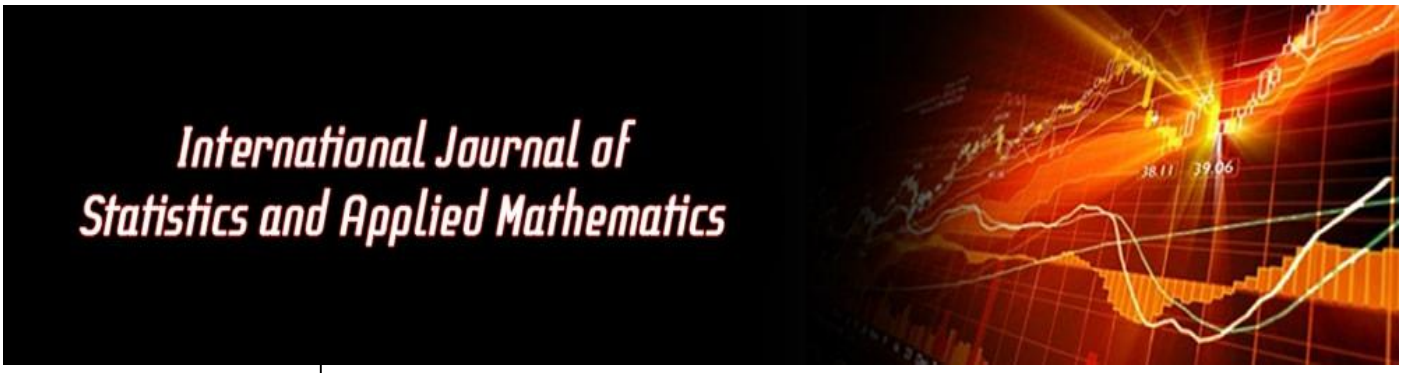
International Journal of Statistics and Applied Mathematics 2018; 3(1): 396-407



ISSN: 2456-1452
March 2018; 3(1): 396-407

www.mathsjournal.com
Received: 20-11-2017
Accepted: 21-12-2017

R Uma Maheswari
Research Scholar, Department of Statistics, Loyola College, Chennai, India

T Leo Alexander
Associate Professor, Department of Statistics, Loyola College, Chennai, India

Correspondence
T Leo Alexander
Associate Professor, Department of Statistics, Loyola College, Chennai, India

Three identical mixture distributions approach to analyze composite survival data

R Uma Maheswari and T Leo Alexander

Abstract
In this paper, a parametric mixture model of three identical (same) distributions of Exponential, Gamma, Log-normal, Weibull and Gompertz is considered to model composite or heterogeneous survival data. Mixtures of these three identical distributions were tested for the best fit by the simulated datasets as well as real time survival dataset. Some properties of the proposed parametric mixture of Exponential, Gamma, Weibull, Lognormal and Gompertz are investigated. The Expectation Maximization Algorithm (EM) is employed to estimate parameters of mixture models based on Maximum Likelihood method. Simulations are performed by generating data, sampled from a population of three component parametric mixtures of three identical distributions and the simulations have been repeated 500, 1000, 5000 times with samples of size 100 observations for each mixture model to investigate the consistency and stability of the EM algorithm. The repetitions of the simulation give estimators closer to the postulated models, as the number of repetitions increases with relatively small standard errors. Akaike's information criterion (AIC) and goodness of fit tests are used for the comparison of model performances. Results revealed that the proposed model fits the real data better than the pure classical survival models corresponding to each component.

Keywords: EM-algorithm, maximum likelihood estimation, mixture distribution, simulation survival analysis, three identical mixture

Introduction
Survival data analysis is a collection of Statistical procedures to analyze the occurrence of a particular event of interest over a given time T. The outcome variable of interest is time until an event occurs. The event can be the development of a disease, treatment outcome, relapse or death. Survival analysis uses the data which are related to Clinical research, laboratory tests such as testing life time of some electronic devices. Historically, Nonparametric and Classical Parametric Survival models are commonly employed in analyzing lifetime data. The most commonly used parametric distributions include the Exponential, Gamma and Weibull among others. [1-3].

Situations where the data are to be believed heterogeneous in nature, Survival mixture models are more convenient for modeling such type of data. Recently, many researchers employing survival mixture models for analyzing survival data.

Mixture models can be used to analyze failure – time data in a variety of situations. Particularly, Mixture distributions provide a way of modeling time to failure in the case of competing risks or failures. Mixture of two different distributions such as Exponential-Gamma, Exponential-Weibull and Gamma-Weibull was proposed by Erisogku Ulku to model heterogeneous survival data [4]. Mixture models of Gamma, Lognormal and Weibull distributions were examined by Erisogku Ulku to model heterogeneous Survival data [5]. Ayça Hatice Tùrkan.showed a comparison study of two-component Mixture model distribution for heterogeneous survival time dataset by taking a mixture of two identical (same kind of) distributions of Exponential, Gamma, Lognormal and Weibull and also all pairwise combinations of these distribution and analyzed which kind of mixture model distributions is more appropriate for the heterogeneous survival times [6]. Two component of Identical and Non-Identical mixtures of Exponential, Gamma, Lognormal, Weibull and Gompertz model was developed to analyze heterogeneous Survival data [7, 8].

Three components parametric survival mixture models did not receive much attention. A study was conducted to observe the risk of death after open-heart surgery. The lifetime of a patient after surgery of this nature can be decomposed into three overlapping phases in time. The first phase (early phase) is the period immediately following surgery in which the risk of dying is relatively very high. The second phase (constant phase) refers to the subsequent period in which the hazard function for death is essentially constant. This second phase then merges with a third and final phase (late phase) in which the risk of death starts to increase. Given these three phases, a convenient way of postulating a parametric model for the distribution of the failure time (death) is to adopt a three component mixture model, where the three component correspond to the early, constant and late phases [9, 10]. and [11]. A simulation study on parametric mixture model of three different Distributions of Exponential, Gamma and Weibull was developed by Yusuf A. Mohammed and applied these mixtures models for Kidney Cather data to analyze heterogeneous Survival time [12, 13]. Yusuf A. Mohammed proposed a three components survival mixture model of the Gamma distribution for the analysis of heterogeneous survival data [14].

In this paper, simulated and real data were used to investigate the consistency and stability of EM in estimating the parameters and appropriateness of a three (identical) components survival mixture model of the Exponential, Gamma, Weibull, Lognormal and Gompertz distribution in modelling heterogeneous survival data. The arrangement of the paper is as follows. In Section 2, we define the functions of Survival time and some properties of the Exponential, Gamma, Weibull, Lognormal and Gompertz distributions were highlighted. In Section 3, we devote in discussing mixture model of three (identical) components in the survival analysis and the maximum likelihood estimators of the parameters are obtained by EM algorithm. In Section 4, we simulate data, each randomly sampled from a population of three component parametric mixture model of identical distributions and the simulations has been repeated 500, 1000 and 5000 times with sample size of 100 observations for each mixture model to investigate the convergence of the EM, consistency, stability of EM algorithm and also examined the appropriateness of these mixture model in analyzing the heterogeneous survival time data. Section 5, the summary and conclusion were presented. All computations are performed using R language.

2. Functions of Survival time and Parametric distributions

Lifetime is the duration of life measured from some particular starting point. In applications, other terms such as "failure time" and "survival time" are also often used. Survival time data measure the time taken for a certain event to occur such as failure, death, response, relapse, the development of a given disease, parole, or divorce. These times are subject to random variations, and like any random variables, form a distribution. Let T denote the survival time. The distribution of T can be characterized by three equivalent functions. Survival function, denoted by $S(t)$, is defined as the probability that an individual survives longer than t : $S(t) = P(T > t), 0 < t < \infty.$

Here $S(t)$ is a non-increasing function of time t with the probability of surviving at least at the time zero is 1 and that of surviving an infinite time is zero. Cumulative distribution function $F(t)$, is defined as the probability that an individual fails before t is $F(t) = P(T \le t), 0 < t < \infty.$. The hazard function $h(t)$ of survival time T gives the conditional failure rate. This is defined as the probability of failure during a very small time interval, assuming that the individual has survived to the beginning of the interval, or as the limit of the probability that an individual fails in a very short interval, $t + \Delta t$, given that the individual has survived to time t :

$$h(t) = \lim_{\Delta t \to 0} \left[\frac{P(t \le T < (t + \Delta t)/T \ge t)}{\Delta t} \right] = \frac{f(t)}{S(t)}.$$

The cumulative hazard function is defined as $H(t) = -\log(S(t)) = \int_0^1 h(u)du.$ Given any one of them, the other two can be derived [2] $S(t) = 1 - F(t) = \exp(-H(t)).$ Probability density functions survival functions distribution functions of the parametric distributions used in this study are briefly summarized below:

Distribution	Probability Density Function	Survival function
Exponential	$f_{\exp}(t) = \frac{1}{\lambda} e^{-\frac{t}{\lambda}},\ t > 0, \lambda > 0$	$s_{\exp}(t) = 1 - e^{-\frac{t}{\lambda}}$
Weibull distribution	$f_{wbl}(t) = \frac{\gamma}{\eta}\left(\frac{t}{\eta}\right)^{\gamma-1} e^{-\left(\frac{t}{\eta}\right)^{\gamma}}, t, \eta, \gamma > 0$	$s_{wbl}(t) = e^{-\left(\frac{t}{\gamma}\right)^{\eta}}.$
Gamma distribution	$f_{gam}(t) = \frac{t^{\alpha-1} e^{-\frac{t}{\beta}}}{\beta^{\alpha}\Gamma(\alpha)}, t, \alpha, \beta > 0$	$S_{gam}(t) = 1 - \frac{\Gamma_x(\alpha)}{\Gamma(\alpha)}$
Lognormal distribution	$f_{\log n}(t) = \frac{1}{\sigma t\sqrt{2\pi}} e^{-\frac{(\log t - \mu)^2}{2\sigma^2}},\ t > 0, \mu, \sigma > 0$	$S_{\log n}(t) = 1 - \phi\left(\frac{\log t - \mu}{\sigma}\right)$
Gompertz distribution	$f_{gomp}(t) = be^{at} e^{-\frac{b}{a}(e^{at}-1)}, t > 0, a, b > 0$	$s_{gomp}(t) = e^{-\frac{b}{a}(e^{at}-1)}$

It may be noted that $\Gamma_x(\alpha)$ is called an incomplete Gamma function and ϕ is cumulative distribution function of normal probability distribution function

3. Mixture model of three components in Survival Analysis

In survival analysis, mixture models are often used because they are high flexible. Mixture models have been widely used to model failure – time data in a variety of situations and they are applicable in situations where the adoption of a single parametric family for the distribution of failure time is inadequate. Mixture of three components is used when the data consists of three subpopulation or subgroups. Equation (1) represents mixture model of three components and written as

$$f_{1,2,3}(t;\psi) = \pi_1 f_1(t;\theta_1) + \pi_2 f_2(t;\theta_2) + \pi_3 f_3(t;\theta_3), \qquad (1)$$

where the vector $\boldsymbol{\psi}' = (\pi_1, \pi_2, \pi_3, \theta_1, \theta_2, \theta_3)$ contains all the unknown parameters in the mixture model [15, 16]. The functions $f_1(t;\theta_1)$, $f_2(t;\theta_2)$ and $f_3(t;\theta_3)$ are the probability density function corresponding to each component with some parameters θ_1, θ_2 and θ_3.

In this paper, to model the heterogeneous survival time data, we consider mixture of three identical distributions. The model includes Exponential, Gamma, Weibull, Lognormal and Gompertz mixture and are defined as follows

$$\begin{aligned}
f_{E_E_E}(t) &= \pi_1 f_E(t;\lambda_1) + \pi_2 f_E(t;\lambda_2) + \pi_3 f_E(t;\lambda_3) \\
f_{G_G_G}(t) &= \pi_1 f_G(t;\alpha_1,\beta_1) + \pi_2 f_G(t;\alpha_2,\beta_2) + \pi_3 f_G(t;\alpha_3,\beta_3) \\
f_{W_W_W}(t) &= \pi_1 f_W(t;\eta_1,\gamma_1) + \pi_2 f_W(t;\eta_2,\gamma_2) + \pi_3 f_W(t;\eta_3,\gamma_3) \\
f_{L_L_L}(t) &= \pi_1 f_L(t;\mu_1,\sigma_1) + \pi_2 f_L(t;\mu_2,\sigma_2) + \pi_3 f_L(t;\mu_3,\sigma_3) \\
\text{and} \quad f_{GO_GO_GO}(t) &= \pi_1 f_{GO}(t;a_1,b_1) + \pi_2 f_{GO}(t;a_2,b_2) + \pi_1 f_{GO}(t;a_3,b_3),
\end{aligned}$$

where $\pi_i' s$ represents the mixture weight of the three subpopulation with $\sum_{i=1}^{3} \pi_i = 1$. The maximum likelihood estimators of parameters of these mixture distributions are estimated using Expectation-Maximization (EM) algorithm

3.1 Expectation-Maximization (EM) algorithm

The most effective method employed to estimate the Maximum likelihood estimators in finite mixture models is the Expectation-Maximization algorithm. [15-17]. 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 t_i belongs to k^{th} component and 0 otherwise for $k = 1,2,3$ and $i = 1, \ldots, n$. The EM algorithm is applied to the mixture distributions by treating z_i as unobserved or missing data. It consists of two steps, E (for Expectation) and M (for Maximization).

In E- step, to estimate the hidden variable vector $z_i = (z_{1i}, z_{2i}, z_{3i})$, conditional expectation function $E(Z_{ki} | t_i)$ are used so that

$$\begin{aligned}
\hat{z}_{1i} = E_{\psi_0}(z_{1i}|t_i) &= \frac{\pi_1 f_{1,0}(t_i;\theta_1)}{\pi_1 f_{1,0}(t_i;\theta_1) + \pi_2 f_{2,0}(t_i;\theta_2) + \pi_3 f_{3,0}(t_i;\theta_3)} \\
\hat{z}_{2i} = E_{\psi 0}(z_{2i}|t_i) &= \frac{\pi_2 f_{2,0}(t_i;\theta_2)}{\pi_1 f_{1,0}(t_i;\theta_1) + \pi_2 f_{2,0}(t_i;\theta_2) + \pi_3 f_{3,0}(t_i;\theta_3)} \\
\hat{z}_{3i} = E_{\psi 0}(z_{3i}|t_i) &= \frac{\pi_3 f_{3,0}(t_i;\theta_3)}{\pi_1 f_{1,0}(t_i;\theta_1) + \pi_2 f_{2,0}(t_i;\theta_2) + \pi_3 f_{3,0}(t_i;\theta_3)} \quad .
\end{aligned}$$

In M-step, $E(Z_{1i}|t_i)$, $E(Z_{2i}|t_i)$ and $E(Z_{3i}|t_i)$ functions calculated in E-step will be maximized. The M-step and E- step should be iterated alternatively till the convergence criteria are met. The estimator of π_k $(k = 1,2,3)$ are obtained as

$$\hat{\pi}_k = \frac{\sum_{i=1}^{n} \hat{z}_{ki}}{n} .$$

By choosing $f_{E\_E\_E}(t)$ as the **Exponential** pdf of the k^{th} group, the maximum likelihood estimator of parameter λ_k can be obtained as

$$\hat{\lambda}_k = \frac{\sum_{i=1}^{n} \hat{z}_{ki} t_i}{\sum_{i=1}^{n} \hat{z}_{ki}} \ , k = 1,2,3.$$

Also, by choosing $f_{G\_G\_G}(t)$ as the **Gamma** pdf of the k^{th} group, the maximum likelihood estimator of parameter α_k and β_k can be obtained as

$$\hat{\beta}_k = \frac{\sum_{i=1}^{n} \hat{z}_{ki} t_i}{\hat{\alpha}_k \sum_{i=1}^{n} \hat{z}_{ki}} \quad \text{and} \quad \hat{\alpha}_{k,(r+1)} = \hat{\alpha}_{k,r} - \frac{\log(\hat{\alpha}_{k,r}) - \psi'(\hat{\alpha}_{k,r}) - \log\left(\frac{\sum_{i=1}^{n} \hat{z}_{ki} t_i}{\sum_{i=1}^{n} \hat{z}_{ki}}\right) + \frac{\sum_{i=1}^{n} \hat{z}_{ki} \log t_i}{\sum_{i=1}^{n} \hat{z}_{ki}}}{\frac{1}{\hat{\alpha}_{k,r}} - \psi'(\hat{\alpha}_{k,r})}, \ k = 1,2,3.$$

Here r is the number of Newton-Raphson iterations within EM algorithm. Also $\psi(.)$ and $\psi'(.)$ are a digamma and trigamma functions respectively. Also considering $f_{L\_L\_L}(t)$ as the pdf of Lognormal mixture, the Maximum likelihood estimators of the parameters μ_k and σ_k of the k^{th} group was obtained as

$$\hat{\mu}_k = \frac{\sum_{i=1}^{n} \hat{z}_{ki} \ln t_i}{\sum_{i=1}^{n} \hat{z}_{ki}} \quad and \quad \hat{\sigma}_k^2 = \frac{\sum_{i=1}^{n} \hat{z}_{ki} (\ln t_i - \hat{\mu}_k)^2}{\sum_{i=1}^{n} \hat{z}_{ki}}, k = 1,2,3.$$

Likewise, by choosing $f_{W\_W\_W}(t)$ as the Weibull pdf of the k^{th} group, the maximum likelihood estimator of parameter η_k and γ_k can be obtained as

$$\hat{\eta}_k = \left(\left(\sum_{i=1}^{n} \hat{z}_{ki} \right)^{-1} \sum_{i=1}^{n} \hat{z}_{ki} t_i^{\hat{\gamma}_k} \right)^{1/\hat{\gamma}_k}, \qquad \hat{\gamma}_{k,(r+1)} = \hat{\gamma}_{k,r} + \frac{A_{k,r}^* + (1/\hat{\gamma}_{k,r}) - (C_{k,r}^* / B_{k,r}^*)}{(1/(\hat{\gamma}_{k,r})^2) + (B_{k,r}^* D_{k,r}^* - C_{k,r}^{*2}) / B_{k,r}^{*2}}, k = 1,2,3$$

$$where \ A_{k,r}^* = \left(\sum_{i=1}^{n} \hat{z}_{ki} \right)^{-1} \sum_{i=1}^{n} \hat{z}_{ki} logt_i \ , B_{k,r}^* = \sum_{i=1}^{n} \hat{z}_{ki} t_i^{\hat{\gamma}_{k,r}} \ , C_{k,r}^* = \sum_{i=1}^{n} \hat{z}_{ki} t_i^{\hat{\gamma}_{k,r}} logt_i \ and \ D_{k,r}^* = \sum_{i=1}^{n} \hat{z}_{ki} t_i^{\hat{\gamma}_{k,r}} (logt_i)^2 .$$

Finally, considering $f_{GO\_GO\_GO}(t)$ as the **Gompertz** pdf of the k^{th} group, the maximum likelihood estimator of parameter a_k and b_k can be obtained as,

$$\hat{b}_k = \frac{\hat{a}_k \sum_{i=1}^{n} \hat{z}_{ki}}{\sum_{i=1}^{n} \hat{z}_{ki} e^{\hat{a}_k t_i} - \sum_{i=1}^{n} \hat{z}_{ki}} \quad \text{and}$$

$$\hat{a}_{k,(r+1)} = \hat{a}_{k,r} + \left\{ \frac{E_{k,r}^* + \left\{ (F_{k,r}^* G_{k,r}^* - \hat{a}_{k,r} F_{k,r}^* H_{k,r}^* - (F_{k,r}^*)^2) / \hat{a}_{k,r} (G_{k,r}^* - F_{k,r}^*) \right\}}{\left(F_{k,r}^* (G_{k,r}^*)^2 - 2(F_{k,r}^*)^2 G_{k,r}^* + (\hat{a}_{k,r})^2 G_{k,r}^* F_{k,r}^* I_{k,r}^* - (\hat{a}_{k,r})^2 (F_{k,r}^*)^2 I_{k,r}^* - (\hat{a}_{k,r})^2 F_{k,r}^* (H_{k,r}^*)^2 + (F_{k,r}^*)^3 \right)} \right\},$$

$$\text{where } E_{k,r}^* = \sum_{i=1}^{n} \hat{z}_{ki} t_i \ , F_{k,r}^* = \sum_{i=1}^{n} \hat{z}_{ki} \ , G_{k,r}^* = \sum_{i=1}^{n} \hat{z}_{ki} e^{\hat{a}_{k,r} t_i} \ , H_{k,r}^* = \sum_{i=1}^{n} \hat{z}_{ki} t_i e^{\hat{a}_{k,r} t_i} \text{ and } I_{k,r}^* = \sum_{i=1}^{n} \hat{z}_{ki} t_i^2 e^{\hat{a}_{k,r} t_i} \ , k = 1,2,3.$$

Hence r is the number of Newton-Raphson iterations within EM algorithm. The M-step and E- step should be iterated alternatively till the convergence criterion is met.

4. Simulation

Simulations are performed by simulating data, each randomly sampled from a population of three component parametric mixture model of identical distributions and the simulations has been repeated 500, 1000 and 5000 times with sample size of 100 observations for different postulated model to investigate the convergence of the EM, consistency, stability of EM algorithm. The

three component identical mixture model includes Exponential-Exponential-Exponential, Gamma-Gamma-Gamma, Weibull-Weibull-Weibull, Lognormal-Lognormal-Lognormal and Gompertz-Gompertz-Gompertz distributions. There is no restriction imposed on the maximum number of iterations and convergence was achieved when the differences between successive estimates were less than 10^{-4}.

The results from the simulated data set for the first, second and third postulated models are listed from Tables 1 – 5, which give the averages of the maximum likelihood estimators $av(\hat{\pi},\hat{\theta})$ and standard errors $se(\hat{\pi},\hat{\theta})$. Also, the graphs of mixture of three identical distributions for simulation parameters are shown in the following Figures 1- 5. From Figure 1 - 5, clearly exhibits the comparison between pdf of identical mixture model and pdf of each single distribution..

Table 1: Repeated Simulation for different Postulated Model of Exponential mixture

Parameters	π_1	π_2	π_3	λ_1	λ_2	λ_3
Postulated model 1	0.4	0.5	0.1	0.25	0.45	0.75
500 times $av(\hat{\pi},\hat{\theta})$	0.406	0.497	0.097	0.250	0.448	0.745
$se(\hat{\pi},\hat{\theta})$	0.050	0.052	0.029	0.586	0.297	0.457
1000 time $av(\hat{\pi},\hat{\theta})$	0.403	0.498	0.100	0.250	0.450	0.752
$se(\hat{\pi},\hat{\theta})$	0.050	0.051	0.030	0.600	0.310	0.449
5000 time $av(\hat{\pi},\hat{\theta})$	0.400	0.500	0.100	0.250	0.450	0.750
$se(\hat{\pi},\hat{\theta})$	0.050	0.051	0.031	0.619	0.305	0.751
Postulated model 2	0.3	0.3	0.4	1	1.5	2
500 times $av(\hat{\pi},\hat{\theta})$	0.302	0.302	0.396	0.999	1.506	1.977
$se(\hat{\pi},\hat{\theta})$	0.047	0.048	0.050	0.173	0.121	0.073
1000 time $av(\hat{\pi},\hat{\theta})$	0.302	0.301	0.398	0.999	1.500	2.000
$se(\hat{\pi},\hat{\theta})$	0.047	0.046	0.049	0.176	0.122	0.074
5000 time $av(\hat{\pi},\hat{\theta})$	0.300	0.300	0.400	1.000	1.498	2.000
$se(\hat{\pi},\hat{\theta})$	0.047	0.047	0.050	0.182	0.120	0.076
Postulated model 3	0.65	0.2	0.15	2	3	4
500 times $av(\hat{\pi},\hat{\theta})$	0.653	0.200	0.148	2.001	2.966	3.941
$se(\hat{\pi},\hat{\theta})$	0.047	0.039	0.035	0.057	0.073	0.068
1000 time $av(\hat{\pi},\hat{\theta})$	0.651	0.199	0.151	1.995	3.009	4.014
$se(\hat{\pi},\hat{\theta})$	0.047	0.039	0.036	0.060	0.073	0.067
5000 time $av(\hat{\pi},\hat{\theta})$	0.649	0.200	0.150	1.997	3.009	3.987
$se(\hat{\pi},\hat{\theta})$	0.048	0.040	0.036	0.060	0.074	0.066

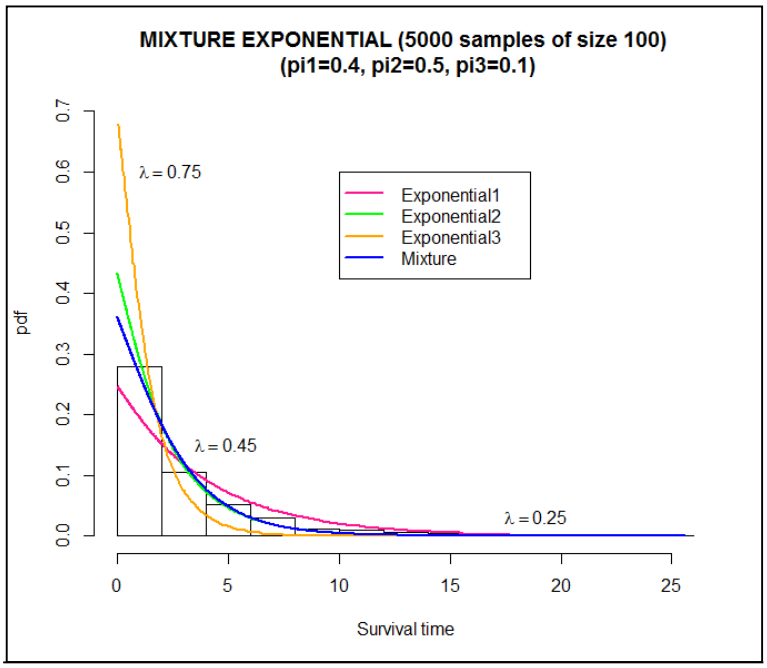


Fig 1: Density function of Exponential mixture versus single distribution

Table 2: Repeated Simulation for different Postulated Model of Lognormal Mixture

Parameters	π_1	π_2	π_3	μ_1	σ_1	μ_2	σ_2	μ_3	σ_3
Postulated model 1	0.38	0.48	0.14	1	0.25	1.5	0.65	2	0.95
500 times $av(\hat{\pi},\hat{\theta})$	0.385	0.477	0.138	1.000	0.249	1.503	0.653	1.987	0.935
$se(\hat{\pi},\hat{\theta})$	0.051	0.051	0.034	0.039	0.028	0.091	0.067	0.263	0.177
1000 times $av(\hat{\pi},\hat{\theta})$	0.383	0.477	0.141	1.000	0.248	1.502	0.651	2.003	0.931
$se(\hat{\pi},\hat{\theta})$	0.050	0.051	0.035	0.039	0.028	0.093	0.067	0.261	0.179
5000 times $av(\hat{\pi},\hat{\theta})$	0.380	0.480	0.141	1.000	0.249	1.500	0.651	1.996	0.945
$se(\hat{\pi},\hat{\theta})$	0.050	0.051	0.035	0.040	0.029	0.091	0.067	0.260	0.176
Postulated model 2	0.10	0.65	0.25	0.22	1.5	0.5	1	1.2	0.5
500 times $av(\hat{\pi},\hat{\theta})$	0.100	0.654	0.247	0.245	1.474	0.502	1.001	1.195	0.497
$se(\hat{\pi},\hat{\theta})$	0.029	0.047	0.043	0.478	0.321	0.115	0.086	0.099	0.069
1000 times $av(\hat{\pi},\hat{\theta})$	0.100	0.651	0.250	0.220	1.461	0.501	1.001	1.202	0.494
$se(\hat{\pi},\hat{\theta})$	0.030	0.047	0.043	0.485	0.331	0.119	0.086	0.097	0.071
5000 times $av(\hat{\pi},\hat{\theta})$	0.100	0.650	0.250	0.221	1.454	0.500	0.999	1.200	0.494
$se(\hat{\pi},\hat{\theta})$	0.030	0.047	0.044	0.500	0.349	0.117	0.087	0.099	0.070
Postulated model 3	0.1	0.1	0.8	0.65	1.4	1	2	2	4
500 times $av(\hat{\pi},\hat{\theta})$	0.100	0.099	0.801	0.673	1.376	0.989	1.944	0.989	4.136
$se(\hat{\pi},\hat{\theta})$	0.029	0.029	0.040	0.446	0.300	0.627	0.461	0.402	0.304
1000 times $av(\hat{\pi},\hat{\theta})$	0.100	0.100	0.801	0.650	1.364	0.988	1.954	2.011	3.999
$se(\hat{\pi},\hat{\theta})$	0.030	0.030	0.040	0.453	0.308	0.662	0.458	0.412	0.313
5000 times $av(\hat{\pi},\hat{\theta})$	0.099	0.100	0.801	0.650	1.357	1.004	1.953	2.001	3.993
$se(\hat{\pi},\hat{\theta})$	0.030	0.030	0.040	0.453	0.326	0.674	0.463	0.408	0.311

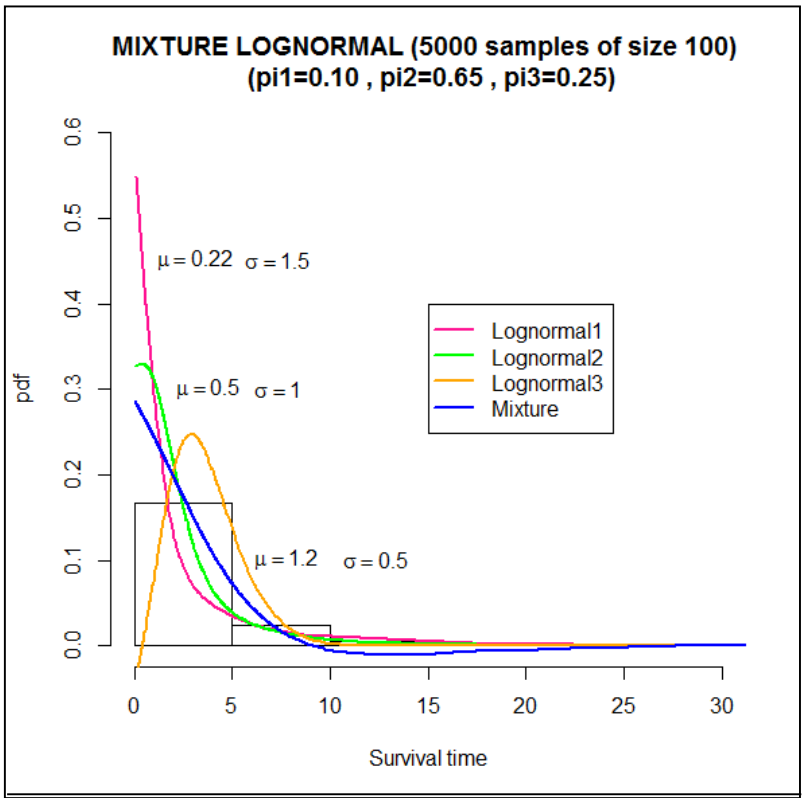


Fig 2: Density function of Lognormal mixture versus single distribution

Table 3: Repeated Simulation for different Postulated Model of Gamma mixture

Parameters	π_1	π_2	π_3	α_1	β_1	α_2	β_2	α_3	β_3
Postulated model 1	0.1	0.7	0.2	3	0.5	6	0.7	8	1.5
500 times $av(\hat{\pi},\hat{\theta})$	0.100	0.703	0.197	3.305	0.457	5.956	0.705	8.079	1.483
$se(\hat{\pi},\hat{\theta})$	0.029	0.045	0.040	0.335	0.087	0.657	0.032	0.926	0.115
1000 times $av(\hat{\pi},\hat{\theta})$	0.100	0.700	0.200	3.336	0.451	5.999	0.700	8.309	1.485
$se(\hat{\pi},\hat{\theta})$	0.030	0.045	0.041	0.239	0.085	0.467	0.032	0.673	0.115
5000 times $av(\hat{\pi},\hat{\theta})$	0.100	0.700	0.200	3.383	0.457	6.000	0.700	8.333	1.487
$se(\hat{\pi},\hat{\theta})$	0.030	0.046	0.041	0.097	0.086	0.467	0.032	0.301	0.116
Postulated model 2	0.15	0.35	0.50	12.00	4.0	8.0	4	4.0	2.0
500 times $av(\hat{\pi},\hat{\theta})$	0.149	0.355	0.496	12.467	3.866	8.098	3.943	3.933	2.036
$se(\hat{\pi},\hat{\theta})$	0.035	0.049	0.052	1.740	0.300	0.842	0.230	0.494	0.138
1000 times $av(\hat{\pi},\hat{\theta})$	0.150	0.353	0.497	12.161	3.793	8.125	3.934	4.025	2.000
$se(\hat{\pi},\hat{\theta})$	0.036	0.048	0.051	1.244	0.287	0.769	0.229	0.356	0.136
5000 times $av(\hat{\pi},\hat{\theta})$	0.150	0.351	0.500	12.109	3.999	8.164	3.922	4.054	2.000
$se(\hat{\pi},\hat{\theta})$	0.035	0.049	0.051	0.566	0.290	0.344	0.231	0.160	0.135
Postulated model 3	0.23	0.32	0.45	15	5.5	10	2.3	5	0.8
500 times $av(\hat{\pi},\hat{\theta})$	0.230	0.324	0.447	15.213	5.422	10.091	2.277	4.901	0.816
$se(\hat{\pi},\hat{\theta})$	0.042	0.048	0.051	2.157	0.288	1.384	0.123	0.629	0.053
1000 times $av(\hat{\pi},\hat{\theta})$	0.230	0.322	0.448	15.459	5.328	10.174	2.260	5.027	0.797
$se(\hat{\pi},\hat{\theta})$	0.043	0.047	0.050	1.543	0.280	0.982	0.125	0.455	0.052
5000 times $av(\hat{\pi},\hat{\theta})$	0.229	0.320	0.450	15.639	5.452	10.243	2.256	5.073	0.788
$se(\hat{\pi},\hat{\theta})$	0.042	0.047	0.051	0.695	0.290	0.440	0.126	0.205	0.051

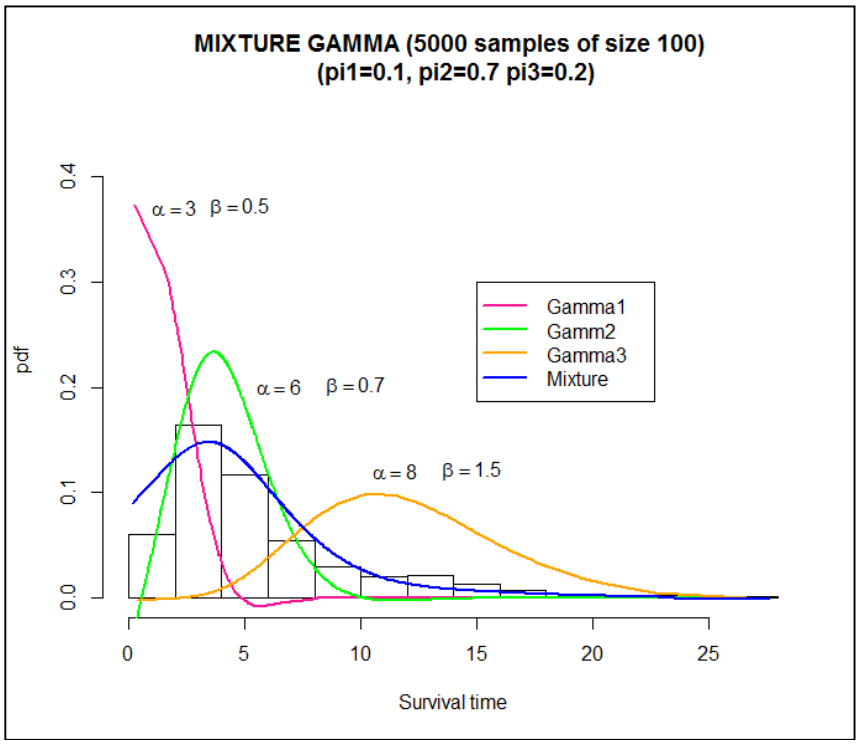


Fig 3: Density function of Gamma mixture versus single distribution

Table 4: Repeated Simulation for different Postulated Model of Weibull Mixture

Parameters	π_1	π_2	π_3	γ_1	η_1	γ_2	η_2	γ_3	η_3
Postulated model 1	0.65	0.12	0.23	15	8	12	6	10	5
500 times $av(\hat{\pi},\hat{\theta})$	0.653	0.120	0.227	15.002	7.994	12.169	5.836	10.053	4.956
$se(\hat{\pi},\hat{\theta})$	0.047	0.031	0.042	0.001	0.064	0.092	0.142	0.029	0.100
1000 time $av(\hat{\pi},\hat{\theta})$	0.651	0.119	0.230	15.003	7.994	12.185	5.816	10.059	4.942
$se(\hat{\pi},\hat{\theta})$	0.047	0.031	0.042	0.001	0.064	0.101	0.148	0.032	0.101
5000 times $av(\hat{\pi},\hat{\theta})$	0.649	0.121	0.230	15.000	7.994	12.102	5.889	10.053	4.920
$se(\hat{\pi},\hat{\theta})$	0.048	0.033	0.042	0.001	0.063	0.085	0.133	0.030	0.098
Postulated model 2	0.75	0.10	0.15	5	2	7	3	9	6
500 times $av(\hat{\pi},\hat{\theta})$	0.753	0.099	0.148	5.007	1.996	7.447	2.796	9.148	5.821
$se(\hat{\pi},\hat{\theta})$	0.043	0.030	0.035	0.003	0.043	0.166	0.121	0.079	0.167
1000 time $av(\hat{\pi},\hat{\theta})$	0.750	0.099	0.151	5.017	1.993	7.531	2.759	9.138	5.817
$se(\hat{\pi},\hat{\theta})$	0.043	0.030	0.036	0.008	0.044	0.199	0.121	0.074	0.171
5000 times $av(\hat{\pi},\hat{\theta})$	0.750	0.099	0.150	5.034	1.989	7.417	2.779	9.138	5.526
$se(\hat{\pi},\hat{\theta})$	0.044	0.030	0.036	0.008	0.043	0.156	0.113	0.074	0.157
Postulated model 3	0.2	0.3	0.5	14	7	10	4.0	8	2.0
500 times $av(\hat{\pi},\hat{\theta})$	0.199	0.305	0.496	14.038	6.951	10.055	3.964	8.014	1.997
$se(\hat{\pi},\hat{\theta})$	0.039	0.047	0.052	0.021	0.107	0.030	0.072	0.007	0.034
1000 time $av(\hat{\pi},\hat{\theta})$	0.200	0.303	0.497	14.047	6.944	10.046	3.969	8.036	1.991
$se(\hat{\pi},\hat{\theta})$	0.040	0.046	0.051	0.026	0.108	0.025	0.072	0.017	0.034
5000 times $av(\hat{\pi},\hat{\theta})$	0.200	0.301	0.500	14.189	6.808	10.147	3.915	8.114	2.000
$se(\hat{\pi},\hat{\theta})$	0.040	0.047	0.051	0.098	0.110	0.066	0.071	0.034	0.033

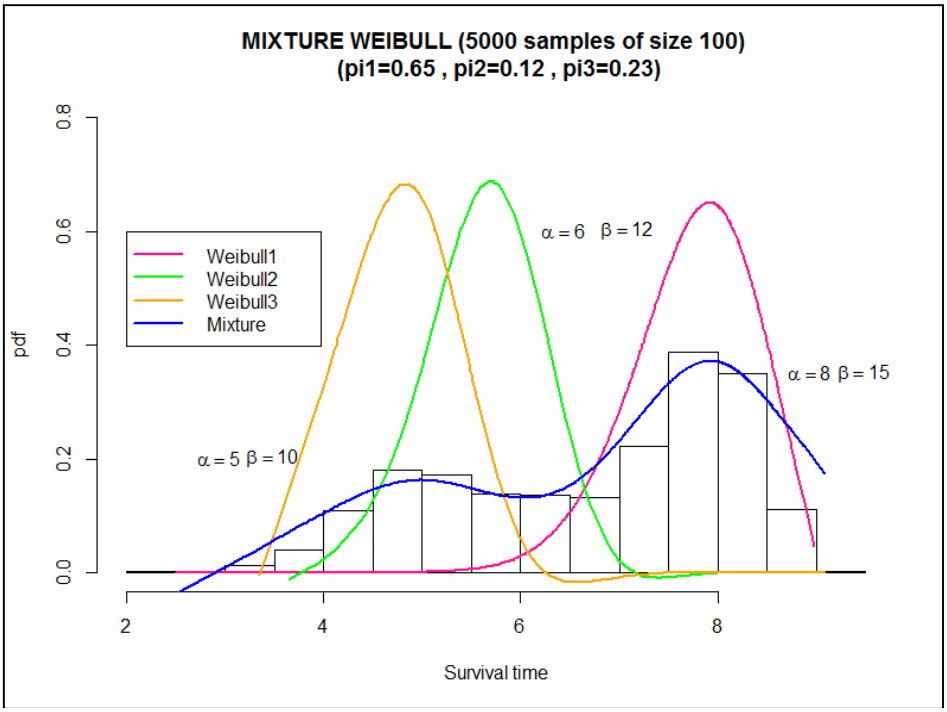


Fig 4: Density function of Weibull mixture versus single distribution

Table 5: Repeated Simulation for different Postulated Model of Gompertz mixture

Parameters	π_1	π_2	π_3	a_1	b_1	a_2	b_2	a_3	b_3
Postulated model 1	0.1	0.3	0.6	0.5	0.04	0.9	0.07	1	0.3
500 times $av(\hat{\pi},\hat{\theta})$	0.100	0.306	0.594	0.423	0.037	0.860	0.069	0.985	0.298
$se(\hat{\pi},\hat{\theta})$	0.029	0.045	0.050	0.004	0.014	0.002	0.013	0.001	0.038
1000 time $av(\hat{\pi},\hat{\theta})$	0.100	0.303	0.597	0.452	0.038	0.861	0.069	0.990	0.300
$se(\hat{\pi},\hat{\theta})$	0.030	0.045	0.050	0.004	0.014	0.002	0.013	0.001	0.039
5000 time $av(\hat{\pi},\hat{\theta})$	0.100	0.300	0.600	0.451	0.038	0.854	0.069	0.992	0.300
$se(\hat{\pi},\hat{\theta})$	0.030	0.047	0.050	0.003	0.015	0.001	0.013	0.001	0.038
Postulated model 2	0.10	0.45	0.45	2	0.8	1.5	0.5	1	0.2
500 times $av(\hat{\pi},\hat{\theta})$	0.100	0.454	0.447	1.839	0.805	1.478	0.503	0.970	0.198
$se(\hat{\pi},\hat{\theta})$	0.029	0.049	0.051	0.005	0.300	0.001	0.073	0.001	0.030
1000 time $av(\hat{\pi},\hat{\theta})$	0.100	0.452	0.448	1.934	0.822	1.480	0.504	0.980	0.199
$se(\hat{\pi},\hat{\theta})$	0.030	0.048	0.050	0.006	0.309	0.001	0.076	0.001	0.030
5000 time $av(\hat{\pi},\hat{\theta})$	0.099	0.451	0.450	1.941	0.832	1.480	0.500	0.983	0.201
$se(\hat{\pi},\hat{\theta})$	0.030	0.051	0.051	0.006	0.333	0.001	0.075	0.001	0.030
Postulated model 3	0.2	0.25	0.55	3	1.5	2.5	1.2	2	1
500 times $av(\hat{\pi},\hat{\theta})$	0.199	0.257	0.544	2.895	1.508	2.482	1.217	1.972	0.995
$se(\hat{\pi},\hat{\theta})$	0.039	0.045	0.053	0.003	0.339	0.002	0.244	0.001	0.134
1000 time $av(\hat{\pi},\hat{\theta})$	0.200	0.253	0.547	2.889	1.521	2.490	1.215	1.983	1.002
$se(\hat{\pi},\hat{\theta})$	0.040	0.044	0.052	0.003	0.356	0.002	0.255	0.001	0.135
5000 time $av(\hat{\pi},\hat{\theta})$	0.200	0.250	0.550	2.895	1.522	2.491	1.211	1.985	1.000
$se(\hat{\pi},\hat{\theta})$	0.040	0.044	0.053	0.003	0.377	0.001	0.252	0.001	0.132

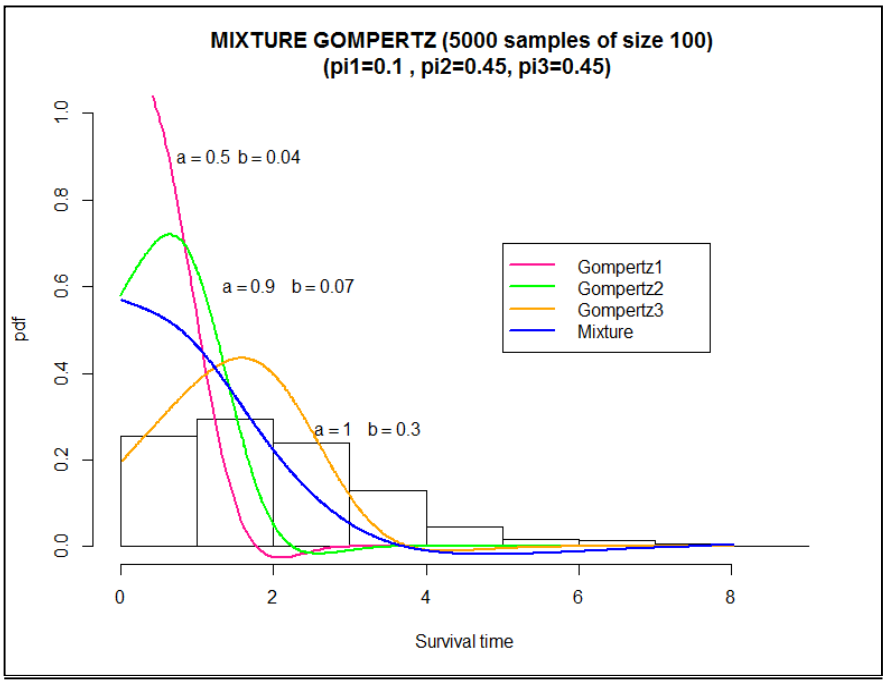


Fig 5: Density function of Gompertz mixture versus single distribution

The averages of the estimated parameters of the three components namely Exponential, Gamma, Weibull, Lognormal, Gompertz mixture model and their corresponding standard errors for the first, second and third postulated models are listed from **Table 1 – 5** respectively**.** It can be observed that the estimators get closer and closer to the true values (postulated model) of the three component mixture model as the number of repetitions increases from 500, 1000 and 5000 times that is the averages of the estimated parameters become exactly as same as that of the postulated models and their standard errors are relatively small which suggests that the estimators obtained through EM algorithm performed consistently. Convergence was achieved in all the cases, even though when the starting values are poor and this emphasizes the numerical stability of the EM algorithm. **Figure 1 - 5**, exhibits the comparison between the probability density function of the parametric mixture model Exponential, Gamma, Weibull, Lognormal and Gompertz distributions and the probability density functions of each single distribution. As it can be seen in the graph, the mixture model fits the simulated data far better than the single distributions. Simulation results revealed that EM algorithm approach works well with three identical mixture proportions.

4.2 Real Data based on Blood and Marrow Transplantation (EBMT)

The real data analyzed in this section is the Blood and Marrow transplantation data (EBMT). The data are included as one of the data sets in the famous *mstate* package which is available in R software. The dataset consists of survival times of 2204 patients. Three identical mixtures have been proposed for this dataset. The estimated parameter, Log-likelihood (LL), K-S test statistic, mean square error (MSE) values, Akaike information Criterion (AIC) for mixture of three Identical distributions such as Exponential-Exponential-Exponential,Gamma-Gamma-Gamma,Weibull-Weibull-Weibull,Lognormal-Lognormal-Lognormal,Gompertz-Gompertz-Gompertz are mentioned in **Table 6**

Table 6: Estimated Parameters, LL, K-S test Statistics, MSE and AIC values for Blood Marrow dataset

MODEL	E_E_E	L_L_L	G_G_G	W_W_W	GO_GO_GO
	$\hat{\pi}_1 = 0.588$	$\hat{\pi}_1 = 0.495$	$\hat{\pi}_1 = 0.448$	$\hat{\pi}_1 = 0.421$	$\hat{\pi}_1 = 0.613$
	$\hat{\pi}_2 = 0.235$	$\hat{\pi}_2 = 0.407$	$\hat{\pi}_2 = 0.305$	$\hat{\pi}_2 = 0.303$	$\hat{\pi}_2 = 0$
			$\hat{\pi}_3 = 0.247$	$\hat{\pi}_3 = 0.276$	$\pi_3 = 0.387$
	$\hat{\pi}_3 = 0.177$	$\hat{\pi}_3 = 0.098$	$\hat{\alpha}_1 = 8.907$	$\hat{\eta}_1 = 28.817$	$\hat{a}_1 = 0.0213$
	$\hat{\lambda}_1 = 40.631$	$\hat{\mu}_1 = 6.077$	$\hat{\beta}_1 = 3.033$	$\hat{\gamma}_1 = 3.438$	$\hat{b}_1 = 0.002$
	$\hat{\lambda}_2 = 1006.97$	$\hat{\sigma}_1 = 1.462$	$\hat{\alpha}_2 = 1.134$	$\hat{\eta}_2 = 147.751$	$\hat{a}_2 = 0.006$
	$\hat{\lambda}_3 = 1000.12$	$\hat{\mu}_2 = 3.192$	$\hat{\beta}_2 = 165.775$	$\hat{\gamma}_2 = 1.123$	$\hat{b}_2 = 0$
		$\hat{\sigma}_2 = 0.317$	$\hat{\alpha}_3 = 7.873$	$\hat{\eta}_3 = 1569.066$	$\hat{a}_3 = 0.011$
		$\hat{\mu}_3 = 3.861$	$\hat{\beta}_3 = 189.592$	$\hat{\gamma}_3 = 2.674$	$\hat{b}_3 = 0.001$
		$\hat{\sigma}_3 = 0.464$			
Log Lik	-14353.94	-14016.19	-13761.99	-13795.29	-14575.21
KS	0.1517	0.0649	0.0188	0.0291	0.1384
MSE	0.0039	0.0005	0	0.0001	0.0033
AIC	28717.89	28048.39	27539.99	27606.58	29166.41

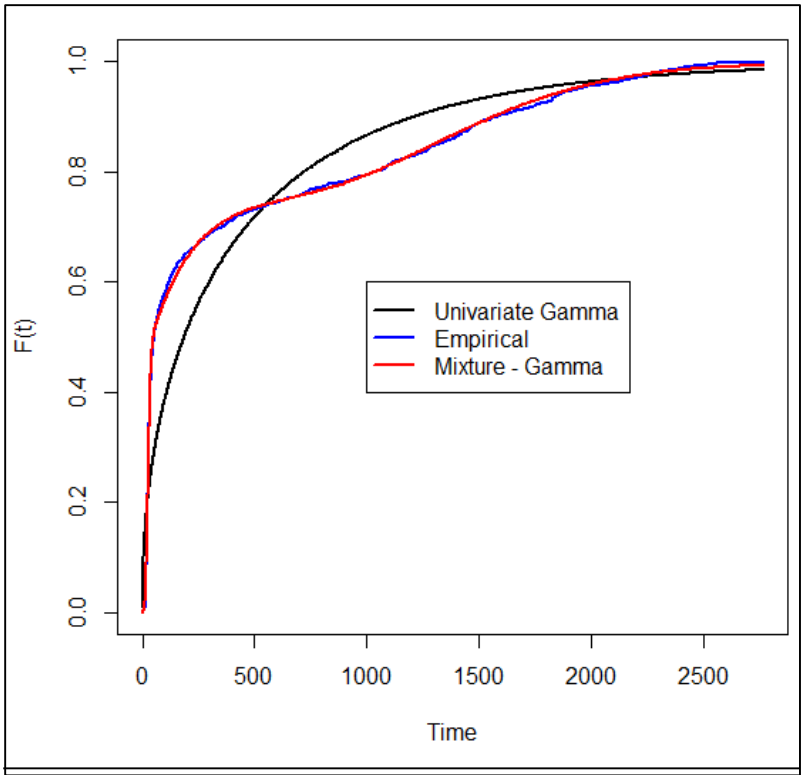


Fig 6: The Empirical distribution function and fitted distribution function for blood marrow dataset

It can be noted from Table 6, based on KS statistic, MSE and AIC values, Gamma mixture has the smallest KS test statistic, smallest MSE value and AIC value. Which suggest that Gamma mixture is the appropriate distribution for modeling Survival times of blood and marrow transplantation data set. Also, it can be viewed from Figure 6 that a graphical comparison of fitted (pure) pdf of Gamma distribution model and the fitted pdf of the Gamma mixture model for blood and marrow transplantation data set, where we can observe that mixture models of three identical Gamma distribution fits much better that the (pure) Gamma distribution to represent Survival times of blood and marrow transplantation dataset.

6. Conclusion

In this paper we proposed three component mixture model of Exponential, Gamma, Lognormal, Weibull and Gompertz distribution to model the heterogeneous survival data. Mixtures of Exponential-Exponential-Exponential, Weibull-Weibull-Weibull, Gamma-Gamma-Gamma, Lognormal-Lognormal-Lognormal and Gompertz-Gompertz-Gompertz distributions were tested for the best fit to the simulated dataset as well as real survival dataset. EM algorithm was employed in estimating the maximum likelihood estimator of the parameters.

The repetitions of the Simulation give estimators closer and closer to the postulated models, as the number of repetitions increases with relatively small standard errors. From Table 1-5, it shows that the EM algorithm converged to the true values (postulated model) of the mixture model parameters in 500, 100, 5000 repetitions and that emphasizes the stability of the algorithm in estimating the parameters with different proportion of mixing probabilities. The averages are close to the true values of the parameters and the standard errors are relatively small which suggest that the EM algorithm estimator performed consistently. Also, the graphs for three component mixture model fits the simulated data far better than the single distributions. According to the simulation results, the EM algorithm successfully estimated the parameters of the three component mixture model of identical distributions.

Also, we employ three identical mixture distributions for modeling Survival times of blood and marrow transplantation dataset. The AIC values, KS test statistics and MSE are calculated to determine the most appropriate distribution for the chosen data set. It can be noted from Table 6 that the best model among the three component mixture models of identical distribution is the mixture of Gamma for blood and marrow transplantation dataset to KS test statistics and MSE value and AIC values. Results revealed that mixture models are more flexible and they are better options to model heterogeneous survival data.

7. References

1. Kleinbanm DG, Klein M. Survival Analysis: A Self-Learning Text, Second Edition, Springer, 2005
2. Lee ET, Wang JW. Statistical Methods for Survival Data Analysis. Fourth Edition, John Wiley &Sons, Inc. All rights reserved, 2013.
3. Lawless JF. Statistics Models and Methods for Lifetime Data, Second Edition, John Wiley & Sons, New Jersey, 2003.
4. Erişoğlu Ü, Erişoğlu M, Erol H. A mixture model of two different distributions approach to the analysis of heterogeneous survival data. International Journal of Computer, Electrical, Automation, Control and Information Engineering. 2011; 5:6.
5. Erişoğlu Ü, Erişoğlu M, Erol H. Mixture model approach to the analysis of heterogeneous Survival time data. Pakistan Journal of Statistics. 2012; 28(1):115-130.

6. Ayça Hatice Türkan, Nazifçalış (2014) Comparison of Two-Component Mixture Distribution Models for Heterogeneous Survival Datasets: A Review Study. ˙ISTAT˙IST˙IK: Journal of the Turkish Statistical Association Vol.7, No. 2, July 2014,pp. 33–42 ISSN 1300-4077 | 14 | 2 | 33 | 42
7. Uma maheswari R, Leo Alexander T. Mixture of Identical Distributions of Exponential, Gamma, Lognormal, Weibull, Gompertz approach to Heterogeneous Survival time Data. International Journal of Current Research, 2017; 9(09):57521-575332.
8. Uma maheswari R, Leo Alexander T. Two- component of Non- Identical Mixture Distribution Models for heterogeneous Survival Data. International Journal of Recent Scientific Research. 2017; 8(10):20813-20824.
9. Blackstone EH, Naftel DC, Turner Jr ME. The decomposition of time-varying hazard into phases, each incorporating a separate stream of concomitant information. Journal of the American statistical Association. 1986; 81(395):615-624.
10. Ng ASK, McLachlan GJ, Yau KKW, Lee AH. Modelling the distribution of ischaemic stroke-specific survival time using an EM-based mixture approach with random effects adjustment. Statistics in Medicine. 2004; 23(17):2729-2744.
11. Phillips N, Coldman A, McBride ML. Estimating cancer prevalence using mixture models for cancer survival. Statistics in Medicine. 2002; 21(9):1257-1270
12. Yusuf A Mohammed, Bidin Yatim, Suzilah Ismail. A Simulation Study of a Parametric Mixture Model of Three Different Distributions to Analyze Heterogeneous Survival Data. Model Applied Science. URL, 2013; 7.
13. Yusuf A Mohammed, Bidin Yatim, Suzilah Ismail. Mixture model of the Exponential, Gamma, and Weibull distributions to Analyze Heterogeneous Survival Data. Journal of Scientific Research and Reports. 2013; 5(2):132-139.
14. Yusuf A Mohammed, Bidin Yatim, Suzilah Ismail. Survival Mixture Model of Gamma Distribution for Modeling Heterogeneous Data. International Journal of Applied Engineering Research ISSN 0973-4562 2016; II(16):8992-8998
15. Mclachlan GJ, Peel D. *Finite Mixture Model*. Wiley, New York, 2000.
16. Hogg Mckean Craig. Introduction to Mathematical Statistics. Sixth Edition, Published by Dorling Kindersley (India) Pvt. Ltd., licensees of Pearson Education in south Asia, 2005.
17. McLachlan GJ, Krishnan T. *The EM Algorithm and Extensions*. Wiley, New York, 1997.