International Journal of Science and Research (IJSR)
ISSN: 2319-7064
ResearchGate Impact Factor (2018): 0.28 | SJIF (2018): 7.426

Mixture Model of Different Distributions: A Simulation Study with Different Censoring and Mixing Probabilities

Yusuf Abbakar Mohammed[1], Suzilah Ismail[2]

[1]University of Maiduguri, Faculty of Science, Department of Mathematical Sciences, Bama Road, PMB 1069, Maiduguri, Nigeria

[2]Universiti Utara Malaysia School of Quantitative Sciences, Sintok, Kedah, Malaysia

Abstract: ***Survival mixture model of three different distributions was proposed. The model consists of a mixture of Exponential, Gamma and Weibull distributions. Simulated data was employed to investigate the performance of the model by considering three different censoring percentages and two sets of mixing probabilities in ascending order and descending order. The simulated data were used to estimate the maximum likelihood estimators of the model by employing Expectation Maximization (EM). Hazard functions corresponding to the censoring percentages were investigated graphically. Parameters of the proposed model were estimated and were all close the values used in generating the data. Simulation was repeated 300 times and 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 percentages revealed that the model performed much better with small percentage of censored observations. Also the model performed well with both the ascending and descending order of the mixing probabilities. However, mixing probabilities in ascending order performed better than the descending order. The hazard function graphs showed that, samples with higher percentage of censored observations seemed to have lower hazard compared to the smaller censored observations. The proposed model showed that survival mixture models are flexible and maintain the features of the pure classical survival model and are better option for modelling heterogeneous survival data.***

Keywords: exponential, gamma, mixing probability, mixture model, Weibull.

1. Introduction

Survival analysis commonly employed to analyse some event that occurs within a particular period of time. The methods of survival analysis are widely used in different fields such as engineering, biological sciences, sociology, economic and engineering to mention few. The nonparametric methods are frequently used to analyse survival data. Pure classical parametric survival models are very powerful methods in survival analysis; they perform better than the nonparametric methods when the chosen distribution fit the data properly. The Exponential, Gamma and Weibull distributions are frequently employed in analysing 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 proposed where the parameters of the model were estimated by the weighted least squares method [5]. A survival mixture model of Weibull distributions with two components was proposed, where the parameters of the model were estimated by graphical approach [6]. A new technique was developed for evaluating the parameters of a two components survival mixture model of Weibull distributions [7].

In another study, Expectation Maximization (EM) was employed to evaluate the parameters of a two-components survival mixture model of the Weibull-Weibull distributions, and the model stability was investigated [8] Two components survival mixture models of Gamma-Gamma, Weibul-Weibull and Lognormal-Lognormal distributions were used to model survival data [9], they implemented model selection technique to select the model which better represents the real data. A survival mixture of mixed distribution was proposed for analysing heterogeneous data. The model consists of two components of the Extended Exponential-Geometric (EEG) distribution [10]. Also a two components survival mixture model of different distributions consisting of an Exponential-Gamma, an Exponential-Weibull and a Gamma-Weibull distributions was employed for analysing heterogeneous survival data [11].

Three components survival mixture models did not receive much attention. A study to observe the risk of death after open-heart surgery was able to classify the risk of death after the surgery by three different time overlapping phases [12]. This type of survival data were better analysed by a three components mixture model [13]and [14]. A parametric survival mixture model of the Exponential, Gamma and Weibull distributions was proposed to model heterogeneous survival data. Simulated data were employed to investigate the stability and consistency of the model [15]. The method of model selection was employed to select the model which fit the data better [16]. Bayesian method was also implemented to analyse a three components survival mixture model of Weibull distributions [17]. In some situations where data consist of some missing or unobserved observations, Expectation Maximization (EM) is appropriate for analysing such data [18]. The Maximum Likelihood parameters of survival mixture models are commonly evaluated by implementing the EM [19]and [20].

In this study simulated data were generated and utilized to

investigate the flexibility and appropriateness of a three components survival mixture model of different distributions consisting of the Exponential, Gamma and Weibull distributions in modelling heterogeneous data. The arrangement of the paper is as follows. In section two the survival analysis and some properties of the Exponential, Gamma and Weibull distributions are highlighted. Section three devoted to discussing the survival mixture model of three components in the survival analysis. Section four highlights the employment of the EM in estimating the maximum likelihood parameters of the proposed model. Section five devoted to data application to evaluate the parameters of the proposed model and compare the different censoring percentages and the two sets of mixing probabilities. Section six devoted for summary and conclusion.

2. Survival Analysis and Probabilty Distributions

Survival analysis concern with the application of some statistical method to model and analyse survival data. The focus of interest is the occurrence of a particular event of interest within a given period of time. The response of 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 probability density function can also be presented graphically, the graph of *f (t)*, is known as the density curve. 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)* can be written as:

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

which gives the probability that an individual will survive beyond a particular time *t*. Note that 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 Exponential, Gamma and Weibull densities are commonly employed in the analysis of survival data. The probability density function $f(t)$ and survival functions $S(t)$ of these distributions are highlighted below.

Exponential Distribution

$$f_E(t)=\lambda e^{-\lambda t} \qquad t,\lambda>0$$

$$S_E(t)=e^{-\lambda t}$$

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.

Weibull Distribution

$$f_W(t)=\frac{\alpha}{\beta}\left(\frac{t}{\beta}\right)^{\alpha-1}\exp\left(-\left(\frac{t}{\beta}\right)^{\alpha}\right) \qquad t \text{ and } \alpha,\beta>0$$

$$S_W(t)=\exp\left(-\left(\frac{t}{\beta}\right)^{\alpha}\right)$$

3. Survival Mixture Model of Different 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 [21]. 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.

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

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 different distributions is proposed to model heterogeneous survival data. The proposed model consists of the Exponential, Gamma and Weibull distributions and is defined as

$$f_{E\_G\_W}(t;\Theta)=\pi_1 f_E(t;\lambda)+\pi_2 f_G(t;\alpha_1,\beta_1)+\pi_2 f_W(t;\alpha_2,\beta_2)$$

where π_i's are the mixing probability and $\sum_{i=1}^{3}\pi_i=1$. The functions f_E, f_G and f_W are the probability density functions of the Exponential, the Gamma and the Weibull distributions respectively corresponding to the components of model.

3.1 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 [20].

Let $t_1, t_2,\ldots,t_n$ be a set of observations of *n* incomplete data and z_1,z_2,z_3 be a set of missing observations, where $z_{ki}=z_K(t_i)=1$, if the observation belongs to the k^{th} component and 0 otherwise for $k=1,2,3$ and $i=1,\ldots,n$. 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

$$\hat{z}_{1i} = E(z_{1i} \mid t_i) = \frac{\pi_1 f_X(t_i;\theta_X)}{\pi_1 f_X(t_i;\theta_X)+\pi_2 f_Y(t_i;\theta_Y)+\pi_3 f_Q(t_i;\theta_Q)}$$

$$\hat{z}_{2i} = E(z_{2i} \mid t_i) = \frac{\pi_2 f_Y(t_i;\theta_Y)}{\pi_1 f_X(t_i;\theta_X)+\pi_2 f_Y(t_i;\theta_Y)+\pi_3 f_Q(t_i;\theta_Q)}$$

$$\hat{z}_{3i} = E(z_{3i} \mid t_i) = \frac{\pi_3 f_Q(t_i;\theta_Q)}{\pi_1 f_X(t_i;\theta_X)+\pi_2 f_Y(t_i;\theta_Y)+\pi_3 f_Q(t_i;\theta_Q)}$$

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 $\pi_i's$ equals to 1. The evaluation of the mixing probabilities $\pi_i's$ and vector of parameter $\theta=[\theta_X,\theta_Y,\theta_Q]$, is by the implementation of the Lagrange method. The mixing probabilities will be obtained by; $\hat{\pi}_1 = \sum_{i=1}^{n}\hat{z}_{1i}/n$, $\hat{\pi}_2 = \sum_{i=1}^{n}\hat{z}_{2i}/n$ and $\hat{\pi}_3 = \sum_{i=1}^{n}\hat{z}_{3i}/n$.

The proposed model can be expressed as defined above: where $f_E(t;\lambda)$ with unknown parameter λ, $f_G(t;\alpha_1,\beta_1)$ with unknown parameters α_1,β_1 and $f_W(t;\alpha_2,\beta_2)$ with unknown parameters α_2,β_2 are the Exponential, Gamma and Weibull density functions respectively. The parameters satisfy the conditions $\lambda>0$, $\alpha_1>0,\beta_1>0$, $\alpha_2>0,\beta_2>0$.

The log-likelihood function of the complete-data of the mixture of the Exponential, Gamma and Weibull distributions is:

$$\log L_c(t;\lambda,\alpha_1,\beta_1,\alpha_2,\beta_2,\pi_i) = \sum_{j=1}^{n} z_{1j}\left[\log\pi_{1i} + \delta_j \log(\lambda e^{-\lambda t_j}) + (1-\delta_j)\log(e^{-\lambda t_j})\right]$$

$$+\sum_{j=1}^{n} z_{2j}\left\{\log\pi_2 + \delta_j \log\left[\frac{1}{\beta_1\Gamma(\alpha_1)}\left(\frac{t_j}{\alpha_1}\right)^{\alpha_1-1} e^{\frac{-t}{\alpha_1}}\right] + (1-\delta_j)\log\left[\frac{\Gamma(\alpha_1,t_j/\beta_1)}{\Gamma\alpha_1}\right]\right\}$$

$$+\sum_{j=1}^{n} z_{3j}\left\{\log\pi_3 + \delta_j \log\left[\left(\frac{\alpha_2}{\beta_2}\right)\left(\frac{t_j}{\beta_2}\right)^{\alpha_2-1} e^{-\left(\frac{t_j}{\beta_2}\right)^{\alpha_2}}\right]\right.$$

The EM 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

$$Q(t;\lambda,\alpha_1,\beta_1,\alpha_2,\beta_2,\pi_i) = \sum_{j=1}^{n} z_{1j}^{(g)}\left[\log\pi_1 + \delta_j(\log\lambda - \lambda t_j) - (1-\delta_j)\lambda t_j\right]$$

$$+\sum_{j=1}^{n} z_{2j}^{(g)}\left\{\log\pi_2 + \delta_j \log\left[\frac{1}{\beta_1\Gamma(\alpha_1)}\left(\frac{t_j}{\alpha_1}\right)^{\alpha_1-1} e^{\frac{-t}{\alpha_1}}\right] + (1-\delta_j)\log\left[\frac{\Gamma(\alpha_1,t_j/\beta_1)}{\Gamma\alpha_1}\right]\right\}$$

$$+\sum_{j=1}^{n} z_{3j}^{(g)}\left\{\log\pi_3 + \delta_j \log\left[\left(\frac{\alpha_2}{\beta_2}\right)\left(\frac{t_j}{\beta_2}\right)^{\alpha_2-1} e^{-\left(\frac{t_j}{\beta_2}\right)^{\alpha_2}}\right] + (1-\delta_j)\left[-\left(\frac{t_j}{\beta_2}\right)^{\alpha_2}\right]\right\} \quad (2)$$

The M-step on the $(g+1)^{th}$ iteration requires the global maximization of Equation (2) with respect to $\lambda,\alpha_1,\beta_1,\alpha_2,\beta_2$ and π_i. The mixing probabilities π_i can be updated by $\hat{\pi}_i^{(g+1)} = \sum_{j=1}^{n} z_{ij}^{(g)}/n,\ i=1,2,3$. In order to get the updated maximum likelihood estimate of the model parameters $\lambda,\alpha_1,\beta_1,\alpha_2,\beta_2$, Equation (2) will be differentiated with respect to each of these parameters.

Now, differentiating Equation (2) with respect to the parameter λ the updated maximum likelihood estimate of the first component model parameter can be obtained in closed form

$$\hat{\lambda}^{(g+1)} = \frac{\sum_{j=1}^{n} z_{1j}^{(g)} t_j}{\sum_{j=1}^{n} z_{1j}^{(g)}\delta_j}$$

This completes the M-step. The E-step on the $(g+1)^{th}$ iteration is to update the current conditional expectation of Z_{1j}, given the observed data, using the current model parameters fit,

$$\hat{z}_{1j}^{(g+1)} = \hat{\pi}_1^{(g)}\left[f_1(t_j;\hat{\lambda}^{(g)})\right]^{\delta_j}\left[S_i(t_j;\hat{\lambda}\right]^{1-\delta_j} / \left[\hat{\pi}_1^{(g)}\left[f\ (t_j;\hat{\lambda}^{(g)}\right]^{\delta_j}\left[S_i(t_j;\hat{\lambda}\right]^{1-\delta_j}\right.$$
$$+\hat{\pi}_2^{(g)}\left[f\ (t_j;\hat{\alpha}_1^{(g)},\hat{\beta}_1^{(g)}\right]^{\delta_j}\left[S\ (t_j;\hat{\alpha}_1^{(g)},\hat{\beta}_1^{(g)}\right]^{1-\delta_j}$$
$$\left.+\hat{\pi}_3^{(g)}\left[f\ (t_j;\hat{\alpha}_2^{(g)},\hat{\beta}_2^{(g)}\right]^{\delta_j}\left[S(t_j;\hat{\alpha}_2^{(g)},\hat{\beta}_2^{(g)}\right]^{1-\delta_j}\right]$$

Again, differentiating Equation (2) with respect to the parameter α_1,β_1 yields:

$$\frac{\partial Q}{\partial\alpha_1} = \sum_{j=1}^{n} z_{2j}^{(g)}\delta_j\left[-\log\beta_1 + \Psi(\alpha_1) + \log t_j\right]$$
$$+\sum_{j=1}^{n} z_{2j}^{(g)}(1-\delta_j)\left[\log\beta_1 + \frac{1}{\Gamma(\alpha_1,t_j/\beta_1)}\frac{\partial}{\partial\alpha_1}\Gamma(\alpha_1,t_j/\beta_1)\right] \quad (3)$$

$$\frac{\partial Q}{\partial\beta_i} = \sum_{j=1}^{n} z_{2}^{(g)}j\left[-\delta_j\left(\frac{\alpha_1}{\beta_1}+\frac{t_i}{\beta_1^2}\right) + \frac{(1-\delta_j)}{\Gamma(\alpha_1,t_j/\beta_1)}\frac{\partial}{\partial\beta_1}\Gamma(\alpha_1,t_j/\beta_1)\right] \quad (4)$$

Now, the incomplete gamma function can be differentiated with respect to β_1 using Leibnitz's rule, and we then obtain from Equation(4) that:

$$\beta_1 = \left[\sum_{j=1}^{n} z_{2j}^{g} t_j/\alpha_1 + \sum_{j=1}^{n} z_{2j}^{g}\delta_j/\alpha_1 - \sum_{j=1}^{n}\frac{t_j^{\alpha_1}e^{-t_j/\beta_1}}{\alpha_1\beta_1^{\alpha_1-1}\Gamma(\alpha_1,t_j/\beta_1)}\right] \quad (5)$$

The RHS of Equation(5) can be evaluated at the current parameter value to obtain the updated parameter estimate $\beta_1^{(g+1)}$. Upon expanding the incomplete gamma function as an infinite series, then differentiating and simplifying the expression, Equation (3) can be expressed as:

$$\frac{\partial Q}{\partial\alpha_1} = \sum_{j=1}^{n} z_{2j}^{g}\partial_j\left[\log t_j - \log\beta_1 - \Psi(\alpha_1)\right]$$
$$+\sum_{j=1}^{n} z_{2j}^{g}(1-\partial_j)\left[\log(t_j/\beta_1) - \log(t_j/\beta_1)/\left\{1-e^{-t_j/\beta_1}\sum_{p=0}^{\infty}\frac{(t_j/\beta_1)^{\alpha_1+p}}{\Gamma(\alpha_1+p+1)}\right\}\right.$$

$$+e^{-t_j/\beta_1}\sum_{p=0}^{\infty}\frac{(t_j/\beta_1)^{\alpha_1+p}\Psi(\alpha_1+p+1)}{\Gamma(\alpha_1+p+1)}\Big/\left\{1-e^{-t_j/\beta_1}\sum_{p=0}^{\infty}\frac{(t_j/\beta_1)^{\alpha_1+p}}{\Gamma(\alpha_1+p+1)}\right\}\Bigg] \quad (6)$$

Setting equating (6) to zero, the equation can be solved numerically for α_1 to obtain the current estimate $\alpha_1^{(g+1)}$ by using $\beta_1^{(g+1)}$ for β_1 .

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,

$$\hat{Z}_{2j}^{(g+1)}=\hat{\pi}_2^{(g)}\left[f(t_j;\hat{\alpha}_1^{(g)},\hat{\beta}_1^{(g)}\right]^{\delta_j}\left[S(t_j;\hat{\alpha}_1^{(g)},\hat{\beta}_1^{(g)}\right]^{1-\delta_j}\Big/\left[\hat{\pi}_1^{(g)}\left[f(t_j;\hat{\lambda}^{(g)}\right]^{\delta_j}\left[S(t_j;\hat{\lambda}\right]^{1-\delta_j}\right.$$
$$+\hat{\pi}_2^{(g)}\left[f(t_j;\hat{\alpha}_1^{(g)},\hat{\beta}_1^{(g)}\right]^{\delta_j}\left[S(t_j;\hat{\alpha}_1^{(g)},\hat{\beta}_1^{(g)}\right]^{1-\delta_j}$$
$$\left.+\hat{\pi}_3^{(g)}\left[f(t_j;\hat{\alpha}_2^{(g)},\hat{\beta}_2^{(g)}\right]^{\delta_j}\left[S(t_j;\hat{\alpha}_2^{(g)},\hat{\beta}_2^{(g)}\right]^{1-\delta_j}\right]$$

Again, differentiating equation (2) with respect to the parameters α_2, β_2 yields:

$$\frac{\partial Q}{\partial\alpha_2}=-\sum_{j=1}^{n}z_{3j}^{(g)}\delta_j\left[\frac{1}{\alpha_2}-\log\beta_2+\log t_j\right]-\sum_{j=1}^{n}z_{3j}^{(g)}\left(\frac{t_j}{\beta_2}\right)^{\alpha_2}(\log t_j-\log\beta_2) \quad (7)$$

$$\frac{\partial Q}{\partial\beta_2}=\sum_{j=1}^{n}z_{3j}^{(g)}\left[-\delta_j\frac{\alpha_2}{\beta_2}+\alpha_2 t_j^{\alpha_2}\beta_2^{-\alpha_2-1}\right] \quad (8)$$

should be solved for the values of the parameters α_2 and β_2 .

The system of Equations (8) can be written as:

$$\beta_2=\exp\left(\frac{1}{\alpha_2}\log\frac{\sum_{j=1}^{n}z_{3j}^{(g)}t_j^{\alpha_2}}{\sum_{j=1}^{n}z_{3j}^{(g)}\delta_j}\right) \quad (9)$$

Plug Equations (9) back to Equation(7) to obtain:

$$\sum_{j=1}^{n}z_{3j}^{(g)}\left[\frac{1}{\alpha_2}-\frac{1}{\alpha_2}\log\frac{\sum_{j=1}^{n}z_{3j}^{(g)}t_j^{\alpha_2}}{\sum_{j=1}^{n}z_{3j}^{(g)}\delta_j}+\log t_j\right]-$$
$$\sum_{j=1}^{n}z_{3j}^{(g)}t_j^{\alpha_2}\frac{\sum_{j=1}^{n}z_{3j}^{(g)}\delta_j}{\sum_{j=1}^{n}z_{3j}^{(g)}t_j^{\alpha}2}\left(\log t_j-\frac{1}{\alpha_2}\log\frac{\sum_{j=1}^{n}z_{3j}^{(g)}t_j^{\alpha_2}}{\sum_{j=1}^{n}z_{3j}^{(g)}\delta_j}\right)=0 \quad (10)$$

Then equations (10) can be solved to obtain the estimates for α_2. Plug the estimates of α_2 back to equations (9) to obtain the estimates for β_2. This completes the M-step. The E-step on the $(g+1)^{th}$ iteration is to update the current conditional expectation of Z_{3j}, given the observed data, using the current model parameters fit:

$$\hat{Z}_{3j}^{(g+1)}=\hat{\pi}_3^{(g)}\left[f(t_j;\hat{\alpha}_2^{(g)},\hat{\beta}_2^{(g)}\right]^{\delta_j}\left[S(t_j;\hat{\alpha}_2^{(g)},\hat{\beta}_2^{(g)}\right]^{1-\delta_j}\Big/\left[\hat{\pi}_1^{(g)}\left[f(t_j;\hat{\lambda}^{(g)}\right]^{\delta_j}\left[S(t_j;\hat{\lambda}\right]^{1-\delta_j}\right.$$
$$+\hat{\pi}_2^{(g)}\left[f(t_j;\hat{\alpha}_1^{(g)},\hat{\beta}_1^{(g)}\right]^{\delta_j}\left[S(t_j;\hat{\alpha}_1^{(g)},\hat{\beta}_1^{(g)}\right]^{1-\delta_j}$$
$$\left.+\hat{\pi}_3^{(g)}\left[f(t_j;\hat{\alpha}_2^{(g)},\hat{\beta}_2^{(g)}\right]^{\delta_j}\left[S(t_j;\hat{\alpha}_2^{(g)},\hat{\beta}_2^{(g)}\right]^{1-\delta_j}\right]$$

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

4. Data Analysis

The performance of the proposed model was investigated by employing simulated data generated from survival mixture model of Exponential, Gamma and Weibull distributions. Three censoring percentages (10%, 20% and 40%) with two different sets of mixing probabilities for the three components were considered to evaluate the model. The first set of mixing probabilities 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 parameter considered for the first component of Exponential distribution is $\lambda=1.5$, the parameters for the second component of Gamma distribution are ($\alpha_1=5$, $\beta_1=2$) and the parameters of the third component of Weibull distribution are ($\alpha_2=9$, $\beta_2=10$). 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, t_j=min(T_j,C_j) was taken as the minimum of the survival time and the censored time of the observed time T where

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

4.1 Mixing Probabilities in Ascending Order

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

$f(t)=0.1\times f_E(t;\lambda=1.5)+0.4\times f_G(t;\alpha_1=5,\beta_1=2)+0.5\times f_W(t;\alpha_2=9,\beta_2=10)$,

where the density functions f_E, f_G and f_W represent the Exponential, the Gamma and the Weibull probability density functions 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 probabilities in ascending order.

Table 1: The Estimated Parameters of the Simulated Data

Sample size 500 observations and 10% censoring							
Parameter	π_1	π_2	λ	α_1	α_2	β_1	β_2
Postulate	0.10	0.40	1.50	5.00	9.00	2.00	10.00
Estimates	0.11	0.40	1.52	5.25	9.00	2.03	10.02
Sample size 500 observations and 20% censoring							
Parameter	π_1	π_2	λ	α_1	α_2	β_1	β_2
Postulate	0.10	0.40	1.50	5.00	9.00	2.00	10.00
Estimates	0.10	0.40	1.49	4.81	9.00	2.04	10.00
Sample size 500 observations and 40% censoring							
Parameter	π_1	π_2	λ	α_1	α_2	β_1	β_2
Postulate	0.10	0.40	1.50	5.00	9.00	2.00	10.00

Estimates	0.10	0.39	1.48	4.84	9.00	1.99	9.94

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 (E, G and W) corresponding to the components of the proposed model are displayed in Figs 1, 2 and 3.

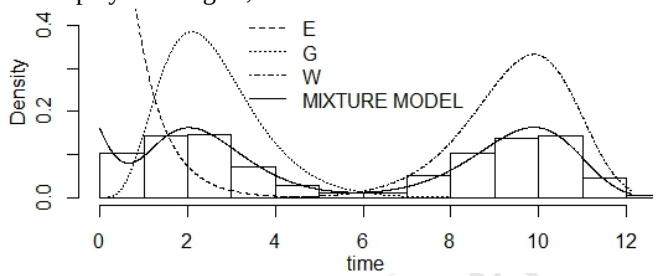

Figure 1: Density Function of the Simulated Data with 10% Censored Observations.

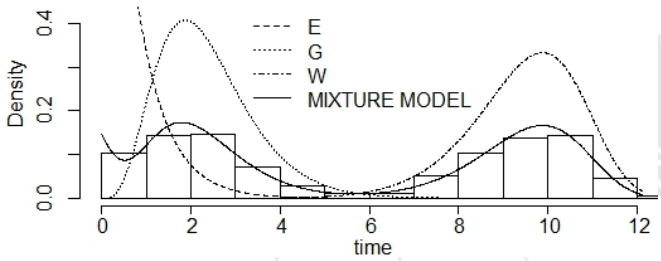

Figure 2: Density Function of the Simulated Data and 20% Censoring.

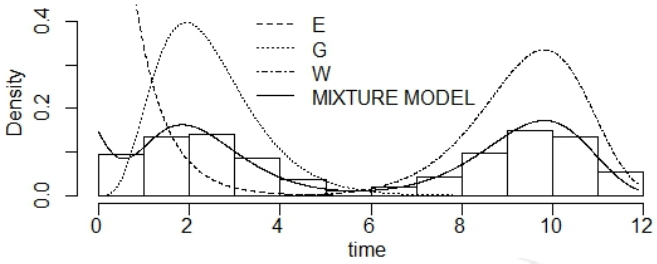

Figure 3: Density Function of the Simulated Data and 40% Censoring.

Table 2: The Repeated Simulation of Set of the Three Samples

Sample size 500 and 10% censoring							
Parameters	π_1	π_2	λ	α_1	α_2	β_1	β_2
Postulates	0.1	0.4	1.5	5.00	9.0	2.00	10.00
Estimates	0.11	0.40	1.45	4.90	9.00	2.00	10.01
MSE	1.84e-7	2.27e-8	2.34e-5	3.82e-4	0.00e+0	5.51e-7	1.50e-5
RMSE	4.30e-4	1.51e-4	4.48e-3	3.82e-4	0.00e+0	7.42e-4	3.87e-3
Sample size 500 and 20% censoring							
Parameters	π_1	π_2	λ	α_1	α_2	β_1	β_2
Postulated	0.1	0.4	1.5	5.00	9.00	2.00	10.00
Estimates	0.10	0.40	1.44	4.69	9.00	2.01	9.99
MSE	1.99e-7	1.78e-7	2.56e-5	4.22e-4	0.00e+0	5.80e-7	1.50e-5
RMSE	4.46e-4	4.21e-4	5.06e-3	2.05e-2	0.00e+0	7.61e-4	3.87e-3
Sample size 500 and 40% censoring							
Parameters	π_1	π_2	λ	α_1	α_2	β_1	β_2
Postulated	0.1	0.4	1.5	5.00	9.00	2.00	10.00
Estimates	0.09	0.39	1.36	4.64	9.00	2.04	9.96
MSE	2.01e-7	1.31e-7	2.38e-5	5.21e-4	0.00e+0	6.19e-7	1.50e-5
RMSE	4.49e-4	3.62e-4	4.88e-3	2.28e-2	0.00e+0	7.87e-4	3.88e-3

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%. 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 Fig. 4.

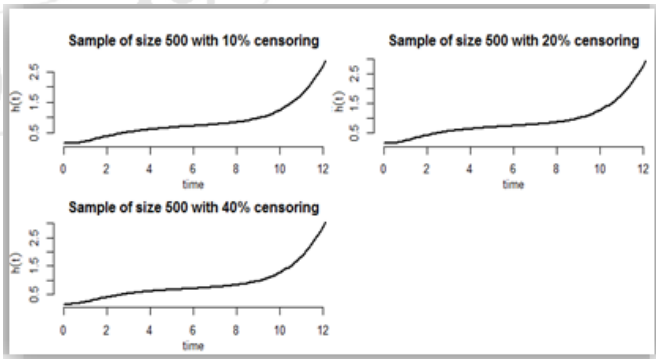

Figure 4: Hazard Functions of Simulated Data for 10%, 20% and 40% Censored Observation

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.

4.2 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:

$$f(t) = 0.5 \times f_E(t; \lambda = 1.5) + 0.3 \times f_G(t; \alpha_1 = 5, \beta_1 = 2) + 0.2 \times f_W(t; \alpha_2 = 9, \beta_2 = 10)$$

where the density functions f_E, f_G and f_W represent the Exponential, the Gamma and the Weibull probability density functions respectively. The simulated data were employed to evaluate the parameters of the proposed model. The mixing probabilities 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	λ	α_1	α_2	β_1	β_2
Postulate	0.50	0.30	1.5	5.00	9.00	2.00	10.00
Estimates	0.51	0.25	1.50	4.63	9.00	2.02	9.81
Sample size 500 observations and 20% censoring							
Parameter	π_1	π_2	λ	α_1	α_2	β_1	β_2
Postulate	0.50	0.30	1.5	5.00	9.00	2.00	10.00
Estimates	0.47	0.29	1.52	4.52	9.00	2.03	9.85
Sample size 500 observations and 40% censoring							
Parameter	π_1	π_2	λ	α_1	α_2	β_1	β_2
Postulate	0.50	0.30	1.5	5.00	9.00	2.00	10.00
Estimates	0.46	0.17	1.56	4.50	9.00	2.65	9.42

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 probabilities 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 (E, G and W) corresponding to the components of the proposed model are displayed in Figures 5,6 and 7.

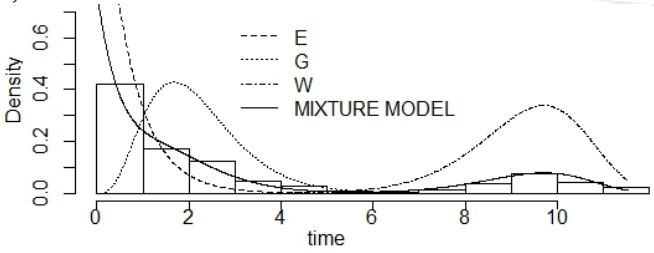

Figure 5: Density Function of the Simulated Data with 10% Censored Observations

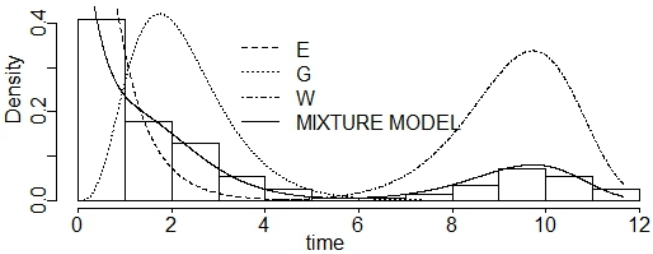

Figure 6: Density Function of the Simulated Data and 20% Censored Observations

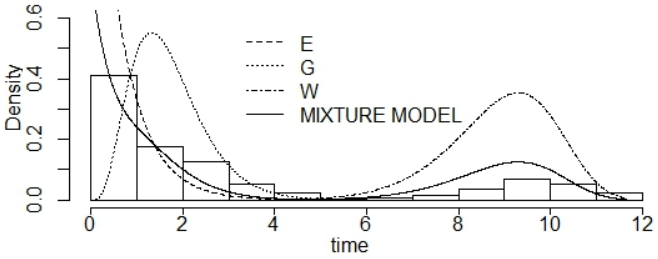

Figure 7: Density Function of the Simulated Data and 40% Censoring.

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.

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%.

Table 4: The Repeated Simulation of Set of 500 Observations

Sample size 500 and 10% censoring							
Parameters	π_1	π_2	λ	α_1	α_2	β_1	β_2
Postulates	0.50	0.30	1.5	5.00	9.00	2.00	10.00
Estimates	0.50	0.26	1.41	4.56	9.00	2.00	9.80
MSE	5.29e-7	2.41e-7	6.65e-6	5.16e-4	0.00e+0	5.34e-7	5.91e-5
RMSE	5.75e-4	4.91e-4	2.57e-3	2.27e-2	0.00+0	7.31e-4	6.25e-3
Sample size 500 and 20% censoring							
Parameters	π_1	π_2	λ	α_1	α_2	β_1	β_2
Postulates	0.50	0.30	1.5	5.00	9.00	2.00	10.00
Estimates	0.48	0.27	1.49	4.55	9.00	2.05	9.79
MSE	4.26e-7	2.65e-7	7.48e-6	5.44e-4	0.00e+0	6.65e-7	4.57e-5
RMSE	6.52e-4	5.12e-4	2.74e-3	2.33e-2	0.00e+0	8.15e-4	6.61e-3
Sample size 500 and 40% censoring							
Parameters	π_1	π_2	λ	α_1	α_2	β_1	β_2
Postulates	0.50	0.30	1.5	5.00	9.00	2.00	10.00
Estimates	0.47	0.17	1.55	4.46	9.00	2.60	9.59
MSE	5.75e-7	2.73e-7	1.28e-5	9.71e-4	0.00e+0	5.75e-6	4.10e-5
RMSE	7.58e-4	5.22e-4	3.58e-3	3.12e-2	0.00e+0	1.95e-3	6.40e-3

The hazard functions of the three simulated data corresponding to 10%, 20% and 40% censoring percentages were presented in Figure 8.

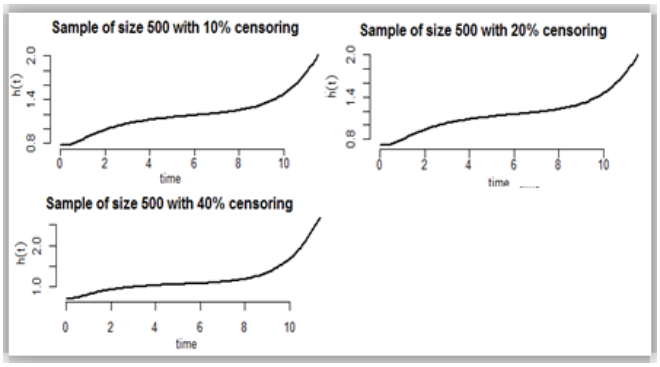

Figure 8: The Hazard Functions of Simulated Data for 10%, 20% and 40% Censored Observations.

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 probabilities 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.

5. Conclusion

The paper proposed a three components survival mixture model of different distributions, namely; the Exponential, the Gamma and the Weibull distributions to model heterogeneous survival data. Simulated data were used to evaluate and assess the performance of the proposed model. The EM algorithm was employed in estimating the maximum likelihood estimator of the parameters of the model. The simulated data used to compare the effect of different censoring percentages revealed 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 probabilities. However the model with mixing probabilities 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. 2001) ISBN 0-387-95277-2.

[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. ISBN 0-471-37215-3

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

[5] Cheng, S. W., Fu, J. C.. Estimation of mixed Weibull parameters in life testing.*Reliability, IEEE Transactions on, R-31*(4), 1982, 377-381. http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=522138210.1109/TR.1982.5221382

[6] Jiang, S., Kececioglu, D., Graphical representation of two mixed-Weibull distributions.*IEEE Transaction on Reliability*, vol. 41 , 1992a,241-247 http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=25778910.1109/24.257789

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

[8] 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

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

[10] Erişoğlu, Ü.,Erol, H. 2010, Modelling heterogeneous survival data using mixture of extended exponential-geometric distributions. *Communications in Statistics - Simulation and Computation, 39*(10), 2012, 1939-1952. http://www.dx.doi.org/ 10.1080/03610918.2010.524335

[11] 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 5: 2.*, 2011, http://www.scopus.com/inward/record.url?eid=2-s2.0-78449258657&partnerID=40&md5=901faa5759d0767b0b2676000e17839c

[12] 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

[13] 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

[14] 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

[15] 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.

[16] Mohammed, Y. A., Yatim, B., Ismail, S., AparametricMixtrue 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, 2014, 1040-1045 (2014); doi: 10.1063/1.4887734.

[17] 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

[18] 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, 39,* 1977, *1-38.* http://www.jstor.org/stable/2984875

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

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

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

Author Profile

Yusuf Abbakar Mohammed received the B.S., MSc. And PhD. degrees in Statistics from University of Maiduguri, 1991, University of Ibadan, 1994 and Universiti Utara Malaysia. Now he is a Senior lecturer and research with the University of Maiduguri, Borno State, Nigeria.