NOVEL COMPUTATIONAL TECHNIQUES FOR SEMIPARAMETRIC ANALYSIS WITH INTERVAL-CENSORED DATA

A Dissertation

by

TONG WANG

Submitted to the Graduate and Professional School of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	Samiran Sinha
Committee Members,	Bani K. Mallick
	Michael Longnecker
	Sanjukta Chakraborty
Head of Department,	Brani Vidakovic

August 2021

Major Subject: Statistics

ABSTRACT

In this dissertation, I focus on the semiparametric analysis for interval-censored data. Two types of interval-censored data, case-I and case-II, are considered. I develop two attractive MM algorithms that converge stably for these two scenarios.

The first problem is statistical inference for clustered current status data, i.e., the case-I interval-censored data. Current status data abounds in the field of epidemiology and public health, where the only observable information is the random inspection time, and the event status at inspection. A unified methodology is proposed to analyze such complex data that are subject to clustering. The time-to-event is assumed to follow the semiparametric generalized odds rate (GOR) model. The non-parametric component of the GOR model is approximated via penalized splines, with a set of knot points that increases with the sample size. The within-subject correlation is accounted for by a random (frailty) effect. For estimation, a novel MM algorithm is developed that allows us to separate the parametric and nonparametric components of the models. This separation eventually makes the problem conducive to the application of the Newton-Raphson algorithm that quickly returns the roots. The work is accompanied by a complexity analysis of the algorithm and a rigorous asymptotic theory and the related semiparametric efficiency of the proposed methodology. The finite sample performance of the proposed method is assessed via simulation studies. Furthermore, the proposed methodology is illustrated via the analysis of a real data on periodontal disease studies accompanied by diagnostic checks to identify influential observations.

The second problem refers modeling case-II interval-censored data via additive risks model. Semiparametric additive risks model is a popular model to assess the relationship between the hazard of an event and a set of covariates. Particularly, it allows to assess the change or the difference in the hazard function for changing the values of the covariates. The model has a nonparametric part and a regression part identified by a finite dimensional parameter. This part contains an efficient approach aided by the MM algorithm to estimate the nonparametric and the finite dimensional components of the model from an interval-censored data. The operating characteristics of

the computational approach is assessed via simulation studies, and the method is illustrated through a real data application. This computational approach will not only make the maximum likelihood method more popular in this particular scenario, but may also simplify the computational burden of other complex likelihoods or models.

DEDICATION

To my friends and family

ACKNOWLEDGMENTS

When I have thought of my graduate study in Texas A&M Universtiy, I admit that I could not finish it without the support of my family, my friends and professors in the Department of Statistics

First of all, I would like to express my appreciation to my advisor Professor Samiran Sinha for being strongly supportive throughout my Ph.D. program. I am honored to have had the opportunity to be Prof. Sinha's student. He has been not only guiding my doctoral work, but also giving encouragement and mental support to withstand some of the difficult periods of the graduate study at Texas A&M. He provides invaluable ideas and help on every single goal I have accomplished. He is the best and the kindest professor I have ever met. He makes me think I want to be a professor like him in the future. I also express my appreciation to Dr. Michael Longnecker, Dr. Bani K. Mallick and Dr. Sanjukta Chakraborty for their kind suggestions and support as members of my advisory committee.

I am grateful to Dr. Kejun He, Dr. Wei Ma and Dr. Dipankar Bandyopadhyay. They provide thoughtful and constructive advice which improved the work in Chapter 2 substantially. I would like to thank Dr. Xiaohui Xu and Dr. Taehyun Roh from the School of Public Health who provided much support in my being here. I am grateful for the research assistantship with them. Thanks also to Ms. Andrea Dawson for her assistance on paperwork.

I would also like to thank my dear friends Huijuan, Lihao, Fei, Huiya, Shirun, Jiangyuan, Guanxun, Honggang, Lin, Fangting and Hanxuan. I honor their friendship and so many good memories throughout my journey at Texas A&M University. I also thank my old friends Tiange, Zuming, Yeda, Bowen, Zehui, Hansen, Zhengkai, Chenwei and Jingwei. They encourage and accompany me through the difficult years in the pandemic.

Finally, I thank my parents who always stand beside me. I could not complete this long journey without their never-ending love and support. I can not find any word to express my deep appreciation to them.

CONTRIBUTORS AND FUNDING SOURCES

Contributors

This work was supported by a dissertation committee consisting of Dr. Samiran Sinha (advisor), Dr. Bani K. Mallick and Dr. Michael Longnecker of the Department of Statistics and Dr. Sanjukta Chakraborty of the College of Medicine. All work conducted for the dissertation was completed by the student independently

Funding Sources

Graduate study was supported by a teaching assistantship from Texas A&M University.

TABLE OF CONTENTS

Page

LIST OF FIGURES

FIGURE Page

LIST OF TABLES

1. INTRODUCTION

1.1 Interval censoring

Interval censoring, which occurs when the failure time is only known to lie in an interval instead of being observed exactly, abounds in epidemiology, clinical trials and longitudinal study. This type of incomplete data structure is usually caused by periodic follow-up. For example, in AIDS study, the exact time of being infected with HIV is only known within a time window since it is detected by blood tests which can only be performed periodically. There are two main types of interval-censored data: case-I and case-II interval-censored data. Case-I interval-censored data, also called current status data, abounds in the field of epidemiology and public health, where the only observable information is the random inspection time, and the event status at inspection. Case-II interval-censored data is a more general type of interval-censored data, that is a mixture of left, interval and censored time to occurrence of an event. There is a generalization of case-II interval-censored data, case-K interval censoring. Instead of only two inspection times in case-II interval-censored data, i.e., left and right endpoints, case-K interval-censored data contains a sequential inspection time for each subject. This general type is not the focus of this dissertation.

1.1.1 Case-I interval censoring

For Case-I interval-censored data, the subject is only inspected once. Suppose C is the inspection time, then the "failure time" T is only known whether it has happened before C or not. In other words, the subject is either left- or right-censored. For an observation, the data consists of (Δ, C, X), where $\Delta = I(T \leq C)$ and X is the covariate. In Chapter 2, I focus on statistical inference for case-I interval-censored data. This study is motivated by periodontal disease (PD) assessment from the Gullah speaking Aferican American Diabetic (GAAD) study (Fernandes et al., 2009). To illustrate the form of case-I interval-censored data, part of the GAAD data is shown in Table 1.1. In the table, InsT denotes the inspection time for each tooth and Δ is the indicator, taking values 0 or 1, for right- or left-censored, respectively.

Table 1.1: GAAD data

ID	tooth	InsT	Δ	gender	smoking	Hba1c	jaw
1	11	43.5	0	1	1	1	1
1	15	42.5	0	1	1	1	1
1	20	43.5	0	1	1	1	1
2	21	41.0	0	1	0	0	0
2	22	42.5	1	1	0	0	0
2	23	44.5	1	1	0	0	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

1.1.2 Case-II interval censoring

Compared to case-I interval-censored data, case-II is a more general scenario which contains left-, interval- and right- censored subjects. For each observation, the data usually consists of $\{L, R, \Delta_L, \Delta_I, \Delta_R, X\}$. If a subject is left censored, then $\Delta_L = 1$ and the unobserved "failure time" T falls in $(0, L]$. If the subject is interval censored, then $\Delta_I = 1$ and T falls in (L, R). If T is right censored, then $\Delta_R = 1$ and T falls in $[R, \infty)$. To be noted, for each subject, $\Delta_L + \Delta_I + \Delta_R = 1$, since T is either left, interval, or right censored. An example of case-II data is presented in Table 1.2. This data is given in Finkelstein and Wolfe (1985), about breast cosmesis study for breast cancer patients. The data is shown as the interval form (V, U). Here $V = 0$ refers to the left censoring, and $U = \infty$ refers to the right censoring.

Table 1.2: Breast cancer data

ID	V	U	Δ_L	Δ_I	Δ_R	X
1	45	∞	0	0	1	0
2	6	10	0	1	0	0
3	0	7	1	0	0	0
4	8	12	0	1	0	1
5	0	22	1	0	0	1
6	32	∞	0	0	1	1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

1.1.3 Nonparametric analysis

In clinical trials and longitudinal studies, one is usually interested in estimating the survival function of the event of interest $S(t) = \text{pr}(T > t)$. Then the log-likelihood function is given

$$l_n(S) = \sum_{i=1}^{n} \left[\Delta_{L,i} \log\{1 - S(L_i)\} + \Delta_{I,i} \log\{S(L_i) - S(R_i)\} + \Delta_{R,i} \log\{S(R_i)\}\right]. \quad (1.1)$$

The log-likelihood (1.1) can incorporate the case-I interval censoring by letting $\Delta_{I,i} = 0$ and $L_i = R_i = C_i$ for $i = 1, \ldots, n$, which will reduce $l_n(S)$ to

$$l_n(S) = \sum_{i=1}^{n} \left[\Delta_i \log\{1 - S(C_i)\} + (1 - \Delta_i) \log\{S(C_i)\}\right],$$

where $\Delta_i = \Delta_{L,i}$. For current status data, the nonparametric maximum likelihood estimator (NPMLE) $\widehat{S}_n(t)$ can be computed by the max-min formula (Huang and Wellner, 1997)

$$\widehat{S}_n(C_{(i)}) = 1 - \max_{j \le i} \min_{k \le i} \frac{\sum_{m=j}^{k} \Delta_{(m)}}{k - j + 1}, \quad (1.2)$$

where $C_{(1)} \leq C_{(2)} \leq \ldots \leq C_{(n)}$ and $\Delta_{(i)}$ is the indicator according to $C_{(i)}$.

For case-II interval-censored data, there is no closed form of $\widehat{S}_n(t)$ like expression (1.2). Several iterative methods have been proposed to compute the NPMLE in this scenario. The first method was given in Turnbull (1976). It is essentially an EM-based method. The estimator is updated iteratively until convergence. Although it converges slowly, it is still commonly-used for easy implementation. Groeneboom and Wellner (1992) proposed an iterative convex minorant (ICM) to maximize the likelihood function, which converges faster than the EM algorithm. Wellner and Zhan (1997) proposed a hybrid algorithm of EM and ICM algorithms, named as EM-ICM algorithm, to compute the NPMLE. Combining the features of those two algorithms, EM-ICM is the fastest one. The sufficient condition for the unique estimate for these algorithms is the log-likelihood function (1.1) is strictly concave (Zhang and Sun, 2010), which can be checked by applying the Karush–Kuhn–Tucker (KKT) conditions (Gentleman and Geyer, 1994).

1.1.4 Regression analysis

Other than the nonparametric analysis, regression analysis is usually conducted to measure the covariate effect and predict the survival probabilities. There are several ways modelling the covariate effect on the time-to-event in regression analysis for interval-censored data. The proportional hazards (PH) model (Cox, 1972) is the most commonly used model. It models the hazard function of the failure time T as

$$\lambda(t|\boldsymbol{Z}) = \lambda_0(t)\exp(\boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{Z}), \tag{1.3}$$

where $\lambda_0(t)$ denotes the unknown baseline hazard function, $\boldsymbol{\beta}$ is the unknown regression parameter and $\boldsymbol{Z}$ is the covariate. Parameter estimation and statistical inference of PH model for interval-censored data can be found in Finkelstein (1986); Sun (2007); Satten (1996); Huang et al. (1996).

Alternative to the multiplicative association between the baseline hazard and the regression

part, additive risks model is another scheme of the hazard function as

$$\lambda(t|\boldsymbol{Z}) = \lambda_0(t) + \boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{Z}. \tag{1.4}$$

In this model, the effect of the covariate can be measured via the difference in the hazard function. To estimate $\boldsymbol{\beta}$ and $\lambda_0(t)$ of (1.4) for interval-censored data, Zeng et al. (2006) and Martinussen and Scheike (2002) proposed maximum likelihood method and a sieve approach, respectively.

Accelerated failure time (AFT) model is also popular in time-to-event data analysis. Instead of modeling the hazard function like (1.3) and (1.4), the AFT model of the survival time T is

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

where ϵ is an unspecified error term. For the AFT model with interval-censored data, Rabinowitz et al. (1995) and Betensky et al. (2001) proposed methods to estimate the regression parameter β along with the distribution of ϵ while Li and Pu (2003) developed a rank-based estimating equation to estimate β only.

An additional, but a flexible class of model is linear transformation model, where I write

$$g(F(t|\boldsymbol{Z})) = h(t) + \boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{Z}, \tag{1.5}$$

where $h(t)$ is an unknown strictly increasing function, g is a known link function and $F(t|\boldsymbol{Z})$ refers to the CDF of T given $\boldsymbol{Z}$. The specific form of linear transformation model depends on the link function g. Some popular models typically belong to the linear transformation model. With $g(s) = \log\{-\log(1-s)\}$ and $g(s) = \log\{s/(1-s)\}$, one obtain the PH model and proportional odds model, respectively. Some statistical inference for linear transformation models with interval-censored data can be seen in Sun and Sun (2005); Younes and Lachin (1997); Zhang et al. (2005); Zeng et al. (2016).

1.1.5 Clustering

In cross-sectional study and clinical trials, the event times are sometimes observed within the clusters, where the clusters may be the patients, families and tumors. Therefore, the event times within the same cluster are naturally correlated. For the current status data, there are plenty of such clustered data. For example, in GAAD data, the teeth of the same subject are clustered. The cataract dataset (Wen and Chen, 2011) is another example of clustered case-I interval-censored data. This dataset is from 2001 National Health Interview Survey (NHIS) Database, which records whether the cataracts for the left and right eyes are observed by the inspection time. It observes that for each patient, the left and right eyes are clustered and the data belongs to the case-I interval censored data. To handle the clustering effect for case-I interval censored data, Wen and Chen (2011) introduced a gamma-frailty to account for the unobserved clustering effect in Cox model. Alternative to frailty-based model, Cook and Tolusso (2009); Feng et al. (2019) considered marginal analysis for clustered current status data.

For clustered case-II interval censored data, an example is the pH1N1 dataset in Taiwan. It is from a cohort study of H1N1 flu in Taiwan during 2009-2010. In this study, several students along with their family members were recruited to take the blood samples in two different time periods to test whether they were infected. Therefore, the samples are clustered with the families and the exact time-to-event is only known to lie in the interval. Similar to the case-I scenario, there are mainly two types of methods to handle the clustering effect for case-II interval-censored data, i.e., frailty-based model and marginal analysis. Yavuz and Lambert (2016) and Li et al. (2012) considered introducing frailty in proportional hazards model and additive risks model, respectively. In Kor et al. (2013), the authors applied a generalized estimating equations approach to estimate the regression parameters in Cox model, which refers to a type of marginal analysis.

1.2 MM algorithm

The MM is short for "Majorize-Minimization" or "Minorize-Maximization" depending on whether the target is minimizing or maximizing the object function. Suppose the goal is maxi-

mizing the objective function $f(\theta)$, which is difficult due to the complicated form and high dimensions. Assume $f_\dagger(\theta|\theta_m)$ is a real-valued function of θ depending on the given value θ_m, then it is said to minorize function $f(\theta)$ at point θ_m if

$$f(\theta) \geq f_\dagger(\theta|\theta_m) \text{ for all } \theta, \text{ and } f(\theta_m) = f_\dagger(\theta_m|\theta_m).$$

In contrast, $-f_\dagger(\theta|\theta_m)$ majorizes function $-f(\theta)$ at point θ_m. In minorize-maximization algorithm, θ_m is updated via $\theta_{m+1} = \arg\max_\theta f_\dagger(\theta|\theta_m)$ until convergence. EM algorithm can be seen as a special case of MM algorithm, since the conditional expectation function in EM is a specific minorization function. To reduce the computational cost, Hunter and Lange (2004) proposed a gradient MM algorithm in the scenario that there is no closed form of maximizer for $f_\dagger(\theta|\theta_m)$. Instead of exactly maximizing the minorization function in each iteration, it updates θ_m via a one-step Newton-Raphson method

$$\theta_{m+1} = \theta_m - \{f_\dagger^{(2)}(\theta|\theta_m)\}^{-1} f_\dagger^{(1)}(\theta|\theta_m),$$

where $f_\dagger^{(2)}(\theta|\theta_m) \equiv \partial^2 f_\dagger(\theta|\theta_m)/\partial\theta\partial\theta^\mathsf{T}$ and $f_\dagger^{(1)}(\theta|\theta_m) \equiv \partial f_\dagger(\theta|\theta_m)/\partial\theta$. By selecting a proper minorization function, MM algorithm can transform a maximization with respect to a high dimensional parameter to a maximization with respect to several low dimensional parameters. Then it successfully avoids inverting a high dimensional matrix whose the computational cost is roughly proportional to the cubic order of the dimensions. The most difficult part of developing an MM algorithm is finding a suitable minorization function that locally approximates the objective function, and which is easier to maximize than directly maximizing the likelihood function.

2. MINORIZE-MAXIMIZE ALGORITHM FOR THE GENERALIZED ODDS RATE MODEL FOR CLUSTERED CURRENT STATUS DATA

2.1 Background and literature review

In epidemiological studies, a subject at risk for an event of interest is often monitored at a particular inspection time, and an indicator of whether the event has occurred is recorded. This generates *current status* information, henceforth CS, also called Case-I interval-censoring, a commonplace in biomedical research (Chen et al., 2012). The CS information implies that the subject (or study unit) is observed only at one time point, with no information between their study entry times and observation time points, leading to a severe form of interval-censoring.

Clustered CS data can arise if multiple time-to-events are recorded from the same observational units, multiple observational units belong to the same family, or twin pairs studies. The particular example I consider arises in periodontal disease (PD) studies, where the mean clinical attachment level (CAL)$\geq$ 3mm is the *landmark event* as it indicates *moderate to severe* PD status of a tooth (Armitage, 1999). Our interest is in modelling the covariate effect on the time to the landmark event. The time to this landmark event is available in the form of CS data, and the time to this landmark event for different teeth are correlated within a mouth, resulting in a clustered data.

I propose to fit a generalized odds rate model, henceforth GOR (Banerjee et al., 2007), to this data. The GOR model is attractive, as it encompasses a variety of models including the popular proportional hazard (PH), and proportional odds (PO) models, and can also be used to predict the survival probability of the onset of the landmark event beyond a given time. I model the nonparametric component of the GOR model via splines. To handle clustering, I introduce subject-specific random effects, and work with the conditional models which are useful in assessing the covariate effects at the subject level. The literature on inferential methods for clustered CS data is sparse; Wen and Chen (2011) considered a semiparametric Cox regression framework with a Gamma distributed cluster (frailty) effect, while some marginal approaches under generalized

estimating equations (Cook and Tolusso, 2009; Feng et al., 2019), and additive hazards models (Su and Chi, 2014) were also considered.

The key novelty of this work is developing a Minorize-Maximization (MM) algorithm (Hunter and Lange, 2004; Wu et al., 2010b) for our clustered CS setup under a GOR model. Aided by the simple Newton-Raphson algorithm, the MM procedure optimizes a complex likelihood function through a number of relatively easier steps. The most difficult part of developing an MM algorithm is finding a suitable minorization function that (a) locally approximates the objective function, and (b) enables easier optimization than direct maximization of the log likelihood function. This minorization step in the MM is built upon recognizing and manipulating mathematical inequalities. On the other hand, the celebrated EM algorithm is often used in a variety of maximum likelihood (ML) estimation scenarios in survival analysis, and the conditional expectation of the log of the complete data likelihood in EM is a specific minorization function. In that regard, the MM algorithm is a more general algorithm (Zhou and Zhang, 2012). Although both algorithms enjoy several advantages, such as achieving computational stability, natural adaptation to parameter constraints, and plausible amenability to big-data scenarios (Henderson and Varadhan, 2019), considering the MM route relieves me from the quintessential missing-data framework as desired in an EM formulation. Nonetheless, I also show that the computational complexity of our proposed MM algorithm is lower than the corresponding EM-based estimation route in our setup. Another major contribution is to provide asymptotic validation of our proposed estimator through consistency and weak convergence results, deemed suitable to handle the interplay between the number of knots and the tuning parameter, thereby achieving the semiparametric efficiency bound.

As a roadmap to the remainder of this chapter, Section 2.2 contains a brief introduction to the GOR model, the associated likelihood, and the regularized semiparametric estimator. Section 2.3 contains the MM algorithm for estimation. The asymptotic properties of our estimator, in light of identifiability, consistency and asymptotic normality, are given in Section 2.4, with their detailed proofs relegated to the Appendix A.2. Using synthetically generated data, the finite sample properties of our estimator are evaluated in Section 2.5. Application of our methodology to a real

data, along with influence diagnostics are presented in Section 2.6 while concluding remarks are given in Section 2.7.

2.2 Statistical model

The observed (clustered) CS data are $(C_{i,j}, \Delta_{i,j}, \boldsymbol{X}_{i,j}, \boldsymbol{Z}_i)$, $j = 1, \ldots, m_i$, $i = 1, \ldots, n$, where i denotes subjects (cluster), m_i denotes cardinality of the cluster i, and $C_{i,j}$ is the (current status) inspection time for the jth tooth of the ith subject. Considering $T_{i,j}$, the unobserved event time of interest corresponding to $C_{i,j}$, I further observe $\Delta_{i,j} = 1$ if $T_{i,j} \leq C_{i,j}$ and $\Delta_{i,j} = 0$ otherwise. Also, $\boldsymbol{X}_{i,j}$ denotes the tooth specific prognostic factors for the ith subject, and $\boldsymbol{Z}_i$ denotes the subject-specific covariates. I assume that both covariates are time-independent. I assume the conditional survival function of T on $\boldsymbol{X}$, $\boldsymbol{Z}$, and the subject-specific cluster effect b follows the GOR model,

$$S(t|\boldsymbol{X}, \boldsymbol{Z}, b) = \text{pr}(T > t|\boldsymbol{X}, \boldsymbol{Z}, b) = \frac{1}{\left\{1 + rH(t)\exp(\boldsymbol{\beta}^\mathsf{T}\boldsymbol{X} + \boldsymbol{\gamma}^\mathsf{T}\boldsymbol{Z} + \theta b)\right\}^{1/r}} \tag{2.1}$$

for $r > 0$. For $r = 1$, I obtain the PO model. When $r = 0$, I have the PH model with

$$S(t|\boldsymbol{X}, \boldsymbol{Z}, b) = \text{pr}(T > t|\boldsymbol{X}, \boldsymbol{Z}, b) = \exp\Big\{-H(t)\exp(\boldsymbol{\beta}^\mathsf{T}\boldsymbol{X} + \boldsymbol{\gamma}^\mathsf{T}\boldsymbol{Z} + \theta b)\Big\}. \tag{2.2}$$

I assume b follows $\text{Normal}(0, 1)$. Here $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are the regression parameters for $\boldsymbol{X}$ and $\boldsymbol{Z}$, θ^2 represents the cluster specific variance after adjusting the covariate effects, and $H(t)$ is a non-negative and non-decreasing function with $H(0) = 0$.

2.2.1 Likelihood construction and estimator

The likelihood function is

$$\begin{aligned}\mathcal{L}_n(\boldsymbol{\alpha}, H) = \prod_{i=1}^{n} \int \prod_{j=1}^{m_i} &\Big\{1 - S(C_{i,j}|\boldsymbol{X}_{i,j}, \boldsymbol{Z}_i, b_i)\Big\}^{\Delta_{i,j}} \Big\{S(C_{i,j}|\boldsymbol{X}_{i,j}, \boldsymbol{Z}_i, b_i)\Big\}^{1-\Delta_{i,j}} \\ &\times \phi(b_i)\mathrm{d}b_i,\end{aligned} \tag{2.3}$$

where ϕ denotes the standard normal density function and $\boldsymbol{\alpha} = (\boldsymbol{\beta}^\mathsf{T}, \boldsymbol{\gamma}^\mathsf{T}, \theta)^\mathsf{T}$ denotes the parameter vector. I approximate $H(t)$ by $H_\psi(t) = \sum_{k=1}^{K} M_k(t) \exp(\psi_k)$, where $M_1(t), \ldots, M_K(t)$ denote K monotone spline basis functions of degree d based on a given set of interior knot points $\tau_1 < \tau_2 < \cdots < \tau_L$ on the compact set $[0, T_0]$, and $\exp(\boldsymbol{\psi})$ is the set of non-negative regression parameters with $\boldsymbol{\psi} = (\psi_0, \psi_1, \ldots, \psi_K)^\mathsf{T}$ and $K = d + L$. In particular, I use I-splines defined as $M_k(t) = \int_0^t \mathcal{B}_k(u)\mathrm{d}u$, with $\mathcal{B}_k$'s being the B-spline basis functions (Ramsay et al., 1988). Consequently, $H_\psi(0) = 0$, since $M_k(0) = 0$, $k = 1, \ldots, K$.

To avoid potential approximation bias due to the specific choices of knots, I use a moderately large number of spline basis to estimate the model components. On the other hand, to overcome the challenge of data over-fitting, a more flexible penalized spline (Rice and Silverman, 1991) is used, serving as a pragmatic compromise between the regression and smoothing splines. The proposed regularized semiparametric estimator is defined as

$$(\widehat{\boldsymbol{\alpha}}_n{}^\mathsf{T}, \widehat{H}_n)^\mathsf{T} = \underset{\{\boldsymbol{\alpha},\ H_\psi(t)\}}{\arg\max} \left(\frac{1}{n} \log \left[\mathcal{L}_n \left\{ \boldsymbol{\alpha}, H_\psi(t) \right\} \right] - \lambda J^2(H_\psi) \right), \tag{2.4}$$

where λ is the penalty parameter and $J^2(\cdot)$ is the roughness penalty function. In particular, $J^2(H_\psi)$ denotes the squared integral of the qth order derivative of the function H_ψ with respect to t, which is assumed to be continuously differentiable up to order q, i.e., $J^2(H_\psi) = \int_0^{T_0} \{H_\psi^{(q)}(t)\}^2 \mathrm{d}t$. Although different q values can be used to define different penalty functions, I shall use $q = 2$ that measures the total curvature of the function (Ruppert et al., 2003) in our numerical studies. In our theoretical study, I investigate general $q \geq 2$. The regularization parameter λ controls how much wiggliness is allowed in the function. Note, a smaller value of λ allows a larger wiggliness resulting in overfitting the data (large variance), while a larger value allows only smaller wiggliness in the fitted curve resulting in under fitting the data (higher bias). In practice, λ should be determined using some data-driven method, such as AIC. See Subsection 2.3.2 for more discussions.

2.3 Estimation Methodology

2.3.1 MM algorithm

I now consider the estimation of $\boldsymbol{\xi} = (\boldsymbol{\alpha}^\mathsf{T}, \boldsymbol{\psi}^\mathsf{T})^\mathsf{T}$, where $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_K)^\mathsf{T}$ denotes the spline coefficients. Define a new function $G_{i,j}(\boldsymbol{\xi}, b_i)$ such that $\log\{G_{i,j}(\boldsymbol{\xi}, b_i)\} = -(1/r)\log\{1 + rH_\psi(C_{i,j})\exp(\boldsymbol{\beta}^\mathsf{T}\boldsymbol{X}_{i,j} + \boldsymbol{\gamma}^\mathsf{T}\boldsymbol{Z}_i + \theta b_i)\}$ when $r > 0$, and $\log\{G_{i,j}(\boldsymbol{\xi}, b_i)\} = -H_\psi(C_{i,j})\exp(\boldsymbol{\beta}^\mathsf{T}\boldsymbol{X}_{i,j} + \boldsymbol{\gamma}^\mathsf{T}\boldsymbol{Z}_i + \theta b_i)$ when $r = 0$. After replacing H by H_ψ and the integral by the Gauss-Hermite quadrature formula (Liu and Pierce, 1994) in expression (2.3), the likelihood becomes

$$\mathcal{L}_n(\boldsymbol{\xi}) = \prod_{i=1}^{n} \sum_{k} \prod_{j=1}^{m_i} \Big\{1 - G_{i,j}(\boldsymbol{\xi}, a_k)\Big\}^{\Delta_{i,j}} \Big\{G_{i,j}(\boldsymbol{\xi}, a_k)\Big\}^{1-\Delta_{i,j}} \omega(a_k), \tag{2.5}$$

where $a_1, a_2, \ldots$ are the quadrature points, with the corresponding weights $\omega(a_1), \omega(a_2), \ldots$. The log-likelihood $\ell(\boldsymbol{\xi}) = \log\{\mathcal{L}_n(\boldsymbol{\xi})\} = \sum_{i=1}^{n} \ell_i(\boldsymbol{\xi})$, where

$$\ell_i(\boldsymbol{\xi}) = \log\Bigg[\sum_{k} \prod_{j=1}^{m_i} \Big\{1 - G_{i,j}(\boldsymbol{\xi}, a_k)\Big\}^{\Delta_{i,j}} \Big\{G_{i,j}(\boldsymbol{\xi}, a_k)\Big\}^{1-\Delta_{i,j}} \omega(a_k)\Bigg]. \tag{2.6}$$

Due to its complex form, the direct maximization of $\ell(\boldsymbol{\xi}) - \lambda\mathcal{P}(\boldsymbol{\psi})$ is difficult, where $\mathcal{P}(\boldsymbol{\psi}) = \int_0^{T_0}\{H_\psi^{(q)}(t)\}^2 \mathrm{d}t$. This computational issue gets more severe as the size of $\boldsymbol{\psi}$ tends to increase with the sample size. Hence, I resort to developing a Minorize-Maximize (MM) algorithm by considering a suitable minorizing function.

Let $\boldsymbol{\xi}_0 = (\boldsymbol{\alpha}_0^\mathsf{T}, \boldsymbol{\psi}_0^\mathsf{T})^\mathsf{T}$ with $\boldsymbol{\alpha}_0 = (\boldsymbol{\beta}_0^\mathsf{T}, \boldsymbol{\gamma}_0^\mathsf{T}, \theta_0)^\mathsf{T}$ and $\boldsymbol{\psi}_0 = (\psi_{1,0}, \ldots, \psi_{K,0})^\mathsf{T}$. Our initial task is to find a minorization function $\ell_\dagger(\boldsymbol{\xi}|\boldsymbol{\xi}_0)$ for $\ell(\boldsymbol{\xi})$ that is relatively easy to maximize, and satisfies $\ell(\boldsymbol{\xi}) \geq \ell_\dagger(\boldsymbol{\xi}|\boldsymbol{\xi}_0)$, for all $\boldsymbol{\xi}_0$ and $\boldsymbol{\xi}$, with equality when $\boldsymbol{\xi} = \boldsymbol{\xi}_0$. For a given λ, $\ell_\dagger$ and $\boldsymbol{\xi}_0$, the aim is to obtain

$$\widehat{\boldsymbol{\xi}}(\boldsymbol{\xi}_0) = \arg\max_{\boldsymbol{\xi}} \ell_\dagger(\boldsymbol{\xi}|\boldsymbol{\xi}_0) - \lambda\mathcal{P}(\boldsymbol{\psi}). \tag{2.7}$$

Next, set $\boldsymbol{\xi}_0 = \widehat{\boldsymbol{\xi}}(\boldsymbol{\xi}_0)$. Then again, $\widehat{\boldsymbol{\xi}}(\boldsymbol{\xi}_0)$ can be obtained through step (2.7). In the general MM

framework, these two steps should be repeated until $\boldsymbol{\xi}_0$ and $\widehat{\boldsymbol{\xi}}(\boldsymbol{\xi}_0)$ are reasonably close.

Now, I present the most important result in Theorem 2.1 that is the key to obtaining the minorization functions. I strongly believe that this result will have future use in other models. The proof of this result is given in the Appendix A.1.1. Also, some well known inequalities used in our calculations are stated in Lemma 2.1, with their proofs omitted.

Theorem 2.1. *For any $u, u_0 > 0$,*

$$i)\ \log\left\{\frac{1-(1+ru)^{-1/r}}{1-(1+ru_0)^{-1/r}}\right\} \geq (u-u_0)A_1(u_0) - (u-u_0)^2 A_2(u_0) + \kappa\left\{\log\left(\frac{u_0}{u}\right) + 1 - \frac{u_0}{u}\right\}, \tag{2.8}$$

$$ii)\ \log\left\{\frac{1-\exp(-u)}{1-\exp(-u_0)}\right\} \geq (u-u_0)A_3(u_0) - (u-u_0)^2 A_4(u_0) + \log\left(\frac{u_0}{u}\right) + \left(1 - \frac{u_0}{u}\right) \tag{2.9}$$

where $\kappa = (1/r)I(0 < r \leq 1) + I(r > 1)$,

$$\begin{aligned}
A_1(u_0) &= \frac{(1+ru_0)^{-1/r-1}}{1-(1+ru_0)^{-1/r}}, \\
A_2(u_0) &= \frac{(1+ru_0)^{-1/r-2}[1+r\{1-(1+ru_0)^{-1/r}\}]}{2\{1-(1+ru_0)^{-1/r}\}^2}, \\
A_3(u_0) &= \frac{\exp(-u_0)}{1-\exp(-u_0)}, \\
A_4(u_0) &= \frac{1}{2}\left[\frac{\exp(-u_0)}{1-\exp(-u_0)} + \frac{\exp(-2u_0)}{\{1-\exp(-u_0)\}^2}\right].
\end{aligned}$$

Inequality (2.8) holds for any given $r > 0$, and in both inequalities, equality holds when $u = u_0$.

Lemma 2.1. *(i) For any $u > 0$, $\log(u) \geq 1 - 1/u$. (ii) For any $u \in \mathcal{R}$, $\exp(u) \geq 1 + u$. (iii) For any arbitrary $x_1, x_2, x_{10}, x_{20} > 0$, $x_1 x_2 \leq 0.5 x_{10} x_{20}\{(x_1/x_{10})^2 + (x_2/x_{20})^2\}$, and this result directly follows from the AM-GM inequality.*

Now, using the Jensen's inequality, and the concavity of the logarithm function on (2.6) I obtain

$$\ell_i(\boldsymbol{\xi}) \geq \ell_i(\boldsymbol{\xi}_0) + \sum_k \omega_i^*(\boldsymbol{\xi}_0, a_k) \log\left[\frac{\prod_{j=1}^{m_i}\{1-G_{i,j}(\boldsymbol{\xi},a_k)\}^{\Delta_{i,j}}\{G_{i,j}(\boldsymbol{\xi},a_k)\}^{1-\Delta_{i,j}}}{\prod_{j=1}^{m_i}\{1-G_{i,j}(\boldsymbol{\xi}_0,a_k)\}^{\Delta_{i,j}}\{G_{i,j}(\boldsymbol{\xi}_0,a_k)\}^{1-\Delta_{i,j}}}\right], \quad (2.10)$$

where $\sum_k \omega_i^*(\boldsymbol{\xi}_0, a_k) = 1$ and

$$\omega_i^*(\boldsymbol{\xi}_0, a_k) = \frac{\omega(a_k)\prod_{j=1}^{m_i}\{1-G_{i,j}(\boldsymbol{\xi}_0,a_k)\}^{\Delta_{i,j}}\{G_{i,j}(\boldsymbol{\xi}_0,a_k)\}^{1-\Delta_{i,j}}}{\sum_{k'}\omega(a_{k'})\prod_{j=1}^{m_i}\{1-G_{i,j}(\boldsymbol{\xi}_0,a_{k'})\}^{\Delta_{i,j}}\{G_{i,j}(\boldsymbol{\xi}_0,a_{k'})\}^{1-\Delta_{i,j}}}.$$

In (2.10), equality holds when $\boldsymbol{\xi} = \boldsymbol{\xi}_0$. Next, I consider two cases, $r > 0$ and $r = 0$ separately. For convenience, I will use the following abbreviated notations $u_{i,j,k}(\boldsymbol{\xi}) = H_\psi(C_{i,j})\exp(\boldsymbol{\alpha}^\mathsf{T}\boldsymbol{W}_{i,j,k})$ and $\boldsymbol{W}_{i,j,k} = (\boldsymbol{X}_{i,j}^\mathsf{T}, \boldsymbol{Z}_i^\mathsf{T}, a_k)^\mathsf{T}$.

Case: $r > 0$

After using the actual expressions of $G_{i,j}(\boldsymbol{\xi}, a_k)$ and $G_{i,j}(\boldsymbol{\xi}_0, a_k)$ for $r > 0$, I re-write inequality (2.10) as

$$\begin{aligned}
\ell_i(\boldsymbol{\xi}) &\geq \ell_i(\boldsymbol{\xi}_0) + \sum_k \omega_i^*(\boldsymbol{\xi}_0, a_k)\sum_{j=1}^{m_i}\left(\Delta_{i,j}\log\left[\frac{1-\{1+ru_{i,j,k}(\boldsymbol{\xi})\}^{-1/r}}{1-\{1+ru_{i,j,k}(\boldsymbol{\xi}_0)\}^{-1/r}}\right]\right. \\
&\quad \left. +(1-\Delta_{i,j})\kappa\log\left\{\frac{1+ru_{i,j,k}(\boldsymbol{\xi}_0)}{1+ru_{i,j,k}(\boldsymbol{\xi})}\right\}\right) \\
&\geq \ell_{\dagger,i}(\boldsymbol{\xi}|\boldsymbol{\xi}_0) = \ell_{\dagger,1,i}(\boldsymbol{\alpha}|\boldsymbol{\xi}_0) + \ell_{\dagger,2,i}(\boldsymbol{\psi}|\boldsymbol{\xi}_0) + \ell_{\dagger,3,i}(\boldsymbol{\xi}_0). \quad (2.11)
\end{aligned}$$

The inequality (2.11) follows after applying the results of Theorem 2.1 and Lemma 2.1, and the

detailed derivation is given in the Appendix A.1.2. Here,

$$
\begin{aligned}
\ell_{\dagger,1,i}(\boldsymbol{\alpha}|\boldsymbol{\xi}_0) &= \sum_k \omega_i^*(\boldsymbol{\xi}_0, a_k) \sum_{j=1}^{m_i} \Bigg[\Delta_{i,j} \Bigg\{ A_1(u_{i,j,k}(\boldsymbol{\xi}_0)) + 2A_2(u_{i,j,k}(\boldsymbol{\xi}_0)) u_{i,j,k}(\boldsymbol{\xi}_0) \Bigg\} \\
&\quad \times u_{i,j,k}(\boldsymbol{\xi}_0)(\boldsymbol{\alpha} - \boldsymbol{\alpha}_0)^\intercal \boldsymbol{W}_{i,j,k} - \left(\frac{\Delta_{i,j}}{2} \right) A_2(u_{i,j,k}(\boldsymbol{\xi}_0)) u_{i,j,k}^2(\boldsymbol{\xi}_0) \exp\{4(\boldsymbol{\alpha} - \boldsymbol{\alpha}_0)^\intercal \boldsymbol{W}_{i,j,k}\} \\
&\quad - \left(\frac{1 - \Delta_{i,j}}{2} \right) \frac{u_{i,j,k}(\boldsymbol{\xi}_0)}{1 + r u_{i,j,k}(\boldsymbol{\xi}_0)} \exp\{2(\boldsymbol{\alpha} - \boldsymbol{\alpha}_0)^\intercal \boldsymbol{W}_{i,j,k}\} \\
&\quad - \left(\frac{\Delta_{i,j}\kappa}{2} \right) \exp\{2(\boldsymbol{\alpha}_0 - \boldsymbol{\alpha})^\intercal \boldsymbol{W}_{i,j,k}\} - \Delta_{i,j}\kappa \boldsymbol{\alpha}^\intercal \boldsymbol{W}_{i,j,k} \Bigg], \\
\ell_{\dagger,2,i}(\boldsymbol{\psi}|\boldsymbol{\xi}_0) &= \sum_k \omega_i^*(\boldsymbol{\xi}_0, a_k) \sum_{j=1}^{m_i} \Bigg[\Delta_{i,j} \Bigg\{ A_1(u_{i,j,k}(\boldsymbol{\xi}_0)) + 2A_2(u_{i,j,k}(\boldsymbol{\xi}_0)) u_{i,j,k}(\boldsymbol{\xi}_0) \Bigg\} \\
&\quad \times u_{i,j,k}(\boldsymbol{\xi}_0) \log \left\{ \frac{H_\psi(C_{i,j})}{H_{\psi_0}(C_{i,j})} \right\} - \left(\frac{\Delta_{i,j}}{2} \right) A_2(u_{i,j,k}(\boldsymbol{\xi}_0)) u_{i,j,k}^2(\boldsymbol{\xi}_0) \left\{ \frac{H_\psi(C_{i,j})}{H_{\psi_0}(C_{i,j})} \right\}^4 \\
&\quad - \left(\frac{1 - \Delta_{i,j}}{2} \right) \frac{u_{i,j,k}(\boldsymbol{\xi}_0)}{1 + r u_{i,j,k}(\boldsymbol{\xi}_0)} \left\{ \frac{H_\psi(C_{i,j})}{H_{\psi_0}(C_{i,j})} \right\}^2 \\
&\quad - \left(\frac{\Delta_{i,j}\kappa}{2} \right) \left\{ \frac{H_{\psi_0}(C_{i,j})}{H_\psi(C_{i,j})} \right\}^2 - \Delta_{i,j}\kappa \log\{H_\psi(C_{i,j})\} \Bigg], \\
\ell_{\dagger,3,i}(\boldsymbol{\xi}_0) &= \ell_i(\boldsymbol{\xi}_0) + \sum_k \omega_i^*(\boldsymbol{\xi}_0, a_k) \sum_{j=1}^{m_i} \Bigg(\Delta_{i,j} A_2(u_{i,j,k}(\boldsymbol{\xi}_0)) u_{i,j,k}^2(\boldsymbol{\xi}_0) + (1 - \Delta_{i,j}) \frac{u_{i,j,k}(\boldsymbol{\xi}_0)}{1 + r u_{i,j,k}(\boldsymbol{\xi}_0)} \\
&\quad + \Delta_{i,j}\kappa \left[1 + \log\{u_{i,j,k}(\boldsymbol{\xi}_0)\} \right] \Bigg).
\end{aligned}
$$

Note that given $\boldsymbol{\xi}_0$, $\ell_{\dagger,1,i}(\boldsymbol{\alpha}|\boldsymbol{\xi}_0)$ involves with only $\boldsymbol{\alpha}$ while $\ell_{\dagger,2,i}(\boldsymbol{\psi}|\boldsymbol{\xi}_0)$ involves with only $\boldsymbol{\psi}$. Thus, the minorization function allows me to separate out the estimation of $\boldsymbol{\psi}$ and $\boldsymbol{\alpha}$. Now, for a given $\boldsymbol{\xi}_0$, I require to solve

$$
S(\boldsymbol{\alpha}|\boldsymbol{\xi}_0) \equiv \sum_{i=1}^n \partial \ell_{\dagger,i}(\boldsymbol{\xi}|\boldsymbol{\xi}_0)/\partial \boldsymbol{\alpha} = \sum_{i=1}^n \partial \ell_{\dagger,1,i}(\boldsymbol{\alpha}|\boldsymbol{\xi}_0)/\partial \boldsymbol{\alpha} = \mathbf{0} \text{ and } S(\boldsymbol{\psi}|\boldsymbol{\xi}_0) - \lambda \mathcal{P}_\psi(\boldsymbol{\psi}) = \mathbf{0},
$$

where

$$
S(\boldsymbol{\psi}|\boldsymbol{\xi}_0) \equiv \sum_{i=1}^n \partial \ell_{\dagger,i}(\boldsymbol{\xi}|\boldsymbol{\xi}_0)/\partial \boldsymbol{\psi} = \sum_{i=1}^n \partial \ell_{\dagger,2,i}(\boldsymbol{\psi}|\boldsymbol{\xi}_0)/\partial \boldsymbol{\psi},
$$

to obtain $\widehat{\boldsymbol{\xi}}(\boldsymbol{\xi}_0)$. Let me further define $S_\alpha(\boldsymbol{\alpha}|\boldsymbol{\xi}_0) = \partial S(\boldsymbol{\alpha}|\boldsymbol{\xi}_0)/\partial \alpha$ and $S_\psi(\boldsymbol{\psi}|\boldsymbol{\xi}_0) = \partial S(\boldsymbol{\psi}|\boldsymbol{\xi}_0)/\partial \boldsymbol{\psi}$.

As understood, $\widehat{\boldsymbol{\xi}}(\boldsymbol{\xi}_0)$ needs to be obtained using an iterative procedure. Thus, to avoid iterations within iterations as instructed in the general MM algorithm, I propose parameter updating via the one-step Newton-Raphson method, also known as the gradient MM algorithm (Hunter and Lange, 2004). Thus, for a given λ, the estimation algorithm can be presented as follows.

Step 1. Initialize the parameters $\boldsymbol{\xi}$.

Step 2. At the mth step, update the parameters as follows:

$$\begin{aligned}\boldsymbol{\alpha}^{(m)} &= \boldsymbol{\alpha}^{(m-1)} - \left\{S_{\alpha}(\boldsymbol{\alpha}^{(m-1)}|\boldsymbol{\xi}^{(m-1)})\right\}^{-1} S(\boldsymbol{\alpha}^{(m-1)}|\boldsymbol{\xi}^{(m-1)}) \\ \boldsymbol{\psi}^{(m)} &= \boldsymbol{\psi}^{(m-1)} - \left\{S_{\psi}(\boldsymbol{\psi}^{(m-1)}|\boldsymbol{\xi}^{(m-1)}) - \lambda\mathcal{P}_{\psi\psi}(\boldsymbol{\psi})\right\}^{-1} \left\{S(\boldsymbol{\psi}^{(m-1)}|\boldsymbol{\xi}^{(m-1)}) - \lambda\mathcal{P}_{\psi}(\boldsymbol{\psi})\right\},\end{aligned} \tag{2.12}$$

where $\mathcal{P}_{\psi}(\boldsymbol{\psi}) = \partial\mathcal{P}(\boldsymbol{\psi})/\partial\boldsymbol{\psi}$ and $\mathcal{P}_{\psi\psi}(\boldsymbol{\psi}) = \partial^2\mathcal{P}(\boldsymbol{\psi})/\partial\boldsymbol{\psi}\partial\boldsymbol{\psi}^{\mathsf{T}}$.

Step 3. Keep repeating Step 2, until $|(\boldsymbol{\xi}^{(m)} - \boldsymbol{\xi}^{(m-1)})/\boldsymbol{\xi}^{(m-1)}|^{\mathsf{T}}\mathbf{1}$ is smaller than a given tolerance ϵ_t.

It should be noted that in (2.12) $\boldsymbol{\alpha}$ and $\boldsymbol{\psi}$ are updated separately. The terms of equation (2.12) are

$$\begin{aligned}S(\boldsymbol{\alpha}^{(m-1)}|\boldsymbol{\xi}^{(m-1)}) &= \sum_{i=1}^{n}\sum_{k}\omega_i^*(\boldsymbol{\xi}^{(m-1)}, a_k)\sum_{j=1}^{m_i}\Bigg\{\Delta_{i,j}A_1(u_{i,j,k}(\boldsymbol{\xi}^{(m-1)})) \\ &\quad -\frac{(1-\Delta_{i,j})}{1+ru_{i,j,k}(\boldsymbol{\xi}^{(m-1)})}\Bigg\}u_{i,j,k}(\boldsymbol{\xi}^{(m-1)})\boldsymbol{W}_{i,j,k}, \\ S_{\alpha}(\boldsymbol{\alpha}^{(m-1)}|\boldsymbol{\xi}^{(m-1)}) &= -\sum_{i=1}^{n}\sum_{k}\omega_i^*(\boldsymbol{\xi}^{(m-1)}, a_k)\sum_{j=1}^{m_i}\Bigg[8\Delta_{i,j}A_2(u_{i,j,k}(\boldsymbol{\xi}^{(m-1)}))u_{i,j,k}^2(\boldsymbol{\xi}^{(m-1)}) \\ &\quad +2\left(1-\Delta_{i,j}\right)\frac{u_{i,j,k}(\boldsymbol{\xi}^{(m-1)})}{1+ru_{i,j,k}(\boldsymbol{\xi}^{(m-1)})} + 2\Delta_{i,j}\kappa\Bigg]\boldsymbol{W}_{i,j,k}^{\otimes 2}, \\ S(\boldsymbol{\psi}^{(m-1)}|\boldsymbol{\xi}^{(m-1)}) &= \sum_{i=1}^{n}\sum_{k}\omega_i^*(\boldsymbol{\xi}^{(m-1)}, a_k)\sum_{j=1}^{m_i}\Bigg\{\Delta_{i,j}A_1(u_{i,j,k}(\boldsymbol{\xi}^{(m-1)}))u_{i,j,k}(\boldsymbol{\xi}^{(m-1)}) \\ &\quad -\left(1-\Delta_{i,j}\right)\frac{u_{i,j,k}(\boldsymbol{\xi}^{(m-1)})}{1+ru_{i,j,k}(\boldsymbol{\xi}^{(m-1)})}\Bigg\}\left[\frac{\partial\log\{H_{\psi}(C_{i,j})\}}{\partial\boldsymbol{\psi}}\right]_{\psi=\psi^{(m-1)}},\end{aligned}$$

$$
\begin{aligned}
S_{\psi}(\boldsymbol{\psi}^{(m-1)}|\boldsymbol{\xi}^{(m-1)}) &= \sum_{i=1}^{n}\sum_{k}\omega_i^*(\boldsymbol{\xi}^{(m-1)},a_k)\sum_{j=1}^{m_i}\Big\{\Delta_{i,j}A_1(u_{i,j,k}(\boldsymbol{\xi}^{(m-1)}))u_{i,j,k}(\boldsymbol{\xi}^{(m-1)}) \\
&\quad -(1-\Delta_{i,j})\,\frac{u_{i,j,k}(\boldsymbol{\xi}^{(m-1)})}{1+ru_{i,j,k}(\boldsymbol{\xi}^{(m-1)})}\Big\}\left[\frac{\partial^2\log\{H_{\psi}(C_{i,j})\}}{\partial\boldsymbol{\psi}\partial\boldsymbol{\psi}^{\mathsf{T}}}\right]_{\psi=\psi^{(m-1)}} \\
&\quad -\sum_{i=1}^{n}\sum_{k}\omega_i^*(\boldsymbol{\xi}^{(m-1)},a_k)\sum_{j=1}^{m_i}\Big\{8\Delta_{i,j}A_2(u_{i,j,k}(\boldsymbol{\xi}^{(m-1)}))u_{i,j,k}^2(\boldsymbol{\xi}^{(m-1)}) \\
&\quad +2\,(1-\Delta_{i,j})\,\frac{u_{i,j,k}(\boldsymbol{\xi}^{(m-1)})}{1+ru_{i,j,k}(\boldsymbol{\xi}^{(m-1)})}+2\Delta_{i,j}\kappa\Big\}\left(\left[\frac{\partial\log\{H_{\psi}(C_{i,j})\}}{\partial\boldsymbol{\psi}}\right]_{\psi=\psi^{(m-1)}}\right)^{\otimes 2},
\end{aligned}
$$

where $a^{\otimes 2}$ denotes aa^{T} for any generic vector or matrix a.

To ascertain the estimate of θ is always positive, I propose to write θ as $\exp(\eta)$, and that require me to simply replace θ by $\exp(\eta)$ in all the expressions involving θ. Furthermore, I redefine $\boldsymbol{\alpha}=(\boldsymbol{\beta}^{\mathsf{T}},\boldsymbol{\gamma}^{\mathsf{T}},\eta)^{\mathsf{T}}$, leading to the revised expressions of $S(\boldsymbol{\alpha}^{(m-1)}|\boldsymbol{\xi}^{(m-1)})$, and $S_{\alpha}(\boldsymbol{\alpha}^{(m-1)}|\boldsymbol{\xi}^{(m-1)})$ as:

$$
\begin{aligned}
S(\boldsymbol{\alpha}^{(m-1)}|\boldsymbol{\xi}^{(m-1)}) &= \sum_{i=1}^{n}\sum_{k}\omega_i^*(\boldsymbol{\xi}^{(m-1)},a_k)\sum_{j=1}^{m_i}\Big\{\Delta_{i,j}A_1(u_{i,j,k}(\boldsymbol{\xi}^{(m-1)})u_{i,j,k}(\boldsymbol{\xi}^{(m-1)}) \\
&\quad -(1-\Delta_{i,j})\,\frac{u_{i,j,k}(\boldsymbol{\xi}^{(m-1)})}{1+ru_{i,j,k}(\boldsymbol{\xi}^{(m-1)})}\Big\}\boldsymbol{W}_{i,j,k}\circ e_{\theta^{(m-1)}}, \\
S_{\alpha}(\boldsymbol{\alpha}^{(m-1)}|\boldsymbol{\xi}^{(m-1)}) &= -\sum_{i=1}^{n}\sum_{k}\omega_i^*(\boldsymbol{\xi}^{(m-1)},a_k)\sum_{j=1}^{m_i}\Big[8\Delta_{i,j}A_2(u_{i,j,k}(\boldsymbol{\xi}^{(m-1)}))u_{i,j,k}^2(\boldsymbol{\xi}^{(m-1)}) \\
&\quad +2\,(1-\Delta_{i,j})\,\frac{u_{i,j,k}(\boldsymbol{\xi}^{(m-1)})}{1+ru_{i,j,k}(\boldsymbol{\xi}^{(m-1)})}+2\Delta_{i,j}\kappa\Big](\boldsymbol{W}_{i,j,k}\circ e_{\theta^{(m-1)}})^{\otimes 2} \\
&\quad +\mathrm{Diag}(0,\ldots,0,v_1^{\mathsf{T}}S(\boldsymbol{\alpha}^{(m-1)}|\boldsymbol{\xi}^{(m-1)})),
\end{aligned}
$$

respectively, where $\circ$ denotes element-wise multiplication, $e_{\theta^{(m-1)}}=(1,\ldots,1,\theta^{(m-1)})^{\mathsf{T}}$, $\theta^{(m-1)}=\exp(\eta^{(m-1)})$, $v_1=(0,0,\ldots,0,1)^{\mathsf{T}}$, and Diag refers to a diagonal matrix.

Case: $r=0$

For $r=0$, starting from inequality (2.10), I can derive $\ell_i(\boldsymbol{\xi})\geq\ell_{\dagger,i}(\boldsymbol{\xi}|\boldsymbol{\xi}_0)\equiv\ell_{\dagger,1,i}(\boldsymbol{\alpha}|\boldsymbol{\xi}_0)+\ell_{\dagger,2,i}(\boldsymbol{\psi}|\boldsymbol{\xi}_0)+\ell_{\dagger,3,i}(\boldsymbol{\xi}_0)$, where

$$
\begin{aligned}
\ell_{\dagger,1,i}(\boldsymbol{\alpha}|\boldsymbol{\xi}_0) &= \sum_k \omega_i^*(\boldsymbol{\xi}_0, a_k) \sum_{j=1}^{m_i} \Big[\Delta_{i,j} \{ A_3(u_{i,j,k}(\boldsymbol{\xi}_0)) + 2A_4(u_{i,j,k}(\boldsymbol{\xi}_0)) u_{i,j,k}(\boldsymbol{\xi}_0) \} \\
&\quad \times u_{i,j,k}(\boldsymbol{\xi}_0)(\boldsymbol{\alpha} - \boldsymbol{\alpha}_0)^\mathsf{T} \boldsymbol{W}_{i,j,k} - \Delta_{i,j} \boldsymbol{\alpha}^\mathsf{T} \boldsymbol{W}_{i,j,k} \\
&\quad - \frac{\Delta_{i,j}}{2} A_4(u_{i,j,k}(\boldsymbol{\xi}_0)) u_{i,j,k}^2(\boldsymbol{\xi}_0) \exp\{4(\boldsymbol{\alpha} - \boldsymbol{\alpha}_0)^\mathsf{T} \boldsymbol{W}_{i,j,k}\} \\
&\quad - \frac{(1 - \Delta_{i,j})}{2} u_{i,j,k}(\boldsymbol{\xi}_0) \exp\{2(\boldsymbol{\alpha} - \boldsymbol{\alpha}_0)^\mathsf{T} \boldsymbol{W}_{i,j,k}\} - \frac{\Delta_{i,j}}{2} \exp\{2(\boldsymbol{\alpha}_0 - \boldsymbol{\alpha})^\mathsf{T} \boldsymbol{W}_{i,j,k}\} \Big], \\
\ell_{\dagger,2,i}(\boldsymbol{\psi}|\boldsymbol{\xi}_0) &= \sum_k \omega_i^*(\boldsymbol{\xi}_0, a_k) \sum_{j=1}^{m_i} \Big[\Delta_{i,j} \{ A_3(u_{i,j,k}(\boldsymbol{\xi}_0)) + 2A_4(u_{i,j,k}(\boldsymbol{\xi}_0)) u_{i,j,k}(\boldsymbol{\xi}_0) \} \\
&\quad \times u_{i,j,k}(\boldsymbol{\xi}_0) \log \left\{ \frac{H_\psi(C_{i,j})}{H_{\psi_0}(C_{i,j})} \right\} - \Delta_{i,j} \log\{H_\psi(C_{i,j})\} \\
&\quad - \Delta_{i,j} A_4(u_{i,j,k}(\boldsymbol{\xi}_0)) u_{i,j,k}^2(\boldsymbol{\xi}_0) \left(\frac{1}{2} \right) \left\{ \frac{H_\psi(C_{i,j})}{H_{\psi_0}(C_{i,j})} \right\}^4 \\
&\quad - (1 - \Delta_{i,j})\, u_{i,j,k}(\boldsymbol{\xi}_0) \left(\frac{1}{2} \right) \left\{ \frac{H_\psi(C_{i,j})}{H_{\psi_0}(C_{i,j})} \right\}^2 - \frac{\Delta_{i,j}}{2} \left\{ \frac{H_{\psi_0}(C_{i,j})}{H_\psi(C_{i,j})} \right\}^2 \Big], \\
\ell_{\dagger,3,i}(\boldsymbol{\xi}_0) &= \ell_i(\boldsymbol{\xi}_0) + \sum_k \omega_i^*(\boldsymbol{\xi}_0, a_k) \sum_{j=1}^{m_i} \Big\{ (1 - \Delta_{i,j}) u_{i,j,k}(\boldsymbol{\xi}_0) + \Delta_{i,j} + \Delta_{i,j} \log\{u_{i,j,k}(\boldsymbol{\xi}_0)\} \Big\}.
\end{aligned}
$$

Applying the similar technique as of the $r > 0$ case, here $\boldsymbol{\alpha}$ and $\boldsymbol{\psi}$ are estimated by the generic Newton-Raphson algorithm given in (2.12) with

$$
\begin{aligned}
S(\boldsymbol{\alpha}^{(m-1)}|\boldsymbol{\xi}^{(m-1)}) &= \sum_{i=1}^{n} \sum_k \omega_i^*(\boldsymbol{\xi}^{(m-1)}, a_k) \sum_{j=1}^{m_i} \Big\{ \Delta_{i,j} A_3(u_{i,j,k}(\boldsymbol{\xi}^{(m-1)})) u_{i,j,k}(\boldsymbol{\xi}^{(m-1)}) \\
&\quad - (1 - \Delta_{i,j})\, u_{i,j,k}(\boldsymbol{\xi}^{(m-1)}) \Big\} \boldsymbol{W}_{i,j,k} \circ e_{\theta^{(m-1)}}, \\
S_\alpha(\boldsymbol{\alpha}^{(m-1)}|\boldsymbol{\xi}^{(m-1)}) &= -\sum_{i=1}^{n} \sum_k \omega_i^*(\boldsymbol{\xi}^{(m-1)}, a_k) \sum_{j=1}^{m_i} \Big[8\Delta_{i,j} A_4(u_{i,j,k}(\boldsymbol{\xi}^{(m-1)})) u_{i,j,k}^2(\boldsymbol{\xi}^{(m-1)}) \\
&\quad + 2\,(1 - \Delta_{i,j})\, u_{i,j,k}(\boldsymbol{\xi}^{(m-1)}) + 2\Delta_{i,j} \Big] (\boldsymbol{W}_{i,j,k} \circ e_{\theta^{(m-1)}})^{\otimes 2} \\
&\quad + \mathrm{Diag}(0, \ldots, 0, v_1^\mathsf{T} S(\boldsymbol{\alpha}^{(m-1)}|\boldsymbol{\xi}^{(m-1)})), \\
S(\boldsymbol{\psi}^{(m-1)}|\boldsymbol{\xi}^{(m-1)}) &= \sum_{i=1}^{n} \sum_k \omega_i^*(\boldsymbol{\xi}^{(m-1)}, a_k) \sum_{j=1}^{m_i} \Big\{ \Delta_{i,j} A_3(u_{i,j,k}(\boldsymbol{\xi}^{(m-1)})) - (1 - \Delta_{i,j}) \Big\} u_{i,j,k}(\boldsymbol{\xi}^{(m-1)}) \\
&\quad \times \left[\frac{\partial \log\{H_\psi(C_{i,j})\}}{\partial \boldsymbol{\psi}} \right]_{\psi = \psi^{(m-1)}},
\end{aligned}
$$

$$
\begin{aligned}
S_\psi(\boldsymbol{\psi}^{(m-1)}|\boldsymbol{\xi}^{(m-1)}) &= \sum_{i=1}^{n}\sum_{k}\omega_i^*(\boldsymbol{\xi}^{(m-1)},a_k)\sum_{j=1}^{m_i}\left\{\Delta_{i,j}A_3(u_{i,j,k}(\boldsymbol{\xi}^{(m-1)}))-(1-\Delta_{i,j})\right\}u_{i,j,k}(\boldsymbol{\xi}^{(m-1)}) \\
&\quad\times\left[\frac{\partial^2\log\{H_\psi(C_{i,j})\}}{\partial\boldsymbol{\psi}\partial\boldsymbol{\psi}^{\mathsf{T}}}\right]_{\psi=\psi^{(m-1)}} \\
&\quad-\sum_{i=1}^{n}\sum_{k}\omega_i^*(\boldsymbol{\xi}^{(m-1)},a_k)\sum_{j=1}^{m_i}\Bigg\{8\Delta_{i,j}A_4(u_{i,j,k}(\boldsymbol{\xi}^{(m-1)}))u_{i,j,k}^2(\boldsymbol{\xi}^{(m-1)})+2\Delta_{i,j} \\
&\quad+2\left(1-\Delta_{i,j}\right)u_{i,j,k}(\boldsymbol{\xi}^{(m-1)})\Bigg\}\left(\left[\frac{\partial\log\{H_\psi(C_{i,j})\}}{\partial\boldsymbol{\psi}}\right]_{\psi=\psi^{(m-1)}}\right)^{\otimes 2}.
\end{aligned}
$$

2.3.2 Choice of the tuning parameter λ

I propose to analyze the data for different choices of λ, and then choose the optimal λ based on one of the classical methods where minimum AIC value is used. Instead of the regular AIC value which is well known for under smoothing, I use the modified AIC defined as $\log\{\mathcal{L}_n(\boldsymbol{\xi})\}+(1+df/n)/\{1-(df+2)/n\}$ (Hurvich et al., 1998). Due to penalized estimation, the degrees of freedom is calculated using the following general formula of Gray (1992)

$$
df=\text{trace}\left[I(\widehat{\boldsymbol{\xi}})\left\{I(\widehat{\boldsymbol{\xi}})+\begin{pmatrix}\mathbf{0} & \mathbf{0}\\ \mathbf{0} & -\lambda\mathcal{P}_{\psi\psi}(\boldsymbol{\psi})\end{pmatrix}\right\}^{-1}\right],
$$

where, $I(\widehat{\boldsymbol{\xi}})=-[\partial^2\log\{\mathcal{L}_n(\boldsymbol{\xi})\}/\partial\boldsymbol{\xi}\partial\boldsymbol{\xi}^{\mathsf{T}}]\big|_{\boldsymbol{\xi}=\widehat{\boldsymbol{\xi}}}$ is the observed information matrix, with $\widehat{\boldsymbol{\xi}}$, the estimator of $\boldsymbol{\xi}$ for a given choice of λ.

2.3.3 The case of non-dependence: $\theta=0$

When all observations are independent, then there is no need to bring in the random frailty. This is a special case of our model with $\theta=0$, i.e., the cluster effect is not present. Till date, there is no easy algorithm to estimate the model parameters of the GOR model with a spline model for H based on such independent CS data. I apply our proposed MM algorithm in this set-up, separating out the estimation of the regression parameters and spline coefficients to develop an efficient algorithm. In absence of clustering, I simplify the notations of the observed data as $\{(C_i,\Delta_i,\boldsymbol{X}_i),i=1,\ldots,n\}$. Under the GOR model, the likelihood is $\mathcal{L}_n(\boldsymbol{\xi})=\prod_{i=1}^n\{1-G_i(\boldsymbol{\xi})\}^{\Delta_i}\{G_i(\boldsymbol{\xi})\}^{1-\Delta_i}$

where, $G_i(\boldsymbol{\xi}) = \{1 + rH_\psi(C_i)\exp(\boldsymbol{\alpha}^\mathsf{T}\boldsymbol{X}_i)\}^{-1/r}$, and $G_i(\boldsymbol{\xi}) = \exp\{-H_\psi(C_i)\exp(\boldsymbol{\alpha}^\mathsf{T}\boldsymbol{X}_i)\}$, for $r > 0$ and $r = 0$, respectively. Then, applying the same inequalities and techniques for the $\theta > 0$ case, I obtain the minorization function $\ell_\dagger(\boldsymbol{\xi}|\boldsymbol{\xi}_0)$ that allows me to separate $\boldsymbol{\alpha}$ and $\boldsymbol{\psi}$. Details of this case are presented in the Appendix A.1.3.

2.3.4 Complexity analysis

I first analyze the computational complexity of the proposed MM algorithm. Let p_α, p_ψ and p_ξ be the dimensions for $\boldsymbol{\alpha}$, $\boldsymbol{\psi}$ and $\boldsymbol{\xi}$, respectively, and $p_\xi = p_\alpha + p_\psi$. In the one-step update using the Newton-Raphson method of MM, $\boldsymbol{\alpha}$ and $\boldsymbol{\psi}$ are updated separately. Before updating $\boldsymbol{\alpha}$, the first step is to calculate $S_\alpha(\boldsymbol{\alpha}^{(m-1)}|\boldsymbol{\xi}^{(m-1)})$ and $S(\boldsymbol{\alpha}^{(m-1)}|\boldsymbol{\xi}^{(m-1)})$. The computational complexity of this step is $O(np_\alpha + np_\alpha^2)$, where n is the sample size (or the number of clusters). Next, the computational complexity of inverting $S_\alpha(\boldsymbol{\alpha}^{(m-1)}|\boldsymbol{\xi}^{(m-1)})$ is $O(p_\alpha^3)$. Therefore, the complexity of one update of $\boldsymbol{\alpha}$ is $O(np_\alpha + np_\alpha^2 + p_\alpha^3)$. Similarly the complexity of one update of $\boldsymbol{\psi}$ is $O(np_\psi + np_\psi^2 + p_\psi^3)$. Hence, the total computational cost for updating $\boldsymbol{\xi}$ is $O(np_\xi + np_\alpha^2 + np_\psi^2 + p_\alpha^3 + p_\psi^3)$.

Now, consider an imaginary scenario where someone develops an EM algorithm for our model. To make things comparable, assume that the M-step of the EM algorithm involves one-step update of the parameters using the Newton-Raphson method, referred to as the gradient EM algorithm. The computational complexity of the E-step of EM will be $O(np_\xi + np_\xi^2)$. Since there will not be any separation of parameters, all components of $\boldsymbol{\xi}$ need to be updated together in the M-step of the EM, and that will require inverting a matrix of order p_ξ, with the complexity of $O(p_\xi^3)$. Therefore, the total computational complexity of updating $\boldsymbol{\xi}$ in the EM will be $O(np_\xi + np_\xi^2 + p_\xi^3)$, which is obviously larger than $O(np_\xi + np_\alpha^2 + np_\psi^2 + p_\alpha^3 + p_\psi^3)$, showing the actual advantage of the MM algorithm over an EM algorithm for the same problem.

2.4 Asymptotic properties

In this section, I present the asymptotic properties of the penalized estimator when $r > 0$. For the case $r = 0$, the asymptotic results are similar, with slight modification of statements due to mild differences in the expression of the survival function.

Define Θ to be a compact subset of $\mathbb{R}^p$, where p is the dimension of $\boldsymbol{\alpha}$, and $\mathcal{H}$ is a class of non-negative and monotonic functions, with zero values at $t = 0$, and continuously differentiable up to order $q \geq 2$ on $[0, T_0]$. Denote $\boldsymbol{\iota} = (\boldsymbol{\alpha}^\mathsf{T}, H)^\mathsf{T}$, where $\boldsymbol{\alpha} \in \Theta$ and $H(\cdot) \in \mathcal{H}$ are the parametric part and the transformation function of the model, respectively. Let $\boldsymbol{\iota}_0 = (\boldsymbol{\alpha}_0^\mathsf{T}, H_0)^\mathsf{T}$ be the true value of $\boldsymbol{\iota}$. The distance between two elements in $\mathcal{H}$ is measured by the Lebesgue L_2-norm. More precisely, for any $H_1, H_2 \in \mathcal{H}$, define $\|H_1 - H_2\|_2^2 = \int_0^{T_0} \{H_1(t) - H_2(t)\}^2 \mathrm{d}t$, and for any $\boldsymbol{\iota}_1 = (\boldsymbol{\alpha}_1{}^\mathsf{T}, H_1)^\mathsf{T}$ and $\boldsymbol{\iota}_2 = (\boldsymbol{\alpha}_2{}^\mathsf{T}, H_2)^\mathsf{T}$ in the space of $\Xi = \Theta \times \mathcal{H}$, define an L_2-metric as follows: $\mathrm{dist}(\boldsymbol{\iota}_1, \boldsymbol{\iota}_2) = \|\boldsymbol{\iota}_1 - \boldsymbol{\iota}_2\|_\Xi = \big(\|\boldsymbol{\alpha}_1 - \boldsymbol{\alpha}_2\|^2 + \|H_1 - H_2\|_2^2\big)^{1/2}$. The order of spline functions I use to approximate $H(\cdot)$ is chosen to satisfy $d \geq q$. Now, I state the results succinctly, however, seven regularity conditions, and the detailed proofs of lemmas and theorems are given in the Appendix A.2.

Lemma 2.2. *Under conditions (C1)–(C4), the parameter component $\boldsymbol{\alpha}$ and the transformation function H are identifiable.*

The following theorem establishes the consistency of the penalized ML estimator $\widehat{\boldsymbol{\iota}}_n = (\widehat{\boldsymbol{\alpha}}_n^\mathsf{T}, \widehat{H}_n)^\mathsf{T}$ given in (2.4) for a general smoothness order q.

Theorem 2.2. *Suppose the regularity conditions (C1)–(C6) hold, $L = O(n^{1/(2q+1)})$, and the tuning parameter λ satisfies $\lambda \asymp n^{-2q/(2q+1)}$. Then, the penalized ML estimator $\widehat{\boldsymbol{\iota}}_n = (\widehat{\boldsymbol{\alpha}}_n^\mathsf{T}, \widehat{H}_n)^\mathsf{T}$ satisfies*

$$\mathrm{dist}(\widehat{\boldsymbol{\iota}}_n, \boldsymbol{\iota}_0) = O_p\big(n^{-q/(2q+1)}\big) \tag{2.13}$$

Theorem 2.2 implies that the penalized estimator achieves the optimal convergence rate (Stone, 1982) in the nonparametric regression setting. Furthermore, when $q = 2$, that is when the transformation function is second-order differentiable, the proposed penalized ML estimator achieves the convergence rate bounded by $O_p(n^{-2/5})$, faster than the convergence rate $O_p(n^{-1/3})$ shown in Huang and Rossini (1997), which considers a nonparametric modeling for current status data. In

the next theorem, I present the asymptotic normality and efficiency for the parametric part of the penalized ML estimator.

Theorem 2.3. *Suppose that all the assumptions given in Theorem 2.2 hold, and the regularity condition (C7') is satisfied. Then,*

$$n^{1/2}\big(\widehat{\boldsymbol{\alpha}}_n - \boldsymbol{\alpha}_0\big) \to \mathcal{N}\big(\mathbf{0}, I^{-1}(\boldsymbol{\alpha}_0)\big) \text{ in distribution,} \tag{2.14}$$

where $I(\boldsymbol{\alpha}_0)$ is the efficient information of $\boldsymbol{\alpha}$ with expected value at $\boldsymbol{\iota}_0$ for the likelihood, and assumed non-singular.

Theorem 2.3 implies that, although the estimators of the transformation function converge at a rate slower than $n^{1/2}$ (as shown in Theorem 2.2), the regularized estimators of the regression parameters converge to the true one at the usual $\sqrt{n}$ rate. Moreover, the estimators from the regularized complete and observed likelihoods are both able to achieve the corresponding semi-parametric efficiency bounds. It is worth noting that I am able to handle a large number of inner knots points under the roughness penalization, and the theoretical results are valid for the function space $\mathcal{H}$, with the distance regularized by the tuning parameter λ.

Since obtaining an analytical form of the efficient information is difficult, I invert the observed information matrix $I(\widehat{\boldsymbol{\xi}})$, and use the submatrix corresponding to the finite dimensional parameters as the asymptotic variance-covariance for the finite dimensional parameters. In calculating the observed information matrix at the estimate, I use numerical differentiation.

2.5 Simulation studies

I simulated cohorts of two different sizes (the number of subjects), $n = 300$ and 1000. For each subject, I simulated Z_i from uniform$(-1, 1)$ distribution. Mimicking the GAAD data, for each subject, I first simulated the cluster size m_i from Poisson(5.47), that is truncated below 1 and above 8. Next, I simulated $X_{i,j}$ from uniform$(-1, 1)$, $j = 1, \ldots, m_i$, and b_i from $\mathrm{Normal}(0, 1)$, $i = 1, \ldots, n$. Then, I simulated unobserved event time $T_{i,j}$ from the following model $\log\{1 - \text{pr}(T_{i,j} \le t|b_i, X_{i,j}, Z_i)\} = -(1/r)\log\{1 + rH(t)\exp(\beta X_{i,j} + \gamma Z_i + \theta b_i)\}$ and $\log\{1 - \text{pr}(T_{i,j} \le$

$t|b_i, X_{i,j}, Z_i)\} = -H(t)\exp(\beta X_{i,j} + \gamma Z_i + \theta b_i)$ for $r > 0$ and $r = 0$, respectively. I set $H(t) = \log(1 + t) + t^{3/2}$, and simulated data for $r = 0, 1, 2$. The inspection time $C_{i,j}$ was simulated according to uniform$(0, C_{\text{upper}})$, where C_{upper} represents the 85% quantiles of $T_{i,j}$. Next, I set $\Delta_{i,j}$ to 1 or 0, depending on $T_{i,j} \leq C_{i,j}$, or not. This resulted in 40% and 30% observations with $\Delta_{i,j} = 0$ for $r = 0$ and $r = 2$, respectively.

Under each scenario, I simulated 500 datasets and fitted the respective data generating model, where the index r was assumed to be known. For the nonparametric H, I transformed observed $C_{i,j}$ into $[0, 1]$, used two equally-spaced inner knots at 0.33 and 0.66, and employed I-splines of degree 2. This resulted in five basis functions. Because of the small number of basis functions, I did not consider the roughness penalty approach to estimate H. For each model parameter, I report the relative mean bias (RB), relative median bias ($\widetilde{\text{RB}}$), empirical standard deviation (SD), median of the estimated standard error (SE), and the 95% coverage probability (CP) based on Wald's confidence interval. The results corresponding to $\theta = 0.5, 1$ and 2 are given in Table 2.1. When $n = 300$, both RB and $\widetilde{\text{RB}}$ for almost all parameters are at most 3% in absolute value. Overall, the bias and SD decrease as n increases from 300 to 1000. There is a reasonable agreement between the empirical standard deviation and the estimated standard error. The CPs are reasonably close to the nominal level, 0.95.

I conducted another simulation study closely resembling the real dataset. The distribution of X, Z, H were the same as before, but set $\beta = 2$, $\gamma = -2$, $\theta = 3.5$, $r = 2$ and $C_{i,j}$s were simulated from the uniform distribution, such that the percentage of $\Delta_{i,j} = 0$ was 75%. The corresponding results are presented in Table 2.2. As expected, due to high percentage of $\Delta_{i,j} = 0$, the bias of the parameter estimators is slightly larger than that in Table 2.1. However, as the sample size increases, the bias and SD decrease. The CPs are also close to the nominal level, 0.95. In all computations, a tolerance of $\epsilon_t = 10^{-7}$ was used. The average computation time for a single dataset were 1.55 mins and 5.05 mins for $n = 300$ and 1000, respectively, on an Intel(R) Xeon(R) CPU E5-2680 v4 @ 2.40GHz machine.

Table 2.1: Results of the simulation study for $\beta = -1$, $\gamma = -1$. Here RB, $\widetilde{\mathrm{RB}}$, SD, SE, CP denote the relative mean bias, the relative median bias, the standard deviation, the median of estimated standard error, and the 95% coverage probability, respectively. PAR: Parameter

	PAR	$\theta = 2$					$\theta = 1$					$\theta = 0.5$				
		RB	$\widetilde{\mathrm{RB}}$	SD	SE	CP	RB	$\widetilde{\mathrm{RB}}$	SD	SE	CP	RB	$\widetilde{\mathrm{RB}}$	SD	SE	CP
$n = 300$																
	r=0															
	β	0.01	0.01	0.14	0.13	0.93	0.01	0.01	0.10	0.10	0.96	0.01	0.00	0.09	0.09	0.95
	γ	-0.01	0.00	0.26	0.25	0.93	0.01	0.01	0.15	0.14	0.93	0.01	0.00	0.11	0.10	0.93
	θ	-0.01	-0.01	0.27	0.20	0.96	0.00	0.00	0.10	0.10	0.97	-0.01	0.00	0.10	0.08	0.96
	r=1															
	β	0.00	-0.01	0.15	0.15	0.94	0.00	0.00	0.13	0.13	0.96	0.00	0.00	0.12	0.12	0.96
	γ	-0.01	-0.01	0.26	0.26	0.95	0.00	0.01	0.18	0.16	0.94	0.00	0.00	0.15	0.13	0.92
	θ	-0.02	-0.02	0.19	0.18	0.95	-0.01	-0.01	0.12	0.12	0.95	-0.02	0.00	0.14	0.14	0.95
	r=2															
	β	-0.03	-0.03	0.21	0.20	0.95	0.01	0.00	0.19	0.18	0.96	0.01	0.01	0.18	0.17	0.95
	γ	-0.02	0.00	0.29	0.28	0.94	0.01	0.01	0.24	0.20	0.92	0.03	0.01	0.21	0.20	0.93
	θ	-0.03	-0.04	0.22	0.20	0.96	0.00	-0.01	0.24	0.18	0.95	0.08	0.06	0.26	0.19	0.91
$n = 1000$																
	r=0															
	β	0.00	0.00	0.07	0.07	0.96	0.00	0.00	0.06	0.06	0.93	0.00	0.00	0.05	0.05	0.95
	γ	0.00	-0.01	0.13	0.14	0.96	0.00	0.00	0.08	0.08	0.94	0.00	0.00	0.06	0.06	0.93
	θ	-0.01	-0.01	0.12	0.11	0.94	-0.01	0.00	0.09	0.06	0.94	-0.01	-0.01	0.05	0.04	0.95
	r=1															
	β	-0.01	-0.01	0.08	0.08	0.96	0.00	-0.01	0.07	0.07	0.95	0.00	-0.01	0.06	0.06	0.97
	γ	-0.01	-0.02	0.14	0.14	0.95	0.00	0.00	0.09	0.09	0.94	0.00	0.00	0.07	0.07	0.95
	θ	-0.03	-0.03	0.12	0.10	0.92	-0.02	-0.02	0.08	0.07	0.91	-0.03	-0.03	0.08	0.07	0.95
	r=2															
	β	-0.01	-0.01	0.10	0.11	0.96	0.00	0.00	0.09	0.10	0.97	0.01	0.00	0.10	0.10	0.96
	γ	-0.02	-0.01	0.15	0.15	0.96	0.01	0.01	0.11	0.11	0.96	0.01	0.01	0.11	0.10	0.95
	θ	-0.02	-0.02	0.10	0.11	0.91	-0.01	-0.01	0.12	0.10	0.96	0.06	0.05	0.16	0.10	0.93

Table 2.2: Results of the simulation study for $\theta = 3.5$, $\beta = 2$, $\gamma = -2$ with $r = 2$. Here RB, $\widetilde{\text{RB}}$, SD, SE, CP denote the relative mean bias, the relative median bias, the standard deviation, the median of estimated standard error, and the 95% coverage probability, respectively. PAR: Parameter

PAR	$n = 300$					$n = 1000$				
	RB	$\widetilde{\text{RB}}$	SD	SE	CP	RB	$\widetilde{\text{RB}}$	SD	SE	CP
β	-0.04	-0.05	0.22	0.24	0.94	-0.02	-0.02	0.12	0.13	0.93
γ	-0.06	-0.06	0.41	0.43	0.94	-0.05	-0.04	0.21	0.23	0.93
θ	-0.06	-0.06	0.22	0.32	0.96	-0.04	-0.04	0.13	0.17	0.92

2.6 Application: GAAD Data

I now illustrate our methodology via application to the Gullah-speaking African American Diabetic (GAAD) data to investigate the association between the time-to-onset of moderate to severe PD of the molars and its prognostic factors. Molars are primarily responsible for mastication and breaking down of food before swallowing. Also, multi-rooted molars affected by PD consequently develop furcation involvement, leading to less favorable response to periodontal therapy, compared to single-rooted teeth (such as canines), or molars without furcation (Wang et al., 1994). Hence, proper risk assessment of the molars in terms of their explanatory variables is necessary to develop targeted therapies that can prolong the lifespan of the tooth. However, due to the cross-sectional nature of the study design, oral hygienists in this study could only assess CAL, the most important biomarker of PD severity, during the clinic visits (also, the inspection time), with no information on when the landmark event actually occurred.

Although the first molar is one of the earliest to erupt, and are lost due to decay or fracture in adult dentition, the (exact) time of eruption of adult molars for a random subject remains vastly unknown. It also varies considerably with respect to tooth-types and locations, i.e., eruption times varies between first and second molars, and also on their jaw locations (mandibular, or maxil-

lary). Hence, in absence of the exact eruption time of the molars (in the GAAD data), I consider the CS inspection time C as the difference between the time of clinic visit and the permanent teeth eruption chart, available at `https://www.mouthhealthy.org/en/az-topics/e/eruption-charts`. Ignoring the third molar (by convention), a subject can have a maximum of 8 molars combining all teeth quadrants, and I consider subjects who have at least one molar, resulting in 234 patients, where 177 are females. Besides gender (1= female, 0= male), the other subject-level covariates $\boldsymbol{Z}$ include smoking status (1 = smoker, 0 = never smoker), and HbA1c status (1 = uncontrolled, 0 = controlled). The jaw indicator (1 = tooth in upper jaw, 0 = tooth in lower jaw) is the only tooth-level covariate X. In addition, the indicator Δ was also recorded, taking values 0 or 1, depending on whether $T > C$, or not, respectively. Instances with $\Delta = 0$ are considered as right-censored. Majority of the teeth did not experience the event of interest by the inspection time, leading to a high percentage of censoring (about 75%).

Figure 2.1 shows the nonparametric (Turnbull) empirical survival curves (Turnbull, 1976) for the time to the landmark event for four groups, combining gender and HbA1c. It shows that overall, the females have a higher survival probability than males, and within a gender, the low HbA1c (controlled) group experiences higher survival than the high HbA1c group, as expected, across the full adult age spectrum. However, all four curves level off at the highest age ranges.

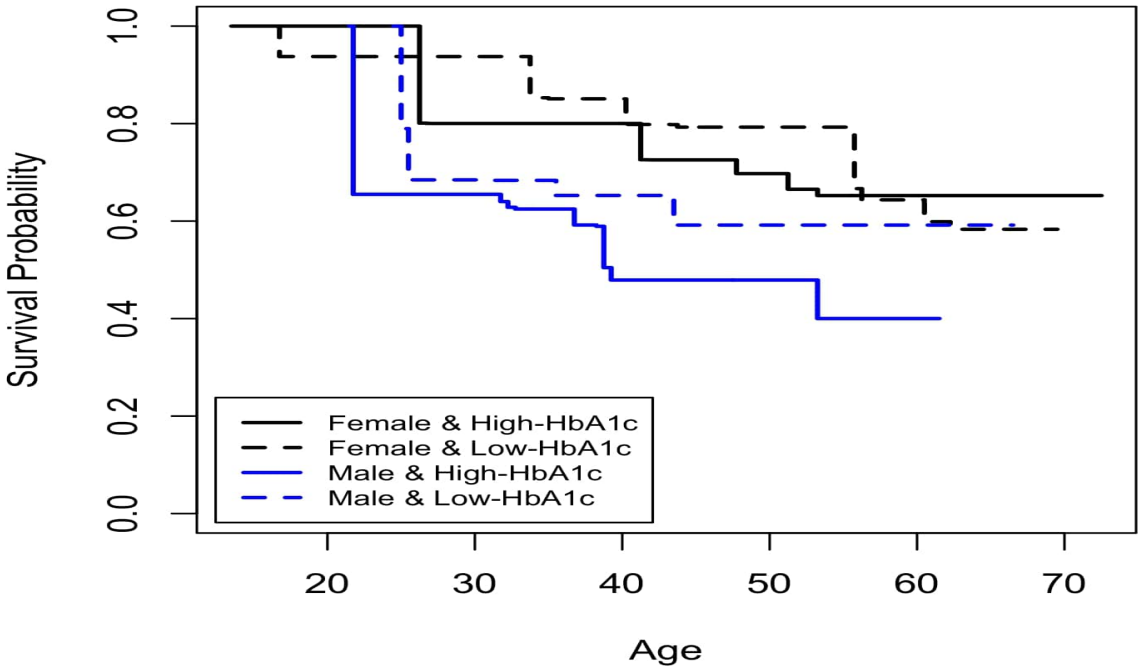

Figure 2.1: Turnbull's nonparametric estimator of the survival function of the time-to-landmark event for the GAAD data, classified by gender and glycemic status.

2.6.1 Model fitting and results

I fitted two models to the dataset, the proposed model (M1) and the model without the random effects (M2, i.e., no clustering, $\theta = 0$). In M1, the random effect b was assumed to follow normal(0, 1), and H was modeled via I-splines of degree 2 with two equispaced knots. In order to determine the best choice of r, I choose an array of r, starting from 0 to 3 with an increments of 0.1, and fit the corresponding models. I observe that $r = 1.9$ yields the maximum log-likelihood. Subsequently, I also fitted the same model ($r = 1.9$) without the frailty term (M2), and the corresponding log-likelihood value was much smaller than the log-likelihood from M1 (-409 versus -702).

The results corresponding to M1 & M2 are summarized in Table 2.3. Under M1, a tooth in the upper jaw experiences significantly higher probability of the event of interest. Compared to males, females have less probability of experiencing the event. Other covariates are not statistically

significant. The coefficient θ has an interpretation, along the lines of the intraclass correlation coefficient. The GOR model can be represented by the linear transformation model $H^*(T_{i,j}) = -X_{i,j}\beta - \boldsymbol{Z}_i{}^{\mathsf{T}}\boldsymbol{\gamma} - \theta b_i + \epsilon_{i,j}$, where $H^*(T_{i,j})$ is a monotonic transformation function, and $\epsilon_{i,j}$ has the survival function $\text{pr}(\epsilon_{i,j} > u) = 1/\{1 + r\exp(u)\}^{1/r}$, and H^* and H of (2.1) are related via $H(t) = \exp\{H^*(t)\}$. Thus, the intraclass correlation (ICC) among the time-to-events within a cluster, adjusted for covariate effects is the $\text{ICC} = \text{var}(\theta b_i)/\{\text{var}(\theta b_i) + \text{var}(\epsilon_{i,j})\}$. For $r = 1.9$, $\text{var}(\epsilon_{i,j}) \approx 6.17$; so the estimated ICC is $3.38^2/(3.38^2 + 6.17) = 0.65$, indicative of a relatively good intraclass correlation.

Table 2.3: GAAD data analysis. In panels 1 and 2, I fit the GOR model with frailty to the full data, and after removing influential subjects, respectively. In panel 3, I fit the GOR model to the full data with the frailty, a moderate number of knot points and a roughness penalty for the nonparametric term. In panel 4, I fit the GOR model without the frailty term and with the same number of knots as of M1. In all four panels, $r = 1.9$. Est: Estimate, SE: Standard error, PV: p-value

	M1			M1*			M1**			M2		
Variable	Est	SE	PV	Est	SE	PV	Est	SE	PV	Est	SE	PV
Jaw	2.13	0.37	0.00	2.47	0.40	0.00	2.15	0.35	0.00	1.15	0.19	0.00
Gender	−2.44	0.70	0.00	−2.46	0.75	0.00	−2.30	0.68	0.01	−1.10	0.22	0.00
Smoking	1.13	0.65	0.08	1.13	0.70	0.10	1.04	0.64	0.10	0.80	0.20	0.00
Hba1c	1.11	0.59	0.06	1.19	0.64	0.06	1.05	0.59	0.07	0.55	0.20	0.00
θ	3.35	0.40	0.00	3.56	0.46	0.00	3.35	0.41	0.00			

2.6.2 Diagnostics

In order to detect if there is any notable local sensitivity, I compared $\widehat{\boldsymbol{\xi}}$ and $\widehat{\boldsymbol{\xi}}_{-j}$, two estimators of the generic parameter vector $\boldsymbol{\xi}$, where $\widehat{\boldsymbol{\xi}}$ denotes the estimator based on the entire dataset, while $\widehat{\boldsymbol{\xi}}_{-j}$ is the estimator of $\boldsymbol{\xi}$ based on the dataset after deleting data from the jth subject. Here, the jth subject is considered influential, if $\left\|\widehat{\boldsymbol{\xi}} - \widehat{\boldsymbol{\xi}}_{-j}\right\|$ is large, compared to the rest of the subjects. The naive approach of computing these differences require fitting the model n times, where n is the number of subjects or clusters. To avoid this lengthy computation, I adopted Cook (1986)'s general

approach in our set-up. Define $l(\boldsymbol{\xi}|\boldsymbol{\kappa}) = \sum_{i=1}^{n} \kappa_i \log\{\mathcal{L}_i(\boldsymbol{\xi})\}$, where $\boldsymbol{\kappa} = (\kappa_1, \ldots, \kappa_n)^{\mathsf{T}}$, and each component of $\boldsymbol{\kappa}$ lies in $[0, 1]$. Next, define $LD(\boldsymbol{\kappa}) = 2[l(\widehat{\boldsymbol{\xi}}) - l\{\widehat{\boldsymbol{\xi}}(\boldsymbol{\kappa})\}]$, where $l(\widehat{\boldsymbol{\xi}}) = \log\{\mathcal{L}_n(\widehat{\boldsymbol{\xi}})\}$, and $\widehat{\boldsymbol{\xi}}(\boldsymbol{\kappa})$ maximizes $l(\boldsymbol{\xi}|\boldsymbol{\kappa})$. Using the Taylor-series expansion, $LD(\boldsymbol{\kappa})$ can be approximated around $\boldsymbol{\kappa}_0 = (1, \ldots, 1)^{\mathsf{T}}$ by

$$\begin{aligned} LD(\boldsymbol{\kappa}) = 2[l(\widehat{\boldsymbol{\xi}}) - l\{\widehat{\boldsymbol{\xi}}(\boldsymbol{\kappa})\}] &\approx \{\widehat{\boldsymbol{\xi}} - \widehat{\boldsymbol{\xi}}(\boldsymbol{\kappa})\}^{\mathsf{T}} \left\{ -\frac{\partial^2 l(\boldsymbol{\xi})}{\partial \boldsymbol{\xi} \partial \boldsymbol{\xi}^{\mathsf{T}}} \right\}_{\boldsymbol{\xi}=\widehat{\boldsymbol{\xi}}} \{\widehat{\boldsymbol{\xi}} - \widehat{\boldsymbol{\xi}}(\boldsymbol{\kappa})\} \\ &\approx (\boldsymbol{\kappa}_0 - \boldsymbol{\kappa})^{\mathsf{T}} \bigtriangledown^{\mathsf{T}} \{\Sigma(\widehat{\boldsymbol{\xi}})\}^{-1} \bigtriangledown (\boldsymbol{\kappa}_0 - \boldsymbol{\kappa}), \end{aligned}$$

where $\bigtriangledown = \{\partial^2 l(\boldsymbol{\xi}|\boldsymbol{\kappa})/\partial \boldsymbol{\xi} \partial \boldsymbol{\kappa}^{\mathsf{T}}\}_{\boldsymbol{\xi}=\widehat{\boldsymbol{\xi}}, \boldsymbol{\kappa}=\boldsymbol{\kappa}_0}$, and $\Sigma(\widehat{\boldsymbol{\xi}}) = \{-\partial^2 l(\boldsymbol{\xi})/\partial \boldsymbol{\xi} \partial \boldsymbol{\xi}^{\mathsf{T}}\}_{\boldsymbol{\xi}=\widehat{\boldsymbol{\xi}}}$. I can re-write $LD(\boldsymbol{\kappa}) = \boldsymbol{d}^{\mathsf{T}} \bigtriangledown^{\mathsf{T}} \{\Sigma(\widehat{\boldsymbol{\xi}})\}^{-1} \bigtriangledown \boldsymbol{d}$, where $\boldsymbol{d} = (\boldsymbol{\kappa} - \boldsymbol{\kappa}_0)$ is a n-dimensional vector, and $\boldsymbol{d} \leq 1$. Suppose that $\boldsymbol{d}^{\mathsf{T}} \bigtriangledown^{\mathsf{T}} \{\Sigma(\widehat{\boldsymbol{\xi}})\}^{-1} \bigtriangledown \boldsymbol{d}$ is maximized at $\boldsymbol{d}_{\max}$, and the corresponding $\boldsymbol{\kappa} = \boldsymbol{\kappa}_0 \pm \boldsymbol{d}_{\max}$ maximizes $LD(\boldsymbol{\kappa})$. Next, I compute the statistic $\sigma_{\max} = \boldsymbol{d}_{\max}{}^{\mathsf{T}} \bigtriangledown^{\mathsf{T}} \{\Sigma(\widehat{\boldsymbol{\xi}})\}^{-1} \bigtriangledown \boldsymbol{d}_{\max}$, and obtain $\sigma_{\max} = 1.45$ for our dataset. Any value of $\sigma_{\max} > 1$ may signal sensitivity in the analyses results (Cook, 1986). In Figure 2.2, I plot the absolute value of each component of the $\boldsymbol{d}_{\max}$ vector against the indices of the observations to detect the local influential observations. The figure reveals that the 24th, 65th and 87th observations (indicated by stars) are the three largest influential observations. After deleting these three subjects, I re-analyzed the data using the proposed model M1, and reported the corresponding estimates (see M1* in Table 2.3). Although the estimates are slightly changed, the removal of the influential points does not change the statistical significance of the covariates.

I further reanalyzed the data using M1, where, H was modeled using cubic I-splines with 5 inner knot points. Here, I estimated the spline coefficients incorporating the roughness penalty for the H function. I varied the tuning parameter λ from 2^{-20} to 2^{10}, where the consecutive tuning parameters were multiples of 2. The penalty term $\mathcal{P}(\boldsymbol{\xi})$ for $q = 2$ was calculated using the function `bsplinepen` available in `R` package `fda`. The estimates corresponding to the minimum AIC value are reported as M1** in Table 2.3. The results are similar to those in M1, and statistical significance of the covariates remain unchanged at the 5% level.

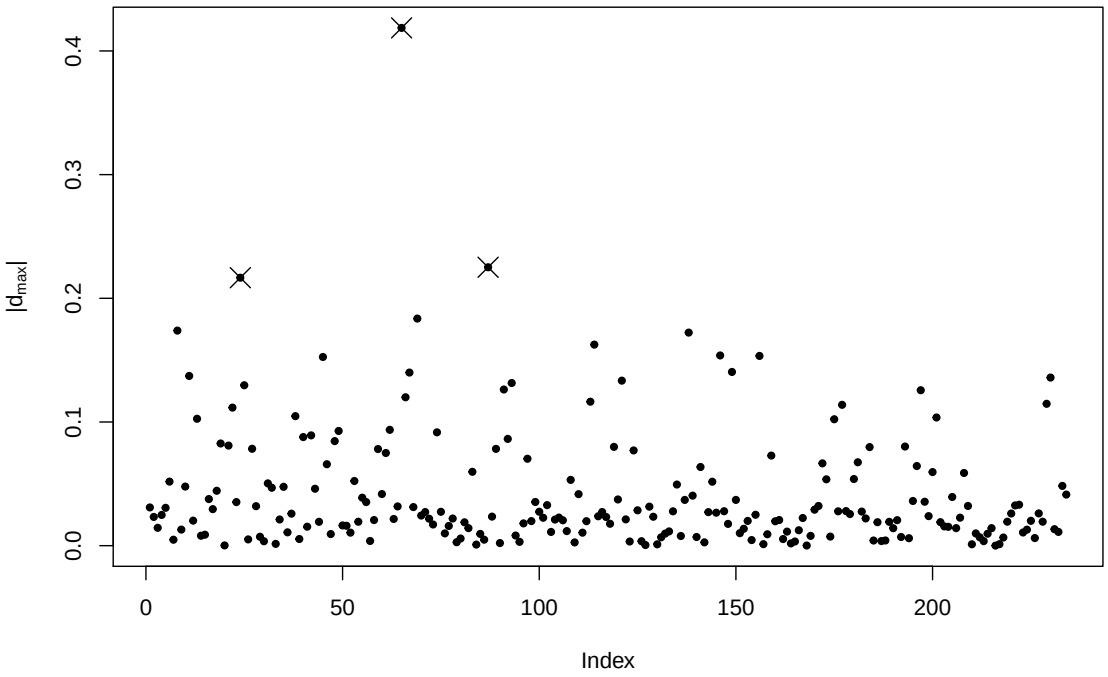

Figure 2.2: Plot of elements of vector $d_{\max}$ against the subject index.

2.7 Conclusion

In summary, this chapter provides a modestly complete solution of analyzing clustered current status data, one that exhibits the most severe interval-censoring patterns, through (a) development of a nice computational algorithm, and (b) thorough asymptotic justifications. The GOR model fitting is attractive, as it encompasses a large class of models. Our MM algorithm works well under both clustered and non-clustered settings, and is expected to be useful for developing similar algorithms for other models.

3. EFFICIENT ESTIMATION OF THE ADDITIVE RISKS MODEL FOR INTERVAL-CENSORED DATA

3.1 Background and literature review

Interval-censoring, which occurs when the failure time is only known to lie in an interval instead of being observed exactly, abounds in finance, epidemiology and longitudinal study. There are two main types of interval-censored data: case-I and case-II interval-censored data. Case-I interval-censored data, also called current status data, is not the focus of this chapter. In this chapter I are concerned about the case-II interval censored data that are mixture of left, interval and right censored time to occurrence of an event. The aim of this chapter is to present an efficient algorithm of estimating the maximum likelihood estimates of the additive risks model for the case-II interval censored data.

In the additive risks model the hazard function is

$$h(t|\boldsymbol{X}(t)) = \lambda(t) + \boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{X}(t), \tag{3.1}$$

where $\boldsymbol{X}(t)$ denotes a vector of possibly time-dependent covariate, $\boldsymbol{\beta}$ is the corresponding regression parameter, and $\lambda(t)$ is the baseline hazard function. In this model, the effect of the covariate can be measured via the difference in the hazard function. Further details on the usefulness of this model can be found in Huffer and McKeague (1991). Lin and Ying (1994) used the additive risks model to analyze right censored data. For the case-II interval censored data, Zeng et al. (2006) first proposed the maximum likelihood method to estimate both baseline hazard function and regression parameters of the model. Wang et al. (2010) proposed a martingale-based estimation procedure, and they focused only on the estimation of the regression parameters but not the baseline hazard function which is also an important component to study the event of interest. Martinussen and Scheike (2002) and Wang et al. (2020) proposed to use a sieve approach to model $\lambda(t)$ that requires an appropriate choice of the sieve parameter space and the number of knots.

In the maximum likelihood approach of fitting the additive model (3.1) to the interval-censored data, the baseline survival function was modeled as a nonparametric step function with jump at the observed inspection time points. The computation of the maximum likelihood estimates through the direct maximization of the observed data likelihood function is problematic due to a large number of parameters. Note that although the regression parameter is of finite dimension, the baseline hazard function contributes a large number of parameters that tends to increase with the sample size when the inspection time is continuous (Zeng et al., 2006). To circumvent this computational difficulty of the high-dimensional maximization, I develop a novel Minorize-Maximization (MM) algorithm (Hunter and Lange, 2004; Wu et al., 2010a) to obtain the maximum likelihood estimates. The proposed method can handle both time-independent and time-dependent covariates. By applying this technique, the original problem of high-dimensional optimization reduces to a simple Newton-Raphson update of the parameters. Moreover, in each step of the Newton-Raphson method, I do not need to invert any high dimensional matrix. All these are possible with a clever choice of the surrogate function, and details of this choice are discussed in the next section. Extensive simulation studies confirm that the proposed MM algorithm can estimate the parameters very well and the computational time is much faster than the method of direct maximization.

The remainder of the chapter is organized as follows. Section 3.2 contains notations and assumptions. The MM algorithm along with the complexity analysis are presented in Section 3.3.1. Simulation results are presented in Section 3.4. The proposed method is applied to analyze the breast cosmesis data and the details are in Section 3.5. Finally, concluding remarks are given in Section 3.6.

3.2 Notations and assumptions

Suppose that T_i denotes the time-to-event for the ith subject. Suppose that I have interval censored data $\{L_i, R_i, \boldsymbol{X}_i, \Delta_{L,i}, \Delta_{I,i}, \Delta_{R,i}\}$, $i = 1, \ldots, n$ on n independent subjects, where $\Delta_{L,i}$, $\Delta_{I,i}$ and $\Delta_{R,i}$ denote the left censoring, interval censoring and right censoring indicators, respectively. If T_i is left censored, then T_i falls in $(0, L_i]$ and $\Delta_{L,i} = 1$ while $\Delta_{I,i} = \Delta_{R,i} = 0$. If T_i is interval censored, then T_i falls in $(L_i, R_i]$ and $\Delta_{L,i} = \Delta_{R,i} = 0$ while $\Delta_{I,i} = 1$. Finally, if

T_i is right censored then T_i falls in (R_i, ∞) and $\Delta_{L,i} = \Delta_{I,i} = 0$ while $\Delta_{R,i} = 1$. As a space holder I can set R_i to any number larger than L_i for left censored time-to-event, and L_i to any number smaller than R_i for right censored time-to-event. Here $\boldsymbol{X}_i$ denotes a $p \times 1$ vector of time-dependent covariates. The hazard rate function is given in (3.1) while the cumulative hazard is $H(t; \boldsymbol{X}) = \Lambda(t) + \beta^T \boldsymbol{Z}_x(t)$, where $\Lambda(t) = \int_0^t \lambda(s)ds$ and $\boldsymbol{Z}_x(t) = \int_0^t \boldsymbol{X}(s)ds$. When the covariate is time independent, $\boldsymbol{Z}_x(t) = \int_0^t \boldsymbol{X}(s)ds = \boldsymbol{X}t$. Given the covariates, the survival probability is

$$S(t; \boldsymbol{X}) = \exp[-\{\Lambda(t) + \boldsymbol{\beta}^T \boldsymbol{Z}_x(t)\}].$$

For nonparametric maximum likelihood estimation, assume that $\Lambda(t)$ is a step function with jump λ_k at t_k $(k = 0, \ldots, m)$, i.e., $\Lambda(t) = \sum_{k:t_k \le t} \lambda_k$, where $t_1 < \cdots < t_m$, denote the unique inspection time points. For our convenience, I take $t_0 = 0$. Let $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_m)^{\mathsf{T}}$, then the observed likelihood and the log-likelihood functions are

$$\mathcal{L}(\boldsymbol{\lambda}, \boldsymbol{\beta}) = \prod_{i=1}^{n} \{1 - S(L_i; \boldsymbol{X}_i)\}^{\Delta_{L,i}} \{S(L_i; \boldsymbol{X}_i) - S(R_i; \boldsymbol{X}_i)\}^{\Delta_{I,i}} \{S(R_i; \boldsymbol{X}_i)\}^{\Delta_{R,i}},$$

and

$$\begin{aligned}
\ell(\boldsymbol{\lambda}, \boldsymbol{\beta}) &= \sum_{i=1}^{n} \Big[\Delta_{L,i} \log\{1 - S(L_i; \boldsymbol{X}_i)\} + \Delta_{I,i} \log\{S(L_i; \boldsymbol{X}_i) - S(R_i; \boldsymbol{X}_i)\} + \Delta_{R,i} \log\{S(R_i; \boldsymbol{X}_i)\} \Big] \\
&= \sum_{i=1}^{n} \Big[\Delta_{L,i} \log\{1 - S(L_i; \boldsymbol{X}_i)\} + \Delta_{I,i} \log\{S(L_i; \boldsymbol{X}_i)\} + \Delta_{I,i} \log\{1 - S^{-1}(L_i; \boldsymbol{X}_i) S(R_i; \boldsymbol{X}_i)\} \\
&\quad + \Delta_{R,i} \log\{S(R_i, \boldsymbol{X}_i)\} \Big] \\
&= \ell_1(\boldsymbol{\lambda}, \boldsymbol{\beta}) + \ell_2(\boldsymbol{\lambda}, \boldsymbol{\beta}) + \ell_3(\boldsymbol{\lambda}, \boldsymbol{\beta}) + \ell_4(\boldsymbol{\lambda}, \boldsymbol{\beta}),
\end{aligned}$$

where

$$
\begin{aligned}
\ell_1(\boldsymbol{\lambda},\boldsymbol{\beta}) &= \sum_{i=1}^{n}\Delta_{L,i}\log\{1-S(L_i|\boldsymbol{X}_i)\} = \sum_{i=1}^{n}\Delta_{L,i}\log[1-\exp\{-\sum_{k:t_k\le L_i}\lambda_k-\boldsymbol{\beta}^T\boldsymbol{Z}_{x_i}(L_i)\}],\\
\ell_2(\boldsymbol{\lambda},\boldsymbol{\beta}) &= \sum_{i=1}^{n}\Delta_{I,i}\log\{S(L_i|\boldsymbol{X}_i)\} = -\sum_{i=1}^{n}\Delta_{I,i}\left\{\sum_{k:t_k\le L_i}\lambda_k+\boldsymbol{\beta}^T\boldsymbol{Z}_{x_i}(L_i)\right\},\\
\ell_3(\boldsymbol{\lambda},\boldsymbol{\beta}) &= \sum_{i=1}^{n}\Delta_{I,i}\log\{1-S^{-1}(L_i|\boldsymbol{X}_i)S(R_i|\boldsymbol{X}_i)\}\\
&= \sum_{i=1}^{n}\Delta_{I,i}\log\left(1-\exp\left[-\sum_{k:L_i<t_k\le R_i}\lambda_k-\boldsymbol{\beta}^T\{\boldsymbol{Z}_{x_i}(R_i)-\boldsymbol{Z}_{x_i}(L_i)\}\right]\right),\\
\ell_4(\boldsymbol{\lambda},\boldsymbol{\beta}) &= \sum_{i=1}^{n}\Delta_{R,i}\log\{S(R_i|\boldsymbol{X}_i)\} = -\sum_{i=1}^{n}\Delta_{R,i}\left\{\sum_{k:t_k\le R_i}\lambda_k+\boldsymbol{\beta}^T\boldsymbol{Z}_{x_i}(R_i)\right\}.
\end{aligned}
$$

In the next section I develop an efficient optimization technique aided by the MM algorithm to estimate $\boldsymbol{\lambda}$ and $\boldsymbol{\beta}$.

3.3 Estimation methodology

3.3.1 MM algorithm

For developing an MM algorithm I need to find a suitable minorization function that determines the usefulness of the algorithm. To develop such a minorization function I use a result from Theorem 2.1 along with some standard mathematical inequalities. Define $\boldsymbol{\lambda}_0 = (\lambda_{10},\ldots,\lambda_{m0})^{\mathsf{T}}$ and $u_0(L_i,\boldsymbol{X}_i) = \sum_{k:t_k\le L_i}\lambda_{k0}+\boldsymbol{\beta}_0^T\boldsymbol{Z}_{x_i}(L_i)$, $u_0(R_i,\boldsymbol{X}_i) = \sum_{k:t_k\le R_i}\lambda_{k0}+\boldsymbol{\beta}_0^T\boldsymbol{Z}_{x_i}(R_i)$ and $u_0(L_i,R_i,\boldsymbol{X}_i) = \sum_{k:L_i<t_k\le R_i}\lambda_{k0}+\boldsymbol{\beta}_0^T\{\boldsymbol{Z}_{x_i}(R_i)-\boldsymbol{Z}_{x_i}(L_i)\}$. Now, I present the main result in the following theorem and its proof is given in the appendix.

Theorem 3.1. *The minorization function for $\ell(\boldsymbol{\lambda},\boldsymbol{\beta})$ is $\ell_\dagger(\boldsymbol{\lambda},\boldsymbol{\beta}|\boldsymbol{\lambda}_0,\boldsymbol{\beta}_0)$ such that $\ell(\boldsymbol{\lambda},\boldsymbol{\beta}) \ge \ell_\dagger(\boldsymbol{\lambda},\boldsymbol{\beta}|\boldsymbol{\lambda}_0,\boldsymbol{\beta}_0)$ $\forall\boldsymbol{\lambda},\boldsymbol{\lambda}_0>0$ and $\boldsymbol{\beta},\boldsymbol{\beta}_0\in\mathcal{R}^p$ and the equality holds when $\boldsymbol{\lambda}=\boldsymbol{\lambda}_0$ and $\boldsymbol{\beta}=\boldsymbol{\beta}_0$, and*

$$
\ell_\dagger(\boldsymbol{\lambda},\boldsymbol{\beta}|\boldsymbol{\lambda}_0,\boldsymbol{\beta}_0) \equiv \sum_{k=1}^{m}\mathcal{M}_{1,k}(\lambda_k|\boldsymbol{\lambda}_0,\boldsymbol{\beta}_0)+\mathcal{M}_2(\boldsymbol{\beta}|\boldsymbol{\lambda}_0,\boldsymbol{\beta}_0)+\mathcal{M}_3(\boldsymbol{\lambda}_0,\boldsymbol{\beta}_0),
$$

where

$$
\begin{aligned}
& \mathcal{M}_{1,k}(\lambda_k|\boldsymbol{\lambda}_0,\boldsymbol{\beta}_0) \\
\equiv\ & -\frac{\lambda_{k0}^2}{\lambda_k}\sum_{i=1}^{n}\left\{\frac{\Delta_{L,i}}{u_0(L_i,\boldsymbol{X}_i)}I(t_k\le L_i)+\frac{\Delta_{I,i}}{u_0(L_i,R_i,\boldsymbol{X}_i)}I(L_i<t_k\le R_i)\right\} \\
& +\lambda_k\sum_{i=1}^{n}\Bigg[\Delta_{L,i}\left\{A_1(u_0(L_i,\boldsymbol{X}_i))+2A_2(u_0(L_i,\boldsymbol{X}_i))u_0(L_i,\boldsymbol{X}_i)-\frac{1}{u_0(L_i,\boldsymbol{X}_i)}\right\}I(t_k\le L_i) \\
& \quad +\Delta_{I,i}\left\{A_1(u_0(L_i,R_i,\boldsymbol{X}_i))+2A_2(u_0(L_i,R_i,\boldsymbol{X}_i))u_0(L_i,R_i,\boldsymbol{X}_i)-\frac{1}{u_0(L_i,R_i,\boldsymbol{X}_i)}\right\} \\
& \quad \times I(L_i<t_k\le R_i)-\Delta_{I,i}I(t_k\le L_i)-\Delta_{R,i}I(t_k\le R_i)\Bigg] \\
& -\frac{\lambda_k^2}{\lambda_{k0}}\sum_{i=1}^{n}\Bigg\{\Delta_{L,i}A_2(u_0(L_i,\boldsymbol{X}_i))u_0(L_i,\boldsymbol{X}_i)I(t_k\le L_i) \\
& \quad +\Delta_{I,i}A_2(u_0(L_i,R_i,\boldsymbol{X}_i))u_0(L_i,R_i,\boldsymbol{X}_i)I(L_i<t_k\le R_i)\Bigg\},\quad k=1,\ldots,m
\end{aligned}
$$

$$
\begin{aligned}
& \mathcal{M}_2(\boldsymbol{\beta}|\boldsymbol{\lambda}_0,\boldsymbol{\beta}_0) \\
\equiv\ & -\sum_{i=1}^{n}\left[\frac{\Delta_{L,i}}{u_0(L_i,\boldsymbol{X}_i)}\times\frac{\{\boldsymbol{\beta}_0^T\boldsymbol{Z}_{x_i}(L_i)\}^2}{\boldsymbol{\beta}^T\boldsymbol{Z}_{x_i}(L_i)}+\frac{\Delta_{I,i}}{u_0(L_i,R_i,\boldsymbol{X}_i)}\times\frac{\{\boldsymbol{\beta}_0^T(\boldsymbol{Z}_{x_i}(R_i)-\boldsymbol{Z}_{x_i}(L_i))\}^2}{\boldsymbol{\beta}^T(\boldsymbol{Z}_{x_i}(R_i)-\boldsymbol{Z}_{x_i}(L_i))}\right] \\
& +\sum_{i=1}^{n}\Bigg[\Delta_{L,i}\left\{A_1(u_0(L_i,\boldsymbol{X}_i))+2A_2(u_0(L_i,\boldsymbol{X}_i))u_0(L_i,\boldsymbol{X}_i)-\frac{1}{u_0(L_i,\boldsymbol{X}_i)}\right\}\boldsymbol{\beta}^T\boldsymbol{Z}_{x_i}(L_i) \\
& \quad +\Delta_{I,i}\left\{A_1(u_0(L_i,R_i,\boldsymbol{X}_i))+2A_2(u_0(L_i,R_i,\boldsymbol{X}_i))u_0(L_i,R_i,\boldsymbol{X}_i)-\frac{1}{u_0(L_i,R_i,\boldsymbol{X}_i)}\right\} \\
& \quad \times\boldsymbol{\beta}^T\{\boldsymbol{Z}_{x_i}(R_i)-\boldsymbol{Z}_{x_i}(L_i)\}-\Delta_{I,i}\boldsymbol{\beta}^T\boldsymbol{Z}_{x_i}(L_i)-\Delta_{R,i}\boldsymbol{\beta}^T\boldsymbol{Z}_{x_i}(R_i)\Bigg] \\
& -\sum_{i=1}^{n}\Bigg(\Delta_{L,i}A_2(u_0(L_i,\boldsymbol{X}_i))\frac{u_0(L_i,\boldsymbol{X}_i)}{\boldsymbol{\beta}_0^T\boldsymbol{Z}_{x_i}(L_i)}\{\boldsymbol{\beta}^T\boldsymbol{Z}_{x_i}(L_i)\}^2 \\
& \quad +\Delta_{I,i}A_2(u_0(L_i,R_i,\boldsymbol{X}_i))\left\{\frac{u_0(L_i,R_i,\boldsymbol{X}_i)}{\boldsymbol{\beta}_0^T(\boldsymbol{Z}_{x_i}(R_i)-\boldsymbol{Z}_{x_i}(L_i))}\right\}[\boldsymbol{\beta}^T\{\boldsymbol{Z}_{x_i}(R_i)-\boldsymbol{Z}_{x_i}(L_i)\}]^2\Bigg),
\end{aligned}
$$

$A_1(u)=\exp(-u)/\{1-\exp(-u)\}$, $A_2(u)=\exp(-u)/[2\{1-\exp(-u)\}^2]$ *and the expression of* $\mathcal{M}_3(\boldsymbol{\lambda}_0,\boldsymbol{\beta}_0)$ *is given in the Appendix B.*

As opposed to a direct maximization of $\ell(\boldsymbol{\lambda},\boldsymbol{\beta})$, in the MM algorithm, for a given $(\boldsymbol{\lambda}_0,\boldsymbol{\beta}_0)$,

$\ell_\dagger(\boldsymbol{\lambda}, \boldsymbol{\beta}|\boldsymbol{\lambda}_0, \boldsymbol{\beta}_0)$ is maximized with respect to $\boldsymbol{\lambda}$ and $\boldsymbol{\beta}$. Then the new estimates are used to replace $(\boldsymbol{\lambda}_0, \boldsymbol{\beta}_0)$, and then again $\ell_\dagger(\boldsymbol{\lambda}, \boldsymbol{\beta}|\boldsymbol{\lambda}_0, \boldsymbol{\beta}_0)$ is maximized with respect to $(\boldsymbol{\lambda}, \boldsymbol{\beta})$, and this process is continued until $(\boldsymbol{\lambda}, \boldsymbol{\beta})$ and $(\boldsymbol{\lambda}_0, \boldsymbol{\beta}_0)$ are sufficiently close. It is important to note that although MM and EM algorithms are similar in iterative structure, they differ in terms of the objective function that is being maximized. In the EM algorithm, a conditional expectation of the complete data likelihood is maximized, whereas in the MM algorithm the minorization function of the log-likelihood is maximized. Most importantly, our specific choice of the minorization function allows separation of the parameters thereby easing the maximization process. Furthermore, $\mathcal{M}_{1,k}(\lambda_k|\boldsymbol{\lambda}_0, \boldsymbol{\beta}_0)$ and $\mathcal{M}_2(\boldsymbol{\beta}|\boldsymbol{\lambda}_0, \boldsymbol{\beta}_0)$ turned out to be concave functions of λ_k and $\boldsymbol{\beta}$ respectively.

To ensure the positivity of $\lambda_k, k = 1, \ldots, m$, I use the transformed parameters $\eta_k = \log(\lambda_k), k = 1, \ldots, m$ in the optimization. Define $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_m)^\mathsf{T}$ and $\boldsymbol{\eta}_0 = (\eta_{10}, \ldots, \eta_{m0})^\mathsf{T}$, and then replace $\boldsymbol{\lambda}$ and $\boldsymbol{\lambda}_0$ by $\exp(\boldsymbol{\eta})$ and $\exp(\boldsymbol{\eta}_0)$, respectively, in $\mathcal{M}_{1,k}$ and $\mathcal{M}_2$ of the minorization function. Also, hereafter, I will refer to $\ell(\boldsymbol{\lambda}, \boldsymbol{\beta})$ by $\ell(\boldsymbol{\eta}, \boldsymbol{\beta})$. Next, I propose to estimate η_k by solving $S_{1,k}(\eta_k|\boldsymbol{\eta}_0, \boldsymbol{\beta}_0) \equiv \partial\mathcal{M}_{1,k}(\exp(\eta_k)|\exp(\boldsymbol{\eta}_0), \boldsymbol{\beta}_0)/\partial\eta_k = 0$ for $k = 1, \ldots, m$ and $\boldsymbol{\beta}$ by solving $S_2(\boldsymbol{\beta}|\boldsymbol{\eta}_0, \boldsymbol{\beta}_0) \equiv \partial\mathcal{M}_2(\boldsymbol{\beta}|\exp(\boldsymbol{\eta}_0), \boldsymbol{\beta}_0)/\partial\boldsymbol{\beta} = \mathbf{0}$. Note that given $(\boldsymbol{\eta}_0, \boldsymbol{\beta}_0)$, $S_{1,k}(\eta_k|\boldsymbol{\eta}_0, \boldsymbol{\beta}_0)$ is a function of only the scalar parameter η_k. Now, following the general strategy of gradient MM algorithm (Hunter and Lange, 2004), given $(\boldsymbol{\eta}_0, \boldsymbol{\beta}_0)$, $(\boldsymbol{\eta}, \boldsymbol{\beta})$ will be updated by one step Newton-Raphson method and the entire method can be summarized in the following steps.

Step 0. Initialize $(\boldsymbol{\eta}, \boldsymbol{\beta})$.

Step 1. At the ιth step of the iteration I update the parameters as follows: where $(\boldsymbol{\eta}^{(\iota-1)}, \boldsymbol{\beta}^{(\iota-1)})$ and $(\boldsymbol{\eta}^{(\iota)}, \boldsymbol{\beta}^{(\iota)})$ denote the parameter estimate at the $(\iota-1)$th and ιth iterations, respectively.

$$\eta_k^{(\iota)} = \eta_k^{(\iota-1)} - S_{1,kk}^{-1}(\eta_k^{(\iota-1)}|\boldsymbol{\eta}^{(\iota-1)}, \boldsymbol{\beta}^{(\iota-1)}) S_{1,k}(\eta_k^{(\iota-1)}|\boldsymbol{\eta}^{(\iota-1)}, \boldsymbol{\beta}^{(\iota-1)}), \text{ for } k = 1, \ldots, m, \quad (3.2)$$

$$\boldsymbol{\beta}^{(\iota)} = \boldsymbol{\beta}^{(\iota-1)} - S_{22}^{-1}(\boldsymbol{\beta}^{(\iota-1)}|\boldsymbol{\lambda}^{(\iota-1)}, \boldsymbol{\beta}^{(\iota-1)}) S_2(\boldsymbol{\beta}^{(\iota-1)}|\boldsymbol{\lambda}^{(\iota-1)}, \boldsymbol{\beta}^{(\iota-1)}), \quad (3.3)$$

Step 2. Repeat Step 1 until $(\boldsymbol{\eta}^{(\iota-1)}, \boldsymbol{\beta}^{(\iota-1)})$ and $(\boldsymbol{\eta}^{(\iota)}, \boldsymbol{\beta}^{(\iota)})$ are sufficiently close.

In the above iteration both $S_{1,k}$ and $S_{1,kk}$ are scalar valued functions, and S_2 is a p-dimensional

vector while S_{22} is $p \times p$ matrix. After the convergence, the final estimate of β and η will be denoted by $\widehat{\beta}$ and $\widehat{\eta}$. The expression of the terms are

$$
\begin{aligned}
& S_{1,k}(\eta_k^{\iota-1}|\boldsymbol{\eta}^{\iota-1}, \boldsymbol{\beta}^{\iota-1}) \\
= \; & \exp(\eta_k^{\iota-1}) \sum_{i=1}^{n} \Delta_{L,i} A_1(u_{(\iota-1)}(L_i, \boldsymbol{X}_i)) I(t_k \le L_i) - \Delta_{I,i} I(t_k \le L_i) - \Delta_{R,i} I(t_k \le R_i) \\
& \quad + \Delta_{I,i} A_1(u_{(\iota-1)}(L_i, R_i, \boldsymbol{X}_i)) I(L_i < t_k \le R_i), \quad k = 1, \ldots, m, \\
& S_{1,kk}(\eta_k^{\iota-1}|\boldsymbol{\eta}^{\iota-1}, \boldsymbol{\beta}^{\iota-1}) \\
= \; & \exp(\eta_k^{\iota-1}) \sum_{i=1}^{n} \Delta_{L,i} \Bigg[A_1(u_{(\iota-1)}(L_i, \boldsymbol{X}_i)) - 2A_2(u_{(\iota-1)}(L_i, \boldsymbol{X}_i)) u_{(\iota-1)}(L_i, \boldsymbol{X}_i) \\
& \quad - \frac{2}{u_{(\iota-1)}(L_i, \boldsymbol{X}_i)} \Bigg] I(t_k \le L_i) - \Delta_{I,i} I(t_k \le L_i) - \Delta_{R,i} I(t_k \le R_i) \\
& \quad + \Delta_{I,i} \Bigg[A_1(u_{(\iota-1)}(L_i, R_i, \boldsymbol{X}_i)) - 2A_2(u_{(\iota-1)}(L_i, R_i, \boldsymbol{X}_i)) u_{(\iota-1)}(L_i, R_i, \boldsymbol{X}_i) \\
& \quad - \frac{2}{u_{(\iota-1)}(L_i, R_i, \boldsymbol{X}_i)} \Bigg] I(L_i < t_k \le R_i), \quad k = 1, \ldots, m,
\end{aligned}
$$

$$
\begin{aligned}
& S_2(\boldsymbol{\beta}^{(\iota-1)}|\boldsymbol{\eta}^{(\iota-1)}, \boldsymbol{\beta}^{(\iota-1)}) \\
= \; & \sum_{i=1}^{n} \Delta_{L,i} A_1(u_{(\iota-1)}(L_i, \boldsymbol{X}_i)) \boldsymbol{Z}_{x_i}(L_i) - \Delta_{I,i} \boldsymbol{Z}_{x_i}(L_i) - \Delta_{R,i} \boldsymbol{Z}_{x_i}(R_i) \\
& \quad + \Delta_{I,i} A_1(u_{(\iota-1)}(L_i, R_i, \boldsymbol{X}_i)) (\boldsymbol{Z}_{x_i}(R_i) - \boldsymbol{Z}_{x_i}(L_i)), \\
& S_{22}(\boldsymbol{\beta}^{(\iota-1)}|\boldsymbol{\eta}^{(\iota-1)}, \boldsymbol{\beta}^{(\iota-1)}) \\
= \; & 2 \sum_{i=1}^{n} -\Delta_{L,i} \left\{ A_2(u_{(\iota-1)}(L_i, \boldsymbol{X}_i)) u_{(\iota-1)}(L_i, \boldsymbol{X}_i) + \frac{1}{u_{(\iota-1)}(L_i, \boldsymbol{X}_i)} \right\} \frac{\boldsymbol{Z}_{x_i}(L_i)^{\otimes 2}}{\boldsymbol{Z}_{x_i}(L_i)^T \boldsymbol{\beta}^{(\iota-1)}} \\
& \quad - \Delta_{I,i} \left\{ A_2(u_{(\iota-1)}(L_i, R_i, \boldsymbol{X}_i)) u_{(\iota-1)}(L_i, R_i, \boldsymbol{X}_i) + \frac{1}{u_{(\iota-1)}(L_i, R_i, \boldsymbol{X}_i)} \right\} \\
& \quad \frac{(\boldsymbol{Z}_{x_i}(R_i) - \boldsymbol{Z}_{x_i}(L_i))^{\otimes 2}}{(\boldsymbol{Z}_{x_i}(R_i) - \boldsymbol{Z}_{x_i}(L_i))^T \boldsymbol{\beta}^{(\iota-1)}},
\end{aligned}
$$

where $u_{\iota-1}(L_i, \boldsymbol{X}_i)$, $u_{\iota-1}(R_i, \boldsymbol{X}_i)$ and $u_{\iota-1}(L_i, R_i, \boldsymbol{X}_i)$ are the $u_0(L_i, \boldsymbol{X}_i)$, $u_0(R_i, \boldsymbol{X}_i)$ and $u_0(L_i, R_i, \boldsymbol{X}_i)$, with $\boldsymbol{\beta}_0$ and $\boldsymbol{\lambda}_0$ replaced by $\boldsymbol{\beta}^{(\iota-1)}$ and $\exp(\boldsymbol{\eta}^{(\iota-1)})$, respectively.

3.3.2 Variance estimation

Zeng et al. (2006) have studied the asymptotic properties of the maximum likelihood estimator, and used the profile likelihood method (Murphy and Van der Vaart, 2000) to calculate the asymptotic standard error of the estimator. I also follow their idea of the standard error calculation which will be aided by our computational tools. Suppose that the estimator of the covariance matrix of $\widehat{\boldsymbol{\beta}}$ is $-D^{-1}$, then the (r,s)th element of the $p \times p$ matrix D is

$$\frac{\mathrm{pl}(\widehat{\boldsymbol{\beta}}) - \mathrm{pl}(\widehat{\boldsymbol{\beta}} + h_n \boldsymbol{e}_r) - \mathrm{pl}(\widehat{\boldsymbol{\beta}} + h_n \boldsymbol{e}_s) + \mathrm{pl}(\widehat{\boldsymbol{\beta}} + h_n \boldsymbol{e}_r + h_n \boldsymbol{e}_s)}{h_n^2},$$

with $\boldsymbol{e}_r$ being the vector with 1 at the rth position and 0 elsewhere and h_n is a constant with an order $n^{-1/2}$, and $\mathrm{pl}(\boldsymbol{\beta})$ stands for the profile log-likelihood function defined as $\mathrm{pl}(\boldsymbol{\beta}) = \ell(\widehat{\boldsymbol{\eta}}^{\beta}, \boldsymbol{\beta})$, where $\widehat{\boldsymbol{\eta}}^{\beta} = \mathrm{argmax}_{\boldsymbol{\eta} \in \mathcal{R}^m} \ell(\boldsymbol{\eta}, \boldsymbol{\beta})$. To obtain $\widehat{\boldsymbol{\eta}}^{\beta}$, I use the proposed minorization function, and specifically use the m equations given in (3.2) after fixing $\boldsymbol{\beta}^{\iota-1}$ to $\boldsymbol{\beta}$. For any given $\boldsymbol{\beta}$, the computation of $\widehat{\boldsymbol{\eta}}^{\beta}$ is very fast when $\widehat{\boldsymbol{\eta}} = (\widehat{\eta}_1, \ldots, \widehat{\eta}_m)^{\mathsf{T}}$, the MLE, is used as the initial value. In contrast, obtaining $\widehat{\boldsymbol{\eta}}^{\beta}$ using any generic optimization of $\ell(\boldsymbol{\eta}, \boldsymbol{\beta})$ is very time consuming.

3.3.3 Complexity analysis

In the proposed method parameters are updated by equations (3.2) and (3.3). Now I inspect the computational complexity (or simply complexity) of a single update. The complexity to calculate $S_2(\boldsymbol{\beta}|\boldsymbol{\eta}, \boldsymbol{\beta})$ and $S_{22}(\boldsymbol{\beta}|\boldsymbol{\eta}, \boldsymbol{\beta})$ is $O(np + np^2)$, where n is the sample size. Next, the complexity of inverting $S_{22}(\boldsymbol{\beta}|\boldsymbol{\eta}, \boldsymbol{\beta})$ is $O(p^3)$. Therefore, the complexity of one update of $\boldsymbol{\beta}$ is $O(np + np^2 + p^3)$. Similarly, for any $k = 1, \ldots, m$, the complexity of one step update of η_k is $O(2n + 1)$. Hence, the total computational cost for updating $\boldsymbol{\eta}$ and $\boldsymbol{\beta}$ is $O((2n + 1)m + np + np^2 + p^3)$.

Now, I look closely the computational complexity of the generic optimization of the log-likelihood $\ell(\boldsymbol{\lambda}, \boldsymbol{\beta})$ using the Newton-Raphson approach. In each step, the computational cost of gradient and the Hessian matrix of the log-likelihood is $O(n(m + p) + n(m + p)^2)$ and inverting a matrix of order $m + p$ will cost $O((p + m)^3)$. The total complexity for a single update is then $O(n(p+m)+n(m+p)^2+(p+m)^3)$, which is obviously larger than $O((2n+1)m+np+np^2+p^3)$.

Since m increases with the sample size n, the difference between the two complexities get wider with n. Alternative to Newton's method, if a quasi-Newton method is used for the generic optimization (such as the BFGS algorithm), the complexity becomes $O(n(p+m)+(n+1)(m+p)^2)$ which is still larger than the complexity of the proposed method.

3.4 Simulation study

In this section, I conducted numerical study to assess the performance of the proposed MM algorithm. I considered two main scenarios, 1) time-independent covariates and 2) time-dependent covariates. For scenario 1, a scalar covariate X was simulated from $\mathrm{Bernoulli}(0.5)$. Conditional on the covariate, I considered the following hazard function $h(t|X) = 0.2 + \beta X$. For scenario 2, the hazard function was $h(t|X) = 0.2 + \beta X \exp(t)$, with $X \sim \mathrm{Bernoulli}(0.5)$. I considered two different values of β, 0.5 and 1. For both scenarios, left censoring time L_i was independently generated from $\mathrm{Uniform}(0.1,\, 2)$ and the right censoring time R_i was generated from $\mathrm{Uniform}(L_i + 0.5, 4)$. The proportion of left censoring was from 30% to 50% and the proportion of right censoring was from 25% to 35% across all the scenarios. For each scenario I considered three sample sizes, $n = 100$, 200 and 500. For the profile likelihood based standard error calculation, I used $h_n = 1.5n^{-1/2}$ because among several trial values of h_n this one yielded good agreement between the standard deviation and the standard error of the estimators. There was no nonconvergence in any of the proposed MM algorithm.

I fit the additive risks model (3.1) to each of the simulated dataset using the proposed MM algorithm. The results of the simulation study with 500 replications are presented in Table 3.1. For each scenario, I report the average of the estimates (Est) for β, empirical standard deviation (SD), the average of the estimated standard error (SE), and the 95% coverage probability (CP) based on Wald's confidence interval. The results indicate that the proposed MM algorithm can estimate the parameters very well, the bias could be up to 8.5% among all the scenarios. Overall, the bias and SD decrease with the sample size n. There is a reasonable agreement between the empirical standard deviation and the estimated standard error. The CPs are quite close to the nominal level, 0.95.

Table 3.1: Results of the simulation study with a scalar covariate.

Time-independent covariate: $h(t\|X) = 0.2 + \beta X$													
		$n = 100$				$n = 200$				$n = 500$			
$\lambda(t)$	β	Est	SD	SE	CP	Est	SD	SE	CP	Est	SD	SE	CP
0.2	0.5	0.495	0.145	0.150	0.956	0.496	0.096	0.099	0.952	0.499	0.059	0.058	0.946
0.2	1.0	1.047	0.222	0.248	0.978	1.005	0.161	0.160	0.944	1.012	0.100	0.091	0.936
Time-dependent covariate: $h(t\|X) = 0.2 + \beta X \exp(t)$													
		$n = 100$				$n = 200$				$n = 500$			
$\lambda(t)$	β	Est	SD	SE	CP	Est	SD	SE	CP	Est	SD	SE	CP
0.2	0.5	0.518	0.134	0.160	0.992	0.504	0.090	0.102	0.980	0.505	0.053	0.059	0.974
0.2	1.0	1.085	0.314	0.317	0.986	1.040	0.200	0.202	0.978	1.013	0.110	0.113	0.950

To assess the performance of the algorithm for multiple covariates scenario, I conducted another simulation study with $h(t|X_1, X_2) = 0.2t^{1/2} + \beta_1 X_1 + \beta_2 X_2$, and both covariates X_1 and X_2 were generated from Bernoulli(0.5) and set $\beta_1 = 0.5$ and $\beta_2 = 1$. After simulating the time-to-event T using the additive hazard $h(t|X_1, X_2)$, the left censoring time L was independently generated from Uniform(0.1, 1.5) and the right censoring time R was generated from $\text{Uniform}(L + 1.5, 4)$. This resulted in 42% left censored, 42% interval censored, and 16% right censored subjects. The results are presented in Table 3.2, and I find that the overall performance of the algorithm is as good as in Table 3.1.

In all computations, the iteration is stopped when the sum of the absolute differences of the estimates for η and β at two successive iterations is less than 10^{-3}. All computations were done in an Intel(R) Xeon(R) CPU E5-2680 v4 @ 2.40GHz machine. In Table 3.3 I provide the average computation time to get parameter estimates and the standard errors for different sample sizes and for the scalar covariate and the two covariates scenarios using the proposed method and the direct optimization of the log-likelihood using the BFGS algorithm. The results show that the proposed method is many times faster than the direct optimization of the log-likelihood function, and the relative gain in the computation increases with the sample size.

Table 3.2: Results of the simulation study with two covariates, $X_1 \sim \text{Bernoulli}(0.5)$ and $X_2 \sim \text{Bernoulli}(0.5)$.

	$n = 100$				$n = 200$				$n = 500$			
	Est	SD	SE	CP	Est	SD	SE	CP	Est	SD	SE	CP
$\beta_1 = 0.5$	0.490	0.193	0.202	0.958	0.493	0.127	0.130	0.950	0.501	0.077	0.076	0.940
$\beta_2 = 1.0$	1.027	0.287	0.287	0.968	1.021	0.181	0.186	0.964	1.010	0.107	0.104	0.934

Table 3.3: The average time (in seconds) to compute estimates (ATE) and standard errors (ATS). Case 1: scalar covariate; Case 2: two covariates; MM: proposed MM algorithm; Direct: direct optimization

		$n = 100$		$n = 200$		$n = 500$	
		ATE	ATS	ATE	ATS	ATE	ATS
Case 1	MM	1.08	0.39	11.92	7.33	78.96	80.04
	Direct	3.50	1.24	37.79	18.88	1587.08	666.62
Case 2	MM	1.91	1.88	13.14	16.93	87.78	208.13
	Direct	8.32	6.23	92.81	65.10	1988.76	1812.97

3.5 Real data analysis

For illustrating the proposed method I analyze the breast cancer data given in Finkelstein and Wolfe (1985). In this breast cosmesis study, the subjects who were under the adjuvant chemotherapy after tumorectomy were periodically followed-up for the cosmetic effect of the therapy. So, patients generally visited the clinic every 4 to 6 months, thus, the time of the appearance of breast retraction was recorded as an interval. Particularly, if the recorded time for a patient is $(0, 4]$, then the breast retraction happened before 4 months, whereas if for any subject the time to occurrence is $(6, 12]$, then it signifies that the event had happened between 6 and 12 months. There were 94 early breast cancer patients in the study, of which 46 patients were given radiation therapy alone and 48 patients were given radiation therapy plus adjuvant chemotherapy.

I set $X = 1$ if a patient had received adjuvant chemotherapy following the initial radiation treatment and 0 otherwise. So X is a time independent covariate and I fit $h(t|X) = \lambda(t) + X\beta$

model to the data using the proposed method. Here β represents the difference in the hazard of breast retraction between $X = 1$ and $X = 0$ groups at any time point. I obtained $\widehat{\beta} = 0.031$. Since the choice of h_n was quite arbitrary in the profile likelihood based method of standard error, I have used different values of h_n, $1.5n^{-1/2}$, $n^{-1/2}/20$, $n^{-1/2}/100$ and $n^{-1/2}/1000$, and obtained 0.09, 0.08, 0.05 and 0.007 as the standard error. Obviously for standard error 0.007, β is significantly different from zero at the 5% level, while for other standard errors β is not significantly different from zero. To investigate the issue further, I calculated bootstrap standard error based on 200 bootstrap samples, and it came out to be 0.06. Figure 3.1 shows the estimated survival functions for the two groups along with the 95% pointwise confidence intervals calculated using the bootstrap method. This analysis shows no significant difference between the two survival functions at any time.

3.6 Conclusions

In this chapter, I proposed an efficient algorithm to obtain the maximum likelihood estimates of a complex likelihood function for the additive risks model with the interval censored data. The attractive feature of the method is enabling the separation of the finite and infinite dimensional parameters. Furthermore, it allows separation of the components of the infinite-dimensional parameter which is a big advantage as the dimension of the infinite dimensional parameter increases with the sample size. The numerical study shows that the algorithm works pretty good. I have not encountered any convergence issue in the simulation or real data analysis section.

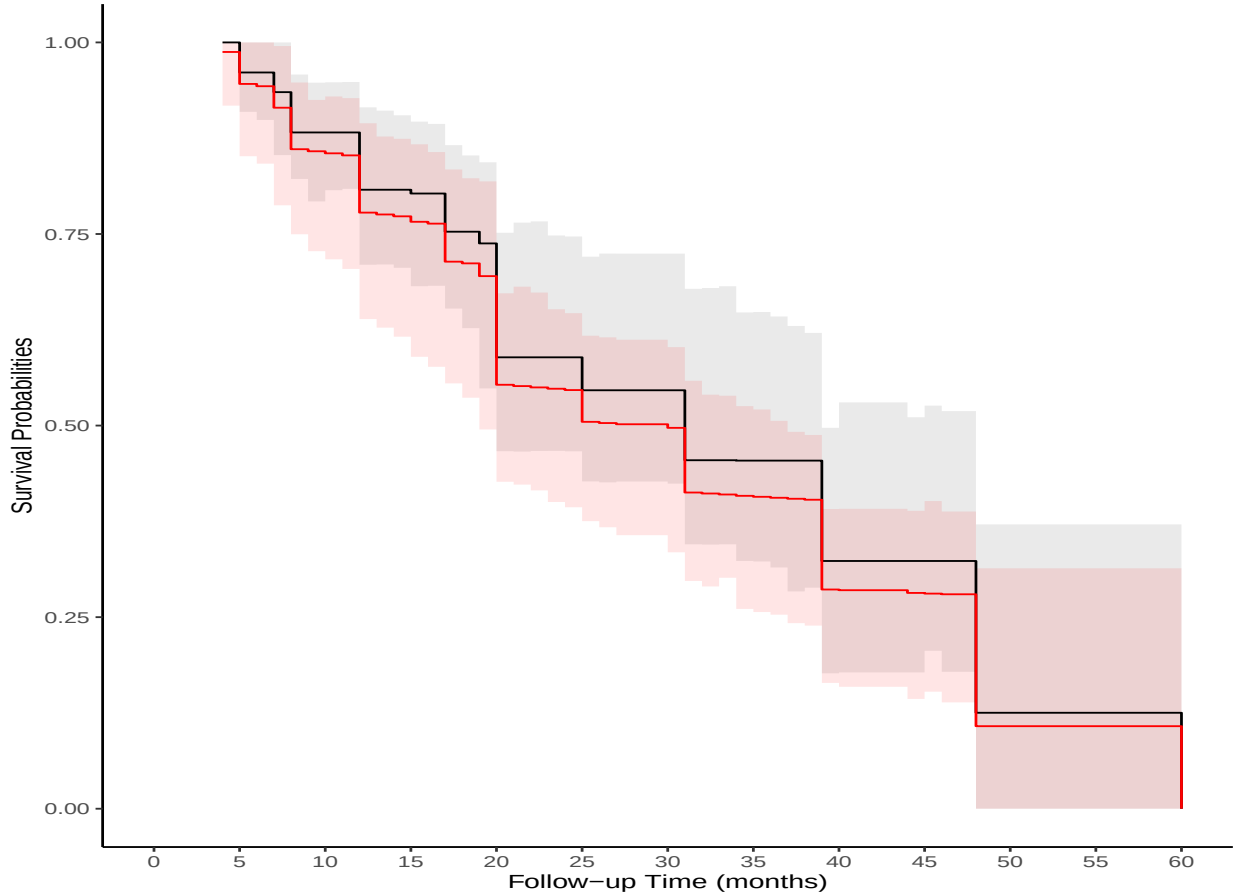

Figure 3.1: Estimated survival curves of the breast cancer data. The red and black curves represent the estimated survival curves for the patients with $X = 1$ (adjuvant chemotherapy + radiation) and $X = 0$ (only radiation), respectively. The pink and gray shaded areas are the confidence bands for red and black curves, respectively.

4. CONCLUSION AND FUTURE WORK

4.1 Summary

In this dissertation, I develop two novel MM algorithms in Chapters 2 and 3, respectively, based on the seminal inequality given in Theorem 2.1. By applying these algorithms, the computational cost of the estimation of the semiparametric model with interval-censored data is reduced significantly without any loss of information.

The proposed MM algorithm in Chapter 2 provides an alternative to EM-based algorithm for the GOR model with clustered current status data. It simplifies the estimation procedure by separating the two parts of parameters, i.e., nonparametric and regression parameters. The algorithm also works for non-clustered data and the details are discussed in Chapter 2. The inequalities presented in Theorem 2.1 are potentially powerful tools for developing MM algorithms for binary data and lifetime data analysis.

In a related development, Zhou et al. (2017) proposed an EM algorithm for parameter estimation in the semiparametric GOR model for interval-censoring, without clustering and without any asymptotic justification. Very recently, Li et al. (2020) proposed an EM-based approach to obtain the NPML estimator for the semiparametric transformation models, of which, the GOR is a special class, to clustered CS data. Particularly, the nonparametric components of the model at different inspection times were considered as distinct parameters. In contrast, I model the nonparametric component of the GOR model using splines, and then develop an MM algorithm that has wider applications and flexible in frameworks where the EM-based development is difficult. The current modeling and computational techniques in Chapter 2 can be advanced in various directions, such as, variable selection in the presence of many covariates, and incorporation of informative cluster size – a commonplace in oral health research. All these will require non-trivial adjustments to the methodology presented here, and will be considered elsewhere.

In Chapter 3, I conduct statistical inference for case-II interval-censored data via a semi-

parametric additive risks model. In the proposed model, the regression parameter $\boldsymbol{\beta}$ and high-dimensional parameter $\boldsymbol{\lambda}$ are entangled and direct maximization of the likelihood function is very time consuming and not guaranteed to produce accurate results when the sample size is large and when the inspection time is continuous. To ease the computational cost of the maximization with respect to a high-dimensional vector, I develop an efficient algorithm to separate not only the non-parametric and regression parameters but also each component of the nonparametric parameter. As a result, in each iteration, I only solve low-dimensional or even scalar equations. I believe that this proposal will help generate new ideas for handling computational bottlenecks of complex models and likelihoods. Some interesting topics of future research include developing efficient computational tools for the clustered right censored or interval censored data that arise in the case of dependent censoring or length-biased sampling. Additionally, developing computationally efficient method for case-I interval censored data (Huang et al., 1996; Wang et al., 2020) could be a direction of future research.

4.2 Dependent inspection

In Chapter 2, the inspection time is assumed to be independent of the time-to-event. However, in the real data world, the observation time is sometimes related to the failure of interest, which is referred as dependent inspection or informative censoring. A famous dependent current status data is tumorigenicity experiments, in which the failure of interest is the time to tumor onset. The animals are only checked once whether the tumor is present at the death. The occurrences of tumors often have some affects on animal death rate which thus cause dependent inspection. Another example is the GAAD data served as an illustration in Chapter 2. In the GAAD data, the inspection time C and the time-to-event T are also likely to be correlated since the teeth are usually inspected by dentists when condition of the teeth or gum becomes worse. There are mainly two methods to model the dependence. One is copula model-based estimation, in which the joint distribution of T and C is modeled as

$$F(t, c) = \text{pr}(T \leq t, C \leq c) = C_\alpha\{F_T(t), F_C(c)\},$$

where C_α is a copula function (Nelsen, 2007) defined on $[0,1] \times [0,1]$, $F_T(t)$ and $F_C(c)$ are the marginal distribution of T and C, respectively. By applying the copula model, Wang et al. (2012) and Hsieh and Chen (2020) proposed nonparametric estimation of the survival function for the failure time. Xu et al. (2019) and Ma et al. (2015) considered copula model-based semiparametric model for current status data with dependent inspection time. Xu et al. (2019) considered modeling the failure time via the linear transformation model, while Ma et al. (2015) used the popular proportional hazards model. Alternative to the copula model-based approach, frailty model-based method is another choice, where the correlation between T and C is accounted by some latent variables. In Li et al. (2017), the authors considered the proportional hazard model for both T and C with the hazard functions

$$\Lambda_T(t|\boldsymbol{X}, b) = \Lambda_1(t) \exp(\boldsymbol{\beta}_1{}^\mathsf{T} \boldsymbol{X}) b$$

and

$$\Lambda_C(c|\boldsymbol{X}, b) = \Lambda_2(c) \exp(\boldsymbol{\beta}_2{}^\mathsf{T} \boldsymbol{X}) b,$$

where Λ_1 and Λ_2 are the cumulative hazard functions for T and C, respectively, $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are the two regression parameters, $\boldsymbol{X}$ is the $p \times 1$ covariate vector, and b is the latent variable with mean one and unknown variance. Other approaches of modeling the failure time and introducing the latent variables can be found in Zhang et al. (2005) and Chen et al. (2012). However, the existing approaches have not considered the clustering effect, which is a crucial component in the GAAD data along with dependent censoring. Therefore, I plan to develop a model to handle dependent current status data in the presence of the clustering effect. Similarly, MM algorithm is potentially a good approach to simplify the estimation procedure for this complex model.

4.3 Length-Biased sampling

In follow-up studies, an economical and popular design is recruiting the subjects who survival at the sampling/recruitment time and have not experienced the failure of interest. Only the subjects whose failure time is more than the sampling time can be included in the study. The failure time of the subjects in the cohort is likely to be longer than that arises from the underlying failure time distribution, leading to a length-biased (LB) data. Consequently, without a proper adjustment, LB data may lead to overestimation of the underlying failure time and inconsistent estimator of the model parameters. For example, in the breast cancer data presented in Chapter 1.1.2, the patients who have died before the sampling/recruitment time cannot be included in the cohort, which results in LB sampling. There is a limited work on length-biased interval-censored data. Recently, Gao and Chan (2019) considered a nonparametric maximum likelihood estimation for the proportional hazards model with LB interval-censored data. It is well known that sometimes the proportional hazards model may not be appropriate and other models including the additive hazards model provide useful alternatives. Thus, I plan to model the LB interval-censored data via the additive risks model and apply the computational techniques shown in this dissertation.

REFERENCES

Armitage, G. (1999). Development of a classification system for periodontal diseases and conditions. *Annals of Periodontology* **4,** 1–6.

Banerjee, T., Chen, M., Dey, D. K., and Kim, S. (2007). Bayesian analysis of generalized odds-rate hazards models for survival data. *Lifetime Data Analysis* **13,** 241–260.

Betensky, R. A., Rabinowitz, D., and Tsiatis, A. A. (2001). Computationally simple accelerated failure time regression for interval censored data. *Biometrika* **88,** 703–711.

Bickel, P. J., Klaassen, C. A., and Wellner, J. A. (1993). *Efficient and adaptive estimation for semiparametric models*. Johns Hopkins University Press, Baltimore.

Billingsley, P. (1995). *Probability and Measure*. Wiley, New York.

Chen, C.-M., Lu, T.-F. C., Chen, M.-H., and Hsu, C.-M. (2012). Semiparametric transformation models for current status data with informative censoring. *Biometrical journal* **54,** 641–656.

Chen, D., Sun, J., and Peace, K. (2012). *Interval-Censored Time-to-Event Data: Methods and Applications*. Chapman and Hall/CRC.

Cook, R. (1986). Assessment of local influence. *Journal of the Royal Statistical Society: Series B (Methodological)* **48,** 133–155.

Cook, R. and Tolusso, D. (2009). Second-order estimating equations for the analysis of clustered current status data. *Biostatistics* **10,** 756–772.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)* **34,** 187–202.

de Boor, C. (1978). *A Practical Guide to Splines*, volume 27. Springer-Verlag.

Feng, Y., Lin, S., and Li, Y. (2019). Semiparametric regression of clustered current status data. *Journal of Applied Statistics* **46,** 1724–1737.

Fernandes, J., Wiegand, R., Salinas, C., Grossi, S., Sanders, J., Lopes-Virella, M., and Slate, E. (2009). Periodontal disease status in gullah african americans with type 2 diabetes living in south carolina. *Journal of Periodontology* **80,** 1062–1068.

Finkelstein, D. M. (1986). A proportional hazards model for interval-censored failure time data. *Biometrics* pages 845–854.

Finkelstein, D. M. and Wolfe, R. A. (1985). A semiparametric model for regression analysis of interval-censored failure time data. *Biometrics* pages 933–945.

Gao, F. and Chan, K. C. G. (2019). Semiparametric regression analysis of length-biased interval-censored data. *Biometrics* **75,** 121–132.

Gentleman, R. and Geyer, C. J. (1994). Maximum likelihood for interval censored data: Consistency and computation. *Biometrika* **81,** 618–623.

Gray, R. (1992). Flexible methods for analyzing survival data using splines, with applications to breast cancer prognosis. *Journal of the American Statistical Association* **87,** 942–951.

Groeneboom, P. and Wellner, J. A. (1992). *Information bounds and nonparametric maximum likelihood estimation*, volume 19. Springer Science & Business Media.

Henderson, N. C. and Varadhan, R. (2019). Damped Anderson acceleration with restarts and monotonicity control for accelerating EM and EM-like algorithms. *Journal of Computational and Graphical Statistics* **28,** 834–846.

Hsieh, J.-J. and Chen, Y.-Y. (2020). Survival function estimation of current status data with dependent censoring. *Statistics & Probability Letters* **157,** 108621.

Huang, J. et al. (1996). Efficient estimation for the proportional hazards model with interval censoring. *Annals of Statistics* **24,** 540–568.

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

Huang, J. and Wellner, J. A. (1997). Interval censored survival data: a review of recent progress. In *Proceedings of the First Seattle Symposium in Biostatistics*, pages 123–169. Springer.

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

Hunter, D. R. and Lange, K. (2004). A tutorial on MM algorithms. *The American Statistician* **58,** 30–37.

Hurvich, C., Simonoff, J., and Tsai, C.-L. (1998). Smoothing parameter selection in nonparametric regression using an improved akaike information criterion. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **60,** 271–293.

Kor, C.-T., Cheng, K.-F., and Chen, Y.-H. (2013). A method for analyzing clustered interval-censored data based on cox's model. *Statistics in medicine* **32,** 822–832.

Li, J., Wang, C., and Sun, J. (2012). Regression analysis of clustered interval-censored failure time data with the additive hazards model. *Journal of nonparametric statistics* **24,** 1041–1050.

Li, L. and Pu, Z. (2003). Rank estimation of log-linear regression with interval-censored data. *Lifetime data analysis* **9,** 57–70.

Li, S., Hu, T., Wang, P., and Sun, J. (2017). Regression analysis of current status data in the presence of dependent censoring with applications to tumorigenicity experiments. *Computational Statistics & Data Analysis* **110,** 75–86.

Li, S., Hu, T., Zhao, S., and Sun, J. (2020). Regression analysis of multivariate current status data with semiparametric transformation frailty models. *Statistica Sinica* **30,** 1117–1134.

Lin, D. and Ying, Z. (1994). Semiparametric analysis of the additive risk model. *Biometrika* **81,** 61–71.

Liu, Q. and Pierce, D. (1994). A note on Gauss-Hermite quadrature. *Biometrika* **81,** 624–629.

Ma, L., Hu, T., and Sun, J. (2015). Sieve maximum likelihood regression analysis of dependent current status data. *Biometrika* **102,** 731–738.

Martinussen, T. and Scheike, T. H. (2002). Efficient estimation in additive hazards regression with current status data. *Biometrika* **89,** 649–658.

Murphy, S. A. and van der Vaart, A. W. (1999). Observed information in semi-parametric models. *Bernoulli* **5,** 381–412.

Murphy, S. A. and Van der Vaart, A. W. (2000). On profile likelihood. *Journal of the American Statistical Association* **95,** 449–465.

Nelsen, R. B. (2007). *An Introduction to Copulas*. Springer Science & Business Media.

Rabinowitz, D., Tsiatis, A., and Aragon, J. (1995). Regression with interval-censored data.

Biometrika **82,** 501–513.

Ramsay, J. O. et al. (1988). Monotone regression splines in action. *Statistical Science* **3,** 425–441.

Rice, J. A. and Silverman, B. W. (1991). Estimating the mean and covariance structure nonparametrically when the data are curves. *Journal of the Royal Statistical Society: Series B (Methodological)* **53,** 233–243.

Ruppert, D., Wand, M., and Carroll, R. (2003). *Semiparametric Regression*, volume 12 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press.

Satten, G. A. (1996). Rank-based inference in the proportional hazards model for interval censored data. *Biometrika* **83,** 355–370.

Stone, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *The Annals of Statistics* **10,** 1040–1053.

Stone, C. J. (1985). Additive regression and other nonparametric models. *The Annals of Statistics* **13,** 689–705.

Su, P.-F. and Chi, Y. (2014). Marginal regression approach for additive hazards models with clustered current status data. *Statistics in Medicine* **33,** 46–58.

Sun, J. (2007). *The statistical analysis of interval-censored failure time data*. Springer.

Sun, J. and Sun, L. (2005). Semiparametric linear transformation models for current status data. *Canadian Journal of Statistics* **33,** 85–96.

Turnbull, B. (1976). The empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society: Series B (Methodological)* **38,** 290–295.

van de Geer, S. A. (2000). *Empirical Processes in M-estimation*. Cambridge University Press, Cambridge, UK; New York.

van der Vaart, A. (2002). Semiparameric statistics. *Lectures on Probability Theory and Statistics* pages 331–457.

van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press, Cambridge, UK.

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes*. Springer Verlag, New York.

Wang, C., Sun, J., Sun, L., Zhou, J., and Wang, D. (2012). Nonparametric estimation of current status data with dependent censoring. *Lifetime data analysis* **18,** 434–445.

Wang, H.-L., Burgett, F., Shyr, Y., and Ramfjord, S. (1994). The influence of molar furcation involvement and mobility on future clinical periodontal attachment loss. *Journal of Periodontology* **65,** 25–29.

Wang, L., Sun, J., and Tong, X. (2010). Regression analysis of case ii interval-censored failure time data with the additive hazards model. *Statistica Sinica* **20,** 1709.

Wang, P., Zhou, Y., and Sun, J. (2020). A new method for regression analysis of interval-censored data with the additive hazards model. *Journal of the Korean Statistical Society* pages 1–17.

Wang, T., He, K., Ma, M., Bandyopadhyay, D., and Sinha, S. (2020). Minorize-maximize algorithm for the generalized odds rate model for clustered current status data. *Preprint* .

Wellner, J. A. and Zhan, Y. (1997). A hybrid algorithm for computation of the nonparametric maximum likelihood estimator from censored data. *Journal of the American Statistical Association* **92,** 945–959.

Wen, C. C. and Chen, Y. H. (2011). Nonparametric maximum likelihood analysis of clustered current status data with the gamma-frailty Cox model. *Computational Statistics and Data Analysis* **55,** 1053–1060.

Wu, T. T., Lange, K., et al. (2010a). The mm alternative to em. *Statistical Science* **25,** 492–505.

Wu, T. T., Lange, K., et al. (2010b). The MM alternative to EM. *Statistical Science* **25,** 492–505.

Xu, D., Zhao, S., Hu, T., Yu, M., and Sun, J. (2019). Regression analysis of informative current status data with the semiparametric linear transformation model. *Journal of Applied Statistics* **46,** 187–202.

Yavuz, A. Ç. and Lambert, P. (2016). Semi-parametric frailty model for clustered interval-censored data. *Statistical Modelling* **16,** 360–391.

Younes, N. and Lachin, J. (1997). Link-based models for survival data with interval and continuous time censoring. *Biometrics* pages 1199–1211.

Zeng, D., Cai, J., and Shen, Y. (2006). Semiparametric additive risks model for interval-censored

data. *Statistica Sinica* **16,** 287–302.

Zeng, D., Lin, D. Y., and Yin, G. (2005). Maximum likelihood estimation for the proportional odds model with random effects. *Journal of the American Statistical Association* **100,** 470–483.

Zeng, D., Mao, L., and Lin, D. (2016). Maximum likelihood estimation for semiparametric transformation models with interval-censored data. *Biometrika* **103,** 253–271.

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

Zhang, Z. and Sun, J. (2010). Interval censoring. *Statistical methods in medical research* **19,** 53–70.

Zhang, Z., Sun, J., and Sun, L. (2005). Statistical analysis of current status data with informative observation times. *Statistics in Medicine* **24,** 1399–1407.

Zhang, Z., Sun, L., Zhao, X., and Sun, J. (2005). Regression analysis of interval-censored failure time data with linear transformation models. *Canadian Journal of Statistics* **33,** 61–70.

Zhou, H. and Zhang, Y. (2012). EM vs MM: A case study. *Computational Statistics & Data Analysis* **56,** 3909–3920.

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

APPENDIX A

APPENDIX FOR CHAPTER 2

A.1 Results of Chapter 2.3

A.1.1 Proof of Theorem 2.1

A.1.1.1 Proof of part i)

Proof. Define $f(u) = \log[\{1-(1+ru)^{-1/r}\}/\{1-(1+ru_0)^{-1/r}\}]$. Then $f(u_0) = 0$. Define $A_1(u) = \partial f(u)/\partial u$ and $A_2(u) = -0.5\partial^2 f(u)/\partial u^2$. Consider the Taylor series expansion of $f(u)$ about $u = u_0$,

$$f(u) = (u-u_0)A_1(u_0) - (u-u_0)^2 A_2(u^*),$$

for some $|u^* - u| < |u_0 - u|$. The fact is $A_2(u)$ is creasing in u for any $r > 0$. Therefore, for $u > u^* > u_0$ and any $r > 0$, $A_2(u) > A_2(u^*) > A_2(u_0)$, and then

$$\begin{aligned} f(u) &\geq (u-u_0)A_1(u_0) - (u-u_0)^2 A_2(u_0) \\ &= (u-u_0)A_1(u_0) - (u-u_0)^2 A_2(u_0) + \kappa\left\{\log\left(\frac{u_0}{u}\right) + \log\left(\frac{u}{u_0}\right)\right\} \\ &\geq (u-u_0)A_1(u_0) - (u-u_0)^2 A_2(u_0) + \kappa\left\{\log\left(\frac{u_0}{u}\right) + \left(1-\frac{u_0}{u}\right)\right\}. \end{aligned} \tag{A.1}$$

In fact the result (A.1) holds for either choices of κ mentioned in Theorem 2.1.

To prove the result when $u < u_0$, let me define $g(u) = f(u) - (u-u_0)A_1(u_0) + (u-u_0)^2 A_2(u_0) - \kappa\{\log(u_0/u) + 1 - u_0/u\}$. Then

$$\begin{aligned} g^{'}(u) &= A_1(u) - A_1(u_0) + 2(u-u_0)A_2(u_0) + \kappa\frac{u-u_0}{u^2} \\ &= (u-u_0)\left\{-2A_2(u_\dagger) + 2A_2(u_0) + \frac{\kappa}{u^2}\right\}, \end{aligned}$$

where the last equality is obtained by applying the Taylor series expansion on $A_1(u)$ about $u = u_0$, $A_1(u) = A_1(u_0) - 2(u - u_0)A_2(u_\dagger)$, for some $u_\dagger \in [u, u_0)$.

Next I consider the case of $0 < r \leq 1$ with $\kappa = 1/r$. Define

$$\begin{aligned} B_1 \equiv \frac{1}{ru^2} - 2A_2(u_\dagger) &= \frac{1}{ru^2} - \frac{(1+ru_\dagger)^{-1/r-2}[1 + r\{1-(1+ru_\dagger)^{-1/r}\}]}{\{1-(1+ru_\dagger)^{-1/r}\}^2} \\ &= \frac{\{1-(1+ru_\dagger)^{-1/r}\}^2 - ru^2(1+ru_\dagger)^{-1/r-2}[1 + r\{1-(1+ru_\dagger)^{-1/r}\}]}{ru^2\{1-(1+ru_\dagger)^{-1/r}\}^2}, \end{aligned}$$

and $B_2 = B_1 ru^2\{1-(1+ru_\dagger)^{-1/r}\}^2$. Then

$$\begin{aligned} B_2 &= \{1-(1+ru_\dagger)^{-1/r}\}\left[1-(1+ru_\dagger)^{-1/r} - r^2u^2(1+ru_\dagger)^{-1/r-2}\right] - ru^2(1+ru_\dagger)^{-1/r-2} \\ &= \left\{\frac{1-(1+ru_\dagger)^{-1/r}}{(1+ru_\dagger)^{1/r+2}}\right\}\left\{(1+ru_\dagger)^{1/r+2} - (1+ru_\dagger)^2 - r^2u^2\right\} - ru^2(1+ru_\dagger)^{-1/r-2} \\ &= \frac{1}{(1+ru_\dagger)^{1/r+2}}\left[\{1-(1+ru_\dagger)^{-1/r}\}\left\{(1+ru_\dagger)^{1/r+2} - (1+ru_\dagger)^2 - r^2u^2\right\} - ru^2\right]. \end{aligned}$$

Using the Bernoulli inequality I have for $0 < r \leq 1$,

$$\begin{aligned} (1+ru_\dagger)^{1/r+2} = (1+ru_\dagger)^{1/r}(1+ru_\dagger)^2 &\geq (1+u_\dagger)(1+ru_\dagger)^2 \\ &= (1+ru_\dagger)^2 + u_\dagger(1+ru_\dagger)^2. \end{aligned}$$

Now, using this inequality in the numerator of B_2 I obtain

$$\begin{aligned} B_2 &\geq \frac{1}{(1+ru_\dagger)^{1/r+2}}\left[\{1-(1+ru_\dagger)^{-1/r}\}\left\{(1+ru_\dagger)^2 + u_\dagger(1+ru_\dagger)^2 - (1+ru_\dagger)^2 - r^2u^2\right\} - ru^2\right] \\ &= \frac{1}{(1+ru_\dagger)^{1/r+2}}\left[\left\{1 - \frac{1}{(1+ru_\dagger)^{1/r}}\right\}\left\{u_\dagger(1+ru_\dagger)^2 - r^2u^2\right\} - ru^2\right] \\ &\geq \frac{1}{(1+ru_\dagger)^{1/r+2}}\left[\left(1 - \frac{1}{1+u_\dagger}\right)\left\{u_\dagger(1+ru_\dagger)^2 - r^2u^2\right\} - ru^2\right] \\ &= \frac{1}{(1+ru_\dagger)^{1/r+2}}\left[\left(\frac{u_\dagger}{1+u_\dagger}\right)\left\{u_\dagger(1+ru_\dagger)^2 - r^2u^2\right\} - ru^2\right] \\ &= \frac{B_3}{(1+ru_\dagger)^{1/r+2}(1+u_\dagger)}, \end{aligned}$$

where $B_3 = u_\dagger^2(1+ru_\dagger)^2 - r^2u^2u_\dagger - ru^2(1+u_\dagger)$. The last inequality in the above display holds due to the application of the Bernoulli inequality for $0 < r \leq 1$ and $u_\dagger(1+ru_\dagger)^2 - r^2u^2 > 0$ for $u < u_\dagger$. Since $0 < r \leq 1$, $u_\dagger(1+ru_\dagger)^2 - r^2u^2 \geq u_\dagger(1+ru_\dagger)^2 - ru_\dagger^2 = u_\dagger + r^2u_\dagger^3 + 2r(u_\dagger^2 - u^2) > 0$. Now, I have

$$\begin{aligned} B_3 &= u_\dagger^2(1+ru_\dagger)^2 - r^2u^2u_\dagger - ru^2(1+u_\dagger) \\ &= u_\dagger^2 + r^2u_\dagger^4 + 2ru_\dagger^3 - r^2u^2u_\dagger - ru^2 - ru^2u_\dagger \\ &= u_\dagger^2\left\{1 - r\left(\frac{u}{u_\dagger}\right)^2\right\} + r^2u_\dagger^4 + ru_\dagger^3\left\{2 - (1+r)\left(\frac{u}{u_\dagger}\right)^2\right\}. \end{aligned}$$

Since $u \leq u_\dagger$, $u/u_\dagger \leq 1$ and consequently $\{1 - r(u/u_\dagger)^2\} \geq 0$ and $2 - (1+r)(u/u_\dagger)^2 \geq 0$ for $r \in (0, 1]$. Hence, $B_3 > 0$ and so are B_2 and B_1. Since $A_2(u_0) > 0$ I have $B_1 + 2A_2(u_0) > 0$ and

$$g^{'}(u) = (u - u_0)\left\{\frac{1}{ru^2} - 2A_2(u_\dagger) + 2A_2(u_0)\right\} < 0$$

for $0 < r \leq 1$ and $u \leq u_0$. This proves $g(u)$ is decreasing for $u \leq u_0$. Note that $g(u_0) = 0$, so $g(u) \geq 0$ for $u \leq u_0$, and together with (A.1) I have $f(u) \geq (u-u_0)A_1(u_0) - (u-u_0)^2A_2(u_0) + (1/r)\{\log(u_0/u) + 1 - u_0/u\}$ for $0 < r \leq 1$.

Next consider the case of $r > 1$ with $\kappa = 1$. Here $g^{'}(u) = (u-u_0)\{-2A_2(u_\dagger) + 2A_2(u_0) + 1/u^2\}$. Our goal is to show $g^{'}(u) < 0$. To prove this it is sufficient to show

$$B_5 = \frac{1}{u_\dagger^2} - 2A_2(u_\dagger) = \frac{1}{u_\dagger^2} - \frac{(1+ru_\dagger)^{-1/r-2}[1 + r\{1 - (1+ru_\dagger)^{-1/r}\}]}{\{1 - (1+ru_\dagger)^{-1/r}\}^2} > 0, \qquad \text{(A.2)}$$

for $u \leq u_0$ because $g^{'}(u) = (u-u_0)\{B_5 + 2A_2(u_0)\}$ and $A_2(u_0) > 0$. Now, consider the following transformation of variable, $t = (1+ru_\dagger)^{1/r}$, so $u_\dagger = (t^r - 1)/r$. Then, showing inequality (A.2)

is equivlent to show the following inequality,

$$\begin{aligned}
&\frac{r^2}{(t^r-1)^2} - \frac{(1+r)t-r}{t^{2r}(t-1)^2} > 0, \quad \forall r>1, t>1 \\
\iff\ & r^2t^{2r}(t-1)^2 - \{(1+r)t-r\}(t^r-1)^2 > 0 \\
\iff\ & r^2(t-1)^2t^{2r} > (t^r-1)^2(t+rt-r) \\
\iff\ & 2\log(r) + 2\log(t-1) + 2r\log(t) - 2\log(t^r-1) > \log(t+rt-r). \qquad \text{(A.3)}
\end{aligned}$$

Obviously $\lim_{t\to 1^+}\log(t+rt-r)=0$, and

$$\lim_{t\to 1^+}\{2\log(t-1)-2\log(t^r-1)\} = 2\lim_{t\to 1^+}\log\left(\frac{t-1}{t^r-1}\right) = 2\log\left(\frac{1}{r}\right).$$

Therefore, $\lim_{t\to 1^+}\{2\log(r)+2\log(t-1)+2r\log(t)-2\log(t^r-1)\}=0$. I thus have

$$\begin{aligned}
& 2\log(r) + 2\log(t-1) + 2r\log(t) - 2\log(t^r-1) \\
=\ & \int_1^t \frac{\partial\{2\log(r)+2\log(s-1)+2r\log(s)-2\log(s^r-1)\}}{\partial s},
\end{aligned}$$

and

$$\log(t+rt-r) = \int_1^t \frac{\partial\{\log(s+rs-r)\}}{\partial s}.$$

Then, to prove (A.3), it suffices to show

$$
\begin{aligned}
& \frac{\partial\{2\log(r)+2\log(t-1)+2r\log(t)-2\log(t^r-1)\}}{\partial t} > \frac{\partial \log(t+rt-r)}{\partial t}, \quad \forall r>1, t>1 \\
\iff & \frac{2}{t-1}-\frac{2r}{t(t^r-1)} > \frac{r+1}{t+rt-r} \\
\iff & \frac{1}{t-1}-\frac{2r}{t(t^r-1)} > (r+1)\left(\frac{1}{t+rt-r}-\frac{1}{(t-1)(r+1)}\right) \\
\iff & \frac{1}{t-1}-\frac{2r}{t(t^r-1)}+\frac{1}{(t+rt-r)(t-1)} > 0 \\
\iff & \frac{t+rt-r+1}{(t-1)(t+rt-r)} > \frac{2r}{t(t^r-1)} \\
\iff & \frac{(t^r-1)t}{(t-1)r} > \frac{2(t+rt-r)}{t+rt-r+1} \\
\iff & \frac{(t^r-1)t}{(t-1)r} > 1+\frac{(t-1)(r+1)}{t+rt-r+1} \\
\iff & \frac{t^{r+1}-t-tr+r}{(t-1)r} > \frac{(t-1)(r+1)}{t+rt-r+1} \\
\iff & \frac{t^{r+1}-t-tr+r}{(t-1)^2r(r+1)} > \frac{1}{t+rt-r+1} \\
\iff & \frac{(t^{r+1}-1)/(t-1)-(r+1)}{(t-1)r(r+1)} > \frac{1}{t+rt-r+1}. \qquad \text{(A.4)}
\end{aligned}
$$

I now provide two useful statements, the first is

$$
(t^{r+1}-1)/(t-1) = (r+1)\xi_1^r \geq (r+1)\left(\frac{t+1}{2}\right)^r \qquad \text{(A.5)}
$$

where the equality is obtained by the mean value theorem with $\xi_1 \in (1,t)$ and the inequality is obtained by

$$
\xi_1^r = \frac{1}{t-1}\int_1^t s^r ds \geq \left(\frac{1}{t-1}\int_1^t s ds\right)^r = \left(\frac{t+1}{2}\right)^r.
$$

The last inequality is obtained by applying Jensen's inequality and noting x^r is a convex function

for $r > 1$ and any generic $x > 0$. The second is

$$\frac{\{(t+1)/2\}^r - \{(t+1)/2\}^0}{r-0} = \left(\frac{t+1}{2}\right)^{\xi_2} \log\left(\frac{t+1}{2}\right) \geq \left(\frac{t+1}{2}\right)^{r/2} \log\left(\frac{t+1}{2}\right), \quad \text{(A.6)}$$

where the equality is obtained by the mean value theorem with $\xi_2 \in (0, r)$ and the inequality is obtained by

$$\left(\frac{t+1}{2}\right)^{\xi_2} = \frac{1}{r-0}\int_0^r \left(\frac{t+1}{2}\right)^s ds \geq \left(\frac{t+1}{2}\right)^{\frac{1}{r-0}\int_0^r sds} = \left(\frac{t+1}{2}\right)^{r/2},$$

where the inequality is obtained by Jensen's inequality and $h(x) = \{(t+1)/2\}^x$ is a convex function for $t > 1$. Applying inequalities (A.5) and (A.6) to the left hand side of inequality (A.4), I have

$$\frac{(t^{r+1}-1)/(t-1) - (r+1)}{(t-1)r(r+1)} \geq \frac{\{(t+1)/2\}^r - 1}{(t-1)r} \geq \frac{\left(\frac{t+1}{2}\right)^{r/2} \log\left(\frac{t+1}{2}\right)}{t-1}.$$

Then, to prove (A.4), it is sufficient to show

$$\frac{\left(\frac{t+1}{2}\right)^{r/2} \log\left(\frac{t+1}{2}\right)}{t-1} > \frac{1}{t+rt-r+1} \iff \frac{t+1}{t-1} + r > \frac{1}{\left(\frac{t+1}{2}\right)^{r/2} \log\left(\frac{t+1}{2}\right)}. \quad \text{(A.7)}$$

Since $\log(x) \geq 1 - 1/x$ for any generic $x > 0$, I get $\log\{(t+1)/2\} \geq (t-1)/(t+1)$ and using this result to the right hand side of (A.7) I get

$$\frac{1}{\{(1+t)/2\}^{r/2} \log\{(t+1)/2\}} \leq \left(\frac{2}{1+t}\right)^{r/2} \times \left(\frac{t+1}{t-1}\right) < \left(\frac{t+1}{t-1}\right) < r + \left(\frac{t+1}{t-1}\right)$$

where the second last inequality follows as $t > 1$. The last inequality follows as $r > 1$. Hence (A.7) follows. Then the inequality (A.2) holds and $1/u^2 - 2A_2(u_\dagger) + 2A_2(u_0) > 0$ for $r > 1$. Consequently $g'(u) < 0$ for $u \leq u_0$, and then $g(u) \geq g(u_0) = 0$ for $u \leq u_0$ and the desired result is obtained. □

A.1.1.2 Proof of part ii)

Proof. To prove the part ii) of Theorem 2.1, I first define $f(u) = \log[\{1 - \exp(-u)\}/\{1 - \exp(-u_0)\}]$. Observe that $f(u_0) = 0$. Let me consider the Taylor series expansion of $f(u)$ about $u = u_0$

$$f(u) = (u - u_0)A_1(u_0) - (u - u_0)^2 A_2(u^*),$$

for some $u^* \in (u_0, u)$, where

$$A_1(u) = \frac{\partial f(u)}{\partial u} = \frac{\exp(-u)}{1 - \exp(-u)},$$

and

$$A_2(u) = -\frac{1}{2}\frac{\partial^2 f(u)}{\partial u^2} = \frac{\exp(-u)}{2\{1 - \exp(-u)\}} + \frac{\exp(-2u)}{2\{1 - \exp(-u)\}^2} = \frac{\exp(-u)}{2\{1 - \exp(-u)\}^2}.$$

For $u \geq u_0$, $A_2(u_0) = \min_{u \geq u_0} A_2(u)$. Hence, for $u \geq u_0$,

$$\begin{aligned} f(u) &\geq (u - u_0)A_1(u_0) - (u - u_0)^2 A_2(u_0) \\ &= (u - u_0)A_1(u_0) - (u - u_0)^2 A_2(u_0) + \log\left(\frac{u_0}{u}\right) + \log\left(\frac{u}{u_0}\right) \\ &\geq (u - u_0)A_1(u_0) - (u - u_0)^2 A_2(u_0) + \log\left(\frac{u_0}{u}\right) + \left(1 - \frac{u_0}{u}\right). \end{aligned} \tag{A.8}$$

To prove the result when $u \leq u_0$, let me define $g(u) = f(u) - (u - u_0)A_1(u_0) + (u - u_0)^2 A_2(u_0) - \log(u_0/u) - 1 + u_0/u$

$$\begin{aligned} g'(u) &= A_1(u) - A_1(u_0) + 2(u - u_0)A_2(u_0) + \frac{u - u_0}{u^2} \\ &= (u - u_0)\left\{-2A_2(u_\dagger) + 2A_2(u_0) + \frac{1}{u^2}\right\}, \end{aligned}$$

where the last equality is obtained by applying the Taylor series expansion on $A_1(u)$ about $u = u_0$, $A_1(u) = A_1(u_0) - 2(u - u_0)A_2(u_\dagger)$, for some $u_\dagger \in (u, u_0)$. As $A_2(u_0) > 0$, then for $u < u_\dagger$

$$\frac{1}{u^2} - 2A_2(u_\dagger) + 2A_2(u_0) > \frac{1}{u_\dagger^2} - 2A_2(u_\dagger) = \frac{1}{u_\dagger^2} - \frac{1}{\{1 - \exp(-u_\dagger)\}^2 \exp(u_\dagger)} \equiv h(u_\dagger).$$

Let me define

$$k(u_\dagger) \equiv \{1 - \exp(-u_\dagger)\}^2 \exp(u_\dagger) - u_\dagger^2 = \exp(u_\dagger) + \exp(-u_\dagger) - u_\dagger^2 - 2,$$

and investigate its properties. Note that

$$k'(u_\dagger) = \frac{\partial k(u_\dagger)}{\partial u_\dagger} = \exp(u_\dagger) - \exp(-u_\dagger) - 2u_\dagger,$$

and

$$k''(u_\dagger) = \frac{\partial^2 k(u_\dagger)}{\partial u_\dagger^2} = \exp(u_\dagger) + \exp(-u_\dagger) - 2 \geq 0,$$

where the inequality is obtained by applying $\exp(u_\dagger) \geq 1 + u_\dagger$ and $\exp(-u_\dagger) \geq 1 - u_\dagger$. This shows that $k'(u_\dagger)$ is an increasing function, and

$$k'(u_\dagger) > \inf_{u_\dagger > 0} k'(u_\dagger) = \lim_{u_\dagger \to 0} k'(u_\dagger) = 0.$$

This result also implies that $k(u_\dagger)$ is increasing, and $k(u_\dagger) > \inf_{u_\dagger > 0} k(u_\dagger) = \lim_{u_\dagger \to 0} k(u_\dagger) = 0$. Observe that $h(u_\dagger) = k(u_\dagger)/u_\dagger^2\{1 - \exp(-u_\dagger)\}^2 \exp(u_\dagger) > 0$ for $u_\dagger > 0$ which proves $g'(u) < 0$ (g is a decreasing function) for any $u \leq u_0$. Thus, $g(u) \geq \min_{u \leq u_0} g(u) = \lim_{u \to u_0} g(u) = 0$. Hence, $g(u) = f(u) - (u - u_0)A_1(u_0) + (u - u_0)^2 A_2(u_0) - \log(u_0/u) - 1 + u_0/u \geq 0$ for $u \leq u_0$ and combined with (A.8) I now have part ii) of the theorem. □

A.1.2 Proof of inequality (2.11)

This is the derivation of the minorization function for the $r > 0$ case with $\theta > 0$. Using part (i) of Theorem 2.1 to the multiplier of $\Delta_{i,j}$ and result (i) of Lemma 2.1 to the multiplier of $(1 - \Delta_{i,j})$, I have

$$
\begin{aligned}
\ell_i(\boldsymbol{\xi}) \;\geq\; & \ell_i(\boldsymbol{\xi}_0) + \sum_k \omega_i^*(\boldsymbol{\xi}_0, a_k) \sum_{j=1}^{m_i} \Bigg(\Delta_{i,j} \{ A_1(u_{i,j,k}(\boldsymbol{\xi}_0)) + 2A_2(u_{i,j,k}(\boldsymbol{\xi}_0)) u_{i,j,k}(\boldsymbol{\xi}_0) \} \\
& \times u_{i,j,k}(\boldsymbol{\xi}_0) \exp\left[(\boldsymbol{\alpha} - \boldsymbol{\alpha}_0)^\mathsf{T} \boldsymbol{W}_{i,j,k} + \log\left\{ \frac{H_\psi(C_{i,j})}{H_{\psi_0}(C_{i,j})} \right\} \right] \\
& - \Delta_{i,j} A_2(u_{i,j,k}(\boldsymbol{\xi}_0)) u_{i,j,k}^2(\boldsymbol{\xi}_0) \exp\left[2(\boldsymbol{\alpha} - \boldsymbol{\alpha}_0)^\mathsf{T} \boldsymbol{W}_{i,j,k} + 2\log\left\{ \frac{H_\psi(C_{i,j})}{H_{\psi_0}(C_{i,j})} \right\} \right] \\
& - (1 - \Delta_{i,j}) \frac{u_{i,j,k}(\boldsymbol{\xi}_0)}{1 + r u_{i,j,k}(\boldsymbol{\xi}_0)} \exp\left[(\boldsymbol{\alpha} - \boldsymbol{\alpha}_0)^\mathsf{T} \boldsymbol{W}_{i,j,k} + \log\left\{ \frac{H_\psi(C_{i,j})}{H_{\psi_0}(C_{i,j})} \right\} \right] \\
& - \Delta_{i,j} \kappa \exp\left[(\boldsymbol{\alpha}_0 - \boldsymbol{\alpha})^\mathsf{T} \boldsymbol{W}_{i,j,k} + \log\left\{ \frac{H_{\psi_0}(C_{i,j})}{H_\psi(C_{i,j})} \right\} \right] \\
& - \Delta_{i,j} \kappa \left[\log\{ H_\psi(C_{i,j}) \} + \boldsymbol{\alpha}^\mathsf{T} \boldsymbol{W}_{i,j,k} \right] \\
& - \Delta_{i,j} A_1(u_{i,j,k}(\boldsymbol{\xi}_0)) u_{i,j,k}(\boldsymbol{\xi}_0) - \Delta_{i,j} A_2(u_{i,j,k}(\boldsymbol{\xi}_0)) u_{i,j,k}^2(\boldsymbol{\xi}_0) + (1 - \Delta_{i,j}) \frac{u_{i,j,k}(\boldsymbol{\xi}_0)}{1 + r u_{i,j,k}(\boldsymbol{\xi}_0)} \\
& + \Delta_{i,j} \kappa + \Delta_{i,j} \kappa \log\{ u_{i,j,k}(\boldsymbol{\xi}_0) \} \Bigg) \\
\geq\; & \ell_i(\boldsymbol{\xi}_0) + \sum_k \omega_i^*(\boldsymbol{\xi}_0, a_k) \sum_{j=1}^{m_i} \Bigg(\Delta_{i,j} \{ A_1(u_{i,j,k}(\boldsymbol{\xi}_0)) + 2A_2(u_{i,j,k}(\boldsymbol{\xi}_0)) u_{i,j,k}(\boldsymbol{\xi}_0) \} \\
& \times u_{i,j,k}(\boldsymbol{\xi}_0) \left[1 + (\boldsymbol{\alpha} - \boldsymbol{\alpha}_0)^\mathsf{T} \boldsymbol{W}_{i,j,k} + \log\left\{ \frac{H_\psi(C_{i,j})}{H_{\psi_0}(C_{i,j})} \right\} \right] \\
& - \Delta_{i,j} A_2(u_{i,j,k}(\boldsymbol{\xi}_0)) u_{i,j,k}^2(\boldsymbol{\xi}_0) \left(\frac{1}{2} \right) \left[\exp\{ 4(\boldsymbol{\alpha} - \boldsymbol{\alpha}_0)^\mathsf{T} \boldsymbol{W}_{i,j,k} \} + \left\{ \frac{H_\psi(C_{i,j})}{H_{\psi_0}(C_{i,j})} \right\}^4 \right] \\
& - (1 - \Delta_{i,j}) \frac{u_{i,j,k}(\boldsymbol{\xi}_0)}{1 + r u_{i,j,k}(\boldsymbol{\xi}_0)} \left(\frac{1}{2} \right) \left[\exp\{ 2(\boldsymbol{\alpha} - \boldsymbol{\alpha}_0)^\mathsf{T} \boldsymbol{W}_{i,j,k} \} + \left\{ \frac{H_\psi(C_{i,j})}{H_{\psi_0}(C_{i,j})} \right\}^2 \right] \\
& - \Delta_{i,j} \left(\frac{1}{2} \right) \kappa \left[\exp\{ 2(\boldsymbol{\alpha}_0 - \boldsymbol{\alpha})^\mathsf{T} \boldsymbol{W}_{i,j,k} \} + \left\{ \frac{H_{\psi_0}(C_{i,j})}{H_\psi(C_{i,j})} \right\}^2 \right] \\
& - \Delta_{i,j} \kappa \left[\log\{ H_\psi(C_{i,j}) \} + \boldsymbol{\alpha}^\mathsf{T} \boldsymbol{W}_{i,j,k} \right]
\end{aligned}
$$

$$
\begin{aligned}
&-\Delta_{i,j}A_1(u_{i,j,k}(\boldsymbol{\xi}_0))u_{i,j,k}(\boldsymbol{\xi}_0) - \Delta_{i,j}A_2(u_{i,j,k}(\boldsymbol{\xi}_0))u^2_{i,j,k}(\boldsymbol{\xi}_0) + (1-\Delta_{i,j})\frac{u_{i,j,k}(\boldsymbol{\xi}_0)}{1+ru_{i,j,k}(\boldsymbol{\xi}_0)} \\
&+\Delta_{i,j}\kappa + \Delta_{i,j}\kappa\log\{u_{i,j,k}(\boldsymbol{\xi}_0)\}\Big) \\
= \;& \ell_{\dagger,i}(\boldsymbol{\xi}|\boldsymbol{\xi}_0).
\end{aligned}
$$

The last inequality is obtained by using result (ii) and (iii) of Lemma 2.1. Thus $\ell_{\dagger,i}$ can be further written as $\ell_{\dagger,i}(\boldsymbol{\xi}|\boldsymbol{\xi}_0) = \ell_{\dagger,1,i}(\boldsymbol{\alpha}|\boldsymbol{\xi}_0) + \ell_{\dagger,2,i}(\boldsymbol{\psi}|\boldsymbol{\xi}_0) + \ell_{\dagger,3,i}(\boldsymbol{\xi}_0)$.

A.1.3 Detailed derivation of Section 2.3.3

This is the details of the non-dependence case ($\theta = 0$). Note that here

$$
\ell(\boldsymbol{\xi}) \;=\; \sum_{i=1}^{n}\ell_i(\boldsymbol{\xi}) = \ell(\boldsymbol{\xi}_0) + \sum_{i=1}^{n}\log\left[\frac{\{1-G_i(\boldsymbol{\xi})\}^{\Delta_i}\{G_i(\boldsymbol{\xi})\}^{1-\Delta_i}}{\{1-G_i(\boldsymbol{\xi}_0)\}^{\Delta_i}\{G_i(\boldsymbol{\xi}_0)\}^{1-\Delta_i}}\right].
$$

Then for $r > 0$, using the actual expressions of $G_i(\boldsymbol{\xi})$ and $G_i(\boldsymbol{\xi}_0)$, I have

$$
\ell(\boldsymbol{\xi}) \;=\; \ell(\boldsymbol{\xi}_0) + \sum_{i=1}^{n}\left(\Delta_i\log\left[\frac{1-\{1+ru_i(\boldsymbol{\xi})\}^{-1/r}}{1-\{1+ru_i(\boldsymbol{\xi}_0)\}^{-1/r}}\right] + (1-\Delta_i)\kappa\log\left\{\frac{1+ru_i(\boldsymbol{\xi})}{1+ru_i(\boldsymbol{\xi}_0)}\right\}\right).
$$

Using the same inequalities and techniques in Section A.1.2, I first obtain the minorization function $\ell_\dagger(\boldsymbol{\xi}|\boldsymbol{\xi}_0)$, such that $\ell(\boldsymbol{\xi}) \geq \ell_\dagger(\boldsymbol{\xi}|\boldsymbol{\xi}_0) \equiv \ell_{\dagger,1}(\boldsymbol{\alpha}|\boldsymbol{\xi}_0) + \ell_{\dagger,2}(\boldsymbol{\psi}|\boldsymbol{\xi}_0) + \ell_{\dagger,3}(\boldsymbol{\xi}_0)$, where

$$
\begin{aligned}
\ell_{\dagger,1}(\boldsymbol{\alpha}|\boldsymbol{\xi}_0) \;=\; & \sum_{i=1}^{n}\Bigg[\Delta_i\Big\{A_1(u_i(\boldsymbol{\xi}_0)) + 2A_2(u_i(\boldsymbol{\xi}_0))u_i(\boldsymbol{\xi}_0)\Big\}u_i(\boldsymbol{\xi}_0)(\boldsymbol{\alpha}-\boldsymbol{\alpha}_0)^{\mathsf{T}}\boldsymbol{X}_i \\
& -\left(\frac{\Delta_i}{2}\right)A_2(u_i(\boldsymbol{\xi}_0))u_i^2(\boldsymbol{\xi}_0)\exp\{4(\boldsymbol{\alpha}-\boldsymbol{\alpha}_0)^{\mathsf{T}}\boldsymbol{X}_i\} \\
& -\left(\frac{1-\Delta_i}{2}\right)\frac{u_i(\boldsymbol{\xi}_0)}{1+ru_i(\boldsymbol{\xi}_0)}\exp\{2(\boldsymbol{\alpha}-\boldsymbol{\alpha}_0)^{\mathsf{T}}\boldsymbol{X}_i\} \\
& -\left(\frac{\Delta_i\kappa}{2}\right)\exp\{2(\boldsymbol{\alpha}_0-\boldsymbol{\alpha})^{\mathsf{T}}\boldsymbol{X}_i\} - \Delta_i\kappa\boldsymbol{\alpha}^{\mathsf{T}}\boldsymbol{X}_i\Bigg],
\end{aligned}
$$

$$
\begin{aligned}
\ell_{\dagger,2}(\boldsymbol{\psi}|\boldsymbol{\xi}_0) &= \sum_{i=1}^{n}\Bigg[\Delta_i\Big\{A_1(u_i(\boldsymbol{\xi}_0)) + 2A_2(u_i(\boldsymbol{\xi}_0))u_i(\boldsymbol{\xi}_0)\Big\} \\
&\quad \times u_i(\boldsymbol{\xi}_0)\log\left\{\frac{H_\psi(C_i)}{H_{\psi_0}(C_i)}\right\} - \left(\frac{\Delta_i}{2}\right)A_2(u_i(\boldsymbol{\xi}_0))u_i^2(\boldsymbol{\xi}_0)\left\{\frac{H_\psi(C_i)}{H_{\psi_0}(C_i)}\right\}^4 \\
&\quad - \left(\frac{1-\Delta_i}{2}\right)\frac{u_i(\boldsymbol{\xi}_0)}{1+ru_i(\boldsymbol{\xi}_0)}\left\{\frac{H_\psi(C_i)}{H_{\psi_0}(C_i)}\right\}^2 \\
&\quad - \left(\frac{\Delta_i\kappa}{2}\right)\left\{\frac{H_{\psi_0}(C_i)}{H_\psi(C_i)}\right\}^2 - \Delta_i\kappa\log\{H_\psi(C_i)\}\Bigg], \\
\ell_{\dagger,3}(\boldsymbol{\xi}_0) &= \ell(\boldsymbol{\xi}_0) + \sum_{i=1}^{n}\Bigg(\Delta_i A_2(u_i(\boldsymbol{\xi}_0))u_i^2(\boldsymbol{\xi}_0) + (1-\Delta_i)\frac{u_i(\boldsymbol{\xi}_0)}{1+ru_i(\boldsymbol{\xi}_0)} \\
&\quad +\Delta_i\kappa\left[1+\log\{u_i(\boldsymbol{\xi}_0)\}\right]\Bigg).
\end{aligned}
$$

Then, $\boldsymbol{\alpha}$ and $\boldsymbol{\psi}$ are estimated by the generic Newton-Raphson algorithm given in (2.12), where the needed quantities are

$$
\begin{aligned}
S(\boldsymbol{\alpha}^{(m-1)}|\boldsymbol{\xi}^{(m-1)}) &= \sum_{i=1}^{n}\left\{\Delta_i A_1(u_i(\boldsymbol{\xi}^{(m-1)})) - \frac{(1-\Delta_i)}{1+ru_i(\boldsymbol{\xi}^{(m-1)})}\right\}u_i(\boldsymbol{\xi}^{(m-1)})\boldsymbol{X}_i, \\
S_\alpha(\boldsymbol{\alpha}^{(m-1)}|\boldsymbol{\xi}^{(m-1)}) &= -\sum_{i=1}^{n}\Bigg[8\Delta_i A_2(u_i(\boldsymbol{\xi}^{(m-1)}))u_i^2(\boldsymbol{\xi}^{(m-1)}) \\
&\quad +2\left(1-\Delta_i\right)\frac{u_i(\boldsymbol{\xi}^{(m-1)})}{1+ru_i(\boldsymbol{\xi}^{(m-1)})} + 2\Delta_i\kappa\Bigg]\boldsymbol{X}_i^{\otimes 2}, \\
S(\boldsymbol{\psi}^{(m-1)}|\boldsymbol{\xi}^{(m-1)}) &= \sum_{i=1}^{n}\Bigg\{\Delta_i A_1(u_i(\boldsymbol{\xi}^{(m-1)})) \\
&\quad -\frac{(1-\Delta_i)}{1+ru_i(\boldsymbol{\xi}^{(m-1)})}\Bigg\}u_i(\boldsymbol{\xi}^{(m-1)})\left[\frac{\partial\log\{H_\psi(C_i)\}}{\partial\boldsymbol{\psi}}\right]_{\psi=\psi^{(m-1)}}, \\
S_\psi(\boldsymbol{\psi}^{(m-1)}|\boldsymbol{\xi}^{(m-1)}) &= \sum_{i=1}^{n}\left\{\Delta_i A_1(u_i(\boldsymbol{\xi}^{(m-1)}))u_i(\boldsymbol{\xi}^{(m-1)}) - (1-\Delta_i)\frac{u_i(\boldsymbol{\xi}^{(m-1)})}{1+ru_i(\boldsymbol{\xi}^{(m-1)})}\right\} \\
&\quad \times\left[\frac{\partial^2\log\{H_\psi(C_i)\}}{\partial\boldsymbol{\psi}\partial\boldsymbol{\psi}^{\mathsf{T}}}\right]_{\psi=\psi^{(m-1)}} \\
&\quad -\sum_{i=1}^{n}\left\{8\Delta_i A_2(u_i(\boldsymbol{\xi}^{(m-1)}))u_i^2(\boldsymbol{\xi}^{(m-1)}) + 2\left(1-\Delta_i\right)\frac{u_i(\boldsymbol{\xi}^{(m-1)})}{1+ru_i(\boldsymbol{\xi}^{(m-1)})} + 2\Delta_i\kappa\right\} \\
&\quad \times\left(\left[\frac{\partial\log\{H_\psi(C_i)\}}{\partial\boldsymbol{\psi}}\right]_{\psi=\psi^{(m-1)}}\right)^{\otimes 2}.
\end{aligned}
$$

On the other hand, for $r = 0$, the minorization function is the total of the following terms

$$\begin{aligned}
\ell_{\dagger,1}(\boldsymbol{\alpha}|\boldsymbol{\xi}_0) &= \sum_{i=1}^{n}\Bigg[\Delta_i\Big\{A_3(u_i(\boldsymbol{\xi}_0)) + 2A_4(u_i(\boldsymbol{\xi}_0))u_i(\boldsymbol{\xi}_0)\Big\} \times u_i(\boldsymbol{\xi}_0)(\boldsymbol{\alpha}-\boldsymbol{\alpha}_0)^{\mathsf{T}}\boldsymbol{X}_i \\
&\quad - \left(\frac{\Delta_i}{2}\right) A_4(u_i(\boldsymbol{\xi}_0))u_i^2(\boldsymbol{\xi}_0)\exp\{4(\boldsymbol{\alpha}-\boldsymbol{\alpha}_0)^{\mathsf{T}}\boldsymbol{X}_i\} - \left(\frac{1-\Delta_i}{2}\right) u_i(\boldsymbol{\xi}_0)\exp\{2(\boldsymbol{\alpha}-\boldsymbol{\alpha}_0)^{\mathsf{T}}\boldsymbol{X}_i\} \\
&\quad - \left(\frac{\Delta_i}{2}\right)\exp\{2(\boldsymbol{\alpha}_0-\boldsymbol{\alpha})^{\mathsf{T}}\boldsymbol{X}_i\} - \Delta_i\boldsymbol{\alpha}^{\mathsf{T}}\boldsymbol{X}_i\Bigg], \\
\ell_{\dagger,2}(\boldsymbol{\psi}|\boldsymbol{\xi}_0) &= \sum_{i=1}^{n}\Bigg[\Delta_i\Big\{A_3(u_i(\boldsymbol{\xi}_0)) + 2A_4(u_i(\boldsymbol{\xi}_0))u_i(\boldsymbol{\xi}_0)\Big\} \times u_i(\boldsymbol{\xi}_0)\log\left\{\frac{H_\psi(C_i)}{H_{\psi_0}(C_i)}\right\} \\
&\quad - \left(\frac{\Delta_i}{2}\right) A_4(u_i(\boldsymbol{\xi}_0))u_i^2(\boldsymbol{\xi}_0)\left\{\frac{H_\psi(C_i)}{H_{\psi_0}(C_i)}\right\}^4 - \left(\frac{1-\Delta_i}{2}\right) u_i(\boldsymbol{\xi}_0)\left\{\frac{H_\psi(C_i)}{H_{\psi_0}(C_i)}\right\}^2 \\
&\quad - \left(\frac{\Delta_i}{2}\right)\left\{\frac{H_{\psi_0}(C_i)}{H_\psi(C_i)}\right\}^2 - \Delta_i\log\{H_\psi(C_i)\}\Bigg], \\
\ell_{\dagger,3}(\boldsymbol{\xi}_0) &= \ell(\boldsymbol{\xi}_0) + \sum_{i=1}^{n}\Bigg(\Delta_i A_4(u_i(\boldsymbol{\xi}_0))u_i^2(\boldsymbol{\xi}_0) + (1-\Delta_i)u_i(\boldsymbol{\xi}_0) + \Delta_i\left[1+\log\{u_i(\boldsymbol{\xi}_0)\}\right]\Bigg),
\end{aligned}$$

and the terms needed in the Newton-Raphson algorithm (2.12) are

$$\begin{aligned}
S(\boldsymbol{\alpha}^{(m-1)}|\boldsymbol{\xi}^{(m-1)}) &= \sum_{i=1}^{n}\Big\{\Delta_i A_3(u_i(\boldsymbol{\xi}^{(m-1)})) - (1-\Delta_i)\Big\}u_i(\boldsymbol{\xi}^{(m-1)})\boldsymbol{X}_i, \\
S_\alpha(\boldsymbol{\alpha}^{(m-1)}|\boldsymbol{\xi}^{(m-1)}) &= -\sum_{i=1}^{n}\Big[8\Delta_i A_4(u_i(\boldsymbol{\xi}^{(m-1)}))u_i^2(\boldsymbol{\xi}^{(m-1)}) + 2\left(1-\Delta_i\right)u_i(\boldsymbol{\xi}^{(m-1)}) + 2\Delta_i\Big]\boldsymbol{X}_i^{\otimes 2}, \\
S(\boldsymbol{\psi}^{(m-1)}|\boldsymbol{\xi}^{(m-1)}) &= \sum_{i=1}^{n}\Big\{\Delta_i A_3(u_i(\boldsymbol{\xi}^{(m-1)})) - (1-\Delta_i)\Big\}u_{i,j}(\boldsymbol{\xi}^{(m-1)})\left[\frac{\partial\log\{H_\psi(C_i)\}}{\partial\boldsymbol{\psi}}\right]_{\psi=\psi^{(m-1)}}, \\
S_\psi(\boldsymbol{\psi}^{(m-1)}|\boldsymbol{\xi}^{(m-1)}) &= \sum_{i=1}^{n}\Big\{\Delta_i A_3(u_i(\boldsymbol{\xi}^{(m-1)})) - (1-\Delta_i)\Big\}u_i(\boldsymbol{\xi}^{(m-1)})\left[\frac{\partial^2\log\{H_\psi(C_i)\}}{\partial\boldsymbol{\psi}\partial\boldsymbol{\psi}^{\mathsf{T}}}\right]_{\psi=\psi^{(m-1)}} \\
&\quad - \sum_{i=1}^{n}\Big\{8\Delta_i A_4(u_i(\boldsymbol{\xi}^{(m-1)}))u_i^2(\boldsymbol{\xi}^{(m-1)}) + 2\left(1-\Delta_i\right)u_i(\boldsymbol{\xi}^{(m-1)}) + 2\Delta_i\Big\} \\
&\quad \times\left(\left[\frac{\partial\log\{H_\psi(C_i)\}}{\partial\boldsymbol{\psi}}\right]_{\psi=\psi^{(m-1)}}\right)^{\otimes 2}.
\end{aligned}$$

Since $H_\psi(C_i) = \sum_l M_l(C_i)\exp(\psi_l), \partial H_\psi(C_i)/\partial\psi_l = M_l(C_i)\exp(\psi_l)$, let me write $\partial H_\psi/\partial\boldsymbol{\psi} = \boldsymbol{D}_i\exp(\boldsymbol{\psi})$, where $\boldsymbol{D}_i = \mathrm{Diag}(M_1(C_i),\ldots,M_K(C_i))$ and $\exp(\boldsymbol{\psi}) = (\exp(\psi_1),\ldots,\exp(\psi_K))^{\mathsf{T}}$.

Then I have

$$\left[\frac{\partial^2 \log\{H_\psi(C_i)\}}{\partial\boldsymbol{\psi}\partial\boldsymbol{\psi}^\mathsf{T}}\right]_{\psi=\psi^{(m-1)}} = \left[\frac{\boldsymbol{D}_i \mathrm{Diag}(\exp(\boldsymbol{\psi}^{(m-1)}))}{H_{\psi^{(m-1)}}} - \frac{\boldsymbol{D}_i \exp(\boldsymbol{\psi}^{(m-1)})\{\exp(\boldsymbol{\psi}^{(m-1)})\}^\mathsf{T}\boldsymbol{D}_i}{H^2_{\psi^{(m-1)}}}\right],$$

and

$$\left(\left[\frac{\partial \log\{H_\psi(C_i)\}}{\partial\boldsymbol{\psi}}\right]_{\psi=\psi^{(m-1)}}\right)^{\otimes 2} = \frac{\boldsymbol{D}_i \exp(\boldsymbol{\psi}^{(m-1)})\{\exp(\boldsymbol{\psi}^{(m-1)})\}^\mathsf{T}\boldsymbol{D}_i}{H^2_{\psi^{(m-1)}}(C_i)}.$$

A.2 Results of Chapter 2.4

A.2.1 Background

Notations:

In order to prove the main theorems more clearly, I first assume the subject specific random effect b is observed, and investigate the asymptotic properties of the penalized complete ML estimator. The rate of convergence (Theorem 2.2) and semiparametric efficiency (Theorem 2.3) of the penalized observed ML estimator (2.4) can be proved with the similar arguments and presented at Subsection A.2.6.

Define $\boldsymbol{O}_* = (C_{*,1}, \ldots, C_{*,m_*}, \Delta_{*,1}, \ldots, \Delta_{*,m_*}\boldsymbol{X}^\mathsf{T}_{*,1}, \ldots, \boldsymbol{X}^\mathsf{T}_{*,m_*}, \boldsymbol{Z}_*)^\mathsf{T}$ as the observed data from a random cluster $*$, where m_* is the cluster size. I also let $\Pr_{\boldsymbol{\iota}}$ be the distribution of the complete data $\boldsymbol{g} = (\boldsymbol{O}_*, b_*)^\mathsf{T}$ from a random cluster $*$ under the parameter vector $\boldsymbol{\iota}$, and $p_{\boldsymbol{\iota}}$ be the corresponding density with the dominating measure μ. For simplicity, I define $\Pr_0 \equiv \Pr_{\boldsymbol{\iota}_0}$ and $p_0 \equiv p_{\boldsymbol{\iota}_0}$. Specifically, let $\mathcal{L}_c(\boldsymbol{\iota}; \boldsymbol{g})$ and $\ell_c(\boldsymbol{\iota}; \boldsymbol{g})$ be the likelihood and log-likelihood for one single

complete observation, respectively. In other words,

$$\begin{aligned}\mathcal{L}_c(\boldsymbol{\iota};\boldsymbol{g}) &= \prod_{j=1}^{m_*}\Big\{1-S(C_{*,j}|\boldsymbol{X}_{*,j},\boldsymbol{Z}_*,b_*)\Big\}^{\Delta_{*,j}}\Big\{S(C_{*,j}|\boldsymbol{X}_{*,j},\boldsymbol{Z}_*,b_*)\Big\}^{1-\Delta_{*,j}}\phi(b_*)\\ &= \phi(b_*)\prod_{j=1}^{m_*}\left(1-\Big[1+rH(C_{*,j})\exp\{\boldsymbol{\beta}^\mathsf{T}\boldsymbol{X}_{*,j}+\boldsymbol{\gamma}^\mathsf{T}\boldsymbol{Z}_*+\theta b_*\}\Big]^{-1/r}\right)^{\Delta_{*,j}} \qquad \text{(A.9)}\\ &\quad\times\left(\Big[1+rH(C_{*,j})\exp\{\boldsymbol{\beta}^\mathsf{T}\boldsymbol{X}_{*,j}+\boldsymbol{\gamma}^\mathsf{T}\boldsymbol{Z}_*+\theta b_*\}\Big]^{-1/r}\right)^{1-\Delta_{*,j}}.\end{aligned}$$

Here I present the asymptotic properties of the penalized estimator when $r>0$, the result for $r=0$ can be similarly obtained with the change of the expression $S(C_{*,j}|\boldsymbol{X}_{*,j},\boldsymbol{Z}_*,b_*)$ in (A.9). Analogous to (2.4), I also define the penalized complete ML estimator as

$$\begin{aligned}\widehat{\boldsymbol{\iota}}_{c,n} &= (\widehat{\boldsymbol{\alpha}}_{c,n}{}^\mathsf{T},\widehat{H}_{c,n})^\mathsf{T}\\ &= \underset{(\boldsymbol{\alpha}^\mathsf{T},\sum_{k=1}^K M_k(t)\exp(\psi_k))^\mathsf{T}}{\arg\min}\left(\frac{1}{n}\sum_{i=1}^n \ell_c\Big\{\boldsymbol{\alpha},\sum_{k=1}^K M_k(t)\exp(\psi_k);\boldsymbol{g}_i\Big\}\right. \qquad \text{(A.10)}\\ &\qquad\left.-\lambda\int_0^{T_0}\Big[\Big\{\sum_{k=1}^K M_k(t)\exp(\psi_k)\Big\}^{(q)}\Big]^2\mathrm{d}t\right).\end{aligned}$$

To study the space spanned by $\{M_k(t)\}$, I let $\mathcal{S}_n(\boldsymbol{\tau}_n,L_n,d-1)$ denote the space of polynomial splines spanned by degree $d-1$ B-spline basis with knots $\boldsymbol{\tau}_n=\{\tau_1,\tau_2,\ldots,\tau_L\}$ where $0=\tau_0<\tau_1<\tau_2<\cdots<\tau_L<\tau_{L+1}=T_0$, $L\equiv L_n=O(n^{1/(2q+1)})$, with $d\geq q$. Furthermore, it is desirable to restrict the knots such that $\max_{0\leq l\leq L}|\tau_{l+1}-\tau_l|=O(n^{-1/(2q+1)})$ as in Stone (1985). I also let $\mathcal{H}_n(\boldsymbol{\tau}_n,L_n,d)$ denote the space of polynomial splines spanned by d-degree I-spline basis, such that each basis function in $\mathcal{H}_n(\boldsymbol{\tau}_n,L_n,d)$ is the integration of the corresponding basis function in $\mathcal{S}_n(\boldsymbol{\tau}_n,L_n,d-1)$ over the domain $[0,T_0]$, and that all the coefficients are positive. In other words,

$$\begin{aligned}\mathcal{H}_n(\boldsymbol{\tau}_n,L_n,d)=\Bigg\{&\sum_{k=1}^K M_k(t)\exp(\psi_k):\ M_k(t)=\int_0^t\mathcal{B}_k(s)\mathrm{d}s,\\ &\mathcal{B}_k(s)\text{ is a basis function of }\mathcal{S}_n(\boldsymbol{\tau}_n,L_n,d-1),\ k=1,\ldots,K\Bigg\},\end{aligned}$$

where $K = L + d$. It is shown in de Boor (1978) that $\mathcal{H}_n(\boldsymbol{\tau}_n, L_n, d) \subset \mathcal{S}_n(\boldsymbol{\tau}_n, L_n, d)$. To simplify the notations, I also denote $\boldsymbol{\varphi} = \exp(\boldsymbol{\psi})$ with positive values, i.e., $\varphi_k = \exp(\psi_k)$, $k = 1, \ldots, K$. I first note that for a fixed n, letting the tuning parameter $\lambda \to 0$ implies an unpenalized estimate in the space spanned by the given polynomial space. On the other hand, letting $\lambda \to \infty$ forces convergence of the qth derivative of the spline function to zero. For example, when $q = 3$, the limiting transformation function will be quadratic with respect to t.

I introduce some further notations to be used in proving results. Given a random sample $\boldsymbol{g}_1, \ldots, \boldsymbol{g}_n$ with the probability measure $\Pr$, for a measurable function f, define $\Pr f = \int f \mathrm{d} \Pr$ as the expectation of f under $\Pr$ and $\mathbb{P}_n f = (1/n) \sum_{i=1}^n f(\boldsymbol{g}_i)$ as the expectation of f under the empirical measure $\mathbb{P}_n$. I write $\mathbb{G}_n f = \sqrt{n}(\mathbb{P}_n - \Pr) f$ for the empirical process $\mathbb{G}_n$ evaluated at f. Denote $\|\mathbb{G}_n\|_{\mathcal{F}} = \sup_{f \in \mathcal{F}} |\mathbb{G}_n f|$. Let $\| \cdot \|$ and $\| \cdot \|_\infty$ be the Euclidean norm of $\mathbb{R}^p$ and the supremum norm, respectively. I will use v to denote a generic constant that may change values from context to context. For two sequences $\{a_{1,n}\}$ and $\{a_{2,n}\}$, I let $a_{1,n} \asymp a_{2,n}$ denote $a_{1,n} = O(a_{2,n})$ and $a_{2,n} = O(a_{1,n})$, simultaneously.

Regularity conditions: Here I present the regularity conditions that are required to study the asymptotic properties of the regularized semiparametric ML estimator.

(C1) The cluster size m_* of a random cluster is completely random, and uniformly bounded above. In addition $\Pr(m_* \geq 1) > 0$.

(C2) The covariates $(\boldsymbol{X}_{*,1}^\mathsf{T}, \ldots, \boldsymbol{X}_{*,m_*}^\mathsf{T}, \boldsymbol{Z}_*^\mathsf{T})^\mathsf{T}$ are uniformly bounded, that is, there exists a scalar v such that $\Pr\{\|(\boldsymbol{X}_{*,1}^\mathsf{T}, \ldots, \boldsymbol{X}_{*,m_*}^\mathsf{T}, \boldsymbol{Z}_*^\mathsf{T})\| \leq v\} = 1$, where $\| \cdot \|$ denotes Euclidean norm. Moreover, all the eigenvalues of $E\Big[\big\{(\boldsymbol{X}_{*,1}^\mathsf{T}, \ldots, \boldsymbol{X}_{*,m_*}^\mathsf{T}, \boldsymbol{Z}_*^\mathsf{T}, b_*)^\mathsf{T}\big\}^{\otimes 2}\Big]$ are bounded away from zero and infinity, where $\boldsymbol{a}^{\otimes 2} = \boldsymbol{a}\boldsymbol{a}^\mathsf{T}$ denotes the gram matrix for any generic vector $\boldsymbol{a}$.

(C3) The conditional joint density of $(\boldsymbol{O}_*|b_*)$ has uniform positive lower and upper bound in the support of the joint random variables $\boldsymbol{O}_*$.

(C4) The L_∞ norm of the true transformation function $H_0(t)$ is bounded away from 0 and ∞. Moreover, $H_0(\cdot)$ belongs to $\mathcal{H}$, a class of non-negative and monotonic functions, with zero

values at $t = 0$ which are also continuously differentiable up to order q, $d \geq q \geq 2$, on $[0, T_0]$.

(C5) Θ is a compact subset of $\mathbb{R}^p$, where p is the dimensionality of $\boldsymbol{\alpha}$. Furthermore, $\boldsymbol{\alpha}_0$ is an interior point of Θ.

(C6) For any cluster size m_*, there exits some $\kappa \in (0, 1)$, such that

$$\begin{aligned}
&\boldsymbol{a}^{\mathsf{T}} \mathrm{var}\Big\{(\boldsymbol{X}_{*,1}^{\mathsf{T}}, \ldots, \boldsymbol{X}_{*,m_*}^{\mathsf{T}}, \boldsymbol{Z}_*^{\mathsf{T}}, b_*)^{\mathsf{T}} \big| C_{*,j}, 1 \leq j \leq m_* \Big\} \boldsymbol{a} \\
&\quad \geq \kappa \boldsymbol{a}^{\mathsf{T}} E\Big[\big\{(\boldsymbol{X}_{*,1}^{\mathsf{T}}, \ldots, \boldsymbol{X}_{*,m_*}^{\mathsf{T}}, \boldsymbol{Z}_*^{\mathsf{T}}, b_*)^{\mathsf{T}}\big\}^{\otimes 2} \big| C_{*,j}, 1 \leq j \leq m_* \Big] \boldsymbol{a}
\end{aligned}$$

uniformly for all $\boldsymbol{a}$ with a suitable length.

Condition (C1), in the use of completely random cluster size, can be found in Zeng et al. (2005). (C2)–(C6) are widely used in semiparametric modeling of survival analysis (see, for example, Huang and Rossini, 1997; Zhang et al., 2010) and usually satisfied in practice. Conditions (C1)–(C4) ensure the proposed model is identifiable. In particular, (C2) implies that for all $(\boldsymbol{\beta}, \boldsymbol{\gamma}, \theta)$ and $v \in \mathbb{R}$,

$$\Pr(\boldsymbol{\beta}^{\mathsf{T}} \boldsymbol{X}_{*,j} + \boldsymbol{\gamma}^{\mathsf{T}} \boldsymbol{Z}_* + \theta b_* \neq v) > 0, \quad \forall j.$$

Condition (C3) suffices to prevent the joint distribution of the covariates and the inspection time from degeneration. For example, under (C3), I are able to show that

$$\begin{cases}
\Pr(\Delta_{*,j} = 1 | C_{*,j} \neq 0) > 0, \\
\Pr(\Delta_{*,j} = 0 | C_{*,j} \neq 0) > 0.
\end{cases}$$

Furthermore, it guarantees that the density function of $(C_{*,j}|b_*)$ is also bounded away from zero and infinity in its support. Condition (C4) regularizes the nonparametric function to be estimated. (C5) and (C6) are technical assumptions used in the proof of rate of convergence and asymptotic normality. Although some of these conditions can be relaxed to a weaker version, it will make the

proofs considerably more difficult and unnecessary to do so.

The following theorem establishes the consistency of the penalized complete ML estimator.

Theorem A.1. *Suppose the regularity conditions (C1)–(C6) hold,* $L = O(n^{1/(2q+1)})$*, and the tuning parameter* λ *satisfies* $\lambda \asymp n^{-2q/(2q+1)}$*. Then*

$$\operatorname{dist}(\widehat{\boldsymbol{\iota}}_{c,n}, \boldsymbol{\iota}_0) = O_p\big(n^{-q/(2q+1)}\big). \tag{A.11}$$

Semiparametric efficiency bound: For notational convenience, for a vector $\boldsymbol{\alpha}$ with suitable length, let $\dot{\ell}_{c,1}(\boldsymbol{\iota}; \boldsymbol{g})$ denote the vector of partial derivatives of $\ell_c(\boldsymbol{\iota}; \boldsymbol{g})$ with respect to $\boldsymbol{\alpha}$. For the nonparametric part, consider a parametric smooth submodel with parameter $(\boldsymbol{\alpha}^\mathsf{T}, H_{(s,w)})^\mathsf{T}$, such that $H_{(s,w)} = H + sw \in \mathcal{H}$ for s in a small interval containing 0, with $H_{(0,w)} = H$ and $\{\partial H_{(s,w)}/\partial s\}|_{s=0} = w$. Let $\mathcal{W}$ be the class of functions w defined by this equation. The score operator for H begins with defining the Gâteaux (directional) derivative at H along w: $\dot{\ell}_{c,2}(\boldsymbol{\iota}; \boldsymbol{g})[w] = \{\partial l(\boldsymbol{\alpha}, H_{(s,w)}; \boldsymbol{g})/\partial s\}|_{s=0}$. In addition, for $\boldsymbol{w} = (w_1, \ldots, w_p)^\mathsf{T}$ with $w_k \in \mathcal{W}$, $k = 1, \ldots, p$, let $\dot{\ell}_{c,2}(\boldsymbol{\iota}; \boldsymbol{g})[\boldsymbol{w}]$ be the p-dimensional vector with its kth element $\dot{\ell}_{c,2}(\boldsymbol{\iota}; \boldsymbol{g})[w_k]$. If $\boldsymbol{w}_c^* \in \mathcal{W}^p$ and satisfies

$$\boldsymbol{w}_c^* = \underset{\boldsymbol{w} \in \mathcal{W}^p}{\arg\min}\, E\big\|\dot{\ell}_{c,1}(\boldsymbol{\iota}; \boldsymbol{g}) - \dot{\ell}_{c,2}(\boldsymbol{\iota}; \boldsymbol{g})[\boldsymbol{w}]\big\|^2, \tag{A.12}$$

then $\boldsymbol{w}_c^*$ is called the *least favorable direction*, and by Theorem 1 in Bickel et al. (1993, pp. 70), the efficient score for $\boldsymbol{\alpha}$ is $\dot{\ell}_{c,1}(\boldsymbol{\iota}; \boldsymbol{g}) - \dot{\ell}_{c,2}(\boldsymbol{\iota}; \boldsymbol{g})[\boldsymbol{w}_c^*]$. According to the result in Bickel et al. (1993), the efficient information matrix of parameter $\boldsymbol{\alpha}$ for the complete likelihood is given by

$$I_c(\boldsymbol{\alpha}) = E\big\{\dot{\ell}_{c,1}(\boldsymbol{\iota}; \boldsymbol{g}) - \dot{\ell}_{c,2}(\boldsymbol{\iota}; \boldsymbol{g})[\boldsymbol{w}_c^*]\big\}^{\otimes 2}. \tag{A.13}$$

Analogously, the efficient information matrix of parameter $\boldsymbol{\alpha}$ for the observed likelihood is given by

$$I(\boldsymbol{\alpha}) = E\big\{\dot{\ell}_1(\boldsymbol{\iota}; \boldsymbol{U}_*) - \dot{\ell}_2(\boldsymbol{\iota}; \boldsymbol{O}_*)[\boldsymbol{w}^*]\big\}^{\otimes 2}, \tag{A.14}$$

where $\dot{\ell}_1$, $\dot{\ell}_2$, and $\boldsymbol{w}^*$ are the partial derivative of ℓ with respect to the parametric component $\boldsymbol{\alpha}$, Gâteaux (directional) derivative of ℓ with respect to the nonparametric component H, and the corresponding least favorable direction, respectively.

The next lemma shows the existence of the least favorable directions $\boldsymbol{w}_c^*$ and $\boldsymbol{w}^*$. Furthermore, the expressions of efficient information matrices $I_c(\boldsymbol{\alpha})$ in (A.13) and $I(\boldsymbol{\alpha})$ in (A.14) can be obtained.

Lemma A.1. *Under conditions (C1)–(C4), the least favorable directions $\boldsymbol{w}_c^*$ and $\boldsymbol{w}^*$ exist.*

To investigate the asymptotic normality and efficiency, the least favorable direction must be estimable in the sense that its roughness penalty is bounded away from infinity, which leads to our last regularity condition.

(C7) The least favorable direction $\boldsymbol{w}_c^*$ for the complete likelihood satisfies $J(\boldsymbol{w}_c^*) < \infty$.

(C7') The least favorable direction $\boldsymbol{w}^*$ for the observed likelihood satisfies $J(\boldsymbol{w}^*) < \infty$.

Theorem A.2. *Suppose that all the assumptions given in Theorem 2.2 hold and the regularity condition (C7) is satisfied. Then, $n^{1/2}\big(\widehat{\boldsymbol{\alpha}}_{c,n} - \boldsymbol{\alpha}_0\big)$ converges to $\mathcal{N}\big(\mathbf{0}, I_c^{-1}(\boldsymbol{\alpha}_0)\big)$ in distribution, where $I_c(\boldsymbol{\alpha}_0)$ is the efficient information of $\boldsymbol{\alpha}$ with expected value at $\boldsymbol{\iota}_0$ for the complete likelihood, and assumed non-singular.*

A.2.2 Proof of Lemma 2.2

Proof of Lemma 2.2. Suppose $(\widetilde{\boldsymbol{\beta}}, \widetilde{\boldsymbol{\gamma}}, \widetilde{\theta}, \widetilde{H})$ gives the same observed likelihood function $(\boldsymbol{\beta}_0, \boldsymbol{\gamma}_0, \theta_0, H_0)$. Due to Condition (C1), it implies that

$$\begin{aligned}
&\left(1 - \left[1 + r\widetilde{H}(C_{*,j}) \exp\{\widetilde{\boldsymbol{\beta}}^{\mathsf{T}} \boldsymbol{X}_{*,j} + \widetilde{\boldsymbol{\gamma}}^{\mathsf{T}} \boldsymbol{Z}_* + \widetilde{\theta} b_*\}\right]^{-1/r}\right)^{\Delta_{*,j}} \\
&\qquad \times \left(\left[1 + r\widetilde{H}(C_{*,j}) \exp\{\widetilde{\boldsymbol{\beta}}^{\mathsf{T}} \boldsymbol{X}_{*,j} + \widetilde{\boldsymbol{\gamma}}^{\mathsf{T}} \boldsymbol{Z}_* + \widetilde{\theta} b_*\}\right]^{-1/r}\right)^{1-\Delta_{*,j}} \\
&= \left(1 - \left[1 + rH_0(C_{*,j}) \exp\{{\boldsymbol{\beta}_0}^{\mathsf{T}} \boldsymbol{X}_{*,j} + {\boldsymbol{\gamma}_0}^{\mathsf{T}} \boldsymbol{Z}_* + \theta_0 b_*\}\right]^{-1/r}\right)^{\Delta_{*,j}} \\
&\qquad \times \left(\left[1 + rH_0(C_{*,j}) \exp\{{\boldsymbol{\beta}_0}^{\mathsf{T}} \boldsymbol{X}_{*,j} + {\boldsymbol{\gamma}_0}^{\mathsf{T}} \boldsymbol{Z}_* + \theta_0 b_*\}\right]^{-1/r}\right)^{1-\Delta_{*,j}}.
\end{aligned} \tag{A.15}$$

After using (C3) and choosing $\Delta_{*,j} = 0$ in (A.15), I then obtain

$$\begin{aligned}
&\left[1 + r\widetilde{H}(C_{*,j}) \exp\{\widetilde{\boldsymbol{\beta}}^{\mathsf{T}} \boldsymbol{X}_{*,j} + \widetilde{\boldsymbol{\gamma}}^{\mathsf{T}} \boldsymbol{Z}_* + \widetilde{\theta} b_*\}\right]^{1/r} \\
&\qquad = \left[1 + rH_0(C_{*,j}) \exp\{{\boldsymbol{\beta}_0}^{\mathsf{T}} \boldsymbol{X}_{*,j} + {\boldsymbol{\gamma}_0}^{\mathsf{T}} \boldsymbol{Z}_* + \theta_0 b_*\}\right]^{1/r}.
\end{aligned}$$

From the monotonicity of $(1 + rx)^{1/r}$ $(r > 0)$ w.r.t. x, the aforementioned equation implies that

$$\widetilde{H}(C_{*,j}) \exp\{\widetilde{\boldsymbol{\beta}}^{\mathsf{T}} \boldsymbol{X}_{*,j} + \widetilde{\boldsymbol{\gamma}}^{\mathsf{T}} \boldsymbol{Z}_* + \widetilde{\theta} b_*\} = H_0(C_{*,j}) \exp\{{\boldsymbol{\beta}_0}^{\mathsf{T}} \boldsymbol{X}_{*,j} + {\boldsymbol{\gamma}_0}^{\mathsf{T}} \boldsymbol{Z}_* + \theta_0 b_*\}. \tag{A.16}$$

I use Conditions (C3) and (C4) to get that with positive probability, I can fix $C_{*,j} \neq 0$ such that both $\widetilde{H}(C_{*,j})$ and $H_0(C_{*,j})$ are not equal to zero. (A.16) together with (C3) then imply that

$$\widetilde{\boldsymbol{\beta}}^{\mathsf{T}} \boldsymbol{X}_{*,j} + \widetilde{\boldsymbol{\gamma}}^{\mathsf{T}} \boldsymbol{Z}_* + \widetilde{\theta} b_* = {\boldsymbol{\beta}_0}^{\mathsf{T}} \boldsymbol{X}_{*,j} + {\boldsymbol{\gamma}_0}^{\mathsf{T}} \boldsymbol{Z}_* + \theta_0 b_* + v$$

for some v. Using (C2), it shows that $(\widetilde{\boldsymbol{\beta}}, \widetilde{\boldsymbol{\gamma}}, \widetilde{\theta}) = (\boldsymbol{\beta}_0, \boldsymbol{\gamma}_0, \theta_0)$. The conclusion of $\widetilde{H} = H_0$ follows after plugging this result into (A.16).

□

A.2.3 Proof of Theorem A.1

To prove Theorem A.1, I first need the following technical lemmas.

Lemma A.2. *If Conditions (C1)–(C7) hold, then, for a sufficiently small $\delta > 0$, there exists a constant $v > 0$ depending on $\Pr_0$ such that $\|H\|_\infty \leq v\{J(H) + 1\}$ whenever $H \in \mathcal{H}$ and $\|H - H_0\|_2 < \delta$.*

Proof of Lemma A.2. Because $\|H - H_0\|_2 < \delta$ for a sufficiently small $\delta > 0$, it implies that there exist disjoint intervals $[a_i, b_i] \subset [0, T_0]$ such that $H(a_i) < H(b_i)$ and $\int_{[a_i,b_i]}\{H(t)-H_0(t)\}^2\mathrm{d}t < \delta^2$ for each $i = 1, \ldots, k$. Therefore, there exists $t_i \in [a_i, b_i]$ satisfying $\{H(t_i) - H_0(t_i)\}^2 \leq v\delta^2$. In view of the fact that H_0 is uniformly bounded on $[0, T_0]$, it follows that $H(t_i) \leq K_\delta$ for some constant K_δ depending on δ. For any $H \in \mathcal{H}$ with $J(H) < \infty$, Condition (C4) and $\|H - H_0\|_2 < \delta$ with sufficiently small δ imply that $J(H)$ is also bounded away from 0. Thus there exists a polynomial spline $\widetilde{H} \in \mathcal{S}(\boldsymbol{\tau}, L, d)$ such that $\|H - \widetilde{H}\|_\infty \leq vq^{-d} \leq J(H)$ (see, for example, the proof of Lemma 7.2 of Murphy and van der Vaart, 1999) with d large enough. It follows that $\widetilde{H}(t_i) \leq J(H) + H(t_i) \leq J(H) + K_\delta$. Using the approximation property of polynomial spline (de Boor, 1978), $\|\widetilde{H}\|_\infty \leq v\{J(H)+K_\delta\}$, and $\|H\|_\infty$ is bounded by $v\{J(H)+1\}$ accordingly. □

Lemma A.3. *If Conditions (C1)–(C7) hold, then there exists a constant $v > 0$ such that*

$$\Pr\{\ell_c(\boldsymbol{\iota}; \boldsymbol{g}) - \ell_c(\boldsymbol{\iota}_0; \boldsymbol{g})\}^2 \geq v\|\boldsymbol{\iota} - \boldsymbol{\iota}_0\|_\Xi^2$$

for $\boldsymbol{\iota}$ in a neighborhood of $\boldsymbol{\iota}_0$.

Proof of Lemma A.3. From the complete likelihood function (A.9), it is shown that

$$\begin{aligned}
&\Pr\{\ell_c(\boldsymbol{\iota};\boldsymbol{g}) - \ell_c(\boldsymbol{\iota}_0;\boldsymbol{g})\}^2 \\
&= \int \Bigg(\textstyle\sum_{j=1}^{m_*}(1-\Delta_{*,j})\big[\log\{S_{\boldsymbol{\iota}}(C_{*,j}|\boldsymbol{X}_{*,j},\boldsymbol{Z}_*,b_*)\} - \log\{S_{,\boldsymbol{\iota}_0}(C_{*,j}|\boldsymbol{X}_{*,j},\boldsymbol{Z}_*,b_*)\}\big] \\
&\quad + \textstyle\sum_{j=1}^{m}\Delta_{*,j}\big[\log\{1-S_{\boldsymbol{\iota}}(C_{*,j}|\boldsymbol{X}_{*,j},\boldsymbol{Z}_*,b_*)\} - \log\{1-S_{\boldsymbol{\iota}_0}(C_{*,j}|\boldsymbol{X}_{*,j},\boldsymbol{Z}_*,b_*)\}\big] \\
&\qquad + \big\{\log\,\phi(b_*) - \log\,\phi(b_*)\big\}\Bigg)^2 \mathrm{d}\Pr, \qquad \text{(A.17)}
\end{aligned}$$

where $S_{\boldsymbol{\iota}}(C_{*,j}|\boldsymbol{X}_{*,j},\boldsymbol{Z}_*,b_*)$ and $\phi(b)$ respectively denote the survival function of the time-to-event in the susceptible population given in (2.1) with parameter $\boldsymbol{\iota}$and probability density function of b which is $\mathcal{N}(0,1)$. Using Conditions (C3) and (C5), to show (A.17) greater than or equal to $\|\boldsymbol{\iota}-\boldsymbol{\iota}_0\|_{\Xi}^2$, up to a constant, it suffices to show that

$$\begin{aligned}
&\int \Bigg(\sum_{j=1}^{m}\big[\log\{1-S_{,\boldsymbol{\iota}}(C_{*,j}|\boldsymbol{X}_{*,j},\boldsymbol{Z}_*,b_*)\} - \log\{1-S_{,\boldsymbol{\iota}_0}(C_{*,j}|\boldsymbol{X}_{*,j},\boldsymbol{Z}_*,b_*)\}\big]\Bigg)^2 \mathrm{d}\Pr \\
&\quad \geq v\big\{\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|^2 + \|\boldsymbol{\gamma}-\boldsymbol{\gamma}_0\|^2 + (\theta-\theta_0)^2 + \|H-H_0\|_2^2\big\},
\end{aligned} \qquad \text{(A.18)}$$

for some constant $v>0$.

Next, I first show the following simplified version of (A.18)

$$\begin{aligned}
&\int \big[\log\{1-S_{,\boldsymbol{\iota}}(C_{*,j}|\boldsymbol{X}_{*,j},\boldsymbol{Z}_*,b_*)\} - \log\{1-S_{,\boldsymbol{\iota}_0}(C_{*,j}|\boldsymbol{X}_{*,j},\boldsymbol{Z}_*,b_*)\}\big]^2 \mathrm{d}\Pr \\
&\quad \geq v\big\{\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|^2 + \|\boldsymbol{\gamma}-\boldsymbol{\gamma}_0\|^2 + (\theta-\theta_0)^2 + \|H-H_0\|_2^2\big\}.
\end{aligned} \qquad \text{(A.19)}$$

Let $g_1(s)$ denote

$$\log\Big[1-\big\{1+rH_s(C_{*,j})\exp(\boldsymbol{\beta}_s{}^{\mathsf{T}}\boldsymbol{X}_{*,j}+\boldsymbol{\gamma}_s{}^{\mathsf{T}}\boldsymbol{Z}_*+\theta_s b_*)\big\}^{-1/r}\Big],$$

where $H_s(C_{*,j}) = sH(C_{*,j}) + (1-s)H_0(C_{*,j})$, $\boldsymbol{\beta}_s = s\boldsymbol{\beta}+(1-s)\boldsymbol{\beta}_0$, $\boldsymbol{\gamma}_s = s\boldsymbol{\gamma}+(1-s)\boldsymbol{\gamma}_0$, and $\theta_s = s\theta+(1-s)\theta_0$, respectively. The term inside the integral of the left hand side of (A.19) is then equal to $\{g_1(1)-g_1(0)\}^2$. Application of the mean value theorem leads to $g_1(1)-g_1(0) = g_1'(\epsilon)$

for some $0 \leq \epsilon \leq 1$. It is shown that

$$
\begin{aligned}
g_1'(\epsilon) = & \left(\frac{\left\{1 + rH_\epsilon(C_{*,j}) \exp(\boldsymbol{\beta}_\epsilon{}^\mathsf{T} \boldsymbol{X}_{*,j} + \boldsymbol{\gamma}_\epsilon{}^\mathsf{T} \boldsymbol{Z}_* + \theta_\epsilon b_*)\right\}^{-1/r-1} \exp(\boldsymbol{\beta}_\epsilon{}^\mathsf{T} \boldsymbol{X}_{*,j} + \boldsymbol{\gamma}_\epsilon{}^\mathsf{T} \boldsymbol{Z}_* + \theta_\epsilon b_*)}{1 - \left[1 + r\big\{H_0 + \epsilon(H - H_0)\big\}(C_{*,j}) \exp(\boldsymbol{\beta}_\epsilon{}^\mathsf{T} \boldsymbol{X}_{*,j} + \boldsymbol{\gamma}_\epsilon{}^\mathsf{T} \boldsymbol{Z}_* + \theta_\epsilon b_*)\right]^{-1/r}} \right) \\
& \times \Big[(H - H_0)(C_{*,j}) + \big\{H_0 + \epsilon(H - H_0)\big\}(C_{*,j}) \\
& \quad \times \big\{(\boldsymbol{\beta} - \boldsymbol{\beta}_0)^\mathsf{T} \boldsymbol{X}_{*,j} + (\boldsymbol{\gamma} - \boldsymbol{\gamma}_0)^\mathsf{T} \boldsymbol{Z}_* + (\theta - \theta_0) b_*\big\} \Big] \\
= & \left(\frac{\left\{1 + rH_\epsilon(C_{*,j}) \exp(\boldsymbol{\beta}_\epsilon{}^\mathsf{T} \boldsymbol{X}_{*,j} + \boldsymbol{\gamma}_\epsilon{}^\mathsf{T} \boldsymbol{Z}_* + \theta_\epsilon b_*)\right\}^{-1/r-1} \exp(\boldsymbol{\beta}_\epsilon{}^\mathsf{T} \boldsymbol{X}_{*,j} + \boldsymbol{\gamma}_\epsilon{}^\mathsf{T} \boldsymbol{Z}_* + \theta_\epsilon b_*)}{1 - \left[1 + r\big\{H_0 + \epsilon(H - H_0)\big\}(C_{*,j}) \exp(\boldsymbol{\beta}_\epsilon{}^\mathsf{T} \boldsymbol{X}_{*,j} + \boldsymbol{\gamma}_\epsilon{}^\mathsf{T} \boldsymbol{Z}_* + \theta_\epsilon b_*)\right]^{-1/r}} \right) \\
& \times \Big[(H - H_0)(C_{*,j}) \big\{1 + \epsilon(\boldsymbol{\beta} - \boldsymbol{\beta}_0)^\mathsf{T} \boldsymbol{X}_{*,j} + \epsilon(\boldsymbol{\gamma} - \boldsymbol{\gamma}_0)^\mathsf{T} \boldsymbol{Z}_* + \epsilon(\theta - \theta_0) b_*\big\} \\
& \quad + H_0(C_{*,j}) \big\{(\boldsymbol{\beta} - \boldsymbol{\beta}_0)^\mathsf{T} \boldsymbol{X}_{*,j} + (\boldsymbol{\gamma} - \boldsymbol{\gamma}_0)^\mathsf{T} \boldsymbol{Z}_* + (\theta - \theta_0) b\big\} \Big] \\
:= & \, g_{1,\epsilon}(C_{*,j}, \boldsymbol{X}_{*,j}, \boldsymbol{Z}_*, b_*) \cdot \Big[(H - H_0)(C_{*,j}) \big\{1 + \epsilon(\boldsymbol{\beta} - \boldsymbol{\beta}_0)^\mathsf{T} \boldsymbol{X}_{*,j} + \epsilon(\boldsymbol{\gamma} - \boldsymbol{\gamma}_0)^\mathsf{T} \boldsymbol{Z}_* + \epsilon(\theta - \theta_0) b_*\big\} \\
& \quad + H_0(C_{*,j}) \big\{(\boldsymbol{\beta} - \boldsymbol{\beta}_0)^\mathsf{T} \boldsymbol{X}_{*,j} + (\boldsymbol{\gamma} - \boldsymbol{\gamma}_0)^\mathsf{T} \boldsymbol{Z}_* + (\theta - \theta_0) b_*\big\} \Big],
\end{aligned}
$$

where $g_{1,\epsilon}$ is a function of random variables $(C_{*,j}, \boldsymbol{X}_{*,j}, \boldsymbol{Z}_*, b_*)$. From the application of the mean value theorem and Conditions (C2)–(C5), I have

$$
\begin{aligned}
& \int \big[\log\{1 - S_{,\boldsymbol{\iota}}(C_{*,j} | \boldsymbol{X}_{*,j}, \boldsymbol{Z}_*, b_*)\} - \log\{1 - S_{,\boldsymbol{\iota}_0}(C_{*,j} | \boldsymbol{X}_{*,j}, \boldsymbol{Z}_*, b_*)\} \big]^2 \mathrm{d}\Pr \\
& \quad \geq \int \Big[(H - H_0)(C_{*,j}) \big\{1 + \epsilon(\boldsymbol{\beta} - \boldsymbol{\beta}_0)^\mathsf{T} \boldsymbol{X}_{*,j} + \epsilon(\boldsymbol{\gamma} - \boldsymbol{\gamma}_0)^\mathsf{T} \boldsymbol{Z}_* + \epsilon(\theta - \theta_0) b_*\big\} \\
& \qquad + H_0(C_{*,j}) \big\{(\boldsymbol{\beta} - \boldsymbol{\beta}_0)^\mathsf{T} \boldsymbol{X}_{*,j} + (\boldsymbol{\gamma} - \boldsymbol{\gamma}_0)^\mathsf{T} \boldsymbol{Z}_* + (\theta - \theta_0) b_*\big\} \Big]^2 \mathrm{d}\Pr
\end{aligned} \tag{A.20}
$$

up to a constant. To simplify the notations, I let $g_2(C_{*,j}, \boldsymbol{X}_{*,j}, \boldsymbol{Z}_*) = \big\{(\boldsymbol{\beta} - \boldsymbol{\beta}_0)^\mathsf{T} \boldsymbol{X}_{*,j} + (\boldsymbol{\gamma} - \boldsymbol{\gamma}_0)^\mathsf{T} \boldsymbol{Z}_* + (\theta - \theta_0) b_*\big\} H_0(C_{*,j})$, $g_3(C_{*,j}) = (H - H_0)(C_{*,j})$, and $\vartheta(C_{*,j}) = 1 + \epsilon(H - H_0)(C_{*,j}) / H_0(C_{*,j})$, respectively. To show (A.19), it thus suffices to verify

$$
\Pr(g_2 \vartheta + g_3)^2 \geq \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|^2 + \|\boldsymbol{\gamma} - \boldsymbol{\gamma}_0\|^2 + (\theta - \theta_0)_2^2 + \|H - H_0\|_2^2 \tag{A.21}
$$

up to a constant. To apply Lemma 25.86 of van der Vaart (1998), I need to bound $\{\Pr(g_2 g_3)\}^2$ by

a constant less than one times $\Pr(g_2^2)\Pr(g_3^2)$. By then computing conditionally on $C_{*,j}$, I have

$$\begin{aligned}
\{\Pr(g_2 g_3)\}^2 &= \big[\Pr\{\Pr(g_2 g_3|C_{*,j})\}\big]^2 \\
&\le \Pr(g_3^2)\Pr\big[\{\Pr(g_2^2|C_{*,j})\}^2\big] \\
&= \Pr(g_3^2)\Pr\Big[H_0^2(C_{*,j})\big\{\big((\boldsymbol{\beta}-\boldsymbol{\beta}_0)^\mathsf{T},(\boldsymbol{\gamma}-\boldsymbol{\gamma}_0)^\mathsf{T},\theta-\theta_0\big) \\
&\qquad \times \big[\{\Pr(\boldsymbol{X}_{*,j}^\mathsf{T},\boldsymbol{Z}_*^\mathsf{T},b_*)^\mathsf{T}|C_{*,j}\}^{\otimes 2}\big]\big((\boldsymbol{\beta}-\boldsymbol{\beta}_0)^\mathsf{T},(\boldsymbol{\gamma}-\boldsymbol{\gamma}_0)^\mathsf{T},\theta-\theta_0\big)^\mathsf{T}\big\}\Big] \\
&\le (1-\kappa)\Pr(g_3^2)\Pr\Big\{H_0^2(C_{*,j})\big((\boldsymbol{\beta}-\boldsymbol{\beta}_0)^\mathsf{T},(\boldsymbol{\gamma}-\boldsymbol{\gamma}_0)^\mathsf{T},\theta-\theta_0\big) \\
&\qquad \times \Pr\big[\{(\boldsymbol{X}_{*,j}^\mathsf{T},\boldsymbol{Z}_*^\mathsf{T},b_*)^\mathsf{T}\}^{\otimes 2}|C_{*,j}\big]\big((\boldsymbol{\beta}-\boldsymbol{\beta}_0)^\mathsf{T},(\boldsymbol{\gamma}-\boldsymbol{\gamma}_0)^\mathsf{T},\theta-\theta_0\big)^\mathsf{T}\Big\} \\
&= (1-\kappa)\Pr(g_3^2)\Pr(g_2^2),
\end{aligned}$$

where the first and second inequalities follow from the Cauchy-Schwarz inequality and Condition (C6), respectively. Thus by Lemma 25.86 of van der Vaart (1998) and Conditions (C2)–(C4),

$$\Pr(g_2\vartheta+g_3)^2\Pr(g_2^2)+\Pr(g_3^2)\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|^2+\|\boldsymbol{\gamma}-\boldsymbol{\gamma}_0\|^2+(\theta-\theta_0)^2+\|H-H_0\|_2^2,$$

where $\lesssim$ denotes $\ge$ up to a constant.

The last step is to show (A.18) from its simplified version (A.19). Indeed, it can be completed by using Condition (C1) and the similar arguments as shown in the proof of (A.21). □

Proof of Theorem A.1. To prove the stated rate of convergence, I first show the consistency of the penalized estimator. Define

$$m_{\boldsymbol{\iota},\lambda} = \log\left(\frac{p_{\boldsymbol{\iota}}+p_0}{2p_0}\right) - \frac{\lambda}{2}\{J^2(H)-J^2(H_0)\}.$$

Under the order assumption of λ, I may assume that $\lambda \in \boldsymbol{\lambda}_n = [\widetilde{\lambda}_n,\infty)$ for

$$\widetilde{\lambda}_n = n^{-2q/(1+2q)}. \tag{A.22}$$

By the concavity of the logarithmic function, the relationship between $p_{\boldsymbol{\iota}}$ and $\ell_c(\boldsymbol{\iota}; \boldsymbol{g})$, and the definition of $\widehat{\boldsymbol{\iota}}_{c,n}$,

$$\mathbb{P}_n m_{\widehat{\boldsymbol{\iota}}_{c,n},\lambda} \geq \frac{1}{2}\mathbb{P}_n \log\left(\frac{p_{\widehat{\boldsymbol{\iota}}_{c,n}}}{p_0}\right) - \frac{\lambda}{2}\{J^2(\widehat{H}_{c,n}) - J^2(H_0)\} \geq 0 = \mathbb{P}_n m_{\boldsymbol{\iota}_0,\lambda}.$$

It can also be shown that

$$\Pr_0(m_{\boldsymbol{\iota},\lambda} - m_{\boldsymbol{\iota}_0,\lambda}) = \int \log\frac{p_{\boldsymbol{\iota}} + p_0}{2p_0} p_0 \mathrm{d}\mu - \frac{\lambda}{2}\{J^2(H) - J^2(H_0)\}.$$

Since $\log(x) \leq 2(x^{1/2} - 1)$ for $x > 0$, it follows that

$$\frac{1}{2}\int \log\left(\frac{p_{\boldsymbol{\iota}} + p_0}{2p_0}\right) p_0 \mathrm{d}\mu \leq \int \left(\frac{p_{\boldsymbol{\iota}} + p_0}{2p_0}\right)^{1/2} p_0 \mathrm{d}\mu - 1 = -\frac{1}{4}h^2(p_{\boldsymbol{\iota}} + p_0, 2p_0),$$

where $h(p_{\boldsymbol{\iota}}, p_0)$ is the Hellinger distance defined as $h^2(p_{\boldsymbol{\iota}}, p_0) = \int (p_{\boldsymbol{\iota}}^{1/2} - p_0^{1/2})^2 \mathrm{d}\mu$. Hence,

$$\Pr_0(m_{\boldsymbol{\iota},\lambda} - m_{\boldsymbol{\iota}_0,\lambda}) \leq -\frac{1}{2}h^2(p_{\boldsymbol{\iota}} + p_0, 2p_0) - \frac{\lambda}{2}\{J^2(H) - J^2(H_0)\}.$$

Using page 328 of van der Vaart and Wellner (1996), I have that

$$h(p_{\boldsymbol{\iota}} + p_0, 2p_0) \leq h(p_{\boldsymbol{\iota}}, p_0) \leq 2h(p_{\boldsymbol{\iota}} + p_0, 2p_0).$$

Thus the squared Hellinger distance $h^2(p_{\boldsymbol{\iota}} + p_0, 2p_0)$ is equivalent $h^2(p_{\boldsymbol{\iota}}, p_0)$, up to a constant. Theorem 3.4.4 of van der Vaart and Wellner (1996) and Condition (C3) imply that

$$\Pr_0\{\log(p_{\boldsymbol{\iota}}) - \log(p_0)\}^2 \leq v h^2(p_{\boldsymbol{\iota}}, p_0),$$

for some constant v. Hence, in view of Lemma A.3 and Condition (C3), it follows that

$$\Pr_0(m_{\boldsymbol{\iota},\lambda} - m_{\boldsymbol{\iota}_0,\lambda}) - \|\boldsymbol{\iota} - \boldsymbol{\iota}_0\|_\Xi^2 - \lambda J^2(H) + \lambda,$$

where $\asymp$ denotes $\leq$ up to a constant. This suggests the choice of

$$d_\lambda(\boldsymbol{\iota}-\boldsymbol{\iota}_0)=\left\{\|\boldsymbol{\iota}-\boldsymbol{\iota}_0\|_\Xi^2+\lambda J^2(H)|\right\}^{1/2} \tag{A.23}$$

in Theorem 25.81 of van der Vaart (1998). Next, using the same arguments as those in Lemma 7.2 of Murphy and van der Vaart (1999), it can be shown that

$$\sup_Q \log N_{[]}\big(\epsilon,\{m_{\boldsymbol{\iota},0},\boldsymbol{\alpha}\in\Theta,J(H)\leq M\},L_2(Q)\big)\leq v\left(\frac{1+M}{\epsilon}\right)^{1/q}. \tag{A.24}$$

Under the choice of (A.23), $d_\lambda(\boldsymbol{\iota}-\boldsymbol{\iota}_0)<\delta$ implies that $J(H)\leq\delta/\widetilde{\lambda}_n^{1/2}$. Using this fact, Lemma 2.1 of van de Geer (2000), Theorem 2.14.1 of van der Vaart and Wellner (1996), and (A.24) imply that

$$\Pr_0\sup_{d_\lambda(\boldsymbol{\iota}-\boldsymbol{\iota}_0)<\delta,\lambda\in\boldsymbol{\lambda}_n}\left|\mathbb{G}_n(m_{\boldsymbol{\iota},\lambda}-m_{\boldsymbol{\iota}_0,\lambda})\right|\leq v\left(1+\frac{\delta}{\widetilde{\lambda}_n^{1/2}}\right)^{1/(2q)}.$$

Theorem 25.81 of van der Vaart (1998) yields $d_\lambda(\widehat{\boldsymbol{\iota}}_{c,n}-\boldsymbol{\iota}_0)=O_p(\delta_n+n^{-q/(1+2q)})$ for any $\delta_n\downarrow 0$ and $\delta_n\geq(n^{2q}\widetilde{\lambda}_n)^{-1/(8q-2)}$, which concludes the consistency of $\widehat{\boldsymbol{\iota}}_{c,n}$ by (A.22).

To show the rate of convergence, using Lemma A.2, it is reasonable to restrict H to the set $\mathcal{H}_n=\{H:\|H\|_\infty\leq v(J(H)+1)\}$ for a large constant v. If $d_\lambda(\boldsymbol{\iota}-\boldsymbol{\iota}_0)<\delta$ and $\lambda\in\boldsymbol{\lambda}_n$, then $\|\boldsymbol{\iota}-\boldsymbol{\iota}_0\|_\Xi<\delta$, $J(H)<\delta/\widetilde{\lambda}_n^{1/2}$, and hence, $\|H\|_\infty\leq v(\delta/\widetilde{\lambda}_n^{1/2}+1)$. Using Taylor expansion along with condition (C1), (C2) and (C5), it can be shown that the parametric part of $m_{\boldsymbol{\iota},0}$ is essentially Lipschitz with respect to $\boldsymbol{\alpha}$. The above two facts and Example 19.10 of van der Vaart (1998) imply that

$$\log N_{[]}\big(\epsilon,\{m_{\boldsymbol{\iota},0}:\lambda\in\boldsymbol{\lambda}_n,H\in\mathcal{H}_n,d_\lambda(\boldsymbol{\iota}-\boldsymbol{\iota}_0)<\delta\},L_2(\Pr_0)\big)\leq v\left(\frac{1+\delta/\widetilde{\lambda}_n^{1/2}}{\epsilon}\right)^{1/q}.$$

Thus, Lemma 19.36 of van der Vaart (1998) shows that

$$\Pr_0\sup_{d_\lambda(\boldsymbol{\iota}-\boldsymbol{\iota}_0)<\delta,\lambda\in\boldsymbol{\lambda}_n}\left|\mathbb{G}_n(m_{\boldsymbol{\iota},\lambda}-m_{\boldsymbol{\iota}_0,\lambda})\right|\leq vJ_n(\delta)\left\{1+\frac{J_n(\delta)}{\delta^2n^{1/2}}\right\},$$

where

$$J_n(\delta) = \int_0^{\delta} \left(\frac{1 + \delta/\widetilde{\lambda}_n^{1/2}}{\epsilon} \right)^{1/(2q)} \mathrm{d}\epsilon = v \left(1 + \frac{\delta}{\widetilde{\lambda}_n^{1/2}} \right)^{1/(2q)} \delta^{1-1/(2q)} = v \big\{ \delta^{1-1/(2q)} + \delta n^{1/2(2q+1)} \big\},$$

for some constant v. Therefore, Theorem 25.81 of van der Vaart (1998) implies

$$\|\widehat{\boldsymbol{\iota}}_{c,n} - \boldsymbol{\iota}_0\|_{\Xi} = O_p(\delta_n + \widetilde{\lambda}_n) = O_p(\delta_n + n^{-q/(1+2q)}), \tag{A.25}$$

with δ_n satisfying

$$J_n(\delta_n) \left\{ 1 + \frac{J_n(\delta_n)}{\delta^2 n^{1/2}} \right\} \leq n^{1/2} \delta_n^2.$$

Brief calculation shows that the optimal rate of δ_n in the aforementioned equation is $n^{-q/(1+2q)}$. This result together with (A.25) completes the proof of Theorem A.1. □

A.2.4 Proof of Lemma A.1

Using Condition (C1), I direct calculate that

$$\begin{aligned}
&\dot{\ell}_{c,1}(\boldsymbol{\iota}; \boldsymbol{g}) \\
&\quad = \sum_{j=1}^{m_*} (1 - \Delta_{*,j}) \left\{ \frac{-H(C_{*,j}) \exp(\boldsymbol{\beta}^{\mathsf{T}} \boldsymbol{X}_{*,j} + \boldsymbol{\gamma}^{\mathsf{T}} \boldsymbol{Z}_* + \theta b_*) (\boldsymbol{X}_{*,j}{}^{\mathsf{T}}, \boldsymbol{Z}_*{}^{\mathsf{T}}, b_*)^{\mathsf{T}}}{1 + rH(C_{*,j}) \exp(\boldsymbol{\beta}^{\mathsf{T}} \boldsymbol{X}_{*,j} + \boldsymbol{\gamma}^{\mathsf{T}} \boldsymbol{Z}_* + \theta b_*)} \right\} \\
&\qquad + \sum_{j=1}^{m_*} \Delta_{*,j} \left(\frac{\big\{ 1 + rH(C_{*,j}) \exp(\boldsymbol{\beta}^{\mathsf{T}} \boldsymbol{X}_{*,j} + \boldsymbol{\gamma}^{\mathsf{T}} \boldsymbol{Z}_* + \theta b_*) \big\}^{-1/r-1}}{1 - \big\{ 1 + rH(C_{*,j}) \exp(\boldsymbol{\beta}^{\mathsf{T}} \boldsymbol{X}_{*,j} + \boldsymbol{\gamma}^{\mathsf{T}} \boldsymbol{Z}_* + \theta b_*) \big\}^{-1/r}} \right) \\
&\qquad\quad \times H(C_{*,j}) \exp(\boldsymbol{\beta}^{\mathsf{T}} \boldsymbol{X}_{*,j} + \boldsymbol{\gamma}^{\mathsf{T}} \boldsymbol{Z}_* + \theta b_*) (\boldsymbol{X}_{*,j}{}^{\mathsf{T}}, \boldsymbol{Z}_*{}^{\mathsf{T}}, b_*)^{\mathsf{T}} \\
&\quad = \sum_{j=1}^{m_*} H(C_{*,j}) \exp(\boldsymbol{\beta}^{\mathsf{T}} \boldsymbol{X}_{*,j} + \boldsymbol{\gamma}^{\mathsf{T}} \boldsymbol{Z}_* + \theta b_*) (\boldsymbol{X}_{*,j}{}^{\mathsf{T}}, \boldsymbol{Z}_*{}^{\mathsf{T}}, b_*)^{\mathsf{T}} \\
&\qquad \times \left[\frac{\Delta_{*,j} \big\{ 1 + rH(C_{*,j}) \exp(\boldsymbol{\beta}^{\mathsf{T}} \boldsymbol{X}_{*,j} + \boldsymbol{\gamma}^{\mathsf{T}} \boldsymbol{Z}_* + \theta b_*) \big\}^{1/r}}{1 - \big\{ 1 + rH(C_{*,j}) \exp(\boldsymbol{\beta}^{\mathsf{T}} \boldsymbol{X}_{*,j} + \boldsymbol{\gamma}^{\mathsf{T}} \boldsymbol{Z}_* + \theta b_*) \big\}^{-1/r}} \right. \\
&\qquad\quad \left. - \frac{1}{1 + rH(C_{*,j}) \exp(\boldsymbol{\beta}^{\mathsf{T}} \boldsymbol{X}_{*,j} + \boldsymbol{\gamma}^{\mathsf{T}} \boldsymbol{Z}_* + \theta b_*)} \right].
\end{aligned} \tag{A.26}$$

After denoting

$$
\begin{aligned}
&Q_{c,j}(C_{*,j}, X_{*,j}, Z_*, b_*; \boldsymbol{\iota}) \\
&\quad = \exp(\boldsymbol{\beta}^\mathsf{T}\boldsymbol{X}_{*,j} + \boldsymbol{\gamma}^\mathsf{T}\boldsymbol{Z}_* + \theta b_*)\left[\frac{\Delta_{*,j}\{1 + rH(C_{*,j})\exp(\boldsymbol{\beta}^\mathsf{T}\boldsymbol{X}_{*,j} + \boldsymbol{\gamma}^\mathsf{T}\boldsymbol{Z}_* + \theta b_*)\}^{1/r}}{1 - \{1 + rH(C_{*,j})\exp(\boldsymbol{\beta}^\mathsf{T}\boldsymbol{X}_{*,j} + \boldsymbol{\gamma}^\mathsf{T}\boldsymbol{Z}_* + \theta b_*)\}^{-1/r}}\right. \\
&\qquad\qquad \left. - \frac{1}{1 + rH(C_{*,j})\exp(\boldsymbol{\beta}^\mathsf{T}\boldsymbol{X}_{*,j} + \boldsymbol{\gamma}^\mathsf{T}\boldsymbol{Z}_* + \theta b_*)}\right],
\end{aligned}
$$

(A.26) can be written as

$$
\dot{\ell}_{c,1}(\boldsymbol{\iota}; \boldsymbol{g}) = \sum_{j=1}^{m_*} H(C_{*,j}) Q_{c,j}(C_{*,j}, X_{*,j}, Z_*, b_*; \boldsymbol{\iota})(\boldsymbol{X}_{*,j}{}^\mathsf{T}, \boldsymbol{Z}_*{}^\mathsf{T}, b_*)^\mathsf{T}. \tag{A.27}
$$

Similarly, differentiating the complete log-likelihood function $\ell_c(\boldsymbol{\iota}; \boldsymbol{g})$ at H along w yields

$$
\dot{\ell}_{c,2}(\boldsymbol{\iota}; \boldsymbol{g})[w] = \sum_{j=1}^{m_*} w(C_{*,j}) Q_{c,j}(C_{*,j}, X_{*,j}, Z_*, b_*; \boldsymbol{\iota}), \tag{A.28}
$$

where $w \in \mathcal{W}$ be the class of functions such that $H + sw \in \mathcal{H}$ for s in a small interval containing 0. Moreover, for $\boldsymbol{w} = (w_1, \ldots, w_p)^\mathsf{T}$ with $w_k \in \mathcal{W}$, $k = 1, \cdots, p$, let $\dot{\ell}_{c,2}(\boldsymbol{\iota}; \boldsymbol{g})[\boldsymbol{w}] = (\dot{\ell}_{c,2}(\boldsymbol{\iota}; \boldsymbol{g})[w_1], \ldots, \dot{\ell}_{c,2}(\boldsymbol{\iota}; \boldsymbol{g})[w_p])^\mathsf{T}$. To see that $\boldsymbol{w}_c^*$ exists in (A.12), I only need to show the normal equation

$$
E\dot{\ell}_{c,2}^*(\boldsymbol{\iota}; \boldsymbol{g})\dot{\ell}_{c,1}(\boldsymbol{\iota}; \boldsymbol{g}) - E\dot{\ell}_{c,2}^*(\boldsymbol{\iota}; \boldsymbol{g})\dot{\ell}_{c,2}(\boldsymbol{\iota}; \boldsymbol{g})[\boldsymbol{w}_c^*] = 0
$$

has a solution, where $\dot{\ell}_{c,2}^*(\boldsymbol{\iota}; \boldsymbol{g})$ is the adjoint operator of $\dot{\ell}_{c,2}(\boldsymbol{\iota}; \boldsymbol{g})$ (van der Vaart, 2002). (A.28) implies that $\dot{\ell}_{c,2}(\boldsymbol{\iota}; \boldsymbol{g})$ is self-adjoint, and thus, writing $\boldsymbol{C}_* = (C_{*,1}, \ldots, C_{*,m_*})^\mathsf{T}$,

$$
\boldsymbol{w}_c^* = \frac{E\big[\dot{\ell}_{c,1}(\boldsymbol{\iota}; \boldsymbol{g})\{\sum_{j=1}^{m_*} Q_{c,j}(C_{*,j}, X_{*,j}, Z_*, b_*; \boldsymbol{\iota})\}|\boldsymbol{C}_*\big]}{E\big[\{\sum_{j=1}^{m_*} Q_{c,j}(C_{*,j}, X_{*,j}, Z_*, b_*; \boldsymbol{\iota})\}^2|\boldsymbol{C}_*\big]} \tag{A.29}
$$

exists, provided (C1)–(C4), where $\dot{\ell}_{c,1}(\boldsymbol{\iota}; \boldsymbol{g})$ is given in (A.27).

The efficient score of $\boldsymbol{\alpha}$ for the complete likelihood is $\dot{\ell}_{c,1}(\boldsymbol{\iota}; \boldsymbol{g}) - \dot{\ell}_{c,2}(\boldsymbol{\iota}; \boldsymbol{g})[\boldsymbol{w}_c^*]$. The efficient

information of $\boldsymbol{\alpha}$ for the complete likelihood thus takes the form of

$$\begin{aligned} I_c(\boldsymbol{\alpha}) &= E\big\{\dot{\ell}_{c,1}(\boldsymbol{\iota};\boldsymbol{g}) - \dot{\ell}_{c,2}(\boldsymbol{\iota};\boldsymbol{g})[\boldsymbol{w}_c^*]\big\}^{\otimes 2} \\ &= E\left[\sum_{j=1}^{m_*} Q_{c,j}(C_{*,j}, X_{*,j}, Z_*, b_*;\boldsymbol{\iota})\big\{H(C_{*,j})(\boldsymbol{X}_{*,j}{}^{\mathsf{T}}, \boldsymbol{Z}_*{}^{\mathsf{T}}, b_*)^{\mathsf{T}} - \boldsymbol{w}_c^*\big\}\right]^{\otimes 2}, \end{aligned}$$

with $\boldsymbol{w}_c^*$ given in (A.29).

The existence of $\boldsymbol{w}^*$ and the form of $I(\boldsymbol{\alpha})$ can be proved with similar arguments above with different expression of $Q_{c,j}$ to be used in (A.29) based on the observed likelihood function. The proof is thus omitted.

A.2.5 Proof of Theorem A.2

Proof of Theorem A.2. I first notice that $\widehat{\boldsymbol{\iota}}_{c,n}$ maximizes the penalized (complete) likelihood (A.10) rather than an ordinary likelihood, thus $\widehat{\boldsymbol{\iota}}_{c,n}$ does not satisfy the efficient score equation

$$\mathbb{P}_n\big\{\dot{\ell}_{c,1}(\boldsymbol{\iota};\boldsymbol{g}) - \dot{\ell}_{c,2}(\boldsymbol{\iota};\boldsymbol{g})[\boldsymbol{w}_c^*]\big\} = 0.$$

However, if I can show that the distance between $\widehat{\boldsymbol{\alpha}}_{c,n}$ and the efficient estimator is bounded above by $o_p(n^{-1/2})$, then the result follows.

To show this, I first show that

$$\mathbb{P}_n\big\{\dot{\ell}_{c,1}(\widehat{\boldsymbol{\iota}}_{c,n};\boldsymbol{g}) - \dot{\ell}_{c,2}(\widehat{\boldsymbol{\iota}}_{c,n};\boldsymbol{g})[\boldsymbol{w}_c^*]\big\} = o_p(n^{-1/2}) \tag{A.30}$$

which can begin with studying the upper bound of the penalization term. Indeed, if I plug $\big((\widehat{\boldsymbol{\alpha}}_{c,n} + s\boldsymbol{a})^{\mathsf{T}}, \widehat{H}_{c,n} - sw\big)^{\mathsf{T}}$ with $w \in \mathcal{W} \cap \mathcal{H}_n$ satisfying $J(w) < \infty$, into the penalized log-likelihood function (A.10), where $\boldsymbol{a}$ is a p-dimensional vector. Differentiating at $s = 0$, it is shown that

$$\mathbb{P}_n\big\{\dot{\ell}_{c,1}(\widehat{\boldsymbol{\iota}}_{c,n};\boldsymbol{g})^{\mathsf{T}}\boldsymbol{a} - \dot{\ell}_{c,2}(\widehat{\boldsymbol{\iota}}_{c,n};\boldsymbol{g})[w]\big\} + \lambda\int(\widehat{H}_{c,n})^{(q)}(t)w^{(q)}(t)\mathrm{d}t = 0. \tag{A.31}$$

Using the Cauchy-Schwarz inequality, the $\lambda\int(\widehat{H}_{c,n})^{(q)}(t)w^{(q)}(t)\mathrm{d}t$ is bounded by $\lambda J(\widehat{H}_{c,n})J(w)$.

In Theorem A.1, it has been shown

$$J(\widehat{H}_{c,n}) = O_p(1).$$

The reader are also referred to Lemma 7.1 of Murphy and van der Vaart (1999) for extra auxiliary results. Moreover, it is assumed that $\lambda = o_p(n^{-1/2})$, it follows that

$$\lambda J(\widehat{H}_{c,n}) J(w) = o_p(n^{-1/2}). \tag{A.32}$$

As a result, the penalized estimator $\widehat{\boldsymbol{\iota}}_{c,n}$ satisfies the efficient score equation, up to a negligible $o_p(n^{-1/2})$ term. It is obvious to show that (A.31) is free of $\boldsymbol{a}$ and thus

$$\mathbb{P}_n\{\dot{\ell}_{c,1}(\widehat{\boldsymbol{\iota}}_{c,n}; \boldsymbol{g})\} = \mathbf{0}. \tag{A.33}$$

(A.31) and (A.32) together imply that for any $w \in \mathcal{W} \cap \mathcal{H}_n$,

$$\mathbb{P}_n\big\{\dot{\ell}_{c,2}(\widehat{\boldsymbol{\iota}}_{c,n}; \boldsymbol{g})[w]\big\} = o_p(n^{-1/2}). \tag{A.34}$$

I next only need to verify $\mathbb{P}_n\{\dot{\ell}_{c,2}(\widehat{\boldsymbol{\iota}}_{c,n}; \boldsymbol{g})[\boldsymbol{w}_c^*]\} = o_p(n^{-1/2})$ for least favorable direction $\boldsymbol{w}_c^*$. Because each component of $\boldsymbol{w}_c^*$ has a bounded derivative, it is also a function with bounded variation. Using the arguments in Billingsley (1995, pp. 415–416) for functions with bounded variation and the Jackson's Theorem in de Boor (1978, pp. 149), it can be shown that there exits a $\boldsymbol{w}_n \in (\mathcal{W} \cap \mathcal{H}_n)^p$ such that $\|\boldsymbol{w}_n - \boldsymbol{w}_c^*\|_2 = O(n^{-1/(2q+1)})$. Furthermore, I have

$$\Pr\big\{\ell_c\big(\boldsymbol{\alpha}_0, H_0 + s\boldsymbol{a}^{\mathsf{T}}(\boldsymbol{w}_c^* - \boldsymbol{w}_n); \boldsymbol{g}\big)\big\} \leq \Pr\big\{\ell_c(\boldsymbol{\alpha}_0, H_0; \boldsymbol{g})\big\}$$

for s with small absolute value and $\boldsymbol{a} \in \mathbb{R}^p$, then $\Pr\{\dot{\ell}_{c,2}(\boldsymbol{\iota}_0; \boldsymbol{g})[\boldsymbol{w}_c^* - \boldsymbol{w}_n]\} = \mathbf{0}$. Therefore I can write

$$\mathbb{P}_n\{\dot{\ell}_{c,2}(\widehat{\boldsymbol{\iota}}_{c,n}; \boldsymbol{g})[\boldsymbol{w}_c^*]\big\} = I_{1,n} + I_{2,n},$$

where

$$I_{1,n} = (\mathbb{P}_n - \Pr)\big\{\dot{\ell}_{c,2}(\widehat{\boldsymbol{\iota}}_{c,n}; \boldsymbol{g})[\boldsymbol{w}_c^* - \boldsymbol{w}_n]\big\}$$

and

$$I_{2,n} = \Pr\big\{\dot{\ell}_{c,2}(\widehat{\boldsymbol{\iota}}_{c,n}; \boldsymbol{g})[\boldsymbol{w}_c^* - \boldsymbol{w}_n] - \dot{\ell}_{c,2}(\boldsymbol{\iota}_0; \boldsymbol{g})[\boldsymbol{w}_c^* - \boldsymbol{w}_n]\big\}.$$

Let $I_{1,n,k}$ be k-th component of $I_{1,n}$ and denote

$$\boldsymbol{A}_{1,k} = \big\{\dot{\ell}_{c,2}(\boldsymbol{\iota}; \boldsymbol{g})[w_{c,k}^* - w_{n,k}] : \boldsymbol{\iota} \in \Theta \times \mathcal{H}_n,\ w_{n,k} \in \mathcal{W} \cap \mathcal{H}_n \text{ and } \|w_{c,k}^* - w_{n,k}\|_2 \leq v n^{-1/(2q+1)}\big\},$$

$k = 1, \ldots, p$. It can be argued that the ϵ-bracketing numbers associated with $L_2(\Pr)$-norm for Θ, $\mathcal{H}_n$, and $\{w_{n,k} \in \mathcal{W} \cap \mathcal{H}_n : \|w_{c,k}^* - w_{n,k}\|_2 \leq v n^{-1/(2q+1)}\}$ are $v(1/\epsilon)^p$, $v(1/\epsilon)^{vn^{1/(2q+1)}}$, and $v(1/\epsilon)^{vn^{1/(2q+1)}}$, respectively. Therefore, the ϵ-bracketing number for $\boldsymbol{A}_{1,k}$ is bounded by $v(1/\epsilon)^p (1/\epsilon)^{vn^{1/(2q+1)}} (1/\epsilon)^{vn^{1/(2q+1)}}$, which results in $\Pr$-Donsker class for $\boldsymbol{A}_{1,k}$ by Theorem 19.5 in van der Vaart (1998), $k = 1, \ldots, p$. Since

$$\dot{\ell}_{c,2}(\widehat{\boldsymbol{\iota}}_{c,n}; \boldsymbol{g})[w_{c,k}^* - w_{n,k}] \in \boldsymbol{A}_{1,k}$$

and as $n \to \infty$,

$$\Pr\big\{\dot{\ell}_{c,2}(\widehat{\boldsymbol{\iota}}_{c,n}; \boldsymbol{g})[w_{c,k}^* - w_{n,k}]\big\}^2 \leq v \|w_{c,k}^* - w_{n,k}\|_\infty^2 \to 0,$$

then by Corollary 2.3.12 of van der Vaart and Wellner (1996) I have

$$I_{1,n,k} = o_p(n^{-1/2}) \quad k = 1, \ldots, p. \tag{A.35}$$

By the Cauchy-Schwarz inequality and Conditions (C2)–(C5), it can be shown that each compo-

nent of $I_{2,n}$,

$$\begin{aligned} I_{2,n,k} &= \Pr\left\{\dot{\ell}_{c,2}(\widehat{\boldsymbol{\iota}}_{c,n};\boldsymbol{g})[w^*_{c,k}-w_{n,k}] - \dot{\ell}_{c,2}(\boldsymbol{\iota}_0;\boldsymbol{g})[w^*_{c,k}-w_{n,k}]\right\} \\ &\leq v\cdot \operatorname{dist}(\widehat{\boldsymbol{\iota}}_{c,n},\boldsymbol{\iota}_0)\|w^*_{c,k}-w_{n,k}\|_\infty = o_p(n^{-1/2}), \end{aligned} \tag{A.36}$$

$k = 1, \ldots, p$. (A.35) and (A.36) imply that

$$\mathbb{P}_n\{\dot{\ell}_{c,2}(\widehat{\boldsymbol{\iota}}_{c,n};\boldsymbol{g})[w^*_{c,k}]\} = I_{1,n,k} + I_{2,n,k} = o_p(n^{-1/2}), \quad k = 1, \ldots, p. \tag{A.37}$$

Thus, (A.31), (A.33), (A.34), and (A.37) together show that (A.30) holds.

I then show the asymptotic normality and efficiency of the estimator $\widehat{\boldsymbol{\alpha}}_{c,n}$ using Theorem 25.54 in van der Vaart (1998). For notational convenience, in the following, let $\widetilde{\ell}_{c,\boldsymbol{\alpha},H}(\boldsymbol{g})$ denote the semiparametric efficient score function under general $\boldsymbol{\alpha}$ and H for the complete data likelihood. I also write $\Pr_{\boldsymbol{\iota}}\widetilde{\ell}_{c,\widehat{\boldsymbol{\alpha}},\widehat{H}}$ as an abbreviation for $\int \widetilde{\ell}_{c,\widehat{\boldsymbol{\alpha}},\widehat{H}}(\boldsymbol{g})\mathrm{d}\Pr_{\boldsymbol{\iota}}$, which is an integration taken with respect to $\boldsymbol{g}$ only and not with respect to $\widehat{\boldsymbol{\alpha}}$ nor $\widehat{H}$. Under the result of (A.30), I only need to verify conditions

$$\Pr_{\widehat{\boldsymbol{\alpha}}_{c,n},H_0}\widetilde{\ell}_{c,\widehat{\boldsymbol{\alpha}}_{c,n},\widehat{H}_{c,n}} = o_p(n^{-1/2} + \|\widehat{\boldsymbol{\alpha}}_{c,n}-\boldsymbol{\alpha}_0\|), \tag{A.38}$$

and

$$\Pr_0\left\|\widetilde{\ell}_{c,\widehat{\boldsymbol{\alpha}}_{c,n},\widehat{H}_{c,n}} - \widetilde{\ell}_{c,\boldsymbol{\alpha}_0,H_0}\right\|^2 \xrightarrow{\Pr} 0, \qquad \Pr_{\widehat{\boldsymbol{\alpha}}_{c,n},H_0}\|\widetilde{\ell}_{\widehat{\boldsymbol{\alpha}}_{c,n},\widehat{H}_{c,n}}\|^2 = O_p(1). \tag{A.39}$$

For (A.38), in view of the fact that $\Pr_{\boldsymbol{\alpha},H}\widetilde{\ell}_{\boldsymbol{\alpha},H} = 0$ for all $(\boldsymbol{\alpha}, H)$, write

$$\begin{aligned} \Pr_{\widehat{\boldsymbol{\alpha}}_{c,n},H_0}\widetilde{\ell}_{c,\widehat{\boldsymbol{\alpha}}_{c,n},\widehat{H}_{c,n}} &= (\Pr_0 - \Pr_{\boldsymbol{\alpha}_0,\widehat{H}_{c,n}})\widetilde{\ell}_{c,\boldsymbol{\alpha}_0,H_0} + (\Pr_{\widehat{\boldsymbol{\alpha}}_{c,n},H_0} - \Pr_{\widehat{\boldsymbol{\alpha}}_{c,n},\widehat{H}_{c,n}})(\widetilde{\ell}_{c,\widehat{\boldsymbol{\alpha}}_{c,n},\widehat{H}_{c,n}} - \widetilde{\ell}_{c,\boldsymbol{\alpha}_0,H_0}) \\ &\quad + (\Pr_{\boldsymbol{\alpha}_0,\widehat{H}_{c,n}} - \Pr_0 - \Pr_{\widehat{\boldsymbol{\alpha}}_{c,n},\widehat{H}_{c,n}} + \Pr_{\widehat{\boldsymbol{\alpha}}_{c,n},H_0})\widetilde{\ell}_{c,\boldsymbol{\alpha}_0,H_0} \\ &= I_{3,n} + I_{4,n} + I_{5,n}. \end{aligned} \tag{A.40}$$

The definition of efficient score in van der Vaart (1998, pp. 369) shows that $\widetilde{\ell}_{c,\boldsymbol{\alpha}_0,H_0}$ is orthogonal

to all functions in the span of $\dot{\ell}_{c,2}(\boldsymbol{\iota}_0)$. It then yields

$$(\Pr_0 - \Pr_{\boldsymbol{\alpha}_0, \widehat{H}_{c,n}})\widetilde{\ell}_{c,\boldsymbol{\alpha}_0,H_0} = \Pr_0 \widetilde{\ell}_{c,\boldsymbol{\alpha}_0,H_0} \left\{ \frac{p_0 - p_{\boldsymbol{\alpha}_0,\widehat{H}_{c,n}}}{p_0} - \dot{\ell}_{c,2}(\boldsymbol{\alpha}_0, H_0)(H_0 - \widehat{H}_{c,n}) \right\}.$$

Using the Taylor expansion, it is able to show that

$$\left|(\Pr_0 - \Pr_{\boldsymbol{\alpha}_0, \widehat{H}_{c,n}})\widetilde{\ell}_{c,\boldsymbol{\alpha}_0,H_0}\right| \leq \int \left|\widetilde{\ell}_{c,\boldsymbol{\alpha}_0,H_0}\right| \left| \frac{\mathrm{d}^2}{\mathrm{d}s^2} p_{\boldsymbol{\alpha}_0, H_0 + s(\widehat{H}_{c,n} - H_0)} \right| \mathrm{d}\mu$$

for $0 < s < 1$. Straightforward differentiation and Condition (C3) imply that

$$\frac{\mathrm{d}^2}{\mathrm{d}s^2} p_{\boldsymbol{\alpha}_0, H_0 + s(\widehat{H}_{c,n} - H_0)}$$

can be upper bounded by $v(\widehat{H}_{c,n} - H_0)^2$ for a positive constant v independent with $\boldsymbol{g}$ and all s. It follows that $I_{3,n} = O_p(1)\|\widehat{H}_{c,n} - H_0\|_2^2$. By the Taylor expansion, $I_{4,n}$ can be written as

$$\begin{aligned}\int (\widetilde{\ell}_{c,\widehat{\boldsymbol{\alpha}}_{c,n},\widehat{H}_{c,n}} - \widetilde{\ell}_{c,\boldsymbol{\alpha}_0,H_0}) \dot{\ell}_2(\widehat{\boldsymbol{\alpha}}_{c,n}, H_0)(H_0 - \widehat{H}_{c,n}) p_0 \mathrm{d}\mu \\ - \frac{1}{2} \int (\widetilde{\ell}_{c,\widehat{\boldsymbol{\alpha}}_{c,n},\widehat{H}_{c,n}} - \widetilde{\ell}_{c,\boldsymbol{\alpha}_0,H_0}) \frac{\mathrm{d}^2}{\mathrm{d}s^2} p_{\widehat{\boldsymbol{\alpha}}_{c,n}, H_0 + s(\widehat{H}_{c,n} - H_0)} \mathrm{d}\mu.\end{aligned}$$

Since $\widehat{\boldsymbol{\alpha}}_{c,n}$ converges to $\boldsymbol{\alpha}_0$ as shown in Theorem A.1, $|\dot{\ell}_{c,2}(\widehat{\boldsymbol{\alpha}}_{c,n}, H_0)(H_0 - \widehat{H}_{c,n})|$ is upper bounded by $|\widehat{H}_{c,n} - H_0|$, up to a constant not depending on $\boldsymbol{g}$, with probability approaching 1. By Conditions (C2) and (C5), it implies that

$$|\widetilde{\ell}_{c,\widehat{\boldsymbol{\alpha}}_{c,n},\widehat{H}_{c,n}} - \widetilde{\ell}_{c,\boldsymbol{\alpha}_0,H_0}| \leq v \cdot \mathrm{dist}(\widehat{\boldsymbol{\iota}}_{c,n}, \boldsymbol{\iota}_0)^2$$

on an event with probability approaching 1. Moreover, $(\mathrm{d}^2/\mathrm{d}s^2) p_{\widehat{\boldsymbol{\alpha}}_{c,n}, H_0 + s(\boldsymbol{w}_{c,n} h H_{c,n} - H_0)}$ is bounded above by $(\widehat{H}_{c,n} - H_0)^2$, up to a constant, with probability approaching 1. It thus concludes that $I_{4,n} = O_p\big(\|\widehat{H}_{c,n} - H_0\|_2^2 + \|\widehat{\boldsymbol{\alpha}}_{c,n} - \boldsymbol{\alpha}_0\| \|\widehat{H}_{c,n} - H_0\|_2\big)$. I further use the Taylor expansion and the Cauchy-Schwarz inequality to obtain that $I_{5,n} = O_p\big(\|\widehat{\boldsymbol{\iota}}_{c,n} - \boldsymbol{\iota}_0\|_2^2 + \|\widehat{\boldsymbol{\alpha}}_{c,n} - \boldsymbol{\alpha}_0\| \|\widehat{H}_{c,n} - H_0\|_2\big)$.

Therefore, (A.38) follows from the rate of convergence of $\widehat{\boldsymbol{\alpha}}_{c,n}$ and $\widehat{H}_{c,n}$ as shown in Theorem A.1.

For (A.39), I first use the dominated convergence theorem and the consistency of $\widehat{\boldsymbol{\iota}}_{c,n}$ to obtain that $\Pr_0 \left\|\widetilde{\ell}_{\widehat{\boldsymbol{\alpha}}_{c,n},\widehat{H}_{c,n}} - \widetilde{\ell}_{\boldsymbol{\alpha}_0,H_0}\right\|^2 \to 0$ in probability. Furthermore, by the consistency of $\widehat{\boldsymbol{\alpha}}_{c,n}$, it can be shown that $\Pr_{\widehat{\boldsymbol{\alpha}}_{c,n},H_0} \|\widetilde{\ell}_{\widehat{\boldsymbol{\alpha}}_{c,n},\widehat{H}_{c,n}}\|^2 = O_p(1)$ with the similar arguments as to show (A.35). As a result, (A.39) holds. To sum up, it is able to use the results in Theorem 25.54 of van der Vaart (1998), and thus $\widehat{\boldsymbol{\alpha}}_{c,n}$ is efficient. □

A.2.6 Proof of Theorems 2.2 and 2.3

I only need to show the similar result as in Lemma A.3 such that

$$\Pr\{\ell(\boldsymbol{\iota};\boldsymbol{g}) - \ell(\boldsymbol{\iota}_0;\boldsymbol{g})\}^2 \geq v\|\boldsymbol{\iota} - \boldsymbol{\iota}_0\|_\Xi^2\,, \tag{A.41}$$

whenever $\operatorname{dist}(\boldsymbol{\iota},\boldsymbol{\iota}_0) < \varepsilon$ for some constant $\varepsilon > 0$. Indeed, the left hand side of (A.41) can be written as

$$\Pr\left[\log\left\{\int \mathcal{L}_c(\boldsymbol{\iota};\boldsymbol{g})\mathrm{d}b_*\right\} - \log\left\{\int \mathcal{L}_c(\boldsymbol{\iota}_0;\boldsymbol{g})\mathrm{d}b_*\right\}\right]^2 \geq v\|\boldsymbol{\iota} - \boldsymbol{\iota}_0\|_\Xi^2\,. \tag{A.42}$$

Next consider $\mathcal{L}_c\{s\boldsymbol{\iota} + (1-s)\boldsymbol{\iota}_0;\boldsymbol{g}\}$, and then following the proof of Lemma A.3, it can be shown that the left hand side of (A.42) is bounded below by

$$\Pr\left(\frac{(\partial/\partial s)\big[\int \mathcal{L}_c\{s\boldsymbol{\iota} + (1-s)\boldsymbol{\iota}_0;\boldsymbol{g}\}\mathrm{d}b_* - \int \mathcal{L}_c(\boldsymbol{\iota}_0;\boldsymbol{g})\mathrm{d}b_*\big]\big|_{s=\epsilon}}{\int \mathcal{L}_c\{\epsilon\boldsymbol{\iota} + (1-\epsilon)\boldsymbol{\iota}_0;\boldsymbol{g}\}\mathrm{d}b_*}\right)^2,$$

for some $\epsilon \in [0,1]$. By Conditions (C3)–(C5), it thus suffices to show

$$\Pr\left(\int \frac{\partial}{\partial s}\big[\mathcal{L}_c\{s\boldsymbol{\iota} + (1-s)\boldsymbol{\iota}_0;\boldsymbol{g}\} - \mathcal{L}_c(\boldsymbol{\iota}_0;\boldsymbol{g})\big]\Big|_{s=\epsilon}\mathrm{d}b_*\right)^2 \geq v\|\boldsymbol{\iota} - \boldsymbol{\iota}_0\|_\Xi^2.$$

Using the mean value theorem and the proof in van der Vaart (2002, pp. 431), the aforementioned equation is satisfied, which completes the proof of (A.41) as a consequence. The rest of the proof follows the same arguments as in Theorems A.1 and A.2, and thus omitted.

APPENDIX B

APPENDIX FOR CHAPTER 3

B.1 Proof of Theorem 3.1

In the proof I use the second part of Theroem 2.1, and I present it in the following proposition.

Proposition 1. *For any* $\tau, \tau_0 \geq 0$

$$\log\left\{\frac{1-\exp(-\tau)}{1-\exp(-\tau_0)}\right\} \geq (\tau-\tau_0)A_1(\tau_0) - (\tau-\tau_0)^2 A_2(\tau_0) + \log\left(\frac{\tau_0}{\tau}\right) + 1 - \frac{\tau_0}{\tau},$$

where $A_1(\tau_0) = \exp(-\tau_0)/\{1-\exp(-\tau_0)\}$ *and* $A_2(\tau_0) = \exp(-\tau_0)/2\{1-\exp(-\tau_0)\}^2$.

For the proof of proposition 1, please see Appendix A.1.1. Define $u(L_i, \boldsymbol{X}_i) = \sum_{k:t_k \leq L_i} \lambda_k + \boldsymbol{\beta}^T \boldsymbol{Z}_{x_i}(L_i)$, $u(R_i, \boldsymbol{X}_i) = \sum_{k:t_k \leq R_i} \lambda_k + \boldsymbol{\beta}^T \boldsymbol{Z}_{x_i}(R_i)$ and $u(L_i, R_i, \boldsymbol{X}_i) = \sum_{k:L_i < t_k \leq R_i} \lambda_k + \boldsymbol{\beta}^T\{\boldsymbol{Z}_{x_i}(R_i) - \boldsymbol{Z}_{x_i}(L_i)\}$. Now, I can re-write

$$\begin{aligned}
\ell_1(\boldsymbol{\lambda}, \boldsymbol{\beta}) &= \sum_{i=1}^{n} \Delta_{L,i} \log[1 - \exp\{-\sum_{k:t_k \leq L_i} \lambda_k - \boldsymbol{\beta}^T \boldsymbol{Z}_{x_i}(L_i)\}] \\
&= \sum_{i=1}^{n} \Delta_{L,i} \log[1 - \exp\{-u(L_i, \boldsymbol{X}_i)\}] \\
&= \sum_{i=1}^{n} \Delta_{L,i} \left(\log[1 - \exp\{-u_0(L_i, \boldsymbol{X}_i)\}] + \log\left[\frac{1-\exp\{-u(L_i, \boldsymbol{X}_i)\}}{1-\exp\{-u_0(L_i, \boldsymbol{X}_i)\}}\right]\right).
\end{aligned}$$

Now applying proposition 1 to the second term of the above display, I obtain

$$\begin{aligned}
\ell_1(\boldsymbol{\lambda},\boldsymbol{\beta}) \;\geq\; & \sum_{i=1}^{n}\Delta_{L,i}\Bigg(\log[1-\exp\{-u_0(L_i,\boldsymbol{X}_i)\}]++\{u(L_i,\boldsymbol{X}_i)-u_0(L_i,\boldsymbol{X}_i)\}A_1(u_0(L_i,\boldsymbol{X}_i)) \\
& -\{u(L_i,\boldsymbol{X}_i)-u_0(L_i,\boldsymbol{X}_i)\}^2A_2(u_0(L_i,\boldsymbol{X}_i))+\log\left\{\frac{u_0(L_i,\boldsymbol{X}_i)}{u(L_i,\boldsymbol{X}_i)}\right\}+1-\frac{u_0(L_i,\boldsymbol{X}_i)}{u(L_i,\boldsymbol{X}_i)}\Bigg) \\
= & \sum_{i=1}^{n}\Delta_{L,i}\Bigg[\{A_1(u_0(L_i,\boldsymbol{X}_i))+2A_2(u_0(L_i,\boldsymbol{X}_i))u_0(L_i,\boldsymbol{X}_i)\}u(L_i,\boldsymbol{X}_i) \\
& -A_2(u_0(L_i,\boldsymbol{X}_i))u^2(L_i,\boldsymbol{X}_i)+\log\left\{\frac{u_0(L_i,\boldsymbol{X}_i)}{u(L_i,\boldsymbol{X}_i)}\right\}-\frac{u_0(L_i,\boldsymbol{X}_i)}{u(L_i,\boldsymbol{X}_i)}+C_1(u_0(L_i,\boldsymbol{X}_i))\Bigg] \\
= & \sum_{i=1}^{n}\Delta_{L,i}\Bigg[\{A_1(u_0(L_i,\boldsymbol{X}_i))+2A_2(u_0(L_i,\boldsymbol{X}_i))u_0(L_i,\boldsymbol{X}_i)\}\left(\sum_{k:t_k\leq L_i}\lambda_k+\boldsymbol{\beta}^T\boldsymbol{Z}_{x_i}(L_i)\right) \\
& -A_2(u_0(L_i,\boldsymbol{X}_i))\left(\sum_{k:t_k\leq L_i}\lambda_k+\boldsymbol{\beta}^T\boldsymbol{Z}_{x_i}(L_i)\right)^2+\log\left(\frac{u_0(L_i,\boldsymbol{X}_i)}{\sum_{k:t_k\leq L_i}\lambda_k+\boldsymbol{\beta}^T\boldsymbol{Z}_{x_i}(L_i)}\right) \\
& -\left(\frac{u_0(L_i,\boldsymbol{X}_i)}{\sum_{k:t_k\leq L_i}\lambda_k+\boldsymbol{\beta}^T\boldsymbol{Z}_{x_i}(L_i)}\right)+C_1(u_0(L_i,\boldsymbol{X}_i))\Bigg] \qquad \text{(B.1)}
\end{aligned}$$

where $C_1(u_0(L_i,\boldsymbol{X}_i))$ is the constant term that only depends on $u_0(L_i,\boldsymbol{X}_i)$ and it is $C_1(u_0(L_i,\boldsymbol{X}_i))=\log[1-\exp\{-u_0(L_i,\boldsymbol{X}_i)\}]-A_1(u_0(L_i,\boldsymbol{X}_i))u_0(L_i,\boldsymbol{X}_i)-A_2(u_0(L_i,\boldsymbol{X}_i))u_0^2(L_i,\boldsymbol{X}_i)+1$. Next, I look into the following three terms of (B.1). First,

$$\begin{aligned}
-\left(\sum_{t_k\leq L_i}\lambda_k+\boldsymbol{\beta}^T\boldsymbol{Z}_{x_i}(L_i)\right)^2 &= -\left(\sum_{t_k\leq L_i}\frac{\lambda_{k0}}{u_0(L_i,\boldsymbol{X}_i)}\frac{u_0(L_i,\boldsymbol{X}_i)}{\lambda_{k0}}\lambda_k+\frac{\boldsymbol{\beta}_0^T\boldsymbol{Z}_{x_i}(L_i)}{u_0(L_i,\boldsymbol{X}_i)}\frac{u_0(L_i,\boldsymbol{X}_i)}{\boldsymbol{\beta}_0^T\boldsymbol{Z}_{x_i}(L_i)}\boldsymbol{\beta}^T\boldsymbol{Z}_{x_i}(L_i)\right)^2 \\
&\geq -\left\{\sum_{t_k\leq L_i}\frac{u_0(L_i,\boldsymbol{X}_i)}{\lambda_{k0}}\lambda_k^2+\frac{u_0(L_i,\boldsymbol{X}_i)}{\boldsymbol{\beta}_0^T\boldsymbol{Z}_{x_i}(L_i)}(\boldsymbol{\beta}^T\boldsymbol{Z}_{x_i}(L_i))^2\right\},
\end{aligned}$$

where the inequality is obtained by applying Jensen's inequality on the concave function $f(x)=-x^2$ and noting that $\sum_{k:t_k\leq L_i}\lambda_{k0}/u_0(L_i,\boldsymbol{X}_i)+\boldsymbol{\beta}_0^T\boldsymbol{Z}_{x_i}(L_i)/u_0(L_i,\boldsymbol{X}_i)=1$. Second, applying the standard inequality $\log(x)\geq 1-1/x$ for any generic $x>0$ I have

$$\log\left(\frac{u_0(L_i,\boldsymbol{X}_i)}{\sum_{t_k\leq L_i}\lambda_k+\boldsymbol{\beta}^T\boldsymbol{Z}_{x_i}(L_i)}\right)\geq 1-\frac{\sum_{t_k\leq L_i}\lambda_k+\boldsymbol{\beta}^T\boldsymbol{Z}_{x_i}(L_i)}{u_0(L_i,\boldsymbol{X}_i)},$$

and third

$$
\begin{aligned}
-\frac{u_0(L_i,\boldsymbol{X}_i)}{\sum_{t_k\le L_i}\lambda_k+\boldsymbol{\beta}^T\boldsymbol{Z}_{x_i}(L_i)} &= -u_0(L_i,\boldsymbol{X}_i)\Big\{\sum_{t_k\le L_i}\frac{\lambda_{k0}}{u_0(L_i,\boldsymbol{X}_i)}\frac{u_0(L_i,\boldsymbol{X}_i)}{\lambda_{k0}}\lambda_k \\
&\quad +\frac{\boldsymbol{\beta}_0^T\boldsymbol{Z}_{x_i}(L_i)}{u_0(L_i,\boldsymbol{X}_i)}\frac{u_0(L_i,\boldsymbol{X}_i)}{\boldsymbol{\beta}_0^T\boldsymbol{Z}_{x_i}(L_i)}\boldsymbol{\beta}^T\boldsymbol{Z}_{x_i}(L_i)\Big\}^{-1} \\
&\ge -\Big[\sum_{t_k\le L_i}\frac{\lambda_{k0}^2}{u_0(L_i,\boldsymbol{X}_i)}\lambda_k^{-1}+\frac{\{\boldsymbol{\beta}_0^T\boldsymbol{Z}_{x_i}(L_i)\}^2}{u_0(L_i,\boldsymbol{X}_i)}\{\boldsymbol{\beta}^T\boldsymbol{Z}_{x_i}(L_i)\}^{-1}\Big],
\end{aligned}
$$

where the last inequality is obtained by applying Jensen's inequality on the concave function $f(x)=-1/x$. Then using the last three inequalities in (B.1) I obtain $\ell_1(\boldsymbol{\lambda},\boldsymbol{\beta})\ge\ell_{1,\dagger}(\boldsymbol{\lambda},\boldsymbol{\beta}|\boldsymbol{\lambda}_0,\boldsymbol{\beta}_0)\equiv\sum_{k=1}^m\mathcal{M}_{1,1,k}(\lambda_k|\boldsymbol{\lambda}_0,\boldsymbol{\beta}_0)+\mathcal{M}_{1,2}(\boldsymbol{\beta}|\boldsymbol{\lambda}_0,\boldsymbol{\beta}_0)+\mathcal{M}_{1,3}(\boldsymbol{\lambda}_0,\boldsymbol{\beta}_0)$, where for $k=1,\ldots,m$,

$$
\begin{aligned}
\mathcal{M}_{1,1,k}(\lambda_k|\boldsymbol{\lambda}_0,\boldsymbol{\beta}_0) &= \sum_{i=1}^n\Delta_{L,i}\Big[\{A_1(u_0(L_i,\boldsymbol{X}_i))+2A_2(u_0(L_i,\boldsymbol{X}_i))u_0(L_i,\boldsymbol{X}_i)\}\lambda_k \\
&\quad -A_2(u_0(L_i,\boldsymbol{X}_i))\left\{\frac{u_0(L_i,\boldsymbol{X}_i)}{\lambda_{k0}}\right\}\lambda_k^2-\frac{\lambda_k}{u_0(L_i,\boldsymbol{X}_i)}-\frac{\lambda_{k0}^2}{u_0(L_i,\boldsymbol{X}_i)}\lambda_k^{-1}\Big]I(t_k\le L_i),
\end{aligned}
$$

$$
\begin{aligned}
\mathcal{M}_{1,2}(\boldsymbol{\beta}|\boldsymbol{\lambda}_0,\boldsymbol{\beta}_0) &= \sum_{i=1}^n\Delta_{L,i}\Big[\{A_1(u_0(L_i,\boldsymbol{X}_i))+2A_2(u_0(L_i,\boldsymbol{X}_i))u_0(L_i,\boldsymbol{X}_i)\}\boldsymbol{\beta}^T\boldsymbol{Z}_{x_i}(L_i) \\
&\quad -A_2(u_0(L_i,\boldsymbol{X}_i))\frac{u_0(L_i,\boldsymbol{X}_i)}{\boldsymbol{\beta}_0^T\boldsymbol{Z}_{x_i}(L_i)}\{\boldsymbol{\beta}^T\boldsymbol{Z}_{x_i}(L_i)\}^2-\frac{\boldsymbol{\beta}^T\boldsymbol{Z}_{x_i}(L_i)}{u_0(L_i,\boldsymbol{X}_i)} \\
&\quad -\frac{\{\boldsymbol{\beta}_0^T\boldsymbol{Z}_{x_i}(L_i)\}^2}{u_0(L_i,\boldsymbol{X}_i)}\{\boldsymbol{\beta}^T\boldsymbol{Z}_{x_i}(L_i)\}^{-1}\Big],
\end{aligned}
$$

and $\mathcal{M}_{1,3}(\boldsymbol{\lambda}_0,\boldsymbol{\beta}_0)=\sum_{i=1}^n\Delta_{L,i}\{\log[1-\exp\{-u_0(L_i,\boldsymbol{X}_i)\}]-A_1(u_0(L_i,\boldsymbol{X}_i))u_0(L_i,\boldsymbol{X}_i)-A_2(u_0(L_i,\boldsymbol{X}_i))u_0^2(L_i,\boldsymbol{X}_i)+1\}$. Next, consider

$$
\ell_3(\boldsymbol{\lambda},\boldsymbol{\beta}) = \sum_{i=1}^n\Delta_{I,i}\log\left(1-\exp\left[-\sum_{k:L_i<t_k\le R_i}\lambda_k-\boldsymbol{\beta}^T\{\boldsymbol{Z}_{x_i}(R_i)-\boldsymbol{Z}_{x_i}(L_i)\}\right]\right).
$$

Then using the same strategy as that used for finding minorization function for $\ell_1(\boldsymbol{\lambda},\boldsymbol{\beta})$, I obtain

$\ell_3(\boldsymbol{\lambda},\boldsymbol{\beta}) \geq \ell_{3,\dagger}(\boldsymbol{\lambda},\boldsymbol{\beta}|\boldsymbol{\lambda}_0,\boldsymbol{\beta}_0) \equiv \sum_{k=1}^m \mathcal{M}_{3,1,k}(\lambda_k|\boldsymbol{\lambda}_0,\boldsymbol{\beta}_0)+\mathcal{M}_{3,2}(\boldsymbol{\beta}|\boldsymbol{\lambda}_0,\boldsymbol{\beta}_0)+\mathcal{M}_{3,3}(\boldsymbol{\lambda}_0,\boldsymbol{\beta}_0)$, where

$$\begin{aligned}\mathcal{M}_{3,1,k}(\lambda_k|\boldsymbol{\lambda}_0,\boldsymbol{\beta}_0) &= \sum_{i=1}^n \Delta_{I,i}\Bigg[\{A_1(u_0(L_i,R_i,\boldsymbol{X}_i))+2A_2(u_0(L_i,R_i,\boldsymbol{X}_i))u_0(L_i,R_i,\boldsymbol{X}_i)\}\lambda_k \\ &\quad -A_2(u_0(L_i,R_i,\boldsymbol{X}_i))\left\{\frac{u_0(L_i,R_i,\boldsymbol{X}_i)}{\lambda_{k0}}\right\}\lambda_k^2 \\ &\quad -\frac{\lambda_k}{u_0(L_i,R_i,\boldsymbol{X}_i)}-\frac{\lambda_{k0}^2}{u_0(L_i,R_i,\boldsymbol{X}_i)}\lambda_k^{-1}\Bigg]I(L_i<t_k\leq R_i), \quad k=1,\ldots,m,\end{aligned}$$

$$\begin{aligned}\mathcal{M}_{3,2}(\boldsymbol{\beta}|\boldsymbol{\lambda}_0,\boldsymbol{\beta}_0) &= \sum_{i=1}^n \Delta_{I,i}\Bigg(\{A_1(u_0(L_i,R_i,\boldsymbol{X}_i))+2A_2(u_0(L_i,R_i,\boldsymbol{X}_i))u_0(L_i,R_i,\boldsymbol{X}_i)\} \\ &\quad \times\boldsymbol{\beta}^T\{\boldsymbol{Z}_{x_i}(R_i)-\boldsymbol{Z}_{x_i}(L_i)\}-A_2(u_0(L_i,R_i,\boldsymbol{X}_i))\frac{u_0(L_i,R_i,\boldsymbol{X}_i)[\boldsymbol{\beta}^T\{\boldsymbol{Z}_{x_i}(R_i)-\boldsymbol{Z}_{x_i}(L_i)\}]}{\boldsymbol{\beta}_0^T\{\boldsymbol{Z}_{x_i}(R_i)-\boldsymbol{Z}_{x_i}(L_i)\}} \\ &\quad -\frac{\boldsymbol{\beta}^T\{\boldsymbol{Z}_{x_i}(R_i)-\boldsymbol{Z}_{x_i}(L_i)\}}{u_0(L_i,R_i,\boldsymbol{X}_i)}-\frac{[\boldsymbol{\beta}_0^T\{\boldsymbol{Z}_{x_i}(R_i)-\boldsymbol{Z}_{x_i}(L_i)\}]^2}{u_0(L_i,R_i,\boldsymbol{X}_i)\boldsymbol{\beta}^T\{\boldsymbol{Z}_{x_i}(R_i)-\boldsymbol{Z}_{x_i}(L_i)\}}\Bigg),\end{aligned}$$

and

$$\begin{aligned}\mathcal{M}_{3,3}(\boldsymbol{\lambda}_0,\boldsymbol{\beta}_0) &= \sum_{i=1}^n \Delta_{I,i}\Bigg[\log\left\{1-\exp\left(-\left[\sum_{L_i<t_k\leq R_i}\lambda_k+\boldsymbol{\beta}^T\{\boldsymbol{Z}_{x_i}(R_i)-\boldsymbol{Z}_{x_i}(L_i)\}\right]\right)\right\} \\ &\quad -A_1(u_0(L_i,R_i,\boldsymbol{X}_i))u_0(L_i,R_i,\boldsymbol{X}_i)-A_2(u_0(L_i,R_i,\boldsymbol{X}_i))u_0^2(L_i,R_i,\boldsymbol{X}_i)+1\Bigg].\end{aligned}$$

Finally, I obtain

$$\begin{aligned}
\ell(\boldsymbol{\lambda},\boldsymbol{\beta}) &= \ell_1(\boldsymbol{\lambda},\boldsymbol{\beta}) + \ell_2(\boldsymbol{\lambda},\boldsymbol{\beta}) + \ell_3(\boldsymbol{\lambda},\boldsymbol{\beta}) \\
&\geq \ell_\dagger(\boldsymbol{\lambda},\boldsymbol{\beta}|\boldsymbol{\lambda}_0,\boldsymbol{\beta}_0) \\
&\equiv \ell_{1,\dagger}(\boldsymbol{\lambda},\boldsymbol{\beta}|\boldsymbol{\lambda}_0,\boldsymbol{\beta}_0) + \ell_2(\boldsymbol{\lambda},\boldsymbol{\beta}) + \ell_{3,\dagger}(\boldsymbol{\lambda},\boldsymbol{\beta}|\boldsymbol{\lambda}_0,\boldsymbol{\beta}_0) \\
&= \sum_{k=1}^{m} \mathcal{M}_{1,1,k}(\lambda_k|\boldsymbol{\lambda}_0,\boldsymbol{\beta}_0) + \mathcal{M}_{1,2}(\boldsymbol{\beta}|\boldsymbol{\lambda}_0,\boldsymbol{\beta}_0) + \mathcal{M}_{1,3}(\boldsymbol{\lambda}_0,\boldsymbol{\beta}_0) + \ell_2(\boldsymbol{\lambda},\boldsymbol{\beta}) \\
&\quad + \sum_{k=1}^{m} \mathcal{M}_{3,1,k}(\lambda_k|\boldsymbol{\lambda}_0,\boldsymbol{\beta}_0) + \mathcal{M}_{3,2}(\boldsymbol{\beta}|\boldsymbol{\lambda}_0,\boldsymbol{\beta}_0) + \mathcal{M}_{3,3}(\boldsymbol{\lambda}_0,\boldsymbol{\beta}_0) \\
&\equiv \sum_{k=1}^{m} \mathcal{M}_{1,k}(\lambda_k|\boldsymbol{\lambda}_0,\boldsymbol{\beta}_0) + \mathcal{M}_2(\boldsymbol{\beta}|\boldsymbol{\lambda}_0,\boldsymbol{\beta}_0) + \mathcal{M}_3(\boldsymbol{\lambda}_0,\boldsymbol{\beta}_0),
\end{aligned}$$

where $\mathcal{M}_{1,k}(\lambda_k|\boldsymbol{\lambda}_0,\boldsymbol{\beta}_0) = \mathcal{M}_{1,1,k}(\lambda_k|\boldsymbol{\lambda}_0,\boldsymbol{\beta}_0) + \mathcal{M}_{3,1,k}(\lambda_k|\boldsymbol{\lambda}_0,\boldsymbol{\beta}_0) - \lambda_k \sum_{i=1}^{n} \Delta_{I,i} I(t_k \leq L_i)$, $\mathcal{M}_2(\boldsymbol{\beta}|\boldsymbol{\lambda}_0,\boldsymbol{\beta}_0) = \mathcal{M}_{1,2}(\boldsymbol{\beta}|\boldsymbol{\lambda}_0,\boldsymbol{\beta}_0) + \mathcal{M}_{3,2}(\boldsymbol{\beta}|\boldsymbol{\lambda}_0,\boldsymbol{\beta}_0) - \sum_{i=1}^{n} \Delta_{I,i} \boldsymbol{\beta}^T \boldsymbol{Z}_{x_i}(L_i)$, and $\mathcal{M}_3(\boldsymbol{\lambda}_0,\boldsymbol{\beta}_0) = \mathcal{M}_{1,3}(\boldsymbol{\lambda}_0,\boldsymbol{\beta}_0) + \mathcal{M}_{3,3}(\boldsymbol{\lambda}_0,\boldsymbol{\beta}_0)$.