Bayesian Imputation of Right Censored Data in Time-to-Event Studies

PhD thesis

by

Shirin Moghaddam

Supervisors:

Prof. John Hinde

Prof. John Newell

School of Mathematics, Statistics and Applied Mathematics

National University of Ireland, Galway

February 2018

Contents

1 Introduction **1**

1.1 Datasets 3

1.1.1 Galaxy Data 3

1.1.2 6-MP Data 3

1.1.3 Metastatic Renal Carcinoma Data 4

1.1.4 Bronchopulmonary Dysplasia (BPD) Data 4

1.2 Structure of the Thesis 5

2 Survival Analysis **8**

2.1 Introduction 8

2.2 Censoring 9

2.3 The Survivor and Hazard Function 13

2.4 Likelihood Function 16

2.5 Parametric Estimation of the Survivor Function 18

2.5.1 Exponential Distribution 18

2.5.2 Weibull Distribution 19

2.5.3 Gamma Distribution 20

2.5.4 Lognormal Distribution 21

2.5.5 Regression models 22

2.6 Non-parametric Estimation of the Survivor Function 23

2.7 Semi-parametric Estimation of the Survivor Function 26

2.8 Bayesian Modelling of Time to Event Data 28

2.8.1 Exponential Example 31

2.9 Chapter Summary 34

3 Non-parametric Bayesian Approach 35

3.1 Introduction 35

3.2 Non-parametric Bayes 36

3.3 Mixture Models 37

3.4 Dirichlet Process 39

3.4.1 Definition of Dirichlet Distribution 40

3.4.2 Definition of Dirichlet Process 41

3.5 Constructive Definition of Dirichlet Process 42

3.5.1 Stick-breaking Method 42

3.5.2 Pólya Urn Method 43

3.5.3 Illustrative Web Application 43

3.6 Dirichlet Process Mixture Model 45

3.7 Simulation Based Model Fitting 48

3.8 Posterior Predictive Distribution 51

3.9 Example of Dirichlet Process Mixture Model 52

3.9.1 Illustration 56

3.10 Chapter Summary 58

4 A Bayesian Approach to Imputation of Survival Data 60

4.1 Introduction 60

4.2 Imputation Methods for Missing Data 61

4.2.1 Case Deletion 62

4.2.2 Single Imputation 62

4.2.3 Multiple Imputation 65

4.3 Imputing Censored Observations 67

4.4 Parametric Approach 69

4.5 A Parametric Bayesian Approach 72

4.5.1 Illustration 76

4.6 Non-parametric Bayesian Approach 85

4.6.1 Illustration 92

4.7 Comparing Different Imputation Methods 95

4.7.1 Assuming a Correct Model 96

4.7.2 Mis-specified model 99

4.8 Chapter Summary 102

5 Simulation Study 104

5.1 Introduction 104

5.2 Methods of Generating Censored Observations 105

5.3 Generating Censored Observations 106

5.4 Simulation Scheme 110

5.5 Parametric Bayesian Imputation 112

5.6 Non-parametric Bayesian Imputation 118

5.7 Comparing Imputation Methods 122

5.7.1 Using a Correct Model 124

5.7.2 Using a Mis-specified Model 124

5.8 Chapter Summary 128

6 Applications **129**

6.1 Introduction 129

6.2 Estimating a Density Function 129

6.3 6-MP Data 133

6.4 Metastatic Renal Carcinoma Data 138

6.5 Bronchopulmonary Dysplasia (BPD) Data 142

6.6 Chapter Summary 147

7 Conclusions and Future Work **148**

7.1 Goal of the Thesis and Proposed Methods 148

7.2 Review of Simulation and Application Studies 149

7.3 Future Work 150

Appendix: Rcode **152**

Shiny Application for Sampling From a Dirichlet Process 152

Location Normal Dirichlet Process Mixture Model 157

Royston Parametric Approach 164

Bibliography **165**

List of Figures

2.1 Different censoring forms 10

2.2 Example of Type I censored data. 11

2.3 Example of Type II censored data. In this experiment, the investigator decides to terminate the study after four of the six patients have experienced the event. 12

2.4 Survivor function 14

2.5 Different hazard functions 16

2.6 lognormal density functions with identical parameter μ but differing σ parameters 21

2.7 Kaplan-Meier plot for the 6-MP data 25

2.8 The posterior means of survivor function based on 5000 simulated survivor functions compared to the Kaplan-Meier plot for the 6-MP data 34

3.1 Mixture of two normal distributions with parameters $(\mu_1 = 0, \sigma_1 = 1)$ and $(\mu_2 = 3, \sigma_2 = 0.5)$ with $\omega = 0.6$ 39

3.2 These graphs shows the sample path from DP process using stick-breaking method with $G_0 \equiv N(0, 1)$. The heavy smooth line indicates $N(0, 1)$. 46

3.3 These graphs shows the sample path from DP process using Pólya Urn method where G_0 has mixture of normal distribution with the weight 0.6 ($G_0 \equiv 0.6N(0,1) + 0.4N(3,0.5)$). The heavy smooth line indicates the mixture of normal distribution. 47

3.4 Posterior density for Galaxy data using model described in (3.20) . . 57

3.5 Posterior density for Galaxy data using `Dpdensity` function in `DPpackage` 58

4.1 The Kaplan-Meier survivor plots: for the full data with censoring; the true failure times; omitting censored observations; and using parametric Bayesian imputation for the censored values. 82

4.2 Kaplan-Meier plots of the censored data compared with the mean of the Kaplan-Meier estimates from one hundred imputations using a parametric Bayesian approach. 83

4.3 Boxplots of the true failure times and when omitting censored observations are compared to the 30 draws of parametric Bayesian imputations. 84

4.4 Comparing the density of the true-failure data to the density of the imputed dataset using a parametric Bayesian approach based on one single imputation. 85

4.5 Kaplan-Meier survivor plots: for the full data with censoring; the true failure times; omitting censored observations; and using the non-parametric Bayesian imputation for the censored values. 93

4.6 Kaplan-Meier plots of the censored data compared with the mean of the Kaplan-Meier estimates from one hundred imputations using a non-parametric Bayesian approach. 93

4.7 Boxplots of the true failure times and when omitting censored observations compared to 30 sets based on non-parametric Bayesian imputations. 94

4.8 Comparison of density estimates of the true-failure data to the density of the imputed dataset using non-parametric Bayesian approach based on one single imputation. 95

4.9 The survivor plot for Kaplan Meier estimates of the data, true failure times, Royston imputation, parametric Bayesian imputation, and non-parametric Bayesian imputation. 97

4.10 Boxplots for data generated from $lognormal(1, 0.5)$ imputing censored observations using parametric, parametric Bayes, and non-parametric Bayesian approaches. 97

4.11 Comparing the density estimates of the true-failure times to the imputed datasets using parametric, parametric Bayesian, and non-parametric Bayesian approaches to impute the censored observations. 98

4.12 The survivor plot for Kaplan Meier estimates of the data, true failure times, Royston imputation, parametric Bayesian imputation and non-parametric Bayesian imputation. 100

4.13 Boxplots for data generated from a $Weibull(2, 4)$ with censored observations imputed using parametric, parametric Bayes and non-parametric Bayesian approaches and compare to the true failure times. 101

4.14 Comparing the density plots of the true-failure to the imputed datasets including parametric, parametric Bayesian and non-parametric Bayesian approaches. 102

5.1 Bound range for difference of true survival and KM estimated value . 107

5.2 True-Estimate values based on $Weibull(2, 4)$ for 100 iterations using binomial censoring. 108

5.3 True-Estimate values based on $Weibull(2, 4)$ for 100 iterations using uniform distribution for censored observations. The calculated value for uniform parameter b is 35.45 for 10 percent, 17.73 for 20 percent and 6.99 for 50 percent censoring 109

5.4 True-Estimate values based on $Weibull(2, 4)$ for 100 iterations using exponential distribution for censored observations. The calculated value for exponential parameter λ is 0.031 for 10 percent, 0.065 for 20 percent and 0.216 for 50 percent censoring 111

5.5 True-Estimate values using the parametric Bayesian approach for 100 datasets generated from $Weibull(1, 4)$ using exponential censoring . . 115

5.6 True-Estimate values using parametric Bayesian approach for 100 datasets generated from a $Weibull(2, 4)$ using exponential censoring . 116

5.7 True-Estimate values using parametric Bayesian approach for 100 datasets generated from a $Weibull(0.25, 4)$ using exponential censoring.117

5.8 True-Estimate values using the non-parametric Bayesian approach for 100 datasets generated from $Weibull(1, 4)$ using exponential censoring.120

5.9 True-Estimate values using non-parametric Bayesian approach for 100 datasets generated from a $Weibull(2, 4)$ using exponential censoring. . 121

5.10 True-Estimate values using non-parametric Bayesian approach for 100 datasets generated from $Weibull(0.25, 4)$ using exponential censoring. 123

5.11 True-Estimate values based on a $Weibull(1, 4)$ for 100 iterations using parametric Bayesian and non-parametric Bayesian approach for n=200125

5.12 True-Estimate values based on a $Weibull(1, 4)$ for 100 iterations using parametric Bayesian (assuming lognormal distribution in imputation method) and non-parametric Bayesian approach for n=200 127

6.1 Density plot for dataset simulated from $Weibull(3, 4)$ using `logspline` package and `epdfplot` in `Envstat` package in R compared to true density of $Weibull(3, 4)$. 133

6.2 Density plot for dataset simulated from $Weibull(3, 4)$ using `logspline` package and `epdfplot` in `Envstat` package in R for transformed data using 6.4 compared to true density for $Weibull(3, 4)$. 134

6.3 Kaplan-Meier plot for treatment and control group in 6-MP dataset. . 135

6.4 Kaplan-Meier plot for complete dataset in treatment group using parametric and non-parametric Bayesian approaches compared to the Kaplan-Meier of treatment and control group in 6-MP dataset. 136

6.5 Comparing the boxplots of time to reoccurrence in the control and treatment group using the parametric and non-parametric Bayesian imputation approaches. 137

6.6 Comparing the density plot of control and treatment group using parametric and non-parametric Bayesian approaches to impute censored observations in treatment group. 138

6.7 Kaplan-Meier curves for the survival of patients by treatment group in the RE01 trial. 139

6.8 Kaplan-Meier plot by treatment group in renal carcinoma dataset before and after imputation using parametric Bayesian, non-parametric Bayesian and Royston parametric imputation approaches 140

6.9 Boxplots of survival times for the interferon-alpha and MPA treatment groups after imputing censored observations using the parametric, parametric Bayesian and non-parametric Bayesian approaches. 141

6.10 Comparing the density plot of the interferon-alpha and MPA treatment groups using the parametric Bayesian approach to impute censored observations. 142

6.11 Kaplan-Meier curves for the survival of patients by treatment group in RE01 trial using alpha blending on the lines to show how many patients are at risk at the time. 143

6.12 Kaplan-Meier plot for treatment and control group in the BPD dataset.144

6.13 Kaplan-Meier plot for complete dataset in treatment and control group using parametric and non-parametric Bayesian approaches compared to the Kaplan-Meier estimates for the original data. 145

6.14 Boxplots of the survival time for the control and treatment groups after imputing censored observations using the parametric and non-parametric Bayesian approaches. 145

6.15 Comparing the density plot of cotrol and treatment group using the parametric and non-parametric Bayesian approaches to impute censored observations in both groups. 146

List of Tables

4.1 $E(T \mid T \geq c)$ for a $Weibull(2, 2)$ distribution. 79

4.2 Check of the predicted values provided using WinBUGS based on 200 observations generated from a $Weibull(\alpha = 1, \lambda = 2)$ with ten percent censoring. 81

Abstract

In time-to-event studies subjects are followed until the event of interest has happened. Subjects who do not experience the event are referred to as censored. Due to censoring, methods of plotting individual survival time, such as density plots, are invalid. The graphical displays of time-to-event data usually take the form of a Kaplan-Meier survival plot. However, using a Kaplan-Meier survival plot might not be the most informative way to present the data to answer the typical questions of interest. The median survival is often used as a summary of the survival experience of a patients' population and it is easily read off the Kaplan-Meier plot. It is unlikely however that the median is a relevant summary at the patient level and a density plot of the data is perhaps more informative for communication than a single summary statistic. A fundamental idea in this thesis is to consider censored data as a form of missing, incomplete, data and use approaches from the missing data literature to handle this issue. In particular, we will use the idea of imputing the censored observations, based on the other information in the dataset and some form of assumed model. By imputing values for the censored observations and combining the original complete and imputed incomplete data, it is possible to plot the density of the full data to complement the information given by Kaplan-Meier plots. In this thesis, we consider using parametric Bayesian and non-parametric Bayesian methods to impute right censored survival data to achieve this aim. The imputation of censored observations not only allows more interpretable graphics to be produced for a wider general audience (physicians and patients), but it opens up the possibility of the use of standard formal methods of analysis for continuous responses.

Acknowledgements

First of all, I am deeply indebted to my supervisors Prof. John Hinde and Prof. John Newell for their dedication, instruction and encouragement, without which I would not have been able to complete this work. Thank you for your support, especially in the difficult times.

Many thanks to the Irish Research Council (IRC) for awarding me a postgraduate scholarship which gave me the opportunity to pursue my PhD in Ireland but also learn from Irish culture.

I would also like to thank my friends who made the stay in Galway a much nicer experience.

My parents, Effat and Mohammad, you have always supportive in all my decisions. It was you that encouraged me to start this project, and without your love, I would not have been able to complete the PhD. You have made me the person I am today.

I also want to thank my aunts Hakimeh and Azimeh, who have been supportive in every way possible. You have always been there to listen to me during times when I was finding things tough.

Finally, I owe special thanks to my husband, Amir, for his never ending patience, love and encouragement during these years. Amir, this thesis would not have been possible without your unconditional support.

I would like to dedicate this thesis to my family, for everything.

Chapter 1

Introduction

The main interest in survival analyses is studying the time to an event (usually death, recurrence or remission). In survival analysis subjects are followed until the event of interest has happened. Subjects who do not experience the event are referred to as censored. Unless waiting till everyone in a trial has died before analysing the data, some patients will be still alive at the end of the study and hence their time to death will remain unknown. Due to censoring, methods of plotting individual survival times such as the histogram or the density plot are less useful. Therefore, graphical displays of time-to-event data usually take the form of a plot of the survivor function (typically using the Kaplan-Meier estimator). However, using a Kaplan-Meier survival plot might not be the most informative way to present the data to answer the typical questions of interest.

The median survival is often used as a summary of the survival experience of a patients' population. It is unlikely however that the median is a relevant summary at the patient level. Stephen Gould in his paper "The median isn't the message" [**48**] recounts his personal experience as a statistician suffering from abdominal mesothelioma, a rare and severe cancer when given a diagnosis of having eight months left to live. On reading the literature, he discovered that the estimate corresponded to the median survival time from a right-skewed distribution and questioned the practice of reporting the median in isolation to a patient. The median gives an estimate of the survival time for half the cohort in question and is unlikely to apply to any individual.

In comparing survivor functions between treatment groups, these curves may be similar at the onset but diverge considerably as time progresses. This wide space between the curves may not represent a dramatic benefit of treatment, but may likely indicate variability rising from the small sample sizes available for the analysis at the end of the study. The smaller the number of people available to experience an event, the larger the drop when there is an occurrence of an event. Two estimated survivor functions may look very different with a small log-rank p-value but the actual distribution of time to events may overlap considerably. So, providing graphical summaries of the underlying distribution of the time to the event is clearly an attractive alternative to avoid these concerns and may temper the enthusiasm of physicians and patients to see diverging survival curves.

A key idea in this thesis is to consider censored data as a form of missing, incomplete, data and use approaches from the missing data literature to handle this issue. In particular, we will use the idea of imputing the censored observations, based on the other information in the dataset and some form of model. However, in using imputation for censored data there is an important difference from standard missing data imputation, with censored data we have some, albeit limited, information on the censored value. For example, with right censoring we know that the true failure times exceeds the recorded censored time and we obviously need to make use of this information in our imputations. By imputing values for the censored observations and combining the original complete and imputed data, it is possible to plot the histogram or density of the completed data to complement the information given by Kaplan-Meier plots. In this thesis, we proposed two novel methods to impute right censored survival data including a parametric Bayesian and non-parametric Bayesian approaches. The imputation of censored observation not only allows more interpretable graphics to be produced for the wider audience (physicians and patients), but it opens the possibility of the use of formal methods of analysis for continuous responses.

1.1 Datasets

Four datasets will be used throughout this thesis to demonstrate the methods proposed in this work.

1.1.1 Galaxy Data

The first dataset is the well-known galaxy data from the astronomy literature. It is believed that after the Big Bang, matter expanded at an enormous rate and due to the local attraction of matter, the galaxies formed. Astronomers predicted that gravitational pull would lead to some clustering of galaxies. Historically, astronomers have mapped galaxies by the latitude and longitude with respect to the earth. The third component of position which is the distance from our galaxy to others is estimated using the red shift in the light spectrum in a fashion analogous to the way the Doppler effect measures changes in speed via changes in sound. Given the expansion scenario of the universe, points furthest from our galaxy must be moving at greater velocities. Distance, then, is proportional to and can be estimated from velocity. If the galaxies are clumped, the distribution of velocities would be multimodal where each mode describes a cluster as it moves away at its own speed. In an unfilled survey of the Corona Borealis region, velocities of 82 distant galaxies from 6 well-separated conic sections of space were measured. The data are the recession velocities in units of 10^3 km/s. The astronomers Postman et al. [**90**] gave the full data by region. The Galaxy dataset was first described by Roeder [**91**]. These data can also be found as a part of the R package `MASS` as dataset `galaxies`.

1.1.2 6-MP Data

The second dataset is based on the historical 6-MP trial. This dataset has been widely used in the survival analysis literature as an illustrative example in theoretical and applied work. The data is from a prospective clinical trial where 6-mercaptopurine (6-MP) was compared to a placebo in the maintenance of remission in acute leukaemia. (Gehan [**44**]; Freireich et al. [**43**]). The study was confined to

patients under 20 years of age who had received no chemotherapy before admission. The first patient was enrolled in April 1959, and the last patient was entered in the study in April 1960. In the study, there are 42 leukaemia patients where 21 patients were allocated to a 6-MP treatment group and 21 to control (placebo) group.
The variable of interest is the time spent in remission by each patient. Some remission durations were censored due to the loss of follow up, and many patients were alive with no recurrence of disease at the end of the study. There is no censoring in the control group while there are 12 censored observations in the intervention group.

1.1.3 Metastatic Renal Carcinoma Data

The third dataset is from the Medical research council RE01 trial in metastatic renal carcinoma [**22**]. In the duration between February 1992 and November 30, 1997, 350 eligible patients who had historically or cytologically confirmed renal carcinoma with metastases were recruited to the RE01 trial which was conducted at 31 centers in the United Kingdom.
Patients were randomly assigned to treatment with interferon alpha or with medroxyprogesterone acetate (MPA). Patients were followed up every four weeks until 12 weeks after randomisation. Minimum follow-up was then at six months and one year from randomisation and after that every six months until death.
The variable of interest is time to death and was defined as the time from randomisation to death. As June 21, 2001, of the 347 patients with available data, 25 (7%) were censored, and the remainder had died. 175 of these patients were treated by MPA and 172 of them with interferon alpha. There are 8 censored observations in the MPA group and 17 censored observations among patients who treated with interferon alpha.

1.1.4 Bronchopulmonary Dysplasia (BPD) Data

The fourth dataset used in this thesis is the Bronchopulmonary Dysplasia (BPD) data [**53**]. This study is about low birth weight infants (< 1500 grams) with Bronchopulmonary Dysplasia who are treated with oxygen. The data were collected

between December 1987 to March 1991.
A total of 78 infants were in this study including 35 babies receiving surfactant therapy and 43 of them not receiving this treatment. There are five censored observations in this study as five babies were still on oxygen at their last follow-up visit. Two of these censored observations are in the treatment group and three of them are in the control group.
The purpose of the study was to determine factors that predict the length of time for these infants to be on oxygen and the outcome is the total number of hours an infant needed oxygen therapy. The data are also available at Wiley's FTP site ftp://ftp.wiley.com/public/ `sci_tech_med`/survival/.

1.2 Structure of the Thesis

This thesis consists of five main chapters (along with this introduction and the future work and discussion in Chapter 7). In Chapter 2 a review of survival analysis is presented. Chapter 3 includes an overview of the non-parametric Bayesian approach to inference. In Chapter 4 two new Bayesian approaches for imputing censored observations in survival analysis are introduced. In Chapter 5 the results of a comprehensive simulation study are presented to study the performance of these approaches. Finally, in Chapter 6 the results of applying the Bayesian imputation approaches to the example datasets are reported.
The following paragraphs contain a brief introduction of each of these main chapters.

Chapter 2: Survival Analysis

Survival analysis is a set of statistical methods to analyse data where the outcome of interest is the time until an event occurs. A particular source of difficulty in analysing survival data is censoring which arises when some of the observations are incomplete due to causes that are not under the control of the investigator.
The chapter begins with the definition of censoring. It continues with a brief introduction of the survivor and hazard functions. The definition of the likelihood

in the context of censored survival data is explained. Also, an overview of various commonly used parametric, non-parametric and also semi-parametric techniques in estimating the survivor function is presented. Finally, the Bayesian modelling of time to event data is discussed which introduces the approach that will be used in this thesis.

Chapter 3: Non-parametric Bayesian Approach

Chapter 3 is devoted to the framework of non-parametric Bayesian methods. The intention is not to be complete but rather to touch on areas of interest for the thesis. Non-parametric Bayesian methods enhance the flexibility of standard parametric models while providing a fully probabilistic framework for inference. Under the non-parametric Bayesian paradigm, the unknown functions or distributions of the model are treated as random parameters with stochastic non-parametric priors, such as the Dirichlet process.
The chapter begins with a definition of non-parametric Bayes continued with a brief introduction of mixture models. An overview of the definition and constructive procedure of Dirichlet process are described. Moreover, the Dirichlet process mixture model is presented followed by a discussion on the associated simulation procedures and the definition of posterior predictive distribution. Finally, a location normal Dirichlet process mixture model is presented as an illustration.

Chapter 4: A Bayesian Approach to Imputation of Survival Data

In the presence of right censoring, standard methods of plotting individual survival times are limited. Therefore, graphical display of time-to-event data usually takes the form of a Kaplan-Meier survival plot. Based on the Kaplan-Meier plot, the median survival time is the classical summary reported to the patients and is often mis-interpreted. If there is no censoring in the data set, standard graphical and numerical summaries can be used. By imputing the censored observations and combining the original and imputed data a 'complete' data set can be constructed, and it is then possible to plot the histogram or density of the data to complement the

information given by Kaplan- Meier plots.
The chapter begins with an overview of imputing methods for missing data as an introductory concept for the rest of the chapter where we try to impute the censored observations. A parametric approach (Royston [**92**], [**93**]) to impute censored observations is introduced. Two new approaches have been proposed for imputing censored observations, including a parametric Bayesian approach and a non-parametric Bayesian method. These methods are compared to the Royston parametric approach for imputing censored observations.

Chapter 5: Simulation Study

The main goal of this simulation study is to understand if the proposed imputation methods create plausible complete (imputed) datasets. The benefit of using simulated data is that we know not only the true underlying density function but also the real value of the sample data before the censoring was applied.
This chapter starts with a review of methods for generating censored observations. In the simulation studies, both parametric Bayesian and non-parametric Bayesian approaches are used to impute censored observations. The study considers different percentages of random right censoring and also situations with decreasing, increasing and constant hazard functions. Finally, the parametric Bayesian method and non-parametric Bayesian approach are compared for a specific sample size.

Chapter 6: Applications

This chapter begins with a review of methods for the estimation of an empirical density. The primary focus of this chapter is to apply the parametric Bayesian and non-parametric Bayesian approaches proposed in this thesis to the three datasets to motivate the benefits of considering these methods for imputing censored observations in time to event studies.

Chapter 2

Survival Analysis

2.1 Introduction

In many applications the question of interest is often the time to the occurrence of a given event. In medical studies, this event may be the time to response to treatment, tumour-free time, the length of remission, or time to death [**70**]. Analysis of such data is usually described as survival analysis and originates in studies of time to death, but the approaches are more generally applicable to any time to event, or lifetime, data. Similar approaches for the analysis of lifetime data are used in many different disciplines, including biology, epidemiology, engineering, demography, public health, etc. Although the outcome is usually measured on a continuous scale, the standard methods for analysing continuous responses cannot be used for various reasons. First, time data are intrinsically positive and so the survival distribution is often right-skewed and far from being normal, therefore standard normal methods are not generally appropriate. The second reason is the presence of censoring. An observation is censored if there is some information about the individual survival time, but the exact survival time is unknown (incomplete observation), this may be due to causes that are not under the control of the investigator or to limited periods of follow-up. An important and common form of censoring is when the event of interest happens at some point in the future, but the investigator does not know when, giving right censored data. It is also possible to have left censoring (the

event of interest occurs at some unknown time in the past) or interval censoring (the event occurs at an unknown time in an interval). In fact, all of the various forms of censoring can be considered as special cases of interval censoring, although in this work only right censoring will be considered. To describe censored data the observed outcome is comprised of a continuous measurement (the, possibly, incompletely observed time) and a binary indicator that specifies if an observation has been censored or not [**64**]. To accommodate censoring in the analysis of survival data specific methods have been developed, which differ from those used with general forms of complete response data.

In this chapter, an overview of survival analysis and its main methods is presented. This chapter is organized as follows: It starts with a definition of censoring in Section 2.2. Then in Section 2.3 a brief introduction is given to the survivor function and the hazard function. In Section 2.4 the definition of the likelihood for censored survival data is explained. In Section 2.5 we consider parametric models for estimating the survivor function. In Section 2.6 we move on to non-parametric methods of estimating the survivor function. Extension to semi-parametric survivor estimates is discussed in Section 2.7. Finally, in Section 2.8 Bayesian modelling of time to event data is described, which introduces the approach that we will use in this thesis.

2.2 Censoring

A special source of difficulty in analysing survival data is censoring. In reality, censoring happens when the time recorded does not correspond with the time at which the event of interest occurs. Censoring could happen in one of the following three situations: In right censoring, a patient is enrolled at the beginning of the study however they may be lost to follow-up, or withdrawn before the study ends, or the patient may still be alive at the end of study, so the true survival time is not known (i.e., censored); all that is known is that it exceeds the observed time and so on a time-axis the censored survival time is to the right of the observed time. Left censoring occurs when the event of interest has already happened at the point of observation and so the patient's true survival time is less than the observed time, i.e.

to the left. In interval censoring, the event occurs at an unknown time between two observation times and so the true survival time is known to lie within an interval. These different forms of censoring are shown in Figure 2.1.

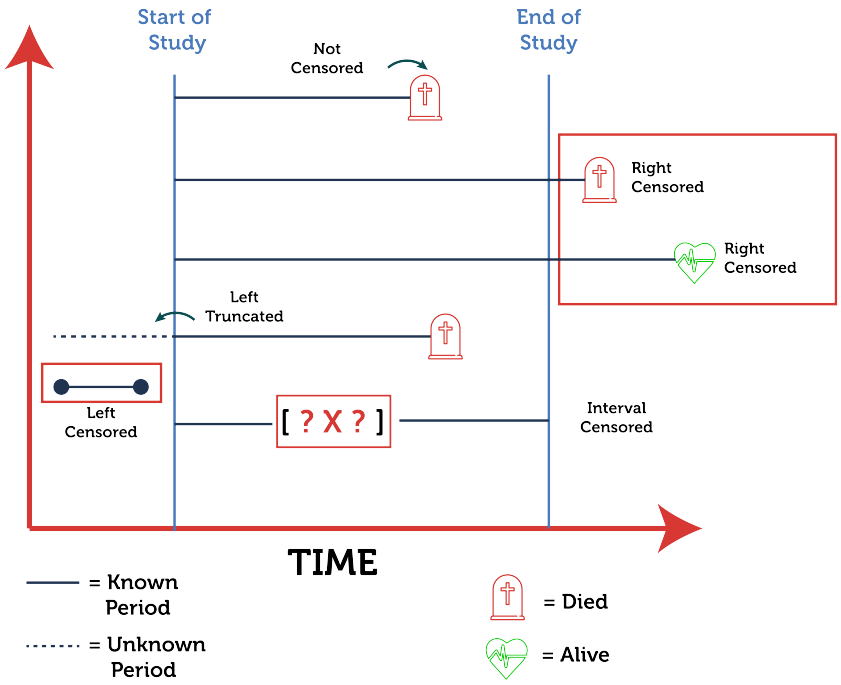

Figure 2.1: Different censoring forms

In this thesis only right censoring will be considered. There are four common reasons for right censoring:

1. A patient does not experience the event before the study ends;
2. A patient is lost to follow-up during the study period;
3. A patient withdraws from the study (drop out);
4. A patient may have experienced another event (competing risk) and no longer be observed (e.g. death by accident).

Right censoring can be classified into three forms.

- **Type I censoring** :

  Often, because of time or cost limitations, the researcher cannot wait for the

event of interest to occur in all patients. The length of an experiment, or observation period, is set up in advance and censoring may occur as a consequence of this. Survival times recorded for patients who had the event during the study period are times from the start of the experiment to the event time. Subjects who did not experience the event of interest before the end of the experiment are considered right censored where the censored times correspond to the length of the experiment. More generally, recruitment is progressive and not all subjects are followed from the start of the study/observation period and the censored times may be different. There may also be other right censored observations where some patients are lost to follow-up or die accidentally (or from causes unrelated to the study) and their censored survival times are from the start of observation until the loss or death (see Random Censoring). In Type I censoring the number of censored observations will be random.

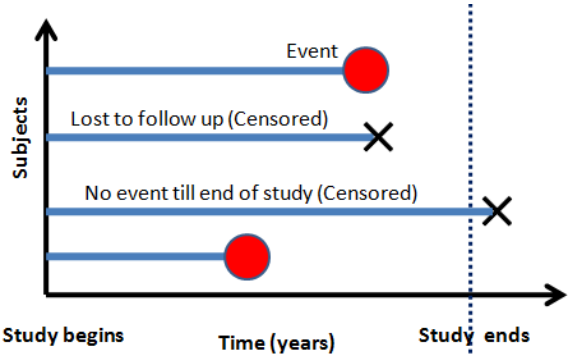

Figure 2.2: Example of Type I censored data.

- **Type II censoring** :

  In type II censoring, the required number of complete observations is set in advance. In other words, when a pre-specified number of events have occurred the experiment stops. The subjects who did not experience the event are treated as right censored. In this case, if there are no accidental losses and all subjects are recruited at the start of the study, then the censored observation times are equal to the largest uncensored observation. Here we have a fixed number of censored observations but the observation period is now random.

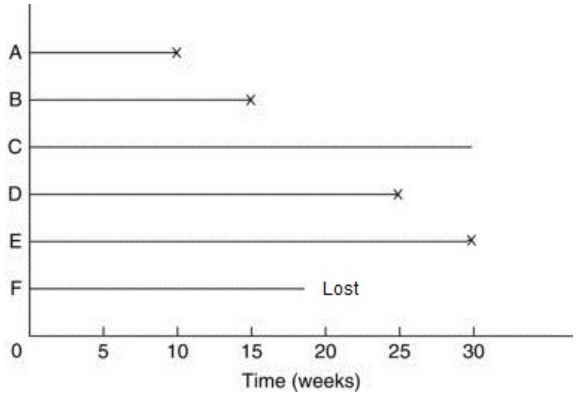

Figure 2.3: Example of Type II censored data. In this experiment, the investigator decides to terminate the study after four of the six patients have experienced the event.

- **Random censoring** :

  A random censoring process is one in which each individual is assumed to have an event time T and a censoring time C where T and C are independent random variables. If the censoring time happens before the failure time the observation is right censored otherwise the true event time is observed. This form of censoring is common in medical studies.

In this thesis, T will represent the random variable for the survival time of an individual, which is the time until an event occurs. So T is called the survival, or event, time. Since T denotes time, its possible values include all non-negative numbers and t denotes any specific value for the random variable T. Similarly we let C be a random variable for the censoring time. In survival analysis we observe either the event time T or the censoring time C, whichever is smaller, i.e. whichever occurs first. So the observed time random variable is $Y = \min\{T, C\}$. Finally, a censoring indicator δ is defined as a 0/1 random variable indicating whether failure or censoring has occurred:

$$\delta = \begin{cases} 1 & T \leq C \\ 0 & T > C \end{cases} \tag{2.1}$$

So $\delta = 1$ if we observe an actual event time and $\delta = 0$ for censored observations. An essential assumption is that survival times T and censored times C are independent. Another assumption is that the distribution of C does not depend on any parameters of the event time T distribution. This is called *uninformative censoring*.

Obviously, in considering survival processes the ideal situation would be to have no censoring and hence complete observed data. However, in practice this is unrealistic and the challenge in analysing survival data is how to handle the incomplete observations. The survival time for a patient who is censored is incomplete but still informative and should be incorporated into the analysis using the information up until censoring. Simple approaches such as dropping subjects with censored survival times from the analysis lead to biased estimates of the survival time as this is just based on analysing observed failure times and long-term survivors are ignored [**21**]. General methods in survival analysis are concerned with how to make use of the limited additional information in the censored observations. However, an alternative would be to obtain a complete dataset, not by deleting the censored observations, but rather by treating them as missing values and using imputation methods. This is the main focus of this thesis, with the aim of complementing traditional formal inferential methods for survival data and providing more interpretable displays for physicians and patients. Different methods of imputing right censored observations are described in Chapter 4.

2.3 The Survivor and Hazard Function

One of the important functions in survival analysis is the survivor function, $S(t)$, which provides a full summary of the survival distribution. The survivor function is specified as the probability that a patient will survive beyond time t.

$$S(t) = P(T > t) = 1 - F(t) = \int_t^\infty f(u)du \tag{2.2}$$

where $F(t)$ is the cumulative distribution function for the random variable T.
Theoretically, as "t" ranges from 0 up to infinity, the survivor function for a continuous random variable can be graphed as a smooth curve, as shown in the Figure 2.4.

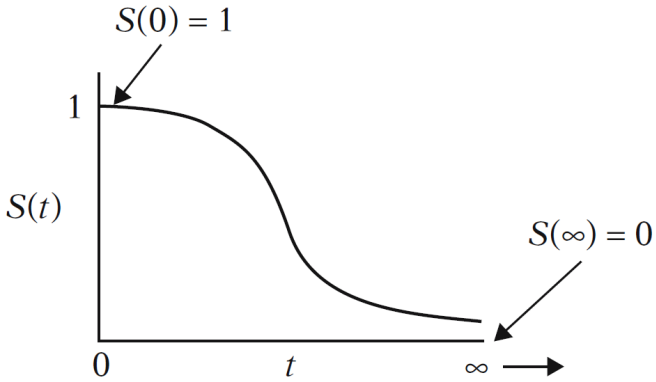

Figure 2.4: Survivor function

All survivor functions have the following characteristics:

- They are non-increasing functions.
- $S(0) = 1$, Since at the start of the study everybody is alive.
- $S(t)$ tends to 0 as t gets large. However, it is possible for $S(\infty) > 0$ as in a cure model [**10**], [**51**] in long term survival data.

In practice, the survivor function can be described parametrically using specific distributions (Section 2.5) or by using some non-parametric estimator such as the Kaplan Meier estimate (Section 2.6).
Compared to the survivor function, which focuses on the probability of not having an event, the hazard function concentrates on the event occurring. The hazard function $h(t)$ gives the instantaneous rate of failure per unit time, given that the individual has survived up to time t, and is defined by:

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t} = \frac{f(t)}{S(t)} \tag{2.3}$$

$$= \frac{-S'(t)}{S(t)} = \frac{-d(log(S(t)))}{dt} \tag{2.4}$$

For a specified value of t, the hazard function $h(t)$ has the following features:

- It is always non-negative (i.e. $h(t) \geq 0$ for all $t > 0$).

- It has no upper bound.

By using (2.4), the survivor function can be written in terms of hazard function as follows:

$$S(t) = exp\left(-\int_0^t h(x)dx\right) = exp(-\Lambda(t)) \tag{2.5}$$

where $\Lambda(t) = \int_0^t h(x)dx$ is the cumulative hazard function.
Further from (2.3) and (2.5) $f(t)$ can be written in terms of the hazard function:

$$f(t) = h(t)exp\left(-\int_0^t h(x)dx\right) = h(t)exp(-\Lambda(t)) \tag{2.6}$$

So defining any one of the probability density function, survivor function, or hazard function enables the other two functions to be determined and they all give equivalent descriptions of the distribution.

The hazard function can have different shapes as shown in Figure 2.5 .

- **Constant over time**; note that for a person who continues to be healthy throughout the study period, his/her instantaneous potential for becoming ill at any time during the period remains constant throughout the follow-up period. Constant hazards correspond to an exponential distribution.

- **Increasing over time**; this might be expected for leukaemia patients not responding to treatment, where the event of interest is death. As survival time increases for such a patient, and as the prognosis accordingly worsens, the patients potential for dying from the disease also increases. Probability models with increasing hazard include the Weibull distribution with a shape parameter greater than 1.

- **Decreasing over time**; this might be expected when the event is death in persons who are recovering from surgery, because the potential for dying after surgery usually decreases as the time after surgery increases. Probability models with deceasing hazard include the Weibull distribution with a shape parameter less than 1.

- **Increasing and then decreasing**; this can be expected for tuberculosis patients, since their potential for dying increases early in the disease and decreases later. The lognormal distribution gives hazards of this form.

- **Decreasing and then increasing**; commonly called a bathtub shape: for example when a child is born there is a chance of early death, but as it escapes childhood diseases, the hazard of failure decreases with age. The hazard is flat and low during the adult life (except for males who have a spike in the early 20s.). Finally, with increased age, the hazard of death is increased. The beta distribution with p.d.f $f(t) = \frac{1}{B(p,q)} t^{p-1}(1-t)^{q-1}$ has the bathtub hazard function when $p < 1$ [**47**].

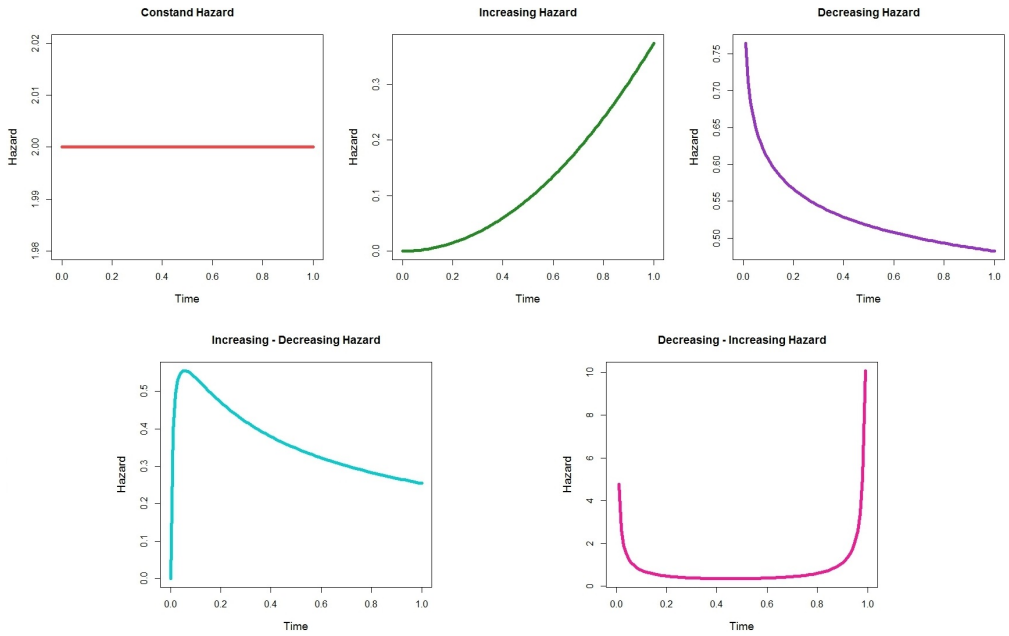

Figure 2.5: Different hazard functions

2.4 Likelihood Function

For a random sample of n individuals we assume independent observations (y_i, δ_i), $i = 1, ..., n$, of the (possibly censored) survival times and associated censoring indicator. The full dataset is defined as $D \equiv \{\mathbf{y}, \boldsymbol{\delta}\}$. For each underling true event time t_i, we assume that the density $f(t_i|\boldsymbol{\theta})$ is known except for a parameter vector $\boldsymbol{\theta}$,

so the survivor function $S(t_i|\boldsymbol{\theta})$ and hazard function $h(t_i|\boldsymbol{\theta})$ are similarly specified. The full likelihood based on the n observations is given by

$$L(\boldsymbol{\theta}|D) = \prod_{i=1}^{n} L(\boldsymbol{\theta}|y_i, \delta_i) \tag{2.7}$$

and to find the explicit form of the likelihood function in a simple way, we assume initially that both the event time and censoring distributions are discrete. As above, we have $f(.|\boldsymbol{\theta})$ and $S(.|\boldsymbol{\theta})$ as the discrete density and survivor function for T, and similarly let $g(.)$ and $G(.)$ be the discrete density and survivor function for C. By using (2.1) the likelihood function for individual i is

$$\begin{aligned}
L(\boldsymbol{\theta}|y_i, \delta_i) &= \begin{cases} P(T_i = y_i, \delta_i = 1|\boldsymbol{\theta}) & \delta_i = 1 \\ P(C_i = y_i, \delta_i = 0|\boldsymbol{\theta}) & \delta_i = 0 \end{cases} \\
&= \begin{cases} P(T_i = y_i, T_i \leq C_i|\boldsymbol{\theta}) & \delta_i = 1 \\ P(C_i = y_i, T_i > C_i|\boldsymbol{\theta}) & \delta_i = 0 \end{cases} \\
&= \begin{cases} P(T_i \leq C_i|T_i = y_i, \boldsymbol{\theta}) P(T_i = y_i|\boldsymbol{\theta}) & \delta_i = 1 \\ P(T_i > C_i|C_i = y_i, \boldsymbol{\theta}) P(C_i = y_i|\boldsymbol{\theta}) & \delta_i = 0 \end{cases}
\end{aligned}$$

Using the independence of C_i and T_i and also the uninformative censoring assumption (the censoring distribution does not depend on parameters of the event time T) and since $\boldsymbol{\theta}$ is the vector of parameters of the distribution of T, we can write

$$\begin{aligned}
L(\boldsymbol{\theta}|y_i, \delta_i) &= \begin{cases} P(y_i \leq C_i) P(T_i = y_i|\boldsymbol{\theta}) & \delta_i = 1 \\ P(T_i > y_i) P(C_i = y_i|\boldsymbol{\theta}) & \delta_i = 0 \end{cases} \\
&= \begin{cases} \left[G_(y_i) + g(y_i)\right] f(y_i|\boldsymbol{\theta}) & \delta_i = 1 \\ S_(y_i|\boldsymbol{\theta}) g(y_i) & \delta_i = 0 \end{cases} \\
&\propto [f(y_i|\boldsymbol{\theta})]^{\delta_i} [S(y_i|\boldsymbol{\theta})]^{(1-\delta_i)}
\end{aligned}$$

This argument can be modified to handle continuous cases [**21**].
Hence, the likelihood function based on the data is

$$L(\boldsymbol{\theta}|D) = \prod_{i=1}^{n} [f(y_i|\boldsymbol{\theta})]^{\delta_i} [S(y_i|\boldsymbol{\theta})]^{(1-\delta_i)} \tag{2.8}$$

which, using the definition of the hazard function can be written as

$$L(\boldsymbol{\theta}|D) = \prod_{i=1}^{n} [h(y_i|\boldsymbol{\theta})]^{\delta_i} [S(y_i|\boldsymbol{\theta})] \tag{2.9}$$

2.5 Parametric Estimation of the Survivor Function

Sometimes there is information about the failure process in a population that suggests a particular distribution, though this information is rarely specific enough to justify one particular family of models [**69**]. In the case that the assumed parametric model is correct, parameter estimates which are obtained from this approach can completely specify the survival and hazard functions. This simplicity and completeness are the primary interests of using a parametric approach. Once a probability density function $f(t)$ is defined for survival time, the corresponding survival and hazard functions can be determined using (2.2) and (2.3). In a parametric approach, the survival time is assumed to follow a known distribution. The commonly used distributions for survival time are exponential, Weibull, gamma,and lognormal. In this section, the characteristics and application of some of these distributions are discussed [**70**].

2.5.1 Exponential Distribution

The simplest distribution which is used in survival analysis is the exponential distribution. This distribution plays a central role in survival analysis as most of the useful survival distributions are related directly to the exponential distribution. The density function of the exponential distribution with parameter λ for survival time

T is defined as:

$$f(t) = \begin{cases} \lambda e^{-\lambda t} & t \geq 0, \lambda > 0 \\ 0 & t < 0 \end{cases} \tag{2.10}$$

The survivor function and the hazard function are:

$$S(t) = e^{-\lambda t} \qquad t \geq 0 \tag{2.11}$$

$$h(t) = \lambda \qquad t \geq 0 \tag{2.12}$$

The hazard rate for exponential distribution is constant so, the probability of death at a time t does not depend on the length of the previous lifetime. This distribution is famous for its "lack of memory" which means the instantaneous probability of failure is the same and not related to how long the item has already survived therefore this distribution represents the lifetime of an item which does not age or wear. A large λ indicates high risk and short-term survival while a small λ indicates low risk and long survival.

2.5.2 Weibull Distribution

The Weibull distribution is a generalisation of the exponential distribution with broader application. This distribution was proposed by Weibull in 1939 [**109**]. Later in 1951 [**110**], Weibull discussed the distribution's applicability to various failure situations. The Weibull distribution is characterised by two parameters, α and λ which are called shape and scale parameter, respectively. The density function, survivor function and hazard function are:

$$f(t) = \alpha\lambda t^{\alpha-1} e^{-\lambda t^{\alpha}} \qquad t \geq 0 \qquad \alpha, \lambda > 0$$

$$S(t) = e^{-\lambda t^{\alpha}} \qquad t \geq 0 \qquad \alpha, \lambda > 0$$

$$h(t) = \lambda\alpha t^{\alpha-1} \qquad t \geq 0 \qquad \alpha, \lambda > 0$$

When the $\alpha = 1$, it reduces to an exponential distribution, as the hazard rate is constant. When $\alpha > 1$, as time t increases, the hazard rate increases, as for patients with lung cancer, while for $\alpha < 1$, the hazard decreases over time, as may apply for patients who undergo successful major surgery.

2.5.3 Gamma Distribution

Suppose that a failure takes place in k stages or as soon as k sub-failures have happened. At the end of the first stage which is after time T_1, the first sub-failure occurs, and so on. Total failures happen at the end of nth stage when kth sub-failure happens. The survival time T is then $T_1 + T_2 + ... + T_k$. If the times $T_1, T_2, ..., T_k$ spent in each stage are assumed to be independently exponentially distributed, then the distribution of T is called the Erlang distribution with parameters λ and k. If the parameter k, restricted here to integer values, is replaced by α which takes any real positive value then the gamma distribution is obtained.
The gamma distribution is characterized by the shape parameter α and a scale parameter λ. The probability density function and survivor function are:

$$f(t) = \frac{\lambda^\alpha}{\Gamma(\alpha)} t^{\alpha-1} e^{-\lambda t} \qquad t > 0 \qquad \alpha > 0, \lambda > 0$$
$$S(t) = \int_t^\infty \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x} dx \qquad \alpha, \lambda > 0$$

When α is integer-valued and equal to k, the hazard function can be calculated as:

$$h(t) = \frac{\lambda(\lambda t)^{k-1}}{(k-1)! \sum_{j=0}^{k-1} \frac{(\lambda t)^j}{j!}} \tag{2.13}$$

while for general α the hazard involves incomplete gamma functions, but is readily calculated in R using the `pgamma` and `dgamma` functions. As time increases from 0 to infinity, when $0 < \alpha < 1$, the gamma hazard rate decreases monotonically from infinity to λ. When $\alpha > 1$, the hazard rate increases monotonically from 0 to λ and when $\alpha = 1$, the hazard rate is a constant, equal to λ, again corresponding to the exponential distribution.

2.5.4 Lognormal Distribution

The lognormal distribution is defined as the distribution of a random variable whose logarithm follows the normal distribution. The distribution of survival times for some diseases, such as Hodgkin's disease and chronic leukaemia, may be rather closely approximated using a lognormal distribution as they are markedly right skewed and the logarithm of survival times are approximately normally distributed. For the lognormal distribution $\Lambda(\mu, \sigma^2)$, the probability density function, survivor function and hazard function are, respectively,

$$f(t) = \frac{1}{t\sigma\sqrt{2\pi}} exp\left[-\frac{1}{2\sigma^2}(logt - \mu)^2\right] \qquad t > 0, \sigma > 0$$

$$S(t) = \frac{1}{\sigma\sqrt{2\pi}} \int_t^\infty \frac{1}{x} exp\left[-\frac{1}{2\sigma^2}(logx - \mu)^2\right] dx$$

$$h(t) = \frac{\frac{1}{t\sigma\sqrt{2\pi}} exp\left[-\frac{1}{2\sigma^2}(logt - \mu)^2\right]}{\frac{1}{\sigma\sqrt{2\pi}} \int_t^\infty \frac{1}{x} exp\left[-\frac{1}{2\sigma^2}(logx - \mu)^2\right] dx}$$

This distribution is positively skewed and the greater the value of σ^2 the greater the skewness (Figure 2.6).

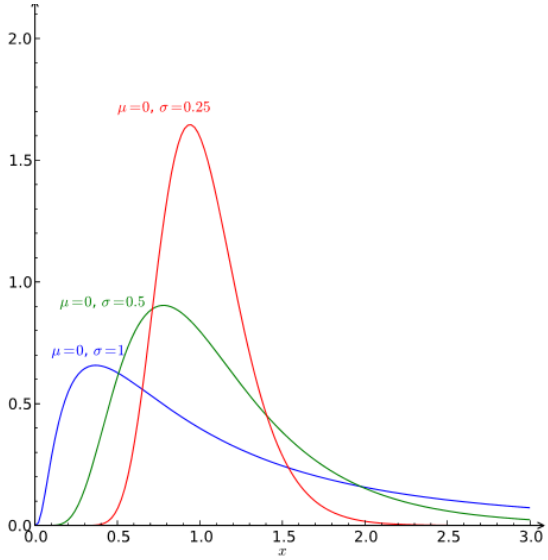

Figure 2.6: lognormal density functions with identical parameter μ but differing σ parameters

The hazard function increases to a maximum and then decreases (almost as soon

as the median is passed) to 0, as time goes to infinity. So, it is suitable for survival studies with increasing and then decreasing hazard rate such as tuberculosis disease.

2.5.5 Regression models

The data in survival analysis often includes covariates, such as age and general condition of the patient, that can be related to lifetime. In many studies, the goal is to understand the relationship between lifetime and these covariates. Regression models for survival data can be defined using parametric models. In parametric survival models, normally either the scale parameter or the shape parameter of a distribution is taken to depend upon a set of covariates, typically by relating it through some appropriate link function to a linear predictor of variables and associated regression parameters while the other parameter of the distribution is held fixed [**69**].
More specifically, if we assume that each individual has a column vector $\mathbf{x} = (x_1, ..., x_p)'$ of covariates then the linear predictor $\eta = \boldsymbol{\beta}'\mathbf{x}$ is related to the parameter of interest, θ, through $h(\theta) = \eta$. For example, if we assume that the distribution of T is exponential, the survivor and hazard functions, for a specific set of covariates $\mathbf{x}$, can be written as

$$S(t|\mathbf{x}) = \exp(-\lambda(\mathbf{x})t)$$
$$h(t|\mathbf{x}) = \lambda(\mathbf{x})$$

where using a log link function for the hazard $\lambda(\mathbf{x})$, gives $\lambda(\mathbf{x}) = \exp(\boldsymbol{\beta}'\mathbf{x})$ where $\boldsymbol{\beta}$ is a $p \times 1$ column vector of regression coefficients, typically including an intercept. Using this log link for the hazard function has the attractive property that $\lambda(\mathbf{x}) \geq 0$ for all real vectors $\mathbf{x}$ and $\boldsymbol{\beta}$, as is required for the exponential distribution, although the canonical link function is the reciprocal. This use of an exponential function of the linear predictor is analagous to the very popular Cox regression model, which will be described in Section 2.7. Although in the exponential model the baseline hazard is constant, while in the Cox model it is assumed to take some unspecified form.
The distributions described in this section are reasonable and commonly used mod-

els for survival times. However, they may not be appropriate for many practical situations which leads us to consider other methods of estimating the survivor function.

2.6 Non-parametric Estimation of the Survivor Function

In using the parametric models described above specific functional forms are assumed for the survivor and hazard functions. However, in some practical situations these may not be appropriate or sufficiently flexible and then non-parametric approaches are often used based on the empirical survivor function, life table, Nelson-Aalen, and Kaplan-Meier estimates. In the case of censored observations, the life-table, Kaplan-Meier[**62**] and Nelson-Aalen[**1**],[**2**],[**85**],[**86**] estimates are the three most common methods that allow estimation of the survivor function with censoring. These techniques do not make any specific assumptions about the shape of the underlying survivor function.

If there are not any censored observations in the dataset, then a simple survivor function estimate would be the empirical survivor function:

$$S_n(t) = \frac{\text{number of individuals with } T > t}{\text{total sample size}} = \frac{\sum_{i=1}^{n} I\{T_i > t\}}{n} \tag{2.14}$$

where I is the indicator function, and defined as:

$$I\{A\} = \begin{cases} 1 & A \text{ is true} \\ 0 & A \text{ is false} \end{cases} \tag{2.15}$$

Note that $F_n(t) = 1 - S_n(t)$ is the empirical CDF. It is easily shown that $S_n(t)$ is a consistent estimate of the true underlying survivor function $S(t)$. We note that $I\{T_i > t\} \sim Bernoulli(S(t))$ and by using the strong law of large numbers which states that the sample average converges almost surely to the expected value, the sample average of $I\{T_i > t\}$, which is $S_n(t)$, converges almost surely to the expected value $S(t)$ (using the expected value in Bernoulli distribution). As $S_n(t)$ converges

almost surely to $S(t)$, it also converges in probability to $S(t)$ and is consistent.
When there are censored observations in the dataset, some modification of (2.14) is required, as the number of lifetimes greater than or equal to t will not be known exactly. The simple modification described here is called the Kaplan-Meier estimate[**62**].
The Kaplan-Meier (KM) estimator, among all the non-parametric approaches, is the most commonly used for estimating the survivor function and has been in widespread use since its development by Kaplan and Meier in 1958. To compute KM curves, the failure times must first be ordered from smallest to largest. We then identify the distinct ordered failure times $t_{(1)} < t_{(2)} < ...$, and associated inter-event time intervals $I_j = [t_{(j)}, t_{(j+1)})$. For each interval $I_{(j)}$ we count e_j, the number of individuals with events at t_j, and r_j, the number of individuals at risk at the start of the interval, that is the number of individuals who have survived until at least time t_j.
The Kaplan-Meier estimator is written as follows:

$$\begin{aligned}\hat{S}(t) &= P(T > t) \\ &= \prod_{j\,:\,t_{(j)} \le t} P(\text{Survive beyond j-th interval } I_j \mid \text{Survived to the start of } I_j) \\ &= \prod_{j\,:\,t_{(j)} \le t} \left(\frac{r_j - e_j}{r_j}\right) \end{aligned} \tag{2.16}$$

To estimate the survival probability at a given time, the KM method uses the risk set at that time to include the information on a censored person up to the time of censorship. Although the KM estimate of survivor function does not change at censoring times, the effect of censoring times is reflected in the size of the risk set r_j. In the absence of censoring the KM estimator is simply the empirical survivor function.
The Kaplan-Meier estimate is a decreasing step function that starts with a horizontal line at a survival probability of 1 and then steps down as we move from one ordered failure time to another.
For illustration, the Kaplan-Meier plot for the 6-MP dataset, described in Chapter 1, is shown in Figure 2.7. The response here is the time to remission in leukaemia patients, with some patients being treated with the drug 6-Mp and others serving as a control group. Plotting the Kaplan-Meier estimate for two or more groups allow a

graphical examination of possible differences between the groups. Figure 2.7 shows that the survivor function for the treatment group consistently lies above that for the control group; this difference indicates that the treatment appears effective at all points of follow-up.

Survival data relate to a non-negative random variable and are typically not symmetrically distributed but rather highly positively skewed, so the median survival time is the preferred summary measure that is often used [**23**]. The median is easily read off the plot of the Kaplan-Meier estimates by looking at the time where the probability of survival of 0.5 meets the Kaplan-Meier estimate, $\hat{S}^{-1}(0.5)$. In the same way other summary quantiles can be calculated. While the full estimated survivor function gives the complete summary of these values it may not be the simplest graphical and a density plot of the data may be more easily interpreted and more informative than any single summary statistic. However, presence of censoring makes the construction of this plot difficult so in Chapter 4 we consider replacing the censored observations by suitably imputed values allowing us to plot histograms, or smoothed density functions, of the completed dataset to complement the information given by Kaplan-Meier plots.

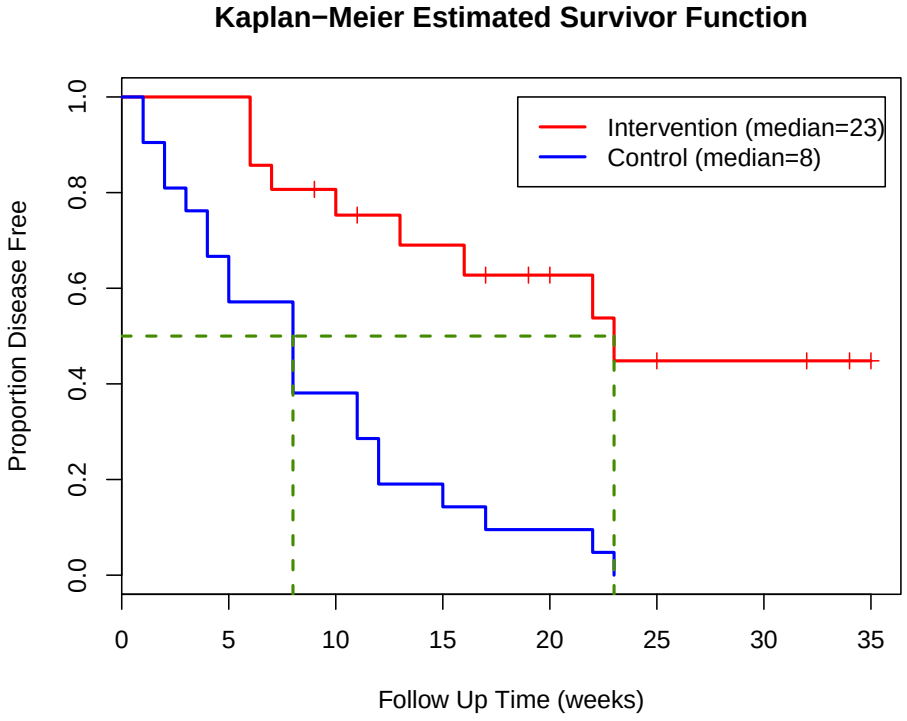

Figure 2.7: Kaplan-Meier plot for the 6-MP data

2.7 Semi-parametric Estimation of the Survivor Function

Survival data often include explanatory variables that might be related to lifetime such as age, sex, type of tumour and type of treatment assigned to the patient. Although in survival analysis, the outcome of interest is the time to an event, by knowing the values of these other explanatory variables, referred to as covariates, a better understanding of the survival experience can be obtained and effects of covariates on survival can be studied.

The Cox proportional hazards model is the most popular regression model for survival data. This model does not require knowledge of the underlying distribution, however a key assumption is that hazard functions of different individuals are assumed to be proportional [**25**] with explanatory variables being used to model this proportionality.

One way of describing the variation in survival among individuals is to consider a particular hazard function $h_i(t)$ for each individual and to assume a proportional hazard assumption which says that

$$h_i(t) = c_i h_0(t) \tag{2.17}$$

where c_i is a time-independent constant for individual i and $h_0(t)$ is an unspecified hazard function. The effect of covariates on the hazard can then be modeled by taking $c_i = exp(\boldsymbol{\beta}'\mathbf{x}_i)$, with the use of the exponential linear form (corresponding to a log link function for the c_i) being designed to ensure that $c_i > 0$, leading to the Cox regression model, which was originally introduced by Cox in 1972 [**26**]. The Cox proportional hazards model is a regression model that enables the information provided by the covariates to explain the variation in survival times across the individuals and identify important variables. The hazard function for the ith individual is then

$$h_i(t|\mathbf{x}_i) = h_0(t)exp(\boldsymbol{\beta}'\mathbf{x}_i) \tag{2.18}$$

Here $\mathbf{x}_i$ is the column vector of the covariates for an individual i and $\boldsymbol{\beta}$ is a column vector of associated regression coefficients. The key feature of this assumption is the

separation of the shape of the hazard function, through the baseline hazard $h_0(t)$, and the multiplicative covariate effects. There is no intercept term included in $\boldsymbol{\beta}'\mathbf{x}$ as it is subsumed into $h_0(t)$.

By using (2.5), the survivor function implied by the model is given by

$$S_i(t|\mathbf{x}_i) = exp(-\Lambda_0(t)exp(\boldsymbol{\beta}'\mathbf{x}_i)) = S_0(t)^{exp(\boldsymbol{\beta}'\mathbf{x}_i)} \tag{2.19}$$

where, $\Lambda_0(t)$ is the cumulative baseline hazard and $S_0(t) = exp(-\Lambda_0(t))$ is the baseline survivor function.

The Cox model is similar to the parametric regression when assuming exponential distribution as described in Section 2.5. The main difference between theses two regression models is that in exponential distribution the hazard is constant for each pattern of covariates which is much stronger assumption than the proportional hazard assumption. If the hazards are constant, consequently the ratio of them is constant as well. But in the Cox model, the hazard ratio is assumed constant which does not necessarily mean that each hazard is constant. In fact, the form of the baseline hazard in the Cox model is not even specified and is essentially considered as a nuisance parameter.

When there are no covariates in the model the Cox model reduces to a common baseline hazard. Thus, $h_0(t)$ may be considered as a starting or baseline version of the hazard function, prior to considering the effect of the covariates $\mathbf{x}$. The fact that in the Cox model the baseline hazard, $h_0(t)$, is an undefined function makes it a semi-parametric model. Even though the baseline hazard is unspecified, it is still possible to estimate the $\boldsymbol{\beta}$ in the exponential part of the model and for a wide variety of datasets, reasonably good estimates of regression coefficients, survival curves and hazard ratios of interest can be obtained. In other words, the Cox proportional hazard model is robust to the precise form of the underlying hazard and avoids the need for a parametric specification of $h_0(t)$, as would happen in a parametric model, and therefore the Cox model results can be considered as closely approximate to the results that would be obtained for a correctly specified parametric model. This is a fundamental reason for the popularity of the Cox model. However, the assumption of proportional hazards is key and may not always hold in practice.

2.8 Bayesian Modelling of Time to Event Data

In the Bayesian approach, the parameters are considered as random variables and inference uses a probability model conditional on the observed data. This contrasts with the frequentist approach where a parameter is considered as a fixed but unknown constant, and inference is based on repeated sampling.

In the Bayesian paradigm, the initial uncertainty about a parameter θ is represented by a probability distribution $p(\theta)$ which is unconditional on any observed data. $p(\theta)$ is called *prior distribution* for θ. By using Bayes theorem [**8**], the prior distribution $p(\theta)$ is updated to a posterior distribution $p(\theta|D)$ which represents uncertainty about θ in light of the observed data.

By specifying a probability model for the observed data D, given the unknown parameter θ, we obtain the likelihood function $L(\theta|D)$. By using the likelihood function (2.8) with prior $p(\theta)$, the posterior of θ is obtained by Bayes' theorem as follows:

$$p(\theta|D) = \frac{L(\theta|D)p(\theta)}{\int_\Theta L(\theta|D)p(\theta)d\theta} \tag{2.20}$$

where Θ denotes the parameter space for θ. The posterior distribution is proportional to the product of likelihood and prior:

$$p(\theta|D) \propto L(\theta|D)p(\theta) \tag{2.21}$$

and $m(D) = \int_\Theta L(\theta|D)p(\theta)d\theta$ is the normalizing constant for $p(\theta|D)$.

If there is little prior information about θ, or an inference based solely on the data is desired, a highly diffuse prior distribution could be chosen, which is referred to as non-informative prior. In choosing a prior belonging to a particular distributional family, some choices may be more computationally convenient than others. In particular, it may be possible to choose a prior distribution which is conjugate to the likelihood, in which case the prior and posterior distribution will be in the same family and have the same distributional form. For example, if the data follow a Binomial distribution it will be convenient, but no means necessary, to use a Beta distribution for the prior. The Beta distribution is conjugate to the Binomial distribution, so the posterior distribution is another Beta distribution [**17**].

In many cases, the posterior distribution does not have closed form because the normalizing constant $m(D)$ does not have simple analytic form. A solution to this is to use simulation methods to obtain sample realizations from the posterior distribution, for example through Gibbs sampling when full conditional distributions are available as is often the case in simpler problems. While in principle providing a general and widely applicable approach, in practice Bayesian inference can be computationally demanding. The increased application of Bayesian methods over the last 30 years can be attributed to the development of Markov Chain Monte Carlo (MCMC) algorithms that allow simulation of draws from the posterior distribution and any associated quantities of interest [**45**].

In survival data, Bayesian analysis has received much recent attention because of these advances in computational techniques. There are some advantages in Bayesian paradigm compared to the frequentist strategy in the area of survival analysis. First, some survival models are generally quite hard to fit, especially in the presence of complex censoring schemes. By using the Gibbs sampler and other MCMC techniques, fitting complex survival models is fairly straightforward, and the availability of software like BUGS greatly eases the implementation. Second, MCMC allows inference from the model for any sample size and does not depend on large sample asymptotics. Also, for many models frequentist inference can be obtained as a special case of Bayesian inference by using non-informative priors, as in this case it can be argued that all of the information resulting in the posterior arose from the data and all resulting inferences are completely objective [**57**].

A major aspect of the Bayesian paradigm is prediction. Prediction is also important in survival modeling. Consider a general parametric model for the data that depends on parameter θ. The posterior predictive density of a future event time t_{new} given the data is defined as:

$$f\left(t_{new}|D\right) = \int f\left(t_{new}|\theta\right) p(\theta|D) d\theta \tag{2.22}$$

where $f\left(t_{new}|\theta\right)$ denotes the sampling density of t_{new}. In (2.22) the dependence of t_{new} on θ is removed by integrating out θ using its posterior distribution $p(\theta|D)$, which summarises the full information on θ given the observed data D.

The predictive survivor function of a future failure time t_{new} given the data is ob-

tained as

$$S\left(t_{new}|D\right) = \int S\left(t_{new}|\theta\right) p(\theta|D) d\theta \tag{2.23}$$

where $S\left(t_{new}|\theta\right)$ is the sampling survivor function of t_{new}.
From (2.22), the predictive mean can be written as the following integral

$$\text{Predictive mean} = \int t_{new} f\left(t_{new}|D\right) dt_{new} \tag{2.24}$$

and also by using (2.23), the predictive median (t_{pmed}) can be obtained by solving the following equation

$$S\left(t_{pmed}|D\right) = 0.5 \tag{2.25}$$

However, in general such direct calculations are not possible, unless the posterior distribution is in a known recognizable form, and hence we rely on simulations. Therefore m posterior samples $\{\theta^k : k = 1, ..., m\}$ could be obtained through Gibbs sampling. So for any function of θ , say $\gamma = g(\theta)$, the posterior $p(\gamma|D)$ is approximated numerically using the sample $\{\gamma^k \equiv g(\theta^k) : k = 1, ..., m\}$. In the case that the posterior distribution does not have closed form, the posterior survivor function $S\left(t|\theta, D\right)$ for all $t > 0$, can be estimated using a MCMC method to simulate $\theta^1, ..., \theta^m$ from the posterior distribution of θ and for each θ^k compute $S(t|\theta^k)$. As we can not estimate the survivor function at all $t > 0$, it is evaluated over a fine grid, possibly $t = 0, 0.01, 0.02, ...T^*$ where T^* is just bigger than the largest observed time in the data. From $\{S(t|\theta^1), ..., S(t|\theta^m)\}$ the posterior median and other percentiles of the survivor function can be calculated. The posterior mean of the survivor function is approximated by:

$$\hat{S}(t|\theta, D) = \frac{1}{m} \sum_{k=1}^{m} S(t|\theta^k) \tag{2.26}$$

Although it might be convenient to simply put the posterior mean of θ into $S\left(t|\theta\right)$, the result is not the posterior mean of $S\left(t|\theta\right)$ because $E(S(t|\theta)|D) \neq S(t|E(\theta|D))$ [**21**].

2.8.1 Exponential Example

As an example [**63**] assume that the survival time has an exponential distribution, which is the most basic parametric model in survival analysis. The exponential arises naturally as the waiting time between events in a Poisson process. The density, survivor and hazard function of the exponential distribution are defined in (2.10), (2.11) and (2.12) respectively. For observed survival data $D = \{(t_i, \delta_i); i = 1, ..., n\}$, subject to right censoring, the likelihood function of λ can be written as:

$$\begin{aligned} L(\boldsymbol{\theta}|D) &= \prod_{i=1}^{n} \left[f(y_i|\boldsymbol{\theta})\right]^{\delta_i} \left[S(y_i|\boldsymbol{\theta})\right]^{(1-\delta_i)} \\ &= \prod_{i=1}^{n} \left[h(y_i|\boldsymbol{\theta})\right]^{\delta_i} \left[S(y_i|\boldsymbol{\theta})\right] \\ &= \lambda^d \exp\left(-\lambda \sum_{i=1}^{n} t_i\right) \end{aligned}$$

Where $d = \sum_{i=1}^{n} \delta_i$. In order to have a conjugate prior for λ, the gamma prior with hyperparameter (α_0, λ_0) is assumed with density:

$$p(\lambda|\alpha_0, \lambda_0) \propto \lambda^{\alpha_0 - 1} \exp\left(-\lambda_0 \lambda\right)$$

So, the posterior distribution of λ is given by:

$$p(\lambda|D) \propto \lambda^{\alpha_0 + d - 1} \exp\left\{-\lambda\left(\lambda_0 + \sum_{i=1}^{n} t_i\right)\right\} \tag{2.27}$$

The kernel of the posterior distribution in (2.27) is a $Gamma(\alpha_0 + d, \lambda_0 + \sum_{i=1}^{n} t_i)$ distribution. So, having an identified form of $p(\lambda|D)$ the normalizing constant is known. The posterior mean and variance of λ are obtained as:

$$\begin{aligned} E(\lambda|D) &= \frac{\alpha_0 + d}{\lambda_0 + \sum_{i=1}^{n} t_i} \\ Var(\lambda|D) &= \frac{\alpha_0 + d}{(\lambda_0 + \sum_{i=1}^{n} t_i)^2} \end{aligned}$$

The posterior predictive density of a future failure time t_{new} is given by

$$\begin{aligned} f(t_{new}|D) &= \int_0^\infty f\left(t_{new}|\lambda\right) p(\lambda|D) d\lambda \\ &\propto \int_0^\infty \lambda^{\alpha_0+d+1-1} \exp\left\{-\lambda\left(t_{new}+\lambda_0+\sum_{i=1}^n t_i\right)\right\} d\lambda \\ &\propto \left(\lambda_0+\sum_{i=1}^n t_i + t_{new}\right)^{-(d+\alpha_0+1)} \end{aligned}$$

Thus the normalized posterior predictive density is given by

$$f(t_{new}|D) = \begin{cases} \dfrac{(d+\alpha_0)(\lambda_0+\sum_{i=1}^n t_i)^{(d+\alpha_0)}}{(\lambda_0+\sum_{i=1}^n t_i + t_{new})^{(d+\alpha_0+1)}} & t_{new} > 0 \\ 0 & \text{otherwise} \end{cases} \tag{2.28}$$

which has the Pareto distribution.

By using (2.27) and the fact that gamma densities integrate to 1, the predictive survivor function for a future survival time t_{new} can be calculated as follows:

$$\begin{aligned} S(t_{new}|D) &= \int_0^\infty S\left(t_{new}|\lambda\right) p(\lambda|D) d\lambda \\ &= \int_0^\infty \exp(-\lambda t_{new}) \frac{(\lambda_0+\sum_{i=1}^n t_i)^{\alpha_0+d}}{\Gamma(\alpha_0+d)} \lambda^{\alpha_0+d-1} \exp\left\{-\lambda\left(\lambda_0+\sum_{i=1}^n t_i\right)\right\} d\lambda \\ &= \frac{(\lambda_0+\sum_{i=1}^n t_i)^{\alpha_0+d}}{\Gamma(\alpha_0+d)} \int_0^\infty \lambda^{\alpha_0+d-1} \exp\left\{-\lambda\left(\lambda_0+t_{new}+\sum_{i=1}^n t_i\right)\right\} d\lambda \\ &= \frac{(\lambda_0+\sum_{i=1}^n t_i)^{\alpha_0+d}}{\Gamma(\alpha_0+d)} \frac{\Gamma(\alpha_0+d)}{(\lambda_0+t_{new}+\sum_{i=1}^n t_i)^{\alpha_0+d}} \\ &= \left(\frac{\lambda_0+\sum_{i=1}^n t_i}{\lambda_0+t_{new}+\sum_{i=1}^n t_i}\right)^{\alpha_0+d} \end{aligned} \tag{2.29}$$

By using (2.24) and (2.28), the predictive mean for the exponential distribution is calculated as follows:

$$\begin{aligned} \text{Predictive mean} &= \int_0^\infty t_{new} f\left(t_{new}|D\right) dt_{new} \\ &= \int_0^\infty t_{new} \frac{(d+\alpha_0)(\lambda_0+\sum_{i=1}^n t_i)^{(d+\alpha_0)}}{(\lambda_0+\sum_{i=1}^n t_i + t_{new})^{(d+\alpha_0+1)}} dt_{new} \\ &= (d+\alpha_0)(\lambda_0+\sum_{i=1}^n t_i)^{(d+\alpha_0)} \int_0^\infty t_{new} (\lambda_0+\sum_{i=1}^n t_i + t_{new})^{-(d+\alpha_0+1)} dt_{new} \end{aligned}$$

Writing $\lambda_0 + \sum_{i=1}^{n} t_i + t_{new} = u$, so the above equation reduces to

$$
\begin{aligned}
&(d+\alpha_0)(\lambda_0 + \sum_{i=1}^{n} t_i)^{(d+\alpha_0)} \int_{\lambda_0+\sum_{i=1}^{n} t_i}^{\infty} (u - \lambda_0 - \sum_{i=1}^{n} t_i)(u)^{-(d+\alpha_0+1)} du \\
&= (d+\alpha_0)(\lambda_0 + \sum_{i=1}^{n} t_i)^{(d+\alpha_0)} \int_{\lambda_0+\sum_{i=1}^{n} t_i}^{\infty} (u)^{-(d+\alpha_0)} - (\lambda_0 + \sum_{i=1}^{n} t_i)(u)^{-(d+\alpha_0+1)} du \\
&= (d+\alpha_0)(\lambda_0 + \sum_{i=1}^{n} t_i)^{(d+\alpha_0)} \left\{ \frac{(u)^{-(d+\alpha_0)+1}}{-(d+\alpha_0)+1} - (\lambda_0 + \sum_{i=1}^{n} t_i) \frac{(u)^{-(d+\alpha_0)}}{-(d+\alpha_0)} \right\}_{\lambda_0+\sum_{i=1}^{n} t_i}^{\infty} \\
&= (d+\alpha_0)(\lambda_0 + \sum_{i=1}^{n} t_i)^{(d+\alpha_0)} \\
&\times \left\{ \frac{(u)^{-(d+\alpha_0)} \left[(\lambda_0 + \sum_{i=1}^{n} t_i)(d+\alpha_0-1) - (d+\alpha_0)u \right]}{(d+\alpha_0-1)(d+\alpha_0)} \right\}_{\lambda_0+\sum_{i=1}^{n} t_i}^{\infty} \\
&= \frac{\lambda_0 + \sum_{i=1}^{n} t_i}{d+\alpha_0-1}
\end{aligned}
\tag{2.30}
$$

Also, the predictive median (t_{pmd}) can be calculated using (2.25) and (2.29) as follows:

$$
\begin{aligned}
& S\left(t_{pmed}|D\right) = 0.5 \\
\Rightarrow \quad & \left(\frac{\lambda_0 + \sum_{i=1}^{n} t_i}{\lambda_0 + t_{pmed} + \sum_{i=1}^{n} t_i} \right)^{\alpha_0+d} = 0.5 \\
\Rightarrow \quad & \left(1 + \frac{t_{pmed}}{\lambda_0 + \sum_{i=1}^{n} t_i} \right)^{\alpha_0+d} = 2 \\
\Rightarrow \quad & t_{pmed} = \left(\sqrt[\alpha_0+d]{2} - 1 \right) \left(\lambda_0 + \sum_{i=1}^{n} t_i \right)
\end{aligned}
\tag{2.31}
$$

For illustration, the Bayesian method is applied to the 6-MP dataset which is described in Chapter 1. The exponential distribution is assumed for the survival time with a gamma prior for the parameter of the exponential distribution. Figure 2.8 shows the posterior mean of Bayesian survivor function, as described in (2.26), for treatment and control groups. It can be seen that the exponential distribution can fit well in this example.

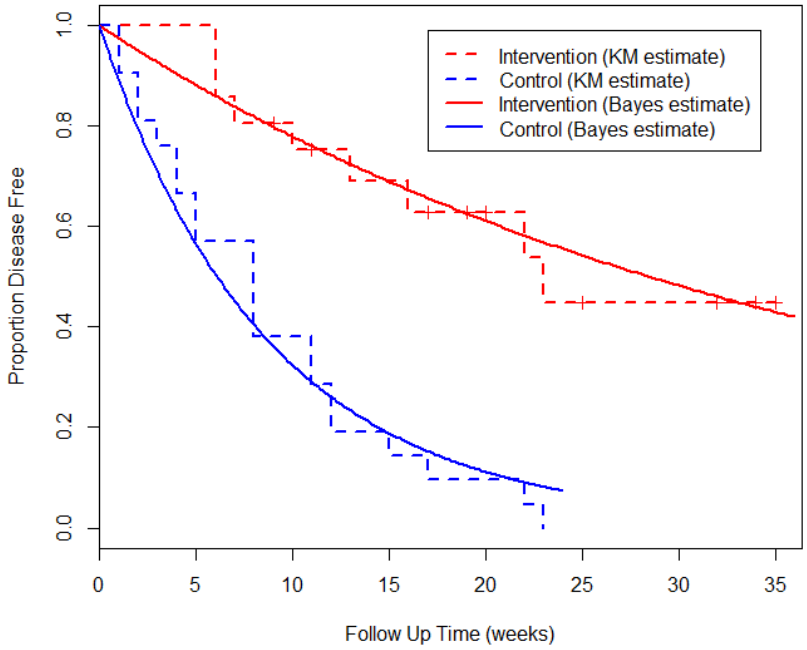

Figure 2.8: The posterior means of survivor function based on 5000 simulated survivor functions compared to the Kaplan-Meier plot for the 6-MP data

2.9 Chapter Summary

This chapter has presented survival analysis and its common feature called censoring. Survivor function and hazard function are introduced as two important functions in the area of survival analysis. Additionally, we have provided an overview of various commonly used parametric, non-parametric and also semi-parametric techniques in estimating the survivor function. In the last part of this chapter, the Bayesian modelling of time to event data is discussed which is the main topic in the context of this thesis to be used as an imputation method to impute the censored observations in the dataset.

In the next chapter, the Bayesian framework is continued by focusing on Bayesian non-parametric models where the Dirichlet process is used as a non-parametric prior.

Chapter 3

Non-parametric Bayesian Approach

3.1 Introduction

A standard way to develop a parametric regression model is to allow distribution parameters to depend on covariates in a pre-specified way. Classical semi-parametric methods specify the regression relationship between the response and covariates parametrically, but leave the actual survival distribution unspecified. The disadvantages of these methods stem from inflexible functional forms of parametric models and limited inference of classical semi-parametric techniques. In particular, a fixed specification of distributional properties for the random or error terms in the model, while typically mathematically convenient, may be inadequate for the actual data. In this chapter non-parametric Bayesian methods (NPB) will be discussed as they will be used in Chapter 4 as a tool to impute censored observations. These methods enhance the flexibility of standard parametric models while providing a fully probabilistic framework for inference. Under the non-parametric Bayesian paradigm, the unknown functions or distributions of the model are treated as random parameters with stochastic non-parametric priors, such as Dirichlet or Gaussian processes.
This chapter is organized as follows: It starts with a definition of non-parametric Bayes in Section 3.2. Then in Section 3.3, a brief introduction is given to mixture

models. In Section 3.4 the definition of Dirichlet distribution and Dirichlet process are explained. In Section 3.5, the constructive definition of the Dirichlet process is described using stick-breaking and Pólya Urn methods. In Section 3.6 we present the Dirichlet process mixture model followed by a discussion of the associated simulation procedures in Section 3.7. In Section 3.8 the posterior predictive distribution is described. Finally, in Sections 3.9 a location normal Dirichlet process mixture model is presented as an illustration.

3.2 Non-parametric Bayes

A common motivation in using non-parametric Bayesian methods is to account for model uncertainty about the choice of a parametric distribution. For example, a normal distribution is often used as the error distribution in regression. But sampling distributions for data or priors often do not follow any standard parametric shape. In contrast to classical non-parametric methods such as the rank test or Kaplan-Meier survivor estimate, non-parametric Bayesian methods can provide full probability models for the data-generating process which can account for uncertainty about distributional shape [**76**].

An early overview of non-parametric Bayesian methods is given in Ferguson in 1973 [**38**]. Under a non-parametric Bayesian perspective, a prior probability model is considered for the unknown density F, in some infinite dimensional function space. This requires the definition of probability measures on a collection of distribution functions. Such probability measures are generically referred to as random probability measures (RPM) [**83**].

Ferguson [**38**] states two important desirable properties for this class of measures; first, the posterior inference should be analytically manageable, although the development of MCMC methods largely overcomes this potential barrier. Second, their support should be large in respect of some suitable topology for the space of probability distributions on the sample space. This means the prior can generate sampling distributions within arbitrary small neighbourhoods of any true data-generating likelihood across a broad class. For example, consider the simple case in which data

consist of a scalar continuous variable y_i, $i = 1, ..., n$ and the density function f is unknown. From a frequentist non-parametric perspective, we could define some density estimator for example through kernel smoothing. From a parametric Bayes perspective, we would choose some parametric form for the density having finitely many parameters θ and we would induce a prior on density f through a prior for θ. Let $\mathcal{F}$ indicate the set of all densities on the real line with respect to Lebesgue measure. Suppose that the true density that generated the data is f_0. Define neighbourhoods around f_0 using some distance $d(f, f_0)$. Parametric priors for f will in general always generate densities on a vanishingly small subset of $\mathcal{F}$. If the true density does not exactly follow the parametric form, then the parametric prior assigns probability zero to small neighbourhoods around "f_0". The idea of large support priors is to define a prior that assigns non-zero probability around "f_0" for any "f_0" in a large subset of $\mathcal{F}$ (perhaps only ruling out weird or irregular densities) and for any neighbourhood size. The non-parametric Bayesian method defines large support priors that are as simple and as interpretable as possible, incorporating any prior knowledge as far as possible, and that lead to tractable (ideally efficient and easy) posterior computation.

Ferguson [**38**] presents a class of such prior distributions, called Dirichlet process (DP) priors, as a random probability measure which is broad in the sense of large support and its posterior inference is analytically manageable.

Before describing the properties of the Dirichlet process, mixture models are reviewed in the next section as these are needed in the rest of this chapter.

3.3 Mixture Models

Mixture models arise naturally as flexible alternatives to standard parametric families when the measurements of a random variable are taken under two, or more, different conditions. For instance, the distribution of heights in a population of adults reflects the mixture of females and males in the population. One of the first analyses using mixture models was by Karl Pearson [**89**]. Finite mixture models make a broad class of interesting statistical models which are less restrictive than

the usual distributional assumptions. The basic idea of these models is that the data arise from two or more underlying groups with the same, or different, distributional form and different parameters [**4**].

Suppose that it is desired to model the distribution of a random sample $\mathbf{y} = (y_1, ..., y_n)$ as a mixture of K components. For $k = 1, ..., K$, the k-th component distribution, $f_k(y_i|\theta_k)$, is assumed to depend on a parameter vector $\boldsymbol{\theta}_k$. So the sampling distribution of $\mathbf{y}$ is:

$$f(y_i|\boldsymbol{\theta}, p) = \sum_{k=1}^{K} p_k f_k(y_i|\boldsymbol{\theta}_k) \quad i = 1, ..., n \tag{3.1}$$

where the p_k denote the proportions of the population from component k. These p_k are non-negative and sum to one, so there are only $K - 1$ identifiable proportion parameters [**28**].

The finite mixture is a special discrete case of the more general compound form, $f(y_i) = \int f(y_i|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}$. This can be considered as a continuous mixture in the sense that each y_i is a random variable with distribution depending on random parameters $\boldsymbol{\theta}$. The prior distribution or population distribution of the parameter $\boldsymbol{\theta}$ is given by the mixing distribution $p(\boldsymbol{\theta})$ [**46**].

A Bayesian finite mixture model with two components and Gaussian densities can be written as follows:

$$\begin{aligned} y_i|\omega, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2 &\overset{ind}{\sim} \omega N(y_i; \mu_1, \sigma_1^2) + (1 - \omega)N(y_i; \mu_2, \sigma_2^2) \\ &= \int N(y_i; \mu, \sigma^2) dG(\mu, \sigma^2) \end{aligned}$$

where $G(.) = \omega\delta_{(\mu_1, \sigma_1^2)}(.) + (1 - \omega)\delta_{(\mu_2, \sigma_2^2)}(.)$ corresponds to a discrete mixing (compound) distribution.

Figure 3.1 shows the mixture of two normal distributions with parameters $(\mu_1 = 0, \sigma_1 = 1)$ and $(\mu_2 = 3, \sigma_2 = 0.5)$ with $\omega = 0.6$.

In finite mixture models, there is uncertainty about the number of mixture components K to include in the model. Models with large numbers of components can be computationally expensive, while models with the small number of components may not reflect all features of the data. Instead of managing the enormous number

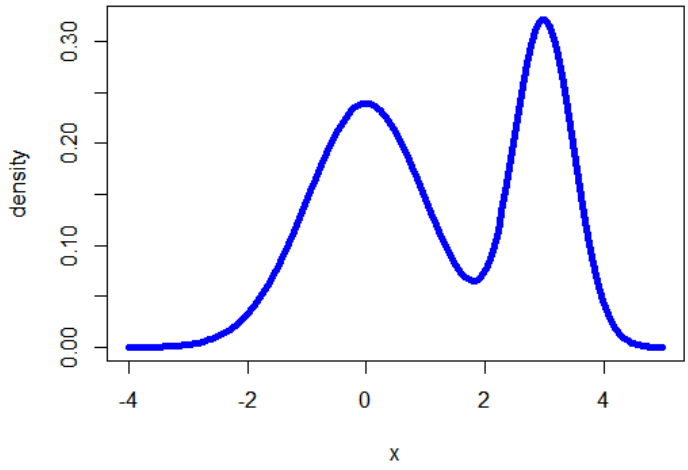

Figure 3.1: Mixture of two normal distributions with parameters $(\mu_1 = 0, \sigma_1 = 1)$ and $(\mu_2 = 3, \sigma_2 = 0.5)$ with $\omega = 0.6$

of parameters of finite mixture models with a large number of components, it could be simpler to work with an infinite dimensional specification by assuming a random mixing distribution which is not limited to a specified parametric family. The Dirichlet process (DP) has been the most widely used prior for the random mixing distribution. Using a DP prior for G, results in a Dirichlet process mixture (DPM) model.

3.4 Dirichlet Process

Instead of using particular parametric forms of prior, non-parametric options have been proposed, such as Dirichlet process priors. The Dirichlet process, developed by Ferguson [**38**],[**39**], is a tool that is used in the non-parametric Bayesian analysis which directly specifies a random probability distribution for the response data $\mathbf{y}$. Specifying a random probability distribution is also what the traditional Bayesian parametric approach does, but it does it in two steps. First, the traditional Bayesian methodology chooses a random distribution for the data by specifying a fixed parametric density $f(y|\theta)$ conditional on the parameter θ and second making θ random by specifying a prior density $p(\theta)$ [**21**].

At first, we review the properties of the Dirichlet distribution needed for the de-

scription of the Dirichlet process.

3.4.1 Definition of Dirichlet Distribution

The Dirichlet distribution arises as a natural family of distributions for point in a simplex and can be used in problems involving order statistics. The Dirichlet distribution is also used as a conjugate prior for the parameters of a multinomial distribution, a set of probabilities $\{\pi_1, \pi_2, \ldots, \pi_k\}$ subject to the constraint $\sum_k \pi_k = 1$, i.e. points in a unit k-dimensional simplex.

Assume $\mathbf{X} = (X_1, ..., X_k)$ be independent random variables, where $X_j \overset{iid}{\sim} Gamma(\alpha_j, 1)$, $j = 1, ..., k$. Define $Y_j = \dfrac{X_j}{\sum_{l=1}^{k} X_l}$ then $(Y_1, ..., Y_k)$ have a Dirichlet distribution with parameters $(\alpha_1, ..., \alpha_k)$ and it is denoted as $(Y_1, ..., Y_k) \sim D(\alpha_0, ..., \alpha_k)$ with the following density function

$$f(y_1, ..., y_k) = \frac{\Gamma(\sum_{i=1}^{k} \alpha_i)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} (1 - \sum_{i=1}^{k-1} y_i)^{\alpha_k - 1} \prod_{i=1}^{k-1} y_i^{\alpha_i - 1} \tag{3.2}$$

where $y_i \geq 0, i = 1, ..., k$ and $\sum_{i=1}^{k} y_i \leq 1$. By defining $\alpha = \sum_{l=1}^{k} \alpha_l$, the first two moments of Dirichlet distribution is as follows

$$\begin{aligned} E(Y_i) &= \frac{\alpha_i}{\alpha} \\ E(Y_i^2) &= \frac{\alpha_i(\alpha_i + 1)}{\alpha(\alpha + 1)} \\ E(Y_i Y_j) &= \frac{\alpha_i \alpha_j}{\alpha(\alpha + 1)} \quad i \neq j \\ V(Y_i) &= \frac{\alpha_i(\alpha - \alpha_i)}{\alpha^2(\alpha + 1)} \end{aligned}$$

If $(Y_1, ..., Y_k) \sim D(\alpha_0, ..., \alpha_k)$ and $r_1, ..., r_k$ are integers such that $0 < r_1 < ... < r_l = k$ then

$$\left(\sum_{1}^{r_1} Y_i, \sum_{r_1+1}^{r_2} Y_i, ..., \sum_{r_{l-1}+1}^{r_l} Y_i \right) \sim D \left(\sum_{1}^{r_1} \alpha_i, \sum_{r_1+1}^{r_2} \alpha_i, ..., \sum_{r_{l-1}+1}^{r_l} \alpha_i \right)$$

Note that Beta distribution is a special case of Dirichlet distribution where $k = 1$ so, $D(\alpha_0, \alpha_1) \equiv Beta(\alpha_0, \alpha_1)$.

3.4.2 Definition of Dirichlet Process

The Dirichlet process (DP) was anticipated in the work of Freedman [**42**] and Fabius [**36**], and developed by Ferguson [**38**],[**39**]. It is the first prior defined for spaces of distribution functions.

A random probability distribution G is generated by a DP if for any finite measurable partition $B_1, ..., B_k$ of the sample space the vector of random probabilities $G(B_k)$ follows a Dirichlet distribution as follows:

$$(G(B_1), ..., G(B_k)) \sim D\left(\alpha G_0(B_1), ..., \alpha G_0(B_k)\right) \tag{3.3}$$

It is denoted as $G \sim DP(\alpha, G_0)$. The DP is characterised by two parameters; G_0, a specific distribution on $\mathcal{X}$ and α, a precision parameter that defines variance. For large α there is small variability in DP realizations, and the larger the α is, the closer the realization of G to G_0.

The DP is an infinite-dimensional analogue of the finite-dimensional Dirichlet prior, which has its roots in the one-dimensional Beta distribution. Therefore, most of the properties of DP arise as an extension of the properties of the Dirichlet distribution. For any measurable subset $B \in \mathcal{X}$, $G(B) \sim Beta(\alpha G_0(B), \alpha G_0(B^c))$ and thus

$$E(G(B)) = G_0(B) \tag{3.4}$$

$$Var(G(B)) = \frac{G_0(B)\left\{1 - G_0(B)\right\}}{\alpha + 1} \tag{3.5}$$

The DP prior distribution also has a conjugacy property. Assume $y_i \sim G$ for $i = 1, ..., n$, and $G \sim DP(\alpha, G_0)$. Let $\delta_y(.)$ denote a point mass at y. Therefore, the posterior distribution is $G|y_1, ..., y_n \sim DP(\alpha + n, G_1)$ with $G_1 \propto G_0 + \sum_{i=1}^{n} \delta_{y_i}$ [**83**],[**46**].

In original forms of the DP prior, G_0 is assumed to be known (fixed). One problem with a DP when G_0 is known is that it produces distributions that are discrete with probability one. Another option is to assume that the parameters in G_0 are unknown and following a set of parametric distributions, with possibly unknown hyperparameters, resulting in a mixture of Dirichlet process model [**24**]. The Dirichlet process mixture model will be reviewed in Section 3.6.

As the DP prior is defined through the marginal probabilities allocated to finite partitions, it does not provide any intuition for what realisations $G \sim DP(\alpha, G_0)$ actually look like. Therefore in the next section constructive representations of Dirichlet process will be described.

3.5 Constructive Definition of Dirichlet Process

Many authors examine various representations for the Dirichlet process including the gamma process, stick-breaking, Chinese restaurant process and also Pólya Urn. In this section the stick-breaking method and Pólya Urn scheme are described in detail.

3.5.1 Stick-breaking Method

The stick-breaking method is one of the constructive methods to generate realizations from Dirichlet process. It was first introduced by Sethuraman and Tiwari 1982 [**97**] and Sethuraman 1994 [**96**].
Assume $\{z_r : r = 1, 2, ..,\}$ and $\{\theta_l : l = 1, 2, ...\}$ are independent sequences of independent and identically distributed (i.i.d.) random variables. Generate z_r and atom θ_l as i.i.d. random variables as follows:

$$\begin{aligned} z_r &\overset{iid}{\sim} Beta(1, \alpha) \qquad r = 1, 2, \dots \\ \theta_l &\overset{iid}{\sim} G_0 \qquad\qquad\quad l = 1, 2, \dots \end{aligned}$$

Define $\omega_1 = z_1$ and $\omega_l = z_l \prod_{r=1}^{l-1}(1 - z_r)$ for $l = 2, 3, ...$ Construction of ω can be thought as a stick-breaking procedure. At each stage, we independently and randomly, break what is left of the unit length and assign the length of this break to the current ω_l. In other words, a random piece of length z_1 is broken off with the length generated from a $Beta(1, \alpha)$ distribution and allocate this $\omega_l = z_l$ probability weight to the randomly generated first atom $\theta_l \sim G_0$. So, $1 - z_l$ of the stick remains to be assigned to the other atoms. Then, a proportion $z_2 \sim Beta(1, \alpha)$ of the $1 - z_l$

is broken off and allocated to the probability $\omega_2 = z_2(1 - z_1)$ to the second atom $\theta_2 \sim G_0$. Consequently, the sticks get shorter so that the length which is allocated to the latest atom decreases stochastically, with a rate of decrease that depends on α. Hence a realization G from $DP(\alpha, G_0)$ using stick-breaking method is (almost surely) of the form:

$$G(.) = \sum_{l=1}^{\infty} \omega_l \delta_{\theta_l}(.),$$

where $\delta_\theta(.)$ denotes a degenerate distribution with all its mass at θ and ω_l is the probability mass at atom θ_l. In our simulations which will describe in Section 3.5.3, we work with a finite number of terms (N) in the above sum.

3.5.2 Pólya Urn Method

The Pólya urn is another method of representing Dirichlet process. Blackwell and MacQueen [**11**] describe the connections between Pólya sequences and the Ferguson distribution.

Assume $\theta_i|G$ for $i = 1, 2, .., N$ are i.i.d from G, and $G \sim DP(\alpha, G_0)$, then, a sequence of θ_i follows a generalized Pólya urn scheme with:

- $\theta_1 \sim G_0$
- For $i = 2, ..., N$, $\theta_i|\theta_1, ..., \theta_{i-1}$ distributed according to the mixed distribution that puts point mass $(\alpha + i - 1)^{-1}$ at θ_j, $j = 1, ..., i-1$ and mass $\alpha(\alpha + i - 1)^{-1}$ on the continuous G_0.

Hence:

$$p(\theta_1, ..., \theta_N) = G_0(\theta_1) \prod_{i=2}^{N} \left\{ \frac{\alpha}{\alpha + i - 1} G_0(\theta_i) + \frac{1}{\alpha + i - 1} \sum_{j=1}^{i-1} \delta_{\theta_j(\theta_i)} \right\}$$

3.5.3 Illustrative Web Application

As a preliminary learning step for developing MCMC samplers for *a posteriori* inference and prediction from the Dirichlet process, a `Shiny` (CRAN-R) application

has been implemented. `Shiny` is a new package from RStudio that makes it easier to build interactive web applications with R. The `Shiny` package allows the creation of user-friendly graphical interfaces for R code to be displayed as a webpage (Shiny) [**19**]. Basically, the Shiny framework is for making the input values entered in a web page readily accessible to R and taking the R output results back to the web page. A Shiny application consists of two main components; a user-interface script (`ui`), and a server script (`server`). The `ui` controls the layout and appearance of the application while the `server` specifies all the R actions. The power of Shiny arises from the ability to use reactive expressions, which are R objects that take inputs, apply R code for the required actions and return another R object of interest.
Our application provides for sampling from a Dirichlet process (DP) and a simple DP location mixture model. It illustrates how the sample paths from the DP process are affected by changes in the main parameters, as listed below. In this application G_0 is assumed to have a normal mixture distribution. In the shiny implementation the following inputs can be set.

- Simulation representation method for Dirichlet process — stick-breaking or Pólya urn;
- the number of θ atoms, N;
- the number of simulations;
- the precision parameter α;
- the weight parameter ω for the mixture of normals in G_0;
- the location and scale parameters of the mixture distributions in G_0.

By changing these values and clicking the `Refresh` button, the app instantly updates the inputs and subsequently draws a new plot. The Shiny R code is listed in the Appendix.
Figure 3.2 shows the effect of changing α, N and the number of simulations in generating data from DP based on the stick-breaking method using the `Shiny` application. The sample paths from a DP process are affected by changing N, the number of simulations, and the value of α. Since $E(z_l) = \frac{1}{1+\alpha}$, values of α near zero

lead to high weight on the first couple of atoms with the remaining atoms being assigned small probabilities. For large α, the DP prior effectively draws from the base parametric distribution G_0.

Figure 3.3 shows the effect of changing N, α and the number of simulations in generating data from a DP based on a Pólya urn method using the Shiny application. In Figure 3.3 we take G_0 to have a normal mixture distribution with parameters μ_1, σ_1, μ_2, σ_2 and ω that can be set interactively to show different mixtures. Again based on Figure 3.3, the sample paths from a DP process are affected by changing N, the number of simulations, and the size of α.

3.6 Dirichlet Process Mixture Model

A Dirichlet process prior as described in the previous section is a simple and computationally tractable prior for an unknown distribution. However, it produces distributions that are discrete with probability one, making it unsuitable for density modelling. To solve this problem, the distribution can be convolved with some continuous kernel, or more generally, by using a DP to define a mixture distribution with infinitely many components, of some simple parametric form.

A Dirichlet process mixture (DPM) model is a mixture with a parametric kernel and a random mixing distribution modelled with a DP prior, see Ferguson [**38**], [**39**], Antoniak [**6**], Escobar and West [**35**]. A DP mixture model is given by

$$F(.;G) = \int K(.|\theta)G(d\theta) \tag{3.6}$$

where $K(.|\theta)$ is the distribution function of the parametric kernel of the mixing and $G \sim DP(\alpha, G_0)$. The corresponding density function is given by

$$f(.;G) = \int k(.|\theta)G(d\theta) \tag{3.7}$$

where $k(.|\theta)$ is the density function corresponding to $K(.|\theta)$.

Choosing an appropriate kernel depends on the underlying sample space. If the underlying density function is defined on the entire real line, a location-scale kernel

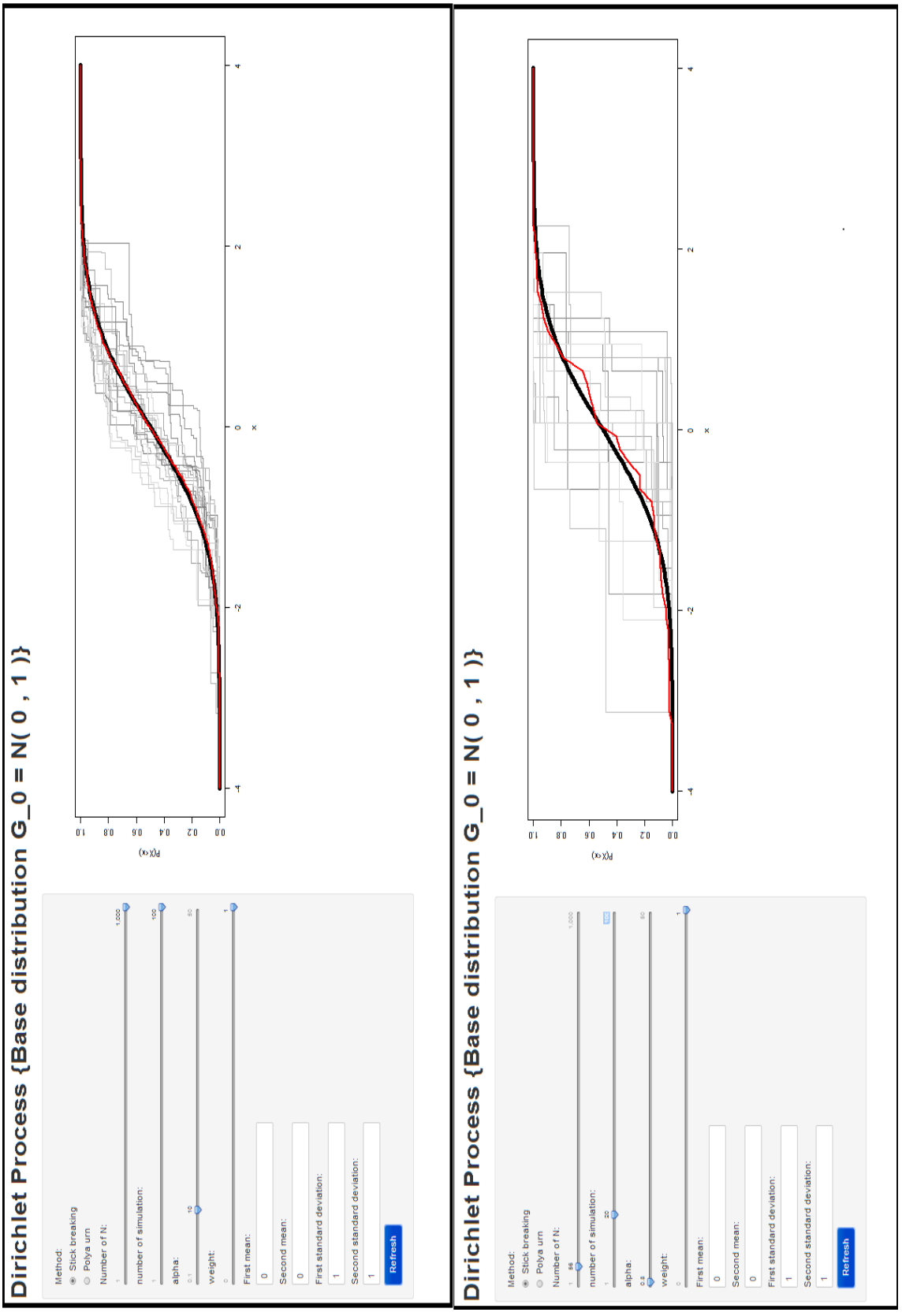

Figure 3.2: These graphs shows the sample path from DP process using stick-breaking method with $G_0 \equiv N(0, 1)$. The heavy smooth line indicates $N(0, 1)$.

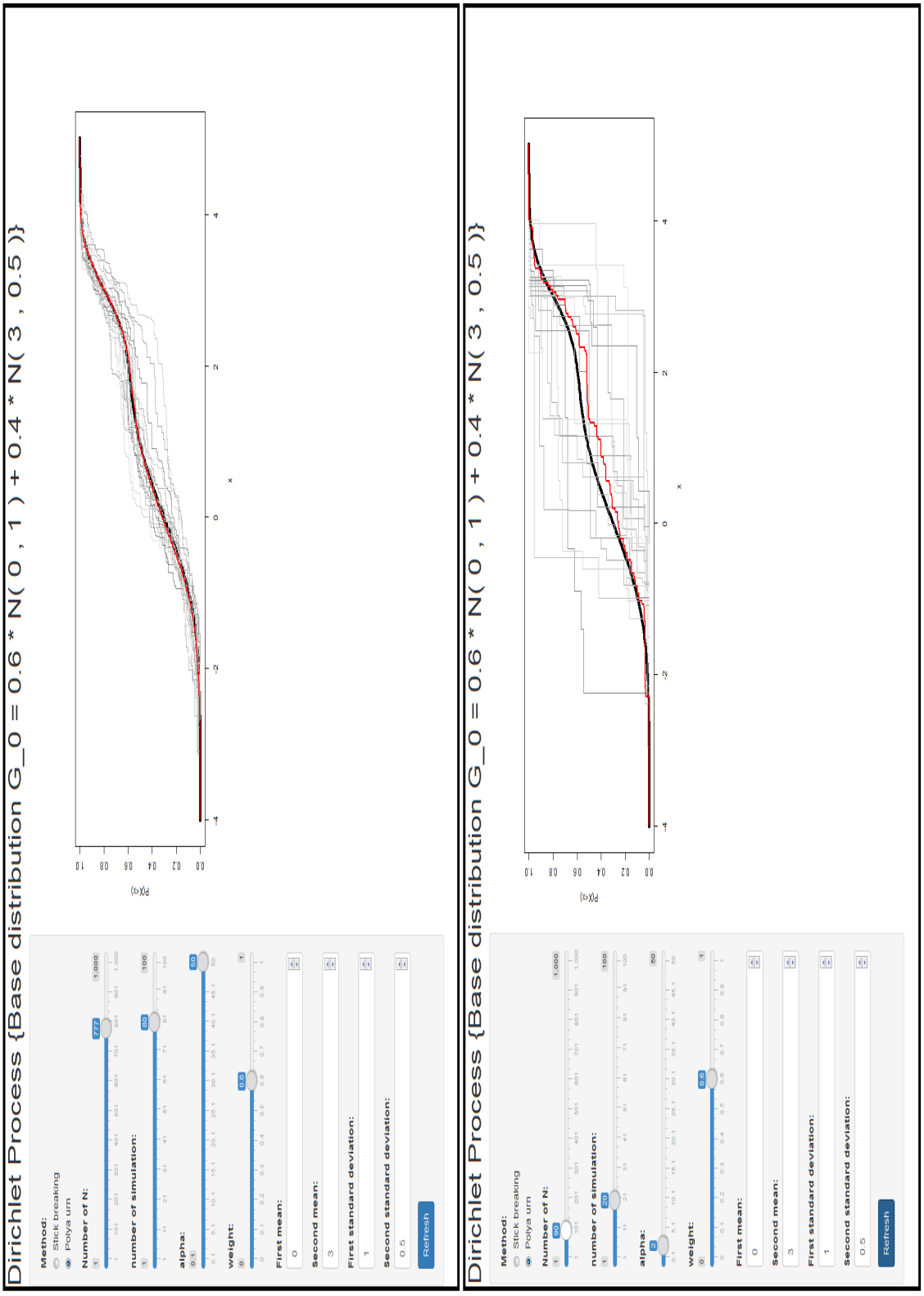

Figure 3.3: These graphs shows the sample path from DP process using Pólya Urn method where G_0 has mixture of normal distribution with the weight 0.6 ($G_0 \equiv 0.6N(0,1) + 0.4N(3, 0.5)$). The heavy smooth line indicates the mixture of normal distribution.

can be used while on the positive half-line, mixtures of gamma, Weibull or lognormal may be appropriate.

The model (3.6) is essentially a Bayesian hierarchical model that can be expressed in the following way. If $Y_1, ..., Y_n$ are i.i.d. realizations from $f(.; G)$, for each Y_i a latent θ_i can be introduced and the model can be written as follows:

$$\begin{aligned} &Y_i|\theta_i, \phi \overset{ind}{\sim} k(\theta_i, \phi), \quad i = 1, ..., n, \\ &\theta_i|G \overset{i.i.d}{\sim} G, \quad i = 1, ..., n, \\ &G|\alpha, \psi \sim DP(\alpha G_0), G_0 = G_0(.|\psi) \\ &\alpha, \psi, \phi \sim p(\alpha)p(\psi)p(\phi), \end{aligned} \tag{3.8}$$

where $p(\alpha)p(\psi)p(\phi)$ denotes the assumption of independent priors for α, ψ and ϕ, with $p(\cdot)$ denoting a generic distribution.

In the next section, posterior inference for DPM models is described, based on MCMC posterior simulation.

3.7 Simulation Based Model Fitting

A break-through in fitting Dirichlet process mixture model is based on the work of Escobar [**34**] which is extended in Escobar and West [**35**], who realized the potential of using MCMC methods in this context.

In Bayesian non-parametric methods the discreteness of the random distribution G induces a clustering of θ. Assume n^* is the number of distinct elements (clusters) in the vector $(\theta_1, ..., \theta_n)$ defined as $\theta_j^*, j = 1, ..., n$. let $s = s_1, ..., s_n$ be the vector of configuration indicator, $s_i = j$ if and only if $\theta_i = \theta_j^*, i = 1, ..., n$. Also, assume n_j be the size of cluster j. It is obvious that $(n^*, s, (\theta_1, ..., \theta_{n^*}))$ gives the same information as $(\theta_1, ..., \theta_n)$.

The main goal is to draw from the conditional distribution $p(\theta_1, ..., \theta_n, \alpha, \psi, \phi|data)$. A standard Gibbs sampling approach is based on following full conditionals:

(a) $p((\theta_i, s_i)|\{(\theta_{i'}, s_{i'}), i' \neq i\}, \alpha, \psi, \phi, data)$, for $i = 1, ..., n$.

(b) $p(\theta_j^*|s, n^*, \psi, \phi, data)$, for $j = 1, ..., n^*$.

(c) $p(\alpha|n^*, data)$ and $p(\psi|\{\theta_j^*, j = 1, ..., n^*\}, n^*)$.

(d) $p(\phi|\{\theta_i, i = 1, ..., n\}, data)$.

Each of the full conditionals in (a) is obtained by multiplying the likelihood term for θ_i, $k(y_i; \theta_i, \phi)$, and the full conditional prior $p(\theta_i|\{\theta_{i'}, i' \neq i\}, \alpha, \phi)$.
A superscript "-" denotes all relevant quantities when θ_i is removed from the vector $(\theta_1, ..., \theta_n)$. So the n^{*-} is defined as the number of clusters in $\{\theta_{i'} i' \neq i\}$ when θ_i is removed. Also, n_{j^-} is the number of elements in cluster $j, j = 1, ..., n^{*-}$ when θ_i is removed and define θ_j^* as the distinct cluster values.
For each $i = 1, ..., n$, the full conditional in (a) is the mixed distribution as follows:

$$p(\theta_i|\{\theta_{i'}, i' \neq i\}, \alpha, \phi, data) = \frac{q_0 h(\theta_i|\gamma, \phi, y_i) + \sum_{j=1}^{n^{*-}} n_j^- q_j \delta_{(\theta_j^*)}(\theta_i)}{q_0 + \sum_{j=1}^{n^{*-}} n_j^- q_j} \tag{3.9}$$

Where

- $q_j = k(y_i; \theta_j^*, \phi)$
- $q_0 = \alpha \int k(y_i; \theta, \phi) g_0(\theta|\psi) d\theta$
- $h(\theta_i|\gamma, \phi, y_i) \propto k(y_i; \theta_i, \phi) g_0(\theta_i|\psi)$

where g_0 is the density of G_0.
Once step (a) is finished, a specific number of clusters (n^*), specific configuration s and the associated cluster locations θ_j^* are produced.
In step(b) by sampling, and thus moving cluster locations, the mixing of the chain is improved [**15**]. For each $j = 1, ..., n^*$

$$p(\theta_j^*|s, n^*, \psi, \phi, data) \propto g_0(\theta_j^*|\psi) \prod_{\{i:s_i=j\}} k(y_i; \theta_j^*, \phi) \tag{3.10}$$

Recall that for a Dirichlet process $DP(\alpha, G_0)$, α controls how close a realisation G is to G_0. In the Dirichlet process mixture model α controls the distribution of the number of distinct elements n^* of the vector $\theta = (\theta_1, ..., \theta_n)$ and hence the number

of distinct components of the mixture [**6**], [**35**] [**73**].
In using the result of Antoniak [**6**],

$$P(n^*|\alpha, n) = c_n(n^*)n!\alpha^{n^*}\frac{\Gamma(\alpha)}{\Gamma(\alpha+n)} \quad n^* = 1, ..., n \tag{3.11}$$

where $c_n(n^*) = P(n^*|\alpha = 1, n)$, not involving α. It can be computed using recurrence formulas for Stirling numbers. The full conditional for α, the precision parameter of the Dirichlet process, is generated using the augmentation method in Escobar and West [**35**]. The full conditional for α depends only on n^* and on the data only through n. Assume a $Gamma(a_\alpha, b_\alpha)$ prior for α with mean a_α/b_α also by knowing

$$\frac{\Gamma(\alpha)}{\Gamma(\alpha+n)} = \frac{(\alpha+n)\beta(\alpha+1, n)}{\alpha\Gamma(n)} \tag{3.12}$$

we have

$$\begin{aligned} p(\alpha|n^*, data) &\propto p(\alpha)\alpha^{n^*}\frac{\Gamma(\alpha)}{\Gamma(\alpha+n)} \\ &\propto p(\alpha)\alpha^{n^*-1}(\alpha+n)\beta(\alpha+1, n) \\ &\propto p(\alpha)\alpha^{n^*-1}(\alpha+n)\int_0^1 x^\alpha(1-x)^{n-1}dx \end{aligned} \tag{3.13}$$

Now if an auxiliary variable u is introduced such that

$$p(\alpha, u|n^*, data) \propto p(\alpha)\alpha^{n^*-1}(\alpha+n)u^\alpha(1-u)^{n-1} \tag{3.14}$$

This implies that $p(\alpha|n^*, data)$ is the marginal distribution for α arising from (3.14). The Gibbs sampler can be extended to draw from the full conditionals for α and u resulting from(3.14). Specifically, $p(u|\alpha, data) \propto Beta(\alpha+1, n)$ and

$$p(\alpha|u, n^*, data) \propto p\,Gamma(a_\alpha+n^*, b_\alpha-log(u))+(1-p)\,Gamma(a_\alpha+n^*-1, b_\alpha-log(u))$$

where

$$p = \frac{a_\alpha + n^* - 1}{n(b_\alpha - log(u)) + a_\alpha + n^* - 1}$$

The full conditional for ψ is as follows:

$$p(\psi|\{\theta_j^*, j = 1, ..., n^*\}, n^*) \propto p(\psi) \prod_{j=1}^{n^*} g_0(\theta_j^*|\psi) \tag{3.15}$$

Finally, the full conditional for the last step is simple and as follows, since it does not involve the non-parametric part of the model.

$$p(\phi|\{\theta_i, i = 1, ..., n\}, data) \propto p(\phi) \prod_{i=1}^{n} k(y_i; \theta_i, \phi) \tag{3.16}$$

We will describe this process below for the particular case of normal Dirichlet process mixture model.

3.8 Posterior Predictive Distribution

By using the MCMC methods described in the previous section, posterior samples can be obtained. A Bayesian density estimate is based on the posterior predictive density $p(y_{new}|data)$ with associated mixing parameter θ_{new}. Using the Polya urn structure for the Dirichlet process,

$$p(\theta_{new}|n^*, s, \theta^*, \alpha, \psi) = \frac{\alpha}{\alpha + n} G_0(\theta_{new}|\psi) + \frac{1}{\alpha + n} \sum_{j=1}^{n^*} n_j \delta_{\theta_j^*}(\theta_{new}) \tag{3.17}$$

The posterior predictive distribution for y_{new} is given by

$$\begin{aligned} p(y_{new}|data) &= \int p(y_{new}|n^*, s, \theta^*, \alpha, \psi, \phi) p(n^*, s, \theta^*, \alpha, \psi, \phi|data) \\ &= \iint k(y_{new}|\theta_{new}, \phi) p(\theta_{new}|n^*, s, \theta^*, \alpha, \psi) p(n^*, s, \theta^*, \alpha, \psi, \phi|data) \end{aligned} \tag{3.18}$$

By using (3.17) we have

$$\begin{aligned} p(y_{new}|data) = \int \left(\frac{\alpha}{\alpha + n} \int k(y_{new}|\theta, \phi) g_0(\theta|\psi) d\theta + \frac{1}{\alpha + n} \sum_{j=1}^{n^*} n_j k(y_{new}; \theta_j^*, \phi) \right) \\ p(n^*, s, \theta^*, \alpha, \psi, \phi|data) \end{aligned} \tag{3.19}$$

The above integral is a mixture of n^*+1 components. The last n^* components (that dominate when α is small relative to n) yield a discrete mixture (in θ) of $k(.;\theta,\phi)$ with the mixture parameters defined by the distinct θ_j^*. The posterior predictive density for y_{new} is achieved by averaging this mixture with respect to the posterior of n^*, s, θ^* and all other parameters.

3.9 Example of Dirichlet Process Mixture Model

As an example, the location normal Dirichlet process mixture model is used to illustrate how the simulation method is applied.

$$\begin{aligned} &Y_i|\theta_i,\phi \overset{ind}{\sim} N(y_i;\theta_i,\phi), \quad i=1,...,n, \\ &\theta_i|G \overset{i.i.d}{\sim} G, \quad i=1,...,n, \\ &G|\alpha,\mu,\sigma^2 \sim DP(\alpha G_0), G_0 = N(\mu,\tau^2) \end{aligned} \tag{3.20}$$

with hyperpriors:

$$\begin{aligned} p(\alpha) &= Gamma(a_\alpha, b\alpha) \\ p(\mu) &= N(a_\mu, b_\mu) \\ p(\tau^2) &= IGamma(a_{\tau^2}, b_{\tau^2}) \\ p(\phi) &= IGamma(a_\phi, b_\phi) \end{aligned}$$

At first we need to find $p(\theta_i|\{\theta_{i'}, i' \neq i\}, \alpha, \phi, \mu, \tau^2, \phi, data)$ for $i = 1, ..., n$ using (3.9). So we need to find q_0 and $h(\theta_i|\mu, \tau^2, \phi, y_i)$.

$$
\begin{aligned}
q_0 &= \alpha \int k(y_i; \theta, \phi) g_0(\theta|\mu, \tau^2) d\theta \\
&= \alpha \int_{-\infty}^{\infty} N(y_i; \theta_i, \phi) N(\theta|\mu, \tau^2) d\theta \\
&= \alpha \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\phi}} exp\left(\frac{-(y_i-\theta)^2}{2\phi}\right) \frac{1}{\sqrt{2\pi\tau^2}} exp\left(\frac{-(\theta-\mu)^2}{2\tau^2}\right) d\theta \\
&= \frac{\alpha}{\sqrt{2\pi\phi}\sqrt{2\pi\tau^2}} exp\left(\frac{-y_i^2}{2\phi}\right) exp\left(\frac{-\mu^2}{2\tau^2}\right) \int_{-\infty}^{\infty} exp\left(\frac{-\theta^2(\tau^2+\phi) + 2\theta(y_i\tau^2+\mu\phi)}{2\phi\tau^2}\right) d\theta \\
&= \frac{\alpha}{\sqrt{2\pi\phi}\sqrt{2\pi\tau^2}} exp\left(\frac{-y_i^2}{2\phi}\right) exp\left(\frac{-\mu^2}{2\tau^2}\right) exp\left(\frac{(y_i\tau^2+\mu\phi)^2}{2\phi\tau^2(\phi+\tau^2)}\right) \\
&\times \int_{-\infty}^{\infty} exp\left(\frac{-\left(\theta - \frac{y_i\tau^2+\mu\phi}{\phi+\tau^2}\right)^2}{\frac{2\phi\tau^2}{(\phi+\tau^2)}}\right) d\theta \\
&= \frac{\alpha}{\sqrt{2\pi(\phi+\tau^2)}} exp\left\{\frac{-(y_i-\mu)^2}{2(\phi+\tau^2)}\right\} \qquad (3.21)
\end{aligned}
$$

Also

$$
\begin{aligned}
h(\theta_i|\mu, \tau^2, \phi, y_i) &= k(y_i; \theta, \phi) g_0(\theta|\mu, \tau^2) \\
&= \frac{1}{\sqrt{2\pi\phi}} exp\left(\frac{-(y_i-\theta)^2}{2\phi}\right) \frac{1}{\sqrt{2\pi\tau^2}} exp\left(\frac{-(\theta-\mu)^2}{2\tau^2}\right) \\
&\propto exp\left(\frac{-\left(\theta - \frac{y_i\tau^2+\mu\phi}{\phi+\tau^2}\right)^2}{\frac{2\phi\tau^2}{(\phi+\tau^2)}}\right) \\
&\sim N\left(\frac{y_i\tau^2+\mu\phi}{\phi+\tau^2}, \frac{\phi\tau^2}{\phi+\tau^2}\right) \qquad (3.22)
\end{aligned}
$$

So, by substituting (3.21),(3.22) and $q_j = N(y_i; \theta_j^*, \phi)$ in (3.9), $p(\theta_i|\{\theta_{i'}, i' \neq i\}, \alpha, \phi, \mu, \tau^2, \phi, data)$ can be found.

By using (3.10), $p(\theta_j^*|s, n^*, \phi, \mu, \tau^2, data)$ which is the resampling of the cluster lo-

cations θ_j^*.

$$
\begin{aligned}
p(\theta_j^*|s, n^*, \phi, \mu, \tau^2, data) &\propto g_0(\theta_j^*|\mu, \tau^2) \prod_{\{i:s_i=j\}} k(y_i; \theta_j^*, \phi) \\
&\propto N(\theta_j^*|\mu, \tau^2) \prod_{\{i:s_i=j\}} N(y_i; \theta_j^*, \phi) \\
&\propto \frac{1}{\sqrt{2\pi\tau^2}} exp\left(\frac{-(\theta_j^* - \mu)^2}{2\tau^2}\right) \times \left(\frac{1}{\sqrt{2\pi\phi}}\right)^{n_j} exp\left(\frac{-\sum_{i:s_i=j}(y_i - \theta_j^*)^2}{2\phi}\right) \\
&\propto exp\left(\frac{-(\theta_j^{*2} - 2\mu\theta_j^*)}{2\tau^2}\right) \times exp\left(\frac{-(n_j\theta_j^{*2} - 2\theta_j^* n_j \bar{y_j^*})}{2\phi}\right) \\
&\propto exp\left(\frac{-1}{2\phi\tau^2}\left\{(\phi + n_j\tau^2)\left[\theta_j^* - \left(\frac{\mu\phi + n_j\tau^2\bar{y_j^*}}{\phi + n_j\tau^2}\right)\right]^2\right\}\right) \\
&\sim N\left(\frac{\mu\phi + n_j\tau^2\bar{y_j^*}}{\phi + n_j\tau^2}, \frac{\phi\tau^2}{\phi + n_j\tau^2}\right) \qquad (3.23)
\end{aligned}
$$

where n_j is the size of cluster j and $\bar{y_j^*} = \frac{1}{n_j}\sum_{i:s_i=j} y_i$.
$p[\alpha|n^*, data]$ is calculated in (3.13). By using (3.15), $p(\mu|\{\theta_j^*, j = 1, ..., n^*\}, n^*)$ and $p(\tau^2|\{\theta_j^*, j = 1, ..., n^*\}, n^*)$ can be found as follows:

$$
\begin{aligned}
p(\mu|\{\theta_j^*, j = 1, ..., n^*\}, n^*) &\propto N(\mu|a_\mu, b_\mu) \prod_{j=1}^{n^*} N(\theta_j^*|\mu, \tau^2) \\
&\propto \frac{1}{\sqrt{2\pi b_\mu}} exp\left(\frac{-(\mu - a_\mu)^2}{2b_\mu}\right) \prod_{j=1}^{n^*} \frac{1}{\sqrt{2\pi\tau^2}} exp\left(\frac{-(\theta_j^* - \mu)^2}{2\tau^2}\right) \\
&\propto exp\left(\frac{-1}{2b_\mu\tau^2}\left\{\mu^2(\tau^2 + n^* b_\mu) - 2\mu(a_\mu\tau^2 + b_\mu n^*\bar{\theta^*})\right\}\right) \\
&\propto exp\left(\frac{-1}{\frac{2b_\mu\tau^2}{\tau^2 + n^* b_\mu}}\left\{\mu - \frac{a_\mu\tau^2 + b_\mu n^*\bar{\theta^*}}{\tau^2 + n^* b_\mu}\right\}^2\right) \\
&\sim N\left(\frac{a_\mu\tau^2 + b_\mu n^*\bar{\theta^*}}{\tau^2 + n^* b_\mu}, \frac{b_\mu\tau^2}{\tau^2 + n^* b_\mu}\right) \qquad (3.24)
\end{aligned}
$$

where $\bar{\theta^*} = \dfrac{\sum_{j=1}^{n^*} \theta_j^*}{n^*}$.

$$
\begin{aligned}
p(\tau^2|\{\theta_j^*, j=1,...,n^*\}, n^*) &\propto IGamma(a_{\tau^2}, b_{\tau^2}) \prod_{j=1}^{n^*} N(\theta_j^*|\mu, \tau^2) \\
&\propto \frac{b_{\tau^2}^{a_{\tau^2}}}{\Gamma(a_{\tau^2})} (\tau^2)^{-a_{\tau^2}-1} exp\left(\frac{-b_{\tau^2}}{\tau^2}\right) \prod_{j=1}^{n^*} \frac{1}{\sqrt{2\pi\tau^2}} exp\left(\frac{-(\theta_j^*-\mu)^2}{2\tau^2}\right) \\
&\propto (\tau^2)^{-a_{\tau^2}-1} exp\left(\frac{-b_{\tau^2}}{\tau^2}\right) \left(\frac{1}{\sqrt{2\pi\tau^2}}\right)^{n^*} exp\left(\frac{-\sum_{j=1}^{n^*}(\theta_j^*-\mu)^2}{2\tau^2}\right) \\
&\propto (\tau^2)^{-(a_{\tau^2}+\frac{n^*}{2})-1} exp\left(\frac{-1}{\tau^2}\left\{b_{\tau^2} + \frac{1}{2}\sum_{j=1}^{n^*}(\theta_j^*-\mu)^2\right\}\right) \\
&\sim IGamma\left(a_{\tau^2} + \frac{n^*}{2}, b_{\tau^2} + \frac{1}{2}\sum_{j=1}^{n^*}(\theta_j^*-\mu)^2\right) \qquad (3.25)
\end{aligned}
$$

$p(\phi|\{\theta_i, i=1,...,n\}, data)$ can be calculated using (3.16) as follows:

$$
\begin{aligned}
p(\phi|\{\theta_i, i=1,...,n\}, data) &\propto IGamma(a_\phi, b_\phi) \prod_{i=1}^{n} N(y_i; \theta_i, \phi) \\
&\propto \frac{b_\phi^{a_\phi}}{\Gamma(a_\phi)} (\phi)^{-a_\phi-1} exp\left(\frac{-b_\phi}{\phi}\right) \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\phi}} exp\left(\frac{-(y_i-\theta_i)^2}{2\phi}\right) \\
&\propto (\phi)^{-a_\phi-1} exp\left(\frac{-b_\phi}{\phi}\right) (\phi)^{\frac{-n}{2}} exp\left(\frac{-\sum_{i=1}^{n}(y_i-\theta_i)^2}{2\phi}\right) \\
&\propto (\phi)^{-(a_\phi+\frac{n}{2})-1} exp\left(\frac{-1}{\phi}\left\{b_\phi + \frac{1}{2}\sum_{i=1}^{n}(y_i-\theta_i)^2\right\}\right) \\
&\sim IGamma\left(a_\phi + \frac{n}{2}, b_\phi + \frac{1}{2}\sum_{i=1}^{n}(y_i-\theta_i)^2\right) \qquad (3.26)
\end{aligned}
$$

Finally, by using (3.19) and (3.21) the posterior predictive distribution can be calculated using the following formula:

$$\begin{aligned} p(y_{new}|data) &= \int \left(\frac{\alpha}{\alpha+n} \int k(y_{new}|\theta,\phi) g_0(\theta|\psi) d\theta \right. \\ &\quad \left. + \frac{1}{\alpha+n} \sum_{j=1}^{n^*} n_j k(y_{new};\theta_j^*,\phi) \right) p(n^*, s, \theta^*, \alpha, \psi, \phi|data) \\ &= \int \left(\frac{\alpha}{\alpha+n} N(y_0|\mu, \phi+\tau^2) + \frac{1}{\alpha+n} \sum_{j=1}^{n^*} n_j N(y_{new};\theta_j^*,\phi) \right) \\ &\quad p(n^*, s, \theta^*, \alpha, \psi, \phi|data) \end{aligned} \tag{3.27}$$

Related R code can be found in the Appendix.

`DPpackage` in R which was developed by Jara [**60**] can also be used for implementation of some non-parametric Bayesian models.

3.9.1 Illustration

As an example, we discuss the well-known galaxy data, which was described in Section 1.1.1. The `Galaxy` dataset considers physical information on the recession velocities (km/second) for 82 galaxies. The data is sampled from six well-separated conic sections of the Corona Borealis region. The question of interest is whether these galaxies form distinct super-clusters surrounded by voids in space. As there is a jump between the seven smallest observations around 10 and also another jump in the central body of observations between 16 and 26 and moreover there is a gap between 27 and 32, for the three largest observations, so we would expect to find at least three components in the data. For ease of presentation the velocities were divided by 1000.

Fitting the location-normal Dirichlet process mixture model, as described in (3.20), the corresponding posterior density for the data is displayed in Figure 3.4. Based on Figure 3.4, five components are suggested for the Galaxy data.

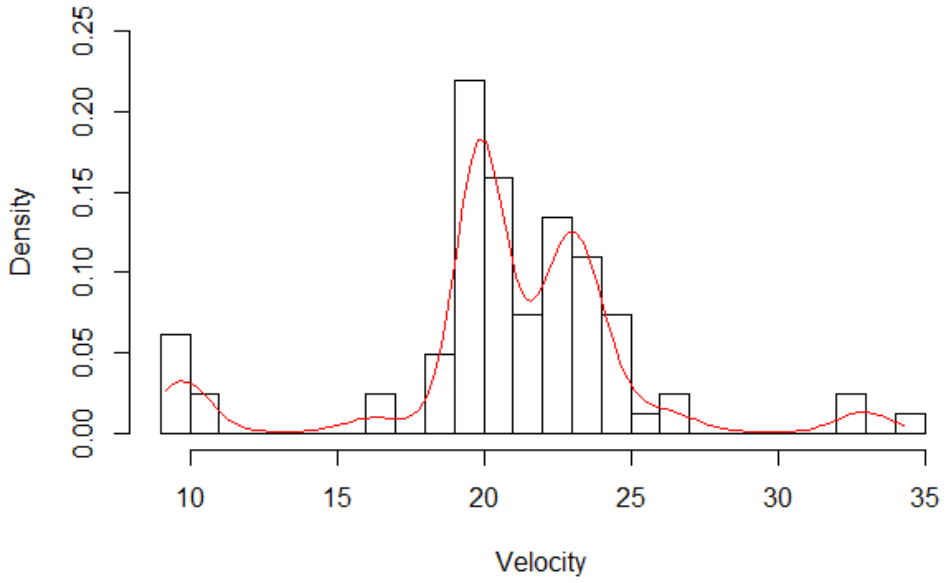

Figure 3.4: Posterior density for Galaxy data using model described in (3.20)

`Dpdensity` function in `DPpackage` is also used to generate a posterior density for a Dirichlet process of a mixture of normals models. The following model is assumed to estimate the density:

$$
\begin{aligned}
y_i \mid \mu_i, \Sigma_i &\sim N(\mu_i, \Sigma_i) \\
(\mu_i, \Sigma_i) \mid G &\sim G \\
G \mid \alpha, G_0 &\sim DP(\alpha, G_0)
\end{aligned}
$$

where the baseline distribution is the conjugate normal-inverted-Wishart

$$G_0 = N(\mu \mid m_1, (1/k_0)\Sigma) IW(\Sigma \mid \nu_1, \psi_1)$$

with hyperpriors

$$\alpha \mid a_0, b_0 \sim Gamma(a_0, b_0)$$
$$m_1 \mid m_2, s_2 \sim N(m_2, s_2)$$
$$k_0 \mid \tau_1, \tau_2 \sim Gamma(\tau_1/2, \tau_2/2)$$
$$\psi_1 \mid \nu_2, \psi_2 \sim IW(\nu_2, \psi_2)$$

Figure 3.5 shows the posterior density for Galaxy data using `Dpdensity` function in `DPpackage`. Based on Figure 3.5, five components are suggested for the data.

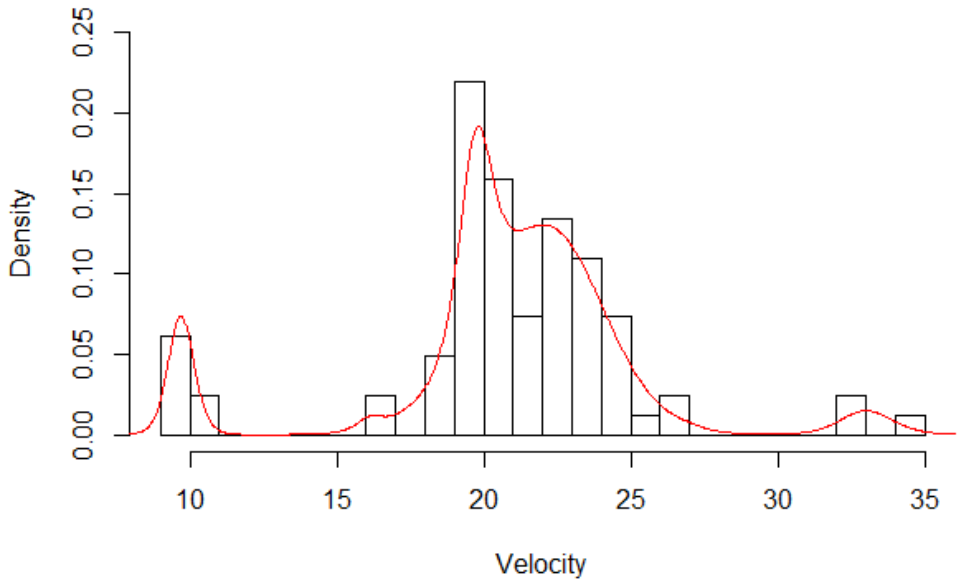

Figure 3.5: Posterior density for Galaxy data using `Dpdensity` function in `DPpackage`

3.10 Chapter Summary

In this chapter, we provide an overview of some important aspects of non-parametric Bayesian inference. The intention is not to give a complete account but rather to touch on areas of interest for the thesis. The non-parametric Bayesian approach is described briefly and how it can account for model uncertainty relating to the choice

of a parametric distribution. This is in contrast to classical non-parametric methods which may be distribution-free, here uncertainty about distributional shape is provided by a flexible probability model for the data-generating process. In this context the Dirichlet process prior is used as a flexible alternative to standard parametric priors. As Dirichlet priors produce, with probability one, posterior distributions that are discrete these are unsuitable for density modelling. The Dirichlet process mixture (DPM) model provides a solution to this that will be used as a non-parametric prior for survival distributions. The simulation method and posterior predictive distribution for DPM are discussed as they are required in imputing the censored observations.

As a preliminary learning step for developing MCMC samplers for *a posteriori* inference and prediction from the Dirichlet process, a `Shiny` (CRAN-R) application has been implemented. Our application provides for sampling from a Dirichlet process (DP) and a simple DP location mixture model. It illustrates how the sample paths from the DP process are affected by changes in the main parameters.

In the next chapter, the Bayesian framework is considered as a flexible approach to impute the censored observations using predictive distributions.

Chapter 4

A Bayesian Approach to Imputation of Survival Data

4.1 Introduction

In the presence of right censoring, standard methods of plotting individual survival times are invalid. By treating the censored observations as missing and using imputation methods, a complete dataset can be formed. Then standard graphics may usefully complement Kaplan-Meier plots.

In this chapter we consider using a Bayesian framework to present a flexible approach to impute the censored observations using predictive distributions. The method is intended to be used for the visual exploration and presentation of survival data. We illustrate its use for standard survivor and hazard function plots, which give a simple, interpretable display for physicians and patients to understand the results from clinical trials.

This chapter is organized as follows: we start with a review on imputation of missing data in Section 4.2. Then in Section 4.3 we give a brief introduction to imputing censored observations. In Section 4.4 the Royston parametric approach [**92**],[**93**] is introduced. In Section 4.5 we consider using a parametric Bayesian framework to impute the censored observations. In Section 4.6 non-parametric Bayesian methods (NPB) are used as a second approach to Bayesian imputation, as considering a fixed

distributional specification for the *random or error components* in the model may be inappropriate for the actual data. Finally, in Section 4.7 the Royston parametric approach is compared to our two proposed imputation methods utilising parametric Bayesian and non-parametric Bayesian methods.

4.2 Imputation Methods for Missing Data

Missing data appear in almost all statistical analyses. The missing data mechanism which describes the probability that a response is observed or missing is not under the control of the investigator. Instead, assumptions are made about the missing data mechanism, and the validity of analysis depends on these assumptions. A hierarchy of three different types of missing data can be distinguished [**80**].

The first and simplest type of missing data is *missing completely at random* (MCAR). The data are said to be missing completely at random if the probability of being missing is the same for all cases. This effectively implies that causes of the missing data are unrelated to the data. The primary aspect of MCAR is that the incomplete observed data can be thought of a random sample of the complete data. Accordingly, all moments, and even the joint distribution of the observed data do not differ from the corresponding moments or joint distribution of the complete data.

The second type of missing data is called *missing at random* (MAR) where the probability that responses are missing depends on the observed responses but is unrelated to the specific missing values that should have been obtained. MAR is a much broader class than MCAR.

The third missingness mechanism is known as *Missing Not At Random* (MNAR), also referred to as "non-ignorable" in much-published research. If missingness on the predictor is MNAR, it depends on the actual level of the predictor and potentially other variables not available in the data. Note that MNAR does not mean that missingness lacks a random component, only that its systematic component is a function of the actual values of the variable with missingness [**52**].

Researchers use different ways to deal with missing data including case deletion, single imputation and multiple imputation. A brief introduction of each method is

explained in the following subsections.

4.2.1 Case Deletion

Deleting the missing values is an obvious simple solution for dealing with missingness and this approach includes *listwise* and *pairwise* deletion methods.
In *listwise deletion*, also known as the complete case analysis, only full cases that have no missing data in any of the recorded variables are entered in the analysis. If the data are not MCAR, listwise deletion can severely bias estimates of means, regression coefficients and correlations. Although in the case of MCAR listwise deletion does not add any bias, it does decrease the power of the analysis by decreasing the effective sample size. Also, in real life applications, when the number of variables is large, there can be more than half of the original sample that is lost, and clearly a small subsample leads to a loss of power and degrades the ability to detect the effects of interest [**104**].
Pairwise deletion, also known as available case analysis, attempts to remedy the data loss problem in listwise deletion. This statistical procedure uses cases that contain some missing data. The process deletes a case when it is missing on a variable that is required for a particular analysis, but includes that case in analyses for other variables with non-missing values. The estimates can be biased if the data are not MCAR. Pairwise deletion enables you to use more of your data, but, each computed statistic may be based on a different subset of cases, which can be problematic. For example, a sample correlation matrix calculated using pairwise deletion may not be positive definite, which is a requirement for most multivariate procedures. Correlations outside the range [1,+1] can also occur, a problem that comes from different subsets used for the covariances and the variances. Such problems are more severe for highly correlated variables [**71**].

4.2.2 Single Imputation

Instead of discarding the unit entirely from the study, it is tempting to replace the missing items with substituted values, commonly referred to as *imputation*. Im-

putation has several desirable features. As no units are sacrificed, it is potentially more efficient than case deletion and, for example, the estimated sample correlation matrix remains positive definite. Moreover, retaining the full sample helps to prevent loss of power resulting from a diminished sample size. In single imputation, a variety of approaches are used to replace missing data.

A quick fix for the missing data is to replace them by the mean of that variable for all other complete cases, while the mode can be used for categorical data [**29**]. *Mean imputation* is a fast and simple way for imputing the missing data. However, it underestimates the variance, as the variability in the data is reduced with all replaced missing values being identical. The overall correlation estimate decreases due to filling in the data with a set of uncorrelated cases. Also, it leads to bias of almost any estimate other than the mean, and may even bias the estimate of the mean when data are not MCAR [**7**]. Mean imputation should perhaps only be used as a rapid fix when a small number of values are missing, and it should be avoided in general.

Another single imputation method which is often used is *regression imputation*. Regression imputation uses regression models to predict values for the missing entries of a variable based on other variables that have been measured for the subjects in the study. The first step involves building a model from the observed data. Predictions for the incomplete cases are then calculated under the fitted model and serve as replacements for the missing data. Regression imputation yields unbiased estimates of the means under MCAR, just as in mean imputation. Regression imputation is better than mean imputation as it considers other information that has been collected on a subject when imputing a value for that subject and gives different imputed missing values for different subjects. Nevertheless, it does not solve the problem associated with mean imputation of underestimated standard errors as any values which have to be imputed will lie along the fitted regression line and not reflect the variability of individual data values.

Stochastic regression imputation is a refinement of regression imputation that produces unbiased parameter estimates under a MAR missing data mechanism. This method uses regression equations to predict the incomplete variables from the complete variables, but it takes the extra step of augmenting each predicted score with a

normally distributed residual term. Adding residuals to the imputed values restores lost variability to the data and efficiently eliminates the biases associated with standard regression imputation schemes [**33**].

Hot-deck imputation is a collection of techniques replacing missing values with similar responding units in the sample [**72**]. In the basic form of hot-deck imputation, missing values are imputed with the scores of other similar respondents where a random draw from the observed data replaces each missing value. Hot-deck approaches are not well suited for estimating measures of association and can produce substantially biased estimates of correlations and regression coefficients.

The *last observation carried forward* (LOCF) is a missing data technique that is specific to longitudinal designs. This procedure involves filling in the missing values for a subject with their last recorded value for that particular measurement. Although this approach is simple, it makes the strong assumption that the value of the outcome remains unchanged after dropout, which seems likely to be unrealistic. Molenberghs and Kenward [**81**] show that LOCF can yield biased estimates even under MCAR. Generally, LOCF is not recommended for use.

Another simple idea developed before the modern computing era is the *indicator* method [**61**]. This method is popular in public health and epidemiology. Suppose that in the regression model there are missing values in one of the explanatory variables. The indicator method modifies the original regression model by adding a missingness indicator and, possibly, interactions between this indicator and the covariates. In other words, the indicator method replaces each missing value by a zero and extends the regression model by the response indicator. The procedure is applied to each incomplete variable. The user analyses the extended regression model instead of the original. An advantage is that the indicator method retains the full dataset. Also, it allows for systematic differences between the observed and the unobserved data by inclusion of the response indicator. The indicator method can yield estimates with much reduced standard errors relative to the complete-subject method. Unfortunately, the method can also yield severely biased regression estimates, even under MCAR and for low amounts of missing data [**102**] [**49**] [**65**].

Single imputation methods generally cause standard errors to be too small because they fail to consider the fact that we are uncertain about the missing values. In the

next section, multiple imputation techniques are discussed, which have the advantage that they can reflect the uncertainty of the missing data.

4.2.3 Multiple Imputation

The concept of multiple imputation (MI) was first described by Donald Rubin in a 1977 manuscript prepared for the United States Social Survey, and it is also represented in his 1987 book [**94**]. This approach has become an important method for dealing with the statistical analysis of incomplete data.

Multiple imputation is an extension of the single imputation method as it replaces each missing value by a vector of $D \geq 2$ imputed values. These D values give D completed data sets which can be formed from the vectors of imputations; the first completed dataset can be created by replacing each missing value with the first component in the vector of imputations, etc. Then each completed dataset is analysed using the standard complete-data procedure for the question of interest. As each analysis is performed D times, once for each imputed data set, the analysis phase yields D sets of parameter estimates and standard errors, and the D complete-data inferences can be combined to form a single inference that properly reflects uncertainty due to nonresponse under that model [**71**] using so-called Rubin's rule [**94**].

Multiple imputation seems a better alternative to single imputation, because by imputing a single value and treating that value as known, single imputation cannot reflect sampling variability under one model for nonresponse or uncertainty about the correct model for nonresponse. Also, by generating repeated randomly drawn imputations under more than one model, multiple imputation provides the study of the sensitivity of inferences to various models for nonresponse by using complete-data methods repeatedly.

Different approaches to performing multiple imputation can be described within three major categories [**103**]. Univariate (generally regression based) imputation methods for single variables and monotone missing patterns; data-augmentation imputation using a joint probability model; and the method of fully conditional univariate regression specification (also known as chained equations or sequential regression imputation) [**80**].

In *monotone missing data imputation*, variables that are subject to missing data can be ordered where the last variable in the ordering has the largest number of missing values. Therefore, imputations are created by drawing from a sequence of univariate conditional distributions $P(Y_j|Y_1, ..., Y_{j-1})$ for $j = 1, ..., p$. This is a powerful approach, but problems in which missingness has a monotone structure seem to be relatively uncommon in practice, motivating the need for more general approaches. *Joint modelling* assumes that the hypothetically complete data can be described by a multivariate distribution, so imputations are created as draws under the assumed model. By using the joint modelling method, the data Y is described by the multivariate distribution $P(Y|\theta)$, where θ is a vector of unknown parameters of the distribution. A major advance in the practical application was Schafer's development of computational algorithms for imputation under joint probability models, in particular, the multivariate normal distribution [**95**]. The response is assumed to have a normal distribution conditional on the covariates along with a multivariate normal distribution for the covariates. This approach uses Markov Chain Monte Carlo (MCMC) to fit this model to the observed data by treating missing values as parameters and updating them using an MCMC algorithm until the MCMC sampler converges. So the first imputed dataset contains current draws of missing values together with the observed values. Successive imputed datasets should be independent draws from the distribution of missing values given the observed data. The algorithm for this method is computationally stable and generates reasonable results for quite large numbers of variables. But users of this approach need to make a decision on how to handle variables which do not follow a normal distribution [**80**]. The term *fully conditional specification* (FCS) refers to a class of imputation models for non-monotone multivariate missing data. This approach relies on specifying univariate regression models for each variable, conditioning on all other variables in the dataset. In other words, the user specifies a conditional distribution $P(Y_j|Y_{-j})$ directly for each variable Y_j conditioning on all other variables in the dataset Y_{-j} (where $-j$ refers to the deletion of the jth variable) and assumes this distribution to be the same for the observed and missing Y_j. Imputations are created by iteratively drawing from these conditional distributions. The multivariate model $P(Y)$ is implicitly specified by the given sets of conditional models. The main reason

for using this method is the flexibility in allowing an appropriate univariate regression specification for each variable, which not only allows appropriate scaling and modelling of univariate error but also allows the univariate models to incorporate non-linear terms and interaction effects. Nevertheless, when multilevel structures with unbalanced data and partially observed variables at several levels of the hierarchy are considered, FCS loses its simplicity and computational attraction. For an appropriate choice of priors, FCS and joint modelling are equivalent for unstructured multivariate normal models[**18**].

To sum up, over the past decade multiple imputation methods have become a valuable tool for the incomplete data analysis. However, their assumptions and limitations need to be considered to prevent misleading application and conclusions.

In the next section, right censored observations in survival analysis are treated as missing values and a brief introduction to imputation methods for censored observations is given.

4.3 Imputing Censored Observations

In the presence of censoring the Kaplan-Meier estimator, described in Section 2.6, is an estimator of the survivor function. As described in Section 2.6, based on the Kaplan-Meier plot, the median survival time is the classical summary reported to the patients and is often mis-interpreted. However, the median may not be unique as if the Kaplan-Meier survival estimate is horizontal at $\hat{S}(t) = 0.5$ any value in the interval $I_j = [t_{(j)}, t_{(j+1)})$ is a reasonable estimate of the median where $t_{(1)} < t_{(2)} < ...$ are ordered observed event times. Also, in some data sets, the median survival time can not be calculated as the Kaplan-Meier survival curve does not reach the median because of extensive censoring [**12**], [**70**].

If there is no censoring in the data set, standard graphical and numerical summaries can be used. By imputing the censored observations and combining the original and imputed data a full data set can be constructed and it is possible to plot the histogram or density of the data to complement the information given by Kaplan-Meier plots.

The idea of imputing right censored survival data has, to some extent, already been considered in the literature. Wei and Tanner [**107**] proposed a method to impute failure times for the censored observations drawn from the conditional normal distribution in the context of regression analysis with censored data. Also in their other paper [**108**], they used the Poor Man's data augmentation algorithm and an asymptotic data augmentation algorithm for censored regression data. Furthermore, Pan and Connett [**88**] have extended these results to clustered censored data in semi-parametric linear regression models. Ageel [**3**] considers using an estimated value of $E[T_i|T_i > C_i]$ as a pseudo-value to replace the ith censored observation using both an empirical distribution and the Weibull distribution. In the empirical approach,

$$E[T_i|T_i > C_i] \approx \sum_{j=1, j\neq i}^{n} \{T_j|T_j \geq C_i \text{ and } T_j \text{ is an uncensored datum } \} / n_{ui}$$

where n_{ui} is the number of events in the set

$$\{T_j|T_j \geq C_i \text{ and } T_j \text{ is uncensored, } j = 1, ..., n, j \neq i\}.$$

In using the Weibull distribution approach, the continuous random variable T is assumed to follow a Weibull distribution, $Weibull(\alpha, \lambda)$, and then

$$E[T_i|T_i > C_i] = \alpha^{-1} exp\left[(\alpha C_i)^{\lambda}\right] \Gamma\left(1 + \frac{1}{\lambda}\right) \left[1 - I\left((\alpha C_i)^{\lambda}, 1 + \frac{1}{\lambda}\right)\right]$$

where Γ is the gamma function and I the cumulative distribution function of a gamma distribution. Also, Chiou [**20**] proposes another method where the censored observation is imputed by Median$[T_i \mid T_i > C_i]$. Cantor [**16**] presents a method of imputation for right censored survival data in which the possibility that an individual with a censored survival time is cured is explicitly accommodated. For each right censored observation, first a cure/non-cure indicator is imputed and then for the non-cured a survival time is imputed based on a Gompertz model. Lue et al. [**75**] addresses the issue of reducing the dimension of predictors in survival regression without requiring a prespecified parametric model. They replace each right censored survival time with its conditional expectation by using Buckley and James [**13**] pseudo-random variables $Y_i = T_i\delta_i + E[T_i \mid T_i > C_i](1 - \delta_i)$. Jackson et al. [**58**]

impute failure times for censored participants from the entire sample of observed failure times that are greater than their corresponding censoring times in order to quantify the sensitivity of the conclusions from fitted Cox proportional hazards models. Taylor et al. [**101**] described non-parametric multiple imputation methods, including risk set imputation and Kaplan-Meier imputation, to handle missing event times for censored observations in the context of non-parametric survival estimation and testing. Faucett, Schenker and Taylor [**37**] and also Hsu in collaboration with Taylor and other colleagues [**56**] used auxiliary variables to recover information from censored observations by using Markov chain Monte Carlo methods to impute event times for censored cases. Hsu and Taylor [**54**] extended this work on estimation using auxiliary variables to adjust, via multiple imputation, for dependent censoring in the comparison of two survival distributions. Hsu, Taylor and Hu [**55**] further adapted their previous method to the situation where the event and censoring times follow accelerated failure time models.

The approach proposed in this work is similar to that proposed by Royston [**92**][**93**] in respect to using imputations of censored observations in *visualising tools* of survival in time to event studies. In Royston's approach, each censored survival time is imputed by assuming a lognormal distribution for the unobserved actual failure time. Royston's method is inherently parametric and is described in detail in Section 4.4.

4.4 Parametric Approach

Royston [**92**][**93**] shows the practical use of the lognormal distribution for imputation in prognostic models for time-to-event in datasets from breast and ovarian cancer. This method introduces an appropriate amount of random variation to simulate realistic individual survival times. Values of the censored survival time are imputed by substituting values randomly sampled from a lognormal distribution by drawing a value for each patient who was censored conditional on their survival to at least the point of censoring. The mean and standard deviation of the distribution are estimated taking into account the values of prognostic factors for each patient.

The survivor function of the lognormal distribution of a random variable T with location parameter μ and variance σ^2 is given by

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

where $\Phi(.)$ is the standard normal distribution function. Typically, μ is modelled as a linear function of covariates $\mathbf{x}$, therefore the regression model for the log-survival times can be written as

$$\ln t = \beta_0 + \boldsymbol{\beta}^T \mathbf{x} + \varepsilon\sigma$$

where $\varepsilon \sim N(0,1)$. Parameters β_0 and $\boldsymbol{\beta}$ can be estimated from a sample of n_U uncensored and $n - n_U$ censored observations by maximising the following log-likelihood function as described in Section 2.4:

$$\ln L = \sum_{i=1}^{n_U} \ln f(t_i; \beta_0, \boldsymbol{\beta}) + \sum_{i=n_U+1}^{n} \ln S(t_i; \beta_0, \boldsymbol{\beta})$$

Let m be the expected log-survival time for a given patient with a real survival time of T, which depends on the patient's prognostic factors. Once a lognormal model has been estimated, the expected value $m_i = E(\ln T_i | \mathbf{x}_i)$ is available for all $i = 1, ..., n$. Assume s to be the residual standard deviation of the log-survival times, that is, the standard deviation of $(\ln T - m)$ over the sample. The lognormal model specfies that $\left(\frac{\ln T - m}{s}\right)$ has a standard normal distribution. For censored observations, the actual survival time, which is unknown, always exceeds the censored time. Assume that τ_i is the unobserved true survival time related to a censored observation c_i. The model assumes that $\left(\frac{\ln \tau_i - m_i}{s}\right)$ has a standard normal distribution, but taking account of the known censoring time c_i we have to consider the right-hand tail of a normal distribution truncated at $\left(\frac{\ln c_i - m_i}{s}\right)$.
In order to impute τ_i stochastically, a random draw u_i is sampled from a standard normal truncated at $k_i = \left(\frac{\ln c_i - m_i}{s}\right)$. Obtaining u_i is straightforward. Assume that u has a standard normal distribution truncated at k, the distribution of u is effectively an $N(0,1)$ with all observations less than k discarded. this is a simple, but wasteful, approach to simulation. Alternatively, the cumulative distribution

function of the truncated distribution can be expressed as:

$$\Phi_k(u) = \frac{\Phi(u) - \Phi(k)}{1 - \Phi(k)}, \quad u \geq k \tag{4.1}$$

where $\Phi(u)$ is the cdf of the standard normal. This can now be used in a standard inverse cdf simulation approach, where to draw a random observation from the truncated distribution we simply take $u_i = \Phi_k^{-1}(p_i)$, with p_i as a realisation of the $uniform(0, 1)$ distribution. Thus, in order to get the value of u_i :

$$\begin{aligned} p_i = \Phi_k(u_i) &= \frac{\Phi(u_i) - \Phi(k)}{1 - \Phi(k)} \\ \Rightarrow \quad \Phi(u_i) &= p_i((1 - \Phi(k)) + \Phi(k) \\ \Rightarrow \quad u_i &= \Phi^{-1}\left(p_i(1 - \Phi(k)) + \Phi(k)\right), \end{aligned}$$

which corresponds to an inversion method for the right-hand tail, truncated at k, of a normal distribution. After sampling a random draw u_i from a standard normal truncated at k_i the pseudo sample $\{t_i^+\}$ is defined as follows:

$$t_i^+ = \begin{cases} t_i, & \text{if } i \leq n_U. \\ \exp(m_i + u_i s), & \text{otherwise, i.e. censored.} \end{cases} \tag{4.2}$$

The process is repeated for all of the individuals with censored times, giving one complete imputation of the censored observations. The uncensored survival times are not altered, as in 4.2. The related R code can be found in the Appendix. In the case of a correctly assumed model, the $\{t_i^+\}$ should act as a sample of uncensored observations from the same distribution as the dataset. The imputation can be repeated several times to test the sensitivity.

This single imputation approach has some limitations. With a high degree of censoring there is uncertainty in the real shape of the survival distribution and producing sensible imputations becomes harder, relying on a small set of observed survival times that may all be relatively short and thus containing only limited information on the underlying survival distribution. Therefore, estimates of extreme survival times can be obtained, including possibly values that extend beyond the human lifetime. Accordingly, Royston recommended that it might be unwise to use this

method when more than approximately 50% of the data are censored, since it is unacceptable to get a reliable feel for the shape of the survival distribution when more than half of the data are being imputed. Although the imputation method that is described by Royston helps to gain an impression of the survival times of an individual, it does not follow that a parametric survival model such as lognormal distribution will necessarily be a satisfactory model for the data. Additionally, even if the lognormal, or any other supposed distribution, were correct the true μ and σ are unknown and their point estimates, based on a regression model, are used, which could lead to underestimating the variability.

To overcome this problem, we propose a Bayesian paradigm as an alternative approach to impute censored observations, which naturally incorporates the uncertainty about the parameters of the distribution.

4.5 A Parametric Bayesian Approach

In the previous section, the Royston method for imputing censored observation is reviewed where a parametric model is assumed for imputing censored observations. In this section a parametric Bayesian approach is introduced as another option to impute censored observations. The introduction of the Bayesian framework for time to event data is explained in Section 2.8. For convenience, conjugate priors for the parameters of the distribution are chosen to give a computationally convenient posterior. Also, non-informative priors could be a good choice for the hyperpriors in situations where there is little prior information about parameters.

By applying MCMC methods [**76**] we obtain simulated draws for the predicted values of the censored observations, conditional on the observed censoring times. We can then use these predicted values as imputed values to give complete datasets, as in standard multiple imputation methods. Standard graphics can then be used to explore many aspects; such as treatment effects and hazard functions, with some indication of the uncertainty due to the censoring and uncertainty in the fitted model. Censoring in survival analysis is a key feature in defining the likelihood and for obtaining a posterior sample. However, censoring plays no role in determining a prior

or in interpreting the results once we have our posterior and predictive sample.

In order to impute censored observations, first, a specific distribution needs to be considered for the survival times. To represent the imputation method, we consider a general parametric model for the data that depends on a parameter θ, say $f(t|\theta)$. Also $p(\theta)$ is assumed as a prior distribution for θ, which is unconditional on any observed data. By using Bayes theorem, the prior distribution $p(\theta)$ is updated to a posterior distribution $p(\theta|D)$.

Second, we need to find the density for the survival time conditional on a (right) censoring time, because it is evident that the imputed value for the censored observation needs to be bigger than its censoring time. Taking c as the value of the censored observation the conditional density is defined as:

$$f(t \mid T \geq c, \theta) = \frac{f(t \mid \theta)}{S(c \mid \theta)}, \quad t \geq c \tag{4.3}$$

and the conditional expected value of the censored observation is:

$$E(T \mid T \geq c) = \frac{\int_c^\infty t f(t \mid \theta) dt}{S(c \mid \theta)} \tag{4.4}$$

Moreover, by using $S(t \mid \theta)$, the survivor function of survival times, the conditional median t_{med}, of a censored observation can be found by solving $S(T|T \geq c, \theta) = 0.5$ as follows:

$$\begin{aligned} 0.5 = S(T \mid T \geq c, \theta) &= \frac{P(T \geq t_{med} \mid \theta)}{P(T \geq c \mid \theta)} \\ \Rightarrow \frac{S(t_{med} \mid \theta)}{S(c \mid \theta)} &= 0.5 \\ \Rightarrow t_{med} = S^{-1}\left(0.5 \times S(c \mid \theta)\right) & \end{aligned} \tag{4.5}$$

Finally, samples from a predictive distribution conditional on the observed censoring time are generated using the following density:

$$f(t \mid T \geq c, D) = \int f(t \mid T \geq c, \theta) p(\theta \mid D) d\theta \tag{4.6}$$

Suppose we have a posterior density $p(\theta|D)$ and that we can generate a random sample from this distribution, say $\theta^1, \theta^2, ..., \theta^s$ where θ^s is the s-th component of the

sample. Then to sample t^k, observations from the predictive conditional distribution, first of all, θ^k are sampled from the posterior density $p(\theta \mid D)$, then t^k are generated from the density $f(t \mid T \geq c, \theta^k)$. The sampled t^k is a draw from the predictive distribution conditional on the observed censoring time. This is done repeatedly and independently to obtain a Monte Carlo sample $\{t^k : k = 1, ..., s\}$. The mean of the predictive distribution can be numerically approximated by $\sum_{k=1}^{s} t^k/s$, and the median of the predictive distribution by the median of the sample.

Here the WinBUGS software is used to impute censored observations. WinBUGS is a menu-driven Windows program that generates samples from the posterior distribution. Samples are generated in WinBugs using the probabilistic idea of a Markov chain with a stationary distribution that corresponds to the desired posterior distribution. WinBUGS requires that the data be entered in a different form from the $D = \{y, \delta\}$ described above, here NA is used to indicate missing (censored) data and so we transform a (y, δ) observation into a form (t, c) where for $i = 1, ..., n$

$$t_i = \begin{cases} y_i, & \text{if } \delta_i = 1 \\ NA, & \text{if } \delta_i = 0 \end{cases} \qquad c_i = \begin{cases} 0, & \text{if } \delta_i = 1 \\ y_i, & \text{if } \delta_i = 0 \end{cases}$$

These define separate data vectors for censored and uncensored observations. In t NA is used for each y_i that corresponds to censored observation and the vector c has the censoring times for all censored observations.

To indicate the censored time in the model, a general distribution is defined as $t \sim ddist(theta)I(a, b)$ where a and b give information about censoring. If there is no available data for a specific t (i.e it is an NA), then it is assumed that the observation is from the specified distribution but censored in the interval $a < t < b$. $I(a,)$ is used to define right-censoring and $I(, b)$ for left-censoring. When WinBUGS recognises the NA for an individual, it knows that the data for that individual is censored and generates an appropriate term for the likelihood, in the case of right-censoring $S(c; \theta)$. When there is an actual value for t, it neglects any information in $I(a, b)$ and generates the appropriate term for the likelihood, $f(t; \theta)$.

As an example, if the Weibull model is assumed for some right-censored survival data, the model can be written in WinBugs as follows:

```
model{
#likelihood
for(i in 1:n){t[i]~dweib(alpha,lambda)I(c[i],)}
#prior
alpha~dlnorm(a,b)
lambda~dgamma(c,d)
}
#initial
list(n=100, a=0.001, b=0.001, c=0.001, d=0.001)
#Data
t[]  c[]
1     0
NA    2.5
# more rows of data would be listed here
END
```

For easy access in R, instead of using Winbugs directly the `R2WinBUGS` package [**98**] can be used. In `R2WinBUGS` package, the `bugs` function takes data and initial values as input and writes a WinBUGS script, calls the model, and saves the simulations in R.

```
bugs(data, inits, parameters.to.save, model.file="model.bug",
n.chains="", n.iter="", n.burnin="",n.thin="",
 n.sims = "", bin="", debug=FALSE, DIC=TRUE,
 digits="", codaPkg=FALSE,bugs.directory="",
 program=c("WinBUGS"), working.directory=NULL, clearWD=FALSE,
  useWINE=FALSE, WINE=NULL,  newWINE=TRUE,
  WINEPATH=NULL, bugs.seed="", summary.only=FALSE,
save.history=TRUE, over.relax = FALSE)
```

In the `bugs` function, the `data` and initial values, `inits`, which provides starting

values for each unknown parameter can be saved in a text file the same format as they are used in WinBUGS in the working directory. `Parameters.to.save` is a vector of the names of parameters of interest that need to be saved. Also, `model.file` is the file containing the model written in WinBUGS code saved as text file. `n.chains` shows the number of parallel Markov chains to be used with a default of 3 and `n.iter` is the number of iterations per chain, which also includes burn-in. `n.burnin` is the length of burn-in which determines the number of iterations to be removed from the start, so that from then on the chain is, hopefully, in a stationary state and providing realizations from the required posterior distributions. `n.thin` shows the thinning rate, which is the way to improve the efficiency of the posterior sample whereby only every ν the value from Gibbs sampler is actually retained for inference, this breaks the autocorrelation of subsequent draws from the chain. `n.sim` is the approximate number of simulations to be kept after thinning. The address of the directory containing WinBUGS needs to be specified with the `bugs.directory` argument.

4.5.1 Illustration

For illustration, the Weibull distribution is considered for the survival times. Weibull distribution features, including its survivor and hazard function, are explained in detail in Section 2.5. The Weibull distribution has a different form of hazard function depending on the value of the shape parameter and this feature makes it a good candidate distribution for time to event data. In order to have a posterior distribution from the same family, conjugate priors are preferred for the parameters. For a Weibull distribution, if the shape parameter is known, a gamma prior is conjugate for the scale parameter. Although when the shape parameter is unknown, no joint prior can be placed on (α, λ) such that it leads to a recognisable, analytically tractable joint posterior distribution, therefore, we rely on simulations. Here, the Lognormal distribution is assumed as a prior for the shape of the Weibull distribution, and the gamma distribution is assumed as a prior for the scale parameter [**21**]. The properties of the lognormal distribution and gamma distributions are explained in Section 2.5. The model can be summarised in the following hierarchical

specification:

$$t_i \sim Weibull(\alpha, \lambda)$$
$$\alpha \sim lognormal(a, b)$$
$$\lambda \sim gamma(d, e)$$

where a,b,d and e are hyper-parameters. Using MCMC methods we can obtain simulated draws for the predicted values of the censored observations, conditional on the observed censoring times.
For the Weibull distribution using the WinBugs definition, $f(t)$ and $S(t)$ are expressed as follows:

$$f(t) = \alpha\lambda t^{\alpha-1} e^{-\lambda t^{\alpha}}$$
$$S(t) = e^{-\lambda t^{\alpha}}$$

Assuming c to be the value of a censored observation and by using the conditional density formula in (4.3), the density function of a survival time conditionally on being censored at time c is:

$$f(t \mid T \geq c) = \frac{\alpha\lambda t^{\alpha-1} e^{-\lambda t^{\alpha}}}{e^{-\lambda c^{\alpha}}}, \quad t \geq c \tag{4.7}$$

Also, the conditional expected value of a censored observation based on a Weibull distribution is given by:

$$\begin{aligned} E(T \mid T \geq c) &= \frac{\int_c^{\infty} t\alpha\lambda t^{\alpha-1} e^{-\lambda t^{\alpha}} dt}{S(c)} \\ &= \frac{1}{S(c)} \left[-te^{-\lambda t^{\alpha}}\right]_c^{\infty} - \frac{1}{S(c)} \int_c^{\infty} -e^{-\lambda t^{\alpha}} dt \\ &= \frac{cS(c)}{S(c)} + \frac{1}{S(c)} \int_c^{\infty} -e^{-\lambda t^{\alpha}} dt \\ &= c + \frac{1}{S(c)} \int_c^{\infty} S(t) dt \\ &= c + \frac{1}{e^{-\lambda c^{\alpha}}} \int_c^{\infty} e^{-\lambda t^{\alpha}} dt \end{aligned} \tag{4.8}$$

So, in general, when the density comes from the Weibull family of distributions, to obtain conditional expected values the following integral needs to be evaluated:

$$\int_c^\infty e^{-\lambda t^\alpha} dt \tag{4.9}$$

Since (4.9) does not have a closed analytic form, we need to resort to a numerical procedure such as the `integrate` function in R.

If $\alpha = 1$ the Weibull distribution is the same as the exponential distribution with survivor function $S(t) = e^{-\lambda t}$. So the conditional expected value of a censored observation is then:

$$E(T \mid T \geq c) = c + \frac{1}{e^{-\lambda t}} \int_c^\infty e^{-\lambda t} dt = c + \frac{1}{\lambda} = c + E(T)$$

where $E(T)$ is the unconditional expectation of T. However, for other values of α the `integrate` function in R needs to be used.

To check the performance of `integrate` for obtaining these conditional expected we consider another special case with $\alpha = 2$ where:

$$\begin{aligned}
&\int_c^\infty e^{-\lambda t^2} dt = \int_c^\infty e^{-2\lambda \frac{t^2}{2}} dt \\
&= \sqrt{\frac{2\pi}{2\lambda}} \frac{1}{\sqrt{\frac{2\pi}{2\lambda}}} \int_c^\infty e^{-\frac{t^2}{2(\frac{1}{2\lambda})}} dt \\
&= \sqrt{\frac{\pi}{\lambda}} \left[1 - \Phi(c\sqrt{2\lambda})\right]
\end{aligned} \tag{4.10}$$

So when $\alpha = 2$ we have an expression in terms of the normal cdf Φ :

$$E(T \mid T \geq c) = c + \frac{1}{e^{-\lambda c^2}} \sqrt{\frac{\pi}{\lambda}} \left[1 - \Phi(c\sqrt{2\lambda})\right] \tag{4.11}$$

Table 4.1 shows the result of solving (4.9) using (4.11) and also using integration in R for $\lambda = 2$ for different values of c.

c	Using (4.11)	Using `integrate`
0	0.6266571	0.6266571
1	1.2106846	1.2106846
2	2.1183262	2.1183262

Table 4.1: $E(T \mid T \geq c)$ for a $Weibull(2, 2)$ distribution.

This gives some confidence that we can use `integrate` to evaluate (4.8) for a general Weibull distribution, and indeed a similar approach could be taken for other parametric survival distrubutions.

Finding the conditional median, t_{med} of a censored observation based on a Weibull distribution reduces to solving $S(T \mid T \geq c) = 0.5$ where c is the censoring time so

$$\begin{aligned} S(T|T \geq c) &= \frac{P(T \geq t_{med})}{P(T \geq c)} = 0.5 \\ \Rightarrow \quad & \frac{S(t_{med})}{S(c)} = \frac{e^{-\lambda t_{med}^{\alpha}}}{e^{-\lambda c^{\alpha}}} = 0.5 \\ \Rightarrow \quad & exp\left\{-\lambda(t_{med}^{\alpha} - c^{\alpha})\right\} = 0.5 \\ \Rightarrow \quad & t_{med}^{\alpha} - c^{\alpha} = \frac{ln(0.5)}{-\lambda} \\ \Rightarrow \quad & t_{med} = \sqrt[\alpha]{\frac{-ln(0.5)}{\lambda} + c^{\alpha}} \end{aligned} \tag{4.12}$$

using WinBugs, samples from the predictive distribution conditional on the observed censoring times are generated based on (4.6). So, in order to check these predicted values provided with WinBugs, a Weibull distribution $Weibull(\alpha = 1, \lambda = 2)$ (i.e. exponential) is assumed for the survival times together with ten percent censoring of the data. Two hundred data are generated from $Weibull(\alpha = 1, \lambda = 2)$ where there is nineteen censored observations in the dataset. The mean of random draws from WinBugs are compared to the explicitly calculated values from (4.8). We write $t_1^*, ..., t_B^*$ for Winbugs random draws from the predictive distribution $f(t \mid T \geq c, D)$ for a specific censored value. We also denote $\theta_1^*, ..., \theta_B^*$ as the posterior values of the Weibull parameters (α, λ) where $\theta_j^* = (\alpha_j^*, \lambda_j^*)$ for $j = 1, ..., B$. Based on the Weibull distribution $E(T|T \geq c) = c + \frac{1}{e^{-\lambda c^{\alpha}}} \int_c^{\infty} e^{-\lambda t^{\alpha}} dt$, so for each posterior parameter value we define

$$I(\theta_j^*) = c + \frac{1}{e^{-\lambda_j^* c^{\alpha_j^*}}} \int_c^{\infty} e^{-\lambda_j^* t^{\alpha_j^*}} dt \tag{4.13}$$

In order to check the WinBugs results, we need to compare $A = \dfrac{\sum_j I(\theta_j^*)}{n_1}$ and $C = \dfrac{\sum_j t_j^*}{n_2}$ where n_1 is the number of random draws of posterior values θ_j^* which is assumed as 100 and n_2 is the number of random draws from the predictive distribution $f(t \mid T \geq c, D)$. We also calculate standard deviations of the random draws $\sigma_1 = sd(I(\theta_j^*))$ and $\sigma_2 = sd(t_j^*)$ based on 100 draws. Then in order to realistically compare the target means, A and C, the standard deviations of these means need to be the same. Taking $n_1 = 100$, an appropriate n_2 can be calculated as follows

$$\begin{aligned} \frac{\sigma_1}{\sqrt{n_1}} &= \frac{\sigma_2}{\sqrt{n_2}} \\ \Rightarrow \frac{\sigma_1}{\sqrt{100}} &= \frac{\sigma_2}{\sqrt{n_2}} \\ \Rightarrow n_2 &= \left(\frac{10\sigma_2}{\sigma_1}\right)^2 \end{aligned}$$

So for each censored value, we found a comparable sample size n_2 and based on that determined C. The results are shown in Table 4.2.

From Table 4.2, after equalising the standard errors of the means, the differences of means are less than 0.05, so it can be concluded that the results from imputing censored observations using WinBugs are in agreement with what we would theoretically expect.

To illustrate how well the parametric Bayesian imputation is working, one hundred event times were generated from a Weibull distribution ($Weibull(\alpha = 2, \lambda = 4)$) with twenty percent of them taken to be censored. In the particular realised example below there are twenty two censored observations. By taking the censored observations as incomplete data and using the posterior predictive distribution as in (2.22) conditional on the observed censored time, simulated draws for the predicted values of the censored observations were produced using WinBugs. After imputing censored values and combining them with the observed failure times the complete dataset can be formed. As a result of having a completed dataset due to the imputed values of censored observations, the survival data transform to the standard dataset and standard graphics, such as a histogram, boxplot, or density plot, can be drawn. As the data are simulated and then the censoring is applied, here we know the real value of the data before the censoring and in what follows we refer to this as *true-*

censored observation	censored value	σ_1	σ_2	n_2	A	C	A-C
1	0.257	0.178	2.348	17400	2.405	2.389	0.01
2	3.997	0.269	2.268	7108	6.082	6.054	0.02
3	1.395	0.215	2.24	10854	3.512	3.543	−0.03
4	4.626	0.278	1.841	4385	6.706	6.708	−0.002
5	1.09	0.206	2.019	9605	3.213	3.189	0.02
6	0.008	0.169	2.162	16365	2.171	2.176	−0.005
7	5.462	0.290	1.78	3767	7.536	7.547	−0.01
8	3.735	0.264	1.615	3742	5.822	5.795	0.02
9	3.39	0.258	1.812	4932	5.48	5.47	0.01
10	1.013	0.204	2.051	10108	3.137	3.122	0.01
11	0.480	0.186	2.215	14181	2.619	2.585	0.03
12	6.481	0.302	2.175	5186	8.549	8.509	0.04
13	1.518	0.218	2.079	9094	3.632	3.621	0.01
14	2.959	0.251	2.008	6400	5.053	5.05	0.003
15	0.913	0.201	2.43	14615	3.04	3.031	0.009
16	0.889	0.201	2.215	12143	3.017	2.978	0.03
17	7.346	0.312	1.911	3751	9.409	9.441	−0.03
18	3.577	0.262	2.317	7820	5.665	5.671	−0.006
19	0.791	0.197	1.861	8924	2.921	2.95	-0.02

Table 4.2: Check of the predicted values provided using WinBUGS based on 200 observations generated from a $Weibull(\alpha = 1, \lambda = 2)$ with ten percent censoring.

failure. Therefore, in this artificial situation the imputed values can be compared to the true original data. In Figure 4.1, a Kaplan-Meier plot of the data after imputation is compared to the Kaplan-Meier plot of the true data, the Kaplan-Meier plot of the full data including censoring (the usual plot in practice) and also the Kaplan-Meier plot of the data when the censored observations are omitted from the dataset (deletion).

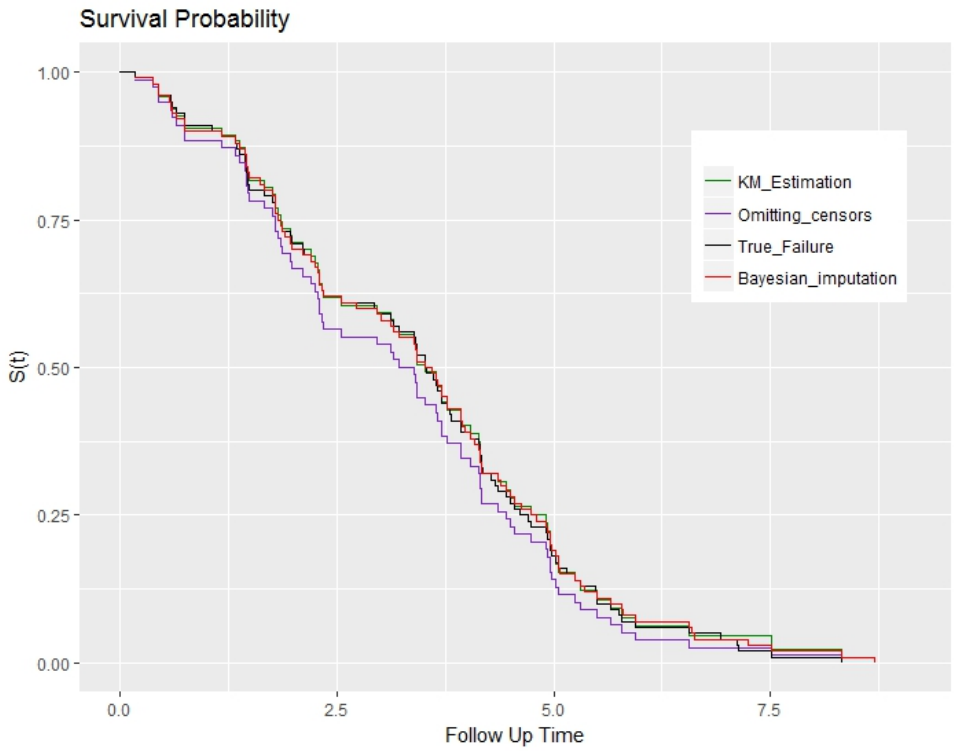

Figure 4.1: The Kaplan-Meier survivor plots: for the full data with censoring; the true failure times; omitting censored observations; and using parametric Bayesian imputation for the censored values.

Based on Figure 4.1, the Bayesian imputation method for censored observations is not only in agreement with the Kaplan-Meier estimate of the data but also in agreement with the Kaplan-Meier estimate of true-failure times with the additional benefit of being able to plot an estimate of the underlying failure time distribution. As discussed at the start of this chapter, one of the ways of dealing with censored observations is to delete them from the dataset, however based on Figure 4.1 it is apparent that by omitting censored observations the results become biased.

Rather than using only one single imputation as in Figure 4.1, we imputed the censored observations multiple times using parametric Bayesian approach. Figure 4.2 compares the Kaplan-Meier estimation of the censored data with the Kaplan-Meier estimates from one hundred imputations using the parametric Bayesian approach. Figure 4.2 shows that the Kaplan-Meier plots for all of the one hundred imputations are around the Kaplan-Meier estimate of the data, and as is expected, the mean of these one hundred Kaplan-Meier plots is near to the Kaplan-Meier of the data before imputation.

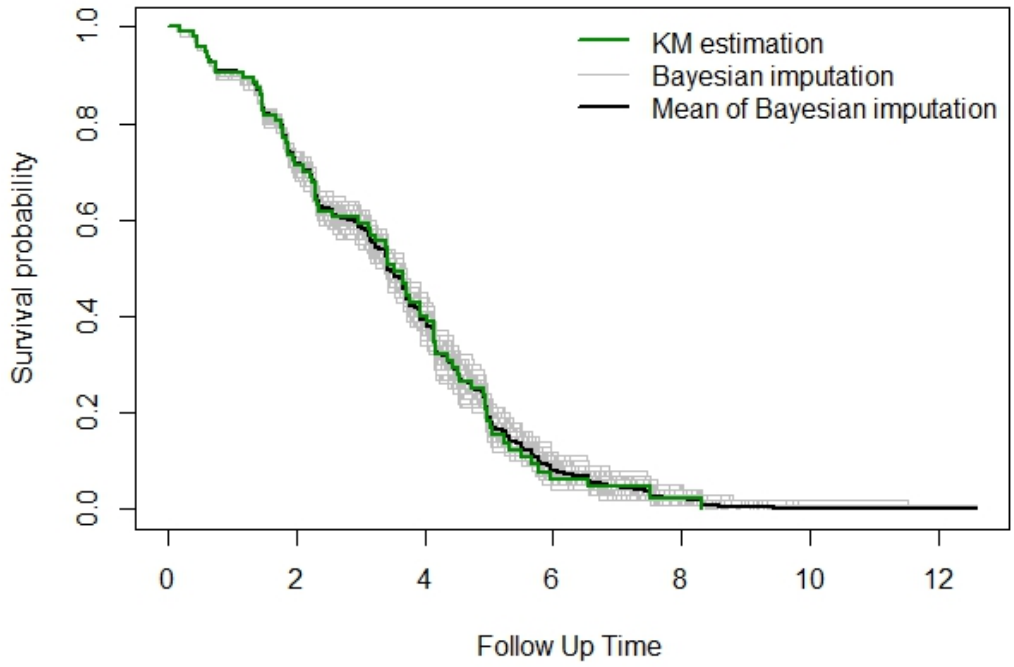

Figure 4.2: Kaplan-Meier plots of the censored data compared with the mean of the Kaplan-Meier estimates from one hundred imputations using a parametric Bayesian approach.

Figure 4.3 shows the boxplot for thirty imputed datasets comparing them to the true-failure times and the dataset omitting the censored values. It can be seen that the ranges of most of the imputation boxplots are near the range of true-failure times. Also, medians are near to the true median of the data. This Figure shows that the distribution of data may not be sensitive to different imputations, however some extreme values may be imputed, as noted by Royston [**93**].

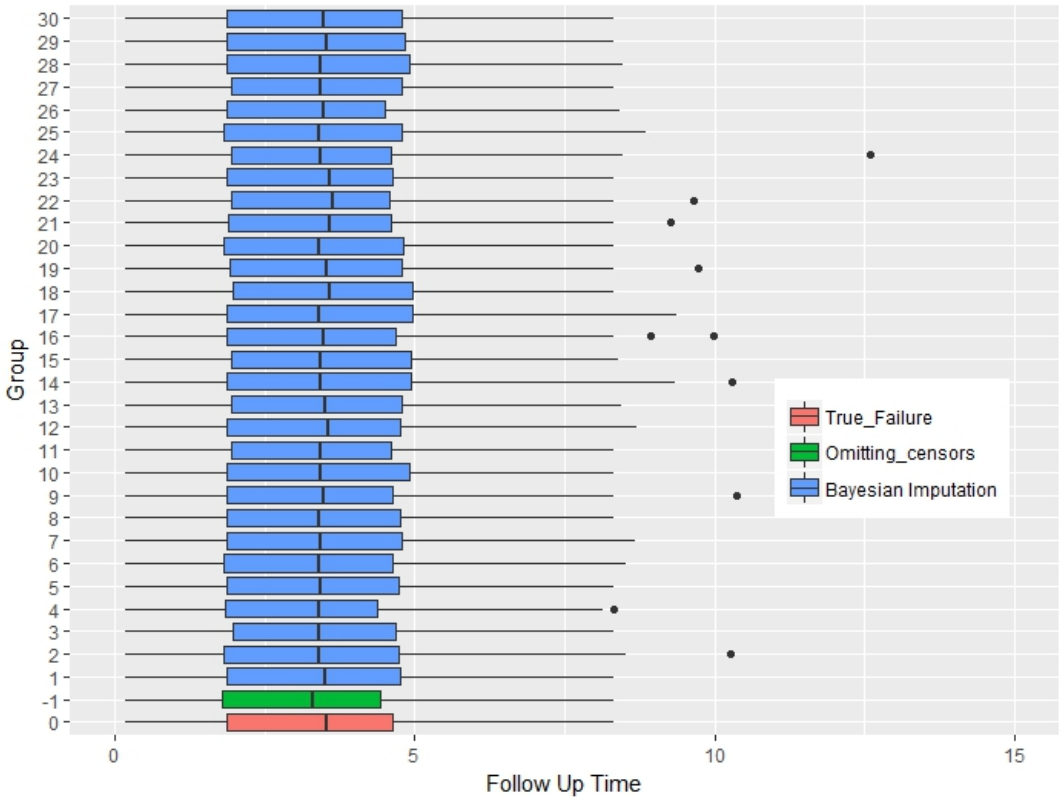

Figure 4.3: Boxplots of the true failure times and when omitting censored observations are compared to the 30 draws of parametric Bayesian imputations.

One goal in this thesis is to impute the censored observations to enable plots of the density of the complete dataset as an additional visualizing tool to complement the common Kaplan-Meier plot. Figure 4.4 compares the density of the true-failure data to the density of the imputed dataset based on one single imputation. It is shown that the density plots are close to each other most of the time. The `logspline` function [**66**] in R is used to plot these densities and will be described in Section 6.2.

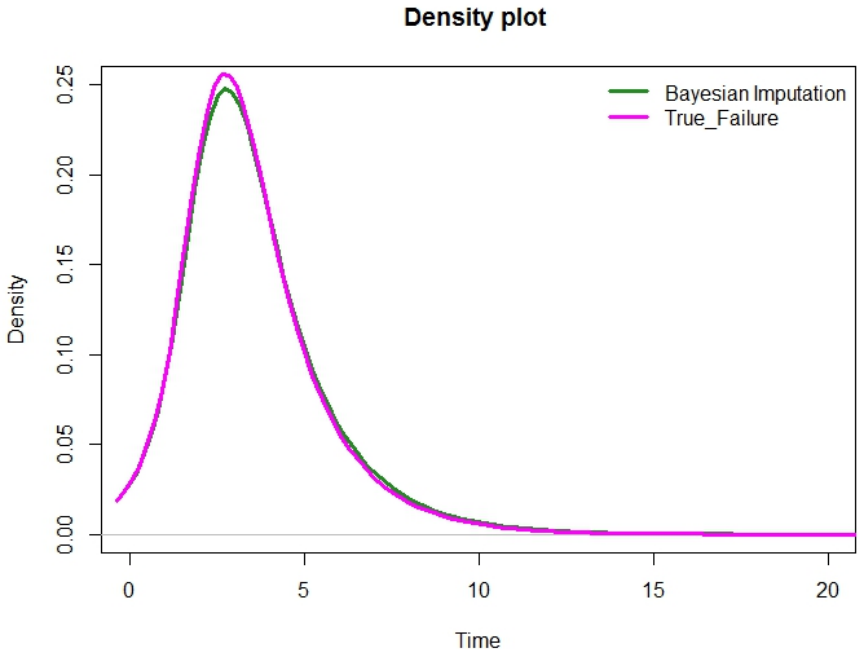

Figure 4.4: Comparing the density of the true-failure data to the density of the imputed dataset using a parametric Bayesian approach based on one single imputation.

The parametric Bayesian approach which is discussed in this section could be extended to other survival distributions and other Bayesian survival models by changing the assumed distribution of the data and also the prior distributions. In the case that there is no recognisable and analytically tractable posterior, then, as above, MCMC methods can be used to draw from the posterior and predictive distributions.

In the next section, a non-parametric Bayesian approach is introduced as a second method for imputing censored observations.

4.6 Non-parametric Bayesian Approach

A potential problem in parametric Bayesian approach is that a fixed specification of distributional properties for the error terms in a model may be inadequate for the real data. Under the non-parametric Bayesian paradigm the unknown distribution of the model is treated as a random parameter with stochastic non-parametric

priors, such as the Dirichlet process. The non-parametric Bayesian methods are explained in detail in Chapter 3.

This section describes using a non-parametric Bayesian framework to present a flexible approach to impute censored observations using predictive distributions. Recently, non-parametric Bayesian models in survival analysis have become popular because of the advances in computing technology and improvement of the efficient computational algorithm. In non-parametric Bayesian inference the typical approach is to specify a prior distribution over the space of all possible cumulative distribution functions $F(t) = 1 - S(t)$ [**57**]. The Dirichlet process $DP(\alpha, G_0)$, which is described in Section 3.4, is perhaps the most celebrated and popular prior process in non-parametric Bayesian inference. In a Dirichlet process G_0 is a specific distribution on Θ and is referred to as the base distribution while α is a precision parameter.

Early work on the Dirichlet process in the context of survival analysis dates back to the work of Susarla and Van Ryzin [**99**] and also Ferguson and Phadia [**40**]. Susarla and Van Ryzin [**99**] in the case of right censoring, derive the Bayes estimator of the survivor function under the Dirichlet process prior and obtained a closed form for the mean and other moments of the posterior distribution. The Bayes estimator of $S(t)$ is obtained under squared error loss :

$$L(\hat{S}, S) = \int_0^\infty (\hat{S}(t) - S(t))^2 dw(t)$$

where w is a weight function (which is nonnegative and nondecreasing on $(0, \infty)$) and $\hat{S}(t)$ is an estimator of $S(t)$. Suppose there are n observations, $y_1, ..., y_n$, Let $y_1, ..., y_k$ denote the uncensored observations and $y_{k+1}, y_{k+2}, ..., y_n$ indicate the censored observations, also let $y_{(k+1)}, y_{(k+2)}, ..., y_{(m)}$ denote the distinct observed times among the censored observations $y_{k+1}, y_{k+2}, ..., y_n$. Write λ_j to represent the number of censored observations that are equal to $y_{(j)}$, for $j = k+1, k+2, ..., m$ and $N(t)$ and $N^+(t)$ for the number of observations (censored or not censored) greater than or equal to t and the number greater than t, respectively. The Bayes estimator of

$\hat{S}(u)$ under square error loss is given by

$$\hat{S}(u) = \frac{\alpha(1 - G_0(u)) + N^+(u)}{\alpha + n} \times \prod_{j=k+1}^{l} \left(\frac{(\alpha(1 - G_0(y_{(j)}))) + N(y_{(j)})}{\alpha(1 - G_0(y_{(j)})) + N(y_{(j)}) - \lambda_j} \right) \tag{4.14}$$

in the interval $y_{(j)} \leq u \leq y_{(l+1)}$, $l = k, k+1, ..., m$, with $y_{(k)} = 0$ and $y_{(m+1)} = \infty$. The estimator (4.14) is based on computing the conditional first moment of the survival probability $S(u) = 1 - F(u)$ given (y_i, δ_i), $i = 1, ..., n$ where F is distributed as a Dirichlet process.

It is shown by Susarla and Van Ryzin [**99**], that the Kaplan-Meier estimator of $S(u)$ is a limiting case of (4.14) which is gained when $G_0 \longrightarrow 1$. To show the relation between the Kaplan-Meier estimator defined in (2.16) and the Bayes estimator defined in (4.14), requires that the number of events e_j in an interval be as small as unity. In other words we take the intervals sufficiently short and numerous to only have one death in each interval. So assuming the n observed y values are labelled as $0 \leq y\acute{}_1 \leq y\acute{}_2 \leq ... \leq y\acute{}_n$ in increasing magnitude, based on the Kaplan-Meier estimator, the product limit of estimator of $S(u)$ is defined as [**62**]

$$\hat{S}(u) = 1 - \hat{F}(u) = \prod_r \frac{(n - r)}{(n - r + 1)} \tag{4.15}$$

where r are values such that $y\acute{}_r \leq u$ and $y\acute{}_r$ is an uncensored observation. As $G_0 \longrightarrow 1$, the (4.14) converges to

$$\frac{N^+(u)}{n} \times \prod_{j=k+1}^{l} \left(\frac{N(y_{(j)})}{N(y_{(j)}) - \lambda_j} \right) \tag{4.16}$$

Assume $i(u)$ is the largest integer to have $\acute{y}_{(i(u))} \leq u$. Therefore

$$\frac{N^+(u)}{n} = \prod_{j \leq i(u)} \frac{N^+(y\acute{}_j)}{N(y\acute{}_j)} \tag{4.17}$$

By replacing (4.17) into (4.16) and cancelling the ratios common to both the products, which are the ratios related to the censored observations, (4.16) reduces to

$$\prod_j \frac{N^+(y_j')}{N(y_j')} \tag{4.18}$$

where the product is taken over those j that $y_j' \leq u$ and y_j' is an uncensored observation. This is exactly (4.15) provided that the uncensored observations are distinct. For calculating the Kaplan-Meier estimator in (4.15), actual censored observations do not need to be known and only the number of censored observations between two uncensored observations is enough. However, both censored and uncensored observations need to be known in order to calculate the Bayes estimator in (4.14). Hence, the Bayes estimate uses all of the data and might be preferable to the Kaplan-Meier estimator. The larger the value of α, the smoother the Bayes estimator becomes in comparison to Kaplan-Meier estimator, as the jumps at the event points are smaller [**57**].

Susarla and Van Ryzin [**99**] also obtained the posterior p^{th} moment of $S(u)$ as

$$\begin{aligned} E(S(u)^p|D) = &\prod_{s=0}^{p-1}\left(\frac{G_0(u,\infty)+s+N^+(u)}{G_0(R^+)+s+n}\right) \\ &\times \prod_{j=k+1}^{l}\left(\frac{G_0(y_{(j)},\infty)+s+N(y_{(j)})}{G_0(y_{(j)},\infty)+s+N(y_{(j)})-\lambda_j}\right) \end{aligned} \tag{4.19}$$

where $G_0(u,\infty) = \alpha(1-G_0(u))$ and $D = (y,\delta)$. It could be seen that when $p = 1$ the (4.19) reduces to (4.14).

Although the Dirichlet process prior is a simple and computationally tractable prior for an unknown distribution, it produces distributions that are discrete with probability one, making it unsuitable for density modelling. To overcome this problem, the distribution can be convolved with some continuous kernel, or more generally, by using a Dirichlet process to define a mixture distribution with infinitely many components, of some simple parametric form. A Dirichlet Process Mixture (DPM) model is a mixture with a parametric kernel and a random mixing distribution modelled with a DP prior. The Dirichlet process Mixture model is described in Section 3.6 (for more detail see Ferguson [**38**], [**39**], Antoniak [**6**]).

Doss [**30**], Doss and Huffer [**31**] and Doss and Narasimhan [**32**] discussed using mixtures of Dirichlet process priors for $F(t) = 1 - S(t)$ with a Gibbs sampler in the presence of right censored data. A non-parametric Bayesian method based on

mixtures of Dirichlet priors gives a reasonable compromise between purely non-parametric and purely parametric models. In the presence of censored data, there is no closed form for the posterior distribution of F given the data. Therefore, Monte Carlo methods need to be used.

De Iorio [**27**] proposed an unconstrained model for survival regression based on a DP prior that allows for the introduction of covariates in an interpretable manner for right censored data. Assume $t_1, ..., t_n$ are n independent and identically distributed survival times, where t_i is the survival times for individual i, and δ_i is the associated censoring indicator, and also $\mathbf{x}_i$ is a p-dimensional vector of categorical and continuous covariates for individual i. Let $k(.|\mu, \sigma^2)$ denote an arbitrary family of location-scale parametric kernel densities. For each individual with covariate $\mathbf{x}_i$ a mixture model for the data is defined as:

$$f_{\mathbf{x}_i=x}(t_i; G) = \int k(t_i|\mu, \sigma^2) G_{\mathbf{x}}(d\mu\sigma^2)$$
$$G_{\mathbf{x}} \sim DP(\alpha, G_{0\,\mathbf{x}}) \tag{4.20}$$

The choice of an appropriate kernel density depends on the underlying sample space. As in survival analysis the sample space defined on the positive half-line, mixtures of gamma, log extreme value (Weibull), log-logistic, or lognormal distributions may be appropriate. The choice of lognormal kernel is equivalent to using a normal kernel based on log-transformed data. In comparison to kernels like log-logistic and Weibull, the lognormal kernel is more convenient in practice [**74**] and computationally easier due to conjugacy of the base measure G_0 and the kernel. In this case, any MCMC scheme for DP mixture models can be used. The MCMC algorithm for implementing posteriors are described in detail in Section 3.7 also a modified version that accounts for the presence of censored observations in the data can be found in Web Appendix of De Iorio [**27**].

The `LDDPsurvival` function in the `DPpackage` in R can be used for this approach to survival modelling and here it is used to impute the censored observations in a non-parametric manner. This package was developed by Jara [**60**] for the implementation of some non-parametric Bayesian and semiparametric models in R. The `LDDPsurvival` function for fitting survival models has the following call:

```
LDDPsurvival(formula,zpred,prior,mcmc,state,status,grid,
data=sys.frame(sys.parent()),na.action=na.fail,work.dir=NULL)
```

where `formula` is a two-sided linear formula object describing the model to be fitted, with the response on the left of a $\sim$ operator and the terms, separated by `+` operators, on the right. The resulting design matrix is used to model the distribution of the response in the LDPP mixture of normals model. The response matrix for right-censored data is a two-column matrix such that for each failure time the two columns are the same, equal to the value of the failure, and for censored observation, the first column is filled with the censored value and the second column should be written as -999 as it is an unknown limit. Constructing the design matrix in R can be done as follows:

```
for (i in 1:n){
if (data$status[i]==1){
data$left[i]=data$time[i]
data$right[i]=data$time[i]
}
if (data$status[i]==0){
data$left[i]=data$time[i]
data$right[i]= -999
}
}
```

The model which is used in the `LDDPsurvival` function is based on the model defined by De Iorio et al. [**27**]. To understand the model specified by the `LDDPsurvival` function, for survival times $t_i, i = 1, ..., n$ the full model can be written in the

hierarchical form:

$$\begin{aligned}
&\log(t_i) \mid \beta, \sigma^2 \sim K_N(. \mid \beta, \sigma^2) \\
&\beta, \sigma^2 \mid G \sim G \\
&G \mid \alpha, G_0 \sim DP(\alpha, G_0)
\end{aligned}$$

by specifying a conditionally conjugate base measure

$$G_0 = N(\beta \mid \mu_b, s_b) Gamma(\sigma^2 \mid \tau_1/2, \tau_2/2) \tag{4.21}$$

with conjugate hyperpriors

$$\begin{aligned}
&\alpha \mid a_0, b_0 \sim Gamma(a_0, b_0) \\
&\mu_b \mid m_0, s_0 \sim N(m_0, s_0) \\
&S_b \mid \nu, \psi \sim IW_q(\nu, \psi) \\
&\tau_2 \mid \tau_{s_1}, \tau_{s_2} \sim Gamma\left(\frac{\tau_{s_1}}{2}, \frac{\tau_{s_2}}{2}\right)
\end{aligned}$$

where the inverted Wishart distribution (IW) is parametrized so that

$$A \sim IW_q(\nu, \psi) \Rightarrow E(A) = \frac{\psi^{-1}}{\nu - q - 1}$$

The computational implementation used in `DPpackage` is based on the MCMC method described in [**77**] and [**84**].

The posterior feature of greatest interest to us here is the predictive distribution, which is used to impute the censored observations. Take ϕ to represent the entire set of parameters, then the posterior predictive distribution for an observation t_{new} is given by

$$p(t_{new}|T \geq c, data) = \iint k(t_{new}|T \geq c, \beta_{new}, \sigma_{new}) p(\beta_{new}, \sigma_{new}|\phi) p(\phi|data) \tag{4.22}$$

Samples from the predictive distribution conditional on the observed censoring time are generated using (4.22). Specifically, ϕ^k are sampled from the posterior density $p(\phi|data)$, then conditional on these values (β^k, σ^k) are generated from the

density $p(\beta_{new}, \sigma_{new}|\phi^k)$ and finally t^k are sampled from the conditional density $k(t_{new}|T \geq c, \beta^k, \sigma^k)$. In this way the sampled t^k is a draw from the predictive distribution conditional on the observed censoring time. This is done repeatedly and independently to obtain a Monte Carlo sample $\{t^k : k = 1, ..., s\}$. More details on the posterior predictive distribution are given in Section 3.8.

4.6.1 Illustration

To illustrate the effectiveness of the non-parametric Bayesian imputation approach, we generate one hundred event times from a Weibull distribution, $(Weibull(\alpha = 2, \lambda = 4))$ with twenty percent censoring. In this illustrative realization there are in fact seventeen censored observations. Simulated draws for the imputed values of the censored observations are produced using `DPpackage`. Because the data are first simulated and then the censoring applied, the real value of the sample data before the censoring is known and again referred to as *true-failure*. Therefore, the imputed values can be compared to the *true* unobserved. In Figure 4.5, the Kaplan-Meier plot of the data after imputation is compared to the Kaplan-Meier of the true-failure data, the standard Kaplan-Meier plot of the censored data, and also a Kaplan-Meier plot of the data when the censored observations are omitted (deletion).

Based on Figure 4.5, the non-parametric Bayesian imputation method for censored observations is in agreement with the Kaplan-Meier estimate of the data and also in agreement with the Kaplan-Meier estimate of the true-failure times. It is also again apparent that by omitting censored observations the results become biased.

It is of course possible to repeat the imputation process and Figure 4.6 compares the Kaplan-Meier plot of the censored data with the Kaplan-Meier estimates of one hundred imputations using the non-parametric Bayesian approach. As can be seen, the Kaplan-Meier plots for all of the one hundred imputations are around the Kaplan-Meier estimate of the data, and the mean of these imputed Kaplan-Meier plots is close to the Kaplan-Meier of the data.

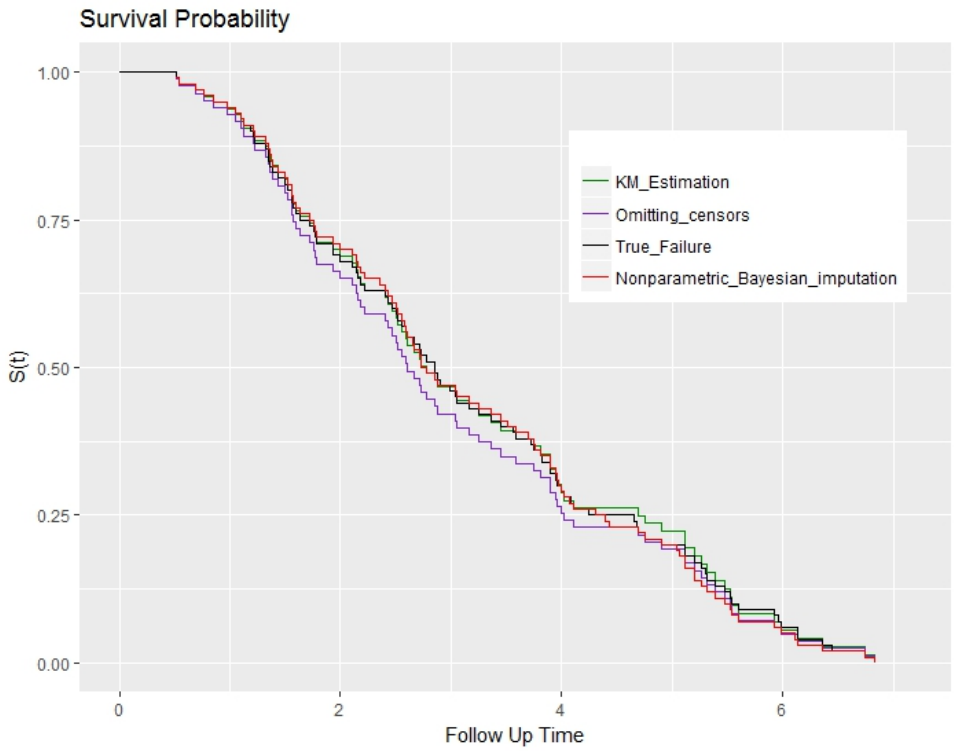

Figure 4.5: Kaplan-Meier survivor plots: for the full data with censoring; the true failure times; omitting censored observations; and using the non-parametric Bayesian imputation for the censored values.

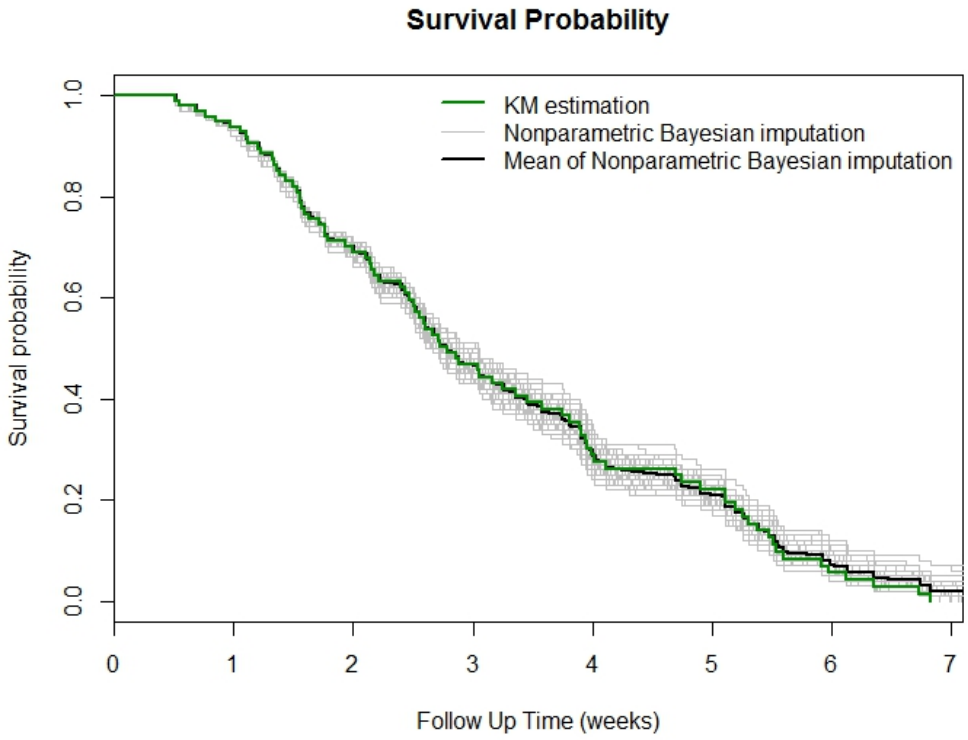

Figure 4.6: Kaplan-Meier plots of the censored data compared with the mean of the Kaplan-Meier estimates from one hundred imputations using a non-parametric Bayesian approach.

Figure 4.7 shows a boxplot for thirty of the imputed datasets comparing them to the true-failure times and the data omitting censored values. It can be seen that medians of the imputation boxplots are close to the true median of the data, however some extreme values may be imputed.

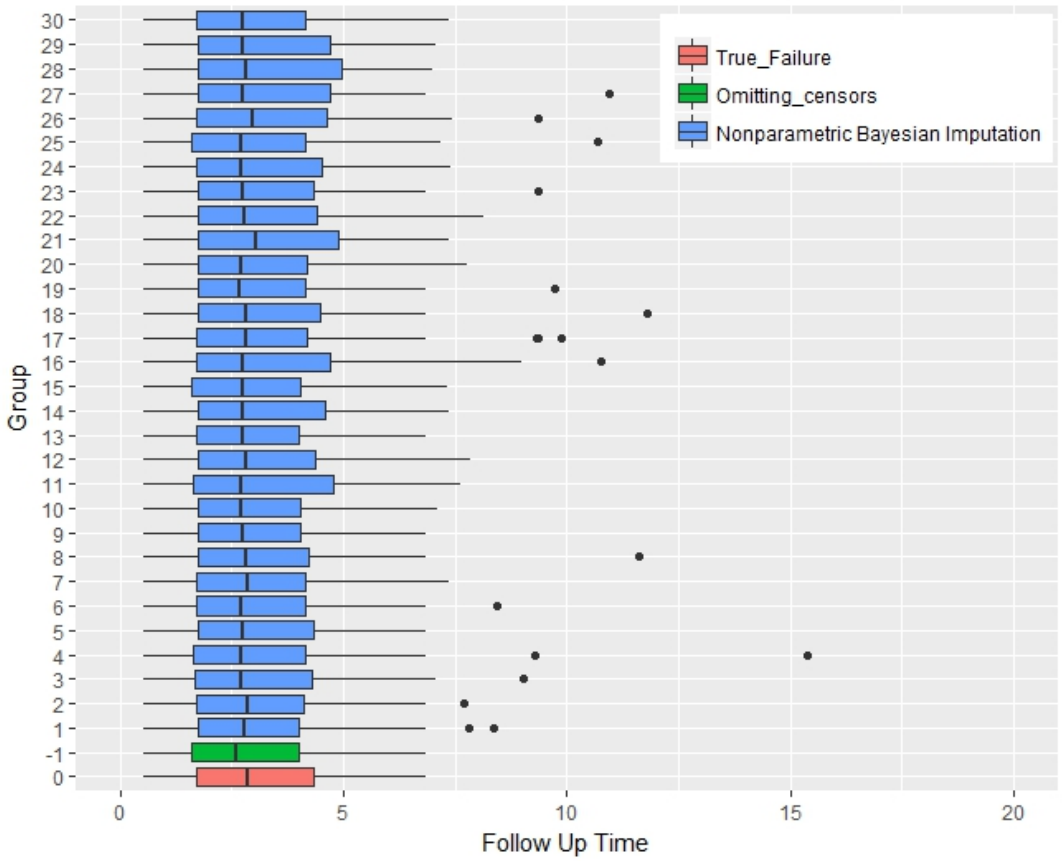

Figure 4.7: Boxplots of the true failure times and when omitting censored observations compared to 30 sets based on non-parametric Bayesian imputations.

Finally, Figure 4.8 compares the estimated density (using the `logspline` function [**66**] in R) of the true-failure data to that of the imputed dataset based on one single imputation. It is clear that the density plots are very close to each other most of the time T.

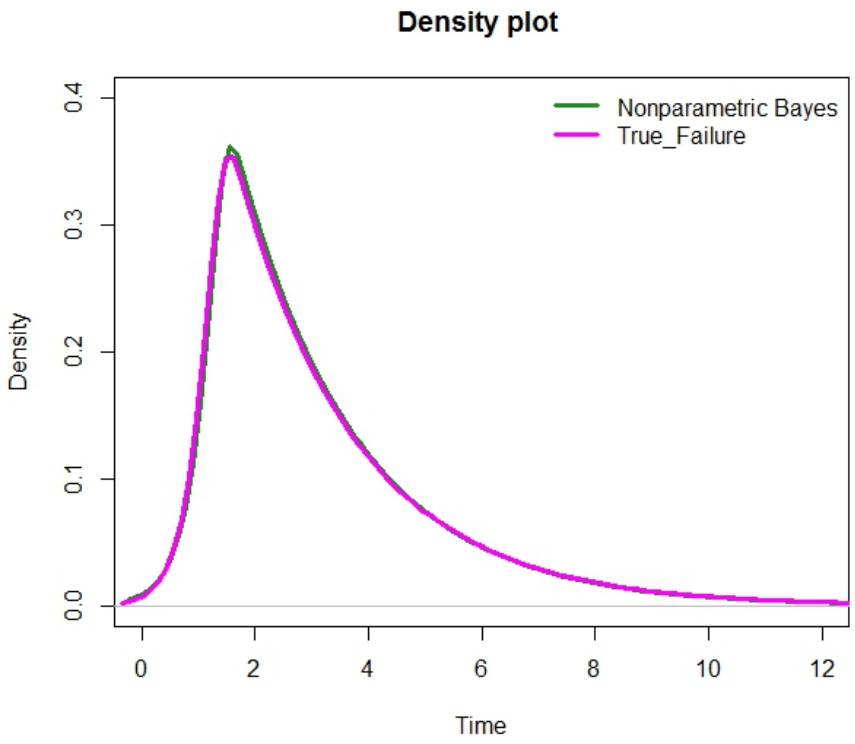

Figure 4.8: Comparison of density estimates of the true-failure data to the density of the imputed dataset using non-parametric Bayesian approach based on one single imputation.

In the next section the parametric approach is compared to the parametric Bayesian and non-parametric Bayesian approaches.

4.7 Comparing Different Imputation Methods

We now compare the Royston parametric approach, described in Section 4.4, with the parametric and non-parametric Bayesian methods that were described in Sections 4.5 and 4.6 using results based on a single imputation. This comparison is divided into two parts. First, a correctly specified distribution of the failure time data is used in imputation approaches, while in the second part the censored observations are imputed based on a mis-specified model.

4.7.1 Assuming a Correct Model

In this part, one dataset is generated from a specific distribution with an assumed desired percentage of censoring. Then based on our knowledge of the true distribution of the data, censored observations are imputed. As Royston's method assumed a lognormal distribution to impute censored observations, we generate one hundred data from a lognormal ($lognormal(1, 0.5)$) distribution with a twenty percent censoring rate. In our example, there are twenty-two censored observations from the 100 data values. The censored observations are then imputed based on Royston's approach, a non-parametric Bayesian method, and a parametric Bayesian approach where a lognormal distribution is assumed for the survival data. As we know the true failure times, these approaches could be compared not only with each other, but also to the true failure times.

Figure 4.9 compare the survivor plot across the different methods. As can be seen, for all of the methods the survivor plot is close to the survivor plot of true failure time.

Figure 4.10 compares this same information in a boxplot for the different imputation methods and the true data. It is apparent that all of the three methods are near to the true failure times, however the median of the parametric Bayesian approach is closer to the median of the true failure, which is perhaps not surprising as it is based on the correct model and also incorporates the added uncertainty in the imputation process that is missing in Royston's method.

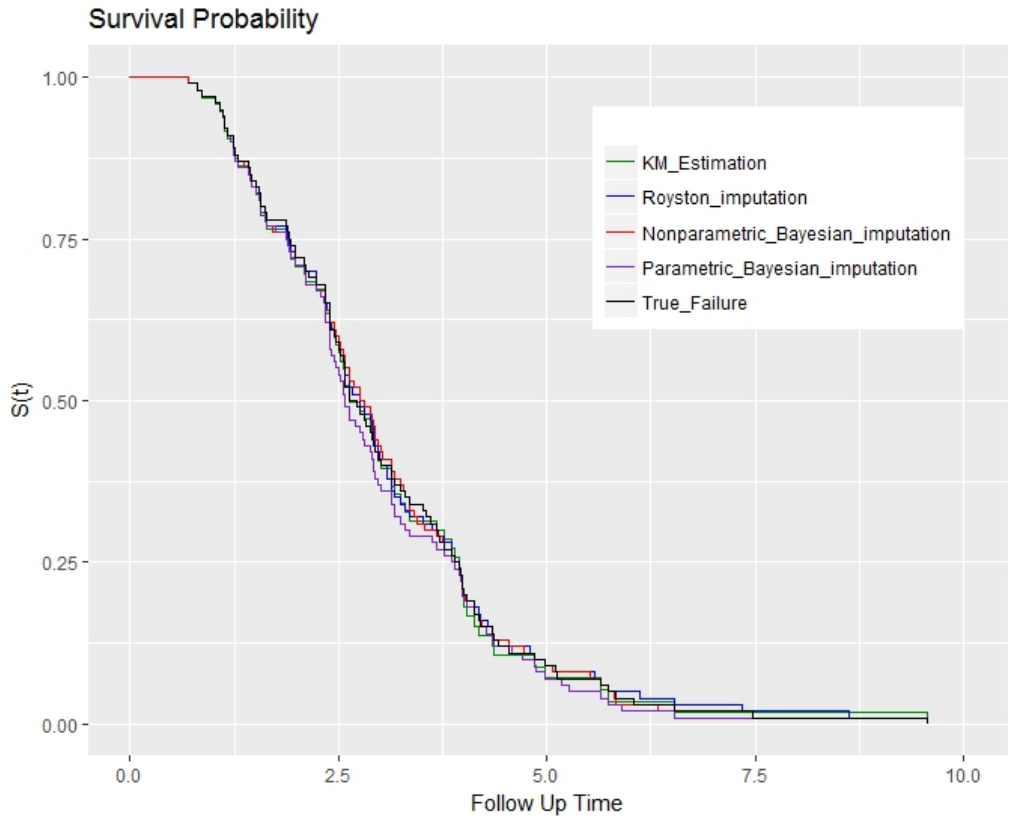

Figure 4.9: The survivor plot for Kaplan Meier estimates of the data, true failure times, Royston imputation, parametric Bayesian imputation, and non-parametric Bayesian imputation.

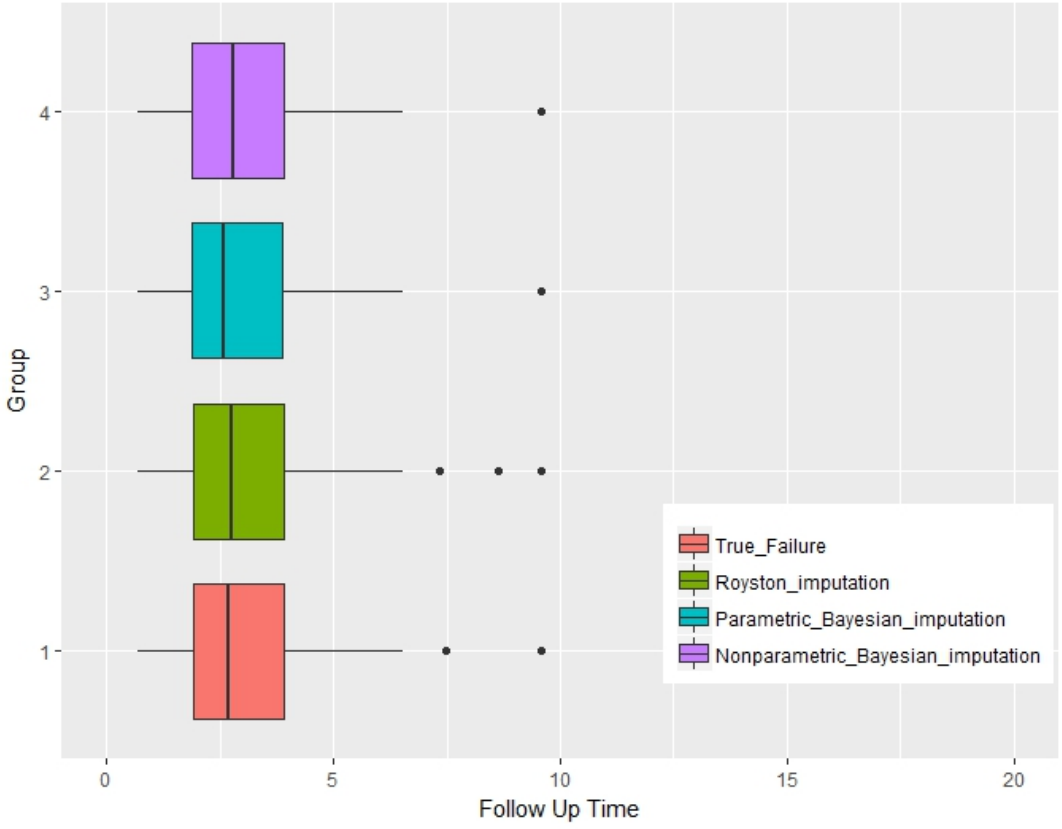

Figure 4.10: Boxplots for data generated from $lognormal(1, 0.5)$ imputing censored observations using parametric, parametric Bayes, and non-parametric Bayesian approaches.

Figure 4.11 compares the density estimates of the true-failure data to those of the imputed datasets based on one single imputation using parametric, parametric Bayesian and non-parametric Bayesian approaches. It shows that the density plots of different imputation methods are near each other most of the time and they are also near to the density of true failure times. However among all methods, the density of parametric Bayesian approach is much closer to the density of true failure time, possibly for the reasons that we noted above.

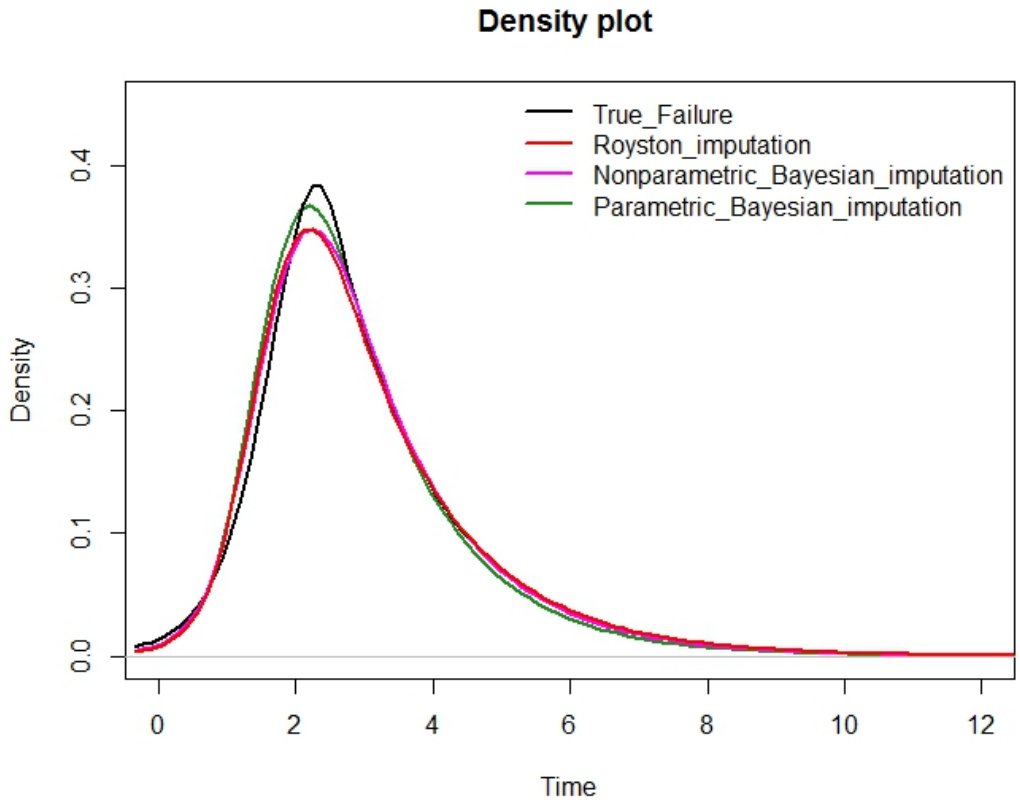

Figure 4.11: Comparing the density estimates of the true-failure times to the imputed datasets using parametric, parametric Bayesian, and non-parametric Bayesian approaches to impute the censored observations.

Based on Figure 4.9, Figure 4.10 and Figure 4.11 when we use the assumption of the correct model in imputing the censored observations all parametric, parametric Bayesian and non-parametric Bayesian methods impute the censored observations around the true failure times. Although it should be noted that the above is just based on one single imputation and only provides us with some visual reassurance and is not sufficient to make firm conclusions.

4.7.2 Mis-specified model

In the previous part, the true assumption about the density used to generate the data is used in different methods of imputation. However, in reality, this distribution of the data may not be known. In this part we consider the effect of mis-specification of this model.

For illustration, one hundred data values are generated from a Weibull ($Weibull(2, 4)$) distribution with twenty percent censoring. In our specific example, there happen to be twenty censored observations among the 100 data. In order to consider the mis-specification of the distribution of the data, we assume that they were generating from a lognormal distribution. The censored observations are then imputed based on the Royston approach, which assumes a lognormal distribution. They are also imputed using a parametric Bayesian approach where a lognormal distribution is assumed for the survival distribution. Finally, the non-parametric Bayesian method is used which makes no specific assumption for the distribution of the data. Again as we know the true failure times from the simulation, these approaches can be compared to the true failure times.

Figure 4.12 compares the survivor plot over the different methods based on one single imputation. It is apparent that the survivor plot using Royston's parametric approach and the Bayesian parametric approach are both far from the survivor plot of the true failure data. Based on one single imputation, among all these three methods the survivor plot of the non-parametric Bayesian approach is much closer to the survivor plot of the true failure times.

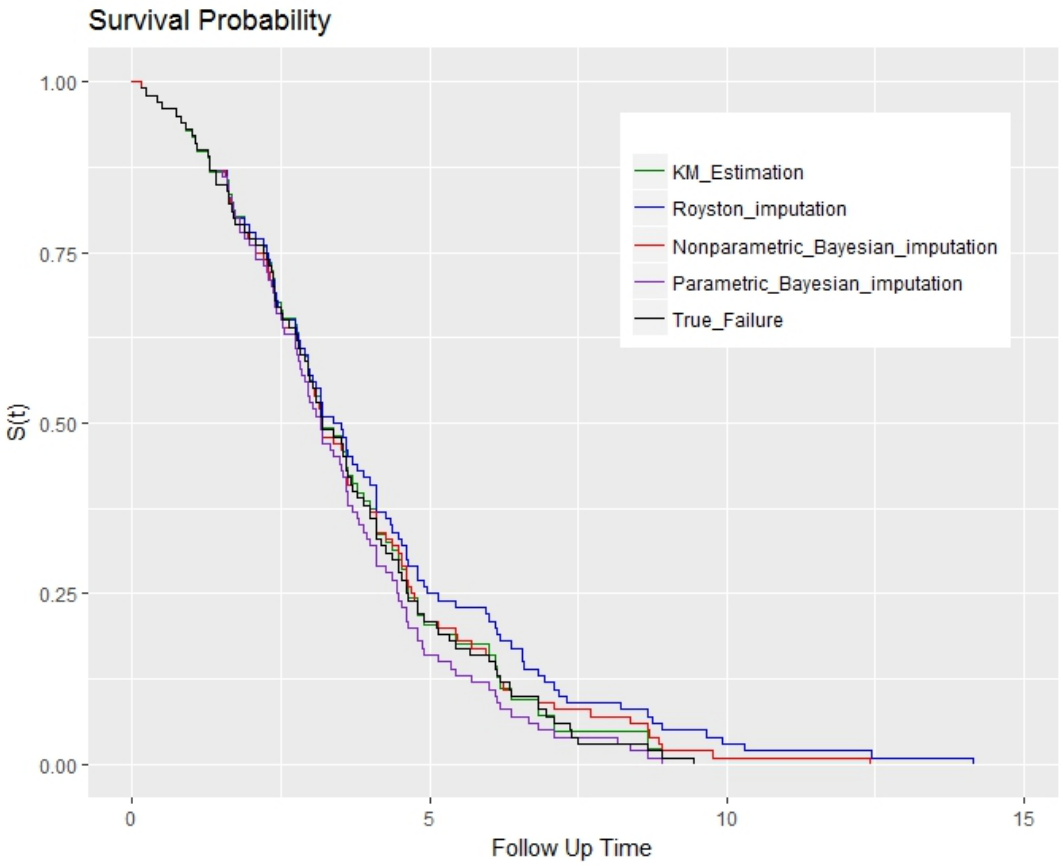

Figure 4.12: The survivor plot for Kaplan Meier estimates of the data, true failure times, Royston imputation, parametric Bayesian imputation and non-parametric Bayesian imputation.

Figure 4.13 compares the boxplots for the different imputation methods to that of the true data. It can be seen that the median and the range of the parametric Bayesian approach and non-parametric Bayesian approach are near to the median and the range of the true failure time. However, some extreme values are generated using the non-parametric Bayesian approach. Compared to the parametric Bayesian and non-parametric Bayesian approaches, the median and the range in the Royston method are greater than the median and the range of true failure data and it also exhibits some extreme imputed values.

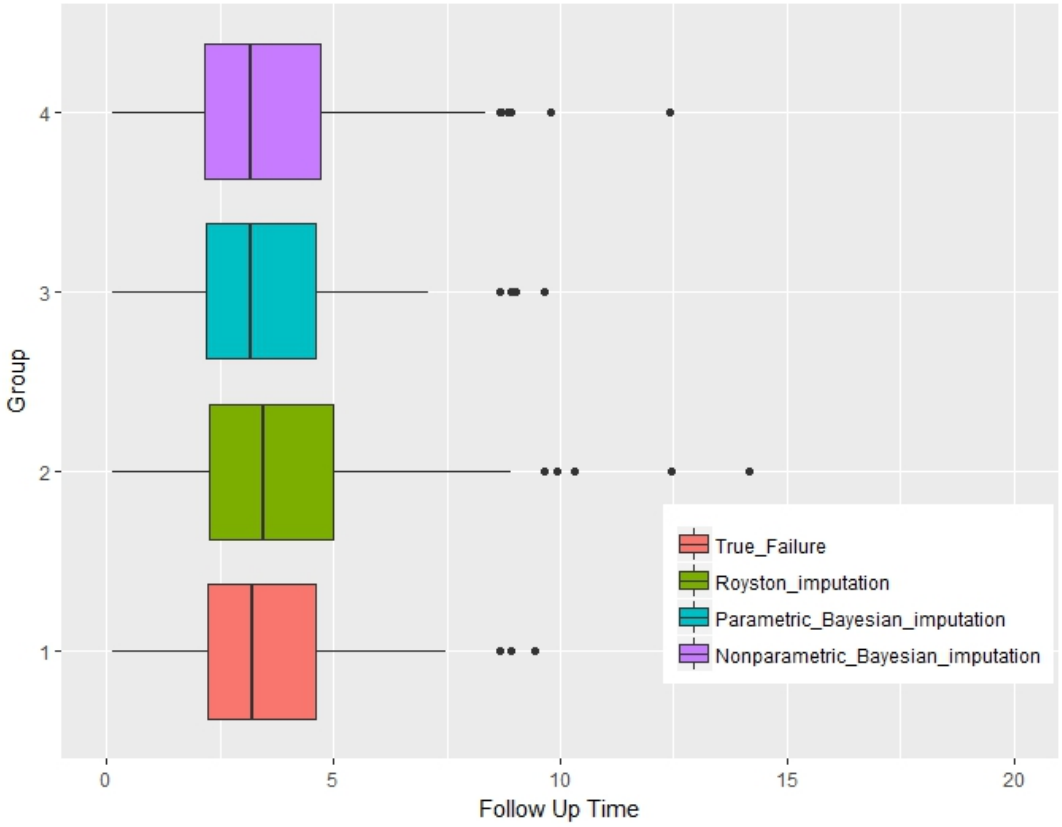

Figure 4.13: Boxplots for data generated from a $Weibull(2, 4)$ with censored observations imputed using parametric, parametric Bayes and non-parametric Bayesian approaches and compare to the true failure times.

Finally Figure 4.14 compares the density estimates of the true-failure data to the density of the imputed dataset based on one single imputation using parametric, parametric Bayesian and non-parametric Bayesian approaches in this mis-specified setting. It seems that the density plots of non-parametric imputation method is close to the density of true failure time across most of the range. The Royston method seems to perform worst with a marked right-skewness, while parametric Bayesian method seems to underestimate the right-hand tail, although is closer to the true density.

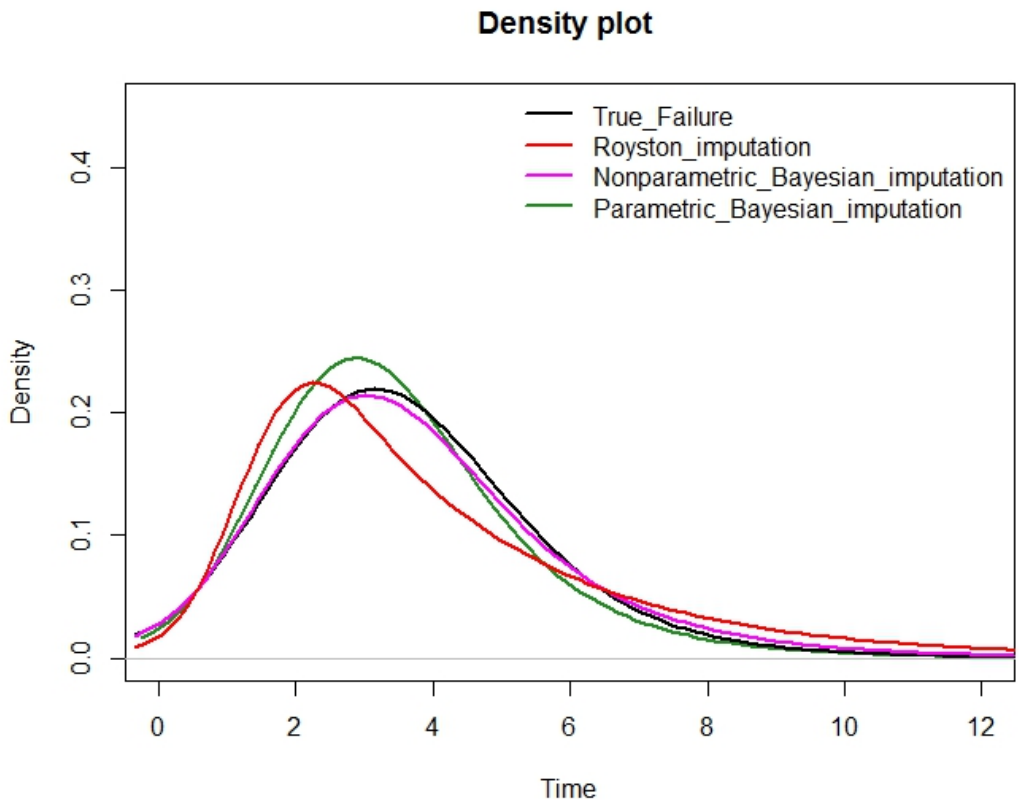

Figure 4.14: Comparing the density plots of the true-failure to the imputed datasets including parametric, parametric Bayesian and non-parametric Bayesian approaches.

According to Figure 4.12, Figure 4.13 and Figure 4.14 which are based on one single imputation when we mis-specified the assumption of the model in imputing the censored observations, the non-parametric Bayesian approach impute the censored observations closer to the true failure times.

4.8 Chapter Summary

In this chapter, we provide an overview on imputing methods for missing data as an introductory concept for the rest of the chapter where we try to treat the censored observations as missing values and used different methods of imputation. By imputing the censored observations, it is possible to present the actual distributions to give simple and interpretable graphical displays for physicians and patients.

Two new approaches have been proposed for imputing censored observations, including a parametric Bayesian approach and a non-parametric Bayesian method. These methods are compared to the Royston parametric approach for imputing censored observations. Based on our preliminary test of concept results, generally limited to one single imputation, when the true distribution of the data is used to impute the censored observations all three of these methods provide imputed censored observation near the true failure values. However when we use a mis-specified model, the imputation results based on non-parametric Bayesian approach are closer to the true failure values. This suggests that when, as in practice, the true distribution is unknown, there may be some advantage to using the more flexible non-parametric Bayesian method and that it might also be considered as a useful diagnostic for studying robustness to model assumptions in the parametric approaches. In the next chapter, we explore these comparisons in more detail through a simulation study. The results of the simulation study are discussed and specifically address aspects such as the effect of differing percentages of censoring and different forms of survival distribution, including situations with decreasing, increasing and constant hazard functions.

Chapter 5

Simulation Study

5.1 Introduction

The primary purpose of this simulation study is to understand if the proposed imputation methods create plausible complete (imputed) datasets. Simulating data sets needs an assumed distribution for the data and full specification of the specified parameters. The simulated data sets should have some correspondence to reality so that any results can be generalized to real life situations. We will compare approaches based on using parametric and non-parametric Bayesian models for survival data simulated from some known survival functions. The benefit of using simulated data, especially for examples that include censoring, is that we know not only the true underlying density function, which the data was generated from, but also the real value of the sample data before the censoring was applied. The study considers different percentages of random right censoring and also situations with decreasing, increasing and constant hazard functions.

This chapter is arranged as follows: we start with a review to the methods of generating censored observations in Section 5.2, followed in Section 5.3 where the chosen method of generating censored observations in this thesis is described. In Section 5.4 the simulation methodology is reviewed. In Section 5.5 the results of parametric Bayesian imputation are reported and in Sections 5.6 the non-parametric Bayesian imputations are used as a second approach to imputation. Finally, in Section 5.7 the

parametric Bayesian method and non-parametric Bayesian approaches are compared for a specific sample size.

5.2 Methods of Generating Censored Observations

Simulating survival data in comparison to other types of data requires specific consideration. To simulate censored survival data, two survival distributions are required, one for representing the censored mechanism and another for the uncensored survival times that would be observed if the follow-up time had been sufficiently long enough to reach the event [**14**]. The observation times, Y_i, incorporating both events and censored observations are calculated for both cases by combining the uncensored survival times, T_i, and the censoring times, C_i. If the uncensored survival time for an observation is less than or equal to the censoring time, then the event is considered to be uncensored and the observation time equals the uncensored survival time, otherwise, the event is considered censored and the observation time equals the censoring time. In other words, the observed times can be defined as $Y_i = \min(C_i, T_i)$ with $\delta_i = I(T_i < C_i)$ a censoring indicator.

There are many different ways to simulate censored observations. One way is to use some specific R packages for the simulation of survival data. Package `survsim` [**82**] could be used for simulation of simple and complex survival data, such as multiple event survival data and recurrent event survival data. `survsim` allows simulation of times using Weibull, lognormal and log-logistic distributions.

Another R Package is `PermAlgo`, which is a permutational algorithm to simulate survival data. This algorithm is a flexible tool to generate a dataset in which event and censoring times follow user-specified distributions and also they can be conditional on a user-specified list of covariates [**100**].

Additionally, Bender et al. [**9**] show how survival times can be generated to simulate Cox models with known regression coefficients and with any non-zero baseline hazard rate. In this article survival times generated from a variety of survival distributions including the exponential distribution for constant hazards, the Weibull distribution for monotone increasing or decreasing hazards, and the Gompertz dis-

tribution to model human mortality. Also, Milkoslavsky et al. [**79**] discuss how to simulate survival times with time-dependent covariates and dependent informative censoring.

Halabi and Singh [**50**] provide formulae to obtain the specific amount of censoring in both unstratified and stratified cases. In other words, for a given censoring probability p, the proportion of patients P_j allocated to treatment j and failure time survivor function $S_j(t)$ in group j, the parameter of the censoring density is obtained. For a given overall proportion p of censoring, the parameter of the censoring distribution is calculated by solving the following equation.

$$p = \sum_{j=1}^{k} P_j \int_0^{\infty} S_j(t) g_c(t) dt \tag{5.1}$$

where $g_c(t)$ is the censoring density.

For the case when stratification is present, the overall censoring proportion p is given by

$$p = \sum_{s=1}^{l} \sum_{j=1}^{k} P_{js} \int_0^{\infty} S_{js}(t) g_c(t) dt \tag{5.2}$$

where P_{js} is the proportion of patients who received treatment j in stratum s, S_{js} is the survivor function for the failure time in stratum s and treatment j. And $g_c(t)$ is the censoring density across the strata and treatment groups.

Since in our simulation study we need to prespecify the percentage of censoring, in the next section we are going to use the Halabi and Singh [**50**] method to generate a desired percentage of censored observations.

5.3 Generating Censored Observations

As discussed above, we use the Halabi and Singh [**50**] method to generate a desired percentage of censored observations. We asssume the survival data, T, are generated from a distribution with failure distribution $F(t)$ and survivor function $S(t)$. The goal is to obtain a percentage p of survival data as censored observations. Let C denote the censored time with density $g(c)$. If there is only one treatment group,

the exact amount of censoring is calculated as follows;

$$p = \int_0^\infty S(c)g(c)dc \tag{5.3}$$

The goal is to choose the proper distribution for censored observations. To achieve it, we will find a bound range when there is no censoring to use as a comparison tool. First, we ran 100 iterations with sample size equal to one hundred from the Weibull distribution ($Weibull(2, 4)$). We divided the time span into 100 intervals. Then for each iteration, the differences between the true value and the estimated value based on Kaplan-Meier estimator are evaluated at each time point which we called True-Estimate. Finally, we plot True-Estimate for all iterations and draw the bound for the maximum and minimum value of iterations in Figure 5.1.

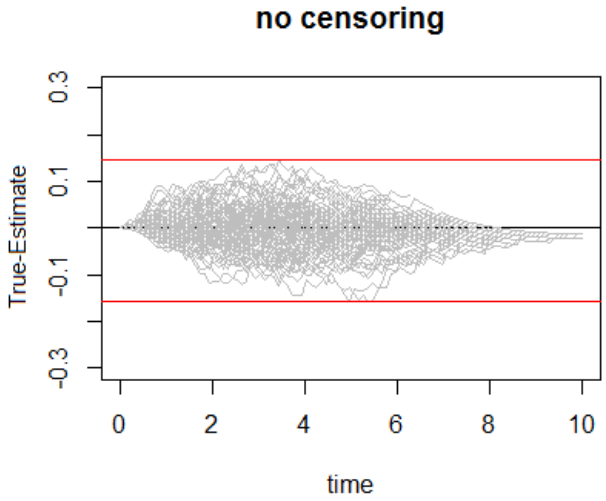

Figure 5.1: Bound range for difference of true survival and KM estimated value

We consider three different censoring mechanisms including binomial, uniform and exponential distributions to investigate the best mechanism regarding the similarity of Kaplan-Meier estimation of survivor plot to the true failure values. To explore whether the assumed censoring distributions are good choices we see how these methods are working in studies with a low, medium or a high percentage of censoring.

Our comparison among different censoring distributions starts by assuming a binomial distribution for censored observations. First, 100 samples are generated from a

binomial distribution $B(1, p)$, where p equals to the desired percentage of censoring, to give the censoring indicators. Second, for each selected observation, a sample is generated from a uniform distribution $U(0, T_i)$ where T_i is the failure value for the ith observation. Finally, we set the generated value from this uniform distribution as the censored observation. The difference between the true Weibull and KM estimate based on binomial censoring is shown in Figure 5.2 according to different percentages of censoring. Based on the results in Figure 5.2, binomial censoring does not work well in the presence of a high percentage of censoring, as the Kaplan-Meier estimate overestimates the survivor function.

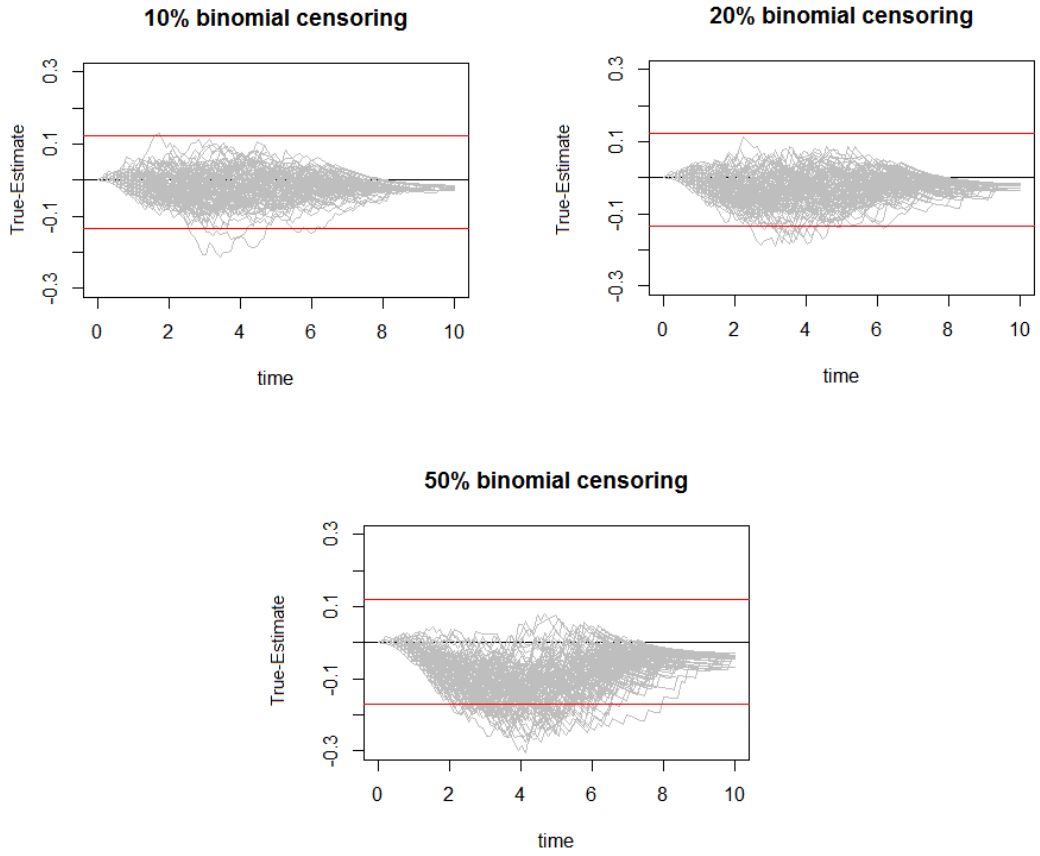

Figure 5.2: True-Estimate values based on $Weibull(2, 4)$ for 100 iterations using binomial censoring.

As another option for censoring, we assume a uniform distribution $U(0, b)$ for the censored observations. To achieve a desired percentage of censoring, p, the parameter b of the uniform distribution needs to be calculated. In general, for any failure

distribution $f(t)$;

$$p = p(C < T) = \frac{1}{b}\int_0^b \int_c^\infty f(t)dtdc = \frac{1}{b}\int_0^b \left[1 - \int_0^c f(t)dt\right] dc \tag{5.4}$$

In our simulations, we use a Weibull distribution for the failure observations. The difference between a true Weibull and the KM estimate based on uniform censoring can be found in Figure 5.3.

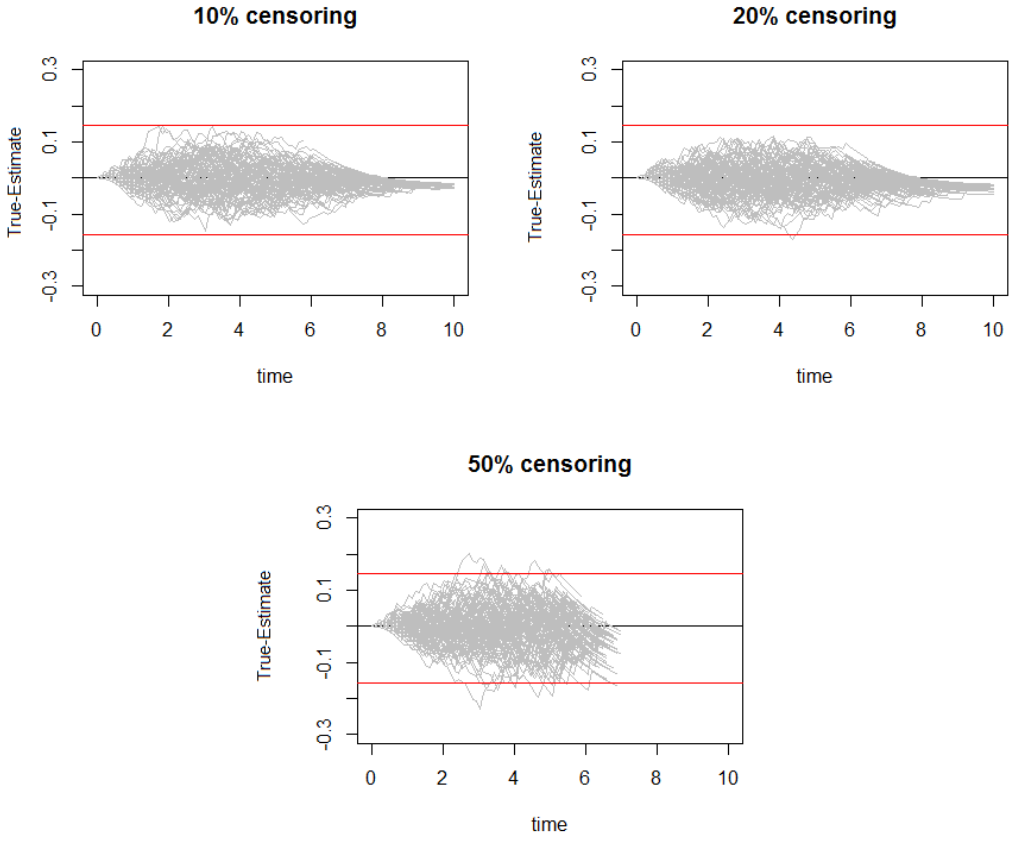

Figure 5.3: True-Estimate values based on $Weibull(2, 4)$ for 100 iterations using uniform distribution for censored observations. The calculated value for uniform parameter b is 35.45 for 10 percent, 17.73 for 20 percent and 6.99 for 50 percent censoring

Although for different percentages of censoring all of the iterations are in the bound range, there is evidence of truncation under 50 percent censoring. It is more evident in this scenario because the value for b is 6.99, which is smaller than the follow-up time. In uniform censoring when we have large value for the failure time, $P(C < T)$ becomes large therefore long survival time values are more likely to be censored. So, by using uniform censoring the data can be truncated.

Finally, an exponential distribution is considered for the censored observations. Consequently to achieve the desired percentage of censoring, the parameter of the exponential distribution needs to be calculated. Failure times are assumed to have a Weibull distribution with parameters α and λ and censored observations have an exponential distribution with parameter β , then by defining a percentage of censoring, the parameter of censored distribution is calculated as follows:

$$p = p(C < T) = \int_0^\infty \int_c^\infty \alpha\lambda t^{\alpha-1} e^{-\lambda t^\alpha} \beta e^{-\beta c} dt dc = \int_0^\infty e^{-\lambda c^\alpha} \beta e^{-\beta c} dc \qquad (5.5)$$

By changing the target percentage of censoring, the required parameter of the censored distribution β will need to change in (5.5). In general, by defining different values for β, the integral can be evaluated. So the integration values could be plotted against the β values and, for any desired percentage of censoring, which is equivalent to the value of integral, the related value for β can be found. For explicit calculation the `uniroot` function in R can be used. The `uniroot` function searches for a root (i.e., zero) of a function f over its first argument in a specified interval, from lower to upper limits. Here the function which is used is the integral minus the desired percentage of censoring considered as a function of β, i.e.

$$f(\beta) = \int_0^\infty e^{-\lambda c^\alpha} \beta e^{-\beta c} dc - p$$

As can be seen from Figure 5.4 all of the iterations for different percentages of censoring are in the bound range.

In summary, the simulations presented here suggest that the best results are obtained when the exponential distribution is chosen as the censoring distribution. Therefore it is used in implementing the censoring mechanism in the remainder of our simulation studies.

5.4 Simulation Scheme

In this part, the simulation methodology is described. Samples from a desired distribution are generated. Then some observations are assigned as censored values

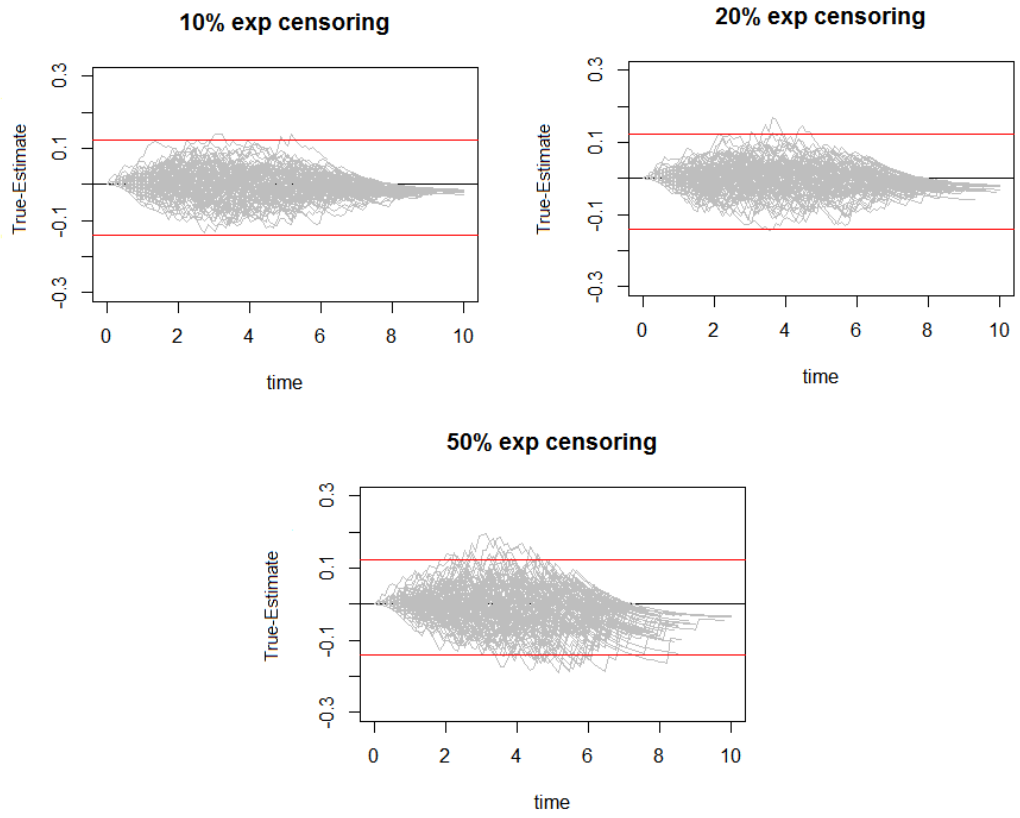

Figure 5.4: True-Estimate values based on $Weibull(2, 4)$ for 100 iterations using exponential distribution for censored observations. The calculated value for exponential parameter λ is 0.031 for 10 percent, 0.065 for 20 percent and 0.216 for 50 percent censoring

by drawing a random values from the censoring distribution, taken here to be an exponential distribution based on the discussions above. If the value that is generated from the censoring distribution is smaller than the actual survival value in the sample, then that individual is assumed to be censored with the censoring value obtained from the exponential distribution. Then by using different methods of imputation, including the parametric Bayesian and nonparametric Bayesian methods which were described in the previous chapter, values for the incomplete censored observations are imputed. By including these imputed values together with the uncensored failure times, a complete data set is created.

These steps are described as follows:

- Draw a random sample $\{t_1, ..., t_n\}$ of size n from the Weibull distribution as failure times.

- Draw a random sample $\{c_1, ..., c_n\}$ of size n from the exponential distribution with parameter β as potential censoring times.
- Obtain the observed value as $y_i = \min(c_i, t_i)$ for $i = 1, ..., n$.
- Obtain the censoring indicator as $\delta_i = 1_{\{t_i \leq c_i\}}$.
- Use the parametric Bayesian method and non-parametric Bayesian method to impute the censored observations in the data set (y_i, δ_i).
- Make a complete dataset by replacing censored times with the imputed failure times.
- Calculate the difference between the survivor curves of the true and the completed dataset at different time points.
- Repeat these steps 100 times.

In the next section, the results of the parametric Bayesian imputation are described for different sample sizes and different percentages of censoring.

5.5 Parametric Bayesian Imputation

In this section, the results of imputing censored observations using the parametric Bayesian approach will be discussed. This method is described in detail in Section 4.5. The samples are simulated from a Weibull distribution, which is a good flexible candidate for modelling survival data as, depending upon the shape parameter, it can have an increasing, decreasing or constant hazard function. As the sample size may affected the imputation results, we consider the effects of different sample sizes. We consider three cases, small ($n = 50$), medium ($n = 100$), and large ($n = 200$) sample sizes. The percentage of censoring, as well as sample size, has a significant impact in analysing survival data and also on any imputation. In the case of a high percentage of censoring the variation in imputations may be increased. To study this samples are generated using different percentages of censoring; low ($p = 10\%$), medium ($p = 20\%$) and high ($p = 50\%$). Furthermore, we want to investigate if

there is any difference in the imputation results for samples from Weibull distributions with different hazard shapes. Hence, the data are simulated from constant, increasing, and decreasing hazards to see the impact, if any, on the performance of imputation.

The general procedure is as follows: a sample is generated from a particular Weibull distribution and the resulting values are considered as the true failure times. Then the desired percentage of censoring is applied, where for the censored observations the true failure values are replaced with values from an exponential distribution, as described in Section 5.3. Winbugs is used to fit the parametric Bayesian model and simulated draws of the imputed values for the censored observations are generated. Using 100 sets of these simulated imputed values and combining them with the uncensored event times, 100 completed datasets are generated. The difference between the survivor function based on the true values from the Weibull distribution and the survivor function based on the completed (imputed) datasets is calculated over a range of different time points. The main goal is to show that our method is at least as good as the Kaplan-Meier estimate, so it is sensible to compare our results to Kaplan-Meier estimates. Consequently, as a visual measure, to see how well the model is working, these 100 completed datasets are also compared to the range and quantiles of the Kaplan-Meier estimates. If the difference between the true failures and our completed dataset are within theses bounds for each of the 100 sets of data, it could be concluded that our method is as good as Kaplan-Meier estimates. The steps for plotting the bound and quantile ranges are as follows:

- 100 sets of data with the desired sample size are simulated from the specific Weibull distribution with no censoring.
- The survivor curve of these data is estimated using a Kaplan-Meier estimator.
- The Kaplan-Meier estimated value for all of the 100 simulation sets is subtracted from the true value of the Weibull distribution survivor function at specific time points.
- The maximum and minimum of these difference values are used to define a bound range over the time period of interest.

The steps for plotting quantile range are as follows:

- 100 sets of the data with the desired sample sizes are simulated from the specific Weibull distribution with the desired percentage of censoring.
- The survivor curve of these data is estimated using a Kaplan-Meier estimator.
- The Kaplan-Meier estimated value for all of the 100 simulation sets is subtracted from the true value of the Weibull distribution survivor function at specific time points.
- Quantiles of these differences are calculated at each of the time points and plotted as a quantile range.

We will start with the simplest constant hazard scenario. As described in Chapter 2, the Weibull distribution has the constant hazard when $\alpha = 1$ and it reduces to an exponential distribution. In Figure 5.5 the results are shown for different sample sizes and different percentages of censoring where the samples are simulated from a $Weibull(1, 4)$ distribution.
As it can be seen from Figure 5.5, by increasing the sample size, the variability decreases and the bounds become tighter as is to be expected. An interesting and good property to note is that in all of the graphs, the medians of the boxplot at each time point are nearly zero, and the boxplots are symmetric around the zero line. For the same sample size increasing the percentage of censoring causes some of the iterations to stay outside the bound, which means it is better not to use imputation methods when there is a high percentage of censoring.
In the second simulation study, the data are simulated from a Weibull with $\alpha > 1$ which leads to an increasing hazard situation. As it can be seen in Figure 5.6, again, by increasing the sample size, the imputations look reasonable as the difference between the true failure and Kaplan-Meier estimates for 100 the complete datasets are symmetrical around the zero line. However, at the later time points all of the iterations are below the zero line in nearly all of the graphs, which means that there is overestimation for the large censored values and this becomes worse when there is a high percentage of censoring.

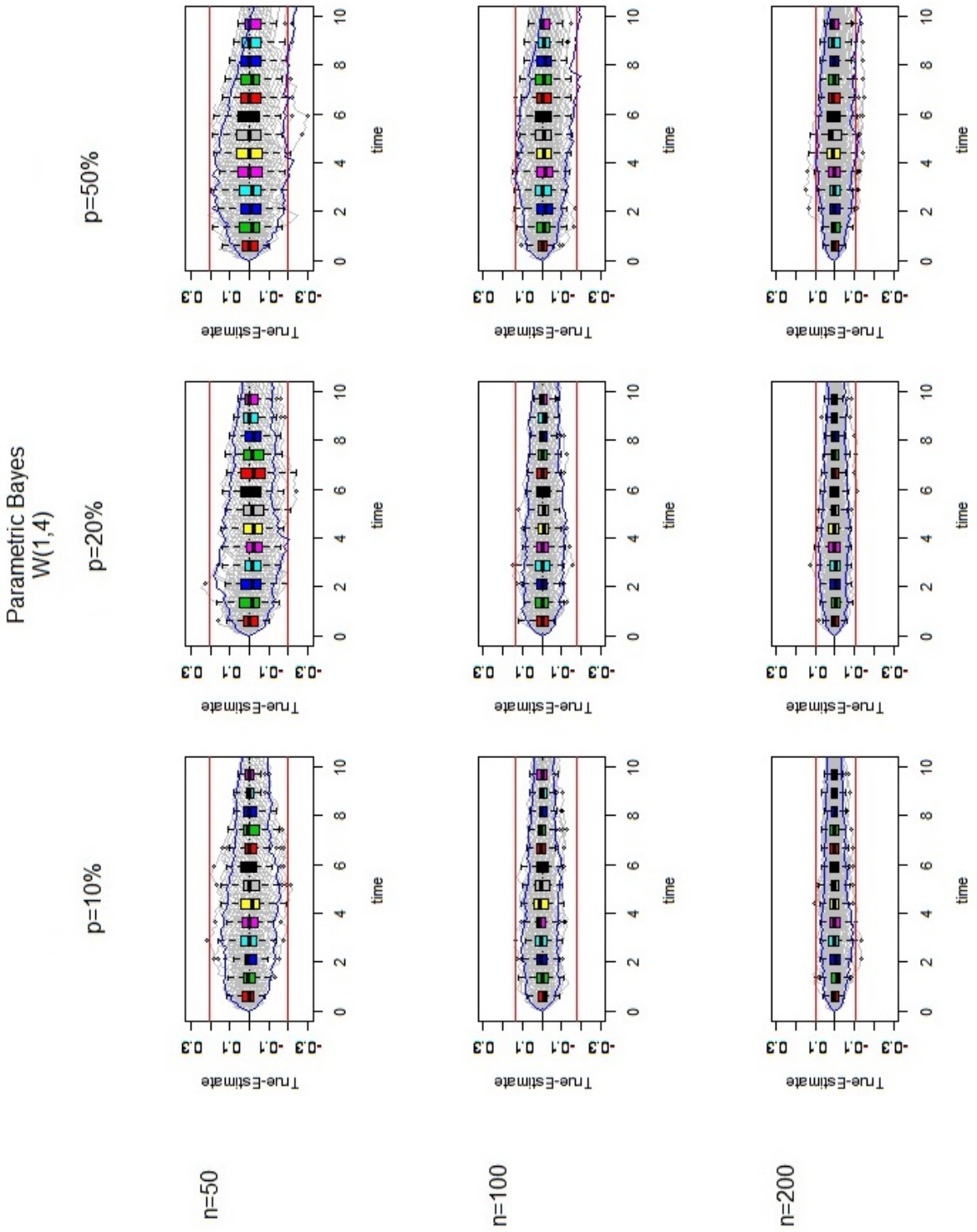

Figure 5.5: True-Estimate values using the parametric Bayesian approach for 100 datasets generated from $Weibull(1, 4)$ using exponential censoring

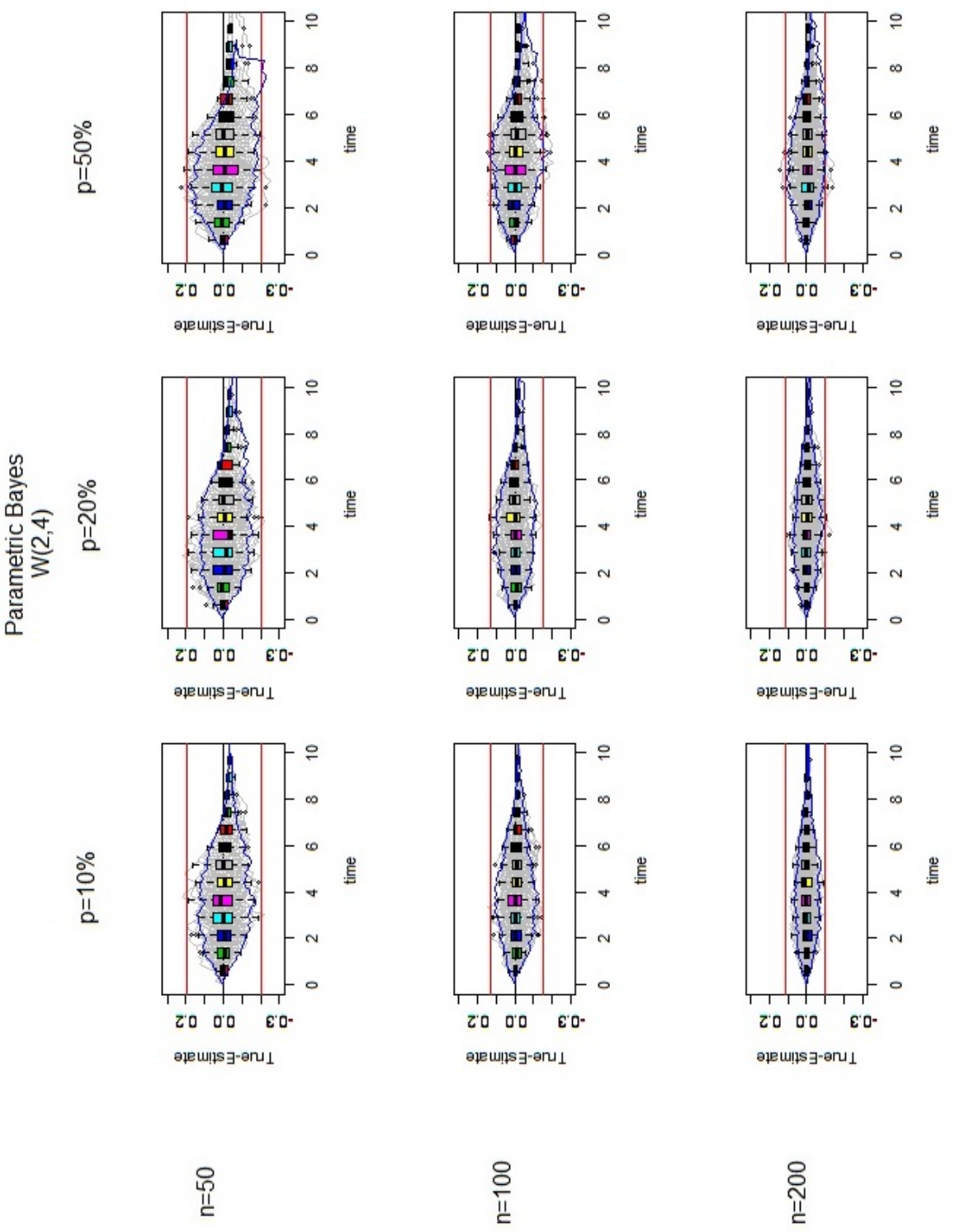

Figure 5.6: True-Estimate values using parametric Bayesian approach for 100 datasets generated from a $Weibull(2, 4)$ using exponential censoring

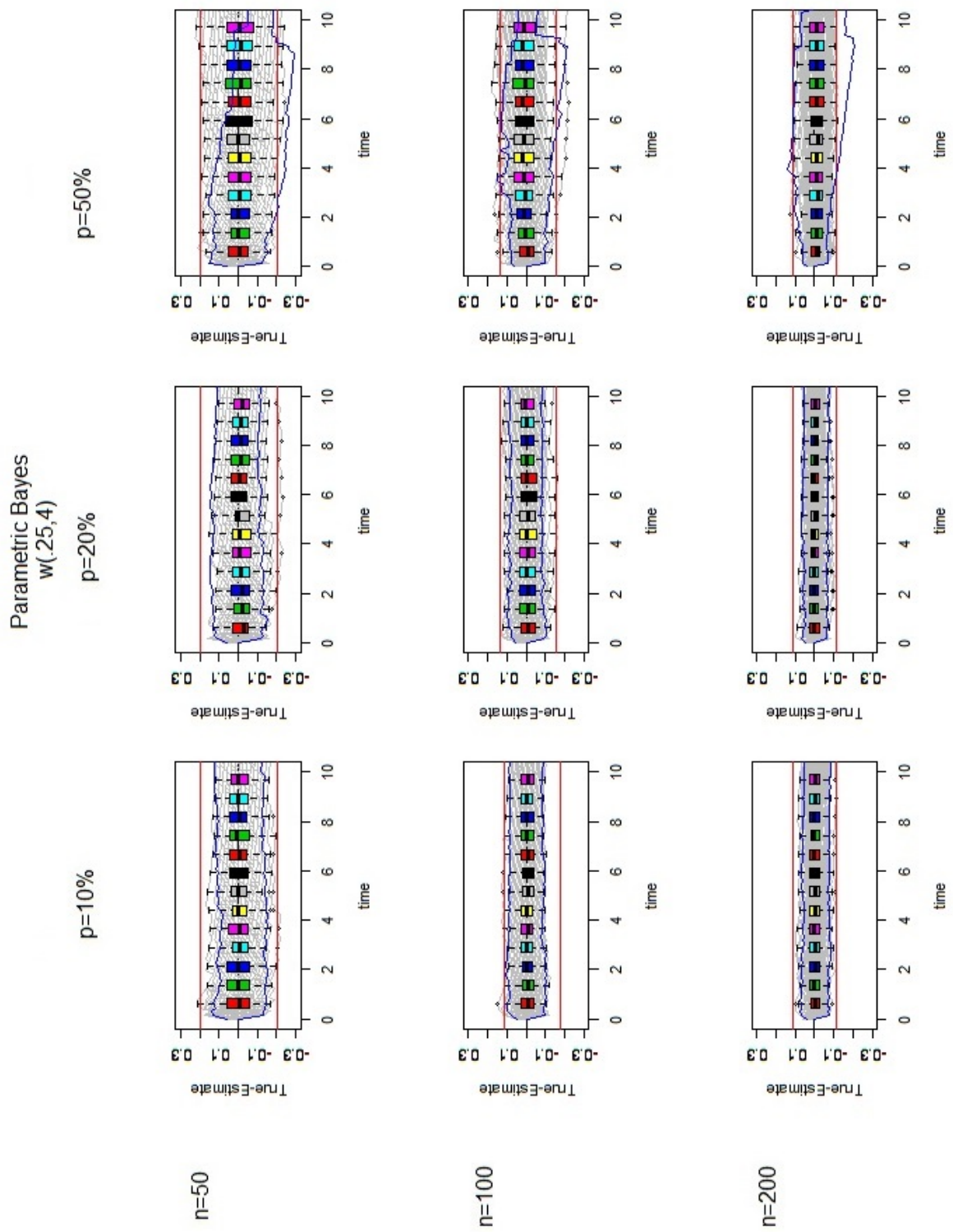

Figure 5.7: True-Estimate values using parametric Bayesian approach for 100 datasets generated from a $Weibull(0.25, 4)$ using exponential censoring.

Finally, we consider a Weibull distribution with decreasing hazard ($\alpha < 1$). In Figure 5.7, as in Figures 5.5 and 5.6, by increasing the sample size the imputed values becomes closer to the true values. Also in comparison to the increasing hazard case, at the later time points, the differences between the true failure and Kaplan-Meier estimates are still symmetric around the zero line, which means there is no significant under or over estimation of the censored observations. Moreover, in nearly all of the combinations of sample size and censoring percentage graphs, the boxplots are symmetric with medians near zero lines.

In conclusion, in all of the these three different situations, with increasing, decreasing and constant hazards there are some common features. First, by increasing the sample size the imputed values become closer to the true values. Second, we get reasonable imputed values when there is only a low degree of censoring. Third, in all of the 27 combination graphs, different draws are within the Kaplan-Meier bound, which means our parametric Bayesian approach is within the variability expected for the Kaplan-Meier estimate. However, across the different hazard scenarios, the imputed values seem more reliable in the case of decreasing or constant hazard, as nearly all of the iterations are placed symmetrically around the zero line, in contrast to the increasing hazard case where there is an indication of overestimation for high censored times.

In the next section, our second approach of imputation is discussed. In this method, the censored observations are imputed using the non-parametric Bayesian approach which was introduced in Section 4.6.

5.6 Non-parametric Bayesian Imputation

In this section, the results of imputing censored observations using a non-parametric Bayesian approach will be discussed. Here, we again examine our method using different Weibull distributions with different hazard function forms. Again, we explore the effect of various percentages of censoring and different sample sizes. As before, 100 draws are simulated from the desired Weibull with the assumed percentage of censoring. Since from our simulations the true failure times are initially known

and only then potentially censored, the imputed values can be compared with the original values before censoring. So after imputing censored observations using the non-parametric Bayesian approach, these imputed values are subtracted from true failure values at different time points. The closer these repeated realizations are to the zero line, the better the estimates are. Additionally, these differences are compared to the bound range of Kaplan-Meier estimate with no censoring and the quantile ranges of the Kaplan-Meier estimate under the desired percentage of censoring.

The simulation study is started with a constant hazard Weibull (exponential) distribution $Weibull(1, 4)$ where the shape parameter of the distribution is equal to one ($\alpha = 1$). The results for combinations of different sample sizes and percentages of censoring can be found in Figure 5.8. It can be seen from Figure 5.8 that when the sample size increases the results of different draws of imputed values are within a smaller range. Also, by increasing the percentage of censoring the imputed values become less accurate as there is a sign of overestimation at the end of follow-up time. The medians of boxplots are near zero in the case of 10% and 20% of censoring, but in 50% censoring case after some time point (approximately after half of the study) the medians fall below the zero line which corresponds to overestimation. This is not surprising as it makes little sense to impute more than half of the data and in that situation there is not enough information in the observed failures to determine the failure distribution.

We now consider simulating from a Weibull distribution with increasing hazard. The data are simulated across the 9 different scenarios from a $Weibull(2, 4)$. The results are displayed in Figure 5.9. It can be noticed from Figure 5.9 that after some time the median falls below zero line in all of the scenarios, which indicates an overestimation at those time points. This becomes even more pronounced under a high percentage of censoring, where nearly all of the realizations are below the zero line which means that the censored observations at those time points are definitely overestimated. In general, in the case of increasing hazard the imputed values are reliable if the sample size is large enough along with a low percentage of censoring.

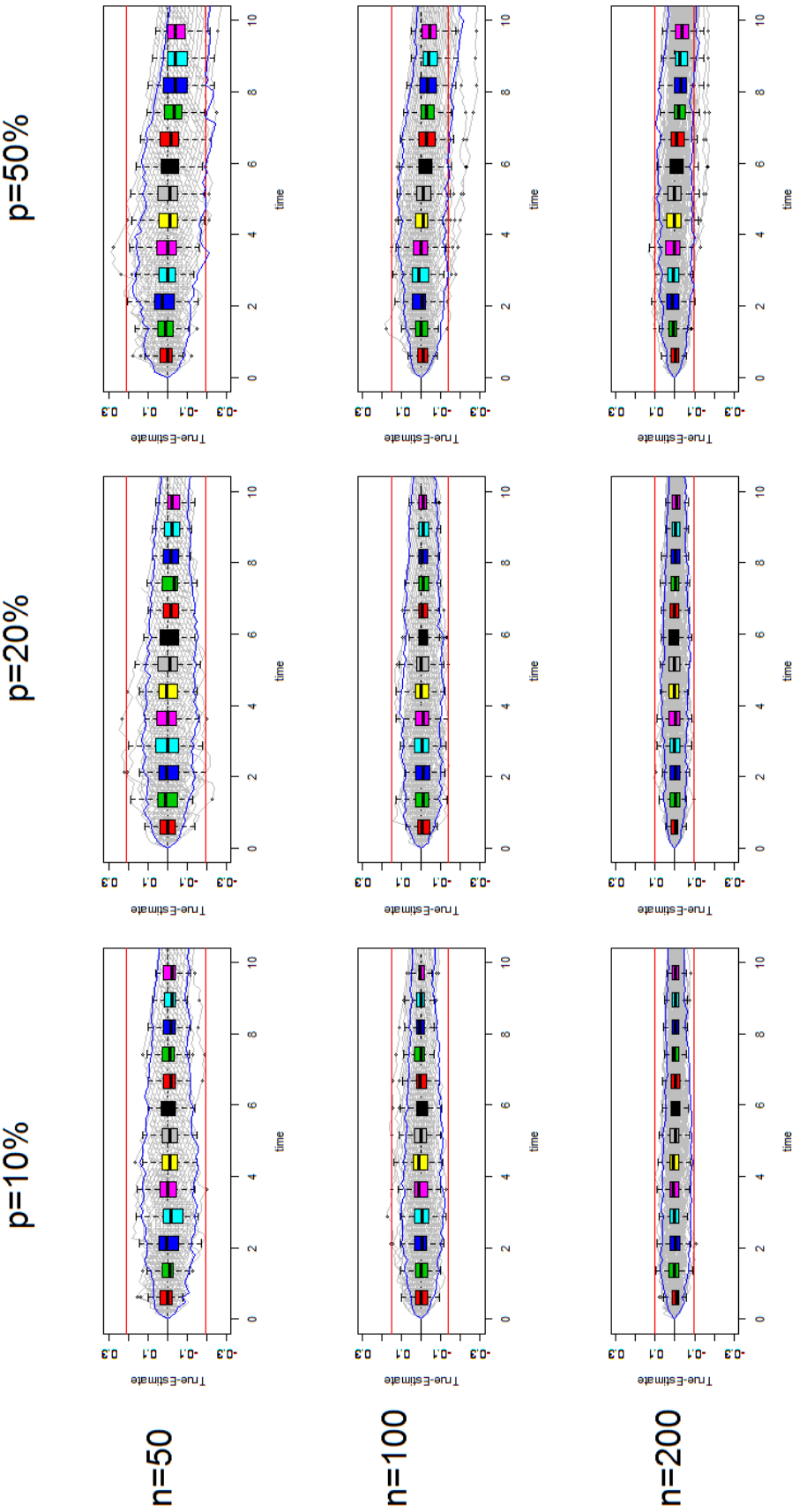

Figure 5.8: True-Estimate values using the non-parametric Bayesian approach for 100 datasets generated from $Weibull(1, 4)$ using exponential censoring.

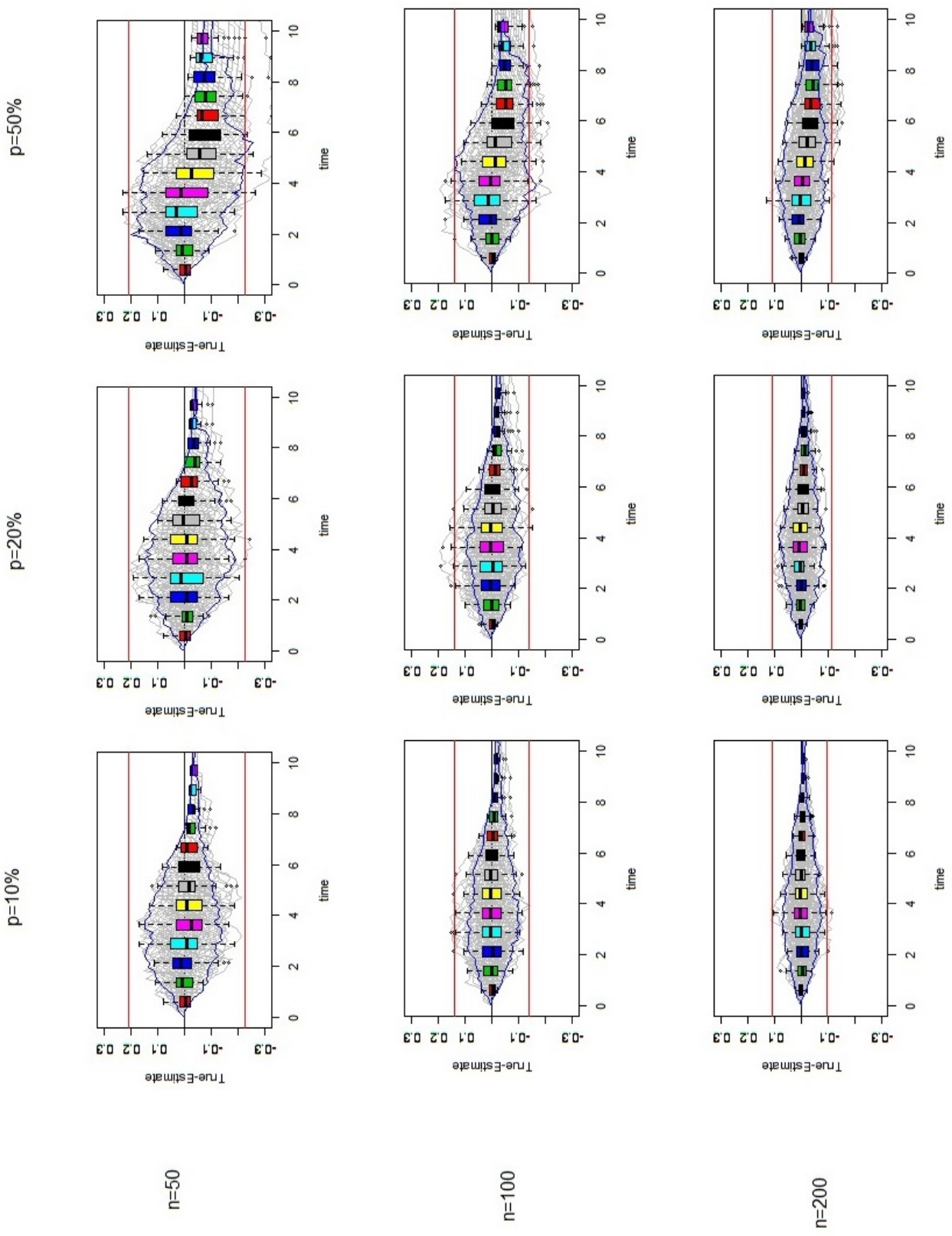

Figure 5.9: True-Estimate values using non-parametric Bayesian approach for 100 datasets generated from a $Weibull(2, 4)$ using exponential censoring.

Finally, we look at the simulation results for a Weibull distribution with decreasing hazard as in Figure 5.10 where the data are simulated under the 9 different scenarios from a $Weibull(0.25, 4)$.

As it can be seen from Figure 5.10, all of the medians across the different time points are around zero and the boxplots are mostly symmetric around the zero-line. However, under 50% censoring some of the trajectories are outside the Kaplan-Meier bound.

To sum up, if the failure times are generated from a Weibull distribution, the non-parametric Bayes approach imputes plausible values for censored observations in the case of a decreasing hazard function. In the case of increasing hazard this method gives overestimated values for long-term censored observations. In the case of constant hazard, the non-parametric Bayesian approach gives good estimates when there is at most a medium percentage of censoring. Hence, this approach is not recommended for imputation in the presence of a high percentage of censoring.

5.7 Comparing Imputation Methods

Here, the parametric Bayesian and non-parametric Bayesian approaches are compared using their simulation results. At first, the assumption of a correct assumed distribution for the data is considered in the parametric Bayesian imputation approach, while in the second part the censored observations are imputed using the parametric Bayesian approach based on a mis-specified probablity model. These comparisons are made based on data generated from a Weibull distribution with constant hazard $Weibull(1, 4)$ with sample size of 200. The difference between the survivor functions based on the true Weibull values and the imputed values over different time points is presented for the two imputation approaches and also under different percentages of censoring. These True-Estimate values are compared to the range and quantiles of the Kaplan-Meier estimate.

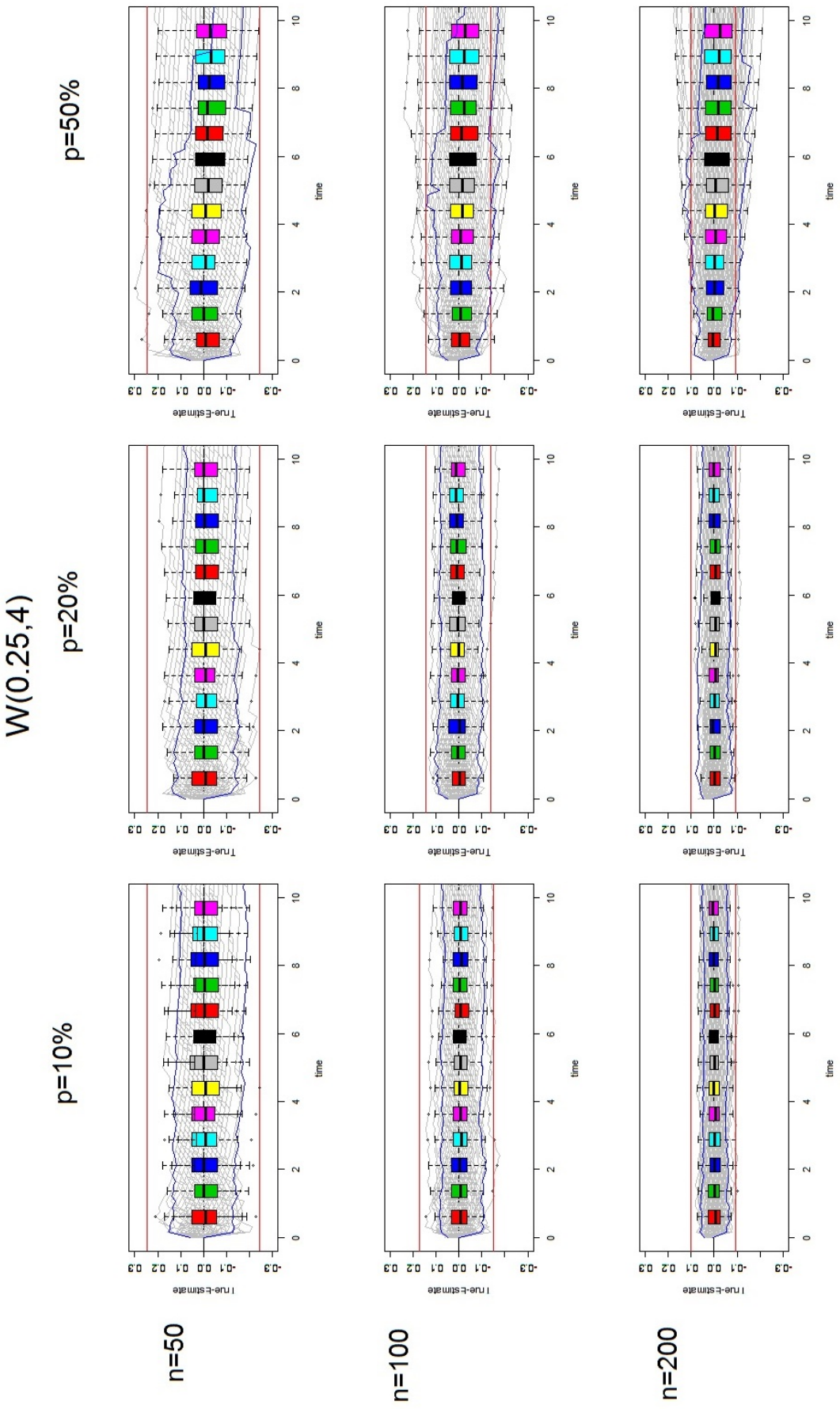

Figure 5.10: True-Estimate values using non-parametric Bayesian approach for 100 datasets generated from $Weibull(0.25, 4)$ using exponential censoring.

5.7.1 Using a Correct Model

In this part, one hundred datasets are generated from a $Weibull(1, 4)$ with different percentages of censoring. In the parametric Bayesian approach, the distribution of the data needs to be pre-specified so based on the simulation setting a Weibull distribution is assumed for the data together with lognormal and Gamma distributions as priors for the shape and scale parameters of the Weibull distribution, respectively. Censored is then applied as above and these observations are then imputed using parametric Bayesian and non-parametric Bayesian methods. The results are presented in Figure 5.11. Based on Figure 5.11, for large sample sizes ($n = 200$) the parametric Bayesian approach and non-parametric Bayesian method are nearly the same under low and medium percentages of censoring. However, with a high percentage of censoring , $p = 50\%$, the parametric Bayesian approach works quite well as the boxplots are symmetric around zero lines and all the medians are nearly zero. However, when using the non-parametric Bayesian approach, we obtained imputed values which are much higher than the true failure times and so overestimation occurs.

In the case of a high percentage of censoring, when the actual distribution of the data is known, imputed values for censored observations using parametric Bayesian approach are closer to the true failure times in comparison to the non-parametric Bayesian imputed values. This is not unexpected as we assuming the correct distribution in the parametric Bayesian approach.

5.7.2 Using a Mis-specified Model

In the previous part, the correct assumption about the distribution of the data is used in the parametric Bayesian approach. However, in reality, the distribution of the data may not be known. In this part we consider the effect of mis-specifying the survival distribution in the parametric Bayesian model. One hundred datasets are generated from a $Weibull(1, 4)$ with different percentages of censoring.

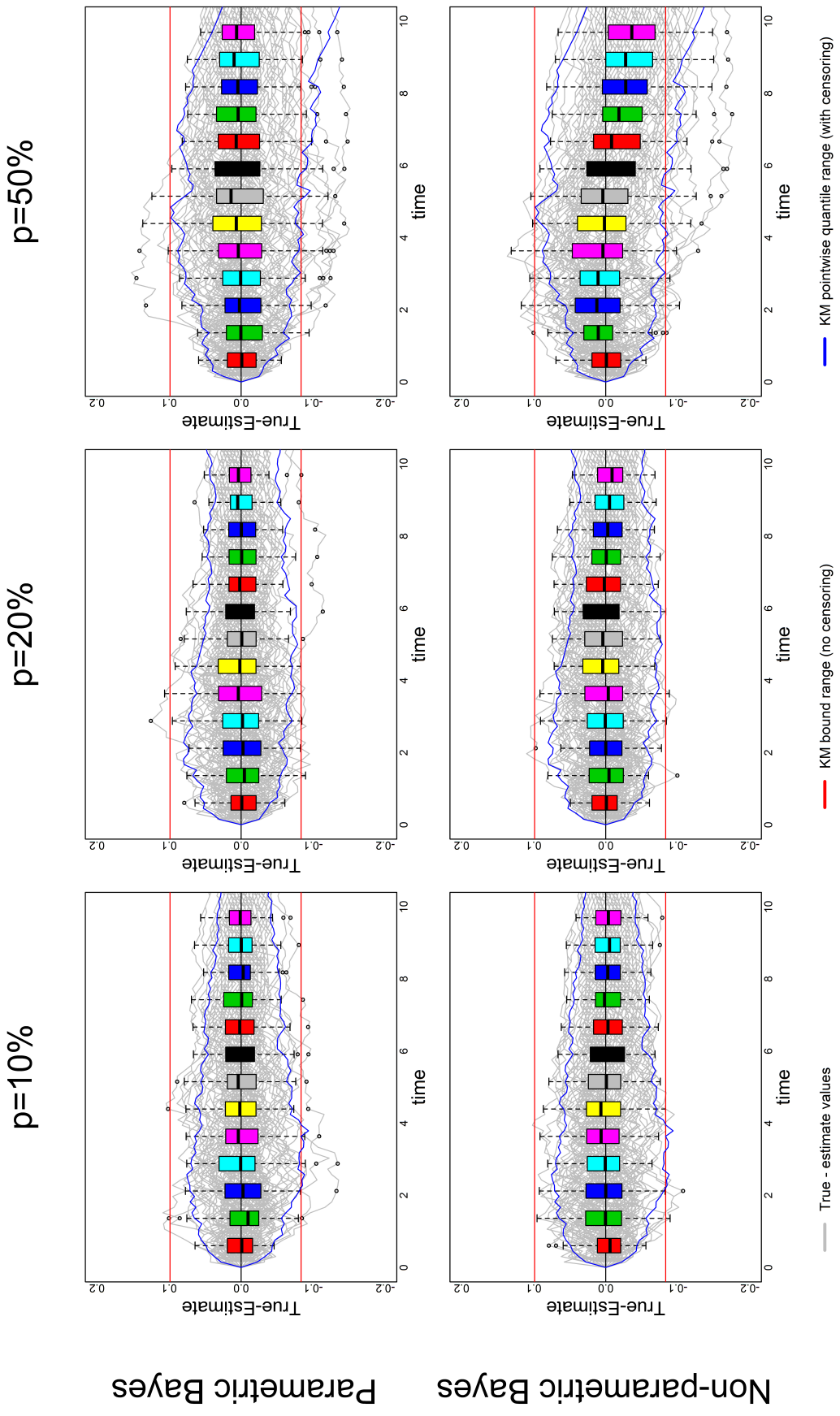

Figure 5.11: True-Estimate values based on a $Weibull(1, 4)$ for 100 iterations using parametric Bayesian and non-parametric Bayesian approach for n=200

In the parametric Bayesian approach, the distribution of the data is assumed to have a lognormal distribution with normal and Gamma distribution priors for the mean and standard deviation of the lognormal distribution, respectively. The results are presented in Figure 5.12.

Based on Figure 5.12, non-parametric Bayesian methods provide better estimation of the censored observations in comparison to the parametric Bayesian approach, as the medians are near zero especially for low and medium percentages of censoring. However, in the parametric Bayesian approach, the results of True-Estimate values are still within the Kaplan-Meier range for low and medium percentages of censoring. This means that even by mis-specifying the model assumption in the parametric Bayesian approach it is still as good as the Kaplan-Meier estimates under low and medium percentages of censoring. For a high degree of censoring both parametric and non-parametric Bayesian methods overestimates the censored observations as the medians fall below zero line. As we have already noted this is not surprising as it is unwise to use imputation methods when more than approximately half of the observations are censored. Under a high percentage of censoring the non-parametric Bayesian method provides better estimation for censored observations.

In general, in comparison to the parametric Bayesian method, the non-parametric Bayesian imputation approach is very time-consuming. For example, for a single simulation realisation with 10000 MCMC iterations it took around 4 to 10 minutes to run, depending on the sample size and the percentage of censoring. So on average, each scenario with 100 sets of data took around 13 hours to run (there are 27 different scenarios in our simulation studies). The parametric Bayesian approach is much faster taking around one minute, or less, to run, so each scenario with 100 sets of data could be completed in around 1.5 hours (again there are 27 scenarios to be considered in parametric Bayesian approach).

In conclusion, if we know the true distribution of the data it is better to use the parametric Bayesian approach as it is not only faster but also it provides appropriate imputations for censored values. In the situations where the true distribution of the data is unknown the non-parametric Bayesian approach is recommended, but it needs considerably more time to run.

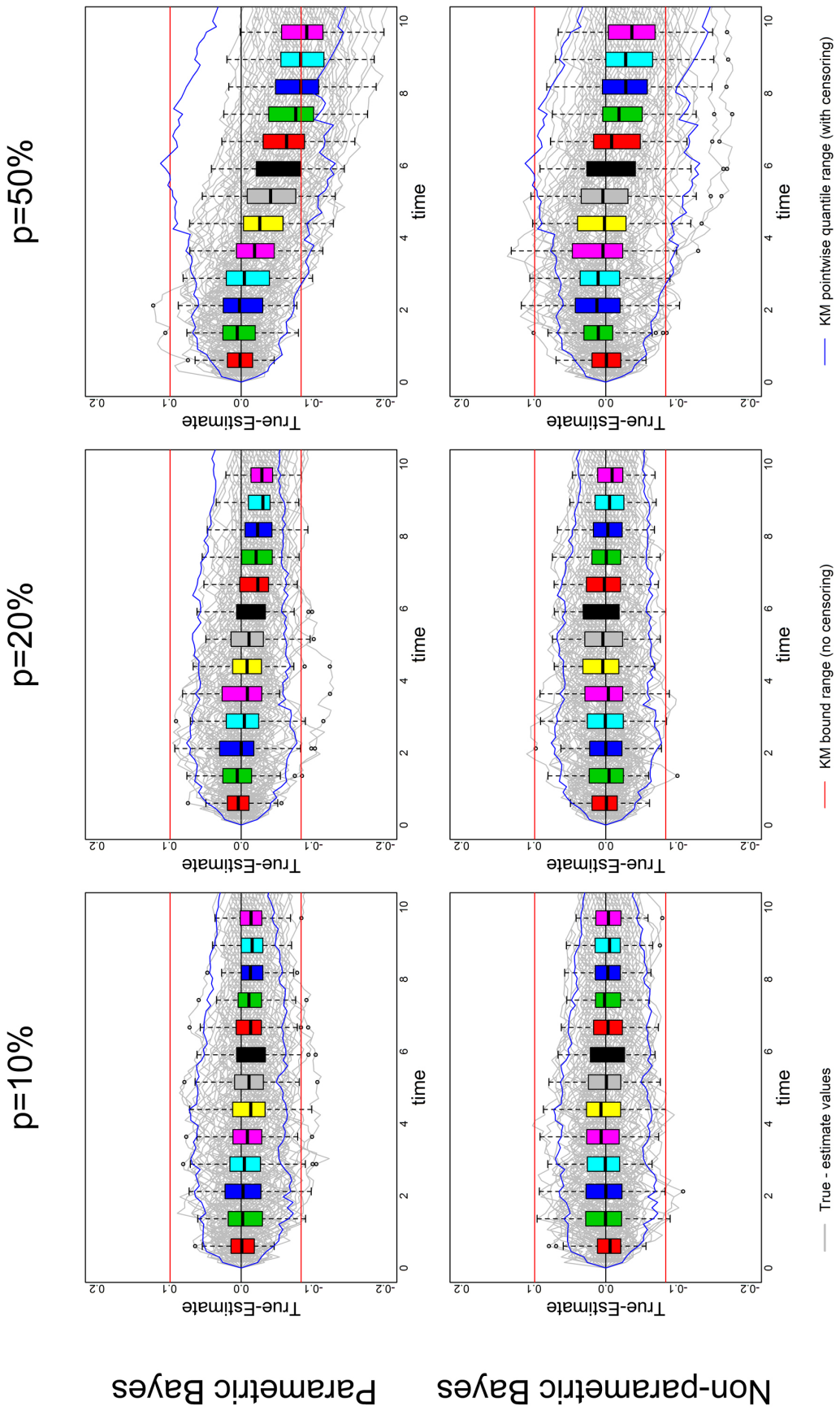

Figure 5.12: True-Estimate values based on a $Weibull(1, 4)$ for 100 iterations using parametric Bayesian (assuming lognormal distribution in imputation method) and non-parametric Bayesian approach for n=200

5.8 Chapter Summary

In this chapter, we mainly focused on our simulation results. At the beginning of the chapter, we reviewed different methods of generating censored observations. From these methods, we decided to use the Halabi and Singh method to generate realizations with the desired percentage of censoring. Based on our simulation results the exponential distribution is chosen as a suitable distribution for the censored mechanism. Therefore the exponential distribution is used in the rest of the simulation studies as the censoring distribution.

In the simulation studies both parametric Bayesian and non-parametric Bayesian approaches are used to impute censored observations. The imputation results are generated based on different sample sizes and different percentages of censoring. Moreover, the data are simulated from the Weibull distribution with different hazard shapes. In both parametric and non-parametric Bayesian methods, as the sample size increased the imputed values became closer to the true values. Also, for the same sample size by increasing the percentage of censoring the imputed values become less accurate. Furthermore, for most of the time, the imputation values are within the Kaplan-Meier bound, which means our imputation methods are at least as good as the Kaplan-Meier estimates, especially in the cases of low or medium percentages of censoring. In the case of constant or decreasing hazard, both methods exhibit better performance in comparison to the increasing hazard case as the realizations are symmetrical around zero line.

If we know the true distribution of the data, it is better to use the parametric Bayesian approach as it provides appropriate imputations for censored values. In the situations where the true distribution of the data is unknown, the non-parametric Bayesian approach is recommended and potentially superior, but this needs to be balanced against the fact that it is much more computationally intensive and this more time-consuming. In general, these imputation methods are not recommended for a high percentage of censoring, as when half or more of the data are censored there is insufficient information in the observed failure times to provide meaningful imputations.

Chapter 6

Applications

6.1 Introduction

The simulation chapter gave an insight into the likely performance of the proposed Bayesian imputation methods. Building on these results the methods developed in this thesis were applied to the three example datasets introduced in Chapter 1 to motivate the benefits of considering parametric Bayesian and non-parametric Bayesian methods for imputing censored observations in time to event studies.
This chapter is arranged as follows: we start with a review of the methods of estimating an empirical density in Section 6.2. In Section 6.3 the results of parametric Bayesian and non-parametric Bayesian approaches are discussed when applied to the 6-MP dataset. Then in Section 6.4, our proposed methods are investigated for imputing censored observations in the metastatic renal carcinoma dataset. In Section 6.5 the BPD dataset is interpreted. Finally, in Section 6.6 concluding remarks are made in terms of the applicability of these approaches.

6.2 Estimating a Density Function

One of the goals of this thesis is to impute the censored observations in order to generate the density plot of the complete dataset as an additional graphical repre-

sentation of the data to complement the widely used Kaplan-Meier plot. To start, a commonly used method to estimate and plot an estimate of a density of a distribution is described.

In practice, the common way of modelling an unknown distribution is to assume the data in question were generated from one of the classical distributions such as the Weibull, gamma, normal, lognormal, or beta. But in reality, the distribution that generated the data is unknown so a histogram, kernel or other non-parametric estimate of the unknown density function, based on the data provided, can be used to generate a plot that may be a useful guide.

Constructing a histogram is easy and also displaying data in the form of a histogram can provide an experimenter with useful information at a glance. But before constructing a histogram, the number of classes, the width of each class and also the lower limit of first class should be considered, which are not easy to choose. For example, if the widths are too small, distracting features are introduced and if they are too large, important features may be lost [**106**].

The empirical density function, which belongs to a large class of non-parametric density estimators, is a simple modification and improvement of the usual histogram. It is estimated directly from sample data, without assuming an underlying form of the distribution model [**41**]. The definition of the distribution function is $F(t) = P(X \leq t)$, so by assuming $X_1, ..., X_n$ to be a random sample from a distribution function F on the real line, the underlying cumulative distribution function (CDF) can be estimated using the following formula:

$$F_n(t) = \frac{\text{number of sample values} \leq t}{n} = \frac{1}{n} \sum_{i=1}^{n} 1_{\{X_i \leq t\}} \tag{6.1}$$

where 1_A is the indicator of event A. This empirical CDF is an estimate of the true CDF which can be found without making any assumption about the underlying distribution provided that it satisfies the definition of a CDF. The empirical CDF is not only an unbiased estimate of the population CDF but it is also a consistent estimator for the true CDF at any value of x [**105**]. The approximate derivative of

$F_n(t)$, refered to as the empirical density function, is as follows [**106**]:

$$g_n(x) = \frac{F_n(x+\lambda) - F_n(x-\lambda)}{2\lambda} \tag{6.2}$$

where $\lambda > 0$.

One of the packages in R which is used to plot the empirical density is the `Envstat` package [**78**]. The function `epdfplot` produces an empirical probability density function plot. When considering a discrete distribution, the empirical pdf plot is the same as the standard relative frequency histogram which means that each bar of the histogram denotes the proportion of the sample equal to that particular category. In the case of a continuous distribution, the function `epdfplot` calls the R function `density` to compute the estimated probability density at a number of evenly spaced points between the minimum and maximum values. The resulting empirical probability density function (epdf) plot is a graphical tool that could be used in conjunction with other graphical tools such as boxplots and histograms to estimate characteristics of the variable of interest (e.g. symmetry, middle, spread).

Another package in R which estimates an unknown density function is `logspline` [**66**]. In this method, the logarithm of the unknown density function is approximated by a polynomial spline, where the unknown coefficients are estimated using maximum likelihood. Also, logspline density estimation was developed to handle data that may be right, left or interval censored [**68**]. Consider estimating an unknown density function f based on sample data, $l = \log(f)$ could be estimated using a function of the form $\hat{l} = \hat{s} + c(\hat{s})$ where $c(\hat{s})$ is a normalizing constant such that $\int \exp(\hat{l}) = 1$ and the maximum likelihood method is used to choose $\hat{s}$ from a finite-dimensional linear space S of functions on $\mathbb{R}$. Therefore the corresponding density estimate is $\hat{f} = \exp(\hat{l})$ which is positive and integrates to one. If $\hat{s}$ is restricted to the subspace of S_0 of cubic splines, the corresponding logspline density is estimated. Assume Y is a random variable with positive and continuous density function. Let $Y_1, ..., Y_n$ be independent random variables having the same distribution as Y. The identifiable p-parameter exponential family $f(.;\theta), \theta \in \Theta$ of a positive twice differentiable density function on $\mathbb{R}$ is referred to as a logspline family. The log-likelihood

function related to the logspline family is given by

$$l(\theta) = \sum_i log(f(Y_i; \theta)) \tag{6.3}$$

The maximum likelihood estimate $\hat{\theta}$ is obtained by maximising the log-likelihood function and $\hat{f} = f(.; \hat{\theta})$ is referred to as the logspline density estimate [**67**].
An illustration of the difference between these two methods will be given by applying them to the illustrative datasets and the methods compared are also using a data simulated from a known distribution. Here 1000 data values are simulated from a Weibull distribution ($Weibull(3, 4)$). As the data are simulated from a known distribution, the `epdfplot` and `logspline` approaches can be compared to the actual density plot of ($Weibull(3, 4)$).
In Figure 6.1 a histogram of the simulated data from ($Weibull(3, 4)$) is displayed, and three density plots of the data are superimposed: these are (i) true population density of ($Weibull(3, 4)$) and the estimated density plots using (ii) `epdfplot` and (iii) `logspline`.
As the datasets that are discussed in this thesis relate to time to event studies area where the response of interest is a time, the domain of the density is the positive real line. As it can be seen in Figure 6.1, the estimated density function using the `epdfplot` function in `Envstat` package is not restrained to positive values. In order to solve the problem of negative values in a density plot generated using `epdfplot`, the logarithm of the response will be plotted, and then the empirical density plot is back transformed to the density plot of the variable on the raw scale.
By assuming $\nu = \log(y)$ instead of plotting the data y, the plot of $\exp(\nu)$ could be drawn using the following transformation formula:

$$\begin{aligned} y &= \exp(\nu) \\ f_y &= f_\nu \left| \frac{d\nu}{dy} \right| = \frac{f_\nu}{\exp(\nu)} \end{aligned} \tag{6.4}$$

So using `epdfplot` to draw $(\exp(\nu), \frac{f_\nu}{\exp(\nu)})$ and the related density plot is compared to the `logspline` and true Weibull density ($Weibull(3, 4)$) in Figure 6.2.
Based on Figure 6.2, the values of density estimates using `epdfplot` show different

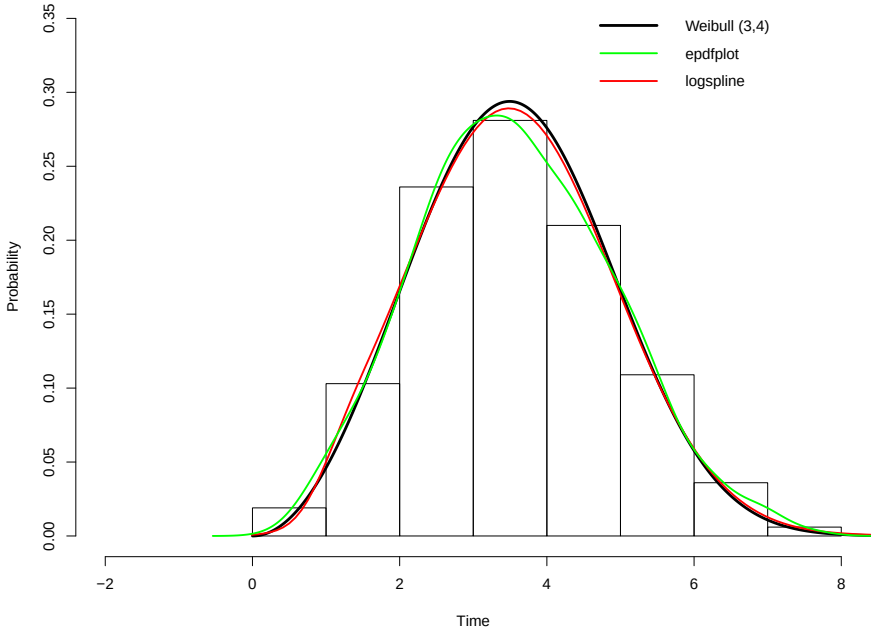

Figure 6.1: Density plot for dataset simulated from $Weibull(3, 4)$ using `logspline` package and `epdfplot` in `Envstat` package in R compared to true density of $Weibull(3, 4)$.

modes near zero. In general, the estimated density plot using the `logspline` package is closer to the true Weibull density in comparison to the `epdfplot`.

The logspline function applied to the response looks like a practical approach, albeit based on a simple example, for generating an estimate of a density function which can then be visualised.

In the following sections, this technique will be used to provide estimates of the underlying density function for the example datasets when censored observations are imputed using the proposed Bayesian methods developed in this thesis.

6.3 6-MP Data

The first dataset considered is the 6-MP dataset, introduced and described in Section 1.1.2 . Leukaemia patients were randomised to a treatment and a control group and the variable of interest was the duration of remission in weeks (i.e. the time until the cancer reoccurred). There are no censored observations in the control group, therefore, all of the graphical plots, including a density plot could be used

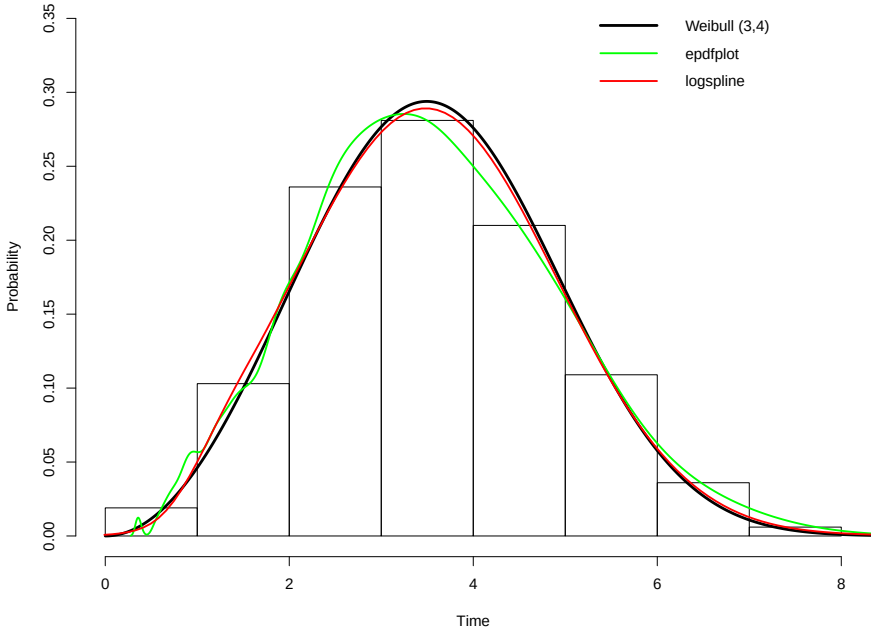

Figure 6.2: Density plot for dataset simulated from $Weibull(3, 4)$ using `logspline` package and `epdfplot` in `Envstat` package in R for transformed data using 6.4 compared to true density for $Weibull(3, 4)$.

as a visualising tool to summarise time in remission for the control group. In the treatment group, which contained 21 subjects, there are 12 censored observations i.e. more than half of the data. Due to presence of censoring in the intervention group, a boxplot or density plot of the variable of interest will be uninformative and biased if the censored observations are simply ignored. The classical approach is to plot a non-parametric estimate of the underlying survival function (i.e. 1- the CDF) typically using the Kaplan-Meier estimator (i.e. a Kaplan-Meier plot).

The approach taken is to impute the censored observations in the treatment group using the parametric Bayesian and non-parametric Bayesian techniques as introduced in Sections 4.5 and 4.6 and, using these estimates, a plot of the estimated density function can be drawn as a useful complement to the typical Kaplan-Meier plot.

Figure 6.3 compares the Kaplan-Meier estimated survivor function for both treatment and control group. It is apparent that the time to remission is shorter (i.e. worse prognosis) for the control group as the corresponding survivor function lies below the treatment group. An estimate of the median time to re-occurrence for

the intervention group is 23 weeks while the median time for the control group is 8 weeks, which is further evidence that the treatment is effective.

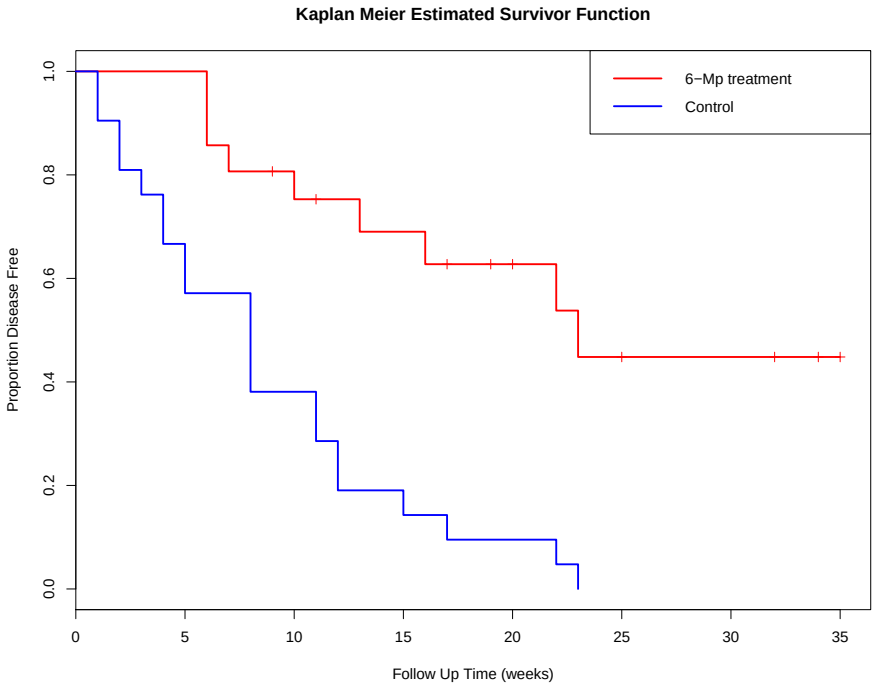

Figure 6.3: Kaplan-Meier plot for treatment and control group in 6-MP dataset.

For completeness, a plot of the estimated survivor functions for the control and treatment groups is given in Figure 6.4 where the censored observations have been imputed using the parametric and non-parametric Bayesian methods. The assumed distribution in the parametric approach to impute censored observations is a Weibull distribution.

Based on Figure 6.4, the estimates arising from the parametric Bayesian imputation and non-parametric Bayesian imputation methods are in good agreement with the Kaplan-Meier estimate. The real benefit in imputing the observations is when the complete data are used to generate graphical displays such as boxplot and the density plot. In Figure 6.5, boxplots of the time to re-occurrence for both groups are presented. When comparing the imputation methods, it can be seen the quantiles of treatment group for both parametric and non-parametric Bayesian approaches are nearly the same, but the tail is greater in the non-parametric Bayesian approach. It might be due to overestimation. As we mentioned in Section 5.7, if there is a high percentage of censoring, the non-parametric Bayesian approach can impute values which are much higher than the true failure times.

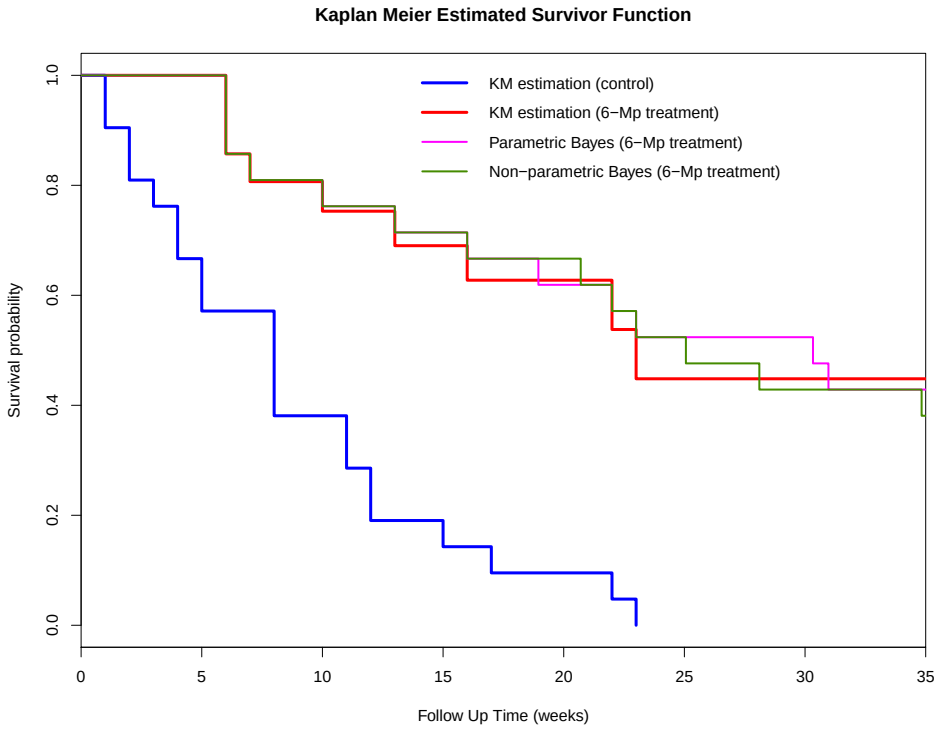

Figure 6.4: Kaplan-Meier plot for complete dataset in treatment group using parametric and non-parametric Bayesian approaches compared to the Kaplan-Meier of treatment and control group in 6-MP dataset.

As a complete dataset is available upon imputation a plot of the estimated density function can be made using the `logspline` function, for example. The plots, using the two imputation approaches, are given in Figure 6.6. There is little to choose between the imputation methods when comparing the plots for the treatment group (i.e. where censoring was present) so there is no substantial evidence against using a parametric assumption for the response distribution.

It could be argued that this plot may be easier to glean relevant information for the clinician administering the treatment and equally, if not more importantly, for the patient. The underlying distribution for time to re-occurrence for the controls is right skewed with an estimated mean of 8.6 weeks compared to the estimated median (based on the KM plot) of 8 weeks with estimated minimum of 1 and maximum of 23. The standard deviation is 6.46 and using the modified Chebycheff inequality, a rough estimate can be made to the range of time to occurrence that 75% of individuals are likely to experience. Therefore 75% of the population are likely to be in (0, 21.6). A more reliable estimate could be calculated by generating a 95% tolerance interval for 95% coverage using an appropriate transformation

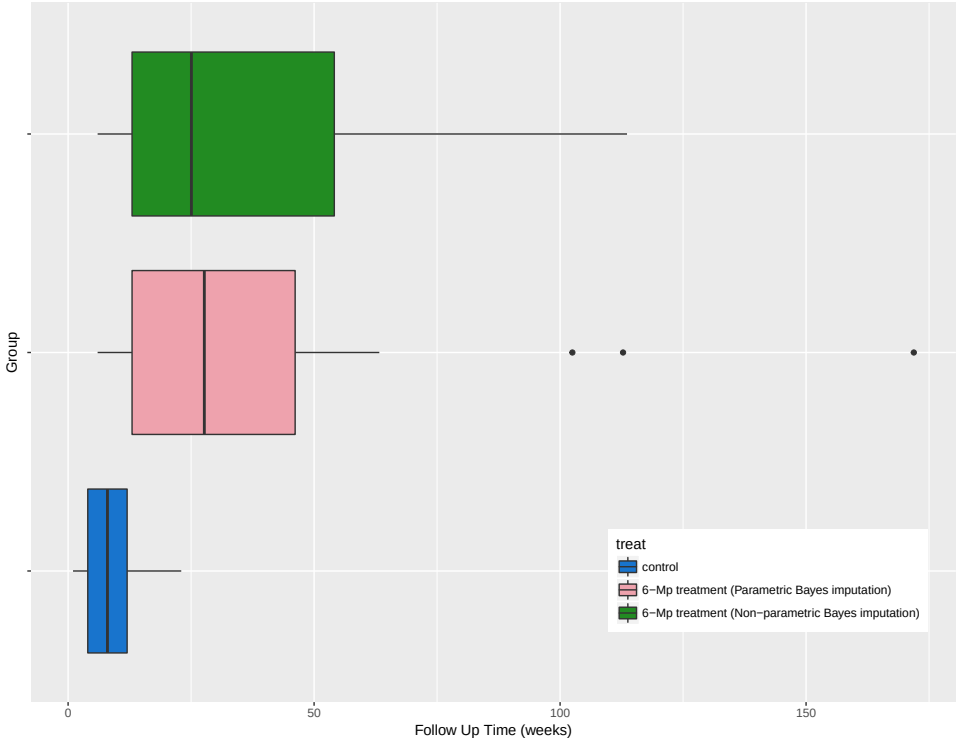

Figure 6.5: Comparing the boxplots of time to reoccurrence in the control and treatment group using the parametric and non-parametric Bayesian imputation approaches.

or non-parametric approach to account for the skewed nature of the distribution. Rather than providing an estimate of a parameter like a mean or median the 95% tolerance interval gives a range of values that apply to the individual in terms of what they are likely to experience. The tolerance interval in control group is (1, 23). Using the parametric Bayesian imputation, the distribution of the time to re-occurrence in the treatment group is shifted to the right (highlighting the treatment effect) with an estimated mean of 40.77 minimum of 6 and maximum of 171. The distribution is also right skewed with a larger standard deviation of 41.9 (compared to the controls) suggesting that considerably longer times to re-occurrence are plausible on treatment. Using modified Chebycheff's inequality, 75% of the population are likely to be in (0, 124.59) and a 95% tolerance interval for 95% coverage is (6,171) . The distribution also suggests that if the disease has not re-occurred by 20 weeks the time to re-occurrence may be quite a while off.

In conclusion, based on Figures 6.3, 6.5 and 6.6 the treatment group appears more effective than the control group as the distribution of time to re-occurrence is shifted

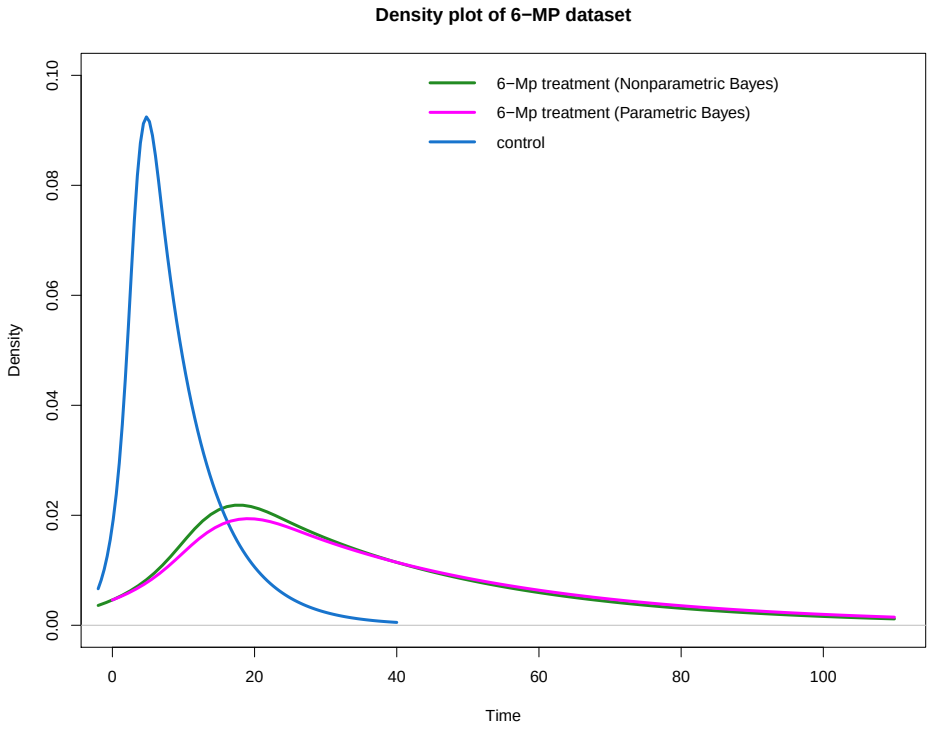

Figure 6.6: Comparing the density plot of control and treatment group using parametric and non-parametric Bayesian approaches to impute censored observations in treatment group.

in the positive direction with a heavier (right) tail suggesting that the typical time to re-occurrence is greater for those on treatment with a possibility of a considerably longer time in remission compared to the controls. An individual in the control groups is likely to have a remission time of between 1 to 23 while an individual on the 6-MP treatment is likely to have a remission time of between 6 to 171 (based on a 95% tolerance interval).

6.4 Metastatic Renal Carcinoma Data

The metastatic renal carcinoma data was introduced in Section 1.1.3 and will be used to further illustrate how the imputation of censored observations can be a useful translational tool. This dataset was also used by Royston [**93**] as an example of his parametric-based imputation approach for censored data. In total 347 patients were randomly assigned to two treatments, where 172 of them were treated with interferon-alpha and 175 of them with medroxyprogesterone acetate (MPA). The

variable of interest is time to death. At the end of the study, there were 8 censored observations in the MPA group and 17 censored observations in the interferon-alpha group, a relatively low degree of censoring.

The estimated survival functions for each group are presented in a Kaplan-Meier plot in Figure 6.7. The Kaplan-Meier estimates suggest an improvement in survival time for the patients in the interferon-alpha group compared to the standard MPA treatment, and there are only a few deaths observed after 48 months. This difference was deemed significant on the basis of a log rank test ($p = 0.008$). Based on Figure 6.7, the median survival of patients treated with interferon-alpha was 10 months, which was 3 months longer than the median for the MPA group.

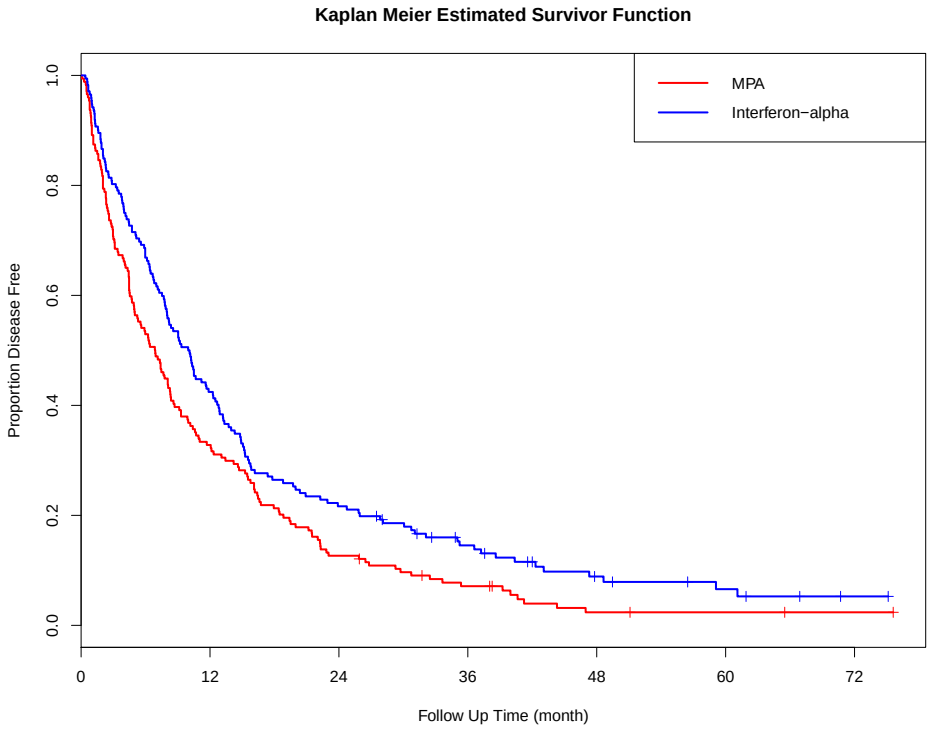

Figure 6.7: Kaplan-Meier curves for the survival of patients by treatment group in the RE01 trial.

The Kaplan-Meier plot (Figure 6.7) gives an indication of the conditional probability of survival over time rather than an estimate of the likely time to death in units of time. Our aim is to impute the censored observations to generate reliable estimates of the time to event in general and at the individual level; both of these estimates can be gained from plots of the underlying density function for survival time.

Using parametric Bayesian and non-parametric Bayesian approaches, we imputed

an estimated time of death for each patient with a censored survival time. On substituting an imputed value for each censored observations and combining these with the original data a complete dataset is formed.

Before focussing on the density plots, KM plots were created using the original and 'complete' data generated using the methods introduced in this thesis and also the Royston method . These plots provide a useful reference in terms of the influence the imputed observations have on the survivor function when compared to the unbiased estimate provided by the KM estimator using the censored data.

Figure 6.8, displays the survivor function estimated by the parametric Bayesian, non-parametric Bayesian and Royston parametric imputed values by treatment group. In the parametric Bayesian approach, a Weibull distribution is assumed for the data and a lognormal distribution for the Royston method as outlined in his paper [**93**]. The estimates appear comparable and in agreement.

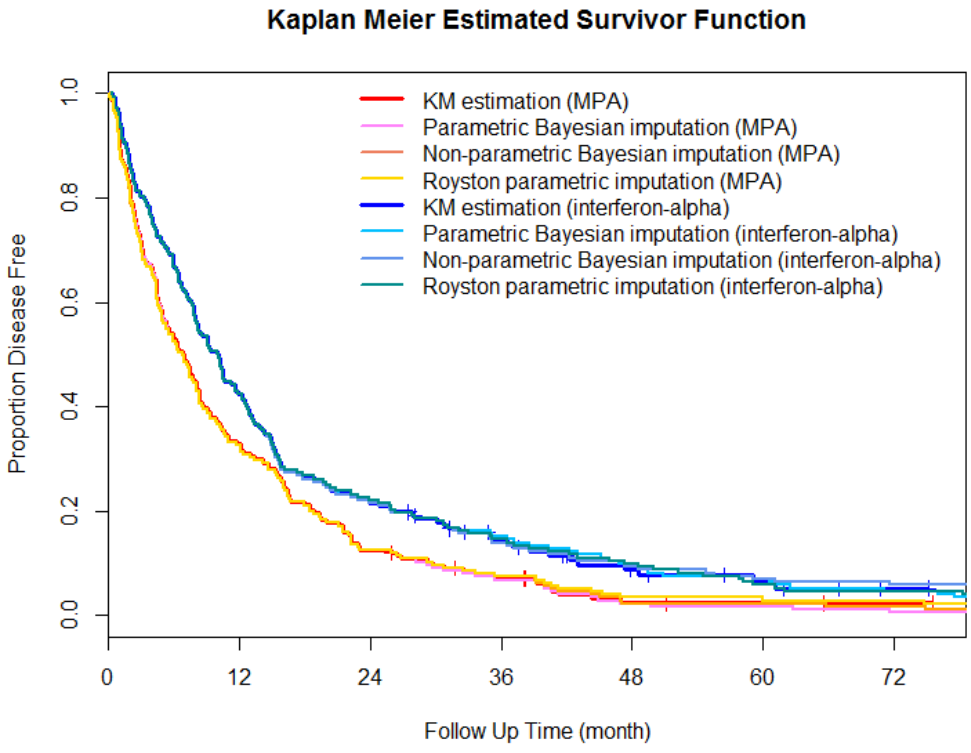

Figure 6.8: Kaplan-Meier plot by treatment group in renal carcinoma dataset before and after imputation using parametric Bayesian, non-parametric Bayesian and Royston parametric imputation approaches

Boxplots of the individual (observed and imputed) survival times are also in agree-

ment across the imputation methods (Figure 6.9). It can be seen that the median survival time amongst those on the interferon-alpha treatment is greater than the median in MPA group.

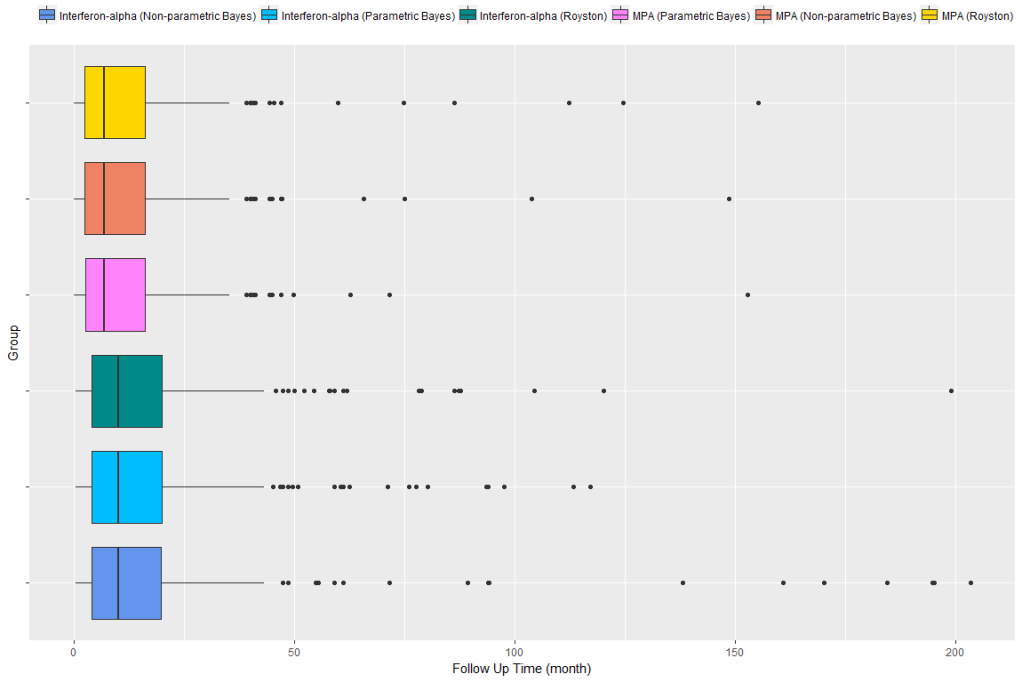

Figure 6.9: Boxplots of survival times for the interferon-alpha and MPA treatment groups after imputing censored observations using the parametric, parametric Bayesian and non-parametric Bayesian approaches.

A more informative interpretation is available when considering the variability in the individual survival times as evidenced by the density plot shown in Figure 6.10 using the `logspline` package. As in Figures 6.8 and 6.9 the parametric Bayesian, non-parametric Bayesian and Royston parametric methods are comparable; for brevity only the parametric Bayesian imputation method was used to estimate the underlying density.

The density plot in Figure 6.10 shows that the substantial difference between Kaplan-Meier curves for two treatment groups in Figure 6.7 which becomes more evident after 12 months in favour of the Interferon therapy, may correspond to a considerable overlap in the distribution of survival times. The mean survival time in MPA therapy is 12 while the survival mean in Interferon alpha treatment is 17.8.

Based on Figure 6.10, the density plots look similar at the start with a small difference between the two with a longer tail in the interferon-alpha treatment. Apparent differences at later follow up times in the Kaplan-Meier plot (Figure 6.7) may be due

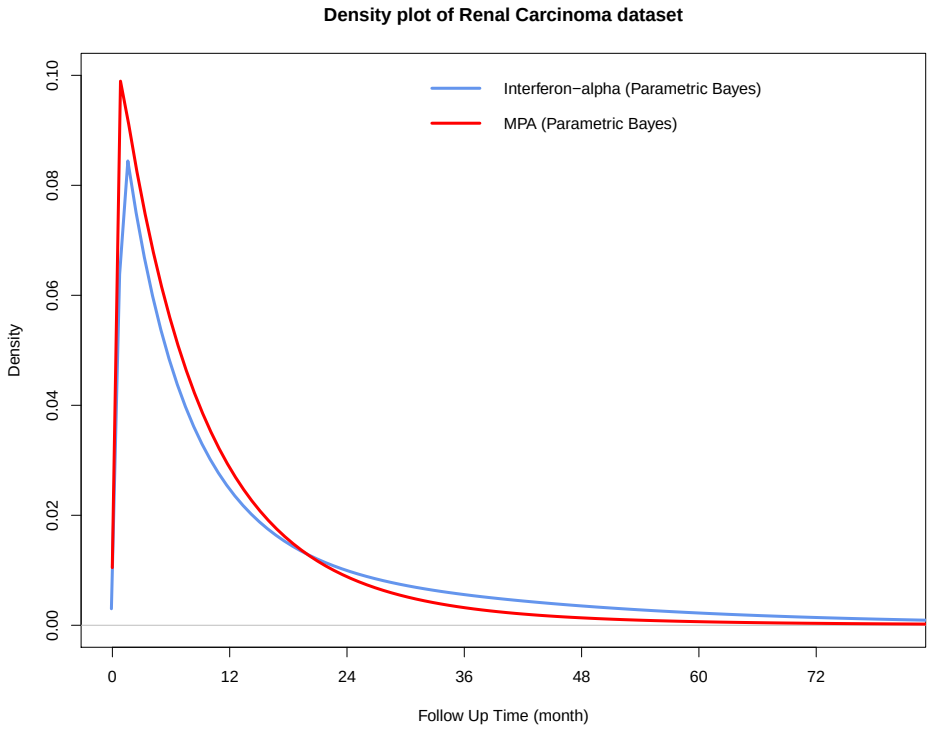

Figure 6.10: Comparing the density plot of the interferon-alpha and MPA treatment groups using the parametric Bayesian approach to impute censored observations.

to variability caused by smaller samples being observed rather than real differences. Obviously, as time progresses the number of patients at risk to calculate the survival estimates is decreasing as patients either experience the event or are censored. Alpha Blending from the `ggplot2` library [**111**] can be used to incorporate the number of patients at risk where the thickness of the line represents the diminishing number of patients at risk over time, the lighter the colour the lower the number of patients at risk. Figure 6.11 shows the Alpha Blending enhanced plot for the interferon-alpha and MPA groups. It is clear that as time increases the information for estimation of the Kaplan-Meier curves is much reduced, since many patients have experienced the event or been censored, and this is reflected in the lighter colour.

6.5 Bronchopulmonary Dysplasia (BPD) Data

The Bronchopulmonary Dysplasia (BPD) Data, described in Section 1.1.4, is the third dataset used to provide an illustration of the use of our imputation methods.

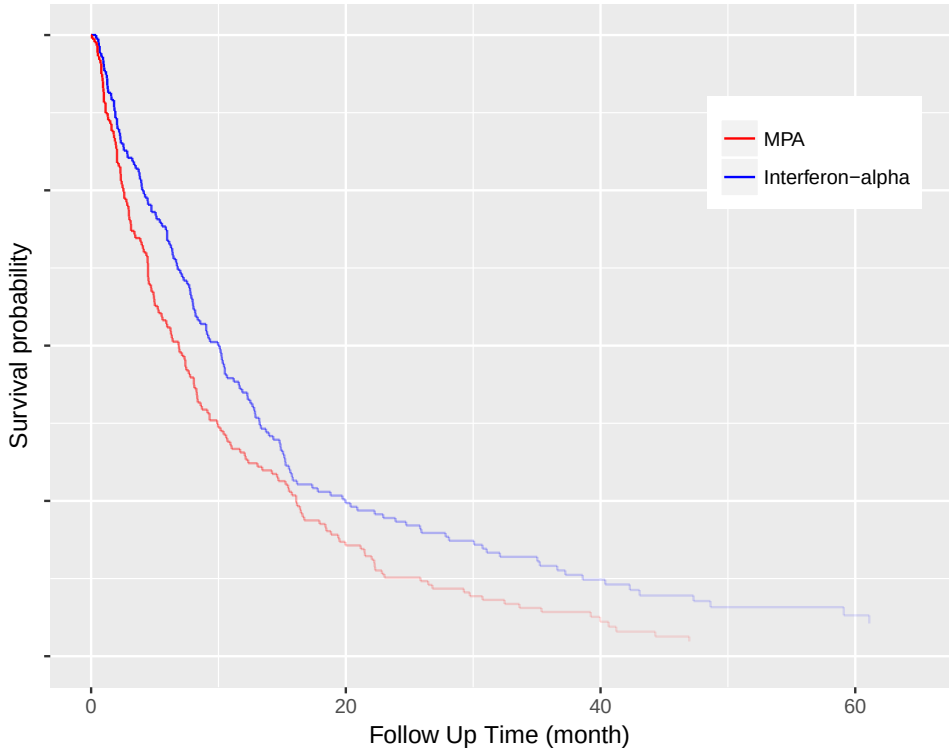

Figure 6.11: Kaplan-Meier curves for the survival of patients by treatment group in RE01 trial using alpha blending on the lines to show how many patients are at risk at the time.

This dataset relates to low-birth weight newborns and the total number of hours needed on (oxygen) treatment for a chronic lung disorder to abate. Note that low values are associated with a good outcome. Infants were randomised into either a treatment (i.e. surfactant therapy) or a control group. There are two censored observations in the treatment group of 35 infants, and three censored observations in the control group of 43 infants.

Based on the Kaplan-Meier plots (Figure 6.12), the estimated median number of hours on oxygen therapy for those infants who did not have surfactant therapy is 107 and the estimated median number of hours for those who had the therapy is 71, suggesting that surfactant is a beneficial therapy.

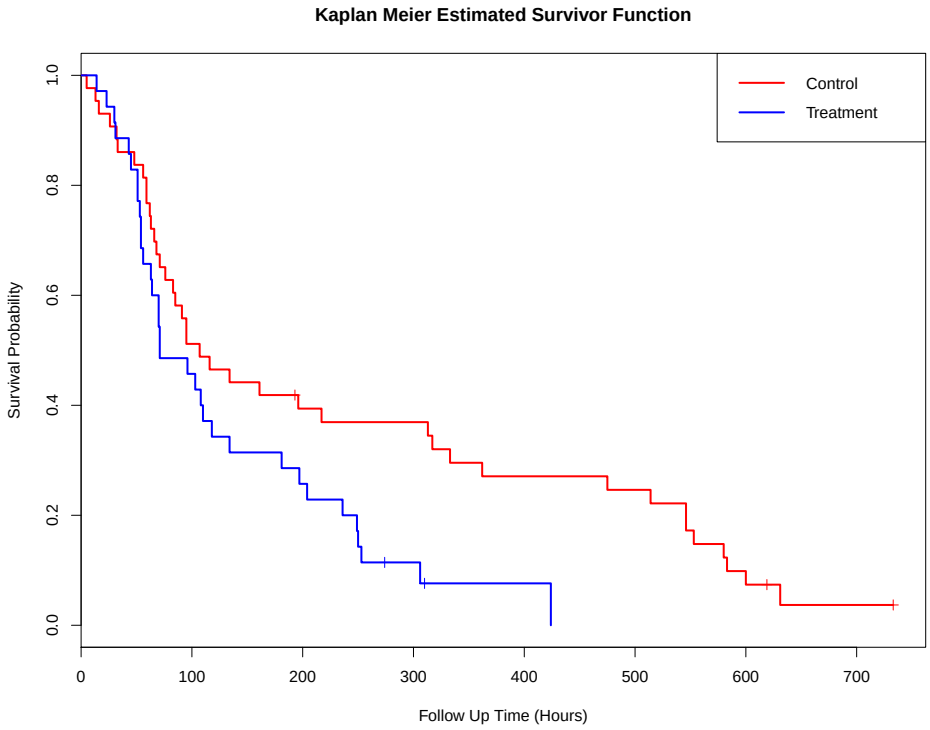

Figure 6.12: Kaplan-Meier plot for treatment and control group in the BPD dataset.

Although there are only a small number of censored observations in the dataset, ignoring them is not statistical best practice. Instead, censored observations in both groups were imputed using parametric Bayesian and non-parametric Bayesian approaches. As the percentage of censored observations in both treatment and control group is small, based on our simulation study the imputation results should be good estimates of their true unknown event times.

The Kaplan-Meier estimates of the survival function for the treatment and control groups, along with the corresponding ones using imputation, are shown in Figure 6.13. The assumed response distribution used in the parametric Bayesian imputation is the Weibull distribution. Based on Figure 6.13, in both the treatment and control group, the survivor estimates of the "complete" dataset after imputation using parametric Bayesian and non-parametric Bayesian methods are in agreement with the Kaplan-Meier estimate before imputation, which is as expected as there are only a small number of censored observations in each group. The boxplot of the completed datasets is shown in Figure 6.14 where the treatment effect is quite evident.

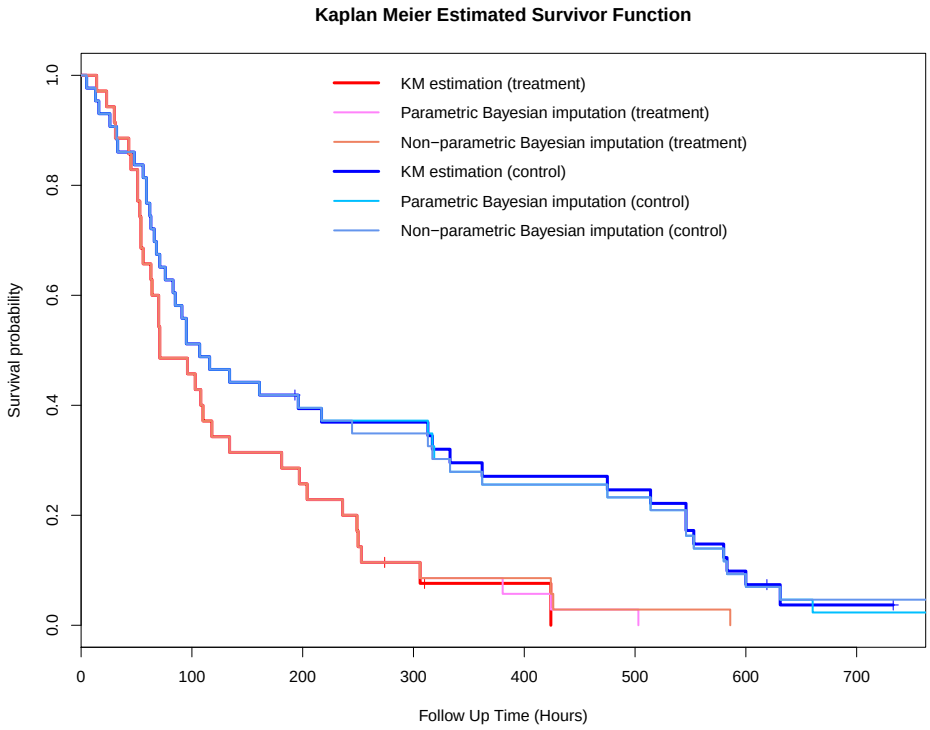

Figure 6.13: Kaplan-Meier plot for complete dataset in treatment and control group using parametric and non-parametric Bayesian approaches compared to the Kaplan-Meier estimates for the original data.

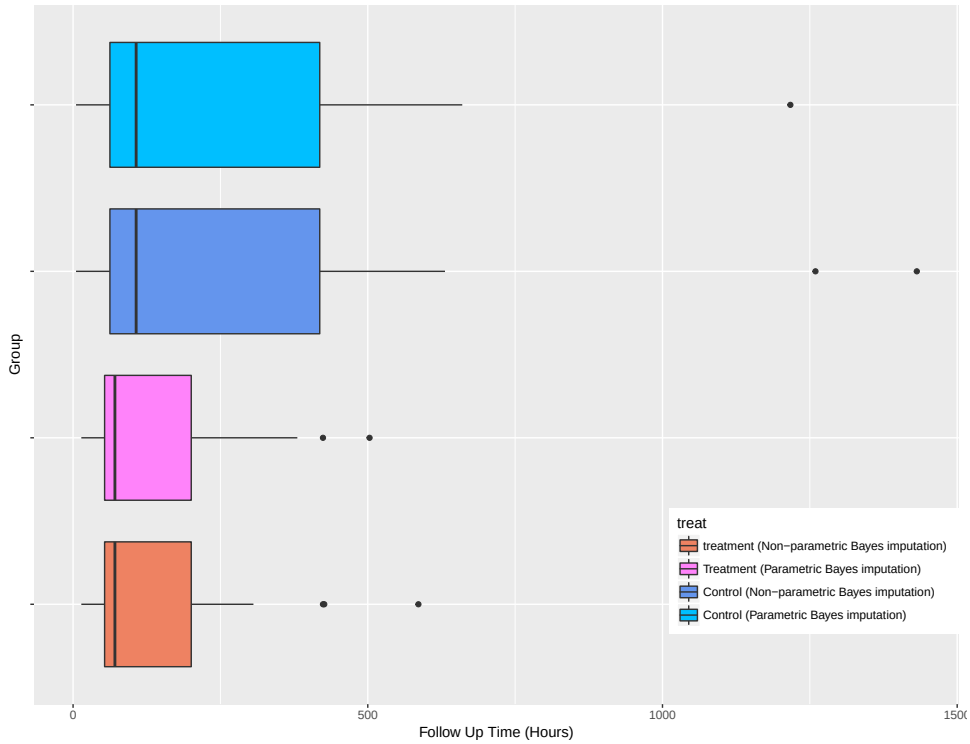

Figure 6.14: Boxplots of the survival time for the control and treatment groups after imputing censored observations using the parametric and non-parametric Bayesian approaches.

The density plot using the `logspline` package in R is shown in Figure 6.15. The densities look quite similar initially while a separation occurs beyond 100 hours. here the effect of treatment becomes more evident with the distribution of the time to come off treatment having a shorter tail than for the control group.

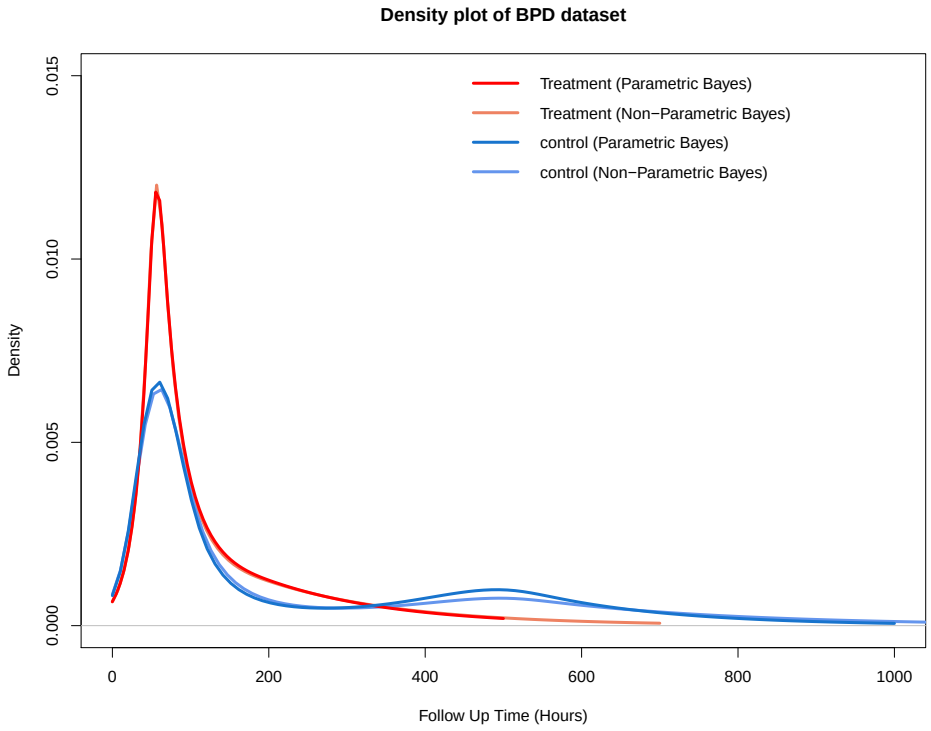

Figure 6.15: Comparing the density plot of cotrol and treatment group using the parametric and non-parametric Bayesian approaches to impute censored observations in both groups.

In summary, from Figures 6.12, 6.14 and 6.15 it is clear that surfactant therapy plays an important role in reducing the number of hours of oxygen therapy required among low birth weight infants who require treatment. A possible explanation for the results seen in the density plot is that there are two subgroups of infants, those who will recover quickly, for whom there is no real treatment effect, and a second group of sicker infants where the additional surfactant treatment is effective in reducing the time required on oxygen therapy.

6.6 Chapter Summary

The main focus of this chapter was to apply the parametric Bayesian and non-parametric Bayesian imputation methods to the three example datasets with different percentages of censoring introduced in Chapter 1. These imputation approaches are used to motivate the benefits of considering imputing censored observations in time to event studies. Also, we illustrate the usefulness of these methods in an applied context. By imputing values for the censored observations and combining the original complete and imputed incomplete data, it allows more interpretable graphics to be produced for a wider general audience (physicians and patients). For instance, it is possible to plot the density of the full data to complement the information given by Kaplan-Meier plots. These density functions were able to show features of the data that were very hard to discern from the Kaplan-Meier plot.
Also, in this chapter, a discussion on different methods of estimating the empirical density function was given where the `logspline` package in R was proposed as a useful approach for estimating a density function of the data.
The benefit of being able to generate an estimate of the underlying density function for the time to event response of interest was highlighted as a complement to plots of the survivor function and a potentially useful translational tool.

Chapter 7

Conclusions and Future Work

7.1 Goal of the Thesis and Proposed Methods

In time-to-event studies subject to censoring methods of plotting individual survival times, such as the histogram or the density plot, are not available and the graphical display of time-to-event data usually takes the form of a plot of the survivor function (typically using the Kaplan-Meier estimator). Based on the Kaplan-Meier plot, the median survival time is the classical summary reported to the patients. The median gives an estimate of the mid-population survival time for the cohort in question and is unlikely to apply to any particular individual. Moreover, the median may not be unique, and in some datasets, it cannot be calculated.

The principal aim of this thesis was to consider censored data as a form of missing, incomplete, data and to propose Bayesian approaches to impute these partially observed values. In this thesis, the imputed values of censored observations were used to produce more interpretable graphical summaries of time-to-event data, such as a density plot, which may usefully complement Kaplan-Meier plots. The imputation approach is intended to be used for the visual exploration and presentation of survival data and give a simple, interpretable display for physicians and patients to better understand summaries generated from time to event models.

The first new approach taken in this thesis was to use a parametric Bayesian framework to impute the censored observations. The Bayesian perspective can be in-

teresting as maybe historical data from similar past studies can be very helpful in interpreting the results of the current study, or used to inform prior distributions for model parameters in the analysis. A second new approach is to use non-parametric Bayesian methods for imputation, as assuming a fixed distributional specification for the random or error terms in the model may be inadequate for the actual data. Mainly, we used the idea of imputing the censored observations based on the other information in the dataset and some form of Bayesian model. However, in imputing censored data, we have some additional limited information on the censored values. For instance, with right censoring, we know that the true failure time exceeds the observed censored time. Knowing this information is an important difference from standard missing data imputation, and we used this knowledge in our imputations to provide imputed values making use of all available information. By repeating the imputation process we are able to obtain some indication of the uncertainty due to the censored (partially observed) values.

7.2 Review of Simulation and Application Studies

To study the performance of our approach we carried out simulation studies, using both parametric Bayesian and non-parametric Bayesian approaches to impute the censored observations. The simulation study results were obtained for different sample sizes and different percentages of censoring.

Based on our simulation results, if the sample size is large enough (for example greater than a hundred) and also in the presence of at most a moderate percentage of censoring (not more than twenty percent) both the parametric Bayesian and non-parametric Bayesian imputation methods work quite well when estimating a survivor function and are comparable to the Kaplan-Meier estimator. For both parametric and non-parametric Bayesian methods, as the sample size increased the imputed values became more accurate, but for a given sample size become less so as the proportion of censoring increases. This shows that the approach behaves as may be expected under conditions of more, or less, information.

Also as expected, when the true survival distribution is known the parametric approach performed better and so should be used when possible. For example, in tuberculosis disease, since the potential for dying increases early in the disease and then decreases later, the lognormal distribution might be an appropriate distribution. However, in situations where the true distribution of the data is unknown, as is often the case in practice, the non-parametric Bayesian approach is recommended and potentially superior, but this needs to be balanced against the fact that it is much more computationally intensive and more time-consuming. In general, these imputation methods are not recommended for a high percentage of censoring, as when half or more of the data are censored there is insufficient information in the observed failure times to provide meaningful imputations.

The parametric Bayesian and non-parametric Bayesian methods were applied to three datasets including 6-MP data, metastatic renal carcinoma data and bronchopulmonary dysplasia data. These datasets are chosen because they have different percentages of censoring. The benefit of being able to generate an estimate of the underlying density function for the time to event response of interest was highlighted as a complement to plots of the survivor function and a potentially useful translational tool. These density functions were able to show features of the data that were very hard to discern from the Kaplan-Meier plot.

7.3 Future Work

The development of the imputation ideas proposed in this thesis opens up new opportunities for further research. The simulation study will be extended to compare and contrast parametric Bayesian and non-parametric Bayesian methods for different sample sizes using mis-specified models. An R package will be developed to impute censored observations using both parametric Bayesian and non-parametric Bayesian methods, which would make the methods developed here accessible for others to use. The current development has been confined to right-censoring, however, the basic idea could be extended to interval and left censoring through appropriate changes in the conditional distributions used for the imputation. Other censoring

schemes, such as Type II censoring, where there are a fixed specified number of failures, could also be considered by imputing suitable conditional order statistics for the censored values.

Another application of the imputation approach could be in estimating the mean residual life (MRL) function, which, it has recently been argued may be an informative alternative summary to the survival function, especially when communicating to a non-statistical audience (see Newell et al. [**87**], Alvarez et al. [**5**], Jalali et al. [**59**]). When there is no censored data the MRL estimation would be straightforward using the empirical estimate for the MRL function [**112**]. The presence of censoring is the main challenge in estimating the MRL function since the survivor function needs to be known. Alvarez et al. [**5**] proposed a method to estimate the MRL function for right censored data using a hybrid estimator. By using Bayesian imputation methods for right censored data a complete dataset can be created so that the MRL function could be estimated using an empirical estimator.

Finally, the use of Bayesian imputation methods may lead to more flexibility in how time to event data are modelled and analysed, as methods for modelling a continuous response may now be applicable. In this the censored data would again be treated as missing and imputed to give a completed dataset for analysis and the usual consideration for analyses with imputed data would apply, such as the use of multiple imputations and associated analyses with the results being combined by Rubin's rules.

Appendix: Rcode

Shiny Application for Sampling From a Dirichlet Process

```
library("shiny")
runApp(list(
ui = bootstrapPage(pageWithSidebar( headerPanel(uiOutput("headerp")),
sidebarPanel(radioButtons("method", "Method:",
c("Stick breaking" = "stick","Polya urn" = "polya")),
sliderInput("N", "Number of N:",min = 1,max = 1000, value = 100),
sliderInput("nsim", "number of simulation:",min=1, max=100, value=50),
sliderInput("alpha", "alpha:",min = 0.1,max = 100, value = 20, step= 0.1),
sliderInput("w", "weight:",min = 0,max = 1, value = 1, step= 0.01),
numericInput("mu1", "First mean:", 0),numericInput("mu2", "Second mean:", 0),
numericInput("sigma1", "First standard deviation:", min=0, 1),
numericInput("sigma2", "Second standard deviation:", min=0, 1),
submitButton("Refresh")),mainPanel( plotOutput("cum.plot"))
)),
server = function(input, output){
library("dynpred")
library("splines")
library("survival")
output$headerp <- renderUI({
hpanel=list(paste("Dirichlet Process {G_0=",
```

```
input$w, "* N(", input$mu1, ",", input$sigma1, ") + ",
1-input$w, "* N(", input$mu2, ",", input$sigma2, ")}", sep=" "))
do.call(tagList, hpanel)
})
LL = 200
output$cum.plot <- renderPlot({
w=input$w
mu1=input$mu1
mu2=input$mu2
sigma1=input$sigma1
sigma2=input$sigma2
nsim=input$nsim
alpha<-input$alpha
N<-input$N
rng= c( min(mu1-4*sigma1, mu2-4*sigma2), max(mu1+4*sigma1, mu2+4*sigma2) )
gen.w = function( N, alpha ){
zz = rbeta( N, 1, alpha )
ww = zz
temp = 1 - zz[1]
for( k in 2:(N-1) ){
ww[k] = temp * zz[k]
temp = temp * (1 - zz[k])
}
ww[N] = 1 - sum( ww[1:(N-1)] )
return( sample( ww ))
}
sample.mix.norm.g0 = function( N, alpha, LL=200 ){
ww = gen.w( N, alpha )
N1 = round( w * N )
psi = c( rnorm( N1, mean=mu1, sd=sigma1 ), rnorm( N-N1, mean=mu2, sd=sigma2))
cpsi = sort( psi, index.return=T )
cdf = cumsum( ww )
```

```
cbind( cpsi$x, cdf )
}
mean.cdf = function( alpha, nsim, LL=200, N=1000, rng=c(-4,4) ){
newtm  = seq( rng[1], rng[2], len = N )
newcdf = matrix( 0, nrow=nsim, ncol=N )
for( ii in 1:nsim ) {
res = sample.mix.norm.g0( N, alpha, LL )
newcdf[ii,] = evalstep( time = res[,1],
stepf = res[,2],
newtime = newtm,
subst = 0)
}
meancdf = apply( newcdf, 2, mean )
list( tm=newtm, cdf=newcdf, mcdf=meancdf )
}
if (input$method=="stick"){
label.plot<-"Stick breaking"
res = mean.cdf( alpha, nsim, LL=200, N=N, rng=rng  )
plot( res$tm, res$mcdf, xlim = rng, ylab="P(X<x)", xlab="x", col="red",
 lwd=2, type='l' )
grayc = seq( .5, .9, len=20 )
for( j in 1:nsim ){
for( i in 1:(N-1) ) {
segments( res$tm[i], res$cdf[j,i],
res$tm[i+1], res$cdf[j,i], col = gray( grayc[j%%20]) )  # horiz
segments( res$tm[i+1], res$cdf[j,i],
res$tm[i+1], res$cdf[j,i+1], col = gray( grayc[j%%20]) )  # horiz
}
if( j > 20 ) break
}
x = seq( rng[1], rng[2], len=200 )
lines( x, w*pnorm(x, mean=mu1, sd=sigma1) + (1-w)*pnorm(x, mean=mu2, sd=sigma2),
```

```
 col = "black", lwd=5, type='l' )
lines( res$tm, res$mcdf, xlim = rng, ylab="P(X<x)", col="red",
 lwd=2, type='l' )
 }
if (input$method=="polya"){
label.plot<-"Polya urn"
mat.x<-NULL
x = seq(rng[1], rng[2],len=200)
plot(x, w*pnorm(x, mean=mu1, sd=sigma1) + (1-w)*pnorm(x, mean=mu2, sd=sigma2),
 ylab="P(X<x)", col="black", lwd=5, type='l')
grayc = seq( .5, .9, len=20 )
for(j in 1:nsim){
theta = rep(0,N)
rn1<-rnorm(1,mu1,sigma1)
rn2<-rnorm(1,mu2,sigma2)
u1<-runif(1)
if (u1<w) {theta[1]=rn1} else {theta[1]=rn2}
theta.star = rep(0,N)
theta.star[1] = theta[1]
for( ii in 2:N ) {
uu = runif( 1, 0, 1 )
if( uu < alpha/(alpha+ii-1)) {
rn1<-rnorm(1,mu1,sigma1)
rn2<-rnorm(1,mu2,sigma2)
u1<-runif(1)
if (u1<w) {theta[ii]=rn1} else {theta[ii]=rn2}
}
else {
theta[ii] = sample( theta[1:(ii-1)], 1, 1/(alpha+ii-1) )
}}
thetau = unique( theta )
L = length( thetau )
```

```
thetap = rep(0,L)
for( ii in 1:L ) {
thetap[ii] = sum( thetau[ii] == theta ) / N
}
stheta = sort( thetau, index.return=T )
cdf = cumsum( thetap[ stheta$ix ] )
res = list( stheta = stheta$x, cdf=cdf )
g.x<-evalstep(time=res$stheta, stepf=cdf, newtime=x, subst=0)
mat.x<-cbind(mat.x,g.x)
if(j<21) {
segments( x[1], 0, res$stheta[1], 0, lwd=.02,col = gray( grayc[j%%20]) )
segments( res$stheta[1], 0, res$stheta[1], res$cdf[1], lwd=.02,
col = gray( grayc[j%%20]) )
for( i in 1:L ) {
segments( res$stheta[i], res$cdf[i],
res$stheta[i+1], res$cdf[i], lwd=.02,col = gray( grayc[j%%20]) )
segments( res$stheta[i+1], res$cdf[i],
res$stheta[i+1], res$cdf[i+1], lwd=.02,col = gray( grayc[j%%20]) )
}
segments( res$stheta[L], res$cdf[L], res$stheta[L], 1, ,
 lwd=.02,col = gray( grayc[j%%20]) )
segments( res$stheta[L], 1, x[LL], 1, , lwd=.02,col = gray( grayc[j%%20]) )
}}
mean.mat<-apply(mat.x,1,mean)
for(ii in 2:200){
segments(x[ii-1], mean.mat[ii-1], x[ii-1], mean.mat[ii],lwd=2,col="red")
segments(x[ii-1], mean.mat[ii], x[ii], mean.mat[ii],lwd=2,col="red")
}}})}))
```

Location Normal Dirichlet Process Mixture Model

```
update.configs = function( new.theta, ii )
{
old.theta.ii = theta[ii]
theta[ii] <<- new.theta
yy.nn=length(theta)
n.star=length(theta.star)
clust.jj = ( new.theta == theta.star )
if( length(which(clust.jj)) == 0 ){
if( n.star == yy.nn ) {
theta.star[ s.indic[ii] ] <<- new.theta
return}
else{
if( nj[ s.indic[ii] ] == 1 ){
theta.star[s.indic[ii]] <<- new.theta}
else{
theta.star[ n.star+1 ] <<- new.theta
nj[s.indic[ii]] <<- nj[s.indic[ii]] - 1
nj[n.star+1] <<- 1
s.indic[ii] <<- n.star+1
n.star <<- n.star + 1}}}
else{
if( new.theta != old.theta.ii ){
theta.star <<- sort( unique( theta ))
s.indic <<- match( theta, theta.star )
n.star <<- length( theta.star )
temp = table( theta )
nj <<- as.vector( temp )}
else{
}}}
update.m.configs = function( ii ){
```

```
theta.star.m <<- theta.star
nj.m <<- nj
n.star.m <<- n.star
if( nj.m[ s.indic[ii]] == 1 ){
theta.star.m <<- theta.star.m[ -s.indic[ ii ]]
nj.m <<- nj.m[ -s.indic[ ii ]]
n.star.m <<- n.star.m - 1}
else{
nj.m[ s.indic[ii] ] <<- nj.m[ s.indic[ii] ] - 1}}
fullcond.theta = function( iter ){
qj = rep( 1, n.star.m )
for( ii in 1:yy.nn ){
update.m.configs( ii )
temp = exp( -0.5*( yy[ii] - mu)^2 / (phi2 + tau2 ))
q0 = alpha.DP / ( SQRT2PI * sqrt(phi2+tau2)) * temp
qj = dnorm( yy[ii], mean = theta.star.m, sd = sqrt( phi2 ))
sum.q.all = sum( nj.m * qj )
uu = runif( 1, 0, 1 )
if ( uu <= q0 / (sum.q.all+ q0)){
mu.h = (yy[ii]*tau2 + mu*phi2) / (tau2+phi2)
tau2.h = phi2 * tau2 / (tau2+phi2)
new.theta = rnorm( 1, mu.h, sqrt( tau2.h ))}
else{
temp = nj.m * qj
idx = sample( 1:n.star.m, size=1, prob = nj.m * qj/ (sum.q.all+ q0)) #)
new.theta = theta.star.m[ idx ]}
update.configs( new.theta, ii )}
theta.out[ iter, ] <<- theta
theta.star.out[[iter]] <<- theta.star
nj.out[[iter]] <<- nj}
fullcond.theta.star = function( iter ){
for( jj in 1:n.star ){
```

```
yy.sum = sum( yy[ (s.indic == jj) ])
temp1 = nj[jj] * tau2 + phi2;
mu.theta.star = (mu*phi2 + yy.sum*tau2) / temp1;
s2.theta.star = tau2 * phi2 / temp1;
theta.star[jj] = rnorm( 1, mu.theta.star, sqrt(s2.theta.star) );}
for( ii in 1:yy.nn ){
theta[ii] <<- theta.star[s.indic[ii]];}
theta.out[ iter, ] <<- theta
theta.star.out[[iter]] <<- theta.star}
fullcond.alpha = function( iter ){
uu = rbeta( 1, alpha.DP +1.0, yy.nn )
ww = (a.alpha.DP + n.star -1) /
( yy.nn*(b.alpha.DP - log(uu)) + alpha.DP + n.star -1 )
if( runif( 1, 0, 1 ) < ww ){
alpha.DP <<- rgamma( 1, shape = a.alpha.DP + n.star,
rate = b.alpha.DP -log(uu) )}
else{
alpha.DP <<- rgamma( 1, shape = a.alpha.DP + n.star -1,
rate = b.alpha.DP -log(uu) )}
alpha.DP.out[ iter ] <<- alpha.DP}
fullcond.phi2 = function( iter ){
temp.sum = sum( (yy-theta)^2 )
phi2 <<- 1.0 / rgamma( 1, shape = c.phi2 + 0.5*yy.nn,
rate = d.phi2 + 0.5*temp.sum )
phi2.out[ iter ] <<- phi2}
fullcond.mu = function( iter ){
theta.star.sum = sum( theta.star )
temp1 = tau2 + n.star*var.mu
mu.res = (mu.mu*tau2 + var.mu*theta.star.sum) / temp1
s2.res = tau2*var.mu / temp1
mu <<- rnorm( 1, mu.res, sqrt(s2.res) )
mu.out[ iter ] <<- mu}
```

```
fullcond.tau2 = function( iter ){
temp.sum = sum( (theta.star - mu )^2 )
tau2 <<- 1.0 / rgamma( 1, shape = c.tau2 + n.star/2,
rate = d.tau2 + 0.5*temp.sum )
tau2.out[ iter ] <<- tau2}
init.all = function( nnn ){
theta <<- rep( mean( yy ), yy.nn )
s.indic <<- rep( 1, yy.nn )
theta.star <<- theta[1]
n.star <<- 1
nj <<- yy.nn
theta.star.m <<- theta[1]
n.star.m <<- 1
nj.m <<- yy.nn-1
alpha.DP <<- 0.1 * yy.nn
mu <<- mean( yy )
tau2 <<- 10*var( yy )
phi2 <<- 10*var( yy )
SQRT2PI <<- sqrt( pi )
theta.out <<- matrix( 0, nrow=nnn, ncol=yy.nn )
theta.star.out <<- list()
nj.out <<- list()
alpha.DP.out <<- rep( 0, nnn )
mu.out <<- rep( 0, nnn )
phi2.out <<- rep( 0, nnn )
tau2.out <<- rep( 0, nnn )}
gibbs.sampler = function( iter ){
fullcond.theta( iter )
fullcond.theta.star( iter )
fullcond.alpha( iter )
fullcond.phi2( iter )
fullcond.mu( iter )
```

```
fullcond.tau2( iter )}
read.input.files =function( fname.par, fname.dat ){
dd = read.table( fname.par )
burnin <<- dd[ dd[,1] == "burnin", 2 ]
monitor <<- dd[ dd[,1] == "monitor", 2 ]
thin <<- dd[ dd[,1] == "thin", 2 ]
a.alpha.DP <<- dd[ dd[,1] == "a.alpha.DP", 2 ]
b.alpha.DP <<- dd[ dd[,1] == "b.alpha.DP", 2 ]
c.phi2 <<- dd[ dd[,1] == "c.phi2", 2 ]
d.phi2 <<- dd[ dd[,1] == "d.phi2", 2 ]
c.tau2 <<- dd[ dd[,1] == "c.tau2", 2 ]
d.tau2 <<- dd[ dd[,1] == "d.tau2", 2 ]
yy <<- as.matrix( read.table( fname.dat ))
yy.nn <<- dim( yy )[1]
yy <<-  yy[,1] / 1000
mu.mu <<- mean(yy)
var.mu <<- var(yy) / sqrt( yy.nn )
if( 0 ) {
print( yy )
print( burnin )
print( monitor )
print( thin )
print( a.alpha.DP )
print( b.alpha.DP )
print( c.phi2 )
print( d.phi2 )
print( c.tau2 )
print( d.tau2 )
print( theta )
print( s.indic )
print( theta.star )
print( n.star )
```

```
print( nj )
print( theta.star.m)
print( n.star.m )
print( nj.m )
print( alpha.DP )
print( mu )
print( tau2 )
print( phi2 )}
post.pred = function(){
y0 = seq( min(yy), max(yy), len = 100 )
y0.nn = length( y0 )
y0.pred = rep( 0, y0.nn )
temp = rep( 0, monitor )
for( ii in 1:y0.nn ) {
for( ll in (burnin+1):(burnin+monitor)) {
temp[ll] = sum( nj.out[[ll]] *
dnorm( y0[ii], theta.star.out[[ll]], sqrt(phi2.out[ll])) )
}
y0.pred[ii] =
mean( alpha.DP.out / (alpha.DP.out + yy.nn ) *
dnorm( y0[ii], mu.out, sqrt( tau2.out + phi2.out )) +
1/(alpha.DP.out + yy.nn) * temp )
}
list( y0 = y0, y0.pred = y0.pred )}
main.mcmc = function( fname.par, fname.dat ){
read.input.files( fname.par, fname.dat )
n.run = burnin + monitor
init.all( n.run )
for ( ii in 1:n.run )
{
if (ii %% (20) == 0)
print(ii)
```

```
gibbs.sampler( ii )}}
dinvgamma = function( x, shape, rate ){
tt = shape*log(rate) - (shape+1)*log(x) -rate/x -lgamma(shape)
return( exp(tt) )}
##if( 0 )
{
system.time( replicate( 1,
main.mcmc( "dpm.par.txt", "galaxies.txt" )))
}
####if( 0 )
{
res = post.pred()
y0.axis = ( res$y0 )
y0.pred = ( res$y0.pred )
par( mfrow=c(2,3) )
acf( theta.out[,38] )
acf( alpha.DP.out )
acf( mu.out )
acf( phi2.out )
acf( tau2.out )
dev.new()
par( mfrow=c(2,3) )
hist( theta.out[,38], prob=T )
hist( alpha.DP.out, prob=T )
x = seq( min( alpha.DP.out ), max(alpha.DP.out), len = 100)
lines( x, dgamma( x, shape = a.alpha.DP, rate = b.alpha.DP ), col=2 )
hist( mu.out, prob=T )
x = seq( min( mu.out ), max(mu.out), len = 100)
lines( x, dnorm( x, mean = mu.mu, sd = sqrt( var.mu )), col=2)
hist( phi2.out, prob=T )
x = seq( min( phi2.out ), max(phi2.out), len = 100)
lines( x, dinvgamma( x, shape = c.phi2, rate = d.phi2 ), col=2 )
```

```
hist( tau2.out, prob=T )
x = seq( min( tau2.out ), max(tau2.out), len = 100)
lines( x, dinvgamma( x, shape = c.tau2, rate = d.tau2 ), col=2 )
hist( yy, prob=T, breaks=30 )
lines( y0.axis, y0.pred, col=2 )
}
hist( yy, prob=T, breaks=30 ,main = "",ylim=c(0,.25),xlab="Velocity")
lines( y0.axis, y0.pred, col=2 )
```

Royston Parametric Approach

```
dat.KC<- data.frame(time=mydata$survtime, status=mydata$cens, prior=mydata$trt)
library(survival)
fit <- survreg(Surv(time,status) ~ prior, data = dat.KC, dist = "lognormal")
summary(fit)
mu_i <- log(predict(fit, dat.KC[dat.KC[, "status"] == 0, ]))
sigm <- fit$scale
k <- (log(dat.KC[dat.KC[, "status"] == 0, "time"]) - mu_i)/sigm
le <- length(k)
u <- runif(le, 0, 1)
F_k <- pnorm(k, 0, 1)
x_i <- qnorm(u * (1 - F_k) + F_k, 0, 1)
tau_i = exp(sigm * x_i + mu_i)
dat.KC[dat.KC[, "status"] == 0, "time"] <- tau_i
Royston<- dat.KC
```

Bibliography

[1] Odd Aalen. Nonparametric estimation of partial transition probabilities in multiple decrement models. *The Annals of Statistics*, pages 534–545, 1978. (Cited on page 23.)

[2] Odd Aalen. Nonparametric inference for a family of counting processes. *The Annals of Statistics*, pages 701–726, 1978. (Cited on page 23.)

[3] MI Ageel. A novel means of estimating quantiles for 2-parameter Weibull distribution under the right random censoring model. *Journal of Computational and Applied Mathematics*, 149(2):373–380, 2002. (Cited on page 68.)

[4] Murray Aitkin and Donald B Rubin. Estimation and hypothesis testing in finite mixture models. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 67–75, 1985. (Cited on page 38.)

[5] Alberto Alvarez-Iglesias, John Newell, Carl Scarrott, and John Hinde. Summarising censored survival data using the mean residual life function. *Statistics in Medicine*, 34(11):1965–1976, 2015. (Cited on page 151.)

[6] Charles E Antoniak. Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The Annals of Statistics*, pages 1152–1174, 1974. (Cited on pages 45, 50, and 88.)

[7] Amanda N Baraldi and Craig K Enders. An introduction to modern missing data analyses. *Journal of School Psychology*, 48(1):5–37, 2010. (Cited on page 63.)

[8] Thomas Bayes, Richard Price, and John Canton. *An essay towards solving a problem in the doctrine of chances*. C. Davis, Printer to the Royal Society of London, 1763. (Cited on page 28.)

[9] Ralf Bender, Thomas Augustin, and Maria Blettner. Generating survival times to simulate Cox proportional hazards models. *Statistics in Medicine*, 24(11):1713–1723, 2005. (Cited on page 105.)

[10] Joseph Berkson and Robert P Gage. Survival curve for cancer patients following treatment. *Journal of the American Statistical Association*, 47(259):501–515, 1952. (Cited on page 14.)

[11] David Blackwell and James B MacQueen. Ferguson distributions via Pólya urn schemes. *The Annals of Statistics*, pages 353–355, 1973. (Cited on page 43.)

[12] Ron Brookmeyer and John Crowley. A confidence interval for the median survival time. *Biometrics*, pages 29–41, 1982. (Cited on page 67.)

[13] Jonathan Buckley and Ian James. Linear regression with censored data. *Biometrika*, 66(3):429–436, 1979. (Cited on page 68.)

[14] Andrea Burton, Douglas G Altman, Patrick Royston, and Roger L Holder. The design of simulation studies in medical statistics. *Statistics in Medicine*, 25(24):4279–4292, 2006. (Cited on page 105.)

[15] Christopher A Bush and Steven N MacEachern. A semiparametric Bayesian mo*d*el for randomised block designs. *Biometrika*, 83(2):275–285, 1996. (Cited on page 49.)

[16] Alan Cantor. Imputation for censored observations in survival studies allowing for a positive cure rate. *University of Alabama*. (Cited on page 68.)

[17] Bradley P Carlin and Thomas A Louis. *Bayesian methods for data analysis*. CRC Press, 2008. (Cited on page 28.)

[18] James Carpenter and Michael Kenward. *Multiple imputation and its application*. John Wiley & Sons, 2012. (Cited on page 67.)

[19] W Chang, J Cheng, J Allaire, Y Xie, and J McPherson. Shiny: web application framework for R, R package version 0.12. 2, 2015. (Cited on page 44.)

[20] Kuo-Ching Chiou. A study of imputing censored observations for 2-parameter weibull distribution based on random censoring. *Department of Accounting and Statistics, The Overseas Chinese Institute of Technology*, pages 1–5, 2003. (Cited on page 68.)

[21] Ronald Christensen, Wesley Johnson, Adam Branscum, and Timothy E Hanson. *Bayesian ideas and data analysis: an introduction for scientists and statisticians*. CRC Press, 2011. (Cited on pages 13, 18, 30, 39, and 76.)

[22] Medical Research Council Renal Cancer Collaborators. Interferon-α and survival in metastatic renal carcinoma: early results of a randomised controlled trial. *The Lancet*, 353(9146):14–17, 1999. (Cited on page 4.)

[23] David Collett. *Modelling survival data in medical research*. CRC press, 2015. (Cited on page 25.)

[24] Peter D Congdon. *Applied Bayesian hierarchical methods*. CRC Press, 2010. (Cited on page 41.)

[25] David Roxbee Cox and David Oakes. *Analysis of survival data*, volume 21. CRC Press, 1984. (Cited on page 26.)

[26] Cox R David. Regression models and life tables (with discussion). *Journal of the Royal Statistical Society*, 34:187–220, 1972. (Cited on page 26.)

[27] Maria De Iorio, Wesley O Johnson, Peter Müller, and Gary L Rosner. Bayesian nonparametric nonproportional hazards survival modeling. *Biometrics*, 65(3):762–771, 2009. (Cited on pages 89 and 90.)

[28] Jean Diebolt and Christian P Robert. Estimation of finite mixture distributions through Bayesian sampling. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 363–375, 1994. (Cited on page 38.)

[29] Allan Donner. The relative effectiveness of procedures commonly used in multiple regression analysis for dealing with missing values. *The American Statistician*, 36(4):378–381, 1982. (Cited on page 63.)

[30] Hani Doss. Bayesian nonparametric estimation for incomplete data via successive substitution sampling. *The Annals of Statistics*, pages 1763–1786, 1994. (Cited on page 88.)

[31] Hani Doss and Fred W. Huffer. Monte carlo methods for Bayesian analysis of survival data using mixtures of Dirichlet process priors. Technical report, 1998. (Cited on page 88.)

[32] Hani Doss and B Narasimhan. Dynamic display of changing posterior in Bayesian survival analysis. *Practical Nonparametric and Semiparametric Bayesian Statistics*, pages 63–87, 1998. (Cited on page 88.)

[33] Craig K Enders. *Applied missing data analysis*. Guilford Press, 2010. (Cited on page 64.)

[34] Michael D Escobar. Estimating normal means with a Dirichlet process prior. *Journal of the American Statistical Association*, 89(425):268–277, 1994. (Cited on page 48.)

[35] Michael D Escobar and Mike West. Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, 90(430):577–588, 1995. (Cited on pages 45, 48, and 50.)

[36] J Fabius et al. Asymptotic behavior of Bayes' estimates. *The Annals of Mathematical Statistics*, 35(2):846–856, 1964. (Cited on page 41.)

[37] Cheryl L Faucett, Nathaniel Schenker, and Jeremy MG Taylor. Survival analysis using auxiliary variables via multiple imputation, with application to AIDS clinical trial data. *Biometrics*, 58(1):37–47, 2002. (Cited on page 69.)

[38] Thomas S Ferguson. A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, pages 209–230, 1973. (Cited on pages 36, 37, 39, 41, 45, and 88.)

[39] Thomas S Ferguson. Prior distributions on spaces of probability measures. *The Annals of Statistics*, pages 615–629, 1974. (Cited on pages 39, 41, 45, and 88.)

[40] Thomas S Ferguson and Eswar G Phadia. Bayesian nonparametric estimation based on censored data. *The Annals of Statistics*, pages 163–186, 1979. (Cited on page 86.)

[41] Catherine Forbes, Merran Evans, Nicholas Hastings, and Brian Peacock. *Statistical distributions*. John Wiley & Sons, 2011. (Cited on page 130.)

[42] David A Freedman. On the asymptotic behavior of Bayes' estimates in the discrete case. *The Annals of Mathematical Statistics*, pages 1386–1403, 1963. (Cited on page 41.)

[43] Emil J Freireich, Edmund Gehan, Emil Frei, Leslie R Schroeder, Irving J Wolman, Rachad Anbari, E Omar Burgert, Stephen D Mills, Donald Pinkel, Oleg S Selawry, et al. The effect of 6-mercaptopurine on the duration of steroid-induced remissions in acute leukemia: A model for evaluation of other potentially useful therapy. *Blood*, 21(6):699–716, 1963. (Cited on page 3.)

[44] Edmund A Gehan. A generalized Wilcoxon test for comparing arbitrarily singly-censored samples. *Biometrika*, 52(1-2):203–224, 1965. (Cited on page 3.)

[45] Alan E Gelfand and Adrian FM Smith. Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85(410):398–409, 1990. (Cited on page 29.)

[46] Andrew Gelman, John B Carlin, Hal S Stern, and David B Dunson. *Bayesian data analysis*, volume 2. 2014. (Cited on pages 38 and 41.)

[47] ME Ghitany. The monotonicity of the reliability measures of the beta distribution. *Applied Mathematics Letters*, 17(11):1277–1283, 2004. (Cited on page 16.)

[48] Stephen Jay Gould. The median isn't the message. *Discover*, 6(6):40–42, 1985. (Cited on page 1.)

[49] Sander Greenland and William D Finkle. A critical look at methods for handling missing covariates in epidemiologic regression analyses. *American Journal of Epidemiology*, 142(12):1255–1264, 1995. (Cited on page 64.)

[50] Susan Halabi and Bahadur Singh. Sample size determination for comparing several survival curves with unequal allocations. *Statistics in Medicine*, 23(11):1793–1815, 2004. (Cited on page 106.)

[51] Jerry Halpern, Wm. Byron, and Jun Brown. Cure rate models: power of the logrank and generalized Wilcoxon tests. *Statistics in Medicine*, 6(4):483–489, 1987. (Cited on page 14.)

[52] Frank Harrell. *Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis*. Springer, 2015. (Cited on page 61.)

[53] David W Hosmer, Stanley Lemeshow, and Susanne May. *Applied Survival Analysis: Regression Modeling of Time-to-Event Data, Second Edition*. Hoboken, N.J. : Wiley-Interscience, New York, 2008. (Cited on page 4.)

[54] Chiu-Hsieh Hsu and Jeremy MG Taylor. Nonparametric comparison of two survival functions with dependent censoring via nonparametric multiple imputation. *Statistics in Medicine*, 28(3):462–475, 2009. (Cited on page 69.)

[55] Chiu-Hsieh Hsu, Jeremy MG Taylor, and Chengcheng Hu. Analysis of accelerated failure time data with dependent censoring using auxiliary variables via nonparametric multiple imputation. *Statistics in Medicine*, 34(19):2768–2780, 2015. (Cited on page 69.)

[56] Chiu-Hsieh Hsu, Jeremy MG Taylor, Susan Murray, and Daniel Commenges. Survival analysis using auxiliary variables via non-parametric multiple imputation. *Statistics in Medicine*, 25(20):3503–3517, 2006. (Cited on page 69.)

[57] Joseph G Ibrahim, Ming-Hui Chen, and Debajyoti Sinha. *Bayesian survival analysis*. Wiley Online Library, 2005. (Cited on pages 29, 86, and 88.)

[58] Dan Jackson, Ian R White, Shaun Seaman, Hannah Evans, Kathy Baisley, and James Carpenter. Relaxing the independent censoring assumption in the Cox proportional hazards model using multiple imputation. *Statistics in Medicine*, 33(27):4681–4694, 2014. (Cited on page 68.)

[59] Amirhossein Jalali, Alberto Alvarez-Iglesias, John Hinde, and John Newell. A visualisation tool summarising time-to-event data. *Proceedings of the Conference on Applied Statistics in Ireland (CASI)*, 2017. (Cited on page 151.)

[60] Alejandro Jara, Timothy Hanson, Fernando Quintana, Peter Müller, and Gary Rosner. DPpackage: Bayesian semi- and nonparametric modeling in R. *Journal of Statistical Software*, 40(5):1–30, 2011. (Cited on pages 56 and 89.)

[61] Michael P Jones. Indicator and stratification methods for missing explanatory variables in multiple linear regression. *Journal of the American Statistical Association*, 91(433):222–230, 1996. (Cited on page 64.)

[62] Edward L Kaplan and Paul Meier. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282):457–481, 1958. (Cited on pages 23, 24, and 87.)

[63] John P Klein, Hans C Van Houwelingen, Joseph G Ibrahim, and Thomas H Scheike. *Handbook of survival analysis*. Chapman and Hall/CRC, 2013. (Cited on page 31.)

[64] David G Kleinbaum and Mitchel Klein. *Survival analysis: a self-learning text*. Springer Science & Business Media, 2006. (Cited on page 9.)

[65] Mirjam J Knol, Kristel JM Janssen, A Rogier T Donders, Antoine CG Egberts, E Rob Heerdink, Diederick E Grobbee, Karel GM Moons, and Mirjam I Geerlings. Unpredictable bias when using the missing indicator method or complete case analysis for missing confounder values: an empirical example. *Journal of Clinical Epidemiology*, 63(7):728–736, 2010. (Cited on page 64.)

[66] Charles Kooperberg. *logspline: Logspline Density Estimation Routines*, 2016. R package version 2.1.9. (Cited on pages 84, 94, and 131.)

[67] Charles Kooperberg and Charles J Stone. A study of logspline density estimation. *Computational Statistics & Data Analysis*, 12(3):327–347, 1991. (Cited on page 132.)

[68] Charles Kooperberg and Charles J Stone. Logspline density estimation for censored data. *Journal of Computational and Graphical Statistics*, 1(4):301–328, 1992. (Cited on page 131.)

[69] Jerald F Lawless. *Statistical models and methods for lifetime data*, volume 362. John Wiley & Sons, 2011. (Cited on pages 18 and 22.)

[70] Elisa T Lee and John Wang. *Statistical methods for survival data analysis*, volume 476. John Wiley & Sons, 2003. (Cited on pages 8, 18, and 67.)

[71] Roderick JA Little. Regression with missing x's: a review. *Journal of the American Statistical Association*, 87(420):1227–1237, 1992. (Cited on pages 62 and 65.)

[72] Roderick JA Little and Donald B Rubin. Statistical analysis with missing data. *Statistical Analysis with Missing Data, 2nd ed., by RJA Little and DB Rubin. Wiley series in probability and statistics. New York, NY: Wiley, 2002*, 2002. (Cited on page 64.)

[73] Jun S Liu. Nonparametric hierarchical Bayes via sequential imputations. *The Annals of Statistics*, pages 911–930, 1996. (Cited on page 50.)

[74] Albert Y Lo et al. On a class of Bayesian nonparametric estimates: I. density estimates. *The Annals of Statistics*, 12(1):351–357, 1984. (Cited on page 89.)

[75] Heng-Hui Lue, Chen-Hsin Chen, and Wei-Hwa Chang. Dimension reduction in survival regressions with censored data via an imputed spline approach. *Biometrical Journal*, 53(3):426–443, 2011. (Cited on page 68.)

[76] David Lunn, Chris Jackson, Nicky Best, Andrew Thomas, and David Spiegelhalter. *The BUGS book: A practical introduction to Bayesian analysis*. CRC press, 2012. (Cited on pages 36 and 72.)

[77] Steven N MacEachern and Peter Müller. Estimating mixture of Dirichlet process models. *Journal of Computational and Graphical Statistics*, 7(2):223–238, 1998. (Cited on page 91.)

[78] Steven P. Millard. *EnvStats: An R Package for Environmental Statistics*. Springer, New York, 2013. (Cited on page 131.)

[79] Maja Miloslavsky, Sündüz Keleş, Mark J Laan, and Steve Butler. Recurrent events analysis in the presence of time-dependent covariates and dependent censoring. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(1):239–257, 2004. (Cited on page 106.)

[80] Geert Molenberghs, Garrett Fitzmaurice, Michael G Kenward, Anastasios Tsiatis, and Geert Verbeke. *Handbook of missing data methodology*. CRC Press, 2014. (Cited on pages 61, 65, and 66.)

[81] Geert Molenberghs and Michael Kenward. *Missing data in clinical studies*, volume 61. John Wiley & Sons, 2007. (Cited on page 64.)

[82] David Morina and Albert Navarro. survsim: Simulation of simple and complex survival data. *R Package Version*, 1(2), 2014. (Cited on page 105.)

[83] Peter Müller and Fernando A Quintana. Nonparametric Bayesian data analysis. *Statistical Science*, pages 95–110, 2004. (Cited on pages 36 and 41.)

[84] Radford M Neal. Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9(2):249–265, 2000. (Cited on page 91.)

[85] Wayne Nelson. Hazard plotting for incomplete failure data. *Journal of Quality Technology*, 1(1), 1969. (Cited on page 23.)

[86] Wayne Nelson. A short life test for comparing a sample with previous accelerated test results. *Technometrics*, 14(1):175–185, 1972. (Cited on page 23.)

[87] John Newell, Amirhossein Jalali, Alberto Alvarez-Iglesias, Martin O'Donnell, and John Hinde. Translational statistics and dynamic nomograms. *Proceedings of the Conference on Applied Statistics in Ireland (CASI)*, 2014. (Cited on page 151.)

[88] Wei Pan and John E Connett. A multiple imputation approach to linear regression with clustered censored data. *Lifetime Data Analysis*, 7(2):111–123, 2001. (Cited on page 68.)

[89] Karl Pearson. Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London. A*, 185:71–110, 1894. (Cited on page 37.)

[90] M Postman, JP Huchra, and MJ Geller. Probes of large-scale structure in the corona borealis region. *The Astronomical Journal*, 92:1238–1247, 1986. (Cited on page 3.)

[91] Kathryn Roeder. Density estimation with confidence sets exemplified by superclusters and voids in the galaxies. *Journal of the American Statistical Association*, 85(411):617–624, 1990. (Cited on page 3.)

[92] P Royston. The lognormal distribution as a model for survival time in cancer, with an emphasis on prognostic factors. *Statistica Neerlandica*, 55(1):89–104, 2001. (Cited on pages 7, 60, and 69.)

[93] Patrick Royston, Mahesh KB Parmar, and Douglas G Altman. Visualizing length of survival in time-to-event studies: a complement to Kaplan–Meier plots. *Journal of the National Cancer Institute*, 100(2):92–97, 2008. (Cited on pages 7, 60, 69, 83, 138, and 140.)

[94] Donald B Rubin. Multiple imputation for nonresponse in surveys (wiley series in probability and statistics). 1987. (Cited on page 65.)

[95] Joseph L Schafer. *Analysis of incomplete multivariate data*. CRC press, 1997. (Cited on page 66.)

[96] Jayaram Sethuraman. A constructive definition of Dirichlet priors. *Statistica Sinica*, pages 639–650, 1994. (Cited on page 42.)

[97] Jayaram Sethuraman and Ram C Tiwari. Convergence of Dirichlet measures and the interpretation of their parameter. Technical report, DTIC Document, 1981. (Cited on page 42.)

[98] Sibylle Sturtz, Uwe Ligges, and Andrew Gelman. R2WinBUGS: A package for running WinBUGS from R. *Journal of Statistical Software*, 12(3):1–16, 2005. (Cited on page 75.)

[99] V Susarla and John Van Ryzin. Nonparametric Bayesian estimation of survival curves from incomplete observations. *Journal of the American Statistical Association*, 71(356):897–902, 1976. (Cited on pages 86, 87, and 88.)

[100] MP Sylvestre, T Evans, T MacKenzie, and M Abrahamowicz. Permalgo: Permutational algorithm to generate event times conditional on a covariate matrix including time-dependent covariates. *R Package Version*, 1, 2013. (Cited on page 105.)

[101] Jeremy MG Taylor, Susan Murray, and Chiu-Hsieh Hsu. Survival estimation and testing via multiple imputation. *Statistics & Probability Letters*, 58(3):221–232, 2002. (Cited on page 69.)

[102] Werner Vach and Mana Blettner. Biased estimation of the odds ratio in case-control studies due to the use of ad hoc methods of correcting for missing values for confounding variables. *American Journal of Epidemiology*, 134(8):895–907, 1991. (Cited on page 64.)

[103] Stef Van Buuren. Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical Methods in Medical Research*, 16(3):219–242, 2007. (Cited on page 65.)

[104] Stef Van Buuren. *Flexible imputation of missing data*. CRC press, 2012. (Cited on page 62.)

[105] Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000. (Cited on page 130.)

[106] MS Waterman and DE Whiteman. Estimation of probability densities by empirical density functions. *International Journal of Mathematical Education in Science and Technology*, 9(2):127–137, 1978. (Cited on pages 130 and 131.)

[107] Greg CG Wei and Martin A Tanner. Posterior computations for censored regression data. *Journal of the American Statistical Association*, 85(411):829–839, 1990. (Cited on page 68.)

[108] Greg CG Wei and Martin A Tanner. Applications of multiple imputation to the analysis of censored regression data. *Biometrics*, pages 1297–1309, 1991. (Cited on page 68.)

[109] W. Weibull. *A Statistical Theory of the Strength of Materials*. Ingeniörsvetenskapsakademiens handlingar. Generalstabens litografiska anstalts förlag, 1939. (Cited on page 19.)

[110] Waloddi Weibull et al. A statistical distribution function of wide applicability. *Journal of Applied Mechanics*, 18(3):293–297, 1951. (Cited on page 19.)

[111] Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2009. (Cited on page 142.)

[112] Grace L Yang. Estimation of a biometric function. *The Annals of Statistics*, pages 112–116, 1978. (Cited on page 151.)