterior samples of θ and ϕ. Essentially a single $t_{i,rep}$ is sampled from the gamma distribution at each posterior iteration $b = 1, ..., B$ integrating out all possible values of θ_{l_i} and ϕ_{l_i}. After obtaining B $t_{i,rep}$, we compute the mean ($E(t_{i,rep}|\boldsymbol{t}_{obs}, m_2)$) and variance ($var(t_{i,rep}|\boldsymbol{t}_{obs}, m_2)$) at each i^{th} replicate. For the ad libitum group we obtained $G(m_2)^a = 318919.2$ and $P(m_2)^a = 684342.1$, and for the restricted group $G(m_2)^r = 739435$ and $P(m_2)^r = 2247120$. Figure 2.5 is a plot of the criterion values over a grid of k values. For both groups the nonparametric gamma DPMM performs significantly better than the exponentiated Weibull model. The results of the formal model comparison support our earlier argument that the nonparametric gamma DPMM is indeed a better model for these data.

2.3.3 Analysis of survival times of patients with small cell lung cancer

As an example of obtaining inference in the presence of right censoring, we fit a gamma DPMM to the survival times, in days, of two treatment groups of patients with small cell lung cancer Ying et al. (1995). The patients were randomly assigned to one of two treatments referred to as Arm A and Arm B. Arm A patients received cisplatin (P) followed by etoposide (E), while Arm B patients received (E) followed by (P). There were a total of 62 patients in Arm A with 15 right censored survival times, while Arm B consisted of 59 patients with 8 right censored survival times. We fit a gamma DPMM independently to the two groups. We allocated 60% of the mean to $exp(a_{\mu_1} - a_{\mu_2})$ and 2.25% to $exp(a_{\mu_1} - 2a_{\mu_2})$ resulting in the following priors for Arm A: $a_\mu = (2.5, -3)$, $B_\mu = B_\sigma = ((0.21, 0)', (0, 0.21)')$. Analoguously, for Arm B: $a_\mu = (2.6, -2.9)$, $B_\mu = B_\sigma = ((0.21, 0)', (0, 0.21)')$. We use $a_\alpha = 3$, $b_\alpha = 1$, and $N = 25$. The effective posterior sample size is 2000.

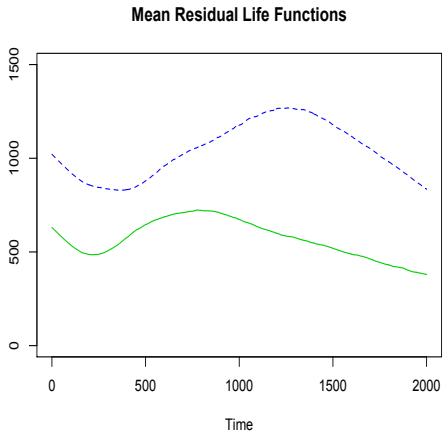

Figure 2.6: Point estimates for the mrl functions of Arm A (blue dashed) and Arm B (green solid).

The point estimates for mrl functions of the two treatment groups show Arm A to have a consistently higher mean residual life compared to Arm B in Figure 2.6. The result leads us to believe that Arm A treatment is more effective than the Arm B treatment. We take a closer look at the difference of the mean residual life survival times at a number of fixed time points.

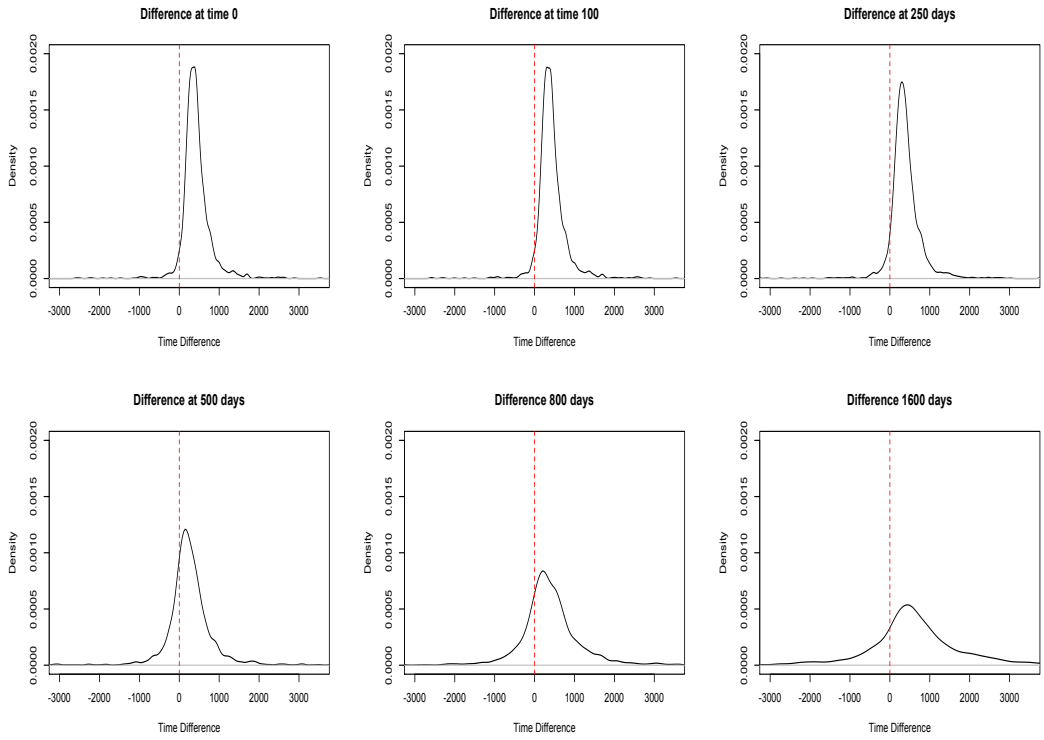

Figure 2.7: Densities of the mrl of Arm A minus Arm B at a number of fixed time points.

Specifically, we explore the posterior density of the mrl of Arm A minus Arm B at a particular time point, Figure 2.7. At time zero, which is the estimated difference of the overall mean of the two distributions, shows a strong difference between the two treatments in favor of Arm A. The same is true at 100 days, and just a slightly less

significance in the difference for day 250. At day 1600, there the difference becomes much less significant. In conclusion, Arm A has a significantly higher mrl then Arm B at lower time points. For larger time points, the difference is visible in the point estimate, but there is much uncertainty surrounding the estimate.

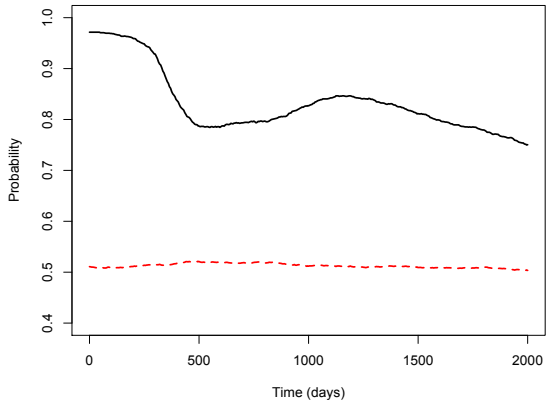

Figure 2.8: The posterior (black solid) and prior (red dashed) probability of the mrl function of Arm A being higher than the mrl function of Arm B over a grid of survival times (days).

A useful result to obtain for comparing the mrl functions of Arm A and Arm B is the probability of the mrl function of one group being higher than that of the other over a grid of survival times. In Figure 2.8, we look at the prior probability, $Pr(m_A(t) > m_B(t))$, and posterior probability, $Pr(m_A(t) > m_B(t)|data)$, under the gamma DPMM, where $m_A(t)$ and $m_B(t)$ are the mrl functions of Arm A and Arm B, respectively. The prior shows no favoritism of one group being higher than the other as the probability is relatively constant around 0.5 across the grid. The posterior probabilities suggest Arm A as having the higher mrl function across all survival time.

In particular, the probability of the mrl function of Arm A being higher than that of Arm B is larger during the earlier time period, reaching nearly probability 1. The probability decreases slightly around 500 days to about 0.8, followed by another peak of about 0.9 around 1200 days. The probability remains above 0.7 across the range of the data.

2.4 Discussion

We have reviewed basic properties and essential characteristics of the mrl function. We presented an easy-to-work with (yet limiting) class of distributions that correspond to a linear mrl function. We provided methods for obtaining the mrl function of several common distributions allowing us to study the various shapes of the mrl function. We find that the form of the mrl function for these distributions is again limited. Knowledge of the form of the mrl function would need to be available in order to select a proper model for mrl inference. The exponentiated Weibull model shows more promise in inference for the mrl function. The mrl function corresponding to the exponentiated Weibull distribution is able to take on several forms, namely constant, linear, increasing, decreasing, BT, and UBT. Another benefit of the exponentiated Weibull distribution is that it has a closed form for its survival function. This helps lower numerical error in estimating the mrl function, and provides ease in inference methods for censored data.

We discussed the benefits of fitting a DPMM to obtain flexible inference for the mrl function. However, when the focus is on inference for this particular functional,

the choice of the kernel plays an important role. Under the sufficient condition given in Section 2.2.1, we studied restrictions that need be placed on the mixture model in order to ensure the mrl function of the mixture distribution is well defined. In addition, we provided a result on the tail behavior of the mixture mrl function based on the corresponding property for the mrl function of the kernel distribution. The gamma kernel was shown to possess the most desirable properties out of the distributions that we investigated. We showed that under a gamma mixture, the resulting mrl function is dense in the pointwise sense on the space of continuous mrl functions. The practical utility of the proposed nonparametric mixture model was demonstrated through analysis of simulated data examples and real data sets from the literature.

A practically important extension involves methods for inclusion of covariates. Mean residual life regression modeling can be explored under either a structured semiparametric setting, such as the proportional mrl setting, or a fully nonparametric framework, for instance, based on mixture modeling for the joint response-covariate distribution. Chapter 3 discloses the model structure for the later.

Chapter 3

Bayesian nonparametric regression modeling for survival response distributions

In this chapter, we maintain the perspective of achieving sound inference for the MRL function under a regression setting. Often, there are a number of covariates associated with survival data. These covariates may be continuous (e.g., blood pressure, weight, age, etc.) or discrete/categorical (e.g., gender, race, profession, etc.). The researcher would like to model the survival times by incorporating the covariates, so that they may predict measurements such as survival time of a new patient given a set of covariates. In Section 3.2, we address this idea using the curve-fitting approach. Another type of covariate frequently seen in survival analysis is a fixed covariate indicating the subject was in a particular experimental group, for example, treatment and control

settings. For data under this scenario, we might expect the experimental groups to have underlining characteristics that tie them together, while still capturing the effect of the treatment. In Section 3.3, we have developed a model that measures the dependency between two experimental groups as well as incorporating individual group effect on the survival times. In Section 3.4, we revisit the small cell lung cancer data analyzed in Section 2.3.3. We compare the graphical results obtained under independent gamma DPMMs for the two experimental groups with those obtained when dependency across groups is incorporated within the model. The dependent model is compared formally, via conditional predictive ordinate values, with two competing models. Combining the idea of the curve fitting approach with modeling dependency across groups, we include a continuous covariate in the small cell lung cancer dataset and present our inferential results.

3.1 Literature review of Bayesian survival regression

The literature on Bayesian regression analysis can be categorized into three sets, fully parametric, semiparametric, and nonparametric. In regards to fully parametric Bayesian regression models, the literature is quite extensive. A basic approach being to consider a parametric model, such as the exponential, Weibull, gamma, or lognormal, and use a link function on one of the parameters to equate a linear structure on the set of covariates. For example, in the exponential model the rate parameter, λ, can be used to incorporate the set of covariates $\mathbf{x}$ with effects $\boldsymbol{\beta}$ by setting $\lambda = \exp(\mathbf{x}'\boldsymbol{\beta})$.

A multivariate normal prior or reference prior is typically placed on the $\boldsymbol{\beta}$ parameters. Several commonly used parametric models also fall into the category of the proportional hazards (PH) models, Cox (1972), such that the model can be separated into a baseline distribution that is independent of covariates, $h_0(t)$, and a function of the covariates independent of time, $c(\mathbf{x}'\boldsymbol{\beta})$. Hence, given a second set of covariates, the hazard function is proportional to that of the original set. Literature on fully parametric Bayesian regression models for survival data include, but is not limited to, Achcarar et al. (1985), Dellaportas & Smith (1993), Scurrah et al. (2000), Kuo et al. (1992), and Ibrahim et al. (2001).

As Hendersoni (1995) points out, the PH model is not famous for it's model formation, but rather for the potential for a nonparametric element through the baseline function. The PH model can be written in terms of the survival function by $S_0(t)^{c(\mathbf{x}'\boldsymbol{\beta})}$, where $S_0(t)$ is the baseline survival function (Klein & Moeschberger, 1997). Thus a nonparametric prior may be placed on the baseline hazard, cumulative hazard, or survival function directly. Once prior processes such as the Dirichlet process prior for the survival distribution (Ferguson, 1973) and the gamma process for the cumulative hazard distribution (Kalbfleisch, 1978) entered the scene, a shift towards semiparametric modeling occurred. While the DP is a stochastic process that produces realizations of a distribution function (Section 2.2.1), a gamma process is a Lévy is stochastic process with independent positive stationary increments. Specifically, the Lévy process can be used as a prior for sequential differences of the cumulative hazard function. For a neutral to the right Beta process, the differences follow a gamma distribution with mean

and variance that are functions of the times that define the increments. Dystra & Laud (1981) provide an absolutely continuous structure for the hazard function by extending gamma process prior with mixing over the reciprocal of a known continuous function and giving the mixing distribution a gamma process prior. The nonparametric element of PH structures can be extended by letting the link function and the baseline distribution be random. An interesting example is Gelfand & Mallick (1995), who utilize the the monotonicity property of the cumulative hazard and the link function to model the functions on a transformed scale of mixtures of beta distribution functions. They argue for the use of beta mixtures, since they are dense in the space of positive supported and valued monotonic functions. A DP prior is placed on the weights of the mixtures, while using Jeffreys' prior on the covariate effects given the weights.

Another approach in Bayesian semiparametric survival regression is to consider the traditional linear regression structure on the responses, t_i, and the covariates, $t_i - \mathbf{x}'\boldsymbol{\beta} = \epsilon_i$, then place a nonparametric prior on the errors, ϵ_i. Accelerated failure time (AFT) structures, $log(t_i) - \mathbf{x}'\boldsymbol{\beta} = \epsilon_i$, have also been studied in the Bayesian semiparametric framework. In regards to median regression, the literature includes Walker & Mallick (1999) and Hanson & Johnson (2002) place Pólya tree mixture on the distribution of the errors, while Kottas & Gelfand (2001) and use a DP prior. As an extension to their median regression modeling, Gelfand & Kottas (2003) present a Bayesian semiparametric model for median residual life induced by a semiparametric AFT regression model. A general semiparametric quantile regression model is developed in Kottas & Krnjajić (2009). A fully nonparametric quantile inference is constructed using DP priors

in Hjort & Petrone (2007) with a discussion on the extension to the regression setting.

Although the Bayesian semiparametric literature has made great strides in flexible Bayesian regression modeling, there is still a parametric assumption being made that places restriction on the inferential potential of the model. In particular, an obvious restriction is seen in PH and AFT models is the inability of the survival curves to cross for any two sets of covariates. Also, up until now, priors for the covariate coefficients have remained parametric. With this motivation, DeIorio et al. (2004) presents the ANOVA dependent Dirichlet process (DDP) model, which is later extended to include continuous covariates to the linear DDP DeIorio et al. (2009). The linear DDP uses a DP mixture model for the log survival responses with a DDP prior on the mixing distribution. The covariates enter the model in the centrality parameter of the kernel distribution, and indexes the mixing distribution. Essentially, the model is a DP mixture of log linear models. To date, the linear DDP is a leading model in Bayesian nonparametric survival regression. A more detailed description and discussion on the linear DDP can be found in Section 3.4.

A second fully nonparametric model that has been more recently developed is the extension of DP mixture models to include random covariates in the kernel distribution. This modeling technique is referred to as curve fitting regression, and uses a joint response-covariate kernel distribution (e.g. Taddy & Kottas (2010) and Kottas et al. (2013)). The curve fitting regression has the ability to capture non-standard relationships across the covariate space, providing flexible inference across the conditional survival and density functions. While the linear DDP possesses the novelty of inter-

pretable parameters and a straight forward posterior sampling algorithm, it is limited to a linear, or log linear, relationship across the covariates. The curve fitting regression approach is a more robust model, but does run into trouble when the number of random covariates becomes large. Typically survival data consists of a small to moderate number of covariates, so the curve fitting regression is attractive model choice. In the following section, we explore the curve fitting regression approach from the perspective of inference for mrl regression.

3.2 Curve fitting with random covariates

When a covariate can be considered to be random, by which we mean the covariate is not fixed such as a patient being assigned to treatment or control groups, it makes sense to model the covariate jointly with the survival response variable. The benefit of this modeling approach is the simplicity of obtaining any conditional or marginal distributions and functionals that are desired by the researcher. In addition, we are not restricted to any particular shape in regards to functionals of the survival responses, given a set of covariates, within and across the covariate space.

3.2.1 Model formulation

Let $\boldsymbol{x}$ be a vector of random covariates and $t > 0$ the survival time of a subject. We model the joint response-covariate density using a DPMM, $f(t, \boldsymbol{x}|G) = \int_{\boldsymbol{\Theta}} k(t, \boldsymbol{x}|\boldsymbol{\theta}) dG(\boldsymbol{\theta})$. We will work with the truncated version of the SB constructive

definition of the DP,

$$f(t, \boldsymbol{x}|G) = \int_{\boldsymbol{\Theta}} k(t, \boldsymbol{x}|\boldsymbol{\theta}) dG(\boldsymbol{\theta}) \approx \sum_{l=1}^{L} p_l k(t, \boldsymbol{x}|\boldsymbol{\theta}_l) \tag{3.1}$$

In Section 2.2.1, we demonstrate the importance of the kernel choice in ensuring the finiteness of the mean. In the regression setting, we are interested in the conditional mrl at any fixed set of covariates, $m(t|\boldsymbol{x}_0|G_L)$. The sufficiency condition that ensures the finiteness for the mean, see Section 2.2.1, can be extended for $E(t|\boldsymbol{x}_0|G_L)$. Let $T(\boldsymbol{x}_0, \boldsymbol{\theta}) = \int_0^\infty t k(t|\boldsymbol{x}_0|\boldsymbol{\theta}) dt$ where $k(t, \boldsymbol{x}_0|\boldsymbol{\theta}) = k(t|\boldsymbol{x}_0|\boldsymbol{\theta}) k(\boldsymbol{x}_0|\boldsymbol{\theta})$. The condition states that if $A(\boldsymbol{x}_0, \boldsymbol{\psi}) = \int_{\boldsymbol{\Theta}} T(\boldsymbol{x}_0, \boldsymbol{\theta}) dG_0(\boldsymbol{\theta}) < \infty$ for all $\boldsymbol{\theta}$, then $E(t|\boldsymbol{x}_0|G_L) < \infty$, almost surely. The form of the mean regression is given by:

$$\begin{aligned} E(t|\boldsymbol{x}_0|G_L) &= \frac{\int_0^\infty t f(t, \boldsymbol{x}_0|G_L) dt}{f(\boldsymbol{x}_0|G_L)} = \frac{\int_0^\infty t \int_{\boldsymbol{\Theta}} k(t, \boldsymbol{x}_0|\boldsymbol{\theta}) dG_L(\boldsymbol{\theta}) dt}{f(\boldsymbol{x}_0|G_L)} \\ &= \frac{\int_{\boldsymbol{\Theta}} \int_0^\infty t k(t, \boldsymbol{x}_0|\boldsymbol{\theta}) dt dG_L(\boldsymbol{\theta})}{f(\boldsymbol{x}_0|G_L)} = \frac{\sum_{l=1}^{L} p_l \int_0^\infty t k(t, \boldsymbol{x}_0|\boldsymbol{\theta}_l) dt}{\sum_{l=1}^{L} p_l k(\boldsymbol{x}_0|\boldsymbol{\theta}_l)} \\ &= \sum_{l=1}^{L} q_l(\boldsymbol{x}_0|\boldsymbol{\theta}_l) E(t|\boldsymbol{x}_0|\boldsymbol{\theta}_l) \end{aligned} \tag{3.2}$$

where $q_l(\boldsymbol{x}_0|\boldsymbol{\theta}_l) = p_l k(\boldsymbol{x}_0|\boldsymbol{\theta}_l) / \{\sum_{l=1}^{L} p_l k(\boldsymbol{x}_0|\boldsymbol{\theta}_l)\}$ are covariate dependent weights (Müller et al., 1996). If we chose independent kernel distributions for T and $\boldsymbol{X}_0$, then $k(t, \boldsymbol{x}_0|\boldsymbol{\theta}) = k(t|\boldsymbol{x}_0|\boldsymbol{\theta}) k(\boldsymbol{x}_0|\boldsymbol{\theta}) = k(t|\boldsymbol{\theta}) k(\boldsymbol{x}_0|\boldsymbol{\theta})$. We need only choose $k(\boldsymbol{x})$ such that the support is consistent with the support of the covariates. In the case where the covariates are continuous and take values on the real line (possibly after transformation), the multivariate normal distribution for $k(\boldsymbol{x})$ is a natural choice. Turning to $k(t)$, under the independent scenario, $A(\boldsymbol{x}_0, \boldsymbol{\psi})$ becomes the same with $A(\boldsymbol{\psi})$ in Section 2.2.1. In that discussion, the gamma distribution was the clear winner for the kernel choice, completing justifica-

tion for the following kernel distribution: $k(t, \boldsymbol{x}_0|\boldsymbol{\theta}) = \Gamma(t|\theta, \phi)N_d(\boldsymbol{x}|\beta, \kappa^2)$, where d is the number of covariates. In our simulation example, we consider a single continuous real-valued covariate ($d = 1$). Let t_i be the survival time and x_i be the corresponding real valued covariate for subject i, for $i = 1, ..., n$. We consider the following DPMM,

$$\begin{aligned} t_i, x_i|\boldsymbol{\theta}, \mathrm{w}_i &\overset{ind}{\sim} \Gamma(t_i|e^{\theta_{\mathrm{w}_i}}, e^{\phi_{\mathrm{w}_i}})N(x_i|\beta_{\mathrm{w}_i}, \kappa^2_{\mathrm{w}_i}) \\ \mathrm{w}_i|\mathbf{p} &\overset{iid}{\sim} \sum_{l=1}^{L} p_l\delta_l(\mathrm{w}_i) \\ (\theta_l, \phi_l, \beta_l, \kappa_l^2)'|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \lambda, \tau^2, \rho &\overset{iid}{\sim} N_2((\theta_l, \phi_l)'|\boldsymbol{\mu}, \boldsymbol{\Sigma})N(\beta_l|\lambda, \tau^2)\Gamma^{-1}(\kappa_l^2|a, \rho) \end{aligned} \tag{3.3}$$

where the DP implied prior for $\boldsymbol{p}$ is the same with model (2.14). We place the following priors: $\alpha \sim \Gamma(\alpha|a_\alpha, b_\alpha(\text{rate}))$, $\boldsymbol{\mu} \sim N_2(\boldsymbol{\mu}|a_\mu, B_\mu)$, $\boldsymbol{\Sigma} \sim IWish(\boldsymbol{\Sigma}|a_\Sigma, B_\Sigma)$, $\lambda \sim N(\lambda|a_\lambda, b_\lambda)$, $\tau^2 \sim \Gamma^{-1}(\tau^2|a_\tau, b_\tau)$, and $\rho \sim \Gamma(\rho|a_\rho, b_\rho)$. The MCMC is relatively straight forward, only requiring one Metropolis-Hastings step for the parameters of the gamma kernel, just as in the model with no covariates. The rest of the parameters can be sampled via Gibbs steps.

Note that under the under the curve fitting approach, as the number of covariates increases, the more computationally expensive fitting the model becomes. This is due mainly to the dimension of the covariance matrix for the random covariates in the kernel $\boldsymbol{\Sigma}$. If d is the number of random covariates, the number of parameters in $\boldsymbol{\Sigma}$ that will have to be updated is $Ld(d+1)/2$. One can see how quickly this number can grow with growing d. Thus, curve fitting under the DP mixture framework is best suited for situations in which the number of random covariates is small to moderate.

3.2.2 Prior selection and posterior inference

For prior specification, we use the same ideas discussed in Section 2.2.2. In particular, by means of imagining one component covering the prior range believed by the expert, here, for both the survival responses and the covariates. Under the product kernel, we specify the prior parameters associated with the survival times independently of the prior parameters for the covariate values.

We can obtain posterior point and interval estimates for the desired survival functionals using the same methodology as in Section 2.2.3. We compute the value of the functional at each posterior sample of the parameters over a grid of survival times at a particular value of the covariate, x_0, and save the desired quantiles. The expressions of the density, survival, hazard, and mrl functions are given respectively below:

$$\begin{aligned}
f(t|\boldsymbol{x}_0, G_L) &= \frac{f(t, \boldsymbol{x}_0|G_L)}{f(\boldsymbol{x}_0|G_L)} = \frac{\sum_{l=1}^{L} p_l k(t, \boldsymbol{x}_0|\boldsymbol{\theta}_l)}{\sum_{l=1}^{L} p_l k(\boldsymbol{x}_0|\boldsymbol{\theta}_l)} \\
S(t|\boldsymbol{x}_0, G_L) &= 1 - \int_0^t f(u|\boldsymbol{x}_0, G_L)du = 1 - \frac{\sum_{l=1}^{L} p_l k(\boldsymbol{x}_0|\boldsymbol{\theta}_l) K(t|\boldsymbol{x}_0, \boldsymbol{\theta}_l)}{\sum_{l=1}^{L} p_l k(\boldsymbol{x}_0|\boldsymbol{\theta}_l)} \\
h(t|\boldsymbol{x}_0, G_L) &= \frac{f(t|\boldsymbol{x}_0, G_L)}{S(t|\boldsymbol{x}_0, G_L)} \\
m(t|\boldsymbol{x}_0, G_L) &= \frac{\int_t^{\infty} S(u|\boldsymbol{x}_0, G_L)du}{S(t|\boldsymbol{x}_0, G_L)} = \frac{E(t|\boldsymbol{x}_0, G_L) - \int_0^t S(u|\boldsymbol{x}_0, G_L)du}{S(t|\boldsymbol{x}_0, G_L)}
\end{aligned}$$

where $K(t|\boldsymbol{x}_0|\boldsymbol{\theta}_l)$ is the conditional kernel distribution function (a gamma cdf under the product kernel of model (3.3)).

Of important note is the regression structure of the functionals under this model. Looking back to the structure for the conditional mean of survival (3.2), the $q_l(\cdot)$ functions can be thought of as a new set of weights, such that the functional is

a weighted mixture of the mean of the conditional kernel components. The structure provides a convenient interpretation of the functional. Moreover, the new weights, $q_l(\cdot)$ are functions of the covariate. This property illustrates the potential for seeing a non-standard relationship of the mean regression across the covariate space. Similarly, the mrl function can be written as,

$$m(t|\boldsymbol{x}_0, G_L) \quad = \quad \sum_{l=1}^{L} q_l(t, \boldsymbol{x}_0|\boldsymbol{\theta}_l) m(t|\boldsymbol{x}_0, \boldsymbol{\theta}_l) \tag{3.4}$$

where $q_l(\boldsymbol{x}_0|\boldsymbol{\theta}_l) = p_l k(\boldsymbol{x}_0|\boldsymbol{\theta}_l) S(t|\boldsymbol{x}_0, \boldsymbol{\theta}_l) / \{\sum_{l=1}^{L} p_l k(\boldsymbol{x}_0|\boldsymbol{\theta}_l) S(t|\boldsymbol{x}_0, \boldsymbol{\theta}_l)\}$. Therefore, we can think of the mrl regression function as a finite weighted sum of the mrl functions associated with the conditional kernel components, with weights that are dependent on time as well at the covariate values. Aside from the nice interpretation on the form of the mrl regression function under this model, (3.4) shows the potential of the model to capture non-standard relationships across the covariate space as well as unique features within the mrl regression function.

3.2.3 Simulation example

We simulate 1500 data values from a population having the following density: $f(t, x) = \sum_{l=1}^{M} q_l \Gamma(t|a_l, b_l) N(x|m_l, s_l^2)$, where $M = 6$, $a = (45, 3, 125, 0.4, 0.5, 4)'$, $b = (3, 0.2, 3.8, 0.2, 0.3, 5)'$, $m = (-12, -8, 0, 12, 18, 21)'$, $s = (6, 5, 4, 5, 3, 2)'$, and $q = (0.28, 0.1, 0.25, 0.21, 0.11, 0.05)'$. The simulated data is shown in the left panel of Figure 3.1.

This population was constructed to have various shapes of the mrl function

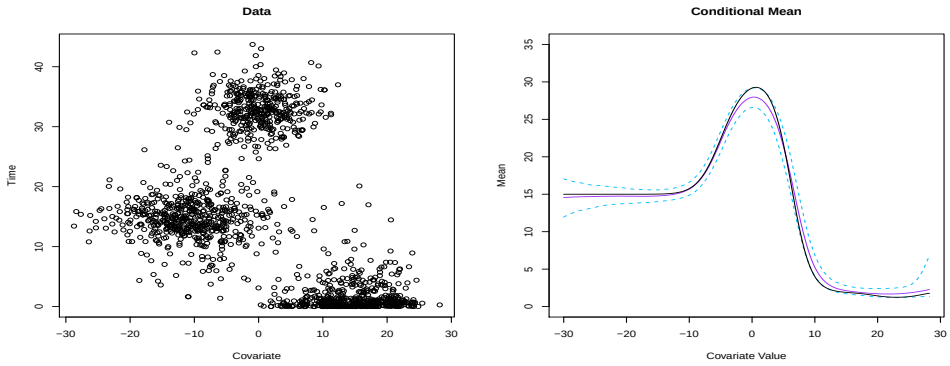

Figure 3.1: Simulated data (left), and point (purple solid) and interval estimates (light blue dashed) of the mean overlaying the truth (black solid) (right).

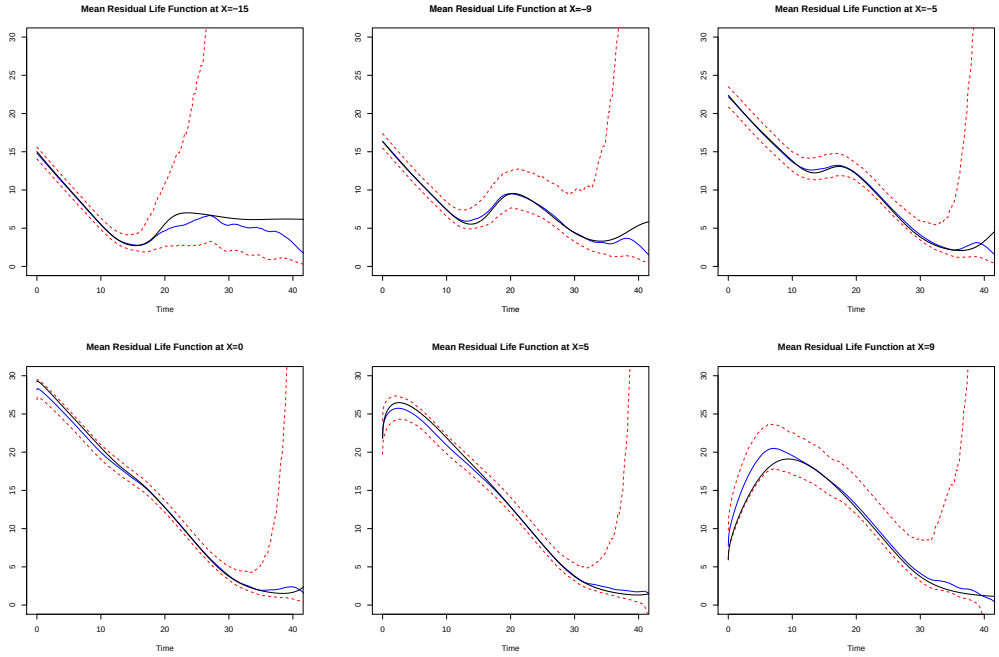

Figure 3.2: Point (blue solid) and 95% interval estimates (red dashed) of the mrl function for the specified covariate value overlaying the true mrl function of the population (black solid).

across different covariate values. The shapes include mrl functions with multiple change points, nonlinearly decreasing, UBT, and nonlinearly increasing. We demonstrate the

ability of the joint DPMM in (3.3) to capture these various mrl functional forms at appropriate covariate values. The following priors were used: $a_\mu = (0.59, -2.12)$, $B_\mu = B_\sigma = ((0.019, 0.019)', (0.019, 0.019)')$, $a_\lambda = 0$, $a_\tau = 2$, $a_\rho = 1$, $b_\lambda = b_\tau = 88$, $b_\rho = 1/88$, $a_\alpha = 3$, $b_\alpha = 0.1$. The DP truncation level was set at $L = 80$.

The mean of the survival times across a grid of covariate values is shown in Figure 3.1 (right panel). In general, the model is able to capture the non-linear trend of the mean over the covariate values. The point estimate is almost on top of the truth for a large portion of the grid.

The results for mrl functional inference is shown in Figure 3.2. We provide point and interval estimates for the mrl function at six different covariate values. The model is able to capture the overall shape of the true mrl functions, despite the variety of shapes. At covariate values where the data is most dense, such as $x = -5$, $x = 0$, and $x = 5$, the inference is more precise as is seen in the narrow interval bands. As we move to covariate values away from zero, where data is more sparse, the wide interval bands reflect the uncertainty of the mrl functional shape.

3.3 Dependent Dirichlet Process Mixture Model Across Experimental Groups

Often in clinical trials, researchers are interested in modeling survival times of patients under treatment and control groups. In Section 2.4, we model the survival times of rats data under two experimental conditions, ad libitum and restricted. The

inference for the groups was obtained independently. An extension would be to model the groups jointly. The benefit of modeling the groups jointly is to be able to capture dependencies amongst groups, and to borrow strength from the group with a larger sample size for more precise inference.

3.3.1 Dirichlet process prior with dependent weights

Let $s \in S$ represent in general the index of dependence. In our case, this indicates the experimental group, that is $S = \{T, C\}$ where (T) is the treatment group and (C) is the control group. The DPMM under the regression setting in (3.3) can be extended to $f(t, \boldsymbol{x}|G_s) = \int_{\boldsymbol{\Theta}} k(t, \boldsymbol{x}|\boldsymbol{\theta}) dG_s(\boldsymbol{\theta})$ for $s \in S$, where now we are modeling a pair of dependent random mixing distributions $\{G_s : s \in S\}$. We desire to model the distributions in such a way as to incorporate dependencies across experimental groups, while maintaining marginally the DP prior, $G_s \sim DP$, for each $s \in S$. MacEachern (2000) develops the dependent DP prior in generality with both the weights and locations in the DP SB definition dependent on experimental group: $G_s(\cdot) = \sum_{l=1}^{\infty} \omega_{ls} \delta_{\boldsymbol{\theta}_{ls}}(\cdot)$. Marginally, $G_s \sim DP(\alpha_s, G_{0s})$ for each $s \in S$. MacEachern (2000) goes on to describe the computational difficulties in modeling dependencies in the weights across groups, thus motivating development of the "single p" model. In this model, the weights do not change over the groups, only the locations vary, $G_s(\cdot) = \sum_{l=1}^{\infty} \omega_l \delta_{\boldsymbol{\theta}_{ls}}(\cdot)$. We have studied such dependent DP mixture modeling for comparison of neuronal intensities under distinct experimental conditions in Kottas et al. (2012). Applications of "single p" dependent DP models include DeIorio et al. (2004), Gelfand et al. (2005), Rodriguez

& ter Horst (2008), Kottas & Krnjajić (2009), DeIorio et al. (2009), and Fronczyk & Kottas (2010).

While computationally convenient and a useful extension from the basic DP prior, assuming the same weights has potential disadvantages in our setting. A practical disadvantage of the "single p" dependent DP construction involves applications with a moderate to large number of covariates. For such cases, the "single p" prior requires building dependence across $s \in S$ for a large number of kernel parameters, whereas modeling dependency through the weights is not affected by the dimensionality of the mixture kernel. In situations where we might expect similar locations across groups, modeling dependency through the weights is more attractive.

Recall that we are interested in the case where we have two groups, treatment and control. In the mixture modeling, we might expect the two groups to be comprised of similar components, but these components may have varying prevalence. Thus we want the mixing distribution to have the form $G_s(\cdot) = \sum_{l=1}^{\infty} \omega_{ls} \delta_{\boldsymbol{\theta}_l}(\cdot)$. We will again use the truncated version of $G_s \approx \sum_{l=1}^{L} p_{ls} \delta_{\boldsymbol{\theta}_l}$ for $s \in \{T, C\}$ representing the treatment and control groups, respectively. Therefore, we propose the following model:

$$f(t, \boldsymbol{x}|G_s) \quad = \quad \int_{\boldsymbol{\Theta}} k(t, \boldsymbol{x}|\boldsymbol{\theta}) dG_s(\boldsymbol{\theta}) \approx \sum_{l=1}^{L} p_{ls} k(t, \boldsymbol{x}|\boldsymbol{\theta}_l) \quad \text{for } s \in \{T, C\} \qquad (3.5)$$

Under the stick-breaking method in obtaining the weights, we sample independently the latent parameters, $\upsilon_r \sim Beta(1, \alpha)$, which is equivalent to using $\zeta_r = (1 - \upsilon_r) \sim Beta(\alpha, 1)$ for $r = 1, ..., L-1$. If we use a bivariate beta distribution for (ζ_{Tr}, ζ_{Cr}), we can incorporate the dependency between the two groups. Minimally, we

need the marginals to be $\zeta^*_{sr} \sim Beta(\alpha, 1)$ for $s \in \{T, C\}$. By applying a bivariate beta distribution to the dependent DPMM with common locations and dependent weights, the following model for survival regression data that specify two experimental groups is presented,

$$
\begin{aligned}
t_{iC}, \boldsymbol{x}_{iC} | G_C &\overset{ind}{\sim} f(t_{iC}, \boldsymbol{x}_{iC} | G_C) = \int_{\boldsymbol{\Theta}} k(t_{iC}, \boldsymbol{x}_{iC} | \boldsymbol{\theta}) dG_C(\boldsymbol{\theta}), \quad i = 1, .., n_C \\
t_{iT}, \boldsymbol{x}_{iT} | G_T &\overset{ind}{\sim} f(t_{iT}, \boldsymbol{x}_{iT} | G_T) = \int_{\boldsymbol{\Theta}} k(t_{iT}, \boldsymbol{x}_{iT} | \boldsymbol{\theta}) dG_T(\boldsymbol{\theta}), \quad i = 1, .., n_T \\
(G_C, G_T) | \boldsymbol{\phi}, \boldsymbol{\psi} &\sim DDP(\boldsymbol{\phi}, G_0(\cdot | \boldsymbol{\psi})) \\
&\quad \text{where } \; G_S = \sum_{l=1}^{\infty} \omega_{lS} \delta_{\boldsymbol{\theta}_S} \approx \sum_{l=1}^{L} p_{ls} \delta_{\boldsymbol{\theta}_l} \\
\boldsymbol{\theta}_l, \boldsymbol{\psi} &\overset{iid}{\sim} G_0(\cdot | \boldsymbol{\psi}), \;\; l = 1, 2, ... \\
&\quad \omega_{1s} = 1 - \zeta_{1s} | \; \omega_{ls} = (1 - \zeta_{ls}) \prod_{r=1}^{l-1} \zeta_{rs} \;\; l = 2, 3, ... \\
&\quad \text{with } \; (\zeta_{lC}, \zeta_{lT}) | \boldsymbol{\phi} \overset{ind}{\sim} \text{biv-beta}(\cdot | \boldsymbol{\phi}), l = 1, 2, ... \\
&\quad \text{such that marginally, } \; \zeta_{lC} \sim beta(\alpha_C, 1) \; \text{ and } \; \zeta_{lT} \sim beta(\alpha_T, 1)
\end{aligned}
$$

There are a number of bivariate beta distributions to consider, however, some exhibit more favorable properties for our purposes than others. In particular, more flexible marginals would allow for different α values, i.e., $\zeta_{Tr} \sim Beta(\alpha_T, 1)$ and $\zeta_{Cr} \sim Beta(\alpha_C, 1)$. Naturally, we also want the bivariate beta to correspond to reasonable computation of the MCMC, have a relatively simple density form, full support for the correlation between ζ_{Cr} and ζ_{Tr}, and ideally an analytic expression for the correlation. The correlation of the bivariate beta distribution for (ζ_{Tr}, ζ_{Cr}) will be important for the study of the implied dependence structure in the dependent DP prior for (G_T, G_C).

The bivariate beta distribution presented by Michael & Schucany (2011) has

a simple analytic form for the correlation structure and while the density does not have an easily obtainable form, sampling from the density is straight forward. A bigger problem is that we can not obtain marginal distributions for ζ_{Cr} and ζ_{Tr} with different α parameters. Another possible bivariate beta is provided by Olkin & Liu (2003). This bivariate beta has a reasonable density form and the appropriate beta marginals for ζ_{Cr} and ζ_{Tr}, allowing different α values. However, the correlation does not have an analytic form, and the support of the correlation is restricted to positive values.

The bivariate beta that we chose to implement in our model is that of Nadarajah & Kotz (2005). They construct a bivariate beta distribution through products of beta distributions. Start with beta random variables, $U \sim Beta(a_1, b_1)$, $V \sim Beta(a_2, b_2)$, and $W \sim Beta(b, c)$, subject to the constraint, $c = a_1 + b_1 = a_2 + b_2$. The bivariate beta distribution we are interested in is defined for random variables, X and Y, where $X = UW$ and $Y = VW$. The marginals are given by $X \sim Beta(a_1, b_1 + b)$ and $Y \sim Beta(a_2, b_2 + b)$. We can obtain the desired beta marginals for ζ_{Cr} and ζ_{Tr} by setting $b_1 + b = b_2 + b = 1$, although the marginals will have the same α parameter. The density has a complicated form, but it can be sampled from using latent variables. The correlation has an analytic expression, however, has positive support, which may or may not be a reasonable assumption. Induced correlations in the model under this bivariate beta distribution is discussed in Section 3.3.2.

3.3.2 Properties of the DDP mixture model

In what follows, we explore what correlation structures are induced by the Nadarajah & Kotz (2005) bivariate beta distribution. At the end of the section, we provide the full hierarchical DDP mixture model. Under this bivariate beta construction, the groups have a common $\alpha = \alpha_C = \alpha_T$, and the correlation is driven by both parameters, α and b. The construction is based off of the product of independent beta distributions. Start with sampling the independent latent variables: $U \sim \beta(\alpha, 1-b)$, $V \sim \beta(\alpha, 1-b)$, $W \sim \beta(\alpha+1-b, b)$. Let $\zeta_C = UW$ and $\zeta_T = VW$. The weights are defined by $w_{s1} = 1 - \zeta_{1s}$, $w_{ls} = (1-\zeta_{ls})\prod_{r=1}^{l-1}\zeta_{rs}$, for $l \in \{2, 3, ...\}$.

We are interested in obtaining the correlation between the two mixing distributions, G_C and G_T, implied under this bivariate beta distribution. We first start with the correlation between ζ_C and ζ_T, $Cor(\zeta_C, \zeta_T)$. We omit the component subscript in the latent variables, since results are the same for each $l \in \{1, 2, ...\}$. The covariance can be written as, $Cov(\zeta_C, \zeta_T) = E(\zeta_C\zeta_T) - E(\zeta_C)E(\zeta_T) = E((UW)(VW)) - E(UW)E(VW)$. Using the fact that U, V, W are independent the covariance becomes, $E(U)E(V)E(W^2) - E(U)E(V)E^2(W) = E(U)E(V)Var(W)$, which gives the following resulting covariance,

$$Cov(\zeta_C, \zeta_T) = \left(\frac{\alpha}{\alpha+1-b}\right)^2 \left(\frac{(\alpha+1-b)b}{(\alpha+1)(\alpha+2)}\right) = \frac{\alpha^2 b}{(\alpha+1-b)(\alpha+1)^2(\alpha+2)} \quad (3.6)$$

By definition, the correlation is $Cor(\zeta_C, \zeta_T) = Cov(\zeta_C, \zeta_T)/\sqrt{Var(\zeta_C)Var(\zeta_T)}$. Note that ζ_C and ζ_T have the same marginal distribution, $Beta(\alpha, 1)$, so they have the same expression for the variance. The correlation is therefor given by,

$$Cor(\zeta_C, \zeta_T) = \left(\frac{\alpha^2 b}{(\alpha+1-b)(\alpha+1)^2(\alpha+2)}\right)\left(\frac{(\alpha+1)^2(\alpha+2)}{\alpha}\right) = \frac{\alpha b}{\alpha+1-b} \quad (3.7)$$

The correlation between ζ_C and ζ_T can take values on the interval $(0, 1)$. As $b \to 0$ and/or $\alpha \to 0$, the correlation goes to 0. As $b \to 1$ and/or $\alpha \to \infty$, the correlation tends to 1. Figure 3.3 below shows a surface plot of the correlation over a grid of α and b values.

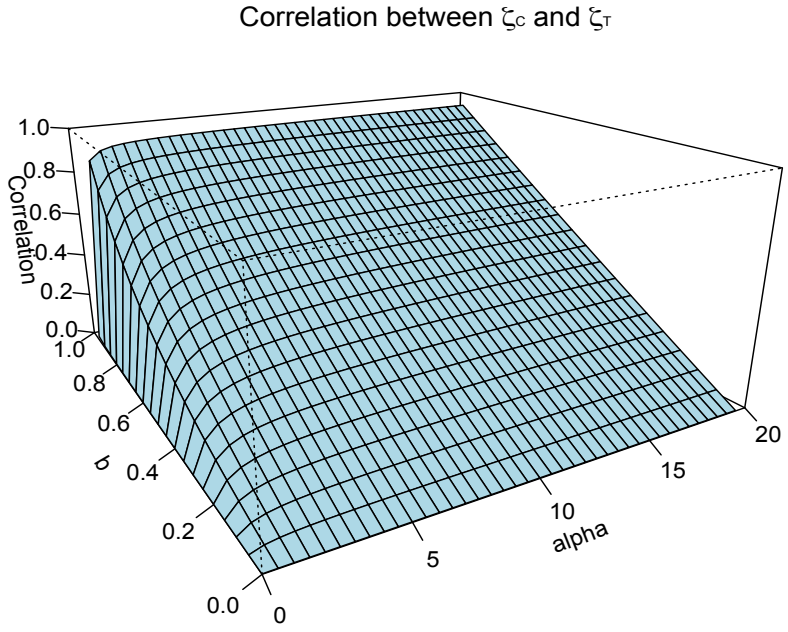

Figure 3.3: Correlation between ζ_C and ζ_T over a grid of α and b values.

The next step is to explore the correlation of the weights, $Cor(w_{lC}, w_{lT})$ for $l \in \{1, 2, ...\}$. When $l = 1$, $w_{1s} = 1 - \zeta_{1s}$, which is simply a linear operation, hence the covariance and correlation are the same as before. The $Cov(w_{1C}, w_{1T}) = Cov(\zeta_C, \zeta_T)$ and $Cor(w_{1C}, w_{1T}) = Cor(\zeta_C, \zeta_T)$ are given by (3.6) and (3.7), respectively. The case is different for $l = \{2, 3, ...\}$. In this case, the covariance is defined as

$$Cov(w_{lC}, w_{lT}) = E\left[\left((1-\zeta_{lC})\prod_{r=1}^{l-1}\zeta_{rC}\right)\left((1-\zeta_{lT})\prod_{r=1}^{l-1}\zeta_{rT}\right)\right] - E\left[(1-\zeta_{lC})\prod_{r=1}^{l-1}\zeta_{rC}\right]E\left[(1-\zeta_{lT})\prod_{r=1}^{l-1}\zeta_{rT}\right].$$

Using the fact that ζ_{ls} are independent across $l = 1, ..., L$, for each $s \in \{C, T\}$ (see Appendix for details), the covariance, for $l \in \{2, 3, ...\}$, can be expressed as,

$$\begin{aligned} Cov(w_{lC}, w_{lT}) &= \frac{(\alpha+1-b)(\alpha+2)+\alpha^2 b}{(\alpha+1-b)(\alpha+1)^2(\alpha+2)}\left(\frac{\alpha^2 b+\alpha^2(\alpha+1-b)(\alpha+2)}{(\alpha+1-b)(\alpha+1)^2(\alpha+2)}\right)^{l-1} \\ &\quad -\frac{1}{(\alpha+1)^2}\left(\frac{\alpha^2}{(\alpha+1)^2}\right)^{l-1} \end{aligned} \tag{3.8}$$

The variance for the weights are independent of group, see Appendix, and can be expressed as $Var(w_{ls}) = \frac{2}{(\alpha+1)(\alpha+2)}\left(\frac{\alpha+\alpha^2(\alpha+2)}{(\alpha+1)^2(\alpha+2)}\right)^{l-1} - \frac{1}{(\alpha+1)^2}\left(\frac{\alpha^2}{(\alpha+1)^2}\right)^{l-1}$. Therefore, the correlation, for $l \in \{2, 3, ...\}$, can be expressed as,

$$\begin{aligned} Cor(w_{lC}, w_{lT}) &= \left[\frac{(\alpha+1-b)(\alpha+2)+\alpha^2 b}{(\alpha+1-b)(\alpha+1)^2(\alpha+2)}\left(\frac{\alpha^2 b+\alpha^2(\alpha+1-b)(\alpha+2)}{(\alpha+1-b)(\alpha+1)^2(\alpha+2)}\right)^{l-1}\right. \\ &\quad \left.-\frac{1}{(\alpha+1)^2}\left(\frac{\alpha^2}{(\alpha+1)^2}\right)^{l-1}\right] \Big/ \left[\frac{2}{(\alpha+1)(\alpha+2)}\left(\frac{\alpha+\alpha^2(\alpha+2)}{(\alpha+1)^2(\alpha+2)}\right)^{l-1}\right. \\ &\quad \left.-\frac{1}{(\alpha+1)^2}\left(\frac{\alpha^2}{(\alpha+1)^2}\right)^{l-1}\right] \end{aligned} \tag{3.9}$$

The correlation between the weights for $l \in \{2, 3, ...\}$ also takes values on the interval $(0, 1)$ and behaves the same in terms of the limits of α and b as in the case when $l = 1$. The component value, l, plays a slight role in the correlation, specifically as l get larger, the rate of change for smaller α values becomes less extreme. Figure 3.4 provides surface plots of the correlation for components 20 and 80.

We can now address the covariance and correlation between the two mixing distributions, $Cov(G_C, G_T)$ and $Cor(G_C, G_T)$. Let $B \in \mathbf{\Theta}$ represent a subset of the space of the mixing parameters. In the model we present, $\mathbf{\Theta}$ is equivalent to $\mathbb{R}^2$, so B is simply a subset of $\mathbb{R}^2$. Recall that the mixing distribution for group s has form

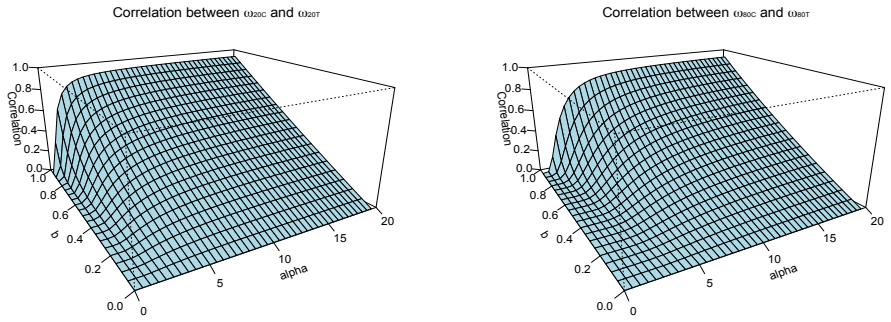

Figure 3.4: Correlation between ω_{20C} and ω_{20T} (left) and ω_{80C} and ω_{80T} (right) over a grid of α and b values.

$G_s(B) = \sum_{l=1}^{\infty} w_{ls}\delta_{\boldsymbol{\theta}_l}(B)$. Marginally, $G_s(B)$ follows a DP, so the expectation and variance of $G_s(B)$ is $G_0(B)$ and $G_0(B)[1-G_0(B)]/(\alpha+1)$, respectively. The covariance between $G_C(B)$ and $G_T(B)$ is given by $Cov\left(\sum_{l=1}^{\infty} w_{lC}\delta_{\boldsymbol{\theta}_l}(B), \sum_{l=1}^{\infty} w_{lT}\delta_{\boldsymbol{\theta}_l}(B)\right)$, which boils down to the expression, $G_0(B)\sum_{l=1}^{\infty} w_{lC}w_{lT} + 2G_0^2(B)\sum_{l=1}^{\infty}\sum_{m=l+1}^{\infty} w_{lC}w_{mT} - G_0^2(B)$. The infinite series converges under geometric series (see appendix for details), and the covariance simplifies to be:

$$Cov(G_C(B), G_T(B)) \quad = \quad G_0(B)(1-G_0(B))\left(\frac{(\alpha-2)b+\alpha+2}{\alpha(2\alpha-3b+5)-2b+2}\right) \tag{3.10}$$

The correlation, therefore, does not depend on the choice of B or G_0, it is driven by α and b alone:

$$Cor(G_C(B), G_T(B)) \quad = \quad \frac{(\alpha+1)((\alpha-2)b+\alpha+2)}{\alpha(2\alpha-3b+5)-2b+2} \tag{3.11}$$

The correlation of the mixing distribution lives on the interval $(1/2, 1)$, see Figure 3.5 for a visual. As $\alpha \to 0$ and/or $b \to 1$, the correlation tends to 1. When $\alpha \to \infty$ the correlation tends to $(b+1)/2$ and as $b \to 0$ the correlation tends to $(\alpha+1)/(2\alpha+1)$, so when $\alpha \to \infty$ and $b \to 0$ the correlation goes to $1/2$. Although this correlation space

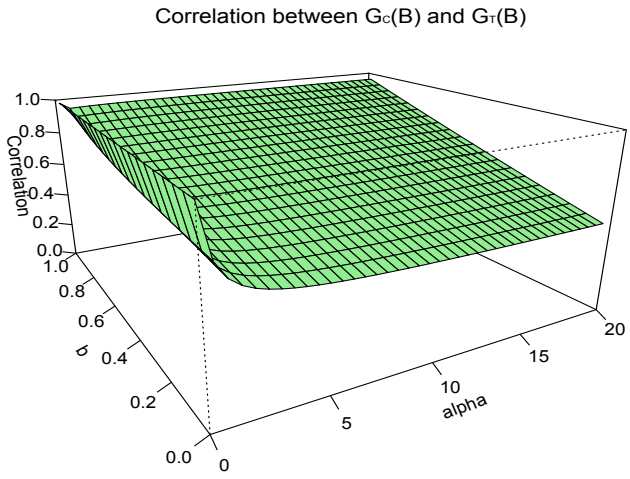

Figure 3.5: Correlation between $G_C(B)$ and $G_T(B)$ over a grid of α and b values.

is limited, it is a typical range seen in the literature (e.g. McKenzie (1985)). It can easily be shown that the correlation of the survival distributions between the two groups given G_C and G_T also live on $(1/2, 1)$, which demonstrates the importance of prior knowledge of the relationship between the distributions of the two group survival times. While the possible values of correlation on the distributions of the survival times are restricted to $(1/2, 1)$, the correlation between the survival times across the two groups, $Cor(T_C, T_T)$, takes on values in $(0, 1)$. The correlation between T_C and T_T is found by marginalizing over the mixing distributions (full details available in appendix), G_C and G_T. Starting with the covariance, $Cov(T_C, T_T) = E[T_C T_T] - E[T_C]E[T_T] = E[E[T_C|G_C]E[T_T|G_T]] - E[E[T_C|G_C]]E[E[T_T|G_T]]$. Under the gamma kernel with bivariate normal G_0 that we have previously discussed, the covariance is given by the following,

$$Cov(T_C, T_T) = \left(e^{t_2'\boldsymbol{\mu}+\frac{1}{2}t_2'\boldsymbol{\Sigma}t_2} - e^{2(t_3'\boldsymbol{\mu}+\frac{1}{2}t_3'\boldsymbol{\Sigma}t_3}\right)\left(\frac{(\alpha-2)b+\alpha+2}{\alpha(2\alpha-3b+5)-2b+2}\right) \quad (3.12)$$

where $t_2 = (2, -2)'$ and $t_3 = (1, -1)'$. The variance of T_s, for both $s \in \{C, T\}$, is given by, $e^{t_1'\boldsymbol{\mu}+\frac{1}{2}t_1'\boldsymbol{\Sigma}t_1} + e^{t_2'\boldsymbol{\mu}+\frac{1}{2}t_2'\boldsymbol{\Sigma}t_2} - e^{2(t_3'\boldsymbol{\mu}+\frac{1}{2}t_3'\boldsymbol{\Sigma}t_3)}$. Recall that $t_1 = (1, -2)'$. Therefore the correlation is given by,

$$\begin{aligned} Cor(T_C, T_T) &= \left[\left(e^{t_2'\boldsymbol{\mu}+\frac{1}{2}t_2'\boldsymbol{\Sigma}t_2} - e^{2(t_3'\boldsymbol{\mu}+\frac{1}{2}t_3'\boldsymbol{\Sigma}t_3)}\right)\left(\frac{(\alpha-2)b+\alpha+2}{\alpha(2\alpha-3b+5)-2b+2}\right)\right] / \\ &\quad \left[e^{t_1'\boldsymbol{\mu}+\frac{1}{2}t_1'\boldsymbol{\Sigma}t_1} + e^{t_2'\boldsymbol{\mu}+\frac{1}{2}t_2'\boldsymbol{\Sigma}t_2} - e^{2(t_3'\boldsymbol{\mu}+\frac{1}{2}t_3'\boldsymbol{\Sigma}t_3)}\right] \end{aligned} \tag{3.13}$$

As the $e^{t_1'\boldsymbol{\mu}+\frac{1}{2}t_1'\boldsymbol{\Sigma}t_1} = E[e^{\theta-2\phi}] \to 0$ the correlation simplifies to $((\alpha-2)b+\alpha+2)/(\alpha(2\alpha-3b+5)-2b+2)$. In this case, as $\alpha \to 0$ the correlation tends to 1 and as $\alpha \to \infty$ the correlation tends to 0. Also, as $b \to 0$ the correlation tends to $1/(2\alpha+1)$ and as $b \to 1$ the correlation tends to $1/(\alpha+1)$. These results are scaled down as $E[e^{\theta-2\phi}]$, the expectation of the kernel variance, gets larger. In Figure 3.6, the surface plot of the correlation between the survival times are shown for different values of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are shown over a grid of values for b and α.

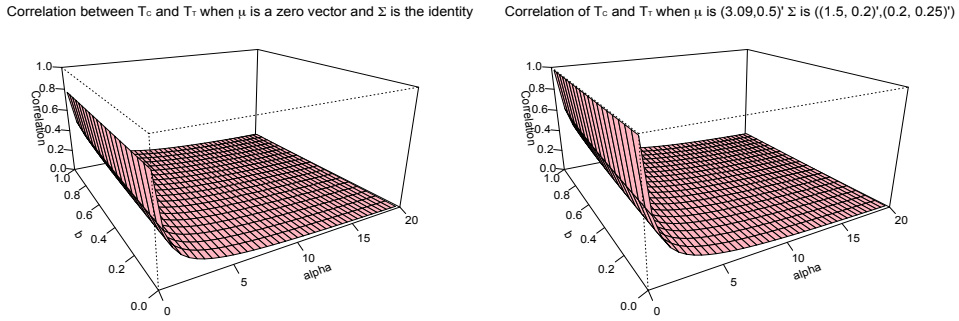

Figure 3.6: Correlation between T_C and T_T when $\mu = (0, 0)'$ and $\Sigma = ((1, 0)'(0, 1)')$ (left) and when $\mu = (3.09, 0.5)'$ and $\Sigma = ((1.5, 0.2)'(0.2, 0.25)')$ (right) over a grid of α and b values.

Having thoroughly explored the DDP prior, we now present the full hierarchical version of the proposed mixture model (3.14). The full hierarchical model can be written

upon introducing the latent configuration variables, $\mathbf{w} = \{\mathrm{w}_{is} : i = 1, ..., n_s | s = C, T\}$, such that $w_{is} = l$ if the i^{th} observation at group s is assigned to mixture component l. Keeping the same kernel structure, baseline distribution, and priors as in (3.3), the hierarchical version of the model with the chosen bivariate beta distribution may be written as follows,

$$
\begin{aligned}
\{t_{is}\}|\mathbf{w}, \{\boldsymbol{\theta}_l\} &\sim \prod_{s\in\{C,T\}1} \prod_{i=1}^{n_s} \Gamma(t_{is}|e^{\theta_{\mathrm{w}_{is}}}, e^{\phi_{\mathrm{w}_{is}}}) \qquad (3.14)\\
\{x_{is}\}|\mathbf{w}, \{\boldsymbol{\theta}_l\} &\sim \prod_{s\in\{C,T\}1} \prod_{i=1}^{n_s} N(x_{is}|\beta_{\mathrm{w}_{is}}, \kappa^2_{\mathrm{w}_{is}})\\
\mathrm{w}_{is}|\{(\zeta_{ls})\} &\overset{ind}{\sim} \sum_{l=1}^{L}\{(1-\zeta_{ls})\prod_{r=1}^{l-1}\zeta_{rs}\}\delta_l(\mathrm{w}_{is}),\\
&\quad \text{for } i = 1, ..., n_s \text{ and } s \in \{C, T\}\\
\{(\zeta_{lC}, \zeta_{lT})\}|\alpha, b &\sim Biv - Beta(\{(\zeta_{lC}, \zeta_{lT})\}|\alpha, b)\\
&\quad \zeta_{lC} = UW, \quad \zeta_{lT} = VW, \text{ for } l = 1, ..., L,\\
&\quad U \overset{iid}{\sim} Beta(\alpha, 1-b),\ V \overset{iid}{\sim} Beta(\alpha, 1-b),\\
&\quad W \overset{iid}{\sim} Beta(1+\alpha-b, b)\\
(\theta_l, \phi_l, \beta_l, \kappa_l^2)'|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \lambda, \tau^2, \rho &\overset{iid}{\sim} N_2((\theta_l, \phi_l)'|\boldsymbol{\mu}, \boldsymbol{\Sigma})N(\beta_l|\lambda, \tau^2)\Gamma^{-1}(\kappa_l^2|a, \rho)
\end{aligned}
$$

We place the following priors: $\alpha \sim \Gamma(\alpha|a_\alpha, b_\alpha(\text{rate}))$, $b \sim Unif(0,1)$, $\boldsymbol{\mu} \sim N_2(\boldsymbol{\mu}|a_\mu, B_\mu)$, $\boldsymbol{\Sigma} \sim IWish(\boldsymbol{\Sigma}|a_\Sigma, B_\Sigma)$, $\lambda \sim N(\lambda|a_\lambda, b_\lambda)$, $\tau^2 \sim \Gamma^{-1}(\tau^2|a_\tau, b_\tau)$, and $\rho \sim \Gamma(\rho|a_\rho, b_\rho)$. The posterior sampling algorithm details are provided in Appendix D.1.

3.3.3 Simulation

In this section, we construct two sets of populations from which to sample from. The first set of populations is were constructed using a mixture of Weibull distributions that shared the same set of locations, but having different weights. Given that model (3.14) has DDP prior has the same construction, we would expect the model perform well under this situation. The populations for the first simulation is shown in the left panel in Figure 3.7. The panel shows how the two populations look similar having modes at the same locations, just differing prevalences for each mode. The second set of populations is also constructed using a mixture of Weibull distributions, however, this time we use different weights as well as locations. The intention is to test the model's inferential ability for populations that have quite different features. Figure 3.7 shows the density populations of the second simulation in the right panel. The second population exhibits a single mode in between the two modes of the first population. The panel indicates the that the two densities are quite dissimilar.

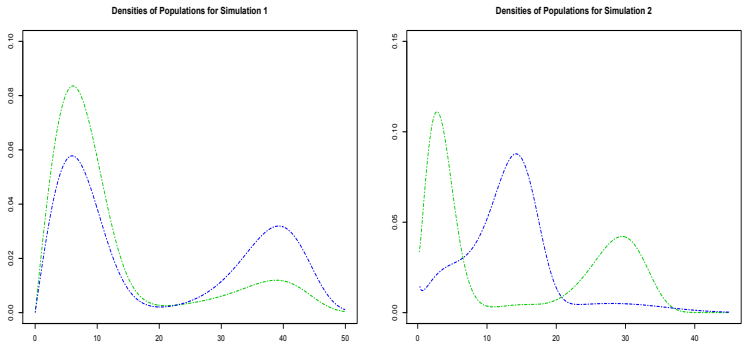

Figure 3.7: Simulation 1 population densities (left) and Simulation 2 population densities (right). The green curve represents the first population (T_1) while the blue represents the second (T_2) in each simulation.

Simulation 1

In Simulation 1, we demonstrate the models' ability to perform under circumstances in which resembles the structure of our model. Specifically, we simulate from two Weibull mixture distributions that share mixture locations, but have different weights:

$$T_1 \sim 0.7Weib(2,8) + 0.1Weib(3,10) + 0.05Weib(4,30) + 0.15Weib(8,40)$$

$$T_2 \sim 0.5Weib(2,8) + 0.05Weib(3,10) + 0.025Weib(4,30) + 0.425Weib(8,40)$$

The populations are comprised of four components each. We sample 250 survival times from the first population and 100 survival times from the second population. We do not consider censoring or covariates here. The histogram of the simulated survival data is shown in Figure 3.8

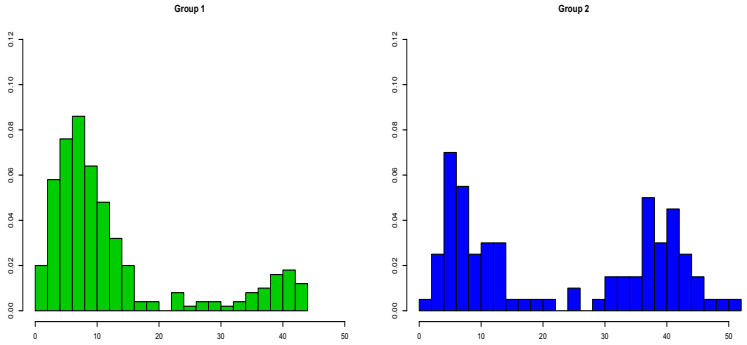

Figure 3.8: Simulation 1. Simulated survival times from mixture of Weibull having the same locations and different weights.

We place a Uniform prior on the b parameter and a Gamma prior on α with shape parameter 2 and rate parameter 0.8. The number of components is conservatively set at 40. Using prior specification methods discussed in Section 2.2.2, we place a bivariate normal prior on μ with mean vector $(1.87, 0.25)'$ and covariance matrix

$((0.27, 0)', (0, 0.27)')$, and an inverse Wishart with 4 degrees of freedom and scale matrix $((0.27, 0)', (0, 0.27)')$. We update b and α together using a bivariate normal on the logit and log scale, respectively. The proposal is centered around the previous iteration, and initial MCMC runs are done to obtain an appropriate covariance matrix. After burn in and thinning, we obtain 2000 independent posterior samples.

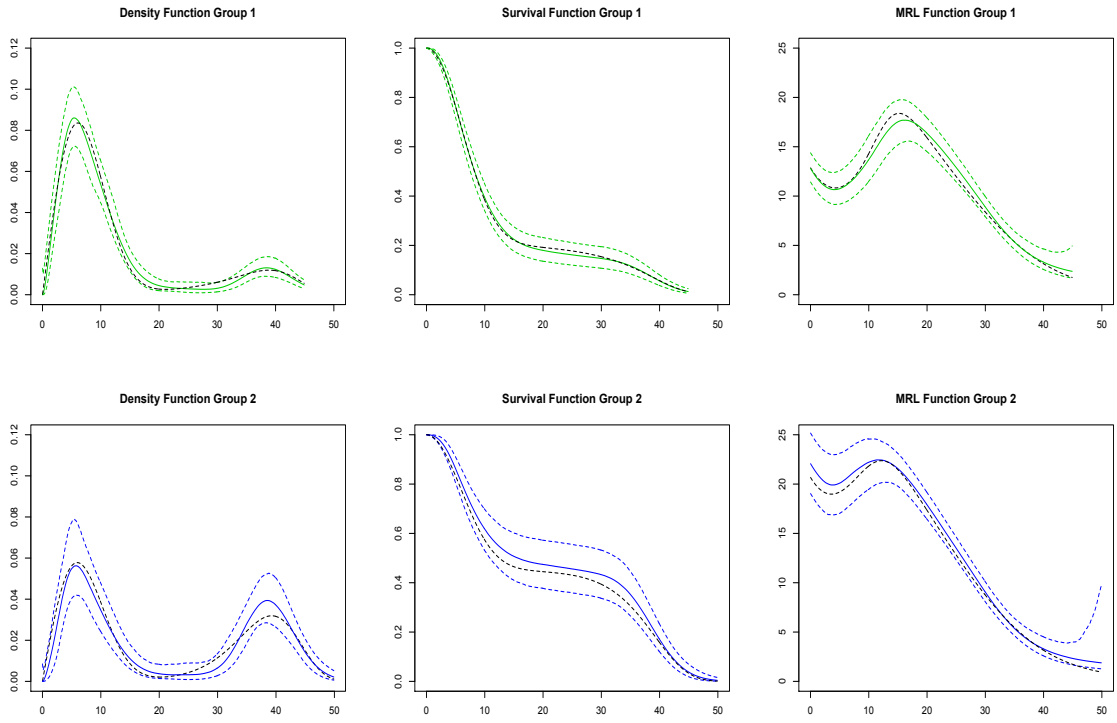

Figure 3.9: Simulation 1. Posterior point and 95% interval estimates for density (left), survival (middle), and mrl (right) functions. The truth is given by the black dashed curves.

Inference for the density, survival, and mrl functions are provided in Figure 3.9. The top panels are results for Group 1, while the bottom panels are that of Group 2. The colored solid and dashed curves represent the poster point and 95% interval

estimates. The truth is plotted, in a dashed black curve, over the posterior results. The model is able to express the features of the functionals, and the true population density is captured within the 95% interval estimates. In particular, the flexibility of the model is demonstrated in the mrl function. The true mrl is non-standard in both groups: initially decreasing, followed by an increase after about time 5, and then decreasing again after about time 12. The difference in sample size between the two groups is indicated by the slightly larger interval bands in Group 2 for the majority of the support of the data.

Simulation 2

The second simulation example is intended to be more of a challenge to the model. The populations consist of mixtures of Weibull distributions, however, here we use different weights, locations, and number of components. Group 1 is comprised of four components, while Group 2 is comprised of five:

$$T_1 \sim 0.5Weib(2,4) + 0.05Weib(0.6,4) + 0.025Weib(5,15) + 0.425Weib(8,30)$$

$$T_2 \sim 0.02Weib(0.6,1) + 0.02Weib(2,4) + 0.66Weib(5,15) + 0.2Weib(2,8) + 0.1Weib(4,30)$$

We simulate 250 observations from each population. All observations are fully observed, and no covariates are considered. The histogram of the simulated survival data is shown in Figure 3.10.

Once again, we use a uniform prior on b, and gamma prior on α with shape parameter 2 and rate parameter 0.8. The number of components is set at 40, which is a

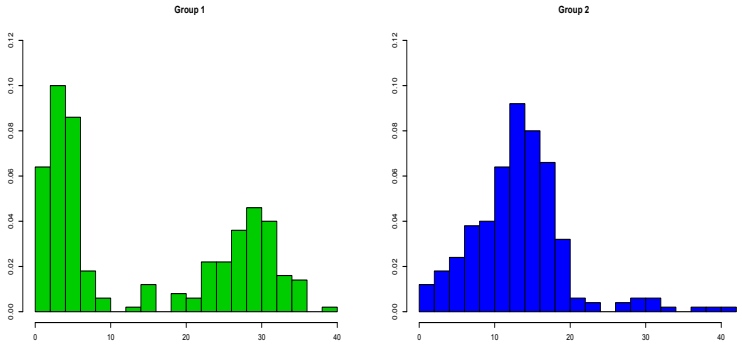

Figure 3.10: Simulation 2. Simulated survival times from mixture of Weibull having the same locations and different weights.

conservative value for these data. Using the same prior specification approach discussed in Section 2.2.2, a bivariate normal prior is placed on $\boldsymbol{\mu}$ with mean vector $(3.02, 0.54)'$ and covariance matrix $((0.1, 0)', (0, 0.1)')$. We update α and b the same way as in the first simulation. After burn in and thinning, we obtain 2000 independent posterior samples.

The posterior results for α, b, and the correlation between the mixing distributions are shown in Figure 3.12. The prior densities are shown in the plots as the red dashed line. The model favors smaller α values, which is not surprising since the number of components in the populations are small. The posterior for b also favors smaller values. Recall that in general, smaller b indicates smaller correlation. The model is likely trying to reflect the difference between the populations. On the other hand, smaller α values lead to a higher correlation. The posterior correlation between the mixing distributions seemw to settle between the competing values of α and b at around 0.7. The 95% credible interval for the correlation is given by $(0.619, 0.855)$.

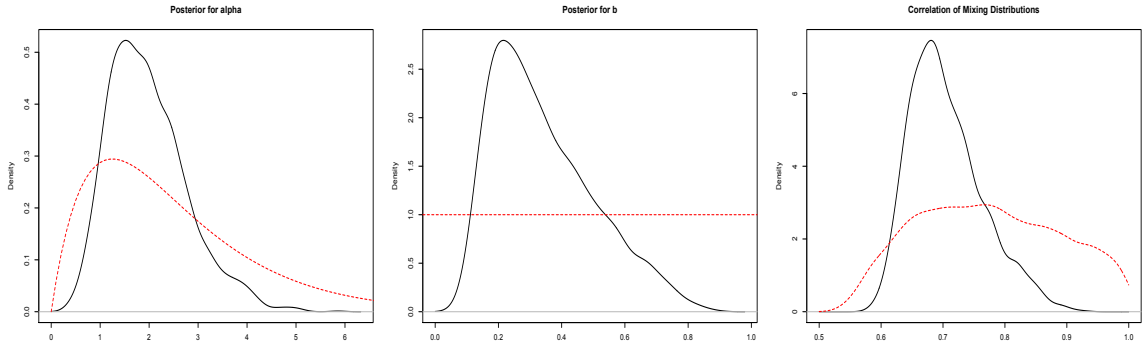

Figure 3.11: Simulation 2. Posterior point and 95% interval estimates for α (left), b (middle), and the correlation between the mixing distributions (right). The priors for α and b are given by the red dashed line.

The posterior results for the density, survival, and mrl functions are shown in Figure 3.12. Despite the difference in the features of the functionals between the two groups, the model is able to capture the features of each group with accuracy. This is especially exciting for the mrl functions. The mrl functions are quite different from one another, and both are non-standard shapes. The model has no problem capturing both shapes of the mrl functions. The only area where we can see struggle in the model for the mrl function inference is in the tails of the functionals. The true mrl function of Group 1 is slightly above the upper interval estimate of the model. This may be just due to the random nature of simulated data; this simulated data may suggest a lower mrl function in the tail. Another possibility is the extreme difference between the mrl functions of the two groups in the tails. Group 1 shoots up sharply, while Group 2 remains gradually decreasing. A third contributor to the tail struggle is that the sparsity of the data in this area, so models in general a have a tougher time achieving accuracy. Even with these elements against the model, the struggle is not significant.

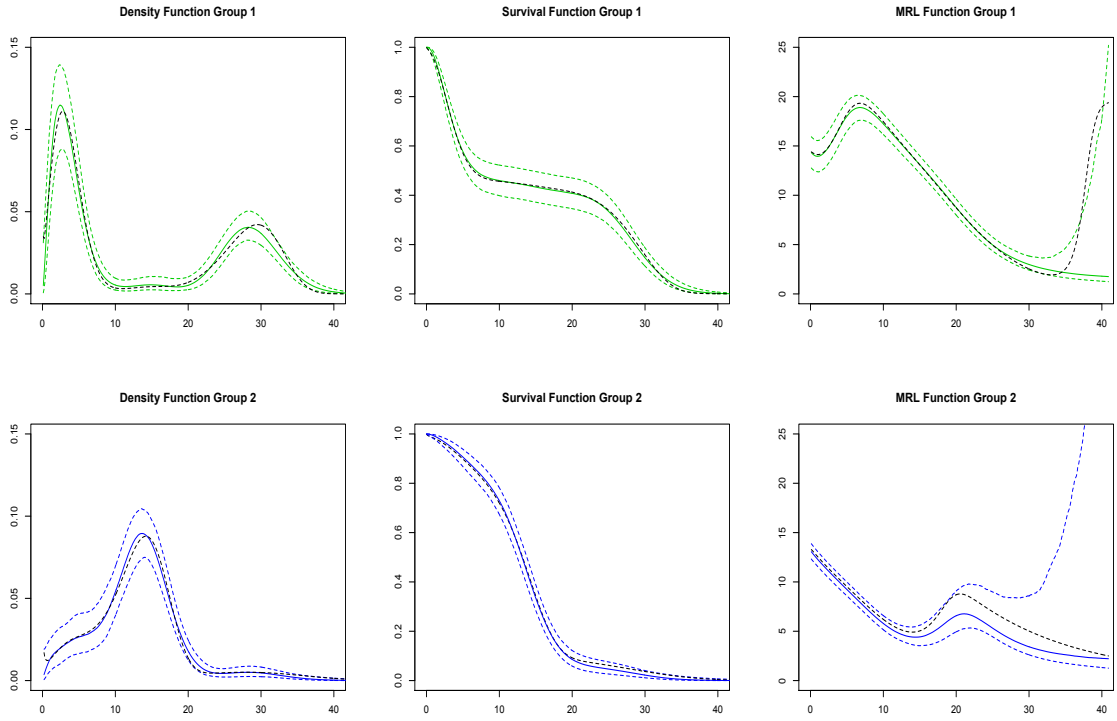

Figure 3.12: Simulation 2. Posterior point and 95% interval estimates for density (left), survival (middle), and mrl (right) functions. The truth is given by the black dashed curves.

The results from the two simulations demonstrates the utility of the gamma DDPMM. The model is able to incorporate dependency across two populations to achieve accurate inference in the functionals of each population. In particular, the model is able to capture provide flexible mrl inference for two groups that exhibit mrl functions having different features across the range of survival. In the following section, we apply the gamma DDPMM to a real dataset, and provide inferential results.

3.4 Data example: small sell lung cancer

We revisit the data set analyzed in Section 2.3.3 under the gamma DPMM. These data describe the survival times of patients with small cell lung cancer under two treatment groups, Arm A and Arm B. Arm A consists of 62 survival times, 15 of which are right censored. Arm B consists of 59 survival times, 8 of which are right censored. The age of each patient upon entry is also available, however, in Section 3.4.1, we work with the treatment as the only covariate. The age covariate is later incorporated in Section 3.4.2 by combining both methodologies described in Sections 3.2 and 3.3.

3.4.1 Dependency across treatment groups

We fit a DDPMM using a gamma kernel to these data. Priors were specified using an analogous approach as described in Section 2.2.2, i.e., using the range and midrange of the observed survival times, which, in practice, would be specified by the expert. We place a uniform prior on b and a gamma prior with shape parameter 2 and rate parameter 0.5 is placed on α, and set $L = 80$. In Figure 3.13, the prior and posterior densities are overlaid for α, b, and the correlation between the mixing distributions, $Cor(G_C, G_T)$. The posterior densities for both α and b indicate learning for these parameters. Consequently, the model is able to learn about the correlation between the G_C and G_T. These data imply a fairly strong correlation between the mixing distributions as well as between the population distributions of the survival times under Arm A and Arm B.

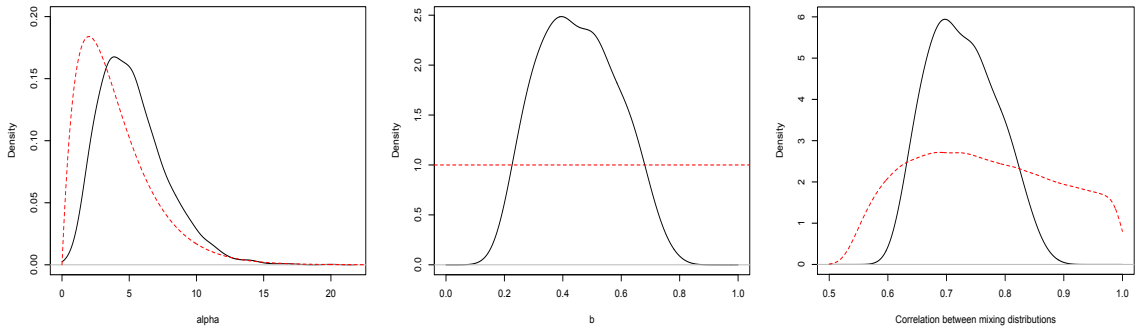

Figure 3.13: Prior (red dashed) and posterior (black solid) densities for α (left), b (middle), and $Cor(G_C, G_T)$ (right).

Inference for the density, survival, and mrl functions are provided in Figure 3.14. The point estimates for the density have the same general shape to the point estimates obtained by Kottas & Krnjajić (2009), who employ a semiparametric regression model. Both models indicate a mode at about 450 days for Arm A and 350 days for Arm B. However, the point estimates under the gamma DDPMM are smoother than under the semi-parametric regression model for both groups. The difference is seen more obviously in the Arm B treatment. The point estimates for the two survival curves indicates that Arm A has a higher survival rate across the range of the data starting from about 200 days. When comparing the results under the gamma DDPMM from under the independent gamma DPMM, the general conclusion regarding favorability of Arm A over Arm B remains the same, however, there an obvious change in the mrl functions. Although, the point estimates for the mrl functions maintain the same non-standard shape under both models, the separation between the two groups is far less under DDPMM compared to the DPMM (Figure 2.6). Arm B is the group that appears

to be most affected by the model change. Specifically, the point estimate for Arm B is shifted up. The shift is most drastic in the tail where data become more sparse.

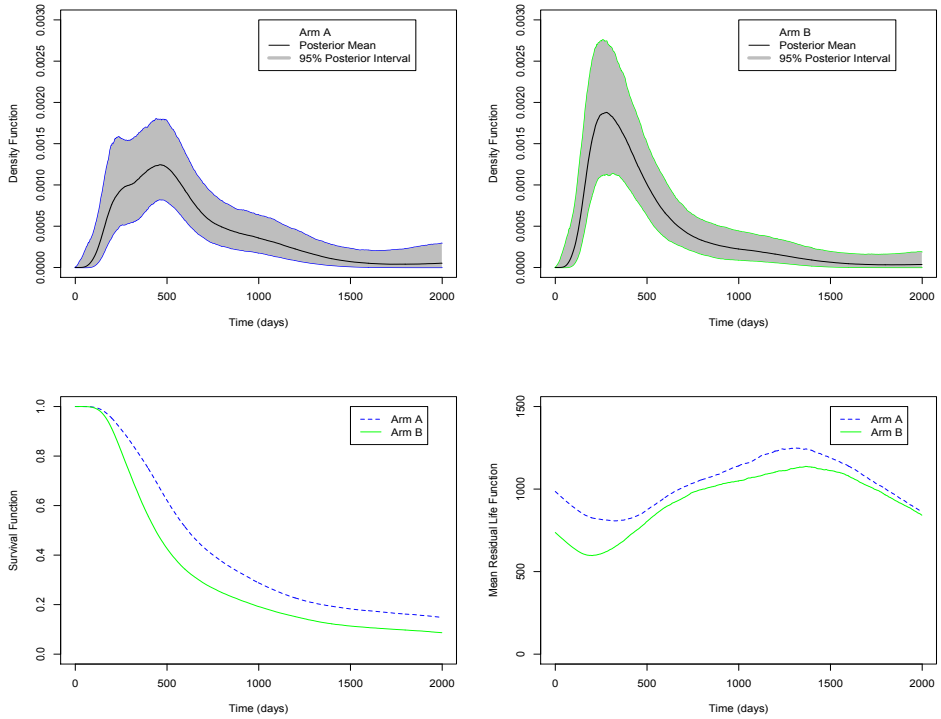

Figure 3.14: Posterior point and 95% interval estimates of the density function for Arm A (upper left) and Arm B (upper right). Posterior point estimate of the survival function (bottom left) and the mean residual life function (bottom right) for Arm A (blue dashed) and Arm B (green solid).

In Figure 3.15, we look at the prior probability, $Pr(m_A(t) > m_B(t))$, and posterior probability, $Pr(m_A(t) > m_B(t)|data)$, under the gamma DDPMM. This Figure is analogous to Figure 2.8, which provides results under the independent gamma DPMM. The prior probabilities under both models do not favor one mrl function over the other at any time point. We also see from Figure 2.8 and Figure 3.15 that the posterior probability changes in a similar fashion as we move across the time space. Specifically, the

probability is highest at smaller survival times then dips down followed by an increase and then then tapers back down. The range in probabilities is larger in Figure 3.15, with some probabilities reaching below 0.6. In particular, Figure 3.15 indicates a lower probability of the mrl function of Arm A being higher than the mrl function of Arm B after about 500 days when compared to Figure 2.8.

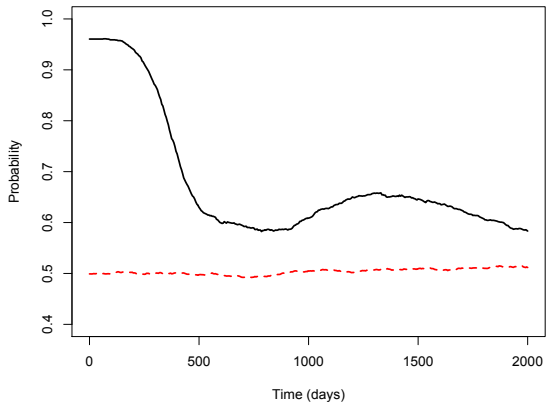

Figure 3.15: The posterior (black solid) and prior (red dashed) probability of the mrl function of Arm A being higher than the mrl function of Arm B over a grid of survival times (days).

We formally compare the performance of the gamma DDPMM to a parametric regression model that is able to obtain constant, increasing, decreasing, UBT, and BT shaped mrl functions. Namely, we extend the Exponentiated Weibull model (EWM), see Section 2.1.3, to a regression model through the scale parameter, σ, by setting $log(\sigma) = \beta_0 + \beta_1 x$. The survival function thus becomes, $S(t) = 1 - [1 - exp\left(-t^{\alpha} exp\left(\beta_0 + \beta_1 x\right)\right)]^{\theta}$. Here x takes value 0 when the survival time corresponds to Arm A, and a 1 when the survival time corresponds to Arm B. A normal prior with mean -10 and standard

deviation 10 was placed on β_0, and a normal prior with mean0 and standard deviation 10 was placed on β_1. Exponential priors with rate parameters 1/1.1 and 1/0.9 were placed on α and θ, respectively. Prior parameters were specified by taking advantage of the closed form survival function of the EWM. The 10^{th}, 50^{th}, and 90^{th} quantiles of the data, which in practice would be specified by an expert, were used as arguments of the inverse survival function to obtain prior point estimates of the parameters.

The Conditional Predictive Ordinate (CPO), originally proposed as the leave-one-out cross-validation predictive density by Geisser & Eddy (1979), is a useful tool is assessing the performance of a model for a particular dataset. The CPO value of the i^{th} observation, CPO_i, is the marginal posterior probability of observing the i^{th} observation, t_i, when the model is fit to the data with t_i omitted:

$$CPO_i = f(t_i|data_{(-i)}) = \int f(t_i|\mathbf{\Psi}, x_i)\pi(\mathbf{\Psi}|data_{(-i)})d\mathbf{\Psi}$$

where $data_{(-i)}$ represents the data with the i^{th} observation removed, $\mathbf{\Psi}$ are the parameters of the model, x_i is the set of covariates associated with t_i, $f(\cdot)$ is the likelihood, and $\pi(\cdot)$ is the joint posterior distribution of the model parameters. A higher CPO value indicates a better model fit. A benefit of omitting the data value is that the data is only used once in assessing the model. The downfall, having to fit the model for each desired CPO value. However, CPO_i can be expressed in terms of the joint posterior distribution of the model parameters given ALL the observations:

$$CPO_i = \left(\int \frac{1}{f(t_i|\mathbf{\Psi}, x_i)}\pi(\mathbf{\Psi}|data)d\mathbf{\Psi}\right)^{-1}$$

In either case, the expression for CPO_i often does not have a closed form, so the MCMC

approximation is used (see, for example, Chen et al. (2000)). The approximation must be monitored to ensure convergence is obtained. In some cases, the inverted likelihood can cause the estimator to be unstable. For these CPO values, the original definition must be used for the MCMC approximation. If the number of values of instability is small, this is not a large inconvenience, but if there are numerous instability cases, this model assessment technique may be practically unfeasible.

Obtaining the CPO values for the EWM is straight forward. The likelihood is given by $f(t_i|\alpha,\theta,\beta_0,\beta_1,x_i) = \theta\alpha t_i^{\alpha-1} exp\{\beta_0+\beta_1 x_i\}\left[1-exp\left(-t_i^{\alpha}exp\{\beta_0+\beta_1 x_i\}\right)\right]^{\theta-1}$ $\left[exp\left(-t_i^{\alpha}exp\{\beta_0+\beta_1 x_i\}\right)\right]$. Recall that x_i is 0 if t_i is an observation from Arm A and 1 if t_i is an observation from Arm B. Let M represent the number of MCMC iterations, then the CPO values for the EWM are estimated using the posterior samples of the parameters via the Harmonic mean estimator:

$$CPO_i \approx \left(\sum_{j=1}^{M}\frac{1}{f(t_i|\alpha_j,\theta_j,\beta_{0j},\beta_{1j},x_i)}\right)^{-1} \tag{3.15}$$

If the i^{th} observation is right censored, the likelihood is replaced by the survival function. The MCMC for the EWM was ran for 500000 iterations with a burn in of 10000 for an effective posterior sample size of 2000. The CPO values are plotted in red in Figure 3.14.

The DDPMM requires a slightly different expression for the CPO values. Under the DDPMM, each observation is believed to come from a particular component of the mixture, so while the density function is a mixture, the likelihood contribution of the i^{th} observation is a single component. Recall that the full posterior distribution from

which we obtain samples from is $p(\boldsymbol{\theta}, \mathbf{w}, \boldsymbol{p}, \boldsymbol{\Psi})$, where $\boldsymbol{\Psi} = (\boldsymbol{\mu}, \boldsymbol{\Sigma})$. In the small cell lung cancer dataset, we have two experimental groups indexedby $s \in \{C, T\}$, thus for these data we compute CPO values for the i^{th} observation in group s, CPO_{is}. Again, let M represent the number of MCMC iterations, then the expression we need to approximate the CPO values using the posterior parameter samples is given by:

$$CPO_{is} \approx \frac{A}{B_{is}} \left(\sum_{j=1}^{M} \frac{\sum_{l=1}^{L} \Gamma(t_{is}|\boldsymbol{\theta}_{lj})}{\Gamma(t_{is}|\boldsymbol{\theta}_{\mathbf{w}_{isj}})} \right) \tag{3.16}$$

$$\text{where} \quad \frac{A}{B_{is}} = \left(\sum_{j=1}^{M} \frac{1}{\Gamma(t_{is}|\boldsymbol{\theta}_{\mathbf{w}_{isj}})} \right)$$

The details of this derivation is provided in Appendix D.2. The CPO values of the gamma DDPMM are plotted in blue (Arm A) and green (Arm B) in Figure 3.16.

With respect to the CPO values, the gamma DDPMM performs better than the EWM around the mode and the far tail of each density, but collectively it's a close race between the models. The shape of densities of each group are fairly standard, so it is not surprising that a flexible parametric model performs well. The EWM has higher CPO values in the tail of the fully observed data values, but struggles once the it enters the territory of the right censored values. This feature along with the lower CPO values around the mode is indicative of parametric nature of the EWM. Although the EWM performs well, it unable to capture the curvature of the mode while at the same time accurately estimating the skewness of the tail. An obvious compromise is made between capturing features of one aspect and not the other versus losing a little on both but not doing bad on either. The gamma DDPMM does not perform as well as the EWM in the tail of the observed data values, which may be attributed to the

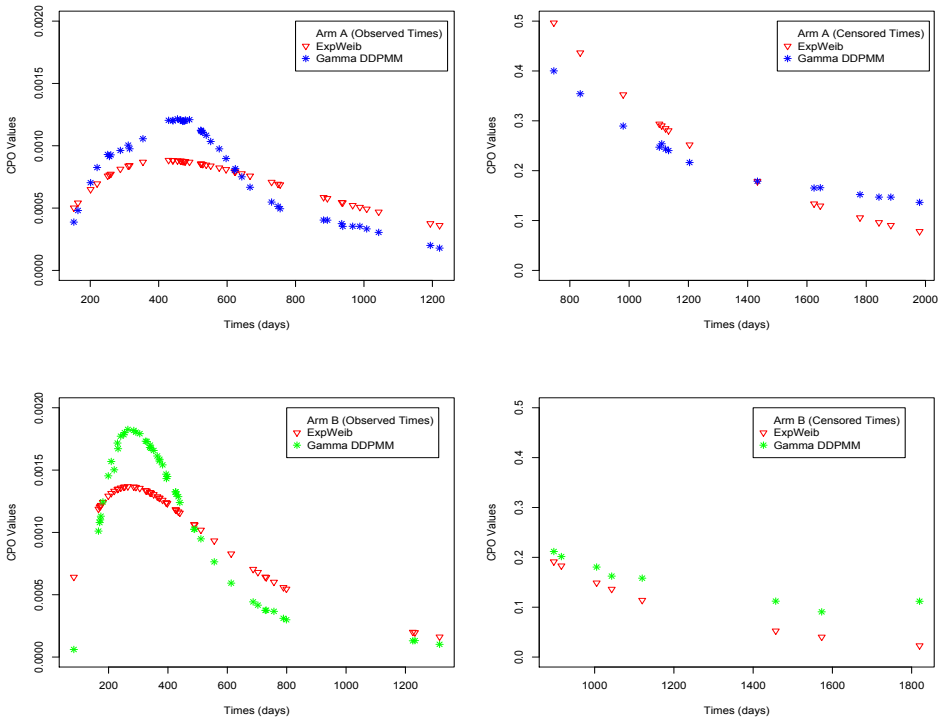

Figure 3.16: The CPO values under the EWM (red) and gamma DDPMM (blue and green) for the small cel lung cancer data. The top panels represent Arm A, and the bottom represent Arm B. The right column are the right censored survival times, and the left column are the fully observed survival times.

right censored observations that share some of the same time space as some of the last few fully observed survival times. A summary of the CPO values were obtained by averaging over the log of the CPO values, ALPML, in each group s:

$$ALPML_s = \frac{1}{n_s} \sum_{i=1}^{n_s} log(CPO_{is})$$

The same rubric applies, meaning a higher ALPML indicates better model performance. The EWM scored a -6.09, while the gamma DDPMM scored a -6.05. The gamma DDPMM performs slightly better than the EWM. We are not too surprised the EWM

scored so close to the gamma DDPMM being as the parametric model was selected after seeing previous inferential results for these data, and was chosen based off the posterior shape exhibited in the density. Kottas & Krnjajić (2009), provide a Bayesian semiparametric model for quantile regression that is a DP scale mixture of uniform densities. The model is fitted to the linear regression errors of the survival times, $\epsilon = t - \boldsymbol{x}^T\boldsymbol{\beta}$, assuming that the median and the mode of the error density is 0. They fit the model to the small cell lung cancer data, without the age covariate, and compute the CPO values. The ALPML that they reported is -6.91. Kottas and Krnjajić also consider a DP scale mixture of Laplace densities and a basic Weibull model for these data. Table 3.1 shows the ALPML value under each of the models.

Model	Summary Value
EWM	-6.09
DP scale mixture of uniform densities	-6.91
gamma DDPMM	-6.05
DP scale mixture of Laplace densities	-8.01
Weibull Model	-11.56

Table 3.1: Summary of the CPO values.

3.4.2 Incorporating the age covariate

Here, we incorporate the age (in years) of the subjects, upon entrance into the study, that is available to us in the small cell lung cancer dataset. The researchers

did not select subjects from particular ages, so it is not a fixed covariate, and can be thought of as being random. Therefore, we model the age covariate on the log scale jointly with the survival response. Specifically, we used independent gamma and normal distributions for the survival times and age, respectively, see model (3.13). We use prior specifications methods discussed in 3.2.2 for parameters associated with the survival times, and use a similar idea to specify parameters associated with the age covariate (see, e.g., Poynor (2010) for details). Appendix D.1 show the details of the posterior sampling algorithm. We run the MCMC to obtain an effective posterior sample size of 2000.

In Figure 3.17, we plot the conditional mean of survival across a grid of ages. This inference was obtained by computing (3.2) at each posterior sample of the parameters for each experimental group. Specifically, the form of the mean regression for group $s \in \{A, B\}$ (A representing Arm A and B representing Arm B) at age x_0 under the gamma DDPMM, is given by,

$$E(t_s|x_0, G_s^L) = \sum_{l=1}^{L} \underbrace{p_{ls}\left(\frac{LN(x_0|\beta_l, \kappa_l^2)}{\sum_{r=1}^{L} p_{rs} LN(x_0|\beta_r, \kappa_r^2)}\right)}_{q_{ls}(x_0|\beta_l, \kappa_l^2)} e^{\theta_l - \phi_l} \tag{3.17}$$

Recall from Section 3.2, the set of functions $q_{ls}(\cdot)$ can be thought of as a new set of weights such that the mean regression is a finite weighted sum of the kernel component means. Moreover, the weights are functions of the covariate, indicating the potential of the model to capture non-standard relationships across the covariate space. This ability is demonstrated in Figure 3.17 where we an increase in the mean survival from about 36 to just after 50, followed by a steeper decline, particularly in Arm B, and then leveling

out at higher ages.

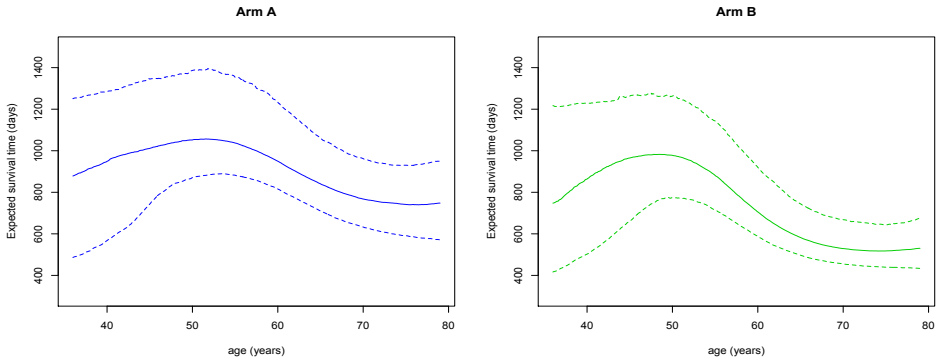

Figure 3.17: Point and 80% interval estimates of the conditional mean of the survival distribution of Arm A (blue) and Arm B (green) across a grid of age values (in years).

We also look at the mrl function at age 50, 60, and 78, see Figure 3.18. At age 50, the mrl function for Arm A appear monotonic while the mrl of Arm B has a very shallow dip at about 400 days then becomes indistinguishable from Arm A. At age 60, the separation becomes more apparent towards in the earlier survival range, and the dips are more pronounced and present in both groups. At age 78, we see a similar curvature as in our past analysis: a dip around $300 - 400$ and a shallow mode around $1000 - 1200$. While the shapes and range of the mrl functions change across the covariate space, Arm A remains as high or higher than Arm B.

The linear DDP by DeIorio et al. (2009) is a leading model for Bayesian nonparametric survival regression. The linear DDP model works with a dependent DP structure for which the collection of random distributions, indexed by a set of covariates, x, denoted $\{G_x,\ x \in X\}$, is almost surely have the form of $G_x(\cdot) = \sum_{l=1}^{\infty} \omega_l \delta_{\theta_{xl}}(\cdot)$. Thus, the weights are shared across the covariate space while the locations differ. One

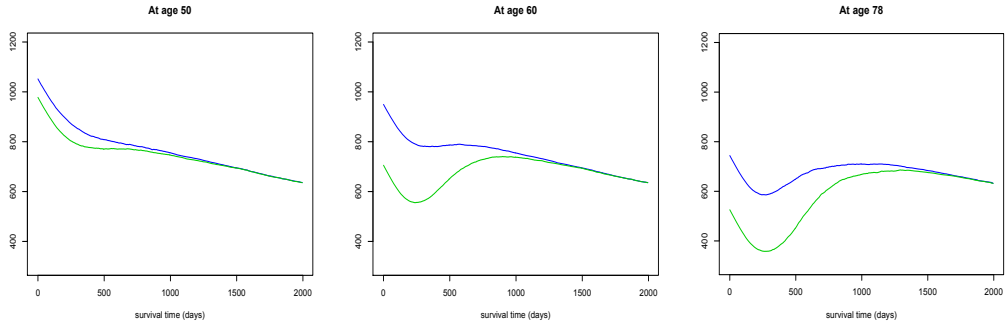

Figure 3.18: Estimates of the mrl function of Arm A (blue) and Arm B (green) for ages 50 (left), 60 (middle), and 78 (right), age is in years.

of difference to note is that under the linear DDP, the covariates are not random, such that in the small cell lung cancer data set, the experimental group and the age would both be part of the set of covariates x. Specifically, under the linear DDP, if we let d_i be the design vector of the i^{th} survival time and α is the vector of coefficients, then $\theta_{x_i l} = \alpha_l d_i = m_l + A_{\nu l} + \beta_l z$. Here, m_l can be thought of the baseline effect, $A_{\nu l}$ represents the effect of a categorical covariate with outcome indicated by ν (experimental group indicated by s in our case), and β_l represents the effect of a continuous covariate z. They chose a normal kernel with mean αd_i and variance σ^2, mixing on both parameters. For $\mathbb{R}^+$ valued responses, modeling is performed on the log transformed data, which is the same as using a lognormal kernel (Poynor, 2010). Thus the model is given by,

$$\begin{aligned} f_{x_i}(t_i|G) &\sim \int LN(t_i|\alpha d_i, \sigma^2) dG(\alpha, \sigma^2) \\ G &\sim DP(M, G_0) \end{aligned} \tag{3.18}$$

where we can write $G \sim DP(M, G_0)$ since $\theta = \alpha d_i$ implies $(\theta, \sigma^2) \sim G_{0x_i}$ by definition of

base measure. Now, if we look at the implications of the mean regression function using the truncated version of the stick-breaking definition, we obtain (3.19) shown below.

$$E_x(t|G_L) = \sum_{l=1}^{L} p_l e^{(\alpha d+\sigma^2/2)} \tag{3.19}$$

The set of covariates only enters into the function as an exponential power, so the mean regression is a strictly increasing function in x (or equivalently d). While the linear DDP provides flexible inference for the survival and density functions, it may not be appropriate in case for which the regression mean or mrl is of interest. Even under a gamma kernel, which is a more appropriate kernel choice for mrl inference, reparameterizing to set the mean to αd would result in a regression mean function that is a weighted sum of linear functions: $E_x(t|G_L) = \sum_{l=1}^{L} p_l \alpha d$. Here, the relationship across the covariate space is subject to a linear trend. Therefore, under such circumstances, the gamma DDPMM would be preferable. In the discussion below, we describe extensions to include potential categorical covariates, however, if the number of experimental group exceeds two, a multivariate beta would have to be explored to extend the DDP.

3.5 Discussion

We have provided an example of a product kernel for a DPMM that jointly models survival times with a continuous covariate, and discussed the model in generality in the case of multiple continuous covariates. We provided a simulation study for a single continuous covariate to demonstrate the utility of the model. We developed a DDP prior using a bivariate beta distribution to model dependency between two

experimental groups. Below we discuss a couple further considerations/extensions to our modeling structure.

Discrete covariates, such as indication of male or female, number of cystic masses, level of pain intensity, are very common in survival analysis. One way to extend the model to include discrete covariates, $\boldsymbol{w}_i$, is to introduce an appropriate density for the kernel product, such that the product kernel becomes $k(t|\boldsymbol{\theta}_T)k(\boldsymbol{x}|\boldsymbol{\theta}_X)k(\boldsymbol{w}|\boldsymbol{\theta}_W)$, where $k(\boldsymbol{w}|\boldsymbol{\theta}_W)$ can be a product of densities such that the categorical covariates are assumed independent in the kernel. In the case where the covariate has a finite upper bound, such as a pain level scale, the binomial is the obvious choice. When the covariate support has no upper bound, the Poisson or negative binomial distribution may be considered.

For added model flexibility, we may seek to incorporate a dependency between the covariates and the survival times within the kernel. A structured approach to build dependence would be to write the kernel as a product of the conditional distribution of the survival times on the covariate and the marginal of the covariates, $k(t|\boldsymbol{x})k(\boldsymbol{x})$. We have made a clear argument for modeling the survival times with a gamma kernel, so one possibility is to incorporate the covariates in one of the parameters of a gamma distribution for $k(t|\boldsymbol{x})$. For example we can write the covariates as a linear combination such as $k(t|\boldsymbol{x}) = \Gamma(t|exp(\theta), exp(\boldsymbol{x}^T\boldsymbol{\beta}))$, such that $E(T|\boldsymbol{x}) = exp(\theta - \boldsymbol{x}^T\boldsymbol{\beta})$. An appropriate marginal kernel, $k(\boldsymbol{x})$, will then complete the joint kernel.

Chapter 4

Modeling and inference for order constrained MRL functions

When a researcher believes that the mean residual life function of one population is higher than the other across the survival domain, then it is desirable to incorporate a mrl order constraint in the model. In this chapter, we present a Bayesian nonparametric model for mrl ordering inference. In Section 4.1, we provide background and motivation for mrl ordering, including a brief literature review on Bayesian work involving ordering constraints. Section 4.2 presents the model formulation and properties. Section 4.3 illustrates the capacity of the model through a data analysis that includes right censored observations.

4.1 Motivation and background

There are certain applications in comparison of survival distributions in which the researcher expects that the average remaining lifetime for one population is higher than that of the other population given survival up to time t, for every t in the domain. For instance, ordered mrl functions may arise in treatment and control groups, in which the population that receives the treatment has a longer remaining life expectancy at all times.

Let T_1 and T_2 represent two continuous random variables on $\mathbb{R}^+$ with mrl functions m_1 and m_2 and distribution functions F_1 and F_2, respectively. We say that T_1 is smaller then T_2 in the mrl order, denoted by $T_1 \leq_{mrl} T_2$, if and only if $m_1(t) \leq m_2(t)$, for all t. For modeling purposes, a useful characterization of the mrl order is the following: $T_1 \leq_{mrl} T_2$ if and only if,

$$M(t) \quad = \quad \frac{\int_t^\infty (1 - F_1(u))du}{\int_t^\infty (1 - F_2(u))du} \text{ is decreasing in } t, \tag{4.1}$$

for all t such that $\int_t^\infty (1 - F_2(u))du > 0$.

Stochastic and hazard rate orders are also of interest in survival analysis. We say that T_1 is stochastically smaller than T_2, denoted by $T_1 \leq_{st} T_2$, if $F_1(t) \geq F_2(t)$, for all t. Similarly, T_1 is smaller than T_2 in the hazard rate order (denoted by $T_1 \leq_{hr} T_2$) if the respective hazard rate functions satisfy $h_1(t) \geq h_2(t)$, for all t. Relationships between the three orders are summarized by Nanda et al. (2010). The hazard rate order is the strongest of the three stochastic constraints, that is, if $T_1 \leq_{hr} T_2$, then $T_1 \leq_{st} T_2$ and $T_1 \leq_{mrl} T_2$. The converse is not true with respect to either mrl ordering

nor stochastic ordering. Also, it is not necessarily the case that mrl order implies stochastic order or vice versa. Therefore, mrl ordering is of interest independently of the more commonly studied stochastic order.

Probability orders have been extensively studied with respect to theoretical properties and classical nonparametric methods for estimation and hypothesis testing (e.g., Shaked & Shanthikumar, 2007). However, when placing an order restriction in a model is appropriate, the Bayesian framework has many advantages. A primary benefit is reduced uncertainty and improved precision in inferential results. In cases where the sample size is small to moderate, there is potential for the data to not accurately reflect the ordering between the two populations. For such cases, there is much to benefit from by incorporating the order constraint in the model. In general, the Bayesian modeling framework is appealing for situations in which ordering constraints are believed, because any probability constraint placed on the prior is carried through to the posterior.

In terms of both probabilistic exploration and statistical modeling, the stochastic order constraint has been the most widely studied in the Bayesian literature. Arjas & Gasbarra (1996) develop a model for stochastically order survival functions by working with a nonparametric specification for the hazard rate functions. Evans et al. (1997) provide a Bayesian testing method for assessing evidence of stochastic ordering in distributions of categorical variables. For stochastic order and partial stochastic order constraints in the Bayesian nonparametric framework, including use of DP priors, see Gelfand & Kottas (2001); Hoff (2003); and Dunson & Peddada (2008). Pólya tree based priors in stochastic ordering have also been explored (Karabatsos & Walker, 2007 and

Hanson et al., 2008). The weaker restriction of stochastic precedence order, introduced by Arcones et al. (2002), has more recently been explored by Chen & Dunson (2004) and Kottas (2011). To our knowledge, nonparametric prior models for mrl order have not been studied in the literature.

4.2 Model formulation

In this section, we walk the reader through our process of developing a prior probability model that provides inference for ordered mrl functions. We begin Section 4.2.1 by further discussing the need to develop a model for mrl ordering specifically, as opposed to usage of models for hazard rate and/or stochastic ordering inferences. We introduce a structured mixture of Erlang distributions and discuss its benefits for our model development. In Section 4.2.2, we develop a hazard rate order constraint using DP priors, which forms a key component of our model for mrl ordering. Finally, in Section 4.2.3, we present the full hierarchical version of the model, and discuss prior specification and posterior inference.

4.2.1 Model properties

As mentioned in Section 4.1, hazard rate order implies mrl order, hence, models that infer hazard rate order also infer mrl order. However, hazard rate order is a stronger constraint, so achieving mrl order through such models is more restrictive than needed for our intentions. Namely, these models will only have support in the space of ordered mrl functions that are also ordered in the hazard rate sense. We have also pointed out

the relationship between stochastic order and mrl order; while hazard rate order implies both stochastic and mrl order, it is not the case that either of the latter two imply the other. Models that have been developed for stochastic order are thus not guaranteed to satisfy the mrl order. Hence, we seek to develop a Bayesian nonparametric model for inference on two mrl ordered populations.

First, we review some equivalence relationships that will be useful tools for our model development (e.g., Shaked & Shanthikumar (2007)). Let Y_1 and Y_2 be non-negative random variables with absolutely continuous distribution functions, G_1 and G_2, survival functions, $\overline{G_1}$ and $\overline{G_2}$, and density functions, g_1 and g_2, respectively. The following are equivalent definitions for the hazard rate order:

$$Y_1 \leq_{hr} Y_2 \quad \Leftrightarrow \quad g_1(t)\overline{G_2}(t) \geq g_2(t)\overline{G_1}(t), \quad \forall t \in \mathbb{R} \tag{4.2}$$

$$\Leftrightarrow \quad \frac{\overline{G_2}(t)}{\overline{G_1}(t)} \text{ is in increasing in } t, \forall t \in \mathbb{R} \tag{4.3}$$

$$\Leftrightarrow \quad \overline{G_1}(x)\overline{G_2}(y) \geq \overline{G_1}(y)\overline{G_2}(x), \quad \forall x \leq y \tag{4.4}$$

By taking the derivative of (4.1), we obtain an equivalent criteria for mrl ordering:

$$Y_1 \leq_{mrl} Y_2 \quad \Leftrightarrow \quad -\overline{G_1}(t)\int_t^\infty \overline{G_2}(u)du + \overline{G_2}(t)\int_t^\infty \overline{G_1}(u)du \leq 0, \quad \forall t \geq 0 \tag{4.5}$$

We take our model inspiration from the structure of the Bernstein polynomial prior model (Petrone, 1999; Petrone & Wasserman, 2002). The Bernstein polynomial model for the density function, $f_l(t)$, where $t \in [0,1]$, of population T_l, for $l = 1, 2$, is

given by

$$\begin{aligned} f_l(t) &= \sum_{j=1}^{M} \omega_j^{(l)} be(t|j, M-j+1) \\ \omega_j^{(l)} &= H_l\left(\frac{j}{M}\right) - H_l\left(\frac{j-1}{M}\right),\ j = 1, ...M. \end{aligned} \tag{4.6}$$

Here, M may be fixed or random, $be(t|j, M-j+1)$ is the density of the beta distribution with mean $j/(M+1)$, and $H_l(\cdot)$ are random distribution functions over $[0, 1]$, for $l = 1, 2$. The key feature of Bernstein priors for our purposes is that the parameters of the beta density basis functions are fixed, which facilitates study of conditions to obtain the mrl order under the prior structure. The implicit restriction is that of an upper bound for the support of the survival distributions.

To overcome the limitation of bounded support for the survival distributions, we consider a mixture of Erlang distributions with fixed shape parameter and random scale parameter. Specifically, if we denote the density of the Erlang distribution with shape parameter $m \in \mathbb{Z}^+$, and scale $\theta > 0$ as $e_m(t|\theta) = t^{m-1}e^{-t/\theta}/((m-1)!\theta^m)$, and survival function as $E_m^S(t|\theta) = \int_t^\infty t^{m-1}e^{-t/\theta}/((m-1)!\theta^m)dt = \sum_{r=0}^{m-1} t^r e^{-t/\theta}/(r!\theta^r)$. The model for the density and survival functions for $t \in \mathbb{R}^+$, $f_l(t) \equiv f_l(t|\theta, M, H_l)$ and $S_l(t) \equiv S_l(t|\theta, M, H_l)$, respectively, with $l = 1, 2$ is given by

$$\begin{aligned} f_l(t) &= \{1 - H_l((M-1)\theta)\}e_M(t|\theta) + \sum_{m=1}^{M-1}\{H_l(m\theta) - H_l((m-1)\theta)\}e_m(t|\theta) \\ S_l(t) &= \{1 - H_l((M-1)\theta)\}E_M^S(t|\theta) + \sum_{m=1}^{M-1}\{H_l(m\theta) - H_l((m-1)\theta)\}E_m^S(t|\theta). \end{aligned} \tag{4.7}$$

Another benefit of using Erlang mixture is that we have a result for denseness, in the weak sense, on the space of continuous mrl functions, see Lemma 2 in Chapter 2. The Bernstein prior has the property of being weakly dense on the space of densities on

$[0,1]$ (Petrone, 1999), but after a transformation, the result is not guaranteed to hold. Furthermore, one would need to obtain a denseness result for space of mrl functions under this prior. Having the denseness result for the Erlang mixture guarantees flexible mrl inference provided M is allowed to be large enough and θ small enough.

Karabatsos & Walker (2007) show that the Bernstein prior can accommodate stochastic ordering for two populations when the distributions that drive the weights are stochastically ordered. Namely, if $H_1 \leq_{st} H_2 \Rightarrow F_1 \leq_{st} F_2$. An analogous result can be obtained for the mixture of Erlang distributions. In particular, assume that $H_1 \leq_{st} H_2$. Then using the stochastic order property of Erlang distributions, that is, $E_m^S(t|\theta) \leq E_{m+1}^S(t|\theta)$ (Marshall & Olkin, 2007), we obtain:

$$\begin{aligned}
S_1(t) &= \{1 - H_1((M-1)\theta)\}E_M^S(t|\theta) + \sum_{m=1}^{M-1}\{H_1(m\theta) - H_1((m-1)\theta)\}E_m^S(t|\theta) \\
&= E_M^S(t|\theta) - H_1((M-1)\theta)E_M^S(t|\theta) + H_1((M-1)\theta)E_{M-1}^S(t|\theta) \\
&\quad -H_1((M-2)\theta)E_{M-1}^S(t|\theta) \pm ... + H_1(2\theta)E_2^S(t|\theta) - H_1(\theta)E_2^S(t|\theta) \\
&\quad +H_1(\theta)E_1^S(t|\theta) - 0 \\
&= E_M^S(t|\theta) + \sum_{m=1}^{M-1} \underbrace{\{E_m^S(t|\theta) - E_{m+1}^S(t|\theta)\}}_{\leq 0} H_1(m\theta) \\
&\leq E_M^S(t|\theta) + \sum_{m=1}^{M-1}\{E_m^S(t|\theta) - E_{m+1}^S(t|\theta)\}H_2(m\theta) \\
&= S_2(t). \qquad (4.8)
\end{aligned}$$

Therefore, $S_1(t) \leq S_2(t), \forall\, t$, and thus $F_1 \leq_{st} F_2$, where F_1 and F_2 are the distributions represented through the Erlang mixture structure in (4.7).

We next turn our attention to the desired model for mrl ordering. In what

follows we explore the mixture of Erlang distributions in (4.7) to see what restrictions can be placed to develop a prior that has the mrl ordering constraint. Using defintion (4.5), we need to show $-S_1(t)\int_t^\infty S_2(u)du + S_2(t)\int_t^\infty S_1(u)du \leq 0$ for all $t \geq 0$. To help simplify this expression, we provide an alternative form for the survival distributions in (4.7). Let m be a generic value in the set $\{2, 3, .., M-1\}$. Then,

$$\begin{aligned} E_M^S(t|\theta) &= \sum_{r=0}^{M-1} \frac{1}{r!}\left(\frac{t}{\theta}\right)^r e^{-t/\theta} = \sum_{r=0}^{M-1} \theta e_{r+1}(t|\theta) \qquad (4.9) \\ E_m^S(t|\theta) - E_{m+1}^S(t|\theta) &= \sum_{r=0}^{m-1} \frac{1}{r!}\left(\frac{t}{\theta}\right)^r e^{-t/\theta} - \sum_{r=0}^{m} \frac{1}{r!}\left(\frac{t}{\theta}\right)^r e^{-t/\theta} \\ &= -\frac{1}{m!}\left(\frac{t}{\theta}\right)^m e^{-t/\theta} = -\theta e_{m+1}(t|\theta). \qquad (4.10) \end{aligned}$$

Using the substitutions (4.9) and (4.10), we obtain,

$$\begin{aligned} S_l(t) &= E_M^S(t|\theta) + \sum_{m=1}^{M-1} \{E_m^S(t|\theta) - E_{m+1}^S(t|\theta)\} H_l(m\theta) \\ &= \theta \sum_{r=0}^{M-1} e_{r+1}(t|\theta) - \theta \sum_{m=1}^{M-1} e_{m+1}(t|\theta) \\ &= \theta e_1(t|\theta) + \theta \sum_{m=1}^{M-1} e_{m+1}(t|\theta)(1 - H_l(m\theta)) \\ &= \theta \sum_{m=0}^{M-1} e_{m+1}(t|\theta)(1 - H_l(m\theta)). \qquad (4.11) \end{aligned}$$

Now, using (4.11) we can write the expression in (4.5) as follows,

$$\begin{aligned} & -\overline{G_1}(t)\int_t^\infty \overline{G_2}(u)du + \overline{G_2}(t)\int_t^\infty \overline{G_1}(u)du \qquad (4.12) \\ = & -\left(\theta \sum_{m=0}^{M-1} e_{m+1}(t|\theta)(1 - H_1(m\theta))\right)\left(\theta \int_t^\infty \sum_{m=0}^{M-1} e_{m+1}(u|\theta)(1 - H_2(m\theta))du\right) \\ & + \left(\theta \sum_{m=0}^{M-1} e_{m+1}(t|\theta)(1 - H_2(m\theta))\right)\left(\theta \int_t^\infty \sum_{m=0}^{M-1} e_{m+1}(u|\theta)(1 - H_1(m\theta))du\right) \end{aligned}$$

$$
\begin{aligned}
&= -\left(\theta \sum_{m=0}^{M-1} e_{m+1}(t|\theta)(1-H_1(m\theta))\right)\left(\theta \sum_{m=0}^{M-1} E^S_{m+1}(t|\theta)(1-H_2(m\theta))\right) \\
&\quad + \left(\theta \sum_{m=0}^{M-1} e_{m+1}(t|\theta)(1-H_2(m\theta))\right)\left(\theta \sum_{m=0}^{M-1} E^S_{m+1}(t|\theta)(1-H_1(m\theta))\right) \\
&= \theta^2 \sum_{m=0}^{M-1}\sum_{n=0}^{M-1} \big[-e_{m+1}(t|\theta)(1-H_1(m\theta))E^S_{n+1}(t|\theta)(1-H_2(n\theta)) \\
&\qquad\qquad + e_{m+1}(t|\theta)(1-H_2(m\theta))E^S_{n+1}(t|\theta)(1-H_1(n\theta))\big] \\
&= \theta^2 \sum_{m=0}^{M-1}\sum_{n=0}^{M-1} e_{m+1}(t|\theta)E^S_{n+1}(t|\theta)[(1-H_1(n\theta))(1-H_2(m\theta)) - (1-H_1(m\theta))(1-H_2(n\theta))]
\end{aligned}
$$

(Note: when $m = n$, the terms are equal to zero)

$$
\begin{aligned}
&= \theta^2 \sum_{m=0}^{M-2}\sum_{n=m+1}^{M-1} e_{m+1}(t|\theta)E^S_{n+1}(t|\theta)[(1-H_1(n\theta))(1-H_2(m\theta)) - (1-H_1(m\theta))(1-H_2(n\theta))] \\
&\quad + \theta^2 \sum_{n=0}^{M-2}\sum_{m=n+1}^{M-1} e_{m+1}(t|\theta)E^S_{n+1}(t|\theta)[(1-H_1(n\theta))(1-H_2(m\theta)) - (1-H_1(m\theta))(1-H_2(n\theta))] \\
&= \theta^2 \sum_{m=0}^{M-2}\sum_{n=m+1}^{M-1} \underbrace{[e_{m+1}(t|\theta)E^S_{n+1}(t|\theta) - e_{n+1}(t|\theta)E^S_{m+1}(t|\theta)]}_{(*)} \\
&\qquad\qquad \times \underbrace{[(1-H_1(n\theta))(1-H_2(m\theta)) - (1-H_1(m\theta))(1-H_2(n\theta))]}_{(**)} \qquad (4.13)
\end{aligned}
$$

The Erlang distribution has increasing hazard rate in the shape parameter (Marshall & Olkin, 2007), so from the definition in (4.2) we know that $(*) \geq 0$ for $t \geq 0$. For the expression denoted by $(**)$, we consider what structure on H_1 and H_2 results in nonpositive summands. Conveniently, $(**)$ is of the same form of the hazard rate ordering definition in (4.4), so if we place the hazard rate ordering constraint on the H functions, namely, $H_1 \leq_{hr} H_2$, then $(**) \leq 0$. With $(*) \geq 0$ and $(**) \leq 0$, each term of the sum is nonpositve, thus the sum of the terms is less than or equal to zero, giving us the desired probabilistic result for mrl ordering under the Erlang mixture model. In the

next section, we develop the full Bayesian model based on the Erlang mixture structure with a nonparametric prior for $\{H_1, H_2\}$ that satisfies the hazard rate order restriction.

4.2.2 DP-based prior for hazard rate order

In this section, we construct a prior for $\{H_1, H_2\}$ through two independent DP priors that ensures $H_1 \leq_{hr} H_2$. The use of independent DP priors is stimulated by the stochastic ordering result that has been established via independent DP priors (Gelfand & Kottas, 2001; Hanson et al., 2008; Kottas, 2011). In particular, consider two latent distribution functions, G_1 and G_2. If we define $H_1(\cdot) = G_1(\cdot)$ and $H_2(\cdot) = G_1(\cdot)G_2(\cdot)$, then $H_1 \leq_{st} H_2$. It is convenient to think of H_1 and H_2 as the distribution of w and $max\{\mathrm{w}, z\}$, respectively, where $\mathrm{w} \sim G_1$ and, independently, $z \sim G_2$. By placing independent DP priors on G_1 and G_2, we have a nonparametric prior for stochastic ordering .

Note that this prior model for stochastic ordering does not guarantee hazard rate ordering. A counterexample is obtained when G_1 and G_2 are defined as independent exponential distributions. Let $G_1(t) \equiv Exp(t|\lambda)$ and independently, $G_1(t) \equiv Exp(t|\mu)$, and define H_1 and H_2 as discussed in the previous paragraph. The ratio of the survival functions for H_1 and H_2 is written as,

$$\frac{(1-H_2(t))}{(1-H_1(t))} = \frac{(1-G_2(t))}{(1-G_1(t))(1-G_2(t))} = \frac{e^{-\mu t}}{1-(1-e^{-\lambda t})(1-e^{-\mu t})} = \frac{e^{-\mu t}}{e^{-\mu t}+e^{-\lambda t}-e^{-(\mu+\lambda)t}} = \frac{1}{1+e^{-(\lambda-\mu)t}-e^{-\lambda t}} \Rightarrow \frac{d}{dt}\left(\frac{(1-H_2(t))}{(1-H_1(t))}\right) = \frac{e^{-\lambda t}\big((\lambda-\mu)e^{\mu t}-\lambda\big)}{\big(1+e^{-(\lambda-\mu)t}-e^{-\lambda t}\big)^2}$$

The derivative shows that the ratio of survival functions is increasing for $0 < \lambda \leq \mu$, however, the function is non-monotonic for $\lambda > \mu$. For example, the ratio is decreasing

for $t > log[\lambda/(\lambda - \mu)]/\mu$ when $1 \leq \mu < \lambda$, hence, (4.3) does not hold so we do not have hazard rate ordering.

As an alternative (albeit related) construction, consider the following. Let $U = min\{\mathrm{w}, z\}$ where, again, $\mathrm{w} \sim G_1$ and $z \sim G_2$ for generic distributions, G_1 and G_2, on $\mathbb{R}^+$. The distribution function on U is given by $H_1(u) = Pr(U \leq u) = 1 - Pr(U > u) = 1 - Pr(\{\mathrm{w} > u\} \cap \{z > u\}) = 1 - Pr(\mathrm{w} > u)Pr(z > u) = 1 - (1 - G_1(u))(1 - G_2(u))$. Therefore, the survival distribution of U is given by the product of the of the survival distributions of G_1 and G_2, namely, $(1 - H_1(u)) = (1 - G_1(u))(1 - G_2(u))$. Define $H_2(\cdot) = G_2(\cdot)$.

We now look at the ratio of the survival functions,

$$\frac{(1 - H_2(t))}{(1 - H_1(t))} = \frac{(1 - G_2(t))}{(1 - G_1(t))(1 - G_2(t))} = \frac{1}{(1 - G_1(t))} \tag{4.14}$$

which is an increasing function of $t \in \mathbb{R}$, since the survival function of G_1 is by definition a decreasing function in t. Therefore, by (4.3), $H_1 \leq_{hr} H_2$. Finally, if we let $G_1 \sim DP(\alpha_1, G_{01})$ and $G_2 \sim DP(\alpha_2, G_{02})$, then we have a nonparametric prior model for hazard rate ordering.

An important theoretical aspect to point out is that this model is also a prior model for stochastic ordering since $H_1 \leq_{hr} H_2 \Rightarrow H_1 \leq_{st} H_2$ (however, we provided a counterexample under our construction for which the reverse relation is not true). In general, mrl ordering does not imply stochastic ordering, thus this model is restricted to the space of set of ordered mrl functions that have corresponding stochastically order distributions. A more general prior model for mrl order would not imply stochastic

order as well.

4.2.3 Implementation details

Let t_{1j}, for $j = 1, ..., n_1$, be the observed survival times of one population, and t_{2k}, for $k = 1, ..., n_2$, be the observed survival times of another population. By introducing two sets of latent variables, $\mathbf{w} = \{\mathrm{w}_k : k = 1, ..., n_2\}$, $\boldsymbol{z} = \{z_j : j = 1, ..., n_1\}$, we can write the hierarchical version of our fully nonparametric Bayesian model for inference for mrl ordering between to groups (4.14).

$$
\begin{aligned}
t_{1j}|\mathrm{w}_j, z_j, \theta &\overset{ind}{\sim} \sum_{m=1}^{M-1} E_m(t_{1j}|\theta)\mathbb{1}_{((m-1)\theta, m\theta]}(min\{\mathrm{w}_j, z_j\}) \\
&\quad + E_M(t_{1j}|\theta)\mathbb{1}_{((M-1)\theta,\infty)}(min\{\mathrm{w}_j, z_j\}), \quad j = 1, ..., n_1 \\
t_{2k}|\mathrm{w}_{n_1+k}, \theta &\overset{ind}{\sim} \sum_{m=1}^{M-1} E_m(t_{2k}|\theta)\mathbb{1}_{((m-1)\theta, m\theta]}(\mathrm{w}_{n_1+k}) \\
&\quad + E_M(t_{2k}|\theta)\mathbb{1}_{((M-1)\theta,\infty)}(\mathrm{w}_{n_1+k}), \quad k = 1, ..., n_2 \\
z_j|G_1 &\overset{iid}{\sim} G_1, \quad j = 1, ..., n_1 \\
\mathrm{w}_k|G_2 &\overset{iid}{\sim} G_2, \quad k = 1, ..., n_1 + n_2 \\
G_l|\alpha_l, \phi_l &\overset{ind}{\sim} DP(\alpha_l, G_{0l} \equiv LN(\mu_l, \sigma_l^2)), \quad l = 1, 2 \\
\alpha_l &\overset{ind}{\sim} \Gamma(a_{\alpha_l}, b_{\alpha_l}), \quad l = 1, 2 \\
\mu_l &\overset{ind}{\sim} N(a_{\mu_l}, b_{\mu_l}), \quad l = 1, 2 \\
\sigma_l^2 &\overset{ind}{\sim} \Gamma^{-1}(a_{\sigma_l}, b_{\sigma_l}), \quad l = 1, 2 \\
\theta &\sim \Gamma^{-1}(a_\theta, b_\theta) \\
M &\sim Unif(2, M_{max})
\end{aligned}
\tag{4.15}
$$

We obtain posterior inference by writing the full posterior distribution is the representation of Antoniak (1974):

$p(G_1, G_2, \boldsymbol{z}, \mathbf{w}, \theta, \alpha_1, \alpha_2, \phi_1, \phi_2, M|data) =$

$p(G_1|\boldsymbol{z}, \alpha_1, \phi_1, \theta, data)p(G_2|\mathbf{w}, \alpha_2, \phi_2, \theta, data)p(\boldsymbol{z}, \mathbf{w}, \theta, \alpha_1, \alpha_2, \phi_1, \phi_2, M|data)$. The marginal posterior is obtained upon marginalizing G_1 and G_2 over their respective DP priors (Blackwell & MacQueen, 1973). Appendix E provides the details of the posterior sampling algorithm.

The posterior conditional distributions for G_1 and G_2 follow an updated DP. Specifically, $p(G_1|\boldsymbol{z}, \alpha_1, \mu_1, \sigma_1^2, \theta, data) = DP(\tilde{\alpha}_1, \tilde{G}_{01})$ where $\tilde{\alpha}_1 = \alpha_1 + n_1$ and

$$\tilde{G}_{01} = \tfrac{\alpha_1}{\alpha_1+n_1} LN(\mu_1, \sigma_1^2)) + \tfrac{1}{\alpha_1+n_1} \textstyle\sum_{j=1}^{n_1} \delta_{z_j}(\cdot)$$

Similarly, $p(G_2|\mathbf{w}, \alpha_2, \mu_2, \sigma_2^2, \theta, data) = DP(\tilde{\alpha}_2, \tilde{G}_{02})$ where $\tilde{\alpha}_2 = \alpha_2 + n_1 + n_2$ and

$$\tilde{G}_{02} = \tfrac{\alpha_2}{\alpha_2+n_1+n_2} LN(\mu_2, \sigma_2^2)) + \tfrac{1}{\alpha_2+n_1+n_2} \textstyle\sum_{k=1}^{n_1+n_2} \delta_{w_k}(\cdot)$$

Therefore, once we obtain samples from the marginal posterior,

$p(\boldsymbol{z}, \mathbf{w}, \theta, \alpha_1, \alpha_2, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2, M|data)$, we can obtain the posterior distribution for the weights in (4.7). Specifically, we sample from a Dirichlet distribution, using the associated cdf, for $l = 1, 2$, at each posterior iteration of M and θ:

$$(u_{l1}, u_{l2}, ..., u_{l(M-1)}, u_{lM})$$

$$\sim Dir(\tilde{\alpha}_1\tilde{G}_{0l}(\theta), \tilde{\alpha}_l(\tilde{G}_{0l}(2\theta) - \tilde{G}_{0l}(\theta)), \tilde{\alpha}_l(\tilde{G}_{0l}(3\theta) - \tilde{G}_{0l}(2\theta)), ..., \tilde{\alpha}_l(1 - \tilde{G}_{0l}((M-1)\theta))$$

For the second population, the vector $(u_{21}, u_{22}..., u_{2M})$ is all we need since the vector represents the successive differences of the cdf, H_2, over the grid set $\{0\theta, 1\theta, 2\theta, ..., (M -$

$1)\theta, \lim_{t\to\infty} t\}$. However, to obtain the desired differences from the cdf, H_1, we have to construct the cdf through the product of the survival functions of G_1 and G_2. Essentially, we compute:

$$H_1(n\theta) = 1 - (1 - G_1(n\theta))(1 - G_2(n\theta)) = 1 - (1 - \sum_{m=1}^{n} u_{1m})(1 - \sum_{m=1}^{n} u_{2m})$$

for $n = 1, 2, ..., M - 1$. Once we have the values of the cdf, then we can plug the values into (4.7), and obtain inference for the desired functionals.

For prior specification, we continue the theme of specifying the prior parameters using a range of survival specified by the expert. Let L denote the lower range value and U denote the upper. The most difficult aspect is specifying an appropriate upper bound for the uniform prior on M. One can think about specifying M_{max} based on how fine of a grid might be desirable over the range specified by the expert. Being that θ and M are closely related, one may specify M_{max} while also considering a prior point estimate for θ. If we consider the means of the Erlang components, the lowest mean θ and the highest mean $(M-1)\theta$ should be inclusive of L and U, i.e., $\theta < L$ and $(M-1)\theta > U$. Thus, consider $E(\theta) < L$ and $M_{max}E(\theta) > U$. Once these values have been conservatively chosen, we can specify b_θ via $E(\theta) = b_\theta/(a_\theta - 1)$. Using $a_\theta = 2$ to provide an infinite prior variance for θ, b_θ simply becomes the prior point estimate for θ.

A range and midpoint can be used to specify the parameters of the prior distributions for μ_l and σ_l^2 for $l = 1, 2$. If the approximate mean or median is known, then those values could be used instead. Let C represent the prior centering estimate

of the survival responses. Since the latent variables will correspond with a Erlang component with mean equal to the ceiling value of the latent value, we can use C as prior point estimate for the latent values. In another words we can use the rough estimate similar to the prior specification technique discussed in Chapter 2,

$$\begin{aligned} C &= E_{G_{0l}}(u) = E(E(u|\mu_l, \sigma_l^2)) = E(e^{\mu_l + 0.5\sigma_l^2}) = E(e^{\mu_l})E(e^{0.5\sigma_l^2}) \\ &\approx e^{(a_{\mu_l} + 0.5b_{\mu_l})} e^{(0.5b_{\sigma_l}/(a_{\sigma_l}-1))} \end{aligned} \tag{4.16}$$

Again, similar to techniques provided in Chapter 2, the variance of the latent values, $Var(u)$, can be approximated using $([U-L]/4)^2$. The variance in terms of the hyper-parameters can in turn be approximated by the following,

$$\begin{aligned} Var(u) &= Var(E(u|\mu_l, \sigma_l^2)) + E(Var(u|\mu_l, \sigma_l^2)) \\ &= Var(e^{\mu_l + 0.5\sigma_l^2}) + E((e^{\sigma_l^2} - 1)e^{2\mu_l + \sigma_l^2}) \\ &= E(e^{2\mu_l})E(e^{\sigma_l^2}) - E^2(e^{\mu_l})E^2(e^{0.5\sigma_l^2}) + E((e^{\sigma_l^2} - 1)e^{\sigma_l^2})E(e^{2\mu_l}) \\ &= E(e^{2\sigma_l^2})E(e^{2\mu_l}) - E^2(e^{\mu_l})E^2(e^{0.5\sigma_l^2}) \\ &\approx e^{2b_{\sigma_l}/(a_{\sigma_l}-1)} e^{2a_{\mu_l} + 2b_{\mu_l}} - e^{2a_{\mu_l} + b_{\mu_l}} e^{b_{\sigma_l}/(a_{\sigma_l}-1)} \end{aligned} \tag{4.17}$$

To make this expression obtainable, yet conservative, set $a_{\sigma_1} = a_{\sigma_2} = 2$. By setting $b_{\sigma_1} = b_{\sigma_2} = b_{\mu_1} = b_{\mu_2}$ and $a_{\mu_1} = a_{\mu_2}$, all the hyperparameters in (4.15) and (4.16) can be specified.

The prior parameters for α_1 and α_2 can be specified using the same expression in Chapter 2, provided that the data set is moderately large. Specifically, an appropriate range for α_1 can be obtained by considering the number of distinct latent variables in

$\{z_1, z_2, ..., z_{n_1}\}$ and using the approximation, $E(n_z^*|\alpha_1) \approx \alpha_1 \log\left(\frac{\alpha_1+n_1}{\alpha_1}\right)$. Similarly, for α_2 we can use $E(n_w^*|\alpha_2) \approx \alpha_2 \log\left(\frac{\alpha_2+n_1+n_2}{\alpha_2}\right)$. Note that the sample size associated with α_2 is necessarily larger than that of α_1, so specifying priors for each makes more sense than assuming the same prior on both.

4.3 Data example: small cell lung cancer

The small size and relatively high number of right censored values, makes the small cell lung cancer data set an especially interesting data set for which to apply our model for mrl ordering. Small sample sizes and right censoring yield larger uncertainty in the mrl function estimates. By incorporating the mrl order constraint, we anticipate a reduction in the uncertainty in the functional inference when compared to the uncertainty we obtained under fitting independent gamma DPMMs (Chapter 2) as well as the results obtained in the gamma DDPMM (Chapter 3).

Using the prior specification discussed in the previous section as a guide, we analyze the data under a number of priors to gage its sensitivity. Under priors that provided less information, we found that the model struggled to learn about the α parameters and θ, which is likely due to the smaller sample size and larger number of censoring in this dataset. Therefore, we considered fixed α parameters and θ. We also implemented the model under informative priors on α_1, α_2, and θ centered about the fixed values we report here. The inferential results under the informative priors were analogous to those obtain under the fixed values. Here, we report our results under

fixed values for α_1, α_2, and θ. Priors were specified by considering a conservative range of the data since there a high number of right censored values. Under the fixed prior set, we set $M_{max} = 700$, and fixed $\theta = 15$. For the α parameters we set $\alpha_1 = 5$ and $\alpha_2 = 20$. The hyperparameters of the DP priors were set to be the following, $a_{\mu_1} = a_{\mu_2} = 6.3$, $b_{\mu_1} = b_{\mu_2} = 0.5$, $a_{\sigma_1} = a_{\sigma_2} = 2$, and $b_{\sigma_1} = b_{\sigma_2} = 0.5$. We ran the model and obtain 2000 independent posterior samples.

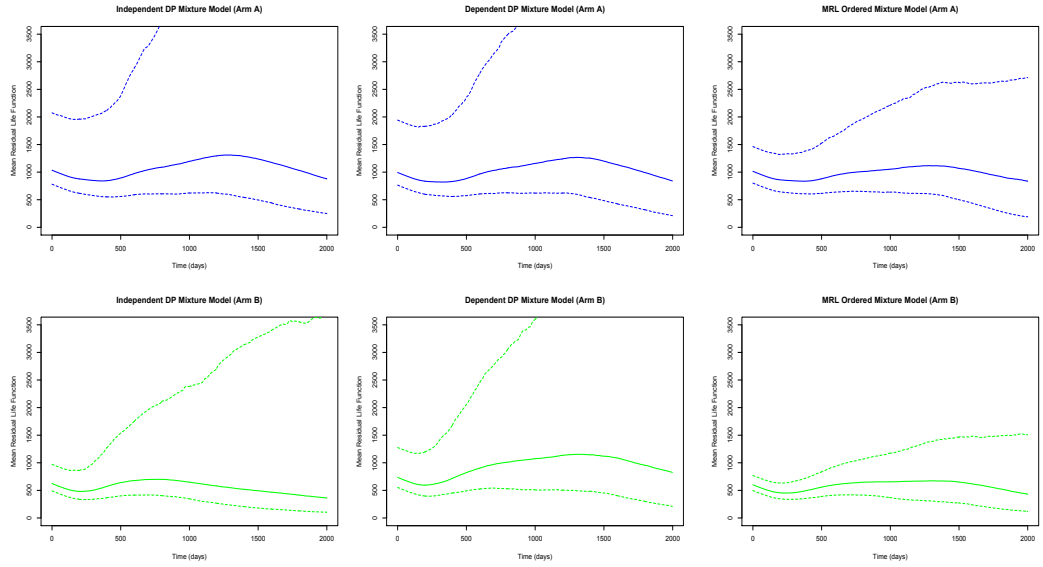

Figure 4.1: Posterior point estimate and 95% interval bands for the mrl functions of Arm A (green) and Arm B (blue) under two independent gamma DPMMs (left), the gamma DDPMM (middle), and the ordered mrl model (right)

The posterior point estimate and 95% interval bands for the mrl functions of Arm A and Arm B, for our three comparison models are shown in Figure 4.1. Results for Arm A are displayed across the top panels and results for Arm B are displayed across the bottom. The point estimates across the three models are very similar in terms of the range and, more importantly, the shape. The order constrained model maintains the

same dip and mode over the same range that we have seen in the gamma DPMM and the gamma DDPMM. The apparent difference between the order constrained model and the other two is the significant reduction in uncertainty for both groups. Incorporating the mrl ordering in the prior provides for narrower posterior interval bands in the mrl functions.

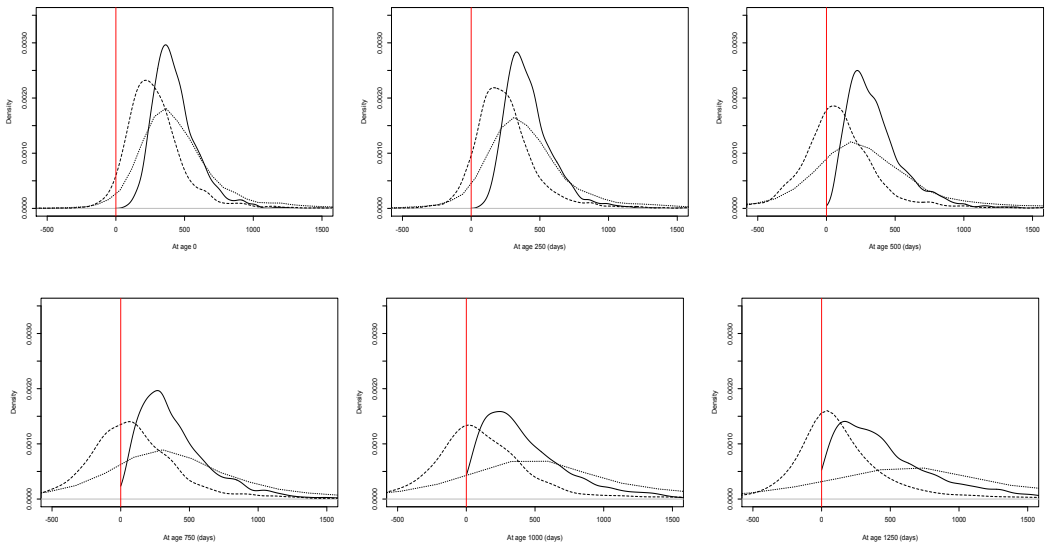

Figure 4.2: Densities of the difference between the mrl functions of Arm A and Arm B under independent gamma DPMMs (dotted), gamma DDPMM (dashed), and the ordered mrl model (solid), are provided at 0 days (top left), 250 days (top middle), 500 days (top right), 750 days (bottom left),1000 days (bottom middle), and 1250 days (bottom right).

In Figure 4.2, we look at the posterior densities of the difference between mrl functions for the two experimental groups at particular time points. For early time points such 0 days and 250 days, the difference appears to be significant across all three models, however, as time goes on, the non-constrained models are not indicating a significant difference between the treatment groups. The variance in the density of

the differences for particularly under the independent gammaDPMMS gets increasingly large as we move to later time periods. The ability to captures dependency groups in the gamma DDPMM seems to help the variance from getting as large as the independent gamma DPMMs. The mrl constraint in the ordered mrl model reduces this variance even further. In fact, across all time points shown in Figure 4.2, the variance in the density of the difference is smaller for the constrained mrl model compared to the variance under the other two models. From these results, we can see the benefits of using the ordered mrl model when assuming the constraint is appropriate.

Chapter 5

Conclusions

The mean residual life function is a valuable functional for multiple fields of research. In the medical field, the mrl function can be used to determine the more effective treatment in two experimental groups. In actuarial studies, insurance companies utilize the mrl function given a set of covariates to help develop different types of policies. The mrl function can also be useful in marketing and setting up warranty policies in selling electronic devices or appliances. Another interesting area in application is in the field of econometrics, where, for example, one could may be interested in model the expected remaining of time of unemployment given that a person is currently unemployed. However, despite the utility of the mrl function, current models in the literature either present restrictions on the shape of the mrl functional, or inference on the mrl function is not even considered. When approached with the mindset of mrl inference, Bayesian nonparametric mixture models for the survival density (or distribution) induce very flexible priors for mrl functions that result in rich posterior functional

inference. Our work provides a comprehensive study of mrl functional inference under the Bayesian nonparametric framework.

Of both methodological and practical interest is development of nonparametric prior models directly for the mrl function. The support of the nonparametric prior would have to be functions that satisfy the properties in the characterization theorem of the mrl function. Under such a prior, inference for the entire distribution would be obtainable by using the inversion formula. Although not a direct prior model for the mrl function, the gamma DP mixture model for the survival density, in Chapter 2, yields desirable results for mrl inference. We defend our choice of kernel by providing key results with respect to the mrl function under various kernels. In particular, we offer a set of criteria based implications of tail behavior for the mrl function, as well as ensuring existence of the posterior mean of the distribution, in kernel selection. Another key result that we present is the denseness of mixtures of gamma distributions on the space of mrl functions. This result implies that under a flexible framework, such as Bayesian nonparametric methods, a mixture of gamma distributions will be able to accurately estimate any form of mrl functions. We demonstrate this capability in both simulated and real data examples.

We extended our flexible inferential results to the regression setting using the curve fitting approach. The curve fitting approach is an attractive framework for the regression setting because by modeling the random covariates jointly with the response we can easily obtain inference the regression functionals. Moreover, we are able to capture nonlinear or nonstandard relationships between the regression functions as we

move across the covariate space. In terms of mrl inference, this has not been achieved under any modeling framework in the literature.

The treatment control setting is common scene in survival analysis. The curve fitting approach is not applicable for fixed covariates such as experimental groups, thus we develop a dependent DP prior that allows use to model the effect of each group in addition to modeling the correlation between the groups. We construct the DDP prior using a bivariate beta distribution that drives the weights of the mixture components for each group. The locations are shared across the experimental groups. This construction is different from most DDP constructions in the literature, where the weights are shared across the groups and locations differ. However, in treatment control settings the range of survival is usually the about the same. Differences in the populations are generally seen in the form of varying prevalence over the range. Thus, the shared locations and differing weights construction for the DDP is more appropriate for the treatment control setting. The DDP prior and the curve fitting regression approach can be combined in a single model to obtain to incorporate both types of covariates, and achieve flexible inferential results for mrl regression functions for two correlated populations.

Contrary to stochastic, hazard rate, and a number of other types of ordering constraints, modeling under mrl order constraint has been comparably neglected. In settings where the researcher believe that the remaining life expectancy of one population is higher than that of another, then this information should surely be incorporated with the model. Particularly in the presence of right censoring and/or a small sample size, the mrl order may not be suggested by the data and uncertainty bands can be

quite large. Using a prior model for mrl ordering can substantially improve inferential accuracy and certainty. In Chapter 4, we present a Bayesian nonparametric model for inference on mrl ordered populations. We use a mixture of Erlang distributions which is a gamma mixture, so our earlier results for mrl inference under gamma mixtures holds for our Erlang mixture model for mrl ordering as well. The Erlang mixture as the same structural idea of the DDP prior model we developed, with having the same kernel locations (although fixed in the mrl ordering case) with weights dependent on the population. The weights in the Erlang mixture are constructed using two independent DP priors. We are able to demonstrate the posterior benefits of our model by comparing results for the same data set under the gamma DPMM and the gamma DDPMM.

Bibliography

A. Antoniadis, G. N., G. Gregoire (1999). Density and hazard rate estimation for right-censored data by using wavelet methods. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **61**, 63–84.

Abdous, B. & Berred, A. (2005). Mean residual life estimation. *Statistical Planning and Inference* **132**, 3–19.

Achcarar, J. A., Brookmeyer, R. & Hunter, W. G. (1985). An application of Bayesian analysis to medical follow-up data. *Statistics in medicine* **4**, 509–520.

Antoniak, C. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The Annals of Statistics* **2**, 1152–1174.

Arajas, E. & Gasbarra, D. (1996). Bayesian inference of survival probabilities, under stochastic ordering constraints. *Journal of the American Statistical Association* **91**.

Arcones, M., Kvam, P. & Samaniego, F. (2002). Nonparametric estimation of a distribution subject to a stochastic precedence constraint. *Journal of the American Statistical Association* **97**, 170–182.

Arjas, E. & Gasbarra, D. (1996). Bayesian inference of survival probabilities, under stochastic ordering constraints. *Journal of the American Statistical Association* **91**, 1101–1109.

Asmussen, S., Frey, A., Rolski, T. & Schmidt, V. (1995). Does markov-modulation increase the risk? *ASTIN Bull* .

Berger, R. L., Boos, D. D. & Guess, F. M. (1988). Tests and confidence sets for comparing two mean residual life functions. *Biometrics* **44**, 103–115.

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

Chen, M., Shao, Q. & Ibrahim, J. (2000). *Monte Carlo Methods in Bayesian Computation*. Statistics. Springer.

Chen, Y. Q. & Cheng, S. (2005). Semiparametric regression analysis of mean residual life with censored data. *Biometrika* **92**, 19–29.

Chen, Z. & Dunson, D. (2004). Bayesian estimation of survival functions under stochastic precedence. *Lifetime Data Analysis* **10**, 159–173.

Conner, R. & Mosemann, J. (1969). Concepts of independence for proportions with a generalization of the Dirichlet distribution. *Journal of the American Statistical Association* **64**, 194–206.

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

Cox, D. & Oakes, D. (1984). *Analysis of Survival Data*. Chapman and Hall.

DeIorio, M., Müller, P., Rosner, G. & MacEachern, S. (2004). An ANOVA model for dependent random measures. *Journal of the American Statistical Association* **99**, 205–215.

DeIorio, M., Müller, P., Rosner, W. & Rosner, G. (2009). Bayesian nonparametric nonproportional hazards survival modelings. *Biometrics* **65**, 762–771.

Dellaportas, P. & Smith, A. (1993). Bayesian inference for generalozed linear and proportional hazards models via Gibbs sampling. *Journal of the Royal Statistical Society Series C (Applied Statistics)* **42**, 443–459.

Dunson, D. & Peddada, S. (2008). Bayesian nonparametric inference on stochastic ordering. *Biometrika* **95**, 859–874.

Dystra, R. L. & Laud, P. (1981). A Bayesian nonparametric approach to reliability. *Annals of Statistics* **9**, 356–367.

Evans, M., Gilula, Z., Guttman, I. & Swartz, T. (1997). Bayesian analysisof stochastically ordered distributions of categorical variables. *Journal of the American Statistical Association* **92**, 208–214.

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

Finkelstein, M. S. (2002). On the shape of the mean residual lifetime function. *Applied Stochastic Models in Business and Industry* **18**, 135–146.

Fronczyk, K. & Kottas, A. (2010). A bayesian nonparametric modeling framework for developmental toxicity studies. Technical report, University of California, Santa Cruz.

Geisser, S. & Eddy, W. (1979). A predictive approach to model selection. *American Statistical Association* **74**, 153–160.

Gelfand, A., Kottas, A. & MacEachern, S. (2005). Bayesian nonparametric spatial modeling with Dirichlet process mixing. *Journal of the American Statistical Association* **100**, 1021–1035.

Gelfand, A. E. & Ghosh, S. K. (1998). Model choice: A minimum posterior predictive loss approach. *Biometrika* **85**.

Gelfand, A. E. & Kottas, A. (2001). Nonparametric Bayesian modeling for stochastic order. *Ann. Inst. Statist. Math.* **53**, 865–876.

Gelfand, A. E. & Kottas, A. (2003). Bayesian semiparametric regression for median residual life. *Scandinavian Journal of Statistics* **30**, 651–665.

Gelfand, A. E. & Mallick, B. K. (1995). Bayesian analysis of proportional hazards models built from monotone functions. *Biometrics* **51**, 843–852.

Govilt, K. & Aggarwal, K. (1983). Mean residual life function of normal, gamma and lognormal densities. *Reliability Engineering* **5**, 47–51.

Guess, F. & Proschan, F. (1985). Mean residual life: theory and applications. Tech-

nical Report 85-178, North Carolina State University and Florida State University, Tallahassee,Florida.

Gupta, R. C. & Akman, H. O. (1995). Mean residual life functions for certain types of non-monotonic ageing. *Commun. Statist.-Stochastic Models* **11**, 219–225.

Gupta, R. C., Akman, O. & Lvin, S. (1999). A study of log-logistic model in survival analysis. *Biometrical Journal* **41**, 431–443.

Hall, W. J. & Wellner, J. A. (1979). Estimation of mean residual life. University of Rochester.

Hall, W. J. & Wellner, J. A. (1981). Mean residual life. In *Statistics and Related Topics*, Eds. M. Csörgö, D. Dawson, J. Rao & A. Saleh. North-Holland Publishing Company.

Hanson, T. & Johnson, W. (2002). Modeling regression error with a mixture of Pólya tress. *Journal of American Statistical Association* **97**, 1020–1033.

Hanson, T., Kottas, A. & Branscum, A. (2008). Modeling stochastic order in the analysis of receiver operating characteristic data: Bayesian nonparametric approaches. *Journal of the Royal Statistical Society Series C Applied Statistics* **57**, 207–225.

Haupt, G. & Mansmann, U. (1995). CART for survival data. *Statlib Archive, http://lib.stat.cmu.edu/S/survcart* **64**, 194–206.

Hendersoni, R. (1995). Problems and prediction in survival data analysis. *Statistics in medicine* **14**, 161–184.

Hjort, N. & Petrone, S. (2007). Nonparametric quantile inference using Dirichlet processes. In *In advances in statistical modeling and inference: essays in honor of Kjell A. Doksum*, Ed. V. Nair, pp. 463–492. World scientific publishing company, Singapore.

Hoff, P. (2003). Bayesian methods for partial stochastic orderings. *Biometrika* **90**, 303–317.

Ibrahim, J., Chen, M. & Sinha, D. (2001). *Bayesian Survival Analysis*. Statistics. Springer.

Iliopoulos, G., Karlis, D. & Ntzoufras, I. (2005). Bayesian estimation in Kibble's bivariate gamma distribution. *The Canadian Journal of Statistics* **33**, 571–589.

Ishwaran, H. & James, L. F. (2001). Gibbs sampling methods for stick-breaking priors. *American Statistical Association* **96**, 161–173.

Johnson, M. A. & Taaffe, M. R. (1998). The denseness of Phase distributions. Research Memorandum No 88-20.

Johnson, W. O. (1999). Survival analysis for interval data. In *Institute for Mathematics and Its Applications: Statistics in the Health Sciences: Diagnosis and Prediction*, Eds. P. Grambsch & S. Geisser, volume 114, pp. 75–90. Springer-Verlag.

Kalbfleisch, J. D. (1978). Nonparametric Bayesian analysis of survival times data. *Journal of the Royal Statistical Society Series B* **40**, 214–221.

Karabatsos, G. & Walker, S. (2007). Bayesian nonparametric inference of stochastically ordered distributions, with Pólya trees of Bernstein polynomials. *Statistics and Probability Letters* **77**, 907–913.

Kim, S. & Ibrahim, J. (2000). On bayesian inference for parametric proportional hazards models using noninformative priors. *Lifetime Data Analysis* **6**, 331–341.

Klein, J. P. & Moeschberger, M. L. (1997). *Survival Analysis: Techniques for Censored and Truncated Data*. Statistics for Biology and Health. Springer.

Kochar, S. C., Mukerjee, H. & Samaniego, F. J. (2000). Estimation of a monotone mean residual life. *The Annals of Statistics* **28**, 905–921.

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

Kottas, A. (2011). Bayesian semiparametric modeling for stochastic precedence, with applications in epidemiology and survival analysis. *Lifetime Data Anal* **17**, 135–155.

Kottas, A., Behseta, S., Moorman, D., Poynor, V. & Olson, C. (2012). Bayesian nonparametric analysis of neuronal intensity rates. *Journal of Neuroscience Methods* **203**, 241–253.

Kottas, A., DeYoreo, M. & Poynor, V. (2013). Bayesian nonparametric inference

- why and how. *Bayesian Analysis* Contributed discussion on paper by P. Ḿuller and R. Mitra.

Kottas, A. & Gelfand, A. E. (2001). Bayesian semiparametric median regression modeling. *Journal of American Statistical Association* **96**, 1458–1468.

Kottas, A. & Krnjajić, M. (2009). Bayesian semiparametric modeling in quantile regression. *Scandinavian Journal of Statistics* **36**, 297–319.

Kuo, L., Smith, A. F., MacEachern, S. & West, M. (1992). *Bayesian computations in survival models via the Gibbs sampler*. Springer Netherlands.

Lahiri, P. & Park, D. H. (1991). Nonparametric Bayes and empirical Bayes estimators of mean residual life. *Journal of Statistical Planning and Inference* **29**, 125–136.

MacEachern, S. N. (2000). Dependent Dirichlet processes. Technical report, Department of Statistics, Ohio State University.

Maguluri, G. & Zhang, C.-H. (1994). Estimation in the mean residual life regression model. *Journal of the Royal Statistical Society Series B* **56**, 477–489.

Marshall, A. & Olkin, I. (2007). *Life Distributions: Structure of nonparametric, semiparametric, and parametric families*. Springer Science and Business Media, LLC.

Mathai, A. M. & Moschopoulos, P. G. (1992). A form of multivariate gamma distribution. *Ann. Inst. Statist. Math.* **44**, 97–106.

McKenzie, E. (1985). An autoregessive process for beta random variables. *Management Science* **31**, 988–997.

Michael, J. R. & Schucany, W. R. (2011). The mixture approach for simulating new families of bivariate distributions with specified correlations. Technical report, Southern Methodist University.

Mudholkar, G. S. & Strivasta, D. K. (1993). Exponentiated Weibull family for analyzing bathtub failure-rate data. *IEEE Transactions of Reliability* **42**, 299–302.

Müller, P., Erkanli, A. & West, M. (1996). Bayesian curve fitting using multivariate normal mixtures. *Biometrika* **83**, 67–79.

Müller, P. & Quintana, F. (2004). Nonparametric Bayesian data analysis. *Statistical Science* **19**, 95–110.

Nadarajah, S. & Gupta, A. (2006). Some bivariate gamma distributions. *Applied Mathematics Letters* **19**, 767–774.

Nadarajah, S. & Kotz, S. (2005). Some bivariate beta distributions. *Journal of Theoretical and Applied Statistics* **39**, 457–466.

Nanda, A., Bhattacharjee, S. & Balakrishnan, N. (2010). Mean residual life function, associated ordering and properties. *Reliability, IEEE Transactions* **59**, 55–56.

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

Oakes, D. & Dasu, T. (1990). A note on mean residual life. *Biometrika* **77**, 409–410.

Olkin, I. & Liu, R. (2003). A bivariate beta distribution. *Statistics and Probability Letters* **62**, 407–412.

Petrone, S. (1999). Random Bernstein polynomials. *Scandinavian Journal of Statistics* **26**, 373–393.

Petrone, S. & Wasserman, L. (2002). Consistency of Bernstein polynomial posteriors. *Journal of the Royal Statistical Society Series B* **64**, 79–100.

Petroni, G. & Wolf, R. (1994). A two-sample test for stochastic ordering with interval censored data. *Biometrics* **50**, 77–87.

Pham, H. & Lai, C.-D. (2007). On recent generalizations of the Weibull distribution. *IEEE Transactions on Reliability* **56**, 454–458.

Poynor, V. (2010). *Bayesian inference for mean residual life functions in survival analysis.* Master's thesis, University of California, Santa Cruz.

Poynor, V. & Kottas, A. (2012). Nonparametric Bayesian inference for mean residual life functions in survival analysis. Draft manuscript in preparation for submission to *Biostatistics.*

Rodriguez, A. *&* ter Horst, E. *(2008). Bayesian dynamic density estimation.* Bayesian Analysis ***3**, 339–366.*

Rojo, J. & Ghebremichael, M. *(2006). Estimation of two ordered bivariate mean residual life functions.* Journal of multivariate analysis ***97****, 431–454.*

Scheike, T. & Jensen, T. *(1997). A discrete survival model with random effects: an application to time to pregnancy.* Biometrics ***53****, 318–329.*

Schmeiser, B. W. & Lal, R. *(1982). Bivariate gamma random vectors.* Operations Research ***30****, 355–374.*

Scurrah, K. J., Palmer, L. J. & Burton, P. R. *(2000). Variance components nalysis for pedigree-based censored survival data using generalized linear mixture models in BUGS.* Genetic Epidemiology ***19****, 127–148.*

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

Shaked, M. & Shanthikumar, J. G. *(2007).* Stochastic Orders*. Springer.*

Singer, J. & Willet, J. *(1993). It's about time: using discrete-time survival analysis to study duration and the timing os events.* Journal of Educational Statistics ***18****, 155–195.*

Sinha, D. & Dey, D. K. *(1997). Semiparametric Bayesian analysis of survival analysis.* Journal of the American Statistical Association ***92****, 1195–1212.*

Smith, P. J. *(2002). Analysis of failure and survival data. In* Texts in Statistical Science Series*, Eds. C. Chatfield, J. Lindsey, M. Tanner & J. Zidek, pp. –. Chapman and Hall/CRC.*

Taddy, M. *(2008).* Bayesian nonparametric analysis of conditional distributions and inference for Poisson point processes. *Ph.D. thesis, University of California, Santa Cruz.*

Taddy, M. & Kottas, A. *(2010). A bayesian nonparametric approach to inference for quantile regression.* Journal of Business and Economic Statistics ***28****, 357–369.*

Walker, S. & Mallick, B. *(1999). A Bayesian semiparametric accelerated failure time model.* Biometrics ***55****, 477–483.*

Willet, J. & Singer, J. *(1993). Investigating onset, cessation, relapse, and recovery: why you should, and how you can, use discrete-time survival analysis to examine event occurence.* Journal of Consulting and Clinical Psychology ***61****, 952–965.*

Xie, M., Goh, T. & Tang, Y. *(2004). On changing points of mean residual life and failure rate function for some generalized Weibull distributions.* Reliability Engineering and System Safety ***84****, 293–299.*

Yang, G. L. *(1978). Estimation of a biometric function.* The Annals of Statistics ***6****, 112–116.*

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

Yue, S., Ouarda, T. & Bobée, B. *(2001). A review of bivariate gamma distributions for hydrological application.* Journal of Hydrology ***246****, 1–18.*

Appendix A

Proof of Properties of MRL

Below we provide the proofs for the properties of the mrl function stated in Section 2.1.

(a) m is a nonnegative and right-continuous, and $m(0) = \mu > 0$:

NON-NEGATIVE: Since $0 \leq F(t) \leq 1 \Rightarrow 0 \leq 1 - S(t) \leq 1 \Rightarrow 0 \leq S(t) \leq 1$. Therefore, $S(t)$ is non-negative. Now consider when $t \geq X$, then $S(t) \equiv 0$, so $m(t) \equiv 0$. For $t < X \Rightarrow S(t) > 0$ thus $\int_t^\infty S(u)du > 0$. Hence $m(t) = \frac{\int_t^\infty S(u)du}{S(t)} \geq 0$.

RIGHT-CONTINUITY: We know that $F(t)$ is right-continuous (ie. $\lim_{h\to 0^+} F(t+h) = F(t)$). Now, $\lim_{h\to 0^+} S(t+h) = \lim_{h\to 0^+} (1 - F(t+h)) = 1 - lim_{h\to 0^+} F(t+h) = 1 - F(t) = S(t)$. Hence $S(t)$ is right-continuous as well. If $S(t)$ is right-continuous, then its integral must also be right-continuous (i.e., the limit, $lim_{h\to 0^+}\left[\int_{t+h}^\infty S(u)du\right] = \int_t^\infty S(u)du$). Finally, $lim_{h\to 0^+} m(t+h) = lim_{h\to 0^+}\left[\frac{\int_{t+h}^\infty S(u)du}{S(t+h)}\right] = \frac{\int_t^\infty S(u)du}{S(t)} = m(t)$, thus $m(t)$ is right-continuous.

FIRST MOMENT STRICTLY POSITIVE: From equation (4) we have established that $\mu = m(0)$. Further, $m(0) = \frac{\int_0^\infty S(u)du}{S(0)} = \int_0^\infty S(u)du$, which must be greater than 0 because $S(u)$ is nonnegative for all $0 \leq u < \infty$ and $S(u+\epsilon) - S(u-\epsilon) > 0$ for at least one value of u and $\epsilon > 0$ in the domain. Therefore, $m(0) \equiv \mu > 0$.

(b) $v(t) \equiv m(t) + t$ is non-decreasing:

Let $h > 0$. **Case 1** $(t + h < X)$: $\Rightarrow v(t+h) - v(t) = m(t+h) + (t+h) - m(t) - t = m(t+h) - m(t) + h = \frac{\int_{t+h}^{\infty} S(u)du}{S(t+h)} - \frac{\int_{t}^{\infty} S(u)du}{S(t)} + h$. Since $S(t)$ is monotone decreasing then $S(t+h) \leq S(t)$ so the former expression is $\geq \frac{\int_{t+h}^{\infty} S(u)du}{S(t)} - \frac{\int_{t}^{\infty} S(u)du}{S(t)} + h = -\frac{\int_{t}^{t+h} S(u)du}{S(t)} + h$ we need to show that this expression is nonnegative. Assume that it is, $\Leftrightarrow h \geq \frac{\int_{t}^{t+h} S(u)du}{S(t)} \Leftrightarrow \int_{t}^{t+h} S(u)du \leq hS(t)$ this is true since the survival function is non-increasing. Hence, $v(t+h) - v(t) \geq 0 \Rightarrow v(t)$ is non-decreasing. **Case 2** $(t < X \leq t + h)$: $\Rightarrow v(t+h) - v(t) \overset{\text{from Case 1}}{=} \frac{\int_{t+h}^{\infty} S(u)du}{S(t+h)} - \frac{\int_{t}^{\infty} S(u)du}{S(t)} + h$, but the first integral is 0 since $t + h > X$. Thus, the expression becomes $-\frac{\int_{t}^{\infty} S(u)du}{S(t)} + h = -\frac{\int_{t}^{X} S(u)du}{S(t)} + h$. Assuming that it is $\Leftrightarrow \int_{t}^{t+h} S(u)du \leq hS(u)$, which is true since the survival function in non-increasing. Therefore, $v(t+h) - v(t) \geq 0 \Rightarrow v(t)$ is non-decreasing. **Case 3** $(X \leq t < t + h)$: $\Rightarrow v(t+h) - v(t) = m(t+h) + (t+h) - m(t) - t$, but since $X \leq t < t + h \Rightarrow m(t+h) = m(t) = 0$. Thus, $v(t+h) - v(t) = h > 0 \Rightarrow v(t)$ is non-decreasing.

(c) $m(t^-) > 0$ for $t \in (0, X)$; if$X < \infty m(X^-) = 0$ and m is continuous at X:

Part 1: Let $t \in (0, X)$, then $m(t^-) = \frac{\int_{t^-}^{X} S(u)du}{S(t^-)}$. Since $S(t^-) < S(X) \leq 1 \Rightarrow \frac{\int_{t^-}^{X} S(u)du}{S(t^-)} > \int_{t^-}^{X} S(u)du$ which is > 0. Therefore, $m(t^-) > 0$.

Part 2: Let $t < X < \infty \Rightarrow v(t) \overset{\text{from (b)}}{\leq} v(X) = m(X) + X = X \Rightarrow v(t) = m(t) + t \leq X \Leftrightarrow m(t) \leq X - t \Rightarrow lim_{t \to X^-} m(t) \leq lim_{t \to X^-}(X - t) = X - X^- = 0 \Rightarrow m(X^-) = m(X) = 0$ proving that $m(t)$ is continuous at X.

(e) $\int_0^t \frac{1}{m(u)} du \to \infty$ as $t \to X$:

Using (e) $\lim_{t \to X} \int_0^t \frac{-k'(u)}{k(u)} du = -\lim_{t \to X} [log(k(t)) - log(k(0))] = -\lim_{t \to X} log\left[\frac{k(t)}{(k(0)}\right] = log\left[\frac{lim_{t \to X} k(t)}{lim_{t \to X} k(0)}\right]$. The limit of the numerator can be found by, $lim_{t \to X} k(t) = lim_{t \to X} S(t) lim_{t \to X} m(t) = 0$, and the denominator is $k(0) = \mu$ which is strictly positive from (a), so the limit inside the log function is 0 with convergence from the right. $\Rightarrow lim_{t \to 0^+ log(t)} = -\infty$, hence $\lim_{t \to X} \int_0^t \frac{1}{m(u)} du = -(-\infty) = \infty$.

Appendix B

Proof of the Lemmas

B.1 Proof of the Lemma 1

Let $f(\cdot)$ and $S(\cdot)$ be the kernel density and survival functions, respectively, of a DPMM. Assume that $f\cdot) > 0$ for all t$\geq$ 0 or for all $t > t_0$ where $t_0 \geq 0$ is some finite value. The corresponding mrl function of the DPMM with L components is given by:

$$m(t|G_L) = \frac{\int_t^\infty \sum_{l=1}^L p_l S(u|\boldsymbol{\theta}_l)du}{\sum_{l=1}^L p_l S(t|\boldsymbol{\theta}_l)} = \frac{\sum_{l=1}^L p_l \int_t^\infty S(u|\boldsymbol{\theta}_l)du}{\sum_{l=1}^L p_l S(t|\boldsymbol{\theta}_l)}$$

Note that $\lim_{t\to\infty} \sum_{l=1}^L p_l \int_t^\infty S(u|\boldsymbol{\theta}_l)du = 0$ and $\lim_{t\to\infty} \sum_{l=1}^L p_l S(t|\boldsymbol{\theta}_l) = 0$, so by L'Hopital's Rule we have:

$$\lim_{t\to\infty} m(t|G_L) = \lim_{t\to\infty} \frac{\frac{d}{dt}\sum_{l=1}^L p_l \int_t^\infty S(u|\boldsymbol{\theta}_l)du}{\frac{d}{dt}\sum_{l=1}^L p_l S(t|\boldsymbol{\theta}_l)} = \lim_{t\to\infty} \frac{-\sum_{l=1}^L p_l S(t|\boldsymbol{\theta}_l)}{-\sum_{l=1}^L p_l f(t|\boldsymbol{\theta}_l)}$$

Once again the limit as t goes to infinity of both the numerator and denominator is zero, so applying L'Hopital once more we have the following:

$$\lim_{t\to\infty} m(t|G_L) = \lim_{t\to\infty} \frac{-\sum_{l=1}^L p_l f(t|\boldsymbol{\theta}_l)}{\sum_{l=1}^L p_l f'(t|\boldsymbol{\theta}_l)}$$

Suppose that the mrl function, $m(t|\boldsymbol{\theta})$ of the kernel distribution tends to infinity as $t \to \infty$. Then, we have:

$$\lim_{t\to\infty} m(t|\boldsymbol{\theta}) = \lim_{t\to\infty} \frac{\int_t^\infty S(u|\boldsymbol{\theta})du}{S(t|\boldsymbol{\theta})} = \lim_{t\to\infty} \frac{-S(t|\boldsymbol{\theta})}{-f(t|\boldsymbol{\theta})} = \lim_{t\to\infty} \frac{-f(t|\boldsymbol{\theta}_l)}{f'(t|\boldsymbol{\theta})} = \infty$$

In other words, $lim_{t\to\infty}(-f(t|\boldsymbol{\theta})/f'(t|\boldsymbol{\theta})) = 0$, so $-f(x|\boldsymbol{\theta})$ grows at a much faster rate the $f'(t|\boldsymbol{\theta})$. Hence $f'(t|\boldsymbol{\theta})$ is " little-o " of $-f(t|\boldsymbol{\theta})$: $f'(t|\boldsymbol{\theta}) \in o(-f(t|\boldsymbol{\theta}))$. Returning now to the mrl function of the DPMM, for each component $l = 1, ..., L$: $f'(t|\boldsymbol{\theta}_l) \in o(-f(t|\boldsymbol{\theta}_l)) \Leftrightarrow p_l f'(t|\boldsymbol{\theta}_l) \in o(-p_l f(t|\boldsymbol{\theta}_l))$ since by definition of "little - o" $\forall \epsilon > 0$ there exists $t_{0l} \in \mathbb{R}$ such that $|f'(t|\boldsymbol{\theta}_l)| \leq \epsilon| - f(t|\boldsymbol{\theta}_l)| \forall t \geq t_{0l}$ for $l = 1, ...L$. Multiplying by $p_l > 0$ on either side gives $p_l|f'(t|\boldsymbol{\theta}_l)| \leq \epsilon p_l| - f(t|\boldsymbol{\theta}_l)| \Leftrightarrow |p_l f'(t|\boldsymbol{\theta}_l)| \leq \epsilon| - p_l f(t|\boldsymbol{\theta}_l)|$. Since this last inequality holds for all $l = 1, ..., L$ we can apply the sum over l on both sides obtaining $\sum_{l=1}^L |p_l f'(t|\boldsymbol{\theta}_l)| \leq \sum_{l=1}^L \epsilon| - p_l f(t|\boldsymbol{\theta}_l)|$. We can bound the left side of the inequality below using the triangle inequality, $|\sum_{l=1}^L p_l f'(t|\boldsymbol{\theta}_l)| \leq \sum_{l=1}^L |p_l f'(t|\boldsymbol{\theta}_l)|$. Meanwhile, the right side of the inequality can be written as $\sum_{l=1}^L \epsilon| - p_l f(t|\boldsymbol{\theta}_l)| = \epsilon| - \sum_{l=1}^L p_l f(t|\boldsymbol{\theta}_l)|$ since $-p_l f(t|\boldsymbol{\theta}_l) \leq 0$ for each $l = 1, ..., L$. Thus, we can make the following statement: $\forall \epsilon > 0 \exists t_0 \in \mathbb{R}$ such that $|\sum_{l=1}^L p_l f'(t|\boldsymbol{\theta}_l)| \leq \epsilon| - \sum_{l=1}^L p_l f(t|\boldsymbol{\theta}_l)| \forall t \geq max\{t_{01}, t_{02}, ..., t_{0L}\}$. In other words, $\sum_{l=1}^L p_l f'(t|\boldsymbol{\theta}_l) \in o(-\sum_{l=1}^L p_l f(t|\boldsymbol{\theta}_l))$, and therefore,

$$\lim_{t\to\infty} m(x|G_L) = \lim_{t\to\infty} \frac{-\sum_{l=1}^L p_l f(t|\boldsymbol{\theta}_l)}{\sum_{l=1}^L p_l f'(t|\boldsymbol{\theta}_l)} = \infty$$

Now suppose that the mrl function of the kernel distribution tends to zero as $t \to \infty$. Hence, we can say that $S(t|\boldsymbol{\theta}) \in o(f(t|\boldsymbol{\theta}))$, since we have the following:

$$\lim_{t\to\infty} m(t|G_L) = \lim_{t\to\infty} \frac{S(t|\boldsymbol{\theta})}{f(t|\boldsymbol{\theta})} = 0$$

Thus, in the DPMM we have for each component, $l = 1, ..., L$, $S(t|\boldsymbol{\theta}_l) \in o(f(t|\boldsymbol{\theta}_l)) \Leftrightarrow p_l S(t|\boldsymbol{\theta}_l) \in o(p_l f(t|\boldsymbol{\theta}_l))$. From the definition $\forall \epsilon > 0 \exists t_{0l} \in \mathbb{R}$ such that $|p_l S(t|\boldsymbol{\theta}_l)| \leq \epsilon |p_l f(t|\boldsymbol{\theta}_l))| \forall t \geq t_{0l}$. Applying the sum over the components gives us $\sum_{l=1}^{L} |p_l S(t|\boldsymbol{\theta}_l)| \leq \sum_{l=1}^{L} \epsilon |p_l f(t|\boldsymbol{\theta}_l))|$. The left side of the inequality can be written as $\sum_{l=1}^{L} |p_l S(t|\boldsymbol{\theta}_l)| = |\sum_{l=1}^{L} p_l S(t|\boldsymbol{\theta}_l)|$, similarly, the right side can be written as $\sum_{l=1}^{L} \epsilon |p_l f(t|\boldsymbol{\theta}_l))| = \epsilon |\sum_{l=1}^{L} p_l f(t|\boldsymbol{\theta}_l))|$. Hence, $\forall \epsilon > 0 \exists t_0 \in \mathbb{R}$ such that $|\sum_{l=1}^{L} p_l S(t|\boldsymbol{\theta}_l)| \leq \epsilon |\sum_{l=1}^{L} p_l f(t|\boldsymbol{\theta}_l))| \forall t \geq max\{t_{01}, t_{02}, ..., t_{0L}\}$. In other words, $\sum_{l=1}^{L} p_l S(t|\boldsymbol{\theta}_l) \in o(\sum_{l=1}^{L} p_l f(t|\boldsymbol{\theta}_l))$, and therefore,

$$\lim_{t\to\infty} m(t|G_L) = lim_{t\to\infty} \frac{\sum_{l=1}^{L} p_l S(t|\boldsymbol{\theta}_l)}{\sum_{l=1}^{L} p_l f(t|\boldsymbol{\theta}_l)} = 0$$

B.2 Proof of the Lemma 2

Let $\mathscr{F}$ be the space of absolutely continuous distribution functions on $\mathbb{R}^+$ with finite mean, $\mu < \infty$. Let $\mathscr{M}$ be the space of continuous mrl functions. Consider the class of gamma mixture distributions, $\mathcal{C}$. Now, let $m(t)$, for $t \geq 0$, be any mrl function in $\mathscr{M}$. We can obtain the survival function corresponding to $m(t)$ via the Inversion Formula:

$$S(t) = \frac{m(0)}{m(t)} exp\left(-\int_0^t \frac{1}{m(u)} du\right)$$

Hence the corresponding distribution function is defined by $F(t) = 1 - S(t)$. Now, we know that $\mathcal{C}$ is dense in $\mathscr{F}$. Particularly, if we define a sequence of distribution functions, $\{F_n(t)\} \subseteq \mathcal{C}$, as follows:

$$F_n(t) = \sum_{l=1}^{\infty} \left[F\left(\frac{l}{n}\right) - F\left(\frac{l-1}{n}\right)\right] F_\Gamma(t|l, n)$$

where $\left[F\left(\frac{l}{n}\right) - F\left(\frac{l-1}{n}\right)\right]$ are the corresponding weights of the gamma cumulative distributions functions, $F_\Gamma(t|l,n)$, with shape parameter l and rate parameter n.. Johnson & Taaffe (1998) show that for any $t_0 \geq 0$, $\lim_{n\to\infty} F_n(t_0) = F(t_0)$. That is the sequence $\{F_n(t)\}$ converges weakly (or pointwise) to $F(t)$. For the case of a finite mixture, the sequence is defined such that the limit of the sequence as the number of mixture components tends to infinity is also taken. Note that since $\{F_n(t)\}$ converges weakly (or pointwise) to $F(t)$, then the sequence of survival functions, $\{S_n(t)\}$ converges pointwise to $S(t)$, since $\lim_{n\to\infty} S_n(t) = \lim_{n\to\infty}(1-F_n(t)) = 1-\lim_{n\to\infty} F_n(t) = 1-F(t) = S(t)$.

Define the sequence of mrl functions, $\{m_n(t)\}$, through the sequence of survival function, $\{S_n(t)\}$ by the following,

$$m_n(t) = \frac{\int_t^\infty S_n(u)du}{S_n(t)}$$

Consider any $t_0 \geq 0$, then take the limit of the sequence,

$$\lim_{n\to\infty} m_n(t_0) \quad = \quad \lim_{n\to\infty} \frac{\int_{t_0}^\infty S_n(u)du}{S_n(t_0)} = \frac{\lim_{n\to\infty}\int_{t_0}^\infty S_n(u)du}{\lim_{n\to\infty} S_n(t_0)}$$

The limit can be distributed in the last step as a basic property of limits provided the limits exist and the limit of the denominator is not zero. Upon evaluating these limits, we will show all these requirements are met. The bottom limit is trivial since $\{S_n(t)\}$ converges pointwise to $S(t)$ which is bounded by $0 < S(t) \leq 1$. The nontrivial step is being able to move the limit inside the integral in the numerator. We can rewrite $\int_t^\infty S_n(u)du$ as $\mu_n - \int_0^t S_n(u)du$, where μ_n is the mean of the nth distribution in the sequence.

$$\Rightarrow \lim_{n\to\infty} m_n(t_0) = \frac{\lim_{n\to\infty}\left[\mu_n - \int_0^{t_0} S_n(u)du\right]}{\lim_{n\to\infty} S_n(t_0)} = \frac{\lim_{n\to\infty}\mu_n - \lim_{n\to\infty}\int_0^{t_0} S_n(u)du}{\lim_{n\to\infty} S_n(t_0)}$$

Now, we will establish that μ_n is finite for every n, and that $\lim_{n\to\infty}\mu_n = \mu$:

$$\begin{aligned}
\mu_n &= \int_0^\infty S_n(u)du = \int_0^\infty 1 - F_n(u)du = \int_0^\infty 1 - \sum_{l=1}^{\infty}\left[F\left(\frac{l}{n}\right) - F\left(\frac{l-1}{n}\right)\right]F_\Gamma(u|l,n)du \\
&= \int_0^\infty \sum_{l=1}^{\infty}\left[F\left(\frac{l}{n}\right) - F\left(\frac{l-1}{n}\right)\right] - \sum_{l=1}^{\infty}\left[F\left(\frac{l}{n}\right) - F\left(\frac{l-1}{n}\right)\right]F_\Gamma(u|l,n)du \\
&= \int_0^\infty \sum_{l=1}^{\infty}\left[F\left(\frac{l}{n}\right) - F\left(\frac{l-1}{n}\right)\right](1 - F_\Gamma(u|l,n))du \\
&= \int_0^\infty \sum_{l=1}^{\infty}\left[F\left(\frac{l}{n}\right) - F\left(\frac{l-1}{n}\right)\right]S_\Gamma(u|l,n)du
\end{aligned}$$

where $S_\Gamma(u|l,n)$ is the survival function of the Gamma distribution with shape l and rate n. Now, since $\left[F\left(\frac{l}{n}\right) - F\left(\frac{l-1}{n}\right)\right]S_\Gamma(u|l,n) \geq 0$, by Tonelli's Theorem, we can exchange the summation and integral:

$$\begin{aligned}
\Rightarrow \mu_n &= \sum_{l=1}^{\infty}\left[F\left(\frac{l}{n}\right) - F\left(\frac{l-1}{n}\right)\right]\int_0^\infty S_\Gamma(u|l,n)du = \sum_{l=1}^{\infty}\left[F\left(\frac{l}{n}\right) - F\left(\frac{l-1}{n}\right)\right]\left(\frac{l}{n}\right) \\
&= \sum_{l=1}^{\infty}\left[\int_{(l-1)/n}^{l/n} f(u)\,du\right]\left(\frac{l}{n}\right) = \sum_{l=1}^{\infty}\int_{(l-1)/n}^{l/n}\left(\frac{l}{n}\right)f(u)\,du \\
&\leq \sum_{l=1}^{\infty}\int_{(l-1)/n}^{l/n}\left(u + \frac{1}{n}\right)f(u)\,du = \sum_{l=1}^{\infty}\int_{(l-1)/n}^{l/n}(u)\,f(u)\,du + \sum_{l=1}^{\infty}\int_{(l-1)/n}^{l/n}\left(\frac{1}{n}\right)f(u)\,du \\
&= \int_0^\infty uf(u)du + \frac{1}{n}\int_0^\infty f(u)du = \mu + \frac{1}{n} < \infty
\end{aligned}$$

where the last inequality holds since we are assuming F has finite mean. The inequality, $\mu_n \leq \mu + (1/n)$, also provides the following upper bound for the limit:

$$\Rightarrow \lim_{n\to\infty}\mu_n \leq \lim_{n\to\infty}\left(\mu + \frac{1}{n}\right) = \mu$$

We can also establish the following lower bound for the limit,

$$\lim_{n\to\infty}\mu_n = \lim_{n\to\infty}\sum_{l=1}^{\infty}\int_{(l-1)/n}^{l/n}\left(\frac{l}{n}\right)f(u)du \geq \lim_{n\to\infty}\sum_{l=1}^{\infty}\int_{(l-1)/n}^{l/n}uf(u)du = \lim_{n\to\infty}\int_0^\infty uf(u)du = \mu$$

Therefore, by Squeeze Theorem, $\lim_{n\to\infty}\mu_n = \mu$.

Using the fact that $S_n(t) \leq 1$ and $S_n(t)$ converges pointwise as $n \to \infty$ to $S(t)$, by the dominated convergence theorem,

$$\lim_{n\to\infty} \int_0^{t_0} S_n(u)du = \int_0^{t_0} \lim_{n\to\infty} S_n(u)du = \int_0^{t_0} S(u)du$$

Returning to the limit of the sequence of mrl functions,

$$\begin{aligned} \Rightarrow \lim_{n\to\infty} m_n(t_0) &= \frac{\lim_{n\to\infty} \mu_n - \lim_{n\to\infty} \int_0^{t_0} S_n(u)du}{\lim_{n\to\infty} S_n(t_0)} = \frac{\mu - \int_0^{t_0} S(u)du}{S(t_0)} \\ &= \frac{\int_{t_0}^{\infty} S(u)du}{S(t_0)} = m(t_0) \end{aligned}$$

Hence, $\{m_n(t)\}$ convergence pointwise to $m(t)$, providing the denseness result for continuous mrl functions under gamma mixture distributions.

Appendix C

Posterior sampling from the gamma DPMM

As we stated in the text, posterior samples of the unknown parameters can easily be obtained using the block Gibbs sampler for DP mixtures described in Ishwaran & James (2001). Recall that our full hierarchical model is given by,

$$\begin{aligned} t_i|\boldsymbol{\theta}, \mathrm{w}_i &\overset{ind}{\sim} \Gamma(t_i|e^{\theta_{\mathrm{w}_i}}, e^{\phi_{\mathrm{w}_i}}) \\ \mathrm{w}_i|\mathbf{p} &\overset{iid}{\sim} \sum_{l=1}^{L} p_l \delta_l(\mathrm{w}_i) \\ \mathbf{p}|\alpha &\sim f(\mathbf{p}|\alpha) \quad (SB) \\ \boldsymbol{\theta}_l = (\theta_l, \phi_l)'|\boldsymbol{\mu}, \boldsymbol{\Sigma} &\overset{iid}{\sim} N_2((\theta_l, \phi_l)'|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \end{aligned}$$

with priors: $\alpha \sim \Gamma(\alpha|a_\alpha, b_\alpha(\text{rate}))$, $\boldsymbol{\mu} \sim N_2(\boldsymbol{\mu}|a_\mu, B_\mu)$, and $\boldsymbol{\Sigma} \sim IWish(\boldsymbol{\Sigma}|a_\Sigma, B_\Sigma)$, where $f(\boldsymbol{p}|\alpha) = \alpha^{L-1} p_L^{\alpha-1}(1-p_1)^{-1}(1-(p_1+p_2))^{-1} \times ... \times (1-\sum_{l=1}^{L-2} p_l)^{-1}$ is a special case of the generalized Dirichlet distribution as is Connor and Mosimann Con-

ner & Mosemann (1969). Let n^* be the number of distinct components of $\mathbf{w}$ where $\mathbf{w}^* = \{\mathrm{w}_j^* : j = 1, ..., n^*\}$ are the distinct components. Let Ψ represent the vector of the most recent iteration of all other parameters. For $i = 1, ..., n$, let $\delta_i = 0$ if t_i is observed and $\delta_i = 1$ if t_i is right censored. Finally, let $b = 1, ..., B$ be the number of iterations in the MCMC. Then B samples from the joint posterior distribution, $p(\boldsymbol{\mu}, \boldsymbol{\sigma}^2, \mathbf{w}, \boldsymbol{p}, \lambda, \tau^2, \rho, \alpha)|data)$ are obtained for $b = 1, ..., B + 1$:

Sample from the posterior conditional distribution for $\boldsymbol{\theta}_l$ for $l = 1, ..., L$: If l is not already a component: $l \notin \{\mathrm{w}_j^{*(b)} : j = 1, ..., n^{*(b)}\}$

$$p(\theta_l^{(b+1)}, \phi_l^{(b+1)}|data, \Psi) \overset{\text{draw}}{\sim} N_2(\boldsymbol{\mu}^{(b)}, \boldsymbol{\Sigma}^{(b)})$$

If l is an active component: $l \in \{\mathrm{w}_j^{*(b)} : j = 1, ..., n^{*(b)}\}$

$$p(\theta_l, \phi_l|data, \Psi) \propto N_2((\theta_l, \phi_l)'|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \prod_{\{i:l=\mathrm{w}_i\}} \left[\Gamma(t_i|e^{\theta_l}, e^{\phi_l})\right]^{1-\delta_i} \left[\int_{t_i}^{\infty} \Gamma(u_i|e^{\theta_l}, e^{\phi_l}) du_i\right]^{\delta_i}$$

We use a Metropolis-Hastings step for this update. We sample from the proposal distribution $(\theta_l', \phi_l')' \sim N_2((\theta_l^{(b)}, \phi_l^{(b)})', cS^2)$, where S^2 is updated from the average posterior samples of $\boldsymbol{\Sigma}$ under initial runs, and $c > 1$. Draw $\eta \sim Unif(0, 1)$.

If $\eta < min \left\{ 1, \dfrac{N_2((\theta_l', \phi_l')'|\boldsymbol{\mu}^{(b)}, \boldsymbol{\Sigma}^{(b)}) \prod_{\{i:l=\mathrm{w}_i^{(b)}\}} \left[\Gamma(t_i|e^{\theta_l'}, e^{\phi_l'})\right]^{1-\delta_i} \left[\int_{t_i}^{\infty} \Gamma(u_i|e^{\theta_l'}, e^{\phi_l'}) du_i\right]^{\delta_i}}{N_2((\theta_l^{(b)}, \phi_l^{(b)})'|\boldsymbol{\mu}^{(b)}, \boldsymbol{\Sigma}^{(b)}) \prod_{\{i:l=\mathrm{w}_i^{(b)}\}} \left[\Gamma(t_i|e^{\theta_l^{(b)}}, e^{\phi_l^{(b)}})\right]^{1-\delta_i} \left[\int_{x_i}^{\infty} \Gamma(u_i|e^{\theta_l^{(b)}}, e^{\phi_l^{(b)}}) du_i\right]^{\delta_i}} \right\}$

set $(\theta_l^{(b+1)}, \phi_l^{(b+1)})' = (\theta_l', \phi_l')'$

else $(\theta_l^{(b+1)}, \phi_l^{(b+1)})' = (\theta_l^{(b)}, \phi_l^{(b)})'$.

Turning to the update for $\boldsymbol{p}$, we have:

$$p(\boldsymbol{p}|data, \Psi) \propto f(\boldsymbol{p}|\alpha) \prod_{l=1}^{L} p_l^{M_l} \;\; M_l = |\{i : \mathrm{w}_i = l\}| \;, l = 1, ..., L$$

$$p(\boldsymbol{p}^{(b+1)}|data,\Psi) \overset{\text{draw}}{\sim} \text{Generalized Dirichlet Distribution}$$

To sample from this distribution, for $l = 1, ..., L$ draw latent variable: $V_l^{*(b+1)} \overset{ind}{\sim} Beta(1+M_l^{(b)}, \alpha^{(b)}+\sum_{r=l+1}^{L} M_r^{(b)})$. Now set $p_1^{(b+1)} = V_1^{*(b+1)}, p_l^{(b+1)} = V_l^{*(b+1)} \prod_{r=1}^{l-1}(1-V_r^{*(b+1)})$ $(l = 2, ..., L-1)$, and $p_L^{(b+1)} = 1 - \sum_{l=1}^{L-1} p_l^{(b+1)}$.

Now we update w_i for $i = 1, .., n$:

$$p(\mathrm{w}_i^{(b+1)}|data,\Psi) \overset{\text{draw}}{\sim} \sum_{l=1}^{L} \tilde{p}_{li}\delta_{(l)}(\cdot)$$

where $\tilde{p}_{li} = \dfrac{p_l^{(b+1)}\left[\Gamma(t_i|e^{\theta_l^{(b+1)}}, e^{\phi_l^{(b+1)}})\right]^{1-\delta_i}\left[\int_{t_i}^{\infty}\Gamma(u_i|e^{\theta_l^{(b+1)}}, e^{\phi_l^{(b+1)}})du_i\right]^{\delta_i}}{\sum_{l=1}^{L} p_l^{(b+1)}\left[\Gamma(t_i|e^{\theta_l^{(b+1)}}, e^{\phi_l^{(b+1)}})\right]^{1-\delta_i}\left[\int_{t_i}^{\infty}\Gamma(u_i|e^{\theta_l^{(b+1)}}, e^{\phi_l^{(b+1)}})du_i\right]^{\delta_i}}$, $l = 1, ..., L$.

For the update for $\boldsymbol{\mu}$ we have:

$$p(\boldsymbol{\mu}|data,\Psi) \propto \prod_{\{l\in\mathbf{w}^*\}} N_2((\theta_l,\phi_l)'|\boldsymbol{\mu},\boldsymbol{\Sigma})N_2(\boldsymbol{\mu}|a_\mu, B_\mu)$$

$$p(\boldsymbol{\mu}^{(b+1)}|data,\Psi) \overset{\text{draw}}{\sim} N_2(m_\mu, S_\mu^2)$$

where $m_\mu = S_\mu^2\left(B_\mu^{-1}a_\mu + \boldsymbol{\Sigma}^{-1}\sum_{\{l\in\mathbf{w}^{*(b+1)}\}}\boldsymbol{\theta}_l^{(b+1)}\right)$, $S_\mu^2 = \left(B_\mu^{-1} + n^{*(b+1)}\boldsymbol{\Sigma}^{-1(b)}\right)^{-1}$.

Turning to the update of $\boldsymbol{\Sigma}$, we have:

$$p(\boldsymbol{\Sigma}|data,\Psi) \propto \prod_{\{l\in\mathbf{w}^*\}} N_2((\theta_l,\phi_l)'|\boldsymbol{\mu},\boldsymbol{\Sigma})IWish(\boldsymbol{\Sigma}|a_\Sigma, B_\Sigma)$$

$$p(\boldsymbol{\Sigma}^{(b+1)}|data,\Psi) \overset{\text{draw}}{\sim}$$

$$IWish(n^{*(b+1)} + a_\Sigma, B_\Sigma + \sum_{\{l\in\mathbf{w}^{*(b+1)}\}}(\boldsymbol{\theta}_l^{(b+1)} - \boldsymbol{\mu}^{(b+1)})(\boldsymbol{\theta}_l^{(b+1)} - \boldsymbol{\mu}^{(b+1)})')$$

Lastly, the update for α is given by:

$$p(\alpha|data,\Psi) \propto \Gamma(\alpha|a_\alpha, b_\alpha)f(\mathbf{p}|\alpha)$$

$$p(\alpha^{(b+1)}|data,\Psi) \overset{\text{draw}}{\sim} \Gamma\left(L + a_\alpha - 1, -\sum_{s=1}^{L-1} log(1 - V_s^{*(b+1)}) + b_\alpha\right)$$

Appendix D

Posterior sampling and Conditional Predictive Ordinate for gamma DDPMM

D.1 Posterior sampling from the gamma DDPMM with random covariates

Here we show the algorithm used to obtain the posterior samples of the parameters in a gamma DDPMM is the presence of a single random continuous real-valued covariate. We use the blocked Gibbs sampler described in Ishwaran & James (2001), with Metropolis-Hastings steps when conjugacy is not obtainable. The full hierarchical version of the model is written as,

$$
\begin{aligned}
(t_{is}, x_{is})|\mathrm{w}_{is}, \boldsymbol{\theta}_l &\overset{ind}{\sim} \Gamma(t_{is}|e^{\theta_{\mathrm{w}_{is}}}, e^{\phi_{\mathrm{w}_{is}}}) N(x_{is}|\beta_{\mathrm{w}_{is}}, \kappa^2_{\mathrm{w}_{is}}), \\
&\qquad \text{for } i = 1, ..., n_s \text{ and } s \in \{C, T\} \\
\mathrm{w}_{is}|\{(\zeta_{ls})\} &\overset{ind}{\sim} \sum_{l=1}^{L} \{(1-\zeta_{ls}) \prod_{r=1}^{l-1} \zeta_{rs}\} \delta_l(\mathrm{w}_{is}), \quad \text{for } i = 1, ..., n_s \text{ and } s \in \{C, T\} \\
\{(\zeta_{lC}, \zeta_{lT})\}|\alpha, b &\sim Biv - Beta(\{(\zeta_{lC}, \zeta_{lT})\}|\alpha, b) \\
&\qquad \zeta_{lC} = UW, \quad \zeta_{lT} = VW, \text{ for } l = 1, ..., L \\
&\qquad U \overset{iid}{\sim} Beta(\alpha, 1-b), \; V \overset{iid}{\sim} Beta(\alpha, 1-b), \\
&\qquad W \overset{iid}{\sim} Beta(1+\alpha-b, b) \\
\alpha &\sim \Gamma(\alpha|a_\alpha, b_\alpha) \\
b &\sim Unif(b|0, 1) \\
(\theta_l, \phi_l)'|\boldsymbol{\mu}, \boldsymbol{\Sigma} &\overset{iid}{\sim} N_2((\theta_l, \phi_l)'|\boldsymbol{\mu}, \boldsymbol{\Sigma}), \quad \text{for } l = 1, ..., L \\
\boldsymbol{\mu} &\sim N_2(\boldsymbol{\mu}|a_\mu, B_\mu) \\
\boldsymbol{\Sigma} &\sim IWish(\boldsymbol{\Sigma}|a_\Sigma, B_\Sigma) \\
\beta_l|\lambda, \tau^2 &\overset{iid}{\sim} N(\beta_l|\lambda, \tau^2), \text{ for } l = 1, ..., L \\
\kappa^2_l|a, \rho &\overset{iid}{\sim} \Gamma^{-1}(\kappa^2|a, \rho), \text{ for } l = 1, ..., L \\
\lambda &\sim N(\lambda|a_\lambda, b^2_\lambda) \\
\tau^2 &\sim \Gamma^{-1}(\tau^2|a_\tau, b_\tau) \\
\rho &\sim \Gamma(\rho, a_\rho, b_\rho)
\end{aligned}
$$

Let L^*_s be the number of distinct components, and $\mathbf{w}^*_s \equiv \{\mathrm{w}^*_{js} : j = 1, ..., L^*_s\}$ be the vector of distinct components for group $s \in \{C, T\}$. For $i = 1, ..., n_s$, let $\delta_{is} = 0$ if t_{is}

is observed and $\delta_{is} = 1$ if t_{is} is right censored for $s \in \{C, T\}$. Let Ψ represent the vector of the most recent iteration of all other parameters. Let $b = 1, ..., B$ be the number of iterations in the MCMC. The posterior samples of $p(\boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{\beta}, \boldsymbol{\kappa}^2, \mathbf{w}, \boldsymbol{\zeta}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \lambda, \tau^2, \rho, \alpha, b|data)$ can be obtained by the following algorithm:

Sample from the posterior conditional distribution for $(\theta_l, \phi_l)'$,β_l, and κ_l^2 for $l = 1, ..., L$:

If l is not already a component: $l \notin \mathbf{w}_C^{*(b)} \cup \mathbf{w}_T^{*(b)}$

$$p(\theta_l^{(b+1)}, \phi_l^{(b+1)}|data, \Psi) \overset{\text{draw}}{\sim} N_2(\boldsymbol{\mu}^{(b)}, \boldsymbol{\Sigma}^{(b)})$$

$$p(\beta_l^{(b+1)}|data, \Psi) \overset{\text{draw}}{\sim} N(\lambda^{(b)}, \kappa_l^{2(b)})$$

$$p(\kappa_l^{2(b+1)}|data, \Psi) \overset{\text{draw}}{\sim} \Gamma^{-1}(a, \rho^{(b)})$$

If l is an active component in either or both: $l \in \mathbf{w}_C^{*(b)} \cup l \in \mathbf{w}_T^{*(b)}$

$$p(\theta_l, \phi_l|data, \Psi) \propto$$

$$N_2((\theta_l, \phi_l)'|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \prod_{s\in\{C,T\}} \prod_{\{i:l=\mathrm{w}_{is}\}} \left[\Gamma(t_{is}|e^{\theta_l}, e^{\phi_l})\right]^{1-\delta_{is}} \left[\int_{t_{is}}^{\infty} \Gamma(u_i|e^{\theta_l}, e^{\phi_l}) dt_i\right]^{\delta_{is}}$$

We use a Metropolis-Hastings step for this update. We sample from the proposal distribution $(\theta_l', \phi_l')' \sim N_2((\theta_l^{(b)}, \phi_l^{(b)})', cS^2)$, where S^2 is updated from the average posterior samples of $\boldsymbol{\Sigma}$ under initial runs, and $c > 1$. Draw $\eta \sim Unif(0, 1)$.

If $\eta < min\left\{1, \right.$

$$\left. \frac{N_2((\theta_l', \phi_l')'|\boldsymbol{\mu}^{(b)}, \boldsymbol{\Sigma}^{(b)}) \prod_{s\in\{C,T\}} \prod_{\{i:l=\mathrm{w}_{is}^{(b)}\}} \left[\Gamma(t_{is}|e^{\theta_l'}, e^{\phi_l'})\right]^{1-\delta_{is}} \left[\int_{t_{is}}^{\infty} \Gamma(u_{is}|e^{\theta_l'}, e^{\phi_l'}) dt_i\right]^{\delta_{is}}}{N_2((\theta_l^{(b)}, \phi_l^{(b)})'|\boldsymbol{\mu}^{(b)}, \boldsymbol{\Sigma}^{(b)}) \prod_{s\in\{C,T\}} \prod_{\{i:l=w_{is}^{(b)}\}} \left[\Gamma(t_{is}|e^{\theta_l^{(b)}}, e^{\phi_l^{(b)}})\right]^{1-\delta_{is}} \left[\int_{t_{is}}^{\infty} \Gamma(u_{is}|e^{\theta_l^{(b)}}, e^{\phi_l^{(b)}}) dt_i\right]^{\delta_{is}}} \right\}$$

set $(\theta_l^{(b+1)}, \phi_l^{(b+1)})' = (\theta_l', \phi_l')'$

else $(\theta_l^{(b+1)}, \phi_l^{(b+1)})' = (\theta_l^{(b)}, \phi_l^{(b)})'$.

$$p(\beta_l|data,\Psi) \propto N(\beta_l|\lambda,\tau^2)\prod_{s\in\{C,T\}}\prod_{\{i:l=\mathrm{w}_{is}\}} N(x_{is}|\beta_l,\kappa_l^2)$$

$$p(\beta_l^{(b+1)}|data,\Psi) \overset{\text{draw}}{\sim} N(m_\beta, s_\beta^2)$$

where $m_\beta = s_\beta^2\left(\kappa_l^{-2(b)}\left[\sum_{s\in\{C,T\}}\sum_{\{i:l=\mathrm{w}_{is}\}} x_{is}\right] + \tau^{-2(b)}\lambda^{(b)}\right)$,
and $s_\beta^2 = \left(\tau^{-2(b)} + \kappa_l^{-2(b)}\left[\sum_{s\in\{C,T\}}\sum_{\{i:l=\mathrm{w}_{is}\}} 1\right]\right)^{-1}$.

$$p(\kappa_l^2|data,\Psi) \propto \Gamma^{-1}(\kappa_l^2|a,\rho)\prod_{s\in\{C,T\}}\prod_{\{i:l=\mathrm{w}_{is}\}} N(x_{is}|\beta_l,\kappa_l^2)$$

$$p(\kappa_l^{2(b+1)}|data,\Psi) \overset{\text{draw}}{\sim}$$

$$\Gamma^{-1}\left(0.5\left[\sum_{s\in\{C,T\}}\sum_{\{i:l=\mathrm{w}_{is}\}} 1\right] + a, 0.5\left[\sum_{s\in\{C,T\}}\sum_{\{i:l=\mathrm{w}_{is}\}}(x_{is}-\beta_l^{(b+1)})^2\right] + \rho^{(b)}\right)$$

To obtain samples from $p(\boldsymbol{\zeta}|\Psi.data)$ we work with the latent variables $\{U_l, V_l, W_l\}$. The posterior $p(\{(U_l,V_l,W_l)\}|\Psi,data)$ is proportional to,

$$p(\{(U_l,V_l,W_l)\}|\Psi,data) \propto$$
$$\prod_{l=1}^{L-1} U_l^{(\sum_{r=l+1}^{L} M_{rC})+\alpha-1}(1-U_l)^{-b}(1-U_lW_l)^{M_{lC}} V_l^{(\sum_{r=l+1}^{L} M_{rT})+\alpha-1}(1-V_l)^{-b}(1-$$
$$V_lW_l)^{M_{lT}} W_l^{(\sum_{r=l+1}^{L} M_{rC}+M_{rT})+\alpha-b}(1-W_l)^{b-1}$$

Using slice sampling, we can introduce latent variables ν_l and γ_l for $l = 1,...,L$, such that we have Gibbs steps for for each parameter. The joint posterior of interest becomes:

$$p(\{(U_l,V_l,W_l,\nu_l,\gamma_l)\}|\Psi,data) \propto$$
$$\prod_{l=1}^{L-1} U_l^{(\sum_{r=l+1}^{L} M_{rC})+\alpha-1}(1-U_l)^{-b}\mathbf{1}_{\left(0<\nu_l\leq(1-U_lW_l)^{M_{lC}}\right)} V_l^{(\sum_{r=l+1}^{L} M_{rT})+\alpha-1}(1-V_l)^{-b}$$
$$\times\mathbf{1}_{\left(0<\gamma_l\leq(1-V_lW_l)^{M_{lT}}\right)} W_l^{(\sum_{r=l+1}^{L} M_{rC}+M_{rT})+\alpha-b}(1-W_l)^{b-1}$$

Therefore, we have the following Gibbs steps for $l = 1,...,L-1$

$p(\nu_l^{(b+1)}|\Psi,data) \sim Unif\left(0,(1-U_l^{(b)}W_l^{(b)})^{M_{lC}^{(b)}}\right)$

$$p(\gamma_l^{(b+1)}|\Psi, data) \sim Unif\left(0, (1-V_l^{(b)}W_l^{(b)})^{M_{lT}^{(b)}}\right)$$

$$p(U_l^{(b+1)}|\Psi, data) \sim Beta\left((\sum_{r=l+1}^{L} M_{rC}^{(b)}) + \alpha, 1-b\right)\mathbf{1}_{\left(0, \frac{1}{W_l^{(b)}}\left[1-exp\left(\frac{log(\nu_l^{(b+1)})}{M_{lC}^{(b)}}\right)\right]\right)}$$

$$p(V_l^{(b+1)}|\Psi, data) \sim Beta\left((\sum_{r=l+1}^{L} M_{rT}^{(b)}) + \alpha, 1-b\right)\mathbf{1}_{\left(0, \frac{1}{W_l^{(b)}}\left[1-exp\left(\frac{log(\gamma_l^{(b+1)})}{M_{lT}^{(b)}}\right)\right]\right)}$$

$$p(W_l^{(b+1)}|\Psi, data) \sim Beta\left((\sum_{r=l+1}^{L} M_{rT}^{(b)} + M_{rC}^{(b)}) + \alpha + 1 - b, b\right)\mathbf{1}_{(0,m^*)}$$

where $m^* = min\left\{\frac{1}{U_l^{(b+1)}}\left[1-exp\left(\frac{log(\nu_l^{(b+1)})}{M_{lT}^{(b)}}\right)\right], \frac{1}{V_l^{(b+1)}}\left[1-exp\left(\frac{log(\gamma_l^{(b+1)})}{M_{lT}^{(b)}}\right)\right]\right\}$

Set $\zeta_{lC}^{(b+1)} = U_l^{(b+1)}W_l^{(b+1)}$

$\zeta_{lT}^{(b+1)} = V_l^{(b+1)}W_l^{(b+1)}$

For the update for w_{is} for $i = 1, ..., n_s$ and $s \in \{C, T\}$ we have:

$$p(\mathrm{w}_{is}|data, \Psi) \propto \Gamma(t_{is}|e^{\theta_{\mathrm{w}_{is}}}, e^{\phi_{\mathrm{w}_{is}}})N(x_{is}|\beta_{\mathrm{w}_{is}}, \kappa^2_{\mathrm{w}_{is}}) \sum_{l=1}^{L}\{(1-\zeta_{ls})\prod_{r=1}^{l-1}\zeta_{rs}\}\delta_l(\mathrm{w}_{is})$$

$$p(\mathrm{w}_{is}^{(b+1)}|data, \Psi) \overset{\text{draw}}{\sim} \sum_{l=1}^{L}\tilde{p}_{lis}\delta_{(l)}(\mathrm{w}_{is})$$

where $\tilde{p}_{lis} =$

$$\frac{p_{ls}\left[\Gamma(t_{is}|e^{\theta_l^{(b+1)}}, e^{\phi_l^{(b+1)}})\right]^{1-\delta_{is}}\left[\int_{t_{is}}^{\infty}\Gamma(u_{is}|e^{\theta_l^{(b+1)}}, e^{\phi_l^{(b+1)}})du_{is}\right]^{\delta_{is}}N(x_{is}|\beta_l^{(b+1)}, \kappa_l^{2(b+1)})}{\sum_{l=1}^{L}p_{ls}\left[\Gamma(t_{is}|e^{\theta_l^{(b+1)}}, e^{\phi_l^{(b+1)}})\right]^{1-\delta_{is}}\left[\int_{t_{is}}^{\infty}\Gamma(u_{is}|e^{\theta_l^{(b+1)}}, e^{\phi_l^{(b+1)}})du_{is}\right]^{\delta_{is}}N(x_{is}|\beta_l^{(b+1)}, \kappa_l^{2(b+1)})}$$

with $p_{1s} = 1 - \zeta_{1s}$ and $p_{ls} = (1-\zeta_{ls})\prod_{r=1}^{l-1}\zeta_{rs}$ for $l = 2, ..., L-1$.

For the update for $\boldsymbol{\mu}$ we have:

$$p(\boldsymbol{\mu}|data, \Psi) \propto N_2(\boldsymbol{\mu}|a_\mu, B_\mu)\prod_{l=1}^{L}N_2((\theta_l, \phi_l)'|\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

$$p(\boldsymbol{\mu}^{(b)}|data, \Psi) \overset{\text{draw}}{\sim} N_2(m_\mu, S_\mu^2)$$

where $m_\mu = S_\mu^2\left(B_\mu^{-1}a_\mu + \boldsymbol{\Sigma}^{-1}\sum_{l=1}^{L}\boldsymbol{\theta}_l^{(b)}\right)$, $S_\mu^2 = \left(B_\mu^{-1} + L\boldsymbol{\Sigma}^{-1(b)}\right)^{-1}$.

Turning to the update of $\boldsymbol{\Sigma}$, we have:

$$p(\mathbf{\Sigma}|data,\Psi) \propto \prod_{l=1}^{L} N_2((\theta_l,\phi_l)'|\boldsymbol{\mu},\mathbf{\Sigma}) IWish(\mathbf{\Sigma}|a_\Sigma, B_\Sigma)$$

$$p(\mathbf{\Sigma}^{(b+1)}|data,\Psi) \overset{\text{draw}}{\sim} IWish(L+a_\Sigma, B_\Sigma + \sum_{l=1}^{L}(\boldsymbol{\theta}_l^{(b+1)} - \boldsymbol{\mu}^{(b+1)})(\boldsymbol{\theta}_l^{(b+1)} - \boldsymbol{\mu}^{(b+1)})')$$

For the update for λ we have:

$$p(\lambda|data,\Psi) \propto N(\lambda|a_\lambda, b_\lambda^2) \prod_{l=1}^{L} N(\beta_l|\lambda,\tau^2)$$

$$p(\lambda^{(b+1)}|data,\Psi) \overset{\text{draw}}{\sim} N(m_\lambda, s_\lambda^2)$$

where $m_\lambda = s_\lambda^2 \left(b_\lambda^{-2} a_\lambda + \tau^{-2} \sum_{l=1}^{L} \beta_l \right)$

and $s_\lambda^2 = \left(b_\lambda^{-2} + \tau^{-2(b)} L \right)^{-1}$.

For the update for τ^2 we have:

$$p(\tau^2|data,\Psi) \propto \Gamma^{-1}(\tau^2|a_\tau, b_\tau) \prod_{l=1}^{L} N(\beta_l|\lambda,\tau^2)$$

$$p(\tau^{2(b+1)}|data,\Psi) \overset{\text{draw}}{\sim} \Gamma^{-1}\left(0.5L + a_\tau, 0.5\left[\sum_{l=1}^{L}(\beta_l^{(b+1)} - \lambda^{(b+1)})^2\right] + b^\tau\right)$$

For the update for ρ we have:

$$p(\rho|data,\Psi) \propto \Gamma(\rho|a_\rho, b_\rho) \prod_{l=1}^{L} \Gamma^{-1}(\kappa_l^2|a,\rho)$$

$$p(\rho^{(b+1)}|data,\Psi) \overset{\text{draw}}{\sim} \Gamma\left(aL + a_\rho, \left[\sum_{l=1}^{L} \kappa_l^{-2(b+1)}\right] + b_\rho\right)$$

We do not have conjugacy for α and b, so we turn to the Metropolis-Hastings algorithm to update these parameters. The Bivariate Beta density of (ζ_c, ζ_T), has a complicated form, however, we can work with the density of the latent variables, (U, V, W):

$$p(\alpha, b|data,\Psi) \propto$$

$$Unif(b|0,1)\Gamma(\alpha|a_\alpha, b_\alpha) \prod_{l-1}^{L-1} Beta(U_l|\alpha, 1-b) Beta(V_l|\alpha, 1-b) Beta(W_l|1+\alpha-b, b)$$

We sample from the proposal distribution,

$(log(\alpha'), logit(b'))' \sim N_2((log(\alpha^{(b)}), logit(b^{(b)})), cS^2_{\alpha b})$, where $S^2_{\alpha b}$ is updated from the average variances and covariance of posterior samples of $((log(\alpha), logit(b))$ under initial runs, and c is updated from initial runs to optimize mixing. Draw $\eta \sim Unif(0,1)$.

If $\eta < min\{1,$

$$\left.\frac{\Gamma(\alpha'|a_\alpha,b_\alpha)\prod_{l-1}^{L-1} Beta(U_l|\alpha',1-b')Beta(V_l|\alpha',1-b')Beta(W_l|1+\alpha'-b',b')\alpha'b'(1-b')}{\Gamma(\alpha^{(b)}|a_\alpha,b_\alpha)\prod_{l-1}^{L-1} Beta(U_l|\alpha^{(b)},1-b^{(b)})Beta(V_l|\alpha^{(b)},1-b^{(b)})Beta(W_l|1+\alpha^{(b)}-b^{(b)},b^{(b)})\alpha^{(b)}b^{(b)}(1-b^{(b)})}\right\}$$

set $(\alpha^{(b+1)}, b^{(b+1)})' = (\alpha', b')'$

else $(\alpha^{(b+1)}, b^{(b+1)})' = (\alpha^{(b)}, b^{(b)})'$

D.2 Conditional Predictive Ordinate for gamma DDPMM

Here we provide the details of how we arrived to the expression necessary for computing the CPO values under the gamma DDPMM. As our data example in Section 3.4.1 does not contain any random covariates, we will derive the expression without covariates, however, the derivation can easily be extended to include random covariates in the curve-fitting setting.

Recall that the model, under the truncated version, can be written in hierar-

chical form as,

$$
\begin{aligned}
t_{iC}|\mathrm{w}_{iC},\boldsymbol{\theta} &\overset{ind}{\sim} \Gamma(t_{iC}|\boldsymbol{\theta}_{\mathrm{w}_{iC}}) \quad \text{for} i=1,...,n_C \\
t_{iT}|\mathrm{w}_{iT},\boldsymbol{\theta} &\overset{ind}{\sim} \Gamma(t_{iT}|\boldsymbol{\theta}_{\mathrm{w}_{iT}}) \quad \text{for} i=1,...,n_T \\
\mathbf{w}|\{\zeta_{lC},\zeta_{lT}\} &\sim \prod_{s\in\{C,T\}}\prod_{i=1}^{n_s}\sum_{l=1}^{L}\left[(1-\zeta_{ls})\prod_{r=1}^{l-1}\zeta_{rs}\right]\delta_l(\mathrm{w}_{is}) \\
\boldsymbol{\theta}_l|\boldsymbol{\mu},\boldsymbol{\Sigma} &\overset{iid}{\sim} N_2(\boldsymbol{\theta}_l|\boldsymbol{\mu},\boldsymbol{\Sigma}) \\
(\zeta_{lC},\zeta_{lT})|\alpha,b &\overset{iid}{\sim} Biv-Beta((\zeta_{lC},\zeta_{lT})|\alpha,b) \quad \text{for} l=1,...,L-1
\end{aligned}
$$

with priors, $\alpha\sim\Gamma(\alpha|a_\alpha,b_\alpha)$, $b\sim Unif(b|0,1)$, $\boldsymbol{\mu}\sim N_2(\boldsymbol{\mu}|a_\mu,B_\mu)$,

and $\boldsymbol{\Sigma}\sim IWish(\boldsymbol{\Sigma}|a_\Sigma,B_\Sigma)$.

Let $\Psi=(\alpha,b,\boldsymbol{\mu},\boldsymbol{\Sigma})$. The predictive density for a new survival time from group $s\in\{C,T\}$, t_{0s}, is given by:

$$
\begin{aligned}
p(t_{0s}|data) &= \int\int \Gamma(t_{0s}|\boldsymbol{\theta}_{\mathrm{w}_{0s}})\left(\textstyle\sum_{l=1}^{L}p_{ls}\delta_l(\mathrm{w}_{0s})\right)p(\boldsymbol{\theta},\boldsymbol{p},\mathbf{w},\Psi|data)d\mathrm{w}_{0s}d\boldsymbol{\theta}d\mathbf{w}d\boldsymbol{p}d\Psi \\
&= \int\left(\textstyle\sum_{l=1}^{L}p_{ls}\Gamma(t_{0s}|\boldsymbol{\theta}_l)\right)p(\boldsymbol{\theta},\boldsymbol{p},\mathbf{w},\Psi|data)d\boldsymbol{\theta}d\mathbf{w}d\boldsymbol{p}d\Psi
\end{aligned}
$$

Let s' be the experimental group that s is not, $data=\{\boldsymbol{t}_s,\boldsymbol{t}_{s'}\}$, and A be the normalizing constant for $p(\boldsymbol{\theta},\boldsymbol{p},\mathbf{w},\Psi|data)$. Namely, $p(\boldsymbol{\theta},\boldsymbol{p},\mathbf{w},\Psi|data)=$

$$
\frac{\left\{\prod_{i=1}^{n_s}\Gamma(t_{is}|\boldsymbol{\theta}_{\mathrm{w}_{is}})\right\}\left\{\prod_{i=1}^{n_{s'}}\Gamma(t_{is'}|\boldsymbol{\theta}_{\mathrm{w}_{is'}})\right\}p(\boldsymbol{\theta},\boldsymbol{p},\mathbf{w},\Psi)}{\int\left\{\prod_{i=1}^{n_s}\Gamma(t_{is}|\boldsymbol{\theta}_{\mathrm{w}_{is}})\right\}\left\{\prod_{i=1}^{n_{s'}}\Gamma(t_{is'}|\boldsymbol{\theta}_{\mathrm{w}_{is'}})\right\}p(\boldsymbol{\theta},\boldsymbol{p},\mathbf{w},\Psi)d\boldsymbol{\theta}d\mathbf{w}d\boldsymbol{p}d\Psi}
$$

Note that $p(\boldsymbol{\theta},\boldsymbol{p},\mathbf{w},\Psi)=N_2(\boldsymbol{\theta}|\boldsymbol{\mu},\boldsymbol{\Sigma})\left(\prod_{i=1}^{n_s}\sum_{l=1}^{L}p_{ls}\delta_l(\mathrm{w}_{is})\right)\left(\prod_{i=1}^{n_{s'}}\sum_{l=1}^{L}p_{ls'}\delta_l(\mathrm{w}_{is'})\right)$

$Biv-Beta(\boldsymbol{p}\equiv(\zeta_s,\zeta_{s'})|\alpha,b)\Gamma(\alpha|a_\alpha,b_\alpha)Unif(b|0,1)N_2(\boldsymbol{\mu}|a_\mu,B_\mu)IWish(\boldsymbol{\Sigma}|a_\Sigma,B_\Sigma)$.

The CPO of the ith survival time in group s is defined as,

$$
\begin{aligned}
CPO_{is} &= p(t_{is}|\boldsymbol{t}_{(-i)s},\boldsymbol{t}_{s'}) \\
&= \int\Gamma(t_{is}|\boldsymbol{\theta}_{\mathrm{w}_{0s}})\left(\textstyle\sum_{l=1}^{L}p_{ls}\delta_l(\mathrm{w}_{0s})\right)p(\boldsymbol{\theta},\boldsymbol{p},\mathbf{w}_{(-i)s},\Psi)d\boldsymbol{\theta}d\mathbf{w}_{(-i)s}d\boldsymbol{p}d\Psi d\mathrm{w}_{0s}
\end{aligned}
$$

where $\mathbf{w}_{(-i)s}$ is the vector $\mathbf{w}$ with the i^{th} member of group s removed. Similarly, $data_{(-i)s}$ represents $data$ with the i^{th} member in group s removed.

Now, consider $p(\boldsymbol{\theta}, \boldsymbol{p}, \mathbf{w}_{(-i)s}, \Psi | data_{(-i)s}) =$

$$\frac{p(data_{(-i)s}|\boldsymbol{\theta}, \mathbf{w}_{(-i)s}) p(\boldsymbol{\theta}, \mathbf{w}_{(-i)s}, \boldsymbol{p}, \Psi)}{\int p(data_{(-i)s}|\boldsymbol{\theta}, \mathbf{w}_{(-i)s}) p(\boldsymbol{\theta}, \mathbf{w}_{(-i)s}, \boldsymbol{p}, \Psi) d\mathbf{w}_{(-i)s} d\boldsymbol{p} d\Psi}$$
$$= \frac{\left\{\prod_{j \neq i}^{n_s} \Gamma(t_{js}|\boldsymbol{\theta}_{\mathrm{w}_{js}})\right\} \left\{\prod_{i=1}^{n_{s'}} \Gamma(t_{is'}|\boldsymbol{\theta}_{\mathrm{w}_{is'}})\right\} p(\boldsymbol{\theta}, \boldsymbol{p}, \mathbf{w}_{(-i)s}, \Psi)}{\int \left\{\prod_{j \neq i}^{n_s} \Gamma(t_{js}|\boldsymbol{\theta}_{\mathrm{w}_{js}})\right\} \left\{\prod_{i=1}^{n_{s'}} \Gamma(t_{is'}|\boldsymbol{\theta}_{\mathrm{w}_{is'}})\right\} p(\boldsymbol{\theta}, \boldsymbol{p}, \mathbf{w}_{(-i)s}, \Psi) d\boldsymbol{\theta} d\mathbf{w}_{(-i)s} d\boldsymbol{p} d\Psi}$$

Let B_{is} be the normalizing constant of $p(\boldsymbol{\theta}, \boldsymbol{p}, \mathbf{w}_{(-i)s}, \Psi | data_{(-i)s})$:

$$B_{is} = \int \left\{\prod_{j \neq i}^{n_s} \Gamma(t_{js}|\boldsymbol{\theta}_{\mathrm{w}_{js}})\right\} \left\{\prod_{i=1}^{n_{s'}} \Gamma(t_{is'}|\boldsymbol{\theta}_{\mathrm{w}_{is'}})\right\} p(\boldsymbol{\theta}, \boldsymbol{p}, \mathbf{w}_{(-i)s}, \Psi) d\boldsymbol{\theta} d\mathbf{w}_{(-i)s} d\boldsymbol{p} d\Psi$$

Then, we can write

$$\begin{aligned} p(\boldsymbol{\theta}, \boldsymbol{p}, \mathbf{w}_{(-i)s}, \Psi | data_{(-i)s}) &= \frac{\left\{\prod_{i=1}^{n_s} \Gamma(t_{is}|\boldsymbol{\theta}_{\mathrm{w}_{is}})\right\} \left\{\prod_{i=1}^{n_{s'}} \Gamma(t_{is'}|\boldsymbol{\theta}_{\mathrm{w}_{is'}})\right\} p(\boldsymbol{\theta}, \boldsymbol{p}, \mathbf{w}, \Psi)}{B_{is} \Gamma(t_{is}|\boldsymbol{\theta}_{\mathrm{w}_{is}}) p(\mathrm{w}_{is}|\boldsymbol{p})} \\ &= \frac{A}{B_{is}} \frac{p(\boldsymbol{\theta}, \boldsymbol{p}, \mathbf{w}, \Psi | data)}{\Gamma(t_{is}|\boldsymbol{\theta}_{\mathrm{w}_{is}}) p(\mathrm{w}_{is}|\boldsymbol{p})} \end{aligned}$$

Thus,

$$\begin{aligned} CPO_{is} &= \int \Gamma(t_{is}|\boldsymbol{\theta}_{\mathrm{w}_{0s}}) p(\mathrm{w}_{0s}|\boldsymbol{p}) p(\boldsymbol{\theta}, \boldsymbol{p}, \mathbf{w}_{(-i)s}, \Psi) d\boldsymbol{\theta} d\mathbf{w}_{(-i)s} d\boldsymbol{p} d\Psi d\mathrm{w}_{0s} \\ &= \int \Gamma(t_{is}|\boldsymbol{\theta}_{\mathrm{w}_{0s}}) \left(\int p(\mathrm{w}_{0s}, \mathrm{w}_{is}|\boldsymbol{p}) d\mathrm{w}_{is} \right) p(\boldsymbol{\theta}, \boldsymbol{p}, \mathbf{w}_{(-i)s}, \Psi) d\boldsymbol{\theta} d\mathbf{w}_{(-i)s} d\boldsymbol{p} d\Psi d\mathrm{w}_{0s} \\ &= \frac{A}{B_{is}} \int \frac{\Gamma(t_{is}|\boldsymbol{\theta}_{\mathrm{w}_{0s}}) p(\mathrm{w}_{0s}, \mathrm{w}_{is}|\boldsymbol{p})}{\Gamma(t_{is}|\boldsymbol{\theta}_{\mathrm{w}_{is}}) p(\mathrm{w}_{is}|\boldsymbol{p})} p(\boldsymbol{\theta}, \boldsymbol{p}, \mathbf{w}, \Psi | data) d\mathrm{w}_{0s} d\boldsymbol{\theta} d\mathbf{w} d\boldsymbol{p} d\Psi \\ &\quad \text{(Note: } p(\mathrm{w}_{0s}|\mathrm{w}_{is}\boldsymbol{p}) = p(\mathrm{w}_{0s}|\boldsymbol{p})\text{)} \\ &= \frac{A}{B_{is}} \int \frac{\sum_{l=1}^{L} p_{ls} \Gamma(t_{is}|\boldsymbol{\theta}_l)}{\Gamma(t_{is}|\boldsymbol{\theta}_{\mathrm{w}_{is}})} p(\boldsymbol{\theta}, \boldsymbol{p}, \mathbf{w}, \Psi | data) d\mathrm{w}_{0s} d\boldsymbol{\theta} d\mathbf{w} d\boldsymbol{p} d\Psi \end{aligned}$$

All that is left is to be able to evaluate A/B_{is}:

$$\begin{aligned}
\left(\frac{A}{B_{is}}\right)^{-1} &= \frac{1}{A}\int\left\{\prod_{j\neq i}^{n_s}\Gamma(t_{js}|\boldsymbol{\theta}_{\mathrm{w}_{js}})\right\}\left\{\prod_{i=1}^{n_{s'}}\Gamma(t_{is'}|\boldsymbol{\theta}_{\mathrm{w}_{is'}})\right\}\underbrace{\left(\int p(\mathrm{w}_{is}|\mathbf{w}_{(-i)s},\boldsymbol{p})d\mathrm{w}_{is}\right)}_{1} \\
&\quad \times p(\mathbf{w}_{(-i)s}|\boldsymbol{p})p(\boldsymbol{p},\boldsymbol{\theta},\Psi)d\boldsymbol{\theta}d\mathbf{w}_{(-i)s}d\boldsymbol{p}d\Psi \\
&= \frac{1}{A}\int\left\{\prod_{j\neq i}^{n_s}\Gamma(t_{js}|\boldsymbol{\theta}_{\mathrm{w}_{js}})\right\}\left\{\prod_{i=1}^{n_{s'}}\Gamma(t_{is'}|\boldsymbol{\theta}_{\mathrm{w}_{is'}})\right\}p(\boldsymbol{\theta},\boldsymbol{p},\mathbf{w},\Psi)d\boldsymbol{\theta}d\mathbf{w}d\boldsymbol{p}d\Psi \\
&= \frac{1}{A}\int\frac{\left\{\prod_{j\neq i}^{n_s}\Gamma(t_{js}|\boldsymbol{\theta}_{\mathrm{w}_{js}})\right\}\left\{\prod_{i=1}^{n_{s'}}\Gamma(t_{is'}|\boldsymbol{\theta}_{\mathrm{w}_{is'}})\right\}}{\Gamma(t_{is}|\boldsymbol{\theta}_{\mathrm{w}_{is}})}p(\boldsymbol{\theta},\boldsymbol{p},\mathbf{w},\Psi)d\boldsymbol{\theta}d\mathbf{w}d\boldsymbol{p}d\Psi \\
&= \int\frac{1}{\Gamma(t_{is}|\boldsymbol{\theta}_{\mathrm{w}_{is}})}p(\boldsymbol{\theta},\boldsymbol{p},\mathbf{w},\Psi)d\boldsymbol{\theta}d\mathbf{w}d\boldsymbol{p}d\Psi
\end{aligned}$$

Appendix E

Posterior sampling for model for mrl ordered populations

Here we show the algorithm used to obtain the posterior samples of the parameters under the nonparametric Erlang mixture model for mrl ordered populations. We use a Pólya-urn based MCMC, with Metropolis-Hastings steps when conjugacy is not obtainable. By introducing two sets of latent variables, $\mathbf{w} = \{\mathrm{w}_k : k = 1, ..., n_2\}$, $\boldsymbol{z} = \{z_j : j = 1, ..., n_1\}$, we can write the hierarchical version of our fully nonparametric Bayesian model for inference for mrl ordering between to groups,

$$
\begin{aligned}
t_{1j}|\mathrm{w}_j, z_j, \theta &\overset{ind}{\sim} \sum_{m=1}^{M-1} E_m(t_{1j}|\theta)\mathbb{1}_{((m-1)\theta, m\theta]}(min\{\mathrm{w}_j, z_j\}) \\
&\quad + E_M(t_{1j}|\theta)\mathbb{1}_{((M-1)\theta,\infty)}(min\{\mathrm{w}_j, z_j\}), \quad j = 1, ..., n_1 \\
t_{2k}|\mathrm{w}_{n_1+k}, \theta &\overset{ind}{\sim} \sum_{m=1}^{M-1} E_m(t_{2k}|\theta)\mathbb{1}_{((m-1)\theta, m\theta]}(\mathrm{w}_{n_1+k}) \\
&\quad + E_M(t_{2k}|\theta)\mathbb{1}_{((M-1)\theta,\infty)}(\mathrm{w}_{n_1+k}), \quad k = 1, ..., n_2 \\
z_j|G_1 &\overset{iid}{\sim} G_1, \quad j = 1, ..., n_1 \\
\mathrm{w}_k|G_2 &\overset{iid}{\sim} G_2, \quad k = 1, ..., n_1 + n_2 \\
G_l|\alpha_l, \phi_l &\overset{ind}{\sim} DP(\alpha_l, G_{0l} \equiv LN(\mu_l, \sigma_l^2)), \quad l = 1, 2
\end{aligned}
$$

with the following priors $M \sim Unif(2, M_{max})$,$\alpha_l \overset{ind}{\sim} \Gamma(a_\alpha, b_\alpha)$, $\mu_l \overset{ind}{\sim} N(a_{\mu_l}, b_{\mu_l})$, $\sigma_l^2 \overset{ind}{\sim} \Gamma^{-1}(a_{\sigma_l}, b_{\sigma_l})$, for $l = 1, 2$, and $\theta \sim \Gamma^{-1}(a_\theta, b_\theta)$. Let $\boldsymbol{\alpha} = \{\alpha_1, \alpha_2\}$, and $\boldsymbol{\phi} = \{\mu_1, \mu_2, \sigma_1^2, \sigma_2^2\}$. We need to obtain samples from the full posterior distribution, $p(G_1, G_2, \boldsymbol{z}, \mathbf{w}, \theta, \boldsymbol{\alpha}, \boldsymbol{\phi}, M|data)$. In order to do so, we break the full posterior distribution into the posterior joint marginal of the posterior parameters and the posterior conditional of the random distributions given the parameters Antoniak (1974):

$p(G_1, G_2, \boldsymbol{z}, \mathbf{w}, \theta, \boldsymbol{\alpha}, \boldsymbol{\phi}, M|data)$

$= p(G_1|\boldsymbol{z}, \alpha_1, \mu_1, \sigma_1^2, \theta, data)p(G_2|\mathbf{w}, \alpha_2, \mu_2, \sigma_2^2, \theta, data)p(\boldsymbol{z}, \mathbf{w}, \theta, \boldsymbol{\alpha}, \boldsymbol{\phi}, M|data)$.

Repeated sequential sampling from the following conditionals yields the algorithm used to obtain posterior samples from $p(\boldsymbol{z}, \mathbf{w}, \theta, \boldsymbol{\alpha}, \boldsymbol{\phi}, M|data)$. For $i = 1, ..., n_l$, let $\delta_{li} = 0$ if t_{li} is observed and $\delta_{li} = 1$ if t_{li} is right censored for $l \in \{1, 2\}$. Let n_{w}^* be the number of distinct components in $\mathbf{w}$, and $\mathbf{w}_k^* \equiv \{\mathrm{w}_k^* : k = 1, ..., n_2\}$ be the vector of distinct components in $\mathbf{w}$. Let n_z^* be the number of distinct components in $\boldsymbol{z}$, and

$\boldsymbol{z}_j^* \equiv \{\boldsymbol{z}_j^* : j = 1, ..., n_1\}$ be the vector of distinct components in $\boldsymbol{z}$ For $i = 1, ..., n_s$, let $\delta_{is} = 0$ if t_{is} is observed and $\delta_{is} = 1$ if t_{is} is right censored for $s \in \{1, 2\}$. Let Ψ represent the vector of the most recent iteration of all other parameters. Let $b = 1, ..., B$ index the iterations of the MCMC.

Define the function $k(t|\theta, M, min\{\mathrm{w}, z\}) =$

$\sum_{m=1}^{M-1} e_m(t_{1j}|\theta)\mathbb{1}_{((m-1)\theta,m\theta]}(min\{\mathrm{w}_j, z_j\}) + e_M(t_{1j}|\theta)\mathbb{1}_{((M-1)\theta,\infty)}(min\{\mathrm{w}_j, z_j\})$,

and define $K^S(t|\theta, M, min\{\mathrm{w}, z\}) =$

$\sum_{m=1}^{M-1} E_m^S(t_{1j}|\theta)\mathbb{1}_{((m-1)\theta,m\theta]}(min\{\mathrm{w}_j, z_j\}) + E_M^S(t_{1j}|\theta)\mathbb{1}_{((M-1)\theta,\infty)}(min\{\mathrm{w}_j, z_j\})$.

For $j = 1, ..., n_1$, we update $\{\mathrm{w}_j, z_j\}$ by:

$$p(\{\mathrm{w}_j, z_j\}|\{\mathrm{w}_r, z_r : r \neq j\}, \Psi, data) \propto$$

$$[k(t_{1j}|\theta, M, min\{\mathrm{w}_j, z_j\})]^{1-\delta_{1j}}[K^S(t_{1j}|\theta, M, min\{\mathrm{w}_j, z_j\})]^{\delta_{1j}} p(z_j|\{z_r : r \neq$$

$$j\}, \Psi)p(\mathrm{w}_j|\{\mathrm{w}_r : r \neq j\}, \Psi)$$

where $p(z_j|\{z_r : r \neq j\}, \Psi) = \frac{\alpha_1}{\alpha_1+n_1-1} LN(z_j|\mu_1, \sigma_1^2) + \frac{1}{\alpha_1+n_1-1}\sum_{r\neq j}^{n_1-1} \delta_{z_r}(z_j)$

and $p(\mathrm{w}_j|\{\mathrm{w}_r : r \neq j\}, \Psi) = \frac{\alpha_2}{\alpha_2+n_1-1} LN(\mathrm{w}_j|\mu_2, \sigma_2^2) + \frac{1}{\alpha_2+n_1-1}\sum_{r\neq j}^{n_1-1} \delta_{\mathrm{w}_r}(\mathrm{w}_j)$

We use a Metropolis-Hastings step for this update. We sample from the prior as a proposal distribution, $\mathrm{w}'_j \sim p(\mathrm{w}_j^{(b)}|\{\mathrm{w}_r^{(b+1)} : r < j\}, \{\mathrm{w}_r^{(b)} : r > j\}\Psi)$ and $\mathrm{z}'_j \sim p(\mathrm{z}_j^{(b)}|\{\mathrm{z}_r^{(b+1)} : r < j\}, \{\mathrm{z}_r^{(b)} : r > j\}\Psi)$. Draw $\eta \sim Unif(0, 1)$.

If $\eta < min\left\{1, \frac{[k(t_{1j}|\theta^{(b)},M^{(b)},min\{\mathrm{w}'_j,z'_j\})]^{1-\delta_{1j}}[K^S(t_{1j}|\theta^{(b)},M^{(b)},min\{\mathrm{w}'_j,z'_j\})]^{\delta_{1j}}}{[k(t_{1j}|\theta^{(b)},M^{(b)},min\{\mathrm{w}_j^{(b)},z_j^{(b)}\})]^{1-\delta_{1j}}[K^S(t_{1j}|\theta^{(b)},M^{(b)},min\{\mathrm{w}_j^{(b)},z_j^{(b)}\})]^{\delta_{1j}}}\right\}$

set $\{\mathrm{w}_j^{(b+1)}, z_j^{(b+1)}\} = \{\mathrm{w}'_j, z'_j\}$

else $\{\mathrm{w}_j^{(b+1)}, z_j^{(b+1)}\} = \{\mathrm{w}_j^{(b)}, z_j^{(b)}\}$

For $k = 1, ..., n_2$, we update w_{n_1+k} by:

$$p(\mathrm{w}_{n_1+k}|\{\mathrm{w}_r : r \neq n_1+k\}, \Psi, data) =$$
$$\frac{q_0 h(\mathrm{w}_{n_1+k}|\theta,M,\alpha_2,\mu_2,\sigma_2^2)+\sum_{r\neq n_1+k}^{n_{\mathrm{w}}^*-1} n_{\mathrm{w}r}^{reps}\left([k(t_{2k}|\theta,M,\mathrm{w}_{n_1+k}]^{1-\delta_{2k}}[K^S(t_{2k}|\theta,M,\mathrm{w}_{n_1+k})]^{\delta_{2k}}\right)\delta_{\mathrm{w}_r^*}(\mathrm{w}_{n_1+k})}{q_0+\sum_{r\neq n_1+k}^{n_{\mathrm{w}}^*-1} n_{\mathrm{w}r}^{reps}\left([k(t_{2k}|\theta,M,\mathrm{w}_{n_1+k}]^{1-\delta_{2k}}[K^S(t_{2k}|\theta,M,\mathrm{w}_{n_1+k})]^{\delta_{2k}}\right)\delta_{\mathrm{w}_r^*}(\mathrm{w}_{n_1+k})}$$

where $q_0 =$

$\alpha_2 \left(\sum_{m=1}^{M-1}[e_m(t_{2k}|\theta)]^{1-\delta_{2k}}[E_m^S(t_{2k}|\theta)]^{\delta_{2k}}\left[LN(m\theta|\mu_2,\sigma_2^2) - LN((m-1)\theta|\mu_2,\sigma_2^2)\right]\right)$

$+\alpha_2\left([e_M(t_{2k}|\theta)]^{1-\delta_{2k}}[E_M^S(t_{2k}|\theta)]^{\delta_{2k}}\left[1 - LN((M-1)\theta|\mu_2,\sigma_2^2)\right]\right)$, with $LN(\cdot)$ representing the cdf of the gamma distribution), and $h(\mathrm{w}_{n_1+k}|\theta, M, \alpha_2, \mu_2, \sigma_2^2) =$

$\alpha_2 \left[k(t_{2k}|\theta, M, \mathrm{w}_{n_1+k})\right]^{1-\delta_{2k}}\left[K^S(t_{2k}|\theta, M, \mathrm{w}_{n_1+k})\right]^{\delta_{2k}} g_0(\mathrm{w}_{n_1+k}|\mu_2,\sigma_2^2)/q_0$, (where $g_0(\cdot)$ is the lognormal density function).

We update M and θ together, since the two parameter are highly correlated:

$$p(M,\theta|\Psi,data) \propto \left(\prod_{j=1}^{n_1}[k(t_{1j}|\theta,M,min\{\mathrm{w}_j,z_j\})]^{1-\delta_{1j}}[K^S(t_{1j}|\theta,M,min\{\mathrm{w}_j,z_j\})]^{\delta_{1j}}\right)$$
$$\times\left(\prod_{k=1}^{n_2}[k(t_{2k}|\theta,M,\mathrm{w}_{n_1+k})]^{1-\delta_{2k}}[K^S(t_{2k}|\theta,M,\mathrm{w}_{n_1+k})]^{\delta_{2k}}\right)\Gamma^{-1}(\theta|a_\theta,b_\theta)$$

We use a Metropolis-Hastings step for this update. We sample $M' \sim Unif(max\{2, M^{(b)} - S_M\}, min\{M_{max}, M^{(b)} + S_M)$, and $log(\theta') \sim N(log(\theta^{(b)}, S_\theta^2)$. Draw $\eta \sim Unif(0,1)$

$$\text{If } \eta < min\left\{1, \frac{\left(\prod_{j=1}^{n_1}[k(t_{1j}|\theta',M',min\{\mathrm{w}_j^{(b+1)},z_j^{(b+1)}\})]^{1-\delta_{1j}}[K^S(t_{1j}|\theta',M',min\{\mathrm{w}_j^{(b+1)},z_j^{(b+1)}\})]^{\delta_{1j}}\right)}{\left(\prod_{j=1}^{n_1}[k(t_{1j}|\theta^{(b)},M^{(b)},min\{\mathrm{w}_j^{(b+1)},z_j^{(b+1)}\})]^{1-\delta_{1j}}[K^S(t_{1j}|\theta^{(b)},M^{(b)},min\{\mathrm{w}_j^{(b+1)},z_j^{(b+1)}\})]^{\delta_{1j}}\right)}\right.$$
$$\left.\times\frac{\left(\prod_{k=1}^{n_2}[k(t_{2k}|\theta',M',\mathrm{w}_{n_1+k}^{(b+1)})]^{1-\delta_{2k}}[K^S(t_{2k}|\theta',M',\mathrm{w}_{n_1+k}^{(b+1)})]^{\delta_{2k}}\right)\Gamma^{-1}(\theta'|a_\theta,b_\theta)\theta'}{\left(\prod_{k=1}^{n_2}[k(t_{2k}|\theta^{(b)},M^{(b)},\mathrm{w}_{n_1+k}^{(b+1)})]^{1-\delta_{2k}}[K^S(t_{2k}|\theta^{(b)},M^{(b)},\mathrm{w}_{n_1+k}^{(b+1)})]^{\delta_{2k}}\right)\Gamma^{-1}(\theta^{(b)}|a_\theta,b_\theta)\theta^{(b)}}\right\}$$

set $\{M^{(b+1)}, \theta^{(b+1)}\} = \{M', \theta'\}$

else $\{M^{(b+1)}, \theta^{(b+1)}\} = \{M^{(b)}, \theta^{(b)}\}$

For the update of α_1:

$$p(\alpha_1|n_z^*, \Psi, data) \propto \Gamma(\alpha_1|a_\alpha, b_\alpha)\alpha_1^{n_z^*}\frac{\Gamma(\alpha_1)}{\Gamma(\alpha_1+n_1)}$$

Introducing the latent variable η_1, we can update α_1 via:

Sample $\eta_1 \sim Beta(\alpha_1^{(b)}+1, n_1)$

Sample $\alpha_1^{(b+1)} \sim p\Gamma(a_\alpha + n_z^{*(b+1)}, b_\alpha - log(\eta_1)) + (1-p)\Gamma(a_\alpha + n_z^{*(b+1)} - 1, b_\alpha - log(\eta_1))$

where $p = (a_\alpha + n_z^{*(b+1)} - 1)/\{n_1(b_\alpha - log(\eta_1)) + a_\alpha + n_z^{*(b+1)} - 1\}$

For the update of α_2:

$$p(\alpha_2|n_{\mathrm{w}}^*, \Psi, data) \propto \Gamma(\alpha_2|a_\alpha, b_\alpha)\alpha_2^{n_{\mathrm{w}}^*}\frac{\Gamma(\alpha_2)}{\Gamma(\alpha_2+n_1+n_2)}$$

Introducing the latent variable η_2, we can update α_1 via:

Sample $\eta_2 \sim Beta(\alpha_2^{(b)}+1, n_1+n_2)$

Sample $\alpha_2^{(b+1)} \sim p\Gamma(a_\alpha + n_{\mathrm{w}}^{*(b+1)}, b_\alpha - log(\eta_2)) + (1-p)\Gamma(a_\alpha + n_{\mathrm{w}}^{*(b+1)} - 1, b_\alpha - log(\eta_2))$

where $p = (a_\alpha + n_{\mathrm{w}}^{*(b+1)} - 1)/\{(n_1+n_2)(b_\alpha - log(\eta_2)) + a_\alpha + n_{\mathrm{w}}^{*(b+1)} - 1\}$

For the update of μ_1:

$$p(\mu_1|\boldsymbol{z}^*, n_z^*, \Psi, data) \propto N(\mu_1|a_{\mu_1}, b_{\mu_1})\left(\prod_{j=1}^{n_z^*} LN(z_j^*|\mu_1, \sigma_1^2)\right)$$

Sample $\mu_1^{(b+1)} \sim N(m_1, s_1^2)$

where $s_1^2 = \left(\frac{1}{b_{\mu_1}} + \frac{n_z^{*(b+1)}}{\sigma_1^{2(b)}}\right)^{-1}$ and $m = \left(\frac{a_{\mu_1}}{b_{\mu_1}} + \frac{\sum_{j=1}^{n_z^{*(b+1)}} log(z_j^{*(b+1)})}{\sigma_1^{2(b)}}\right)s_1^2$

For the update of μ_2:

$$p(\phi_2|\mathbf{w}^*, n_{\mathrm{w}}^*, \Psi, data) \propto N(\mu_2|a_{\mu_2}, b_{\mu_2}) \left(\prod_{k=1}^{n_{\mathrm{w}}^*} LN(\mathrm{w}_k^*|\mu_2, \sigma_2^2)\right)$$

Sample $\mu_2^{(b+1)} \sim N(m_2, s_2^2)$

where $s_2^2 = \left(\frac{1}{b_{\mu_2}} + \frac{n_{\mathrm{w}}^{*(b+1)}}{\sigma_2^{2(b)}}\right)^{-1}$ and $m = \left(\frac{a_{\mu_2}}{b_{\mu_2}} + \frac{\sum_{k=1}^{n_{\mathrm{w}}^{*(b+1)}} log(\mathrm{w}_j^{*(b+1)})}{\sigma_2^{2(b)}}\right) s_1^2$

For the update of σ_1^2:

$$p(\sigma_1^2|\boldsymbol{z}^*, n_z^*, \Psi, data) \propto \Gamma^{-1}(\sigma_1^2|a_{\sigma_1}, b_{\sigma_1}) \left(\prod_{j=1}^{n_z^*} LN(z_j^*|\mu_1, \sigma_1^2)\right)$$

Sample $\sigma_1^{2(b+1)} \sim \Gamma^{-1}(a_{\sigma_1} + 0.5 n_z^{*(b+1)}, b_{\sigma_1} + 0.5 \sum_{j=1}^{n_z^{*(b+1)}} log(z^{*(b+1)} - \mu_1^{(b+1)})^2)$

For the update of σ_2^2:

$$p(\sigma_2^2|\mathbf{w}^*, n_{\mathrm{w}}^*, \Psi, data) \propto \Gamma^{-1}(\sigma_2^2|a_{\sigma_1}, b_{\sigma_1}) \left(\prod_{k=1}^{n_{\mathrm{w}}^*} LN(\mathrm{w}_k^*|\mu_2, \sigma_2^2)\right)$$

Sample $\sigma_2^{2(b+1)} \sim \Gamma^{-1}(a_{\sigma_1} + 0.5 n_{\mathrm{w}}^{*(b+1)}, b_{\sigma_2} + 0.5 \sum_{k=1}^{n_{\mathrm{w}}^{*(b+1)}} log(\mathrm{w}^{*(b+1)} - \mu_2^{(b+1)})^2)$