Available online at http://www.journalcra.com

International Journal of Current Research
Vol. 9, Issue, 09, pp.57521-57532, September, 2017

INTERNATIONAL JOURNAL OF CURRENT RESEARCH

ISSN: 0975-833X

RESEARCH ARTICLE

MIXTURE OF IDENTICAL DISTRIBUTIONS OF EXPONENTIAL, GAMMA, LOGNORMAL, WEIBULL, GOMPERTZ APPROACH TO HETEROGENEOUS SURVIVAL DATA

[1]Uma maheswari, R. and *,[2]Leo Alexander, T.

[1]Research Scholar, Department of Statistics, Loyola College, Chennai–34, India
[2]Associate Professor, Department of Statistics, Loyola College, Chennai–34, India

ARTICLE INFO

Article History:
Received 12th June, 2017
Received in revised form
20th July, 2017
Accepted 23rd August, 2017
Published online 29th September, 2017

Key words:
Exponential, Gamma,
Weibull, Lognormal, Gompertz,
Mixture of identical distributions.

ABSTRACT

In this paper, a parametric mixture model of two identical distributions is proposed to analyze heterogeneous survival data. Mixtures of Exponential-Exponential, Weibull-Weibull, Gamma-Gamma, Lognormal-Lognormal and Gompertz-Gompertz distributions were tested for the best fit to the simulated datasets as well as real survival datasets. Various properties of the proposed mixture models were discussed. The Expectation Maximization Algorithm (EM) is implemented to estimate the maximum likelihood estimators of the parameters of mixture models. Simulations were performed by simulating data, each randomly sampled from a population of two component parametric mixture model of identical distributions and the simulations has 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 and closer to the postulated models, as the number of repetitions increases with relatively small standard errors. Model performances are compared using goodness of fit tests and Akaike's information criterion(AIC). Results revealed that the proposed model fits the real data better than the pure classical survival models corresponding to each component.

Citation: Uma maheswari, R. and Leo Alexander, 2017. "Mixture of identical distributions of exponential, gamma, lognormal, weibull, gompertz approach to heterogeneous survival data", *International Journal of Current Research*, 9, (09), 57521-57532.

INTRODUCTION

Survival time data analysis is a set of procedures to analyze the occurrence of a particular event over the 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 devices. It plays a vital role in the field of Social Sciences, Medical Studies, Economics and Engineering, etc. Historically, Nonparametric and Classical Parametric methods are commonly used to handle survival data. It is also used in the analysis of lifetime data. Among the univariate models, a few distributions like Exponential, Gamma and Weibull are mainly used for their demonstrated applications in a wide range of situations distributions (Kleinbanm, 2005 and Lee, 2013 and Lawless, 2003).Apart from these, there are other novel models for survival data that have been developed recently.

Mixture distributions are highly recommended to handle data with heterogeneous structure. Similarly, Mixture models are used to model failure-time data in a variety of situations. As a flexible way of modeling data, the mixture approach is directly applicable in situations where the adoption of a single parametric family is inadequate. Also, now-a-days, researchers would like to use mixture model technique to analyze survival time data. HirotuguAkaike (1974) introduced a new estimator which is called Akaike's theoretic criterion (AIC) for the purpose of statistical identification (Akaike, 1974).

Cheng *et al*., (1982) recommended a parametric mixture model of two Weibull distributions for the data that are grouped and censored by employing the weighted least squares method to estimate the parameters (Cheng, 1982). Blackstone *et al*., (1986) identified three overlapping phases of death after an open-heart surgery. This could be modeled by a three component parametric mixture model than the conventional parametric survival model (Blackstone, 1986). A mixture of Weibull component and a survival fraction has been used by Quiang (1994) in the context of a lung cancer trial (Quiang, 1994).

Corresponding author: **Leo Alexander, T.**
Associate Professor, Department of Statistics, Loyola College, Chennai–34, India

Angelis *et al*., (1999) proposed a parametric mixture model with an application to relative Survival model to individual data on colon cancer patients (Angelis, 1999). DankmarBohning *et al*., (2003) discussed the special issues on mixture models (DankmarBohning and Wilfried Seidel, 2003). Marín *et al*., (2005) used Bayesian analysis to fit a Weibull mixture survival model with an unknown number of components to possibly right-censored survival data (Marín, 2005). A parametric mixture model approach has been proposed and studied by Zhang (2008) for the analysis of Survival data (Zhang, 2008). Ersioglu *et al*., (2010) proposed a two component mixture model of the Extended Exponential-Geometric (EEG) distribution to model heterogeneous survival time data (Erişoğlu, 2010). Erisoglu.M *et al*., (2011) showed that the mixture of the identical and non – identical distributions of Weibull, Gamma, and Exponential are appropriate distributions for the earthquake inter occurrence times (Eri¸so˘glu, 2011a) Erisoglu *et al*., (2011) proposed a mixture of two different distributions such as Exponential-Gamma, Exponential-Weibull and Gamma-Weibull are the appropriate distributions to model heterogeneous survival data (Erişoğlu, 2011).

AyçaHaticeTürkan *et al*., (2014) 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 (AyçaHaticeTürkan, 2014). Muhammad Aslam *et al*., (2015) studied a 3-component mixture of the Rayleigh distributions in Bayesian perspective (Muhammad Aslam, 2015). Yusuf A. Mohammed *et al*., (2016) proposed a three components survival mixture model of the Gamma distribution for the analysis of heterogeneous survival data (Yusuf, 2016).

The purpose of this paper is to investigate the consistency and stability of EM in estimating the parameters and also show the appropriateness of a mixture of two identical (same) distributions in analyzing the heterogeneous survival time data. Our paper is organized as follows: In Section 2, we define the basics of survival analysis. Furthermore, several theoretical distributions have been used widely to describe survival time such as Exponential, Gamma, Weibull, lognormal, Gompertz distributions are discussed and their properties are highlighted. In Section 3, we define mixture of two identical distributions (ID) in survival analysis and the maximum likelihood estimators of the parameters are obtained by EM algorithm. In Section 4, Simulations were performed by simulating data, each randomly sampled from a population of two 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 examine the appropriateness of these mixture model distributions in analyzing the heterogeneous survival time data. The data got from National Institute for Research in Tuberculosis, Chetput, Chennai. Finally in Section 5, the summary and conclusion were presented. All computations are performed using R language.

2. Basics of Survival Analysis

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

(12) $S(t) = 1 - F(t) = \exp(-H(t)).$

2.1 Parametric survival distributions

A parametric survival model is a model in which survival time, thus the outcome, is assumed to follow a known distribution. By reviewing the literature about modeling the survival data, it can be seen that the Exponential, Gamma, Weibull, Lognormal and Gompertz probability distribution functions are commonly used in survival analysis. The probability density function $f(t)$ and the survival function $S(t)$ of Exponential, Gamma, Weibull, Lognormal and Gompertz distributions are summarized below (Kleinbanm, 2005 and Lee, 2013).

Exponential distribution

$f_{\exp}(t) = \frac{1}{\lambda} e^{-\frac{t}{\lambda}}$, $t > 0, \lambda > 0,$ where λ is a scale parameter of the distribution and reciprocal of it defined as rate parameter.

$s_{\exp}(t) = 1 - e^{-\frac{t}{\lambda}}$.

Weibull distribution

$f_{wbl}(t) = \frac{\gamma}{\eta}\left(\frac{t}{\eta}\right)^{\gamma-1} e^{-\left(\frac{t}{\eta}\right)^{\gamma}}$, $t, \eta, \gamma > 0,$ where η is a scale parameter and γ is called shape parameter.

$s_{wbl}(t) = e^{-\left(\frac{t}{\gamma}\right)^{\eta}}$.

Gamma distribution

$f_{gam}(t) = \frac{t^{\alpha-1} e^{-\frac{t}{\beta}}}{\beta^{\alpha}\Gamma(\alpha)}$, $t, \alpha, \beta > 0,$ where α is the shape parameter and β is the scale parameter.

$S_{gam}(t) = 1 - \frac{\Gamma_x(\alpha)}{\Gamma(\alpha)}$, where $\Gamma_x(\alpha)$ is called an incomplete Gamma function and calculated by $\Gamma_x(\alpha) = \int_0^x t^{\alpha-1} \exp(-t)dt$.

Lognormal distribution

$f_{\text{logn}}(t) = \frac{1}{\sigma t\sqrt{2\pi}} e^{-\frac{(\log t-\mu)^2}{2\sigma^2}}$, $t > 0, \mu, \sigma > 0,$ where μ is the location parameter and σ is the scale parameter.

$s_{\text{logn}}(t) = 1 - \phi\left(\frac{\log t - \mu}{\sigma}\right)$, where ϕ is cumulative distribution function of normal probability distribution function and is defined by $\phi\left(\frac{\log t - \mu}{\sigma}\right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\frac{\log t-\mu}{\sigma}} \exp\left(-\frac{u^2}{2}\right) du$.

Gompertz distribution

$f_{gomp}(t) = be^{at} e^{-\frac{b}{a}(e^{at}-1)}$, $t > 0, a, b > 0,$ where a is the shape parameter and b is the rate parameter.

$s_{gomp}(t) = e^{-\frac{b}{a}(e^{at}-1)}$.

3. Parametric Mixture models Approach in survival analysis

In finite mixture models, it is assumed that population consists of $g(\geq 2)$ distinct subgroups or subclasses. Finite mixture distribution function can be written as $f_{1,2}(t;\psi) = \pi f_1(t;\theta_1) + (1-\pi) f_2(t;\theta_2)$, where the vector $\psi' = (\pi, \theta)$ contains all the unknown parameters for π and $\theta = (\theta_1, \theta_2)'$ in the mixture model(14,10).The function $f_1(t;\theta_1)$ is called mixture component density function for the first population with parameter θ_1 and $f_2(t;\theta_2)$ is called mixture component density function for the Second population with parameter θ_2.

In this study, to model the heterogeneous survival time data, we use the mixture of two identical distributions. The Identical distributions are Exponential-Exponential, Gamma-Gamma, Weibull-Weibull, Lognormal-Lognormal and Gompertz –Gompertz which are defined as follows,

$$f_{\exp-\exp}(t) = \pi f_{\exp}(t;\lambda_1)+(1-\pi)f_{\exp}(t;\lambda_2) \quad (3.1)$$

$$f_{gam-gam}(t) = \pi f_{gam}(t;\alpha_1,\beta_1)+(1-\pi)f_{gam}(t;\alpha_2,\beta_2) \quad (3.2)$$

$$f_{wbl-wbl}(t) = \pi f_{wbl}(t;\eta_1,\gamma_1)+(1-\pi)f_{wbl}(t;\eta_2,\gamma_2) \quad (3.3)$$

$$f_{\text{logn}-\text{logn}}(t) = \pi f_{\text{logn}}(t;\mu_1,\sigma_1)+(1-\pi)f_{\text{logn}}(t;\mu_2,\sigma_2) \quad (3.4)$$

and $$f_{gomp-gomp}(t) = \pi f_{gomp}(t;a_1,b_1)+(1-\pi)f_{gomp}(t;a_2,b_2), \quad (3.5)$$

where π is the mixture weight of the distributions and $\pi \in (0,1)$. The maximum likelihood estimators of parameters of these mixture distributions are estimated using Expectation-Maximization (EM) algorithm.

3.1 Parameter estimation in mixture model

In finite mixture models, the EM (Expectation-Maximization) algorithm has been used as an effective method to find the unknown parameters by maximum likelihood (Mclachlan and Peel, 2000; Hogg MckeanCraig, 2005; McLachlan and Krishnan, 1997). In EM framework, the observed data $t_1, t_2, ..., t_n$ is considered as an incomplete data and latent class variables z_1, z_2 to be missing where $z_{1i} = z_1(t_i) = 1$ for $i = 1,..., n$ if observation t_i belongs to 1st class and 0 otherwise. The EM algorithm is applied to the mixture distributions by treating z_i as unobserved or missing data. It has two iterative steps, E (for Expectation) and M (for Maximization).

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

So, $$\hat{z}_{1i} = E_{\psi_0}(z_{1i}|t_i) = \frac{\pi f_{1,0}(t_i;\theta_1)}{\pi f_{1,0}(t_i;\theta_1)+(1-\pi)f_{2,0}(t_i;\theta_2)}$$

and $$\hat{z}_{2i} = E_{\psi 0}(z_{2i}|t_i) = \frac{(1-\pi)f_{2,0}(t_i;\theta_2)}{\pi f_{1,0}(t_i;\theta_1)+(1-\pi)f_{2,0}(t_i;\theta_2)} \quad .$$

In M-step, $E(Z_{1i}|t_i)$ and $E(Z_{2i}|t_i)$ function which are calculated in E-step is maximized. The M-step and E- step should be iterated alternatively till the convergence criterion is met. The estimator of π_k $(k = 1,2)$ is obtained as $\hat{\pi}_k = \frac{\sum_{i=1}^{n}\hat{z}_{ki}}{n}$.

By using equation (3.1), we can evaluate mixture of exponential distribution and the maximum likelihood estimator for Exponential-Exponentialis given by,

$$\hat{\lambda}_1 = \frac{\sum_{i=1}^{n}\hat{z}_{1i}t_i}{\sum_{i=1}^{n}\hat{z}_{1i}} \quad \text{and} \quad \hat{\lambda}_2 = \frac{\sum_{i=1}^{n}\hat{z}_{2i}t_i}{\sum_{i=1}^{n}\hat{z}_{2i}}.$$

By using equation (3.2), we can evaluate mixture of Gamma distribution and the maximum likelihood estimator for Gamma-Gamma is given by,

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}\hat{z}_{1i}t_i}{\hat{\alpha}_1\sum_{i=1}^{n}\hat{z}_{1i}} \quad \text{and} \quad \hat{\alpha}_1^{r+1} = \hat{\alpha}_1^{r} - \frac{\log(\hat{\alpha}_1^r)-\psi'(\hat{\alpha}_1^r)-\log\left(\frac{\sum_{i=1}^{n}\hat{z}_{1i}t_i}{\sum_{i=1}^{n}\hat{z}_{1i}}\right)+\frac{\sum_{i=1}^{n}\hat{z}_{1i}\log t_i}{\sum_{i=1}^{n}\hat{z}_{1i}}}{\frac{1}{\hat{\alpha}_1^r}-\psi'(\hat{\alpha}_1^r)} \quad .$$

Similarly,

$$\hat{\beta}_2 = \frac{\sum_{i=1}^{n} \hat{z}_{2i} t_i}{\hat{\alpha}_2 \sum_{i=1}^{n} \hat{z}_{2i}} \quad \text{and} \quad \hat{\alpha}_2^{r+1} = \hat{\alpha}_2^{r} - \frac{\log(\hat{\alpha}_2^{r}) - \psi'(\hat{\alpha}_2^{r}) - \log\left(\frac{\sum_{i=1}^{n} \hat{z}_{2i} t_i}{\sum_{i=1}^{n} \hat{z}_{2i}}\right) + \frac{\sum_{i=1}^{n} \hat{z}_{2i} \log t_i}{\sum_{i=1}^{n} \hat{z}_{2i}}}{\frac{1}{\hat{\alpha}_2^{r}} - \psi'(\hat{\alpha}_2^{r})} .$$

Here r is the number of Newton-Raphson iterations within EM algorithm. Also $\psi(.)$ and $\psi'(.)$ are a digamma and trigamma functions respectively. By using equation (3.3), we can evaluate mixture of Weibull distribution and the maximum likelihood estimator for Weibull-Weibull is given by,

$$\hat{\eta}_1 = \left(\left(\sum_{i=1}^{n} \hat{z}_{1i} \right)^{-1} \sum_{i=1}^{n} \hat{z}_{1i} t_i^{\hat{\gamma}_1} \right)^{1/\hat{\gamma}_1}, \qquad \hat{\gamma}_1^{r+1} = \hat{\gamma}_1^{r} + \frac{A_r^* + \left(1/\hat{\gamma}_1^{r}\right) - \left(C_r^*/B_r^*\right)}{(1/(\hat{\gamma}_1^{r})^2) + (B_r^* D_r^* - C_r^{*2})/B_r^{*2}}$$

$$\hat{\eta}_2 = \left(\left(\sum_{i=1}^{n} \hat{z}_{2i} \right)^{-1} \sum_{i=1}^{n} \hat{z}_{2i} t_i^{\hat{\gamma}_2} \right)^{1/\hat{\gamma}_2} \text{ and } \quad \hat{\gamma}_2^{r+1} = \hat{\gamma}_2^{r} + \frac{A_r^{**} + \left(1/\hat{\gamma}_2^{r}\right) - \left(C_r^{**}/B_r^{**}\right)}{(1/(\hat{\gamma}_2^{r})^2) + (B_r^{**} D_r^{**} - C_r^{**2})/B_r^{**2}},$$

where $A_r^* = \left(\sum_{i=1}^{n} \hat{z}_{1i} \right)^{-1} \sum_{i=1}^{n} \hat{z}_{1i} \log t_i$, $B_r^* = \sum_{i=1}^{n} \hat{z}_{1i} t_i^{\hat{\gamma}_1^{r}}$, $C_r^* = \sum_{i=1}^{n} \hat{z}_{1i} t_i^{\hat{\gamma}_1^{r}} \log t_i$ and $D_r^* = \sum_{i=1}^{n} \hat{z}_{1i} t_i^{\hat{\gamma}_1^{r}} (\log t_i)^2$,

$A_r^{**} = \left(\sum_{i=1}^{n} \hat{z}_{2i} \right)^{-1} \sum_{i=1}^{n} \hat{z}_{2i} \log t_i$, $B_r^{**} = \sum_{i=1}^{n} \hat{z}_{2i} t_i^{\hat{\gamma}_2^{r}}$, $C_r^{**} = \sum_{i=1}^{n} \hat{z}_{2i} t_i^{\hat{\gamma}_2^{r}} \log t_i$ and $D_r^{**} = \sum_{i=1}^{n} \hat{z}_{2i} t_i^{\hat{\gamma}_2^{r}} (\log t_i)^2$.

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.

By using equation (3.4), we can evaluate mixture of Lognormal distribution and the maximum likelihood estimator for Lognormal-Lognormal is given by,

$$\hat{\mu}_1 = \frac{\sum_{i=1}^{n} \hat{z}_{1i} \ln t_i}{\sum_{i=1}^{n} \hat{z}_{1i}} \quad \text{and} \quad \hat{\sigma}_1^2 = \frac{\sum_{i=1}^{n} \hat{z}_{1i} (\ln t_i - \hat{\mu}_1)^2}{\sum_{i=1}^{n} \hat{z}_{1i}}.$$

Similarly,

$$\hat{\mu}_2 = \frac{\sum_{i=1}^{n} \hat{z}_{2i} \ln t_i}{\sum_{i=1}^{n} \hat{z}_{2i}} \quad \text{and} \quad \hat{\sigma}_2^2 = \frac{\sum_{i=1}^{n} \hat{z}_{2i} (\ln t_i - \hat{\mu}_2)^2}{\sum_{i=1}^{n} \hat{z}_{2i}}.$$

By using equation (3.5), we can evaluate mixture of Gompertz distribution and the maximum likelihood estimator for Gompertz-Gompertzis given by,

$$\hat{b}_1 = \frac{\hat{a}_1 \sum_{i=1}^{n} \hat{z}_{1i}}{\sum_{i=1}^{n} \hat{z}_{1i} e^{\hat{a}_1 t_i} - \sum_{i=1}^{n} \hat{z}_{1i}}, \quad \hat{b}_2 = \frac{\hat{a}_2 \sum_{i=1}^{n} \hat{z}_{2i}}{\sum_{i=1}^{n} \hat{z}_{2i} e^{\hat{a}_2 t_i} - \sum_{i=1}^{n} \hat{z}_{2i}},$$

$$\hat{a}_1^{r+1} = \hat{a}_1^{r} + \frac{E_r^* + \left\{ (F_r^* G_r^* - \hat{a}_1^{r} F_r^* H_r^* - (F_r^*)^2) / \hat{a}_1^{r} (G_r^* - F_r^*) \right\}}{\left\{ \frac{(F_r^* (G_r^*)^2 - 2(F_r^*)^2 G_r^* + (\hat{a}_1^{r})^2 G_r^* F_r^* I_r^* - (\hat{a}_1^{r})^2 (F_r^*)^2 I_r^* - (\hat{a}_1^{r})^2 F_r^* (H_r^*)^2 + (F_r^*)^3)}{(\hat{a}_1^{r} (G_r^* - F_r^*))^2} \right\}},$$

where $E_r^* = \sum_{i=1}^{n} \hat{z}_{1i} t_i$, $F_r^* = \sum_{i=1}^{n} \hat{z}_{1i}$, $G_r^* = \sum_{i=1}^{n} \hat{z}_{1i} e^{\hat{a}_1^r t_i}$, $H_r^* = \sum_{i=1}^{n} \hat{z}_{1i} t_i e^{\hat{a}_1^r t_i}$ and $I_r^* = \sum_{i=1}^{n} \hat{z}_{1i} t_i^2 e^{\hat{a}_1^r t_i}$ and

$$\hat{a}_2^{r+1} = \hat{a}_2^r + \frac{E_r^{**} + \left\{ (F_r^{**} G_r^{**} - \hat{a}_2^r F_r^{**} H_r^{**} - (F_r^{**})^2) / \hat{a}_2^r (G_r^{**} - F_r^{**}) \right\}}{\left\{ \frac{(F_r^{**}(G_r^{**})^2 - 2(F_r^{**})^2 G_r^{**} + (\hat{a}_2^r)^2 G_r^{**} F_r^{**} I_r^{**} - (\hat{a}_2^r)^2 (F_r^{**})^2 I_r^{**} - (\hat{a}_2^r)^2 F_r^{**} (H_r^{**})^2 + (F_r^{**})^3)}{(\hat{a}_2^r (G_r^{**} - F_r^{**}))^2} \right\}},$$

where $E_r^{**} = \sum_{i=1}^{n} \hat{z}_{2i} t_i$, $F_r^{**} = \sum_{i=1}^{n} \hat{z}_{2i}$, $G_r^{**} = \sum_{i=1}^{n} \hat{z}_{2i} e^{\hat{a}_2^r t_i}$, $H_r^{**} = \sum_{i=1}^{n} \hat{z}_{2i} t_i e^{\hat{a}_2^r t_i}$ and $I_r^{**} = \sum_{i=1}^{n} \hat{z}_{2i} t_i^2 e^{\hat{a}_2^r t_i}$.

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.

3.2 Model Selection Criteria

To find the appropriate distribution, we use two different goodness of fit tests: the mean square error (MSE) test and the Kolmogorov-Smirnov (KS) test. Let us first use the MSE test. The MSE value is defined as

$$MSE = \frac{\sum_{i=1}^{n} \left[F_e(t_i) - F(t_i) \right]^2}{n-k},$$

where $F_e(t)$ is the empirical distribution and $F(t)$ is the cumulative distribution function that is proposed to model the heterogeneous survival data set. Here k is the number of free parameters in the distribution. As it is known, the smallest MSE value reveals the most appropriate distribution. Then The Kolmogorov-Smirnov statistic KS is defined by

$$KS = \max \left| F_e(t) - F(t) \right| .$$

It is known that the preferred distribution has the smallest value of KS. Also, we use AIC as goodness of fit test for model selection criteria. AIC value is as follows $AIC = -2 \log L + 2d$, where d represents estimated parameter (Mclachlan and Peel, 2000). The smallest AIC value represents the best model.

4. Simulation and Application

4.1 Simulation

Simulations are performed by simulating data, each randomly sampled from a population of two 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. The mixture model includes Exponential-Exponential, Gamma-Gamma, Weibull-Weibull, Lognormal-Lognormal and Gompertz-Gompertz. 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 sets are listed in the following Tables 1 –5, which gives 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 two identical distributions for simulation parameters are shown in the following Figures 1- 5. From Figure 1 - 5, it displays the comparison between pdf of identical mixture model and pdf of each single distribution.

4.1.1 Mixture model of exponential-exponential

Table 4.1

Exponential- Exponential			
Parameters	π	λ_1	λ_2
Postulated model	$\pi = 0.5$	$\lambda_1 = 0.25$	$\lambda_2 = 1.0$
$av(\hat{\pi}, \hat{\theta})$ 5000 times	**0.500**	**0.250**	0.999
$se(\hat{\pi}, \hat{\theta})$	**0.051**	**0.553**	0.135

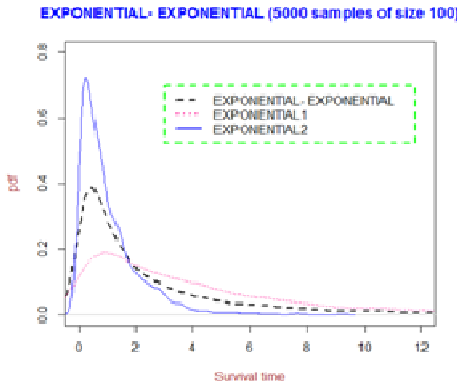

Figure 4.1

4.1.2 Mixture model of Gamma - Gamma

Table 4.2

Gamma- Gamma					
Parameters	π	α_1	β_1	α_2	β_2
Postulated model	$\pi = 0.3$	$\alpha_1 = 9$	$\beta_1 = 0.5$	$\alpha_2 = 3$	$\beta_2 = 2$
5000 times $av(\hat{\pi}, \hat{\theta})$	**0.300**	9.259	0.486	3.030	1.982
$se(\hat{\pi}, \hat{\theta})$	**0.045**	0.197	0.029	0.036	0.133

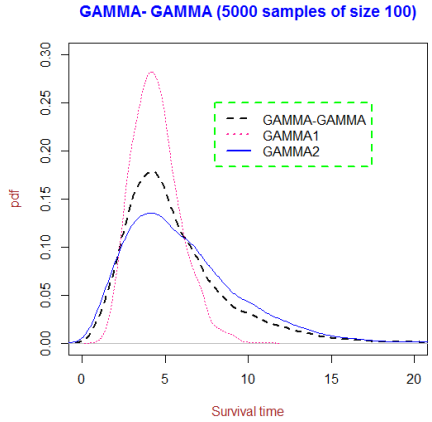

Figure 4.2

4.13 Mixture Model of Weibull-Weibull

Table 4.3

Weibull –Weibull					
Parameters	π	η_1	γ_1	η_2	γ_2
Postulated model	$\pi = 0.4$	$\eta_1 = 4$	$\gamma_1 = 8$	$\eta_2 = 2$	$\gamma_2 = 6$
5000 times $av(\hat{\pi}, \hat{\theta})$	0.400	3.997	8.084	1.998	6.077
$se(\hat{\pi}, \hat{\theta})$	0.050	0.075	0.037	0.040	0.019

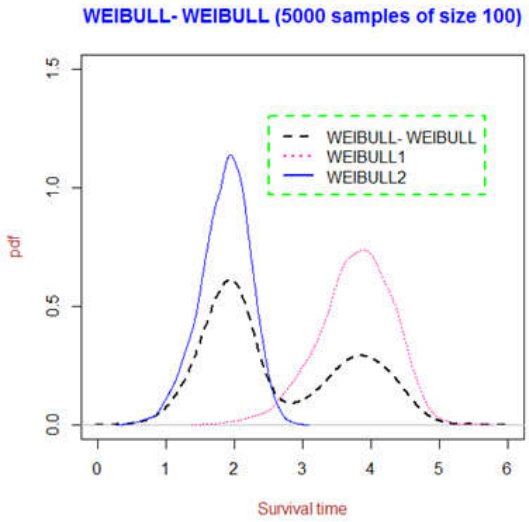

Figure 4.3

4.1.4 Mixture model of Lognormal- Lognormal

Table 4.4

Lognormal-Lognormal					
Postulated model	π	μ_1	σ_1	μ_2	σ_2
Parameters	$\pi = 0.7$	$\mu_1 = 2$	$\sigma_1 = 0.5$	$\mu_2 = 2$	$\sigma_2 = 1$
5000 times $av(\hat{\pi},\hat{\theta})$	0.700	2.000	0.499	2.001	0.993
$se(\hat{\pi},\hat{\theta})$	0.046	0.056	0.041	0.178	0.128

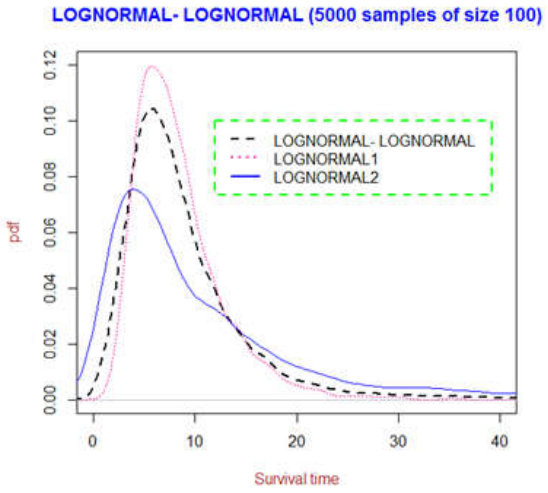

Figure 4.4

4.1.5 Mixture Model of Gompertz- Gompertz

Table 4.5

Gompertz- Gompertz					
Parameters	π	a_1	b_1	a_2	b_2
Postulated model	$\pi = 0.4$	$a_1 = 2$	$b_1 = 1$	$a_2 = 6$	$b_2 = 2$
5000 times $av(\hat{\pi},\hat{\theta})$	0.400	1.971	1.005	5.955	2.017
$se(\hat{\pi},\hat{\theta})$	0.050	0.001	0.161	0.001	0.253

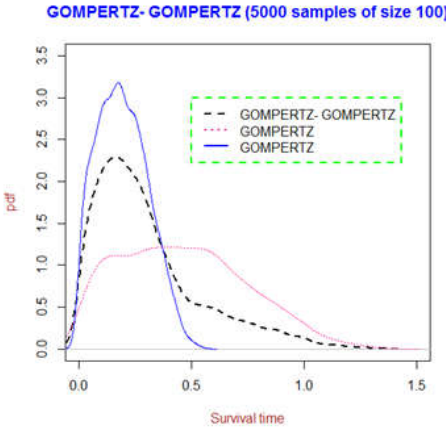

Figure 4.5

The result of the parameter estimation listed from Table4.1-4.5, shows the averages of the estimated parameters of Exponential-Exponential,Gamma-Gamma, Weibull-Weibull, Lognormal-Lognormal, Gompertz-Gompertzmixture model and its corresponding standard error respectively. It can be observed that the estimators get closer to the true values (postulated model) of the mixture model as the number of repetitions increases i.e., the averages of the estimators are very close to the true values of the parameters 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 4.1- 4.5, exhibits the comparison between the probability density function of the parametric mixture model Exponential, Gamma, Weibull, Lognormal and Gompertzdistributions 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 identical mixture proportions

4.2Illustrativeexample based on bone marrow Survival Data

Bone marrow survival data set consists of survival times of 137 patients. The data has been collected from National Institute for Research in Tuberculosis (NIRT), Chennai. Mixtures of identical distributions have been proposed for the data set. The estimated parameter, Log-likelihood (LL), K-S test statistic, mean square error (MSE) values, Akaike information Criterion (AIC) for mixture of identical distributions such as Exponential-Exponential, Gamma-Gamma, Weibull-Weibull, Lognormal-Lognormal, Gompertz-Gompertzare mentioned in Table 4.6

Table 4.6 The estimated Parameters, Ll values, k-s test statistics, mse values and aic for bone marrow dataset

S.NO	MODELS	ESTIMATES		π_1	π_2	LL	KS	MSE	AIC
1	Exp-Exp	$\hat{\lambda}_1 = 335.0073$	$\hat{\lambda}_2 = 334.0071$	0.4633	0.5367	-933.5389	0.2174	0.0160	1873.07
2	Gam-Gam	$\hat{\alpha}_1 = 0.9673$	$\hat{\alpha}_2 = 15.7347$	0.7010	0.2990	-921.7987	0.0276	0.0002	1853.60
		$\hat{\beta}_1 = 192.5961$	$\hat{\beta}_2 = 43.4484$						
3	Wbl-Wbl	$\hat{\eta}_1 = 740.5906$	$\hat{\eta}_2 = 179.4835$	0.3145	0.6855	-920.3076	0.0280	0.0002	1850.62
		$\hat{\gamma}_1 = 4.3974$	$\hat{\gamma}_2 = 0.9967$						
4	Logn-Logn	$\hat{\mu}_1 = 4.6016$	$\hat{\mu}_2 = 6.4400$	0.6818	0.3182	- 931.5698	0.0643	0.0011	1873.14
		$\hat{\sigma}_1 = 1.3459$	$\hat{\sigma}_2 = 0.2814$						
5	Gomp-Gomp	$\hat{a}_1 = 0.0020$	$\hat{a}_2 = 0.0074$	0.9207	0.0793	-930.6647	0.0776	0.0026	1871.33
		$\hat{b}_1 = 0.0011$	$\hat{b}_2 = 0.0026$						

From Table 4.6,it can be viewed that based on KS statistic, MSE and AIC values, Gamma-Gamma mixture has the smallest KS test statistic and smallest MSE value and it is the best model for bone marrow survival data set. According to MSE and AIC comparison values Weibull-Weibull mixture has least value which is also considered another best model for the same data set. Therefore, it can be observed from Table 6, Gamma-Gamma and Weibull-Weibull mixture models are best models for bone marrow survival data set.

A graphical comparison of the fitted (pure) pdf of Gamma and Weibull distribution and fitted pdf of the mixture models of Gamma- Gamma and Weibull- Weibull for survival times of bone marrow data set is mentioned in Figure 4.6(a) and Figure 4.6(b)

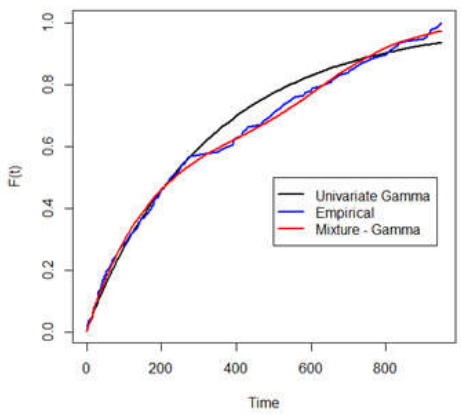

Figure 4.6(a).

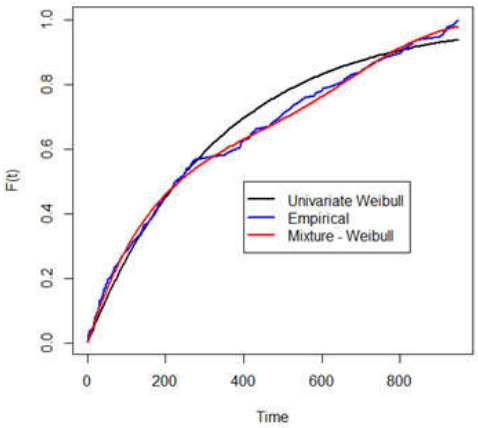

Figure 4.6(b)

From Figure 6(a) and 6(b), mixture models of Gamma-Gamma and Weibull-Weibull fit much better than (pure) Gamma and Weibull distributions for survival times of bone marrow data set.

5. Conclusion

In this paper we proposed the mixture models of two identical distributions such as Exponential- Exponential, Gamma-Gamma, Weibull-Weibull, Lognormal-Lognormal and Gompertz-Gompertz to represent the heterogeneous survival data sets. Heterogeneous survival time data can have two different distributions before and after a certain time due to many factors which affects the life of the creatures. For instance, a slowly growing tumor can grow faster after a particular process and this can affect the life time. Each of the different phases of life will generate a peak in the mixture distribution. Therefore, we try to model the heterogeneous survival time data with the most appropriate distributions among the mixture models. Mixtures of Exponential-Exponential, Weibull-Weibull, Gamma-Gamma, Lognormal-Lognormal and Gompertz-Gompertz distributions were tested for the best fit to the simulated dataset as well as real survival dataset. The maximum likelihood estimations of parameters of the mixture models obtained with EM algorithm. 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 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 entire the two component mixture model fit the simulated data far better than the single distributions. According to the simulation results, the EM algorithm successfully estimated the parameters of the two component mixture model of identical distributions.

Also, we employ mixture of identical distributions for modeling Survival times for bone marrow dataset. The AIC values, KS test statistics and MSE are calculated to determine the most appropriate distribution for the present data set. It can be noted from Table 6 that the best model among the two component mixture models of identical distribution is the mixture of Gamma-Gamma for Survival times of bone marrow patients according to KS test statistics and MSE value. And according to MSE and AIC values, Weibull-Weibull mixture is another best model for the same data set respectively. The histogram and the two probability densities of Gamma-Gamma and Weibull-Weibull mixture fit better than others for the survival times of 137 Bone marrow patients that isgiven in Figure 4.7(a). The empirical distribution function and two distribution functions Gamma-Gamma and Weibull-Weibull fitbetter is shown in Figure 4.7(b).

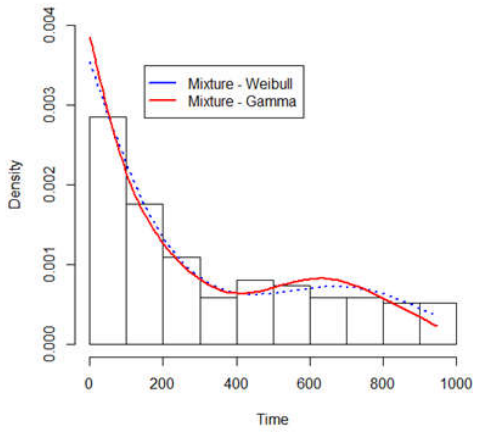

Figure 4.7(a). The probability densities of the fitted distributions and a histogram

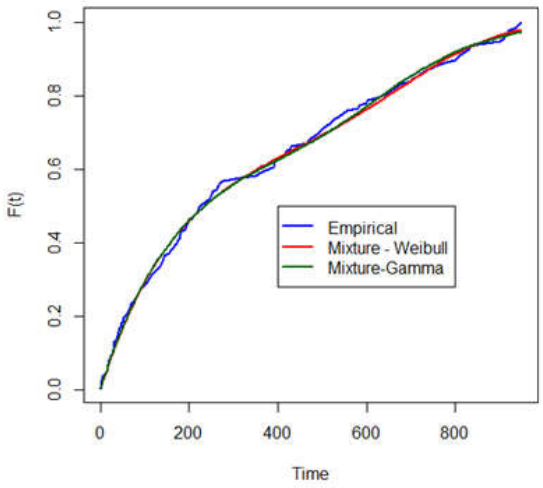

Figure 4.7(b). The empirical distribution function and the fitted distribution function for bone marrow dataset

Acknowledgement

We would like to thank Dr.C.Ponnuraja, Scientist–D, National Institute for Research in Tuberculosis, ICMR, Chetput, Chennai who helped in providing data.

REFERENCES

Akaike H. 1974. A New Look at the Statistical Model Identification. *IEEE. Transactionson Automatic Control*, Ac 19, 716-723..

Angelis R. De, Capocaccia R., Hakulinen T., Soderman B. and Verdecchia A., Mixture Models for Cancer Survival Analysis: Application to Population-Based Data With Covariates, Statistics in Medicine, 18, 441-454, 1999.

AyçaHaticeTürkan and Nazifçaliş 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

Blackstone, E. H., Naftel, D. C., & Turner Jr., M. E. 1986. The decomposition of time-varying hazard into phases, each incorporating a separate stream of concomitant information. *Journal of the American statistical Association* 81(395), 615-624 http://dx.doi.org/10.1080/01621459.1986.10478314

Cheng, S. W., & Fu, J. C. 1982. Estimation of mixed Weibull parameters in life testing. Reliability,IEEE Transactions on, R-31(4), 377-381. http://dx.doi.org/10.1109/ TR.1982.5221382

DankmarBohningand Wilfried Seidel 2003. Editorial: Recent developments in mixture models. Computational Statistics & Data Analysis 41 (2003) 349 – 357.

Eri¸so˘glu M, C¸ alı¸s N, Servi T, Eri¸saho˘glu ¨ U, Topaksu M 2011a. The mixture distribution models for interoccurence times of earthquakes. Russian Geology and Geophysics 52(2011):685-692.

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

Erişoğlu, Ü.,Erişoğlu, M., & Erol, H. 2011. 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 Vol:5, No:6,

Hogg MckeanCraig 2005. Introduction to Mathematical Statistics. Sixth Edition, Published by Dorling Kindersley (India) Pvt.Ltd.,licensees of Pearson Education in south Asia.

Kleinbanm D.G. and Klein M., Survival Analysis: A Self-Learning Text, Second Edition, Springer, 2005

Lawless J.F., Statistics Models and Methods for Lifetime Data, Second Edition, John Wiley & Sons, New Jersey, 2003

Lee E.T. and Wang J.W. 2013. Statistical Methods for Survival Data Analysis. Fourth Edition, John Wiley &Sons, Inc.All rights reserved

Marín,J. M., Rodríguez-Bernal, M. T.and Wiper, M. P. 2005. Using Weibull Mixture Distributions to Model Heterogeneous Survival Data, Communications in Statistics - Simulation and Computation, 34:3, 673-684, DOI: 10.1081/SAC-200068372

McLachlan G.J. and Krishnan T. 1997. *The EM Algorithm and Extensions*. Wiley, New York.

Mclachlan G.J. and Peel D. 2000. *Finite Mixture Model*. Wiley, New York.

Muhammad Aslam, Muhammad Tahir, Zawar Hussain, Bander Al-Zahrani 2015. A 3-Component Mixture of Rayleigh Distributions: Properties and Estimation in Bayesian Framework . PLOS ONE | DOI:10.1371/journal.pone.0126183

Quiang J., A Bayesian Weibull Survival Model. Unpublished Ph.D. Thesis, Institute of Statistical and Decision Sciences, Duke University: North Corolina, 1994.

Yusuf A. Mohammed, BidinYatim and Suzilah Ismail 2016. Survival Mixture Model Of Gamma Distribution For Modeling Heterogeneous Data. International Journal of Applied Engineering Research ISSN 0973-4562 volume II, number 16(2016) pp 8992-8998.http://www.ripublication.com

Zhang, Y. 2008. Parametric mixture models in survival analysis with application, (Doctoral Dissertation) UMI : 3300387, Graduate School, Temple University.

****