International Journal of Mathematical Research
2022 Vol. 11, No. 1, pp. 1-9.
ISSN(e): 2306-2223
ISSN(p): 2311-7427
DOI: 10.18488/24.v11i1.2924

A MIXTURE OF GAMMA-GAMMA, LOGLOGISTIC-GAMMA DISTRIBUTIONS FOR THE ANALYSIS OF HETEROGENOUS SURVIVAL DATA

Othman Musa Yakubu[1+]
Yusuf Abbakar Mohammed[2]
Akeyede Imam[3]

[1,2]*Department of Mathematical Sciences, University of Maiduguri, Nigeria.*
[1]*Email: othmanyakub@gmail.com Tel: +2348028281177*
[2]*Email: yusufabbakarm@gmail.com Tel: +23433903822*
[3]*Department of Mathematics, Federal University Lafia, Lafia, Nigeria.*
[3]*Email: imamakeyede@gmail.com Tel: +23439371735*

(+ Corresponding author)

Article History
Received: 8 December 2021
Revised: 17 January 2022
Accepted: 2 February 2022
Published: 15 February 2022

Keywords
Survival analysis
Log-logistic
Gamma
Expectation maximization
Algorithm
Akaike information criteria
Maximum likelihood estimate
Mean square error
Root mean square error.

ABSTRACT

Survival analysis deals with failure time data. The presence of censoring makes the application of the classical parametric and nonparametric methods of survival analysis inadequate and as such need's modifications. Parametric mixture models are applied where a single classical model may not suffice. The parametric mixture needs to be made more robust to address the heterogeneity of survival data. This paper proposed a mixture of two distributions for the analysis of survival data, the models consist of Gamma-Gamma, and Loglogistic-Gamma distributions. Data was simulated to investigate the performance of the models, and used to estimate the maximum likelihood parameters of the models by employing Expectation Maximization (EM). Parameters of the models were estimated and were all close the postulated values. Simulations were repeated to test the consistency and stability of the models through mean square error (MSE) and root mean square error (RMSE), and were all found to be stable and consistent. Real data was applied to determine the best fit among the mixture models and classical distributions using information criteria. Mixture models were found to model the data and the mixture of two different distributions gives the best fit.

Contribution/Originality: The study is significant in the sense that, all existing studies of mixture models need to be extended to enrich the study of the analysis of heterogeneous survival data, because of its wide application, such as, in biomedical sciences, industrial reliability or reliability engineering, social sciences, and business.

1. INTRODUCTION

Survival or reliability study is an area with its unique characteristic; it deals with the statistical methods of analysing survival data obtained from clinical studies of humans, laboratory study of animals and investigation of the durability of manufactured items, among other appropriate applications. Survival time can broadly be defined as the time to the occurrence of the event of interest, the event of interest can be the time to failure of a manufactured item, the time to occurrence of a disease, time to relapse, response to treatment, death, etc. The study of survival data has paid attention on predicting the probability of response, survival, or mean lifetime, comparing the survival distributions of experimental animals or, of human patients and the identification of risk and/or prognostic factors related to response, survival, and the development of a disease, [1]. Parametric and nonparametric methods are usually employed [2, 3].

Mixture models are being explored in survival and reliability analysis in recent times. Mixture models can be used to analyze failure-time data in a variety of ways. As a flexible way of modelling data, mixture models have several applications in situations where a single model may not suffice. They are applied where the data is heterogenous. Some authors proposed a mixture of three classical distributions [4] to analyze survival data that has three different time overlapping phases [5].

Similarly, two component mixture models of different distributions were also studied [6] therefore, this paper wish to enrich the study of two component, different distributions, mixture models for the analysis of survival data.

2. SURVIVAL ANALYSIS

Let *T* denote the survival time, which is a non-negative absolutely continuous random variable that represents the life time of individuals. If *F(t)* is the cumulative distribution of *T*, the survival function is defined to be;

S(t) = *P* (an individual survives longer than *t).*

$$= p(T > t)$$

$$= 1 - F(t) \qquad (1)$$

Equation 1 represent the survival function of an individual with survival time T.

2.1. The Probability Density Function

Similar to any other continuous random variable, the survival time *T* has a probability density function defined as the limit of the probability that an individual fails in the short interval *(t, t + Δt)* per unit width Δ*t*.

It can be expressed as;

$$f(t) = \lim_{\Delta \to t} \frac{P\{an\ individual\ dying\ in\ the\ interval(t,\ \ t{+}\Delta t)\}}{\Delta t} \qquad (2)$$

Equation 2 is the probability density function as explained above in 2.1.

2.2. The Hazard Function

The hazard function *h* (t) of survival time T gives the conditional failure rate; it is defined as the probability of failure during small interval of time, given that the individual has survived to the beginning of the interval. It can be expressed as;

$$H(t) = \lim_{\Delta \to 0} \frac{P\{an\ individual\ of\ age\ t\ fails\ in\ the\ interval(t,\ \ t{+}\Delta t)\}}{\Delta t} \qquad (3)$$

The hazard function *h* (t) can also be defined in terms of the cumulative distribution function *F* (t) and the probability density function *f* (t).

$$h(t) = \frac{f(t)}{1-F(t)} \qquad (4)$$

The hazard function is also known as the *force of mortality*, *conditional mortality rate*, and *age-specific failure rate* [1]. The survival function, *S*(t), probability density function, *f*(t), and hazard function, *h*(t), are mathematically equivalent. The relationship is expressed thus:

$$h(t) = \frac{f(t)}{S(t)} \qquad (5)$$

$$f(t) = \frac{d}{dt}[1 - S(t)] \qquad (6)$$

$$h(t) = \frac{S'(t)}{S(t)} = -\frac{d}{dt} loge\ S(t) \qquad (7)$$

$$S(t) = \exp\,[-\int_0^t h(x)dx] \qquad (8)$$

$$f(t) = h(t)\exp\left[-\int_0^t h(t)dx\right] \tag{9}$$

Equations 5 through 9 presents the relationship between the hazard function, the probability density function and the survival function.

2.3. Gamma Distribution

The Gamma distribution has the pdf of the form:

$$f(t) = \frac{\lambda(\lambda t)^{k-1}e^{-\lambda t}}{\Gamma(k)} \quad t > 0 \tag{10}$$

Where k > 0 and t > 0 are parameters, λ^{-1} is the scale parameter and k is the shape parameter.
The survival function 1 - F(t) is

$$S(\mathrm{t}) = \int_t^\infty \frac{\lambda}{\Gamma(k)}(\lambda t)^{k-1}e^{-\lambda t}dx. \tag{11}$$

The hazard function,

$$h(\mathrm{t}) \text{ is } f(\mathrm{t})/S(\mathrm{t}) = \frac{\lambda(\lambda)^{n-1}}{(n-1)!\sum_{k=0}^{n-1}(1/k!)(\lambda t)^k} \tag{12}$$

It can be shown to be monotone increasing for k > 1.

2.4. Log-Logistic Distribution

The log-logistic distribution is related to the logistic distribution in an identical fashion to how the log-normal and normal distributions are related with each other.

A logarithmic transformation on the logistic distribution generates the log-logistic distribution. Because of its flexible shapes, the log-logistic distribution has been illustrated to provide useful fits to data from many different fields, including engineering, economics, hydrology, and medical sciences.

The log logistic distribution is characterized by two parameters, γ (positive shape parameter) and η (positive scale parameter). A log-logistic random variable X with parameters γ and η has probability density function and survival function as follows;

$$f(x) = \frac{\Upsilon\eta x^{\eta-1}}{(1+\Upsilon x^{\eta})^2}$$

$$S(x) = \frac{1}{(1+\Upsilon x^{\eta})} \tag{13}$$

For ϒ > 0, η > 0. The log logistic distribution can be used to model the lifetime of an object, the lifetime of an organism, or a service time [1].

The cumulative distribution function,

$$\mathrm{F}\ (x) = \mathrm{P}(\mathrm{X} \le \mathrm{x}) = \frac{(\Upsilon x)^{\eta}}{1+(\Upsilon x)^{\eta}} \qquad \mathrm{x} > 0 \tag{14}$$

The hazard function is

$$h(x) = \frac{f(x)}{S(x)} = \frac{\Upsilon\eta(\Upsilon x)^{\eta-1}}{\left(1+(\Upsilon x)\right)^{\eta}} \qquad \mathrm{x} > 0 \tag{15}$$

The cumulative hazard function,

$$H(x) = -\ln\left(S(x)\right) = \ln[1 + (\Upsilon x)^{\eta}] \qquad \mathrm{x} > 0. \tag{16}$$

The log-logistic distribution was used in survival analysis [7] Similarly, it was used to model economic data [8].

3. METHODOLOGY

The Expectation maximization algorithm (EM) was used to achieve the maximum likelihood estimation of the parameters of the model, it is a way to find maximum-likelihood estimates for model parameters when your data is incomplete, has missing data points, or has unobserved variables. It is an iterative way to approximate the maximum likelihood function. A model selection criterion based on the Akaike Information (AIC) was employed to find the mixture model that gives the best fit.

Data is simulated from a two-component parametric mixture of Gamma-Gamma, and Loglogistic-Gamma distributions. The model is evaluated by simulated data set, before it's applied to real dataset.

3.1. Parametric Mixture Model

Let $T_1, \dots, T_n$ be n independent random variables, where T_j is the survival time of the jth subject. We assume that the probability density function $f(t)$ of T_j is a mixture.

$$f(t) = \sum_{i=1}^{a} \pi_i \, f_i(t;\theta) \tag{17}$$

Where $f_i(t;\boldsymbol{\theta}_i)$ are the component densities of the mixture, $\boldsymbol{\theta}_{\mathrm{i}}$ are the corresponding parameters of the ith density and the $\boldsymbol{\pi}_{\mathrm{i}}$ are nonnegative probabilities that sum to one. That is,

$$\sum \pi_i = 1 \quad \text{and} \quad 0 \leq \boldsymbol{\pi}_i \leq 1 \qquad (i = 1,\dots\dots,a). \tag{18}$$

The quantities $\boldsymbol{\pi}_1, \dots {}_{\mathrm{a}}$ are called mixing proportions or weights. Since the components f_1 (t;$_1$),…,f_{a} (t;$\boldsymbol{\theta}_{\mathrm{a}}$) are densities, the mixture (3.1) is a density.

It follows the survival function of failure-time T_j under mixture model is also a mixture,

$$S(t) = \sum \pi_i S_i(t;\theta_i) \tag{19}$$

where $S_i(t;\theta)$ denotes the i^{th} component survival function.

3.2. Gamma-Gamma Mixture Model

The mixture model of distributions assumes the population consist of two distinct sub groups or classes, therefore, the gamma mixture can be written as:

$$f_{\text{gm-gm}}(\mathrm{t}) = \pi f_{\text{gm}}(\mathrm{t};\lambda_1, \mathrm{k}_1) + (1\text{-}\pi) f_{\text{gm}}(\mathrm{t};\lambda_2, \mathrm{k}_2) \tag{20}$$

where π is the mixture weight of the distributions and λ_1, λ_2, k_1, and k_1 are shapes and scales of the two components respectively.

3.3. Log-logistic Gamma Mixture Model

The mixture of the densities of Log-logistic and Gamma can be represented as:

$$f_{\text{ll-gm}}(\mathrm{t}) = \pi f_{\text{ll}}(\mathrm{t};\gamma_1, \eta_1) + (1\text{-}\pi) f_{\text{gm}}(\mathrm{t};\lambda_2, \mathrm{k}_2) \tag{21}$$

where is the weights of the mixture and γ_1, η_1 are the shape and scale of the loglogistic component and λ_2, k_2 are the shape and scale of the gamma component of the distribution. The Expectation Maximization (EM) algorithm [9] is a general approach to maximum likelihood estimation for problems of finite mixture models [10]. Starting value initialisation is very important in the EM algorithm because the likelihood surface of mixture models tend to have multiple modes [11]. The EM algorithm typically produces improved result when started from reasonable initial values [12]. The EM algorithm is a broadly applicable algorithm that provides an iterative procedure for computing Maximum Likelihood Estimates (MLE), [10].

Suppose the density of a random variable $\boldsymbol{Y}$ has an a component mixture form.

$$f(y_i;\Psi) = \sum \pi_i f_i(y;\theta_i) \tag{22}$$

where $\Psi = (\boldsymbol{\pi}_1, \dots, \boldsymbol{\pi}_{a-1}, \boldsymbol{\theta}'_1, \dots, \boldsymbol{\theta}'_a)'$ is the vector containing all the unknown parameters in the mixture model. Let $\boldsymbol{\pi} = (\boldsymbol{\pi}_1, \boldsymbol{\pi}_2, \dots, {}_{a})'$ be the vector of mixing proportions. Suppose $\boldsymbol{y_1}, \boldsymbol{y_2}, \dots, \boldsymbol{y_n}$ is an observed sample of size n, the likelihood for Ψ can be written as:

$$L(\Psi) = \prod_{j=1}^{n} f(y_i;\Psi)$$

$$= \prod_{j=1}^{n}[\sum \pi_i\, f_i(y_i;\theta_i)_i\,] \tag{23}$$

Maximum likelihood estimation of mixture model is cumbersome to solve using the traditional method of taking derivative with respect to each parameter. These, and some other difficulties made modelling heterogeneous data unattractive for a long time [13].

3.4. Mixture Model Selection

Regarding the mixture density estimation problem, the problem of determining the proper number of components and proper mathematical form of each component is faced. In other words, one need to determine which mixture model fits the data better. The question we try to answer here is a model selection problem.

The statistic Akaike Information Criterion (AIC) appears to be adequate for model selection in the mixture density estimation [13]. Here, we follow the model selection approach using AIC proposed by Akaike [14].

$$\text{AIC} = -2logL(\Psi) + 2d \tag{24}$$

where d is the total number of independent parameters, n is the number of observation and $\boldsymbol{\Psi}$ is the estimate of the vector containing all the parameters.

4. RESULTS

The simulated data contains $n = 100$ observations. The Maximum Likelihood Estimates (MLE), of the parameters of each component of a mixture, and the Akaike Information Criteria (AIC) are used for model selection in each case, mean square error (MSE) and root mean square error (RMSE) are also employed, the MSE and RMSE are among the many ways to quantify the difference between an estimator and the true value of the quantity being estimated.

4.1. Gamma-Gamma

Dataset is simulated from a two-component parametric mixture of Gamma model. The simulated data contains $n = 100$ observations and the proportion parameters for each component is λ_1=0.5, λ_2 =0.5. Table 1 list the MLEs of the parameters and the mixing proportions of the mixture model, it can be seen that, the parameters of the model were estimated successfully, because the estimated parameters are closed to the postulated values. Table 2 shows that, the EM is also consistent and stable in its estimation because of the relatively small values of MSE and RMSEs.

Table 1. MLEs of gamma-gamma model with no repetitions.

Gamma-Gamma						
Parameter	λ_1	λ_2	α_1	α_2	β_1	β_2
Postulate	0.5	0.5	40	6	20	1
Estimate	0.485	0.515	39.999	6.007	19.752	0.963

Table 2. MLEs of gamma-gamma model with 300 repetitions.

Gamma-Gamma						
Parameter	λ_1	λ_2	α_1	α_2	β_1	β_2
Postulate	0.5	0.5	40	6	20	1
Estimate	0.495	0.515	41.412	5.787	20.661	0.946
MSE	1.46e-06	1.45e-06	1.70e-01	2.78e-03	4.00e-02	1.14e-04
RMSE	0.001	0.001	0.412	0.053	0.200	0.011

To assess the consistency of the model, simulation was repeated 300 times, the Mean Square Error (MSE), and the Root Mean Square Error were obtained. It can be seen that, both MSE and the RMSE values are very small which shows the model is consistent. Figure 1 compares the density function of classical parametric model and the density of the mixture model, it can be seen that, the mixture model fits the data better.

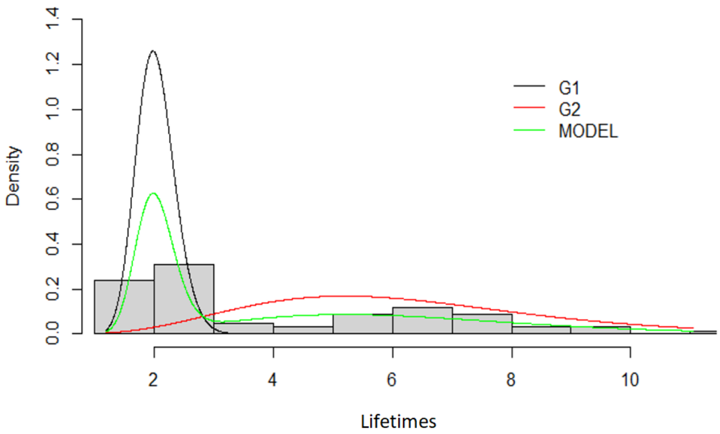

Figure 1. Density of survival mixture model of gamma-gamma.

Table 3 shows that, the model estimates are very close to the postulated values, hence the model can be said to be efficient.

Table 3. MLEs of log-logistic-gamma model with no repetitions.

Log-logistic-Gamma						
Parameter	λ_1	λ_2	α_1	α_2	β_1	β_2
Postulate	0.5	0.5	10	2	4	0.1
Estimate	0.521	0.480	10.523	1.887	3.956	0.111

Simulations were repeated 300 times to test for consistency of the model and it is found to be consistent as can be seen from the MSE and RMSE in Table 4, which are very close to zero.

Table 4. MLEs of log-logistic - gamma model with 300 repetitions.

Log-logistic-Gamma						
Parameter	λ_1	λ_2	α_1	α_2	β_1	β_2
Postulate	0.5	0.5	10	2	4	0.1
Estimate	0.440	0.561	10.692	1.556	4.341	0.111
MSE	4.19e-06	4.19e-06	1.34e-02	2.24e-04	2.26e-03	1.43e-06
RMSE	0.002	0.002	0.116	0.015	0.0478	0.001

The densities of the two classical distributions were also compared to the model as shown in Figure 2 it can be seen that, the model fits the data better.

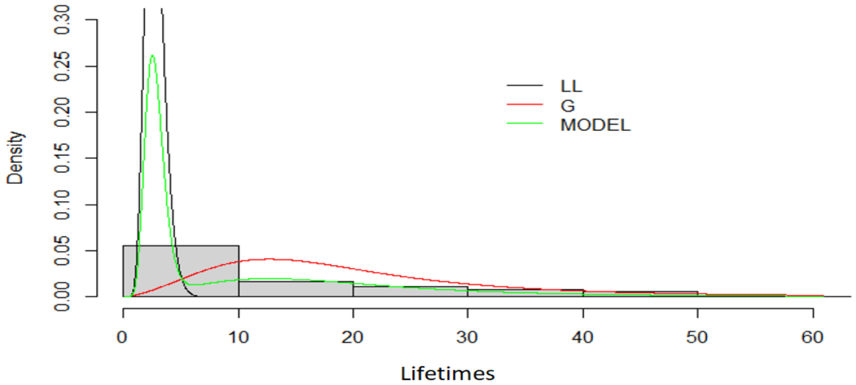

Figure 2. Density of survival mixture model of Gamma – log-logistic model.

4.2. Real Data Application

The data used is the Acute Myelogenous Leukemia (aml) dataset in R statistical software, it was first used by Rupert [15]. The question at the time was whether the standard course of chemotherapy should be extended ("maintained") by additional cycles.

The MLEs of aml dataset using Gamma-Gamma model is in Table 5 and Figure 3, it shows the survival function alongside the K-M of the model.

Table 5. MLEs and AIC of the AML dataset using gamma-gamma model.

Model	Estimate	LL	AIC
Gamma-Gamma	$\hat{\lambda}_1 = 0.791$, $\hat{\lambda}_2 = 0.213$	-98.934	209.866
	$\hat{\alpha}_1 = 1.335$ $\hat{\alpha}_2 = 18.094$		
	$\hat{\beta}_1 = 21.333$ $\hat{\beta}_2 = 1.867$		

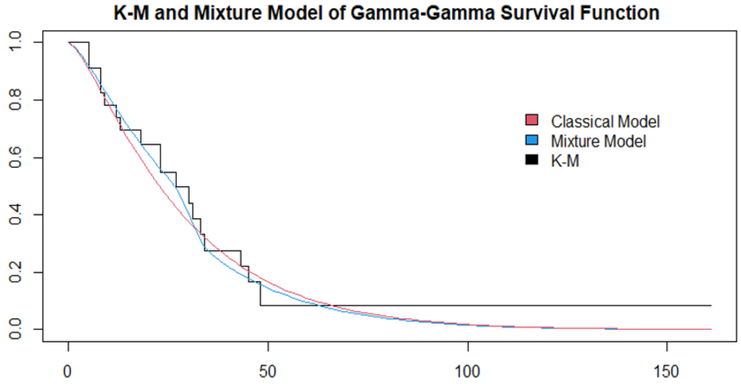

Figure 3. Kaplan meier (K-M) survival curve of gamma-gamma.

Figure 3 shows that, the mixture model fits the K-M curve better than the classical model as it is closer to the K-M curve.

Table 6. MLEs and AIC of loglogistic-gamma model.

Model	Estimate	LL	AIC
Loglogistic-Gamma	$\hat{\lambda}_1 = 0.133$ $\hat{\lambda}_2 = 0.877$	-98.842	209.690
	$\hat{\alpha}_1 = 179.512$ $\hat{\alpha}_2 = 1.442$		
	$\hat{\beta}_1 = 0.173$ $\hat{\beta}_2 = 20.691$		

Table 6 presents the results of the loglogistic-gamma model after it's application on real data, the aml dataset, $\hat{\lambda}_1$ and $\hat{\lambda}_2$ are the mixing proportions, while $\hat{\alpha}_1$, $\hat{\alpha}_2$ and $\hat{\beta}_1$, $\hat{\beta}_2$ are the shapes and scale parameters of the model respectively. The AIC measures the amount of information lost in using the model to fit the data.

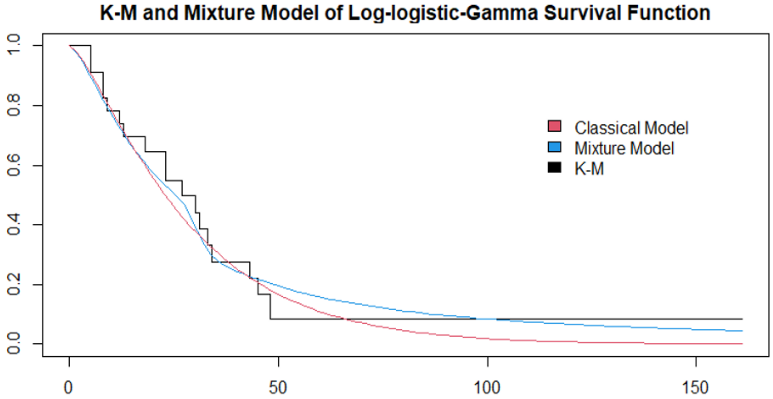

Figure 4. Kaplan meier (K-M) survival curve of loglogistic-gamma model.

Figure 4 shows the Kaplan meier survival curve of the loglogistic-gamma distribution.

It is observed that, the Gamma-Gamma fits both the simulated and real data well, however Loglogistic-Gamma model gives the best fit to both datasets, because it has the lowest Akaike Information Criteria (AIC), and comparing the Kaplan Meier (K-M) survival curves it can also be seen that, the K-M survival curve of the loglogistic-Gamma in gives a better fit than the single classical distributions. The estimation of the parameters of the model were successful in both the simulated and the real data, as estimated values were relatively close to postulated values. As can be seen the mixture model of Loglogistic-Gamma gives the best fit for the real data among the proposed models.

5. CONCLUSION

The paper proposed a two-component mixture model of classical distributions, namely, Gamma-Gamma and Loglogistic-Gamma distributions to analyze heterogenous survival data, simulated and real data were employed to assess the performance of the models, the models were found to estimate the parameters successfully as the estimates were close to the postulated values. It is found that, the mixture of two different distributions i.e. Loglogistic-Gamma gives a best fit to the real data applied.

Funding: The study did not receive any financial support.
Competing Interests: The authors declare no competing interest.
Authors' Contributions: All authors contributed to the conception and design of the study.

REFERENCES

[1] E. T. Lee and J. W. Wang, *Statistical methods for survival data analysis*, 3rd ed.: John Wiley & Son, 2003.

[2] D. A. Kouassi and J. Singh, "A semiparametric approach to hazard estimation with randomly censored observations," *Journal of the American Statistical Association*, vol. 92, pp. 1351-1355, 1997.Available at: https://doi.org/10.1080/01621459.1997.10473656.

[3] Y. A. Mohammed, B. Yatim, and S. Ismail, "A parametric mixture model of three different distributions: An approach to analyse heterogeneous survival data," in *Proceedings of the 21st National Symposium on Mathematical Sciences (SKSM21) AIP Conf. Proc. 1605*, 2014, pp. 1040-1045.

[4] Y. A. Mohammed, B. Yatim, and S. Ismail, "A simulation study of parametric mixture model of three different distributions to analyse heterogeneous survival data," *Modern Applied Science*, vol. 7, pp. 1-9, 2013.Available at: http://dx.doi.org/10.5539/mas.v7n7p1.

[5] E. H. Blackstone, D. C. Naftel, and M. E. Turner Jr, "The decomposition of time-varying hazard into phases, each incorporating a separate stream of concomitant information," *Journal of the American Statistical Association*, vol. 81, pp. 615-624, 1986.Available at: https://doi.org/10.1080/01621459.1986.10478314.

[6] U. Erisoglu, M. Erisoglu, and H. Erol, "Mixture model approach to the analysis of heterogeneous survival time data," *Pakistan Journal of Statistics*, vol. 28, pp. 115-30, 2011.

[7] R. C. Gupta, O. Akman, and S. Lvin, "A study of log-logistic model in survival analysis," *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, vol. 41, pp. 431-443, 1999.Available at: https://doi.org/10.1002/(sici)1521-4036(199907)41:4%3C431::aid-bimj431%3E3.0.co;2-u.

[8] P. R. Fisk, "The graduation of income distributions," *Econometrica: Journal of the Econometric Society*, vol. 29, pp. 171-185, 1961.Available at: https://doi.org/10.2307/1909287.

[9] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 39, pp. 1-22, 1977.

[10] G. J. McLachlan and T. Krishnan, *EM algorithm and extensions*, 2nd ed.: A John Wiley & Sons Publication, 2007.

[11] Y. Zhang, "Parametric mixture models in survival analysis with application," Doctoral Dissertation, UMI Number: 3300387, Graduate School, Temple University, 2008.

[12] C. Fraley and A. E. Raftery, "Model-based clustering, discriminant analysis, and density estimation," *Journal of the American Statistical Association*, vol. 97, pp. 611-631, 2002.Available at: https://doi.org/10.1198/016214502760047131.

[13] G. McLachlan and D. Peel, *Finite mixture models*: John Wiley & Son, 2000.

[14] H. Akaike, "Information theory and extension of the maximum likelihood principle," presented at the 2nd International Symposium on Information Theory, B.N. Petov and F. Csaki (Eds.), Akademiai Kiado, Budapest, 1973.

[15] G. M. Rupert, *Survival analysis*, 2nd ed.: Wiley Inter Science, 1998.

Views and opinions expressed in this article are the views and opinions of the author(s), International Journal of Mathematical Research shall not be responsible or answerable for any loss, damage or liability etc. caused in relation to/arising out of the use of the content.