

**ΟΙΚΟΝΟΜΙΚΟ
ΠΑΝΕΠΙΣΤΗΜΙΟ
ΑΘΗΝΩΝ**



**ATHENS UNIVERSITY
OF ECONOMICS
AND BUSINESS**

**Department of Statistics
Full time Master in Statistics**

**Mixture of Weibull regressions as parametric
models for flexible hazard functions**

By

AIKATERINI VERVITA

M.Sc. Thesis

Submitted to the Department of Statistics
of the Athens University of Economics and Business
in partial fulfilment of the requirements for
the degree of Master of Science in Statistics

Athens, Greece

February 2017





ACKNOWLEDGEMENTS

Foremost, I would like to express my sincere gratitude to my thesis advisor Prof. Dimitris Karlis for his continuous guidance, unwavering support, indispensable advice and incessant encouragement during the preparation of this thesis, and also during all my undergraduate and postgraduate studies at the Athens University of Economics and Business. Also, I would like to extend my heartfelt gratitude towards all the Professors of my academic life thus far, for all the valuable knowledge they have instilled in me these past years.

Furthermore, I would like to deeply thank all the employees at the Frontier Science Foundation Hellas, who helped me to fully understand the purpose of HERA trial -a clinical trial which make up the motivation of this thesis-.

Lastly, a huge thanks goes to my family and my friends for all their unconditional support and confidence in me. They have been a great encouragement throughout my academic years and unreservedly continue to do so.



CURRICULUM VITAE

Name: Aikaterini Vervita

Address: Lidorikiou 27
 Agios Dimitrios, 17342
 Athens, Greece

E-mail: kat_vrta@hotmail.com

Telephone: +30 21 0 995 6709

Mobile: +30 69 7 670 9823

Date of birth: May 20th, 1992

Family: *Single*

Education:

2007-2010	3 rd General Lyceum of Agios Dimitrios, Athens, Greece
2010–2015	Athens University of Economics and Business, B.Sc. Statistics
2015–2017	Athens University of Economics and Business, M.Sc. Statistics

Languages: Greek(Native Speaker) Fluent speak, write, read
English: Full Professional Proficiency,
 FCE University of Lancaster, December 2007
French: Minimum Professional Proficiency,
 Prepadelf A2, May 2007

Computer Skills:

Excellent knowledge of Microsoft Word, Excel and Powerpoint
 Excellent knowledge of Statistical Software SPSS, SAS and R/S-PLUS

Present position:
 Biostatistician at Frontier Science Foundation Hellas

Professional experiences:

October 2014– January 2015	Frontier Science Foundation Hellas, Biostatistics Intern (Intership)
September 2015–June 2016	Accounting Clerk
September 2015–June 2016	A' Surgery Clinic Hospital METAXA, Statistical Analyst of Clinical Protocols



ABSTRACT

In survival analysis, the idea of hazard function is might be of more interest than the probability density function to a patient who had survived a certain time period and wanted to know something about its survival progress. Thus, in clinical trials the attention is focused on the estimation of hazard function and not in p.d.f. It is easily understood that a hazard function can not be always increasing or decreasing, which happen in many known survival models. It is necessary to be find parametric models which can describe the behavior of the hazard function properly. Most of the distributions that are used very common for the analysis of survival data (e.g. Weibull and exponential) are not able to give hazard functions with flexible shapes.

HERA is a clinical trial about breast cancer which compares patients who had received treatment and patients in the observation group. However, in both cases the hazard function is non-monotonic, but it increases for some period and then turns decreasing. A two Weibull mixture model with shape parameters 1.38, 1.10 and scale parameters 1.46, 34.2 for the observation group, another Weibull mixture model with shape parameters 1.74, 0.96 and scale parameters 1.59, 47.18 for the one year treatment group and a mixture model with shape parameters 1.26, 1.51 and scale parameters 3.20, 52.9 for the two year treatment group can describe the hazard functions sufficiently in all cases. It seems that the first Weibull distribution of the model corresponds to patients with small survival probability until a certain time and the other to patients with longer survival time. Also the mixture model can be used in cases where we are not interested in survival time only but we include some patients' characteristics..

It is known that the hazard function from a Weibull distribution can be either increasing or decreasing. So a finite mixture of Weibull distributions can be considered as proper for survival data with non-monotonic hazard rates.





Contents

1	Introduction	1
1.1	The importance of clinical trials	1
1.2	Motivation of the thesis - HERA trial	4
1.3	Problem and structure of the thesis	8
2	Survival models	11
2.1	Survival analysis and survival data	11
2.2	Model fitting	13
2.2.1	Common families of survival distributions	14
2.3	Hazard function estimation with kernels	23
3	Mixture of Weibull distributions	29
3.1	Mixture models	30
3.1.1	Weibull mixture model	31
3.2	Estimation of model parameters	35
3.2.1	Maximum-Likelihood estimation for mixture models	35
3.2.2	MLEs for a 2-finite Weibull mixture model	37
3.3	EM algorithm	38
3.3.1	EM algorithm for mixture models	38
3.3.2	EM algorithm for a 2-finite Weibull mixture model	41
3.3.3	Convergence of the algorithm	42
3.3.4	Advantages and disadvantages of the algorithm	42
3.3.5	Simulation study about the EM algorithm	43



4	Data analysis with HERA data	49
4.1	Descriptive statistics and crosstabs	51
4.2	Log-rank test and multivariate Cox model	57
4.2.1	Log-rank test	57
4.2.2	Cox PH model	60
4.3	Estimation of hazard function	63
4.3.1	No covariates in the model	63
4.3.2	Adding covariates to the model	67
5	Conclusions	75



List of Tables

3.1	Estimated values with EM algorithm for simulated data from a 2-finite Weibull mixture model	44
3.2	Estimated values with EM algorithm for simulated data from a 2-finite Gaussian mixture model	47
4.1	Frequencies about patients' characteristics	52
4.2	Frequencies about patients' characteristics by group	55
4.3	Log-rank test for Disease Free Survival and Overall Survival	58
4.4	Multivariate Cox proportional hazard model for DFS (we needed 5 steps until the final model)	62
4.5	Results from fitting some parametric models in observation group .	64
4.6	Results from fitting some parametric models in 1 year treatment group	65
4.7	Results from fitting some parametric models in 2 year treatment group	65
4.8	Estimated values and standard errors for the parameters of the 2-finite Weibull mixture model as they are obtained from the EM algorithm (for each of 3 groups)	66
4.9	Results from fitting some parametric models with covariate the patients' group	70
4.10	Estimated values and standard errors for the parameters of the 2-finite Weibull mixture model as they are obtained from the EM algorithm with covariate the patients' group	70





List of Figures

1.1	Profile of HERA trial	5
1.2	Non parametric estimator of the hazard function for the patients in the observation group and the 1 and 2 year treatment groups. (We have excluded patients with DFS time equal to 0 to avoid numerical problems)	8
2.1	Density, cumulative, survival and hazard functions for exponential distribution for different values of parameter λ	15
2.2	Density, cumulative, survival and hazard functions for Weibull distribution for different values of shape parameter a and scale parameter $b = 1$	16
2.3	Density, cumulative, survival and hazard functions for gamma distribution for different values of shape parameter a and scale parameter $b = 1$	18
2.4	Density, cumulative, survival and hazard functions for log-normal distribution for different values of location and scale parameters	19
2.5	Density, cumulative, survival and hazard functions for log-logistic distribution for different values of shape parameter b and scale parameter $a = 1$	21
2.6	Estimation of the hazard function using kernels for different values of (global) bandwidth (kernel function=Epanechnikov)	27
3.1	Probability density function for 2-finite Weibull mixture model for different values of the parameters a_1, a_2, b_1, b_2 and p	32



3.2	Cumulative distribution and Survival functions for 2-finite Weibull mixture model for different values of the parameters a_1, a_2, b_1, b_2 and p	33
3.3	Hazard function for 2-finite Weibull mixture model for different values of the parameters a_1, a_2, b_1, b_2 and p	34
3.4	Convergence of EM algorithm about each parameter of the 2-finite Weibull mixture model	45
3.5	Converge of EM algorithm about each parameter of the 2-finite Gaussian mixture model	48
4.1	Kaplan-Meier plot for Disease-free Survival by patients' group	59
4.2	Kaplan-Meier plot for Overall Survival by patients' group	60
4.3	Kaplan-Meier survival curves and the fitted survival models for patients in the observation group	71
4.4	Nonparametric hazard functions and parametric estimated functions in the observation group	71
4.5	Kaplan-Meier survival curves and the fitted survival models for patients in the 1 year treatment group	72
4.6	Nonparametric hazard functions and parametric estimated functions in the 1 year treatment group	72
4.7	Kaplan-Meier survival curves and the fitted survival models for patients in the 2 year treatment group	73
4.8	Nonparametric hazard functions and parametric estimated functions in the 2 year treatment group	73
4.9	Non parametric hazard function and parametric fuctions estimated from mixture of 2 and 3 Weibull distributions	74
4.10	Kaplan-Meier survival curves and the fitted survival model as well as the fitted survival functions for each of the 2 Weibull distributions separately	74



Chapter 1

Introduction

1.1 The importance of clinical trials

Health and fighting off diseases are two major issues that concern humanity every day. As time passes, the necessity of the increasing of life expectancy and the development of quality of life has been growing. Clinical trials play a great role in this effort and, as a result, they are considered essential. Clinical trials are research studies that explore whether a medical strategy, treatment, or device is safe and effective for humans. These studies also may show which medical approaches work best for certain illnesses or groups of people (see Pocock, 2013).

We need to know: Does a treatment work? Does it work better than other treatments? Does it have any side effects? Clinical trials are designed to answer these questions and improve health and quality of life for patients. Until well-designed trials have been carried out, we simply do not have enough evidence to know if a treatment is both effective and safe. Without trials, there is a risk that people will be given treatments which do not work and which may even be harmful.

It may come as a surprise to many people that sometimes doctors do not know which treatment is the best. When doctors make decisions about how to treat a particular illness or condition, they use their medical knowledge, based on the textbooks they have read, the results they have observed in previous patients, similar observations by their colleagues, what they have heard at conferences and



what they have read in medical journals. Clinical trials produce the best data available for health care decision-making and provide a different kind of knowledge, based on statistics.

Experimenting and testing have long been a part of medicine, and there are many different kinds of trials (Phase I, Phase II, Phase III and Phase IV trials) (see Friedman et al., 2015). Phase I trials are designed to determine the maximum amount of the drug that can be given to a person before adverse effects become intolerable or dangerous. Usually, in Phase I trials, a small group of 2-100 healthy volunteers is recruited. Once a dose or range of doses is determined, the next goal is to evaluate whether the drug has any biological activity or effect. Phase II trials are performed on larger groups (100-300) and are designed to assess how well the drug works, as well as to continue Phase I safety assessments in a larger group of volunteers and patients. Phase III trials are the full scale evaluation of treatment and are designed to compare efficacy of the new treatment with the standard treatment. Phase III clinical trials are presented below particularly. Phase IV studies include all studies performed after drug approval and related to the approved indication. These are post-marketing surveillance studies. The focus of these trials is on how drugs work in the real world.

Phase III clinical trials are the most rigorous and extensive type of scientific clinical investigation of a new treatment. These are usually the most expensive and time-consuming of the trials. Phase III trials are usually large, and may include hundreds or even thousands of patients. They often compare the effects of newer drugs or treatments with standard treatments if there are any. They provide a better test of whether new treatments work better than existing treatments, and firmer evidence about how common and serious any side effects are.

To make sure that each group contains a similar mix of people, many trials are randomized. This means that people are allocated at random to one of the groups in the trial, often by using a computer program. When people are randomized they have an equal chance of being in either of the trial groups. Random allocation helps ensure we are comparing two very similar groups of patients, so if one group does better than another, it is very likely to be because the treatments being compared have different effects, and not because of differences between the people in the



groups. Randomized clinical trials have been generally recognized as the best way to compare different approaches to preventing and treating illness.

Almost all Phase III trials are randomized. In randomized clinical trials, one group of people, the experimental group, is given the new treatment. The other group, called the control group, is given the standard treatment. If no standard treatment exists, the control group may not be given any specific treatment or may be given a placebo. Some trials may compare more than two groups.

A placebo is a treatment, with no active ingredient, which is designed to appear very like the treatment being tested. By comparing peoples responses to the placebo and to the treatment being tested, researchers can tell whether the treatment is having any real benefit, rather than patients simply feeling better because something is being done. There are several ways in which the results of trials can be made as reliable and accurate as possible. One of these is to make the trial a blind trial. In a blind trial the participants are not told which group they are in. This is because if they knew which treatment they were getting it might influence how they felt or reported their symptoms. Some trials are double-blind, which means that neither participants nor the doctors and others treating them know which people are getting which treatments. This also avoids the doctors hopes and expectations influencing the results of the trial

Many people believe that clinical trials are only related to new drug treatments, especially in the field of cancer. But clinical trials can also be used to test and compare all sorts of different types of treatments across a range of conditions, including surgery, physiotherapy and rehabilitation programs, screening, prevention (such as vaccines), complementary therapies, radiotherapy and chemotherapy.

Clinical trials are designed by doctors, scientists and others, and are conducted together with patients. The first step is to decide which questions need answering, and then to look carefully at the results of any trials that have already been done and any other research evidence. Then doctors, nurses, patients and researchers work together with statisticians and trial managers to design the trial. This is written down in the trial protocol. All trial protocols have to be approved by a research ethics committee which checks that the trial is ethical. In particular they should check that the questions being addressed in the trial have not already been



answered, and that people are asked to take part in an appropriate way, with clear information to help them decide whether to take part.

A clinical trial may find that a new strategy, treatment, or device can improve patient outcomes or offer no benefit or cause unexpected harm. However, whatever the result the conduct of the trial is very important because the result of it is able to advance medical knowledge and help improve patient care. In our days, people from all over the world have the opportunity thought the Clinicaltrials.gov (a free online database) to learn more about clinical trials as descriptions, locations, and other vital information about more than 100,000 clinical trials.

1.2 Motivation of the thesis - HERA trial

Besides skin cancer, breast cancer is the most commonly diagnosed cancer among American women. As of March 2017, there are more than 3.1 million women with a history of breast cancer in the U.S (this includes women currently being treated and women who have finished treatment). Also it is estimated that in 2017 just under 30% of newly diagnosed cancers in women will be breast cancers (<http://www.breastcancer.org> accessed at 12/04/2017). Thus more and more trials about breast cancer treatment are conducted the last few years.

Breast cancer and especially a clinical trial about breast cancer in women, which called HERA, is the motivation of this thesis. HERA (HERceptin Adjuvant) trial is an international, multicentre, randomized, open-label, phase 3 trial. HERA enrolled 5102 women with HER2-positive early-stage invasive breast cancer who had completed all locoregional therapy (surgery with or without radiotherapy) and (neo)adjuvant chemotherapy. Patients were randomly allocated to three groups: observation, trastuzumab treatment for 1 year and trastuzumab treatment for 2 years. trastuzumab is a monoclonal antibody that interferes with the HER2/neu receptor and it has established efficacy against breast cancer. The standard of care is 1 year of trastuzumab, but the optimum duration of treatment is unknown.

Initial trials (see Piccart-Gebhart et al., 2005; Romond et al., 2005; Slamon et al., 2011) compared 1 year of trastuzumab treatment with a no trastuzumab control group and showed that there is a persistent benefit of 1 year of treatment



compared with observation alone. So the purpose of HERA trial is firstly to update the comparison of 1 year of trastuzumab versus observation at a median follow-up of 8 years and secondly is the unique trial which allow comparison of two different durations of treatment (1 year vs. 2 years of trastuzumab).

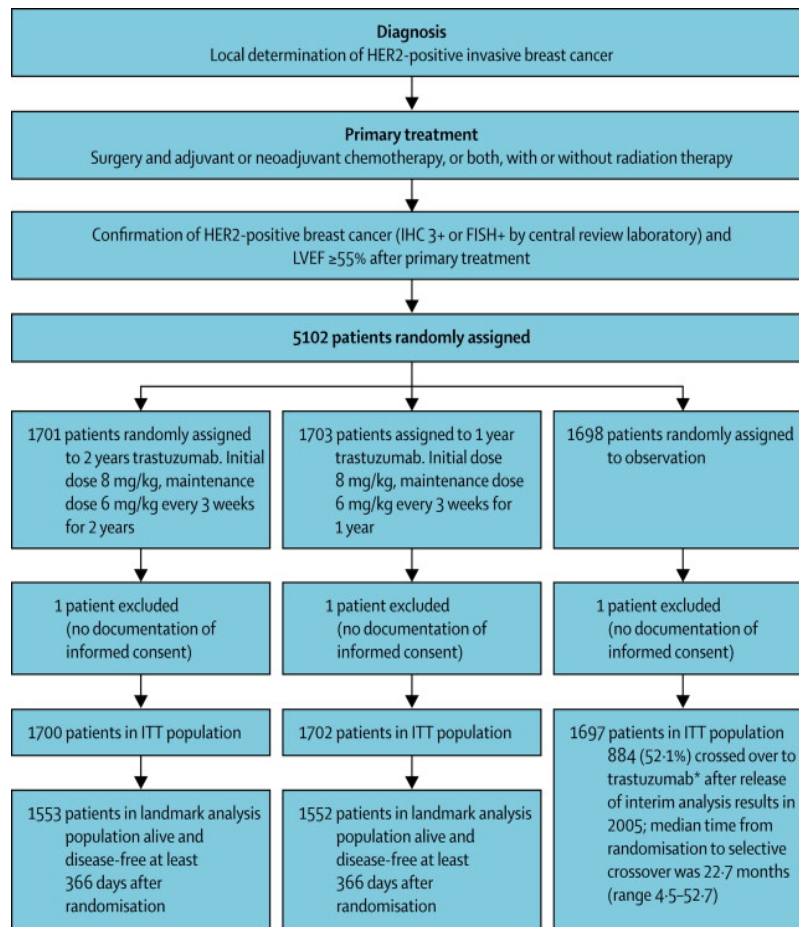


Figure 1.1: Profile of HERA trial

(IHC=immunohistochemistry, FISH=fluorescence in-situ hybridisation, LVEF=left ventricular ejection fraction, ITT=intention to treat.)

The updated comparison of 1 year of trastuzumab versus observation is based on 3399 patients who enrolled in the two groups (1702 patients in 1 year of trastuzumab group and 1697 patients in observation group). Also because of the

initial positive results of the 1 year treatment, in event-free patients in the observation group were offered crossover to receive trastuzumab and finally 884 of the 1697 patients in the observation group selectively crossed over to trastuzumab. The comparison of 2 years versus 1 year of trastuzumab is based on a 12-month landmark analysis (see Dafni, 2011) of the 3105 women who remained alive and disease free for at least 12 months after randomisation to one of the two trastuzumab treatment groups (1553 patients in 1 year of trastuzumab group and 1552 patients in 1 year of trastuzumab group). Figure 1.1 show the profile of HERA trial (see Goldhirsch et al., 2013).

The primary end point was disease-free survival (DFS), defined as time from randomization to the first occurrence of any of the following disease-free survival events: recurrence of breast cancer at any site, the development of ipsilateral or contralateral breast cancer (including ductal carcinoma in situ but not lobular carcinoma in situ), second nonbreast malignant disease other than basal-cell or squamous-cell carcinoma of the skin or carcinoma in situ of the cervix, or death from any cause without documentation of a cancer-related event. Secondary endpoints includes overall survival, sites of first relapse and adverse events (particularly cardiac safety).

The final results of HERA (see Goldhirsch et al., 2013) confirmed that one year of trastuzumab treatment remains the standard of care for people with early-stage HER2-positive breast cancer. The results also showed that after a median follow-up of eight years, the improvements in disease-free survival and overall survival for women who received trastuzumab remained statistically significant compared with observation.

The story of HERA trial is interesting since after the first interim analysis where the new drug showed statistically significant improvement, it was decided to give the opportunity to the observation arm to switch, i.e. to receive therapy with trastuzumab. This, in a great extent created some ambiguity later on since any analyses was blurred due to this crossover. A natural question that arised is what the effect of trastuzumab would be if no crossover was present. Such problems related to crossover are not new in the literature and there are several other trials with similar questions. The literature contains several procedures to handle the



crossover but no method is considered as perfect so far, leaving a challenge to the researchers. One approach can be based on estimating the survival function for the observation arm before the crossover and then simulating the hypothetical survival paths of all those that crossed over in order to compare what the effect would be at this case. To this direction, it is very important to determine the survival function in a parametric manner, i.e. by assuming a specific parametric form in order to be able to simulate from. For HERA data Regan et al. (2012) used a simplistic Weibull assumption for the survival functions but simple plots of the survival function reveal a non-monotonic behavior which cannot be captured by the Weibull distribution. So, the purpose of current thesis is to examine more deeply this assumption and in particular to examine in detail the survival function for all arms of HERA trial, using different parametric models.

Therefore, our own interest focuses on finding the best parametric model for the relevant survival functions in HERA data. For the thesis we use the most recent HERA database as elaborated after the latest data cleaning 2012 (clinical cut-off date April 12,2012). We will try to estimate the hazard functions of 1694, 1698 and 1967 patients who belong to the observation group and the 1 and 2 year treatment arms respectively. R version 3.3.1 software will be used for the analysis of data.

Figure 1.2 display the non-parametric estimator of the hazard function for each of the three groups of the trial. The function estimates the hazard function using kernel-based methods (see section 2.3). The statistical properties of many of these estimators are reported and compared in Hess et al. (1999). As we can see in Figure 1.2 the hazard function for the 1964 patients in the observation group is non-monotonic and it increases at the beginning and after about 1 year falls continuously. Also the hazard function of 1698 patients in the 1 year trastuzumab group increases at the beginning and after almost 2 years decreases. The same happens and for patients in 2 year trastuzumab group with the difference that here the hazard increases again at 7 years but falls very quickly after some days. For the results we have excluded 10 patients totally (3 from observation group, 4 for 1-year treatment group and 3 from 2-year treatment group) with DFS time equal to 0 to avoid numerical problems. That means that distributions such as



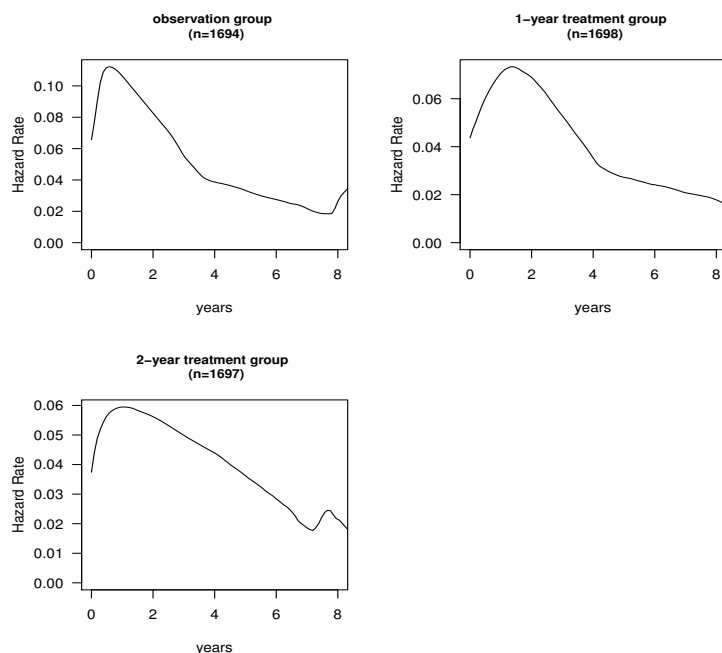


Figure 1.2: Non parametric estimator of the hazard function for the patients in the observation group and the 1 and 2 year treatment groups. (We have excluded patients with DFS time equal to 0 to avoid numerical problems)

Weibull or exponential which are used very often in survival analysis are not able to describe our data successfully. For example Weibull distribution has only monotonic behavior and hence does not seem to be suitable. We try to examine if the idea of a mixture of Weibull distributions can make good fit in our data.

1.3 Problem and structure of the thesis

In many applied sciences such as medicine, modelling and analyzing lifetime data are of central interest. Several lifetime distributions have been used to model such kind of data. The quality of the procedures used in a statistical analysis depends heavily on the assumed probability model or distributions. Because of it, considerable effort has been expended in the development of large classes of standard

probability distributions along with relevant statistical methodologies. However, there still remain many important problems where the real data does not follow any of the classical or standard probability models.

The parametric survival distributions most commonly used in regression modelling (e.g. Weibull, log-normal, log-logistic etc.) have unimodal or monotone hazard rate functions and as a result, these distributions are incapable of modelling hazard functions with more complicated shapes.

There are many cases in survival analysis where hazard rate decrease at the beginning and afterwards increase and look like a bathtub, or first increase and then decrease and look like an inverted (upside down) bathtub. We can say that these hazard functions can be described by three phases in which the hazard rate initially increases/decreases, then becomes essentially constant, and ultimately decreases/increases.

In diseases, such as cancer, many times the hazard rate increases for a finite period of time until reaches a peak and then declines gradually. The main purpose of this thesis is to examine whether mixtures of Weibull distributions, which lead to flexible hazard function, can be considered as models for survival data with non-monotonic survival function.

So in Chapter 2 of this paper we will describe some survival models (exponential, Weibull, gamma and generalized gamma, log-normal and log-logistic) which are used commonly in the analysis of failure time but they have not very flexible hazard functions. Also we will describe the poly-Weibull distribution which differs from the others because it starts from the hazard function and then are calculated the other functions. At the end of this chapter we will refer to the idea of kernels that constitute a very popular non-parametric way for estimation of a hazard function.

In Chapter 3 we will introduce the finite mixture of Weibull distributions. At first we will mention in mixture models generally and after we see the structure of a 2-finite Weibull mixture model. We will describe the Maximum-Likelihood Estimation (MLE) approach about the parameters of the model and then we will bring in the EM algorithm which is used for finding MLEs parameters. Finally, using simulated data we will see how efficient can be the algorithm.



In Chapter 4 we will apply the idea of mixture of Weibull distributions in real data. As we mentioned before, HERA trial is the motivation of this thesis. Except for the Weibull mixture model we will apply all the parametric models which are used commonly in survival analysis and we will compare the models between them. Moreover we will see the characteristics of the sample and with Cox model we will examine which of them influence the survival time. Also with the log-rank test we will confirm the results of HERA trial which indicate that a 1-year treatment with trastuzumab can improve the duration of life for patients with breast cancer.

Concluding in Chapter 5 we will sum up if the Weibull mixture model is suitable for describing our data and what additionally there is about cases with non-monotonic hazard functions.



Chapter 2

Survival models

2.1 Survival analysis and survival data

Survival analysis is used to analyze data in which the outcome variable is the time until the occurrence of an event of interest. The event can be death, occurrence or recurrence of a disease, length of stay in a hospital e.t.c. The response is often referred to as a failure time, survival time, or event time. Subjects are usually followed over a specified time period and the focus is on the time at which the event of interest occurs.

Survival times are typically positive numbers and can be measured in days, weeks, years, etc. Incompletely observed responses are censored. Censoring is present when we have some information about a subject's event time, but we don't know the exact event time. The censoring can be either left or right. Right censoring occurs when a subject leaves the study before an event occurs (for example the subject is lost to follow-up during the study period or withdraws from the study) or the study ends before the event has occurred. Left censoring occurs when the event of interest has already occurred before enrolment but this is very rarely encountered. Essentially the presence of censoring is the most important difference between survival analysis and other statistical methods such as logistic regression.



The dependent variable in survival analysis is composed of two parts:

- the survival time T_i
- the event status, which records if the event of interest occurred or not (indicator) δ_i .

That is for a random sample of size n , with X_i denotes the failure time of the i th subject and C_i denotes the censoring time for the i th subject, the observation is of the form (T_i, δ_i) where,

$T_i = \min(X_i, C_i)$ and

$$\delta_i = \begin{cases} 1 & \text{if } X_i \leq C_i \text{ (not censored observation)} \\ 0 & \text{if } X_i > C_i \text{ (right-censored observation)} \end{cases}$$

One can then estimate two functions that are dependent on time, the survival and hazard functions. The survival and hazard functions are key concepts in survival analysis for describing the distribution of event times. More detailed:

- **Survivor Function**

$$S(t) = Pr(T \geq t) = 1 - F_T(t)$$

where $F_T(t)$ is the cumulative distribution function. The survival function gives the probability that a subject will survive beyond a specified time t . The survival function is non-increasing. At time $t = 0$, $S(t) = 1$. In other words, the probability of surviving past time 0 is 1. At time $t = \infty$, $S(t) = S(\infty) = 0$, i.e. as time goes to infinity, the survival curve goes to 0. In theory, the survival function is smooth but in practice, we observe events on a discrete time scale

For a continuous random variable:

$$S(t) = \int_t^{\infty} f(u) du$$

where $f(t) = \lim_{\Delta t \rightarrow 0} \frac{1}{\Delta t} Pr(t \leq T \leq t + \Delta t)$, i.e. is the p.d.f

- **Hazard Function**

For a continuous random variable:

$$\begin{aligned}
 h(t) &= \lim_{\Delta t \rightarrow 0} \frac{1}{\Delta t} Pr(t \leq T \leq t + \Delta t | T \geq t) \\
 &= \lim_{\Delta t \rightarrow 0} \frac{1}{\Delta t} \frac{Pr([t \leq T \leq t + \Delta t] \cap [T \geq t])}{Pr(T \leq t)} \\
 &= \lim_{\Delta t \rightarrow 0} \frac{1}{\Delta t} \frac{Pr(t \leq T \leq t + \Delta t)}{Pr(T \leq t)} = \frac{f(t)}{S(t)}
 \end{aligned}$$

As it is obvious, if we know one of the above functions ($f(t)$, $S(t)$, $h(t)$ or $H(t)$) we can calculate the others. In survival analysis except for these functions there are many other quantities of interest such as mean and median survival time which subsequently are estimated from knowing either the hazard or survival function. Also it is generally of interest in survival studies to describe the relationship of a factor of interest (e.g. treatment) to the time to event, in the presence of several covariates, such as age, gender, race, e.t.c.. A huge number of models are available to analyze the relationship of a set of predictor variables with the survival time. Methods include parametric, non-parametric and semi-parametric approaches.

2.2 Model fitting

There are essentially three approaches to fitting survival models:

1. Parametric survival models where we assume a specific functional form for the hazard function. Some common distributions which are used widely are exponential, Weibull, gamma and generalized gamma, log-normal and log-logistic and we describe them below.



2. Semi-parametric survival models, where we make mild assumptions about the hazard function.
3. Non-parametric strategy that focuses on estimation of the regression coefficients leaving the baseline hazard function completely unspecified. Kernels are used very often for the estimation of the coefficients and we introduce this way of estimation below.

In this thesis we will focus on parametric survival models and we will try to find out the distribution which describe our data as well as possible. Also we will deal with the idea of kernels because we first need a non-parametric estimator to examine if the assumption about the distribution in parametric survival analysis is proper.

2.2.1 Common families of survival distributions

- **Exponential distribution**

Let's denote $T \sim Exp(\lambda)$. For $t > 0$ we have:

Probability density function: $f(t) = \lambda e^{-\lambda t}$,
where $\lambda > 0$ (rate or inverse scale parameter)

Cumulative distribution function: $F(t) = 1 - e^{-\lambda t}$

Survival function: $S(t) = e^{-\lambda t}$

Hazard function: $h(t) = \lambda$

The exponential model is the simplest parametric model because as we can see in Figure 2.1 one of the characteristics of exponential distribution is that the hazard function is constant over time and equals to parameter λ , i.e. the probability to die within a particular time interval depends only on the length but not on the location of this interval. Also exponential is a memoryless probability distribution, i.e. if an event has not occurred after



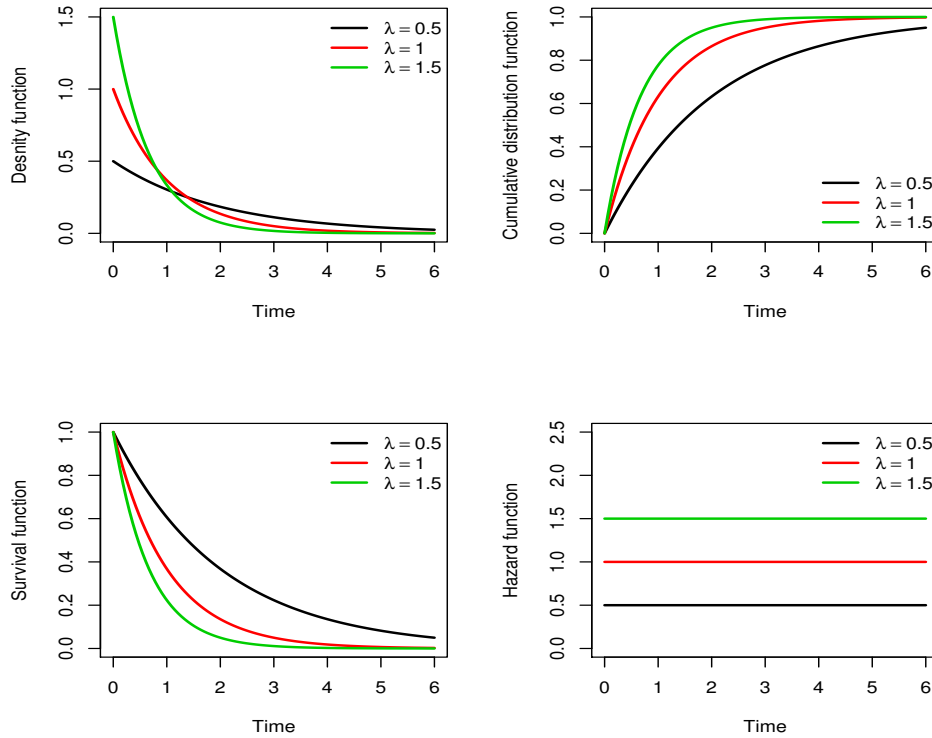


Figure 2.1: Density, cumulative, survival and hazard functions for exponential distribution for different values of parameter λ

30 seconds, the conditional probability that occurrence will take at least 10 more seconds is equal to the unconditional probability of observing the event more than 10 seconds relative to the initial time. The model is very sensitive to even a modest variation because it has only one adjustable parameter, the inverse of which is both mean and standard deviation.

- **Weibull distribution**

Let's denote $T \sim W(a,b)$. For $t > 0$ we have:

Probability density function: $f(t) = \frac{a}{b} \left(\frac{t}{b}\right)^{a-1} e^{-\left(\frac{t}{b}\right)^a}$,

where $a > 0$ (shape parameter) and $b > 0$ (scale parameter)

Cumulative distribution function: $F(t) = 1 - e^{-\left(\frac{t}{b}\right)^a}$

Survival function: $S(t) = e^{-\left(\frac{t}{b}\right)^a}$

Hazard function: $h(t) = a\left(\frac{1}{b}\right)^a t^{a-1}$

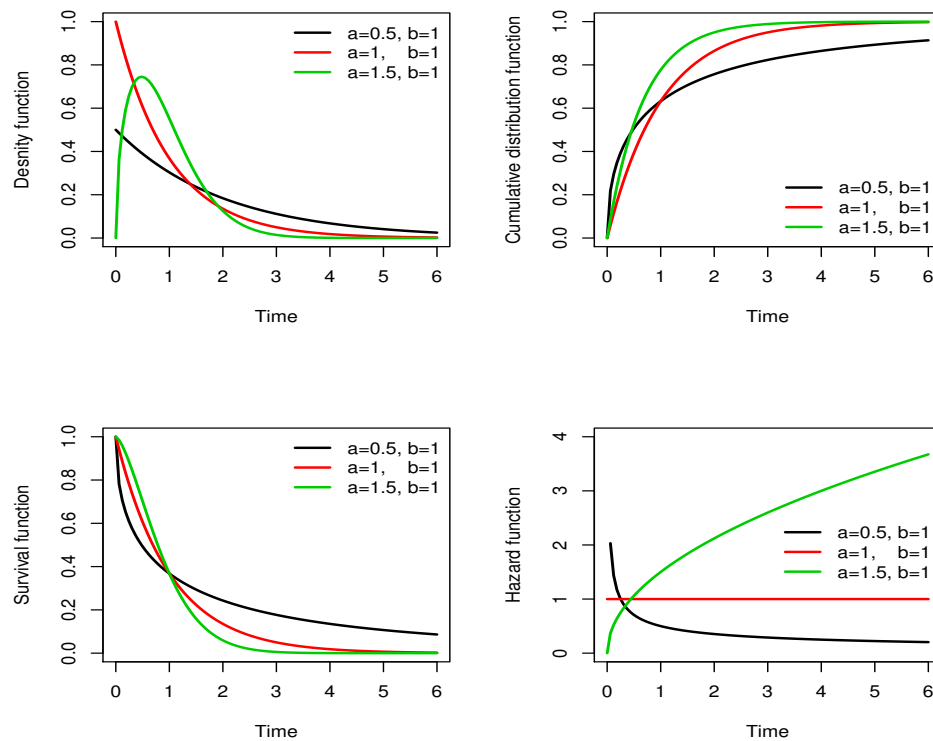


Figure 2.2: Density, cumulative, survival and hazard functions for Weibull distribution for different values of shape parameter a and scale parameter $b = 1$

We can say that Weibull distribution is a generalization of the exponential distribution and except for the scale parameters it also depends on a second shape parameter. The second parameter in the model allows great flexibility of the model and different shapes of the hazard function. The convenience of the Weibull model stems on the one hand from this flexibility and on the other from the simplicity of the hazard and survival functions. As it is obvious from Figure 2.2 a value of $a > 1$ indicates that the failure rate increases with time and a value of $a < 1$ indicates that failure rate decreases over time. For $a = 1$ the Weibull distribution reduces to an exponential distribution with scale parameter $= \frac{1}{\lambda}$, where the failure rate is constant over time as we see above. The Weibull distribution is inappropriate when the hazard rate is indicated to be unimodal or bathtub-shaped

- **Gamma distribution**

Let's denote $T \sim G(a, b)$. For $t > 0$ we have:

Probability density function: $f(t) = \frac{t^{a-1}e^{-\frac{t}{b}}}{b^a\Gamma(a)}$,

where $a > 0$ (shape parameter) and $b > 0$ (scale parameter)

and $\Gamma(a) = \int_0^\infty t^{a-1}e^{-t} dt$ is the gamma function

Cumulative distribution function: $F(t) = 1 - \frac{\Gamma(a, \frac{t}{b})}{\Gamma(a)} = \frac{\gamma(a, \frac{t}{b})}{\Gamma(a)}$,

where $\Gamma(a)$ is the Gamma function evaluated at a

and $\gamma(a, t) = \int_0^t x^{a-1}e^{-x} dx$ is the lower incomplete gamma function

Survival function: $S(t) = 1 - \frac{\gamma(a, \frac{t}{b})}{\Gamma(a)}$

Hazard function: $h(t) = \frac{t^{a-1}e^{-\frac{t}{b}}}{b^a(\Gamma(a) - \gamma(a, \frac{t}{b}))}$

Gamma distribution can be considered as a generalization of the exponential because with shape parameter $a=1$ and scale parameter $b=\frac{1}{\lambda}$ identify with exponential distribution. The gamma distribution is of limited use in survival analysis because the gamma models do not have closed form expressions for survival and hazard functions because both include the incomplete



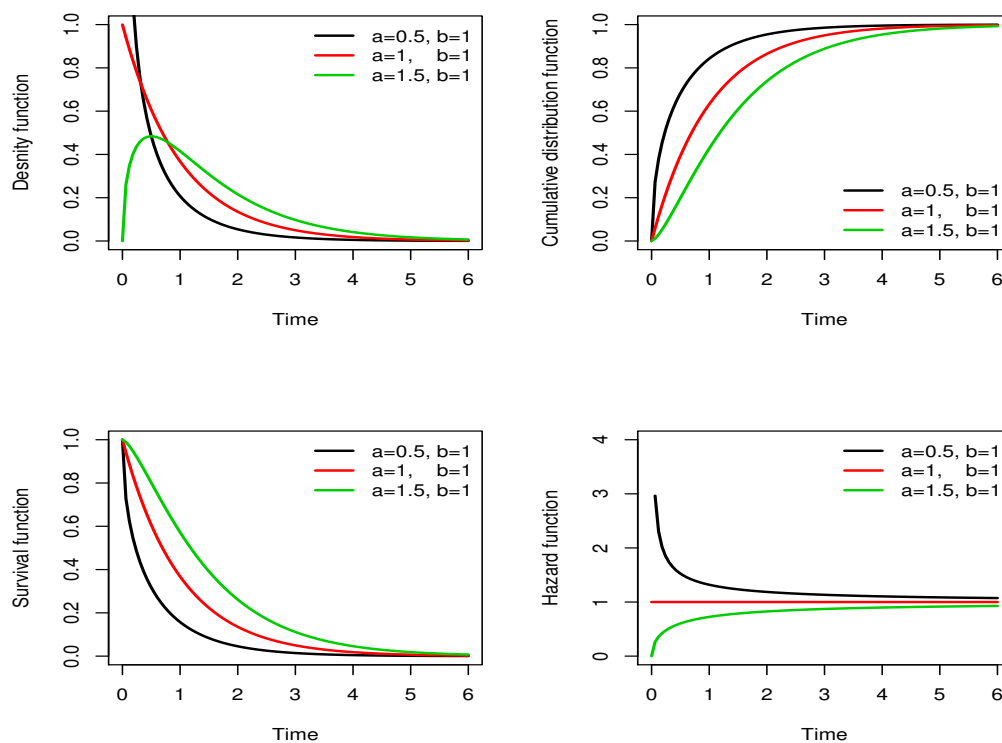


Figure 2.3: Density, cumulative, survival and hazard functions for gamma distribution for different values of shape parameter a and scale parameter $b = 1$

gamma integral. However, as we can see in Figure 2.3 gamma distribution sometimes gives flexible shapes about hazard function. When $a > 1$, the hazard function is concave and increasing and when $a < 1$ the hazard function is convex and decreasing. The case $a = 1$ corresponds to the exponential distribution, where the hazard function is constant.

- **Log-normal distribution**

Let's denote $T \sim LN(\mu, \sigma)$. For $t \in (-\infty, +\infty)$ we have:

Probability density function: $f(t) = \frac{1}{\sqrt{2\pi\sigma t}} e^{-\frac{(\ln t - \mu)^2}{2\sigma^2}}$,

where $\mu \in (-\infty, +\infty)$ (location parameter) and $\sigma > 0$ (scale parameter)

Cumulative distribution function: $F(t) = \Phi\left(\frac{\ln t - \mu}{\sigma}\right)$,

where Φ is the cumulative function of the standard normal distribution,

$$\Phi(t) = \int_{-\infty}^t \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx$$

Survival function: $S(t) = 1 - \Phi\left(\frac{\ln t - \mu}{\sigma}\right)$

Hazard function: $h(t) = \frac{f(t)}{S(t)} = \frac{\frac{1}{\sqrt{2\pi\sigma t}} e^{-\frac{(\ln t - \mu)^2}{2\sigma^2}}}{\Phi\left(\frac{\ln t - \mu}{\sigma}\right)}$

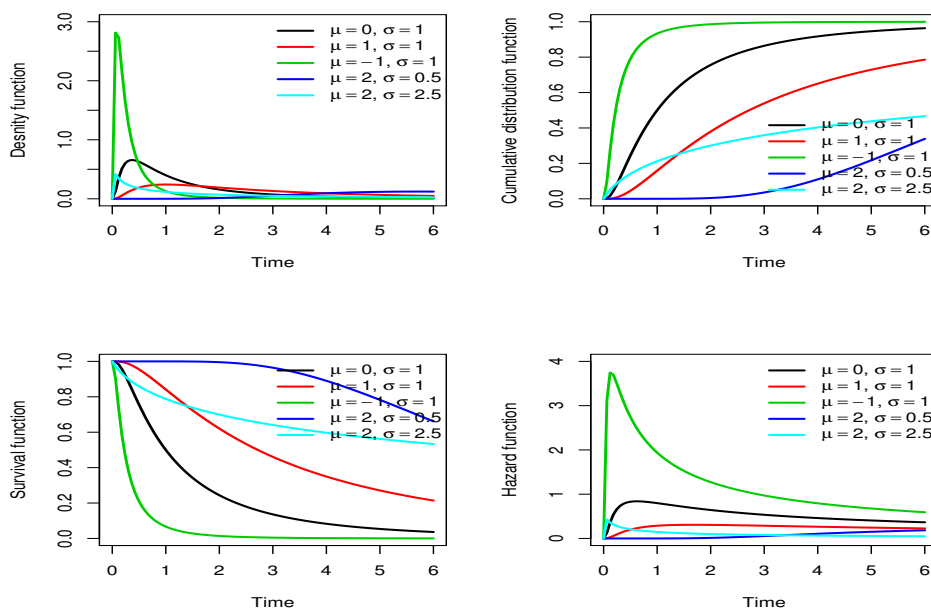


Figure 2.4: Density, cumulative, survival and hazard functions for log-normal distribution for different values of location and scale parameters

Log-normal distribution is not used very much in survival analysis because it hasn't got closed form expressions for survival and hazard functions and it is essential to compute the above integrals which make estimation a little bit difficult. The log-normal distribution may be convenient to use with non-censored data, but when this distribution is applied to censored data, the computations quickly become formidable. As we can see in Figure 2.4 the hazard function has a strange form. It has value zero at $t = 0$, increases to a maximum and then decreases, approaching zero as t heads to infinity. Because of the decreasing form of the hazard function for older ages, the distributions seem implausible as a lifetime model in most situations. Nevertheless, it makes sense if interest is focused on time periods of younger ages. Despite its unattractive features, the log-normal distribution has been widely used as failure distribution in diverse situations, such as the analysis of electrical insulation or time to occurrence of lung cancer among smokers (see Stone et al., 2004; Tai et al., 2007).

- **Log-logistic distribution**

Let's denote $T \sim LL(A, B)$. For $t > 0$ we have:

Probability density function: $f(t) = \frac{\frac{b}{a}(\frac{t}{a})^{b-1}}{(1+(\frac{t}{a})^b)^2}$,
where $a > 0$ (scale parameter) and $b > 0$ (shape parameter)

Cumulative distribution function: $F(t) = \frac{1}{(1+(\frac{t}{a})^b)^{-b}} = \frac{(\frac{t}{a})^b}{1+(\frac{t}{a})^b}$

Survival function: $S(t) = (1 + (\frac{t}{a})^b)^{-1}$

Hazard function: $h(t) = \frac{f(t)}{S(t)} = \frac{\frac{b}{a}(\frac{t}{a})^{b-1}}{1+(\frac{t}{a})^b}$

Log-logistic distribution is an alternative model to the Weibull distribution. The general shape of the hazard function of a log-logistic distribution is very similar to that of the log-normal distribution. As we can see in Figure 2.5 the log-logistic distribution has a fairly flexible functional form and it is one of the parametric survival time models in which the hazard rate may be



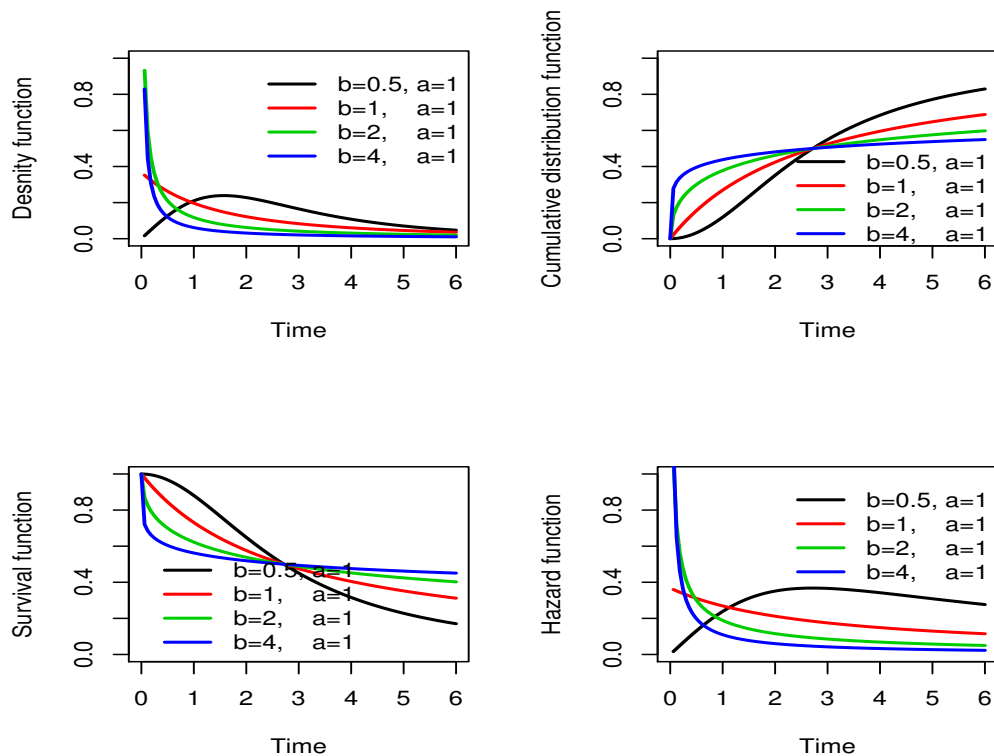


Figure 2.5: Density, cumulative, survival and hazard functions for log-logistic distribution for different values of shape parameter b and scale parameter $a = 1$

decreasing, increasing or even hump-shaped, which mean that it initially increases and then decreases.

- **Generalized Gamma distribution**

Let's denote $T \sim GG(a, d, p)$. For $t > 0$ we have:

Probability density function:
$$f(t) = \frac{\left(\frac{p}{a^d}\right)t^{d-1}e^{-\left(\frac{t}{a}\right)^p}}{\Gamma\left(\frac{d}{p}\right)},$$

where $a, d, p > 0$

Cumulative distribution function: $F(t) = \frac{\gamma(\frac{d}{p}, (\frac{t}{p})^a)}{\Gamma(\frac{d}{p})}$

where $\Gamma(\cdot)$ is the Gamma function

and $\gamma(\cdot, t)$ is the lower incomplete Gamma function

Survival function: $S(t) = 1 - \frac{\gamma(\frac{d}{p}, (\frac{t}{p})^a)}{\Gamma(\frac{d}{p})}$

Hazard function: $h(t) = \frac{(\frac{p}{a})t^{d-1}e^{-(\frac{t}{p})^a}}{\Gamma(\frac{d}{p}) - \gamma(\frac{d}{p}, (\frac{t}{p})^a)}$

The Generalized Gamma distribution can be viewed as a generalization of the Exponential, Weibull and Gamma distributions and differs from the other ones because it has 3 parameters. For $p=1$, Generalized Gamma distribution correspond to Gamma distribution, for $d=p$ identifies with Weibull distribution and for $a=d=1$ it is an Exponential distribution. The number of the parameters declares that the hazard function can be severally flexible and this is a big advantage about survival analysis. The distribution has already been used to fit the survival curve for breast cancer data (see Ardoino et al., 2012, Abadi et al., 2012)

- **Poly-Weibull distribution** (see Demiris et al., 2015)

Poly-Weibull distribution is an extension of Weibull distribution. The Poly-Weibull distributions arises in applications involving competing risks in survival analysis. The idea is that an individual is subject to m independent sources of risk that operate additively. We assume further that the distribution of each of the components may be sufficiently described by a Weibull form with density function $f(t) = \nu\lambda t^{\nu-1}e^{-t^\nu}$. (This is an alternative parameterization about Weibull distribution, which is used mostly in Medical Statistics and its form is simpler than the form we describe above. In this form of distribution ν is the scale parameter and is the same as parameter a above and λ which corresponds to $\frac{1}{b^a}$ is the scale parameter. Also the form of hazard function is $h(t) = \nu\lambda t^{\lambda-1}$). Then, the observed event time is said to follow a poly-Weibull distribution.

The hazard function of a Poly-Weibull distribution arises as the sum of the



m independent Weibull hazards functions:

$$h(t) = \sum_{i=1}^m \nu_i \lambda_i t^{\nu_i - 1}$$

So, we can find that the form of survival function is:

$$S(t) = \exp\left(-\int_0^t h(s) ds\right) = \exp\left(-\sum_{i=1}^m \lambda_i t^{\nu_i}\right)$$

and the density function has the following form:

$$f(t) = h(t)S(t) = \sum_{i=1}^m \nu_i \lambda_i t^{\nu_i - 1} \exp\left(-\sum_{i=1}^m \lambda_i t^{\nu_i}\right)$$

The difference with the other distributions which we introduced before and are used very commonly in survival analysis is that all the idea begin from the Hazard function and the other functions arise from it. Conversely, in all cases above we started from the probability density function and after we found the type of the Hazard and the Survival function.

The advantage over the Weibull distribution is that allows not only increasing, constant or decreasing hazard functions with zero or non-zero asymptotes but also non-monotone hazard functions, like hazards with "bathtub" shapes. Unfortunately, the main disadvantage of Poly-Weibull distribution is the inability to work when causes of death is unknown and not reported.

2.3 Hazard function estimation with kernels

While parametric models provide convenient ways to analyze lifetime data, the necessary model assumptions, when violated, can lead to erroneous analyses and thus need to be checked carefully. In these cases a non-parametric approach is considered essential, as the estimation is more flexible, model-free and data-driven. Also the non-parametric approach is useful because it is one of the ways to examine whether the assumptions in a parametric survival analysis are logical. So



one method to estimate the Hazard function without assumptions about the distribution of the data is to use the kernel estimators which are used very often at Survival analysis.

The estimation of hazard rates for continuously observed data is conceptually close to density estimation. As probability density function is the derivative of cumulative distribution function, we can also consider the hazard rate function as the derivative of the cumulative hazard function, i.e. $H(t) = \int_0^t h(x) dx$. A hazard rate estimate can thus be obtained, analogous to a density estimate, by smoothing the increments of an estimate of $H(t)$.

Watson and Leadbetter (1964a) were the first who propose and study such a smoothed hazard estimator using the cumulative hazard estimate based on an independent and identically distributed sample of lifetimes. They propose the following type hazard estimator:

$$\widehat{h}_n(t) = \int W_n(t-x) dH_n(t),$$

where W_n is a sequence of smooth functions approaching the Dirac delta function for large n . This delta-sequence method is quite general and covers several types of smoothing methods, including the kernel method. Hazard estimators in this situation are ordinarily obtained by smoothing the increments of the Nelson-Aalen estimator $H_n(\cdot)$ for the cumulative hazard function $H(t)$. The Nelson-Aalen estimator is a non-parametric estimator of the cumulative hazard rate function in case of censored data and it is used in survival analysis to estimate the cumulative number of expected events. The estimator is given by the following type:

$$\widehat{H}(t) = \sum_{t_i \leq t} \frac{d_i}{n_i},$$

where d_i is the number of events at t_i and n_i is the total number of individuals at risk at t_i .

Thus, using the Nelson-Aalen estimator for the cumulative hazard function $H(t)$ and choosing $W_n(t) = \frac{1}{b} K(\frac{t-x}{b})$ for a particular choice of kernel K and band-



width $b = b_n$, we end up at the following kernel hazard estimator:

$$\begin{aligned}\widehat{h(t)} &= \int \frac{1}{b} K\left(\frac{t-x}{b}\right) H_n(t) \\ &= \sum_{t_i \leq t} \int \frac{1}{b} K\left(\frac{t-x_i}{b}\right) \frac{d_i}{n_i}\end{aligned}$$

Asymptotic properties on consistency are typically obtained under the following assumptions:

- i the true hazard rate is k -times differentiable for $k \leq 0$
- ii for the bandwidth force that $b_n \rightarrow 0$ and $nb_n \rightarrow \infty$
- iii the kernel is of order k and defined as:

$$\int K(x)dx = 1, \quad \int K^2(x)dx < \infty, \quad \int x^j K(x)dx = 0 \text{ for } 1 < j < k,$$

$$\text{and } \int x^k K(x)dx \text{ is finite but nonzero}$$

Kernels have interesting statistical interpretation and lead to smooth estimations. Obviously questions are created about the choice of the function, the order and the bandwidth of the kernel. Also the rate of convergence of the kernel hazard estimation depends on the order of the kernel, the bandwidth and the differentiability of the hazard function.

The choice of the kernel is important but not crucial. Often non-negative kernels are used in practice.

Epanechnikov kernel: $K(x) = 0.75(1 - x^2)$, $-1 \leq x \leq 1$ and

Gaussian kernel : $K(x) = (2\pi)^{-\frac{1}{2}} \exp(-\frac{x^2}{2})$

constitute the most common choices. Typically, the order k of the kernel is chosen to be an even number with $k = 2$ being the standard choice.

The choice of the bandwidth is of crucial importance and regulates the trade off between the bias and variance of the estimator of hazard function $h(t)$. A small bandwidth yields a less smooth curve, with smaller bias but larger variance, as compared to a larger bandwidth. Bandwidth choice is particularly crucial for



hazard estimation near the right boundary of the data as the variance increases to infinity there.

The bandwidth for a kernel hazard estimate can be fixed at all points (global bandwidth b) or can vary for different points (local bandwidth $b(t)$). Usually a global bandwidth is employed for a smooth density or regression. However, for the hazard estimation situation discussed here there are compelling reasons to adopt local rather than global bandwidth choices. The main reason about preference in local bandwidth is that the variance of the kernel hazard estimate extends to infinity as t approaches the right boundary of the data and thus the variance tends to dominate the bias in the right tail and this needs to be compensated for by a larger bandwidth.

The optimal local bandwidth of $\widehat{h}(t)$ which minimizes the leading term of $\text{MSE}(\widehat{h}(t))$ is:

$$b(t) = n^{-1/(2k+1)} \left\{ \frac{1}{2k} \frac{h(t)}{[1 - F(t)][1 - G(t)]} \frac{V}{[h^{(k)}(t)B_k]^2} \right\}^{1/(2k+1)}$$

where $B_k = (-1)^k k! \int x^k K(x) dx$ and $V = \int K^2(x) dx < \infty$

For the optimal global bandwidth, we have to restrict the range of t to a compact interval $[0, \tau]$ with $F(\tau) < 1$ and $G(\tau) < 1$. The global optimal bandwidth which minimizes the leading term of

$\text{MISE}(\widehat{h}(t)) = E \int_0^\tau [\widehat{h}(x) - h(x)]^2 dx$ is:

$$b_{opt} = n^{-1/(2k+1)} \left\{ \frac{1}{2k} \int_0^\tau \frac{h(x)}{[1 - F(x)][1 - G(x)]} dx \frac{V}{B_k^2 \int_0^\tau [h^{(k)}(y)]^2 dy} \right\}^{1/(2k+1)}$$

In order to understand the significance of the bandwidth choice we will use the 'ovarian' data (Survival in a randomized trial comparing two treatments for ovarian cancer with $n=26$ patients) from the package 'survival' in R software (see Therneau, 2015) and with the assistance of package 'muhaz' (a package for estimation of the hazard function from right-censored data using kernel-based methods, see Hess and Gentleman, 2015) we will estimate the hazard rate. In Figure 2.6 we can see how different is the estimation of the hazard function for various values of the bandwidth. For the value 40 of the bandwidth we observe



that is far from the reality but the optimal bandwidth gives a great estimation.

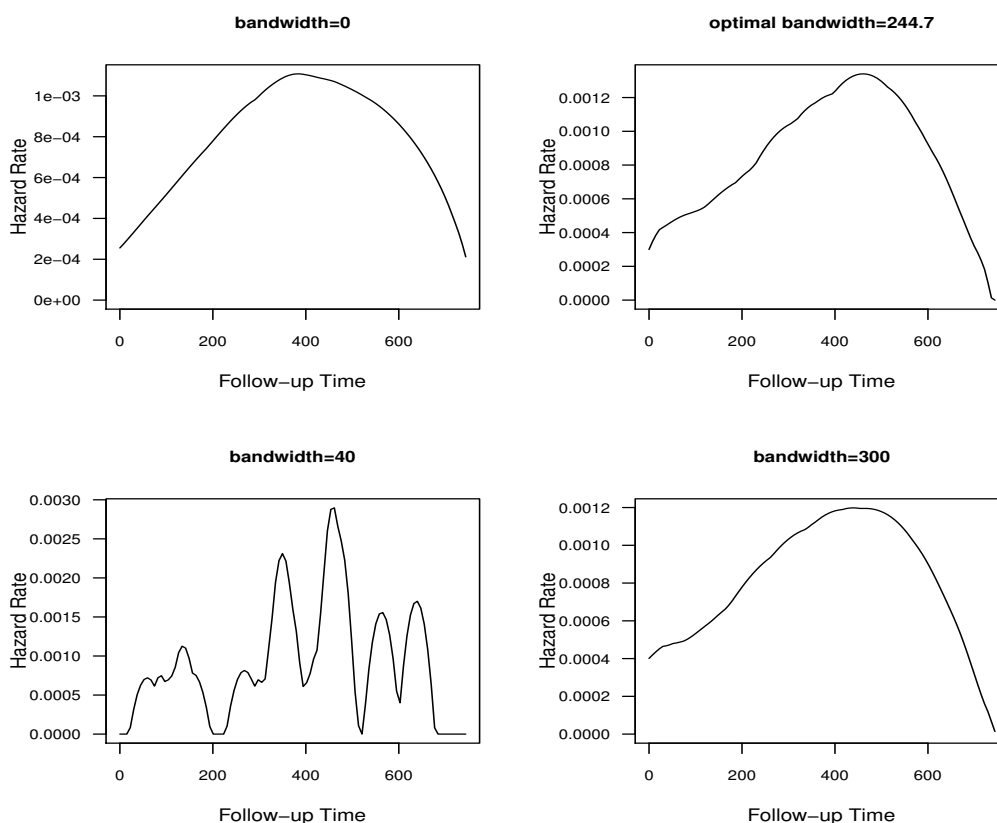


Figure 2.6: Estimation of the hazard function using kernels for different values of (global) bandwidth (kernel function=Epanechnikov)

Also, another issue that is worth to mention is that the kernel smoothing method needs to be employed very carefully near the boundary as there is a bias problem in such regions. This problem usually referred to in the literature as boundary effects. Boundary effects may be attributed to the fact that the support of the kernel exceeds the available range of data and are not unique to hazard estimates.

An unmodified kernel estimate is unreliable in the boundary region, which

is the region within one bandwidth of the largest or smallest observations. To remedy the boundary effects, different kernels, referred to as 'boundary kernels' can be used within the boundary region. As a consequence, varying kernels are employed at each location t and the bandwidths are affected accordingly. The resulting kernel estimate with varying kernels and varying local bandwidths takes the form below:

$$\widehat{h}(t) = \int \frac{1}{b(t)} K_t \left(\frac{t-x}{b(t)} \right) dh(x) \quad (2.1)$$

So, it is obvious that kernels are very useful in statistics because it is not need to assume anything about the distribution of the population, while even the distribution will be estimated by data. In case of hazard function, a kernel smoother which is a statistical technique for estimating the real valued function $h(x)$ by using its noisy observations, when no parametric model for the function is known. The estimated function is smooth, and the level of smoothness is set by a single parameter. Some examples about kernel smoothers are the Gaussian kernel smoother, the nearest neighbor smoother, the kernel average smoother, the local linear regression and the local polynomial regression. Library `muhaz` in R, which we use about a non-parametric estimation of the hazard function of our data, makes use of nearest neighbor smoother.



Chapter 3

Mixture of Weibull distributions

One of the problems that is encountered in survival analysis is the inappropriateness of the common distributions which are described in Chapter 2 and used widely. Most of these distributions can not fit well data with strange (non-monotonic) shapes about the hazard function. So in parametric survival analysis, the necessity of distributions with more flexible shapes about hazard function is considered huge.

An idea is to use a mixture of one of the known distributions or a mixture with more than one of them. We know that the hazard function in Weibull distribution can be either ascending (shape parameter >1) or descending (shape parameter <1) (see Figure 2.2). So, for example, if the hazard rate increases at the beginning and after a few time decreases, it is a good idea to use a mixture of two Weibull distributions, where the first one will have shape parameter greater than 1 and the second one will have shape parameter smaller than 1. We prefer Weibull distribution rather than other distributions because it has simple forms about hazard and survival functions and it has been used in many surveys in the past.

The usage of mixture model for parametric survival models is no new. See for example the work of Mc Lachlan and Mc Giffin (1994) about finite mixture models in survival analysis. More specifically, for mixtures of Weibull one can see



the early work of Farawell (1982), Greenhouse and Silliman (1996) and recently by Farcomeni and Nardi (2010). The latter model is similar to the one used in this thesis. A thorough examination on the failure rate of the mixture of Weibull distributions can be found in Jiang and Murthy (1998).

3.1 Mixture models

Before we introduce the idea of finite mixture of Weibull distributions, we refer to mixture models generally. All the construction of mixture models rely on the Law of Total Probability according to:

if B_1, B_2, \dots is a partition of the sample space S , then for any event A we have $P(A) = \sum_{i=1}^{\infty} P(A|B_i)P(B_i)$

This theory can be used and in cases where instead of probabilities we have random variables or density probability functions.

Supposing that the population consists of K sub-populations and for each sub-population we know its density, we can calculate the probability density function, the cumulative distribution function, the survival and the hazard functions of all the population.

Probability density function (pdf):

$$g(t) = \sum_{j=1}^K p_j f_j(t),$$

where $f_j(t)$ is the pdf of the j -th sub-population and p_j is the probability that a random selected individual comes from the j -th population, $p_j > 0$ and $\sum_{j=1}^K p_j = 1$

Cumulative distribution function (cdf):

$$G(t) = \sum_{j=1}^K p_j F_j(t),$$

where $F_j(t)$ is the cdf of the j -th sub-population associated with $f_j(t)$



Survival function:

$$S(t) = \sum_{j=1}^K p_j S_j(t)$$

Hazard function:

$$h(t) = \sum_{j=1}^K w_j(t) h_j(t),$$

where $h_j(t)$ is the hazard function of the j -th sub-population and

$$w_j(t) = \frac{p_j S_j(T)}{\sum_{j=1}^K p_j S_j}, \quad \sum_{j=1}^K w_j(t) = 1 \quad \text{with} \quad S_j(t) = 1 - F_j(t)$$

We can see that the hazard rate for a general K -finite mixture model is a weighted mean of the hazard rate for the sub-populations with weights varying with t , $t \geq 0$.

Mixture models find huge applications in many fields of Statistics (see for example Mc Lachlan and Basford, 1987; Frhwirth-Schnatter, 2006; Schlattmann, 2009; Karlis and Santourian, 2009; Huang Y. et al, 2015). The most important application of mixture models is that beginning from a simple model of one distribution we can construct very flexible models. Mixture models of known distributions can have a variety of properties and thus they are very useful in real data. For example we can have models with skewness functions, or functions with fat tails and more than one modes. Hence, models like those offer great flexibility, because we can model phenomena with various properties. Other marked applications of mixture models are in Cluster analysis, in Random effect models and also they are very useful as a way of simulation.

3.1.1 Weibull mixture model

As we had described the idea of mixture models, we can now introduce the special case of the **2-finite Weibull mixture model** (see Farcomeni and Nardi, 2010) . If $f_1(t)$ follows Weibull(a_1, b_1) and $f_2(t)$ follows Weibull(a_2, b_2) distributions where $a_1, a_2 > 0$ are shape parameters and $b_1, b_2 > 0$ are scale parameters, the



Probability density function (pdf) for 2-finite Weibull mixture model is:

$$g(t) = pf_1(t) + (1 - p)f_2(t)$$

$$= p\left[\frac{a_1}{b_1}\left(\frac{x}{b_1}\right)^{a_1-1}e^{-\left(\frac{x}{b_1}\right)^{a_1}}\right] + (1 - p)\left[\frac{a_2}{b_2}\left(\frac{x}{b_2}\right)^{a_2-1}e^{-\left(\frac{x}{b_2}\right)^{a_2}}\right]$$

The corresponding Cumulative distribution function is:

$$G(t) = pF_1(t) + (1 - p)F_2(t)$$

$$= \left[1 - e^{-\left(\frac{t}{b_2}\right)^{a_2}}\right] + p\left[e^{-\left(\frac{t}{a_2}\right)^{b_2}} - e^{-\left(\frac{t}{b_1}\right)^{a_1}}\right]$$

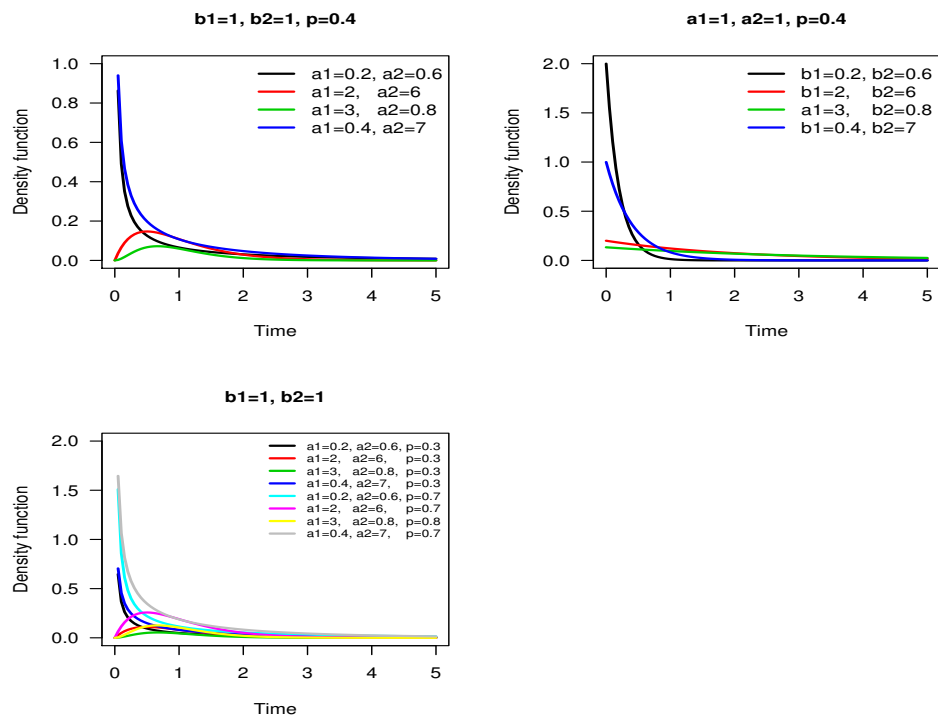


Figure 3.1: Probability density function for 2-finite Weibull mixture model for different values of the parameters a_1, a_2, b_1, b_2 and p

The Survival function is:

$$S(t) = p[e^{-\left(\frac{t}{b_1}\right)^{a_1}}] + (1 - p)[e^{-\left(\frac{t}{b_2}\right)^{a_2}}]$$

And the hazard function is:

$$h(t) = \frac{g(t)}{S(t)} = \frac{pf_1(t) + (1 - p)f_2(t)}{p[e^{-\left(\frac{t}{b_1}\right)^{a_1}}] + (1 - p)[e^{-\left(\frac{t}{b_2}\right)^{a_2}}]} \quad \text{with } t \geq 0$$

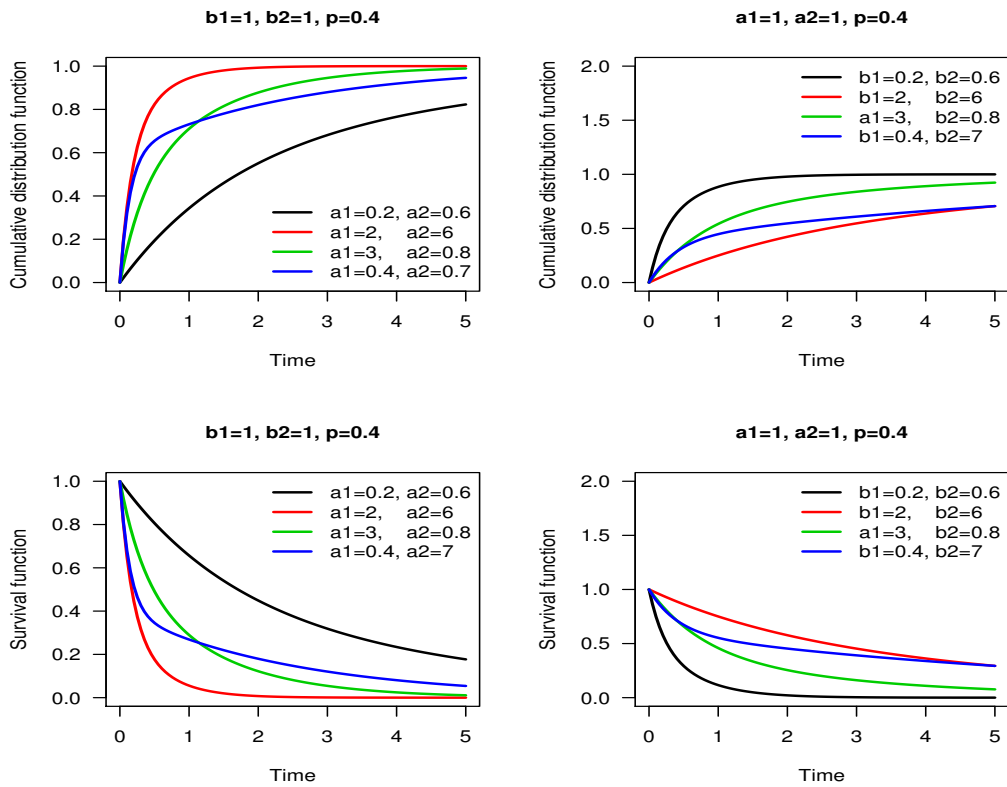


Figure 3.2: Cumulative distribution and Survival functions for 2-finite Weibull mixture model for different values of the parameters a_1, a_2, b_1, b_2 and p



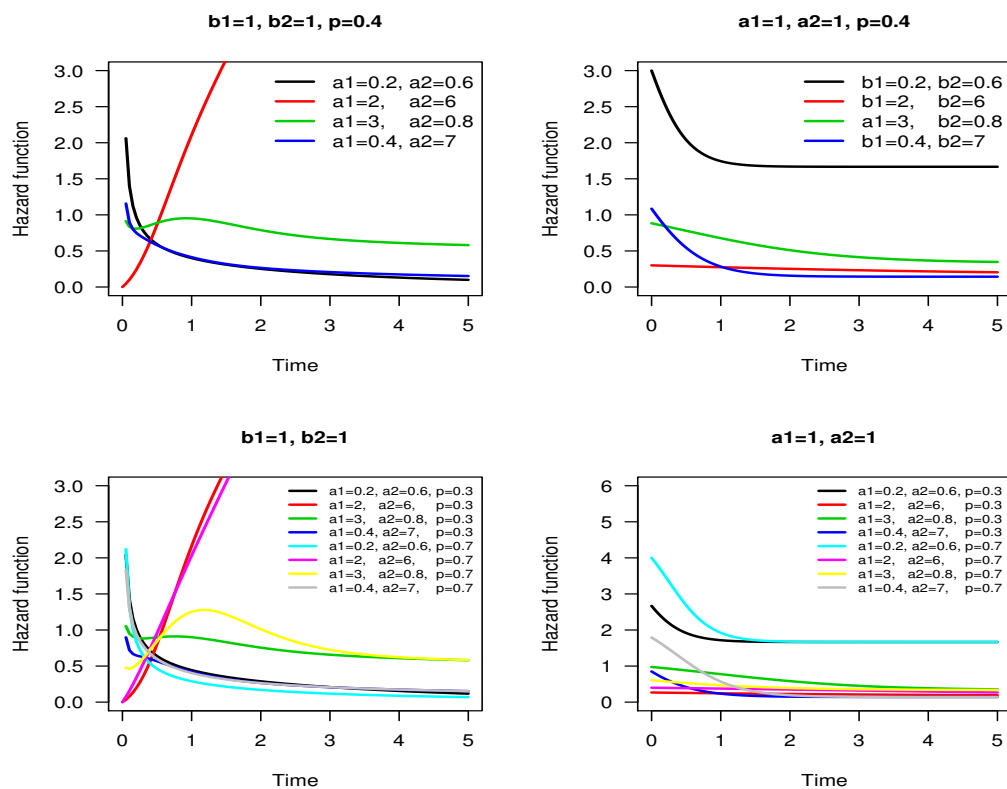


Figure 3.3: Hazard function for 2-finite Weibull mixture model for different values of the parameters a_1, a_2, b_1, b_2 and p

In Figures 3.1, 3.2 and 3.3 we can see the Density function, the Cumulative distribution function, the Survival function and the Hazard function for the 2-finite Weibull mixture model respectively. As we can notice the Hazard function of the model can take various shapes and especially the green line which corresponds to shape parameters $a_1 = 3 > 1$ and $a_2 = 0.8 < 1$ and scale parameters b_1, b_2 equal to 1 looks like the non-parametric estimation of the hazard function of HERA data (see Figure 1.2, Chapter 1).

3.2 Estimation of model parameters

There are many methods for estimating the parameters of mixture distributions. Both graphical and analytical approaches have been used. About the analytical methods we know that these methods start from Pearsons (1894) method of moments, through the formal maximum likelihood approaches, general curve fitting, Bayesian and so on. In this thesis we will describe the maximum likelihood method and it will be used for estimating the parameters of the mixture Weibull distribution.

3.2.1 Maximum-Likelihood estimation for mixture models

We know that the likelihood is the probability of observing our data, as a function of the parameters. For observations x_1, x_2, \dots, x_n the likelihood function is:

$$L(\theta) = \prod_{i=1}^n f(x_i; \theta)$$

For easiness, we use the log-likelihood function:

$$l(\theta) = \sum_{i=1}^n \log f(x_i; \theta)$$

and in case of mixture models the log-likelihood function becomes as below:

$$l(\theta) = \sum_{i=1}^n \log g(x_i) = \sum_{i=1}^n \log \left[\sum_{k=1}^K p_k f(x_i; \theta_k) \right]$$

The maximum likelihood estimates of the parameters are obtained by taking the partial derivatives of the log-likelihood function with respect to each of the parameters of the model and setting to zero. The derivative of $l(\theta)$ with respect to one



parameter, say θ_j is:

$$\begin{aligned} \frac{dl}{d\theta_j} &= \sum_{i=1}^n \frac{1}{\sum_{k=1}^K p_k f(x_i; \theta_k)} p_j \frac{df(x_i; \theta_j)}{d\theta_j} \\ &= \sum_{i=1}^n \frac{p_j f(x_i; \theta_j)}{\sum_{k=1}^K p_k f(x_i; \theta_k)} \frac{1}{f(x_i; \theta_j)} \frac{df(x_i; \theta_j)}{d\theta_j} \\ &= \sum_{i=1}^n \frac{p_j f(x_i; \theta_j)}{\sum_{k=1}^K p_k f(x_i; \theta_k)} \frac{d \log f(x_i; \theta_j)}{d\theta_j} \end{aligned}$$

In an ordinary parametric model the derivative of the log-likelihood is

$$\sum_{i=1}^n \frac{d \log f(x_i; \theta_j)}{d\theta_j}$$

So maximizing the likelihood for a mixture model is like doing a weighted likelihood maximization, where the weight of x_i depends on cluster, is

$$w_{ij} = \frac{p_j f(x_i; \theta_j)}{\sum_{k=1}^K p_k f(x_i; \theta_k)}$$

The problem is that these weights depend on the parameters we are trying to estimate. Lets look at these weights a bit more. We define the discrete random variable Z which says, which component X (data) is drawn from, so $Z \sim \text{Multinomial}(p_1, p_2, \dots, p_K)$, where p_j is the probability that the hidden class variable Z is j . Now we can see that the numerator in the weights is the joint probability of getting $Z = j$ and $X = x_i$ and the denominator is the marginal probability of getting $X = x_i$, so the ratio is the conditional probability of $Z = j$ given $X = x_i$

$$w_{ij} = \frac{p_j f(x_i; \theta_j)}{\sum_{k=1}^K p_k f(x_i; \theta_k)} = \frac{P(Z = j, X = x_i)}{P(X = x_i)} = P(Z = j | X = x_i; \theta).$$

Thus if we try to estimate the mixture model, then, were doing weighted maximum likelihood, with weights given by the posterior cluster probabilities.



3.2.2 MLEs for a 2-finite Weibull mixture model

Our interest is to estimate the hazard function of the data using the Maximum Likelihood method. In Survival analysis, we assume that the data contain the failure/censored random variable T_i and the failure/censored indicator δ_i (if the i -th observation is failure, i.e. the event has occurred, then $\delta_i = 1$ and if it is censored, i.e. the event has not occurred, then $\delta_i = 0$), $i=1,2, \dots, n$. Then the likelihood function under random censoring scheme for the data is given by

$$L(\theta) = \prod_{i=1}^n [f(t_i)]^{\delta_i} [1 - F(t_i)]^{1-\delta_i}$$

where θ is the parameter vector for the assumed model.

Taking log on both sides of the equation we have the log-likelihood function

$$\log(L(\theta)) = \sum_{i=1}^n [\delta_i \log f(t_i) + (1 - \delta_i) \log \{1 - F(t_i)\}]$$

In the case of the mixture of 2 Weibull distributions with $\theta = (a_1, a_2, b_1, b_2, p)$ the log-likelihood function becomes

$$\begin{aligned} \log(L(\theta)) &= \sum_{i=1}^n [\delta_i \log g(t_i) + (1 - \delta_i) \log \{1 - G(t_i)\}] \\ &= \sum_{i=1}^n \{ \delta_i \log [p \left[\frac{a_1}{b_1} \left(\frac{x}{b_1} \right)^{a_1-1} e^{-\left(\frac{x}{b_1} \right)^{a_1}} \right] + (1 - p) \left[\frac{a_2}{b_2} \left(\frac{x}{b_2} \right)^{a_2-1} e^{-\left(\frac{x}{b_2} \right)^{a_2}} \right] \right] \\ &\quad + (1 - \delta_i) \log [e^{-\left(\frac{x}{b_2} \right)^{a_2}} - p [e^{-\left(\frac{x}{b_2} \right)^{a_2}} - e^{-\left(\frac{x}{b_1} \right)^{a_1}}]] \} \end{aligned}$$

The maximum likelihood estimates of the parameters are obtained by taking the partial derivatives with respect to a_1, a_2, b_1, b_2 , and p and setting to zero. However, the maximum likelihood estimating equations that we obtain do not give closed form solutions for the parameters $\theta = (a_1, a_2, b_1, b_2, p)$. So it requires a numerical iterative procedure for finding the MLEs of the model parameters. EM algorithm is used very common for this operation.



3.3 EM algorithm

EM algorithm (see Dempster et al., 1977) is a moderate new iterative algorithm for estimation with the Maximum Likelihood (ML) method. In fact, is a mathematical maximizing method with great statistical interpretation and offers important help in simplification of estimation problems with the ML method. The algorithm is known in Statistics for many years and it has been described in its general form.

The algorithm is used when we have missing data or we can express the problem like we have missing data. Such examples are:

- trimmed distributions (distributions that we can not observe some values)
- mixture models (as missing data we consider the labels that will say us from which sub-population originate each observation)
- censored data (many times in survival analysis for some patients we know that they live more than a specific time T but we do not know the exact time)
- missing data (in many data sets there are missing observations that we want to know)

Algorithm owes its name in the two steps that constitute it. The E-step (Expectation step) and the M-step (Maximization step). The main idea is that in E-step we estimate missing data with the information we have until that moment (i.e. the observations and the values of estimation until that moment) and in M-step we use these estimations in order to maximize the likelihood function, renewing the estimations' values.

3.3.1 EM algorithm for mixture models

It is very important when we want to use the EM algorithm to find a way to define proper missing data that enable us to maximize the likelihood function.

Here we discuss the EM algorithm for finding the MLEs of the parameters of a general K -finite mixture model with parameters $\Theta = (p_1, \dots, p_K, \theta_1, \dots, \theta_K)$, where p_j are the mixing parameters and θ_j are the parameters for the density



function f_j , $j = 1, 2, \dots, K$.

- Let $t = (t_1, \dots, t_n)$ denotes the observed random sample obtained from the mixture density.
- In this case the missing data is the knowledge of which distribution each observation in the sample comes from. These missing data can be represented by the random vector $Z = (Z_1, Z_2, \dots, Z_n)$ where $Z_i = (Z_{i1}, Z_{i2}, \dots, Z_{iK})$ and

$$Z_{ij} = \begin{cases} 1 & \text{if } T_i \text{ belongs to distribution } j \\ 0 & \text{otherwise} \end{cases}$$

($i = 1, 2, \dots, n$ and $j = 1, 2, \dots, K$).

- The EM algorithm handles the unobservable data to the problem by working with the current conditional expectation of the complete-data log likelihood given the observed data. Based on $T = (T_1, \dots, T_n)$ and Z , the complete data likelihood function will be

$$L_c(\theta|t, z) = \prod_{i=1}^n \prod_{j=1}^K \{ [p_j f_j(t_i; \theta_j)]^{Z_{ij} \delta_i} [p_j S_j(t_i; \theta_j)]^{Z_{ij} (1-\delta_i)} \}.$$

Each iteration of the EM algorithm constitutes of two steps.

1 Expectation or E-step

At E-step we compute the conditional expectation of the complete-data log-likelihood for Θ given the current estimate of the parameter vector and the observed data, which at the $(m + 1)$ th iteration can be expressed as below:

$$\begin{aligned} Q(\Theta, \Theta^{(m)}) &= E_{\theta^{(m)}} [\log L_c(\theta; T)] \\ &= \sum_{i=1}^n \sum_{j=1}^K E_{\Theta^{(m)}} (Z_{ij}|t) \delta_i \log [p_j^{(m)} f_j(t_i|\theta_j^{(m)})] \\ &+ \sum_{i=1}^n \sum_{j=1}^K E_{\Theta^{(m)}} (Z_{ij}|t) (1 - \delta_i) \log [p_j^{(m)} S_j(t_i|\theta_j^{(m)})] \end{aligned}$$



And we end up that

$$Q(\Theta, \Theta^{(m)}) = \sum_{i=1}^n \sum_{j=1}^K [\log(p_j) E_{\Theta^{(m)}}(Z_{ij}|t) + \delta_i \log[f_j(t_i|\theta_j^{(m)})] E_{\Theta^{(m)}}(Z_{ij}|t) + (1 - \delta_i) \log[f_j(t_i|\theta_j^{(m)})] E_{\Theta^{(m)}}(Z_{ij}|t)]$$

The equation is linear in the unobservable data z_{ij} and the E-step (on the $(m + 1)$ th iteration) simply requires the calculation of the current conditional expectation of Z_{ij} given the observed data t , where Z_{ij} is the random variable corresponding to z_{ij} . Thus,

$$E_{\Theta^{(m)}}(Z_{ij}|t) = z_{ij}^{(m)}$$

And as we saw before, at description of Maximum Likelihood Estimation here z_{ij} are the posterior probabilities which can be expressed using the Bayess theorem as

$$z_{ij}^{(m)} = \frac{p_j^{(m)} [\delta_j f_j(t_i|\theta_j^{(m)}) + (1 - \delta_i) S_j(t_i|\theta_j^{(m)})]}{\sum_{j=1}^m p_j^{(m)} [\delta_j f_j(t_i|\theta_j^{(m)}) + (1 - \delta_i) S_j(t_i|\theta_j^{(m)})]}$$

2 Maximization or M-step

At M-step we maximize the conditional expectation of the complete-data log-likelihood $Q(\Theta, \Theta^{(m)})$ with respect to the parameters in order to obtain new parameter estimations $\Theta^{(m+1)}$. We can maximize the term containing p_j and the term containing θ_j independently since they are not related. For each parameter we take the derivative of $Q(\Theta, \Theta^{(m)})$ with respect to each of them severally and we set equal to zero. For some distributions it is possible to get closed-form analytical expressions for θ_j but for others, like Weibull distribution, there is not this possibility. To find the expression for p_j , we use the Lagrange multiplier λ with the constraint $\sum_{j=1}^K p_j = 1$. Under this constraint, if we take the derivative of $Q(\Theta, \Theta^{(m)})$ with respect to p_j and



setting equal to zero, we get

$$\sum_{i=1}^n \frac{1}{p_j} z_{ij}^{(m)} + \lambda = 0$$

Summing both sides over j and using $\sum_{j=1}^m z_{ij}^{(m)} = 1$, we get that $\lambda = -n$. So we end up to

$$p_j^{(m+1)} = \frac{1}{n} \sum_{i=1}^n z_{ij}^{(m)} \quad (3.1)$$

The E-step and M-step are iterated until the algorithm converges.

3.3.2 EM algorithm for a 2-finite Weibull mixture model

Now, we are going to describe the EM algorithm for estimating the parameters of a K -finite Weibull mixture distribution step by step. In this way we will construct the procedure for the EM algorithm in statistical package R. The procedure extensively is:

- *Step 1* Begin with initial values of $p_j^{(0)}$, $a_j^{(0)}$ and $b_j^{(0)}$ for $j = 1, 2, \dots, K$ (a =shape parameter and b =scale parameter > 0)
- *Step 2* Using the initial values of $p_j^{(0)}$, $a_j^{(0)}$ and $b_j^{(0)}$ at m -th iteration calculate the conditional expectation of z_{ij} from the type above
- *Step 3* At the $(m+1)$ -th iteration, find the MLEs of $p_j^{(m+1)}$, $a_j^{(m+1)}$ and $b_j^{(m+1)}$ as follows:
 - 1 Find the MLE for $p_j^{(m+1)}$ from (3.1)
 - 2 Estimate $a_j^{(m+1)}$, and $b_j^{(m+1)}$ maximizing the weighted log-likelihood function as it arises from the type of $Q(\Theta, \Theta^{(m)})$ with $z_{ij}^{(m)}$ as weights.
- *Step 4* Repeat Steps 2 and 3 until the algorithm converges with a desired accuracy.



3.3.3 Convergence of the algorithm

The algorithm stops when a convergence criterion is satisfied. There are many convergence criteria.

The first category of criteria stops the iterations, when the relative increase of log-likelihood between two consecutive iterations is smaller than a very very small value tol (eg. $tol=10^{-10}$). So the criterion has the below form:

$$\left| \frac{L^{(r+1)} - L^{(r)}}{L^{(r+1)}} \right| \leq tol,$$

where $L^{(r)}$ is the log-likelihood function after the r iteration.

The other category of criteria stops the iterations, when the values of the parameters do not change between two consecutive iterations, i.e. the biggest difference between each parameter in two consecutive iterations is smaller than a very very small value tol . So the criterion has the below form:

$$\max_j (|\theta_j^{(r+1)} - \theta_j^{(r)}|) \leq tol$$

Virtually, both of convergence criteria examine if there is a change from iteration to iteration and no if the algorithm convergence. Unfortunately, there is not a criterion that check clearly if the EM algorithm convergences.

3.3.4 Advantages and disadvantages of the algorithm

A few of the advantages of EM algorithm generally (applying to our case too) are the following:

- Monotonous convergence: in every iteration the likelihood function increases. This situation allow us to decide when we will stop the repetitions, but under no circumstances it is not certain that we have not been entrapped in a local and no in a total maximum of likelihood function.
- Estimations are in the allowed limits if the initial values are in the allowed limits. This is no assured, when we use other methods of maximization. For



instance, the algorithm Newton-Raphson is able to lead in a solution out of the allowed space.

- Easy programming: virtually the algorithm we just described can be easily implemented in almost every statistical package.

The disadvantages of algorithm are the following:

- The result is determined by the initial values, hence we need good primary values. We have to emphasize that in the case of EM, good initial values are equivalent to fast convergence in few iterations, while in other methods, if we don not have good initial values, this may could be mean that the algorithm never converges
- We can find local and no total maximum. Therefore we need to start from different initial values in order to be sure. As a consequence the calculating load is bigger.
- Slow convergence: the convergence of algorithm is much slower than other algorithms. Of course, there are ways to accelerate the algorithm, but this is not very easy be occurred.
- Other algorithms use the second derivatives of log-likelihood function. This means that we can easily estimate standard errors of the estimators. Something like this is not true for the EM algorithm.

3.3.5 Simulation study about the EM algorithm

It is very important to examine how well the algorithm can work and how close to reality are the results that we obtain. The only way to check the adequacy of the algorithm is to simulate data from a known distribution with known parameters and after the implementation of the algorithm to check how close are the values of parameters that we obtain from the algorithm with true values. Also another key issue is how the algorithm is affected by the initial values and how this influences the convergence of it.



In this section of the thesis we will simulate data from:

- i) a 2 finite Weibull mixture model
- ii) a 2 finite Gaussian mixture model.

and we will deal with all the above issues. For the Weibull mixture model we have described the steps of the algorithm in detail in previous section and for the Gaussian mixture model we will describe briefly the steps of the algorithm.

- **Two-finite Weibull mixture model**

We generate 1000 values from a Weibull distribution with shape parameter $a_1 = 2$ and scale parameter $b_1 = 3$ and 1000 values from a Weibull distribution with shape parameter $a_2 = 1$ and scale parameter $b_2 = 4$. The probability an observation comes from the first distribution is 0.3 and from the second Weibull distribution is 0.7.

As we can see in Table 3.1 the algorithm operates very well, because the estimated values for all the 5 parameters are very close to the true values. Also in Figure 3.4, we see that the algorithm converges very quickly and more specific at about first 40 iterations.

Parameters	True values	Estimated values with EM algorithm
a_1	2	1.9822
b_1	3	2.9192
a_2	1	0.9888
b_2	4	3.9959
p	0.3	0.2793

Table 3.1: Estimated values with EM algorithm for simulated data from a 2-finite Weibull mixture model

- **Two-finite Gaussian mixture model**

Every normal distribution is a version of the standard normal distribution whose domain has been stretched by a factor σ (the standard deviation) and then translated by μ (the mean value). So we can express the pdf of a Normal distribution as:

$$f(x|\mu, \sigma) = \frac{1}{\sigma} \phi\left(\frac{x - \mu}{\sigma}\right)$$



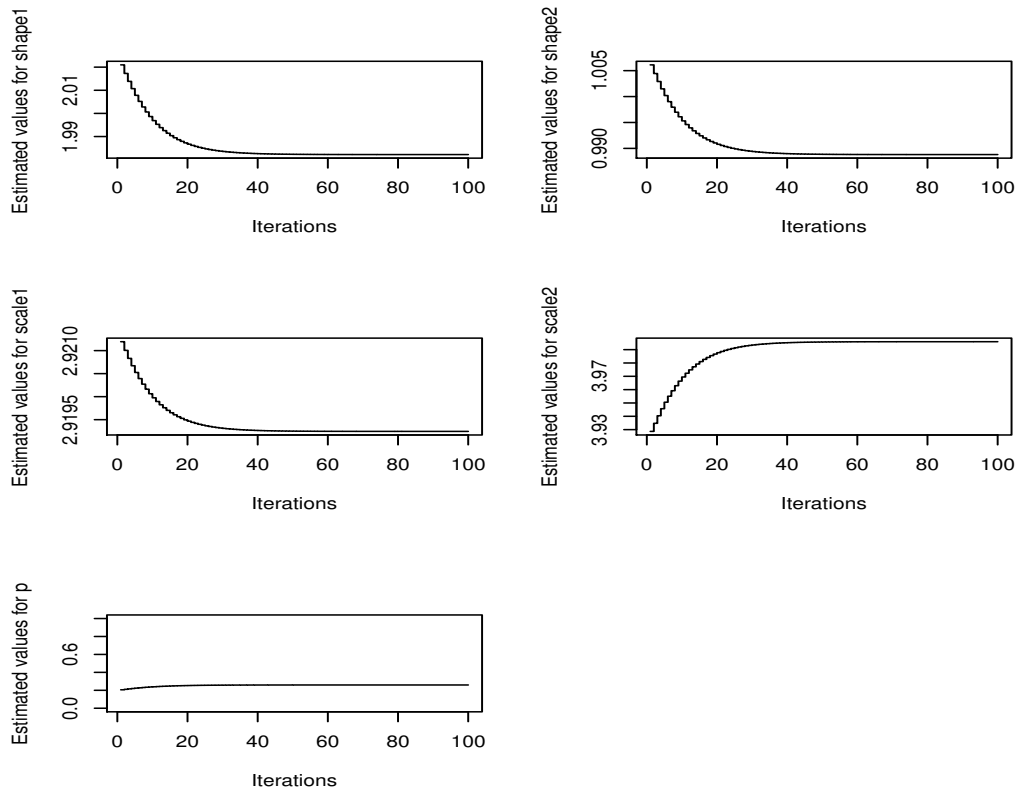


Figure 3.4: Convergence of EM algorithm about each parameter of the 2-finite Weibull mixture model

For simplification of the mathematical forms which will follow below, we will use this type for the Normal distribution and not the analytical.

Hence, the density for the mixture of two Gaussian populations is

$$f(x|\theta) = p \frac{1}{\sigma_1} \phi\left(\frac{x - \mu_1}{\sigma_1}\right) + (1 - p) \frac{1}{\sigma_2} \phi\left(\frac{x - \mu_2}{\sigma_2}\right),$$

where $\theta = (p, \mu_1, \mu_2, \sigma_1, \sigma_2)$ and more specific: μ_1 and σ_1 are the mean and the standard deviation of the first normal distribution, μ_2 and σ_2 are the

mean and the standard deviation of the second normal distribution and p represents the proportion an observation comes from the first normal distribution. We don't know in which of the two populations an observation comes from, so we let

$$Z_i = \begin{cases} 1 & \text{if } X_i \text{ belongs to the first Normal distribution} \\ 0 & \text{if } X_i \text{ belongs to the second Normal distribution} \end{cases}$$

This indicator variable is the missing data in this case of the mixture model and $Z_i \sim \text{Bernoulli}$ with parameter p .

Therefore, the complete likelihood is given by

$$L_c(\theta|X, Z) = \prod_i^n p_i^{z_i} \frac{1}{\sigma_1^{z_i}} \phi\left(\frac{x_i - \mu_1}{\sigma_1}\right)^{z_i} (1 - p_i)^{1-z_i} \frac{1}{\sigma_2^{1-z_i}} \phi\left(\frac{x_i - \mu_2}{\sigma_2}\right)^{1-z_i}$$

At the E-step:

We compute the expectation of Z_i conditional on the observed information and the current parameter estimates $\theta^{(k)}$

$$w_i^{(k)} = E_{Z_i|X_i, \theta^{(k)}}[Z_i] = \frac{p^{(k)} \frac{1}{\sigma_1^{(k)}} \phi\left(\frac{x_i - \mu_1^{(k)}}{\sigma_1^{(k)}}\right)}{p^{(k)} \frac{1}{\sigma_1^{(k)}} \phi\left(\frac{x_i - \mu_1^{(k)}}{\sigma_1^{(k)}}\right) + (1 - p^{(k)}) \frac{1}{\sigma_2^{(k)}} \phi\left(\frac{x_i - \mu_2^{(k)}}{\sigma_2^{(k)}}\right)}$$

At the M-step:

We set the first derivatives of $Q(\theta|\theta^{(k)}) = E_{Z_i|X_i, \theta^{(k)}}[L_0(\theta|X, Z)]$ with respect to each parameter equal to zero and we obtain the following results about the parameters:

$$p^{(k+1)} = \frac{1}{n} \sum_{i=1}^n w_i^{(k)}$$

$$\mu_1^{(k+1)} = \frac{\sum_{i=1}^n w_i^{(k)} x_i}{\sum_{i=1}^n w_i^{(k)}}$$



$$\mu_2^{(k+1)} = \frac{\sum_{i=1}^n (1 - w_i^{(k)}) x_i}{\sum_{i=1}^n (1 - w_i^{(k)})}$$

$$\sigma_1^{(k+1)} = \frac{\sum_{i=1}^n w_i^{(k)} (x_i - \mu_1^{(k+1)})^2}{\sum_{i=1}^n w_i^{(k)}}$$

$$\sigma_2^{(k+1)} = \frac{\sum_{i=1}^n (1 - w_i^{(k)}) (x_i - \mu_2^{(k+1)})^2}{\sum_{i=1}^n (1 - w_i^{(k)})}$$

We generate 1000 values from a Normal distribution with $\mu_1 = 2$ and $\sigma_1^2 = 1.5$ and 1000 values from a Normal distribution with $\mu_2 = 6$ and $\sigma_2^2 = 0.3$. The probability an observation comes from the first Normal distribution is 0.3 and the probability an observation comes from the second Normal distribution is 0.7.

As we can see in Table 3.2 the estimated values that we obtain from the EM algorithm about the parameters of the Normal mixture model is very close to the real values. We used as initial values, (a) values close to and (b) far from them that we simulate, but the result was exactly the same. The only difference as we can see in Figure 3.6 is that the algorithm converges to the final estimated value slower in case that we had started with values far from the true.

Parameters	True values	Initial values		Estimated values with EM algorithm
		case a	case b	
μ_1	2	3	10	2.0041
μ_2	6	9	40	5.9915
σ_1^2	1.5	1	7	1.4648
σ_2^2	0.3	0.5	17	0.3057
p	0.3	0.5	0.9	0.3042

Table 3.2: Estimated values with EM algorithm for simulated data from a 2-finite Gaussian mixture model



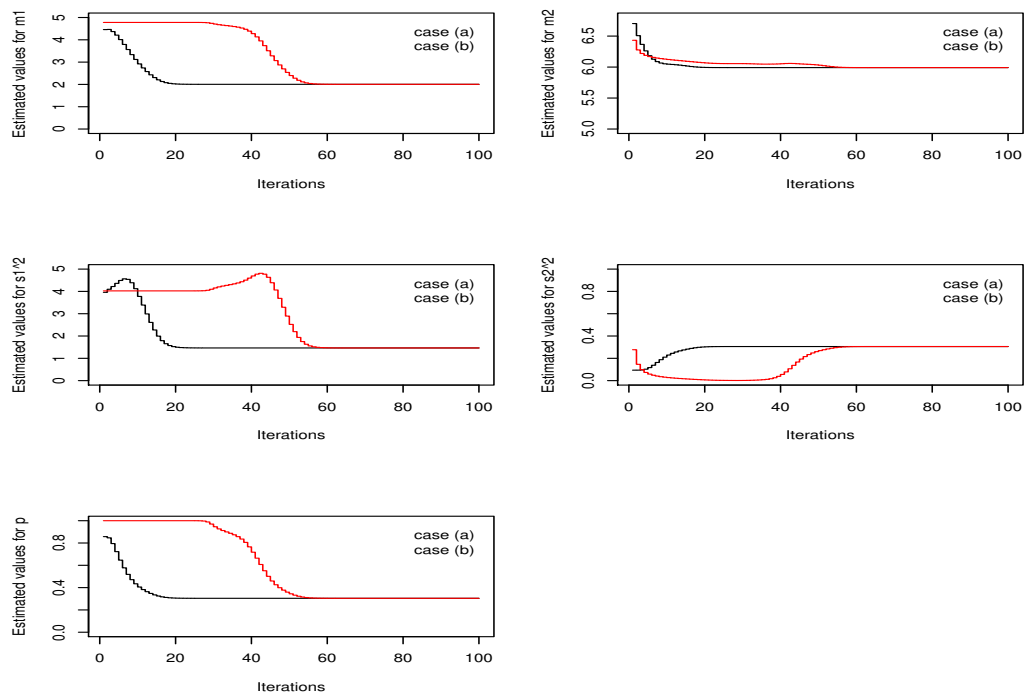


Figure 3.5: Converge of EM algorithm about each parameter of the 2-finite Gaussian mixture model

Chapter 4

Data analysis with HERA data

In order to understand how efficient can be the idea of a Weibull mixture model is necessary to fit this model in real data, which has the attribute of a non-monotonic hazard function. In HERA trial the hazard function in each of the three groups looks like an inverse "bathtub" (see Figure 1.2). HERA (Herceptin Adjuvant) trial is a phase III randomized trial involving women with HER2-positive early-stage invasive breast cancer. For this thesis, we use the most recent HERA database as elaborated after the latest data cleaning during 2012 (clinical cut-off date April 12,2012). Between Dec 7, 2001, and June 20, 2005, a total of 5102 patients were randomly allocated to three groups: observation, trastuzumab for 1 year, and trastuzumab for 2 years. More specific, the HERA trial population consists of 1677 patients randomly assigned to the observation group and 3402 to trastuzumab, receiving treatment for one year (1702 patients) or two years (1700 patients) (see Figure 1.1). We included these two groups for three reasons: a major peak in the rate of relapse occurs 18 to 24 months after surgery, effective treatment of HER2-positive breast cancer may require prolonged attenuation of HER2 activity and tamoxifen, which is an effective targeted therapy for breast cancer, is most beneficial when given for longer than one year.

For each patient was recorded some baseline and tumor characteristics, which described below thoroughly.

- Region of patient (1: Western and Northern Europe, Canada, South Africa,



Australia, New Zealand 2: Asia Pacific, Japan 3: Eastern Europe 4: Central and South America)

- Race of patient (1: Caucasian 2: Oriental 3: Other)
- Age of patient (1: <35 yrs 2: 35-49 yrs 3: 50-59 yrs 4: ≥ 60 yrs)
- Menopausal status (1: pre-menopausal (regular period) 2: peri-menopausal (irregular period) 3: post-menopausal (no longer having period))
- Nodal status (1: Not Assessed 2: Negative (lymph nodes do not contain cancer) 3: 1-3 positive (contain cancer) lymph nodes 4: ≥ 4 positive (contain cancer) lymph nodes)
- Estrogen-Receptor (1: ER Positive (cancer has receptors for estrogen, which means that cancer cells, like normal breast cells, may receive signals from estrogen that could promote their growth) 2: ER Negative (cancer has not receptors for estrogen))
- Progesterone-Receptor (1: PR Positive (cancer has progesterone receptors, which means that cancer cells may receive signals from progesterone that could promote their growth) 2: PR Negative (cancer has not progesterone receptors))
- Adjuvant Chemotherapy (1: No Anthracyclines 2: Anthracyclines, no Taxanes 3: Anthracyclines and Taxanes, (Anthracyclines and Taxanes are chemotherapeutic agents))
- ECOG Performance Status (1: 0 \equiv Fully active, able to carry on all pre-disease performance without restriction 2: 1 \equiv Restricted in physically strenuous activity but ambulatory and able to carry out work of a light or sedentary nature, e.g., light house work, office work)
- Grade of tumor (1: G1 \equiv the tumor cells and the organization of the tumor tissue appear close to normal and they tend to grow and spread slowly 2: G2 \equiv the tumor cells and the organization of the tumor tissue look slightly



abnormal and they are slow-growing 3: G3 \equiv the tumor cells and the organization of the tumor tissue are malignant and they tend to grow and spread rapidly 4: GX \equiv Grade cannot be assessed)

- Tumor size in cm (1: 0-2cm 2: 3-5cm 3: >5cm)

At first, we will present the frequency distributions for all the characteristics above. Consequently applying with the long-rank test, we will examine if the survival time differs between patients in the observation group and the patients who received trastuzumab (either for one or two years), as well as we will find which of the characteristics of patients affect their survival time using a semi-parametric multivariate Cox model. Ultimately we will try to examine if a finite mixture model of Weibull distributions is more appropriate than other parametric models commonly used in survival analysis, in order to estimate the patients' hazard function in each of the three groups of HERA trial.

4.1 Descriptive statistics and crosstabs

It is apparent that, what concerns us is first and foremost is to obtain a quick general image of the sample which we use for the realization of the research. Thus, in this section we will initially present frequency tables for all the characteristics of patients in the trial who suffer from breast cancer.

Initially, in HERA trial enrolled 5102 patients but after the random assignment in one of the three groups, 3 patients excluded and the results are obtained by the remaining patients (see Figure 1.1). As we can observe from Tables 4.1 and 4.2 our sample consists of 5099 patients from whom 1697 are in the observation group and 3402 are in the treatment- trastuzumab group (we consider 1-year treatment and 2-year treatment groups as one). In Table 4.1 are presented the number of patients according to their characteristics overall and in Table 4.2 by treatment group.



Table 4.1: Frequencies about patients' characteristics

Characteristic	Number of patients (n%)
Region	
W., N. Europe, Canada and Australia	3645 (71%)
Asia Pacific, Japan	609 (12%)
Eastern Eu-ropce	561 (11%)
Central and South America	284 (6%)
Age at study entry (years)	
<35	378 (8%)
35-49	2264 (44%)
50-59	1639 (32%)
≥60	818 (16%)
Race	
Caucasian	4254 (83%)
Oriental	644 (13%)
Other	201 (4%)
Menopausal status	
Post-menopausal	2319 (45%)
Pre-menopausal	717 (14%)
Peri-menopausal	2063 (41%)
Nodal status	
Not assessed	563 (11%)
Negative	1646 (32%)
1-3 positive nodes	1464 (29%)
≥4 positive nodes	1426 (28%)
Estrogen-Receptor status	
Negative	2789 (55%)
Positive	2309(45%)
Unknown	1 (0%)
Progesterone-Receptor status	

Continued on next page



Table 4.1 – continued from previous page

Characteristic	Number of patients (n%)
Negative	3098 (61%)
Positive	1774 (35%)
Unknown	227 (4%)
Adjuvant chemotherapy	
No Anthracyclines	302 (6%)
Anthracyclines, no Taxanes	3469(68%)
Anthracyclines and Taxanes	1328 (26%)
ECOG Performance status	
0	4683 (92%)
1	414 (8%)
Missing	2 (0%)
Grade of tumor	
G1	106 (2%)
G2	1652 (32%)
G3	3089 (61%)
GX	225 (4%)
Missing	27 (1%)
Size of tumor (cm)	
0-2	2233 (44%)
3-5	2417 (47%)
> 5	311 (6%)
Missing	138 (3%)

From the total patients, 3645 (71%) come from Western and Northern Europe, Canada, South Africa, Australia or New Zealand, 609 (12%) come from Asia Pacific or Japan, 561 (11%) from Eastern Europe and 284 (6%) from Central and South America. The majority of the women with breast cancer (44%) are between the age of 35 and 49, 32% of them are among 50-59 years old, 16% are more than 60 years and only 8% of women are smaller than 35 years old. Also 83% of the



patients are Caucasian, 13% are Oriental and only 4% of them are from other races.

From the women in the trial 2319 are in post-menopausal situation, 717 are in the pre-menopausal situation and 2063 of them have irregular period. 32% of the patients have negative nodal status, 57% have positive nodal status and for 563 patients the lymph nodes has not been assessed. Also 55% and 61% of the patients have negative estrogen and progesterone receptor status respectively whereas 45% and 35% of the patients have positive estrogen and progesterone receptor status respectively. From the total of patients, 302 have been in a procedure of adjuvant chemotherapy with no-Anthracyclines, 3469 have been in a procedure of adjuvant chemotherapy with Anthracyclines and 1328 have been in a procedure of adjuvant chemotherapy with Anthracyclines and Taxanes. Almost all the patients (92%) have 0 ECOG performance status and only 8% of the patients have 1 as ECOG performance status.

The majority of the patients (61%) have a tumor of grade G3, 32% have tumor of grade G2, 2% have tumor of grade G1 and 4% of them have tumor of grade GX. For 2233 patients the size of tumor is between 0 and 2 cm, 2417 patients have tumor with size among 2 and 5 cm and only 311 patients have a tumor with size more than 5cm.

Finally, from Table 4.2 we can understand that cohorts for comparison in analysis are well balanced in terms of demographics and baseline disease characteristics. This is important because in this way, results are more dependable, as the comparison not be affected by the characteristics of patients.



Table 4.2: Frequencies about patients' characteristics by group

Characteristic	Observation group (n%)	Treatment group (n%)
Region		
W., N. Europe, Canada and Australia	1224 (72%)	2421 (71%)
Asia Pacific, Japan	202 (12%)	407 (12%)
Eastern Eu-rope	177 (10%)	384 (11%)
Central and South America	94 (6%)	190 (6%)
Age at study entry (years)		
<35	126 (8%)	252 (8%)
35-49	752 (44%)	1512 (44%)
50-59	546 (32%)	1093 (32%)
≥60	273 (16%)	545 (16%)
Race		
Caucasian	1415 (83%)	2839 (83%)
Oriental	213 (13%)	431 (13%)
Other	69 (4%)	132 (4%)
Menopausal status		
Post-menopausal	770 (45%)	1549 (45%)
Pre-menopausal	234 (14%)	483 (14%)
Peri-menopausal	693 (41%)	1370 (41%)
Nodal status		
Not assessed	178 (10%)	385 (11%)
Negative	555 (33%)	1091 (32%)
1-3 positive nodes	490 (29%)	974 (29%)
≥4 positive nodes	474 (28%)	952 (28%)
Estrogen-Receptor status		
Negative	928 (55%)	1861 (55%)
Positive	769 (45%)	1540 (45%)
Unknown	0 (0%)	1 (0%)

Continued on next page



Table 4.2 – continued from previous page

Characteristic	Observation group (n%)	Treatment group (n%)
Progesterone-Receptor status		
Negative	1065 (63%)	2033 (60%)
Positive	545 (32%)	1229 (36%)
Unknown	87 (5%)	140 (4%)
Adjuvant chemotherapy		
No Anthracyclines	99 (6%)	203 (6%)
Anthracyclines, no Taxanes	1158 (68%)	2311 (68%)
Anthracyclines and Taxanes	440 (26%)	888 (26%)
ECOG Performance status		
0	1546 (91%)	3137 (92%)
1	149 (9%)	265 (8%)
Missing	2 (0%)	0 (0%)
Grade of tumor		
G1	38 (2%)	68 (2%)
G2	559 (33%)	1093 (32%)
G3	1015 (60%)	2074 (61%)
GX	77 (5%)	148 (4%)
Missing	8 (0%)	19 (1%)
Size of tumor (cm)		
0-2	764 (44%)	1469 (43%)
3-5	781 (47%)	1636 (48%)
> 5	107 (6%)	204 (6%)
Missing	45 (3%)	93 (3%)



4.2 Log-rank test and multivariate Cox model

We are interested in examining if survival time differs between patients in the observation group and patients in the treatment group (trastuzumab group for 1 year and trastuzumab group for 2 years are considered as one group for this analysis). Also we need a model which will describe the relationship between survival time and exploratory variables, i.e. the characteristics of patients. It should be noted that for all survival analysis, which follows, we have excluded 10 patients with DFS and OS time equal to 0 to avoid numerical problems, so we conduct the analysis with a sample of 5089 patients (observation group n=1694, trastuzumab for one year n=1968, trastuzumab for two years n=1967).

4.2.1 Log-rank test

The Log-rank test (see Mantel , 1966) is the most well-known and widely used test for comparison of survival curves. It is a nonparametric test and appropriate to use when the data are right skewed and censored. It is used to test the null hypothesis that there is no difference between the population survival curves (i.e. the probability of an event occurring at any time point is the same for each population). In other words, Logrank test examines the difference between observed and expected number of events among groups The test statistic is calculated as follows (for two groups):

$$X_{log-rank}^2 = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2}$$

Where O_1 and O_2 are the total numbers of observed events in groups 1 and 2, respectively, and E_1 and E_2 the total numbers of expected events. In our trial the first group, include the patients in the observation cohort and the second group include the patients who received trastuzumab for 1 or 2 years.

Exactly the hypothesis testing is:

$$H_0: S_1(t) = S_2(t)$$

$H_1: S_1(t) \neq S_2(t)$ where $S_1(t)$ and $S_2(t)$ are the survival functions for the first and the second group respectively



Before we move on, it is important to define the endpoints on which we will base our log-rank test. The endpoint is a direct indicator of the disease progression that is used to describe a health effect (or the possibility of this particular health effect) occurring from the expose at a risk factor. In our analysis the clinical outcome is presented by:

- Disease Free Survival (DFS) is determined as the time from date of surgery until first relapse of tumor or death from any cause
- Overall Survival (OS) is defined as the time from surgery until death from any cause

Group	Number of patients	Number of events	Log-rank test (p-value)
Disease Free Survival			
Observation group	1964	570	0.00
Treatment group	3395	943	
Overall Survival			
Observation group	1964	350	0.00
Treatment group	3395	552	

Table 4.3: Log-rank test for Disease Free Survival and Overall Survival

As we can see in Table 4.3 the number of RFS events for patients in the observation group is 570 and 943 for patients in treatment group. Also the number of OS events for patients in the observation group is 350 and 552 for patients in treatment group. The p-value for the log-rank test for both DFS and OS is equal to 0.00 which mean that at a level of significance $\alpha = 5\%$ we reject the null hypothesis, i.e. there is difference between the survival curves of patients in the observation and treatment group. Therefore, the results of the trial are confirmed and actually the medicine administration seems to improve the lifetime of patients.

With the help of Kaplan-Meier plot we will acquire an image on how the survival functions of the patients in the observation and treatment group differ and behave in time.

The Kaplan-Meier estimator (see Kaplan and Meier, 1958), also known as the product limit estimator, is an estimator for estimating the survival function



from lifetime data. The KaplanMeier method can be used to estimate the survival curve from the observed survival times without the assumption of an underlying probability distribution. The method is based on the basic idea that the probability of surviving k or more periods from entering the study is a product of the k observed survival rates for each period, i.e. the cumulative proportion surviving, given by the following:

$$S(t) = p_1 \times p_2 \times p_3 \times \cdots \times p_k$$

Here, p_1 is the proportion surviving the first period, p_2 is the proportion surviving beyond the second period conditional on having survived up to the second period, and so on. The proportion surviving period i having survived up to period i is given by: $p_i = (r_i - d_i)/r_i$, where r_i is the number alive at the beginning of the period and d_i the number of deaths within the period.

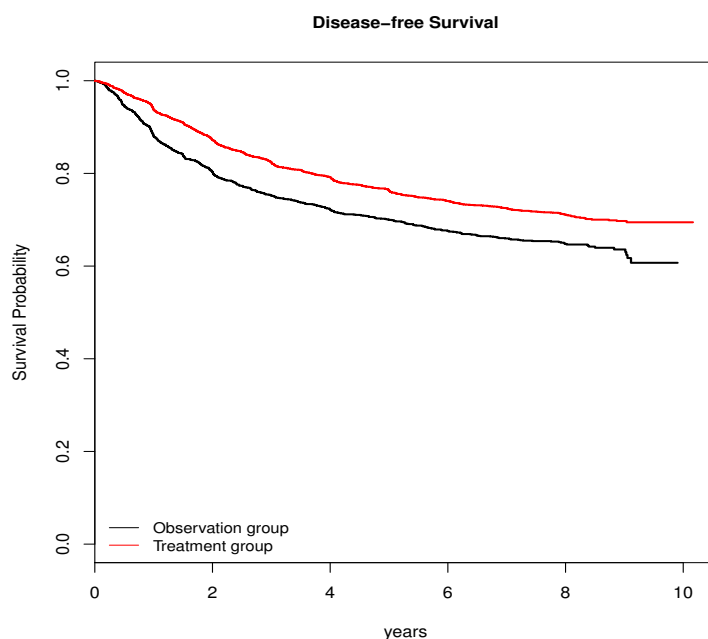


Figure 4.1: Kaplan-Meier plot for Disease-free Survival by patients' group

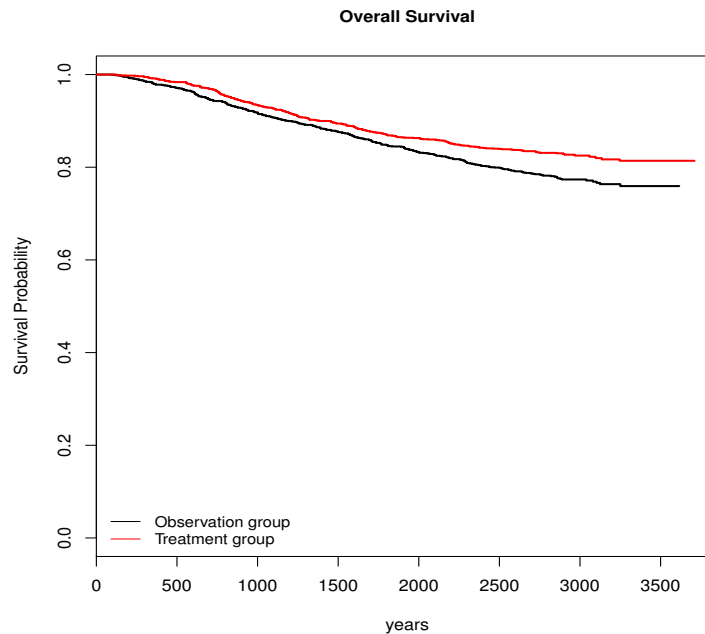


Figure 4.2: Kaplan-Meier plot for Overall Survival by patients' group

Previously, based on the Log-rank test we saw that there is a difference between the survival curves of patients in the observation and treatment groups. Graphically this can be seen, using the Kaplan-Meier plot, where the horizontal axis shows time in days and the vertical axis, the estimated survival probability. (see Figure 4.1 and 4.2 for DFS and OS respectively)

4.2.2 Cox PH model

The log rank test is used to test whether there is a difference between the survival times of different groups but it does not allow other explanatory variables to be taken into account. Cox's proportional hazard model is analogous to a multiple regression model and enables the difference between survival times of particular groups of patients to be tested while allowing for other factors (see Cox, 1972). In this model, the response (dependent) variable is the hazard, i.e. the risk event

occurs. Coxs method does not assume any particular distribution for the survival times, but it rather assumes that the effects of the different variables on survival are constant over time, in other words, the hazard ratio does not depend on time.

The model can be written as:

$$h(t; Z) = h_0(t) \exp(bZ) = h_0(t) \exp(b_1 Z_1 + b_2 Z_2 + \dots + b_p Z_p)$$

where $h(t)$ is the hazard at time t , $Z_1, Z_2 \dots Z_p$ are the explanatory variables, and $h_0(t)$ is the baseline hazard when all the explanatory variables are zero. The coefficients $b_1, b_2 \dots, b_p$ are estimated from the data.

We consider likelihood methods for estimating the model parameters. Specifically, Cox used the idea of a partial likelihood. However, partial likelihood assumes that all of the survival (or censor) times are distinct. In reality, this is probably true in most cases, but in practice, there are often datasets that have many tied survival times. There needs to be a way of accounting for these tied data. There are several proposed modifications to the likelihood to adjust for ties. For example see Efron (1974) and Breslow N.E (1975). In our study we will rely on Breslow method that works well when the ties are relatively few.

Now, we will present the Cox model for the DFS endpoint, which is the primary endpoint in HERA trial. After having included all the characteristics and based on a stepwise method, we will arrive at the final model, including the most important characteristics. In Table 4.4, we can see that the characteristics, which are statistically significant and seem to influence the survival time of patients, are first of all the group they belong to, their race, the menopausal and the nodal status as well as the estrogen-receptor status of them, the type of adjuvant chemotherapy and the size of tumor (all the p-values for these characteristics are smaller than the significant level $\alpha = 0.05$).



Characteristics	Parameter Estimate	Hazard Ratio	Standard Error	p-value
Group (ref=Observation)				
Treatment	-0.3060	0.7364	0.0531	0.00
Race (ref=Caucasian)				
Oriental	-0.2465	0.7815	0.0844	0.0034
Other	-0.1983	0.8201	0.1357	0.14
Menopausal-status (ref=Post-menopausal)				
Pre-menopausal	0.2076	1.2307	0.0758	0.0062
Peri-menopausal	-0.0288	0.9716	0.0572	0.61
Nodal-status (ref=Negative)				
Not Assessed	0.9092	2.4822	0.1004	0.00
1-3 positive nodes	0.4281	1.5344	0.0784	0.00
≥4 positive nodes	1.0302	2.8016	0.0750	0.00
Estrogene-Receptor (ref=Negative)				
Positive	-0.2837	0.7530	0.0530	0.00
Unknown	1.9224	6.8375	1.0036	0.055
Adjuvant-Chemotherapy (ref=No Anthracyclines)				
Anthracyclines	-0.1597	0.8524	0.1204	0.18
ANthracyclines and Taxanes	-0.2676	0.7652	0.1299	0.039
Tumor size in cm (ref=0-2)				
2-5	0.1883	1.2072	0.0568	0.0009
> 5	0.6040	1.8293	0.0944	0.00
Missing	0.2106	1.2344	0.1548	0.17

Table 4.4: Multivariate Cox proportional hazard model for DFS (we needed 5 steps until the final model)

4.3 Estimation of hazard function

4.3.1 No covariates in the model

The problem in HERA trial is that none of the known distributions which are used in parametric survival analysis can estimate the hazard function of neither the patients in the observation group nor the patients who received Transtuzumab for 1 or 2 years. So it is important to find a proper model which can describe the failure time of our data properly. In previous section we use the non-parametric Cox PH model but may not be a good choice about patients who had experienced a very high early risk of relapse/death followed by several years of reduced, but not still negligible.

The idea is to use a mixture of Weibull distributions in order to estimate the hazard function and the parameters of the model, because as we saw before a Weibull mixture model can give hazard functions with flexible shapes. Also, we will examine how some of the most commonly used models in survival analysis can fit the hazard function both for patients in the observation group and patients in the two treatment groups. Totally, we will fit the following parametric models:

- Generalized Gamma
- Weibull
- Log-Normal
- Log-Logistic
- Gamma
- Mixture of 2 Weibull
- Mixture of 3 Weibull

For the common distributions (gamma, generalized gamma, Weibull, log-normal and log-logistic), in order to estimate the hazard function, i.e. the parameters of the model, we use the R package "flexsurv" (see Jackson, 2017) and especially the



function `flexsurvreg`. This function fits parametric models for time-to-event survival data. On the other hand for the estimation of parameters in the finite mixture model we use the EM algorithm as it was described in the previous chapter.

In Tables 4.5, 4.6 and 4.7 are presented all these models and their log-likelihood as well as the number of estimated parameters and the value of Akaike Information Criterion for patients in the observation and the treatment groups respectively. The Akaike information criterion (AIC) is a measure of the relative quality of statistical models for a given set of data and it is used very often as a method for model selection. Given a collection of models for the data, AIC estimates the quality of each model, relative to each of the other models. Hence, AIC provides a means for model selection. The AIC value of a model is given by:

$$AIC = 2k - 2\log(\widehat{L}),$$

where \widehat{L} is the maximized value of the likelihood function of the model and k are the number of free parameters to be estimated. Given a set of candidate models for the data, the preferred model is the one with the minimum AIC value.

We can see that based on the value of AIC the Generalized Gamma model is selected as the best among the basic models, because it has the smallest value in all cases (for the observation and the two treatment groups). This was expected since the shape was closed to the observed one. The second best parametric model seems to be the Log-Normal.

Distribution	Log-likelihood	No. of parameters	AIC
Generalized Gamma	-2072.303	3	4150.607
Weibull	-2124.277	2	4252.555
Log-Normal	-2110.698	2	4225.396
Log-Logistic	-2131.107	2	4266.214
Gamma	-2089.456	2	4182.911
Mixture of 2 Weibull	-2065.341	5	4140.682
Mixture of 3 Weibull	-2063.09	7	4541.180

Table 4.5: Results from fitting some parametric models in observation group



Distribution	Log-likelihood	No. of parameters	AIC
Generalized Gamma	-1881.454	3	3768.909
Weibull	-1912.495	2	3828.990
Log-Normal	-1903.939	2	3811.878
Log-Logistic	-1915.730	2	3835.460
Gamma	-1888.691	2	3881.878
Mixture of 2 Weibull	-1866.05	5	3742.100
Mixture of 3 Weibull	-1865.164	7	3744.328

Table 4.6: Results from fitting some parametric models in 1 year treatment group

Distribution	Log-likelihood	No. of parameters	AIC
Generalized Gamma	-1909.02	3	3824.041
Weibull	-1926.922	2	3857.843
Log-Normal	-1920.159	2	3844.317
Log-Logistic	-1928.845	2	3861.689
Gamma	-1910.201	2	3824.401
Mixture of 2 Weibull	-1901.481	5	3812.962
Mixture of 3 Weibull	-1900.000	7	3814.000

Table 4.7: Results from fitting some parametric models in 2 year treatment group

Nevertheless, while in Figures 4.3, 4.5 and 4.7 the Survival function looks quite satisfactory in case of Generalized Gamma distribution (for some models like Weibull and Gamma, which have also the worst AIC value, the fitted curves lies outside the confidence band as it is obtained from the Kaplan-Meier curves), in Figures 4.4, 4.6 and 4.8, where we plot the hazard rate against time, we can see certain deviations even for the Generalized Gamma model and the Log-Normal model which have the smallest AIC value. Unfortunately, we can see that the peak of the hazard is much earlier in both parametric models for all cohorts.

However, the mixture of 2 Weibull distributions improves the AIC (see Tables 4.5, 4.6 and 4.7). Also we can see that the estimated hazard function from the 2 Weibull mixture model is more close to the non-parametric and the peak of the hazard function in this model is not far from real peak of the data. This is true for



patients in the observation group as well as the patients in two treatment groups (see Figures 4.4, 4.6 and 4.8)

Well, the best model for our data seems to be the 2-finite Weibull mixture model with AIC=4140.682 in the observation group, AIC=3742.100 in the 1 year treatment group and AIC=3813.268 in the 2 year treatment group. In Table 4.8 the estimated values for all the parameters of the model are presented as well as their standard errors. Standard errors were derived using bootstrap. The bootstrap method introduced in Efron (1979) is a very general resampling procedure for estimating the distributions of statistics based on independent observations.

Parameter	Estimated value	Standard error
Observation group		
shape 1	1.3882	0.1112
scale 1	1.4674	0.2001
shape 2	1.1035	2.0596
scale 2	34.2052	6.2372
p	0.2111	0.0306
1 year Treatment group		
shape 1	1.7449	0.3865
scale 1	1.9331	0.1484
shape 2	0.9631	0.2053
scale 2	47.1835	9.4137
p	0.1482	0.0292
2 year Treatment group		
shape 1	1.2518	0.0861
scale 1	3.6612	0.5011
shape 2	7.4669	0.0204
scale 2	15.5827	3.5892
p	0.3072	0.0448

Table 4.8: Estimated values and standard errors for the parameters of the 2-finite Weibull mixture model as they are obtained from the EM algorithm (for each of 3 groups)

We observe that the estimated values for the two shapes parameters are close to 1, which mean that shape parameters from 2 Weibull distributions close to 1 can give a mixture model in which the hazard function can have a shape like



”inverse bathtub” as happen in our data. Also in all groups we can see that the the scale parameter of first Weibull distribution have a small value while the scale parameter of second distribution is much bigger.

Another important issue is to check if a mixture of more than two Weibull distributions can describe our data properly. Now, if we apply a mixture o 3 Weibull distributions we can see from Tables 4.5, 4.6 and 4.7 that the model with 3 distributions does not improve the value of AIC. In Figure 4.8 it is obvious that adding a distribution has not to offer something better in the estimation of hazard function. So, it is preferable to use a model with 2 distributions because the model is more simple and also we have to estimate 5 parameters instead of 7.

Ending, it is important to try to give an interpretation in two distributions which constitute the model. As we observe in Figure 4.10 the first Weibull distribution corresponds to patients who have important probability to survive for many years while the second Weibull distribution represents the patients who don’t survive more than almost 4 years.

The improvement offered by the mixture model can have interesting consequences. The superiority of the 2-finite model reveals the inhomogeneity of the patients. It would be very interesting to examine further the characteristics of the two groups of women, i.e. early responders versus the late responders. Also shows that the simple Weibull used for simulation in Regan et al. (2012) can be improved further by assuming a 2-finite mixture of Weibull distributions for better approximating the behavior of patients.

4.3.2 Adding covariates to the model

Until now, we hadn’t used covariates in our models. Merely, we tried to estimate the survival time of patients without thinking of factors that may influence their survival time. So, it is a good idea to extent the idea of the mixture of Weibull distributions when we allow covariates at the model. The extension is rather simple because only few things change.

The basic different is that now the scale parameters b_1 and b_2 of two Weibull distributions are not just values but vectors with different values for each observation. Generally let \mathbf{X}_{ij} be the j^{th} covariate associated with patient i, for



$j = 1, 2, \dots, d$ and $i = 1, 2, \dots, n$. The covariates now can be included in the mixture model with m components as follows:

$$\log(b_m) = X_i^T \beta_m,$$

where $X_i = X_{i1}, X_{i2}, \dots, X_{id}$,

$b_m = b_{1m}, b_{2m}, \dots, b_{dm}$

and $\beta_m = \beta_{1m}, \beta_{2m}, \dots, \beta_{dm}$.

Therefore it is needed to estimate d more parameters about the coefficient of the covariates in each of the 2 Weibull distributions. Of course covariates can be either qualitative or quantitative. This assumption about scale parameters corresponds to the classical accelerated failure time models, which are used very much in survival analysis, where the effect of covariates is to accelerate or decelerate a baseline survival time T_0 by a factor $\exp(-\beta_i^T \mathbf{X}_j)$. Thus negative values of $\beta_i^T \mathbf{X}_j$ lead to a better survival prognosis while positive values result in an increased risk of failure. We assume that the probability p an observation belongs to first Weibull distribution does not be affected by the exploratory variables. Also the shape parameters of the model do not be influenced.

About the EM algorithm the basic difference between the model without covariates and the model which allow covariates is that now the parameter θ in $Q(\Theta, \Theta^{(m)})$ is not a vector with length equal to the number of parameters but a matrix with n rows like the sample size and columns equal to the number of parameters. As we said before only the scale parameter will be different for each observation. Thus the we have that $Q(\Theta, \Theta^{(m)})$ at E-step is given by the following type:

$$\begin{aligned} Q(\Theta, \Theta^{(m)}) &= E_{\theta^{(m)}}[\log L_c(\theta; T)] \\ &= \sum_{i=1}^n \sum_{j=1}^K E_{\Theta^{(m)}}(Z_{ij}|t) \delta_i \ln[p_j^{(m)} f_j(t_i|\theta_{ij}^{(m)})] \\ &+ \sum_{i=1}^n \sum_{j=1}^K E_{\Theta^{(m)}}(Z_{ij}|t) (1 - \delta_i) \ln[p_j^{(m)} S_j(t_i|\theta_{ij}^{(m)})] \end{aligned}$$



And we end up that

$$Q(\Theta, \Theta^{(m)}) = \sum_{i=1}^n \sum_{j=1}^K [\log(p_j) E_{\Theta^{(m)}}(Z_{ij}|t) + \delta_i \ln[f_j(t_i|\theta_{ij}^{(m)})] E_{\Theta^{(m)}}(Z_{ij}|t) + (1 - \delta_i) \ln[f_j(t_i|\theta_{ij}^{(m)})] E_{\Theta^{(m)}}(Z_{ij}|t)].$$

As before, at M-step we maximize $Q(\Theta, \Theta^{(m)})$ with respect to the parameters to obtain new parameter estimations Θ^{m+1} . We can again maximize the term containing p_j and the term containing θ_j independently since they are not related. Estimation of p_j is given from

$$p_j^{(m+1)} = \frac{1}{n} \sum_{i=1}^n z_{ij}^{(m)}$$

$$\text{where } z_{ij}^{(m)} = \frac{p_j^{(m)} [\delta_j f_j(t_i|\theta_{ij}^{(m)}) + (1 - \delta_i) S_j(t_i|\theta_{ij}^{(m)})]}{\sum_{j=1}^m p_j^{(m)} [\delta_j f_j(t_i|\theta_{ij}^{(m)}) + (1 - \delta_i) S_j(t_i|\theta_{ij}^{(m)})]}.$$

Estimations about the other parameters is given by maximizing the second and third terms of $Q(\Theta, \Theta^{(m)})$ which constitute a weighted log-likelihood function with weights z_{ij} which are given by the above form.

Now we will try to apply such a model in our data for women with breast cancer from HERA trial. As we saw before in Cox PH model (see Table 4.4) the group that each patient belongs to influence the survival time. Thus we will apply for all 5089 patients (we have excluded 10 patients from 5099 with DFS time equal to 0 to avoid numerical problems) a mixture of two Weibull distributions, taking into account the qualitative variable which indicate the patients' group.

From Table 4.9 we can see that a mixture of 2 Weibull distributions is the model which can describe our data more properly even in case that we assume that the survival time of patients depends on their clinical group. The mixture model has the smallest value for Akaike Information Criterion and equals to 116895.56 and the second better model is the Generalized Gamma model with AIC= 11745.67. With the application of Weibull mixture model in our data the value of criterion is improved a lot in contrast with the other popular survival models.



Distribution	Log-likelihood	No. of parameters	AIC
Generalized Gamma	5867.83	5	11745.67
Weibull	-5968.448	4	11944.90
Log-Normal	-5889.96	4	11787.93
Log-Logistic	-5938.36	4	11884.74
Gamma	-5980.47	4	11968.96
Mixture of 2 Weibull (without covariate)	-5865.873	5	11741.75
Mixture of 2 Weibull	-5835.78	9	11689.56

Table 4.9: Results from fitting some parametric models with covariate the patients' group

Parameter	Estimated value	Standard error
scale 1 intercept	0.3392	0.1002
1 year trmnt vs. observation	0.4374	0.1196
2 year trmnt vs. observation	0.6781	0.1213
shape 1	1.4269	0.0482
scale 2 intercept	3.5177	0.1834
1 year treatment vs. observation	0.4833	0.1241
2 year treatment vs. observation	0.4773	0.1243
shape 2	1.0496	0.1050
p	0.1919	0.0134

Table 4.10: Estimated values and standard errors for the parameters of the 2-finite Weibull mixture model as they are obtained from the EM algorithm with covariate the patients' group

Furthermore, in Table 4.10 are presented the estimated parameters as arises from the EM algorithm after the usage of the patients' group as covariate in the mixture model and their standard errors that we obtain from Bootstrap method.



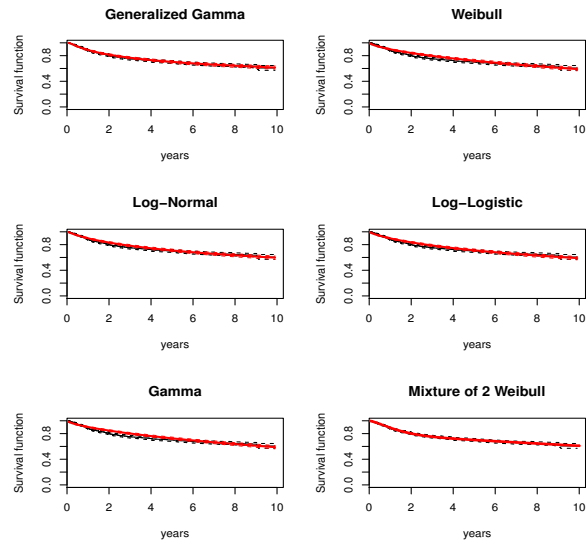


Figure 4.3: Kaplan-Meier survival curves and the fitted survival models for patients in the observation group

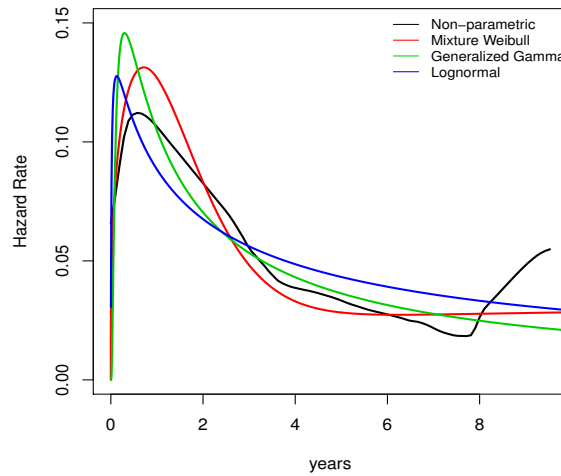


Figure 4.4: Nonparametric hazard functions and parametric estimated functions in the observation group



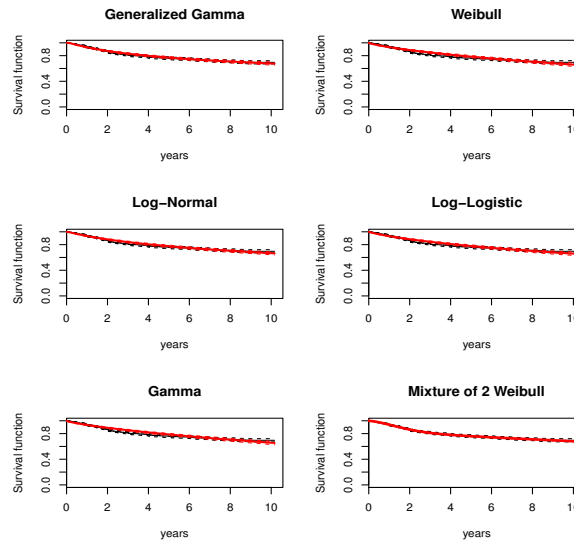


Figure 4.5: Kaplan-Meier survival curves and the fitted survival models for patients in the 1 year treatment group

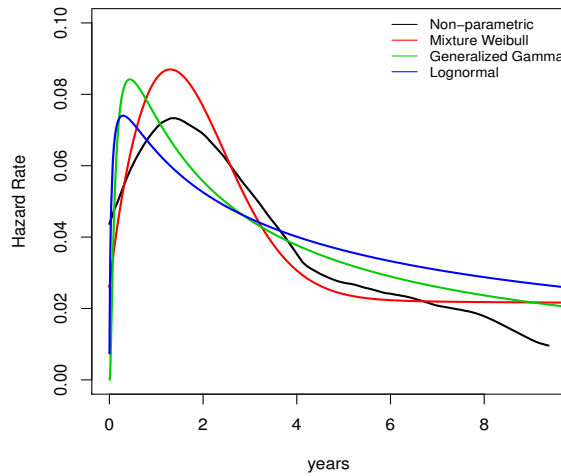


Figure 4.6: Nonparametric hazard functions and parametric estimated functions in the 1 year treatment group



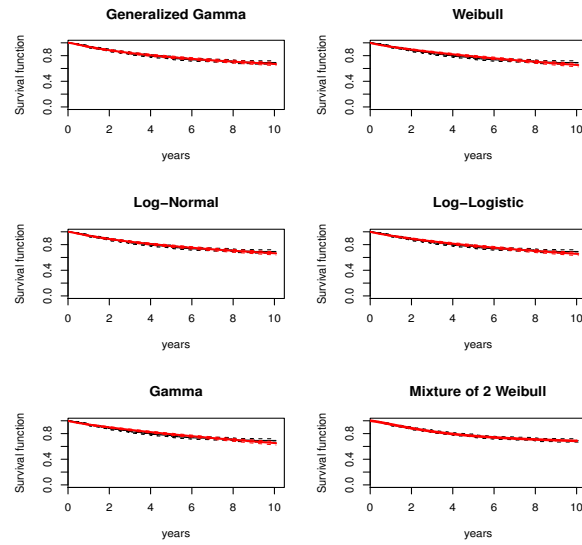


Figure 4.7: Kaplan-Meier survival curves and the fitted survival models for patients in the 2 year treatment group

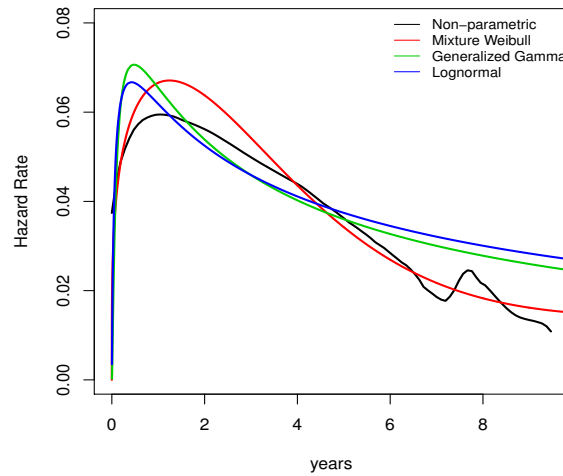


Figure 4.8: Nonparametric hazard functions and parametric estimated functions in the 2 year treatment group



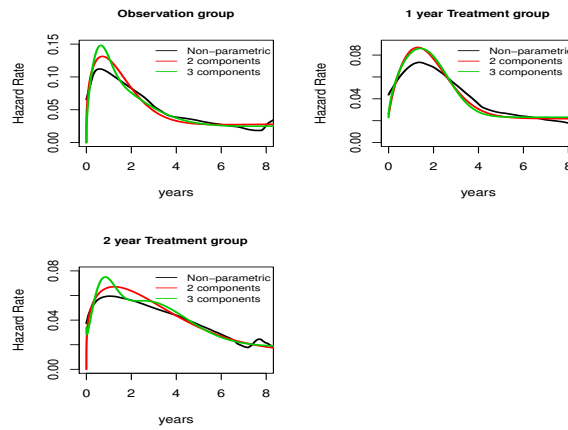


Figure 4.9: Non parametric hazard function and parametric functions estimated from mixture of 2 and 3 Weibull distributions

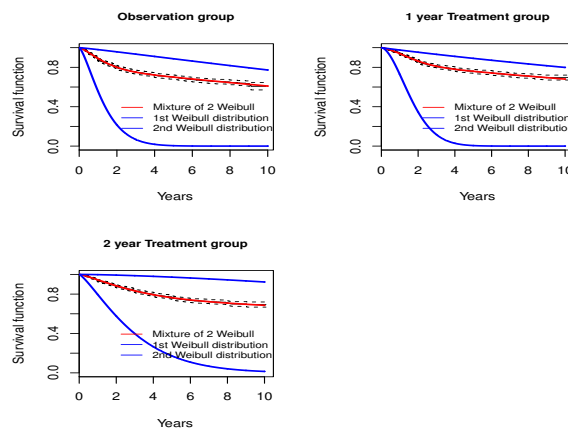


Figure 4.10: Kaplan-Meier survival curves and the fitted survival model as well as the fitted survival functions for each of the 2 Weibull distributions separately

Chapter 5

Conclusions

Survival analysis and general Biostatistics play an important role in our days because field of health and improvement of people's lifestyle are of great interest. The last few years, the number of clinical trials is increasing in a great extent for the above reasons.

In all clinical trials, the basic necessity is the estimation of patients' survival time. In HERA trial -in which this thesis lean on- the estimation of survival and hazard functions is not a simple issue, as survival time of patients does not come from a known distribution. The hazard function of patients in each of three groups that there are in trial is non-monotonic. More specific the hazard function is increasing for some period and then turns decreasing.

Unfortunately, few parametric models can handle this. So we aim at producing flexible models for such cases starting from the widely used Weibull regression model. In this thesis we examined the different shapes that can be obtained by a mixture model of Weibull distributions.

Actually, with the idea of mixture of Weibull distributions we manage to estimate the non-monotonic hazard function successfully and better than using other parametric survival models which are used in survival analysis widely. Also the model has the opportunity to allow covariate information and this is a big asset since permit its using in many cases and especially in data for cancer where a variety of factors affect the survival time of patients.



So it is obvious that a mixture of Weibull distributions outweighs other popular survival models such as Weibull, Log-Normal or Generalized Gamma, because Weibull distribution can have only increasing or decreasing hazard function depending on the value of its shape parameter. Therefore a Weibull mixture model can give more flexible shapes in hazard function and the number of distributions in the model depend on the data and the actually shape of the function.

For sure, there are and other models in bibliography which can be used for the estimation of hazard functions that have a shape like "bathtub" or "inverse bathtub". An extension in our attempt with the Weibull mixture model is to use in model Weibull distributions with one more parameter or mixture model with more than one distributions. These ideas may give more complicated models but it is very possible to achieve better estimations.



References

- Abadi A., Amanpour F., Bajdik C., Yavari P. (2012)**. Breast cancer survival analysis: Applying the generalized gamma distribution under different conditions of the proportional hazards and accelerated failure time assumptions, *International journal of preventive medicine*, **3**, 644-651.
- Ardoino I., Biganzoli E.M., Bajdi, C., Lisboa P.J., Boracchi P., Ambrog, F. (2012)**. Flexible parametric modelling of the hazard function in breast cancer studies, *Journal of Applied Statistics*, **39**, 1409-1421.
- Breslow N.E. (1975)**. Analysis of Survival Data under the Proportional Hazards Model, *International Statistical Review*, **43**, 45-57.
- Cox D.R. (1972)**. Regression Models and Life-Tables, *Journal of the Royal Statistical Society, Series B*, **34**, 187-220.
- Dafni U. (2011)**. Hazard Analysis I, *Cardiovascular Quality and Outcomes*, **4**, 363-371.
- Demiris N., Lunn D., Sharples LD. (2015)**. Survival extrapolation using the poly-Weibull model, *Statistical Methods in Medical Research*, **24**, 287-301.
- Dempster A.P., Laird N.M., Rdin D.B. (1977)**. Maximum Likelihood from Incomplete Data via the EM Algorithm , *Journal of the Royal Statistical Society. Series B (Methodological)*, **39**, 1-38.
- Efron B. (1974)**. The Efficiency of Cox's Likelihood Function for Censored Data, *Journal of the American Statistical Association*, **72**, 557-565.



- Efron B. (1979).** Bootstrap methods: Another look at the Jackknife, *Annals of Statistics*, **7**, 1-26.
- Farcomeni A., Nardi A. (2010).** A two-component Weibull mixture to model early and late mortality in a Bayesian framework, *Computational Statistics Data Analysis*, **54**, 416-428.
- Farewell V.T. (1982).** The use of mixture models for the analysis of survival data with longterm survivors, *Biometrics*, **38**, 1041-1046.
- Friedman L.M., Curt D.F., DeMets D.L., Reboussin D.M., Granger C.B. (2015).** *Fundamentals of Clinical Trials (5th Edition)*. Springer International Publishing, Switzerland
- Fruhwirth-Schnatter S. (2006).** *Finite mixture and Markov switching models*. Springer, New York
- Goldhirsch A., Gelber R.D., Piccart-Gebhart M.J., de Azambuja E., Procter M., Suter T.M., Jackisch C., Cameron D., Weber H.A., Heinzmann D., Dal Lago L., McFadden E., Dowsett M., Untch M., Gianni L., Bell R., Khne C.H., Vindevoghel A., Andersson M., Brunt A.M., Otero-Reyes D., Song S., Smith I., Leyland-Jones B., Baselga J., Herceptin Adjuvant (HERA) Trial Study Team (2013).** 2 years versus 1 year of adjuvant trastuzumab for HER2-positive breast cancer (HERA): an open-label, randomised controlled trial, *Lancet*, **382**, 1021-1028.
- Greenhouse J.B., Silliman N.P. (1996).** Applications of a mixture survival model with covariates to the analysis of a depression prevention trial, *Statistics in Medicine*, **15**, 2077-2094.
- Hess K.R., Serachitopol D.M., Brown B.W. (1999).** Hazard Function Estimators: A Simulation Study, *Statistics in Medicine*, **18**, 3075-3088.
- Hess K., Gentleman R. (2015).** *Hazard Function Estimation in Survival Analysis using the muhaz package*.
<https://cran.r-project.org/web/packages/muhaz/muhaz.pdf>



- Huang Y, Qiu H, Yan C. (2015).** Semiparametric mixture modeling for skewed longitudinal data: A Bayesian approach, *Annals of Biometrics and Biostatistics*, **2**, 1011.
- Jackson C. (2017).** *Flexible Parametric Survival and Multi-State Models*.
<https://cran.r-project.org/web/packages/flexsurv/flexsurv.pdf>
- Jiang R., Murthy D.N.P. (1998).** Mixture of Weibull distributions Parametric characterization of failure rate function, *Applied Stochastic Models and Data Analysis*, **14**, 47-65.
- Kaplan E.L., Meier P. (1958).** Nonparametric estimation from incomplete observations, *Journal of the American Statistical Association*, **53**, 457-481.
- Karlis D., Santourian A. (2009).** Model-based clustering with non-elliptically contoured distributions, *Statistics and Computing*, **19**, 73-83.
- Mantel N. (1966).** Evaluation of survival data and two new rank order statistics arising in its consideration, *Cancer Chemotherapy Reports*, **50**, 163-170.
- Mc Lachlan G.J, Basford K.E. (1987).** *Mixture models. Inference and applications to clustering. Statistics: Textbooks and Monographs*. New York: Dekker
- Mc Lachlan G.J., Mc Giffin D.C. (1994).** On the role of finite mixture models in survival analysis, *Statistical Methods in Medical Research*, **3**, 211-226.
- Piccart-Gebhart M.J., Procter M., Leyland-Jones B., Goldhirsch A., Untch M., Smith I., Gianni L., Baselga J., Bell R., Jackisch C., Cameron D., Dowsett M., Barrios C.H., Steger G., Huang C.S., Andersson M., Inbar M., Lichinitser M., Lang I., Nitz U., Iwata H., Thomssen C., Lohrisch C., Suter T.M., Rschoff J., Suto T., Grotzer V., Ward C., Straehle C., McFadden E., Dolci M.S., Gelber R.D. (2005).** trastuzumab after Adjuvant Chemotherapy in HER2-Positive Breast Cancer, *New England Journal of Medicine*, **353**, 1659-1672.



Pocock S.J. (2013). *Clinical Trials. A Practical Approach.* John Wiley & Sons, England

Regan M.M., Dafni U., Karlis D., A Goldhirsch, Untch M., Smith M., Gianni L., Jackisch C., de Azambuja E., Heinzmann D., Cameron D., Bell R., Dowsett M., Baselga J., Leyland-Jones B., Piccart-Gebhart M.J., Gelber R.D (2012). Selective Crossover in Randomized Trials of Adjuvant trastuzumab for Breast Cancer: Coping with Success, *Cancer Research*, **72**, Abstract nr P5-18-02.

Romond E.H., Perez E.A., Bryant J., Suman V.J., Geyer C.E., Davidson N.E., Tan-Chiu E., Martino S., Paik S., Kaufman P.A., Swain S.M., Pisansky T.M., Fehrenbacher L., Kutteh L.A., Vogel V.G., Visscher D.W., Yothers G., Jenkins R.B., Brown A.M., Dakhil S.R., Mamounas E.P., Lingle W.L., Klein P.M., Ingle J.N., Wolmark N. (2005). trastuzumab plus Adjuvant Chemotherapy for Operable HER2-Positive Breast Cancer, *New England Journal of Medicine*, **353**, 1673-1684.

Schlattmann P. (2008). *Medical Applications of Finite Mixture Models.* Springer Berlin Heidelberg

Slamon D., Eiermann W., Robert N., Pienkowski T., Martin M., Press M., Mackey J., Glaspy J., Chan A., Pawlicki M., Pinter T., Valero V., Liu M-C., Sauter G., von Minckwitz G., Visco F., Bee V., Buyse M., Bendahmane B., Tabah-Fisch I., Lindsay M-A., Riva A., Crown J. (2011). Adjuvant trastuzumab in HER2-Positive Breast Cancer, *New England Journal of Medicine*, **365**, 1273-1283.

Stone G.C., Culbert I., Boulter E.A., Dhirani H. (2004). *Electrical Insulation for Rotating Machines: Design, Evaluation, Aging, Testing, and Repair, 2nd Edition.* John Wiley & Sons, England

Tai P., Chapman J-A. W., Yu E., Jones D., Yu C., Yuan F., Sang-Joon L. (2007). Disease-specific survival for limited-stage small-cell lung cancer affected by statistical method of assessment, *Bio Med Central Cancer*, **7:31**.



Therneau T. (2015). *A Package for Survival Analysis in S. version 2.38.*

<https://CRAN.R-project.org/package=survival>

Watson G.S., Leadbetter M.R. (1964a). Hazard Analysis I, *Biometrika*,
51, 175-184.

