Cox-Type Regression and Transformation Models with Change-Points Based on Covariate Thresholds

Dissertation zur Erlangung des Doktorgrades der Naturwissenschaften (Dr. rer. nat.)

Fakultät Naturwissenschaften
Universität Hohenheim

Institut für Angewandte Mathematik und Statistik

vorgelegt von
Constanze Lütkebohmert-Marhenke

aus Stadthagen

September 2007

Dekan:	Prof. Dr. H. Breer
1. berichtende Person:	Prof. Dr. U. Jensen
2. berichtende Person:	Prof. Dr. U. Stadtmüller
Eingereicht am:	27. September 2007
Mündliche Prüfung am:	20. Dezember 2007

Die vorliegende Arbeit wurde am 10. Dezember 2007 von der Fakultät Naturwissenschaften der Universität Hohenheim als "Dissertation zur Erlangung des Doktorgrades der Naturwissenschaften" angenommen.

Contents

Chapter 1

Introduction

The studying of Survival Analysis has a long tradition. The most popular model is the proportional hazards model introduced by Cox (1972), but many other regression models have been proposed since then. The Cox model is a semiparametric regression model which is useful to calculate a survival function of survival time data depending on covariates which influence this function. It suggests that the underlying regression function is linear in the covariates. But of course, the question arises whether this assumption is always true. It may occur that a covariate has another functional form like a logarithmic or a quadratic one. On the other hand the influence of a covariate can change at a certain threshold of the covariate. But how to investigate the correct functional form? Plots of residuals can be used to obtain an educated guess. But is there also an analytical way? The aim of this thesis is to provide a more flexible Cox model with bent-line change-points according to thresholds of the covariates, i.e. the underlying regression function is continuous but not differentiable in the change-points. Thresholds in time are also interesting, but will not be discussed in this thesis. The Cox model with change-points and certain goodness-of-fit tests enable us to rebuild the functional form of a covariate as piecewise linear. A further goal is to introduce a more complex transformation model with a bent-line change-point.

Survival analysis has its origin in biostatistics. Usually, it is concerned with the analysis of individuals experiencing events over time. The aim of regression models is to relate the events to certain covariates. A classical application is the study of patients that undergo some type of surgery. One is interested in which way covariates like age or a special treatment influence the length of survival. However, in most cases the patients are still alive when the study ends and the statistical analysis of the gathered data is made. For those patients it is only known that they survive up to a certain time. This phenomenon is called censoring and has to be taken into account in survival analysis studies. But of course such data can not only be found in biometrics. Other fields of application

are for instance system or software reliability and actuarial mathematics. In reliability theory one individual can be a machine or a motor and the interest lies in predicting the survival until a failure occurs. Considering a piece of software the events are incoming bug reports. In actuarial mathematics one can think of different applications. An individual can be represented by an insurance contract and the events are claims made by the insurance holder. On the other hand it is conceivable to investigate the cancellation of contracts by means of survival analysis.
In classical survival analysis only one event per individual occurs. The modern interpretation of the models also allow more than one event. Different examples where this is the case are mentioned above. Theoretically, for the ith individual a stochastic process $N_i(t)$ is given which counts the number of events for the individual up to time t. Regression models are designed to connect certain covariates with the rate of occurrence of events. Usually, such models are described in form of a so-called intensity $\lambda_i(t)$, which can be defined in the following way. Under certain regularity conditions in martingale theory there exists a predictable increasing process $\Lambda_i(t)$ such that $N_i(t) - \Lambda_i(t)$ is a local martingale. If the paths of the compensator $\Lambda_i(t)$ are absolutely continuous with respect to Lebesgue measure, then a predictable process $\lambda_i(t)$ such that $\Lambda_i(t) = \int_0^t \lambda_i(s)\,\mathrm{d}s$ holds is called an intensity.

1.1 Cox Models

In the Cox model the intensity is assumed to be

$$\lambda_i(t) = \lambda_0(t) R_i(t) \exp\{\boldsymbol{\beta}^\top \boldsymbol{Z}_i(t)\},$$

where the observable covariates $\boldsymbol{Z}_i$ are combined in a vector of predictable processes, λ_0 is a deterministic function, the so-called baseline intensity and the vector of regression parameters is denoted by $\boldsymbol{\beta} \in \mathbb{R}^p$. The observable stochastic process R_i is called the at-risk indicator which indicates whether an individual is at risk or not by taking only values 1 and 0. In the most basic model the covariates are not time-dependent and hence, the covariates are simply given as a vector of random variables. The unknown function $\lambda_0(t)$ and the regression parameter vector $\boldsymbol{\beta}$ have to be estimated. This model is called semiparametric, since it contains an infinite-dimensional parameter λ_0 and a finite-dimensional parameter that consists of the regression parameters.
An extension of this model is the Cox model with one single change-point at an unknown threshold of a covariate, i.e. one covariate, say Z_2, is misspecified in the sense that the ordinary linear unchanging influence of the covariate is not given. The intensity of this

model can be written as

$$\lambda_i(t) = \lambda_0(t) R_i(t) \exp\{\boldsymbol{\beta}^\top \boldsymbol{Z}_{1i}(t) + \beta_2 Z_{2i} + \beta_3 (Z_{2i} - \xi)^+\}, \tag{1.1}$$

where a^+ means the maximum of a and 0, the parameter $\xi \in \mathbb{R}$ represents the change-point and the other parameters are as in the classical Cox model. Thus, the influence of the covariate Z_{2i} changes from β_2 to $\beta_2 + \beta_3$ when $\boldsymbol{Z}_{2i}$ exceeds the change-point. The unknown change-point parameter has to be estimated as well as the regression parameter and the baseline intensity function. In the usual Cox model the regression parameter is estimated by a partial log likelihood

$$\log L(\boldsymbol{\beta}) = \sum_{i=1}^{n} \int_0^{\tau} \boldsymbol{\beta}^\top \boldsymbol{Z}_i(t)\, \mathrm{d}N(t) - \int_0^{\tau} \log \left(\sum_{i=1}^{n} R_i(t) \exp\{\boldsymbol{\beta}^\top \boldsymbol{Z}_i(t)\} \right) \mathrm{d} \left(\sum_{i=1}^{n} N_i(t) \right)$$

and instead of the baseline function $\lambda_0(t)$ the cumulative intensity function $\Lambda_0(t) = \int_0^t \lambda_0(s)\, \mathrm{d}s$ is estimated by the so-called Breslow estimator (see Andersen et al. (1993))

$$\hat{\Lambda}_0(t) = \int_0^t \frac{\mathrm{d}(n^{-1} \sum_{i=1}^{n} N_i(s))}{n^{-1} \sum_{j=1}^{n} R_j(s) \exp\{\hat{\boldsymbol{\beta}}_n^\top \boldsymbol{Z}_j(s)\}}.$$

For the Cox model with a change-point the partial likelihood is determined by the intensity given in (1.1). Hence, the likelihood depends on the regression parameter $\boldsymbol{\beta}$ and on the change-point parameter ξ. We obtain estimates by maximizing the likelihood with respect to the parameters, which is done in a two-phase maximization. These estimates possess certain desirable properties. We show that our combined estimator of $\boldsymbol{\theta} = (\xi, \boldsymbol{\beta}^\top)^\top$ is $\sqrt{n}-$consistent and asymptotically normal. Since our underlying regression function is continuous but not differentiable in the change-point parameter ξ the usual approach of considering a Taylor expansion can not be made. Therefore, we use techniques developed for the theory of empirical processes. Some authors stated that the rate of convergence of the change-point estimate in our model should be n. However, our simulation studies as well as our analytical proofs do not support this claim, see Figure 1.1. The cumulative baseline intensity is estimated by the Breslow estimator depending on the estimates of $\boldsymbol{\beta}$ and ξ. We show that $\sqrt{n}(\hat{\Lambda}_n(t) - \Lambda_0(t))$ converges weakly to a Gaussian process.
The Cox model with one change-point in one covariate can be further extended to a model with a general risk function and with more than one change-point. This model is given by

$$\lambda_i(t, \boldsymbol{\theta}) = \lambda_0(t) R_i(t)\, r \left\{ \boldsymbol{\beta}_1^\top \boldsymbol{Z}_{1i}(t) + \boldsymbol{\beta}_2^\top \boldsymbol{Z}_{2i}(t) + \boldsymbol{\beta}_3^\top (\boldsymbol{Z}_{2i}(t) - \boldsymbol{\xi})^+ \right\},$$

where $r : \mathbb{R} \to [0, \infty)$ is a twice continuously differentiable nonnegative link function. The change-points are indicated by $\boldsymbol{\xi}$, which is a vector of parameters lying in a rectangle

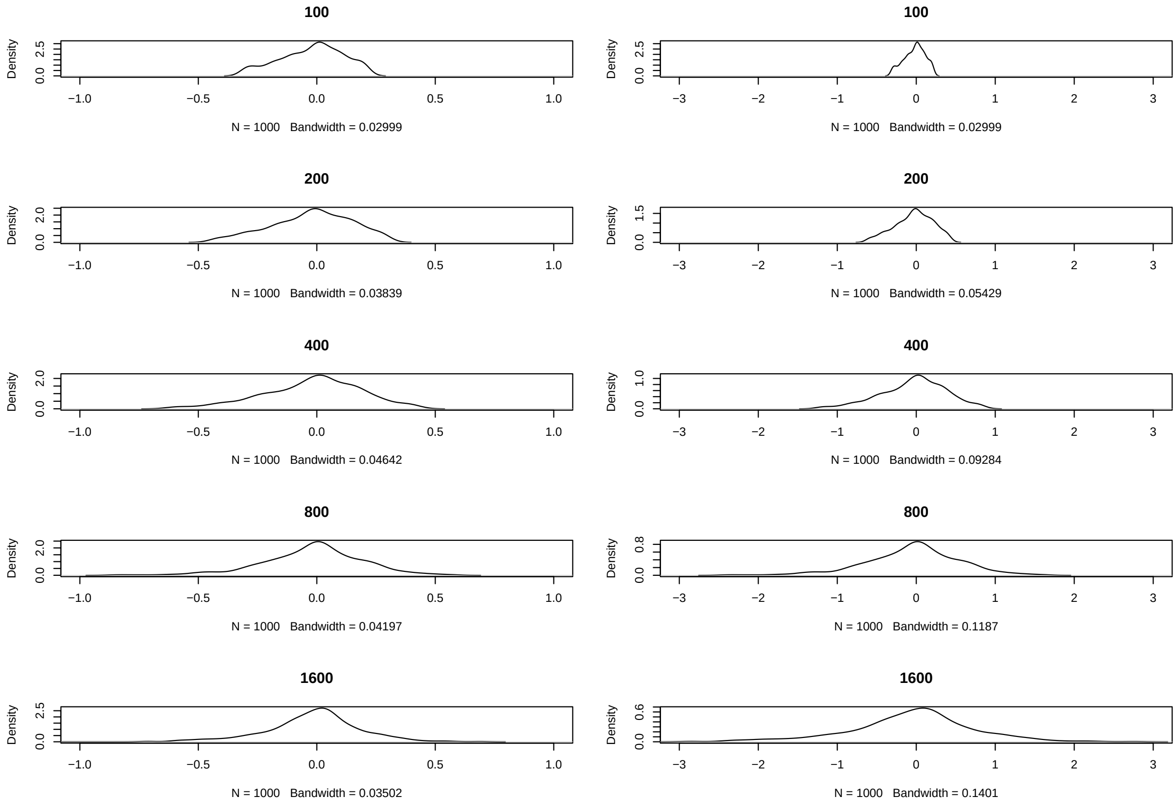

Figure 1.1: Simulation study on the rate of convergence. On the left-hand side the empirical density of $\sqrt{n}(\hat{\xi}_n - \xi_0)$ and on the right-hand side the empirical density of $n(\hat{\xi}_n - \xi_0)$ is shown with $n = 100, 200, 400, 800, 1600$ and with 1000 replicates for each n.

$\Xi = [\xi_{11}, \xi_{21}] \times [\xi_{12}, \xi_{22}] \times \cdots \times [\xi_{1m}, \xi_{2m}]$. The parameters $\xi_{11}, \xi_{21}, \xi_{12}, \xi_{22}, ..., \xi_{1m}, \xi_{2m}$ are assumed to be known. The ideas of the proofs are similar to the ones in the univariate case. But the replacement of the exponential function by a general risk function requires additional conditions for the function r and makes the proofs more complex.

1.2 Transformation Model

In applications, it is possible that the Cox model does not represent the data well enough. It may occur, that groups of related survival times are correlated due to an unobservable risk factor. Groups sharing some risk factor might be a family or electric motors from the same plant. One way to describe such kind of data is in terms of so-called frailty models. In that case an unobservable random variable acts multiplicatively on the intensity. If this intensity is given by a Cox model then the new intensity can be written as

$$\lambda_i(t) = W_i \lambda_0(t) R_i(t) \exp\{\boldsymbol{\beta}^\top \boldsymbol{Z}_i\},$$

where W_i is the unobservable positive random variable. Usually, one assumes that the distribution of W_i belongs to some specific class of distributions. The most frequently used is the Gamma distribution but other choices are possible, too. Among others these frailty models and the Cox model are submodels of the linear transformation model, which can be written in terms of the survival function of a survival time conditional on the covariates $\boldsymbol{Z}$ as

$$S_{\boldsymbol{Z}}(t) = \Lambda\left(\int_0^t \exp\{\boldsymbol{\beta}^\top \boldsymbol{Z}(u)\}\,\mathrm{d}A(u)\right),$$

where $\boldsymbol{\beta} \in \mathbb{R}^p$ is a parameter vector, the function Λ is known, thrice differentiable and decreasing with $\Lambda(0) = 1$ and A is an unknown increasing function restricted to $[0, \tau]$. Different choices of Λ produce different models. Especially, $\Lambda(u) = \exp\{-u\}$ results in the Cox model. As in the Cox model we introduce change-points at thresholds of covariates in the underlying regression function. Thus, we study the model

$$S_{\boldsymbol{Z}}(t) = \Lambda\left(\int_0^t \exp\{\boldsymbol{\beta}_1^\top \boldsymbol{Z}_1(u) + \boldsymbol{\beta}_2^\top \boldsymbol{Z}_2 + \boldsymbol{\beta}_3^\top (\boldsymbol{Z}_2 - \boldsymbol{\xi})^+\}\,\mathrm{d}A(u)\right),$$

where $\boldsymbol{\xi}$ denotes the vector of change-points. The estimation of the parameter is much more complex in this model than in the usual Cox model, since the estimation of the infinite-dimensional parameter, the integrated baseline hazard $A(t)$, can no longer be separated from the estimation of the finite-dimensional parameters by a partial likelihood method. We use a nonparametric maximum likelihood to obtain estimates. Again we can show that the estimates of the finite-dimensional parameters $\boldsymbol{\beta}$ and $\boldsymbol{\xi}$ as well as the estimate of the infinite-dimensional parameter $A(t)$ are $\sqrt{n}-$consistent and asymptotically normal. The theory of empirical processes was a helpful tool for our proofs, since as in the Cox model with change-points the underlying regression function is not differentiable in $\boldsymbol{\xi}$ and hence, the classical approach using a Taylor expansion fails. Especially, we use techniques developed for the general class of M-Estimators. Furthermore, proving the asymptotic properties of the infinite-dimensional parameter involves the theory of linear operators and Fréchet differentiabillity.

1.3 Outline

The thesis is structured as follows.

In Chapter 2, we review the main ideas of survival analysis and present some models from it. Moreover, we discuss properties of the estimates and we introduce some notation we will need in the main chapters.

In Chapter 3, we give an overview of change-point models. Different kinds of change-point models are described and properties of the estimates are discussed.

One of the main chapters is Chapter 4. It contains the Cox model with one single change-point. The estimation procedure for the change-point parameter and the other parameters is illustrated. Furthermore, consistency and asymptotic normality of the estimates are proved and the rate of convergence of the estimates is derived.
In Chapter 5 we extend the Cox model with one change-point to a Cox model with general risk function and more than one change-point. Again, we prove the usual properties of the estimates.
Chapter 6 contains the more complex transformation model with change-points. We present a nonparametric likelihood, which is used to estimate the finite- and infinite-dimensional parameters. The consistency of all estimates is proved in subsection 6.4. Furthermore, the score and information operators are calculated after a reparametrization of our model. As a result, the rate of convergence $\sqrt{n}$ is obtained again and asymptotic normality of the estimates can be established.
In Chapter 7 we apply the Cox model with change-points to different datasets. The first dataset consists of insurance contracts, for which different attributes were recorded. The purpose of our study was to investigate how the different attributes influence the cancellation of a contract. It can be seen that one covariate is misspecified using the classical Cox model such that our model yields a better fit than the classical one. Furthermore, by using a special goodness-of-fit test we are able to describe the functional form of these covariates in a piecewise linear form.
The second dataset contains information about lifetimes of electric motors and covariates like load, current, nominal voltage and r.p.m. The aim of this analysis is to determine the influence of the covariates on the lifetime and to obtain survival functions of different types of motors, including estimates for untested configurations.
The last dataset is the well known PBC dataset described in Fleming & Harrington (1991), which contains data about the survival of patients with primary biliary cirrhosis (PBC). We suggest to use the Cox model with change-points to get a better fit compared to the model used in Fleming & Harrington (1991).
In Chapter 8 some remarks about the models and the applications are given. Furthermore, some open problems are discussed.
In the Appendix some general definitions and results from the theory of empirical processes are stated.
The contents of Chapter 3 will be published in Jensen & Lütkebohmert (2007a). Some of the basic ideas of Chapter 4 have already been published in Gandy et al. (2005) whereas the ideas of Chapter 5 are submitted for publication (see Jensen & Lütkebohmert (2007b)).
The application concerning the electric motor dataset is published in Lütkebohmert et al. (2007).

Chapter 2

Regression Models for Survival Data

2.1 Survival Data

A typical dataset in Survival Analysis is obtained from a collection of individuals which are observed from an entry time of a study until the occurrence of a particular event. A classical example is the observation of patients in a clinical trial. The interest lies in the time period from a certain surgery until death. Recording such data involves some problems. Sometimes the dataset must be analyzed before all patients have died. Patients can die due to other reasons or the individuals can leave the study such that the data is lost for follow-up studies. Therefore, most often such a dataset is not complete and the question is how to handle the data. Leaving out some data would falsify the outcome and thus the concept of right censoring is used. Right censoring is the phenomenon that it is only known that the event of death has not yet happened until a specified time. Hence, one assumes that for each individual i there are two random times: T_i the time of an event and C_i the censoring time. Actually, one observes the minimum V_i of T_i and C_i. The status of the ith individual will be denoted by the indicator $\delta_i = I_{\{T_i \leq C_i\}}$, which is 1 if T_i is observed and 0 if the observation is censored. This setup can also be expressed in terms of counting processes. Let $N_i(t) = \delta_i I_{\{V_i \leq t\}}$ be the counting process which stays 0 if the ith individual does not experience an event and jumps from 0 to 1 at the observed event time T_i. This formulation of the problem is useful, since a counting process admits an intensity $\lambda_i(t)$ which enables us to define different models. Furthermore, by definition the difference between the counting process and the cumulative intensity $\Lambda_i(t) = \int_0^t \lambda_i(s)\,\mathrm{d}s$ is a martingale. Hence, martingale theory can be used for analyzing the data. Another advantage of this setup is that it generalizes immediately to multiple events during a finite interval $[0, \tau]$, where $\tau < \infty$.

In many studies the main interest is in how explanatory variables, so-called covariates, influence the survival of a patient. Explanatory variables can be a certain medical treat-

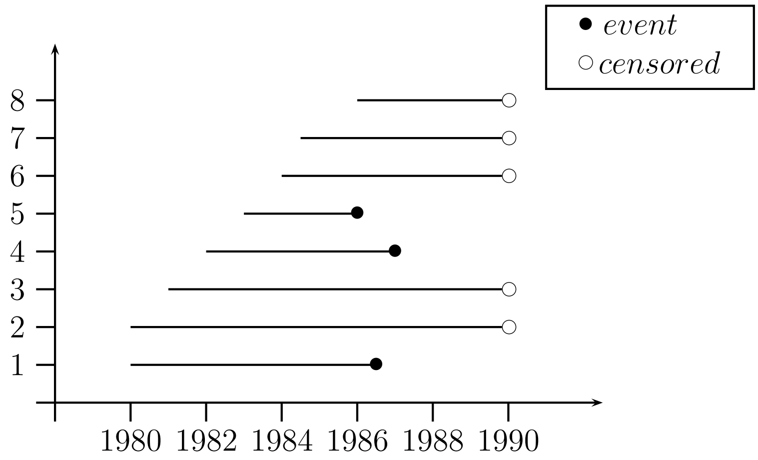

Figure 2.1: Eight observations from a clinical trial, calendar years

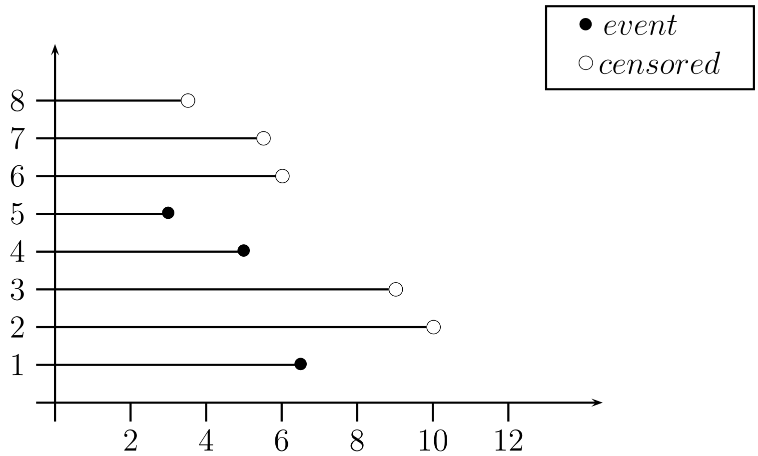

Figure 2.2: Eight observations from a clinical trial, years since operation

ment or known characteristics of a patient like age, blood pressure, etc. Therefore, a typical dataset in survival analysis consists of iid triples (V_i, Z_i, δ_i), $i = 1, \ldots, n.$, where V_i is the observed right censored lifetime, δ_i indicates whether the time is censored or not and Z_i is a covariate. Typical models are regression models since they allow to incorporate covariates and explore their effects on the lifetimes. Moreover, they take right-censoring into account.

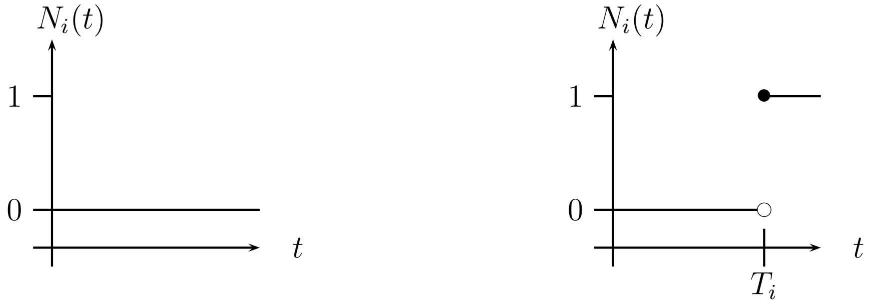

Figure 2.3: Counting processes with at most one event per individual. Left: no event, Right: event at time T_i.

2.2 Notation

In this section we want to introduce some notation. Some of these conventions will be used only in later chapters, whereas others are already needed in the next section. Nevertheless, we will give an exhaustive overview here and will refer to this section, whenever it is useful. One main aspect of this thesis is to provide the asymptotic properties of the estimates. Therefore, stochastic convergence is denoted by $\xrightarrow{\mathrm{P}}$ and convergence in distribution is denoted by $\xrightarrow{\mathrm{d}}$. Matrices and vectors are written in bold face ($\boldsymbol{H}, \boldsymbol{z}$) and their entries are described by H_{ij} and z_i, respectively. If not stated otherwise, the vectors are column vectors. The transpose of a vector is denoted by $\boldsymbol{z}^\top$ and $\mathbf{1} \in \mathbb{R}^k$ is an k-dimensional vector of 1's. By the partial derivatives with respect to a vector $\boldsymbol{x} \in \mathbb{R}^k$ and $\boldsymbol{y} \in \mathbb{R}^m$ the following is meant

$$\frac{\partial}{\partial \boldsymbol{x}} = \left(\frac{\partial}{\partial x_1}, \dots, \frac{\partial}{\partial x_k} \right)^\top$$

and

$$\frac{\partial}{\partial \boldsymbol{x}} \frac{\partial}{\partial \boldsymbol{y}} = \left(\frac{\partial}{\partial x_\nu} \frac{\partial}{\partial y_\mu} \right) \text{ for } \nu = 1, \dots, k,\ \mu = 1, \dots, m.$$

Moreover, if $x_i : \mathbb{R}_+ \to \mathbb{R}$, $i = 1, \dots, k$ is a real valued function, then

$$\int \boldsymbol{x}(t)\,\mathrm{d}t = \left(\int x_1(t)\,\mathrm{d}t, \dots, \int x_k(t)\,\mathrm{d}t \right)^\top$$

and considering a matrix $\boldsymbol{H} \in \mathbb{R}^k \times \mathbb{R}^m$ of functions on $\mathbb{R}_+$

$$\int \boldsymbol{H}(t)\,\mathrm{d}t = \begin{pmatrix} \int h_{11}(t)\,\mathrm{d}t & \cdots & \int h_{1m}(t)\,\mathrm{d}t \\ \vdots & \ddots & \vdots \\ \int h_{k1}(t)\,\mathrm{d}t & \cdots & \int h_{km}(t)\,\mathrm{d}t \end{pmatrix}.$$

The indicator function is written in the form $I_{\{.\}}$. Let a^+ denote the maximum of a and 0. For vectors, this expression has to be read componentwise. Especially, for vectors

$\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{a} \in \mathbb{R}^k$ the following $\boldsymbol{x} I_{\{\boldsymbol{z}>\boldsymbol{a}\}}$ is a short notation for $(x_1 I_{\{z_1>a_1\}}, \cdots, x_k I_{\{z_k>a_k\}})^\top$. The expectation of a vector conditional on a vector also has to be read elementwise. We denote by $\|\cdot\|$ the Euclidean norm of a vector, by $\|\cdot\|_v$ the total variation norm and by $\|\cdot\|_\infty$ the uniform norm. The notation $\Delta N(t)$ is used for $N(t) - N(t-)$, where $N(t-) = \lim_{s\uparrow t} N(s)$.
The symbol $O(x)$ is the usual Landau symbol. Whereas, we write $X_n = O_{\mathrm{P}}(1)$ for sequences (X_n), $n \in \mathbb{N}$ of random variables, if for each $\varepsilon > 0$ there exists a constant $K > 0$, such that $\sup_{n\in\mathbb{N}} \mathrm{P}(|X_n| > K) < \varepsilon$. For sequences $(X_n), n \in \mathbb{N}$, and $(a_n), n \in \mathbb{N}$, of random variables, we say $X_n = o_{\mathrm{P}}(a_n)$ if $X_n/a_n \xrightarrow{\mathrm{P}} 0$ and $X_n = o_{a.s.}(a_n)$ if $X_n/a_n \to 0$ almost surely.
The following notation is used in Chapter 6. Let $X_1, \ldots, X_n$ be a random sample from a probability distribution P on a measurable space $(\mathcal{X}, \mathcal{A})$. The empirical distribution is the discrete uniform measure on the observations. We denote it by $\mathbb{P}_n = n^{-1} \sum_{i=1}^n \delta_{X_i}$, where δ_x is the probability distribution that is degenerate at x. Given a measurable function $f : \mathcal{X} \to \mathbb{R}$, we write $\mathbb{P}_n f$ for the expectation of f under the empirical measure, and $\mathrm{P} f$ for the expectation under P. Thus

$$\mathbb{P}_n f = \frac{1}{n} \sum_{i=1}^n f(X_i), \qquad \mathrm{P}\, f = \int f \, \mathrm{d}P.$$

In this context the $L_r(\mathrm{P})$-norm is often used, which is defined as follows:

$$\|f\|_{P,r} = \left(\mathrm{P}\, |f|^r\right)^{1/r}.$$

The empirical process is given by $\mathbb{G}_n f = \sqrt{n}(\mathbb{P}_n f - \mathrm{P}\, f)$.

2.3 Regression Models

In Survival Analysis the objective may be to compare different treatment effects on the survival time including the information which is available for each individual such as age, sex and various clinical data. This leaves us with a regression problem. Various regression models were proposed in the last 30 years. The most famous ones are the Cox model and the Aalen model. There also exist various other models. We will review some of them here.

2.3.1 The Cox Model

A classical survival time model was proposed by Cox (1972). A counting process approach was first presented by Andersen & Gill (1982). Let $t \in [0, \tau]$, $0 < \tau < \infty$. In this model the vector of covariates $\boldsymbol{Z}_i(t)$ is related to the counting process $\boldsymbol{N}(t) = (N_1(t), \dots, N_n(t))^\top$ by the intensity $\boldsymbol{\lambda}(t) = (\lambda_1(t), \dots, \lambda_n(t))^\top$ which is specified as follows:

$$\lambda_i(t) = \lambda_0(t) R_i(t) \exp\{\boldsymbol{\beta}^\top \boldsymbol{Z}_i(t)\} \tag{2.1}$$

The observed vector of covariates is a p-dimensional predictable and locally bounded stochastic process, the parameter $\boldsymbol{\beta} \in \mathbb{R}^p$ is an unknown parameter vector, $R_i(t)$ is a process taking only values 1 or 0 indicating whether an individual is at risk or not and $\lambda_0(t)$ is the baseline hazard function. The vector of regression parameters $\boldsymbol{\beta}$ and the integrated baseline hazard $\Lambda_0(t) = \int_0^t \lambda_0(u)\,\mathrm{d}u$ have to be estimated. Using a partial likelihood it is possible to estimate the parameter $\boldsymbol{\beta}$ separately from the cumulative baseline intensity $\Lambda_0(t)$. The partial likelihood is given by

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \prod_{t=0}^{\tau} \left\{ \frac{\exp\{\boldsymbol{\beta}^\top \boldsymbol{Z}_i(t)\}}{\sum_{j=1}^n R_j(t) \exp\{\boldsymbol{\beta}^\top \boldsymbol{Z}_j(t)\}} \right\}^{\Delta N_i(t)} \tag{2.2}$$

with $\Delta N_i(t) = N_i(t) - N_i(t-)$. Clearly, $\hat{\boldsymbol{\beta}}_n$ also maximizes the log partial likelihood

$$\log L(\boldsymbol{\beta}) = \sum_{i=1}^{n} \int_0^\tau \boldsymbol{\beta}^\top \boldsymbol{Z}_i(t)\,\mathrm{d}N(t) - \int_0^\tau \log\left(\sum_{j=1}^n R_j(t) \exp\{\boldsymbol{\beta}^\top \boldsymbol{Z}_j(t)\} \right) \mathrm{d} \sum_{i=1}^n N_i(t).$$

The cumulative hazard function $\Lambda_0(t)$ can be estimated by the well known Breslow estimator (see Andersen et al. (1993)) which is given by

$$\hat{\Lambda}_0(t) = \int_0^t \frac{\mathrm{d}(n^{-1} \sum_{i=1}^n N_i(s))}{n^{-1} \sum_{j=1}^n R_j(s) \exp\{\hat{\boldsymbol{\beta}}_n^\top \boldsymbol{Z}_j(s)\}}.$$

Under some regularity conditions it can be shown that the estimate $\hat{\boldsymbol{\beta}}_n$ is $\sqrt{n}$-consistent and that $\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0)$ converges in distribution to a normal distribution. Furthermore, weak convergence of $\sqrt{n}(\hat{\Lambda}_n(t) - \Lambda_0(t))$ can be established.

In the classical model the link function is given as an exponential function. An extension is to consider instead a general known function $r : \mathbb{R} \to [0, \infty)$ which has been studied by Prentice & Self (1983). The hypothesis of a linear functional form of the covariates has to be abandoned for several applications. An alternative is to consider a general unknown function $\psi(\boldsymbol{Z})$ instead of $\boldsymbol{\beta}^\top \boldsymbol{Z}$ in (2.1). This approach was presented by Chen & Zhou (2007). Another way to obtain the functional form is to use an underlying regression function with bent-line change-points according to the covariates. Details will be given in Chapter 4.

2.3.2 The Aalen Model

The difference between the Cox model and the Aalen model is the construction of their intensity processes which link covariates to counting processes. In the Aalen model the covariates are given in the form of a matrix $\boldsymbol{Y}(t) = (Y_{ij}(t)), i = 1, \ldots, n,\ j = 1, ..., p,\ p \leq n$, of locally bounded predictable processes. The covariate $Y_{ij}(t)$ is set equal to 0 if the individual i is not at risk. The model is characterized by the intensity

$$\lambda_i(t) = \sum_{j=1}^{p} Y_{ij}(t)\alpha_j(t), \quad t \in [0, \tau],$$

where $\alpha_j(t)$ are unknown deterministic baseline intensities, which need to be estimated. An estimator for the integrated baseline intensity $\mathbf{A}(t) = \int_0^t \boldsymbol{\alpha}(s)\,\mathrm{d}s$ is given by a generalized Nelson-Aalen estimator

$$\hat{\mathbf{A}}(t) = \int_0^t \mathbf{Y}^{-}(s)\,\mathrm{d}\boldsymbol{N}(s),$$

where $\mathbf{Y}^{-}(t)$ is a generalized inverse of $\mathbf{Y}(t)$. In the case that $\mathbf{Y}(t)$ has full rank, we can choose $\mathbf{Y}^{-}(t) = (\mathbf{Y}^{\top}(t)\mathbf{Y}(t))^{-1}\mathbf{Y}^{\top}(t)$. Now the baseline intensity $\boldsymbol{\alpha}(t)$ can be estimated by smoothing procedures, e.g. one can use kernel smoothers.

2.3.3 Multiplicative-Additive Hazards Models

The previous intensity models postulate different relationships between the covariates and the hazard. But sometimes it is not clear, which one is to be preferred. Maybe a combination of both models presents specific data in a better way. There exist various ways to combine the two models above. One possibility is to add up the basic models, which leads to the proportional excess hazard model, where the additive part can be thought of as modeling the baseline mortality while the multiplicative part describes the excess risk due to different exposure levels. Various authors have studied such a combination, e.g. Lin & Ying (1995) considered the following intensity model

$$\lambda(t) = R(t)\left[g(\boldsymbol{X}^{\top}(t)\boldsymbol{\alpha}) + \lambda_0(t)h(\boldsymbol{Z}^{\top}\boldsymbol{\beta})\right],$$

where $R(t)$ is an at risk indicator, $(\boldsymbol{X}^{\top}(t), \boldsymbol{Z}^{\top}(t))^{\top}$ is a covariate vector, $(\boldsymbol{\alpha}^{\top}, \boldsymbol{\beta}^{\top})^{\top}$ is a vector of regression coefficients and $\lambda_0(t)$ is an unspecified baseline hazard. Both functions g and h are assumed to be known. Sasieni (1996) studied the model

$$\lambda(t) = R(t)\left[\alpha(t, X) + \lambda_0(t)\exp\{\boldsymbol{Z}^{\top}\boldsymbol{\beta}\}\right],$$

where $\alpha(t, X)$ is the background rate of the mortality in a control population and is assumed to be known. Estimates for the unknown parameter $\boldsymbol{\beta}$ and $\Lambda_0(t) = \int_0^\tau \lambda_0(s)\,\mathrm{d}s$ can be derived. Martinussen & Scheike (2002) consider a more flexible model

$$\lambda(t) = R(t)\left[\boldsymbol{X}^\top(t)\alpha(t) + \rho(t)\lambda_0(t)\exp\{\boldsymbol{Z}^\top(t)\boldsymbol{\beta}\}\right],$$

where both $R(t)$ and $\rho(t)$ are at risk indicators, $\alpha(t)$ is a time varying regression function, $\lambda_0(t)$ is the baseline hazard of the excess term and $\boldsymbol{\beta}$ is a vector of relative risk coefficients. A different way of combining the two models is to multiply them. Other approaches were made by Dabrowska (1997) and Scheike & Zhang (2002).

2.3.4 Frailty Models and Transformation Models

In frailty models the intensity process depends partly on an unobservable random variable. Usually, the frailty is modeled by an unobservable random variable acting multiplicatively on the intensity. One has to distinguish between two cases of frailty models. In the univariate case we consider life times of independent individuals where the frailty describes the influence of unobserved risk factors, i.e. we observe survival times $T_1, \ldots, T_n$ and these have, conditional on frailty variables $W_1, \ldots, W_n$, hazards $W_i\alpha(t)$, for some baseline $\alpha(t)$. The frailties are supposed to be unknown and hence their distributions have to be deduced from the hazard functions by integration. Thus the observed hazards are given by $E[W_i|T_i > t]\alpha(t)$. Since the first term is time dependent the observed hazard can be quantitatively different from the conditional hazard we describe below. In the multivariate case, where the frailty is common to a group of individuals like families, the frailty induces a correlation between the individuals in the group. Such a shared frailty model is given as follows: Let T_{ij}, $i = 1, \ldots, n$ be the survival times of the jth individual in the ith group. Then the model is given by

$$\lambda_{ij}(t|W_i) = W_i\alpha(t),$$

where $W_1, \ldots, W_n$ are the frailty variables and $\alpha(t)$ is a baseline hazard. The most common choice for the distribution of W_i is a gamma distribution with mean one and an unknown variance. But other distributions are also possible, see Hougaard (2000). The multivariate case is the more common approach (see Clayton (1978), Nielsen et al. (1992)). In terms of the Cox model we have a semiparametric frailty model for which the conditional hazard function for independent, possibly censored survival times $V_1, \ldots, V_n$ is given by

$$\lambda_i(t|\boldsymbol{Z}_i) = W_i R_i(t)\exp\{\boldsymbol{\beta}^\top \boldsymbol{Z}_i\}\lambda_0(t),$$

where $\boldsymbol{Z}_i$ is a covariate vector for the ith individual, $\boldsymbol{\beta}$ the regression parameter vector, $\lambda_0(t)$ the baseline function and $R_i(t)$ an at-risk indicator. In this case the frailty term represents the neglected common covariates.

There has been a number of suggestions on how to estimate the parameters. One approach is via the EM-algorithm, whereas another approach is via a nonparametric maximum likelihood method.

A more general class of models is that of so-called transformation models which involve one or more monotone transformations $\phi : \mathbb{R} \to \mathbb{R}$. Semiparametric frailty models are examples of this class. One special case is a linear transformation regression model. Suppose that $\boldsymbol{Z}$ is a random vector on $\mathbb{R}^p$ and ε is some nuisance random variable with distribution function F. Furthermore, assume that for some $\boldsymbol{\beta} \in \mathbb{R}^p$ the following linear relationship holds

$$Y = -\boldsymbol{\beta}^\top \boldsymbol{Z} + \varepsilon.$$

Now, the random vector $(\boldsymbol{Z}, U)$ can be observed, where $U = \phi^{-1}(Y)$ for some transformation function ϕ. Thus, we can write equivalently

$$\phi(U) = -\boldsymbol{\beta}^\top \boldsymbol{Z} + \varepsilon.$$

Different choices of the distribution function F of ε result in different models. For example, if $e^\varepsilon \sim$Pareto(η), i.e. $P(\varepsilon \geq t) = (1 + \eta e^t)^{-1/\eta}$ then we obtain a semiparametric Pareto regression model, which was studied by Clayton & Cuzick (1985). This model can also be viewed as a Cox regression model with frailty $W > 0$. To see this relationship consider the hazard function conditional on $\boldsymbol{Z}, W$

$$\lambda(u|\boldsymbol{Z}, W) = W \exp\{\boldsymbol{\beta}^\top \boldsymbol{Z}\}\lambda_0(u)$$

or equivalently

$$\Lambda(u|\boldsymbol{Z}, W) = W \exp\{\boldsymbol{\beta}^\top \boldsymbol{Z}\}\Lambda_0(u).$$

Thus the survival function is given by $S(u|\boldsymbol{Z}, W) = \exp\{-W \exp\{\boldsymbol{\beta}^\top \boldsymbol{Z}\}\Lambda_0(u)\}$. If W follows a $\Gamma(1/\eta, 1/\eta)$ distribution, then

$$S(u|\boldsymbol{Z}) = \left[1 + \eta \exp\{\boldsymbol{\beta}^\top \boldsymbol{Z}\}\Lambda_0(u)\right]^{-1/\eta}.$$

Consequently, if $e^\varepsilon \sim$Pareto(η), then

$$\exp\{\boldsymbol{\beta}^\top \boldsymbol{Z}\}\Lambda_0(U) = \exp\{\varepsilon\} \iff \log \Lambda_0(U) = -\boldsymbol{\beta}^\top \boldsymbol{Z} + \varepsilon. \tag{2.3}$$

Choosing $\phi(U) = \log \Lambda_0(U)$ we get the linear relationship as described above again. In the Cox model (2.3) holds, if $e^{\varepsilon} \sim$Exp(1).
A general description of linear transformation models is in terms of the survival function conditional on some covariate vector $\boldsymbol{Z}$. For example, Kosorok et al. (2004) considered the model

$$S(t|\boldsymbol{Z}) = \Lambda_\gamma \left(\int_0^t \exp\{\boldsymbol{\beta}^\top \boldsymbol{Z}(s)\} \, \mathrm{d}A(s) \right),$$

where $A(s)$ denotes the cumulative baseline hazard function, Λ_γ is the Laplace transform of a random variable W and γ is an unknown parameter. They use a nonparametric likelihood method to obtain estimates and to prove their asymptotical properties. Another approach is given by Slud & Vonta (2004). They prove consistency of the nonparametric maximum likelihood estimator for the model

$$S(t|\boldsymbol{Z}) = \exp\{-G(\exp\{\boldsymbol{Z}^\top \boldsymbol{\beta}\} A(t))\},$$

where G is assumed known and the other functions and parameters are given as in the model of Kosorok et al. (2004). Bagdonavičius & Nikulin (2002) consider among other models

$$S(t|\boldsymbol{Z}) = G \left(\int_0^t \exp\{\boldsymbol{\beta}^\top \boldsymbol{Z}_1(u)\} \, \mathrm{d}G^{-1}(S_0(u)) + \int_0^t \boldsymbol{\gamma}^\top \boldsymbol{Z}_2(u) \, \mathrm{d}u \right),$$

where G is some survival function and S_0 is an unknown baseline survival function. They use estimating equations to study the asymptotical properties of their estimates.
Other transformation models can be set up without using the linear relationship. One example is a Copula model. Suppose that $C_{\boldsymbol{\theta}}$ is a distribution on $[0,1]^2$ for $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^k$. Moreover, we assume that $C_{\boldsymbol{\theta}}$ has uniform marginals. If $(U,V) \sim C_{\boldsymbol{\theta}}$ for some $\boldsymbol{\theta} \in \Theta$, we observe $X = (S,T) = (\sigma^{-1}(U), \phi^{-1}(V)) = G^{-1}(U), H^{-1}(V))$, where G and H are distribution functions on $\mathbb{R}$ and hence σ and ϕ are transformation functions. Thus the joint distribution function of X is given by

$$F_{S,T}(s,t) = C_{\boldsymbol{\theta}}(G(s), H(t)).$$

For more details we refer to Bickel et al. (1998).

Chapter 3

Change-Point Models

This chapter gives an overview over different change-point models and is based on Jensen & Lütkebohmert (2007a). Our main interest in this thesis lies on change-point models in hazard rate and regression models, but due to completeness we will depict the more familiar change-point models form quality control as well.

Change-point models have originally been developed in connection with applications in quality control, where a change from the *in-control* to the *out-of-control* state has to be detected based on the available random observations. Up to now various change-point models have been suggested for a broad spectrum of applications like quality control, reliability, econometrics or medicine.

The general change-point problem can be described as follows: A random process indexed by time is observed and we want to investigate whether a change in the distribution of the random elements occurs. In other words we are interested in determining whether the observed stochastic process is homogeneous or not. Formally, in the discrete time case, let $X_1, X_2, \ldots$ denote a sequence of independent random variables, where the elements $X_1, \ldots, X_{\theta-1}$ have an identical distribution function F_0 and $X_\theta, X_{\theta+1}, \ldots$ are distributed according to F_1 and the change-point θ is unknown. Several statistical tests of the null hypothesis $F_0 = F_1$ against the alternative $F_0 \neq F_1$ for some $\theta > 1$ have been suggested. In addition, estimates for the change-point have been proposed and their properties have been investigated.

Change-point problems can be classified in different ways. Approaches in the classical framework as in the Bayesian framework have been made. Also, there exist models in continuous time as well as in discrete time. Furthermore, the analysis of change-points can be partitioned in sequential and posteriori detection models (ex post analysis). And of course, the problem can be viewed at in a parametric or nonparametric context. Another characterization is whether only one change-point exists or more than one. The early work in change-point analysis is described in the survey article of Zacks (1982). Other comprehensive reviews are given in Bhattacharya (1994) and in the book of Brodsky & Darkhovsky

(1993) for nonparametric models. For an overview of limit theorems in change-point problems we refer to Csörgő & Horváth (1997). Here we want to concentrate shortly on a review of models and methods for sequentially observed data and will explain in more detail change-points in regression and hazard rate models.

3.1 Detection of a Change-Point in Sequentially Observed Data

The aim of so-called disorder or detection problems is to detect the change-point "as soon as possible" but avoiding too many false alarms. We distinguish discrete time and continuous time models.

3.1.1 Discrete Time Models

In the discrete case let $X_1, X_2, \ldots$ be a sequence of independent random variables which are observed sequentially. The first $X_1, \ldots, X_{\theta-1}$, $\theta > 1$ are distributed according to some known distribution F_0 while $X_\theta, X_{\theta+1}, \ldots$ have some known distribution function $F_1 \neq F_0$. The change-point θ is unknown. The time of alarm (a change-point has occurred) is determined by a stopping rule which takes the random observations into account. Concerning the change-point there exist Bayesian and non-Bayesian approaches. One of the first non-Bayesian methods is the CUSUM-procedure proposed by Page (1954), which was further investigated by Lorden (1971) and Moustakides (1986).
A Bayesian formalization of the disorder problem goes back to Shiryaev (1963). He postulated that the change-point θ has a geometric a priori distribution with some parameter p and considered the following risk function $R(\tau)$ for stopping at τ: $R(\tau) = P(\tau < \theta) + cE(\tau - \theta)^+$. Here the penalty costs of a false alarm are normed to 1, and the costs for the delay of stopping after the change-point are c per time unit. Now the Bayes stopping rule is to stop at the smallest n for which the posterior probability $\Pi_n = P(\theta \leq n | X_1, ..., X_n)$ of a change up to n is greater than some threshold A for some $0 < A < 1$.

3.1.2 Continuous Time Model

Shiryaev (1963) was also one of the first to present a model in continuous time in a Bayesian framework: the change-point is assumed to be a random variable with some prior distribution. Shiryaev considered the following observation process

$$W_t = B_t + r(t - \theta)^+, \quad t \in [0, \infty),$$

where B denotes a standard Brownian motion, r is a known fixed constant and θ is an unknown (random) change-point, which is assumed to be independent of B and to have a mixed exponential prior distribution: $P(\theta = 0) = p$ and $P(\theta > t) = (1-p)e^{-\lambda t}$, $p \in [0,1)$, $\lambda > 0$, $t \geq 0$. The stopping time τ with respect to the filtration generated by W should signal the change in the drift as soon as possible. The speed of detection is measured by the risk function $R(\tau)$, which is the same as in the discrete time case. A Bayes solution τ^* should minimize the risk

$$R(\tau^*) = \inf_{\tau} R(\tau).$$

The optimal stopping time τ^* can be determined by means of the posterior distribution $\Pi_t = P(\theta \leq t \,|\, \mathcal{F}_t^W)$, with $\mathcal{F}_t^W = \sigma\{W_s : s \leq t\}$. Then the optimal stopping time is

$$\tau^* = \inf\{t > 0 \,|\, \Pi_t \geq p^*\},$$

for some properly chosen $p^* \in [0,1)$. Details about this approach can be found in Shiryaev (1978). An explicit expression for the optimal threshold p^*, and further ramifications can be found in Beibel (1994, 1996).
Another type of change-point problems has been studied in recent years, namely the Poisson disorder problem. Instead of considering a Wiener process with changing drift a Poisson process with changing intensity is observed. Then the problem is to determine a stopping time which signals a change of the intensity of the observed Poisson process. Formally, a point process (T_n), $n \in \mathbb{N}$ and its corresponding counting process $N_t = \sum_{n=1}^{\infty} I_{\{T_n \leq t\}}$ are observed, where I is the indicator function. At an unknown random time θ the intensity of N switches from μ_0 to $\mu_1 > \mu_0$. This means, that if θ is given, N is a Poisson process with intensity μ_0 up to θ and a Poisson process with intensity μ_1 after θ. For more details we refer to Peskir & Shiryaev (2002), Brown & Zacks (2006) and Herberts & Jensen (2004).

3.2 Change-Points in Regression and Hazard Rate Models

3.2.1 Regression Models

In the literature two different types of change-point regression models can be found: On the one hand so called time-varying regression models, in which the model parameters change at some unknown point in time, and on the other hand two-phase regression models. Both models are presented briefly in the following.

In the time-varying model a change in the regression coefficients takes place from the early to the late observations of a sequence (X_n), $n \in \mathbb{N}$. Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be independent random vectors. Then the model in the random design is given by

$$Y_i = \begin{cases} \alpha_0 + \alpha_1 X_i + \epsilon_i, & i \leq \tau \\ \beta_0 + \beta_1 X_i + \epsilon_i, & i \geq \tau + 1 \end{cases}$$

where (X_i) and (ϵ_i), $i = 1, \ldots, n$ are mutually independent iid sequences with $E(\epsilon_i) = 0$ and $E(\epsilon_i^2) = 1$ and $(\alpha_0, \alpha_1) \neq (\beta_0, \beta_1)$. If $1 \leq \tau \leq n-1$, then τ is a change-point. This design is called fixed if the sequence (X_i), $i = 1, \ldots, n$ is non-stochastic.

A two-phase regression model is a regression model with piecewise linear regression functions over two different domains of the design-variable. The random design of a two-phase regression model is given by

$$Y_i = (\alpha_0 + \alpha_1 X_i) I_{\{X_i \leq \tau\}} + (\beta_0 + \beta_1 X_i) I_{\{X_i > \tau\}} + \epsilon_i = m(X_i) + \epsilon_i. \tag{3.1}$$

The two-phase regression models can be classified further into a restricted and an unrestricted case. In the restricted case the regression function f is continuous but not differentiable, whereas in the unrestricted case the regression function is discontinuous. The discontinuity can be expressed in form of a fixed jump size or a contiguous jump size, in which the jump size tends to zero as the sample size tends to infinity.

Hinkley (1971) was one of the first authors to investigate a maximum likelihood estimator of the point of intersection for the special case of two line segments under normally distributed errors. A generalization of his model with multiple change-points was considered by Feder (1975a,b), who investigated least squares estimates and showed that these estimates are consistent under suitable identifiability assumptions and the asymptotic distributions of these estimates are obtained by "classical" methods.

Koul & Qian (2002), Koul et al. (2003) considered M-estimators in the unrestricted two-phase random design with a fixed jump size. The M-process corresponding to a function $\phi : \mathbb{R} \to [0, \infty)$ is defined as

$$M_n(\boldsymbol{\theta}) = \sum_{i=1}^{n} \phi(Y_i - m(X_i, \boldsymbol{\theta})),$$

where $m(X; \boldsymbol{\theta})$ is the linear regression function of model (3.1) and the M-estimator $\hat{\boldsymbol{\theta}}$ is given as the minimizer of the M-process:

$$M_n(\hat{\boldsymbol{\theta}}) = \inf_{\boldsymbol{\theta}} M_n(\boldsymbol{\theta}) \quad \text{a.s.}$$

They showed that the estimate of the jump point converges with rate $O_p(n^{-1})$, whereas the rate of convergence of the coefficient parameters is $O_p(n^{-1/2})$. The normalized M-process is asymptotically equivalent to the sum of two processes. One is a quadratic form in the standardized coefficient parameter vector, the other is a jump point process in the change-point parameter. This result can be exploited to show weak convergence. The suitably standardized M-estimator of the change-point converges weakly to the minimizer of a compound Poisson process. The estimates of the regression coefficients are asymptotically normal and independent of the jump point M-estimator. This is remarkable because the results differ from the restricted and unrestricted contiguous non-random design cases. All models considered above assumed a parametric setting. Of course, there exist various nonparametric models as well. Müller (1992) studied the following fixed design nonparametric regression model

$$Y_{in} = g(t_{in}) + \epsilon_{in}, \quad t_{in} \in [0,1],\ 1 \le i \le n,$$

where Y_{in} are noisy measurements of the regression function g taken at points t_{in} and $\epsilon_{in} \sim \mathcal{N}(0, \sigma^2)$ are iid errors. The assumption is made that there is a change-point for the νth derivative $g^{(\nu)}$ at τ, $0 < \tau < 1$ in the following sense: There exists a function $f \in \mathcal{C}^{k+\nu}([0,1])$ with $\nu \ge 0$ and $k \ge 2$ an even integer, such that

$$g^{(\nu)}(t) = f^{(\nu)}(t) + \Delta_\nu I_{[\tau,1]}(t), \quad \Delta_\nu > 0,\ 0 \le t \le 1.$$

The case $\Delta_\nu < 0$ can be treated analogously. Now, the jump size at the possible change-point τ of the νth derivative of g is given by

$$\Delta_\nu = g_+^{(\nu)}(\tau) - g_-^{(\nu)}(\tau),$$

where $g_+^{(\nu)}(x) = \lim_{y \downarrow x} g^{(\nu)}(y)$ and $g_-^{(\nu)}(x) = \lim_{y \uparrow x} g^{(\nu)}(y)$ are the one-sided limits of the derivative $g^{(\nu)}(x)$. Hence, the idea is to base the inference of the change-points on differences between the left and right sided estimates of $g^{(\nu)}(t)$, which can be done by suitably chosen one sided kernel estimates. The location of the maximum of these differences is a reasonable estimator of the location of the change-point. Let τ be an element of a closed interval $T \subset (0,1)$. Then the estimator is

$$\hat{\tau} = \inf\{\rho \in T : \hat{\Delta}_\nu(\rho) = \sup_{x \in T} \hat{\Delta}_\nu(x)\}.$$

In this setting Müller (1992) proved weak convergence of the estimator $\hat{\tau}$. Loader (1996) considered a similar nonparametric regression model in which the mean function may have a discontinuity at an unknown point. His estimate is similar in principle to that

studied by Müller (1992). But since he imposed different conditions on the kernel K, his estimate has different properties. It is shown that the change-point estimate converges in probability with rate $O_P(n^{-1})$ and that it has the same asymptotic distribution as maximum likelihood estimates in parametric models.
The same rate of convergence is attained by Müller & Song (1997) in a two-step estimation of the change-point in a nonparametric fixed design regression model with fixed jump size, whereas the rate of convergence in the contiguous case is $O_P(n^{-1}\Delta_n^{-2})$, where Δ_n is a sequence of jump sizes which tends to zero.
Another important problem in modeling data is the question whether an unknown function, which cannot be specified parametrically, should be modeled as a globally smooth function or a smooth function with isolated change-points. Müller & Stadtmüller (1999) proposed statistics which provide relevant information for this decision.

3.2.2 Hazard Rate Models

Hazard rate models often occur in medical follow up studies after major surgery. The simplest one with a change-point can be expressed as follows

$$\lambda(t) = \begin{cases} \lambda_1 & t \leq \tau \\ \lambda_2 & t > \tau \end{cases}, \ t \geq 0\,,\ \tau \geq 0$$

with constants $\lambda_1, \lambda_2 > 0$ and change-point τ. A first attempt to estimate these three parameters was made by Anderson & Senthilselvan (1982). They investigated as a special case of this simple model an extended Cox model with $\lambda_1 = e^{\boldsymbol{\alpha}^\top \boldsymbol{Z}}$, $\lambda_2 = e^{\boldsymbol{\gamma}^\top \boldsymbol{Z}}$, where $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$ are parameter vectors and $\boldsymbol{Z}$ is a vector of covariates. The parameters are estimated by the conditional log-likelihood given the value of τ and then the baseline hazard $\lambda(t)$ is estimated by a penalized maximum likelihood method conditioning on the parameter estimates. Liang et al. (1990) proposed a slightly different Cox model

$$\lambda(t) = \lambda_0(t)\exp\{(\beta + \theta I_{\{t \leq \tau\}})Z + \boldsymbol{\gamma}^\top \boldsymbol{X}\},$$

where Z is a one-dimensional covariate which should be included in possibly different magnitudes over time, $\boldsymbol{X}$ is another confounding covariate vector and the change-point at an unknown time is given by τ. They tested the hypothesis of $H_0:\ \theta = 0$ by using a test statistic

$$M = \sup_{\tau \in [a,b]} S(\tau),$$

where S is a function of the first two derivatives of the partial log likelihood function with respect to β and γ.

A further Cox model is presented by Luo & Boyett (1997):

$$\lambda(t) = \lambda_0(t) \exp\{\beta I_{\{X \leq \theta\}} + \boldsymbol{\alpha}^\top \boldsymbol{Z}\},$$

where a constant is added to a covariate beyond a threshold, which is characterized by a random variable X. They proved consistency of their partial MLE.
A Cox model for independent and identically distributed right censored survival times with a change-point according to the unknown threshold of a covariate was introduced by Pons (2003):

$$\lambda(t) = \lambda_0(t) \exp\{\boldsymbol{\alpha}^\top \boldsymbol{Z}_1(t) + \boldsymbol{\beta}^\top \boldsymbol{Z}_2(t) I_{\{Z_3 \leq \zeta\}} + \boldsymbol{\gamma}^\top \boldsymbol{Z}_2(t) I_{\{Z_3 > \zeta\}}\}.$$

In this model it is shown that the partial MLE of the change-point ζ is n-consistent, i.e. the rate of convergence is $O_p(n^{-1})$. Such a rate was also attained for the change-point in the unrestricted two-phase random linear regression design with a fixed jump size (see Koul & Qian (2002)). Furthermore, Pons (2003) proved that the estimates of the regression parameter vectors $\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}$ are $\sqrt{n}$-consistent and that $n(\hat{\zeta}_n - \zeta)$ converges weakly to a random variable $\hat{\nu}_Q$ which is a maximizer of a certain jump process. The estimates of the regression parameters are asymptotically normal.
Gandy et al. (2005) also investigated an extended Cox model. But instead of a discontinuous underlying regression function they considered a continuous underlying regression function, which is not differentiable at the change-point ξ. Such a change-point is called a bent-line change-point. The intensity of this model is the following

$$\lambda(t) = \lambda_0(t) \exp\{\boldsymbol{\beta}_1^\top \boldsymbol{Z}_1(t) + \beta_2 Z_2 + \beta_3 (Z_2 - \xi)^+\}.$$

As the model of Pons (2003) can be compared with an unrestricted two-phase random linear regression model with fixed jump, the model of Gandy et al. (2005) can be compared with a restricted two-phase regression model. Hence, the rate of convergence of the bent-line change-point parameter is different to the one which is obtained in the model with a jump. In the model with a bent-line change-point the rate of convergence of all parameters is $O_{\mathrm{P}}(n^{-1/2})$ and all parameter are asymptotically normal. More details about the estimation of the parameter and the properties of the estimates are given in Chapter 4, which is one of the main chapter in this thesis.
Other approaches relying on the Cox model have recently been suggested by Dupuy (2006). He considered a model with a change-point in both hazard and regression parameters. Estimates of the change-point, hazard and regression parameters are proposed and shown to be consistent.

Chapter 4

Cox Model with a Bent-Line Change-Point

In this chapter, we consider a new extended Cox model with a single change-point in one of the covariates. The change-point is a bent-line change-point, i.e. the underlying regression function is linear and continuous but not differentiable at that point. Thus, a change-point specifies the unknown threshold at which the influence of a covariate shifts. In contrast to this notion there exist change-points at which the corresponding regression function jumps, see Pons (2003). For a better understanding of the following chapters we first include only one change-point and thoroughly describe the techniques used.
We derive estimates of the regression and change-point parameters and prove their asymptotic properties, namely consistency and asymptotic normality. Furthermore, we investigate the rate of convergence of the estimates. Especially, the rate of convergence of the change-point parameter differs from the one obtained in a model, in which the underlying regression function is discontinuous. Moreover, we provide a proof of weak convergence of $\sqrt{n}(\hat{\Lambda}_n(t) - \Lambda_0(t))$, where $\hat{\Lambda}_n(t)$ is the Breslow estimator of the cumulative baseline intensity.

4.1 Model Setup

Let $[0, \tau]$, $0 < \tau < \infty$ be a fixed time interval on which all stochastic processes are defined and let all stochastic elements be given on a filtered probability space $(\Omega, \mathcal{F}, \mathrm{P})$. We assume that the n-variate counting process $\boldsymbol{N}(t) = (N_i(t), i = 1, ..., n)$ has independent and identically distributed elements which have no common jumps. The counting process admits an intensity $\boldsymbol{\lambda}$ and counts only one event per subject. More precisely, we assume that $(N, R, \boldsymbol{Z}_1(t), Z_2)$, $(N_i, R_i, \boldsymbol{Z}_{1i}(t), Z_{2i})$, $i = 1, ..., n$ are iid vectors of random quantities, where $\boldsymbol{Z}_1, R$ are adapted left continuous processes with right-hand limits. Furthermore,

$\boldsymbol{M}(t) = \boldsymbol{N}(t) - \int_0^t \boldsymbol{\lambda}(s)\,\mathrm{d}s$ is a vector of martingales on the time interval $[0,\tau]$. The components of $\boldsymbol{\lambda}$ are defined by

$$\lambda_i(t,\boldsymbol{\theta}) = \lambda_0(t)R_i(t)\exp\left\{\boldsymbol{\beta}_1^\top \boldsymbol{Z}_{1i}(t) + \beta_2 Z_{2i} + \beta_3(Z_{2i}-\xi)^+\right\},$$

where $\boldsymbol{\theta} = (\xi, \boldsymbol{\beta}^\top)^\top$ with $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^\top, \beta_2, \beta_3)^\top \in \mathcal{B} \subset \mathbb{R}^{p+2}$ is the vector of regression parameters and $\xi \in \mathbb{R}$ indicates the change-point. The baseline intensity is denoted by $\lambda_0(t)$ and $R_i(t)$ is the at-risk indicator, which is 1 if the individual is under risk and 0 otherwise. Thus, we have a regular Cox model for the covariate vector $\boldsymbol{Z}_{1i}(t)$ and a change of the influence of the covariate Z_{2i} at ξ from β_2 to $\beta_2+\beta_3$. For brevity, we consider

$$\tilde{\boldsymbol{Z}}_i(t;\xi) = \left(\boldsymbol{Z}_{1i}^\top(t), Z_{2i}, (Z_{2i}-\xi)^+\right)^\top.$$

The change-point ξ is a parameter, which lies in a compact interval $[\xi_1,\xi_2]$ of known parameters ξ_1 and ξ_2. For applications this is not a great constraint. Moreover, the true parameter values $\boldsymbol{\theta}_0 = (\xi_0, \boldsymbol{\beta}_0^\top)^\top$ are supposed to be identifiable, meaning that $\beta_{30} \neq 0$. In this model the regression parameter $\boldsymbol{\beta}$, the change-point parameter ξ and the baseline intensity function $\lambda_0(t)$ have to be estimated. Using a partial likelihood the estimation of the finite-dimensional parameter can be separated from the estimation of the infinite-dimensional parameter $\lambda_0(t)$.

4.2 Estimation

In this new Cox model $\boldsymbol{\theta}_0$ is estimated by the value $\hat{\boldsymbol{\theta}}_n$ that maximizes the logarithm of the partial likelihood

$$\log L(\boldsymbol{\theta}) = \sum_{i=1}^n \int_0^\tau \boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}_i(t;\xi)\,dN_i(t) - \int_0^\tau \log\left(\sum_{i=1}^n R_i(t)\exp(\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}_i(t;\xi))\right)\mathrm{d}\left(\sum_{i=1}^n N_i(t)\right).$$

The maximization can be carried out in two phases:
For fixed ξ, let $\hat{\boldsymbol{\beta}}_n(\xi) = \arg\max_{\boldsymbol{\beta}\in\mathcal{B}} \log L(\xi,\boldsymbol{\beta})$ and $\log L(\xi) = \log L(\xi, \hat{\boldsymbol{\beta}}_n(\xi))$. Then ξ_0 can be estimated by $\hat{\xi}_n$ satisfying

$$\hat{\xi}_n = \underset{\xi\in[\xi_1,\xi_2]}{\arg\max}\ \log L(\xi).$$

Hence, the partial maximum likelihood estimator of $\boldsymbol{\theta}_0$ is $\hat{\boldsymbol{\theta}}_n = (\hat{\xi}_n, \hat{\boldsymbol{\beta}}_n^\top)^\top$, where $\hat{\boldsymbol{\beta}}_n = \hat{\boldsymbol{\beta}}_n(\hat{\xi}_n)$.
In the usual approach to show consistency and asymptotic normality of the parameters the score function of $\log L(\boldsymbol{\theta})$ has to be calculated. In our case this is impossible, since

$\log L(\boldsymbol{\theta})$ is not differentiable in $\boldsymbol{\theta}$, i.e. in particular not in ξ. Therefore, one could try to consider the (possibly differentiable) limit of $\log L(\boldsymbol{\theta})$ as $n \to \infty$. Unfortunately, $\log L(\boldsymbol{\theta})$ does not converge to a finite limit as $n \to \infty$, instead we can contemplate the process

$$\begin{aligned} X_n(\boldsymbol{\theta}) &= \frac{1}{n}\left(\log L(\boldsymbol{\theta}) + (\log n)\sum_{i=1}^{n} N_i(\tau)\right) \qquad (4.1) \\ &= \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}_i(t;\xi)\, dN_i(t) - \int_0^\tau \log\{\frac{1}{n}\sum_{i=1}^{n} R_i(t)\exp(\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}_i(t;\xi))\}\, \mathrm{d}\bar{N}(t), \end{aligned}$$

where $\bar{N}(t) = \frac{1}{n}\sum_{i=1}^{n} N_i(t)$. Obviously, the estimate $\hat{\boldsymbol{\theta}}_n$ not only maximizes $\log L(\boldsymbol{\theta})$ but also $X_n(\boldsymbol{\theta})$. We can show that the limit of $X_n(\boldsymbol{\theta})$ as $n \to \infty$ is given by

$$x(\boldsymbol{\theta}) \;=\; E\left[\int_0^\tau \left(\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}(t;\xi) - \log\left(s(t;\boldsymbol{\theta})\right)\right)\lambda(t,\boldsymbol{\theta}_0)\,\mathrm{d}t\right], \qquad (4.2)$$

where $s(t;\boldsymbol{\theta}) = E[R(t)\exp(\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}(t;\xi))]$.
The cumulative hazard function $\Lambda_0(t) = \int_0^t \lambda_0(u)\,\mathrm{d}u$ is estimated by the Breslow estimator

$$\hat{\Lambda}_n(t) = \int_0^t \frac{\mathrm{d}\left(n\bar{N}(u)\right)}{S(u,\hat{\boldsymbol{\theta}}_n)},$$

where $S(u,\boldsymbol{\theta}) = \sum_{i=1}^{n} R_i(u)\exp\{\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}_i(u,\xi)\}$.

4.3 Conditions

The following conditions are needed to establish the asymptotic properties of the estimates. To ease up notation it is convenient to define the probability measure P^t with

$$\mathrm{d}\,\mathrm{P}^t = q_t^{-1}\,\mathrm{d}\,\mathrm{Q}^t, \quad \mathrm{Q}^t(A) = \int_A R(t)\exp(\boldsymbol{\beta}_0^\top \tilde{\boldsymbol{Z}}(t;\xi_0))\,\mathrm{d}\,\mathrm{P} \text{ and } q_t = \int \mathrm{d}\,\mathrm{Q}^t,$$

provided that $q_t < \infty$.
The conditions are based on the existence of a convex and compact set $\Theta = [\xi_1,\xi_2]\times\mathcal{B} \subset \mathbb{R}^{p+3}$ with $\boldsymbol{\theta}_0$ in its interior.

Conditions.

A.1 [Finite baseline intensity] $\sup_{t\in[0,\tau]} \lambda_0(t) < \infty$.

A.2 The random variable Z_2 has an absolutely continuous distribution with density f_{Z_2} which is strictly positive, bounded and continuous in a neighborhood of ξ_0. Moreover, $EZ_2 < \infty$.

A.3 For k=0,1,2,

$$E \sup_{t\in[0,\tau]} \sup_{\boldsymbol{\theta}\in\Theta} \left\{ \left(\|\boldsymbol{Z}_1(t)\|^k + |Z_2|^k \right) \exp(\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}(t;\xi)) \right\}^2 < \infty.$$

A.4 [Asymptotic regularity conditions] *The function* $s(t;\boldsymbol{\theta}) = E[R(t)\exp(\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}(t;\xi))]$ *is bounded away from zero on* $[0,\tau]\times\Theta$ *and the first two partial derivatives of* $s(t;\boldsymbol{\theta})$ *with respect to* $\boldsymbol{\beta}$ *are continuous on* Θ*, uniformly in* $t\in[0,\tau]$*.*

A.5 a)For all $t\in[0,\tau]$ *there exists a neighborhood* $V(\boldsymbol{\theta}_0)$ *of* $\boldsymbol{\theta}_0$ *such that the covariance matrix* $\mathrm{Cov}_{\mathrm{P}^t}(\boldsymbol{Y}(t))$*, where* $\boldsymbol{Y}(t) = \left(-\beta_{30} I_{\{Z_2>\xi_0\}}, \boldsymbol{Z}_1^\top(t), Z_2, (Z_2-\xi_0)^+\right)^\top$ *is positive definite.*
b)For k=0,1,2, j=1,2,

$$\sup_{z\in[\xi_1,\xi_2]} E\left[\sup_{t\in[0,\tau]} \sup_{\boldsymbol{\theta}\in\Theta} \left\{ \left(\|\boldsymbol{Z}_1(t)\|^k + |Z_2|^k \right) \exp(\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}(t;\xi)) \right\}^j |Z_2 = z \right] < \infty$$

and

$$\sup_{z,z'} \sup_{t\in[0,\tau]} \sup_{\boldsymbol{\theta}\in\Theta} \left| E\left\{ \exp(\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}(t;\xi)) | Z_2 = z \right\} - E\left\{ \exp(\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}(t;\xi)) | Z_2 = z' \right\} \right| \overset{|z-z'|\to 0}{\longrightarrow} 0,$$

where z *and* z' *vary in* $[\xi_1,\xi_2]$*.*

Condition A.5 is used for interchanging integrability and differentiability. Note that conditions A.3 is a consequence of A.5. But this more stringent condition is not needed for the first proofs.

4.4 Consistency of the Estimates

In this subsection we establish the consistency of $\hat{\boldsymbol{\theta}}_n$. The proof is based on the uniform convergence of X_n to x, see (4.1) and (4.2), and on properties of x in a neighborhood of $\boldsymbol{\theta}_0$.

Lemma 4.1. *Under conditions A.1-A.4,* $\sup_{\boldsymbol{\theta}\in\Theta} |X_n(\boldsymbol{\theta}) - x(\boldsymbol{\theta})|$ *converges in probability to zero as* $n\to\infty$*.*

Proof. $X_n(\boldsymbol{\theta})$ can be rewritten in the following way:

$$\begin{aligned} X_n(\boldsymbol{\theta}) =& (\boldsymbol{\beta}_1^\top, \beta_2)^\top \frac{1}{n} \sum_{i=1}^n \int_0^\tau \begin{pmatrix} \boldsymbol{Z}_{1i}(t) \\ Z_{2i} \end{pmatrix} \, \mathrm{d}M_i(t) \\ &+ (\boldsymbol{\beta}_1^\top, \beta_2)^\top \frac{1}{n} \sum_{i=1}^n \int_0^\tau \begin{pmatrix} \boldsymbol{Z}_{1i}(t) \\ Z_{2i} \end{pmatrix} R_i(t) \exp(\boldsymbol{\beta}_0^\top \tilde{\boldsymbol{Z}}(t;\xi_0)) \, \mathrm{d}\Lambda_0(t) \\ &+ \beta_3 \frac{1}{n} \sum_{i=1}^n (Z_{2i} - \xi)^+ N_i(\tau) \\ &- \int_0^\tau \log \left(\frac{1}{n} \sum_{i=1}^n R_i(t) \exp(\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}_i(t;\xi)) \right) \mathrm{d}\bar{N}(t). \end{aligned} \tag{4.3}$$

Due to A.2 and A.3 and since Θ is compact the first term has mean zero and hence converges in probability to zero by the law of large numbers. Similarly, the second term converges to

$$(\boldsymbol{\beta}_1^\top, \beta_2)^\top E \left[\int_0^\tau R(t) \begin{pmatrix} \boldsymbol{Z}_1(t) \\ Z_2 \end{pmatrix} \exp(\boldsymbol{\beta}_0^\top \tilde{\boldsymbol{Z}}(t;\xi_0)) \, \mathrm{d}\Lambda_0(t) \right].$$

The third term in (4.3) can be handled as follows. By condition A.3, for all $\xi \in [\xi_1, \xi_2]$,

$$\begin{aligned} E\left[(Z_2 - \xi)^+ \int_0^\tau \lambda(t, \boldsymbol{\theta}_0) \, \mathrm{d}t \right] &\le E\left[(Z_2 - \xi_1)^+ \int_0^\tau \lambda(t, \boldsymbol{\theta}_0) \, \mathrm{d}t \right] \\ &\le E \int_0^\tau |Z_2| \lambda(t, \boldsymbol{\theta}_0) \, \mathrm{d}t + |\xi_1| E \int_0^\tau \lambda(t, \boldsymbol{\theta}_0) \, \mathrm{d}t < \infty. \end{aligned}$$

Hence by condition A.2,

$$\int_0^t (Z_2 - \xi)^+ \, \mathrm{d}M(s)$$

is a martingale and

$$E\left[(Z_2 - \xi)^+ N(\tau) \right] = E\left[\int_0^\tau (Z_2 - \xi)^+ dN(t) \right] = E\left[\int_0^\tau (Z_2 - \xi)^+ \lambda(t, \boldsymbol{\theta}_0) \, \mathrm{d}t \right] < \infty.$$

We want to apply the Glivenko-Cantelli theorem given in Theorem 19.4 and Example 19.8 in Van der Vaart (1998), (see also Appendix Theorem A.5). Clearly, $(Z_2 - \xi_1)^+ N(\tau)$ is an envelope function for $(Z_2 - \xi)^+ N(\tau)$. Since $(Z_2 - \xi)^+ N(\tau)$ is continuous in ξ, we get

$$\sup_{\xi \in [\xi_1, \xi_2]} \left| \frac{1}{n} \sum_{i=1}^n (Z_{2i} - \xi)^+ N_i(\tau) - E\left[\int_0^\tau (Z_2 - \xi)^+ \lambda(t, \boldsymbol{\theta}_0) \, \mathrm{d}t \right] \right| \xrightarrow{\mathrm{P}} 0.$$

Multiplying by the bounded parameter β_3 gives the convergence of the third term in (4.3).

To show uniform stochastic convergence of the fourth term in (4.3) one can argue as follows: By the strong law of large numbers given by Andersen & Gill (1982) (see also Theorem A.1),

$$\sup_{\boldsymbol{\theta}\in\Theta}\sup_{t\in[0,\tau]}\left|\frac{1}{n}\sum_{i=1}^{n}R_i(t)\exp(\boldsymbol{\beta}^\top\tilde{\boldsymbol{Z}}_i(t,\xi))-s(t,\boldsymbol{\theta})\right|\xrightarrow{\mathrm{P}}0,$$

where we used the integrability condition A.3. Since $s(t,\boldsymbol{\theta})$ is bounded away from 0 by condition A.4, it follows immediately that

$$\sup_{\boldsymbol{\theta}\in\Theta}\sup_{t\in[0,\tau]}\left|\log\left(\frac{1}{n}\sum_{i=1}^{n}R_i(t)\exp(\boldsymbol{\beta}^\top\tilde{\boldsymbol{Z}}_i(t,\xi))\right)-\log(s(t,\boldsymbol{\theta}))\right|\xrightarrow{\mathrm{P}}0.$$

Since

$$\frac{1}{n}\sum_{i=1}^{n}N_i(\tau)\xrightarrow{\mathrm{P}}EN(\tau)=E\left[\int_0^\tau\lambda(t,\boldsymbol{\theta}_0)\,\mathrm{d}t\right]<\infty,$$

the difference between

$$\int_0^\tau\log(s(t,\boldsymbol{\theta}))\,\mathrm{d}\bar{N}(t)$$

and the fourth term in (4.3) converges uniformly to 0 in probability. Using the Glivenko-Cantelli theorem and Example 19.8 in Van der Vaart (1998) as before, we get

$$\sup_{\boldsymbol{\theta}\in\Theta}\sup_{t\in[0,\tau]}\left|\int_0^\tau\log(s(t,\boldsymbol{\theta}))\,\mathrm{d}\bar{N}(t)-E\left[\int_0^\tau\log(s(t,\boldsymbol{\theta}))\lambda(t,\boldsymbol{\theta}_0)\,\mathrm{d}t\right]\right|\xrightarrow{\mathrm{P}}0,$$

where the envelope function

$$\sup_{\boldsymbol{\theta}\in\Theta}\sup_{t\in[0,\tau]}\log(s(t,\boldsymbol{\theta}))N(\tau)$$

is bounded by A.3 and A.4. □

In order to prove the next theorem it is beneficial to show concavity of $x(\boldsymbol{\theta})$ in a neighborhood of $\boldsymbol{\theta}_0$. For this we need to consider the score function $U(\boldsymbol{\theta})=\frac{\partial}{\partial\boldsymbol{\theta}}x(\boldsymbol{\theta})$ and the Hessian matrix of $x(\boldsymbol{\theta})$ at $\boldsymbol{\theta}_0$.

Lemma 4.2. *Under conditions A.1-A.5 the score function $U(\boldsymbol{\theta}_0)=0$ and the Hessian matrix $\boldsymbol{H}(\boldsymbol{\theta}_0)$ of $x(\boldsymbol{\theta}_0)$ is given by*

$$\boldsymbol{H}(\boldsymbol{\theta}_0)=-\int_0^\tau q_s\,\mathrm{Cov}_{\mathrm{P}^s}(\boldsymbol{Y}(t))\lambda_0(t)\,\mathrm{d}t,$$

where $\boldsymbol{Y}(t)=\left(-\beta_{30}I_{\{Z_2>\xi_0\}},\boldsymbol{Z}_1^\top(t),Z_2,(Z_2-\xi_0)^+\right)^\top$.

Moreover, the matrix $\boldsymbol{H}(\boldsymbol{\theta}_0)$ *is negative definite.*

Proof. To compose the score function and the Hessian matrix of the function $x(\boldsymbol{\theta}_0)$ we need to calculate several partial derivatives. First of all we consider the derivatives of $s(t;\boldsymbol{\theta}) = E[R(t)\exp(\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}(t,\xi))]$. Note that the density of f_{Z_2} of the distribution of Z_2 exists by condition A.2. Furthermore, let $\frac{\partial}{\partial \boldsymbol{\beta}_1} = \left(\frac{\partial}{\partial \beta_{11}}, \cdots, \frac{\partial}{\partial \beta_{1p}}\right)^\top$ as described in Section 2.2 .

$$
\begin{aligned}
\frac{\partial}{\partial \xi} s(t;\boldsymbol{\theta}) &= E[R(t)(-\beta_3 I_{\{Z_2>\xi\}})\exp(\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}(t,\xi))] \\
\frac{\partial}{\partial \boldsymbol{\beta}_1} s(t;\boldsymbol{\theta}) &= E[R(t)\boldsymbol{Z}_1(t)\exp(\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}(t,\xi))] \\
\frac{\partial}{\partial \beta_2} s(t;\boldsymbol{\theta}) &= E[R(t)Z_2\exp(\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}(t,\xi))] \\
\frac{\partial}{\partial \beta_3} s(t;\boldsymbol{\theta}) &= E[R(t)(Z_2-\xi)I_{\{Z_2>\xi\}}\exp(\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}(t,\xi))] \\
\frac{\partial^2}{(\partial \xi)^2} s(t;\boldsymbol{\theta}) &= E[R(t)(-\beta_3 I_{\{Z_2>\xi\}})^2\exp(\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}(t,\xi))] \\
&\quad -E[R(t)(-\beta_3)\exp(\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}(t,\xi))|Z_2=\xi]f_{Z_2}(\xi) \\
\frac{\partial^2}{(\partial \boldsymbol{\beta}_1)^2} s(t;\boldsymbol{\theta}) &= E[R(t)\boldsymbol{Z}_1(t)\boldsymbol{Z}_1^\top(t)\exp(\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}(t,\xi))] \\
\frac{\partial^2}{(\partial \beta_2)^2} s(t;\boldsymbol{\theta}) &= E[R(t)(Z_2)^2\exp(\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}(t,\xi))] \\
\frac{\partial^2}{(\partial \beta_3)^2} s(t;\boldsymbol{\theta}) &= E[R(t)(Z_2-\xi)^2 I_{\{Z_2>\xi\}}\exp(\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}(t,\xi))] \\
\frac{\partial}{\partial \boldsymbol{\beta}_1}\frac{\partial}{\partial \xi} s(t;\boldsymbol{\theta}) &= E[R(t)(-\beta_3 I_{\{Z_2>\xi\}})\boldsymbol{Z}_1(t)\exp(\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}(t,\xi))] \\
\frac{\partial}{\partial \beta_2}\frac{\partial}{\partial \xi} s(t;\boldsymbol{\theta}) &= E[R(t)(-\beta_3 I_{\{Z_2>\xi\}})Z_2\exp(\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}(t,\xi))] \\
\frac{\partial}{\partial \beta_3}\frac{\partial}{\partial \xi} s(t;\boldsymbol{\theta}) &= E[R(t)(-\beta_3 I_{\{Z_2>\xi\}})(Z_2-\xi)\exp(\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}(t,\xi))] \\
&\quad +E[R(t)(-1)I_{\{Z_2>\xi\}}\exp(\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}(t,\xi))].
\end{aligned}
$$

The first partial derivatives of $x(\boldsymbol{\theta})$ are given by

$$
\begin{aligned}
\frac{\partial}{\partial \xi} x(\boldsymbol{\theta}) &= \int_0^\tau \left[E\left[R(t)(-\beta_3 I_{\{Z_2>\xi\}})\exp(\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}(t,\xi))\right]\right. \\
&\quad \left. -\frac{s(t;\boldsymbol{\theta}_0)}{s(t;\boldsymbol{\theta})}\frac{\partial}{\partial \xi}s(t;\boldsymbol{\theta})\right]\lambda_0(t)\,\mathrm{d}t \\
\frac{\partial}{\partial \boldsymbol{\beta}_1} x(\boldsymbol{\theta}) &= \int_0^\tau \left[E\left[R(t)\boldsymbol{Z}_1(t)\exp(\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}(t,\xi))\right] - \frac{s(t;\boldsymbol{\theta}_0)}{s(t;\boldsymbol{\theta})}\frac{\partial}{\partial \boldsymbol{\beta}_1}s(t;\boldsymbol{\theta})\right]\lambda_0(t)\,\mathrm{d}t
\end{aligned}
$$

$$\frac{\partial}{\partial\beta_2}x(\boldsymbol{\theta}) = \int_0^\tau \left[E\left[R(t)Z_2\exp(\boldsymbol{\beta}^\top\tilde{\boldsymbol{Z}}(t,\xi))\right] - \frac{s(t;\boldsymbol{\theta}_0)}{s(t;\boldsymbol{\theta})}\frac{\partial}{\partial\beta_2}s(t;\boldsymbol{\theta})\right]\lambda_0(t)\,\mathrm{d}t$$

$$\frac{\partial}{\partial\beta_3}x(\boldsymbol{\theta}) = \int_0^\tau \left[E\left[R(t)(Z_2-\xi)^+\exp(\boldsymbol{\beta}^\top\tilde{\boldsymbol{Z}}(t,\xi))\right] - \frac{s(t;\boldsymbol{\theta}_0)}{s(t;\boldsymbol{\theta})}\frac{\partial}{\partial\beta_3}s(t;\boldsymbol{\theta})\right]\lambda_0(t)\,\mathrm{d}t$$

Differentiation and integration can be interchanged because of conditions A.3 and A.5. It follows that $\frac{\partial}{\partial\xi}x(\boldsymbol{\theta}_0) = \frac{\partial}{\partial\boldsymbol{\beta}_1}x(\boldsymbol{\theta}_0) = \frac{\partial}{\partial\beta_2}x(\boldsymbol{\theta}_0) = \frac{\partial}{\partial\beta_3}x(\boldsymbol{\theta}_0) = 0$ and hence $U(\boldsymbol{\theta}_0) = \frac{\partial}{\partial\xi}x(\boldsymbol{\theta}_0) + \frac{\partial}{\partial\boldsymbol{\beta}_1}x(\boldsymbol{\theta}_0) + \frac{\partial}{\partial\beta_2}x(\boldsymbol{\theta}_0) + \frac{\partial}{\partial\beta_3}x(\boldsymbol{\theta}_0) = 0$.
To calculate the Hessian matrix we need the second derivatives of $x(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}_0$, which exist because of conditions A.2-A.5 and the Lebesgue differentiation theorem. We use the notation $\mathrm{Q}^t(A)$ and consequently, $s(t;\boldsymbol{\theta}_0) = \int \mathrm{d}\,\mathrm{Q}^t$. Thus,

$$\begin{aligned}
\frac{\partial}{\partial\xi}s(t;\boldsymbol{\theta}_0) &= \int(-\beta_{30}I_{\{Z_2>\xi_0\}})\,\mathrm{d}\,\mathrm{Q}^t\\
\frac{\partial}{\partial\boldsymbol{\beta}_1}s(t;\boldsymbol{\theta}_0) &= \int \boldsymbol{Z}_1(t)\,\mathrm{d}\,\mathrm{Q}^t\\
\frac{\partial}{\partial\beta_2}s(t;\boldsymbol{\theta}_0) &= \int Z_2\,\mathrm{d}\,\mathrm{Q}^t\\
\frac{\partial}{\partial\beta_3}s(t;\boldsymbol{\theta}_0) &= \int (Z_2-\xi_0)I_{\{Z_2>\xi_0\}}\,\mathrm{d}\,\mathrm{Q}^t\\
\frac{\partial^2}{(\partial\xi)^2}s(t;\boldsymbol{\theta}_0) &= \int(-\beta_{30}I_{\{Z_2>\xi_0\}})^2\,\mathrm{d}\,\mathrm{Q}^t - \int(-\beta_{30}I_{\{Z_2>\xi_0\}})\,\mathrm{d}\,\mathrm{Q}^t\\
\frac{\partial^2}{(\partial\boldsymbol{\beta}_1)^2}s(t;\boldsymbol{\theta}_0) &= \int \boldsymbol{Z}_1(t)\boldsymbol{Z}_1^\top(t)\,\mathrm{d}\,\mathrm{Q}^t\\
\frac{\partial^2}{(\partial\beta_2)^2}s(t;\boldsymbol{\theta}_0) &= \int (Z_2)^2\,\mathrm{d}\,\mathrm{Q}^t\\
\frac{\partial^2}{(\partial\beta_3)^2}s(t;\boldsymbol{\theta}_0) &= \int (Z_2-\xi_0)^2 I_{\{Z_2>\xi_0\}}\,\mathrm{d}\,\mathrm{Q}^t\\
\frac{\partial}{\partial\boldsymbol{\beta}_1}\frac{\partial}{\partial\xi}s(t;\boldsymbol{\theta}_0) &= \int(-\beta_{30}I_{\{Z_2>\xi_0\}})\boldsymbol{Z}_1(t)\,\mathrm{d}\,\mathrm{Q}^t\\
\frac{\partial}{\partial\beta_2}\frac{\partial}{\partial\xi}s(t;\boldsymbol{\theta}_0) &= \int(-\beta_{30}I_{\{Z_2>\xi_0\}})Z_2\,\mathrm{d}\,\mathrm{Q}^t\\
\frac{\partial}{\partial\beta_3}\frac{\partial}{\partial\xi}s(t;\boldsymbol{\theta}_0) &= \int(-\beta_{30}I_{\{Z_2>\xi_0\}})(Z_2-\xi_0)\,\mathrm{d}\,\mathrm{Q}^t + \int I_{\{Z_2>\xi_0\}}\,\mathrm{d}\,\mathrm{Q}^t\,.
\end{aligned}$$

Therefore, it follows that

$$
\begin{aligned}
\frac{\partial^2}{(\partial\xi)^2}x(\boldsymbol{\theta}_0) &= \int_0^\tau \left[\frac{1}{\int \mathrm{d}\,\mathrm{Q}^s}\left(\int(-\beta_{30}I_{\{Z_2>\xi_0\}})\,\mathrm{d}\,\mathrm{Q}^s\right)^2 \right.\\
&\quad \left. - \int(-\beta_{30}I_{\{Z_2>\xi_0\}})^2\,\mathrm{d}\,\mathrm{Q}^s\right]\lambda_0(s)\,\mathrm{d}s \\
\frac{\partial^2}{(\partial\boldsymbol{\beta}_1)^2}x(\boldsymbol{\theta}_0) &= \int_0^\tau \left[\frac{1}{\int \mathrm{d}\,\mathrm{Q}^s}\left(\int \boldsymbol{Z}_1(s)\,\mathrm{d}\,\mathrm{Q}^s\right)\left(\int \boldsymbol{Z}_1(s)\,\mathrm{d}\,\mathrm{Q}^s\right)^\top \right.\\
&\quad \left. - \int \boldsymbol{Z}_1(s)\boldsymbol{Z}_1^\top(s)\,\mathrm{d}\,\mathrm{Q}^s\right]\lambda_0(s)\,\mathrm{d}s \\
\frac{\partial^2}{(\partial\beta_2)^2}x(\boldsymbol{\theta}_0) &= \int_0^\tau \left[\frac{1}{\int \mathrm{d}\,\mathrm{Q}^s}\left(\int Z_2\,\mathrm{d}\,\mathrm{Q}^s\right)^2 - \int(Z_2)^2\,\mathrm{d}\,\mathrm{Q}^s\right]\lambda_0(s)\,\mathrm{d}s \\
\frac{\partial^2}{(\partial\beta_3)^2}x(\boldsymbol{\theta}_0) &= \int_0^\tau \left[\frac{1}{\int \mathrm{d}\,\mathrm{Q}^s}\left(\int(Z_2-\xi_0)^+\,\mathrm{d}\,\mathrm{Q}^s\right)^2 \right.\\
&\quad \left. - \int((Z_2-\xi_0)I_{\{Z_2>\xi_0\}})^2\,\mathrm{d}\,\mathrm{Q}^s\right]\lambda_0(s)\,\mathrm{d}s \\
\frac{\partial}{\partial\boldsymbol{\beta}_1}\frac{\partial}{\partial\xi}x(\boldsymbol{\theta}_0) &= \int_0^\tau \left[\frac{1}{\int \mathrm{d}\,\mathrm{Q}^s}\int \boldsymbol{Z}_1(s)\,\mathrm{d}\,\mathrm{Q}^s\int(-\beta_{30}I_{\{Z_2>\xi_0\}})\,\mathrm{d}\,\mathrm{Q}^s\right]\lambda_0(s)\,\mathrm{d}s \\
&\quad - \int_0^\tau\left[\int \boldsymbol{Z}_1(s)(-\beta_{30}I_{\{Z_2>\xi_0\}})\,\mathrm{d}\,\mathrm{Q}^s\right]\lambda_0(s)\,\mathrm{d}s \\
\frac{\partial}{\partial\beta_2}\frac{\partial}{\partial\xi}x(\boldsymbol{\theta}_0) &= \int_0^\tau \left[\frac{1}{\int \mathrm{d}\,\mathrm{Q}^s}\int Z_2\,\mathrm{d}\,\mathrm{Q}^s\int(-\beta_{30}I_{\{Z_2>\xi_0\}})\,\mathrm{d}\,\mathrm{Q}^s\right]\lambda_0(s)\,\mathrm{d}s \\
&\quad - \int_0^\tau\left[\int Z_2(-\beta_{30}I_{\{Z_2>\xi_0\}})\,\mathrm{d}\,\mathrm{Q}^s\right]\lambda_0(s)\,\mathrm{d}s \\
\frac{\partial}{\partial\beta_3}\frac{\partial}{\partial\xi}x(\boldsymbol{\theta}_0) &= -\int_0^\tau\left[(-1)I_{\{Z_2>\xi_0\}}\,\mathrm{d}\,\mathrm{Q}^s\right]\lambda_0(s)\,\mathrm{d}s \\
&\quad + \int_0^\tau \left[\frac{1}{\int \mathrm{d}\,\mathrm{Q}^s}\int(Z_2-\xi_0)^+\,\mathrm{d}\,\mathrm{Q}^s\int(-\beta_{30}I_{\{Z_2>\xi_0\}})\,\mathrm{d}\,\mathrm{Q}^s\right]\lambda_0(s)\,\mathrm{d}s \\
&\quad - \int_0^\tau\left[\int(Z_2-\xi_0)^+(-\beta_{30})\,\mathrm{d}\,\mathrm{Q}^s\right]\lambda_0(s)\,\mathrm{d}s.
\end{aligned}
$$

The Hessian matrix $\boldsymbol{H}$ of $x(\boldsymbol{\theta}_0)$ is given by

$$
\boldsymbol{H} = \begin{pmatrix}
\left(\frac{\partial^2}{\partial^2\xi^2}\right)x(\boldsymbol{\theta}_0) & \left(\frac{\partial^2}{\partial\xi\partial\boldsymbol{\beta}_1}\right)x(\boldsymbol{\theta}_0) & \left(\frac{\partial^2}{\partial\xi\partial\beta_2}\right)x(\boldsymbol{\theta}_0) & \left(\frac{\partial^2}{\partial\xi\partial\beta_3}\right)x(\boldsymbol{\theta}_0) \\
\left(\frac{\partial^2}{\partial\boldsymbol{\beta}_1\partial\xi}\right)x(\boldsymbol{\theta}_0) & \left(\frac{\partial^2}{\partial^2\boldsymbol{\beta}_1^2}\right)x(\boldsymbol{\theta}_0) & \left(\frac{\partial^2}{\partial\boldsymbol{\beta}_1\partial\beta_2}\right)x(\boldsymbol{\theta}_0) & \left(\frac{\partial^2}{\partial\boldsymbol{\beta}_1\partial\beta_3}\right)x(\boldsymbol{\theta}_0) \\
\left(\frac{\partial^2}{\partial\beta_2\partial\xi}\right)x(\boldsymbol{\theta}_0) & \left(\frac{\partial^2}{\partial\beta_2\partial\boldsymbol{\beta}_1}\right)x(\boldsymbol{\theta}_0) & \left(\frac{\partial^2}{\partial^2\beta_2^2}\right)x(\boldsymbol{\theta}_0) & \left(\frac{\partial^2}{\partial\beta_2\partial\beta_3}\right)x(\boldsymbol{\theta}_0) \\
\left(\frac{\partial^2}{\partial\beta_3\partial\xi}\right)x(\boldsymbol{\theta}_0) & \left(\frac{\partial^2}{\partial\beta_3\partial\boldsymbol{\beta}_1}\right)x(\boldsymbol{\theta}_0) & \left(\frac{\partial^2}{\partial\beta_3\partial\beta_2}\right)x(\boldsymbol{\theta}_0) & \left(\frac{\partial^2}{\partial^2\beta_3^2}\right)x(\boldsymbol{\theta}_0)
\end{pmatrix}.
$$

It remains to show that $\boldsymbol{H}$ is negative definite. Therefore, we consider the following

$$\boldsymbol{H}(\boldsymbol{\theta}_0) = \int_0^\tau (\boldsymbol{H}_1(\boldsymbol{\theta}_0) - \boldsymbol{H}_2(\boldsymbol{\theta}_0))\lambda_0(t)\,\mathrm{d}t$$

with the symmetric matrices

$$\boldsymbol{H}_1(\boldsymbol{\theta}_0) = \frac{1}{\int \mathrm{d}\,\mathrm{Q}^s} \int \boldsymbol{Y}\,\mathrm{d}\,\mathrm{Q}^s \left(\int \boldsymbol{Y}\,\mathrm{d}\,\mathrm{Q}^s\right)^\top \quad \text{and } \boldsymbol{H}_2(\boldsymbol{\theta}_0) = \int \boldsymbol{Y}\boldsymbol{Y}^\top\,\mathrm{d}\,\mathrm{Q}^s,$$

where $\boldsymbol{Y} = \left(-\beta_{30} I_{\{Z_2>\xi_0\}}, \boldsymbol{Z}_1^\top(s), Z_2, (Z_2-\xi_0)^+\right)^\top$. Using the notation $q = \int \mathrm{d}\,\mathrm{Q}^s$ and $\mathrm{d}\,\mathrm{P}^s = q^{-1}\,\mathrm{d}\,\mathrm{Q}^s$ we get

$$\begin{aligned}
-(\boldsymbol{H}_1 - \boldsymbol{H}_2) &= \int \boldsymbol{Y}\boldsymbol{Y}^\top\,\mathrm{d}\,\mathrm{Q}^s - q^{-1}\int \boldsymbol{Y}\,\mathrm{d}\,\mathrm{Q}^s \left(\int \boldsymbol{Y}\,\mathrm{d}\,\mathrm{Q}^s\right)^\top \\
&= \int \boldsymbol{Y}\boldsymbol{Y}^\top\,\mathrm{d}\,\mathrm{Q}^s - 2q^{-1}\int \boldsymbol{Y}\,\mathrm{d}\,\mathrm{Q}^s \left(\int \boldsymbol{Y}\,\mathrm{d}\,\mathrm{Q}^s\right)^\top \\
&\quad + q\left(q^{-1}\int \boldsymbol{Y}\,\mathrm{d}\,\mathrm{Q}^s\right)\left(q^{-1}\int \boldsymbol{Y}\,\mathrm{d}\,\mathrm{Q}^s\right)^\top \\
&= \int \left(\boldsymbol{Y} - q^{-1}\int \boldsymbol{Y}\,\mathrm{d}\,\mathrm{Q}^s\right)\left(\boldsymbol{Y} - q^{-1}\int \boldsymbol{Y}\,\mathrm{d}\,\mathrm{Q}^s\right)^\top \mathrm{d}\,\mathrm{Q}^s \\
&= q\,\mathrm{Cov}_{\mathrm{P}^s}(\boldsymbol{Y}(s)).
\end{aligned}$$

Hence, $\boldsymbol{H}$ is negative semidefinite. By means of condition A.5 this result can be strengthened to ensure that $\boldsymbol{H}(\boldsymbol{\theta}_0)$ is negative definite. □

Theorem 4.2. *Under conditions A.1-A.5 there exists a neighborhood $V(\boldsymbol{\theta}_0)$ of $\boldsymbol{\theta}_0$ such that if $\hat{\boldsymbol{\theta}}_n$ lies in $V(\boldsymbol{\theta}_0)$, it follows that $\hat{\boldsymbol{\theta}}_n \xrightarrow{\mathrm{P}} \boldsymbol{\theta}_0$ as $n \to \infty$.*

Proof. By Lemma 4.1 we know that X_n converges uniformly to x. Furthermore, Lemma 4.2 yields that $x(\boldsymbol{\theta})$ is strictly concave in a neighborhood $V(\boldsymbol{\theta}_0) \subset \Theta$. Together with $U(\boldsymbol{\theta}_0) = 0$ this gives the unique maximum of x at $\boldsymbol{\theta}_0$. Now, the assertion follows. □

4.5 Rate of Convergence

Usually, when a change-point model with a jump is considered the rate of convergence of the change-point estimator is n. In our case it turns out that the rate of convergence of the change-point estimator is not better than $\sqrt{n}$. The difference between a jump and a bent-line change-point is the continuity of the process $X_n(\boldsymbol{\theta})$ (cf. (4.1)) in ξ in the bent-line model. The continuity causes the limit of $X_n(\boldsymbol{\theta})$ to be differentiable in ξ.

Let $V_\epsilon(\boldsymbol{\theta}_0) = \{\boldsymbol{\theta} \in \Theta : \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| < \epsilon\}$ be a neighborhood of $\boldsymbol{\theta}_0$ and let W_n be the process

$$W_n(\boldsymbol{\theta}) = \sqrt{n}(X_n(\boldsymbol{\theta}) - x(\boldsymbol{\theta})).$$

The following lemmas are needed to establish the rate of convergence.

Lemma 4.3. *Under conditions A.1 and A.5, for ϵ sufficiently small there exists a constant $\alpha > 0$ such that for all $\boldsymbol{\theta}$ in $V_\epsilon(\boldsymbol{\theta}_0)$, $x(\boldsymbol{\theta}) - x(\boldsymbol{\theta}_0) \leq -\alpha\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|^2$.*

Proof. For $x(\boldsymbol{\theta}) = E\left[\int_0^\tau \left\{\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}(t,\xi) - \log(s(t,\boldsymbol{\theta}))\right\} \lambda(t,\boldsymbol{\theta}_0)\,\mathrm{d}t\right]$ we know that $\frac{\partial}{\partial \xi} x(\boldsymbol{\theta}_0) = \mathbf{0}$ and $\frac{\partial}{\partial \boldsymbol{\beta}} x(\boldsymbol{\theta}_0) = \mathbf{0}$. Hence, by a Taylor expansion of $x(\boldsymbol{\theta})$ for ϵ sufficiently small and for $\boldsymbol{\theta}$ in $V_\epsilon(\boldsymbol{\theta}_0)$,

$$\begin{aligned} x(\boldsymbol{\theta}) - x(\boldsymbol{\theta}_0) &= \frac{\partial}{\partial \xi} x(\boldsymbol{\theta}_0)(\xi - \xi_0) + \frac{\partial}{\partial \boldsymbol{\beta}} x(\boldsymbol{\theta}_0)(\boldsymbol{\beta} - \boldsymbol{\beta}_0) \\ &\quad + \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)^\top \boldsymbol{H}(\boldsymbol{\theta}_0)(\boldsymbol{\theta} - \boldsymbol{\theta}_0) + o(\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|^2) \\ &\leq -\alpha\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|^2, \end{aligned}$$

since $\boldsymbol{H}(\boldsymbol{\theta}_0)$ is negative definite. □

Lemma 4.4. *Under conditions A.1-A.5, for every $\epsilon > 0$ there exists a constant $\kappa > 0$ such that $E[\sup_{\boldsymbol{\theta} \in V_\epsilon(\boldsymbol{\theta}_0)} |W_n(\boldsymbol{\theta}) - W_n(\boldsymbol{\theta}_0)|] \leq \kappa\epsilon$, for all $n \in \mathbb{N}$.*

Proof. Let $\overline{\boldsymbol{\beta}} = (\boldsymbol{\beta}_1^\top, \boldsymbol{\beta}_2^\top)^\top$ and $\bar{S}(t;\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^n R_i(t)\exp(\boldsymbol{\beta}^\top \tilde{\mathbf{Z}}_i(t))$. Rewrite $W_n(\boldsymbol{\theta}) - W_n(\boldsymbol{\theta}_0) = W_{1n}(\boldsymbol{\theta}) - W_{2n}(\boldsymbol{\theta})$, where

$$\begin{aligned} W_{1n}(\boldsymbol{\theta}) &= n^{-1/2}(\overline{\boldsymbol{\beta}} - \overline{\boldsymbol{\beta}}_0)^\top \sum_{i=1}^n \left[\int_0^\tau \begin{pmatrix} \boldsymbol{Z}_{1i}(t) \\ Z_{2i} \end{pmatrix} \mathrm{d}N_i(t) - \int_0^\tau E\begin{pmatrix} \boldsymbol{Z}_1(t) \\ Z_2 \end{pmatrix} \lambda(t,\boldsymbol{\theta}_0)\,\mathrm{d}t\right] \\ &\quad + n^{-1/2}\beta_3 \sum_{i=1}^n \left[\int_0^\tau (Z_{2i} - \xi)^+\,\mathrm{d}N_i(t) - \int_0^\tau E(Z_2 - \xi)^+ \lambda(t,\boldsymbol{\theta}_0)\,\mathrm{d}t\right] \\ &\quad - n^{-1/2}\beta_{30} \sum_{i=1}^n \left[\int_0^\tau (Z_{2i} - \xi_0)^+\,\mathrm{d}N_i(t) - \int_0^\tau E(Z_2 - \xi_0)^+ \lambda(t,\boldsymbol{\theta}_0)\,\mathrm{d}t\right] \\ &= n^{-1/2}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)^\top \sum_{i=1}^n \left[\int_0^\tau \tilde{\mathbf{Z}}_i(t,\xi_0)\,\mathrm{d}N_i(t) - \int_0^\tau E[\tilde{\mathbf{Z}}(t,\xi_0)]\lambda(t,\boldsymbol{\theta}_0)\,\mathrm{d}t\right] \\ &\quad + n^{-1/2}\beta_3 \left[\sum_{i=1}^n \int_0^\tau (Z_{2i} - \xi)^+ - (Z_{2i} - \xi_0)^+\,\mathrm{d}N_i(t)\right. \\ &\quad \left. - \int_0^\tau [E(Z_2 - \xi)^+ - E(Z_2 - \xi_0)^+]\lambda(t,\boldsymbol{\theta}_0)\,\mathrm{d}t\right] \end{aligned}$$

and

$$\begin{aligned}
W_{2n}(\boldsymbol{\theta}) &= \sqrt{n}\left(\int_0^\tau \log(\bar{S}(t,\boldsymbol{\theta}))\,\mathrm{d}\bar{N}(t) - \int_0^\tau \log(s(t,\boldsymbol{\theta}))s(t,\boldsymbol{\theta}_0)\,\mathrm{d}\Lambda_0(t)\right.\\
&\quad \left. - \int_0^\tau \log(\bar{S}(t,\boldsymbol{\theta}_0))\,\mathrm{d}\bar{N}(t) + \int_0^\tau \log(s(t,\boldsymbol{\theta}_0))s(t,\boldsymbol{\theta}_0)\,\mathrm{d}\Lambda_0(t)\right)\\
&= n^{-1/2}\sum_{i=1}^n \left[\int_0^\tau \log\left(\frac{s(t,\boldsymbol{\theta})}{s(t,\boldsymbol{\theta}_0)}\right)\mathrm{d}N_i(t) - \int_0^\tau \log\left(\frac{s(t,\boldsymbol{\theta})}{s(t,\boldsymbol{\theta}_0)}\right)s(t,\boldsymbol{\theta}_0)\,\mathrm{d}\Lambda_0(t)\right]\\
&\quad + n^{-1/2}\sum_{i=1}^n \left[\int_0^\tau \log\left(\frac{\bar{S}(t,\boldsymbol{\theta})}{s(t,\boldsymbol{\theta})}\right) - \log\left(\frac{\bar{S}(t,\boldsymbol{\theta}_0)}{s(t,\boldsymbol{\theta}_0)}\right)\mathrm{d}N_i(t)\right]
\end{aligned}$$

Consider $W_{1n}(\boldsymbol{\theta})$. The expectation of the supremum of the absolute value of the first term is $O(\epsilon)$, since $\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\| < \epsilon$. In the second term we consider the sets of functions $\{f_\xi : \xi \in [\xi_1,\xi_2]\}$ and $\{g_\xi : \xi \in [\xi_1,\xi_2]\}$ with $f_\xi(a,b) = abI_{\{b>\xi\}}$ and $g_\xi(a,b) = a\xi I_{\{b>\xi\}}$. These sets form Vapnik-Cervonenkis classes. The function $\int_0^\tau (Z_{2i}-(\xi_0-\epsilon))^+\,\mathrm{d}N_i(t)$ is an envelope function for $\int_0^\tau (Z_{2i}-\xi)^+\,\mathrm{d}N_i(t)$ in $V_\epsilon(\boldsymbol{\theta}_0)$. Therefore, the $L_2(P)$ norm of the envelope function is bounded by

$$\begin{aligned}
&E\sup_{\xi\in V_\epsilon(\xi_0)}\left|\int_0^\tau [(Z_2-\xi)^+ - (Z_2-\xi_0)^+]\,\mathrm{d}N(t)\right|\\
&\qquad \le \left\{E\int_0^\tau |(Z_2-(\xi_0-\epsilon))^+ - (Z_2-\xi_0)^+|^2\,\mathrm{d}N(t)\right\}^{1/2} = O(\epsilon).
\end{aligned}$$

The boundedness of

$$\begin{aligned}
E\sup_{\xi\in V_\epsilon(\xi_0)}&\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n \int_0^\tau [(Z_{2i}-\xi)^+ - (Z_{2i}-\xi_0)^+]\,\mathrm{d}N_i(t)\right.\\
&\left. - \int_0^\tau [E(Z_2-\xi)^+ - E(Z_2-\xi_0)^+]\lambda(t,\boldsymbol{\theta}_0)\,\mathrm{d}t\right|
\end{aligned}$$

is a consequence of Theorem 2.14.1 of Van der Vaart & Wellner (1996), see Appendix Theorem A.7.

Now consider $W_{2n}(\boldsymbol{\theta})$. For the class of functions $\left\{\log\left(\frac{s(t,\boldsymbol{\theta})}{s(t,\boldsymbol{\theta}_0)}\right) : \boldsymbol{\theta}\in V_\epsilon(\boldsymbol{\theta}_0)\right\}$ it can be shown that it has an envelope function with $L_2(P)$ norm of order $O(\epsilon)$ and that its $L_2(P)$ bracketing integral is finite by Theorem 2.7.11 in Van der Vaart & Wellner (1996), see Appendix Theorem A.3. Hence, as a consequence of Theorem 2.14.2 of Van der Vaart & Wellner (1996) (Appendix Theorem A.8) the bound of

$$E\left[\sup_{\boldsymbol{\theta}\in V_\epsilon(\boldsymbol{\theta}_0)}\left|n^{-1/2}\sum_{i=1}^n \left[\int_0^\tau \log\left(\frac{s(t,\boldsymbol{\theta})}{s(t,\boldsymbol{\theta}_0)}\right)\mathrm{d}N_i(t) - \int_0^\tau \log\left(\frac{s(t,\boldsymbol{\theta})}{s(t,\boldsymbol{\theta}_0)}\right)s(t,\boldsymbol{\theta}_0)\,\mathrm{d}\Lambda_0(t)\right]\right|\right]$$

is of order $O(\epsilon)$. The second term can be treated as follows: Using a Taylor expansion of $\log\left(\frac{\bar{S}(t,\boldsymbol{\theta})}{s(t,\boldsymbol{\theta})}\right)$ at 1 yields

$$\log\left(\frac{\bar{S}(t,\boldsymbol{\theta})}{s(t,\boldsymbol{\theta})}\right) = \frac{\bar{S}(t,\boldsymbol{\theta})}{s(t,\boldsymbol{\theta})} - 1 + o_{a.s.}(1).$$

For the second logarithm we use a similar Taylor expansion, such that the second term of $W_{2n}(\boldsymbol{\theta})$ can be approximated by

$$n^{-3/2}\sum_{i,j}\int_0^\tau \left(\frac{R_j(t)\exp\{\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}_j(t,\xi)\}}{s(t,\boldsymbol{\theta})} - \frac{R_j(t)\exp\{\boldsymbol{\beta}_0^\top \tilde{\boldsymbol{Z}}_j(t,\xi_0)\}}{s(t,\boldsymbol{\theta}_0)}\right) \mathrm{d}N_i(t)\left(1 + o_{a.s.}(1)\right).$$

If $i = j$,

$$n^{-3/2}\sum_{i=1}^n E \sup_{\boldsymbol{\theta}\in V_\epsilon(\boldsymbol{\theta}_0)}\int_0^\tau \left(\frac{\exp\{\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}_i(t,\xi)\}}{s(t,\boldsymbol{\theta})} - \frac{\exp\{\boldsymbol{\beta}_0^\top \tilde{\boldsymbol{Z}}_i(t,\xi_0)\}}{s(t,\boldsymbol{\theta}_0)}\right) \mathrm{d}N_i(t) = o(1).$$

Otherwise,

$$E \sup_{\boldsymbol{\theta}\in V_\epsilon(\boldsymbol{\theta}_0)} n^{-3/2}\sum_{i\neq j}\int_0^\tau \left(\frac{R_j(t)\exp\{\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}_j(t,\xi)\}}{s(t,\boldsymbol{\theta})} - \frac{R_j(t)\exp\{\boldsymbol{\beta}_0^\top \tilde{\boldsymbol{Z}}_j(t,\xi_0)\}}{s(t,\boldsymbol{\theta}_0)}\right) \mathrm{d}N_i(t)$$

$$\begin{aligned}
&= E\Bigg[E\int_0^\tau \sup_{\boldsymbol{\theta}\in V_\epsilon(\boldsymbol{\theta}_0)} n^{-3/2}\sum_{i\neq j}\Bigg(\frac{R_j(t)\exp\{\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}_j(t,\xi)\}}{s(t,\boldsymbol{\theta})} \\
&\qquad\qquad - \frac{R_j(t)\exp\{\boldsymbol{\beta}_0^\top \tilde{\boldsymbol{Z}}_j(t,\xi_0)\}}{s(t,\boldsymbol{\theta}_0)}\Bigg) \mathrm{d}N_i(t) | R_j, \boldsymbol{Z}_j\Bigg] \\
&\le E\Bigg[\int_0^\tau \sup_{\boldsymbol{\theta}\in V_\epsilon(\boldsymbol{\theta}_0)} n^{-1/2}\sum_{j}\Bigg(\frac{R_j(t)\exp\{\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}_j(t,\xi)\}}{s(t,\boldsymbol{\theta})} \\
&\qquad\qquad - \frac{R_j(t)\exp\{\boldsymbol{\beta}_0^\top \tilde{\boldsymbol{Z}}_j(t,\xi_0)\}}{s(t,\boldsymbol{\theta}_0)}\Bigg) s(t,\boldsymbol{\theta}_0)\,\mathrm{d}\Lambda_0(t) | R_j, \boldsymbol{Z}_j\Bigg].
\end{aligned}$$

The integrand of the last term can be divided into four terms according to the location of ξ and ξ_0. Thus for $r \in \{0,1\}$ and $\boldsymbol{z} = (\boldsymbol{z}_1, \boldsymbol{z}_2)$ we consider the following families of functions:

$$\begin{aligned}
\phi_{1,t,\boldsymbol{\theta}}(r,\boldsymbol{z}) = r\Bigg\{&\frac{\exp\{\boldsymbol{\beta}_1^\top \boldsymbol{z} + \beta_2 z_2 + \beta_3(z_2 - \xi)\}}{s(t,\boldsymbol{\theta})} \\
&- \frac{\exp\{\boldsymbol{\beta}_{10}^\top \boldsymbol{z} + \beta_{20} z_2 + \beta_{30}(z_2 - \xi_0)\}}{s(t,\boldsymbol{\theta}_0)}\Bigg\} I_{\{z_2>\xi_0\}}
\end{aligned}$$

$$\phi_{2,t,\boldsymbol{\theta}}(r,\boldsymbol{z}) = r\left\{\frac{\exp\{\boldsymbol{\beta}_1^\top \boldsymbol{z} + \beta_2 z_2\}}{s(t,\boldsymbol{\theta})} - \frac{\exp\{\boldsymbol{\beta}_{10}^\top \boldsymbol{z} + \beta_{20} z_2\}}{s(t,\boldsymbol{\theta}_0)}\right\} I_{\{z_2<\xi_0\}}$$

$$\phi_{3,t,\boldsymbol{\theta}}(r,\boldsymbol{z}) = r\left\{\frac{\exp\{\boldsymbol{\beta}_1^\top \boldsymbol{z} + \beta_2 z_2\}}{s(t,\boldsymbol{\theta})} - \frac{\exp\{\boldsymbol{\beta}_{10}^\top \boldsymbol{z} + \beta_{20} z_2 + \beta_{30}(z_2-\xi_0)\}}{s(t,\boldsymbol{\theta}_0)}\right\} I_{\{\xi>z_2>\xi_0\}}$$

$$\phi_{4,t,\boldsymbol{\theta}}(r,\boldsymbol{z}) = r\left\{\frac{\exp\{\boldsymbol{\beta}_1^\top \boldsymbol{z} + \beta_2 z_2 + \beta_3(z_2-\xi)\}}{s(t,\boldsymbol{\theta})} - \frac{\exp\{\boldsymbol{\beta}_{10}^\top \boldsymbol{z} + \beta_{20} z_2\}}{s(t,\boldsymbol{\theta}_0)}\right\} I_{\{\xi_0>z_2>\xi\}}$$

For $k = 1, 2$, the functions $\phi_{k,t,\boldsymbol{\theta}}$ are continuously differentiable with respect to $\boldsymbol{\theta}$ and their derivatives are uniformly square integrable on $[0,\tau] \times V_\epsilon(\boldsymbol{\theta}_0)$. For $k = 3, 4$, the functions $\phi_{k,t,\boldsymbol{\theta}}$ are products of indicator functions $I_{(\xi,\xi_0)}$ with $\xi \in [\xi_0 - \epsilon^2, \xi_0]$, and of continuously differentiable functions with respect to $\boldsymbol{\theta}$. These continuously differentiable functions also have uniformly square integrable derivatives on $[0,\tau] \times V_\epsilon(\boldsymbol{\theta}_0)$. Furthermore, the class of functions $\{\phi_{k,t,\boldsymbol{\theta}} : \boldsymbol{\theta} \in V_\epsilon(\boldsymbol{\theta}_0)\}$ has a finite L_2-bracketing integral which does not depend on t. Hence, using Theorem 2.14.2 in Van der Vaart & Wellner (1996) we know that for $k = 1, \ldots, 4$

$$\int_0^\tau \left[E \sup_{\boldsymbol{\theta}\in V_\epsilon(\boldsymbol{\theta}_0)} \left|\frac{1}{\sqrt{n}}\sum_{i=1}^n \phi_{k,t,\boldsymbol{\theta}}(R,\boldsymbol{Z}) - E\phi_{k,t,\boldsymbol{\theta}}(R,\boldsymbol{Z})\right|\right] s(t,\boldsymbol{\theta}_0)\,\mathrm{d}\Lambda_0(t) = O(\epsilon).$$

Consequently, the sum is bounded by ϵ times a constant. Hence, the assertion of the lemma is proved. □

Theorem 4.3. *Under conditions A.1-A.5,* $\sqrt{n}\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\| = O_P(1)$.

Proof. Let $\epsilon > 0$ be sufficiently small to ensure that Lemma 4.3 holds on $V_\epsilon(\boldsymbol{\theta}_0)$. Because of Theorem 4.2 we know that $\hat{\boldsymbol{\theta}}_n$ converges to $\boldsymbol{\theta}_0$ in a neighborhood of $\boldsymbol{\theta}_0$, i.e. $\mathrm{P}(\hat{\boldsymbol{\theta}}_n \in V_\epsilon(\boldsymbol{\theta}_0)) > 1 - \eta$ for n sufficiently large and some $\eta > 0$.

Now, for each n, the parameter set $V_\epsilon(\boldsymbol{\theta}_0) \setminus \{\boldsymbol{\theta}_0\}$ can be partitioned into subsets $H_{n,j} = \{\boldsymbol{\theta} \in V_\epsilon(\boldsymbol{\theta}_0) : 2^j < \sqrt{n}\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\| \le 2^{j+1}\}$, $j \in \mathbb{Z}$. Based on ideas of Ibragimov & Has'minskii (1981) and for n sufficiently large we get by using Lemma 4.3 and Lemma 4.4 the following

$$\begin{aligned}
&\mathrm{P}\left(\sqrt{n}\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\| > M\right)\\
&\qquad \le \mathrm{P}\left(\sup_{\substack{\boldsymbol{\theta}\in V_\epsilon(\boldsymbol{\theta}_0)\\ M\le\sqrt{n}\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|}} X_n(\boldsymbol{\theta}) \ge X_n(\boldsymbol{\theta}_0)\right) + \eta\\
&\qquad \le \sum_{\{j:2^j>M\}} \mathrm{P}\left(\sup_{H_{n,j}} X_n(\boldsymbol{\theta}) - X_n(\boldsymbol{\theta}_0) \ge 0\right) + \eta
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{\{j:2^j>M\}} \mathrm{P}\left(\sup_{H_{n,j}}(W_n(\boldsymbol{\theta}) - W_n(\boldsymbol{\theta}_0)) \geq -\sqrt{n}(x(\boldsymbol{\theta}) - x(\boldsymbol{\theta}_0))\right) + \eta \\
&\leq \sum_{\{j:2^j>M\}} \mathrm{P}\left(\sup_{H_{n,j}}\left(W_n(\boldsymbol{\theta}) - W_n(\boldsymbol{\theta}_0)\right) \geq \sqrt{n}\alpha\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|^2\right) + \eta \\
&= \sum_{\{j:2^j>M\}} \mathrm{P}\left(\sup_{H_{n,j}}\sqrt{n}\left(W_n(\boldsymbol{\theta}) - W_n(\boldsymbol{\theta}_0)\right) \geq n\alpha\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|^2\right) + \eta \\
&\leq \sum_{\{j:2^j>M\}} \mathrm{P}\left(\sup_{H_{n,j}}\sqrt{n}\left(W_n(\boldsymbol{\theta}) - W_n(\boldsymbol{\theta}_0)\right) \geq \alpha\, 2^{2j}\right) + \eta \\
&\leq \sum_{\{j:2^j>M\}} \frac{E\left[\sup_{H_{n,j}}|W_n(\boldsymbol{\theta}) - W_n(\boldsymbol{\theta}_0)|\right]}{\alpha\, n^{-1/2}\, 2^{2j}} + \eta \\
&\leq \sum_{\{j:2^j>M\}} \frac{\kappa}{\alpha 2^{j-1}} + \eta.
\end{aligned}
$$

The last step holds because of Markov's inequality. This concludes the proof. □

4.6 Asymptotic Normality

In this section we prove the asymptotic normality of our estimates. Standard methods fail, since they use the differentiability of the partial likelihood function with respect to its parameters. We use a theorem which establishes the asymptotic normality of M-estimators in the case the criterion function is Lipschitz and its limit function admits a second order Taylor expansion. Consider the criterion function

$$
m_{\boldsymbol{\theta}} = m_{\boldsymbol{\theta}}(\boldsymbol{z}) = \int_0^\tau \left[\left(\boldsymbol{\beta}_1^\top, \beta_2, \beta_3\right)^\top \begin{pmatrix} \boldsymbol{z}_1(t) \\ z_2 \\ (z_2 - \xi)^+ \end{pmatrix} - \log(s(t, \boldsymbol{\theta}))\right] \mathrm{d}N(t)
$$

and the matrix $\Delta(\boldsymbol{\theta}_0) = E\dot{m}_{\boldsymbol{\theta}_0}\dot{m}_{\boldsymbol{\theta}_0}^\top$, where $\dot{m}_{\boldsymbol{\theta}_0}$ is given by

$$
\dot{m}_{\boldsymbol{\theta}_0} = \begin{pmatrix} -\int_0^\tau \left(\beta_{30} I_{\{z_2>\xi_0\}} + \frac{1}{s(t,\boldsymbol{\theta}_0)}\frac{\partial}{\partial\xi}s(t, \boldsymbol{\theta}_0)\right) \mathrm{d}N(t) \\ \int_0^\tau \left(\boldsymbol{z}_1(t) - \frac{1}{s(t,\boldsymbol{\theta}_0)}\frac{\partial}{\partial\boldsymbol{\beta}_1}s(t, \boldsymbol{\theta}_0)\right) \mathrm{d}N(t) \\ \int_0^\tau \left(z_2 - \frac{1}{s(t,\boldsymbol{\theta}_0)}\frac{\partial}{\partial\beta_2}s(t, \boldsymbol{\theta}_0)\right) \mathrm{d}N(t) \\ \int_0^\tau \left((z_2 - \xi_0)^+ - \frac{1}{s(t,\boldsymbol{\theta}_0)}\frac{\partial}{\partial\beta_3}s(t, \boldsymbol{\theta}_0)\right) \mathrm{d}N(t) \end{pmatrix}.
$$

Theorem 4.4. *Under conditions A.1-A.5 and under the assumption that $\hat{\boldsymbol{\theta}}_n$ is a consistent estimator of $\boldsymbol{\theta}_0$,*

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \xrightarrow{d} N(0, \boldsymbol{H}(\boldsymbol{\theta}_0)^{-1}\Delta(\boldsymbol{\theta}_0)\boldsymbol{H}(\boldsymbol{\theta}_0)^{-1}) \text{ as } n \to \infty,$$

where $\boldsymbol{H}(\boldsymbol{\theta}_0)$ is given as in Lemma 4.2.

Proof. Rewrite the criterion function as follows

$$m_{\boldsymbol{\theta}}(\boldsymbol{z}) = (\boldsymbol{\beta}_1^\top, \beta_2)^\top \int_0^\tau \begin{pmatrix} \boldsymbol{z}_1(t) \\ z_2 \end{pmatrix} dN(t) + \beta_3 \int_0^\tau (z_2 - \xi)^+ dN(t) - \int_0^\tau \log\left(s(t, \boldsymbol{\theta})\right) dN(t)$$

The function $\boldsymbol{z} \mapsto m_{\boldsymbol{\theta}}(\boldsymbol{z})$ is a measurable function such that $\boldsymbol{\theta} \mapsto m_{\boldsymbol{\theta}}(\boldsymbol{z})$ is differentiable at $\boldsymbol{\theta}_0$ for P-almost every $\boldsymbol{z}$ because of condition A.2. It can easily be seen that the first term of $m_{\boldsymbol{\theta}}$ is Lipschitz in $\boldsymbol{\beta}_1$ and β_2 since it is linear in $\boldsymbol{\beta}_1$ and β_2. The second term is Lipschitz in a neighborhood of $\boldsymbol{\theta}_0$ since

$$\begin{aligned}
&\left| \int_0^\tau \tilde{\beta}_3(z_2 - \tilde{\xi})^+ \, \mathrm{d}N(t) - \int_0^\tau \beta_3(z_2 - \xi)^+ \, \mathrm{d}N(t) \right| \\
&\leq |\xi - \tilde{\xi}| N(\tau) |\beta_3| + |\tilde{\beta}_3 - \beta_3| \int_0^\tau |(z_2 - \tilde{\xi})| \, \mathrm{d}N(t).
\end{aligned}$$

Now, for the third term, by a Taylor expansion at $\boldsymbol{\theta}$

$$\log(s(t, \tilde{\boldsymbol{\theta}})) - \log(s(t, \boldsymbol{\theta})) = \frac{\frac{\partial}{\partial \boldsymbol{\beta}} s(t, \boldsymbol{\theta}')}{s(t, \boldsymbol{\theta}')} (\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}) + \frac{\frac{\partial}{\partial \xi} s(t, \boldsymbol{\theta}')}{s(t, \boldsymbol{\theta}')} (\tilde{\xi} - \xi)$$

where $\boldsymbol{\theta}'$ is on the line segment between $\boldsymbol{\theta}$ and $\tilde{\boldsymbol{\theta}}$. The partial derivatives are uniformly bounded and bounded away form zero by conditions A.3 and A.4. Hence, the last term is Lipschitz in $\boldsymbol{\theta}$.

Furthermore, the map $\boldsymbol{\theta} \mapsto Em_{\boldsymbol{\theta}} = x(\boldsymbol{\theta})$ admits a second order Taylor expansion at $\boldsymbol{\theta}_0$ with the nonsingular symmetric Hessian matrix $\boldsymbol{H}(\boldsymbol{\theta}_0)$ given in Lemma 4.2.

Finally, since $\hat{\boldsymbol{\theta}}_n$ is consistent for $\boldsymbol{\theta}_0$ in a neighborhood of $\boldsymbol{\theta}_0$, it follows that $\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)$ is asymptotically normal with covariance matrix $\boldsymbol{H}(\boldsymbol{\theta}_0)^{-1}\Delta(\boldsymbol{\theta}_0)\boldsymbol{H}(\boldsymbol{\theta}_0)^{-1}$ by Theorem 5.23 in Van der Vaart (1998). □

Since $\boldsymbol{H}(\boldsymbol{\theta})$ and $\Delta(\boldsymbol{\theta})$ are continuous in $\boldsymbol{\theta}_0$, they can be consistently estimated by $\boldsymbol{H}(\hat{\boldsymbol{\theta}}_n)$ and $\Delta(\hat{\boldsymbol{\theta}}_n)$.

Using the approach of Andersen & Gill (1982) the weak convergence of $\sqrt{n}(\hat{\Lambda}_n(t) - \Lambda_0(t))$ can be established. Its asymptotic behavior follows from Theorem 4.4 and from the next result, which is the same if the underlying change-point ξ_0 was known.

Theorem 4.5. *Under conditions A.1-A.5 the process*

$$\sqrt{n}(\hat{\Lambda}_n(t) - \Lambda_0(t)) + \sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0)^\top \int_0^t \frac{E[R(u)\tilde{\boldsymbol{Z}}(u,\xi_0)\exp\{\boldsymbol{\beta}_0^\top \tilde{\boldsymbol{Z}}(u,\xi_0)\}]}{s(u,\boldsymbol{\theta}_0)} \,\mathrm{d}\Lambda_0(u)$$

converges weakly to a mean zero Gaussian process with covariance $\int_0^{s\wedge t} \frac{1}{s(u,\boldsymbol{\theta}_0)} \,\mathrm{d}\Lambda_0(u),\ s,t \in [0,\tau]$ *and* $\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0)$ *and the process above are asymptotically independent.*

Proof. Note that $S(u,\boldsymbol{\theta}) = \sum_{i=1}^n R_i(u)\exp\{\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}_i(u,\xi)\}$. Consider

$$\begin{aligned}
&\sqrt{n}(\hat{\Lambda}_n(t) - \Lambda_0(t)) \\
&\quad= \sqrt{n}\left\{\int_0^t \frac{\mathrm{d}(n\bar{N})(u)}{S(u,\hat{\boldsymbol{\theta}}_n)} - \Lambda_0(t)\right\} \\
&\quad= \sqrt{n}\left\{\int_0^t \frac{\mathrm{d}(n\bar{N})(u)}{S(u,\hat{\boldsymbol{\theta}}_n)} - \int_0^t \frac{S(u,\boldsymbol{\theta}_0)}{S(u,\hat{\boldsymbol{\theta}}_n)}\,\mathrm{d}\Lambda_0(u)\right. \\
&\qquad\qquad\left. + \int_0^t \frac{S(u,\boldsymbol{\theta}_0)}{S(u,\hat{\boldsymbol{\theta}}_n)}\,\mathrm{d}\Lambda_0(u) - \int_0^t \frac{S(u,\hat{\boldsymbol{\theta}}_n)}{S(u,\hat{\boldsymbol{\theta}}_n)}\,\mathrm{d}\Lambda_0(u)\right\} \\
&\quad= \sqrt{n}\left\{\int_0^t \frac{\mathrm{d}(n\bar{N})(u) - S(u,\boldsymbol{\theta}_0)\,\mathrm{d}\Lambda_0(u)}{S(u,\hat{\boldsymbol{\theta}}_n)} - \int_0^t \frac{S(u,\hat{\boldsymbol{\theta}}_n) - S(u,\boldsymbol{\theta}_0)}{S(u,\hat{\boldsymbol{\theta}}_n)}\,\mathrm{d}\Lambda_0(u)\right\} \\
&\quad= \int_0^t \frac{\mathrm{d}[n^{1/2}\bar{M}(u)]}{n^{-1}S(u,\hat{\boldsymbol{\theta}}_n)} - \int_0^t \frac{n^{-1/2}[S(u,\hat{\boldsymbol{\theta}}_n) - S(u,\boldsymbol{\theta}_0)]}{n^{-1}S(u,\hat{\boldsymbol{\theta}}_n)}\,\mathrm{d}\Lambda_0(u)
\end{aligned}$$

where $\bar{M}(u) = \frac{1}{n}\sum_{i=1}^n M_i(u)$. The first term in the last expression converges to a centered Gaussian process with covariance $\int_0^{s\wedge t} \frac{1}{s(u,\boldsymbol{\theta}_0)}\,\mathrm{d}\Lambda_0(u)$ by Rebolledos theorem (see Rebolledo (1980)). For the second term consider a Taylor expansion at $\boldsymbol{\beta}_0$

$$\begin{aligned}
n^{-1/2}\left(S(u,\hat{\boldsymbol{\theta}}_n) - S(u,\boldsymbol{\theta}_0)\right) &= n^{-1/2}\left(S(u,\boldsymbol{\beta}_0,\hat{\xi}_n) - S(u,\boldsymbol{\beta}_0,\xi_0)\right) \\
&\quad + n^{-1/2}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0)^\top \left(\sum_{i=1}^n R_i(u)\tilde{\boldsymbol{Z}}_i(u,\hat{\xi}_n)\exp\left\{\boldsymbol{\beta}_*^\top \tilde{\boldsymbol{Z}}_i(u,\hat{\xi}_n)\right\}\right),
\end{aligned}$$

where $\boldsymbol{\beta}_*$ is on the line segment between $\boldsymbol{\beta}_0$ and $\hat{\boldsymbol{\beta}}_n$. The first term of the Taylor expansion converges uniformly in $u \in [0,\tau]$ to zero in probability by using the continuous mapping theorem, since S is a continuous function in ξ. By the strong law of large numbers given by Andersen & Gill (1982) the following difference

$$\sup_{u\in[0,\tau]}\sup_{\boldsymbol{\theta}\in\Theta}\left\|\frac{1}{n}\sum_{i=1}^n R_i(u)\tilde{\boldsymbol{Z}}_i(u,\xi)\exp\{\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}_i(u,\xi)\} - E[R(u)\tilde{\boldsymbol{Z}}(u,\xi)\exp\{\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}(u,\xi)\}]\right\|$$

converges to zero in probability and

$$\sup_{u\in[0,\tau]}\sup_{\boldsymbol{\theta}\in\Theta}|n^{-1}S(u,\boldsymbol{\theta})-s(u,\boldsymbol{\theta})|\xrightarrow{\mathrm{P}}0.$$

The asymptotic independence follows from the approximation

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_n-\boldsymbol{\beta}_0)=\left(n^{-1}\frac{\partial^2}{(\partial\boldsymbol{\beta})^2}\log L(\boldsymbol{\theta}_0)\right)^{-1}\cdot n^{-1/2}\frac{\partial}{\partial\boldsymbol{\beta}}\log L(\boldsymbol{\theta}_0)+o_P(1),$$

where

$$\frac{\partial}{\partial\boldsymbol{\beta}}\log L(\boldsymbol{\theta}_0)=\sum_{i=1}^n\int_0^\tau\tilde{\boldsymbol{Z}}_i(u,\xi_0)\,\mathrm{d}M_i(u)-\int_0^\tau\frac{E[R(u)\tilde{\boldsymbol{Z}}(u,\xi_0)\exp\{\boldsymbol{\beta}_0^\top\tilde{\boldsymbol{Z}}(u,\xi_0)\}]}{S(u,\boldsymbol{\theta}_0)}\,\mathrm{d}\bar{M}(u),$$

since $n^{-1/2}\frac{\partial}{\partial\boldsymbol{\beta}}\log L(\boldsymbol{\theta}_0)$ and $\int_0^t\frac{n^{1/2}\,\mathrm{d}\bar{M}(u)}{s(u,\boldsymbol{\theta}_0)}$ are asymptotically Gaussian with mean zero and

$$E\left[\int_0^t\frac{n^{1/2}\,\mathrm{d}\bar{M}(u)}{s(u,\boldsymbol{\theta}_0)}\cdot n^{-1/2}\frac{\partial}{\partial\boldsymbol{\beta}}\log L(\boldsymbol{\theta}_0)\right]=0$$

for all $t\in[0,\tau]$. $\square$

Chapter 5

Cox Model with Change-Points and a General Risk Function

In this chapter we examine a further extended version of the Cox model. In contrast to the last chapter we allow a general risk function, multiple change-points of the type we have discussed before and a counting process, which may jump more than once. This model can be seen in two different ways. Firstly, it allows us to insert several change-points in a single covariate , secondly one can use it to describe multiple change-points in different covariates.
An example of a different risk function is $r(x) = 1 + x^2$. The main problem that arises is that the function r is not so well-behaved as the exponential function, such that some work-around has to be made.
Again we show the asymptotic properties such as consistency and asymptotic normality of our estimates and the rate of convergence of the change-point parameter vector. The ideas, results and proofs are similar to those stated in Chapter 4.

5.1 Model and Estimation

We use nearly the same setup as we did in the last chapter. Our random quantities $(N, R, \boldsymbol{Z}_1(t), \boldsymbol{Z}_2(t))\,, (N_i, R_i, \boldsymbol{Z}_{1i}(t), \boldsymbol{Z}_{2i}(t))\,, i = 1, ..., n$ are given on a filtered probability space and they are independently identically distributed. Both, $\boldsymbol{Z}_1(t)$ and $\boldsymbol{Z}_2(t)$ are predictable and adapted stochastic processes taking values in $\mathbb{R}^p$ and $\mathbb{R}^q$, respectively. Consider a multivariate counting process $\boldsymbol{N}(t) = (N_1(t), \ldots, N_n(t))$, where $N_i(t)$ counts observed events in the lifetime of the ith individual, $i = 1, \ldots, n$, over the time interval $[0, \tau]$. The sample paths of $\boldsymbol{N}(t)$ are step functions, zero at time zero with jumps of size one only and two arbitrary components do not jump at the same time. The counting process $\boldsymbol{N}(t)$ admits an intensity $\boldsymbol{\lambda}(t) = (\lambda_1(t), \ldots, \lambda_n(t))$ such that the processes $M_i(t) =$

$N_i(t) - \int_0^t \lambda_i(u)\,\mathrm{d}u$, $i = 1, \ldots, n$, and $t \in [0, \tau]$ are martingales. Then, the model involving time-dependent covariates with possible q change-points is given by:

$$\lambda_i(t, \boldsymbol{\theta}) = \lambda_0(t) R_i(t)\, r\left\{\boldsymbol{\beta}_1^\top \boldsymbol{Z}_{1i}(t) + \boldsymbol{\beta}_2^\top \boldsymbol{Z}_{2i}(t) + \boldsymbol{\beta}_3^\top (\boldsymbol{Z}_{2i}(t) - \boldsymbol{\xi})^+\right\},$$

where $\boldsymbol{\theta} = (\boldsymbol{\xi}^\top, \boldsymbol{\beta}^\top)^\top$ with $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^\top, \boldsymbol{\beta}_2^\top, \boldsymbol{\beta}_3^\top)^\top \in \mathcal{B} \subset \mathbb{R}^{p+2q}$ is the vector of regression parameters, $\lambda_0(t)$ is the baseline intensity and $R_i(t)$ is a process taking only values 1 or 0 to indicate whether a subject is at risk or not. The function $r : \mathbb{R} \to [0, \infty)$ is a twice continuously differentiable nonnegative known function. Again, we use for brevity,

$$\tilde{\boldsymbol{Z}}_i(t; \boldsymbol{\xi}) = \left(\boldsymbol{Z}_{1i}^\top(t), \boldsymbol{Z}_{2i}^\top(t), ((\boldsymbol{Z}_{2i}(t) - \boldsymbol{\xi})^+)^\top\right)^\top.$$

The vector of change-points is indicated by $\boldsymbol{\xi} \in \mathbb{R}^q$, which is a vector of parameters lying in a rectangle $\Xi = [\xi_{11}, \xi_{21}] \times [\xi_{12}, \xi_{22}] \times \cdots \times [\xi_{1q}, \xi_{2q}]$. The parameters $\xi_{11}, \xi_{21}, \xi_{12}, \xi_{22}$,..., ξ_{1q}, ξ_{2q} are assumed to be known. The true parameter values $\boldsymbol{\theta}_0 = (\boldsymbol{\xi}_0^\top, \boldsymbol{\beta}_0^\top)^\top$ are supposed to be identifiable, meaning that at least one component of $\boldsymbol{\beta}_{30}$ is unequal to 0. The parameter $\boldsymbol{\theta}_0$ is estimated by the value $\hat{\boldsymbol{\theta}}_n$ that maximizes the logarithm of the partial likelihood

$$\begin{aligned}\log L(\boldsymbol{\theta}) =& \sum_{i=1}^n \int_0^\tau \log\left(r\{\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}_i(t; \boldsymbol{\xi})\}\right) \mathrm{d}N_i(t)\\ &- \int_0^\tau \log\left(\sum_{i=1}^n R_i(t)\, r\{\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}_i(t; \boldsymbol{\xi})\}\right) \mathrm{d}\left(\sum_{i=1}^n N_i(t)\right).\end{aligned}$$

The maximization is carried out in two phases again:
For fixed $\boldsymbol{\xi}$, let $\hat{\boldsymbol{\beta}}_n(\boldsymbol{\xi}) = \arg\max_{\boldsymbol{\beta} \in \mathcal{B}} \log L(\boldsymbol{\xi}, \boldsymbol{\beta})$ and $\log L(\boldsymbol{\xi}) = \log L(\boldsymbol{\xi}, \hat{\boldsymbol{\beta}}_n(\boldsymbol{\xi}))$. Then $\boldsymbol{\xi}_0$ can be estimated by $\hat{\boldsymbol{\xi}}_n$ satisfying

$$\hat{\boldsymbol{\xi}}_n = \arg\max_{\boldsymbol{\xi} \in \Xi} \log L(\boldsymbol{\xi}).$$

The partial likelihood estimate of $\boldsymbol{\theta}_0$ is $\hat{\boldsymbol{\theta}}_n = (\hat{\boldsymbol{\xi}}_n, \hat{\boldsymbol{\beta}}_n)$, where $\hat{\boldsymbol{\beta}}_n = \hat{\boldsymbol{\beta}}_n(\hat{\boldsymbol{\xi}}_n)$.
Since $\log L(\boldsymbol{\theta})$ does not converge to a finite limit as in the univariate case we consider the process

$$\begin{aligned}X_n(\boldsymbol{\theta}) \quad = \quad & \frac{1}{n} \sum_{i=1}^n \int_0^\tau \log\left(r\{\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}_i(t; \boldsymbol{\xi})\}\right) \mathrm{d}N_i(t) \\ & - \int_0^\tau \log\{\frac{1}{n} \sum_{i=1}^n R_i(t)\, r\{\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}_i(t; \boldsymbol{\xi})\}\, \mathrm{d}\left(\frac{1}{n} \sum_{i=1}^n N_i(t)\right).\end{aligned} \tag{5.1}$$

Obviously, the estimate $\hat{\boldsymbol{\theta}}_n$ not only maximizes $\log L(\boldsymbol{\theta})$ but also $X_n(\boldsymbol{\theta})$ and the limit of $X_n(\boldsymbol{\theta})$ as $n \to \infty$ will be given by

$$\begin{aligned} x(\boldsymbol{\theta}) := & E\left[\int_0^\tau \log\left(r\{\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}(t;\boldsymbol{\xi})\}\right) r\{\boldsymbol{\beta}_0^\top \tilde{\boldsymbol{Z}}(t;\boldsymbol{\xi}_0)\}R(t)\lambda_0(t)\,\mathrm{d}t\right] \\ & -E\left[\int_0^\tau \log\left(s(t;\boldsymbol{\theta})\right) r\{\boldsymbol{\beta}_0^\top \tilde{\boldsymbol{Z}}(t;\boldsymbol{\xi}_0)\}R(t)\lambda_0(t)\,\mathrm{d}t\right], \end{aligned} \tag{5.2}$$

where $s(t;\boldsymbol{\theta}) = E[R(t)\, r\{\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}(t;\boldsymbol{\xi})\}]$.
Now, the cumulative hazard function $\Lambda_0(t) = \int_0^t \lambda_0(u)\,\mathrm{d}u$ can be estimated by the Breslow estimator

$$\hat{\Lambda}_n(t) = \int_0^t \frac{\mathrm{d}\left(\sum_{i=1}^n N_i(u)\right)}{S(u,\hat{\boldsymbol{\theta}}_n)},$$

where $S(u,\boldsymbol{\theta}) = \sum_{i=1}^n R_i(u) r\{\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}_i(u,\xi)\}$.

5.2 Conditions

Similar conditions as in the last chapter are needed to establish the asymptotic properties of the estimates. Condition C.6 refers to properties of the general link function. We use the notation

$$\mathrm{d}\,\mathrm{P}^t = q_t^{-1}\,\mathrm{d}\,\mathrm{Q}^t, \quad \mathrm{Q}^t(A) = \int_A R(t) r(\boldsymbol{\beta}_0^\top \tilde{\boldsymbol{Z}}(t;\xi_0))\,\mathrm{d}\,\mathrm{P} \text{ and } q_t = \int \mathrm{d}\,\mathrm{Q}^t,$$

provided that $q_t < \infty$.
There exists a convex and compact set $\Theta \subset \mathbb{R}^{p+3q}$ with $\boldsymbol{\theta}_0$ in its interior such that the following holds:

Conditions.

C.1 [Finite baseline intensity] $\sup_{t\in[0,\tau]} \lambda_0(t) < \infty$.

C.2 The random vector $\boldsymbol{Z}_2(t)$ has an absolutely continuous distribution with density $f_{\boldsymbol{Z}_2(t)}$ which is strictly positive, bounded and continuous in a neighborhood of $\boldsymbol{\xi}_0$ for every $t \in [0,\tau]$.

C.3 The expectation $E\sup_{\boldsymbol{\theta}\in\Theta}\left\{r\{\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}(t;\boldsymbol{\xi})\}\right\} < \infty$.

C.4 [Asymptotic regularity conditions] *The function $s(t;\boldsymbol{\theta}) = E[R(t)\, r\{\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}(t;\boldsymbol{\xi})\}]$ is bounded away from zero on $[0,\tau]\times\Theta$ and the first two partial derivatives of $s(t;\boldsymbol{\theta})$ with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\xi}$ are bounded on $[0,\tau]\times\Theta$ and continuous on Θ, uniformly in $t \in [0,\tau]$.*

C.5 a) For all $(\boldsymbol{\xi}, \boldsymbol{\beta}) \in \Theta$ *and* $t \in [0, \tau]$ *the covariance matrix* $\mathrm{Cov}_{\mathrm{P}^t}(\boldsymbol{Y}(t))$, *where* $\boldsymbol{Y}(t) = \tilde{Y}(t, \boldsymbol{\theta}_0)\left((-\boldsymbol{\beta}_{30} I_{\{\boldsymbol{Z}_2 > \boldsymbol{\xi}_0\}})^\top, \tilde{\boldsymbol{Z}}^\top(t, \boldsymbol{\xi})\right)^\top$ *with* $\tilde{Y}(t, \boldsymbol{\theta}_0) = \frac{r'(\boldsymbol{\beta}_0^\top \tilde{\boldsymbol{Z}}_i(t, \boldsymbol{\xi}_0))}{r(\boldsymbol{\beta}_0^\top \tilde{\boldsymbol{Z}}_i(t, \boldsymbol{\xi}_0))}$ *is positive definite.*

b)Furthermore,

$$E\left[\sup_{t\in[0,\tau]} \sup_{\boldsymbol{\theta}\in\Theta} \left\{ \frac{r'\{\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}(t;\boldsymbol{\xi})\}}{r\{\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}(t;\boldsymbol{\xi})\}} r\{\boldsymbol{\beta}_0^\top \tilde{\boldsymbol{Z}}(t;\boldsymbol{\xi}_0)\} \right\}^2\right] < \infty$$

and for k=1,2,

$$E\left[\sup_{t\in[0,\tau]} \sup_{\boldsymbol{\theta}\in\Theta} \left\{ (\|\boldsymbol{Z}_1(t)\|^k + \|\boldsymbol{Z}_2(t)\|^k) \frac{r^{(k)}\{\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}(t;\boldsymbol{\xi})\}}{r\{\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}(t;\boldsymbol{\xi})\}} r\{\boldsymbol{\beta}_0^\top \tilde{\boldsymbol{Z}}(t;\boldsymbol{\xi}_0)\} \right\}^2\right] < \infty,$$

$$\sup_{\boldsymbol{z},\boldsymbol{z}'} \sup_{t\in[0,\tau]} \sup_{\boldsymbol{\theta}\in\Theta} \left| E\left\{ r^{(k)}\{\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}(t;\boldsymbol{\xi})\} | \boldsymbol{Z}_2(t) = \boldsymbol{z} \right\} - E\left\{ r^{(k)}\{\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}(t;\boldsymbol{\xi})\} | \boldsymbol{Z}_2(t) = \boldsymbol{z}' \right\} \right|$$

converges to zero as $\|\boldsymbol{z} - \boldsymbol{z}'\| \to 0$, *where* $\boldsymbol{z}$ *and* $\boldsymbol{z}'$ *vary in* Ξ.

C.6 [Regression function positivity] There exists a neighborhood Θ_0 *of* $\boldsymbol{\theta}_0$ *such that, for* $\boldsymbol{\theta} \in \Theta_0$, $r\{\boldsymbol{\beta}^\top \boldsymbol{Z}_i(t;\boldsymbol{\xi})\}$ *is locally bounded away form zero for all* $i = 1, ..., n$.

The last statement in condition C.5 is used for interchanging integration and differentiation.

5.3 Consistency of the Estimator

As before the proof of consistency relies on the uniform convergence of X_n to x, see (5.1) and (5.2), and on properties of x in a neighborhood of Θ_0. Of course, the proofs are similar to those with the exponential function as link function. Therefore, we want to clarify the differences.

Lemma 5.1. *Under conditions C.1-C.4 and C.6,*

$$\sup_{\boldsymbol{\theta}\in\Theta} |X_n(\boldsymbol{\theta}) - x(\boldsymbol{\theta})| \xrightarrow{\mathrm{P}} 0 \text{ as } (n \to \infty).$$

Proof. $X_n(\boldsymbol{\theta})$ can be rewritten in the following way:

$$\begin{aligned} X_n(\boldsymbol{\theta}) =& \frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau} \log\left(r\{\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}_i(t;\boldsymbol{\xi})\}\right) \mathrm{d}N_i(t) \\ & - \int_0^{\tau} \log\left(n^{-1}\sum_{i=1}^{n} R_i(t)\, r\{\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}_i(t;\boldsymbol{\xi})\}\right) d\left(\frac{1}{n}\sum_{i=1}^{n} N_i(t)\right) \end{aligned} \tag{5.3}$$

The predictability of each $\boldsymbol{Z}_i$, the continuity of r and condition C.6 ensure that

$$\log\left(r\{\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}_i(t;\boldsymbol{\xi})\}\right) \text{ and } \log\left\{\frac{1}{n}\sum_{i=1}^{n} R_i(t)\, r\{\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}_i(t;\boldsymbol{\xi})\}\right\}$$

are predictable and locally bounded for each $\boldsymbol{\theta} \in \Theta_0$. Consider the first term of (5.3). By conditions C.3 and C.6 and by the continuity of $r(\cdot)$ and $\log(\cdot)$ for all $\boldsymbol{\xi} \in \Xi$,

$$E\left[\int_0^{\tau} \log(r\{\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}(t;\boldsymbol{\xi})\})\lambda(t,\boldsymbol{\theta}_0)\,\mathrm{d}t\right] \leq E\left[\int_0^{\tau} \sup_{\boldsymbol{\theta}\in\Theta_0} \log(r\{\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}(t;\boldsymbol{\xi})\})\lambda(t,\boldsymbol{\theta}_0)\,\mathrm{d}t\right] < \infty,$$

Hence,

$$\int_0^{t} \log\left(r\{\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}(t;\boldsymbol{\xi})\}\right) dM(s)$$

is a martingale and

$$E\left[\int_0^{\tau} \log\left(r\{\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}(t;\boldsymbol{\xi})\}\right) \mathrm{d}N(t)\right] = E\left[\int_0^{\tau} \log\left(r\{\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}(t;\boldsymbol{\xi})\}\right) \lambda(t,\boldsymbol{\theta}_0)\,\mathrm{d}t\right] < \infty.$$

We want to apply the Glivenko-Cantelli Theorem given in Theorem 19.4 and Example 19.8 in Van der Vaart (1998). Clearly, $\int_0^{\tau} \sup_{\boldsymbol{\beta}\in\mathcal{B}_0} \log\left(r\{\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}(t;\boldsymbol{\xi}_1)\}\right) \mathrm{d}N(t)$ is an envelope function for $\int_0^{\tau} \log\left(r\{\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}(t;\boldsymbol{\xi})\}\right) \mathrm{d}N(t)$. Since $\int_0^{\tau} \log\left(r\{\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}(t;\boldsymbol{\xi})\}\right) \mathrm{d}N(t)$ is continuous in $\boldsymbol{\xi}$, we get

$$\sup_{\boldsymbol{\theta}\in\Theta}\left|\frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau} \log\left(r\{\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}(t;\boldsymbol{\xi})\}\right) \mathrm{d}N_i(t) - E\left[\int_0^{\tau} \log\left(r\{\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}(t;\boldsymbol{\xi})\}\right) \lambda(t,\boldsymbol{\theta}_0)\,\mathrm{d}t\right]\right| \xrightarrow{\mathrm{P}} 0.$$

For the second term we can use the same argumentation as in the last chapter. The strong law of large numbers given by Andersen & Gill (1982) can be used, since $R(t) r\{\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}(t,\boldsymbol{\xi})\}$ are caglad (left continuous with right hand limits) functions due to the fact that r is continuous.

Theorem 5.2. *Under conditions C.1-C.6 there exists a neighborhood Θ_0 of $\boldsymbol{\theta}_0$ such that if $\hat{\boldsymbol{\theta}}_n$ lies in Θ_0, it follows that $\hat{\boldsymbol{\theta}}_n$ converges in probability to $\boldsymbol{\theta}_0$ as $n \to \infty$.*

Proof. By Lemma 5.1 we know that X_n converges uniformly to x. Hence it suffices to show that x is strictly concave in a neighborhood $\Theta_0 \subset \Theta$ and attains a maximum at $\boldsymbol{\theta}_0$. Consider the derivatives with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\xi}$. To simplify notation we will use as described in Section 2.2 $\frac{\partial}{\partial \boldsymbol{\beta}}$ and $\frac{\partial}{\partial \boldsymbol{\xi}}$ as a short form for the derivatives with respect to components of $\boldsymbol{\beta}$ and $\boldsymbol{\xi}$, respectively. Furthermore, let

$$\tilde{Y}(t, \boldsymbol{\theta}) = \frac{r'\{\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}(t; \boldsymbol{\xi})\}}{r\{\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}(t; \boldsymbol{\xi})\}}$$

$$\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\beta}} x(\boldsymbol{\theta}) = & \int_0^\tau \Bigg[E\left[\tilde{\boldsymbol{Z}}(t, \boldsymbol{\xi}) \tilde{Y}(t, \boldsymbol{\theta}) r\{\boldsymbol{\beta}_0^\top \tilde{\boldsymbol{Z}}(t; \boldsymbol{\xi}_0)\} R(t) \right] \\
& - \frac{s(t; \boldsymbol{\theta}_0)}{s(t; \boldsymbol{\theta})} \left(\frac{\partial}{\partial \boldsymbol{\beta}} s(t; \boldsymbol{\theta}) \right) \Bigg] \lambda_0(t) \, \mathrm{d}t, \\
\frac{\partial}{\partial \boldsymbol{\xi}} x(\boldsymbol{\theta}) = & \int_0^\tau \Bigg[E\left[(-\boldsymbol{\beta}_3) I_{\{\boldsymbol{Z}_2(t) > \boldsymbol{\xi}\}} \tilde{Y}(t, \boldsymbol{\theta}) r\{\boldsymbol{\beta}_0^\top \tilde{\boldsymbol{Z}}(t; \boldsymbol{\xi}_0)\} R(t) \right] \\
& - \frac{s(t; \boldsymbol{\theta}_0)}{s(t; \boldsymbol{\theta})} \left(\frac{\partial}{\partial \boldsymbol{\xi}} s(t; \boldsymbol{\theta}) \right) \Bigg] \lambda_0(t) \, \mathrm{d}t,
\end{aligned}$$

where

$$\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\xi}} s(t; \boldsymbol{\theta}) = & E[R(t)(-\boldsymbol{\beta}_3) I_{\{\boldsymbol{Z}_2(t) > \boldsymbol{\xi}\}} r'\{\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}(t; \boldsymbol{\xi})\}], \\
\frac{\partial}{\partial \boldsymbol{\beta}} s(t; \boldsymbol{\theta}) = & E[R(t) \tilde{\boldsymbol{Z}}(t, \boldsymbol{\xi}) r'\{\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}(t; \boldsymbol{\xi})\}].
\end{aligned}$$

Differentiation and integration can be interchanged because of condition C.5. Hence, $\frac{\partial}{\partial \boldsymbol{\xi}} x(\boldsymbol{\theta}_0) = \frac{\partial}{\partial \boldsymbol{\beta}} x(\boldsymbol{\theta}_0) = 0$ follows.
Next, we calculate the Hessian matrix of x

$$\boldsymbol{H}(\boldsymbol{\theta}_0) = \begin{pmatrix} \frac{\partial^2}{(\partial \boldsymbol{\xi})^2} x(\boldsymbol{\theta}_0) & \frac{\partial^2}{\partial \boldsymbol{\xi} \partial \boldsymbol{\beta}} x(\boldsymbol{\theta}_0) \\ \frac{\partial^2}{\partial \boldsymbol{\beta} \partial \boldsymbol{\xi}} x(\boldsymbol{\theta}_0) & \frac{\partial^2}{(\partial \boldsymbol{\beta})^2} x(\boldsymbol{\theta}_0) \end{pmatrix}.$$

Therefore, we need the second partial derivatives of $x(\boldsymbol{\theta})$ which exist because of condition C.5 and the Lebesgue differentiation theorem. Using the notation

$$\mathrm{Q}^t(A) = \int_A R(t) r(\boldsymbol{\beta}_0^\top \tilde{\boldsymbol{Z}}(t; \xi_0)) \, \mathrm{d}\,\mathrm{P}, \; q_t = \int \mathrm{d}\,\mathrm{Q}^t$$

and

$$\tilde{Y}^{(1)}(t, \boldsymbol{\theta}) = \left[\frac{r''\{\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}(t, \boldsymbol{\xi})\}}{r\{\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}(t, \boldsymbol{\xi})\}} - \frac{(r'\{\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}(t, \boldsymbol{\xi})\})^2}{(r\{\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}(t, \boldsymbol{\xi})\})^2} \right]$$

we get

$$
\begin{aligned}
\frac{\partial^2}{(\partial\boldsymbol{\beta})^2}x(\boldsymbol{\theta}) = \int_0^\tau \Bigg\{ & \int \tilde{\boldsymbol{Z}}(t,\boldsymbol{\xi})(\tilde{\boldsymbol{Z}}(t,\boldsymbol{\xi}))^\top \tilde{Y}^{(1)}(t,\boldsymbol{\theta})\,\mathrm{d}Q^t \\
& + q_t \left[\frac{\left(\frac{\partial}{(\partial\boldsymbol{\beta})}s(t,\boldsymbol{\theta})\right)\left(\frac{\partial}{(\partial\boldsymbol{\beta})}s(t,\boldsymbol{\theta})\right)^\top}{(s(t,\boldsymbol{\theta}))^2} - \frac{\frac{\partial^2}{(\partial\boldsymbol{\beta})^2}s(t,\boldsymbol{\theta})}{s(t,\boldsymbol{\theta})} \right] \Bigg\} \lambda_0(t)\,\mathrm{d}t \\
\frac{\partial^2}{(\partial\boldsymbol{\beta})^2}x(\boldsymbol{\theta}_0) = \int_0^\tau \Bigg\{ & -\int \tilde{\boldsymbol{Z}}(t,\boldsymbol{\xi}_0)(\tilde{\boldsymbol{Z}}(t,\boldsymbol{\xi}_0))^\top (\tilde{Y}(t,\boldsymbol{\theta}_0))^2\,\mathrm{d}Q^t \\
& + \frac{1}{q_t}\left(\int \tilde{\boldsymbol{Z}}(t,\boldsymbol{\xi}_0)\tilde{Y}(t,\boldsymbol{\theta}_0)\,\mathrm{d}Q^t\right)\left(\int \tilde{\boldsymbol{Z}}(t,\boldsymbol{\xi}_0)\tilde{Y}(t,\boldsymbol{\theta}_0)\,\mathrm{d}Q^t\right)^\top \Bigg\} \lambda_0(t)\,\mathrm{d}t \\
\frac{\partial^2}{(\partial\boldsymbol{\xi})^2}x(\boldsymbol{\theta}) = \int_0^\tau \Bigg\{ & \int (-\boldsymbol{\beta}_3 I_{\{\boldsymbol{Z}_2>\boldsymbol{\xi}\}})(-\boldsymbol{\beta}_3 I_{\{\boldsymbol{Z}_2>\boldsymbol{\xi}\}})^\top \tilde{Y}^{(1)}(t,\boldsymbol{\theta})\,\mathrm{d}Q^t \\
& + q_t \left[\frac{\left(\frac{\partial}{\partial\boldsymbol{\xi}}s(t,\boldsymbol{\theta})\right)\left(\frac{\partial}{\partial\boldsymbol{\xi}}s(t,\boldsymbol{\theta})\right)^\top}{(s(t,\boldsymbol{\theta}))^2} - \frac{\frac{\partial^2}{(\partial\boldsymbol{\xi})^2}s(t,\boldsymbol{\theta})}{s(t,\boldsymbol{\theta})} \right] \Bigg\} \lambda_0(t)\,\mathrm{d}t \\
\frac{\partial^2}{(\partial\boldsymbol{\xi})^2}x(\boldsymbol{\theta}_0) = \int_0^\tau \Bigg\{ & -\int (-\boldsymbol{\beta}_{30} I_{\{\boldsymbol{Z}_2>\boldsymbol{\xi}_0\}})(-\boldsymbol{\beta}_{30} I_{\{\boldsymbol{Z}_2>\boldsymbol{\xi}_0\}})^\top (\tilde{Y}(t,\boldsymbol{\theta}_0))^2\,\mathrm{d}Q^t \\
& + \frac{1}{q_t}\left(\int (-\boldsymbol{\beta}_{30} I_{\{\boldsymbol{Z}_2>\boldsymbol{\xi}_0\}})\tilde{Y}(t,\boldsymbol{\theta}_0)\,\mathrm{d}Q^t\right)\left(\int (-\boldsymbol{\beta}_{30} I_{\{\boldsymbol{Z}_2>\boldsymbol{\xi}_0\}})\tilde{Y}(t,\boldsymbol{\theta}_0)\,\mathrm{d}Q^t\right)^\top \Bigg\} \lambda_0(t)\,\mathrm{d}t \\
\frac{\partial^2}{\partial\boldsymbol{\xi}\partial\boldsymbol{\beta}}x(\boldsymbol{\theta}) = \int_0^\tau \Bigg\{ & \int (-\boldsymbol{\beta}_3 I_{\{\boldsymbol{Z}_2>\boldsymbol{\xi}\}})(\tilde{\boldsymbol{Z}}(t,\boldsymbol{\xi}))^\top \tilde{Y}^{(1)}(t,\boldsymbol{\theta})\,\mathrm{d}Q^t \\
& + q_t \left[\frac{\left(\frac{\partial}{\partial\boldsymbol{\xi}}s(t,\boldsymbol{\theta})\right)\left(\frac{\partial}{\partial\boldsymbol{\beta}}s(t,\boldsymbol{\theta})\right)^\top}{(s(t,\boldsymbol{\theta}))^2} - \frac{\frac{\partial^2}{\partial\boldsymbol{\xi}\partial\boldsymbol{\beta}}s(t,\boldsymbol{\theta})}{s(t,\boldsymbol{\theta})} \right] \Bigg\} \lambda_0(t)\,\mathrm{d}t \\
\frac{\partial^2}{\partial\boldsymbol{\xi}\partial\boldsymbol{\beta}}x(\boldsymbol{\theta}_0) = \int_0^\tau \Bigg\{ & -\int (-\boldsymbol{\beta}_{30} I_{\{\boldsymbol{Z}_2>\boldsymbol{\xi}_0\}})(\tilde{\boldsymbol{Z}}(t,\boldsymbol{\xi}_0))^\top (\tilde{Y}(t,\boldsymbol{\theta}_0))^2\,\mathrm{d}Q^t \\
& + \frac{1}{q_t}\left(\int (-\boldsymbol{\beta}_{30} I_{\{\boldsymbol{Z}_2>\boldsymbol{\xi}_0\}})\tilde{Y}(t,\boldsymbol{\theta}_0)\,\mathrm{d}Q^t \int \tilde{\boldsymbol{Z}}(t,\boldsymbol{\xi}_0)\tilde{Y}(t,\boldsymbol{\theta}_0)\,\mathrm{d}Q^t\right) \Bigg\} \lambda_0(t)\,\mathrm{d}t,
\end{aligned}
$$

where

$$
\begin{aligned}
\frac{\partial^2}{(\partial\boldsymbol{\beta})^2}s(t;\boldsymbol{\theta}) =& E[R_i(t)\tilde{\boldsymbol{Z}}(t,\boldsymbol{\xi})(\tilde{\boldsymbol{Z}}(t,\boldsymbol{\xi}))^\top r''\{\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}(t;\boldsymbol{\xi})\}] \\
\frac{\partial^2}{(\partial\boldsymbol{\beta})^2}s(t;\boldsymbol{\theta}_0) =& \int \tilde{\boldsymbol{Z}}(t,\boldsymbol{\xi}_0)(\tilde{\boldsymbol{Z}}(t,\boldsymbol{\xi}_0))^\top \frac{r''\{\boldsymbol{\beta}_0^\top \tilde{\boldsymbol{Z}}(t,\boldsymbol{\xi}_0)\}}{r\{\boldsymbol{\beta}_0^\top \tilde{\boldsymbol{Z}}(t,\boldsymbol{\xi}_0)\}}\,\mathrm{d}Q^t
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial^2}{(\partial \boldsymbol{\xi})^2} s(t;\boldsymbol{\theta}) =& E[R_i(t)(-\boldsymbol{\beta}_3 I_{\{\boldsymbol{Z}_2(t)>\boldsymbol{\xi}\}})(-\boldsymbol{\beta}_3 I_{\{\boldsymbol{Z}_2(t)>\boldsymbol{\xi}\}})^\top r''\{\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}(t;\boldsymbol{\xi})\}] \\
& - E[R_i(t)(-\boldsymbol{\beta}_3) r'\{\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}(t;\boldsymbol{\xi})\} | \boldsymbol{Z}_2(t) = \boldsymbol{\xi}] f_{\boldsymbol{Z}_2}(\boldsymbol{\xi}), \\
\frac{\partial^2}{(\partial \boldsymbol{\xi})^2} s(t;\boldsymbol{\theta}_0) =& \int (-\boldsymbol{\beta}_{30} I_{\{\boldsymbol{Z}_2>\boldsymbol{\xi}_0\}})(-\boldsymbol{\beta}_{30} I_{\{\boldsymbol{Z}_2>\boldsymbol{\xi}_0\}})^\top \frac{r''\{\boldsymbol{\beta}_0^\top \tilde{\boldsymbol{Z}}(t,\boldsymbol{\xi}_0)\}}{r\{\boldsymbol{\beta}_0^\top \tilde{\boldsymbol{Z}}(t,\boldsymbol{\xi}_0)\}} \, \mathrm{d}Q^t \\
& - \int (-\boldsymbol{\beta}_{30} I_{\{\boldsymbol{Z}_2>\boldsymbol{\xi}_0\}}) \tilde{Y}(t,\boldsymbol{\theta}_0) \, \mathrm{d}Q^t \\
\frac{\partial^2}{\partial \boldsymbol{\xi} \partial \boldsymbol{\beta}} s(t;\boldsymbol{\theta}) =& E[R_i(t)(-\boldsymbol{\beta}_3 I_{\{\boldsymbol{Z}_2>\boldsymbol{\xi}\}})(\tilde{\boldsymbol{Z}}(t,\boldsymbol{\xi}))^\top r''\{\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}(t;\boldsymbol{\xi})\}] \\
\frac{\partial^2}{\partial \boldsymbol{\xi} \partial \boldsymbol{\beta}} s(t;\boldsymbol{\theta}_0) =& \int (-\boldsymbol{\beta}_{30} I_{\{\boldsymbol{Z}_2>\boldsymbol{\xi}\}})(\tilde{\boldsymbol{Z}}(t,\boldsymbol{\xi}_0))^\top \frac{r''\{\boldsymbol{\beta}_0^\top \tilde{\boldsymbol{Z}}(t,\boldsymbol{\xi}_0)\}}{r\{\boldsymbol{\beta}_0^\top \tilde{\boldsymbol{Z}}(t,\boldsymbol{\xi}_0)\}} \, \mathrm{d}Q^t
\end{aligned}
$$

As before we can show that

$$
\boldsymbol{H}(\boldsymbol{\theta}_0) = \int_0^\tau (\boldsymbol{H}_1(\boldsymbol{\theta}_0) - \boldsymbol{H}_2(\boldsymbol{\theta}_0)) \lambda_0(t) \, \mathrm{d}t
$$

with $\boldsymbol{H}_1(\boldsymbol{\theta}_0) = \int \boldsymbol{Y} \, \mathrm{d}Q^t$ and $\boldsymbol{H}_2(\boldsymbol{\theta}_0) = q_t^{-1} \left(\int \boldsymbol{Y} \, \mathrm{d}Q^t \right)^2$, where in this case

$$
\boldsymbol{Y} = \tilde{Y}(t,\boldsymbol{\theta}_0) \begin{pmatrix} -\boldsymbol{\beta}_{30} I_{\{\boldsymbol{Z}_2>\boldsymbol{\xi}_0\}} \\ \tilde{\boldsymbol{Z}}(t,\boldsymbol{\xi}_0) \end{pmatrix}.
$$

Hence, it follows that $\boldsymbol{H}(\boldsymbol{\theta}_0)$ is negative semidefinite. Condition C.5 ensures that $\boldsymbol{H}(\boldsymbol{\theta}_0)$ is negative definite. Furthermore, $\boldsymbol{H}$ is continuous in $\boldsymbol{\theta}$. We conclude, since $\boldsymbol{H}$ is continuous and negative definite at $\boldsymbol{\theta}_0$, there exists a neighborhood Θ_0 of $\boldsymbol{\theta}_0$ on which $\boldsymbol{H}(\boldsymbol{\theta})$ is negative definite for all $\boldsymbol{\theta}$ in Θ_0. □

5.4 Rate of Convergence

As before, we can show that the rate of convergence is $\sqrt{n}$ for all parameters.
Let $V_\epsilon(\boldsymbol{\theta}_0) = \{\boldsymbol{\theta} : \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| < \epsilon\}$ be an ϵ-neighborhood of $\boldsymbol{\theta}_0$ again and consider the process $W_n(\boldsymbol{\theta}) = \sqrt{n}(X_n(\boldsymbol{\theta}) - x(\boldsymbol{\theta}))$.

Lemma 5.2. *Under conditions C.1-C.6, for ϵ sufficiently small there exists a constant $\alpha > 0$ such that for all $\boldsymbol{\theta}$ in $V_\epsilon(\boldsymbol{\theta}_0)$, $x(\boldsymbol{\theta}) - x(\boldsymbol{\theta}_0) \leq -\alpha \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|^2$.*

Proof. The proof is a perfect analogy to the proof of Lemma 4.3. □

Lemma 5.3. *Under conditions C.1-C.6, for every $\epsilon > 0$ there exists a constant $\kappa > 0$ such that $E[\sup_{\boldsymbol{\theta} \in V_\epsilon(\boldsymbol{\theta}_0)} |W_n(\boldsymbol{\theta}) - W_n(\boldsymbol{\theta}_0)|] \leq \kappa\epsilon$, for all n.*

Proof. Rewrite $W_n(\boldsymbol{\theta}) - W_n(\boldsymbol{\theta}_0) = W_{1n}(\boldsymbol{\theta}) - W_{2n}(\boldsymbol{\theta})$ with W_{1n} and W_{2n} as follows

$$\begin{aligned}W_{2n}(\boldsymbol{\theta}) =& \sqrt{n}\left(\int_0^\tau \log(\bar{S}(t,\boldsymbol{\theta}))d\bar{N}(t) - \int_0^\tau \log(s(t,\boldsymbol{\theta}))s(t,\theta_0)d\Lambda_0(t)\right.\\ &\left.-\int_0^\tau \log(\bar{S}(t,\boldsymbol{\theta}_0))d\bar{N}(t) - \int_0^\tau \log(s(t,\boldsymbol{\theta}_0))s(t,\theta_0)d\Lambda_0(t)\right)\\ =& n^{-1/2}\sum_{i=1}^n\left[\int_0^\tau \log\left(\frac{\bar{S}(t,\boldsymbol{\theta})}{\bar{S}(t,\boldsymbol{\theta}_0)}\right)\mathrm{d}N_i(t) - \int_0^\tau \log\left(\frac{s(t,\boldsymbol{\theta})}{s(t,\boldsymbol{\theta}_0)}\right)s(t,\boldsymbol{\theta}_0)d\Lambda_0(t)\right],\end{aligned}$$

where $\bar{S}(t;\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^n R_i(t) r\{\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}_i(t,\boldsymbol{\xi})\}$ and

$$\begin{aligned}W_{1n}(\boldsymbol{\theta}) \;=\;& n^{-1/2}\sum_{i=1}^n\left[\int_0^\tau \log\left(r\{\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}_i(t;\boldsymbol{\xi})\}\right) - \log\left(r\{\boldsymbol{\beta}_0^\top \tilde{\boldsymbol{Z}}_i(t;\boldsymbol{\xi})\}\right)\mathrm{d}N_i(t)\right.\\ &+\int_0^\tau \log\left(r\{\boldsymbol{\beta}_0^\top \tilde{\boldsymbol{Z}}_i(t;\boldsymbol{\xi})\}\right) - \log\left(r\{\boldsymbol{\beta}_0^\top \tilde{\boldsymbol{Z}}_i(t;\boldsymbol{\xi}_0)\}\right)\mathrm{d}N_i(t)\\ &-E\int_0^\tau \log\left(r\{\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}(t;\boldsymbol{\xi})\}\right) - \log\left(r\{\boldsymbol{\beta}_0^\top \tilde{\boldsymbol{Z}}(t;\boldsymbol{\xi})\}\right)\lambda(t,\boldsymbol{\theta}_0)\,\mathrm{d}t\\ &\left.-E\int_0^\tau \log\left(r\{\boldsymbol{\beta}_0^\top \tilde{\boldsymbol{Z}}(t;\boldsymbol{\xi})\}\right) - \log\left(r\{\boldsymbol{\beta}_0^\top \tilde{\boldsymbol{Z}}(t;\boldsymbol{\xi}_0)\}\right)\lambda(t,\boldsymbol{\theta}_0)\,\mathrm{d}t\right]\end{aligned}$$

The expectation of the supremum of the absolute value of W_{2n} is $O(\epsilon)$ using the same arguments as in Lemma 4.4.
Consider W_{1n}. A Taylor expansion of $\log\left(r\{\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}_i(t;\boldsymbol{\xi})\}\right)$ at $\boldsymbol{\beta}_0$ yields

$$\log\left(r\{\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}_i(t;\boldsymbol{\xi})\}\right) - \log\left(r\{\boldsymbol{\beta}_0^\top \tilde{\boldsymbol{Z}}_i(t;\boldsymbol{\xi})\}\right) = (\boldsymbol{\beta}-\boldsymbol{\beta}_0)^\top \tilde{\boldsymbol{Z}}_i(t;\boldsymbol{\xi})\frac{r'\{\boldsymbol{\beta}_*^\top \tilde{\boldsymbol{Z}}_i(t;\boldsymbol{\xi})\}}{r\{\boldsymbol{\beta}_*^\top \tilde{\boldsymbol{Z}}_i(t;\boldsymbol{\xi})\}},$$

where $\boldsymbol{\beta}_*$ is on the line segment between $\boldsymbol{\beta}$ and $\boldsymbol{\beta}_0$. Substituting this into $W_{1n}(\boldsymbol{\theta})$ we get

$$\begin{aligned}W_{1n}(\boldsymbol{\theta}) =& n^{-1/2}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)^\top\sum_{i=1}^n\left\{\int_0^\tau \tilde{\boldsymbol{Z}}_i(t;\boldsymbol{\xi})\frac{r'\{\boldsymbol{\beta}_*^\top \tilde{\boldsymbol{Z}}_i(t;\boldsymbol{\xi})\}}{r\{\boldsymbol{\beta}_*^\top \tilde{\boldsymbol{Z}}_i(t;\boldsymbol{\xi})\}}\mathrm{d}N_i(t)\right.\\ &\left.-E\int_0^\tau \tilde{\boldsymbol{Z}}(t;\boldsymbol{\xi})\frac{r'\{\boldsymbol{\beta}_*^\top \tilde{\boldsymbol{Z}}(t;\boldsymbol{\xi})\}}{r\{\boldsymbol{\beta}_*^\top \tilde{\boldsymbol{Z}}(t;\boldsymbol{\xi})\}}\lambda(t,\boldsymbol{\theta}_0)\,\mathrm{d}t\right\}\\ &+n^{-1/2}\sum_{i=1}^n\left\{\int_0^\tau \log\left(r\{\boldsymbol{\beta}_0^\top \tilde{\boldsymbol{Z}}_i(t;\boldsymbol{\xi})\}\right) - \log\left(r\{\boldsymbol{\beta}_0^\top \tilde{\boldsymbol{Z}}_i(t;\boldsymbol{\xi}_0)\}\right)\mathrm{d}N_i(t)\right.\\ &\left.-E\int_0^\tau \log\left(r\{\boldsymbol{\beta}_0^\top \tilde{\boldsymbol{Z}}(t;\boldsymbol{\xi})\}\right) - \log\left(r\{\boldsymbol{\beta}_0^\top \tilde{\boldsymbol{Z}}(t;\boldsymbol{\xi}_0)\}\right)\lambda(t,\boldsymbol{\theta}_0)\,\mathrm{d}t\right\}.\end{aligned}$$

The expectation of the supremum of the first term is $O(\epsilon)$ using C.3 and C.6. For the second difference, note that

$$\int_0^\tau \log\left(r\{\boldsymbol{\beta}_0^\top \tilde{\boldsymbol{Z}}(t;\boldsymbol{\xi}_0-\epsilon)\}\right)\,\mathrm{d}N_i(t)$$

is an envelope for the class of functions

$$\mathcal{F}_\epsilon = \left\{\int_0^\tau \log\left(r\{\boldsymbol{\beta}_0^\top \tilde{\boldsymbol{Z}}(t;\boldsymbol{\xi})\}\right)\,\mathrm{d}N_i(t) : \|\boldsymbol{\xi}-\boldsymbol{\xi}_0\| \le \epsilon\right\}.$$

Furthermore, since r and $\log(\cdot)$ are Lipschitz functions,

$$\left|\log(r(\boldsymbol{\beta}_0^\top \tilde{\boldsymbol{Z}}(t,\boldsymbol{\xi}))) - \log(r(\boldsymbol{\beta}_0^\top \tilde{\boldsymbol{Z}}(t,\boldsymbol{\xi}_0)))\right| \le K\left|\boldsymbol{\beta}_0^\top \tilde{\boldsymbol{Z}}(t,\boldsymbol{\xi}) - \boldsymbol{\beta}_0^\top \tilde{\boldsymbol{Z}}(t,\boldsymbol{\xi}_0)\right|, \text{for some } K<\infty.$$

Again we will consider the components of the vectors individually as discussed in Section 2.2. The sets of functions $\{f_\xi : \xi \in [\xi_1,\xi_2]\}$ and $\{g_\xi : \xi \in [\xi_1,\xi_2]\}$ with $f_\xi(b) = bI_{\{b>\xi\}}$ and $g_\xi(b) = \xi I_{\{b>\xi\}}$ form Vapnik-Cervonenkis classes. By Theorem 2.6.7 in Van der Vaart & Wellner (1996) we know that the class $\mathcal{F}_\epsilon$ has a finite entropy number. Moreover,

$$\begin{aligned}
&E\sup_{\boldsymbol{\xi}\in V_\epsilon(\boldsymbol{\xi}_0)}\left|\int_0^\tau \log\left(r\{\boldsymbol{\beta}_0^\top \tilde{\boldsymbol{Z}}(t;\boldsymbol{\xi})\}\right) - \log\left(r\{\boldsymbol{\beta}_0^\top \tilde{\boldsymbol{Z}}(t;\boldsymbol{\xi}_0)\}\right)\,\mathrm{d}N(t)\right| \\
&\quad\le \left\{E\int_0^\tau \left|\log\left(r\{\boldsymbol{\beta}_0^\top \tilde{\boldsymbol{Z}}(t;\boldsymbol{\xi}-\mathbf{1}\cdot\epsilon)\}\right) - \log\left(r\{\boldsymbol{\beta}_0^\top \tilde{\boldsymbol{Z}}(t;\boldsymbol{\xi}_0)\}\right)\right|^2\,\mathrm{d}N(t)\right\}^{1/2} = O(\epsilon).
\end{aligned}$$

Thus,

$$\begin{aligned}
&E\sup_{\boldsymbol{\xi}\in V_\epsilon(\boldsymbol{\xi}_0)}\Bigg|n^{-1/2}\sum_{i=1}^n\int_0^\tau \log\left(r\{\boldsymbol{\beta}_0^\top \tilde{\boldsymbol{Z}}_i(t;\boldsymbol{\xi})\}\right) - \log\left(r\{\boldsymbol{\beta}_0^\top \tilde{\boldsymbol{Z}}_i(t;\boldsymbol{\xi}_0)\}\right)\,\mathrm{d}N_i(t) \\
&\qquad - E[\int_0^\tau \log\left(r\{\boldsymbol{\beta}_0^\top \tilde{\boldsymbol{Z}}(t;\boldsymbol{\xi})\}\right) - \log\left(r\{\boldsymbol{\beta}_0^\top \tilde{\boldsymbol{Z}}(t;\boldsymbol{\xi}_0)\}\right)\,\mathrm{d}N(t)]\Bigg| = O(\epsilon)
\end{aligned}$$

as a consequence of Theorem 2.14.1 in Van der Vaart & Wellner (1996). □

Using the theorems and lemmas above the following theorem can be proved in analogy to Theorem 4.3. Hence, the rate of convergence is established.

Theorem 5.3. *Under conditions C.1-C.6,* $\sqrt{n}\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\| = O_P(1)$.

5.5 Asymptotic Normality

In this section we prove the asymptotic normality of our estimates. We use similar arguments as in Section 4.6.

Consider the criterion function

$$m_{\boldsymbol{\theta}}(\boldsymbol{z}) = \int_0^\tau \log\left(r\{\boldsymbol{\beta}^\top \tilde{\boldsymbol{z}}(t;\boldsymbol{\xi})\}\right) \mathrm{d}N(t) - \int_0^\tau \log\left(s(t,\boldsymbol{\theta})\right) \mathrm{d}N(t)$$

and the matrix $\Delta(\boldsymbol{\theta}_0) = E\dot{m}_{\boldsymbol{\theta}_0}\dot{m}_{\boldsymbol{\theta}_0}$, where $\dot{m}_{\boldsymbol{\theta}_0}$ is given by

$$\dot{m}_{\boldsymbol{\theta}_0} = \begin{pmatrix} \int_0^\tau (-\boldsymbol{\beta}_{30} I_{\{\boldsymbol{z}_2(t)>\boldsymbol{\xi}_0\}}) \frac{r\{\boldsymbol{\beta}_0^\top \tilde{\boldsymbol{z}}(t,\boldsymbol{\xi}_0)\}}{r'\{\boldsymbol{\beta}_0^\top \tilde{\boldsymbol{z}}(t,\boldsymbol{\xi}_0)\}} - \frac{\frac{\partial}{\partial \boldsymbol{\xi}} s(t,\boldsymbol{\theta}_0)}{s(t,\boldsymbol{\theta}_0)} \mathrm{d}N(t) \\ \int_0^\tau \tilde{\boldsymbol{z}}(t,\boldsymbol{\xi}_0) \frac{r\{\boldsymbol{\beta}_0^\top \tilde{\boldsymbol{z}}(t,\boldsymbol{\xi}_0)\}}{r'\{\boldsymbol{\beta}_0^\top \tilde{\boldsymbol{z}}(t,\boldsymbol{\xi}_0)\}} - \frac{\frac{\partial}{\partial \boldsymbol{\beta}} s(t,\boldsymbol{\theta}_0)}{s(t,\boldsymbol{\theta}_0)} \mathrm{d}N(t) \end{pmatrix}.$$

Theorem 5.4. *Under conditions C.1-C.6 and under the assumption that $\hat{\boldsymbol{\theta}}_n$ is a consistent estimator of $\boldsymbol{\theta}_0$, $\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)$ is asymptotically normal with mean zero and covariance matrix $\boldsymbol{H}(\boldsymbol{\theta}_0)^{-1}\Delta(\boldsymbol{\theta}_0)\boldsymbol{H}(\boldsymbol{\theta}_0)^{-1}$.*

Proof. The function $\boldsymbol{z} \mapsto m_{\boldsymbol{\theta}}(\boldsymbol{z})$ is a measurable function such that $\boldsymbol{\theta} \mapsto m_{\boldsymbol{\theta}}(\boldsymbol{z})$ is differentiable at $\boldsymbol{\theta}_0$ for P-almost every $\boldsymbol{z}$ because of condition C.1. The first term of the function m is Lipschitz in $\boldsymbol{\theta} = (\boldsymbol{\xi}, \boldsymbol{\beta})$ since piecewise linear functions are Lipschitz and the composition of functions which are Lipschitz is Lipschitz again. The second term is Lipschitz in a neighborhood of $\boldsymbol{\theta}_0$, which follows from a Taylor expansion

$$\log(s(t,\tilde{\boldsymbol{\theta}})) - \log(s(t,\boldsymbol{\theta})) = \frac{\frac{\partial}{\partial \boldsymbol{\beta}} s(t,\boldsymbol{\theta}')}{s(t,\boldsymbol{\theta}')}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}) + \frac{\frac{\partial}{\partial \boldsymbol{\xi}} s(t,\boldsymbol{\theta}')}{s(t,\boldsymbol{\theta}')}(\tilde{\boldsymbol{\xi}} - \boldsymbol{\xi})$$

where $\boldsymbol{\theta}'$ is on the line segment between $\boldsymbol{\theta}$ and $\tilde{\boldsymbol{\theta}}$. The partial derivatives are uniformly bounded and bounded away form zero by conditions C.3, C.4 and C.6. Hence, the last term is Lipschitz in $\boldsymbol{\theta}$.
Moreover, the map $\boldsymbol{\theta} \mapsto Em_{\boldsymbol{\theta}} = x(\boldsymbol{\theta})$ admits a second order Taylor expansion at $\boldsymbol{\theta}_0$ with nonsingular symmetric second derivative matrix $\boldsymbol{H}(\boldsymbol{\theta}_0)$, which has been calculated in Theorem 5.2. Finally, since $\hat{\boldsymbol{\theta}}_n$ is a consistent estimator of $\boldsymbol{\theta}_0$ in a neighborhood of $\boldsymbol{\theta}_0$, the assertion of the theorem follows using Theorem 5.23 in Van der Vaart (1998). Note that $\boldsymbol{H}(\boldsymbol{\theta}_0)$ and $\Delta(\boldsymbol{\theta}_0)$ are continuous in $\boldsymbol{\theta}$, thus they can be estimated consistently by $\boldsymbol{H}(\hat{\boldsymbol{\theta}}_n)$ and $\Delta(\hat{\boldsymbol{\theta}}_n)$. □

The weak convergence of $\sqrt{n}(\hat{\Lambda}_n(t) - \Lambda_0(t))$ can be established as in Chapter 4. Therefore, we state the theorem without proof.

Theorem 5.5. *Under conditions C.1-C.6 the process*

$$\sqrt{n}(\hat{\Lambda}_n(t) - \Lambda_0(t)) + \sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0)^\top \int_0^t \frac{E[R(u)\tilde{\boldsymbol{Z}}(u,\boldsymbol{\xi}_0) r\{\boldsymbol{\beta}_0^\top \tilde{\boldsymbol{Z}}(u,\boldsymbol{\xi}_0)\}]}{s(u,\boldsymbol{\theta}_0)} \mathrm{d}\Lambda_0(u)$$

converges weakly to a mean zero Gaussian process with covariance $\int_0^{s\wedge t} \frac{1}{s(u,\boldsymbol{\theta}_0)}\,\mathrm{d}\Lambda_0(u),\ s,t \in [0,\tau]$ *and* $\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0)$ *and the process above are asymptotically independent.*

Chapter 6

A Transformation Model with a Bent-Line Change-Point

In this chapter a further generalization of the models before is discussed. We consider a linear transformation model with bent-line change-points in the covariates. The main difference between the models explained earlier and this new model is, that the cumulative intensity function and the other parameters can not be estimated separately. Therefore, some new techniques are involved. We use a nonparametric maximum likelihood method instead of a partial likelihood method. In this way we can estimate the finite-dimensional regression and change-point parameters as well as the infinite-dimensional parameter of the cumulative baseline intensity function. The parameters are estimable with the same precision as if the true threshold of the covariates were known. This model includes the Cox model with a change-point but also so-called frailty models.
Kosorok & Song (2007) considered a similar model. But they included in their model a change-point in which the underlying regression function is discontinuous. Some of our proofs are based on the same techniques they used.

6.1 Model

Consider a linear transformation model for a nonnegative survival time T which is given by

$$\log A(T) = -\boldsymbol{\beta}^\top \boldsymbol{Z} + \epsilon,$$

where A is an unspecified monotone increasing transformation and ϵ follows a known error distribution not depending on the covariates $\boldsymbol{Z}$. If $S_{\boldsymbol{Z}}(t) = \mathrm{P}[T > t|\boldsymbol{Z}]$ denotes the survival function of T given the covariates $\boldsymbol{Z}$ and $S_\epsilon(t) = 1 - F_\epsilon(t)$, where $F_\epsilon(t)$ is the distribution function of ϵ, the model can be equivalently written in the form

$$S_{\boldsymbol{Z}}(t) = S_\epsilon\left(\log A(t) + \boldsymbol{\beta}^\top \boldsymbol{Z}\right).$$

Choosing $S_\epsilon(u) = \Lambda(e^u)$ results in the model

$$S_{\boldsymbol{Z}}(t) = \Lambda\left(\int_0^t \exp\{\boldsymbol{\beta}^\top \boldsymbol{Z}\}\, \mathrm{d}A(u)\right),$$

where the function Λ is known, thrice differentiable and decreasing with $\Lambda(0) = 1$. Several choices of the function $\Lambda(u)$ satisfy the model conditions, which we will give in Section 6.3. One example is $\Lambda(u) = \exp\{-u\}$. Choosing this extreme value distribution in the model results in a Cox model. Another choice could be $\Lambda(u) = (1+cu)^{-1/c}$, $c \in (0,\infty)$, which belongs to the family of log-Pareto distributions and results in an odds-rate transformation family. The limit $c \to 0$ leads to a Cox model and if $c = 1$ then we obtain a proportional odds model. A further different possibility is to choose $\Lambda(u) = E[\exp\{-Wu\}]$, where W is a positive frailty with $E[W^{-c}] < \infty$, for some $c > 0$ and $E[W^4] < \infty$. Thus, we are able to consider a family of frailty transformations. Especially, the conditions are fulfilled for the inverse Gaussian and log-normal families. Verification of this last statement is given in Kosorok & Song (2007).

We refine this general model setup. Consider censored survival time data given by $(V, \delta, \boldsymbol{Z})$, where $V = T \wedge C$ and $\delta = I_{\{T \leq C\}}$ for a survival time T and a censoring time C. Furthermore, let $0 < \tau < \infty$ and let $\boldsymbol{Z} = \{\boldsymbol{Z}(t), t \in [0,\tau]\}$ denote a left continuous covariate process with right hand limits and with $\boldsymbol{Z}(t) = (\boldsymbol{Z}_1(t), \boldsymbol{Z}_2) \in \mathbb{R}^p \times \mathbb{R}^q$. The data $(V_i, \delta_i, \boldsymbol{Z}_i)$, $i = 1, \dots, n$ consists of n iid copies of $(V, \delta, \boldsymbol{Z})$.

Thus, the transformation model for a survival time T conditionally on $\boldsymbol{Z}$ is given by

$$\begin{aligned} S_{\boldsymbol{Z}}(t) &= \Lambda\left(\int_0^t \exp\{\boldsymbol{\beta}_1^\top \boldsymbol{Z}_1(u) + \boldsymbol{\beta}_2^\top \boldsymbol{Z}_2 + \boldsymbol{\beta}_3^\top (\boldsymbol{Z}_2 - \boldsymbol{\xi})^+\}\, \mathrm{d}A(u)\right) \\ &= \Lambda\left(\int_0^t \exp\{\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}(u, \boldsymbol{\xi})\}\, \mathrm{d}A(u)\right), \end{aligned} \tag{6.1}$$

where $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^\top, \boldsymbol{\beta}_2^\top, \boldsymbol{\beta}_3^\top)^\top \in \mathbb{R}^{p+2q}$ is a parameter vector and $\boldsymbol{\xi} \in \mathbb{R}^q$ is a change-point vector. The function Λ is known, thrice differentiable and decreasing with $\Lambda(0) = 1$ and A is an unknown increasing function restricted to $[0, \tau]$.

For convenience, let $G = -\log(\Lambda)$ and denote the derivatives $\Lambda'(u) = \frac{\partial \Lambda(u)}{\partial u}$, $\Lambda''(u) = \frac{\partial \Lambda'(u)}{\partial u}$, $G'(u) = \frac{\partial G(u)}{\partial u}$, $G''(u) = \frac{\partial G'(u)}{\partial u}$ and $G'''(u) = \frac{\partial G''(u)}{\partial u}$. Moreover, define the combined parameter $\boldsymbol{\theta} = (\boldsymbol{\xi}, \boldsymbol{\beta}, A) = (\boldsymbol{\psi}, A)$. The true parameter values and the true probability measure are denoted with a subscript 0.

6.2 Nonparametric Maximum Likelihood Estimation

In this model setup we cannot separate the estimation of the finite-dimensional regression and change-point parameters and the estimation of the infinite-dimensional cumulative intensity function parameter as in the previous chapters. Therefore, a different concept is needed. To obtain estimates we use a nonparametric maximum likelihood.

Consider the model (6.1) with $\Lambda = \exp\{-G\}$. The standard likelihood for right-censored survival data and A being absolutely continuous with respect to the Lebesgue measure is

$$\prod_{i=1}^{n} \left\{ G'\left(H^{\boldsymbol{\theta}}(V_i)\right) \exp\{\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}_i\} a(V_i) \exp\{-G\left(H^{\boldsymbol{\theta}}(V_i)\right)\} \right\}^{\delta_i} \left\{\exp\{-G\left(H^{\boldsymbol{\theta}}(V_i)\right)\}\right\}^{1-\delta_i},$$

where $a(t) = \frac{\mathrm{d}}{\mathrm{d}t} A(t)$, $H^{\boldsymbol{\theta}}(s) = \int_0^s R_i(t) \exp\{\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}_i(t, \boldsymbol{\xi})\} \,\mathrm{d}A(t)$ and $R(t) = I_{\{V \geq t\}}$. Hence, the nonparametric log-likelihood scaled by $1/n$ is given by

$$\tilde{L}_n(\boldsymbol{\theta}) = \mathbb{P}_n \left[\delta \log(a(V)) + \int_0^\tau \left\{ \log\left(G'(H^{\boldsymbol{\theta}}(s)) \right) + \boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}(s, \boldsymbol{\xi}) \right\} \mathrm{d}N(s) - G(H^{\boldsymbol{\theta}}(V)) \right] \tag{6.2}$$

where $N(t) = \delta I_{\{V \leq t\}}$ and $\mathbb{P}_n$ denotes the empirical measure, i.e. $\mathbb{P}_n f(V) = \frac{1}{n} \sum_{i=1}^n f(V_i)$. For more details on constructing the nonparametric likelihood we refer to Slud & Vonta (2004). A problem, which arises by considering the nonparametric log-likelihood is, that there exists no maximizer if A is continuous, as every unrestricted maximizer of (6.2) puts mass at observed failure times and is thus not a continuous hazard. Therefore, we extend the set of hazard functions and allow also discrete hazard functions. Thus, we replace $a(t)$ with $n\Delta A(t)$ as suggested in Parner (1998), where $\Delta A(s) = A(s) - A(s-)$ and $A(s-) = \lim_{t \uparrow s} A(t)$. In this case we get an estimate $\hat{A}$ for A_0 instead of an estimate for $a_0(t)$. The likelihood (6.2) with $a(t)$ replaced by $n\Delta A(t)$ is denoted by $L_n(\boldsymbol{\theta})$.

The estimates are obtained by two-phase maximization. For fixed $\boldsymbol{\xi}$, maximize the full nonparametric log-likelihood over $\boldsymbol{\gamma} = (\boldsymbol{\beta}, A)$ to obtain the profile likelihood $pL_n(\boldsymbol{\xi}) = \sup_{\boldsymbol{\gamma}} L_n(\boldsymbol{\theta})$. Then maximize $pL_n(\boldsymbol{\xi})$ over $\boldsymbol{\xi}$ to obtain $\hat{\boldsymbol{\xi}}_n$ and compute

$$\hat{\boldsymbol{\gamma}}_n = \arg\max_{\boldsymbol{\gamma}} L_n(\boldsymbol{\gamma}, \hat{\boldsymbol{\xi}}_n),$$

which yields the nonparametric maximum likelihood estimator (NPMLE) $\hat{\boldsymbol{\theta}}_n = (\hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\xi}})$ for $\boldsymbol{\theta}_0$. For the construction of an estimate of A we need to consider the following one-dimensional submodels of A:

$$u \mapsto A_u = \int_0^{(\cdot)} (1 + ug(s)) dA(s),$$

where $g : [0, \tau] \to \mathbb{R}$ is an arbitrary measurable bounded nonnegative function and $(\cdot)$ denotes an argument ranging over $[0, \tau]$. Using this submodel we can define a score function for A as the derivative of $L_n(\boldsymbol{\psi}, A_u)$ with respect to u at $u = 0$, which is given by

$$\frac{\partial}{\partial u} L_n(\boldsymbol{\theta})\big|_{u=0} =$$
$$\mathbb{P}_n \left[\delta g(V) - \left\{ G^{'}(H^{\boldsymbol{\theta}}(V)) - \delta \frac{G^{''}(H^{\boldsymbol{\theta}}(V))}{G^{'}(H^{\boldsymbol{\theta}}(V))} \right\} \cdot \int_0^\tau R(s) \exp\{\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}(s, \boldsymbol{\xi})\} g(s) \,\mathrm{d}A(s) \right]. \tag{6.3}$$

Now, for any fixed $\boldsymbol{\psi}$, we will denote the maximizer of $A \mapsto L_n(\boldsymbol{\theta})$ by $\hat{A}_{\boldsymbol{\psi}}$, and $\hat{\boldsymbol{\theta}}_{\boldsymbol{\psi}} = (\boldsymbol{\psi}, \hat{A}_{\boldsymbol{\psi}})$. Choosing $g(s) = I_{\{s \le t\}}$ and $t \in [0, \tau]$ in (6.3), we solve the equation $\frac{\partial}{\partial u} L_n(\boldsymbol{\theta})\big|_{u=0} = 0$. A solution to this equation is given by the recursive formula

$$\hat{A}_{\boldsymbol{\psi}}(s) = \int_0^s \left[\mathbb{P}_n W(u; \hat{\boldsymbol{\theta}}_{\boldsymbol{\psi}}) \right]^{-1} \mathbb{P}_n(\,\mathrm{d}N(u)), \tag{6.4}$$

where

$$W(u, \hat{\boldsymbol{\theta}}_{\boldsymbol{\psi}}) = R(u) \exp\{\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}(u, \boldsymbol{\xi})\} \left\{ G^{'}(H^{\hat{\boldsymbol{\theta}}_{\psi}}(V)) - \delta \frac{G''(H^{\hat{\boldsymbol{\theta}}_{\psi}}(V))}{G^{'}(H^{\hat{\boldsymbol{\theta}}_{\psi}}(V))} \right\}. \tag{6.5}$$

6.3 Conditions

Some special conditions are needed to prove identifiability, consistency and asymptotic normality of the parameter and the estimates, respectively.

Conditions.

A.1 $\mathrm{P}_0[C = 0] = 0$, $\mathrm{P}_0[C \ge \tau | \boldsymbol{Z}] = \mathrm{P}_0[C = \tau | \boldsymbol{Z}] > 0$ *almost surely, and censoring is independent of* T *given* $\boldsymbol{Z}$.

A.2 The total variation of $\boldsymbol{Z}_1(\cdot)$ *on* $[0, \tau]$ *is almost surely less than some* $m_0 < \infty$ *and* $\boldsymbol{Z}_2$ *is bounded almost surely.*

B.1 The vector $\boldsymbol{\xi}_0$ *lies in the interior of a compact set* $\Xi = [\xi_{11}, \xi_{21}] \times [\xi_{12}, \xi_{22}] \times \cdots \times [\xi_{1q}, \xi_{2q}]$ *with known* $\xi_{11}, \xi_{21}, \ldots, \xi_{1q}, \xi_{2q}$.

B.2 For some neighborhood $V(\boldsymbol{\xi}_0)$ *of* $\boldsymbol{\xi}_0$ *the density of* $\boldsymbol{Z}_2$, $f_{\boldsymbol{Z}_2}$, *exists and is strictly positive, bounded and continuous.*

B.3 For some $t_1 \in (0, \tau]$, $\mathrm{Var}[\boldsymbol{Z}_1(t_1) | \boldsymbol{Z}_2 = \boldsymbol{\xi}_0]$ *is positive definite.*

C.1 The parameter vector is given by $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_{10}^\top, \boldsymbol{\beta}_{20}^\top, \boldsymbol{\beta}_{30}^\top)^\top \in B = B_1 \times B_2 \times B_3 \subset \mathbb{R}^{p+2q}$, *where* B_1, B_2 *and* B_3 *are open, convex, bounded and known.*

C.2 At least one component of $\boldsymbol{\beta}_{30}$ *is unequal to* 0.

C.3 The function $A_0 \in \mathcal{A}$, *where* $\mathcal{A}$ *is the set of all increasing functions* $A : [0, \tau] \to [0, \infty)$ *with* $A(0) = 0$ *and* $A(\tau) < \infty$. *Moreover,* A_0 *has derivative* a_0 *satisfying* $0 < a_0(t) < \infty$ *for all* $t \in [0, \tau]$.

D.1 The function $G = -\log \Lambda : [0, \infty) \to [0, \infty)$ *is thrice continuously differentiable, with* $G(0) = 0$, *and for each* $u \in [0, \infty)$, $0 < G^{'}(u) < \infty$, $0 < \Lambda^{''}(u) < \infty$ *and* $\sup_{s \in [0,u]} |G^{'''}(s)| < \infty$.

D.2 For some $c_0 > 0$, *both* $\sup_{u \geq 0} |u^{c_0} \Lambda(u)| < \infty$ *and* $\sup_{u \geq 0} |u^{1+c_0} \Lambda'(u)| < \infty$.

The conditions A.1, A.2, C.1 and C.3 are general conditions required for the use of non-parametric maximum likelihood methods. They provide identifiability in right censored transformation models. Especially, condition A.1 expresses that there is no censoring after τ such that all functions only matter on the interval $[0, \tau]$. Moreover, we need the conditions B.1, B.2, B.3 and C.2 to show that the change-points are identifiable. The conditions D.1 and D.2 are used to establish asymptotic normality. All these conditions are similar to the ones stated in Kosorok & Song (2007).

6.4 Consistency

In this section we derive consistency of the finite and infinite-dimensional estimates. First of all we show that our parameters are identifiable, if the conditions of Section 6.3 are satisfied. Afterwards we prove that the NPMLE is bounded, i.e. $\limsup_{n \to \infty} \hat{A}_n(\tau) < \infty$. Consistency will then follow.

Lemma 6.1. *Under the regularity conditions A.1, B.2, B.3, C.2, C.3, D.1, the transformation model with a change-point is identifiable.*

Proof. Consider the Kullback-Leibler divergence $\int \log \frac{\mathrm{d}\,\mathrm{P}_{\boldsymbol{\theta}}}{\mathrm{d}\,\mathrm{P}_0} \mathrm{d}\,\mathrm{P}_0$ in Van der Vaart (1998). Since in general

$$\mathrm{P}_0 \left(\log(\,\mathrm{d}P_{\boldsymbol{\theta}}) - \log(\,\mathrm{d}P_0)\right) \leq -\int \left((\,\mathrm{d}P_{\boldsymbol{\theta}})^{1/2} - (\,\mathrm{d}P_0)^{1/2}\right)^2 \,\mathrm{d}\mu,$$

for a dominating measure μ, the Kullback-Leibler divergence is only zero if $\mathrm{d}\,\mathrm{P}_{\boldsymbol{\theta}} = \mathrm{d}\,\mathrm{P}_0$. Therefore, it is enough to show that for all $\boldsymbol{\theta} \in \Theta$ the equality $G(H^{\boldsymbol{\theta}}(t)) = G(H^{\boldsymbol{\theta}_0}(t))$

implies $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ almost surely. Hence, suppose that almost surely under P_0 for all $t \in [0, \tau]$

$$G\left(\int_0^t R(s)\exp\{\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}(s,\boldsymbol{\xi})\}\,\mathrm{d}A(s)\right) = G\left(\int_0^t R(s)\exp\{\boldsymbol{\beta}_0^\top \tilde{\boldsymbol{Z}}(s,\boldsymbol{\xi}_0)\}\,\mathrm{d}A_0(s)\right). \quad (6.6)$$

Since the conditions D.1 and A.1 hold, (6.6) implies

$$\int_0^t \exp\{\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}(s,\boldsymbol{\xi})\}\,\mathrm{d}A(s) = \int_0^t \exp\{\boldsymbol{\beta}_0^\top \tilde{\boldsymbol{Z}}(s,\boldsymbol{\xi}_0)\}\,\mathrm{d}A_0(s).$$

Taking the Radon-Nikodym derivative with respect to A_0 and the logarithm on both sides yields

$$\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}(t,\boldsymbol{\xi}) + \log(\frac{\mathrm{d}A}{\mathrm{d}A_0}) - \boldsymbol{\beta}_0^\top \tilde{\boldsymbol{Z}}(t,\boldsymbol{\xi}_0) = 0. \quad (6.7)$$

Choose a $\boldsymbol{z} \in V(\boldsymbol{\xi}_0)$ according to B.2 and consider the left-hand side of (6.7) conditioned on $\boldsymbol{Z}_2 = \boldsymbol{z}$. Calculating the variance yields

$$(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{10})^\top \operatorname{Var}[\boldsymbol{Z}_1(t)|\boldsymbol{Z}_2 = \boldsymbol{z}](\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{10}) = 0.$$

Since for some $t_1 \in (0, \tau]$ the variance $\operatorname{Var}[\boldsymbol{Z}_1(t_1)|\boldsymbol{Z}_2 = \boldsymbol{z}]$ is assumed to be positive definite because of condition B.3, the last equation yields $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_{10}$.
The assertion $\boldsymbol{\xi} = \boldsymbol{\xi}_0$ is proved by contradiction.
Assume that $\boldsymbol{\xi} > \boldsymbol{\xi}_0$ and consider the rest of equation (6.7) conditioned on $\boldsymbol{Z}_2 < \boldsymbol{\xi}_0 < \boldsymbol{\xi}$

$$\mathrm{P}\left((\boldsymbol{\beta}_{20} - \boldsymbol{\beta}_2)^\top \boldsymbol{Z}_2 = \log(\frac{\mathrm{d}A}{\mathrm{d}A_0})\middle|\boldsymbol{Z}_2 < \boldsymbol{\xi}_0 < \boldsymbol{\xi}\right) = 1.$$

Since $\operatorname{Var}[\boldsymbol{Z}_2|\boldsymbol{Z}_2 < \boldsymbol{\xi}_0 < \boldsymbol{\xi}]$ is positive definite in a neighborhood of $\boldsymbol{\xi}_0$ by B.2, the term $(\boldsymbol{\beta}_{20} - \boldsymbol{\beta}_2)^\top \boldsymbol{Z}_2$ can only be a constant if $\boldsymbol{\beta}_{20} = \boldsymbol{\beta}_2$. Now let $\boldsymbol{\xi}_0 < \boldsymbol{Z}_2 < \boldsymbol{\xi}$ and

$$\mathrm{P}\left((\boldsymbol{\beta}_{20} - \boldsymbol{\beta}_2)^\top \boldsymbol{Z}_2 + \boldsymbol{\beta}_{30}^\top(\boldsymbol{Z}_2 - \boldsymbol{\xi}_0) = \log(\frac{\mathrm{d}A}{\mathrm{d}A_0})\middle|\boldsymbol{\xi}_0 < \boldsymbol{Z}_2 < \boldsymbol{\xi}\right) = 1.$$

Since $\operatorname{Var}[\boldsymbol{Z}_2|\boldsymbol{\xi}_0 < \boldsymbol{Z}_2 < \boldsymbol{\xi}]$ is positive definite in a neighborhood of $\boldsymbol{\xi}_0$ because of B.2 and $\boldsymbol{\beta}_2 = \boldsymbol{\beta}_{20}$, the term $\boldsymbol{\beta}_{30}^\top(\boldsymbol{Z}_2 - \boldsymbol{\xi}_0)$ can only be a constant almost surely, if all elements of $\boldsymbol{\beta}_{30}$ are equal to zero. This is a contradiction to C.2.
The case $\boldsymbol{\xi} < \boldsymbol{\xi}_0$ can be shown analogously. Hence, $\boldsymbol{\xi} = \boldsymbol{\xi}_0$.
Using the obtained results $\boldsymbol{\xi} = \boldsymbol{\xi}_0$, $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_{10}$ and $\boldsymbol{\beta}_2 = \boldsymbol{\beta}_{20}$ and conditioning the left-hand side of (6.7) on $\boldsymbol{Z}_2 > \boldsymbol{\xi}_0$ yields

$$\mathrm{P}\left((\boldsymbol{\beta}_3 - \boldsymbol{\beta}_{30})^\top(\boldsymbol{Z}_2 - \boldsymbol{\xi}_0) = \log\left(\frac{\mathrm{d}A}{\mathrm{d}A_0}\right)\middle|\boldsymbol{Z}_2 > \boldsymbol{\xi}_0\right) = 1.$$

Since $\text{Var}[\boldsymbol{Z}_2|\boldsymbol{Z}_2 > \boldsymbol{\xi}_0]$ is positive definite, the last equation implies $\boldsymbol{\beta}_3 = \boldsymbol{\beta}_{30}$ and hence $A(t) = A_0(t)$. □

Lemma 6.2. *Under the regularity conditions A.2, D.1 and D.2, $\hat{A}_n = \hat{A}_{\boldsymbol{\psi}}(\tau)$ is asymptotically bounded, i.e.*

$$\limsup_{n\to\infty} \hat{A}_n(\tau) < \infty \quad \textit{almost surely.}$$

Proof. Let $(\Omega^\infty, \mathcal{F}^\infty, \mathrm{P}_0^\infty)$ be the probability space for infinite sequences of observations and let $W \subset \Omega^\infty$ be a set of inner probability 1 for which $\mathbb{P}_n N(t) \to \mathrm{P}_0\, N(t)$ uniformly in t. The conclusion of this lemma is shown by contradiction.
Assume that

$$\limsup_{n\to\infty} \hat{A}_n(\tau) = \infty \tag{6.8}$$

with positive probability. Define $\boldsymbol{\theta}_n = (\boldsymbol{\xi}_0, \boldsymbol{\beta}_0, A_n) = (\boldsymbol{\psi}_0, A_n)$ with A_n chosen as $A_n = \mathbb{P}_n N(t)$. If $\hat{\boldsymbol{\theta}}_n = (\boldsymbol{\psi}_0, \hat{A}_n)$ maximizes the likelihood then the difference $L_n(\hat{\boldsymbol{\theta}}_n) - L_n(\boldsymbol{\theta}_n)$ should be nonnegative. Our goal is to show that under assumption (6.8) this is not the case. We can find a subsequence $\{n_k\}$, such that $\lim_{k\to\infty} \hat{A}_{n_k}(\tau) = \infty$ for some fixed $\omega \in W$ due to (6.8) and we will prove that the likelihood difference $L_{n_k}(\hat{\boldsymbol{\theta}}_{n_k}) - L_{n_k}(\boldsymbol{\theta}_{n_k})$ diverges to negative infinity as n_k tends to infinity. This yields the intended contradiction since $\hat{\boldsymbol{\theta}}_{n_k}$ should maximize $L_{n_k}(\boldsymbol{\theta})$.
For the subsequence described above consider

$$\begin{aligned} &L_{n_k}(\hat{\boldsymbol{\theta}}_{n_k}) - L_{n_k}(\boldsymbol{\theta}_{n_k}) \\ &\le O(1) + \mathbb{P}_{n_k}\left[\delta \log\left(n_k\, \Delta\hat{A}_{n_k}(V)\right)\right] + \mathbb{P}_{n_k}\left[\delta \log\left(G'(H^{\hat{\boldsymbol{\theta}}_{n_k}}(V))\right)\right] - \mathbb{P}_{n_k}\left[G(H^{\hat{\boldsymbol{\theta}}_{n_k}}(V))\right], \end{aligned} \tag{6.9}$$

since $\mathbb{P}_{n_k}\left[\delta \log(n_k \Delta A_{n_k}(V))\right] = 0$ and $\mathbb{P}_{n_k}\left[\delta \log\left(G'(H^{\boldsymbol{\theta}_{n_k}}(V))\right)\right] - \mathbb{P}_{n_k}\left[G(H^{\boldsymbol{\theta}_{n_k}}(V))\right] = O(1)$ using the fact, that $\boldsymbol{\psi}_0$ is bounded and $\mathbb{P}_{n_k} N(t)$ converges uniformly to $\mathrm{P}_0\, N(t)$. For the last two terms in the above inequality we know by condition D.2 that for all $u > 0$ and some $c_0 > 0$,

$$\log(-\Lambda'(u)) = \log(-u^{1+c_0}\, \Lambda'(u)) - (1 + c_0) \log(u) \le O(1) - (1 + c_0) \log(u)$$

and

$$\log(\Lambda(u)) = \log(u^{c_0} \Lambda(u)) - c_0 \log(u) \le O(1) - c_0 \log(u).$$

Hence,

$$
\begin{aligned}
&\mathbb{P}_{n_k}\left[\delta \log\left(G'(H^{\hat{\boldsymbol{\theta}}_{n_k}}(V))\right)\right] - \mathbb{P}_{n_k}\left[G(H^{\hat{\boldsymbol{\theta}}_{n_k}}(V))\right] \\
&\qquad = \mathbb{P}_{n_k}\left[\delta \log\left(-\Lambda'(H^{\hat{\boldsymbol{\theta}}_{n_k}}(V))\right) - \delta \log\left(\Lambda(H^{\hat{\boldsymbol{\theta}}_{n_k}}(V))\right)\right] \\
&\qquad\quad - \mathbb{P}_{n_k}\left[-\log\left(\Lambda(H^{\hat{\boldsymbol{\theta}}_{n_k}}(V))\right)\right] \\
&\qquad = \mathbb{P}_{n_k}\left[\delta \log\left(-\Lambda'(H^{\hat{\boldsymbol{\theta}}_{n_k}}(V))\right)\right] - \mathbb{P}_{n_k}\left[(1-\delta)(-\log \Lambda(H^{\hat{\boldsymbol{\theta}}_{n_k}})(V))\right] \\
&\qquad \le O(1) - \mathbb{P}_{n_k}\left[(\delta + c_0)\log(H^{\hat{\boldsymbol{\theta}}_{n_k}}(V))\right] \\
&\qquad \le O(1) - \mathbb{P}_{n_k}\left[(\delta + c_0)\log(\hat{A}_{n_k}(V))\right].
\end{aligned}
\tag{6.10}
$$

The last inequality is valid, since

$$
\begin{aligned}
\log(H^{\hat{\boldsymbol{\theta}}_{n_k}}(V)) &= \log\left(\int_0^V \exp\{\boldsymbol{\beta}_0^\top \tilde{\boldsymbol{Z}}(s, \boldsymbol{\xi}_0)\}\, \mathrm{d}\hat{A}_{n_k}(s)\right) \\
&\ge \log\left(\int_0^V \exp\{-K_0\}\, \mathrm{d}\hat{A}_{n_k}(s)\right) \\
&\ge -K_0 + \log(\hat{A}_{n_k}(V)),
\end{aligned}
$$

where $\exp\{-K_0\}$ is a lower bound of $\exp\{\boldsymbol{\beta}_0^\top \tilde{\boldsymbol{Z}}(t, \boldsymbol{\xi}_0)\}$ by condition A.2.
Next consider the second term of (6.9).
Choose a partition of $[0, \tau]$, $0 = u_0 < u_1 < \cdots < u_J = \tau$ for some finite J and let $N^j(s) = N(s) I_{\{V \in [u_{j-1}, u_j]\}}$, $j = 1, \ldots, J$. Then, $\mathbb{P}_{n_k}\left[\int_0^\tau \log\left(n_k\, \Delta \hat{A}_{n_k}(s)\right)\, \mathrm{d}N(s)\right]$ can be split in the following way

$$
\mathbb{P}_{n_k}\left\{\int_0^\tau \log(n_k \Delta \hat{A}_{n_k}(s))\, \mathrm{d}N^J(s)\right\} + \sum_{j=1}^{J-1} \mathbb{P}_{n_k}\left\{\int_0^\tau \log(n_k \Delta \hat{A}_{n_k}(s))\, \mathrm{d}N^j(s)\right\}.
$$

The $j-$th term is equal to

$$
\mathbb{P}_{n_k} N^j(\tau) \left(\int_0^\tau \log(n_k \Delta \hat{A}_{n_k}(s))\, \mathrm{d}\mathbb{P}_{n_k} N^j(s)\right) / \mathbb{P}_{n_k} N^j(\tau)
$$

Using Jensen's inequality it is dominated by

$$
\begin{aligned}
&\leq \mathbb{P}_{n_k} N^j(\tau) \log \left(\int_0^{u_j} (n_k \Delta \hat{A}_{n_k}(s)) \, \mathrm{d}\mathbb{P}_{n_k} N^j(s) / \mathbb{P}_{n_k} N^j(\tau) \right) \\
&= \mathbb{P}_{n_k} N^j(\tau) \left(\log([\mathbb{P}_{n_k} N^j(\tau)]^{-1}) + \log \left(\int_0^{u_j} (n_k \Delta \hat{A}_{n_k}(s)) \, \mathrm{d}\mathbb{P}_{n_k} N^j(s) \right) \right) \\
&= \mathbb{P}_{n_k} N^j(\tau) \log([\mathbb{P}_{n_k} N^j(\tau)]^{-1}) + \mathbb{P}_{n_k} N^j(\tau) \cdot \log \left(\int_0^{u_j} (n_k \Delta \hat{A}_{n_k}(s)) \, \mathrm{d}\mathbb{P}_{n_k} N^j(s) \right) \\
&\leq O(1) + \mathbb{P}_{n_k} N^j(\tau) \cdot \log \left(\int_0^{u_j} n_k \Delta \hat{A}_{n_k}(s) \, \mathrm{d} \frac{1}{n_k} \sum_{i=1}^{n_k} N_i^j(s) \right) \\
&= O(1) + \mathbb{P}_{n_k} \delta I_{\{V \in [u_{j-1}, u_j]\}} \cdot \log \left(\int_0^{u_j} \Delta \hat{A}_{n_k}(s) \, \mathrm{d} \sum_{i=1}^{n_k} N_i^j(s) \right) \\
&\leq O(1) + \mathbb{P}_{n_k} \delta I_{\{V \in [u_{j-1}, u_j]\}} \cdot \log(\hat{A}_{n_k}(u_j)).
\end{aligned} \tag{6.11}
$$

The last inequality holds, since $\hat{A}_{n_k}$ jumps at the same time as N. Hence, (6.9) is upper bounded by

$$
\begin{aligned}
O(1) &+ \log(\hat{A}_{n_k}(\tau)) \mathbb{P}_{n_k} \left[\delta I_{\{V \in [u_{J-1}, u_J]\}} \right] - \log(\hat{A}_{n_k}(\tau)) \mathbb{P}_{n_k} \left[(\delta + c_0) I_{\{V \in [u_J, \infty]\}} \right] \\
&+ \sum_{j=1}^{J-1} \left\{ \log(\hat{A}_{n_k}(u_j)) \mathbb{P}_{n_k} \left[\delta I_{\{V \in [u_{j-1}, u_j]\}} \right] - \log(\hat{A}_{n_k}(u_j)) \mathbb{P}_{n_k} \left[(\delta + c_0) I_{\{V \in [u_j, u_{j+1}]\}} \right] \right\}.
\end{aligned} \tag{6.12}
$$

For a fixed constant $c > 1$, ($\frac{c_0}{c} < c_0$) we can choose the partition such that

$$
\mathrm{P}_0 \left[N(\tau) I_{\{V \in [u_{J-1}, u_J]\}} \right] = \mathrm{P}_0 \left[\left(N(\tau) + \frac{c_0}{c} \right) I_{\{V \in [u_J, \infty]\}} \right]
$$

and for $j = 1, \ldots, J-1$

$$
\mathrm{P}_0 \left[N(\tau) I_{\{V \in [u_{j-1}, u_j]\}} \right] = \mathrm{P}_0 \left[\left(N(\tau) + \frac{c_0}{c} \right) I_{\{V \in [u_j, u_{j+1}]\}} \right].
$$

With this choice it is not hard to see that the summands of (6.12) tend to negative infinity, since $\mathbb{P}_{n_k} N(t) \to \mathrm{P}_0 \, N(t)$ uniformly, $\hat{A}_{n_k} \in \mathcal{A}$ and because of assumption (6.8). Hence, we obtain that (6.12) tends to negative infinity as $n_k \to \infty$. This is the desired contradiction, which shows that $\limsup_{n \to \infty} \hat{A}_n(\tau) < \infty$ almost surely. □

The following lemma shows that the class of functions

$$
\mathcal{F}_{(k)} = \{ W(t; \boldsymbol{\theta}) : t \in [0, \tau], \boldsymbol{\xi} \in \Xi, \boldsymbol{\beta} \in B, A \in \mathcal{A}_{(k)} \}
$$

with $W(t;\boldsymbol{\theta})$ given in (6.5) and $\mathcal{A}_{(k)} = \{A \in \mathcal{A} : A(\tau) \leq k\}$, is P_0-Donsker. This fact will be needed in the proof of consistency.

Lemma 6.3. *Assume the regularity conditions A.2 and D.1. The class $\mathcal{F}_{(k)}$, is P_0-Donsker $\forall\, k < \infty$.*

Proof. Consider $H^{\boldsymbol{\theta}}(s) = \int_0^s \exp\{\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}(u,\boldsymbol{\xi})\}\,\mathrm{d}A(u)$. The classes $\{\boldsymbol{\beta}_1^\top \boldsymbol{Z}_1(t) : \boldsymbol{\beta}_1 \in B_1, t \in [0,\tau]\}$, $\{\boldsymbol{\beta}_2^\top \boldsymbol{Z}_2 : \boldsymbol{\beta}_2 \in B_2\}$ and $\{\boldsymbol{\beta}_3^\top(\boldsymbol{Z}_2 - \boldsymbol{\xi})^+ : \boldsymbol{\beta}_3 \in B_3, \boldsymbol{\xi} \in \Xi\}$ are Donsker classes. Since $\exp(\cdot)$ is Lipschitz on compacts and a sum of Donsker classes is Donsker, the class $\{\exp\{\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}(t,\boldsymbol{\xi})\} : \boldsymbol{\beta} \in B, \boldsymbol{\xi} \in \Xi, t \in [0,\tau]\}$ is Donsker. Consider the map

$$h \in D[0,\tau] \mapsto \left\{ \int_0^t h(s)\,\mathrm{d}A(s) :\, t \in [0,\tau], A \in \mathcal{A} \right\} \in l^\infty([0,\tau]) \times \mathcal{A}$$

where $l^\infty([0,\tau])$ denotes the set of bounded functions on $[0,\tau]$. Note that this map is uniformly equicontinuous and linear. Thus the class

$$\mathcal{F} = \left\{ \int_0^t \exp\{\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}(u,\boldsymbol{\xi})\}\,\mathrm{d}A(u) :\, t \in [0,\tau], \boldsymbol{\beta} \in B, \boldsymbol{\xi} \in \Xi, A \in \mathcal{A} \right\}$$

is Donsker by the continuous mapping theorem. Now condition D.1 ensures that G' and $\frac{G''}{G}$ are Lipschitz on compacts. The sum of Donsker classes is Donsker and the product of bounded Donsker classes is Donsker. $\square$

Theorem 6.2. *Under the regularity conditions of Section 6.3, $\hat{\boldsymbol{\theta}}_n$ converges outer almost surely to $\boldsymbol{\theta}_0$.*

Proof. Note that almost sure convergence of $\hat{\boldsymbol{\theta}}_n$ is equivalent to outer almost sure convergence in our setup, since the uniform distance between $\hat{\boldsymbol{\theta}}_n$ and $\boldsymbol{\theta}_0$ is measurable.
By Lemma 6.3 we know that the class of functions $\mathcal{F}_{(k)}$ is Donsker and hence Glivenko-Cantelli for all $k < \infty$. Using the same arguments as in the lemma above one can also show that the classes $\{G(H^{\boldsymbol{\theta}}(V)) : \boldsymbol{\xi} \in \Xi, \boldsymbol{\beta} \in B, A \in \mathcal{A}_{(k)}\}$ and $\{G'(H^{\boldsymbol{\theta}}(s)) : \boldsymbol{\xi} \in \Xi, \boldsymbol{\beta} \in B, A \in \mathcal{A}_{(k)}, s \in [0,\tau]\}$ are Glivenko Cantelli for all $k < \infty$. Thus, with probability 1 the following two expressions $(\mathbb{P}_n - \mathrm{P}_0)\Big(W(\cdot;\hat{\boldsymbol{\theta}}_n)\Big)$ and $(\mathbb{P}_n - \mathrm{P}_0)\Big(G(H^{\hat{\boldsymbol{\theta}}_n}(V)) - G(H^{\boldsymbol{\theta}_n}(V))\Big)$ converge to zero. Furthermore, by Lemma 6.2 $\{\hat{A}_n(\tau)\}$ is asymptotically bounded and $\mathbb{P}_n N(t) \to \mathrm{P}_0\, N(t)$ uniformly in t.
For the rest of the proof we fix some ω for which the last asymptotics hold. By Helly's lemma (Theorem A.4) we can find a subsequence $\{\hat{A}_{n_k}\}$ with the property that $\hat{A}_{n_k}(t) \to A(t)$ at each continuity point $t \in [0,\tau]$ of some function A. From the construction of the estimator in (6.4) we know that $\hat{A}_{n_k}(t)$ jumps at the same time as $N(t)$. Since $\mathbb{P}_n N(t)$ converges to $\mathrm{P}_0\, N(t)$ uniformly and since $\mathrm{P}_0\, N(t)$ is continuous, $A(t)$ is continuous for all

$t \in [0, \tau]$. This convergence is uniform, since the sequence $\hat{A}_n$ is monotone, A is continuous and by the application of Dinis theorem, (see LeCam (1986)). Now assume without loss of generality that along this subsequence $\{\hat{A}_{n_k}\}$ there exists a sequence $\hat{\boldsymbol{\psi}}_{n_k}$ which converges to some $\boldsymbol{\psi} \in \Psi$. Moreover, let $\boldsymbol{\theta}_n = (\boldsymbol{\psi}_0, A_n)$, where

$$A_n(t) = \int_0^t \frac{1}{\mathrm{P}_0 W(s, \boldsymbol{\theta}_0)} \, \mathrm{d}\mathbb{P}_n N(s).$$

For all $t \in [0, \tau]$ derive

$$A_0(t) = \int_0^t \frac{1}{\mathrm{P}_0 W(s, \boldsymbol{\theta}_0)} \, \mathrm{d}\,\mathrm{P}_0\, N(s)$$

with the same technique we used to develop the estimator $\hat{A}_n(t)$ only with the true parameter $\boldsymbol{\theta}_0$. Hence, A_{n_k} converges uniformly to A_0 , since $\mathbb{P}_{n_k} N(s)$ converges uniformly to $\mathrm{P}_0\, N(s)$ as $k \to \infty$.
Since $\hat{\boldsymbol{\theta}}_{n_k}$ maximizes $L_{n_k}(\boldsymbol{\theta})$, we get

$$\begin{aligned}
0 &\leq L_{n_k}(\hat{\boldsymbol{\theta}}_{n_k}) - L_{n_k}(\boldsymbol{\theta}_{n_k}) \\
&= \mathbb{P}_{n_k} \int_0^\tau \log \left(\frac{\mathrm{P}_0 W(s, \boldsymbol{\theta}_0)}{\mathbb{P}_{n_k} W(s, \hat{\boldsymbol{\theta}}_{n_k})} \right) \mathrm{d}N(s) \\
&\quad + \mathbb{P}_{n_k} \left[\int_0^\tau \log \left(G'(H^{\hat{\boldsymbol{\theta}}_{n_k}}(s)) \right) + \hat{\boldsymbol{\beta}}_{n_k}^\top \tilde{\boldsymbol{Z}}(s, \hat{\boldsymbol{\xi}}_{n_k}) \, \mathrm{d}N(s) \right] \\
&\quad - \mathbb{P}_{n_k} \left[\int_0^\tau \log \left(G'(H^{\boldsymbol{\theta}_{n_k}}(s)) \right) + \boldsymbol{\beta}_0^\top \tilde{\boldsymbol{Z}}(s, \boldsymbol{\xi}_0) \, \mathrm{d}N(s) \right] \\
&\quad - \mathbb{P}_{n_k} \left[G(H^{\hat{\boldsymbol{\theta}}_{n_k}}(V)) - G(H^{\boldsymbol{\theta}_{n_k}}(V)) \right] \\
&\to \mathrm{P}_0 \int_0^\tau \log \left(\frac{\mathrm{d}A(s)}{\mathrm{d}A_0(s)} \right) \mathrm{d}N(s) + \mathrm{P}_0 \left[\int_0^\tau \log \left(G'(H^{\boldsymbol{\theta}}(s)) \right) + \boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}(s, \boldsymbol{\xi}) \, \mathrm{d}N(s) \right] \\
&\quad - \mathrm{P}_0 \left[\int_0^\tau \log \left(G'(H^{\boldsymbol{\theta}_0}(s)) \right) + \boldsymbol{\beta}_0^\top \tilde{\boldsymbol{Z}}(s, \boldsymbol{\xi}_0) \, \mathrm{d}N(s) \right] \\
&\quad - \mathrm{P}_0 \left[G(H^{\boldsymbol{\theta}}(s)) - G(H^{\boldsymbol{\theta}_0}(s)) \right] \\
&= \int \log \frac{\mathrm{d}\,\mathrm{P}_{\boldsymbol{\theta}}}{\mathrm{d}\,\mathrm{P}_0} \, \mathrm{d}\,\mathrm{P}_0 \\
&\leq 0.
\end{aligned}$$

The last inequality holds, since the Kullback-Leibler divergence is negative. Therefore, the last calculations force $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, since the parameters are identifiable as shown in Lemma 6.1. To summarize the results: Since we chose a sequence arbitrarily and found a subsequence $\{n_k\}$ such that $\hat{\boldsymbol{\theta}}_{n_k} \to \boldsymbol{\theta}_0$, it follows that all convergent subsequences of $\hat{\boldsymbol{\theta}}_n$ converge to $\boldsymbol{\theta}_0$. □

6.5 Local Behavior of the Limit Function

In this section we will consider the derivatives of the nonparametric likelihood, which are needed later to determine the rate of convergence and the asymptotic normality of the parameters. Three different problems arise when it comes to calculating the derivatives of the nonparametric likelihood with respect to the parameters. First of all we have to deal with the same problem as in the previous chapters. The empirical likelihood is not differentiable in the parameter vector $\boldsymbol{\xi}$, such that we have to consider the limit function of the empirical likelihood. This brings up the second problem. The limit function as well as the empirical likelihood evaluated at the true parameter A_0 are negative infinity, since A_0 is continuous and hence ΔA_0 is zero. The third problem is given by the infinite-dimensional parameter included in the nonparametric likelihood function.

One way to solve the second problem is a reparametrization of the estimator $\hat{A}_n$. We use the reparametrization $\Gamma(\cdot) \mapsto A_n^\Gamma(\cdot) = \int_0^{(\cdot)} \exp\{-\Gamma(s)\} \,\mathrm{d}\mathbb{P}_n N(s)$ as suggested in Kosorok & Song (2007). This reparametrization yields the same NPMLE as before by maximizing the process $L_n(\boldsymbol{\xi}, \boldsymbol{\beta}, \Gamma)$ over the parameter $\boldsymbol{\vartheta} = (\boldsymbol{\xi}, \boldsymbol{\beta}, \Gamma)$, where

$$L_n(\boldsymbol{\vartheta}) = \mathbb{P}_n \left[\int_0^\tau \left[-\Gamma(s) + \boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}(s, \boldsymbol{\xi}) + \log\left(G'(H^{\boldsymbol{\theta}^{(n)}}(s)) \right) \right] \mathrm{d}N(s) - G(H^{\boldsymbol{\theta}^{(n)}}(V)) \right]$$

and $\boldsymbol{\theta}^{(n)} = (\boldsymbol{\xi}, \boldsymbol{\beta}, A_n^\Gamma)$.

The parameter Γ is estimated by $\hat{\Gamma}_n(\cdot) = \log(\mathbb{P}_n W(\cdot; \hat{\boldsymbol{\theta}}_n))$, where $W(\cdot; \hat{\boldsymbol{\theta}}_n)$ is given in (6.5). For the construction of the limit function define $\Gamma_0(\cdot) = \log(\mathrm{P}_0\, W(\cdot; \boldsymbol{\theta}_0))$. The limit function of the empirical likelihood has the form

$$L(\boldsymbol{\vartheta}) = \mathrm{P}_0 \left[\int_0^\tau \left[-\Gamma(s) + \boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}(s, \boldsymbol{\xi}) + \log\left(G'(H^{\boldsymbol{\theta}^{(0)}}(s)) \right) \right] \mathrm{d}N(s) - G(H^{\boldsymbol{\theta}^{(0)}}(V)) \right],$$

where $\boldsymbol{\theta}^{(0)} = (\boldsymbol{\xi}, \boldsymbol{\beta}, A_0^\Gamma)$ with $A_0^\Gamma(s) = \int_0^s \exp\{-\Gamma(u)\} \,\mathrm{d}\,\mathrm{P}_0\, N(u)$ and the function $H^{\boldsymbol{\theta}^{(0)}}(t) = \int_0^t R(s) \exp\{\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}(s, \boldsymbol{\xi})\} \,\mathrm{d}A_0^\Gamma(s)$.

The third problem can be handled by considering the setup in a different way. Let $D[0, \tau]$ denote the space of all cadlag functions on $[0, \tau]$ and consider the modified space $\bar{\Theta} = \Xi \times B \times BV$ with elements $(\boldsymbol{\xi}, \boldsymbol{\beta}, \Gamma)$ instead of the parameter space Θ, where the space BV is a subspace of $D[0, \tau]$ containing all functions that are of bounded variation on the interval $[0, \tau]$. The space $\mathcal{H}$ consists of elements $\boldsymbol{h} = (\boldsymbol{h}_1, \boldsymbol{h}_2, \boldsymbol{h}_3, \boldsymbol{h}_4, h_5)$ with $\boldsymbol{h}_1 \in \mathbb{R}^q, \boldsymbol{h}_2 \in \mathbb{R}^p, \boldsymbol{h}_3 \in \mathbb{R}^q, \boldsymbol{h}_4 \in \mathbb{R}^q, h_5 \in BV$ and we define the norm

$$\rho(\boldsymbol{h}) = \left(\|\boldsymbol{h}_1\|^2 + \|\boldsymbol{h}_2\|^2 + \|\boldsymbol{h}_3\|^2 + \|\boldsymbol{h}_4\|^2 + \|h_5\|_\infty^2 \right)^{1/2}$$

where $\|\cdot\|_\infty$ is the uniform norm. Furthermore, the set $\mathcal{H}_r$ is defined by $\mathcal{H}_r = \{\boldsymbol{h} \in \mathcal{H} : \rho(\boldsymbol{h}) \leq r, r \in (0,\infty)\}$ and $\mathcal{H}_\infty = \{\boldsymbol{h} \in \mathcal{H}_r : r < \tilde{r}\}$ for some sufficiently large $\tilde{r} < \infty$. In this way the parameter $\boldsymbol{\vartheta} \in \bar{\Theta}$ can be viewed as a linear functional on $\mathcal{H}_\infty$ by defining $\boldsymbol{\vartheta}(\boldsymbol{h}) = \boldsymbol{h}_1^\top \boldsymbol{\xi} + \boldsymbol{h}_2^\top \boldsymbol{\beta}_1 + \boldsymbol{h}_3^\top \boldsymbol{\beta}_2 + \boldsymbol{h}_4^\top \boldsymbol{\beta}_3 + \int_0^\tau h_5(u)\,\mathrm{d}A_0^\Gamma(u)$. Hence, elements of $\bar{\Theta}$ can be identified as elements of $l^\infty(\mathcal{H}_\infty)$ with uniform norm $\|\boldsymbol{\vartheta}\|_\infty = \sup_{\boldsymbol{h}\in\mathcal{H}_\infty} |\boldsymbol{\vartheta}(\boldsymbol{h})|$, where $l^\infty(\mathcal{H}_\infty)$ is the set of all bounded functionals on $\mathcal{H}_\infty$. Note that $\mathcal{H}_1$ is rich enough to extract all components of $\boldsymbol{\vartheta}$. This is easy to see for the Euclidean parameters. For the infinite-dimensional parameter choose $\{\boldsymbol{h} : \boldsymbol{h}_1 = 0, \boldsymbol{h}_2 = 0, \boldsymbol{h}_3 = 0, \boldsymbol{h}_4 = 0, h_5 = I_{\{u<t\}}, t \in [0,\tau]\} \subset \mathcal{H}_1$. Thus we will use the concept of operators.

6.5.1 The Score Operator

Before we study the score operator note that for any $g \in BV$

$$\left.\frac{\partial A_0^{(\Gamma+tg)}(\cdot)}{\partial t}\right|_{t=0} = -\int_0^{(\cdot)} g(s)\,\mathrm{d}A_0^\Gamma(s).$$

The one-dimensional submodel $t \mapsto \boldsymbol{\vartheta}_t = \boldsymbol{\vartheta} + t(\boldsymbol{h}_1, \boldsymbol{h}_2, \boldsymbol{h}_3, \boldsymbol{h}_4, \int_0^{(\cdot)} h_5(u)\,\mathrm{d}A_0^\Gamma(u))$, $\boldsymbol{h} \in \mathcal{H}_\infty$ is needed for the calculation of the score operator. Using the abbreviation

$$\tilde{R}^{\boldsymbol{\theta}^{(0)}} = G'(H^{\boldsymbol{\theta}^{(0)}}(V)) - \delta \frac{G''(H^{\boldsymbol{\theta}^{(0)}}(V))}{G'(H^{\boldsymbol{\theta}^{(0)}}(V))}$$

the score operator is given by

$$\left.\frac{\partial L(\boldsymbol{\vartheta}_t)}{\partial t}\right|_{t=0} = \mathrm{P}_0\, U^\tau(\boldsymbol{\vartheta})(\boldsymbol{h}),$$

where $U^\tau(\boldsymbol{\vartheta})(\boldsymbol{h}) = U_1^\tau(\boldsymbol{\vartheta})(\boldsymbol{h}_1) + U_2^\tau(\boldsymbol{\vartheta})(\boldsymbol{h}_2) + U_3^\tau(\boldsymbol{\vartheta})(\boldsymbol{h}_3) + U_4^\tau(\boldsymbol{\vartheta})(\boldsymbol{h}_4) + U_5^\tau(\boldsymbol{\vartheta})(h_5)$, and

$$\begin{aligned}
U_1^\tau(\boldsymbol{\vartheta})(\boldsymbol{h}_1) &= \int_0^\tau (-\boldsymbol{\beta}_3 I_{\{\boldsymbol{Z}_2>\boldsymbol{\xi}\}})^\top \boldsymbol{h}_1\,\mathrm{d}N(s) \\
&\qquad - \tilde{R}^{\boldsymbol{\theta}^{(0)}} \int_0^\tau R(s)(-\boldsymbol{\beta}_3 I_{\{\boldsymbol{Z}_2>\boldsymbol{\xi}\}})^\top \boldsymbol{h}_1 \exp\{\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}(s,\boldsymbol{\xi})\}\,\mathrm{d}A_0^\Gamma(s) \\
U_2^\tau(\boldsymbol{\vartheta})(\boldsymbol{h}_2) &= \int_0^\tau \boldsymbol{Z}_1^\top(s)\boldsymbol{h}_2\,\mathrm{d}N(s) - \tilde{R}^{\boldsymbol{\theta}^{(0)}} \int_0^\tau R(s)\boldsymbol{Z}_1^\top(s)\boldsymbol{h}_2 \exp\{\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}(s,\boldsymbol{\xi})\}\,\mathrm{d}A_0^\Gamma(s) \\
U_3^\tau(\boldsymbol{\vartheta})(\boldsymbol{h}_3) &= \int_0^\tau \boldsymbol{Z}_2^\top \boldsymbol{h}_3\,\mathrm{d}N(s) - \tilde{R}^{\boldsymbol{\theta}^{(0)}} \int_0^\tau R(s)\boldsymbol{Z}_2^\top \boldsymbol{h}_3 \exp\{\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}(s,\boldsymbol{\xi})\}\,\mathrm{d}A_0^\Gamma(s) \\
U_4^\tau(\boldsymbol{\vartheta})(\boldsymbol{h}_4) &= \int_0^\tau \left((\boldsymbol{Z}_2 - \boldsymbol{\xi})^+\right)^\top \boldsymbol{h}_4\,\mathrm{d}N(s) \\
&\qquad - \tilde{R}^{\boldsymbol{\theta}^{(0)}} \int_0^\tau R(s)\left((\boldsymbol{Z}_2 - \boldsymbol{\xi})^+\right)^\top \boldsymbol{h}_4 \exp\{\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}(s,\boldsymbol{\xi})\}\,\mathrm{d}A_0^\Gamma(s)
\end{aligned}$$

$$U_5^\tau(\boldsymbol{\vartheta})(h_5) = -\int_0^\tau h_5(s)\,\mathrm{d}N(s) + \tilde{R}^{\boldsymbol{\theta}^{(0)}} \int_0^\tau R(s) h_5(s) \exp\{\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}(s,\boldsymbol{\xi})\}\,\mathrm{d}A_0^\Gamma(s).$$

6.5.2 The Information Operator

To prove weak convergence the score operator has to be Fréchet differentiable, since we want to use Theorem A.10 given in the Appendix A. Therefore, we calculate the Gateaux derivative of the score operator first and then we show that this derivative can be strengthened to a Fréchet derivative. The Gateaux derivative at $\boldsymbol{\vartheta} \in \Theta$ has the following form

$$-\frac{\partial}{\partial t} \mathrm{P}_0\, U^\tau(\boldsymbol{\vartheta} + t\boldsymbol{\vartheta}_1)(\boldsymbol{h})\Big|_{t=0} = \boldsymbol{\vartheta}_1(\sigma_{\boldsymbol{\theta}^{(0)}}(\boldsymbol{h})), \quad \text{for every } \boldsymbol{h} \in \mathcal{H}_\infty,$$

where $\boldsymbol{\vartheta}_1 \in \bar{\Theta}$ and $\sigma_{\boldsymbol{\theta}^{(0)}} : \mathcal{H}_\infty \to \mathcal{H}_\infty$ represents the information operator. Hence, the derivative can be interpreted as a linear functional. Using

$$\tilde{R}_1^{\boldsymbol{\theta}^{(0)}} = G''(H^{\boldsymbol{\theta}^{(0)}}(V)) - \delta\left(\frac{G'''(H^{\boldsymbol{\theta}^{(0)}}(V))G'(H^{\boldsymbol{\theta}^{(0)}}(V)) - G''(H^{\boldsymbol{\theta}^{(0)}}(V))^2}{G'(H^{\boldsymbol{\theta}^{(0)}}(V))^2}\right)$$

the information operator is the 5×5 matrix $\sigma_{\boldsymbol{\theta}^{(0)}}(\boldsymbol{h}) = \tilde{\boldsymbol{V}}(\boldsymbol{h})$, where

$$\begin{aligned}
\tilde{V}_{\boldsymbol{\xi},1}(\boldsymbol{h}_1) = &- \mathrm{P}_0\left[\int_0^\tau (-\boldsymbol{\beta}_3)^\top \boldsymbol{h}_1\,\mathrm{d}N(s) | \boldsymbol{Z}_2 = \boldsymbol{\xi}\right] f_{\boldsymbol{Z}_2}(\boldsymbol{\xi}) \\
&+ \mathrm{P}_0\left[\tilde{R}^{\boldsymbol{\theta}^{(0)}} \int_0^\tau R(s)(-\boldsymbol{\beta}_3)^\top \boldsymbol{h}_1 \exp\{\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}(s,\boldsymbol{\xi})\}\,\mathrm{d}A_0^\Gamma(s) | \boldsymbol{Z}_2 = \boldsymbol{\xi}\right] f_{\boldsymbol{Z}_2}(\boldsymbol{\xi}) \\
&+ \mathrm{P}_0\Bigg[\tilde{R}_1^{\boldsymbol{\theta}^{(0)}} \int_0^\tau R(s)(-\boldsymbol{\beta}_3 I_{\{\boldsymbol{Z}_2 > \boldsymbol{\xi}\}}) \exp\{\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}(s,\boldsymbol{\xi})\}\,\mathrm{d}A_0^\Gamma(s) \\
&\times \int_0^\tau R(s)(-\boldsymbol{\beta}_3)^\top \boldsymbol{h}_1 \exp\{\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}(s,\boldsymbol{\xi})\}\,\mathrm{d}A_0^\Gamma(s) \\
&+ \tilde{R}^{\boldsymbol{\theta}^{(0)}} \int_0^\tau R(s)(-\boldsymbol{\beta}_3)^\top \boldsymbol{h}_1 (-\boldsymbol{\beta}_3 I_{\{\boldsymbol{Z}_2 > \boldsymbol{\xi}\}}) \exp\{\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}(s,\boldsymbol{\xi})\}\,\mathrm{d}A_0^\Gamma(s)\Bigg] \\
\tilde{V}_{\boldsymbol{\xi},2}(\boldsymbol{h}_2) = &\,\mathrm{P}_0\left[\tilde{R}^{\boldsymbol{\theta}^{(0)}} \int_0^\tau R(s)\boldsymbol{Z}_1^\top(s)\boldsymbol{h}_2 \exp\{\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}(s,\boldsymbol{\xi})\}\,\mathrm{d}A_0^\Gamma(s) | \boldsymbol{Z}_2 = \boldsymbol{\xi}\right] f_{\boldsymbol{Z}_2}(\boldsymbol{\xi}) \\
&+ \mathrm{P}_0\Bigg[\tilde{R}_1^{\boldsymbol{\theta}^{(0)}} \int_0^\tau R(s)(-\boldsymbol{\beta}_3 I_{\{\boldsymbol{Z}_2 > \boldsymbol{\xi}\}}) \exp\{\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}(s,\boldsymbol{\xi})\}\,\mathrm{d}A_0^\Gamma(s) \\
&\times \int_0^\tau R(s)\boldsymbol{Z}_1^\top(s)\boldsymbol{h}_2 \exp\{\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}(s,\boldsymbol{\xi})\}\,\mathrm{d}A_0^\Gamma(s) \\
&+ \tilde{R}^{\boldsymbol{\theta}^{(0)}} \int_0^\tau R(s)(-\boldsymbol{\beta}_3 I_{\{\boldsymbol{Z}_2 > \boldsymbol{\xi}\}})\boldsymbol{Z}_1^\top(s)\boldsymbol{h}_2 \exp\{\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}(s,\boldsymbol{\xi})\}\,\mathrm{d}A_0^\Gamma(s)\Bigg]
\end{aligned}$$

$$
\begin{aligned}
\tilde{V}_{\boldsymbol{\xi},3}(\boldsymbol{h}_3) =& \mathrm{P}_0\left[\tilde{R}^{\boldsymbol{\theta}^{(0)}}\int_0^\tau R(s)\boldsymbol{Z}_2^\top\boldsymbol{h}_3\exp\{\boldsymbol{\beta}^\top\tilde{\boldsymbol{Z}}(s,\boldsymbol{\xi})\}\,\mathrm{d}A_0^\Gamma(s)|\boldsymbol{Z}_2=\boldsymbol{\xi}\right]f_{\boldsymbol{Z}_2}(\boldsymbol{\xi})\\
&+\mathrm{P}_0\left[\tilde{R}_1^{\boldsymbol{\theta}^{(0)}}\int_0^\tau R(s)(-\boldsymbol{\beta}_3 I_{\{\boldsymbol{Z}_2>\boldsymbol{\xi}\}})\exp\{\boldsymbol{\beta}^\top\tilde{\boldsymbol{Z}}(s,\boldsymbol{\xi})\}\,\mathrm{d}A_0^\Gamma(s)\right.\\
&\times\int_0^\tau R(s)\boldsymbol{Z}_2^\top\boldsymbol{h}_3\exp\{\boldsymbol{\beta}^\top\tilde{\boldsymbol{Z}}(s,\boldsymbol{\xi})\}\,\mathrm{d}A_0^\Gamma(s)\\
&\left.+\tilde{R}^{\boldsymbol{\theta}^{(0)}}\int_0^\tau R(s)(-\boldsymbol{\beta}_3 I_{\{\boldsymbol{Z}_2>\boldsymbol{\xi}\}})\boldsymbol{Z}_2^\top\boldsymbol{h}_3\exp\{\boldsymbol{\beta}^\top\tilde{\boldsymbol{Z}}(s,\boldsymbol{\xi})\}\,\mathrm{d}A_0^\Gamma(s)\right]\\
\tilde{V}_{\boldsymbol{\xi},4}(\boldsymbol{h}_4) =& \mathrm{P}\left[\tilde{R}^{\boldsymbol{\theta}^{(0)}}\int_0^\tau R(s)\left((\boldsymbol{Z}_2-\boldsymbol{\xi})^+\right)^\top\boldsymbol{h}_4\exp\{\boldsymbol{\beta}^\top\tilde{\boldsymbol{Z}}(s,\boldsymbol{\xi})\}\,\mathrm{d}A_0^\Gamma(s)|\boldsymbol{Z}_2=\boldsymbol{\xi}\right]f_{\boldsymbol{Z}_2}(\boldsymbol{\xi})\\
&+\mathrm{P}_0\left[\tilde{R}_1^{\boldsymbol{\theta}^{(0)}}\int_0^\tau R(s)(-\boldsymbol{\beta}_3 I_{\{\boldsymbol{Z}_2>\boldsymbol{\xi}\}})\exp\{\boldsymbol{\beta}^\top\tilde{\boldsymbol{Z}}(s,\boldsymbol{\xi})\}\,\mathrm{d}A_0^\Gamma(s)\right.\\
&\times\int_0^\tau R(s)((\boldsymbol{Z}_2-\boldsymbol{\xi})^+)^\top\boldsymbol{h}_4\exp\{\boldsymbol{\beta}^\top\tilde{\boldsymbol{Z}}(s,\boldsymbol{\xi})\}\,\mathrm{d}A_0^\Gamma(s)\\
&\left.+\tilde{R}^{\boldsymbol{\theta}^{(0)}}\int_0^\tau R(s)(-\boldsymbol{\beta}_3 I_{\{\boldsymbol{Z}_2>\boldsymbol{\xi}\}})((\boldsymbol{Z}_2-\boldsymbol{\xi})^+)^\top\boldsymbol{h}_4\exp\{\boldsymbol{\beta}^\top\tilde{\boldsymbol{Z}}(s,\boldsymbol{\xi})\}\,\mathrm{d}A_0^\Gamma(s)\right]\\
\tilde{V}_{\boldsymbol{\xi},5}(h_5) =& \mathrm{P}_0\left[\tilde{R}^{\boldsymbol{\theta}^{(0)}}\int_0^\tau R(s)h_5(s)\exp\{\boldsymbol{\beta}^\top\tilde{\boldsymbol{Z}}(s,\boldsymbol{\xi})\}\,\mathrm{d}A_0^\Gamma(s)|\boldsymbol{Z}_2=\boldsymbol{\xi}\right]f_{\boldsymbol{Z}_2}(\boldsymbol{\xi})\\
&+\mathrm{P}_0\left[\tilde{R}_1^{\boldsymbol{\theta}^{(0)}}\int_0^\tau R(s)(-\boldsymbol{\beta}_3 I_{\{\boldsymbol{Z}_2>\boldsymbol{\xi}\}})\exp\{\boldsymbol{\beta}^\top\tilde{\boldsymbol{Z}}(s,\boldsymbol{\xi})\}\,\mathrm{d}A_0^\Gamma(s)\right.\\
&\times\int_0^\tau R(s)h_5(s)\exp\{\boldsymbol{\beta}^\top\tilde{\boldsymbol{Z}}(s,\boldsymbol{\xi})\}\,\mathrm{d}A_0^\Gamma(s)\\
&\left.+\tilde{R}^{\boldsymbol{\theta}^{(0)}}\int_0^\tau R(s)(-\boldsymbol{\beta}_3 I_{\{\boldsymbol{Z}_2>\boldsymbol{\xi}\}})h_5(s)\exp\{\boldsymbol{\beta}^\top\tilde{\boldsymbol{Z}}(s,\boldsymbol{\xi})\}\,\mathrm{d}A_0^\Gamma(s)\right]\\
\tilde{V}_{\boldsymbol{\beta}_1,1}(\boldsymbol{h}_1) =& \mathrm{P}_0\left[\tilde{R}_1^{\boldsymbol{\theta}^{(0)}}\int_0^\tau R(s)\boldsymbol{Z}_1(s)\exp\{\boldsymbol{\beta}^\top\tilde{\boldsymbol{Z}}(s,\boldsymbol{\xi})\}\,\mathrm{d}A_0^\Gamma(s)\right.\\
&\times\int_0^\tau R(s)(-\boldsymbol{\beta}_3 I_{\{\boldsymbol{Z}_2>\boldsymbol{\xi}\}})^\top\boldsymbol{h}_1\exp\{\boldsymbol{\beta}^\top\tilde{\boldsymbol{Z}}(s,\boldsymbol{\xi})\}\,\mathrm{d}A_0^\Gamma(s)\\
&\left.+\tilde{R}^{\boldsymbol{\theta}^{(0)}}\int_0^\tau R(s)\boldsymbol{Z}_1(s)(-\boldsymbol{\beta}_3 I_{\{\boldsymbol{Z}_2>\boldsymbol{\xi}\}})^\top\boldsymbol{h}_1\exp\{\boldsymbol{\beta}^\top\tilde{\boldsymbol{Z}}(s,\boldsymbol{\xi})\}\,\mathrm{d}A_0^\Gamma(s)\right]\\
\tilde{V}_{\boldsymbol{\beta}_1,2}(\boldsymbol{h}_2) =& \mathrm{P}_0\left[\tilde{R}_1^{\boldsymbol{\theta}^{(0)}}\int_0^\tau R(s)\boldsymbol{Z}_1(s)\exp\{\boldsymbol{\beta}^\top\tilde{\boldsymbol{Z}}(s,\boldsymbol{\xi})\}\,\mathrm{d}A_0^\Gamma(s)\right.\\
&\times\int_0^\tau R(s)\boldsymbol{Z}_1^\top(s)\boldsymbol{h}_2\exp\{\boldsymbol{\beta}^\top\tilde{\boldsymbol{Z}}(s,\boldsymbol{\xi})\}\,\mathrm{d}A_0^\Gamma(s)\\
&\left.+\tilde{R}^{\boldsymbol{\theta}^{(0)}}\int_0^\tau R(s)\boldsymbol{Z}_1(s)\boldsymbol{Z}_1^\top(s)\boldsymbol{h}_2\exp\{\boldsymbol{\beta}^\top\tilde{\boldsymbol{Z}}(s,\boldsymbol{\xi})\}\,\mathrm{d}A_0^\Gamma(s)\right]\\
\tilde{V}_{\boldsymbol{\beta}_1,3}(\boldsymbol{h}_3) =& \mathrm{P}_0\left[\tilde{R}_1^{\boldsymbol{\theta}^{(0)}}\int_0^\tau R(s)\boldsymbol{Z}_1(s)\exp\{\boldsymbol{\beta}^\top\tilde{\boldsymbol{Z}}(s,\boldsymbol{\xi})\}\,\mathrm{d}A_0^\Gamma(s)\right.\\
&\times\int_0^\tau R(s)\boldsymbol{Z}_2^\top\boldsymbol{h}_3\exp\{\boldsymbol{\beta}^\top\tilde{\boldsymbol{Z}}(s,\boldsymbol{\xi})\}\,\mathrm{d}A_0^\Gamma(s)\\
&\left.+\tilde{R}^{\boldsymbol{\theta}^{(0)}}\int_0^\tau R(s)\boldsymbol{Z}_1(s)\boldsymbol{Z}_2^\top\boldsymbol{h}_3\exp\{\boldsymbol{\beta}^\top\tilde{\boldsymbol{Z}}(s,\boldsymbol{\xi})\}\,\mathrm{d}A_0^\Gamma(s)\right]
\end{aligned}
$$

$$
\begin{aligned}
\tilde{V}_{\boldsymbol{\beta}_1,4}(\boldsymbol{h}_4) = &\mathrm{P}_0 \left[\tilde{R}_1^{\boldsymbol{\theta}^{(0)}} \int_0^\tau R(s)\boldsymbol{Z}_1(s) \exp\{\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}(s,\boldsymbol{\xi})\}\, \mathrm{d}A_0^\Gamma(s) \right.\\
&\times \int_0^\tau R(s)((\boldsymbol{Z}_2-\boldsymbol{\xi})^+)^\top \boldsymbol{h}_4 \exp\{\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}(s,\boldsymbol{\xi})\}\, \mathrm{d}A_0^\Gamma(s)\\
&\left. + \tilde{R}^{\boldsymbol{\theta}^{(0)}} \int_0^\tau R(s)\boldsymbol{Z}_1(s)((\boldsymbol{Z}_2-\boldsymbol{\xi})^+)^\top \boldsymbol{h}_4 \exp\{\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}(s,\boldsymbol{\xi})\}\, \mathrm{d}A_0^\Gamma(s) \right]\\
\tilde{V}_{\boldsymbol{\beta}_1,5}(h_5) = &\mathrm{P}_0 \left[\tilde{R}_1^{\boldsymbol{\theta}^{(0)}} \int_0^\tau R(s)\boldsymbol{Z}_1(s) \exp\{\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}(s,\boldsymbol{\xi})\}\, \mathrm{d}A_0^\Gamma(s) \right.\\
&\times \int_0^\tau R(s) h_5(s) \exp\{\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}(s,\boldsymbol{\xi})\}\, \mathrm{d}A_0^\Gamma(s)\\
&\left. + \tilde{R}^{\boldsymbol{\theta}^{(0)}} \int_0^\tau R(s)\boldsymbol{Z}_1(s) h_5(s) \exp\{\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}(s,\boldsymbol{\xi})\}\, \mathrm{d}A_0^\Gamma(s) \right]\\
\tilde{V}_{\boldsymbol{\beta}_2,1}(\boldsymbol{h}_1) = &\mathrm{P}_0 \left[\tilde{R}_1^{\boldsymbol{\theta}^{(0)}} \int_0^\tau R(s)\boldsymbol{Z}_2 \exp\{\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}(s,\boldsymbol{\xi})\}\, \mathrm{d}A_0^\Gamma(s) \right.\\
&\times \int_0^\tau R(s)(-\boldsymbol{\beta}_3 I_{\{\boldsymbol{Z}_2>\boldsymbol{\xi}\}})^\top \boldsymbol{h}_1 \exp\{\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}(s,\boldsymbol{\xi})\}\, \mathrm{d}A_0^\Gamma(s)\\
&\left. + \tilde{R}^{\boldsymbol{\theta}^{(0)}} \int_0^\tau R(s)\boldsymbol{Z}_2(-\boldsymbol{\beta}_3 I_{\{\boldsymbol{Z}_2>\boldsymbol{\xi}\}})^\top \boldsymbol{h}_1 \exp\{\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}(s,\boldsymbol{\xi})\}\, \mathrm{d}A_0^\Gamma(s) \right]\\
\tilde{V}_{\boldsymbol{\beta}_2,2}(\boldsymbol{h}_2) = &\mathrm{P}_0 \left[\tilde{R}_1^{\boldsymbol{\theta}^{(0)}} \int_0^\tau R(s)\boldsymbol{Z}_2 \exp\{\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}(s,\boldsymbol{\xi})\}\, \mathrm{d}A_0^\Gamma(s) \right.\\
&\times \int_0^\tau R(s)\boldsymbol{Z}_1^\top(s) \boldsymbol{h}_2 \exp\{\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}(s,\boldsymbol{\xi})\}\, \mathrm{d}A_0^\Gamma(s)\\
&\left. + \tilde{R}^{\boldsymbol{\theta}^{(0)}} \int_0^\tau R(s)\boldsymbol{Z}_2\boldsymbol{Z}_1^\top(s) \boldsymbol{h}_2 \exp\{\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}(s,\boldsymbol{\xi})\}\, \mathrm{d}A_0^\Gamma(s) \right]\\
\tilde{V}_{\boldsymbol{\beta}_2,3}(\boldsymbol{h}_3) = &\mathrm{P}_0 \left[\tilde{R}_1^{\boldsymbol{\theta}^{(0)}} \int_0^\tau R(s)\boldsymbol{Z}_2 \exp\{\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}(s,\boldsymbol{\xi})\}\, \mathrm{d}A_0^\Gamma(s) \right.\\
&\times \int_0^\tau R(s)\boldsymbol{Z}_2^\top \boldsymbol{h}_3 \exp\{\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}(s,\boldsymbol{\xi})\}\, \mathrm{d}A_0^\Gamma(s)\\
&\left. + \tilde{R}^{\boldsymbol{\theta}^{(0)}} \int_0^\tau R(s)\boldsymbol{Z}_2\boldsymbol{Z}_2^\top \boldsymbol{h}_3 \exp\{\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}(s,\boldsymbol{\xi})\}\, \mathrm{d}A_0^\Gamma(s) \right]\\
\tilde{V}_{\boldsymbol{\beta}_2,4}(\boldsymbol{h}_4) = &\mathrm{P}_0 \left[\tilde{R}_1^{\boldsymbol{\theta}^{(0)}} \int_0^\tau R(s)\boldsymbol{Z}_2 \exp\{\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}(s,\boldsymbol{\xi})\}\, \mathrm{d}A_0^\Gamma(s) \right.\\
&\times \int_0^\tau R(s)((\boldsymbol{Z}_2-\boldsymbol{\xi})^+)^\top \boldsymbol{h}_4 \exp\{\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}(s,\boldsymbol{\xi})\}\, \mathrm{d}A_0^\Gamma(s)\\
&\left. + \tilde{R}^{\boldsymbol{\theta}^{(0)}} \int_0^\tau R(s)\boldsymbol{Z}_2(\boldsymbol{Z}_2-\boldsymbol{\xi})^+ \boldsymbol{h}_4 \exp\{\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}(s,\boldsymbol{\xi})\}\, \mathrm{d}A_0^\Gamma(s) \right]\\
\tilde{V}_{\boldsymbol{\beta}_2,5}(h_5) = &\mathrm{P}_0 \left[\tilde{R}_1^{\boldsymbol{\theta}^{(0)}} \int_0^\tau R(s)\boldsymbol{Z}_2 \exp\{\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}(s,\boldsymbol{\xi})\}\, \mathrm{d}A_0^\Gamma(s) \right.\\
&\times \int_0^\tau R(s) h_5(s) \exp\{\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}(s,\boldsymbol{\xi})\}\, \mathrm{d}A_0^\Gamma(s)\\
&\left. + \tilde{R}^{\boldsymbol{\theta}^{(0)}} \int_0^\tau R(s)\boldsymbol{Z}_2 h_5(s) \exp\{\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}(s,\boldsymbol{\xi})\}\, \mathrm{d}A_0^\Gamma(s) \right]
\end{aligned}
$$

$$
\begin{aligned}
\tilde{V}_{\boldsymbol{\beta}_3,1}(\boldsymbol{h}_1) = \mathrm{P}_0 \Bigg[& \int_0^\tau -I_{\{\boldsymbol{Z}_2 > \boldsymbol{\xi}\}} \boldsymbol{h}_1 \, \mathrm{d}N(s) + \tilde{R}_1^{\boldsymbol{\theta}^{(0)}} \int_0^\tau (-\boldsymbol{\beta}_3 I_{\{\boldsymbol{Z}_2 > \boldsymbol{\xi}\}})^\top \boldsymbol{h}_1 \exp\{\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}(s, \boldsymbol{\xi})\} \, \mathrm{d}A_0^\Gamma(s) \\
& \times \int_0^\tau R(s)(\boldsymbol{Z}_2 - \boldsymbol{\xi})^+ \exp\{\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}(s, \boldsymbol{\xi})\} \, \mathrm{d}A_0^\Gamma(s) \\
& + \tilde{R}^{\boldsymbol{\theta}^{(0)}} \Bigg[\int_0^\tau R(s)(-I_{\{\boldsymbol{Z}_2 > \boldsymbol{\xi}\}}) \boldsymbol{h}_1 \exp\{\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}(s, \boldsymbol{\xi})\} \, \mathrm{d}A_0^\Gamma(s) \\
& + \int_0^\tau R(s)(\boldsymbol{Z}_2 - \boldsymbol{\xi})^+ (-\boldsymbol{\beta}_3 I_{\{\boldsymbol{Z}_2 > \boldsymbol{\xi}\}})^\top \boldsymbol{h}_1 \exp\{\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}(s, \boldsymbol{\xi})\} \, \mathrm{d}A_0^\Gamma(s) \Bigg] \Bigg] \\
\tilde{V}_{\boldsymbol{\beta}_3,2}(\boldsymbol{h}_2) = \mathrm{P}_0 \Bigg[& \tilde{R}_1^{\boldsymbol{\theta}^{(0)}} \int_0^\tau R(s)(\boldsymbol{Z}_2 - \boldsymbol{\xi})^+ \exp\{\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}(s, \boldsymbol{\xi})\} \, \mathrm{d}A_0^\Gamma(s) \\
& \times \int_0^\tau R(s) \boldsymbol{Z}_1^\top(s) \boldsymbol{h}_2 \exp\{\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}(s, \boldsymbol{\xi})\} \, \mathrm{d}A_0^\Gamma(s) \\
& + \tilde{R}^{\boldsymbol{\theta}^{(0)}} \int_0^\tau R(s)(\boldsymbol{Z}_2(s) - \boldsymbol{\xi})^+ \boldsymbol{Z}_1^\top(s) \boldsymbol{h}_2 \exp\{\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}(s, \boldsymbol{\xi})\} \, \mathrm{d}A_0^\Gamma(s) \Bigg] \\
\tilde{V}_{\boldsymbol{\beta}_3,3}(\boldsymbol{h}_3) = \mathrm{P}_0 \Bigg[& \tilde{R}_1^{\boldsymbol{\theta}^{(0)}} \int_0^\tau R(s)(\boldsymbol{Z}_2 - \boldsymbol{\xi})^+ \exp\{\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}(s, \boldsymbol{\xi})\} \, \mathrm{d}A_0^\Gamma(s) \\
& \times \int_0^\tau R(s) \boldsymbol{Z}_2^\top \boldsymbol{h}_3 \exp\{\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}(s, \boldsymbol{\xi})\} \, \mathrm{d}A_0^\Gamma(s) \\
& + \tilde{R}^{\boldsymbol{\theta}^{(0)}} \int_0^\tau R(s)(\boldsymbol{Z}_2 - \boldsymbol{\xi})^+ \boldsymbol{Z}_2^\top \boldsymbol{h}_3 \exp\{\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}(s, \boldsymbol{\xi})\} \, \mathrm{d}A_0^\Gamma(s) \Bigg] \\
\tilde{V}_{\boldsymbol{\beta}_3,4}(\boldsymbol{h}_4) = \mathrm{P}_0 \Bigg[& \tilde{R}_1^{\boldsymbol{\theta}^{(0)}} \int_0^\tau R(s)(\boldsymbol{Z}_2 - \boldsymbol{\xi})^+ \exp\{\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}(s, \boldsymbol{\xi})\} \, \mathrm{d}A_0^\Gamma(s) \\
& \times \int_0^\tau R(s)(\boldsymbol{Z}_2 - \boldsymbol{\xi}) + \boldsymbol{h}_4 \exp\{\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}(s, \boldsymbol{\xi})\} \, \mathrm{d}A_0^\Gamma(s) \\
& + \tilde{R}^{\boldsymbol{\theta}^{(0)}} \int_0^\tau R(s)(\boldsymbol{Z}_2 - \boldsymbol{\xi})^+ (\boldsymbol{Z}_2 - \boldsymbol{\xi})^+ \boldsymbol{h}_4 \exp\{\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}(s, \boldsymbol{\xi})\} \, \mathrm{d}A_0^\Gamma(s) \Bigg] \\
\tilde{V}_{\boldsymbol{\beta}_3,5}(h_5) = \mathrm{P}_0 \Bigg[& \tilde{R}_1^{\boldsymbol{\theta}^{(0)}} \int_0^\tau R(s)(\boldsymbol{Z}_2 - \boldsymbol{\xi})^+ \exp\{\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}(s, \boldsymbol{\xi})\} \, \mathrm{d}A_0^\Gamma(s) \\
& \times \int_0^\tau R(s) h_5(s) \exp\{\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}(s, \boldsymbol{\xi})\} \, \mathrm{d}A_0^\Gamma(s) \\
& + \tilde{R}^{\boldsymbol{\theta}^{(0)}} \int_0^\tau R(s)(\boldsymbol{Z}_2 - \boldsymbol{\xi})^+ h_5(s) \exp\{\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}(s, \boldsymbol{\xi})\} \, \mathrm{d}A_0^\Gamma(s) \Bigg] \\
\tilde{V}_{A_0,1}(\boldsymbol{h}_1) = \mathrm{P}_0 \Big[& \tilde{R}_1^{\boldsymbol{\theta}^{(0)}} R(s) \exp\{\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}(s, \boldsymbol{\xi})\} \\
& \times \int_0^\tau R(s)(-\boldsymbol{\beta}_3 I_{\{\boldsymbol{Z}_2 > \boldsymbol{\xi}\}})^\top \boldsymbol{h}_1 \exp\{\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}(s, \boldsymbol{\xi})\} \, \mathrm{d}A_0^\Gamma(s) \\
& + \tilde{R}^{\boldsymbol{\theta}^{(0)}} R(s)(-\boldsymbol{\beta}_3 I_{\{\boldsymbol{Z}_2 > \boldsymbol{\xi}\}})^\top \boldsymbol{h}_1 \exp\{\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}(s, \boldsymbol{\xi})\} \Big] \\
\tilde{V}_{A_0,2}(\boldsymbol{h}_2) = \mathrm{P}_0 \Big[& \tilde{R}_1^{\boldsymbol{\theta}^{(0)}} R(s) \exp\{\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}(s, \boldsymbol{\xi})\} \\
& \times \int_0^\tau R(s) \boldsymbol{Z}_1^\top(s) \boldsymbol{h}_2 \exp\{\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}(s, \boldsymbol{\xi})\} \, \mathrm{d}A_0^\Gamma(s) \\
& + \tilde{R}^{\boldsymbol{\theta}^{(0)}} R(s) \boldsymbol{Z}_1^\top(s) \boldsymbol{h}_2 \exp\{\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}(s, \boldsymbol{\xi})\} \Big]
\end{aligned}
$$

$$
\begin{aligned}
\tilde{V}_{A_0,3}(\boldsymbol{h}_3) = & \mathrm{P}_0 \Big[\tilde{R}_1^{\boldsymbol{\theta}^{(0)}} R(s) \exp\{\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}(s,\boldsymbol{\xi})\} \\
& \times \int_0^\tau R(s) \boldsymbol{Z}_2^\top \boldsymbol{h}_3 \exp\{\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}(s,\boldsymbol{\xi})\} \,\mathrm{d}A_0^\Gamma(s) \\
& + \tilde{R}^{\boldsymbol{\theta}^{(0)}} R(s) \boldsymbol{Z}_2^\top \boldsymbol{h}_3 \exp\{\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}(s,\boldsymbol{\xi})\}\Big] \\
\tilde{V}_{A_0,4}(\boldsymbol{h}_4) = & \mathrm{P}_0 \Big[\tilde{R}_1^{\boldsymbol{\theta}^{(0)}} R(s) \exp\{\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}(s,\boldsymbol{\xi})\} \\
& \times \int_0^\tau R(s) ((\boldsymbol{Z}_2 - \boldsymbol{\xi})^+)^\top \boldsymbol{h}_4 \exp\{\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}(s,\boldsymbol{\xi})\} \,\mathrm{d}A_0^\Gamma(s) \\
& + \tilde{R}^{\boldsymbol{\theta}^{(0)}} R(s) ((\boldsymbol{Z}_2 - \boldsymbol{\xi})^+)^\top \boldsymbol{h}_4 \exp\{\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}(s,\boldsymbol{\xi})\}\Big] \\
\tilde{V}_{A_0,5}(h_5) = & \mathrm{P}_0 \Big[\tilde{R}_1^{\boldsymbol{\theta}^{(0)}} R(s) \exp\{\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}(s,\boldsymbol{\xi})\} \\
& \times \int_0^\tau R(s) h_5(s) \exp\{\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}(s,\boldsymbol{\xi})\} \,\mathrm{d}A_0^\Gamma(s) \\
& + \tilde{R}^{\boldsymbol{\theta}^{(0)}} R(s) h_5(s) \exp\{\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}(s,\boldsymbol{\xi})\}\Big].
\end{aligned}
$$

The next lemma yields that the Gateaux differentiability of the score operator can be strengthened to Fréchet differentiability.

Lemma 6.4. *Under the regularity condition of Section 6.3, the operator* $\boldsymbol{\vartheta} \mapsto \mathrm{P}_0\, U^\tau(\boldsymbol{\vartheta})$ *is Fréchet differentiable at* $\boldsymbol{\vartheta}^*$ *for any* $\boldsymbol{\vartheta}^* \in \bar{\Theta}$ *with derivative* $-\boldsymbol{\vartheta}(\sigma_{\boldsymbol{\vartheta}^*}(\boldsymbol{h}))$*, where* $\boldsymbol{h}$ *ranges over* $\mathcal{H}_\infty$ *and* $\boldsymbol{\vartheta}$ *ranges over the linear span* $\operatorname{lin}\bar{\Theta}$ *of* $\bar{\Theta}$.

Proof. Consider the expression

$$
\begin{aligned}
& \lim_{t\to 0^+} \sup_{\boldsymbol{h}^* \in \operatorname{lin}\bar{\Theta}:\rho(\boldsymbol{h}^*)\le 1} \sup_{\boldsymbol{h}\in\mathcal{H}_\infty} \Big|\int_0^1 \boldsymbol{h}^*(\sigma_{\boldsymbol{\vartheta}^*+st\boldsymbol{h}^*}(\boldsymbol{h}) - \sigma_{\boldsymbol{\vartheta}^*}(\boldsymbol{h}))\,\mathrm{d}s\Big| \\
= & \lim_{t\to 0^+} \sup_{\boldsymbol{h}^* \in \operatorname{lin}\bar{\Theta}:\rho(\boldsymbol{h}^*)\le 1} \sup_{\boldsymbol{h}\in\mathcal{H}_\infty} \Big|\frac{1}{t}\int_0^t \boldsymbol{h}^*(\sigma_{\boldsymbol{\vartheta}^*+u\boldsymbol{h}^*}(\boldsymbol{h}) - \sigma_{\boldsymbol{\vartheta}^*}(\boldsymbol{h}))\,\mathrm{d}u\Big|
\end{aligned} \tag{6.13}
$$

Note that $u \mapsto \boldsymbol{h}^*(\sigma_{\boldsymbol{\vartheta}^*+u\boldsymbol{h}^*}(\boldsymbol{h}))$ is integrable with antiderivative $G^*(u)$ due to the smoothness condition D.1. Furthermore, note that

$$
\begin{aligned}
0 &= \boldsymbol{h}^*(\sigma_{\boldsymbol{\vartheta}^*}(\boldsymbol{h})) - \boldsymbol{h}^*(\sigma_{\boldsymbol{\vartheta}^*}(\boldsymbol{h})) \\
&= \lim_{t\to 0^+} \frac{G^*(t) - G^*(0)}{t} - \boldsymbol{h}^*(\sigma_{\boldsymbol{\vartheta}^*}(\boldsymbol{h})) \\
&= \lim_{t\to 0^+} \frac{1}{t}\int_0^t \boldsymbol{h}^*(\sigma_{\boldsymbol{\vartheta}^*+u\boldsymbol{h}^*}(\boldsymbol{h}) - \sigma_{\boldsymbol{\vartheta}^*}(\boldsymbol{h}))\,\mathrm{d}u.
\end{aligned}
$$

which implies that (6.13) equals to zero. Hence, the assertion of the lemma is proved. □

The following lemma shows the continuous invertibility of the operators $\sigma_{\boldsymbol{\theta}_0}$ and $\boldsymbol{\vartheta} \mapsto \boldsymbol{\vartheta}(\sigma_{\boldsymbol{\theta}_0}(\cdot))$. Note that $\sigma_{\boldsymbol{\theta}_0} = \sigma_{\boldsymbol{\vartheta}_0}$ and $A_0^{\Gamma_0} = A_0$.

Lemma 6.5. *Under the regularity conditions of Section 6.3, the information operator $\sigma_{\boldsymbol{\theta}_0} : \mathcal{H}_\infty \to \mathcal{H}_\infty$ is continuously invertible and onto with inverse $\sigma_{\boldsymbol{\theta}_0}^{-1}$. The linear operator $\boldsymbol{\vartheta} \mapsto \boldsymbol{\vartheta}(\sigma_{\boldsymbol{\theta}_0}(\cdot))$ from $\operatorname{lin}\bar{\Theta}$ into itself is also continuously invertible and onto with inverse $\boldsymbol{\vartheta} \mapsto \boldsymbol{\vartheta}(\sigma_{\boldsymbol{\theta}_0}^{-1}(\cdot))$.*

Proof. For any $\boldsymbol{h} = (\boldsymbol{h}_1, \boldsymbol{h}_2, \boldsymbol{h}_3, \boldsymbol{h}_4, h_5) \in \mathcal{H}_\infty$ we can write $\sigma_{\boldsymbol{\theta}_0}(\boldsymbol{h})$ as a sum $K(\boldsymbol{h}) + C(\boldsymbol{h})$, where $K(\boldsymbol{h}) = (\boldsymbol{h}_1, \boldsymbol{h}_2, \boldsymbol{h}_3, \boldsymbol{h}_4, g_0 h_5)$ and $C(\boldsymbol{h}) = \sigma_{\boldsymbol{\theta}_0}(\boldsymbol{h}) - K(\boldsymbol{h})$ and $g_0(s) = \mathrm{P}_0 \left[R(s) \exp\{\boldsymbol{\beta}_0^\top \tilde{\boldsymbol{Z}}(s, \boldsymbol{\xi}_0)\} \tilde{R}^{\boldsymbol{\theta}_0} \right]$. Since g_0 is bounded, K is one-to-one and onto with continuous inverse defined by $K^{-1}(\boldsymbol{h}) = (\boldsymbol{h}_1, \boldsymbol{h}_2, \boldsymbol{h}_3, \boldsymbol{h}_4, h_5/g_0)$. We need to show that the operator $C(\boldsymbol{h})$ is compact. Because of the construction of σ and since a bounded linear operator with finite-dimensional range is compact, we will only consider $C_5 : BV \mapsto BV$ given by

$$C_5(h_5) = \mathrm{P}_0 \left[\tilde{R}_1^{\boldsymbol{\theta}^{(0)}} R(s) \exp\{\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}(s, \boldsymbol{\xi})\} \int_0^\tau R(s) h_5(s) \exp\{\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}(s, \boldsymbol{\xi})\} \, \mathrm{d}A_0^\Gamma(s) \right].$$

To prove compactness of C_5 we show that for an arbitrary bounded sequence of functions $\{\bar{h}_{5n}\}$, there exists a convergent subsequence of $\{C(\bar{h}_{5n})\}$ such that the limit point is an element of BV. Now, C_5 is a linear operator with $\|C_5(h_5)\|_v \le M \int_0^\tau |h_5| \, \mathrm{d}A_0^\Gamma(s)$ for every h_5 and for a fixed constant M. Hence, it suffices to show that there exists a subsequence of $\bar{h}_{5n}$ that converges in L_1. By Helly's selection theorem we can find a subsequence $\{\bar{h}_{5n_k}\}$, such that $\bar{h}_{5n_k}$ converges pointwise to some function $\bar{h}_5$ as $k \to \infty$. Using the dominated convergence theorem $\bar{h}_{5n_k}$ converges to a limit in L_1. Hence, the operator $C(\boldsymbol{h})$ is compact.

Now we will show that $\sigma_{\boldsymbol{\theta}_0}$ is one-to-one, i.e. if $\|\boldsymbol{h}\| > 0$, then $\|\sigma_{\boldsymbol{\theta}_0}(\boldsymbol{h})\| > 0$. Suppose this is not the case, i.e. $\sigma_{\boldsymbol{\theta}_0}(\boldsymbol{h}) = 0$ for some $\boldsymbol{h} \in \mathcal{H}_\infty$. Then, for the one-dimensional submodel defined by the map $s \mapsto \boldsymbol{\vartheta}_{0s} = \boldsymbol{\vartheta}_0 + s(\boldsymbol{h}_1, \boldsymbol{h}_2, \boldsymbol{h}_3, \boldsymbol{h}_4, \int_0^\tau h_5(u) \, \mathrm{d}A_0(u))$ we have that

$$\mathrm{P}_0 \left[U^\tau(\boldsymbol{\vartheta}_0)(h) \right]^2 = 0. \tag{6.14}$$

Define a random set $S(n, r, t) = \{\omega : (N(u), R(u))(\omega) = (n(u), r(u)), u \in [t, \tau]\}$. Because of the positivity of $[U^\tau(\boldsymbol{\vartheta}_0)(h)]^2$ equality (6.14) implies that $\mathrm{P}_0 \left[U^\tau(\boldsymbol{\vartheta}_0)(\boldsymbol{h}) | S(n, r, t) \right]^2 = 0$ for all S such that $\mathrm{P}_0(S(n, r, t)) > 0$. Hence, $U^t(\boldsymbol{\vartheta}_0)(\boldsymbol{h})$ equals zero almost surely for all $t \in [0, \tau]$. Now consider the set on which the observation (V, δ, Z) is censored at time $t \in [0, \tau]$. The equality $U^t(\boldsymbol{\vartheta}_0)(\boldsymbol{h}) = 0$ at a censoring time t yields

$$\begin{aligned} \int_0^t R(s) \left[\boldsymbol{h}_1(-\boldsymbol{\beta}_{30}) I_{\{\boldsymbol{Z}_2 > \boldsymbol{\xi}_0\}} + \boldsymbol{h}_2^\top \boldsymbol{Z}_1(s) + \boldsymbol{h}_3^\top \boldsymbol{Z}_2 + \boldsymbol{h}_4^\top (\boldsymbol{Z}_2 - \boldsymbol{\xi}_0)^+ + h_5(s) \right] & \\ \times \exp\{\boldsymbol{\beta}_0^\top \tilde{\boldsymbol{Z}}(s, \boldsymbol{\xi}_0)\} \, \mathrm{d}A_0(s) = 0. & \end{aligned}$$

Taking the Radon-Nikodym derivative with respect to A_0 and dividing the equality by $\exp\{\boldsymbol{\beta}_0^\top \tilde{\boldsymbol{Z}}(t, \boldsymbol{\xi}_0)\}$ yields

$$R(t)\left[\boldsymbol{h}_1(-\boldsymbol{\beta}_{30})I_{\{\boldsymbol{Z}_2>\boldsymbol{\xi}_0\}} + \boldsymbol{h}_2^\top \boldsymbol{Z}_1(t) + \boldsymbol{h}_3^\top \boldsymbol{Z}_2 + \boldsymbol{h}_4^\top(\boldsymbol{Z}_2 - \boldsymbol{\xi}_0)^+ + h_5(t)\right] = 0.$$

Using the same arguments as in the proof of Lemma 6.1 we get that $\boldsymbol{h} = \boldsymbol{0}$. Hence, $\sigma_{\boldsymbol{\theta}_0}(\boldsymbol{h}) = \boldsymbol{0}$ implies $\boldsymbol{h} = \boldsymbol{0}$, thus $\sigma_{\boldsymbol{\theta}_0}$ is one-to-one. Using the assertion of Lemma 25.93 in Van der Vaart (1998) yields that $\sigma_{\boldsymbol{\theta}_0}$ is onto and continuously invertible with $\sigma_{\boldsymbol{\theta}_0}^{-1}$. Using this fact, we know that for each $0 < r < \infty$, there exists an $s > 0$ with $\sigma_{\boldsymbol{\theta}_0}^{-1}(\mathcal{H}_s) \subset \mathcal{H}_r$. Hence,

$$\inf_{\boldsymbol{\vartheta}\in \operatorname{lin}\bar{\Theta}} \frac{\|\boldsymbol{\vartheta}(\sigma_{\boldsymbol{\theta}_0}(\cdot))\|_{(r)}}{\|\boldsymbol{\vartheta}\|_{(r)}} \geq \inf_{\boldsymbol{\vartheta}\in \operatorname{lin}\bar{\Theta}} \frac{\sup_{\boldsymbol{h}\in\sigma_{\boldsymbol{\theta}_0}^{-1}(\mathcal{H}_s)} |\boldsymbol{\vartheta}(\sigma_{\boldsymbol{\theta}_0}(\boldsymbol{h}))|}{\|\boldsymbol{\vartheta}\|_{(r)}} = \inf_{\boldsymbol{\vartheta}\in \operatorname{lin}\bar{\Theta}} \frac{\|\boldsymbol{\vartheta}\|_{(s)}}{\|\boldsymbol{\vartheta}\|_{(r)}} \geq \frac{s}{5r}.$$

Since $\|\boldsymbol{\vartheta}(\sigma_{\boldsymbol{\theta}_0}(\cdot))\|_{(r)} \geq \frac{s}{5r}\|\boldsymbol{\vartheta}\|_{(r)}$, the linear operator $\boldsymbol{\vartheta} \mapsto \boldsymbol{\vartheta}(\sigma_{\boldsymbol{\theta}_0}(\cdot))$ is continuously invertible using Proposition A.1.7 in Bickel et al. (1998). The operator $\boldsymbol{\vartheta}(\sigma_{\boldsymbol{\theta}_0}(\cdot))$ is onto with inverse $\boldsymbol{\vartheta} \mapsto \boldsymbol{\vartheta}(\sigma_{\boldsymbol{\theta}_0}^{-1}(\cdot))$, since $\sigma_{\boldsymbol{\theta}_0}$ is onto. This proves the lemma. □

Before we study the local behavior of $L(\boldsymbol{\vartheta})$ define

$$\bar{B}_\epsilon^k = \{\boldsymbol{\vartheta} \in \bar{\Theta} : \rho(\boldsymbol{\vartheta} - \boldsymbol{\vartheta}_0) < \epsilon,\ \|\Gamma\|_\nu < k\},$$

where $\|\cdot\|_v$ is the total variation norm on BV. Note that $\hat{\boldsymbol{\vartheta}} = (\hat{\boldsymbol{\xi}}_n, \hat{\boldsymbol{\beta}}_n, \hat{\Gamma}_n)$ lies in $\bar{B}_\epsilon^k$ for all n large enough, since $\hat{\boldsymbol{\theta}}_n$ is consistent and because of the next lemma.

Lemma 6.6. *There exists a $k_0 < \infty$ such that*

$$\limsup_{n\to\infty} \|\hat{\Gamma}_n\|_v \leq k_0 \text{ and } \lim_{n\to\infty} \|\hat{\Gamma}_n - \Gamma_0\|_\infty = 0 \text{ outer almost surely.}$$

Proof. Consider $\hat{\Gamma}(\cdot) = \log(\mathbb{P}_n W(\cdot, \hat{\boldsymbol{\theta}}_n))$ with

$$W(s, \hat{\boldsymbol{\theta}}_n) = R(s)\exp\{\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}(s, \boldsymbol{\xi})\}\left\{G'(H^{\hat{\boldsymbol{\theta}}_n}(V)) - \delta\,\frac{G''(H^{\hat{\boldsymbol{\theta}}_n}(V))}{G'(H^{\hat{\boldsymbol{\theta}}_n}(V))}\right\}.$$

The total variation of $t \mapsto R(t)$ is bounded by 1. Due to condition A.2, the total variation of $t \mapsto \exp\{\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}(t, \boldsymbol{\xi})\}$ is bounded by a constant K_0. Thus,

$$\|\mathbb{P}_n W(\cdot, \hat{\boldsymbol{\theta}}_n)\|_v \leq K_0 \mathbb{P}_n \left| G'(H^{\hat{\boldsymbol{\theta}}_n}(V)) - \delta \frac{G''(H^{\hat{\boldsymbol{\theta}}_n}(V))}{G'(H^{\hat{\boldsymbol{\theta}}_n}(V))}\right|.$$

Since all functions, which are involved are smooth and since the logarithm is Lipschitz on compacts bounded away from zero, the first result follows. The second part follows from Lemma 6.3 combined with the consistency of the estimate $\hat{\boldsymbol{\theta}}_n$, the continuity of $\boldsymbol{\theta} \mapsto \mathrm{P}\, W(\cdot, \boldsymbol{\theta})$, and reapplication of the Lipschitz continuity of $u \mapsto \log(u)$. □

Now, by the definition of the score and information operators the first derivative of $\boldsymbol{\vartheta} \mapsto L(\boldsymbol{\vartheta})$ in the direction $\boldsymbol{h} \in \mathcal{H}_\infty$ is zero at the point $\boldsymbol{\vartheta}_0$, while the second derivative in the same direction is < 0, which is proved in Lemma 6.5. The conditions D.1 and D.2 ensure the smoothness of the score and information operators. This smoothness and the arbitrariness of $\boldsymbol{h}$ yield that $L(\boldsymbol{\vartheta})$ is concave for every $\boldsymbol{\vartheta} \in \bar{B}^k_\epsilon$, for sufficiently small ϵ.

6.6 Rate of Convergence

In a recent paper by Kosorok & Song (2007) it has been claimed that our model is a submodel of theirs. However, when the rate of convergence is concerned the two models have to be recognized as different. Usually, when a change-point model with a jump is considered the rate of convergence of the change-point estimator is n. In our case it turns out that the rate of convergence of the change-point estimate is not better than $\sqrt{n}$. The difference between a jump and a bent-line change-point is the continuity of the likelihood in $\boldsymbol{\xi}$ as described in the simpler models of the previous chapters. The continuity causes the limit of the likelihood to be differentiable in $\boldsymbol{\xi}$. Therefore, our case leads to a rate of convergence different from that one in Kosorok & Song (2007).
Define the process $X_n(\boldsymbol{\xi}, \boldsymbol{\beta}, \Gamma) = L_n(\boldsymbol{\vartheta}) - L_n(\boldsymbol{\vartheta}_0)$ and the function $X(\boldsymbol{\xi}, \boldsymbol{\beta}, \Gamma) = L(\boldsymbol{\vartheta}) - L(\boldsymbol{\vartheta}_0)$. Moreover, let $D_n = \sqrt{n}(X_n(\boldsymbol{\vartheta}) - X(\boldsymbol{\vartheta}))$.

Lemma 6.7. *Under the conditions in Section 6.3, for ϵ sufficiently small there exists a constant $\alpha > 0$ such that for all $\boldsymbol{\vartheta} \in \bar{B}^k_\epsilon$, $X(\boldsymbol{\vartheta}) \leq -\alpha\rho(\boldsymbol{\vartheta} - \boldsymbol{\vartheta}_0)^2$.*

Proof. For $L(\boldsymbol{\vartheta})$ we know that $\mathrm{P}_0\, U^\tau(\boldsymbol{\vartheta}_0) = 0$ by definition and the second derivative $\boldsymbol{I}(\boldsymbol{\vartheta}_0) = \frac{\partial}{\partial \boldsymbol{\vartheta}} \mathrm{P}_0\, U^\tau(\boldsymbol{\vartheta})\big|_{\boldsymbol{\vartheta}=\boldsymbol{\vartheta}_0}$ is negative definite. Hence, by a Taylor expansion of $L(\boldsymbol{\vartheta})$ for ϵ sufficiently small and for $\boldsymbol{\vartheta} \in \bar{B}^k_\epsilon(\boldsymbol{\vartheta}_0)$,

$$\begin{aligned} X(\boldsymbol{\vartheta}) &= \mathrm{P}_0\, U^\tau(\boldsymbol{\vartheta}_0)(\boldsymbol{\vartheta} - \boldsymbol{\vartheta}_0) + \frac{1}{2}(\boldsymbol{\vartheta} - \boldsymbol{\vartheta}_0)^\top \boldsymbol{I}(\boldsymbol{\vartheta}_0)(\boldsymbol{\vartheta} - \boldsymbol{\vartheta}_0) + o(\rho(\boldsymbol{\vartheta} - \boldsymbol{\vartheta}_0)^2) \\ &\leq -\alpha\rho(\boldsymbol{\vartheta} - \boldsymbol{\vartheta}_0)^2, \end{aligned}$$

since $\boldsymbol{I}(\boldsymbol{\vartheta}_0)$ is negative definite. □

Lemma 6.8. *Under the regularity conditions in Section 6.3, for every $\epsilon > 0$ there exists a constant $\kappa > 0$ such that*

$$E[\sup_{\boldsymbol{\vartheta} \in \bar{B}^k_\epsilon} |D_n(\boldsymbol{\vartheta})|] \leq \kappa\epsilon, \quad \textit{for all } n \in \mathbb{N}.$$

Proof. Consider the process

$$
\begin{aligned}
D_n(\boldsymbol{\vartheta}) &= \sqrt{n}\,\{L_n(\boldsymbol{\vartheta}) - L(\boldsymbol{\vartheta}) - L_n(\boldsymbol{\vartheta}_0) + L(\boldsymbol{\vartheta}_0)\} \\
&= \sqrt{n}\Big\{\mathbb{P}_n\Big[\int_0^\tau \Big(-\Gamma(t) + \boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}(t,\boldsymbol{\xi}) + \log(G'(H^{\boldsymbol{\theta}^{(n)}}(t)))\Big)\,\mathrm{d}N(t) - G(H^{\boldsymbol{\theta}^{(n)}}(V))\Big] \\
&\quad - \mathrm{P}_0\Big[\int_0^\tau \Big(-\Gamma(t) + \boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}(t,\boldsymbol{\xi}) + \log(G'(H^{\boldsymbol{\theta}^{(0)}}(t)))\Big)\,\mathrm{d}N(t) - G(H^{\boldsymbol{\theta}^{(0)}}(V))\Big] \\
&\quad -\mathbb{P}_n\Big[\int_0^\tau \Big(-\Gamma_0(t) + \boldsymbol{\beta}_0^\top \tilde{\boldsymbol{Z}}(t,\boldsymbol{\xi}_0) + \log(G'(H^{\boldsymbol{\theta}_0^{(n)}}(t)))\Big)\,\mathrm{d}N(t) - G(H^{\boldsymbol{\theta}_0^{(n)}}(V))\Big] \\
&\quad + \mathrm{P}_0\Big[\int_0^\tau \Big(-\Gamma_0(t) + \boldsymbol{\beta}_0^\top \tilde{\boldsymbol{Z}}(t,\boldsymbol{\xi}_0) + \log(G'(H^{\boldsymbol{\theta}_0}(t)))\Big)\,\mathrm{d}N(t) - G(H^{\boldsymbol{\theta}_0}(V))\Big]\Big\} \\
&= \sqrt{n}(\mathbb{P}_n - \mathrm{P}_0)\Big[\int_0^\tau (-\Gamma(t) + \Gamma_0(t))\,\mathrm{d}N(t)\Big] \\
&\quad + \sqrt{n}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)^\top(\mathbb{P}_n - \mathrm{P}_0)\Big[\int_0^\tau \tilde{\boldsymbol{Z}}(t,\boldsymbol{\xi}_0)\,\mathrm{d}N(t)\Big] \\
&\quad + \sqrt{n}\boldsymbol{\beta}_3^\top(\mathbb{P}_n - \mathrm{P}_0)\int_0^\tau (\boldsymbol{Z}_2 - \boldsymbol{\xi})^+ - (\boldsymbol{Z}_2 - \boldsymbol{\xi}_0)^+\,\mathrm{d}N(t) \\
&\quad + \sqrt{n}\Big\{\mathbb{P}_n\Big[\int_0^\tau \log(G'(H^{\boldsymbol{\theta}^{(n)}}(t))) - \log(G'(H^{\boldsymbol{\theta}_0^{(n)}}(t)))\,\mathrm{d}N(t)\Big] \\
&\qquad - \mathrm{P}_0\Big[\int_0^\tau \log(G'(H^{\boldsymbol{\theta}^{(0)}}(t))) - \log(G'(H^{\boldsymbol{\theta}_0}(t)))\,\mathrm{d}N(t)\Big]\Big\} \\
&\quad - \sqrt{n}\Big\{\mathbb{P}_n\Big[G(H^{\boldsymbol{\theta}^{(n)}}(V)) - G(H^{\boldsymbol{\theta}_0^{(n)}}(V))\Big] - \mathrm{P}_0\Big[G(H^{\boldsymbol{\theta}^{(0)}}(V)) - G(H^{\boldsymbol{\theta}_0}(V))\Big]\Big\}.
\end{aligned}
$$

The expectation of the first term is of order $O(\epsilon)$. To establish this we use Theorem 2.14.1 in Van der Vaart & Wellner (1996) with the envelope function $\int_0^\tau (\Gamma_0(s) + \epsilon)\,\mathrm{d}N(s)$ for $\int_0^\tau \Gamma(s)\,\mathrm{d}N(s)$ in $\bar{B}_\epsilon^k$. The second and the third term can be handled as in the proof of Lemma 4.4 in Chapter 4 and hence, it is $O(\epsilon)$. To obtain the order of the fourth term note that the functions $\log(\cdot)$, $\exp(\cdot)$, G' and G'' are Lipschitz. Hence,

$$
\begin{aligned}
&\left|\int_0^\tau \log(G'(H^{\boldsymbol{\theta}^{(n)}}(t))) - \log(G'(H^{\boldsymbol{\theta}_0^{(n)}}(t)))\,\mathrm{d}N(t)\right| \\
&\le \int_0^\tau K\left|H^{\boldsymbol{\theta}^{(n)}}(t) - H^{\boldsymbol{\theta}_0^{(n)}}(t)\right|\,\mathrm{d}N(t) \\
&\le \int_0^\tau \left(K\int_0^t |\exp\{\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}(s,\boldsymbol{\xi}) - \Gamma(s)\} - \exp\{\boldsymbol{\beta}_0^\top \tilde{\boldsymbol{Z}}(s,\boldsymbol{\xi}_0) - \Gamma_0(s)\}|\,\mathrm{d}\mathbb{P}_n N(s)\right)\,\mathrm{d}N(t) \\
&\le \int_0^\tau \left(K\int_0^t K_2\,|\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}(s,\boldsymbol{\xi}) - \Gamma(s) - \boldsymbol{\beta}_0^\top \tilde{\boldsymbol{Z}}(s,\boldsymbol{\xi}_0) - \Gamma_0(s)|\,\mathrm{d}\mathbb{P}_n N(s)\right)\,\mathrm{d}N(t),
\end{aligned}
$$

where $K > 0$ and $K_2 > 0$ are some constants. Similarly, we can bound

$$\left| \int_0^\tau \log(G'(H^{\boldsymbol{\theta}^{(0)}}(t))) - \log(G'(H^{\boldsymbol{\theta}_0}(t)))\, \mathrm{d}N(t) \right|.$$

Since it is possible to find an envelope function for $\boldsymbol{\beta}^\top \tilde{\boldsymbol{Z}}(s, \boldsymbol{\xi}) - \Gamma(s) - \boldsymbol{\beta}_0^\top \tilde{\boldsymbol{Z}}(s, \boldsymbol{\xi}_0) - \Gamma_0(s)$ by using the arguments as above, we can apply Theorem 2.14.1 in Van der Vaart & Wellner (1996) again and obtain that

$$E\left[\sup_{\boldsymbol{\vartheta} \in B_\epsilon^k} \sqrt{n} \left| \mathbb{P}_n \int_0^\tau \log\left(\frac{G'(H^{\boldsymbol{\theta}^{(n)}}(t))}{G'(H^{\boldsymbol{\theta}_0^{(n)}}(t))} \right) \mathrm{d}N(t) - \mathrm{P}_0 \int_0^\tau \log\left(\frac{G'(H^{\boldsymbol{\theta}^{(0)}}(t))}{G'(H^{\boldsymbol{\theta}_0}(t))} \right) \mathrm{d}N(t) \right| \right]$$

is bounded by ϵ times a constant.

Since G is also Lipschitz the last term can be treated similarly as the fourth term. This yields the desired result. □

Theorem 6.3. *Under the conditions 6.3,* $\sqrt{n}\rho(\hat{\boldsymbol{\vartheta}}_n - \boldsymbol{\vartheta}_0) = O_{\mathrm{P}}(1)$.

Proof. Since Lemma 6.7 and Lemma 6.8 hold, the assertion can be proved using the same techniques as in the proof of Theorem 4.3 in Chapter 4. □

6.7 Asymptotic Normality

In this section the asymptotic normality is proved. We use Hoffmann-Jørgensen weak convergence as described in Van der Vaart & Wellner (1996).

Theorem 6.4. *Under the conditions described in Section 6.3,* $\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)$ *is asymptotically linear, with influence function* $\tilde{m}(\boldsymbol{h}) = U^\tau(\boldsymbol{\theta}_0)(\sigma_{\boldsymbol{\theta}_0}^{-1}(\boldsymbol{h}))$, $\boldsymbol{h} \in \mathcal{H}$, *converging weakly in the uniform norm to a tight, mean zero Gaussian process* $\mathcal{Z}$ *with covariance* $E[\tilde{m}(\boldsymbol{g})\tilde{m}(\boldsymbol{h})]$, *for all* $\boldsymbol{g}, \boldsymbol{h} \in \mathcal{H}$.

Proof. We use theorem A.10 in the Appendix to show the assertion of the lemma. The class of functions $\{U_{\boldsymbol{\theta},\boldsymbol{h}} : \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| < \epsilon, h \in \mathcal{H}\}$ is P_0 –Donskser for some $\epsilon > 0$ by the same arguments as in Lemma 6.3. Moreover, the continuity of the functions involved yields that $\sup_{h\in\mathcal{H}} \mathrm{P}_0 \left(U_{\boldsymbol{\theta},\boldsymbol{h}} - U_{\boldsymbol{\theta}_0,\boldsymbol{h}}\right)^2 \to 0$ as $\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| \to 0$. The map $\boldsymbol{\theta} \mapsto \mathrm{P}_0\, U^\tau(\boldsymbol{\theta})$ is twice continuously differentiable at $\boldsymbol{\theta}_0$ with nonsingular derivative matrix. The other conditions of Theorem A.10 are proved in Lemma 6.4, Lemma 6.5 and Theorem 6.3. □

The estimate $\hat{\boldsymbol{\theta}}_n$ is regular and hence as sufficient as if the change-point parameter were known, since $\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)$ is asymptotically linear and provides an influence function which is contained in the closed linear span of the tangent space. The latter is the case, since $\sigma_{\boldsymbol{\theta}_0}$ is continuously invertible.

Chapter 7

Applications

We investigate three different datasets. The first is chosen from an actuarial context, while the second contains electric motor data. The last considers the well known PBC dataset described in Fleming & Harrington (1991). Especially, the first two datasets will be discussed in detail. The last dataset is mentioned, since it enables us to see the difference between a piecewise linear approximation of the functional form and a function guessed from a plot.
Furthermore, we will give a short overview of a goodness-of-fit test that we used to determine which model fits best. Also, we will describe a heuristic development of the martingale residual plots used.

7.1 Goodness-of-Fit Tests

Goodness-of-fit tests are used to examine whether a model is adequate. We describe two tests. The first investigates whether an Aalen model fits better than a Cox model, whereas the second determines whether a Cox model with change-point has a better fit than a Cox model without a change-point. The following methods for testing goodness-of-fit in the Aalen model are based on Gandy & Jensen (2005).
Assume that $\boldsymbol{c}(t) = (c_1(t), ..., c_n(t))^\top$ is a vector of predictable stochastic processes such that $\boldsymbol{c}(t)$ is perpendicular to the columns of the matrix of covariates $\mathbf{Y}(t)$ in the Aalen model, i.e. $\mathbf{Y}^\top(t)\mathbf{c}(t) = 0$ for all $t \geq 0$. Then under some regularity conditions

$$\hat{T}(t) := \frac{1}{\sqrt{n}} \int_0^t \boldsymbol{c}^\top(s)\, \mathrm{d}\boldsymbol{N}(s)$$

is a local martingale. The process $\mathbf{c}(t)$ can be defined by a projection of some vector $\mathbf{d}(t)$ onto the orthogonal complement of the column space of $\mathbf{Y}(t)$. With the corresponding projection matrix $\mathbf{P}(t)$ we get $\mathbf{c}(t) = \mathbf{P}(t)\mathbf{d}(t)$. If $\mathbf{Y}(t)$ has full rank we can set $\mathbf{P}(t) = \mathbf{I} -$

$\mathbf{Y}(t)(\mathbf{Y}^\top(t)\mathbf{Y}(t))^{-1}\mathbf{Y}^\top(t)$. Gandy & Jensen (2005) suggest different choices of $\mathbf{d}$ to detect the following alternatives. The first alternative is a completely known fixed alternative. The second alternative is that there is an additional covariate. The third alternative is that the Cox model (2.1) holds. The covariates $\boldsymbol{Z}(t)$ in the Cox model do not have to be the same as in the Aalen model. The suggested choice for detecting Cox's model is

$$d_i(s) := R_i(s)\exp\{\hat{\boldsymbol{\beta}}^\top \boldsymbol{Z}_i(s)\}, \tag{7.1}$$

where $\hat{\boldsymbol{\beta}}$ is the maximum partial likelihood estimator for Cox's model, see (2.2). Note that even though in this case d_i is not predictable due to $\hat{\boldsymbol{\beta}}$, the following asymptotics still hold.

If the Aalen Model (2.3.2) is the correct model and some additional assumptions hold then $\hat{T}(t)$ converges to a mean zero Gaussian martingale whose variance can be estimated consistently by

$$[\hat{T}](t) = \frac{1}{n}\int_0^t \mathbf{d}^\top(s)\mathbf{P}(s)\ \mathrm{diag}(\,\mathrm{d}\boldsymbol{N}(s))\mathbf{P}(s)\mathbf{d}(s),$$

where $\mathrm{diag}(\,\mathrm{d}\boldsymbol{N}(s))$ is a diagonal matrix with entries $\mathrm{d}N_i(s), i \in \{1, ..., n\}$. The simplest test statistic that can be constructed based on this asymptotic behavior is

$$T := \frac{1}{\sqrt{[\hat{T}](\tau)}}\hat{T}(\tau)$$

which converges as $n \to \infty$ in distribution to a standard normal random variable. With the choice of $\mathbf{d}$ given by (7.1), Gandy & Jensen (2005) showed that a test that rejects for large values of T is consistent against Cox's model.

The second test decides whether a Cox model with change-point is more adequate than a Cox model without a change-point. We use tests which were developed by Gandy & Jensen (2006) for an extended version of a Cox-type regression model $\lambda_i(t) = \lambda_0(t)\rho_i(\boldsymbol{\beta}, t)$, where $\lambda_0(t)$ is an unknown baseline and $\rho_i(\boldsymbol{\beta}, t)$ is an observable stochastic process which may depend on a finite-dimensional parameter vector $\boldsymbol{\beta}$. In our case we have just the basic Cox model with time-dependent covariates as a null hypothesis such that $\rho_i(\boldsymbol{\beta}, t) = R_i(t)\exp\{\boldsymbol{\beta}^\top \boldsymbol{Z}_i(t)\}$.

The test is based on sums of weighted martingale residuals and the test statistic is given by

$$T(c(\hat{\boldsymbol{\vartheta}}, \cdot)) = n^{-1/2}\sum_{i=1}^{n}\int_0^\tau c_i(\hat{\boldsymbol{\vartheta}}, s)\,\mathrm{d}N(s).$$

The weights $c_i(\cdot, \cdot)$ are chosen such that a simple asymptotic distribution can be derived and secondly such that the test is powerful against certain alternatives, which are called competing models. Therefore, the test statistic does not only contain the parameter of the null model but also that of the competing model. The parameter vector $\boldsymbol{\vartheta} := (\boldsymbol{\beta}^\top, \boldsymbol{\gamma}^\top)^\top$, where $\boldsymbol{\beta}$ is a parameter vector of the null model and $\boldsymbol{\gamma}$ a parameter vector of the competing model, is estimated by the maximum likelihood estimator $\hat{\boldsymbol{\vartheta}} = (\hat{\boldsymbol{\beta}}^\top, \hat{\boldsymbol{\gamma}}^\top)^\top$. In our case we consider the null hypothesis

$$H_0 : \lambda_i(t) = \lambda_0(t)\rho_i(\boldsymbol{\beta}, t).$$

The competing models are given by

$$\lambda_i(t) = a(t)h_i(\boldsymbol{\gamma}, t),$$

where $a(t)$ is an unspecified baseline, $\boldsymbol{\gamma}$ is an unknown parameter vector and the stochastic processes $h_i(\boldsymbol{\gamma}, t), i = 1, ..., n$ are observable. In our case we have the basic Cox model $\rho_i(\boldsymbol{\beta}, t) = R_i(t)\exp\{\boldsymbol{\beta}^\top \boldsymbol{Z}_i\}$ as a null hypothesis and the change-point model $h_i(\boldsymbol{\gamma}, t) = R_i(t)\exp\{\boldsymbol{\gamma}^\top \tilde{\boldsymbol{Z}}_i(t, \boldsymbol{\xi})\}$ as a competing model. Under the null hypothesis (and some mild technical conditions) the test statistic is asymptotically normal:

$$T(c(\hat{\boldsymbol{\vartheta}}, t)) \xrightarrow{d} N(0, \sigma^2),$$

where σ^2 can be estimated consistently by $\hat{\sigma}^2(c) = n^{-1}\sum_{i=1}^n \int_0^\tau c_i^2(\hat{\boldsymbol{\vartheta}}, s)\,\mathrm{d}N_i(s)$. Simulation studies show that this test performs well even for moderate sample sizes. For the explicit choice of the weights and all further details we refer to Gandy & Jensen (2006).

7.2 Martingale Residuals and Functional Form in the Cox Model

The Cox model heavily relies on the functional form of the covariates $\mathbf{Z}$. In applications it is not clear whether one of the covariates, say X, should better be included in a different functional form like X^2 or $\log X$. Therneau et al. (1990) suggested to use martingale residuals to determine the functional form of covariates graphically. Arguing differently we derive similar results.

We consider only one individual and drop the index i. Let X and $\boldsymbol{Z}$ be stochastically independent random covariates constant over time. We assume that the counting process N admits the following intensity

$$\lambda(t) = h(X)\exp\{\boldsymbol{\beta}^\top \boldsymbol{Z}\}R(t)\lambda_0(t) = h(X)\lambda^*(t),$$

where h is an unknown positive function. Hence,

$$M(t) = N(t) - \int_0^t h(X)\lambda^*(s)\,\mathrm{d}s = N(t) - h(X)\Lambda^*(t)$$

is a mean zero martingale. Forming conditional expectation with respect to X we get

$$E[M(t)|X] = E[N(t)|X] - h(X)E[\Lambda^*(t)|X].$$

Since this is again a mean zero martingale we set, heuristically, $E[M(\tau)|X]$ equal to zero and get

$$h(X) \approx \frac{E[N(\tau)|X]}{E[\Lambda^*(\tau)|X]} = \left(1 - \frac{E[N(\tau) - \Lambda^*(\tau)|X]}{E[N(\tau)|X]}\right)^{-1}.$$

In particular, we are interested in $f(X) := \log h(X)$. Using a first order Taylor expansion we get

$$f(X) \approx -\log\left(1 - \frac{E[N(\tau) - \Lambda^*(\tau)|X]}{E[N(\tau)|X]}\right) \approx \frac{E[N(\tau) - \Lambda^*(\tau)|X]}{E[N(\tau)|X]}.$$

Treating $c = E[N(\tau)|X]$ as constant it remains to estimate $E[N(\tau) - \Lambda^*(\tau)|X]$ for which we use the martingale residuals

$$\hat{M}(\tau) = N(\tau) - \int_0^\tau R(s)\exp\{\hat{\boldsymbol{\beta}}^\top \boldsymbol{Z}\}\,\mathrm{d}\hat{\Lambda}_0(s)$$

resulting from the Cox model ignoring X. To do so we smooth a scatterplot of $\hat{M}_i(\tau)$ against X_i via robust locally weighted regression (see Cleveland (1979)). To sum up, plotting the martingale residuals against X should give an idea of the functional form of X. A linear scatterplot indicates that no further transformation of X is necessary. We have carried out several simulation studies, which supported the validity of this heuristic method.

7.3 Insurance Dataset

The dataset we are considering stems from a German insurance company and contains information about private accident insurance contracts. Generally, in survival analysis the time to death of an individual or the time to failure of a technical system is examined. We investigate the time from the conclusion until the cancellation of a contract. Our dataset does not only consist of information about the time to cancellation. There are several other attributes given about the insurance holder and the person insured: age, number of persons insured, amount of the annual premium, insurance sums covering death or disablement, etc. Our main goal is to investigate in which way the attributes

influence the cancellation of contracts. The dataset consists of more than 100 000 private accident insurance contracts. Special features of these contracts are that more than one person can be insured in a contract and that the insurance holder does not have to be insured in it. There also exist aggregated covariates like the average insurance sum per insured person in each contract. In total, each contract offers about 70 attributes. For our analysis we have transformed some attributes into numerical covariates and deleted some due to too small frequency, e.g. the covariate which is 1 if the premium is paid in advance appeared only once. The cancellation of a contract could only be observed during the period of May 1st, 2002 to April 30th, 2003. About 91 percent of all contracts were censored meaning that they were not canceled during this period. Since this dataset is quite big, we reduced our analysis to a smaller dataset. There we only contemplated contracts belonging to insurance holders working in similar professions which were 31298 contracts. In our analysis we focused on 43 covariates since some of the covariates were redundant. In the smaller dataset in the first, second, third and ninth year no contracts have been canceled. Furthermore, there are only few contracts which have a duration longer than 30 years. The longest duration of a contract is given by 44 years.
The models we use to examine the data are the Cox model, the Aalen model and the Cox model with change-points.

7.3.1 Results

The conclusions we want to present are drawn from the smaller dataset containing 43 covariates. All computations were done in `SAS` and `R`.
First we analyzed our dataset by using two different variable selection methods to exclude the least significant covariates. Here we confine ourselves to the forward selection method since the backward selection method has produced similar results. Conducting the forward selection method in the Cox model we first estimate parameters for covariates forced into the model (see Krall et al. (1975)). Then we compute adjusted χ^2-statistics for each covariate and examine the largest of these statistics. If it is significant at a 5 percent level the corresponding covariate is added to the model and stays in the model in all the following steps. In the Aalen model we include as a first covariate in the forward selection the baseline covariate, i.e. the covariate which is 1 for all contracts under risk. Then we test the hypothesis that the Aalen model (2.3.2) holds as described in Section 7.1 against the hypothesis that there exists an additional covariate, i.e. we test against all other variables and include the covariate having the smallest p-value into the model. We stop our selection when the remaining covariates are not significant on the 5 percent level.

Covariate Description	Parameter Estimate	Standard Error
first insured person is female	0.08593	0.04808
paying the premium every 6 months	0.23833	0.05627
paying the premium every month	0.19082	0.04340
paying the annual premium by direct debit	-0.43466	0.05085
executive employee	-0.26966	0.11709
employee	-0.10646	0.04210
standardized single insurance	-0.24778	0.05623
risk group B of first ins. person	-0.18729	0.05398
dynamic in the contract	0.21613	0.04258
insurance holder equals first person insured	-0.22325	0.05460
insurance sum for disability of first ins. person	-5.193E-6	1.024E-6
risk premium of first insured person	0.00284	0.00036
age of the insurance holder	-0.01330	0.00173
average accident benefits per person insured	-0.000365	0.000102
average daily benefits per person insured	-0.03044	0.00657
average hospital daily benefits per person insured	-0.00551	0.00156
number of adults insured	0.26230	0.05245

Table 1: Parameter estimation after forward selection in the Cox model

The analysis of our dataset yields nearly the same significant covariates by using the forward selection method in both models. In the Cox model as well as in the Aalen model the forward selection methods suggest to include 17 covariates into the models, see Table 1 for the covariates in the Cox model. Except for the first and the fifth covariate all p-values are less than 0.01. The first ten covariates displayed in Table 1 are categorical covariates taking only values 1 or 0. One may be tempted to compare the influence of certain covariates by the value of their parameter estimator. This may yield a false conclusion because of the different values of the covariates.
Variables like the risk premium, the insurance sums, paying with direct debit show effects as one would have expected. For example, a higher risk premium leads to an increasing churn rate. The intensity of a contract being canceled declines as the insurance sum grows. Furthermore, insurance holders paying with direct debit are less likely to cancel their contracts. A closer look at the martingale residuals, following the procedure described in Section 7.2, reveals that we obtain a nearly linear smoothed scatterplot for all investigated covariates that are not 0-1 variables except for the one indicating the age of the insurance holder. Recall that a linear smoothed scatterplot provides evidence that the corresponding covariate has been introduced into the model adequately. The smoothed plots of the martingale residuals against the variable age of the insurance holder and against the

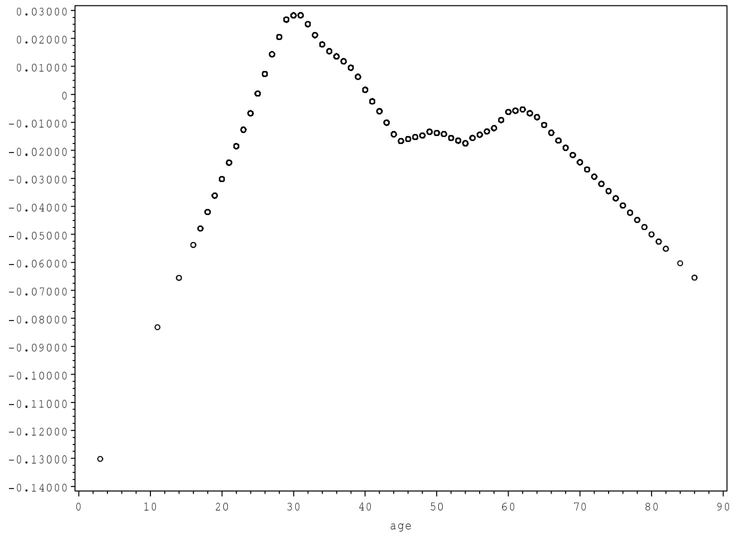

Figure 7.1: Martingale Residual plot of the covariate age

variable for average accident benefits are given in the Figures 7.1 and 7.2. As Figure 7.1 shows, the plot of the residuals of the covariate age is nonlinear. Therefore, we fit our extended model with change-points to the data. The estimated change-points are at the age of 29.5, 45.7 and 59.5. Furthermore, the influence before the first change-point (i.e. β_2) is positive and the influence after the first change-point (i.e. $\beta_2+\beta_3$) is negative. This means that the intensity of cancellation for insurance holders increases with age up to the first change-point and declines afterwards until the second change-point. As an empirical affirmation, we judged by the partial likelihood that the model with change-points has a better fit than the original model.

The Aalen model provides nearly the same trends of the variables as those indicated by the Cox model. Whenever the parameter estimate of β for a covariate is positive (negative) in the Cox model, then the estimated integrated intensity $\hat{B}(t)$ of this covariate is increasing (decreasing). This can be seen for example in Figure 7.3. There the estimated integrated intensity of the covariate for paying the annual premium by direct debit is plotted with its pointwise confidence intervals. Testing goodness-of-fit of the Aalen model against the Cox model using the test explained in Section 7.1 the hypothesis that Aalen's model is the true model is rejected (p-value< 0.001).

To sum up the results, we can state that several covariates have been found to be of significant influence. They have been identified by using forward and backward selection methods. The influence of the covariates can be interpreted in a reasonable way. Even in the bigger dataset we are able to observe similar parameter estimates for the covariates in the Cox model. But further investigations are needed to reveal the effect of the same

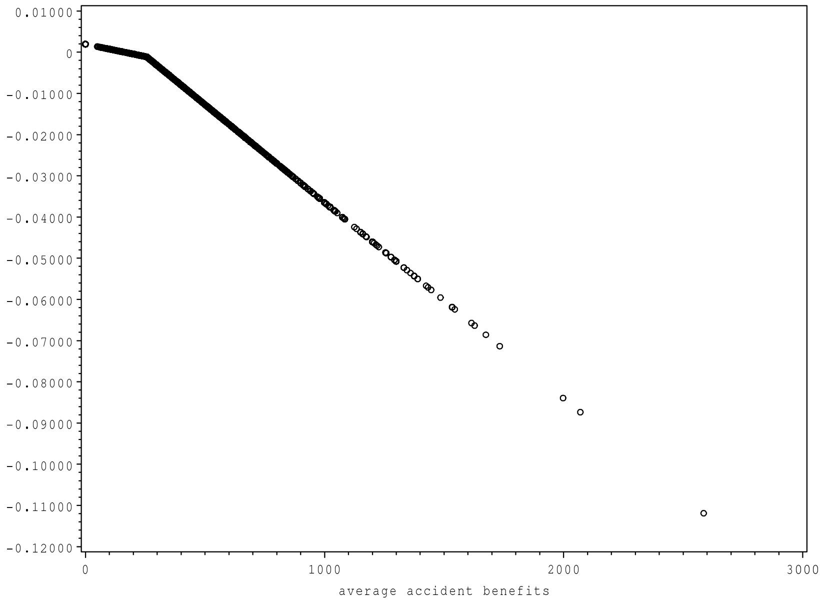

Figure 7.2: Martingale Residual plot of the average accident benefits

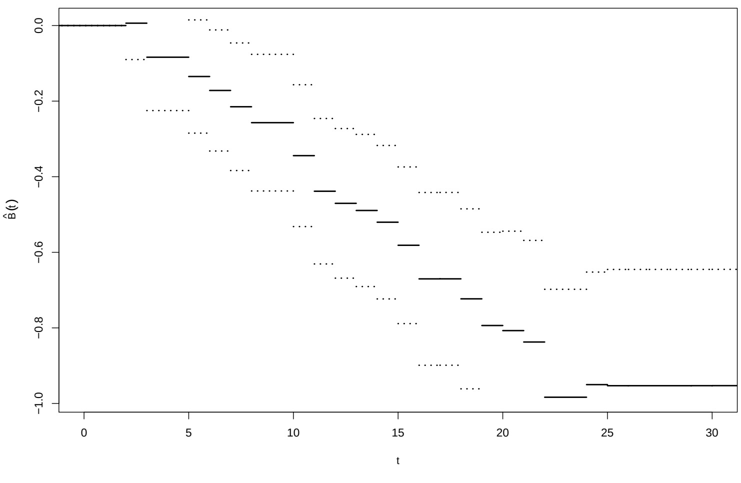

Figure 7.3: Estimated integrated intensity $\hat{B}(t)$ of the covariate for paying the annual premium by direct debit

occurrence times of several events in both models. Furthermore, our analysis shows that the functional form of one of the covariates seems to be misspecified. Therefore, it is reasonable to use the Cox model with change-points.

7.3.2 Electric Motor Dataset

As part of a DFG (Deutsche Forschungsgemeinschaft) research project long term experiments with electric motors were conducted. A total of 200 sample objects were observed on a test bed and failure- and censoring times as well as the covariates load, current, nominal voltage and r.p.m. were recorded. The goal of the analysis was to quantify the influence of the different covariates on the survival times. By observing the smoothed martingale residuals a change-point in nominal voltage seems to exist, see Figure 7.4. Fitting our change-point model to the data and using the goodness-of-fit test of Section 7.1 a change-point at 18V is suggested. The covariates load, current and r.p.m seem to

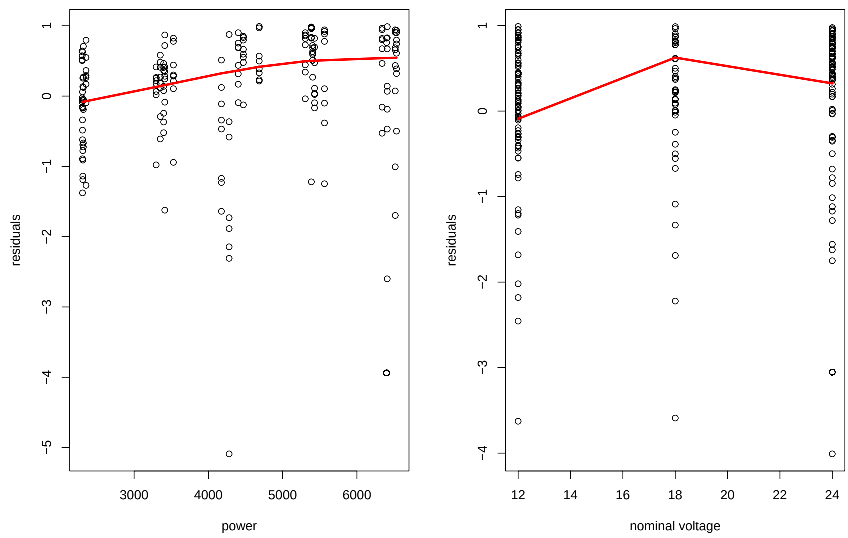

Figure 7.4: Martingale Residuals

have a linear influence. Now a predicted survival function can be calculated with its 95 % confidence intervals, see Figure 7.5.

null hypothesis	competing model	p-value
Model without CP	Model CP in nominal voltage	0.0001

Table 2: Table contains the p-value calculated based on Section 7.1

For the validity of the model we estimated the survival function based on our model out of two different datasets and compared the functions in a plot. One dataset consists of all data from 12V and 24V electric motors. Based on these data we estimate the survival function for a covariate value of 18 V. On the other hand we determine the survival function for a covariate value of 18 V out of the dataset which only contains data from 18V electric motors. In Figure 7.6 we compare the two functions. It can be seen that the

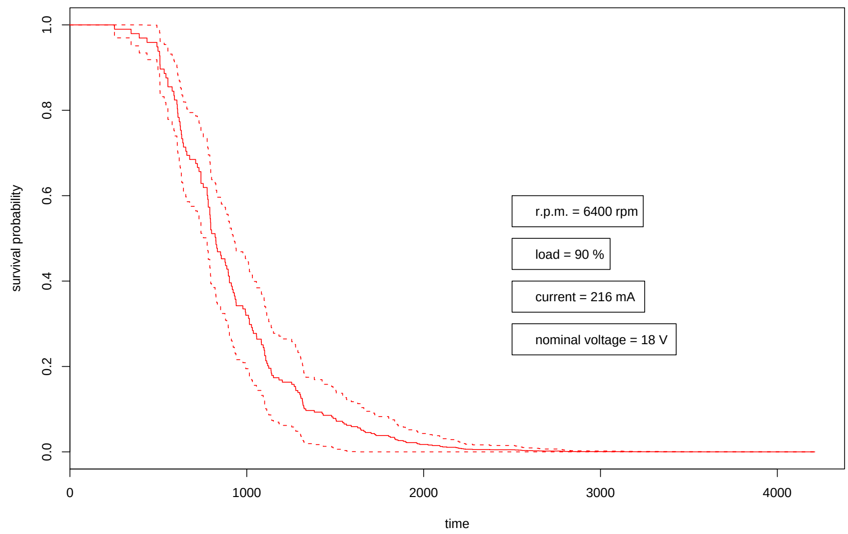

Figure 7.5: Predicted survival function for an 18 V electric motor

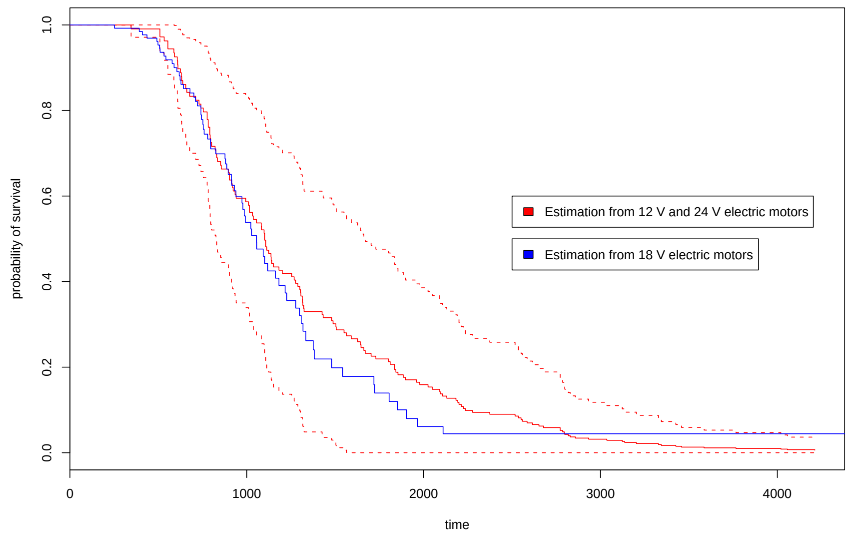

Figure 7.6: Validation

survival function based on 12V and 24V electric motors is inside the pointwise confidence intervals of the survival curve of 18V electric motors. This suggests that the estimation with the Cox model with change-points is close to the survival curve for other covariate values.

7.3.3 PBC Dataset

The third dataset we looked at is the well known PBC dataset described in Fleming & Harrington (1991). It contains survival data of 312 patients with primary biliary cirrhosis. We use the corrected dataset given by Fleming & Harrington (1991), p. 81.
First they developed a Cox model which included the covariates age, albumin, bilirubin, edema, hepatomegaly and prothrombin time. Hereby, the covariate albumin describes the amount of a certain protein in the blood, bilirubin is the level of a liver bile pigment, edema is an indicator for the presence of a swelling or enlargement of the liver, and prothrombin time is the amount of time it takes the blood sample to begin coagulation in a certain laboratory test. To get a better fit they used model selection methods and transformed some covariates. At the end they suggested to use the following covariates in the model: age, edema, log(albumin), log(bilirubin) and log(prothrombin time). After considering the martingale residual plots we suggest to use a different model with a change-point in the covariate bilirubin instead of log(bilirubin). A change-point is obtained at a level of 3. Using the directed goodness-of-fit tests for Cox type regression models described in Section 7.1 with the model of Fleming & Harrington (Model F&H) as null hypothesis and the change-point model (Model CP) as alternative we get a clear rejection of the model of Fleming & Harrington.

null hypothesis	competing model	p-value
Model F& H	Model CP	0.0088

Table 1: Table contains the p-value calculated based on Section 7.1 with respect to the PBC dataset

Chapter 8

Conclusions and Remarks

We introduce new Cox-type regression and transformation models with change-points according to covariate thresholds. For the Cox-type regression model with bent-line change-points in the underlying regression function we proposed an estimation procedure and we proved that the regression parameters and change-point parameters are $\sqrt{n}$-consistent and asymptotically normal. Furthermore, we applied this model to different data sets and showed that using a goodness-of-fit test the new model is superior compared to the classical Cox model.
Moreover, this model enables us to study the unknown functional form of different covariates. In practice, the true nature of the functional form of a covariate is often opaque. Usually, martingale residuals are used for analyzing the functional form as plotted in Figure 7.1. But it is still a problem to select the most accurate function by simply looking at these plots. In fact, we suggest that a piecewise linear functional form of the covariate often fits the data in a better way. Therefore, we recommend to proceed as follows:
Starting with the classical Cox model, change-points can be added successively to the model, resulting in a piecewise linear functional form. With each new change-point a goodness-of-fit test must be made to check whether the assumption that another change of influence occurs is accurate. The procedure ends when the last goodness-of-fit test is rejected.
A further generalization of the model can be made if the exponential function is replaced by a general known risk function, for example $1+x$, $x>0$. In Chapter 5 it is shown that the same asymptotic properties as in the usual case with the exponential function still hold.
To cover even more survival time models we introduced the transformation model with change-points according to covariate thresholds. We showed that in this model the finite and infinite-dimensional parameters can be estimated $\sqrt{n}$-consistently and that they are asymptotically normal. Although Kosorok & Song (2007) stated that our model is a submodel of theirs, we have to recognize the two models as different when the rate of

convergence and the asymptotic properties of the change-point estimates are concerned. The study of the transformation model presented in Chapter 6 is of interest for applications since it enables us to insert bent-line change-points in frailty models.
However, adapting the transformation model to data is a task of noticeable complexity. For the proportional hazards and the proportional odds model Kosorok & Song (2007) conducted simulation studies to justify their approach. A similar analysis would also be conceivable for our model.
Another interesting extension could be the multiplicative-additive transformation model with survival function

$$S_{\boldsymbol{Z}}(t) = \Lambda\left(\int_0^t \exp\{\beta(Z_1 - \xi)^+\}\,\mathrm{d}A(u) + \int_0^t \boldsymbol{\gamma}^T \boldsymbol{Z}_2(u)\,\mathrm{d}u\right).$$

Using the same approach as in our case seems reasonable at first sight but turns out to be difficult because of the additive term.
Summarizing, we suggest new survival time models with bent-line change-points which are superior compared to the classical Cox model. Furthermore, the new models combined with a goodness-of-fit test give an analytical impression of the functional form of the covariates. Namely, the functional form is modeled as piecewise linear by the procedure described above. Many reasonable applications for our model exist. For example, in biological settings bent-line change-points are in many cases more realistic than complex nonlinear or jump effects. This was already noted by Chapell (1989). Various other applications where a piecewise linear functional form yields a better fit to the given data than the classical approach were studied in Chapter 8

Appendix A

Some Results from the Theory of Empirical Processes

In this chapter we collect some results from the theory of empirical processes used in this thesis. Most definitions and theorems can be found in Van der Vaart & Wellner (1996) and Van der Vaart (1998).

A.1 Empirical Process

The empirical measure $\mathbb{P}_n$ of a sample of random elements $X_1, \dots, X_n$ on a measurable space $(\mathcal{X}, \mathcal{A})$ is the discrete random measure given by $\mathbb{P}_n(C) = n^{-1}\#(1 \leq i \leq n : X_i \in C)$. If the points are measurable, it can be described as the random measure, that puts mass $1/n$ at each observation. The empirical measure can be written in the form $\mathbb{P}_n = n^{-1}\sum_{i=1}^n \delta_{X_i}$ of the dirac measures at the observation. Given a collection $\mathcal{F}$ of measurable functions $f : \mathcal{X} \to \mathbb{R}$, the empirical measure induces a map from $\mathcal{F}$ to $\mathbb{R}$ given by $f \mapsto \mathbb{P}_n f$. If P is the common distribution of the X_i, then the centered and scaled version of the given map is the $\mathcal{F}$-indexed empirical process $\mathbb{G}_n$ given by

$$f \mapsto \mathbb{G}_n f = \sqrt{n}(\mathbb{P}_n - P)f = \frac{1}{\sqrt{n}}\sum_{i=1}^n (f(X_i) - \mathrm{P}\, f),$$

where $\mathrm{P}\, f = \int f \, \mathrm{d}\,\mathrm{P}$ for a signed measure P.

A.2 Measurability

The classical theory of weak convergence requires that the random elements are Borel measurable. This measurability usually holds when we consider a separable metric space such as $\mathbb{R}^k$ with the supremum metric. This requirement can fail when the metric space is not separable. One example is the Skorohod space $D[0,1]$ of all right-continuous functions on $[0,1]$ with left hand limits endowed with the metric induced by the supremum norm. One approach to deal with this difficulty was made by Billingsley (1968). He endowed $D[0,1]$ with the Skorohod metric under which $D[0,1]$ is separable and complete. In this context the extension of the strong law of large numbers introduced by Andersen & Gill (1982) has to be considered. They extended the strong law of large numbers to cadlag processes in Banach spaces. Let $D_B[0,\tau]$ the space which contains right continuous functions on $[0,1]$ with left hand limits taking values in a separable Banach space B.

Theorem A.1. *Let $X_1, X_2, \ldots$ be i.i.d. random elements of $D_B[0,\tau]$ (endowed with the Skorohod topology). If $E\sup_{t\in[0,\tau]}\|X_1(t)\| < \infty$, then*

$$\sup_{t\in[0,\tau]}\left\|\frac{1}{n}\sum_{i=1}^{n} X_i(t) - EX_1(t)\right\| \to 0 \quad \textit{almost surely.}$$

Another approach concerning the weak convergence put forward by Hoffmann-Jørgensen and which is described in detail in Van der Vaart & Wellner (1996) is to drop the requirement of Borel measurability of each X_n, meanwhile upholding the requirement $Ef(X_n) \to Ef(X)$, for all function f in the set of all bounded, continuous, real functions on a metric space. The expectations are now to be interpreted as outer expectations and the X_n may be arbitrary maps. Denote the extended real line by $\bar{\mathbb{R}} = \mathbb{R} \cup \{-\infty, \infty\}$.

Definition A.1. (outer expectation and outer probability)
Let $(\Omega, \mathcal{A}, \mathrm{P})$ be an arbitrary probability space and $Z : \Omega \mapsto \bar{\mathbb{R}}$ an arbitrary map. The outer expectation of Z with respect to P is defined as

$$E^*Z = \inf\left\{EU : U \geq Z, U : \Omega \mapsto \bar{\mathbb{R}} \, \textit{measurable}, EU \textit{ exists}\right\}.$$

The outer probability of a set $B \subset \Omega$ is

$$\mathrm{P}^*(B) = E^* I_B.$$

Inner expectation and inner probability can be defined in a similar way. They can also be defined by $E_*Z = -E^*(-Z)$ and $\mathrm{P}_*(B) = 1 - \mathrm{P}^*(\Omega - B)$, respectively.
We use the approach of Billingsley (1968) in Chapter 4 and Chapter 5 and the one of Hoffmann-Jørgensen in Chapter 6.

A.3 Entropy Numbers

One of the main concerns in the theory of empirical processes is to obtain results about the convergence of

$$\sup_{f\in\mathcal{F}} |\mathbb{P}_n f - \mathrm{P}\, f|$$

where $\mathcal{F}$ is a class of functions. Therefore, it is essential to measure the size of the class $\mathcal{F}$. A relatively simple way is the use of entropy numbers.
Let $(\mathcal{F}, \|\cdot\|)$ be a subset of a normed space of real functions $f : \mathcal{X} \mapsto \mathbb{R}$ on some set.

Definition A.2. (Envelope function)
An envelope function of a class $\mathcal{F}$ is any function $x \mapsto F(x)$ such that $|f(x)| \leq F(x) < \infty$, for every x and f.

Note that, the minimal envelope function is $x \mapsto \sup_f |f(x)|$.

Definition A.3. (Covering numbers)
The covering number $N(\epsilon, \mathcal{F}, \|\cdot\|)$ is the minimal number of balls $\{g : \|g - f\| < \epsilon\}$ of radius ϵ needed to cover the set $\mathcal{F}$. The entropy is defined as the logarithm of the covering number.

Definition A.4. (Bracketing numbers)
Given two functions l and u, the bracket $[l, u]$ is the set of all functions f with $l \leq f \leq u$. An ϵ-bracket in $L_r(\mathrm{P})$ is a bracket $[l, u]$ with $\mathrm{P}(u - l)^r < \epsilon$. The bracketing number $N_{[]}(\epsilon, \mathcal{F}, L_r(\mathrm{P}))$ is the minimal number of ϵ-brackets needed to cover $\mathcal{F}$. The entropy with bracketing is the logarithm of the bracketing number.

Using the notion of Vapnik-Červonenkis classes of sets it is possible to derive upper bounds for uniform covering numbers, which are needed to show that a class is Glivenko-Cantelli.

Definition A.5. (Vapnik-Červonenkis class)
For a collection of subsets $\mathcal{C}$ of a set $\mathcal{X}$, and points $x_1, \ldots, x_n \in \mathcal{X}$, define

$$\Delta_n^{\mathcal{C}}(x_1, \ldots, x_n) = \#\{C \cap \{x_1, \ldots, x_n\} : C \in \mathcal{C}\},$$

so that $\Delta_n^{\mathcal{C}}(x_1, \ldots, x_n)$ is the number of subsets of $\{x_1, \ldots, x_n\}$ picked out by the collection $\mathcal{C}$. Define moreover,

$$m^{\mathcal{C}}(n) = \max_{x_1,\ldots,x_n\in\mathcal{X}} \Delta_n^{\mathcal{C}}(x_1, \ldots, x_n)$$

and

$$V(\mathcal{C}) = \inf\{n \geq 1 : m^{\mathcal{C}}(n) < 2^n\}.$$

We call $V(\mathcal{C})$ the index of the class $\mathcal{C}$. The collection $\mathcal{C}$ is a Vapnik-Červonenkis class (VC-class) if $V(\mathcal{C}) < \infty$.

Definition A.6. (Vapnik-Červonenkis class of functions)
The subgraph of a function $f : \mathcal{X} \mapsto \mathbb{R}$ is the subset of $\mathcal{X} \times \mathbb{R}$ given by

$$\{(x,t) \in \mathcal{X} \times \mathbb{R} : t < f(x)\}.$$

For a class of functions $\mathcal{F}$, let $V(\mathcal{F})$ be the index of the collection of subgraphs. A collection of functions $\mathcal{F}$ is called a Vapnik-Červonenkis subgraph class if $V(\mathcal{F}) < \infty$.

Theorem A.2. *(Van der Vaart & Wellner (1996), Theorem 2.6.7)*
For a Vapnik-Červonenkis subgraph class $\mathcal{F}$ with measurable envelope function F and $r \geq 1$, one has for any probability measure Q *with $\|F\|_{\mathrm{Q},r} = \left(\int |F|^r \, \mathrm{d}Q\right)^{1/r} > 0$,*

$$N(\epsilon \|F\|_{\mathrm{Q},r}, \mathcal{F}, L_r(\mathrm{Q})) \leq K V(\mathcal{F}) (16e)^{V(\mathcal{F})} \left(\frac{1}{\epsilon}\right)^{r(V(\mathcal{F})-1)},$$

for a universal constant K and $0 < \epsilon < 1$.

A.4 Lipschitz Functions and Helly's Lemma

For classes of functions $x \mapsto f_t(x)$ that are Lipschitz in the index parameter $t \in T$ the following holds for every x:

$$|f_s(x) - f_t(x)| \leq d(s,t) F(x),$$

where d is some metric on the index set and F is a function on the sample space. Then the diameter of T times F is an envelope function for the class $\{f_t - f_{t_0} : t \in T\}$ for any fixed t_0.

Theorem A.3. *(Van der Vaart & Wellner (1996), Theorem 2.7.11)*
Let $\mathcal{F} = \{f_t : t \in T\}$ be a class of functions satisfying the preceding display for every s and t and some fixed function F. Then for any norm $\|\cdot\|$,

$$N_{[]}(2\epsilon \|F\|, \mathcal{F}, \|\cdot\|) \leq N(\epsilon, T, d).$$

A quite useful theorem which we needed in several proofs in Chapter 6 is Helly's lemma and corollaries from it. We present here the lemma as stated in Van der Vaart (1998).

Theorem A.4. *(Van der Vaart (1998), Lemma 2.5)*
Each given sequence F_n of cumulative distribution functions on $\mathbb{R}^k$ possesses a subsequence F_{nj} with the property that $F_{nj}(x) \to F(x)$ at each continuity point x of a possibly defective distribution function F.

A.5 Glivenko-Cantelli and Donsker Classes

By the law of large numbers, the sequence $\mathbb{P}_n f$ converges almost surely to $\mathrm{P}\, f$, for every f such that $\mathrm{P}\, f$ is defined. The abstract Glivenko-Cantelli theorems make this result uniform in f ranging over a class of functions.

Definition A.7. (Glivenko-Cantelli class)
A class $\mathcal{F}$ of measurable functions $f : \mathcal{X} \mapsto \mathbb{R}$ is called P-*Glivenko-Cantelli if*

$$\|\mathbb{P}_n f - \mathrm{P}\, f\|_{\mathcal{F}} = \sup_{f\in\mathcal{F}} |\mathbb{P}_n f - \mathrm{P}\, f| \to 0, \quad \textit{almost surely.}$$

Theorem A.5. *(Van der Vaart (1998), Theorem 19.4)*
Every class $\mathcal{F}$ of measurable functions such that $N_{[]}(\epsilon, \mathcal{F}, L_1(\mathrm{P})) < \infty$ for every $\epsilon > 0$ is P-*Glivenko-Cantelli.*

Definition A.8. (Donsker class)
A class $\mathcal{F}$ of measurable functions $f : \mathcal{X} \mapsto \mathbb{R}$ is called P-*Donsker if the sequence of processes $\mathbb{G}_n f : f \in \mathcal{F}$ converges in distribution to a tight limit process in the space l^∞.*

The abstract Donsker theorem is a uniform version of the central limit theorem.

Theorem A.6. *(Van der Vaart (1998) Theorem 19.5)*
Every class $\mathcal{F}$ of measurable functions with $J_{[]}(1, \mathcal{F}, L_2(\mathrm{P})) < \infty$ is P-*Donsker.*

A.6 Finite Entropy Integrals

In this section we derive bounds on moments for the supremum $\|\mathbb{G}_n\|_{\mathcal{F}}$ of the empirical process for classes $\mathcal{F}$ that possess a finite uniform-entropy or bracketing entropy integral. For a class of functions $\mathcal{F}$ with envelope function F and $\delta > 0$, let the uniform-entropy integral be

$$J(\delta, \mathcal{F}) = \sup_Q \int_0^\delta \sqrt{1 + \log N(\epsilon\|F\|_{Q,2}, \mathcal{F}, L_2(Q))}\, \mathrm{d}\epsilon,$$

where the supremum is over all discrete probability measures Q with $\|F\|_{Q,2} > 0$. Furthermore, define the $L_2(\mathbb{P}_n)$-seminorm by

$$\|f\|_n = \sqrt{\frac{1}{n}\sum_{i=1}^n f^2(X_i)}.$$

The symbol $\lesssim$ indicates that the left side is bounded by a constant times the right side.

Theorem A.7. *(Van der Vaart & Wellner (1996), Theorem 2.14.1)*
Let $\mathcal{F}$ be a P*-measurable class of functions with measurable envelope function F. Then*

$$\| \|\mathbb{G}_n\|_{\mathcal{F}}^* \|_{\mathrm{P},p} \lesssim \| J(\rho_n, \mathcal{F}) \|F\|_n \|_{\mathrm{P},p} \lesssim J(1, \mathcal{F}) \|F\|_{\mathrm{P},2\vee p}, \quad p \geq 1.$$

Here $\rho_n = (\sup_{f\in\mathcal{F}} \|f\|_n^*)/\|F\|_n$ where $\|\cdot\|_n$ is the $L_2(\mathbb{P}_n)$-seminorm and the inequalities are valid up to constants depending only on p. In particular, when $p = 1$

$$E\|\mathbb{G}_n\|_{\mathcal{F}}^* \lesssim E\left[J(\rho_n, \mathcal{F})\|F\|_n\right] \lesssim J(1, \mathcal{F})\|F\|_{\mathrm{P},2}.$$

For a given norm $\|\cdot\|$, define a bracketing integral of a class of functions $\mathcal{F}$ as

$$J_{[]}(\delta, \mathcal{F}, \|\cdot\|) = \int_0^\delta \sqrt{1 + \log N_{[]}(\epsilon\|F\|, \mathcal{F}, \|\cdot\|)}\, \mathrm{d}\epsilon.$$

For most classes of interest, the bracketing numbers $N_{[]}(\epsilon, \mathcal{F}, L_r(\mathrm{P}))$ grow to infinity as $\epsilon \downarrow 0$. A sufficient condition for a class to be Donsker is that they do not grow too fast. The speed can be measured in terms of the bracketing integral described above. The integrand in the integral is a decreasing function of ϵ. Hence, the convergence of the integral depends only on the size of the bracketing numbers for $\epsilon \downarrow 0$. Roughly speaking, the integral condition requires that the entropies grow of slower order than $(1/\epsilon)^2$.

Theorem A.8. *(Van der Vaart & Wellner (1996), Theorem 2.14.2)*
Let $\mathcal{F}$ be a class of measurable functions with measurable envelope function F. For a fixed $\eta > 0$ define

$$\alpha(\eta) = \frac{\eta\|F\|_{\mathrm{P},2}}{\sqrt{1 + \log N_{[]}(\eta\|F\|_{\mathrm{P},2}, \mathcal{F}, L_2(\mathrm{P}))}}.$$

Then, for every $\eta > 0$,

$$\begin{aligned} E^*\|\mathbb{G}_n\|_{\mathcal{F}} \lesssim & J_{[]}(\eta, \mathcal{F}, L_2(\mathrm{P}))\|F\|_{\mathrm{P},2} + \sqrt{n}\,\mathrm{P}\, F I_{\{F > \sqrt{n}\alpha(\eta)\}} \\ & + \| \|f\|_{P,2} \|_{\mathcal{F}} \sqrt{1 + \log N_{[]}(\eta\|F\|_{\mathrm{P},2}, \mathcal{F}, L_2(\mathrm{P}))}. \end{aligned}$$

If $\|f\|_{P,2} < \delta\|F\|_{P,2}$ for every $f \in \mathcal{F}$, then taking $\eta = \delta$ in the last display yields

$$E^*\|\mathbb{G}_n\|_{\mathcal{F}} \lesssim J_{[]}(\delta, \mathcal{F}, L_2(\mathrm{P}))\|F\|_{\mathrm{P},2} + \sqrt{n}\,\mathrm{P}\, F I_{\{F > \sqrt{n}\alpha(\delta)\}}$$

Hence, for any class $\mathcal{F}$,

$$E^*\|\mathbb{G}_n\|_{\mathcal{F}} \lesssim J_{[]}(1, \mathcal{F}, L_2(\mathrm{P}))\|F\|_{\mathrm{P},2}$$

A.7 Limit Theorems

The following theorems are concerned with the asymptotic normality of the estimators. Theorem A.9 is used in Chapter 4 and Chapter 5, whereas Theorem A.10 is applied in Chapter 6.

The following theorem concerns M-estimators defined as maximizers of a criterion function $\boldsymbol{\theta} \mapsto \mathbb{P}_n m_{\boldsymbol{\theta}}$, which are assumed to be consistent for a point of maximum $\boldsymbol{\theta}_0$ of the function $\boldsymbol{\theta} \mapsto \mathrm{P}\, m_{\boldsymbol{\theta}}$.

Theorem A.9. *(Van der Vaart (1998), Theorem 5.23)*
For each $\boldsymbol{\theta}$ in an open subset of Euclidean space let $x \mapsto m_{\boldsymbol{\theta}}(x)$ be a measurable function such that $\boldsymbol{\theta} \mapsto m_{\boldsymbol{\theta}}(x)$ is differentiable at $\boldsymbol{\theta}_0$ for P- *almost every x with derivative $\dot{m}_{\boldsymbol{\theta}_0}(x)$ and such that, for every $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ in a neighborhood of $\boldsymbol{\theta}_0$ and a measurable function $\dot{m}$ with $\mathrm{P}\,\dot{m}^2 < \infty$*

$$|m_{\boldsymbol{\theta}_1}(x) - m_{\boldsymbol{\theta}_2}(x)| \leq \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|.$$

Furthermore, assume that the map $\boldsymbol{\theta} \mapsto \mathrm{P}\, m_{\boldsymbol{\theta}}$ admits a second-order Taylor expansion at a point of maximum $\boldsymbol{\theta}_0$ with nonsingular symmetric second derivative matrix $I_{\boldsymbol{\theta}_0}$. If $\mathbb{P}_n m_{\hat{\boldsymbol{\theta}}_n} \geq \sup_{\boldsymbol{\theta}} \mathbb{P}_n m_{\boldsymbol{\theta}} - o_{\mathrm{P}}(n^{-1})$ and $\hat{\boldsymbol{\theta}}_n \xrightarrow{\mathrm{P}} \boldsymbol{\theta}_0$, then

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) = -I_{\boldsymbol{\theta}_0}^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \dot{m}_{\boldsymbol{\theta}_0}(X_i) + o_{\mathrm{P}}(1).$$

In particular, the sequence $\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)$ is asymptotically normal with mean zero and covariance matrix $I_{\boldsymbol{\theta}_0}^{-1}\,\mathrm{P}\,\dot{m}_{\boldsymbol{\theta}_0}\dot{m}_{\boldsymbol{\theta}_0}^{\top} I_{\boldsymbol{\theta}_0}^{-1}$.

For the treatment of semiparametric models, for which the infinite-dimensional parameter can not be handled separately, it is useful to extend the results on Z-estimators to the case of infinite-dimensional parameters. A differentiability or Lipschitz condition on the maps, which are equal to zero at the estimate $\hat{\boldsymbol{\theta}}_n$, would preclude most applications. However, if we use the language of Donsker classes, the extension is straightforward.

If the parameter $\boldsymbol{\theta}$ ranges over a subset of infinite-dimensional normed space, then we use an infinite number of estimating equations, which we label by some set $\mathcal{H}$. Thus the estimator $\hat{\boldsymbol{\theta}}_n$ solves an equation $\mathbb{P}_n U(\boldsymbol{\theta})(\boldsymbol{h}) = \mathbb{P}_n U_{\boldsymbol{\theta},\boldsymbol{h}} = 0$ for every $\boldsymbol{h} \in H$. In our case $\mathbb{P}_n U(\boldsymbol{\theta})$ represents the score function.

Theorem A.10. *(Van der Vaart & Wellner (1996), Theorem 3.3.1 and Van der Vaart (1998) Theorem 19.26)*
For each $\boldsymbol{\theta}$ in a subset Θ of a normed space and every h in an arbitrary set $\mathcal{H}$, let $x \mapsto U_{\boldsymbol{\theta},\boldsymbol{h}}(x)$ be a measurable function such that:

1. *The class of functions* $\{U_{\boldsymbol{\theta},\boldsymbol{h}} : \|\boldsymbol{\theta}-\boldsymbol{\theta}_0\| < \epsilon, \boldsymbol{h} \in \mathcal{H}\}$ *is* P_0*-Donsker for some* $\epsilon > 0$, *with finite envelope function.*

2. $\sup_{\boldsymbol{h}\in\mathcal{H}} \mathrm{P}_0 \left(U_{\boldsymbol{\theta},\boldsymbol{h}} - U_{\boldsymbol{\theta}_0,\boldsymbol{h}}\right)^2 \to 0$ *as* $\boldsymbol{\theta} \to \boldsymbol{\theta}_0$.

3. *The map* $\boldsymbol{\theta} \mapsto \mathrm{P}_0\, U(\boldsymbol{\theta})$ *is Fréchet-differentiable at a zero* $\boldsymbol{\theta}_0$, *with a derivative* $\sigma_{\boldsymbol{\theta}_0} : \operatorname{lin}\Theta \mapsto l^\infty(\mathcal{H})$ *that has a continuous inverse on its range.*

4. $\mathrm{P}_0\, U(\boldsymbol{\theta}_0) = 0$ *and* $\hat{\boldsymbol{\theta}}_n$ *satisfies* $\mathbb{P}_n U(\hat{\boldsymbol{\theta}}_n) = o_{\mathrm{P}}(n^{-1/2})$ *and converges to* $\boldsymbol{\theta}_0$ *in probability.*

Then $\sigma_{\boldsymbol{\theta}_0}\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) = -\mathbb{G}_n[U(\boldsymbol{\theta}_0)] + o_{\mathrm{P}}(1)$.

Bibliography

Andersen, P. & Gill, R. (1982). Cox's regression model for counting processes: A large sample study. *Ann. Stat.* **10**, 1100–1120.

Andersen, P., Ø.B., Gill, R. & Keiding, N. (1993). *Statistical Models Based on Counting Processes*. Springer-Verlag, New York.

Anderson, J. & Senthilselvan, A. (1982). A two-step regression model for hazard functions. *Appl. Statist.* **31**, 44–51.

Bagdonavičius, V. & Nikulin, M. (2002). *Accelerated Life Models*. Monographs on Statistics and Applied Probability 94, Boca Raton.

Beibel, M. (1994). Bayes problems in change-point models for the Wiener process. In *Change-point Problems* (eds. E. Carlstein, H. Müller & D. Siegmund), 1–6, Institute of Mathematical Statistics, Hayward.

Beibel, M. (1996). A note on Ritov's Bayes approach to the minimax property of the CUSUM-procedure. *Ann. Stat.* **24**, 1804–1812.

Bhattacharya, P. (1994). Some aspects of change-point analysis. In *Change-point Problems* (eds. E. Carlstein, H. Müller & D. Siegmund), 28–56, Institute of Mathematical Statistics, Hayward.

Bickel, P., Klaassen, A., Ritov, Y. & Wellner, J. (1998). *Efficient and Adaptive estimation for Semiparametric Models*. Springer, New York.

Billingsley, P. (1968). *Convergence of Probability Measures*. John Wiley, New York.

Brodsky, B. & Darkhovsky, B. (1993). *Nonparametric Methods in Change-Point Problems*. Kluwer Academic Publishers, Dordrecht.

Brown, M. & Zacks, S. (2006). A note on optimal stopping for possible change in the intensity of an ordinary Poisson process. *Stat. Probab. Lett.* **76**, 1417–1425.

Chapell, R. (1989). Fitting bent lines to data, with applications to allometry. *Journal of Theoretical Biology* .

Chen, S. & Zhou, L. (2007). Local partial likelihood estimation in proportional hazards regression. *Ann. Stats.* **35**, 888–916.

Clayton, D. (1978). A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika* **65**, 141–151.

Clayton, D. & Cuzick, J. (1985). Multivariate generalization of the proportional hazards model. *J. Roy. Statist. Soc. Ser. A* **148**, 82–117.

Cleveland, W. (1979). Robust locally weighted regression and smoothing scatterplots. *J. Am. Statist. Assoc.* **74**, 829–836.

Cox, D. (1972). Regression models and life tables. *J. Roy. Statist. Soc. Ser. B* **34**, 187–200.

Csörgő, M. & Horváth, L. (1997). *Limit Theorems in Change-Point Analysis*. Wiley, Chichester.

Dabrowska, D. M. (1997). Smoothed Cox regression. *Ann. Stat.* **25**, 1510–1540.

Dupuy, J.-F. (2006). Estimation in a change-point hazard regression model. *Stat. Probab. Lett.* **76**, 182–190.

Feder, P. (1975a). The log likelihood ratio in segmented regression. *Ann. Stat.* **3**, 84–97.

Feder, P. (1975b). On asymptotic distribution theory in segmented regression problems. *Ann. Stat.* **3**, 49–83.

Fleming, T. & Harrington, D. (1991). *Counting processes and survival analysis*. Wiley & Sons, Hoboken.

Gandy, A. & Jensen, U. (2005). On goodness of fit tests for Aalen's additive risk model. *Scand. J. Stat.* **32**, 425–445.

Gandy, A. & Jensen, U. (2006). Model checks for Cox-type regression models based on optimally weighted martingale residuals. Tech. rep., www.uni-hohenheim.de/inst110/jensen.

Gandy, A., Jensen, U. & Lütkebohmert, C. (2005). A Cox model with a change-point applied to an actuarial problem. *Braz. J. Probab. Stat.* **19**, 93–109.

Herberts, T. & Jensen, U. (2004). Optimal detection of a change point in a Poisson process for different observation schemes. *Scand. J. Stat.* **31**, 347–366.

Hinkley, D. (1971). Inference in two-phase regression. *J. Am. Stat. Assoc.* **66**, 736–743.

Hougaard, P. (2000). *Analysis of mulitvariate survival data*. Springer, New York.

Ibragimov, I. & Has'minskii, R. (1981). *Statistical estimation. Asymptotic theory*. Springer, New York.

Jensen, U. & Lütkebohmert, C. (2007a). Change-point models. In *Encyclopedia of Statistics in Quality and Reliability*, Wiley.

Jensen, U. & Lütkebohmert, C. (2007b). A Cox-type regression model with change-points in covariates. Tech. rep., www.uni-hohenheim.de/spp.

Kosorok, M., Lee, B. & Fine, J. (2004). Robust inference for univariate proportional hazards frailty regression models. *Ann. Stat.* **32**, 1448–1491.

Kosorok, M. & Song, R. (2007). Inference under right censoring for transformation models with a change-point based on a covariate threshold. *Ann. Stat.* **35**, 957–989.

Koul, H. & Qian, L. (2002). Asymptotics of maximum likelihood estimator in a two-phase linear regression model. *J. Stat. Plann. Inference* **108**, 99–119.

Koul, H., Quian, L. & Surgailis, D. (2003). Asymptotics of M-estimators in two-phase linear regression models. *Stochastic Processes Appl.* **103**, 123–154.

Krall, J., Uthoff, V. & Harley, J. (1975). A step-up procedure for selecting variables associated with survival. *Biometrics* **31**, 49–57.

LeCam, L. (1986). *Asymptotic Methods in Statistical Decision Theory*. Springer, New York.

Liang, K.-Y., Self, S. & Liu, X. (1990). The Cox proportional hazard model with change-point: An epidemiologie application. *Biometrics* **46**, 783–793.

Lin, D. & Ying, Z. (1995). Semiparametric analysis of general additive-multiplicative hazard models for counting porcesses. *Ann. Stat.* **23**, 1712–1734.

Loader, C. (1996). Change point estimation using nonparametric regression. *Ann. Stat.* **24**, 1667–1678.

Lorden, G. (1971). Procedures for reacting to a change in distribution. *Ann. Math. Stat.* **42**, 1897–1908.

Luo, X. & Boyett, J. (1997). Estimations of a threshold parameter in Cox regression. *Commun. Stat., Theory Methods* **26**, 2329–2346.

Lütkebohmert, C., Jensen, U., Beier, M. & Schinköthe, W. (2007). Wie lange lebt mein kleinmotor? *Mechatronik* **9**.

Martinussen, T. & Scheike, T. H. (2002). A flexible additive multiplicative hazard model. *Biometrika* **89**, 283–298.

Moustakides, G. (1986). Optimal stopping times for detecting changes in distribution. *Ann. Stat.* **14**, 1379–1387.

Müller, H.-G. (1992). Change-points in nonparametric regression analysis. *Ann. Stat.* **20**, 737–761.

Müller, H.-G. & Song, K.-S. (1997). Two-stage change-point estimators in smooth regression. *Stat. Probab. Lett.* **34**, 323–335.

Müller, H.-G. & Stadtmüller, U. (1999). Discontinuous versus smooth regression. *Ann. Stat.* **27**, 299–337.

Nielsen, G., Gill, R. D., Andersen, P. & Sørensen, T. (1992). A counting process approach to maximum likelihood estimation in frailty models. *Scand. J. Stat.* **19**, 25–43.

Page, E. (1954). Continuous inspection schemes. *Biometrika* **41**, 100–115.

Parner, E. (1998). Asymptotic theory for correlated gamma-frailty model. *Ann. Stat.* **26**, 183–214.

Peskir, G. & Shiryaev, A. (2002). Solving the Poisson disorder problem. In *Advances in Finance and Stochastics* (eds. K. Sandmann & P. Schönbacher), 295–312, Springer, Berlin.

Pons, O. (2003). Estimation in a Cox regression model with a change-point according to a threshold in a covariate. *Ann. Stat.* **31**, 442–463.

Prentice, R. & Self, S. (1983). Asymptotic distribution theory for Cox-type regression models with general relative risk form. *Ann. Stat.* **11**, 804–813.

Rebolledo, R. (1980). Central limit theorems for local martingales. *Z. Wahrsch. Verw. Gebiete* **51**, 269–286.

Sasieni, P. (1996). Proportional excess hazards. *Biometrika* **83**, 127–141.

Scheike, T. & Zhang, M. (2002). An additive-mulitplicative Cox-Aalen survival model. *Scand. J. Statist.* **28**, 75–88.

Shiryaev, A. (1963). On optimum methods in quickest detection problems. *Theor. Probab. Appl.* **8**, 22–46.

Shiryaev, A. (1978). *Optimal Stopping Rules*. Springer, New York.

Slud, E. & Vonta, F. (2004). Consistency of the NPMLE esimator in the right-censored transformation model. *Scand. J. Stat.* **31**, 21–41.

Therneau, T., Grambsch, P. & Fleming, T. (1990). Martingale-based residuals for survival models. *Biometrika* **77**, 147–160.

Van der Vaart, A. (1998). *Asymptotic statistics*. Cambridge University Press, New York.

Van der Vaart, A. & Wellner, J. (1996). *Weak convergence and empirical processes*. Springer, New York.

Zacks, S. (1982). Classical and bayesian approaches to the change-point problem: Fixed sample and sequential procedures. *Stat. Anal. Donnees* **7**, 48–81.

Zusammenfassung

In dieser Arbeit werden verschiedene Regressionsmodelle aus der Lebensdaueranalyse betrachtet. Speziell das Cox Modell, welches auf proportionalen Ausfallraten basiert, wird erweitert.
Eine klassische Anwendung der Lebensdaueranalyse gibt es im Bereich der Medizin. Bei Patienten in einer Studie wird beobachtet, zu welchem Zeitpunkt ein bestimmtes Ereignis eintritt. Dieses Ereignis kann der Tod des Patienten oder ein Rückfall nach einer Operation sein. Das Ziel von Regressionsmodellen ist es, einen Zusammenhang zwischen dem Risiko für das Auftreten eines Ereignisses und bestimmten Attributen einzelner Individuen zu modellieren. Ein besonderes Interesse besteht darin herauszufinden, welchen Einfluss Medikamente oder Merkmale eines Patienten wie zum Beispiel das Alter auf das Überleben des Patienten haben. Ein Problem ist, dass oftmals eine Auswertung der Studie erfolgen muss, bevor bei allen Patienten der Tod registriert wurde. Das heißt, es ist nur bekannt, dass die Patienten bis zu einem bestimmten Zeitpunkt überlebt haben. Diese so genannten Zensierungseffekte müssen in einer statistischen Auswertung berücksichtigt werden. Dieses Phänomen tritt jedoch nicht nur im medizinischen Umfeld auf. Zum Beispiel werden Dauerlaufversuche von verschiedenen Motoren aus Kostengründen ebenso nach einer bestimmten Zeit abgebrochen, so dass für die Motoren nicht nur Ausfallzeiten sondern auch Zensierungen beobachtet werden. Trotzdem möchte man Lebensdauerverteilungen der Motoren unter Berücksichtigung verschiedener Merkmalkombinationen angeben. Ein weiteres Beispiel für zensierte Daten ergibt sich aus einer Studie über Versicherungsverträge. Die Stornierung von Verträgen, die verschiedene Attribute aufweisen, kann nur in einem bestimmten Zeitfenster beobachtet werden. Mit Hilfe von Regressionsmodellen der Lebensdaueranalyse kann man Attribute bestimmen, die einen Effekt auf die Stornierung eines Vertrages haben.
Während in der klassischen Lebensdaueranalyse davon ausgegangen wird, dass nur ein Ereignis auftritt, kann mit der Theorie der Zählprozesse auch mehr als ein Ereignis betrachtet werden. Ein Beispiel dafür ist die Berechnung des Risikos für das Auftreten von Schadensfällen mehrerer Versicherungsverträge. Theoretisch wird für ein Individuum i ein stochastischer Prozess $N_i(t)$ beobachtet, der die Anzahl der Ereignisse bis zu einem Zeitpunkt t zählt. Die Regressionsmodelle werden typischerweise über die Intensität der

Zählprozesse definiert. Das am weitesten verbreitete Modell ist das Cox Modell. Es wird beschrieben durch die Intensität

$$\lambda_i(t) = \lambda_0(t) R_i(t) \exp\{\boldsymbol{\beta}^\top \boldsymbol{Z}_i(t)\}, \tag{A.1}$$

wobei $\lambda_0(t)$ die Baseline-Intensität bezeichnet, die für alle Individuen gleich ist. Des Weiteren ist $R_i(t)$ ein Faktor, der angibt, ob ein Individuum unter Risiko steht oder nicht und deshalb nur die Werte 1 oder 0 annimmt. Die Merkmale eines Individuums werden in dem so genannten Kovariablenvektor $\boldsymbol{Z}_i$ zusammengefasst und $\boldsymbol{\beta}$ stellt die zu schätzenden Regressionsparameter dar. In diesem Modell wird angenommen, dass alle Kovariablen linear eingehen. Wir haben Datensätze untersucht, in denen diese Annahme verletzt ist. Daher haben wir ein neues flexibleres Modell entwickelt.
In dieser Arbeit betrachten wir ein Cox Modell mit einem so genannten Change-Point. Ein Change-Point beschreibt den Wert, an dem sich der Einfluss einer Kovariablen ändert. Wir nehmen dazu an, dass unsere zugrunde liegende Regressionsfunktion in dem Change-Point stetig, aber nicht differenzierbar ist. Im Gegensatz zum klassischen Cox Modell erhalten wir auf diese Weise einen weiteren zu schätzenden Parameter. Das Modell hat die folgende Form

$$\lambda_i(t) = \lambda_0(t) R_i(t) \exp\{\boldsymbol{\beta}_1^\top \boldsymbol{Z}_{1i}(t) + \beta_2 Z_{2i} + \beta_3 (Z_{2i} - \xi)^+\}, \tag{A.2}$$

wobei a^+ das Maximum von a und 0 ist, $\lambda_0(t)$ wieder die Baseline-Intensität beschreibt und $R_i(t)$ wie im klassischen Cox Modell definiert ist. Der Kovariablenvektor teilt sich in $\boldsymbol{Z}_{1i}$ und Z_{2i} auf, wobei sich der Einfluss von Z_{2i} im Change-Point $\xi \in \mathbb{R}$ verändert. Die Regressionsparameter sind in der Form eines Vektors $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^\top, \beta_2, \beta_3)^\top \in \mathbb{R}^{p+2}$ gegeben. Zu schätzen sind der endlichdimensionale Parametervektor $\boldsymbol{\beta}$, der Change-Point ξ und die unbekannte Baseline-Intentsität $\lambda_0(t)$. In dem klassischen Cox Modell benutzen wir für die Schätzung der Parameter einen Partial Likelihood der auf der Intensität (A.1) basiert. Anstelle der Baseline-Intensität schätzen wir die kumulierte Baseline-Intensität $\Lambda_0(t) = \int_0^t \lambda_0(s)\,\mathrm{d}s$ mit Hilfe des Breslow Schätzers.
Im Cox Modell mit einem Change-Point schätzen wir die Parameter ebenfalls mit einem Partial Likelihood, der jetzt jedoch auf (A.2) basiert, das heißt die Likelihood-Funktion hängt von $\boldsymbol{\beta}$ und ξ ab. In unserer Arbeit zeigen wir, dass die Schätzer die üblichen asymptotischen Eigenschaften aufweisen, das heißt sie sind konsistent und asymptotisch normalverteilt. Für den Nachweis dieser Eigenschaften benötigen wir Techniken aus der Theorie der Empirischen Prozesse und verwenden insbesonder Methoden, die für allgemeine M-Schätzer entwickelt wurden.
Die kumulierte Baseline-Intensität wird mit Hilfe des Breslow Schätzers, der jetzt von $\boldsymbol{\beta}$ und ξ abhängt, geschätzt. Wir zeigen, dass $\sqrt{n}(\hat{\Lambda}_n(t) - \Lambda_0(t))$ gegen einen Gaußschen

Prozess konvergiert.
In einem zweiten Schritt haben wir das Cox Modell mit einem Change-Point erweitert, indem wir anstatt der Exponentialfunktion eine allgemeine Risikofunktion $r : \mathbb{R} \mapsto [0, \infty)$ un mehrere Change-Points zugelassen haben. Unter geeigneten Bedingungen an die Funktion r lassen sich dieselben Resultate erzielen wie im Fall der Exponentialfunktion. Allerdings zeigt sich, dass die Beweise deutlich aufwändiger sind, wenn die Exponentialfunktion ersetzt wird.
Ein weiteres Kernthema der Arbeit ist die Betrachtung eines linearen Transformationsmodells mit Change-Points. Dieses Modell beinhaltet nicht nur unter bestimmten Voraussetzungen das oben genannte Cox Modell, sondern auch allgemeinere Modelle wie Frailty-Modelle. In Frailty-Modellen werden Gruppen von Ausfallzeiten betrachtet, die aufgrund eines nicht beobachtbaren Risikofaktors korreliert sind. Gruppen, die den gleichen Risikofaktor teilen, können Familien oder auch Motoren aus demselben Werk sein. In diesen Modellen wirkt eine unbeobachtbare Zufallsvariable multiplikativ auf die Intensität ein. Für die unbekannte Zufallsvariable wird normalerweise eine Verteilungsannahme getroffen. Typischerweise betrachtet man eine Klasse von Gamma-Verteilungen. Solche Frailty-Modelle und andere Modelle lassen sich in einem linearen Transformationsmodell zusammenfassen. Ein lineares Modell bezüglich einer Überlebenszeit T hat die Form

$$\log A(T) = -\boldsymbol{\beta}^\top \boldsymbol{Z} + \epsilon,$$

wobei A eine unbekannte monoton wachsende Funktion ist, $\boldsymbol{\beta}$ einen Regressionsparameter beschreibt und ϵ einer Fehlerverteilung folgt, die nicht von den Kovariablen $\boldsymbol{Z}$ abhängt. Das Modell kann deshalb äquivalent beschrieben werden durch

$$S_{\boldsymbol{Z}}(t) = S_\epsilon \left(\log A(t) + \boldsymbol{\beta}^\top \boldsymbol{Z} \right).$$

Hier stellt $S_{\boldsymbol{Z}}$ die Überlebensfunktion von T bei einem gegebenen Kovariablenvektor $\boldsymbol{Z}$ dar und S_ϵ beschreibt die Funktion $1 - F_\epsilon$, wobei F_ϵ die Verteilungsfunktion von ϵ bezeichnet. Wählt man nun $S_\epsilon(u) = \Lambda(e^u)$, so erhält man die folgende Darstellung

$$S_{\boldsymbol{Z}}(t) = \Lambda \left(\int_0^t \exp\{\boldsymbol{\beta}^\top \boldsymbol{Z}\} \, \mathrm{d}A(u) \right).$$

Die Funktion $\Lambda(t)$ ist eine bekannte, dreimal differenzierbare, fallende Funktion mit $\Lambda(0) = 1$ und $A(t)$ beschreibt eine unbekannte kumulierte Baseline-Intensität. Dieses Transformationsmodell, das in der letzten Zeit von verschiedenen Autoren untersucht wurde, erweitern wir, indem wir zensierte Daten zulassen und in die zugrunde liegende

Regressionsfunktion Change-Points aufnehmen, so dass wir das folgende Modell erhalten

$$S_{\boldsymbol{Z}}(t) = \Lambda\left(\int_0^t \exp\{\boldsymbol{\beta}_1^\top \boldsymbol{Z}_1(t) + \boldsymbol{\beta}_2^\top \boldsymbol{Z}_2 + \boldsymbol{\beta}_3^\top (\boldsymbol{Z}_2 - \boldsymbol{\xi})^+\}\, \mathrm{d}A(u)\right).$$

Die Schwierigkeit bei der Anwendung dieses Modells liegt darin, dass die Schätzung der Baseline-Intensität nicht mehr von der Schätzung der Regressionsparameter und Change-Point Parameter mit Hilfe des Partial Likelihoods getrennt werden kann. Deshalb wird für die Schätzung ein nichtparametrischer Likelihood verwendet, der die Komplexität des Problems deutlich erhöht. Mit der Hilfe von modernen Beweismethoden aus dem Bereich der empirischen Prozesse, die in Van der Vaart & Wellner (1996) zusammengefasst sind, können wir die üblichen asymptotischen Eigenschaften für unsere Schätzer wie $\sqrt{n}$-Konsistenz und asymptotische Normalität nachweisen. Insbesondere benötigen wir lineare Operatoren und Fréchet Differenzierbarkeit, um die Eigenschaften des unendlichdimensionalen Schätzers für die Baseline-Intensität zu erhalten.
Zum Abschluss unserer Arbeit haben wir das Cox-Modell mit Change-Points auf drei verschiedene Datensätze angewendet. Mit verschiedenen Selektionsmethoden haben wir die für die Lebensdauer signifikanten Kovariablen ermittelt. Außerdem konnten wir mit einem speziellen Anpassungstest nachweisen, dass in einigen Kovariablen Change-Points vorhanden sind. Dieser Test ermöglicht uns auch, deren Anzahl zu bestimmen und zu entscheiden, ob das Cox Modell mit Change-Points eine bessere Anpassung gegenüber dem klassischen Modell liefert. Ein weiterer Vorteil unseres Modells besteht darin, dass die funktionale Form einer Kovariablen durch die sukzessive Anpassung von Change-Points beschrieben werden kann. Bei der Auswertung des bekannten PBZ-Datensatzes aus Fleming & Harrington (1991), in dem Daten über Patienten mit primärer biliärer Zirrhose zusammengefasst sind, konnten wir mit dem Change-Point Modell eine bessere Anpassung erzielen als mit dem von Fleming & Harrington (1991) vorgeschlagenen Modell.
In dieser Arbeit haben wir neue Überlebensdauermodelle mit Change-Points entwickelt und für die zu schätzenden endlich- und unendlich-dimensionalen Parameter die üblichen asymptotischen Eigenschaften nachgewiesen. Außerdem haben wir reale Datensätze untersucht und gezeigt, dass sich durch die Verwendung der Change-Point Modelle eine Möglichkeit ergibt, die funktionale Form einer Kovariablen stückweise linear zu beschreiben.

Acknowledgements

First I want to thank my advisor Prof. Dr. Uwe Jensen for his support, encouragement and his valuable comments.
Furthermore, I would like to thank Dr. Axel Gandy for his very useful suggestions, comments and disscussions. Many thanks to my husband Jörg Marhenke for his continued support and very helpful disscussions.
Moreover, I would like to thank Prof. Dr. Ulrich Stadtmüller for serving on my thesis committee.
I also want to thank my colleagues from the Institute of Applied Mathematics and Statistics at the University of Hohenheim.

Curriculum Vitae

Persönliche Daten

Name:	Constanze Lütkebohmert-Marhenke, geb. Lütkebohmert
Geburtsdatum,-ort:	15. März 1978, Stadthagen, Deutschland
Staatsangehörigkeit:	deutsch
Adresse:	Fleinsbachstr.21, 73765 Neuhausen a.d.F.

Ausbildung

04/2003-	Universität Hohenheim Promotionsstudentin
10/1997-03/2003	Universität Ulm Diplom Wirtschaftsmathematik
08/2001-05/2002	University of Wisconsin at Milwaukee Master of Science in Mathematics
02/2000-08/2000	Universidad de La Laguna, Teneriffa Studentin
1990-1997	Gymnasium Bad Nenndorf allgemeine Hochschulreife

Berufserfahrung

04/2003-	Wissenschaftliche Mitarbeiterin an der Universität Hohenheim und in dem Teilprojekt "Stochastische Modelle komplexer Systeme und die Analyse ihrer Zuverlässigkeit in frühen Entwicklungsphasen" der DFG-Forschergruppe 460
08/2001-05/2002	Teaching Assistant an der University of Wisconsin at Milwaukee
11/1999-03/2003	Hilfskraft an der Universität Ulm

Erklärung

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbständig angefertigt habe, nur die angegebenen Quellen und Hilfsmittel benutzt und wörtlich oder inhaltlich übernommene Stellen als solche gekennzeichnet habe.

Hohenheim, den 27. September 2007 Constanze Lütkebohmert-Marhenke