Semiparametric sieve maximum likelihood estimation for interval censored data with/without cure fraction

Szabo, Zsolt

2018

Szabo, Z. (2018). Semiparametric sieve maximum likelihood estimation for interval censored data with/without cure fraction. Doctoral thesis, Nanyang Technological University, Singapore.

<http://hdl.handle.net/10356/75048>

<https://doi.org/10.32657/10356/75048>

**NANYANG
TECHNOLOGICAL
UNIVERSITY**

SINGAPORE

**SEMIPARAMETRIC SIEVE
MAXIMUM LIKELIHOOD
ESTIMATION FOR INTERVAL
CENSORED DATA
WITH/WITHOUT CURE
FRACTION**

ZSOLT SZABO

**DIVISION OF MATHEMATICAL SCIENCES
SCHOOL OF PHYSICAL AND MATHEMATICAL
SCIENCES**

2018

**SEMIPARAMETRIC SIEVE
MAXIMUM LIKELIHOOD
ESTIMATION FOR INTERVAL
CENSORED DATA
WITH/WITHOUT CURE
FRACTION**

ZSOLT SZABO

Division of Mathematical Sciences
School of Physical and Mathematical Sciences

A thesis submitted to Nanyang Technological University
in partial fulfillment of the requirement for the degree of
Doctor of Philosophy in Mathematical Sciences

2018

Abstract

This thesis focuses on semiparametric sieve maximum likelihood estimation of interval censored survival data. It consists of two parts.

In the first part, we discuss the accelerated hazards (AH) model, which provides an alternative to the popular proportional hazards (PH) model when the proportionality does not hold. We explore the difficulties that arise when one fits the AH model to interval censored data. We develop a semiparametric sieve maximum likelihood estimator and provide an algorithm for its implementation. We also establish consistency results and set up the rate of convergence.

In the second part, we propose a new double semiparametric mixture cure model for analyzing interval censored data with possible cure fraction. The proposed model incorporates semiparametric latency and incidence parts. Unlike existing works in the literature, where the incidence follows a parametric model, the proposed model allows the incidence to be semiparametric. We develop a spline-based sieve maximum likelihood estimation approach to analyze such data. Using modern empirical process techniques we establish large sample properties of the estimator, including the consistency, rate of convergence and the asymptotic normality of the finite dimensional parameters.

For both parts of the thesis we provide extensive simulation studies to show the finite sample size performance of the proposed estimation algorithm. For illustration purpose, the proposed method are applied to real data.

Acknowledgements

First and foremost, I would like to express my deepest gratitude to my supervisor Associate Professor Xiang Liming for her guidance through my PhD studies and her feedback on my research. This work would not have been possible with her encouragement and her hiring me as a research assistant.

I would like to thank A*STAR for providing me with the Singapore International Graduate Award, a generous scholarship that supported my studies.

I would like to express my gratitude to Associate Professor Pan Guangmin and Associate Professor Nicholas Privault at the Division of Mathematical Sciences who helped extend my scope of knowledge in probability and statistics through their valuable graduate courses and seminar. I am really grateful for all the support from fellow students and staff at SPMS. Especially, the current and past members of our group, Wang Chunjie, Mark Hannay, Peng Mengjiao, Hu Tao, Yu Tonghui, Liu Xiaoyu and Huang Rui. I am also indebted to Yang Qing and Bao Zhigang from the statistics group who helped me become more familiar with the domain statistical research in general.

I am also thankful to Biswaroop Maiti, and Vu Van Khu for being close friends, helpful classmates and nice housemates.

I am deeply grateful for my girlfriend, Vicky Tran, who supported me through the toughest period of my work. I really appreciate my brother, Istvan's help in checking my work. Last but not least, I am grateful to my parents who provided me great support during all the years of my studies.

Contents

Abstract	i
Acknowledgements	iii
1 Introduction	1
1.1 Interval censored survival data	1
1.2 Semiparametric models for interval censored data	3
1.2.1 The Accelerated Hazards Model	4
1.2.2 Mixture Cure Models	6
1.3 Preliminaries	8
1.3.1 Maximum Likelihood Estimator	8
1.3.2 Modeling the cumulative baseline hazard function with splines	10
1.3.3 Sieve semiparametric maximum likelihood estimation	14
1.4 Outline	15
2 Semiparametric Sieve Maximum Likelihood Estimation for Accelerated Hazards Model with Interval-Censored Data	17
2.1 Accelerated hazards model with interval censored data . . .	17

2.2	Estimation Procedure	19
2.2.1	Sieve Semiparametric Maximum Likelihood Estimator	19
2.2.2	Implementation	21
2.2.3	Choosing the number of base splines	24
2.3	Large sample properties	25
2.4	Simulations	28
2.5	Application	32
2.6	Technical proofs	38
2.6.1	Proof of Theorem 2.1	38
2.6.2	Proof of Theorem 2.2	41
2.6.3	Proof of Theorem 2.3	44
3	Double Semiparametric Mixture Cure Model with Interval Censored Data	47
3.1	Double semiparametric model with interval censored data . .	47
3.2	Estimation procedure	50
3.2.1	Semiparametric Sieve Maximum Likelihood Estimator	50
3.2.2	Estimation algorithm	52
3.2.3	Implementation	54
3.2.4	Inference	55
3.3	Asymptotic Theory	57
3.4	Simulations	59
3.5	Application to the hypobaric decompression sickness data . .	65
3.6	Technical proofs	66
3.6.1	Proof of identifiability	66
3.6.2	Proof of large sample properties	69
4	Discussion and Future Research	83

CONTENTS

vii

Bibliography

87

List of Figures

2.1	Estimated cumulative baseline hazard functions in scenario 1 with censoring rates 0% to 5% (left panel), 12% to 18% (middle panel) and 20% to 25% (right panel). In each plot, the solid line gives the true cumulative baseline hazard, the dashed line corresponds to the mean of estimated baseline cumulative hazards, and the dotted lines are the upper and lower 2.5% quantiles of the estimates.	31
2.2	Estimated cumulative baseline hazard functions in scenario 2 with censoring rates 0% to 5% (left panel), 12% to 18% (middle panel) and 20% to 25% (right panel). In each plot, the solid line gives the true cumulative baseline hazard, the dashed line corresponds to the mean of estimated baseline cumulative hazards, and the dotted lines are the upper and lower 2.5% quantiles of the estimates.	34
2.3	Estimated cumulative baseline hazard curves under the AH and PH models in comparison with the empirical nonparametric cumulative baseline hazard function.	38

2.4	Estimated cumulative hazard functions based on ICM non-parametric estimate, and estimates given by the PH and the AH models	39
2.5	Estimated cumulative hazard functions based on nonparametric estimate (NP), and estimates given AH models for the control group on the left and for the intervention group on the right.	40
3.1	Estimated cumulative baseline hazard function. The solid line is the true, the dashed line is the mean, and the dotted lines are the 2.5% and 97.5% quantiles of the estimated curves. The subfigures from left to right shows the results for 15%, 35% and 50% of cure rate respectively.	62
3.2	Estimated functional part of the cure probability, scenario 1. The solid line is the true $f(w) = \sin(w * 2) + 1/2$, the dashed line is the mean, the dotted lines are the 2.5% and 97.5% quantiles of the estimated curves. The subfigures from left to right shows the results for 15%, 35% and 50% of cure rate respectively.	63
3.3	Estimated functional part of the cure probability. The solid line is the true $f(w) = 0.91w + 1/2$, the dashed line is the mean, the dotted lines are the 2.5% and 97.5% quantiles of the estimated curves. The subfigures from left to right shows the results for 15%, 35% and 50% of cure rate respectively.	64
3.4	Estimate of f , the effect of Age on the incidence for the HDSD data set.	67

List of Tables

2.1	Simulation results of the estimated regression coefficient in scenario 1 with the Weibull cumulative baseline hazard function	30
2.2	Simulation results of the estimated regression coefficient in scenario 2 with concave cumulative baseline hazard function	33
2.3	Analysis of the diabetes conversion data	36
3.1	Numerical results for estimators in the first scenario.	80
3.2	Numerical results for estimators in the second scenario.	81
3.3	The estimated coefficient for the HDSD data set	82

Chapter 1

Introduction

1.1 Interval censored survival data

Survival analysis is the study of a non-negative real-valued random variable T , which is considered to be the time to a specific event, usually a failure event. Such failure events include the death of a patient due to a disease or the time to onset of a disease of a high-risk individual. Because of the vast applications in clinical studies, various models based on such random variables have been created in the last decades. One can describe T with the usual cumulative distribution function (CDF) $F(t) = P(T \leq t)$, however, survival analysis studies traditionally use the survival function $S(t) = 1 - F(t)$ instead. In this thesis, we always assume that the survival function is absolutely continuous, hence the density $f(t) = -dS(t)/dt$ exists. Another function which is used to describe the distribution of T is the hazard function

$$\lambda(t) = \lim_{h \rightarrow 0} \frac{P(t < T \leq t + h | T > t)}{h}.$$

The hazard function is always non-negative and $\lambda(t)h$ can be regarded as a way to approximate the probability that a subject who survived until time t experiencing the failure event in the next instant (Klein and Moeschberger, 1997, p.27). If T is a non-negative continuous random variable, $\lambda(t)$ uniquely determines the distribution of T , and we can write $\lambda(t) = d\{-\log[S(t)]\}/dt$. In this thesis, we mostly refer to the cumulative version of the hazard function:

$$\Lambda(t) = \int_0^t \lambda(s) ds.$$

The relationship between the survival function and the cumulative hazard function can be written as

$$S(t) = \exp(-\Lambda(t)).$$

We remark that $\Lambda(t)$ is the integral of a nonnegative function, hence it is non-negative and monotone increasing. Standard statistical methodology can be used to analyze survival data if all the survival times are observed. However, a distinctive feature of survival data is that survival times may not be observed directly, they may be only known to fall in a specified interval. Such data is called censored data and it requires a different set of tools to analyze. The most common type of censoring is right censoring, which means that the unobserved T is larger than an observed censoring time C . Data that contain exact observations and right censored observations are referred to as right censored data. Right censoring naturally occurs in clinical studies, because such studies have to be finished within a stipulated time. At the end of the study if a subject survives, the failure time of that subject is censored at that time. Statistical techniques exist to make inference of right censored data, including the Kaplan–Meier (Kaplan and Meier, 1958) estimator and the Nelson–Aalen estimator (Nelson,

1972 and Aalen, 1978) for nonparametric estimate and the partial likelihood estimator for the Proportional Hazards (PH) model (Cox, 1972).

Another type of censoring is interval censoring which means that the failure time T is only known that $L < T \leq R$, where L and R are two known time points. We remark that it is possible that $L = 0$ or $R = \infty$. We distinguish case 1 and case 2 interval censored data. Case 1 data are also called current status data, where either $L = 0$ or $R = \infty$. In practice, the observation time C is known along with indicator δ for whether the subject has experienced the failure by that time. If $\delta = 1$ then $L = 0$ and $R = C$, otherwise $L = C$ and $R = \infty$.

In case 2, there are two theoretical censoring times U and V . The pair (L, R) of observed interval endpoints that contain T is $(0, U)$ if $T < U$, (U, V) if $U < T \leq V$ and (V, ∞) if $V < T$. Throughout this thesis, we investigate models based on case 2 interval censored data, and we use the term “interval censoring” for case 2 interval censoring. For example, interval censored data arises from studies in which patients go to laboratory tests according to a schedule. Their status are taken at each tests. The left observation is the time of the last negative test result, the right observation is the time of the first positive result.

1.2 Semiparametric models for interval censored data

In practice, it has become common to include explanatory variables, or covariates to the survival data. In such cases, a covariate vector $\mathbf{Z}$ is added to the interval endpoints L and R as a component of the observed data.

That changes the main problem from estimating the survival function of T to a regression problem, that is the estimation of the conditional survival function $S(t|\mathbf{Z}) = P(T > t|\mathbf{Z} = \mathbf{z})$. Commonly used survival regression models are semiparametric, such as the Cox proportional hazards (PH) model (Cox, 1972), the proportional odds (PO) model and the accelerated hazard (AH) model (Chen and Wang, 2000). These semiparametric models incorporate a linear form of the covariates $\mathbf{Z}$ with the coefficient $\boldsymbol{\beta}$ of finite dimension and an unspecified baseline hazard function λ_0 . For example, the PH model assumes the hazard function

$$\lambda(t|\mathbf{Z}) = \lambda_0(t) \exp(\boldsymbol{\beta}^T \mathbf{Z}). \quad (1.1)$$

In the analysis of interval censored data, we more often refer to the cumulative baseline hazard function $\Lambda_0(t)$, which is defined as $\int_0^t \lambda_0(s) ds$. Since Λ_0 is the integral of a nonnegative function, it is also nonnegative and non-decreasing.

In the literature, regression analysis of interval censored data has been investigated based on various semiparametric models. An elaborate summary can be found in the book of Sun (2006). In particular, the proportional hazards (PH) model studied by Huang and Wellner (1995), Rosenberg (1995) and Pan (1999) using maximum likelihood estimation (MLE) and Zhang et al. (2010) using the sieve MLE method; the proportional odds (PO) and accelerated failure time (AFT) models investigated by Zhang and Davidian (2008) through a smooth parametric estimation method.

1.2.1 The Accelerated Hazards Model

Existing models in the literature impose strict assumptions on the effect of the covariates on the hazard curve. As Chen and Wang (2000) argues, in

the PH, the PO and the AFT models the effect of a covariate is immediate. For example, in some clinical trials, subjects are randomly divided into a treatment group and a control group, and the treatment may only have long term effects. Chen and Wang (2000) proposed a new survival model, the accelerated hazards (AH) model as an alternative to the PH and PO models to allow the covariate to have less immediate effect on the hazard function.

For the failure time of interest T , the AH model assumes that the hazard function of T given covariate vector $\mathbf{Z}$ is scaled by the factor $\exp(\boldsymbol{\beta}^T \mathbf{Z})$ and has the form:

$$\lambda(t|\mathbf{Z}) = \lambda_0(t \exp(\boldsymbol{\beta}^T \mathbf{Z})), \quad (1.2)$$

where λ_0 is a nonnegative baseline hazard function and $\boldsymbol{\beta}$ is a vector of regression parameters.

Regression analysis of right censored data based on the AH model was initially studied by Chen and Wang (2000) and Chen (2001), who developed the method of estimating equations to find the parameter vector $\boldsymbol{\beta}$ with a note that the estimating equations in general do not have a unique solution due to nonsmooth estimating functions. This problem has attracted broad attention in recent years. Zhang et al. (2011) developed a smoothing estimating equation approach for the AH model based on a kernel smoothing method, alleviating computational challenges in the earlier works. Li et al. (2012) further addressed to solve nonsmoothing estimating equations using an induced smoothing procedure so that a unique solution for estimation can be achieved. Chen et al. (2014) studied a Bayesian AH model with a spline prior for smooth survival densities based on standard parametric distribution families. Despite significant recent development in the estimation of the AH regression, the statistical theory of the aforementioned methods

generally cannot be applied to the situation under interval censoring, where two related random variables corresponding to interval endpoints have to be considered.

When estimating semiparametric survival models, a challenge is to fit the unknown cumulative baseline hazard function, which is even more complicated under interval censoring than in the right censored case. One possible method is to model the cumulative baseline hazard function nonparametrically with B-splines, proposed by Rosenberg (1995) for the PH model. The B-spline estimator for the nonparametric component in the model gives a natural way to generalize the maximum likelihood estimator to a sieve estimator, and consequently large sample properties can be established. In this thesis, we aim to approximate the cumulative baseline hazard function using B-splines and then develop a sieve maximum likelihood estimation procedure for the AH model with interval censored data. To implement it, we propose a two-step maximization algorithm in which the interval that the B-splines are defined on, can be updated in each iteration step. An appealing feature of the proposed algorithm is that it allows us to adjust the support interval of the B-splines at each iteration step so that the model fits better to observed interval endpoints.

1.2.2 Mixture Cure Models

The assumption that T can only take finite values is not always realistic. In some cases, the data or the clinical specialist suggests that a subgroup of the subjects never experience the failure event. For example, in the dataset extracted from NASA's Hypobaric Decompression Sickness Data-bank (HDSD) study by Conkin et al. (1992) the failure time of interest is the time to onset of grade IV venous gas emboli (VGE). However, not every

subject experience grade IV VGE, as noted by Thompson and Chhikara (2003). In order to account for a large proportion of subjects that may never experience the failure event, cure models have been proposed. The two most common cure models are the non-mixture cure models and the mixture cure models. The former model, originally proposed by Yakovlev and Tsodikov (1996), uses a single model to describe an improper survival function such that $\lim_{t \rightarrow \infty} S(t) > 0$. The mixture cure model (Farewell, 1982) assumes that the survival time is a mixture of a non-negative real-valued random variable and ∞ . The population survival function can be written as

$$S_{\text{pop}}(t|\mathbf{Z}, \mathbf{X}) = \pi(\mathbf{X})S(t|\mathbf{Z}) + 1 - \pi(\mathbf{X}), \quad (1.3)$$

where $\pi(\mathbf{X})$ is a mixing variable called the incidence, $S(t|\mathbf{X})$ is a proper survival function and $(\mathbf{X}, \mathbf{Z})$ are covariate vectors that may overlap. In existing works $\pi(\mathbf{X})$ is a transformation of the linear form of a covariate vector $\mathbf{X}$ through a link function. The function $S(t|\mathbf{Z})$ is the survival function corresponding to a semiparametric survival model and it is called the latency. Semiparametric mixture cure models have been studied with interval censored data, for example by Kim and Jhun (2008) and Ma (2010) based on the PH survival model.

The common feature of the MC model is that the incidence probability follows a parametric linear model through a link function. The most commonly used link functions are the logit link, and the complementary log-log link. However, there may not be enough evidence to assume such parametric models and more flexible models are necessary. To account for non-linear effects, Xu and Peng (2014) provided a simple way to estimate the incidence rate non-parametrically for a one-dimensional covariate, and it was extended by López-Cheda et al. (2017) for multidimensional covari-

ates. Wang et al. (2012) introduced a two-component nonparametric model for right censored data. However, if the estimator for the incidence is nonparametric, it is not easy to interpret the effect of the risk factors thus it is not suitable to identify the significant factors. Hu and Xiang (2016) proposed a semiparametric transformation cure model for interval censored survival data, where one covariate may not have a linear effect of the survival. Although they use the usual logistic model for the incidence, they mention the possibility of a semiparametric model in their discussion of future work.

In this thesis, we propose a double semiparametric mixture cure model that provides a way to incorporate a non-linear effect of certain covariates on the incidence probability. We perform the estimation of the nonparametric terms using B-splines, which results in a smooth estimator to describe the nonlinear effect of the given covariate.

1.3 Preliminaries

In this section we present some techniques used in the thesis. We introduce some likelihood-based approaches estimation techniques of semiparametric survival models.

1.3.1 Maximum Likelihood Estimator

Let us consider a sample of n independent subjects:

$$\mathcal{O} = \{L_i, R_i, \mathbf{Z}_i, i = 1, \dots, n\}.$$

Additional auxiliary indicator variables $\delta_{1i} = I(L_i = 0)$ and $\delta_{2i} = I(0 < L_i < R_i < \infty)$ and $\delta_{3i} = I(R_i = \infty)$ are introduced, where I is the in-

indicator function. The main objective of the analysis is to estimate the conditional survival function $S(t|\mathbf{Z}_i)$. In semiparametric survival models the conditional survival function can be described by combining the cumulative baseline hazard function $\Lambda_0(t)$ and the regression coefficient $\boldsymbol{\beta}$. We illustrate the semiparametric estimation procedures by the example of the PH model. In that case, the conditional survival function can be written as

$$S(t|\mathbf{Z}_i) = \exp \left[- \Lambda_0(t) \exp(\boldsymbol{\beta}^T \mathbf{Z}_i) \right].$$

In the literature of survival regression analysis, most inference for interval censored data are developed based on likelihood approaches. To obtain the likelihood function we need the following two assumptions.

(A1) Conditional on $\mathbf{Z}$, T is independent of the underlying censoring variables (U, V) .

(A2) The joint distribution of $(U, V, \mathbf{Z})$ does not depend on $(\boldsymbol{\beta}, \Lambda_0)$

Both assumptions are common in the literature of survival regression analysis. Assumption (A1) ensures that the likelihood function contains the joint distribution of (U, V) as a constant multiplicative term. Condition (A2) ensures that the joint likelihood of $(U, V, \mathbf{Z})$ is a constant term with respect to the regression parameters in the full likelihood. Thus it is usually omitted from the likelihood function.

Given assumptions (A1) and (A2), the likelihood function is proportional to

$$\begin{aligned} L(\boldsymbol{\beta}, \Lambda_0 | \mathcal{O}) \propto \prod_{i=1}^n \left\{ 1 - \exp \left[- \Lambda_0(R_i) \exp(\boldsymbol{\beta}^T \mathbf{Z}_i) \right] \right\}^{\delta_{1i}} \\ \left\{ \exp \left[- \Lambda_0(L_i) \exp(\boldsymbol{\beta}^T \mathbf{Z}_i) \right] - \exp \left[- \Lambda_0(R_i) \exp(\boldsymbol{\beta}^T \mathbf{Z}_i) \right] \right\}^{\delta_{2i}} \end{aligned} \quad (1.4)$$

$$\left\{ \exp \left[- \Lambda_0(L_i) \exp(\boldsymbol{\beta}^T \mathbf{Z}_i) \right] \right\}^{\delta_{3i}}.$$

In the rest of this thesis, we write equality instead of proportionality to simplify the notation. In general, the goal is to find the pair $(\hat{\boldsymbol{\beta}}, \hat{\Lambda}_0)$ that maximizes (1.4) over $\boldsymbol{\beta} \in B$, where B is usually a compact subset of a Euclidian space and Λ_0 in some function space. Since the cumulative hazard function is a nonnegative and nondecreasing function and $\exp(\boldsymbol{\beta}^T \mathbf{Z}_i)$ is positive, $\Lambda_0(t)$ is also a nonnegative and nondecreasing function. Finkelstein (1986) provided the first maximum likelihood estimator for the PH model with interval censored data. She used the fact that Λ_0 is only evaluated at the finite observation points L_i and R_i , hence it is enough to know the value at those points. As Huang and Wellner (1995) pointed out, to avoid ambiguity, the estimator $\hat{\Lambda}_0$ of Λ_0 can be regarded as a right continuous step function, where the steps are at the nonzero finite L_i and R_i . However, as Rosenberg (1995) argues, sometimes a continuous estimator of $\hat{\Lambda}_0$ is desired, and proposed to use B-splines for it.

1.3.2 Modeling the cumulative baseline hazard function with splines

In this section we present the B-splines, a method to create a finite dimensional space of nonnegative, nondecreasing piecewise polynomial functions defined on the finite closed interval $[a, b]$. First, let us define the space of real polynomials of order m as

$$\mathcal{P}_m = \left\{ p(x) = \sum_{i=1}^m c_i x^{i-1} \right\}.$$

We remark that the maximum degree of polynomials in $\mathcal{P}_m$ is $m - 1$. Let us define a partition $\Delta = \{\tau_i\}_{i=1}^Q$ where $a = \tau_0 < \tau_1 < \cdots < \tau_Q < \tau_{Q+1} = b$.

The partition defines the subintervals $I_i = [\tau_i, \tau_{i+1})$ for $i = 0, \dots, Q-1$ and $I_Q = [\tau_Q, \tau_{Q+1}]$. Let m be a positive integer and $\mathcal{M} = (m_1, m_2, \dots, m_Q)$ vector of positive integers such that $m_i \leq m$ for all $i = 1, \dots, Q$. Let D^j be the j th differential operator of a univariate function. We define the space

$$\begin{aligned} \mathcal{S}(\mathcal{P}_m, \mathcal{M}, \Delta) = \{s(t) : & \text{there exist polynomials } s_i(t) \in \mathcal{P}_m, i = 1, \dots, Q+1 \\ & \text{such that } s(t) = s_i(t) \text{ for } t \in I_i, i = 0, \dots, Q, \text{ and} \\ & D^j s_i(\tau_i) = D^j s_{i+1}(\tau_i) \text{ for } j = 0, \dots, m - m_i - 1, \\ & \text{and } i = 0, \dots, Q.\} \end{aligned}$$

The space $\mathcal{S}(\mathcal{P}_m, \mathcal{M}, \Delta)$ is called the polynomial spline space of order m with knots $\tau_1, \dots, \tau_Q$ of multiplicities $m_1, \dots, m_Q$. The vector $\mathcal{M}$ is called the multiplicity vector. It controls the smoothness of the spline space. When $m_i = m$, we interpret the definition so that $s_i(t)$ and $s_{i+1}(t)$ need not have any relation. Namely, when $m_1 = m_2 = \dots = m_Q = m$, the space is the piecewise polynomials of order m . Since our aim is to approximate the cumulative baseline hazard function, we always assume that $\mathcal{M} = (1, 1, \dots, 1)$, which means that functions in the spline space are continuously differentiable up to order $m-2$. For simplicity, we omit $\mathcal{M}$ from $\mathcal{S}(\mathcal{P}_m, \Delta)$. Another important feature of the polynomial spline spaces is that they are finite dimensional linear spaces. In particular, we have the following theorem.

Theorem 1.1 (Theorem 4.4 of Schumaker (2007)). The space $\mathcal{S}(\mathcal{P}_m, \Delta)$ is a linear space of dimension $q = m + Q$.

We need to give an efficient basis of $\mathcal{S}(\mathcal{P}_m, \Delta)$, so that we can use it in the estimation procedure. We introduce some technical terms in the following.

Definition 1.1. The matrix associated with the functions $\{u_i\}_{i=1}^r$ and values $\{t_i\}_{i=1}^r$ is

$$M \begin{pmatrix} t_1, t_2, \dots, t_r \\ u_1, u_2, \dots, u_r \end{pmatrix} = \begin{pmatrix} u_1(t_1) & u_2(t_1) & \dots & u_r(t_1) \\ u_1(t_2) & u_2(t_2) & \dots & u_r(t_2) \\ \vdots & \vdots & \ddots & \vdots \\ u_1(t_r) & u_2(t_r) & \dots & u_r(t_r) \end{pmatrix}$$

The determinant associated with $\{u_i\}_{i=1}^r$ and $\{t_i\}_{i=1}^r$ is

$$D \begin{pmatrix} t_1, t_2, \dots, t_r \\ u_1, u_2, \dots, u_r \end{pmatrix} = \det M \begin{pmatrix} t_1, t_2, \dots, t_r \\ u_1, u_2, \dots, u_r \end{pmatrix}.$$

Definition 1.2. Let f be a function of t and $t_1, t_2, \dots, t_{r+1}$ be in the domain of f . The r th order divided difference of f over the points $t_1, t_2, \dots, t_{r+1}$ is

$$[t_1, t_2, \dots, t_{r+1}]f = \frac{D \begin{pmatrix} t_1, t_2, \dots, t_{r+1} \\ 1, t, t^2, \dots, t^{r-1}, f \end{pmatrix}}{D \begin{pmatrix} t_1, t_2, \dots, t_{r+1} \\ 1, t, t^2, \dots, t^{r-1}, t^r \end{pmatrix}}.$$

Definition 1.3. An extended partition is defined as $\tilde{\Delta} = \{y_i\}_{i=1}^{2m+Q}$ so that $y_1 \leq y_2 \leq \dots \leq y_{2m+Q}$, and

$$y_1 \leq y_2 \leq \dots \leq y_m \leq a, \quad (1.5)$$

$$b \leq y_{m+Q+1} \leq \dots \leq y_{2m+Q}, \quad (1.6)$$

and $y_{m+i} = \tau_i$ for $i = 1, \dots, Q$.

We remark that the extended partition $\tilde{\Delta}$ of Δ is unique up to the first m and last m elements, which can be chosen subject to the properties given in (1.5) and (1.6).

For an arbitrary function f let us denote the positive part of f as f_+ , that is

$$f_+(t) = \begin{cases} 0 & \text{if } f(t) \leq 0 \\ f(t) & \text{if } f(t) > 0 \end{cases}$$

In the next theorem we give an efficient basis of the spline space.

Theorem 1.2 (Theorem 4.9 of Schumaker (2007)). Let $\tilde{\Delta}$ be an extended partition associated with $\mathcal{S}(\mathcal{P}_m, \Delta)$, and assume $b < y_{2m+Q}$. Let us define the functions $B_i(t)$ for $i = 1, \dots, q$ as

$$B_i(t) = (-1)^m (y_{i+m} - y_i) [y_i, \dots, y_{i+m}](t - y)_+^{m-1}, \quad a \leq t \leq b,$$

and $(t - y)_+$ is considered to be a function of y when the divided difference operator is applied to it. Then the functions $\{B_i(t)\}_{i=1}^q$ form a basis of $S(\mathcal{P}_m, \Delta)$ with the following properties.

$$B_i(t) = 0 \quad \text{for } t \notin [y_i, y_{i+m}],$$

and

$$B_i(t) > 0 \quad \text{for } t \in (y_i, y_{i+m}),$$

finally,

$$\sum_{i=1}^q B_i(t) = 0 \quad \text{for } t \in [a, b].$$

As a consequence of Theorem 1.2 one can describe a function in $S(\mathcal{P}_m, \Delta)$ as a linear combination of the spline basis function. Hence we can write

$$\mathcal{S}(\mathcal{P}_m, \Delta) = \left\{ f : f = \sum_{i=1}^q c_i B_i(t) \right\}.$$

Corollary 1.1. Let $c_1, c_2, \dots, c_q$ be a sequence of nonnegative real numbers.

The function $\sum_{i=1}^q c_i B_i(t)$, which is in $\mathcal{S}(\mathcal{P}_m, \Delta)$, is non-negative.

With Corollary 1.1, it is easy to constrain the splines to be non-negative. We also need restrict the spline to be monotone, which is asserted by the following proposition.

Proposition 1.1 (Example 4.75 of Schumaker (2007)). Let $c_1 < c_2 < \cdots < c_q$ be an increasing sequence. Then the spline $\sum_{j=1}^q c_j B_j(t)$ is monotone increasing.

This proposition allows us to define the space

$$\mathcal{M}(\mathcal{P}_m, \Delta) = \left\{ f : f(t) = \sum_{j=1}^q c_j B_j(t), \mathbf{c} \in C_q \right\}, \quad (1.7)$$

where C_q is the space of length- q non-negative monotone increasing real sequences. Since $\mathcal{M}$ contains the non-negative, monotone increasing functions on a specific interval, we can use it to approximate Λ_0 . With that approximation, the maximum likelihood problem is reduced to finding such $(\hat{\beta}, \hat{\Lambda}_0)$ that maximizes the likelihood (1.4) over $\beta \in B$ and $\Lambda_0 \in \mathcal{M}(\mathcal{P}_m, \Delta)$. In the next section, we discuss a framework to generalize this estimator.

1.3.3 Sieve semiparametric maximum likelihood estimation

The sieve semiparametric maximum likelihood estimation (Geman and Hwang, 1982) has become an attractive tool in nonparametric and semiparametric inference of interval censored data, due to its ability to simultaneously estimate the parametric and nonparametric components in the framework of complicated semi/nonparametric models. Shen and Wong (1994) established the theory for the convergence rate of the sieve estimation and showed that the resulting estimators for both components achieved optimal

convergence rates, respectively. Their theory has consequently been applied to the development of large sample properties for maximum likelihood estimators in the analysis of interval censored data under various models, including the proportional odds model by Huang and Rossini (1997), the proportional hazards model by Zhang et al. (2010), the PH-based copula joint model for dependent current status data by Ma et al (2015) and a class of semiparametric transformation models for bivariate data by Zhou et al. (2017) more recently.

A sieve maximum likelihood estimator is an estimator θ_n that maximizes a likelihood $l(\theta|\mathcal{O})$, over $\theta_n \in \Theta_n$, where $\mathcal{O}$ is the sample of size n and Θ_n is a finite dimensional subspace of an infinite dimensional space Θ . The space Θ_n is called the sieve space and its dimension increases with n . In case of a spline approximation of Λ_0 , the sieve can be defined as

$$\Theta_n = B \times \mathcal{M}_n(\mathcal{P}_m, \Delta_{Q_n}),$$

where Δ_{Q_n} is a partition of size Q_n and $Q_n \approx n^\nu$, $0 < \nu < 1$. With the estimator coming from such sieve space, one can get the asymptotic properties of θ_n . More precise definitions of the sieve space specific to each model are found in Chapters 2 and 3.

1.4 Outline

The rest of the thesis is organized as follows.

In Chapter 2 we investigate the accelerated hazards model, and it is organised as follows. In Section 2.1 we describe the AH model for interval censored data. In Section 2.2 we give estimation procedure using the B-splines to estimate the cumulative baseline hazard function. We establish

large sample properties in Section 2.3, and report simulation results to show the finite sample efficiency of the proposed method in Section 2.4, followed by applying the method to a real example in Section 2.5. We give the proof of theoretical results in Section 2.6.

In Chapter 3 we focus on the mixture cure model and it is organized as follows. In Section 3.1 we give a detailed description of the model. We provide the estimating procedure in Section 3. Large sample properties included in Section 4. Simulation studies and analysis of a real data set are in Section 5 and 6 respectively. We give proof of the theorems in Section 3.6

Chapter 4 concludes this thesis by summarizing the major results and addressing a few problems for future research.

Chapter 2

Semiparametric Sieve Maximum Likelihood Estimation for Accelerated Hazards Model with Interval-Censored Data

2.1 Accelerated hazards model with interval censored data

Under interval censoring, failure time T_i of subject i is known to be in the interval $(L_i, R_i]$ with observed endpoints L_i and R_i . The observed data is given as $O_i = (L_i, R_i, \mathbf{Z}_i)$, $i = 1, \dots, n$, where $\mathbf{Z}_i$ is a covariate vector of dimension p . $L_i = 0$ and $R_i = \infty$ correspond to left and right censored observations, respectively. We assume that $O_i \in \mathfrak{X} \subset \mathbb{R}^+ \times \mathbb{R}^+ \times \mathbb{R}^p$, where $\mathbb{R}^+$ is the set of nonnegative real numbers including ∞ and $\mathbb{R}^p$ is the Euclidean space of dimension p and $\mathfrak{X}$ is the space of possible observations. According

to the AH model (1.2), the cumulative hazard function of individual i is

$$\Lambda(t|\mathbf{Z}_i, \boldsymbol{\beta}) = \Lambda_0(t \exp(\boldsymbol{\beta}^T \mathbf{Z}_i)) \exp(-\boldsymbol{\beta}^T \mathbf{Z}_i), \quad (2.1)$$

where $\Lambda_0(t) = \int_0^t \lambda_0(s) ds$ is the cumulative baseline hazard function. Given the pair $(\boldsymbol{\beta}, \Lambda_0)$, one can obtain the cumulative hazard function of each individual from model (2.1) with different time scales depending on covariate $\mathbf{Z}_i$, which determines the stochastic order of observations.

We define δ_{1i} , δ_{2i} and δ_{3i} as left, interval and right censoring indicators as in Section 1.3. By assumptions (A1) and (A2) the loglikelihood function based on a sample of n observations is

$$\begin{aligned} l(\boldsymbol{\beta}, \Lambda_0|\mathcal{O}) = \sum_{i=1}^n \bigg[& \delta_{1i} \log \{1 - \exp(-\Lambda_0(R_i \exp(\boldsymbol{\beta}^T \mathbf{Z}_i)) \exp(-\boldsymbol{\beta}^T \mathbf{Z}_i))\} \\ & + \delta_{2i} \log \{ \exp(-\Lambda_0(L_i \exp(\boldsymbol{\beta}^T \mathbf{Z}_i)) \exp(-\boldsymbol{\beta}^T \mathbf{Z}_i)) \\ & \quad - \exp(-\Lambda_0(R_i \exp(\boldsymbol{\beta}^T \mathbf{Z}_i)) \exp(-\boldsymbol{\beta}^T \mathbf{Z}_i)) \} \\ & - \delta_{3i} \Lambda_0(L_i \exp(\boldsymbol{\beta}^T \mathbf{Z}_i)) \exp(-\boldsymbol{\beta}^T \mathbf{Z}_i) \bigg]. \end{aligned} \quad (2.2)$$

Now we show that the model defined in (2.2) is identifiable.

Theorem 2.1. Assume that Λ_0 is not a linear function but it is continuous and $\mathbf{Z}$ is a non-degenerate variable. Then the model defined by (2.1) is identifiable.

The proof of the theorem can be found in Section 2.6.

By the method of maximum likelihood, the estimation of $\boldsymbol{\beta}$ and Λ_0 can be achieved through maximizing the loglikelihood (2.2) over some restricted space, in which $\boldsymbol{\beta} \in B$ with B being a compact subset of $\mathbb{R}^p$ and $\Lambda_0(t)$ is a nonnegative and nondecreasing function of t . In a nonparametric survival

model or the proportional hazards model with interval censored data, estimation of the cumulative baseline hazard Λ_0 often assumes that Λ_0 is a nondecreasing step-function with possible jumps at the observation points L_i and R_i . See for example Groeneboom and Wellner (1992) and Huang and Wellner (1995). The existing estimation methods in this case rely on the fixed stochastic order of observation points, at which unknown Λ_0 is evaluated in the likelihood function. In the AH model, however, the observation points used to evaluate Λ_0 become $L_i \exp(\beta^T Z_i)$ and $R_i \exp(\beta^T Z_i)$, whose order may change with different values of β . In the next section, we therefore propose to approximate the cumulative baseline hazard function Λ_0 using the B-spline method and develop a sieve maximum likelihood estimation procedure that works for such observations with changing stochastic order.

2.2 Estimation Procedure

2.2.1 Sieve Semiparametric Maximum Likelihood Estimator

Given the n -element sample $\mathcal{O}$, we aim to develop an estimation procedure to find the spline-based sieve semiparametric maximum likelihood estimator $(\hat{\beta}_n, \hat{\Lambda}_n)$ that maximizes $l(\beta, \Lambda_0 | \mathcal{O})$ with the constraint that $\beta \in B$ and Λ_0 is in some spline space defined later in this subsection.

Existing spline-based estimators make use of the fixed interval $[\min\{L_i, R_i\}, \max\{L_i, R_i I(R_i < \infty)\}]$, on which the spline space is defined. To account for the changing observed points in the AH model, it is essential to develop a flexible approach to approximate the cumulative baseline haz-

and function on an appropriate interval $[a, b]$ that fits all the observations for updated values of β over the iteration of the estimation procedure.

For a fixed $\beta \in B$, we define a space of polynomial splines in the following way. Denote the set of finite interval endpoints $L_i \exp(\beta^T Z_i)$ and $R_i \exp(\beta^T Z_i) I(R_i < \infty)$ as $\tau_1^\beta < \tau_2^\beta < \dots < \tau_N^\beta$. Then we consider the interval $[\tau_1^\beta, \tau_N^\beta]$ and partition points $d_0 = \tau_1^\beta < d_1 < \dots < d_{K_n} = \tau_N^\beta$ so as to partition the interval into K_n subintervals, where the number of partition points $K_n = O(n^\nu)$ such that $\max_i |d_{i+1} - d_i| = O(n^{-\nu})$ for some constant ν , $0 < \nu < 1$. Denote the set of all partition points by Δ_n^β . First, we define the general spline space as in Section 1.3.2 the following way.

$$\mathcal{S}_n^\beta(\mathcal{P}_m, \Delta_n^\beta) = \left\{ f_n : f_n(t) = \sum_{j=1}^{q_n} c_j s_j^\beta(t) \right\},$$

where $s_j^\beta(t)$ are the polynomial base splines of order m and c_j are corresponding spline coefficients, $j = 1, \dots, q_n$. We need to construct a spline space of continuous, increasing functions similar to (1.7). The monotone spline space for a particular β is defined as

$$\mathcal{M}_n^\beta = \mathcal{M}_n^\beta(\mathcal{P}_m, \Delta_n^\beta) = \left\{ f_n : f_n(t) = \sum_{j=1}^{q_n} c_j s_j^\beta(t), \mathbf{c} \in C_{q_n} \right\},$$

where $C_{q_n} = \{\mathbf{c} = (c_1, \dots, c_{q_n}) : 0 < c_1 < c_2 < \dots < c_{q_n}\}$. Each $f_n \in \mathcal{M}_n^\beta$ is a monotone increasing spline, which can be used for approximating Λ_0 . Plugging the spline estimate of Λ_0 in the log-likelihood l in (2.2), we have

$$\begin{aligned}
l(\boldsymbol{\beta}, c|\mathcal{O}) = \sum_{i=1}^n \Bigg\{ & \delta_{1i} \log \left\{ 1 - \exp \left[- \sum_{j=1}^{q_n} c_j s_j^\beta (R_i \exp(\boldsymbol{\beta}^T \mathbf{Z}_i)) \exp(-\boldsymbol{\beta}^T \mathbf{Z}_i) \right] \right\} \\
& + \delta_{2i} \log \left\{ \exp \left[- \sum_{j=1}^{q_n} c_j s_j^\beta (L_i \exp(\boldsymbol{\beta}^T \mathbf{Z}_i)) \exp(-\boldsymbol{\beta}^T \mathbf{Z}_i) \right] \right. \\
& \quad \left. - \exp \left[- \sum_{j=1}^{q_n} c_j s_j^\beta (R_i \exp(\boldsymbol{\beta}^T \mathbf{Z}_i)) \exp(-\boldsymbol{\beta}^T \mathbf{Z}_i) \right] \right\} \\
& \left. - \delta_{3i} \sum_{j=1}^{q_n} c_j s_j^\beta (L_i \exp(\boldsymbol{\beta}^T \mathbf{Z}_i)) \exp(-\boldsymbol{\beta}^T \mathbf{Z}_i) \right\}.
\end{aligned} \tag{2.3}$$

Denote $\mathcal{M}_n = \cup_{\boldsymbol{\beta} \in B} \mathcal{M}_n^\beta$. The sieve semiparametric maximum likelihood estimator is the pair $(\hat{\boldsymbol{\beta}}_n, \hat{\Lambda}_n)$ that maximizes the loglikelihood $l(\boldsymbol{\beta}, \Lambda_0|\mathcal{O})$ in (2.2) over $B \times \mathcal{M}_n$. Equivalently, it can be obtained by maximizing the loglikelihood $l(\boldsymbol{\beta}, c|\mathcal{O})$ in (2.3) over $B \times C_{q_n}$.

2.2.2 Implementation

Unlike the work of Zhang et al. (2010), the sieve space for Λ_0 here is not a single spline space but a set of spline spaces indexed by $\boldsymbol{\beta} \in B$. Therefore, the existing algorithm to jointly estimate Λ_0 and $\boldsymbol{\beta}$ cannot be directly applied to the current estimation. To facilitate the maximization of (2.3), we propose a modified two-step algorithm allowing one to update the base splines at each iteration step of the estimation procedure. Particularly, we

first rewrite (2.3) in the following form:

$$\begin{aligned}
 l(\boldsymbol{\beta}, \mathbf{c} | \boldsymbol{\beta}', \mathcal{O}) = & \sum_{i=1}^n \left\{ \delta_{1i} \log \left\{ 1 - \exp \left[- \sum_{j=1}^{q_n} c_j s_j^{\boldsymbol{\beta}'} (R_i \exp(\boldsymbol{\beta}^T \mathbf{Z}_i)) \exp(-\boldsymbol{\beta}^T \mathbf{Z}_i) \right] \right\} \right. \\
 & + \delta_{2i} \log \left\{ \exp \left[- \sum_{j=1}^{q_n} c_j s_j^{\boldsymbol{\beta}'} (L_i \exp(\boldsymbol{\beta}^T \mathbf{Z}_i)) \exp(-\boldsymbol{\beta}^T \mathbf{Z}_i) \right] \right. \\
 & \quad \left. - \exp \left[- \sum_{j=1}^{q_n} c_j s_j^{\boldsymbol{\beta}'} (R_i \exp(\boldsymbol{\beta}^T \mathbf{Z}_i)) \right] \exp(-\boldsymbol{\beta}^T \mathbf{Z}_i) \right\} \\
 & \left. - \delta_{3i} \sum_{j=1}^{q_n} c_j s_j^{\boldsymbol{\beta}'} (L_i \exp(\boldsymbol{\beta}^T \mathbf{Z}_i)) \exp(-\boldsymbol{\beta}^T \mathbf{Z}_i) \right\}, \tag{2.4}
 \end{aligned}$$

where the argument $\boldsymbol{\beta}$ corresponds to the regression coefficient presented in the loglikelihood (2.2), and the argument $\boldsymbol{\beta}'$ is used to generate the base splines. Then our algorithm iteratively alternates between two steps: 1) compute estimates $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{c}}$ by maximizing (2.4) over $\boldsymbol{\beta}$ and $\mathbf{c}$ given the spline space with $\boldsymbol{\beta}'$ fixed, and 2) update the spline space $\mathcal{M}_n^{\boldsymbol{\beta}'}$ based on estimated points $L_i \exp(\hat{\boldsymbol{\beta}}^T \mathbf{Z}_i)$ and $R_i \exp(\hat{\boldsymbol{\beta}}^T \mathbf{Z}_i)$. More specifically, in step 1) at the m th iteration of the algorithm, we update the value of the spline coefficients $\hat{\mathbf{c}}^{(m)}$ with fixed observation points $L_i \exp(\hat{\boldsymbol{\beta}}^{(m-1)T} \mathbf{Z}_i)$ and $R_i \exp(\hat{\boldsymbol{\beta}}^{(m-1)T} \mathbf{Z}_i)$, and then compute the updated coefficient $\hat{\boldsymbol{\beta}}^{(m)}$ given the estimated cumulative baseline hazard function $\hat{\Lambda}^{(m-1)}(t) = \sum_{j=1}^{q_n} \hat{c}_j^{(m)} s_j^{\hat{\boldsymbol{\beta}}^{(m-1)}}(t)$. Detailed steps of the estimation procedure are given below.

Step 0 Initialize $\hat{\boldsymbol{\beta}}^{(0)}$ and $\hat{\mathbf{c}}^{(0)}$.

Step 1 In the m th iteration of the algorithm with the fixed value of $\hat{\boldsymbol{\beta}}^{(m-1)}$, consider the loglikelihood function $l(\hat{\boldsymbol{\beta}}^{(m-1)}, \mathbf{c} | \hat{\boldsymbol{\beta}}^{(m-1)}, \mathcal{O})$ and maximize it over the constraint that $\mathbf{c} \in C_{q_n}$. Let $\hat{\mathbf{c}}^{(m)}$ be the vector that maximizes the log-likelihood (2.4).

Step 2 Based on the value $\hat{\mathbf{c}}^{(m)}$, compute updated estimate $\hat{\boldsymbol{\beta}}^{(m)}$ through maximizing the loglikelihood function $l(\boldsymbol{\beta}, \hat{\mathbf{c}}^{(m)} | \hat{\boldsymbol{\beta}}^{(m-1)}, \mathcal{O})$ with respect to $\boldsymbol{\beta}$.

Step 3 Consequently update $\hat{\Lambda}^{(m)}(t) = \sum_{j=1}^{q_n} \hat{c}_j^{(m)} s_j^{\hat{\boldsymbol{\beta}}^{(m)}}(t)$.

Step 4 Iterate Steps 1-3 until convergence, that is, the absolute difference of the loglikelihoods between two consecutive iterations is less than a given threshold ϵ , where ϵ is a small positive value and is typically chosen as 10^{-6} .

The two-stage algorithm above is in line with the general semiparametric iterative estimation approach presented by Cheng (2013), which aims to formalize the approach that is commonly used in semiparametric models. By using the asymptotic parabolic form of $Pl_n(\theta)$ established by Murphy and van der Vaart (2000), Cheng (2013) argues that the sequence of the estimators $\theta_n^{(m)}$ resulting from the iterations of the algorithm tends to the estimator $\hat{\theta}_n$ that maximizes $Pl_n(\theta)$ with respect to $\theta \in \Theta_n$.

Let $\hat{\boldsymbol{\beta}}_n$ and $\hat{\mathbf{c}}_n$ be the values of $\hat{\boldsymbol{\beta}}^{(m)}$ and $\hat{\mathbf{c}}^{(m)}$, respectively, at convergence. Then, $\hat{\Lambda}_n = \sum_{j=1}^{q_n} \hat{c}_{nj} s_j^{\hat{\boldsymbol{\beta}}_n}$ is obtained in the last iteration. As a result, we have $(\hat{\boldsymbol{\beta}}_n, \hat{\Lambda}_n)$ as the sieve semiparametric maximum likelihood estimator for $(\boldsymbol{\beta}, \Lambda_0)$. In our numerical analysis we use the standard quasi Newton optimization method in Step 2. In Step 1, we need to restrict the optimization to the space C_{q_n} of nonnegative increasing sequences of length q_n by applying a logarithmic barrier to the likelihood function. Since the constraints of $\mathbf{c}$ are $c_1 > 0$ and $c_j - c_{j-1} > 0$ for $j = 2, \dots, q_n$, we can

express them in form of $\mathbf{R}\mathbf{c} > 0$ with matrix

$$\mathbf{R} = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ -1 & 1 & 0 & \dots & 0 \\ 0 & -1 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & -1 & 1 \end{bmatrix}.$$

Reparametrizations similar to this are a usual way to calculate parameters under constraint. For example, Wang et al. (2015) compared constrained optimization and unconstrained optimization with transformed variables.

Note that in case of convergence, the values $\hat{\beta}^{(m)}$ and $\hat{\beta}^{(m-1)}$ are very close to each other, hence the difference between $s_j^{\hat{\beta}^{(m)}}$ and $s_j^{\hat{\beta}^{(m-1)}}$ is marginal, so as the difference between $\sum_{j=1}^{q_n} c_j^{(m-1)} s_j^{\beta^{(m-1)}}$ and $\sum_{j=1}^{q_n} c_j^{(m)} s_j^{\beta^{(m)}}$. This ensures the convergence of the spline estimator $\hat{\Lambda}_n$ when $\hat{\beta}^{(m)}$ and $\hat{\mathbf{c}}^{(m)}$ converge.

2.2.3 Choosing the number of base splines

The implementation of the proposed estimation requires the number of base splines as well as the degree of base polynomials. Following Zhang and Davidian (2008), we determine the number of base splines in this work by the Hannan–Quinn (HQ) criterion, which is of the form

$$HQ = -2l_{\max} + 2k \log(\log(n)), \quad (2.5)$$

where $l_{\max}$ is the loglikelihood function defined in (2.3) and evaluated at the maximum values $(\hat{\beta}_n, \hat{c}_n)$, and k is the number of free parameters $k = q_n + p$. As noted by Zhang and Davidian (2008), the HQ criterion tends to select an

intermediate model compared to the standard Akaike information criterion (AIC) and the Bayesian information criterion (BIC), which often choose larger and smaller models, respectively.

Based on our experience in numerical studies, we suggest the estimation of the cumulative baseline hazard function using quadratic or cubic base splines, while the number of base splines is selected within a range of 4 to 10.

2.3 Large sample properties

In this section, express the likelihood in terms of U_i and V_i , instead of L_i and R_i to allow conditional independence.

For theoretical reasons, we work with the logarithm of the spline estimators. Denote the natural logarithm of all functions in $\mathcal{M}_n$ and $\mathcal{M}_n^\beta$ by $\overline{\mathcal{M}}_n$ and $\overline{\mathcal{M}}_n^\beta$, respectively. For any $\Lambda \in \mathcal{M}_n$ with any subscript, let $\phi = \log(\Lambda)$. In order to set up the convergence, let $\theta_n = (\hat{\beta}_n, \log \hat{\Lambda}_n)$ be the estimated parameters and function, and $\theta_0 = (\beta_0, \log \Lambda_0)$ be their corresponding true values. The distance between any two sets of parameters $\theta_1 = (\beta_1, \phi_1)$ and $\theta_2 = (\beta_2, \phi_2)$ in $B \times \overline{\mathcal{M}}_n$ is defined as

$$d(\theta_1, \theta_2) = \left(\|\beta_1 - \beta_2\|^2 + \|\phi_1 - \phi_2\|_{\Phi(\beta_1, \beta_2)}^2 \right)^{1/2},$$

where

$$\begin{aligned} \|\phi_1 - \phi_2\|_{\Phi(\beta_1, \beta_2)}^2 &= E \left[\left(\phi_1(U \exp(\beta_1^T \mathbf{Z})) - \phi_2(U \exp(\beta_2^T \mathbf{Z})) \right)^2 \right] \\ &\quad + E \left[\left(\phi_1(V \exp(\beta_1^T \mathbf{Z})) - \phi_2(V \exp(\beta_2^T \mathbf{Z})) \right)^2 \right] \end{aligned}$$

and $\Phi(\beta_1, \beta_2)$ is the distance between the two functions ϕ_1 and ϕ_2 and depends on parameters β_1 and β_2 . It is worth to notice that, besides the

first term, the second term in distance d is also related to parameters β , and thus it makes sense to compare ϕ_1 and ϕ_2 incorporating β_1 and β_2 .

The following conditions are necessary and most of them are often used regularity conditions for convergence of the maximum likelihood estimator under interval censoring.

- (B1) There exists a bound M_Z such that $|Z| \leq M_Z$ almost surely, and $E(ZZ^T)$ is nonsingular.
- (B2) The parameter space B is compact, therefore, there exists a bound M_B such that for any $\beta \in B$, $\|\beta\| < M_B$.
- (B3) There exists a finite interval $[a, b] \subset R^+$ such that the supports of U and V are subsets of $[a, b]$ and there exists a constant $\zeta > 0$ such that $P(V - U > \zeta) = 1$.
- (B4) The first derivative of $\log \Lambda_0$ is strictly positive. Furthermore, $\log \Lambda_0$ has bounded r th derivatives on the interval $[a \exp(-M_B M_Z), b \exp(M_B M_Z)]$.

Condition (B1) to (B4) are usual conditions in the analysis of interval censored data. They are generally needed to make sure that both the likelihood function and its derivative are bounded, similar to those conditions used in Zhang et al. (2010).

Now we establish the consistency for estimator θ_n in the following theorem.

Theorem 2.2. Given Conditions (B1)–(B4) the estimator θ_n converges to θ_0 in probability.

Theorem 2.3 (Rate of convergence). Assume that Conditions (B1) – (B4) hold then $d(\theta_n, \theta_0) = O_P(n^{-\min((1-\nu)/2, r\nu)})$, where r is an order of smoothness introduced in Condition (B4).

The techniques used for proving these theorems are mainly based on the empirical process theory developed in van der Vaart and Wellner (1996) and van der Vaart (1998). Details of the proof can be found in Section 2.6.

We show the consistency of the resulting estimator and its convergency rate in Theorems 2-3, but have not obtained the explicit form of the asymptotic variance yet. Note that the variance estimate of $\hat{\beta}$ can be obtained through bootstrap sampling. There are some other typical methods developed in the literature for estimating the information matrix for the finite-dimension parameter of a semiparametric model, see Huang et al. (2008). For instance, a straightforward method is to directly use the second derivative of the profile likelihood function of β , while our simulation experience shows that the profile likelihood function of the AH model is not smooth enough for numerically calculating derivatives. Another way is to use a least square calculation, which requires the derivative of the loglikelihood function with respect to β . In the case of the AH model, however, the least square calculation involves an untractable derivative of the nonparametric part.

In our numerical analysis, the standard errors of parameters are estimated based on bootstrap samples $\mathcal{O}^{B_i}$ for $i = 1, \dots, n_B$, which can be generated by randomly drawing n observations from the sample data with replacement. We compute the estimates $\hat{\beta}^{B_i}$ through the proposed estimation procedure and use the sample standard deviation and sample distribution of $\hat{\beta}^{B_i}$, $i = 1, \dots, n_B$ to estimate the standard error and approximate the distribution of $\hat{\beta}_n$, respectively. In addition, the 0.025 and 0.975 quan-

tiles of the bootstrap estimates $\hat{\beta}^{B_i}$ can be used to construct an approximate 95% confidence interval of $\hat{\beta}_n$.

2.4 Simulations

We carry out simulation studies to evaluate the performance of the proposed estimation procedure under two scenarios with convex and concave cumulative baseline hazards, respectively. The maximization of log-likelihood in steps 1 and 2 of the proposed estimation procedure can be performed using the standard optimization functions in R, such as `optimize` and `optim` for one-dimensional and multi-dimensional parameters, respectively. Existing R package `Splines` is also used for generating the base spline polynomials.

In the first scenario, the true survival times T_i have been generated independently based on a convex cumulative baseline hazard function $\Lambda_0(t) = (2/3)t^{3/2}$, corresponding to Weibull random variables T_i . We set regression parameter $\beta = 0.5$ and generate covariates Z_i according to the Bernoulli distribution with parameter 0.5. For the purpose of censoring, two independent random variables are generated: W_{i1} from the uniform distribution on interval $[0.2, 0.5]$ and W_{i2} of the exponential distribution with rates 0.035, 0.23, and 0.38, providing the desired right censoring rates of less than 5%, between 12% and 18%, between 25% and 30%, respectively. If $W_{i1} + W_{i2} < T_i$, then the individual i is right censored with censoring time $L_i = W_{i1} + W_{i2}$ and $R_i = \infty$. If $T_i \leq W_{i1} + W_{i2}$, then the left and right censoring times are selected as $L_i = kW_{i1}$ and $R_i = (k+1)W_{i1}$, respectively, with an appropriate integer k such that $L_i < T_i \leq R_i$.

We then apply the estimation procedure proposed in Section 2.2 to each simulated sample. The parameters of the splines are the following. The

degree of the base spline polynomials are set to 2 and 3 corresponding to quadratic and cubic base splines, respectively. The number of internal knots are set to 1, 2 and 3. After fitting the models with these 6 combinations of degrees and number of knots, we use the HQ information criterion to select the best polynomial design.

500 samples were generated with sample sizes of $n = 200, 500, 1000$. The simulation results for the first scenario are summarized in Table 2.1 in terms of relative bias (Rbias) defined by $meanbias(\beta)/\beta$, standard deviation (SD) of $\hat{\beta}$ and coverage probability (CP) computed as the proportion of estimates $\hat{\beta}$ being at most 1.96 times SD from the true value β . The table shows that the relative bias and SD decrease with increased sample size, while the relative bias increases as the level of right censoring rate goes up. In Figure 2.1, we can see that the average estimate of the cumulative baseline hazard function is very close to the corresponding true function in each simulation setting and the estimate improves with increasing sample size. However, increased right censoring rate seems have no significant effects on the estimated cumulative baseline hazard function.

In the second scenario of the simulation, we demonstrate the performance of the proposed method with a concave cumulative baseline hazard. We generate T_i according to the AH model (1) with $\beta = 1$ and $\Lambda_0(t) = 2\sqrt{0.01 + t}$. The censoring variables W_{i1} were generated based on the uniform distribution on interval $[0.02, 0.05]$. W_{i2} are generated as minimum of a random variable ω and a constant c , where ω follows the exponential distribution with rate $\mu = 0.02, 0.22$ and 0.45 and $c = 10, 5$ and 4 for different levels of right censoring. Interval censored data (L_i, R_i) are then generated similarly to those in the first scenario with sample sizes of 200, 500 and 1000.

Table 2.1: *Simulation results of the estimated regression coefficient in scenario 1 with the Weibull cumulative baseline hazard function*

n	Censoring	Rbias	SD	CP
200	0%-5%	-0.102	0.328	0.958
	12%-18%	-0.084	0.330	0.964
	20%-25%	-0.116	0.356	0.970
500	0%-5%	0.001	0.209	0.932
	12%-18%	0.008	0.217	0.946
	20%-25%	0.019	0.239	0.932
1000	0%-5%	0.004	0.090	0.948
	12%-18%	0.004	0.103	0.948
	20%-25%	0.006	0.096	0.944

(a) $n=200$ (b) $n=500$ (c) $n=1000$

Figure 2.1: *Estimated cumulative baseline hazard functions in scenario 1 with censoring rates 0% to 5% (left panel), 12% to 18% (middle panel) and 20% to 25% (right panel). In each plot, the solid line gives the true cumulative baseline hazard, the dashed line corresponds to the mean of estimated baseline cumulative hazards, and the dotted lines are the upper and lower 2.5% quantiles of the estimates.*

We implement the proposed method for 500 replications in each simulation setting, and report results in Table 2.2. As seen from the table, when Λ_0 is concave, the estimates of β tend to have bigger relative biases and SDs in comparison with the corresponding results obtained from scenario 1 with a convex Λ_0 . However, the relative bias and the standard deviation decrease fast as the sample size increases. We further demonstrate in Figure 2.2 the estimated cumulative baseline hazard curves under all simulation settings within scenario 2. The mean of estimated cumulative baseline hazard curves is comparable with the corresponding true function. Although it is slightly worse than that in scenario 1 for $n = 200$, it gets closer to the truth with larger sample size. Findings in scenario 2 indicate that a large sample size would be required to achieve reasonably good estimates for both regression parameter β and the cumulative baseline hazard Λ_0 when Λ_0 is concave.

2.5 Application

We apply our method to the diabetes conversion data from a screening and intervention study of 1519 Danish individuals with high-risk of type 2 diabetes. The study was a part of the ADDITION project (Rasmussen et al., 2008), aiming to examine if certain intervention measures can reduce the conversion risk from prediabetes to type 2 diabetes mellitus. The time to event of interest in this study is the time (in days) of conversion to type 2 diabetes. However, it cannot be observed exactly in practice. Laboratory tests were carried out to measure the fasting capillary blood glucose (FBG) and 2 hour capillary blood glucose (2hBG) levels for each individual. If either the FBG is higher than the critical level of 6.1 mmol/l or the 2hBG is greater than 11.1 mmol/l, the individual would be diagnosed with diabetes.

Table 2.2: *Simulation results of the estimated regression coefficient in scenario 2 with concave cumulative baseline hazard function*

n	Censoring Rate	Rbias	SD	CP
200	0-5%	0.183	0.475	0.884
	12-18%	0.190	0.486	0.890
	20-25%	0.202	0.503	0.880
500	0-5%	0.119	0.305	0.926
	12-18%	0.098	0.313	0.930
	20-25%	0.128	0.341	0.920
1000	0-5%	0.081	0.201	0.928
	12-18%	0.074	0.225	0.946
	20-25%	0.094	0.223	0.942

Figure 2.2: *Estimated cumulative baseline hazard functions in scenario 2 with censoring rates 0% to 5% (left panel), 12% to 18% (middle panel) and 20% to 25% (right panel). In each plot, the solid line gives the true cumulative baseline hazard, the dashed line corresponds to the mean of estimated baseline cumulative hazards, and the dotted lines are the upper and lower 2.5% quantiles of the estimates.*

The event of interest took place between the last negative and the first positive tests, hence the survival time was interval censored. The dataset can be extracted from the R package Epi.

At the beginning of the study, individuals were classified into two categories through a screen program. 908 individuals in one category were found exhibiting signs of impaired fasting glycaemia (IFG) only, while 611 individuals in the other category had both impaired glucose tolerance (IGT) and IFG. The isolated IFG was diagnosed by $5.6 \text{ mmol/l} \leq \text{FBG} < 6.1 \text{ mmol/l}$ and $2\text{hBG} < 7.8 \text{ mmol/l}$, while the combined IFG and IGT was defined by $\text{FBG} < 6.1 \text{ mmol/l}$ and $7.8 \text{ mmol/l} \leq 2\text{hBG} < 11.1 \text{ mmol/l}$. Through a survey conducted in the study, the investigators also obtained a dichotomized variable of lifestyle changes of the individuals to indicate health improvements vs. no improvement (or worsening), which were referred to as the intervention and control groups with 871 and 648 individuals, respectively.

In the work of Rasmussen et al. (2008), a preliminary analysis was performed based on a relatively simple incidence model as well as a multiplicative relative risk model by Farrington (1996). They found that the incidence rate of diabetes was much higher in the first 1.5 years than that in the rest of the study period, indicating a decreasing hazard function. Note that their model was a simplified version of the proportional hazards model with a piecewise constant baseline hazard function and their estimated baseline hazard function was based on two distinct intervals only, instead of the entire sample of interval censored event data. We now further analyze the data with the proposed AH model.

We fit the AH model using the estimation method developed in Section 3 in comparison with the PH model for interval censored data using the method proposed by Pan (1999). Two binary covariates are considered:

impaired glucose tolerance (IGT = 1 for an individual having both IFG and IGT, 0 for IFG only), and intervention (1 for a subject in the intervention group, 0 otherwise). By the HQ criterion, we found that the AH model with spline approximations for the baseline hazard fits the data best when using 7 quadratic spline basis functions. The fitting results of both the AH and PH models are summarized in Table 2.3, where the estimated confidence intervals of the regression parameters in the AH model are calculated using the bootstrap approach given in Section 4. The estimated regression coefficients indicate that the intervention leads to a slight reduction of the risk for type 2 diabetes, while the subject having syndrome of both IFG and IGT tends to have a higher risk. However, all covariate effects are not statistically significant in both models since the corresponding 95% confidence intervals include zero.

Table 2.3: *Analysis of the diabetes conversion data*

	PH		AH	
	IGT	Intervention	IGT	Intervention
Estimated value	0.355	-0.123	0.041	-0.068
95% CI	(-0.104,0.215)	(-0.225,0.264)	(-0.112,0.174)	(-0.221,0.251)
HQ	4088.28		2424.32	

Following the suggestion of Zhang and Davidian (2008), we also apply the HQ criterion to compare the fittings of the AH and PH models. The HQ for the PH model is calculated similarly to (2.5) with the value of the loglikelihood evaluated at its maxima and the number of free parameters k defined as the sum of the number of regression coefficients and the num-

ber of distinct values of the estimated cumulative baseline hazard function. We obtain that $HQ=2421$ for the AH model and 4088 for the PH model, indicating that the AH model fits the data better than the PH model. To illustrate the performance of the proposed method, we show in Figure 2.3 the estimated cumulative baseline hazard function under the AH model in comparison with that under the PH model and its empirically nonparametric estimate, that we obtained by the ICM algorithm by Groeneboom and Wellner (1992). It can be seen that our estimated cumulative baseline hazard curve is closer to the nonparametric estimate of Λ_0 than that obtained in the PH model.

To further examine the impact of covariates on cumulative hazard, we fit the data using different models with intervention as the single covariate for the IGT+IFG and IFG groups, respectively. The estimated cumulative baseline hazard curves are presented in Figure 2.4. We observe that in the IGT+IFG group the AH estimated cumulative hazard goes closer to the nonparametric estimate but the estimated curve under the PH model behaves differently from them in the later study period. However, all three estimated curves are quite similar for the IFG group. This finding indicates that the proposed AH model fits the data better than the PH model. We also performed separate analyses for the control group and the intervention group. Figure 2.5 shows the estimated cumulative hazards obtained by the AH model and the empirical nonparametric method. Both the IGT + IFG and IFG groups produce similar cumulative hazards in AH and nonparametric models type of model and whether the intervention is received. However, as expected, subjects receiving the intervention in the IFG group tend to have slightly lower cumulative hazard towards the end of the study than subjects under control.

Figure 2.3: *Estimated cumulative baseline hazard curves under the AH and PH models in comparison with the empirical nonparametric cumulative baseline hazard function.*

2.6 Technical proofs

2.6.1 Proof of Theorem 2.1

Suppose that we have two parameter sets (β, Λ_0) and (β^*, Λ_0^*) . We need to show that for every O_i in $\mathfrak{X}$ and for the likelihood l of one element sample, $l(\beta, \Lambda_0|O_i) = l(\beta^*, \Lambda_0^*|O_i)$ implies $(\beta, \Lambda_0) = (\beta^*, \Lambda_0^*)$. First, we note that the underlying censoring process that yields L_i and R_i is conditionally independent of T_i given $\mathbf{Z}_i$. From Proposition 1 of Oller et al. (2004), we

Figure 2.4: *Estimated cumulative hazard functions based on ICM nonparametric estimate, and estimates given by the PH and the AH models*

have that the model satisfies the constant-sum property defined in that paper. By Theorem 1 of Oller et al. (2007) the constant-sum property implies that if $l(\beta, \Lambda_0 | O_i) = l(\beta^*, \Lambda_0^* | O_i)$ in the interval censored setting, then $S(t | \mathbf{Z}_i; \beta, \Lambda_0) = S(t | \mathbf{Z}_i; \beta^*, \Lambda_0^*)$ for almost every t . So we only need to show that $S(t | \mathbf{Z}_i; \beta, \Lambda_0) = S(t | \mathbf{Z}_i; \beta^*, \Lambda_0^*)$ implies $(\beta, \Lambda_0) = (\beta^*, \Lambda_0^*)$.

In the proof, we make use of the assumption that the parameters Λ_0 and Λ_0^* are differentiable functions so their derivatives are continuous. By taking the negative logarithm of $S(t | \mathbf{Z}_i; \beta, \Lambda_0) = S(t | \mathbf{Z}_i; \beta^*, \Lambda_0^*)$ and differentiating both sides with respect to t , we have

$$\lambda_0(t \exp(\beta^T \mathbf{Z}_i)) = \lambda_0^*(t \exp(\beta^{*T} \mathbf{Z}_i)).$$

Observe that λ_0 and λ_0^* only differ by a scale. By substituting $t' = t \exp(-\beta^T \mathbf{Z}_i)$

Figure 2.5: *Estimated cumulative hazard functions based on nonparametric estimate (NP), and estimates given AH models for the control group on the left and for the intervention group on the right.*

into the equation above, we have

$$\lambda_0(t') = \lambda_0^*(t' \exp((\beta^* - \beta)^T \mathbf{Z}_i)). \quad (2.6)$$

The left hand side of (2.6) does not depend on $\mathbf{Z}_i$, so for a fixed t' , the right hand side must be the same for all possible values of $\mathbf{Z}_i$. Namely, take a $\mathbf{Z}'_i$ that differs from $\mathbf{Z}_i$ in a component whose corresponding coefficient in $\beta^* - \beta$ is non-zero. This can be done because $\mathbf{Z}$ is a non-degenerate variable by the conditions of the theorem. Then using $\mathbf{Z}_i$ and $\mathbf{Z}'_i$ we have

$$\lambda_0^*(t' \exp((\beta^* - \beta)^T \mathbf{Z}_i)) = \lambda_0^*(t' \exp((\beta^* - \beta)^T \mathbf{Z}'_i)),$$

by applying the same scaling method as above, we get

$$\lambda_0^*(t') = \lambda_0^*(t' \exp\{(\beta^* - \beta)^T (\mathbf{Z}'_i - \mathbf{Z}_i)\}).$$

Let $\eta = (\beta^* - \beta)^T(\mathbf{Z}'_i - \mathbf{Z}_i)$. By the choice of $\mathbf{Z}'_i$, $\eta \neq 0$, and $\lambda_0^*(t') = \lambda_0^*(t' \exp(\eta))$ for all R_i . By using this equation several times, it is easy to show that $\lambda_0^*(t) = \lambda_0^*(t \exp(r\eta))$ for all positive rational r . That means λ_0^* takes a constant value on a dense subset of possible R_i values. By using the property that λ_0^* is continuous, it is a constant function. This contradicts the assumption of the theorem, hence $\beta = \beta^*$. That, along with (2.6) implies $\lambda_0^* = \lambda_0$.

2.6.2 Proof of Theorem 2.2

The main technique and notations used in this proof basically follow the works of van der Vaart and Wellner (1996) and van der Vaart (1998). Namely, $l(\theta|O)$ is the likelihood of the one element sample, as introduced in Section 2.3. $\mathbb{P}l(\theta|O)$ denotes the expectation of the likelihood function of one element sample with respect to the true model and $\mathbb{P}_n l(\theta|O)$ the expectation with respect to the empirical distribution based on the n -element sample. Throughout the proof, we denote an arbitrary positive constant by C , which does not depend on n , θ_0 .

We aim to show that the estimator of θ is consistent by using Theorem 5.7 of van der Vaart (1998). To this end, we need to check if the three conditions in Theorem 5.7 are satisfied for $M(\theta) = \mathbb{P}l(\theta|O)$ and $M_n(\theta) = \mathbb{P}_n l(\theta|O)$. The first condition to verify is that

$$\sup_{\theta \in \Theta} |M_n(\theta) - M(\theta)| \rightarrow 0.$$

One can construct ϵ -brackets in a way similar to that of Zhang et al. (2010) where ϵ is a small positive number. Since B is compact by condition (B2), it can be covered by balls of radius ϵ . So for every $\beta \in B$, there exists some β_s such that $|\beta - \beta_s| < \epsilon$, where β_s is one of the ball centres, $1 \leq s \leq C(1/\epsilon)^p$

for positive constant C . Since $|\mathbf{Z}|$ is bounded by $M_{\mathbf{Z}}$, $|\boldsymbol{\beta}^T \mathbf{Z} - \boldsymbol{\beta}_s^T \mathbf{Z}| < \epsilon M_{\mathbf{Z}}$. Moreover, $\exp(\boldsymbol{\beta}_s^T \mathbf{Z})\xi^{-1} \leq \exp(\boldsymbol{\beta}^T \mathbf{Z}) \leq \exp(\boldsymbol{\beta}_s^T \mathbf{Z})\xi$ by the mean value theorem for some $1 < \xi < 1 + C\epsilon$ and for small enough $\epsilon > 0$. From the work of Shen and Wong (1994, p. 597), using $1/(2r+2) < \nu < 1/(2r)$, for each $\overline{\mathcal{M}}_n^{\beta_s}$, $1 \leq s \leq C(1/\epsilon)^p$ there exists a set of brackets $\{\phi_{si}^1, \phi_{si}^2\}$, $1 \leq i \leq \epsilon^{-Cq_n}$ with the following properties. First, for every $\phi \in \overline{\mathcal{M}}_n^{\beta_s}$ there is an i such that $\phi_{si}^1(u) \leq \phi(u) \leq \phi_{si}^2(u)$ for every u . Second, for every i , $1 \leq i \leq \epsilon^{-Cq_n}$, and for every u , $\phi_{si}^2(u) - \phi_{si}^1(u) < \epsilon$. That implies $\mathbb{P}_n|\phi_{si}^2(U) - \phi_{si}^1(U)| \leq C\epsilon$ and $\mathbb{P}_n|\phi_{si}^2(V) - \phi_{si}^1(V)| \leq C\epsilon$.

Now we can construct the brackets $[l_{si}^1, l_{si}^2]$, $1 \leq s \leq C(1/\epsilon)^p$, $1 \leq i \leq \epsilon^{-Cq_n}$ for $\mathfrak{L}_1 = \{l(\theta|O), \theta \in \Theta_n\}$, where

$$\begin{aligned} l_{si}^1(O) &= \delta_1 \log [1 - \exp\{-\exp(-\boldsymbol{\beta}_s^T \mathbf{Z} + \phi_i^1(U \exp(\boldsymbol{\beta}_s^T \mathbf{Z})\xi^{-1}) - C\epsilon)\}] \\ &\quad + \delta_2 \log [\exp\{-\exp(-\boldsymbol{\beta}_s^T \mathbf{Z} + \phi_i^2(U \exp(\boldsymbol{\beta}_s^T \mathbf{Z})\xi) + C\epsilon)\} \\ &\quad - \exp\{-\exp(-\boldsymbol{\beta}_s^T \mathbf{Z} + \phi_i^1(U \exp(\boldsymbol{\beta}_s^T \mathbf{Z})\xi^{-1}) - C\epsilon)\}] \\ &\quad - \delta_3 \exp(-\boldsymbol{\beta}_s^T \mathbf{Z} + \phi_i^2(U \exp(\boldsymbol{\beta}_s^T \mathbf{Z})\xi) + C\epsilon) \end{aligned}$$

and

$$\begin{aligned} l_{si}^2(O) &= \delta_1 \log [1 - \exp\{-\exp(-\boldsymbol{\beta}_s^T \mathbf{Z} + \phi_i^2(U \exp(\boldsymbol{\beta}_s^T \mathbf{Z})\xi) + C\epsilon)\}] \\ &\quad + \delta_2 \log [\exp\{-\exp(-\boldsymbol{\beta}_s^T \mathbf{Z} + \phi_i^1(U \exp(\boldsymbol{\beta}_s^T \mathbf{Z})\xi^{-1}) - C\epsilon)\} \\ &\quad - \exp\{-\exp(-\boldsymbol{\beta}_s^T \mathbf{Z} + \phi_i^2(U \exp(\boldsymbol{\beta}_s^T \mathbf{Z})\xi) + C\epsilon)\}] \\ &\quad - \delta_3 \exp(-\boldsymbol{\beta}_s^T \mathbf{Z} + \phi_i^1(U \exp(\boldsymbol{\beta}_s^T \mathbf{Z})\xi^{-1}) - C\epsilon). \end{aligned}$$

These brackets satisfy that there exist i and s such that $l_{si}^1(O) \leq l(\theta|O) \leq l_{si}^2(O)$ for any $l(\theta, O) \in \mathfrak{L}_1$. By the mean value theorem and the Taylor expansion of some terms in $l_{si}^2(O) - l_{si}^1(O)$, one can prove that $|\mathbb{P}_n[l_{si}^1(O) - l_{si}^2(O)]| < C\epsilon$ for all i and s . Let $N_{[]}(\epsilon, \mathfrak{L}_1, L_1(\mathbb{P}_n))$ be the ϵ -bracketing number and $N(\epsilon, \mathfrak{L}_1, L_1(\mathbb{P}_n))$ the ϵ -covering number of $\mathfrak{L}_1$ with respect to

the $L_1(\mathbb{P}_n)$ measure. Then the ϵ -bracketing number of $\mathfrak{L}_1$ with $L_1(\mathbb{P}_n)$ norm is bounded by $C(1/\epsilon)^{p+q_n}$. Using the inequality $N(\epsilon, \mathfrak{L}_1, L_1(\mathbb{P}_n)) \leq N_{[]} (2\epsilon, \mathfrak{L}_1, L_1(\mathbb{P}_n))$ and Theorem 2.4.3 of van der Vaart and Wellner (1996), it follows that $\mathfrak{L}_1$ is Glivenko-Cantelli and hence $\sup_{\theta \in \Theta} |M_n(\theta) - M(\theta)| \rightarrow 0$.

The second condition we need to verify is that $\sup_{\theta, d(\theta, \theta_0) > \epsilon} M(\theta) < M(\theta_0)$. By Gibbs' inequality, we have that $M(\theta) \leq M(\theta_0)$ for all $\theta \in \Theta_n$ with equality holds if and only if $l(\theta, O) = l(\theta_0, O)$ for almost every O in $\mathfrak{X}$. Assume that $\sup_{\theta, d(\theta, \theta_0) > \epsilon} M(\theta) = M(\theta_0)$. Then there exists a sequence θ_m such that $M(\theta_m) \rightarrow \sup_{\theta, d(\theta, \theta_0) > \epsilon} M(\theta)$ and $d(\theta_m, \theta_0) > \epsilon$. Since B is compact and the coefficients of the splines in $\overline{\mathcal{M}}_n$ are uniformly bounded, θ_m has a subsequence θ_{mk} converging to θ_{m0} , where θ_{m0} may not be in Θ_n , but in Θ . It is easy to see that $M(\theta)$ is continuous in θ , so $M(\theta_{m0}) = \sup_{\theta, d(\theta, \theta_0) > \epsilon} M(\theta)$. By Theorem 2.1, $\theta_{m0} = \theta_0$. However, $d(\theta_{mk}, \theta_0) > \epsilon$, so θ_{mk} cannot converge to θ_0 . Because of this contradiction, the condition $\sup_{\theta, d(\theta, \theta_0) > \epsilon} M(\theta) < M(\theta_0)$ holds.

The third condition to verify is $M_n(\theta_n) \geq M_n(\theta_0) - o_p(1)$. We use the same argument as in Zhang et al. (2010). By Corollary 6.21 of Schumaker (2007), there exists a $\phi_{0,n}$ such that $\|\phi_0 - \phi_{0,n}\|_{\Phi(\beta_0)} \leq cq_n^{-p\nu}$.

$$\begin{aligned} M_n(\theta_n) - M_n(\theta) &\geq M_n(\theta_n) - M_n(\theta_{0,n}) + M_n(\theta_{0,n}) - M_n(\theta_0) \\ &\geq \mathbb{P}_n l(\theta_{0,n}|O) - \mathbb{P}_n l(\theta_0|O) \\ &= (\mathbb{P}_n - \mathbb{P})(l(\theta_{0,n}|O) - l(\theta_n|O)) + M(\theta_{0,n}) - M(\theta_0), \end{aligned}$$

where the first inequality holds because $M_n(\theta_n) - M_n(\theta_{0,n}) \geq 0$.

Furthermore, we show that $0 \geq M(\theta_{0,n}) - M(\theta_0) = -o_p(1)$. It is easy to see that $|l(\beta_0, \phi_{0,n}|O) - l(\beta_0, \phi_0|O)|$ tends to 0 at every O . Given the definition of $\phi_{0,n}$, there exists a ϕ_M such that for every $n > 0$, $|l(\beta_0, \phi_M|O) - l(\beta_0, \phi_0|O)| \geq |l(\beta_0, \phi_{0,n}|O) - l(\beta_0, \phi_0|O)|$ for every O except possibly a set

of measure 0. The claim follows from the dominated convergence theorem.

Let us define the class $\mathfrak{L}_2 = \{l(\beta_0, \phi|O) - l(\beta_0, \phi_0|O), \phi \in \mathcal{M}_n(\beta_0), \|\phi - \phi_0\| < cn^{-p\nu}\}$, so that $l(\beta_0, \phi_{0,n}|O) - l(\beta_0, \phi_0|O)$ is also in $\mathfrak{L}_2$. Using the brackets $l_{si}^1(O) - l(\beta_0, \phi_0|O)$ and $l_{si}^2(O) - l(\beta_0, \phi_0|O)$ as before, the bracketing number for $\mathfrak{L}_2$ with $L_2(\mathbb{P})$ -norm is bounded by $(1/\epsilon)^{cq_n}$. So the bracketing integral

$$\begin{aligned} J(\delta, \mathfrak{L}_2, L_2(\mathbb{P})) &= \int_0^\delta \sqrt{1 + \log N_{[]}(\epsilon, \mathfrak{L}_2, L_2(\mathbb{P}))} \, d\epsilon \\ &\leq \int_0^\delta \sqrt{1 - cq_n \log \epsilon} \, d\epsilon < \infty \end{aligned}$$

is finite. Hence by the Donsker Theorem 19.5 in van der Vaart (1998), $\mathfrak{L}_2$ is $\mathbb{P}$ -Donsker. By the first two statements of Corollary 2.3.12 in van der Vaart and Wellner (1996),

$$(\mathbb{P}_n - \mathbb{P})[l(\beta_0, \phi_{0,n}|O) - l(\beta_0, \phi_0|O)] = o_p(n^{-1/2}).$$

Therefore

$$M(\theta_n) - M(\theta_0) \geq o_p(n^{-1/2}) - o_p(1) = -o_p(1),$$

Thus we obtain the convergence $d(\theta_n, \theta_0) \rightarrow 0$ in probability.

2.6.3 Proof of Theorem 2.3

We prove the rate of convergence using Theorem 3.4.1 of van der Vaart and Wellner (1996). Let M_n be the random function and M be the deterministic function in the theorem for all n . In the following, we verify the conditions of that theorem along the same line as that of Theorem 2 in Zhou et al. (2017).

As in the proof of in Theorem 2, by Lu et al. (2007) there exists function $\phi_{0,n} \in \overline{\mathfrak{M}}_n$ such that $\|\phi - \phi_{0,n}\|_\infty = Cq_n^{-r} = O(n^{-r\nu})$. Let $\theta_{0,n} = (\beta_0, \phi_{0,n})$,

we have $d(\theta_0, \theta_{0,n}) = O(n^{-r\nu})$. By similar calculations to the number of brackets in the proof of Theorem 2.2, one can show that $M(\theta_0) - M(\theta_{0,n}) \leq cd^2(\theta_0, \theta_{0,n}) = O(n^{-2r\nu})$.

Let $\eta > 0$ be a small constant and define $\mathfrak{L}_\eta = \{l(\theta|O) - l(\theta_{0,n}|O) : \theta \in B \times \overline{\mathcal{M}}_n, \eta/2 < d(\theta, \theta_{0,n}) < \eta\}$. Using the previous statement, and the fact that $M(\theta_0) - M(\theta) > Cd^2(\theta_0, \theta)$ in a neighborhood of θ_0 , for all $l(\theta, O) - l(\theta_{0,n}, O) \in \mathfrak{L}_\eta$ and for large enough n ,

$$\begin{aligned} \mathbb{P}(l(\theta|O) - l(\theta_{0,n}|O)) &= M(\theta) - M(\theta_{0,n}) \\ &= M(\theta) - M(\theta_0) + M(\theta_0) - M(\theta_{0,n}) \\ &\leq -C\eta^2 + Cn^{-2r\nu} \leq -c\eta^2, \end{aligned}$$

and the two sides of the inequality verifies first condition.

Secondly, we need to find the appropriate function $\psi_n(\eta)$ so that for small enough η

$$E \left[\sup_{\theta: \eta/2 < d(\theta, \theta_0) < \eta} |(M_n - M)(\theta) - (M_n - M)(\theta_{0,n})| \right] < \frac{\psi_n(\eta)}{\sqrt{n}}.$$

Under Conditions (B1) – (B3), $\mathfrak{L}_\eta$ is uniformly bounded. Moreover, with some algebraic manipulations similar to the one in case of brackets and for small enough η , $\mathbb{P}(l(\theta|O) - l(\theta_{0,n}|O))^2 \leq c\eta^2$ for some positive constant c . By applying Lemma 3.4.2 of van der Vaart and Wellner (1996) we get

$$E[\|\sqrt{n}(M_n - M)\|_{\mathfrak{L}_\eta}] \leq CJ_{[]}(\eta, \mathfrak{L}_\eta, L_2(\mathbb{P})) \left[1 + \frac{J_{[]}(\eta, \mathfrak{L}_\eta, L_2(\mathbb{P}))}{\eta^2 n^{1/2}} \right],$$

where $\|\cdot\|_{\mathfrak{L}_\eta}$ is the uniform bound of the operator over the space $\mathfrak{L}_\eta$ and $J_{[]}(\eta, \mathfrak{L}_\eta, L_2(\mathbb{P})) = \int_0^\eta [1 + \log N_{[]}(\epsilon, \mathfrak{L}_\eta, L_2(\mathbb{P}))]^{1/2} d\epsilon$. We use the calculations similar to that of Shen and Wong (1994, p. 597) to obtain

$$\log N_{[]}(\epsilon, \mathfrak{L}_\eta, L_2(\mathbb{P})) \leq C(q_n + p) \log(\eta/\epsilon)$$

for every $0 < \epsilon < \eta$. So $J_{[\cdot]}(\eta, \mathfrak{L}_\eta, L_2(\mathbb{P})) \leq C\eta(q_n + p)^{1/2}$. That yields $\psi_n(\eta) = C\eta(q_n + p)^{1/2} + c(q_n + p)n^{-1/2}$. The condition that $\psi_n(\eta)/\eta$ is a decreasing function of η is satisfied. If we choose r_n to be $(q_n + p)^{-1/2}n^{1/2}$ then $r_n^2\psi_n(1/r_n) = Cr_n(q_n + p)^{1/2} + cr_n^2(q_n + p)n^{-1/2} \leq Cn^{1/2}$.

The condition $M_n(\theta_n) \geq M_n(\theta_{0,n})$ is clearly satisfied, as θ_n is maximizes the likelihood. Finally, $d(\theta_n, \theta_{0,n}) < d(\theta_n, \theta_0) + d(\theta_0, \theta_{0,n})$, and both of the two terms in the right hand side tends to 0 as $n \rightarrow \infty$. By Theorem 3.4.1 of van der Vaart and Wellner (1996) we get that $r_nd(\theta_n, \theta_{0,n}) = O_P(1)$. Together with $d(\theta_0, \theta_{0,n}) = O(n^{-r\nu})$, we obtain that $d(\theta_n, \theta_0) = O_P((q_n + p)^{1/2}n^{-1/2} + n^{-r\nu}) = O_P(n^{-(1-\nu)/2} + n^{-r\nu})$.

Chapter 3

Double Semiparametric Mixture Cure Model with Interval Censored Data

3.1 Double semiparametric model with interval censored data

We assume that the non-negative survival time T_i of subject i cannot be observed directly, it is subject to case 2 interval censoring. As introduced in Section 1.1, the status of the subject is known at two random censoring times U_i and V_i , such that $U_i < V_i$. If $T_i \leq U_i$ then the subject is left censored, $L_i = 0$ and $R_i = U_i$. If $U_i < T_i \leq V_i$ then the subject is interval censored $L_i = U_i$, $R_i = V_i$. Finally, if $V_i < T_i$, then the subject is right censored, $L_i = V_i$ and $R_i = \infty$. We define δ_{1i} , δ_{2i} and δ_{3i} as left, interval and right censoring indicators as in Section 1.3. Let $\mathbf{X}_i$, $\mathbf{Z}_i$ be covariate vectors of dimension p_1 , p_2 associated with the incidence and latency. Besides, W_i is a one-dimensional covariate. We remark that $\mathbf{Z}_i$ and $\mathbf{X}_i$ may or may not be the same. The difference between the covariates is discussed later in this

section. The observed data is the following:

$$\mathcal{O} = \{L_i, R_i, \delta_{1i}, \delta_{2i}, \delta_{3i}, \mathbf{X}_i, W_i, \mathbf{Z}_i, \quad i = 1, \dots, n\}$$

We assume that the subjects come from two disjoint subgroups. Those who are in the susceptible subset experience the failure event eventually, even if it is not observed. Subjects in the cured subset never experience the event hence their survival time is always right censored. We create the incidence model by defining the latent random variable Y_i as the indicator of subject i is susceptible, and assume that $P(Y_i = 1) = \pi_i = \pi(\mathbf{X}_i, W_i)$, so the probability π_i only depends on the covariates $\mathbf{X}_i$ and W_i . $Y_i = 1$ is known when the subject is left censored or interval censored.

In the presence of a cured subgroup we assume the mixture cure model for the survival function of the overall population as follows

$$S_{pop}(t|\mathbf{Z}_i, \mathbf{X}_i, W_i) = \pi(\mathbf{X}_i, W_i)S(t|\mathbf{Z}_i) + 1 - \pi(\mathbf{X}_i, W_i), \quad (3.1)$$

where $\mathbf{Z}_i$ and $\mathbf{X}_i$ are vectors of covariates that have linear effects on the latency and the incidence respectively, W_i is the main exposure variable of interest, whose effects on the incidence might be non-linear and $S(t|\mathbf{Z}_i)$ is the survival function of the variable T_i of a susceptible subject given $\mathbf{Z}_i$, i.e., $S(t|\mathbf{Z}_i) = P(T_i > t|\mathbf{Z}_i, Y_i = 1)$. We remark that, as mentioned above, the covariates are separated into two subsets, and usually previous studies or clinical doctors' suggestion indicate which covariates should be included in which set. The covariates $\mathbf{X}_i$ and W_i are also separated in the proposed model because W_i , which is a covariate of continuous support may not have a linear effect on the logit transformed probability π_i . Therefore we assume that $\pi(\mathbf{X}_i, W_i)$ follows a semiparametric model, specifically, its logit follows

a partially linear form,

$$\pi(\mathbf{X}_i, W_i) = \pi(\mathbf{X}_i, W_i | \boldsymbol{\alpha}, f) = \frac{\exp(\boldsymbol{\alpha}^T \mathbf{X}_i + f(W_i))}{1 + \exp(\boldsymbol{\alpha}^T \mathbf{X}_i + f(W_i))}, \quad (3.2)$$

where $\boldsymbol{\alpha} \in A$ is a coefficient vector and f is an unknown univariate function. Note that in this case we omit the usual intercept term due to identifiability issues. We further assume that the survival function of the susceptible subjects follows the proportional hazards model, thus the negative logarithm of the survival function S is defined as

$$\Lambda(t | \mathbf{Z}_i) = \Lambda_0(t) \exp(\boldsymbol{\beta}^T \mathbf{Z}_i), \quad (3.3)$$

where $\Lambda_0(t)$ is the unknown cumulative baseline hazard function and $\boldsymbol{\beta} \in B$ is an unknown parameter of dimension p_2 . The observed log-likelihood function of the n -element sample is

$$\begin{aligned} l_{\text{obs}}(\boldsymbol{\alpha}, \boldsymbol{\beta}, \Lambda_0, f | \mathcal{O}) = & \sum_{i=1}^n \delta_{1i} \left[\log [\pi(\mathbf{X}_i, W_i | \boldsymbol{\alpha}, f)] \right. \\ & + \log [1 - \exp(-\Lambda_0(R_i) \exp(\boldsymbol{\beta}^T \mathbf{Z}_i))] \Big] \\ & + \delta_{2i} \left[\log [\pi(\mathbf{X}_i, W_i | \boldsymbol{\alpha}, f)] + \log [\exp(-\Lambda_0(L_i) \exp(\boldsymbol{\beta}^T \mathbf{Z}_i)) \right. \\ & \quad \left. - \exp(-\Lambda_0(R_i) \exp(\boldsymbol{\beta}^T \mathbf{Z}_i))] \Big] \\ & + \delta_{3i} \left[\log [\pi(\mathbf{X}_i, W_i | \boldsymbol{\alpha}, f)] \right. \\ & \quad \left. \exp(-\Lambda_0(L_i) \exp(\boldsymbol{\beta}^T \mathbf{Z}_i)) + 1 - \pi(\mathbf{X}_i, W_i | \boldsymbol{\alpha}, f) \right]. \end{aligned} \quad (3.4)$$

We establish that the model defined in (3.1) is identifiable.

Theorem 3.1. Assume that the conditions (C1) and (C4) in Section 3.3 hold. Then the model (3.1) is identifiable.

The proof of the theorem can be found in Section 3.6.

By the method of maximum likelihood, estimation of $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, f and Λ_0 can be achieved by maximizing (3.4) over a restricted space, in which $\boldsymbol{\alpha} \in A$,

$\beta \in B$, with A and B being in a compact subspace of $\mathbb{R}^{p_1}$ and $\mathbb{R}^{p_2}$ respectively, $\Lambda_0(t)$ is a nonnegative and nondecreasing function of t and $f(w)$ is a continuous function of w . With two unknown functions involved, finding the semiparametric maximum likelihood estimator is a challenging task. The true semiparametric maximum likelihood estimator, however, contains the value of Λ_0 at the endpoints in each distinct interval, and the value of f at each distinct W_i , so the dimension of the estimator may linearly increase with the sample size n . Specifically, when there are no ties among the censoring variables and the covariates W_i , then the dimension of the estimator can go up to $3n$. Therefore, in the next section we propose to approximate Λ_0 and f using B-splines that reduces computational complexity, and provide a sieve maximum likelihood estimation procedure.

3.2 Estimation procedure

3.2.1 Semiparametric Sieve Maximum Likelihood Estimator

Our proposed estimator is obtained using a spline-based semiparametric sieve maximum likelihood approach. Specifically, let $(\hat{\alpha}_n, \hat{\beta}_n, \hat{\Lambda}_n, \hat{f}_n)$ maximize the log-likelihood $l_{\text{obs}}(\alpha, \beta, \Lambda_0, f | \mathcal{O})$, where the finite parameters $\alpha \in A$ and $\beta \in B$, as in Section 3.1 and f and Λ_0 are unknown and approximated by B-splines.

To estimate f , we construct the spline space as in shown in Section 1.3.2. We consider the interval $[\min_i(W_i), \max_i(W_i)]$. Denote the order of the spline space by m . We define the partition Δ_n^f as the $q_n + m$ quantiles of $W_1, W_2, \dots, W_n$ of the sample of size n , where $q_n = \lfloor n^\nu \rfloor$ and $0 < \nu < 1$ is

fixed real number. Define the spline space as

$$\mathcal{S}_n = \mathcal{S}_n(\mathcal{P}_m, \Delta_n^f) = \left(f(w) : f(w) = \sum_{j=1}^{q_n} b_j s_j(w), b_j \in \mathbb{R} \right),$$

where s_j , $j = 1, \dots, q_n$ are B-spline basis functions of order d . The model for $f(w)$ is $\sum_{j=1}^{q_n} b_j s_j(w)$, where s_j 's are the B-spline base functions defined above and b_j are unknown real parameters. We denote the vector of b_j 's by $\mathbf{b}$.

Let τ_i , $i = 1, \dots, N$, be the pooled set of the non-zero L_i 's and the finite R_i . We define the spline space over which we aim to estimate Λ_0 on the interval $[\min_i \tau_i, \max_i \tau_i]$ and we put the subinterval endpoints as the quantiles of τ_i , $i = 1, \dots, N$. As the cumulative baseline hazard function is a nonnegative non-decreasing function, we need to define the following space of monotone splines as in (1.7).

$$\mathcal{M}_n = \mathcal{M}(\mathcal{P}_m, \Delta_n^\tau) = \left(\sum_{j=1}^{q_n} \tilde{c}_j r_j(w) : 0 < \tilde{c}_1 < \tilde{c}_2 < \dots < \tilde{c}_{q_n} \right).$$

In practice, one can get the nonnegative increasing sequence $\{\tilde{c}_j\}_{j=1}^{q_n}$, by performing a one-to-one transformation from an arbitrary real sequence $\{c_j\}_{j=1}^{q_n}$, so that we get the necessary properties. For example, $\tilde{\mathbf{c}}$ can be $\mathbf{R}(\exp(c_1), \exp(c_2), \dots, \exp(c_{q_n}))^T$, where $\mathbf{R}$ is the lower triangle matrix whose all nonzero entries are one. We denote the vectors of c_j 's and $\tilde{c}_j$'s by $\mathbf{c}$ and $\tilde{\mathbf{c}}$.

We introduce the following notations to describe the incidence and the survival functions in terms of the B-spline models.

$$\pi(\mathbf{X}, W | \boldsymbol{\alpha}, \mathbf{b}) = \left(1 + \exp \left(- \sum_{j=1}^{q_n} b_j s_j(W) - \boldsymbol{\alpha}^T \mathbf{X} \right) \right)^{-1}$$

and

$$S(t|Z, \boldsymbol{\beta}, \mathbf{c}) = \exp \left(- \sum_{j=1}^{q_n} \tilde{c}_j r_j(t) \exp(\boldsymbol{\beta}^T \mathbf{Z}) \right).$$

The model log-likelihood is written as

$$\begin{aligned} l(\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{b}, \mathbf{c}|\mathcal{O}) = & \sum_{i=1}^n \delta_{1i} [\log \pi(\mathbf{X}_i, W_i, |\boldsymbol{\alpha}, \mathbf{b}) + \log[1 - S(R_i|\mathbf{Z}_i, \boldsymbol{\beta}, \mathbf{c})]] \\ & + \delta_{i2} [\log \pi(\mathbf{X}_i, W_i, |\boldsymbol{\alpha}, \mathbf{b}) \\ & + \log(S(L_i|\mathbf{Z}_i, \boldsymbol{\beta}, \mathbf{c}) - S(R_i|\mathbf{Z}_i, \boldsymbol{\beta}, \mathbf{c}))] \\ & + (1 - \delta_{1i} - \delta_{2i}) \log [\pi(\mathbf{X}_i, W_i|\boldsymbol{\alpha}, \mathbf{b}) S(L_i|\mathbf{Z}_i, \boldsymbol{\beta}, \mathbf{c}) \\ & + 1 - \pi(\mathbf{X}_i, W_i|\boldsymbol{\alpha}, \mathbf{b})]. \end{aligned} \tag{3.5}$$

We denote the sieve space $A \times B \times \mathcal{S}_n \times \mathcal{M}_n$ by Θ_n . To maximize (3.5) with respect to the parameters $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, $\mathbf{b}$ and $\mathbf{c}$, is equivalent to find $\hat{\theta}_n = (\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}, \hat{f}_n, \hat{\Lambda}_n) \in \Theta_n$ that maximizes $l_{obs}(\boldsymbol{\alpha}, \boldsymbol{\beta}, f, \Lambda_0|\mathcal{O})$ over $(\boldsymbol{\alpha}, \boldsymbol{\beta}, f, \Lambda_0) \in \Theta_n$. After we find maximum values $\hat{\mathbf{b}}$ and $\hat{\mathbf{c}}$, we obtain $\hat{f}_n(w) = \sum_{i=1}^{q_n} \hat{b}_j s_j(w)$ and $\hat{\Lambda}_n(t) = \sum_{j=1}^{q_n} \hat{c}_j r_j(t)$, where $\hat{b}_j$ is the j th component of $\hat{\mathbf{b}}$ and $\hat{c}_j$ is the j th component of the transformed $\hat{\mathbf{c}}$.

3.2.2 Estimation algorithm

We employ an expectation–maximization (EM) algorithm to incorporate the unknown Y_i 's. Let us denote the parameters $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{b}, \mathbf{c})$ by θ and let $\mathbf{Y}$ be the vector of Y_i 's. The former notation is only valid in this subsection to avoid confusion with the notation introduced in Subsection 3.2.1. The

complete log-likelihood l^* , which involves the latent $\mathbf{Y}$, is defined as

$$\begin{aligned} l^*(\theta|\mathcal{O}, \mathbf{Y}) = & \sum_{i=1}^n Y_i \log \pi(\mathbf{X}_i, W_i|\boldsymbol{\alpha}, \mathbf{b}) + (1 - Y_i) \log(1 - \pi(\mathbf{X}_i, W_i|\boldsymbol{\alpha}, \mathbf{b})) \\ & + Y_i \left\{ \delta_{1i} \log[1 - S(R_i|\mathbf{Z}_i, \boldsymbol{\beta}, \mathbf{c})] + \delta_{i2} \log[S(L_i|\mathbf{Z}_i, \boldsymbol{\beta}, \mathbf{c}) \right. \\ & \left. - S(R_i|\mathbf{Z}_i, \boldsymbol{\beta}, \mathbf{c})] + \delta_{3i} \log[S(L_i|\mathbf{Z}_i, \boldsymbol{\beta}, \mathbf{c})] \right\}. \end{aligned} \quad (3.6)$$

We introduce the conditional expectation $Q(\theta|\theta^{(m)}) = E(l^*(\theta|\mathcal{O}, \mathbf{Y})|\theta^{(m)})$, which has to be calculated at each iteration of the algorithm, where $\theta^{(m)}$ is the estimator after m iterations. The steps of the EM algorithm are defined as the following.

Initialization Initialize the parameters, that is set $\theta^{(0)}$ as some initial value. Details about the setting of initial value are provided later in Subsection 3.2.3.

E-step Given the the estimator $\theta^{(m)}$ from the m -th iteration, calculate the function $Q(\theta|\theta^{(m)})$. Since the likelihood $l^*(\theta|\mathcal{O}, \mathbf{Y})$ is a linear function of Y_i 's, we have the conditional expectation

$$\begin{aligned} E(Y_i|\mathcal{O}, \theta^{(m)}) = & \delta_i \\ & + \frac{(1 - \delta_i)\pi(\mathbf{X}_i, W_i|\boldsymbol{\alpha}^{(m)}, \mathbf{b}^{(m)})S(L_i|\mathbf{Z}_i, \boldsymbol{\beta}^{(m)}, \mathbf{c}^{(m)})}{\pi(\mathbf{X}_i, W_i|\boldsymbol{\alpha}^{(m)}, \mathbf{b}^{(m)})S(L_i|\mathbf{Z}_i, \boldsymbol{\beta}^{(m)}, \mathbf{c}^{(m)}) + 1 - \pi(\mathbf{X}_i, W_i|\boldsymbol{\alpha}^{(m)}, \mathbf{b}^{(m)})}, \end{aligned}$$

Then substitute $\omega_i = E(Y_i|\mathcal{O}, \theta^{(m)})$ in $Q(\theta|\theta^{(m)})$ for Y_i .

M-step Update θ by $\theta^{(m+1)}$ which is the maxima of $Q(\theta|\theta^{(m)})$ with respect to θ . From the likelihood (3.6), to maximize Q with respect to $(\boldsymbol{\alpha}, \mathbf{b})$ and $(\boldsymbol{\beta}, \mathbf{c})$ is equivalent to maximize $Q_1^{(m)}$ and $Q_2^{(m)}$ respectively, where

$$Q_1^{(m)}(\theta) = \sum_{i=1}^n \omega_i \log \pi(\mathbf{X}_i, W_i|\boldsymbol{\alpha}^{(m)}, \mathbf{b}^{(m)})$$

$$+ (1 - \omega_i) \log (1 - \pi(\mathbf{X}_i, W_i | \boldsymbol{\alpha}^{(m)}, \mathbf{b}^{(m)}))$$

and

$$\begin{aligned} Q_2^{(m)}(\theta) = & \sum_{i=1}^n \omega_i \left\{ \delta_{1i} \log(1 - S(R_i | \mathbf{Z}_i, \boldsymbol{\beta}^{(m)}, \mathbf{c}^{(m)})) \right. \\ & + \delta_{2i} \log [S(L_i | \mathbf{Z}_i, \boldsymbol{\beta}^{(m)}, \mathbf{c}^{(m)}) - S(R_i | \mathbf{Z}_i, \boldsymbol{\beta}^{(m)}, \mathbf{c}^{(m)})] \\ & \left. + \delta_{3i} \log S(L_i | \mathbf{Z}_i, \boldsymbol{\beta}, \mathbf{c}) \right\} \end{aligned}$$

The algorithm repeats E-step and M-step until convergence is achieved, that is, $\|\theta^{(m+1)} - \theta^{(m)}\|_2 < \epsilon$ for a small positive ϵ . The optimization can be carried out by a Newton–Raphson iteration algorithm. The constrained optimization with respect to $\mathbf{c}$ can be performed as in Section 2.2. We denote the estimators obtained at the end of the iteration by $(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}, \hat{\mathbf{b}}, \hat{\mathbf{c}})$, and the estimator for the functional parameters by $\hat{f}_n = \sum_{j=1}^{q_n} \hat{b}_j s_j$ and $\hat{\Lambda}_n = \sum_{j=1}^{q_n} \hat{c}_j r_j$, as in Subsection 3.2.1. This makes the final estimator $\hat{\theta}_n = (\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}, \hat{f}_n, \hat{\Lambda}_n)$.

3.2.3 Implementation

The implementation of the proposed estimation procedure requires to fix the number of base splines as well as degree of the polynomials. In our simulations and data analysis we follow the work of Zhang and Davidian (2008), in which the authors propose the Hannan–Quinn (HQ) information criterion defined as $HQ = -2l_{\max} + 2k \log \log n$, where $l_{\max}$ is the value of the observed the log-likelihood function (3.5) evaluated at the estimator $\hat{\theta}_n$ given by the algorithm and k is the sum of the dimensions of all finite parameters $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, $\mathbf{b}$ and $\mathbf{c}$ to be estimated. Based on our experience, in practice, we suggest to use quadratic or cubic spline basis and the number of base splines range from 4 to 8.

Since the algorithm converges relatively fast, one can initialize the EM algorithm by reasonable values, say the finite parameters are 0, $\mathbf{b}$ can be chosen as constant 0, and $\mathbf{c}$ can be chosen as some small negative value, say -0.1 .

3.2.4 Inference

Asymptotic normality of the finite parameters is asserted in Section 3.3, so we provide a way to approximate the information matrix provided by the semiparametric theory. The calculation of the information for the sieve semiparametric model uses the notation and the methodology of Huang et al. (2008). In order to separate the finite and the infinite dimensional parameters, we denote $\boldsymbol{\sigma}_0 = (\boldsymbol{\alpha}_0, \boldsymbol{\beta}_0)$, $\phi_0 = (f_0, \Lambda_0)$ the true parameters and $\hat{\boldsymbol{\sigma}} = (\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}})$, $\hat{\phi} = (\hat{f}_n, \hat{\Lambda}_n)$ the estimators. The information matrix evaluated at a parameter $\boldsymbol{\sigma}$ is denoted as $I(\boldsymbol{\sigma})$. We define $\dot{l}_1(\boldsymbol{\sigma}, \phi, O)$ as the vector containing partial derivatives of $l(\boldsymbol{\theta}|O)$ with respect to $\boldsymbol{\sigma}$, where

$$\begin{aligned} l(\boldsymbol{\theta}|O) = & \delta_1 [\log \pi(\mathbf{X}, W|\boldsymbol{\alpha}, f) + \log(1 - \exp(-\Lambda_0(R) \exp(\boldsymbol{\beta}^T \mathbf{Z})))] \\ & + \delta_2 [\log \pi(\mathbf{X}, W|\boldsymbol{\alpha}, f) + \log[\exp(-\Lambda_0(L) \exp(\boldsymbol{\beta}^T \mathbf{Z})) \\ & \quad - \exp(-\Lambda_0(R) \exp(\boldsymbol{\beta}^T \mathbf{Z}))]] \\ & + \delta_3 \log[\pi(\mathbf{X}, W|\boldsymbol{\alpha}, f) \exp(-\Lambda_0(L) \exp(\boldsymbol{\beta}^T \mathbf{Z})) + 1 - \pi(\mathbf{X}, W|\boldsymbol{\alpha}, f)], \end{aligned}$$

which is the observed log-likelihood of the one-element sample. For each pair $(\boldsymbol{\sigma}, \phi)$ we consider a smooth parametric submodel $(\boldsymbol{\sigma}, \phi_{(s)})$ with a scalar parameter s that is defined so that $\phi_{(0)} = \phi$. Let $\frac{\partial \phi_s}{\partial s}|_{s=0} = h$. The functions h satisfy this equation form the space $\mathcal{H}$. Since the functional parameters belong to a product space, we also consider $\mathcal{H}$ to be a product space itself.

The score operator for ϕ is

$$\dot{l}_2(\boldsymbol{\sigma}, \phi, O)(h) = \left. \frac{\partial l((\boldsymbol{\sigma}, \phi_s), O)}{\partial s} \right|_{s=0}, \quad (3.7)$$

which is a linear operator mapping $\mathcal{H}$ to $L_2(P)$. A separate score operator is defined for each component of σ , which yields the vector $\dot{l}_2(\boldsymbol{\sigma}, \phi, O)(\mathbf{h}) = (\dot{l}_2(\boldsymbol{\sigma}, \phi, O)(h_1), \dots, \dot{l}_2(\boldsymbol{\sigma}, \phi, O)(h_{p_1+p_2}))$. The effective information $I(\boldsymbol{\sigma})$ is the minimum of

$$E \left[\left\| \dot{l}_1(\boldsymbol{\sigma}, \phi, O) - \dot{l}_2(\boldsymbol{\sigma}, \phi, O)(\mathbf{h}) \right\|^2 \right],$$

over the values $\mathbf{h} \in \mathcal{H}^{p_1+p_2}$. The minimizer ξ_0 is called the least favourable direction and the effective score function is $l^*(\boldsymbol{\sigma}, \phi) = \dot{l}_1(\boldsymbol{\sigma}, \phi, O) - \dot{l}_2(\boldsymbol{\sigma}, \phi, O)(\xi_0)$. Calculating the effective information is complicated. Therefore, we use the least square approximation method proposed by Huang et al. (2008). Since we approximate the derivative space of f and Λ_0 respectively, we use the same spline base polynomials as a basis of approximation. For each k , the approximation of the suitable h_k , $k = 1, \dots, p_1 + p_2$ can also be considered from the space $\mathfrak{F}_n \times \mathfrak{B}_n$. To approximate the information for $\hat{\boldsymbol{\sigma}}$, we introduce

$$\begin{aligned} \dot{l}_2(\hat{\boldsymbol{\sigma}}, \hat{\phi}, O)(\mathfrak{F}_n \times \mathfrak{B}_n) = & (\dot{l}_2(\hat{\boldsymbol{\sigma}}, \hat{\phi}, O)(s_1), \dot{l}_2(\hat{\boldsymbol{\sigma}}, \hat{\phi}, O)(s_2), \dots, \dot{l}_2(\hat{\boldsymbol{\sigma}}, \hat{\phi}, O)(s_{q_n}), \\ & \dot{l}_2(\hat{\boldsymbol{\sigma}}, \hat{\phi}, O)(r_1), \dots, \dot{l}_2(\hat{\boldsymbol{\sigma}}, \hat{\phi}, O)(r_{q_n})), \end{aligned}$$

which contains components corresponding to all the base splines. We define the sample version of the score components as in Huang et al. (2008).

$$\begin{aligned} A_{11} &= P_n(\dot{l}_1(\hat{\boldsymbol{\sigma}}, \hat{\phi}, O))^{\otimes 2}, & A_{12} &= P_n \dot{l}_1(\hat{\boldsymbol{\sigma}}, \hat{\phi}, D) \dot{l}_2(\hat{\boldsymbol{\sigma}}, \hat{\phi}, O)(\mathfrak{F}_n \times \mathfrak{B}_n)^T \\ A_{21} &= A_{12}^T, & A_{22} &= P_n(\dot{l}_2(\hat{\boldsymbol{\sigma}}, \hat{\phi}, O))(\mathfrak{F}_n \times \mathfrak{B}_n)^{\otimes 2} \end{aligned}$$

where $P_n f(O) = \sum_{i=1}^n f(O_i)$. The approximation of $I(\boldsymbol{\sigma}_0)$ is then given by

$$\hat{I}_n = A_{11} - A_{12}A_{22}^-A_{21},$$

with A_{22}^- being the generalized inverse of A_{22} .

3.3 Asymptotic Theory

In this section we establish the large sample properties of the estimator $\hat{\theta}_n$. In order to assert conditions (A1) and (A2), we express the likelihood in terms of U_i and V_i instead of L_i and R_i in this section and in the proofs. Denote the Euclidean distance between the vectors $(\boldsymbol{\alpha}_1^T, \boldsymbol{\beta}_1^T)^T$ and $(\boldsymbol{\alpha}_2^T, \boldsymbol{\beta}_2^T)^T$ by $\|(\boldsymbol{\alpha}_1^T, \boldsymbol{\beta}_1^T)^T - (\boldsymbol{\alpha}_2^T, \boldsymbol{\beta}_2^T)^T\|_2$. Furthermore, we define the distance between two sets of spline estimates as the L_2 distance between the functions. The distance between θ_1 and θ_2 is

$$d(\theta_1, \theta_2) = [\|(\boldsymbol{\alpha}_1^T, \boldsymbol{\beta}_1^T)^T - (\boldsymbol{\alpha}_2^T, \boldsymbol{\beta}_2^T)^T\|_2^2 + \|f_1 - f_2\|_2^2 + \|\Lambda_1 - \Lambda_2\|_2^2]^{1/2}.$$

We need the following conditions to show main theoretical results.

- (C1) There exists a bound M_Z such that $|\mathbf{Z}| \leq M_Z$ almost surely, and $E(\mathbf{Z}\mathbf{Z}^T)$ is nonsingular. Similarly, there exists a bound M_X such that $|\mathbf{X}| < M_X$ and $E(\mathbf{X}\mathbf{X}^T)$ is nonsingular.
- (C2) The parameter spaces A and B are compact, in particular there is an upper bound M such that $\|\boldsymbol{\alpha}\| + \|\boldsymbol{\beta}\| < M$
- (C3) There exists a finite interval $[a, b] \subset R^+$ such that the supports of U and V are subsets $[a, b]$, and there exists $\zeta > 0$ such that $P(V - U > \zeta) = 1$. There exists a bounded interval $[a_1, b_1]$ that contains all the possible values of W .

(C4) The true Λ_0 and f_0 are bounded and continuously differentiable up to order r . In addition, the first derivative of Λ_0 is strictly positive.

The conditions above are usual in case of semiparametric estimators for interval censored data. Condition (C1) is needed to ensure that the covariate is non-degenerate, (C1) – (C3) ensure that the likelihood has finite derivatives. The first part of (C4) is a smoothness criterion that affects the rate of convergence, and the second part ensures that Λ_0 is a proper cumulative baseline hazard function.

Theorem 3.2. Given the assumptions (C1) – (C4), the estimator $\hat{\theta}_n$ converges almost surely to the true value θ_0 , that is $d(\hat{\theta}_n, \theta_0) \rightarrow 0$ almost surely.

Theorem 3.3. Assume that conditions in Theorem 3.2 hold. Then $d(\hat{\theta}_n, \theta_0) = O(n^{-\min(\nu r, (1-\nu)/2)})$.

Theorem 3.4. Given assumptions (C1) – (C4), the estimator $(\hat{\alpha}, \hat{\beta})$ asymptotically follows normal distribution, that is

$$\sqrt{n}[(\hat{\alpha}, \hat{\beta}) - (\alpha_0, \beta_0)] \rightarrow N(\mathbf{0}, I^{-1}(\sigma_0))$$

in distribution, where $I(\sigma_0)$ is defined in Section 3.2.4.

The proofs of the theorems can be found in Section 3.6. The convergence rate in Theorem 3.3 is the same as for example in Zhang et al. (2010), that is in case $\nu = 1/(1 + 2r)$ the convergence rate is $O(r/(1 + 2r))$, which is the optimal convergence rate for non-parametric regression. The idea used to prove Theorem 3.3 is similar to that of Zhou et al. (2017) but we explain them in a more elaborate way and rely on one additional lemma and give a simpler proof of Theorem 1. The additional lemma aims to remove conditions A3 and A5 of Zhou et al. (2017), that are difficult to check in

practice. As such, our asymptotic theory requires quite mild conditions only.

3.4 Simulations

We carry out simulation studies to evaluate the proposed model under two different scenarios. Under the first scenario we apply our method to data that are generated in a way that the incidence has a covariate with a non-linear effect, while under the second scenario, we check the efficiency of the proposed model in case the covariate effect on the incidence is linear, and show that our method does not lose much efficiency.

In the first scenario, the one dimensional covariates Z_i are generated from the standard normal distribution, X_i are generated from the uniform distribution on $[0, 1]$ and W_i are generated from the uniform distribution on $[0, 2]$. The incidence model followed the model in (3.2), with true parameter $\alpha = -1.8, -3.5$ and -8.4 to achieve approximate right censoring rates of 15%, 35% and 50%, and $f(w) = \sin(w * 2) + 1/2$ to incorporate a nonlinear effect that has an increasing part and a decreasing part. The indicator Y_i for susceptibility are generated from the Bernoulli distribution with $P(Y_i = 1) = \pi_i(X_i, W_i | \alpha, f)$. For subjects with $Y_i = 0$ the survival time T_i is set to ∞ . For susceptible subjects, i.e. $Y_i = 1$, finite survival time are generated according to the proportional hazards model with $\Lambda_0(t) = \frac{2}{3}t^{3/2}$, and $\beta = 1$. We chose Λ_0 to be a monomial because generating the survival times can be done by generating a sample from the Weibull distribution.

The following censoring procedure is applied to the simulated survival times. We generate two auxiliary random variables, U_{1i} from the uniform distribution on $[0.5, 1]$ and U_{2i} from the uniform distribution on $[1, 6]$. If

$T_i > U_{1i} + U_{2i}$ then the subject is right censored with censoring time $U_{1i} + U_{2i}$. Otherwise we set the two interval endpoints to be kU_{1i} and $(k+1)U_{1i}$ so that $kU_{1i} \leq T_i < (k+1)U_{1i}$. With the method above we generated 500 samples of size $n = 500, 1000$ and 2000 for each of the three censoring rates. We applied the method given in section 3.2 to find the regression parameters as well as the unknown functions. We employed quadratic splines to approximate both Λ_0 and f with 5 base polynomials. We compare our results with a similar model, that does not incorporate nonlinear random effect, by Zhou et al. (2017). Their method contains the proportional hazards mixture cure model as a special case with linear effect on the incidence. Although we do not expect their method to be correct, we would like to show that our estimator provides better results on the regression parameters compared to their method.

500 simulation runs were conducted, with sample sizes of $n = 500, 1000, 2000$. The numerical results are summarized in Table 3.1, where the DSP denotes the proposed method and the SSP denotes the method by Zhou et al. (2017). We provide the is the mean of the bias over all repetitions (Mean Bias), the mean of estimated standard error (SE), the standard deviation of the estimated coefficients (SD) and the 95% coverage probability (CP) defined as the proportion of the 0.95 confidence intervals containing the true parameter values. The table shows that the bias, the standard error and the standard deviation decreases for the both methods with increased sample size. The SSP method provides relatively robust estimates for β , with more unbiased as the sample size grows, but it shows mixed results for α , the bias does not seem to decrease, which is expected. For $n = 500$ the bias is less than in case of proposed method, but it does not decrease significantly as n increases. In Figure 3.1 the mean and quantiles of

the estimated cumulative baseline hazard functions show that the proposed estimator work well for larger sample sizes. When cure rate increases, there is an increased variation among the estimators, which is expected, as there are less data available for the survival part. Figure 3.2 shows the mean, the 2.5% and the 97.5% quantiles of the estimated f curves. One can observe that in case of larger sample sizes, the estimator works quite well. However, in case of $n = 500$ and relatively high cure rate, the estimator yields high variance.

Under the second scenario, we generated the data in a similar way to scenario 1, but using a different $f(w) = 0.91w + 1/2$, making the effect linear. Therefore the generated data would fit the model by Zhou et al. (2017). Hence we can compare the performance of our estimator, when the true model is less complex. The numerical results for the second scenario are summarized in Table 3.2. One can see that the simpler model is slightly less biased in most cases but the difference is not that much. The standard deviation given by the two methods are very close to each other. The SD in the proposed method is slightly higher for α and slightly lower for β in most cases. So even if f is linear in the true model, we do not lose much accuracy of the parameter estimations. And also, one can observe that the bias, the standard deviation and the estimated standard errors decrease as sample sizes increase. As one can see in Figure 3.3, the estimates for f are approximately linear. Increased cure rate has an adverse effect on the bias.

In general, we suggest to use the proposed method in case where the sample size is large enough, and there is empirical evidence that the effect of a covariate on the logit of the probability of susceptibility is not linear.

Figure 3.1: *Estimated cumulative baseline hazard function. The solid line is the true, the dashed line is the mean, and the dotted lines are the 2.5% and 97.5% quantiles of the estimated curves. The subfigures from left to right shows the results for 15%, 35% and 50% of cure rate respectively.*

Figure 3.2: *Estimated functional part of the cure probability, scenario 1. The solid line is the true $f(w) = \sin(w * 2) + 1/2$, the dashed line is the mean, the dotted lines are the 2.5% and 97.5% quantiles of the estimated curves. The subfigures from left to right shows the results for 15%, 35% and 50% of cure rate respectively.*

Figure 3.3: *Estimated functional part of the cure probability. The solid line is the true $f(w) = 0.91w + 1/2$, the dashed line is the mean, the dotted lines are the 2.5% and 97.5% quantiles of the estimated curves. The subfigures from left to right shows the results for 15%, 35% and 50% of cure rate respectively.*

3.5 Application to the hypobaric decompression sickness data

We consider a data set obtained from the Hypobaric Decompression Sickness Data Bank (HDSB) by National Aeronautics and Space Administration, Conkin et al. (1992). The dataset contains 549 records of 238 subjects. The event of interest is the development of grade IV venous gas emboli (VGE), which can be a cause of serious decompression sickness. The time to event, which is measured in hours, cannot be observed directly, it is known to be between two examination times. The covariates of the recorded data are the following:

Age – the age of the subject in years.

Sex – the gender of the subject (male: Sex=1, 177 subjects, female: Sex=0, 61 subjects).

TR360 – this covariate measures the decompression stress.

Noadyn – the experimental indicator variable, the subject was either ambulatory (Noadyn = 1) or lower body adynamic (Noadyn = 0).

Since the data contains repeated measurements, we extracted the first observation of each subject with respect to their ID. The chosen sample consists of 238 observations, out of which 70% are right censored data. We set the covariates in the following way. Let $\mathbf{X}$ = Sex, W = Age, and $\mathbf{Z}$ = (Noadyn, TR360, Sex). We note that we leave out other measurements from the covariate associated with the incidence because only individual characteristics can influence one's susceptibility to grade IV VGE (Thompson and Chhikara, 2003).

We fit the double semiparametric model using the method developed in Section 3.2 with 5, 6 and 7 base splines, with degrees 2 and 3. We selected a cubic spline with 5 basis functions both for both f and for Λ_0 that fits the best according to the HQ criterion discussed in Section 3.2.3. The estimated parameters are summarized in Table 3.3. As one can see, TR360 level and NOADYN has an adverse effect on the survival, which is expected. However, Sex does not seem to have a significant effect on the survival, just like in case of previous analysis of the data by Ma (2010). The incidence is affected by Sex, namely, with a positive coefficient, males are more likely to experience grade IV VGE which coincides with previous studies. Analyses involving linear models have concluded that older subjects are more likely to experience the failure event. However, according to Figure 3.4, the effect of age increases until around the age of 40, which means higher chance of experiencing grade IV VGE. While for subjects older than 40 become less susceptible as they become older. Since the number of subjects whose age is above 40 is 35, so we cannot analyze separately that subset. However, it is still interesting to remark that age may not affect the probability of susceptibility linearly. Further data may be required to establish a more accurate estimate for individuals who are older than 40 years.

3.6 Technical proofs

3.6.1 Proof of identifiability

Proof. Let $(\alpha, \beta, \Lambda_0, f)$ and $(\alpha^*, \beta^*, \Lambda_0^*, f^*)$ be two sets of estimators. We need to prove for the likelihood of one element sample O that if

$$l_{\text{obs}}(\alpha, \beta, \Lambda_0, f|O) = l_{\text{obs}}(\alpha^*, \beta^*, \Lambda_0^*, f^*|O) \quad (3.8)$$

Figure 3.4: *Estimate of f , the effect of Age on the incidence for the HDSD data set.*

for almost everywhere in the possible O values, then $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \Lambda_0, f)$
 $= (\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*, \Lambda_0^*, f^*)$. First, we show that (3.8) implies

$$\begin{aligned} \pi(\mathbf{X}, W | \boldsymbol{\alpha}, f) \exp(-\Lambda_0(t) \exp(\boldsymbol{\beta}^T \mathbf{Z})) + 1 - \pi(\mathbf{X}, W | \boldsymbol{\alpha}, f) \\ = \pi(\mathbf{X}, W | \boldsymbol{\alpha}^*, f^*) \exp(-\Lambda_0^*(t) \exp(\boldsymbol{\beta}^{*T} \mathbf{Z})) + 1 - \pi(\mathbf{X}, W | \boldsymbol{\alpha}^*, f^*). \end{aligned} \quad (3.9)$$

Note that independent censoring implies the constant-sum property of the model by Proposition 1 of Oller et al. (2004), where the constant-sum property is defined in that paper. By similar arguments to Theorem 1 of Oller et al. (2007) we get (3.9). Also note that the left hand side and the right hand side of (3.9) are monotone decreasing in $\exp(\boldsymbol{\beta}^T \mathbf{Z})$ and $\exp(\boldsymbol{\beta}^{*T} \mathbf{Z})$ respectively. Since the exponential function is monotone increasing, $\boldsymbol{\beta}^T \mathbf{Z}$ and $\boldsymbol{\beta}^{*T} \mathbf{Z}$ must be monotone in each component of $\mathbf{Z}$, with the same sign. This implies that $\boldsymbol{\beta}$ and $\boldsymbol{\beta}^*$ must have the same sign in each component.

We subtract 1 from both sides of (3.9), take the negative and let $c = \pi(\mathbf{X}, W | \boldsymbol{\alpha}, f) / \pi(\mathbf{X}, W | \boldsymbol{\alpha}^*, f^*)$, a non-negative constant that does not depend on Z , and we get

$$1 - \exp(-\Lambda_0^*(t) \exp(\boldsymbol{\beta}^{*T} \mathbf{Z})) = c [1 - \exp(-\Lambda_0(t) \exp(\boldsymbol{\beta}^T \mathbf{Z}))].$$

After rearranging the previous equation, we get

$$\Lambda_0^*(t) = - \frac{\log(1 - c + c \exp(-\Lambda_0(t) \exp(\boldsymbol{\beta}^T \mathbf{Z})))}{\exp(\boldsymbol{\beta}^{*T} \mathbf{Z})} \quad (3.10)$$

The right hand side of (3.10) is monotone increasing in both $\exp(\boldsymbol{\beta}^T \mathbf{Z})$ and $\exp(\boldsymbol{\beta}^{*T} \mathbf{Z})$. Therefore, using condition (C1), for a non-zero coordinate of $\boldsymbol{\beta}$ and $\boldsymbol{\beta}^*$, if we change the corresponding value of $\mathbf{Z}$, it must change the value of the right hand side of (3.10), while the left hand side remains unchanged. It is easy to see that the only way the right hand side remains

unchanged if $c = 1$ and $\beta = \beta^*$. We then get the equation $\Lambda_0^*(U) = \Lambda_0(U)$, which is true for all U . From $c = 1$, we know that

$$\pi(\mathbf{X}, W | \alpha, f) = \pi(\mathbf{X}, W | \alpha^*, f^*).$$

After some algebra, we get

$$f(W) - f^*(W) = (\alpha - \alpha^*)^T \mathbf{X},$$

which, along with conditions (C1) and (C4) imply $\alpha = \alpha^*$ and $f(W) = f^*(W)$ almost everywhere. $\square$

3.6.2 Proof of large sample properties

We introduce some notations that are used throughout the proofs. Let $\mathfrak{L}_n = \{l(\theta|O) : \theta \in \Theta_n\}$, a class of likelihood functions based on the one-element sample. We denote $\mathbb{P}g(O) = \int f(O)dP(O)$, which is the expectation of g based on the true model, and $\mathbb{P}_n g(\mathcal{O}) = \frac{1}{n} \sum_{i=1}^n g(O_i)$, the sample mean of $g(O)$. We denote a suitable constant by C , which is independent of n and the estimator $\hat{\theta}_n$.

Let $N(\epsilon, \mathfrak{L}_n, L_1(\mathbb{P}_n))$ be the covering number of $\mathfrak{L}_n$ with respect to the $L_1(\mathbb{P}_n)$ norm.

Lemma 3.1 (Covering number). Under conditions (C1) – (C4) the covering number of $\mathfrak{L}_n$ has the following upper bound: $N(\epsilon, \mathfrak{L}_n, L_1(\mathbb{P}_n)) \leq C\epsilon^{-(Cq_n + p_1 + p_2)}$.

Proof. Let ϵ be a small positive number. Since α and β are finite dimensional parameters from compact sets A and B respectively and by (C2), there exists $\lfloor C(1/\epsilon)^{p_1} \rfloor$ balls of radius ϵ that cover A and there exists $\lfloor C(1/\epsilon)^{p_2} \rfloor$ balls of radius ϵ that cover B . Let the centers of the sets of

balls be $\{\alpha_{\tilde{s}}, \tilde{s} = 1, \dots, \lfloor C(1/\epsilon)^{p_1} \rfloor\}$ and $\{\beta_{\tilde{r}}, \tilde{r} = 1, \dots, \lfloor C(1/\epsilon)^{p_2} \rfloor\}$ respectively. Then for every $\alpha \in A$ and $\beta \in B$ there exists such α_r and β_s , $1 \leq r \leq \lfloor C(1/\epsilon)^{p_1} \rfloor$, $1 \leq s \leq \lfloor C(1/\epsilon)^{p_2} \rfloor$ so that $\|\alpha - \alpha_s\| < \epsilon$ and $\|\beta - \beta_r\| < \epsilon$. It follows from (C1) that $|\alpha^T \mathbf{X} - \alpha_s^T \mathbf{X}| < \epsilon M_X$ and $|\beta^T \mathbf{Z} - \beta_r^T \mathbf{Z}| < \epsilon M_Z$ for all $\mathbf{X}$ and $\mathbf{Z}$ respectively. Therefore, $\beta_s^T \mathbf{Z} - M_Z \epsilon \leq \beta^T \mathbf{Z} \leq \beta_s^T \mathbf{Z} + M_Z \epsilon$ and $\alpha_r^T \mathbf{X} - M_X \epsilon \leq \alpha^T \mathbf{X} \leq \alpha_r^T \mathbf{X} + M_X \epsilon$ for all $\mathbf{X}$ and $\mathbf{Z}$.

By Shen and Wong (1994, p. 597) and (C4), there exists a set of brackets $\{[\Lambda_i^l, \Lambda_i^u] : \tilde{i} = 1, \dots, C(1/\epsilon)^{Kq_n}\}$ such that for every $\Lambda \in \mathcal{M}_n$ there is an $i \in \{1, \dots, C(1/\epsilon)^{Kq_n}\}$ such that $\Lambda_i^l(t) \leq \Lambda(t) \leq \Lambda_i^u(t)$ for every t and $\|\Lambda_i^u - \Lambda_i^l\|_1 < \epsilon$. Similarly, there exists a set of brackets $\{[f_j^l, f_j^u] : \tilde{j} = 1, \dots, \lfloor C(1/\epsilon)^{Kq_n} \rfloor\}$ such that for every $f \in \mathcal{S}_n$ there is a $j \in \{1, \dots, \lfloor C(1/\epsilon)^{Kq_n} \rfloor\}$, such that $f_j^l(w) \leq f(w) \leq f_j^u(w)$ for every w and for all $\tilde{j} = 1, \dots, \lfloor C(1/\epsilon)^{Kq_n} \rfloor$, $\|f_j^u - f_j^l\|_1 < \epsilon$.

First, denote

$$\begin{aligned}\pi_{j,s}^l(\mathbf{X}, W) &= \frac{1}{1 + \exp(-f_j^l(W) - \alpha_s^T \mathbf{X} + M_X \epsilon)} \\ \pi_{j,s}^u(\mathbf{X}, W) &= \frac{1}{1 + \exp(-f_j^u(W) - \alpha_s^T \mathbf{X} - M_X \epsilon)}.\end{aligned}$$

Now we show that for every j and s , $\mathbb{P}_n(|\pi_{j,s}^u(\mathbf{X}, W) - \pi_{j,s}^l(\mathbf{X}, W)|) < C\epsilon$ for some positive constant C . By the mean value theorem there exists a $\xi \in (f_j^l(W) - \alpha_s^T \mathbf{X} + M_X \epsilon, f_j^u(W) + \alpha_s^T \mathbf{X} + M_X \epsilon)$, such that

$$\begin{aligned}\pi_{j,s}^u(\mathbf{X}, W) - \pi_{j,s}^l(\mathbf{X}, W) &= \frac{\exp(\xi)}{1 + \exp(\xi)} (f_j^u(W) + \alpha_s^T \mathbf{X} + M_X \epsilon \\ &\quad - f_j^l(W) - \alpha_s^T \mathbf{X} + M_X \epsilon) \\ &\leq f_j^u(W) - f_j^l(W) + 2M_X \epsilon \leq C\epsilon,\end{aligned}$$

which is true for every $\mathbf{X}$ and W , hence true in $L_1(\mathbb{P}_n)$ norm. Similarly, we define the functions

$$S_{i,r}^l(t|\mathbf{Z}) = \exp(-\Lambda_i^u(t) \exp(\beta_r^T \mathbf{Z} + M_Z \epsilon))$$

$$S_{i,r}^u(t|\mathbf{Z}) = \exp(-\Lambda_i^l(t) \exp(\boldsymbol{\beta}_r^T \mathbf{Z} - M_Z \epsilon)).$$

Again, by the mean value theorem we have

$$\begin{aligned} S_{i,r}^u(t|\mathbf{Z}) - S_{i,r}^l(t|\mathbf{Z}) &= \exp(\xi) [\Lambda_i^u(t) \exp(\boldsymbol{\beta}_r^T \mathbf{Z} + M_Z \epsilon) \\ &\quad - \Lambda_i^l(t) \exp(\boldsymbol{\beta}_r^T \mathbf{Z} - M_Z \epsilon)] \end{aligned}$$

for some $\xi \in (-\Lambda_i^u(t) \exp(\boldsymbol{\beta}_r^T \mathbf{Z} + M_Z \epsilon), -\Lambda_i^l(t) \exp(\boldsymbol{\beta}_r^T \mathbf{Z} - M_Z \epsilon))$. Since ξ is negative, $\exp(\xi) < 1$. Furthermore, $\Lambda_i^u(t) - \Lambda_i^l(t) < \epsilon$ and

$$\begin{aligned} \exp(\boldsymbol{\beta}_r^T \mathbf{Z} + M_Z \epsilon) - \exp(\boldsymbol{\beta}_r^T \mathbf{Z} - M_Z \epsilon) &= \exp(\boldsymbol{\beta}_r^T \mathbf{Z}) (\exp(M_Z \epsilon) - \exp(-M_Z \epsilon)) \\ &\leq 2\epsilon \exp(M_Z) \exp(M M_Z). \end{aligned}$$

By the mean value theorem, one can obtain

$$\begin{aligned} &|\Lambda_i^u(t) \exp(\boldsymbol{\beta}_r^T \mathbf{Z} + M_Z \epsilon) - \Lambda_i^l(t) \exp(\boldsymbol{\beta}_r^T \mathbf{Z} - M_Z \epsilon)| \\ &\leq \epsilon \exp(M_Z) \exp(M M_Z) + 2\epsilon M_\Lambda \exp(M_Z) \exp(M M_Z) \\ &\quad + 2\epsilon^2 M_\Lambda \exp(M_Z) \exp(M M_Z) \\ &< C\epsilon, \end{aligned}$$

where the last inequality is true for $\epsilon < 1$ and a constant $C > 0$ that does not depend on ϵ . From all the above,

$$|S_{i,r}^u(t|\mathbf{Z}) - S_{i,r}^l(t|\mathbf{Z})| < C\epsilon.$$

Next we construct the following brackets for $\mathfrak{L}_n$.

$$\begin{aligned} l_{i,j,s,r}^l(O) &= \delta_1 \log(1 - S_{i,r}^u(U|\mathbf{Z})) + \delta_2 \log(S_{i,r}^l(U|\mathbf{Z}) - S_{i,r}^u(V|\mathbf{Z})) \\ &\quad + \delta_3 \log S_{i,r}^l(V|\mathbf{Z}) + (\delta_1 + \delta_2) \log \pi_{i,r}^l(\mathbf{X}, W) + \delta_3 \log(1 - \pi_{i,r}^u(\mathbf{X}, W)) \\ l_{i,j,s,r}^u(O) &= \delta_1 \log(1 - S_{i,r}^l(U|\mathbf{Z})) + \delta_2 \log(S_{i,r}^u(U|\mathbf{Z}) - S_{i,r}^l(V|\mathbf{Z})) \\ &\quad + \delta_3 \log S_{i,r}^u(V|\mathbf{Z}) + (\delta_1 + \delta_2) \log \pi_{j,s}^u(\mathbf{X}, W) + \delta_3 \log(1 - \pi_{j,s}^l(\mathbf{X}, W)). \end{aligned}$$

By applying similar arguments as above and using (C3), we can obtain $l_{i,j,s,r}^u(O) - l_{i,j,s,r}^l(O) < C\epsilon$, which is true for all the choice of O . Hence it is

true in $L_1(\mathbb{P}_n)$ norm as well, thus the set of pairs $[l_{i,j,s,r}^l, l_{i,j,s,r}^u]$ for $1 \leq i \leq C(1/\epsilon)^{Kq_n}$, $1 \leq j \leq C(1/\epsilon)^{Kq_n}$, $1 \leq s \leq \lfloor C(1/\epsilon)^{p_1} \rfloor$ and $1 \leq r \leq \lfloor C(1/\epsilon)^{p_2} \rfloor$ form a bracket for $\mathfrak{L}_n$. Finally, from the fact that $N(\epsilon, \mathfrak{L}_n, L_1(\mathbb{P}_n)) \leq N_{[]} (2\epsilon, \mathfrak{L}_n, L_1(\mathbb{P}_n))$ true in general, the proof is finished. $\square$

Lemma 3.2 (Uniform convergence). If conditions (C1), (C2) and (C4) hold, and $\mathfrak{L}_n$ defined above has the covering number given by Lemma 3.1, then

$$\sup_{\theta \in \Theta_n} |\mathbb{P}_n l(\theta|O) - \mathbb{P} l(\theta|O)| \rightarrow 0$$

almost surely, as $n \rightarrow \infty$.

Proof. We follow the proof of Lemma 2 of Zhou et al. (2017) with some simplification and elaborate explanation. We begin by defining a sequence $\epsilon_n = \epsilon n^{-1/2+\phi_1}(\log n)^{1/2}$, where $\nu/2 < \phi_1 < 1/2$, and $\epsilon > 0$ is an arbitrary small number. Note that because of the choice of ϕ_1 , the sequence always tends to 0. By the conditions (C1), (C2) and (C4), it follows that $|l(\theta, O)|$ is bounded. We can assume without loss of generality that $\sup_{\theta \in \Theta_n} |l(\theta|O)| \leq 1$. From which we have $\mathbb{P} l^2(\theta|O) \leq \mathbb{P}(\sup_{\theta \in \Theta_n} |l(\theta|O)|^2) \leq 1$.

By the fact that $\text{Var}(\mathbb{P}_n l(\theta|O)) \leq E[(\mathbb{P}_n l(\theta|O))^2] = \frac{1}{n} \mathbb{P} l^2(\theta|O)$ we can put

$$\frac{\text{Var}(\mathbb{P}_n l(\theta|O))}{(4\epsilon_n)^2} \leq \frac{\mathbb{P} l^2(\theta|O)}{16n\epsilon_n^2} \leq \frac{1}{16n\epsilon_n^2} \leq \frac{1}{2}, \quad (3.11)$$

where the last inequality holds for large enough n .

We introduce the signed measure $\mathbb{P}_n^0$ that puts $\pm 1/n$ at each observation $\{O_1, \dots, O_n\}$ with random signs. From Pollard (1984, p. 31) and (3.11) we have the symmetrization inequality

$$\mathbb{P}(\sup_{\theta \in \Theta_n} |\mathbb{P}_n l(\theta|O) - \mathbb{P} l(\theta|O)| \geq 8\epsilon_n) \leq 4\mathbb{P}(\sup_{\theta \in \Theta_n} |\mathbb{P}_n^0 l(\theta|O)| \geq 2\epsilon_n) \quad (3.12)$$

Let $\theta^{(1)}, \theta^{(2)}, \dots, \theta^{(N(\epsilon_n/2, \mathfrak{L}_n, L_1(\mathbb{P}_n)))}$ be the centers of each balls of the $\epsilon_n/2$ covering of $\mathfrak{L}_n$, that is

$$\min_j \mathbb{P}_n |l(\theta|O) - l(\theta^{(j)}, O)| < \epsilon_n/2$$

for all $\theta \in \Theta_n$, moreover, one can choose θ^* as one of the $\theta^{(j)}$ for each $\theta \in \Theta_n$ such that the difference is that small. Note that because of the construction of the brackets, those $\theta^{(j)}$ do not depend on the actual data $\mathcal{O}$ that the $\mathbb{P}_n$ measure is based on. Because of the triangle inequality,

$$\begin{aligned} P\left(\sup_{\theta \in \Theta_n} |P_n^0 l(\theta|O)| \geq 2\epsilon_n\right) \\ \leq P\left(\sup_{\theta \in \Theta_n} [|P_n^0 l(\theta^*|O)| + |P_n^0 l(\theta|O) - P_n^0 l(\theta^*|O)|] \geq 2\epsilon_n\right). \end{aligned} \quad (3.13)$$

We use the following calculation

$$\begin{aligned} |\mathbb{P}_n^0(l(\theta|O) - l(\theta^*|O))| &= \left| \frac{1}{n} \sum_{i=1}^n \pm(l(\theta|O) - l(\theta^*|O)) \right| \\ &\leq \frac{1}{n} \sum_{i=1}^n |l(\theta|O) - l(\theta^*|O)| \\ &= \mathbb{P}_n |l(\theta|O) - l(\theta^*|O)| \leq \epsilon_n/2 \end{aligned} \quad (3.14)$$

to substitute the right handside of (3.13) and get a bigger quantity.

$$\begin{aligned} &P\left(\sup_{\theta \in \Theta_n} [|P_n^0 l(\theta^*|O)| + |P_n^0 l(\theta|O) - P_n^0 l(\theta^*|O)|] \geq 2\epsilon_n\right) \\ &\leq P\left(\sup_{\theta \in \Theta_n} [|P_n^0 l(\theta^*|O)| + \epsilon_n/2] \geq 2\epsilon_n\right) \\ &\leq P\left(\sup_{\theta \in \Theta_n} [|P_n^0 l(\theta^*|O)|] \geq 3\epsilon_n/2\right) \\ &= P\left(\max_j [|P_n^0 l(\theta^{(j)}|O)|] \geq 3\epsilon_n/2\right) \\ &\leq N(\epsilon_n/2, \mathfrak{L}_n, L_1(\mathbb{P}_n)) \max_j P(|P_n^0 l(\theta^{(j)}|O)| \geq 3\epsilon_n/2) \\ &\leq N(\epsilon_n/2, \mathfrak{L}_n, L_1(\mathbb{P}_n)) \max_j P(|P_n^0 l(\theta^{(j)}|O)| \geq \epsilon_n), \end{aligned} \quad (3.15)$$

where the first inequality comes from substituting (3.14), the second one comes from adding a constant to both sides, the equality comes from the fact that $\theta^{(j)}$ can have finitely many values. The third inequality comes from the fact that the probability of the union of events is smaller than the number of events times the maximum of each probabilities. Finally, the last inequality is true, because we consider a larger event.

We use Hoeffding's inequality to show

$$\begin{aligned} P(|\mathbb{P}_n^0 l(\theta^{(j)}|O)| \geq \epsilon_n) &\leq 2 \exp \left(- \frac{2n^2 \epsilon_n^2}{\sum_{i=1}^n (2l(\theta^{(j)}|O_i))^2} \right) \\ &\leq 2 \exp \left(- n \epsilon_n^2 / 2 \right), \end{aligned} \quad (3.16)$$

where the last inequality holds because of the assumption that $\sup_{\theta \in \Theta_n} |l(\theta|O)| \leq 1$. By combining (3.13), (3.15) and (3.16)

$$\begin{aligned} P \left(\sup_{\theta \in \Theta_n} |\mathbb{P}_n^0 l(\theta|O)| \geq 2\epsilon_n, \right) &\leq 2N(\epsilon_n/2, \mathfrak{L}_n, L_1(\mathbb{P}_n)) \exp \left(- n \epsilon_n^2 / 2 \right) \\ &= 2C(\epsilon_n/2)^{-(2q_n+p_1+p_2)} \exp \left(- n \epsilon_n^2 / 2 \right). \end{aligned}$$

Next, we combine the result above with (3.12) and apply some algebra

$$\begin{aligned} P \left(\sup_{\theta \in \Theta_n} |\mathbb{P}_n l(\theta|O) - \mathbb{P} l(\theta|O)| \geq 8\epsilon_n \right) &\leq 4P \left(\sup_{\theta \in \Theta_n} |\mathbb{P}_n^0 l(\theta|O)| \geq 2\epsilon_n \right) \\ &\leq C(\epsilon_n/2)^{-(2Kq_n+p_1+p_2)} \exp \left(- n \epsilon_n^2 / 2 \right) \\ &\leq C(\epsilon n^{-1/2+\phi_1} (\log n)^{1/2} / 2)^{-(2Kq_n+p_1+p_2)} \exp \left(- n(\epsilon n^{-1/2+\phi_1} (\log n)^{1/2})^2 / 2 \right) \\ &= C(\epsilon n^{-1/2+\phi_1} (\log n)^{1/2} / 2)^{-(2Kq_n+p_1+p_2)} \exp \left(- \epsilon^2 n^{2\phi_1} (\log n) / 2 \right) \\ &= C \exp \left\{ - [(\phi_1 - 1/2) \log n + (\log \log n) / 2 + \log \epsilon - \log 2] \right. \\ &\quad \left. \times (2Kq_n + p_1 + p_2) - \epsilon^2 n^{2\phi_1} (\log n) / 2 \right\} \\ &\leq \exp(-C n^{2\phi_1} \log n), \end{aligned}$$

where the last inequality holds for large enough n because $n^{2\phi_1} > q_n = n^\nu$. Since the sum $\sum_{i=1}^\infty \exp(-n^{2\phi_1} \log n)$ is finite for all $\phi_1 > 0$, we have the

following

$$\sum_{n=0}^{\infty} P(\sup_{\theta \in \Theta_n} |\mathbb{P}_n l(\theta|O) - \mathbb{P}l(\theta|O)| \geq 8\epsilon_n) < \infty.$$

Therefore, by Borel–Cantelli lemma, only finitely many events in the argument of P are true, hence the proof is done. $\square$

Lemma 3.3. Let $\epsilon > 0$ be a fixed small number and denote $K_\epsilon = \{\theta : d(\theta, \theta_0) > \epsilon, \theta \in \Theta_n\}$. If conditions (C2) and (C4) hold, then

$$\mathbb{P}l(\theta_0|O) - \sup_{\theta \in K_\epsilon} \mathbb{P}l(\theta|O) > 0.$$

Proof. By similar arguments as in the proof of the Gibbs inequality, one can see that $\mathbb{P}l(\theta_0|O) \geq \mathbb{P}l(\theta|O)$ for any θ , with equality if and only if $l(\theta_0|O) = l(\theta|O)$ for almost every O . So, we need to see why the equality $\mathbb{P}l(\theta_0|O) = \sup_{\theta \in K_\epsilon} \mathbb{P}l(\theta|O)$ is not possible. Suppose that the equality holds. Then there exists a subsequence $\theta_n \in K_\epsilon$ such that $\mathbb{P}l(\theta_n|O) \rightarrow \mathbb{P}l(\theta_0|O)$. By conditions (C2) and (C4), the closure $\bar{\Theta}_n$ of Θ_n is compact, one can choose a subsequence n_k such that θ_{n_k} is convergent and tends to a value $\tilde{\theta} \in \bar{\Theta}_n$. By the continuity of the expectation and the likelihood function, $\mathbb{P}l(\tilde{\theta}|O) = \sup_{\theta \in K_\epsilon} \mathbb{P}l(\theta|O)$.

However, by the identifiability of the model, $l(\theta_0|O) = l(\tilde{\theta}|O)$ for almost every O is only possible when $\theta_0 = \tilde{\theta}$. But θ_{n_k} is in K_ϵ , so it cannot converge to θ_0 by the definition of K_ϵ . $\square$

Proof of Theorem 3.2. Let $\epsilon > 0$ be a fixed small number and denote $K_\epsilon = \{\theta : d(\theta, \theta_0) > \epsilon, \theta \in \Theta_n\}$. Let $\delta_\epsilon = \mathbb{P}l(\theta_0|O) - \sup_{\theta \in K_\epsilon} \mathbb{P}l(\theta|O)$. By Lemma 3.3, $\delta_\epsilon > 0$. By Lemma 3.2, there exists a threshold N such that if $n > N$ then $\sup_{\theta \in \Theta_n} |\mathbb{P}l(\theta|O) - \mathbb{P}_n l(\theta|O)| < \delta_\epsilon/2$. Let us assume that $\hat{\theta}_n \not\rightarrow \theta_0$. Then there is an infinite subsequence $\hat{\theta}_{n_k}$ such that $\hat{\theta}_{n_k} \in K_\epsilon$. Consider a

fixed $\hat{\theta}_{n_{k'}}$ in the subsequence such that $n_{k'} > N$. In that case

$$\begin{aligned}\mathbb{P}_{n_{k'}}l(\hat{\theta}_{n_{k'}}|O) &< \mathbb{P}l(\hat{\theta}_{n_{k'}}|O) + \frac{\delta_\epsilon}{2} \\ &\leq \sup_{\theta \in K_\epsilon} \mathbb{P}l(\theta|O) + \frac{\delta_\epsilon}{2} \\ &= \mathbb{P}l(\theta_0|O) - \frac{\delta_\epsilon}{2} < \mathbb{P}_{n_{k'}}l(\theta_0|O),\end{aligned}$$

where the first and the third inequalities are true because $n_{k'} > N$. At the two sides of the inequalities we get that $\mathbb{P}_{n_{k'}}l(\hat{\theta}_{n_{k'}}|O) < \mathbb{P}_{n_{k'}}l(\theta_0)$, which contradicts the likelihood maximization property of $\hat{\theta}_{n_{k'}}$ based on the sample of size $n_{k'}$, hence the assumption is false. $\square$

Proof of Theorem 3.3. By Lu et al. (2007) and Theorem 6.2.5 of Schumaker (2007) there exists $\Lambda_{0n} \in \mathfrak{B}_n$ and $f_{0n} \in \mathfrak{F}_n$ such that $\|\Lambda_{0n} - \Lambda_0\|_2 < n^{(\nu-1)/2}$ and $\|f_0 - f_{0n}\|_2 < n^{-\nu r}$, where Λ_0 and f_0 are the true functions of interest and ν is the growth rate of the sieve space defined in section 3.2.1. Using those two functions, we introduce $\theta_{0n} = (\alpha_0, \beta_0, \Lambda_{0n}, f_{0n})$, and note that $d(\theta_0, \theta_{0n}) = O(n^{-\min(\nu r, (1-\nu)/2)})$. Similarly to the arguments of Lemma 3.1, one can show that if η is small enough and $\eta/2 < d(\theta) \leq \eta$, then $\mathbb{P}l(\theta_0|O) - \mathbb{P}l(\theta_{0n}|O) < Cn^{-\min(\nu r, (1-\nu)/2)}$.

We give the convergence rate of $d(\hat{\theta}_n, \theta_{0n})$ by verifying the five conditions in Theorem 3.4.1 of van der Vaart and Wellner (1996). To see the first condition, let $\mathbb{M}_n = \mathbb{P}_n l(\theta|O)$ and $M_n(\theta) = M(\theta) = \mathbb{P}l(\theta|O)$ as used by van der Vaart and Wellner (1996). For any θ such that $\eta/2 \leq \theta \leq \eta$ and large enough n ,

$$\begin{aligned}M(\theta) - M(\theta_{0n}) &= M(\theta) - M(\theta_0) + M(\theta_0) - M(\theta_{0n}) \\ &\leq -c\eta^2 + cn^{-2r\nu} = -c\eta^2,\end{aligned}$$

which satisfies the first condition of Theorem 3.4.1 of van der Vaart and Wellner (1996). Next, we need to verify the existence of $\phi_n(\eta)$ such that that for small enough η

$$E \left[\sup_{\theta: \eta/2 < d(\theta, \theta_0) < \eta} |(\mathbb{M}_n - M)(\theta) - (\mathbb{M}_n - M)(\theta_{0,n})| \right] < \frac{\phi_n(\eta)}{\sqrt{n}}.$$

Define the space $\mathcal{L}_\eta = \{l(\theta|O) - l(\theta_{0,n}|O) : \eta/2 < d(\theta, \theta_{0,n}) \leq \eta\}$. Under Conditions (C1), (C2) and (C4), $\mathcal{L}_\eta$ is uniformly bounded. Moreover, with some algebraic manipulations similar to the one in case of brackets and for small enough η and for all $l(\theta|O) - l(\theta_{0,n}|O) \in L_\eta$, $\mathbb{P}(l(\theta|O) - l(\theta_{0,n}|O))^2 \leq C\eta^2$ for some positive constant C . By applying Lemma 3.4.2 of van der Vaart and Wellner (1996) we get

$$E[|\sqrt{n}(\mathbb{M}_n - M)|_{\mathcal{L}_\eta}] \leq C J_{[]}(\eta, \mathcal{L}_\eta, L_2(\mathbb{P})) \left[1 + \frac{J_{[]}(\eta, \mathcal{L}_\eta, L_2(\mathbb{P}))}{\eta^2 n^{1/2}} \right], \quad (3.17)$$

where $J_{[]}(\eta, \mathcal{L}_\eta, L_2(\mathbb{P})) = \int_0^\eta [1 + \log N_{[]}(\epsilon, \mathcal{L}_\eta, L_2(\mathbb{P}))]^{1/2} d\epsilon$, and $\|\cdot\|_{\mathcal{L}_\eta}$ is the uniform bound over the the space $\mathcal{L}_\eta$. To get an estimate for $J_{[]}(\eta, \mathcal{L}_\eta, L_2(\mathbb{P}))$, we need to estimate $\log N_{[]}(\epsilon, \mathcal{L}_\eta, L_2(\mathbb{P}))$. Similar calculations to Shen and Wong (1994, p.597) leads to the estimate $\log N_{[]}(\epsilon, \mathcal{L}_\eta, L_2(\mathbb{P})) \leq C(2q_n + p_1 + p_2) \log(\epsilon/\eta)$. This yields $J_{[]}(\eta, \mathcal{L}_\eta, L_2(\mathbb{P})) \leq C\eta(2q_n + p_1 + p_2)^{1/2}$. By substituting that into the right hand side of (3.17), we can select $\phi_n(\eta)$ as $C\eta(2q_n + p_1 + p_2)^{1/2} + C(2q_n + p_1 + p_2)n^{-1/2}$. $\phi_n(\eta)/\eta$ is a decreasing function of η . We choose $r_n = (2q_n + p_1 + p_2)^{-1/2}n^{1/2}$, which satisfies the third condition, $r_n^2 \phi_n(1/r_n) \leq Cn^{1/2}$.

The fourth condition, that $M_n(\hat{\theta}_n) \geq M_n(\theta_{0n})$ is always satisfied because of the likelihood maximization property of $\hat{\theta}_n$. Finally, from Theorem 3.2, the distance $d(\hat{\theta}_n, \theta_0)$ converges to 0, as well as $d(\theta_{0n}, \theta_0)$, which asserts $d(\hat{\theta}_n, \theta_{0n})$ tends to zero. Therefore by Theorem 3.4.1 of van der Vaart and Wellner (1996), $r_n d(\hat{\theta}_n, \theta_{0n}) = O_P(1)$. That gives $d(\hat{\theta}_n, \theta_{0n}) \leq O(n^{-\min(\nu r, (1-\nu)/2)}) + O_P(n^{-(1-\nu)/2}) = O_P(n^{-\min(\nu r, (1-\nu)/2)})$. $\square$

Proof of Theorem 3.4. We refer to Theorem 3 of Zhang et al. (2010), which provides a general way to prove asymptotic normality of a sieve semiparametric estimator. Our aim is to verify conditions (B1) – (B3) of that theorem.

We only need to prove the second part of condition (B1) because the function $\mathbb{P}_n \dot{l}_1(\hat{\boldsymbol{\sigma}}_n, \hat{\phi}_n, O) \equiv 0$, due to the properties of the maximum likelihood estimator. So we need to show that $\mathbb{P}_n \dot{l}_2(\hat{\boldsymbol{\sigma}}_n, \hat{\phi}_n, O)(\xi_0) = o(n^{-1/2})$. By Theorem 6.25 of Schumaker (2007) there exists a function ξ_{0n} such that $\|\xi_n - \xi_{0n}\| < Cn^{-r\nu}$, and $\mathbb{P}_n \dot{l}_2(\hat{\theta}_n, O)(\xi_{0n}) = 0$. Then $\mathbb{P}_n \dot{l}_2(\hat{\boldsymbol{\sigma}}_n, \hat{\phi}_n, O)(\xi_0)$ can be written as

$$\begin{aligned} \mathbb{P}_n \dot{l}_2(\hat{\boldsymbol{\sigma}}_n, \hat{\phi}_n, O)(\xi_0) - \mathbb{P}_n \dot{l}_2(\hat{\boldsymbol{\sigma}}_n, \hat{\phi}_n, O)(\xi_{0n}) \\ &= (\mathbb{P}_n - \mathbb{P})[\dot{l}_2(\hat{\boldsymbol{\sigma}}_n, \hat{\phi}_n, O)(\xi_0 - \xi_{0n})] \\ &+ \mathbb{P}[\dot{l}_2(\hat{\boldsymbol{\sigma}}_n, \hat{\phi}_n, O)(\xi_0 - \xi_{0n}) - \dot{l}_2(\boldsymbol{\sigma}_0, \phi_0, O)(\xi_0 - \xi_{0n})] \\ &= I_{1,n} + I_{2,n} \end{aligned} \tag{3.18}$$

We define $\mathcal{L}_{n2} = \{\dot{l}_2(\boldsymbol{\sigma}, \phi, O)(\xi_0 - \xi) : \xi \in \mathcal{S}_n \times \mathcal{M}_n, \|\xi - \xi_0\| < O(n^{-\nu})\}$. By conditions (C1) – (C4), similarly to the proof of Lemma 1, $\mathcal{L}_{n2}$ has a finite covering number as $\mathcal{L}_n$ in Lemma 1, hence by Theorem 2.5.6 of van der Vaart and Wellner (1996), $\mathcal{L}_{n2}$ is $\mathbb{P}$ -Donsker. Therefore by Corollary 2.3.12 of van der Vaart and Wellner (1996), $I_{1,n} = o(n^{-1/2})$. By the Cauchy–Schwartz inequality, conditions (C1)–(C4), and Theorem 3,

$$\begin{aligned} I_{2,n} &\leq Cd(\hat{\theta}_n, \theta_0)\|\xi_0 - \xi_{0n}\| \\ &= O(n^{-\min(\nu r, (1-\nu)/2) - \nu}) = O(n^{-\min(\nu(r+1), (1+\nu)/2)}) = o(n^{-1/2}). \end{aligned}$$

Condition (B2) can be proven similarly, by defining the class

$$\mathcal{L}_{3n} = \{l^*(\boldsymbol{\sigma}, \phi, O) - l^*(\boldsymbol{\sigma}_0, \phi_0, O) : \theta \in \mathcal{S}_n \times \mathcal{M}_n, d(\theta, \theta_0) < \eta\}.$$

For a small enough η , $\mathcal{L}_{3n}$ is a $\mathbb{P}$ -Donsker class and the upper bound of $(\mathbb{P} - \mathbb{P}_n)[l^*(\hat{\boldsymbol{\sigma}}_n, \hat{\phi}_n, O) - l^*(\boldsymbol{\sigma}_0, \phi_0, O)]$ follows. Condition (B3) can be proven by Taylor expansion and the convergence rate derived in Theorem 3. $\square$

n	Cure Rate	Method	Variable	True_Value	MeanBias	SE	SD	CP
500	15%	DSP	α	-1.8	-0.061	0.403	0.375	0.928
			β	1.0	0.022	0.108	0.101	0.952
		SSP	α	-1.8	0.026	0.386	0.361	0.920
			β	1.0	0.019	0.108	0.105	0.958
	35%	DSP	α	-3.4	-0.083	0.473	0.448	0.952
			β	1.0	0.029	0.131	0.121	0.928
		SSP	α	-3.4	0.052	0.457	0.429	0.920
			β	1.0	0.022	0.136	0.128	0.952
	50%	DSP	α	-8.5	-0.595	1.673	1.298	0.956
			β	1.0	0.022	0.292	0.193	0.868
		SSP	α	-8.5	-0.306	1.458	1.187	0.946
			β	1.0	-0.232	0.522	0.216	0.681
1000	15%	DSP	α	-1.8	-0.014	0.276	0.261	0.942
			β	1.0	0.010	0.073	0.070	0.948
		SSP	α	-1.8	0.063	0.267	0.253	0.940
			β	1.0	0.008	0.072	0.074	0.960
	35%	DSP	α	-3.4	-0.014	0.303	0.310	0.946
			β	1.0	0.016	0.086	0.083	0.944
		SSP	α	-3.4	0.104	0.290	0.300	0.934
			β	1.0	0.013	0.086	0.090	0.960
	50%	DSP	α	-8.5	-0.235	0.912	0.818	0.944
			β	1.0	0.031	0.168	0.131	0.900
		SSP	α	-8.5	-0.037	0.914	0.789	0.928
			β	1.0	-0.072	0.349	0.142	0.844
2000	15%	DSP	α	-1.8	-0.010	0.167	0.183	0.958
			β	1.0	0.006	0.052	0.049	0.938
		SSP	α	-1.8	0.064	0.160	0.178	0.956
			β	1.0	0.005	0.052	0.052	0.956
	35%	DSP	α	-3.4	-0.019	0.221	0.218	0.950
			β	1.0	0.008	0.064	0.058	0.916
		SSP	α	-3.4	0.095	0.209	0.211	0.920
			β	1.0	0.006	0.064	0.063	0.954
	50%	DSP	α	-8.5	-0.074	0.567	0.553	0.936
			β	1.0	0.015	0.097	0.088	0.934
		SSP	α	-8.5	0.135	0.550	0.534	0.926
			β	1.0	-0.003	0.156	0.097	0.948

Table 3.1: Numerical results for estimators in the first scenario.

n	Cure Rate	Method	Variable	True_Value	MeanBias	SE	SD	CP
500	15%	DSP	α	-1.8	-0.037	0.404	0.389	0.948
			β	1.0	0.016	0.092	0.093	0.962
		SSP	α	-1.8	-0.021	0.399	0.383	0.940
			β	1.0	0.014	0.093	0.094	0.972
	35%	DSP	α	-3.4	-0.114	0.448	0.435	0.942
			β	1.0	0.026	0.113	0.108	0.938
		SSP	α	-3.4	-0.085	0.445	0.428	0.946
			β	1.0	0.023	0.113	0.113	0.942
	50%	DSP	α	-8.5	-0.430	1.188	1.087	0.946
			β	1.0	0.045	0.190	0.166	0.914
		SSP	α	-8.5	-0.362	1.119	1.055	0.940
			β	1.0	-0.158	0.443	0.182	0.774
1000	15%	DSP	α	-1.8	-0.018	0.279	0.271	0.956
			β	1.0	0.007	0.065	0.064	0.954
		SSP	α	-1.8	-0.011	0.276	0.270	0.958
			β	1.0	0.005	0.064	0.066	0.958
	35%	DSP	α	-3.4	-0.048	0.301	0.301	0.958
			β	1.0	0.016	0.080	0.074	0.932
		SSP	α	-3.4	-0.035	0.298	0.300	0.952
			β	1.0	0.013	0.080	0.078	0.956
	50%	DSP	α	-8.5	-0.193	0.789	0.728	0.942
			β	1.0	0.036	0.131	0.112	0.908
		SSP	α	-8.5	-0.164	0.806	0.713	0.936
			β	1.0	-0.010	0.237	0.121	0.906
2000	15%	DSP	α	-1.8	-0.021	0.194	0.190	0.958
			β	1.0	0.003	0.046	0.045	0.944
		SSP	α	-1.8	-0.016	0.193	0.190	0.958
			β	1.0	0.002	0.046	0.047	0.954
	35%	DSP	α	-3.4	-0.042	0.209	0.211	0.950
			β	1.0	0.007	0.056	0.051	0.922
		SSP	α	-3.4	-0.034	0.207	0.210	0.952
			β	1.0	0.005	0.056	0.055	0.948
	50%	DSP	α	-8.5	-0.101	0.523	0.501	0.948
			β	1.0	0.016	0.087	0.077	0.924
		SSP	α	-8.5	-0.084	0.517	0.497	0.950
			β	1.0	0.011	0.094	0.085	0.952

Table 3.2: Numerical results for estimators in the second scenario.

	Latency			Incidence
Variable	TR360	Noadyn	Sex	Sex
Estimate	1.289	1.841	-0.311	1.581
SE	0.365	0.490	0.375	0.434

Table 3.3: *The estimated coefficient for the HDSD data set*

Chapter 4

Discussion and Future Research

In this thesis two semiparametric survival models were studied in the interval censored setting. In Chapter 2 we considered the AH model and developed a semiparametric sieve maximum likelihood estimator for interval censored data. A two-step estimation algorithm was provided to implement the estimator. Simulation studies showed that the proposed estimator is suitable to use in practice for a variety of situations.

The proposed method with some modifications can be applied for estimation of other semiparametric survival models with different censoring schemes. For example, one can apply it to the AFT model with interval censored data or to the same accelerated hazards model with right censored data. A more general model, the extended hazard model (Etezadi-Amoli and Ciampi, 1987) contains the AH model as a special case. No regression analysis of the extended hazard model with interval censored data has been conducted. The challenges of the AH model are also present in the extended hazard model as well. Therefore, an estimator similar to the one proposed in Chapter 2 can be considered. Extending our method to an AH-based mixture cure model with interval censored data is another issue

which deserves future research.

A limitation of the proposed estimator is that the standard error has to be approximated by a bootstrap estimator. Finding a suitable standard error estimator by theoretical approach can be the subject of future research.

In Chapter 3 we considered the semiparametric mixture cure model and generalized it in a way that both the latency and the incidence parts are semiparametric. The model incorporates some covariates that do not have a linear effect on the incidence. We developed a semiparametric sieve maximum likelihood estimator and provide an implementation of it. The double semiparametric model gives excellent results even if the true covariate effect is linear. In such cases, the standard logit model may be considered to obtain the estimator of the coefficient. Real data analysis was conducted by both proposed estimating algorithms.

In Chapter 3 the covariate that may not have a linear effect on the incidence was known. As we saw in the real data analysis, a continuous variable is good candidate for the non-linear variable. In case there is only one continuous covariate that affect the incidence according to the doctor's opinion, one can fit the double semiparametric model to the data and can graphically check the linearity. However, if there are more continuous covariates, it might be difficult to select only one variable for the non-linear effect. In this case, future work on selecting variables for the linear and nonlinear components would be desirable in the future.

Another interesting direction for future study could be to accommodate potentially non-linear explanatory variables W_i in the survival model part to describe the event time for susceptible subjects. For example,

$$\lambda(t|Z_i, W_i) = \lambda_0(t) \exp(\beta^T Z_i + \psi(W_i)),$$

which yields a partially linear Cox PH model for uncured subgroup. Alternatively, a partially linear time-varying coefficient model can also be specified for susceptible subjects. An intuitively appealing approach to fit these models is to approximate the nonparametric functions by another set of splines, and then construct the sieve maximum likelihood estimation as studied in this chapter. Further work is warranted to elaborate on the theoretical properties of the approach and its numerical implementation.

References

- Aalen, O. “Nonparametric Inference for a Family of Counting Processes.” *Annals of Statistics*, 6(4):701–726 (1978).
- Chen, Y., Hanson, T., and Zhang, J. “Accelerated hazards model based on parametric families generalized with Bernstein polynomials.” *Biometrics*, 70(1):192–201 (2014).
URL <http://www.ncbi.nlm.nih.gov/pubmed/24261450>
- Chen, Y. Q. “Accelerated hazards regression model and its adequacy for censored survival data.” *Biometrics*, 57(3):853–860 (2001).
- Chen, Y. Q. and Wang, M.-C. “Analysis of Accelerated Hazards Models.” *Journal of the American Statistical Association*, 95(450):608–618 (2000).
URL <http://www.jstor.org/stable/2669403>{%}5Cnfile:///Users/CYHuang/Documents/Dropbox/Missing/D3253A9D-6DBD-4C2E-81F0-EBBCE133ABA5{%}5Cnpapers2://publication/uuid/5AB9B32A-07FC-4E88-BECF-14A01C24502E
- Cheng, G. “How Many Iterations are Sufficient for Efficient Semiparametric Estimation?” *Scandinavian Journal of Statistics*, 40(2):592–618 (2013).

- Conkin, J., Bedhal, S. R., and Van Liew, H. D. “A computerized data-bank of decompression sickness incidence in altitude chambers.” *Aviation, Space, and Environmental Medicine*, 63(10):819–824 (1992).
- Cox, D. R. “Regression Models and Life-Tables.” *Journal of the Royal Statistical Society, Series B*, 34(2):187–220 (1972).
- Etezadi-Amoli, J. and Ciampi, A. “Extended Hazard Regression for Censored Survival Data with Covariates : A Spline Approximation for the Baseline Hazard Function.” *Biometrics*, 43(1):181–192 (1987).
- Farewell, V. T. “The Use of Mixture Models for the Analysis of Survival Data with Long-Term Survivors.” *Biometrics*, 38(4):1041–1046 (1982).
- Farrington, C. P. “Interval censored survival data: a generalized linear modelling approach.” *Statistics in Medicine*, 15(3):283–292 (1996).
- Finkelstein, D. M. “Hazards Model for Interval-Censored Failure Time Data.” *Biometrics*, 42(4):845–854 (1986).
- Geman, S. and Hwang, C.-r. “Nonparametric Maximum Likelihood Estimation by the Method Of Sieves.” *Annals of Statistics*, 10(2):401–414 (1982).
- Groeneboom, P. and Wellner, J. A. *Information Bounds and Nonparametric Maximum Likelihood Estimation*, volume 19. Springer (1992).
- Hu, T. and Xiang, L. “Partially linear transformation cure models for interval-censored data.” *Computational Statistics and Data Analysis*, 93(1):257–269 (2016).
- URL <http://dx.doi.org/10.1016/j.csda.2014.08.014>

Huang, J. and Rossini, A. J. “Sieve estimation for the proportional-odds failure-time regression model with interval censoring.” *American Statistical Association. Journal*, 92(439):960–967 (1997).

URL <http://amstat.tandfonline.com/doi/abs/10.1080/01621459.1997.10474050>{%}5Cnpapers2://publication/uuid/E60201EF-2BAB-4CC1-B537-E412482FD389

Huang, J. and Wellner, J. A. “Efficient Estimation for the Cox Model with ”Case 2” Interval Censoring.” (1995).

Huang, J., Zhang, Y., and Hua, L. “A least-squares approach to consistent information estimation in semiparametric models.” Technical report, Department of Biostatistics, University of Iowa (2008).

URL <http://cph.uiowa.edu/biostat/research/documents/2008-3-Huang-Zhang-Hua.pdf>

Kaplan, E. L. and Meier, P. “Nonparametric Estimation from Incomplete Observations.” *Journal of the American Statistical Association*, 52(282):457–481 (1958).

Kim, Y.-J. and Jhun, M. “Cure rate model with interval censored data.” *Statistics in medicine*, 27(1):3–14 (2008).

Klein, J. P. and Moeschberger, M. M. *Survival Analysis. Techniques for censored and Truncated data*. New York: Springer (1997).

Li, H., Zhang, J., and Tang, Y. “Induced Smoothing for the Semiparametric Accelerated Hazards Model.” *Computational statistics & data analysis*, 56(12):4312–4319 (2012).

URL <http://www.sciencedirect.com/science/article/pii/S0167947312001491>

López-Cheda, A., Cao, R., Jácome, M. A., and Van Keilegom, I. “Nonparametric incidence estimation and bootstrap bandwidth selection in mixture cure models.” *Computational Statistics and Data Analysis*, 105:144–165 (2017).

URL <http://dx.doi.org/10.1016/j.csda.2016.08.002>

Lu, M., Zhang, Y., and Huang, J. “Estimation of the mean function with panel count data using monotone polynomial splines.” *Biometrika*, 94(3):705–718 (2007).

Ma, S. “Mixed Case Interval Censored Data with a Cured Subgroup.” *Statistica Sinica*, 20:1165–1181 (2010).

Murphy, A. S. A. and van der Vaart, A. W. “On Profile Likelihood.” *Journal of the American Statistical Association*, 95(450):449–465 (2000).

Nelson, W. “Theory and Applications of Hazard Plotting for Censored Failure Data.” *Technometrics*, 14(4):945–966 (1972).

Oller, R., Gómez, G., and Calle, M. “Interval censoring: Model characterizations for the validity of the simplified likelihood.” *Canadian Journal of Statistics*, 32(3):315–326 (2004).

URL <http://doi.wiley.com/10.2307/3315932>

Oller, R., Gómez, G., and Calle, M. L. “Interval censoring: Identifiability and the constant-sum property.” *Biometrika*, 94(1):61–70 (2007).

Pan, W. “Extending the Iterative Convex Minorant Algorithm to the Cox Model for Interval-Censored Data.” *Journal of Computational and Graphical Statistics*, 8(1):109–120 (1999).

- Pollard, D. *Convergence of Stochastic Processes*. New York: Springer (1984).
- Rasmussen, S. S., Glümer, C., Sandbaek, A., Lauritzen, T., and Borch-Johnsen, K. “Determinants of progression from impaired fasting glucose and impaired glucose tolerance to diabetes in a high-risk screened population: 3 year follow-up in the ADDITION study, Denmark.” *Diabetologia*, 51(2):249–257 (2008).
- Rosenberg, P. S. “Hazard Function Estimation Using B-Splines.” *Biometrics*, 51(3):874–887 (1995).
- Schumaker, L. *Spline Functions: Basic Theory*. Cambridge: Cambridge University Press (2007).
- Shen, X. and Wong, W. H. “Convergence Rate of Sieve Estimates.” *The Annals of Statistics*, 22(2):580–615 (1994).
- Sun, J. *The statistical analysis of interval-censored failure time data*. New York: Springer (2006).
- URL <http://link.springer.com/content/pdf/10.1007/0-387-37119-2.pdf>
- Thompson, L. A. and Chhikara, R. S. “A Bayesian cure rate model for repeated measurements and interval censoring.” In *Proceedings of Joint Statistical Meeting, San Francisco* (2003).
- van der Vaart, A. *Asymptotic Statistics*. Cambridge: Cambridge University Press (1998).
- van der Vaart, A. W. and Wellner, J. A. *Weak convergence and empirical processes*. New York: Springer-Verlag (1996).

- Wang, L., Du, P., and Liang, H. “Two-Component Mixture Cure Rate Model with Spline Estimated Nonparametric Components.” *Biometrics*, 68(3):726–735 (2012).
- Wang, N., Wang, L., and McMahan, C. S. “Regression analysis of bivariate current status data under the Gamma-frailty proportional hazards model using the EM algorithm.” *Computational Statistics & Data Analysis*, 83:140–150 (2015).
- URL <http://www.sciencedirect.com/science/article/pii/S0167947314003053>
- Xu, J. and Peng, Y. “A generalized class of skew distributions and associated robust quantile regression models.” *Canadian Journal of Statistics*, 42(1):1–17 (2014).
- Yakovlev, A. Y. and Tsodikov, A. D. *Stochastic Models of Tumor Latency and Their Biostatistical Applications*. Singapore: World Scientific (1996).
- Zhang, J., Peng, Y., and Zhao, O. “A new semiparametric estimation method for accelerated hazard model.” *Biometrics*, 67(4):1352–1360 (2011).
- Zhang, M. and Davidian, M. “Smooth Semiparametric Regression Analysis for Arbitrarily Censored Time-to-Event Data.” *Biometrics*, 64(2):567–576 (2008).
- Zhang, Y., Hua, L., and Huang, J. “A Spline-Based Semiparametric Maximum Likelihood Estimation Method for the Cox Model with Interval-Censored Data.” *Scandinavian Journal of Statistics*, 37(2):338–354 (2010).

- Zhou, J., Zhang, J., and Lu, W. “An Expectation Maximization algorithm for fitting the generalized odds-rate model to interval censored data.” *Statistics in Medicine*, 36(7):1157–1171 (2017).