Bayesian Nonparametric Survival Analysis using mixture of Burr XII distributions

S. B. Hajjar

bohlurihajjar.soghra@razi.ac.ir

S. Khazaei

s.khazaei@razi.ac.ir

Abstract

Recently, the Bayesian nonparametric approach in survival studies attracts much more attentions. Because of multi modality in survival data, the mixture models are very common in this field. One of the famous priors on Bayesian nonparametric models is Dirichlet process prior. In this paper we introduce a Bayesian nonparametric mixture model with Burr distribution(Burr type XII) as the kernel of mixture model. Since the Burr distribution shares good properties of common distributions on survival analysis, it has more flexibility than other distributions. By applying this model to simulated and real failure time data sets, we show the preference of this model and compare it with other Dirichlet process mixture models with different kernels. And also we show that this model can be applied for the right censored data. For calculating the posterior of the parameters for inference and modeling, we used the MCMC simulation methods, especially Gibbs sampling.

Keywords: Bayesian nonparametric, Dirichlet process, Burr XII distribution, survival analysis, right censored data.

1. Introduction

Many distributions are commonly used for modeling failure time data, whether for reliability studies in manufacturing or survival analysis in the health domain. These include the exponential, Weibull, log-normal and log-logistic distributions. The log-normal distribution is an attractive distribution for modeling component failure times because it has non-monotone failure rates. Also, the log-logistic distribution has been used because it shares this property. The Burr(XII) distribution has recently emerged as a promising distribution for use with failure time data ([17],[11]). It shares many properties with the more traditional distributions used with failure time data: the log-logistic distribution is a special case of the Burr(XII) distribution; the Burr(XII) distribution can be a good approximation to the Weibull distribution; and the Weibull distribution is the limiting distribution of the Burr(XII) distribution [21]. While the Burr(XII) distribution shares many of the advantages of these other distributions, it is more flexible.

While parametric distributions offer convenience in modeling, it is often difficult to choose and/or justify an appropriate parametric form for lifetime data. This may be particularly difficult with complex systems with multimodal survival times or when modeling failure time data from products manufactured using emerging technologies with unfamiliar mechanisms [3]. Such multimodal data can be more flexibly modeled using a mixture model rather than a single parametric distribution. The mixture model can be written as a summation of a coefficient and the probability of choosing that coefficient (kernel) [13]. The Bayesian mixture model is obtained by allowing these coefficients to be random. The nonparametric Bayesian mixture model gains even greater flexibility by considering the space of parameters to be infinite.

Nonparametric Bayesian mixture models can be characterized by the choice of mixture distribution. One of the most common choices for the distribution of coefficients is the Dirichlet process, which yields the Dirichlet Process mixture model(DPMM) [3]. The DPMM can be further tailored to the application through the choice of distributional form for the kernel. A DPMM with a normal kernel is often used ([5], [14]); however, it is unsuitable for survival data given the positive support required for failure times. A DPMM with a Weibull kernel has been studied for survival analysis by Kottas [10] and a DPMM with a log-normal kernel was developed for reliability applications by Cheng and Yuan [3]. This work proposes the use of a DPMM with a Burr(XII) distribution for the kernel which also accommodates the positive support required for failure time data.

In the next section we describe the general framework for DPMMs. Section 3 describes the use of the Burr(XII) kernel in the DPMMs. The Gibbs sampling estimation algorithm for the DPMM with Burr(XII) kernel will be described in Section 4. Section 5 presents simulation results and an application of the DPMM with Burr(XII) kernel to failure time data. Finally, the obtained results are discussed in section 6.

2. Dirichlet process mixture models

2.1. Dirichlet Process

Let $(\Theta, \mathcal{A}, G)$ be a probability space, let G_0 be a distribution over Θ and υ be a positive real number, then for any finite measurable partition $A_1, A_2, ..., A_k$ of Θ, G will be called a Dirichlet process (DP) with base distribution G_0 and concentration parameter υ, $G \sim DP(\upsilon, G_0)$, if for every vector $(G(A_1), G(A_2), ..., G(A_k)), k \in N$, we have

$$\Big(G(A_1), G(A_2), ..., G(A_k)\Big) \sim Dir(\upsilon G_0(A_1), ..., \upsilon G_0(A_k)).$$

An important constructive definition of a DP is given by Sethuraman [19] and it is based on the discrete nature of the process. In this construction, θ_h's are i.i.d from the centering measure G_0 and each weight w_h is defined as a fraction of $1 - \sum_{l<h} w_h$. Let $w_h = \nu_h \prod_{l<h}(1 - \nu_l)$ where $\nu_h \overset{i.i.d}{\sim} Be(1, \upsilon)$, $\theta_h \overset{i.i.d}{\sim} G_0$, and also assume υ and θ_h are independent, then

$$G(.) = \sum_{h=1}^{\infty} w_h \delta_{\theta_h}(.),$$

define a DP(υ, G_0) random probability measure. This representation of a DP is known as "Stick-Breaking" representation.

2.2. Dirichlet Process Mixture Models

A parametric mixture model for function $f(x)$ can be written as

$$f(x|\pi_1, ..., \pi_M, \theta_1, ..., \theta_M) = \sum_{j=1}^{M} k(x|\theta_j)\pi_j; \;\; \theta_j \in \Theta, \;\; j = 1, ..., M, \qquad (1)$$

where $k(x|\theta)$ is a parametric kernel and π_j's are the values that for all j; $\pi_j > 0$ and $\sum_{j=1}^{M} \pi_j = 1$.

By the Bayesian approach we can consider the weights(π_j's) as the probability of choosing θ_j's and then by putting a nonparametric Bayesian prior like Dirichlet process on the wights, we have a mixture model named by

Dirichlet process mixture models(DPMMs). A DPMM can be illustrated as bellow:

$$f_G(x) = \int_{\Theta} k(x|\theta) dG(\theta)$$

where $\theta \sim G$ and $G \sim DP(\upsilon, G_0)$.

By a hierarchical representation, a DPMM can be presented as the following form, which is more popular of these type models,

$$\begin{array}{rcl} x_i|\theta_i & \sim & k(x_i|\theta_i) \quad i = 1, ..., n \\ \theta_i|G & \sim & G(\theta) \\ G & \sim & DP(\upsilon, G_0). \end{array} \qquad (2)$$

Through this hierarchical representation, a marginalization step is often used in implementation of the model. Blackwell and McQueen [1] showed that by integrating out of G, the joint distribution of $\theta_1, ..., \theta_n$ may be written into a product of conditional distribution of the following form

$$\theta_i|\theta_1, ..., \theta_{i-1}, G_{0,\upsilon} \sim \frac{1}{i-1+\upsilon} \sum_{j=1}^{i-1} \delta_{\theta_j} + \frac{\upsilon}{i-1+\upsilon} G_0, \qquad i = 1, ..., n \qquad (3)$$

where $\delta(\theta)$ denotes the distribution concentrated at point θ, and as we see, this formula has the form a mixture model.

2.3. Model implementation

Our aim here is to formulate how to sample from the DPMMs by Gibbs sampling. According to [10], Gibbs sampling for drawing sample from $[(\theta_1, ..., \theta_n), \upsilon, ...|t]$ is based on the following full conditionals (we use the bracket([]) to show the conditional and marginal distributions):

$$\begin{array}{l} (1)\ [(\theta_i)|(\theta_{-i}, z_{-i}), \upsilon, ..., t], \quad for\ i = 1, ..., n \\ (2)\ [(\theta^*_j)|z, n^*, \upsilon, ..., t], \ \ for\ j = 1, ..., n^* \\ (3)\ [\upsilon|\{(\theta^*_j), j = 1, ..., n^*\}, n^*, t], [...|\{(\theta^*_j), j = 1, ..., n^*\}, n^*, t]. \end{array} \qquad (4)$$

where t is the failure time data. Here, θ_i's are the parameters of kernel in DPMMs that we want to analyze them. For example if the kernel of a DPMM is two-parameter Burr(XII) distribution with (c,k) parameters, as it will be discussed by details in the next section, then $\theta = (\theta_1, \theta_2) = (c, k)$. The model (3) and discreteness property of DP, exhibit a clustering kind of effect.

We present n^* as the number of the clusters between θ's that denote by θ^*'s. The vector of indicators $z = (z_1, ..., z_n)$ indicates the clustering configuration such that, $z_i = j$ when $\theta_i = \theta_j^*$. Also, the θ_{-i} that used in (4), it will be defined by $\theta_{-i} = (\theta_1, \theta_2, ..., \theta_{i-1}, \theta_{i+1}, ..., \theta_d)$.

By combining formula (3) with the likelihood of data[14], the posterior distribution can be presented as

$$f(\theta_i|\theta_{-i}, t_i) = b\upsilon f(t_i;\theta_i)G_0(\theta) + b\sum_{j=1, j\neq i}^{n} f(t_i;\theta_i)\delta_{\theta_j^*}(\theta_j) \tag{5}$$

where

$$b = (q_0 + \sum_{j=1, j\neq i}^{n} f(t_i;\theta_i))^{-1}$$

and

$$q_0 = \upsilon \int f(t_i|\theta)G_0(\theta)d\theta.$$

Note that by [3], we have the posterior of θ_i as

$$h(\theta_i|t_i) = \frac{f(t_i;\theta_i)G_0(\theta_i)}{\int f(t_i|\theta)G_0(\theta)d\theta},$$

then we can rewrite equation (5) as follows

$$f(\theta_i|\theta_{-i}, t_i) = bq_0h(\theta_i|t_i) + b\sum_{j=1, j\neq i}^{n} f(t_i;\theta_i)\delta_{\theta_j^*}(\theta_j). \tag{6}$$

Regarding the form of the model (5) or (6), we can draw from the distribution by Gibbs sampling such that $f(\theta_i|\theta_{-i}, t_i)$ takes the one of the previous values with probability of $bf(t_i;\theta_i)$ and takes a new θ from $h(\theta_i|t_i)$ with probability of bq_0.

2.4. Simulation-based parameter estimation algorithm

The ability of MCMC methods to compute integrals caused that statisticians widely use it in the implementation of the Bayesian approach. An MCMC method constructs a Markov chain which is a stochastic process that the current value generated from step will always depend on the value in the previous step. This algorithm is a preferable method to sample from conditional distributions in Bayesian approach. We may summarize this method as follows

```
$Algorithm : Gibbs\ sampler$
    1. $Initialize\ \ \theta^{(0)} \sim f(\theta)$
    2. $For\ iteration\ \ i = 1, 2, ...\ do$
        $\theta_1^{(i)} \sim f(\theta_1|\theta_2^{(i-1)}, \theta_3^{(i-1)}, ..., \theta_d^{(i-1)}, D),$
        $\vdots$
        $\theta_d^{(i)} \sim f(\theta_d|\theta_1^{(i)}, \theta_2^{(i)}, ..., \theta_{d-1}^{(i)}, D),$
```

where $\theta_1, ..., \theta_d$ represent the parameters in the model and D is the observation set. The values of iteration i would be sampled from the last version of the other values.

In our model with Dirichlet process prior for the infinite parameters, Gibbs sampling is more appropriate simulation method.

3. Dirichlet process Burr(XII) mixture model

The Burr distribution is the renowned distribution in the probability. Burr[2] has propose a numbers of forms of c introduced a system of distributions by considering distribution functions F(x) satisfying the differential equation

$$\frac{dF}{dx} = A(F)g(x).$$

In the special case if $A(F) = F(1 - F)$ then the solution of Burr's equation is

$$F(x) = \frac{1}{1 + exp(-G(x))}$$

where $G(x) = \int_{-\infty}^{x} g(t)dt$. Twelve distributions that named by Burr distributions are particular form of this solution. Here we use the type XII of Burr distributions that is named by Burr(XII) distribution [18].

Let $k_B(t|c, k)$ and $K_B(t|c, k)$ denote the p.d.f and c.d.f of the Burr(XII) distribution respectively, which defined by

$$k_B(t|c, k) = ck\frac{t^{c-1}}{(1 + t^c)^{k+1}} \qquad c, k > 0,\ t > 0$$

and

$$K_B(t|c,k) = 1 - (1+t^c)^{-k}$$

which the parameter space is $\Theta = \{(c,k); 0 < c < \infty, 0 < k < \infty\}$. Suppose base distribution G_0 as the prior guess for the joint distribution of c and k. When we choose with Burr(XII) kernel, the G_0 that yields closed form expression for $\int k_B(.|c,k)G_0(dc,dk)$ is not available, but we choose a multiple distributions of uniform$(0,\phi)$ and exponential with parameter γ for the G_0,

$$G_0(c,k|\phi,\gamma) = Unif(c|0,\phi) \times Exp(k|\gamma). \tag{7}$$

This choice achieves aforementioned goals. Suppose γ and ϕ are random and have $Pareto(a_\phi, b_\phi)$ and $IGamma(a_\gamma, b_\gamma)$ distributions respectively (Here $IGamma(.|a,b)$ denotes the inverse-gamma distribution with mean $\frac{b}{a-1}$ provided $a > 1$).
We set the $a_\phi = a_\gamma = d$ and then we choose d=2, since this value makes the variance of Pareto distribution infinite that cover all values in R. We determine the values of b_ϕ and b_γ by the data that the details are presented in the appendix A.

Finally, by considering Burr(XII) as the kernel of the DPMM we introduce the hierarchical model of Dirichlet process Burr(XII) mixture model (DPBMM) as the following form

$$\begin{aligned} t_i|c_i,k_i &\sim Burr(t_i|c_i,k_i), \quad i=1,...,n, \\ (c_i,k_i)|G &\sim G, \\ G &\sim DP(\nu, G_0), \\ \nu,\gamma,\phi &\sim Gamma(a_\nu,b_\nu) \times IGamma(a_\gamma,b_\gamma) \times Pareto(a_\phi,b_\phi). \end{aligned} \tag{8}$$

4. Modeling based on DPMMs with Burr(XII) kernel

Now we describe how to estimate parameters with the DPBMMs. Data contains rightly censored failure times that are indicated by a characteristic function:

$$\delta_i = \{\begin{matrix} 0 & \text{if } t_i \text{ is an uncensored failure time} \\ 1 & \text{if } t_i \text{ is a censored failure time} \end{matrix}.$$

As is denoted before, n^* is the number of distinct cluster of parameters (c_i, k_i) in which $i = 1,...,n$ and let $(c_j^*, k_j^*),\ j = 1,...,n^*$ denote the cluster locations.

Then the clustering indicator z_i indicates the cluster which (c_i, k_i) belongs to $z_i = j$ if $(c_i, k_i) = (c_j^*, k_j^*)$. Let n_j^* be the number of members of cluster j.

A simulation algorithm to implement Gibbs sampling is drawing a sample for one parameter from its conditional posterior density conditioned on the last estimates of all the other parameters[12]. Kottas [10] is derived the conditional posterior densities for the Dirichlet process Weibull mixture model. So by replacing the Weibull with Burr(XII) as the kernel, we derive the conditional posterior densities in DPBMMs.

At first we draw a sample of (c_i, k_i) and update the z_i for each failure time t_i. The sample can be a new value from G_0 or be equal to an existing values. Let $n^{*(i)}$ be the number of clusters when (c_i, k_i) is removed from the sample and let $n_j^{*(i)}$ be the number of elements in cluster j after removing.

We denote the quantities related to an observed data by letter "o" and censored data with letter "c". If t_{io} is an observed (uncensored) failure time, then by [10] and equation (6) the conditional posterior density of (c_i, k_i) is the mixed distribution

$$f(c_i, k_i|(c_i, k_i); i \neq i', \nu, \gamma, \phi, t_{io}) = \frac{q_0^o h^o(c_i, k_i|\phi, \gamma, t_{io}) + \sum_{j=1}^{n^{*(i)}} n_j^{*(i)} q_j^o \delta_{c_j^*, k_j^*}}{q_0^o + \sum_{j=1}^{n^{*(i)}} n_j^{*(i)} q_j^o},$$

where $q_j^o = k_B(t_{io}|c_j^*, k_j^*)$ and q_0^o in is given by

$$\begin{aligned} q_0^o &= \nu \int_0^\phi \int_0^\infty k(t_{io}|c, k) G_0(c, k) dcdk \\ &= \frac{\nu}{\phi} \int_0^\phi \frac{ct_{io}^{c-1}}{(1 + t_{io}^c)} (\int_0^\infty \frac{ke^{-\frac{k}{\gamma}}}{(1 + t_{io}^c)^k} dk) dc \\ &= \frac{\nu}{\phi} \int_0^\phi \frac{ct_{io}^{c-1}}{(1 + t_{io}^c)(ln(1 + t_{io}^c) + \frac{1}{\gamma})} dc \end{aligned}$$

in which the last integration can be computed numerically and

$$h^o(c_i, k_i|\gamma, \phi, t_{io}) \propto k_B(t_i|c_i, k_i) G_0(c_i, k_i) \propto [c_i|\gamma, \phi, t_{io}][k_i|c_i, \gamma, \phi, t_{io}]$$

where

$$[c_i|\gamma, \phi, t_{io}] \propto c_i t_i^{c_i - 1} I_{(0,\phi)}(c_i) \qquad i = 1, ..., n$$

and

$$[k_i|c_i, \gamma, \phi, t_{io}] \propto Gamma(.|2, \frac{1}{[\frac{1}{\gamma} + ln(1 + t_i^{c_i})]}).$$

To sample from $h^o(c_i, k_i|\phi, \gamma, t_{io})$, we first draw a value from $[c_i|\gamma, \phi, t_{io}]$ using slice sampling [22]. Then draw a value for k_i from the Gamma distribution with parameters 2 and $\frac{1}{[\frac{1}{\gamma}+ln(1+t_{io}^{c_i})]}$.

For rightly censored data (t_{ic}) the conditional posterior density of (c_i, k_i) is

$$f(c_i, k_i|(c_i, k_i); i \neq i', \nu, \gamma, \phi, t_{ic}) = \frac{q_0^c h^c(c_i, k_i|\phi, \gamma, t_{ic}) + \sum_{j=1}^{n^{*(i)}} n_j^{*(i)} q_j^c \delta_{c_j^*, k_j^*}}{q_0^c + \sum_{j=1}^{n^{*(i)}} n_j^{*(i)} q_j^c}$$

where $q_j^c = 1 - K_B(t_{ic}|c_j^*, k_j^*)$

$$\begin{aligned}
q_0^c &= \nu \int_0^\phi \int_0^\infty (1 - K(t_{ic}|c, k)) G_0(c, k) dc dk \\
&= \frac{\nu}{\phi\gamma} \int_0^\phi \int_0^\infty \frac{e^{-\frac{k}{\gamma}}}{(1 + t_{ic}^c)^k} dk dc \\
&= \frac{\nu}{\phi\gamma} \int_0^\phi (\frac{1}{\gamma} + ln(1 + t_{ic}^c)) dc
\end{aligned}$$

which can again be computed using numerical integration and

$$\begin{aligned}
h^c(c_i, k_i|\gamma, \phi, t_{ic}) &\propto (1 - K_B(t_i|c_i, k_i)) G_0(c_i, k_i) \\
&\propto [c_i|\gamma, \phi, t_{ic}][k_i|c_i, \gamma, \phi, t_{ic}] \\
&= \frac{I_{(0,\phi)}(c_i)}{\phi\gamma} \frac{1}{\frac{1}{\gamma} + ln(1 + t_{ic}^{c_i})} k_i e^{-k_i(\frac{1}{\frac{1}{\gamma}+ln(1+t_{ic}^{c_i})})} \\
&= \frac{I_{(0,\phi)}(c_i)}{\phi\gamma} \frac{1}{\frac{1}{\gamma} + ln(1 + t_{ic}^{c_i})} \times Gamma(k_i|2, \frac{1}{\frac{1}{\gamma} + ln(1 + t_{ic}^{c_i})}).
\end{aligned}$$

Using the slice sampling, we can draw a sample from $h^c(c_i, k_i|\phi, \gamma, t_{ic})$.

Now we can the update (c_i, k_i) for $i = 1, ..., n$, for observed and censored data. Then in a general form, $(c_j^*, k_j^*)'s$ can be updated on ϕ, γ and t as the following

$$\begin{aligned}
f(c_j^*, k_j^*|\phi, \gamma, t, n^*) &\propto G_0(c_j^*, k_j^*|\gamma, \phi) \prod_{\{io:s_{io}=j\}} k_B(t_{io}|c_j^*, k_j^*) \prod_{\{ic:s_{ic}=j\}} (1 - K_B(t_{io}|c_j^*, k_j^*)) \\
&\propto [c_j^*|\gamma, \phi, t_{io}][k_j^*|c_j^*, \gamma, \phi, t_{ic}] \prod_{\{ic:s_{ic}=j\}} \frac{1}{(1 + t_{ic}^{c_j^*})^{k_j^*}} \\
&\propto c_j^{*n_j^o} I_{(0,\phi)}(c_j^*) \prod_{\{io:s_{io}=j\}} \frac{t_{io}^{c_j^*-1}}{1 + t_{io}^{c_j^*}} \times Gamma(n_j^o + 1, B^*) \qquad (9)
\end{aligned}$$

where $B^* = \sum_{\{io:s_{io}=j\}}(\frac{1}{\gamma} + ln(1+t_{io}^{c_j^*})) + \sum_{\{ic:s_{ic}=j\}} ln(1+t_{ic}^{c_j^*})$ and n_j^o is the number of observed data which located in cluster j. To generate a sample from equation (9), we just need to draw from the first part. Sampling from the gamma distribution is simple. To sample from

$$[c_j^*|\phi,\gamma,t] \propto c_j^{*n_j^o} I_{(o,\phi)} c_j^* \prod_{\{io:s_{io}=j\}} \frac{t_{io}^{c_j^*-1}}{1+t_{io}^{c_j^*}}$$

$$\propto c_j^{*n_j^o} I_{(o,\phi)} c_j^* \prod_{\{io:s_{io}=j\}} (t_{io}^{c_j^*-1}) \frac{1}{1+t_{io}^{c_j^*}}$$

we introduce auxiliary variables $W = \{w_{io}; \{io : s_{io} = j\}\}$ such that

$$[c_j^*, W|\phi, t_{io}] = c_j^{*n_j^o} 1_{(c_j^* \leq \phi)} \prod_{\{io:s_{io}=j\}} 1_{(w_{io} \leq \frac{t_{io}^{c_j^*-1}}{1+t_{io}^{c_j^*}})}.$$

By marginalization over the auxiliary variables, we get the $[c_j^*|\phi, t_{io}]$ for $j = 1, ..., n^*$. Moreover, w_{io}'s are uniform on $(0, \frac{t_{io}^{c_j^*-1}}{1+t_{io}^{c_j^*}})$. Therefore we have

$$[c_j^*|\phi, t] = c_j^{*n_j^o} I_{(B,\phi)}(c_j^*)$$

where $B = max\{0, \frac{ln(w_{io})}{1+t_{io}}\}$ and now drawing from $[c_j^*|\phi, t]$ is straightforward.

Afterwards, using the methods in [5], we update ϕ, γ and ν. If we take u as a latent variable such that

$$[u|\nu, t] = Beta(\nu+1, n)$$

then

$$[\nu|u, n^*, t] = pGamma(a_\nu + n^*, b_\nu - log(u)) + (1-p)Gamma(a_\nu + n^* - 1, b_\nu - log(u))$$

where $p = \frac{a_\nu + n^* - 1}{n(b_\nu - log(u)) + a_\nu + n^* - 1}$.

And finally to update ϕ we have

$$[\phi|c^*, k^*] = [\phi][c^*, k^*|\phi] = \frac{2b_\phi^2}{\phi^3} I_{(b_\phi,\infty)}(\phi) \prod_{j=1}^{n^*} \frac{1}{\phi} I_{(0,\phi)}(c^*) = \frac{2b_\phi^2}{\phi^{n^*+3}} I_{(b^*,\infty)}(\phi)$$

where $b^* = max\{b_\phi, max_{1\leq j\leq n^*} c_j^*\}$. So

$$[\phi|c^*, k^*] = Pareto(\phi|2 + n^*, b^*)$$

and by repeating this technique , we can update γ

$$[\gamma|c^*, k^*] = [\gamma]\prod_{j=1}^{n^*}[k_j^*|\gamma] = IGamma(n^* + 2, b_\gamma + \sum_{j=1}^{n^*} k_j^*).$$

Now, we can calculate all conditional distributions on the equation (3).

5. Data illustration and case study

We are going to work with two data sets to show the validity of the proposed model. First data set is a simulated data that we generate it from two mixture of Burr(XII) distributions. Another data set is a real data of breast cancer that was used in [16]. For each of the data sets we fit DPMM with different kernel distributions. Exactly Weibull, Log-normal and Burr(XII) distributions and then compare these models based on some indexes and also by their plots.

To see how well the models capture the true distribution, we use some goodness-of-fit(gof) metrics [3]. Three of these metrics included,

$$\frac{1}{n}\sum_{i=1}^{n}|\frac{Q_T(t_i) - \tilde{Q}_T(t_i)}{Q_T(t_i)}|,\ \frac{1}{n}\sum_{i=1}^{n}|Q_T(t_i) - \tilde{Q}_T(t_i)|\ and\ \frac{1}{n}\sum_{i=1}^{n}(Q_T(t_i) - \tilde{Q}_T(t_i))^2,$$

where $Q_T(t)$ can be any function of data. We can calculate these metrics for the probability density, cumulative probability and hazard rate (HR) functions by replacing them with Q in the proposed gof functions. To show the adaptivity of the model in survival applications, we apply it for comparison of two treatments.

5.1. Simulated data

Based on a finite mixture of Burr(XII) distributions, we simulate a complete life time data of the size n=200. The finite mixture model is the mixture of two Burr(XII) distributions, exactly

$$pBurr(5, 1) + (1 - p)Burr(2, 6)$$

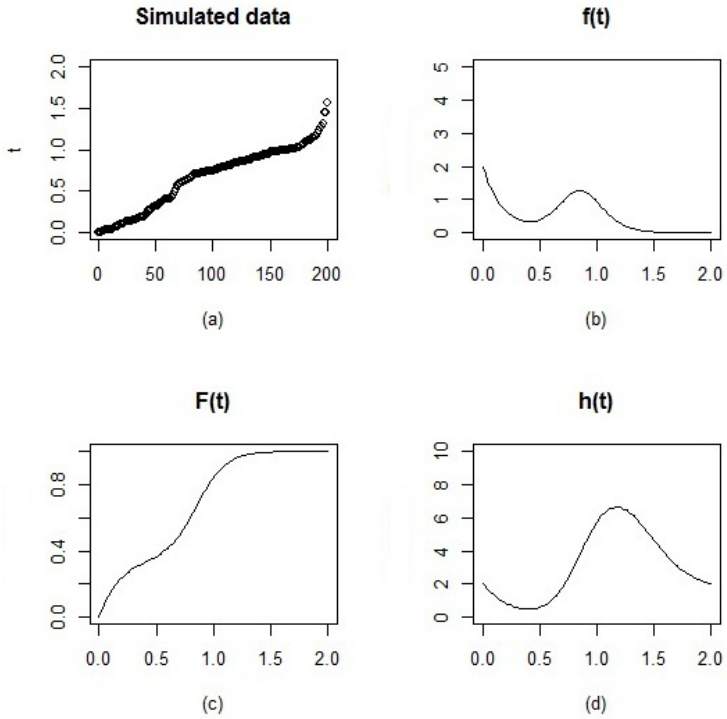

Figure 1: Simulated data from mixture of Burr(XII) distributions (a)Empirical plotting position;(b)density function;(c) distribution function; (d) hazard function

where Burr(c,k) denotes the Burr(XII) distribution with the scale parameter c and the shape parameter k. We use p=0.2 to weight the Burr(XII) distributions.

As figure 1 shows simulated data has a bimodal probability density function that it is common in survival data. Also plots (c) and (d) of figure 1 illustrate cumulative and hazard rate functions of simulated data respectively. Table 1 illustrates indexes that we introduced in the beginning of this section and also we calculated values for DPMMs with different kernels. In this table calculated indexes are for probability density functions of the models.

The values in the table 1 signify that DPBMM has a better fitness for the simulated data, also cumulative probability and hazard rate functions, they both proved preference of the DPBMM for this simulated data set.

indexes	Weibull	Log-Noraml	Burr(XII)
$\frac{1}{n}\sum_{i=1}^{n} \lvert \frac{f_T(t_i)-\tilde{f}_T(t_i)}{f_T(t_i)} \rvert$	1.7544	0.6346	0.3041
$\frac{1}{n}\sum_{i=1}^{n} \lvert f_T(t_i) - \tilde{f}_T(t_i) \rvert$	157.1424	110.4019	26.8968
$\frac{1}{n}\sum_{i=1}^{n} (f_T(t_i) - \tilde{f}_T(t_i))^2$	0.7898	0.4587	0.0296

Table 1: Computed metrics for DPMMs with different kernels

Figure 2 shows the posterior mean of f_T, F_T and h_T respectively, using DPBM, DPWM and DPLNM models with the $\upsilon \sim Gamma(1, 0.001)$. By [10] we know that with different choices of υ's parameters, there is a learning on n^*, so we select a approximately non-informative prior for υ.In this figure density function of DPBMM fits the exact curve of simulated data very well.Also the cumulative distribution and hazard functions of DPBMM has a good fitness than other DPMMs.

5.2. Case study

To show the preference of DPBM model rather than DPLNM and DPWM models we apply these models for a real survival time data and compare them. We use a breast cancer data set that the data is obtained from [16]. The data is the incidence and death dates of about 1000 breast cancer patients in Gaza strip within a period of 5 years starting from the being of 2009 to the end of 2013. The survival times for those patients that remained valid is 242 patients.

In figure 3 we plot the histogram of data in scale of 1000 days and the probability density of DPBM, DPLNM and DPWM models. The dashed line is DPBMM, the dotted line is DPLNMM and another one is DPWMM. The DPBMM captures the shape of the histogram very well Figure 4 shows the reliability of data with step line and DPMMs with different kernels. As we see, DPBM model is the closest line to the reliability of data.

5.3. Comparison of two treatments

Here we consider the leukemia patients data on remission times, in weeks that taken from [12]. Our purpose of choosing this data is the comparison of survival functions across their populations with DPBMM. The data set involve two treatments, A and B, each of them included 20 patients. The data set is presented in table 2.
Lawless [12] tested the equality of the hazard functions of two treatments A

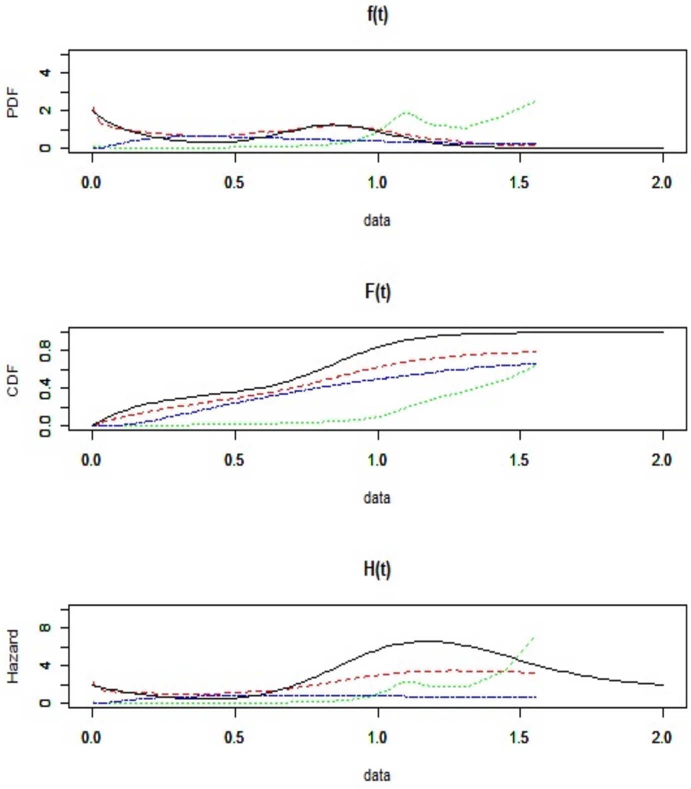

Figure 2: (a) the density, (b) the cumulative distribution function and (c) the hazard function of simulated data with different kernels, solid line is true density, dashed line DPBMM, dotted line, DPWMM and dash-dotted line DPLNMM.

and B based on classical test procedures and concluded that " there is no evidence of a difference in distributions". In the another work Damien and walker [4] compared these two treatments with a Bayesian nonparametric approach. Their approach did not assume any functional relationship between the distribution functions associated with the treatments and yielded a result they regarded " far from conclusive of no difference". But Kottas [10] employ the DPWMM for each of the underlying population distributions forcing no particular relation between the corresponding distribution distribution functions. He plots the 95% confidence interval for the mean of posterior and compare the difference between the treatments by posterior means of density functions and hazard functions.

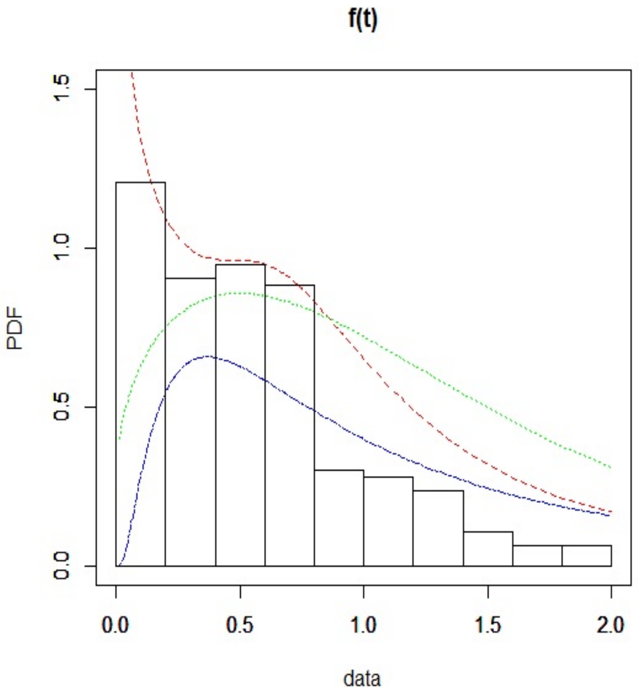

Figure 3: Histogram of Ghaza data set and the estimated density function with DPBMM (dashed line), DPLNMM(dotted line) and DPLNMM(dash-dotted line)

Here we apply the DPWMM to the two data set and calculated the mean of posterior for density and hazard functions. Figure (5) shows the calculated quantities for the two treatment. As we see the difference between the treatments is obviously.

6. Conclusion

Here in this article we developed a mixture model for survival inference based on Burr XII as the kernel. Since we put priors on the space of parameter and the number of components is infinite, so we have a Bayesian nonparametric model. In this article we have applied our model i.e. Dirichlet process Burr(XII) mixture model for modeling of survival populations. At first we applied the proposed model to the simulated data and then used it for modeling a real survival data. For both of them the DPBMM was fitted

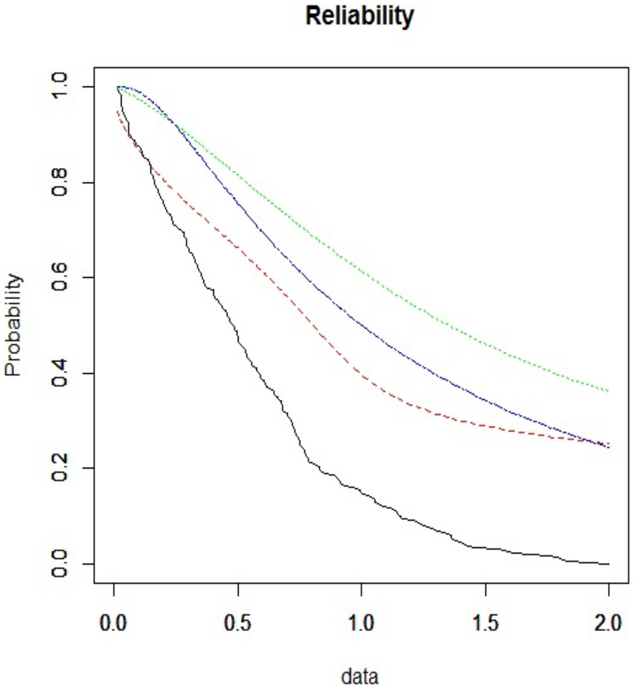

Figure 4: Empirical reliability curve of Ghaza data set and the estimated reliability function with DPBMM (dashed line), DPLNMM(dotted line) and DPLNMM(dash-dotted line)

much more better than other models. Also for the data of two treatments the DPBMM showed the difference between two treatments. For a further work we can propose using another kernel and introducing a new model.

Appendix A. Choosing b_γ and b_ϕ in the equation (8)

Since $\gamma \sim IG(2, b_\gamma)$ and $\phi \sim Pareto$ and by (7) we have

$$c|\phi \sim Unif(0, \phi)$$

$$k|\gamma \sim Exp(\gamma)$$

so we can calculate the marginal distributions of c and k as bellow

$$[c] = \frac{2b_\phi^2}{3} \frac{2}{(max\{c, b_\phi\})^3} \qquad [k] = \frac{2b_\gamma^2}{(k + b_\gamma)^3}. \tag{A.1}$$

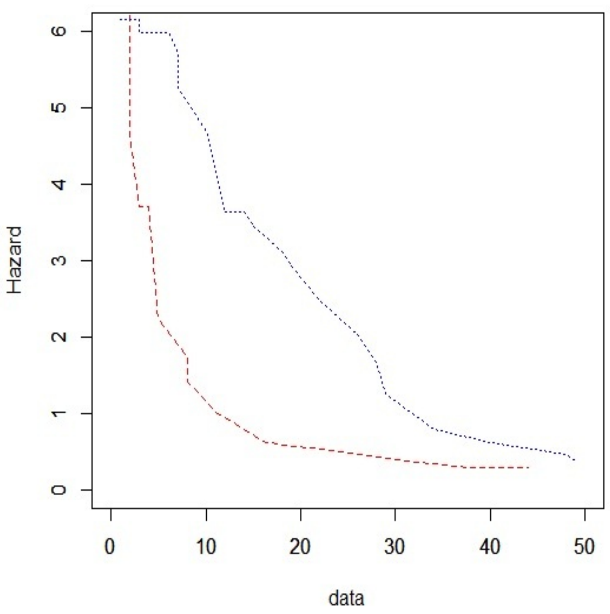

Figure 5: Hazard rates for treatment A (dashed line) and treatment B (dotted line) with the mean of posterior in DPBMM

Then the medians of c and k is obtained as the following

$$m_c = \frac{3}{4} b_\phi \qquad m_k = (\sqrt{2} - 1) b_\gamma .$$

If we follow the methods that is given in [10], we use the median and interquartile range of Burr(XII) distribution for obtaining a prior guess for (c,k) which denoted by $\tilde{c}$ and $\tilde{k}$.
At first we know

$$F_B^{-1}(p) = ((1-p)^{-1/k} - 1)^{1/c}$$

and we have the quartiles by replacing $p = \frac{1}{4}, \frac{2}{4}, \frac{3}{4}$.

$$Q_1 = ((4/3)^{1/k} - 1)^{1/c} \quad Q_2 = ((2)^{1/k} - 1)^{1/c} \quad Q_3 = ((2)^{2/k} - 1)^{1/c}$$

By putting equal the empirical and theoretical quartiles, the values of $\tilde{c}$ and

Treatment A		Treatment B	
1	3	1	1
3	6	2	2
7	7	3	4
10	12	5	8
14	15	8	9
18	19	11	12
22	26	14	16
28+	29	18	21
34	40	27+	31
48+	49+	38+	44

Table 2: Leukemia patients data for two treatments A and B

$\tilde{k}$ are obtained and then to specify b_γ and b_ϕ, set medians of [c] and [k] equal to $\tilde{c}$ and $\tilde{k}$ respectively.

References

[1] Blackwell D, MacQueen J(1973) ”Ferguson distributions via Polya urn schemes.” The annals of statistics 353-355.

[2] Burr IW (1942) ”Cumulative frequency functions.” The Annals of mathematical statistics 13(2): 215-232.

[3] Cheng N, Yuan T(2013) ”Nonparametric Bayesian lifetime data analysis using Dirichlet process lognormal mixture model.” Naval Research Logistics (NRL) 60(3): 208-221.

[4] Damien P, Walker S(2002) ”A Bayesian Non-parametric Comparison of Two Treatments.” Scandinavian journal of statistics 29(1): 51-56.

[5] Escobar MD, West M(1995) ”Bayesian density estimation and inference using mixtures.” Journal of the american statistical association 90(430): 577-588.

[6] Ferguson TS (1983)”Bayesian density estimation by mixtures of normal distributions.” Recent advances in statistics 1983(24): 287-302.

[7] Gelman A, et al.(2014) Bayesian data analysis. Vol. 2. Boca Raton, FL, USA: Chapman & Hall/CRC.

[8] Giudici P, et al. (2009) ”Bayesian modeling using WinBUGS.” .

[9] Ghosh JK, Ramamoorthi RV (2003) ”Bayesian nonparametrics.” Springer Series in Statistics. Springer-Verlag, New York.

[10] Kottas A(2006) ”Nonparametric Bayesian survival analysis using mixtures of Weibull distributions.” Journal of Statistical Planning and Inference 136(3): 578-596.

[11] Lanjoni B R, et al. (2016) ”Extended Burr XII regression models: theory and applications.” Journal of Agricultural, Biological, and Environmental Statistics 21(1): 203-224.

[12] Lawless JF (2011) Statistical models and methods for lifetime data. Vol. 362. John Wiley & Sons.

[13] McLachlan et al.(2004) Finite mixture models. John Wiley & Sons.

[14] Muller P,et al.(2004) ”Nonparametric Bayesian data analysis.” Statistical science 95-110.

[15] Nelson W (1982) ”Applied Life Data Analysis, New York: JohnWiley.” NelsonApplied Life Data Analysis.

[16] Okasha MK, Matter MY (2008)”On the three-parameter burr type XII distribution and its application to heavy tail lifetime data.” Journal of advances in mathematics 10(4).

[17] Rao et al.(2015) ”Burr-XII Distribution Parametric Estimation and Estimation of Reliability of Multicomponent Stress-Strength.” Communications in Statistics-Theory and Methods 44(23): 4953-4961.

[18] Rodriguez RN (1977)”A guide to the Burr type XII distributions.” Biometrika 64(1): 129-134.

[19] Sethuraman J (1994)”A constructive definition of Dirichlet priors.” Statistica sinica 639-650.

[20] Singpurwalla ND(2006) Reliability and risk: a Bayesian perspective. John Wiley & Sons.

[21] Tadikamalla PR(1980) "A look at the Burr and related distributions." International Statistical Review/Revue Internationale de Statistique 337-344.

[22] Walker SG (2007)"Sampling the Dirichlet mixture model with slices." Communications in StatisticsSimulation and Computation 36(1): 45-54.

[23] Zimmer WJ, et al.(1998)"The Burr XII distribution in reliability analysis." Journal of Quality Technology 30(4): 386.