IOSR Journal of Mathematics (IOSR-JM)
e-ISSN: 2278-5728, p-ISSN: 2319-765X. Volume 15, Issue 6 Ser. VI (Nov – Dec 2019), PP 39-48
www.iosrjournals.org

A Simulation Study of Survival Mixture Model of Gamma Distributions with Different Set of Censoring Percentages and Mixing Proportions

Yusuf Abbakar Mohammed[1], Suzilah Ismail[2]
[1](*Department of Mathematical Sciences, University of Maiduguri, PMB1069, Maiduguri, Nigeria)*
[2](*School of Quantitative Sciences, Universiti Utara Malaysia, Sintok, Malaysia)*

Abstract: *This study proposes a three components survival mixture model of Gamma distribution. The performance of the model was investigated using simulated survival data. The data were simulated based on two different sets of mixing proportion arranged in ascending and descending order with three different censoring proportions (10%, 20% and 40%). The Expectation Maximization (EM) Algorithm was used to estimate the maximum likelihood estimates of the model parameters. The hazard functions of each of the censoring proportions were investigated graphically. The parameters of the model were estimated successfully and they were all close the initial values used in generating the data. Three hundred times repetitions of the simulation were run. The mean square error (MSE) and root mean square error (RMSE) were estimated to assess the consistency and stability of the model. The simulated data used to compare the effect of different censoring proportions revealed that the model performed much better with small proportion of censored observations. Also the model performed well with both the ascending and descending order of the mixing proportions. However, mixing proportions in ascending order performed better than the descending order. The hazard function graphs showed that, samples with higher proportion of censored observations have lower hazard compared to the smaller censored observations. The model showed that survival mixture models are flexible and maintain the features of the classical survival model and are better option for modelling heterogeneous survival data.*

Keywords: *Gamma; Proportion; Survival function; Mixture model; Expectation maximization;*

Date of Submission: 04-12-2019 Date of Acceptance: 19-12-2019

I. Introduction

Survival analysis commonly employed to analyse some event that occurs within a particular period of time. The methods of survival analysis are employed in several fields such as medicine, social sciences, economic and industrial production to mention few. Classical parametric survival models are very powerful techniques in survival analysis; they preferred over the nonparametric methods when the chosen distribution fit the data properly. The Lognormal, Gamma and logistic distributions are frequently employed in the analysis of survival data. [1], [2], [3] and [4]. Mixture models are normally utilised for analysing survival data which are heterogeneous in nature. In the recent decades, many authors employed mixture model technique to analyse survival data. A survival mixture model of Weibull distributions with two components was used where the parameters of the model were estimated by graphical approach [5]. A new technique was developed for evaluating the parameters of a two components survival mixture model of Weibull distributions [6].

Expectation Maximization (EM) was used to assess performance of a two-component survival mixture model of the Weibull-Weibull distributions, and also, the model stability was evaluated [7]. A survival mixture models of Gamma-Gamma, Weibul-Weibull and Lognormal-Lognormal distributions were proposed to model survival data [8], model selection technique was used to select the model which better represents the real data. A survival mixture of mixed distribution was used to analyse heterogeneous data; the study proposed a two components survival model of the Extended Exponential-Geometric (EEG) distribution [9]. Also, a two components survival mixture models of different distributions was proposed, the study considered an Exponential-Gamma, an Exponential-Weibull and a Gamma-Weibull models for analysing heterogeneous survival data[10].

Three components survival mixture models did not receive much attention. In a study to observe the risk of death after open-heart surgery [11], the article was able to classify the risk of death after the surgery by three different time overlapping phases which are better analysed by a three components mixture model, as was pointed out by [12] and [13]. Another study proposed a parametric survival mixture model of the Exponential,

Gamma and Weibull distributions to model heterogeneous survival data, where they used simulated data to investigate the stability and consistency of the model [14]. The method of model selection was employed in another article, to select the model which fit the data better [15]. Bayesian method was utilized to analyse a three components survival mixture model of Weibull distributions [16]. In several cases where data include missing or unobserved observations, Expectation Maximization (EM) is frequently employed to analyse such data [17]. The Maximum Likelihood parameters of survival mixture models are usually estimated by the EM method [18].

Simulated survival data were generated and utilized to evaluate the flexibility and appropriateness of a three components survival mixture model of the Gamma distributions in analysing heterogeneous data. This article is arrangement in the following order. Section two is devoted to outline some important functions in survival analysis and some properties of Gamma distribution are discussed. Section three used to highlight the implementation of the survival mixture model of three components in the survival analysis. Section four elaborates the implementation of the EM method in estimating the maximum likelihood parameters of the three component survival mixture model. Section five devoted for data analysis to estimate the parameters of the proposed model and compare the different censoring percentages and the two sets of mixing proportions. Section six is used for summary and conclusion.

II. Probability Distributions Used In Survival Analysis

Survival analysis concerned with the implementation of a collection of statistical techniques in the modelling and analyse of survival data. The point of importance is the occurrence of a specified event of interest within an interval of time. The response variable *T* is a non-negative random variable which gives the survival time of an object or an individual which can be expressed as a probability density function (pdf) denoted by $f(t)$, which is written as

$f(t) = \frac{dF(t)}{dt}$

Where $F(t)$ is the distribution function of response variable *T*. The density function $f(t)$ is a nonnegative function and the area between the curve and the *t* axis is equal to 1. The survival function denoted by $S(t)$ is one of the important function and is expressed as

$S(t) = 1 - F(x)$

It specify the probability that an individual will survive beyond a particular time *t*. The survival function $S(t)$ is a monotonic decreasing continuous function with $S(0) = 1$ and $S(\infty) = 0$. The hazard function can be represented by $h(t)$, and is given by

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

which gives the probability of an individual to fail within a small interval $(t, t + \Delta t)$, provided that the individual was a life until the beginning of that interval.

Pure classical parametric survival models are powerful method in survival analysis; when the chosen probability distribution appropriately represents the data. The Gamma density is commonly employed in the analysis of survival data. [2], [3] and [4]. The probability density function $f(t)$ and survival functions $S(t)$ of the Gamma distributions are highlighted below.

Gamma distribution

$f_G(t) = t^{\alpha-1} \frac{e^{-t/\beta}}{\beta^{\alpha}\Gamma(\alpha)} \qquad t \text{ and } \alpha, \beta > 0$

$S_G(t) = 1 - \frac{\Gamma_x(\alpha)}{\Gamma(\alpha)}$

Where $\Gamma_x(\alpha)$ is known as the incomplete Gamma function.

III. Parametric Mixture Model Of Three Distributions

Mixture models are commonly employed in survival analysis for their flexibility. They are preferred over the pure classical parametric survival models when the data are of heterogeneous nature [19] and [20]. Survival mixture model of three components is used when it is believed that the data consist of three subpopulation or subgroups. Equation (1) represents a parametric survival mixture model of three components.

$$f_{X,Y,Q}(t;\Theta) = \pi_1 f_X(t;\theta_X) + \pi_2 f_Y(t;\theta_Y) + \pi_3 f_Q(t;\theta_Q) \tag{1}$$

Where the vector $\Theta = (\pi_1, \pi_2, \theta_X, \theta_Y, \theta_Q)$, represents the vector the parameters of the mixture model. The functions

$f_X(t;\theta_X), f_Y(t;\theta_Y)$ and $f_Q(t;\theta_Q)$ are the probability density functions corresponding to each component with some parameters θ_X, θ_Y and θ_Q respectively.

In this paper a three components survival mixture model of Gamma distributions is proposed to model heterogeneous survival data. The proposed model consists of three component of the Gamma distributions and is defined as

$$f_{G1\_G2\_G3}(t;\Theta) = \pi_1 f_{G1}(t;\alpha_1,\beta_1) + \pi_2 f_{G2}(t;\alpha_2,\beta_2) + \pi_3 f_{G3}(t;\alpha_3,\beta_3),$$

where π_i's are the mixing proportions and $\sum_{i=1}^{3}\pi_i = 1$. The functions f_{G1} ,f_{G2} and f_{G3} are the probability density functions of the Gamma distributions of respective three components of the model.

IV. Expectation Maximization (EM) And Survival Mixture Model

One of the most efficient and effective methods commonly employed to estimate the maximum likelihood estimators of finite mixture models is the EM [19]. On the implementation of the EM to the mixture model, the variables z`s are considered as missing values. The EM consists of two different steps, the first one is the Expectation step or the E-step and the second one is the Maximization step or the M-step.

The z_i variables are treated as missing observations in the E-step, the hidden variable vector $z_i=[z_{1i},z_{2i},z_{3i}]$ are estimated by the evaluation of the expectation $E(z_{ki}|t_i)$.

Thus

$$\sum_{i=1}^{3}\hat{z}_{ij} = E(z_{ij}\,|\,t_j) = \frac{\pi_i f_X(t_j;\theta_X)}{\pi_1 f_X(t_i;\theta_X)+\pi_2 f_Y(t_i;\theta_Y)+\pi_3 f_Q(t_i;\theta_Q)}$$

The functions $E(z_{1i}|t_i)$, $E(z_{2i}|t_i)$ and $E(z_{3i}|t_i)$ calculated in the E-step will be maximized in the M-step of the EM under the condition the sum of π_i's equals to 1.

The proposed model can be expressed as follows

$$f_{E\_G\_W}(t;\Theta) = \pi_1 f_{G1}(t;\alpha_1,\beta_1) + \pi_2 f_{G2}(t;\alpha_2,\beta_2) + \pi_3 f_{G3}(t;\alpha_3,\beta_3),$$

where $f_G(t;\alpha_1,\beta_1)$ with unknown parameters α_1,β_1, $f_G(t;\alpha_2,\beta_2)$ with unknown parameters α_2,β_2 and $f_W(t;\alpha_3,\beta_3)$ with unknown parameters α_3,β_3 are the Gamma density functions for the three distributions (G1, G2 and G3).

The general Gamma mixture model is defined as

$$f(t) = \sum_{i=1}^{k} \pi_i f_i(t;\alpha_i,\beta_i) \tag{2}$$

where $f_i(t;\alpha_i,\beta_i)$ represent the density function of Gamma distribution as highlighted in section two with unknown parameters α_i,β_i, with $\alpha_i > 0$, $\beta_i > 0$.

The log-likelihood function is expressed as

$$\log L_c(\alpha_i,\beta_i,\pi_i) = \sum_{i=1}^{k}\sum_{j=1}^{n} z_{ij}\left\{\log \pi_i + \delta_j \log\left[\frac{1}{\beta_i\Gamma(\alpha_i)}\left(\frac{t_j}{\alpha_i}\right)^{\alpha_i-1} e^{\frac{-t}{\alpha_i}}\right] + (1-\delta_j)\log\left[\frac{\Gamma(\alpha_i,t_j/\beta_i)}{\Gamma\alpha_i}\right]\right\} \tag{3}$$

The EM algorithm starts with the E-step. After the g^{th} iteration, $z_{ij}^{(g)}$ is the conditional expectation of Z_{ij} given the observed data. Then the current conditional expectation of the complete-data log-likelihood is given by

$$Q(\alpha_i,\beta_i,\pi_i) = \sum_{i=1}^{k}\sum_{j=1}^{n} z_{ij}^{(g)}\left\{\log \pi_i + \delta_j \log\left[\frac{1}{\beta_i\Gamma(\alpha_i)}\left(\frac{t_j}{\alpha_i}\right)^{\alpha_i-1} e^{\frac{-t}{\alpha_i}}\right] + (1-\delta_j)\log\left[\frac{\Gamma(\alpha_i,t_j/\beta_i)}{\Gamma\alpha_i}\right]\right\} \tag{4}$$

The M-step on the $(g+1)^{th}$ iteration requires the global maximization of (3) with respect to α_i,β_i and π_i . The

mixing proportions π_i can be updated by $\hat{\pi}_i^{(g+1)} = \sum_{j=1}^{n} z_{ij}^{(g)} / n, \quad i = 1,..., \quad k$. In order to get the updated maximum likelihood estimate of the component model parameters α_i, β_i , the partial differentiation of equation (3) was taken with respect to the parameters the α_i, β_i , thus

$$\frac{\partial Q}{\partial \alpha_i} = \sum_{j=1}^{n} z_{ij}^{(g)} \delta_j \left[-\log \beta_i + \Psi(\alpha_i) + \log t_j \right] + \sum_{j=1}^{n} z_{ij}^{(g)} (1-\delta_j) \left[\log \beta_i + \frac{1}{\Gamma(\alpha_i, t_j/\beta_i)} \frac{\partial}{\partial \alpha_i} \Gamma(\alpha_i, t_j/\beta_i) \right] \tag{5}$$

$$\frac{\partial Q}{\partial \beta_i} = \sum_{j=1}^{n} z_{ij}^{(g)} \left[-\delta_j \left(\frac{\alpha_i}{\beta_i} + \frac{t_i}{\beta_i^2} \right) + \frac{(1-\delta_j)}{\Gamma(\alpha_i, t_j/\beta_i)} \frac{\partial}{\partial \beta_i} \Gamma(\alpha_i, t_j/\beta_i) \right] \tag{6}$$

Now, the upper incomplete gamma function can be differentiated with respect to β_i using Leibnitz's rule, and we then obtain from (5) that

$$\beta_i = \left[\sum_{j=1}^{n} z_{ij}^{g} t_j / \alpha_i + \sum_{j=1}^{n} z_{ij}^{g} \delta_j / \alpha_i - \sum_{j=1}^{n} \frac{t_j^{\alpha_i} e^{-t_j/\beta_i}}{\alpha_i \beta_i^{\alpha_i - 1} \Gamma(\alpha_i, t_j/\beta_i)} \right] \tag{7}$$

The RHS of (6) can be evaluated at the current parameter value to obtain the updated parameter estimate $\beta_i^{(g+1)}$. Upon expanding the incomplete gamma function as an infinite series, then differentiating and simplifying the expression, (4) can be expressed as

$$\frac{\partial Q}{\partial \alpha_i} = \sum_{j=1}^{n} z_{ij}^{g} \partial_j \left[\log t_j - \log \beta_i - \Psi(\alpha_i) \right] + \sum_{j=1}^{n} z_{ij}^{g} (1-\partial_j) \left[\log(t_j/\beta_i) - \log(t_j/\beta_i) \Big/ \left\{ 1 - e^{-t_j/\beta_i} \sum_{p=0}^{\infty} \frac{(t_j/\beta_i)^{\alpha_i + p}}{\Gamma(\alpha_i + p + 1)} \right\} + e^{-t_j/\beta_i} \sum_{p=0}^{\infty} \frac{(t_j/\beta_i)^{\alpha_i + p} \Psi(\alpha_i + p + 1)}{\Gamma(\alpha_i + p + 1)} \Big/ \left\{ 1 - e^{-t_j/\beta_i} \sum_{p=0}^{\infty} \frac{(t_j/\beta_i)^{\alpha_i + p}}{\Gamma(\alpha_i + p + 1)} \right\} \right] \tag{8}$$

Equating (7) to zero, the equation can be solved numerically for α_i to obtain the current estimate $\alpha_i^{(g+1)}$ by using $\beta_i^{(g+1)}$ for β_i .

The E-step on the *(g+1)th* iteration is to update the current conditional expectation of z_{ij} , given the observed data, using the current model parameters fit,

$$\hat{z}_{ij}^{(g+1)} = \frac{\hat{\pi}_i^{(g)} \left[f_i(t_j; \hat{\alpha}_i^{(g)}, \hat{\beta}_i^{(g)} \right]^{\delta_j} \left[S_i(t_j; \hat{\alpha}_i^{(g)}, \hat{\beta}_i^{(g)} \right]^{1-\delta_j}}{\sum_{i=1}^{k} \hat{\pi}_i^{(g)} \left[f_i(t_j; \hat{\alpha}_i^{(g)}, \hat{\beta}_i^{(g)} \right]^{\delta_j} \left[S_i(t_j; \hat{\alpha}_i^{(g)}, \hat{\beta}_i^{(g)} \right]^{1-\delta_j}} \tag{9}$$

The E-step and M-step iterate alternatively till the convergence criterion is met.

V. Data analysis

The performance of the proposed model was investigated by employing simulated data generated from survival mixture model of Gamma distributions. Three censoring percentages (10%, 20% and 40%) with two different sets of mixing proportions for the three components were considered to evaluate the model. The first set of mixing proportion in ascending order (10% , 40% and 50%) and the second one in descending order (50%, 30% and 20%). Survival data of size 500 observations were generated based on each of the three censoring percentages and the two sets of the mixing probabilities. The parameters of the first component Gamma distribution are (α_1=40, β_1=20) respectively, the parameters for the second component Gamma

distribution are (α_2=6 , β_2=1) and the parameters of the third component Gamma distribution are (α_3=200, β_3=20). Samples of size 500 were generated from the Exponential distribution for the censored time C with (b), where the value of b depends solely of the percentage of the observations that are censored. In this study 10%, 20% and 40% censoring observations were considered for each of the sample generated in which, tj =min(Tj,Cj) was taken as the minimum of the survival time and the censored time of the observed time T where

$$T = \begin{cases} \delta_i = 1, & if \quad X \le C, \\ \delta_i = 0, & if \quad X > C. \end{cases}$$

The proposed model corresponding to mixing proportions in ascending order was formed by substituting the values of the parameters mentioned earlier. Thus,

$f(t) = 0.1 * f_G(t; \alpha_1 = 40, \beta_1 = 20) + 0.4 * f_G(t; \alpha_2 = 6, \beta_2 = 1) + 0.5 * f_G(t; \alpha_3 = 200, \beta_3 = 20)$, where the density functions f_G represent the Gamma probability density functions of the three components respectively.

The simulated data were used to estimate the parameters of the proposed model by employing the EM. Table 1 displays the result of the estimates of the parameters of the proposed model for the three different censoring percentages with mixing proportions in ascending order.

Table 1: The Estimated Parameters of the Simulated Data

Sample size 500 observations and 10% censoring								
Parameter	π_1	π_2	α_1	α_2	α_3	β_1	β_2	β_3
Postulate	0.10	0.40	40	6	200	20	1	20
Estimates	0.09	0.38	40.00	6.00	200.00	19.52	0.96	19.91
Sample size 500 observations and 20% censoring								
Parameter	π_1	π_2	α_1	α_2	α_3	β_1	β_2	β_3
Postulate	0.10	0.40	40	6	200	20	1	20
Estimates	0.09	0.37	40.00	6.01	200.00	19.60	0.94	19.74
Sample size 500 observations and 40% censoring								
Parameter	π_1	π_2	α_1	α_2	α_3	β_1	β_2	β_3
Postulate	0.10	0.40	40	6	200	20	1	20
Estimates	0.08	0.32	40.06	4.79	199.96	19.59	0.71	19.36

The parameters of the three samples with different censoring percentages were estimated successfully. Table 1 showed that the estimated parameters are all close to the postulated parameters used in the data generation. Also the parameter for the simulated set of data with 10% censoring are more closer the true parameters compared to that of the 20% and 40% censored observations. The probability density function of the simulated data of the proposed model, with 10%, 20%, 40% censoring percentages respectively, and the probability density functions of pure classical survival model (G1, G2 and G3) corresponding to the components of the proposed model are displayed in Figures 1, 2 and 3.

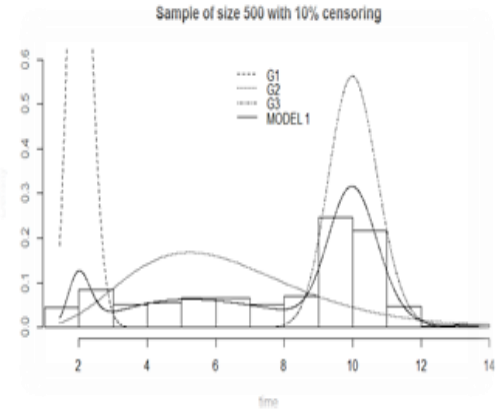

Figure 1: Density Function of the Simulated Data with 10% Censored Observations.

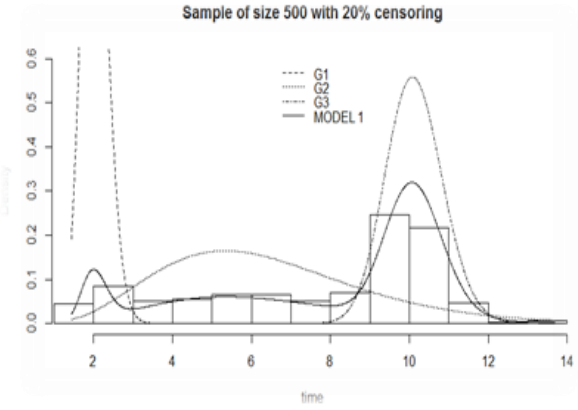

Figure 2: Density Function of the Simulated Data and 20% Censoring.

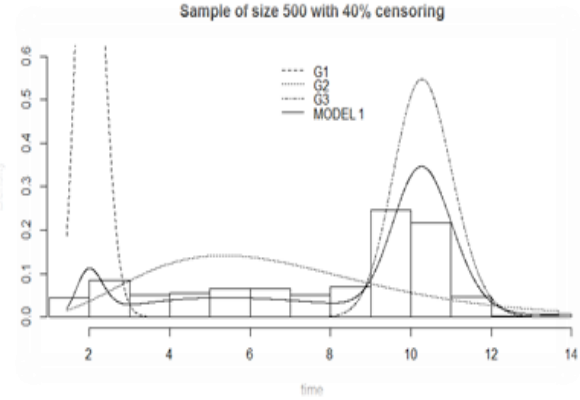

Figure 3: Density Function of the Simulated Data and 40% Censoring.

It can be seen that Model 1 fits the simulated data better than the individual pure classical parametric survival models which indicates that the simulated data is better modelled by the proposed Model than the pure classical parametric survival model.

The simulation of the three sets of the generated data with 10%, 20% and 40% censored observations were repeated 300 times to check the consistency and stability of the proposed model. The averages, the mean square errors (MSE) and root mean square error (RMSE) of estimated parameters were listed in Table 2.

The averages of the estimated parameters are close to the parameters of the postulated model with MSE and RMSE relatively small, which suggests that, the EM performed consistently in estimating the parameters. The MSE and RMSE corresponding to the mixing probabilities are relatively smaller for the 10% censoring as compared to the 20% and 40% censoring. Also the MSE for the parameters of the components are smaller for the 10% censoring compared to that of the 20% and 40%.

Table 2: The Repeated Simulation of Set of 500 Observations

Model 1 with sample size 500 and 10% censoring								
parameters	π_1	π_2	α_1	α_2	α_3	β_1	β_2	β_3
Postulates	0.10	0.40	40	6	200	20	1	20
Estimates	0.10	0.38	40.59	5.61	198.89	20.27	0.92	19.75
MSE	1.44e-7	5.57e-7	7.40e-02	5.91e-4	9.01e-2	1.74e-2	2.05e-5	9.16e-4
RMSE	3.80e-4	7.46e-4	0.27298	0.02430	0.30013	0.13190	0.00453	0.03026
Model 1 with sample size 500 and 20% censoring								
parameters	π_1	π_2	α_1	α_2	α_3	β_1	β_2	β_3
postulated	0.10	0.40	40	6	200	20	1	20
Estimates	0.10	0.37	41.40	5.18	195.67	20.68	0.83	19.29
MSE	1.43e-7	6.23e-7	1.53e-1	6.32e-4	3.40e-1	3.63e-2	2.15e-5	3.37e-3
RMSE	3.79e-4	7.89e-4	3.91e-01	2.51e-2	5.83e-1	1.91e-1	4.64e-3	5.80e-2
Model 1 with sample size 500 and 40% censoring								
parameters	π_1	π_2	α_1	α_2	α_3	β_1	β_2	β_3

postulated	0.10	0.40	40	6	200	20	1	20
estimates	0.08	0.34	40.86	4.29	194.03	20.45	0.65	18.81
MSE	1.53e-7	8.23e-7	1.56e-1	6.23e-4	8.50e-1	3.85e-2	1.85e-4	8.12e-3
RMSE	3.91e-4	9.01e-4	3.94e-1	2.29e-2	9.22e-1	1.96e-01	1.36e-2	9.01e-2

The averages of the parameters are close to the parameters of the postulated parametric survival mixture model with MSE relatively small, which suggests that, the EM performed consistently in estimating the parameters. Generally, the estimation of the mixing probabilities and the parameters are seemed to be closer to the true value with smaller censoring percentage 10% than with 20% and 40%.
The hazard functions of the three simulated data corresponding to the 10%, 20% and 40% censoring percentages were presented in Figure 4.

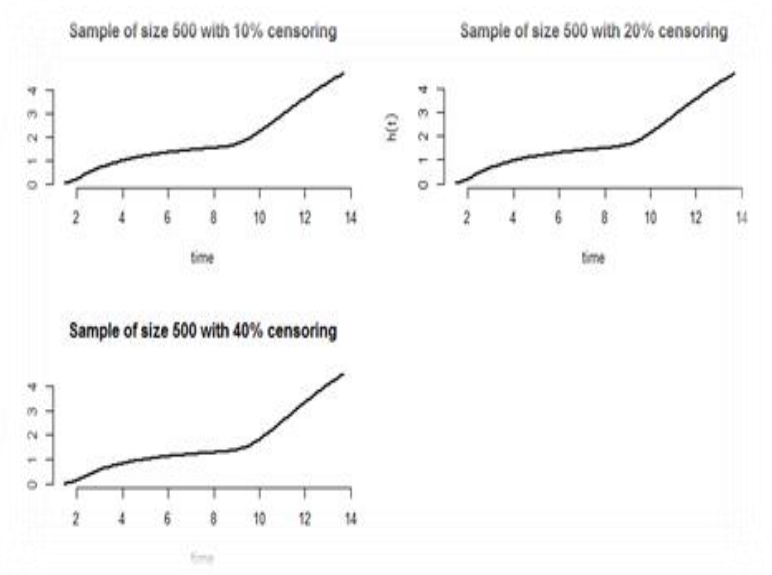

Figure 4: The Hazard Functions of the Simulated Data Corresponding to 10%, 20% and 40% Censored Observation.

The hazard function of the set of simulated data with 10% censoring observations is higher when compared with that of 20% and 40% censoring. As the number of censored observations increases the hazard tends to be lower and lower.

5.2 **Mixing probabilities in descending order**
The proposed model corresponding to mixing probabilities in descending order was formed by substituting the values of the parameters mentioned earlier. Thus,
$f(t) = 0.5 \times f_{G1}(t; \alpha_1 = 40, \beta_1 = 20) + \times f_{G2}(t; \alpha_2 = 6, \beta_2 = 1) + 0.2 \times f_{G3}(t; \alpha_3 = 200, \beta_3 = 20)$,
where the density functions f_{G1}, f_{G2} and f_{G3} represent the Gamma distribution for the first, second and third component of the model respectively.
The simulated data were employed to evaluate the parameters of the proposed model. The mixing proportions considered are in descending order. Table 3 displays the result of the estimates of the parameters of the proposed model for the three different censoring percentages.

Table 3: The Estimated Parameters the Simulated Data of size 500 Observations

Sample size 500 observations and 10% censoring								
Parameter	π_1	π_2	α_1	α_2	α_3	β_1	β_2	β_3
Postulate	0.50	0.30	40	6	200	20	1	20
Estimates	0.48	0.30	40.00	6.00	200.00	20.05	1.00	19.70
Sample size 500 observations and 20% censoring								
Parameter	π_1	π_2	α_1	α_2	α_3	β_1	β_2	β_3
Postulate	0.50	0.30	40	6	200	20	1	20
Estimates	0.46	0.28	40.00	6.00	200.00	19.63	1.00	19.64
Sample size 500 observations and 40% censoring								
Parameter	π_1	π_2	α_1	α_2	α_3	β_1	β_2	β_3

Postulate	0.50	0.30	40	6	200	20	1	20
Estimates	0.43	0.26	40.01	4.48	199.96	20.17	0.68	19.16

The parameters for the three sets of the simulated data were estimated successfully. From Table 3, it can be seen that the estimated parameters are all close to the postulated parameters used in the data generation. Also the parameter for the simulated set of data with 10% censoring are more closer the true parameters compared to that of the 20% and 40% censored observations.

The estimation of the mixing proportions was more accurate in sample with 10% censoring. The probability density function of the simulated data of the proposed model, with 10%, 20%, and 40% censoring percentages respectively, and the probability density functions of pure classical survival model (G1, G2 and G3) corresponding to the components of the proposed model are displayed in Figures 5,6 and 7.

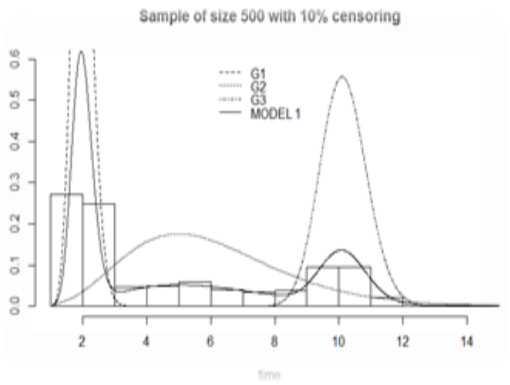

Figure 5: Density Function of the Simulated Data with 10% Censored Observations.

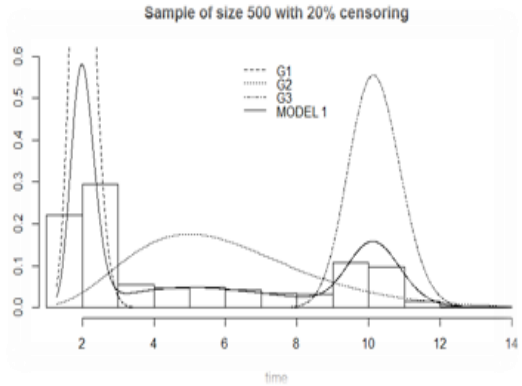

Figure 6: Density Function of the Simulated Data and 20% Censored Observations.

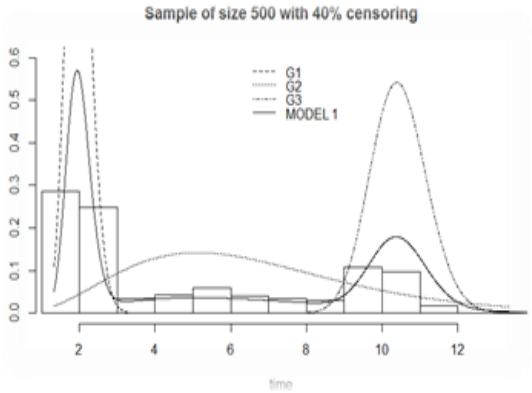

Figure 7: Density Function of the Simulated Data and 40% Censoring.

The simulation of the three sets of the generated data with 10%, 20% and 40% censored observations were repeated 300 times to check the consistency and stability of the proposed model. The averages, the mean square errors (MSE) and root mean square error (RMSE) of estimated parameters were listed in Table 4.

Table 4: The Repeated Simulation of Set of 500 Observations

Model 1 with sample size 500 and 10% censoring								
Parameters	π_1	π_2	α_1	α_2	α_3	β_1	β_2	β_3
Postulates	0.50	0.30	40	6	200	20	1	20
Estimates	0.48	0.29	40.37	5.46	199.49	20.18	0.88	19.74
MSE	2.29e-7	5.68e-7	2.17e-2	7.47e-4	4.77e-2	5.33e-3	2.75e-6	5.44e-4
RMSE	4.79e-4	7.54e-4	1.47e-01	2.73e-2	2.18e-1	7.30e-2	1.66e-3	2.33e-2
Model 1 with sample size 500 and 20% censoring								
Parameters	π_1	π_2	α_1	α_2	α_3	β_1	β_2	β_3
Postulated	0.50	0.30	40	6	200	20	1	20
Estimates	0.47	0.29	40.25	4.6	198.39	20.11	0.71	19.38
MSE	2.62e-7	3.84e-6	2.35e-2	6.15e-3	2.27e-1	5.73e-3	2.08e-5	2.24e-3
RMSE	5.12e-4	1.96e-4	1.53e-1	2.48e-2	4.77e1	7.57e-2	4.56e-3	4.73e-2
Model 1 with sample size 500 and 40% censoring								
Parameters	π_1	π_2	α_1	α_2	α_3	β_1	β_2	β_3
Postulated	0.50	0.30	40	6	200	20	1	20
Estimates	0.43	0.26	39.99	3.79	196.18	20.01	0.54	18.63
MSE	5.25e-7	9.56e-7	3.14e-2	5.96e-3	5.75e-1	7.68e-3	2.31e-5	5.36e-3
RMSE	7.24e-4	9.78e-4	1.77e-1	2.44e-2	7.58e-1	8.77e-2	4.80e-3	7.32e-2

The averages of the parameters are close to the parameters of the postulated with MSE and RMSE relatively small, which suggests that, the EM performed consistently in estimating the parameters. The MSE corresponding to the mixing probabilities are relatively smaller for the 10% censoring as compared to the 20% and 40% censoring. Also the MSE for the parameters of the components are smaller for the 10% censoring compared to that of 20% and 40%. Generally, the estimation of the mixing probabilities and the parameters are seemed to be closer to the true value with smaller censoring percentage 10% than with 20% and 40%.
The hazard functions of the three simulated data corresponding to 10%, 20% and 40% censoring percentages were presented in Figure 8.

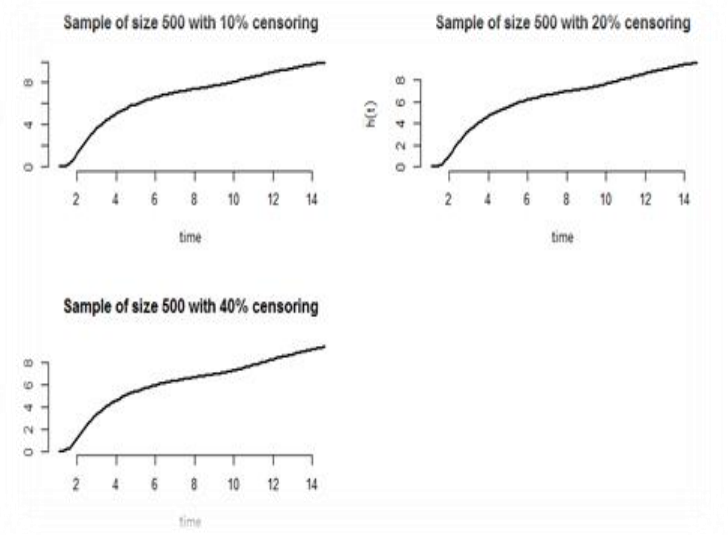

Figure 8: The Hazard Functions of the Simulated Data Corresponding to 10%, 20% and 40% Censored Observations.

The hazard function of the set of simulated data with 10% censoring observation is higher when compared with that of 20% and 40% censoring. As the number of censored observations increases the hazard tends to be decrease.

The estimation of the parameters of the model was successful for both the ascending and descending order of the mixing probabilities. For both the sets of mixing probabilities the estimation of parameters were closer the true postulate parameters which indicates the stability of the proposed model. It is also observed that

the estimates of the parameters were much better for small censoring percentages. The estimation of the mixing proportions for the ascending order was better than that of the descending with relatively small values for MSE. In general, it was observed that the mixing probabilities of ascending order performed better than the descending order as the censoring percentages increase.

VI. Conclusion

The study proposed a survival mixture model of Gamma distributions, with three components to model heterogeneous survival data. Simulated data were employed to investigate and assess the performance of the proposed model. The EM algorithm was employed in estimating the maximum likelihood estimators of the parameters of the model. The simulated data used to compare the effect of different censoring percentages showed that the model performed much better with small percentage of censored observations. It was also observed that the model performed well with both the ascending and descending order of the mixing proportions. However the model with mixing proportions in ascending order performed better the descending order. Samples with higher percentage of censored observations seemed to have lower hazard compared to the smaller censored observations. The proposed model showed that the survival mixture models are flexible and maintain the feature of pure classical parametric survival models and they are better options to model heterogeneous survival data.

References

[1]. Ibrahim, J. G., Chen, M. H., Sinha, D., *Bayesian survival analysis*. New York: Springer-verlag.ISBN 0-387-95277-2.
Jiang, S., Kececioglu, D., 1992a, Graphical representation of two mixed-Weibull distributions. *IEEE Transaction on Reliability*, vol. 41, 2001, 241-247, 2001. http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=257789 10.1109/24.257789

[2]. Kalbfleisch J. D., Prentice R. L., *The statistical analysis of failure time data* (2nd ed.). John Wiley & Sons, Inc. Hoboken, New Jersey. 2002, ISBN 0-471-36357-X.

[3]. Lawless J. F., *Statistical models and methods of lifetime data*, (2dn ed.). John Wiley and Sons, Inc. Hoboken, New Jersey. 2003, ISBN 0-471-37215-.

[4]. Lee, E. T., Wang, J. W., *Statistical methods for survival time data analysis* (3rd ed.). John Wiley & son. New Jersey.,2003, ISBN 0-471-36997-7.

[5]. Jiang, S., Kececioglu, D., Graphical representation of two mixed-Weibull distributions. *IEEE Transaction on Reliability*, vol. 41, 1992, 241-247. http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=257789 10.1109/24.257789

[6]. Jiang, S., Kececioglu, D., Maximum likelihood estimates, from censored data, for mixed-Weibull distributions. *IEEE Transaction on Reliability*, vol. 41, 1992, 248-255. http://www.dx.doi.org/ 10.1109/24.257791

[7]. Zhang Y., Parametric mixture models in survival analysis with application, (Doctoral Dissertation) UMI Number: 3300387, Graduate School, Temple University, 2008.
http://proquest.umi.com/pqdlink?did=1472138111&Fmt=7&clientI

[8]. Erişoğlu, Ü., Erişoğlu, M., Erol, H., Mixture model approach to the analysis of heterogeneous survival time data. *Pakistan Journal of Statistics 28*(1), 2012, 115-130. www.pakjs.com/journals/28(1)/28(1)8.pdf

[9]. Erişoğlu, Ü., Erol, H. , Modelling heterogeneous survival data using mixture of extended exponential-geometric distributions. *Communications in Statistics - Simulation and Computation, 39*(10), 2010, 1939-1952. http://www.dx.doi.org/ 10.1080/03610918.2010.524335

[10]. Erişoğlu, Ü., Erişoğlu, M., Erol, H., A mixture model of two different distributions approach to the analysis of heterogeneous survival data. *International Journal of Computational and Mathematical Sciences, 2011, 5: 2.* http://www.scopus.com/inward/record.url?eid=2-s2.0-78449258657&partnerID=40&md5=901faa5759d0767b0b2676000e17839c

[11]. Blackstone, E. H., Naftel, D. C., Turner M. E., The decomposition of time-varying hazard into phases, each incorporating a separate stream of concomitant information. *Journal of the American Statistical Association, 81*(395),1986, 615-624. http://www.jstor.org/stable/2288989

[12]. Ng, A. S. K., McLachlan, G. J., Yau, K. K. W., Lee, A. H., Modelling the distribution of ischaemic stroke-specific survival time using an EM-based mixture approach with random effects adjustment. *Statistics in Medicine, 23*(17), 2004, 2729-2744. http://www.scopus.com/inward/record.url?eid=2-s2.0-4444257472&partnerID=40&md5=139f3273239b2d5525b68c728faf99e3

[13]. Phillips, N., Coldman, A., McBride, M. L., Estimating cancer prevalence using mixture models for cancer survival. *Statistics in Medicine, 21*(9), 2002, 1257-1270. http://dx.doi.org/10.1002/sim.1101 DO - 10.1002/sim.1101

[14]. Mohammed, Y. A., Yatim, B., Ismail, S., A simulation study of parametric mixture model of three different distributions to analyse heterogeneous survival data. Modern Applied Science, 7(7),2013, 1-9 http://dx.doi.org/10.5539/mas.v7n7p1.

[15]. Mohammed, Y. A., Yatim, B., Ismail, S., A parametric Mixture Model of Three Different distributions: An approach to Analyse Heterogeneous Survival Data. Proceedings of the 21st National Symposium on Mathematical Sciences (SKSM21) AIP Conf. Proc. 1605, 1040-1045 (2014); doi: 10.1063/1.4887734.

[16]. Marín, J. M., Rodríguez-Bernal, M. T., Wiper, M. P., Using Weibull mixture distributions to model heterogeneous survival data. *Communications in Statistics: Simulation and Computation, 34*(3), 2005,673-684. http://www.researchgate.net/publication/4849603_Using_Weibull_Mixture_Distributions_To_Model_Heterogeneous_Survival_Data

[17]. Dempster, A. P. Laird, N. M., Rubin, D. B., Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion)". *Journal of Royal Statistical Society. Series B 1977, 39, 1-38.* http://www.jstor.org/stable/2984875

[18]. McLachlan, G. J., Krishnan, T., *The EM algorithm and extensions* (2nd ed.). (Hoboken New Jersey: John Wiley & Sons, Inc. ISBN 978-0-471-20170-0,2008).

[19]. McLachlan, G. J., Peel, D., *Finite mixture models* (John Wiley & Sons, Inc. New York. ISBN 0-471-00626-2, 2000).

[20]. Fruhwirth-Schnatter, S., *Finite mixture and markovs switching models* (Springer. New York. ISBN013:978-0387-32909-3, 2006.