Sahand Communications in Mathematical Analysis (SCMA) Vol. ··· No. ··· (20??), ···-···
http://scma.maragheh.ac.ir
DOI: ...

Analysis of Survival Data Using Dirichlet Process Mixture Models with Kernels of Exponentiated Weibull Distribution

Ali Salih Ali Al-Bajalan[1], Soleiman Khazaei[2*] and Bahram Haji Joudaki[3]

Abstract. The Dirichlet process mixture model, a prominent member of the Bayesian nonparametric family, stands out as one of the most favored methods for automatically inferring the number of clusters in complex datasets. This article introduces a novel approach to this model by incorporating an Exponentiated Weibull kernel, which provides additional flexibility in modeling diverse survival patterns. By carefully establishing the prior distribution of parameters, MCMC algorithms are employed to efficiently sample from the joint posterior distribution. Evaluation of the proposed model's performance includes estimating density, survival and hazard rate functions, as well as clustering survival data, through detailed analysis of both real-world and simulated datasets. The results highlight the robust and desirable performance of the proposed model, demonstrating its potential as a reliable tool for statistical analysis and survival modeling in biomedical and other applied research domains.

1. Introduction

One of the most important concerns of a statistician is choosing a model to perform statistical inference. The issue of choosing the model is significant but challenging from both theoretical and practical perspectives. Most statisticians believe that the major problems in statistical inference can be considered as issues related to statistical modeling. In fact, translating the objective problem into the statistical model is often the most critical part of a statistical analysis. In statistical science, to

2020 *Mathematics Subject Classification.* 62F10, 62N01.
Key words and phrases. Bayesian nonparametric, Density estimation, Time censoring, Clustering, MCMC.
Received: 2025-09-30, Accepted: 2025-12-01.
* Corresponding author.

have easy access to a set of statistical inferences, we often opt for a family of probabilistic models, which consists of models with an indexed set of parameters of a finite number. If the distribution of model parameter's selection preferences are given, then the posterior distribution of the parameters can be calculated.

Moreover, any inference regarding the parameters is based on the posterior distribution of the parameters, a method referred to as the parametric Bayesian method. Indeed, the capability to address complex data distributions, such as multimodal data, using simple probability models, often makes working with these models easier. However, there may be significant disparities between the results obtained from these models and the actual results. To accommodate complex data distributions, it is necessary to employ more sophisticated statistical models.

In general, the complexity of a statistical model is relevant to the number of model parameters. To achieve a more complex model and consequently obtain better results, it is reasonable to choose a solution for performing statistical inference that allows for the utilization of probability models. In such a scenario, a richer and broader class of probability models becomes available. For Bayesian approaches to be applicable to such probabilistic models, they should be implementable for models with an infinite number of parameters. Therefore, the selection of a prior distribution model becomes crucial. As these parameters are infinite, they are of the type of random processes and such priors are termed non-parametric priors. Utilizing such an approach for statistical inference leads to the adoption of a non-parametric Bayesian approach.

The most significant predecessor of the nonparametric Bayesian Dirichlet process (DP) was introduced by [4]. Ferguson outlined and proved several essential features of the DP prior distribution. Subsequently, [6, 13, 20] utilized the DP prior distribution to model the survival distribution function. Despite the appeal of the DP prior distribution, it remains a discrete process. Therefore, approximating continuous functions such as density, survival and hazard rate with it will inevitably encounter challenges. By augmenting the DP prior distribution with certain continuous distributions, this issue can be addressed, leading to the development of Dirichlet Process Mixture Models (DPMM) [1]. One of the key advantages of DPMMs is their ability to encompass multidimensional data distributions and conduct data clustering. In this article, these two advantages of DPMMs are exploited.

For further insights into nonparametric Bayes priors, refer to [4, 5, 7]. The selection of a kernel is a pivotal consideration when working with DPMMs. In essence, based on the nature of the data under examination, a flexible distribution is chosen as the DPMM kernel. For real-valued

data, the kernel of choice is typically the normal distribution or a closely related variant. In the case of survival or longevity data, distributions such as Weibull, Gamma, Lognormal, Burr XII, or other longevity distributions are more suitable selections for the kernel. For data confined within the range [0,1], the beta distribution is commonly employed as the kernel.

Researchers such as [2, 8–10, 12] have utilized DPMM with Gamma, Weibull, Beta, Lognormal, Burr XII and Generalized Weibull Inverse kernels for analyzing right-censored survival data. Despite the flexibility offered by DPMM with these kernels, none of these models consistently yield optimal results. Introducing new mixed models by selecting alternative flexible distributions as DPMM kernels may yield improved results in certain datasets compared to other DPMMs.

One of the distributions gaining recent attention in the analysis of survival and reliability data is the Exponentiated Weibull (EW). This distribution serves as a generalization of the Weibull distribution, derived from transforming the exponential function of the collective Weibull distribution. This simple transformation introduces a second shape parameter to the Weibull distribution, significantly enhancing the distribution's flexibility. The Exponentiated Weibull distribution is characterized by one scale parameter and two shape parameters. Depending on the values assigned to these three parameters, the distribution can exhibit constant, increasing, decreasing, one-shaped, or monotonic hazard rates [15, 17].

The objective of this article is to introduce a novel DPM model with an EW kernel and explore some of its potentials, particularly in demonstrating clustering and modeling of multivariate survival data. Censoring is a prevalent occurrence in survival analysis, manifesting in various forms such as right-censoring, left censoring and interval censoring, among others. In survival data, the most typical form of censoring is right-censoring, which encompasses two subtypes: time censoring (Type 1) and failure censoring (Type 2) [14]. In this article, the focus is on time censoring, which involves the termination of observation at a specific time point, regardless of whether the event of interest has occurred.

The subsequent sections of this article are structured as follows: The second section begins with the presentation of key definitions, followed by a discussion on the Exponentiated Weibull (EW) distribution. In the third section, the DPMM with an EW kernel is introduced, along with an explanation of the process for sampling from the posterior distribution of parameters. The fourth and fifth sections delve into the utilization of the proposed model for analyzing both simulated and real datasets.

Finally, the concluding section of the article presents the results derived from the study.

2. Definitions

Definition 2.1. Let (A, B, P) be a probability space, where P_0 is a probability measure on A and ν is a positive real number. For any finite measurable partition $B_1, B_2, \ldots, B_K$ of A, if the following relation holds:

$$(P(B_1), P(B_2), \ldots, P(B_K)) \sim \text{Dir}(\nu P_0(B_1), \nu P_0(B_2), \ldots, \nu P_0(B_K)),$$

then P is called a Dirichlet process with precision parameter ν and base distribution P_0, denoted by $P \sim \text{DP}(\nu, P_0)$. Here, Dir denotes the Dirichlet distribution.

The DPM model is often utilized to model the distribution of multimodal survival data. This model offers a flexible and powerful approach for capturing the complexities inherent in survival data that exhibit multiple modes or distinct patterns. This model was introduced by [1]. In general, a DPM model is defined as follows:

$$f(t; P) = \int k(t|\theta) dP(\theta), \quad \theta \in \Theta$$

where k is the kernel of the mixture model, θ is the parameter vector, Θ is the parameter space and P is the prior distribution function of the θ, which plays the role of the weights of the mixture model. In this framework, the probability measure is random and we consider $P \sim DP(\alpha, P_0)$. As stated earlier in this article, our objective is to employ the EW distribution as a kernel for the DPMM. We aim to express the density functions, survival functions and hazard rates of the EW distribution in the following manner:

$$k(t|\alpha, \beta, \lambda) = \alpha\beta\lambda(\alpha t)^{\beta-1} \exp\left(-(\alpha t)^{\beta}\right) \left\{1 - \exp\left(-(\alpha t)^{\beta}\right)\right\}^{\lambda-1},$$
$$t > 0, \alpha > 0, \beta > 0, \lambda > 0$$

$$S(t|\alpha, \beta, \lambda) = 1 - \left\{1 - \exp\left(-(\alpha t)^{\beta}\right)\right\}^{\lambda}, \quad t > 0$$

$$h(t|\alpha, \beta, \lambda) = \frac{\alpha\beta\lambda(\alpha t)^{\beta-1} \exp\left(-(\alpha t)^{\beta}\right) \left\{1 - \exp\left(-(\alpha t)^{\beta}\right)\right\}^{\lambda-1}}{1 - \left\{1 - \exp\left(-(\alpha t)^{\beta}\right)\right\}^{\lambda}}, \quad t > 0$$

where α represents a scale parameter, while β and λ denote distribution shape parameters. The EW distribution, characterized by the parameters α, β and λ, is denoted as EW (α, β, λ).

Figure 1 illustrates the density, survival and hazard rate functions of the EW distribution which are plotted for various values of α, β and λ. The EW distribution encompasses several distributions as special cases.

Specifically, if $\lambda = 1$, the EW distribution transforms into the Weibull distribution; when $\beta = 1$, it becomes the exponential distribution. For $\beta = 2$, the EW distribution converges to the Burr Type X distribution. Moreover, when both $\beta = 2$ and $\lambda = 1$, the EW distribution yields the Riley distribution [17].

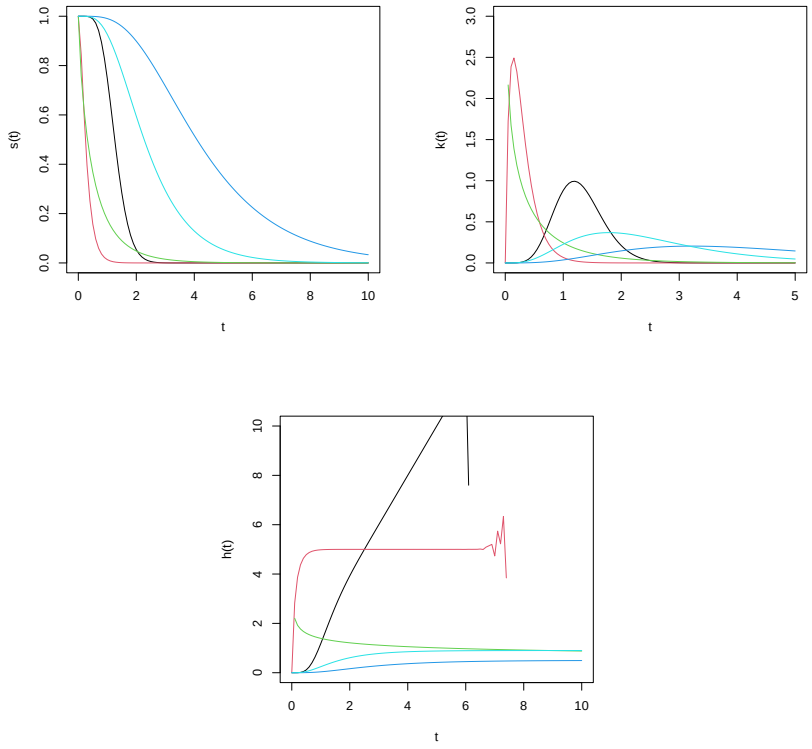

FIGURE 1. Graphs of the survival function (top left), probability density function (top right) and hazard rate function (bottom) for the EW distribution under different values of α, β and λ.

3. The Modelling

In this section, the proposed model followed by a detailed discussion of its components. The model involves a DPMM with the EW distribution serving as the kernel, formulated as

$$f(t;P) = \int_{\alpha}\int_{\beta}\int_{\lambda} k(t|\alpha,\beta,\lambda)dP(\alpha,\beta,\lambda)$$

Here, k represents the density functions of the EW distribution and P as $DP(\nu, P_0)$. The parameter ν can either be assumed as a constant or

treated as a random quantity. In the latter case, a prior distribution is defined for it.

Similar to [3], we will consider the prior distribution for the parameter ν as $Ga(a_\nu, b_\nu)$, where Ga denotes the gamma distribution. Suppose p_0 is the density function corresponding to the distribution function P_0. Let's consider prior distributions for the parameters α, β and λ as follows

$$\lambda|\gamma \sim \text{Exp}(1/\gamma), \qquad \beta \sim U(0,5), \qquad \alpha|\psi \sim \text{Exp}(1/\psi)$$

Assuming the independence of these priors, the joint prior density function for the parameters will be expressed as:

$$p_0(\alpha, \beta, \lambda|\psi, \gamma) = \pi(\alpha|\psi)\pi(\beta)\pi(\lambda|\gamma)$$

For the hyperparameters ψ and γ, we select conjugate prior distributions as follows: ψ follows an Inverse Gamma distribution $IG(a_\psi, b_\psi)$ and γ follows an Inverse Gamma distribution $IG(a_\gamma, b_\gamma)$. If $a_\psi = a_\gamma = 2$, then the variance of ψ and γ would be infinite, indicating that the prior distributions would be noninformative. Values for the hyperparameters b_ψ and b_γ are selected using the same method as described by [12]. Here, we are considering data that is right-censored. Under this type of censorship, the dataset consists of pairs (r_i, t_i^*), where $i = 1, 2, \ldots, n$. Here, $t_i^* = \min(t_i, C_i)$, representing the observed survival time (lifetime) of person i. t_i denotes the actual survival time and C_i represents the time of censorship for person i. r_i also specifies the censorship status of person i. If $r_i = 1$, it means that the lifetime of person i is observed at time t_i and if $r_i = 0$, it indicates that the lifetime of person i is right-censored at time c_i. The Hierarchical representation of the DPMM with an EW kernel can be expressed as follows:

$$\begin{aligned}
&t_i^*|\alpha_i, \beta_i, \lambda_i \sim \{k\left(t_i^*|\alpha_i, \beta_i, \lambda_i\right)\}^{r_i} \{S\left(t_i^*|\alpha_i, \beta_i, \lambda_i\right)\}^{1-r_i}; \quad i = 1, 2, \ldots, n \\
&(\alpha_i, \beta_i, \lambda_i)|P \sim P; \quad i = 1, 2, \ldots, n \\
&P|\psi, \gamma \sim DP(\nu, P_0(\cdot|\psi, \gamma)) \\
&p_0(\alpha, \beta, \lambda|\psi, \gamma) = \pi(\alpha|\psi)\pi(\beta)\pi(\lambda|\gamma) \\
&\alpha|\psi \sim \text{Exp}(1/\psi), \qquad \beta \sim U(0,5), \qquad \lambda|\gamma \sim \text{Exp}(1/\gamma) \\
&\psi \sim IG(a_\psi, b_\psi), \qquad \gamma \sim IG(a_\gamma, b_\gamma), \qquad \nu \sim Ga(a_\nu, b_\nu).
\end{aligned}$$

The proposed model's hierarchy indicates that parameters $(\alpha_i, \beta_i, \lambda_i)$ are associated with the ith dataset and $\{(\alpha_i, \beta_i, \lambda_i); i = 1, 2, \ldots, n\}$ comprises the parameter set corresponding to n datasets. If we designate n^* as the count of distinct vectors in the recent set, then the set of distinct vectors will be $\{\left(\alpha_j^*, \beta_j^*, \lambda_j^*\right); j = 1, 2, \ldots, n^*\}$. Indeed, n* represents the number of clusters and the parameters $\left(\alpha_j^*, \beta_j^*, \lambda_j^*\right)$ correspond to

cluster j. The clustering state of specified data is determined by the relationship between the values $(\alpha_i, \beta_i, \lambda_i)$ and $\left(\alpha_j^*, \beta_j^*, \lambda_j^*\right)$.

Let's consider the vector $L = (L_1, \ldots, L_n)$ representing the clustering state of n data points. This vector is commonly referred to as the configuration vector. The data point i is assigned to the cluster j, denoted as $L_i = j$, if and only if the relation $(\alpha_i, \beta_i, \lambda_i) = \left(\alpha_j^*, \beta_j^*, \lambda_j^*\right)$ is satisfied. Continuing with the definitions, let's denote the vectors as follows: $\alpha = (\alpha_1, \ldots, \alpha_n), \beta = (\beta_1, \ldots, \beta_n), \lambda = (\lambda_1, \ldots, \lambda_n), t^* = (t_1^*, \ldots, t_n^*)$ and $r = (r_1, \ldots, r_n)$, respectively. Similarly, α_{-i}, β_{-i} and λ_{-i} denote the α, β and λ vectors with their ith component removed.

The parameters related to clusters, components of the configuration vector, hyperparameters of the prior distributions and the precision parameter are updated using the Gibbs sampling algorithm. During its implementation, the following five steps are executed iteratively:

Algorithm 1-Gibbs Sampling

(i) Update the components of the configuration vector.
(ii) Update the cluster parameters based on the data within each cluster.
(iii) Update the hyperparameters ψ and γ.
(iv) Update the parameter ν.
(v) Update mixture model weights.

Next, we explain how to implement the 5 steps of the Gibbs sampling algorithm. The first two steps follow Algorithm 8 from [16].

- For $i = 1, \ldots, n$: Let $n^{*(-i)}$ be the number of distinct L_j for $j \neq i$ and let $z = n^{*(-i)} + m$. Label these L_j with values in $\{1, \ldots, n^{*(-i)}\}$. If $L_i = L_j$ for some $j \neq i$, draw values independently from P_0 for those $(\alpha_l, \beta_l, \lambda_l)$ for which $n^{*(-i)} < l \leq z$. If $L_i \neq L_j$ for all $j \neq i$, let L_i have the label $n^{*(-i)} + 1$ and draw values independently from P_0 for those $(\alpha_l, \beta_l, \lambda_l)$ for which $n^{*(-i)} + 1 < l \leq z$. Draw a new value for L_i from $\{1, \ldots, z\}$ using the following probabilities:

$$P(L_i = l \mid L_{-i}, t_i, (\alpha_1, \beta_1, \lambda_1), \ldots, (\alpha_z, \beta_z, \lambda_z)) = \begin{cases} \frac{n_{-i,l} k(t_i \mid \alpha_l, \beta_l, \lambda_l)}{\text{cons}}, & l = 1, \ldots, n^{*(-i)} \\ \frac{\nu/m}{\text{cons}}, & l = n^{*(-i)} + 1, \ldots, z \end{cases}$$

where $n_{-i,l}$ is the number of L_j for $j \neq i$ that are equal to l and cons is equal to $\nu + \sum_{l=1}^{z} n_{-i,l} k(t_i \mid \alpha_l, \beta_l, \lambda_l)$. Change the state to contain only those $(\alpha_l, \beta_l, \lambda_l)$ that are now associated with one or more observations. After updating the components

of the configuration vector, the state of the data clustering is updated. Then, based on the data belonging to each cluster, the parameters of that cluster are updated as follows.

- For $j = 1, \cdots, n^*$: Draw a new value for $\left(\alpha_j^*, \beta_j^*, \lambda_j^*\right)$ from posterior distribution

$$\pi\left[\alpha_j^*, \beta_j^*, \lambda_j^* | \mathbf{t}^*, \mathbf{r}\right] \propto \pi\left[\alpha_j^* | \psi\right] \pi\left[\beta_j^*\right] \pi\left[\lambda_j^* | \gamma\right] \times \prod_{i; L_i = j} \left\{k\left(t_i^* | \alpha_j^*, \beta_j^*, \lambda_j^*\right)\right\}^{r_i} \left\{S\left(t_i^* | \alpha_j^*, \beta_j^*, \lambda_j^*\right)\right\}^{1-r_i}$$

  The recent distribution has a complex form and sampling from it is not straightforward. To generate samples from this distribution, the Metropolis-Hastings algorithm can be employed.
- The conditional posterior distributions of hyperparameters ψ and γ are as follows:

$$\pi\left[\psi | \alpha^*, n^*\right] \propto \pi[\psi] \prod_{j=1}^{n^*} \pi\left[\alpha_j^* | \psi\right],$$

$$\pi\left[\gamma | \lambda^*, n^*\right] \propto \pi[\gamma] \prod_{j=1}^{n^*} \pi\left[\lambda_j^* | \gamma\right].$$

  After undergoing some calculations and simplifications, these conditional distributions take the following form:

$$\psi | \alpha^*, n^* \sim IG\left(\text{shape} = a_\psi + n^*, \text{scale} = b_\psi + \sum_{j=1}^{n^*} \alpha_j^*\right),$$

$$\gamma | \lambda^*, n^* \sim IG\left(\text{shape} = a_\gamma + n^*, \text{scale} = b_\gamma + \sum_{j=1}^{n^*} \lambda_j^*\right).$$

  The precision parameter ν of the DP is updated according to the algorithm proposed by [3]. The values of the DPM model weights, namely $\omega_j; j = 1, 2, \ldots, n^*$, are updated as follows. According to [19] representation of the DP, we have

$$P(.) = \sum_{j=1}^{\infty} \omega_j I_{(\alpha_j, \beta_j, \lambda_j)(.)}.$$

  The aforementioned sum is approximated as

$$P(.) \approx \sum_{j=1}^{n^*} \omega_j I_{(\alpha_j, \beta_j, \lambda_j)}(\cdot)$$

so that $\sum_{j=1}^{n^*} \omega_j = 1$. The sequence $\omega_j; j = 1, 2, \ldots, n^*$ constitutes the weight values of the model. To calculate these weights, we employ the stick-breaking approach [11]. To generate the weights $\omega_j; j = 1, 2, \ldots, n^*$ according to the mentioned method, first a sample of size $n^* - 1$ is generated from the $Beta(1, \nu)$ distribution, $v_j; j = 1, 2, \ldots, n^* - 1$. Then, the weight values are calculated as follows

$$\omega_1 = v_1, \qquad \omega_j = v_j \prod_{i=1}^{j-1}(1 - v_i), \quad j = 2, 3, \ldots, n^* - 1,$$

where $\omega_{n^*} = 1 - \sum_{j=1}^{n^*-1} \omega_j$. Additionally, values $(\alpha_j, \beta_j, \lambda_j)$ for $j = 1, 2, \ldots, n^*$ are generated from a mixture distribution

$$\frac{\nu}{\nu + n} p_0(\cdot|\psi, \gamma) + \frac{1}{\nu + n} \sum_{j=1}^{n^*} I_{(\alpha_j^*, \beta_j^*, \lambda_j^*)}(\cdot).$$

At each iteration of the Gibbs algorithm, after generating the values of $\left\{ \left(\alpha_j^*, \beta_j^*, \lambda_j^* \right); j = 1, 2, \ldots, n^* \right\}$ and $\omega_j; j = 1, 2, \ldots, n^*$, the density, survival and hazard rate functions at point t are approximated as follows:

$$\hat{f}(t) = \sum_{j=1}^{n^*} \omega_j^* k\left(t|\alpha_j^*, \beta_j^*, \lambda_j^*\right),$$

$$\hat{S}(t) = \sum_{j=1}^{n^*} \omega_j^* S\left(t|\alpha_j^*, \beta_j^*, \lambda_j^*\right),$$

$$\hat{h}(t) = \frac{\hat{f}(t)}{\hat{S}(t)}.$$

One of the biggest challenges in fitting Bayesian models, especially DPM models, is the computationally intensive nature of Gibbs sampling algorithms, which often require long runtimes. Over the last twenty years, significant efforts have been made to tackle this problem, leading to the development of specialized software tools like Nimble, Dpweibull and DPpackage. [18] introduced a valuable package for fitting DPMMs within the R software, known as the Dirichletprocess package. The default choice for the kernel in the DPMM is the Weibull distribution. However, it's worth noting that within this package, users have the flexibility to select any arbitrary distribution as the kernel for the DPMM.

The package also accommodates censored data, enhancing its versatility and applicability. In [16], eight algorithms for implementing MCMC

in DPMMs are presented, among which algorithm number 8 stands out as particularly important and widely applicable. The Dirichletprocess package provides the choice to employ algorithms 4 and 8 from [16]. In this article, we specifically utilize algorithm number 8.

4. Simulation Study

In this section, to assess the performance of the proposed model, samples of size $n = 100$ are generated from a mixture of two log-normal distributions.

(4.1)
$$0.4\ \text{Lnorm}\left(\mu_1 = 0, \sigma_1 = \sqrt{0.25}\right) + 0.6\ \text{Lnorm}\left(\mu_2 = 1.2, \sigma_2 = \sqrt{0.02}\right)$$

where $\text{Lnorm}(\mu, \sigma)$ represents a log-normal distribution with μ and σ parameters.

In this article, the continued development of the DPM model is represented by the Exponentiated Weibull kernel (DPEWM). To compare the performance of the proposed model with other DPM models, we considered DPMM with the Weibull kernel (DPWM) and Burr XII with three parameters as the kernel (DPB3M). They are fitted to the dataset generated from model (4.1). Details about DPWM and DPB3M models can be found in [12] and [9], respectively.

We simulate 100 datasets from model (4.1). For each dataset, Gibbs' algorithm is repeated $N = 5,000$ times and $B = 1000$ repetitions are considered as burn-in period iterations. To demonstrate the insensitivity of the inferences to the selection of the prior distribution parameters, two sets of values for the parameter ν are considered: $(a_\nu, b_\nu) = (2.0, 5)$ and $(3.0, 0.05)$. Corresponding to the two sets of the most recent values, we let $\nu \sim Ga(2.0, 5)$ and $\nu \sim Ga(3.0, 0.05)$. The corresponding results of these two prior distributions are shown in Figures 2 and 3. These figures display the estimated density, survival and hazard rate functions of model (4.1) using DPMM with various kernels across 100 simulated datasets.

These figures depict the results of the DPEWM, DPWM and DPB3M models, presented in the right, middle and left columns, respectively. The density function, survival and hazard rate estimates for model (4.1) are displayed in the first through third rows, respectively. The color red is also employed in these plots to represent the actual curves of model (4.1). Upon comparing the findings depicted in Figures 2 and 3 it becomes evident that inferences are not significantly affected by the selection of a_ν and b_ν values.

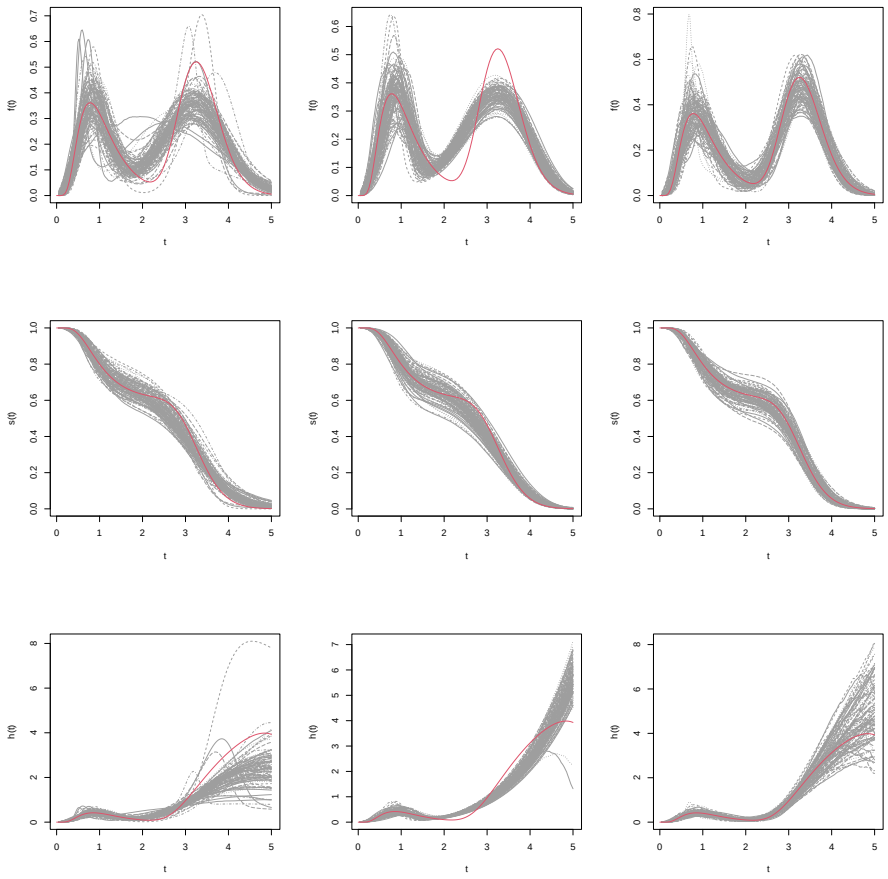

FIGURE 2. Estimating density functions (first row), survival (middle row) and hazard rate (lower row) mixture model (4.1) using DPEWM (right column), DPWM (middle column) and DPB3M (left column) models in 100 sets of models (4.1) and under the prior $\nu \sim Ga(2, 5)$.

The results generated by the DPWM model for estimating the density and survival functions of model (4.1) may exhibit both overestimation and underestimation. Nevertheless, when estimating the hazard rate function of the actual model, the DPWM model tends to exhibit overestimation for large values of t. The observed discrepancies in overestimation and underestimation under the DPB3M model are reduced in the estimation of the density and survival functions compared to the DPWM model. However, there is a notable underestimation in the estimation of the risk rate function of the actual model under DPB3M.

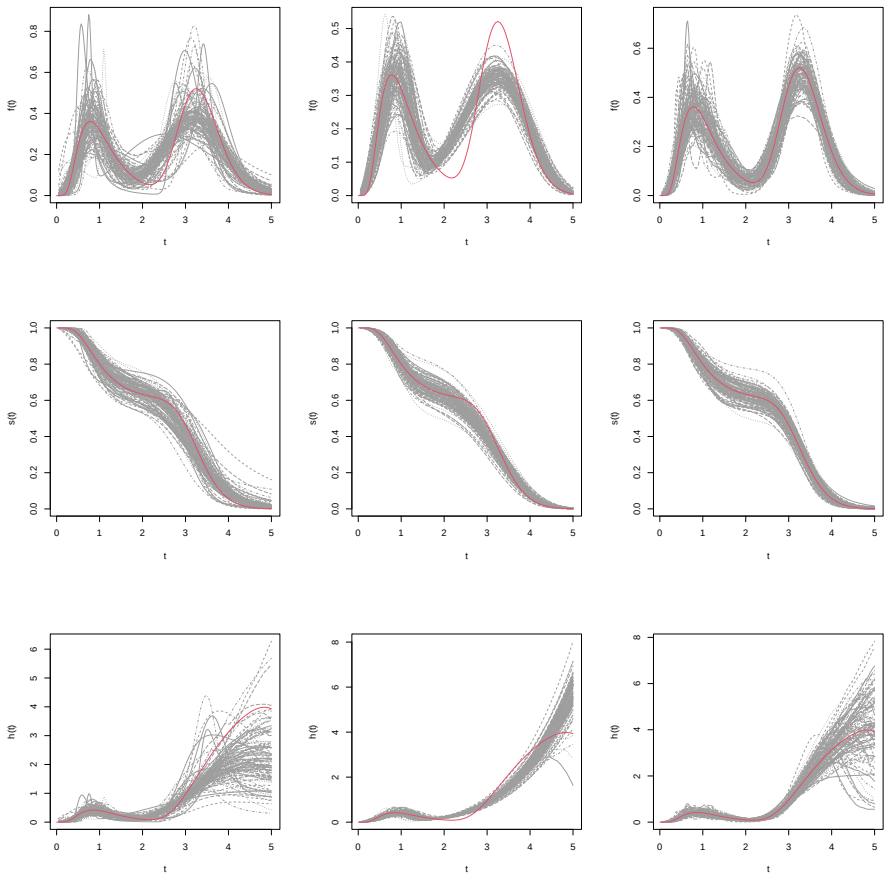

FIGURE 3. Estimating density functions (first row), survival (middle row) and hazard rate (lower row) mixture model (4.1) using DPEWM models (right column), DPWM (middle column) and DPB3M (left column) in 100 sets of models (4.1) and under the prior $\nu \sim Ga(3.0, 0.05)$.

The results obtained from the DPEWM model demonstrate neither underestimation nor overestimation in estimating the curves of the actual model. Consequently, it can be concluded that the DPEWM model yields superior results in estimating the actual curves of the model compared to the other two models it has produced.

To facilitate a more precise comparison of the proposed model's performance with the other two models, the Mean Squared Error (MSE) index values in estimating the mentioned functions are computed at various points, as illustrated in Figure 4. In this figure, the results of the DPEWM, DPWM and DPB3M models were respectively colored black,

blue and red. The results corresponding to the priors $\nu \sim Ga(2.0, 5)$ and $\nu \sim Ga(3.0, 0.05)$ are displayed in the first and second rows, respectively in Figure 4. Comparing the results of these two rows, it's clear they are nearly identical, confirming the insensitivity of the inferences to the prior distribution of the parameter ν.

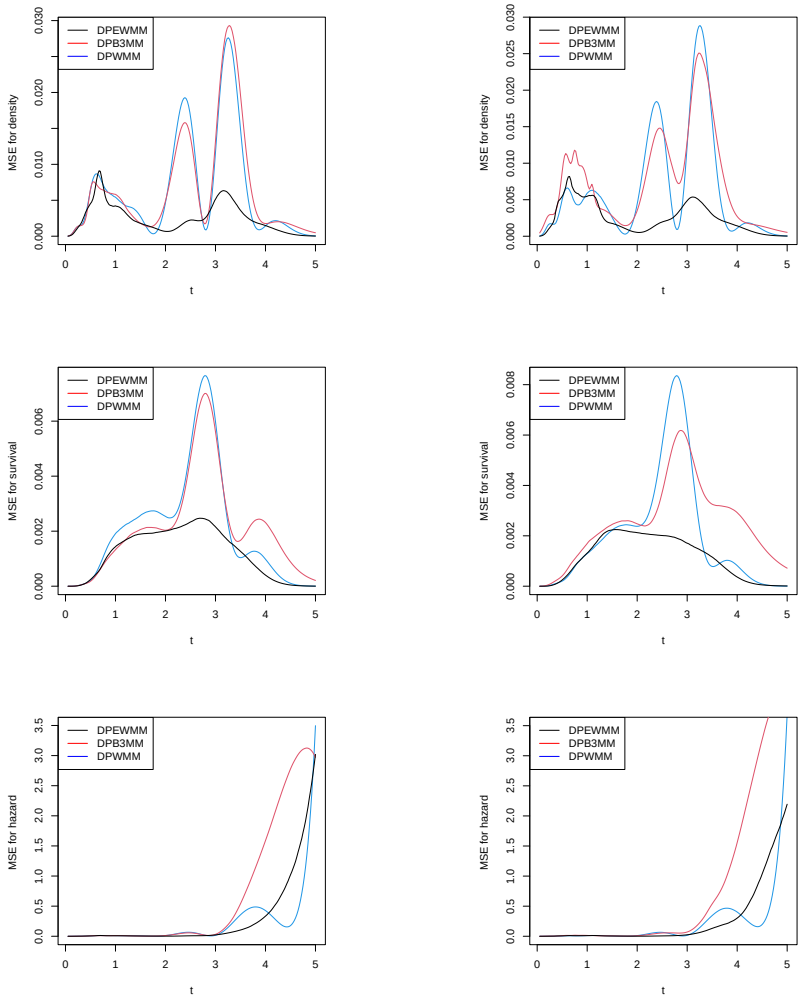

FIGURE 4. MSE values for estimating density functions (top), survival functions (middle) and hazard rate functions (bottom) of model (4.1) at different points using DPMM with different kernels and under the Priors: $\nu \sim Ga(2.0, 5)$ (Left column) and $\nu \sim Ga(3.0, 0.05)$ (Right column).

It's also notable that the total MSE values obtained in the estimation of the density functions, survival and hazard rate of model (4.1) under the DPEWM model are smaller than those of the other two models. Consequently, it can be inferred that the DPEWM model has estimated the actual curves of model (4.1) better than the other two models.

5. Analysis of the Real Dataset

In this section, the Faithful dataset in R software was analyzed. This dataset contains data on two variables, namely "eruptions" and "waiting". These two variables represent the time of eruption and the waiting time between eruptions of the old volcano, measured in minutes. The waiting time between volcano eruptions is a significant variable of interest in this analysis. The primary objective in analyzing this dataset is to estimate the real curves, including the density, survival and hazard rate functions.

Aside from this primary objective, the secondary aim involves clustering this dataset. During the implementation of the Gibbs sampling algorithm, the total number of iterations, the burn-in period iterations and the thinning interval were set as follows: $N = 15,000$, $B = 5,000$ and $T = 20$, respectively.

Figure 5 illustrates the Bayesian estimates (posterior median) of density functions, survival rates, data hazard rate and cumulative hazard rate under the DPEWM, DPWM and DPB3M models.

The data histogram along with the posterior medians for the density function is shown on the right of the upper row of Figure 5. In this diagram, the blue curve represents the results obtained by estimating the data density using the *density* command in R software. It's evident from the diagram that the proposed DPEWM model effectively captures the distribution of data, particularly the peaks, in comparison to other models. When analyzing survival data, nonparametric estimates like the Kaplan-Meier method are frequently employed [14]. In the left graph of the upper row in Figure 5, you'll find the results of posterior point estimation for the survival function using the three mixture models mentioned. The color blue is used to denote values associated with the Kaplan-Meier estimate. Except for the DPB3M model, the results of the other three models closely resemble each other in estimating the survival function.

The Bayesian estimates of data hazard rate are depicted in the lower right graph of Figure 5. Various methods exist to determine which model provides a better estimate of the actual hazard rate of data. One of these methods involves comparing the cumulative hazard rate values presented by the models with the values of the Nelson-Aalen cumulative

hazard rate. A model that exhibits cumulative hazard rate values closer to the values of the Nelson-Aalen cumulative hazard rate is considered to better estimate the real hazard rate of the data. The outcomes from the DPEWM and DPWM models closely align with the Nelson-Aalen cumulative hazard rate values. Overall, the DPEWM model outperforms other models in this dataset, yielding superior results in estimating density functions, survival rates and generated hazard rates.

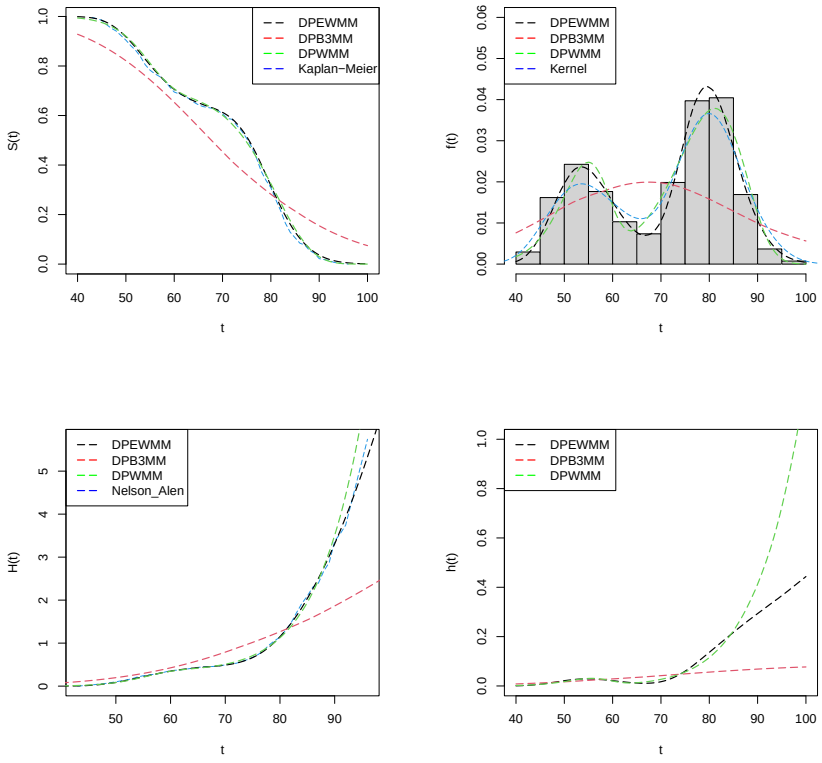

FIGURE 5. Results related to calculation of density functions estimation (top right), survival (top left), hazard rate (lower right) and cumulative hazard rate (lower left) of the Faithful dataset and under the Prior: $\nu \sim Ga(2, 2)$.

Clustering stands as a fundamental data analysis method, where data is partitioned into groups according to their shared characteristics, ensuring that members within each group exhibit similarity. This method aids in uncovering hidden patterns and relationships within data, thereby facilitating a deeper understanding of both data and information.

Through categorizing data into distinct clusters, similar observations are grouped together within each cluster, enabling varied analyses to be conducted on each cluster.

DPMMs offer the advantage of efficiently clustering data. Figure 6 displays the outcomes of data clustering employing DPMM with various kernels.

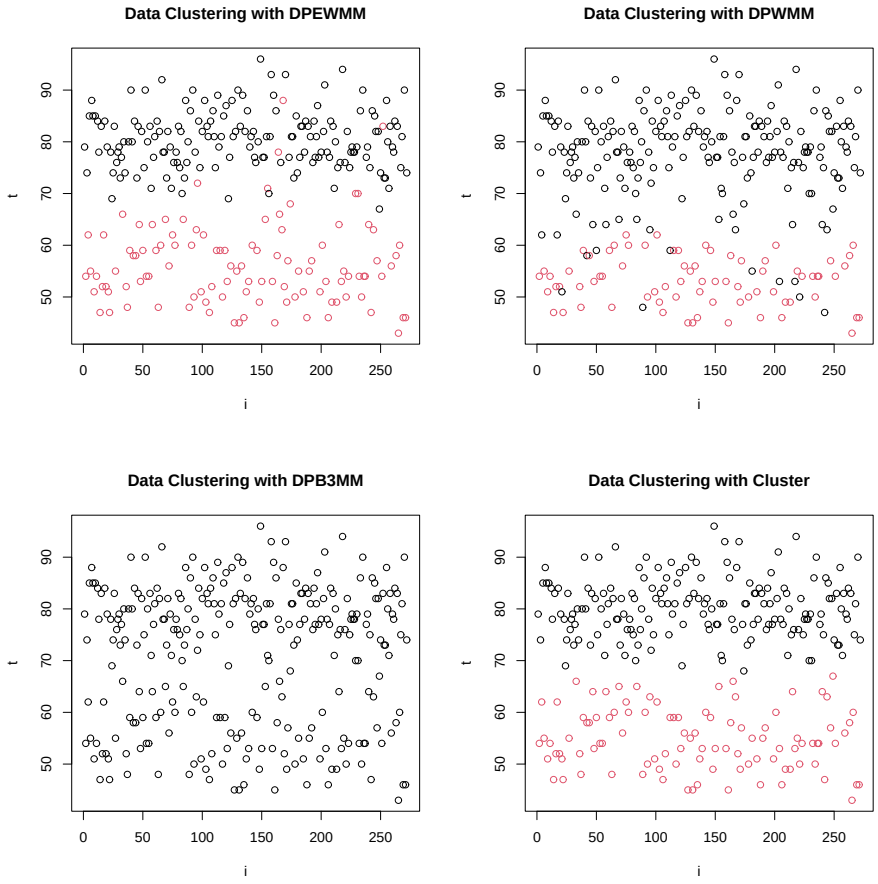

FIGURE 6. Results of data clustering of the Faithful dataset using different models and under the prior $\nu \sim Ga(2, 2)$.

These outcomes are derived from the clustering status of the data in the final iteration of the Gibbs algorithm. Within R software, numerous packages facilitate data clustering, among which is the Cluster package. In the latest version of the package, data clustering is conducted utilizing the K-Means command, with the corresponding outcomes depicted in the final diagram of Figure 6. Given the presence of two peaks in the

histogram of the data, it's highly likely that the data naturally forms clusters. As depicted in Figure 6, both the DPEWM and DPWM models separated the data into two clusters, while the DPB3M model grouped the data into a single cluster.

In Table 1, the values for calculating the empirical probability function of the number of clusters n^* in the DPEWM, DPWM and DPB3M models are provided. These results are based on the number of clusters observed in 10,000 end iterations of the Gibbs algorithms. According to these findings, the most probable number of data clusters under the DPEWM, DPWM and DPB3M models is 2, 2 and 1, respectively. The probabilities corresponding to the last three values are 0.9374, 0.6604 and 0.9943, respectively. These values indicate that the DPEWM model is significantly more likely than the other two models to accurately capture the data in two-cluster models. This results show that the DPEWM model clustered data better than the other two models.

TABLE 1. Empirical probability function values of the number of clusters n^* in DPEWM, DPWM and DPB3M models of the Faithful dataset and under the Prior: $\nu \sim Ga(2, 2)$.

n^*	DPEWM	DPWM	DPB3M
1	0.0000	0.0000	0.9943
2	0.9374	0.6604	0.0056
3	0.0600	0.2576	0.0001
4	0.0024	0.0613	0.0000
5	0.0002	0.0166	0.0000
6	0.0000	0.0034	0.0000
7	0.0000	0.0006	0.0000
8	0.0000	0.0001	0.0000

6. Conclusion and Discussion

This paper introduces a novel mixed model for analyzing survival data, which combines Dirichlet processes characterized by an Exponentiated Weibull kernel. Through the analysis of both simulated and real datasets, the performance of this proposed model was assessed. The findings reveal that the new model demonstrates optimal performance in estimating the density, survival and hazard rate of survival data. One of the key advantages of mixture models employing Dirichlet processes

is their ability to cluster data, a feature highlighted in this study. The utilization of real data in this research further reinforces the satisfactory results obtained.

Although the current model provides a powerful framework for nonparametric analysis, several extensions could enhance its practical applicability. First, incorporating regression components would strengthen the model's ability to analyze covariate effects. Second, extending the methodology to accommodate other censoring types (such as interval-censoring) would broaden its scope. Finally, developing more efficient computational algorithms would facilitate more extensive simulation studies, further validating the model's robustness. Notwithstanding these potential extensions, the proposed model in its current form remains a robust and flexible tool for analyzing survival data with complex underlying distributions.

Acknowledgment. We would like to express our gratitude to Razi University for its support during the development of this work.

References

1. C.E. Antoniak, *Mixtures of Dirichlet processes with applications to nonparametric problems*, Ann. Stat., 2 (1974), pp. 1152-1174.
2. N. Cheng and T. Yuan, *Nonparametric Bayesian lifetime data analysis using Dirichlet process lognormal mixture model*, Naval Res. Logist., 60 (2013), pp. 208-221.
3. M.D. Escobar and M. West, *Bayesian density estimation and inference using mixtures*, J. Amer. Statist. Assoc., 90 (1995), pp. 577-588.
4. T.S. Ferguson, *Bayesian analysis of some nonparametric problems*, Ann. Stat., 1 (1973), pp. 209-230.
5. T.S. Ferguson, *Prior distributions on spaces of probability measures*, Ann. Stat., 2 (1974), pp. 615-629.
6. T.S. Ferguson and E.G. Phadia, *Bayesian nonparametric estimation based on censored data*, Ann. Stat., 7 (1979), pp. 163-186.
7. J.K. Ghosh and R.V. Ramamoorthi, *Bayesian Nonparametrics*, Springer Series in Statistics, Springer-Verlag, New York, 2003.
8. B. Haji Joudaki, R. Hashemi and S. Khazaei, *Analysis of censored data using Dirichlet process mixture model with generalized inverse Weibull distribution as kernel*, J. Statist. Sci., 17 (2023), pp. 275-298.
9. B. Haji Joudaki, R. Hashemi and S. Khazaei, *Survival analysis using Dirichlet process mixture model with three-parameter Burr XII distribution as kernel*, Comm. Statist. Simulation Comput., 53 (2024), pp. 2406-2424.

10. T. Hanson, *Modeling censored lifetime data using a mixture of gammas baseline*, Bayesian Anal., 1 (2006), pp. 575-594.
11. H. Ishwaran and L.F. James, *Approximate Dirichlet process computing in finite normal mixtures: Smoothing and prior information*, J. Comput. Graph. Statist., 11 (2002), pp. 508-532.
12. A. Kottas, *Dirichlet process mixtures of beta distributions, with applications to density and intensity estimation*, in: Proc. 23rd International Conference on Machine Learning (ICML), Pittsburgh, USA, 2006, 47.
13. P. Lahiri and D.H. Park, *Nonparametric Bayes and empirical Bayes estimators of mean residual life at age t*, J. Statist. Plann. Inference, 29 (2000), pp. 125-136.
14. J.F. Lawless, *Statistical Models and Methods for Lifetime Data*, John Wiley and Sons, New York, 2011.
15. G.S. Mudholkar and D.K. Srivastava, *Exponentiated Weibull family for analyzing bathtub failure-rate data*, IEEE Trans. Reliab., 42 (1993), pp. 299-302.
16. R.M. Neal, *Markov chain sampling methods for Dirichlet process mixture models*, J. Comput. Graph. Statist., 9 (2000), pp. 249-265.
17. M. Pal, M.M. Ali and J. Woo, *Exponentiated Weibull distribution*, Statistica, 66 (2006), pp. 139-147.
18. G.J. Ross and D. Markwick, *dirichletprocess: An R Package for Fitting Complex Bayesian Nonparametric Models*, University of Edinburgh and University College London, 2018.
19. J. Sethuraman, *A constructive definition of Dirichlet priors*, Statist. Sinica, 4 (1994), pp. 639-650.
20. V. Susarla and J. Van Ryzin, *Nonparametric Bayesian estimation of survival curves from incomplete observations*, J. Amer. Statist. Assoc., 71 (1976), pp. 897-902.

---

[1] Department of Statistics, Faculty of Sciences, Razi University, Kermanshah, Iran.
Email address: alisalih95@gmail.com

[2] Department of Statistics, Faculty of Sciences, Razi University, Kermanshah, Iran.
Email address: s.khazaei@razi.ac.ir

[3] Department of Statistics, Faculty of Basic Sciences, Lorestan University, Khorramabad, Iran.
Email address: joudaki.b@lu.ac.ir