Parameter Estimation For Additive Hazards Model With Partly Interval-Censored Failure Time Data using a Penalized Likelihood Approach

Kasun Dananjaya Rathnayake

A thesis submitted in fulfilment of the requirements for the degree of Doctor of Philosophy

Department of Statistics

Faculty of Science and Engineering

Macquarie University, Sydney, Australia

January, 2019

"The mind is everything. What you think, you become"

- Lord Buddha

This thesis is dedicated,

To my ever loving appachchi,

For bringing me up to what I am today, even though you are physically not with me, you are in my heart always...

To my dearest amma, my mother-in-law and father-in-law,

For all the encouragement throughout the years, for your unflagging love and faith...

To my beloved wife Sujanie,

For your unconditional love, unwavering support, inexhaustible patience and strength that made me go on...

This journey would not have been possible without you...

Abstract

In the context of failure time data, interval censoring is a censoring type which has become increasingly prevalent in many areas, including medical, financial, actuarial and sociological studies. In interval-censored data, the actual failure time is neither exactly observed nor right-censored nor left-censored, but one can establish boundaries of an interval within which the survival event has occurred. The aim of this thesis is to develop a maximum penalized log-likelihood (MPL) method which estimates model parameters of the semiparametric additive hazards model with partly interval-censored failure time data. This data will contain exactly observed, left-censored, finite interval-censored and right-censored data. This MPL method estimates the regression coefficients and the underlying non-parametric baseline hazard function, simultaneously, by imposing non-negativity constraints on the baseline hazard and the overall hazard function. We approximate infinite dimensional baseline hazard from a finite number of non-negative basis functions.

The chosen MPL method guarantees the smoothness of the baseline hazard estimates, which clearly depicted the trend of how the estimates changed over time. We adopted the augmented Lagrangian method to solve this constraint optimization problem, and the estimates were obtained simultaneously using the Newton and multiplicative iterative (Newton-MI) algorithm followed by line-search steps. The asymptotic properties of these derived constrained MPL estimators were studied when the number of basis functions was fixed and when it went to infinity. We investigated the performance of this proposed MPL method by conducting simulation studies for both right-censored data and partly interval-censored data. Both of the simulation studies demonstrated that our method worked well for small and large datasets as well as small and large censoring proportions. The derived asymptotic standard deviation formula was generally accurate in approximating the standard deviation of the constrained MPL estimates. In addition, we also made comparisons between our MPL method and existing parameter estimation methods developed by Aalen (1980) and Lin & Ying (1994). Results show that our MPL method provided better estimates. In a real data analysis, we applied our MPL method to fit the additive hazards model to a melanoma data set with all types of censoring, which was provided by the Melanoma Institute of Australia.

Acknowledgements

As I approach the completion of my PhD, I realize that I am deeply indebted to the support and assistance of many people. I wish to thank everyone of those, with whom I shared experiences in life, who have helped me survive the birthing of this thesis.

Foremost, I would like to express my heartfelt gratitude to my principal supervisor, Associate Professor Jun Ma, for his tremendous time and effort spent in leading, advising and encouraging me during the time passed with this research. His passion for challenges and his enthusiasm has given me inspiration, his vast knowledge has given me guidance, his perseverance in research has given me motivation. I feel extremely lucky to work under the supervision of such a gifted statistician.

I owe a special gratitude to Dr. Maurizio Manuguerra, not only for being a wonderful associate supervisor, but also for guiding me with R coding and Latex. He has been a constant source of knowledge and support, guiding me with tremendous patience throughout the years.

I gratefully acknowledge the support of everyone in the Department of Statistics, Macquarie University specifically, Head of the department Professor Gillian Heller, the former Head of the department Professor Ian Marschner, all the academic staff, all the non-academic staff, especially Victoria Park and Susan Pe for their kindness and support in all administration issues. I would like to acknowledge the Macquarie University for providing me with necessary financial support mainly through the international Macquarie University Research Excellence Scholarship (iMQRS). I am deeply indebted to Dr. Serigne N. Lo from Melanoma institute of Australia for giving me opportunity to use his Melanoma data in my research. I would also like to thank my fellow PhD candidates, colleagues and students at Macquarie University. I especially thank everyone in the biostatistics research group including Manjula, Jinqing, Andrew L., Andrew G. and Gigi who were always very friendly, kind and willing to help.

These past few years have been less stressful and more enjoyable because of the companionship of my friends. My friends in Sydney made my PhD life more bearable through many gatherings, trips and food. I would like to thank them, the extended UWS crew and the Macquarie gang, for making it much easier to live so far from my beloved motherland.

I am tempted to thank all of my friends who from my childhood until this point, have joined me in the discovery of what life is about and how to make the best of it. Especially my uni buddies: Amal, Anushka, Harithra, Waruni, Laknilusha, Dulanjali and Kenath, who made the four years at University of Peradeniya one of the most enjoyable and cherished periods of my life. Since then they have always cheered me up bringing me closer to my motherland and helped me get through the tough times by being the best friends I will ever find in my life and I am grateful to their friendship more than they will ever know.

I sincerely thank all my teachers since childhood for their contribution in shaping me up to what I am today, my school teachers and the lecturers from the University of Peradeniya who gave me confidence in my abilities that made me climb up the ladder of education. I feel truly lucky to have many supportive relatives and I would like to acknowledge all of them who offered many words of encouragement throughout the years. I would especially like to express my deepest gratitude and love to my akki who always believed in me.

An ocean of thanks goes out to my awesome parents. My parents who were the guiding stars in my life, have always made me believe that education is the key to success. They have made so many sacrifices in life, put hold to so many plans of theirs to give way to ours and always made their children the first priority in life. I am forever indebted to all that they have done for me. Without their endless support and unconditional love I would not have achieved this much.

I thank my wonderful extended family, my mother-in-law, father-in-law, sister-in-law, brother-in-law and cute Dylan baba for their love, continuous support and encouragement, but more specifically for their lovely daughter, my beloved wife Sujanie.

Last but not the least, my wife has been a constant source of love, caring and support. She is my best friend, my better half and the best thing that ever happened to me. She is tremendously patient throughout my downs, she always listens to me and offers me a helping hand, she instils my confidence and faith whenever I loose it and loves me unconditionally during my good and bad times. Her complete and unflagging love carries me through always. I get through my worst moments simply by thinking about her and her love. She has always understood me the best and has taken care of whatever needed tending without any complaints so that I could complete my thesis. Words cannot simply express my love and gratitude for her. This PhD journey has not been an easy ride. Through her love, patience, support and unwavering faith in me I was able to make it out unscathed.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Survival analysis describes the analysis of data selected from a well-defined time origin until the occurrence of some event of interest or end-point (Collett 2003). For that reason survival analysis is often referred to as *time-to-event* analysis. It is widely used in research areas, such as medicine, biological studies, financial studies, epidemiological studies, sociological studies and engineering. In engineering sciences, survival analysis is largely known as reliability theory or reliability analysis, because modeling the lifetime of mechanical or electronic components is the main purpose of the analysis (Lawless 2002). In the field of economics, survival analysis is referred to as event history analysis or duration analysis. The time taken to commit a crime, or recidivism, after former convicts have been released from jail, could be considered as an example. Even though survival analysis is known by different names in many scientific fields, it uses the same analytical technique to investigate those survival models (Lee & Wang 2003).

Generally, the occurrences of events of interest in survival analysis are referred to as *failures*, even though the event may not actually be a failure; for instance, the birth of a child. When the response variable is time to failure, then we can use the term *failure time data* to represent it. A basic overview of failure time data is presented in the next section.

1.1 Failure time data

Failure time data concerns positive random variables which represent the time taken to a certain event. It appears extensively in biomedical studies and technical reliability studies, but there are many other disciplines where it may also occur. In a medical research context, the starting time, diagnosis date or the time of origin often corresponds to the enrolment of an individual into a study. The data is referred to as a failure time if the end-point of a study is the death of an individual. However, a similar kind of data can be obtained when the end-point of a study is not a terminal event, such as the diagnosis of a secondary disease or recurrence of symptoms. Therefore, sometimes it is possible to use the term survival data and can refer the variable of interest as survival time. As a generic term, the time from the beginning of the study up to the event of interest can be denoted as survival time (Aalen et al. 2008). While these discussions refer to survival times and patients, and thereby imply survival analysis in human subjects, the methodology discussed applies broadly to all time-to-event data with experimental units that are not necessarily human subjects. Therefore, the terms survival data, failure data, and time-to-event data are used interchangeably throughout this thesis. Generally, survival data is not symmetrically distributed, and it tends to be positively skewed implying that there is a longer tail to the right which is generally caused by censoring. Therefore, it is inappropriate to use standard statistical modeling methods to analyse survival data (Torben & Scheike 2006).

The main reason which differentiates the failure time data from other types of data is the existence of censoring. A failure time is known to be exactly observed when the actual failure time is observed. When the end-point of interest has not been observed for an individual, the failure time is referred to as censored (Therneau & Grambsch 2000) which implies that the failure time is known to fall within a certain range instead of knowing the exact time. Censoring might be due to an individual being still alive at the end of the study or the survival status of that person might not be known as he or she has not been located for follow-up.

There are three types of censoring: right-censoring, left-censoring and finite interval-censoring. When the failure time of interest is greater than the study end time, it is referred to as right-censored failure time data. This often occurs when an individual has dropped out, or a survival study finished before the occurrence of the event of interest. In such situations, the corresponding failure time is known to be greater than the censoring time. Another form of censoring is known as left-censoring, which occurs when the actual failure time of an individual is less than the observed time. Yet another censoring method is available known as finite interval-censoring. If the individuals of interest are not under continuous monitoring, the actual failure time is neither exactly observed nor right-censored nor left-censored, but one observes the interval within which the survival event has occurred.

Consider a survival study which focuses on time to recurrence of a tumor after removal of the primary site and assume that there are follow-up consultations at six and twelve months. If an individual is found to have had a recurrence at the first follow-up, then it is left-censored. If the patient is free of a tumor recurrence at six months, but it is found to have had a recurrence at the second follow-up, then it is finite interval-censored and if the recurrence has not occurred by the second and final follow-up, then it is considered as right-censored. Leiderman et al. (1973) provides a classic example of censoring types using some data on children's ability to learn. Right-censored or left-censored failure times can be considered as special cases of interval-censored failure times (Sun 2006). Thus, through out this thesis we use the term interval-censored data to represent a collection of left, right and finite interval-censored data. In some survival studies, the failure time of an individual is restricted to knowledge of whether or not it exceeds a time point. This type of data is generally referred to as current status data (e.g. Rossini & Tsiatis 1996) and they are just equivalent to left and right-censoring.

Truncation is another feature which distinguishes the failure time data from the other types of data. In truncation, individuals enter the study only if they survive a sufficient length of time or individuals are included in the study only if the event has occurred by a given date (Mandel 2007). Analyzing failure time data with censoring is the main focus of this thesis, and truncation is not considered.

1.2 Notations and formulations

Some notation for failure time data is defined in this section. Suppose in a given survival study, there are n independent subjects. For each subject denoted by i, $i = 1, \dots, n$, the failure time is denoted by T_i, and the censoring time is denoted by C_i. For interval-censored failure time data, T_i is not always exactly observed and we may only observe a time interval within which the failure event has occurred. Thus, according to the definition of interval-censored data discussed in Section 1.1, for T_i there is an observed time interval $(L_i, R_i]$ such that

$$T_i \in (L_i, R_i], \quad \text{for} \quad i = 1, 2, \dots, n, \tag{1.1}$$

where $L_i \leq R_i$. Throughout this thesis, it is assumed that T_i, L_i, R_i and C_i are continuous random variables. T_i is exactly observed, when $L_i = R_i$. If $L_i = 0$, it implies that T_i is left-censored and if $R_i = \infty$, then T_i is right-censored. T_i is said to be finite interval-censored if $(L_i, R_i]$ satisfies the requirement $0 < L_i \leq R_i < \infty$. Throughout this thesis, partly interval-censored data, which is defined as a collection of fully observed and interval-censored data, is considered.

Sun (2006) presented an alternative method for representing interval-censored data based on the assumption that each subject i in a given survival study is observed twice and the suggested functional form is given by:

$$\Big\{C_{i1}, C_{i2}; \delta_{i1} = I(T_i \leq C_{i1}), \delta_{i2} = I(C_{i1} < T_i \leq C_{i2}), \delta_{i3} = 1-\delta_{i1}-\delta_{i2}; i = 1, 2, \ldots, n\Big\}, \tag{1.2}$$

where C_{i1} and C_{i2} are the two monitoring random variables satisfying $C_{i1} < C_{i2}$, and δ_{i1}, δ_{i2} and δ_{i3} are the indicators. Here $I(\cdot)$ is an indicator function. Note that the formulation in equation (1.1) can be easily obtained by equation (1.2).

It is possible to obtain a generalized form of interval-censored failure time data representation by assuming that each individual is observed more than twice. Therefore, a sequence of monitoring time points exists, for an instance $C_{i1} \leq C_{i2} \leq, \ldots, \leq C_{iM_i}$, where M_i is the number of monitoring random times for the individual i. Then, the observed information can be expressed as:

$$\{M_i, C_{ij}; \delta_{ij} = I(C_{i(j-1)} < T_i \leq C_{ij});\ i = 1, 2, \ldots, n;\ j = 1, 2, \ldots, M_i\}, \tag{1.3}$$

where $C_{i0} = 0$. Failure time data of this type is often described as mixed case interval-censored data (Wellner 1995, Schick & Yu 2000). Subsequently, Wang et al. (2010) used functional formulations similar to equations (1.2) and (1.3) in developing a counting process approach for interval-censored data, which will be discussed under Section 2.3.1.

1.3 Hazard function and survival function

The hazard and survival functions are two common functions used in survival studies. Those two functions along with relevant existing relationships between survival measures will be discussed in this section. Denote $h(t)$ and $S(t)$ as the hazard and the survival function of failure time T respectively. $h(t)$ is commonly used to express the hazard or risk of death at time t. It is obtained using the conditional probability that an individual fails at time t, given that the subject has survived up to that time. Thus, $h(t)$ defines the instantaneous probability that a subject fails at time t given that the subject has not failed before t, i.e.,

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t}. \tag{1.4}$$

Note that $h(t)$ must be a non-negative function.

Suppose T has a probability density function $f(t)$, then the cumulative distribution function of T is given by,

$$F(t) = P(T < t) = \int_0^t f(u)du, \tag{1.5}$$

and it represents the probability that the survival time T is less than some random value t. On the other hand, the survival function $S(t)$ is defined as the probability that the failure time T is greater than, or equal to, t, and therefore,

$$S(t) = P(T \geq t) = 1 - F(t). \tag{1.6}$$

A useful relationship between $h(t)$ and $S(t)$ can be obtained by considering equations (1.4), (1.5) and (1.6) as follows,

$$h(t) \equiv \frac{f(t)}{S(t)}. \tag{1.7}$$

Then, the hazard function can be re-expressed as

$$h(t) = -\frac{d}{dt}\{\log S(t)\}, \tag{1.8}$$

and

$$S(t) = \exp\{-H(t)\}, \tag{1.9}$$

where $H(t) = \int_0^t h(u)du$, which is known as the cumulative hazard function.

1.4 Hazard model

In survival analysis, the main concern is to calculate the hazard to evalaute the risk of failure at any time (after the time origin) of the survival study. Therefore, the hazard function is modelled directly. The hazard function is a convenient method of describing the probability distribution for an event time T. Also, there may be explanatory variables which might be used in a survival study as covariates affect the final outcome (Klein & Moeschberger 2003). Two models are mainly used to explore the relationship between the hazard function of an individual and its covariates. They are: (i) the proportional hazards model and (ii) the additive hazards model.

1.4.1 Proportional hazards model

The proportional hazards model (or well-known as Cox PH model) (Cox 1972) is one of the most popular survival models (Kalbfleisch & Prentice 2002). The Cox PH model is based on the proportional hazards assumption, that the hazard ratio of two strata (determined by the particular choices of covariates) is constant over time. For each individual i, the Cox PH model specifies $h(t|\mathbf{x}_i)$ as

$$h(t|\mathbf{x}_i) = h_0(t)\exp(\mathbf{x}_i\boldsymbol{\beta}). \tag{1.10}$$

Here, $h_0(\cdot)$ is an unspecified (non-parametric) baseline hazard function which shows the shape of the hazard rate when covariates are set at baseline values, while $\exp(\mathbf{x}_i\boldsymbol{\beta})$ is the hazard ratio which can be used to explain how the magnitude of the hazard depends on covariates (Aalen et al. 2008). In equation (1.10), $\mathbf{x}_i$ is the time-independent covariate vector, and $\boldsymbol{\beta}$ is a vector of regression coefficients which is usually the primary interest in the model fitting process. Here, $h_0(t)$ and the regression coefficients, $\boldsymbol{\beta}$ can be considered as model parameters. This model assumes that the covariates act multiplicatively on some unknown baseline hazard.

For two individuals with corresponding covariate vectors denoted by $\mathbf{x}_i$ and $\mathbf{x}_j$, the ratio of their hazard functions at time t is given by,

$$\frac{h(t|\mathbf{x}_i)}{h(t|\mathbf{x}_j)} \equiv \exp\{(\mathbf{x}_i - \mathbf{x}_j)\boldsymbol{\beta}\}. \tag{1.11}$$

As discussed under model (1.12), this ratio is called the hazard ratio of an individual with risk factor $\mathbf{x}_i$, compared to an individual with risk factor $\mathbf{x}_j$. Thus, it is possible to conclude that the hazards are proportional to each other as they do not depend on time t.

The cumulative hazard function for model (1.10) is

$$H(t|\mathbf{x}_i) = \int_0^t h(u|\mathbf{x}_i)du = H_0(t)\exp(\mathbf{x}_i\boldsymbol{\beta}), \tag{1.12}$$

where $H_0(t) = \int_0^t h_0(u)du$ is known to be the cumulative baseline hazard function. Then the survival function is given by,

$$S(t|\mathbf{x}_i) = \exp\{-H_0(t)\exp(\mathbf{x}_i\boldsymbol{\beta})\} = [S_0(t)]^{\exp(\mathbf{x}_i\boldsymbol{\beta})} \tag{1.13}$$

where $S_0(t)$ is the baseline survival function.

1.4.2 Additive hazards model

An additive hazards model is a good alternative when the proportionality assumption in the Cox PH model is not satisfied (Huffer & McKeague 1991). In contrast to the Cox PH model, where the covariates are assumed to act multiplicatively on the baseline hazard function, the additive hazards model assumes that the effect of covariates is to additively increase or decrease the hazard function. This model is quite useful when an absolute change in risk is of more interest than a relative change in risk. Also, examples exist where the additive hazards model fits better when compared to the Cox PH model (Breslow & Day 1987).

Similar to the Cox PH model, the baseline hazard $h_0(t)$ and the regression coefficients $\boldsymbol{\beta}$ are the model parameters of interest for the additive hazards model. In general, in order to be a valid additive hazards model, there are two requirements to be satisfied;

$$\textbf{(a)}\text{ non-negative baseline hazard: } \quad h_0(t) \ \geq 0, \quad \text{and} \tag{1.14}$$

$$\textbf{(b)}\text{ non-negative hazard function: } \quad h(t|\mathbf{x}_i(t)) \ \geq 0. \tag{1.15}$$

The requirements specified in (1.14) and (1.15) are referred to as non-negativity constraints for the additive hazards model.

Two forms of additive hazards models are presented in this section. The first was introduced by Aalen (1980, 1989) and the second was introduced by Lin & Ying (1994). Aalen (1980, 1989) considered a more generalized additive hazards model which specifies the hazard function, $h(t|\mathbf{x}_i(t))$ for each individual i at a time point t according to

$$h(t|\mathbf{x}_i(t)) = h_0(t) + \mathbf{x}_i(t)\boldsymbol{\beta}(t), \tag{1.16}$$

where $\boldsymbol{\beta}(t) = (\beta_1(t), \beta_2(t), \ldots, \beta_p(t))^\top$ is a vector of time-varying regression coefficients in which $\beta_j(t)$ denotes the excess risk at time t for the j^{th} covariate and

$\mathbf{x}_i(t) = (x_{i1}(t), x_{i2}(t), \ldots, x_{ip}(t))^\top$ is a $p \times 1$ vector of possibly time-dependent covariates. Since this model does not make any distributional assumptions, this can be identified as a non-parametric model (Aalen et al. 2008). Note that Aalen's additive hazards model allows the regression coefficients to be functions of t, thus it can reveal the changes in the influence of the covariates over time. This is considered as one of the main advantages of Aalen's model. Estimation of the regression parameters, testing the model and assessment of the model fit can be found in Aalen (1989).

Lin & Ying (1994) proposed a simpler additive hazards model. Lin and Ying's additive hazards model replaces $\boldsymbol{\beta}(t)$ in the Aalen's model by time-independent regression coefficients $\boldsymbol{\beta}$ (Lin & Ying 1994, 1995, 1997). Thus, Lin and Ying's additive hazards model is given by,

$$h(t|\mathbf{x}_i(t)) = h_0(t) + \mathbf{x}_i(t)\boldsymbol{\beta}. \tag{1.17}$$

Using an estimating equation approach, it is possible to estimate regression coefficients directly for this model. This method treats $h_0(t)$ as a nuisance parameter and only estimates $\boldsymbol{\beta}$.

Another additive hazards model has been proposed by McKeague & Sasieni (1994) and it is an intermediate between the models (1.16) and (1.17) in which some of the regression coefficients are assumed to be constant over time while others are allowed to be arbitrary regression functions. Note that for both models proposed by

Lin & Ying (1994) and McKeague & Sasieni (1994), no explicit assumptions about the distributional form of the underlying baseline hazard needs to be made and the estimation of the $\boldsymbol{\beta}$ parameters can be conducted independent of the underlying baseline hazard.

Several authors presented parameter estimation techniques for those additive hazards models discussed (see Aalen (1989), Lin & Ying (1994), McKeague & Sasieni (1994), Huffer & McKeague (1991), Andersen et al. (2012)). Most of the studies mainly focused on estimating the regression coefficients while the baseline hazard has been considered as a nuisance parameter. Furthermore, none of the implementations consider both of the non-negativity constraints discussed under (1.14) and (1.15) in their parameter estimation procedures. The development of the novel methodology proposed in this thesis to estimate model parameters and to derive their asymptotic properties is based on Lin and Ying's additive hazards model given in (1.17).

1.5 Existing semi-parametric estimation methods

Limited literature exists on the semi-parametric estimation methods for the additive hazards models (see Section 1.4.2) with right-censored or interval-censored data. For right-censored data, Lin & Ying (1994) develop an estimation function for $\boldsymbol{\beta}$, which mimics the counting process characteristics of the partial likelihood score function in the Cox PH model (Cox 1975). This approach does not facilitate the estimation of $h_0(t)$. The estimator for the cumulative baseline hazard function can be obtained by the method proposed by Lin & Ying (1994) which is similar to the Breslow method (Breslow 1972).

For interval-censored data, Lin et al. (1998) and Wang et al. (2010) proposed a counting process approach, where a partial likelihood type score function for $\boldsymbol{\beta}$ is obtained. Regression coefficient estimates are obtained by finding the root of the score function. The baseline hazard $h_0(t)$ is again considered as a nuisance parameter and is not estimated directly in this method. Thus, this method is not suitable for prediction. A maximum likelihood approach was developed in Ghosh (2001) and Zeng et al. (2006) for the additive hazards model with interval-censored data. Ghosh (2001) developed an estimation procedure for time-independent covariates in model (1.17), whereas Zeng et al. (2006) considered time-dependent covariates (model (1.17)). Using a primal-dual interior point algorithm (Wright 1997), estimates for $\boldsymbol{\beta}$ and $H_0(t)$ were developed by Ghosh (2001), with non-negativity constraint and monotonicity on $H_0(t)$. Zeng et al. (2006) estimate $\boldsymbol{\beta}$, and baseline survival function $S_0(t)$ using Newton's algorithm, where a log function was used to impose the non-negativity constraint on $S_0(t)$. Since the estimates of $h_0(t)$ can be obtained indirectly, this method can be used for prediction purposes. Yet, this method may not be time efficient when the sample size is large, because of the need for estimation of $S_0(t)$ at each distinct observed time point. A generalized linear model (GLM) approach was proposed to estimate the model parameters with interval-censored data in Farrington (1996). This approach considers occurrence of survival observations as independent Bernoulli trials and links the occurrence probability to a linear predictor. For that linear predictor, $\boldsymbol{\beta}$ and $h_0(t)$ are regression coefficients which can be estimated using the statistical software GLIM (Nelder 1975). However, this approach does not satisfy the non-negativity constraint. A detailed discussion of these methods is available in Section 2.3.

1.6 Goal

In this thesis, a novel methodology to fit an additive hazards model with partly interval-censored data (collection of fully observed, left-censored, right-censored and finite interval-censored data) is proposed. We assume that: (i) the observations $(t_i, \mathbf{x}_i)$ from different individuals are independent: (ii) the distribution of covariates does not involve regression coefficients and the covariates are time-independent; and, (iii) censoring time is independent of the failure time.

Estimating the regression coefficients $\boldsymbol{\beta}$ and the baseline hazard $h_0(t)$ for the additive hazards model is the main purpose of our proposed method. The estimates of $\boldsymbol{\beta}$ and $h_0(t)$ are obtained simultaneously by adopting the maximum penalized likelihood (MPL) method. This procedure requires both $h(\cdot|\mathbf{x}_i)$ and $h_0(\cdot)$ constrained to be non-negative (according to (1.14) and (1.15)) during the estimation process. We adopt the augmented Lagrangian method (Hestenes 1969, Powell 1969, Rockafellar 1976, Bertsekas 1996) to perform this constrained optimization. The hazard of failure of most biological mechanisms tends to happen gradually. For example, the underlying hazard of time to death of patients suffering from cancer is likely to increase gradually with time. Thus, a penalty function is used to produce a smoothed estimate of $h_0(t)$ in this MPL method. The asymptotic properties of the MPL estimates are also derived. Therefore, those MPL estimates can be used for predictions, to perform hypothesis testing and to obtain variance estimates.

In summary, the objective of this thesis is to develop a MPL approach to estimate the additive hazards model parameters with partly interval-censored data by considering the two non-negativity constraints. The performance of this method is evaluated through a series of simulation studies. Furthermore, the proposed method is applied to analyze a real data set to further exemplify the significance of the method. Applying the MPL approach for right-censored data is less complex than that of partly interval-censored data. Therefore, the implementation of our MPL method for right-censored data is considered first followed by the implementation of this MPL method for partly interval-censored data. However, for the interval-censored data, there are no publicly available computational developed statistical tools for the parameter estimation process. There are few R packages available for additive hazards models of Aalen (1989) and Lin & Ying (1994) with right-censored data. Thus, a comprehensive simulation study is performed to compare our MPL approach with the results from these R packages for right-censored data. For interval-censored data, a simulation study is only used to demonstrate the effectiveness of the proposed method. Note that, for the simplicity of implementation, here we only consider time-independent covariates. However, the proposed approach can be easily extended to time-dependent covariates.

1.7 Thesis overview

This thesis consists of an Introduction, Literature review, five research chapters and a Conclusion which also suggests future work on this proposed methodology. In Chapter 3, we develop a parameter estimation approach for the additive hazards model and, as an initial attempt, we restrict the failure time data to be either right-censored or fully observed. We used the MPL method to estimate the model parameters and the augmented Lagrangian method was used for constrained optimization.

Chapter 4 considers the same parameter estimation method. In this chapter, we extend the method proposed in Chapter 3 for partly-interval censored data. Asymptotic properties of the derived MPL estimates are considered in Chapter 5. Here, we discuss the asymptotic properties in two ways; one is more important in theoretical nature and the second method is more useful in applications. These asymptotic properties can be used for model checking and the model validation process.

A detailed simulation study is given in the Chapter 6. Firstly, it presents the simulation results for less complex right-censored data by following the method discussed in Chapter 3. These simulation results are compared against the results of two other existing parameter estimation methods. Then, it presents the simulation results based on Chapter 4 for partly interval-censored data. The results from a real dataset are given in Chapter 7.

A discussion for each of these five research chapters is given in their respective chapters, followed by an overall conclusion along with suggestions for future work is provided in Chapter 8.

Chapter 2

Literature Review

2.1 Introduction

In this chapter, some of the existing parameter estimation methods for the semi-parametric additive hazards model are reviewed. Semi-parametric survival models assume that there is a relationship between covariate effects and the hazard function, but the underlying distribution functions for the failure time are not fully specified and it needs to be estimated during the estimation process. Semi-parametric models contain both parametric and non-parametric components and, particularly for the additive hazards model, $h_0(t)$ being the non-parametric component and the contribution of the covariates being the parametric component.

Section 2.2 presents parameter estimation approaches followed by Aalen (1989) and Lin & Ying (1994) when they first introduce the additive hazards models. Section 2.3 discusses three existing parameter estimation approaches for the semi-parametric additive hazards models; (i) counting process approach; (ii) maximum likelihood approach and (iii) generalized linear model approach. Section 2.4 presents the proposed parameter estimation methodology for the additive hazards model using the MPL method.

2.2 Parameter estimation for additive hazards model

This section presents the parameter estimation procedures followed by Aalen (1989) and Lin & Ying (1994) when they introduced additive hazards models given in equations (1.16) and (1.17) respectively. The parameter estimation procedures summarized in this section are only developed for right-censored data.

2.2.1 Aalen's additive hazards model

Aalen (1989) considered a generalized additive hazards model (1.16) in which no finite dimensional parameter is assumed; thus, it can be considered as a non-parametric model. In general, estimating the cumulative distribution function is far easier than estimating the density function. For similar reasons, estimating cumulative regression functions are much easier than estimating regression functions themselves. Thus, a least squares approach is used by Aalen (1989) for model (1.16) to estimate cumulative regression coefficients and the standard errors of those estimates. This method directly estimates the cumulative regression coefficient for the j^{th} covariate, $B_j(t)$ and then the crude estimates of $\beta_j(t)$ can be obtained from the slope of $B_j(t)$ as follows:

$$B_j(t) = \int_0^t \beta_j(u)du, \quad \text{for } j = 1, 2, \cdots, p. \tag{2.1}$$

Better estimates for $\beta_j(t)$ can be acquired indirectly from kernel-smoothing technique (Huffer & McKeague 1991).

According to the approach followed by Aalen (1989), to estimate the cumulative regression coefficient vector $\mathbf{B}(t)$, a design matrix $\mathbf{Z}(t)$ with dimensions $(n \times (p+1))$ should be defined as follows: for the i^{th} row of $\mathbf{Z}(t)$, assign $\mathbf{Z}_i(t) = Y_i(t)(1, \mathbf{x}_i(t))$. Here, $Y_i(t) = 1$ if i^{th} individual is under observation (at risk) at time t, and $Y_i(t) = 0$ if i^{th} individual is not under observation (not at risk) at time t. That is, if the i^{th} individual is a member of the risk set at time t (event has not occurred or the individual has not been censored), then $\mathbf{Z}_i(t) = (1, x_{i1}(t), x_{i2}(t), \ldots, x_{ip}(t))$. If the i^{th} individual is not in the risk set at time t (the event of interest has already occurred or the individual has been censored), then $\mathbf{Z}_i(t)$ is a $(p+1)$ row vector of zeros. Suppose that $\mathbf{I}(t)$ is the $(n \times 1)$ vector with the i^{th} element equal to 1 if the subject i fails at time t and 0 otherwise. Then, the least square estimate of $\mathbf{B}(t)$ is given by

$$\widehat{\mathbf{B}}(t) = \sum_{T_i \leq t} [\mathbf{Z}^\top(T_i)\mathbf{Z}(T_i)]^{-1}\mathbf{Z}^\top(T_i)\mathbf{I}(T_i). \tag{2.2}$$

Aalen (1989) derived the variance-covariance matrix of $\widehat{\mathbf{B}}(t)$ as

$$\mathrm{Var}[\widehat{\mathbf{B}}(t)] = \sum_{T_i \leq t} [\mathbf{Z}^\top(T_i)\mathbf{Z}(T_i)]^{-1}\mathbf{Z}^\top(T_i)\mathbf{I}^{\mathbf{D}}(T_i)\mathbf{Z}(T_i)\{[\mathbf{Z}^\top(T_i)\mathbf{Z}(T_i)]^{-1}\}^\top, \quad (2.3)$$

where $\mathbf{I}^{\mathbf{D}}(t)$ is a $(n \times n)$ diagonal matrix in which diagonal elements are given by $\mathbf{I}(t)$. $\widehat{\mathbf{B}}(t)$ only exists up to the smallest time at which the matrix $\mathbf{Z}^\top(T_i)\mathbf{Z}(T_i)$ becomes singular and it is unbiased up to the time where $\mathbf{Z}_i(T_i)$ loses its full rank (Aalen 1989). Furthermore, this procedure does not obey the two non-negativity constraints discussed under (1.14) and (1.15), which is clearly a disadvantage of this method. This procedure is not suitable for prediction as it does not involve the estimation of $h_0(t)$.

2.2.2 Lin and Ying's additive hazards model

Lin & Ying (1994) proposed a semi-parametric parameter estimation method for model (1.17) and developed an estimation equation approach to estimate the model parameters. This approach directly estimates $\boldsymbol{\beta}$ and treats $h_0(t)$ as a nuisance parameter. To estimate the regression coefficients β_j, the vector of the average values of the covariates at time t, $\bar{\mathbf{x}}(t)$ are constructed by

$$\bar{\mathbf{x}}(t) = \frac{\sum_{i=1}^{n} \mathbf{x}_i Y_i(t)}{\sum_{i=1}^{n} Y_i(t)}. \quad (2.4)$$

The numerator in equation (2.4) is the sum of the covariates for all individuals at risk at time t and the denominator is the number of individuals at risk at time t.

Next, a $(p \times p)$ matrix $\mathbf{A}$, a $(p \times 1)$ vector $\mathbf{B}$ and a $(p \times p)$ matrix $\mathbf{C}$ are constructed as follows:

$$\mathbf{A} = \sum_{i=1}^{n}\sum_{j=1}^{i}(T_j - T_{j-1})[\mathbf{x}_i - \bar{\mathbf{x}}(T_j)]^\top[\mathbf{x}_i - \bar{\mathbf{x}}(T_j)], \tag{2.5}$$

$$\mathbf{B}^\top = \sum_{i=1}^{n}\delta_i[\mathbf{x}_i - \bar{\mathbf{x}}(T_j)], \tag{2.6}$$

$$\mathbf{C} = \sum_{i=1}^{n}\delta_i[\mathbf{x}_i - \bar{\mathbf{x}}(T_j)]^\top[\mathbf{x}_i - \bar{\mathbf{x}}(T_j)]. \tag{2.7}$$

Then, the estimate for $\boldsymbol{\beta}$

$$\hat{\boldsymbol{\beta}} = \mathbf{A}^{-1}\mathbf{B}, \tag{2.8}$$

and the variance-covariance matrix for $\hat{\boldsymbol{\beta}}$

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}) = \mathbf{A}^{-1}\mathbf{C}\mathbf{A}^{-1}. \tag{2.9}$$

An estimator of the cumulative baseline hazard can be obtained indirectly through the Breslow method (Breslow 1972). But this method has some efficiency issues when working with left-truncated and right-censored data by using the conditional estimating equation method. Furthermore, these estimates naturally cannot be extended to interval-censored data.

2.3 Estimation methods for additive hazards models with interval-censored data

In this section, three approaches to fit the semi-parametric additive hazards model with interval-censored data are reviewed. Specifically, the additive hazards model with time-independent regression coefficients is considered here.

2.3.1 Counting process approach

A counting process method to fit the additive hazards model for current status data was introduced in Lin et al. (1998). Subsequently, Wang et al. (2010) adopt this method to study interval-censored data. According to the definition of current status data given in Section 1.1, the exact value of the failure time T_i is never known, but known that it is observed below or above the monitored time variable C_i. Thus, only monitoring time variable, C_i and the indicator variable $\delta_i = I(C_i \leq T_i)$, where $I(\cdot)$ is the indicator function, are observed. Denote the i^{th} observation by $\{C_i, \delta_i, \mathbf{x}_i\}$. Lin et al. (1998) considered two cases: (i) the monitoring time C_i is dependent on $\mathbf{x}_i$; and, (ii) the monitoring time C_i is independent of $\mathbf{x}_i$. Their results indicated that the estimates of $\boldsymbol{\beta}$ obtained in case (i) have smaller standard errors compared to those in case (ii).

This section discusses the counting process approach mentioned under case (i).

Lin et al. (1998) suggest the dependence of C_i on $\mathbf{x}_i$ through a proportional hazards model, and denote the hazard function of C_i at time t as

$$d\tilde{H}(t|\mathbf{x}_i) = e^{\mathbf{x}_i\boldsymbol{\alpha}} d\tilde{H}_0(t), \tag{2.10}$$

where $\tilde{H}_0(\cdot)$ is an unspecified cumulative baseline hazard function, and $\boldsymbol{\alpha}$ is a $(p \times 1)$ vector of unknown regression parameters for the effect of $\mathbf{x}_i$ on C_i. When $\boldsymbol{\alpha} = 0$, the process is simplified into case (ii). Define a counting process for subject i as $N_i(t) = \delta_i I(t \geq C_i)$. $N_i(t)$ increases by one unit at time t whenever the i^{th} individual is monitored at time t and found still to be failure-free. Let $Y_i(t) = I(t \leq C_i)$, and the hazard rate function of $N_i(t)$ is defined as

$$d\tilde{\Lambda}_i(t) = Y_i(t) e^{-\mathbf{x}_i\boldsymbol{\beta}\mathbf{t} + \mathbf{x}_i\boldsymbol{\alpha}} d\tilde{\Lambda}_0(t), \tag{2.11}$$

where $d\tilde{\Lambda}_0(t) = e^{-H_0(t)} d\tilde{H}_0(t)$ with $H_0(t) = \int_0^t h_0(u)du$. Equation (2.11) is the well-known Cox proportional hazards model. Lin et al. (1998) proves that the compensated counting process

$$M_i(t) = N_i(t) - \int_0^t Y_i(u) e^{-\mathbf{x}_i\boldsymbol{\beta}\mathbf{u} + \mathbf{x}_i\boldsymbol{\alpha}} d\tilde{\Lambda}_0(u) \quad \text{for} \quad i = 1, 2, \ldots, n \tag{2.12}$$

is a martingale. By applying the partial likelihood approach to model (2.11) to make inference about $\boldsymbol{\beta}$, the score function for $\boldsymbol{\beta}$ with given $\boldsymbol{\alpha}$ is,

$$U_{\boldsymbol{\beta}}(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \sum_{i=1}^{n} \int_0^{\infty} \left[t\mathbf{x}_i - \frac{W_{\boldsymbol{\beta}}^{(1)}(t; \boldsymbol{\beta}, \boldsymbol{\alpha})}{W_{\boldsymbol{\beta}}^{(0)}(t; \boldsymbol{\beta}, \boldsymbol{\alpha})} \right] dN_i(t), \tag{2.13}$$

where,

$$W_{\boldsymbol{\beta}}^{(j)}(t; \boldsymbol{\beta}, \boldsymbol{\alpha}) = \sum_{i=1}^{n} Y_i(t)(t\mathbf{x}_i)^{(j)} e^{-\mathbf{x}_i\boldsymbol{\beta}\mathbf{t} + \mathbf{x}_i\boldsymbol{\alpha}} \quad \text{for} \quad j = 0, 1.$$

In order to estimate $\boldsymbol{\alpha}$, Lin et al. (1998) apply the partial likelihood approach to model (2.10) which carries information about monitoring time C_i. The partial likelihood score function for $\boldsymbol{\alpha}$ is given by

$$U_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \int_{0}^{\infty} \left[\mathbf{x}_i - \frac{W_{\boldsymbol{\alpha}}^{(1)}(t;\boldsymbol{\alpha})}{W_{\boldsymbol{\alpha}}^{(0)}(t;\boldsymbol{\alpha})} \right] dI(C_i \leq t),$$

where

$$W_{\boldsymbol{\alpha}}^{(j)}(t;\boldsymbol{\alpha}) = \sum_{i=1}^{n} Y_i(t)\mathbf{x}_i^{(j)} e^{\mathbf{x}_i \boldsymbol{\alpha}} \quad \text{for} \quad j = 0, 1.$$

Let $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\alpha}}$ be estimates of $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ respectively, which can be obtained through the Newton's algorithm. This approach requires C_i to follow the proportional hazard model. Note that the dependence of C_i on $\mathbf{x}_i$ through the proportional hazards model may result in dependent censoring, since T_i is also dependent on $\mathbf{x}_i$, but through the additive hazards model.

Following the methodology proposed by Lin et al. (1998), an estimation equation approach has been introduced by Wang et al. (2010) for interval-censored data $\{(L_i, R_i], \mathbf{x}_i; i = 1, 2, \ldots, n\}$. In that approach, it is assumed that for each individual i, there exist only two monitoring times denoted by C_{i1} and C_{i2}, where $C_{i1} < C_{i2}$. Assume that the event time T_i is independent of C_{i1} and C_{i2}. Then, the indicator functions for interval-censored data can be defined as $\delta_{i1} = I(T_i < C_{i1})$, $\delta_{i2} = I(C_{i1} \leq T_i < C_{i2})$ and $\delta_{i3} = 1 - \delta_{i1} - \delta_{i2}$ corresponding to the failure time of individual i has occurred before C_{i1}, during the interval $[C_{i1}, C_{i2})$, or after C_{i2} respectively. Similar to Lin et al. (1998), Wang et al. (2010) also model C_{i1} and

C_{i2} respectively through proportional hazards models as,

$$d\tilde{H}_{i1}(t|\mathbf{x}_i) = e^{\mathbf{x}_i\boldsymbol{\alpha}} d\tilde{H}_{01}(t) \tag{2.14}$$

and

$$d\tilde{H}_{i2}(t|C_{i1}, \mathbf{x}_i) = I(t > C_{i1}) e^{\mathbf{x}_i\boldsymbol{\alpha}} d\tilde{H}_{02}(t), \tag{2.15}$$

where $\tilde{H}_{01}(\cdot)$ and $\tilde{H}_{02}(\cdot)$ are unspecified cumulative baseline hazard functions, and $\boldsymbol{\alpha}$ is a $(p \times 1)$ vector of unknown regression parameters. Then, for each individual i, they define counting processes $N_{i1}(t) = (1 - \delta_{i1})I(t \geq C_{i1})$ and,

$$N_{i2}(t) = \begin{cases} \delta_{i3} I(t \geq C_{i2}) & \text{if} \quad t \geq C_{i1}, \\ 0 & \text{if} \quad t < C_{i1}. \end{cases}$$

Let $Y_{i1}(t) = I(t \leq C_{i1})$ and $Y_{i2}(t) = I(C_{i1} < t \leq C_{i2})$. Adopting similar arguments as Lin et al. (1998), and using models (1.17), (2.14) and (2.15), hazard functions for $N_{i1}(t)$ and $N_{i2}(t)$ are derived respectively as

$$d\tilde{\Lambda}_{i1}(t) = Y_{i1}(t) e^{-\mathbf{x}_i\boldsymbol{\beta}\mathbf{t} + \mathbf{x}_i\boldsymbol{\alpha}} d\tilde{\Lambda}_{01}(t) \tag{2.16}$$

and

$$d\tilde{\Lambda}_{i2}(t) = Y_{i2}(t) e^{-\mathbf{x}_i\boldsymbol{\beta}\mathbf{t} + \mathbf{x}_i\boldsymbol{\alpha}} d\tilde{\Lambda}_{02}(t), \tag{2.17}$$

where

$$d\tilde{\Lambda}_{01}(t) = e^{-H_0(t)} d\tilde{H}_{01}(t) \quad \text{and} \quad d\tilde{\Lambda}_{02}(t) = e^{-H_0(t)} d\tilde{H}_{02}(t).$$

Clearly models (2.16) and (2.17) are Cox proportional hazards type models similar to model (2.11). Also note that the model (2.17) is a conditional model since its starting time point is C_{i1}. For given $\boldsymbol{\alpha}$, the score function for $\boldsymbol{\beta}$ is defined as

$$U_{\boldsymbol{\beta}}(\boldsymbol{\beta},\boldsymbol{\alpha}) = \sum_{i=1}^{n}\left\{\int_0^\infty \left[t\mathbf{x}_i - \frac{W^{(1)}_{1,\boldsymbol{\beta}}(t;\boldsymbol{\beta},\boldsymbol{\alpha})}{W^{(0)}_{1,\boldsymbol{\beta}}(t;\boldsymbol{\beta},\boldsymbol{\alpha})}\right] dN_{i1}(t) + \int_0^\infty \left[t\mathbf{x}_i - \frac{W^{(1)}_{2,\boldsymbol{\beta}}(t;\boldsymbol{\beta},\boldsymbol{\alpha})}{W^{(0)}_{2,\boldsymbol{\beta}}(t;\boldsymbol{\beta},\boldsymbol{\alpha})}\right] dN_{i2}(t)\right\} \quad (2.18)$$

where

$$W^{(j)}_{1,\boldsymbol{\beta}}(t;\boldsymbol{\beta},\boldsymbol{\alpha}) = \sum_{i=1}^{n} Y_{i1}(t)e^{-\mathbf{x}_i\boldsymbol{\beta}t+\mathbf{x}_i\boldsymbol{\alpha}}(t\mathbf{x}_i)^{(j)} \quad \text{for} \quad j=0,1$$

and

$$W^{(j)}_{2,\boldsymbol{\beta}}(t;\boldsymbol{\beta},\boldsymbol{\alpha}) = \sum_{i=1}^{n} Y_{i2}(t)e^{-\mathbf{x}_i\boldsymbol{\beta}t+\mathbf{x}_i\boldsymbol{\alpha}}(t\mathbf{x}_i)^{(j)} \quad \text{for} \quad j=0,1.$$

Then, for each individual i, define 0-1 counting processes $\tilde{N}_{i1}(t) = I(t \geq C_{i1})$ and, conditional on C_{i1},

$$\tilde{N}_{i2}(t) = \begin{cases} I(t \geq C_{i2}) & \text{if} \quad t \geq C_{i1} \\ 0 & \text{if} \quad t < C_{i1}. \end{cases}$$

In order to estimate $\boldsymbol{\alpha}$, it is possible to apply the partial likelihood approach to equations (2.14) and (2.15) to obtain the score function for $\boldsymbol{\alpha}$ as

$$U_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}) = \sum_{i=1}^{n}\left\{\int_0^\infty \left[\mathbf{x}_i - \frac{W^{(1)}_{1,\boldsymbol{\alpha}}(t;\boldsymbol{\alpha})}{W^{(0)}_{1,\boldsymbol{\alpha}}(t;\boldsymbol{\alpha})}\right] d\tilde{N}_{i1}(t) + \int_0^\infty \left[\mathbf{x}_i - \frac{W^{(1)}_{2,\boldsymbol{\alpha}}(t;\boldsymbol{\alpha})}{W^{(0)}_{2,\boldsymbol{\alpha}}(t;\boldsymbol{\alpha})}\right] d\tilde{N}_{i2}(t)\right\},$$

where

$$W^{(j)}_{1,\boldsymbol{\alpha}}(t;\boldsymbol{\alpha}) = \sum_{i=1}^{n} Y_{i1}(t)\mathbf{x}_i^{(j)}e^{\mathbf{x}_i\boldsymbol{\alpha}} \quad \text{for} \quad j=0,1$$

and

$$W^{(j)}_{2,\boldsymbol{\alpha}}(t;\boldsymbol{\alpha}) = \sum_{i=1}^{n} Y_{i2}(t)\mathbf{x}_i^{(j)}e^{\mathbf{x}_i\boldsymbol{\alpha}} \quad \text{for} \quad j=0,1.$$

The root of $U_{\boldsymbol{\alpha}}(\boldsymbol{\alpha})$ gives the estimate of $\boldsymbol{\alpha}$, and then $\boldsymbol{\beta}$ can be estimated by solving $U_{\boldsymbol{\beta}}(\boldsymbol{\beta}, \hat{\boldsymbol{\alpha}}) = 0$. With the resulting estimates denoted by $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\alpha}}$, Wang et al. (2010) showed that the estimates are consistent and asymptotically normal, and the variance-covariance matrix of $\hat{\boldsymbol{\beta}}$ can be consistently estimated.

This procedure does not require estimation of $h_0(t)$ at any stage of the estimation process, which is an advantage of this approach. However, this procedure highly depends on the proportional hazards assumption. The user should check the validity of the proportionality assumption before utilizing this approach. On the other hand, this method is not suitable for prediction as it does not estimate $h_0(t)$.

2.3.2 Maximum Likelihood (ML) approach

Ghosh (2001) and Zeng et al. (2006) developed a maximum likelihood (ML) approach for the additive hazards model, which can be considered as an alternative method to the counting process approach (see Section 2.3.1). Using current status data, Ghosh (2001) developed a primal-dual interior point algorithm (e.g.Wright (1997)) for estimating $\boldsymbol{\beta}$ and the cumulative baseline hazard function $H_0(t)$ simultaneously. His algorithm imposes positivity constraints on both $h_0(\cdot)$ and $h(\cdot)$ and maintains monotonic increments on both $H_0(\cdot)$ and $H(\cdot)$. The maximum likelihood (ML) estimator of $\boldsymbol{\beta}$ is shown to be consistent and it converges to a multivariate normal distribution.

Zeng et al. (2006) developed a parameter estimation procedure for the additive hazards model with interval-censored data based on the ML approach, where the corresponding log-likelihood function is specified using regression coefficient vector $\boldsymbol{\beta}$ and the baseline survival function $S_0(\cdot)$ as

$$\ell(\boldsymbol{\beta}, S_0) = \sum_{i=1}^{n} \log\left\{S_0(L_i)e^{-\mathbf{x}_i\boldsymbol{\beta}L_i} - S_0(R_i)e^{-\mathbf{x}_i\boldsymbol{\beta}R_i}\right\}. \tag{2.19}$$

Let $0 = v_0 < v_1 < v_2 < \cdots < v_m < v_{m+1} = \infty$ be distinct ordered time points of all observed interval end points $\{(L_i, R_i]; i = 1, 2, \ldots, n\}$. Here, $S_0(\cdot)$ is only evaluated at those unique ordered time points. Define $\mathbf{S}_0 = [S_0(v_0), S_0(v_1), \ldots, S_0(v_{m+1})]^\top$, where $S_0(v_0) = 1$ and $S_0(v_{m+1}) = 0$. Let indicator functions be $\xi_{ij} = I(v_j \in (L_i, R_i])$ for $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, (m+1)$, then the log-likelihood function (2.19) can be re-expressed as

$$\ell(\boldsymbol{\beta}, S_0) = \sum_{i=1}^{n} \log\left\{\sum_{j=1}^{m+1} \xi_{ij}\left[S_0(v_{j-1})e^{-\mathbf{x}_i\boldsymbol{\beta}v_{j-1}} - S_0(v_j)e^{-\mathbf{x}_i\boldsymbol{\beta}v_j}\right]\right\}. \tag{2.20}$$

In order to impose the non-negativity constraints and monotonicity of $S_0(\cdot)$ (i.e.: $0 \leq S_0(v_m) < S_0(v_{m-1}) < \cdots < S_0(v_1) \leq 1$), log transformation is applied on $S_0(\cdot)$ as;

$$\omega_j = \begin{cases} \log S_0(v_1) & \text{for} \quad j = 1 \\ \log[S_0(v_{j-1}) - S_0(v_j)] & \text{for} \quad j = 2, 3, \ldots, m. \end{cases}$$

Let the log-transformed vector be $\boldsymbol{\omega} = [\omega_1, \omega_2, \ldots, \omega_m]^\top$. Replacing S_0 in equation (2.20) by $\boldsymbol{\omega}$ gives the updated log-likelihood with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\omega}$. Maximizing the log-likelihood $\ell(\boldsymbol{\beta}, \boldsymbol{\omega})$, the estimates of $\boldsymbol{\beta}$ and $\boldsymbol{\omega}$ can be obtained. Then Newton's algorithm can be applied to find the optimal solution when m is not large. When m is large, the Nelder-Mead method (Nelder & Mead 1965) can be used to solve the score equations for $\boldsymbol{\beta}$ and $\boldsymbol{\omega}$. This method estimates $h_0(\cdot)$ indirectly. Thus, the results from this method can be used for prediction. However, this procedure is not efficient in terms of processing time when the sample size is large as it requires the estimation of the baseline survival at each unique observed time point.

2.3.3 Generalized linear model (GLM) approach

Farrington (1996) introduced another parameter estimation approach for the additive hazards model with interval-censored data using the GLM approach. This approach facilitates the estimation of both $\boldsymbol{\beta}$ and $h_0(t)$.

Let n_L, n_R, n_I be the number of left-censored, right-censored and finite interval-censored individuals respectively, and the total number of individuals in the survival study $n = n_L + n_R + n_I$. Let the observations be ordered as left-censored, right-censored and finite interval-censored and $S(\cdot|\mathbf{x}_i)$ be the survival function conditional on $\mathbf{x}_i$ and (t_i^L, t_i^R) denotes the observed survival time interval. Then, the likelihood can be expressed in terms of $S(\cdot|\mathbf{x}_i)$ as

$$\mathcal{L} = \prod_{i=1}^{n_L}[1 - S_i(t_i^R|\mathbf{x}_i)] \prod_{i=n_L+1}^{n_L+n_R} S_i(t_i^L|\mathbf{x}_i) \prod_{i=n_L+n_R+1}^{n} S_i(t_i^L|\mathbf{x}_i)\left[1 - \frac{S_i(t_i^R|\mathbf{x}_i)}{S_i(t_i^L|\mathbf{x}_i)}\right], \quad (2.21)$$

where the first, second and third product terms on the right-hand side correspond to the likelihoods for the left-censored, right-censored and finite interval-censored observations respectively.

From the GLM point of view, it considers (2.21) as a particular realisation for $n+n_I$ independent Bernoulli trials with probability π_i and response y_i, where $i = 1, 2, \ldots, n+n_I$. The associated model can be defined as follows. If the i^{th} individual is left-censored, it contributes one Bernoulli trial with probability $\pi_i = 1 - S(t_i^R|\mathbf{x}_i)$, response $y_i = 1$ and observed time interval $J_i = (0, t_i^R]$. If the i^{th} individual is right-censored, then it contributes one Bernoulli trial with $\pi_i = 1 - S(t_i^L|\mathbf{x}_i)$, $y_i = 0$ and $J_i = (0, t_i^L]$. Finally, if the i^{th} individual is finite interval-censored (when $n - n_I + 1 \leq i \leq n$), it contributes two Bernoulli trials indexed i and $i + n_I$. The first Bernoulli trial has $\pi_i = 1 - S(t_i^L|\mathbf{x}_i)$, $y_i = 0$ and $J_i = (0, t_i^L]$, and the second Bernoulli trial has $\pi_{i+n_I} = 1 - \frac{S(t_i^R|\mathbf{x}_i)}{S(t_i^L|\mathbf{x}_i)}$, $y_{i+n_I} = 1$ and $J_{i+n_I} = (t_i^L, t_i^R]$. Thus, the likelihood function (2.21) can be rewritten as

$$\mathcal{L} = \prod_{i=1}^{n+n_I} \pi_i^{y_i}(1-\pi_i)^{1-y_i}, \tag{2.22}$$

where

$$\pi_i = 1 - \exp\left\{ - \int_{J_i} h(t|\mathbf{x}_i)dt \right\}. \tag{2.23}$$

Therefore, maximizing (2.21) is equivalent to maximizing (2.22). According to the GLM approach, $h_0(\cdot)$ is assumed to be a piecewise constant over the time intervals, $(\psi_{r-1}, \psi_r]$; $r = 1, 2, \ldots, m$, which divides the time line into m intervals. Thus, it is possible to define $h_0(t) = \theta_r$ for $t \in (\psi_{r-1}, \psi_r]$ and $\boldsymbol{\theta} = [\theta_1, \theta_2, \ldots, \theta_m]^\top$. As per equation (2.23), the probability π_i is related to a linear predictor through a link function

$$\eta_i = -\ln(1 - \pi_i) = \mathbf{x}_i \boldsymbol{\beta} d_i + \sum_{r=1}^{m} \theta_r h_{ir}, \tag{2.24}$$

where d_i is the length of J_i and h_{ir} denotes the length of $J_i \cap (\psi_{r-1}, \psi_r]$. Equation (2.24) defines a GLM with covariates $\mathbf{x}_i d_i$ and h_i, and by fitting the GLM, the parameters $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ can be estimated. During the estimation process, all the linear predictors in (2.24) need to remain positive, which may cause some computational difficulties. Moreover, this estimation approach does not guarantee the non-negativity or smoothness of the $h_0(t)$. This is a disadvantage of this method. On the positive side, since this method estimates both of the model parameters, this can be used for prediction purposes.

2.4 Proposed methodology

In this section, the novel methodology developed in this thesis to fit the additive hazards model is summarized. Here, the additive hazards model proposed by Lin & Ying (1994) is evaluated by estimating the model parameters $\boldsymbol{\beta}$ and $h_0(\cdot)$ simultaneously by adopting the maximum penalized likelihood (MPL) method. One of the major advantage of our MPL is that it imposes the non-negativity constraints on both $h_0(\cdot)$ and $h(\cdot|\mathbf{x}_i)$ when maximizing the penalized log-likelihood function.

Furthermore, this methodology accommodates partly interval-censored data which contains exactly observed, left-censored, finite interval-censored and right-censored failure time data.

Firstly, this new method attempts to maximize the penalized log-likelihood function, which is obtained by approximating the undefined baseline hazard, $h_0(t)$ using a finite number of non-negative basis functions. The estimate of $h_0(t)$ is smoothed through a penalty function and an associated smoothing parameter. For parameter estimations, the Newton-MI algorithm is used, which combines the Newton algorithm and the multiplicative iterative (MI) algorithm of Ma (2010).

Constrained optimization is achieved by using the augmented Lagrangian method; see, for example Bertsekas (2014). The augmented Lagrangian method guarantees non-negativity of $h(\cdot|\mathbf{x}_i)$ and the MI algorithm guarantees non-negativity of $h_0(t)$. Furthermore, the asymptotic properties and the convergence properties of these MPL estimates are considered. The accurate asymptotic covariance matrices for the model parameters are also considered. Since this MPL method produces estimates of both of the parameters during this iterative procedure and the corresponding asymptotic results, this can be used for prediction purposes.

Chapter 3

Penalized likelihood parameter estimation for additive hazards model with right-censored data

3.1 Introduction

This chapter discusses a parameter estimation method for the additive hazards model for right-censored data, where the parameter estimation is conducted via a penalized likelihood approach. The hazard function for an additive hazards model proposed by Lin & Ying (1994) is given by equation (1.17). Unlike the proportional hazards model (1.10) which assumes a relative increase or decrease of magnitude

$\exp(\boldsymbol{\beta}_j)$ to the hazard with a unit change in the j^{th} covariate, $\mathbf{x}_j$, the change in the hazard under an additive hazards model is assumed to be an increase or decrease of $|\boldsymbol{\beta}_j|$. As noted in Section 1.4.2, $h_0(t)$ needs not be explicitly defined. Therefore, this additive hazards model can be described as a semi-parametric model. The goal in this chapter is to develop a parameter estimation method for this semi-parametric model with right-censored survival data.

To this end, Section 3.2 presents the background of the parameter estimation of an additive hazards model with right-censored data. Section 3.3 introduces the Maximum Penalized Log-likelihood (MPL) function for the additive hazards model including aspects related to smoothing, and Section 3.4 discusses the associated parameter estimation method with the constrained optimization algorithm.

3.2 Background of the problem

As discussed in Chapter 1, when survival data contains right censoring, one of the most common parameter estimation approach is the conventional partial likelihood type estimation equation approach proposed by Lin & Ying (1994). It is adequate if one's intent is simply to estimate the $\boldsymbol{\beta}$s. This partial likelihood approach has been widely used since it can accommodate some of the censoring types and the availability of asymptotic properties of the maximum partial likelihood estimates (Lin & Ying 1994). However, this approach does not allow the estimation of $h_0(t)$

which is needed if prediction is the aim of the estimation. A method that allows the estimation of the cumulative baseline hazard function separately has been presented in Lin & Ying (1994), which is similar to the Breslow estimator for the proportional hazard model (Breslow 1972). Separate estimation of $\boldsymbol{\beta}$ and $h_0(t)$ certainly causes difficulties when calculating standard deviations for the predictions. Furthermore, this approach does not impose non-negativity constraints on the baseline hazard (constraint (1.14)) or the hazard function (constraint (1.15)).

This chapter addresses those limitations of the existing approach, and proposes a new methodology using a penalized likelihood approach to facilitate the simultaneous estimation of $\boldsymbol{\beta}$ and $h_0(t)$ by imposing the two non-negativity constraints. The penalized likelihood method was implemented by Good & Gaskins (1971), and also by Silverman (1978), in the context of density estimation. This was extended to the proportional hazards model estimation by Anderson & Senthilselvan (1980). Later, the asymptotic properties of those MPL estimates were investigated by Cox & O'Sullivan (1990). The application of this method in the case of right censoring is discussed in detail in this chapter and more complex partly interval censoring is reserved till Chapter 4.

The MPL method presented in this chapter addresses the scenarios where the failure times are either fully-observed or right-censored. It allows modeling through smoothing of $h_0(t)$ using an appropriate penalty function, while ensuring the non-negativity constraint of $h_0(t)$ (constraint (1.14)). The constraint on $h(t|\mathbf{x}_i)$ (constraint (1.15)) is also satisfied using a Newton-MI algorithm (Ma et al. 2014), which is a combination of the Newton algorithm and a multiplicative-iterative (MI) algorithm (Ma 2010).

3.3 MPL method for the additive hazards model with right-censored data

As discussed earlier in Chapter 1, most of the available techniques for estimating $\boldsymbol{\beta}$ in the additive hazard regression model in the presence of right censoring do not allow the simultaneous estimation of $h_0(t)$. To this end, this section introduces the MPL approach for an additive hazards model with right-censored data, and discusses how this technique can be used to attain simultaneous estimation of both $\boldsymbol{\beta}$ and $h_0(t)$. Furthermore, this section addresses the optimization issues related to this method.

3.3.1 Notations

The notations used in this and subsequent sections of this chapter are briefly describe in this section. Note that the term survival data is used interchangeably with failure time data throughout this chapter as mentioned in Section 1.1.

Consider survival data that are either fully-observed or right-censored. Survival times that are fully-observed will be indicated by the subscript, $\mathcal{O}$, while survival times that are right-censored will be indicated by the subscript, $\mathcal{R}$. Now consider n independent and identically distributed survival times, $t_1, t_2, \ldots, t_n$. Thus, the survival time of subject i will be indicated by $\{t_i; i \in \mathcal{O}\}$ if fully-observed and by $\{t_i; i \in \mathcal{R}\}$ if right-censored. Assume now that associated with subject i, there is a set of p covariates, $x_{i1}, x_{i2}, \ldots, x_{ip}$. Thus, the set $\{(t_i, \mathbf{x}_i), i \in \mathcal{O}\}$ describes a fully-observed survival time including a covariate vector while $\{(t_i, \mathbf{x}_i), i \in \mathcal{R}\}$ is its right-censored equivalent.

The censoring times are assumed to be independent of t_i given $\mathbf{x}_i$. Our developments assume covariates $\mathbf{x}_i$ are time-independent; however, they can be extended to time-dependent covariates. The next section presents the steps taken to estimate the hazard function associated with the failure time of the i^{th} patient, $h(t|\mathbf{x}_i)$.

3.3.2 Imposing non-negativity constraints and estimating the baseline hazard function

This section will discuss how the two non-negativity constraints given by (1.14) and (1.15) are addressed. Since $h_0(t)$ is infinite dimensional, estimation of $h_0(t)$ from a finite number of observations can be ill-conditioned. This problem can be addressed through approximating $h_0(t)$ using a finite number of non-negative basis functions. To this end, assume $\psi_1, \psi_2, ..., \psi_m$ form a basis of this finite dimensional space. Then, $h_0(t)$ can be expressed as

$$h_0(t) = \sum_{u=1}^{m} \theta_u \psi_u(t), \tag{3.1}$$

where m is the dimension of the approximating space and $\psi_u(t)$s are the non-negative basis functions. Since $\psi_u(t) \geq 0$ for all u, if we restrict the basis coefficient vector $\boldsymbol{\theta} = [\theta_1, \theta_2, ..., \theta_m]^\top$ to be non-negative, then $h_0(t) \geq 0$. There are number of choices available for the basis functions. M-splines and B-splines are known to be powerful basis functions where as functions such as indicator and haar known to be simple and easy to implement basis functions. This thesis adopts indicator basis functions to simplify computations and to clearly illustrate the parameter estimation approach. Other basis functions such as M-splines, B-spline will be considered during the further development of this proposed method in future.

For indicator basis functions, $h_0(t)$ becomes a piecewise constant function, which we also termed as discretization of $h_0(t)$. Assume that all the survival times, $t_1, t_2, \ldots, t_n$, are contained within a finite interval $\jmath = [t_{(1)}, t_{(n)}]$, where $t_{(1)} = \min(t_i)$ and $t_{(n)} = \max(t_i)$. Suppose now that there are m bins, $B_1, B_2, ..., B_m$, partitioning the interval $\jmath$ with the bin edges $t_{(1)} = \tau_0 < \tau_1 < ... < \tau_m = t_{(n)}$, where

$$B_u = \{t : \tau_{u-1} < t \leq \tau_u\} \quad \text{for } \ u = 1, 2, ..., m \tag{3.2}$$

with $\cup_{u=1}^{m} B_u = \jmath$ and $B_u \cap B_v = \emptyset$ for $u \neq v$. Thus, at a given time t,

$$h_0(t) = \sum_{u=1}^{m} \theta_u I(\tau_{u-1} < t \leq \tau_u), \tag{3.3}$$

where $I(\cdot)$ is an indicator function. After discretization, estimation of $h_0(t)$ is equivalent to estimating the basis coefficients vector $\boldsymbol{\theta}$, subject to $\theta_u \geq 0$ for all u. We comment that although the baseline hazard is discretized, the non-parametric nature of $h_0(t)$ is preserved up to some extent as there is no restriction on the number of bins m.

Denoting the cumulative baseline hazard function by $H_0(t)$, the cumulative hazard for subject i is

$$H(t|\mathbf{x}_i) = \int_0^t h(s|\mathbf{x}_i)ds = H_0(t) + \mathbf{x}_i\boldsymbol{\beta}t, \tag{3.4}$$

where

$$H_0(t) = \sum_{u=1}^{m} \theta_u \Psi_u(t), \tag{3.5}$$

with $\Psi_u(t) = \int_0^t \psi_u(v)dv$ being the cumulative basis function. From equation (3.3), the discretized $H_0(t)$ is

$$H_0(t) = \sum_{u=1}^{m} \theta_u[(t - \tau_{u-1})I(\tau_{u-1} < t \leq \tau_u) + \rho_u I(t \geq \tau_u)], \tag{3.6}$$

with $\rho_u = (\tau_u - \tau_{u-1})$ indicating the width of B_u.

Now that $h_0(t)$ and $H_0(t)$ functions have been defined using indicator basis functions, the hazard, the cumulative hazard and the survival functions for the additive hazards model follow directly from equations (3.3) and (3.6) as:

$$h(t|\boldsymbol{\theta}, \boldsymbol{\beta}, \mathbf{x}_i) = \sum_{u=1}^{m} \theta_u I(\tau_{u-1} < t \leq \tau_u) + \mathbf{x}_i \boldsymbol{\beta}, \tag{3.7}$$

$$H(t|\boldsymbol{\theta}, \boldsymbol{\beta}, \mathbf{x}_i) = \sum_{u=1}^{m} \theta_u \Big[(t - \tau_{u-1}) I(\tau_{u-1} < t \leq \tau_u) + \rho_u I(t \geq \tau_u)\Big] + \mathbf{x}_i \boldsymbol{\beta} t, \tag{3.8}$$

$$S(t|\boldsymbol{\theta}, \boldsymbol{\beta}, \mathbf{x}_i) = \exp[-H(t|\boldsymbol{\theta}, \boldsymbol{\beta}, \mathbf{x}_i)]. \tag{3.9}$$

After the discretization of $h_0(t)$, our aim is to estimate the model parameters $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ and the estimation process is discussed in the next section.

3.3.3 Maximum Penalized Likelihood estimation

Based on the survival times $t_1, t_2, \ldots, t_n$, which include fully-observed and right-censored data, the log-likelihood is

$$\ell(\boldsymbol{\theta}, \boldsymbol{\beta}) = \sum_{i \in \mathcal{O}} [\log h(t_i|\boldsymbol{\theta}, \boldsymbol{\beta}, \mathbf{x}_i) - H(t_i|\boldsymbol{\theta}, \boldsymbol{\beta}, \mathbf{x}_i)] - \sum_{i \in \mathcal{R}} H(t_i|\boldsymbol{\theta}, \boldsymbol{\beta}, \mathbf{x}_i). \tag{3.10}$$

Here, the first term on the right captures the log-likelihood for the uncensored failure times while the second term captures the log-likelihood for the right-censored failure times. Substituting the expressions in (3.7) and (3.8) into (3.10) results in the following representation of the log-likelihood function

$$\begin{aligned} \ell(\boldsymbol{\theta}, \boldsymbol{\beta}) = \sum_{u=1}^{m} \sum_{t_i \in B_u, i \in \mathcal{O}} \log(\theta_u + \mathbf{x}_i \boldsymbol{\beta}) - \Bigg[\sum_{u=1}^{m} w_u \sum_{v=1}^{u-1} (\rho_v \theta_v) \\ + \sum_{u=1}^{m} \sum_{t_i \in B_u} (t_i - \tau_u) \theta_u + \sum_{i=1}^{n} \mathbf{x}_i \boldsymbol{\beta} t_i \Bigg], \end{aligned} \tag{3.11}$$

where w_u is the number of observations in bin u.

As explained in Section 1.6, it is possible to assume that $h_0(t)$ is smoothed. This can be achieved by subtracting a penalty function (Anderson & Senthilselvan 1980) from the log-likelihood function introduced in (3.11), which results in a penalized log-likelihood function,

$$\Phi(\boldsymbol{\theta}, \boldsymbol{\beta}) = \ell(\boldsymbol{\theta}, \boldsymbol{\beta}) - \lambda J(h_0). \tag{3.12}$$

Here, $\lambda > 0$ is a smoothing parameter used to balance smoothness of the estimated $h_0(t)$ and reliability of the fitted model, and $J(h_0)$ is a penalty function, which depends only on $h_0(\cdot)$. The penalty function in (3.12) could be, for instance, a roughness penalty function, i.e., $J(h_0) = \int_0^t [h_0''(u)]^2 du$, which measures the total curvature of $h_0(t)$. Since $h_0(\cdot)$ is approximated by a piecewise constant function, a penalty representing the square of the second order differences is adopted:

$$J(h_0) = J(\boldsymbol{\theta}) = \sum_{j=2}^{m-1} (\theta_{j-1} - 2\theta_j + \theta_{j+1})^2. \tag{3.13}$$

This penalty can also be re-written in a quadratic form as

$$J(\boldsymbol{\theta}) = \boldsymbol{\theta}^\top \boldsymbol{R} \boldsymbol{\theta}, \tag{3.14}$$

where $\boldsymbol{R} = \boldsymbol{A}^\top \boldsymbol{A}$ and $\boldsymbol{A}$ is an $m \times m$ matrix with $\boldsymbol{A\theta}$ representing the second order differences of $\boldsymbol{\theta}$. Therefore, the matrix $\boldsymbol{R}$ is given by,

$$\boldsymbol{R} = \begin{pmatrix} 5 & -6 & 1 & 0 & 0 & \cdots & 0 \\ -6 & 9 & -4 & 1 & 0 & \cdots & 0 \\ 1 & -4 & 6 & -4 & 1 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 1 & -4 & 6 & -4 & 1 \\ 0 & \cdots & 0 & 1 & -4 & 9 & -6 \\ 0 & \cdots & 0 & 0 & 1 & -6 & 5 \end{pmatrix}.$$

The penalty term introduced in (3.14) supports the penalization of the variation between the θ_us and the average of its neighborhoods.

For the additive hazards models, both $h_0(t)$ and $h(t|\mathbf{x}_i)$ have to be constrained to be non-negative in obtaining MPL estimates of $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$. Hence, under piecewise constant approximation for $h_0(t)$, the proper constrained optimization problem is;

$$(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\beta}}) = \underset{\boldsymbol{\theta},\boldsymbol{\beta}}{\operatorname{argmax}}\ \Phi(\boldsymbol{\theta}, \boldsymbol{\beta}) \tag{3.15}$$

subject to:

$$\theta_u \geq 0 \quad \text{for } u = 1, 2, ..., m \tag{3.16}$$

and

$$\sum_{u=1}^{m} \theta_u I(t \in B_u) + \mathbf{x}_i \boldsymbol{\beta} \geq 0 \quad \text{for } i = 1, 2, ..., n. \tag{3.17}$$

3.4 Constrained optimization by the augmented Lagrangian method

This section discusses the constrained optimization technique used to solve the optimization problem in (3.15) under the constraints (3.16) and (3.17). In the context of optimization problems, although methods such as gradient descent or conjugate gradient descent are quite common, those methods can only be used for unconstrained optimization problems. Here, we have a constrained optimization problem as it is required to consider the two non-negativity constraints. Then, the projected gradient descent method or the Lagrange multiplier method can be considered as possible candidates. But, depending on the number of time points, n, and the number of bins that are chosen, m, these non-negativity constraints could result in an optimization problem with a large number of constraints. Thus, it is required to use an optimization method which can handle large number of constraints to solve this problem.

3.4.1 Optimization method

Since the involvement of $\mathbf{x}_i\boldsymbol{\beta}$ in the constraints, directly imposing those constrains can be inefficient particularly when n is large. In order to address this issue, the augmented Lagrangian method (Hestenes 1969, Powell 1969, Bertsekas 1996) is used as it is capable of handling a large number of constraints. Similar to penalty methods (Zangwill 1967, Fletcher 1973), this augmented Lagrangian method operates by replacing a constrained optimization problem with a series of more manageable unconstrained problems. The difference of this method with penalty methods is that the augmented Lagrangian performs optimization by adding a term to mimic the Lagrange multiplier (Rockafellar 2015) to discourage solutions which do not satisfy the constraints. Then, those small unconstrained optimization problems can be solved by using a suitable solver.

The approach of solving a global problem by a set of local sub problems makes this approach an efficient optimization method (Boyd et al. 2011). Rather than optimizing all the parameters jointly as in method of multipliers (Nocedal & Wright 2006), in this method, it performs one pass of a Gauss-Seidel method (Golub & Van Loan 2012) for each parameter. That means each parameter is updated by fixing all the other parameters at their most current estimates. As oppose to the Jacobi method (Saad 2003) which iterates the updating procedure until convergence, the augmented Lagrangian method then proceeds directly to a dual update step and then repeats the process until convergence.

Let $\boldsymbol{\eta}$ and $\boldsymbol{\gamma}$ be vectors for $(\eta_1, \ldots, \eta_n)$ and $(\gamma_1, \ldots, \gamma_n)$, respectively. Then the augmented Lagrangian function can be written as follows:

$$\mathcal{L} = \sum_{u=1}^{m} \sum_{t_i \in B_u, i \in \mathcal{O}} \log(\theta_u + \eta_i) - \left\{ \sum_{u=1}^{m} w_u \sum_{v=1}^{u-1} (\rho_v \theta_v) + \sum_{u=1}^{m} \sum_{t_i \in B_u} (t_i - \tau_u)\theta_u \right.$$
$$\left. + \sum_{i=1}^{n} \eta_i t_i \right\} - \lambda\, J(\boldsymbol{\theta}) - \sum_{i=1}^{n} \gamma_i (\mathbf{x}_i \boldsymbol{\beta} - \eta_i) - \frac{\alpha}{2} \sum_{i=1}^{n} (\mathbf{x}_i \boldsymbol{\beta} - \eta_i)^2, \quad (3.18)$$

where γ_i is the Lagrange multiplier for imposing the constraint $\eta_i = \mathbf{x}_i \boldsymbol{\beta}$, and the last term in this augmented Lagrangian is a penalty term with $\alpha > 0$. The reasons for having a penalty term are (i) to relax the constraints $\eta_i = \mathbf{x}_i \boldsymbol{\beta}$ in such a way that they are true at convergence of the algorithm, and (ii) to stabilize estimation of the Lagrangian multipliers γ_i during each iteration. In this algorithm, we gradually change the value of α from a small value (say $\alpha_0 > 1$) to a large quantity (such as α_0^{12}). This is useful in defining the convergence criteria for the outer iterations of this algorithm and details about inner and outer iterations will be discussed at the end of this chapter. The Karush-Kuhn-Tucker (KKT) necessary conditions for the constrained MPL estimation of $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ are

$$\frac{\partial \mathcal{L}}{\partial \theta_u} = 0 \quad \text{if } \theta_u > 0 \quad \text{and} \quad \frac{\partial \mathcal{L}}{\partial \theta_u} < 0 \quad \text{if } \theta_u = 0, \quad \text{for } u = 1, 2, \ldots, m, \quad (3.19)$$

$$\frac{\partial \mathcal{L}}{\partial \beta_j} = 0 \quad \text{for all } j. \quad (3.20)$$

From (3.18), we have the following constraint optimization problem

$$(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\eta}}) = \arg\max_{\boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\eta}} [\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\eta})] \quad (3.21)$$

subject to $\theta_u \geq 0$ for $u = 1, 2, ..., m$ and $\sum_{u=1}^{m} \theta_u I(t \in B_u) + \eta_i \geq 0$ for $i = 1, 2, ..., n$.

3.4.2 Details of the algorithm

Here, we propose an iterative algorithm to solve the optimization problem (3.21) under the two non-negativity constraints. In this algorithm, the two model parameters $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ are updated alternately in each iteration. Let $v^{(k)}$ denote the estimate of v at iteration k. Then, the constraint optimization problem under (3.21) can be solved by the following alternative iterative approach.

$$\boldsymbol{\theta}^{(k+1)} = \underset{\boldsymbol{\theta}\geq 0}{\arg\max}\ \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\beta}^{(k)}, \boldsymbol{\eta}^{(k)}) \tag{3.22}$$

$$\boldsymbol{\beta}^{(k+1)} = \underset{\boldsymbol{\beta}}{\arg\max}\ \mathcal{L}(\boldsymbol{\theta}^{(k+1)}, \boldsymbol{\beta}, \boldsymbol{\eta}^{(k)}) \tag{3.23}$$

$$\boldsymbol{\eta}^{(k+1)} = \underset{\eta_i\geq \eta^*}{\arg\max}\ \mathcal{L}(\boldsymbol{\theta}^{(k+1)}, \boldsymbol{\beta}^{(k+1)}, \boldsymbol{\eta}) \quad \text{where } \eta^* = \max_u(-\theta_u^{(k+1)}) \tag{3.24}$$

$$\boldsymbol{\gamma}^{(k+1)} = \boldsymbol{\gamma}^{(k)} + \alpha\,(\mathbf{x}_i\boldsymbol{\beta}^{(k+1)} - \boldsymbol{\eta}^{(k+1)}) \tag{3.25}$$

Here, the multiplicative-iterative (MI) (Ma 2010) algorithm was used to solve equations (3.22) and (3.24), and Newton's algorithm for equation (3.23). The MI algorithm is an efficient non-negatively constrained algorithm which demands only the first derivative of the objective function, making it very easy to derive and implement. Applying the MI algorithm to equation (3.22) ensures that constraint (3.16) is satisfied and constraint (3.17) is imposed through applying Newton's algorithm and the MI algorithm to equations (3.23) and (3.24) respectively. Obtaining estimates for $\boldsymbol{\theta}$, $\boldsymbol{\beta}$, $\boldsymbol{\eta}$ and $\boldsymbol{\gamma}$ at iteration $(k+1)$, given the estimates at k^{th} iteration (i.e.: $\boldsymbol{\theta}^{(k)}$, $\boldsymbol{\beta}^{(k)}$, $\boldsymbol{\eta}^{(k)}$ and $\boldsymbol{\gamma}^{(k)}$), comprise the four steps described below.

3.4.2.1 $\boldsymbol{\theta}$ updating step

First, $\boldsymbol{\theta}^{(k+1)}$ is obtained with $\boldsymbol{\beta}, \boldsymbol{\eta}$ and $\boldsymbol{\gamma}$ fixed at their updates at the k^{th} iteration by performing one iteration of the MI algorithm followed by a line search. The main purpose of performing a line search step is that it guarantees that the objective function $\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\eta})$ increases as a function of $\boldsymbol{\theta}$ when the other parameter estimates remain at their current iteration values.

The first order derivative of the augmented Lagrangian with respect to θ_u is

$$\frac{\partial \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\eta})}{\partial \theta_u} = \sum_{t_i \in B_u, i \in \mathcal{O}} \frac{1}{(\theta_u + \eta_i)} - \sum_{t_i \in B_u} (t_i - \tau_u) - \lambda \left[\frac{\partial J(\boldsymbol{\theta})}{\partial \theta_u} \right] \tag{3.26}$$

for $u = 1, 2, ..., m$. For the purpose of developing the MI algorithm for this context, equation (3.26) was re-arranged since the derivative is zero, in such a way that the resulting equation does not contain any negative terms on both sides:

$$\theta_u \left\{ \sum_{t_i \in B_u} (t_i - \tau_u) + \lambda \left[\frac{\partial J(\boldsymbol{\theta}^{(k)})}{\partial \theta_u} \right]^+ \right\} = \theta_u \left\{ \sum_{t_i \in B_u, i \in \mathcal{O}} \frac{1}{(\theta_u + \eta_i^{(k)})} - \lambda \left[\frac{\partial J(\boldsymbol{\theta}^{(k)})}{\partial \theta_u} \right]^- \right\}. \tag{3.27}$$

The derivative of $f(a)$ evaluated at $a = a^{(k)}$ is denoted by $\partial f(a^{(k)})/\partial a$ with $[a]^+ = \max\{0, a\}$ and $[a]^- = \min\{0, a\}$ such that $a = a^+ + a^-$. From equation (3.27), we suggest updating θ_u by two steps. In the first step, a temporary estimate $\theta_u^{(k+1/2)}$ is computed from $\theta_u^{(k)}$ via the following updating equation:

$$\theta_u^{(k+1/2)} = \theta_u^{(k)} \frac{\sum_{t_i \in B_u, i \in \mathcal{O}} \frac{1}{(\theta_u + \eta_i^{(k)})} - \lambda \left[\frac{\partial J(\boldsymbol{\theta}^{(k)})}{\partial \theta_u} \right]^- + \xi_{1u}}{\sum_{t_i \in B_u} (t_i - \tau_u) + \lambda \left[\frac{\partial J(\boldsymbol{\theta}^{(k)})}{\partial \theta_u} \right]^+ + \xi_{1u}}. \tag{3.28}$$

In order to avoid zero denominator and to improve the convergence speed of the algorithm, a small constant $\xi_{1u}(>0)$ is introduced to both the numerator and the denominator. The choice of ξ_{1u} does not affect the final solution of the algorithm. It can be seen from (3.28) that $\theta_u^{(k+1/2)}$ satisfies the non-negativity constraint if $\boldsymbol{\theta}^{(k)} \geq 0$. However, this $\theta_u^{(k+1/2)}$ may fail to increase $\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\beta}^{(k)}, \boldsymbol{\eta}^{(k)})$ when $\boldsymbol{\theta}$ is moving from $\theta_u^{(k)}$ to $\theta_u^{(k+1/2)}$, leading to possible divergence. Hence, a line search step is required and first we rewrite equation (3.28) as:

$$\theta_u^{(k+1/2)} = \theta_u^{(k)} + S_1(\boldsymbol{\theta}^{(k)}) \frac{\partial \mathcal{L}(\boldsymbol{\theta}^{(k)}, \boldsymbol{\beta}^{(k)}, \boldsymbol{\eta}^{(k)})}{\partial \theta_u}, \tag{3.29}$$

where $S_1(\boldsymbol{\theta}^{(k)})$ is a diagonal matrix with diagonal elements $s_{1u}^{(k)}$, $u = 1, 2, \ldots, m$, where $s_{1u}^{(k)} = \theta_u^{(k)} / w_{1u}^{(k)}$ with $w_{1u}^{(k)}$ represents the denominator of the right hand side of (3.28). Here, $s_{1u}^{(k)}$ is non-negative due to the constraints discussed earlier. Thus, from equation (3.29), we know that the MI step given by (3.28) proceeds along the gradient direction with a non-negative step size.

Finally, the second step of the MI algorithm, $\boldsymbol{\theta}^{(k+1)}$ is obtained as:

$$\theta_u^{(k+1)} = \theta_u^{(k)} + \omega_1^{(k)} \left[\theta_u^{(k+1/2)} - \theta_u^{(k)} \right], \tag{3.30}$$

where $\left[\theta_u^{(k+1/2)} - \theta_u^{(k)}\right]$ is the search direction and $\omega_1^{(k)} \in (0, 1]$ is the corresponding line search step size which ensures $\mathcal{L}(\boldsymbol{\theta}^{(k+1)}, \boldsymbol{\beta}^{(k)}, \boldsymbol{\eta}^{(k)}) \geq \mathcal{L}(\boldsymbol{\theta}^{(k)}, \boldsymbol{\beta}^{(k)}, \boldsymbol{\eta}^{(k)})$, with equality being achieved when the algorithm has converged. The step respects the non-negativity of $\boldsymbol{\theta}^{(k+1)}$. Armijo's rule (Luenberger et al. 1984) can be used to update ω_1. Note that $0 < \omega_1^{(k)} \leq 1$ guarantees that $\theta_u^{(k+1)} \geq 0$ when $\theta_u^{(k)} \geq 0$.

3.4.2.2 $\boldsymbol{\beta}$ updating step

With $\boldsymbol{\theta}$ being fixed at $\boldsymbol{\theta}^{(k+1)}$ at iteration $(k+1)$, $\boldsymbol{\beta}^{(k+1)}$ is obtained by running one iteration of Newton's algorithm followed by a standard line search. The line search step ensures that the objective function $\mathcal{L}(\boldsymbol{\theta}^{(k+1)}, \boldsymbol{\beta}, \boldsymbol{\eta}^{(k)})$ increases with the updated $\boldsymbol{\beta}$ estimate of $\boldsymbol{\beta}^{(k+1)}$. To deliver the Newton update of $\boldsymbol{\beta}$, firstly, consider the first order derivative of the augmented Lagrangian with respect to β_j as follows:

$$\frac{\partial \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\eta})}{\partial \beta_j} = -\sum_{i=1}^{n} \gamma_i x_{ij} - \alpha \sum_{i=1}^{n} (\mathbf{x}_i \boldsymbol{\beta} - \eta_i) x_{ij} \tag{3.31}$$

for $j = 1, 2, ..., p$. Secondly, the second order derivative of $\mathcal{L}$ with respect to β_j and β_t is:

$$\frac{\partial^2 \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\eta})}{\partial \beta_j \partial \beta_t} = -\alpha \sum_{i=1}^{n} x_{ij} x_{it} \tag{3.32}$$

for $j = 1, 2, ..., p$ and $t = 1, 2, ..., p$. One iteration of the Newton's algorithm for solving the KKT necessary condition given in (3.20) for the optimal $\boldsymbol{\beta}$ with line search gives

$$\boldsymbol{\beta}^{(k+1)} = \boldsymbol{\beta}^{(k)} + \omega_2^{(k)} \left[\frac{\partial^2 \mathcal{L}(\boldsymbol{\theta}^{(k+1)}, \boldsymbol{\beta}^{(k)}, \boldsymbol{\eta}^{(k)})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\top} \right]^{-1} \left[\frac{\partial \mathcal{L}(\boldsymbol{\theta}^{(k+1)}, \boldsymbol{\beta}^{(k)}, \boldsymbol{\eta}^{(k)})}{\partial \boldsymbol{\beta}} \right]. \tag{3.33}$$

In this step also, $\omega_2^{(k)} \in (0, 1]$ is the line search step size and we can use Armijo's rule to determine the $\omega_2^{(k)}$'s. Again, the line search step ensures that $\mathcal{L}(\boldsymbol{\theta}^{(k+1)}, \boldsymbol{\beta}^{(k+1)}, \boldsymbol{\eta}^{(k)}) \geq \mathcal{L}(\boldsymbol{\theta}^{(k+1)}, \boldsymbol{\beta}^{(k)}, \boldsymbol{\eta}^{(k)})$ when $\boldsymbol{\beta}$ is moving from $\boldsymbol{\beta}^{(k)}$ to $\boldsymbol{\beta}^{(k+1)}$, where the equality holds only if the iterations have converged.

Here, note that even though Newton's algorithm requires calculation of the second order derivative with respect to $\boldsymbol{\beta}$'s iteratively to update $\boldsymbol{\beta}$, it does not cause slow computation since the second derivative given in (3.32) does not involve updated $\boldsymbol{\beta}$ and it depends only on the covariates. For any given iteration, it results in a constant for a combination of β_j and β_t. This can be considered as an advantage of this proposed method.

3.4.2.3 $\boldsymbol{\eta}$ updating step

In this step, we impose the non-negativity constraint on the overall hazard (constraint (3.16)). Similar to Section 3.4.2.1, the MI algorithm is applied to solve this non-negatively constrained optimization subproblem. The updating process of the new constraint vector $\boldsymbol{\eta}$, using the updated $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ values as follows. Firstly, the first order derivative of $\mathcal{L}$ with respect to η_i is

$$\frac{\partial \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\eta})}{\partial \eta_i} = \sum_{u=1}^{m} \frac{1}{(\theta_u + \eta_i)} I(t_i \in B_u, i \in \mathcal{O}) - t_i + \gamma_i + \alpha(\mathbf{x}_i \boldsymbol{\beta} - \eta_i). \tag{3.34}$$

The derivative given in (3.34) is a function of single η_i. Therefore, this can be updated by subjecting η_i. But, to keep consistency, here we are using the MI algorithm to update η_i.

Since this step deals with handling the non-negativity constraint on $h(t|\mathbf{x}_i)$, it can be presented as:

$$\eta_i \geq -\min_u \,(\theta_u^{(k+1)}) \quad \text{or} \quad \eta_i + \min_u \,(\theta_u^{(k+1)}) \geq 0.$$

In order to follow the MI updating scheme, first the derivative was re-arranged as in Section 3.4.2.1 in such a way that both sides of the resulting equation does not contain any negative terms. Then, the updating step of the intermediate estimate of η_i: ${\eta_i}^{(k+1/2)}$ from ${\eta_i}^{(k)}$ based on the MI algorithm can be given as:

$$\left[\eta_i^{(k+1/2)} + \min_u \,\theta_u^{(k+1)}\right] = \left[\eta_i^{(k)} + \min_u \,\theta_u^{(k+1)}\right]\left\{\frac{\Delta_1}{\Delta_2}\right\}, \tag{3.35}$$

$$\text{where} \quad \Delta_1 = t_i - [\gamma_i]^- - \alpha[(\mathbf{x}_i\boldsymbol{\beta} - \eta_i)]^- + \zeta_{1i} \quad \text{and}$$
$$\Delta_2 = \sum_{i=1}^{m} \frac{1}{(\theta_u + \eta_i)} I(t_i \in B_u, i \in \mathcal{O}) + [\gamma_i]^+ + \alpha[(\mathbf{x}_i\boldsymbol{\beta} - \eta_i)]^+ + \zeta_{1i}.$$

In order to avoid zero denominator and to improve the convergence speed of this MI algorithm, ζ_{1i} is introduced to both the numerator and the denominator. It is clearly seen from (3.35) that $\{\eta_i^{(k+1/2)} + \min_u \,(\theta_u^{(k+1)})\}$ satisfies the non-negativity constraint if $\{\eta_i^{(k)} + \min_u \,(\theta_u^{(k+1)})\} \geq 0$. Even though this maintains $\{\eta_i^{(k+1/2)} + \min_u \,(\theta_u^{(k+1)})\} \geq 0$, it is necessary to ensure that $\mathcal{L}(\boldsymbol{\theta}^{(k+1)}, \boldsymbol{\beta}^{(k+1)}, \boldsymbol{\eta})$ increases when moving from $\eta_i^{(k)}$ to $\eta_i^{(k+1/2)}$. Thus, similar to Section 3.4.2.1 a line search step is performed and first we rewrite the equation (3.35) as:

$$\left[\eta_i^{(k+1/2)} + \min_u \,\theta_u^{(k+1)}\right] = \left[\eta_i^{(k)} + \min_u \,\theta_u^{(k+1)}\right] + S_2(\boldsymbol{\eta}^{(k)})\frac{\partial \mathcal{L}(\boldsymbol{\theta}^{(k+1)}, \boldsymbol{\beta}^{(k+1)}, \boldsymbol{\eta}^{(k)})}{\partial \eta_i}, \tag{3.36}$$

where $S_2(\boldsymbol{\eta}^{(k)})$ is a diagonal matrix with diagonal elements $s_{2i}^{(k)}$, $i = 1, 2, \ldots, n$, where $s_{2i}^{(k)} = (\eta_i^{(k)} + \min_u\ \theta_u^{(k+1)})/w_{2i}^{(k)}$ with $w_{2i}^{(k)}$ represents the denominator of the right hand side of (3.35). Here, $s_{2i}^{(k)}$ is non-negative due to the earlier discussion. Thus, from equation (3.36), it is clear that the MI step given by (3.35) proceeds along the gradient direction with a non-negative step size.

Then, the updated $\boldsymbol{\eta}$: $\boldsymbol{\eta}^{(k+1)}$ can be obtained by a line search as follows:

$$\eta_i^{(k+1)} = \eta_i^{(k)} + \omega_3^{(k)}\left[\eta_i^{(k+1/2)} - \eta_i^{(k)}\right]. \tag{3.37}$$

In order to guarantee that $\eta_i^{(k+1)} \geq 0$ when $\eta_i^{(k)} \geq 0$, we only restrict $\omega_3^{(k)} \in (0, 1]$ and Armijo's rule can be used to determine the values of $\omega_3^{(k)}$'s.

3.4.2.4 γ updating step

The Lagrangian multiplier, $\boldsymbol{\gamma}$ is updated using the standard dual update equation as the final step of this iterative approach, which is updated as

$$\gamma_i^{(k+1)} = \gamma_i^{(k)} + \alpha[\mathbf{x}_i\boldsymbol{\beta}^{(k+1)} - \eta_i^{(k+1)}]. \tag{3.38}$$

The algorithm of these four parameter updating steps (Sections 3.4.2.1 - 3.4.2.4) is performed iteratively until the convergence conditions are satisfied. Convergence of the algorithm is controlled by an inner loop and an outer loop. The inner loop is responsible for enhancing the accuracy of the two model parameters, $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$. The outer loop focuses on minimizing the effect of the penalty term of the augmented Lagrangian in the equation (3.18).

First, in the inner loop, the absolute difference between two adjacent estimated values of $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ are of concern. Then, the convergence criteria for the $(k+1)^{\text{th}}$ inner loop can be given as

$$|\boldsymbol{\theta}^{(k+1)} - \boldsymbol{\theta}^{(k)}| < \ 1 \times 10^{-6} \quad \text{and} \quad |\boldsymbol{\beta}^{(k+1)} - \boldsymbol{\beta}^{(k)}| < \ 1 \times 10^{-6}.$$

For the outer loop, the minimization of $(\mathbf{x}_i\boldsymbol{\beta} - \eta_i)$ is of interest and for that purpose, the convergence criterion is applied for α as mentioned in Section 3.4.1. In the outer iteration $(s+1)$, the value of α is calculated as $\min\{\alpha_0^{s+1}, \alpha_0^{12}\}$ where α_0 is the α value at the 1^{st} outer iteration. With the selection of appropriate α, the above four updating steps are performed iteratively until the convergence criterion for the inner loop satisfied. Once the inner loop satisfies the convergence criterion, then the convergence criterion for the outer loop can be given as

$$\max|\mathbf{x}_i\boldsymbol{\beta} - \eta_i| < 1 \times 10^{-5}.$$

This algorithm is converged once the convergence conditions for these inner and outer loops are satisfied.

In the next chapter, the parameter estimation procedure discussed in this chapter is extended for more general and complex case of interval-censored data.

Chapter 4

MPL parameter estimation approach for additive hazards model with interval-censored survival time data

4.1 Introduction

This chapter presents a parameter estimation approach for the semiparametric additive hazards models with partly interval-censored failure time data. Partly interval-censored data describes the most general form of censored data that can be encountered in time-to-event analysis. Exactly observed event data, right and left-censored data are special cases of interval-censored data as described in Chapter 1. Therefore, the methodology presented in Chapter 3 for right-censored data is a special case of the estimation methodology that will be presented in this chapter. To this end, the estimation steps are presented in the same order as in the previous chapter, but these are inevitably more complex due to interval censoring data.

4.2 Background of the problem

Several approaches are currently available to fit the semiparametric additive hazards model with partly interval-censored data (Farrington 1996, Ghosh 2001, Zeng et al. 2006, Wang et al. 2010) as discussed in Section 2.3.

Farrington (1996) introduced a generalized linear model (GLM) approach for partly interval-censored data assuming that the occurrences of censored observations are from independent Bernoulli trials and the corresponding probability related to a linear predictor by a negative log link function (see Section 2.3.3). The linear predictor includes the baseline hazard and the GLM model regression coefficients which are directly related to the additive model coefficients. Additive hazards model coefficients are then estimated by the GLM model. However, this approach did not guarantee non-negativity or smoothness of the estimated baseline hazard. Ghosh (2001) applied a maximum likelihood (ML) approach to studying the current status data, a special case of interval-censored data and considered the non-negativity constraints on the cumulative baseline hazard. Zeng et al. (2006) also applied a ML approach to estimate the regression coefficients $\boldsymbol{\beta}$ and the baseline survival function $S_0(t)$ for partly interval-censored data. For the estimation process they considered the logarithm transformation to impose monotonically decreasing and positivity constraints on the baseline survival function $S_0(t)$ as discussed in Section 2.3.2. However, due to the logarithm transformation, when $S_0(t)$ approaches zero this estimation procedure is unstable. Wang et al. (2010) implemented a counting process estimation approach for partly interval-censored data. They focused their attention on estimating the regression coefficients $\boldsymbol{\beta}$ while the baseline hazard $h_0(t)$ was considered to be a nuisance parameter and was not estimated. The method proposed in this chapter addresses the limitations of these existing methods. It simultaneously estimates the baseline hazard $h_0(t)$ and the regression coefficients $\boldsymbol{\beta}$ while ensuring that both of the non-negativity constraints are met.

The remainder of this chapter is laid out as follows. Section 4.3 details the MPL function for the additive hazards model with partly interval-censored data. Section 4.4 discusses the simultaneous estimation of $h_0(\cdot)$ and $\boldsymbol{\beta}$ using a constrained optimization algorithm. Suitable basis function is chosen and the non-negativity constraints are imposed in order to ensure the non-negativity of $h_0(\cdot)$ and $h(\cdot)$

during the estimation process which is conducted via a Newton-MI algorithm.

4.3 MPL function for the additive hazards model with partly interval-censored data

This section discusses the MPL approach with partly interval-censored data which can be used to estimate the $\boldsymbol{\beta}$ and $h_0(t)$ simultaneously by considering the two non-negativity constraints pertaining to the additive hazards model.

4.3.1 Model and notations

Firstly, the mathematical notations used for the survival times considered in this and subsequent sections are defined, including fully observed event times, finite interval censoring times and left or right censoring times. Suppose there are n observations in a study and let $\{Y_i : i = 1, ..., n\}$ be the random variables representing time to event of interest. Let the bivariate random vector $\boldsymbol{C}_i = (C_i^L, C_i^R)^\top$ represent the random censoring interval. Here, $0 \leq C_i^L < C_i^R$. It is assumed that Y_i and $\boldsymbol{C}_i$ are independent and cannot be observed simultaneously. Then, the observed survival time for the i^{th} individual is denoted by $\boldsymbol{T}_i = (T_i^L, T_i^R)^\top$. If $\boldsymbol{C}_i$ is observed, then $T_i^L = C_i^L$ and $T_i^R = C_i^R$ and if Y_i is observed, then $T_i^L = T_i^R = Y_i$. Assume further that the $\boldsymbol{T}_i$'s are independent. Observed values for T_i^L and T_i^R are represented by t_i^L and t_i^R.

Now, the survival information for individual i can be denoted by $(t_i^L, t_i^R, \mathbf{x}_i)$ with (t_i^L, t_i^R) representing the observed survival time and $\mathbf{x}_i$ be the $p \times 1$ vector of covariates. It is now possible to illustrate how the fully observed, left and right censoring are special cases of interval censoring. For example, left censoring occurs when $t_i^L = 0$ in the (t_i^L, t_i^R) vector. Likewise, right censoring happens when $t_i^R = \infty$. Uncensored data arise when $t_i^L = t_i^R$. Let δ_i^L, δ_i^I, δ_i^R represent the censoring indicators for left-censored, finite interval-censored and right-censored observations respectively. Partly interval-censored data for subject i can be written as $\{(t_i^L, t_i^R), \delta_i^L, \delta_i^R, \delta_i^I, \mathbf{x}_i\}$. Note that $\delta_i = 1 - (\delta_i^L + \delta_i^I + \delta_i^R)$ represents the indicator for fully observed subjects. Let n_O, n_L, n_I and n_R be the number of subjects with their failure times exactly observed, left-censored, finite interval-censored, and right-censored respectively. Thus, following the above description we have only one time point each of $(n_O + n_L + n_R)$ number of subjects and for n_I number of subjects we have both left and right censoring time points. The assumptions on time-independence of $\mathbf{x}_i$ and independence between the censoring times and the failure times remain the same as in Section 3.3.1.

The semi-parametric additive hazards model defined under (1.17) is considered in this chapter as well. Then, the non-parametric baseline hazard $h_0(t)$ is treated similarly to Section 3.3.1 in order to approximate the infinite dimensional space by a finite dimensional space. Again, in this thesis, the indicator function is selected as the basis function and it results in a piecewise constant approximation to $h_0(t)$. This discretization procedure of $h_0(t)$ is similar to the procedure discussed under Section 3.3.1 except the number of time points observed. In Section 3.3.1, for n subjects we consider n number of time points, whereas in the case of partly interval-censored data we have $(n + n_I)$ number of time points from n subjects. Recall that under the indicator basis functions, the hazard, the cumulative hazard and the survival functions are (see Section 3.3.2):

$$h(t|\boldsymbol{\theta}, \boldsymbol{\beta}, \mathbf{x}_i) = \sum_{u=1}^{m} \theta_u I(\tau_{u-1} < t \leq \tau_u) + \mathbf{x}_i\boldsymbol{\beta} \tag{4.1}$$

$$H(t|\boldsymbol{\theta}, \boldsymbol{\beta}, \mathbf{x}_i) = \sum_{u=1}^{m} \theta_u \Big[(t - \tau_{u-1})I(\tau_{u-1} < t \leq \tau_u) + \rho_u I(t \geq \tau_u)\Big] + \mathbf{x}_i\boldsymbol{\beta}t \tag{4.2}$$

$$S(t|\boldsymbol{\theta}, \boldsymbol{\beta}, \mathbf{x}_i) = \exp[-H(t|\boldsymbol{\theta}, \boldsymbol{\beta}, \mathbf{x}_i)]. \tag{4.3}$$

4.3.2 Maximum penalized likelihood (MPL) estimation

Using equations (4.1) to (4.3), the log-likelihood function is given as

$$\ell(\boldsymbol{\theta}, \boldsymbol{\beta}) = \sum_{i=1}^{n} \Big\{\delta_i[\log h(t_i|\boldsymbol{\theta}, \boldsymbol{\beta}, \mathbf{x}_i) - H(t_i|\boldsymbol{\theta}, \boldsymbol{\beta}, \mathbf{x}_i)] + \delta_i^L \log[1 - S(t_i^L|\boldsymbol{\theta}, \boldsymbol{\beta}, \mathbf{x}_i)]$$
$$+ \delta_i^I \log[S(t_i^L|\boldsymbol{\theta}, \boldsymbol{\beta}, \mathbf{x}_i) - S(t_i^R|\boldsymbol{\theta}, \boldsymbol{\beta}, \mathbf{x}_i)] - \delta_i^R H(t_i^R|\boldsymbol{\theta}, \boldsymbol{\beta}, \mathbf{x}_i)\Big\}. \tag{4.4}$$

In this log-likelihood function, the first three terms correspond to fully observed event times, left-censored times and finite interval-censored times respectively, while the last negative term is for right-censored times. As per Section 3.3.1, a smoothness constraint is applied to $h_0(t)$ by introducing a penalty term. Then, the resulting penalized log-likelihood is,

$$\Phi(\boldsymbol{\theta}, \boldsymbol{\beta}) = \ell(\boldsymbol{\theta}, \boldsymbol{\beta}) - \lambda\, J(\boldsymbol{\theta}). \tag{4.5}$$

Note that the non-negativity constraints for the additive hazards model are $h_0(t) \geq 0$ and $h(t|\mathbf{x}_i) \geq 0$. Thus, we wish to solve the following constrained optimization problem for $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$:

$$(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\beta}}) = \arg \max_{\boldsymbol{\theta}, \boldsymbol{\beta}} \; \Phi(\boldsymbol{\theta}, \boldsymbol{\beta}) \tag{4.6}$$

subject to the conditions,

$$\theta_u \geq 0 \quad \text{for} \;\; u = 1, 2, ..., m \tag{4.7}$$

and

$$\sum_{u=1}^{m} \theta_u I(t \in B_u) + \mathbf{x}_i \boldsymbol{\beta} \geq 0 \quad \text{for} \;\; i = 1, 2, ..., n. \tag{4.8}$$

Parameter estimation can be performed by using a suitable constraint optimization algorithm taking into account these two non-negativity constraints ((4.7) and (4.8)). In next section, we develop an optimization algorithm which is similar to the one in Chapter 3, and has the capability of estimating model parameters simultaneously by considering these constraints.

4.4 Constrained optimization by the augmented Lagrangian method

Similar to the case of right-censored data, since the sample size n can be large, we may have a large number of constraints for the partly interval-censored data as well. Thus, we may have a constrained optimization problem with a potentially large number of constraints depending on the sample size. Therefore, as per the properties discussed in Section 3.4, we again adopt the augmented Lagrangian method to deal with this constrained optimization problem.

4.4.1 Optimization method

Let $\boldsymbol{\eta}$ and $\boldsymbol{\gamma}$ be the vectors for η_i and γ_i, respectively. Then, the augmented Lagrangian function can be obtained from the penalized log-likelihood equation (4.5) as follows:

$$\mathcal{L} = \sum_{i=1}^{n} \Big\{ \delta_i [\log \tilde{h}(t_i) - \tilde{H}(t_i)] + \delta_i^L \log[1 - \tilde{S}(t_i^L)] + \delta_i^I \log[\tilde{S}(t_i^L) - \tilde{S}(t_i^R)] \\ - \delta_i^R \tilde{H}(t_i^R) \Big\} - \lambda\, J(\boldsymbol{\theta}) - \sum_{i=1}^{n} \gamma_i(\mathbf{x}_i\boldsymbol{\beta} - \eta_i) - \frac{\alpha}{2}\sum_{i=1}^{n}(\mathbf{x}_i\boldsymbol{\beta} - \eta_i)^2, \quad (4.9)$$

where γ_i is the Lagrange multiplier for imposing the constraint $\eta_i = \mathbf{x}_i\boldsymbol{\beta}$ and the functions $\tilde{h}(t_i)$, $\tilde{H}(t_i)$ are defined by substituting η_i for $\mathbf{x}_i\boldsymbol{\beta}$ in (4.1) and (4.2). $\tilde{S}(t_i)$ can be obtained as: $\tilde{S}(t_i) = \exp[-\tilde{H}(t_i)]$.

This augmented Lagrangian function is very similar to (3.18) in Section 3.4.1 and, we also move α from a small value to a large quantity as explained in Chapter 3 to assist the convergence. The KKT necessary conditions for the constrained MPL estimation of $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ discussed in Section 3.4.1 are considered in this algorithm as well. Then, from (4.9), the following constraint optimization problem must be solved:

$$(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\eta}}) = \arg\max_{\boldsymbol{\theta},\boldsymbol{\beta},\boldsymbol{\eta}} [\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\eta})] \tag{4.10}$$

subject to $\theta_u \geq 0$ for $u = 1, 2, ..., m$ and $\sum_{u=1}^{m} \theta_u I(t \in B_u) + \eta_i \geq 0$ for $i = 1, 2, ..., n$.

4.4.2 Details of the algorithm

The constrained optimization problem under (4.10) can be solved by an alternative iterative approach when the estimate of u at iteration k is denoted by $u^{(k)}$ as follows:

$$\begin{aligned}
\boldsymbol{\theta}^{(k+1)} &= \arg\max_{\boldsymbol{\theta}\geq 0} \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\beta}^{(k)}, \boldsymbol{\eta}^{(k)}) && (4.11)\\
\boldsymbol{\beta}^{(k+1)} &= \arg\max_{\boldsymbol{\beta}} \mathcal{L}(\boldsymbol{\theta}^{(k+1)}, \boldsymbol{\beta}, \boldsymbol{\eta}^{(k)}) && (4.12)\\
\boldsymbol{\eta}^{(k+1)} &= \arg\max_{\eta_i \geq \eta^*} \mathcal{L}(\boldsymbol{\theta}^{(k+1)}, \boldsymbol{\beta}^{(k+1)}, \boldsymbol{\eta}) \quad \text{where } \eta^* = \max_u(-\theta_u^{(k+1)}) && (4.13)\\
\boldsymbol{\gamma}^{(k+1)} &= \boldsymbol{\gamma}^{(k)} + \alpha\,(\mathbf{x}_i \boldsymbol{\beta}^{(k+1)} - \boldsymbol{\eta}^{(k+1)}). && (4.14)
\end{aligned}$$

Similar to Section 3.4, $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$ are updated using the MI algorithm since it is efficient in handling non-negativity constraints while Newton's algorithm is used to update $\boldsymbol{\beta}$. A standard Lagrange multiplier updating step is used to update $\boldsymbol{\gamma}$. Detailed parameter updating procedures of $\boldsymbol{\theta}$, $\boldsymbol{\beta}$, $\boldsymbol{\eta}$ and $\boldsymbol{\gamma}$ are given below.

4.4.2.1 $\boldsymbol{\theta}$ updating step

Similar to Section 3.4.2.1, $\boldsymbol{\theta}^{(k+1)}$ is firstly obtained by fixing $\boldsymbol{\beta}, \boldsymbol{\eta}$ and $\boldsymbol{\gamma}$ at their updates at the iteration k by performing one iteration of the MI algorithm followed by a line search. Again this updating step ensures that each updated $\boldsymbol{\theta}$ follows the non-negativity constraint given by (4.7). Again as in Section 3.4.2.1, the line search search step guarantees that the objective function $\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\eta})$ increases as a function of $\boldsymbol{\theta}$ when the other estimates remain at their current iteration values.

The first order derivative of $\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\eta})$ with respect to θ_u is:

$$\begin{aligned}\frac{\partial \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\eta})}{\partial \theta_u} = \sum_{i=1}^{n} \Bigg\{ & \delta_i \frac{1}{\tilde{h}(t_i)} \psi_u(t_i) - \delta_i \Phi_u(t_i) + \delta_i^L \frac{\tilde{S}(t_i^L)\Phi_u(t_i^L)}{1 - \tilde{S}(t_i^L)} \\ & - \delta_i^I \frac{\tilde{S}(t_i^L)\Phi_u(t_i^L) - \tilde{S}(t_i^R)\Phi_u(t_i^R)}{\tilde{S}(t_i^L) - \tilde{S}(t_i^R)} - \delta_i^R \Phi_u(t_i^R) \Bigg\} \\ & - \lambda \left[\frac{\partial J(\boldsymbol{\theta})}{\partial \theta_u} \right]\end{aligned} \tag{4.15}$$

for $u = 1, 2, ..., m$. The Karush-Kuhn-Tucker (KKT) necessary conditions for the constrained MPL estimation of $\boldsymbol{\theta}$ is exactly same as (3.19). Then, the parameter $\boldsymbol{\theta}$ is updated using the MI algorithm. For that purpose, equation (4.15) is rearranged such that the resulting equation contains non-negative terms on both sides as follows:

$$\theta_u\left\{\sum_{i=1}^{n}\left\{\Delta_4+\Delta_6+\Delta_8\right\}+\lambda\left[\frac{\partial J(\theta)}{\partial\theta_u}\right]^{+}\right\}$$
$$=\theta_u\left\{\sum_{i=1}^{n}\left\{\Delta_3+\Delta_5+\Delta_7\right\}-\lambda\left[\frac{\partial J(\theta)}{\partial\theta_u}\right]^{-}\right\}, \quad (4.16)$$

where $\Delta_3=\delta_i\frac{1}{\tilde{h}(t_i)}\psi_u(t_i)$ $\quad \Delta_4=\delta_i\Phi_u(t_i)$ $\quad \Delta_5=\delta_i^L\frac{\tilde{S}(t_i^L)\Phi_u(t_i^L)}{1-\tilde{S}(t_i^L)}$

$\Delta_6=\delta_i^I\frac{\tilde{S}(t_i^L)\Phi_u(t_i^L)}{\tilde{S}(t_i^L)-\tilde{S}(t_i^R)}$ $\quad \Delta_7=\delta_i^I\frac{\tilde{S}(t_i^R)\Phi_u(t_i^R)}{\tilde{S}(t_i^L)-\tilde{S}(t_i^R)}$ $\quad \Delta_8=\delta_i^R\Phi_u(t_i^R).$

Here, the notations $[a]^+$ and $[a]^-$ carry the same meanings as in Section 3.4.2.1. Then, the equation (4.16) leads to the following intermediate $\boldsymbol{\theta}$ updating scheme which calculates the value, $\theta^{(k+1/2)}$,

$$\theta_u^{(k+1/2)}=\theta_u^{(k)}\frac{\sum_{i=1}^{n}\left\{\Delta_3+\Delta_5+\Delta_7\right\}-\lambda\left[\frac{\partial J(\theta)}{\partial\theta_u}\right]^{-}+\xi_{2u}}{\sum_{i=1}^{n}\left\{\Delta_4+\Delta_6+\Delta_8\right\}+\lambda\left[\frac{\partial J(\theta)}{\partial\theta_u}\right]^{+}+\xi_{2u}}. \quad (4.17)$$

The constant $\xi_{2u}\geq 0$ is used to avoid a zero denominator and to improve the convergence speed of the algorithm. It is evident that from (4.17), $\theta_u^{(k+1/2)}$ satisfies the non-negativity constraint given in (4.7) if $\boldsymbol{\theta}^{(k)}\geq 0$. However, in order to guarantee that $\mathcal{L}(\boldsymbol{\theta},\boldsymbol{\beta}^{(k)},\boldsymbol{\eta}^{(k)})$ increases when $\boldsymbol{\theta}$ is moving from $\theta_u^{(k)}$ to $\theta_u^{(k+1/2)}$, the equation (4.17) is rewritten in a form similar to equation (3.29) in Section 3.4.2.1. This ensures that the updated $\boldsymbol{\theta}$ from this intermediate step follows the non-negativity constraint and increment of the likelihood.

Finally, $\boldsymbol{\theta}$ is updated to $\boldsymbol{\theta}^{(k+1)}$ using the MI updating step given by:

$$\theta_u^{(k+1)}=\theta_u^{(k)}+\omega_4^{(k)}\left[\theta_u^{(k+1/2)}-\theta_u^{(k)}\right]. \quad (4.18)$$

The line search step size, $0 \leq \omega_4^{(k)} \leq 1$ possess the same characteristics as the line search step size in Section 3.4.2.1 which guarantees the increment of the likelihood: $\mathcal{L}(\boldsymbol{\theta}^{(k+1)}, \boldsymbol{\beta}^{(k)}, \boldsymbol{\eta}^{(k)}) \geq \mathcal{L}(\boldsymbol{\theta}^{(k)}, \boldsymbol{\beta}^{(k)}, \boldsymbol{\eta}^{(k)})$ and the updated $\theta_u^{(k+1)} \geq 0$ when $\theta_u^{(k)} \geq 0$. Here also Armijo's rule is used to update ω_4.

4.4.2.2 $\boldsymbol{\beta}$ updating step

When $\boldsymbol{\theta}$ is fixed at $\boldsymbol{\theta}^{(k+1)}$, the regression coefficient estimator at iteration $(k+1)$, $\boldsymbol{\beta}^{(k+1)}$ is obtained by running one iteration of Newton's algorithm followed by a line search. This assures that $\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\eta})$ increases with the new $\boldsymbol{\beta}$ estimate $\boldsymbol{\beta}^{(k+1)}$.

The overall first order derivative of the augmented Lagrangian for fully observed, left-censored, finite interval-censored and right-censored data with respect to β_j can be computed as follows:

$$\frac{\partial \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\eta})}{\partial \beta_j} = -\sum_{i=1}^{n} \gamma_i x_{ij} - \alpha \sum_{i=1}^{n} (\mathbf{x}_i \boldsymbol{\beta} - \eta_i) x_{ij} \tag{4.19}$$

for $j = 1, 2, ..., p$. Then, the overall second order derivative of the augmented Lagrangian with respect to β_j and β_t can be computed as follows:

$$\frac{\partial^2 \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\eta})}{\partial \beta_j \partial \beta_t} = -\alpha \sum_{i=1}^{n} x_{ij} x_{it} \tag{4.20}$$

for $j = 1, 2, ..., p$ and $t = 1, 2, ..., p$.

Thus, one iteration of Newton's algorithm is performed to solve the KKT necessary condition for the constrained MPL estimation of $\boldsymbol{\beta}$ given in (3.20). From equations (4.19) and (4.20), the updated estimate of $\boldsymbol{\beta}^{(k+1)}$ derived from $\boldsymbol{\beta}^{(k)}$ is given by

$$\boldsymbol{\beta}^{(k+1)} = \boldsymbol{\beta}^{(k)} + \omega_5^{(k)} \left[\frac{\partial^2 \mathcal{L}(\boldsymbol{\theta}^{(k+1)}, \boldsymbol{\beta}^{(k)}, \boldsymbol{\eta}^{(k)})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\top} \right]^{-1} \left[\frac{\partial \mathcal{L}(\boldsymbol{\theta}^{(k+1)}, \boldsymbol{\beta}^{(k)}, \boldsymbol{\eta}^{(k)})}{\partial \boldsymbol{\beta}} \right]. \tag{4.21}$$

Here, to update the regression parameters $\boldsymbol{\beta}$, Armijo's rule is used and the selection of ω_5 assures that $\mathcal{L}(\boldsymbol{\theta}^{(k+1)}, \boldsymbol{\beta}^{(k+1)}, \boldsymbol{\eta}^{(k)}) \geq \mathcal{L}(\boldsymbol{\theta}^{(k+1)}, \boldsymbol{\beta}^{(k)}, \boldsymbol{\eta}^{(k)})$ when $\boldsymbol{\beta}$ is moving from $\boldsymbol{\beta}^{(k)}$ to $\boldsymbol{\beta}^{(k+1)}$. As mentioned in Section 3.4.2.2, the $\boldsymbol{\beta}$ updating step in partly interval-censored data also does not result slow computation even though it requires the second derivative.

4.4.2.3 $\boldsymbol{\eta}$ updating step

Update of $\boldsymbol{\eta}$ is carried out exactly same as in Section 3.4.2.3 by performing one iteration of the MI algorithm. In this step, we impose the non-negativity constraint on the overall hazard (constraint (4.8)). The first order derivative of the augmented Lagrangian with respect to $\boldsymbol{\eta}_i$ is

$$\frac{\partial \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\eta})}{\partial \eta_i} = \delta_i \frac{1}{\tilde{h}(t_i)} - \delta_i t_i + \delta_i^L \frac{\tilde{S}(t_i^L) t_i^L}{[1 - \tilde{S}(t_i^L)]} - \delta_i^I \frac{[\tilde{S}(t_i^L) t_i^L - \tilde{S}(t_i^R) t_i^R]}{[\tilde{S}(t_i^L) - \tilde{S}(t_i^R)]} - \delta_i^R t_i^R + \gamma_i + \alpha\left(\mathbf{x}_i \boldsymbol{\beta} - \eta_i\right). \tag{4.22}$$

This step requires handling the constraint on $h(t|\mathbf{x}_i)$, which can be presented as: $\eta_i + \min_u\, (\theta_u^{(k+1)}) \geq 0$. In order to perform the MI algorithm to handle this non-negativity constraint, the derivative given in equation (4.22) was first re-arranged as in Section 3.4.2.3 in such a way that the both sides of the resulting equation contains only non-negative terms. Then, the updating step of intermediate estimate $\eta_i^{(k+1/2)}$ based on MI algorithm can be given as:

$$[\eta_i^{(k+1/2)} + \min_u \ \theta_u^{(k+1)}] = [\eta_i^{(k)} + \min_u \ \theta_u^{(k+1)}]\left\{\frac{\Delta_9}{\Delta_{10}}\right\}, \tag{4.23}$$

where $\Delta_9 = \delta_i t_i + \delta_i^I \frac{[\tilde{S}(t_i^L)t_i^L}{[\tilde{S}(t_i^L) - \tilde{S}(t_i^R)]} + \delta_i^R t_i^R - [\gamma_i]^- - \alpha[(\mathbf{x}_i\boldsymbol{\beta} - \eta_i)]^- + \zeta_{2i}$ and $\Delta_{10} = \delta_i \frac{1}{\tilde{h}(t_i)} + \delta_i^L \frac{\tilde{S}(t_i^L)t_i^L}{[1 - \tilde{S}(t_i^L)]} + \delta_i^I \frac{\tilde{S}(t_i^R)t_i^R}{[\tilde{S}(t_i^L) - \tilde{S}(t_i^R)]} + [\gamma_i]^+ + \alpha[(\mathbf{x}_i\boldsymbol{\beta} - \eta_i)]^+ + \zeta_{2i}$.

ζ_{2i} is introduced to avoid a zero denominator and to improve the convergence speed of this MI algorithm. This updated $\left\{\eta_i^{(k+1/2)} + \ \min_u \ \theta_u^{(k+1)}\right\}$ satisfies the non-negativity constraint as explained in Section 3.4.2.3. Then, in order to ensure that $\mathcal{L}(\boldsymbol{\theta}^{(k+1)}, \boldsymbol{\beta}^{(k+1)}, \boldsymbol{\eta})$ increases when moving from $\eta_i^{(k)}$ to $\eta_i^{(k+1/2)}$, equation (4.23) is rewritten in a form similar to equation (3.36) in Section 3.4.2.3.

Then, $\boldsymbol{\eta}^{(k+1)}$ is computed by a line search as

$$\eta_i^{(k+1)} = \eta_i^{(k)} + \omega_6^{(k)}\left[\eta_i^{(k+1/2)} - \eta_i^{(k)}\right]. \tag{4.24}$$

In order to assure that $\eta_i^{(k+1)} \geq 0$ when $\eta_i^{(k)} \geq 0$, we restrict $\omega_6^{(k)} \in (0, 1]$ and Armijo's rule can be used to determine the search step size updates.

4.4.2.4 $\boldsymbol{\gamma}$ updating step

As the final step, the value of the augmented Lagrangian multiplier $\boldsymbol{\gamma}$ is updated by using the standard dual update equation as

$$\gamma_i^{(k+1)} = \gamma_i^{(k)} + \alpha[\mathbf{x}_i\boldsymbol{\beta}^{(k+1)} - \eta_i^{(k+1)}]. \tag{4.25}$$

With this iterative parameter estimation (Sections 4.4.2.1-4.4.2.4), it is guaranteed that this MPL method estimates both $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ simultaneously by ensuring the non-negativity of the two constraints. These four parameter updating steps are performed iteratively until both inner and outer loop convergence conditions are satisfied. The value of $\alpha(> 0)$ is determined by the inner and outer iterations of the algorithm as explained at the end of Chapter 3. The convergence conditions for the algorithm for partly interval-censored data remain the same as the convergence criteria considered in Chapter 3. The asymptotic properties of this parameter estimation method are presented in Chapter 5. Simulation studies for the right-censored and interval-censored survival data are presented in Chapter 6.

Chapter 5

Asymptotic properties

5.1 Introduction

The asymptotic properties of the derived constrained MPL estimates for $\boldsymbol{\beta}$ and $h_0(t)$ are discussed in this chapter. We will investigate the asymptotic convergence and asymptotic distributions of MPL estimates when (i) the number of basis functions, m tends to infinity, and (ii) m is fixed. Section 5.2 presents notation and a mathematical framework which will be used through out this chapter. In Section 5.3, the asymptotic consistency of the estimates $\hat{\boldsymbol{\beta}}$ and $\hat{h}_0(t)$ when the number of distinctive knots m (number of basis functions) $\to \infty$, but $m/n \to 0$ and $\mu_n = \lambda/n \to 0$ when $n \to \infty$, is discussed. Section 5.4 presents the asymptotic results for the constrained MPL estimates when m is fixed.

5.2 Notations and mathematical framework

When $a = \min_i t_i^L$ and $b = \max_i t_i^R$, it is assumed that the baseline hazard $h_0(t)$ is a bounded function and has $q(\geq 1)$ bounded derivatives over $[a, b]$, which is denoted by $C^q[a, b]$. The parameter space for $h_0(t)$ is considered to be $A = \{h_0(t) : h_0(t) \in C^q[a, b], 0 \leq h_0(t) \leq J_1 < \infty, \forall t \in [a, b]\}$. Here we assume that the $h_0(t)$ is bounded and the upper bound J_1 exists. The parameter space for $\boldsymbol{\beta}$ is given by $B = \{\boldsymbol{\beta} : |\beta_j| \leq J_2 < \infty, \forall j\}$, where for any $\boldsymbol{\beta} \in B$, the boundary $J_2 = \max \parallel \boldsymbol{\beta} \parallel$. This is a compact finite dimensional parameter subspace for R^p. Let $\boldsymbol{\pi} = (h_0(t), \boldsymbol{\beta})$. Then, the overall parameter space is given by $\boldsymbol{\Pi} = \{\boldsymbol{\pi} : h_0(t) \in A, \boldsymbol{\beta} \in B\} = A * B$.

In this section, the baseline hazard $h_0(t)$ is approximated by $h_n(t) = \sum_{u=1}^{m} \theta_{un} I(t \in B_u)$, where I is an indicator function for a time point in bin u, and θ_{un} are assumed to be non-negative and bounded. Thus, a finite dimensional space for the approximated baseline hazard $h_n(t)$ is defined by $A_n = \{h_n(t) : 0 \leq h_n(t) \leq J_3 < \infty, \forall t \in [a, b]\}$, where J_3 is the upper bound and it exists. The parameter space for $\boldsymbol{\pi}_n = (h_n(t), \boldsymbol{\beta})$ can be denoted by $\boldsymbol{\Pi}_n = \{\boldsymbol{\pi}_n : h_n \in A_n, \boldsymbol{\beta} \in B\} = A_n * B$. The corresponding MPL estimates for $\boldsymbol{\pi}_n$ is denoted by $\hat{\boldsymbol{\pi}}_n = (\hat{h}_n(t), \hat{\boldsymbol{\beta}})$. Throughout this thesis, we consider $\parallel \boldsymbol{a} \parallel$ to be the Euclidean norm of a vector $\boldsymbol{a}$ and $\parallel \boldsymbol{a}_1 - \boldsymbol{a}_2 \parallel_2^2$ denotes the square of the Euclidean distance between two vectors $\boldsymbol{a}_1$

and $\boldsymbol{a}_2$.

Next, with the use of censoring indicators, left and right survival time vectors and covariates, a combined random vector $\mathbf{Z}_i$ is defined as follows:

$$\mathbf{Z}_i = (\delta_i, \delta_i^L, \delta_i^R, \delta_i^I, T_i^L, T_i^R, \mathbf{x}_i)^\top, \quad \text{for } i = 1, 2, ..., n$$

and the random vectors $\mathbf{Z}_i$ are assumed to be i.i.d. The density function for any $\mathbf{Z}_i$ is

$$f(z_i) = [h_i(t_i)S_i(t_i)]^{\delta_i}[1 - S_i(t_i^L)]^{\delta_i^L} S_i(t_i^R)^{\delta_i^R}[S_i(t_i^L) - S_i(t_i^R)]^{\delta_i^I}\zeta(\mathbf{x}_i), \tag{5.1}$$

where ζ denotes the density function of $\mathbf{x}_i$ which is independent of $\boldsymbol{\beta}$ and $h_0(t)$, and $t_i^L = t_i^R = t_i$ when $\delta_i = 1$. Let $\mathbf{Z}$ represent a general $\mathbf{Z}_i$ and F be the cumulative distribution function of $\mathbf{Z}$. The log-likelihood functions of $\mathbf{Z}$ corresponding to spaces $\boldsymbol{\Pi}$ and $\boldsymbol{\Pi}_n$ are denoted by $\ell(\boldsymbol{\pi}; \mathbf{Z})$ and $\ell(\boldsymbol{\pi}_n; \mathbf{Z})$ respectively and the class $\boldsymbol{\Lambda}_n$ can be defined as $\boldsymbol{\Lambda}_n = \{\ell(\boldsymbol{\pi}_n; \mathbf{Z}) : \boldsymbol{\pi}_n \in \boldsymbol{\Pi}_n\}$. For $\boldsymbol{\pi} \in \boldsymbol{\Pi}$, the actual mean and the empirical mean are denoted as $P\ell(\boldsymbol{\pi}) = \int \ell(\boldsymbol{\pi}; \mathbf{z})dF(\mathbf{z})$ and $P_n\ell(\boldsymbol{\pi}) = \frac{1}{n}\sum_{i=1}^n \ell(\boldsymbol{\pi}; \mathbf{Z}_i)$ respectively. For $\boldsymbol{\pi}_n \in \boldsymbol{\Pi}_n$, the actual mean $P\ell(\boldsymbol{\pi}_n)$ and the empirical mean $P_n\ell(\boldsymbol{\pi}_n)$ are similarly defined.

Let $h_{00}(t)$ and $\boldsymbol{\beta}_0$ to be the true values of $h_0(t)$ and $\boldsymbol{\beta}$ respectively, which maximize $P\ell(\boldsymbol{\pi})$. One common method to parameterize the nonparametric component is to use the sieve estimators (Shen & Wong (1994); Huang & Rossini (1997); Shen (1997)). The sieve method approximates an infinite dimensional parameter space $\boldsymbol{\Pi}$ by a series of finite dimensional parameter spaces $\boldsymbol{\Pi}_n$, which involves estimating parameters on $\mathbf{\Pi}_n$, not on $\mathbf{\Pi}$. The estimation depends on the sample size n and the approximation error must decrease to zero as the sample size increases (Shen & Wong 1994). According to Grenander (1981), such a sequence of approximating spaces $\mathbf{\Pi}_n$ is called a *sieve*, and the maximizer of $P_n\ell(\boldsymbol{\pi})$ over $\mathbf{\Pi}_n$ is called the *sieve MLE*. That means, the maximizer $\hat{\boldsymbol{\pi}}_n$ could be referred to as *sieve MLE* if $\hat{\boldsymbol{\pi}}_n$ maximizes $P_n\ell(\boldsymbol{\pi}_n)$ and satisfies the condition,

$$P_n\ell(\hat{\boldsymbol{\pi}}_n) \geq \sup_{\boldsymbol{\pi}\in\mathbf{\Pi}_n} P_n\ell(\boldsymbol{\pi}) - \xi_n.$$

The rate at which ξ_n reaches 0 does not affect the asymptotic properties of the sieve estimator $\hat{\boldsymbol{\pi}}_n$.

The concepts discussed in this section are used in deriving asymptotic results of the MPL estimates in the following sections. In the next section, the consistency of the MPL estimates $\hat{h}_n(t)$ and $\hat{\boldsymbol{\beta}}$ when $m \to \infty$, $m/n \to 0$ and the penalty parameter $\mu_n = \lambda/n \to 0$ is considered.

5.3 Asymptotic results when $m \to \infty$, but $m/n \to 0$ and $\mu_n \to 0$

In this section, a discussion on the asymptotic properties of the MPL estimates when the number of distinctive knots $m \to \infty$, but the rate is slower than the rate of $n \to \infty$ is presented. In particular, it is assumed that $\mu_n = o(n^{-1/2})$. When the baseline hazard estimate is $\hat{h}_n(t) = \sum_{i=1}^{m} \hat{\theta}_{un} I(t \in B_u)$, let $\hat{\boldsymbol{\theta}}_n$ and $\hat{\boldsymbol{\beta}}$ be the MPL estimates of $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ with $\hat{\boldsymbol{\theta}}_n = (\hat{\theta}_{1n}, ..., \hat{\theta}_{mn})^\top$. The asymptotic results for the estimates $\hat{h}_n(t)$ and $\hat{\boldsymbol{\beta}}$ is presented in Theorem 5.3.1. The results require the following assumptions.

Assumptions:

A1. Design matrix $\mathbf{X}$ is bounded and $E(\mathbf{X}\mathbf{X}^\top)$ is non-singular.

A2. The penalty function J is bounded over the spaces $\mathbf{\Pi}$ and $\mathbf{\Pi}_n$.

A3. For the approximated baseline hazard function $h_n(t)$, assume its coefficient vector $\boldsymbol{\theta}_n$ is in a compact subset of R^m.

A4. The basis functions and their associated knots are selected in a way such that for a $h_0(t) \in A$, there exists $h_n(t) \in A_n$ such that $\max_t |h_n(t) - h_0(t)| \to 0$ as $n \to \infty$.

The claim on Assumption A4 can be achieved under certain regularity conditions. For an example, when $h_0(t)$ is approximated using a B-spline basis function, the difference between the approximated and true baseline hazard values reaches to zero when the bin size reaches to minuscule, which means the number of bins, n reaches to ∞. For more information refer to Marsden (1968) and Proposition 2.8 in De Boor & Daniel (1974).

Theorem 5.3.1 *Suppose that the Assumptions A1 - A4 are satisfied and assume the baseline hazard $h_0(t)$ is a bounded function and has $q(\geq 1)$ bounded derivatives over $[a,b]$. When $\mu_n \to 0$ as $n \to \infty$ and $m = \lfloor n^\nu \rfloor$, where $0 < \nu < 1$, then for $n \to \infty$:*

1. $\sup_{t\in[a,b]}|\hat{h}_n(t) - h_{00}(t)| \xrightarrow{a.s.} 0$,
2. $\| \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 \| \xrightarrow{a.s.} 0$.

Proof:

The proof of the above theorem given below follows the proofs in Huang (1996), Zhang et al. (2010) and Xue et al. (2004). Recall that the combined parameter vector $\boldsymbol{\pi} = (h_0(t), \boldsymbol{\beta})$ with $\boldsymbol{\pi} \in A * B$ and the combined parameter vector in finite dimensional space $\boldsymbol{\pi}_n = (h_n(t), \boldsymbol{\beta})$ with $\boldsymbol{\pi}_n \in A_n * B \subset A * B$. Let $\mathcal{D}(\boldsymbol{\pi}_1, \boldsymbol{\pi}_2)$ be the distance measure for two parameter vectors $\boldsymbol{\pi}_1, \boldsymbol{\pi}_2$ defined by,

$$\mathcal{D}(\boldsymbol{\pi}_1, \boldsymbol{\pi}_2) = \{\| \boldsymbol{\pi}_1 - \boldsymbol{\pi}_2 \|^2\}^{1/2} = \left\{ \sup_{t\in[a,b]} |h_{01}(t) - h_{02}(t)|^2 + \| \boldsymbol{\beta}_1 - \boldsymbol{\beta}_2 \|_2^2 \right\}^{1/2}. \quad (5.2)$$

This proof needs the concept of a covering number in a space. According to Van Der Vaart & Wellner (1996), the *covering number* $N(\varepsilon, \mathcal{F}, \| \cdot \|)$ is the minimal number of balls $\{g : \| g - f \| < \varepsilon\}$ of radius ε needed to cover the set $\mathcal{F}$.

To prove Theorem 5.3.1 , it is required to demonstrate that the distance measure between $\boldsymbol{\pi}_0$ and $\hat{\boldsymbol{\pi}}_n$: $\mathcal{D}(\boldsymbol{\pi}_0, \hat{\boldsymbol{\pi}}_n)$ converges to zero ($\to 0$) almost surely (a.s.). Since the penalty parameter $\mu_n \to 0$ when $n \to \infty$ and the penalty function J is bounded, it is required to focus only on the log-likelihood function. The following steps are used to obtain the results stated in Theorem 5.3.1.

Step 1:

Denote the density function in (5.1) by $f(\boldsymbol{\pi};\mathbf{Z})$ to emphasise its dependence on $\boldsymbol{\pi}$. Let $p(\boldsymbol{\pi};\mathbf{Z})$ be the corresponding Fréchet derivative (Dieudonné 2013) of $f(\boldsymbol{\pi};\mathbf{Z})$ with respect to $\boldsymbol{\pi}$ and ϑ be a point between $\boldsymbol{\pi}_0$ and $\hat{\boldsymbol{\pi}}_n$. Applying the mean value theorem to $f^{\frac{1}{2}}(\boldsymbol{\pi};\mathbf{Z})$ gives,

$$\begin{aligned}
|P\ell(\boldsymbol{\pi}_0;\mathbf{Z}) - P\ell(\hat{\boldsymbol{\pi}}_n;\mathbf{Z})| &= E_0(\ell(\boldsymbol{\pi}_0;\mathbf{Z}) - \ell(\hat{\boldsymbol{\pi}}_n;\mathbf{Z})) \\
&\geq \| f^{\frac{1}{2}}(\boldsymbol{\pi}_0;\mathbf{Z}) - f^{\frac{1}{2}}(\hat{\boldsymbol{\pi}}_n;\mathbf{Z}) \|_2^2 \qquad (5.3) \\
&= \left\| \left. \frac{p(\boldsymbol{\pi};\mathbf{Z})}{2f^{\frac{1}{2}}(\boldsymbol{\pi};\mathbf{Z})} \right|_{\boldsymbol{\pi}=\vartheta} (\boldsymbol{\pi}_0 - \hat{\boldsymbol{\pi}}_n) \right\|_2^2 \\
&\geq J_4 \| \boldsymbol{\pi}_0 - \hat{\boldsymbol{\pi}}_n \|_2^2, \qquad (5.4)
\end{aligned}$$

where the first inequality exists since the Kullback-Leibler distance (Kullback & Leibler 1951) is not less than the square of the Hellinger distance (Wong & Shen 1995). E_0 refers to the expectation with respect to $f(\pi_0;\mathbf{Z})$ and since $f(\boldsymbol{\pi};\mathbf{Z})$ is non-zero and bounded, the function $\left.\frac{p(\boldsymbol{\pi};\mathbf{Z})}{2f^{\frac{1}{2}}(\boldsymbol{\pi};\mathbf{Z})}\right|_{\boldsymbol{\pi}=\vartheta}$ is also bounded. J_4 is the lower bound of $|p(\vartheta;\mathbf{Z})/2f^{\frac{1}{2}}(\vartheta;\mathbf{Z})|$. Thus, in order to show $\mathcal{D}(\boldsymbol{\pi}_0,\hat{\boldsymbol{\pi}}_n) \to 0$ (a.s.), it is sufficient to show that $|P\ell(\boldsymbol{\pi}_0) - P\ell(\hat{\boldsymbol{\pi}}_n)| \to 0$ (a.s.).

Step 2:

This step and the subsequent steps prove $|P\ell(\boldsymbol{\pi}_0) - P\ell(\hat{\boldsymbol{\pi}}_n)| \to 0$ (a.s.). To simplify the proof of this convergence, triangular inequality is obtained as follows:

$$|P\ell(\boldsymbol{\pi}_0) - P\ell(\hat{\boldsymbol{\pi}}_n)| \leq \ |P\ell(\boldsymbol{\pi}_0) - P_n\ell(\hat{\boldsymbol{\pi}}_n)| + |P_n\ell(\hat{\boldsymbol{\pi}}_n) - P\ell(\hat{\boldsymbol{\pi}}_n)|. \tag{5.5}$$

Then, we just need to show that each term on the right hand side of equation (5.5) converges to 0 almost surely. Define $\boldsymbol{\pi}_{0n} \in A_n * B$ which satisfies, for the sieve space chosen, $\mathcal{D}(\boldsymbol{\pi}_{0n}, \boldsymbol{\pi}_0) \to 0$ (a.s.) as $n \to \infty$ according to assumption A4. Since the true parameter vector $\boldsymbol{\pi}_0$ maximizes the true mean $P\ell(\boldsymbol{\pi})$ for $\boldsymbol{\pi} \in A * B$ and the estimate $\hat{\boldsymbol{\pi}}_n$ maximizes the empirical mean $P_n\ell(\boldsymbol{\pi})$ for $\boldsymbol{\pi}_n \in A_n * B$, we get

$$P\ell(\boldsymbol{\pi}_0) - P\ell(\boldsymbol{\pi}_{0n}) + P\ell(\boldsymbol{\pi}_{0n}) - P_n\ell(\boldsymbol{\pi}_{0n}) \geq P\ell(\boldsymbol{\pi}_0) - P_n\ell(\hat{\boldsymbol{\pi}}_n) \geq P\ell(\hat{\boldsymbol{\pi}}_n) - P_n\ell(\hat{\boldsymbol{\pi}}_n). \tag{5.6}$$

In equation (5.6), the middle term is sandwiched by the two terms on both sides. Then, it is necessary to show that the two terms on both sides of the equation (5.6) converge to zero (a.s.) in order to claim that the middle term converges to zero (a.s.).

According to assumption A4, for the selected sieve space above, it is guaranteed that $P\ell(\boldsymbol{\pi}_0) - P\ell(\boldsymbol{\pi}_{0n})$ converges to 0 (a.s.). Next, we need to show that $\{P\ell(\hat{\boldsymbol{\pi}}_n) - P_n\ell(\hat{\boldsymbol{\pi}}_n)\}$ and $\{P\ell(\boldsymbol{\pi}_{0n}) - P_n\ell(\boldsymbol{\pi}_{0n})\}$ converge to 0 (a.s.).

Step 3:

This step demonstrates $\sup_{\boldsymbol{\pi}_n \in A_n * B} |P\ell(\boldsymbol{\pi}_n) - P_n\ell(\boldsymbol{\pi}_n)| \to 0$ (a.s.) using the following steps:

(a) Firstly, it can be shown that $N(\varepsilon, A_n, L_\infty) \leq (6M_0/\varepsilon)^m$ where M_0 is a constant and it is obtained as follows. For any $h_{01}, h_{02} \in A_n$ (where $h_{0i}(t) = \sum_{u=1}^{m} \theta_{iu} I(t \in B_u)$) there exists a relationship such that $\max_t |h_{01}(t) - h_{02}(t)| \leq \max_u |\theta_{1u} - \theta_{2u}| \leq \parallel \boldsymbol{\theta}_1 - \boldsymbol{\theta}_2 \parallel$. This relationship is achieved since θ_{iu} is bounded and consequently bounded $h_{0i}(t)$. Hence, by following Lemma A.2 of Xue et al. (2004), $N(\varepsilon, A_n, L_\infty) \leq N(\varepsilon, \{0 \leq \theta_u \leq M_0, 1 \leq u \leq m\}, L_2)$. Then, by Lemma 4.1 of Pollard (1990), we can obtain $N(\varepsilon, \{0 \leq \theta_u \leq M_0, 1 \leq u \leq m\}, L_2) \leq (6M_0/\varepsilon)^m$.

(b) Next, by following the Taylor's series expansion presented in Xue et al. (2004), it can be shown that $N(\varepsilon, \boldsymbol{\pi}_{0n} \in A_n * B, L_\infty) \leq K/\varepsilon^{m+p}$, where K is a constant and it can be specified as follows: $N(\varepsilon, A_n * B, L_\infty) \leq \Big[N(\varepsilon/2, A_n, L_\infty) \cdot N(\varepsilon/2, B, L_2)\Big] \leq (12M_0/\varepsilon)^m (6M_1/\varepsilon)^p = K/\varepsilon^{m+p}$ where $K = (12M_0)^m (6M_1)^p$.

(c) Then, select $\alpha_n = n^{-1/2+\phi_1}(\log n)^{1/2}$ where $\phi_1 \in (\phi_0/2, 1/2)$ with $\phi_0 < 1$ in such a way that α_n is a non-increasing series. Let $\varepsilon_n = \varepsilon\alpha_n$ for any fixed ε. According to the proof of Theorem 1 of Xue et al. (2004) and by following the results of Pollard (1984), for any $P_n\ell(\boldsymbol{\pi}_n) \in \boldsymbol{\Lambda}_n$ which has been defined in Section 5.2, it can be derived that $\mathrm{var}[P_n\ell(\boldsymbol{\pi}_n)]/(16\varepsilon_n^2) \leq \frac{(1/n)Pl^2}{16\varepsilon^2\alpha_n^2} \leq \frac{C}{16n\varepsilon^2\alpha_n^2}$. Then, it is possible to claim that $\mathrm{var}[P_n\ell(\boldsymbol{\pi}_n)]/(16\varepsilon_n^2) \to 0$ (a.s.) as $n \to \infty$.

(d) From the results in Steps 3((a)-(c)), and following the results of Chapter II equation 31 of Pollard (1984) along with same arguments of Xue et al. (2004), it is possible to show that $\sum_{n=1}^{\infty} P\{\sup_{A_n * B} |P\ell(\boldsymbol{\pi}_n) - P_n\ell(\boldsymbol{\pi}_n)| > 8\varepsilon_n\} < \infty$.

Thus, this is a convergent series and by the Borel-Cantelli lemma (Feller 2008), $\sup_{A_n * B}|P\ell(\boldsymbol{\pi}_n) - P_n\ell(\boldsymbol{\pi}_n)| \to 0$ (a.s.).

Following the above proof in Step 3, $\{P_n\ell(\hat{\boldsymbol{\pi}}_n) - P\ell(\hat{\boldsymbol{\pi}}_n)\}$ in equation (5.5) and $\{P\ell(\boldsymbol{\pi}_{0n}) - P_n\ell(\boldsymbol{\pi}_{0n})\}$, $\{P\ell(\hat{\boldsymbol{\pi}}_n) - P_n\ell(\hat{\boldsymbol{\pi}}_n)\}$ in equation (5.6) converge to 0 (a.s.). By selecting $\boldsymbol{\pi}_{0n}$ as in Step 2, $P\ell(\boldsymbol{\pi}_0) - P\ell(\boldsymbol{\pi}_{0n}) \to 0$ (a.s.). Thus, according to these results, the middle term in equation (5.6): $|P\ell(\boldsymbol{\pi}_0) - P_n\ell(\hat{\boldsymbol{\pi}}_n)| \to 0$ (a.s.). Consequently, the two terms on the right side of equation (5.5) converge to 0 (a.s.). Therefore, it can be shown that $|P\ell(\boldsymbol{\pi}_0) - P\ell(\hat{\boldsymbol{\pi}}_n)| \to 0$ (a.s.) and, since it is the upper bound for $\mathcal{D}(\boldsymbol{\pi}_0, \hat{\boldsymbol{\pi}}_n)$, according to equation (5.4), the distance measure $\mathcal{D}(\boldsymbol{\pi}_0, \hat{\boldsymbol{\pi}}_n) \to 0$ (a.s.).

The consistency results in Theorem 5.3.1 can be further developed to obtain a convergence rate for the two estimates of $\boldsymbol{\beta}$ and $h_n(t)$ and then to show the asymptotic normality of the estimator of $\boldsymbol{\beta}$. This can be done by following the similar type of work for the Cox model by Huang (1996), Xue et al. (2004) and Zhang et al. (2010). But, those results are less useful in practice, even though they are quite important from a theoretical aspect. The reasons these asymptotic results are less popular are mainly: (i) the results do not consider the fact that some baseline hazard estimates can be zero, (ii) the computational inefficiency of calculation of the variance covariance matrix of $\hat{\boldsymbol{\beta}}$. Due to this impracticality of asymptotic normality when $m \to \infty$, useful asymptotic results of the estimators of $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$, when m is fixed and for non-zero μ_n, are developed and given in the next section.

5.4 Asymptotic properties for constrained MPL estimates with fixed m

In this section, the two model parameters are combined into a single vector $\boldsymbol{\sigma} = (\boldsymbol{\theta}^\top, \boldsymbol{\beta}^\top)^\top$, with dimension of $m + p$, in order to show the asymptotic properties. Note that we are using the basis coefficient vector $\boldsymbol{\theta}$ which is equivalent to $h_0(t)$ when it is approximated with finite m. Consider $\boldsymbol{\sigma}_0 = (\boldsymbol{\theta}_0^\top, \boldsymbol{\beta}_0^\top)^\top$ to be its associated true parameter vector and $\hat{\boldsymbol{\sigma}} = (\hat{\boldsymbol{\theta}}^\top, \hat{\boldsymbol{\beta}}^\top)^\top$ to be the MPL estimate of $\boldsymbol{\sigma}$. During this asymptotic derivation, the interval censored data representation discussed in Section 5.2 is considered again. Then, with the inclusion of the new combined parameter vector $\boldsymbol{\sigma}$, the joint density function introduced in (5.1) can be rewritten as:

$$f(z_i|\boldsymbol{\sigma}) = [h_i(t_i|\boldsymbol{\sigma})S_i(t_i|\boldsymbol{\sigma})]^{\delta_i}[1 - S_i(t_i^L|\boldsymbol{\sigma})]^{\delta_i^L} S_i(t_i^R|\boldsymbol{\sigma})^{\delta_i^R}[S_i(t_i^L|\boldsymbol{\sigma}) - S_i(t_i^R|\boldsymbol{\sigma})]^{\delta_i^I}\zeta(\mathbf{x}_i), \tag{5.7}$$

where $\zeta(\mathbf{x}_i)$ represents the density function of $\mathbf{x}_i$. By using the joint density function defined in (5.7), the log-likelihood function in (4.4) can be re-expressed as:

$$\ell(\boldsymbol{\sigma}) = \sum_{i=1}^{n} \ell_i(\boldsymbol{\sigma}) = \sum_{i=1}^{n} \log f(z_i|\boldsymbol{\sigma}). \tag{5.8}$$

Then, by using the above log-likelihood function, the objective function given in (4.5) can be re-expressed as $\Phi(\boldsymbol{\sigma}) = \sum_{i=1}^{n} \phi_i(\boldsymbol{\sigma})$ where $\phi_i(\boldsymbol{\sigma}) = \ell_i(\boldsymbol{\sigma}) - \mu_n J(\boldsymbol{\sigma})$. The penalty function is $J(\boldsymbol{\sigma}) = J(\boldsymbol{\theta})$ as discussed in Section 3.3.2 and $\mu_n = \lambda/n$. Next, $\boldsymbol{\sigma}$ is estimated by maximizing $\Phi(\boldsymbol{\sigma})$ subject to the two non-negativity constraints: $\theta_u \geq 0$ and $\sum_{u=1}^{m} \theta_u I(t \in B_u) + \mathbf{x}_i\boldsymbol{\beta} \geq 0$ introduced by (4.7) and (4.8). During the estimating process, active constraints (i.e., $\theta_u = 0$ or $\sum_{u=1}^{\theta} {}_u I(t \in B_u) + \mathbf{x}_i\boldsymbol{\beta} = 0$) are often observed. Thus, those active constraints also should be considered when developing asymptotic results. The following assumptions are needed to develop the asymptotic properties of $\hat{\boldsymbol{\sigma}}$.

Assumptions:

B1. Observations $(T_i^L, T_i^R, \mathbf{x}_i, \delta_i; 1 \leq i \leq n)$ are independently and identically distributed and the distribution of covariate $\mathbf{x}_i$ is independent of $\boldsymbol{\sigma} = (\boldsymbol{\theta}^\top, \boldsymbol{\beta}^\top)^\top$.

B2. The censoring time is independent of the failure time given $\mathbf{x}_i$. The distribution of C_i is independent of $\boldsymbol{\sigma}$.

B3. The objective function $\Phi(\boldsymbol{\sigma})$ is bounded.

B4. Let Ω be the parameter space for vector $\boldsymbol{\sigma}$. Ω is a compact subset of R^{m+p}.

B5. Corresponding to the true parameter vector $\boldsymbol{\sigma}_0 \in \Omega$, $E_{\boldsymbol{\sigma}_0}[n^{-1}\ell(\boldsymbol{\sigma})]$ exists and a has a unique maximum at $\boldsymbol{\sigma}^* \in \Omega$ which is not necessarily equal to $\boldsymbol{\sigma}_0$ due to the penalty function $J(\boldsymbol{\sigma})$.

B6. $\Phi(\boldsymbol{\sigma})$ is continuous over Ω and is twice differentiable in a neighborhood of $\boldsymbol{\sigma}^*$. The matrices:

$$G(\boldsymbol{\sigma}^*) = -\mathrm{E}_{\boldsymbol{\sigma}_0}\left\{n^{-1}\frac{\partial^2 l(\boldsymbol{\sigma}^*)}{\partial\boldsymbol{\sigma}\partial\boldsymbol{\sigma}^\top}\right\} \quad \text{and} \quad F(\boldsymbol{\sigma}^*) = -\mathrm{E}_{\boldsymbol{\sigma}_0}\left\{n^{-1}\frac{\partial^2 \Phi(\boldsymbol{\sigma}^*)}{\partial\boldsymbol{\sigma}\partial\boldsymbol{\sigma}^\top}\right\} \tag{5.9}$$

exist and are positive definite in a neighborhood of $\boldsymbol{\sigma}^*$. Moreover, the matrices:

$$n^{-1}\frac{\partial^2 l(\boldsymbol{\sigma}^*)}{\partial\boldsymbol{\sigma}\partial\boldsymbol{\sigma}^\top} \quad \text{and} \quad n^{-1}\frac{\partial l(\boldsymbol{\sigma}^*)}{\partial\boldsymbol{\sigma}}\frac{\partial l(\boldsymbol{\sigma}^*)}{\partial\boldsymbol{\sigma}^\top} \tag{5.10}$$

converge uniformly in $\boldsymbol{\sigma}^*$ in a neighborhood of $\boldsymbol{\sigma}_0$.

B7. Penalty function $J(\boldsymbol{\sigma})$ is continuous and bounded over Ω. Both $\frac{\partial J(\boldsymbol{\sigma})}{\partial\boldsymbol{\sigma}}$ and $\frac{\partial^2 J(\boldsymbol{\sigma})}{\partial\boldsymbol{\sigma}\partial\boldsymbol{\sigma}^\top}$ exist for all $\boldsymbol{\sigma} \in \Omega$. Furthermore, $\frac{\partial^2 J(\boldsymbol{\sigma})}{\partial\boldsymbol{\sigma}\partial\boldsymbol{\sigma}^\top}$ is bounded in a neighborhood of $\boldsymbol{\sigma}^*$.

The assumptions **B1-B5** are standard assumptions, which are used in maximum likelihood theory with censored data. Assumption **B6** is useful to prove consistency and asymptotic normality of the estimates and it is similar to Assumption 4 in Yu & Ruppert (2002). In order to control the penalty term, Assumption **B7** is used during the proof of asymptotic normality of the estimates and it is generally applicable for many penalty types of interest.

Crowder (1984) and Moore et al. (2008) worked on asymptotic properties for constrained maximum likelihood estimates. Deriving asymptotic properties for estimates of the additive hazards model are aligned with the method of the latter reference. Recall that there are two non-negativity constraints on the baseline hazard and hazard function according to (4.7) and (4.8) and the dimensions of $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ are m and p respectively. The KKT conditions for both of the non-negativity constraints discussed in Section 3.4.1 also applies here. Those constraints are considered as active constraints when $\theta_u = 0$ and/or $\sum_{u=1}^{m} \theta_u I(t \in B_u) + \mathbf{x}_i \boldsymbol{\beta} = 0$. Note that, for simplicity, we do not consider any active constraints for the constraint on hazards; $\sum_{u=1}^{m} \theta_u I(t \in B_u) + \mathbf{x}_i \boldsymbol{\beta} \geq 0$ and assume that only the first q constraints of $\theta_u \geq 0$ are active. The consideration of active constraints for both of the non-negativity constraints is straightforward and the proof of asymptotic properties of that case is omitted here since it is a simple modification of the method presented here.

Define the matrix $\mathbf{U}$ which takes the following form with q active constraints;

$$\mathbf{U} = [0_{(m-q+p)\times q},\ \mathbf{I}_{(m-q+p)\times(m-q+p)}]^{\top}. \tag{5.11}$$

The $\mathbf{U}$ matrix has the property $\mathbf{U}^{\top}(\boldsymbol{\sigma})\mathbf{U}(\boldsymbol{\sigma}) = \mathbf{I}_{(m-q+p)(m-q+p)}$. Thus, asymptotic results for the constrained MPL estimates of $\boldsymbol{\sigma}$ can be presented as follows.

Theorem 5.4.1 *Suppose that Assumptions B1-B7 are satisfied. Assume that there are q active constraints in the constrained MPL estimate process for both con-*

straints. And the corresponding $\boldsymbol{U}$ matrix can be defined in a similar way as (5.11). When $n \to \infty$,

1. *the estimate $\hat{\boldsymbol{\sigma}}$, the constrained MPL estimate of $\boldsymbol{\sigma}$ is consistent for $\boldsymbol{\sigma}^*$, and*
2. $\sqrt{n}(\hat{\boldsymbol{\sigma}} - \boldsymbol{\sigma}^*) \xrightarrow{D} N(0_{(m+p)\times 1}, \tilde{\mathbf{F}}(\boldsymbol{\sigma}^*)^{-1}\mathbf{G}(\boldsymbol{\sigma}^*)[\tilde{\mathbf{F}}(\boldsymbol{\sigma}^*)^{-1}]^\top)$, *where* $\tilde{\mathbf{F}}(\boldsymbol{\sigma})^{-1} = \mathbf{U}(\mathbf{U}^\top\mathbf{F}(\boldsymbol{\sigma})\mathbf{U})^{-1}\mathbf{U}^\top$.

Let $\bar{\Phi}(\boldsymbol{\sigma}) = E_{\boldsymbol{\sigma}_0}[n^{-1}\Phi(\boldsymbol{\sigma})]$. Then, according to the uniform strong law of large numbers, it follows that $n^{-1}\Phi(\boldsymbol{\sigma}) \to \bar{\Phi}(\boldsymbol{\sigma})$ (a.s.) and uniformly for $\boldsymbol{\sigma}$ in the compact parameter space Ω. This outcome, along with $\boldsymbol{\sigma}^*$ being the unique maximum of $\bar{\Phi}(\boldsymbol{\sigma})$ according to assumption **B5**, leads to $\hat{\boldsymbol{\sigma}} \to \boldsymbol{\sigma}^*$ (a.s.) by following, as an example, Corollary 1 of Honoré & Powell (n.d.). Hence, the asymptotic normality for the distribution of $\sqrt{n}(\hat{\boldsymbol{\sigma}} - \boldsymbol{\sigma}^*)$ can be proved as follows.

From the KKT necessary conditions stated under equations (3.19) and (3.20), it is possible to show that the constrained MPL estimate $\hat{\boldsymbol{\sigma}}$ satisfies:

$$\mathbf{U}^\top \frac{\partial \Phi(\hat{\boldsymbol{\sigma}})}{\partial \boldsymbol{\sigma}} = 0.$$

From the Taylor's series expansion of $\frac{\partial \Phi(\hat{\boldsymbol{\sigma}})}{\partial \boldsymbol{\sigma}}$:

$$\frac{\partial \Phi(\hat{\boldsymbol{\sigma}})}{\partial \boldsymbol{\sigma}} = \frac{\partial \Phi(\boldsymbol{\sigma}^*)}{\partial \boldsymbol{\sigma}} + \frac{\partial^2 \Phi(\tilde{\boldsymbol{\sigma}})}{\partial \boldsymbol{\sigma} \partial \boldsymbol{\sigma}^\top}(\hat{\boldsymbol{\sigma}} - \boldsymbol{\sigma}^*), \tag{5.12}$$

where, $\tilde{\boldsymbol{\sigma}}$ is a vector between $\hat{\boldsymbol{\sigma}}$ and $\boldsymbol{\sigma}^*$. Then, multiplying equation (5.12) by $\mathbf{U}^\top$:

$$\mathbf{U}^\top \frac{\partial \Phi(\hat{\boldsymbol{\sigma}})}{\partial \boldsymbol{\sigma}} = \mathbf{U}^\top \frac{\partial \Phi(\boldsymbol{\sigma}^*)}{\partial \boldsymbol{\sigma}} + \mathbf{U}^\top \frac{\partial^2 \Phi(\tilde{\boldsymbol{\sigma}})}{\partial \boldsymbol{\sigma} \partial \boldsymbol{\sigma}^\top}(\hat{\boldsymbol{\sigma}} - \boldsymbol{\sigma}^*), \tag{5.13}$$

and,

$$0 = \mathbf{U}^\top \frac{\partial \Phi(\boldsymbol{\sigma}^*)}{\partial \boldsymbol{\sigma}} + \mathbf{U}^\top \frac{\partial^2 \Phi(\tilde{\boldsymbol{\sigma}})}{\partial \boldsymbol{\sigma} \partial \boldsymbol{\sigma}^\top}(\hat{\boldsymbol{\sigma}} - \boldsymbol{\sigma}^*). \tag{5.14}$$

Let $\tilde{\hat{\boldsymbol{\sigma}}}$ and $\tilde{\boldsymbol{\sigma}}^*$ be the new vectors for $\boldsymbol{\sigma}$ after deleting active constraints in $\hat{\boldsymbol{\sigma}}$ and $\boldsymbol{\sigma}^*$, i.e., $\mathbf{U}^\top(\hat{\boldsymbol{\sigma}} - \boldsymbol{\sigma}^*) = (\tilde{\hat{\boldsymbol{\sigma}}} - \tilde{\boldsymbol{\sigma}}^*)$. We get,

$$(\hat{\boldsymbol{\sigma}} - \boldsymbol{\sigma}^*) = \mathbf{U}(\tilde{\hat{\boldsymbol{\sigma}}} - \tilde{\boldsymbol{\sigma}}^*). \tag{5.15}$$

Applying equation (5.15) into equation (5.14) results in:

$$\begin{aligned}
\mathbf{U}^\top \frac{\partial^2 \Phi(\tilde{\boldsymbol{\sigma}})}{\partial \boldsymbol{\sigma} \partial \boldsymbol{\sigma}^\top} \mathbf{U}(\tilde{\hat{\boldsymbol{\sigma}}} - \tilde{\boldsymbol{\sigma}}^*) &= -\mathbf{U}^\top \frac{\partial \Phi(\boldsymbol{\sigma}^*)}{\partial \boldsymbol{\sigma}}, \\
(\tilde{\hat{\boldsymbol{\sigma}}} - \tilde{\boldsymbol{\sigma}}^*) &= -\left[\mathbf{U}^\top \frac{\partial^2 \Phi(\tilde{\boldsymbol{\sigma}})}{\partial \boldsymbol{\sigma} \partial \boldsymbol{\sigma}^\top} \mathbf{U}\right]^{-1} \mathbf{U}^\top \frac{\partial \Phi(\boldsymbol{\sigma}^*)}{\partial \boldsymbol{\sigma}}, \\
(\hat{\boldsymbol{\sigma}} - \boldsymbol{\sigma}^*) &= -\mathbf{U}\left[\mathbf{U}^\top \frac{\partial^2 \Phi(\tilde{\boldsymbol{\sigma}})}{\partial \boldsymbol{\sigma} \partial \boldsymbol{\sigma}^\top} \mathbf{U}\right]^{-1} \mathbf{U}^\top \frac{\partial \Phi(\boldsymbol{\sigma}^*)}{\partial \boldsymbol{\sigma}}.
\end{aligned} \tag{5.16}$$

Finally, the following equation is obtained to assess for the asymptotic results:

$$\sqrt{n}(\hat{\boldsymbol{\sigma}} - \boldsymbol{\sigma}^*) = -\mathbf{U}\left[\mathbf{U}^\top \frac{1}{n} \frac{\partial^2 \Phi(\tilde{\boldsymbol{\sigma}})}{\partial \boldsymbol{\sigma} \partial \boldsymbol{\sigma}^\top} \mathbf{U}\right]^{-1} \mathbf{U}^\top \frac{1}{\sqrt{n}} \frac{\partial \Phi(\boldsymbol{\sigma}^*)}{\partial \boldsymbol{\sigma}}. \tag{5.17}$$

Next, from (5.17) we need to show that:

$$\frac{1}{\sqrt{n}} \frac{\partial \Phi(\boldsymbol{\sigma}^*)}{\partial \boldsymbol{\sigma}} \xrightarrow{D} N(0, \mathbf{G}(\boldsymbol{\sigma}^*)) \quad \text{and} \tag{5.18}$$

$$-\mathbf{U}\left[\mathbf{U}^\top \frac{1}{n} \frac{\partial^2 \Phi(\tilde{\boldsymbol{\sigma}})}{\partial \boldsymbol{\sigma} \partial \boldsymbol{\sigma}^\top} \mathbf{U}\right]^{-1} \mathbf{U}^\top \xrightarrow{P} \tilde{\mathbf{F}}(\boldsymbol{\sigma}^*)^{-1}, \tag{5.19}$$

where $\mathbf{G}(\boldsymbol{\sigma}^*)$ and $\tilde{\mathbf{F}}(\boldsymbol{\sigma}^*)^{-1}$ are given by

$$\begin{aligned}\mathbf{G}(\boldsymbol{\sigma}^*) &= -\mathrm{E}_{\boldsymbol{\sigma}_0}\left[n^{-1}\frac{\partial^2 \ell(\boldsymbol{\sigma}^*)}{\partial\boldsymbol{\sigma}\partial\boldsymbol{\sigma}^\top}\right] \quad \text{and}\\ \tilde{\mathbf{F}}(\boldsymbol{\sigma}^*)^{-1} &= \mathbf{U}\left\{\mathbf{U}^\top\mathbf{F}(\boldsymbol{\sigma}^*)\mathbf{U}\right\}^{-1}\mathbf{U}^\top \quad \text{where} \quad \mathbf{F}(\boldsymbol{\sigma}^*) = -\mathrm{E}_{\boldsymbol{\sigma}_0}\left[n^{-1}\frac{\partial^2 \Phi(\boldsymbol{\sigma}^*)}{\partial\boldsymbol{\sigma}\partial\boldsymbol{\sigma}^\top}\right].\end{aligned}$$

The notation, $\xrightarrow{P}$ represents convergence in probability and $\xrightarrow{D}$ represents convergence in distribution.

Proof of equation (5.18):

Consider

$$\begin{aligned}\frac{1}{\sqrt{n}}\frac{\partial\Phi(\boldsymbol{\sigma}^*)}{\partial\boldsymbol{\sigma}} &= \frac{1}{\sqrt{n}}\left\{\sum_{i=1}^{n}\left[\frac{\partial\ell_i(\boldsymbol{\sigma}^*)}{\partial\boldsymbol{\sigma}} - \frac{\lambda}{n}\frac{\partial J(\boldsymbol{\sigma}^*)}{\partial\boldsymbol{\sigma}}\right]\right\}\\ &= \frac{1}{\sqrt{n}}\left\{\sum_{i=1}^{n}\left[\frac{\partial\ell_i(\boldsymbol{\sigma}^*)}{\partial\boldsymbol{\sigma}} - \mu_n\frac{\partial J(\boldsymbol{\sigma}^*)}{\partial\boldsymbol{\sigma}}\right]\right\}.\end{aligned} \tag{5.20}$$

Because of assumption **B7** and $\mu_n = o(n^{-1/2})$, the second term on the right hand side of equation (5.20),

$$\mu_n\frac{\partial J(\boldsymbol{\sigma}^*)}{\partial\boldsymbol{\sigma}} \to 0 \quad \text{as} \quad n\to\infty.$$

Therefore, equation (5.20) leads to:

$$\frac{1}{\sqrt{n}}\frac{\partial\Phi(\boldsymbol{\sigma}^*)}{\partial\boldsymbol{\sigma}} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[\frac{\partial\ell_i(\boldsymbol{\sigma}^*)}{\partial\boldsymbol{\sigma}}\right] = \sqrt{n}\left[\frac{1}{n}\frac{\partial\ell(\boldsymbol{\sigma}^*)}{\partial\boldsymbol{\sigma}}\right]. \tag{5.21}$$

By applying the central limit theorem to (5.21),

$$\sqrt{n}\left[\frac{1}{n}\frac{\partial\ell(\boldsymbol{\sigma}^*)}{\partial\boldsymbol{\sigma}}\right] \xrightarrow{D} N\{0, I_1\}, \tag{5.22}$$

where $I_1 = \mathrm{Var}_{\boldsymbol{\sigma}_0}\left[\frac{1}{n}\frac{\partial \ell(\boldsymbol{\sigma}^*)}{\partial \boldsymbol{\sigma}}\right]$ which can be simplified as,

$$\mathrm{Var}_{\boldsymbol{\sigma}_0}\left[\frac{1}{n}\frac{\partial \ell(\boldsymbol{\sigma}^*)}{\partial \boldsymbol{\sigma}}\right] = \mathrm{E}_{\boldsymbol{\sigma}_0}\left[\frac{1}{n}\frac{\partial \ell(\boldsymbol{\sigma}^*)}{\partial \boldsymbol{\sigma}}\right]\left[\frac{1}{n}\frac{\partial \ell(\boldsymbol{\sigma}^*)}{\partial \boldsymbol{\sigma}}\right]^\top - \mathrm{E}_{\boldsymbol{\sigma}_0}\left[\frac{1}{n}\frac{\partial \ell(\boldsymbol{\sigma}^*)}{\partial \boldsymbol{\sigma}}\right]\mathrm{E}_{\boldsymbol{\sigma}_0}\left[\frac{1}{n}\frac{\partial \ell(\boldsymbol{\sigma}^*)}{\partial \boldsymbol{\sigma}}\right]^\top . \tag{5.23}$$

According to assumption **B5**, $\mathrm{E}_{\boldsymbol{\sigma}_0}[n^{-1}\ell(\boldsymbol{\sigma}^*)] = 0$. Therefore, the second term on right hand side of the equation (5.23) reaches 0. Thus, equation (5.23) can be expressed in a more simple manner as:

$$\begin{aligned} I_1 = \mathrm{Var}_{\boldsymbol{\sigma}_0}\left[\frac{1}{n}\frac{\partial \ell(\boldsymbol{\sigma}^*)}{\partial \boldsymbol{\sigma}}\right] &= \mathrm{E}_{\boldsymbol{\sigma}_0}\left[\frac{1}{n}\frac{\partial \ell(\boldsymbol{\sigma}^*)}{\partial \boldsymbol{\sigma}}\right]\left[\frac{1}{n}\frac{\partial \ell(\boldsymbol{\sigma}^*)}{\partial \boldsymbol{\sigma}}\right]^\top \\ &= -\mathrm{E}_{\boldsymbol{\sigma}_0}\left[n^{-1}\frac{\partial^2 \ell(\boldsymbol{\sigma}^*)}{\partial \boldsymbol{\sigma}\partial \boldsymbol{\sigma}^\top}\right] \qquad (5.24) \\ &= \mathbf{G}(\boldsymbol{\sigma}^*). \qquad (5.25) \end{aligned}$$

Therefore, according to the result in equation (5.24), the distribution under (5.21) can be re-expressed as,

$$\sqrt{n}\left[\frac{1}{n}\frac{\partial \ell(\boldsymbol{\sigma}^*)}{\partial \boldsymbol{\sigma}}\right] \xrightarrow{D} N\{0, \mathbf{G}(\boldsymbol{\sigma}^*)\}, \tag{5.26}$$

and therefore satisfies the requirement under equation (5.18) as well.

Proof of equation (5.19):

Here, it is required to show that

$$-\mathbf{U}\left[\mathbf{U}^\top \frac{1}{n}\frac{\partial^2 \Phi(\tilde{\boldsymbol{\sigma}})}{\partial \boldsymbol{\sigma}\partial \boldsymbol{\sigma}^\top}\mathbf{U}\right]^{-1}\mathbf{U}^\top \xrightarrow{P} \tilde{\mathbf{F}}(\boldsymbol{\sigma}^*)^{-1}.$$

Since $\tilde{\boldsymbol{\sigma}}$ is a vector between $\hat{\boldsymbol{\sigma}}$ and $\boldsymbol{\sigma}^*$:

$$\hat{\boldsymbol{\sigma}} \to \boldsymbol{\sigma}^* \quad \text{and} \quad \tilde{\boldsymbol{\sigma}} \to \boldsymbol{\sigma}^* \quad \text{as} \quad n \to \infty.$$

Hence, using the consistent property of the constrained MPL estimate $\hat{\boldsymbol{\sigma}}$ of $\boldsymbol{\sigma}^*$, and Assumption 4 in Amemiya (1983), the term on the left side of the expression (5.18) converges in probability as follows:

$$-\mathbf{U}\left[\mathbf{U}^\top \frac{1}{n}\frac{\partial^2\Phi(\tilde{\boldsymbol{\sigma}})}{\partial\boldsymbol{\sigma}\partial\boldsymbol{\sigma}^\top}\mathbf{U}\right]^{-1}\mathbf{U}^\top \xrightarrow{P} -\mathbf{U}\left\{\mathbf{U}^\top \mathrm{E}_{\boldsymbol{\sigma}_0}\left[\frac{1}{n}\frac{\partial^2\Phi(\boldsymbol{\sigma}^*)}{\partial\boldsymbol{\sigma}\partial\boldsymbol{\sigma}^\top}\right]\mathbf{U}\right\}^{-1}\mathbf{U}^\top = \tilde{\mathbf{F}}(\boldsymbol{\sigma}^*)^{-1}. \tag{5.27}$$

Thus, the result in equation (5.27) satisfies the requirement under equation (5.19).

With that result, as per Theorem 5.4.1, the expression for asymptotic variance is given by:

$$V(\boldsymbol{\sigma}^*) = \tilde{\mathbf{F}}(\boldsymbol{\sigma}^*)^{-1}\mathbf{G}(\boldsymbol{\sigma}^*)[\tilde{\mathbf{F}}(\boldsymbol{\sigma}^*)^{-1}]^\top \quad \text{where} \quad \tilde{\mathbf{F}}(\boldsymbol{\sigma})^{-1} = \mathbf{U}(\mathbf{U}^\top\mathbf{F}(\boldsymbol{\sigma})\mathbf{U})^{-1}\mathbf{U}^\top. \tag{5.28}$$

The results in Theorem 5.4.1 are quite useful in a practical aspect, since this method accommodates active constraints and non-zero smoothing values. In practice, the unique maximum value $\boldsymbol{\sigma}^*$ is often unavailable. But, the consistency results allow us to replace it by the constrained MPL estimate $\hat{\boldsymbol{\sigma}}$. With these results, inference can be made with respect to regression coefficients or the baseline hazard. Note that the penalty term in the penalized log-likelihood will disappear in proving the asymptotic properties of the constrained MPL estimates, if $\mu_n \to 0$ as $n \to \infty$. Then, the analysis will follow the same approach as in proving asymptotic properties of the maximum likelihood estimator (MLE).

Next, Corollary 5.4.1.1 examines the asymptotic properties for fixed m and when $\mu_n = o(n^{-1/2})$.

Corollary 5.4.1.1 *Suppose the assumptions **B1-B7** hold and there are q active constraints in the constrained optimization and matrix* $\mathbf{U}$ *is defined similar to (5.11). Let* $\boldsymbol{\sigma}_0$ *be the true parameter value associated with fixed m. When* $\mu_n = o(n^{-1/2})$*, the MPL estimate* $\hat{\boldsymbol{\sigma}}$ *is strongly consistent for* $\boldsymbol{\sigma}_0$ *and the distribution of* $\sqrt{n}(\hat{\boldsymbol{\sigma}} - \boldsymbol{\sigma}_0)$ *converges to the normal distribution* $N(0_{(m+p)\times 1}, \mathbf{U}(\mathbf{U}^\top \mathbf{G}(\boldsymbol{\sigma}_0)\mathbf{U})^{-1}\mathbf{U}^\top)$ *as* $n \to \infty$.

The variance formula given in (5.28) for the asymptotic normal distribution is called the sandwich formula, and a consistent estimate can be obtained by replacing $\tilde{\mathbf{F}}(\boldsymbol{\sigma}^*)$ and $\mathbf{G}(\boldsymbol{\sigma}^*)$ by their empirical versions, with $\boldsymbol{\sigma}^*$ replaced by $\hat{\boldsymbol{\sigma}}$. Studying the asymptotic variance of the MLE estimates assume that $\boldsymbol{\sigma} = (\boldsymbol{\theta}^\top, \boldsymbol{\beta}^\top)^\top$ is an interior point of the parameter space Ω, e.g., $\theta_u > 0$ for all u. However, generally it is possible that elements take zero values resulting in active constraints. Then, it is difficult to develop the asymptotic theory for $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\beta}}$ when some $\boldsymbol{\theta}$ elements lie on the boundary. In those situations, it is possible to perform bootstrapping which can provide an approximate variance. This approach requires a large amount of repeated estimations of both of the parameters and it can be computationally intensive. But, in our study, we considered the case of active constraints as well. Thus, it is possible to obtain asymptotic variance for the estimates using the results obtained in Section 5.4. The simulation studies in Chapter 6 indicate that the sandwich formula given in (5.28) for the asymptotic variance is generally accurate for both right-censored and interval-censored survival data.

Chapter 6

Simulation studies

6.1 Introduction

In this chapter, simulation studies are considered to evaluate the MPL method in fitting the additive hazards model. The main objectives of these simulation studies are: (i) to study the effects of sample size and censoring proportion on the constrained MPL estimators of regression coefficient $\boldsymbol{\beta}$ and baseline hazard $h_0(t)$, (ii) to compare the asymptotic standard deviations with the Monte Carlo standard deviations of the constrained MPL estimators, and (iii) to compare our proposed constrained MPL method with the existing parameter estimation methods developed by Aalen (1980) and Lin & Ying (1994) which has been reviewed in Sections 1.4.2 and 2.3.1. These simulation studies were performed to check and assess the performance of the proposed MPL method with right-censored survival data and partly interval-censored survival data separately.

The first objective aims to analyse the sensitivity of the MPL estimators of $\boldsymbol{\beta}$ and $h_0(t)$ with different censoring proportions π_c and the sample sizes n. We considered three sample sizes covering small, intermediate and large samples. For each sample size, different censoring proportions covering small, intermediate and large proportion of censoring were considered to achieve this objective. For the second objective of this study, we investigate whether the asymptotic standard deviations computed by the sandwich formula given by (5.28) are accurate for those of the MPL estimators. Asymptotic standard deviations are calculated by replacing $\boldsymbol{\sigma}^*$ with $\hat{\boldsymbol{\sigma}}$ in equation (5.28) due to the consistency result and those standard deviations were compared against the Monte Carlo standard deviations to assess their accuracy. For the third objective, we demonstrate improvements of our MPL method in estimating $\boldsymbol{\beta}$ and $h_0(t)$ compared with the existing parameter estimation methods developed by Aalen (1980) and Lin & Ying (1994). We used R statistical computing software with relevant R packages to do the required computations and assess these three objectives. For the comparison of the third objective, R packages, '*Survival*'(Therneau & Grambsch 2000) and '*ahaz* '(Gorst-Rasmussen & Scheike 2012) which are available in CRAN were used as they implemented parameter estimation methods developed by Aalen and Lin & Ying for additive hazards models respectively. Note that the methods developed by Aalen and Lin

& Ying are referred as "Aalen method" and "L-Y method" respectively in the subsequent sections of this chapter.

This chapter is organised as follows. Section 6.2 presents the notations and equations used to generate results based on simulations studies. Section 6.3 presents the simulation results for right-censored survival data, and Section 6.4 presents the simulation results for partly interval-censored data.

6.2 Notations

For each combination of n and π_c, we perform Monte Carlo simulations for the MPL method using K = 1,000 repeated samples, and thus obtain 1,000 MPL estimates for $\boldsymbol{\beta}$ and $h_0(t)$. From those 1,000 estimates of $\boldsymbol{\beta}$, we can compute the average estimate (AEST), the estimated bias (BIAS), the Monte Carlo standard deviation (MCSD), the average asymptotic standard deviation (AASD) and the mean squared error (MSE) of $\hat{\boldsymbol{\beta}}$. Specifically, let $\hat{\boldsymbol{\beta}}_k$ be the estimate of $\boldsymbol{\beta}$ from the k^{th} sample, $k = 1, \ldots, K$. Then we have,

$$\mathrm{AEST}(\hat{\boldsymbol{\beta}}) = \frac{\sum_{k=1}^{K} \hat{\boldsymbol{\beta}}_k}{K}, \tag{6.1}$$

$$\mathrm{BIAS}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta} - \mathrm{AEST}(\hat{\boldsymbol{\beta}}), \tag{6.2}$$

$$\mathrm{MCSD}(\hat{\boldsymbol{\beta}}) = \sqrt{\frac{1}{K-1} \sum_{k=1}^{K} \left\{ \hat{\boldsymbol{\beta}}_k - \mathrm{AEST}(\hat{\boldsymbol{\beta}}) \right\}^2}, \tag{6.3}$$

$$\begin{aligned} \text{MSE}(\hat{\boldsymbol{\beta}}) &= \left[\text{BIAS}(\hat{\boldsymbol{\beta}})\right]^2 + \left[\text{MCSD}(\hat{\boldsymbol{\beta}})\right]^2 \quad \text{and} & (6.4)\\ \text{AASD}(\hat{\boldsymbol{\beta}}) &= \frac{\sum_{k=1}^{K} \text{ASD}(\hat{\boldsymbol{\beta}}_k)}{K}, & (6.5) \end{aligned}$$

where $\text{ASD}(\hat{\boldsymbol{\beta}}_k)$ is the asymptotic standard deviation of the estimator $\hat{\boldsymbol{\beta}}_k$ obtained by the sandwich formula given in equation (5.28). The coverage probabilities of the 95% confidence intervals of the $\boldsymbol{\beta}$ estimates are also calculated. Let $(\text{LL}_k, \text{UL}_k)$ denote the confidence interval of β_j, $j = 1, 2, \ldots, p$ for the k^{th} sample. Then, the coverage probability, C_p can be calculated as,

$$C_p = \frac{\sum_{k=1}^{K} I\{\text{LL}_k < \beta_j \quad \text{and} \quad \text{UL}_k > \beta_j\}}{K}. \tag{6.6}$$

Let $\hat{h}_0(t)_k$ be the MPL estimate of the true baseline hazard $h_0(t)$ at time t for the k^{th} sample, $k = 1, \ldots, N$. From the $N = 1,000$ estimates of $h_0(t)$, we calculate the AEST, MCSD and AASD of $\hat{h}_0(t)$ by,

$$\begin{aligned} \text{AEST}(\hat{h}_0(t)) &= \frac{\sum_{k=1}^{K} \hat{h}_0(t)_k}{K}, & (6.7)\\ \text{MCSD}(\hat{h}_0(t)) &= \sqrt{\frac{1}{K-1}\sum_{k=1}^{K}\left\{\hat{h}_0(t)_k - \text{AEST}(\hat{h}_0(t))\right\}^2} \quad \text{and} & (6.8)\\ \text{AASD}(\hat{h}_0(t)) &= \frac{\sum_{k=1}^{K} \text{ASD}(\hat{h}_0(t)_k)}{K}, & (6.9) \end{aligned}$$

where $\text{ASD}(\hat{h}_0(t)_k)$ is the asymptotic standard deviation of the estimator $\hat{h}_0(t)_k$ at time t for the k^{th} sample and obtained from the formula (5.28). For each sample k, we calculate the distances between the true baseline hazard and the MPL estimate of it. The distance used is the integrated squared error (ISE) and it is given by,

$$\text{ISE}_k = \int_0^\infty [h_0(t) - \hat{h}_0(t)_k]^2 dt. \tag{6.10}$$

Hence, the average integrated squared error (AISE) of the 1,000 baseline hazard estimates is

$$\text{AISE} = \frac{\sum_{k=1}^{K} \text{ISE}_k}{K}. \tag{6.11}$$

Variance of the ISE's does not calculate at this stage, but from Figures 1-3 show the distribution of ISE's for different sample sizes and censoring proportions clearly.

6.3 Simulation studies for right-censored survival data

The MPL method was developed for parameter estimation with partly interval-censored data covering all the censoring types. But, firstly we used this MPL method for the simulation studies with only right censoring due to the following reasons: (i) we need to check the performance of the MPL method with the most common censoring type, and (ii) as per the third objective we wish to compare the results of this MPL method with other exisiting AH model parameter estimation approaches, which are not developed to handle interval-censored data. Here, we used survival data simulated from the Weibull distribution as described below. A random sample of data, $\{(t_i, \delta_i), \mathbf{x}_i; i = 1, 2, \ldots, n\}$ was generated by following the below steps:

1. The failure time, T_i for each individual $i = 1, 2, \dots, n$ is generated from the Weibull distribution with the hazard function,

$$h(t_i) = 3t_i^2 + (x_{i1} + 0.8x_{i2} - 0.5x_{i3}), \tag{6.12}$$

   where the baseline hazard function of this additive hazards model is given by $h_0(t_i) = 3t_i^2$. Here, we set the regression coefficients, $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3)^\top$ to be $\beta_1 = 1$, $\beta_2 = 0.8$ and $\beta_3 = -0.5$. Then, the covariate vector $\mathbf{x}_i = (x_{i1}, x_{i2}, x_{i3})^\top$ is generated from the distributions: $x_{i1} \sim \text{Uniform}(-1, 1)$, $x_{i2} \sim \text{Bernoulli}(0.5)$ and $x_{i3} \sim \text{Uniform}(0, 3)$. An inversion method is used to generate T_i based on the relationship $u_i = F_T(t_i)$, where u_i is a standard uniform random variable and $F_T(t_i)$ is the cumulative distribution function of T_i, which is given by,

$$F_T(t_i) = 1 - \exp\Big[- t_i^3 - (x_{i1} + 0.8x_{i2} - 0.5x_{i3})t_i \Big],$$

   corresponding to the hazard model (6.12).

2. Corresponding to each T_i generated from the above step, we generated a censoring time C_i from the exponential distribution $\text{Exp}(\mu_c)$, where μ_c denotes the mean of this exponential distribution. Then, the observed data can be obtained as $\Big\{(t_i, \delta_i); i = 1, \dots, n\Big\}$, where $t_i = \min\Big\{T_i, C_i\Big\}$ and $\delta_i = I(T_i < C_i)$, where I is the indicator function. By adjusting the value of

μ_c, it is possible to control the censoring proportion π_c. Here, the values of C_i in this step were generated independently of t_i.

In order to get an overall idea on how each estimation method is affected by sample size n and censoring proportion π_c, we used samples of sizes $n = 100, 500$, and $1,000$ covering small, intermediate and large sample sizes. For each sample size, approximate censoring proportions of 20%, and 80% were considered. Model parameters $\boldsymbol{\beta}$ and $h_0(t)$ were estimated using the augmented Lagrangian method discussed in Chapter 3. In discretising the $h_0(t)$, the bins are selected in such a way that the number of observations in each bin, n_c, is approximately the same. Some preliminary tests indicates that the MPL estimator of regression coefficient $\boldsymbol{\beta}$ is not very sensitive to the choice of n_c, as long as n_c is not too large and the smoothing value parameter λ is appropriate. We select $n_c = 2$ for $n = 100$, $n_c = 5$ for $n = 500$ and $n_c = 8$ for $n = 1,000$.

The convergence criteria of these simulation studies are the same as the convergence criteria discussed in Chapter 3. Here we obtained the MPL estimates when augmented Lagrangian method is converged or the maximum of 5,000 iterations is reached, whichever occurs first. Here, we set the smoothing parameter to $\lambda = \xi/(1 - \xi)$, where the tuning parameter ξ satisfies $\xi \in [0, 1)$. In the simulation, we determined ξ experimentally so that good estimates of $\boldsymbol{\beta}$ and $h_0(t)$ were obtained. Small samples generally require lighter smoothing while large samples prefer heavier smoothing. This is because there were less bins for small samples and more bins for large samples. From the 1,000 repeated samples, the performance of these estimates was assessed by examining their biases, standard deviations and mean squared errors introduced in Section 6.2.

Tables 1-3 summarise the AEST, BIAS, MCSD, AASD and MSE values for the MPL estimates of $\boldsymbol{\beta}$ with different censoring proportions and sample sizes. We observe that:

i with a fixed sample size n, the MSE increases with censoring proportion, and the MCSD, AASD and absolute value of BIAS follow the same trend,

ii with a fixed censoring proportion, all of these four quantities are decreasing as sample size increases, and

iii comparison between MCSD and AASD demonstrates that the sandwich formula given in (5.28) is generally accurate in approximating the variance of the MPL estimates of $\boldsymbol{\beta}$, particularly when the sample size becomes larger or censoring proportion becomes smaller.

Tables 1-3 also give AISEs for the MPL estimates of $h_0(t)$. It is observed that the AISEs exhibit an increasing trend as the censoring proportion increases, but a decreasing trend as the sample size increases.

Along with these above mentioned simulation studies, we performed a comparison between our MPL method and the two existing methods, the Aalen and

L-Y methods. These two comparisons can be easily implemented for right-censored data using '*ahaz*' and '*aareg*' R functions for the Aalen and L-Y methods respectively. The two methods are evaluated based on the same data sets as the MPL method. The simulated data for both methods have sample sizes of $n = 100, 500$ and $1,000$. Monte Carlo simulations are done with $1,000$ repeated samples. The covariates $\mathbf{x}_i = (x_{i1}, x_{i2}, x_{i3})^\top$ and regression coefficients $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3)^\top$ are the same as the MPL method. Here, we estimated regression coefficients $\boldsymbol{\beta}$ for both of the comparison methods. Note that the baseline hazard estimates were only estimated for the L-Y method, since the Aalen method implemented in '*aareg*' R function in '*Survival*' R package does not provide estimates for $h_0(t)$. Furthermore, the inference methods were not implemented for the Aalen method in '*aareg*' R function. Therefore, the asymptotic standard deviations for $\boldsymbol{\beta}$ were not available for Aalen method. Hence, we only estimated Monte Carlo standard deviations.

Tables 1-3 report the estimation results for $\boldsymbol{\beta}$ from the two comparison methods. We observe that, under each of the three sample sizes and two censoring proportions, the biases of estimates and standard deviations using the MPL method are smaller than those in the Aalen and L-Y methods. Thus, MPL method achieves lower MSE than Aalen and L-Y methods. Tables 1-3 also report values of AISE for the baseline hazard estimate for L-Y method along with that for MPL method. We observe that our MPL method gives much smaller AISE than L-Y method. It can be seen that the values of AISE decrease with sample size but increase with censoring proportion. When comparing the biases and standard deviations of the Aalen and L-Y methods, under each sample size and censoring proportion combination, the biases of estimates in the L-Y method are smaller than those in Aalen method, although they yield roughly similar standard deviations. This leads to, in general, lower MSE values for the L-Y method compared to the Aalen method.

Both the MPL method and L-Y method provide reasonable coverage probabilities of confidence intervals for $\boldsymbol{\beta}$ for all the sample sizes and censoring proportions. When the sample sizes are $n = 100$ and $n = 500$, the regression estimate of β_2 shows a slightly lower coverage probability for both of the methods compared to coverage probabilities of other instances which might be due to small sample sizes. As expected, the coverage probabilities tend to increase when the sample sizes increase and/or the censoring percentage decreases. In general, the coverage probabilities of the MPL confidence intervals for $\boldsymbol{\beta}$ tend to reach the 95% nominal value in all the simulations except for β_2 for sample sizes $n = 100$ and $n = 500$.

Sample size		$n = 100$					
Censoring proportion		$c = 20\%$			$c = 80\%$		
Method		MPL	L-Y	Aalen	MPL	L-Y	Aalen
$\beta_1 = 1$	AEST	0.9821	0.9740	0.9572	0.9719	0.9724	0.9455
	BIAS	0.0179	0.0260	0.0428	0.0281	0.0276	0.0545
	MCSD	0.0724	0.0854	0.0782	0.0808	0.0932	0.0869
	MSE	0.0056	0.0080	0.0079	0.0073	0.0094	0.0105
	AASD	0.0696	0.0879		0.0816	0.0945	
	C_p	0.9492	0.9466		0.9456	0.9402	
$\beta_2 = 0.8$	AEST	0.8219	0.8118	0.8139	0.8157	0.8172	0.7438
	BIAS	-0.0219	-0.0118	-0.0139	-0.0157	-0.0172	0.0562
	MCSD	0.0550	0.0705	0.0654	0.0637	0.0734	0.0813
	MSE	0.0035	0.0051	0.0045	0.0043	0.0057	0.0098
	AASD	0.0574	0.0679		0.0644	0.0748	
	C_p	0.9392	0.9364		0.9322	0.9262	
$\beta_3 = -0.5$	AEST	-0.5704	-0.5872	-0.5962	-0.5769	-0.5784	-0.3975
	BIAS	0.0704	0.0872	0.0962	0.0769	0.0784	-0.1025
	MCSD	0.0851	0.0937	0.0881	0.0908	0.0970	0.0984
	MSE	0.0122	0.0164	0.0170	0.0142	0.0156	0.0202
	AASD	0.0884	0.0985		0.0916	0.0941	
	C_p	0.9622	0.9542		0.9592	0.9510	
$h_0(t)$	AISE	0.2873	1.3424		0.2985	1.4768	

Table 1: Comparisons of estimates of $\boldsymbol{\beta}$ and $h_0(t)$ between the MPL, L-Y and Aalen methods for right-censored data with sample size $n = 100$

Sample size		$n = 500$					
Censoring proportion		$c = 20\%$			$c = 80\%$		
Method		MPL	L-Y	Aalen	MPL	L-Y	Aalen
$\beta_1 = 1$	AEST	0.9819	0.9851	0.9615	0.9797	0.9793	0.9647
	BIAS	0.0181	0.0149	0.0385	0.0203	0.0207	0.0353
	MCSD	0.0440	0.0474	0.0467	0.0449	0.0499	0.0571
	MSE	0.0023	0.0025	0.0037	0.0024	0.0029	0.0045
	AASD	0.0442	0.0450		0.0435	0.0513	
	C_p	0.9582	0.9560		0.9470	0.9410	
$\beta_2 = 0.8$	AEST	0.8142	0.8204	0.8166	0.8199	0.8226	0.8204
	BIAS	-0.0142	-0.0204	-0.0166	-0.0199	-0.0226	-0.0204
	MCSD	0.0405	0.0411	0.0378	0.0437	0.0457	0.0431
	MSE	0.0018	0.0021	0.0017	0.0023	0.0026	0.0023
	AASD	0.0408	0.0420		0.0441	0.0463	
	C_p	0.9402	0.9350		0.9412	0.9362	
$\beta_3 = -0.5$	AEST	-0.5295	-0.5328	-0.5525	-0.5461	-0.5469	-0.5754
	BIAS	0.0295	0.0328	0.0525	0.0461	0.0469	0.0754
	MCSD	0.0456	0.0508	0.0540	0.0481	0.0532	0.0600
	MSE	0.0029	0.0037	0.0057	0.0044	0.0050	0.0093
	AASD	0.0474	0.0519		0.0461	0.0512	
	C_p	0.9656	0.9580		0.9628	0.9538	
$h_0(t)$	AISE	0.0985	0.6587		0.0998	0.6892	

Table 2: Comparisons of estimates of $\boldsymbol{\beta}$ and $h_0(t)$ between the MPL, L-Y and Aalen methods for right-censored data with sample size $n = 500$

Sample size		$n = 1,000$					
Censoring proportion		$c = 20\%$			$c = 80\%$		
Method		MPL	L-Y	Aalen	MPL	L-Y	Aalen
$\beta_1 = 1$	AEST	0.9953	0.9883	0.9725	0.9863	0.9789	0.9664
	BIAS	0.0047	0.0117	0.0275	0.0137	0.0211	0.0336
	MCSD	0.0180	0.0206	0.0226	0.0184	0.0209	0.0266
	MSE	0.0003	0.0006	0.0013	0.0005	0.0009	0.0018
	AASD	0.0171	0.0220		0.0172	0.0213	
	C_p	0.9660	0.9588		0.9490	0.9438	
$\beta_2 = 0.8$	AEST	0.8083	0.8044	0.8112	0.8076	0.8047	0.8114
	BIAS	-0.0083	-0.0044	-0.0112	-0.0076	-0.0047	-0.0114
	MCSD	0.0154	0.0196	0.0169	0.0172	0.0199	0.0207
	MSE	0.0003	0.0004	0.0004	0.0004	0.0004	0.0006
	AASD	0.0146	0.0192		0.0171	0.0190	
	C_p	0.9532	0.9410		0.9480	0.9394	
$\beta_3 = -0.5$	AEST	-0.5220	-0.5287	-0.5894	-0.5335	-0.5507	-0.5893
	BIAS	0.0220	0.0287	0.0894	0.0335	0.0507	0.0893
	MCSD	0.0188	0.0210	0.0251	0.0194	0.0220	0.0275
	MSE	0.0008	0.0013	0.0086	0.0015	0.0031	0.0087
	AASD	0.0190	0.0219		0.0190	0.0233	
	C_p	0.9720	0.9598		0.9680	0.9542	
$h_0(t)$	AISE	0.0398	0.2387		0.0427	0.2674	

Table 3: Comparisons of estimates of $\boldsymbol{\beta}$ and $h_0(t)$ between the MPL, L-Y and Aalen methods for right-censored data with sample size $n = 1,000$

6.4 Simulation studies for partly interval-censored survival data

In this section, we focus on simulation studies based on partly interval-censored survival data. Since our MPL approach can accommodate any type of censoring, for these simulation studies we include right, left and finite interval-censored survival data along with fully observed event data. Here, we performed these simulation studies to check the first two objectives mentioned in Section 6.1. We do not compare the MPL results with results from the L-Y and the Aalen methods for partly interval-censored data due to: (i) the '*ahaz*' R function which implements the L-Y method supports only right-censored data, and (ii) acco1rding to the simulation results from Section 6.3, it has been shown that the MPL method outperformed the Aalen method even for right-censored data.

Simulation for the partly interval-censored data closely follows the method used for right-censored data in Section 6.3. A random sample of data, $\{(L_i, R_i], \mathbf{x}_i; i = 1, 2, \ldots, n\}$ was generated by following the below steps:

i The failure time, T_i, is generated from the Weibull distribution by following the method discussed under step 1 in the right-censored survival data generation process. We used the same hazard function from the Weibull distribution given in equation (6.12). The true regression coefficients take the same values and the covariate vector follows the same distributions we set in right-censored data. The inversion method is used to generate T_i for partly interval-censored data as well.

ii This step explains the process of generating a bivariate random vector for partly interval-censored data with corresponding censoring type. We followed the method used by Cai & Betensky (2003) to generate data for all censoring types. First, we generate two monitoring times independent of T_i for each subject i. Let $C_{i1} \sim \text{Uniform}(0, 1)$ be the left censoring time random variable and $C_{i2} = C_{i1} + \text{Uniform}(0, 1)$ be the right censoring time random variable. Then, a standard uniform random variable, w_i was generated. If $w_i \geq \pi_c$, then the failure time T_i is deemed to be exactly observed and we set $L_i = R_i = T_i$. If $w_i < \pi_c$, the failure time is considered as censored and there are three possible censoring scenarios: if $T_i \leq C_{i1}$, it is left-censored and we set $L_i = 0$ and $R_i = C_{i1}$; if $C_{i1} < T_i \leq C_{i2}$, it is considered as finite interval-censored and $L_i = C_{i1}$ and $R_i = C_{i2}$, and if $T_i > C_{i2}$, it is right-censored and $L_i = C_{i2}$ and $R_i = +\infty$. By using the different values of π_c, it is possible to get the data with different censoring proportions.

Three sample sizes, $n = 100, 500$ and $n = 1,000$ were considered for the simulation studies and for each sample size three approximate censoring proportions, $20\%, 50\%$ and 80% were also considered. Model parameters $\boldsymbol{\beta}$ and $h_0(t)$ were estimated using the augmented Lagrangian method discussed in Chapter 4. All the conditions, namely binning criteria, the number of observations per bin for each sample size, convergence criteria and smoothing parameter are remain the same as the case of right-censored survival data. A simulation study was performed using $1,000$ repeated samples and the performance of the augmented Lagrangian method was assessed by examining the biases, standard deviations and mean squared errors discussed in Section 6.2.

Tables 4-6 summarise the regression estimates, bias, standard deviations and mean square error values for $\hat{\boldsymbol{\beta}}$ with different censoring proportions and sample sizes. Similar to the observations we had for right-censored data, we observe that:

i the BIAS and MCSD increase with censoring proportion with a fixed sample size n, thus the MSE follows the same trend,

ii the four quantities BIAS, MCSD, AASD and MSE are decreasing as n increases with fixed π_c, and

iii the two values MCSD and AASD are approximately the same and it implies that the sandwich formula given in (5.28) is generally accurate in approximating the variance of the MPL estimates of $\boldsymbol{\beta}$.

Furthermore, the AISEs for the MPL estimates of $h_0(t)$ increase with the censoring proportion, but decrease with the sample size.

Sample size		$n = 100$		
Censoring proportion		20%	50%	80%
$\beta_1 = 1$	AEST	0.9987	1.0046	0.9765
	BIAS	0.0013	-0.0046	0.0234
	MCSD	0.1819	0.1970	0.2063
	MSE	0.0335	0.0409	0.0431
	AASD	0.1739	0.1909	0.1990
	C_p	0.9568	0.9454	0.9422
$\beta_2 = 0.8$	AEST	0.8048	0.8010	0.8081
	BIAS	-0.0048	-0.0010	-0.0081
	MCSD	0.1531	0.1780	0.1960
	MSE	0.1585	0.1639	0.1885
	AASD	0.0537	0.0508	0.0539
	C_p	0.9542	0.9434	0.9381
$\beta_3 = -0.5$	AEST	-0.5241	-0.5417	-0.5387
	BIAS	0.0241	0.0417	0.0387
	MCSD	0.1881	0.2024	0.2162
	MSE	0.0360	0.0427	0.0482
	AASD	0.1833	0.1997	0.2234
	C_p	0.9509	0.9526	0.9468
$h_0(t)$	AISE	0.2873	0.3248	0.3585

Table 4: Comparison of AEST, BIAS, MCSD, AASD, MSE and C_p for $\boldsymbol{\beta}$ estimates, and AISE for $h_0(t)$ estimates for partly interval-censored data with sample size $n = 100$

Sample size		$n = 500$		
Censoring proportion		20%	50%	80%
$\beta_1 = 1$	AEST	0.9998	1.0092	0.9848
	BIAS	0.0002	-0.0092	0.0152
	MCSD	0.1403	0.1606	0.1939
	MSE	0.0197	0.0259	0.0378
	AASD	0.1339	0.1536	0.1808
	C_p	0.9552	0.9523	0.9446
$\beta_2 = 0.8$	AEST	0.8044	0.7980	0.8018
	BIAS	-0.0044	0.0020	-0.0018
	MCSD	0.1526	0.1730	0.1858
	MSE	0.0233	0.0299	0.0345
	AASD	0.1437	0.1604	0.1897
	C_p	0.9580	0.9508	0.9485
$\beta_3 = -0.5$	AEST	-0.5238	-0.5385	-0.5573
	BIAS	0.0238	0.0385	0.0573
	MCSD	0.1881	0.1811	0.2005
	MSE	0.0359	0.0343	0.0435
	AASD	0.1833	0.1894	0.1949
	C_p	0.9551	0.9588	0.9464
$h_0(t)$	AISE	0.1273	0.1424	0.1585

Table 5: Comparison of AEST, BIAS, MCSD, AASD, MSE and C_p for $\boldsymbol{\beta}$ estimates, and AISE for $h_0(t)$ estimates for partly interval-censored data with sample size $n = 500$

Sample size		$n = 1,000$		
Censoring proportion		20%	50%	80%
$\beta_1 = 1$	AEST	0.9995	0.9983	0.9771
	BIAS	0.0005	0.0017	0.0229
	MCSD	0.0318	0.0825	0.1267
	MSE	0.0010	0.0068	0.0166
	AASD	0.0326	0.0816	0.1210
	C_p	0.9562	0.9527	0.9503
$\beta_2 = 0.8$	AEST	0.8006	0.8020	0.8074
	BIAS	-0.0006	-0.0020	-0.0074
	MCSD	0.0598	0.0915	0.1059
	MSE	0.0036	0.0084	0.0113
	AASD	0.0534	0.0884	0.0939
	C_p	0.9603	0.9609	0.9545
$\beta_3 = -0.5$	AEST	-0.5344	-0.5339	-0.5382
	BIAS	0.0344	0.0339	0.0382
	MCSD	0.0759	0.0864	0.1163
	MSE	0.0069	0.0086	0.0158
	AASD	0.0708	0.0808	0.1084
	C_p	0.9602	0.9567	0.9561
$h_0(t)$	AISE	0.0473	0.0594	0.0685

Table 6: Comparison of AEST, BIAS, MCSD, AASD, MSE and C_p for $\boldsymbol{\beta}$ estimates, and AISE for $h_0(t)$ estimates for partly interval-censored data with sample size $n = 1,000$

Figures 1-3 exhibit plots for the true baseline hazard, AEST of the baseline hazard estimates, the corresponding 95% Monte Carlo piecewise confidence intervals (PWCI) and the corresponding 95% asymptotic piecewise confidence intervals (PWCI). We detect that:

i AESTs are all very close to the true baseline hazard under different sample sizes and censoring proportions,

ii the 95% Monte Carlo PWCI is close to the 95% asymptotic PWCI, hence the sandwich formula given in (5.28) gives a good variance approximation for the MPL estimates, $\hat{h}_0(t)$,

iii both the 95% Monte Carlo PWCI and the 95% asymptotic PWCI become wider as the censoring proportion increases, but narrower when the sample size increases, and

iv even with the indicator basis function, the baseline hazard is approximated well and more accurate approximations can be obtained by using powerful basis functions such as M-spline.

In this chapter, we have tested a MPL method to fit the additive hazards model with right-censored and partly interval-censored data, where a quadratic penalty term is added to the log-likelihood function to assure smoothness of the estimated baseline hazard. To obtain MPL estimators of $\boldsymbol{\beta}$ and $h_0(t)$, we used the augmented

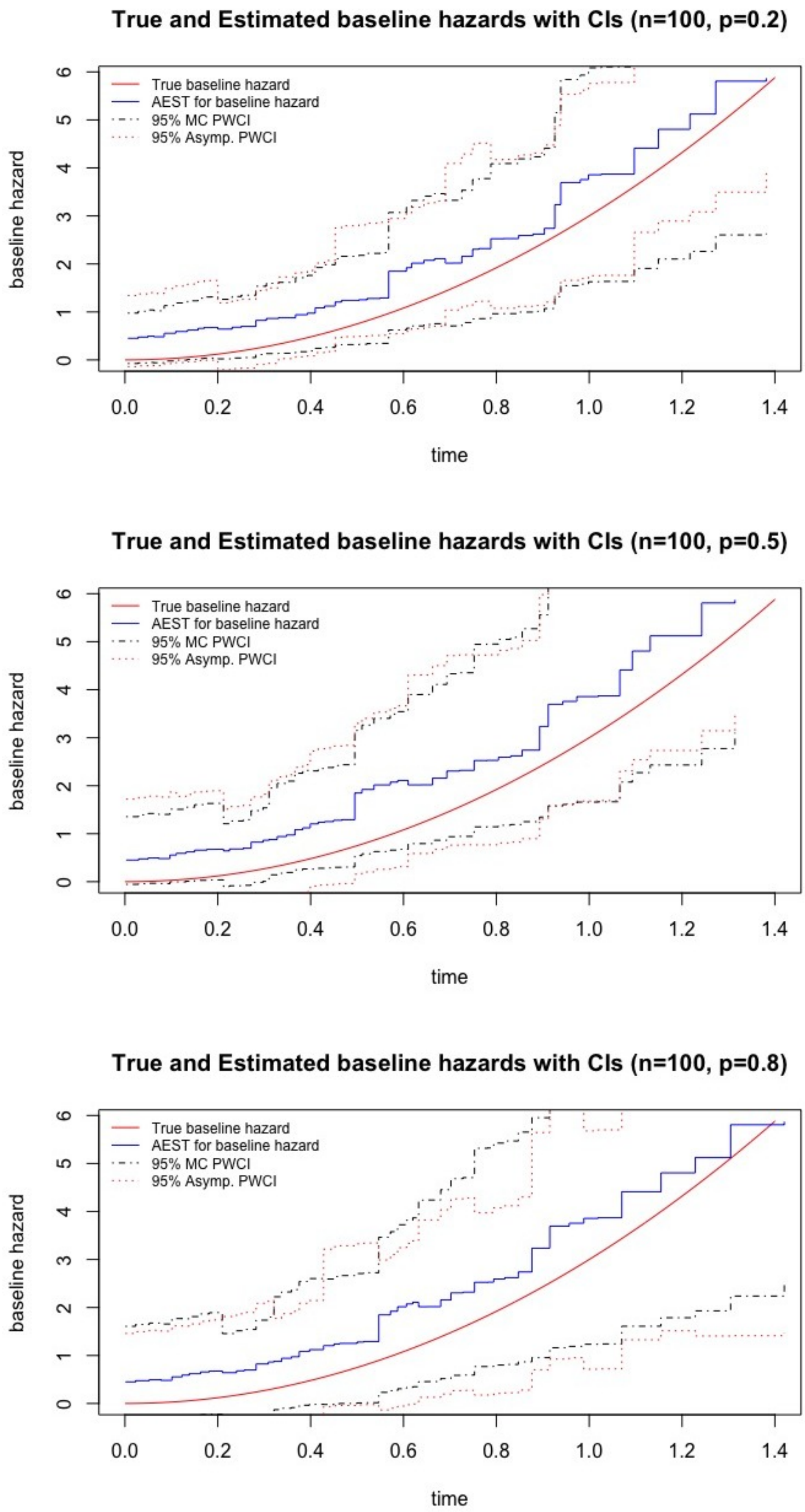

Figure 1: Plot of the true $h_0(t)$, average $h_0(t)$ estimates, Monte Carlo PWCI, and asymptotic PWCI for partly interval-censored survival data for $n = 100$

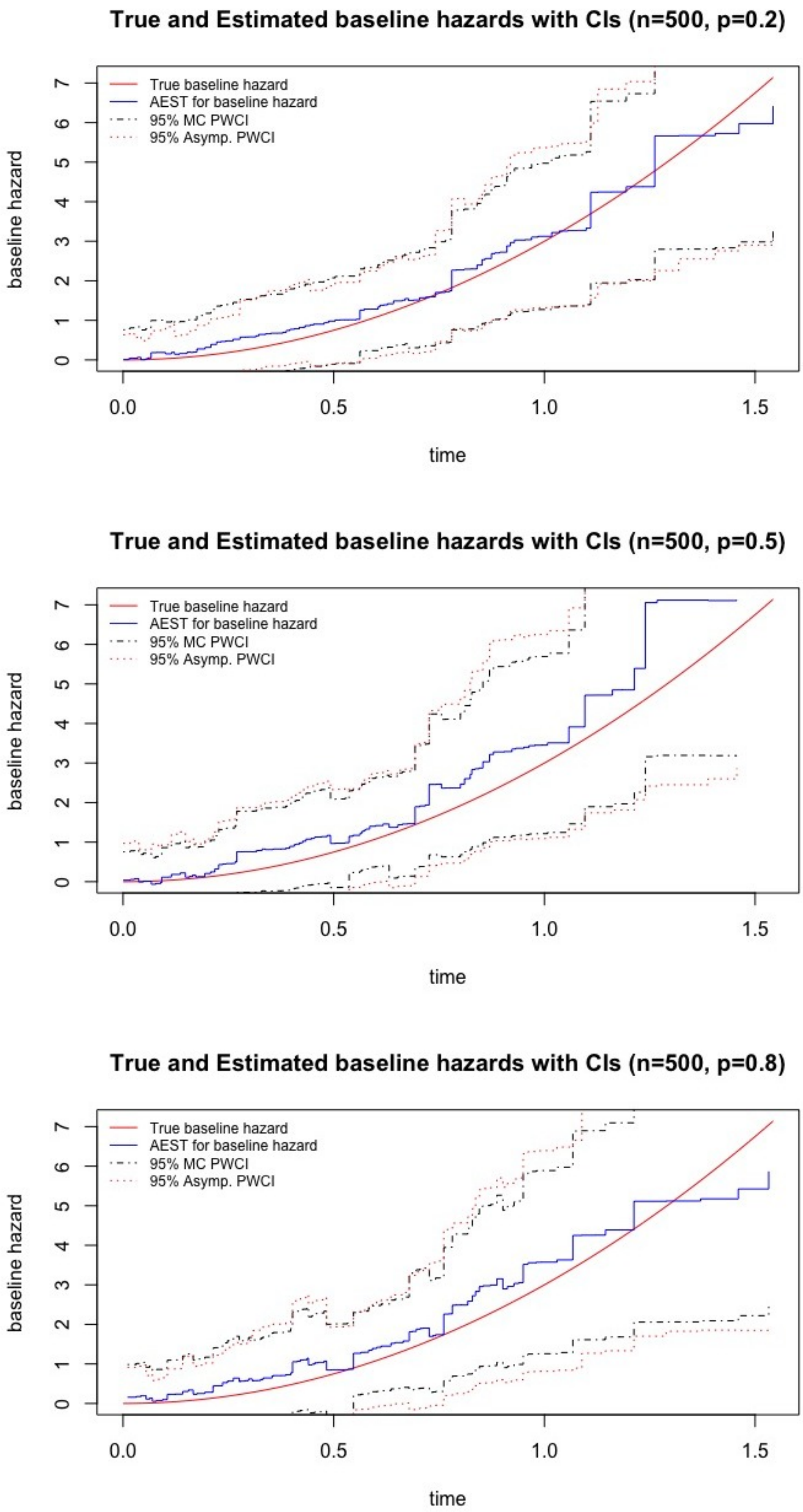

Figure 2: Plot of the true $h_0(t)$, average $h_0(t)$ estimates, Monte Carlo PWCI, and asymptotic PWCI for partly interval-censored survival data for $n = 500$

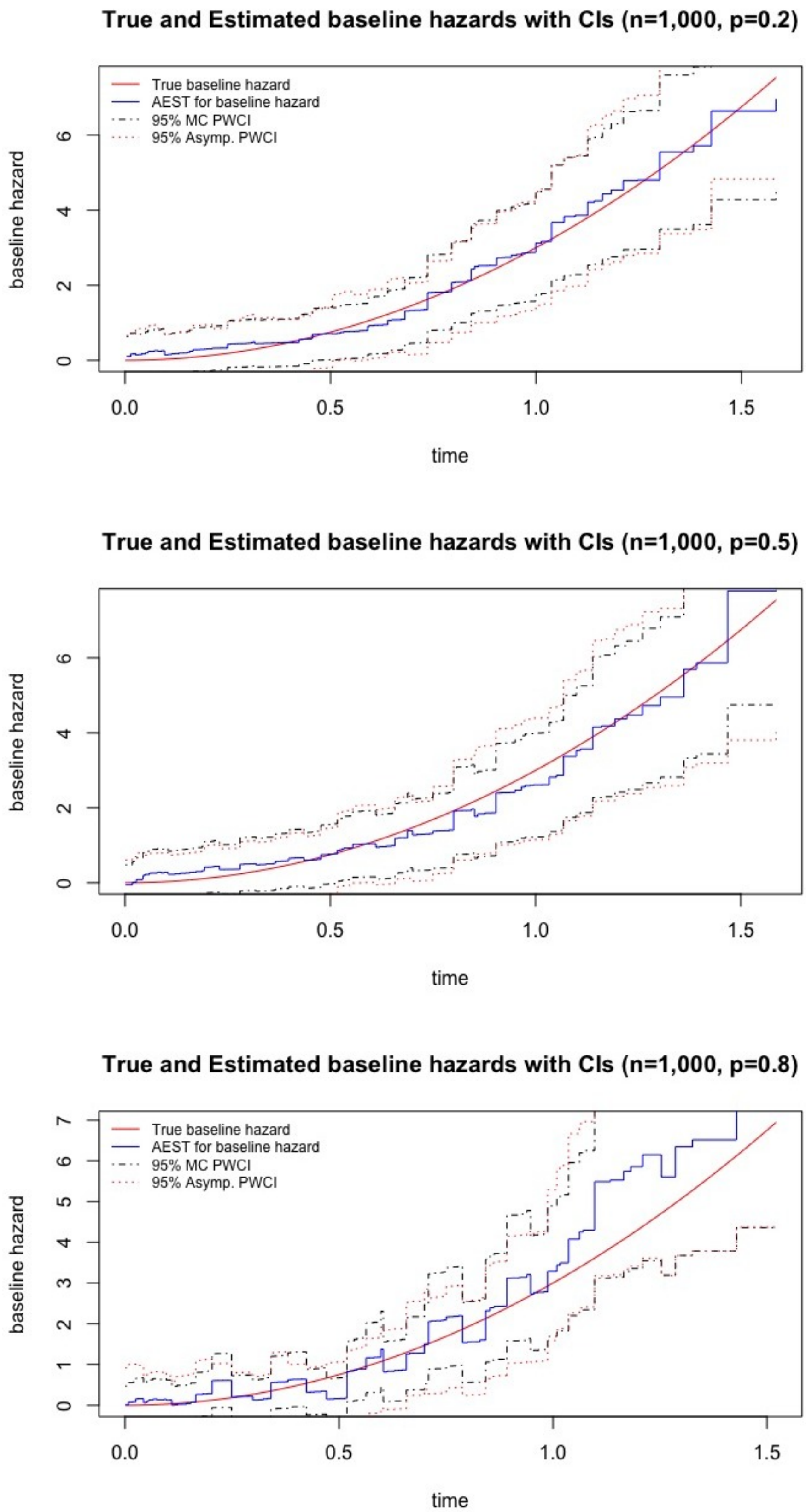

Figure 3: Plot of the true $h_0(t)$, average $h_0(t)$ estimates, Monte Carlo PWCI, and asymptotic PWCI for partly interval-censored survival data for $n = 1,000$

Lagrangian method which computes the augmented Lagrangian with the help of the Newton-MI algorithm and the estimates of hazard and baseline hazard are constrained to be non-negative by using the Newton-MI algorithm. In this study, we considered the piecewise constant approximation to discretize the non-parametric baseline hazard. The two simulation studies demonstrate that our MPL method works well for any combination of small, intermediate and large number of observations and small, moderate and large censoring proportions. Comparatively good estimates can be obtained even with very small sample sizes (such as n=30 or 50) with small censoring proportions (i.e.: π_c=20%) . The sandwich formula given in the asymptotic analysis in Chapter 5 provides accurate variance approximations for both the regression coefficients and baseline hazard. Furthermore, the simulation study on right-censored data shows that the MPL method outperforms the existing parameter estimation methods developed by Aalen (1980) and Lin & Ying (1994). In developing this MPL method, we assume the covariates are time independent and the censoring time is independent of the failure time. However, it is possible to extend our method to fit the additive hazards model with time-dependent covariates and dependent censoring.

Chapter 7

Real data application

In this chapter, we apply our MPL method to fit the additive hazards model to a melanoma data set which was kindly provided by Melanoma Institute of Australia. This data contains information regarding patients who were diagnosed with melanoma. Melanoma is a type of cancer that develops in the skin's pigment cells. It is the most serious form of skin cancer and grows very quickly if left untreated. Melanoma can spread to the inner lower part of the skin which is known as dermis, enter the lymphatic system or bloodstream and then spread to other parts of the body e.g. lungs, liver, brain or bone very quickly. Similar to other cancers, melanoma can also return even after it has been treated, and it is called recurrence. Recurrent melanoma may appear locally (at or near the site of the original primary melanoma tumor), or in a different part of the body. Melanoma can comes back even 10 years after it was first treated.

This melanoma data set contains records for 2,175 patients covering the study period from January 1, 1998 to March 14, 2016. Each patient record consisted of the diagnosis date of melanoma (t_{diag}) and the date of the last follow-up (t_{last}) with the status at the last follow up. Furthermore, for some of the patients, it may contain the first melanoma recurrence observed date (t_{first}) and the date of the last negative check before the first melanoma recurrence (t_{neg}). Records for three patients have been removed from the analysis, since those records did not follow the chronological order of the following dates: t_{diag}, t_{first}, t_{last}. This data set covers the patient diagnostic period from January 1, 1998 to December 30, 2002 and those patients were followed up until March 14, 2016.

In this study, we are mainly interested in the time taken to the first melanoma recurrence, which is generally interval-censored. We set the time of the melanoma diagnosis, t_{diag} as the time origin for each patient. The time to event should be considered differently depending on the availability of relevant time points. For a patient whose t_{neg} and t_{first} are available, the first melanoma recurrence time is interval-censored in $[(t_{neg}$-$t_{diag}), (t_{first}$-$t_{diag})]$. Both of these time points were available for 474 patients. For a patient whose t_{first} is available, but t_{neg} is missing, then the melanoma recurrence time can be left-censored in $[0, (t_{first}$-$t_{diag})]$ and 234 patients were in this category. If both t_{neg} and t_{first} are missing, but the status of t_{last} indicates melanoma, then melanoma recurrence time is left-censored in

[0, (t_{last}-t_{diag})]. 15 patients were belong to this category. All the remaining 1,449 patients were identified as missing recurrence of melanoma and the melanoma recurrence time can be right-censored in [(t_{last}-t_{diag}),+∞].

According to the primary diagnosis, melanoma was identified on the arm, head & neck, leg and trunk. Then, Sentinel Lymph Node Biopsy (SLNB) was performed to check the presence of cancer cells. Lymph Node Dissection (LND) was performed for 193 patients to remove the lymph nodes that have cancer cells. Complete LND was commonly performed (76.7%) in the excision process, followed by elective LND (22.3%) and therapeutic LND (1%). Out of all the melanoma diagnosed patients, there were 33.3% melanoma recurrences and 66.7% non-recurrences. First recurrence information was available for 706 patients and out of them only 3.4% experienced the recurrence tumor at or near the site of the original primary melanoma tumor. The remaining 96.4% experienced the recurrence tumor on a site different to the primary diagnosis site. The least common type of first recurrence is regional field recurrences (1.1%) and the most common recurrence is regional node recurrences(39.1%). As per the status of the last follow up, there were 65.0% alive patients at the end of the study period and the remaining 35.0% were deceased during the study period. Among the alive patients, 92.6% patients showed no sign of recurrence, 3.9% with melanoma and 3.5% with melanoma status unknown. Among the dead patients, 54.0% of them had melanoma, 13.6% of them died without melanoma recurrence and 32.4% died due to an unknown cause.

The impact of the following factors on the risk of recurrence was evaluated using the MPL method for the additive hazards model: patient sex; age at diagnosis of melanoma; primary tumor location; melanoma thickness and status of SLNB. Here, the melanoma thickness or the Breslow thickness was considered as a categorical variable. A summary of those covariates is given in Table 7. For this data set, we assume that the observed time points are independent to each other, and assume those time points follow the additive hazards model specified in 1.17.

Based on the asymptotic normality properties of $\hat{\boldsymbol{\beta}}$ discussed in Chapter 5, we performed a z-test of the null hypothesis, $H_0 : \beta_j = 0$ versus the alternative hypothesis, $H_a : \beta_j \neq 0$ using the estimates of the regression coefficients. Gender is a significant covariate and has a negative regression coefficient for females with reference to males. This implies that the male patients have a higher risk of melanoma recurrence compared to female patients. Similarly, age at diagnosis is significant and the corresponding estimate explains that the chance of recurring melanoma will increase as someone gets older to diagnose melanoma. Since the regression coefficients in the additive hazards model explain the difference of the hazards, it is possible to interpret the regression estimate for 'age at diagnosis' as: the risk of melanoma recurrence increases by 0.0012 when the age at diagnosis increases by one year. Comparing the primary site of melanoma, the risk of recurrence when the trunk is the primary location is insignificant. But, when leg is the primary location, the risk of melanoma recurrence is significant and the corresponding risk is higher compared to that of arm as the primary location by 0.1836 on average. The status of SLNB is highly significant and the patients with biopsy had a lower risk of recurring melanoma as opposed to the patients without biopsy. Thickness between 1 and 2 units is the significant categorical level associated with covariate thickness. These results are summarised in Table 8.

To determine how the risk of melanoma recurrence behaves when all the covariates set to their baseline values for males and females, the baseline hazard estimates were obtained from the MPL method. Estimates of the baseline hazard and the corresponding 95% piecewise confidence intervals for males and females are presented in Figure 4. This plot shows that for both males and females, the risk of recurring melanoma decreases monotonically with time. Significant differences can be seen on the decreasing pattern for male and female patients. It can be clearly seen that the risk of recurring melanoma deceases to a level close to 0 nearly after the first 6 years of melanoma diagnosis for male patients, whereas for female patients, the risk of recurrence decreases gradually over time. Furthermore, it can be observed that the baseline hazard estimate, produced by our MPL method, gave clear patterns of hazard over time. Similarly, the MPL method can be applied to fit additive hazards model for data sets which contains any type of censoring.

Characteristic	**Finding***
Age at diagnosis, years, mean ± SD (range)	57.25±16.77(5, 98)
Melanoma thickness	
<1	191(8.8%)
1-2	961(44.2%)
2-4	678(31.2%)
≥ 4	342(15.8%)
Gender	
Male	1287(59.3%)
Female	885(40.7%)
Primary melanoma diagnosis location	
Arm	330(15.2%)
Head and neck	443(20.4%)
Leg	593(27.3%)
Trunk	806(37.1%)
Sentinel Lymph Node Biopsy	
Performed	953(43.9%)
Not performed	1219(56.1%)

Table 7: Summary of the covariates in the model for 2,172 melanoma patients

*Data are expressed as number (percentage) unless otherwise indicated

	Estimate	Std. Error	z-value	P value
Gender: Female	-0.1356	0.0644	-2.1056	0.0352
Age at diagnosis (years)	0.0012	0.0006	2.1376	0.0325
Location: Head and Neck	0.1385	0.0794	1.7450	0.0810
Location: Leg	0.1836	0.0726	2.5288	0.0114
Location: Trunk	0.7751	0.6801	1.1397	0.2544
SLNB Status: Performed	-0.1879	0.0481	-3.9068	$<$0.0001
Thickness: 1-2	-0.2453	0.0905	-2.7099	0.0067
Thickness: 2-4	0.0394	0.9690	0.0406	0.9676
Thickness: ≥ 4	0.0699	0.1107	0.6316	0.5277

Table 8: Regression coefficient estimates given by the MPL method with p-values, treatment contrasts were considered for the variables '**Gender**', '**Location**' and '**Thickness**' with reference '**Male**', '**Arm**' and '**Thickness $<$1**' respectively

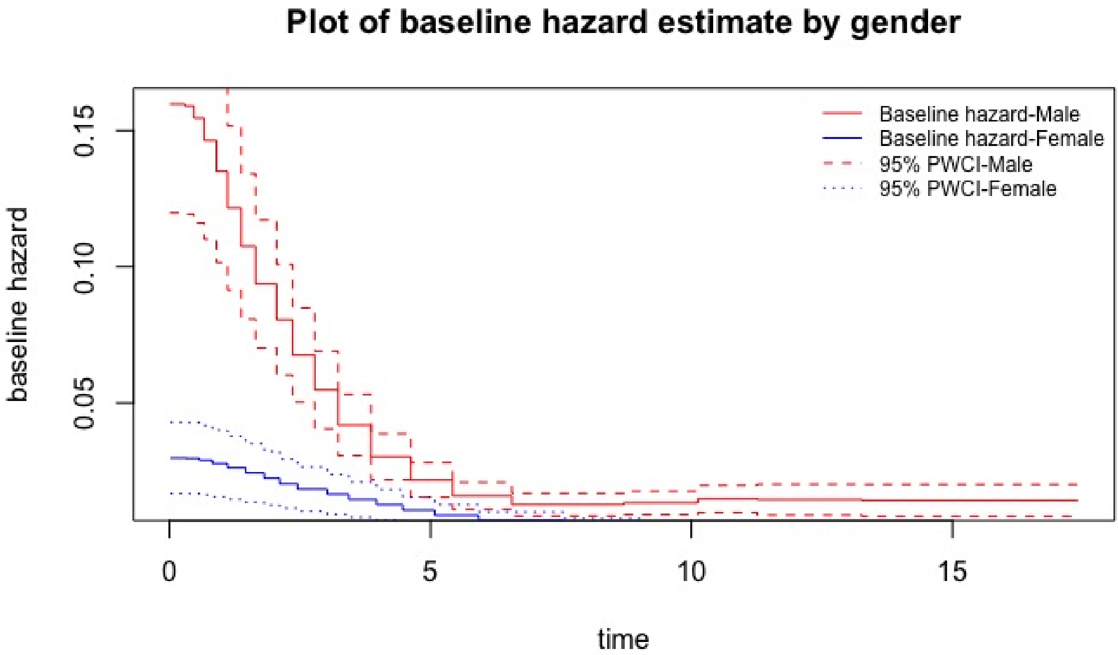

Figure 4: Plot of the estimated $h_0(t)$ against the survival time for males and females

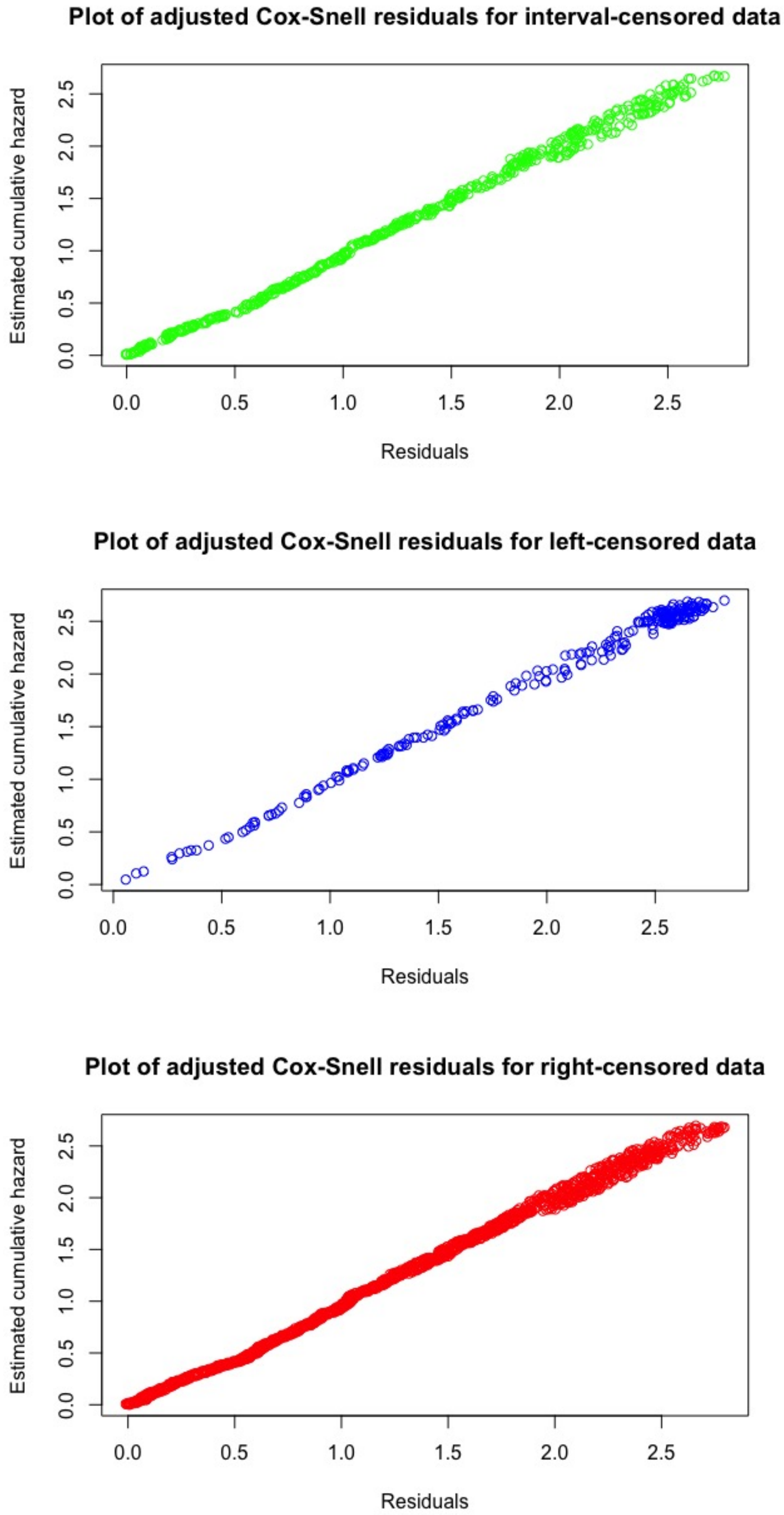

Figure 5: Plots of the adjusted Cox-Snell residuals for different censoring types

Figure 5 shows a plots of the adjusted Cox-Snell residuals versus the estimated cumulative hazards of residuals. If a model fits well, then the graphs will be approximately a $45°$ line. Thus, these figures can be used to demonstrate the goodness of fit of this estimated additive hazards model. Since the estimated cumulative hazard approximately follow the $45°$ degree line with some slight deviations for all three censoring types, this plot suggests that this model fit the data well.

Chapter 8

Conclusion and Future work

8.1 Conclusion

In this thesis, we developed a novel parameter estimation procedure for the additive hazards model with partly interval-censored data, which comprised exactly observed, left- censored, finite interval-censored and right-censored survival data. We fit the additive hazards model by estimating the regression coefficients and the baseline hazard function. We assumed that the observations from different individuals were independent. For each individual, the distribution of covariates did not involve regression coefficients and the censoring time was independent of the failure time. Furthermore, we assumed that the covariates were time independent.

This thesis developed a maximum penalized log-likelihood (MPL) method to estimate the regression coefficients and the underlying non-parametric baseline hazard function, simultaneously, by imposing non-negativity constraints on the baseline hazard and the overall hazard function. We approximated an infinite dimensional baseline hazard from a finite number of non-negative basis functions. We used an indicator basis function which led to a piecewise constant baseline hazard function. We obtained smoothed baseline hazard estimates through the penalty function, which represented the square of the second order differences for the basis coefficients of the baseline hazard.

We encountered a constrained optimization problem as it was necessary to consider the two non-negativity constraints and was essential that we used an optimization method which could handle large numbers of constraints in order to solve this problem. We adopted the augmented Lagrangian method to solve this constraint optimization issue and the estimates were obtained simultaneously using the Newton and multiplicative iterative (Newton-MI) algorithm which combines the Newton algorithm and the MI algorithm (Ma 2010).

During the parameter estimation procedure, the regression coefficients were estimated by the Newton algorithm, while the baseline hazard was estimated by the MI algorithm. The constraints on the baseline hazard and the overall hazard were imposed simultaneously and directly by the MI algorithm. The MI algorithm not only guaranteed the non-negativity of the baseline hazard, but also helped avoid the requirement of the second order derivatives of the augmented Lagrangian with respect to the baseline hazard. Moreover, as per the augmented Lagrangian, the new constraint $\eta_i = \mathbf{x}_i\boldsymbol{\beta}$ was also updated using the MI algorithm, and the corresponding Lagrange multiplier was updated using a standard dual update equation. Each of these parameter updates was followed by a line search, which ensured that the likelihood increased during each parameter update.

The asymptotic properties of these derived constrained MPL estimators can be studied in two ways; one is more important in theoretical nature (when the number of basis functions goes to infinity), and the second method is more useful in applications (when the number of basis functions is fixed). These asymptotic properties can be used for model checking and the model validation process.

According to the limited existing literature on semi-parametric estimation methods for the additive hazards model with interval-censored data, some methods estimate only the regression coefficients by considering the baseline hazard as a nuisance parameter. Some methods estimate the baseline hazard indirectly, and others do not impose the non-negativity constraints on both baseline hazard and overall hazard function. Thus, these factors cause limitations, such as the inability to use these methods for prediction, and an incapacity for obtaining asymptotic variances of these estimates and violations of the two non-negativity constraints. The proposed MPL method, however, addressed all the above limitations, and it estimated the model parameters simultaneously by imposing the two non-negativity constraints.

We investigated our proposed MPL method under the additive hazards model by conducting simulation studies. Firstly, we performed the simulation results for less complex right-censored data. Those results were compared against the results of two other existing parameter estimation methods. Then, the MPL method was used to perform simulation studies on more complex partly interval-censored data. In both of the simulation studies, the results showed that the bias and standard deviation for the MPL estimate of regression coefficients increased with the censoring proportion but decreased with the sample size. The MPL estimate for the baseline hazard also exhibited the same trend. The sandwich formula derived from asymptotic theory provided a good estimate for the Monte Carlo standard deviation, especially when the sample size was large, or the censoring proportion was small.

Comparisons between our MPL method and the other two existing methods were also made in the simulation studies. We compared results of our MPL method with the results of parameter estimation methods by Lin & Ying (1994) and Aalen (1980). Irrespective of the sample size and the censoring proportion, the biases of the estimates and standard deviations in the MPL method were smaller than those in Aalen's method and Lin and Ying's method. Thus, our MPL method achieved a lower mean squared error (MSE) for the estimate of the regression coefficients, compared to those of the other two existing methods. Furthermore, we observed that our MPL method gave a much smaller average integrated squared error (AISE) for the baseline hazard estimate than Lin and Ying's method.

In a real data analysis, we applied our MPL method to fit the additive hazards model to a melanoma data set. In this study, we were primarily interested in the time taken to the first melanoma recurrence, which was generally interval-censored. But, depending on the availability of time points, some of the observations were left- or right-censored. For this dataset, we observed that the smoothed baseline hazard estimate produced good estimated over time, by our MPL method. Furthermore, the Cox-Snell residual plot showed that the estimated model fit the data well.

8.2 Future work

We would like to extend the proposed MPL method of this thesis further, based on the following future research directions.

In developing this MPL method, we assumed that the covariates were time independent and the censoring time was independent of the failure time. However, it is possible to extend this MPL method to estimate the additive hazards model with partly interval-censored data under time-dependent covariates and dependent censoring where the censoring depends on the failure time. The use of time dependent covariates is a simple modification to this proposed method and the use of dependent censoring extends beyond the current work.

Furthermore, when we discretized the baseline hazard, we considered indicator basis function for simplicity. Extending this MPL method, which accommodates basis functions such as M-spline, would be of interest for future research. One of the major advantages of the proposed method is its capability of making prediction. Thus, we are expecting to develop prediction accuracy measures which support the prediction process as a numerical evidence. This proposed method provides a graphical tool to empirically assess the goodness-of-fit. As an extension of assessing the model fitting, it is possible to develop few other residual types such as Martingale and Schoenfeld residuals. Furthermore, we would like to propose an official model assessment procedure for the additive hazards model as well. Finally, we are expecting to develop a R function based on this proposed method with expected further developments and add to the existing R package, $SurvivalMPL$ which uses to estimate model parameters of the Cox proportional hazards model using MPL approach.

Appendix A

Appendix: R code

Note that the following R codes are used to generate and check the results presented in this thesis.

A.1 Data Generation

```
#' @title Additive hazard fitting by MPL with arbitrary censoring,
#'  including interval, left and right censoring
#' @author Kasun Rathnayake
#' @param n the sample size
#' @param p the number of covariates
#' @param beta regression coefficients to simulate the data set
#' @param lambda the smoothing parameter
#' @description This funtion simulates data from weibull distribution
#' @export
data.gen=function(n,beta,cp)
{
  lambda=1;shape=3
  x_up=c(1,2,3); x_low=c(-1,0,0)
  p=length(t.beta)
  temp.x=matrix(nrow=n,ncol=p)
  time.surv=matrix(nrow=n,ncol=1)
  aa=rep((1/lambda)^shape,n) ;cc=runif(n,0,1)
```

```
val=rep(-0.01,n) ; coeff=rep(0,shape+1)
for(count in 1:n)
{
  while(val[count]<0)
  {
    temp.x[count,]=as.matrix(data.frame(x1=runif(1,x_low[1],x_up[1]),
            x2=runif(1,x_low[2],x_up[2]),x3=runif(1,x_low[3],x_up[3])))
    coeff[1]=log(cc[count]) ;coeff[2]=(temp.x[count,]%*%beta) ;
    coeff[shape+1]=aa[count]
    temp.roots=polyroot(coeff)
    time.surv[count]=max(Re(temp.roots[abs(Im(temp.roots))<1e-10]))
    ###Need to consider about abs here
    val[count]=temp.x[count,]%*%beta+
      (shape*((time.surv[count])^(shape-1)/lambda^shape)) }}
###End of While & For loop
ind=rbinom(n,1,(cp))
###(censoring prop) here 0 for observed and 1 for censored
cl=runif(n); cr=cl+runif(n)##left and right r.v's
l=r=xi.temp=matrix(rep(0,n))
time=status=matrix(rep(0,n))
for(ab in 1:n)
```

```
  { if(ind[ab]==0)
  {l[ab]=r[ab]=time.surv[ab];xi.temp[ab]=0  ###For ah_mpl
  time[ab]=time.surv[ab];status[ab]=2}
    else
    { if(time.surv[ab]<=cl[ab])
    {l[ab]=0;r[ab]=cl[ab];xi.temp[ab]=1
    time[ab]=(cl[ab]/2);status[ab]=1}
      else if(cl[ab]<time.surv[ab] && time.surv[ab]<=cr[ab])
      {l[ab]=cl[ab];r[ab]=cr[ab];xi.temp[ab]=2
      time[ab]=(cl[ab]+cr[ab])/2;status[ab]=1}
      else if(time.surv[ab]>cr[ab])
      {l[ab]=cr[ab];r[ab]=Inf;xi.temp[ab]=3
      time[ab]=cr[ab];status[ab]=1}}}###End of For loop
  temp.x=data.frame(temp.x); time.mat=data.frame(cbind(l,r,xi.temp))
  ;mid.point=data.frame(time,status)
  return(list(time.mat=time.mat,temp.x=temp.x,mid.point=mid.point))}
###Combining Survival time is optional
```

A.2 Parameter estimation using MPL method

```
#' @title Additive hazard function to estimate parameters
#' @export
```

```
add.haz=function(n,p,data,bin.count,smooth,maxiter,alpha)
{

  times<<-c(times, proc.time()[1])
  lambda=1;shape=3
  xi=data[,3]
  obs=matrix(data[,1][xi==0]);l.cens=matrix(data[,2][xi==1]);
  il.cens=matrix(data[,1][xi==2]);ir.cens=matrix(data[,2][xi==2])
  ;r.cens=matrix(data[,1][xi==3])
  n.obs=length(obs);nl.cens=length(l.cens);ni.cens=length(il.cens);
  nr.cens=length(r.cens);n.new=n+ni.cens

  ###Centering the covariates
  x=as.matrix(data[,4:length(data)])

  time.obs=rbind(obs,l.cens,il.cens,ir.cens,r.cens)

  xi.order=c(rep(0,n.obs),rep(1,nl.cens),rep(2,ni.cens),
             rep(2,ni.cens),(rep(3,nr.cens)))
  bin.cut=cut2(time.obs,g=bin.count,onlycuts=TRUE)
```

```
m=bin.count

bin.cut=bin.cut+1e-12;bin.cut[m+1]=max(time.obs)+1e-12;

bin.cut[1]=max(0,(min(time.obs)-1e-12))

bin.wid=diff(bin.cut)


id=histc(time.obs,bin.cut)  ###which bin

id.obs=id$bin[xi.order==0]; id.lcens=id$bin[xi.order==1];

id.ilcens=id$bin[xi.order==2][1:ni.cens]

id.ircens=id$bin[xi.order==2][(ni.cens+1):(2*ni.cens)];

id.rcens=id$bin[xi.order==3]

id.ordered=c(id.obs,id.lcens,id.ilcens,id.ircens,id.rcens)

R=matrix(nrow=m,ncol=m,0)

for(i in 1:m){

  if (i<=m-1){R[i,i+1]=-4

             R[i+1,i]=-4}

  if (i<=m-2){R[i,i+2]=1

             R[i+2,i]=1}

  if (i<=m){R[i,i]=6}}

R[1,1]=R[m,m]=5

R[2,2]=R[m-1,m-1]=9

R[1,2]=R[2,1]=R[m,m-1]=R[m-1,m]=-6
```

```
x.obs=matrix(x[xi==0],n.obs,p);xl.cens=matrix(x[xi==1],nl.cens,p)
;xi.cens=matrix(x[xi==2],ni.cens,p);xr.cens=matrix(x[xi==3],nr.cens,p)
x.ordered=rbind(x.obs,xl.cens,xi.cens,xi.cens,xr.cens)
old.theta=rep(0.1,m);old.beta=rep(0.8,p);old.gamma=rep(1e-2,(n+ni.cens))
#old.eta=rep(0,(n+ni.cens))
old.eta=(x.ordered%*%old.beta)
mini=.Machine$double.eps ###The smallest floating poing number
p_like=like0=like1=like2=like3=NULL;alpha_dum=alpha_ini=2
times<<-c(times, proc.time()[1])
psi.obs=matrix(0,n.obs,m)
if(n.obs>0){
  for(i in 1:n.obs){
    psi.obs[i,id.obs[i]]=1}}


phi=matrix(0,nrow=(n+ni.cens),ncol=m)
for(i in 1:(n+ni.cens)){
  for(j in 1:m){
    if(j<(id.ordered)[i])
      phi[i,j]=bin.wid[j]
    else if(j==(id.ordered)[i])
      phi[i,j]=(time.obs[i,1]-bin.cut[j])
```

```
    #else phi[i,j]=0
  }}

phi.mat=phi

phi.obs=phi.mat[(1:n.obs),]
phi.rcens=phi.mat[((n.new-nr.cens+1):n.new),]
phi_u.obs=apply(phi.obs,2,sum)
phi_u.rcens=apply(phi.rcens,2,sum)
phi.lcens=phi.mat[((n.obs+1):(n.obs+nl.cens)),];phi.ilcens=
  phi.mat[((n.obs+nl.cens+1):(n-nr.cens)),];phi.ircens
=phi.mat[((n-nr.cens+1):(n.new-nr.cens)),]
phi_u.lcens=apply(phi.lcens,2,sum);phi_u.ilcens=
  apply(phi.ilcens,2,sum);phi_u.ircens=apply(phi.ircens,2,sum)

i_maxiter=maxiter[1];o_maxiter=maxiter[2]
iter_outter=0;iter_inner=NULL;iter_count=0;iter=0
s.time1=s.time2=s.time3=NULL
s.theta1=s.theta2=s.eta1=s.eta2=s.beta1=s.beta2=NULL

times<<-c(times, proc.time()[1])
```

```
#### Main function ####

for(j in 1:o_maxiter)##for this use FOR loop with number of iterations

{

  times<<-c(times, proc.time()[1])

  #iter_outter=iter_outter+1

  alpha=alpha

  alpha[alpha < (alpha_ini*(alpha_dum^10))]=alpha_ini*(alpha_dum^(j-1))


  #iter=0

  for(s in 1:i_maxiter){

    times<<-c(times, proc.time()[1])

    t1=proc.time()[[3]]

    iter=iter+1

    #iter_count=iter_count+1

    old.theta1=old.theta; old.theta1[old.theta1<mini]=mini

    ###Just like eps

    like0[iter]=likelihood0=like.func(old.theta,old.eta,

        old.beta,old.gamma,alpha,phi.mat,x.ordered,time.obs,

        phi.obs,R,smooth,x.obs,xi.order,ni.cens,psi.obs)


    ####Applying MI algorithm for theta####
```

```
times<<-c(times, proc.time()[1])
t3=proc.time()[[3]]


surv.vec2=exp(-((phi.mat%*%old.theta)+
                 ((x.ordered%*%old.beta)*time.obs)))
surv.obs=surv.vec2[xi.order==0]
surv.lcens=surv.vec2[xi.order==1];surv.ilcens=
  surv.vec2[xi.order==2][1:ni.cens]
surv.ircens=surv.vec2[xi.order==2][(ni.cens+1):(2*ni.cens)];
surv.rcens=surv.vec2[xi.order==3]
left=(t(l.cens*(surv.lcens/(1-surv.lcens)))%*%xl.cens)
leftint=(t(il.cens*(surv.ilcens/(surv.ilcens-surv.ircens)))
%*%xi.cens)
rightint=(t(ir.cens*(surv.ircens/(surv.ilcens-surv.ircens)))
%*%xi.cens)
full=t((1/(psi.obs%*%old.theta+x.obs%*%old.beta))-obs)%*%x.obs
right=t(r.cens)%*%xr.cens
lag_grad=((old.gamma)+
            alpha*(as.vector((x.ordered%*%
            old.beta)-old.eta)))%*%x.ordered
```

```
grad=left-leftint+rightint+full-right-lag_grad


left_h=t(xl.cens)%*%diag(as.vector(surv.lcens*
                      (l.cens/(1-surv.lcens))^2))%*%xl.cens
int_h=t(xi.cens)%*%diag(as.vector(surv.ilcens*
    surv.ircens*((il.cens-ir.cens)/
    (surv.ilcens-surv.ircens))^2))%*%xi.cens
full_h=t(x.obs)%*%diag(as.vector((1/(psi.obs%*%
                  old.theta+x.obs%*%old.beta))^2))%*%x.obs
lag_hess=alpha*t(x.ordered)%*%(x.ordered)


hess=left_h+int_h+full_h+lag_hess


idH=which(diag(hess)==0)
hess[idH,idH]=mini


inch3=as.vector(grad%*%solve(hess))
new.beta=as.vector(old.beta+inch3)
likelihood1=like.func(old.theta,old.eta,new.beta,old.gamma,
     alpha,phi.mat,x.ordered,time.obs,phi.obs,R,smooth,x.obs,
                      xi.order,ni.cens,psi.obs)
```

```
sigma=0.6
s.beta1[iter]=proc.time()[[3]]-t3
t6=proc.time()[[3]]

while((likelihood1 <= likelihood0)){
  new.beta=old.beta+sigma*(inch3)
  likelihood1=like.func(old.theta,old.eta,new.beta,
  old.gamma,alpha,phi.mat,x.ordered,time.obs,phi.obs,
  R,smooth,x.obs,xi.order,ni.cens,psi.obs)
  if(sigma>=1e-2)
    sigma=0.6*sigma
  else if (sigma < 1e-2 & sigma >= 1e-5)
    sigma=5e-2*sigma
  else if (sigma <1e-5 & sigma > 1e-30)
    sigma=1e-5*sigma
  else
    break}  ## end of while for updating theta

s.beta2[iter]=proc.time()[[3]]-t6
s.time3[iter]=proc.time()[[3]]-t3
```

```
####Applying MI algorithm for beta####


times<<-c(times, proc.time()[1])

like1[iter]=likelihood1

surv.vec1=exp(-((phi.mat%*%old.theta)+

                  ((x.ordered%*%new.beta)*time.obs)))

surv.obs=surv.vec1[xi.order==0];surv.lcens=

  surv.vec1[xi.order==1];surv.ilcens=surv.vec1

  [xi.order==2][1:ni.cens]

surv.ircens=surv.vec1[xi.order==2][(ni.cens+1):

     (2*ni.cens)];surv.rcens=surv.vec1[xi.order==3]


penal=(2*R%*%old.theta);penal1=penal2=penal

denom1=as.vector(t(surv.ilcens/(surv.ilcens-surv.ircens))

  %*%phi.ilcens)+(phi_u.obs)+(phi_u.rcens)+

  (smooth*penal1)+c(rep(1e-3,m))

num1=as.vector(t(surv.lcens/(1-surv.lcens))

%*%phi.lcens)+

  as.vector(t(surv.ircens/(surv.ilcens-surv.ircens))

  %*%phi.ircens)+
```

```
  as.vector((t(1/(psi.obs%*%old.theta+x.obs%*%new.beta))
  %*%psi.obs))-
  (smooth*penal2)+c(rep(1e-3,m))
delta1=num1-denom1


est.eta=old.eta
dd=data.table(id.ordered,(-est.eta));setkey(dd,id.ordered)
max.eta=(aggregate(. ~ id.ordered, data=dd, FUN=max)[,2])
max.eta[max.eta<0]=0
est.b=max.eta
ss=rep(0,m)
for(k in 1:m){
  if (old.theta[k]!=0){
    ss[k]=(old.theta[k]-est.b[k])/denom1[k]}
  else if (old.theta[k]==0 && denom1[k]<0){
    ss[k]=0}
  else ss[k]=(1e-5/denom1[k])
}
inch1=ss*delta1
new.theta=old.theta+inch1
likelihood2=like.func(new.theta,old.eta,new.beta,
```

```
old.gamma,alpha,phi.mat,x.ordered,time.obs,phi.obs,R,
smooth,x.obs,xi.order,ni.cens,psi.obs)
sigma=0.6
s.theta1[iter]=proc.time()[[3]]-t1
t4=proc.time()[[3]]
while((likelihood2 <= likelihood1)){
  new.theta=old.theta+sigma*(inch1)
  likelihood2=like.func(new.theta,old.eta,new.beta,
  old.gamma,alpha,phi.mat,x.ordered,time.obs,
  phi.obs,R,smooth,x.obs,xi.order,ni.cens,psi.obs)
  if(sigma>=1e-2)
    sigma=0.6*sigma
  else if (sigma < 1e-2 & sigma >= 1e-5)
    sigma=5e-2*sigma
  else if (sigma <1e-5 & sigma > 1e-30)
    sigma=1e-5*sigma
  else
    break
}  ## end of while for updating beta
s.theta2[iter]=proc.time()[[3]]-t4
s.time1[iter]=proc.time()[[3]]-t1
```

```
####Applying MI algorithm for eta####
times<<-c(times, proc.time()[1])
t2=proc.time()[[3]]
like2[iter]=likelihood2
new.theta1=new.theta; new.theta1[new.theta1<mini]=mini
new.eta=(old.gamma+alpha*(x.ordered%*%new.beta))/alpha
inch2=new.eta-old.eta
likelihood3=like.func(new.theta,new.eta,new.beta,old.gamma,
    alpha,phi.mat,x.ordered,time.obs,phi.obs,R,smooth,
    x.obs,xi.order,ni.cens,psi.obs)
sigma=0.6
s.eta1[iter]=proc.time()[[3]]-t2
t5=proc.time()[[3]]
while((likelihood3 <= likelihood2)){
  new.eta=old.eta+sigma*(inch2)
  likelihood3=like.func(new.theta,new.eta,new.beta,old.gamma,
     alpha,phi.mat,x.ordered,time.obs,phi.obs,R,smooth,
     x.obs,xi.order,ni.cens,psi.obs)
  if(sigma>=1e-2)
    sigma=0.6*sigma
```

```
        else if (sigma < 1e-2 & sigma >= 1e-5)
          sigma=5e-2*sigma
        else if (sigma <1e-5 & sigma > 1e-30)
          sigma=1e-5*sigma
        else
          break
      }  ## end of while for updating eta
      s.eta2[iter]=proc.time()[[3]]-t5
      s.time2[iter]=proc.time()[[3]]-t2

      ####Applying MI algorithm for gamma####
      times<<-c(times, proc.time()[1])
      like3[iter]=likelihood3
      new.gamma=old.gamma+alpha*(x.ordered%*%new.beta-new.eta)
      new.gamma1=new.gamma
      likelihood4=like.func(new.theta,new.eta,new.beta,new.gamma,
            alpha,phi.mat,x.ordered,time.obs,phi.obs,R,smooth,
            x.obs,xi.order,ni.cens,psi.obs)
      p_like[iter]=likelihood4

{if((max(abs(new.theta-old.theta))<1e-6) &
```

```
(max(abs(new.beta-old.beta))<1e-6) ){
  break}
 else
 { old.beta=as.vector(new.beta)
   old.theta=new.theta
   old.eta=new.eta
   old.gamma=as.vector(new.gamma)}}
    }###End of for loop, end of i


print(j);print(s)
cat("Ext.␣iter:",j,"\tInt.␣iter:",s,"\n")
iter_inner[j]=iter


{if(max(abs(x.ordered%*%new.beta-new.eta))<1e-10){
  break}
 else
 { old.beta=as.vector(new.beta)
   old.theta=new.theta
   old.eta=new.eta
   old.gamma=as.vector(new.gamma)}}
  }##end of j
```

```
########################   Hessian matrix calculation
  surv.vec.f=exp(-((phi.mat%*%new.theta)+
                    ((x.ordered%*%new.beta)*time.obs)))
  surv.obs=surv.vec.f[xi.order==0]
  surv.lcens=surv.vec.f[xi.order==1];surv.ilcens=
    surv.vec.f[xi.order==2][1:ni.cens]
  surv.ircens=surv.vec.f[xi.order==2][(ni.cens+1):
           (2*ni.cens)];surv.rcens=surv.vec.f[xi.order==3]

  left_bh=t(xl.cens)%*%diag(as.vector(surv.lcens*
                     (l.cens/(1-surv.lcens))^2))%*%xl.cens
  int_bh=t(xi.cens)%*%diag(as.vector(surv.ilcens*
    surv.ircens*((il.cens-ir.cens)/
    (surv.ilcens-surv.ircens))^2))%*%xi.cens
  full_bh=t(x.obs)%*%diag(as.vector((1/(psi.obs%*%
                   new.theta+x.obs%*%new.beta))^2))%*%x.obs
  lag_bhess=alpha*t(x.ordered)%*%(x.ordered)

  hess_b=as.matrix(left_bh+int_bh+full_bh+lag_bhess)
  # cat("beta",new.beta,"\n")
```

```
left_th= t(phi.lcens)%*%
diag(surv.lcens/(1-surv.lcens)^2)%*%(phi.lcens)
int_th=t(phi.ircens-phi.ilcens)%*%diag((surv.ilcens*
    surv.ircens)/(surv.ilcens-surv.ircens)^2)
    %*%(phi.ircens-phi.ilcens)
full_th=t(psi.obs)%*%diag(as.vector((1/(psi.obs%*%
           new.theta+x.obs%*%new.beta))^2))%*%(psi.obs)


hess_t=hess_tg=left_th+int_th+full_th
hess_tf=hess_t+2*lambda*R


left_bth=t(xl.cens)%*%diag(as.vector((surv.lcens*
                  l.cens)/(1-surv.lcens)^2))%*%phi.lcens
int_bth=t(xi.cens)%*%diag(as.vector((surv.ilcens*surv.ircens)*
    (il.cens-ir.cens)/(surv.ilcens-surv.ircens)^2))
    %*%(phi.ilcens-phi.ircens)
full_bth=t(x.obs)%*%diag(as.vector((1/(psi.obs%*%
                        new.theta+x.obs%*%
                        new.beta))^2))%*%(psi.obs)


hess_bt=left_bth+int_bth+full_bth
```

```
hess_g=as.matrix(rbind(cbind(hess_b,hess_bt),
cbind(t(hess_bt),hess_tg)))
hess_f=rbind(cbind(hess_b,hess_bt),cbind(t(hess_bt),hess_tf))


new.para=c(new.beta,new.theta)
tol=1e-3
id_theta1=which(new.para<tol);id_theta2=which(new.para>=tol)
print(id_theta1);print(id_theta2)
if(length(id_theta1)==0){inv_hess=solve(hess_f)}
else{
  hess_f_mat=hess_f[-id_theta1,-id_theta1]
  inv_hess=solve(hess_f_mat)
}


mat=matrix(0,(m+p),(m+p))
for(i in 1:(m+p)){
  k=1
  for(j in 1:(m+p)){
    if(j==id_theta2[k]){
      mat[j,id_theta2]=inv_hess[k,]
```

```
        k=k+1
        if(k>length(id_theta2)) {break}
      }
      mat=matrix(mat,(m+p),(m+p))
    }
  }
  inv_hess_f=mat
  v_eta=(inv_hess_f)%*%(hess_g)%*%t(inv_hess_f)
#####################
times<<-c(times, proc.time()[1])
s.time=cbind(s.time1,s.time2,s.time3)
s.para=cbind(s.theta1,s.theta2,s.eta1,s.eta2,s.beta1,s.beta2)

surv.vec3=exp(-((phi.mat%*%new.theta)+
((x.ordered%*%new.beta)*time.obs)))
surv.obs=surv.vec3[xi.order==0]
surv.lcens=surv.vec3[xi.order==1];surv.ilcens=
surv.vec3[xi.order==2][1:ni.cens]
surv.ircens=surv.vec3[xi.order==2][(ni.cens+1):
               (2*ni.cens)];surv.rcens=surv.vec3[xi.order==3]
```

```
##  Cox-snell residuals (Revised)

x.ordered.res=rbind(x.obs,xl.cens,xi.cens,xr.cens)

##This is w/o having interval covariates twice

id.l.mid=histc((l.cens/2),bin.cut)$bin

id.i.mid=histc((ir.cens+il.cens)/2,bin.cut)$bin

id.ordered.res2=c(id.obs,id.l.mid,id.i.mid,id.rcens)

###id values for cox resudual




phi.lcens=phi.mat[((n.obs+1):(n.obs+nl.cens)),];phi.ilcens=
  phi.mat[((n.obs+nl.cens+1):(n-nr.cens)),];phi.ircens=
  phi.mat[((n-nr.cens+1):(n.new-nr.cens)),]

phi.lcens.mid=phi.lcens/2

phi.icens.mid=(phi.ilcens+phi.ircens)/2


surv.lcens.mid= exp(-((phi.lcens.mid%*%new.theta)+((xl.cens%*%
                                new.beta)*(l.cens/2))))

surv.icens.mid=exp(-((phi.icens.mid%*%new.theta)+((xi.cens%*%
                          new.beta)*(ir.cens+il.cens)/2)))

right.cox=log(surv.rcens)

full.cox=log(surv.obs)+1
```

```
left.cox=log(surv.lcens.mid)+1

int.cox=log(surv.icens.mid)+1

res.cox=c(full.cox,left.cox,int.cox,right.cox)

times<<-c(times, proc.time()[1])

##############


return(list(Bin.cut=bin.cut,Data=x.ordered,ID=id.ordered,ID.res2=
    id.ordered.res2,Iteration=iter_outter,iter_inner=iter_inner,
    like.mat=cbind(like0,like1,like2,like3,p_like),p_like=p_like,
    theta=new.theta,beta=new.beta,eta=new.eta,gamma=new.gamma,m=m,
    res.cox=res.cox,x.order=x.ordered,x.res=x.ordered.res,
    time.obs=time.obs,sys.time=s.time,s.para=s.para,hess_f=hess_f,
    hess_g=hess_g,v_eta=v_eta))
}



like.func=function(e.theta,e.eta,e.beta,e.gamma,alpha,
phi.mat,x.ordered,time.obs,phi.obs,R,smooth,x.obs,
xi.order,ni.cens,psi.obs){
  mini=.Machine$double.eps ###The smallest floating poing number
  theta1=e.theta; theta1[theta1<mini]=mini
```

```
  surv.vec=exp(-((phi.mat%*%e.theta)+
  ((x.ordered%*%e.beta)*time.obs)))
  #print((surv.vec))
  surv.obs=surv.vec[xi.order==0]
  surv.rcens=surv.vec[xi.order==3]
  surv.lcens=surv.vec[xi.order==1];
  surv.ilcens=surv.vec[xi.order==2]
  [1:ni.cens];surv.ircens=surv.vec[xi.order==2]
  [(ni.cens+1):(2*ni.cens)]
  likelihood1=sum(log(1-surv.lcens))+
  sum(log(surv.ilcens-surv.ircens))+
    sum(log(psi.obs%*%theta1+x.obs%*%e.beta)+
    log(surv.obs))+sum(log(surv.rcens))
  #likelihood1=sum(log(1-surv.lcens))+
  sum(log(surv.ilcens-surv.ircens))+sum(log(surv.rcens))
  likelihood2=-smooth*t(e.theta)%*%R%*%e.theta-(alpha/2)*
    sum((x.ordered%*%e.beta)-e.eta)^2
  return(likelihood1+likelihood2)
}
```

Bibliography

Aalen, O. (1980), A model for nonparametric regression analysis of counting processes, *in* 'Mathematical statistics and probability theory', Springer, pp. 1–25. (Cited on pages vi, 11, 95, 96, 118, and 132.)

Aalen, O., Borgan, O. & Gjessing, H. (2008), *Survival and event history analysis: a process point of view*, Springer Science & Business Media. (Cited on pages 2, 9, and 12.)

Aalen, O. O. (1989), 'A linear regression model for the analysis of life times.', *Statistics in medicine* **8**(July 1988), 907–925. (Cited on pages 11, 12, 13, 17, 19, 20, 21, and 22.)

Amemiya, T. (1983), 'Non-linear regression models', *Handbook of econometrics* **1**, 333–389. (Cited on page 91.)

Andersen, P. K., Borgan, O., Gill, R. D. & Keiding, N. (2012), *Statistical models* *based on counting processes*, Springer Science & Business Media. (Cited on page 13.)

Anderson, J. A. & Senthilselvan, A. (1980), 'Smooth estimates for the hazard function', *Journal of the Royal Statistical Society. Series B (Methodological)* **42**(3), 322–327.

URL: *http://www.jstor.org/stable/2985166* (Cited on pages 37 and 43.)

Bertsekas, D. P. (1996), *Constrained Optimization and Lagrange Multiplier Methods*. (Cited on pages 16 and 45.)

Bertsekas, D. P. (2014), *Constrained optimization and Lagrange multiplier methods*, Academic press. (Cited on page 34.)

Boyd, S., Parikh, N., Chu, E., Peleato, B. & Eckstein, J. (2011), 'Distributed optimization and statistical learning via the alternating direction method of multipliers', *Foundations and Trends in Machine Learning* **3**(1), 1–122. (Cited on page 46.)

Breslow, N. (1972), 'Discussion of the paper by D. R Cox', *Journal of Royal Statistical Society* **B**(34), 216–217. (Cited on pages 14, 23, and 37.)

Breslow, N. & Day, N. (1987), *Statistical Methods in Cancer Research - The Design and Analysis of Cohort Studies*. (Cited on page 11.)

Cai, T. & Betensky, R. A. (2003), 'Hazard regression for interval-censored data with penalized spline', *Biometrics* **59**(3), 570–579. (Cited on page 109.)

Collett, D. (2003), *Modelling survival data in medical research*, CRC press. (Cited on page 1.)

Cox, D. D. & O'Sullivan, F. (1990), 'Asymptotic analysis of penalized likelihood and related estimators', *The Annals of Statistics* pp. 1676–1695. (Cited on page 37.)

Cox, D. R. (1972), 'Regression models and life-tables', *Journal of the Royal Statistical Society. Series B (Methodological)* **34**(2), 187–220.
URL: *http://www.jstor.org/stable/2985181* (Cited on page 9.)

Cox, D. R. (1975), 'Partial likelihood', *Biometrika* pp. 269–276. (Cited on page 14.)

Crowder, M. (1984), 'On Constrained Maximum Likelihood Estimation with non-i.i.d observations', *Ann. Inst. Statist. Math* **36**(A), 239–249. (Cited on page 85.)

De Boor, C. & Daniel, J. W. (1974), 'Splines with nonnegative B-spline coefficients', *Mathematics of Computation* **28**(126), 565–568. (Cited on page 77.)

Dieudonné, J. (2013), *Foundations of modern analysis*, Read Books Ltd. (Cited on page 79.)

Farrington, C. P. (1996), 'Interval censored survival data: a generalized linear modelling approach', *Statistics in medicine* **15**, 283–292. (Cited on pages 15, 31, and 58.)

Feller, W. (2008), *An introduction to probability theory and its applications*, Vol. 2, John Wiley & Sons. (Cited on page 82.)

Fletcher, R. (1973), 'An exact penalty function for nonlinear programming with inequalities', *Mathematical Programming* **5**(1), 129–150. (Cited on page 46.)

Ghosh, D. (2001), 'Efficiency considerations in the additive hazards model with current status data', *Statistica Neerlandica* **55**(3), 367–376. (Cited on pages 14, 29, and 58.)

Golub, G. H. & Van Loan, C. F. (2012), *Matrix computations*, Vol. 3, JHU Press. (Cited on page 46.)

Good, I. J. & Gaskins, R. A. (1971), 'Nonparametric roughness penalties for probability densities', *Biometrika* pp. 255–277. (Cited on page 37.)

Gorst-Rasmussen, A. & Scheike, T. H. (2012), 'Coordinate Descent Methods for the Penalized Semiparametric Additive Hazards Model', *Journal of Statistical Software* **47**(9). (Cited on page 96.)

Grenander, U. (1981), Abstract inference, Technical report. (Cited on page 76.)

Hestenes, M. R. (1969), 'Multiplier and gradient methods', *Journal of optimization theory and applications* **4**(5), 303–320. (Cited on pages 16 and 45.)

Honoré, B. E. & Powell, J. L. (n.d.), *Journal of Econometrics* . (Cited on page 87.)

Huang, J. (1996), 'Efficient estimation for the proportional hazards model with interval censoring', *The Annals of Mathematical Statistics* **24**(2), 540–568. (Cited on pages 78 and 82.)

Huang, J. & Rossini, A. (1997), 'Sieve estimation for the proportional-odds failure-time regression model with interval censoring', *Journal of the American Statistical Association* **92**(439), 960–967. (Cited on page 75.)

Huffer, F. W. & McKeague, I. W. (1991), 'Weighted least squares estimation for aalen's additive risk model', *Journal of the American Statistical Association* **86**(413), 114–129. (Cited on pages 10, 13, and 21.)

Kalbfleisch, J. D. & Prentice, R. L. (2002), *The Statistical Analysis of Failure Time Data*, John Wiley Sons. (Cited on page 9.)

Klein, J. P. & Moeschberger, M. L. (2003), *Survival analysis: techniques for censored and truncated data*, Springer Science & Business Media. (Cited on page 9.)

Kullback, S. & Leibler, R. A. (1951), 'On information and sufficiency', *The annals of mathematical statistics* **22**(1), 79–86. (Cited on page 79.)

Lawless, J. F. (2002), *Statistical models and methods for lifetime data*, Vol. 362, John Wiley & Sons. (Cited on page 1.)

Lee, E. T. & Wang, J. (2003), *Statistical methods for survival data analysis*, Vol. 476, John Wiley & Sons. (Cited on page 2.)

Leiderman, P. H., Babu, B., Kagia, J., Kraemer, H. C. & Leiderman, G. F. (1973), 'African infant precocity and some social influences during the first year', *Nature* **242**(5395), 247–249. (Cited on page 4.)

Lin, D., Oakes, D. & Ying, Z. (1998), 'Additive hazards regression with current status data', *Biometrika* **85**(2), 289–298. (Cited on pages 14, 24, 25, 26, and 27.)

Lin, D. Y. & Ying, Z. (1995), 'Semiparametric Analysis of General Additive-Multiplicative Hazard Models for Counting Processes', *The Annals of Mathematical Statistics* **23**(5), 1712–1734. (Cited on page 12.)

Lin, D. & Ying, Z. (1994), 'Semiparametric analysis of the additive risk model', *Biometrika* **81**(1), 61–71. (Cited on pages vi, 11, 12, 13, 14, 17, 19, 20, 22, 33, 35, 36, 37, 95, 96, 118, and 132.)

Lin, D. & Ying, Z. (1997), Additive hazards regression models for survival data, *in* 'Proceedings of the first Seattle symposium in Biostatistics', Springer, pp. 185–198. (Cited on page 12.)

Luenberger, D. G., Ye, Y. et al. (1984), *Linear and nonlinear programming*, Vol. 2, Springer. (Cited on page 50.)

Ma, J. (2010), 'Positively constrained multiplicative iterative algorithm for maximum penalized likelihood tomographic reconstruction', *IEEE Transactions on Nuclear Science* **57**(1), 181–192. (Cited on pages 34, 38, 48, and 130.)

Ma, J., Heritier, S. & Lô, S. N. (2014), 'On the maximum penalized likelihood approach for proportional hazard models with right censored survival data', *Computational Statistics & Data Analysis* **74**, 142–156. (Cited on page 38.)

Mandel, M. (2007), 'Censoring and truncation - highlighting the differences', *The American Statistician* **61**(4), 321–324. (Cited on page 5.)

Marsden, M. J. (1968), An identity for spline functions with applications to variation diminishing spline approximation., Technical report, Wisconsin University Madison Mathematics Research Centre. (Cited on page 77.)

McKeague, I. W. & Sasieni, P. D. (1994), 'A partly parametric additive risk model', *Biometrika* **81**(3), 501–514. (Cited on pages 12 and 13.)

Moore, T. J., Sadler, B. M. & Kozick, R. J. (2008), 'Maximum-Likelihood Estimation , the Cramér - Rao Bound, and the Method of Scoring with Parameter Constraints', *IEEE Transactions on Signal Processing* **56**(3), 895–908. (Cited on page 86.)

Nelder, J. (1975), 'Announcement by the working party on statistical computing: GLIM (Generalized Linear Interactive Modelling Program)', *Applied Statistics* pp. 259–261. (Cited on page 15.)

Nelder, J. A. & Mead, R. (1965), 'A simplex method for function minimization', *The Computer Journal* **7**(4), 308–313. (Cited on page 31.)

Nocedal, J. & Wright, S. J. (2006), *Sequential quadratic programming*, Springer. (Cited on page 46.)

Pollard, D. (1984), *Convergence of stochastic processes*, Springer Science & Business Media. (Cited on page 81.)

Pollard, D. (1990), Empirical processes: theory and applications, *in* 'NSF-CBMS regional conference series in probability and statistics', JSTOR, pp. i–86. (Cited on page 81.)

Powell, M. (1969), *A method for nonlinear constraints in minimization problems in Optimization*, Academic Press, London. (Cited on pages 16 and 45.)

Rockafellar, R. T. (1976), 'Augmented lagrangians and applications of the proximal point algorithm in convex programming', *Mathematics of operations research* **1**(2), 97–116. (Cited on page 16.)

Rockafellar, R. T. (2015), *Convex analysis*, Princeton university press. (Cited on page 46.)

Rossini, A. & Tsiatis, A. (1996), 'A semiparametric proportional odds regression model for the analysis of current status data', *Journal of the American Statistical Association* **91**(434), 713–721. (Cited on page 5.)

Saad, Y. (2003), *Iterative methods for sparse linear systems*, Vol. 82, siam. (Cited on page 46.)

Schick, A. & Yu, Q. (2000), 'Consistency of the GMLE with mixed case interval-censored data', *Scandinavian Journal of Statistics* **27**(1), 45–55. (Cited on page 6.)

Shen, X. (1997), 'On methods of sieves and penalization', *The Annals of Statistics* pp. 2555–2591. (Cited on page 75.)

Shen, X. & Wong, W. H. (1994), 'Convergence rate of sieve estimates', *The Annals of Statistics* pp. 580–615. (Cited on pages 75 and 76.)

Silverman, B. W. (1978), 'Density ratios, empirical likelihood and cot death', *Applied Statistics* pp. 26–33. (Cited on page 37.)

Sun, J. (2006), *The statistical analysis of interval-censored failure time data*, Springer Science & Business Media. (Cited on pages 4 and 6.)

Therneau, T. M. & Grambsch, P. M. (2000), *Modeling the Survival Data-Extending the Cox model*, Springer. (Cited on pages 3 and 96.)

Torben, M. & Scheike, T. H. (2006), *Dynamic regression models for survival data*, Springer. (Cited on page 3.)

Van Der Vaart, A. W. & Wellner, J. A. (1996), Weak convergence, *in* 'Weak Convergence and Empirical Processes', Springer, pp. 16–28. (Cited on page 78.)

Wang, L., Sun, J. & Tong, X. (2010), 'Regression analysis of case II interval-censored failure time data with the additive hazards model', *Statistica Sinica* **20**(4), 1709–1723. (Cited on pages 6, 14, 24, 26, 29, 58, and 59.)

Wellner, J. A. (1995), 'Interval censoring, case 2: alternative hypotheses', *Lecture Notes-Monograph Series* **27**, 271–291. (Cited on page 6.)

Wong, W. H. & Shen, X. (1995), 'Probability inequalities for likelihood ratios and convergence rates of sieve MLEs'', *The Annals of Statistics* **23**(2), 339–362. (Cited on page 79.)

Wright, S. J. (1997), *Primal-dual interior-point methods*, SIAM. (Cited on pages 14 and 29.)

Xue, H., Lam, K. F. & Li, G. (2004), 'Sieve Maximum Likelihood Estimator for Semiparametric Regression Models With Current Status Data', *Journal of the American Statistical Association* **99**(466), 346–356. (Cited on pages 78, 81, and 82.)

Yu, Y. & Ruppert, D. (2002), 'Penalized Spline Estimation for Partially Linear Single-Index Models', *Journal of the American Statistical Association* **97**(460), 1042–1054. (Cited on page 85.)

Zangwill, W. I. (1967), 'Non-linear programming via penalty functions', *Management science* **13**(5), 344–358. (Cited on page 46.)

Zeng, D., Cai, J. & Shen, Y. (2006), 'Semiparametric additive risks model for interval-censored data', *Statistica Sinica* pp. 287–302. (Cited on pages 14, 29, 30, 58, and 59.)

Zhang, Y., Hua, L. & Huang, J. (2010), 'A spline-based semiparametric maximum likelihood estimation method for the cox model with interval-censored data', *Scandinavian Journal of Statistics* **37**(2), 338–354. (Cited on pages 78 and 82.)