MODELING LONGITUDINAL DATA WITH INTERVAL CENSORED ANCHORING EVENTS

Chenghao Chu

Submitted to the faculty of the University Graduate School

in partial fulfillment of the requirements

for the degree

Doctor of Philosophy

in the Department of Biostatistics,

Indiana University

May, 2018

Accepted by the Graduate Faculty, Indiana University, in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Ying Zhang, Ph.D., Co-chair

Doctoral Committee

Wanzhu Tu, Ph.D., Co-chair

Chunyan He, Sc.D.

March 1, 2018

Giorgos Bakoyannis, Ph.D.

DEDICATION

To:

Nan Jia,

Natalie Chu and Justin Chu,

and my parents.

ACKNOWLEDGMENTS

I am sincerely thankful to Dr. Ying Zhang and Dr. Wanzhu Tu, both of whom are my mentors and friends. As mentors, their solid knowledge, keen insights, and professional skills have benefitted me as a student, and will continue benefitting me as a statistician. As friends, their generosity and intelligence helped me out of those difficult times, and encouraged me to explore new paths in my life-journey.

I am also thankful to Dr. Chunyan He and Dr. Giorgos Bakoyannis for valuable comments on my research, and kindly serving on my research committee.

It has been an enjoyable four years of studying and living in my department. I thank everyone for the friendship.

Dr. Malaz Boustani and the Center for Health Innovation and Implementation Science (CHIIS) provided me financial support and various opportunities to practice my statistical training. I greatly appreciate their generosity.

I thank my parents and my wife for all the support, and my children for all the joy.

Chenghao Chu

MODELING LONGITUDINAL DATA WITH INTERVAL CENSORED ANCHORING EVENTS

In many longitudinal studies, the time scales upon which we assess the primary outcomes are anchored by pre-specified events. However, these anchoring events are often not observable and they are randomly distributed with unknown distribution. Without direct observations of the anchoring events, the time scale used for analysis are not available, and analysts will not be able to use the traditional longitudinal models to describe the temporal changes as desired. Existing methods often make either ad hoc or strong assumptions on the anchoring events, which are unverifiable and prone to biased estimation and invalid inference.

Although not able to directly observe, researchers can often ascertain an interval that includes the unobserved anchoring events, i.e., the anchoring events are interval censored. In this research, we proposed a two-stage method to fit commonly used longitudinal models with interval censored anchoring events. In the first stage, we obtain an estimate of the anchoring events distribution by nonparametric method using the interval censored data; in the second stage, we obtain the parameter estimates as stochastic functionals of the estimated distribution. The construction of the stochastic functional depends on model settings. In this research, we considered two types of models. The first model was a distribution-free model, in which no parametric assumption was made on the distribution of the error term. The second model was likelihood based, which extended the classic mixed-effects models to the situation

that the origin of the time scale for analysis was interval censored. For the purpose of large-sample statistical inference in both models, we studied the asymptotic properties of the proposed functional estimator using empirical process theory. Theoretically, our method provided a general approach to study semiparametric maximum pseudo-likelihood estimators in similar data situations. Finite sample performance of the proposed method were examined through simulation study. Algorithmically efficient algorithms for computing the parameter estimates were provided. We applied the proposed method to a real data analysis and obtained new findings that were incapable using traditional mixed-effects models.

Ying Zhang, Ph.D., Co-Chair

Wanzhu Tu, Ph.D., Co-Chair

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

In many biomedical investigations, the process under study is known to be anchored by certain event of clinical significance. We call such event as anchoring event and its timing of occurrence as anchoring point. Researchers are often interested in quantifying the patterns of the process around the anchoring events. For example, oncologists want to assess the rates of neoplastic growth immediately following the initial tumorigenesis or subsequent tumor recurrence (Fournier et al., 1980; Spratt et al., 1993; Carter et al., 1989). Human growth researchers want to evaluate the rates of skeletal changes before and after pubertal growth spurt (PGS), i.e., the time at which a child's height increase reaches its maximum velocity (Tanner and Whitehouse, 1976). In these examples, tumorigenesis/recurrence and PGS only function as events that anchor the time scale for the analysis. When these events are observed, the time scale for analysis is available, and usual mixed-effects models (Laird and Ware, 1982) or subject-specific smoothing models (Durbán et al., 2005) can be devised to model the trajectory of the process under investigation. In reality, however, the anchoring events are not always directly observed and the individual anchoring points should be regarded as a random quantity with an unknown distribution. In such situation, the time scale for analysis is not available, and all traditional methods can not be directly applied, because the time scale for building those models is no longer available.

We are interested in the situation that the anchoring points are interval censored. Such situation is abundant in real life studies, because investigators are usually able to determine the time intervals within which the anchoring events occur. In literatures, parametric joint models were proposed to analyze data with interval censored anchoring events (van den Hout et al., 2013; Robinson et al., 2010). Imputation methods were also often seen in practice (Shankar et al., 2005; Tu et al., 2009). By assuming all subjects share the same anchoring point, changing-point models (Muller, 1992) was also proposed. But these methods suffer from the following fundamental limitations: (1) The parametric assumptions are not easy to verify, and are prone to biased estimation; (2) They are unable to accommodate the uncertainty associated with the estimation of the anchoring point for each individual subject. Recently, Zhang et al. (2016) proposed a robust nonparametric estimator for post-PGS pubertal growth, which was anchored by the unobservable PGS. Their approach was purely nonparametric and they showed that the estimator was consistent. Although no asymptotic normality theory was established to allow an asymptotic inference procedure, their two-stage estimation method shed a light to overcome the aforementioned difficulties in these problems. First, it completely side-stepped the difficulty of estimating the subject specific anchoring points; Second, by regarding the model estimates as stochastic functionals of the estimated anchoring point distribution, asymptotic distributions of the model estimates were possible to study through the empirical process theory.

In this research, we adopted the two-stage functional estimation method from Zhang et al. (2016) to analyze longitudinal data involving interval censored anchoring points. Two models were considered using the two-stage estimation procedure. The

first model was distribution-free in the sense that we did not make any parametric assumption on the outcomes. The advantages of this method were the robustness against the unknown distribution of the outcomes, and the numerical convenience in computing the parameter estimates. But it completely ignored the possible correlations among the repeated measures within the same subject. Therefore, a second model based on mixed-effects models was proposed. Although statistically more efficient, the second model was numerically more complicated than the first model. For both models, we showed under mild regularity conditions that the model estimates were asymptotically normally distributed with $\sqrt{n}$-convergence rate, and hence statistical inference can be conducted for large samples. Simulation studies were conducted to validate the asymptotic properties, as well as to examine the good finite sample performance. Real data analysis were presented to illustrate the application of the proposed method.

The thesis is structured as follows. Chapter 2 reviews necessary vector calculus needed for studying linear mixed-effects models. Chapter 3 briefly reviews the empirical process theory that is used for this research. In addition, we provide a general theory that is useful to study the large sample property of model estimates, obtained from the proposed two-stage functional estimation procedure. With these theoretic preparations, the distribution-free model is studied in Chapter 4, and the mixed-effects model is studied in Chapter 5. The proposed methods are implemented using R software in Chapter 6. A summary of this work is provided in Chapter 7.

Chapter 2

Introduction to vector calculus

The likelihood function of a mixed-effects model naturally takes the covariance matrix $\boldsymbol{G}$ of the random effects as an argument. To calculate the estimating equations, we must calculate the derivatives with respect to the entries of $\boldsymbol{G}$. One could simply parameterize $\boldsymbol{G}$ by its entries, or some entries since $\boldsymbol{G}$ is symmetric and possibly structured. However, taking derivatives with respect to one entry at a time may be difficult since the likelihood function is complicated, especially when $\boldsymbol{G}$ has a large size. Vector calculus is a mathematical tool that arranges all partial derivatives in an organized manner, so that it takes partial derivatives with respect to all the parameters in $\boldsymbol{G}$ in one step. In other words, vector calculus is a convenient device that helps to calculate the estimating equations for mixed-effects models.

In this chapter, we review some results in vector calculus. These results are well known and can be found in any textbook on vector calculus. We state them in lemmas for the convenience of reference in later chapters.

2.1 Vector calculus

By "vector" we always mean a column vector. If $\boldsymbol{V}$ is a vector, its transpose is denoted by $\boldsymbol{V}^t$, which is a row vector. Both column and row vectors are also regarded as matrices.

In many situations, a scalar function f takes a matrix $\boldsymbol{G} = \left(x_{ij}\right)_{m\times n}$ as an argument. Indeed, f is a multivariate function on the variables $\{x_{ij}\big|1 \leq i \leq m, 1 \leq j \leq n\}$, and hence it makes sense to define the gradient of f with respect to the x_{ij}'s. It is more convenient to write the gradient $\nabla_{\boldsymbol{G}} f$ in a matrix form as

$$\nabla_{\boldsymbol{G}} f = \left(\frac{\partial f}{\partial x_{ij}}\right)_{m\times n}.$$

In particular, if $\boldsymbol{X} = (x_1, \cdots, x_n)$ is a row vector, then we write the gradient

$$\nabla_{\boldsymbol{X}} f = (\frac{\partial f}{\partial x_1}, \cdots, \frac{\partial f}{\partial x_n})$$

as a row vector; if $\boldsymbol{X} = (x_1, \cdots, x_n)^t$ is a column vector, we write the gradient

$$\nabla_{\boldsymbol{X}} f = \left(\frac{\partial f}{\partial x_1}, \cdots, \frac{\partial f}{\partial x_n}\right)^t$$

as a column vector. For a multivariate function $F = (f_1(\boldsymbol{X}), \cdots, f_m(\boldsymbol{X}))^t$, regardless of $\boldsymbol{X}$ being a row or column vector, we always denote the Jacobian of F as

$$\nabla_{\boldsymbol{X}} F = \left(\frac{\partial f_i}{\partial x_j}\right)_{m\times n}$$

which is a matrix. Some useful properties of these operations are summarized in the following lemma, whose validity can be easily checked.

Lemma 2.1.1. (Wand, 2002, Section 3)

- *Product rule. For scalar function f and multivariate functions F, G:*

$$\nabla(F^T G) = G^t(\nabla F) + F^t(\nabla G) \quad \textit{and} \quad \nabla(fG) = G(\nabla f) + f(\nabla G).$$

- *For a constant matrix A and a multivariate function F,*

$$\nabla(AF) = A\nabla F.$$

- *Chain rule. For a multivariate function $\boldsymbol{G}(\boldsymbol{X}) = \big(g_1(\boldsymbol{X}), \cdots, g_m(\boldsymbol{X})\big)^t$ and a multivariable function $F(y_1, \cdots, y_m)$,*

$$\nabla_{\boldsymbol{X}}(F \circ \boldsymbol{G}) = \nabla_{\boldsymbol{Y}} F\Big|_{\boldsymbol{Y}=\boldsymbol{G}(\boldsymbol{X})} \cdot \nabla_{\boldsymbol{X}} G.$$

- *For scalar functions f and h of matrix variable $\boldsymbol{G}$,*

$$\nabla_{\boldsymbol{G}}(fh) = f(\nabla_{\boldsymbol{G}} h) + h(\nabla_{\boldsymbol{G}} f) \textit{ and } \nabla_{\boldsymbol{G}}(f(h(\boldsymbol{G}))) = f'(h(\boldsymbol{G}))(\nabla_{\boldsymbol{G}} h).$$

Another often seen situation is when a matrix is a function of a variable t. That is, $\boldsymbol{A}(t) = \big(a_{ij}(t)\big)$, where the entries $a_{ij}(t)$ are functions of t. The derivative of $\boldsymbol{A}(t)$ is defined as the following matrix.

$$\boldsymbol{A}'(t) = \frac{d\boldsymbol{A}(t)}{dt} = \left(\frac{da_{ij}(t)}{dt}\right).$$

Lemma 2.1.2. (Wand, 2002, Section 3) *For differentiable matrices* $\boldsymbol{A}(t)$ *and* $\boldsymbol{B}(t)$*:*

$$(\boldsymbol{A}(t)+\boldsymbol{B}(t))' = \boldsymbol{A}'(t)+\boldsymbol{B}'(t), \quad (\boldsymbol{A}(t)\boldsymbol{B}(t))' = \boldsymbol{A}(t)\boldsymbol{B}'(t)+\boldsymbol{A}'(t)\boldsymbol{B}(t).$$

The following well-known result is very useful when deriving estimating equations of longitudinal models. A short proof is provided for self-completeness.

Lemma 2.1.3. (Wand, 2002, Section 3) *If* $\boldsymbol{A}(t)$ *is nonsingular, then*

1. **(Jacobi's Formula)** $\frac{d}{dt}(det(\boldsymbol{A}(t))) = det(\boldsymbol{A}(t)) \cdot Tr\left(\boldsymbol{A}(t)^{-1}\frac{d\boldsymbol{A}(t)}{dt}\right)$,
2. $(\boldsymbol{A}^{-1}(t))' = -\boldsymbol{A}^{-1}(t)\boldsymbol{A}'(t)\boldsymbol{A}^{-1}(t)$.

Proof. For the Jacobi's Formula, we have the following direct calculations.

$$\begin{aligned}
& \frac{d}{dt}(det(\boldsymbol{A}(t))) \\
= & \lim_{h\to 0}\Big(det(\boldsymbol{A}(t+h)) - det(\boldsymbol{A}(t))\Big)\Big/ h \\
= & det(\boldsymbol{A}(t)) \lim_{h\to 0}\Big(det(\boldsymbol{A}(t)^{-1}\boldsymbol{A}(t+h)) - 1\Big)\Big/ h \\
= & det(\boldsymbol{A}(t)) \lim_{h\to 0}\Big(det(I + h\boldsymbol{A}(t)^{-1}\frac{d\boldsymbol{A}}{dt} + h^2\boldsymbol{A}^*) - 1\Big)\Big/ h \text{ by Taylor expanson,} \\
& \text{where } \boldsymbol{A}^* \text{ is a bounded matrix on } h. \\
= & det(\boldsymbol{A}(t)) \lim_{h\to 0}\Big(1 + h\cdot Tr(\boldsymbol{A}(t)^{-1}\frac{d\boldsymbol{A}}{dt}) + O(h^2) - 1\Big)\Big/ h \\
= & det(\boldsymbol{A}(t)) \cdot Tr\left(\boldsymbol{A}^{-1}\frac{d\boldsymbol{A}(t)}{dt}\right).
\end{aligned}$$

For the derivative of the inverse matrix, we apply Lemma 2.1.2 to have the following calculations.

$$\begin{aligned} \boldsymbol{I} = \boldsymbol{A}(t)\boldsymbol{A}^{-1}(t) &\implies \frac{d\boldsymbol{I}}{dt} = \frac{d(\boldsymbol{A}(t)\boldsymbol{A}^{-1}(t))}{dt} \\ &\implies \boldsymbol{O} = \boldsymbol{A}'(t)\boldsymbol{A}^{-1}(t) + \boldsymbol{A}(t)\boldsymbol{A}^{-1}(t)' \\ &\implies (\boldsymbol{A}^{-1}(t))' = -\boldsymbol{A}^{-1}(t)\boldsymbol{A}'(t)\boldsymbol{A}^{-1}(t). \end{aligned}$$

□

Chapter 3

A short introduction to empirical process theory

The study of asymptotic properties in our work relies heavily on the modern empirical process theory (Dudley, 1984; Pollard, 1990; van der Vaart and Wellner, 1996; Kosorok, 2007). In this chapter we first review some important definitions and results in empirical process theory, and then present a useful lemma that helps to study the Donsker property of a class of functions. We also provide a general theorem, which provides sufficient conditions for a sequence of stochastic functionals to converge.

By "map" we mean a map between sets. By "metric space" we mean a set together with a notion of distance. By "measurable space" we mean a set together with a σ-algebra, i.e., there is a notion of measurable sets. By "probability space" we mean a set together with a probability measure, i.e, there is a notion of probability of each measurable set. Note that for any metric space, we have a notion of "open sets" (i.e., a topology), and hence a notion of "Borel σ-algebra" as the one generated by the open sets. In other words, any metric space (indeed, topological space) is naturally a measurable space, equipped with the Borel σ-algebra.

3.1 Stochastic process and weak convergence

There are three commonly used equivalent definitions of a stochastic process. Recall that a d-dimensional random variable on a probability space Ω is a measurable function $X : \Omega \to \mathbb{R}^d$. Let T be a set. The first definition of stochastic process simply

regards a stochastic process Γ on Ω indexed by T as a collection of random variables on Ω :

$$\Gamma = \{X_t : \Omega \to \mathbb{R}^d \mid t \in T\}.$$

In different applications, different correlation structures among these random variables can be further imposed.

A second equivalent definition of a stochastic process Γ on Ω indexed by T is to be a map between sets

$$\Gamma : T \times \Omega \to \mathbb{R}^d,$$

which sends (t, ω) to $X_t(\omega)$, such that each restriction X_t, $t \in T$, is a measurable function on Ω. Note that we treat the indexing set T simply as a set without any possible extra structure, such as symmetry or topology.

There is a third equivalent definition. Following Holden et al. (1996, Section 2.1), it is intuitively helpful to think $\omega \in \Omega$ as a particle, $t \in T$ as time, and $X_t(\omega)$ as the position of the particle ω at time t. If we fix ω and let t vary in T, we get a sample path:

$$p_\omega : \ T \to \mathbb{R}^d; \ t \mapsto X_t(\omega).$$

So a stochastic process on Ω can also be identified as a "measurable" function from the probability space Ω to the space of functions $T \to \mathbb{R}^d$, denoted as $(\mathbb{R}^d)^T$. The σ-algebra on $(\mathbb{R}^d)^T$ is generated by the sets:

$$A_{t_1,\cdots,t_k;B_1,\cdots,B_k} \ = \ \{f : T \to \mathbb{R}^d \big| f(t_1) \in B_1, \cdots, f(t_k) \in B_k\}$$

where $k \geq 1$, $t_1, \cdots, t_k$ are points in T, and $B_1, \cdots, B_k$ are Borel sets in $\mathbb{R}^d$.

Each of the above three equivalent definitions has its advantages in different situations. The third definition makes it easier to define weak convergence of a sequence of stochastic processes. For a general discussion of weak convergence, we refer the readers to Pollard (2012, Chpater 5). In this section, we define weak convergence of stochastic processes when there is a metric, denoted as $\|\cdot\|$, on the space $(\mathbb{R}^d)^T$. For example, in most of the problems in this research, the metric is given by the supremum norm $\|\cdot\|_\infty$ on $(\mathbb{R}^d)^T$.

Definition 3.1.1. (Kosorok, 2007, Section 7.2.1) *Let $\Gamma_1, \cdots, \Gamma_n, \cdots$ be a sequence of stochastic processes on Ω indexed by T with values in $\mathbb{R}^d$. The sequence is called weakly convergent to a stochastic process Γ, denoted as*

$$\{\Gamma_i \big| i = 1, \cdots\} \Longrightarrow \Gamma,$$

if for every bounded continuous function $\phi : (\mathbb{R}^d)^T \to \mathbb{R}$, the sequence of random variables $\{\phi(\Gamma_i) \big| i = 1, 2, \cdots\}$ converges to $\phi(\Gamma)$ in distribution. The σ-algebra on $(\mathbb{R}^d)^T$ is the Borel algebra associated with the metric on $(\mathbb{R}^d)^T$.

Example 3.1.2. *(1) A random variable can be thought as a stochastic process indexed by one point. A sequence of random variables $\{X_i \big| i = 1, \cdots\}$ weakly converges to X, viewed as stochastic processes, if and only if they converge to X in distribution, viewed as random variables. This is a special case of the famous Portmanteau Theorem (Klenke, 2008).*

(2) A **Gaussian process** *is a stochastic process such that the random vector $\left(X_{t_1}, \cdots, X_{t_k}\right)$ is multinormal for any finite number of points $t_1, \cdots, t_k$. If $\{\Gamma_i \big| i = 1, \cdots\}$ weakly converges to a Gaussian process Γ, then for any finite number of indices*

$t_1, \cdots, t_m$, the sequence of random vectors $\{(\Gamma_i|_{t_1}, \cdots, \Gamma_i|_{t_m})\big|i = 1, \cdots\}$ converges in distribution to the multinormal random vector $(\Gamma_{t_1}, \cdots, \Gamma_{t_m})$. The converse is true only under certain conditions, such as "tightness" (Kosorok, 2007, Section 7.1).

3.2 Empirical process, Glivenko-Cantelli class, and Donsker class

An empirical process indexed by T is a stochastic process that is defined through random samples (Kosorok, 2007, Page 9). This general definition is rather abstract. In all applications in this thesis, we consider the following special type of empirical processes. Let $\Omega \subset \mathbb{R}^d$ be a probability space with probability measure P. If f is a random variable on Ω, we write the expectation of f as $\mathrm{P}f$. Let $\mathcal{F}$ be a class of measurable functions on Ω, then we have a deterministic functional on $\mathcal{F}$:

$$\mathrm{P}: \ \mathcal{F} \to \mathbb{R}; \ f \mapsto \mathrm{P}f.$$

Now, if we have a random sample of size n on Ω, denoted as $\{\omega_i\big|i = 1, \cdots\}$, we can define an empirical measure $\mathbb{P}_n$ on Ω as

$$\mathbb{P}_n(A) = \frac{1}{n}\sum_{i=1}^{n} 1(\omega_i \in A).$$

For a random variable f on Ω, we write the empirical expectation of f as $\mathbb{P}_n f = \frac{1}{n}\sum_{i=1}^{n} f(\omega_i)$. For a class of measurable functions $\mathcal{F}$ on Ω, we have an empirical process indexed by $\mathcal{F}$:

$$\mathbb{P}_n: \ \mathcal{F} \to \mathbb{R}; \ f \mapsto \mathbb{P}_n f.$$

Empirical process theory mainly focus on studying the weak convergency of the process $\{\mathbb{P}_n \big| n = 1, \cdots\}$ indexed by $\mathcal{F}$, in which the metric on $(\mathbb{R}^d)^{\mathcal{F}}$ is usually the supremum norm. Such a study is not trivial, mainly because the measure $\mathbb{P}_n$ is not absolute continuous with respect to P. Consequently, the Radon-Nikodym derivative $\frac{d\mathbb{P}_n}{d\mathrm{P}}$ does not exist, and hence special analysis tools are required. The following two concepts are extremely useful, as they are used in most applications of empirical process theory. The first concept is related to the consistency of a sequence of empirical processes.

Definition 3.2.1. *A class $\mathcal{F}$ is called* P*-Glivenko-Cantelli if*

$$\|\mathbb{P}_n - \mathrm{P}\|_\infty = \sup_{f \in \mathcal{F}} \left(\left| \mathbb{P}_n f - \mathrm{P} f \right| \right) \xrightarrow{a.s.} 0$$

when $n \to \infty$.

It follows from the above definition that, if $\mathcal{F}$ is P-Glivenko-Cantelli, then for any $\epsilon > 0$, we have:

$$Prob \left(\max_{f \in \mathcal{F}} \{ |\mathbb{P}_n f - \mathrm{P} f| \} < \epsilon \right) \longrightarrow 1.$$

In other words, for n large, the random value $\mathbb{P}_n f$ is close to the fixed value $\mathrm{P} f$ a.s. regardless of the value of f. This is a much stronger statement than point-wise convergence, i.e., for any $f \in \mathcal{F}$, the random variable $\mathbb{P}_n f$ converges to $\mathrm{P} f$ in distribution.

The second concept is related to the limiting process of a sequence of empirical processes. Let $\mathbb{G}_n = \sqrt{n}(\mathbb{P}_n - \mathrm{P})$ be the "centralized" process of $\mathbb{P}_n$. Define a Gaussian process $\mathbb{G}$ indexed by $\mathcal{F}$, to be the process that has a zero mean and covariance

structure given by

$$Cov(\mathbb{G}f_1, \mathbb{G}f_2) = \mathrm{P}(f_1 f_2) - (\mathrm{P}f_1)(\mathrm{P}f_2).$$

Definition 3.2.2. *A class $\mathcal{F}$ is called* P*-Donsker if*

$$\mathbb{G}_n \Longrightarrow \mathbb{G}, \text{ as processes indexed by } \mathcal{F},$$

with respect to the supremum norm on $(\mathbb{R}^d)^{\mathcal{F}}$, when $n \to \infty$.

It follows immediately from the definitions that a P-Donsker class is automatically a P-Glivenko-Cantelli class. Using these two concepts, the study of asymptotic properties of empirical processes reduces to checking whether the indexing class $\mathcal{F}$ is Glivenko-Cantelli or Donsker. To do so, many helpful methods have been developed. For the purpose of our application, the method of "bracketing numbers" and "bracketing entropy" are explained in the remaining part of this section. For more details and more methods in studying the Glivenko-Cantelli property or Donsker property of $\mathcal{F}$, please refer to van der Vaart (1998) for an extensive exploration.

Let $L^p(\Omega)$ be the L^p-space on Ω, which is by definition the set of all measurable functions $f : \Omega \to \mathbb{R}$ such that $E(\left|f\right|^p) < \infty$, together with the L^p-norm

$$\|f\|_p = \left(E(\left|f\right|^p)\right)^{\frac{1}{p}}.$$

An $L^p(\Omega)$-ϵ-bracket (or simply ϵ-bracket if the metric is clear) is a pair of functions (f^-, f^+) such that $f^- \in L^p(\Omega)$, $f^+ \in L^p(\Omega)$, and $f^-(\omega) \leq f^+(\omega)$ for any $\omega \in \Omega$. Let $\mathcal{F}$ be a class of measurable functions and assume that $\mathcal{F} \subset L^p(\Omega)$. A set $\mathcal{X}$ consists

of ϵ-brackets is said to cover $\mathcal{F}$ if for any $f \in \mathcal{F}$, there is a bracket $(f^-, f^+) \in \mathcal{X}$ that covers f, i.e. $f^-(\omega) \leq f(\omega) \leq f^+(\omega)$ for any $\omega \in \Omega$. The $L^p(\Omega)$-ϵ-bracketing number of $\mathcal{F}$ is defined as

$$\min_{\mathcal{X}} \left\{ \text{number of } \epsilon\text{-brackets in } \mathcal{X} \middle| \mathcal{X} \text{ covers } \mathcal{F} \right\}$$

Conventionally, the ϵ-bracketing number of $\mathcal{F}$ is denoted as $N_{[]}(\epsilon, \mathcal{F}, L^p(\mathrm{P}))$. To continue the discussion, we assume that the number $N_{[]}(\epsilon, \mathcal{F}, L^p(\mathrm{P}))$ is finite for any given $\epsilon > 0$. The L^p-bracketing entropy is then defined as the function that sends $\epsilon > 0$ to $\log\left(N_{[]}(\epsilon, \mathcal{F}, L^p(\mathrm{P}))\right)$, which is monotone and hence integrable. The bracketing entropy integral $J_{[]}(\delta, \mathcal{F}, L^p(\mathrm{P}))$ is defined as

$$J_{[]}(\delta, \mathcal{F}, L^p(\mathrm{P})) = \int_0^\delta \sqrt{\log\left(N_{[]}(\epsilon, \mathcal{F}, L^p(\mathrm{P}))\right)}\, d\epsilon.$$

Roughly speaking, $N_{[]}(\epsilon, \mathcal{F}, L^p(\mathrm{P}))$ and $J_{[]}(\delta, \mathcal{F}, L^p(\mathrm{P}))$ measures the size of $\mathcal{F}$ in $L^p(\Omega)$, and the variation of functions in $\mathcal{F}$. These numbers are important because they are useful to study the asymptotic properties of empirical processes indexed by $\mathcal{F}$, as described in the following two results.

Theorem 3.2.3. (Glivenko-Cantelli Theorem (van der Vaart, 1998, Theorem 2.4.1)) *If $N_{[]}(\epsilon, \mathcal{F}, L^1(\mathrm{P})) \leq \infty$ for all $\epsilon > 0$, then $\mathcal{F}$ is* P*-Glivenko-Cantelli.*

Theorem 3.2.4. (Donsker Theorem (van der Vaart, 1998, Theorem 2.5.2)) *If $J_{[]}(1, \mathcal{F}, L^2(\mathrm{P})) \leq \infty$, then $\mathcal{F}$ is* P*-Donsker.*

3.3 Some Donsker classes

In this section, we study the bracketing entropy of serval classes, which is required in the derivation of the asymptotic properties of certain parameter estimates. Readers with no interest in the technique details can skip this section by admitting the results formulated in Lemma 3.3.4.

Let T be a random event time taking values inside a compact interval $[\tau_1, \tau_2]$, where $\tau_1 < \tau_2$ are constants. Let $(L, R]$ be an independent censoring interval of T. That is, there are a sequence of random screening times $\tau_1 = T_1 < T_2 < \cdots < T_n = \tau_2$ that are jointly independent of T, and $L < R$ are the adjacent screening times that bracket T, i.e., $L < T \leq R$. Let F_0 be the cumulative distribution function (CDF) of T. Assume that the censoring interval satisfies the separation condition:

Assumption 3.3.1. (Separation Condition) $P(F_0(R) - F_0(L) \geq c) = 1$ *for some fixed constant* $c > 0$.

The separation condition is almost always satisfied in real studies. It means there is a minimum gap between any adjacent screening times. Unless subjects can be continuously monitored without breaks, such a minimum gap is almost unavoidable.

In real data applications, in addition to the observed censoring interval $(L, R]$, there might be other observed random quantities, for which we denote by $\boldsymbol{W}$. Let H be the joint distribution of the random vector $(\boldsymbol{W}, L, R)$, and P the probability measure associated to H. Assume the following differentiability properties of the distribution of the observable data:

Assumption 3.3.2. *1. the support of* P *is contained in a compact set* Ω.

2. *the Radon–Nikodym derivative* $\frac{dP}{dQ}$ *exists and is bounded over* Ω*, where* Q *is the usual Borel measure.*

3. *the marginal density of* L *and* R *are continuous in* $[\tau_1, \tau_2]$.

In Assumption 3.3.2, the first and last conditions are usually satisfied in biomedical studies, since the event must happen during the subject's life, which is of course bounded. The second condition is hard to verify, but it is implied if H has continuous first order derivatives.

For a small $\delta > 0$, we let $\mathcal{F}_\delta$ denote the class

$$\mathcal{F}_\delta = \Big\{ F \text{ is a CDF over } [\tau_1, \tau_2] \text{ satisfying } \parallel F - F_0 \parallel_\infty < \delta \Big\}.$$

Let $\boldsymbol{\Theta}$ be a compact subset in some Euclidean space, and let $G = G(\boldsymbol{W}, L, R, t; \boldsymbol{\theta})$ be a continuous function where $(\boldsymbol{W}, L, R) \in \Omega$, $t \in [\tau_1, \tau_2]$ and $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. Let $\mathcal{G}_{\boldsymbol{\Theta}, \mathcal{F}_\delta}$ be the following induced class, indexed by $\boldsymbol{\Theta} \times \mathcal{F}_\delta$.

$$\mathcal{G}_{\boldsymbol{\Theta}, \mathcal{F}_\delta} = \left\{ \int_L^R G dF = \int_L^R G(\boldsymbol{W}, L, R, t; \boldsymbol{\theta}) dF(t) \;\Big|\; F \in \mathcal{F}_\delta, \boldsymbol{\theta} \in \boldsymbol{\Theta} \right\}.$$

For the class $\mathcal{G}_{\boldsymbol{\Theta}, \mathcal{F}_\delta}$ to have good properties, we assume

Assumption 3.3.3. $\frac{\partial^2 G}{\partial t^2}$ *exists and is continuous on* $\Omega \times [\tau_1, \tau_2] \times \boldsymbol{\Theta}$.

The following lemma is used in Chapters 4 and 5 to derive the asymptotic properties of our proposed estimators.

Lemma 3.3.4. *Under Assumptions (3.3.2) and (3.3.3), the class* $\mathcal{G}_{\boldsymbol{\Theta}, \mathcal{F}_\delta}$ *is* P*-Donsker for small* δ.

Proof. To prove the theorem, We evaluate $N_{[]}(\epsilon, \mathcal{G}_{\boldsymbol{\Theta},\mathcal{F}_\delta}, L^2(\mathrm{P}))$, the $L^2(\mathrm{P})$-norm ϵ-bracketing number of $\mathcal{G}_{\boldsymbol{\Theta},\mathcal{F}_\delta}$ with respect to the probability measure P. Let K be a constant whose values vary from place to place.

By Theorem 2.7.5 of van der Vaart and Wellner (1996), the family $\mathcal{F}_\delta$ can be covered by N_ϵ number of ϵ-brackets in L^2-norm $\|\cdot\|_2$ with respect to the Borel measure with $N_\epsilon \leq \exp(\frac{K}{\epsilon})$. In other words, there exist pairs of measurable functions

$$\left\{\left(F_i^-(t), F_i^+(t)\right) : i = 1, \cdots N_\epsilon\right\}$$

such that, for any $F \in \mathcal{F}_\delta$, there exists a bracket $\left(F_i^-(t), F_i^+(t)\right)$ satisfying $F_i^-(t) \leq F(t) \leq F_i^+(t)$ and $\| F_i^-(t) - F_i^+(t) \|_2 < \epsilon$. We assume that each bracket contains at least one function F in $\mathcal{F}_\delta$. Otherwise, such a bracket should be removed and results in fewer ϵ-brackets. We can also require that $0 \leq F_i^+ \leq 1$ and $0 \leq F_i^- \leq 1$. It is obvious that there are no more than $\left(\frac{K}{\epsilon}\right)^d$ solid hypercubes $\{Q_1, Q_2, \cdots, Q_K\}$, whose union covers $\boldsymbol{\Theta}$ and whose sides have lengths ϵ.

For any hypercube Q_j and any $t \in [\tau_1, \tau_2]$, define the following functions $S_{j,t}^-$, $S_{j,t}^+$, $S_{j,t}^{\prime -}$ and $S_{j,t}^{\prime +}$ of $(\boldsymbol{Y}, \boldsymbol{W}, L, R) \in \Omega$.

$$\begin{aligned}
S_{j,t}^-(\boldsymbol{Y}, \boldsymbol{W}, L, R) &= \min_{\boldsymbol{\theta} \in Q_j} G(\boldsymbol{Y}, \boldsymbol{W}, L, R, t, \boldsymbol{\theta}), \\
S_{j,t}^+(\boldsymbol{Y}, \boldsymbol{W}, L, R) &= \max_{\boldsymbol{\theta} \in Q_j} G(\boldsymbol{Y}, \boldsymbol{W}, L, R, t, \boldsymbol{\theta}), \\
S_{j,t}^{\prime -}(\boldsymbol{Y}, \boldsymbol{W}, L, R) &= \min_{\boldsymbol{\theta} \in Q_j} \frac{\partial G}{\partial t}(\boldsymbol{Y}, \boldsymbol{W}, L, R, t, \boldsymbol{\theta}), \\
S_{j,t}^{\prime +}(\boldsymbol{Y}, \boldsymbol{W}, L, R) &= \max_{\boldsymbol{\theta} \in Q_j} \frac{\partial G}{\partial t}(\boldsymbol{Y}, \boldsymbol{W}, L, R, t, \boldsymbol{\theta}).
\end{aligned}$$

By Assumption 3.3.3, both G and $\frac{\partial G}{\partial t}$ are continuous on the compact set $\Omega \times \boldsymbol{\Theta}$, and hence absolutely continuous. So $|S^+_{j,t} - S^-_{j,t}| \leq K\epsilon$ and $|S'^+_{j,t} - S'^-_{j,t}| \leq K\epsilon$ for all j and t, where the value of K does not depend on j or t. For any bracket $\left(F^-_i(t), F^+_i(t)\right)$ and hypercube Q_j, we define the following functions of $(\boldsymbol{Y}, \boldsymbol{W}, L, R) \in \Omega$.

$$\begin{aligned} G^-_{ij} &= S^-_{j,R} \cdot \Big(F^-_i(R) \cdot 1(S^-_{j,R} > 0) + F^+_i(R) \cdot 1(S^-_{j,R} \leq 0)\Big) \\ &\quad -S^+_{j,L} \cdot \Big(F^+_i(L) \cdot 1(S^+_{j,L} > 0) + F^-_i(L) \cdot 1(S^+_{j,L} \leq 0)\Big) \\ &\quad - \textstyle\int_L^R S'^+_{j,t} \cdot \Big(F^+_i(t) \cdot 1(S'^+_{j,t} > 0) + F^-_i(t) \cdot 1(S'^+_{j,t} \leq 0)\Big) dt \end{aligned}$$

$$\begin{aligned} G^+_{ij} &= S^+_{j,R} \cdot \Big(F^+_i(R) \cdot 1(S^+_{j,R} > 0) + F^-_i(R) \cdot 1(S^+_{j,R} \leq 0)\Big) \\ &\quad -S^-_{j,L} \cdot \Big(F^-_i(L) \cdot 1(S^-_{j,L} > 0) + F^+_i(L) \cdot 1(S^-_{j,L} \leq 0)\Big) \\ &\quad - \textstyle\int_L^R S'^-_{j,t} \cdot \Big(F^-_i(t) \cdot 1(S'^-_{j,t} > 0) + F^+_i(t) \cdot 1(S'^-_{j,t} \leq 0)\Big) dt. \end{aligned}$$

Although the expressions of G^-_{ij} and G^+_{ij} are complicate, it is easy to see that the summands of these functions bracket the summands of the following integral in order.

$$G_{\boldsymbol{\theta},F} = \int_L^R G\, dF = G\big|_{t=R} \cdot F(R) - G\big|_{t=L} \cdot F(L) - \int_L^R \frac{\partial G}{\partial t} \cdot F\, dt,$$

where $F^-_i \leq F \leq F^+_i$ and $\boldsymbol{\theta} \in Q_j$. So we have $G^-_{ij} \leq G_{\boldsymbol{\theta},F} \leq G^+_{ij}$. In other words, the bracket $\left(G^-_{ij}, G^+_{ij}\right)$ covers $G_{\boldsymbol{\theta},F}$.

Let $\|\cdot\|_{2,\mathrm{P}}$ denote the $L^2(\mathrm{P})$-norm with respect to the probability measure P. The $\|\cdot\|_{2,\mathrm{P}}$-length of the bracket $\left(G^-_{ij}, G^+_{ij}\right)$ is calculated below. Using the fact that $\left|F^+_i\right| \leq 1$, $\left|F^-_i\right| \leq 1$, the obvious relation $AB - CD = A(B - D) + (A - C)D$ and

the regularity condition 3, it follows that

$$
\begin{aligned}
& \|G_{ij}^{+} - G_{ij}^{-}\|_{2,\mathrm{P}} \\
\leq\ & \left\|\max(|S_{j,R}^{+}|)\cdot\left(F_i^{+}(R) - F_i^{-}(R)\right)\right\|_{2,\mathrm{P}} + \left\|\left(S_{j,R}^{+} - S_{j,R}^{-}\right)\right\|_{2,\mathrm{P}} \\
& + \left\|\max(|S_{j,L}^{+}|)\cdot\left(F_i^{+}(L) - F_i^{-}(L)\right)\right\|_{2,\mathrm{P}} + \left\|\left(S_{j,L}^{+} - S_{j,L}^{-}\right)\right\|_{2,\mathrm{P}} \\
& + \left\|\int_L^R \max(|S_{j,t}^{'+}|)\left(F_i^{+}(t) - F_i^{-}(t)\right)dt\right\|_{2,\mathrm{P}} + \left\|\int_L^R \left(S_{j,t}^{+} - S_{j,t}^{-}\right)dt\right\|_{2,\mathrm{P}} \\
\leq\ & \left\|K\cdot\left(F_i^{+}(R) - F_i^{-}(R)\right)\right\|_{2,\mathrm{P}} + \|K\epsilon\|_{2,\mathrm{P}} \\
& + \left\|K\cdot\left(F_i^{+}(L) - F_i^{-}(L)\right)\right\|_{2,\mathrm{P}} + \|K\epsilon\|_{2,\mathrm{P}} \\
& + \left\|\int_L^R K\cdot\left(F_i^{+}(t) - F_i^{-}(t)\right)dt\right\|_{2,\mathrm{P}} + \left\|\int_L^R K\epsilon dt\right\|_{2,\mathrm{P}} \\
\leq\ & K d_R \left\|\left(F_i^{+}(R) - F_i^{-}(R)\right)\right\|_2 + K\epsilon \\
& + K d_L \left\|\left(F_i^{+}(L) - F_i^{-}(L)\right)\right\|_2 + K\epsilon \\
& + \left\|\sqrt{K(R-L)^2\int_L^R \left(F_i^{+}(t) - F_i^{-}(t)\right)^2 dt}\right\|_{2,\mathrm{P}} + K\epsilon\cdot\|R - L\|_{2,\mathrm{P}} \\
\leq\ & K\epsilon + K\epsilon + K\epsilon + K\epsilon + \left\|\sqrt{K\epsilon^2}\right\|_{2,\mathrm{P}} + K\epsilon \leq K\epsilon.
\end{aligned}
$$

where d_L and d_R denote the respective maximum of the marginal densities of L and R.

In summary, we found a total of $(K/\epsilon)^d N_\epsilon$ brackets for $\mathcal{G}_{\boldsymbol{\Theta},\mathcal{F}_\delta}$, each of length $\leq K\epsilon$, where K is independent on ϵ. So $N_{[]}(\epsilon, \mathcal{G}_{\boldsymbol{\Theta},\mathcal{F}_\delta}, L^2(\mathrm{P}))$, the $L^2(\mathrm{P})$-norm ϵ-

bracketing number for $\mathcal{G}_{\boldsymbol{\Theta},\mathcal{F}_\delta}$, is bounded by $(K/\epsilon)^d \exp(K/\epsilon)$. Then it follows that

$$\begin{aligned} & J_{[]}\big(\epsilon, \mathcal{G}_{\boldsymbol{\Theta},\mathcal{F}_\delta}, L^2(\mathrm{P})\big) = \int_0^1 \sqrt{\log\Big(N_{[]}(\epsilon, \mathcal{G}_{\boldsymbol{\Theta},\mathcal{F}_\delta}, L^2(\mathrm{P}))\Big)}d\epsilon \\ \leq \;& \int_0^1 \sqrt{\tfrac{K}{\epsilon} - K\log(\epsilon)}d\epsilon < \infty. \end{aligned}$$

By Theorem 3.2.4, $\mathcal{G}_{\boldsymbol{\Theta},\mathcal{F}_\delta}$ is a P-Donsker class. □

3.4 A general theorem

In this section, we provide a general asymptotic normality theorem for semiparametric maximum pseudo-likelihood estimator. Similar theorems can be found in Kosorok (2007, Theorem 2.11) for nonparametric estimators, and in Wellner and Zhang (2007, Theorem 7.1) for semiparametric maximum likelihood estimators.

We consider a general data situation involving latent variables. Let $\boldsymbol{Y}$ be the random vector of outcomes, and $\boldsymbol{C}$ the random vector of independent variables. Let F_0 denote the unknown CDF of the latent variable T, and $\mathcal{F}$ a class of one dimensional CDF's containing F_0. For example, in the situation considered in Section 5.1, we have $\boldsymbol{C} = (\boldsymbol{W}, L, R)$ and T is the anchoring point. Let $\boldsymbol{\theta}_0$ denote the true value of a d-dimensional parameter of interest, and $\boldsymbol{\Theta}$ a subset in $\mathbb{R}^d$ containing $\boldsymbol{\theta}_0$. Let P denote the probability measure associated with $(\boldsymbol{Y}, \boldsymbol{C})$, and $\mathbb{P}_n$ the empirical measure associated with a random sample of $(\boldsymbol{Y}, \boldsymbol{C})$ of size n. For any P-measurable function f, the integrals $\int f d\mathrm{P}$ and $\int f d\mathbb{P}_n$ are respectively denoted as $\mathrm{P}(f)$ and $\mathbb{P}_n(f)$.

We consider the situation of semiparametric estimation with unbiased estimating equation, i.e., the true parameter $\boldsymbol{\theta}_0$ satisfies

$$\mathrm{P}\big(\boldsymbol{\psi}(\boldsymbol{Y}, \boldsymbol{C}, F_0, \boldsymbol{\theta}_0)\big) = \mathbf{0}$$

where $\boldsymbol{\psi} = \boldsymbol{\psi}(\boldsymbol{Y}, \boldsymbol{C}, F, \boldsymbol{\theta})$ is a d-dimensional estimating function for $\boldsymbol{\theta}$ given a CDF $F \in \mathcal{F}$. When the true CDF F_0 is unknown but a consistent estimator $\hat{F}_n$ of F_0 can be obtained from the data, it leads to an asymptotically unbiased estimating equation

$$\mathbb{P}_n\big(\boldsymbol{\psi}(\boldsymbol{Y}, \boldsymbol{C}, \hat{F}_n, \boldsymbol{\theta})\big) = \mathbf{0},$$

from which a semiparametric maximum pseudo-likelihood estimator $\hat{\boldsymbol{\theta}}_n$ can be obtained.

The following theorem provides sufficient conditions for $\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)$ to converge in distribution. For the sake of convenience in presentation, we define a map $\boldsymbol{\Psi} : \boldsymbol{\Theta} \times \mathcal{F} \to \mathbb{R}$ by setting

$$\boldsymbol{\Psi}(\boldsymbol{\theta}, F) = \mathrm{P}\big(\boldsymbol{\psi}(\boldsymbol{Y}, \boldsymbol{C}, F, \boldsymbol{\theta})\big)$$

for any $(\boldsymbol{\theta}, F) \in \boldsymbol{\Theta} \times \mathcal{F}$. Let $\boldsymbol{\Psi}_n$ be the empirical version of $\boldsymbol{\Psi}$, i.e., $\boldsymbol{\Psi}_n(\boldsymbol{\theta}, F) = \mathbb{P}_n\big(\boldsymbol{\psi}(\boldsymbol{Y}, \boldsymbol{C}, F, \boldsymbol{\theta})\big)$.

Theorem 3.4.1. *Suppose $\boldsymbol{\theta}_0$ satisfies $\boldsymbol{\Psi}(\boldsymbol{\theta}_0, F_0) = \mathbf{0}$. Let $\hat{\boldsymbol{\theta}}_n$ be a solution of $\boldsymbol{\Psi}_n(\boldsymbol{\theta}, \hat{F}_n) = \mathbf{0}$, where $\hat{F}_n$ is an estimate of F_0 from the sample. Assume that*

T1. $\boldsymbol{\theta}_0$ *is an inner point of* $\boldsymbol{\Theta}$*. The function* $\boldsymbol{\theta} \mapsto \boldsymbol{\Psi}(\boldsymbol{\theta}, F_0)$ *has continuous second order derivatives in a neighborhood of* $\boldsymbol{\theta}_0$ *and the matrix* $\boldsymbol{A} = \nabla_{\boldsymbol{\theta}} \boldsymbol{\Psi}(\boldsymbol{\theta}_0, F_0)$ *is nonsingular.*

T2. $\hat{\boldsymbol{\theta}}_n \xrightarrow{P} \boldsymbol{\theta}_0$.

T3. $\sqrt{n}\boldsymbol{\Psi}_n(\boldsymbol{\theta}_0, \hat{F}_n) \xrightarrow{D} \boldsymbol{Z}$ *for some random vector* $\boldsymbol{Z}$.

T4. $\Big(1 + \sqrt{n}\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\|\Big)^{-1} \Big\| \sqrt{n}\big(\boldsymbol{\Psi}(\hat{\boldsymbol{\theta}}_n, F_0) + \boldsymbol{\Psi}_n(\boldsymbol{\theta}_0, \hat{F}_n)\big) \Big\| \xrightarrow{P} 0.$

Then $\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \xrightarrow{D} -\boldsymbol{A}^{-1}\boldsymbol{Z}$.

Proof. Since $\hat{\boldsymbol{\theta}}_n \xrightarrow{P} \boldsymbol{\theta}_0$, the Taylor expansion for $\boldsymbol{\Psi}(\boldsymbol{\theta}, F_0)$ at $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ yields

$$\boldsymbol{\Psi}(\hat{\boldsymbol{\theta}}_n, F_0) = \big(\nabla_{\boldsymbol{\theta}} \boldsymbol{\Psi}(\boldsymbol{\theta}_0, F_0) + \boldsymbol{o}_p(1)\big)(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) = \big(\boldsymbol{A} + \boldsymbol{o}_p(1)\big)(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0). \tag{3.1}$$

and hence

$$\begin{aligned} \sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) &= \big(\boldsymbol{A} + \boldsymbol{o}_p(1)\big)^{-1} \cdot \sqrt{n}\boldsymbol{\Psi}(\hat{\boldsymbol{\theta}}_n, F_0) \\ &= (\boldsymbol{A} + \boldsymbol{o}_p(1))^{-1} \cdot \Big(\sqrt{n}\big(\boldsymbol{\Psi}(\hat{\boldsymbol{\theta}}_n, F_0) + \boldsymbol{\Psi}_n(\boldsymbol{\theta}_0, \hat{F}_n)\big) - \sqrt{n}\boldsymbol{\Psi}_n(\boldsymbol{\theta}_0, \hat{F}_n)\Big). \end{aligned} \tag{3.2}$$

Since $\sqrt{n}\boldsymbol{\Psi}_n(\boldsymbol{\theta}_0, \hat{F}_n) \xrightarrow{D} \boldsymbol{Z}$, we have $\big(\boldsymbol{A}+\boldsymbol{o}_p(1)\big)^{-1} \cdot \big(\sqrt{n}\boldsymbol{\Psi}_n(\boldsymbol{\theta}_0, \hat{F}_n)\big) = \boldsymbol{A}^{-1}\boldsymbol{Z} + \boldsymbol{o}_p(1)$. So Equation (3.2) becomes

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) = \big(\boldsymbol{A}+\boldsymbol{o}_p(1)\big)^{-1} \cdot \sqrt{n}\Big(\boldsymbol{\Psi}(\hat{\boldsymbol{\theta}}_n, F_0) + \boldsymbol{\Psi}_n(\boldsymbol{\theta}_0, \hat{F}_n)\Big) - \boldsymbol{A}^{-1}\boldsymbol{Z} + \boldsymbol{o}_p(1). \tag{3.3}$$

Next we show that the first summand on the right hand side of Equation (3.3) is $\boldsymbol{o}_p(1)$. Note that

$$\begin{aligned}
&\Big\|\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)\Big\| - \Big\|\boldsymbol{A}^{-1}\boldsymbol{Z}\Big\| \le \Big\|\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) + \boldsymbol{A}^{-1}\boldsymbol{Z}\Big\| \\
&= \Big\|\big(\boldsymbol{A} + \boldsymbol{o}_p(1)\big)^{-1} \cdot \sqrt{n}\Big(\boldsymbol{\Psi}(\hat{\boldsymbol{\theta}}_n, F_0) + \boldsymbol{\Psi}_n(\boldsymbol{\theta}_0, \hat{F}_n)\Big) + \boldsymbol{o}_p(1)\Big\| \text{ by Equation (3.3)} \\
&\le o_p(1) \cdot \Big(1 + \Big\|\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)\Big\|\Big) \text{ by Condition T4}
\end{aligned}$$

which implies $\big(1 - o_p(1)\big)\Big\|\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)\Big\| \le \Big\|\boldsymbol{A}^{-1}\boldsymbol{Z}\Big\| + o_p(1)$. So $\Big\|\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)\Big\|$ is bounded in probability, and hence

$$\sqrt{n}\Big(\boldsymbol{\Psi}(\hat{\boldsymbol{\theta}}_n, F_0) + \boldsymbol{\Psi}_n(\boldsymbol{\theta}_0, \hat{F}_n)\Big) = \boldsymbol{o}_p(1) \tag{3.4}$$

by Condition T4 again. Plugging Equation (3.4) into Equation (3.3), we have

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) = \big(\boldsymbol{A} + \boldsymbol{o}_p(1)\big)^{-1} \cdot \boldsymbol{o}_p(1) - \boldsymbol{A}^{-1}\boldsymbol{Z} + \boldsymbol{o}_p(1) = -\boldsymbol{A}^{-1}\boldsymbol{Z} + \boldsymbol{o}_p(1)$$

which completes the proof. □

Remark. Condition T1 in Theorem 3.4.1 is the general regularity condition for parametric models when F_0 is known, which is usually satisfied if the estimating function is a smooth function of $\boldsymbol{\theta}$. Method for verifying Conditions T2 and T3 depends on the specific model setting, and usually requires more efforts with empirical process theory. The following lemma facilitates a set of sufficient conditions to justify Condition T4.

Lemma 3.4.2. *Let $\boldsymbol{\Theta}$ be a compact set that contains $\boldsymbol{\theta}_0$ as an inner point. Let $\|\cdot\|_\infty$ be the supremum norm on $\mathcal{F}$. Assume that*

L1. For any $F \in \mathcal{F}$, the Stieltjes-Lebesgue measure dF exists and is supported in a finite closed interval $[\tau_1, \tau_2]$, where the constants $\tau_1 < \tau_2$ do not depend on F.

L2. $\|\hat{F}_n - F_0\|_\infty = o_p(n^{-1/4})$.

L3. $\sqrt{n}\Big(\boldsymbol{\Psi}_n(\hat{\boldsymbol{\theta}}_n, F) - \boldsymbol{\Psi}(\hat{\boldsymbol{\theta}}_n, F)\Big) - \sqrt{n}\Big(\boldsymbol{\Psi}_n(\boldsymbol{\theta}_0, F) - \boldsymbol{\Psi}(\boldsymbol{\theta}_0, F)\Big) = \boldsymbol{o}_p(1 + \sqrt{n}\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\|)$, uniformly over the class $\mathcal{F}$.

L4. $\boldsymbol{\Psi}(\boldsymbol{\theta}, \hat{F}_n) - \boldsymbol{\Psi}(\boldsymbol{\theta}, F_0) = \int \kappa(\boldsymbol{\theta}, t) d(\hat{F}_n(t) - F_0(t)) + \boldsymbol{O}_p(\|\hat{F}_n - F_0\|_\infty^2)$ uniformly over $\boldsymbol{\Theta}$, where $\kappa(\boldsymbol{\theta}, t) \in C^1\Big(\boldsymbol{\Theta} \times [\tau_1, \tau_2]\Big)$, the set of functions on $\boldsymbol{\Theta} \times [\tau_1, \tau_2]$ that have continuous first-order derivatives.

Then the Condition T4 of Theorem 3.4.1 is satisfied.

Proof. Since $\boldsymbol{\Psi}_n(\hat{\boldsymbol{\theta}}_n, \hat{F}_n) = 0$ and $\boldsymbol{\Psi}(\boldsymbol{\theta}_0, F_0) = 0$, by Condition L3 and triangle inequality, verifying Condition T4 of Theorem 3.4.1 is equivalent to verifying

$$\frac{\Big\|\sqrt{n}(\boldsymbol{\Psi}(\hat{\boldsymbol{\theta}}_n, \hat{F}_n) - \boldsymbol{\Psi}(\hat{\boldsymbol{\theta}}_n, F_0)) - \sqrt{n}(\boldsymbol{\Psi}(\boldsymbol{\theta}_0, \hat{F}_n) - \boldsymbol{\Psi}(\boldsymbol{\theta}_0, F_0))\Big\|}{1 + \sqrt{n}\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\|} \xrightarrow{P} 0.$$

Because $\kappa(\boldsymbol{\theta}, t) \in C^1\Big(\boldsymbol{\Theta} \times [\tau_1, \tau_2]\Big)$, the partial derivative $\nabla_{\boldsymbol{\theta}}\big(\kappa(\boldsymbol{\theta}, t)\big)$ is continuous on the compact set $\boldsymbol{\Theta} \times [\tau_1, \tau_2]$ and hence is uniformly bounded by a constant

K. Using Conditions L2 and L4, we have

$$\begin{aligned}
& \frac{\left\|\sqrt{n}(\boldsymbol{\Psi}(\hat{\boldsymbol{\theta}}_n,\hat{F}_n)-\boldsymbol{\Psi}(\hat{\boldsymbol{\theta}}_n,F_0))-\sqrt{n}(\boldsymbol{\Psi}(\boldsymbol{\theta}_0,\hat{F}_n)-\boldsymbol{\Psi}(\boldsymbol{\theta}_0,F_0))\right\|}{1+\sqrt{n}\|\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0\|} \\
\leq & \frac{\left\|\sqrt{n}\int\big(\kappa(\hat{\boldsymbol{\theta}}_n,t)-\kappa(\boldsymbol{\theta}_0,t)\big)d(\hat{F}_n-F_0)\right\|+\left\|\boldsymbol{O}_p\Big(\sqrt{n}\|\hat{F}_n-F_0\|_\infty^2\Big)\right\|}{1+\sqrt{n}\|\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0\|} \\
\leq & \frac{K\cdot\sqrt{n}\|\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0\|\cdot\|\hat{F}_n-F_0\|_\infty}{1+\sqrt{n}\|\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0\|}+o_p(1) \\
\leq & K\cdot\|\hat{F}_n-F_0\|_\infty+o_p(1)\xrightarrow{P}0,
\end{aligned}$$

which completes the proof. □

Remark. Verifying the conditions in Lemma 3.4.2 is often manageable. In many applications, the class $\mathcal{F}$ consists of CDF's whose densities are supported in a common finite interval. The convergency rate of the estimated distribution function $\hat{F}_n$ is often shown to be faster than $n^{1/4}$. Condition L3 is often satisfied if the class $\mathcal{F}$ is Donsker. Condition L4 is satisfied if $\boldsymbol{\Psi}$ is smooth, which has been verified for applications of interval censored data (Huang and Wellner, 1995; Geskus and Groeneboom, 1999).

Chapter 4

A distribution-free model with interval censored anchoring points

Local rates around the anchoring points are often of clinic importance. For example, tumor growth rate around cancer onset are important to study cancer, and skeleton growth around pubertal growth spurt (PGS) describes the gender differences of adult body shapes. However, there lacks statistical methods to appropriately quantify these local rates around unobserved anchoring points. In this chapter, we introduce our proposed method to model local rates around interval censored anchoring points. The model is distribution-free in the sense that it does not make parametric assumptions for the distributions of the anchoring points and the outcomes. We assume the mean model of the process is locally linear around the anchoring point, which is reasonable when the intervals are not too wide, such as in the real data situation studied in Section 4.4. The model is formulated in Section 4.1, and the asymptotic property of the model estimate is derived in Section 4.2. Finite sample performance of the proposed model is examined through simulation studies in Section 4.3.

4.1 Model formulation and estimation

To illustrate model formulation and for the ease of model interpretation, we consider a pubertal growth study, aimed to quantify the local growth rate in a somatic growth outcome Y, such as wrist circumference or shoulder length, etc, before and after the PGS. More details of this study is provided in Section 4.4.

Suppose there are n independent subjects. For the ith subject, $i = 1, 2, \ldots, n$, the anchoring point T_i is censored by the interval $(U_i, V_i]$. That is, T_i is not observed but is known to satisfy $U_i < T_i \leq V_i$. The outcome of interest Y is assessed at the two end points of the censoring interval, denoted respectively as Y_{U_i} and Y_{V_i}. For convenience, we write the observed data from the ith subject as $W_i = (U_i, V_i, Y_{U_i}, Y_{V_i})$, and we assume that $W_1, W_2, \cdots, W_n$ constitute an independent and identically distributed sample.

The goal of the analysis is to estimate the change rates in Y, immediately before and after the unobserved subject-specific anchoring point T, therein referred to as local rates.

The rates of interest can be modelled by a latent piecewise linear regression model as

$$\begin{cases} Y_U = \lambda + \alpha(U - T) + \epsilon_U, \\ Y_V = \lambda + \beta(V - T) + \epsilon_V, \end{cases} \tag{4.1}$$

where λ is the average value of the response variable Y at the latent time T; α and β are the respective pre and post-anchoring point rates of change; U and V are random observation times bracketing T, and they follow an unspecified joint distribution $H(u, v)$; and ϵ_U and ϵ_V are random errors following an unknown distribution $\psi(\cdot, \cdot)$.

An implicit assumption in the model is that the local growth rates are adequately depicted by this linear model. This assumption is appropriate to address the scientific question pertinent to this study. In the human growth application, growth curves are known to be smooth and the interval that brackets the PGS is relatively tight. An advantage for adopting a linear model is that both pre and post-

anchoring point rates are explicitly specified as model parameters, such as α and β in Model (4.1).

In the absence of observed T, directly fitting Model (4.1) becomes intractable. Let $\boldsymbol{\theta}_0 = (\lambda_0, \alpha_0, \beta_0)^t$ be the true values of the parameters in Model (4.1), and let F_0 be the true distribution of T. Following Zhang et al. (2016), we note that the true parameter $(\boldsymbol{\theta}_0, F_0)$ minimizes the deterministic functional

$$\begin{aligned} & \mathbb{M}(\boldsymbol{\theta}, F) \\ = \; & E_{Y_U, Y_V, U, V}\Big[\big(Y_U - \lambda - \alpha(U - E_{F,U,V}T)\big)^2 + \big(Y_V - \lambda - \beta(V - E_{F,U,V}T)\big)^2\Big] \end{aligned}$$

where $\boldsymbol{\theta} = (\lambda, \alpha, \beta)^t$ ranges over all possible parameters in Model (4.1), F ranges over all cumulative distribution functions (CDF), and $E_{F,U,V}T$ is the conditional expectation of T given $U < T \leq V$ under law F.

An intuitive and logical way to estimate $(\boldsymbol{\theta}_0, F_0)$ is, therefore, to minimize the corresponding stochastic functional

$$\mathbb{M}_n(\boldsymbol{\theta}, F) = \sum_{i=1}^{n}\Big[\big(Y_{U_i} - \lambda - \alpha(U_i - E_{F,U_i,V_i}T)\big)^2 + \big(Y_{V_i} - \lambda - \beta(V_i - E_{F,U_i,V_i}T)\big)^2\Big].$$

Admittedly, minimizing $\mathbb{M}_n(\boldsymbol{\theta}, F)$ jointly over $\boldsymbol{\theta}$ and F is a daunting task computationally. To resolve, we employ a two-stage estimation procedure, which was originally developed by Zhang et al. (2016).

In Stage 1, we obtain the following nonparametric maximum likelihood estimator (NPMLE) of F_0 (Groeneboom and Wellner, 1992), which is denoted as $\hat{F}_n$. By definition, $\hat{F}_n$ is the unique solution that maximizes the following nonparametric

likelihood

$$\hat{F}_n = \arg\max_{F \in \mathcal{F}} \prod_{i=1}^{n} \Big(F(V_i) - F(U_i) \Big),$$

where $\mathcal{F}$ is the class of all stepwise cumulative distribution functions that do not have jumps outside of the set $\{U_1, \cdots, U_n, V_1, \cdots, V_n\}$. The estimate $\hat{F}_n$ can be computed using an efficient hybrid algorithm combining an EM algorithm and an Iterative Convex-Minorant algorithm, recommended by Zhang and Jamshidian (2004).

In Stage 2, we obtain $\hat{\boldsymbol{\theta}}_n = (\hat{\lambda}_n, \hat{\alpha}_n, \hat{\beta}_n)^t$ as an M-estimator of $\boldsymbol{\theta}_0$, by minimizing the plug-in stochastic objective function

$$\begin{aligned} & \mathbb{M}_n(\boldsymbol{\theta}, \hat{F}_n) \\ = & \sum_{i=1}^{n} \Big(Y_{U_i} - \lambda - \alpha \cdot (U_i - E_{\hat{F}_n, U_i, V_i} T) \Big)^2 + \sum_{i=1}^{n} \Big(Y_{V_i} - \lambda - \beta \cdot (V_i - E_{\hat{F}_n, U_i, V_i} T) \Big)^2 \end{aligned}$$

where $E_{\hat{F}_n, U_i, V_i} T$ is the conditional expectation of T given $U_i < T \leq V_i$, under the estimated CDF $\hat{F}_n$. Letting $s_1 < s_2 < \cdots < s_k$ be the set of time points that $\hat{F}_n$ jumps, and letting $\hat{p}_i = \hat{F}_n(s_i) - \hat{F}_n(s_i-)$ be the jump at s_i, we can calculate the expectation term as

$$E_{\hat{F}_n, U_i, V_i} T = \sum_{U_i < s_j \leq V_i} s_j \hat{p}_j \Bigg/ \sum_{U_i < s_j \leq V_i} \hat{p}_j \ .$$

An immediate benefit of using the two-stage model is that $\hat{\boldsymbol{\theta}}_n$ has a closed-form solution. Let

$$\boldsymbol{X}_i(\hat{F}_n) = \begin{pmatrix} 1 & U_i - E_{\hat{F}_n, U_i, V_i} T & 0 \\ 1 & 0 & V_i - E_{\hat{F}_n, U_i, V_i} T \end{pmatrix}, \qquad \boldsymbol{Y}_i = \begin{pmatrix} Y_{U_i} \\ Y_{V_i} \end{pmatrix}.$$

The proposed estimator is essentially the least-square estimator $\hat{\boldsymbol{\theta}}_n$ that minimizes

$$\mathbb{M}_n(\boldsymbol{\theta}, \hat{F}_n) = \sum_{i=1}^{n} \Big(\boldsymbol{Y}_i - \boldsymbol{X}_i(\hat{F}_n)\boldsymbol{\theta}\Big)^t \Big(\boldsymbol{Y}_i - \boldsymbol{X}_i(\hat{F}_n)\boldsymbol{\theta}\Big).$$

It then follows that $\hat{\boldsymbol{\theta}}_n$ has a closed-form solution, given by

$$\hat{\boldsymbol{\theta}}_n = \Big(\sum_{i=1}^{n} \boldsymbol{X}_i(\hat{F}_n)^t \boldsymbol{X}_i(\hat{F}_n)\Big)^{-1} \Big(\sum_{i=1}^{n} \boldsymbol{X}_i(\hat{F}_n)^t \boldsymbol{Y}_i\Big),$$

which can be viewed as a stochastic functional of $\hat{F}_n$, which we denote as $\mathbb{Q}_n(\hat{F}_n)$.

4.2 Asymptotic property

For the purpose of inference, we examine the asymptotic behavior of the stochastic functional estimator $\hat{\boldsymbol{\theta}}_n = \mathbb{Q}_n(\hat{F}_n)$, which is by definition the M-estimator of the stochastic objective function $\mathbb{M}_n(\boldsymbol{\theta}; \hat{F}_n)$.

If the true CDF of the anchoring point F_0 is known, the asymptotic properties of $\tilde{\boldsymbol{\theta}}_n = \mathbb{Q}_n(F_0)$, the M-estimator of $\mathbb{M}_n(\boldsymbol{\theta}; F_0)$, will follow directly from the standard M-estimation theory for parametric models (Huber, 2011).

When F_0 is unknown, as it is the case in the current research, we first obtain its NPMLE $\hat{F}_n$, which converges to F_0 at a rate of $n^{\frac{1}{3}}$ (Groeneboom and Wellner, 1992). In such a situation, development of the asymptotic properties of $\hat{\boldsymbol{\theta}}_n = \mathbb{Q}_n(\hat{F}_n)$, the M-estimator for $\mathbb{M}_n(\boldsymbol{\theta}, \hat{F}_n)$, is more challenging and technically involved with the use of empirical process theory (Kosorok, 2007). The following regularity conditions are required to establish the asymptotic properties of $\hat{\boldsymbol{\theta}}_n$.

C1: There exist constants $\tau_1 < \tau_2 < \infty$ such that the support of the density function f_T of the anchoring point T is contained in $[\tau_1, \tau_2]$.

C2: The true anchoring point T is independent of the random observation interval $(U, V]$ that brackets T.

C3: The support of F_0, the CDF of T, is included in the union of the supports of the CDF of U and the CDF of V.

C4: There exists a constant c such that $P\big(F_0(V) - F_0(U) > c\big) = 1$.

C5: The sum of marginal density functions of U and V, $f_U + f_V$, is strictly positive over $[\tau_1, \tau_2]$.

C6: The joint density function of (U, T, V) is twice differentiable over $[\tau_1, \tau_2]$. In particular, f_U and f_V are differentiable and uniformly bounded over $[\tau_1, \tau_2]$.

C7: The density function f_T is twice differentiable over $[\tau_1, \tau_2]$.

Remark 4.2.1. *Conditions C1-C4 are the general regularity conditions needed to assure consistency and convergence rate of $\hat{F}_n$ (Groeneboom and Wellner, 1992). Conditions C5-C7 are distributional requirements for the observation and anchoring points. These conditions are needed for studying the asymptotic properties of a class of functionals of $\hat{F}_n$ (Geskus and Groeneboom 1999), which helps in the derivation of the asymptotic normality of $\hat{\boldsymbol{\theta}}_n$. In most interval-censored data situations, these conditions are fairly mild and they pose no extra restriction on data analysis.*

Theorem 4.2.2. *Under conditions C1-C7, the functional estimator $\hat{\boldsymbol{\theta}}_n = \mathbb{Q}_n(\hat{F}_n)$ for the parameters in Model (4.1) is consistent and asymptotically normal with a convergence rate of $n^{\frac{1}{2}}$, i.e., $\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \xrightarrow{D} N(\mathbf{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\theta}_0 = (\lambda_0, \alpha_0, \beta_0)^t$ is*

the true value of the parameter vector and

$$\boldsymbol{\Sigma} = \Big[E\left(\boldsymbol{X}(F_0)^{\otimes 2}\right)\Big]^{-1} E\Bigg[\bigg(\Big(\boldsymbol{\Phi}(U,V) + \boldsymbol{X}(F_0)^t \boldsymbol{A}\Big)^t\bigg)^{\otimes 2}\Bigg] \Big[E\Big(\boldsymbol{X}(F_0)^{\otimes 2}\Big)\Big]^{-1},$$

where $\boldsymbol{\Phi}(U,V) = \big(0, \quad \phi_1(U,V), \quad \phi_2(U,V)\big)^t$,

$$\boldsymbol{A} = \begin{pmatrix} \alpha_0(E_{F_0,U,V}T - T) + \epsilon_U \\ \\ \beta_0(E_{F_0,U,V}T - T) + \epsilon_V \end{pmatrix},$$

$$\boldsymbol{X}(F_0) = \begin{pmatrix} 1 & U - E_{F_0,U,V}T & 0 \\ \\ 1 & 0 & V - E_{F_0,U,V}T \end{pmatrix}$$

and we denote $\boldsymbol{M}^t\boldsymbol{M}$ *as* $\boldsymbol{M}^{\otimes 2}$ *for any matrix* $\boldsymbol{M}$. *Functions* ϕ_1 *and* ϕ_2 *are the unique solutions to the following integral equations, respectively.*

$$\begin{aligned} &\int\limits_{U<T\le V} \phi_1(U,V) dH(U,V) \\ = &\int\limits_{U<T\le V} \frac{\Big(\int_U^V s dF_0(s) - T\big(F_0(V) - F_0(U)\big)\Big) E_{F_0,U,V,\boldsymbol{\theta}_0} Y_U}{\big(F_0(V) - F_0(U)\big)^2} dH(U,V|T) \end{aligned}$$

$$\begin{aligned} &\int\limits_{U<T\le V} \phi_2(U,V) dH(U,V) \\ = &\int\limits_{U<T\le V} \frac{\Big(\int_U^V s dF_0(s) - T\big(F_0(V) - F_0(U)\big)\Big) E_{F_0,U,V,\boldsymbol{\theta}_0} Y_V}{\big(F_0(V) - F_0(U)\big)^2} dH(U,V|T) \end{aligned}$$

where $H(U,V|T)$ *is the measure associated with the conditional joint distribution of* U *and* V *for* $U < T \le V$.

Proof. Briefly, the theorem is proved in two steps. First, we show that $\sqrt{n}(\tilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)$ is asymptotically normal. Then, we examine the difference $\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \tilde{\boldsymbol{\theta}}_n) = \sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) - \sqrt{n}(\tilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)$, which is by definition $\sqrt{n}\left(\mathbb{Q}_n(\hat{F}_n) - \mathbb{Q}_n(F_0)\right)$. Using the empirical process theory, we show that this quantity times $E\left(\boldsymbol{X}(F_0)^{\otimes 2}\right)$ is asymptotically equivalent to $\sqrt{n}\big(\boldsymbol{K}(\hat{F}_n) - \boldsymbol{K}(F_0)\big)$, where $\boldsymbol{K}$ is an appropriately defined deterministic smooth functional. Using the general result from Geskus and Groeneboom (1999), we show that $\sqrt{n}\left(\boldsymbol{K}(\hat{F}_n) - \boldsymbol{K}(F_0)\right)$ has an asymptotic linear expansion. Combining the results, we establish the consistency and the asymptotic normality of $\hat{\boldsymbol{\theta}}_n$.

For a given cumulative distribution function (CDF) F, we write

$$\boldsymbol{X}(F) = \begin{pmatrix} 1 & U - E_{F,U,V}T & 0 \\ \\ 1 & 0 & V - E_{F,U,V}T \end{pmatrix},$$

where $(U, V]$ is a random censoring interval with $U < V$ and $F(U) < F(V)$, and $E_{F,U,V}(T) = \int_U^V T dF \Big/ \int_U^V dF$ is the conditional expectation of the event time T given $U < T \leq V$. Let $\boldsymbol{X}_i(F)$ be the $\boldsymbol{X}(F)$ matrix associated with $(U_i, V_i]$ for the ith subject.

When the true CDF (F_0) of the anchoring point T is known, a least square estimator $\tilde{\boldsymbol{\theta}}_n = \mathbb{Q}_n(F_0)$ can be obtained from Stage 2 of the proposed method. By the assumption in Model (1), we have $\boldsymbol{Y}_i = \boldsymbol{X}_i(F_0)\boldsymbol{\theta}_0 + \boldsymbol{A}_i$, where $\boldsymbol{\theta}_0 = (\lambda_0, \alpha_0, \beta_0)^t$ is the true parameter vector and

$$\boldsymbol{A}_i = \left(\alpha_0(E_{F_0,U_i,V_i}T - T_i) + \epsilon_{U_i},\ \beta_0(E_{F_0,U_i,V_i}T - T_i) + \epsilon_{V_i}\right)^t$$

is the vector of $\boldsymbol{A}$, associated with the observation $(U_i, V_i, Y_{U_i}, Y_{V_i})$ and unobserved T_i for the ith subject.

From the explicit formula of $\tilde{\boldsymbol{\theta}}_n = \mathbb{Q}_n(F_0)$ in Section 2, we have

$$\begin{aligned}
\tilde{\boldsymbol{\theta}}_n &= \Big[\sum_{i=1}^{n} \boldsymbol{X}_i(F_0)^t \boldsymbol{X}_i(F_0)\Big]^{-1} \Big[\sum_{i=1}^{n} \boldsymbol{X}_i(F_0)^t \boldsymbol{Y}_i\Big] \\
&= \Big[\sum_{i=1}^{n} \boldsymbol{X}_i(F_0)^t \boldsymbol{X}_i(F_0)\Big]^{-1} \Big[\Big(\sum_{i=1}^{n} \boldsymbol{X}_i(F_0)^t \boldsymbol{X}_i(F_0)\boldsymbol{\theta}_0\Big) + \Big(\sum_{i=1}^{n} \boldsymbol{X}_i(F_0)^t \boldsymbol{A}_i\Big)\Big] \\
&= \boldsymbol{\theta}_0 + \Big[\sum_{i=1}^{n} \boldsymbol{X}_i(F_0)^t \boldsymbol{X}_i(F_0)\Big]^{-1} \Big[\sum_{i=1}^{n} \boldsymbol{X}_i(F_0)^t \boldsymbol{A}_i\Big],
\end{aligned}$$

which implies that

$$\sqrt{n}\big(\tilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\big) = \Big[\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i(F_0)^t \boldsymbol{X}_i(F_0)\Big]^{-1} \Big[\frac{1}{\sqrt{n}}\sum_{i=1}^{n} \boldsymbol{X}_i(F_0)^t \boldsymbol{A}_i\Big].$$

Since $\boldsymbol{X}_i(F_0)$, $i = 1, \cdots, n$, are iid observations of $\boldsymbol{X}(F_0)$, from the Law of Large Numbers we have

$$\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i(F_0)^t \boldsymbol{X}_i(F_0) = E_{F_0}\big(\boldsymbol{X}(F_0)^t \boldsymbol{X}(F_0)\big) + \boldsymbol{o}_P(1).$$

Further, since $\boldsymbol{X}_i(F_0)^t \boldsymbol{A}_i$, $i = 1, \cdots, n$, are iid observations of $\boldsymbol{X}(F_0)^t \boldsymbol{A}$, which has zero mean and finite variance $Var\big(\boldsymbol{X}(F_0)^t \boldsymbol{A}\big)$, from the Central Limit Theorem we have

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} \boldsymbol{X}_i(F_0)^t \boldsymbol{A}_i = \boldsymbol{N}\Big(\boldsymbol{0}, Var\big(\boldsymbol{X}(F_0)^t \boldsymbol{A}\big)\Big) + \boldsymbol{o}_P(1),$$

which is an $\boldsymbol{O}_P(1)$. So we write

$$\begin{aligned} & \sqrt{n}\big(\tilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\big) \\ = \; & \Big[E_{F_0}\big(\boldsymbol{X}(F_0)^t\boldsymbol{X}(F_0)\big)\Big]^{-1}\left[\tfrac{1}{\sqrt{n}}\sum_{i=1}^{n}\boldsymbol{X}_i(F_0)^t\boldsymbol{A}_i\right] + \boldsymbol{o}_P(1), \end{aligned} \tag{4.2}$$

which implies that $\tilde{\boldsymbol{\theta}}_n$ is a consistent estimator of $\boldsymbol{\theta}_0$, and that $\sqrt{n}\big(\tilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\big)$ is asymptotically normally distributed.

When F_0 is unknown, we estimate it by $\hat{F}_n$. The asymptotics of $\hat{\boldsymbol{\theta}}_n = \mathbb{Q}_n(\hat{F}_n)$ must take into account the variation associated with estimation of F_0, in Stage 1 of the procedure. We note that $\sqrt{n}\big(\hat{\boldsymbol{\theta}}_n - \tilde{\boldsymbol{\theta}}_n\big) = (I) + (II)$, where

$$\begin{aligned} (I) \;\; = \;\; & \left[\Big(\tfrac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}_i(\hat{F}_n)^t\boldsymbol{X}_i(\hat{F}_n)\Big)^{-1} - \Big(\tfrac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}_i(F_0)^t\boldsymbol{X}_i(F_0)\Big)^{-1}\right] \\ & \times\left[\tfrac{1}{\sqrt{n}}\sum_{i=1}^{n}\boldsymbol{X}_i(\hat{F}_n)^t\boldsymbol{Y}_i\right] \\ (II) \;\; = \;\; & \left[\tfrac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}_i(F_0)^t\boldsymbol{X}_i(F_0)\right]^{-1}\left[\tfrac{1}{\sqrt{n}}\sum_{i=1}^{n}\Big(\boldsymbol{X}_i(\hat{F}_n) - \boldsymbol{X}_i(F_0)\Big)^t\boldsymbol{Y}_i\right]. \end{aligned}$$

We claim that there exists an influence function $\boldsymbol{\Phi} = \boldsymbol{\Phi}(U, V)$ with a zero mean and a finite variance such that

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Big(\boldsymbol{X}_i(\hat{F}_n) - \boldsymbol{X}_i(F_0)\Big)^t\boldsymbol{Y}_i = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\boldsymbol{\Phi}(U_i, V_i) + \boldsymbol{o}_p(1). \tag{$*$}$$

For narrative convenience, we first complete the proof assuming that Claim $(*)$ is true. We then prove the claim.

Since $\hat{F}_n$ converges uniformly to F_0 (Groeneboom and Wellner, 1992) and the matrix $E_{F_0}\big(\boldsymbol{X}(F_0)^t\boldsymbol{X}(F_0)\big)$ is nonsingular, it follows that the first factor in (I) is an

$\boldsymbol{o}_p(1)$:

$$\left[\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}_i(\hat{F}_n)^t\boldsymbol{X}_i(\hat{F}_n)\right]^{-1} - \left[\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}_i(F_0)^t\boldsymbol{X}_i(F_0)\right]^{-1} = \boldsymbol{o}_p(1).$$

By Claim $(*)$ and the fact that $\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\boldsymbol{X}_i(F_0)^t\boldsymbol{Y}_i = \boldsymbol{O}_P(1)$, the second factor in (I) is an $\boldsymbol{O}_P(1)$:

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\boldsymbol{X}_i(\hat{F}_n)^t\boldsymbol{Y}_i = \boldsymbol{O}_P(1).$$

So we proved that $(I)= \boldsymbol{o}_p(1)\cdot\boldsymbol{O}_p(1) = \boldsymbol{o}_p(1)$, and hence

$$\begin{aligned}
&\sqrt{n}\big(\hat{\boldsymbol{\theta}}_n - \tilde{\boldsymbol{\theta}}_n\big)\\
=\ & \boldsymbol{o}_p(1) + \left[\tfrac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}_i(F_0)^t\boldsymbol{X}_i(F_0)\right]^{-1}\left[\tfrac{1}{\sqrt{n}}\sum_{i=1}^{n}\big(\boldsymbol{X}_i(\hat{F}_n) - \boldsymbol{X}_i(F_0)\big)^t\boldsymbol{Y}_i\right]\\
=\ & \boldsymbol{o}_p(1) + \Big[E_{F_0}\big(\boldsymbol{X}(F_0)^t\boldsymbol{X}(F_0)\big)\Big]^{-1}\left[\tfrac{1}{\sqrt{n}}\sum_{i=1}^{n}\boldsymbol{\Phi}(U_i,V_i)\right].
\end{aligned}$$

Combining the above equation and Equation (4.2), we have

$$\begin{aligned}
&\sqrt{n}\big(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\big) = \sqrt{n}\big(\hat{\boldsymbol{\theta}}_n - \tilde{\boldsymbol{\theta}}_n\big) + \sqrt{n}\big(\tilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\big)\\
=\ & \Big[E_{F_0}\big(\boldsymbol{X}(F_0)^t\boldsymbol{X}(F_0)\big)\Big]^{-1}\left[\tfrac{1}{\sqrt{n}}\sum_{i=1}^{n}\big(\boldsymbol{\Phi}(U_i,V_i) + \boldsymbol{X}_i(F_0)^t\boldsymbol{A}_i\big)\right] + \boldsymbol{o}_p(1).
\end{aligned}$$

Since $\boldsymbol{\Phi}(U_i,V_i) + \boldsymbol{X}_i(F_0)^t\boldsymbol{A}_i$, $i = 1,\cdots,n$, are iid observations of the random vector $\boldsymbol{\Phi}(U,V) + \boldsymbol{X}(F_0)^t\boldsymbol{A}$, which has a zero mean and a finite variance, we see that $\hat{\boldsymbol{\theta}}_n$ is a consistent estimator of $\boldsymbol{\theta}_0$ and $\sqrt{n}\big(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\big)$ is asymptotically normal by the Central Limit Theorem. This completes the proof for Theorem 4.2.2.

The rest of this proof uses the empirical process theory to prove Claim $(*)$.

Denote the empirical and the true probability measures for the random vector (U, V, Y_U, Y_V) by $\mathbb{P}_n$ and P, respectively. Let C be a constant, whose value varies from place to place throughout the proof. For a small $\delta > 0$, we let $\mathcal{F}_\delta$ denote the class

$$\mathcal{F}_\delta = \Big\{ F \text{ is a CDF over } [\tau_1, \tau_2] \text{ satisfying } \| F - F_0 \|_\infty < \delta \Big\}.$$

Considering the stochastic process $\boldsymbol{U}$ indexed by F in the class $\mathcal{F}_\delta$:

$$\begin{aligned} \boldsymbol{U}(F) &= \Big(\big(Y_U, Y_V\big) \boldsymbol{X}(F) \Big)^t \\ &= \Big(Y_U + Y_V, \big(U - E_{F,U,V} T\big) Y_U, \big(V - E_{F,U,V} T\big) Y_V \Big)^t. \end{aligned}$$

Using $\boldsymbol{U}$, we can rewrite the left hand side of the equation in Claim $(*)$ as

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \Big(\boldsymbol{X}_i(\hat{F}_n)^t - \boldsymbol{X}_i(F_0)^t \Big) \boldsymbol{Y}_i = \sqrt{n} \mathbb{P}_n \big(\boldsymbol{U}(\hat{F}_n) - \boldsymbol{U}(F_0) \big) = (III) + (IV),$$

where $(III) = \sqrt{n}(\mathbb{P}_n - \mathrm{P})\big(\boldsymbol{U}(\hat{F}_n) - \boldsymbol{U}(F_0)\big)$ and $(IV) = \sqrt{n}\mathrm{P}\big(\boldsymbol{U}(\hat{F}_n) - \boldsymbol{U}(F_0)\big)$.

First, we show that (III) is an $\boldsymbol{o}_p(1)$. Consider the following class $\mathcal{G}$ induced by the class $\mathcal{F}$:

$$\mathcal{G} = \Big\{ G_F(a, b) = E_{F,a,b} T : \;\; F \in \mathcal{F}_\delta \text{ with } F_0(b) - F_0(a) > c \text{ for } a, b \in [\tau_1, \tau_2] \Big\}$$

where c is the constant given by the regularity condition C4. By Lemma 3.3.4 and van der Vaart and Wellner (1996, Theorem 2.10.6), $\mathcal{G}$ is P-Donsker, which further implies that the class

$$\{ \boldsymbol{U}(F) - \boldsymbol{U}(F_0) : \;\; F \in \mathcal{F}_\delta \}$$

is P-Donsker as well. By the uniform $n^{\frac{1}{3}}$-convergency of $\hat{F}_n$ (Groeneboom and Wellner, 1992, Section 4.3), we have $\hat{F}_n \in \mathcal{F}_\delta$ and $\mathrm{P}\left(\boldsymbol{U}(\hat{F}_n) - \boldsymbol{U}(F_0)\right)^2 \to 0$ as $n \to \infty$. So we have

$$(III) = \sqrt{n}(\mathbb{P}_n - \mathrm{P})\left(\boldsymbol{U}(\hat{F}_n) - \boldsymbol{U}(F_0)\right) = \boldsymbol{o}_p(1)$$

by van der Vaart and Wellner (1996, Corollary 2.3.12).

Second, we evaluate $(IV) = \sqrt{n}\mathrm{P}\big(\boldsymbol{U}(\hat{F}_n) - \boldsymbol{U}(F_0)\big)$. Direct calculation yields

$$\sqrt{n}\mathrm{P}\left(\boldsymbol{U}(\hat{F}_n) - \boldsymbol{U}(F_0)\right) = \sqrt{n}\mathrm{P}\Big(\big(E_{F_0,U,V}T - E_{\hat{F}_n,U,V}T\big)(0, Y_U, Y_V)^t\Big) \tag{4.3}$$

Using Taylor expansion, the regularity condition (C4) and $n^{\frac{1}{3}}$-rate of convergence of $\hat{F}_n$, the second entry of the vector in Equation (4.3) can be written as

$$\begin{aligned}
& \sqrt{n}\mathrm{P}\left((E_{F_0,U,V}T - E_{\hat{F}_n,U,V}T)Y_U\right) \\
= \;& \sqrt{n}\int\left(\frac{\int_U^V T dF_0}{\int_U^V dF_0} - \frac{\int_U^V T d\hat{F}_n}{\int_U^V d\hat{F}_n}\right) Y_U d\mathrm{P} \\
= \;& \sqrt{n}\int\left(\frac{\int_U^V T dF_0 \cdot \int_U^V d\hat{F}_n - \int_U^V dF_0 \cdot \int_U^V T d\hat{F}_n}{\left(\int_U^V dF_0\right)^2}\right) Y_U d\mathrm{P} + o_P(1) \\
= \;& \sqrt{n}\big(\boldsymbol{K}(\hat{F}_n) - \boldsymbol{K}(F_0)\big) + o_p(1)
\end{aligned}$$

where $\boldsymbol{K}$ is the linear functional on $\mathcal{F}_\delta$ defined by

$$\begin{aligned}
\boldsymbol{K}(F) &= \int \frac{\Big(\int_U^V T dF_0 \cdot \int_U^V dF - \int_U^V dF_0 \cdot \int_U^V T dF\Big) E_{F_0,U,V,\boldsymbol{\theta}_0} Y_U}{\left(\int_U^V dF_0\right)^2} dH(U,V) \\
&= \int\left(\int_{U<T\le V} \frac{\Big(\int_U^V s dF_0(s) - T\big(F_0(V) - F_0(U)\big)\Big) E_{F_0,U,V,\boldsymbol{\theta}_0} Y_U}{\big(F_0(V) - F_0(U)\big)^2} dH(U,V|T)\right) dF(T)
\end{aligned}$$

where $H(U, V|T)$ is the measure associated with the conditional joint distribution of U and V given $U < T \leq V$.

To study the asymptotic distribution of $\sqrt{n}\big(\boldsymbol{K}(\hat{F}_n) - \boldsymbol{K}(F_0)\big)$, we verify the regularity conditions M1-M3, D1-D4 and F1-F3 stated in Geskus and Groeneboom (1999): Condition M1 requires the true CDF F_0 to be absolutely continuous with bounded derivative, which is implied by our regularity conditions C1 and C7. Condition M2 requires the censoring interval to be independent on the anchoring point and the joint density of (U, V) is absolutely continuous with respect to the two dimensional Lebesgue measure, which is implied by our regularity conditions C2 and C6. Condition M3 requires F_0 to have no mass where the marginal densities of U and V are simultaneously zero, which is implied by our regularity condition C3. Condition D1 requires the marginal density of U and V not to be simultaneously zero over $[\tau_1, \tau_2]$, which is equivalent to our regularity condition C5. Condition D2 requires the joint density of (U, V) to be differentiable with bounded derivatives, which is implied by our regularity condition C6. Condition D3 requires F_0 to be differentiable except for at most finitely many jumps, and the left/right-derivatives to be bounded, which is implied by our regularity condition C7. Conditions F1 and F2 together require the functional $\boldsymbol{K}$ to be Hellinger differentiable at F_0 with a higher order term controlled by the squared distance, more precisely,

$$\boldsymbol{K}(G) - \boldsymbol{K}(F_0) = \int \kappa_{F_0}(t) d(G - F_0)(t) + O(\|G - F_0\|^2)$$

for some function $\kappa_{F_0}(t)$ (called the canonical gradient) and an appropriate metric $\|\cdot\|$ (in our case, the $\|\cdot\|_\infty$ norm is used). Since our functional $\boldsymbol{K}$ is linear, by the

arguments given on Pages 631-632 in Geskus and Groeneboom (1999), Conditions F1 and F2 are automatically satisfied, and the canonical gradient of $\boldsymbol{K}$ at F_0 can be computed using Proposition A.5.2 in Bickel et al. (1998) as

$$\kappa_{F_0}(T) = \int\limits_{U<T\leq V} \frac{\Big(\int_U^V sdF_0(s) - T\big(F_0(V) - F_0(U)\big)\Big) E_{F_0,U,V,\boldsymbol{\theta}_0} Y_U}{\big(F_0(V) - F_0(U)\big)^2} dH(U,V|T).$$

Condition F3 requires $\kappa_{F_0}(T)$ to have bounded derivatives over $[\tau_1, \tau_2]$, which can be verified through calculus based algebra using our regularity conditions C6 and C7.

With all conditions met, from Theorem 3.2 of Geskus and Groeneboom (1999), we know that there is an influence function ϕ_1 such that

$$\begin{aligned}\sqrt{n}\boldsymbol{P}\Big((E_{F_0,U,V}T - E_{\hat{F}_n,U,V}T)Y_U\Big) &= \sqrt{n}\big(\boldsymbol{K}(\hat{F}_n) - \boldsymbol{K}(F_0)\big) + o_P(1) \\ &= \tfrac{1}{\sqrt{n}}\sum_{i=1}^{n}\phi_1(U_i,V_i) + o_p(1),\end{aligned}$$

where $\phi_1(U,V)$ has a zero mean and a finite variance, and it is uniquely determined by the integral equation

$$\int\limits_{U<T\leq V} \phi_1(U,V)dH(U,V) = \kappa_{F_0}(T)$$

by Geskus and Groeneboom (1999, Corollary 2.1) and van der Vaart (1991, Theorem 3.1).

Similarly, the last entry of the vector in Equation (4.3) can be expressed as

$$\sqrt{n}\boldsymbol{P}\Big((E_{F_0,U,V}T - E_{\hat{F}_n,U,V}T)Y_V\Big) = \tfrac{1}{\sqrt{n}}\sum_{i=1}^{n}\phi_2(U_i,V_i) + o_p(1),$$

for some influence function ϕ_2 such that $\phi_2(U, V)$ has a zero mean and a finite variance, and it is uniquely determined by the integral equation given in Theorem 4.2.2.

In summary, we have shown that

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \Big(\boldsymbol{X}_i (\hat{F}_n)^t - \boldsymbol{X}_i (F_0)^t \Big) \boldsymbol{Y}_i = (III) + (IV) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \boldsymbol{\Phi}(U_i, V_i) + \boldsymbol{o}_p(1)$$

where $\boldsymbol{\Phi} = (0, \phi_1, \phi_2)^t$. This completes the proof of Claim $(*)$, and hence the proof of Theorem 4.2.2. □

We finish this section with a few remarks regarding the application of Theorem 4.2.2.

Remark 4.2.3. *Given its complicated structure, direct evaluation of the asymptotic variance matrix* $\boldsymbol{\Sigma}$ *is difficult. Since asymptotic normality is established and* $\hat{\boldsymbol{\theta}}_n$ *is relatively easy to compute, it is usually more convenient to use a resampling method to estimate* $\boldsymbol{\Sigma}$*. Here we estimate* $\boldsymbol{\Sigma}$ *by using a nonparametric bootstrap method. Specifically, for a data set containing n subjects, we draw bootstrap resamples containing n subjects from the original sample with equal weight and with replacement. We obtain a prespecified number* $(b = 1, \ldots, B)$ *of resamples independently, and from which we calculate* B *estimates* $\hat{\boldsymbol{\theta}}_n^{(b)}, b = 1, \ldots, B$*. We use the sample variance matrix of the estimated sample mean of these estimates* $\hat{\boldsymbol{\theta}}_n^{(b)}, b = 1, \ldots, B$*, to approximate* $\boldsymbol{\Sigma}$*; such a variance estimate is known to be consistent (Efron and Tibshirani, 1994).*

Remark 4.2.4. *Finally, we note that the proposed model is further generalizable in multiple ways. First, covariates could be included. Inclusion of covariates does not fundamentally change the proof of theorem, except for the involvement of more*

complicate algebraic operations. Second, the main theoretical results hold for more complicated functions of $U - T$ and $V - T$. A more general model can be written as

$$\begin{cases} Y_U = \boldsymbol{\lambda}^t \cdot \mathbf{Z} + \boldsymbol{\alpha}^t \cdot \boldsymbol{B}(U - T) + \epsilon_U, \\ Y_V = \boldsymbol{\lambda}^t \cdot \mathbf{Z} + \boldsymbol{\beta}^t \cdot \boldsymbol{B}(V - T) + \epsilon_V, \end{cases}$$

where $\mathbf{Z}$ is a vector of time-invariant covariates, and $\boldsymbol{B}(t) = \big(b_1(t), \cdots, b_q(t)\big)^T$ is a vector of continuous and piecewise smooth functions satisfying $b_k(0) = 0$ for $i = 1, 2, \cdots, q$. Such extensions may provide more enhanced modeling flexibility in some applications.

4.3 Simulation study

To evaluate the operating characteristics of the proposed method, we conducted two sets of simulation studies. The first one used data generated from a model in the form of Model (4.1), while mimicking the data structure of the pubertal growth application. Specifically, for each subject i, we first generated the anchoring point T_i from a Weibull distribution with shape and scale parameters, 80 and 12, respectively. We simulated a series of assessment times uniformly from the non-overlapping intervals $(2j, 2j + 2]$, $j = 0, 1, \cdots$. Based on T_i, we identified U_i and V_i as the adjacent points from the series of the simulated assessment times that bracket T_i, i.e. $U_i < T_i \leq V_i$. From the simulated values of U_i, T_i and V_i, we generated the outcome (Y_{U_i}, Y_{V_i}) from the piecewise linear model:

$$Y_{U_i} = \lambda + \alpha \cdot (U_i - T_i) + \epsilon_{U_i}, \quad Y_{V_i} = \lambda + \beta \cdot (V_i - T_i) + \epsilon_{V_i},$$

with $(\epsilon_{U_i}, \epsilon_{V_i})^t$ being simulated from the bivariate normal distribution $\boldsymbol{N}(\boldsymbol{\mu}, \boldsymbol{\Omega})$, where

$$\boldsymbol{\mu} = \begin{pmatrix} 0 \\ \\ 0 \end{pmatrix}, \quad \boldsymbol{\Omega} = \begin{pmatrix} 5 & 4 \\ \\ 4 & 5 \end{pmatrix}.$$

The true model parameters were chosen to be $\lambda = 50$, $\alpha = 5$ and $\beta = 8$. We considered four different sample sizes, $n = 100, 200, 400$ and 800. For a given sample size, we conducted a Monte-Carlo simulation with 1000 replicates.

For each simulated dataset, three models were fitted. First, the proposed model was fitted using the two-stage estimation method. Then for comparison purposes, we considered two alternative approaches. One was a midpoint imputation model, i.e., imputing T_i by the midpoint of the interval $(U_i, V_i]$, and then estimating the parameters using the ordinary least-squares method. This is a commonly used technique in analytical practice (Shankar et al., 2005). The other was a model assuming the true anchoring point distribution F_0 was known. For this latter model, we used F_0 instead of $\hat{F}_n$ in Stage 2 to obtain parameter estimates. This second model was not realistic for most applications. We use it here simply as a benchmark to investigate the efficiency loss due to the estimation of F_0. The estimated standard errors of these three models were obtained by using the aforementioned bootstrap method, based on $B = 50$ resamples.

For the 1000 replicates of samples of size n, we reported the percentage of average estimation bias (% Bias), Monte-Carlo standard deviations (M-C SD), average bootstrap standard errors (Av. SE), and the empirical coverage probabilities of

the 95% Wald-confidence intervals (95% CP) based on the asymptotical normality described in Theorem 4.2.2. Simulation results were summarized in Table 4.1.

Table 4.1: Simulation result for wider censoring intervals.

	$\lambda = 50$				$\alpha = 5$				$\beta = 8$			
n	100	200	400	800	100	200	400	800	100	200	400	800
Proposed model:												
% Bias	0.182	0.052	0.047	0.017	0.117	0.464	0.115	0.189	0.417	0.212	0.240	0.178
M-C SD	0.453	0.291	0.215	0.147	0.254	0.173	0.128	0.086	0.278	0.193	0.137	0.094
Av. SE	0.425	0.297	0.207	0.146	0.244	0.175	0.124	0.088	0.274	0.192	0.134	0.096
95% CP	0.916	0.944	0.937	0.945	0.939	0.938	0.942	0.948	0.942	0.949	0.939	0.951
Midpoint imputation:												
% Bias	1.195	1.135	1.223	1.160	11.02	10.51	11.03	10.59	7.237	7.005	7.217	7.035
M-C SD	1.041	0.748	0.530	0.360	1.023	0.710	0.508	0.350	1.139	0.774	0.556	0.388
Av. SE	1.031	0.738	0.521	0.371	0.987	0.710	0.504	0.359	1.086	0.782	0.556	0.396
95% CP	0.915	0.870	0.789	0.655	0.907	0.879	0.804	0.682	0.904	0.883	0.818	0.699
Use true F_0:												
% Bias	0.018	0.022	0.006	0.006	0.226	0.121	0.130	0.028	0.013	0.114	0.015	0.006
M-C SD	0.354	0.245	0.177	0.119	0.257	0.176	0.126	0.085	0.277	0.190	0.134	0.092
Av. SE	0.343	0.245	0.172	0.121	0.252	0.178	0.125	0.087	0.269	0.190	0.132	0.094
95% CP	0.947	0.948	0.932	0.944	0.948	0.947	0.944	0.953	0.940	0.949	0.933	0.957
Normal distribution approximation:												
% Bias	0.463	0.255	0.245	0.223	0.170	0.361	0.167	0.345	1.723	1.267	0.978	0.910
M-C SD	0.792	0.568	0.398	0.269	0.261	0.184	0.129	0.088	0.342	0.238	0.169	0.112
Av. SE	0.738	0.555	0.394	0.278	0.257	0.183	0.129	0.090	0.380	0.251	0.172	0.120
95% CP	0.912	0.926	0.941	0.942	0.942	0.944	0.945	0.957	0.964	0.951	0.931	0.930

A second set of simulation was conducted to assess the influence of interval width on the performance of the estimation. We used the same simulation scheme as in the first set of simulation, but generated a series of the assessment times uniformly from non-overlapping intervals $(j, j+1]$, $j = 0, 1, \ldots$. The resulted censoring intervals were narrower than that in the first set of simulation. The simulation results were summarized in Table 4.2.

Tables 4.1 and 4.2 show that the estimation bias is virtually ignorable in the proposed method, even with a moderate sample size ($n = 100$). The average bootstrap standard errors are all close to the corresponding Monte-Carlo standard deviations. In

Table 4.2: Simulation result for narrower censoring intervals.

	$\lambda = 50$				$\alpha = 5$				$\beta = 8$			
n	100	200	400	800	100	200	400	800	100	200	400	800
Proposed model:												
% Bias	0.123	0.050	0.053	0.028	0.118	0.646	0.013	0.094	1.446	0.880	0.786	0.536
M-C SD	0.367	0.241	0.185	0.123	0.456	0.331	0.238	0.161	0.499	0.346	0.247	0.166
Av. SE	0.346	0.248	0.175	0.124	0.434	0.316	0.228	0.162	0.482	0.343	0.243	0.174
95% CP	0.932	0.951	0.931	0.947	0.945	0.937	0.939	0.945	0.925	0.935	0.936	0.952
Midpoint imputation:												
% Bias	0.500	0.446	0.521	0.487	9.143	7.920	9.089	8.510	5.929	5.204	5.880	5.615
M-C SD	0.807	0.589	0.420	0.274	1.479	1.077	0.763	0.500	1.578	1.139	0.810	0.532
Av. SE	0.810	0.578	0.405	0.289	1.470	1.044	0.735	0.527	1.566	1.112	0.783	0.563
95% CP	0.936	0.925	0.895	0.870	0.932	0.922	0.896	0.875	0.924	0.925	0.899	0.872
Use true F_0:												
% Bias	0.006	0.023	0.005	0.001	0.348	0.342	0.201	0.035	0.073	0.177	0.032	0.016
M-C SD	0.344	0.232	0.174	0.114	0.479	0.338	0.241	0.160	0.515	0.343	0.248	0.166
Av. SE	0.331	0.237	0.166	0.117	0.470	0.331	0.234	0.164	0.497	0.350	0.244	0.173
95% CP	0.935	0.954	0.939	0.956	0.947	0.942	0.946	0.953	0.940	0.943	0.939	0.959
Normal distribution approximation:												
% Bias	0.416	0.243	0.258	0.235	0.565	1.349	0.827	1.056	4.296	3.228	2.678	2.569
M-C SD	0.648	0.437	0.315	0.208	0.462	0.328	0.229	0.153	0.676	0.457	0.328	0.212
Av. SE	0.609	0.446	0.316	0.219	0.479	0.331	0.229	0.160	0.816	0.484	0.319	0.221
95% CP	0.917	0.933	0.930	0.946	0.957	0.952	0.945	0.940	0.965	0.938	0.910	0.876

addition, the coverage probabilities of the 95% Wald-confidence intervals approach the nominal level of 0.95 as the sample size increases. Together, the simulation provides a strong numerical evidence in support of the asymptotic normality theory developed in Section 4.2. In comparison, the bias in the estimates using the midpoint imputation method is much larger and the bias is not reduced as sample size increases. Both the Monte-Carlo standard deviations and bootstrap standard errors of the midpoint imputation are markedly larger than those in the proposed method. The 95% Wald-confidence intervals from the midpoint imputation method have decreasing coverage probabilities when sample size increases.

As one would expect, parameter estimation performs best in the hypothetical situation of known F_0. But the bias in the proposed method are fairly small as

well. The bias also decreases with sample size. In situations of moderate to large sample sizes, the bias of the new method is practically ignorable, relatively to the magnitudes of the true parameter values. The Monte-Carlo standard deviation and average bootstrap standard errors of the parameters of the new method are slightly larger that those obtained in the hypothetical situation of known F_0. But importantly, the coverage probabilities of the new method and the case of known F_0 are quite comparable, especially for moderately large samples ($n \geq 200$).

To empirically evaluate the relative efficiency, we calculated the ratio of the Monte-Carlo standard deviations in the proposed model over those from the case of known F_0; see Table 4.3. All ratios are close to 1, especially for the local change rates α and β. Some ratios have values less than 1 due to random errors, since these ratios are simulated relative efficiency.

Narrower censoring intervals produce more accurately estimated $\hat{F}_n$. In those situations, the differences in standard errors between the proposed method and the known F_0 case are even smaller, suggesting that the new method does not lead to substantial loss of efficiency in parameter estimation.

Table 4.3: Empirical relative efficiency: proposed method vs knowing F_0.

	$\lambda = 50$				$\alpha = 5$				$\beta = 8$			
n	100	200	400	800	100	200	400	800	100	200	400	800
Wider censoring intervals												
M-C SD	1.280	1.191	1.213	1.233	0.986	0.986	1.008	1.013	1.005	1.016	1.019	1.023
Narrower censoring intervals												
M-C SD	1.068	1.038	1.060	1.082	0.952	0.978	0.989	1.007	0.969	1.007	0.992	1.000

Finally, we fitted a parametric model under the assumption that F_0 follows a normal distribution with unknown mean μ and variance σ^2. The parameters $(\lambda, \alpha, \beta; \mu, \sigma^2)$ were jointly estimated. Such a model is commonly used in practice

(Robinson et al., 2010). When the true F_0 was not normally distributed, our simulation results show that it produces biased estimates and consequently suboptimal coverage probability. Especially with narrow censoring intervals, the coverage probability of β was only 87.6% when sample size was 800.

In summary, the simulation study provides strong empirical evidence indicating good finite sample performance of the proposed method.

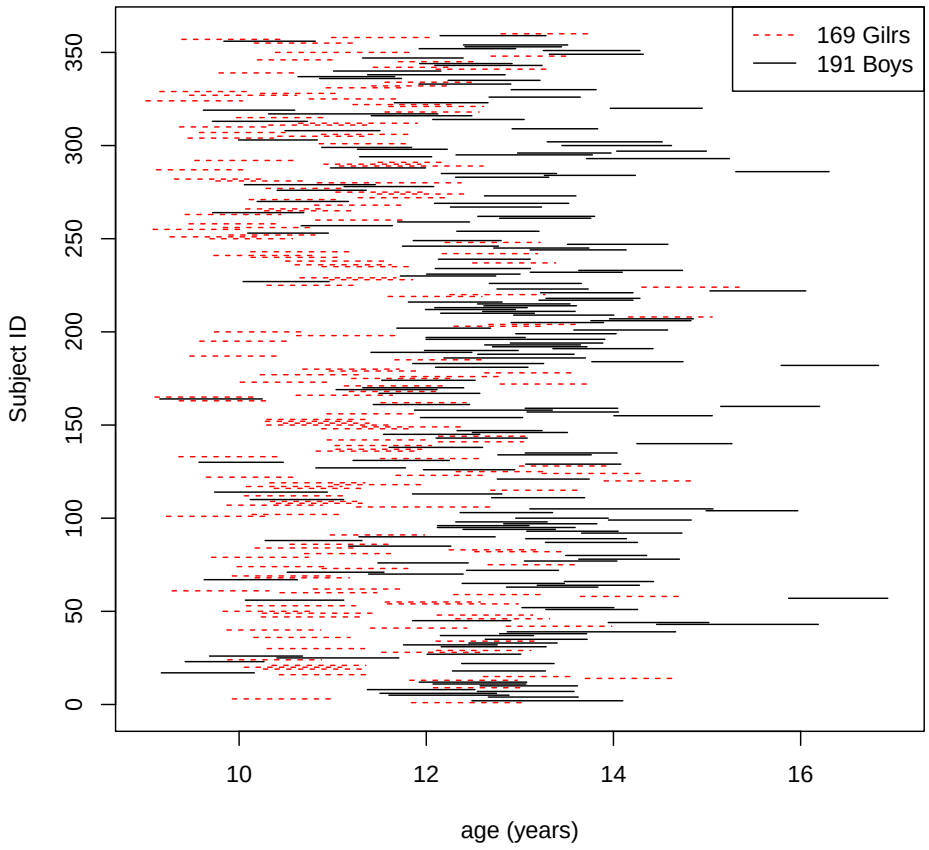
4.4 Analysis of pubertal skeletal growth data

To illustrate the application of the proposed method to real data, we analyze the pubertal growth data from 360 children. The original data came from an observational study of somatic growth and blood pressure development. The study protocol was described elsewhere (Tu et al., 2009, 2014). In the current analysis, we attempt to determine the rates of growth in height, upper body length (i.e., height in sitting position), shoulder length, elbow, wrist, and knee diameters, and to compare the rates between male and female participants, immediately before and after the subject-specific pubertal growth spurt (PGS).

Although the exact PGS time for an individual was not observable, the investigators were able to determine the assessment times that flanked the unobserved PGS (Shankar et al., 2005), which we referred to as the peak growth period. The current analysis included a total of 169 girls and 191 boys. The age range from the youngest and the oldest assessment times were between 9.005 and 16.930 years, thus ensuring the coverage of PGS in all participants. Figure 4.1 shows the peak growth intervals for the study children. Given the skeletal measurements at the endpoints of

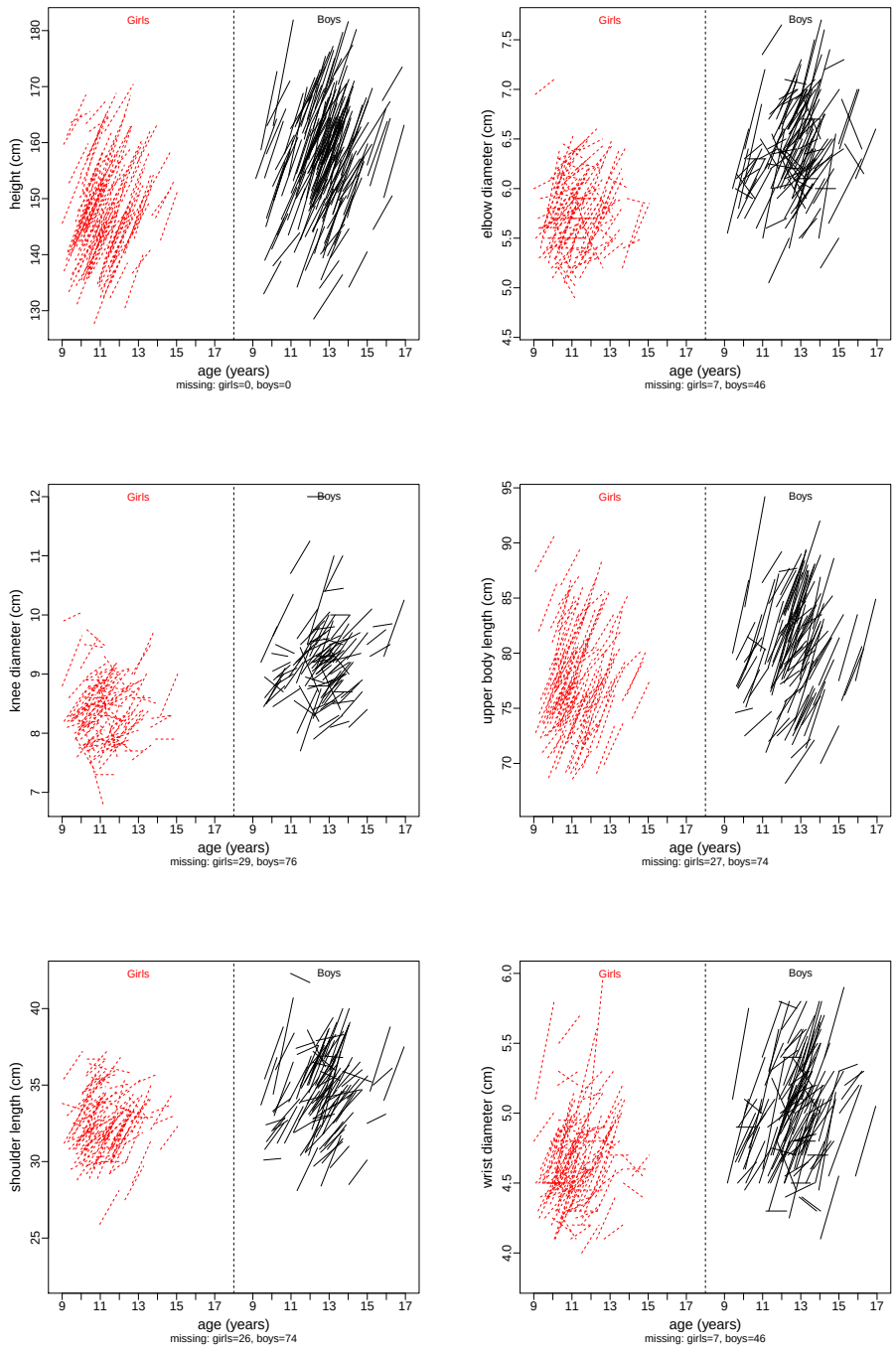
these intervals, we used the proposed method to estimate the change rates in these outcomes before and after the unobserved PGS.

Figure 4.1: Peak growth periods in 360 children



The skeletal measures of interest, including height, upper body length, shoulder length, elbow, wrist and knee diameters of the participants in the peak growth intervals are shown in Figure 4.2, stratified by sex. The figure clearly show that significant changes occur simultaneously in all skeletal dimensions during this peak growth period.

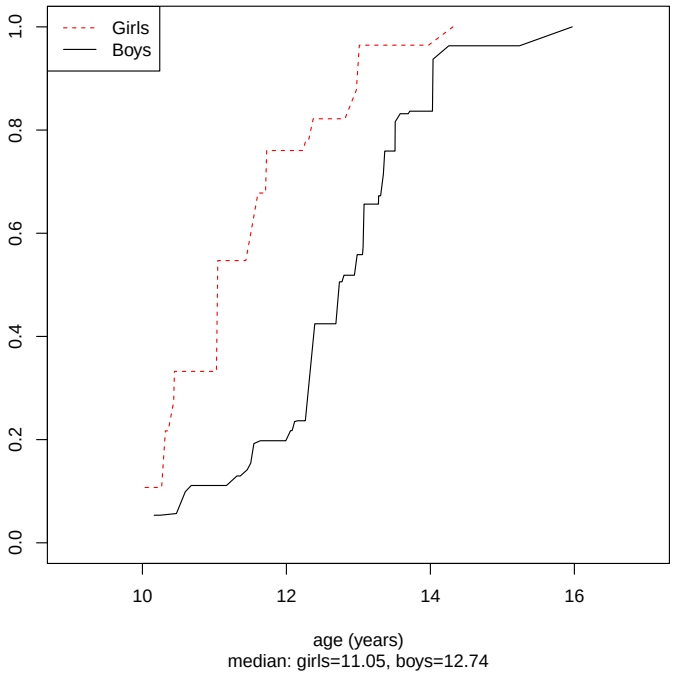
Figure 4.2: Observed data



As proposed, we used the NPMLE of the unknown CDF to depict the PGS distribution in male and female children, as shown in Figure 4.3. From the NPMLE

of the CDFs, we estimated median ages of PGS to be 11.05 years for girls, and 12.74 years for boys.

Figure 4.3: The estimated CDFs of F_0 for males and females

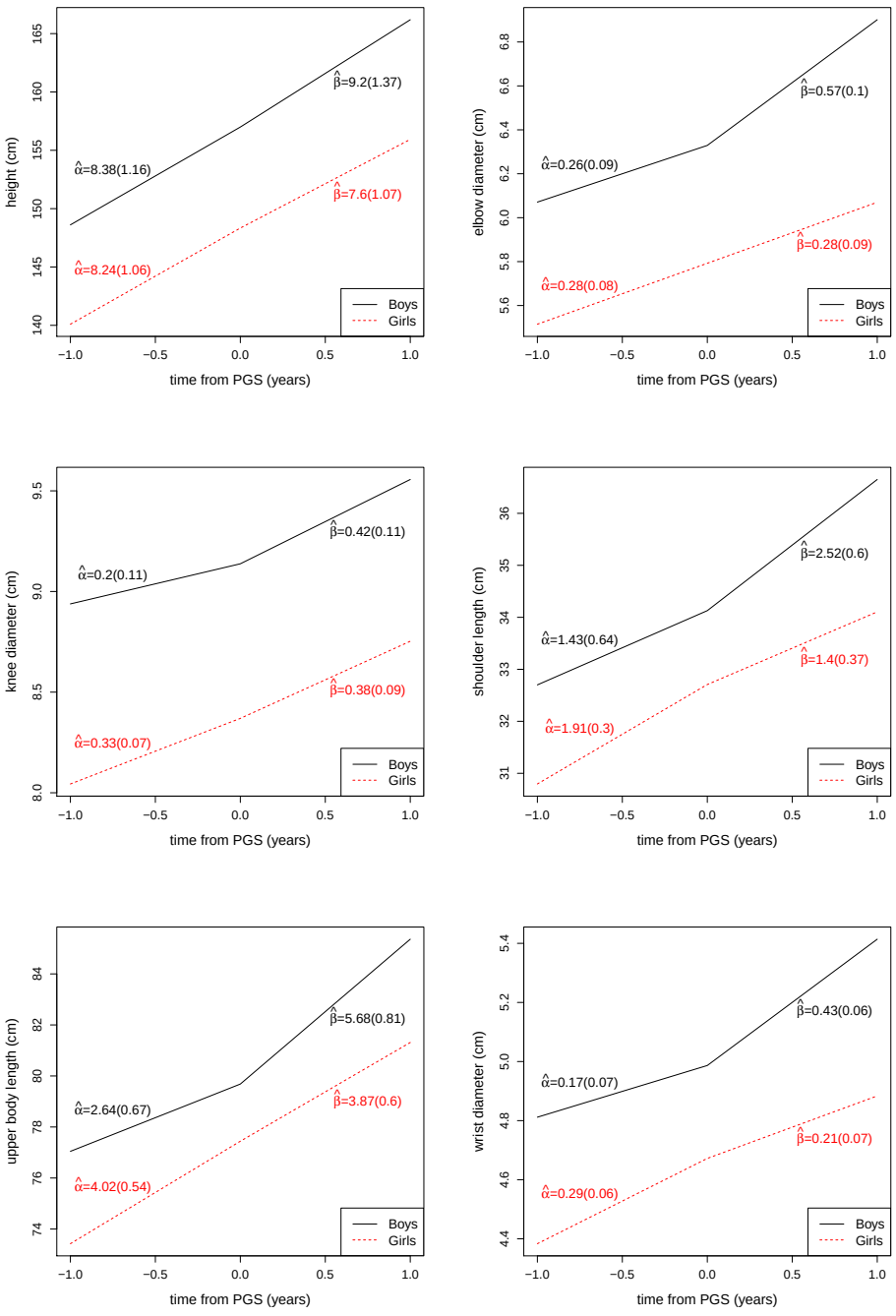


We then fit the following distribution-free model

$$\begin{cases} E(Y_U) = \lambda + \alpha \cdot (U - T), \\ E(Y_V) = \lambda + \beta \cdot (V - T), \end{cases}$$

separately for boys and girls, where Y_U and Y_V were the observed values of the skeletal variables, including height, upper body length, shoulder length, and elbow, wrist, and knee diameters, measured at the two end points of $(U, V]$, respectively. The functional estimates of the pre and post-PGS skeletal growth in the six measures, stratified by sex, are presented graphically in Figure 4.4.

Figure 4.4: The fitted anchoring point models.



Inference on the post-PGS growth rate changes from the pre-PGS period were made based on the asymptotic results of Theorem 4.2.2. Depending the specific need of testing, one could express tests in form of linear contrast $\mathbf{e}^t\boldsymbol{\theta}$, with null

hypothesis written as $H_0 : \quad \mathbf{e}^t\boldsymbol{\theta} = 0$. The two-sided test statistic, therefore, takes the form $n(\mathbf{e}^t\hat{\boldsymbol{\Sigma}}_n\mathbf{e})^{-1}\left(\mathbf{e}^t\hat{\boldsymbol{\theta}}_n\right)^2$, where $\hat{\boldsymbol{\theta}}_n$ is the parameter estimate and $\hat{\boldsymbol{\Sigma}}_n$ is the bootstrap estimate of the asymptotic variance matrix. The test statistic follows the χ^2-distribution with 1 degree of freedom asymptotically according to Theorem 4.2.2. Letting $\mathbf{e} = (0, -1, 1)^t$, it allows for comparison of the pre-PGS rate α against the post-PGS rate β of a given outcome.

Similarly, we can make inference on the difference in growth rates between boys and girls by testing hypothesis $H_0 : \ \mathbf{e}^t(\boldsymbol{\theta_1} - \boldsymbol{\theta_2}) = 0$. The corresponding test statistic is derived from the standard independent two-sample test given by

$$\left[n_1^{-1}(\mathbf{e}^t\hat{\boldsymbol{\Sigma}}_{1,n_1}\mathbf{e}) + n_2^{-1}(\mathbf{e}^t\hat{\boldsymbol{\Sigma}}_{2,n_2}\mathbf{e})\right]^{-1}\left(\mathbf{e}^t(\hat{\boldsymbol{\theta}}_{1,n_1} - \hat{\boldsymbol{\theta}}_{2,n_2})\right)^2$$

with $\mathbf{e} = (0, 0, 1)^t$, where $\hat{\boldsymbol{\theta}}_{1,n_1}$ and $\hat{\boldsymbol{\theta}}_{2,n_2}$ are the parameter estimates, and $\hat{\boldsymbol{\Sigma}}_{1,n_1}$ and $\hat{\boldsymbol{\Sigma}}_{2,n_2}$ are the bootstrap estimates of the asymptotic variance matrices for the respective groups. Again, the test statistic follows a χ^2-distribution with 1 degree of freedom asymptotically.

The analysis represents the first attempt in quantifying the skeletal growth rates in boys and girls around the time of PGS (See Figure 4.4). The analysis clearly showed that boys and girls experienced very different rates of skeletal growth around PGS. Three important observations emerged from the analysis: (1) Skeletal growth continues around PGS in both sexes, as shown by the strictly positive growth rates in all variables. (2) In comparison with girls, boys have greater skeletal measures around PGS. Interestingly, sex differences show not only in the length of the bones but also in the thickness of the bones, in both pre and post-PGS periods. For example,

in the post-PGS period, boy's elbow diameter increased at a rate of 0.57cm/year, significantly greater than girl's 0.28cm/year ($p = 0.03$). In the same period, boy's wrist diameter increases at a rate of 0.43cm/year, significantly greater than girl's 0.21cm/year ($p = 0.01$). (3) Boy's post-PGS growth rates are generally greater than their pre-PGS rates. For example, the growth rate of upper-body length in boys increases from 2.64cm/year in the pre-PGS period to 5.68cm/year in the post-PGS period, a net increase of 3.04cm/year ($p = 0.02$), comparing to a slight decrease in girls from 4.02cm/year pre-PGS to 3.87cm/year post-PGS ($p = 0.88$). The same was true for the bone thickness. For example, the wrist diameter growth rate in boys increases from 0.17cm/year pre-PGS to 0.43cm/year post-PGS ($p = 0.04$).

Viewed as a whole, the analysis provided a clearer picture of the emergence of sexual dimorphism in human skeletal development. Although girls start puberty and reach their peak height growth velocity nearly two years earlier than boys, at the PGS boys exceed girls in all skeletal measures including both bone lengths and bone thickness. Importantly, boy's greater post-PGS growth rates in different body parts set the stage for a stronger and more sustained growth that ultimately led their bigger average body size.

The findings, however, also raised intriguing questions about the regulation of such coordinated patterns of growth. One might speculate, for example, that sex differences around the PGS could be the result of a surging influence of androgenic hormones such as testosterone. In the absence of direct evidence, we simply note the concurrent emergence of accelerated bone growth and male sexual characteristics right after PGS appears to give credence to such a speculation. Of course, variations in timing as well as length of pubertal growth suggest the existence of multiple operators,

including hormonal (Rose et al., 1991), nutritional (Whiting et al., 2004), and genetic (Tu et al., 2015) influences on the rapid skeletal development in puberty.

Chapter 5

Mixed-effects model with interval censored anchoring points

The distribution-free model presented in the previous chapter completely ignores the correlations of the longitudinal observations. The estimation efficiency can be improved if the correlations can be appropriately modeled. Therefore, we consider a likelihood based approach to study parameter estimations involving interval censored anchoring points.

A second motivation comes from the need of extending the traditional mixed-effects models for longitudinal data anchored by interval censored events. Longitudinal data are collected on a predetermined time scale. To conduct analysis, such as using mixed-effects models (Laird and Ware, 1982), the time scale for defining the primary endpoint is needed to place the observations in a proper time context. When the time scale needs to be defined using unobserved anchoring points that are randomly distributed with unknown distributions, standard mixed-effects models are not readily applicable. In this chapter, we apply the two-stage estimation method to extend the standard mixed-effects models for longitudinal data to accommodate interval censored anchoring points.

5.1 Estimation with with interval-censored anchoring points

5.1.1 Parameter estimation

We formally define the notation in a generic longitudinal study setting. From each subject, we observe the response vector $\boldsymbol{Y}$, covariates vector $\boldsymbol{W}$, and a censoring interval $(L, R]$ that contains the unobserved anchoring point T, i.e., $L < T \leq R$. Given $\boldsymbol{W}, L, R$ and unobserved T, the conditional density function of $\boldsymbol{Y}$ can be modeled as

$$\boldsymbol{Y}\big|(\boldsymbol{W}, L, R, T) \sim \phi(\boldsymbol{Y}\big|\boldsymbol{W}, L, R, T; \boldsymbol{\theta}) \tag{5.1}$$

for a known density function ϕ that contains the finite dimensional parameter $\boldsymbol{\theta}$, whose true value $\boldsymbol{\theta}_0$ is of the interest. Here, the conditional density function ϕ can be any continuous or discrete distributions. For the main theoretical result to hold, we only require ϕ to satisfy a set of regularity conditions (see Section 5.1.2). For example, ϕ can be a member of the exponential family of distributions. When ϕ is the density function of a normal distribution, the model can be written in the familiar form of a linear mixed-effects model, which we shall examine as a special case with greater details in Section 5.2.

In traditional longitudinal models, the parameter $\boldsymbol{\theta}$ is well defined only when $\boldsymbol{W}, L, R$, and T are fully observed; when T is not observed, the true value $\boldsymbol{\theta}_0$ cannot be estimated from Model (5.1). An intuitive way to estimate $\boldsymbol{\theta}_0$ in the absence of T is to focus on the conditional density function of $\boldsymbol{Y}$ given $\boldsymbol{W}, L$ and R, while integrating out T:

$$\boldsymbol{Y}\big|(\boldsymbol{W}, L < T \leq R) \sim \int \phi(\boldsymbol{Y}\big|\boldsymbol{W}, L, R, t; \boldsymbol{\theta})\ dF_{T|(\boldsymbol{W}, L<T\leq R; \boldsymbol{\theta})}(t),$$

where $F_{T|(\boldsymbol{W},L<T\leq R;\boldsymbol{\theta})}$ is the conditional cumulative distribution function (CDF) of T given the observed covariates $\boldsymbol{W}, L < T \leq R$, and parameter $\boldsymbol{\theta}$. We assume that T is conditionally independent of $\boldsymbol{W}$, given L and R; $(L, R]$ is an independent censoring interval; and the marginal distribution of T is not informative to $\boldsymbol{\theta}$. These assumptions lead to $F_{T|(\boldsymbol{W},L<T\leq R;\boldsymbol{\theta})}(t) = 1(L < t \leq R)\big(F_0(t) - F_0(L)\big)\Big/\big(F_0(R) - F_0(L)\big)$, where F_0 denotes the true but often unknown CDF of the anchoring point T. So we have

$$\begin{array}{rcl} \boldsymbol{Y}\big|(\boldsymbol{W}, L < T \leq R) & \sim & \int_L^R \frac{\phi(\boldsymbol{Y}\big|\boldsymbol{W},L,R,t;\boldsymbol{\theta})dF_0(t)}{F_0(R)-F_0(L)} \\ & \propto & \int_L^R \phi(\boldsymbol{Y}\big|\boldsymbol{W}, L, R, t; \boldsymbol{\theta})dF_0(t). \end{array} \tag{5.2}$$

When the anchoring point distribution F_0 is known, the true parameter value $\boldsymbol{\theta}_0$ can be estimated directly from Model (5.2). But F_0 is often unknown in real studies. In such a situation, ideally, one would jointly estimate $\boldsymbol{\theta}_0$ and F_0 from Model (5.2). Given a random sample $\{(\boldsymbol{Y}_i, \boldsymbol{W}_i, L_i, R_i)|i = 1, \cdots, n\}$, i.e., without observing the anchoring point T_i, $i = 1, \cdots, n$, one could estimate $\boldsymbol{\theta}_0$ and F_0 by maximizing the marginal log-likelihood

$$\mathcal{L}_n(\boldsymbol{\theta}, F) = \sum_{i=1}^{n} \log\left(\int_{L_i}^{R_i} \phi(\boldsymbol{Y}_i\big|\boldsymbol{W}_i, L_i, R_i, t; \boldsymbol{\theta})dF(t)\right)$$

jointly over $\boldsymbol{\Theta} \times \mathcal{F}$, where $\boldsymbol{\Theta}$ is a parameter space for $\boldsymbol{\theta}$ and $\mathcal{F}$ the class of one-dimensional CDF's. Maximization of the above likelihood function presents a daunting computational challenge because the surface of $\mathcal{L}_n(\boldsymbol{\theta}, F)$ is often very complicated.

To solve the problem, we propose a pseudo-likelihood approach as described in the Introduction. First, we estimate F_0 by the nonparametric maximum likelihood

estimator (NPMLE) $\hat{F}_n$ using the interval censored data $\{(L_i, R_i]|i = 1, \cdots, n\}$, as described in Section 4.1. With the estimated CDF $\hat{F}_n$, we then estimate $\boldsymbol{\theta}_0$ by $\hat{\boldsymbol{\theta}}_n$, the maximizer of the pseudo-likelihood

$$\mathcal{L}_n(\boldsymbol{\theta}, \hat{F}_n) = \sum_{i=1}^{n} \log \left(\int_{L_i}^{R_i} \phi(\boldsymbol{Y}_i | \boldsymbol{W}_i, L_i, R_i, t; \boldsymbol{\theta}) d\hat{F}_n(t) \right). \tag{5.3}$$

The algorithmic efficiency for computing the parameter estimates $\hat{\boldsymbol{\theta}}_n$ from Model (5.3) depends on the specific model ϕ. In Section 5.2, we provide a hybrid algorithm in the case that ϕ represents a linear mixed-effects model.

5.1.2 Asymptotic property

The estimator that we proposed can be regraded as a stochastic functional of the NPMLE $\hat{F}_n$. Consider the following stochastic functional $\mathbb{Q}_n$, which maps a CDF F to

$$\begin{aligned} \mathbb{Q}_n(F) &= \arg\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \left(\mathcal{L}_n(\boldsymbol{\theta}, F) \right) \\ &= \arg\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \left\{ \sum_{i=1}^{n} \log \left(\int_{L_i}^{R_i} \phi(\boldsymbol{Y}_i | \boldsymbol{W}_i, L_i, R_i, t; \boldsymbol{\theta}) dF(t) \right) \right\}. \end{aligned}$$

The proposed estimate $\hat{\boldsymbol{\theta}}_n$ is the value of $\mathbb{Q}_n$ at the estimated distribution $\hat{F}_n$, i.e., $\hat{\boldsymbol{\theta}}_n = \mathbb{Q}_n(\hat{F}_n)$. If F_0 is known, the true parameter $\boldsymbol{\theta}_0$ can be estimated from Model (5.2). Let $\tilde{\boldsymbol{\theta}}_n$ be the estimate under F_0, i.e., $\tilde{\boldsymbol{\theta}}_n = \mathbb{Q}_n(F_0)$. It follows from standard maximum likelihood theory that $\tilde{\boldsymbol{\theta}}_n$ is a consistent and asymptotic normal estimator of $\boldsymbol{\theta}_0$.

The idea in the proposed method is that, $\hat{F}_n$ is a consistent estimate of F_0 and $\mathbb{Q}_n$ is a smooth functional, so $\hat{\boldsymbol{\theta}}_n = \mathbb{Q}_n(\hat{F}_n)$ is potentially asymptotically equivalent to $\tilde{\boldsymbol{\theta}}_n = \mathbb{Q}_n(F_0)$, and hence a possibly consistent estimate of $\boldsymbol{\theta}_0$. Although the

idea is simple, a rigorous study of the asymptotic property of $\hat{\boldsymbol{\theta}}_n$ is much evolved, because of the extra variability associated with the estimation of $\hat{F}_n$. Note that $\hat{F}_n$ has $n^{1/3}$-convergency rate (Groeneboom and Wellner, 1992), which complicates the study of asymptotic distribution of $\hat{\boldsymbol{\theta}}_n$. We adopt techniques from the empirical process theory to accomplish this goal. Throughout the rest of the manuscript, for a measurable function f on a measure space with measure P, let $\mathrm{P}(f)$ denote $\int f d\mathrm{P}$, the integral of f with respect to P. The following regularity conditions are sufficient to justify the forthcoming theorem on the asymptotic properties of $\hat{\boldsymbol{\theta}}_n$.

1. Regularity conditions on the interval-censoring data: $\boldsymbol{W}$, L, R and T.

   F1: There exist constants $\tau_1 < \tau_2 < \infty$ such that the support of the density function f_T of the anchoring point T is contained in $[\tau_1, \tau_2]$.

   F2: The event time T is conditionally independent of $\boldsymbol{W}$, given L and R. The censoring interval $(L, R]$ is independent of T.

   F3: Support of F_0 is included in the union of the supports of the CDF of L and the CDF of R. And F_0 does not depend on $\boldsymbol{\theta}$.

   F4: There exists a constant c such that $P\big(F_0(R) - F_0(L) > c\big) = 1$.

   F5: The sum of density functions of L and R, $f_L + f_R$, is strictly positive in $[\tau_1, \tau_2]$.

   F6: The joint density function of (L, T, R), is twice differentiable in $[\tau_1, \tau_2]$. In particular, f_L and f_R are differentiable and uniformly bounded in $[\tau_1, \tau_2]$.

   F7: The density function of T, f_T, is twice differentiable.

2. Regularity conditions on the longitudinal data model when F_0 is known.

   Let $\nabla^k_{\boldsymbol{\theta}}$ denote the differential operator of taking all k-th order partial derivatives with respect to the vector variable $\boldsymbol{\theta}$. Let $d = \dim(\boldsymbol{\theta})$ denote the dimension of $\boldsymbol{\theta}$. The model parameter space $\boldsymbol{\Theta}$ is a subset of $\mathbb{R}^d$ such that:

M1: $\int_L^R \phi(\boldsymbol{Y}\big|\boldsymbol{W}, L, R, t; \boldsymbol{\theta}_1) dF_0(t) \neq \int_L^R \phi(\boldsymbol{Y}\big|\boldsymbol{W}, L, R, t; \boldsymbol{\theta}_2) dF_0(t)$ for any different parameters $\boldsymbol{\theta}_1 \neq \boldsymbol{\theta}_2$ in $\boldsymbol{\Theta}$.

M2: The true parameter $\boldsymbol{\theta}_0$ is an inner point of $\boldsymbol{\Theta}$.

M3: Support of $\int_L^R \phi(\boldsymbol{Y}\big|\boldsymbol{W}, L, R, t; \boldsymbol{\theta}) dF_0(t)$ does not dependent on $\boldsymbol{\theta} \in \boldsymbol{\Theta}$.

M4: The conditional density function $\phi = \phi(\boldsymbol{Y}\big|\boldsymbol{W}, L, R, T; \boldsymbol{\theta})$ is continuous. The third order partial derivative $\nabla^3_{\boldsymbol{\theta}}(\phi)$ exists and is continuous. Both ϕ and its partial derivative function $\boldsymbol{u} = \nabla_{\boldsymbol{\theta}}(\phi)$ have continuous partial derivatives with respect to T.

M5: Let P be the probability measure associated with $(\boldsymbol{Y}, \boldsymbol{W}, L, R)$, then

$$\nabla_{\boldsymbol{\theta}} \left[\mathrm{P} \left[\log \left(\int \phi dF_0 \right) \right] \right] = \mathrm{P} \left[\nabla_{\boldsymbol{\theta}} \left[\log \left(\int \phi dF_0 \right) \right] \right]$$

$$\nabla^2_{\boldsymbol{\theta}} \left[\mathrm{P} \left[\log \left(\int \phi dF_0 \right) \right] \right] = \mathrm{P} \left[\nabla^2_{\boldsymbol{\theta}} \left[\log \left(\int \phi dF_0 \right) \right] \right]$$

3. The random vector $(\boldsymbol{Y}, \boldsymbol{W}, L, R)$ is bounded with probability 1.

Remark 5.1.1. *The regularity conditions are mild and pose no extra restrictions in most applications. The first set of conditions are usually assumed, in order to have good estimate of smooth functionals on the CDF of interval-censored event time (Geskus and Groeneboom, 1999). The second set of conditions are the usual regularity conditions assumed in the maximum likelihood theory. The third condition is often satisfied in practice. It means that, as long as the data do not contain substantial amount of extreme observations, the parameter estimate is asymptotically normally distributed as described in the following theorem.*

Theorem 5.1.2. *Under the stated regularity conditions, the model estimate $\hat{\boldsymbol{\theta}}_n$ of Model (5.3) is consistent and asymptotically normally distributed. More precisely,*

let P *be the probability measure associated with* $(\boldsymbol{Y}, \boldsymbol{W}, L, R)$*,* $\boldsymbol{\theta}_0$ *the true parameter,* $\boldsymbol{u} = \nabla_{\boldsymbol{\theta}}\phi$ *the gradient of* ϕ *with respect to* $\boldsymbol{\theta}$*,* $\boldsymbol{U}_{\boldsymbol{\theta}_0,F_0} = \left[\left(\int_L^R \phi dF_0\right)^{-1} \int_L^R \boldsymbol{u} dF_0\right]_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}$ *the score function and* $\boldsymbol{A} = -\mathrm{P}\big(\nabla_{\boldsymbol{\theta}}(\boldsymbol{U}_{\boldsymbol{\theta},F_0})\big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\big)$ *the information matrix when* F_0 *is known, and* $\boldsymbol{\Phi} = \boldsymbol{\Phi}(L, R)$ *the multidimensional function that has a zero mean and uniquely solves the following integral equation system*

$$\int_{L<t\leq R} \boldsymbol{\Phi}(L,R) d\mathrm{P} = \int_{S_t} \left[\left(\int_L^R \phi dF_0\right)^{-2} \left(\boldsymbol{u}\int_L^R \phi dF_0 - \phi\int_L^R \boldsymbol{u} dF_0\right)\right]_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} d\mathrm{P}_t,$$

where S_t *denotes the domain of* $(\boldsymbol{Y}, \boldsymbol{W}, L, R)$ *given value* $T = t$*, and* P_t *is the conditional measure of* P *when restricted to* S_t*. Then*

$$\sqrt{n}\big(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\big) = \boldsymbol{A}^{-1} \cdot \sqrt{n}\mathbb{P}_n\big(\boldsymbol{U}_{\boldsymbol{\theta}_0,F_0} + \boldsymbol{\Phi}\big) + \boldsymbol{o}_p(1),$$

where $\mathbb{P}_n$ *denotes the empirical probability measure associated with a random sample of* $(\boldsymbol{Y}, \boldsymbol{W}, L, R)$ *of size* n*. In particular,* $\sqrt{n}\big(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\big) \xrightarrow{P} \boldsymbol{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$*, with the asymptotic variance matrix* $\boldsymbol{\Sigma}$ *given by*

$$\boldsymbol{\Sigma} \;=\; \boldsymbol{A}^{-1} + \boldsymbol{A}^{-1}\mathrm{P}\left(\boldsymbol{U}_{\boldsymbol{\theta}_0,F_0}\boldsymbol{\Phi}^t + \boldsymbol{\Phi}\boldsymbol{U}^t_{\boldsymbol{\theta}_0,F_0} + \boldsymbol{\Phi}^{\otimes 2}\right)\boldsymbol{A}^{-1},$$

where $\boldsymbol{M}^{\otimes 2}$ *denotes* $\boldsymbol{M}\boldsymbol{M}^t$ *for any matrix* $\boldsymbol{M}$*.*

Proof. Let K denote a constant, whose value differs from place to place. The proof is done by applying Theorem 3.4.1. Consider the following multivariate random

functional,

$$\boldsymbol{U}(\boldsymbol{Y},\boldsymbol{W},L,R,F,\boldsymbol{\theta})=\frac{\int_L^R \boldsymbol{u}(\boldsymbol{Y},\boldsymbol{W},L,R,t;\boldsymbol{\theta})\,dF(t)}{\int_L^R \Phi(\boldsymbol{Y},\boldsymbol{W},L,R,t;\boldsymbol{\theta})\,dF(t)}=\frac{\int_L^R \boldsymbol{u}\,dF}{\int_L^R \Phi\,dF}$$

which is the score function of the parameter $\boldsymbol{\theta}$ based on the marginal likelihood of $(\boldsymbol{Y},\boldsymbol{W},L,R)$, obtained by integration on T using CDF F. The true parameter $\boldsymbol{\theta}_0$ of interest is the solution of

$$\mathrm{P}\big(\boldsymbol{U}(\boldsymbol{Y},\boldsymbol{W},L,R,F_0,\boldsymbol{\theta})\big)=\int\frac{\int_L^R \boldsymbol{u}\,dF_0}{\int_L^R \Phi\,dF_0}\,d\mathrm{P}=\mathbf{0}.$$

From the regularity conditions (M2) and (M4), there exists a closed ball $\boldsymbol{\Theta}$ with radius $K>0$ and with center $\boldsymbol{\theta}_0$, over which the function $\mathrm{P}\big(\boldsymbol{U}(\boldsymbol{Y},\boldsymbol{W},L,R,F_0,\boldsymbol{\theta})\big)$ is C^3 and strictly convex. Let $\mathcal{F}_\delta$ denote the space of all CDF supported in $[\tau_1,\tau_2]$, whose $\|\cdot\|_\infty$-distances from F_0 are less than a small number δ. Let $\boldsymbol{U}_{\boldsymbol{\theta},F}$ denote the following function on $(\boldsymbol{Y},\boldsymbol{W},L,R)$

$$\boldsymbol{U}_{\boldsymbol{\theta},F}(\boldsymbol{Y},\boldsymbol{W},L,R)=\boldsymbol{U}(\boldsymbol{Y},\boldsymbol{W},L,R,F,\boldsymbol{\theta}).$$

Define an empirical process $\boldsymbol{\Psi}_n$ by setting $\boldsymbol{\Psi}_n(\boldsymbol{\theta},F)=\mathbb{P}_n\boldsymbol{U}_{\boldsymbol{\theta},F}$, and the corresponding functional $\boldsymbol{\Psi}$ by setting $\boldsymbol{\Psi}(\boldsymbol{\theta},F)=\mathrm{P}\boldsymbol{U}_{\boldsymbol{\theta},F}$, where $(\boldsymbol{\theta},F)\in\boldsymbol{\Theta}\times\mathcal{F}_\delta$.

To prove Theorem 5.1.2, we need two preparations.

The first preparation is to study the properties of several classes of functions. By Lemma 3.3.4, the following two classes of functions are both P-Donsker.

$$\mathcal{NUM}_{\boldsymbol{\Theta},\mathcal{F}_\delta}=\left\{\int_L^R \boldsymbol{u}\,dF\;\middle|\;F\in\mathcal{F}_\delta,\boldsymbol{\theta}\in\boldsymbol{\Theta}\right\},$$

$$\mathcal{DEN}_{\boldsymbol{\Theta},\mathcal{F}_\delta} = \left\{ \int_L^R \Phi \, dF \;\Big|\; F \in \mathcal{F}_\delta, \boldsymbol{\theta} \in \boldsymbol{\Theta} \right\}.$$

The regularity condition 3 implies that both $\mathcal{NUM}_{\boldsymbol{\Theta},\mathcal{F}_\delta}$ and $\mathcal{DEN}_{\boldsymbol{\Theta},\mathcal{F}_\delta}$ are uniformly bounded. By Condition F4, $\mathcal{DEN}_{\boldsymbol{\Theta},\mathcal{F}_\delta}$ is uniformly bounded away from zero when δ is small enough. By van der Vaart and Wellner (1996, Theorem 2.10.6), both the point-wise quotient class

$$\mathcal{U}_{\boldsymbol{\Theta},\mathcal{F}_\delta} = \left\{ \boldsymbol{U}_{\boldsymbol{\theta},F} = \frac{\int_L^R \boldsymbol{u} \, dF}{\int_L^R \Phi \, dF} \Big| \boldsymbol{\theta} \in \boldsymbol{\Theta}, F \in \mathcal{F}_\delta \right\}$$

and the smooth transformation class

$$\mathcal{M}_{\boldsymbol{\Theta},\mathcal{F}_\delta} = \left\{ \boldsymbol{M}_{\boldsymbol{\theta},F} = \log \left(\int_L^R \Phi \, dF \right) \Big| \boldsymbol{\theta} \in \boldsymbol{\Theta}, F \in \mathcal{F}_\delta \right\}$$

are also P-Donsker classes. In particular $\mathcal{M}_{\boldsymbol{\Theta},\mathcal{F}_\delta}$ is a Glivenko–Cantelli class. By van der Vaart and Wellner (1996, Example 2.10.7), the difference class

$$\mathcal{DU}_{\boldsymbol{\Theta},\mathcal{F}_\delta} = \left\{ \boldsymbol{U}_{\boldsymbol{\theta}_1,F_1} - \boldsymbol{U}_{\boldsymbol{\theta}_2,F_2} \Big| \boldsymbol{U}_{\boldsymbol{\theta}_1,F_1}, \boldsymbol{U}_{\boldsymbol{\theta}_2,F_2} \in \mathcal{U}_{\boldsymbol{\Theta},\mathcal{F}_\delta} \right\}$$

is also a P-Donsker class.

Next, we need to study the properties of the functional $\boldsymbol{\Psi}$. For any $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, a direct computation shows $\boldsymbol{\Psi}(\boldsymbol{\theta}, F) - \boldsymbol{\Psi}(\boldsymbol{\theta}, F_0) = (I) - (II) + (III)$, where

$$(I) = \int \left(\int_L^R \phi dF_0 \right)^{-2} \cdot \left(\int_L^R \boldsymbol{u} dF \cdot \int_L^R \phi dF_0 - \int_L^R \boldsymbol{u} dF_0 \cdot \int_L^R \phi dF \right) d\mathrm{P}$$

$$(II) = \int \left(\int_L^R \phi dF_0 \right)^{-2} \cdot \int_L^R \boldsymbol{u} d(F - F_0) \cdot \int_L^R \phi d(F - F_0) \; d\mathrm{P}$$

$$(III) = \int \left(\int_L^R \phi dF \right)^{-1} \cdot \left(\int_L^R \phi dF_0 \right)^{-2} \cdot \left(\int_L^R \boldsymbol{u} dF \cdot \int_L^R \phi d(F - F_0) \right)^2 d\mathrm{P}$$

The first term (I) can be calculated as

$$\begin{aligned}
(I) &= \int \left(\int_L^R \phi dF_0 \right)^{-2} \left(\int_L^R \boldsymbol{u} d(F - F_0) \cdot \int_L^R \phi dF_0 - \int_L^R \boldsymbol{u} dF_0 \cdot \int_L^R \phi d(F - F_0) \right) d\mathrm{P} \\
&= \int \left[\int_L^R \left(\int_L^R \phi dF_0 \right)^{-2} \left(\boldsymbol{u} \int_L^R \phi dF_0 - \phi \int_L^R \boldsymbol{u} dF_0 \right) d(F - F_0) \right] d\mathrm{P} \\
&= \int \left[\int_{S_t} \left(\int_L^R \phi dF_0 \right)^{-2} \left(\boldsymbol{u} \int_L^R \phi dF_0 - \phi \int_L^R \boldsymbol{u} dF_0 \right) d\mathrm{P}_t \right] d(F - F_0),
\end{aligned}$$

where the last equality follows from changing order of integrals, and S_t denotes the domain of $(\boldsymbol{Y}, \boldsymbol{W}, L, R)$ given value t, and P_t denotes the induced conditional probability measure on S_t. By the regularity condition (M4) on the smoothness of ϕ and regularity condition (F6) on H, a straightforward algebra yields that

$$\kappa(\boldsymbol{\theta}, t) = \int_{S_t} \left(\int_L^R \phi dF_0 \right)^{-2} \left(\boldsymbol{u} \int_L^R \phi dF_0 - \phi \int_L^R \boldsymbol{u} dF_0 \right) d\mathrm{P}_t \tag{5.4}$$

is a C^1 function on $\boldsymbol{\Theta} \times [\tau_1, \tau_2]$.

Using the regularity condition 3, the terms (II) and (III) can be controlled by

$$\begin{aligned}
\big| -(II) + (III) \big| &\leq \left(\tfrac{1}{K} \max(\|\boldsymbol{u}\|) \max(\Phi) + \tfrac{1}{K} \max(\|\boldsymbol{u}\|) \max(\Phi)^2 \right) \big\| F_n - F_0 \big\|_\infty^2 \\
&= K \big\| F_n - F_0 \big\|_\infty^2.
\end{aligned}$$

Since the constant K does not depend on $\boldsymbol{\theta}$, we proved that

$$\boldsymbol{\Psi}(\boldsymbol{\theta}, F_n) - \boldsymbol{\Psi}(\boldsymbol{\theta}, F_0) = \int \kappa(\boldsymbol{\theta}, t) d(F_n - F_0) + \boldsymbol{O}_p(\|F_n - F_0\|_\infty^2), \tag{5.5}$$

uniformly on $\boldsymbol{\Theta}$, as required in Condition L4 of Lemma 3.4.2.

We are now ready to verify the four conditions in Theorem 3.4.1 as follows.

1. Condition T1.

   This is included in the regularity condition 2.

2. Condition T2: $\hat{\boldsymbol{\theta}}_n \xrightarrow{P} \boldsymbol{\theta}_0$.

   By definition, $\hat{\boldsymbol{\theta}}_n$ maximizes $\mathbb{P}_n \boldsymbol{M}_{\boldsymbol{\theta}, \hat{F}_n}$, where $\boldsymbol{M}_{\boldsymbol{\theta}, \hat{F}_n} = \log\left(\int_L^R \Phi d\hat{F}_n\right)$. Since $\mathcal{M}_{\boldsymbol{\Theta}, \mathcal{F}_\delta}$ is a P-Glivenko–Cantelli class containing $\boldsymbol{M}_{\boldsymbol{\theta}, \hat{F}_n}$, it follows that

   $$\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} |\big(\mathbb{P}_n - \mathrm{P}\big) \boldsymbol{M}_{\boldsymbol{\theta}, \hat{F}_n}| \xrightarrow{P} 0.$$

   Since $n^{1/3} \|\hat{F}_n - F\|_\infty \xrightarrow{P} 0$ by Groeneboom and Wellner (1992), it can be easily shown by the Dominated Convergence Theorem (DCT) that

   $$\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} |\mathrm{P} \boldsymbol{M}_{\boldsymbol{\theta}, \hat{F}_n} - \mathrm{P} \boldsymbol{M}_{\boldsymbol{\theta}, F_0}| \xrightarrow{P} 0.$$

   Hence $\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} |\mathbb{P}_n \boldsymbol{M}_{\boldsymbol{\theta}, \hat{F}_n} - \mathrm{P} \boldsymbol{M}_{\boldsymbol{\theta}, F_0}| \xrightarrow{P} 0$ as well. It follows from the regularity condition 2 that $\mathrm{P} \boldsymbol{M}_{\boldsymbol{\theta}, F_0}$ is strictly convex over $\boldsymbol{\Theta}$ with local maximum at $\boldsymbol{\theta}_0$, which implies

   $$\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}, \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| > \epsilon} \mathrm{P} \boldsymbol{M}_{\boldsymbol{\theta}, F_0} < \mathrm{P} \boldsymbol{M}_{\boldsymbol{\theta}_0, F_0}.$$

Therefore by Theorem 5.7 of van der Vaart (1998), there exists a $\hat{\boldsymbol{\theta}}_n \in \boldsymbol{\Theta}$ that maximizes $\mathbb{P}_n \boldsymbol{M}_{\boldsymbol{\theta},\hat{F}_n}$ and $\hat{\boldsymbol{\theta}}_n \xrightarrow{P} \boldsymbol{\theta}_0$. In particular, when n is large, $\hat{\boldsymbol{\theta}}_n$ is a maximizer inside $\boldsymbol{\Theta}$ and hence a solution of $\boldsymbol{\Psi}_n(\boldsymbol{\theta}, \hat{F_n}) = \mathbf{0}$, which is the proposed estimate of the model parameter $\boldsymbol{\theta}$.

3. Condition T3: $\sqrt{n}\boldsymbol{\Psi}_n(\boldsymbol{\theta}_0, \hat{F}_n) \xrightarrow{P} \boldsymbol{Z}$ for a zero mean normal distribution $\boldsymbol{Z}$.

   Since $\hat{F}_n$ is the NPMLE with an interval censored data satisfying the regularity condition 1, the Hellinger differentiability (Geskus and Groeneboom, 1999, Pages 631-632) of $\boldsymbol{\Psi}(\boldsymbol{\theta}_0, F)$ with respect to F at F_0, as shown in Equation (5.5), implies that there exists a unique zero mean random variable $\boldsymbol{\Phi}(L, R)$ such that

$$\sqrt{n}\boldsymbol{\Psi}(\boldsymbol{\theta}_0, \hat{F}_n) = \sqrt{n}\boldsymbol{\Psi}(\boldsymbol{\theta}_0, \hat{F}_n) - \sqrt{n}\boldsymbol{\Psi}(\boldsymbol{\theta}_0, F_0) = \sqrt{n}\mathbb{P}_n\big(\boldsymbol{\Phi}(L, R))\big) + \boldsymbol{o}_p(1)$$

   by Corollary 2.1 of Geskus and Groeneboom (1999) and Theorem 3.1 of van der Vaart (1991). The function $\boldsymbol{\Phi}(L, R)$ is characterized as the solution to the integral equation

$$\int_{L<t\leq R} \boldsymbol{\Phi}(L, R) d\mathrm{P} = \kappa(\boldsymbol{\theta}_0, t),$$

   where $\kappa(\boldsymbol{\theta}_0, t)$ is given in Equation (5.4).

   On the other hand, we have

$$\begin{aligned}
\sqrt{n}\boldsymbol{\Psi}_n(\boldsymbol{\theta}_0, \hat{F}_n) - \sqrt{n}\boldsymbol{\Psi}_n(\boldsymbol{\theta}_0, F_0) &= \sqrt{n}\mathbb{P}_n \left(\boldsymbol{U}_{\boldsymbol{\theta}_0,\hat{F}_n} - \boldsymbol{U}_{\boldsymbol{\theta}_0,F_0}\right) \\
&= \sqrt{n}(\mathbb{P}_n - \mathrm{P}) \left(\boldsymbol{U}_{\boldsymbol{\theta}_0,\hat{F}_n} - \boldsymbol{U}_{\boldsymbol{\theta}_0,F_0}\right) + \sqrt{n}\mathrm{P} \left(\boldsymbol{U}_{\boldsymbol{\theta}_0,\hat{F}_n} - \boldsymbol{U}_{\boldsymbol{\theta}_0,F_0}\right) \\
&= \boldsymbol{o}_p(1) + \sqrt{n}\boldsymbol{\Psi}(\boldsymbol{\theta}_0, \hat{F}_n).
\end{aligned}$$

where the first term is $o_p(1)$ by Corollary 2.3.12 of van der Vaart and Wellner (1996), because $\boldsymbol{U}_{\boldsymbol{\theta}_0,\hat{F}_n} - \boldsymbol{U}_{\boldsymbol{\theta}_0,F_0}$ is in the P-Donsker class $\mathcal{DU}_{\boldsymbol{\Theta},\mathcal{F}_\delta}$ and $n^{1/3}\|\hat{F}_n - F\|_\infty \xrightarrow{P} 0$ implies $\mathrm{P}\left(\boldsymbol{U}_{\boldsymbol{\theta}_0,\hat{F}_n} - \boldsymbol{U}_{\boldsymbol{\theta}_0,F_0}\right)^2 \xrightarrow{P} \mathbf{0}$ by DCT. Thus, we have shown that

$$\begin{aligned}\sqrt{n}\boldsymbol{\Psi}_n(\boldsymbol{\theta}_0,\hat{F}_n) &= \sqrt{n}\boldsymbol{\Psi}_n(\boldsymbol{\theta}_0,F_0) + \sqrt{n}\boldsymbol{\Psi}(\boldsymbol{\theta}_0,\hat{F}_n) + \boldsymbol{o}_p(1)\\ &= \sqrt{n}\mathbb{P}_n\big(\boldsymbol{U}_{\boldsymbol{\theta}_0,F_0}\big) + \sqrt{n}\mathbb{P}_n\big(\boldsymbol{\Phi}\big) + \boldsymbol{o}_p(1).\end{aligned}$$

In particular, we have $\sqrt{n}\boldsymbol{\Psi}_n(\boldsymbol{\theta}_0,\hat{F}_n)\xrightarrow{P}\boldsymbol{Z}$, where $\boldsymbol{Z}$ is the limiting distribution of $\sqrt{n}\mathbb{P}_n\big(\boldsymbol{U}_{\boldsymbol{\theta}_0,F_0} + \boldsymbol{\Phi}\big)$, which is normally distributed with a zero mean.

4. Condition T4.

   We verify the four sufficient conditions in Lemma 3.4.2. The first two conditions follow directly from the regularity conditions, and the $n^{1/3}$-convergence rate of $\hat{F}_n$ (Groeneboom and Wellner, 1992). The last condition follows from Equation (5.5). It remains to check the third condition.

   For any $F \in \mathcal{F}$, we have

$$\begin{aligned}\sqrt{n}\Big(\boldsymbol{\Psi}_n(\hat{\boldsymbol{\theta}}_n,F) &- \boldsymbol{\Psi}(\hat{\boldsymbol{\theta}}_n,F)\Big) - \sqrt{n}\Big(\boldsymbol{\Psi}_n(\boldsymbol{\theta}_0,F) - \boldsymbol{\Psi}(\boldsymbol{\theta}_0,F)\Big)\\ &= \sqrt{n}(\mathbb{P}_n - \mathrm{P})\left(\boldsymbol{U}_{\hat{\boldsymbol{\theta}}_n,F} - \boldsymbol{U}_{\boldsymbol{\theta}_0,F}\right).\end{aligned}$$

   Notice that $\boldsymbol{U}_{\hat{\boldsymbol{\theta}}_n,F} - \boldsymbol{U}_{\boldsymbol{\theta}_0,F}$ is a member in the P-Donsker class $\mathcal{DU}_{\boldsymbol{\Theta},\mathcal{F}_\delta}$. Using the regularity condition 3, the consistency of $\hat{\boldsymbol{\theta}}_n$, and the fact that the measure dF has support in the compact set $[\tau_1,\tau_2]$, it follows by DCT that $\mathrm{P}\left(\boldsymbol{U}_{\hat{\boldsymbol{\theta}}_n,F} - \boldsymbol{U}_{\boldsymbol{\theta}_0,F}\right)^2 \xrightarrow{P} \mathbf{0}$. Hence $\sqrt{n}(\mathbb{P}_n - \mathrm{P})\left(\boldsymbol{U}_{\hat{\boldsymbol{\theta}}_n,F} - \boldsymbol{U}_{\boldsymbol{\theta}_0,F}\right) = \boldsymbol{o}_p(1)$ by

Corollary 2.3.12 of van der Vaart and Wellner (1996). This verifies the third condition in Lemma 3.4.2.

So Condition T4 is satisfied by Lemma 3.4.2.

Finally, we complete the proof of Theorem 5.1.2 by applying Theorem 3.4.1 to obtain

$$\sqrt{n}\big(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\big) = \boldsymbol{A}^{-1} \cdot \sqrt{n}\mathbb{P}_n\big(\boldsymbol{U}_{\boldsymbol{\theta}_0,F_0} + \boldsymbol{\Phi}\big) + \boldsymbol{o}_p(1),$$

where $-\boldsymbol{A}$ is the Jacobian of $\mathrm{P}\left(\boldsymbol{U}_{\boldsymbol{\theta},F_0}\right)$ at $\boldsymbol{\theta} = \boldsymbol{\theta}_0$. So we have $\sqrt{n}\big(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\big) \xrightarrow{P} \boldsymbol{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma} = \boldsymbol{A}^{-1}\mathrm{P}\left((\boldsymbol{U}_{\boldsymbol{\theta}_0,F_0} + \boldsymbol{\Phi})^{\otimes 2}\right)\boldsymbol{A}^{-1}$. Since $\boldsymbol{U}_{\boldsymbol{\theta},F_0}$ is the score function of $\boldsymbol{\theta}$, it follows from the classical MLE theory that

$$\boldsymbol{A} = -\mathrm{P}\left(\nabla \boldsymbol{U}_{\boldsymbol{\theta}_0,F_0}\right) = \mathrm{P}\left(\boldsymbol{U}^{\otimes 2}_{\boldsymbol{\theta}_0,F_0}\right).$$

So the asymptotic variance matrix $\boldsymbol{\Sigma}$ can be decomposed as

$$\boldsymbol{\Sigma} = \boldsymbol{A}^{-1} + \boldsymbol{A}^{-1}\mathrm{P}\left(\boldsymbol{U}_{\boldsymbol{\theta}_0,F_0}\boldsymbol{\Phi}^t + \boldsymbol{\Phi}\boldsymbol{U}^t_{\boldsymbol{\theta}_0,F_0} + \boldsymbol{\Phi}^{\otimes 2}\right)\boldsymbol{A}^{-1}$$

as in Theorem 5.1.2. □

Similar to the estimation of the asymptotic variance in Theorem 4.2.2, the asymptotic variance matrix $\boldsymbol{\Sigma}$ has a complicated description. Since the $\sqrt{n}$-convergence rate and the asymptotic normality of the parameter estimate are achieved, we recommend the bootstrap resembling method to estimate $\boldsymbol{\Sigma}$.

5.2 A case study: Linear mixed-effects models

In this section, we apply the general method proposed in Section 5.1 to study parameter estimation in linear mixed-effects models with interval censored anchoring points. Linear mixed-effects models have been a work horse for analysis of longitudinal data with continuous outcomes. Here we discuss the situations where the anchoring points defining the longitudinal time scale are interval censored. A more specific real data application is described in Section 5.4.

5.2.1 Linear mixed-effects model with interval censored anchoring points

As before, we let $\boldsymbol{Y}$ denote the longitudinal outcome, $\boldsymbol{W}$ the covariates, and $(L, R]$ the time interval that brackets the unobserved anchoring point T. We consider a linear mixed-effects model:

$$\boldsymbol{Y}\big|(\boldsymbol{W}, L, R, T, \boldsymbol{r}) \sim \boldsymbol{N}\left(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{r}, \sigma^2\boldsymbol{I}\right), \quad \boldsymbol{r} \sim \boldsymbol{N}(\boldsymbol{0}, \boldsymbol{G}),$$

where $\boldsymbol{G}$ is the fixed but unknown covariance matrix of the random effects $\boldsymbol{r}$, $\boldsymbol{X} = \boldsymbol{X}(\boldsymbol{W}, L, R, T)$ and $\boldsymbol{Z} = \boldsymbol{Z}(\boldsymbol{W}, L, R, T)$ are respectively the design matrices for the fixed effects and random effects. Entries of $\boldsymbol{X}$ and $\boldsymbol{Z}$ are functions of $\boldsymbol{W}$, L, R and T. The parameter of interest is $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma^2, \boldsymbol{G})$. For the unobservable event time T, we assume a conditional independence that $F_{T|(\boldsymbol{W}, L<T\leq R)} = F_{T|L<T\leq R}$, which is not informative to $\boldsymbol{\theta}$. We also assume that $(L, R]$ is an independent censoring interval of T. It follows from these assumptions that $F_{T|(\boldsymbol{W}, L<T\leq R)}(t) = 1(L < t \leq R)\big(F_0(t) - F_0(L)\big)\Big/\big(F_0(R) - F_0(L)\big)$, where F_0 is the true but unknown CDF of the anchoring point T.

To estimate the true parameter $\boldsymbol{\theta}_0$ using the functional estimation process proposed in Section 5.1.1, we first obtain the NPMLE $\hat{F}_n$ of F_0 by using the interval censored data $(L, R]$. Assume that $\hat{F}_n$ has jumps p_j at time s_j, $j = 1, \cdots, k$, then the pseudo-likelihood of $\boldsymbol{Y}\big|(\boldsymbol{W}, L < T \leq R)$ is given as

$$\sum_{L<s_j\leq R} \frac{p_j}{\sum\limits_{L<s_k\leq R} p_k} |\boldsymbol{V}(s_j)|^{-1/2} \exp\left(-\frac{1}{2}\big(\boldsymbol{Y}-\boldsymbol{X}(s_j)\boldsymbol{\beta}\big)^t \boldsymbol{V}(s_j)^{-1}\big(\boldsymbol{Y}-\boldsymbol{X}(s_j)\boldsymbol{\beta}\big)\right),$$

where $\boldsymbol{X}(s_j) = \boldsymbol{X}|_{T=s_j}$, $\boldsymbol{Z}(s_j) = \boldsymbol{Z}|_{T=s_j}$ and $\boldsymbol{V}(s_j) = \sigma^2\boldsymbol{I} + \boldsymbol{Z}(s_j)\boldsymbol{G}\boldsymbol{Z}(s_j)^t$.

Given a random sample: $\Big\{(\boldsymbol{Y}_i, \boldsymbol{W}_i, L_i, R_i)\Big| i = 1, \cdots, n\Big\}$, for the i-th subject and index j such that $L_i < s_j \leq R_i$, we write $\boldsymbol{X}_{ij} = \boldsymbol{X}(\boldsymbol{W}_i, L_i, R_i, s_j)$, $\boldsymbol{Z}_{ij} = \boldsymbol{Z}_i(\boldsymbol{W}_i, L_i, R_i, s_j)$, $\boldsymbol{Vec}_{ij} = \boldsymbol{Y}_i - \boldsymbol{X}_{ij}\boldsymbol{\beta}$, $\boldsymbol{V}_{ij} = \sigma^2\boldsymbol{I} + \boldsymbol{Z}_{ij}\boldsymbol{G}\boldsymbol{Z}_{ij}^t$, and $p_{ij} = p_j \Big/ \sum\limits_{L_i<s_k\leq R_i} p_k$. Under this notational abbreviation, the log pseudo-likelihood for the observed data is given by

$$\mathcal{L}_n^{pl}(\boldsymbol{\theta}) \propto \sum_{i=1}^{n} \log\left[\sum_{L_i<s_j\leq R_i} p_{ij}|\boldsymbol{V}_{ij}|^{-1/2} \exp\left(-\frac{1}{2}\boldsymbol{Vec}_{ij}^t \boldsymbol{V}_{ij}^{-1}\boldsymbol{Vec}_{ij}\right)\right]. \tag{5.6}$$

The parameter estimate $\hat{\boldsymbol{\theta}}_n$ is the maximizer of the above function $\mathcal{L}_n^{pl}(\boldsymbol{\theta})$.

5.2.2 Computation

The function $\mathcal{L}_n^{pl}(\boldsymbol{\theta})$ has a complicated structure. The commonly used computation algorithms in fitting traditional linear mixed-effects models, namely the profile likelihood method and the restricted maximum likelihood method, do not seem to be easily applicable to maximize $\mathcal{L}_n^{pl}(\boldsymbol{\theta})$. Therefore, we propose a hybrid computation algorithm to maximize $\mathcal{L}_n^{pl}(\boldsymbol{\theta})$, combining the Fisher-Scoring (FS) algorithm with an

EM-algorithm. The hybrid algorithm is more robust than the FS-algorithm, and converges faster than the EM-algorithm.

The following notation is defined for the computation algorithm. For a positive definite matrix $\boldsymbol{G}$, there exists a unique lower triangular matrix $\boldsymbol{A}$ with positive diagonal entries such that $\boldsymbol{G} = \boldsymbol{A}\boldsymbol{A}^t$. We re-parameterize $\boldsymbol{G}$ using $\boldsymbol{A}$ for the computational advantage that the boundary condition is easy to check, because $\boldsymbol{G}$ is positive definite if and only if $\boldsymbol{A}$ has positive diagonal entries. To simplify notation, for any parameter value $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma^2, \boldsymbol{A})$, interval $(L_i, R_i]$, and index j such that $L_i < s_j \leq R_i$, let $\tilde{p}_{ij}(\boldsymbol{\theta})$ denote the quantity $\tilde{p}_{ij}(\boldsymbol{\theta}) = p_{ij}|\boldsymbol{V}_{ij}|^{-1/2} \exp\left(-\frac{1}{2}\boldsymbol{Vec}_{ij}^t \cdot \boldsymbol{V}_{ij}^{-1} \cdot \boldsymbol{Vec}_{ij}\right)$, and let $p_{ij}(\boldsymbol{\theta})$ denote the quantity $p_{ij}(\boldsymbol{\theta}) = \tilde{p}_{ij}(\boldsymbol{\theta}) \Big/ \sum_{L_i < s_k \leq R_i} \tilde{p}_{ik}(\boldsymbol{\theta})$. Let $\boldsymbol{X}_i$ and $\boldsymbol{Z}_i$ denote the functions of T, defined as $\boldsymbol{X}_i(T) = \boldsymbol{X}(\boldsymbol{W}_i, L_i, R_i, T)$ and $\boldsymbol{Z}_i(T) = \boldsymbol{Z}(\boldsymbol{W}_i, L_i, R_i, T)$.

The score function $\boldsymbol{U}(\boldsymbol{\theta})$ can be expressed as $\boldsymbol{U}(\boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}}\left(\mathcal{L}_n^{pl}(\boldsymbol{\theta})\right) = \sum_{i=1}^{n} \boldsymbol{U}_i(\boldsymbol{\theta})$, where $\boldsymbol{U}_i(\boldsymbol{\theta})$ is the score function computed from the i-th subject. Using the vector calculus reviewed in Chapter 2, the component functions of $\boldsymbol{U}_i(\boldsymbol{\theta})$ are computed as:

$$
\begin{aligned}
\boldsymbol{U}_i(\boldsymbol{\beta}) &= \sum_{L_i < s_j \leq R_i} \boldsymbol{X}_{ij}^t \boldsymbol{V}_{ij}^{-1} \boldsymbol{Vec}_{ij} \cdot p_{ij}(\boldsymbol{\theta}), \\
U_i(\sigma^2) &= \sum_{L_i < s_j \leq R_i} \frac{1}{2} Tr\left((\boldsymbol{V}_{ij}^{-1}\boldsymbol{Vec}_{ij})^{\otimes 2} - \boldsymbol{V}_{ij}^{-1}\right) \cdot p_{ij}(\boldsymbol{\theta}), \\
U_i(a_{pq}) &= \sum_{L_i < s_j \leq R_i} \boldsymbol{E}_p^t \boldsymbol{Z}_{ij}^t \left((\boldsymbol{V}_{ij}^{-1}\boldsymbol{Vec}_{ij})^{\otimes 2} - \boldsymbol{V}_{ij}^{-1}\right) \boldsymbol{Z}_{ij} \boldsymbol{A} \boldsymbol{E}_q \cdot p_{ij}(\boldsymbol{\theta}),
\end{aligned}
$$

where a_{pq} is the (p, q)-th entry of $\boldsymbol{A}$ and $p \geq q$, and $\boldsymbol{E}_k$ denotes the column vector with all entries equal 0 except that the k-th entry is 1. Using the above formula, the FS-algorithm with step-halving line search strategy is adopted.

To implement the FS-algorithm, a good initial value is very important. We propose to start with the EM-algorithm to obtain a reasonably good initial value. The derivation of the E-step and M-step is lengthy but algebraically straightforward. We provide the essential details in Section 5.2.3. Given the current parameter estimate $\boldsymbol{\theta}^{(k)} = \left(\boldsymbol{\beta}^{(k)}, {\sigma^{(k)}}^2, \boldsymbol{G}^{(k)}\right)$, the EM-algorithm computes the next estimate $\boldsymbol{\theta}^{(k+1)} = \left(\boldsymbol{\beta}^{(k+1)}, {\sigma^{(k+1)}}^2, \boldsymbol{G}^{(k+1)}\right)$ as

$$
\begin{aligned}
\boldsymbol{\beta}^{(k+1)} &= \boldsymbol{\beta}^{(k)} + {\sigma^{(k)}}^2 \cdot \arg\min_{\boldsymbol{\Delta}} \sum_{i=1}^{n} \boldsymbol{B}_i^{(k)}(\boldsymbol{\Delta}), \text{ where } \boldsymbol{\Delta} = \tfrac{\boldsymbol{\beta}-\boldsymbol{\beta}^{(k)}}{{\sigma^{(k)}}^2} \\
{\sigma^{(k+1)}}^2 &= {\sigma^{(k)}}^2 + {\sigma^{(k)}}^4 \left(\min_{\boldsymbol{\Delta}} \sum_{i=1}^{n} \boldsymbol{B}_i^{(k)}(\boldsymbol{\Delta})\right) \Big/ \sum_{i=1}^{n} q_i \\
&\quad -{\sigma^{(k)}}^4 \left(\sum_{i=1}^{n} E_{T_i^{(k)}} \left[Tr\left({\boldsymbol{V}_i^{(k)}}^{-1}\right)\right]\right) \Big/ \sum_{i=1}^{n} q_i \\
\boldsymbol{G}^{(k+1)} &= \boldsymbol{G}^{(k)} + \tfrac{1}{n}\boldsymbol{G}^{(k)} \left(\sum_{i=1}^{n} E_{T_i^{(k)}} \left[\boldsymbol{Z}_i^t \left({\boldsymbol{V}_i^{(k)}}^{-1} \boldsymbol{Vec}_i^{(k)}\right)^{\otimes 2} \boldsymbol{Z}_i\right]\right) \boldsymbol{G}^{(k)} \\
&\quad -\tfrac{1}{n}\boldsymbol{G}^{(k)} \left(\sum_{i=1}^{n} E_{T_i^{(k)}} \left[\boldsymbol{Z}_i^t {\boldsymbol{V}_i^{(k)}}^{-1} \boldsymbol{Z}_i\right]\right) \boldsymbol{G}^{(k)}
\end{aligned}
$$

where $\boldsymbol{B}_i^{(k)}(\boldsymbol{\Delta}) = E_{T_i^{(k)}} \left[\left({\boldsymbol{V}_i^{(k)}}^{-1} \boldsymbol{Vec}_i^{(k)} - \boldsymbol{X}_i \boldsymbol{\Delta}\right)^t\right]^{\otimes 2}$ is a function of $\boldsymbol{\Delta}$; $E_{T_i^{(k)}}$ denotes the expectation with respect to the random variable $T_i^{(k)}$, which has density $p_{ij}\left(\boldsymbol{\theta}^{(k)}\right)$ at $s_j \in (L_i, R_i]$ and 0 elsewhere; $\boldsymbol{V}_i(k) = {\sigma^{(k)}}^2 \boldsymbol{I} + \boldsymbol{Z}_i \boldsymbol{G}^{(k)} \boldsymbol{Z}_i^t$ and $\boldsymbol{Vec}_i^{(k)} = \boldsymbol{Y}_i - \boldsymbol{X}_i \boldsymbol{\beta}^{(k)}$ are functions of $T = T_i^{(k)}$; and q_i is the number of observations for the i-th subject. Note that $\sum_{i=1}^{n} \boldsymbol{B}_i^{(k)}(\boldsymbol{\Delta})$ is a quadratic function of $\boldsymbol{\Delta}$. So minimizing $\sum_{i=1}^{n} \boldsymbol{B}_i^{(k)}(\boldsymbol{\Delta})$ over $\boldsymbol{\Delta}$ is easy to accomplish. The above formula for $\boldsymbol{\theta}^{(k+1)}$ does not guarantee ${\sigma^{(k+1)}}^2$ or $\boldsymbol{G}^{(k+1)}$ to be non-negatively definite. The step-halving line search strategy is built in the algorithm to guarantee that $\boldsymbol{\theta}^{(k+1)}$ is inside $\boldsymbol{\Theta}$.

Regardless of initial values, the FS-algorithm could fail to converge because of poor approximation of the Hessian matrix, especially when the sample size is small. To overcome this algorithmic difficulty, we propose the following hybrid approach: For the current parameter estimate $\boldsymbol{\theta}^{(k)}$, we compute a temporary parameter estimate $\widetilde{\boldsymbol{\theta}}^{(k+1)}$ using the FS-algorithm. If $\mathcal{L}_n^{pl}\left(\widetilde{\boldsymbol{\theta}}^{(k+1)}\right) \geq \mathcal{L}_n^{pl}\left(\boldsymbol{\theta}^{(k)}\right)$, the updated parameter estimate $\boldsymbol{\theta}^{(k+1)}$ is set to be $\widetilde{\boldsymbol{\theta}}^{(k+1)}$. Otherwise, $\boldsymbol{\theta}^{(k+1)}$ is obtained by running the EM-algorithm for N iterations, where $N \geq 2$ is a pre-specified number. In other words, this hybrid algorithm attempts to use FS-algorithm to accelerate the EM-algorithm, while also keep the FS-iterations in the right track of increasing the likelihood with the assist of the EM-steps. The performance of this algorithm was tested in the simulation study and in a real data analysis reported in the following sections. In all these applications, the hybrid algorithm produced algorithmically convergent series of updated parameter estimates.

The proposed algorithm is implemented using the R software. A user friendly R function is provided in Chapter 6.

5.2.3 Derivation of the formula in Section 5.2.1

Using the vector calculus reviewed in Chapter 2, this subsection provides details in the derivation of the formulas in Section 5.2.2.

We first derive the score functions for $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma^2, \boldsymbol{A})$. Using Lemma 2.1.1, the computation of $\boldsymbol{U}_i(\boldsymbol{\beta})$ is straightforward.

$$\begin{aligned}
\boldsymbol{U}_i(\boldsymbol{\beta}) &= \frac{\partial}{\partial \boldsymbol{\beta}} \left\{ \log \left[\sum_{L_i < s_j \le R_i} p_{ij} |\boldsymbol{V}_{ij}|^{-\frac{1}{2}} \exp\left(-\tfrac{1}{2} \boldsymbol{Vec}_{ij}^t \boldsymbol{V}_{ij}^{-1} \boldsymbol{Vec}_{ij} \right) \right] \right\} \\
&= \frac{\sum_{L_i < s_j \le R_i} p_{ij} |\boldsymbol{V}_{ij}|^{-\frac{1}{2}} \exp\left(-\tfrac{1}{2} \boldsymbol{Vec}_{ij}^t \boldsymbol{V}_{ij}^{-1} \boldsymbol{Vec}_{ij} \right) \cdot \frac{\partial}{\partial \boldsymbol{\beta}} \left(-\tfrac{1}{2} \boldsymbol{Vec}_{ij}^t \boldsymbol{V}_{ij}^{-1} \boldsymbol{Vec}_{ij} \right)}{\sum_{L_i < s_j \le R_i} p_{ij} |\boldsymbol{V}_{ij}|^{-\frac{1}{2}} \exp\left(-\tfrac{1}{2} \boldsymbol{Vec}_{ij}^t \boldsymbol{V}_{ij}^{-1} \boldsymbol{Vec}_{ij} \right)} \\
&= \sum_{L_i < s_j \le R_i} \frac{p_{ij} |\boldsymbol{V}_{ij}|^{-\frac{1}{2}} \exp\left(-\tfrac{1}{2} \boldsymbol{Vec}_{ij}^t \boldsymbol{V}_{ij}^{-1} \boldsymbol{Vec}_{ij} \right) \cdot \boldsymbol{X}_{ij}^t \boldsymbol{V}_{ij}^{-1} \boldsymbol{Vec}_{ij}}{\sum_{L_i < s_j \le R_i} p_{ij} |\boldsymbol{V}_{ij}|^{-\frac{1}{2}} \exp\left(-\tfrac{1}{2} \boldsymbol{Vec}_{ij}^t \boldsymbol{V}_{ij}^{-1} \boldsymbol{Vec}_{ij} \right)} \\
&= \sum_{L_i < s_j \le R_i} p_{ij}(\boldsymbol{\theta}) \cdot \boldsymbol{X}_{ij}^t \boldsymbol{V}_{ij}^{-1} \boldsymbol{Vec}_{ij}.
\end{aligned}$$

For $U_i(\sigma^2)$, the same computation as for $\boldsymbol{U}_i(\boldsymbol{\beta})$ leads to the first step in the following calculations.

$$\begin{aligned}
U_i(\sigma^2) &= \sum_{L_i < s_j \le R_i} p_{ij}(\boldsymbol{\theta}) \cdot \frac{\partial}{\partial \sigma^2} \left(-\tfrac{1}{2} \boldsymbol{Vec}_{ij}^t \boldsymbol{V}_{ij}^{-1} \boldsymbol{Vec}_{ij} - \tfrac{1}{2} \log(|\boldsymbol{V}_{ij}|) \right) \\
&= \sum_{L_i < s_j \le R_i} p_{ij}(\boldsymbol{\theta}) \cdot \tfrac{1}{2} \left(\boldsymbol{Vec}_{ij}^t \boldsymbol{V}_{ij}^{-1} \frac{\partial \boldsymbol{V}_{ij}}{\partial \sigma^2} \boldsymbol{V}_{ij}^{-1} \boldsymbol{Vec}_{ij} - Tr\left(\boldsymbol{V}_{ij}^{-1} \frac{\partial \boldsymbol{V}_{ij}}{\partial \sigma^2} \right) \right) \\
&= \sum_{L_i < s_j \le R_i} p_{ij}(\boldsymbol{\theta}) \cdot \tfrac{1}{2} Tr\left(\left(\boldsymbol{V}_{ij}^{-1} \boldsymbol{Vec}_{ij} \right)^{\otimes 2} - \boldsymbol{V}_{ij}^{-1} \right)
\end{aligned}$$

where the second step follows from Lemma 2.1.3 and the last step follows from the fact that $\frac{\partial \boldsymbol{V}_{ij}}{\partial \sigma^2} = \boldsymbol{I}$.

For $U_i(a_{pq})$, the same computation as for $U_i(\sigma^2)$ leads to

$$U_i(a_{pq}) = \sum_{L_i < s_j \leq R_i} p_{ij}(\boldsymbol{\theta}) \tfrac{1}{2} \left(\boldsymbol{Vec}_{ij}^t \boldsymbol{V}_{ij}^{-1} \frac{\partial \boldsymbol{V}_{ij}}{\partial a_{pq}} \boldsymbol{V}_{ij}^{-1} \boldsymbol{Vec}_{ij} - Tr\left(\boldsymbol{V}_{ij}^{-1} \frac{\partial \boldsymbol{V}_{ij}}{\partial a_{pq}} \right) \right).$$

By Lemma 2.1.2, we have

$$\frac{\partial \boldsymbol{V}_{ij}}{\partial a_{pq}} = \frac{\partial}{\partial a_{pq}} \left(\sigma^2 \boldsymbol{I} + \boldsymbol{Z}_{ij} \boldsymbol{A} \boldsymbol{A}^t \boldsymbol{Z}_{ij}^t \right) = \boldsymbol{Z}_{ij} \boldsymbol{E}_p \boldsymbol{E}_q^t \boldsymbol{A}^t \boldsymbol{Z}_{ij}^t + \boldsymbol{Z}_{ij} \boldsymbol{A} \boldsymbol{E}_q \boldsymbol{E}_p^t \boldsymbol{Z}_{ij}^t,$$

which implies

$$\begin{aligned}
& Tr\left(\boldsymbol{V}_{ij}^{-1} \frac{\partial \boldsymbol{V}_{ij}}{\partial a_{pq}} \right) \\
= \; & Tr\left(\boldsymbol{V}_{ij}^{-1} \boldsymbol{Z}_{ij} \boldsymbol{E}_p \cdot \boldsymbol{E}_q^t \boldsymbol{A}^t \boldsymbol{Z}_{ij}^t \right) + Tr\left(\boldsymbol{V}_{ij}^{-1} \boldsymbol{Z}_{ij} \boldsymbol{A} \boldsymbol{E}_q \cdot \boldsymbol{E}_p^t \boldsymbol{Z}_{ij}^t \right) \\
= \; & Tr\left((\boldsymbol{V}_{ij}^{-1} \boldsymbol{Z}_{ij} \boldsymbol{E}_p)^t \cdot (\boldsymbol{E}_q^t \boldsymbol{A}^t \boldsymbol{Z}_{ij}^t)^t \right) + Tr\left(\boldsymbol{E}_p^t \boldsymbol{Z}_{ij}^t \cdot \boldsymbol{V}_{ij}^{-1} \boldsymbol{Z}_{ij} \boldsymbol{A} \boldsymbol{E}_q \right) \\
= \; & \boldsymbol{E}_p^t \boldsymbol{Z}_{ij}^t \boldsymbol{V}_{ij}^{-1} \cdot \boldsymbol{Z}_{ij} \boldsymbol{A} \boldsymbol{E}_q + \boldsymbol{E}_p^t \boldsymbol{Z}_{ij}^t \cdot \boldsymbol{V}_{ij}^{-1} \boldsymbol{Z}_{ij} \boldsymbol{A} \boldsymbol{E}_q = 2 \boldsymbol{E}_p^t \boldsymbol{Z}_{ij}^t \cdot \boldsymbol{V}_{ij}^{-1} \cdot \boldsymbol{Z}_{ij} \boldsymbol{A} \boldsymbol{E}_q
\end{aligned}$$

$$\begin{aligned}
& \boldsymbol{Vec}_{ij}^t \boldsymbol{V}_{ij}^{-1} \frac{\partial \boldsymbol{V}_{ij}}{\partial a_{pq}} \boldsymbol{V}_{ij}^{-1} \boldsymbol{Vec}_{ij} \\
= \; & \boldsymbol{Vec}_{ij}^t \boldsymbol{V}_{ij}^{-1} \boldsymbol{Z}_{ij} \boldsymbol{E}_p \cdot \boldsymbol{E}_q^t \boldsymbol{A}^t \boldsymbol{Z}_{ij}^t \boldsymbol{V}_{ij}^{-1} \boldsymbol{Vec}_{ij} + \boldsymbol{Vec}_{ij}^t \boldsymbol{V}_{ij}^{-1} \boldsymbol{Z}_{ij} \boldsymbol{A} \boldsymbol{E}_q \cdot \boldsymbol{E}_p^t \boldsymbol{Z}_{ij}^t \boldsymbol{V}_{ij}^{-1} \boldsymbol{Vec}_{ij} \\
= \; & \left(\boldsymbol{Vec}_{ij}^t \boldsymbol{V}_{ij}^{-1} \boldsymbol{Z}_{ij} \boldsymbol{E}_p \right)^t \cdot \left(\boldsymbol{E}_q^t \boldsymbol{A}^t \boldsymbol{Z}_{ij}^t \boldsymbol{V}_{ij}^{-1} \boldsymbol{Vec}_{ij} \right)^t \\
& + \boldsymbol{E}_p^t \boldsymbol{Z}_{ij}^t \boldsymbol{V}_{ij}^{-1} \boldsymbol{Vec}_{ij} \cdot \boldsymbol{Vec}_{ij}^t \boldsymbol{V}_{ij}^{-1} \boldsymbol{Z}_{ij} \boldsymbol{A} \boldsymbol{E}_q \\
= \; & 2 \boldsymbol{E}_p^t \boldsymbol{Z}_{ij}^t \cdot \left(\boldsymbol{V}_{ij}^{-1} \boldsymbol{Vec}_{ij} \right)^{\otimes 2} \cdot \boldsymbol{Z}_{ij} \boldsymbol{A} \boldsymbol{E}_q
\end{aligned}$$

Plug-in the above results, we have

$$U_i(a_{pq}) = \sum_{L_i < s_j \leq R_i} p_{ij}(\boldsymbol{\theta}) \cdot \boldsymbol{E}_p^t \boldsymbol{Z}_{ij}^t \left(\left(\boldsymbol{V}_{ij}^{-1} \boldsymbol{Vec}_{ij} \right)^{\otimes 2} - \boldsymbol{V}_{ij}^{-1} \right) \boldsymbol{Z}_{ij} \boldsymbol{A} \boldsymbol{E}_q.$$

Next, we derive the iterative formula used in the EM-algorithm. The complete log-likelihood corresponding to the marginal likelihood in Equation 5.6 for $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma^2, \boldsymbol{G})$ is

$$l_C(\boldsymbol{\theta}) \propto \sum_{i=1}^{n} -q_i \log(\sigma^2) - \sigma^{-2}\big((\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta} - \boldsymbol{Z}_i\boldsymbol{u}_i)^t\big)^{\otimes 2} - \log(|\boldsymbol{G}|) - \boldsymbol{u}_i^t\boldsymbol{G}^{-1}\boldsymbol{u}_i,$$

where q_i is the number of observations of the i-th subject, and the coefficient matrices $\boldsymbol{X}_i$ and $\boldsymbol{Z}_i$ are functions of the anchoring point T_i. There are two sets of missing data in $l_C(\boldsymbol{\theta})$, namely the random effects $\boldsymbol{u} = \{\boldsymbol{u}_1, \cdots \boldsymbol{u}_n\}$ and the anchoring points $T = \{T_1, \cdots, T_n\}$.

In the E-step, we need to compute the Q-function

$$Q\Big(\boldsymbol{\theta}; \boldsymbol{\theta}^{(k)}\Big) = E_{\boldsymbol{u},T\big|\boldsymbol{\theta}^{(k)},\hat{F}_n,Data}\big(l_C(\boldsymbol{\theta})\big),$$

where $\boldsymbol{\theta}^{(k)} = \big(\boldsymbol{\beta}^{(k)}, {\sigma^{(k)}}^2, \boldsymbol{G}^{(k)}\big)$ is the current parameter estimate. The expectation is computed in two steps as

$$\begin{aligned} Q\Big(\boldsymbol{\theta}; \boldsymbol{\theta}^{(k)}\Big) &= E_{\boldsymbol{u},T\big|\boldsymbol{\theta}^{(k)},\hat{F}_n,Data}\big(l_C(\boldsymbol{\theta})\big) \\ &= E_{T\big|\boldsymbol{\theta}^{(k)},\hat{F}_n,Data}\Big(E_{\boldsymbol{u}\big|\boldsymbol{\theta}^{(k)},T,Data}\big(l_C(\boldsymbol{\theta})\big)\Big) \end{aligned}$$

We now write down the last two conditional expectations in the above equation. Based on the usual linear mixed-effects model theory, it can be derived that the

conditional distribution of $\boldsymbol{u}_i \big| (\boldsymbol{\theta}^{(k)}, T_i, \boldsymbol{Y}_i, \boldsymbol{W}_i, L_i, R_i)$ is

$$N\left(\left(\sigma^{(k)^2}\boldsymbol{G}^{(k)^{-1}} + \boldsymbol{Z}_i^t\boldsymbol{Z}_i\right)^{-1}\boldsymbol{Z}_i^t(\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta}^{(k)}),\ \sigma^2(k)\left(\sigma^2(k)\boldsymbol{G}^{(k)^{-1}} + \boldsymbol{Z}_i^t\boldsymbol{Z}_i\right)^{-1}\right).$$

To simplify notation, let $E_{T_i^{(k)}}(-)$ denote the conditional expectation with respect to T_i given the observed data $(\boldsymbol{Y}_i, \boldsymbol{W}_i, L_i < T_i \leq R_i)$, the current parameter value $\boldsymbol{\theta}^{(k)}$, and the plug-in distribution $\hat{F}_n$ of T, which can be shown to have a point mass $p_{ij}\left(\boldsymbol{\theta}^{(k)}\right)$ at s_j, where $L_i < s_j \leq R_i$, and 0 elsewhere. See Section 5.2.2 for the definition of the function $p_{ij}(\boldsymbol{\theta})$. In other words, for any function $G(t)$, we have

$$E_{T_i^{(k)}}\big(G(t)\big) = \sum_{L_i < s_j \leq R_i} p_{ij}\left(\boldsymbol{\theta}^{(k)}\right) G(s_j).$$

Using these conditional distributions, the Q-function can be calculated as

$$\begin{aligned}
Q\big(\boldsymbol{\theta}; \boldsymbol{\theta}^{(k)}\big) &= E_{T\big|\boldsymbol{\theta}^{(k)}, \hat{F}_n, Data}\Big(E_{\boldsymbol{u}\big|\boldsymbol{\theta}^{(k)}, T, Data}\big(l_C(\boldsymbol{\theta})\big)\Big) \\
= \ & -n\log(|\boldsymbol{G}|) - \sigma^{(k)^2}\sum_{i=1}^{n} E_{T_i^{(k)}}\left[Tr\left(\left(\sigma^{(k)^2}\boldsymbol{G}^{(k)^{-1}} + \boldsymbol{Z}_i^t\boldsymbol{Z}_i\right)^{-1}\boldsymbol{G}^{-1}\right)\right] \\
& - \sum_{i=1}^{n} E_{T_i^{(k)}}\left[Tr\left(\left(\boldsymbol{G}^{(k)}\boldsymbol{Z}_i^t\boldsymbol{V}_i^{(k)^{-1}}\boldsymbol{Vec}_i^{(k)}\right)^{\otimes 2}\boldsymbol{G}^{-1}\right)\right] \\
& - \sum_{i=1}^{n} q_i\log(\sigma^2) - \sigma^{-2}\sum_{i=1}^{n} E_{T_i^{(k)}}\left[Tr\Big(\sigma^{(k)^2}\boldsymbol{I} - \sigma^{(k)^4}\boldsymbol{V}_i^{(k)^{-1}}\Big)\right] \\
& - \sigma^{-2}\sum_{i=1}^{n} E_{T_i^{(k)}}\left[Tr\Big(\sigma^{(k)^2}\boldsymbol{V}_i^{(k)^{-1}}\boldsymbol{Vec}_i^{(k)} - \boldsymbol{X}_i\big(\boldsymbol{\beta} - \boldsymbol{\beta}^{(k)}\big)\Big)^{\otimes 2}\right].
\end{aligned}$$

In the M-step, the updated estimate $\boldsymbol{\theta}^{(k+1)}$ is the maximizer of the above Q-function. Using the matrix identity that

$$(\boldsymbol{I}+\boldsymbol{M}_1\boldsymbol{M}_2)^{-1}=\boldsymbol{I}-\boldsymbol{M}_1(\boldsymbol{I}+\boldsymbol{M}_2\boldsymbol{M}_1)^{-1}\boldsymbol{M}_2,$$

one checks easily that

$$\sigma^{(k)^2}\boldsymbol{V}_i^{-1}(k)=\boldsymbol{I}-\boldsymbol{Z}_i\left(\boldsymbol{G}^{(k)^{-1}}+\frac{\boldsymbol{Z}_i^t\boldsymbol{Z}_i}{\sigma^{(k)^2}}\right)^{-1}\frac{\boldsymbol{Z}_i^t}{\sigma^{(k)^2}}.$$

It follows that, as a function of $\boldsymbol{\beta}$, we have

$$\begin{aligned}
& Q=Q(\boldsymbol{\beta}) \\
\propto\ & \sum_{i=1}^{n} E_{T_i^{(k)}}\left[\left(\boldsymbol{Y}_i-\boldsymbol{X}_i\boldsymbol{\beta}-\boldsymbol{Z}_i\left(\boldsymbol{G}^{(k)^{-1}}+\frac{\boldsymbol{Z}_i^t\boldsymbol{Z}_i}{\sigma^{(k)^2}}\right)^{-1}\frac{\boldsymbol{Z}_i^t}{\sigma^{(k)^2}}\left[\boldsymbol{Y}_i-\boldsymbol{X}_i\boldsymbol{\beta}^{(k)}\right]\right)^t\right]^{\otimes 2} \\
=\ & \sum_{i=1}^{n} E_{T_i^{(k)}}\left[\left(\boldsymbol{Y}_i-\boldsymbol{X}_i\boldsymbol{\beta}-\left[\boldsymbol{I}-\sigma^{(k)^2}\boldsymbol{V}_i^{(k)^{-1}}\right]\left[\boldsymbol{Y}_i-\boldsymbol{X}_i\boldsymbol{\beta}^{(k)}\right]\right)^t\right]^{\otimes 2} \\
=\ & \sum_{i=1}^{n} E_{T_i^{(k)}}\left[\left(\sigma^{(k)^2}\boldsymbol{V}_i^{(k)^{-1}}\boldsymbol{Vec}_i^{(k)}-\boldsymbol{X}_i(\boldsymbol{\beta}-\boldsymbol{\beta}^{(k)})\right)^t\right]^{\otimes 2}=\sigma^{(k)^4}\sum_{i=1}^{n}\boldsymbol{B}_i^{(k)}(\boldsymbol{\Delta})
\end{aligned}$$

where $\boldsymbol{\Delta}=(\boldsymbol{\beta}-\boldsymbol{\beta}^{(k)})\Big/\sigma^{(k)^2}$. Since $\sum_{i=1}^{n}\boldsymbol{B}_i^{(k)}(\boldsymbol{\Delta})$ is a quadratic function on $\boldsymbol{\Delta}$, its minimizer can be easily calculated. This shows that

$$\boldsymbol{\beta}^{(k+1)}=\boldsymbol{\beta}^{(k)}+\sigma^{(k)^2}\cdot\arg\min_{\boldsymbol{\Delta}}\left(\sum_{i=1}^{n}\boldsymbol{B}_i^{(k)}(\boldsymbol{\Delta})\right).$$

Plug in $\boldsymbol{\beta} = \boldsymbol{\beta}^{(k+1)}$, as a function of σ^2, we have

$$
\begin{aligned}
& Q = Q(\sigma^2; \boldsymbol{\beta} = \boldsymbol{\beta}^{(k+1)}) \\
\propto \; & -\sum_{i=1}^{n} \dim(\boldsymbol{V}_i) \log(\sigma^2) - \sigma^{-2} {\sigma^{(k)}}^4 \min_{\boldsymbol{\Delta}} \left(\sum_{i=1}^{n} \boldsymbol{B}_i^{(k)}(\boldsymbol{\Delta}) \right) \\
& -\sigma^{-2} \sum_{i=1}^{n} E_{T_i^{(k)}} \left[Tr \left(\boldsymbol{Z}_i \cdot \left({\boldsymbol{G}^{(k)}}^{-1} + \frac{\boldsymbol{Z}_i^t \boldsymbol{Z}_i}{{\sigma^{(k)}}^2} \right)^{-1} \cdot \boldsymbol{Z}_i^t \right) \right] \\
= \; & -\sum_{i=1}^{n} \dim(\boldsymbol{V}_i) \log(\sigma^2) - \sigma^{-2} {\sigma^{(k)}}^4 \min_{\boldsymbol{\Delta}} \left(\sum_{i=1}^{n} \boldsymbol{B}_i^{(k)}(\boldsymbol{\Delta}) \right) \\
& -\sigma^{-2} \sum_{i=1}^{n} E_{T_i^{(k)}} \left[Tr \left({\sigma^{(k)}}^2 \boldsymbol{I}_i - {\sigma^{(k)}}^4 {\boldsymbol{V}_i^{(k)}}^{-1} \right) \right]
\end{aligned}
$$

where $\boldsymbol{I}_i$ is the identity matrix of the same size as $\boldsymbol{V}_i$. Using elementary calculus, we can easily get the minimizer of $Q(\sigma^2; \boldsymbol{\beta} = \boldsymbol{\beta}^{(k+1)})$ as

$${\sigma^{(k+1)}}^2 = {\sigma^{(k)}}^2 + {\sigma^{(k)}}^4 \left\{ \min_{\boldsymbol{\Delta}} \left(\sum_{i=1}^{n} \boldsymbol{B}_i^{(k)}(\boldsymbol{\Delta}) \right) - \sum_{i=1}^{n} E_{T_i^{(k)}} \left[Tr \left({\boldsymbol{V}_i^{(k)}}^{-1} \right) \right] \right\} \Big/ \sum_{i=1}^{n} q_i$$

To compute $\boldsymbol{G}^{(k+1)}$, we simplify the notation by defining

$$
\begin{aligned}
\boldsymbol{M}_{1i} &= \left({\boldsymbol{G}^{(k)}}^{-1} + \boldsymbol{Z}_i^t \boldsymbol{Z}_i / {\sigma^{(k)}}^2 \right)^{-1} \\
\boldsymbol{M}_{2i} &= \left({\boldsymbol{G}^{(k)}}^{-1} + \boldsymbol{Z}_i^t \boldsymbol{Z}_i / {\sigma^{(k)}}^2 \right)^{-1} \frac{\boldsymbol{Z}_i^t}{{\sigma^{(k)}}^2} \big(\boldsymbol{Y}_i - \boldsymbol{X}_i \boldsymbol{\beta}^{(k)} \big).
\end{aligned}
$$

Note that $\boldsymbol{M}_{1i}$ is a symmetric matrix and $\boldsymbol{M}_{2i}$ is a column vector. One checks that

$$\left({\boldsymbol{G}^{(k)}}^{-1} + \boldsymbol{Z}_i^t \boldsymbol{Z}_i / {\sigma^{(k)}}^2 \right) \left(\boldsymbol{G}^{(k)} - \boldsymbol{G}^{(k)} \boldsymbol{Z}_i^t {\boldsymbol{V}_i^{(k)}}^{-1} \boldsymbol{Z}_i \boldsymbol{G}^{(k)} \right) = \boldsymbol{I}$$

and

$$\left(\boldsymbol{G}^{(k)^{-1}} + \boldsymbol{Z}_i^t \boldsymbol{Z}_i / {\sigma^{(k)}}^2\right) \cdot \boldsymbol{G}^{(k)} \boldsymbol{Z}_i^t \boldsymbol{V}_i^{(k)^{-1}} \boldsymbol{Vec}_i^{(k)} = \frac{\boldsymbol{Z}_i^t}{{\sigma^{(k)}}^2}\big(\boldsymbol{Y}_i - \boldsymbol{X}_i \boldsymbol{\beta}^{(k)}\big),$$

which immediately implies that

$$\boldsymbol{M}_{1i} = \boldsymbol{G}^{(k)} - \boldsymbol{G}^{(k)} \boldsymbol{Z}_i^t \boldsymbol{V}_i^{(k)^{-1}} \boldsymbol{Z}_i \boldsymbol{G}^{(k)}, \quad \boldsymbol{M}_{2i} = \boldsymbol{G}^{(k)} \boldsymbol{Z}_i^t \boldsymbol{V}_i^{(k)^{-1}} \boldsymbol{Vec}_i^{(k)}. \tag{5.7}$$

Using $\boldsymbol{M}_{1i}$ and $\boldsymbol{M}_{2i}$, we can rewrite Q as a function of $\boldsymbol{G}$ as

$$Q = Q(\boldsymbol{G}) = -n \log(|\boldsymbol{G}|) - \sum_{i=1}^{n} E_{T_i^{(k)}} Tr\big(\boldsymbol{G}^{-1} \boldsymbol{M}_{1i}\big) - \sum_{i=1}^{n} E_{T_i^{(k)}} {\boldsymbol{M}_{2i}}^t \boldsymbol{G}^{-1} \boldsymbol{M}_{2i}.$$

Let t_{pq} be the (p,q)-th entry, which is also the (q,p)-th entry of $\boldsymbol{G}$. First, we assume $p \neq q$. In this case, we have $\frac{\partial \boldsymbol{G}}{\partial t_{pq}} = \boldsymbol{E}_p \boldsymbol{E}_q^t + \boldsymbol{E}_q \boldsymbol{E}_p^t$. Using Lemma 2.1.3, we have

$$\begin{aligned}
\frac{\partial Q}{\partial t_{pq}} = \; & -nTr\big(\boldsymbol{G}^{-1}(\boldsymbol{E}_p \boldsymbol{E}_q^t + \boldsymbol{E}_q \boldsymbol{E}_p^t)\big) \\
& + \sum_{i=1}^{n} E_{T_i^{(k)}} \big(\boldsymbol{G}^{-1} \boldsymbol{E}_p \boldsymbol{E}_q^t \boldsymbol{G}^{-1} \boldsymbol{M}_{1i} + \boldsymbol{G}^{-1} \boldsymbol{E}_q \boldsymbol{E}_p^t \boldsymbol{G}^{-1} \boldsymbol{M}_{1i}\big) \\
& + \sum_{i=1}^{n} E_{T_i^{(k)}} \big({\boldsymbol{M}_{2i}}^t \boldsymbol{G}^{-1} \boldsymbol{E}_p \cdot \boldsymbol{E}_q^t \boldsymbol{G}^{-1} \boldsymbol{M}_{2i} + {\boldsymbol{M}_{2i}}^t \boldsymbol{G}^{-1} \boldsymbol{E}_q \cdot \boldsymbol{E}_p^t \boldsymbol{G}^{-1} \boldsymbol{M}_{2i}\big) \\
= \; & -n\big(\boldsymbol{E}_q^t \boldsymbol{G}^{-1} \boldsymbol{E}_p + \boldsymbol{E}_p^t \boldsymbol{G}^{-1} \boldsymbol{E}_q\big) \\
& + \sum_{i=1}^{n} E_{T_i^{(k)}} Tr\big(\boldsymbol{E}_q^t \boldsymbol{G}^{-1} \boldsymbol{M}_{1i} \boldsymbol{G}^{-1} \boldsymbol{E}_p + \boldsymbol{E}_p^t \boldsymbol{G}^{-1} \boldsymbol{M}_{1i} \boldsymbol{G}^{-1} \boldsymbol{E}_q\big) \\
& + \sum_{i=1}^{n} E_{T_i^{(k)}} \big(\boldsymbol{E}_q^t \boldsymbol{G}^{-1} {\boldsymbol{M}_{2i}}^{\otimes 2} \boldsymbol{G}^{-1} \boldsymbol{E}_p + \boldsymbol{E}_p^t \boldsymbol{G}^{-1} {\boldsymbol{M}_{2i}}^{\otimes 2} \boldsymbol{G}^{-1} \boldsymbol{E}_q\big) \\
= \; & 2\boldsymbol{E}_p \left(-n\boldsymbol{G}^{-1} + \boldsymbol{G}^{-1} \left(\sum_{i=1}^{n} E_{T_i^{(k)}} \left(\boldsymbol{M}_{1i} + {\boldsymbol{M}_{2i}}^{\otimes 2}\right)\right) \boldsymbol{G}^{-1}\right) \boldsymbol{E}_q^t
\end{aligned}$$

where the last equality follows from the fact that $\boldsymbol{G}^{-1}, \boldsymbol{M}_{1i}$ and $\boldsymbol{M}_{2i}{}^{\otimes 2}$ are all symmetric. Setting $\frac{\partial Q}{\partial t_{pq}} = 0$ we have

$$0 = \boldsymbol{E}_p \left(-n\boldsymbol{G}^{(k+1)^{-1}} + \boldsymbol{G}^{(k+1)^{-1}} \left(\sum_{i=1}^{n} E_{T_i^{(k)}} \left(\boldsymbol{M}_{1i} + \boldsymbol{M}_{2i}{}^{\otimes 2} \right) \right) \boldsymbol{G}^{(k+1)^{-1}} \right) \boldsymbol{E}_q^t.$$

A similar calculation shows that the above equation also holds for $p = q$. Since the above equation holds for all indices (p, q), we have

$$\begin{aligned} & 0 = -n\boldsymbol{G}^{(k+1)^{-1}} + \boldsymbol{G}^{(k+1)^{-1}} \left(\sum_{i=1}^{n} E_{T_i^{(k)}} \left(\boldsymbol{M}_{1i} + \boldsymbol{M}_{2i}{}^{\otimes 2} \right) \right) \boldsymbol{G}^{(k+1)^{-1}} \\ \Rightarrow \quad & 0 = -n\boldsymbol{G}^{(k+1)} + \sum_{i=1}^{n} E_{T_i^{(k)}} \left(\boldsymbol{M}_{1i} + \boldsymbol{M}_{2i}{}^{\otimes 2} \right) \text{ by } \times \boldsymbol{G}^{(k+1)} \text{ from left and right} \\ \Rightarrow \quad & \boldsymbol{G}^{(k+1)} = \tfrac{1}{n} \sum_{i=1}^{n} E_{T_i^{(k)}} \left(\boldsymbol{M}_{1i} + \boldsymbol{M}_{2i}{}^{\otimes 2} \right). \end{aligned}$$

Plug-in the identities in Equations (5.7), the above equation can be rewritten as

$$\boldsymbol{G}^{(k+1)} = \boldsymbol{G}^{(k)} + \frac{1}{n} \boldsymbol{G}^{(k)} \sum_{i=1}^{n} E_{T_i^{(k)}} \left[\boldsymbol{Z}_i^t \left(\left(\boldsymbol{V}_i^{(k)^{-1}} \boldsymbol{Vec}_i^{(k)} \right)^{\otimes 2} - \boldsymbol{V}_i^{(k)^{-1}} \right) \boldsymbol{Z}_i \right] \boldsymbol{G}^{(k)}.$$

5.3 Simulation study

Simulation studies were conducted to investigate the performance of the proposed model in finite-sample situations. Two sample sizes were considered: $n = 200$, and 400. To evaluate the impact of the interval lengths, for each given sample size n, we simulated two scenarios: (1) average censoring interval length $l = 1$; and (2) average censoring interval length $l = 2$.

For a given sample size n and the average interval length l, we generated a total of 1500 simulated data sets as follows: For the i-th subject, the true anchoring point

T_i was independently generated from a Weibull distribution with shape parameter 80 and scale parameter 12. For each non-overlapping time window $(kl, (k+1)l]$, where $k = 0, 1, 2, \cdots$, a uniformly distributed screening time was generated. The censoring interval $(L_i, R_i]$ was identified as the adjacent screening times that bracket T_i, i.e., $L_i < T_i \leq R_i$. To allow covariates in the proposed method, we also simulated a binary covariate X_{1i} with equal probability $P(X_{1i} = 0) = P(X_{2i} = 1) = 1/2$, and a continuous covariate X_{2i} that was $N(0, 1)$ distributed. The observations at the two endpoints of the censoring interval, $Y_{L,i}$ and $Y_{R,i}$, were then generated from the following linear mixed-effects model

$$\begin{cases} Y_{L,i} = \lambda + \beta_1 X_{1i} + \beta_2 X_{2i} + \alpha(L_i - T_i) + \lambda_i + \alpha_i(L_i - T_i) + \epsilon_{L,i} \\ Y_{R,i} = \lambda + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta(R_i - T_i) + \lambda_i + \beta_i(R_i - T_i) + \epsilon_{R,i} \end{cases} \tag{5.8}$$

where $(\lambda, \beta_1, \beta_2, \alpha, \beta)$ were the parameters for the population fixed effects, $(\lambda_i, \alpha_i, \beta_i)$ were the subject-specific random deviations, and $\epsilon_{L,i}$ and $\epsilon_{R,i}$ were the independent error terms. The true values of the parameters were $\lambda = 50, \beta_1 = -2, \beta_2 = -3, \alpha = 5, \beta = 8$. The random effects $(\lambda_i, \alpha_i, \beta_i)$ were generated from $\boldsymbol{N}(\boldsymbol{0}, \boldsymbol{G})$, and the error terms $(\epsilon_{L,i}, \epsilon_{R,i})$ were generated from $\boldsymbol{N}(\boldsymbol{0}, diag(\sigma^2, \sigma^2))$, where $\sigma^2 = 2.25$ and

$$\boldsymbol{G} = \begin{pmatrix} 9 & 1 & -1 \\ 1 & 3 & -1 \\ -1 & -1 & 4 \end{pmatrix}.$$

Model 5.8 was a changing-point model in the sense of van den Hout et al. (2013) and many others. But our method is not restricted to changing-point problems. For example, if the data before the anchoring points are not available, there is no "change" to model at all. Our method can still estimate the average slope β that is "anchored" by the anchoring event, as long as the censoring-intervals are observed.

For each simulated data set, four models were fitted. First, the linear mixed-effects model Model (5.8) was fitted using the proposed two-stage estimation procedure, with standard errors of estimators estimated using Bootstrap method with 50 resamples. Second, the distribution-free model proposed in Chapter 4 was fitted, by ignoring the random effects in Model (5.8). To compare with existing methods used in practice (Shankar et al., 2005), the third method used the "nlme" R-package to fit Model (5.8) with the unobserved anchoring point T_i imputed by $\frac{L_i+R_i}{2}$, the midpoint of the censoring interval. The last method also used the "nlme" R-package to fit Model (5.8), where the true anchoring point T_i was used. Note that our research focuses on the situation when T_i is unobserved, which is often encountered in real studies. So the last model is often not available in practice. It was fitted only to serve as a benchmark model to evaluate the efficiency loss in our proposed method.

It is worth noting that the standard algorithm with the "nlme" R-package for computing the maximum likelihood estimates for longitudinal normal data with midpoint imputation of T failed to algorithmically converge in about 20% of the simulated data sets. For each sample size n and average censoring interval length l, we summarized the simulation result from the first 1000 data sets that provided numerically convergent parameter estimates from all of the four methods. The percentage of bias in parameter estimates (% Bias), Monte-Carlo standard deviation (M-C SD),

average Bootstrap standard error (Av. SE) and coverage probability of the 95% Wald-confidence intervals (95% CP) are reported in Tables 5.1 and 5.2.

Table 5.1: Simulation result for Scenario (1)

	sample size =200					sample size =400				
	λ	β_1	β_2	α	β	λ	β_1	β_2	α	β
linear mixed-effects model:										
% Bias	0.073	0.834	0.976	1.145	0.838	0.067	0.063	0.620	0.640	0.541
M-C SD	0.571	0.469	0.812	0.415	0.432	0.383	0.335	0.559	0.284	0.290
Av. SE	0.565	0.468	0.803	0.400	0.418	0.395	0.328	0.568	0.283	0.294
95% CP	0.947	0.942	0.935	0.921	0.923	0.951	0.943	0.945	0.940	0.948
distribution-free model:										
% Bias	0.053	0.875	0.908	0.162	1.172	0.034	0.017	0.579	0.714	0.670
M-C SD	0.596	0.471	0.809	0.535	0.567	0.395	0.337	0.561	0.375	0.379
Av. SE	0.587	0.467	0.800	0.512	0.539	0.414	0.328	0.569	0.370	0.389
95% CP	0.947	0.943	0.936	0.928	0.928	0.950	0.938	0.943	0.937	0.946
midpoint imputation:										
% Bias	0.438	0.710	0.784	8.076	5.234	0.373	0.341	0.351	6.482	4.100
M-C SD	0.917	0.506	0.878	1.370	1.414	0.620	0.368	0.623	0.931	0.953
Av. SE	0.895	0.502	0.871	1.316	1.359	0.631	0.358	0.621	0.930	0.961
95% CP	0.944	0.947	0.943	0.929	0.938	0.939	0.954	0.950	0.933	0.935
event time is known:										
% Bias	0.021	1.194	0.868	0.048	0.041	0.009	0.078	0.526	0.256	0.087
M-C SD	0.536	0.447	0.769	0.390	0.397	0.353	0.323	0.533	0.253	0.263
Av. SE	0.523	0.442	0.766	0.357	0.368	0.370	0.314	0.544	0.254	0.262
95% CP	0.947	0.938	0.945	0.929	0.930	0.965	0.945	0.954	0.944	0.946

The simulation result in Tables 5.1 and 5.2 show that the proposed method has a generally excellent finite-sample performance. Estimation bias is very small and virtually ignorable when the sample size is large (less than 1%). Both bias and Monte-Carlo standard deviation decrease with sample size. The bootstrap standard error are very close to Monte-Carlo standard deviation, which justifies the use of bootstrap method in estimating the standard errors. The empirical coverage probabilities of the

Table 5.2: Simulation result for Scenario (2)

	sample size =200					sample size =400				
	λ	β_1	β_2	α	β	λ	β_1	β_2	α	β
linear mixed-effects model:										
% Bias	0.119	0.148	0.051	0.607	0.436	0.112	0.063	1.098	0.501	0.328
M-C SD	0.612	0.496	0.840	0.285	0.312	0.420	0.338	0.597	0.202	0.212
Av. SE	0.616	0.492	0.843	0.287	0.305	0.433	0.343	0.597	0.203	0.212
95% CP	0.949	0.938	0.944	0.946	0.938	0.955	0.958	0.947	0.944	0.945
distribution-free model:										
% Bias	0.079	0.318	0.053	0.375	0.517	0.060	0.001	0.966	0.410	0.356
M-C SD	0.657	0.513	0.855	0.368	0.390	0.456	0.345	0.615	0.264	0.263
Av. SE	0.656	0.500	0.860	0.351	0.371	0.466	0.353	0.612	0.260	0.266
95% CP	0.937	0.940	0.940	0.917	0.933	0.951	0.955	0.946	0.936	0.950
midpoint imputation:										
% Bias	0.855	0.155	0.694	7.630	4.895	0.734	0.571	0.792	6.464	3.903
M-C SD	1.029	0.616	1.083	0.811	0.848	0.712	0.439	0.754	0.545	0.573
Av. SE	1.011	0.611	1.060	0.779	0.823	0.712	0.435	0.754	0.550	0.580
95% CP	0.928	0.948	0.950	0.916	0.916	0.917	0.953	0.948	0.915	0.927
event time is known:										
% Bias	0.005	0.198	0.017	0.001	0.126	0.021	0.056	0.915	0.081	0.098
M-C SD	0.569	0.475	0.788	0.281	0.298	0.382	0.328	0.560	0.192	0.199
Av. SE	0.556	0.458	0.794	0.269	0.280	0.393	0.325	0.563	0.191	0.198
95% CP	0.952	0.928	0.944	0.944	0.937	0.957	0.946	0.946	0.951	0.953

95% Wald-confidence intervals are all close to the nominal level. The result clearly validates the large sample theory derived in Theorem 5.1.2.

The same pattern of simulation results using the linear mixed-effects model was also observed for the distribution-free model. To numerically evaluated the gain in estimation efficiency, we calculated the ratio of Monte-Carlo standard deviations of the parameter estimates using the linear mixed-effects model to that of the distribution-free model. See Table 5.3. The ratios showed that, if a covariate is not associated with the unobserved anchoring point T, then there is no noticeably efficiency gain

in estimating the corresponding parameter, such as λ, β_1, and β_2. If a covariate is associated with the unobserved anchoring point T, then there is about 25% efficiency gain in estimating the corresponding parameter, such as α and β.

Table 5.3: Ratio of Monte Carlo standard deviations of linear mixed-effects model vs distribution-free model

	sample size =200					sample size =400				
	λ	β_1	β_2	α	β	λ	β_1	β_2	α	β
Scenario (1)										
ratio	0.959	0.997	1.004	0.776	0.762	0.970	0.996	0.995	0.757	0.764
Scenario (2)										
ratio	0.932	0.966	0.982	0.773	0.800	0.921	0.981	0.972	0.764	0.806

For a fixed sample size n, the simulation study shows an improved model performance with wider censoring intervals, which is expected in view of the regularity condition (F4). When censoring interval is narrow and n is not large enough, the NPMLE $\hat{F}_n$ is possibly not a satisfactory estimate of F_0, which leads to a decreased performance of the proposed method. In such situations, a larger sample size is generally required to have an asymptotic normal distribution for the estimated model parameters.

Not surprisingly, the midpoint imputation method did not perform well. For parameters α and β that are associated with the anchoring point T, the bias did not appear to decrease with sample size; the Monte-Carlo standard deviations almost tripled in comparison with that of the proposed method; the coverage probability of the 95% Wald-confidence intervals noticeably deviated from the nominal level. However, the estimation bias of the parameters $\lambda, \beta_1, \beta_2$ (the parameters that are not associated with the anchoring point T) was ignorable compared to that of α and β.

As expected, the method using the true anchoring point outperformed all its competitors. But it is important to note that the proposed method also had practically ignorable bias and comparable efficiency in parameter estimation. To empirically evaluate the relative efficiency, we computed the ratios of the Monte-Carlo standard deviations of the parameter estimates between the proposed method and the model with known true anchoring points; see Table 5.4. The Monte-Carlo standard deviations of the parameter estimates in the proposed method were 1.4%-12.3% larger, indicating only a mild loss of efficiency compared to the ideal situation of knowing T values.

Table 5.4: Ratio of Monte Carlo standard deviations of linear mixed-effects model vs the model knowing true event time

	sample size =200					sample size =400				
	λ	β_1	β_2	α	β	λ	β_1	β_2	α	β
Scenario (1)										
ratio	1.065	1.049	1.056	1.064	1.088	1.085	1.037	1.049	1.123	1.103
Scenario (2)										
ratio	1.076	1.044	1.066	1.014	1.047	1.099	1.030	1.066	1.052	1.065

To investigate the robustness of the normality assumption on the error term in the proposed method, we repeated the previous simulation with a change that the error terms, $\epsilon_{L,i}$ and $\epsilon_{R,i}$ in Model 5.8, were simulated from a mixture normal distribution

$$Z_1 Z_2 + (1 - Z_1) Z_3$$

where Z_1 was a binary random variable with equal probability $P(Z_1 = 0) = P(Z_1 = 1) = \frac{1}{2}$, Z_2 was $N(-1, 1^2)$ distributed, and Z_3 was $N(1, 0.5^2)$ distributed. The simulation results are summarized in Tables 5.5 and 5.6.

Table 5.5: Simulation result for Scenario (1) with mixture normal errors

	sample size =200					sample size =400				
	λ	β_1	β_2	α	β	λ	β_1	β_2	α	β
linear mixed-effects model:										
% Bias	0.098	0.830	0.679	1.360	0.866	0.084	0.204	0.490	0.953	0.780
M-C SD	0.552	0.462	0.808	0.377	0.388	0.379	0.331	0.552	0.268	0.277
Av. SE	0.554	0.462	0.795	0.371	0.387	0.386	0.323	0.560	0.262	0.271
95% CP	0.944	0.943	0.948	0.939	0.939	0.948	0.939	0.947	0.928	0.935
distribution-free model:										
% Bias	0.064	0.716	0.717	0.062	1.157	0.037	0.153	0.518	0.635	0.733
M-C SD	0.581	0.464	0.808	0.512	0.541	0.394	0.331	0.554	0.363	0.382
Av. SE	0.577	0.461	0.791	0.494	0.524	0.408	0.324	0.562	0.359	0.376
95% CP	0.941	0.943	0.943	0.936	0.929	0.956	0.938	0.950	0.936	0.928
midpoint imputation:										
% Bias	0.438	1.039	0.764	7.965	5.007	0.370	0.131	0.158	6.378	4.066
M-C SD	0.885	0.501	0.871	1.326	1.359	0.622	0.363	0.621	0.927	0.961
Av. SE	0.880	0.495	0.858	1.295	1.336	0.622	0.353	0.611	0.916	0.946
95% CP	0.936	0.943	0.951	0.933	0.936	0.938	0.954	0.956	0.928	0.928
event time is known:										
% Bias	0.014	1.007	0.690	0.008	0.035	0.001	0.039	0.391	0.002	0.117
M-C SD	0.513	0.437	0.763	0.350	0.355	0.345	0.317	0.524	0.236	0.247
Av. SE	0.506	0.431	0.748	0.324	0.337	0.359	0.307	0.533	0.233	0.240
95% CP	0.953	0.940	0.942	0.941	0.939	0.970	0.946	0.954	0.946	0.945

Although the error terms were not normally distributed, the simulation in Tables 5.5 and 5.6 show a similar pattern as those reported in Tables 5.1 and 5.2. Estimation bias is small and ignorable in larger samples; the average Bootstrap standard errors are close to the Monte-Carlo standard deviations; the observed coverage probabilities of the 95% confidence intervals are close to the nominal level. In conclusion, the simulation result indicates that the normal assumption on the error terms in the proposed method is generally robust for larger samples.

Table 5.6: Simulation result for Scenario (2) with mixture normal errors

	sample size =200					sample size =400				
	λ	β_1	β_2	α	β	λ	β_1	β_2	α	β
linear mixed-effects model:										
% Bias	0.134	0.600	0.546	0.658	0.338	0.109	0.120	0.776	0.510	0.364
M-C SD	0.606	0.475	0.858	0.275	0.295	0.417	0.336	0.587	0.196	0.204
Av. SE	0.605	0.486	0.832	0.275	0.293	0.424	0.338	0.588	0.193	0.203
95% CP	0.943	0.944	0.936	0.946	0.940	0.951	0.947	0.952	0.933	0.939
distribution-free model:										
% Bias	0.070	0.537	0.678	0.476	0.346	0.050	0.116	0.809	0.506	0.318
M-C SD	0.650	0.495	0.872	0.364	0.378	0.455	0.342	0.608	0.258	0.260
Av. SE	0.649	0.497	0.853	0.346	0.364	0.460	0.349	0.605	0.255	0.261
95% CP	0.940	0.943	0.937	0.924	0.938	0.947	0.948	0.951	0.936	0.937
midpoint imputation:										
% Bias	0.821	1.322	0.087	7.206	4.452	0.741	0.246	0.795	6.376	3.861
M-C SD	1.006	0.604	1.081	0.804	0.835	0.698	0.436	0.755	0.528	0.560
Av. SE	0.993	0.602	1.044	0.768	0.809	0.702	0.429	0.744	0.542	0.572
95% CP	0.928	0.948	0.938	0.924	0.930	0.924	0.949	0.945	0.921	0.924
event time is known:										
% Bias	0.010	0.551	0.482	0.043	0.007	0.017	0.062	0.671	0.070	0.132
M-C SD	0.545	0.449	0.801	0.266	0.278	0.374	0.318	0.550	0.186	0.191
Av. SE	0.538	0.447	0.775	0.254	0.265	0.382	0.318	0.551	0.182	0.189
95% CP	0.956	0.942	0.948	0.943	0.945	0.963	0.954	0.957	0.938	0.943

For the distribution-free model, similar patterns were also observed. There were noticeably larger estimation variations for α and β, but not for λ, β_1 or β_2, similar to the simulation results for normal errors. See the ratio of Monti-Carlo standard deviation of these two models, calculated in Table 5.7.

Again, the midpoint imputation method showed worse performance, especially in estimating α and β. The relative estimation efficiency of the proposed method v.s. the ideal method using true anchoring point was empirically evaluated using the ratios of the Monte-Carlo standard deviations. See Table 5.8. Similar to the situation with

Table 5.7: Ratio of Monte Carlo standard deviations of linear mixed-effects model vs distribution-free model with mixture normal errors

	sample size =200					sample size =400				
	λ	β_1	β_2	α	β	λ	β_1	β_2	α	β
Scenario (1)										
ratio	0.951	0.995	1.000	0.736	0.718	0.960	1.002	0.997	0.739	0.725
Scenario (2)										
ratio	0.934	0.958	0.984	0.755	0.781	0.916	0.984	0.965	0.760	0.785

normal error terms, the Monte-Carlo standard deviation of the parameter estimates was about 3.4%-13.6% larger, indicating again only a mild loss of efficiency with the non-normal errors.

Table 5.8: Ratio of Monte Carlo standard deviations of proposed model vs the model knowing true event times with mixture normal errors

	sample size =200					sample size =400				
	λ	β_1	β_2	α	β	λ	β_1	β_2	α	β
Scenario (1)										
ratio	1.076	1.057	1.059	1.077	1.093	1.099	1.044	1.053	1.136	1.121
Scenario (2)										
ratio	1.112	1.058	1.071	1.034	1.061	1.115	1.057	1.067	1.054	1.068

5.4 Analysis of pubertal weight growth data

In a longitudinal study of pubertal growth and blood pressure regulation, school children aged from 5 to 17 were recruited for longitudinal assessment of somatic growth and blood pressure. At each assessment, somatic growth measures were taken and recorded. The detailed study protocol was described in Tu et al. (2009, 2014). For each study participant, the investigators identified the interval that showed the greatest rate of height increase and used it as the interval containing the pubertal growth spurt (PGS) (Shankar et al., 2005).

An overarching objective of this research was to quantify the weight changes around the time of PGS, with the goal of improving the existing understanding of the adolescent growth. We focused on growth rates around of the time of PGS, because they are thought to set the trajectory of adolescent development into the adulthood. The outcome of interest in this particular analysis was weight, which is one of the the primary markers of body development. Specifically, we attempted to compare: (1) the pre and post-PGS rates of weight increase, (2) the average weights at the time of PGS between races, and (3) the race difference in weight increase rates around the time of PGS. Because of the known differences in pubertal growth patterns between boys and girls, analyses often proceed in sex-specific groups. The current analysis was based on data from 188 boys.

The peak growth intervals, i.e., the censoring intervals containing the unobserved PGS, are presented in the left panel of Figure 5.1. Weights measured at the endpoints of the censoring intervals are depicted in the right panel of Figure 5.1. The pre and post-PGS weight measures from the same individuals are connected by line segments.

To analyze, we considered the following piece-wise linear mixed-effects model with random intercepts and and post-PGS growth rates. We did not include random pre-PGS slopes because existing literature suggests that heterogeneity in rate of weight increase started at the PGS.

$$\begin{cases} Y_{L,i} = \lambda_1 + \alpha_1(L_i - T_i) + \big(\lambda_2 + \alpha_2(L_i - T_i)\big) I_w + \lambda_i + \epsilon_{L,i} \\ Y_{R,i} = \lambda_1 + \beta_1(R_i - T_i) + \big(\lambda_2 + \beta_2(R_i - T_i)\big) I_w + \lambda_i + \beta_i(R_i - T_i) + \epsilon_{R,i} \end{cases} \tag{5.9}$$

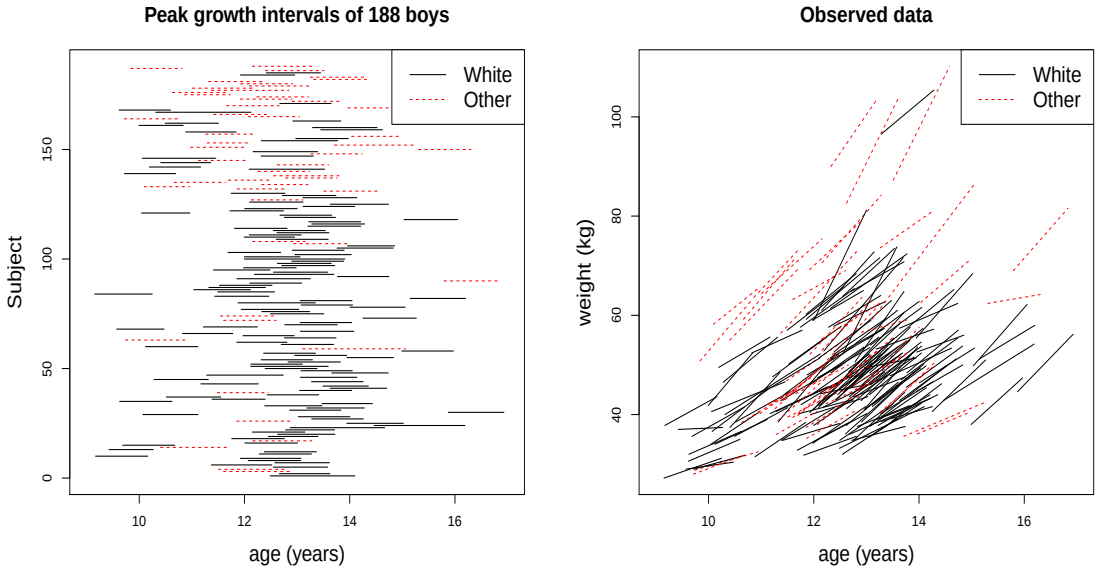


Figure 5.1: Peak growth intervals and observed weight

where $(L_i, R_i]$ was the censoring interval for the i-th subject; T_i was the unobserved PGS time; $Y_{L,i}$ and $Y_{R,i}$ were the respective weights measured at L_i and R_i; I_w was an indicator for being whites; λ_1 was the average weight at PGS in non-whites; and α_1 and β_1 were the respectively average pre and post-PGS weight growth rates in non-whites. Similarly, λ_2, α_2 and β_2 respectively represented the differences in average value, pre and post-PGS weight growth rates between non-whites and whites. Here λ_i and β_i were the random intercept and slope; and $\epsilon_{L,i}$ and $\epsilon_{R,i}$ were the random errors.

All regularity conditions were satisfied in this data application. For example, Condition (F2) was satisfied, as Figure 5.1 did not show substantial differences in PGS distributions between whites and non-whites. So we applied the proposed functional estimation method to fit Model (5.9). The parameter estimates $\hat{\boldsymbol{\theta}}_n = (\hat{\lambda}_{n,1}, \hat{\lambda}_{n,2}, \hat{\alpha}_{n,1}, \hat{\beta}_{n,1}, \hat{\alpha}_{n,2}, \hat{\beta}_{n,2})$ are summarized in Table 5.9.

Table 5.9: Parameter estimates

	λ_1	λ_2	α_1	β_1	α_2	β_2
estimate	54.478	-6.927	7.526	9.730	-1.583	-0.543
se	1.901	1.992	0.811	1.005	0.971	1.256

The following covariance matrix $\hat{\boldsymbol{\Sigma}}_n$ of the parameter estimates was estimated using bootstrap resampling method with 100 resamples.

$$\hat{\boldsymbol{\Sigma}}_n = \begin{bmatrix} 3.613 & -3.409 & 0.278 & 0.769 & -0.329 & -0.659 \\ -3.409 & 3.968 & -0.117 & -0.879 & 0.334 & 0.733 \\ 0.278 & -0.117 & 0.658 & -0.442 & -0.671 & 0.491 \\ 0.769 & -0.879 & -0.442 & 1.009 & 0.450 & -1.004 \\ -0.329 & 0.334 & -0.671 & 0.450 & 0.942 & -0.741 \\ -0.659 & 0.733 & 0.491 & -1.004 & -0.741 & 1.577 \end{bmatrix}.$$

We then proceeded to make inferences on parameters of interest along the lines laid out by Theorem 5.1.2. First, the pre and post-PGS rates of weight increase can be compared by testing hypothesis that $\alpha_1 = \beta_1$ and $\alpha_2 = \beta_2$ in all ethnic groups. Using contrast

$$\boldsymbol{M} = \begin{pmatrix} 0 & 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 \end{pmatrix},$$

the test statistic $(\boldsymbol{M}\hat{\boldsymbol{\theta}}_n)^t(\boldsymbol{M}\hat{\boldsymbol{\Sigma}}_n\boldsymbol{M}^t)^{-1}(\boldsymbol{M}\hat{\boldsymbol{\theta}}_n)$ was approximately $\chi^2(2)$-distributed under the null hypothesis. The observed value of this statistic was calculated to be 10.150, which resulted in a p-value of 0.006 for a χ^2-test. So we had evidence that rate of weight growth was greater in the post-PGS period.

To compare the average weight at PGS between individuals in different ethnic groups, we simply test the hypothesis that $\lambda_2 = 0$. Under the null hypothesis, the Z-score of $\hat{\lambda}_2$ was calculated as $-6.927/1.992 = -3.477$, which resulted in a p-value of 0.0005 for the two-sided Z-test. So we concluded that whites and ethnic minority children had different weights at PGS, with whites had lower weights.

Finally, to compare the rates of weight increase around PGS between individuals of different ethnic groups, we tested the hypothesis $\alpha_2 = \beta_2 = 0$. To implement, we calculated the value of the statistic $(\hat{\alpha}_{n,2}, \hat{\beta}_{n,2}) \cdot \hat{\mathbf{\Sigma}}_n'^{-1} \cdot (\hat{\alpha}_{n,2}, \hat{\beta}_{n,2})^t$, which can be approximated by a $\chi^2(2)$-distribution under the null hypothesis. Here matrix $\hat{\mathbf{\Sigma}}_n'$ is the covariance matrix of $\hat{\alpha}_{n,2}$ and $\hat{\beta}_{n,2}$, which was a submatrix of $\hat{\mathbf{\Sigma}}_n$. The observed value of this statistic was calculated to be 5.877, which led to a p-value of 0.053. So we conclude that there was some indication that whites had slower rate of weight gain around PGS, although the difference did not meet the threshold of 0.05 to be statistically significant. We note that all findings were consistent with the observed data shown in Figure 5.1.

Through this real data example, we demonstrated the operations of parameter estimation and statistical inference using the new method, which handled interval censored PGS times nicely in this application. While we note that the findings are largely consistent with the existing theory of human growth (Hall, 2006), no studies to the best of our knowledge have actually quantified the growth parameters around PGS, because of the traditional longitudinal models' inability to accommodate the unobserved anchoring points, unless strong and unverifiable parametric assumption on the PGS distribution is imposed, which could produce biased estimates.

Chapter 6

An R function for fitting linear mixed-effects model with interval censored anchoring points

In this chapter we provide a user-friendly R function that implements the hybrid algorithm proposed in Chapter 5. We follow the two-stage estimation procedure.

6.1 Nonparametric maximum likelihood estimation of the anchoring point distribution

Although our research focuses on the situation that F_0, the true distribution of the anchoring points, does not depend on covariates, we present here a function that computes the baseline hazard $\Lambda_0(t)$ and the parameters $\boldsymbol{\beta}$, for the situation that the anchoring points follow a proportional hazards model, i.e.,

$$\Lambda(t|\boldsymbol{Z}) = \exp(\boldsymbol{\beta}^t \boldsymbol{Z})\Lambda_0(t)$$

where $\boldsymbol{Z}$ is the covariate that determines the hazard ratio for a subject. We estimate $\Lambda_0(t)$ as a step function. The function is attached below.

```
PHregIC<-function(L, R, Z=NULL,
                  bndry.ctrl.alpha=(1e-10), NR.ctrl=1e-5,
                  CumHazard.max=1e200,  print.hist.phreg=TRUE)
{tryCatch({
    time.start<-proc.time()
    N<-length(R)
    if(is.null(Z)){Z<-matrix(rep(0, N), ncol=1)}

#### Find S over which \Lambda_0 has jumps
```

```
  df<-data.frame(c(L,R), rep(c(0, 1), each=N))
  df<-df[order(df[,1], -df[2]),]
  chs.temp<-df[,2]
  chs<-chs.temp-c(0, chs.temp[-length(chs.temp)])
  S<-df[,1][chs==1];   N.alphas<-length(S)

  X.L<-outer(L, S, ">=");  X.R<-outer(R, S, ">=");

#### initial setup
  beta<-rep(0, dim(Z)[2]);  alpha<-rep(1/N.alphas, N.alphas);
  finished.profile<-N;   iter.profile<-0;

  while(finished.profile>1 & iter.profile<1000){
        iter.profile<-iter.profile+1;

  #Given beta, find Lambda_0:
  if(print.hist.phreg==TRUE)
    {cat("\n\n\n profile step-", iter.profile,
         ":****** Fix beta, find alpha ******\n")}
  weight<-as.vector(exp(Z%*%beta))
  coef.L<-X.L*weight
  coef.R<-X.R*weight
  sum.coef.L<-matrix(colSums(coef.L), ncol=1)
  coef.L.minus.R<-coef.L-coef.R

  Surv.L<-exp(-coef.L%*%alpha)
  Surv.R<-exp(-coef.R%*%alpha)
  NewLogLik<-sum(log(Surv.L-Surv.R ))
  U.score<-matrix(-colSums(as.vector(Surv.R/(Surv.L-Surv.R))*
                  coef.L.minus.R),  ncol=1)-sum.coef.L
  Hessian.part.L<-as.vector(Surv.L/(Surv.L-Surv.R))*coef.L.minus.R
  Hessian.part.R<-as.vector(Surv.R/(Surv.L-Surv.R))*coef.L.minus.R
  Hessian<-t(Hessian.part.L)%*%Hessian.part.R

  ctrl.alpha<-max(abs(U.score[  alpha>=0|U.score>0  ]));
  iter.alpha<-0;
  while(ctrl.alpha>NR.ctrl & iter.alpha<100)
  {   if(print.hist.phreg==TRUE)
      {cat("  iter.alpha=", iter.alpha, ", LogL=", NewLogLik,
           ", ctrl.alpha=", ctrl.alpha, "\n", sep="")}
    on.bndry<-(((alpha<bndry.ctrl.alpha)&(U.score<0))|
               (abs(U.score)<(NR.ctrl/2)));
    if(prod(on.bndry)==0){   alwd.step<-0;
      while(alwd.step < (1e-100)  &  iter.alpha<100 )
      {  iter.alpha<-iter.alpha+1
```

```
    alwd<-(1-on.bndry)
    alwd.alpha<-alpha[alwd==1]
    alwd.U<-U.score[alwd==1]
    alwd.Hessian<-Hessian[alwd==1, alwd==1]
    alwd.vec.temp<-solve(alwd.Hessian, alwd.U)
    if(sum(alwd)>1){
      alwd.vec.temp<-solve(alwd.Hessian/diag(alwd.Hessian),
                           alwd.U/diag(alwd.Hessian))
                  }else{alwd.vec.temp<-alwd.U/alwd.Hessian}
    cond<-(-alwd.alpha[alwd.vec.temp<0]/
            alwd.vec.temp[alwd.vec.temp<0])
    alwd.step<-ifelse(length(cond)==0, 1, min(c(cond,1)))
    new.on.bndry.temp<-(alwd.alpha<bndry.ctrl.alpha)*
                       (alwd.vec.temp<0)
    on.bndry[on.bndry==0]<-new.on.bndry.temp  }# End of while

    alwd.vec<-rep(0, length(alpha))
    alwd.vec[alwd==1]<-alwd.vec.temp
    while(sum(alpha+alwd.step*alwd.vec) >= CumHazard.max)
         {alwd.step=alwd.step/2}
    while(min(alpha+alwd.step*alwd.vec) < 0)
         {alwd.step=alwd.step*0.9999}

    alpha<-alpha+alwd.step*alwd.vec
    Surv.L<-exp(-coef.L%*%alpha);  Surv.R<-exp(-coef.R%*%alpha)
    NewLogLik<-sum(log(Surv.L-Surv.R ))
    U.score<-matrix(-colSums(as.vector(Surv.R/(Surv.L-Surv.R))*
                     coef.L.minus.R), ncol=1)-sum.coef.L
    Hessian.part.L<-as.vector(Surv.L/(Surv.L-Surv.R))*
                               coef.L.minus.R
    Hessian.part.R<-as.vector(Surv.R/(Surv.L-Surv.R))*
                               coef.L.minus.R
    Hessian<-t(Hessian.part.L)%*%Hessian.part.R
    on.bndry<-(alpha<bndry.ctrl.alpha)*(U.score<0)
    alwd.U<-U.score[on.bndry==0]
    ctrl.alpha=max(abs(alwd.U))}else{ctrl.alpha<-0;}
}# End of while
if(print.hist.phreg==TRUE)
{  cat("  iter.alpha=", iter.alpha, "LogL=", NewLogLik,
       "ctrl.alpha=", ctrl.alpha, "\n")   }

#Given Lambda_0, find beta:
if(print.hist.phreg==TRUE)
{  cat("\n profile step-", iter.profile,
       ":****** Fix alpha, find beta ******\n")}
```

```
A.i.tmp<-(-X.L%*%alpha); B.i.tmp<-(-X.R%*%alpha);
A.i<-A.i.tmp*exp(Z%*%beta);  B.i<-B.i.tmp*exp(Z%*%beta)
den.4.UZ<-(exp(A.i)-exp(B.i));  NewLogLik<-sum(log(den.4.UZ))
UZ<-matrix(colSums(as.vector((exp(A.i)*A.i-exp(B.i)*B.i)/
                  den.4.UZ)*Z),  ncol=1)
num.4.HZ<-exp(A.i+B.i)*((A.i-B.i)^2+A.i+B.i)-
          exp(2*A.i)*A.i-exp(2*B.i)*B.i
num.4.HZ[num.4.HZ<0]<-0
scaled.Z.4.HZ<-as.vector(sqrt(num.4.HZ)/den.4.UZ)*Z
HZ<-t(scaled.Z.4.HZ)%*%scaled.Z.4.HZ

ctrl.beta<-max(abs(UZ));   iter.beta<-0;
while(ctrl.beta>NR.ctrl & iter.beta<100){
  if(print.hist.phreg==TRUE)
  {   cat("  iter.beta=", iter.beta, ", LogL=", NewLogLik,
          ", ctrl.beta=", ctrl.beta, "\n", sep="")}
  iter.beta<-iter.beta+1
  beta<-beta+solve(HZ,UZ)
  A.i<-A.i.tmp*exp(Z%*%beta);   B.i<-B.i.tmp*exp(Z%*%beta)
  den.4.UZ<-(exp(A.i)-exp(B.i));  NewLogLik<-sum(log(den.4.UZ))
  UZ<-matrix(colSums(as.vector((exp(A.i)*A.i-exp(B.i)*B.i)/
             den.4.UZ)*Z), ncol=1)
  num.4.HZ<-exp(A.i+B.i)*((A.i-B.i)^2+A.i+B.i)-
            exp(2*A.i)*A.i-exp(2*B.i)*B.i
  num.4.HZ[num.4.HZ<0]<-0
  scaled.Z.4.HZ<-as.vector(sqrt(num.4.HZ)/den.4.UZ)*Z
  HZ<-t(scaled.Z.4.HZ)%*%scaled.Z.4.HZ
  ctrl.beta=max(abs(UZ))
}
if(print.hist.phreg==TRUE)
{ cat("  iter.beta=", iter.beta, ", LogL=", NewLogLik,
      ", ctrl.beta=", ctrl.beta, "\n", sep="")  }

finished.profile<-iter.beta+iter.alpha
}

time.stop<-proc.time()
new.S<-S[alpha>bndry.ctrl.alpha]
new.hr<-alpha[alpha>bndry.ctrl.alpha]

return(list("beta"=beta,
            "S"=new.S,
            "hr"=new.hr,
            "weight"=weight,
            "domain"=c(min(S), max(S)),
```

```
              "used.time"=time.stop-time.start,
              "control"=ctrl.beta+ctrl.alpha))
}, error=function(e){cat("ERROR :", conditionMessage(e), "\n")}
) #### end of tryCatch
}
```

The function "PHregIC()" has the following arguments.

1. L, R

   The vector of left endpoints and right endpoints of the censoring intervals. The i-th entry of L and the i-th entry of R are the endpoints from the i-th censoring interval.

2. Z

   The matrix of covariates. The i-th row of Z are the covariates of the i-th subject. If F_0 does not depend on any covariates, such as in the current research, then Z does not need to be specified.

3. bndry.ctrl.alpha

   Λ_0 is computed as a step function that has jumps over a finite set S. At some $s \in S$, Λ_0 may have no jump, but for numerical stability, the function regard a jump that has value $\leq$bndry.ctrl.alpha as no jump. The default value is 1e-10.

4. NR.ctrl

   A small quantity that controls the accuracy of the numeric solution. Smaller value results in more accurate solution. The default value is 1e-5.

5. CumHazard.max

   For numeric stability, the function forces $\Lambda_0 \leq$CumHazard.max. The default value is 1e200.

6. print.hist.phreg An option for user to investigate details for each iteration. The default value is TRUE.

The function "PHregIC()" returns a list, which contains

1. beta

   The vector of $\boldsymbol{\beta}$ in the proportional hazard model. If F_0 does not depend on any covariate, then the value of beta is the constant 0.

2. S, hr

   S is the vector of points over which Λ_0 has a jump. hr is the vector of jumps of Λ_0. The i-th entry of hr is the jump at the i-th entry of S.

3. weight

   A useful quantity for the use in this research. It computes $\exp(\boldsymbol{\beta}^t Z)$ for each subject.

4. control

   A value for user to check wether the algorithm terminated with convergence. If the convergence is achieved, this value should be no larger than NR.ctrl. Otherwise, the algorithm did not terminate with convergence, and the estimate is not reliable.

6.2 A hybrid algorithm combining Fisher-Scoring algorithm and EM-algorithm

The following two functions decompose a symmetric matrix $\boldsymbol{M}$ as $\boldsymbol{M} = \boldsymbol{A}\boldsymbol{A}^t$ for a lower triangular matrix $\boldsymbol{A}$. The first function "LL.decomp()" returns $\boldsymbol{A}$ as a vector, and the second function "vec.to.lower.mat()" changes the vector into a lower triangular matrix. These two functions are used in the main function to be explained in this section.

```
LL.decomp<-function(M){
  if(length(M)>1)
    {  n<-dim(M)[1]
       if(M[1,1]<(1e-5)){ A.vec<-c(c(1e-5, rep(0, n-1)),
                                   LL.decomp(M[-1,-1]))
       }else{A.vec<-c(M[1,]/sqrt(M[1,1]),
                  LL.decomp(M[-1,-1]-M[2:n,1]%*%t(M[1,2:n])/M[1,1]))}
    }else{ A.vec<-sqrt(max(M,0)) }
return(A.vec) }

vec.to.lower.mat<-function(vec){
  n<-floor(sqrt(2*length(vec)))
  if(n>1){  M.vec<-c(outer(1:n, 1:n, "-"))
            M<-rep(0, n^2); M[M.vec >=0]<-vec
            return( matrix(M, nrow=n) )}else{return(vec)}  }
```

The following function computes linear mixed-effects model parameters, as described in Chapter 5.

```
get.para.EMFS<-function(lst, weight, n.fixed, n.random, initial.G,
                        outlier.control=(1e-50), initial.sigma.sq=1,
                        max.iter=1000, accuracy=0.00001,
                        min.step.ctrl=(2e-4), initial.EM.ctrl=0.1,
                        EM.acc=2,  EM.max.iter=1000, GroupEM.num=25,
                        print.hist.lme=TRUE, plot.hist=TRUE,
                        plot.index=c(1), plot.y=0.2, plot.x=250,
                        plot.pch=".", plot.cex=3, plot.title="")
{tryCatch({
    n.tmp<-length(lst); sum.weight<-sum(weight);
    total.obs<-sum(sapply(1:n.tmp,
                      function(i.tmp) lst[[i.tmp]]$nobs.i)*weight)

    ## get initial values of beta by using LSE method.
    FindXpXXpY<-function(i.tmp)
      {EX.i<-Reduce("+", lapply(1:length(lst[[i.tmp]]$X.ij),
               function(j.tmp) lst[[i.tmp]]$X.ij[[j.tmp]]*
                                lst[[i.tmp]]$Fhat.i[j.tmp,2]))
         return(t(EX.i)%*%cbind(EX.i, lst[[i.tmp]]$Y.i))   }
    XpX.XpY<-Reduce("+", lapply(1:n.tmp, function(i.tmp)
                         FindXpXXpY(i.tmp)*weight[i.tmp]))
    beta.lse<-solve(XpX.XpY[,-n.fixed-1], XpX.XpY[,n.fixed+1])

#### get initial value of all parameters
    all.para.initial<-c(beta.lse, initial.sigma.sq,
```

```
                        LL.decomp(initial.G))
    A.pattern<-LL.decomp(diag(1, n.random))
    need2check.A<-(1:length(A.pattern))[A.pattern==1]
    need2check<-n.fixed+1+c(0, need2check.A)
    select.lower<-as.vector(outer(1:n.random, 1:n.random, ">="))

#### function to compute material for observation i evaluated at s_j
get.mtrl.ij<-function(i.ttmp=1, beta.ttmp,
                      sigma.sq.ttmp, G.ttmp, j.ttmp)
{     V.ij.inv<-solve(diag(sigma.sq.ttmp,  lst[[i.ttmp]]$nobs.i)+
                      (lst[[i.ttmp]]$Z.ij[[j.ttmp]])%*%G.ttmp%*%
                          t(lst[[i.ttmp]]$Z.ij[[j.ttmp]]))
      Vec.ij<-(lst[[i.ttmp]]$Y.i)-
              (lst[[i.ttmp]]$X.ij[[j.ttmp]])%*%beta.ttmp
      V.inv.Vec.ij=V.ij.inv%*%Vec.ij

return(list("tr.V.ij.inv"=sum(diag(V.ij.inv)),
            "V.inv.Vec.ij"=V.inv.Vec.ij,
            "mat.for.G.ij"=t(lst[[i.ttmp]]$Z.ij[[j.ttmp]])%*%
                  (V.inv.Vec.ij%*%t(V.inv.Vec.ij)-V.ij.inv)%*%
                  (lst[[i.ttmp]]$Z.ij[[j.ttmp]]),
            "newp.ij.unscaled"=(lst[[i.ttmp]]$Fhat.i[j.ttmp, 2])*
                               sqrt(det(V.ij.inv))/exp( t(Vec.ij)%*%
                               V.ij.inv%*%Vec.ij/2 )  ))
}##End of get.mtrl.ij()

#### EM-step
NextStep.EM<-function(all.para.tmp=all.para.initial)
{ beta.tmp<-all.para.tmp[1:n.fixed]
  sigma.sq.tmp<-all.para.tmp[1+n.fixed]
  G.sqrt.tmp<-vec.to.lower.mat(
    all.para.tmp[(2+n.fixed):length(all.para.tmp)]  )
    G.tmp<-G.sqrt.tmp%*%t(G.sqrt.tmp)

    newlist<-lapply(1:n.tmp,
             function(i){lapply(1:length(lst[[i]]$X.ij),
                         function(j) {get.mtrl.ij(i.ttmp=i,
                         beta.ttmp=beta.tmp,
                         sigma.sq.ttmp=sigma.sq.tmp,
                         G.ttmp=G.tmp,
                         j.ttmp=j)})})
             lkhd.i<-sapply(1:n.tmp,
             function(i){sum(sapply(1:length(newlist[[i]]),
                function(j){newlist[[i]][[j]]$newp.ij.unscaled}))})
```

```
     possible.outliers<-(1:length(lkhd.i))[lkhd.i<outlier.control]
     if(length(possible.outliers)>0)
       { cat("There are possible outliers,
              recommend to remove subjects:\n")
         print(possible.outliers) }

     newp.ij<-lapply(1:n.tmp,
            function(i){lapply(1:length(newlist[[i]]),
function(j){as.numeric(newlist[[i]][[j]]$newp.ij.unscaled)/
                        lkhd.i[i]})})

               ####compute logl
               logl<-sum(weight*log(lkhd.i))

               ####compute G.next
               mat4G<-Reduce("+", lapply(1:n.tmp,
function(i) {weight[i]*Reduce("+", lapply(1:length(newlist[[i]]),
function(j) {newlist[[i]][[j]]$mat.for.G.ij*
               newp.ij[[i]][[j]]} ))}))
               delta.G<-G.tmp%*%mat4G%*%G.tmp/sum.weight
               G.next<-G.tmp+delta.G; A.next<-LL.decomp(G.next)

               ####compute beta.next
               mat4beta<-Reduce("+", lapply(1:n.tmp,
function(i) weight[i]*Reduce("+", lapply(1:length(newlist[[i]]),
   function(j) t(lst[[i]]$X.ij[[j]])%*%cbind(lst[[i]]$X.ij[[j]],
                   newlist[[i]][[j]]$V.inv.Vec.ij)*
                    newp.ij[[i]][[j]]))))
               Delta.tmp<-solve(mat4beta[,-n.fixed-1],
                                mat4beta[,n.fixed+1])
               delta.beta<-Delta.tmp*sigma.sq.tmp
               beta.next<-beta.tmp+delta.beta

               ####compute sigma.sq.next
               min.C<-sum(sapply(1:n.tmp,
function(i) weight[i]*sum(sapply(1:length(newlist[[i]]),
   function(j) sum((newlist[[i]][[j]]$V.inv.Vec.ij-
           lst[[i]]$X.ij[[j]]%*%Delta.tmp)^2)*newp.ij[[i]][[j]]))))
               ETr<-sum(sapply(1:n.tmp,
function(i) weight[i]*sum(sapply(1:length(newlist[[i]]),
   function(j) sum((newlist[[i]][[j]]$tr.V.ij.inv)*
                     newp.ij[[i]][[j]])))))
               delta.sigma.sq<-sigma.sq.tmp^2*(min.C-ETr)/total.obs
               sigma.sq.next<-max(0.000001, sigma.sq.tmp + delta.sigma.sq)
```

```
    return(list("LogL"=logl,
                "next.para"=c(beta.next, sigma.sq.next, A.next),
                "change"=sum(abs(delta.beta)/abs(beta.tmp))+
                             sum(abs(delta.beta))+
                             sum(abs(c(delta.sigma.sq, delta.G)))/
                             (1+n.random))) }### End of NextStep.EM()

#### FS step
NextStep.FS<-function(all.para.tmp=all.para.initial)
  {   beta.tmp<-all.para.tmp[1:n.fixed]
      sigma.sq.tmp<-all.para.tmp[1+n.fixed]
      G.sqrt.tmp<-vec.to.lower.mat(
                  all.para.tmp[(2+n.fixed):length(all.para.tmp)])
      G.tmp<-G.sqrt.tmp%*%t(G.sqrt.tmp)

      newlist<-lapply(1:n.tmp,
function(i) {lapply(1:length(lst[[i]]$X.ij),
   function(j) {get.mtrl.ij(i.ttmp=i,
        beta.ttmp=beta.tmp,
        sigma.sq.ttmp=sigma.sq.tmp,
        G.ttmp=G.tmp,
        j.ttmp=j)})})

      lkhd.i<-sapply(1:n.tmp,
function(i) {sum(sapply(1:length(newlist[[i]]),
   function(j) {newlist[[i]][[j]]$newp.ij.unscaled}))})

      newp.ij<-lapply(1:n.tmp,
function(i) {lapply(1:length(newlist[[i]]),
   function(j){as.numeric(newlist[[i]][[j]]$newp.ij.unscaled)/
                          lkhd.i[i]})})

      ####compute logl
      logl<-sum(weight*log(lkhd.i))

      #### compute U
      U<-rbind( ####compute U.beta
                Ubeta=sapply(1:n.tmp,
function(i) Reduce("+", lapply(1:length(newlist[[i]]),
   function(j) t(lst[[i]]$X.ij[[j]])%*%
                 newlist[[i]][[j]]$V.inv.Vec.ij*
                 newp.ij[[i]][[j]]))),

      ####compute U.sigma.sq
      Usigma.sq=sapply(1:n.tmp,
```

```
function(i) sum(sapply(1:length(newlist[[i]]),
   function(j)(sum(newlist[[i]][[j]]$V.inv.Vec.ij^2)-
                   newlist[[i]][[j]]$tr.V.ij.inv)*
                   newp.ij[[i]][[j]]))/2),

      ####compute U.A
      UA=sapply(1:n.tmp,
function(i) c(Reduce("+", lapply(1:length(newlist[[i]]),
   function(j) newlist[[i]][[j]]$mat.for.G.ij*
               newp.ij[[i]][[j]]))%*%
               G.sqrt.tmp)[select.lower] )
   ) #### End of U<-rbind(Ubeta=, Usigma.sq, UA)

      return(list("LogL"=logl,
                  "Score"=colSums(weight*t(U)),
                  "Fisher.I"= t(weight*t(U))%*%t(U) ))
} ##End of function NextStep.FS()


#### intialized and run EM for several times
all.para<-all.para.initial
used.EM<-0; EM.change<-1;
    while( (used.EM<GroupEM.num | EM.change>initial.EM.ctrl)&
           (used.EM < EM.max.iter) )
      {   used.EM<-used.EM+1;
          Next.EM<-NextStep.EM(all.para);
          all.para<-Next.EM$next.para;
          EM.change<-Next.EM$change       }

    #### Start Fisher.Scoring-EM mixture algorithm
  if(plot.hist==TRUE)
    {  plot.adj<-all.para[plot.index]; n2plot<-length(plot.index)
       plot(NULL, xlim=c(0, plot.x), ylim=c(-plot.y, plot.y),
            xlab="FS.iter", ylab="", main=plot.title)  }

    FS.infor<-NextStep.FS(all.para)
    LogL.current<-FS.infor$LogL

    n.all.para<-length(all.para); iter<-0; ctrl.FS<-1;
    min.step=ifelse(n.tmp<400, n.tmp/400, 1);
    target.min.step=1; used.FS<-0;
while(iter<max.iter & ctrl.FS>accuracy & min.step>min.step.ctrl){
       iter=iter+1
#### remove paras that are on boundary.
 on.bndry<-rep(0, n.all.para)
```

```
 next.direction<-rep(0, n.all.para)
 for(i.tmp in need2check){on.bndry[i.tmp]=(all.para[i.tmp]<(1e-5)&
                          FS.infor$Score[i.tmp]<(1e-10))}
 Fisher.I<-FS.infor$Fisher.I

 if(min(eigen(Fisher.I[on.bndry==0,on.bndry==0])$values)<(1e-8)){
       cat("\n EM used before FS. \n")
       target.min.step=min.step*0.9
       min.step=min.step/2
       used.EM<-used.EM+GroupEM.num; used.FS<-0;
       for(dummy in 1:GroupEM.num){
           Next.EM<-NextStep.EM(all.para)
           all.para<-all.para+EM.acc*(Next.EM$next.para-all.para) }
       FS.infor<-NextStep.FS(all.para)
       LogL.current<-FS.infor$LogL
       ctrl.FS<-Next.EM$change+abs(LogL.current-Next.EM$LogL)
 }else{
 next.direction[on.bndry==0]<-solve(Fisher.I[on.bndry==0,on.bndry==0],
                                    FS.infor$Score[on.bndry==0])
        need.to.check=c(1,need2check)
        while(length(need.to.check) > 1)
          {   find.bndry<-0
              for(i.tmp in need.to.check[-1])
                {if(all.para[i.tmp]<(1e-8)&next.direction[i.tmp]<0&
                     on.bndry[i.tmp]<1)
                    {on.bndry[i.tmp]<-1; next.direction[i.tmp]<-0;
                      find.bndry<-find.bndry+1 }   }
              if(find.bndry>1){
next.direction[on.bndry==0]<-solve(Fisher.I[on.bndry==0,on.bndry==0],
                                    FS.infor$Score[on.bndry==0])
              }else{need.to.check<-c(1)}
          }#### End of While(length(need.to.check) > 1)

next.step<-min(min.step, sapply(need2check,
function(i.tmp) ifelse(next.direction[i.tmp]>(-1e-100), 1,
                       abs(all.para[i.tmp]/next.direction[i.tmp]))))

tmp.move<-next.step*next.direction;
tmp.all.para<-all.para+tmp.move;
FS.infor<-NextStep.FS(tmp.all.para)

if(FS.infor$LogL>(LogL.current-(1e-10)))
  {ctrl.FS<-sum(abs(c(tmp.move[1:n.fixed],
                      next.direction/sum(1-on.bndry),
                      tmp.move[(1+n.fixed):n.all.para]/(1+n.random),
```

```
                      FS.infor$LogL-LogL.current)))
   used.FS<-used.FS+1;
   LogL.current<-FS.infor$Log
   all.para<-tmp.all.para
   min.step=min.step+(target.min.step-min.step)/10
   if(used.FS>10){target.min.step=target.min.step+
                     min(10*min.step, 1-target.min.step)/(used.FS) }
      }else{ cat("\n EM used. \n")
             target.min.step=min.step*0.9
             min.step=min.step/2
             used.EM<-used.EM+GroupEM.num; used.FS<-0;
             for(dummy in 1:GroupEM.num){
                 Next.EM<-NextStep.EM(all.para)
                 all.para<-all.para+
                           EM.acc*(Next.EM$next.para-all.para)}
             FS.infor<-NextStep.FS(all.para)
             LogL.current<-FS.infor$LogL
             ctrl.FS<-Next.EM$change+abs(LogL.current-Next.EM$LogL)
                }#End of if-else
    }#End of if=else

 if(print.hist.lme==TRUE) {
   cat("\n Iter=",round(iter, 1),". min.step=", min.step,
       ". Ctrl.FS=", round(ctrl.FS, 6),
       ". LogL=", LogL.current, ".   Sigma.sq=",
       round(all.para[1+n.fixed],6),"\n fixed.effects(lse lme)=:",
       "\n", sep="")
    print(rbind(beta.lse=beta.lse, beta.lme=all.para[1:n.fixed]))
    A.tmp<-vec.to.lower.mat(all.para[(2+n.fixed):n.all.para])
    cat("\n G-matrix is: \n");
    print(A.tmp%*%t(A.tmp))   }

    if(plot.hist==TRUE){
        points(x=rep(iter-floor(iter/plot.x)*plot.x, n2plot),
               y=all.para[plot.index]-plot.adj,
               pch=plot.pch, cex=plot.cex, col=plot.index)    }
 } #### End of while()



#### Control the convergence with EM
new.iter<-0;
Next.EM<-NextStep.EM(all.para)
ctrl<-abs(Next.EM$LogL-LogL.current)+Next.EM$change
```

```
if(ctrl>10*accuracy){   cat("\n \n Finalize with EM. \n \n ")
      acc=c(1, EM.acc, EM.acc, EM.acc)
      while(ctrl>10*accuracy & new.iter<EM.max.iter)
      { new.iter <- new.iter +1;
        need.acc<-new.iter-floor(new.iter/4)*4+1
        Next.EM<-NextStep.EM(all.para)
        ctrl<-abs(Next.EM$LogL-LogL.current)+Next.EM$change

        all.para<-all.para+(Next.EM$next.para-all.para)*acc[need.acc]
        LogL.current<-Next.EM$LogL;

 if(print.hist.lme==TRUE) {
   cat("\n new.iter=", round(new.iter, 1),
       ".   Ctrl=", round(ctrl, 6),
       ".    LogL=", LogL.current,
       ".    Sigma.sq=", round(all.para[1+n.fixed], 6),
       "\n", sep="")
   print(c(beta.lse, all.para[1:n.fixed]))
   A.tmp<-vec.to.lower.mat(all.para[(2+n.fixed):n.all.para])
   print(A.tmp%*%t(A.tmp))   }

 if(plot.hist==TRUE) {
    points(x=rep(iter-floor(iter/plot.x)*plot.x, n2plot),
           y=all.para[plot.index]-plot.adj,
           pch=plot.pch, cex=plot.cex, col=plot.index) }
      }#### End of while.
  }####End of if(ctrl>10*accuracy)

    cat("ctrl.final=", ctrl,
        ". total iteration:",iter+new.iter+used.EM,
        "\n", sep="")

    return(c(para.LSE=beta.lse,
             LogL=LogL.current,
             all.para=all.para,
             iter=iter+new.iter+used.EM))

}, error=function(e){cat("ERROR :", conditionMessage(e), "\n")}
) #### end of tryCatch
} #### End of function   get.para.EMFS()
```

Before calling the function "get.para.EMFS()", the anchoring point distribution F_0 needs to be obtained as a step function, such as computed by the "PHregIC()"

function in Section 6.1. For the i-th subject, assume the censoring interval is $(L.i, R.i]$, and assume $T\big|(L_i < T \leq R_i)$ has jumps p_{ij} at s_{ij}, where $j = 1, \cdots, k_i$. Let $S.i$ denote the vector of s_{ij}'s and let $P.i$ denote the vector of the p_{ij}'s. The function "get.para.EMFS()" has the following arguments.

1. lst

   A list of list. The i-th element of lst is a list that contains the following information of the i-th subject: list(intvl, Fhat.i, nobs.i, Y.i, X.ij, Z.ij), where

   (a) intvl=c(L.i, R.i) the censoring interval.
   (b) Fhat.i=cbind(S.i, P.i)
   (c) nobs.i= number of observations
   (d) Y.i= the column matrix of response
   (e) X.ij, Z.ij = lists of matrices, the j-th matrices of X.ij and Z.ij are the respective fixed and random effects matrices when the event time is assumed to be at $T = s_{ij}$

2. weight

   The weight for each subject, if specified. Default weight is 1 for each subject.

3. n.fixed, n.random

   number of parameters in fixed and random effects.

4. initial.G

   initial value for the random effect matrices G. Default value is the identity matrix.

5. outlier.control=(1e-50), initial.sigma.sq=1, max.iter=1000, accuracy=0.00001, min.step.ctrl=(2e-4), initial.EM.ctrl=0.1, EM.acc=2, EM.max.iter=1000,

GroupEM.num=25, print.hist.lme=TRUE, plot.hist=TRUE, plot.index=c(1), plot.y=0.2, plot.x=250, plot.pch=".", plot.cex=3, plot.title=""

These are optional arguments to control the convergence of the algorithm, and obtain details of the updated parameter estimates in each iteration. Please see the definition of the function for details.

When the algorithm terminates because of convergence, the function returns the following vector c(para.LSE, LogL, all.para, iter), where

1. para.LSE = the parameter estimate in the distribution-free model, i.e., the LSE based model.
2. LogL = the log-likelihood at the parameter estimates.
3. all.para = the vector of parameter estimate of fixed effects and the lower triangular entries of $\boldsymbol{A}$ that parameterizes the covariance matrices for the random effects.
4. iter = the number of total iteration used in the algorithm.

6.3 A user-friendly function

Using the functions "PHregIC()" and "get.para.EMFS()", we provide the following function that can be directly used for analyzing longitudinal data anchored by interval censored anchoring point.

```
PHregLMEICA<-function(int.l, int.r, int.subject, phreg.covariate=NULL,
                subject, response, covariates,  fixed.fun, random.fun,
                bs.se=TRUE, bs.seed.start=20170502, bs.num=50,
                outlier.control=(1e-50), initial.sigma.sq=1,
                max.iter=1000, accuracy=(1e-5), min.step.ctrl=(2e-4),
                initial.EM.ctrl=0.1,  EM.acc=2, EM.max.iter=1000,
                GroupEM.num=25, print.hist.phreg=FALSE,
                print.hist.lme=TRUE, plot.hist=TRUE,  plot.index=2:3,
                plot.y=0.2, plot.x=100,  plot.pch=20, plot.cex=1,
```

```
                plot.title=""){

if(!((length(int.l)==length(int.r)) &
    (length(int.r)==length(int.subject))))
   {stop("rows of int.l, int.r, int.subject must match!")}

if(!is.null(phreg.covariate))
{  if(!(is.matrix(phreg.covariate)))
     {stop("phreg.covariate must be in matrix form!")}
   if(length(int.l)!=dim(phreg.covariate)[1])
     {stop("rows of phreg.covariate must match intervals")}}

ID<-1:length(int.subject)

get.covariate<-function(i.tmp=1, F.tmp=F.hat){
  L.i<-int.l[i.tmp];   R.i<-int.r[i.tmp];
  S.i<-F.tmp$S[ L.i<F.tmp$S & F.tmp$S<=R.i ]
  hr.at.S.i<-F.tmp$weight[i.tmp]*
             F.tmp$hr[L.i<F.tmp$S & F.tmp$S<=R.i]
  F.at.S.i<-1-exp(-cumsum(hr.at.S.i))
  P.i.tmp<-F.at.S.i-c(0, F.at.S.i[-length(F.at.S.i)])
  P.i<-P.i.tmp/sum(P.i.tmp)
  Fhat.i<-cbind(S.i, P.i )

  covariates.i.tmp<-covariates[(subject==int.subject[i.tmp]),]
  response.i.tmp<-response[(subject==int.subject[i.tmp])]
  not.missing.i<-!is.na(rowSums(covariates.i.tmp)+response.i.tmp)
  covariates.i<-covariates.i.tmp[not.missing.i,]
  response.i<-response.i.tmp[not.missing.i]
  nobs.i<-sum(not.missing.i)

  X.ij<-lapply(S.i, function(ap.tmp)
   {t(apply(covariates.i, 1,
      function(vec.tmp=covariates[1,]) fixed.fun(vec.tmp, ap.tmp)))})

  Z.ij<-lapply(S.i, function(ap.tmp)
   {Z.ij.tmp<-apply(covariates.i, 1,
      function(vec.tmp) random.fun(vec.tmp, ap.tmp))
                       if(length(Z.ij.tmp)==dim(covariates.i)[1])
      {return(matrix(Z.ij.tmp, ncol=1))}else{return(t(Z.ij.tmp))}})

  list("intvl"=c(L.i, R.i),  "Fhat.i"=Fhat.i, "nobs.i"=nobs.i,
       "Y.i"=matrix(response.i, ncol=1),  "X.ij"=X.ij,  "Z.ij"=Z.ij)
}
```

```
F.hat<-PHregIC(L=int.l, R=int.r, Z=phreg.covariate,
               print.hist.phreg=print.hist.phreg)
if(is.null(F.hat)){stop("Cannot compute CDF.")}

dt.list<-lapply(ID, function(id.tmp)
                    get.covariate(i.tmp=id.tmp, F.tmp=F.hat))
N.fixed <-dim( (dt.list[[1]]$X.ij)[[1]] )[2]
N.random<-dim( (dt.list[[1]]$Z.ij)[[1]] )[2]

#### get para.est
para.est<-get.para.EMFS( lst=dt.list, weight=rep(1, length(dt.list)),
               n.fixed=N.fixed,  n.random=N.random,
               initial.G=diag(N.random),
               outlier.control=outlier.control,
               initial.sigma.sq=initial.sigma.sq,
               max.iter=max.iter, accuracy=accuracy,
               min.step.ctrl=min.step.ctrl,
               initial.EM.ctrl=initial.EM.ctrl,  EM.acc=EM.acc,
               EM.max.iter=EM.max.iter, GroupEM.num=GroupEM.num,
               print.hist.lme=print.hist.lme, plot.hist=plot.hist,
               plot.index=plot.index, plot.y=plot.y,
               plot.x=plot.x, plot.pch=plot.pch,
               plot.cex=plot.cex, plot.title=plot.title)
if(is.null(para.est)){stop("Cannot get parameter estiamte")}

if(bs.se){
  cat("\n Start bootstrap===>:\n")
  bs.paras<-matrix(rep(0, length(para.est)*bs.num), nrow=bs.num)
  bs.seed<-bs.seed.start; dummy<-1;
while(dummy <= bs.num){
  cat("  bs", dummy, ": ", sep="")
  set.seed(bs.seed)
  bs.id.dup<-sample(ID, size=length(ID), replace=TRUE)
  bs.id.dup<-bs.id.dup[order(bs.id.dup)]
  bs.id.unique<-unique(bs.id.dup)
  weight.bs<-as.vector(table(bs.id.dup))
  bs.phreg.covariates<-phreg.covariate[bs.id.dup,]
  if(!is.null(bs.phreg.covariates)){
    if(!is.matrix(bs.phreg.covariates))
      {bs.phreg.covariates<-as.matrix(bs.phreg.covariates)}}
  bs.F.hat<-PHregIC(L=int.l[bs.id.dup], R=int.r[bs.id.dup],
                    Z=bs.phreg.covariates,
                    print.hist.phreg=print.hist.phreg)
  bs.dt.list<-lapply(bs.id.unique, function(id.tmp)
                     get.covariate(i.tmp=id.tmp, F.tmp=bs.F.hat))
```

```
  bs.para.tmp<-get.para.EMFS(lst=bs.dt.list, weight=weight.bs,
 n.fixed=N.fixed, n.random=N.random, initial.G=diag(N.random),
 initial.sigma.sq=initial.sigma.sq, max.iter=max.iter,
accuracy=accuracy, min.step.ctrl=min.step.ctrl,
outlier.control=outlier.control, initial.EM.ctrl=initial.EM.ctrl,
EM.acc=EM.acc, EM.max.iter=EM.max.iter,
GroupEM.num=GroupEM.num, print.hist.lme=print.hist.lme,
plot.hist=plot.hist, plot.index=plot.index,
plot.y=plot.y, plot.x=plot.x, plot.pch=plot.pch,
plot.cex=plot.cex,
plot.title=paste(plot.title, ": seed-", bs.seed, sep=""))

  if(!is.null(bs.para.tmp))
    { bs.paras[dummy,]<-bs.para.tmp; dummy<-dummy+1 }
  bs.seed<-bs.seed+1
}## End of for-loop
cat("\n ===>: Finish Boostrap \n")
chs.tmp<-bs.paras[,length(para.est)]<(max.iter+EM.max.iter)
bs.paras<-bs.paras[chs.tmp, ]
if(dim(bs.paras)[1]<bs.num)
   {cat("Some bootstrap steps do not converge.",
        "Covariance matrix estimated based on convergent steps")}
bs.cov.lse<-cov(bs.paras[, 1:N.fixed])
bs.cov.lme<-cov(bs.paras[, N.fixed+1+(1:N.fixed)])
bs.cov.sigma.and.A<-cov(bs.paras[,(2*N.fixed+2):(length(para.est)-1)])

lse.est<-rbind(lse.est=para.est[1:N.fixed], se=sqrt(diag(bs.cov.lse)))
lme.est<-rbind(lme.est=para.est[N.fixed+1+ (1:N.fixed)],
               se=sqrt(diag(bs.cov.lme)))
ests<-rbind(lse.est, lme.est)
colnames(ests)<-paste("para", 1:N.fixed, sep="")

return(list("data.list"=dt.list,
            "F.hat"=F.hat,
            "para.est"=para.est,
            "bs.paras"=bs.paras,
            "bs.covs"=list("bs.cov.lse"=bs.cov.lse,
                           "bs.cov.lme"=bs.cov.lme,
                           "bs.cov.sigma.and.A"=bs.cov.sigma.and.A)))
}else{return(list("data.list"=dt.list,
                  "F.hat"=F.hat,
                  "para.est"=para.est))}
}
```

The above function takes the following arguments.

1. int.l, int.r

   Vectors. The i-th element of int.l and int.r are the left and right endpoints of the i-th censoring interval.

2. int.subject

   Vector. The i-th element is the subject id for the i-th subject.

3. phreg.covariate

   Covariance matrix for the proportional hazards model. The i-th row are the covariates for the i-th subject.

4. subject response, covariates

   the vector for subject, response, and the covariance matrix for the longitudinal model. int.subject must be a subset of subject. Subjects may have different number of observations. Missing values are allowed. response and covariates must be numerical.

5. fixed.fun, random.fun

   Because the true event time is not observed, fixed.fun and random.fun are functions that defines the fixed and random effects coefficient matrices as a function of event time t. fixed.fun must be a function that returns a vector of length at least 2

6. all other arugment

   These are options to manually control the algorithm, and obtain details of the iterations. Please see the definition of the function for details.

The function returns a list, which contains:

1. data.lst

   Data used to run the model. data.lst is a list of the same structure as lst in the "get.para.EMFS()" function in Section 6.2.

2. F.hat

   Estimate anchoring point distributions. F.hat is a list of the same structure as the list returned by the "PHregIC()" function in Section 6.1.

3. para.est

   The vector of parameter estimates, as returned by the "get.para.EMFS()" function in Section 6.2.

4. bs.paras

   All parameter estimates from the bootstrap steps, whose sample covariance matrix gives the bootstrap covariance matrix.

5. bs.covs

   Bootstrap covariance matrices for the distribution-free model estimates and the linear mixed-effects model estimates.

To avoid numeric problem of extremely small or large residuals, and hence to achieve stable computations, it is advised to scale the covariates vector and/or the response vector if they are large. It is also advised to remove outliers, if possible.

Chapter 7

Summary

Longitudinal data with subject-specific random anchoring events are often encountered in scientific research. Most of the existing methods rely on assumptions of the anchoring event time distribution. These assumptions are usually hard to verify, and hence the analysis is prone to biased estimation and questionable inference.

In this research, we took an initial step toward solving the problem for a general class of longitudinal models, in a situation where the anchoring event times are censored by observable intervals. Given a sample of size n, we proposed to first obtain the anchoring event distribution $\hat{F}_n$ using a nonparametric maximum likelihood method, and then estimate the model parameters as the value of a stochastic functional $\mathbb{Q}_n$ at $\hat{F}_n$. The stochastic functional $\mathbb{Q}_n$ was constructed in two different situations, one with a distribution-free model (Chapter 4) and the other with a linear mixed-effects model (Chapter 5). In the distribution-free model, $\mathbb{Q}_n$ is relative easy to compute since an explicit formula is available. The method is also robust on the outcome distribution. However, since the distribution-free model completely ignores the correlation between repeated observations of the same subject, the estimation efficiency can be improved if the within-subject correlations are appropriately modeled. Therefore, we considered a likelihood based approach. Under a general setting, we focused on the properties of the semiparametric maximum pseudo-likelihood estimators, in which an estimated nonparametric distribution function $\hat{F}_n$ was used to compute the pseudo-likelihood. In this case, the stochastic functional $\mathbb{Q}_n$ is only implicitly defined as sending $\hat{F}_n$

to the maximizer of the corresponding pseudo-likelihood. Usually, the computation of $\mathbb{Q}_n(\hat{F}_n)$ can be complex. For linear mixed-effects model with interval censored anchoring events, we developed an algorithmically efficient algorithm to compute the model estimate. The statistic efficiency gain compared to the distribution-free model was empirically evaluated using simulation study (Section 5.3).

Using the empirical process theory, we showed that under mild regularity conditions, the proposed model estimates were consistent and asymptotically normally distributed with $n^{\frac{1}{2}}$-convergence rate (Theorems 4.2.2 and 5.1.2). A good finite-sample performance was demonstrated through simulation studies (Sections 4.3 and 5.3). Applications of the proposed method were illustrated through real data analysis (Sections 4.4 and 5.4). Based on the rigorously developed large-sample theory and good finite-sample performance, we recommend the application of our proposed method for longitudinal data when the time scale is anchored by interval censored events.

Considering the levels of maturity and variety of the existing longitudinal models, our research is at best an initial attempt towards the goal of a more complete solution. There are certainly important questions that remain to be addressed. Among them are the incorporation of nonlinear modeling components and covariate-dependent anchoring event distributions. With the extensions of the basic modeling structure, new computational algorithms also need to be developed for model fitting.

BIBLIOGRAPHY

Carter, C. L., C. Allen, and D. E. Henson (1989). Relation of tumor size, lymph node status, and survival in 24,740 breast cancer cases. *Cancer 63*(1), 181–187.

Dudley, R. M. (1984). A course on empirical processes. In *Ecole d'été de Probabilités de Saint-Flour XII-1982*, pp. 1–142. Springer.

Durbán, M., J. Harezlak, M. Wand, and R. Carroll (2005). Simple fitting of subject-specific curves for longitudinal data. *Statistics in medicine 24*(8), 1153–1167.

Fournier, D., E. Weber, W. Hoeffken, M. Bauer, F. Kubli, and V. Barth (1980). Growth rate of 147 mammary carcinomas. *Cancer 45*(8), 2198–2207.

Geskus, R. and P. Groeneboom (1999). Asymptotically optimal estimation of smooth functionals for interval censoring, case 2. *The Annals of Statistics 27*(2), 627–674.

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

Hall, S. S. (2006). *Size Matters: How Height Affects the Health, Happiness, and Success of Boys–and the Men They Become*. Houghton Mifflin Harcourt.

Holden, H., B. Øksendal, J. Ubøe, and T. Zhang (1996). Stochastic partial differential equations. In *Stochastic partial differential equations*, pp. 141–191. Springer.

Huang, J. and J. A. Wellner (1995). Asymptotic normality of the npmle of linear functionals for interval censored data, case 1. *Statistica Neerlandica 49*(2), 153–163.

Klenke, A. (2008). Probability theory. universitext.

Kosorok, M. R. (2007). *Introduction to empirical processes and semiparametric inference*. Springer Science & Business Media.

Laird, N. M. and J. H. Ware (1982). Random-effects models for longitudinal data. *Biometrics 38*(4), 963–974.

Muller, H.-G. (1992). Change-points in nonparametric regression analysis. *The Annals of Statistics 20*(2), 737–761.

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

Pollard, D. (2012). *Convergence of stochastic processes*. Springer Science & Business Media.

Robinson, L. F., T. D. Wager, and M. A. Lindquist (2010). Change point estimation in multi-subject fmri studies. *Neuroimage 49*(2), 1581–1592.

Shankar, R. R., G. J. Eckert, C. Saha, W. Tu, and J. H. Pratt (2005). The change in blood pressure during pubertal growth. *The Journal of Clinical Endocrinology & Metabolism 90*(1), 163–167.

Spratt, J. A., D. Von Fournier, J. S. Spratt, and E. E. Weber (1993). Decelerating growth and human breast cancer. *Cancer 71*(6), 2013–2019.

Tanner, J. M. and R. H. Whitehouse (1976). Clinical longitudinal standards for height, weight, height velocity, weight velocity, and stages of puberty. *Archives of disease in childhood 51*(3), 170–179.

Tu, W., G. J. Eckert, T. S. Hannon, H. Liu, L. M. Pratt, M. A. Wagner, L. A. DiMeglio, J. Jung, and J. H. Pratt (2014). Racial differences in sensitivity of blood pressure to aldosterone. *Hypertension 63*(6), 1212–1218.

Tu, W., G. J. Eckert, C. Saha, and J. H. Pratt (2009). Synchronization of adolescent blood pressure and pubertal somatic growth. *The Journal of Clinical Endocrinology & Metabolism 94*(12), 5019–5022.

van den Hout, A., G. Muniz-Terrera, and F. E. Matthews (2013). Change point models for cognitive tests using semi-parametric maximum likelihood. *Computational statistics & data analysis 57*(1), 684–698.

van der Vaart, A. (1991). On differentiable functionals. *The Annals of Statistics 19*(1), 178–204.

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

van der Vaart, A. and J. A. Wellner (1996). *Weak Convergence and Empirical Processes with Application to Statistics*. Springer Verlage.

Wand, M. (2002). Vector differential calculus in statistics. *The American Statistician 56*(1), 55–62.

Wellner, J. A. and Y. Zhang (2007). Two likelihood-based semiparametric estimation methods for panel count data with covariates. *The Annals of Statistics 35*(5), 2106–2142.

Zhang, Y., G. Cheng, and W. Tu (2016). Robust nonparametric estimation of monotone regression functions with interval-censored observations. *Biometrics 72*(3), 720–730.

CURRICULUM VITAE

Chenghao Chu

EDUCATION

- Ph.D. in Biostatistics, Indiana University, Indianapolis, IN, 2018 (minor in Epidemiology)
- Ph.D. in Mathematics, Northwestern University, Evanton, IL, 2008
- B.S. in Mathematics, University of Science and Technology of China, Hefei, Anhui, China, 2000

WORK EXPERIENCE

- Summer Intern, Astellas Pharma US, Inc. Northbrook, IL, 2017
- Associate Professor, University of Science and Technology of China, Hefei, Anhui, China, 2011-2014
- Assistant Professor, The Johns Hopkins University, Baltimore, MD, 2008-2011