PARAMETRIC MIXTURE MODEL OF THREE COMPONENTS FOR MODELLING HETEROGENEOUS SURVIVAL DATA

YUSUF ABBAKAR MOHAMMED

DOCTOR OF PHILOSOPHY
UNIVERSITI UTARA MALAYSIA
2015

Permission to Use

In presenting this thesis in fulfilment of the requirements for a postgraduate degree from Universiti Utara Malaysia, I agree that the Universiti Library may make it freely available for inspection. I further agree that permission for the copying of this thesis in any manner, in whole or in part, for scholarly purpose may be granted by my supervisor(s) or, in their absence, by the Dean of Awang Had Salleh Graduate School of Arts and Sciences. It is understood that any copying or publication or use of this thesis or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to Universiti Utara Malaysia for any scholarly use which may be made of any material from my thesis.

Requests for permission to copy or to make other use of materials in this thesis, in whole or in part, should be addressed to:

Dean of Awang Had Salleh Graduate School of Arts and Sciences
UUM College of Arts and Sciences
Universiti Utara Malaysia
06010 UUM Sintok

Abstrak

Kajian yang lepas menunjukkan model kemandirian campuran dua komponen mencatatkan prestasi yang lebih baik berbanding model kemandirian berparameter klasik tulen. Namun terdapat juga keperluan yang penting bagi model kemandirian campuran tiga komponen kerana tingkah laku data kemandirian heterogen yang lazimnya merangkumi lebih dari dua taburan. Oleh itu dalam kajian ini dua model bagi tiga komponen telah dibina. Model 1 adalah model kemandirian campuran berparameter tiga komponen bertaburan Gamma dan Model 2 adalah model kemandirian campuran berparameter tiga komponen bertaburan Eksponen, Gamma dan Weibull. Kedua-dua model telah dianggar menggunakan Pemaksimuman jangkaan (EM) dan pengesahan prestasi model melalui kajian simulasi dan empirikal. Simulasi telah diulang 300 kali dengan mengambil kira tiga saiz sampel berbeza: 100, 200, 500; tiga peratus penapisan yang berbeza: 10%, 20%, 40%; dan dua set kebarangkalian bercampur secara: menaik (10%, 40%, 50%) dan secara menurun (50%, 30%, 20%). Beberapa set data sebenar telah digunakan dalam kajian empirikal dan perbandingan model-model telah dilaksanakan. Model 1 telah dibandingkan dengan model kemandirian berparameter klasik tulen, model kemandirian berparameter campuran dua dan empat komponen bertaburan Gamma. Model 2 telah dibandingkan dengan model kemandirian berparameter klasik tulen dan model kemandirian berparameter campuran tiga komponen bertaburan sama. Persembahan grafik, *log likelihood* (LL), Kriteria Maklumat Akaike (AIC), Min Ralat Kuasa Dua (MSE) dan Punca Min Ralat Kuasa Dua (RMSE) telah digunakan bagi menilai prestasi. Dapatan simulasi menunjukkan bahawa kedua-dua model mencatatkan prestasi yang baik pada saiz sampel yang besar, peratus tertapis yang kecil dan pada kebarangkalian bercampur secara menaik. Kedua-dua model menghasilkan ralat yang kecil berbanding dengan model kemandirian jenis lain dalam kajian empirikal. Ini menunjukkan bahawa kedua-dua model yang dibina adalah lebih tepat dan merupakan pilihan yang lebih baik untuk menganalisis data kemandirian heterogen.

Kata kunci: data survival, heterogen, tiga komponen, eksponen, Gamma, Weibull, Pengmaksimuman Jangkaan

Abstract

Previous studies showed that two components of survival mixture model performed better than pure classical parametric survival model. However there are crucial needs for three components of survival mixture model due to the behaviour of heterogeneous survival data which commonly comprises of more than two distributions. Therefore in this study two models of three components of survival mixture model were developed. Model 1 is three components of parametric survival mixture model of Gamma distributions and Model 2 is three components of parametric survival mixture model of Exponential, Gamma and Weibull distributions. Both models were estimated using the Expectation Maximization (EM) and validated via simulation and empirical studies. The simulation was repeated 300 times by incorporating three different sample sizes: 100, 200, 500; three different censoring percentages: 10%, 20%, 40%; and two different sets of mixing probabilities: ascending (10%, 40%, 50%) and descending (50%, 30%, 20%). Several sets of real data were used in the empirical study and models comparisons were implemented. Model 1 was compared with pure classical parametric survival model, two and four components parametric survival mixture models of Gamma distribution, respectively. Model 2 was compared with pure classical parametric survival models and three components parametric survival mixture models of the same distribution. Graphical presentations, log likelihood (LL), Akaike Information Criterion (AIC), Mean Square Error (MSE) and Root Mean Square Error (RMSE) were used to evaluate the performance. Simulation findings revealed that both models performed well at large sample size, small percentage of censoring and ascending mixing probabilities. Both models also produced smaller errors compared to other type of survival models in the empirical study. These indicate that both of the developed models are more accurate and provide better option to analyse heterogeneous survival data.

Keywords: survival data, heterogeneous, three components, Exponential, Gamma, Weibull, Expectation Maximization.

Acknowledgement

All praise and gratitude are due to Allah (SWT) the Lord of heavens for providing me the strength and sparing my life to witness the completion of this work. I would like to express my endless thanks to my wonderful supervisors Associate Professor Dr Bidin Yatim and Dr Suzilah Ismail, for their inspiration, motivation, guidance, and commitment. Without their supervision this thesis would not have been completed to this level. Once again, I express my deep gratitude and appreciation. Special thanks go to UUM CAS staff, especially the staff of the School of Quantitative Sciences (SQS) for their kind assistance whenever needed. My special thanks also go to my postgraduate colleagues both the local and international students from different countries which include Nigeria, Yemen, Iraq, Jordan and Sudan to mention a few, for the good spirit of cooperation, motivation and encouragement among them, and wish them all the best. My appreciations go to the Nigerian community for their cooperation and encouragement. Great appreciation and gratitude go to my mother Maryam Mohammed and my father Abbakar Mohammed for their continued prayers and inspiration. Thanks to my brothers and sisters, for their continued encouragement. Thanks and appreciations go to my loving wife, Dr Zainab Abdallah for her prayers, encouragement and patience. My thanks to my wonderful children Mohammed, Abubakr, Rayyan and Abdallah for their inspiration and patience. Finally, I would like to thank and express my appreciation to all those who in one way or the other contributed to the completion of this work and I could not mention their names, but still they are in my heart. Thank you very much.

Table of Contents

List of Tables

List of Figures

List of Appendices

Glossary of Terms

E1_E2_E3 PARAMETRIC SURVIVAL MIXTURE MODEL OF EXPONENTIAL_EXPONENTIAL_EXPONENTIAL DISTRIBUTIONS.

E_G_W PARAMETRIC SURVIVAL MIXTURE MODEL OF EXPONENTIAL_GAMMA_WEIBULL DISTRIBUTIONS.

G1_G2_G3 PARAMETRIC SURVIVAL MIXTURE MODEL OF GAMMA_GAMMA_GAMMA DISTRIBUTIONS.

W1_W2_W3 PARAMETRIC SURVIVAL MIXTURE MODEL OF WEIBULL_WEIBULL_ WEIBULL DISTRIBUTIONS.

List of Abbreviations

AIC	AKAIKE INFORMATION CRITERION
EEG	EXTENDED EXPONENTIAL-GEOMETRIC
EM	EXPECTATION MAXIMIZATION
KM	KAPLAN-MEIER
LL	LOG LIKELIHOOD
MCM	MIXTURE CURE MODELS
ML	MAXIMUM LIKELIHOOD
MSE	MEAN SQUARE ERROR
MTIWD	MIXTURE OF TWO INVERSE WEIBULL DISTRIBUTION
RMSE	ROOT MEAN SQUARE ERROR

CHAPTER ONE
INTRODUCTION

1.1 Background of the Study

Survival data analysis is the analysis of time to occurrence of a particular event of interest. The data are usually related to clinical studies of human, or laboratory studies of animal, or studies to test the life time of some devices. Major applications are in the areas of human clinical studies and industrial life testing (Kalbfleisch & Prentice, 2002).

The event of interest in clinical studies could be death, remission, or some other clinical events. The event of interest could be time taken to learning a new skill, exit from unemployment, divorce of a couple or failure of a device, to mention a few. The variable of interest, the time to occurrence of particular event *T*, which is a positive random variable, should clearly be defined in the study at hand. The start and end with the length of the time period in-between corresponding to *T*, should also be clearly defined prior to the commencement of the study (Lee & Wang, 2003).

Generally, in survival analysis, some individuals or objects do not experience the event of interest for one reason or the other, either they are lost to follow up during the period of the study or they do not experience the event until the end of the study. In such situation, the information about this particular individual will not be exactly known, and such individuals are referred to as censored observations or censored times.

When there is no censored observation the set of survival data is said to be complete. The occurrence of censoring is the reason for the uniqueness of the survival time data, in such case, the classical statistical methods may not be appropriate for analysing such data. Therefore, different statistical procedures have been developed to handle such complexity in the data.

The purposes of applying the survival methods includes predicting the probability of response, comparing the survival distributions of experimental units and identifying the risk and/or prognostic factors related to the development of disease (Lee & Wang, 2003).

Like other branches of statistics the survival statistical methods for data analysis include parametric and non-parametric methods. The parametric methods are more suitable when the data under study follow some specified probability distribution. While on the other hand, the non-parametric methods do not require distributional assumptions; they are more flexible, and are preferred when no particular parametric distribution is appropriate for the data.

1.2 Problem Statement

Parametric probability distributions are commonly employed in statistical analysis; they are very useful if the selected parametric probability distribution fits the data well. In survival analysis with its uniqueness of the presence of censored observations, the most frequently used parametric distributions are the Exponential, Gamma, Weibull, Lognormal and Gompertz. If a particular distribution is found to fit the data well, then analysis, estimation and statistical inference can be based on the selected distribution (Ibrahim Chen & Sinha, 2001; Kalbfleisch & Prentice, 2002; Lawless, 2003; Lee & Wang, 2003).

Parametric survival mixture models provide the flexibility like the non-parametric models, while maintaining the features of parametric models. Many research works proposed parametric survival mixture model of the same probability distribution in survival analysis. Cheng and Fu (1982) proposed a parametric survival mixture model of Weibull distribution where they employed the weighted least squares method to estimate the parameters of the mixture model. Jiang and Kececioglu (1992a) estimated the parameters of a survival mixture model of Weibull distribution using graphical approach. They (Jiang & Kececioglu, 1992b) also developed a new procedure to estimate the parameters of a survival mixture model of Weibull distribution. Jaheen (2005) employed a parametric survival mixture model of Exponential-Exponential to model survival data. Zhang (2008) proposed a two-component parametric survival mixture model of Weibull distribution to model survival data and investigated the suitability of the model in survival data analysis. Also, Erisoglu, Erisoglu & Erol (2012) modelled heterogeneous survival data by a survival mixture model of Gamma-Gamma, a survival mixture of Lognormal-Lognormal and a survival mixture of Weibull-Weibull distributions, where they investigated the best fit model to real survival data. Other literature concerned with employing survival mixtures of same parametric distribution are; Ling, Huang & Liu, (2009); Farcomeni & Nardi (2010); Erisoglu, & Erol, (2010); and Zhang, Wang & Lu, (2011).

Most of the literatures above focused on parametric survival mixture models of same parametric distribution, very few considered survival mixture models of different parametric distributions. Among the few, Abu-Zinadah (2010) proposed a two components parametric survival mixture model of different distributions of Exponentiated pareto and Exponential distributions to model survival data. Recently, Erisoglu, Erisoglu & Erol (2011) proposed a two component parametric survival mixture model of two different distributions, namely: Exponential-Gamma, Exponential-Weibull and Gamma-Weibull, for the analysis of heterogeneous survival data. Their results showed the suitability of the parametric survival mixture models compared to the pure classical parametric survival models.

The parametric survival mixture models are more flexible compared to the pure classical parametric survival models. Therefore, the parametric survival mixture models are better than the pure classical parametric survival model when the survival data come from a heterogeneous population (Lawless, 2003). The problem of heterogeneity arises frequently in survival data analysis, where the pure classical parametric methods become no longer appropriate to model such data. For instance in the case of an open-heart surgery: Blackstone, Naftel, and Turner (1986) were able to classify the risk of death after the surgery by three different times overlapping phases. The phases are defined as an early phase in which the risk is relatively high; a middle phase where the risk becomes constant, and finally a phase in which the risk starts to increase with the advancement of the age of patients. Modelling each time period with a separate pure classical parametric survival model may not be appropriate. Therefore, a three component survival mixture should be an effective way of modelling such data (McLachlan & Peel, 2000; Ng, McLachlan, Yau, & Lee, 2004; Zhang, 2008).

In real life situation we encounter data with tri-modal nature due to some characteristics such as three age groups of some patients in a certain study or three modes of failure of a group of individuals or three geographical or regional or ethnic back grounds associated with some patients or three stages of certain disease of some patients. In some cases the graph of the real data is observed by employing the Exploratory Data Analysis (EDA) to find out what the data can tell about the appropriate model to be used (Tukey, 1977). Therefore, the three components parametric survival mixture model of Gamma distributions in some situations could be appropriate to model such types of data that may arise in real life.

A considerable number of researches work explored survival mixture models of same parametric distribution in terms of parameter estimation and inference (Rider, 1961; Jewell, 1982; Cheng & Fu, 1982; Jiang & Murthy, 1995; Sultan, Ismail, & Al-Moisheer, 2007; Razali & Salih, 2009). Most of them focused on two components survival mixture model. However, very little has been done in choosing mixtures of two different parametric distributions to address the issue of heterogeneous survival data (Erisoglu, et al., 2011). Very few studies that considered three components parametric survival mixture model of the same distribution. In the case of parametric survival mixture model of the same distribution, Marin, Rodríguez-Bernal and Wiper, (2005) proposed a three component parametric survival mixture model of the Weibull distribution. To the best of our knowledge the Gamma distribution in a parametric survival mixture of three components did not receive much attention, despite the importance of the Gamma distribution in survival data analysis (Kalbfeisch & Prentice, 2002; Lawless, 2003; Lee & Wang, 2003). Also, no study considered a three component parametric survival mixture model of the Exponential, Gamma and Weibull distributions.

The Gamma distribution is flexible and closely related to the Exponential and Weibull distribution (Ibrahim, et al. 2002; Lee & Wang, 2003). There is a crucial need to develop parametric survival mixture model of Gamma distributions of three components to model heterogeneous survival data. Also a three components parametric survival mixture model of the Exponential, Gamma and Weibull distributions needs to be developed.

This study focuses on modelling heterogeneous survival data by a parametric survival mixture model of three components of the Gamma distributions (Model 1) and a parametric survival mixture model of three components of different distributions of the Exponential, Gamma and Weibull distributions (Model 2). The proposed models are compared with pure classical parametric distribution models and the parametric survival mixture model of the same distribution.

1.3 Objectives of the Study

In this study, there are five objectives which are as follows

(i) to develop three components parametric survival mixture model of the Gamma distributions (Model 1) to model heterogeneous survival data.

The sub-objective of objective (i) is to employ Model 1 in evaluating the pure classical parametric survival model when the number of components is set to one.

(ii) To develop three components different parametric survival mixture model of the Exponential, Gamma and Weibull distributions (Model 2) to model heterogeneous survival data.

(iii) To evaluate the performance of the models via simulation study with three different samples sizes, three different censoring percentages and two sets of three different mixing probabilities.

(iv) To investigate the effect of the three different censoring percentages on the hazard function of the models via simulation study.

(v) To investigate the survival function in evaluating the fit of the models using empirical study.

1.4 Significance of the Study

The significance of this study is to show the importance and appropriateness of the three components parametric survival mixture models in modelling real life cases involving heterogeneous survival data. The study also shows that some real life situations are better modelled with three components parametric survival mixture model of the same distribution (the Gamma distribution). In some other cases a three components parametric survival mixture model of different distributions (the

Exponential, Gamma and Weibull) would be the appropriate choice to model heterogeneous survival data.

The study highlights the suitability, appropriateness and advantages of parametric survival mixture model over pure classical parametric survival models, when the data is heterogeneous.

1.5 Outline and Summary of the Thesis

Chapter Two is devoted to the literature review where basic concepts of the survival data analysis related to the parametric survival and the non-parametric methods were highlighted. Basic ideas of the parametric survival mixture models were elaborated. The recent expansions and development of the application of parametric survival mixture models were discussed.

Chapter Three outlined the methodology adopted to realize the objectives of the study. The introduction section highlighted the methodology frameworks of Model 1 and Model 2 respectively. The following two sections describe the steps and the procedures employed for both Model 1 and Model 2 using simulated and real data respectively.

Chapter Four is devoted to discuss Model 1. It includes an introduction section followed by a section in which the theoretical development of Model 1 and its transformation into computer coding were explained. In the next two sections the comparison studies of Model 1 were explained in details for both simulated and real data respectively. Finally the last section devoted to the findings summary of Model 1.

In Chapter Five, Model 2 has been discussed and it consists of four sections. The first section was an introduction section followed by a section in which the theoretical development of Model 2 together with the computer coding used in the analysis was explained. In the next two sections the comparison studies of Model 2 were explained in details for both simulated and real data respectively. Finally the last section devoted to the findings summary of Model 2.

Chapter Six is the conclusion chapter that outlines the summary and findings of the thesis, the problems and the limitations that were encountered during the research work and finally suggestion for future research works.

CHAPTER TWO
REVIEW OF THE LITERATURE

2.1 Introduction

Survival data analysis is concerned with implementing certain statistical methods to model and analyse survival data. The primary interest in such data is the endpoint time when an event of interest occurs. Generally, the events of interest are referred to as failures. For example, the time to death of a patient, time to learning a new skill, time to exit from unemployment, time to promotion for employees and time to breakdown of some devices.

From the examples above it is possible that some objects or individuals might not experience the event of interest, either by design or because of random censoring. This happens when some devices do not fail at the end of the experiment; some patients survive to the end of a clinical study or fail to follow up. The presence of censoring in survival data made it necessary to develop methods that can accommodate censored observations. The name survival analysis is given to a collection of statistical methods which are employed to handle survival data with censored observations (Tableman & Kim, 2004).

The survival time is used in connection with clinical studies. However, survival time has different names such as time to event, life time, duration time or failure time, depending on the field of application. In addition to medical studies, these methods have wide application in different fields such as, public health, epidemiology, social sciences, economics and engineering. The survival data analysis witnessed rapid developments and expansions with respect to theory, methodology and field of application (Lawless, 2003).

As mentioned earlier in Chapter One, the response of primary interest, T, is a non-negative continuous random variable representing a survival time of an individual or object from a well-defined population. All the functions characterizing T are defined over the interval [0, ∞). There are many ways to express the distribution of the random variable T. However, the three most useful functions in survival data analysis are; the probability density function (pdf) denoted by $f(t)$, survival function denoted by $S(t)$ and hazard function which is denoted by $h(t)$ (Ibrahim, et al., 2001; Kalbfleisch & Prentice, 2002; Tableman & Kim, 2004).

Let the random variable T represents the survival time of occurrence of the event of interest, and let the cumulative distribution function of T be defined by

$$F(t) = p(T \leq t), \;\; t > 0 \qquad (2.1)$$

which represents, the probability of occurrence of event of interest at time t or less than t. The probability density function (pdf) of the random variable T can be written as

$$f(t) = \frac{dF(t)}{dt} \qquad (2.2)$$

The probability density function can also be presented graphically, the graph of $f(t)$ is known as the density curve. The density function $f(t)$ is a nonnegative function and the area between the curve and the t axis is equal to 1.

The survival function is defined by

$$S(t) = 1 - F(t) \qquad (2.3)$$

which means, the probability that an individual survives beyond time t. Note that the survival function $S(t)$ is a monotonic decreasing continuous function with $S(0) = 1$ and $S(\infty) = \lim_{t\to\infty} S(t) = 0$.

The survival function can be represented graphically, and the graph of $S(t)$ is called the survival curve. This graph is used to estimate the 50th percentile (median) and other percentiles of the survival time and also for comparing survival distributions of two or more groups (Lee & Wang, 2003).

The hazard function can be defined as

$$h(t) = \lim_{\Delta t \to 0} \frac{p(t \le T < t + \Delta t | t)}{\Delta t}$$

$$h(t) = \frac{f(t)}{S(t)} \qquad (2.4)$$

which is the probability that an individual fails within a small interval $(t, t + \Delta t)$, given that the individual survived up to the beginning of the interval. The cumulative hazard function of the survival time T is defined by

$$H(t) = \int_0^t h(u) du \qquad (2.5)$$

Therefore, when $t = 0$ then $S(t) = 1$ and $H(t) = 0$, and when $t = \infty$, then $S(t) = 0$ and $H(t) = \infty$, that is, the cumulative hazard function can assume any value between zero to infinity.

The hazard function specifies the instantaneous rate of failure at time t given that the individual survived up to time t, and sometimes it is known as the instantaneous failure rate, force of mortality, conditional mortality rate, and age-specific failure rate. The hazard function may be also presented graphically (Lee & Wang, 2003).

Those three important functions are equivalent, if one is known the other two can be derived. The following equations illustrate their relationship.

$$f(t) = \frac{d}{dt}[1 - S(t)] = -S^{'}(t) \tag{2.6}$$

$$S(t) = \exp[-H(t)] \tag{2.7}$$

$$h(t) = \frac{f(t)}{S(t)} \tag{2.8}$$

The conventional statistical methods are not appropriate to handle the survival data due to the fact that some observations cannot be exactly observed or they are censored. There are many types of censoring arising in practice (Ibrahim, et al., 2001; Lee & Wang, 2003; Sun, 2006). Some of these types will be discussed in the following sections.

2.2 Censoring

Generally, parametric methods are used if the distribution of the set of data is known to be normal, and nonparametric methods are used if the distribution is unknown. This assumption does not hold in the case of survival data because some of the survival times are not exactly known. The survival distribution is often skewed and far from normal, this is because some objects or individuals have not experienced the event of interest at the end of the study. Such individuals are referred to as censored observations (Lee & Wang, 2003).

In survival data analysis there are three common modes of censoring namely, right censoring, interval censoring and left censoring. If it is known that the survival time *T* exceeds some particular time *U*, where *U* is the follow up time, then the survival time of that individual is said to be right censored, and *U* is called the censoring time. However, if the survival time *T* is not observed, but it is known to be less than or equals to some particular time *U*, then the individual is said to be left censored observation. Moreover, if it is known that the event time *T* is in between the two times *U* and *V*, where $U < V$, then the individual is said to be an interval censored (Ibrahim et al., 2001; Lawless, 2003; Lee & Wang, 2003). The most frequently encountered modes of censoring in survival data analysis are the right censoring. There are different types of right censoring; for example, type I censoring, independent random censoring and type II censoring, to mention some (Lawless, 2003).

2.2.1 Some Types of Right Censoring

Type I Censoring

In type I censoring scheme, the potential censoring time $C_i > 0$ for each individual is fixed in advance, such that T_i observed if $T_i \leq C_i$; otherwise it is only known that $T_i > C_i$. Type I censoring often arises when a study is conducted over a fixed period of time, after that time the study terminates, so all the individual are expected to fail on or before that time (Lawless, 2003). Sometimes individuals or objects under study are divided into subgroups, with a fixed right censoring time for each subgroup, and that is what is known as progressive type I censoring. Generally, type I right censoring arises in engineering and animal studies (Tableman & Kim, 2004; Lee & Wang, 2003).

Independent Random Censoring

A very simple random censoring process that is always realistic is the one in which each individual is considered to have a survival time *T* and a censoring time *C*. Here *T* and *C* are independent continuous random variables. Their survival functions are *S(t)* and *G(t)* respectively. All survival times and censoring times are assumed mutually independent and it is assumed that *G(t)* does not depend on any of the parameters of *S(t)* (Lawless, 2003).

Type II Censoring

The term type II censoring refers to the situation where only the *r* first survival times $t_{(i)} \leq ... \leq t_{(r)}$ in a random sample of *n* objects or individuals are observed. Here *r* is a specified integer between 1 and *n*. This censoring scheme arises when *n* individuals start the study at the same time, with the study terminating once r failures (or survival times) have been observed. Type I censoring is therefore much more common in planned experiments (Lawless, 2003).

The type of censoring normally depends on the nature and the field of study. However, the most frequently encountered modes of censoring is the random independent censoring, which is applied to both the parametric and non-parametric cases as will be discussed in the next few sections.

2.3 Parametric Methods in Survival Analysis

Parametric statistical methods are very powerful tools in survival data analysis provided that the selected parametric distribution fits the data well. Otherwise, nonparametric statistical techniques will be a better choice. There are several theoretical probability distributions that have been widely used in modelling survival data. Table 2.1 displays some important characteristics of these distributions, which include the probability density functions denoted by $f(t)$, cumulative distribution functions denoted by $F(t)$, survival functions denoted by $S(t)$, the mean survival times denoted by $E(t)$ and the set parameter(s) of each distribution. In the following subsections, some important parametric distributions employed in modelling survival data will be highlighted.

2.3.1 The Exponential Distribution

Exponential distribution is one of the most important and simplest distributions in survival data analysis. Researchers used the Exponential distribution to describe the life pattern of electronic systems in the late 1940s of the twentieth century. The distribution is characterized by its constant hazard function, where the instantaneous hazard rate is independent of time *t*, regardless of the age of individuals. This is referred to as the memory-less property of the Exponential distribution (Lee & Wang, 2003).

Table 2.1

Some Important Parametric Distributions in Survival Data Analysis

Distributions	**Parameters**	$\boldsymbol{f(t)}$	$\boldsymbol{F(t)}$	$\boldsymbol{S(t)}$	$\boldsymbol{E(t)}$
Exponential	λ	$\lambda e^{-\lambda t}$	$1 - e^{-\lambda t}$	$e^{-\lambda t}$	λ
Gamma	α and β	$t^{\alpha-1}\dfrac{e^{-\frac{t}{\beta}}}{\beta^{\alpha}\Gamma(\alpha)}$	$\dfrac{\Gamma_x\left(\alpha, \frac{t}{\beta}\right)}{\Gamma(\alpha)}$	$1 - \dfrac{\Gamma_x(\alpha)}{\Gamma(\alpha)}$	$\alpha\beta$
Weibull	α and β	$\dfrac{\alpha}{\beta}\left(\dfrac{t}{\beta}\right)^{\alpha-1} e^{\left(-\left(\frac{t}{\beta}\right)^{\alpha}\right)}$	$1 - e^{\left(-\left(\frac{t}{\beta}\right)^{\alpha}\right)}$	$e^{\left(-\left(\frac{t}{\beta}\right)^{\alpha}\right)}$	$\beta\Gamma\left(1 + \dfrac{1}{\alpha}\right)$
Lognormal	μ and σ	$\dfrac{e^{\left(-\frac{1}{2}\left(\frac{lnt-\mu}{\sigma}\right)^{2}\right)}}{t\sigma\sqrt{2\pi}}$	$\phi\left(\dfrac{lnt - \mu}{\sigma}\right)$	$1 - \phi\left(\dfrac{lnt - \mu}{\sigma}\right)$	$e^{\left(\mu + \frac{\sigma^{2}}{2}\right)}$
Gompertz	α and β	$\lambda e^{\left\{\alpha t - \frac{\lambda}{\alpha}\left[e^{(\alpha t)} - 1\right]\right\}}$	$1 - e^{\left\{-\frac{\lambda}{\alpha}\left[e^{(\alpha t)} - 1\right]\right\}}$	$e^{\left\{-\frac{\lambda}{\alpha}\left[e^{(\alpha t)} - 1\right]\right\}}$	$\frac{1}{\lambda} G\left(\frac{\lambda}{\alpha}\right)$*

*Note: $\mathrm{G(x)} = \frac{1}{\mathrm{y}}\int_{\mathrm{x}}^{\infty} \mathrm{e}^{-\mathrm{y}\mathrm{dy}}$

A number of examples of survival data described by an Exponential distribution were given by Davis (1952), including bank statement and ledger error, payroll check errors, automatic calculating machine failure and radar set components failure. Epstein and Sobel (1953) selected the Exponential distribution over the popular Normal distribution and estimated its parameter when some of the data singly censored. In a paper presented before the Royal Statistical Society, Zelen (1966) enumerated the applications of the Exponential distribution in survival data analysis in the areas of cancer research. Also, in the last decade, Jaheen (2005) used the Exponential distribution to model survival data by a survival mixture model of two components of the Exponential distribution.

The Exponential distribution is characterized by one parameter $\lambda > 0$ (scale parameter) which defines the constant hazard rate as mentioned earlier. When the value of λ is high it means high risk and shorter survival time, when the value of λ is low it means low risk and longer survival time. The graph of a constant hazard function of the Exponential distribution is displayed in Figure 2.1.

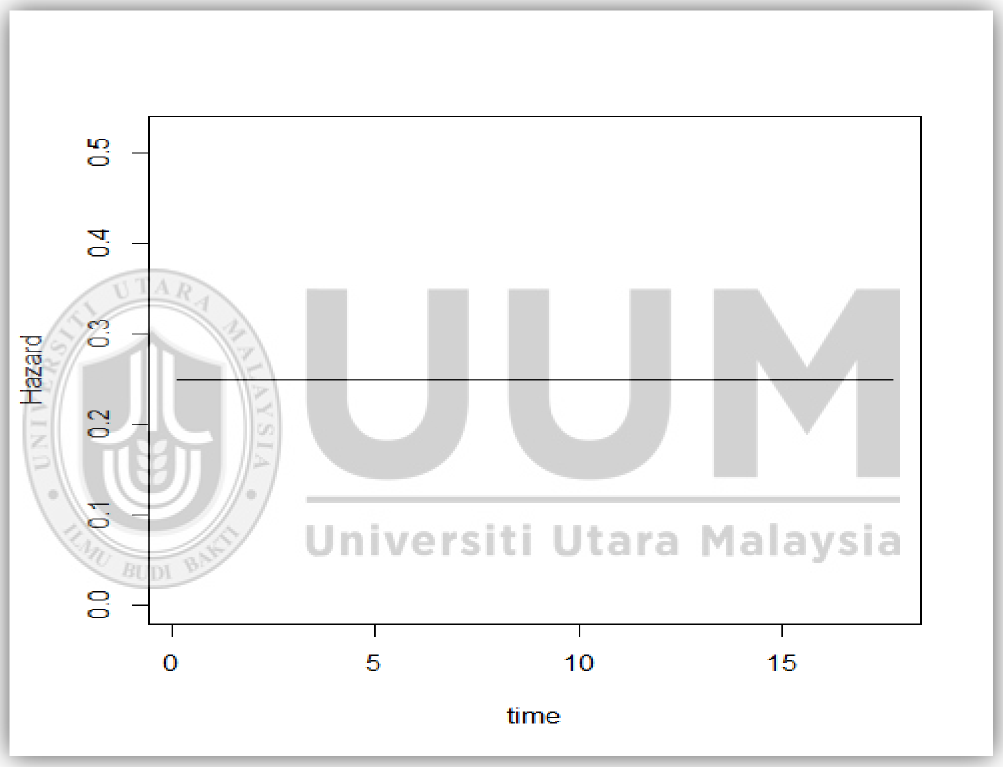


Figure 2.1 Hazard Function of Exponential Distribution λ = 0.25

Let the survival time T be a random variable that follows the Exponential distribution, then its probability density distribution (pdf) denoted by $f(t)$, the cumulative distribution function denoted by $F(t)$, the survival function denoted by $S(t)$ and mean survival time denoted by $E(t)$ are as defined in Table 2.1 where

$\lambda > 0$ and $t > 0$ are the scale parameter and survival time respectively. Figure 2.2 displays the survival function of the Exponential distribution.

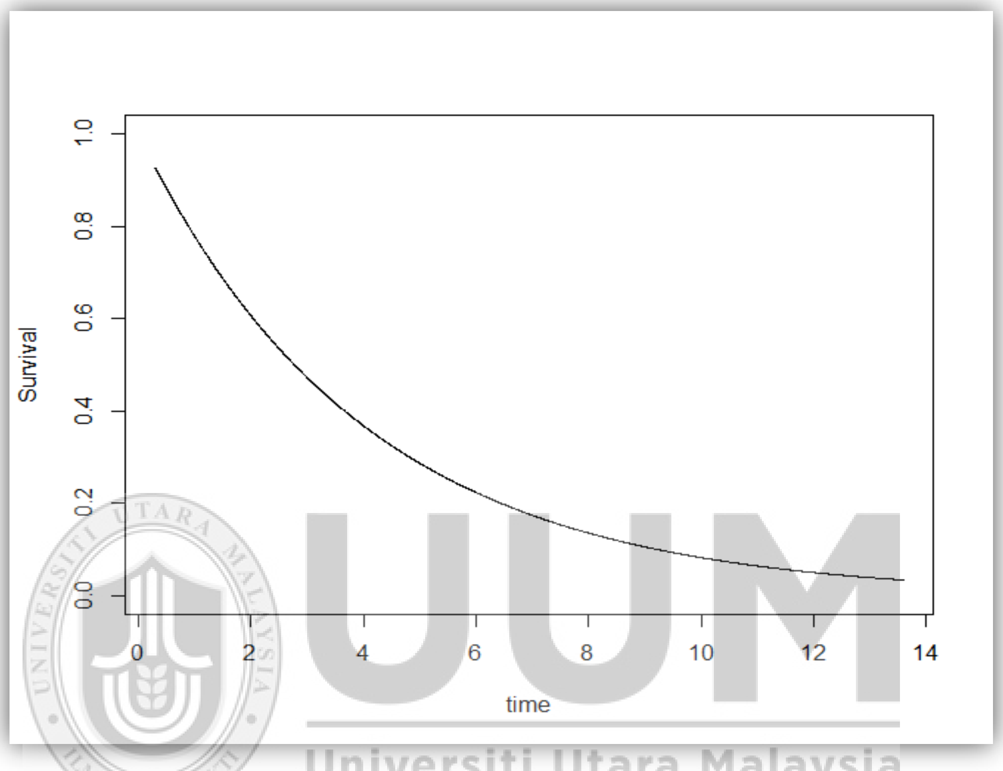


*Figure 2.2.*Survival Function of Exponential Distribution λ = 0.25

2.3.2 The Gamma Distribution

The Gamma distribution has been mentioned in the literature long time ago. Brown and Flood (1947) used Gamma distribution to describe glass tumblers survival time in circulation in a cafeteria. Also, Birnbaum and Saunder (1958) used it to model the life length of materials. The Gamma distribution has been frequently and efficiently used in modelling survival data. The Gamma distribution is characterized by two parameters, namely, $\alpha > 0$ the shape parameter and $\beta > 0$ the scale parameter. Figure 2.3 shows the Gamma density function with parameters α = 3 and β = 1.

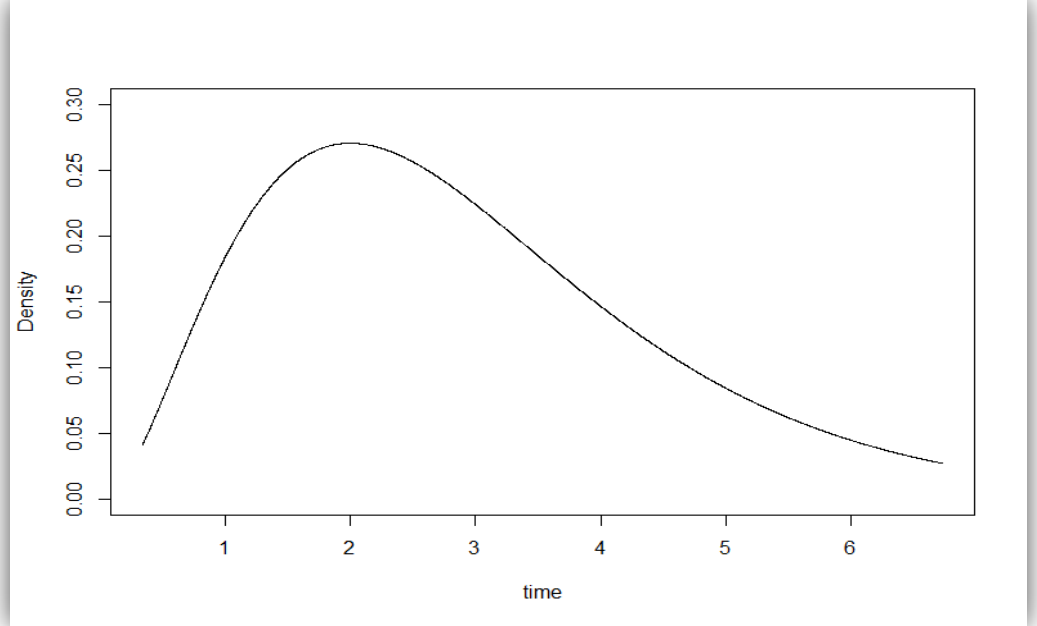


*Figure 2.3.*Probability Density Function of Gamma Distribution $\alpha = 3$ and $\beta = 1$

Let the survival time T be a random variable that follows the Gamma distribution, then its probability density function (pdf) denoted by $f(t)$, the cumulative distribution function denoted by $F(t)$, the survival function denoted by $S(t)$ and mean survival time denoted by $E(t)$ are as defined in Table (2.1), where $t > 0$. Figure 2.4 shows the shape of the survival function of the Gamma distribution with parameters $\alpha = 3$ and $\beta = 1$.

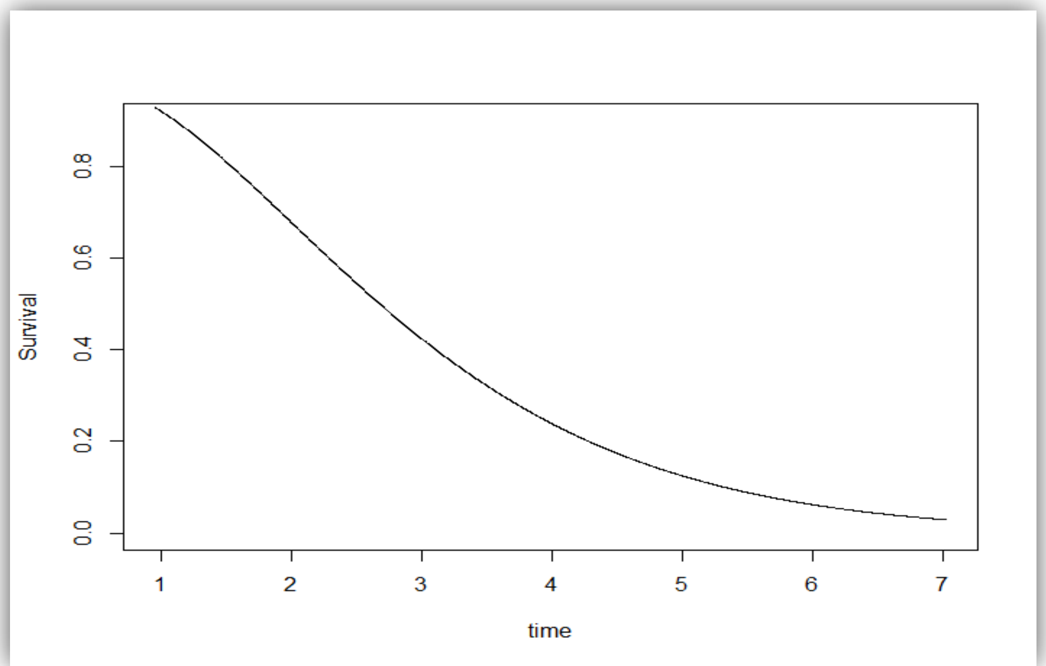


*Figure 2.4.*Survival Function of the Gamma Distribution α =3 and β =1

2.3.3 The Weibull Distribution

The Weibull distribution is characterized by two parameters, $\alpha > 0$ which determines the shape of the distribution, and is known as the shape parameter, and $\beta > 0$ which determines the scaling of the distribution, and is known as the scale parameter. The Weibull distribution is a generalization of Exponential distribution, but does not assume a constant hazard rate and therefore has broader applications (Lee & Wang, 2003).

The Weibull distribution is named after Waloddi Weibull who was the first to promote the usefulness of the distribution for modelling data sets of widely differing applications (Murthy, Xie & Jaing, 2004). The Weibull distribution was first proposed by Weibull (1939), and later the different applications of the model to survival data were discussed by Weibull (1951). Since then, the Weibull distribution has been used frequently in the literature, particularly in reliability and survival data analysis. Murthy et al. (2004) listed examples of some of the applications of the Weibull models in the literature, particularly in reliability data analysis and also generally in some other fields. When the shape parameter $\alpha = 1$, the hazard rate remains constant as time increases, this is the exponential case. When $\alpha < 1$, the hazard rate decreases with time and when $\alpha > 1$, the hazard rate increases with time. Thus, the Weibull distribution may be used to model survival data of population with increasing and decreasing, or constant risk. Figure 2.5 shows the nature of increasing hazard rate when the shape parameter is greater than one. Let the survival time T be a random variable that follows the Weibull distribution, as defined in Table 2.1, where $t > 0$. Figure 2.6 displays the survival function of the Weibull distribution.

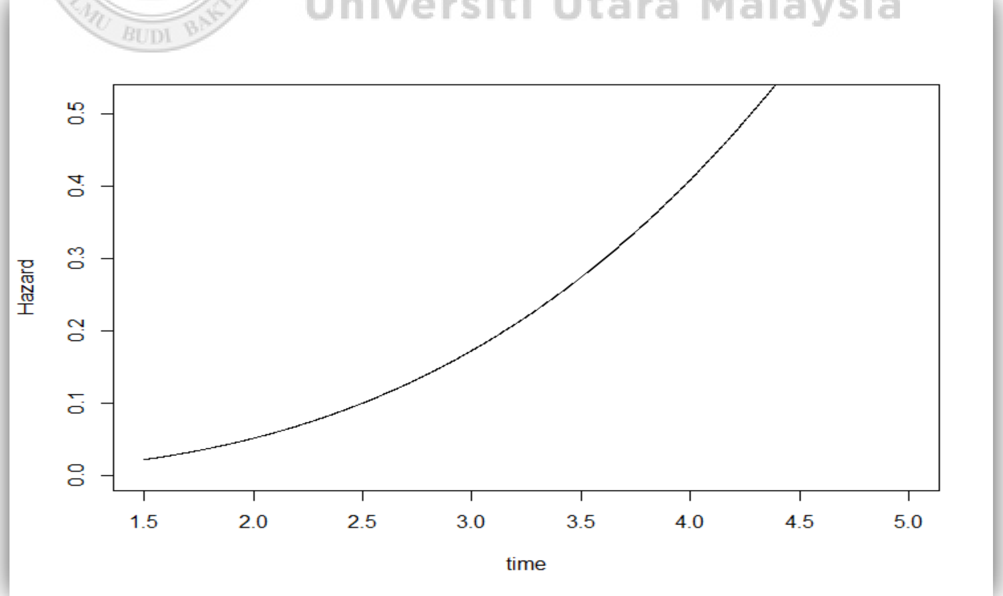


*Figure 2.5.*Hazard Function of the Weibull Distribution $\alpha = 4$

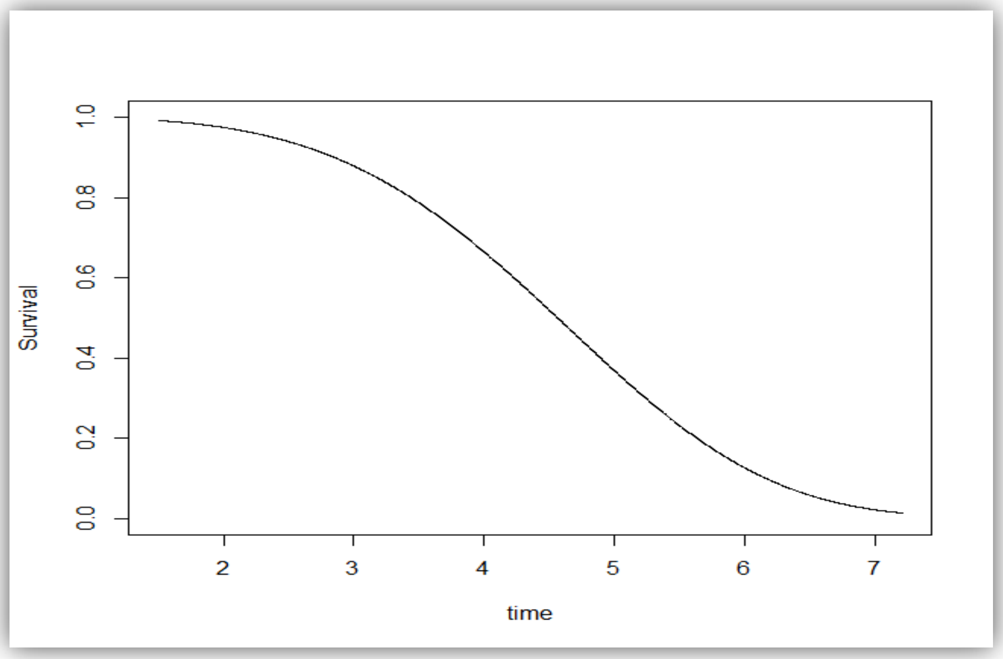


*Figure 2.6.*Survival Function of the Weibull Distribution $\alpha = 4$

2.3.4 The Lognormal Distribution

The Lognormal distribution is defined as the distribution of a random variable whose logarithm is normally distributed. The distribution is markedly positively skewed which makes it a good approximation of several diseases such as Hodgkin's disease and chronic leukaemia since the data of these diseases are skewed to the right (Lee & Wang, 2003). Cohen (1951) and Harte and Moore (1966) discussed the Methods of estimating the parameters μ and σ^2 for complete samples of Lognormal distribution. Recently, Vernic, Teodorescu and Pelican (2009) used insurance data set to fit a survival mixture model of two components of the Lognormal distribution.

Let the survival time T be a random variable, if the distribution of $Y = \log T$ is normal with mean μ and σ^2 variance, then T follows the Lognormal distribution with mean, $\exp\left(\mu + \frac{1}{2}\sigma^2\right)$ and variance $\left[\exp(\sigma^2) - 1\right]\exp\left(2\mu + \sigma^2\right)$. It should be noted that μ and σ^2 are not the mean and variance of the Lognormal distribution. Figure 2.7 shows the nature of the probability density function of the Lognormal distribution. Figure 2.8 displays the survival function of the Lognormal distribution with mean $\mu = 5$ and standard deviation $\sigma = 3$.

Table 2.1 (p. 17) displays the probability density, the cumulative distribution and the survival functions of lognormal distribution where $t > 0$ and $\mu, \sigma > 0$.

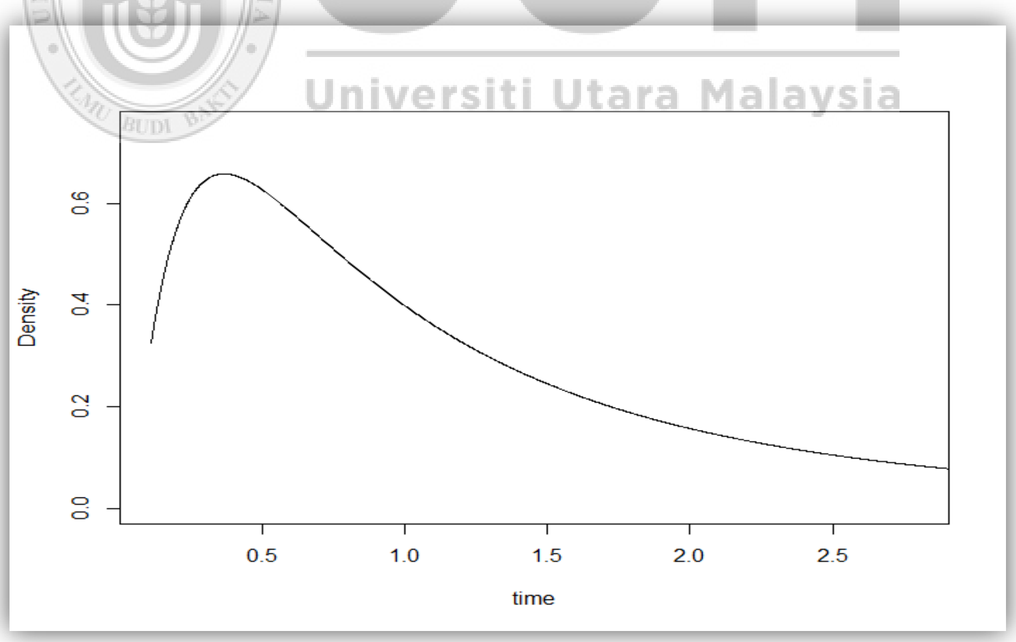


Figure 2.7 Density Function of the Lognormal Distribution $\mu = 0$ and $\sigma = 1$

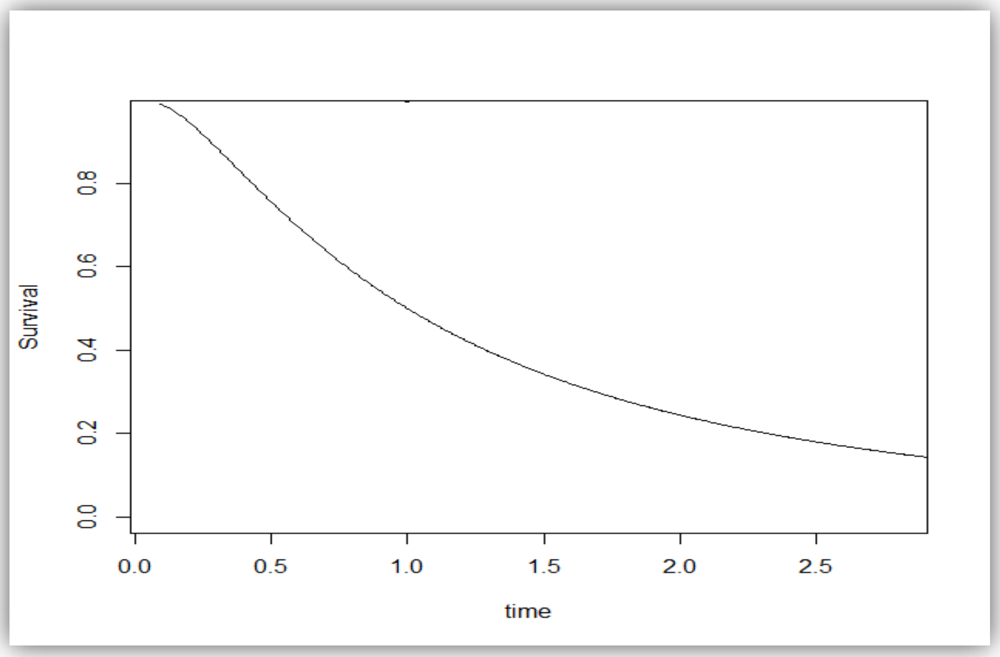


*Figure 2.8.*Survival Function of the Lognormal Distribution $\mu = 5$ and $\sigma = 3$

2.3.5 The Gompertz Distribution

Benjamin Gompertz (1825) was the first to formulate the Gompertz distribution function to fit mortality table. The probability density function is characterized by two parameters α and β, which must be both positive for a proper probability density function. When $0 < \alpha \le \beta$, the derivative of the density function is less than zero for $t \in (0,\infty)$, and the density function is monotone decreasing over $(0,\infty)$ with its mode at $t = 0$, and if $\alpha > \beta$, the density function increases on $(0, t_{\text{mod}e})$ and reaches its maximum at $t_{\text{mod}e}$ and then decreases to zero on $(t_{\text{mod}e},\infty)$ (Al-Hussaini, Al-Dayian & Adham, 2000). The Gompertz distribution is characterized by the fact that it describes the survival pattern that has a constant initial hazard rate. The hazard varies as an exponential function of time or age (Lee & Wang, 2003). Figure 2.9 shows the Gompertz density function with parameters α =1.2 and β =1.

Let the survival time T be a random variable that follows the Gompertz distribution, where $t > 0$. Figure 2.10 displays the survival function of the Gompertz distribution.

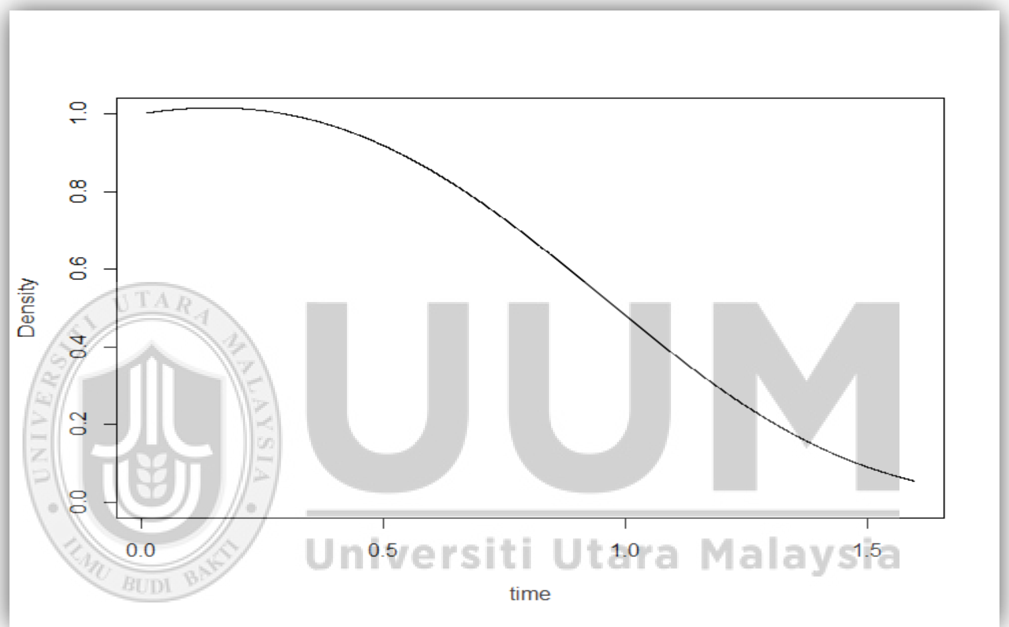


*Figure 2.9.*Probability Density Function of the Gompertz Distribution $\alpha = 1.2$ and $\beta = 1$

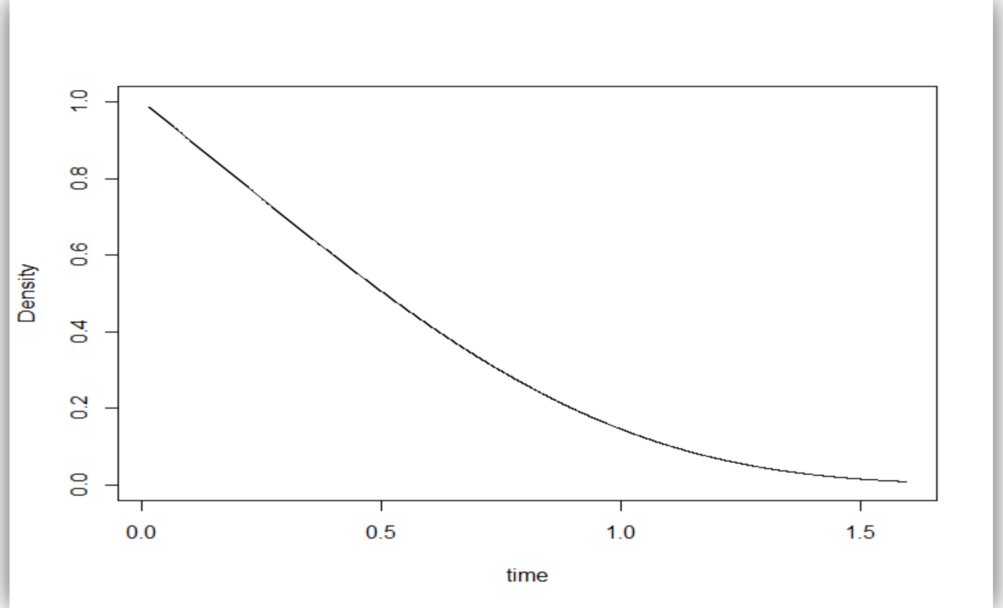


*Figure 2.10.*Survival Function of the Gompertz Distribution $\alpha = 3$ and $\alpha = 1$

2.3.6 Summary of Parametric Methods in Survival Analysis

Among the most frequently used parametric distributions in survival data analysis aare the Exponential, the Weibull, the Gamma, the Lognormal and the Gompertz distributions. Each of these distributions is characterized by a number of parameters and has different patterns in representing the survival and hazard functions (Ibrahim, et al., 2001; Kalbfleisch & Prentice, 2002; Lawless, 2003; Lee & Wang, 2003).

The parametric distributions are always preferred if some particular distribution seems to fit the data in the study as mentioned earlier. However, in cases where it is not appropriate to use a particular distribution, then the non-parametric method or distribution free approach is the alternative choice, which does not require any distributional assumptions, as it will be discussed in the next section.

2.4 Non-Parametric Methods in Survival Analysis

Non-parametric or distribution-free methods are among the oldest methods in survival data analysis (Kalbfleisch & Prentice, 2002; Lawless, 2003; Lee & Wang, 2003). They are flexible, easy to apply and do not require any theoretical distributional assumptions. When the survival data follow some particular theoretical parametric distribution, the non-parametric techniques are less efficient and loss their attraction. The most important non-parametric method in survival data analysis was introduced by the work of Kaplan and Meier (1958), which is used to estimate the survival function of a given survival data, and the estimated probabilities are also represented graphically.

Let $t_1 < t_2 < ... < t_n$ be the survival times for a sample of n individual, and let $t_1 < t_2 < ... < t_k$ be the times of the occurrence of the event of interest, where $(k \le n)$, and let d_j be the number of events of interest occurred at time t_j , and n_j the number of individuals at risk at time t_j, and $(n-k)$ the number of censored observations, then the Kaplan-Meier (K-M) estimate of the survival function $S(t)$ is defined by;

$$\hat{S}(t) = \prod_{j:t_j<t} \frac{n_j - d_j}{n_j} \tag{2.9}$$

The Kaplan-Meier or Product Moment survival function is a step decreasing function with jumps at observed events of interest (Lee & Wang, 2003). Figure 2.11 displays the non-parametric Kaplan-Meier survival function.

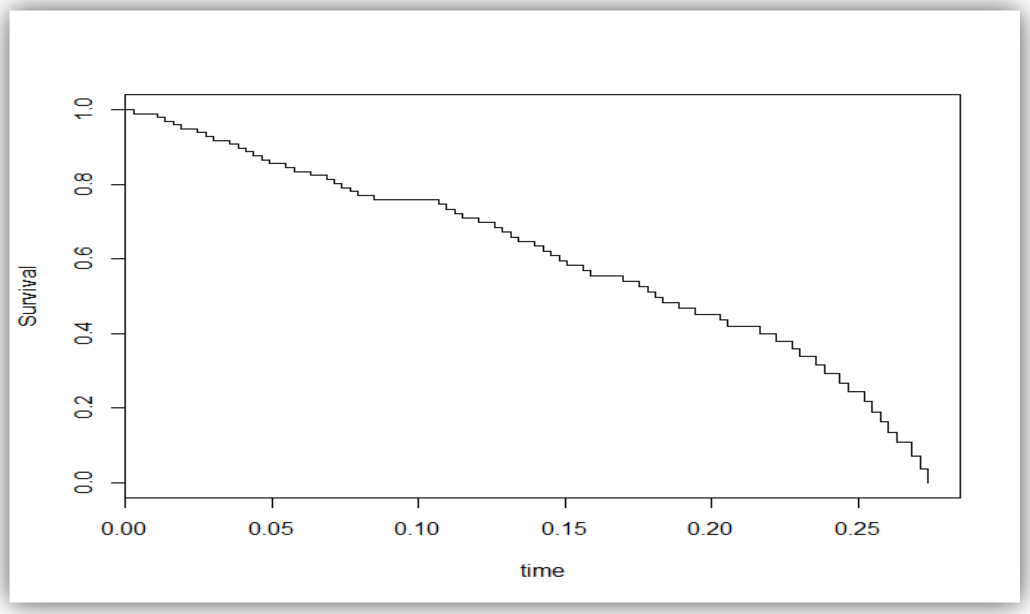


*Figure 2.11.*Kaplan-Meier Survival Function

In the same way, the cumulative hazard function can be estimated by the Nelson-Aalen, which is defined by

$$\hat{H}(t_i) = \sum_{j=1}^{i} \frac{d_j}{n_j} \qquad (2.10)$$

where n_j and d_j are as defined above, and the hazard can also be represented graphically. Figure 2.12 presents the cumulative hazard function.

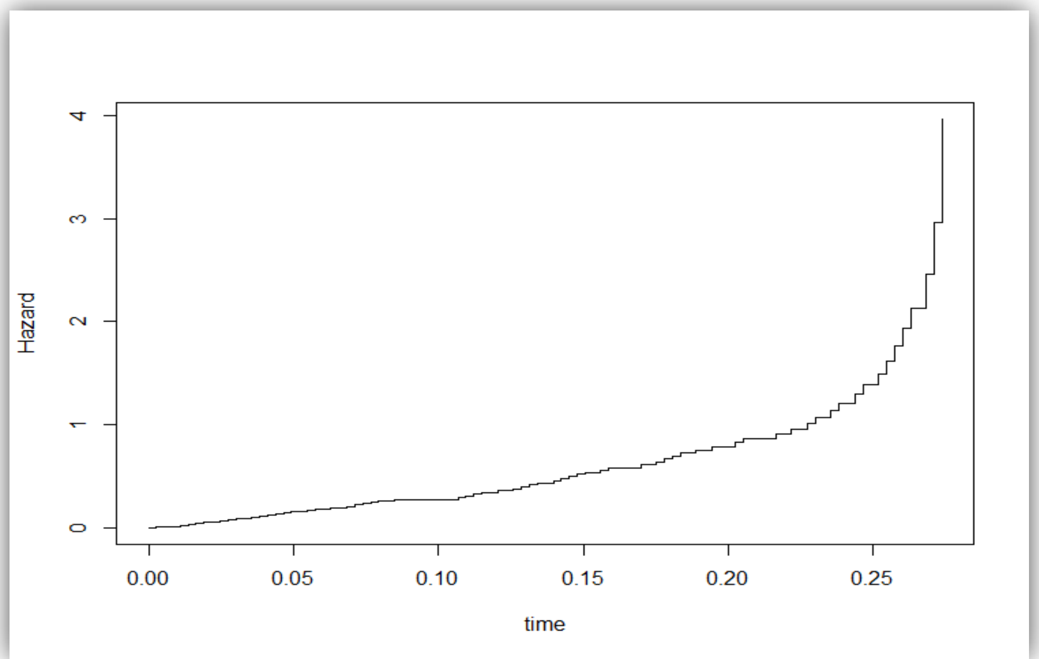


*Figure 2.12.*Nelson-Aalen Cumulative Hazard Function

The usual methods for modelling survival data are the parametric methods, when the distributional assumptions are satisfied or the non-parametric methods when no parametric distribution seems to fit the data. In some cases where the individuals are believed to arise from k distinct types with some proportion π, then survival mixture model methods will be the appropriate choice to model such data (Lawless, 2003). The survival mixture model methods in survival data shall be discussed in the following sections.

2.5 Mixture Models

Mixture models have been known for more than 100 years. In recent decades, the mixture models witnessed a huge expansion due to the rapid development in computing facilities. The application of mixture models covers different fields, including biology, medicine, physics, economics, engineering and marketing to mention a few (Leisch, 2004). Mixture models arise when modelling a set of data drawn from a population that is believed to consist of subpopulations within the main population, without any need to identify the subpopulation to which an individual observation belongs.

The parametric survival mixture models are more flexible compared to the pure classical parametric survival model, and they are the preferred choices for modelling the heterogeneous survival data. Bohning and Seide (2003) pointed out that, the reason why survival mixture models are developing very rapidly in recent decades, is their ability to offer natural models for unobserved population heterogeneity. Under the standard assumptions, the population is homogeneous, and pure classical parametric conventional distributions are used efficiently to model such homogeneous populations. However, when these assumptions are violated due to the heterogeneity of the population, the pure classical parametric survival models lose their attraction. Moreover, survival mixture models can easily handle cases of heterogeneous data.

In survival data analysis, it is sometimes observed that a considerable number of individuals or items do not experience the event of interest; they are referred to as long-term survivors. In such a situation, data analysis of the population under study cannot be appropriately handled by the conventional parametric distribution methods. Instead, the survival mixture model will be the best choice to model the data. Larson and Dinse (1985) proposed a survival mixture of long-term survivors to model survival data with multiple modes of failure and with censored observations.

Kuk and Chen (1992) used a long-term survivor semi-parametric survival model, where they employed proportional hazards for the time of occurrence of the event and logistic regression for the probability of occurrence of an event. Similarly, Taylor (1995) proposed a semi-parametric survival mixture model using logistic regression for the mixing probability and the nonparametric KM approach for the latency portion of the model. The two semi-parametric survival mixtures of parametric and non-parametric distributions (Kuk & Chen, 1992; Taylor, 1995) are a generalization of a parametric survival model proposed by Farewell (1982). Most of the above research works focus on semi-parametric survival modelling using a parametric model (logistic) for the mixing probability and a non-parametric distribution for the survival function of the uncured portion of the mixture model.

Recently, researchers expanded the literature on long-term survivor techniques considering parametric survival mixture models for estimation and making inference, assuming that some part of the population will not experience the event of interest (cured portion of the population). A long-term survivor's model was proposed by Copas and Hedary (1997), in which they used the Exponential distribution to model the re-offending of released prisoners within a given interval of time, during which they would have been in prison. In a study of the duration time until having a second child among Chinese women, where the population of women who have one child is assumed to consist of two subpopulations, women susceptible to having a second child within a given period of time, and those who will never have a second child, Li and Choe (1997), employed long-term survivors mixture model; they used the logistic regression model to assess the effects of covariates on the mixing probability of the mixture model, and the piecewise proportional hazard model to assess the effects of covariates on the conditional survival function. Koti (2001) used Lognormal Distribution to model long-term survival data where Logistic regression was proposed for the incidence part of the model, also the covariate effects were considered. Abu Bakar, Daud and Ibrahim (2006) proposed a long-term survivor logistic Weibull model to evaluate the effect of covariates associated with heart transplant surgery on the survival of the patient.

Yu and Peng (2008) proposed a marginal Mixture Cure Models (MCM) to model multivariate survival data, where they employed the Weibull distribution as the latency survival function for the uncured patients. Also, Khalid and Morgan (2008) used the Weibull distribution in a long-term survivor mixture model to compare the efficiency of longitudinal and cross-sectional settings. Likewise, Othus and Tiwari (2009) proposed a semi-parametric transformation cure model that includes the proportional hazard model and the proportional odds model and allowed for time dependent covariate in the cure mixture model. Seppa, Hakulinen, Kim, and Laara (2010) proposed a Generalized Gamma distribution to model the survival function for non-cured breast cancer patients to identify the regional variation in the cure fraction and in the survival of the non-cured patients. Logistic model was used to model the cured proportion. In the long-term survivors cure model, one distribution is used for survival function of non-cured, and in most cases, the logistic model is employed to model the cured proportion.

Mixture models have been discussed very frequently in the survival and reliability literature in cases of using mixtures of parametric and non-parametric distributions when it is not appropriate to use exclusively parametric or non-parametric distributions. For example, in a compromise between parametric and non-parametric distributions, Olkin and Spiegelman (1987) proposed a semi-parametric survival mixture procedure for estimating the probability density function of a mixture of parametric and non-parametric survival distributions. Also Kouassi and Singh (1997) and Zhang (2008), employing the technique of (Olkin & Spiegelman, 1987), proposed a semi-parametric survival mixture model to model the hazard and survival functions.

Despite the fact that the semi-parametric distributions are very flexible in terms of estimation, it was found that they are computationally intensive in terms of time-dependent mixing probability, the choice of parametric and non-parametric components need to be justified and the fact that the mixing probability does not have closed form (Kouassi & Singh, 1997). Due to these drawbacks, the pure parametric survival mixture models become a better alternative for modelling survival time data. (Olkin & Spiegelman, 1987; Kouassi & Singh, 1997; Zhang, 2008),

In most cases, parametric survival mixture models in the literature are mixtures of same parametric distribution. The method of moments was employed by Rider (1961) for the estimation of parameters of a parametric survival mixture of the Exponential distribution in a population assumed to have come from mixed Exponential distributions. Similarly, Jewell (1982) used the ML method to estimate the parameters of a proposed parametric survival mixture model of the Exponential distributions.

The Weibull distribution has been the most frequently explored distributions by many researchers for modelling survival data. Cheng and Fu (1982) employed the weighted least squares method to estimate the parameters of a two component parametric survival mixture of the Weibull distributions when data are grouped and censored. Jiang and Kececioglu (1992a) estimated the parameters of a two components parametric survival mixture model of the Weibull distribution graphically, exploring six different types of cumulative distribution functions of the survival mixture model. Also, Jiang and Kececioglu (1992b) proposed a new procedure to estimate the parameters of a two components parametric survival mixture model of Weibull distribution through EM with censoring data. Another graphical approach for estimating parameters of parametric survival mixture model of the Weibull distribution was proposed by Jiang and Murthy (1995), where they compared their results with that of an earlier published paper by Jiang and Kececioglu (1992a) and pointed out some errors in the previously published work. Jaheen (2005) proposed a parametric survival mixture of Exponential distributions to compare two different methods of parameter estimation, where the Bayesian and ML methods were compared. A parametric survival mixture model of two inverse Weibull distributions (MTIWD) was proposed by Sultan et al. (2007) where the properties of the parametric survival mixture were investigated and the parameters were estimated. Razli and Salih (2009) proposed a parametric survival mixture of two Weibull distributions: the first component with two parameters Weibull distribution and the second component with three parameters Weibull distribution, to model survival data with multiple modes of failure. They employed ML method to estimate the parameter and found that the mixing probability affects the parameter estimates.

It is observed that most of the parametric survival mixture models employed the same distribution in a two component mixture model. It would have been better to use different distributions to model such heterogeneous survival data in some instances. A parametric survival mixture model of Gompertz distribution was proposed by Al-Hussaini et al. (2000), to model heterogeneous survival data based on type I and type II censored samples, where they compared the Bayesian and ML estimation techniques and concluded that, the Bayesian estimates of the parameters of the parametric survival mixture model are generally better than those of the ML. Another study comparing the parameter estimates of ML and Bayesian method was conducted by Leng and Khalid (2010), where they employed a parametric survival mixture model of long-term survivors, which allowed for a cure fraction and frailty at the same time, and the baseline survival function was assumed to follow a Weibull distribution. They concluded that the ML estimators performed better than the Bayesian estimators.

A parametric survival mixture model of mixed distributions was proposed by Erisoglu and Erol (2010), to model heterogeneous survival time data, where they employed a two component parametric survival mixture model of the Extended Exponential-Geometric (EEG) distribution, and they used a real survival data to estimate the parameters of the survival mixture model. Very recently, Erisoglu et al. (2012) employed two components parametric survival mixture models of Weibull, Gamma and Lognormal to model heterogeneous survival time data; the parameters of the model were estimated using the EM Algorithm. They employed the Akaike Information Criterion (AIC) to test the best fit distribution (Akaike, 1974). Their models were applied to real data. This work is among the few that treated parametric mixture model of the Gamma distribution.

Erisoglu and Erol (2010) and Erisoglu et al. (2012) employed two components parametric survival mixture model of same distribution. In some situations, it would have been appropriate to use three components parametric survival mixture models of the same distributions to model such heterogeneous survival data. The study by Erisoglu et al. (2012) that considered the Gamma distribution could be extended to a three components parametric survival mixture model of the Gamma distribution which did not get thorough attention. Also, in some case parametric survival mixture models of different distribution may be more suitable. As an extension of Ersioglu et al. (2011) which considered a two components parametric survival mixture model of different distributions, a three components parametric survival mixture model could be considered.

2.6 Three Components Parametric Survival Mixture Models of Same Distribution

Many research works have been done on parametric survival mixture models of a same distribution, and various distributions were used for that purpose. However, most of the works concerned two components parametric survival mixture models (Cheng & Fu, 1982; Jiang & Kececioglu, 1992a; Jiang & Kececioglu, 1992b; Zhang, 2008; Erisoglu & Erol, 2010; Farcomeni & Nardi, 2010; Erisoglu, Erisoglu & Erol, 2011; Zang, Wang & Lu, 2011; Erisoglu, Erisoglu & Erol, 2012). Very little has been done concerning parametric survival mixture models of same distribution with three components. For example, mixture model of three components Weibull distributions were employed to model data with three subpopulation categorized as strong, freak and infant mortality (Jensen & Petersen, 1982; Moltoft, 1983). Jiang and Kececioglu (1992) used simulation to model mixed- Weibull distribution. They proposed two, three and five component mixture model all with complete data using EM to estimate the parameters of the models. Jiang and Murthy (1996) used weibull mixture model of three components with the shape parameters and mixing probabilities arranged in ascending order. Marin, et al. (2005) proposed a parametric survival mixture model of three components of the Weibull distribution where they employed Bayesian method to estimate the parameters of the model. Moreover, a very important distribution like the Gamma distribution received very little attention in parametric survival mixture models. Erisoglu, et al. (2012) employed two components parametric survival mixture model of the Gamma distribution. Hanson (2006) used the Bayesian method to analyse lifetime censored data by employing three components mixture model of Gamma distributions. The study treated the survival distributions as a Dirichlet process mixture of Gamma distributions. There is a need to investigate a three component parametric survival mixture model of the Gamma distribution.

2.7 Three Components Parametric Survival Mixture Model of Different Distributions

In the literature, most of the parametric survival mixture models employed mixtures of the same parametric distributions. There are very few cases where parametric survival mixtures of different parametric distributions are investigated. Chang (1998) employed parametric statistical model to analyse mortality structure in Taiwan, where he used a mixture of three different distributions; Weibull, Inverse Weibull and Gompertz distributions. The study was of life table of Taiwan between 1926 and 1991. A parametric survival mixture model of two components of different distributions was proposed by Abu-Zinadah (2010), where Exponentiated pareto and Exponential distributions were used to model survival data, and the Bayesian and ML methods of estimation were employed and compared. They concluded that, generally, the performance of the ML was found to be better than that of the Bayesian estimates. Also, Erisoglu et al. (2011) proposed a parametric survival mixture model of two components of different distributions to model heterogeneous survival time data, where they employed parametric survival mixtures of Exponential-Gamma, Exponential-Weibull and Gamma-Weibull to estimate the parameters of the models.

Finite mixture models are capable of capturing unobserved heterogeneity in survival time data, which make them extremely flexible in representing the density of various k-components of mixture models. That makes them a very good choice for modelling heterogeneous survival data, which arise in real life (Fruhwirth-Schnatter, 2006).

Blackstone, et al. (1986) classified the death after surgery, in the case of open-heart surgery, into three overlapping phases, which could be modelled by a three component parametric mixture model (Ng, McLachlan, Yau, & Lee, 2004; Philips, Coldman & McBride, 2002). In such a situation where different modes of hazard are identified, a parametric survival mixture model of different distributions could be a good choice of modelling such survival data.

Considerable research works have been done on parametric survival mixture models of same distribution, and various distributions were used for that purpose. However, very little has been done concerning parametric survival mixture models of different distributions in terms of estimating the parameters of the survival mixture models and investigating their suitability and appropriateness. Consequently, parametric survival mixture models of different distributions could be a better choice to model such heterogeneous survival time data.

2.8 Summary

Parametric survival mixture model is the preferred choice for modelling heterogeneous survival data over the pure classical parametric survival model. Many studies considered the parametric survival mixture model and most of the studies focused on the two components parametric survival mixture models of the same distribution. The Gamma distribution received less attention compared to the other distributions, and there was no study that investigated the three components parametric survival mixture model of the Gamma distribution. However, very few studies considered two component parametric survival mixture model of different distribution. No study considered a three components parametric survival mixture model of Exponential, Gamma and Weibull distributions.

This study is aimed at filling the gap in the literature by proposing a parametric survival mixture model of three components, where a three components parametric survival mixture model of the Gamma distributions (Model 1) and a three components parametric survival mixture of the Exponential, Gamma and Weibull distributions (Model 2) were considered. The EM was employed in the estimation of the ML parameters of the models.

CHAPTER THREE
METHODOLOGY

3.1 Introduction

This chapter is to outline the methodology employed in realizing the objectives of the study. First objective is to develop three components parametric survival mixture model of the Gamma distributions (Model 1). The second objective is to develop three components parametric survival mixture model of the Exponential, Gamma and Weibull distributions (Model 2). The third objective is to employ simulated data to evaluate the performance of the models with different sample sizes, different censoring percentages and different mixing probabilities. The fourth is to investigate the effect of different censoring percentages on the hazard function of the models using simulated data. The fifth objective is to evaluate the performance of the estimate of the models by investigating the survival function of the models using real data and compare it with the K-M empirical survival function of the real data.

Figure 3.1 and Figure 3.2 outline the methodology framework for both Model 1 and Model 2 respectively. Expectation Maximization (EM) was used to estimate of the two models. Simulation study was used and the performance of Model 1 and Model 2 was investigated using three different sample sizes, three different censoring percentages and two sets of three different mixing probabilities.

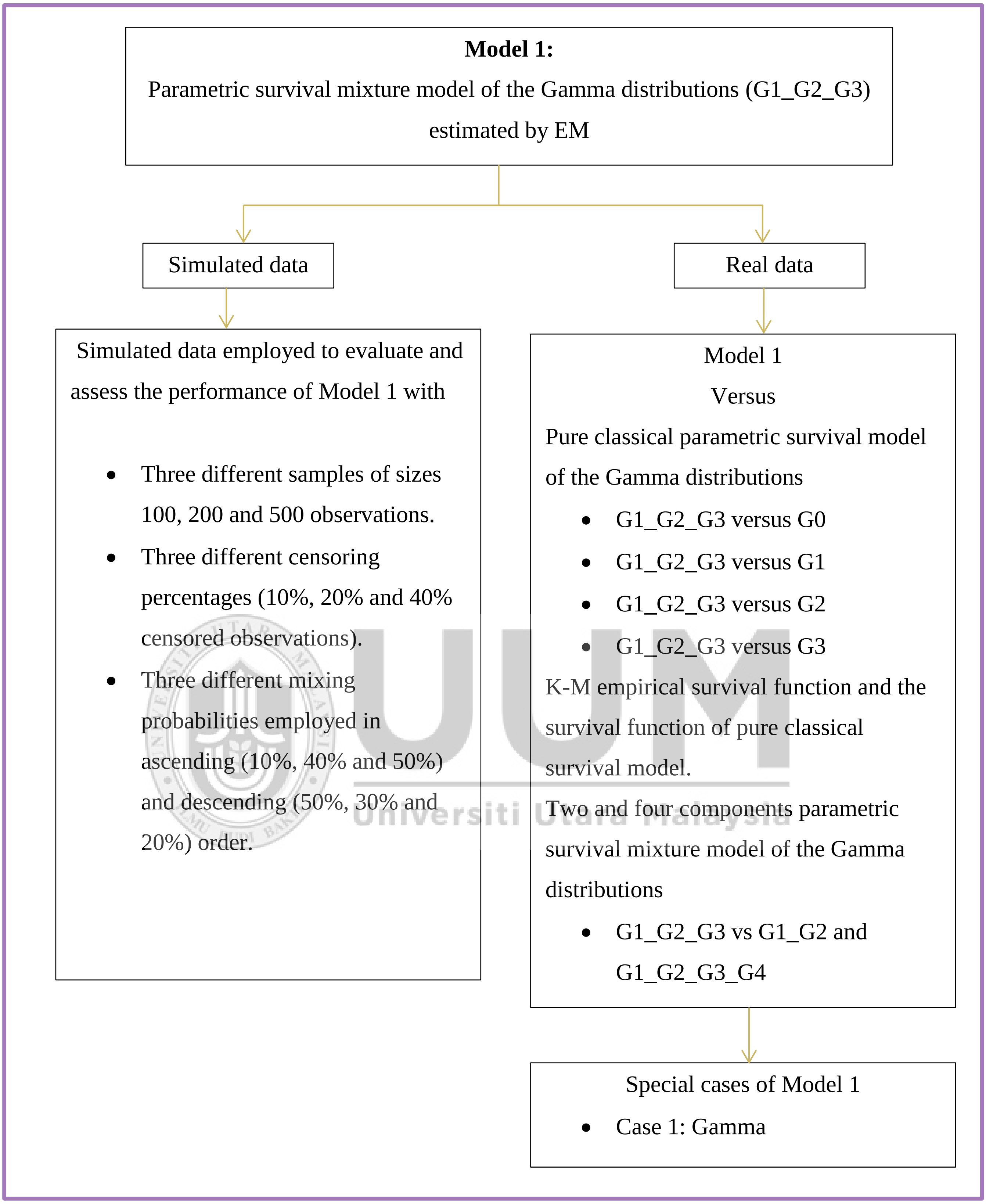
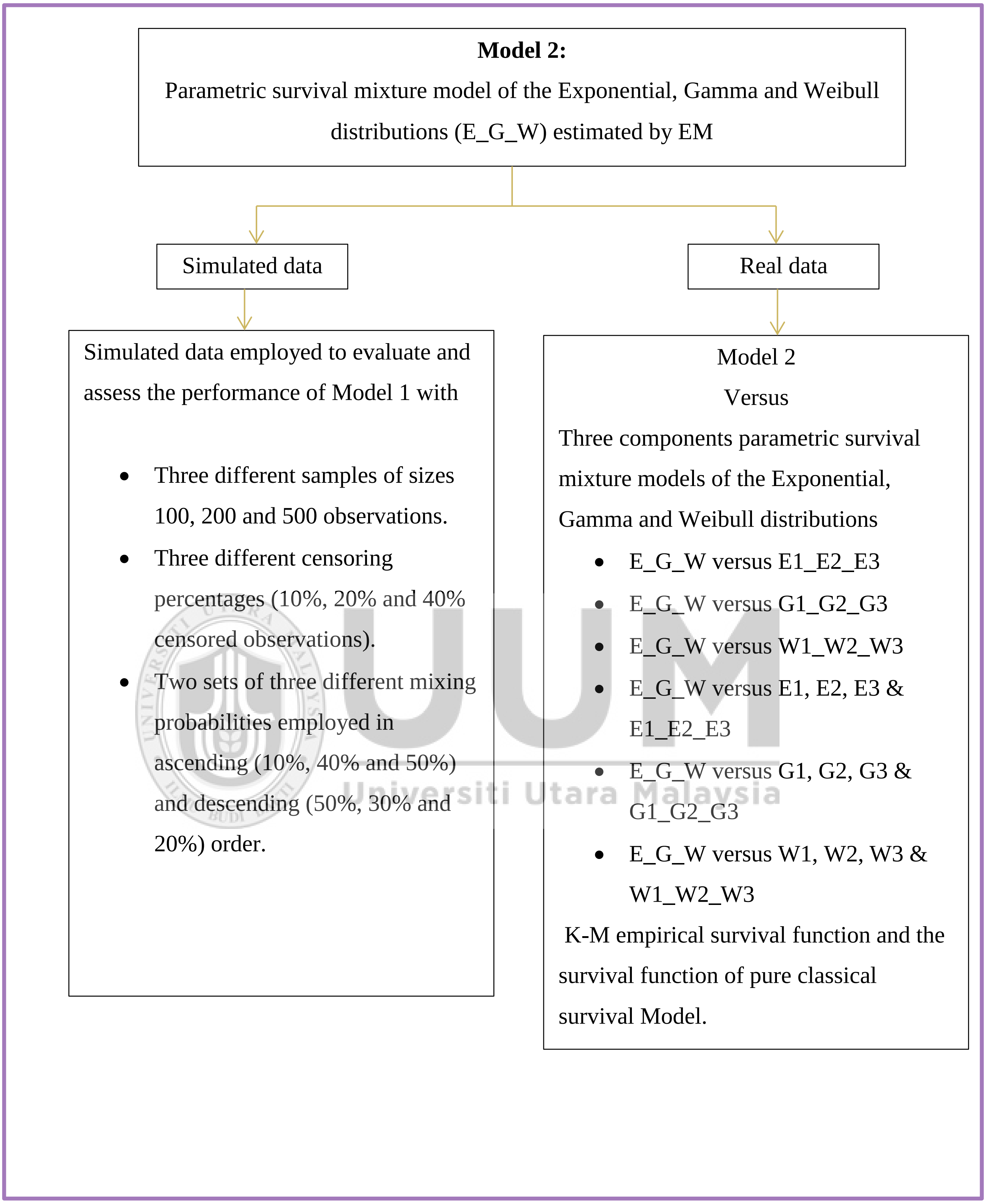

Model 1:
Parametric survival mixture model of the Gamma distributions (G1_G2_G3) estimated by EM
Simulated data
Real data
Simulated data employed to evaluate and assess the performance of Model 1 with
• Three different samples of sizes 100, 200 and 500 observations.
• Three different censoring percentages (10%, 20% and 40% censored observations).
• Three different mixing probabilities employed in ascending (10%, 40% and 50%) and descending (50%, 30% and 20%) order.
Model 1
Versus
Pure classical parametric survival model of the Gamma distributions
• G1_G2_G3 versus G0
• G1_G2_G3 versus G1
• G1_G2_G3 versus G2
• G1_G2_G3 versus G3
K-M empirical survival function and the survival function of pure classical survival model.
Two and four components parametric survival mixture model of the Gamma distributions
• G1_G2_G3 vs G1_G2 and G1_G2_G3_G4
Special cases of Model 1
• Case 1: Gamma

*Figure 3.1.*Methodology Frameworks of Model 1



Model 2:

Parametric survival mixture model of the Exponential, Gamma and Weibull distributions (E_G_W) estimated by EM

Simulated data

Real data

Simulated data employed to evaluate and assess the performance of Model 1 with

- Three different samples of sizes 100, 200 and 500 observations.
- Three different censoring percentages (10%, 20% and 40% censored observations).
- Two sets of three different mixing probabilities employed in ascending (10%, 40% and 50%) and descending (50%, 30% and 20%) order.

Model 2

Versus

Three components parametric survival mixture models of the Exponential, Gamma and Weibull distributions

- E_G_W versus E1_E2_E3
- E_G_W versus G1_G2_G3
- E_G_W versus W1_W2_W3
- E_G_W versus E1, E2, E3 & E1_E2_E3
- E_G_W versus G1, G2, G3 & G1_G2_G3
- E_G_W versus W1, W2, W3 & W1_W2_W3

K-M empirical survival function and the survival function of pure classical survival Model.

*Figure 3.2.*Methodology Frameworks of Model 2

The simulated data were employed to investigate the effect of the different censoring percentages on the hazard function of both of Model 1 and Model 2 respectively.

Empirical study was conducted to validate Model 1 and Model 2 using real data. Graphical, Log Likelihood (LL), Akaike Information Criterion (AIC), Mean Square Error (MSE), Root Mean Square Error (RMSE), Kolmogorov-Smirnov test (K-S) and the mean survival time *E(t)* were employed in the validation the model.

The properties of Model 1 and Model 2 were investigated by comparing the K-M empirical survival function of the real data with the survival function of Model 1 and Model 2 evaluated using real data. Model 1 was compared with pure classical parametric survival models, two and four components parametric survival mixture model of the Gamma distributions respectively. A special case for using Model 1 was also investigated.

Model 2 was compared with pure classical parametric survival models and three components parametric survival mixture models of the Exponential, Gamma and Weibull distributions respectively.

3.2 Development of Model 1 and Model 2

In this section a three components parametric survival mixture model of the Gamma distributions (Model 1) and a three components parametric survival mixture model of the Exponential, Gamma and Weibull distributions (Model 2) were developed to model heterogeneous survival data. EM was employed to estimate the parameters of the two models. The two models consist of three mixing probabilities to connect the three components corresponding to each of the three distributions. Note that the mixing probabilities sum up to one. The procedures of employing simulated and real data in estimating the parameters of Model 1 and Model 2 were explained in Chapters Four and Five respectively. Empirical study was employed using real data to compare Model 1 and Model 2 with pure classical survival models and validate the two models.

3.2.1 Validating the Performance of the Models Using Simulated Data

To evaluate the performance of the two models, data were generated based on three different sample sizes. The samples were categorized into small sample (100 observations or less), medium sample (200 observations) and large sample (500 observations or more). Each of these three samples consists of three different censoring percentage (10%, 20% and 40% censored observations) and two different sets mixing probabilities. The two different sets of mixing probabilities were arranged in ascending order (10%, 40% and 50%) and descending order (50%, 30% and 20%). Based on this simulation study, eighteen (18) different random samples of survival data were generated from a population of three components parametric survival mixture model each for Model 1 and Model 2. The data were generated based on random censoring procedure which was explained in section 4.2.2 of Chapter Four.

The choice of the three different sample sizes, the three different censoring percentages and the three mixing probabilities were based on some previous simulation and real data applications as summarized in Table 3.1.

Table 3.1

Some Previous Simulation and Real Data Applications

No	Author/s	simulation			Real data		
		Sample size	Mixing probabilities	Censored Items	Sample size	Mixing probabilities	censored
1	Marin et al. (2005)	150	60%, 30% and 10%	10%	87	-	40%
2	Larson and Dinse (1985)				65		37%
3	Zhang and Wang (2011)	50 100 150 200	60% and 40%	20%			
4	Erisoglu and Hamza (2010)	500	-70% and 30% -65% and 35% - 85% and 15% - 80% and 20%	-	50	46% and 54%	-
5	Siyuan Jiang and Kececioglu	200 500	10%, 40% and 50%	-			
6	Zhang (2008)	500	40% and 60%	20%	51	18% and 82%	
7	Erisoglu et al. (2011)	100	- 60% and40% -80% and20% -30% and 70%	-		71% and 29% 66% and 44% 65% and 45%	
8	Wiper, et al. (2001)	400	20%, 60% and 20%	-	203	-	
9	Moltoft (1983)	-	-	-	-	2%, 10% and 88%	

Sample sizes employed in simulation studies ranged between 50 - 500 observations. Marin et al. (2005) employed sample size of 150 observations in the simulating survival data for three component survival mixture model of Weibull distribution via Bayesian method. Zhang and wang (2011) simulated survival data to estimate the parameters of a two component Mixed Weibull Distributions. They used samples of sizes 50, 100, 150 and 200 observations. Three and five components survival mixture model of Weibull distribution of complete data was simulated with sample size of 200 and 500 observations respectively (Siyuan Jiang and Kececioglu, 1992).

McGilchrist and Aisbett, (1991) used the Kidney catheter data which consist of 76 observations in their work. Marin et al., (2005) analyzed the Lupus nephritis survival data which include 87 individuals. The Bone Marrow transplant data used by Kersey, et al., (1987) consist of 91 observations. The lung cancer data consist of 137 observations and the Heart transplant data consist of 184 observations (Kalbfleisch and Prentice, 2002).

Based on the simulation studies and real data employed in some research works, this study considers three different sample sizes (small, moderate and large), for simulating survival data. The sample size of 100 observations which cover observation less than or equal to 100 was considered as small sample size. The sample size of 200 observations was considered as moderate sample size. The sample size of 500 observations which covers observations greater than or equal to 500 was considered as large sample size.

Different censoring percentages were considered in simulating data of survival mixture models. Marin, et al. (2005) simulated survival data of three components Weibull distribution with 10% censored observations. In other research works, 20% censored observations was considered (Zhang and Wang, 2011; Zhang 2008).

In real survival data, the Lung Cancer data consist of 9 censored observations out of 137 individuals, which make the censoring about 7% of the data (Kalbfleisch and Prentice, 2002). In vaginal cancer data, four out of the forty observations (10% censoring) were censored (Kalbfleisch and Prentice, 2002). Twenty four percent of the Kidney catheter data are censored (McGilchrist and Aisbett, 1991). The real data employed by Marin et al. (2005), Larson and Dinse (1985) and Stanford heart transplant data have 40%, 37% and 38% censored observations respectively. Therefore, since the most frequently encountered censoring percentage were between 10 to 40 percent. The censoring percentages considered in this simulation were 10%, 20% and 40%.

Mixing probabilities specify the proportion of the observation that belongs to each component of the survival mixture model. Most of the simulation studies considered unequal mixing probability either in ascending or descending order. Zhang (2008) simulated data with mixing probability of 40% for component one and 60% for component two of a survival mixture of Weibull distribution. Erisoglu et al. (2011) simulated data with 30% and 70% mixing probability for mixture of Gamma-Weibull distribution and 60% and 40% for Exponential and Weibull distributions. Zhang and Wang (2011) used 60% and 40% for a survival mixture model of Weibull distributions. In the simulation of three components mixture model of Weibull distribution, 60%, 40% and 10% was used for mixing probabilities (Marin et al, 2005). Also, Siyuan Jiang and Kececioglu (1992) simulate survival data with 10%, 40% and 50% mixing probabilities. In real data application, the Kidney catheter data were used to model a three components survival mixture mode of Exponential, Gamma and Weibll distribution with 52%, 29% and 19% mixing probabilities. Therefore, this study will adapt the mixing probabilities 10% 40% and 50% for ascending order and 50%, 30% and 20% for descending order.

EM was employed to estimate the ML estimators of the parameters of the postulated parametric survival mixture models. The EM has been efficiently applied in cases of data with missing or unobserved observations. The EM can be traced back to the famous work presented by Dempster, Laird and Rubin (1977), where they gave the general formulation of the EM. Subsequently, its application continued to expand and develop (McLachlan & Krishnan, 2008).

The EM is an iterative procedure of computing ML estimators when observations are viewed as incomplete in the presence of missing or hidden data. The iteration consists of two processes, namely, the Expectation step or the E-step and the Maximization step or the M-step. In the Expectation step, the missing data are estimated given the observed data and the current estimate of the model parameters. This is achieved using the conditional expectations. In the Maximization step, the likelihood function is maximized under the assumption that the missing data are known. The estimates of the missing data from the Expectation step are used instead of the actual data. The convergence is assured since the EM is guaranteed to increase the value of the likelihood at the end of each of the iterations (Dempster, Laird & Rubin, 1977; McLachlan & Krishnan, 2008).

The parameter estimates of the postulated Model 1 and Model 2 were evaluated using the EM and the estimated values were compared with the true values used in the simulation of the data.

To investigate the consistency and stability of the EM in estimating the parameters of Model 1 and Model 2, the data regeneration was repeated 300 times for each set of data. The parameters of the postulated Model 1 and Model 2 were estimated. The averages of the estimated parameters along with the mean square errors (MSE) and root mean square errors (RMSE) were reported.

3.2.2 Validating the Models Using Real Data

Three sets of real data were involved in the empirical study to validate Model 1 and Model 2. The set of survival real data include the Bone Marrow Transplant data, the Vaginal Cancer data and the Kidney Catheter data.

The first set of data is the Bone Marrow Transplant. The data are among the sets of data included in the *smcure* package developed by Cai, et al. (2012) of R statistical software (Team, 2005). The set of data were originally used in the study for the refractory acute lymphoblastic leukaemia patient (Kersey, et al., 1987). The data consist of 91 observations with 21 censored observations (approximately 23% censoring).

The second set of data is the Kidney Catheter data. The data are included as one of the data set in the famous *survival* package developed by Therneau (1999) of the R statistical software (Team, 2005). This data were studied originally by McGilchrist and Aisbett (1991). The data give the recurrence times to infection, at the point of insertion of catheters, of kidney patients using portable dialysis equipment. It consists of 76 observations and 7 variables as presented in Appendix B. The data constitute of 18 censored observations which makes the censoring percentage approximately 24%.

The third set of data is the Vaginal Cancer data. The data give the time to death of vaginal cancer of some rats as a result of insult with the carcinogen (Kalbfleisch & Prentice, 2002). The data constitute of 40 observations of survival time of two groups of rats distinguished by a pre-treatment regiment; there are four censored observations which make the censoring percentage approximately 10%. See Appendix B.

The first and third sets of the data were employed in validating Model 1 and the second set of the data were used to validate both Model 1 and Model 2 respectively.

For the validation of the models several graphical representations were employed. The probability density function of Model 1 and Model 2, the probability density function of the pure classical survival model corresponding to component evaluated using the real data and the histogram of the real data was presented. Also the K-M empirical survival function of the real data were presented graphically together with the survival function of the two models and the survival function of the pure classical survival model corresponding to each component.

For graphical comparison of Model 2 with E1_E2_E3 and W1_W2_W3 survival mixture models, the graph of the probability density functions of Model 2, the probability density function of E1_E2_E3, the probability density function of W1_W2_W3 and the histogram of the Kidney Catheter data was presented.

Model 2 was compared, on separate graphs, with the parametric survival mixture models of E1_E2_E3, W1_W2_W3; each evaluated using the real data together with their pure classical survival parametric models corresponding to each component of the mixture models and the histogram of the Kidney Catheter data.

To validate the models the LL values of each model were evaluated and compared with the LL values corresponding to the pure classical survival models corresponding to each component of the mixture models.

The most frequently used method in model selection, the AIC recommended by Akaike (1974) is employed in the model selection.

The AIC criterion is defined by

$$AIC = -2\log L + 2d \tag{3.1}$$

where d is the number of free parameters in the finite parametric survival mixture model. The Bayesian Information Criterion (BIC) may be used for model selection (Fraley and Raftery, 2002). However, Burnham and Anderson (2002, 2004) pointed out that the AIC has theoretical advantage over the BIC. To investigate the appropriateness of Model 1 and Model 2 compared to the pure classical survival models and the other survival mixture model the AIC value were evaluated.

In many instances, the Mean Square error (MSE) has been used as one of the methods of quantifying the difference between an estimator and the true value of the quantity being estimated. MSE is obtained with

$$MSE = \frac{\sum \left\{ Emp(t_i) - F(t_i) \right\}^2}{n - p}$$

(3.2)

where $Emp(t_i) = (i - 0.3)/(n + 0.4)$ for $i = 1,2,\ldots,n$ is the empirical distribution, p is the number of free parameters in the distribution and $F(t_i)$ is the theoretical distribution function (Erisoglu et al., 2011). The MSE values of Model 1 and Model 2 were evaluated and employed in comparing the models with the pure classical survival models corresponding to each distribution. The RMSE values which are defined as the square root of the MSE were also evaluated.

The Kolmogorov-Smirnov (K-S) test is frequently employed to test whether the cumulative distribution of set of data comes for a particular parametric distribution. The K-S test was employed to test the fitness of Model 1 and Model 2 compared to the pure classical survival models corresponding to each component.

The mean survival time *E(t)* was evaluated for both Model 1 and Model 2 and compared with those of the pure classical survival models corresponding to the each component. The mean survival time E(t) was defined in Chapter Two, Table 2.1 (p. 17).

Everitt and Hand, (1981) considered investigating the histogram of set of data to decide whether a mixture model structure is more appropriate than a pure classical model. The histogram shows the sign of multimodality in the data which suggests the appropriateness of mixture model with subpopulation. To decide the number of components which indicates the sup-population in the set of data, model selection method AIC is used to decide the number of sup-population appropriate to the data (McLachlan and Peel, 2004; Fruhwirth-Schnatter, 2006; Erisoglu et al., 2012). Based on this model selection were performed by evaluating the values of LL and AIC to show that the three component sub-population is the most appropriate to represent the real data for Model 1.

Model selection was also employed to select the model that represents the Kidney Catheter data better among Model 2, the parametric survival mixture model of the Exponential distributions (E1_E2_E3) and the parametric survival mixture model of the Weibull distributions (W1_W2_W3) respectively. The LL and the AIC values were computed and used for the model selection.

3.3 Summary

In this chapter the procedure of employing Model 1 and Model 2 to model heterogeneous survival data was explained. The EM was used in the estimation of the parameters of the two models. The performance of the two models was investigated via simulation and empirical studies.

Simulation study showed the procedure used to evaluate the two models by generating three different samples of size (100, 200 and 500) observations. Each sample constituted of three different censoring percentages (10%, 20% and 40%) and two sets of three different mixing probabilities. The first set of the mixing probabilities was arranged in ascending order (10%, 40% and 50%) and the second set was in descending order (50%, 30% and 20%). The simulations were repeated

300 times where the MSE and RMSE were estimated. The simulated data were used to investigate the effect of the different censoring percentages on the nature of the hazard function of the two models. The simulation study was explained in detail for Model 1 and Model 2 in Chapters Four and Five respectively.

Empirical study carried out to validate the performance of the two models using three sets of real data was explained. The empirical study considered the estimation of the parameters of the model. Graphical representation were used as well as the LL, AIC, MSE, RMSE, K-S and E(t) to evaluate the performance of the models. The K-M empirical survival function was used to evaluate the fitness the two models. The K-M of the real data was compared with the survival function of the two models and that of the pure classical survival model corresponding to each distribution graphically. The empirical study was presented in detail for both of Model 1 and Model 2 in Chapter Four and Five respectively.

CHAPTER FOUR
THREE COMPONENTS PARAMETRIC SURVIVAL MIXTURE MODEL OF THE GAMMA DISTRIBUTIONS

4.1 Introduction

This chapter is dedicated to develop a three components parametric survival mixture model of Gamma distributions (G1_G2_G3, noted as Model 1). The outlines of the chapter are as follows. The first section highlighted the theoretical development of Model 1 by applying the Expectation Maximization (EM). The section also includes the explanation regarding the algorithm transformation to computer coding. The next two sections are validation of Model 1 conducted based on simulated and real data respectively. The last section summarized the outcomes and findings of this chapter.

4.2 Theoretical Development of Model 1

Since the first objective of this study proposes a three components parametric survival mixture model of the Gamma distributions (Model 1), then Model 1 can be expressed as follows

$$f_{G1\_G2\_G3}(t;\Theta) = \pi_1 f_{G1}(t;\alpha_1,\beta_1) + \pi_2 f_{G2}(t;\alpha_2,\beta_2) + \pi_3 f_{G3}(t;\alpha_3,\beta_3)\ , \qquad (4.1)$$

where π_i are the mixing probability and $\sum_{i=1}^{3}\pi_i = 1$. The functions f_{G1}, f_{G2} and f_{G3} are the probability density functions of the Gamma distributions corresponding to each component of Model 1. The EM employed to estimate the ML estimates of the parameters of the Model 1 proceeds by considering the survival mixture model of Gamma distribution. To explain the derivation of the parameters of the mixture model, the general formulation of three component parameteric survival mixture model was highlighted in the next subsection.

4.2.1 General Formulation of Three Component Parametric Survival Mixture Model

The general formulation of a three components parametric survival mixture model which is assumed to consist of three subpopulations; each subpopulation corresponds to a component in the parametric survival mixture model, can be expressed as follows

$$f_{X,Y,Q}(t;\Theta) = \pi_1 f_X(t;\theta_X) + \pi_2 f_Y(t;\theta_Y) + \pi_3 f_Q(t;\theta_Q)\ , \qquad (4.2)$$

where the vector $\Theta = (\ \pi_1, \pi_2, \theta_X, \theta_Y, \theta_Q)$, contains all the unknown parameters in the parametric survival mixture model. The functions $f_X(t;\theta_X), f_Y(t;\theta_Y)$and $f_Q(t;\theta_Q)$ are known as the probability density functions corresponding to each component of the parametric survival mixture model for some parameters $\theta_X, \theta_Y$and θ_Q respectively. The π_i are the mixing probability of the survival mixture mode and $\sum_{i=1}^{3} \pi_i = 1$.

To highlight the derivation of the parameters of survival mixture model in (4.2), there is a need to highlight the development of the EM and its implementation in the estimation of the parameters of survival mixture model with censored observations.

Consider the random variable $T_1, T_2, ..., T_n$ of size *n* observations to represent survival data of some *n* objects or individuals, where T_j denotes the survival time of the j^{th}

object or individual. The probability density function $f(t)$ of T_j is assumed to be a linear mixture denoted by

$$f(t) = \sum_{i=1}^{k} \pi_i f_i(t;\theta_i) \tag{4.3}$$

where $f_i(t;\theta_i)$ is the probability density function of the i^{th} component of the mixture, θ_i represents the parameters corresponding to the i^{th} density and π_i's are the mixing probabilities corresponding to each component of the mixture and satisfy the condition,

$$\sum_{i=1}^{k} \pi_i = 1 \text{ and } 0 \le \pi_i \le 1 \quad (i = 1,2,\ldots,k) \tag{4.4}$$

The mixture (4.3) is considered a density function since $f_1(t;\theta_1), f_2(t;\theta_2),\ldots, f_k(t;\theta_k)$ are probability density function corresponding to the *k*-components.

The survival function of random variable T_j can be expressed as a linear mixture of the survival function corresponding to the *k*-components and is denoted by,

$$S(t) = \sum_{i=1}^{k} \pi_i S_i(t;\theta_i) \tag{4.5}$$

where $S_i(t;\theta_i)$ is the survival function of the i^{th} component.

4.2.2 Estimation of Parameters Using the EM

Assume the density of a *k*-component mixture of a random variable *Y* is defined by

$$f(y;\Theta)=\sum_{i=1}^{k}\pi_i f_i(y;\theta_i) \tag{4.6}$$

where $\Theta=(\pi_1,\pi_2,...,\pi_{k-1},\theta_1^{'},\theta_2^{'},...,\theta_k^{'})^{'}$ is the vector of the unknown parameters of the mixture model. Consider the vector $\pi=(\pi_1,\pi_2,...\pi_k)^{'}$ to be the vector of mixing probabilities which sum up to one. Assuming the vector $y_1,y_2,...y_n$ is a vector of an observed sample of size *n* then the likelihood for Θ can be expressed as

$$L(\Theta)=\prod_{j=1}^{n}f(y_j;\Theta)$$

$$=\prod_{j=1}^{n}\left[\sum_{i=1}^{k}\pi_i f_i(y_j;\theta_i)\right] \tag{4.7}$$

It is difficult to computationally estimate maximum likelihood of mixture model using classical method, that is, by taking derivative with respect to each parameter. This difficulty has been simplified considerably by the use of EM introduced by Dempster, Laird, and Rubin (1977). The estimation is simplified by considering the data to be incomplete data (McLachlan 2000). In order to pose this problem as an incomplete-data, assume that the vector $y=(y_1^{'},y_2^{'},...,y_n^{'})^{'}$, denotes the observed random sample obtained from the mixture density (4.6). Now a vector of an unobservable or missing data is introduced,

$$Z = (Z_1^{'}, Z_2^{'}, \ldots, Z_k^{'})^{'} \tag{4.8}$$

where Z_j, $j = 1,2,\ldots,n$ is k-dimensional component-label vector.

In mixture models, the EM framework considers the observed values $y = (y_1^{'}, y_2^{'}, \ldots, y_n^{'})^{'}$ to be the incomplete data and latent class variables in (4.8) to be the missing data, where $z_{ji} = z_j(y_i) = 1$ if observation y_i belongs to j^{th} class and 0 otherwise for $j = 1,2,\ldots,k$ and $i = 1,2,\ldots,n$. Thus Z_j is distributed according to a multinomial distribution consisting of one draw on k categories with probabilities $\pi_1, \pi_2, \ldots, \pi_k$; that is $P(Z_j = z_j) = \pi_1^{z_{1j}}, \pi_2^{z_{2j}}, \ldots, \pi_k^{kj}$. This can be written as $Z_1, \ldots, Z_n \overset{i.i.d}{\sim}$ Mult$_k(1, \pi)$.

Suppose the density of an observation y_i given $Z_{ij} = 1$ is $f_i(y_j)$ and the unconditional density of y_j is $f(y_j)$. Since the i^{th} mixing probabilities π_i can be viewed as the prior probability that the entity belongs to the i^{th} component of the mixture, the posterior probability that the entity belongs to the i^{th} component with y_j having been observed on it, can be written as

$$P(Z_{ij} = 1 \mid y_j) = \frac{\pi_i f_i(y_j)}{f(y_j)} \tag{4.9}$$

Therefore the complete data vector can be written as $y_c = (y^{'}, z^{'})^{'}$. Thus the complete data likelihood function for Θ can be expressed as

$$L_c(\Theta) = \prod_{i=1}^{k} \prod_{j=1}^{n} \left[f_i(y_j; \theta_i) \right]^{z_{ij}} \pi_i^{z_{ij}} \tag{4.10}$$

From (4.10) the log-likelihood can be expressed as,

$$\log L_c(\Theta) = \sum_{i=1}^{k}\sum_{j=1}^{n} z_{ij} \log \pi_i f_i(y_j;\theta_i)$$

$$= \sum_{i=1}^{k}\sum_{j=1}^{n} z_{ij} \left\{\log \pi_i + \log f_i(y_j;\theta_i)\right\} \qquad (4.11)$$

The EM algorithm is employed in solving this problem by treating z_{ij} as unobservable or missing data. It proceeds in two iterative steps, E (for expectation) and M (for maximization). The addition of the unobservable data to the problem is handled by E-step, which takes the conditional expectation of the complete data log-likelihood, $\log L_c(\Theta)$, given the observed data y, using the current fit for Θ.

To be specific, on the g^{th} iteration, the E-step requires the calculation of the conditional expectation of $\log L_c(\Theta)$ given y, which is given by

$$Q(\Theta;\Theta^{(g)}) = E_{\Theta^{(g)}}\left[\log L_c(\Theta \mid y)\right] \qquad (4.12)$$

The expectation operator E has the subscript $\Theta^{(g)}$ to explicitly convey that this expectation is being applied using $\Theta^{(g)}$ for Θ, where $\Theta^{(g)}$ is the current fit of the parameters from the previous M-step. As the complete-data log-likelihood is linear in the unobservable data z_{ij}, the E-step simply requires to calculate the current conditional expectation of Z_{ij} given the observed data y,

$$E_{\Theta^{(g)}}(Z_{ij} \mid y) = P_{\Theta^{(g)}}\left\{z_{ij} = 1 \mid y\right\} = \tau_i(y_j;)\Theta^{(g)} \qquad (4.13)$$

from (4.9),

$$\tau_i(y_j;\Theta^{(g)}) = \frac{\pi_i^{(g)} f_i(y_j;\theta_i^{(g)})}{f(y_j;\Theta^{(g)})}$$

$$= \frac{\pi_i^{(g)} f_i(y_j;\theta_i^{(g)})}{\sum_{i=1}^{k} \pi_i^{(g)} f_i(y_j;\Theta_i^{(g)})} \tag{4.14}$$

for $i = 1,2,\ldots,k$; $j = 1,2,\ldots,n$. The quantity $\tau_i(y_j;\Theta^{(g)})$ is the posterior probability that the j^{th} member of the sample with observed value y_j belongs to the i^{th} component of the mixture. Using (4.14) the conditional expectation of the complete-data log-likelihood given the observed data *y* is

$$Q(\Theta;\Theta^{(g)}) = \sum\sum \tau_i(y_j;\Theta^{(g)})\left[\log\pi_i + \log f_i(y_j;\theta_i)\right] \tag{4.15}$$

The M-step on the *(g+1)*th iteration requires the global maximization of (4.15) with respect to Θ, to give the updated estimate $\Theta^{(g+1)}$. For the mixture model, the updated estimates $\pi_i^{(g+1)}$ of the mixing probabilities π_i 's are calculated independently of the updated estimates $\theta_i^{(g+1)}$ of the parameter vector θ_i containing the unknown parameters in the i^{th} component density.

If the z_{ij} were observable, then the complete data maximum likelihood estimate of π_i would be given by

$$\hat{\pi}_i = \sum_{j=1}^{n} \frac{z_{ij}}{n} \text{ for } (i = 1,2,\ldots,k) \tag{4.16}$$

As the E-step simply involves replacing each z_{ij} with its current conditional expectation $\tau_i(y_j;\Theta^{(g)})$ in the complete-data log-likelihood, the updated estimate of π_i is given by replacing each z_{ij} in (4.16) by $\tau_i(y_j;\Theta^{(g)})$ to give

$$\pi_i^{(g+1)} = \frac{\sum_{j=1}^{n} \tau_i(y_j;\Theta^{(g)})}{n} \text{ for } i = (1,2,\ldots,k) \tag{4.17}$$

Thus in forming the estimate of π_i on the *(g+1)*th iteration, there is a contribution from each observation y_j equal to its posterior probability of membership of the *i*th component of the mixture model.

Next, we can estimate the model parameters θ_i by maximizing the log likelihood (4.15) with respect to the model parameters θ_i .

The EM algorithm iterates between the E-step and the M-step until the difference

$$L(\Theta^{(g+1)}) - L(\Theta^{(g)}) \tag{4.18}$$

changes by an arbitrarily small amount in the case of convergence of the sequence of likelihood values $\{L(\Theta^{(g)})\}$ (McLachlan & Krishnan, 2008).

4.2.3 EM for Parametric Survival Mixture Model

EM is frequently been employed in solving complicated maximum likelihood estimation problem in mixture model. The presence of censored observations in survival analysis makes it more complicated. Therefore, there is a need to make arrangement to accommodate censoring problem when applying the EM to the mixture model estimation in survival analysis. In this derivation, random censoring procedure was considered. In this type of censoring, it is assumed that each object or individual has survival time *T* and censoring time *C*, where *T* and *C* are independent continuous random variables. The survival function corresponding to both *T* and *C* are *S(t)* and *G(t)* respectively. The survival time and the censoring time are assumed to be independent. The survival time t_i of each individual is given by t_j =min(T_j,C_j) and δ_j = 1 if $T_j \le C_j$ and δ_i =0 if $T_j > C_j$. then the data from observations on *n* individuals is assumed to consist of the pair (t_j, δ_j) for *j=1,…..,n*. Given the probability density function and the survival function of the survival time and censoring time, the likelihood function can be expressed as

$$L = \prod_{j=1}^{n} f(t_j)^{\delta_j} S(t_j)^{1-\delta_j} \tag{4.19}$$

For implementing the EM on survival data with censored observations, let $T_1, T_2, \ldots, T_n$ be *n* independent random variables, where T_j is the survival time of the j^{th} subject. We assume that the probability density function $f(t)$ and survival function $S(t)$ of T_j are defined in (4.3) and (4.5).

Let $(t_1,\delta_1),(t_2,\delta_2),\ldots,(t_n,\delta_n)$ be n pairs of survival time and censoring indicator observed. The likelihood function is expressed as

$$L(\Theta)=\prod_{j=1}^{n}\left[\sum_{i=1}^{k}\pi_i f_i^{\delta_j}(t_j;\theta_i)S_i^{1-\delta_j}(t_j;\theta_i)\right] \tag{4.20}$$

The unobservable vector Z can be defined as in (4.8) where Z_j is a variable that indicates whether T_j came or did not come from the i^{th} component of the mixture. The observable vector (t_j,δ_j) together with the unobservable z_j are considered as the complete data. The likelihood function for the complete-data is expressed as

$$L_c(\Theta)=\prod_{i=1}^{k}\prod_{j=1}^{n}\left[f_i^{\delta_j}(t_j;\theta_i)S_i^{1-\delta_j}(t_j;\theta_i)\right]^{z_{ij}}\pi_i^{z_{ij}} \tag{4.21}$$

and the log-likelihood function can be written as,

$$\log L_c(\Theta)=\sum_{i=1}^{k}\sum_{j=1}^{n}z_{ij}\left[\log\pi_i+\delta_j\log f_i(t_i;\theta_i)+(1-\delta_j)S_i(t_j;\theta_i)\right] \tag{4.22}$$

Since the Gamma distribution was considered for Model 1, the derivation of the parameters of this mixture model was explained as follows.

Survival Mixture of the Gamma Distribution

The Gamma distribution is one of the frequently used probability distributions in modelling survival data. The Gamma mixture model is defined as

$$f(t) = \sum_{i=1}^{k} \pi_i f_i(t;\alpha_i,\beta_i) \tag{4.23}$$

where $f_i(t;\alpha_i,\beta_i)$ represent the density function of Gamma distribution as in Table 2.1 with unknown parameters α_i,β_i, with $\alpha_i > 0$, $\beta_i > 0$.

From (4.22), the log-likelihood function of the complete-data is

$$\log L_c(\alpha_i,\beta_i,\pi_i) = \sum_{i=1}^{k}\sum_{j=1}^{n} z_{ij} \left\{ \log \pi_i + \delta_j \log\left[\frac{1}{\beta_i \Gamma(\alpha_i)} \left(\frac{t_j}{\alpha_i} \right)^{\alpha_i - 1} e^{\frac{-t}{\alpha_i}} \right] \right.$$

$$\left. + (1-\delta_j)\log\left[\frac{\Gamma(\alpha_i, t_j/\beta_i)}{\Gamma \alpha_i} \right] \right\} \tag{4.24}$$

The EM algorithm starts with the E-step. After the g^{th} iteration, $z_{ij}^{(g)}$ is the conditional expectation of z_{ij} given the observed data, as defined in (4.13) and (4.14). Then the current conditional expectation of the complete-data log-likelihood is given by

$$Q(\alpha_i,\beta_i,\pi_i) = \sum_{i=1}^{k}\sum_{j=1}^{n} z_{ij}^{(g)} \left\{ \log \pi_i + \delta_j \log\left[\frac{1}{\beta_i \Gamma(\alpha_i)} \left(\frac{t_j}{\alpha_i} \right)^{\alpha_i - 1} e^{\frac{-t}{\alpha_i}} \right] \right.$$

$$\left. + (1-\delta_j)\log\left[\frac{\Gamma(\alpha_i, t_j/\beta_i)}{\Gamma \alpha_i} \right] \right\} \tag{4.25}$$

The M-step on the *(g+1)*th iteration requires the global maximization of (4.25) with respect to α_i, β_i and π_i. The mixing probabilities π_i can be updated by $\hat{\pi}_i^{(g+1)} = \sum_{j=1}^{n} z_{ij}^{(g)} / n,\ \boldsymbol{i = 1, \ldots, k}$. In order to get the updated maximum likelihood estimate of the component model parameters α_i, β_i, the partial differentiation of equation (4.25) was taken with respect to the parameters the α_i, β_i, thus

$$\frac{\partial Q}{\partial \alpha_i} = \sum_{j=1}^{n} z_{ij}^{(g)} \delta_j \left[-\log \beta_i + \Psi(\alpha_i) + \log t_j \right]$$

$$+ \sum_{j=1}^{n} z_{ij}^{(g)} (1-\delta_j) \left[\log \beta_i + \frac{1}{\Gamma(\alpha_i, t_j / \beta_i)} \frac{\partial}{\partial \alpha_i} \Gamma(\alpha_i, t_j / \beta_i) \right] \quad (4.26)$$

$$\frac{\partial Q}{\partial \beta_i} = \sum_{j=1}^{n} z_{ij}^{(g)} \left[-\delta_j \left(\frac{\alpha_i}{\beta_i} + \frac{t_i}{\beta_i^2} \right) + \frac{(1-\delta_j)}{\Gamma(\alpha_i, t_j / \beta_i)} \frac{\partial}{\partial \beta_i} \Gamma(\alpha_i, t_j / \beta_i) \right] \quad (4.27)$$

Now, the upper incomplete gamma function can be differentiated with respect to β_i using Leibnitz's rule, and we then obtain from (4.27) that

$$\beta_i = \left[\sum_{j=1}^{n} z_{ij}^{g} t_j / \alpha_i + \sum_{j=1}^{n} z_{ij}^{g} \delta_j / \alpha_i - \sum_{j=1}^{n} \frac{t_j^{\alpha_i} e^{-t_j / \beta_i}}{\alpha_i \beta_i^{\alpha_i - 1} \Gamma\left(\alpha_i, t_j / \beta_i\right)} \right] \quad (4.28)$$

The RHS of (4.28) can be evaluated at the current parameter value to obtain the updated parameter estimate $\beta_i^{(g+1)}$.

Upon expanding the incomplete gamma function as an infinite series, then differentiating and simplifying the expression, (4.26) can be expressed as

$$\frac{\partial Q}{\partial \alpha_i}=\sum_{j=1}^{n} z_{ij}^{g}\partial_j\left[\log t_j-\log\beta_i-\Psi\left(\alpha_i\right)\right]$$

$$+\sum_{j=1}^{n} z_{ij}^{g}(1-\partial_j)\left[\log(t_j/\beta_i)-\log(t_j/\beta_i)/\left\{1-e^{-t_j/\beta_i}\sum_{p=0}^{\infty}\frac{(t_j/\beta_i)^{\alpha_i+p}}{\Gamma(\alpha_i+p+1)}\right\}\right.$$

$$\left.+e^{-t_j/\beta_i}\sum_{p=0}^{\infty}\frac{(t_j/\beta_i)^{\alpha_i+p}\Psi(\alpha_i+p+1)}{\Gamma(\alpha_i+p+1)}/\left\{1-e^{-t_j/\beta_i}\sum_{p=0}^{\infty}\frac{(t_j/\beta_i)^{\alpha_i+p}}{\Gamma(\alpha_i+p+1)}\right\}\right] \tag{4.29}$$

Equating (4.29) to zero, the equation can be solved iteratively for α_i to obtain the current estimate $\alpha_i^{(g+1)}$ by using $\beta_i^{(g+1)}$ for β_i .

The E-step on the *(g+1)*th iteration is to update the current conditional expectation of z_{ij}, given the observed data, using the current model parameters fit,

$$\hat{Z}_{ij}^{(g+1)}=\frac{\hat{\pi}_i^{(g)}\left[f_i(t_j;\hat{\alpha}_i^{(g)},\hat{\beta}_i^{(g)}\right]^{\delta_j}\left[S_i(t_j;\hat{\alpha}_i^{(g)},\hat{\beta}_i^{(g)}\right]^{1-\delta_j}}{\sum_{i=1}^{k}\hat{\pi}_i^{(g)}\left[f_i(t_j;\hat{\alpha}_i^{(g)},\hat{\beta}_i^{(g)}\right]^{\delta_j}\left[S_i(t_j;\hat{\alpha}_i^{(g)},\hat{\beta}_i^{(g)}\right]^{1-\delta_j}} \tag{4.30}$$

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

The EM procedure outlined was employed on both of the simulated and real data to evaluate the parameters of Model 1. These derivations and equations were transformed into computer coding. The R statistical software (Team, 2005) was employed to develop functions to generate survival data used for the simulated data.

EM functions were developed to estimate the parameters of the Model 1. The R functions regarding the estimation were presented in Appendix C. All the computations were performed using R statistical software version 3.0.2 (2013-09-25). Also, among the R statistical software packages used are package *survival* developed by Therneau (1999), package *Mixtools* developed by Young, et al. (2007), and package *Smcure* developed by Cai, et al. (2012).

4.3 Validation of the performance of Model 1 Using Simulated Data

In this section two sets of mixing probabilities were considered for the validation of Model 1 using simulated data. The two sets were arranged in ascending order (10%, 20% and 40%) and descending order (50%, 30% and 20%). The performance of Model 1 (G1_G2_G3) was validated by simulating data with three different sample sizes (100, 200 and 500) and three different censoring percentages.

4.3.1 Model 1 with Mixing Probabilities in Ascending Order

In this section survival data for Model 1 were generated based on mixture model of three well separated components of Gamma distribution. The parameters of the first component Gamma distribution (G1) are ($\alpha_1 = 40,\ \beta_1 = 20$) respectively, the parameters for the second component Gamma distribution (G2) are ($\alpha_2 = 6\ ,\beta_2 = 1$) and the parameters of the third component Gamma distribution (G3) are ($\alpha_3 = 200, \beta_3 = 20$). Wiper, Insua and Ruggeeri (2001) employed the Bayesian estimation method to analyse the mixture model of Gamma distribution with those parameters. Based on these three components Gamma distribution, survival data were generated for the three different sample sizes (100, 200 and 500) each with three different censoring percentages (10%, 20% and 40%). The mixing probabilities employed were in the ascending order (10%, 40% and 50%). Three sets of survival data of sample size of 100 observations each. The same samples size were generated from the Exponential distribution for the censored time C with (b), where the value of b depends solely of the percentage of the observations that are censored. In this study 10%, 20% and 40% censoring observations were considered for each of the sample generated. t_j =min(T_j,C_j) was taken as the minimum of the survival time and the censored time of the observed time T where

$$T = \begin{cases} \delta_i = 1, & if \;\; X \le C, \\ \delta_i = 0, & if \;\; X > C. \end{cases} \tag{4.33}$$

The postulated Model 1 can be formed by substituting the values of the parameters in equation (4.1), which is expressed as

$$f(t) = 0.1 * f_G(t;\alpha_1 = 40, \beta_1 = 20) + 0.4 * f_G(t;\alpha_2 = 6, \beta_2 = 1) + 0.5 * f_G(t;\alpha_3 = 200, \beta_3 = 20) \tag{4.34}$$

where the density function f_G represents the Gamma distribution probability density functions corresponding to each component of Model 1.

4.3.1.1 Sample of Size 100 Observations

The simulated data for samples of size 100 observations with 10%, 20% and 40% censored observations were used to estimate the parameters of the postulated Model 1 by employing the EM. The estimates of the parameters together with the parameters of the postulated models were reported. Figures 4.1, 4.2 and 4.3 display the probability density function of simulated data of Model 1, with 100 observations and 10%, 20% 40% censoring percentages respectively, and the probability density functions of pure classical parametric survival models (G1, G2 and G3) corresponding to each component of Model 1.

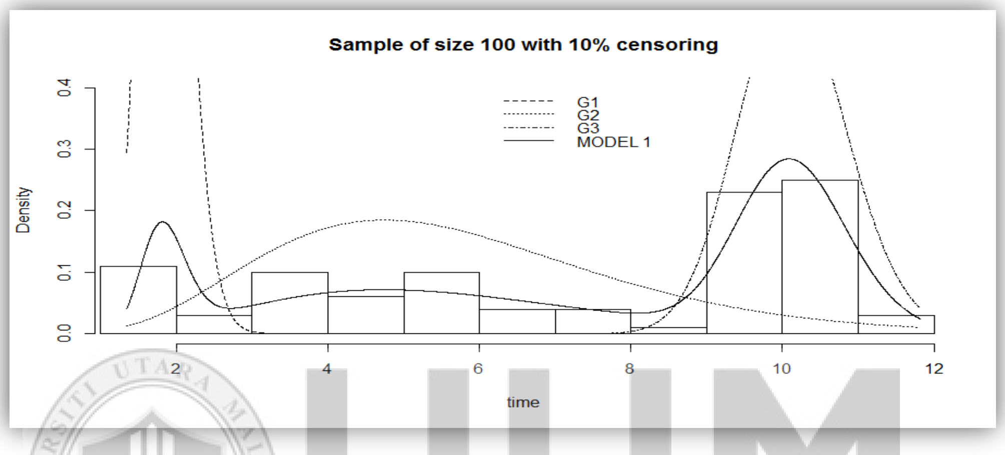


Figure 4.1 Probability Density Function of the Simulated Data of Model 1 with 100 Observations and 10% Censoring.

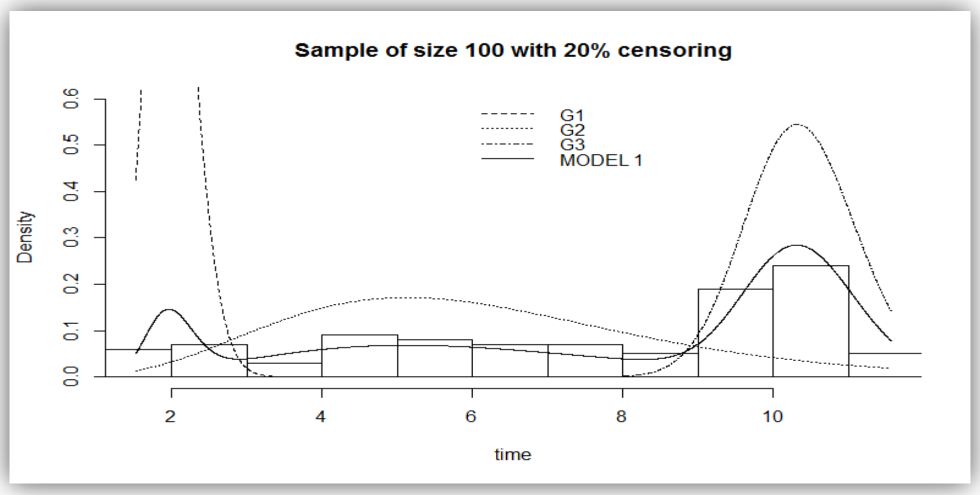


Figure 4.2 Probability Density Function of the Simulated Data of Model 1 with 100 Observations and 20% Censoring.

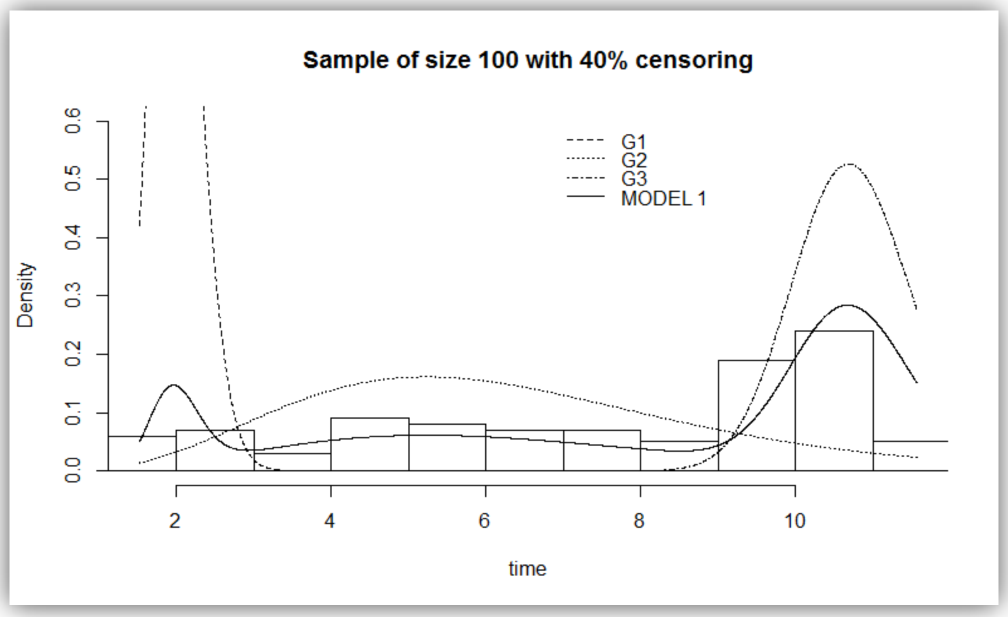


Figure 4.3 Probability Density Function of the Simulated Data of Model 1 with 100 Observations and 40% Censoring.

It can be seen that Model 1 fits the simulated data better than the individual pure classical parametric survival models which indicates that the simulated data is better modelled by Model 1 than the pure classical parametric survival model.

The estimated parameters of the sample of size 100 with 10% 20% and 40% censoring percentages are displayed in Table 4.1.

Table 4.1

The Estimated Parameters the Simulated Data of Postulated Model 1 with 10%, 20% and 40% Censoring Observations

Model 1 with sample size 100 observations and 10% censoring

Parameter	π_1	π_2	α_1	α_2	α_3	β_1	β_2	β_3
Postulate	0.10	0.40	40	6	200	20	1	20
Estimates	0.12	0.39	40.00	5.99	200.00	21.60	1.05	19.71

Model 1 with sample size 100 observations and 20% censoring

Parameter	π_1	π_2	α_1	α_2	α_3	β_1	β_2	β_3
Postulate	0.10	0.40	40	6	200	20	1	20
Estimates	0.10	0.40	40.00	6.01	200.00	20.03	0.98	19.27

Model 1 with sample size 100 observations and 40% censoring

Parameter	π_1	π_2	α_1	α_2	α_3	β_1	β_2	β_3
Postulate	0.10	0.40	40	6	200	20	1	20
Estimates	0.11	0.38	40.05	5.64	199.99	20.04	0.89	18.62

The estimated parameters of the three set of the simulated data are all close to the postulated parameters used in the data generation. From Table 4.1 it can be observed that the parameter for the simulated set of data with 10% and 20% censored observations are closer to the true parameters compared to that of the 40% censored observations.

The hazard functions of the three simulated data corresponding to the 10%, 20% and 40% censoring were presented in Figure 4.4.

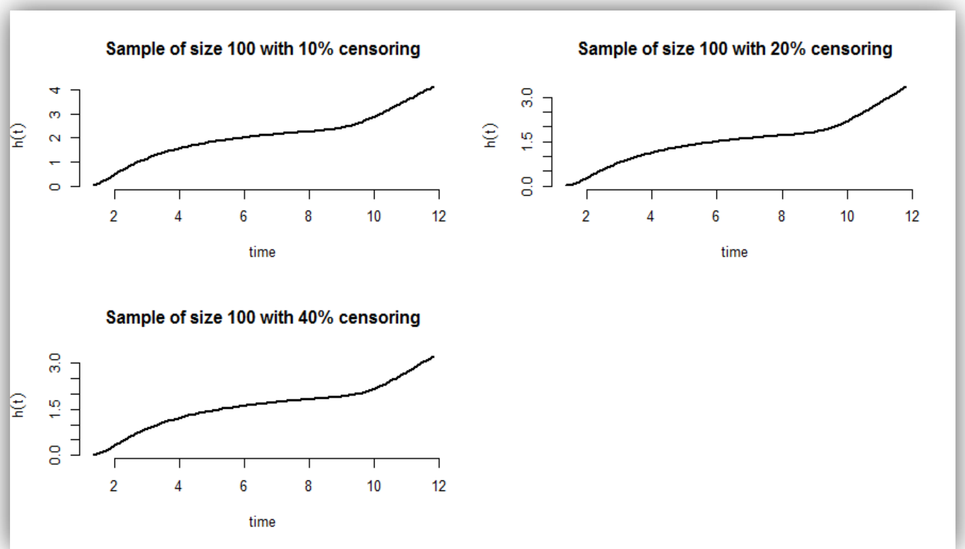


Figure 4.4 The Hazard Functions of the Simulated Data of Model 1 Corresponding to 10%, 20% and 40% Censored Observation.

The hazard function of simulated data of size 100 observations with 10% of the observations censored is higher than that of 20% and 40% censoring. As the number of censored observations increases the hazard tends to be lower and lower.

The simulation of the three sets of generated data were repeated 300 times to check the consistency and stability of the EM in estimating the model parameters. The averages, the mean square errors and root mean square error of estimated parameters of the postulated Model 1 are listed in Table 4.2.

Table 4.2

The Repeated Simulation of Set of 100 Observations

Model 1 with sample size 100 and 10% censoring								
parameters	π_1	π_2	α_1	α_2	α_3	β_1	β_2	β_3
Postulates	0.10	0.40	40	6	200	20	1	20
Estimates	0.10	0.39	41.51	5.60	197.88	20.76	0.91	19.62
MSE	6.00e-07	1.95e-06	7.35e-02	2.65e-4	3.46e-02	1.82e-02	6.82e-05	3.87e-4
RMSE	8.00e-4	0.0014	0.2711	0.0162	0.1860	0.1349	0.0083	1.69e-2
Model 1 with sample size 100 and 20% censoring								
parameters	π_1	π_2	α_1	α_2	α_3	β_1	β_2	β_3
postulated	0.10	0.40	40	6	200	20	1	20
Estimates	0.10	0.38	42.74	5.76	198.88	21.35	0.95	19.72
MSE	8.11e-7	3.85e-5	5.72e-1	2.33e-3	3.11e-1	1.35e-1	7.67e-5	3.22e-3
RMSE	9.01e-4	1.96e-3	7.56e-1	4.83e-2	5.58e-1	3.67e-1	8.75e-3	5.67e-2
Model 1 with sample size 100 and 40% censoring								
Parameters	π_1	π_2	α_1	α_2	α_3	β_1	β_2	β_3
Postulated	0.10	0.40	40	6	200	20	1	20
Estimates	0.09	0.36	43.02	5.31	196.96	21.63	0.85	19.37
MSE	8.42e-7	3.91e-6	7.53e-1	2.84e-3	3.31e-1	2.11e-1	9.22e-5	3.32e-3
RMSE	9.18e-4	1.98e-3	8.68e-1	5.33e-2	5.75e-1	4.59e-1	9.60e-3	5.76e-2

The averages of the parameters are close to the parameters of the postulated parametric survival mixture model with mean square errors and root mean square relatively small, which suggests that, the EM performed consistently in estimating the parameters. The MSE corresponding to the mixing probabilities are relatively smaller for the 10% censoring as compared to the 20% and 40% censoring.

4.3.1.2 Sample of Size 200 Observations

Three sets of survival data of size 200 observations with 10%, 20% and 40% censored observations, respectively, were generated and employed to estimate the parameters of the postulated Model 1 by the EM. The estimated parameters corresponding to each set of data and the true parameters of the postulated models were reported. The probability density function of simulated data of size 200 observations and 10%, 20% and 40% censored observations were presented in Figures 4.5, 4.6 and 4.7 respectively. The graphs also display the probability density functions of pure classical parametric survival models (G1, G2 and G3) corresponding to each component of Model 1.

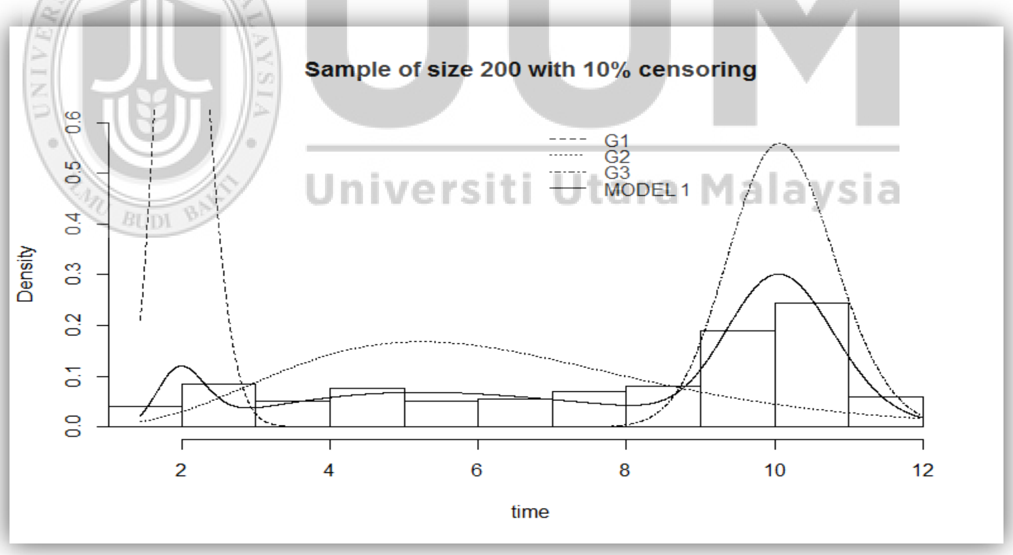


Figure 4.5 Probability Density Function of the Simulated Data of Size 200 Observations and 10% Censoring

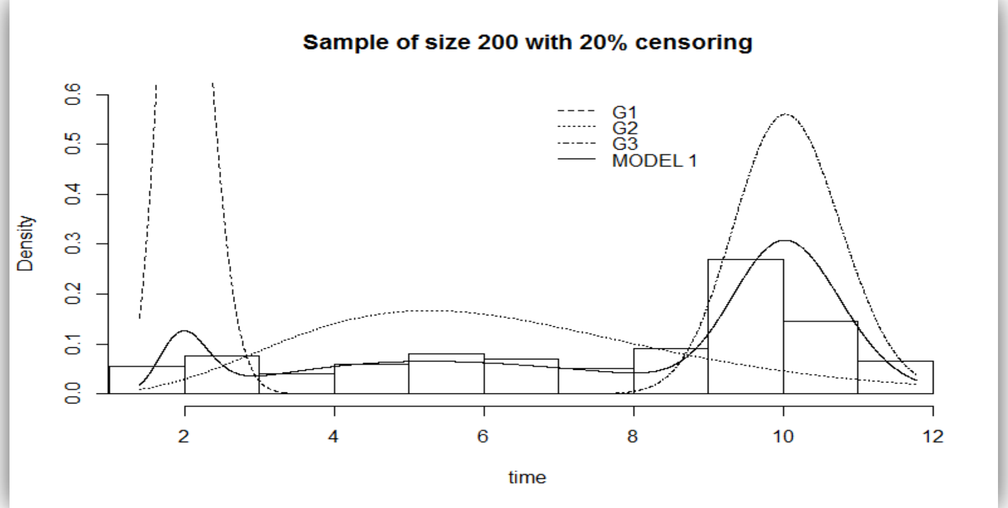


Figure 4.6 Probability Density Function of the Simulated Data of Size 200 and 20% Censored Observations.

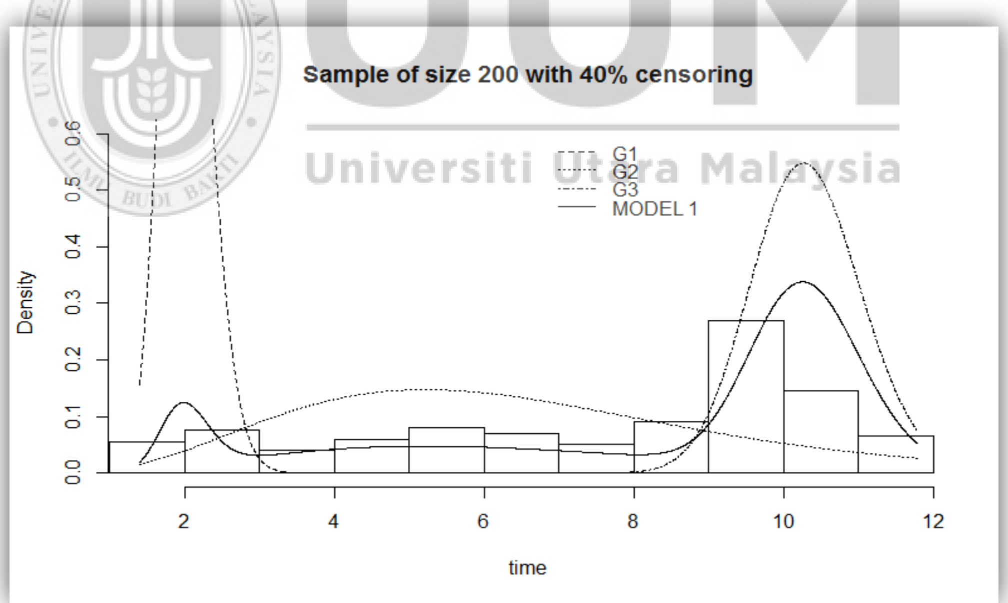


Figure 4.7 Probability Density Function of the Simulated Data of Size 200 Observations and 40% Censoring.

It can be seen that Model 1 fits the simulated data better than the individual pure classical parametric survival models which indicates that the simulated data is better modelled by Model 1 than the pure classical parametric survival model.

The estimated parameters of the set of simulated data of size 200 with 10%, 20% and 40% censored observations were presented in Table 4.3 together with true parameters of Model 1.

Table 4.3

The Estimated Parameters the Simulated Data of size 200 with 10% Censoring Observations

Model 1 with sample size 200 observations and 10% censoring								
Parameter	π_1	π_2	α_1	α_2	α_3	β_1	β_2	β_3
Postulate	0.10	0.40	40	6	200	20	1	20
Estimates	0.09	0.41	40.00	6.01	200.00	19.72	0.96	19.77
Model 1 with sample size 200 observations and 20% censoring								
Parameter	π_1	π_2	α_1	α_2	α_3	β_1	β_2	β_3
Postulate	0.10	0.40	40	6	200	20	1	20
Estimates	0.09	0.39	40.00	6.01	200.00	19.70	0.95	19.83
Model 1 with sample size 200 observations and 40% censoring								
Parameter	π_1	π_2	α_1	α_2	α_3	β_1	β_2	β_3
Postulate	0.10	0.40	40	6	200	20	1	20
Estimates	0.11	0.38	40.05	5.64	199.99	20.04	0.89	18.62

The estimated parameters of the three set of the simulated data are all close to the postulated parameters used in the data generation. From Table 4.3, it can be observed that the parameter for the simulated set of data with 10% censoring are more closer the true parameters compared to that of the 20% and 40% censoring observations.

The estimation of the mixing probabilities was more accurate in sample with 10% censoring. It can be observed that the estimation of the mixing probabilities is better compared to that of the sample with 100 observations.

The hazard functions of the three simulated data of size 200 observations corresponding to the 10%, 20% and 40% censoring percentages were presented in Figure 4.8.

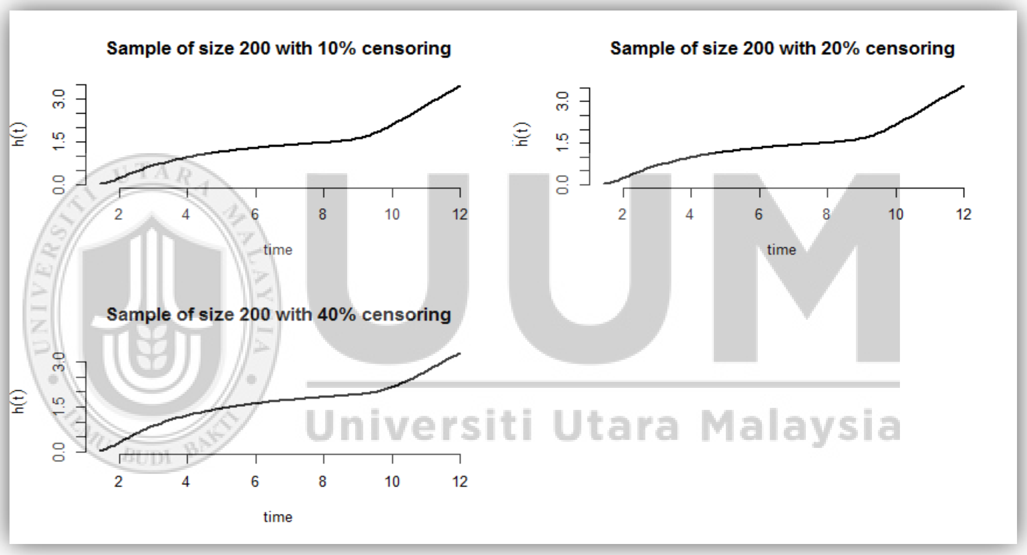


Figure 4.8 The Hazard Functions of the Simulated Data of Size 200 Corresponding to 10%, 20% and 40% Censored Observation.

The hazard function of simulated data of size 200 observations with 10% of the observations censored is higher than that of 20% and 40% censoring.

The simulation of the three set of generated data of 200 observations with 10%, 20% and 40% censoring observations were repeated 300 times to check the consistency and stability of the EM in estimating the model parameters. The averages, the mean

square errors and root mean square error of estimated parameters of the postulated Model 1 are listed in Table 4.4.

Table 4.4

The Repeated Simulation of Set of 200 Observations

Model 1 with sample size 200 and 10% censoring								
Parameters	π_1	π_2	α_1	α_2	α_3	β_1	β_2	β_3
Postulates	0.10	0.40	40	6	200	20	1	20
Estimates	0.10	0.38	42.33	5.60	198.19	21.16	0.91	19.67
MSE	3.89e-7	1.86e-7	2.62e-1	1.06e-3	2.44e-1	6.28e-2	3.88e-5	2.59e-3
RMSE	6.24e-4	1.36e-3	5.13e-1	3.26e-2	4.94e-1	2.51e-1	6.23e-3	5.08e-2
Model 1 with sample size 200 and 20% censoring								
parameters	π_1	π_2	α_1	α_2	α_3	β_1	β_2	β_3
postulated	0.10	0.40	40	6	200	20	1	20
Estimates	0.09	0.37	42.21	5.25	195.44	21.10	0.84	19.26
MSE	3.8e-6	2.2e-06	2.82e-1	2.00e-3	4.55e-1	6.9e-2	5.4e-05	4.6e-3
RMSE	6.5e-4	1.40e-1	5.31e-1	4.0e-2	6.71e-1	2.64e-1	7.00e-3	6.8e-2
Model 1 with sample size 200 and 40% censoring								
Parameters	π_1	π_2	α_1	α_2	α_3	β_1	β_2	β_3
Postulated	0.10	0.40	40	6	200	20	1	20
Estimates	0.09	0.33	42.22	4.46	194.26	21.16	0.68	18.80
MSE	5.19e-7	2.42e-6	3.67e-1	1.64e-3	6.64e-1	8.91e-2	5.76e-5	6.44e-3
RMSE	7.21e-4	1.56e-3	6.05e-1	4.05e-2	8.15e-1	2.30e-1	7.59e-3	8.02e-2

The averages of the parameters are close to the true values of the parameters of parametric survival mixture model with mean square errors relatively small, which suggests that, the EM performed consistently in estimating the parameters. For the small censoring percentage (10%) the values of the MSE tend to be smaller than that of the 20% and 40% censoring percentages. Also, it can be observed that, the mixing probabilities tend to be closer to the true value with the smaller censoring percentage (10%).

4.3.1.3 Sample of Size 500 observations

Three sets of survival data of size 500 observations with 10%, 20% and 40% censored observations, respectively, were generated and employed to estimate the parameters of the postulated Model 1 by using the EM. The estimated parameters of corresponding to each set of data and the parameters of the postulated models were reported. The probability density function of simulated data of size 500 with 10%, 20% and 40% censored observations were presented in Figures 4.9, 4.10 and 4.11 respectively. The graph also, displays the probability density functions of pure classical parametric survival models (G1, G2 and G3) corresponding to each component of Model 1.

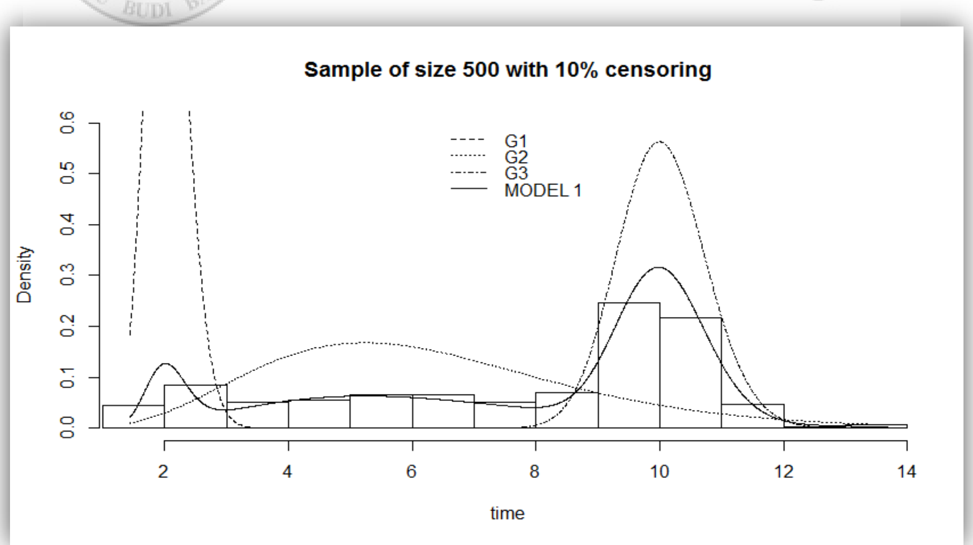


Figure 4.9 Probability Density Function of the Simulated Data of Size 500 Observations and 10% Censoring.

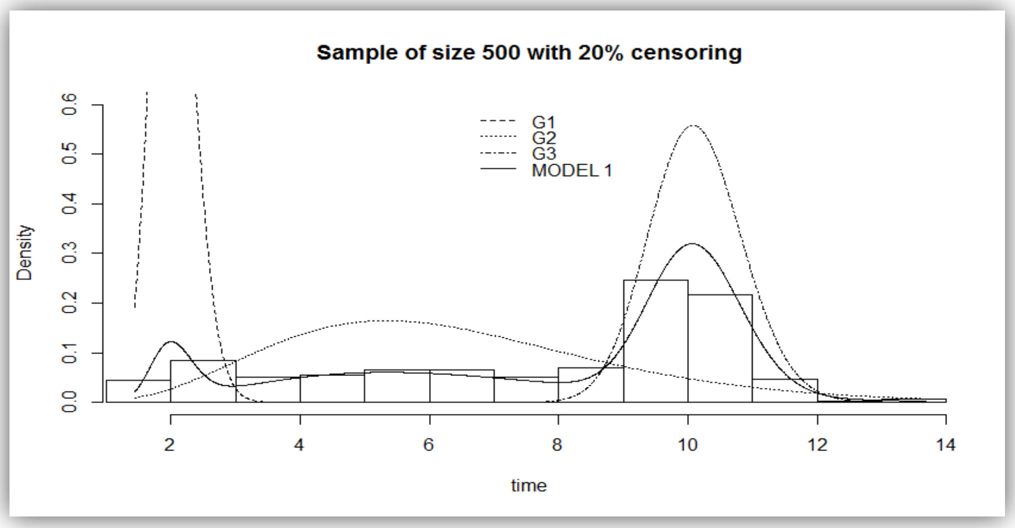


Figure 4.10 Probability Density Function of the Simulated Data of Size 500 Observations and 20% Censoring.

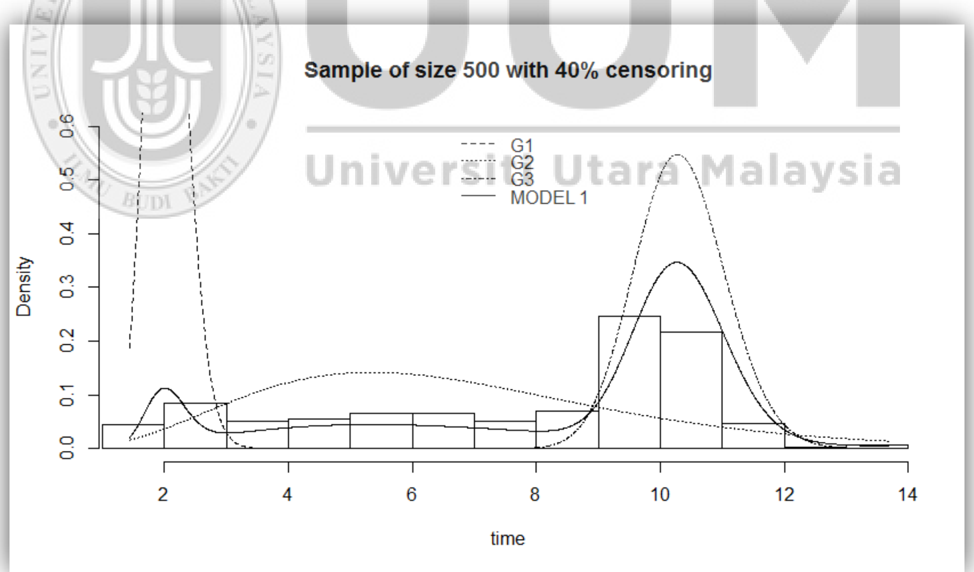


Figure 4.11 Probability Density Function of the Simulated Data of Size 500 Observations and 40% Censoring.

It can be seen that Model 1 fits the simulated data better than the individual pure classical parametric survival models which indicates that the simulated data is better modelled by Model 1 than the pure classical parametric survival model.

The estimated parameters of the set of simulated data of size 200 with 10%, 20% and 40% censored observations were presented in Table 4.5. The estimated parameters are close to the values of the parameters of the postulated model. It can be observed that the parameter for the simulated set of data with 10% censoring are more closer the postulated parameters compared to that of the 20% and 40% censoring observations.

Table 4.5

The Estimated Parameters the Simulated Data of size 500 with 10% Censoring Observations

Model 1 with sample size 500 observations and 10% censoring								
Parameter	π_1	π_2	α_1	α_2	α_3	β_1	β_2	β_3
Postulate	0.10	0.40	40	6	200	20	1	20
Estimates	0.09	0.38	40.00	6.00	200.00	19.52	0.96	19.91
Model 1 with sample size 500 observations and 20% censoring								
Parameter	π_1	π_2	α_1	α_2	α_3	β_1	β_2	β_3
Postulate	0.10	0.40	40	6	200	20	1	20
Estimates	0.09	0.37	40.00	6.01	200.00	19.60	0.94	19.74
Model 1 with sample size 500 observations and 40% censoring								
Parameter	π_1	π_2	α_1	α_2	α_3	β_1	β_2	β_3
Postulate	0.10	0.40	40	6	200	20	1	20
Estimates	0.08	0.32	40.06	4.79	199.96	19.59	0.71	19.36

The hazard functions of the three simulated data of size 500 observations corresponding to the 10%, 20% and 40% censoring percentages were presented in Figure 4.12.

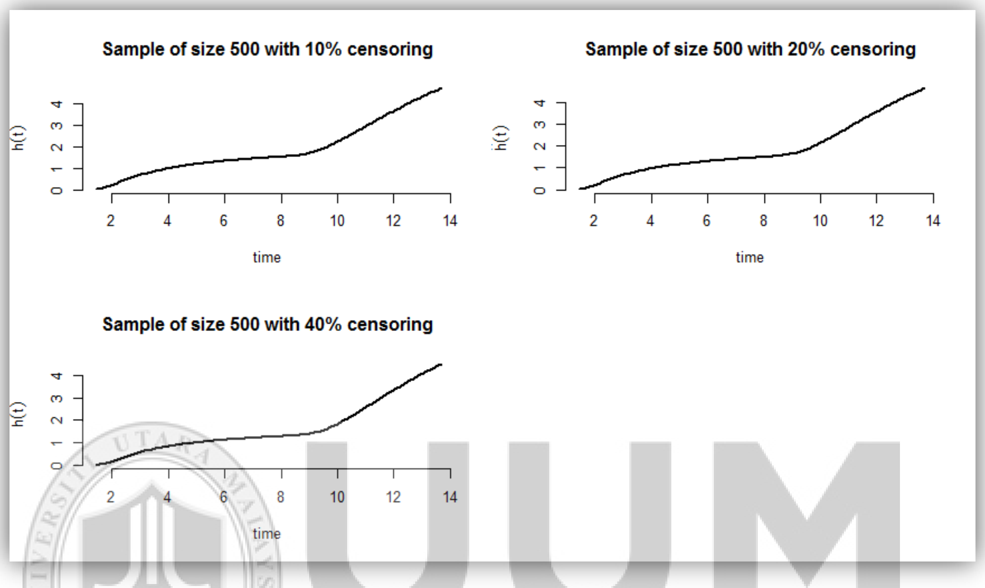


Figure 4.12 The Hazard Functions of the Simulated Data of Size 500 Corresponding to 10%, 20% and 40% Censored Observation.

The hazard function of simulated data of size 500 observations with 10% of the observations censored is higher than that of 20% and 40% censoring. As the number of censored observations increases the hazard tends to be lower and lower.

The simulation of the three sets of generated data of 500 observations with 10%, 20% and 40% censoring observations were repeated 300 times to check the consistency and stability of the EM in estimating the model parameters. The averages, the mean square errors and root mean square error of estimated parameters of the postulated Model 1 are listed in Table 4.6.

Table 4.6

The Repeated Simulation of Set of 500 Observations

Model 1 with sample size 500 and 10% censoring								
parameters	π_1	π_2	α_1	α_2	α_3	β_1	β_2	β_3
Postulates	0.10	0.40	40	6	200	20	1	20
Estimates	0.10	0.38	40.59	5.61	198.89	20.27	0.92	19.75
MSE	1.44e-7	5.57e-7	7.40e-02	5.91e-4	9.01e-2	1.74e-2	2.05e-5	9.16e-4
RMSE	3.80e-4	7.46e-4	0.27298	0.02430	0.30013	0.13190	0.00453	0.03026
Model 1 with sample size 500 and 20% censoring								
parameters	π_1	π_2	α_1	α_2	α_3	β_1	β_2	β_3
postulated	0.10	0.40	40	6	200	20	1	20
Estimates	0.10	0.37	41.40	5.18	195.67	20.68	0.83	19.29
MSE	1.43e-7	6.23e-7	1.53e-1	6.32e-4	3.40e-1	3.63e-2	2.15e-5	3.37e-3
RMSE	3.79e-4	7.89e-4	3.91e-01	2.51e-2	5.83e-1	1.91e-1	4.64e-3	5.80e-2
Model 1 with sample size 500 and 40% censoring								
parameters	π_1	π_2	α_1	α_2	α_3	β_1	β_2	β_3
postulated	0.10	0.40	40	6	200	20	1	20
estimates	0.08	0.34	40.86	4.29	194.03	20.45	0.65	18.81
MSE	1.53e-7	8.23e-7	1.56e-1	6.23e-4	8.50e-1	3.85e-2	1.85e-4	8.12e-3
RMSE	3.91e-4	9.01e-4	3.94e-1	2.29e-2	9.22e-1	1.96e-01	1.36e-2	9.01e-2

The averages of the parameters are close to the parameters of the postulated parametric survival mixture model with MSE relatively small, which suggests that, the EM performed consistently in estimating the parameters. It can be observed that the estimation of the parameters were closer to the true parameters as the sample size increases from 100 to 500 observations for the three censoring percentages. Also, the estimation of the mixing probabilities were closer to the true values when the censoring percentages get smaller and smaller.

Generally, the MSE for the sample with 10% censoring are smaller than that of the samples with 20% and 40% which shows that the parameters are better estimated with smaller censoring percentages. The estimation of the mixing probabilities get distorted with the increase in the censoring percentages of the samples. Observing Tables 4.3, 4.4 and 4.5 the parameter estimation improve with the increase in the sample size of the simulated data.

4.3.2 Model 1 with Mixing Probabilities in Descending Order

Survival data for Model 1 were generated based on mixture model of three well separated components of Gamma distribution as described in section 4.4.1. In this section the mixing probabilities were arranged in descending order. The mixing probabilities employed were 50%, 30% and 20% for the first, second and third components respectively. The same parameters of the three components gamma distribution used earlier in section 4.4.1 were employed to generate survival data for the three different sample sizes (100, 200 and 500) each with three different censoring proportions (10%, 20% and 40% censored observations). The mixing probabilities employed were in the descending order. The postulated Model 1 was formed by substituting the values of the parameters as in equation (4.44).

4.3.2.1 Sample of Size 100 observations

Data of size 100 with 10%, 20% and 40% censored observations were generated and used to estimate the parameters of the postulated Model 1 by employing the EM. The estimates of the parameters together with the true parameters of the postulated models were reported.

Figures 4.13, 4.14 and 4.15 display the probability density function of simulated data of Model 1, with 100 observations and 10%, 20% and 40% censoring percentages. The probability density functions of pure classical parametric survival models (G1, G2 and G3) corresponding to each component of Model 1 were also presented in the same graph.

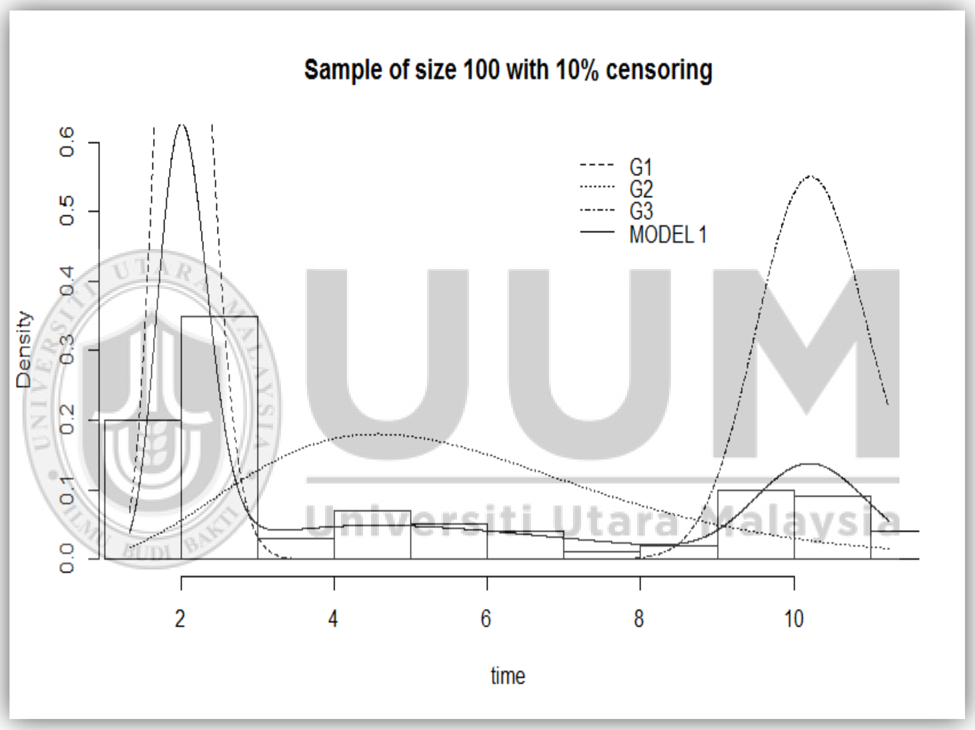


Figure 4.13 Probability Density Function of the Simulated Data of Model 1 with 100 Observations and 10% Censoring.

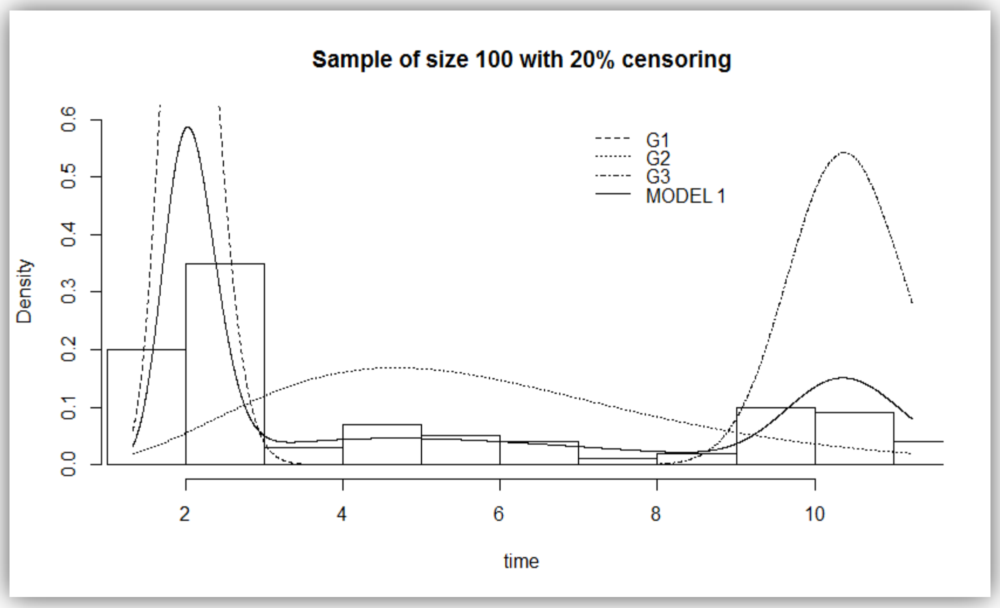


Figure 4.14 Probability Density Function of the Simulated Data of Model 1 with 100 Observations and 20% Censoring.

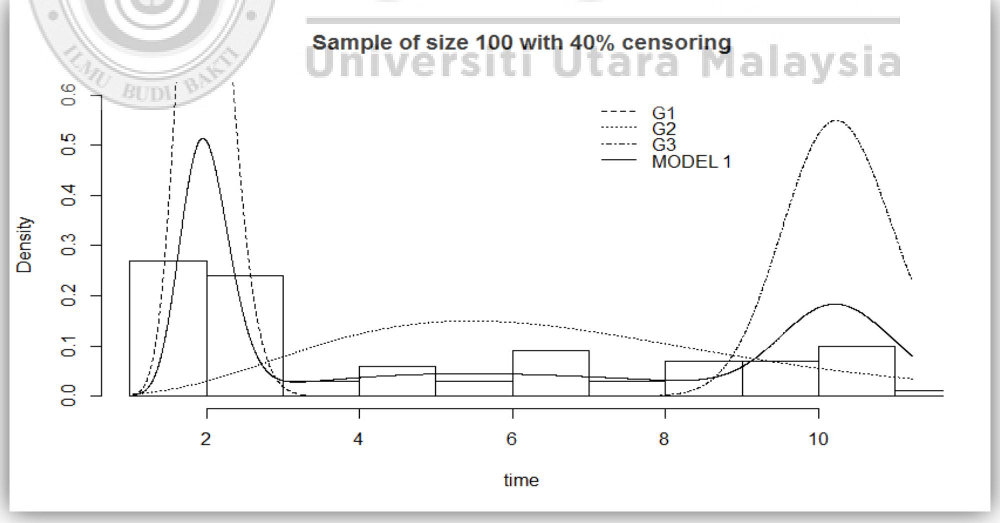


Figure 4.15 Probability Density Function of the Simulated Data of Model 1 with 100 Observations and 40% Censoring

It can be observed that Model 1 fits the simulated data better than the individual pure classical parametric survival models which indicate that the simulated data is better modelled by Model 1 than the pure classical parametric survival model.

The parameters of simulated data consisting of 100 observations with 10%, 20% and 40% censored observations were evaluated. Table 4.7 displays these estimates together with the true parameters used in generating the data.

Table 4.7

The Estimated Parameters the Simulated Data of Postulated Model 1 with 10% Censoring Observations

Model 1 with sample size 100 observations and 10% censoring								
Parameter	π_1	π_2	α_1	α_2	α_3	β_1	β_2	β_3
Postulate	0.50	0.30	40	6	200	20	1	20
Estimates	0.49	0.27	40.07	5.41	199.98	19.51	1.00	19.47
Model 1 with sample size 100 observations and 20% censoring								
Parameter	π_1	π_2	α_1	α_2	α_3	β_1	β_2	β_3
Postulate	0.50	0.30	40	6	200	20	1	20
Estimates	0.46	0.27	40.14	5.05	199.98	19.35	0.87	19.18
Model 1 with sample size 100 observations and 40% censoring								
Parameter	π_1	π_2	α_1	α_2	α_3	β_1	β_2	β_3
Postulate	0.50	0.30	40	6	200	20	1	20
Estimates	0.39	0.30	40.30	5.37	199.91	20.13	0.80	19.44

The estimated parameters of the three set of the simulated data are all close to the postulated parameters used in the data generation. From Table 4.7 it can be observed that the parameter for the simulated set of data with 10% censoring are more closer the postulated parameters compared to that of the 20% and 40% censoring observations. The mixing probabilities were estimated more accurately estimated with lower number of censored observations.

To investigate the effect of changing censoring percentage on the Model 1, the hazard functions of the three simulated data corresponding to the 10%, 20% and 40% censoring were evaluated and presented in Figure 4.16.

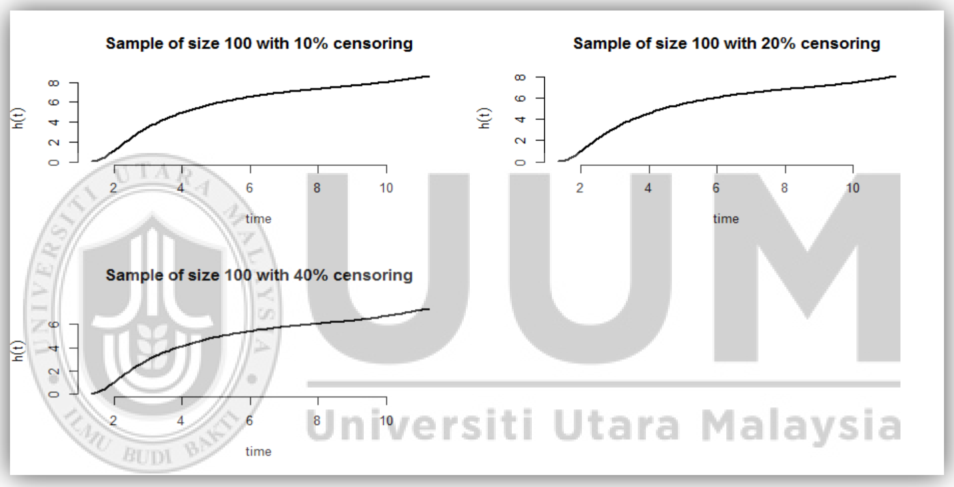


Figure 4.16 the Hazard Functions of the Simulated Data of Size 100 Corresponding to 10%, 20% and 40% Censored Observations.

The hazard function of simulated data of size 100 observations with 10% of the observations censored is higher than that of 20% and 40% censoring. As the number of censored observations increases the hazard tends to be lower and lower.

The simulation of the three set of generated data were repeated 300 time to check the consistency and stability of the EM in estimating the model parameters. The averages, the mean square errors and root mean square error of estimated parameters of the postulated Model 1 are listed in Table 4.8.

The averages of the parameters are close to the parameters of the postulated parametric survival mixture model with mean square errors and root mean square relatively small, which suggests that, the EM performed consistently in estimating the parameters.

Table 4.8

The Repeated Simulation of Set of 100 Observations

Model 1 with sample size 100 and 10% censoring								
Parameters	π_1	π_2	α_1	α_2	α_3	β_1	β_2	β_3
Postulates	0.50	0.30	40	6	200	20	1	20
Estimates	0.48	0.28	42.21	5.41	198.49	21.11	0.87	19.54
MSE	1.23e-6	3.31e-6	1.82e-1	3.10e-4	2.30e-1	4.55e-2	1.16e-4	2.56e-3
RMSE	1.11e-3	1.82e-3	4.26e-1	1.76e-2	4.79e-1	2.13e-1	1.08e-2	5.06e-2
Model 1 with sample size 100 and 20% censoring								
Parameters	π_1	π_2	α_1	α_2	α_3	β_1	β_2	β_3
Postulated	0.50	0.30	40	6	200	20	1	20
Estimates	0.46	0.27	42.58	4.80	197.07	21.29	0.75	19.11
MSE	1.73e-6	4.06e-6	2.12e-1	3.53e-3	6.12e-1	5.34e-2	1.37e-4	6.57e-3
RMSE	1.31e-3	2.01e-3	4.61e-1	5.95e-2	7.82e-1	2.31e-1	1.17e-2	8.10e-2
Model 1 with sample size 100 and 40% censoring								
Parameters	π_1	π_2	α_1	α_2	α_3	β_1	β_2	β_3
Postulated	0.50	0.30	40	6	200	20	1	20
Estimates	0.42	0.26	42.34	4.01	193.60	21.12	0.58	18.26
MSE	3.18e-6	5.59e-6	2.08e-1	2.90e-3	1.23e+00	5.25e-2	1.15e-3	1.20e-2
RMSE	1.78e-3	2.36e-3	4.56e-1	5.38e-2	1.11e+00	2.29e-1	3.39e-2	1.10e-1

The MSE corresponding to the mixing probabilities are relatively smaller for the 10% censoring as compared to the 40% censoring. Also the MSE for the parameters of the components are smaller for the 10% censoring compared to that of the 40% with exception of the third component. Generally, the estimation of the mixing probabilities and the parameters seemed to be closer to the true value with smaller censoring percentage 10% than with 40%.

4.3.2.2 Sample of Size 200 observations

Three sets of survival data of size 200 observations with 10%, 20% and 40% censored observations, respectively, were generated and employed to estimate the true parameters of the postulated Model 1 by using the EM. The estimated parameters of corresponding to each set of data and the true parameters of the postulated models were reported.

The probability density function of simulated data of size 200 observations with 10%, 20% and 40% censored observations was presented in Figures 4.17, 4.18 and 4.19 respectively. The graphs also, display the probability density functions of pure classical parametric survival models (G1, G2 and G3) corresponding to each component of Model 1.

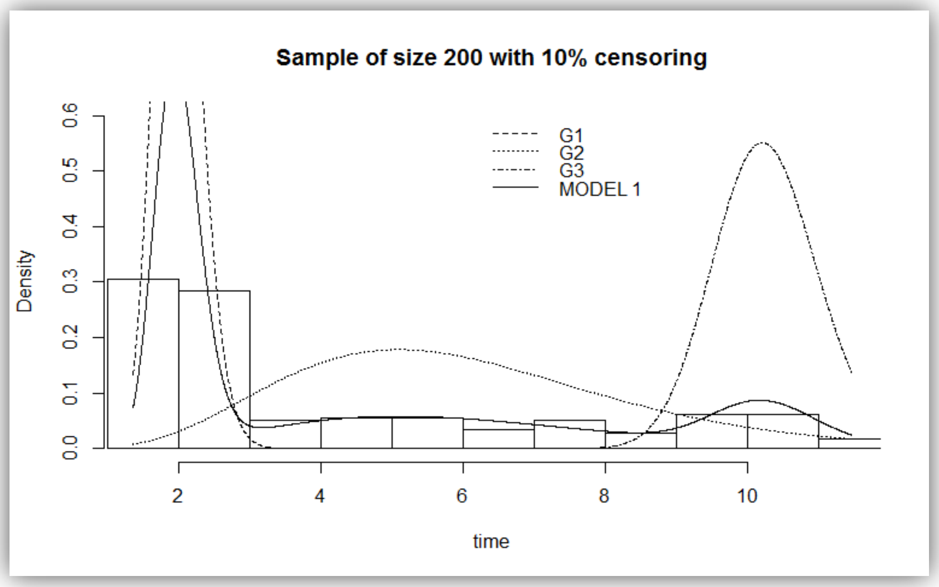


Figure 4.17 Probability Density Function of the Simulated Data of Size 200 Observations and 10% Censoring.

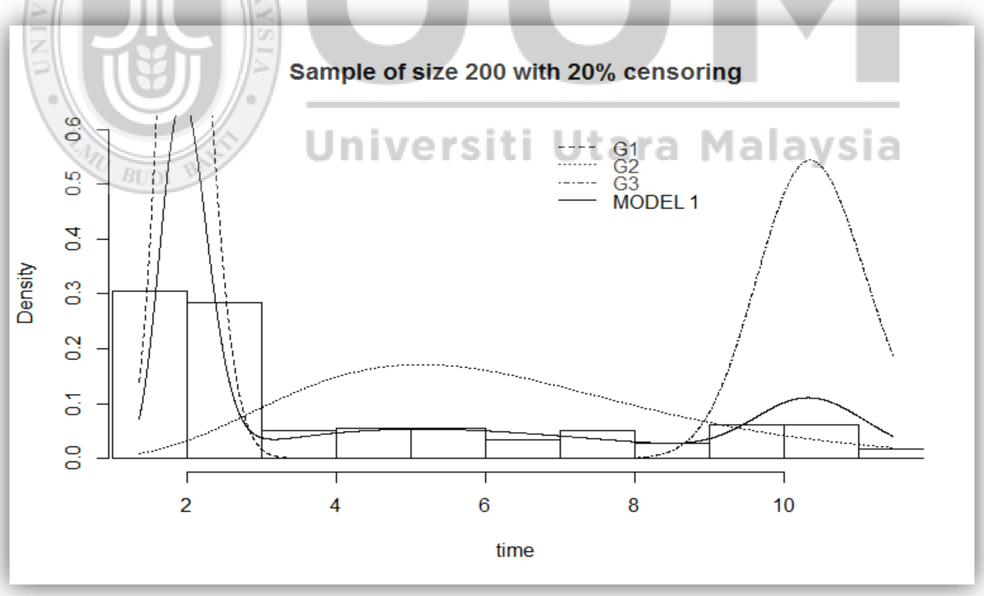


Figure 4.18 Probability Density Function of the Simulated Data of Size 200 Observations and 20% Censoring.

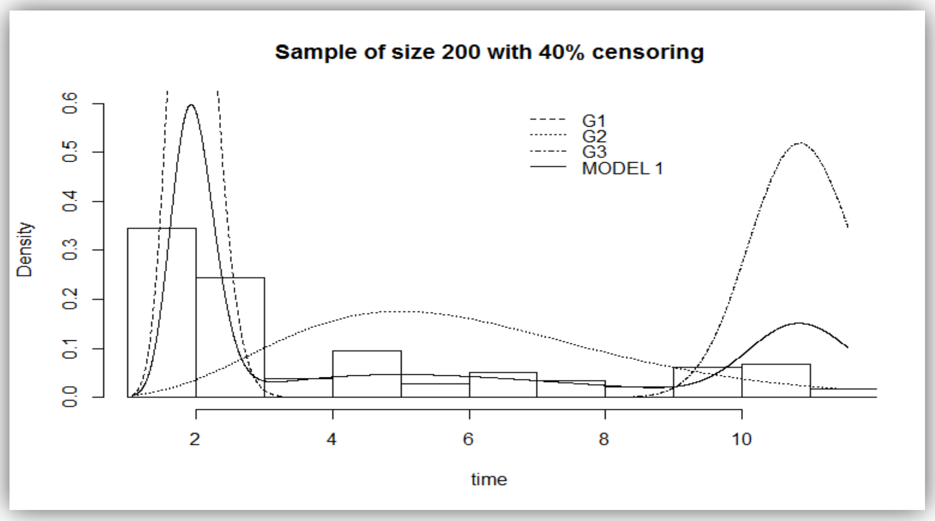


Figure 4.19 Probability Density Function of the Simulated Data of Size 200 Observations and 40% Censoring.

It can be seen that Model 1 fits the simulated data better than the individual pure classical parametric survival models which indicates that the simulated data is better modelled by Model 1 than the pure classical parametric survival model.

The parameters of the set of simulated data of size 200 observations with 10%, 20% and 40% censored observations were estimated and the values together with the true parameters of the postulated Model 1 were presented in Table 4.9.

Table 4.9

The Estimated Parameters the Simulated Data of size 200 with 10% Censoring Observations

Model 1 with sample size 200 observations and 10% censoring								
Parameter	$\boldsymbol{\pi_1}$	$\boldsymbol{\pi_2}$	$\boldsymbol{\alpha_1}$	$\boldsymbol{\alpha_2}$	$\boldsymbol{\alpha_3}$	$\boldsymbol{\beta_1}$	$\boldsymbol{\beta_2}$	$\boldsymbol{\beta_3}$
Postulate	0.50	0.30	40	6	200	20	1	20
Estimates	0.49	0.29	40.00	6.00	200.00	20.01	1.00	19.41
Model 1 with sample size 200 observations and 20% censoring								
Parameter	$\boldsymbol{\pi_1}$	$\boldsymbol{\pi_2}$	$\boldsymbol{\alpha_1}$	$\boldsymbol{\alpha_2}$	$\boldsymbol{\alpha_3}$	$\boldsymbol{\beta_1}$	$\boldsymbol{\beta_2}$	$\boldsymbol{\beta_3}$
Postulate	0.50	0.30	40	6	200	20	1	20
Estimates	0.47	0.28	40.05	5.69	199.99	20.0	0.91	19.25
Model 1 with sample size 200 observations and 40% censoring								
Parameter	$\boldsymbol{\pi_1}$	$\boldsymbol{\pi_2}$	$\boldsymbol{\alpha_1}$	$\boldsymbol{\alpha_2}$	$\boldsymbol{\alpha_3}$	$\boldsymbol{\beta_1}$	$\boldsymbol{\beta_2}$	$\boldsymbol{\beta_3}$
Postulate	0.50	0.30	40	6	200	20	1	20
Estimates	0.41	0.23	40.09	4.80	199.95	20.15	0.74	19.19

The estimated parameters of the three set of the simulated data are all close to the postulated parameters used in the data generation. From Table 4.9 it can be observed that the parameter for the simulated set of data with 10% censoring are more closer the postulated parameters compared to that of the 20% and 40% censoring observations. The mixing probabilities were estimated more accurately estimated with lower number of censored observations.

To investigate the effect of changing censoring percentage on the Model 1, the hazard functions of the three simulated data of size 200 observations corresponding to the 10%, 20% and 40% censoring percentages were presented in Figure 4.20.

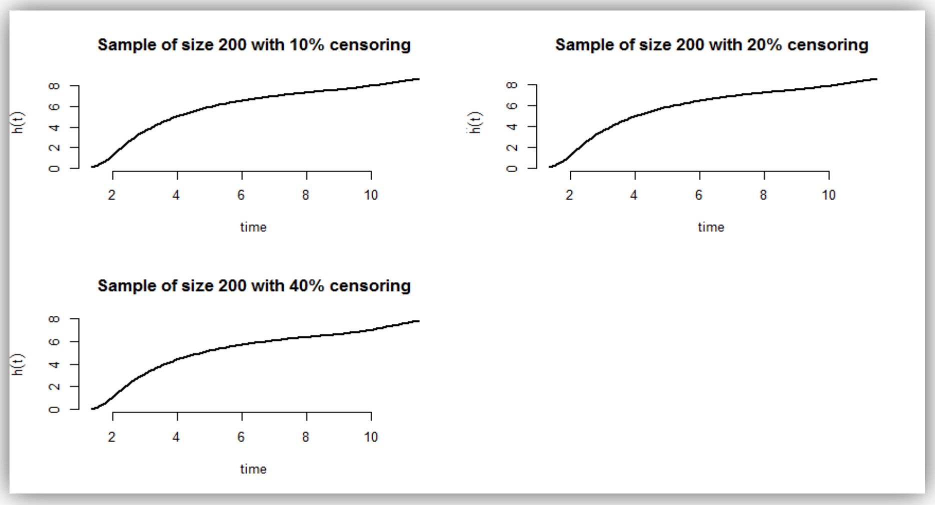


Figure 4.20 The Hazard Functions of the Simulated Data of Size 200 Corresponding to 10%, 20% and 40% Censored Observation.

The hazard function of simulated data of size 200 observations with 10% of the observations censored is higher than that of 20% and 40% censoring. As the number of censored observations increases the hazard tends to be lower and lower.

The simulation of the three set of generated data of 200 observations with 10%, 20% and 40% censoring observations were repeated 300 times to check the consistency and stability of the EM in estimating the model parameters. The averages, the mean square errors and root mean square error of estimated parameters of the postulated Model 1 are listed in Table 4.10.

Table 4.10

The Repeated Simulation of Set of 200 Observations

Model 1 with sample size 200 and 10% censoring								
Parameters	π_1	π_2	α_1	α_2	α_3	β_1	β_2	β_3
Postulates	0.50	0.30	40	6	200	20	1	20
Estimates	0.48	0.29	40.61	5.38	199.28	20.32	0.87	19.65
MSE	5.68e-7	1.63e-6	6.28e-2	1.48e-4	1.70e-1	1.59e-2	5.63e-5	1.90e-3
RMSE	7.53e-4	1.28e-3	2.51e-1	1.21e-2	4.12e-1	1.26e-1	7.51e-3	4.36e-2
Model 1 with sample size 200 and 20% censoring								
Parameters	π_1	π_2	α_1	α_2	α_3	β_1	β_2	β_3
Postulated	0.50	0.30	40	6	200	20	1	20
Estimates	0.45	0.27	40.50	4.34	196.87	20.25	0.65	18.93
MSE	1.09e-6	1.97e-6	8.30e-2	1.64e-3	5.08e-1	2.07e-2	6.57e-5	5.03e-3
RMSE	1.04e-3	1.40e-3	2.87e-1	4.05e-2	7.13e-1	1.44e-1	8.11e-3	7.09e-2
Model 1 with sample size 200 and 40% censoring								
Parameters	π_1	π_2	α_1	α_2	α_3	β_1	β_2	β_3
Postulated	0.50	0.30	40	6	200	20	1	20
Estimates	0.41	0.26	41.13	3.68	195.98	20.56	0.52	18.40
MSE	1.55e-6	2.45e-6	7.81e-2	1.61e-3	6.71e-1	1.98e-2	5.72e-5	6.36e-3
RMSE	1.24e-3	1.57e-3	2.80e-1	3.75e-2	8.19e-1	1.41e-1	7.57e-3	7.97e-2

The averages of the parameters are close to the parameters of the postulated parametric survival mixture model with MSE and RMSE relatively small, which suggests that, the EM performed consistently in estimating the parameters. The MSE corresponding to the mixing probabilities are relatively smaller for the 10% censoring as compared to the 40% censoring. Also the MSE for the parameters of the components are smaller for the 10% censoring compared to that of the 40% with exception of the third component. Generally, the estimation of the mixing probabilities and the parameters are seemed to be closer to the true value with smaller censoring percentage 10% than with 40%.

4.3.2.3 Sample of Size 500 observations

Three sets of survival data of size 500 observations with 10%, 20% and 40% censored observations, respectively, were generated and employed to estimate the parameters of the postulated Model 1 by using the EM. The estimated parameters of corresponding to each set of data and the parameters of the postulated models were reported. The probability density function of simulated data of size 500 observations with 10%, 20% and 40% censored observations were presented in Figures 4.21, 4.22 and 4.23 respectively. The graphs also, display the probability density functions of pure classical parametric survival models (G1, G2 and G3) corresponding to each component of Model 1.

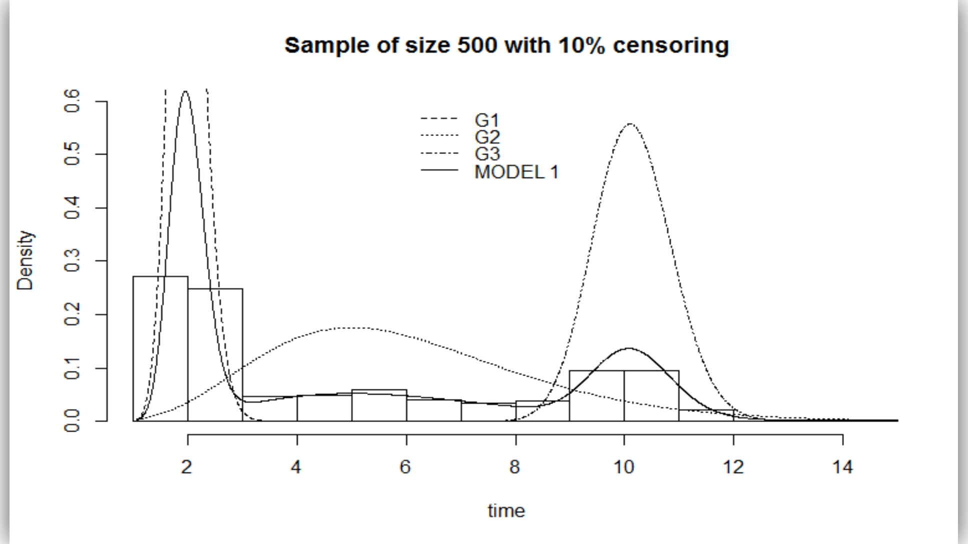


Figure 4.21 Probability Density Function of the Simulated Data of Size 500 Observations and10% Censoring.

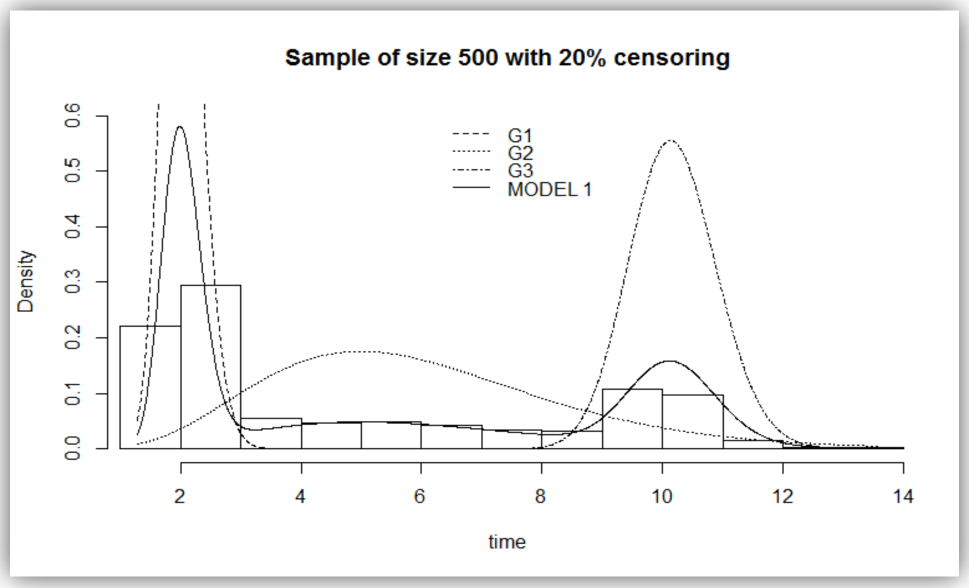


Figure 4.22 Probability Density Function of the Simulated Data of Size 500 Observations and 20% Censoring.

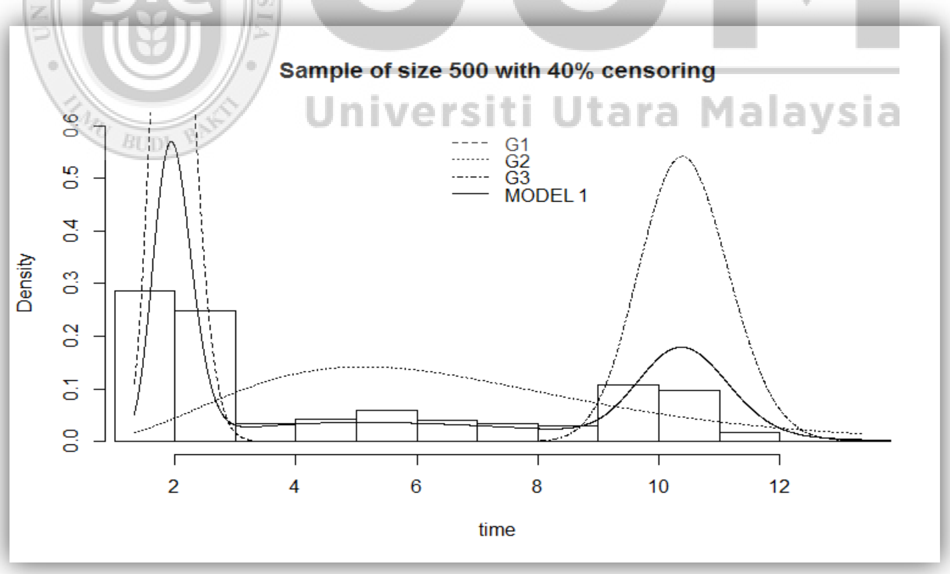


Figure 4.23 Probability Density Function of the Simulated Data of Size 500 Observations and 40% Censoring.

It can be seen that Model 1 fits the simulated data better than the individual pure classical parametric survival models which indicates that the simulated data is better modelled by Model 1 than the pure classical parametric survival model.

The estimated parameters of the set of simulated data of size 500 with 10%, 20% and 40% censored observations were presented in Table 4.11. The estimated parameters are close to the values of the parameters of the postulated model.

Table 4.11

The Estimated Parameters the Simulated Data of size 500 with 10% Censoring Observations

Model 1 with sample size 500 observations and 10% censoring								
Parameter	π_1	π_2	α_1	α_2	α_3	β_1	β_2	β_3
Postulate	0.50	0.30	40	6	200	20	1	20
Estimates	0.48	0.30	40.00	6.00	200.00	20.05	1.00	19.70
Model 1 with sample size 500 observations and 20% censoring								
Parameter	π_1	π_2	α_1	α_2	α_3	β_1	β_2	β_3
Postulate	0.50	0.30	40	6	200	20	1	20
Estimates	0.46	0.28	40.00	6.00	200.00	19.63	1.00	19.64
Model 1 with sample size 500 observations and 40% censoring								
Parameter	π_1	π_2	α_1	α_2	α_3	β_1	β_2	β_3
Postulate	0.50	0.30	40	6	200	20	1	20
Estimates	0.43	0.26	40.01	4.48	199.96	20.17	0.68	19.16

The estimated parameters of the three set of the simulated data are all close to the postulated parameters used in the data generation. From Table 4.11 it can be observed that the parameter for the simulated set of data with 10% censoring are more closer the postulated parameters compared to that of the 20% and 40% censoring observations.

The hazard functions of the three simulated data of size 500 observations corresponding to the 10%, 20% and 40% censoring percentages were presented in Figure 4.24.

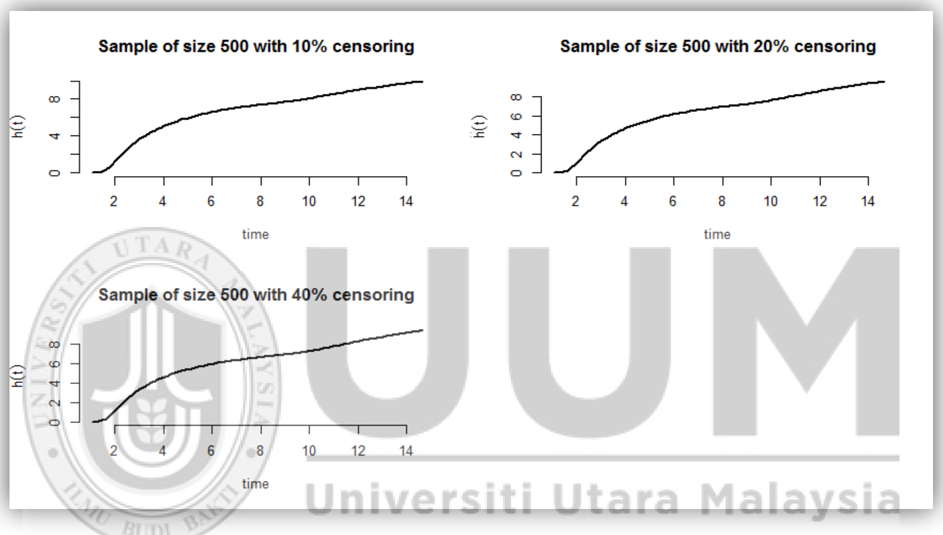


Figure 4.24 The Hazard Functions of the Simulated Data of Size 500 Corresponding to 10%, 20% and 40% Censored Observations.

The hazard function of simulated data of size 500 observations with 10% of the observations censored is higher than that of 20% and 40% censoring. As the number of censored observations increases the hazard tends to be lower and lower.

The simulation of the three set of generated data of 500 observations with 10%, 20% and 40% censoring observations were repeated 300 times to check the consistency and stability of the EM in estimating the model parameters. The averages, the mean

square errors and root mean square error of estimated parameters of the postulated Model 1 are listed in Table 4.12.

Table 4.12

The Repeated Simulation of Set of 500 Observations

Model 1 with sample size 500 and 10% censoring								
Parameters	π_1	π_2	α_1	α_2	α_3	β_1	β_2	β_3
Postulates	0.50	0.30	40	6	200	20	1	20
Estimates	0.48	0.29	40.37	5.46	199.49	20.18	0.88	19.74
MSE	2.29e-7	5.68e-7	2.17e-2	7.47e-4	4.77e-2	5.33e-3	2.75e-6	5.44e-4
RMSE	4.79e-4	7.54e-4	1.47e-01	2.73e-2	2.18e-1	7.30e-2	1.66e-3	2.33e-2
Model 1 with sample size 500 and 20% censoring								
Parameters	π_1	π_2	α_1	α_2	α_3	β_1	β_2	β_3
Postulated	0.50	0.30	40	6	200	20	1	20
Estimates	0.47	0.29	40.25	4.6	198.39	20.11	0.71	19.38
MSE	2.62e-7	3.84e-6	2.35e-2	6.15e-3	2.27e-1	5.73e-3	2.08e-5	2.24e-3
RMSE	5.12e-4	1.96e-4	1.53e-1	2.48e-2	4.77e1	7.57e-2	4.56e-3	4.73e-2
Model 1 with sample size 500 and 40% censoring								
Parameters	π_1	π_2	α_1	α_2	α_3	β_1	β_2	β_3
Postulated	0.50	0.30	40	6	200	20	1	20
Estimates	0.43	0.26	39.99	3.79	196.18	20.01	0.54	18.63
MSE	5.25e-7	9.56e-7	3.14e-2	5.96e-3	5.75e-1	7.68e-3	2.31e-5	5.36e-3
RMSE	7.24e-4	9.78e-4	1.77e-1	2.44e-2	7.58e-1	8.77e-2	4.80e-3	7.32e-2

The averages of the parameters are close to the true parameters of the postulated parametric survival mixture model with mean square errors relatively small, which suggests that, the EM performed consistently in estimating the parameters.

The estimation of the parameters of the model was successful for both the ascending and descending order of the mixing probabilities. The estimated parameters were closer the true postulate parameters as the sample size increases from 100 to 500 observations for both the ascending and descending order mixing probabilities. It is also observed that the estimates of the parameters were much better for small censoring percentages. The mixing probabilities were better estimated for small number of censored observation. It could be due to the fact that with the increase in number of censored observation considerable amount of information were lost. Generally, the mixing probabilities for the ascending order have smaller MSE value corresponding to the estimates of the parameters which shows that they were better than the parameters estimated with the descending order, especially with the increase in the censoring percentages. In general, it is observed that the mixing probabilities of ascending order performed better than the descending order.

4.3.3 Special Case of Model 1

The pure classical parametric survival model of Gamma distribution can be looked at as a special case of Model 1 by setting the number of components of the parametric survival mixture model of the Gamma distributions equals to one as was shown in Figure 3.1. Pure classical parametric survival model of the Gamma distribution has been simulated. The predetermined parameters of pure classical parametric survival model of the Gamma distribution is given by

$$f(t) = f_G(t; \alpha_1 = 20, \beta_1 = 6) \tag{4.2}$$

The simulated data were used to estimate the parameters of the pure classical parametric survival model of the Gamma distribution and the parameter estimates are reported.

The probability density function of the estimated pure classical parametric survival model of the Gamma was plotted together with the histogram of the simulated data in Figure 4.25. The density function plotted indicates that the model fits the data well.

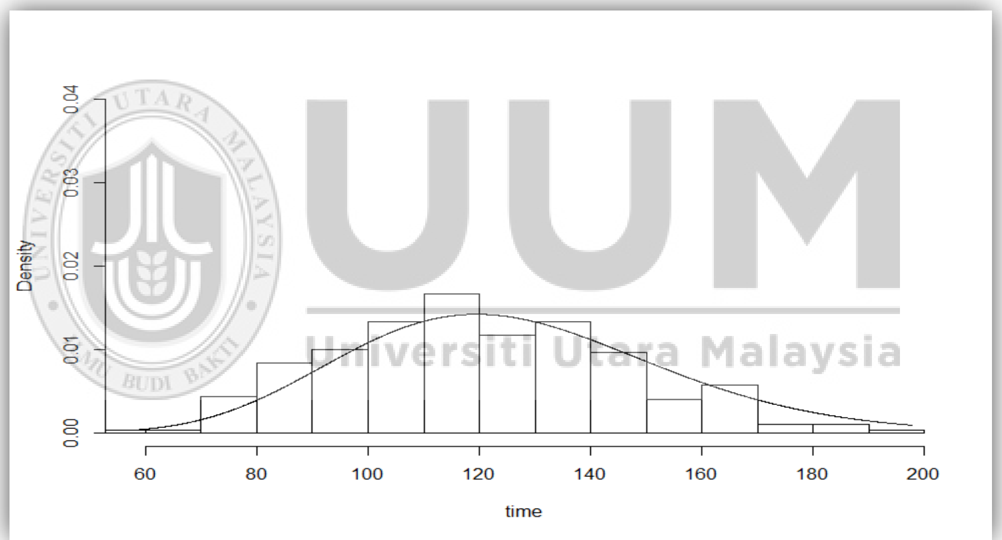


*Figure 4.25*The Density Functions of Pure Classical Parametric Survival Model of the Gamma Distribution

Table 4.13 displays the result of the estimates of the parameters of the simulated data of the pure classical parametric survival model of the Gamma distribution. The results show that the estimates of the parameters are close to the original parameters of the postulated pure classical parametric survival model.

Table 4.13

The Estimated Parameters of Simulated Data of the Pure Classical Survival Model of Gamma Distribution

Parameter	α	β
Postulate model	20	6
Estimates	19.29	6.54

4.4 Validating Model 1 Using Real Data

This section is devoted to the application of Model 1 on real data. The parameters of such data were estimated using the EM. Also the graphical representations of the probability density function of the parametric survival mixture models were presented together with the probability density function of the pure classical survival models of each component. The survival function of Model 1 was compared with the K-M empirical survival function to validate the model. The estimated parameters were presented together with the LL, AIC, MSE, RMSE, Kolmogorov- Simonov test (K-S) and the mean survival time *E(t)* values. The pure classical parametric survival model of the Gamma distribution was presented as a special case when the number of components of the proposed parametric survival mixture model is set to one.

4.4.1 Bone Marrow Transplant Data

The Bone Marrow Transplant data had been used as real data. The three components parametric survival mixture model of the Gamma distributions (Model 1) were applied on the Marrow Transplant Study data.

The probability density function of Model 1 and the probability density function of the pure classical survival models of the Gamma distribution (G1, G2 and G3) corresponding to each of the components of Model 1 were plotted together with the histogram of the Marrow Transplant data in Figure 4.26. The graph indicates that Model 1 fits the data better than the individual pure classical survival models of the Gamma distributions corresponding to each component.

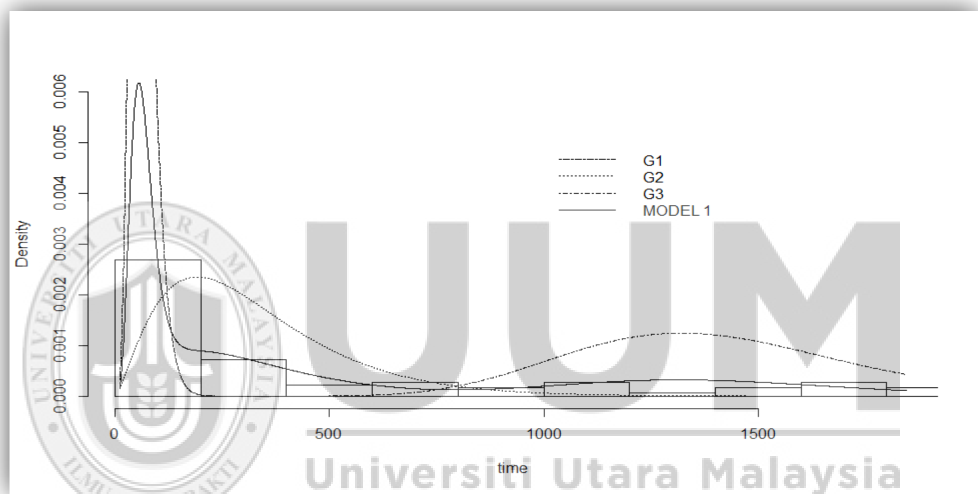


Figure 4.26 probability density function of Model 1 using Bone Marrow Transplant Data

The values of estimated parameters of Model 1of the Marrow Transplant data were displayed in Table 4.14.

Table 4.14

The Estimated Parameters of Model 1 of Marrow Transplant Data

Parameter	$\boldsymbol{\pi_1}$	$\boldsymbol{\pi_2}$	$\boldsymbol{\alpha_1}$	$\boldsymbol{\alpha_2}$	$\boldsymbol{\alpha_3}$	$\boldsymbol{\beta_1}$	$\boldsymbol{\beta_2}$	$\boldsymbol{\beta_3}$
Estimates	0.37	0.38	5.81	2.43	18.06	11.39	134.25	77.42

The probability density function of Model 1 and the probability density function of the pure classical survival model of the Gamma distribution (G0) evaluated using the Marrow Transplant data, were plotted together with the histogram of the Marrow Transplant data in Figure 4.27. The graph indicates that Model 1 fits the data better than the pure classical survival model of the Gamma distribution (G0).

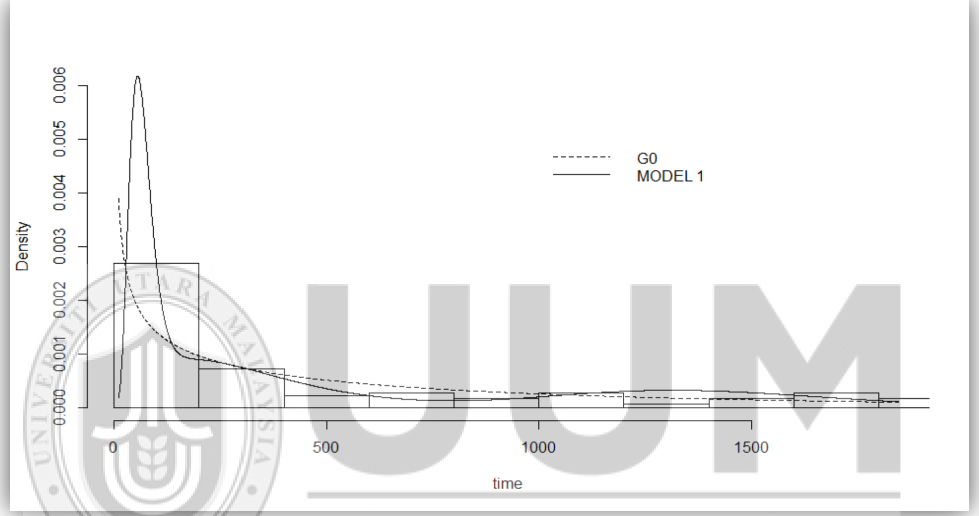


*Figure 4.27.*Model 1 vs the Pure Classical Survival Model of the Gamma Distribution (G0) Using Bone Marrow Transplant Data

The parameters, LL, AIC, MSE, RMSE, Kolmogorov-Smirnov K-S test and the mean survival time E(t) values were estimated and reported. Table 4.15 shows that, Model 1 scored higher value for the LL (-493.01) than the values (-501.13) scored by the pure classical survival parametric model of the Gamma distribution (G0). Also, the AIC value (1002.02) of Model 1 was smaller compared to corresponding value (1006.26) of the pure classical parametric survival model of the Gamma distribution (G0). The MSE of the fitted Model 1 (0.0016) is smaller than that of the pure classical model (0.0117). This result indicates that the Marrow Transplant Study data seem to be appropriately fitted by Model 1.

Table 4.15

The Estimated Parameters of Model 1 for Bone Marrow Transplant Data

Model	Estimates	LL	AIC	MSE	RMSE	K-S	E(T)
G0	$\hat{\alpha}$ = 0.57, $\hat{\beta}$ = 1243.48	-501.13	1006.26	0.0117	0.1082	0.02 (0.05)	708.78
Model 1	$\hat{\alpha}_1$=5.81, $\hat{\beta}_1$= 11.39 $\hat{\alpha}_2$ = 2.43, $\hat{\beta}_2$= 134.25 $\hat{\alpha}_3$ = 18.06, $\hat{\beta}_3$= 77.42 $\hat{\pi}_1$ = 0.37, $\hat{\pi}_2$= 0.38	**-493.01**	**1002.02**	**0.0016**	**0.0400**	**0.09** **(0.87)**	**504.42**

The K-S test statistic of Model 1 (0.09) with the p-value in bracket shows that Model 1 fits the data better than the pure classical survival distribution.

The survival function graph of the fitted Bone Marrow Transplant data used to validate the fit of Model 1. The survival function graph was compared with the K-M empirical survival function of the real data to investigate the fit of Model 1. The survival function of Model 1 and the K-M empirical survival function were presented in Figure 4.28.

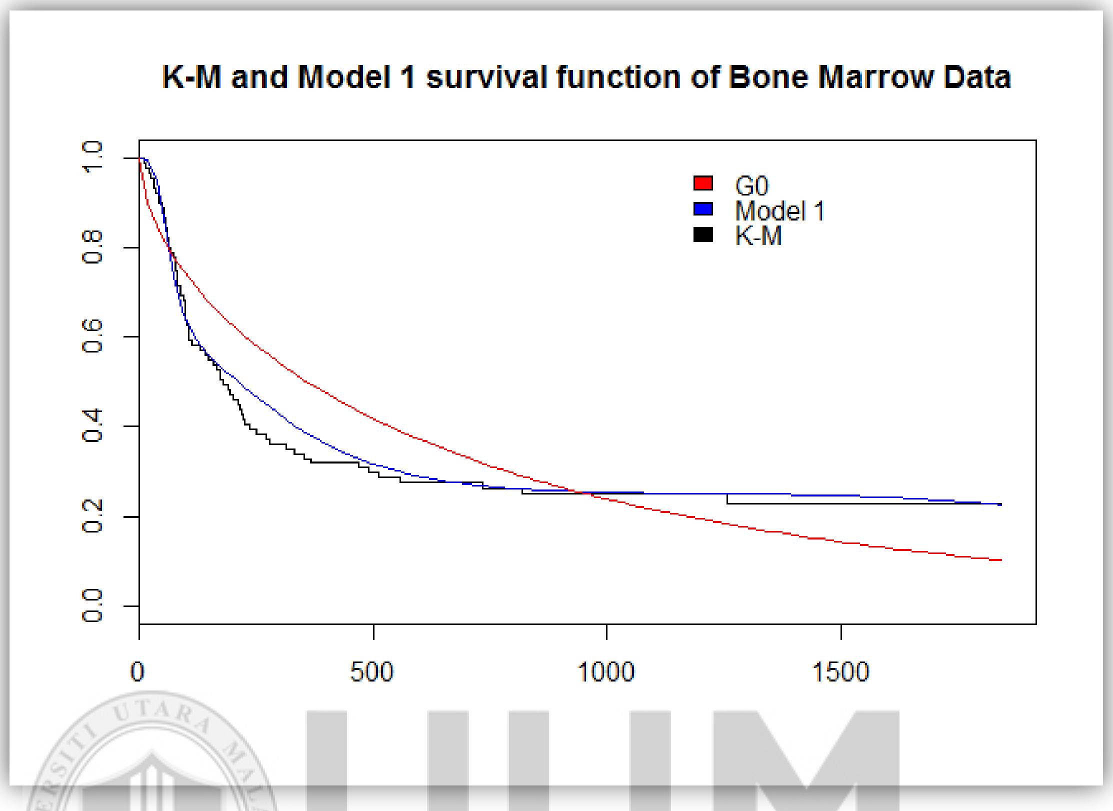


Figure 4.28 K-M, the Survival function of Model 1 and the Pure Survival Model

In Figure 4.28 the K-M empirical survival function is in solid black, the survival function of Model 1 is in dark blue, the pure classical survival model of the Gamma distribution is in red. From the Figure it can be observed that the survival function of Model 1 is in full agreement with the K-M empirical survival function and much better than the pure classical survival model.

Model selection was performed among Model 1, the two components parametric survival mixture models of the Gamma distributions (G1_G2) and the four components parametric survival mixture models of the Gamma distributions (G1_G2_G3_G4) to select the model that represents the Bone Marrow Transplant

Study data better by applying the LL and AIC criterion. The estimates of the parameters of (G1_G2) are presented in Table 4.16.

Table 4.16

The Estimated Parameters of (G1_G2) of the Bone Marrow Transplant Data

Parameter	$\boldsymbol{\pi_1}$	$\boldsymbol{\alpha_1}$	$\boldsymbol{\alpha_2}$	$\boldsymbol{\beta_1}$	$\boldsymbol{\beta_2}$
Estimates	0.80	0.71	186.86	446.08	8.95

Also the estimates of the parameters of the (G1_G2_G3_G4) are presented in Table 4.17.

Table 4.17

The Estimated Parameters of (G1_G2_G3_G4) of the Bone Marrow Transplant Data

Parameter	$\boldsymbol{\pi_1}$	$\boldsymbol{\pi_2}$	$\boldsymbol{\pi_3}$	$\boldsymbol{\alpha_1}$	$\boldsymbol{\alpha_2}$	$\boldsymbol{\alpha_3}$	$\boldsymbol{\alpha_4}$	$\boldsymbol{\beta_1}$	$\boldsymbol{\beta_2}$	$\boldsymbol{\beta_3}$	$\boldsymbol{\beta_3}$
Estimates	0.23	0.23	0.26	6.10	13.13	4.52	10.85	8.06	8.48	71.94	118.55

Table 4.18 Gives the LL and the AIC corresponding to each parametric survival mixture models of the Gamma distributions. The LL value of Model 1 (-493.01) is higher than that of the two, four components parametric survival mixture model of the Gamma distributions (-497.41), (-494.39) respectively. The AIC criterion value of Model 1 (1002.02) is smaller than that of the two and four components parametric survival mixture model of the Gamma distributions respectively.

Table 4.18

The LL and AIC Values of the Parametric Survival Mixture Models of the Gamma Distribution

Number of components		**2 (G1_G2)**	**3 (Model 1: G1_G2_G3)**	**4 (G1_G2_G3_G4)**
Mixture of Gamma	**LL**	-497.41	**-493.01**	-494.39
	AIC	1004.82	**1002.02**	1010.77

The result shows that both the LL and AIC are in support of Model 1. Three sub-populations fit the Bone Marrow data much better than the two, four sub-populations survival mixture model and the pure classical survival model.

4.4.2 Kidney Catheter Data

The set of real data analysed in this section is the Kidney Catheter data which is included as one of the data set in the famous *survival* package developed by Therneau (1999) of the R statistical software (Team, 2005). This data were studied originally by McGilchrist and Aisbett (1991). The data give the recurrence times to infection, at the point of insertion of catheters, of kidney patients using portable dialysis equipment. It consists of 76 observations and 7 variables as presented in Appendix B. The data constitutes of 18 censored observations which makes the censoring percentage approximately 24%. The data were used to fit Model 1.

The probability density function of Model 1 and the probability density function of the pure classical survival models of the Gamma distribution (G1, G2 and G3) corresponding to each of the components of Model 1 were plotted together with the histogram of the Kidney Catheter data in Figure 4.29. The graph indicates that Model 1 fits the data better than the individual pure classical survival models of the Gamma distributions corresponding to each component.

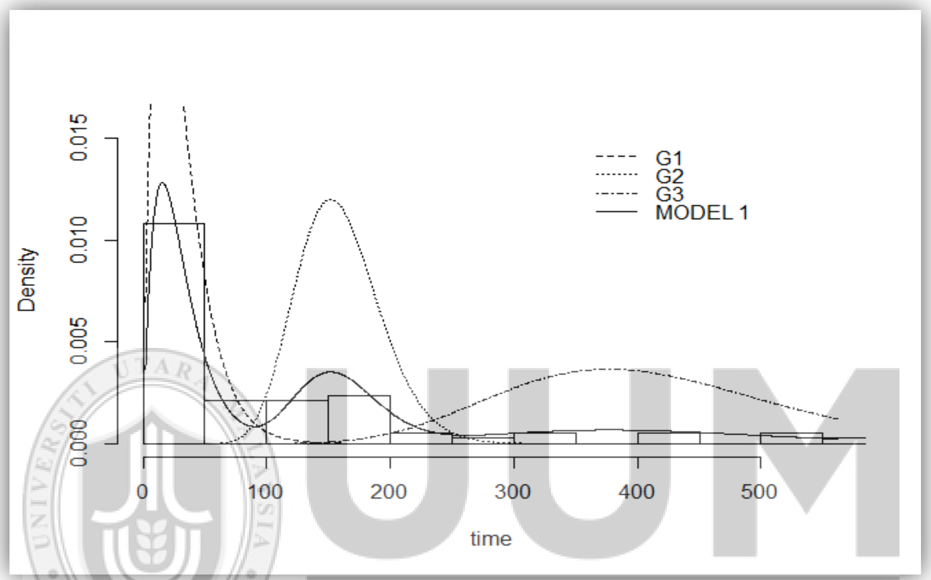


Figure 4.29 the probability density function of Model 1 using Kidney Catheter Data

The values of estimated parameters of Model 1of the Kidney Catheter data were displayed in Table 4.19.

Table 4.19

The Estimated Parameters of Model 1 of Kidney Catheter Data

Parameter	π_1	π_2	α_1	α_2	α_3	β_1	β_2	β_3
Estimates	0.53	0.29	2.06	21.97	13.05	14.75	7.14	31.28

The probability density function of Model 1 and the probability density function of the pure classical survival model of the Gamma distribution (G0) evaluated using the Kidney Catheter data; were plotted together with the histogram of the Kidney Catheter data in Figure 4.30. The graph indicates that Model 1 fits the data better than the pure classical survival model of the Gamma distribution (G0).

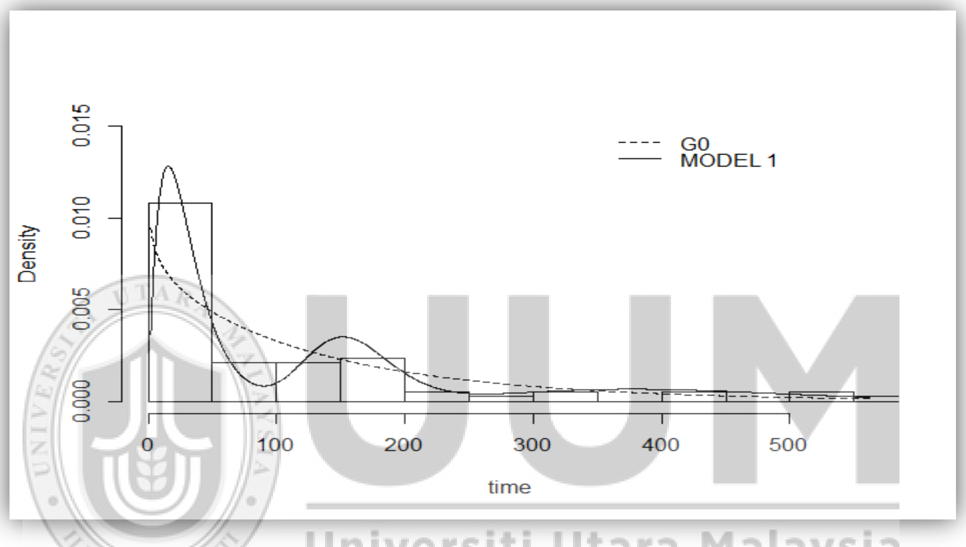


*Figure 4.30.*Model 1 vs the Pure Classical Survival Model of the Gamma Distribution (G0) Using Kidney Catheter Data

The parameters, LL, AIC, MSE, RMSE, Kolmogorov-Smirnov K-S test and the mean survival time E(t) values were estimated and reported. Table 4.20 shows that, Model 1 scored higher value for the LL (-331.57) than the values (-341.20) scored by the pure classical survival parametric model of the Gamma distribution (G0). Also, the AIC value (679.13) of Model 1 was smaller compared to corresponding value (686.40) of the pure classical parametric survival model of the Gamma distribution (G0). The MSE of the fitted Model 1 (0.0108) is smaller than that of the pure classical model (0.0194). This result indicates that the Kidney Catheter data seem to be appropriately fitted by Model 1.

Table 4.20

The Estimated Parameters of Model 1 for Kidney Catheter Data

Model	Estimates	LL	AIC	MSE	RMSE	K-S	E(T)
G0	$\hat{\alpha}$ = 0.89, $\hat{\beta}$ = 156.96	-341.20	686.40	0.0194	0.1392	0.25 (0.02)	139.69
Model 1	$\hat{\alpha}_1$=2.06, $\hat{\beta}_1$= 14.75 $\hat{\alpha}_2$ = 21.97, $\hat{\beta}_2$= 7.14 $\hat{\alpha}_3$ = 13.05, $\hat{\beta}_3$= 31.28 $\hat{\pi}_1$ = 0.53, $\hat{\pi}_2$= 0.29	**-331.57**	**679.13**	**0.0108**	**0.1038**	**0.16** **(0.30)**	**137.00**

The K-S test statistic of Model 1 (0.16) with the p-value in bracket shows that Model 1 fits the data better than the pure classical survival distribution.

The survival function graph of the fitted Kidney Catheter data used to validate the fit of Model 1. The survival function graph was compared with the K-M empirical survival function of the real data to investigate the fit of Model 1. The survival function of Model 1 and the K-M empirical survival function were presented in Figure 4.31.

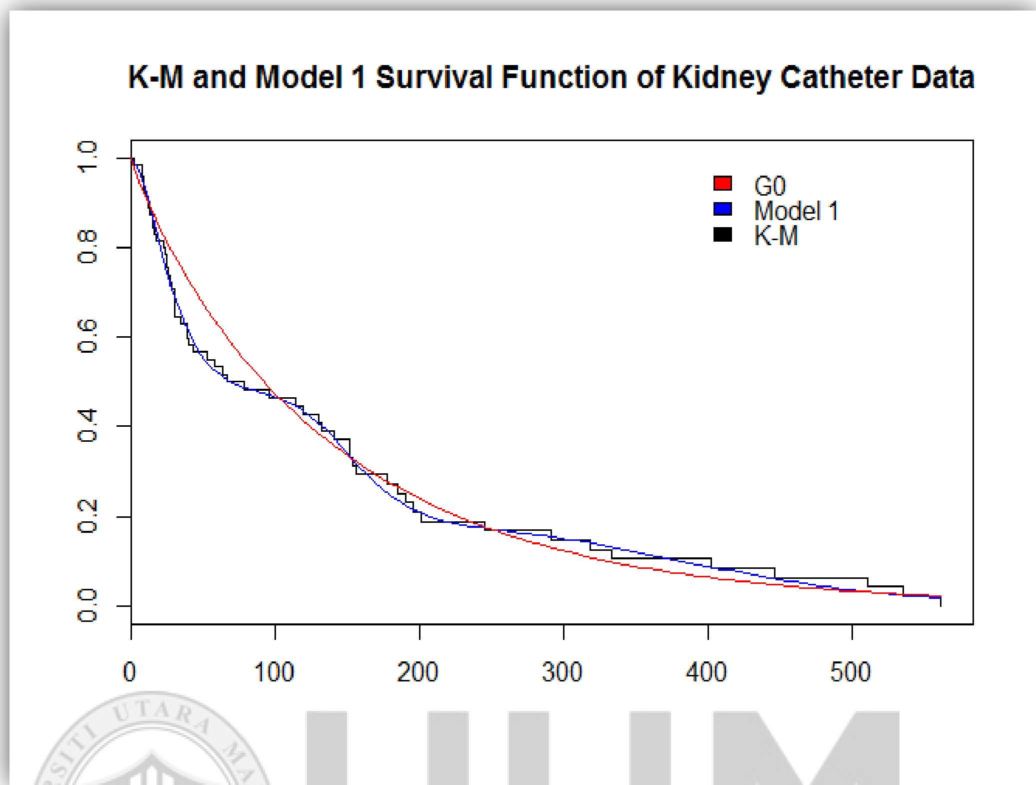


Figure 4.31 K-M, the Survival function of Model 1 and the Pure Survival Model

In Figure 4.31 the K-M empirical survival function is in solid black, the survival function of Model 1 is in dark blue, the pure classical survival model of the Gamma distribution is in red. Form the Figure it can be observed that the survival function of Model 1 is in full agreement with the K-M empirical survival function much better than the pure classical survival model.

The histogram of the Kidney Catheter data shows that mixture structure is appropriate for the data; hence the AIC model selection was used to determine the sub-population that fits the data. The Kidney Catheter data were used to model a two components parametric survival mixture model of the Gamma distributions (G1_G2) and four components parametric survival mixture model of the Gamma distributions

(G1_G2_G3_G4). The estimates of the parameters of (G1_G2) are presented in Table 4.21.

Table 4.21

The Estimated Parameters of (G1_G2) of the Kidney Catheter Data

Parameter	$\boldsymbol{\pi_1}$	$\boldsymbol{\alpha_1}$	$\boldsymbol{\alpha_2}$	$\boldsymbol{\beta_1}$	$\boldsymbol{\beta_2}$
Estimates	0.44	2.41	2.30	10.29	98.26

Also the estimates of the parameters of the (G1_G2_G3_G4) are presented in Table 4.22.

Table 4.22

The Estimated Parameters of of (G1_G2_G3_G4) of the Kidney Catheter Data

Parameter	$\boldsymbol{\pi_1}$	$\boldsymbol{\pi_2}$	$\boldsymbol{\pi_3}$	α_1	α_2	α_3	α_4	$\boldsymbol{\beta_1}$	$\boldsymbol{\beta_2}$	$\boldsymbol{\beta_3}$	$\boldsymbol{\beta_3}$
Estimates	0.16	0.24	0.18	5.21	17.82	5.60	3.97	1.81	1.48	15.27	69.48

Model selection was performed among Model 1, the two components parametric survival mixture models of the Gamma distributions (G1_G2) and the four components parametric survival mixture models of the Gamma distributions (G1_G2_G3_G4) to select the model that represents Kidney Catheter Study data better by applying the LL and AIC criterion.

Table 4.23 gives the LL and the AIC corresponding to each parametric survival mixture model of the Gamma distributions. The LL value of Model 1 (-331.57) is higher than that of the two, four components parametric survival mixture model of the Gamma distributions (-334.90), (-336.88) respectively. The AIC criterion value of Model 1 (679.13) is smaller than that of the two and four components parametric survival mixture model of the Gamma distributions respectively.

Table 4.23

The LL and AIC Values of the Parametric Survival Mixture Models of the Gamma Distribution

Number of components		**2** **(G1_G2)**	**3** **(Model 1: G1_G2_G3)**	**4** **(G1_G2_G3_G4)**
Mixture of Gamma	**LL**	-334.90	**-331.57**	-336.88
	AIC	681.80	**679.13**	695.76

The result shows that both the LL and AIC are in support of Model 1. Three sub-populations fit the Kidney Catheter data much better than the two, four sub-populations survival mixture model and the pure classical survival model.

As a special case of the parametric survival mixture model of the Gamma distributions, the pure classical parametric survival model of the Gamma distribution has been used to model Vaginal Cancer data in the next sub-section.

4.4.3 The Special Case of Model 1

The pure classical parametric survival model of the Gamma distribution was applied to the Vaginal Cancer data by setting the number of components of Model 1 to one. The Vaginal Cancer data set is one of the data sets included in *survival* package developed by Therneau (1999) which is one of the packages of the R statistical software (Team, 2005).

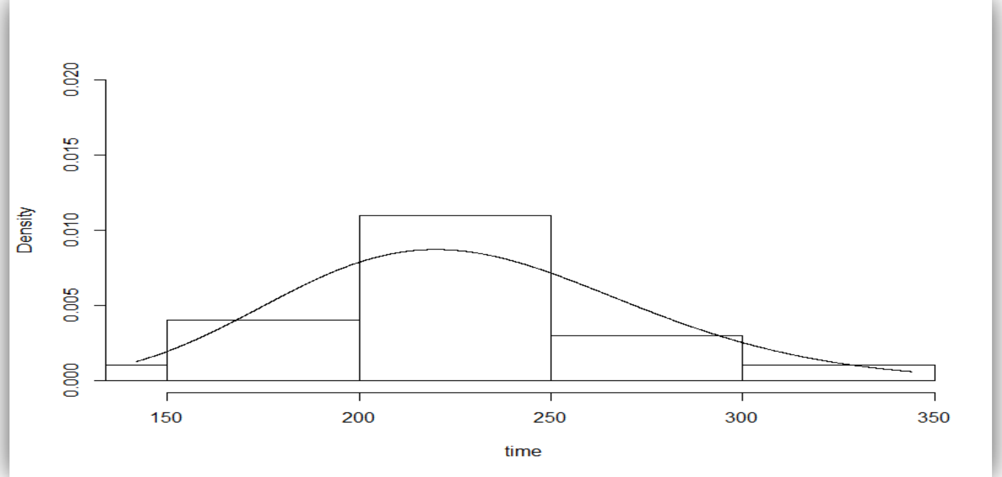


Figure 4.32 The Pure Classical Parametric Survival Model of the Gamma Distribution for Vaginal Cancer Data

The probability density function of the pure classical parametric survival model of the Gamma distribution was plotted together with the histogram of the Vaginal Cancer data in Figure 4.32. The density function plotted indicates that the model fits the data well.

The estimated parameters of the pure classical parametric survival Gamma distribution model are shown in Table 4.24 along with the LL value.

Table 4.24

The Estimated Parameters of Vaginal Cancer Data of the Pure Classical Gamma Distribution

Parameter	α	β
Estimates	24.22	9.47
LL	-191.68	

4.5 Summary

This chapter discussed the development of a three components parametric survival mixture model of the Gamma distributions (Model 1). The EM, the model derivation and the estimation of the parameters of Model 1 were highlighted.

Simulation study was carried out to investigate and validate the performance of Model 1. The simulated data constituted of the three different samples of size 100, 200 and 500 observations respectively. Each of the samples was based on three different censoring percentages. Also the generated samples were based on two sets of different mixing probabilities arranged in ascending and descending order. Simulated data of 18 different samples were generated from the parametric survival mixture model of Gamma distribution. The EM was employed in estimating the parameters of Model 1 and the consistency and stability of EM was investigated by repeating the simulation 300 times.

Generally, the parameters estimated from the data were closed to the true parameters used in the simulation of the data. The simulation was repeated 300 times and the MSE and RMSE were obtained. Validating the performance of Model 1 using the three different sample sizes showed that the estimation of the parameters was better as the sample size increases. Comparing the three censoring percentages showed that the parameter estimation of the Model 1 was better with smaller censoring percentages for both the ascending and descending order of the mixing probabilities. However, the performance of Model 1 with the mixing probabilities in ascending order was better than that of the mixing probabilities in descending order. The hazard functions for different samples of Model 1 with different censoring percentages were investigated and the graphical representations were provided. Generally, it was found that the hazard function tends to be higher with small censoring percentage. As the censoring percentage increases more individual or items survive which reduce the value of the hazard function.

Empirical study was also employed to validate Model 1. The parameters of Model 1 were estimated and reported. Model 1 was compared with pure classical parametric survival model of Gamma distribution (G0) evaluated using real data and pure classical parametric survival model of Gamma distribution (G1, G2 and G3) corresponding to the distribution to each component of Model 1 graphically. To validate Model 1, the LL, AIC, MSE, RMSE, K-S test and E(t) were evaluated and compared with those of the pure classical Gamma distribution. The K-M empirical survival function was better represented by the survival function of Model 1 compared to the pure classical survival model.

Model 1 was also compared with the two and four components parametric survival mixture model of the Gamma distributions using LL and AIC values to select the number of component that better represents the real data. The comparison showed that real data were better modelled with three component mixture model, and the data constitute of three sub-populations. Model 1 was used to evaluate the pure classical parametric survival model of the Gamma distribution when the number of components of Model 1 set to one. The application of the simulation and empirical studies showed that Model 1 is preferred over the pure classical survival models in modelling survival data when the data seem to come from population of heterogeneous nature.

CHAPTER FIVE
THREE COMPONENTS PARAMETRIC SURVIVAL MIXTURE MODEL OF THE EXPONENTIAL, GAMMA AND WEIBULL DISTRIBUTIONS

5.1 Introduction

A three components parametric survival mixture model of the Exponential, Gamma and Weibull distributions (E_G_W, referred to as Model 2) was developed. The chapter was arranged as follows. The first section highlighted the implementation of EM in the theoretical development of Model 2. The section also includes the explanation regarding the algorithm transformation to computer coding. The next two sections are validation of Model 2 based on simulated and real data respectively. The last section summarized the outcomes and findings of this chapter.

5.2 Theoretical Development of Model 2

The second objective of the study is about developing a three components parametric survival mixture model of the Exponential, Gamma and Weibull distributions (Model 2). The implementation of the EM for estimating the parameters of the model was based on random censoring procedure. The general formulation of a three components parametric survival mixture model which is assumed to consist of three sub-populations; each sub-population corresponds to a component in the parametric survival mixture model, as was highlighted in section 4.2.1 in Chapter Four.

Since the second objective proposes a three components parametric survival mixture model of the Exponential, Gamma and Weibull distributions (Model 2), then Model 2 can be expressed as follows

$$f_{E\_G\_W}(t;\Theta) = \pi_1 f_E(t;\lambda) + \pi_2 f_G(t;\alpha_1,\beta_1) + \pi_3 f_W(t;\alpha_2,\beta_2), \tag{5.1}$$

where π_i's are the mixing proportion or mixing probability and $\sum_{i=1}^{3}\pi_i = 1$. The functions f_E ,f_G and f_W are the probability density functions of the Exponential, the Gamma and the Weibull distributions respectively corresponding to the components of Model 2. The EM employed to estimate the parameters of Model 2 proceeds as mentioned earlier in section 4.2.1 of Chapter Four. Since Model 2 consists of different distribution, the estimation procedure consider the Exponential distribution, the Gamma distribution and the Weibull distribution for the first, second and third component respectively. The derivation of the parameters of mixture of the survival model the Exponential, Gamma and Weibull distribution are given below.

The probability density function of the mixture of Exponential, Gamma and Weibull is as given in equation (5.1), where $f_E(t;\lambda)$ with unknown parameter λ, $f_G(t;\alpha_1,\beta_1)$ with unknown parameters α_1,β_1 and $f_W(t;\alpha_2,\beta_2)$ with unknown parameters α_2,β_2 are the Exponential, Gamma and Weibull distributions density functions as in Table 2.1. The parameters satisfy the conditions $\lambda > 0$, $\alpha_1 > 0, \beta_1 > 0, \alpha_2 > 0, \beta_2 > 0$.

From (4.22), the log-likelihood function of the complete-data of the mixture of the Exponential, Gamma and Weibull distributions is

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

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

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

$$\left.+(1-\delta_j)\log\left[e^{-\left(\frac{t_j}{\beta_2}\right)^{\alpha_2}}\right]\right\} \tag{5.2}$$

The EM algorithm starts with the E-step. After the g^{th} iteration, $z_{ij}^{(g)}$ is the conditional expectation of z_{ij} given the observed data, as defined in (4.13) and (4.14). Then the current conditional expectation of the complete-data log-likelihood is given by

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

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

$$+\sum_{j=1}^{n} z_{3j}^{(g)}\left\{\log\pi_3+\delta_j\log\left[\left(\frac{\alpha_2}{\beta_2}\right)\left(\frac{t_j}{\beta_2}\right)^{\alpha_2-1}e^{-\left(\frac{t_j}{\beta_2}\right)^{\alpha_2}}\right]+(1-\delta_j)\left[-\left(\frac{t_j}{\beta_2}\right)^{\alpha_2}\right]\right\} \tag{5.3}$$

The M-step on the *(g+1)*th iteration requires the global maximization of (5.3) with respect to $\lambda,\alpha_1,\beta_1,\alpha_2,\beta_2$ and π_i . The mixing probabilities π_i can be updated by $\hat{\pi}_i^{(g+1)}=\sum_{j=1}^{n} z_{ij}^{(g)}/n,\ i=1,2,3$. In order to get the updated maximum likelihood estimate of the component model parameters $\lambda,\alpha_1,\beta_1,\alpha_2,\beta_2$, equation (5.3) will be differentiated with respect to each of the parameters.

Now, differentiating equation (5.3) with respect to the parameter λ the updated maximum likelihood estimate of the component model parameter can be obtained in closed form

$$\hat{\lambda}^{(g+1)}=\frac{\sum_{j=1}^{n} z_{1j}^{(g)}t_j}{\sum_{j=1}^{n} z_{1j}^{(g)}\delta_j} \tag{5.4}$$

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

$$\hat{Z}_{1j}^{(g+1)} = \hat{\pi}_1^{(g)} \left[f_1(t_j ; \hat{\lambda}^{(g)} \right]^{\delta_j} \left[S_i(t_j ; \hat{\lambda} \right]^{1-\delta_j} / \left[\hat{\pi}_1^{(g)} \left[f \ (t_j ; \hat{\lambda}^{(g)} \right]^{\delta_j} \left[S_i(t_j ; \hat{\lambda} \right]^{1-\delta_j} \right.$$

$$+ \hat{\pi}_2^{(g)} \left[f \ (t_j ; \hat{\alpha}_1^{(g)}, \hat{\beta}_1^{(g)} \right]^{\delta_j} \left[S \ (t_j ; \hat{\alpha}_1^{(g)}, \hat{\beta}_1^{(g)} \right]^{1-\delta_j} \qquad (5.5)$$

$$\left. + \hat{\pi}_3^{(g)} \left[f \ (t_j ; \hat{\alpha}_2^{(g)}, \hat{\beta}_2^{(g)} \right]^{\delta_j} \left[S(t_j ; \hat{\alpha}_2^{(g)}, \hat{\beta}_2^{(g)} \right]^{1-\delta_j} \right]$$

Again, differentiating equation (5.3) with respect to the parameter α_1, β_1 yields

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

$$+ \sum_{j=1}^{n} z_{2j}^{(g)} (1-\delta_j) \left[\log \beta_1 + \frac{1}{\Gamma(\alpha_1, t_j / \beta_1)} \frac{\partial}{\partial \alpha_1} \Gamma(\alpha_1, t_j / \beta_1) \right] \qquad (5.6)$$

$$\frac{\partial Q}{\partial \beta_i} = \sum_{j=1}^{n} z_{2j}^{(g)} \left[-\delta_j \left(\frac{\alpha_1}{\beta_1} + \frac{t_i}{\beta_1^2} \right) + \frac{(1-\delta_j)}{\Gamma(\alpha_1, t_j / \beta_1)} \frac{\partial}{\partial \beta_1} \Gamma(\alpha_1, t_j / \beta_1) \right] \qquad (5.7)$$

Now, the incomplete gamma function can be differentiated with respect to β_1 using Leibnitz's rule, and we then obtain from (5.7) that

$$\beta_1 = \left[\sum_{j=1}^{n} z_{2j}^{g} t_j / \alpha_1 + \sum_{j=1}^{n} z_{2j}^{g} \delta_j / \alpha_1 - \sum_{j=1}^{n} \frac{t_j^{\alpha_1} e^{-t_j / \beta_1}}{\alpha_1 \beta_1^{\alpha_1 - 1} \Gamma\left(\alpha_1, t_j / \beta_1\right)} \right] \qquad (5.8)$$

The RHS of (5.8) can be evaluated at the current parameter value to obtain the updated parameter estimate $\beta_i^{(g+1)}$.

Upon expanding the incomplete gamma function as an infinite series, then differentiating and simplifying the expression, (5.6) can be expressed as

$$\frac{\partial Q}{\partial \alpha_1} = \sum_{j=1}^{n} z_{2j}^{g} \partial_j \left[\log t_j - \log \beta_1 - \Psi(\alpha_1) \right]$$

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

$$\left. + e^{-t_j / \beta_1} \sum_{p=0}^{\infty} \frac{(t_j / \beta_1)^{\alpha_1 + p} \Psi(\alpha_1 + p + 1)}{\Gamma(\alpha_1 + p + 1)} / \left\{ 1 - e^{-t_j / \beta_1} \sum_{p=0}^{\infty} \frac{(t_j / \beta_1)^{\alpha_1 + p}}{\Gamma(\alpha_1 + p + 1)} \right\} \right] \quad (5.9)$$

Equating (5.9) to zero, the equation can be solved numerically for α_i to obtain the current estimate $\alpha_1^{(g+1)}$ by using $\beta_1^{(g+1)}$ for β_1 .

The E-step on the *(g+1)*th iteration is to update the current conditional expectation of z_{ij} , given the observed data, using the current model parameters fit,

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

$$+ \hat{\pi}_2^{(g)} \left[f\ (t_j; \hat{\alpha}_1^{(g)}, \hat{\beta}_1^{(g)}) \right]^{\delta_j} \left[S\ (t_j; \hat{\alpha}_1^{(g)}, \hat{\beta}_1^{(g)}) \right]^{1-\delta_j} \quad (5.10)$$

$$\left. + \hat{\pi}_3^{(g)} \left[f\ (t_j; \hat{\alpha}_2^{(g)}, \hat{\beta}_2^{(g)}) \right]^{\delta_j} \left[S\ (t_j; \hat{\alpha}_2^{(g)}, \hat{\beta}_2^{(g)}) \right]^{1-\delta_j} \right]$$

Again, differentiating equation (5.3) with respect to the parameters α_2, β_2 yields

$$\frac{\partial Q}{\partial \alpha_2} = \sum_{j=1}^{n} z_{3j}^{(g)} \delta_j \left[\frac{1}{\alpha_2} - \log \beta_2 + \log t_j \right] - \sum_{j=1}^{n} z_{3j}^{(g)} \left(\frac{t_j}{\beta_2} \right)^{\alpha_2} \left(\log t_j - \log \beta_2 \right) \quad (5.11)$$

$$\frac{\partial Q}{\partial \beta_2} = \sum_{j=1}^{n} z_{3j}^{(g)} \left[-\delta_j \frac{\alpha_2}{\beta_2} + \alpha_2 t_j^{\alpha_2} \beta_2^{-\alpha_2 - 1} \right] \quad , \quad (5.12)$$

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

The system of equations (5.12) can be written as

$$\beta_2 = \exp \left(\frac{1}{\alpha_2} \log \frac{\sum_{j=1}^{n} z_{3j}^{(g)} t_j^{\alpha_2}}{\sum_{j=i}^{n} z_{3j}^{(g)} \delta_j} \right) \quad (5.13)$$

Plug equations (5.13) back to (5.11) to obtain

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

$$\sum_{j=1}^{n} z_{3j}^{(g)} t_j^{\alpha_2} \frac{\sum_{j=1}^{n} z_{3j}^{(g)} \delta_j}{\sum_{j=1}^{n} z_{3j}^{(g)} t_j^{\alpha}2} \left(\log t_j - \frac{1}{\alpha_2} \log \frac{\sum_{j=1}^{n} z_{3j}^{(g)} t_j^{\alpha_2}}{\sum_{j=1}^{n} z_{3j}^{(g)} \delta_j} \right) = 0 \quad (5.14)$$

Then equations (5.14) can be solved to obtain the estimates for α_2. Plug the estimates of α_2 back to equations (5.13) to obtain the estimates for β_2. This completes the M-step. The E-step on the *(g+1)*th iteration is to update the current conditional expectation of Z_{3j}, given the observed data, using the current model parameters fit,

$$\begin{aligned}\hat{Z}_{3j}^{(g+1)} = \hat{\pi}_3^{(g)}\left[f\ (t_j;\hat{\alpha}_2^{(g)},\hat{\beta}_2^{(g)}\right]^{\delta_j}\left[S\ (t_j;\hat{\alpha}_2^{(g)},\hat{\beta}_2^{(g)}\right]^{1-\delta_j} / \Big[\hat{\pi}_1^{(g)}\left[f\ (t_j;\hat{\lambda}^{(g)}\right]^{\delta_j}\left[S\ (t_j;\hat{\lambda}\right]^{1-\delta_j} \\ +\hat{\pi}_2^{(g)}\left[f\ (t_j;\hat{\alpha}_1^{(g)},\hat{\beta}_1^{(g)}\right]^{\delta_j}\left[S\ (t_j;\hat{\alpha}_1^{(g)},\hat{\beta}_1^{(g)}\right]^{1-\delta_j} \\ +\hat{\pi}_3^{(g)}\left[f\ (t_j;\hat{\alpha}_2^{(g)},\hat{\beta}_2^{(g)}\right]^{\delta_j}\left[S\ (t_j;\hat{\alpha}_2^{(g)},\hat{\beta}_2^{(g)}\right]^{1-\delta_j}\Big]\end{aligned} \tag{5.15}$$

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

The EM procedure outlined earlier was employed on both of the simulated and real data to evaluate the parameters of Model 2. These derivations and equations were transformed into computer coding. The R statistical software (Team, 2005) was employed to develop functions to generate survival data of the parametric mixture model of Exponential, Gamma and Weibull distributions. EM functions were developed to estimate the parameters of the Model 2. The R functions regarding the estimation are presented in Appendix C. All the computations were performed using R statistical software version 3.0.2 (2013-09-25). Also among the R statistical software packages used are package *survival* developed by (Therneau ,1999) and package *Mixtools* developed by (Young, et al. ,2007).

5.3 Validation of the performance of Model 2 Using Simulated Data

Survival data of three component parametric survival mixture model of the Exponential, Gamma and Weibull distributions were generated to represent Model 2. Two validation procedures were considered to analyse the simulated data. The first was to validate the performance of Model 2 (E_G_W) using simulated data from three different sample sizes, three different censoring percentages and three mixing probabilities in ascending order. The second was to validate the performance of Model 2 (E_G_W), by simulating data from three different samples, three different censoring percentages and three mixing probabilities in descending order. The three sample sizes, the three censoring percentages and the mixing proportions employed in section 4.3 of Chapter Four were used to simulate the survival data for Model 2.

5.3.1 Model 2 with Mixing Probabilities in Ascending Order

Survival data for Model 2 were generated based on mixture model of three components of mixture model of the Exponential, Gamma and Weibull distributions. The parameters of the first component of Exponential distribution are $\lambda = 1.5$, the parameters for the second component of Gamma distribution are ($\alpha_1 = 5$, $\beta_1 = 2$) and the parameters of the third component of Weibull distribution are ($\alpha_2 = 9$, $\beta_2 = 10$). Based on these three components of the Exponential, Gamma and Weibull distributions, survival data were generated for the three different sample sizes (100, 200 and 500) each with three different censoring percentages (10%, 20% and 40%). The mixing probabilities employed were in the ascending order (10%, 20% and 50%).

Three sets of survival data of sample size of 100, 200 and 500 observations each. The same samples size were generated from the Uniform distribution for the censored time C with (0,b), where the value of b depends solely on the percentage of the observations that are censored. In this study 10%, 20% and 40% censoring observations were considered for each of the sample generated. t_j =min(T_j,C_j) was taken as the minimum of the survival time and the censored time of the observed time T where is as in (4.33). The postulated Model 2 was formed by substituting the values of the parameters in equation (5.1), which is expressed as

$$f(t) = 0.1 \times f_E(t; \lambda = 1.5) + 0.4 \times f_G(t; \alpha_1 = 5, \beta_1 = 2) + 0.5 \times f_W(t; \alpha_2 = 9, \beta_2 = 10)\ , \quad (5.2)$$

where the density functions f_E, f_G and f_W represent the Exponential, the Gamma and the Weibull probability density functions respectively.

5.3.1.1 Sample of Size 100 observations

Survival data of size 100 with 10%, 20% and 40% censoring observations were generated and used to estimate the parameters of the postulated Model 2 by employing the EM. The estimates of the parameters together with the parameters of the postulated models were reported.

Figures 5.1, 5.2 and 5.3 display the probability density function of simulated data of Model 2, with 100 observations and 10%, 20% and 40% censored observations respectively, and the probability density functions of pure classical parametric survival models (E, G and W) corresponding to each component of Model 2.

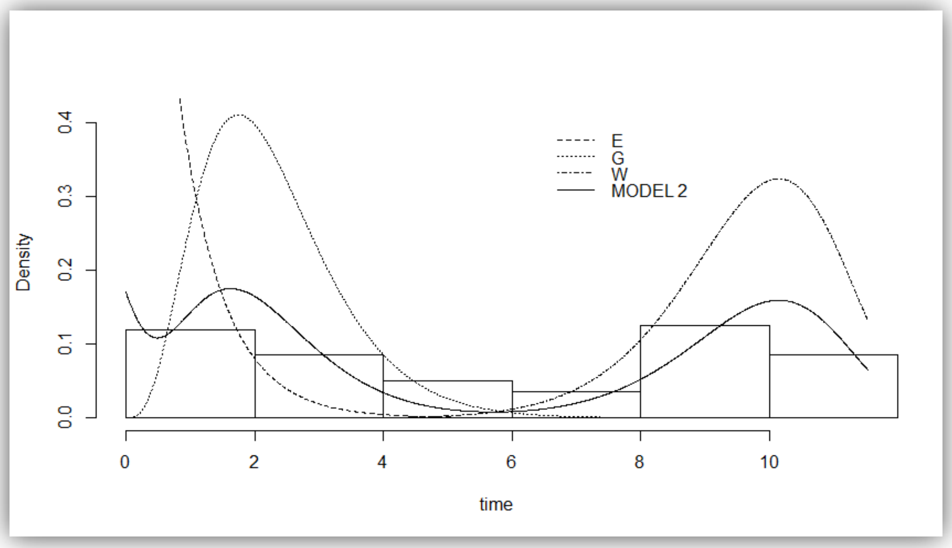


Figure 5.1 Probability Density Function of the Simulated Data of Model 2 with 100 Observations and 10% Censoring.

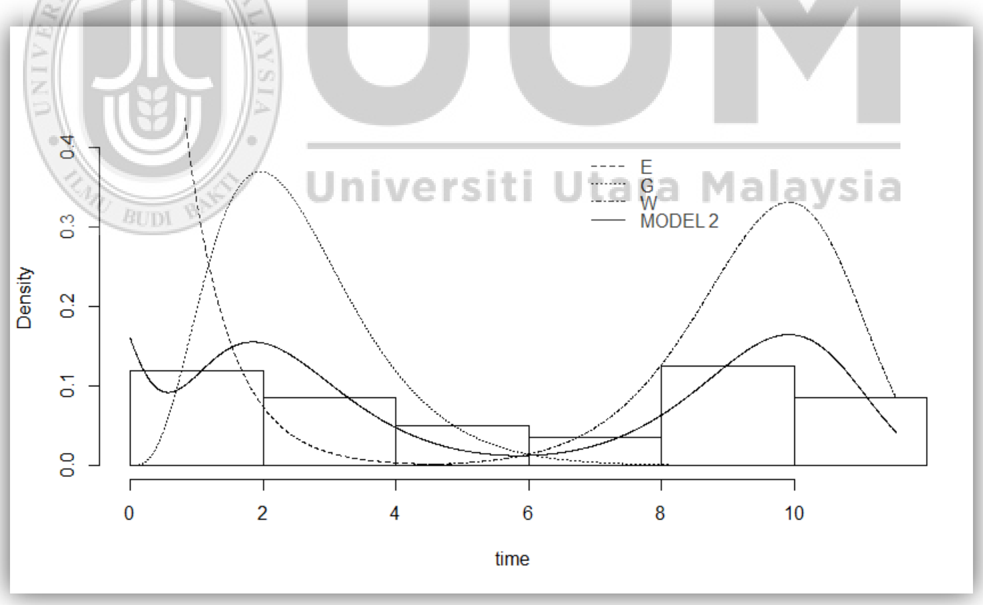


Figure 5.2 Probability Density Function of the Simulated Data of Model 2 with 100 Observations and 20% Censored Observations.

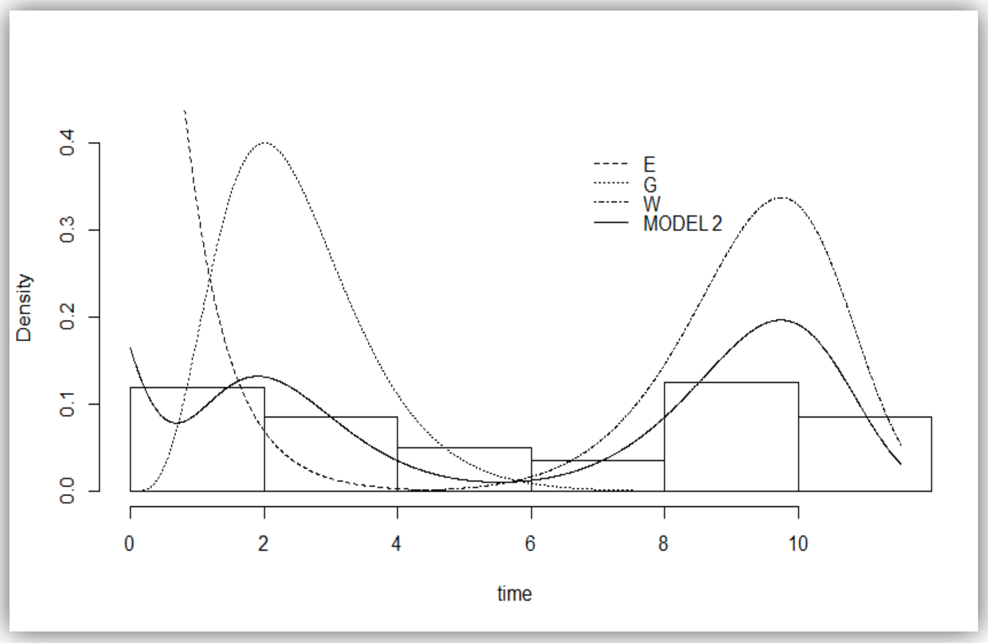


Figure 5.3 Probability Function of the Simulated Data of Model 2 of size 100 Observations and 40% Censored Observations.

It can be seen that Model 2 fits the simulated data better than the individual pure classical parametric survival models which indicates that the simulated data is better modelled by Model 2 than the pure classical parametric survival model.

Table 5.1 displays the result of the estimated parameters of the simulated data of 100 observations with 10%, 20% and 40% censored observations respectively.

Table 5.1

The Estimated Parameters of the Simulated Data of Postulated Model 2 with 10% Censoring Observations

<table>
<tr><th colspan="8">Model 2 with sample size 100 observations and 10% censoring</th></tr>
<tr><td>Parameter</td><td>π_1</td><td>π_2</td><td>λ</td><td>α_1</td><td>α_2</td><td>β_1</td><td>β_2</td></tr>
<tr><td>Postulate</td><td>0.10</td><td>0.40</td><td>1.5</td><td>5</td><td>9</td><td>2</td><td>10</td></tr>
<tr><td>Estimates</td><td>0.12</td><td>0.39</td><td>1.44</td><td>4.40</td><td>9.00</td><td>1.95</td><td>10.28</td></tr>
<tr><td colspan="8">Model 2 with sample size 100 observations and 20% censoring</td></tr>
<tr><td>Parameter</td><td>π_1</td><td>π_2</td><td>λ</td><td>α_1</td><td>α_2</td><td>β_1</td><td>β_2</td></tr>
<tr><td>Postulate</td><td>0.10</td><td>0.40</td><td>1.5</td><td>5</td><td>9</td><td>2</td><td>10</td></tr>
<tr><td>Estimates</td><td>0.10</td><td>0.40</td><td>1.51</td><td>4.47</td><td>9.00</td><td>1.77</td><td>10.05</td></tr>
<tr><td colspan="8">Model 2 with sample size 100 observations and 40% censoring</td></tr>
<tr><td>Parameter</td><td>π_1</td><td>π_2</td><td>λ</td><td>α_1</td><td>α_2</td><td>β_1</td><td>β_2</td></tr>
<tr><td>Postulate</td><td>0.10</td><td>0.40</td><td>1.5</td><td>5</td><td>9</td><td>2</td><td>10</td></tr>
<tr><td>Estimates</td><td>0.10</td><td>0.31</td><td>1.55</td><td>5.21</td><td>9.00</td><td>2.10</td><td>9.87</td></tr>
</table>

The parameters of the three sets of the simulated data of size 100 observations were all estimated successfully. The values of the parameter were close to the postulated parameters used in the data generation.

The hazard functions of the three simulated data corresponding to the 10%, 20% and 40% censoring were presented in Figure 5.4.

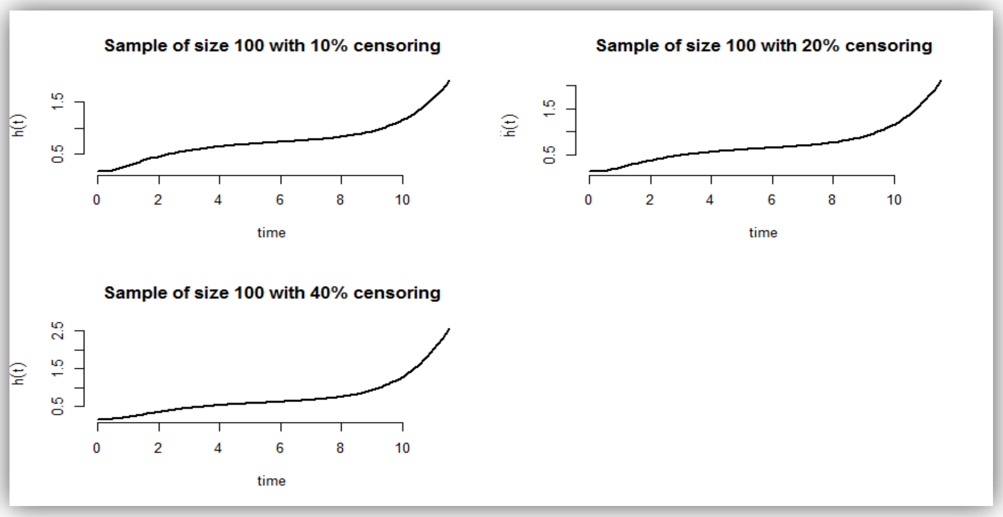


Figure 5.4 The Hazard Functions of the Simulated Data of Size 100 Corresponding to 10%, 20% and 40% Censored Observations.

The hazard function of simulated data of size 100 observations with 10% of the observations censored is higher than that of 20% and 40% censoring. As the number of censored observations increases the hazard tends to be lower and lower.

The simulation of the three set of generated data were repeated 300 time to check the consistency and stability of the EM in estimating the model parameters. The averages, the MSE and RMSE of estimated parameters of the postulated Model 1 are listed in Table 5.2.

Table 5.2

The Repeated Simulation of Set of 100 Observations

Model 1 with sample size 100 and 10% censoring							
Parameters	π_1	π_2	λ	α_2	α_3	β_1	β_2
Postulates	0.1	0.4	1.5	5	9	2	10
Estimates	0.11	0.40	1.39	4.97	9.00	1.98	10.27
MSE	9.56e-8	1.25e-6	1.04e-5	1.60e-3	0.00e+0	2.49e-6	1.62e-4
RMSE	3.09e-4	1.12e-3	1.02e-3	4.00e-2	0.00e+0	1.58e-3	1.27e-2
Model 1 with sample size 100 and 20% censoring							
parameters	π_1	π_2	λ	α_2	α_3	β_1	β_2
postulated	0.1	0.4	1.5	5	9	2	10
Estimates	0.11	0.40	1.33	5.03	9.00	2.00	10.00
MSE	8.92e-7	1.00e-5	9.03e-5	1.79e-3	0.00e+0	3.14e-6	8.30e-4
RMSE	9.44e-4	3.16e-3	9.50e-3	4.24e-2	0.00e+0	1.77e-3	9.11e-3
Model 1 with sample size 100 and 40% censoring							
Parameters	π_1	π_2	λ	α_2	α_3	β_1	β_2
Postulated	0.1	0.4	1.5	5	9	2	10
Estimates	0.11	0.37	1.31	4.74	9.00	2.07	10.41
MSE	7.61e-7	1.88e-6	9.29e-5	5.71e-3	0.00e+0	4.67e-6	1.77e-4
RMSE	8.72e-4	1.37e-3	9.64e-3	7.55e-2	0.00e+0	2.16e-3	1.33e-2

The averages of the parameters are close to the parameters of the postulated parametric survival mixture model with MSE and RMSE relatively small, which suggests that, the EM performed consistently in estimating the parameters. Generally, the value of MSE for the 10% censoring tend to be smaller than that of 20% and 40% which shows that the parameters were estimated better with 10% censored observation. The estimation of the mixing probabilities seems to be better with small censoring percentage.

5.3.1.2 Sample of Size 200 observations

Three sets of survival data of size 200 observations with 10%, 20% and 40% censored observations, respectively, were generated. The data were used to estimate the parameters of the postulated Model 2 by the EM. The estimated parameters corresponding to each set of data and the true parameters of the postulated models were reported.

The probability density function of simulated data of sample of size 200 observations with 10%, 20% and 40% censored observations was presented in Figure 5.5, 5.6 and 5.7 respectively. The graphs also, display the probability density functions of pure classical parametric survival models E, G and W corresponding to each component of Model 2.

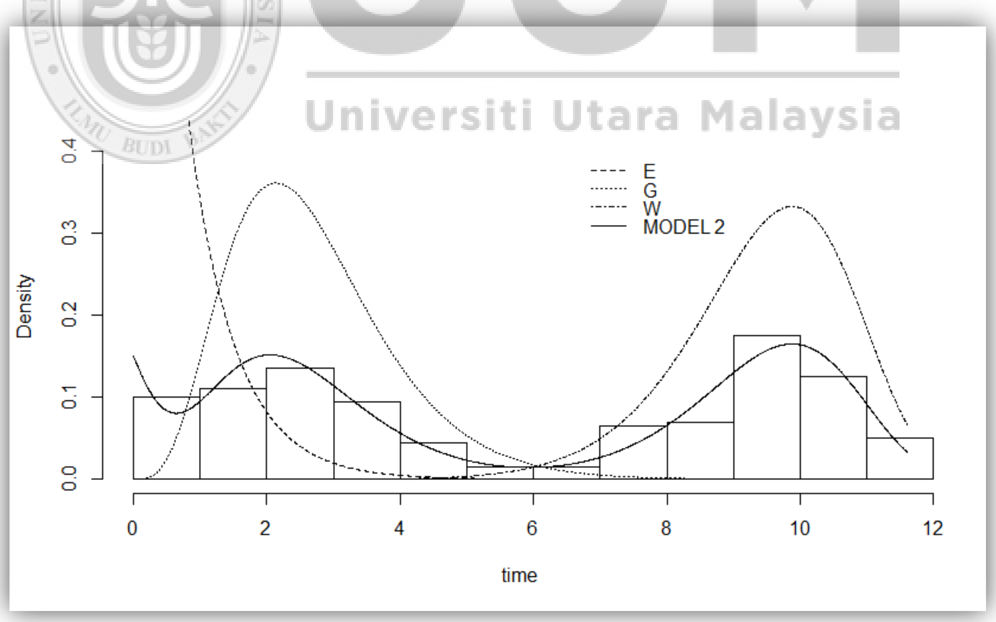


Figure 5.5 Probability Density Function of the Simulated Data of Size 200 Observations and 10% Censoring.

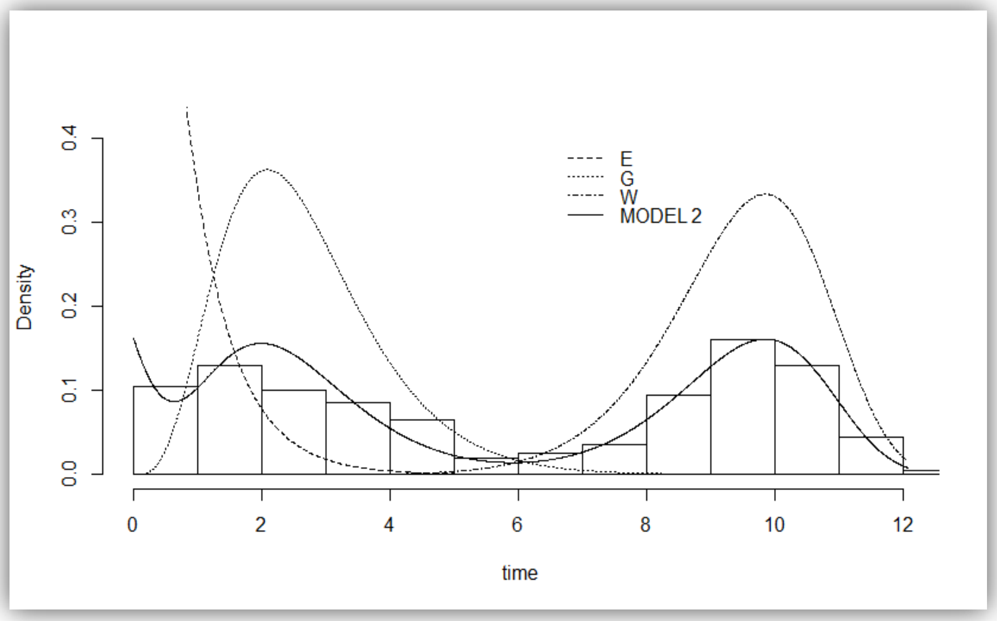


Figure 5.6 Probability Density Function of the Simulated Data of Size 200 and 20% Censored Observations.

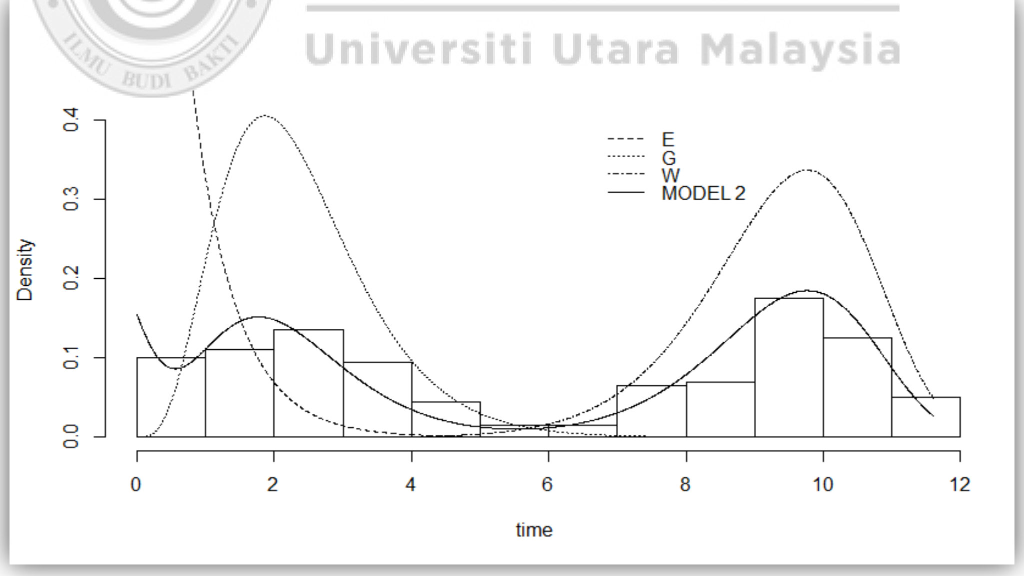


Figure 5.7 Probability Density Function of the Simulated Data of Size 200 Observations and 40% Censoring.

It can be seen that Model 2 fits the simulated data better than the individual pure classical parametric survival models which indicates that the simulated data is better modelled by Model 2 than the pure classical parametric survival model.

The estimated parameters of the set of simulated data of sample of size 200 with 10%, 20% and 40% censored observations respectively were presented in Table 5.3.

Table 5.3

The Estimated Parameters the Simulated Data of size 200 with 10% Censoring Observations

Model 2 with sample size 200 observations and 10% censoring							
Parameter	π_1	π_2	λ	α_1	α_2	β_1	β_2
Postulate	0.10	0.40	1.5	5	9	2	10
Estimates	0.11	0.40	1.43	4.92	9.00	1.83	10.01
Model 2 with sample size 200 observations and 20% censoring							
Parameter	π_1	π_2	λ	α_1	α_2	β_1	β_2
Postulate	0.10	0.40	1.5	5	9	2	10
Estimates	0.11	0.40	1.47	4.75	9.00	1.80	9.98
Model 2 with sample size 500 observations and 40% censoring							
Parameter	π_1	π_2	λ	α_1	α_2	β_1	β_2
Postulate	0.10	0.40	1.5	5	9	2	10
Estimates	0.10	0.35	1.56	4.76	9.00	2.02	9.89

The estimated parameters of the three set of the simulated data are all close to the postulated parameters used in the data generation. From Table 5.5 it can be observed that the parameter for the simulated set of data with 10% censoring are more closer the postulated parameters compared to that of the 20% and 40% censoring observations.

The hazard functions of the three simulated data of size 200 observations corresponding to the 10%, 20% and 40% censoring percentages were presented in Figure 5.8.

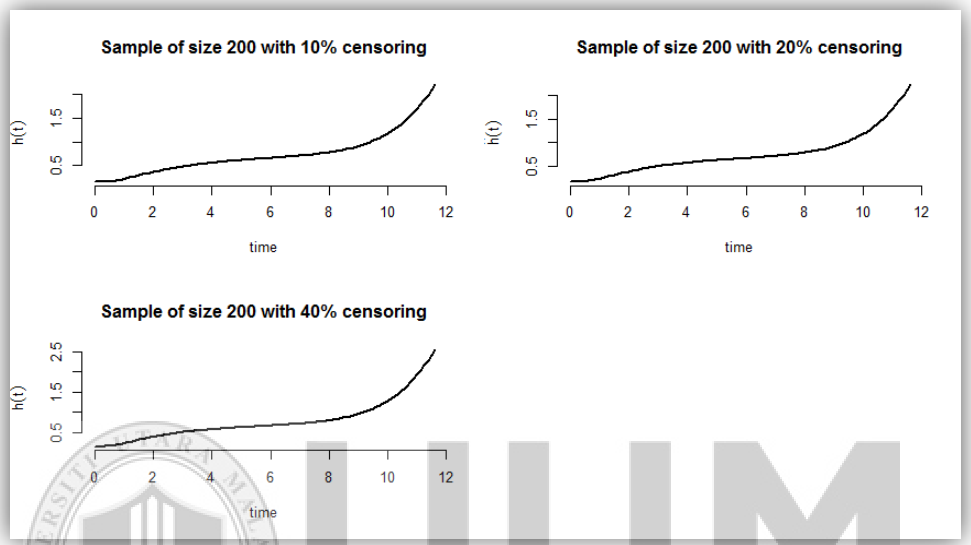


Figure 5.8 The Hazard Functions of the Simulated Data of Size 200 Corresponding to 10%, 20% and 40% Censored Observation.

The hazard function of the set of simulated data consisting of 200 observations with 10% censoring observation is higher when compared with that of 20% and 40% censoring. As the number of censored observations increases the hazard tends to be lower and lower.

The simulation of the three set of generated data of 200 observations with 10%, 20% and 40% censoring observations were repeated 300 times to check the consistency and stability of the EM in estimating the model parameters. The averages, the mean

square errors and root mean square error of estimated parameters of the postulated Model 2 are listed in Table 5.4.

Table 5.4

The Repeated Simulation of Set of 200 Observations

Model 2 with sample size 200 and 10% censoring							
parameters	π_1	π_2	λ	α_2	α_3	β_1	β_2
Postulates	0.1	0.4	1.5	5	9	2	10
Estimates	0.11	0.40	1.33	4.98	9.00	1.99	10.08
MSE	4.29e-7	4.94e-7	3.56e-5	8.58e-4	0.00e+0	1.30e-6	3.44e-5
RMSE	6.55e-4	7.03e-4	5.96e-3	2.92e-2	0.00e+0	1.14e-3	5.87e-3
Model 2 with sample size 200 and 20% censoring							
parameters	π_1	π_2	λ	α_2	α_3	β_1	β_2
postulated	0.1	0.4	1.5	5	9	2	10
Estimates	0.11	0.40	1.34	4.94	9.00	2.00	10.02
MSE	4.55e-7	5.67e-7	4.47e-5	1.18e-3	0.00e+0	1.48e-6	4.48e-5
RMSE	6.75e-4	7.53e-4	6.68e-3	3.43e-2	0.00e+0	1.22e-3	6.69e-3
Model 2 with sample size 200 and 40% censoring							
parameters	π_1	π_2	λ	α_2	α_3	β_1	β_2
postulated	0.1	0.4	1.5	5	9	2	10
estimates	0.10	0.37	1.44	5.17	9.00	2.16	9.96
MSE	4.49e-7	8.50e-7	6.72e-5	1.18e-3	0.00e+0	1.14e-5	4.19e-5
RMSE	6.70e-4	9.22e-4	8.20e-3	3.43e-2	0.00e+0	3.38e-3	6.47e-3

The averages of the parameters are close to the parameters of the postulated parametric survival mixture model with mean square errors relatively small, which suggests that, the EM performed consistently in estimating the parameters. The MSE value of the sample with 10% censoring were smaller than that of 20% and 40% censoring. The estimation of the parameter was better with small censoring percentage. Also, the mixing probabilities were better estimated using samples with smaller censoring percentage (10%).

5.3.1.3 Sample of Size 500 observations

Three sets of survival data of size 500 observations with 10%, 20% and 40% censored observations, respectively, were generated and employed to estimate the parameters of the postulated Model 2 by using the EM. The estimated parameters of corresponding to each set of data and the parameters of the postulated models were reported. The probability density function of simulated data of sample size 500 observations with 10%, 20% and 40% censored observations were presented in Figure 5.9, 5.10 and 5.11 respectively. The graphs also, display the probability density functions of pure classical parametric survival models E, G and W corresponding to each component of Model 2.

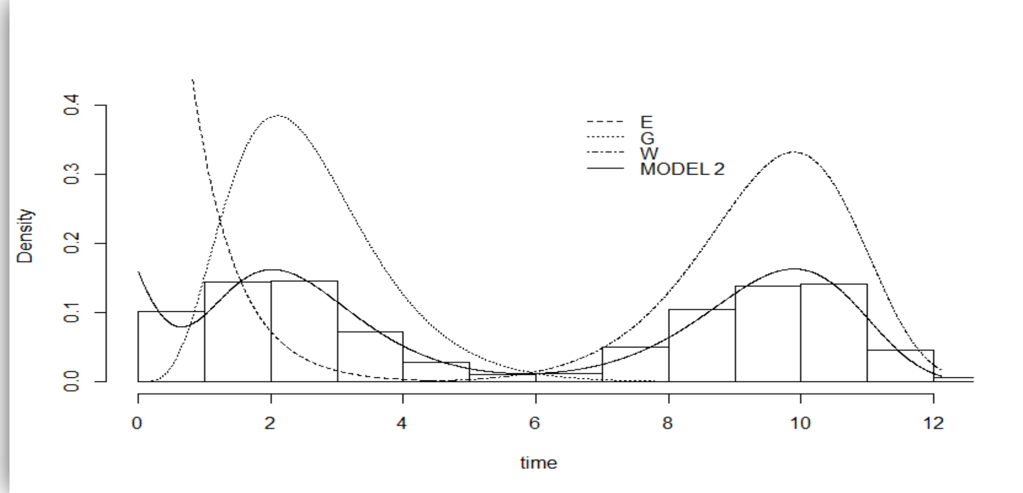


Figure 5.9 Probability Density Function of the Simulated Data of Size 500 Observations and 10% Censoring.

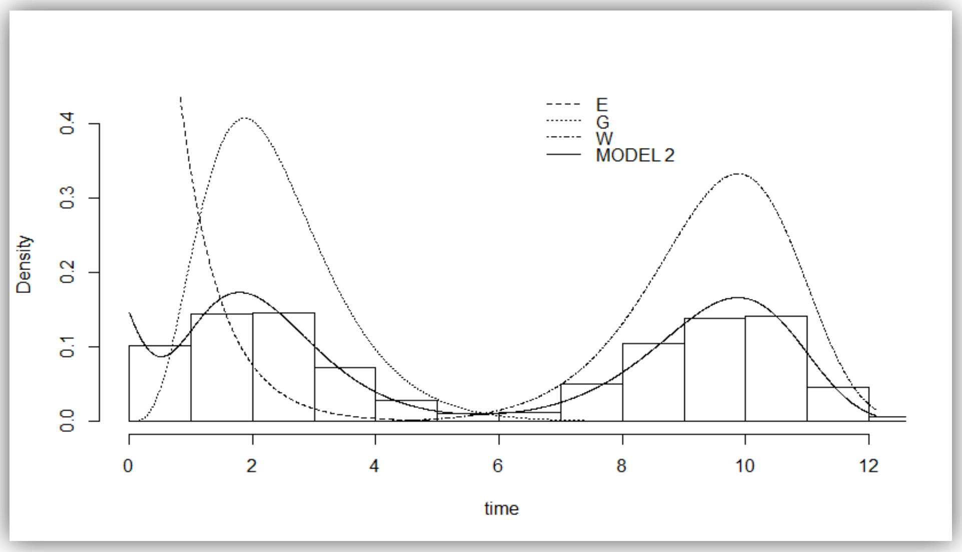


Figure 5.10 Probability Density Function of the Simulated Data of Size 500 Observations and 20% Censoring.

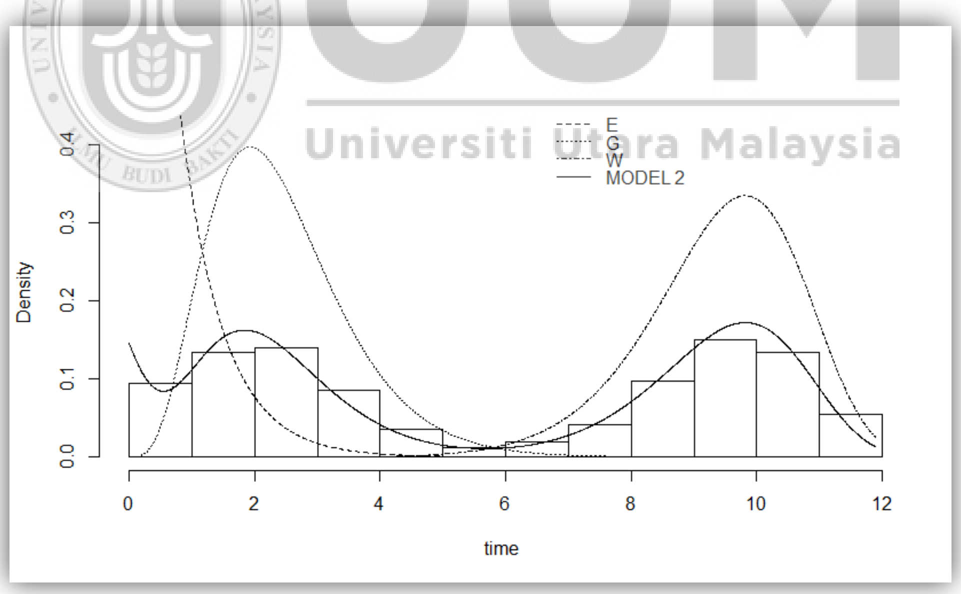


Figure 5.11 Probability Density Function of the Simulated Data of Size 500 Observations and 40% Censoring.

It can be seen that Model 2 fits the simulated data better than the individual pure classical parametric survival models which indicates that the simulated data is better modelled by Model 2 than the pure classical parametric survival model.

The estimated parameters of the set of simulated data of sample of size 500 with 10%, 20% and 40% censored observations were presented in Table 5.5. The estimated parameters are close to the values of the parameters of the postulated model.

Table 5.5

The Estimated Parameters the Simulated Data of size 500 with 10% Censoring Observations

Model 2 with sample size 500 observations and 10% censoring							
Parameter	π_1	π_2	λ	α_1	α_2	β_1	β_2
Postulate	0.10	0.40	1.50	5	9	2	10
Estimates	0.11	0.40	1.52	5.25	9.00	2.03	10.02
Model 2 with sample size 500 observations and 20% censoring							
Parameter	π_1	π_2	λ	α_1	α_2	β_1	β_2
Postulate	0.10	0.40	1.50	5	9	2	10
Estimates	0.10	0.40	1.49	4.81	9.00	2.04	10.00
Model 2 with sample size 500 observations and 40% censoring							
Parameter	π_1	π_2	λ	α_1	α_2	β_1	β_2
Postulate	0.10	0.40	1.50	5	9	2	10
Estimates	0.10	0.39	1.48	4.84	9.00	1.99	9.94

It can be seen that the estimation of the parameters improved with the increase in the sample size from 100 through to 500 observations.

The hazard functions of the three simulated data of size 500 observations corresponding to the 10%, 20% and 40% censoring percentages were presented in Figure 5.12.

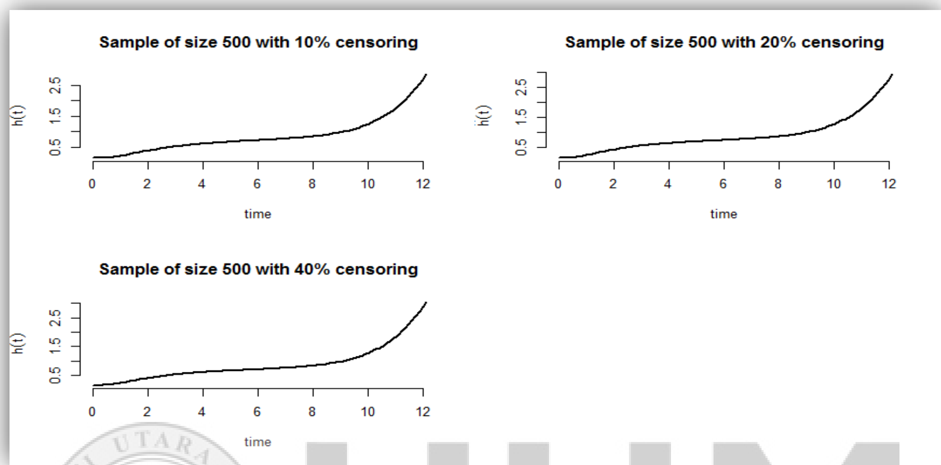


Figure 5.12 The Hazard Functions of the Simulated Data of Size 500 Corresponding to 10%, 20% and 40% Censored Observation.

The hazard function of the set of simulated data consisting of 500 observations with 10% censoring observation is higher when compared with that of 20% and 40% censoring. As the number of censored observations increases the hazard tends to be lower and lower.

The simulation of the three set of generated data of 500 observations with 10%, 20% and 40% censoring observations were repeated 300 times to check the consistency and stability of the EM in estimating the model parameters. The averages, the mean square errors and root mean square error of estimated parameters of the postulated Model 2 are listed in Table 5.6.

Table 5.6

The Repeated Simulation of Set of 500 Observations

Model 2 with sample size 500 and 10% censoring							
parameters	π_1	π_2	λ	α_2	α_3	β_1	β_2
Postulates	0.1	0.4	1.5	5	9	2	10
Estimates	0.11	0.40	1.45	4.90	9.00	2.00	10.01
MSE	1.84e-7	2.27e-8	2.34e-5	3.82e-4	0.00e+0	5.51e-7	1.50e-5
RMSE	4.30e-4	1.51e-4	4.48e-3	3.82e-4	0.00e+0	7.42e-4	3.87e-3
Model 2 with sample size 500 and 20% censoring							
parameters	π_1	π_2	λ	α_2	α_3	β_1	β_2
postulated	0.1	0.4	1.5	5	9	2	10
Estimates	0.10	0.40	1.44	4.69	9.00	2.01	9.99
MSE	1.99e-7	1.78e-7	2.56e-5	4.22e-4	0.00e+0	5.80e-7	1.50e-5
RMSE	4.46e-4	4.21e-4	5.06e-3	2.05e-2	0.00e+0	7.61e-4	3.87e-3
Model 2 with sample size 500 and 40% censoring							
parameters	π_1	π_2	λ	α_2	α_3	β_1	β_2
postulated	0.1	0.4	1.5	5	9	2	10
estimates	0.09	0.39	1.36	4.64	9.00	2.04	9.96
MSE	2.01e-7	1.31e-7	2.38e-5	5.21e-4	0.00e+0	6.19e-7	1.50e-5
RMSE	4.49e-4	3.62e-4	4.88e-3	2.28e-2	0.00e+0	7.87e-4	3.88e-3

The averages of the parameters are close to the parameters of the postulated parametric survival mixture model with mean square errors relatively small, which suggests that, the EM performed consistently in estimating the parameters.

Generally, the estimated parameter for Model 2 with ascending order mixing probabilities were successfully estimated. It can be observed that the estimated parameter were better as the sample size increases. The mixing probabilities for the three sample sizes were closer to the true values of the postulated parameters when the censoring percentages were smaller.

5.3.2 Model 2 with Mixing Probabilities in Descending Order

The parameters of the components of the survival mixture employed in the simulation in section 5.3.1 were used to generate Survival data for Model 2. The data were generated based on the three different samples of sizes (100, 200 and 500 observations) each with three different censoring percentages (10%, 20% and 40%). The mixing probabilities employed were in the descending order (50%, 30% and 20%). The postulated Model 2 can be formed by substituting the values parameters in equation (5.1). The estimations of the parameters were discussed in next subsection.

5.3.2.1 Sample of Size 100 observations

Data of size 100 with 10%, 20% and 40% censoring observations were generated and used to estimate the parameters of the postulated Model 2 by employing the EM. The estimates of the parameters together with the parameters of the postulated models were reported.

Figure 5.13, 5.14 and 5.15 display the probability density function of simulated data of Model 2 respectively, with 100 observations and 10%, 20% and 40% censored observations. The probability density functions of pure classical parametric survival models (E, G and W) corresponding to each component of Model 2 were also presented in the same graph.

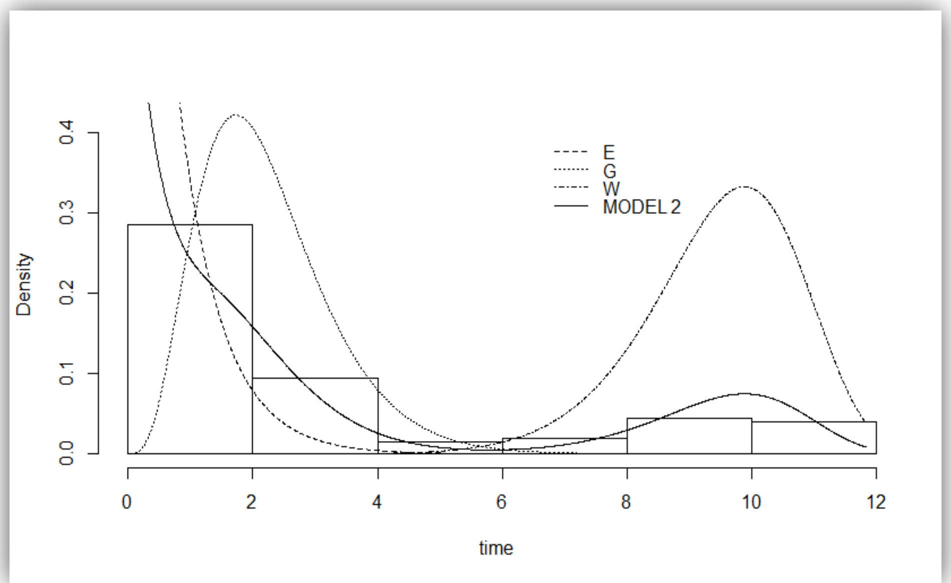


Figure 5.13 Probability Density Function of the Simulated Data of Model 2 with 100 and 10% Censored Observations.

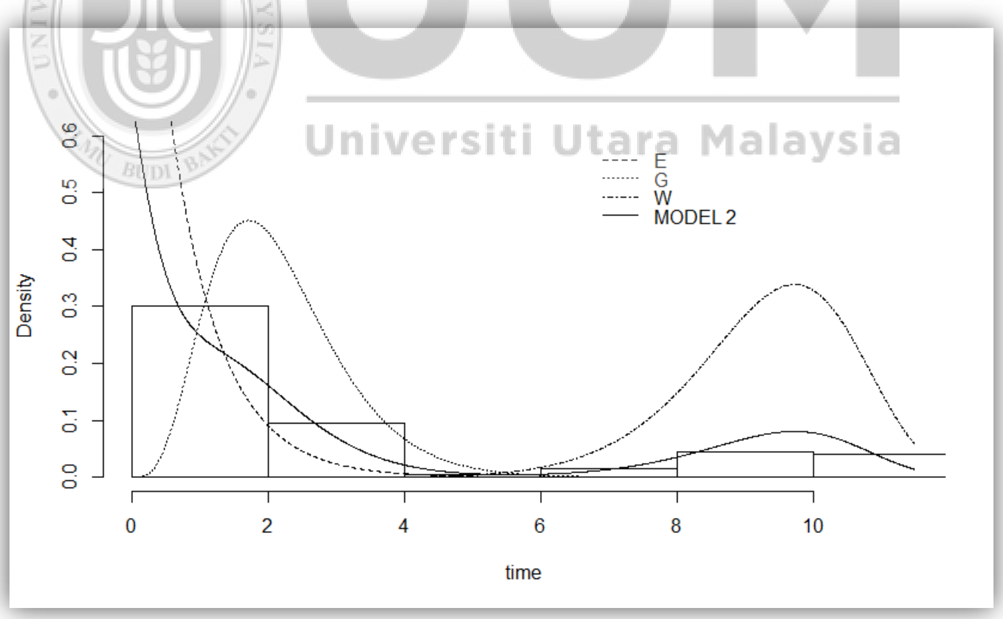


Figure 5.14 Probability Density Function of the Simulated Data of Model 2 with 100 Observations and 20% Censoring.

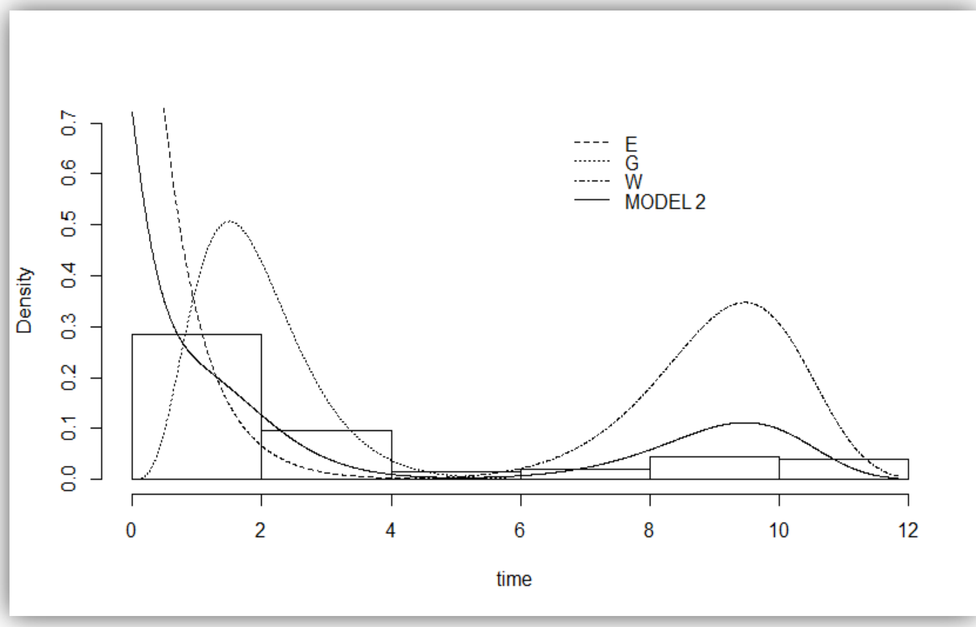


Figure 5.15 Probability Density Function of the Simulated Data of Model 2 with 100 and 40% Censored Observations.

It can be seen that Model 2 fits the simulated data better than the individual pure classical parametric survival models which indicates that the simulated data is better modelled by Model 2 than the pure classical parametric survival model.

Table 5.7 shows the values of estimated parameter corresponding to the simulated data of 100 observations with 10%, 20% and 40% censored observations respectively and the true parameters of the postulated model used in generating the data set.

Table 5.7

The Estimated Parameters the Simulated Data of size 100 with 10% Censored Observations

Model 2 with sample size 100 observations and 10% censoring							
Parameter	π_1	π_2	λ	α_1	α_2	β_1	β_2
Postulate	0.50	0.30	1.5	5	9	2	10
Estimates	0.48	0.30	1.46	4.50	9.00	2.03	10.01
Model 2 with sample size 100 observations and 20% censoring							
Parameter	π_1	π_2	λ	α_1	α_2	β_1	β_2
Postulate	0.50	0.30	1.5	5	9	2	10
Estimates	0.49	0.27	1.34	4.90	9.00	2.28	9.84
Model 2 with sample size 100 observations and 40% censoring							
Parameter	π_1	π_2	λ	α_1	α_2	β_1	β_2
Postulate	0.50	0.30	1.5	5	9	2	10
Estimates	0.45	0.23	1.60	4.78	9.00	2.53	9.59

From Table 5.7 it can be observed that the estimates of the mixing probabilities were much distorted when the censoring percentage increases.

The hazard functions of the three simulated data corresponding to the 10%, 20% and 40% censoring were presented in Figure 5.16.

The hazard function of simulated data of size 100 observations with 10% of the observations censored is higher than that of 20% and 40% censoring. As the number of censored observations increases the hazard tends to be lower and lower.

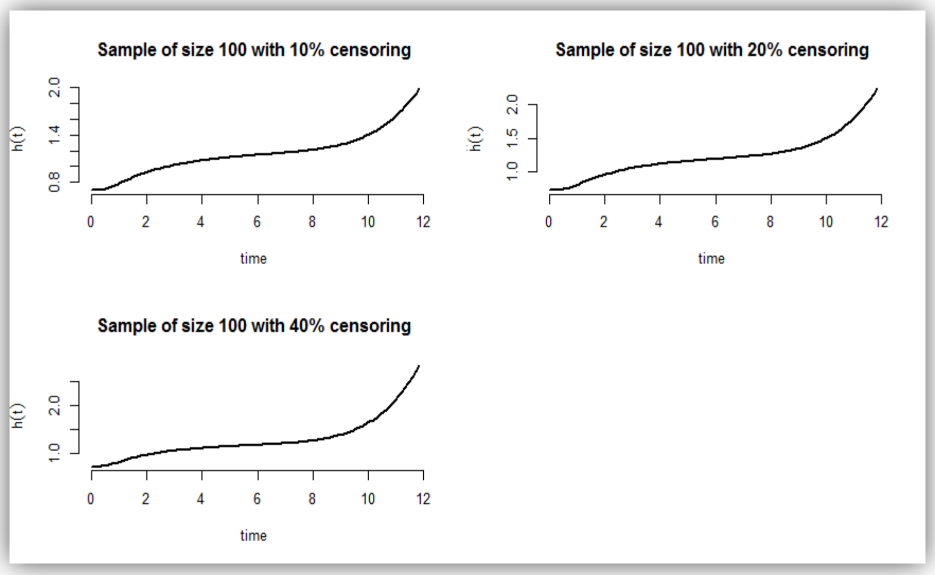


Figure 5.16 The Hazard Functions of the Simulated Data of Size 100 Corresponding to 10%, 20% and 40% Censored Observation.

The simulations of the three sets of generated data were repeated 300 times to check the consistency and stability of the EM in estimating the model parameters. The averages, the mean square errors and root mean square error of estimated parameters of the postulated Model 2 are listed in Table 5.8.

Table 5.8

The Repeated Simulation of Set of 100 Observations

Model 2 with sample size 100 and 10% censoring							
parameters	π_1	π_2	λ	α_2	α_3	β_1	β_2
Postulates	0.50	0.30	1.5	5	9	2	10
Estimates	0.49	0.29	1.48	4.23	9.00	2.06	9.88
MSE	1.95e-6	1.57e-7	3.41e-5	3.10e-3	0.00e+0	2.89e-6	2.24e-4
RMSE	1.40e-3	3.96e-4	5.58e-3	5.57e-2	0.00e+0	1.70e-3	1.50e-2
Model 2 with sample size 100 and 20% censoring							
parameters	π_1	π_2	λ	α_2	α_3	β_1	β_2
postulated	0.50	0.30	1.5	5	9	2	10
Estimates	0.49	0.27	1.53	5.41	9.00	2.13	9.78
MSE	2.40e-6	1.53e-5	4.91e-5	3.63e-3	0.00e+0	3.17e-6	1.58e-2
RMSE	1.55e-3	3.91e-3	7.03e-3	6.02e-2	0.00e+0	1.78e-3	1.58e-2
Model 2 with sample size 100 and 40% censoring							
parameters	π_1	π_2	λ	α_2	α_3	β_1	β_2
postulated	0.50	0.30	1.5	5	9	2	10
estimates	0.45	0.23	1.60	4.22	9.00	2.62	9.44
MSE	1.99e-6	1.45e-4	6.41e-5	7.05e-3	0.00e+0	1.42e-5	2.01e-2
RMSE	1.41e-3	1.20e-2	8.01e-3	8.39e-2	0.00e+0	3.77e-3	1.42e-1

The averages of the parameters are close to the parameters of the postulated parametric survival mixture model with mean square errors and root mean square relatively small, which suggests that, the EM performed consistently in estimating the parameters. The MSE value of the mixing probabilities for sample with 10% censored observations was smaller compared to samples with 20% 40% censored observations. This shows that the model performed better with sample with smaller censoring percentage.

5.3.2.2 Sample of Size 200 observations

Three sets of survival data of size 200 observations with 10%, 20% and 40% censored observations, respectively, were generated. The data were employed to estimate the parameters of the postulated Model 2 by using the EM. The estimated parameters of corresponding to each set of data and the parameters of the postulated models were reported.

The probability density function of simulated data of sample of size 200 observations with 10%, 20% and 40% censored observations were presented in Figures 5.17, 5.18 and 5.19 respectively. The graphs also, display the probability density functions of pure classical parametric survival models E, G and W corresponding to each component of Model 2.

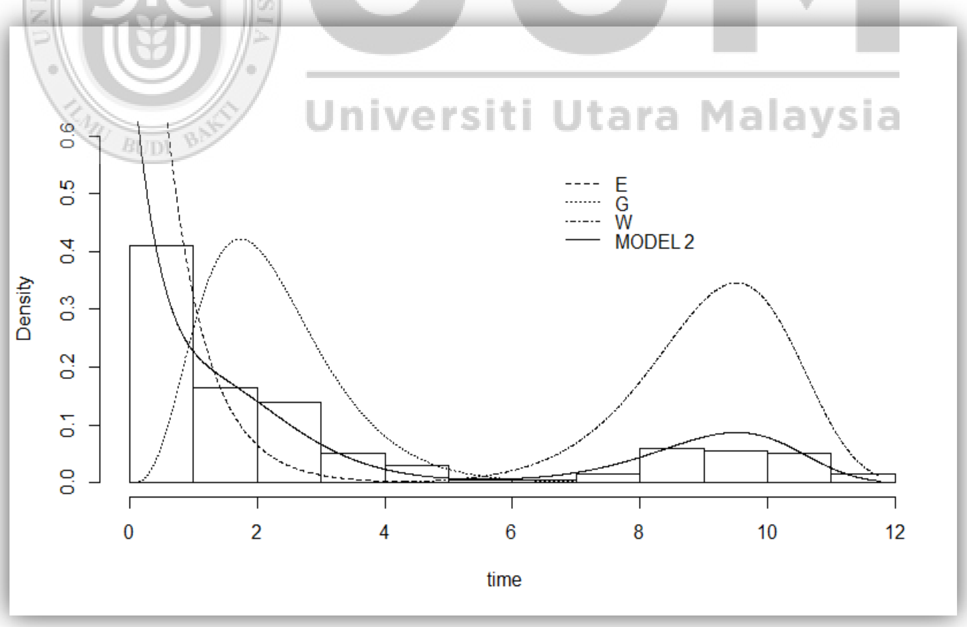


Figure 5.17 Probability Density Function of the Simulated Data of Size 200 and 10% Censored Observations.

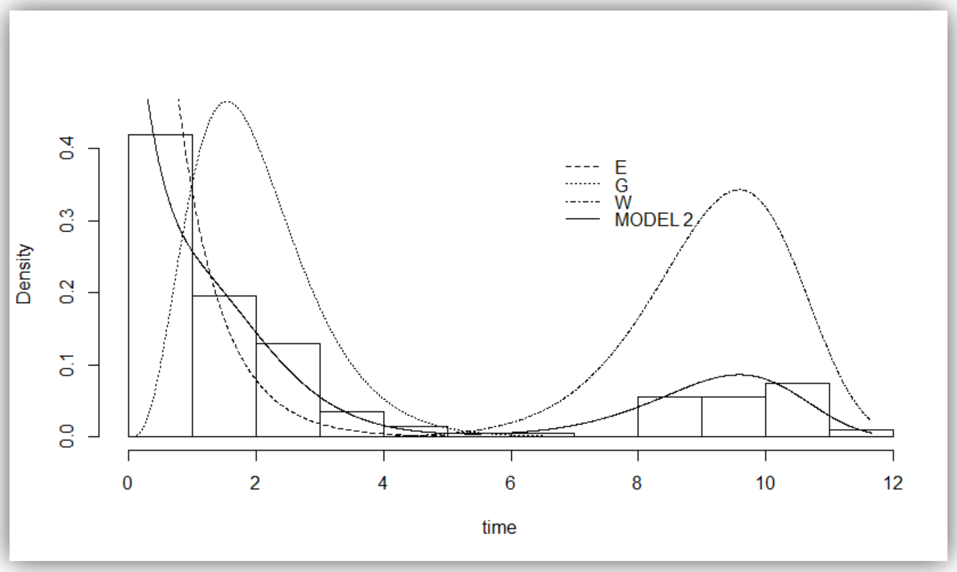


Figure 5.18 Probability Density Function of the Simulated Data of Size 200 and 20% Censored Observations.

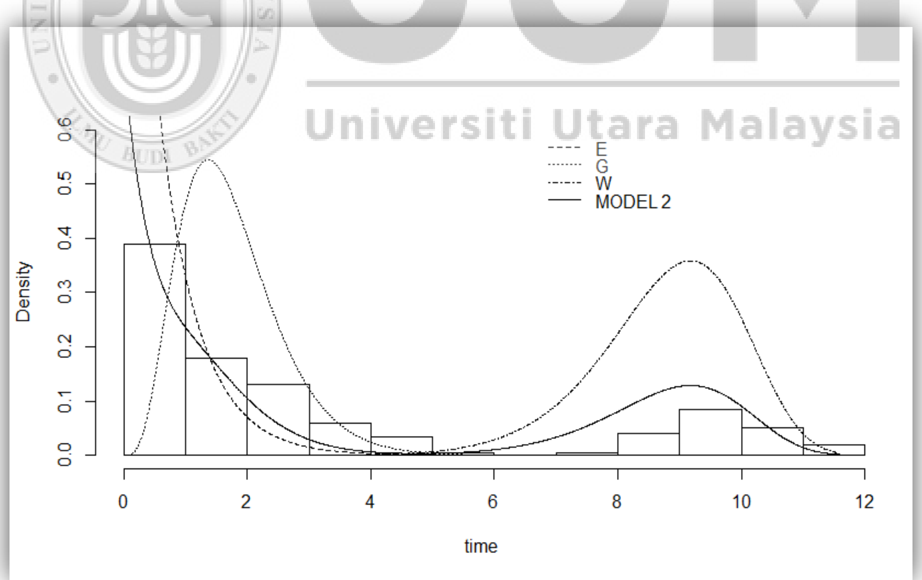


Figure 5.19 Probability Density Function of the Simulated Data of Size 200 and 40% Censored Observations.

It can be seen that Model 2 fits the simulated data better than the individual pure classical parametric survival models which indicates that the simulated data is better modelled by Model 2 than the pure classical parametric survival model.

The estimated parameters of the set of simulated data of sample of size 200 observations with 10%, 20% and 40% censored observations were presented in Table 5.9.

Table 5.9

The Estimated Parameters the Simulated Data of size 200 with 10% Censoring Observations

Model 2 with sample size 200 observations and 10% censoring							
Parameter	π_1	π_2	λ	α_1	α_2	β_1	β_2
Postulate	0.50	0.30	1.5	5	9	2	10
Estimates	0.48	0.27	1.60	4.53	9.00	2.03	9.63
Model 2 with sample size 200 observations and 20% censoring							
Parameter	π_1	π_2	λ	α_1	α_2	β_1	β_2
Postulate	0.50	0.30	1.5	5	9	2	10
Estimates	0.48	0.26	1.46	4.40	9.00	2.20	9.71
Model 2 with sample size 200 observations and 40% censoring							
Parameter	π_1	π_2	λ	α_1	α_2	β_1	β_2
Postulate	0.50	0.30	1.5	5	9	2	10
Estimates	0.46	0.18	1.55	4.63	9.00	2.66	9.30

The hazard functions of the three simulated data of size 200 observations corresponding to the 10%, 20% and 40% censoring percentages were presented in Figure 5.20.

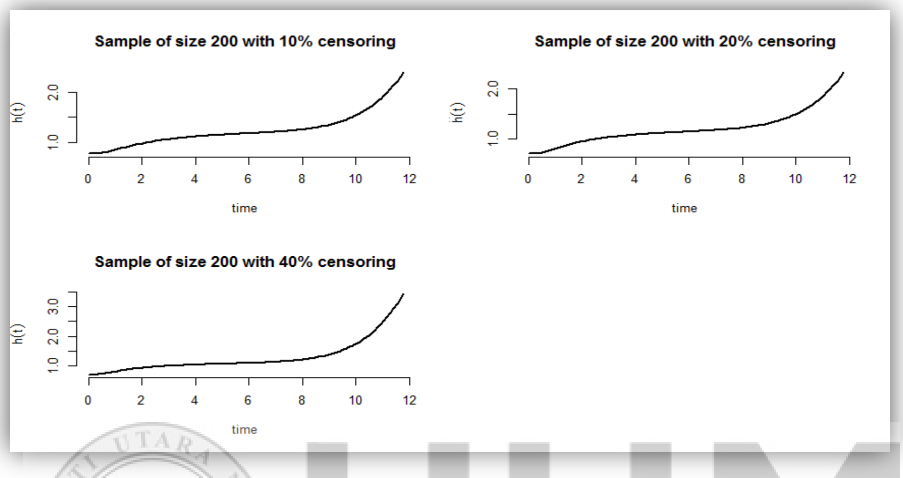


Figure 5.20 The Hazard Functions of the Simulated Data of Size 200 Corresponding to 10%, 20% and 40% Censored Observation.

The hazard function of the set of simulated data consisting of 200 observations with 10% censoring observation is higher when compared with that of 20% and 40% censoring. As the number of censored observations increases the hazard tends to be slightly lower and lower.

The simulation of the three set of generated data of 200 observations with 10%, 20% and 40% censoring observations were repeated 300 times to check the consistency and stability of the EM in estimating the model parameters. The averages, the mean square errors and root mean square error of estimated parameters of the postulated Model 2 are listed in Table 5.10.

Table 5.10

The Repeated Simulation of Set of 200 Observations

Model 2 with sample size 200 and 10% censoring

parameters	π_1	π_2	λ	α_2	α_3	β_1	β_2
Postulates	0.50	0.30	1.5	5	9	2	10
Estimates	0.48	0.28	1.50	4.40	9.00	2.02	9.77
MSE	1.40e-7	8.11e-8	2.21e-5	1.41e-3	0.00e+0	1.37e-6	1.09e-4
RMSE	3.74e-4	2.85e-4	4.70e-3	3.75e-2	0.00e+0	1.17e-3	1.05e-2

Model 2 with sample size 200 and 20% censoring

parameters	π_1	π_2	λ	α_2	α_3	β_1	β_2
postulated	0.50	0.30	1.5	5	9	2	10
Estimates	0.48	0.26	1.42	3.851	9.00	2.040	9.70
MSE	1.14e-6	6.01e-7	1.61e-6	1.86e-3	0.00e+0	1.55e-6	1.13e-4
RMSE	1.07e-3	7.75e-4	1.37e-3	4.31e-2	0.00e+0	1.25e-3	1.06e-2

Model 2 with sample size 200 and 40% censoring

parameters	π_1	π_2	λ	α_2	α_3	β_1	β_2
postulated	0.50	0.30	1.5	5	9	2	10
estimates	0.47	0.18	1.53	4.50	9.00	2.61	9.39
MSE	1.28e-6	6.05e-7	2.87e-6	2.69e-3	0.00e+0	9.38e-6	9.91e-3
RMSE	1.13e-3	7.78e-4	1.69e-3	5.19e-2	0.00e+0	3.06e-3	9.95e-2

The averages of the parameters are close to the parameters of the postulated parametric survival mixture model with mean square errors relatively small, which suggests that, the EM performed consistently in estimating the parameters. The Mixing probabilities were distorted as the censoring percentage increases with MSE values relatively high.

5.3.2.3 Sample of Size 500 observations

Three sets of survival data of size 500 observations with 10%, 20% and 40% censored observations, respectively, were generated and employed to estimate the parameters of the postulated Model 2 by using the EM. The estimated parameters of corresponding to each set of data and the parameters of the postulated models were reported.

The probability density function of simulated data of sample of size 500 observations with 10%, 20% and 40% censored observations were presented in Figures 5.21, 5.22 and 5.23 respectively. The graphs also, display the probability density functions of pure classical parametric survival models E, G and W corresponding to each component of Model 2.

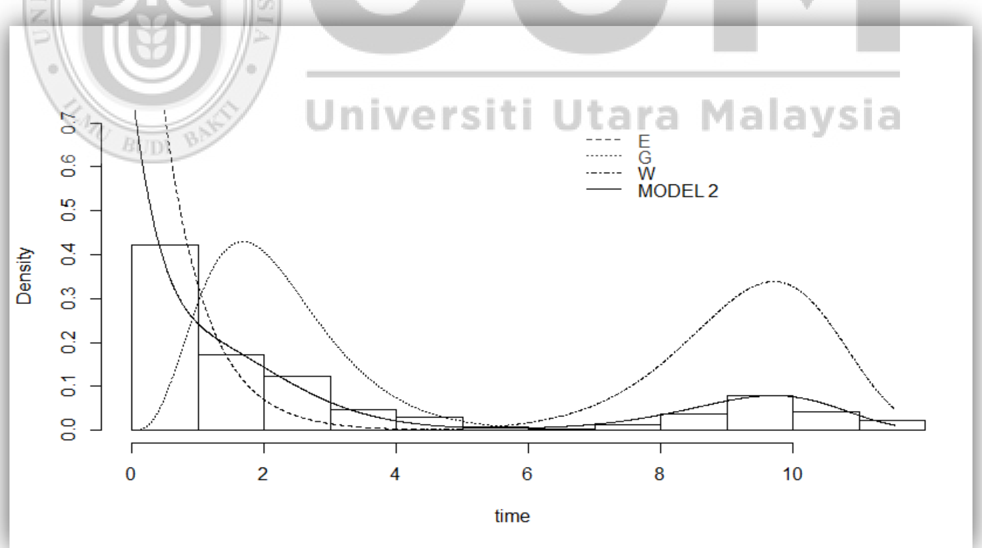


Figure 5.21 Probability Density Function of the Simulated Data of Size 500 and 10% Censored Observations.

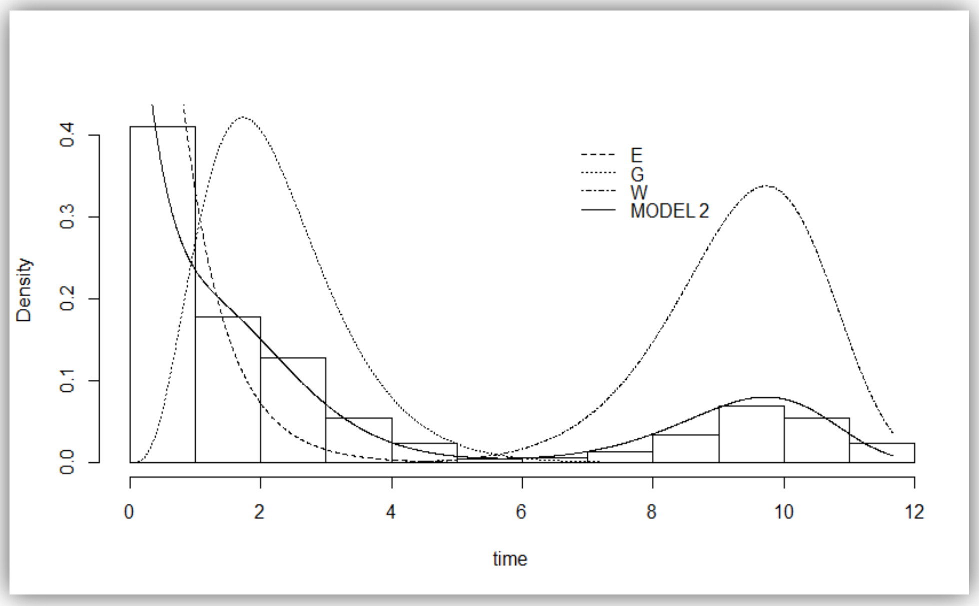


Figure 5.22 Probability Density Function of the Simulated Data of Size 500 and 20% Censored Observations.

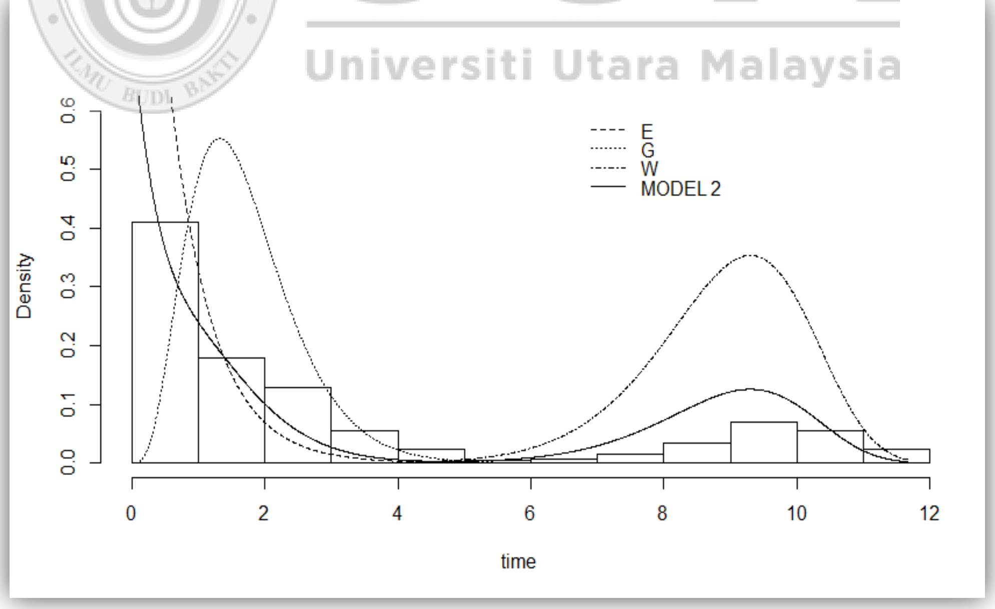


Figure 5.23 Probability Density Function of the Simulated Data of Size 500 Observations and 40% Censoring.

It can be seen that Model 2 fits the simulated data better than the individual pure classical parametric survival models which indicates that the simulated data is better modelled by Model 2 than the pure classical parametric survival model.

The estimated parameters of the set of simulated data of size 200 with 10%, 20% and 40% censored observations were presented in Table 5.11. The estimated parameters are close to the values of the parameters of the postulated model.

Table 5.11

The Estimated Parameters the Simulated Data of size 500 with 10% Censoring Observations

Model 2 with sample size 500 observations and 10% censoring							
Parameter	π_1	π_2	λ	α_1	α_2	β_1	β_2
Postulate	0.50	0.30	1.5	5	9	2	10
Estimates	0.51	0.25	1.50	4.63	9.00	2.02	9.81
Model 2 with sample size 500 observations and 20% censoring							
Parameter	π_1	π_2	λ	α_1	α_2	β_1	β_2
Postulate	0.50	0.30	1.5	5	9	2	10
Estimates	0.47	0.29	1.52	4.52	9.00	2.03	9.85
Model 2 with sample size 500 observations and 40% censoring							
Parameter	π_1	π_2	λ	α_1	α_2	β_1	β_2
Postulate	0.50	0.30	1.5	5	9	2	10
Estimates	0.46	0.17	1.56	4.50	9.00	2.65	9.42

The hazard functions of the three simulated data of size 500 observations corresponding to the 10%, 20% and 40% censoring percentages were presented in Figure 5.24.

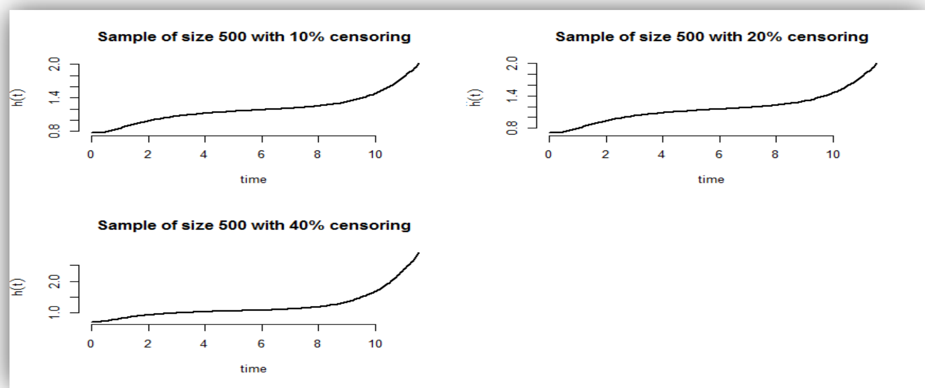


Figure 5.24 The Hazard Functions of the Simulated Data of Size 500 Corresponding to 10%, 20% and 40% Censored Observation.

The parameters of the three sets of the simulated data of size 500 observations were all estimated successfully. The values of the parameter were close to the postulated parameters used in the data generation. Comparing the results in Tables 5.11 showed that, the estimated parameters of the simulated set of data with 10% censoring are closer to the postulated parameters compared to that of the 20% and 40% censoring observations. The hazard function of the set of simulated data consisting of 500 observations with 10% censoring observation is higher when compared with that of 20% and 40% censoring. As the number of censored observations increases the hazard tends to be lower and lower.

The simulation of the three set of generated data of 500 observations with 10%, 20% and 40% censoring observations were repeated 300 times to check the consistency and stability of the EM in estimating the model parameters. The averages, the mean square errors and root mean square error of estimated parameters of the postulated Model 2 are listed in Table 5.12

Table 5.12

The Repeated Simulation of Set of 500 Observations

Model 1 with sample size 500 and 10% censoring

parameters	π_1	π_2	λ	α_2	α_3	β_1	β_2
Postulates	0.50	0.30	1.5	5	9	2	10
Estimates	0.50	0.26	1.41	4.56	9.00	2.00	9.80
MSE	3.29e-7	2.41e-7	6.63e-6	5.16e-4	0.00e+0	5.34e-7	3.91e-5
RMSE	5.73e-4	4.91e-4	2.57e-3	2.27e-2	0.00+0	7.31e-4	6.25e-3

Model 1 with sample size 500 and 20% censoring

parameters	π_1	π_2	λ	α_2	α_3	β_1	β_2
postulated	0.50	0.30	1.5	5	9	2	10
Estimates	0.48	0.27	1.49	4.38	9.00	2.03	9.79
MSE	4.26e-7	2.63e-7	7.48e-6	5.44e-4	0.00e+0	6.65e-7	4.37e-5
RMSE	6.52e-4	5.12e-4	2.74e-3	2.33e-2	0.00e+0	8.15e-4	6.61e-3

Model 1 with sample size 500 and 40% censoring

parameters	π_1	π_2	λ	α_2	α_3	β_1	β_2
postulated	0.50	0.30	1.5	5	9	2	10
estimates	0.47	0.17	1.53	4.46	9.00	2.60	9.39
MSE	5.75e-7	2.73e-7	1.28e-5	9.72e-4	0.00e+0	3.75e-6	4.10e-5
RMSE	7.58e-4	5.22e-4	3.58e-3	3.12e-2	0.00e+0	1.93e-3	6.40e-3

The averages of the parameters are close to the parameters of the postulated parametric survival mixture model with MSE relatively small, which suggests that, the EM performed consistently in estimating the parameters.

The estimation of the parameters of the model was successful for both the ascending and descending order of the mixing probabilities. For both the sets of mixing probabilities the estimation of parameters were closer the true postulate parameters as the sample size increases from 100 to 500 observations. It is also observed that the estimates of the parameters were much better for small censoring percentages. The estimation of the mixing probabilities for the ascending order was better than that of the descending with relatively small value for MSE. In general, it is observed that the mixing probabilities of ascending order performed better than the descending order.

5.4 Kidney Catheter Data

The set of real data analysed in this section is the Kidney Catheter data which were employed for Model1 in Chapter Four were used for Model 2. The estimate and the graphs were presented.

5.4.1 Model 2 versus the Pure Classical Parametric Survival Models

The Kidney Catheter data were modelled by Model 2 and also by the pure classical parametric survival model of the Exponential distribution (E0), the pure classical parametric survival model of the Gamma distribution (G0) and the pure classical parametric survival model of the Weibull distribution (W0) respectively.

Figure 5.25 display the probability density functions of Model 2 and the probability density functions of the pure classical parametric survival models of Exponential, Gamma and Weibull distributions (E1, G2, and W3) corresponding to each component of Model 2 along with the histogram of the Kidney Catheter data.

It can be seen that Model 2 fits the Kidney Catheter data better than the pure classical parametric survival models. This shows that, the Kidney Catheter data were better modelled by the parametric survival mixture model of Exponential, Gamma and Weibull distributions (Model 2) instead of the pure classical parametric survival model of Exponential, Gamma and Weibull distributions respectively.

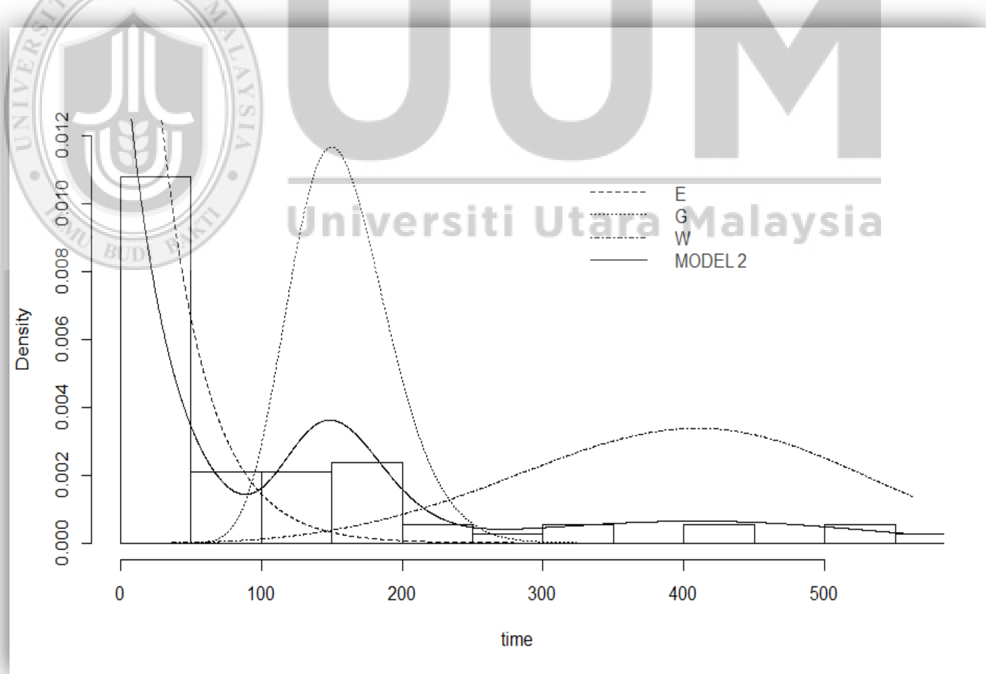


*Figure 4.25.*Model 2 vs the Pure Classical Parametric Survival Models (E, G and W) for the Kidney Catheter Data

The estimated parameters of Model 2 were presented in Table 5.13.

Table 5.13

The Estimated Parameters of Model 2 Using Kidney Catheter Data

Parameter	π_1	π_2	λ	α_1	α_2	β_1	β_2
Estimates	0.52	0.29	32.79	20.38	3.91	7.73	441.56

Figure 5.26 shows the probability density function of Model 2 and the pure classical parametric survival model of the Exponential distribution (E0), the pure classical parametric survival model of the Gamma distribution (G0) and the pure classical parametric survival model of the Weibull distribution (W0) together with the histogram of the Kidney Catheter data. The graph indicates that the Kidney Catheter data fit Model 2 better than the pure classical parametric survival models.

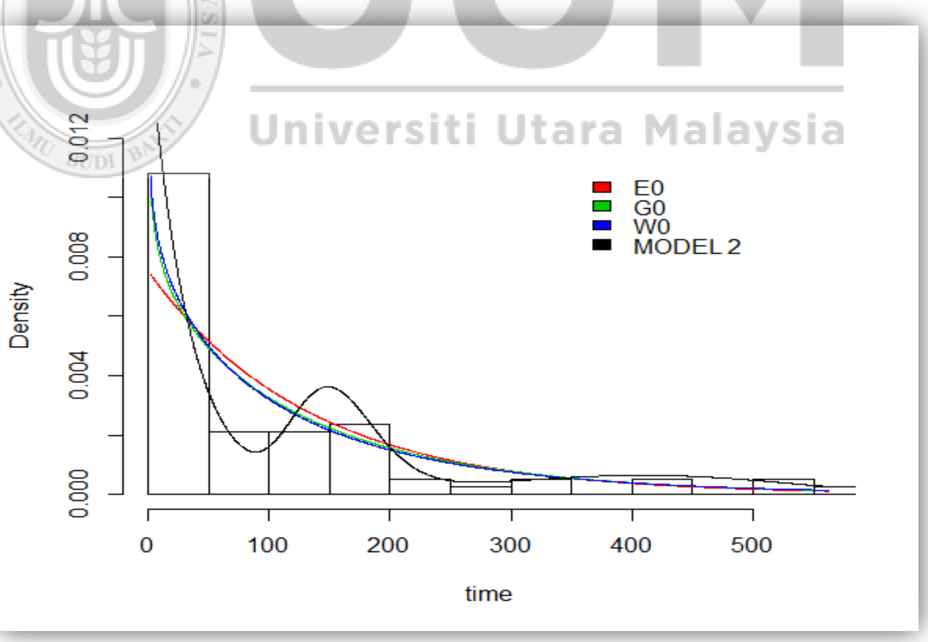


*Figure 4.26.*Model 2 vs the Pure Classical Parametric Survival Models (E0, G0 and W0) for the Kidney Catheter Data

The parameters, the LL, AIC, MSE, RMSE, K-S and E(t) values were estimated and presented in Table 5.14. The estimated result shows that, Model 2 has higher LL value (-331.50) compared to the values of the LL of the individual pure classical survival parametric model of the Exponential, Gamma and Weibull distributions (E0, G0 and W0) respectively. Also, the estimated values of the AIC (677.01) support the selection of Model 2 as the model of that better represents the Kidney Catheter data more than the pure classical parametric survival models.

Table 5.14

The LL and AIC Values for the Kidney Catheter Data

Model	**Estimates**	**LL**	**AIC**	**MSE**	**RMSE**	**K-S**	**E(T)**
E0	$\hat{\lambda}$ = 132.95	-341.70	685.40	0.0211	0.1454	0.26 (0.01)	132.95
G0	$\hat{\alpha}$ = 0.89, $\hat{\beta}$ = 156.96	-341.20	686.40	0.0194	0.1392	0.25 (0.02)	139.69
W0	$\hat{\alpha}$ = 0.86, $\hat{\beta}$ = 128.00	-340.90	685.80	0.0137	0.1171	0.21 (0.04)	138.26
Model 2	$\hat{\lambda}$ = 32.79, $\hat{\alpha}_1$ = 20.38, $\hat{\beta}_1$= 7.73, $\hat{\alpha}_2$ = 3.91, $\hat{\beta}_2$= 441.56, $\hat{\pi}_1$ = 0.52, $\hat{\pi}_2$= 0.29	**-331.50**	**677.01**	**0.0109**	**0.1046**	**0.16** **(0.30)**	**137.25**

The MSE and RMSE values for Model 2 (0.0109), (0.1046) respectively show that Model 2 fit the data better than the pure classical survival model of the Exponential, Gamma and Weibull distributions. The K-S test statistic value for model 2 (0.16) with the p-value of 0.30 shows that the model is adequately represented by the data. Comparing the Model 1 and Model 2 for the Kidney Catheter data shows that the data were slightly better represented by Model 2 with LL (-331.50) and AIC (677.01) compared to Model 1 with LL (-331.57) and AIC (679.13).

The survival function graph of the fitted read data was used to validate the fit of Model 2. The survival function graph was compared with the K-M empirical survival function of the real data to investigate the fit of Model 2. The survival function and the K-M graph of Model 2 were presented in the Figure 5.27.

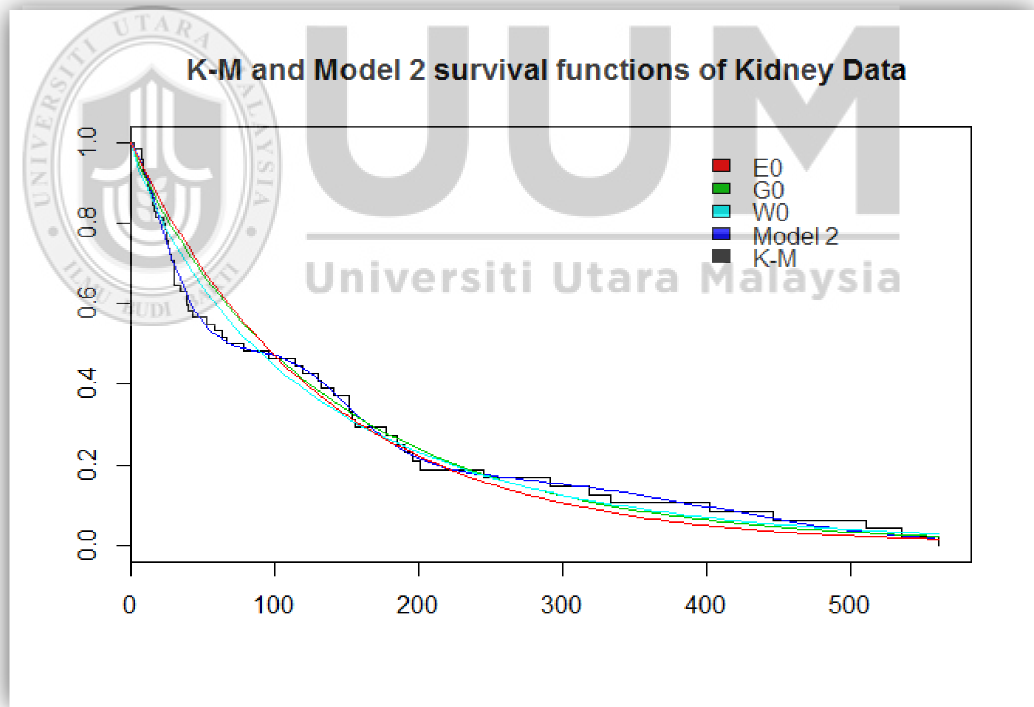


Figure 5.27 K-M, the Survival function of Model 2 and the Pure Survival Models Corresponding to Each Component

In Figure 5.27 the K-M empirical survival function is in solid black, the survival function of Model 2 is in dark blue, the pure classical survival model of the

Exponential in red, the pure classical survival model of the Gamma distribution is in green and the pure classical survival model of the Weibull in light blue. Form the Figure it can be observed that the survival function of Model 2 is in full agreement with the K-M empirical survival function much better than the three other pure classical survival models.

5.4.2 Model 2 versus Three Components Parametric Survival Mixture Models of the Same Distributions

The Kidney Catheter data were used to estimate the parameters of a parametric survival mixture model of the Exponential distributions (E1_E2_E3) and a parametric a parametric survival mixture model of the Weibull distributions (W1_W2_W3).

The probability density functions of Model 2, the parametric survival mixture model of the Exponential distributions and the parametric survival mixture model of the Weibull distributions are plotted graphically along with the histogram of the Kidney Catheter data. Figure 5.28 displays the graphical comparison of the probability density functions of Model 2, the parametric survival mixture model of the Exponential distribution and the parametric survival mixture model of the Weibull distributions. It can be seen that Model 2 represents the Kidney Catheter data much better than the remaining two parametric survival mixture models of Exponential and Weibull respectively.

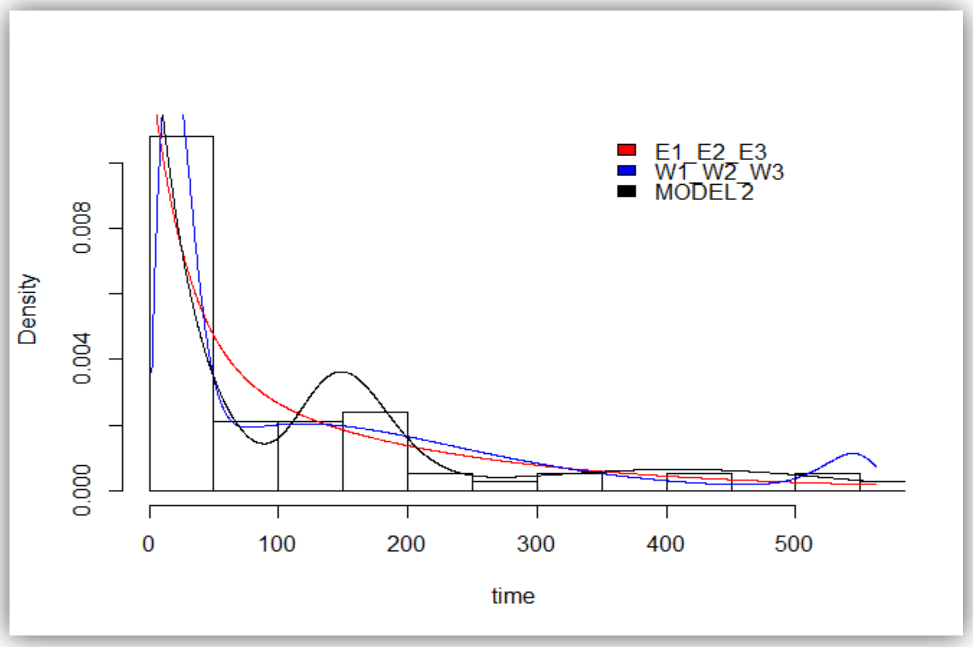


*Figure 5.28.*Model 2 vs the Parametric Survival Mixture Models Corresponding to Each Component Model 2

Model selection was performed to select the model that represents the Kidney Catheter data best among Model 2 and the two parametric survival models of the E1_E2_E3 and W1_W2_W3 and Model 2. Table 5.15 displays the parameters of the fitted parametric survival mixture models along with the LL and AIC values. Model 2 scored the highest value of LL compared to the other parametric survival mixture models. The values of AIC of the two parametric survival mixture models of E1_E2_E3 and W1_W2_W3 are bigger than the value scored by the Model 2. This suggests that, Model 2 represents the Kidney Catheter data better the other three parametric survival mixture models of E1_E2_E3 and W1_W2_W3 models.

Table 5.15

Parameters, LL and AIC of Model 2 and Parametric Survival Mixture Models Corresponding to Each Component of Model 2

	Model 2		**E1_E2_E3**		**W1_W2_W3**
$\boldsymbol{\pi_1}$	0.52	$\boldsymbol{\pi_1}$	0.26	$\boldsymbol{\pi_1}$	0.41
$\boldsymbol{\pi_2}$	0.29	$\boldsymbol{\pi_2}$	0.22	$\boldsymbol{\pi_2}$	0.53
$\boldsymbol{\lambda}$	32.79	$\boldsymbol{\lambda_1}$	0.04	$\boldsymbol{\alpha_1}$	1.82
$\boldsymbol{\alpha_1}$	20.38	$\boldsymbol{\lambda_2}$	0.01	$\boldsymbol{\alpha_2}$	1.69
$\boldsymbol{\alpha_2}$	3.91	$\boldsymbol{\lambda_3}$	0.01	$\boldsymbol{\alpha_3}$	26.56
$\boldsymbol{\beta_1}$	7.73			$\boldsymbol{\beta_1}$	26.19
$\boldsymbol{\beta_2}$	441.56			$\boldsymbol{\beta_2}$	202.39
				$\boldsymbol{\beta_3}$	545.34
LL	***-331.50***	**LL**	-339.46	**LL**	-331.91
AIC	***677.01***	**AIC**	682.91	**AIC**	679.83

The Kidney Catheter data were used to model the parametric survival mixture models; E1_E2_E3 and W1_W2_W3 corresponding to each component of Model 2. Model 2 has been plotted with each of the parametric survival mixture model of E1_E2_E3 and W1_W2_W3 and their pure classical parametric survival models graphically. In Figure 5.29 Model 2 and the parametric survival mixture model of the Exponential distributions (E1_E2_E3) were plotted together with the pure classical parametric survival model of the Exponential distributions (E1, E2 and E3) corresponding to each component. The graph shows that Model 2 fits the Kidney Catheter data better than the parametric survival mixture model of Exponential distribution and the pure classical survival Exponential models corresponding to each component.

Figure 5.30 displays probability density functions of Model 2 and the parametric survival mixture model of the Weibull distributions (W1_W2_W3) along with the pure classical parametric survival models of the Weibull distribution (W1, W2 and W3). The graph shows that the Kidney Catheter data were better modelled by Model 2 than by the parametric survival mixture model of the Weibull distribution and the pure classical parametric survival models of the Weibull distribution corresponding to each of the components of the parametric survival mixture model of the Weibull distributions (W1_W2_W3).

The Kidney Catheter data showed that the developed EM estimated the parameters of Model 2 successfully and the model selection revealed that Model 2 represents the Kidney Catheter data better than the pure classical parametric survival models corresponding to each component of Model 2, the parametric survival mixture model of the Exponential distributions and the parametric survival mixture model of the Weibull distributions.

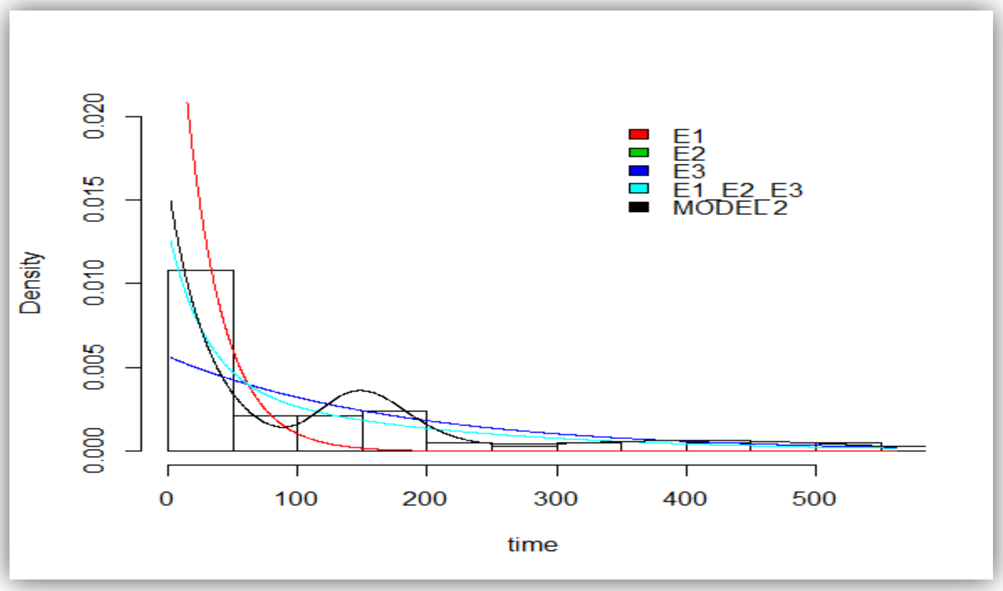


*Figure 5.29.*Model 2 vs the Parametric Survival Mixture of Exponential and the Pure Classical Distribution of Each Component

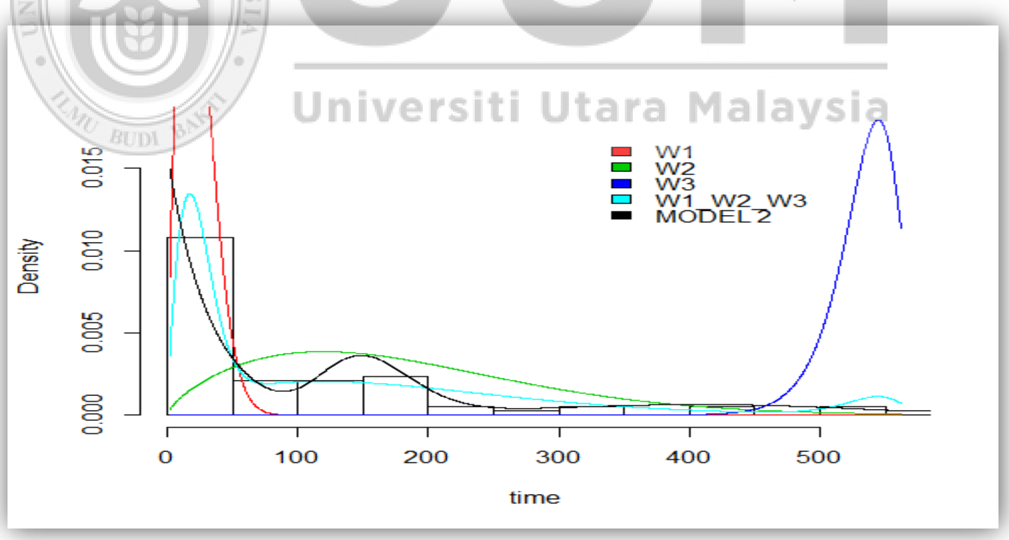


*Figure 5.31.*Model 2 vs the Parametric Survival Mixture of Weibull and the Pure Classical Distribution of Each Component

5.5 Summary

The development of a three components parametric survival mixture model of the Exponential, Gamma and Weibull distributions (Model 2) was discussed in this chapter. The implementation of EM in survival mixture model and the derivation of the parameters of Model 2 was highlighted.

Simulated survival data were used to validate the performance of Model 2 by generating 18 different samples from the parametric survival mixture model of the Exponential, Gamma and Weibull distributions. The generated data constitute of the three different samples of size 100, 200 and 500 observations. Each of the samples was generated based on three different censoring percentages. Also the generated samples were based on two different mixing probabilities arranged in ascending and descending order. The parameters of Model 2 were estimated by employing the EM and the consistency and stability of EM was investigated by repeating the simulation 300 times.

Generally, the parameters estimated from the data were closed to the true parameters used in the simulation of the data. Validating the performance of Model 2 using the three different sample sizes showed that the estimation of the parameters was better as the sample size increases. The comparison of the three censoring percentages showed that the Model 2 performed better with smaller censoring percentages for both the ascending and descending order of the mixing probabilities. However, the performance of Model 2 with the mixing probabilities in ascending order was better than that of the mixing probabilities in descending order. Graphical representations of the hazard functions for different samples of Model 2 with different censoring percentages were investigated. Generally, it was found that the hazard function tends to be higher with small censoring percentage. As the censoring percentage increases more individual or items survive which reduce the value of the hazard function.

Model 2 was validated by employing empirical study. The parameters of Model 2 were estimated and reported. Model 2 was compared with pure classical parametric survival distributions corresponding to the distribution of each component graphically. To validate Model 2, the LL, AIC, MSE, RMSE, K-S test and E(t) were computed and compared with those of the pure classical Gamma distribution.

The K-M empirical survival function of the real data was compared with the survival function of Model 2 graphically to evaluate the fitness of the model. The graph showed that Model 2 fit the data better the pure classical parametric survival models. Model 2 was also compared with the E1_E2_E3 and W1_W2_W3 survival mixture models corresponding to distribution of each of the components of Model 2. The simulation and the empirical study showed that Model 2 is preferred over the pure classical survival models in modelling survival data when the data seem to come from population of heterogeneous nature.

CHAPTER SIX
CONCLUSION

6.1 Summary

The pure classical parametric survival models are the conventional method for analysing survival data when the data are believed to be homogeneous and follow some particular parametric probability distribution. In some situations the survival data come from populations that are believed to be heterogeneous in nature. This thesis proposed parametric survival mixture model of three components as a useful and flexible tool for analysing survival data of heterogeneous nature, instead of the pure classical parametric survival models. The study proposed two models; the first one is a parametric survival mixture model of the Gamma distributions referred to as Model 1. The second is a parametric survival mixture model of the Exponential, Gamma and Weibull distributions referred to as Model 2.

Simulation study was employed to validate performance of the two models with three different samples sizes, three different censoring percentages and two sets of mixing probabilities arranged in ascending and descending orders. The simulation study performed well in validating the performance of the models. Both models perfomed well with large sample compared to small sample. Also, the models performed better with small censoring percentages. The mixing probabilities were better estimated with small censoring percentage compared to the samples with large censoring. The hazard function of both models was investigated using different censoring percentages and represented graphically. It was found that the hazard function tends to decrease with the increase in the censoring percentage of the data.

Empirical study was carried out to validate the two models. Graphical representations were used where the probability densities of the models were plotted together with the pure classical parametric survival models, parametric survival mixture models and the histogram of the real data. To compare the performance of the two models with the pure classical survival models and parametric survival mixture models the LL, AIC, MSE, RMSE, K-S test and E(t) were provided. The comparison showed that Model 1 and Model 2 fit the real data better than the other models. The K-M empirical survival function was compared with the survival function of both Model 1 and Model 2 graphically. The graphs showed that the two models fit the real data better than the pure classical survival models corresponding to each component of the survival mixture models.

In conclusion, the simulated and real data application of Model 1 and Model 2 demonstrated that the parametric survival mixture models are flexible tools and maintain the feature of the classical parametric survival distribution. This result shows that the three components parametric survival mixture model is an appropriate alternative for modelling heterogeneous survival data.

6.2 Problems and Limitations

The main problem with the EM is the problem of the starting or initial point. The Algorithm is very sensitive to starting point. In the case of the parametric survival mixture model of different distributions (the Exponential, the Gamma and the Weibull) one has to have a very good guess of the starting point. So the problem will still be how to get a very good convenient starting point. The researcher needs to specify the mixing probabilities together with the initial points of the parameters of each component of the parametric survival mixture model. Unlike the case of the parametric survival mixture model of Gamma distribution which consists of same distribution, the EM can be modified to generate the set of initial starting point by the method of moments. The researcher has the option of specifying the number of components or he/she may just allow the EM to generate the initial or starting point of both the mixing probabilities and parameters of distribution of the components.

Also, among the problems and limitations is obtaining real data that can be used for applications. It is a very difficult task to obtain convenient real data for survival analysis. Most of the published researches do not provide the data used to validate their work. Very few authors respond to request for some data they used in their research. The main sources are published text books, some statistical software, and very few websites where they make set of data available.

6.3 Future Research

In the concluding chapter there is needed to point out some areas that require some further research in the future. Chapter Four discussed a three component parametric survival mixture model of same distribution with a particular reference to the Gamma distribution. There are some other important distributions that have not been addressed in the literature. Three components parametric survival mixture models of some important distributions such as Lognormal, Logistic and Gompertz to mention a few could be investigated in future research.

In Chapter Five a parametric survival mixture model of Exponential, Gamma and Weibull distributions was discussed. The survival data analysis is not limited to only these distributions discussed in this thesis; they are many other important distributions. There is need to extend the idea of three components parametric survival mixture model of different distributions to some other distribution, such as Lognormal, Logistic, Log logistic to mention a few which could be a good and effective tool for modelling heterogeneous survival data

Also, in many instances, survival data are collected together with some covariates that are believed to influence the survival time of the observation, such as age, treatments, gender and race to mention a few. Model 1 and Model 2 did not consider covariates in the analysis. Further research needs to include the covariates in the analysis to assess their effect on the survival time of the observations.

REFERENCES

Abu Bakar, M. Z., Daud, I, & Ibrahim, N. A. (2006). Estimating a logistic Weibull mixture models with long-Term survivors. *Jurnal Tecknologi*, 45(C) , 57-66.

Abu -Zinadah, H. H. (2010). A study on mixture of exponentiated pareto and exponential distributions. *Journal of Applied Sciences Research, 6*(4), 358-376.

Akaike, H. (1974). A new look at the statistical model identification. *Automatic Control, IEEE Transactions on, 19*(6), 716-723.

Al-Hussaini, E. K., Al-Dayian, G. R., & Adham, S. A. (2000). On finite mixture of two-component Gompertz lifetime model. *Journal of Statistical Computation and Simulation, 67*(1), 1-20.

Birnbaun, Z. W. & Saunders S. C. (1958).“A statistical model for life-length of materials”. *Journal of the American Statistical Association.* 53, 151-160.

Blackstone, E. H., Naftel, D. C., & Turner, M. E. Jr. (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.

Bohning, D., & Seidel, W. (2003). Editorial: recent developments in mixture models. *Computational Statistics &amp; Data Analysis, 41*(3-4), 349-357.

Brown, G. W., & Flood, M. M. (1947). Tumbler mortality. *Journal of the American Statistical Association*. 42, 562-574.

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference: a practical information-theoretic approach*. Springer.

Burnham, K. P., & Anderson, D. R. (2004). Multimodel inference understanding AIC and BIC in model selection. *Sociological methods & research*, *33*(2), 261-304.

Cai, C., Zou, Y., Peng, Y., Zhang, J., & Cai, M. C. (2012). Package ‘smcure’.

Chang, S. C. (1998). Using parametric statistical models to estimate mortality structure: The case of Taiwan. *Journal of Actuarial Practice*, *6*(1).

Cheng, S. W., & Fu, J. C. (1982). Estimation of mixed Weibull parameters in life testing. *Reliability, IEEE Transactions on, R-31*(4), 377-381.

Cohen, A. C., Jr. (1951). Estimating parameters of logarithmic-normal distributions by maximum likelihood. *Journal of the American Statistical Association, 46*(254), 206-212.

Copas, J. B., & Heydari, F. (1997). Estimating the risk of reoffending by using exponential mixture models. *Journal of the Royal Statistical Society. Series A (Statistics in Society), 160*(2), 237-252.

Davis, D. J. (1952). An analysis of some failure data. *Journal of the American Statistical Association, 47*(258), 113-150.

Dempster, A. P. Laird, N. M., & Rubin, D. B.(1977). Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion)". *Journal of Royal Statistical Society. Series B, 39, 1-38.*

Epstein, B. & Sobel, M. (1953). Life testing. *Journal of the American Statistical Association*, 48, 486-502.

Erisoglu, U., Erisoglu, M. & Erol, H. (2011). A mixture model of two different distributions approach to the analysis of heterogeneous survival data. *International Journal of Computational and Mathematical Sciences 5(2).*

Erisoglu, U., Erisoglu, M., & Erol, H. (2012). Mixture model approach to the analysis of heterogeneous survival data. *Pakistan Journal of Statistics 28*(1), 115-130.

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.

Escobar, L. A., & Meeker, W. Q., Jr. (1992). Assessing Influence in Regression Analysis with Censored Data. *Biometrics, 48*(2), 507-528. doi: 10.2307/2532306.

Everitt, B. S., & Hand, D. J., (1981). *Finite mixture distributions*. Chapman and Hall Inc. New York

Farcomeni, A., & Nardi, A. (2010). A two-component Weibull mixture to model early and late mortality in a Bayesian framework. *Computational Statistics & amp; Data Analysis, 54*(2), 416-428.

Farewell, V. T. (1982). The use of mixture models for the analysis of survival data with long-term survivors. *Biometrics, 38*(4), 1041-1046.

Fraley, C., & Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association, 97*(458), 611-631.

Fruhwirth-Schnatter, S. (2006). *Finite mixture and markovs switching models*. New York: Springer.

Harter, H. L., & Moore, A. H. (1966). Local-maximum-likelihood estimation of the parameters of three-parameter Lognormal populations from complete and censored samples. *Journal of the American Statistical Association, 61*(315), 842-851.

Ibrahim, J. G., Chen, M. H., & Sinha, D. (2001). *Bayesian survival analysis*. New York: Springer-verlag.

Jaheen, Z. (2005). On record Statistics from a mixture of two exponential distributions. *Journal of Statistical Computation & Simulation, 75*(1), 1-11.

Jensen, J. & Petersen, N. E. (1982). *Burn-in: an engineering approach to the design and analysis of burn-in procedures,* wiley , New york.

Jewell, N. P. (1982). Mixtures of exponential distributions. *The Annals of Statistics, 10*(2), 479-484.

Jiang, S. & Kececioglu, D (1992a). Graphical representation of two mixed-Weibull distributions. *IEEE Transaction on Reliability*, vol. 41,241-247.

Jiang, S. & Kececioglu, D (1992b). Maximum likelihood estimates, from censored data, for mixed-Weibull distributions. *IEEE Transaction on Reliability*, vol. 41,248-255.

Jiang, R., & Murthy, D. N. P. (1995). modelling failure-data by mixture of 2 Weibull distributions: a graphical approach. *Reliability, IEEE Transactions on, 44*(3), 477-488.

Jiang, R., & Murthy, D. N. P. A mixture model involving three weibull distributions, Proceedings of the Second Australia-Japan Workshop on Stochastic Models in Engineering. Technology and Management (Gold Coast, Australia).

Kalbfleisch J. D. & Prentice R. L. (2002). *The statistical analysis of failure time data* (second ed.), John Wiley & Sons, Inc. Hoboken, New Jersey.

Kaplan, E. L., & Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association, 53*(282), 457-481.

Kersey, J. H., Weisdorf, D., Nesbit, M. E., LeBien, T. W., Woods, W. G., McGlave, P. B., . . . Bostrom, B. (1987). Comparison of autologous and allogeneic bone marrow transplantation for treatment of high-risk refractory acute lymphoblastic leukemia. *New England Journal of Medicine, 317*(8), 461-467.

Khalid, Z. M. & Morgan, J. T.(2008). Cross-sectional and longitudinal approaches in a survival mixture model, *Matematika*, Vol. 24, 231-242.

Koti, K. M. (2001). Failure-time mixture models: yet another way to establish efficacy. *Drug Information Journal, 35*(4), 1253-1260.

Kouassi, D. A. & Singh J. (1997). A semi-parametric approach to hazard estimation with randomly censored observations. *Journal of American Statistical Association* 92, pp.1351-1355.

Kuk, A. Y. C., & Chen, C.-H. (1992). A mixture model combining logistic regression with proportional hazards regression. *Biometrika, 79*(3), 531-541.

Larson, M. G., & Dinse, G. E. (1985). A mixture model for the regression analysis of competing risks data. *Journal of the Royal Statistical Society. Series C (Applied Statistics), 34*(3), 201-211.

Lawless J. F. (2003). *Statistical models and methods of lifetime data*, (2nd ed.) John Wiley and Sons, Inc. Hoboken, New Jersey.

Lee, E. T. & Wang, J. W.(2003). *Statistical methods for survival data analysis* (3rd ed.). John Wiley & son.

Leisch, F. (2004). Exploring the structure of mixture model components. In J Antoch (ed.), “Compstat 2004- proceedings in Computational Statistics”, pp. 1405-1412. Physica Verlag, Heidelberg. ISBN 3-7908-1554-3.

Leng, O. Y., & Khalid, Z. M. (2010). *A comparative study of maximum likelihood and Bayesian estimation approaches in estimating frailty mixture survival model parameters.* Paper presented at the Proceedings of the 6th IMT-GT Conference on Mathematics, Statistics and its Applications (ICMSA2010), Universiti Tunku Abdul Rahman, Kuala Lumpur, Malaysia.

Li, L., & Choe, M. K. (1997). A mixture model for duration data: analysis of second births in China. *Demography, 34*(2), 189-197.

Ling, D., Huang, H.-Z., & Liu, Y. (26, 26-29 Jan. 2009). *A method for parameter estimation of mixed Weibull distribution.* Paper presented at the Reliability and Maintainability Symposium, 2009. RAMS 2009. Annual.

Marín, J. M., Rodríguez-Bernal, M. T., & Wiper, M. P. (2005). Using Weibull mixture distributions to model heterogeneous survival data. *Communications in Statistics: Simulation and Computation, 34*(3), 673-684.

McGilchrist, C. A., & Aisbett, C. W. (1991). Regression with frailty in survival analysis. *Biometrics, 47*, 461-466.

McLachlan, G. J., & Peel, D. (2000). *Finite mixture models*: John Wiley & Sons, Inc.

McLachlan, G. J., & Krishnan, T. (2008). *The EM algorithm and extensions* (Second ed.). Hoboken New Jersey: John Wiley & Sons, Inc.

Moltoft, J. (1983). Behind the “bathtub” curve, a new model and its consequences, *Microeclectonics & Reliability*, 23, 489-500.

Murthy D. N. P., Xie, M. & Jiang, R. (2004). *Weibull models.* John Wiley & son.

Ng, A. S. K., McLachlan, G. J., Yau, K. K. W., & Lee, A. H. (2004). Modelling the distribution of ischaemic stroke-specific survival time using an EM-based mixture approach with random effects adjustment. *Statistics in Medicine, 23*(17), 2729-2744.

Olkin, I., & Spiegelman, C. H.(1987). A semi-parametric approach to density estimation, *Journal of the American Statistical Association, 82, 858-865.*

Othus, M. Li, Y & Tiwari, R. C. (2009). A class of semi-paramertic mixture cure survival models with dependent censoring. *Journal of American Statistical Association,* 104(487). 1241-1250.

Phillips, N., Coldman, A., & McBride, M. L. (2002). Estimating cancer prevalence using mixture models for cancer survival. *Statistics in Medicine, 21*(9), 1257-1270.

Razali, A. M., & Salih, A. A. (2009). Combining two Weibull distributions using a mixing parameter. *European Journal of Scientific Research, 31*(2), 296-305.

Rider, P. R. (1961). The method of moments applied to a mixture of two exponential distributions. *The Annals of Mathematical Statistics, 32*(1), 143-147.

Seppa, K. Hakulinen, T., Kim, J. J. & Laara, E. (2010). Cure fraction model with random effects for regional variation in cancer survival. *Statistics in Medicine*, 29. 2781-2793.

Sultan, K. S., Ismail, M. A., & Al-Moisheer, A. S. (2007). Mixture of two inverse Weibull distributions: properties and estimation. *Computational Statistics &amp; Data Analysis, 51*(11), 5377-5387.

Sun, J. (2006). *The statistical analysis of interval-cencored failure time data*. New York: Springer Science, Business Media.

Tableman, M., & Kim, J. S. (2004). *Survival analysis using S: analysis of time- to-event data*: Chapman & Hall/CRC.

Taylor, J. M. G. (1995). Semi-parametric estimation in failure time mixture models. *Biometrics, 51*(3), 899-907.

Team, R. C. (2005). R: A language and environment for statistical computing: ISBN 3-900051-07-0. R Foundation for Statistical Computing. Vienna, Austria, 2013. url: http://www. R-project. org.

Therneau, T. (2013). A Package for Survival Analysis in S. R package version 2.37-4. Retrieved from http://CRAN.R-project.org/package=survival

Tukey, J. W. (1977) *Explanatory data analysis*. Addison Wesley publishing company Inc. Philippines.

Vernic, R., Teodorescu, S., & Pelican, E. (2009). Two Lognormal models for real data. *Annals of Statistics Ovidius Constanta, 17*(3), 263-279.

Weibull, W. (1939). A statistical theory of strength of materials. *Ingeniorsvetens Kapsakadeniens Handlingar.*

Weibull, & W. (1951). A statistical distribution function for wide applicability. *Journal of Applied Mathematics*(18), 293-297.

Wiper, M., Insua, D. R., & Ruggeri, F. (2001). Mixtures of gamma distributions with applications. *Journal of Computational and Graphical Statistics, 10*(3).

Young, D. S., Benaglia, T., Chauveau, D., Hunter, D. R., Elmore, R. T., Xuan, F., ... & Thomas, H. (2007). The mixtools package: tools for mixture models. *R Package Version 0.2. 0*.

Yu, B., & Peng, Y. (2008). “Mixture cure models for multivariate survival time data” *Computational Statistics & Data Analysis*, 52, 1524-1532.

Zelen, M. (1966). Application of exponential models to problems in cancer research. *Journal of the Royal Statistical Society. Series A (General), 129*(3), 368-398.

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

Zhang, X., Wang, Y., & Lu, D. (2011, 26-28 July 2011). *A new algorithm for parameters estimations of multivariate mixed Weibull distributions with censoring data.* Paper presented at the 2011 International Conference on Multimedia Technology, ICMT 2011.