Delft University of Technology
Faculty of Electrical Engineering, Mathematics and Computer Science
Delft Institute of Applied Mathematics

Bootstrapping in the Cox-model with interval censored observations

A thesis submitted to the
Delft Institute of Applied Mathematics
in partial fulfilment of the requirements

for the degree

MASTER OF SCIENCE
in
APPLIED MATHEMATICS

by

SIGUR GOUWENS

Delft, The Netherlands
December 2019

MSc thesis APPLIED MATHEMATICS

"Bootstrapping in the Cox-model with interval censored observations"

SIGUR GOUWENS

Delft University of Technology

Supervisor

Prof. Dr. Ir. G. Jongbloed

Committee

Dr. H. P. Lopuhaä Dr. Ir. F. J. Vermolen

December, 2019 Delft

Abstract

In this study the interval censoring case 2 model combined with the Cox model is considered. The event-time distribution function is modelled non-parametrically. Two algorithms are proposed to estimate the event-time distribution function together with the Cox coefficients. Kernel smoothing is applied to the non-parametric MLE of the event-time distribution resulting in the smoothed MLE (SMLE). A two-step method for choosing the smoothing bandwidth based on minimising the MSE is introduced. Given the SMLE, the precision of the MLE is tested using bootstrap simulations. New event-times are sampled from the SMLE which are then used to compute bootstrap estimates of the event-time distribution. This is done for multiple sample sizes to observe large sample behaviour. This study suggests that larger sample sizes lead to better estimates. Monte Carlo simulations and the bootstrap simulations agree on the bandwidth and the large sample distribution of pointwise estimates of the event-time distribution.

Contents

Chapter 1

Introduction

Human population size has been documented for centuries. In 1760, Daniel Bernoulli wrote an essay reviewing the effectiveness of inoculation against smallpox, a method for immunisation. Based on data containing individuals ages, the population that is alive and the decrease in population at each age group, Bernoulli stated that the prevention method was advantageous if the procedure itself did not cause too many deaths as inoculation itself had a risk of infecting with the disease. Bernoulli's essay is one of the early articles in which the ideas of what is currently called survival analysis are used.

Since then, the branch of statistics has seen many developments. Applications are not only found in epidemiology but any field where the time elapsed until an event of interest is measured. Such data are called time-to-event data. The name survival analysis suggests its application in medicine. Consider a study on breast cancer, in the Netherlands, women aged between 50 and 75 years old are asked for examination. In a study like this, the event of interest is the starting of the development of breast cancer. During the screening, the patient either has the disease or not. If malignant tissue is discovered during the examination, the exact time of the event is only known to be before the inspection time. If the inspection time is used as the true event time, the estimate of the time of developing breast cancer is overestimated. This issue motivates to consider *censored data*, that is data of which the precise time of event but only known to lie in a time interval. Survival analysis deals with this type of data and is used extensively in medicine. Within actuarial science, the theory of survival analysis is applied to make predictions of the payout. Insurance companies need to estimate the risk of a client to determine the premium.

In survival analysis, important functions are the survival function $S(t) = 1 - F(t)$ describing the probability that an event has not yet occurred at time t. The hazard function $\lambda(t)$ is interpreted as a measure of danger a subject is exposed to at time t. Research on lifetimes of a subject arise in many different fields. A common complication seen in research is that time-to-event data are censored. The time interval in which the event has taken place can be of infinite length. In some cases only a single inspection is done. This often happens when testing is destructive so only a single inspection can be done. An example of this is in animal testing where death of the animal is required for observation. There also exist cases where continuous inspection of the subject is impractical, think for example of medical checkups. Such interval censored data contain less information than exact measurements of the event times. This problem brings complications to the analysis and should be treated carefully.

Considerable research on survival analysis has been done within both the frequentist and the Bayesian paradigm. Multiple articles have been published on Bayesian survival analysis, one of them is (Ibrahim, Chen, & Sinha, 2014). However, the majority of the literature studies survival analysis in a frequentist paradigm. Recently, more articles have been published using nonparametric regression to estimate hazard functions rather than parametric regression. This type of inference does not assume the target distribution belongs to a parametric family. It attempts to fit data in a larger function space, adding more flexibility to the model. This can be useful when the underlying parametric model is unknown or when the event distribution is not restricted to a parametric family. The likelihood function has been derived for interval censored data. Maximising the likelihood function where the input is a function that does not belong to a parametric family leads to a nonparametric optimisation problem. In short, this means that the nonparametric maximum likelihood estimator (NPMLE) lets the data speak for themselves and does not depend on strong restrictions. Non-parametric estimation does not solely depend on parametrised families but it can make use of parametric distributions in some cases. Different algorithms exist for computing the NPMLE for interval censored data, see for example (Groeneboom, Jongbloed, & Wellner, 2008), (Groeneboom & Jongbloed, 2014) and (Dempster, Laird, & Rubin, 1977).

In this study it will be seen that the NPMLE is a step function even though the true event-time distribution is continuous. In (Banerjee & Wellner, 2005), kernel smoothing has been applied to the NPMLE to produce a continuous estimator by smearing out the jumps using a kernel function. Because the survival function is defined on the positive real axis inconsistency can occur near the boundary at $t = 0$ as kernel density estimation adds mass on the negative time axis. A method to fix inconsistency around boundaries of the domain is studied in (Groeneboom & Jongbloed, 2015). It is natural to question the accuracy of this smoothed NPMLE (SMLE). (Banerjee & Wellner, 2005) proposed point-wise confidence intervals for $F(t_0)$ using subsampling techniques. More research is done in (Sen & Xu, 2015) where a simulation study for a the mixed case interval censoring model is done and proved inconsistency for bootstrap confidence intervals using the NPMLE.

The event time distribution in survival analysis can differ per subject. Other covariates can play an important role. The lifetime of a machine can for example be increased by using a material with a higher hardness, then, hardness is a covariate. A medical example would be to measure the blood pressure of patients when doing research on cardiovascular diseases. The Cox model is a way to incorporate covariates into the analysis. In this model, each covariate is weighted which need estimation together with the cumulative hazard function. The hazard function within the Cox model is defined as $\Lambda(t|s) = \Lambda_0(t)\exp(\beta^\top s)$ where Λ_0 is the baseline, which is equal for each subject. Covariates s are weighted by the vector of Cox coefficients β. Both the baseline hazard and β can be estimated using the the maximum likelihood estimator. The optimisation is in the cross product of a function space and a Euclidean space. This means that there is a parametric as well as a nonparametric component. Such models are called semiparametric. The amount of literature on the Cox model combines with interval censored models is limited. A study is done in (Pan, 1999) on the Cox model using current status data. Asymptotic results on convergence of both the parametric part and nonparametric parts in semiparametric regression are proven in (Ma & Kosorok, 2005). This article works under technical assumptions and proves a rate $\sqrt{n}$-consistency and asymptotic normality is proven for weights β in the current status model. Moreover, (Groeneboom & Wellner, 1992) proved an optimal convergence rate $n^{1/3}$ for $\hat{\Lambda}_n$ not using the Cox model. Ma and Kosorok state that one cannot improve on this rate in their article. Results for the interval censoring case 2 using the Cox model have not been found.

The aim of this study is to be able to generate confidence intervals for $F(t_0)$ in the Cox model where $t_0 > 0$ for interval case 2 censored data. Furthermore, large sample behaviour is observed of both the NPMLE $\hat{F}_n$ and the estimator of the Cox coefficients. Let $(\hat{\beta}_n, \hat{\Lambda}_{0n})$ be the NPMLE for (β, Λ_0), the true Cox coefficients and the baseline cumulative hazard function. Estimators are said to be (weakly) consistent if $(\hat{\beta}_n, \hat{\Lambda}_{0n}) \to (\beta, \Lambda_0)$ in probability as $n \to \infty$. In the semiparametric model, the cumulative hazard function of a subject depends on a function Λ_0 and parameters β in a d-dimensional Euclidean space. Monte Carlo simulations are done to see if one can speak of consistency of β and Λ_0. Multiple sample sizes are used to check whether the Monte Carlo estimates become closer to the true β and Λ_0 the baseline cumulative hazard function. Because in many cases, the target function Λ_0 is expected to be continuous, kernel smoothing is applied to find a smoothed NPMLE. A bandwidth parameter becomes part of the estimator. In order to obtain a good estimator, finding a good bandwidth parameter is essential. Because in practice only a single sample is given, the bootstrap is a useful tool for estimating the bandwidth. (Sen & Xu, 2015) proved that the bootstrapped NPMLE yields inconsistent results. This motivates to use smoothing for estimation of the bandwidth. However, in (Groeneboom & Hendrickx, 2017) the opposite is proven and it is stated that the bootstrap method is consistent.

A method is proposed to select a good bandwidth in terms of having a low mean squared error. For different sample sizes, this method is applied to select bandwidths. These bandwidths can be used to compute statistics of interest. A bootstrap study is done on large sample behaviour of bootstrap samples. For the sample large sample sizes, results are also obtained using Monte Carlo, sampling from the true distributions. The Monte Carlo and bootstrap results are then compared. Large sample behaviour is also checked to see the effect that the sample size has on estimators in terms of precision. Furthermore, confidence intervals are constructed and the length of the interval will be observed compared to the used sample size.

1.1 Thesis Outline

An overview of the thesis is given in order to see the connections between the topics that are treated. First general theory on survival analysis is given. In this chapter it is also explained what interval censored data means. The next chapter gives understanding on how the likelihood function is derived and what properties of this function are stated. Next, the Cox proportional hazards model is introduced allowing to add more variables to the analysis. The newly introduced variables are then incorporated into the previously derived likelihood function. The maximiser of this function is called the maximum likelihood estimator (MLE). Algorithms are introduced to find the MLE. Because the output turns out to be a step function while the underlying distribution function is continuous, smoothing is applied to the MLE. To check for the accuracy of the estimator, bootstrap procedures are used. Figure 1.1 shows a flowchart of the order of the chapters of this thesis and their connections.

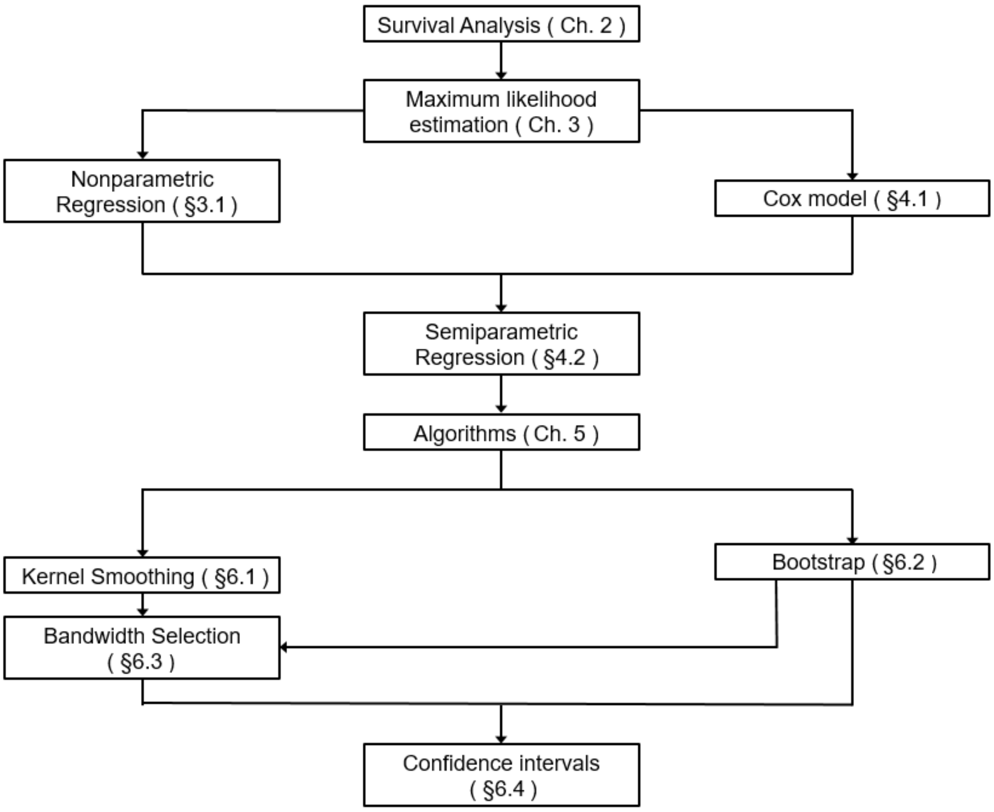
Survival Analysis (Ch. 2)
Maximum likelihood estimation (Ch. 3)
Nonparametric Regression (§3.1)
Cox model (§4.1)
Semiparametric Regression (§4.2)
Algorithms (Ch. 5)
Kernel Smoothing (§6.1)
Bootstrap (§6.2)
Bandwidth Selection (§6.3)
Confidence intervals (§6.4)

Figure 1.1: This flowchart gives an overview of the topics treated in the thesis and the connections between them.

Chapter 2

Survival Analysis and Data Censoring

2.1 Survival Analysis

In survival analysis, “time-to-event” data are analysed. That is, one tries to extract information from data about the time elapsed until a specific event happens. For a long time life tables have been used to make estimations of peoples survival times, see for example (Cox, 1972). The theory has found applications in many other fields such as engineering, actuarial science, biology and many other fields where time-to-event data play a role. Using survival analysis, many things can be quantified, e.g., how well a new medical treatment performs or whether the lifetime of a machine improves when using a more durable material. This quantification is in a measure called *hazard*. As the name suggests, the higher the hazard, the more likely the machine is to fail. Important functions in survival analysis are the survival function and the hazard function, which turn out to be closely connected.

Let the time at which an event occurs be denoted by the continuous random variable X taking values in $[0, \infty]$. This random variable describes the event-times and is distributed according to the distribution function F having density f. The survival function at time $t > 0$ is the probability that an event will not occur until time t and is denoted by

$$S(t) = 1 - F(t) = \mathbb{P}(X > t).$$

From the frequentist interpretation of probability, the survival function evaluated at time t is seen as the proportion of the subjects for which the event has not occurred before time t when infinitely many samples are taken. Another quantity which turns out to be in one-to-one correspondence to the survival function is the hazard function, λ. This quantity is defined by

$$\lambda(t) = \lim_{dt \downarrow 0} \frac{\mathbb{P}(X \in [t, t + dt] | X \geq t)}{dt}. \tag{2.1}$$

The numerator is the probability the event happens in the interval $[t, t+dt]$ given it had not occurred before. This definition is not valid if X is a discrete random variable. The hazard function of a discrete distribution is defined as the conditional probability $\mathbb{P}(X = t | X \geq t)$, however, in this study only continuous distributions are treated. Consider a medical situation. The probability in (2.1) can be seen as the fraction of the population that is still alive at time t that dies during the time interval $[t, t + dt]$. Dividing this quantity by the duration of the time interval under consideration results in a rate. The hazard function can be understood as a measure of danger at time t. Integrating the hazard function accumulates the amount of hazard a subject has been exposed to until time t. Assuming that F is differentiable in a neighbourhood of t, the definition of the hazard function can be expressed differently by observing that $\{X \in [t, t+dt]\} \subset \{X \geq t\}$, so

$$\frac{\mathbb{P}(X \in [t, t+dt] | X \geq t)}{dt} = \frac{\mathbb{P}(X \in [t, t+dt])}{dt \cdot \mathbb{P}(X \geq t)} = \frac{1}{S(t)} \frac{F(t+dt) - F(t)}{dt} \overset{dt \to 0}{\rightarrow} \frac{f(t)}{S(t)}, \tag{2.2}$$

by the definition of the derivative. The chain rule motivates to express the survival function in a more elegant way once it is observed that the numerator is proportional to the derivative of the denominator:

$$\lambda(t) = \frac{f(t)}{S(t)} = -\frac{(1 - F(t))'}{1 - F(t)} = -\frac{d}{dt} \log S(t). \tag{2.3}$$

Integrating this quantity results in the cumulative hazard function, a measure of "how much risk there has been up to time t", using that $S(0) = 1$ this gives

$$\Lambda(t) = \int_0^t \lambda(s) ds = -\int_0^t \frac{d}{ds} \log S(s) ds = -\log S(t) = -\log\big(1 - F(t)\big). \tag{2.4}$$

By convention, when $F(t) = 1$, then $\Lambda(t) = \infty$. Likewise, it can be formulated inversely as

$$F(t) = 1 - S(t) = 1 - \exp(-\Lambda(t)). \tag{2.5}$$

Within survival analysis, the family of Weibull distributions is frequently used. This makes it a good example to show what a cumulative hazard function looks like. The distribution function is

$$F_{\rho,k}(t) = 1 - e^{-(\rho t)^k}, \tag{2.6}$$

where $\rho, k > 0$ are the shape and scale parameters respectively. Then the cumulative hazard function is computed using (2.4),

$$\Lambda_{\rho,k}(t) = -\log(1 - 1 + \exp(-(\rho t)^k)) = (\rho t)^k. \tag{2.7}$$

Taking $k = 1$ in (2.6) results in the exponential distribution. From (2.7) it is seen that the cumulative hazard function is a linear function $\Lambda_{\rho,1}(t) = \rho t$, making the hazard function $\lambda_{\rho,1}$ constant with respect to time.

The hazard rate itself does not have much meaning. When integrated it does have an interpretation. For a practical understanding, the idea of (Cleves, Gould, & Marchenko Yulia, 2002) is followed. For the so-called count-data interpretation of the cumulative hazard function, the underlying event-time process is observed. Consider a machine with a bearing. When the bearing breaks, it is replaced by a working one that is as old as the broken one. The time to the next replacement is needed is measured. Let N_t be the process counting the number of replacements needed during the time interval $(0, t)$. Then the cumulative hazard function at t can be interpreted as $\mathbb{E}(N_t) = \Lambda(t)$. The failure time of the bearing can be modelled by a Weibull distribution where the time t is given in months. The hazard function is found in (2.6), let the parameters for the Weibull distribution be $\rho = \frac{1}{80}$ and $k = 7$. This could be a model for failure times of bearings. Then at 80 months, $\Lambda_{\rho,k}(80) = 1$ the expected number of replacements is 1. A bearing that has lasted for 100 months has been exposed more risk as $\Lambda_{\rho,k}(100) \approx 4.7$. This translates to the bearing on average has been replaced 4.7 times if it was replaced at failure by an equally old working one. A code to simulate how many replacements are needed before 80 months in `R` together with a simulation are found in the Appendix in 8.0.3.

Research on nonparametric U-shaped hazard functions λ is done by (Wang & Fani, 2017). This type of function is also called *bathtub shaped*. At time $t = 0$ the function value is high, then it decreases to a minimum before becoming increasing. Why this type of function may appear as a hazard function can be exemplified the following. A newborn is prone to many medical complications such as birth defects and lower immune system which may be fatal. This translates to a relatively high value of the cumulative hazard function for small values of t. After infancy the hazard decreases the child becomes stronger, then, as it becomes older, the probability of death starts increasing with age, for example due to elderly diseases. This results in a U-shaped curve. This shape is also studied parametrically in (El-Gohary, Alshamrani, & Al-Otaibi, 2013) where the Gompertz distribution is fitted, of which hazard function is U-shaped.

A quantity often used in clinical trials is the hazard ratio. Given two sub-populations having hazard ratios λ_1 and λ_2. The hazard ratio at time t is defined the following $\mathrm{HR}(t) = \frac{\lambda_1(t)}{\lambda_2(t)}$. A link with the count-data interpretation can be drawn. Given t, a hazard ratio of 2 indicates that the expected number of events that occurred is twice as high for the first subject compared to the second.

2.2 Incomplete Data and Censoring

Data is a fundamental part in statistics and should be used correctly. The question is how to use it properly. There are multiple possible issues the data can be subject to. When collecting data, the data obtained can be biased, corrupted or duplicated or may have other issues. One specific type of problem that is often seen, within survival analysis is censoring of data. If not addressed correctly, all these complications can lead to poor conclusions making it crucial to handle them correctly. In the following paragraphs these types of censoring are explained and will be motivated by practical examples.

2.3 Right Censoring and Left Censoring

In the right censoring model, it is assumed that when the event had occurred, the precise event-time is known. If it had not during the study, then it may or may not happen in the future, that means it happens at either a finite t or, by convention, at $t = \infty$. When this happens is unknown. In the last case, the event of the subject is censored on the right side of the time axis, i.e., it is only known that the event occurred in the time interval $[t, \infty]$. The following example illustrates this. Figure 2.1 shows this data visually.

Example 1. *Right censoring can be illustrated with strokes. When a person experiences one, it is (usually) known when this had happened making it an exact measurement. If the person being observed, has not had a heart attack or stroke, then if may happen in the future, making the observation right censored.*

Similar to the above, there is left censoring. In this case, the data are observed as follows: A part of the event times is known exact and another part is known only to lie in an interval in the past. The occurred event of the ith subject is only known to lie in the interval $[0, t_i]$.

Example 2. *Consider the Epstein-Barr virus of the herpes family. At young ages, infections usually cause no symptoms or are indistinguishable from the common mild diseases. It is estimated that 90 percent of adults in the US are infected, so it is assumed in this example that eventually, everyone is infected. Because no unusual symptoms occur, no time of infection can be indicated. If presence of the virus is measured, the only thing known is that the time of infection lies between birth and the current age. Such data is left censored.*

The following paragraph elaborates another type of censoring which will be the main focus of this study.

2.4 Current Status Model and Interval Censoring Case 2

Right and left censoring are not the only type of censoring mechanisms. A natural extension to the previously mentioned types of censoring is where the event-times are never exact. In this situation, the event either happened before the inspection time or after. The model described here only

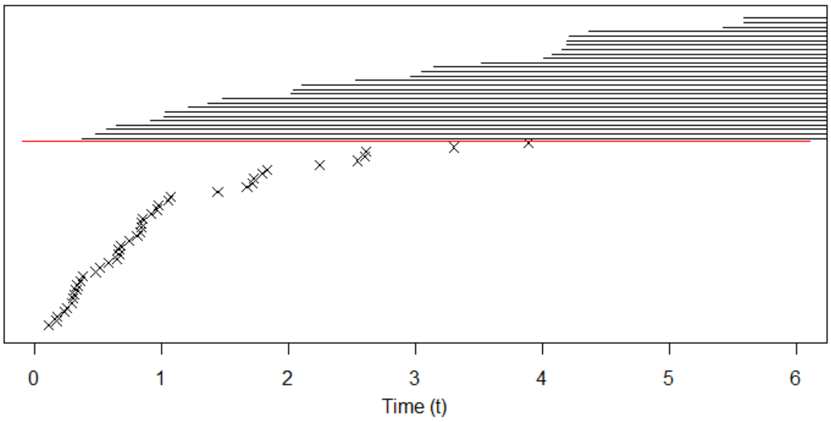

Figure 2.1: Data in the right censoring model consists partly of exact observations and the rest is censored, i.e., the event had not happened before the time of measurement. The red line separates the exact observations (crosses) from the intervals of the censored data (lines).

measures the state of the subject at the current time, giving its name *current status model* but also goes under the name of interval censoring case 1. Let again $X \sim F$ be a continuous random variable describing the event-times. Application of this model often happens when observations on the event requires destructive testing, for example in animal testing. For each subject i, X_i is unknown. The only thing known is the first and only inspection time $T_{1,i}$ and the fact whether X_i had taken place before $T_{1,i}$ or may have happened after the inspection. Here it can be checked whether the onset of a disease has already taken place or not at a certain time where inspection means death of the animal. Another type of censoring, on which this study is focussed is the interval censoring type 2 (IC2). The difference here is that there are two moments of inspection whereas the current status model has only one. This can be applied when inspections cannot be done continuously and it is possible to do them multiple times. It is possible that the event has occurred before the first inspection, after the last, or in between the two.

Example 3. *In medical studies one could be interested in a survival time after a given treatment for a certain disease. There may be risk of the disease to return. Because of this, two check-ups are planned. The first check-up is after 3 years after the treatment, and another one 6 years later. If the patient appears to be ill again during the first check-up, then the only thing known is that this happened at some point during the first three years after the treatment. It is also possible that signs of the diseases have returned between the first and second check-up. Another possibility is that no signs have returned until the second check-up, then it may or may not happen in the future.*

Throughout the study, the focus lies on case 2 interval censoring. For an easier discussion, by left censored it is meant that the event happened before the first inspection time and middle censored when the event happened between the two inspection times. The last possibility is called right censored, not to be confused with the type of censoring.

Before starting the analysis, the concept of interval censoring needs translation to the language of mathematics. Starting with the general mixed case interval censoring model. In this model, for each subject the event-time is interval censored with a different number of inspections. Denote the random vector containing K_i inspection times of the i'th subject by $T_i = (T_{1,i}, \dots, T_{K_i,i}) \sim H$ such that almost surely $T_{k_1,i} < T_{k_2,i}$ when $k_1 < k_2$. A very general form would be to take K_i to be a positive integer valued random variable for each subject. An indicator is defined for every interval of the form $(T_{j,i}, T_{j+1,i}]$ for $j = 0, \dots K_i + 1$. By convention $T_{0,i} = 0$ and $T_{K_i+1,i} = \infty$. Define the indicator $\Delta_{j,i} = 1_{X_i \in (T_{j,i}, T_{j+1,i}]}$ which is one if the event occurred in the given interval and zero if

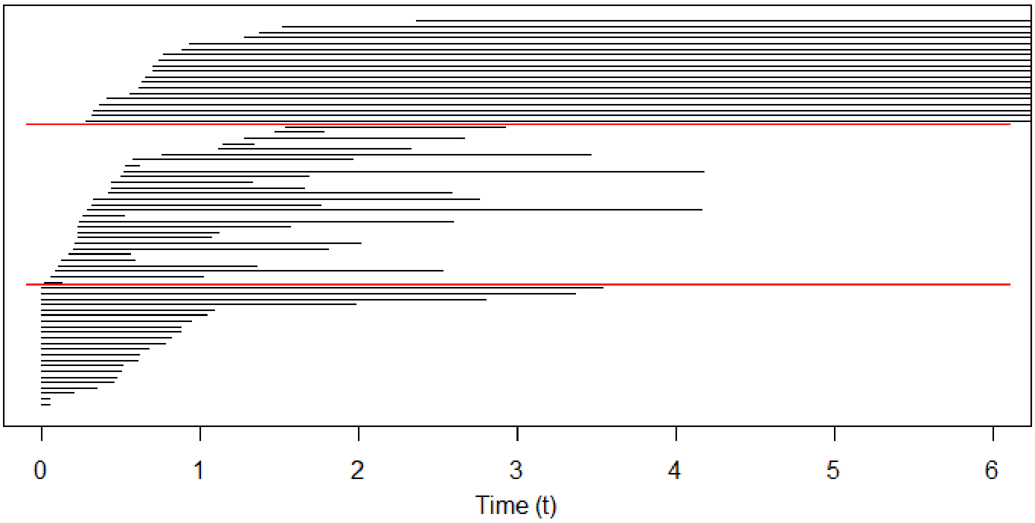

Figure 2.2: A visualisation of type 2 interval censored data. The red lines separate the events of $\delta_i = 1$ (bottom), $\gamma_i = 1$ (middle) and the events with $\mu_i = 1$ that have not occurred yet (top). Each line represents the time-interval in which the event had taken place.

this is not the case.

For the current status model, $K_i = 1$ for all i and the event X_i happened either before time $T_{1,i}$ or after. Only a single indicator variable is used to represent in which interval the event X_i occurred. This interval is either in $[0, T_{1,i}]$ or in $(T_{1,i}, \infty]$.

One speaks of the interval censoring case 2 model if $K_i = 2$ for each subject i. In that case one speaks of interval censoring case 2 (IC2). This model has for each subject two inspections and thus three intervals of the form $(T_{j,i}, T_{j+1,i}]$ in which the event can happen. This model will be of main interest during the thesis so here, special notation for the variables are therefore introduced. Instead of $T_{1,i}$ and $T_{2,i}$ the two inspection times are denoted by $(T, U) \sim H$ such that $T < U$ almost surely. In this study, the distribution of inspection times H is assumed to be continuous implying that, with probability one, ties in inspection times do not occur. The indicators variables are defined as

$$\Delta = 1_{X \in (0,T]} \quad \Gamma = 1_{X \in (T,U]} \quad M = 1 - \Delta - \Gamma = 1_{X \in (U,\infty]}. \tag{2.8}$$

The indicators in (2.8) are linked to the maximum likelihood estimator in the upcoming chapter. To some extent, the current status model can be seen as a special case of the interval case 2 model. Suppose the distribution of the second inspection times U returns high values, so that $\mathbb{P}(M = 1)$ has very low probability, then the observed event i happens only before T_i or in the interval T_i, U_i. If U is approximately infinite almost surely, that is $\mathbb{P}(X_i \in (T_i, U_i)) \approx 1 - \mathbb{P}(X_i \in [0, T_i])$, then IC2 reduces to the current status model. A visualisation of this type of data is presented in Figure 2.2.

Chapter 3

The Maximum Likelihood Estimator

Various approaches to estimating F based on censored data exist. One of the most frequently used ones is the Maximum Likelihood Estimator (MLE). The philosophy behind it is to estimate a parameter of function that is most likely according to the observations. The MLE is found by maximising the likelihood function, a function that is proportional to the probability of obtaining the data. To put this in more mathematical terms, let $x_1, \dots, x_n$ be event times from $X \sim F$ and assume that the information available is: $x_i \in I_i$ for $1 \leq i \leq n$. For estimating F first the probability of the data is written down and then maximised for the parameter given the data. The likelihood function is defined as a function of F by

$$L(F) := \mathbb{P}(X_i \in I_i \text{ for all } i = 1, \dots, n). \tag{3.1}$$

The nature of I_i is dependent on context. L is called the likelihood function. Maximising this probability, or more commonly, the logarithm of this probability as a function of F yields the MLE. From a set of distribution functions, the distribution function F is chosen for which the sample X is the most likely. These sets can be further specified if more is known about the data generating mechanism. Independence of samples is an assumption used frequently when estimating parameters or functions. Suppose F_θ with continuously differentiable density f_θ belongs to a parametric family and one wishes to estimate a parameter θ of a parametric distribution using exact independent samples $x_1, \dots, x_n$, by proportionality and setting $I_i = (x_i - \delta/2, x_i + \delta/2)$ for $\delta > 0$, then the likelihood function becomes

$$\begin{aligned} L(\theta|x) &= \delta^{-n}\mathbb{P}(X_i \in B(x_i, \delta/2) \text{ for all } i = 1, \dots, n) = \prod_{i=1}^{n} \delta^{-1}\mathbb{P}(X_i \in I_i) \\ &= \prod_{i=1}^{n} \delta^{-1}\big(F_\theta(x_i + \delta/2) - F_\theta(x_i - \delta/2)\big) \overset{(\dagger)}{=} \prod_{i=1}^{n} \left[f_\theta(x_i) + O(\delta)\right] \overset{\delta \to 0}{\to} \prod_{i=1}^{n} f_\theta(x_i)). \end{aligned} \tag{3.2}$$

It is used in (†) that $F(x \pm \delta/2) = F(x) \pm \frac{1}{2}f(x)(\delta/2) + O(\delta^2)$. Then the maximum likelihood estimator for θ can be found by maximising

$$\hat{\theta} \in \underset{\theta \in \Theta}{\operatorname{argmax}} \prod_{i=1}^{n} f_\theta(x_i)),$$

if it exists. The notation $\operatorname{argmax}_{x \in D} f(x)$ outputs the argument $x' \in D$ so that it attains the maximum of f, that is, $f\left(\operatorname{argmax}_{x \in D} f(x)\right) = \max_{x \in D} f(x)$. Twice differentiability of the distribution function F is required for this estimator due to the application of Taylor's theorem in (3.2). Not only parametric distributions can be estimated using the likelihood function. Rather than estimating parameters of a parametric distribution, one can drop the parametric assumption and maximise the likelihood in a more general function space. Estimation becomes more complicated but a broader range of distribution functions can be approximated using so called nonparametric inference.

3.1 Estimating Functions with Censored Data

In the context of recovering a survival function from data, equation (3.1) can be used again. Let the sets I_i denote the intervals in which an event happened for the i'th subject. More formally, let $X \sim F$, a random variable that is not directly observable. Define the random variables $(T, U) \sim H$ such that $T < U$ almost surely. The variables T and U denote two inspection times which are done subsequently and are assumed to be independent of X. Now, variables indicating in which interval event X had happened can be defined. The event occurred before the first inspection t, between the inspection times or after u.

Let $x_1, \dots, x_n$ be n independent draws from F with inspection times $(t_1, u_1), \dots (t_n, u_n)$. By definition, this gives rise to $(\delta_1, \dots, \delta_n)$ and $(\gamma_1, \dots, \gamma_n)$. In case 2 censoring the sets I_i have the bounds t_i, u_i and ±∞, where t_i and u_i are the first and second inspection times of the subject i respectively. Then

$$I_i = \begin{cases} (-\infty, t_i], & \text{if} \quad \delta_i = 1 \\ (t_i, u_i], & \text{if} \quad \gamma_i = 1 \\ (u_i, \infty), & \text{if} \quad \mu_i = 1 \end{cases} \tag{3.3}$$

From data censored this way, a distribution function is estimated. In survival analysis one is interested in the event time which is positive. Therefore, the set of interest is the set of distributions having mass only on the positive time axis

$$\mathcal{F} = \{F : F \text{ is a distribution function, } F(t) = 0 \text{ if } t \leq 0\}$$

This set is refined when more is known about the likelihood function.

Let $X_1, \dots, X_n$ be independent identically distributed exact samples from F. A good nonparametric estimator for F based on exact data is the empirical distribution function (ECDF), which is given by

$$\check{F}_n(t) = \frac{1}{n} \sum_{i=1}^{n} 1_{X_i < t} \tag{3.4}$$

This form can be motivated by the fact that the probability of an event can be viewed as its limiting relative frequency, that is, the number of times the event occurred divided by the number of trials. The expression for $\check{F}_n(t)$ is the relative frequency of the event $\{X < t\}$. The Glivenko-Cantelli theorem states that the ECDF converge uniformly to the true distribution, that is $\sup_{t>0} |\check{F}_n(t) - F(t)| \to 0$ as $n \to \infty$, see (Wasserman, 2006). The ECDF is however a rather simple example of a nonparametric estimator, but it gives some insight in how estimation is possible without using parametric families. When one has censored data instead of exact data, the ECDF cannot be applied. In the next section a method for treating such data is shown.

3.2 Derivation of the Likelihood Function

In case 2 censoring there are two inspection time variables and two indicators which imply in which interval the event happens. Each observation is an independent draw of the random variable

$$Z = (T, U, \Delta, \Gamma) = (T_i, U_i, 1_{X_i \leq T_i}, 1_{T_i < X_i \leq U_i})_{i=1}^{n}. \tag{3.5}$$

Deriving the density function of Z is done by first obtaining the density function. In order to calculate this, it is useful to split into cases $\delta_i = 0$ or $\delta_i = 1$ and $\gamma_i = 0$ or $\gamma_i = 1$ and calculate separately. The probability of the event $\{T \leq t, U \leq u, 1, 1\}$ is zero because the inspection time cannot be an element of the intersection of two disjoint sets. Rewriting the event corresponding to

T, U, Δ and Γ in terms of X, T and U gives

$$\begin{aligned} Z = \{T \leq t, U \leq u, \Delta = 0, \Gamma = 1\} &= \{T \leq t, U \leq u, 1_{X \leq T} = 0, 1_{t \leq X \leq u} = 1\} \\ &= \{T \leq t, U \leq u, t < X \leq u\} \\ &= \{T \leq t, U \leq u, T < X, X \leq U\}. \end{aligned}$$

Using the fact that X is independent of (T, U), the probability of this event under joint probability measure $\mathbb{P}_{X,H}$ is

$$\begin{aligned} \mathbb{P}_{X,H}(T \leq t, U \leq u, \Delta = 0, \Gamma = 1) &= \int_0^t \int_0^u \int_t^u f(\bar{x}) h(\bar{t}, \bar{u}) d\bar{x} d\bar{u} d\bar{t} \\ &= \int_0^t \int_t^u (F(u) - F(t)) h(\bar{t}, \bar{u}) d\bar{u} d\bar{t}. \end{aligned}$$

Differentiating the last expression with respect to t and u respectively yields the density given $\Gamma = 1$ and is

$$g(t, u, \delta = 0, \gamma = 1) = (F(u) - F(t)) h(t, u).$$

In a similar fashion the probabilities of the events $\{T \leq t, U \leq u, \Delta = 1, \Gamma = 0\}$ and $\{T \leq t, U \leq u, \Delta = 0, \Gamma = 0\}$ can be computed. These computations show that

$$\begin{aligned} \mathbb{P}(T \leq t, U \leq u, \Delta = 1, \Gamma = 0) &= \int_0^t \int_0^u F(t) h(\bar{t}, \bar{u})) d\bar{u} d\bar{t}, \\ \mathbb{P}(T \leq t, U \leq u, \Delta = 0, \Gamma = 0) &= \int_0^t \int_0^u (1 - F(u)) h(\bar{t}, \bar{u})) d\bar{u} d\bar{t}. \end{aligned}$$

Again, by differentiation, the density for Z is given by

$$\begin{aligned} g(t, u, \delta = 1, \gamma = 0) &= F(t) h(t, u) \\ g(t, u, \delta = 0, \gamma = 1) &= (F(u) - F(t)) h(t, u) \\ g(t, u, \delta = 0, \gamma = 0) &= (1 - F(u)) h(t, u) \end{aligned}$$

Using that any non-zero number to the power zero equals one, the three densities can be written in a single expression which will be the key ingredient for the likelihood function

$$g(t, u, \delta, \gamma) = F(t)^{\delta} (F(u) - F(t))^{\gamma} (1 - F(t))^{1 - \delta - \gamma} h(t, u), \tag{3.6}$$

where $0 < t < u, \delta, \gamma \in \{0, 1\}, \delta + \gamma \leq 1$. This is the density of Z defined in (3.5).

There exists a σ-finite measure κ such that for all measurable sets A it holds that if $\kappa(A) = 0$, then $\mathbb{P}(A) = 0$. Let ξ be a measure having finite mass at the points $\{0, 1\}$ and let ψ be the Lebesgue measure on $[0, \infty)$. Define the product measure $\kappa = \psi \times \psi \times \xi \times \xi$ on $[0, \infty)^2 \times \{0, 1\}^2$. It is clear that κ is σ-finite. For this measure, define that for $A = [a_1, b_1] \times [a_2, b_2] \times \{0\} \times \{0, 1\}$, then

$$\int_A d\mathbb{P} = \int_A g(t, u, \delta, \gamma) d(\psi \times \psi \times \xi \times \xi) = \int_{a_1}^{b_1} \int_{a_2}^{b_2} \sum_{\delta=0}^{0} \sum_{\gamma=0}^{1} g(t, u, \delta, \gamma) dt du.$$

The sums arise from the discreteness of the measures ξ. Applying independence of the samples, the log-likelihood function basic on a sample of size n becomes

$$L(F) = \prod_{i=1}^{n} F(t_i)^{\delta_i} (F(u_i) - F(t_i))^{\gamma_i} (1 - F(t_i))^{1 - \delta_i - \gamma_i} h(t_i, u_i). \tag{3.7}$$

The nonparametric maximum likelihood estimator (NPMLE) is a function $\hat{F}_n$ maximising L over the set of distribution functions $\mathcal{F}$. In the maximisation with respect to F, the values of h play no role, the function h can therefore be dropped. Taking the logarithm of L the log-likelihood becomes

$$\mathcal{L}(F) = \sum_{i=1}^{n} \delta_i \log F(t_i) + \gamma_i \log(F(u_i) - F(t_i)) + \mu_i \log(1 - F(t_i)), \tag{3.8}$$

where $\mu_i = 1 - \delta_i - \gamma_i$. Note that the NPMLE must satisfy certain conditions. It must be a distribution function. that is only non-zero on the positive real line. If there exists a maximiser, it is given by

$$\hat{F}_n = \operatorname*{argmax}_{F \in \mathcal{F}} \mathcal{L}(F) = \operatorname*{argmax}_{F \in \mathcal{F}} L(F),$$

The space of all distribution functions could be too large. The key to optimise $\mathcal{L}$ is to appropriately reduce the size of the function space $\mathcal{F}$. The likelihood (3.8) only depends on the values of F at the inspection times t_i and u_i. Its behaviour between these inspection times will not change the output of the likelihood. Because the MLE is a distribution, the function may not be decreasing. Therefore, restrictions to the class of piecewise constant functions is appropriate. Define the set of unique ordered times

$$V = \{t_1, u_1, \ldots, t_n, u_n, t_\infty\} = \{v_1, \ldots v_m, \text{ such that } v_1 < \ldots, < v_m < v_\infty\}.$$

Here v_∞ is a large value which is defined later. By defining V this way, the inspection times u_i and v_i are no longer connected. In proofs and implementation it is essential to keep them tied despite the new ordering of V, for example for computing the second term in (3.8). To solve this issue, a link function ℓ connecting the left and right inspection times is essential for further analysis. Based on this set V, function class in which $\mathcal{L}$ will be optimised becomes the class of piecewise constant distribution functions with possible jumps at V, defined as

$$\mathcal{F} := \left\{ F : F(t) = \sum_{i=1}^{m} \omega_i \mathbb{1}_{(v_i \le t)}, \sum_{i=1}^{m} \omega_i = 1, \text{ with } \omega_i \in [0,1] \text{ for } 1 \le i \le m \right\}. \tag{3.9}$$

Collecting indices of V of left censored data, that is, all data points with $\delta_i = 1$ makes the set $\mathcal{I}_1$. Similarly the indices that correspond to the events that occurred between the inspection times form $\mathcal{I}_2$ and the same is done for the right censored events. To finish this, the set $\mathcal{I}_2$ is subdivided in sets $\mathcal{I}_{2a}$ and $\mathcal{I}_{2b}$ such that they correspond with the left and right bounds of the interval respectively. More formally,

$$\begin{aligned} \mathcal{I}_1 &= \{j \in \{1, \ldots, m\} : v_j = t_i \text{ for some } i \text{ in } \{1, \ldots, n\}, \text{ such that } \delta_i = 1\}, \\ \mathcal{I}_2 &= \{j \in \{1, \ldots, m\} : v_j = t_i \text{ or } v_j = u_i \text{ for some } i \in \{1, \ldots, n\}, \text{ such that } \gamma_i = 1\}, \\ \mathcal{I}_3 &= \{j \in \{1, \ldots, m\} : v_j = u_i \text{ for some } i \in \{1, \ldots, n\}, \text{ such that } \mu_i = 1\}. \end{aligned} \tag{3.10}$$

And divide the set $\mathcal{I}_2$ into $\mathcal{I}_{2a}$ and $\mathcal{I}_{2b}$ as mentioned above,

$$\begin{aligned} \mathcal{I}_{2a} &= \{j \in I_2 : v_j = t_i \text{ for some } i\} \\ \mathcal{I}_{2b}, &= \{j \in I_2 : v_j = u_i \text{ for some } i\}. \end{aligned} \tag{3.11}$$

Define the link function $\ell : \mathcal{I}_{2a} \to \mathcal{I}_{2b}$ that links the left inspection times in V with the corresponding right inspection times. This means that if $v_j = t_i \in \mathcal{I}_{2a}$ implies $v_{\ell(j)} = u_i \in \mathcal{I}_{2b}$ and inversely, $v_j = u_i$ implies $v_{\ell^{-1}(j)} = t_i$. A property following from monotonicity of V is that $v_{\ell(i)} \ge v_i$. With this, equation (3.8) becomes

$$\mathcal{L}(F) = \underbrace{\sum_{i \in \mathcal{I}_1} \delta_i \log F(v_i)}_{\text{(I)}} + \underbrace{\sum_{i \in \mathcal{I}_{2a}} \gamma_i \log(F(v_{\ell(i)}) - F(v_i))}_{\text{(II)}} + \underbrace{\sum_{i \in \mathcal{I}_3} \mu_i \log(1 - F(v_i))}_{\text{(III)}}. \tag{3.12}$$

Writing the likelihood in terms of (3.12) makes implementation of the algorithms that will be introduced easier.

3.3 Properties of the Likelihood Function

For interval censoring where each subject has the same number of inspections, (Van der Vaart & Wellner, 1996) states the inspection time distribution drops in the likelihood thus need no estimation. To gain more intuition about the likelihood, it helps to look at the function carefully. All three terms will give important information on the shape of the maximiser. Since $\mathcal{L}$ is to be maximised, it is essential to know whether an maximum exists. Therefore, the following proposition is important before maximisation.

Proposition 1. *The set $\mathcal{F}$ is convex and the function $\mathcal{L}$ is concave on $\mathcal{F}$.*

Proof. Fix two different functions $G_1, G_2 \in \mathcal{F}$ and $\epsilon \in [0,1]$. Then $\epsilon G_1 + (1-\epsilon)G_2$ is a convex combination of G_1 and G_2. Both functions only have jumps at V so the problem can be reduced this set. By non-decreasingness of the elements of $\mathcal{F}$ it holds that $\epsilon G_1(v_i) \leq \epsilon G_1(v_j)$ and $(1-\epsilon)G_2(v_i) \leq (1-\epsilon)G_2(v_j)$ if $i < j$. Let $\epsilon \in (0,1)$. Adding the two inequalities together yields

$$\epsilon G_1(v_i) + (1-\epsilon)G_2(v_i) \leq \epsilon G_1(v_j) + (1-\epsilon)G_2(v_j).$$

Convex combinations are increasing. It is clear that $\lim_{t\to\infty} \epsilon G_1(t) + (1-\epsilon)G_2(t) = 1$ and $t = 0$ results in zero. The convex combination is also right continuous, hence a cumulative distribution function. It can be concluded that $\epsilon G_1 + (1-\epsilon)G_2 \in \mathcal{F}$. The next step is to show concavity of $\mathcal{L}$ in $\mathcal{F}$. First note that the second derivative the logarithm is strictly negative and so the logarithm is strictly concave. This means that $\log(\epsilon x + (1-\epsilon)y) \geq \epsilon \log x + (1-\epsilon)\log y$ where equality only holds when $x = y$. Concavity is shown for each part of $\mathcal{L}$. Because $G_1 \neq G_2$ there exists at least one v_i on which they differ. It follows that

$$\log(\epsilon G_1(v_i) + (1-\epsilon)G_2(v_i)) > \epsilon \log G_1(v_i) + (1-\epsilon)\log G_2(v_i).$$

The same fact follows for the second and third term $\log(G_1(v_{\ell(i)}) - G_1(v_i))$ and $\log(1 - G_1(v_{\ell(i)}))$ of $\mathcal{L}$ for $i \leq n$. This follows again from concavity of the logarithm.

$$\begin{aligned}
&\log(\epsilon G_1(v_k) + (1-\epsilon)G_2(v_k) - \epsilon G_1(v_{\ell(k)}) + (1-\epsilon)G_2(v_{\ell(k)}))\\
=&\log(\epsilon(G_1(v_k) - G_1(v_{\ell(k)})) + (1-\epsilon)\log(G_2(v_k) - G_2(v_{\ell(k)}))\\
\geq&\epsilon\log(G_1(v_k) - G_1(v_{\ell(k)})) + (1-\epsilon)\log(G_2(v_k) - G_2(v_{\ell(k)})).
\end{aligned}$$

Note that in this case one cannot speak of strict concavity because it is possible that $G_1(v_k) - G_1(v_{\ell(k)}) = G_2(v_k) - G_2(v_{\ell(k)})$. Furthermore

$$\begin{aligned}
\log(1 - \epsilon G_1(v_i) - (1-\epsilon)G_2(v_i))) &= \log(\epsilon(1 - G_1(v_i) - (1-\epsilon)(1 - G_2(v_i)))\\
&> \epsilon\log(1 - G_1(v_i)) + (1-\epsilon)\log(1 - G_1(v_i))
\end{aligned}$$

Applying these inequalities to every term of the log-likelihood given in (3.8) yields

$$\mathcal{L}(\epsilon G_1 + (1-\epsilon)G_2) \geq \epsilon\mathcal{L}(G_1) + (1-\epsilon)\mathcal{L}(G_2),$$

hence $\mathcal{L}$ is concave, completing the proof. □

Of course, concavity is not a sufficient condition for a function to have a maximiser. If a function is not upper bounded, then there exists no meaningful maximiser. Note that for values in $c \in (0,1)$ it is true that $\log(c) \in (-\infty, 0)$. With this in mind, the log-likelihood function is upper bounded by a sum of zeros, so that $\mathcal{L}(F) \leq 0$. Before showing that there exists a unique maximiser, first a few results are shown about which elements in V can contain jumps. These results are used for proving uniqueness of the NPMLE.

3.4 Reduction of the Support of the NPMLE

The set $\mathcal{F}$ consists of piecewise constant distribution functions of which the support is the positive real axis. Its functions contain jumps only at V, that is, for $G \in \mathcal{F}$,

$$G(x) - G(x^-) \begin{cases} = 0 \text{ if } x \notin V, \\ \geq 0 \text{ if } x \in V. \end{cases}$$

It will turn out that the MLE $\hat{F}_n$ does not necessarily have jumps at all $v \in V$. This section will cover some propositions that prove on what subset of V, the jumps are concentrated. The function G can be expressed in terms of an integral. Let ν be the measure generating G by $G(t) = \int_{[0,t]} d\nu(\tau)$. Because ν can only have positive mass at points in V, it allows for a representation of a Dirac mixture. Denote the Dirac measure with all of its mass at v by η_v, then one can write $\nu = \sum_{i=1}^m \omega_i \eta_{v_i}$, where for all i, $\omega_i \in [0,1]$ are weights such that $\sum_{i=1}^m \omega_i = 1$. In measure theory, the support of ν is the smallest set of points at which ν has positive measure. Why the true support is smaller than V is explained, then propositions are stated and proven to make the discussion formal.

Some of the elements of V can be immediately removed from the support or ν due to their absence in the likelihood. Two cases exist. Take $\delta_i = 1$, then the effect this observation has on the likelihood (3.8) is $\log F(t_i)$. It is clear that u_i has no influence on the likelihood and can therefore be ignored in further inference. If this point has mass, that is, $\nu(\{u_i\}) > 0$ yields precisely the same likelihood value as putting its mass on the closest point to the right. What is meant by this is seen in Figure 3.1 and is shown to be true in the proof of Proposition 2. A similar argument holds for an observation having $\mu_i = 1$, then the time t_i has no influence on the likelihood. Therefore $\nu(\{t_i\}) = 0$ for these times. Since $\mathcal{L}(F)$ is to be maximised, it first needs to be finite. With this in mind, part (II) can shed light on where the discontinuities of the estimator $\hat{F}_n$ are located. Suppose that for some $k \in I_{2a}$ the step-function G has no increase on the interval $(v_k, v_{\ell(k)}]$, then $G(v_{\ell(i)}) - G(v_i) = 0$. Since $\log(0)$ does not exist it must be true that for all $i \in I_{2a}$ that $G(v_{\ell(i)}) - G(v_i) > 0$. This implies that for every $k \in I_{2a}$, the NPMLE must have at least one jump on the interval $(v_k, v_{\ell(k)}]$ for $\mathcal{L}(G)$ to be finite.

For all i such that $\delta_i = 1$, it can be seen that u_i is no part $\mathcal{L}$. However, such argument does not hold when $\gamma_i = 1$, then both t_i and u_i are part of the likelihood. It will be shown in proposition 2 this argument can be used to show that the nonparametric maximum likelihood estimator has no jumps on I_{2a}, reducing the set (3.9) of possible functions. In the remainder of this section, propositions are stated and proven that reduce the support of functions in $\mathcal{F}$ even more. A close look at the log-likelihood function shows that for right censored observations, the times u_i play no role. This suggests that in the maximisation algorithms proposed in chapter 5 these points need not to be taken into account. The same can be said for $i \in I_3$. If $\mu_i = 1$, then the contribution to the likelihood of this inspection does not involve t_i. The following proposition shows that the MLE $\hat{F}_n$ contains no jumps at these points.

Proposition 2. *Let $(t_1, u_1, \delta_1, \gamma_1), \ldots, (t_n, u_n, \delta_n, \gamma_n)$ be independent copies of (T, U, Δ, Γ). Let $\hat{G}$ maximise $\mathcal{L}(G)$ and let the measure ν be such that it generates $\hat{G}$ by $\hat{G}(t) = \int_{[0,t]} d\nu(t)$. If $\delta_i = 1$ and $\nu(\{u_i\}) > 0$, the maximiser is not unique. The same holds if $\mu_i = 1$ and $\nu(\{t_i\}) > 0$. Moreover, there exists a maximiser having no mass to these points.*

Proof. Fix i such that $\delta_i = 1$ and suppose $\hat{G}$ is such that $\nu(\{u_i\}) > 0$. Note that the term corresponding to $\delta_i = 1$ in $\mathcal{L}$ does not contain the value $G(u_i)$. Therefore changing the value $G(u_i)$ itself does not change the value of the log likelihood because it is not dependent on $G(u_i)$. Let $v_q = u_i$. The proof is based on using the measure ν to construct another measure $\tilde{\nu}$ by moving mass from v_q to v_{q+1}. The goal is to generate a measure $\tilde{\nu}$ such that it generates $\tilde{G}$ such that $\mathcal{L}(G) = \mathcal{L}(\tilde{G})$ having no mass at u_i. Define $\tilde{\nu}$ as

$$\tilde{\nu}(\{v_{q+1}\}) = \nu(\{v_q, v_{q+1}\}), \quad \tilde{\nu}(\{v_q\}) = 0. \tag{3.13}$$

For the rest of the support let $\tilde{\nu}|_{V\setminus\{v_q,v_{q+1}\}} = \nu|_{V\setminus\{v_q,v_{q+1}\}}$. By $\nu|_B$ means that the measure ν is restricted to B.[1] A visualisation of the movement of mass is shown in Figure 3.1. Doing so, the new distribution function $\tilde{G}$ is only changed between on the interval (v_q, v_{q+1}). A simple calculation shows this,

$$\begin{aligned} G(v_{q+1}) = \int_{[0,v_{q+1}]} d\nu(x) &= \int_{[0,v_q)} d\nu(x) + \nu(\{v_q, v_{q+1}\}) \\ &= \int_{[0,v_q)} d\tilde{\nu}(x) + \tilde{\nu}(\{v_{q+1}\}) = \int_{[0,v_{q+1}]} d\tilde{\nu}(x) = \tilde{G}(v_{q+1}). \end{aligned} \tag{3.14}$$

It is clear that at all times $v \in V\setminus\{v_q\}$ it is also true that $G(v) = \tilde{G}(v)$, hence, the likelihood is unchanged, i.e., $\mathcal{L}(G) = \mathcal{L}(\tilde{G})$. This implies that, if G is an MLE, another maximiser of $\mathcal{L}$ is generated, thus the maximiser of $\mathcal{L}$ is not unique. The same argument holds when i is such that $\mu_i = 1$ and $\nu(\{t_i\}) > 0$. This can be done for any of such cases where $\delta_j = 1$ and $\nu(u_j) > 0$, so there exists a function $\tilde{G}$ that has all mass removed from the inspection times u_i with $\delta_i = 1$ and t_i with $\mu_i = 1$, resulting in a maximiser without mass on these points. □

Because these inspection times are no part of the likelihood function, they do not need to be included in the optimisation, this is useful in the algorithmic procedure in chapter 5. Instead of observing which points are not part of the likelihood, one can also try to see which elements v the MLE cannot have mass while being part of $\mathcal{L}$. Suppose that G is the MLE. If G has mass on specific points, it will be shown that moving masses as done in (3.13) results in a strict increase in the likelihood. To argue why this is possible, the monotonicity of the logarithm is used. Consider i such that $\mu_i = 1$. Its contribution to the log-likelihood is $\mu_i \log(1 - F(u_i))$. Because the goal is to maximise $\mathcal{L}$, finding a way to increase $\mu_i \log(1 - F(u_i))$ will result in a strict increase. Decreasing $F(u_i)$ will result in a strict decrease of the likelihood. Let $v_q = u_i$, moving the mass $\nu(\{u_i\})$ to the first left point v_{q-1} as in Figure 3.2 yields a decrease of $F(u_i)$. A formal statement is found in the following proposition.

Proposition 3. *Let $(t_1, u_1, \delta_1, \gamma_1), \ldots, (t_n, u_n, \delta_n, \gamma_n)$ be independent copies of (T, U, Δ, Γ). The measure ν generating the NPMLE G by $G(t) = \int_{[0,t]} d\nu(t)$ maximising $\mathcal{L}$ has no mass at u_i when $\mu_i = 1$. Similarly, when $\gamma_i = 1$, then $\nu(\{t_i\}) = 0$.*

Proof. Let G be the NPMLE and fix i such that $\mu_i = 1$. Let ν be the measure generating $G \in \mathcal{F}$ such that $\nu(\{t_i\}) > 0$, then $\int_{[0,v_k]} d\nu(t) = G(v_k)$. As in the previous proof, a new measure is constructed using ν that increases the likelihood. Move the mass of ν at u_i to the first inspection on the right of u_i. This means, if $v_q = u_i$, then the new measure is defined as

$$\tilde{\nu}(\{v_{q+1}\}) = \nu(\{v_q, v_{q+1}\}), \quad \tilde{\nu}(\{v_q\}) = 0,$$

and $\tilde{\nu}|_{V\setminus\{v_q,v_{q+1}\}} = \nu|_{V\setminus\{v_q,v_{q+1}\}}$. Measures $\tilde{\nu}$ and ν only differ on the points v_q and v_{q+1}. Now it

[1] For any ν-measurable sets A, B it holds that $\nu|_B(A) = \nu(B \cap A)$. It is easy to show that $\nu|_B$ is a measure.

can be shown that this procedure will always increase the likelihood. Denote $v_{q+1} = t_r$, then

$$\begin{aligned}\mathcal{L}(G) =& \sum_{j=1}^{n} \delta_j \log G(t_j) + \gamma_j \log(G(u_j) - G(t_j)) + \mu_j \log(1 - G(u_j)) \qquad (3.15)\\ =& \left(\sum_{\substack{j=1\\ j\neq i,r}}^{n} \delta_j \log G(t_j) + \gamma_j \log(G(u_j) - G(t_j)) + \mu_j \log(1 - G(u_j)) \right) + \log(1 - G(u_i))\\ &+ \delta_r \log G(t_r) + \gamma_r \log(G(u_r) - G(t_r)) + \mu_r \log(1 - G(u_r))\\ <& \left(\sum_{\substack{j=1\\ j\neq i,r}}^{n} \delta_j \log \tilde{G}(t_j) + \gamma_j \log(\tilde{G}(u_j) - \tilde{G}(t_j)) + \mu_j \log(1 - \tilde{G}(u_j)) \right) + \log(1 - \tilde{G}(u_i))\\ &+ \delta_r \log \tilde{G}(t_r) + \gamma_q \log(\tilde{G}(u_r) - \tilde{G}(t_r)) + \mu_q \log(1 - \tilde{G}(u_r))\\ =&\mathcal{L}(\tilde{G}).\end{aligned}$$

the strict inequality in the third line is true because $G(u_i) > \tilde{G}(u_i)$ implies $\log(1 - G(u_i)) < \log(1 - \tilde{G}(u_i))$. This relocation procedure can be applied for every i with $\mu_i = 1$. If there is no inspection time on the right, the mass can be relocated to v_∞. This procedure results in a strict increase in the likelihood, thus, G was not the MLE. Consequently, the MLE cannot have mass at u_i if $\mu_i = 1$. In the case $\gamma_i = 1$ on the time t_i can have its mass moved to the first inspection time on the left. Then a new measure is found

$$\tilde{\nu}(\{v_{q-1}\}) = \nu(\{v_{q-1}, v_q\}), \quad \tilde{\nu}(\{v_q\}) = 0,$$

such that $G(t_i) < \tilde{G}(t_i)$ so that $\log(G(u_i) - G(t_i)) < \log(\tilde{G}(u_i) - \tilde{G}(t_i))$, increasing the likelihood. The movement of the mass is in this case different, see Figure 3.1 for an illustration. The rest of the proof remains the same. □

In the previous propositions two types of inspection times have been considered. One type of observation that play no role in the likelihood function. The other proposition looked at how to increase the likelihood by moving masses to the right. This makes it natural to ask if it is possible to increase the likelihood by *taking mass* from a point on the right. By looking at the likelihood function, it can be seen that some terms increase by increasing the value of F. Take for example $\delta_i = 1$, then the contribution to the likelihood is $\delta_i \log F(t_i)$. If a relocation of mass can find a distribution function $\tilde{F}$ such that $\tilde{F}(t_i) > F(t_i)$, then a strict increase is found. This is also the case when $\gamma_i = 1$ and a new distribution function $\tilde{F}$ is generated such that $\tilde{F}(u_i) > F(u_i)$ such that $\log(\tilde{F}(u_i) - F(t_i)) > \log(F(u_i) - F(t_i))$. The next proposition states that such points do exist.

Proposition 4. *Let $(t_1, u_1, \delta_1, \gamma_1), \ldots, (t_n, u_n, \delta_n, \gamma_n)$ be independent copies of (T, U, Δ, Γ). Let $G \in \mathcal{F}$ be a distribution function such that for $\delta_i = 1$ and $v_q = t_i$ and $\nu(\{v_{q+1}\}) > 0$, then $G \notin \operatorname{argmax}_{G' \in \mathcal{F}} \mathcal{L}(G')$. The same holds when $\gamma_i = 1$ and $v_q = u_i$ and $\nu(\{v_{q+1}\}) > 0$.*

Proof. Similar type of argument holds as used in the proof for Proposition 3. Consider the case where $\delta_i = 1$ and $v_q = t_i$ and $\nu(\{v_{q+1}\}) > 0$. By constructing a new measure $\tilde{\nu}$ such that

$$\tilde{\nu}(\{v_q\}) = \nu(\{v_q, v_{q+1}\}), \quad \tilde{\nu}(\{v_{q+1}\}) = 0.$$

The new distribution function $\tilde{G}$ is equal to G on the set $[0, v_q) \cup [v_{q+1}, \infty)$. The difference in the function values playing a role in the log likelihood is that $\tilde{G}(t_i) > G(t_i)$ so that $\delta_i \log \tilde{G}(t_i) > G(t_i)$, increasing the likelihood. For all other elements of $v \in V\backslash\{t_i\}$ it holds that $\tilde{G}(v) = G(v)$. Precisely the same idea is used as in (3.15) for the case when $\gamma_i = 1$ and $v_q = u_i$ where $\nu(\{v_{q+1}\}) > 0$. □

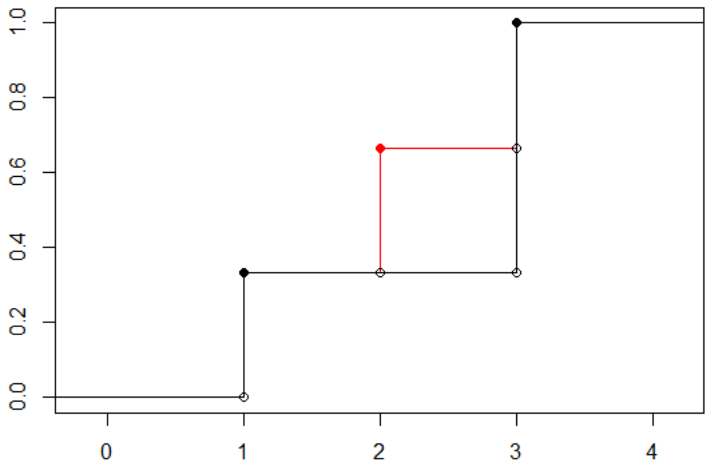

Figure 3.1: The original distribution function is displayed in red, for the biggest part coinciding with the black function. Moving the mass of $x = 2$ to $x = 3$ yields the black step function.

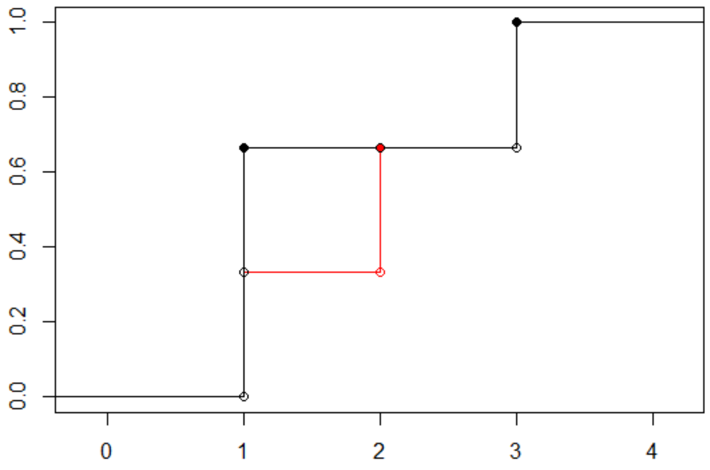

Figure 3.2: The original distribution function is displayed in red, for the biggest part coinciding with the black function. Moving the mass of $x = 2$ to $x = 1$ yields the black step function.

Suppose data are collected such that for the smallest inspection time $v_1 = t_q$ has $\delta_q = 0$. When maximising $\mathcal{L}(F)$ it follows that $F(t_q) = 0$. This is because $F(t_q)$ must be larger or equal than zero. There exist two cases. Take $\gamma_q = 1$, then F can be optimised such that $\log(F(u_q) - F(t_q))$ is maximal. This is done by letting $F(t_q) = 0$. If $\mu_q = 1$, by Proposition 2 the mass can be moved to v_2. A similar statement follows from data with $v_m = u_q$ having $\mu_q = 1$. Then the contribution to the log-likelihood is $\log(1 - F(u_n))$. If the MLE is a distribution function then $F(\infty) = 1$. If $F(u_q) = 1$ then $\log(1 - F(u_q)) = \log(0) = -\infty$. In this case, the MLE is a subdistribution since $F(u_q) < 1$, the leftover mass is then put at the point v_∞.

After data are collected, Proposition 2 allows to reduce the number of potential support points. Computationally this is advantageous as the complexity of the algorithms is an increasing function of the size of the support. Propositions 3 and 4 give more intuition on how the log likelihood behaves. They show that all the mass of the NPMLE is located at the inspection times t_i of $\delta_i = 1$ and u_i's of $\gamma_i = 1$. Next to this, these propositions allow for more reductions on the set of support points.

Now that the log likelihood and some of its properties are known, behaviour of its maximiser is studied. First uniqueness of the MLE is proven. Now that the support of the jumps of the functions in $\mathcal{F}$ can be reduced using the propositions above, the final support V of the measures generating the distribution functions in $\mathcal{F}$ is reduced to

$$V = \{v \in \mathcal{I}_1 \cup \mathcal{I}_{2b} \cup \{v_\infty\},\ v_1 < v_2 < \cdots < v_m < v_\infty\}, \tag{3.16}$$

having $m = |V|$ elements. For proving uniqueness, the fact that jumps are only located on V will be used.

3.5 Uniqueness of the Maximiser

In the previous paragraph properties for *the* maximiser, or should be spoken about *a* maximiser? Suppose in a study the event time distribution F needs to be found and conclusions are based upon it. If there are multiple cumulative hazard functions, conclusions based upon such function cannot be consistent. Because of this, uniqueness of the NPMLE is important in the analysis.

Theorem 1. *The log likelihood function $\mathcal{L}(F)$ defined in (3.8) has a unique maximiser on $\mathcal{F}$.*

Proof. Let the measures ν_1, ν_2 be the measures generating F_1, F_2 respectively by $F_i(t) = \int_{[0,t]} d\nu_i(t)$, $i = 1, 2$. The support of the measures is V, the set of all inspection times reduced using Propositions 2 and 3, that is, for both measures ν_1 and ν_2 we have $\nu(\{t_q\}) = 0$ if $\delta_q = 0$ and $\nu(\{u_q\}) = 0$ if $\gamma_q = 0$. First the easy case is considered. The proof is based on strict concavity of the logarithm, i.e., if $x, y \in \mathbb{R}^+$ such that $x \neq y$, then for $\varepsilon \in (0, 1)$,

$$\log(\varepsilon x + (1-\varepsilon)y) > \varepsilon \log x + (1-\varepsilon)\log y. \tag{3.17}$$

The distributions from above are chosen such that $F_1 \neq F_2$, otherwise there is nothing to prove. There must exist at least one $v_k \in V$ such that $F_1(v_k) \neq F_2(v_k)$. If v_k corresponds to an inspection time t_q with $\delta_q = 1$ then (3.17) shows that

$$\log(\varepsilon F_1(t_q) + (1-\varepsilon)F_2(t_q)) > \varepsilon \log F_1(t_q) + (1-\varepsilon)\log F_2(t_q). \tag{3.18}$$

If v_k corresponds to an inspection time u_q with $\mu_q = 1$ then

$$\begin{aligned}\log\left(1 - \varepsilon F_1(u_q) + (1-\varepsilon)F_2(u_q)\right) &= \log\left(\varepsilon(1 - F_1(u_q)) + (1-\varepsilon)(1 - F_2(u_q))\right) \\ &> \varepsilon \log(1 - F_1(u_q)) + (1-\varepsilon)\log(1 - F_2(u_q)).\end{aligned} \tag{3.19}$$

Only a single occurrence of either (3.18) or (3.19) is needed for strict concavity of $\mathcal{L}$ because a strictly concave function plus a concave function is strictly concave. A case that requires more attention is when $t_q = v_k$ corresponds to an inspection time such that $\gamma_q = 1$. When $F_1(u_q) - F_1(t_q) \neq$

$F_2(u_q) - F_2(t_q)$, (3.17) results in strict concavity, however, it may happen that $F_1(u_q) - F_1(t_q) = F_2(u_q) - F_2(t_q)$. Then (3.17) becomes an equality if $x = F_1(u_q) - F_1(t_q)$ and $y = F_2(u_q) - F_2(t_q)$. However, it this equality occurs, then it is possible to show that this local behaviour of F_1 and F_2 change non-local behaviour of the functions too as a consequence on where the jumps are located.

By definition of $\mathcal{F}$, no jumps occurs on $t_q = v_k$. Because if $F_1(v_k) \neq F_2(v_k)$, it is therefore also true that $F_1(v_{k-1}) \neq F_2(v_{k-1})$. If the time v_{k-1} corresponds to an inspection time with indicator variable $\gamma_{q_2} = 0$. Strict inequalities (3.18) and (3.19) finish the proof. If $\gamma_{q_2} = 1$ the same situation occurs, i.e.,

(*i*) $F_1(u_{q_2}) - F_1(t_{q_2}) \neq F_2(u_{q_2}) - F_2(t_{q_2},)$

(*ii*) $F_1(u_{q_2}) - F_1(t_{q_2}) = F_2(u_{q_2}) - F_2(t_{q_2})$,

where $F_1(u_{q_2}) \neq F_2(u_{q_2})$. If case (*i*) occurs, the proof is done. The problem becomes nested in case of (*ii*). By following thr last step multiple times with no occurrence of (*i*), one finds a q_0 such that $t_{q_0} = v_1$, for $\mu_{q_0} = 0$. If $\gamma_{q_0} = 1$ and (*ii*) occurs, then

$$F_1(u_{q_0}) - F_1(t_{q_0}) = F_2(u_{q_0}) - F_2(t_{q_0}) \tag{3.20}$$

and $F_1(u_{q_0}) \neq F_2(u_{q_0})$ so that $F_1(t_{q_0}) \neq F_2(t_{q_0})$. However, $t_{q_0} = v_1$ and the functions in $\mathcal{F}$ have no mass in t_{q_0} when $\gamma_{q_0} = 1$, this forces $F_1(t_{q_0}) = F_2(t_{q_0}) = 0$. Consequently, $F_1(u_{q_0}) - F_1(t_{q_0}) \neq F_2(u_{q_0}) - F_2(t_{q_0})$, contradicting (3.20). Thus, there must always exist some point in which the strict concavity of the logarithm can be applied. If $\delta_{q_0} = 1$ and $t_{q_0} \neq v_1$ and (*ii*) occurs, then there exists a smaller $t_{q_{b+1}} = v_1$. By the same reasoning it holds true that $F_1(t_{q_{b+1}}) \neq F_2(t_{q_{b+1}})$. This does not lead to a contradictory situation, but it does hold true that for a $\delta_{q_{b+1}} = 1$ that $F_1(t_{q_{b+1}}) \neq F_2(t_{q_{b+1}})$, so a single application of (3.18) yields strict concavity of $\mathcal{L}$. □

Uniqueness of the maximiser does not necessarily result in a good estimate. In simple cases it does. Suppose a mean of a random variable $X' \sim F'$ is to be estimated if it exists. Let $X'_1, \dots, X'_n$ be i.i.d. samples from F', then the sample mean converges to the mean, i.e., $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X'_i \to \mathbb{E}X'$. In this case, the estimator for the mean $\bar{X}$ converges to where it should. This motivates to introduce the concept of consistency. Let $Y_i, \dots, Y_n$ be independent identically distributed from G according to distribution $\mathbb{P}_Y$. An estimator $\hat{\Theta}_n := \hat{\Theta}_n(Y_1, \dots, Y_n)$ for θ is said to be consistent if $\hat{\Theta}_n$ converges to θ in $\mathbb{P}_Y-$probability, that is, for all fixed $\varepsilon > 0$

$$\mathbb{P}_Y(|\hat{\Theta}_n - \theta| > \varepsilon) \to 0,$$

in probability as $n \to \infty$. The consistency is called strong when the convergence in the case of almost sure convergence.

An interpretation of this definition is that for a fixed value $\varepsilon > 0$, as the sample size increases, it becomes less probable that the absolute error in the estimation is greater than ε. In estimation problems consistency is an important property as it guarantees that the estimator converges to the right value as the sample size tends to infinity. Aside from the mathematical definition, it can be seen as a property that shows whether an estimator converges to where it should. For standard examples of the maximum likelihood estimator, for example if samples are drawn i.i.d. from a density with finite mean, the estimator for the mean μ becomes

$$\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^n x_i \to \mu \qquad \text{as } n \to \infty$$

by the law of large numbers. There also exist cases where an estimator is unbiased but does not converge to its mean. An easy example illustrating this is the statistic $\Theta(X) = X_n$. Then $\mathbb{E}\Theta(X) = \mu$ but no convergence takes place.

However, in less trivial cases, things may become more complicated. In (Schick & Yu, 2000) state that the $\mathcal{L}(F)$ has a maximiser. Research has been done on consistency of the NPMLE for interval censored data in (Van der Vaart & Wellner, 2000) (Song, 2004). Vaart concludes for the mixed case interval censoring in Theorem 8 that under the L_1 distance, the NPMLE converges to the true distribution almost surely, thus, the NPMLE is consistent. This result holds for the IC2 model. In the next chapter, the IC2 model is extended to the Cox model.

Chapter 4

Cox Proportional Hazards Model

The hazard function λ is a function of only time. It can be expected that the hazard is different for every subject having different covariates, resulting in a different hazard. This chapter introduces a model to include other variables into survival models other than time. For example, mortality does not merely depend on time but can also be influenced by other factors that can quantify health such as blood pressure. For instance the drugs taken by a subject during medical study influence the time-behaviour of the hazard function in a way. Effect of drugs can be beneficial as well as it could be harmful, dosage of such drugs can be a covariate in a model. Then the model can help to gain insight in how the hazard is being influenced by the dosage. If the drugs lead to a decrease in the hazard function, it may be possible that the drugs are beneficial.

Let s be a n by d real matrix of explanatory variables and let s_j denote the vector of variables of subject j. Different models exist that allow additional variables to be introduced. One important model is the accelerated failure time model. As the name suggests, it models an accelerated deterioration by a covariate-dependent scale parameter. This model is described mathematically as

$$F(t|s) = \mathbb{P}(T \leq t|S = s) = F_0(t/c(\beta^\top s)),$$

where $c(\cdot)$ is a function describing how the covariates s affect the "speed of time". Another important model on which this study is focussed is the Cox proportional hazard model, named after Sir David Cox due to his study in (Cox, 1972). In the general proportional hazards model, the hazard function is assumed to be proportional to a function of covariates, this assumption is called the proportional hazards assumption. This function is of the form $c(\beta^\top s)$ such that $c(0) = 1$. From this point the Cox proportional hazards model is abbreviated as the Cox model. The function c is defined in the Cox model as

$$c(\beta^\top s_i) = \exp(\beta^\top s_i) = \exp\left(\sum_{j=1}^{d} \beta_j s_{ij}\right).$$

The vector $\beta \in \mathbb{R}^d$ is referred to as the Cox coefficients. Since proportionality of $\exp(\beta^\top s)$ with the baseline is assumed, the following expression for the (cumulative) hazard function is obtained,

$$\begin{aligned} \lambda(t|s_i) &= \exp(\beta^\top s_i)\lambda_0(t), \qquad (4.1)\\ \Lambda(t|s_i) &= \exp(\beta^\top s_i)\Lambda_0(t). \end{aligned}$$

If $s_i = \mathbf{0}$, i.e. $\exp(\beta^\top s_i) = \exp(0) = 1$, then the hazard function is equal to the so-called baseline hazard function, denoted by λ_0. It needs to be noted that unlike for distribution functions, positive multiples of (cumulative) hazard functions are again a (cumulative) hazard function. If a set has this property, then the set is called a *cone*. The right hand sides of both expressions in (4.1) consist of a parametric part, where β is the parameter and a nonparametric part, λ_0. Proper estimation of β can lead to useful conclusions. The proportional hazards assumption also has meaning in terms of the survival function. Some algebra shows that

$$\Lambda(t|s) = \exp(\beta^\top s)\Lambda_0(t) = -\exp(\beta^\top s)\log S_0(t) = -\log\left(S_0(t)^{\exp(\beta^\top s)}\right) \tag{4.2}$$

Now, using equation (2.3) by applying the mapping $x \mapsto -\log x$ on both sides yields

$$S(t) = S_0(t)^{c(\beta^\top s)}. \tag{4.3}$$

The latter also allows to write F in terms of the baseline distribution F_0. From equation (4.3) the following is obtained

$$F(t) = 1 - (1 - F_0(t))^{\exp(\beta^\top s)}. \tag{4.4}$$

If s_i and β are such that $\exp(\beta^\top s_i) > 1$, then the hazard function becomes higher than the baseline hazard. When $\exp(\beta^\top s_i) < 1$ it is implied that subject i has a lower risk of failure than the baseline hazard. Using the proportional hazards assumption should not be done without verifying whether the assumption is appropriate. Even though testing this assumption is not part of this study, it is important to mention.

In (Minami et al., 1998) concludes that the a high blood pressure increases the risk on a myocardial infarction. A study can be done on the influence of blood pressure $(s^\top)_1$ on the risk of having a myocardial infarction. Of course there is a link between these medical conditions, but proper estimation of β under the proportional hazards assumption can quantify how blood pressure can increase the risk of a heart attack once the proportional hazard assumption is verified. If enough data are available, more covariates can have their weight β estimated accurately. This allows for introducing gender as a second covariate. It is possible that the weight β_2 for genders s_2 is estimated to be close to zero, i.e., $\beta_2 \approx 0$ which implies that $\exp(\beta^\top s_j) \approx \exp(\beta_1 s_{1j})$. Then one could study whether gender is an explanatory variable.

The hazard ratio is earlier discussed in chapter 2. In the Cox model, comparing two different sub-populations, the hazard ratio simplifies elegantly. Continuing with the medical example above, suppose weights β have been fitted to the data and the cumulative hazard function Λ_0 is estimated. Let one subject have covariates s_j and another subject have s_k. The hazard ratio becomes entirely dependent on the covariates and not on time,

$$\mathrm{HR}(t) = \frac{\lambda(t|s_j)}{\lambda(t|s_k)} = \frac{\exp(\beta \cdot s_j)\lambda_0(t)}{\exp(\beta \cdot s_k)\lambda_0(t)} = \exp(\beta \cdot (s_j - s_k)).$$

It is possible to estimate the baseline hazard function parametrically. Generalised Gamma, Weibull and Gompertz are distributions that are commonly used in applications, see (El-Gohary et al., 2013) and (Bradburn, Clark, Love, & Altman, 2003). Assuming a parametric distribution on the data strongly decreases the size of the function space in which the optimisation is done, which possibly introduces a bias. In many situations it is desirable to let the data speak for themselves, and therefore opt for a nonparametric approach. Parametric estimation only needs approximation of parameters. Because the nonparametric approach needs estimation of its function values such an approach becomes more complicated and computationally intensive compared to parametric approximation where only a few parameters need estimation. Another downside to a non-parametric approach is that convergence rates are usually lower than for parametric methods. For example, the ECDF has a convergence rate to the true distribution of $\sqrt{n}$, see (Wasserman, 2006). Theorem 4.2 in (Groeneboom, Jongbloed, & Witte, 2010) states that the smoothed version of the NPMLE has a convergence rate of $n^{2/5}$.[1]

[1] The concept of smoothing is elaborated in paragraph 6.1.

4.1 Semiparametric Regression

Additional to estimating Λ_0, in the Cox model one also needs to estimate the Cox coefficients. Maximising equation (3.8) gives the most likely estimator according to the data. To apply this to Cox models, a translation is required, expressing the distribution function F in terms of the Cox hazard will do this. The relation between the distribution function and the Cox cumulative hazard is required and is easily found using equation (2.5) together with (4.1). It follows that

$$F(t|s) = 1 - S(t|s) = 1 - \exp(-\Lambda(t|s)) = 1 - \exp\left(-e^{\beta^\top s}\Lambda_0(t)\right). \tag{4.5}$$

Plugging this expression into the log-likelihood function given in equation (3.8) yields

$$\begin{aligned}\mathcal{L}(\beta, \Lambda_0) = \sum_{i=1}^{n} \Big[&\delta_i \log\left(1 - \exp(-\Lambda_0(t_i)e^{\beta^\top s_i})\right) \\ &+\gamma_i \log\left(\exp\left(-\Lambda_0(t_i)e^{\beta^\top s_i}\right) - \exp\left(-\Lambda_0(u_i)e^{\beta^\top s_i}\right)\right) - \mu_i\Lambda_0(u_i)e^{\beta^\top s_i}\Big].\end{aligned} \tag{4.6}$$

There is a slight abuse of notation as $\mathcal{L}$ was already defined. From this point, the log-likelihood refers to this expression. Note that the cumulative hazard function does not in general belong to $\mathcal{F}$. Because the distribution function and the cumulative hazard are bijective, the class of cumulative hazard functions can be defined as

$$\mathcal{H} := \{\Lambda : \Lambda(t) = -\log(1 - F(t)) \text{ for all } F(t) \in [0, 1)\} \tag{4.7}$$

with the convention that $\log(0) = -\infty$. The difficulty is that the maximising argument needs to be found for two objects simultaneously, one is parametric and the other is non-parametric, such problem is called semi-parametric regression. Next, a method to deal with such a problem is explained.

The method explained here is called the profile likelihood approach. This approach allows to split the problem into multiple non-parametric problems. Fix $\beta \in \mathbb{R}^d$ so that only the nonparametric problem needs to be solved for this particular β. The profile likelihood is defined by

$$p\mathcal{L}_\beta(\Lambda_0) = \mathcal{L}(\beta, \Lambda_0). \tag{4.8}$$

Let $\hat{\Lambda}_0^\beta$ be the maximiser of $p\mathcal{L}_\beta$. If this can be computed for every β, then the NPMLE is the tuple $(\beta, \hat{\Lambda}_0^\beta)$ maximising $\mathcal{L}$.

$$(\hat{\beta}, \hat{\Lambda}_0) = \underset{\beta' \in \mathbb{R}^d}{\operatorname{argmax}}\, \mathcal{L}(\beta', \hat{\Lambda}_0^{\beta'}) = \underset{\beta' \in \mathbb{R}^d, \Lambda_0 \in \mathcal{H}}{\operatorname{argmax}}\, \mathcal{L}(\beta', \Lambda_0).$$

This method is called profile likelihood approach. A simple example is given to make the idea more clear.

Example 4. *Consider the function $f : [0, 2] \times [0, 1]$ defined by $f(x, y) = \frac{xy^2}{x^2+1}$. The profile likelihood approach works by first fixing $x = x_0 \in [0, 2]$. Proceed by maximising $f_0(y) = f(x_0, y)$ in which the only variable is y. It is easy to see that $y = 1$ maximises f_0 independent of which x is fixed. As the second step of the approach is to maximise $f(x, 1)$ over x. Equating the derivative of $f(x, 1)$ to zero yields $x = 1$. The maximiser of f using the profile likelihood approach is then (1,1).*

A problem that arises is the computation time of estimating the baseline for a given β so a fine enough discretisation of a bounded subset of $\mathbb{R}^d$ and maximise the profile likelihood is implausible. It is desirable to decrease the number of β's for which the profile likelihood function needs to be optimised. Efficient choices of β for which the baseline is estimated are crucial. For this, an algorithm is proposed in Chapter 5 that chooses such β limiting the computation time.

A useful property of $\mathcal{L}(\beta, \Lambda_0)$ is concavity with respect to β. This is shown through convexity of the exponential function. A negative multiplication of a convex function is concave. That makes it clear that $-\Lambda_0(v_i)e^{\beta^\top z_i}$ is concave with respect to β. Compositions of concave or convex functions preserve this property under non-decreasingness. (Tibshirani & Wasserman, 2015) state that if $g : \mathbb{R}^n \to \mathbb{R}$, $h : \mathbb{R} \to \mathbb{R}$, then for the composition $f = h \circ g$, the following holds:

(i) If h is concave and non-decreasing and g is concave, then f is concave.

(ii) If h is concave and non-increasing and g is convex, then f is concave.

The first term of the likelihood function $\log\big(1 - \exp(-\Lambda_0(t_i)e^{\beta^\top s_i})\big)$ is shown to be concave using these facts. It follows from (ii) that $-\Lambda_0(t_i)e^{\beta^\top s_i}$ is concave. Since the mapping $x \mapsto 1 - \exp(x)$ is concave non-increasing mapping, $\beta \mapsto 1 - \exp(-\Lambda_0(t_i)e^{\beta^\top s_i})$ is concave by (i). The middle part is shown to be concave using the previous two results and the fact that a sum of concave functions is again concave. It follows that

$$\begin{aligned}
&\log\Big(\exp(-\Lambda_0(t_i)e^{\beta^\top s_i}) - \exp(-\Lambda_0(u_i)e^{\beta^\top s_i})\Big) \qquad (4.9)\\
=&\log\Big(1 - \exp\big(-\big[\Lambda_0(u_i) - \Lambda(t_i)\big]e^{\beta^\top s_i}\big)\Big) + \log\Big(\exp(-\Lambda_0(u_i)e^{\beta^\top s_i})\Big)\\
=&\log\Big(1 - \exp\big(-\big[\Lambda_0(u_i) - \Lambda(t_i)\big]e^{\beta^\top s_i}\big)\Big) + \big(-\Lambda_0(u_i)e^{\beta^\top s_i}\big),
\end{aligned}$$

and both terms of the last line are concave, hence, $\mathcal{L}$ is concave in β. Concavity with respect to Λ_0 can be shown using the same method as done for proving strict concavity of $\mathcal{L}(F)$.

It is proven that the maximiser of $\mathcal{L}(F)$ is unique. Because the goal of this study is to maximise in the Cox model, it is natural to ask whether the NPMLE $\hat{\Lambda}_{0n}$ is unique. Fix β and s. By (2.5) it is true that

$$\mathcal{L}(F) = \mathcal{L}\big(1 - \exp(-\Lambda_0 e^{\beta^\top s})\big)$$

which equals the log-likelihood function in the Cox model. Both sides share the same maximiser as the function $\mathcal{L}$ itself remains unchanged. Suppose $\hat{\Lambda}_{0n} \in \mathcal{H}$ maximises the profile likelihood. Again, by the bijective relation in (2.5), the maximiser $\hat{\Lambda}_{0n}(t)$ has a unique $\hat{F}_n \in \mathcal{F}$, defined by

$$\hat{F}_n(t) = 1 - \exp\big(\hat{\Lambda}_{0n}(t)e^{\beta^\top s}\big)$$

for $t \in V$. $\hat{F}_n$ is the unique maximiser of $\mathcal{L}(F)$. Because the cumulative hazard and the distribution function are one-to-one and only one maximiser $\hat{F}_n$ exists, $\hat{\Lambda}_{0n}$ is the unique maximiser of $p\mathcal{L}_\beta\big(\Lambda_0\big)$. Until now it is not known how the maximisers can be computed. The next chapter will introduce algorithms that are able to do this.

Chapter 5

Algorithms for Maximising the Likelihood Function

Because estimation of hazard function outside the Cox model is difficult, adding an extra parameters makes things even more difficult. No method is known to estimate β and Λ_0 simultaneously. The profile likelihood method will be used to make the problem easier, the MLE can then be solved in two parts. Fix $\beta \in \mathbb{R}^n$. An algorithm able to solve for $\hat{\Lambda}_0^\beta = \text{argmax}_{\Lambda_0 \in \mathcal{H}} p\mathcal{L}_\beta(\Lambda_0)$ is proposed. This algorithm is called the iterative convex minorant algorithm and is studied in (Groeneboom & Wellner, 1992). More algorithms that can solve the NPMLE exist, see e.g. expectation maximisation and the support reduction algorithm. All these two are studied in (Dempster et al., 1977) and (Groeneboom et al., 2008). These algorithms are of iterative nature. This brings a problem when searching for the optimal β. The idea of the profile likelihood is to fix multiple β's and for each of them, $\hat{\Lambda}_0^\beta$ is computed iteratively. This can become very costly in terms of computation time. The profile likelihood implicitly becomes a function of β so the profile likelihood needs optimisation in $\mathbb{R}^d$. Many algorithms that optimise functions in Euclidean spaces make use of derivatives. The derivative of $p\mathcal{L}_\beta$ with respect to β is time consuming to compute as no explicit formula is available. Numerical approximations can be computed for derivatives with respect to every direction of β. In order to do this, the profile likelihood needs to be optimised multiple times for computing single derivative, making derivative based algorithms a infeasible option. Due to this, so-called derivative-free methods may be a better alternative for optimising β. Two well-known examples of such algorithms are Hooke-Jeeves and Nelder-Mead, studied in (Hooke & Jeeves, 1961) and (Nelder & Mead, 1965). Both algorithms use a set of points in the domain of a multidimensional objective function. The objective function $\mathcal{L}$ is evaluated at these points and based on the output, the algorithm attempts to find a direction increasing the objective function. It will turn out that optimising the profile likelihood is computationally expensive. Therefore one should choose a derivative-free method that has needs the least function evaluations.

First the ideas behind the algorithms solving the profile likelihood are treated. The remaining part of this chapter will cover the derivative-free algorithms.

5.1 Isotonic Regression

Consider the regression context with a set of independent observations $(X_i, Y_i)_{i=1}^n$ for which a regression function r is to be estimated. Each observation Y_i represents a drawing from the random variable that can be expressed as

$$Y_i = r(X_i) + \varepsilon_i, \tag{5.1}$$

where ε_i models the noise and X_i is the explanatory variable for Y_i and r is an unknown function. The random variables modelling noise ε_i are independent following the same distribution with a finite second moment. In standard linear regression, one restricts r to be of the form $r(x) = ax + b$

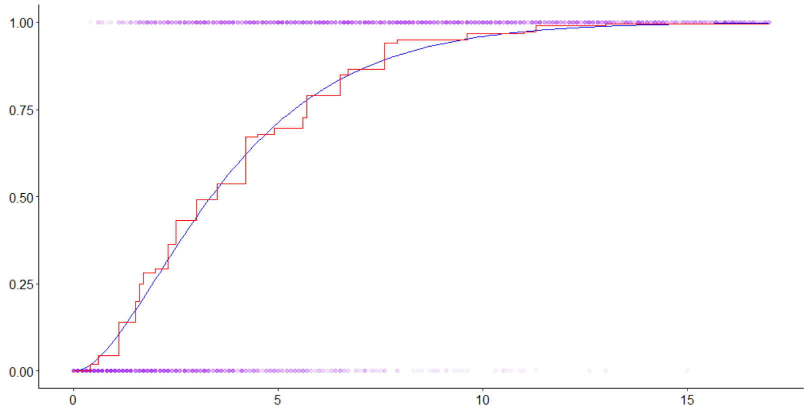

Figure 5.1: A dataset is simulated of size $n = 4000$. The times t_i are simulated from an exponential distribution. Event-times y_i are then sampled from $\mathrm{Ber}(F_{gamma}(t_i))$ where F_{gamma} is the distribution of $\mathrm{Gamma}(2, 1/2)$ displayed in blue. Each purple dot represents a coordinate t_i, y_i. The dots are transparent for better visual interpretation. The red line is the output of isotonic regression.

where $a, b \in \mathbb{R}$. Choice of the set of basis function depends on context. When periodic behaviour is expected, a suitable basis could contain a sine function so that $r(x) = a + b\sin(\omega x)$. Sometimes, when not much is known about the regression function r, a functional assumption is already limiting the function space too much. The only constraint isotonic regression puts on r is that it must be non-decreasing. An important property used in the algorithms introduced in chapter 5 is stated in the next lemma. Only assuming that r is non-decreasing, the data will be used to make a nonparametric estimation for this function. This comes useful when estimating for example distribution functions, hazard functions, growth curves, or other monotone relations. Non-increasing functions can also be estimated by approximating $-r$ using the non-increasing data $-(v_i)_{i=1}^m$. Applications of isotonic regression are first given in the following examples.

Example 5. *A microbiologist is interested in the growth curve of an unknown species of bacteria. Bacteria are self-reproductive and therefore growth is measured by the number of bacteria per square centimetre. This quantity is measured up to an uncertainty which is assumed to be symmetric around zero. Since the colony is cultivated in a Petri-dish with appropriate nutrition, it can be said that the density is an non-decreasing function of time. This gives rise to a regression context that fits in the model of (5.1) where r is non-decreasing and the density is fitted as a function of time. After a certain time the colony starts dying out. From this time, isotonic regression is no longer valid as it does not allow decrease.*

Example 6. *Binary regression is commonly applied in practice. One is interested in the relationship between an (exact) time-to-event and a set of explanatory variables x. A common technique is called logistic regression in which the logistic distribution is assumed to be the event time distribution of the form $F(x) = (1 + \exp(-\beta \cdot x))^{-1}$. Isotonic can offer a nonparametric alternative. Consider data $y_1, \ldots, y_n$ having values in the set $\{0, 1\}$ and times t_i. Figure 5.1 displays this situation. A free-form is fitted between the graph (t, y).*

Regression is an optimisation problem, so a error metric is required. The weighted sum of squared residuals $\varepsilon_i^2 = (y_i - r(x_i))^2$ is defined as a function of r by

$$Q(r) = \sum_{i=1}^{n} \big(y_i - r(x_i)\big)^2 w_i. \tag{5.2}$$

Here the weights w_i, $i = 1, \ldots, n$ are non-negative. Minimising this functions in $\mathcal{H}$ results in

$$\hat{r} = \operatorname*{argmin}_{r \in \mathcal{H}} \sum_{i=1}^{n} \big(y_i - r(x_i)\big)^2 w_i \tag{5.3}$$

where $w \in [0,1]^n$ is a vector of weights. These weights can be chosen depending on context, usually $w_i = 1$ for all i is used. Uniqueness of the minimiser of Q is shown through its (strict) convexity. Second derivatives with respect to $r(x_i)$ are given by

$$\frac{\partial^2}{\partial r(x_i)^2} Q(r) = \frac{\partial^2}{\partial r(x_i)^2} \sum_{i=1}^{n} \big(y_i - r(x_i)\big)^2 w_i = 2r(x_i)w_i.$$

If all $r(x_i)$'s are positive, this results in a positive definite matrix of second derivatives of Q implying strict convexity of Q in r. Suppose that there are two minimisers $r^{(1)}$ and $r^{(2)}$ with $r^{(1)} \neq r^{(2)}$, then for $\epsilon \in (0,1)$,

$$Q(\epsilon r^{(1)} + (1-\epsilon)r^{(2)}) \leq \epsilon Q(r^{(1)}) + \epsilon Q(r^{(2)}) = Q(r^{(1)}) = Q(r^{(2)}).$$

This would imply that any convex combination $\epsilon r_1 + (1-\epsilon)r_2$ results in a new minimiser for Q, which contradicts the assumption of them being minimisers. An important result of this theory is given in (E., Robertson, Wright, & Dykstra, 1990). The authors state that the solution $r = \big(r(x_1), \ldots, r(x_n)\big)$ of (5.3) is given by the left derivative of the greatest convex minorant of the diagram that connects the lines of the coordinates. The diagram is defined by the points

$$P_0 = (0,0) \text{ and } P_j = \left(\sum_{i=1}^{j} w_i, \sum_{i=1}^{j} w_i y_i\right) \text{ for } 1 \leq j \leq n. \tag{5.4}$$

More specifically, $\hat{r}(x_j)$ is given by the left derivative evaluated at $\sum_{i=1}^{j} w_i$. The left derivative of a function ϕ is equivalent to the standard derivative if ϕ is continuously differentiable, if it is only piecewise continuously differentiable, the left derivative is defined as the slope on the left side of the jump. The greatest convex minorant of a diagram is the greatest function that lies completely under or on the diagram, restricted to being a convex function. This can be thought of the diagram as nails in a wall, then a thread is lifted from the bottom up, see Figure 5.2. The left derivative of a function $\phi : D \to \mathbb{R}$ at point x is defined as the limit

$$\lim_{h \to 0^-} \frac{\phi(x+h) - \phi(x)}{h},$$

for every $x \in D^o$. In the context of estimating event-time distributions or hazard functions, the function r is restricted to the set $\mathcal{F}$ or $\mathcal{H}$, that is, jumps occur only at the set V. The left derivatives of the diagram (5.4) at the point P_k corresponds to the value of $\hat{r}_k = \hat{r}(x_k)$, this holds for each $k = 1, \ldots, n$. Because the points x_k are the elements V in which the jumps lie, the solution $\hat{r}$ is fully determined by isotonic regression. Because the functions behaviour between the elements of V is constant, $\hat{r}$ has its function value defined on $[0, \max(x)]$. Let $x_{(k)}$ denote the kth ordered statistic of x, that is, the set is put in order from smallest to largest, such that $x_{(1)} \leq x_{(2)} \leq \ldots, x_{(n)}$. Knowing $r(x_{(k)})$ and $r(x_{(k+1)})$, then the function values of $\hat{r}$ on the interval $(x_{(k)}, x_{(k+1)}]$ is equal to $r(x_{(k+1)})$. A close look at the image shows that the two red points above the line do not influence the output.

5.2 Iterative Convex Minorant

The algorithm introduced in this chapter is strongly related to Newton's method. Newton's method is a well-known root-finding algorithm that makes use of Taylor's theorem. Let $\varphi : \mathbb{R} \to \mathbb{R}$ be a convex function that is at least twice differentiable. Standard use of Newton's method finds a $y' \in \mathbb{R}$ such that $\varphi(y') \approx 0$. If y' is obtained by finding the root of the derivative, φ', then $\varphi(y')$ is should

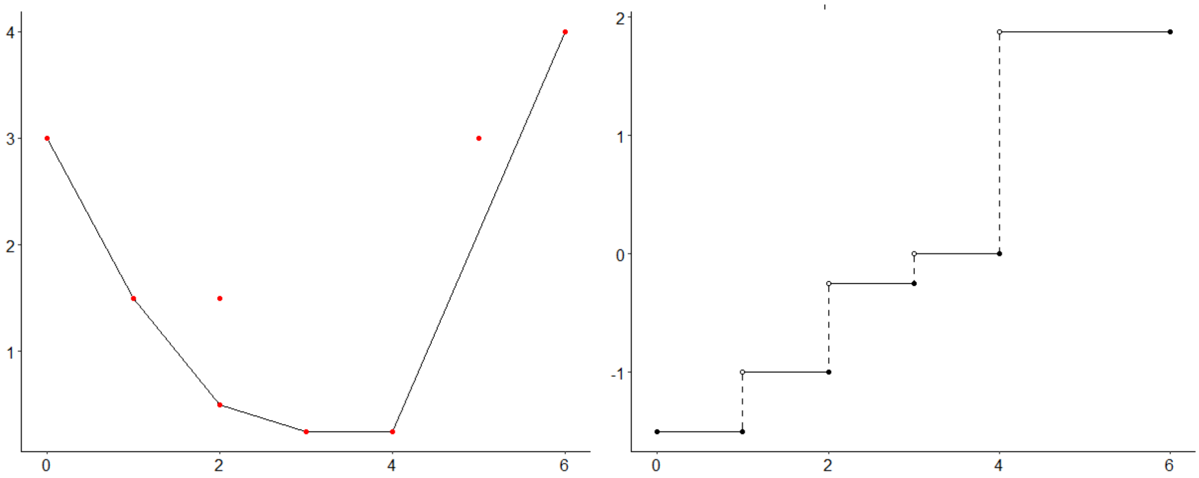

Figure 5.2: The left image shows the greatest convex minorant based on the cumulative sum diagram. The right image shows the left derivative of the cumulative sum diagram. At a jump, the solid dot denotes the value of the left derivative at the jump.

be close to an extreme of φ. Since the profile likelihood needs maximisation, the latter procedure is of interest. Let y_{root} be the root of φ'. Then

$$0 = \varphi'(y_{root}) \approx \varphi'(y') + (y' - y_{root})\varphi''(y')$$

by applying Taylor's theorem and truncate the polynomial after the second term. Isolating y_{root} yields

$$y_{root} \approx y' - \frac{\varphi'(y')}{\varphi''(y')}. \tag{5.5}$$

Updating y' as the right hand side of (5.5) iteratively results in the Newton algorithm. This procedure iterates to the true y_{root}. Essentially, Newton's method finds the best linear approximation every step, not taking into account second and higher order terms.

Let $\hat{\Lambda}_0$ be the the NPMLE and $\bar{\Lambda}_0 \in \mathcal{H}\backslash\{\hat{\Lambda}_0\}$ another estimate. It will turn out that the difference of these in the log-likelihood

$$\mathcal{L}(\beta, \hat{\Lambda}_0) - \mathcal{L}(\beta, \bar{\Lambda}_0) \tag{5.6}$$

can be approximated by a quadratic form. Clearly, the difference is positive because $\mathcal{L}(\beta, \hat{\Lambda}_0) > \mathcal{L}(\beta, \bar{\Lambda}_0)$ and $\mathcal{L}(\beta, \hat{\Lambda}_0)$ is the maximum of $\mathcal{L}(\beta, \Lambda_0)$ for fixed β. To approximate the NPMLE it makes sense to minimise the quantity above in the set $\mathcal{H}$. Finding algorithmic sequence iterating over $\bar{\Lambda}_0$ to make the quantity in (5.6) small. Because $\bar{\Lambda}_0$ is chosen or known and β is fixed, the only unknown in (5.6) is $\hat{\Lambda}_0$. Evaluating the Taylor expansion of $\mathcal{L}(\beta, \hat{\Lambda}_0)$ around $\bar{\Lambda}_0$ up to the third term yields

$$\begin{aligned} \mathcal{L}(\beta, \hat{\Lambda}_0) - \mathcal{L}(\beta, \bar{\Lambda}_0) &= \mathcal{L}(\beta, \bar{\Lambda}_0) - \mathcal{L}(\beta, \bar{\Lambda}_0) + (\hat{\Lambda}_0 - \bar{\Lambda}_0)^\top \nabla \mathcal{L}(\beta, \bar{\Lambda}_0) \\ &+ \frac{1}{2}(\hat{\Lambda}_0 - \bar{\Lambda}_0)^\top \mathbf{Hess}(\mathcal{L})(\beta, \bar{\Lambda}_0)(\hat{\Lambda}_0 - \bar{\Lambda}_0) + O\big(|\hat{\Lambda}_0 - \bar{\Lambda}_0|^3\big) \\ &\approx (\hat{\Lambda}_0 - \bar{\Lambda}_0)^\top \nabla \mathcal{L}(\beta, \bar{\Lambda}_0) + \frac{1}{2}(\hat{\Lambda}_0 - \bar{\Lambda}_0)^\top \mathbf{Hess}(\mathcal{L})(\beta, \bar{\Lambda}_0)(\hat{\Lambda}_0 - \bar{\Lambda}_0) \\ &=: \tilde{q}(\hat{\Lambda}_0, \bar{\Lambda}_0) \end{aligned} \tag{5.7}$$

Where $\nabla \mathcal{L}$ is the gradient with respect to $\Lambda_0(v_j)$ for $j = 1, \ldots, m$ and $\mathbf{Hess}(\mathcal{L})$ is the matrix of second derivatives. These objects exist because there are only a finite number of function values $\Lambda_0(v_i)$ to which derivatives need to be calculated. The j'th element of the gradient and j, k'th element of the Hessian are given by

$$\nabla \mathcal{L}(\beta, \Lambda_0)_j = \frac{\partial \mathcal{L}(\beta, \Lambda_0)}{\partial \Lambda_0(v_j)} \quad \text{and} \quad \mathbf{Hess}(\mathcal{L})(\beta, \Lambda_0)_{jk} = \frac{\partial^2 \mathcal{L}(\beta, \Lambda_0)}{\partial \Lambda_0(v_j) \partial \Lambda_0(v_k)}.$$

Minimising $\tilde{q}$ with respect to the variable $\bar{\Lambda}_0$ seems difficult at this point. When maximising an expression above, the maximiser does not change when a quantity independent of the maximiser is added to the expression. Choosing an appropriate candidate, $-\nabla\mathcal{L}(\beta,\bar{\Lambda}_0)^\top\mathbf{Hess}(\mathcal{L})(\beta,\bar{\Lambda}_0)^{-1}\nabla\mathcal{L}(\beta,\bar{\Lambda}_0)$ which is independent of $\hat{\Lambda}_0$, allows to rewrite the difference in a quadratic form. This form turns out to be suitable for isotonic regression. Adding this constant to $\tilde{q}$ yields

$$q(\hat{\Lambda}_0,\bar{\Lambda}_0) := \Big[\big\{\hat{\Lambda}_0-\bar{\Lambda}_0+\mathbf{Hess}(\mathcal{L})(\bar{\Lambda}_0)^{-1}\nabla\mathcal{L}(\bar{\Lambda}_0)\big\}^\top\mathbf{Hess}(\mathcal{L})(\bar{\Lambda}_0) \\ \times\big\{\hat{\Lambda}_0-\bar{\Lambda}_0+\mathbf{Hess}(\mathcal{L})(\bar{\Lambda}_0)^{-1}\nabla\mathcal{L}(\bar{\Lambda}_0)\big\}\Big] \tag{5.8}$$

leads to a quadratic approximation of the MLE. To relate this expression to the earlier mentioned Newton step in (5.5), note that in q, the inverse Hessian is multiplied by the gradient. A minimisation step of q is in fact a step in Newton's method as the error of the second order polynomial is minimised and thus a second order correct method. That means that the error is of order $\mathcal{O}(|\hat{\Lambda}_0-\bar{\Lambda}_0|^2)$. A similarity between q and isotonic regression (5.3) can be discovered. If $\mathbf{Hess}(\mathcal{L})(\bar{\Lambda}_0)$ is a diagonal matrix, then q takes the form

$$\sum_{i=1}^{m}\Big[\hat{\Lambda}_{0i}-\bar{\Lambda}_{0i}+\big[\mathbf{Hess}(\mathcal{L})(\bar{\Lambda}_0)_{ii}\big]^{-1}\nabla\mathcal{L}(\bar{\Lambda}_0)_i\Big]^2\mathbf{Hess}(\mathcal{L})(\bar{\Lambda}_0)_{ii}. \tag{5.9}$$

With proper choices of y_i, $r(x_i)$ and w_i, this equation fits perfectly in the theory of isotonic regression, see (5.2). Take $y_i=\hat{\Lambda}_{0_i}$ as the value of the maximiser which is unknown and let $r(v_j)=r_j=\bar{\Lambda}_{0i}-\big[\mathbf{Hess}(\mathcal{L})(\beta,\bar{\Lambda}_0)_{ii}\big]^{-1}\nabla\mathcal{L}(\beta,\bar{\Lambda}_0)_i$. Finally let $w_i=\big[\mathbf{Hess}(\mathcal{L})(\beta,\bar{\Lambda}_0)_{ii}\big]^{-1}$.

However, first the required derivatives need to be computed. These are found in equations (5.10) and (5.11). Recall that the the likelihood function is given by

$$\begin{aligned}\mathcal{L}(\beta,\Lambda_0) = \sum_{i=1}^{n}\Big[&\delta_i\log\Big(1-\exp(-\Lambda_0(t_i)e^{\beta^\top s_i})\Big)\\ &+\gamma_i\log\Big(\exp\Big(-\Lambda_0(t_i)e^{\beta^\top s_i}\Big)-\exp\Big(-\Lambda_0(u_i)e^{\beta^\top s_i}\Big)\Big)-\mu_i\Lambda_0(u_i)e^{\beta^\top s_i}\Big]\\ &:=\mathcal{L}_1(\beta,\Lambda_0)+\mathcal{L}_2(\beta,\Lambda_0)+\mathcal{L}_3(\beta,\Lambda_0).\end{aligned}$$

The first derivatives with respect to $\Lambda_0(v_j)$ are

$$\begin{aligned}\frac{\partial\mathcal{L}_1(\beta,\Lambda_0)}{\partial\Lambda_0(t_i)} &= \delta_i\frac{e^{\beta^\top s_i}}{\exp(\Lambda_0(t_i)e^{\beta^\top s_i})-1}\\ \frac{\partial\mathcal{L}_2(\beta,\Lambda_0)}{\partial\Lambda_0(t_i)} &= \gamma_i\begin{cases}\dfrac{e^{\beta^\top s_i}}{\exp((\Lambda_0(t_i)-\Lambda_0(u_i))e^{\beta^\top s_i})-1}\\ -e^{\beta^\top s_i}\end{cases}\\ \frac{\partial\mathcal{L}_2(\beta,\Lambda_0)}{\partial\Lambda_0(u_i)} &= \gamma_i\begin{cases}\dfrac{e^{\beta^\top s_i}}{\exp((\Lambda_0(u_i)-\Lambda_0(t_i))e^{\beta^\top s_i})-1}\\ \dfrac{e^{\beta^\top s_i}\exp(-\Lambda_0(u_i)e^{\beta^\top s_i})}{1-\exp(-\Lambda_0(u_i)e^{\beta^\top s_i})}\end{cases}\\ \frac{\partial\mathcal{L}_3(\beta,\Lambda_0)}{\partial\Lambda_0(u_i)} &= -\mu_i e^{\beta^\top s_i}.\end{aligned} \tag{5.10}$$

Note that also $\frac{\partial\mathcal{L}_1(\beta,\Lambda_0)}{\partial\Lambda_0(u_i)}=0$ and $\frac{\partial\mathcal{L}_3(\beta,\Lambda_0)}{\partial\Lambda_0(t_i)}=0$, because $\mathcal{L}_1$ and $\mathcal{L}_3$ do not depend on $\Lambda_0(u_i)$ and

$\Lambda_0(t_i)$ respectively. and the second derivatives are

$$\begin{aligned}
\frac{\partial^2 \mathcal{L}_1(\beta,\Lambda_0)}{\partial^2 \Lambda_0(t_i)} &= -\delta_j \frac{e^{2\beta^\top s_i}\exp(\Lambda_0(t_j)e^{\beta^\top s_i})}{(\exp(\Lambda_0(t_j)e^{\beta^\top s_i})-1)^2} \qquad (5.11)\\
\frac{\partial^2 \mathcal{L}_2(\beta,\Lambda_0)}{\partial^2 \Lambda_0(t_i)} &= -\gamma_j \begin{cases} \dfrac{e^{2\beta^\top s_i}\exp(-(\Lambda_0(u_j)-\Lambda_0(t_j))e^{\beta^\top s_i})}{(1-\exp(-(\Lambda_0(u_j)-\Lambda_0(t_j))e^{\beta^\top s_i}))^2} & \text{if } \Lambda_0(u_i)<\infty \\ 0 & \text{otherwise.} \end{cases}\\
\frac{\partial^2 \mathcal{L}_2(\beta,\Lambda_0)}{\partial^2 \Lambda_0(u_i)} &= -\gamma_j \begin{cases} \dfrac{e^{2\beta^\top s_i}\exp(-(\Lambda_0(u_j)-\Lambda_0(t_j))e^{\beta^\top s_i})}{(1-\exp(-(\Lambda_0(u_j)-\Lambda_0(t_j))e^{\beta^\top s_i}))^2} & \text{if } \Lambda_0(t_i)>0 \\ \dfrac{e^{2\beta^\top s_i}\exp(-\Lambda_0(u_j)e^{\beta^\top s_i})}{(1-\exp(-\Lambda_0(u_j)e^{\beta^\top s_i}))^2} & \text{if } \Lambda_0(t_i)=0 \end{cases}\\
\frac{\partial^2 \mathcal{L}_3(\beta,\Lambda_0)}{\partial^2 \Lambda_0(u_i)} &= 0
\end{aligned}$$

The Hessian does not have off-diagonal entries if $\delta_i = 1$ or $\mu_i = 1$. This is due to the fact that $\frac{\partial \mathcal{L}_1(\beta,\Lambda_0)}{\partial \Lambda_0(t_i)}$ does not depend on $\Lambda_0(v_k)$ if $v_k \neq t_i$. The same argument holds for $\mu_i = 1$. Due to this property, non-zero second derivatives with respect to different variables only occur when $\gamma_i = 1$, and only when derivatives are taken w.r.t. to $\Lambda_0(t_i)$ and $\Lambda_0(u_i)$ for fixed i. Off-diagonal elements corresponding to $\gamma_i = 1$ are

$$\mathbf{Hess}(\mathcal{L})(\beta,\Lambda_0)_{i\ell(i)} = \gamma_i \frac{\partial^2 \mathcal{L}_2(\beta,\Lambda_0)}{\partial \Lambda_0(u_i)\partial \Lambda_0(t_i)} = \frac{e^{2\beta^\top s_i}\exp(\{\Lambda_0(t_i)-\Lambda_0(u_i)\}e^{\beta^\top s_i})}{(1-\exp(\{\Lambda_0(t_i)-\Lambda_0(u_i)\}e^{\beta^\top s_i}))^2} = -\mathbf{Hess}(\mathcal{L})(\beta,\Lambda_0)_{ii} \tag{5.12}$$

For every $\gamma_i = 1$, two off-diagonal elements are non-zero in the Hessian. It should be clear that there are not many off-diagonal elements in the Hessian. Recall (5.9), because the Hessian is not a diagonal matrix, the minimiser of (5.9) is only an approximation of the minimiser of (5.7). By this approximation, the second order part of the approximation has an error, so the method becomes only first order correct with this additional approximation. This approximation cannot be expected to work without any justification. The justification is due to the working hypothesis (Groeneboom & Wellner, 1992) which is explained after the precise statement of the later introduced ICM algorithm. Now that the theory is established, it can be shown how isotonic regression is applied to (5.8). After the Hessian is approximated by its diagonal, the correct values for y_i, $r(x_i)$ and w_i given under (5.9) can be substituted into the cumulative sum diagram (5.4) which then becomes

$$P_0 = (0,0) \text{ and } P_j = \left(\sum_{i=1}^{j} \mathbf{Hess}(\mathcal{L})(\beta,\bar{\Lambda}_0)_{ii}, \sum_{i=1}^{j} \mathbf{Hess}(\mathcal{L})(\beta,\bar{\Lambda}_0)_{ii}\bar{\Lambda}_{0i} - \nabla \mathcal{L}(\beta,\bar{\Lambda}_0)_i\right). \tag{5.13}$$

for $1 \leq i \leq n$. Here $\bar{\Lambda}_{0i}$ means $\bar{\Lambda}_0(v_i)$. The left derivative of the greatest convex minorant is computed, resulting in numerical values of $r(x_i)$ which is an estimate of $\Lambda_0(x_i)$.

The ICM algorithm is in principle Newton's method where in each step isotonic regression is applied to do a linear approximation like the method in (5.5). Iterative use of isotonic regression itself is not enough to obtain the maximiser of $p\mathcal{L}_\beta$. Let $\bar{\Lambda}_0$ be the current iteration of a maximisation algorithm. Each iteration contains a minimisation problem, given by

$$B(\bar{\Lambda}_0) = \underset{\hat{\Lambda}_0 \in \mathcal{H}}{\operatorname{argmin}}\, q\left(\hat{\Lambda}_0, \bar{\Lambda}_0\right)$$

which is solved using isotonic regression. To see why this works, the statement of the following lemma from (Jongbloed, 1998) guarantees convergence of the algorithm.

Lemma 1. *Let $\mathcal{H}$ be a closed convex cone and $p\mathcal{L}_\beta : \mathbb{R}^m \to \mathbb{R}$ be a concave function such that it has a unique maximiser $\hat{\Lambda}_0$. Suppose also that $p\mathcal{L}_\beta$ is continuously differentiable on the set $\{\Lambda_0 \in \mathbb{R}^m : p\mathcal{L}_\beta(\Lambda_0) > -\infty\}$. If additionally $\bar{\Lambda}_0 \in \mathcal{H}\backslash\{\hat{\Lambda}_0\}$ such that $p\mathcal{L}_\beta(\hat{\Lambda}_0) > -\infty$, then for $1 \geq \epsilon_0 > 0$ sufficiently small*

$$p\mathcal{L}_\beta(\bar{\Lambda}_0 + \epsilon(B(\bar{\Lambda}_0) - \hat{\Lambda}_0)) > p\mathcal{L}_\beta(\bar{\Lambda}_0)$$

for all $\epsilon \in (0, \epsilon_0]$.

In words, $B(\bar{\Lambda}_0)$ is a search direction such that there exists an element between $\bar{\Lambda}_0$ and $B(\bar{\Lambda}_0)$ that has a larger log likelihood value $p\mathcal{L}_\beta$ than the current iterate $\bar{\Lambda}_0$. In this application, ϵ must be an element of the set $[0, 1]$. Suppose $\epsilon > 1$, then it may happen that $\bar{\Lambda}_0 + \epsilon(B(\bar{\Lambda}_0) - \bar{\Lambda}_0)$ which not an element of the cone. Take for instance $\bar{\Lambda}_0 \in \mathcal{H}$, strictly increasing, then $2\Lambda_0 \in \mathcal{H}$. It is clear that $\bar{\Lambda}_0 - 2\bar{\Lambda}_0$ is a decreasing function and hence not in $\mathcal{H}$. Lemma 1 guarantees the existence of an interval $[0, \epsilon_0]$ for which any $\epsilon \in (0, \epsilon_0]$ results in a decrease in the likelihood. Pseudocode of the algorithm is given in Algorithm 1.

```
Set $\varepsilon \in (0, \frac{1}{2})$, line search parameter;
Set $\eta > 0$, accuracy parameter;
Set $x^{(0)} \in \mathcal{C}$ such that $f(x^{(0)}) = \infty$;
while $[|x \cdot \nabla f(x)| > \eta$ or $|1 \cdot \nabla f(x))| > \eta$ or $\min_{1 \leq j \leq n} \sum_{i=j}^{n} (\nabla f(x))_i]$ do
    $\tilde{y} := \mathrm{argmin}_{y \in \mathcal{C}} (y - x + W(x)^{-1} \nabla f(x))^\top W(x)(y - x + W(x)^{-1} \nabla f(x))$;
    if $f(\tilde{y}) < f(x) + \varepsilon \nabla f(x)^\top (\tilde{y} - x)$ then
        $x := \tilde{y}$
    end
    else
        $\lambda := 1$, $s := \frac{1}{2}$. $z := \tilde{y}$;
        while $(f(\tilde{y}) < f(x) + (1 - \varepsilon) \nabla f(x)^\top (\tilde{y} - x))$ (I)
              $(f(\tilde{y}) > f(x) + \varepsilon \nabla f(x)^\top (\tilde{y} - x))$ (II) do
            if (I) then
                $\lambda := \lambda - s$
            end
            if (II) then
                $\lambda := \lambda + s$
            end
            $z := x + \lambda(\tilde{y} - x)$;
            $s := s/2$;
        end
    end
end
```

Algorithm 1: Pseudocode for the Iterative Convex Minorant

Application of the Armijo rule guarantees convergence of ICM. This rule is found in the second while loop in Algorithm 1. This rule roughly states how fast the iterates must result in de decrease of the objective function in term of its derivatives. It ensures that the function values of the iterates do decrease sufficiently as well as the iterates themselves are "different enough" from the previous one, preventing trapping behaviour. Outputs of the ICM algorithm for different sample sizes are found in Figure 7.1.

The working hypothesis states the following. Let $\hat{F}_n^{(k)}$ be the k'th iteration of the ICM algorithm. Normally, the ICM runs until its convergence criterion is reached. The working hypothesis states that the one-step estimator $\hat{F}_n^{(1)}$ and $\hat{F}_n$ are asymptotically equivalent. That is, suppose for some scaling sequence α_n, if it is true that $\alpha_n(\hat{F}_n^{(1)}(t_0) - F_0(t_0)) \to G$ in distribution then it is also true that $\alpha_n(\hat{F}_n(t_0) - F_0(t_0)) \to G$ in distribution. Here, G is a non-degenerate distribution. For this so-called one-step estimator, a limiting distribution is proven in (Groeneboom & Wellner, 1992). Because in the ICM algorithm, the Hessian is approximated by its diagonal to compute $\hat{F}_n^{(1)}$ in which the off-diagonal elements are not used. Asymptotically, the estimators $\hat{F}_n^1$ and $\hat{F}_n$ are equivalent. For large n this implies the information of the off-diagonal elements can be neglected when estimating the MLE $\hat{F}_n$ as its information is ignored in the ICM step. Even though the hypothesis remains unproven, it is supported by computer simulations. Under the assumption of the working hypothesis being true, the estimator obtained using the ICM algorithm is close to the true NPMLE.

Frankly, implementing the ICM is more than copying these lines into the programming language of preference. The next paragraph is devoted to the problems that are encountered in implementing the algorithm in the Cox model.

5.3 Difficulties Implementing the ICM Algorithm

This paragraph covers some issues that were encoutered during the implementation of the ICM algorithm that took time to be solved. Parametrising the time-to-event distribution in terms of the Cox model comes with problems making the implementation of algorithms less easy. Aside from the mathematical problems, some of them involve the limitations of the computer or used `R` packages. These difficulties are discussed in this paragraph together with their solution.

The isotonic regression step of the ICM algorithm makes is done using the function `gcmlcm` from the package `fdrtools` by (Klaus & Strimmer, 2012). Recall that isotonic regression returns the left derivative of the greatest convex minorant of a set of points. `gcmlcm` is a function that finds the right derivative of the cumulative sum diagram. After constructing the cumulative sum diagram stated in equation (5.13), it is not always possible to directly apply the function sometimes it is possible that multiple points of the cusum diagram share the same x-coordinate. This problem is caused by parametrisation of the Cox model. If an observation is right censored, then the diagonal element of the Hessian is zero. The cumulative sum diagram is given by

$$P_0 = (0,0) \text{ and } P_j = \left(\sum_{i=1}^{j} \mathbf{Hess}(\mathcal{L})(\beta, \Lambda_0^{(k)})_{ii}, \sum_{i=1}^{j} \mathbf{Hess}(\mathcal{L})(\beta, \Lambda_0^{(k)})_{ii} {\Lambda_0^{(k)}}_{ii} - \nabla\mathcal{L}(\beta, \hat{\Lambda}_0^{(k)})_i \right).$$

Suppose $\mu_j = 1$, then

$$\sum_{i=1}^{j} \mathbf{Hess}(\mathcal{L})(\beta, \Lambda_0^{(k)})_{ii} = \sum_{i=1}^{j-1} \mathbf{Hess}(\mathcal{L})(\beta, \Lambda_0^{(k)})_{ii}.$$

In this case, the y-coordinate decreases because $\frac{\partial\mathcal{L}(\beta,\Lambda_0)}{\partial\Lambda_0(u_i)} = -e^{\beta^\top s_k}$, which is always negative. If a coordinate on the x-axis is shared by multiple points, only the lowest value determines the slope for all these points. The two points at the x-axis at $x = 2$ in Figure 5.2 visualise this idea. When multiple subsequential right censored observations exist, then by the same argument, the last point, in terms of index, is the point that determines the slope. This insight allows to temporarily remove the points of the cusum diagram on the same x-coordinate except for the last before applying `gcmlcm`.

Not only right-censoring causes issues with `gcmlcm`. In some cases it occurs that second derivatives become very small, close to the machine precision. During the iterative process, the Hessian matrix evaluated to some iteration $\hat{\Lambda}_0^{(k)}$ can have very small diagonal elements. This caused some subsequential x-coordinates of the cusum diagram to become close to- or smaller than the machine precision which is approximately $2 \cdot 10^{-16}$. When this is the case, the two x-values are not distinguishable by the computer in regular settings. In this case, `gcmlcm` fails as described above. Setting Hessian elements that are close to the machine precision to a small value of 10^{-8} solves this issue while keeping the convergence rate.

5.4 Nelder-Mead Algorithm for Estimating Cox Coefficients

The previous paragraphs of this chapter were devoted to solve the profile likelihood. This resulted in the ICM algorithm which is able to compute $\hat{\Lambda}_0^\beta$ for any β up to computational limitations. However, the ICM algorithm can be time-consuming, especially when the sample size is large. For sample sizes larger than, say, $n = 4000$, maximising a single profile likelihood function takes more than half a minute when $\eta = 10^{-2}$ and the initial hazard function is set as the true hazard function.

A clever method is necessary to prevent the computation time from becoming unnecessarily high. The Nelder-Mead algorithm introduced in (Nelder & Mead, 1965) is a derivative-free method that can optimise a multidimensional function. The proposed algorithm explores the domain of β based on a chosen set of points. New iterates are found using the initial set of points. For each element $\beta^{(i)}$ in this set, the profile likelihood needs to be maximised. For each of these, an iterative procedure is required to find the corresponding hazard function $\hat{\Lambda}_0^{\beta^{(i)}}$ and $\mathcal{L}\big(\beta^{(i)}, \hat{\Lambda}_0^{\beta^{(i)}}\big)$ can be computed. Nelder-Mead will be applied together with profile-likelihood and the iterative convex minorant algorithm to find new β's that would improve upon the previous β's until the MLE $\hat{\beta}_n$ is found. Initial $d+1$ distinct candidates for β spanning $\mathbb{R}^d$ are determined as the initial set of βs to start the algorithm. These are denoted by $\beta^{(1)}, \ldots, \beta^{(d+1)}$, then for each $\beta^{(i)}$, the profile likelihood $p\mathcal{L}_{\beta^{(i)}}(\Lambda_0)$ is maximised to find $\hat{\Lambda}_0^{\beta^{(i)}}$ for $i = 1, \ldots, d+1$. These $(d+1)$ values for the log likelihood $\mathcal{L}\big(\beta^{(i)}, \hat{\Lambda}_0^{\beta^{(i)}}\big)$ are then compared, and based on the output, new candidates for β are generated. This means that for each candidate β, the ICM algorithm is applied to optimise the profile likelihood.

There are multiple derivative free optimisation algorithms. One commonly used algorithm is the Hooke-Jeeves algorithm published in (Hooke & Jeeves, 1961). A description of the algorithm can be found in the Appendix 8.0.2. This algorithm has been tested against the Nelder-Mead algorithm. In a speed comparison between the two algorithms, Nelder-Mead was faster because it required less function evaluations in most cases. Because it will be used in the actual simulations of chapter 7, the idea behind the Nelder-Mead algorithm is explained.

Consider an optimisation problem where a function $\varphi : \mathbb{R}^d \to \mathbb{R}$ needs to be maximised. The Nelder-Mead algorithm requires $d+1$ distinct initial points that span $\mathbb{R}^d$, denote these points by $x^{(0)} = (x_1^{(0)}, \ldots, x_{d+1}^{(0)})$. In order to update an iterate $x^{(k)}$ to $x^{(k+1)}$ the following procedure is followed. Figure 5.3 shows a visualisation of a single iteration of the algorithm. The values of $\varphi(x^{(k)}) = (\varphi_1^{(k)}, \ldots, \varphi_{d+1}^{(k)}) := (\varphi(x_1^{(k)}), \ldots, \varphi(x_d^{(k)}))$ are computed. The lowest and the two highest values are denoted using ordered statistics by $\varphi_{(1)}^{(k)}$, $\varphi_{(d)}^{(k)}$ and $\varphi_{(d+1)}^{(k)}$. Let x_L be the element of $x^{(k)}$ corresponding to $\varphi_{(1)}^{(k)}$, i.e., $\varphi(x_L^{(k)}) = \varphi(x_{(1)}^{(k)})$. Furthermore denote the best and second best values by $x_G^{(k)}$ and $x_B^{(k)}$, defined such that $\varphi(x_G^{(k)}) = \varphi_{(d)}^{(k)}$ and $\varphi(x_B^{(k)}) = \varphi_{(d+1)}^{(k)}$. Figure 5.3 illustrates a single iteration in the algorithm in $\mathbb{R}^2$. The algorithm works as follows. Fix a threshold for convergence $\gamma > 0$. The element $x_L^{(k)}$ yields the least value of $\varphi(x^{(k)})$ while $x_G^{(k)}$ and $x_B^{(k)}$ yield the highest values. Reflecting $x_L^{(k)}$ through the line between the two best points may lead to an increase of φ. Let $x_M^{(k)} = \frac{1}{2}(x_G^{(k)} + x_B^{(k)})$ be the mirror point between the two best points. Then $x_R^{(k)} = 2x_M^{(k)} - x_L^{(k)}$ is the reflection $x_L^{(k)}$ through $x_M^{(k)}$. Another point $x_E^{(k)}$ extend the search direction $x_R^{(k)}$ by defining $x_E^{(k)} = x_R^{(k)} + (x_M^{(k)} - x_L^{(k)})$.

If $\varphi(x_M^{(k)}) > \max(\varphi(x_R^{(k)}), \varphi(x_E^{(k)}))$, then φ decreases in function value compared to $x_M^{(k)}$, the algorithm will look closer near $x_M^{(k)}$. Two points are generated around $x_M^{(k)}$. Let $x_{M1}^{(k)} = \frac{1}{2}(x_M^{(k)} + x_R^{(k)})$ and $x_{M2}^{(k)} = \frac{1}{2}(x_M^{(k)} + x_L^{(k)})$, then $x_L^{(k)}$ is updated by $x_{M1}^{(k)}$ if $\varphi(x_{M1}^{(k)}) > \varphi(x_{M2}^{(k)})$ and is updated by $x_{M2}^{(k)}$ otherwise if $\varphi(x_M^{(k)}) \leq \max\big(\varphi(x_R^{(k)}), \varphi(x_E^{(k)})\big)$, then $x_L^{(k)}$ is updated by $x_R^{(k)}$. If $\varphi(x_R^{(k)}) > \varphi(x_E^{(k)})$ and is updated by $x_E^{(k)}$ otherwise. This replacement results in a new iteration $x^{(k+1)}$. The algorithm stops when $|\varphi(x_L^{(k)}) - \varphi(x_L^{(k+1)})| < \gamma$. Pseudocode for Nelder-Mead is found in Algorithm 2. Note

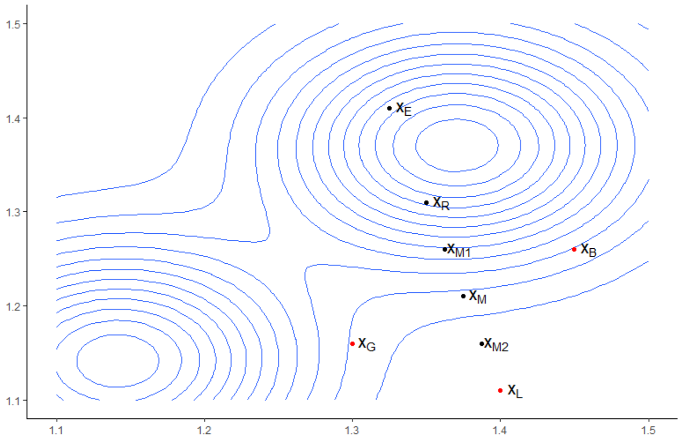

Figure 5.3: A contour plot is given. The contour lines represent function values where the smallest rings represent higher values. The points are plotted that are relevant for a single iteration of the Nelder-Mead algorithm.

that the if statements avoid unnecessary function evaluations of φ, saving calculation time. This is especially useful when computation of a single $\varphi^{(k)}$ is costly.

```
Data: Initial estimate $x^0 \in \mathbb{R}^d$ and a function to maximise $\varphi$
Generate $d+1$ distinct points $x^{(0)} = (x_1^{(0)}, \dots, x_{d+1}^{(0)})$;
Set $k = 0$;
Set $\gamma > 0$;
Set $\varphi(x^{(-1)}) = -\infty$;
while $\varphi(x^{(k)}) - \varphi(x^{(k-1)}) > \gamma$ do
    Calculate $\varphi_i = \varphi(x_i^{(k)})$ for all $i = 1, \dots, d+1$;
    Find $x_i^{(k)}$'s correspronding to $\varphi_{(1)}, \varphi_{(d)}$ and $\varphi_{(d+1)}$, call these $x_L^{(k)}, x_G^{(k)}, x_B^{(k)}$ respectively;
    Generate $x_M^{(k)} = \frac{1}{2}(x_G^{(k)} + x_B^{(k)})$, $x_R^{(k)} = x_M^{(k)} + (x_M^{(k)} - x_L^{(k)})$ and $x_E^{(k)} = x_R^{(k)} + (x_M^{(k)} - x_L^{(k)})$;
    if $\varphi(x_M^{(k)}) > \max(\varphi(x_R^{(k)}), \varphi(x_E^{(k)}))$ then
        Generate $x_{M1}^{(k)} = \frac{1}{2}(x_M^{(k)} + x_W^{(k)})$ and $x_{M2}^{(k)} = \frac{1}{2}(x_M^{(k)} + x_R^{(k)})$;
        if $\varphi(x_{M1}^{(k)}) > \varphi(x_{M2}^{(k)})$ then
            Replace $x_L^{(k)}$ with $x_{M1}^{(k)}$;
        else
            Replace $x_L^{(k)}$ with $x_{M2}^{(k)}$;
        end
    else
        if $\varphi(x_E^{(k)}) > \varphi(x_R^{(k)})$ then
            Replace $x_L^{(k)}$ with $x_E^{(k)}$;
        else
            Replace $x_L^{(k)}$ with $x_R^{(k)}$;
        end
    end
    Set $x^{(k+1)} := x^{(k)}$;
    Set $k \mapsto k+1$;
end
```

Algorithm 2: Pseudocode for the Nelder-Mead algorithm.

Chapter 6

The Bootstrap and Interval Censored Data

In the previous chapters theory and applications of survival analysis are explained. The chapters that followed introduced theory on the likelihood function for which no known explicit expression for the maximiser seems to exist. Algorithms are proposed to solve for the NPMLE iteratively. The solution for $(\hat{\beta}_n, \hat{\Lambda}_{0,n}) = \operatorname{argmax}_{\beta\in\mathbb{R}^d,\Lambda_0\in\mathcal{H}} \mathcal{L}(\beta, \Lambda_0)$ can be found using these methods. Because these approximations are made using a single sample, the output of the regression $(\hat{\beta}_n, \hat{\Lambda}_{0,n})$ is also a single estimate. Because the estimator is simply a function of the data, the estimator itself is random. This makes it natural to question the precision of the estimator.

Precision can be interpreted as a measure on how close estimates are to each other. A simple example is given to illustrate the idea. Let $Y_1, \dots, Y_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$. Calculating the mean $\bar{Y}_n = \frac{1}{n}\sum_{i=1}^n Y_i$ then gives the mean of the sample. It can be seen that $\bar{Y}_n$ is random as it is a sum of random variables, it is easily shown that $\bar{Y}_n \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$. Precision of the estimator $\bar{Y}_n$ is then given by $\frac{n}{\sigma^2}$. Randomness also appears in the NPMLE because it is generated using random variables. Precision of the NPMLE can be quantified using simulations.

In the IC2 model, no limiting distribution has been derived for the NPMLE of the pointwise function values $F(t_0)$. Under assumption of the unproven working hypothesis, introduced in (Groeneboom & Wellner, 1992), a limiting distribution is derived. Consider the current status model, an interval censoring model with only one inspection time per subject. Let F and G be the distribution functions of X and T respectively. The log likelihood function for the distribution function of the IC1 model is given by

$$\mathcal{L}_{IC1}(F) = \sum_{i=1}^{n} \delta_i \log F(t_i) + (1-\delta_i)\log(1-F(t_i)).$$

If the densities f and g of F and G are continuously differentiable at time t_0 such that $f(t_0), g(t_0) > 0$, then an asymptotic distribution is derived,

$$\Xi_n := n^{\frac{1}{3}}(\hat{F}_n(t_0) - F(t_0)) \to \kappa\mathbb{C} \tag{6.1}$$

in distribution, where $\kappa = [4F(t_0)(1-F(t_0))f(t_0)g(t_0)]^{\frac{1}{3}}$ and $\mathbb{C} = \operatorname{argmin}_{h\in\mathbb{R}}(W(h)+h^2)$ with W a standard two-sided Brownian Motion, see (Groeneboom & Wellner, 1992). This distribution can be used as a basis to compute asymptotic confidence intervals. It can however be difficult to estimate the values of $g(t_0)$ and $f(t_0)$. Theorem 3.1 in (Sen & Xu, 2015) shows an asymptotic distribution for IC2 data. Conditional on the inspection times, time evaluations of bootstrapped smoothed NPMLE $\tilde{F}_{nh}(t_0)$ yields consistent results. This method involves kernel smoothing and which is elaborated in paragraph 6.1. Limiting distributions are only known to be true under the working hypothesis.

Construction of confidence intervals based on this result involves estimation of $f(t_0)$ and $g(t_0)$. To prevent estimation of κ, the bootstrap will be used for computing confidence intervals.

As discussed above, in situation of the censoring model theoretical results on the MLE of the Cox coefficients β or function values of F are lacking in the IC2 model. Let $\theta = \theta(F)$ be the statistic of interest. This quantity can be estimated by $\theta(\bar{F}_n)$ where $\bar{F}_n$ is an estimate of the distribution function F. Without knowing the finite sample or asymptotic distribution of $\theta(\bar{F}_n)$, it becomes difficult to measure the accuracy of the initial estimate. Typically, when one knows the underlying distributions generating the data, Monte Carlo simulations can be used to approximate the distribution of an estimator. In general, Monte Carlo simulations cannot be done in practice to estimate the distribution of $\theta(\bar{F}_n)$ because it requires sampling from the true distribution which is unavailable. The bootstrap is a technique introduced in (Efron, 1979) that can be used to solve this issue at the cost of an error. Instead of Monte Carlo sampling from the true distribution F, this can be done from $\bar{F}_n$. Then Monte Carlo is used to estimate the distribution of $\theta(\bar{F}_n)$ by sampling new data from $\bar{F}_n$ multiple times and compute estimates $\theta(\bar{F}_n^*)$. The distribution of $\theta(\bar{F}_n^*)$ can then be approximated arbitrarily well when more estimates are computed. Note that this is not the distribution of $\theta(\bar{F}_n)$ but an approximation of it. As (Efron & Tibshirani, 1993) phrased it: "The bootstrap requires no theoretical calculations and is available no matter how complicated the estimator may be". The idea behind the bootstrap is to assume that the data can represent of the true underlying distribution well so that bootstrap samples generated using the observed data represent samples of the true distribution. These bootstrap samples can for example be made using the ECDF of the data. This is supported by the Glivenko-Cantelli theorem, stating that the ECDF converges to the true distribution see for example (Wasserman, 2006). With these bootstrap samples, the statistic of interest can be approximated as many times as needed.

The easiest case of a bootstrap sample is to consider data $Y = (Y_1, \dots, Y_n)$ sampled from the distribution function G. A sample $Y^* = (Y_1^*, \dots, Y_n^*)$ from the ECDF is generated by choosing for each entry of Y^*, one from Y with equal probability. This means that in a bootstrap sample, it is possible that some Y_k does not occur in the bootstrap sample, and some occur multiple times in Y^*. Given the statistic of interest, $\theta(G)$, a bootstrap realisation is calculated based on Y^* using the same calculations as $\theta(G)$. Let G_n^* be the ECDF made using Y^*, then as many realisations of $\theta(G_n^*)$ can be computed to find an approximation of the distribution of the bootstrap distribution.

Let $\hat{\theta}_n \to \theta$ in probability. Using the definition of (Sen & Xu, 2015), the bootstrap procedure is weakly consistent if $\hat{\theta}_n^* \to \hat{\theta}_n$ in probability as $n \to \infty$. One needs to be aware of the fact that the bootstrap does not work in certain situations. When an estimator $\hat{\theta}_n$ for θ is consistent does not mean that the bootstrap is consistent. The following example by Efron and Tibshirani (1993) illustrates this.

Example 7. *Let $X_1, \dots, X_n$ be independent samples of $X \sim Unif(0, \theta)$. The maximum likelihood estimator $X_{(n)} := max_{1 \leq i \leq n} X_i$ for θ converges to θ almost surely. By independence of the samples it follows that $\mathbb{P}(X_{(n)} \leq x) = \left(\frac{x}{\theta}\right)^n$ for $x \in [0, \theta]$. Then it holds also that*

$$\mathbb{P}(n(\theta - X_{(n)}) \leq x) = 1 - \mathbb{P}\left(X_{(n)} \leq \frac{x}{n} - 1\right) = 1 - \left(1 - \frac{x}{n\theta}\right)^n \to 1 - \exp(-x/\theta).$$

Therefore, the limiting distribution of $n(\theta - X_{(n)})$ is $Exp(\frac{1}{\theta})$. Given the same sample $X_1, \dots, X_n$, a bootstrap procedure can be followed by replacing θ by $X_{(n)}$ and resample $X_{(n)}^*$ by taking the maximum of a bootstrap sample.

$$\mathbb{P}(X_{(n)}^* = X_{(n)}) = 1 - \mathbb{P}(X_{(n)}^* \neq X_{(n)}) = 1 - \left(\frac{n-1}{n}\right)^n \to 1 - \exp(-1).$$

So that $\mathbb{P}(n(X_{(n)}^* - X_{(n)}) = 0) = 1 - \exp(-1)$. Hence, asymptotically both distributions are not the same. This can be seen easily by observing that the bootstrap distribution has a strictly positive mass at zero whereas the limiting distribution of $X_{(n)}$ is a continuous distribution. This implies that the bootstrap procedure is not consistent.

Kernel	$K(t)$
Rectangular	$\frac{1}{2}1_{\lvert t\rvert\leq 1}$
Biweight	$\frac{15}{16}(1-t^2)^2 1_{\lvert t\rvert\leq 1}$
Tricube	$\frac{70}{81}(1-\lvert t^3\rvert)^3 1_{\lvert t\rvert\leq 1}$
Epanechnikov	$\frac{3}{4}(1-t^2)1_{\lvert t\rvert\leq 1}$
Gaussian	$\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}t^2}$

Table 6.1: A table of commonly used kernels with the corresponding formulas.

6.1 Kernel Smoothing

The discrete nature of the NPMLE is not always desirable. It can be the case that the true distribution is expected to be continuous. One way to make a discrete function continuous is by a method called smoothing. To exemplify the idea of smoothing, first observe a discrete distribution given by $\check{F}_n = \sum_{i=1}^n w_i \eta_{\tau_i}$ such that $\sum_{i=1}^n w_i = 1$, see Figure 6.1. The underlying continuous density can be estimated by smoothing it. Here, η_τ is a Dirac measure with its mass at τ and τ_i is the i'th inspection time. Smoothing generates a continuous distribution based on $\check{F}_n$ by replacing the Dirac mixture to another mixture with the same weights, but with different measures which called are called *kernels*. A kernel is a continuous non-negative function K which integrates to one and has a centre of mass at zero, i.e., $\int_{\mathbb{R}} tK(t)dt = 0$. Kernels are usually symmetric. It can be thought of as a probability density function. Common functions used for smoothing are polynomial or Gaussian, see Table 6.1. The first four kernels in this table have a bounded support. The Gaussian kernel has an unbounded support. The kernel K is scaled by a bandwidth h that determines how much a (discrete) distribution is smoothed. Scaled kernels are expressed as $\frac{1}{h}K(t/h)$. The constant h^{-1} ensures that the rescaled kernel still integrates to one. Substituting $t = \frac{t'-\tau}{h}$ yields

$$\int K(t)dt = \int \frac{1}{h}K\left(\frac{t'-\tau}{h}\right)du = 1.$$

Smoothing a general discrete density function can be seen as convolution of a kernel with a Dirac mixture,

$$\check{f}_{nh}(t) = \int \frac{1}{h}K\left(\frac{t-\tau}{h}\right)d\check{F}_n(\tau) = \int \frac{1}{h}K\left(\frac{t-\tau}{h}\right)d\left(\sum_{i=1}^m w_i\eta_{\tau_i}\right) = \frac{1}{h}\sum_{i=1}^m w_i K\left(\frac{t-\tau_i}{h}\right). \quad (6.2)$$

The *bandwidth* h in (6.2) will turn out to be important in smoothing. This shows that for Dirac mixtures, the smoothed function becomes a sum of the rescaled kernels centred around the jumps of $\check{F}_n$ and the jump heights of $\check{F}_n$ determine the weights. The bandwidth determines the support or the rescaled kernels. This parameter needs optimisation. Choosing a good value for the bandwidth can result in better estimation of the distribution function, more details are found in paragraph 6.3.

Instead of estimating a density, the goal here is to estimate a smoothed distribution function. Since the distribution is simply the integrated density function it follows that

$$\begin{aligned}\check{F}_{nh}(t) &= \int_{-\infty}^t d\check{F}_{nh}(t') = \int_{-\infty}^t \check{f}_{nh}(t')dt' = \int_{-\infty}^t \int_{\mathbb{R}} \frac{1}{h}K\left(\frac{t'-\tau}{h}\right)d\check{F}_n(\tau)dt' \quad (6.3)\\ &= \int_{\mathbb{R}}\left[\int_{-\infty}^t \frac{1}{h}K\left(\frac{t'-\tau}{h}\right)dt'\right]d\check{F}_n(\tau) = \int_{\mathbb{R}}\mathbb{K}\left(\frac{t-\tau}{h}\right)d\check{F}_n(\tau) = \sum_{i=1}^m w_i\mathbb{K}\left(\frac{t-\tau_i}{h}\right),\end{aligned}$$

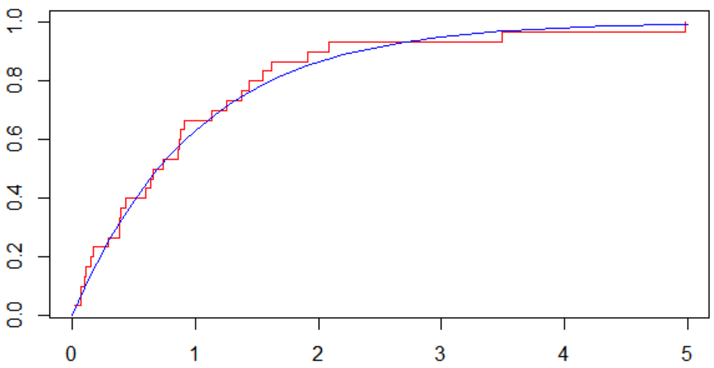

Figure 6.1: Empirical distribution function (red) of 30 samples of an Exp(1) distribution (blue).

using Fubini's theorem and differentiability of $\tilde{F}_{nh}$. The integrated kernel is denoted by $\mathbb{K}(t) = \int_{-\infty}^{t} K(t')dt'$.

Due to the nature of $\mathcal{F}$, the NPMLE is a Dirac mixture where the weights are generally not equal, that is, the NPMLE for F can be expressed $\hat{F}_n = \sum_{i=1}^{m} \omega_i \eta_{v_i}$ such that the weights sum to one. The smoothed version of the NPMLE (SMLE) $\hat{F}_n$ is defined as

$$\tilde{F}_{nh}(t) = \int_{\mathbb{R}} \mathbb{K}\left(\frac{t-v_i}{h}\right) d\hat{F}_n(v) = \frac{1}{h}\sum_{i=1}^{m} \omega_i \mathbb{K}\left(\frac{t-v_i}{h}\right). \tag{6.4}$$

A visualisation of the formula above is shown in Figure 6.2. When the domain of the true distribution F is the real line, i.e., it has no boundaries, then kernel smoothing with the proper bandwidth can reduce the error in the estimation. In the case of survival analysis, there is always a starting time of an experiment which is usually set to zero. Suppose that the Epanechnikov kernel is chosen, then

$$K(t) = \frac{3}{4}(1-t^2)1_{[-1,1](t)}, \tag{6.5}$$

and $\hat{F}_n$ has a jump close enough to the boundary $t = 0$ of the domain of F at t'. In this case, smoothing can put mass outside the domain of F, allowing unwanted behaviour of the estimator. This is the case when $1_{(\frac{t-t'}{h} \in [-1,1])} = 1$ and this can happen when $t' \in [0, h)$. Figure 6.3 shows what goes wrong when not applying a boundary correction.

Because kernel smoothing can place mass outside the domain, pointwise estimators can be inconsistent. Methods for correcting for this inconsistency at boundaries is studied in literature. A boundary correction method is proposed by (Groeneboom & Jongbloed, 2015) to prevent mass outside the true domain and will be used in the simulations. Let $[M_1, M_2]$ be the boundaries of the support of $\hat{F}_n$.

$$\tilde{F}_{nh}(t) = \int \left\{ \mathbb{K}\left(\frac{t-u}{h}\right) + \mathbb{K}\left(\frac{t+u+2M_1}{h}\right) - \mathbb{K}\left(\frac{2M_2-t-u}{h}\right) \right\} d\hat{F}_n(u) \tag{6.6}$$

When $t \in [M_1 + h, M_2 - h]$, equations (6.4) and (6.6) are equal, see (Groeneboom & Jongbloed, 2015). Applications in survival analysis often use that $M_1 = 0$ and the upper bound M_2 is infinite.

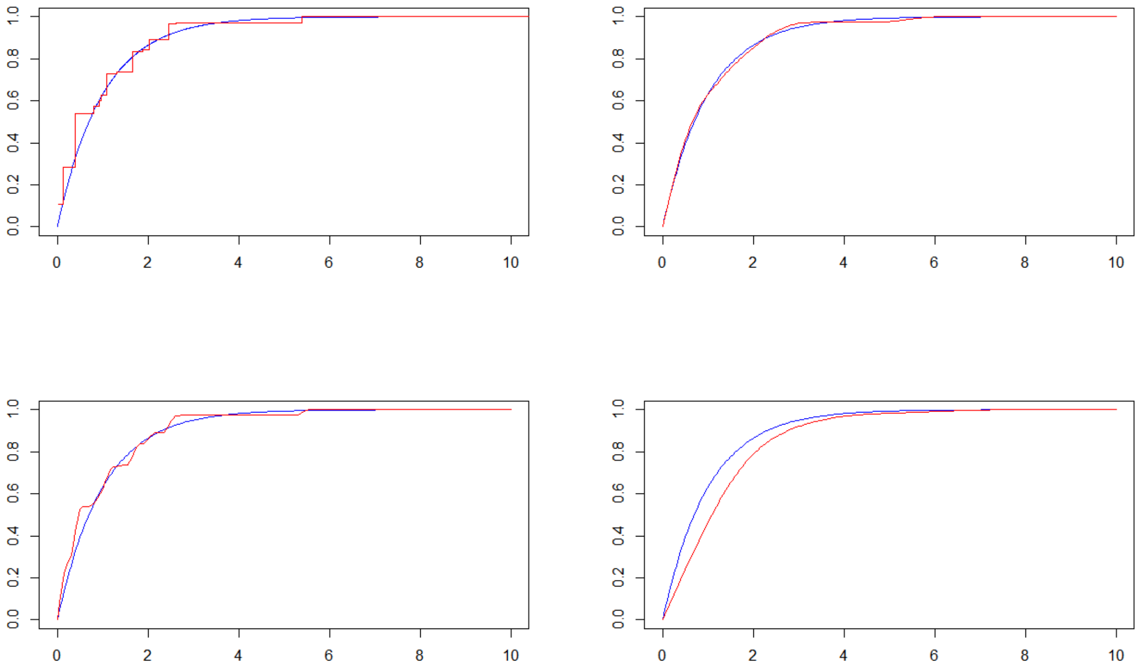

Figure 6.2: Using a sample $n = 500$, the NPMLE is computed. Top left shows the NPMLE (red) approximating the true distribution (blue in every figure). Top right is the SMLE with a proper choice of the bandwidth. On the bottom, the red curves are under and over smoothed estimates.

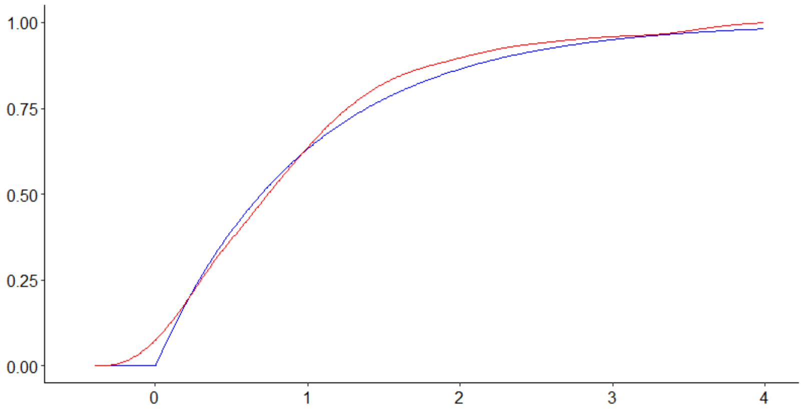

Figure 6.3: This figure shows that kernel smoothing can fail at the boundaries of the domain. Mass is found on the negative interval $(-h, 0)$ which is not allowed by definition of the survival function. The blue line is the true event-time distribution and the red line is an SMLE without boundary correction.

6.2 Bootstrap Methods

Multiple methods exist for generating bootstrap samples. Suppose some statistic of a real-valued distribution G is estimated. These statistics can be quantiles of G, the mean, the variance or any other functional $\theta(G)$. Let $Y = (Y_1, \ldots, Y_n)$ be independent samples from G. Take $\hat{G}_n$ to be the ECDF $\hat{G}_n(t) = \frac{1}{n}\sum_{i=n}^{n} \eta_{Y_i}$. The goal is to estimate $\theta(G)$, but only a sample of G is available. From the sample, $\theta(\hat{G}_n)$ is computed which itself is a random variable. The distribution of $\theta(\hat{G}_n)$ can be estimated using the finite sample Y. A bootstrap sample $(Y_1^*, \ldots, Y_n^*)$ drawn uniformly resample from the ECDF $\hat{G}_n$ from which a bootstrapped ECDF $\hat{G}_n^*$ can be constructed. By resampling the bootstrapped ECDF, the statistic $\theta(\hat{G}_n^*)$ can be computed as often as needed. This way the distribution of $\theta(\hat{G}_n)$ can be approximated to arbitrary precision for a fixed sample Y. This method is called the empirical bootstrap.

Another way to resample is called the parametric bootstrap. If it is assumed that G belongs to a parametric family, then its parameter π can be approximated according to the initial sample Y, generating an approximation $G_{\hat{\pi}}$ of G. The bootstrap would not be needed if the functional $\theta(G_{\hat{\pi}})$ is easy to compute. If the functional gets more complicated, Monte Carlo can be applied. New samples $(Y_1^*, \ldots, Y_n^*)$ are then drawn from $\hat{G}_{\hat{\pi}}$ to compute as many bootstrap samples $\theta(\hat{G}_n^*)$ as needed which are then used to approximate the distribution of $\theta(\hat{G}_n)$. Next to the parametric bootstrap, there exists another way of obtaining increasing continuous estimations of F by smoothing.

In the IC2 model, more resampling methods can be explored. The time-to-event distribution F is of interest so it is desirable to resample variables Z. Pairs of observations can be drawn uniformly from the initial sample. In this method, the n samples are seen as representative the true joint distribution of Z. This way, B bootstrap samples $Z_i^{*b} = (T_i^{*b}, U_i^{*b}, \Delta_i^{*b}, \Gamma_i^{*b})_{i=1}^n$ are generated by resampling pairs for $b = 1, \ldots, B$. The corresponding bootstrap measure becomes

$$\mathbb{P}^* = \sum_{i=1}^{n} M_i \eta_{Z_i}, \tag{6.7}$$

where M_j is the number of times that the observation z_j is drawn. By this interpretation, M is a multinomial distribution with parameters

$$M = (M_i, \ldots, M_n) \sim \text{Multinomial}(n, n^{-1}, \ldots, n^{-1}).$$

With these samples, bootstrapped NPMLEs $\hat{F}_n^{*b}$ can be computed as described in chapter 5. The empirical bootstrap is advocated in (Groeneboom & Hendrickx, 2017). Another method of resampling is to generate event times X^* from an estimate of the underlying distribution function F and use the inspection times T and U to determine the indicator variables Δ^*, Γ^*, M^* using (2.8). In selecting the inspection times T and U, one can choose the values originally observed (so one conditions on the inspections T and U) or one can sample from an estimate of the joint distribution of these. Based on the (smoothed) NPMLE, indicator variables are determined using the inspection times following the argument

$$\mathbb{P}(\Delta = 1) = \mathbb{P}(X < T) = F(T) \approx \tilde{F}_{nh}(T),$$

similarly,

$$\mathbb{P}(\Gamma = 1) = \mathbb{P}(T \leq X < T) = F(U) - F(T) \approx \tilde{F}_{nh}(U) - \tilde{F}_{nh}(T).$$

Consequently,

$$\mathbb{P}(M = 1) = 1 - \mathbb{P}(\Delta = 1) - \mathbb{P}(\Gamma = 1) \approx 1 - \tilde{F}_{nh}(U),$$

for some choice of h. New indicator variables are then drawn using the SMLE by

$$(\Delta_i^*, \Gamma_i^*, M_i^*) \sim \text{Multinomial}(1, \tilde{F}_{nh}(T_i), \tilde{F}_{nh}(U_i) - \tilde{F}_{nh}(T_i), 1 - \tilde{F}_{nh}(U_i)), \tag{6.8}$$

for $i = 1, \ldots, n$. This method is used in (Sen & Xu, 2015). It should be noted that their approach conditions on the inspection times T and U, implying that the results of the study can only be used if precisely the same times $(T_i, U_i)_{i=1}^n$ are in the dataset.

In this study the method by Sen and Xu is adapted. Given h, this resampling procedure is described by the following steps.

1. Using data $(T_i, U_i, \Delta_i, \Gamma_i)_{i=1}^n$, generate an estimator $\tilde{F}_{nh}$ for F.
2. From the existing oberservation times, calculate $\tilde{F}_{nh}(T_i)$ and $\tilde{F}_{nh}(U_i)$ for $i = 1, \ldots, n$.
3. Resample new indicator variables Δ_i^*, Γ_i^* and M_i^* for each $i = 1, \ldots, n$ using equation (6.8).
4. Using the new dataset $(T_i, U_i, \Delta_i^*, \Gamma_i^*)_{i=1}^n$, compute the NPMLE $\hat{F}_n^{*b}$.
5. Repeat steps (2)-(4) B times generating $\hat{F}_n^{*b}$ for $b = 1, \ldots, B$.

After these steps, the samples are used for making confidence intervals. However, Sen and Xu have concluded in Theorem 2.2 in their article that the bootstrap does not yield consistent results when using the bootstrap samples of the NPMLE. Because of this, the SMLE is used making it is important to estimate the bandwidth h before starting the actual bootstrapping procedure.

6.3 Bandwidth Selection

The problems occurring when selecting the unsuitable bandwidth have been discussed. In this paragraph a method is presented to find a proper bandwidth by minimising the mean squared error in (6.9) as a function of h. In many situations, h follows a power law dependent on n. In the cases of the ECDF or in the current status model the bandwidth is of the form $h = cn^{-\tau}$ for some constant c and $\tau > 0$. Results of this nature are given in (Wasserman, 2006) and (Groeneboom & Hendrickx, 2017). Proofs of such results in the Cox model have not been found in during study. No assumptions are made on a constant c and rate τ and h is found directly.

If the constant h is not chosen properly, estimators will have undesirable properties. Choosing the bandwidth h to be too small, then one speaks of under-smoothing, i.e., kernels have a too narrow peak in kernel density estimation. Roughly stated, the kernel smoothing function in that case almost like a point mass. Since the goal of kernel estimates is to smooth estimators, such behaviour is impractical. On the other hand there is over-smoothing when h is too big, then a kernel smears out the mass too much and local behaviour disappears. Both under- and oversmoothing are displayed in Figure 6.4. An SMLE with properly chosen bandwidth is shown in Figure 6.5. How such bandwidth is selected for the SMLE is the aim of this paragraph.

Instead of kernel density estimation, in this study, the distribution is being estimated by applying integrated kernels to the NPMLE. Because the smoother function is the integrated kernel $\mathbb{K}(t) = \int_{-\infty}^{t} K(t')dt'$, the problems occurring in kernel distribution estimation can be explained through observing under- and oversmoothing of density estimation. The narrow peaks of an undersmoothed NPMLE translate to a continuous SMLE but rather steep slopes. A case of oversmoothing in kernel distribution smoothing is depicted in Figure 6.2 on the bottom right image and bottom left for undersmoothing.

Related to this issue is the bias-variance trade-off. Let θ be the functional of interest and $\hat{\theta}$ an estimator for θ. Denote the error of the estimator $\hat{\theta}$ of θ by $\hat{\theta} - \theta$. The MSE is defined as the second moment of the error.

$$\mathrm{MSE}_\theta(\hat{\theta}) = \mathbb{E}_\theta(\hat{\theta} - \theta)^2.$$

A simple calculation shows that this quantity is directly related to the variance of the estimator and

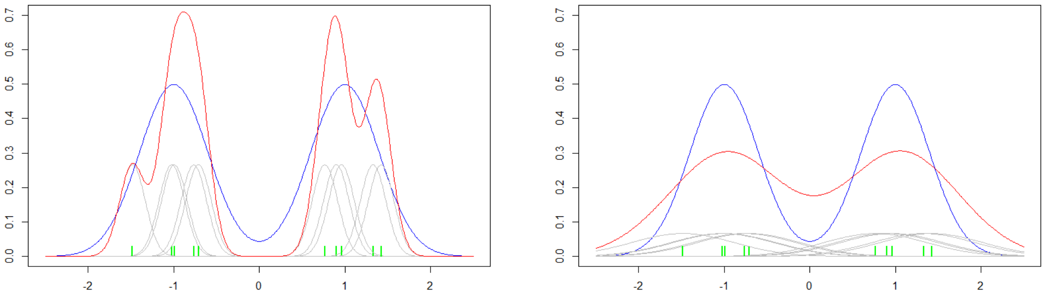

Figure 6.4: A mixture distribution $\big(\mathcal{N}(-1, 0.4) + \mathcal{N}(1, 0.4)\big)/2$ is used for illustration. Two kernel density estimations with the Epanechnikov kernel showing the importance of a well-chosen bandwidth. In the left image, $h = 0.15$ is chosen, which is too small. In the right image, the bandwidth $h = 0.6$ is chosen which is too large. Both estimators are clearly not good.

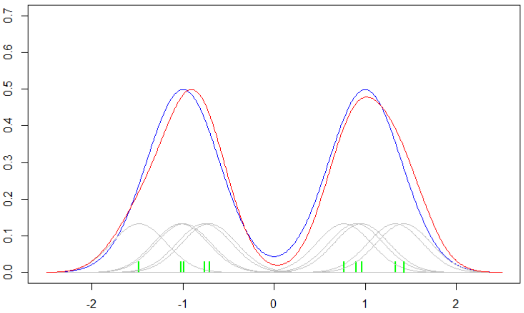

Figure 6.5: Kernel density estimation of the same data as used in Figure 6.4. Here, a proper bandwidth $h = 0.3$ is chosen.

its bias squared.

$$\begin{aligned}\mathrm{MSE}_\theta(\hat\theta) &= \mathbb{E}_\theta(\hat\theta-\theta)^2 = \mathbb{E}_\theta\left[(\hat\theta-\mathbb{E}_\theta(\hat\theta)) + (\mathbb{E}_\theta(\hat\theta)-\theta)^2\right] \\ &= \mathbb{E}_\theta(\hat\theta-\theta)^2 + 2\big[\mathbb{E}_\theta(\hat\theta)-\theta\big]\big[\mathbb{E}_\theta(\mathbb{E}_\theta(\hat\theta)-\hat\theta)\big] + (\mathbb{E}_\theta(\hat\theta)-\theta)^2 \\ &= \mathbb{V}_\theta(\hat\theta) + Bias_\theta(\hat\theta)^2,\end{aligned}$$

where $\mathbb{V}_\theta$ denotes the variance of $\hat\theta_n$ with respect to the distribution of θ. The bias is defined as the first moment of the error $Bias_\theta(\hat\theta) = \mathbb{E}_\theta(\hat\theta-\theta) = \mathbb{E}_\theta(\hat\theta)-\theta$, the average error. Finding the estimator with the lowest MSE should result in an estimator with a low bias and variance. When using the SMLE, the only variable is h so this can be chosen to minimise the MSE. Choosing the bandwidth too large introduces more bias but reduces the variance. It is argued in (Silverman, 1986) that in kernel density estimation of the ECDF, the variance of a kernel density estimate is of order $O(h^{-1})$, that is, the variance decreases as the bandwidth increases. On the other hand, the bias should be kept small in order to improve the accuracy of the estimator. Accuracy is obtained at the cost of variance, so a carefully selected bandwidth is important. Note that both the variance and the squared bias are positive, therefore, minimising the MSE makes both of these quantities smaller. Substituting $\hat\theta$ by an SMLE $\tilde F_{nh}$ at time t_0 and θ by the true F in the MSE yields

$$\mathrm{MSE}(h) = \mathbb{E}_F(\tilde F_{nh}(t_0) - F(t_0))^2. \tag{6.9}$$

In this case, the MSE becomes a function of h. The h minimising MSE(h) will be searched for as it results in the estimator with a small bias and variance and is denoted by h_{opt}. It should be clear that the true quantity cannot be calculated as F is unknown in practice. An estimate $\bar F_n$ of F is required for further estimation such that

$$\mathbb{E}_F(\tilde F_{nh}(t_0) - F(t_0))^2 \approx \mathbb{E}_{\bar F_n}(\tilde F_{nh}(t_0) - \bar F_n(t_0))^2. \tag{6.10}$$

Different options for $\bar F_n$ exist of which two will be discussed. If $\bar F_n$ is close to F, then the estimated MSE should be close to the true MSE.

A reasonable choice for $\bar F_n$ would be the NPMLE $\hat F_n$. However, (Sen & Xu, 2015) have shown that this choice would lead to inconsistent estimates in the current status model. Even though this does not necessarily imply inconsistency in the IC2 model it is decided not to use it. Options that could be more promising are an SMLE or a parametric estimation of F. The latter is useful when there is an indication whether the time-to-event is distributed according to some parametric family. In the simulations in chapter 7, samples are drawn from $X \sim Exp(\rho)$. Real world data does not always follow a parametric distribution. If it does, it may not always be easy to recognise a parametric family from the NPMLE for smaller sample sizes. If there is a parametric distribution that is close to the true F, it can nevertheless be used for modelling the true MSE. Keep in mind that this parametric assumption would then only be used for selecting the optimal bandwidth for the smoothing kernel and not for the final SMLE $\tilde F_{nh}$. If F does not belong to a parametric family, then this may result in poor modelling of the MSE. This would motivate a nonparametric approximation of F by using an SMLE $\tilde F_{nh_1}$ for some preselected h_1.

Like the parametric assumption, one can say that this approach may also lead to poor modelling of the MSE. As discussed earlier, there is under- and oversmoothing. The choice of h_1 can be arbitrary which can lead to improper smoothing. An SMLE is chosen to sample from to estimate the MSE. It is however difficult to find h_1 minimising the right hand side of (6.10).

A method to select h_1 is proposed. First the NPMLE $\hat F_n$ can be calculated. From the shape of the distribution, an appropriate parametric family can be selected. For multiple parametric families, parameter(s) $\hat\rho_n$ can be estimated to the data. Goodness of fit tests can then help to select the final parametric distribution $F_{\hat\rho_n}$ which is then used to model F in (6.10). The MSE approximated using $\bar F_n = F_{\hat\rho_n}$ is then approximated using Monte Carlo with the true underlying event-time distribution.

Fix a number of bootstrap samples B, then the BMSE approximates the right hand side of (6.10) by

$$\text{BMSE}_{par}(h) = \frac{1}{B}\sum_{b=1}^{B}\left(\tilde{F}_{nh}^{*b}(t_0) - F_{\hat{\rho}_n}(t_0)\right)^2 \approx \mathbb{E}_{F_{\hat{\rho}_n}}(\tilde{F}_{nh}(t_0) - F_{\hat{\rho}_n}(t_0))^2. \tag{6.11}$$

The bootstrap samples $\tilde{F}_{nh}^{*b}$ for $b = 1, \dots, B$ are generated by drawing event-time samples from $F_{\hat{\rho}_n}$ and generating new indicator variables, as described in chapter 6.2. Until this point, h is not fixed. The idea is that h is chosen to minimise BMSE_{par}, obtaining h_{par}. Then h_{par} can be used to compute the SMLE $\tilde{F}_{nh_{par}}$ that will be used for estimation of the final bandwidth h_{smo}. Let

$$h_{par} = \operatorname*{argmin}_{h\in\mathbb{R}^+} \text{BMSE}_{par}(h). \tag{6.12}$$

A bandwidth is obtained that does not require an arbitrary choice but is instead data-driven. From this point, the method of (Sen & Xu, 2015) is followed. Instead of a parametric distribution to generate event-times from, $\bar{F}_n = \tilde{F}_{nh_{par}}$ is used to resample event times. Then the BMSE using this function is defined as

$$\text{BMSE}_{smo}(h) = \frac{1}{B}\sum_{b=1}^{B}\left(\tilde{F}_{nh}^{*b}(t_0) - \tilde{F}_{nh_{par}}(t_0)\right)^2 \approx \mathbb{E}_{\tilde{F}_{nh_{par}}}\left((\tilde{F}_{nh}(t_0) - \tilde{F}_{nh_{par}}(t_0))\right)^2. \tag{6.13}$$

To obtain the functions $\tilde{F}_{nh}^{*b}$, new samples need to be drawn from $\tilde{F}_{nh_{par}}$ and new indicator variables are made using the multinomial distribution found in (6.8). As in (6.12), h_{smo} is obtained by minimising BMSE_{smo}, which is given by

$$h_{smo} = \operatorname*{argmin}_{h\in\mathbb{R}^+} \text{BMSE}_{smo}(h). \tag{6.14}$$

A step by step description to the above method is formulated the following:

1. Fix an equidistant grid $\mathcal{G}$ of candidates for h. Using the data, propose an estimate $\bar{F}_n$ of F
2. Using $\bar{F}_n$, generate new indicator variables $\Delta_i^{*b}, \Gamma_i^{*b}$ for $i = 1, \dots, n$ and compute the NPMLE $\hat{F}_n^{*b}$. Repeat this step B times.
3. For each $h \in \mathcal{G}$, compute $\tilde{F}_{nh}(t_0)$.
4. Using equation (6.11) or (6.13) for all $h \in \mathcal{G}$ and choose $h_{opt} = \text{argmin}_{h\in\mathcal{G}} \text{BMSE}(h)$.

Once $h_{smo} = \text{argmin}_{h\in\mathcal{G}} \text{BMSE}_{smo}(h)$ is computed using the steps above, the final estimator $\tilde{F}_{nh_{smo}}(t_0)$ is calculated. From this point, bootstrap samples can be drawn to do inference. For example, approximating the distribution of estimators of generating bootstrapped confidence intervals. The latter is done in the chapter 7.

If in the log-likelihood, F is replaced by a distribution function of a parametric family, its parameter ρ can be estimated from censored data. Because the distribution function and cumulative hazard function are one-to-one. The log-likelihood function is maximised for β and Λ_0 and the estimator $\hat{\Lambda}_{0n}$ is transformed to $\hat{F}_n$ according to (2.5). Estimating of ρ requires to reparametrise the log-likelihood by replacing Λ_0 by the cumulative hazard function of the assumed parametric family. In practical application of survival analysis, it is common to assume that F_ρ follows a general distribution like the Weibull distribution. In our simulation it is known that samples are drawn from the exponential distribution so only a single parameter is estimated for simplicity. To estimate ρ, use that the cumulative hazard function of $\text{Exp}(\rho)$ is $\Lambda(t) = \rho t$. Then the log-likelihood (4.6) becomes

$$\begin{aligned}\mathcal{L}_{parametric}(\beta,\rho) = \sum_{i=1}^{n}\Big[&\delta_i \log\left(1 - \exp(-\rho t_i e^{\beta^\top s_i})\right) \\ &+\gamma_i \log\left(\exp\left(-\rho t_i e^{\beta^\top s_i}\right) - \exp\left(-\rho u_i e^{\beta^\top s_i}\right)\right) - \mu_i \rho u_i e^{\beta^\top s_i}\Big].\end{aligned}$$

Finding the maximising argument $(\hat{\beta}_n, \hat{\rho}_n)$ or $\mathcal{L}_{parametric}$ becomes an optimisation problem in $\mathbb{R}^{d+1}$ since $\beta \in \mathbb{R}^d$ and $\rho \in \mathbb{R}$. This can be solved easily using only the Nelder-Mead algorithm, or again the profile likelihood using Nelder-Mead on β and `optimize` function in `R` compute the optimal ρ for each β. With this procedure $\hat{\rho}_n, \hat{\beta}_n$ is estimated as $\text{argmax}_{(\beta,\rho)\in\mathbb{R}^{d+1}} \mathcal{L}(\beta, \rho)$. The estimated distribution function becomes $\bar{F}_n(t) = 1 - \exp(-\hat{\rho}_n t)$ from which one can sample. Now the methodology of the bootstrap is set-up and the bandwidth can be selected. Before continuing with simulations, some information is given about confidence intervals.

6.4 Confidence Intervals

From data Z, only a single estimate is obtained. Because Z is only a sample of the true population, an error is made. Confidence intervals can reveal insight on the sampling error. These intervals consist of a random lower and upper bound such that with a given probability $1 - \alpha$, the true parameter lies between the random bounds. Often, $\alpha = 5\%$ is chosen. Confidence intervals are used to express uncertainty of an estimator. When from different datasets, 95% confidence intervals are generated, then it should be true that for approximately 95% of the intervals, the true parameter is included in the interval. The percentage of intervals in which the parameter is included is called the coverage. If the promised coverage is achieved, the intervals can be informative about the uncertainty.

Suppose that one attempts to estimate a probability of some event. The estimate always lies in the interval $[0, 1]$. From the experiment, a 95% confidence interval is given of $[0.02, 0.97]$. Because the interval is very wide, not much information is given as estimate do not seem concentrate anywhere in the interval $[0, 1]$. An interval with correct coverage of $[0.5, 0.55]$ would however be more informative because it is narrower. With 95% confidence, the estimate is contained in this interval, expressing confidence on in which range the true parameter is located. Narrow intervals suggest lower variability of the estimator.

Confidence intervals cannot be expected to work in all cases. Asymptotic results are usually applied to derive them, see for example (6.1). Because asymptotic results are used, it is not surprising that they cannot guarantee the coverage they should have. This is especially the case when the sample size is low. It can be shown that if asymptotic results cannot be used due to having a low sample size, these intervals can fail. Limiting distributions have been proven for the current status model, and under the working hypothesis, also for the interval censoring case 2 model. For the current status model, (Groeneboom & Hendrickx, 2017) have proven a limiting distribution under the unconditional bootstrap. It is used to generate confidence intervals in the current status model. Let $h_n \sim cn^{-1/5}$ for some fixed c, then

$$n^{2/5} \left[\tilde{F}^*_{nh}(t_0) - \int \mathbb{K}\left(\frac{t-u}{h}\right) d\tilde{F}_{nh}(u) \right] \to \mathcal{N}(0, \sigma^2).$$

Other asymptotic results can also be used to construct asymptotically correct confidence intervals. Let

$$\Xi^*_n = n^{-1/3}\left(F^*_n(t_0) - F(t_0)\right) \to \kappa\mathbb{C},$$

where κ is a non-negative constant and $\mathbb{C}$ is a symmetric continuous distribution. Denote the lower boundary of an interval by $\theta_{(lo)}$ and the upper bound by $\theta_{(up)}$. From this asymptotic result, it follows that

$$\begin{aligned} \mathbb{P}(\theta_{(lo)} \leq \kappa\mathbb{C} \leq \theta_{(up)}) &\approx \mathbb{P}(L \leq n^{1/3}(F^*_n(t_0) - F_n(t_0)) \leq U) \\ &= \mathbb{P}\left(\hat{F}_n(t) - n^{-1/3}L \leq F(t_0) \leq \hat{F}_n(t) - n^{-1/3}U\right). \end{aligned}$$

This result is used by (Banerjee & Wellner, 2005). Let $\hat{Q}^*_n$ be bootstrap samples of γ^*_n, as a consequence of this calculation the confidence interval becomes

$$[\theta_{(lo)}, \theta_{(up)}] = \left[\hat{F}_n(t_0) - n^{-1/3}\hat{Q}_{(1-\alpha/2)}, \hat{F}_n(t_0) - n^{-1/3}\hat{Q}_{(\alpha/2)}\right],$$

where $\hat{Q}_{(\alpha B)}$ is a consistent estimator or the α'th quantile of Q. This interval is also used by Sen and Xu, except their estimator is the SMLE, (Sen & Xu, 2015). Asymptotic results like these have not yet been proven within the Cox model. Because of this, more general standard confidence intervals such as the percentile bootstrap or studentised intervals are generated. The remainder of this chapter will cover these types of intervals.

Using the observations (T, U, Δ, Γ), estimators for the baseline distribution and the Cox-coefficients are computed. In the previous chapter multiple resampling techniques are discussed. These methods involve drawing pairs from the original sample, and also using a smoothed version of the MLE to obtain new indicator variables. Applying the previously introduced algorithms results in bootstrap samples of the SMLE. For fixed times t_0 one can construct confidence intervals for $F(t_0)$.

Denote the residual by $R_n = \hat{\theta}_n - \theta$ and define the function $H(r) = \mathbb{P}_R(R_n \leq r)$. A $100 - \alpha\%$ confidence interval is constructed by finding an interval $[a, b]$ such that $H(R_n \leq a) \approx \alpha/2$ and $H(R_n \leq b) \approx 1 - \alpha/2$. This can be done by generating B bootstrap samples so that the distribution of R_n can be approximated. The estimate $\tilde{F}_{nh}$ is assumed to be a good representation of the underlying F. This makes the assumption that the distributions of $\hat{\theta} - \theta$ is approximately the distribution of $\hat{\theta}^* - \hat{\theta}$. The percentile bootstrap confidence interval will now be constructed. let $\hat{H}$ be the cumulative distribution function of R_n^*.

Because $H(r) \approx \hat{H}(r) = \frac{1}{B}\sum_{b=1}^{B} 1_{R_n^{*b} \leq r} = \frac{1}{B}\sum_{b=1}^{B} 1_{R_n^{*(b)} \leq r}$, quantiles of R_n can be expressed in terms of the index of ordered statistics. This becomes clear if consider $H(R^*_{(\alpha B/2)})$ is considered such that $\alpha B/2$ is an integer, then there exist precisely $\alpha B/2$ values of $\hat{r}^*$ that are smaller or equal to $r^{*(\alpha B/2)}$. It follows that

$$\hat{H}(r^{*(\alpha B/2)}) = \frac{1}{B}\sum_{b=1}^{B} 1_{R_n^{*(b)} \leq r^{*(\alpha B/2)}} = \frac{\alpha B/2}{B} = \frac{\alpha}{2}. \tag{6.15}$$

Then the inverse of H is defined by $\hat{H}^{-1}_{(\alpha)} = r^*_{(\alpha B)}$.[1] The following computation approximates the quantiles $r^{(\alpha)}$ and $r^{(1-\alpha)}$. Start by finding r such that $\alpha \leq H(r)$, approximate H by $\hat{H}$, then $\hat{H}^{-1}(\alpha/2) \leq r$ from here it follows

$$\begin{aligned}\hat{H}^{-1}(\alpha/2) \leq r \implies \hat{r}^{*(\alpha B/2)} \leq r &\implies (\hat{\theta}^* - \hat{\theta})^{(\alpha B/2)} \leq \hat{\theta} - \theta \\ &\implies \hat{\theta}^{*(\alpha B/2)} - \hat{\theta} \leq \hat{\theta} - \theta \implies \hat{\theta}^{*(\alpha B/2)} - 2\hat{\theta} \leq -\theta.\end{aligned}$$

Then $\theta \leq 2\hat{\theta} - \hat{\theta}^{*(\alpha B/2)}$. Similarly one can find that $\hat{H}((1-\alpha)/2) \leq r \implies 2\hat{\theta} - \theta^*_{((1-\alpha)/2)}$. The $1 - \alpha$ confidence interval is then given by

$$[\theta_{(lo)}, \theta_{(up)}] = [2\hat{\theta} - \hat{\theta}^{*((1-\alpha/2)B)}, 2\hat{\theta} - \hat{\theta}^{*(\alpha B/2)}], \tag{6.16}$$

such that approximately, $\mathbb{P}(\theta \in [\theta_{(lo)}, \theta_{(up)}]) \approx 1 - \alpha$. This construction of confidence intervals is used in the chapter on simulations. It is possible for percentile confidence intervals of $F(t_0)$ to include elements outside the set $[0, 1]$. Because the true value of $F(t_0)$ cannot be outside of $[0, 1]$, it makes sense to restrict the interval $[\theta_{(lo)}, \theta_{(up)}]$ to $[0, 1]$ by updating the interval to $[\theta_{(lo)}, \theta_{(up)}] \cap [0, 1]$.

A bootstrap method for constructing confidence intervals that has a stronger accuracy at the cost of more computation power is the studentised confidence interval. If instead a confidence interval is based on the residual $R_n = \hat{\theta}_n - \theta$, an interval is constructed based on the residual divided by the approximated standard error of the estimate. The studentised bootstrap makes use of the standardised form. Let Z be the initial sample and Z^* a bootstrap sample according to the method of (Sen & Xu, 2015). The quantity of interest is

$$\mathcal{Z}^*_b = \frac{\theta(Z^*_i) - \theta(Z)}{\hat{se}(\theta(Z^*_b))}.$$

[1] Note that this inverse is actually in an interval. The minimum of the interval is taken as the inverse.

The denominator is what makes studentised confidence interval so computationally expensive, especially when the statistic itself is difficult to compute. Each bootstrap sample Z^* of the data generates a sample of the statistic. The studentised interval is constructed such that each bootstrap sample of the statistic is standardised. To do so, the term $\theta(Z_i^*) - \theta(Z)$ needs to be divided by the standard error of $\theta(Z_i^*)$. This means that bootstrap samples of Z^* need to be drawn and the statistic for needs to be computed multiple times to estimate the standard error of the sample. For clarity, this means that for each separate bootstrap sample, one needs to bootstrap, say B_1 times from the bootstrap sample to estimate its standard error. Let Z^{**} a bootstrap sample from Z^* according to the method of (Sen & Xu, 2015). The standard error for B_1 resamples of the bootstrap sample is given by

$$\widehat{se}^*(\theta(Z^*)) = \left(\frac{1}{B_1 - 1} \sum_{b=1}^{B_1} \left[\theta(Z_{i,b}^{**}) - \overline{\theta(Z^{**})} \right]^2 \right)^{\frac{1}{2}}.$$

Suppose $B_1 = 30$ to estimate the standard error of the bootstrap samples accurately enough. If one wants to base a confidence interval on $B = 1000$ bootstrap samples, 30000 total samples are needed. This is a strong limitation when computing a single sample is time consuming. If the computation is feasible, then the studentised confidence interval is given by

$$\left[\ \theta(Z) - \mathcal{Z}_{(1-\alpha/2)}\widehat{se}, \theta(Z) - \mathcal{Z}_{(\alpha/2)}\widehat{se}\ \right],$$

where $\widehat{se}$ is the standard deviation of $\theta(Z)$ defined by

$$\widehat{se}(\theta(Z)) = \left(\frac{1}{B - 1} \sum_{b=1}^{B} \left[\theta(Z_{i,b}^{*}) - \overline{\theta(Z^{*})} \right]^2 \right)^{\frac{1}{2}}.$$

The studentised confidence interval is used in literature on the current status model, see for example (Groeneboom & Hendrickx, 2017). Due to the complexity that is added by the Cox model, this is not feasible. Therefore, in the simulation section, only percentile intervals are generated.

Chapter 7

Simulations

7.1 Sampling and Simulation Setup

After implementation of the algorithms, it is natural to check how they perform in practice. Testing performance of the algorithms and accuracy of estimators is central to this chapter. Before applying the algorithms to actual data, computer generated data samples are used to investigate the behaviour of the SMLE. First a lemma is proven that helps to understand how samples from the Cox model can be generated.

Lemma 2. *Let $X \sim F$ with density f. Let g be an strictly monotone invertible function such that g^{-1} is differentiable. Then the density function of $g(X)$ is given by*

$$h(t) = f(g^{-1}(t)) \left| \frac{d}{dt} g^{-1}(t) \right| .$$

Proof. The assumptions in the statement are used. A simple calculation gives the distribution function H. The density is then found through the derivative, which exists by assumption. First it is assumed that g is strictly increasing, then

$$H(t) = \mathbb{P}(g(X) \leq t) = \mathbb{P}(X \leq g^{-1}(t)) = F(g^{-1}(t)).$$

Differentiating both sides by applying the chain rule yields the desired result.

$$h(t) = \frac{d}{dt} H(t) = \frac{d}{dt} F(g^{-1}(t)) = f(g^{-1}(t)) \frac{d}{dt} g^{-1}(t). \tag{7.1}$$

If g is strictly decreasing one must observe that if $x \leq y$, then $g(x) \geq g(y)$, so that $H(t) = \mathbb{P}(g(X) \leq t) = \mathbb{P}(X \geq g^{-1}(t)) = 1 - F(g^{-1}(t))$. Differentiating as above and combining the result with (7.1) yields

$$h(t) = \frac{d}{dt} H(t) = \frac{d}{dt} F(g^{-1}(t)) = f(g^{-1}(t)) \left| \frac{d}{dt} g^{-1}(t) \right| ,$$

completing the proof. □

This lemma makes sampling in the Cox hazard model easier. Let $U \sim \text{Unif}(0,1)$ and $g(x) = -\log(x)$ be a decreasing function. Clearly $g^{-1}(x) = e^{-x}$. Applying this to Lemma 2 the following is obtained

$$h(t) = 1_{[0,1]}(e^{-t})(|-e^{-t}|) = e^{-t} 1_{[0,\infty)}(t)$$

which is the density function of the standard exponential distribution, so $-\log(1-U) \overset{d}{=} V$ where $V \sim Exp(1)$ and $\overset{d}{=}$ means that two objects follow the same distribution. The fact that $1-U$ follows the same distribution as U is not surprising due to its symmetry around $\frac{1}{2}$. This lemma helps to formalise this statement. These two facts are used for the sampling in the Cox hazard model.

The uniform distribution is used to sample event times in the Cox model. Given the true event-time distribution F, a sample is drawn by taking the inverse of F of a uniform sample, i.e., $F^{-1}(U)$. This is easily shown by the following. Let $\mathbb{P}_U$ be the measure of U, then

$$\mathbb{P}_U(F^{-1}(U) \leq t) = \mathbb{P}_U(U \leq F(t)) = F(t).$$

The above equation implies that a sample of F is drawn by taking a realisation u of a uniform random variable U and take $F^{-1}(u)$. This fact is also useful for sampling from the SMLE. First F is expressed in terms of the cumulative hazard function. If one wishes to sample from the Cox model, the inverse of F needs to be computed.[1] Fix the Cox coefficients β and covariates s, then

$$u = F(x) = 1 - S(x) = 1 - \exp(-\Lambda(x|s)) = 1 - \exp\{-\Lambda_0(x)\exp(\beta^\top s)\}, \tag{7.2}$$

by equation (2.5). Inverting (7.2) to isolate x yields

$$x = \Lambda_0^{-1}\left(\frac{-\log(1-u)}{\exp(\beta^\top s)}\right),$$

when u is not a realisation but the standard uniform random variable, it follows

$$X \overset{d}{=} \Lambda_0^{-1}\left(\frac{-\log(1-U)}{\exp(\beta^\top s)}\right) \overset{d}{=} \Lambda_0^{-1}\left(\frac{V}{\exp(\beta^\top s)}\right). \tag{7.3}$$

The right hand side of (7.3) is used for sampling. As stated in Chapter 2, common distributions used in Survival analysis are Weibull, log-logistic, generalised Gamma. The distribution function of a Weibull random variable is given by

$$F_{\rho,k}(t) = 1 - \exp\left(-\left(\frac{t}{\rho}\right)^k\right).$$

By equation (2.3), it follows

$$\Lambda(t) = -\log(1 - F(t)) = \left(\frac{t}{\rho}\right)^k.$$

For sampling event-times, (7.3) states that the inverse of the cumulative hazard function is required. For the Weibull, the inverse is given by

$$\Lambda^{-1}(t) = \rho t^{\frac{1}{k}}.$$

The functions above correspong to the Exp($\frac{1}{\rho}$) distribution when $k = 1$. In the simulation done in this chapter, the exponential distribution will be used for its explicit formula of the inverse cumulative hazard function. In each simulation the setup (1)-(4) is as follows unless otherwise specified

1. $X \sim Exp(1)$, the time-to-event distribution.

2. $T \sim Exp(1/3)$, the first inspection time.

3. $U \sim T + Exp(2)$, the second inspection time.

4. $\beta = (0, 0)$, the Cox coefficients.

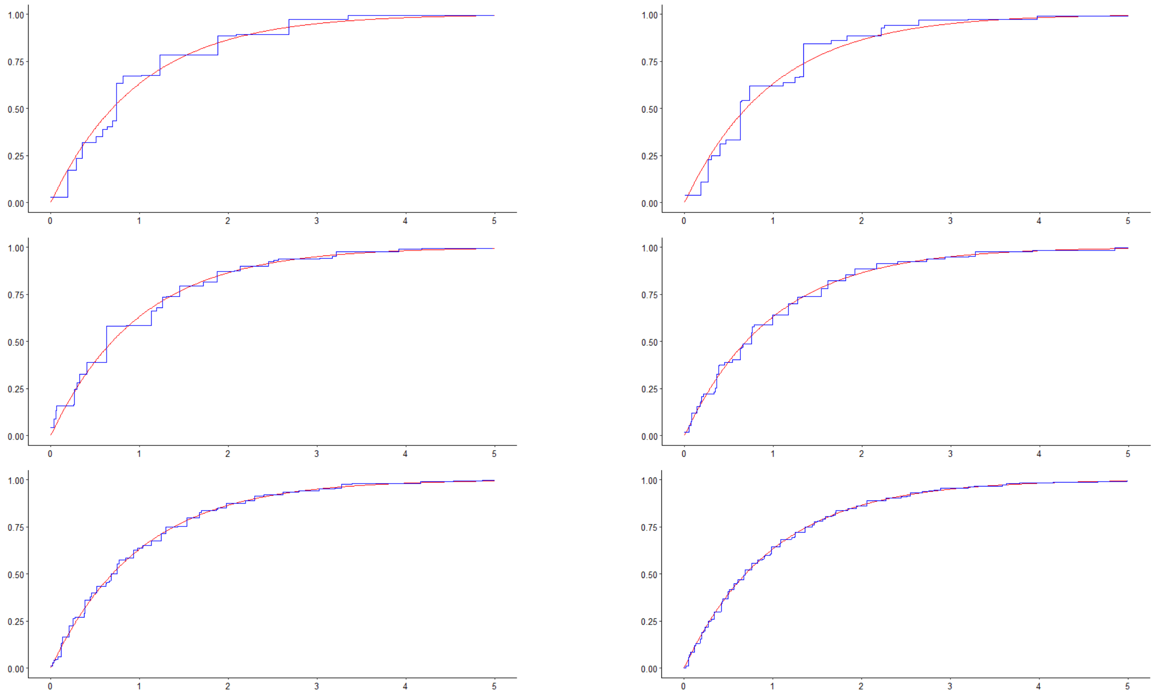

Figure 7.1: From top left to bottom right, NPMLE (blue) with sample sizes n of 500, 1000, 2000, 4000, 8000 and 16000 approximate the true distribution (red) better as n increases. The x-axis ranges from 0 to 5.

7.2 Simulations on Bandwidth Selection in the Cox Model

Example 1 in (Van der Vaart & Wellner, 2000) it is shown that under mild conditions the NPMLE converges to the true event-time distribution. Figure 7.1 contains a six realisations of the NPMLE for different sample sizes to illustrate that larger samples tend to have better estimates. The figure shows that, as the sample size increases, the estimator gets closer to the true distribution, this supports their result on convergence.

In 6.3 an almost automated method for selecting the bandwidth is described minimising the MSE. The performance is tested in this chapter. Samples of sizes $n = 500$, $n = 1000$ and $n = 2000$ are drawn according to the sampling scheme (1)-(4). Because the MSE in (6.9) is approximated first using a fitted exponential distribution, the model is well-specified in the sense that the approximation of the MSE by BMSE_{par}, uses the parametric family containing F to approximate F. Recall that

$$\text{BMSE}_{par}(h) = \frac{1}{B}\sum_{b=1}^{B}\left(\tilde{F}_{nh}^{*b}(t_0) - F_{\hat{\rho}}(t_0)\right)^2.$$

The distribution from which bootstrap samples X^* are drawn is $F_\rho(t) = 1-\exp(-\hat{\rho}_n t)$ and parameter ρ is estimated by $\hat{\rho}_n$ by the procedure described on page 48. The value h_{par} is defined by the bandwidth h that minimises BMSE_{par}. This bandwidth is used to find a bandwidth parameter h_{smo} by minimising BMSE_{smo}, which is defined by

$$\text{BMSE}_{smo}(h) = \frac{1}{B}\sum_{b=1}^{B}\left(\tilde{F}_{nh}^{*b}(t_0) - \tilde{F}_{nh_{par}}(t_0)\right)^2,$$

[1] A distribution function F does not always have an inverse. If F is constant on some interval, one uses the generalised inverse, $F^{-1}(t) = \inf\{x \in \mathbb{R} : F(x) > t\}$.

n	$\hat{\rho}_n$	h_{par}	$\text{BMSE}_{par}(\tilde{F}_{nh_{par}}(1))$	h_{smo}	$\text{BMSE}_{smo}(\tilde{F}_{nh_{smo}}(1))$
500	1.052471	0.75	0.004244447	0.70	0.004462109
1000	0.9758408	0.60	0.00246531	0.6	0.002593156
2000	0.9983666	0.45	0.001338151	0.4	0.001475884

Table 7.1: This table shows results of a bootstrap simulation to find the optimal smoothing parameter.

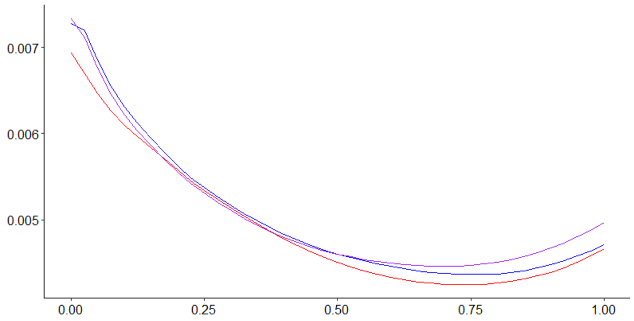

Figure 7.2: From the true distribution, a sample of size $n = 500$ is drawn. An exponential model is fitted to the censored data and $B = 1000$ bootstrap samples are computed. The graph shows the BMSE$_{par}$ (red), the value as a function of the smoothing parameter h_{par} on the x-axis. BMSE$_{smo}$ (purple) is calculated using draws from $\tilde{F}_{nh_{par}}$. A Monte Carlo simulation is done to estimate the true MSE (blue). 1000 datasets of $n = 500$ samples are drawn calculating the NPMLE. Sampled estimators are smoothed using values h on the x-axis.

where $F_{nh_{par}}(t) = \sum_{i=1}^{m} \omega_i \mathbb{K}\left(\frac{t-v_i}{h_{par}}\right)$ and $\mathbb{K}$ is the integrated Epanechnikov kernel. The kernel itself is defined by

$$K(t) = \frac{3}{4}(1-t^2)1_{t\in[-1,1]}(t).$$

If the approximation of the MSE by BMSE$_{par}$ is done using the correct parametric family, it is expected that $\min_h \text{BMSE}_{par}(h) < \min_h \text{BMSE}_{smo}(h)$. In such case, h_{par} already yields good results. However, in practice such parametric assumptions are not always correct, making sampling from the SMLE with $h = h_{smo}$ a better choice than sampling from a parametric approximation. Results of the MSE estimation in which the correct parametric assumption is made is found in Table 7.1.

The results in the table support the claim that the MSE is better approximated by BMSE$_{par}$ than by BMSE$_{smo}$. However, one must note that these quantities are generated using a single sample for each n of the true distribution. This means that repeating the experiment many times may lead to different results, this is however too time consuming. The MSE appears to decrease as the sample size increase. This could imply that estimation becomes more accurate with more data. The quantities found in Table 7.1 are not expected to be much different with new datasets in the same sampling setup. More computational power is required to verify this. It is however possible to suggest the correctness of the output of the bootstrap with a Monte Carlo simulation using the sampling sceme (1)-(4). Table 7.2 shows that the bandwidth differ but the BMSE appears to approximate the MSE well. The true MSE is approximated as well as BMSE$_{par}$ and BMSE$_{smo}$. Figures 7.2, 7.3 and 7.4 show the BMSE as a function of h for single samples of sizes $n = 500$, $n = 1000$ and $n = 2000$ respectively. The figures show that BMSE$_{smo}$ approximate the MSE well.

A simulation is done using a parametric family which does not contain Exp(1). This is to

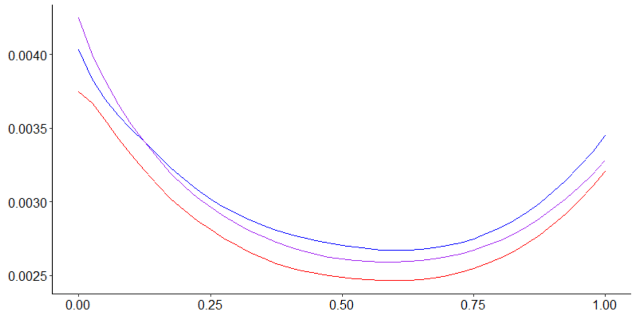

Figure 7.3: From the true distribution, a sample of size $n = 1000$ is drawn. An exponential model is fitted to the censored data and $B = 1000$ bootstrap samples are computed. The graph shows the BMSE$_{par}$ (red), the value as a function of the smoothing parameter h_{par} on the x-axis. BMSE$_{smo}$ (purple) is calculated using draws from $\tilde{F}_{nh_{par}}$. A Monte Carlo simulation is done to estimate the true MSE (blue). 1000 datasets of $n = 1000$ samples are drawn calculating the NPMLE. Sampled estimators are smoothed using values h on the x-axis.

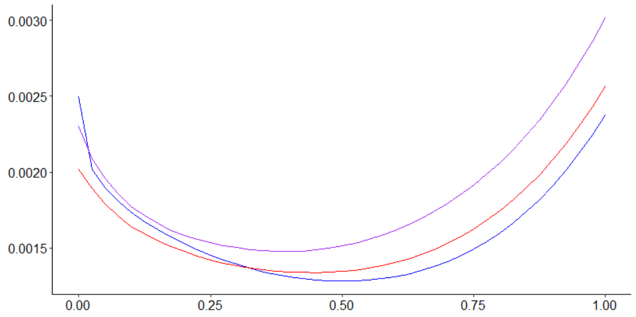

Figure 7.4: From the true distribution, a sample of size $n = 2000$ is drawn. An exponential model is fitted to the censored data and $B = 1000$ bootstrap samples are computed. The graph shows the BMSE$_{par}$ (red), the value as a function of the smoothing parameter h_{par} on the x-axis. BMSE$_{smo}$ (purple) is calculated using draws from $\tilde{F}_{nh_{par}}$. A Monte Carlo simulation is done to estimate the true MSE (blue). 1000 datasets of $n = 2000$ samples are drawn calculating the NPMLE. Sampled estimators are smoothed using values h on the x-axis.

n	h_{opt}	MSE($\tilde{F}_{nh_{opt}}(1)$)
500	0.75	0.004364859
1000	0.6	0.002671733
2000	0.5	0.001281583

Table 7.2: The true MSE in (6.9) is approximated using 1000 samples of sizes $n = 500$, $n = 1000$ and $n = 2000$. The bandwidth that resulted in the smallest MSE is chosen as h_{opt}.

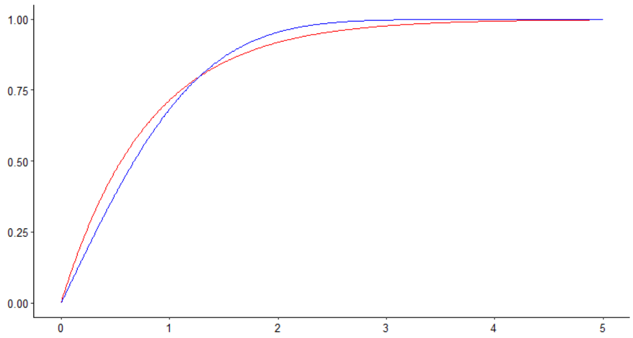

Figure 7.5: A plot showing the distribution function of $\text{Exp}\big(\sqrt{2/\pi}\big)$ (red) and the standard half normal distribution (blue).

n	1000
$\hat{\rho}_n$	1.434622
h_{par}	0.225
$\text{BMSE}_{par}(\tilde{F}_{nh_{par}}(1))$	0.005280483
h_{smo}	0.45
$\text{BMSE}_{smo}(\tilde{F}_{nh_{smo}}(1))$	0.003555854
h_{opt}	0.575
$\text{MSE}_F(\tilde{F}_n(1))$	0.002972139

Table 7.3: Results under misspecification of the model are shown where data generated from a half-normal distribution $X \sim |\mathcal{N}(0,1)|$ was assumed to be exponentially distributed.

show the performance when the chosen parametric model is incorrect. Because the true distribution is not an element of the assumed parametric family, it is expected that $\min_h \text{BMSE}_{par}(h) > \min_h \text{BMSE}_{smo}(h)$. The event time distribution used in this simulation is the half-normal distribution, i.e., $X \sim |\mathcal{N}(0,1)|$. Given censored data, one could think the data are drawn from an exponential distribution. Figure 7.5 shows the cumulative distribution of both the standard half normal distribution and the exponential distribution with parameter $\sqrt{2/\pi}$. It appears that, when the wrong family is chosen, the parametric approach of optimising the bandwidth does on itself not perform well. However, using h_{par} to compute $\text{BMSE}_{smo}(h)$ yields more promising results. It is seen that with the data, a proper bandwidth is obtained as the BMSE approach resulted in $h_{smo} = 0.45$ whereas the Monte Carlo approach resulted in $h_{opt} = 0.575$. The BMSE as a function of h is found in Figure 7.6. Table 7.3 shows the output. It appears that the MSE of $\tilde{F}_{nh_{opt}}(1)$ obtained with an incorrect underlying distribution is not much bigger than when the correct family is chosen. The final SMLE uses the bandwidth that optimised $\text{BMSE}_{smo}(\tilde{F}_{nh}(1))$. Figure 7.7 shows the true distribution, the NPMLE and the SMLE for different sample sizes and misspecification. It can be seen that BMSE_{par} does not match the MSE. It does however result in a reasonable bandwidth h_{par} because as BMSE_{smo} does match the MSE well.

7.3 Simulations on Large Sample Behaviour of the SMLE in the Cox Model

This paragraph contains a simulation study on the large sample performance of the SMLE in the Cox model. Recall the definition of consistency. Let $\hat{\theta}_n$ be an estimator for θ. The estimator $\hat{\theta}_n$ is said to be (weakly) consistent if $\hat{\theta}_n$ converges to θ in probability, that is, for any $\varepsilon > 0$

$$\mathbb{P}(||\hat{\theta}_n - \theta||_N > \varepsilon) \to 0 \text{ as } n \to \infty, \tag{7.4}$$

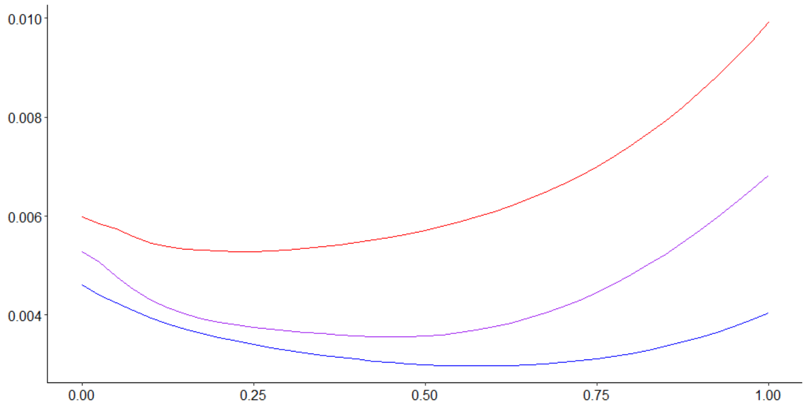

Figure 7.6: From the standard half-normal distribution, a sample of size $n = 1000$ is drawn. An exponential model is fitted to the censored data and $B = 1000$ bootstrap samples are computed. The graph shows the BMSE$_{par}$ (red), the value as a function of the smoothing parameter h_{par} on the x-axis. BMSE$_{smo}$ (purple) is calculated using draws from $\tilde{F}_{nh_{par}}$. A Monte Carlo simulation is done to estimate the true MSE (blue). 1000 datasets of $n = 1000$ samples are drawn calculating the NPMLE. Sampled estimators are smoothed using values h on the x-axis.

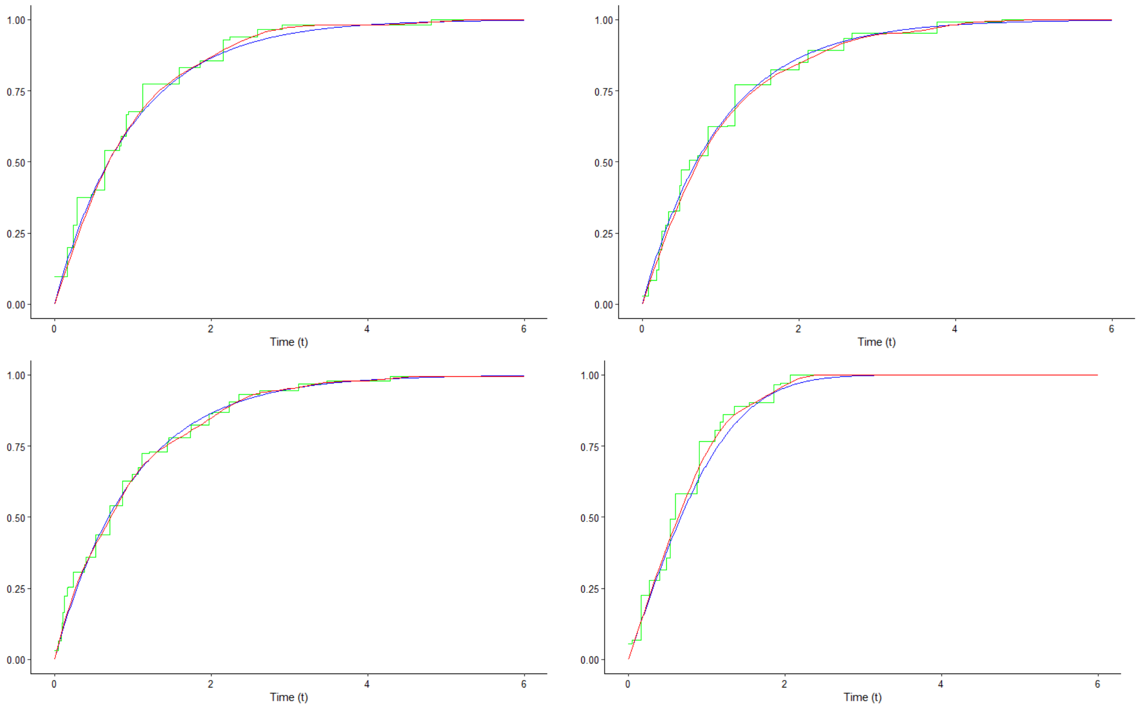

Figure 7.7: In all plots, blue represents the true distribution, the NPMLE is visualised in green and the SMLE is displayed in red. Top left is the estimate for $n = 500$, top right shows the estimate for $n = 1000$, bottom left shows the estimate for $n = 2000$ and bottom right shows the estimate in which the wrong parametric model was assumed when estimating the MSE. The sample of $n = 1000$ is drawn.

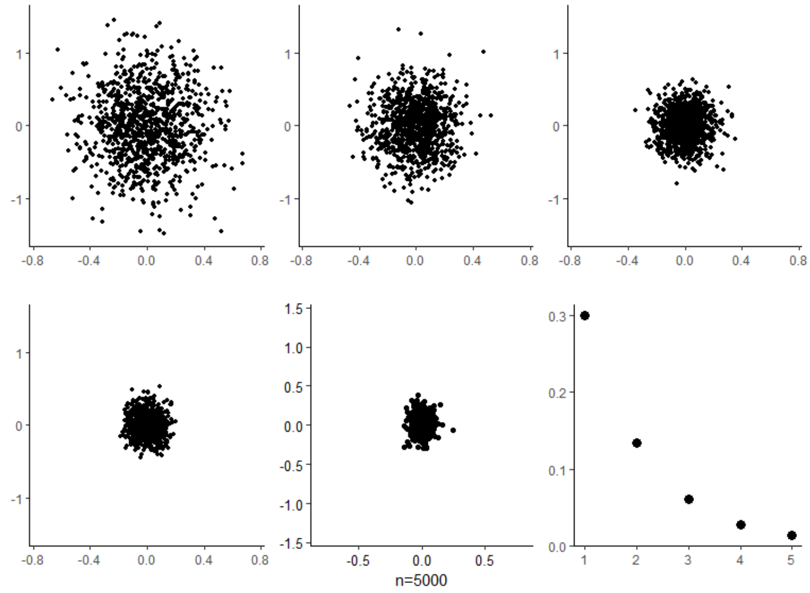

Figure 7.8: For different sample sizes, 1000 Monte Carlo samples of $\hat{\beta}_n$ are generated. From top left to middle bottom, the sample size n are 250, 500, 1000, 2000, 4000 respectively. Scatter plots are plotted and variance of the x-axis. The bottom right image shows the mean squared distances of the estimates of β to the origin for all sample sizes in the same order.

n	250	500	1000	2000	4000
$\xi_{\hat{\beta}_n}(0.2)$	0.82	0.682	0.481	0.213	0.0813

Table 7.4: Approximated probabilities of (7.5) are shown.

where N is some appropriate norm. Because β and $F(t_0)$ belong to a Euclidean space, the Euclidean norm is chosen. In this study, point-wise function estimates of $F(t_0)$ are of interest. In words (7.4) means that if it can be shown that the probability that the error distance surpasses any ε decreases to zero when the sample size increases, the estimator is said to be consistent. Figure 7.8 shows a scatter plot 1000 Monte Carlo estimates of β from data sets consisting of varying sample sizes. It is suggested that the points cluster around the true β. The probability given in (7.4) can be approximated by the relative frequency. To indicate that the estimator $\hat{\beta}_n$ may be consistent, define

$$\xi_{\hat{\beta}_n}(\varepsilon) = \frac{1}{C}\sum_{c=1}^{C} 1_{||\hat{\beta}_n^c-\beta||_{L^2}>\varepsilon} \approx \mathbb{P}(||\hat{\beta}_n-\beta||_{L^2}>\varepsilon). \tag{7.5}$$

For the simulations, $C = 1000$ samples of sizes n are drawn from the model described in paragraph 7.1. For each sample c, the estimates of β and $F(1)$ are calculated and denoted by $\hat{\beta}_n^c$ and $\tilde{F}^c_{nh_{smo}}(1)$. Figure 7.9 plots $\xi_{\hat{\beta}_n}(\varepsilon)$ for ε ranging from 0 to 1. It is shown that for larger n, $\xi_{\hat{\beta}_n}$ decreases from 1 towards zero faster, suggesting that for a fixed ε, the probability does decrease when n increases. Take $\varepsilon = 0.2$, for this sample the numerical values of the probabilities are found in Table 7.4. This behaviour is precisely what is expected when the estimator $\hat{\beta}_n$ is consistent. This gives an indication that the estimator $\hat{\beta}_n$ could be consistent. However, to make this rigorous, more research is needed.

With each $\hat{\beta}_n^c$ comes a sample of $\tilde{F}^c_{nh_{smo}}(1)$. The smoothing parameter h_{smo} found in paragraph 7.2 is used depending on the size of the sample. Possible consistency of $\tilde{F}_{nh_{smo}}(1)$ is now indicated

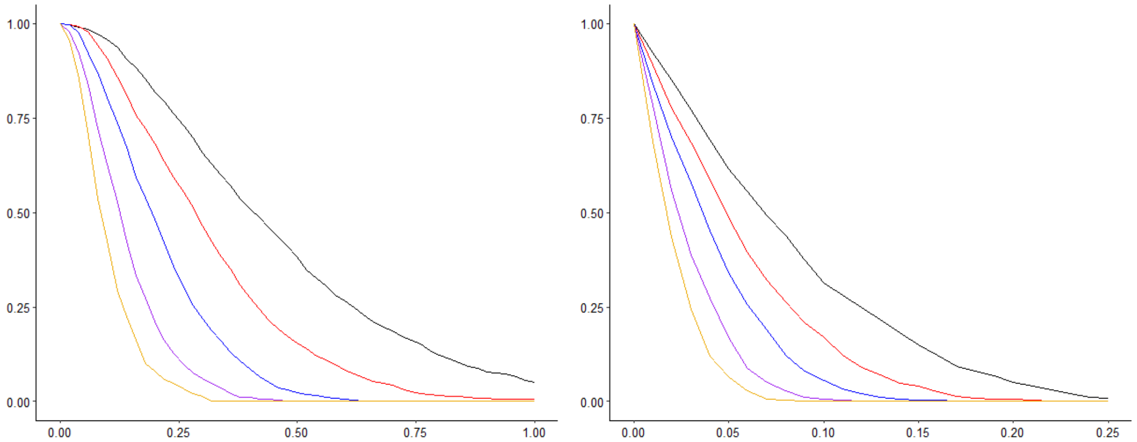

Figure 7.9: For sample sizes $n = 250$ (black), $n = 500$ (red), $n = 1000$ (blue), $n = 2000$ (purple) and $n = 4000$ (yellow), the functions $\xi_{\hat{\beta}_n}(\varepsilon)$ in (7.5) (left) and $\xi_{\tilde{F}_{nh_{smo}}}(\varepsilon)$ in (7.6) (right) are displayed. It can be seen that as the sample size gets bigger, both functions starts to concentrate around zero.

again by approximating (7.4). Define a new function

$$\xi_{\tilde{F}_{nh_{smo}}}(\varepsilon) = \frac{1}{C}\sum_{i=1}^{C} 1_{|\tilde{F}^c_{nh_{smo}}(1)-F(1)|>\varepsilon} \approx \mathbb{P}(|\tilde{F}_{nh_{smo}}(1) - F(1)| > \varepsilon). \tag{7.6}$$

Figure 7.9 shows plots of this function for multiple values of n. The same conclusions on consistency are made as done for $\hat{\beta}_n$.

If $\alpha_n(\tilde{F}_{nh_{smo}}(1) - F(1))$ converges to some distribution, then one speaks of bootstrap consistency if $\alpha_n(\tilde{F}^*_{nh_{smo}}(1) - F_{nh_{smo}}(1))$ converges to the same distribution. Figure 7.10 shows approximated distributions for both $\tilde{F}^*_{nh_{smo}})(1)$, sampled from the initial sampling setup and the distribution of $\tilde{F}^*_{nh_{smo}}(1)$ when resampling from $\tilde{F}_{nh_{smo}}$. For each sample size, the histograms are very close to each other. This suggests bootstrap consistency.

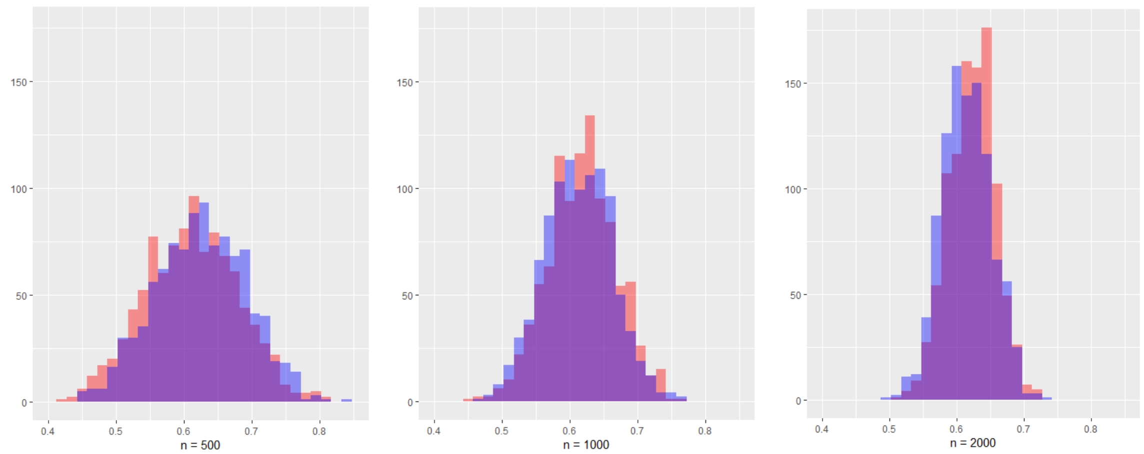

Figure 7.10: For sample sizes $n = 500$, $n = 1000$ and $n = 2000$. In all three figures, the red histogram represents the distribution of $\tilde{F}_{nh_{smo}}(1)$ from the true sampling setup. The blue histogram represents the distribution of $\tilde{F}^*_{nh_{smo}}(1)$ where event times were resampled from $F_{nh_{smo}}$. The overlap is coloured in purple. The histograms are very close to eachother, indicating consistency of the bootstrap procedure.

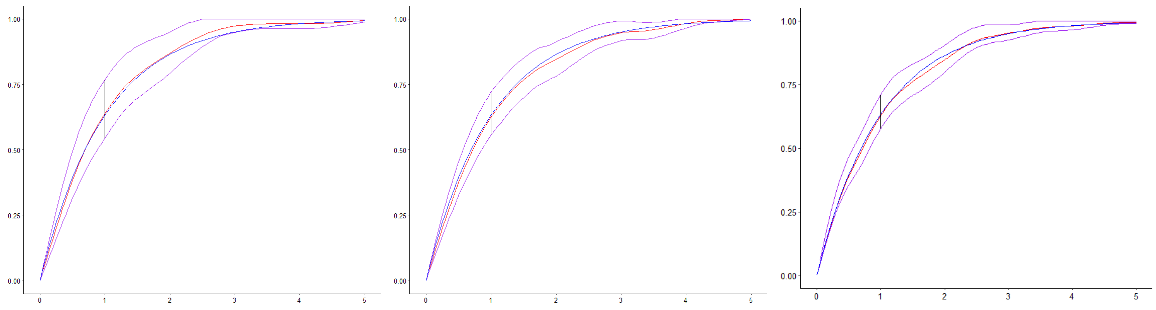

Figure 7.11: Confidence bands for the event-time distribution F made using samples of sizes (from left to right) $n = 500$, $n = 1000$ and $n = 2000$. The blue lines are the true function F, red is the SMLE and purple the confidence intervals with the bandwidth h being optimised for estimation of $F(1)$. The black line helps visualising the length of the confidence interval at $F(1)$

Finally, confidence intervals are constructed for $F(1)$ using different sample sizes. Note that an α confidence interval is a random interval which asymptotically has the true parameter included with probability α. With the setup used for the simulations, generating a single confidence interval can take more than a day. Lengths of intervals are found in Table 7.5. The confidence bands of F are shown in Figure 7.11, using the same h_smo for the whole function. With more computer power it would be possible to generate more confidence intervals so that coverage of the confidence intervals can be tested. If the confidence intervals are consistent, then the coverage should become closer to α as more data becomes available.

n	500	1000	2000
length CI	0.2677453	0.1977452	0.1391965

Table 7.5: The lengths of the 95% single confidence intervals are given.

7.4 Application to Real Data

In the period from 1996 to 2001, data has been collected from 4386 randomly sampled children in Flanders and is studied in (Vanobbergen, Martens, Lesaffre, & Declerck, 2000). The outcome of interest was the time (in age) of the right central incisor to come through (tooth 24). The data is given in tenths of years. The inspection times in the data range from 1.1 to 7.4 years old. Together with the time interval in which the emergence of the tooth occurred, two extra variables are given. The extra variables are the gender of the child, male is indicated by 0 and female by a 1. The second variable is "dmf", which is one if the tooth is delayed, missing or filled and zero otherwise. The result of the NPMLE is found in Figure 7.12 and the Cox vector is given by $\beta = (0.3239359, 0.3366725)$. The output β_1 implies that girls have a higher chance of having their tooth 24 to emerge prior to boys.

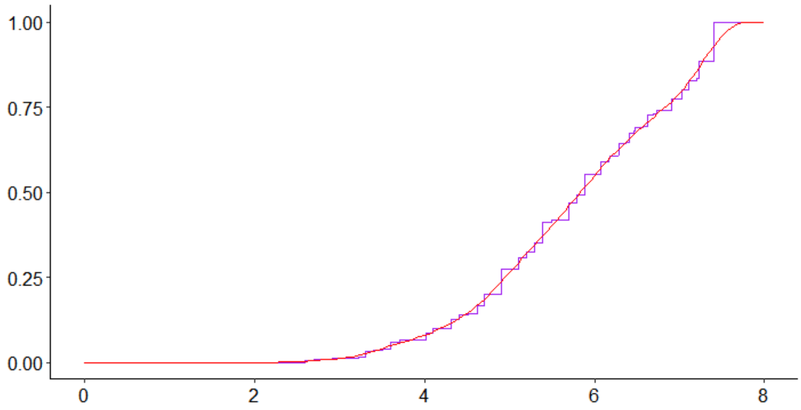

Figure 7.12: Baseline distribution function based on the tooth data from dynsurv. The blue line is the NPMLE and the red line is the SMLE with $h = 0.4$.

Bootstrapping according to the method of Sen and Xu is not possible with this dataset as the data was incomplete. Only the intervals in which the emergence of the tooth occurred are provided. For a right censored event i, only u_i and ∞ were in the dataset and t_i was missing. For this reason, no new indicator variables can be resampled according to (6.8). The data would however be suitable for the method where pairs of observations are resampled. Because resampling was in this case not possible, $h = 0.4$ is used, the bandwidth chosen based in the results found in Table 7.2.

The positive Cox coefficient, weighting the gender implies that girls generally have their central incisor emerged earlier than boys. This result is supported in (Diamanti & Townsend, 2003) who conclude the same using data on Australian children. The variable "dmf" stands for *decay, missing or filled*. This means that the primary tooth decays from caries, fell out or had to be filled. It appears from this analysis that if this condition is linked the permanent tooth to emerge faster. However, no supporting articles have been found.

Chapter 8

Appendix

8.0.1 Gateaux Derivatives

Let F and G be two functions in $\mathcal{K}$. The Gateaux derivative of $\mathcal{L}$ at F in the direction G is defined as

$$[D_G(\mathcal{L})](F) = \lim_{\varepsilon \downarrow 0} \frac{1}{\varepsilon}\big(\mathcal{L}(F + \varepsilon G) - \mathcal{L}(F)\big). \tag{8.1}$$

Instead of F and G being variables as in standard calculus, Gateaux derivatives have a slightly different interpretation. As in the normal case, the derivative indicates how fast the function output changes with the variables. Now it is of interest how functions themselves change when shifted on a line segment between two functions. Computation of the derivative is split into three parts. L'Hôpital's theorem is key for calculating the derivative. A short notation is used to improve readability. In the following the notation $F_x = F(x)$, $G_x = G(x)$ and $c = \exp(\beta^\top s)$ is used. The Gateaux derivative of $\mathcal{L}$ given in (4.6) is computed in parts,

$$\lim_{\varepsilon \downarrow 0} \frac{1}{\varepsilon}\Big(\log\big(1 - \exp(-c(F + \varepsilon G)\big) - \log\big(1 - \exp(-cF)\big)\Big) = \lim_{\varepsilon \downarrow 0} \frac{1}{\varepsilon} \log\left(\frac{1 - \exp(-c(F + \varepsilon G))}{1 - \exp(-cF)}\right).$$

Writing the right hand side as a fraction of the logarithm divided by ε yields an expression to which L'hôpital's theorem can be applied. Note that the inside of the logarithm converges to one as $\varepsilon \to 0$ and $\log(1) = 0$. Therefore the right hand side becomes

$$\lim_{\varepsilon \downarrow 0} \frac{cG_t \exp(-c(F_t + \varepsilon G_t))}{1 - \exp(-c(F_t + \varepsilon G_t))} = \frac{cG_t \exp(-cF_t)}{1 - \exp(-cF_t)} = \frac{cG_t}{\exp(cF_t) - 1} = \frac{\exp(\beta^\top s) G(t)}{\exp(e^{\beta^\top s} F(t)) - 1}. \tag{8.2}$$

Derivatives of the second and third factor of the likelihood are computed in a similar fashion.

$$\begin{aligned}
&\lim_{\varepsilon \downarrow 0} \frac{1}{\varepsilon} \log\left(\frac{\exp(-c(F_t + \varepsilon G_t)) - \exp(-c(F_u + \varepsilon G_u))}{\exp(-cF_t) - \exp(-cF_u)}\right) \\
=& \lim_{\varepsilon \downarrow 0} \frac{-cG_t \exp(-c(F_t + \varepsilon G_t)) + cG_u \exp(-c(F_u + \varepsilon G_u))}{\exp(-c(F_t + \varepsilon G_t)) - \exp(-c(F_u + \varepsilon G_u))} \\
=& \frac{-cG_t \exp(-cF_t) + cG_u \exp(-cF_u)}{\exp(-cF_t) - \exp(-cF_u)} \\
=& \frac{-e^{\beta^\top s} G(t) \exp(-e^{\beta^\top s} F(t)) + e^{\beta^\top s} G(u) \exp(-e^{\beta^\top s} F(u))}{\exp(-e^{\beta^\top s} F(t)) - \exp(-e^{\beta^\top s} F(u))}
\end{aligned} \tag{8.3}$$

And the third derivative becomes

$$\lim_{\varepsilon \downarrow 0} \frac{1}{\varepsilon}\Big(- c(F_u + \varepsilon G_u) + cF_u\Big) = -cG_u = -e^{\beta^\top s} G(u). \tag{8.4}$$

Substituting the true quantities: $c = e^{\beta^\top s_i}$, $F_t = \Lambda_0(u_i), F_u = \Lambda_0(u_i)$. Here, G is the function to which the derivative is taken. final result:

$$[D_G(\mathcal{L})](F) = \sum_{i=1}^{n} \delta_i \frac{cG_t}{\exp(cF_t) - 1} + \gamma_i \frac{-cG_t \exp(-cF_t) + cG_u \exp(-cF_u)}{\exp(-cF_t) - \exp(-cF_u)} - \mu_i cG_u. \tag{8.5}$$

The Gateaux derivative can be used to test optimality. If for an estimate $\hat{\Lambda}_{0n}$ it holds that $[D_{\Lambda_0}(\mathcal{L})](\hat{\Lambda}_{0n}) \geq 0$ for all $\Lambda_0 \in \mathcal{H}$ then $\hat{\Lambda}_{0n}$ is the maximiser.

8.0.2 Hooke-Jeeves Algorithm

An alternative to the Nelder-Mead is the Hooke-Jeeves algorithm. Consider an optimisation problem of the function $\varphi : \mathbb{R}^d \to \mathbb{R}$. Given an initial point $x^{(0)}$, Hooke-Jeeves algoritm evaluates φ at the points around $x^{(0)}$. Define a set of unit vectors, for example the standard unit vectors $e_i = (0, \ldots, 1, \ldots, 0)$ with a 1 at entry i. Then based on the current estimate of the maximiser $x^{(k)}$, the algorithm generated $2d$ new points. Denote these points by $x_j^{(k)}$ where $j = 1, \ldots, 2d$ and $x_j^{(k)} = x^{(k)} + e_j$ if $j = 1, \ldots, d$ and $x_j^{(k)} = x^{(k)} - e_{j-d}$ if $j = d+1, \ldots, 2d$. A direction with the best increase if searched for. $\varphi(x^{(k)}) > \varphi(x_i^{(k)})$ for all i, the algorithm decreases the searching area by replacing $x_j^{(k)} = x^{(k)} \pm e_j$ by $x_j^{(k)} = x^{(k)} \pm \alpha e_j$ and attempts to find vector with higher value of φ. Pseudocode for Hooke-Jeeves is found in Algorithm 3.

```
Data: Initial estimate $x^{(0)} \in \mathbb{R}^d$
Set $k = 0$;
Set $d = 1$;
Set $\gamma > 0$;
Set $a > 0$ Set $\varphi(x^{(-1)}) = \infty$;
while $\varphi(x^{(k)}) - \varphi(x^{(k-1)}) > \gamma$ do
    Generate $2d$ distinct points around $x^k$ by $x_i^{(k)} = x^{(k)} \pm \epsilon e_i$, $i = 1 \ldots d+1$;
    Calculate $\ell_i = \varphi(x_i^{(k)})$ for all $i = 1 \ldots 2d$;
    Find $i$ correspronding to $\ell_{(2d)}$;
    while $\varphi(x^{(k)}) > \ell_{(2d)}$ do
        Set $d \mapsto \alpha d$;
        Generate $2d$ distinct points around $x^0$: $x_i^{(k)} = x^{(0)} \pm \epsilon e_i$, $i = 1, \ldots, d+1$;
        Calculate $\ell_i = \varphi(x_i^{(k)})$ for all $i = 1 \ldots 2d$;
        Find $i$ correspronding to $\ell_{(2d)}$;
    end
    Set $d \mapsto \xi d$;
    Set $x^{(k+1)} = \text{argmax}_{x^{(k)}} \varphi(x_i^{(k)})$;
    Set $k \mapsto k+1$;
end
```

Algorithm 3: Hooke-Jeeves

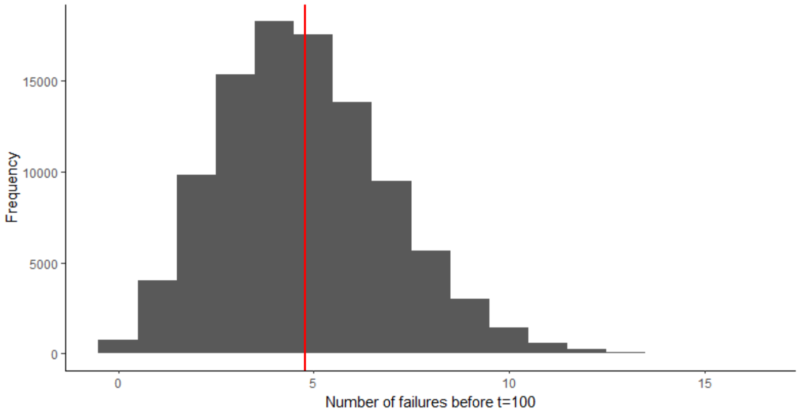

Figure 8.1: N is the number of deaths before time $t = 100$ in years if the mortality of people is modelled by a Weibull($\frac{1}{80}$, 7) distribution.. 100000 samples of N are generated.. The red line is the mean of the samples and is approximately 400. This supports that $\mathbb{E}(N) = \Lambda_{\rho,k}(100) = \left(\frac{100}{80}\right)^7 \approx 4.768$ deaths should be the mean of the experiment. The vertical red line displays the mean, coinciding with the theoretical prediction.

8.0.3 Count-data Interpretation of Cumulative Hazard Funtion

```
#### Simulation number of human deaths ####
Lambda = function(t, scale = 1, shape = 1) (t/scale)^shape       # CHF
Lambda_inv = function(t, scale = 1, shape = 1) scale*t^(1/shape) # Inverse CHF
t_0 = 100              # Age of interest
N_max = 20             # Maximum of events for a single subject
scale = 80             # Scale parameter of the Weibull distribution
shape = 7              # Shape parameter of the Weibull distribution
samples_N = NA         # Vector will contain the number of events before t_0
for ( j in 1:10000){
unif = runif(40)
cu_unif = cumsum(-log(1-unif))
event_times = Lambda_inv(cu_unif, scale, shape)
samples_N[j] = max(which( event_times < t_0 ))
}
isInf = which(is.infinite(samples_N))
samples_N[isInf] = 0   # People who did not die was set to -Inf is set to 0 deaths
```

References

Banerjee, M., & Wellner, A., Jon. (2005). *Confidence intervals for current status data.* doi: 10.1111/j.1467-9469.2005.00454.x

Bradburn, M. J., Clark, T. G., Love, S. B., & Altman, D. G. (2003). Survival analysis part ii: Multivariate data analysis – an introduction to concepts and methods. *British Journal of Cancer*, *89*(3), 431–436. doi: 10.1038/sj.bjc.6601119

Cleves, M., Gould, W., William, & Marchenko Yulia, V. (2002). An introduction to survival analysis using stata. , 428.

Cox, D. R. (1972). Regression models and life-tables. *Springer Series in Statistics Breakthroughs in Statistics*, 527–541. doi: 10.1007/978-1-4612-4380-9_37

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, *39*(1), 1–22. doi: 10.1111/j.2517-6161.1977.tb01600.x

Diamanti, J., & Townsend, G. C. (2003). New standards for permanent tooth emergence in australian children. *Australian dental journal*, *48*(1), 39–42.

E., Robertson, T., Wright, F. T., & Dykstra, R. L. (1990). Order restricted statistical inference. *Biometrics*, *46*(3), 878. doi: 10.2307/2532111

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, *7*(1), 1–26. doi: 10.1214/aos/1176344552

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap.* doi: 10.1007/978-1-4899-4541-9

El-Gohary, A., Alshamrani, A., & Al-Otaibi, A. N. (2013). The generalized gompertz distribution. *Applied Mathematical Modelling*, *37*(1-2), 13–24. doi: 10.1016/j.apm.2011.05.017

Groeneboom, P., & Hendrickx, K. (2017). Confidence intervals for the current status model. *Scandinavian Journal of Statistics*, *45*(1), 135–163. doi: 10.1111/sjos.12294

Groeneboom, P., & Jongbloed, G. (2014). *Nonparametric estimation under shape constraints.* doi: 10.1017/cbo9781139020893

Groeneboom, P., & Jongbloed, G. (2015). Nonparametric confidence intervals for monotone functions. *The Annals of Statistics*, *43*(5), 2019–2054. doi: 10.1214/15-aos1335

Groeneboom, P., Jongbloed, G., & Wellner, J. A. (2008). The support reduction algorithm for computing non-parametric function estimates in mixture models. *Scandinavian Journal of Statistics*, *35*(3), 385–399. doi: 10.1111/j.1467-9469.2007.00588.x

Groeneboom, P., Jongbloed, G., & Witte, B. I. (2010). Maximum smoothed likelihood estimation and smoothed maximum likelihood estimation in the current status model. *The Annals of Statistics*, *38*(1), 352–387. doi: 10.1214/09-aos721

Groeneboom, P., & Wellner, J. A. (1992). *Information bounds and nonparametric maximum likelihood estimation.* doi: 10.1007/978-3-0348-8621-5

Hooke, R., & Jeeves, T. A. (1961). Direct search solution of numerical and statistical problems. *Journal of the ACM*, *8*(2), 212–229. doi: 10.1145/321062.321069

Ibrahim, J. G., Chen, M.-H., & Sinha, D. (2014). Bayesian survival analysis. *Wiley StatsRef: Statistics Reference Online*. doi: 10.1002/9781118445112.stat06003

Jongbloed, G. (1998). The iterative convex minorant algorithm for nonparametric estimation. *Journal of Computational and Graphical Statistics*, *7*(3), 310–321. doi: 10.1080/10618600.1998.10474778

Klaus, B., & Strimmer, K. (2012). Signal identification for rare and weak features: higher criticism or false discovery rates? *Biostatistics*, *14*(1), 129–143. doi: 10.1093/biostatistics/kxs030

Ma, S., & Kosorok, M. R. (2005). Robust semiparametric m-estimation and the weighted bootstrap. *Journal of Multivariate Analysis*, *96*(1), 190–217. doi: 10.1016/j.jmva.2004.09.008

Minami, J., Kawano, Y., Nonogi, H., Ishimitsu, T., Matsuoka, H., & Takishita, S. (1998). Blood pressure and other risk factors before the onset of myocardial infarction in hypertensive patients. *Journal of Human Hypertension*, *12*(10), 713–718. doi: 10.1038/sj.jhh.1000697

Nelder, J. A., & Mead, R. (1965). A simplex method for function minimization. *The Computer Journal*, *7*(4), 308–313. doi: 10.1093/comjnl/7.4.308

Pan, W. (1999). Extending the iterative convex minorant algorithm to the cox model for interval-censored data. *Journal of Computational and Graphical Statistics*, *8*(1), 109. doi: 10.2307/1390923

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

Sen, B., & Xu, G. (2015). Model based bootstrap methods for interval censored data. *Computational Statistics and Data Analysis*, *81*, 121–129. doi: 10.1016/j.csda.2014.07.007

Silverman, B. W. (1986). Density estimation for statistics and data analysis. doi: https://doi.org/10.1201/9781315140919

Song, S. (2004). Estimation with univariate mixed case interval censored data. *Statistical Sinica*, *14*, 269-282.

Tibshirani, R., & Wasserman, L. (2015). *Convexity and optimization.*

Van der Vaart, A. W., & Wellner, J. A. (1996). *Weak convergence and empirical processes.* doi: 10.1007/978-1-4757-2545-2_3

Van der Vaart, A. W., & Wellner, J. A. (2000). Preservation theorems for glivenko-cantelli and uniform glivenko-cantelli classes. *High Dimensional Probability II*, 115–133. doi: 10.1007/978-1-4612-1358-1_9

Vanobbergen, J., Martens, L., Lesaffre, E., & Declerck, D. (2000). The signal-tandmobiel project a longitudinal intervention health promotion study in flanders (belgium): baseline and first year results. *European Journal of Paediatric Dentistry*, *2*, 87–96.

Wang, Y., & Fani, S. (2017). Nonparametric maximum likelihood computation of a u-shaped hazard function. *Statistics and Computing*, *28*(1), 187–200. doi: 10.1007/s11222-017-9724-z

Wasserman, L. (2006). *All of nonparametric statistics.* doi: 10.1007/0-387-30623-4

Nomenclature

F	Event-time distribution function
S	Survival function
Λ	Cumulative hazard function
Λ_0	Baseline cumulative hazard function
$\check{F}_n$	Empirical distribution function based on n samples
$\hat{F}_n$	Nonparametric maximum likelihood estimator based on n samples
$\hat{F}_n^{(k)}$	The k'th iteration of an algorithm approximating F
$\tilde{F}_{nh}$	Smoothed MLE based on n samples, with bandwidth h
f	Event-time density function
λ	Hazard function
λ_0	Baseline hazard function
H	Joint distribution of T and U
T	First inspection time random variable
U	Second inspection time random variable
X	Event-time random variable
X^*	Bootstrap resample of random variable X
Δ	Random variable indicating the event happened before first inspection
Γ	Random variable indicating the event happened between two inspections
M	Random variable indicating the event happened after the last inspection
Z	Random variable of the observations (T, U, Δ, Γ)
V	The set containing the times in which $\hat{F}_n$ can contain jumps
$\mathbb{R}^+$	The set of positive real numbers
$\mathcal{F}$	The set of distribution functions of interest
$\mathcal{H}$	The set of cumulative hazard functions of interest
$\mathcal{I}_1$	The set of t_i's of left censored events
$\mathcal{I}_{2a}$	The set of t_i's of middle censored events

$\mathcal{I}_{2b}$	The set of u_i's of middle censored events
$\mathcal{I}_3$	The set of u_i's of right censored events
β	Cox coefficients
h	Bandwidth
s	Matrix containing vectors of covariates of all subjects
K	Kernel function
L	Likelihood function
$\mathbb{K}$	Integrated kernel function
$\mathcal{L}$	Log likelihood function
$p\mathcal{L}_\beta$	Profile likelihood function with fixed β
BMSE	Bootstrapped mean squared error
MSE	Mean squared error
∇	Gradient operator
Hess	Hessian operator
$\mathbb{E}$	Expected value operator
$\mathbb{V}$	Variance operator
$\mathbb{P}$	Measure of random variable X
η_v	Dirac measure with mass at v
$\mathcal{N}$	Normal distribution
P_j	Index j of the cumulative sum diagram P
β	Vector containing Cox coefficients
ℓ	Link function connecting two inspection times of the same subject
$\operatorname{argmax}_{x\in D} f(x)$	The element in $x \in D$ such that x maximises f
$\operatorname{argmin}_{x\in D} f(x)$	The element in $x \in D$ such that x minimises f