World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:5, No:6, 2011

A Mixture Model of Two Different Distributions Approach to the Analysis of Heterogeneous Survival Data

Ülkü Erişoğlu, Murat Erişoğlu and Hamza Erol

Open Science Index, Computer and Information Engineering Vol:5, No:6, 2011 publications.waset.org/2675.pdf

***Abstract*—**In this paper we propose a mixture of two different distributions such as Exponential-Gamma, Exponential-Weibull and Gamma-Weibull to model heterogeneous survival data. Various properties of the proposed mixture of two different distributions are discussed. Maximum likelihood estimations of the parameters are obtained by using the EM algorithm. Illustrative example based on real data are also given.

***Keywords*—**Exponential-Gamma, Exponential-Weibull, Gamma-Weibull, EM Algorithm, Survival Analysis.

I. INTRODUCTION

SURVIVAL analysis is a collection of statistical procedures for data analysis for which the outcome variable of interest is time until an event occurs. Historically, survival analysis has usually been carried out using nonparametric methods or via the classical statistical analysis of parametric survival models. Survival analysis datasets have been usually represented with the classical statistical distributions such as Gamma, Exponential and Weibull distributions [7]-[10]. Besides these pure classical statistical distribution models, other novel models for survival data have been developed recently. Especially in the heterogeneous structure of the data model in the use of mixed distribution has become widespread. Frequency distribution in case of single-mode data with the standard model of probability distribution is useful and helpful. Mixture of distributions is even useful because it is applied to represent heterogeneous data set which there is evidence of multimodality or simply unimodality [5]. Chen et al. [3] used a two-component mixture model for the analysis of cancer survival data generalizing an earlier idea in Berkson and Gage [2]. In Quiang [13], a similar model of a mixture of a Weibull component and a surviving fraction in the context of a lung cancer trial is considered. Angelis et al.[1] proposed an application of a parametric mixture model to relative survival rates of colon cancer patients from the Finnish population-based cancer registry, and including major survival determinants as explicative covariates. Marin et al.[11] have illustrated how Bayesian methods can be used to fit a mixture of Weibulls model with an unknown number of components to heterogeneous, possibly right-censored survival data using a birth death MCMC algorithm.

The purpose of this paper is to show that the mixture of the different distributions is the appropriate distribution for the heterogeneous survival times. The article is organized as follows. In Sec. 2, we define the functions of survival time. Also several theoretical distributions that have been used widely to describe survival time are discussed and their characteristics summarized. In Sec. 3, we define mixture of two different distributions in survival analysis and the maximum likelihood estimations of the parameters are obtained by the EM algorithm. In Sec. 4, mixture of two different distributions is applied on illustrative examples based on heterogeneous survival real dataset successfully.

II. FUNCTIONS OF SURVIVAL TIME

Survival time data measure the time to a certain event, 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:

$$S(t) = P(T > t), 0 < t < \infty \tag{1}$$

Here $S(t)$ is a nonincreasing 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

$$F(t) = P(T \le t)\ ,\ 0 < t < \infty \tag{2}$$

U.E Department of Statistics, Faculty of Science and Letters, Çukurova University, 01330 Adana, Turkey (corresponding author to provide phone: 90-338 60 84; fax: 90-338 60 70; e-mail: ugokal@ cu.edu.tr).

M.E. Department of Statistics, Faculty of Science and Letters, Çukurova University, 01330 Adana, Turkey (e-mail: merisoglu@cu.edu.tr).

H. E. Department of Statistics, Faculty of Science and Letters, Çukurova University, 01330 Adana, Turkey (e-mail: herol@cu.edu.tr)

The hazard function (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:

$$h(t) = \lim_{\Delta t \to 0}\left[\frac{P(t \le T < (t+\Delta t)/T \ge t}{\Delta t}\right] = \frac{f(t)}{S(t)} \qquad (3)$$

The cumulative hazard function is defined as

$$H(t) = -\log(S(t)) = \int_0^t h(u)du \qquad (4)$$

Given any one of them, the other two can be derived [9].

$$S(t) = 1 - F(t) = \exp(-H(t)) \qquad (5)$$

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 and Weibull probability distribution functions are commonly used in survival analysis. The $f(t)$ probability density function (pdf), $S(t)$ survival function and mean lifetime denoted by $E(t)$ form of these distribution models can be summarized below.

Exponential Distribution:

$$f_{\exp}(t) = \frac{1}{\lambda} e^{-\frac{t}{\lambda}} \quad t > 0,\ \lambda > 0 \qquad (6)$$

$$S_{\exp}(t) = 1 - e^{-\frac{t}{\lambda}} \qquad (7)$$

$$E_{\exp}(t) = \lambda \qquad (8)$$

Gamma Distribution:

$$f_{gm}(t) = t^{\alpha_1 - 1}\frac{e^{-t/\beta_1}}{\beta_1^{\alpha_1}\Gamma(\alpha_1)},\ t \text{ and } \alpha_1, \beta_1 > 0 \qquad (9)$$

$$S_{gm}(t) = 1 - \frac{\Gamma_x(\alpha_1)}{\Gamma(\alpha_1)} \qquad (10)$$

$$E_{gm}(t) = \alpha_1 \beta_1 \qquad (11)$$

where $\Gamma_x(\alpha_1)$ is called an incomplete Gamma function and calculated by $\Gamma_x(\alpha_1) = \int_0^x t^{\alpha_1 - 1} e^{-t} dt$

Weibull Distribution:

$$f_{wbl}(t) = \frac{\beta_2}{\alpha_2}\left(\frac{t}{\alpha_2}\right)^{\beta_2 - 1} e^{-\left(\frac{t}{\alpha_2}\right)^{\beta_2}},\ t \text{ and } \alpha_2, \beta_2 > 0 \qquad (12)$$

$$S_{wbl}(t) = e^{-\left(\frac{t}{\beta_2}\right)^{\alpha_2}} \qquad (13)$$

$$E_{wbl}(t) = \beta_2 \Gamma\left(1 + \frac{1}{\alpha_2}\right) \qquad (14)$$

III. MIXTURE MODEL OF TWO DIFFERENT DISTRIBUTIONS

3.1. Model

Mixture model of two different distributions assumed that the population consists of *g*=2 distinct subgroups or subclasses. Mixture model of two different distributions can written as

$$f_{\mathrm{X,Y}}(t;\psi) = \pi\ f_X(t;\theta_X) + (1-\pi) f_Y(t;\theta_Y) \qquad (15)$$

where the vector $\psi = (\pi, \theta)$ contains all unknown parameters π and $\theta = (\theta_X, \theta_Y)$ in the mixture model. The function $f_X(t;\theta_X)$ is called mixture component density function for some parameter θ_X, $f_Y(t;\theta_Y)$ is defined similarly.

In this study, to model the heterogeneous survival times, we used the mixture of two different distributions of Exponential-Gamma, Exponential-Weibull and Gamma-Weibull which are defined as

$$f_{\exp-gm}(t) = \pi\ f_{\exp}(t;\lambda) + (1-\pi) f_{gm}(t;\alpha_1,\beta_1) \qquad (16)$$

$$f_{\exp-wbl}(t) = \pi\ f_{\exp}(t;\lambda) + (1-\pi) f_{wbl}(t;\alpha_2,\beta_2) \qquad (17)$$

$$f_{gm-wbl}(t) = \pi\ f_{gm}(t;\alpha_1,\beta_1) + (1-\pi) f_{wbl}(t;\alpha_2,\beta_2) \qquad (18)$$

where π is the mixture weight of the distributions and $\pi \in (0,1)$. $f_{\exp}(t)$, $f_{gm}(t)$ and $f_{wbl}(t)$ are defined as in Eqs. (6), (9) and (12) respectively. The maximum likelihood estimators of parameters of these mixture distributions are estimated using Expectation-Maximization (EM) algorithm

3.2. Parameter Estimation in Mixture Models of Two Different Distributions using with EM Algorithm

In finite mixture models, the EM (Expectation-Maximization) algorithm has been used as an effective method to find maximum likelihood parameters estimation [12]. In EM framework, the observed data $t_1,\ldots,t_n$ is considered as an incomplete data and latent class variables

z_1, z_2 to be missing where $z_{1i} = z_1(x_i) = 1$ if observation t_i belongs to 1 th class and 0 otherwise and $i = 1,\ldots,n$. The EM algorithm is applied to the mixture distributions by treating z as missing data. EM algorithm can be preceded with two steps, E- and M- steps.

In E step, to estimate the hidden variable vector $z_i = [z_{1i} \quad z_{2i}]$, Conditional expectation funciton $E(z_{1i} \mid t_i)$ and $E(z_{2i} \mid t_i)$ are used.

$$\hat{z}_{1i} = E(z_{1i} \mid t_i) = \frac{\pi f_X(t_i;\theta_X)}{\pi f_X(t_i;\theta_X) + (1-\pi) f_Y(t_i;\theta_Y)} \tag{19}$$

$$\hat{z}_{2i} = E(z_{2i} \mid t_i) = \frac{(1-\pi)\pi f_Y(t_i;\theta_Y)}{\pi f_X(t_i;\theta_X) + (1-\pi) f_Y(t_i;\theta_Y)} \tag{20}$$

In M step, $E(z_{1i} \mid t_i)$ and $E(z_{2i} \mid t_i)$ function which are calculated in E step is maximized under the conditional of on mixture weights as $\pi_k \in (0,1)$. To estimate the mixture weights and parameters vectors which are denoted by π and $\theta = (\theta_X, \theta_Y)$ respectively, Lagrange method can be used. The estimated mixture weight is defined by

$$\hat{\pi} = \frac{1}{n}\sum_{i=1}^{n} \hat{z}_{1i} \tag{21}$$

The maximum likelihood estimator of λ parameter can be obtained with Eq. (22) for Exponential-Gamma and Exponential-Weibull distributions. The maximum likelihood estimator of λ parameter is given by

$$\hat{\lambda} = \left(\sum_{i=1}^{n} \hat{z}_{1i}\right)^{-1} \sum_{i=1}^{n} \hat{z}_{1i} t_i \tag{22}$$

The maximum likelihood estimators of α_1 and β_1 parameters can be obtained with Eqs. (23) and (24) respectively for Exponential-Gamma mixture distribution. The maximum likelihood estimators of α_1 and β_1 parameters are given by

$$\hat{\alpha}_1^{r+1} = \hat{\alpha}_1^{r} - \frac{\ln(\hat{\alpha}_1^{r}) - \psi(\hat{\alpha}_1^{r}) - \ln\left(\dfrac{\sum_{i=1}^{n} \hat{z}_{2i} t_i}{\sum_{i=1}^{n} \hat{z}_{2i}}\right) + \dfrac{\sum_{i=1}^{n} \hat{z}_{2i} \ln(t_i)}{\sum_{i=1}^{n} \hat{z}_{2i}}}{\dfrac{1}{\hat{\alpha}_1^{r}} - \psi'\left(\hat{\alpha}_1^{r}\right)} \tag{23}$$

$$\hat{\beta}_1 = \left(\hat{\alpha}_1 \sum_{i=1}^{n} \hat{z}_{2i}\right)^{-1} \sum_{i=1}^{n} \hat{z}_{2i} t_i \tag{24}$$

where r is number of Newton-Raphson iteration within EM algorithm and $\psi(.)$ with $\psi'(.)$ are a digamma and trigamma functions respectively. The maximum likelihood estimators of α_1 and β_1 parameters of Gamma-Weibull mixture distribution are estimated using $\hat{z}_{1i}$ instead of $\hat{z}_{2i}$ in the Eqs. (23) and (24).

The maximum likelihood estimators of β_2 and α_2 parameters can be obtained with Eqs. (25) and (26) respectively for Exponential-Weibull and Gamma-Weibull mixture distribution. The maximum likelihood estimators of β_2 and α_2 parameters are given by

$$\hat{\beta}_2^{r+1} = \hat{\beta}_2^{r} + \frac{A_r + \left(1/\hat{\beta}_2^{r}\right) - \left(C_r / B_r\right)}{\left(1/(\hat{\beta}_2^{r})^2\right) + \left(B_r D_r - C_r^2\right)/B_r^2} \tag{25}$$

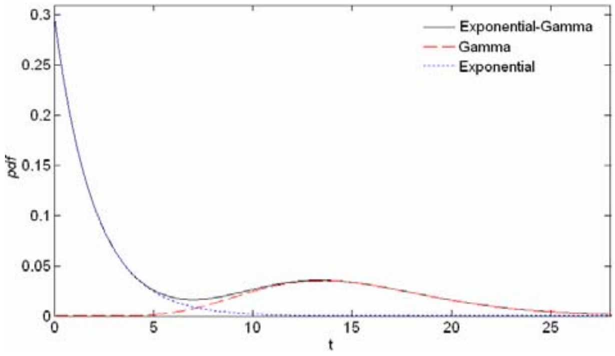
$$\hat{\alpha}_2 = \left(\left(\sum_{i=1}^{n} \hat{z}_{2i}\right)^{-1} \sum_{i=1}^{n} \hat{z}_{2i} t_i^{\hat{\beta}_2}\right)^{1/\hat{\beta}_2} \tag{26}$$

where $A_r = \left(\sum_{i=1}^{n} \hat{z}_{2i}\right)^{-1} \sum_{i=1}^{n} \hat{z}_{2i} \ln t_i$, $B_r = \sum_{i=1}^{n} \hat{z}_{2i} t_i^{\hat{\beta}_2^r}$, $C_r = \sum_{i=1}^{n} \hat{z}_{2i} t_i^{\hat{\beta}_2^r} \ln t_i$, $D_r = \sum_{i=1}^{n} \hat{z}_{2i} t_i^{\hat{\beta}_2^r} (\ln t_i)^2$ and r is number of Newton-Raphson iteration within EM algorithm.

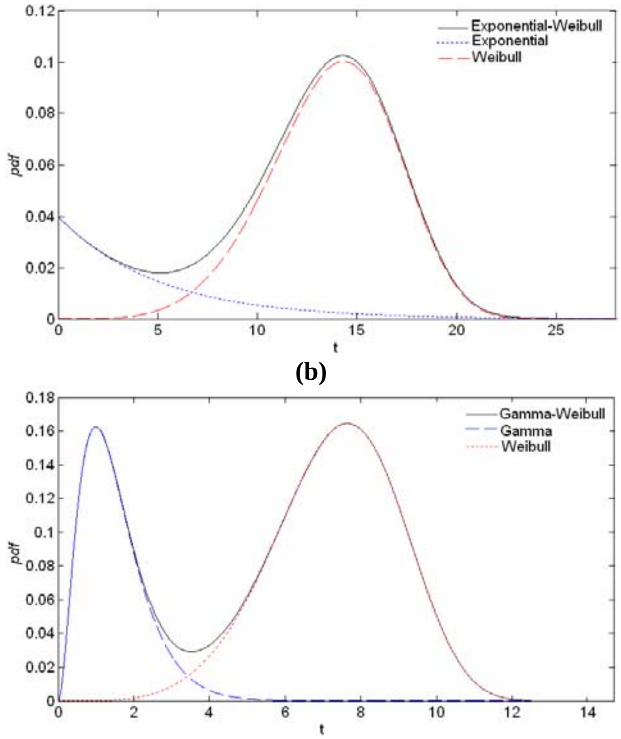
IV. APPLICATIONS

4.1. Simulation

Simulations were performed to investigate the convergence of the proposed EM scheme. We generated 1000 samples of size 100, each randomly sampled from the mixture of two different distributions. The graphs of the mixture of two different distributions for simulation parameters are shown in Figure 1.



(a)



(b)

(c)

Fig 1. The graphs of the mixture models of two different distributions for simulation parameters

No restriction was imposed on the maximum number of iterations and convergence was assumed when the absolute differences between successive estimates were less than 10^{-4}. The results from the simulated data sets are reported in Table 1, which gives the averages of the maximum likelihood estimators $av(\hat{\pi},\hat{\theta})$ and standard errors $se(\hat{\pi},\hat{\theta})$.

TABLE I
SIMULATION RESULTS

	Exponential-Gamma :				
	$\pi_1 = 0.6$	$\lambda = 2$	$\alpha_1 = 10$	$\beta_1 = 1.5$	
	$\hat{\pi}_1$	$\hat{\lambda}$	$\hat{\alpha}_1$	$\hat{\beta}_1$	
$av(\hat{\pi},\hat{\theta})$	0.6041	2.0502	11.7699	1.4165	
$se(\hat{\pi},\hat{\theta})$	0.0018	0.0133	0.1443	0.0143	
	Exponential-Weibull				
	$\pi_1 = 0.8$	$\lambda = 5$	$\alpha_2 = 15$	$\beta_2 = 5$	
	$\hat{\pi}_1$	$\hat{\lambda}$	$\hat{\alpha}_2$	$\hat{\beta}_2$	
$av(\hat{\pi},\hat{\theta})$	0.8110	4.4460	14.9555	5.0691	
$se(\hat{\pi},\hat{\theta})$	0.0019	0.0538	0.0128	0.0206	
	Gamma–Weibull :				
	$\pi_1 = 0.3$	$\alpha_1 = 3$	$\beta_1 = 0.5$	$\alpha_2 = 8$	$\beta_2 = 5$
	$\hat{\pi}_1$	$\hat{\alpha}_1$	$\hat{\beta}_1$	$\hat{\alpha}_2$	$\hat{\beta}_2$
$av(\hat{\pi},\hat{\theta})$	0.3075	3.4077	0.5446	8.0142	5.2434
$se(\hat{\pi},\hat{\theta})$	0.0018	0.0405	0.0108	0.0073	0.0226

Convergence was achieved in all cases, even when the starting values were poor and this emphasizes the numerical stability of the EM algorithm. The values of $av(\hat{\pi},\hat{\theta})$ and $se(\hat{\pi},\hat{\theta})$ suggest that the EM estimates performed consistently. According to the simulation results, the EM approach works well with different mixture proportions.

4.2. Failure times for oral irrigators

The studied real dataset in this paper is failure times for oral irrigators dataset in [4] and studied by Jiang and Murthy [6]. This dataset is dealing with failure times for oral irrigators. The estimated parameters, K-S test statistics and mean square error (MSE) values for the (pure) pdf of Exponential distribution, the (pure) pdf of Gamma distribution model, the (pure) pdf of Weibull distribution model and the pdf of Exponential-Gamma, Exponential-Weibull and Gamma-Weibull for failure times for oral irrigators data are given in Table 2. The mean square error is one of many ways to quantify the difference between an estimator and the true value of the quantity being estimated. MSE is obtained with

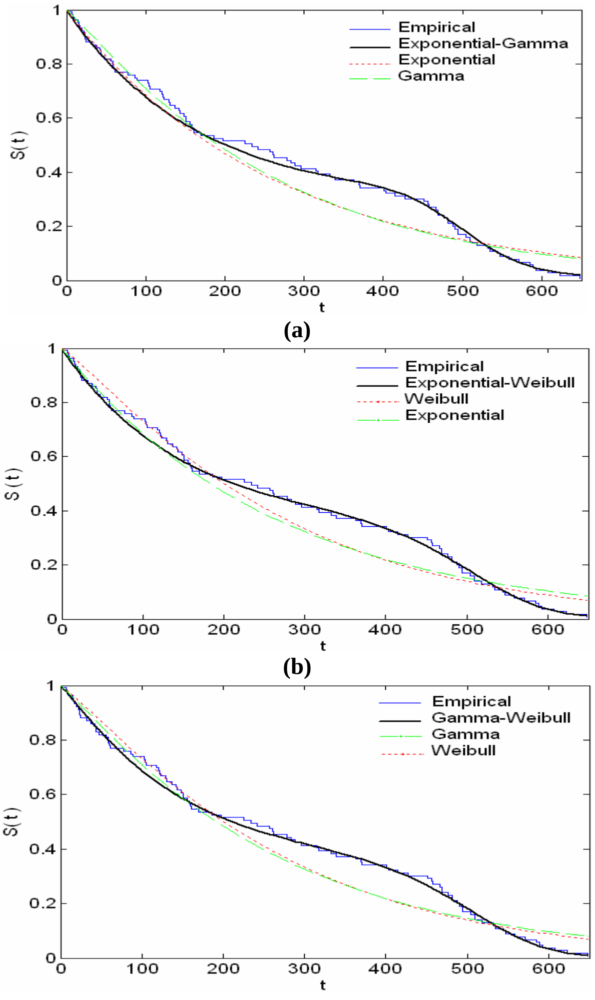
$$\text{MSE} = \frac{\sum_{i=1}^{n}\{Emp(t_i) - F(t_i)\}^2}{n-m} \tag{27}$$

where $Emp(t_i) = (i-0.3)/(n+0.4)$ for $i = 1,...,n$ is empirical distribution, m is the number of free parameters in the distribution and $F(t_i)$ is theoretical distribution function.

TABLE II
PARAMETER ESTIMATIONS, K-S and MSE VALUES

	Parameter Estimations	K-S	MSE
Exponential	$\hat{\lambda} = 264.5593$	0.1127	0.0034
Gamma	$\hat{\alpha}_1 = 1.1219,\ \hat{\beta}_1 = 235.8211$	0.1151	0.0032
Weibull	$\hat{\alpha}_2 = 276.3050,\ \hat{\beta}_2 = 1.1464$	0.1182	0.0033
Exponential & Gamma	$\hat{\pi}_1 = 0.7109$ $\hat{\lambda} = 165.7352$ $\hat{\alpha}_1 = 65.5052,\ \hat{\beta}_1 = 7.7484$	0.0591	0.00047
Exponential & Weibull	$\hat{\pi}_1 = 0.6561$ $\hat{\lambda} = 148.1692$ $\hat{\alpha}_2 = 520.9826,\ \hat{\beta}_2 = 6.8394$	0.0575	0.00041
Gamma & Weibull	$\hat{\pi}_1 = 0.6507$ $\hat{\alpha}_1 = 1.1412,\ \hat{\beta}_1 = 127.1518$ $\hat{\alpha}_2 = 521.5024,\ \hat{\beta}_2 = 6.8419$	0.0500	0.00034

A graphical comparison of the fitted (pure) pdf of Exponential, Gamma and Weibull distributions model and the fitted pdf of the mixture models of Exponential-Gamma, Exponential-Weibull and Gamma-Weibull for failure times for oral irrigators data is given in Figure 2.



(a)

(b)

(c)

Fig 2. The survival functions of fitting models for failure times for oral irrigators data

This can be seen graphically from Figure 2 that the mixture models of two different distributions such as Exponential-Gamma, Exponential-Weibull and Gamma-Weibull fits much better than the (pure) Exponential, Gamma and Weibull distributions to represent failure times for oral irrigators dataset.

V.CONCLUSIONS

In this paper we proposed the mixture models of two different distributions such as Exponential-Gamma, Exponential-Weibull and Gamma-Weibull to represent the heterogeneous survival data sets. The maximum likelihood estimations of parameters of the mixture models obtained with EM algorithm. Simulations were performed to investigate the convergence of the proposed EM algortihm. According to the simulation results, the EM algorithm was succesfull in estimation of parameters of the mixture models. The mixture models of two different distributions such as Exponential-Gamma, Exponential-Weibull and Gamma-Weibull successfully applied for modeling failure times for oral irrigators dataset. Gamma –Weibull mixture distribution is best distribution for modeling failure times for oral irrigators within the mixture models of two different distributions.

ACKNOWLEDGEMENTS

We would like to thank the editor and referees whose comments significantly improved this manuscript. This work is supported by Çukurova University Scientific Research Project Unit (No. FEF 2009D7 LTP).

REFERENCES

[1] 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.

[2] Berkson, J., Gage, R.P, Survival cure for cancer patients following treatment. Journal of the American Statistical Association 47, 501-515, 1952.

[3] Chen W.C., Hill B.M., Greenhouse J.B. and Fayos J.V.,. Bayesian Analysis of Survival Curves for Cancer Patients Following Treatment. Bayesian Statistics 2, 299-328, 1985.

[4] Colvert, R.E. and Boardman, T.J., Estimation in the piece-wise constant hazard rate model. Communication in Statistics-Theory. Methods. 11:1013-1029, 1976.

[5] Everitt B.S. and Hand D.J.,. Finite Mixture Distributions, Chapman and Hall, London, 1981.

[6] Jiang, R. and Murthy, D.N.P., Two sectional models involving three Weibull distributions. Quality and Reliability Engineering İnternational 13:83-96, 1997.

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

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

[9] Lee E.T. and Wang J.W., Statistical Methods For Survival Data Analysis, Third Edition, John Wiley &Sons, New York, 2003.

[10] Machin D, Cheung Y.B. and Parmar M.K,. Survival Analysis: A Practical Approach, Second Edition, John Wiley & Sons, 2006.

[11] Marin J.M., Rodriguez-Bernal M.T. and Wiper M.P., Using Weibull Mixture Distributions to Model Heterogeneous Survival Data, Communication in Statistics-Simulation and Computation, 34, 673-684, 2005.

[12] Mclachlan G.J. and G.J. Peel D., Finite Mixture Model, Wiley, New York. 2001.

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