New Statistical Issues for Censored Survival Data: High-Dimensionality and Censored Covariate

by

Shengchun Kong

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biostatistics)
in The University of Michigan
2014

Doctoral Committee:

Professor Bin Nan, Chair
Professor Sioban D. Harlow
Professor John D. Kalbfleisch
Professor Naisyin Wang

To Shan and my parents

ACKNOWLEDGEMENTS

First and foremost, I would like to express my sincere gratitude to my advisor Dr. Bin Nan, who has always been insightful and helpful, encouraging and inspiring. I have benefited so much from his insights into statistics and constructive suggestions. Without his support and guidance, not only would this thesis be started or completed, the improvement on my statistical knowledge and many professional skills could not have been gained.

My thanks go out to Dr. Jack Kalbfleisch, Dr. Naisyin Wang and Dr. Sioban Harlow for serving on my dissertation committee and providing valuable comments and suggestions on my research. Special thanks are due to Dr. Jack Kalbfleisch who gave me my first survival lecture. He was and remains willing to help and gave me great advice on many things ranging from problem solving, teaching, and job hunting.

I am deeply grateful to Dr. Catherine Kim, who provided a great opportunity for me to work on women's health-related problems and broaden my knowledge in medical science. I would also like to thank Dr. Brisa Sánchez, who gave me financial support through her research grant. It has been a very pleasant experience working as a graduate student research assistant under their supervision.

I am thankful for all the knowledgeable faculty members of the Department of Biostatistics at the University of Michigan for their excellent lectures, and for all the smart fellow students for their support and company throughout graduate school.

I must thank my parents for their unconditional love and absolute faith in me, and my husband Shan, who can always make me happy from my heart and enables

more meaning in life for me. This thesis is dedicated to you - my beloved family.

TABLE OF CONTENTS

LIST OF TABLES

Table

ABSTRACT

New Statistical Issues for Censored Survival Data: High-Dimensionality and Censored Covariate

by

Shengchun Kong

Chair: Bin Nan

Censored survival data arise commonly in many areas including epidemiology, engineering and sociology. In this dissertation, we explore several emerging statistical issues for censored survival data.

In Chapter II, we consider finite sample properties of the regularized high-dimensional Cox regression via lasso. Existing literature focuses on linear or generalized linear models with Lipschitz loss functions, where the empirical risk functions are the summations of independent and identically distributed (iid) losses. The summands in the negative log partial likelihood function for censored survival data, however, are neither iid nor Lipschitz. We first approximate the negative log partial likelihood function by a sum of iid non-Lipschitz terms, then derive the non-asymptotic oracle inequalities for the lasso penalized Cox regression, using pointwise arguments to tackle the difficulties caused by lacking iid Lipschitz losses.

In Chapter III, we consider generalized linear regression analysis with a left-censored covariate due to the limit of detection. The complete case analysis yields valid estimates for regression coefficients, but loses efficiency. Substitution meth-

ods are biased; the maximum likelihood method relies on parametric models for the unobservable tail probability, thus may suffer from model misspecification. To obtain robust and more efficient results, we propose a semiparametric likelihood-based approach for the regression parameters using an accelerated failure time model for the left-censored covariate. A two-stage estimation procedure is considered. The proposed method outperforms the existing methods in simulation studies. Technical conditions for asymptotic properties are provided.

In Chapter IV , we consider longitudinal data analysis with a terminal event. The existing methods include the joint modeling approach and the marginal estimating equation approach, and both assume that the relationship between the response variable and a set of covariates is the same no matter whether the terminal event occurs or not. This assumption, however, is not reasonable for many longitudinal studies. Therefore we directly model event time as a covariate, which provides intuitive interpretation. When the terminal event times are right-censored, a semiparametric likelihood-based approach similar to Chapter III is proposed for the parameter estimations. The proposed method outperforms the complete case analysis in simulation studies and its asymptotic properties are provided.

CHAPTER I

Introduction

This dissertation primarily explores survival analysis and spans several important areas of statistics including high-dimensional data analysis and longitudinal data analysis. There are three different topics in this dissertation, and each focuses on a different problem with its own features. The last two topics concern the censored covariate issue in two different setups.

1.1 High-dimensional Cox Regression via LASSO

Since it was introduced by Tibshirani (1996), the lasso regularized method for high-dimensional regression models with sparse coefficients has received a great deal of attention in the literature; see, for instance, Bickel, Ritov, and Tsybakov (2009) and van de Geer (2008). Properties of interest for such regression models include the finite sample oracle inequalities. For censored survival data, the Cox regression model is the most widely used method and is of great interest whether the oracle inequalities hold. Unlike the linear models and generalized linear models, however, the finite sample non-asymptotic statistical properties for the Cox model are extremely difficult, mainly due to lacking independent and identically distributed (iid) Lipschitz losses in the partial likelihood.

To address this problem, in Chapter II, we first approximate the negative log partial likelihood function by a sum of iid non-Lipschitz terms. With the Lipschitz condition replaced by a less restrictive boundedness assumption for the regression parameters, we tackle the problem using pointwise arguments to obtain the oracle bounds of two types of errors: one is between the empirical loss and the expected loss, and one is between the negative log partial likelihood and the empirical loss. We show that the non-asymptotic error bounds for the lasso penalized Cox regression have the same order as if the set of underlying non-zero coefficients were given ahead by an oracle, which is typically of the order $\log m \times dim_\theta/n$ where m is the number of covariates, n is the number of observations, and dim_θ is the number of non-zero coefficients.

1.2 Covariate Subject to Limit of Detection

Detection limit is a threshold below which measured values are not considered significantly different from background noise. Hence, values measured below this threshold are unreliable. A variety of statistical tools have been developed to tackle this problem with the response variable subject to limit of detection (LOD). For example, standard semiparametric survival models can be applied because LOD is in fact left censoring, which can be easily transformed to right censoring by changing the sign of the variable. Estimation for regression models with a covariate subject to LOD is more difficult, and many ad hoc methods have been implemented in practice but found to be inappropriate. The complete case analysis, which eliminates observations with covariate values below LOD, yields valid estimates for regression coefficients, but loses efficiency. Substitution methods are easily implementable, but can yield large bias. Maximum likelihood methods rely on parametric models for the unobservable tail probability, and therefore may suffer from model misspecification.

In Chapter III, to obtain more efficient and yet robust results, we propose a semi-parametric likelihood-based approach to fit generalized linear models with covariate subject to LOD. The tail distribution of the covariate beyond its LOD is estimated from a semiparametric accelerated failure time (AFT) model, conditional on all the fully observed covariates. A two-stage estimation procedure is considered, where the conditional distribution of the covariate with LOD given other variables is estimated prior to maximizing the likelihood function. The estimation based on the proposed method is proved to be consistent and asymptotically normal, and outperforms existing methods in simulations. In the anti-Mullerian hormone data analysis, the proposed two-stage method yields similar point estimates with smaller variances compared to the complete case analysis, indicating the efficiency gain of the proposed method.

1.3 Longitudinal Data Analysis with Terminal Event

Chapter IV considers the longitudinal data analysis with terminal events. In longitudinal studies, the collection of information can be stopped at the end of the study, or at the time of dropout of a study participant, or at the time that a terminal event occurs. For example, death, the most common terminal event, often occurs in aging cohort studies and cancer studies. Existing methods include the joint modeling approach using latent frailty and the marginal estimating equation approach using inverse probability weighting. Neither approach directly models the effect of terminal event to the response variable or to the relationship between response variable and covariates. These type of modeling strategies, however, are not reasonable for many longitudinal studies, where the explicit effect of terminal event time is of interest.

We propose to directly model event time as a covariate, which provides intuitive interpretation. When the terminal event times are right-censored, a semiparametric likelihood-based approach is proposed for the parameter estimation, where the Cox regression model is used for the censored terminal event time. We consider a two-stage estimation procedure, where the conditional distribution of the right-censored

terminal event time given other variables is estimated prior to maximizing the likelihood function for the regression parameters. The proposed method outperforms the complete case analysis in simulation studies, which simply eliminates the subjects with censored terminal event times. Desirable asymptotic properties are provided.

CHAPTER II

Non-Asymptotic Oracle Inequalities for the High-Dimensional Cox Regression via Lasso

2.1 Introduction

Since it was introduced by Tibshirani (1996), the lasso regularized method for high-dimensional regression models with sparse coefficients has received a great deal of attention in the literature. Properties of interest for such regression models include the finite sample oracle inequalities. Among the extensive literature of the lasso method, Bunea, Tsybakov, and Wegkamp (2007) and Bickel, Ritov, and Tsybakov (2009) derived the oracle inequalities for prediction risk and estimation error in a general nonparametric regression model, including the high-dimensional linear regression as a special example, and van de Geer (2008) provided oracle inequalities for the generalized linear models with Lipschitz loss functions, e.g., logistic regression and classification with hinge loss. Bunea (2008) and Bach (2010) also considered the lasso regularized logistic regression. For censored survival data, the lasso penalty has been applied to the regularized Cox regression in the literature, see e.g. Tibshirani (1997) and Gui and Li (2005), among others. Recently, Bradic, Fan, and Jiang (2011) studied the asymptotic properties of the lasso regularized Cox model. However, its finite sample non-asymptotic statistical properties have not yet been established in the literature to the best of our knowledge, largely due to lacking iid Lipschitz losses from the partial likelihood. Nonetheless, the lasso approach has been studied extensively in the literature for other models, see e.g. Martinussen and Scheike (2009) and Gaiffas and Guilloux (2012), among others, for the additive hazards model.

We consider the non-asymptotic statistical properties of the lasso regularized high-dimensional Cox regression. Let T be the survival time and C the censoring time. Suppose we observe a sequence of iid observations $(\mathbf{X}_i, Y_i, \Delta_i)$, $i = 1, \dots, n$, where $\mathbf{X}_i = (X_{i1}, \cdots, X_{im})$ are the m-dimensional covariates in $\mathscr{X}$, $Y_i = T_i \wedge C_i$, and $\Delta_i = I_{\{T_i \leq C_i\}}$. Due to a large amount of parallel material, we follow closely the notation in van de Geer (2008). Let

$$\mathscr{F} = \left\{ f_\theta(\mathbf{x}) = \sum_{k=1}^{m} \theta_k x_k,\ \theta \in \Theta \subset \mathbf{R}^{\mathbf{m}} \right\}.$$

Consider the Cox model (Cox (1972)):

$$\lambda(t|\mathbf{X}) = \lambda_0(t) e^{f_\theta(\mathbf{X})},$$

where θ is the parameter of interest and λ_0 is the unknown baseline hazard function. The negative log partial likelihood function for θ is

$$l_n(\theta) = -\frac{1}{n} \sum_{i=1}^{n} \left\{ f_\theta(\mathbf{X}_i) - \log \left[\frac{1}{n} \sum_{j=1}^{n} 1(Y_j \geq Y_i) e^{f_\theta(\mathbf{X}_j)} \right] \right\} \Delta_i. \tag{2.1}$$

The corresponding estimator with lasso penalty is denoted by

$$\hat{\theta}_n := \arg\min_{\theta \in \Theta} \{ l_n(\theta) + \lambda_n I(\theta) \},$$

where $I(\theta) := \sum_{k=1}^{m} \sigma_k |\theta_k|$ is the weighted l_1 norm of the vector $\theta \in \mathbf{R}^{\mathbf{m}}$. van de Geer (2008) considered σ_k to be the square-root of the second moment of the k-th covariate X_k, either at the population level (fixed) or at the sample level (random). For normalized X_k, $\sigma_k = 1$. We consider fixed weights σ_k, $k = 1, \cdots, m$. The results for random weights can be easily obtained from the case with fixed weights following van de Geer (2008), and we leave the detailed calculation to interested readers.

Clearly the negative log partial likelihood (2.1) is a sum of non-iid random variables. For ease of calculation, consider an intermediate function as a "replacement" of the negative log partial likelihood function

$$\tilde{l}_n(\theta) = -\frac{1}{n}\sum_{i=1}^{n} \{f_\theta(\mathbf{X}_i) - \log \mu(Y_i; f_\theta)\} \Delta_i \tag{2.2}$$

that has the iid structure, but with an unknown population expectation

$$\mu(t; f_\theta) = E_{\mathbf{X},Y}\Big\{1(Y \geq t)e^{f_\theta(\mathbf{X})}\Big\}.$$

The negative log partial likelihood function (2.1) can then be viewed as a "working" model for the empirical loss function (2.2). The corresponding loss function is

$$\gamma_{f_\theta} = \gamma(f_\theta(\mathbf{X}), Y, \Delta) := -\{f_\theta(\mathbf{X}) - \log \mu(Y; f_\theta)\}\Delta, \tag{2.3}$$

with expected loss

$$l(\theta) = -E_{Y,\Delta,\mathbf{X}}[\{f_\theta(\mathbf{X}) - \log \mu(Y; f_\theta)\}\Delta] = P\gamma_{f_\theta}, \tag{2.4}$$

where P denotes the distribution of $(Y, \Delta, \mathbf{X})$. Define the target function $\bar{f}$ as

$$\bar{f} := \arg\min_{f \in \mathscr{F}} P\gamma_f := f_{\bar{\theta}},$$

where $\bar{\theta} = \arg\min_{\theta\in\Theta} P\gamma_{f_\theta}$. It is well-known that $P\gamma_{f_\theta}$ is convex with respect to θ for the regular Cox model, see for example, Andersen and Gill (1982). Thus, the above minimum is unique if the Fisher information matrix of θ at $\bar{\theta}$ is non-singular. Define the excess risk of f by

$$\mathcal{E}(f) := P\gamma_f - P\gamma_{\bar{f}}.$$

It is desirable to show similar non-asymptotic oracle inequalities for the Cox regression model as in, for example, van de Geer (2008) for generalized linear models. That is, with large probability,

$$\mathcal{E}(f_{\hat{\theta}_n}) \leq const. \times \min_{\theta \in \Theta} \left\{ \mathcal{E}(f_\theta) + \mathcal{V}_\theta \right\}.$$

Here $\mathcal{V}_\theta$ is called the "estimation error", which is typically proportional to λ_n^2 times the number of nonzero elements in θ.

Note that the summands in the negative log partial likelihood function (2.1) are not iid, and the intermediate loss function $\gamma(\cdot, Y, \Delta)$ given in (2.3) is not Lipschitz. Hence the general result of van de Geer (2008) that requires iid Lipschitz loss functions does not apply to the Cox regression. We tackle the problem using pointwise arguments to obtain the oracle bounds of two types of errors: one is between empirical loss (2.2) and expected loss (2.4) without involving the Lipschitz requirement of van de Geer (2008), and one is between the negative log partial likelihood (2.1) and empirical loss (2.2) which establishes the iid approximation of non-iid losses. These steps distinguish our work from that of van de Geer (2008); we rely on the Mean Value Theorem with van de Geer's Lipschitz condition replaced by the similar, but much less restrictive, boundedness assumption for regression parameters in Bühlmann (2006).

The article is organized as follows. In Section II.2, we provide assumptions that are used throughout the paper. In Section II.3, we define several useful quantities followed by the main result. We then provide a detailed proof in Section II.4 by introducing a series of lemmas and corollaries useful for deriving the oracle inequalities for the Cox model. To avoid duplicate material as much as possible, we refer to the preliminaries and some results in van de Geer (2008) from place to place in the proofs without providing much detail.

2.2 Assumptions

We impose five basic assumptions. Let $\|\cdot\|$ be the $L_2(P)$ norm and $\|\cdot\|_\infty$ the sup norm.

Assumption II.A. $K_m := \max_{1\le k\le m}\{\|X_k\|_\infty/\sigma_k\} < \infty$.

Assumption II.B. There exists an $\eta > 0$ and strictly convex increasing G, such that for all $\theta \in \Theta$ with $\|f_\theta - \bar{f}\|_\infty \le \eta$, one has $\mathcal{E}(f_\theta) \ge G(\|f_\theta - \bar{f}\|)$.

In particular, G can be chosen as a quadratic function with some constant C_0, i.e., $G(u) = u^2/C_0$, then the convex conjugate of function G, denoted by H, such that $uv \le G(u) + H(v)$ is also quadratic.

Assumption II.C. There exists a function $D(\cdot)$ on the subsets of the index set $\{1, \cdots, m\}$, such that for all $\mathscr{K} \subset \{1, \cdots, m\}$, and for all $\theta \in \Theta$ and $\tilde{\theta} \in \Theta$, we have $\sum_{k\in\mathscr{K}} \sigma_k|\theta_k - \tilde{\theta}_k| \le \sqrt{D(\mathscr{K})}\|f_\theta - f_{\tilde{\theta}}\|$. Here, $D(\mathscr{K})$ is chosen to be the cardinal number of $\mathscr{K}$.

Assumption II.D. $L_m := \sup_{\theta\in\Theta} \sum_{k=1}^{m} |\theta_k| < \infty$.

Assumption II.E. The observation time stops at a finite time $\tau > 0$, with $\xi := P(Y \ge \tau) > 0$.

Assumptions II.A, II.B, and II.C are identical to those in van de Geer (2008) with her ψ_k the identity function. Assumptions II.B and II.C can be easily verified for the random design setting where $\mathbf{X}$ is random (van de Geer (2008)) together with the usual assumption of non-singular Fisher information matrix at $\bar{\theta}$ (and its neighborhood) for the Cox model. Assumption II.D has a similar flavor to the assumption (A2) in Bühlmann (2006) for the persistency property of boosting method in high-dimensional linear regression models, but is much less restrictive in the sense that L_m is allowed to depend on m in contrast with the fixed constant in Bühlmann (2006). Here it replaces the Lipschitz assumption in van de Geer (2008). Assumption II.E is commonly used for survival models with censored data, see for example, Andersen and Gill (1982). A straightforward extension of Assumption II.E is to allow τ (thus ξ) to depend on n.

From Assumptions II.A and II.D, we have, for any $\theta \in \Theta$,

$$e^{|f_\theta(\mathbf{X}_i)|} \le e^{K_m L_m \sigma_{(m)}} := U_m < \infty \tag{2.5}$$

for all i, where $\sigma_{(m)} = \max_{1\le k\le m} \sigma_k$. Note that U_m is allowed to depend on m.

2.3 Main result

Let $I(\theta) := \sum_{k=1}^m \sigma_k|\theta_k|$ be the l_1 norm of θ. For any θ and $\tilde{\theta}$ in Θ, denote

$$I_1(\theta|\tilde{\theta}) := \sum_{k:\tilde{\theta}_k\neq 0} \sigma_k|\theta_k|, \;\; I_2(\theta|\tilde{\theta}) := I(\theta) - I_1(\theta|\tilde{\theta}).$$

Consider the estimator

$$\hat{\theta}_n := \arg\min_{\theta\in\Theta}\{l_n(\theta) + \lambda_n I(\theta)\}.$$

2.3.1 Useful quantities

We first define a set of useful quantities that are involved in the oracle inequalities.

- $\bar{a}_n = 4a_n, \quad a_n = \sqrt{\frac{2K_m^2 \log(2m)}{n}} + \frac{K_m \log(2m)}{n}$.
- $r_1 > 0$, $b > 0$, $d > 1$, and $1 > \delta > 0$ are arbitrary constants.
- $d_b := d\left(\frac{b+d}{(d-1)b} \vee 1\right)$.

- $\bar{\lambda}_{n,0} = \bar{\lambda}^A_{n,0} + \bar{\lambda}^B_{n,0}$, where

$$\bar{\lambda}^A_{n,0} := \bar{\lambda}^A_{n,0}(r_1) := \bar{a}_n \left(1 + 2r_1\sqrt{2\left(K_m^2 + \bar{a}_n K_m\right)} + \frac{4r_1^2 \bar{a}_n K_m}{3}\right),$$
$$\bar{\lambda}^B_{n,0} := \bar{\lambda}^B_{n,0}(r_1) := \frac{2K_m U_m^2}{\xi}\left(2\bar{a}_n r_1 + \sqrt{\frac{\log(2m)}{n}}\right).$$

- $\lambda_n := (1+b)\bar{\lambda}_{n,0}$.

- $\delta_1 = (1+b)^{-N_1}$ and $\delta_2 = (1+b)^{-N_2}$ are arbitrary constants for some N_1 and N_2, where $N_1 \in \mathbf{N} := \{1, 2, \dots\}$ and $N_2 \in \mathbf{N} \cup \{0\}$.

- $d(\delta_1, \delta_2) = 1 + \frac{1+(d^2-1)\delta_1}{(d-1)(1-\delta_1)}\delta_2$.

- W is a fixed constant given in Lemma II.3 for a class of empirical processes.

- $D_\theta := D(\{k : \theta_k \neq 0, k = 1, \dots, m\})$ is the number of nonzero θ_k's, where $D(\cdot)$ is given in Assumption II.C.

- $\mathcal{V}_\theta := 2\delta H\left(\frac{2\lambda_n\sqrt{D_\theta}}{\delta}\right)$, where H is the convex conjugate of function G defined in Assumption II.B.

- $\theta^*_n := \arg\min_{\theta\in\Theta}\{\mathcal{E}(f_\theta) + \mathcal{V}_\theta\}$.

- $\epsilon^*_n := (1+\delta)\mathcal{E}(f_{\theta^*_n}) + \mathcal{V}_{\theta^*_n}$.

- $\zeta^*_n := \frac{\epsilon^*_n}{\bar{\lambda}_{n,0}}$.

- $\theta(\epsilon^*_n) := \arg\min_{\theta\in\Theta, I(\theta-\theta^*_n)\le d_b\zeta^*_n/b}\{\delta\mathcal{E}(f_\theta) - 2\lambda_n I_1(\theta - \theta^*_n|\theta^*_n)\}$.

In the above, the dependence of θ^*_n on the sample size n is through $\mathcal{V}_\theta$ that involves the tuning parameter λ_n. We also impose conditions as in van de Geer (2008):

Condition II.I(b, δ). $\|f_{\theta^*_n} - \bar{f}\|_\infty \le \eta$.

Condition II.II(b, δ, d). $\|f_{\theta(\epsilon^*_n)} - \bar{f}\|_\infty \le \eta$.

In both conditions, η is given in Assumption II.B.

2.3.2 Oracle inequalities

We now provide our theorem on oracle inequalities for the Cox model lasso estimator, with detailed proof given in the next section. The key idea of the proof is to find bounds of differences between empirical errors of the working model (2.2) and between approximation errors of the partial likelihood, denoted as Z_θ and R_θ in the next section.

Theorem II.1. *Suppose Assumptions II.A-II.E and Conditions II.I(b, δ) and II.II(b, δ, d) hold. With*

$$\Delta(b, \delta, \delta_1, \delta_2) := d(\delta_1, \delta_2)\frac{1-\delta^2}{\delta b} \vee 1,$$

we have, with probability at least

$$1 - \left\{\log_{1+b} \frac{(1+b)^2\Delta(b,\delta,\delta_1,\delta_2)}{\delta_1\delta_2}\right\} \times \left\{\left(1+\frac{3}{10}W^2\right)\exp\left(-n\bar{a}_n^2 r_1^2\right) + 2\exp\left(-n\xi^2/2\right)\right\},$$

that

$$\mathcal{E}(f_{\hat{\theta}_n}) \leq \frac{1}{1-\delta}\epsilon_n^* \quad \textit{and} \quad I(\hat{\theta}_n - \theta_n^*) \leq d(\delta_1, \delta_2)\frac{\zeta_n^*}{b}.$$

2.4 Proofs

2.4.1 Preparations

Denote the empirical probability measure based on the sample $\{(\mathbf{X}_i, Y_i, \Delta_i) : i = 1, \ldots n\}$ by P_n. Let $\varepsilon_1, \cdots, \varepsilon_n$ be a Rademacher sequence, independent of the training data $(\mathbf{X}_1, Y_1, \Delta_1), \cdots, (\mathbf{X}_n, Y_n, \Delta_n)$. For some fixed $\theta^* \in \Theta$ and some $M > 0$, denote $\mathscr{F}_M := \{f_\theta : \theta \in \Theta, I(\theta - \theta^*) \leq M\}$. Later we take $\theta^* = \theta_n^*$, which is the case of interest. For any θ where $I(\theta - \theta^*) \leq M$, denote

$$Z_\theta(M) := |(P_n - P)[\gamma_{f_\theta} - \gamma_{f_{\theta^*}}]| = \left| \left[\tilde{l}_n(\theta) - l(\theta)\right] - \left[\tilde{l}_n(\theta^*) - l(\theta^*)\right] \right|.$$

Note that van de Geer (2008) sought to bound $\sup_{f \in \mathscr{F}_M} Z_\theta(M)$, thus the contraction theorem of Ledoux and Talagrand (1991) (Theorem A.3 in van de Geer (2008)) was needed, which holds for Lipschitz functions. We find that the calculation in van de Geer (2008) does not apply to the Cox model due to the lack of Lipschitz property. However, the pointwise argument is adequate for our purpose because only the lasso estimator or the difference between the lasso estimator $\hat{\theta}_n$ and the oracle θ_n^* is of interest. Note the notational difference between an arbitrary θ^* in the above $Z_\theta(M)$ and the oracle θ_n^*.

Lemma II.1. *Under Assumptions II.A, II.D, and II.E, for all θ satisfying $I(\theta - \theta^*) \leq M$, we have $EZ_\theta(M) \leq \bar{a}_n M$.*

Proof. By the symmetrization theorem, see e.g. van der Vaart and Wellner (1996) or Theorem A.2 in van de Geer (2008), for a class of only one function we have

$$\begin{aligned} EZ_\theta(M) &\leq 2E\left(\left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i \{[f_\theta(\mathbf{X}_i) - \log \mu(Y_i; f_\theta)]\Delta_i - [f_{\theta^*}(\mathbf{X}_i) - \log \mu(Y_i; f_{\theta^*})]\Delta_i\} \right| \right) \\ &\leq 2E\left(\left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i \{f_\theta(\mathbf{X}_i) - f_{\theta^*}(\mathbf{X}_i)\}\Delta_i \right| \right) \\ &\qquad + 2E\left(\left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i \{\log \mu(Y_i; f_\theta) - \log \mu(Y_i; f_{\theta^*})\}\Delta_i \right| \right) \\ &= A + B. \end{aligned}$$

For A we have

$$A \leq 2\left(\sum_{k=1}^m \sigma_k |\theta_k - \theta_k^*| \right) E\left(\max_{1 \leq k \leq m} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i \Delta_i X_{ik} / \sigma_k \right| \right).$$

Applying Lemma A.1 in van de Geer (2008), we obtain

$$E\left(\max_{1\le k\le m} \left| \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i \Delta_i \frac{X_{ik}}{\sigma_k} \right| \right) \le a_n.$$

Thus we have

$$A \le 2a_n M. \tag{2.6}$$

For B, instead of using the contraction theorem that requires Lipschitz, we use the Mean Value Theorem:

$$\begin{aligned}
&\left| \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i \{\log \mu(Y_i; f_\theta) - \log \mu(Y_i; f_{\theta^*})\}\Delta_i \right| \\
&\quad = \left| \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i \Delta_i \sum_{k=1}^{m} \frac{1}{\mu(Y_i; f_{\theta^{**}})} \iint_{Y_i \mathscr{X}}^{\infty} (\theta_k - \theta_k^*) x_k e^{f_{\theta^{**}}(x)} dP_{\mathbf{X},Y}(\mathbf{x}, y) \right| \\
&\quad = \left| \sum_{k=1}^{m} \sigma_k (\theta_k - \theta_k^*) \frac{1}{n}\sum_{i=1}^{n} \frac{\varepsilon_i \Delta_i}{\mu(Y_i; f_{\theta^{**}})\sigma_k} \iint_{Y_i \mathscr{X}}^{\infty} x_k e^{f_{\theta^{**}}(x)} dP_{\mathbf{X},Y}(\mathbf{x}, y) \right| \\
&\quad \le \left| \sum_{k=1}^{m} \sigma_k (\theta_k - \theta_k^*) \right| \max_{1\le k\le m} \left| \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i \Delta_i F_{\theta^{**}}(k, Y_i) \right| \\
&\quad \le M \max_{1\le k\le m} \left| \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i \Delta_i F_{\theta^{**}}(k, Y_i) \right|,
\end{aligned}$$

where θ^{**} is between θ and θ^*, and

$$F_{\theta^{**}}(k,t) = \frac{E\left[1(Y \ge t) X_k e^{f_{\theta^{**}}(\mathbf{X})}\right]}{\mu(t; f_{\theta^{**}})\sigma_k} \tag{2.7}$$

satisfying

$$|F_{\theta^{**}}(k,t)| \le \frac{(\|X_k\|_\infty/\sigma_k) E\left[1(Y \ge t) e^{f_{\theta^{**}}(\mathbf{X})}\right]}{\mu(t; f_{\theta^{**}})} \ \le\ K_m.$$

Since for all i,

$$E[\varepsilon_i \Delta_i F_{\theta^{**}}(k, Y_i)] = 0,\ \|\varepsilon_i \Delta_i F_{\theta^{**}}(k, Y_i)\|_\infty \leq K_m, \text{ and}$$
$$\frac{1}{n}\sum_{i=1}^{n} E[\varepsilon_i \Delta_i F_{\theta^{**}}(k, Y_i)]^2 \leq \frac{1}{n}\sum_{i=1}^{n} E[F_{\theta^{**}}(k, Y_i)]^2 \leq EK_m^2 = K_m^2,$$

following Lemma A.1 in van de Geer (2008), we obtain

$$B \leq 2a_n M. \tag{2.8}$$

Combining (2.6) and (2.8), the upper bound for $EZ_\theta(M)$ is achieved. □

We now can bound the tail probability of $Z_\theta(M)$ using the Bousquet's concentration theorem noted as Theorem A.1 in van de Geer (2008).

Corollary II.1. *Under Assumptions II.A, II.D, and II.E, for all $M > 0$, $r_1 > 0$ and all θ satisfying $I(\theta - \theta^*) \leq M$, it holds that*

$$P\left(Z_\theta(M) \geq \bar{\lambda}_{n,0}^A M\right) \leq \exp\left(-n\bar{a}_n^2 r_1^2\right).$$

Proof. Using the triangular inequality and the Mean Value Theorem, we obtain

$$\begin{aligned}
|\gamma_{f_\theta} - \gamma_{f_{\theta^*}}| &\leq |f_\theta(\mathbf{X}) - f_{\theta^*}(\mathbf{X})|\Delta + |\log\mu(Y; f_\theta) - \log\mu(Y; f_{\theta^*})|\Delta \\
&\leq \sum_{k=1}^{m} \sigma_k |\theta_k - \theta_k^*| \frac{|X_k|}{\sigma_k} + |\log\mu(Y; f_\theta) - \log\mu(Y; f_{\theta^*})| \\
&\leq MK_m + \sum_{k=1}^{m} \sigma_k |\theta_k - \theta_k^*| \cdot \max_{1\leq k\leq m} |F_{\theta^{**}}(k, Y)| \\
&\leq 2MK_m,
\end{aligned}$$

where θ^{**} is between θ and θ^*, and $F_{\theta^{**}}(k, Y)$ is defined in (2.7). So we have

$$\|\gamma_{f_\theta} - \gamma_{f_{\theta^*}}\|_\infty \leq 2MK_m, \quad \text{and} \quad P(\gamma_{f_\theta} - \gamma_{f_{\theta^*}})^2 \leq 4M^2K_m^2.$$

Therefore, in view of Bousquet's concentration theorem and Lemma II.1, for all $M > 0$ and $r_1 > 0$,

$$P\left(Z_\theta(M) \geq \bar{a}_n M \left(1 + 2r_1\sqrt{2\left(K_m^2 + \bar{a}_n K_m\right)} + \frac{4r_1^2 \bar{a}_n K_m}{3}\right)\right) \leq \exp\left(-n\bar{a}_n^2 r_1^2\right).$$

□

Now for any θ satisfying $I(\theta - \theta^*) \leq M$, we bound

$$\begin{aligned} R_\theta(M): &= \left|\left[l_n(\theta) - \tilde{l}_n(\theta)\right] - \left[l_n(\theta^*) - \tilde{l}_n(\theta^*)\right]\right| \\ &= \frac{1}{n}\sum_{i=1}^n \left|\left[\log\frac{1}{n}\sum_{j=1}^n \frac{1(Y_j \geq Y_i)e^{f_\theta(\mathbf{X}_j)}}{\mu(Y_i; f_\theta)} - \log\frac{1}{n}\sum_{j=1}^n \frac{1(Y_j \geq Y_i)e^{f_{\theta^*}(\mathbf{X}_j)}}{\mu(Y_i; f_{\theta^*})}\right]\Delta_i\right| \\ &\leq \sup_{0\leq t\leq\tau}\left|\log\frac{1}{n}\sum_{j=1}^n \frac{1(Y_j \geq t)e^{f_\theta(\mathbf{X}_j)}}{\mu(t; f_\theta)} - \log\frac{1}{n}\sum_{j=1}^n \frac{1(Y_j \geq t)e^{f_{\theta^*}(\mathbf{X}_j)}}{\mu(t; f_{\theta^*})}\right|. \end{aligned}$$

Here recall that τ is given in Assumption II.E. By the Mean Value Theorem, we have

$$\begin{aligned} R_\theta(M) &\leq \sup_{0\leq t\leq\tau}\left|\sum_{k=1}^m(\theta_k - \theta_k^*)\left\{\frac{\sum_{j=1}^n 1(Y_j \geq t)e^{f_{\theta^{**}}(\mathbf{X}_j)}}{\mu(t; f_{\theta^{**}})}\right\}^{-1}\right. \\ &\quad \left\{\frac{\sum_{j=1}^n 1(Y_j \geq t)X_{jk}e^{f_{\theta^{**}}(\mathbf{X}_j)}}{\mu(t; f_{\theta^{**}})}\right. \\ &\quad \left.\left. - \frac{\sum_{j=1}^n 1(Y_j \geq t)e^{f_{\theta^{**}}(\mathbf{X}_j)}E\left[1(Y \geq t)X_k e^{f_{\theta^{**}}(\mathbf{X})}\right]}{\mu(t; f_{\theta^{**}})^2}\right\}\right| \\ &= \sup_{0\leq t\leq\tau}\left|\sum_{k=1}^m \sigma_k(\theta_k - \theta_k^*)\left\{\frac{\sum_{j=1}^n 1(Y_j \geq t)(X_{jk}/\sigma_k)e^{f_{\theta^{**}}(\mathbf{X}_j)}}{\sum_{j=1}^n 1(Y_j \geq t)e^{f_{\theta^{**}}(\mathbf{X}_j)}}\right.\right. \\ &\quad \left.\left. - \frac{E\left[1(Y \geq t)(X_k/\sigma_k)e^{f_{\theta^{**}}(\mathbf{X})}\right]}{E\left[1(Y \geq t)e^{f_{\theta^{**}}(\mathbf{X})}\right]}\right\}\right| \\ &\leq M \sup_{0\leq t\leq\tau}\left[\frac{1}{n}\sum_{i=1}^n 1(Y_i \geq t)e^{f_{\theta^{**}}(\mathbf{X}_i)}\right]^{-1} \qquad (2.9) \\ &\quad \sup_{0\leq t\leq\tau}\left\{\max_{1\leq k\leq m}\left|\frac{1}{n}\sum_{i=1}^n 1(Y_i \geq t)(X_{ik}/\sigma_k)e^{f_{\theta^{**}}(\mathbf{X}_i)}\right.\right. \end{aligned}$$

$$- E\left[1(Y \geq t)(X_k/\sigma_k)e^{f_{\theta^{**}}(\mathbf{X})}\right] \Bigg|$$

$$+ K_m \left| \frac{1}{n}\sum_{i=1}^{n} 1(Y_i \geq t)e^{f_{\theta^{**}}(\mathbf{X}_i)} - E\left[1(Y \geq t)e^{f_{\theta^{**}}(\mathbf{X})}\right] \right| \Bigg\},$$

where θ^{**} is between θ and θ^* and, by (2.5), we have

$$\sup_{0\leq t\leq\tau} \left[\frac{1}{n}\sum_{i=1}^{n} 1(Y_i \geq t)e^{f_{\theta^{**}}(\mathbf{X}_i)}\right]^{-1} \leq U_m \left[\frac{1}{n}\sum_{i=1}^{n} 1(Y_i \geq \tau)\right]^{-1}. \qquad (2.10)$$

Lemma II.2. *Under Assumption II.E, we have*

$$P\left(\frac{1}{n}\sum_{i=1}^{n} 1(Y_i \geq \tau) \leq \frac{\xi}{2}\right) \leq 2e^{-n\xi^2/2}.$$

Proof. This is obtained directly from Massart (1990) for the Kolmogorov statistic by taking $r = \xi\sqrt{n}/2$ in the following:

$$\begin{aligned} P\left(\frac{1}{n}\sum_{i=1}^{n} 1(Y_i \geq \tau) \leq \frac{\xi}{2}\right) &\leq P\left(\sqrt{n}\left|\frac{1}{n}\sum_{i=1}^{n} 1(Y_i \geq \tau) - \xi\right| \geq r\right) \\ &\leq P\left(\sup_{0\leq t\leq\tau} \sqrt{n}\left|\frac{1}{n}\sum_{i=1}^{n} 1(Y_i \geq t) - P(Y \geq t)\right| \geq r\right) \\ &\leq 2e^{-2r^2}. \end{aligned}$$

□

Lemma II.3. *Under Assumptions II.A, II.D, and II.E, for all θ we have*

$$P\left(\sup_{0\leq t\leq\tau}\left|\frac{1}{n}\sum_{i=1}^{n} 1(Y_i \geq t)e^{f_\theta(\mathbf{X}_i)} - \mu(t; f_\theta)\right| \geq U_m \bar{a}_n r_1\right) \leq \frac{1}{5}W^2 e^{-n\bar{a}_n^2 r_1^2}, \qquad (2.11)$$

where W is a fixed constant.

Proof. For a class of functions indexed by t, $\mathscr{F} = \{1(y \geq t)e^{f_\theta(x)}/U_m : t \in [0, \tau], y \in \mathbf{R}, e^{f_\theta(x)} \leq U_m\}$, we calculate its bracketing number. For any nontrivial ϵ satisfying

$1 > \epsilon > 0$, let t_i be the i-th $\lceil 1/\varepsilon \rceil$ quantile of Y, so

$$P(Y \le t_i) = i\varepsilon, \quad i = 1, \cdots, \lceil 1/\varepsilon \rceil - 1,$$

where $\lceil x \rceil$ is the smallest integer that is greater than or equal to x. Furthermore, take $t_0 = 0$ and $t_{\lceil 1/\varepsilon \rceil} = +\infty$. For $i = 1, \cdots, \lceil 1/\varepsilon \rceil$, define brackets $[L_i, U_i]$ with

$$L_i(x, y) = 1(y \ge t_i)e^{f_\theta(x)}/U_m, \; U_i(x, y) = 1(y > t_{i-1})e^{f_\theta(x)}/U_m$$

such that $L_i(x, y) \le 1(y \ge t)e^{f_\theta(x)}/U_m \le U_i(x, y)$ when $t_{i-1} < t \le t_i$. Since

$$\begin{aligned}\left\{E[U_i - L_i]^2\right\}^{1/2} &\le \left\{E\left[\frac{e^{f_\theta(\mathbf{X})}}{U_m}\{1(Y \ge t_i) - 1(Y > t_{i-1})\}\right]^2\right\}^{1/2} \\ &\le \{P(t_{i-1} < Y \le t_i)\}^{1/2} = \sqrt{\varepsilon},\end{aligned}$$

we have $N_{[]}(\sqrt{\varepsilon}, \mathscr{F}, L_2) \le \lceil 1/\varepsilon \rceil \le 2/\varepsilon$, which yields

$$N_{[]}(\varepsilon, \mathscr{F}, L_2) \le \frac{2}{\varepsilon^2} = \left(\frac{K}{\varepsilon}\right)^2,$$

where $K = \sqrt{2}$. Thus, from Theorem 2.14.9 in van der Vaart and Wellner (1996), we have for any $r > 0$,

$$\begin{aligned}P\left(\sqrt{n} \sup_{0 \le t \le \tau}\left|\frac{1}{n}\sum_{i=1}^{n}\frac{1(Y_i \ge t)e^{f_\theta(\mathbf{X}_i)}}{U_m} - \frac{\mu(t; f_\theta)}{U_m}\right| \ge r\right) &\le \frac{1}{2}W^2 r^2 e^{-2r^2} \\ &\le \frac{1}{5}W^2 e^{-r^2},\end{aligned}$$

where W is a constant that only depends on K. Note that $r^2 e^{-r^2}$ is bounded by e^{-1}. With $r = \sqrt{n}\bar{a}_n r_1$, we obtain (2.11). □

Lemma II.4. *Under Assumptions II.A, II.D, and II.E, for all θ we have*

$$\begin{aligned} P\Bigg(\sup_{0\le t\le\tau} \max_{1\le k\le m} \Bigg| \frac{1}{n}\sum_{i=1}^n 1(Y_i \ge t)\frac{X_{ik}}{\sigma_k} e^{f_\theta(\mathbf{X}_i)} \\ - E\left[1(Y\ge t)\frac{X_k}{\sigma_k}e^{f_\theta(\mathbf{X})}\right] \Bigg| \ge K_m U_m \left[\bar{a}_n r_1 + \sqrt{\frac{\log(2m)}{n}}\right]\Bigg) \\ \le \frac{1}{10} W^2 e^{-n\bar{a}_n^2 r_1^2}. \end{aligned} \tag{2.12}$$

Proof. Consider the classes of functions indexed by t,

$$\begin{aligned} \mathscr{G}^k &= \Big\{1(y\ge t)e^{f_\theta(x)}x_k/(\sigma_k K_m U_m) : t\in[0,\tau], y \in \mathbf{R}, \\ & \qquad \left|e^{f_\theta(x)}x_k/\sigma_k\right| \le K_m U_m\Big\}, \quad k = 1,\dots,m. \end{aligned}$$

Using the argument in the proof of Lemma II.3, we have

$$N_{[\,]}(\varepsilon, \mathscr{G}^k, L_2) \le \left(\frac{K}{\varepsilon}\right)^2,$$

where $K = \sqrt{2}$, and then for any $r > 0$,

$$\begin{aligned} P\left(\sqrt{n}\sup_{0\le t\le\tau}\left|\frac{1}{n}\sum_{i=1}^n \frac{1(Y_i\ge t)e^{f_\theta(\mathbf{X}_i)}X_{ik}}{\sigma_k K_m U_m} - E\left[\frac{1(Y\ge t)e^{f_\theta(\mathbf{X})}X_k}{\sigma_k K_m U_m}\right]\right| \ge r\right) \\ \le \frac{1}{5}W^2 e^{-r^2}. \end{aligned}$$

Thus we have

$$\begin{aligned} P\Bigg(\sqrt{n}\sup_{0\le t\le\tau}\max_{1\le k\le m}\Bigg|\frac{1}{n}\sum_{i=1}^n 1(Y_i\ge t)e^{f_\theta(\mathbf{X}_i)}X_{ik}/\left(\sigma_k U_m K_m\right) \\ - E\left[1(Y\ge t)e^{f_\theta(\mathbf{X})}X_k/\left(\sigma_k U_m K_m\right)\right]\Bigg| \ge r\Bigg) \\ \le P\Bigg(\bigcup_{k=1}^m \sqrt{n}\sup_{0\le t\le\tau}\Bigg|\frac{1}{n}\sum_{i=1}^n 1(Y_i\ge t)e^{f_\theta(\mathbf{X}_i)}X_{ik}/\left(\sigma_k U_m K_m\right) \end{aligned}$$

$$
\begin{aligned}
&\quad - E\left[1(Y \geq t)e^{f_\theta(\mathbf{X})}X_k/\left(\sigma_k U_m K_m\right)\right]\Bigg| \geq r\Bigg) \\
&\leq\ mP\Bigg(\sqrt{n}\sup_{0\leq t\leq\tau}\Bigg|\frac{1}{n}\sum_{i=1}^{n}1(Y_i \geq t)e^{f_\theta(\mathbf{X}_i)}X_{ik}/\left(\sigma_k U_m K_m\right) \\
&\quad - E\left[1(Y \geq t)e^{f_\theta(\mathbf{X})}X_k/\left(\sigma_k U_m K_m\right)\right]\Bigg| \geq r\Bigg) \\
&\leq\ \frac{m}{5}W^2 e^{-r^2} = \frac{1}{10}W^2 e^{\log(2m)-r^2}.
\end{aligned}
$$

Let $\log(2m) - r^2 = -n\bar{a}_n^2 r_1^2$, so $r = \sqrt{n\bar{a}_n^2 r_1^2 + \log(2m)}$. Since

$$
\sqrt{\bar{a}_n^2 r_1^2 + \frac{\log(2m)}{n}} \leq \bar{a}_n r_1 + \sqrt{\frac{\log(2m)}{n}},
$$

we obtain (2.12). □

Corollary II.2. *Under Assumptions II.A, II.D, and II.E, for all $M > 0$, $r_1 > 0$, and all θ that satisfy $I(\theta - \theta^*) \leq M$, we have*

$$
P\left(R_\theta(M) \geq \bar{\lambda}_{n,0}^B M\right) \leq 2\exp\left(-n\xi^2/2\right) + \frac{3}{10}W^2\exp\left(-n\bar{a}_n^2 r_1^2\right). \tag{2.13}
$$

Proof. From (2.9) and (2.10) we have

$$
P\left(R_\theta(M) \leq \bar{\lambda}_{n,0}^B \cdot M\right) \geq P\left(E_1^c \cap E_2^c \cap E_3^c\right),
$$

where the events E_1, E_2 and E_3 are defined as

$$
\begin{aligned}
E_1 &= \left\{\frac{1}{n}\sum_{i=1}^{n}1(Y_i \geq \tau) \leq \xi/2\right\}, \\
E_2 &= \left\{\sup_{0\leq t\leq\tau}\left|\frac{1}{n}\sum_{i=1}^{n}1(Y_i \geq t)e^{f_{\theta^{**}}(\mathbf{X}_i)} - \mu(t; f_{\theta^{**}})\right| \geq U_m\bar{a}_n r_1\right\},
\end{aligned}
$$

$$E_3 = \left\{ \max_{1\le k\le m} \sup_{0\le t\le\tau} \left| \frac{1}{n}\sum_{i=1}^{n} 1(Y_i \ge t)\frac{X_{ik}}{\sigma_k} e^{f_{\theta^{**}}(\mathbf{X}_i)} - E\left[1(Y \ge t)\frac{X_k}{\sigma_k} e^{f_{\theta^{**}}(\mathbf{X})}\right] \right| \ge K_m U_m(\bar{a}_n r_1 + \sqrt{\frac{\log(2m)}{n}}) \right\}.$$

Thus

$$P\left(R_\theta(M) \ge \bar{\lambda}^B_{n,0} \cdot M\right) \le P(E_1) + P(E_2) + P(E_3),$$

and the result follows from Lemmas II.2, II.3 and II.4. □

Now with $\theta^* = \theta^*_n$, we have the following results.

Lemma II.5. *Suppose Conditions II.I(b, δ) and II.II(b, δ, d) are met. Under Assumptions II.B and II.C, for all $\theta \in \Theta$ with $I(\theta - \theta^*_n) \le d_b\zeta^*_n/b$, it holds that*

$$2\lambda_n I_1(\theta - \theta^*_n) \le \delta\mathcal{E}(f_\theta) + \epsilon^*_n - \mathcal{E}(f_{\theta^*_n}).$$

Proof. The proof is exactly the same as that of Lemma A.4 in van de Geer (2008), with the λ_n defined in Subsection 2.3.1. □

Lemma II.6. *Suppose Conditions II.I(b, δ) and II.II(b, δ, d) are met. Consider any random $\tilde{\theta} \in \Theta$ with $l_n(\tilde{\theta}) + \lambda_n I(\tilde{\theta}) \le l_n(\theta^*_n) + \lambda_n I(\theta^*_n)$. Let $1 < d_0 \le d_b$. It holds that*

$$P\left(I(\tilde{\theta} - \theta^*_n) \le d_0\frac{\zeta^*_n}{b}\right) \le P\left(I(\tilde{\theta} - \theta^*_n) \le \left(\frac{d_0 + b}{1 + b}\right)\frac{\zeta^*_n}{b}\right) + \left(1 + \frac{3}{10}W^2\right)\exp\left(-n\bar{a}_n^2 r_1^2\right) + 2\exp\left(-n\xi^2/2\right).$$

Proof. The idea is similar to the proof of Lemma A.5 in van de Geer (2008). Let $\tilde{\mathcal{E}} = \mathcal{E}(f_{\tilde{\theta}})$ and $\mathcal{E}^* = \mathcal{E}(f_{\theta^*_n})$. We will use short notation: $I_1(\theta) = I_1(\theta|\theta^*_n)$ and $I_2(\theta) = I_2(\theta|\theta^*_n)$. Since $l_n(\tilde{\theta}) + \lambda_n I(\tilde{\theta}) \le l_n(\theta^*_n) + \lambda_n I(\theta^*_n)$, on the set where $I(\tilde{\theta} - \theta^*_n) \le d_0\zeta^*_n/b$

and $Z_{\tilde{\theta}}(d_0\zeta_n^*/b) \le d_0\zeta_n^*/b \cdot \bar{\lambda}_{n,0}^A$, we have

$$\begin{aligned} R_{\tilde{\theta}}(d_0\zeta_n^*/b) &\ge [l_n(\theta_n^*) + \lambda_n I(\theta_n^*)] - [l_n(\tilde{\theta}) + \lambda_n I(\tilde{\theta})] - \lambda_n I(\theta_n^*) + \lambda_n I(\tilde{\theta}) \\ &\quad -[\tilde{l}_n(\theta_n^*) - \tilde{l}_n(\tilde{\theta})] \\ &\ge -\lambda_n I(\theta_n^*) + \lambda_n I(\tilde{\theta}) - [\tilde{l}_n(\theta_n^*) - \tilde{l}_n(\tilde{\theta})] \\ &\ge -\lambda_n I(\theta_n^*) + \lambda_n I(\tilde{\theta}) - [l(\theta_n^*) - l(\tilde{\theta})] - d_0\zeta_n^*/b \cdot \bar{\lambda}_{n,0}^A \\ &\ge -\lambda_n I(\theta_n^*) + \lambda_n I(\tilde{\theta}) - \mathcal{E}^* + \tilde{\mathcal{E}} - d_0\bar{\lambda}_{n,0}^A\zeta_n^*/b. \end{aligned} \tag{2.14}$$

By (2.13) we know that $R_{\tilde{\theta}}(d_0\zeta_n^*/b)$ is bounded by $d_0\bar{\lambda}_{n,0}^B\zeta_n^*/b$ with probability at least $1 - \frac{3}{10}W^2\exp(-n\bar{a}_n^2 r_1^2) - 2\exp(-n\xi^2/2)$, then we have

$$\tilde{\mathcal{E}} + \lambda_n I(\tilde{\theta}) \le \bar{\lambda}_{n,0}^B d_0\zeta_n^*/b + \mathcal{E}^* + \lambda_n I(\theta_n^*) + \bar{\lambda}_{n,0}^A d_0\zeta_n^*/b.$$

Since $I(\tilde{\theta}) = I_1(\tilde{\theta}) + I_2(\tilde{\theta})$ and $I(\theta_n^*) = I_1(\theta_n^*)$, using the triangular inequality, we obtain

$$\begin{aligned} &\tilde{\mathcal{E}} + (1+b)\bar{\lambda}_{n,0}I_2(\tilde{\theta}) \\ &\qquad \le \bar{\lambda}_{n,0}d_0\zeta_n^*/b + \mathcal{E}^* + (1+b)\bar{\lambda}_{n,0}I_1(\theta_n^*) - (1+b)\bar{\lambda}_{n,0}I_1(\tilde{\theta}) \\ &\qquad \le \bar{\lambda}_{n,0}d_0\zeta_n^*/b + \mathcal{E}^* + (1+b)\bar{\lambda}_{n,0}I_1(\tilde{\theta} - \theta_n^*). \end{aligned} \tag{2.15}$$

Adding $(1+b)\bar{\lambda}_{n,0}I_1(\tilde{\theta} - \theta_n^*)$ to both sides and from Lemma II.5,

$$\begin{aligned} \tilde{\mathcal{E}} + (1+b)\bar{\lambda}_{n,0}I(\tilde{\theta} - \theta_n^*) &\le \bar{\lambda}_{n,0}d_0\frac{\zeta_n^*}{b} + \mathcal{E}^* + 2(1+b)\bar{\lambda}_{n,0}I_1(\tilde{\theta} - \theta_n^*) \\ &\le (\bar{\lambda}_{n,0}d_0 + b\bar{\lambda}_{n,0})\frac{\zeta_n^*}{b} + \delta\tilde{\mathcal{E}} \\ &= (d_0 + b)\bar{\lambda}_{n,0}\frac{\zeta_n^*}{b} + \delta\tilde{\mathcal{E}}. \end{aligned}$$

Because $0 < \delta < 1$, it follows that

$$I(\tilde{\theta} - \theta_n^*) \le \frac{d_0 + b}{1 + b} \frac{\zeta_n^*}{b}.$$

Hence,

$$P\left(\left\{ I(\tilde{\theta} - \theta_n^*) \le d_0 \frac{\zeta_n^*}{b} \right\} \cap \left\{ Z_{\tilde{\theta}}(d_0 \zeta_n^*/b) \le d_0 \bar{\lambda}_{n,0}^A \frac{\zeta_n^*}{b} \right\} \cap \left\{ R_{\tilde{\theta}}(d_0 \zeta_n^*/b) \le d_0 \bar{\lambda}_{n,0}^B \frac{\zeta_n^*}{b} \right\} \right)$$
$$\le \; P\left(I(\tilde{\theta} - \theta_n^*) \le \frac{d_0 + b}{1 + b} \frac{\zeta_n^*}{b} \right),$$

which yields the desired result. □

Corollary II.3. *Suppose Conditions II.I(b, δ) and II.II(b, δ, d) are met. Consider any random $\tilde{\theta} \in \Theta$ with $l_n(\tilde{\theta}) + \lambda_n I(\tilde{\theta}) \le l_n(\theta_n^*) + \lambda_n I(\theta_n^*)$. Let $1 < d_0 \le d_b$. It holds that*

$$P\left(I(\tilde{\theta} - \theta_n^*) \le d_0 \frac{\zeta_n^*}{b} \right)$$
$$\le P\left(I(\tilde{\theta} - \theta_n^*) \le \left[1 + (d_0 - 1)(1 + b)^{-N} \right] \frac{\zeta_n^*}{b} \right)$$
$$+ \; N \left\{ \left(1 + \frac{3}{10} W^2 \right) \exp\left(-n \bar{a}_n^2 r_1^2 \right) + 2 \exp(-n \xi^2/2) \right\}.$$

Proof. Repeat Lemma II.6 N times. □

Lemma II.7. *Suppose Conditions II.I(b, δ) and II.II(b, δ, d) hold. If $\tilde{\theta}_s = s\hat{\theta}_n + (1 - s)\theta_n^*$, where*

$$s = \frac{d\zeta_n^*}{d\zeta_n^* + bI(\hat{\theta}_n - \theta_n^*)},$$

then for any integer N, with probability at least

$$1 - N \left\{ \left(1 + \frac{3}{10} W^2 \right) \exp\left(-n \bar{a}_n^2 r_1^2 \right) + 2 \exp\left(-n \xi^2/2 \right) \right\},$$

we have

$$I(\tilde{\theta}_s - \theta_n^*) \leq \left(1 + (d-1)(1+b)^{-N}\right) \frac{\zeta_n^*}{b}.$$

Proof. Since the negative log partial likelihood $l_n(\theta)$ and the lasso penalty are both convex with respect to θ, applying Corollary II.3, we obtain the above inequality. This proof is similar to the proof of Lemma A.6 in van de Geer (2008). □

Lemma II.8. *Suppose Conditions II.I(b, δ) and II.II(b, δ, d) are met. Let $N_1 \in \mathbf{N} := \{1, 2, \dots\}$ and $N_2 \in \mathbf{N} \cup \{0\}$. With $\delta_1 = (1+b)^{-N_1}$ and $\delta_2 = (1+b)^{-N_2}$, for any n, with probability at least*

$$1 - (N_1 + N_2)\left\{\left(1 + \frac{3}{10}W^2\right)\exp\left(-n\bar{a}_n^2 r_1^2\right) + 2\exp(-n\xi^2/2)\right\},$$

we have

$$I(\hat{\theta}_n - \theta_n^*) \leq d(\delta_1, \delta_2)\frac{\zeta_n^*}{b},$$

where

$$d(\delta_1, \delta_2) = 1 + \frac{1 + (d^2-1)\delta_1}{(d-1)(1-\delta_1)}\delta_2.$$

Proof. The proof is the same as that of Lemma A.7 in van de Geer (2008), with a slightly different probability bound. □

2.4.2 Proof of Theorem II.1

Proof. The proof follows the same ideas in the proof of Theorem A.4 in van de Geer (2008), with exceptions of pointwise arguments and slightly different probability bounds. Since this is our main result, we provide a detailed proof here despite the amount of overlaps.

Define $\hat{\mathcal{E}} := \mathcal{E}(f_{\hat{\theta}_n})$ and $\mathcal{E}^* := \mathcal{E}(f_{\theta_n^*})$; use the notation $I_1(\theta) := I_1(\theta|\theta_n^*)$ and $I_2(\theta) := I_2(\theta|\theta_n^*)$; set $c := \delta b/(1-\delta^2)$. Consider the cases (a) $c < d(\delta_1, \delta_2)$ and (b) $c \geq d(\delta_1, \delta_2)$.

(a) $c < d(\delta_1, \delta_2)$. Let J be an integer satisfying $(1+b)^{J-1}c \le d(\delta_1, \delta_2)$ and $(1+b)^J c > d(\delta_1, \delta_2)$. We consider the cases (a1) $c\zeta_n^*/b < I(\hat{\theta}_n - \theta_n^*) \le d(\delta_1, \delta_2)\zeta_n^*/b$ and (a2) $I(\hat{\theta}_n - \theta_n^*) \le c\zeta_n^*/b$.

(a1) If $c\zeta_n^*/b < I(\hat{\theta}_n - \theta_n^*) \le d(\delta_1, \delta_2)\zeta_n^*/b$, then

$$(1+b)^{j-1}c\frac{\zeta_n^*}{b} < I(\hat{\theta}_n - \theta_n^*) \le (1+b)^j c\frac{\zeta_n^*}{b}$$

for some $j \in \{1, \cdots, J\}$. Let $d_0 = c(1+b)^{j-1} \le d(\delta_1, \delta_2) \le d_b$. From Corollary II.1, with probability at least $1 - \exp(-n\bar{a}_n^2 r_1^2)$ we have $Z_{\hat{\theta}_n}((1+b)d_0\zeta_n^*/b) \le (1+b)d_0\bar{\lambda}_{n,0}^A\zeta_n^*/b$. Since $l_n(\hat{\theta}_n) + \lambda_n I(\hat{\theta}_n) \le l_n(\theta_n^*) + \lambda_n I(\theta_n^*)$, from (2.14) we have

$$\hat{\mathcal{E}} + \lambda_n I(\hat{\theta}_n) \le R_{\hat{\theta}_n}\left((1+b)d_0\frac{\zeta_n^*}{b}\right) + \mathcal{E}^* + \lambda_n I(\theta_n^*) + (1+b)\bar{\lambda}_{n,0}^A d_0\frac{\zeta_n^*}{b}.$$

By (2.13), $R_{\hat{\theta}_n}((1+b)d_0\zeta_n^*/b)$ is bounded by $(1+b)\bar{\lambda}_{n,0}^B d_0\zeta_n^*/b$ with probability at least

$$1 - \frac{3}{10}W^2 \exp\left(-n\bar{a}_n^2 r_1^2\right) - 2\exp\left(-n\xi^2/2\right).$$

Then we have

$$\begin{aligned}\hat{\mathcal{E}} + (1+b)\bar{\lambda}_{n,0}I(\hat{\theta}_n) &\le (1+b)\bar{\lambda}_{n,0}^B d_0\frac{\zeta_n^*}{b} + \mathcal{E}^* + (1+b)\bar{\lambda}_{n,0}I(\theta_n^*) + (1+b)\bar{\lambda}_{n,0}^A d_0\frac{\zeta_n^*}{b} \\ &\le (1+b)\bar{\lambda}_{n,0}I(\hat{\theta}_n - \theta_n^*) + \mathcal{E}^* + (1+b)\bar{\lambda}_{n,0}I(\theta_n^*).\end{aligned}$$

Since $I(\hat{\theta}_n) = I_1(\hat{\theta}_n) + I_2(\hat{\theta}_n)$, $I(\hat{\theta}_n - \theta_n^*) = I_1(\hat{\theta}_n - \theta_n^*) + I_2(\hat{\theta}_n)$, and $I(\theta_n^*) = I_1(\theta_n^*)$, by triangular inequality we obtain $\hat{\mathcal{E}} \le 2(1+b)\bar{\lambda}_{n,0}I_1(\hat{\theta}_n - \theta_n^*) + \mathcal{E}^*$. From Lemma II.5, $\hat{\mathcal{E}} \le \delta\hat{\mathcal{E}} + \epsilon_n^* - \mathcal{E}^* + \mathcal{E}^* = \delta\hat{\mathcal{E}} + \epsilon_n^*$. Hence, $\hat{\mathcal{E}} \le \epsilon_n^*/(1-\delta)$.

(a2) If $I(\hat{\theta}_n - \theta_n^*) \le c\zeta_n^*/b$, from (2.15) with $d_0 = c$, with probability at least

$$1 - \left\{\left(1 + \frac{3}{10}W^2\right)\exp\left(-n\bar{a}_n^2 r_1^2\right) + 2\exp(-n\xi^2/2)\right\},$$

we have

$$\hat{\mathcal{E}} + (1+b)\bar{\lambda}_{n,0} I(\hat{\theta}_n) \le \frac{\delta}{1-\delta^2}\bar{\lambda}_{n,0}\zeta_n^* + \mathcal{E}^* + (1+b)\bar{\lambda}_{n,0} I(\theta_n^*).$$

By the triangular inequality and Lemma II.5,

$$\begin{aligned}
\hat{\mathcal{E}} &\le \frac{\delta}{1-\delta^2}\bar{\lambda}_{n,0}\zeta_n^* + \mathcal{E}^* + (1+b)\bar{\lambda}_{n,0} I_1(\hat{\theta}_n - \theta_n^*) \\
&\le \frac{\delta}{1-\delta^2}\bar{\lambda}_{n,0}\frac{\epsilon_n^*}{\bar{\lambda}_{n,0}} + \mathcal{E}^* + \frac{\delta}{2}\hat{\mathcal{E}} + \frac{1}{2}\epsilon_n^* - \frac{1}{2}\mathcal{E}^* \\
&= \left(\frac{\delta}{1-\delta^2} + \frac{1}{2}\right)\epsilon_n^* + \frac{1}{2}\mathcal{E}^* + \frac{\delta}{2}\hat{\mathcal{E}} \\
&\le \left(\frac{\delta}{1-\delta^2} + \frac{1}{2}\right)\epsilon_n^* + \frac{1}{2(1+\delta)}\epsilon_n^* + \frac{\delta}{2}\hat{\mathcal{E}}.
\end{aligned}$$

Hence,

$$\hat{\mathcal{E}} \le \frac{2}{2-\delta}\left[\frac{\delta}{1-\delta^2} + \frac{1}{2} + \frac{1}{2(1+\delta)}\right]\epsilon_n^* = \frac{1}{1-\delta}\epsilon_n^*.$$

Furthermore, by Lemma II.8, we have with probability at least

$$1 - (N_1 + N_2)\left\{\left(1 + \frac{3}{10}W^2\right)\exp\left(-n\bar{a}_n^2 r_1^2\right) + 2\exp\left(-n\xi^2/2\right)\right\}$$

that $I(\hat{\theta}_n - \theta_n^*) \le d(\delta_1, \delta_2)\frac{\zeta_n^*}{b}$, where

$$N_1 = \log_{1+b}\left(\frac{1}{\delta_1}\right), \; N_2 = \log_{1+b}\left(\frac{1}{\delta_2}\right).$$

(b) $c \ge d(\delta_1, \delta_2)$. On the set where $I(\hat{\theta}_n - \theta_n^*) \le d(\delta_1, \delta_2)\zeta_n^*/b$, from equation (2.15) we have with probability at least

$$1 - \left\{\left(1 + \frac{3}{10}W^2\right)\exp\left(-n\bar{a}_n^2 r_1^2\right) + 2\exp\left(-n\xi^2/2\right)\right\}$$

that

$$\begin{aligned}\hat{\mathcal{E}} + (1+b)\bar{\lambda}_{n,0} I(\hat{\theta}_n) &\leq \bar{\lambda}_{n,0} d(\delta_1,\delta_2)\frac{\zeta_n^*}{b} + \mathcal{E}^* + (1+b)\bar{\lambda}_{n,0} I(\theta_n^*) \\ &\leq \frac{\delta}{1-\delta^2}\bar{\lambda}_{n,0}\zeta_n^* + \mathcal{E}^* + (1+b)\bar{\lambda}_{n,0} I(\theta_n^*),\end{aligned}$$

which is the same as (a2) and leads to the same result.

To summarize, let

$$A = \left\{\hat{\mathcal{E}} \leq \frac{1}{1-\delta}\epsilon_n^*\right\}, \quad B = \left\{I(\hat{\theta}_n - \theta_n^*) \leq d(\delta_1,\delta_2)\frac{\zeta_n^*}{b}\right\}.$$

Note that

$$J + 1 \leq \log_{1+b}\left(\frac{(1+b)^2 d(\delta_1,\delta_2)}{c}\right).$$

Under case (a), we have

$$\begin{aligned}P(A\cap B) &= P(\text{a1}) - P(A^c \cap \text{a1}) + P(\text{a2}) - P(A^c\cap \text{a2}) \\ &\geq P(\text{a1}) - J\left\{\left(1+\frac{3}{10}W^2\right)\exp\left(-n\bar{a}_n^2 r_1^2\right) + 2\exp(-n\xi^2/2)\right\} \\ &\quad + P(\text{a2}) - \left\{\left(1+\frac{3}{10}W^2\right)\exp(-n\bar{a}_n^2 r_1^2) + 2\exp\left(-n\xi^2/2\right)\right\} \\ &= P(B) - (J+1)\left\{\left(1+\frac{3}{10}W^2\right)\exp\left(-n\bar{a}_n^2 r_1^2\right) + 2\exp\left(-n\xi^2/2\right)\right\} \\ &\geq 1 - (N_1+N_2+J+1)\left\{\left(1+\frac{3}{10}W^2\right)\exp\left(-n\bar{a}_n^2 r_1^2\right) 2\exp\left(-n\xi^2/2\right)\right\} \\ &\geq 1 - \log_{1+b}\left\{\frac{(1+b)^2}{\delta_1\delta_2}\cdot\frac{d(\delta_1,\delta_2)(1-\delta^2)}{\delta b}\right\} \\ &\qquad \left\{\left(1+\frac{3}{10}W^2\right)\exp\left(-n\bar{a}_n^2 r_1^2\right) + 2\exp(-n\xi^2/2)\right\}.\end{aligned}$$

Under case (b),

$$
\begin{aligned}
P(A \cap B) &= P(B) - P(A^c \cap B) \\
&\geq P(B) - \left\{ \left(1 + \frac{3}{10} W^2\right) \exp\left(-n\bar{a}_n^2 r_1^2\right) + 2\exp\left(-n\xi^2/2\right) \right\} \\
&\geq 1 - (N_1 + N_2 + 2) \left\{ \left(1 + \frac{3}{10} W^2\right) \exp\left(-n\bar{a}_n^2 r_1^2\right) + 2\exp\left(-n\xi^2/2\right) \right\} \\
&= 1 - \log_{1+b} \left\{ \frac{(1+b)^2}{\delta_1 \delta_2} \right\} \\
&\qquad \times \left\{ \left(1 + \frac{3}{10} W^2\right) \exp\left(-n\bar{a}_n^2 r_1^2\right) + 2\exp(-n\xi^2/2) \right\}.
\end{aligned}
$$

We thus obtain the desired result. $\square$

CHAPTER III

Semiparametric Approach for Regression with Covariate Subject to Limit of Detection

3.1 Introduction

Detection limit is a threshold below which measured values are not considered significantly different from background noise (Helsel, 2005). Hence, values measured below this threshold are unreliable. In environmental epidemiology, particularly exposure analysis, when exposure levels are low, measurement of chemicals has a large percentage falling below the limit of detection L due to inadequate instrument sensitivity. For example, in the National Health and Nutrition Examination Survey (Crainiceanu et al., 2008), the limit of detection for blood cadmium was 2.67 nmol/L in Nutrition Examination Survey 1999-2002 and 1.78 nmol/L in Nutrition Examination Survey 2003-2004, and 21.6% and 13.4% of the subjects had blood cadmium levels below limit of detection, respectively. Similar limit of detection issue exists in other studies, for example, the Diabetes Prevention Program, where of the 301 eligible participants 66 had a testosterone level below a detection limit of 8.0 ng/dl (Kim et al., 2012). In this article, we consider an analysis for the Michigan Bone Health and Metabolism Study, which examines the relationship between anti-Mullerian hormone and time to the final menstrual period (Sowers et al., 2008). The data set consists of fifty women that had six consecutive annual visits. The levels of anti-Mullerian hormone were recorded during the time period with a limit of detection of 0.05ng/ml that is the smallest available anti-Mullerian hormone value in Sowers et al. (2008). As a result, 6% of the 50 study participants had anti-Mullerian hormone below limit of detection in the first visit and the percentage increases in later visits, with 66% below limit of detection in the last visit. For illustration purpose, we focus on visit 3 where 18% of participants had anti-Mullerian hormone measures below limit of detection. Several other covariates were also recorded, including age, body mass index and Follicle-Stimulating Hormone.

A variety of statistical tools have been developed to deal with the problem of response variable subject to limit of detection, see for examples Thompsonand and Nelson (2003), Lubin et al. (2004) and Helsel (2005). Limit of detection is in fact left censoring, which can be easily transformed to right censoring by multiplying the variable by -1. As a result, standard semiparametric survival models can be applied.

Statistical methods for regression models with a covariate subject to limit of detection, however, are yet to be thoroughly studied (Schisterman and Little, 2010), even though many ad hoc methods have been implemented in practice. The complete case analysis, of simply eliminating observations with values below limit of detection, yields consistent estimates of the regression coefficients (Nie et al., 2010; Little and Rubin, 2002), but loses efficiency. Substitution methods are frequently used, see for examples, Hornung and Reed (1990), Moulton et al. (2002), Richardson and Ciampi (2003), and Nie et al. (2010), among many others, where the values below limit of detection L are substituted by L, or $L/\sqrt{2}$, or zero, or $E(X|X \leq L)$ that is obtained from an assumed distribution of X. These methods are easily implementable, but found to be inappropriate and can yield large biases, see for example Helsel (2006).

Another widely used method is the maximum likelihood estimation based on a parametric distributional assumption to the unobservable tail probability of the covariate that is subject to limit of detection. For examples, Cole et al. (2009) and Nie et al. (2010) considered logistic and linear regression, respectively, based on a normal distribution for the tail probability of the covariate subject to limit of detection; D'Angelo and Weissfeld (2008) applied this approach to the Cox regression. In practice, however, the underlying covariate distribution is unknown. Both Lynn (2001) and Nie et al. (2010) noted that a parametric assumption can yield large bias if misspecified and argued that such an approach should not be attempted. Nie et al. (2010) recommended the complete case analysis despite the fact that simply dropping data below the limit of detection can lose a significant amount of information.

To obtain more efficient and yet robust results, we propose a semiparametric likelihood-based approach to fit generalized linear models with covariate subject to limit of detection. The tail distribution of the covariate beyond its limit of detection is estimated from a semiparametric accelerated failure model, conditional on all the fully observed covariates. Model checking can be done using martingale residuals for semiparametric accelerated failure time models. The proposed method is shown to be consistent and asymptotically normal, and outperforms existing methods in simulations. The proofs of the asymptotic properties rely heavily on empirical process theory.

3.2 A semiparametric approach

For a single observation, denote the response variable by Y, the covariate subject to limit of detection by Z, and the fully observed covariates by $X = (X_1, \ldots, X_p)'$, where p is the number of fully observed covariates. For simplicity, we only consider one covariate that is subject to limit of detection. Consider a generalized linear model with

$$E(Y) = \mu = g^{-1}(D'\theta), \tag{3.1}$$

where g is the link function, $D'\theta$ is the linear predictor with $D = (1, X', Z)'$ and $\theta = (\beta', \gamma)'$, here β is a $(p+1)$-dimensional vector and γ is a scalar. The variance V is typically a function of the mean denoted by

$$var(Y) = W(\mu) = W\{g^{-1}(D'\theta)\}.$$

We consider the exponential dispersion family in the natural form (Agresti, 2002; McCullagh and Nelder , 1989) given (Z, X)

$$f_{\varpi,\phi}(Y|Z,X) = \exp\left\{\frac{Y\varpi - b(\varpi)}{a(\phi)} + c(Y,\phi)\right\}, \tag{3.2}$$

where ϕ is the dispersion parameter and ϖ is the natural parameter. We have $\mu = E(Y) = \dot{b}(\varpi)$, and $var(Y) = \ddot{b}(\varpi)a(\phi)$, where $\dot{b}$ is the first derivative of b and $\ddot{b}$ is the second derivative of b.

The actual value of Z is not observable when $Z < L$, where the constant L denotes the limit of detection, which is an example of left-censoring. In practice Z is a concentration measure of certain substance and thus non-negative. Consider a monotone decreasing transformation h that yields $Z = h(T)$, for example, $h(T) = \exp(-T)$. Denote $D(T) = (1, X', h(T))'$. If $T \le C = h^{-1}(L)$, then T is observed; otherwise T is right-censored by C. We denote the observed value by $V = \min(T, C)$ and the censoring indicator by $\Delta = I(T \le C)$.

The proposed methodology works for a broad family of link functions defined by the regularity conditions given in Subsection 3.6.1. For notational simplicity, we present the main material using canonical link function g, where $g = (\dot{b})^{-1}$. Then, when T is observed, model (3.2) becomes

$$f_{\theta,\phi}(Y|T,X) = \exp\left\{\frac{YD'(T)\theta - b(D'(T)\theta)}{a(\phi)} + c(Y,\phi)\right\}. \tag{3.3}$$

Denote the conditional cumulative distribution function of T given X by $F_1(t|X)$ with density $f_1(t|X)$. The likelihood function for the observed data (V, Δ, Y, X) can be factorized into

$$f(V, \Delta, Y, X) = f_2(V, \Delta|Y, X)f_3(Y|X)f_4(X),$$

where f denotes the joint density of (V, Δ, Y, X), f_2 denotes conditional density of (V, Δ) given (Y, X), f_3 denotes conditional density of Y given X, and f_4 denotes marginal density of X. Going through conditional arguments using the Bayes' rule and dropping $f_4(X)$, we obtain the likelihood function

$$L(V, \Delta, Y, X) = \{f_{\theta,\phi}(Y|T, X)f_1(T|X)\}^{\Delta} \left\{ \int_C^{\infty} f_{\theta,\phi}(Y|t, X)dF_1(t|X) \right\}^{1-\Delta}, \qquad (3.4)$$

where only $f_{\theta,\phi}$ contains the parameter of interest θ, whereas f_1 is a nuisance parameter in addition to ϕ.

There are two parts in (4.4): (i) $\{f_{\theta,\phi}(Y|T, X)f_1(T|X)\}^{\Delta}$ for fully observed subject, and (ii) $\left\{\int_C^{\infty} f_{\theta,\phi}(Y|t, X)dF_1(t|X)\right\}^{1-\Delta}$ for subject with covariate below limit of detection. Complete case analysis is only based on the first part and, although it yields a consistent estimate of θ, it clearly loses efficiency. We see from the second part of (4.4) that the efficiency gain comparing to the complete case analysis depends on how well we can recover the right tail of the conditional distribution $F_1(t|X)$ beyond C. Parametric models for $F_1(t|X)$ are often considered in the literature, see Nie et al. (2010), but it may suffer from model misspecification. The nonparametric method degenerates to the complete case analysis because there is no actual observation beyond censoring time C. We consider a semiparametric approach that allows reliable extrapolation beyond C and is robust against any parametric assumption.

Among all the commonly used semiparametric models for right-censored data, only the accelerated failure time model allows extrapolation beyond C, and model checking can be done by visualizing the cumulative sums of the martingale-based residuals (Lin et al., 1993, 1996; Peng and Fine, 2006). We hence propose a semi-parametric accelerated failure time model for the transformed covariate subject to limit of detection given by

$$T = X'\alpha + \varsigma, \tag{3.5}$$

where ς follows some unknown distribution, denoted by η, and is independent of X. We only consider a fixed h for T in this article. More flexible transformation, for example, the Box-Cox transformation (Box and Cox, 1964; Foster et al., 2001; Cai et al., 2005), is worth further investigation. Note that X appears in both models (3.1) and (3.5), but it may refer to different forms of covariates in these models. For example, X_1 is a covariate in (3.1) whereas X_1^2 is a covariate in (3.5). We use the same X to denote all fully observed covariates for notational simplicity. The log-likelihood function then becomes

$$\begin{aligned} \log L &= \Delta \log f_{\theta,\phi}(Y|T,X) + \Delta \log \dot{\eta}(T - X'\alpha) \\ &\quad + (1-\Delta)\log\left\{ \int_{C-X'\alpha}^{\tau} f_{\theta,\phi}(Y|t + X'\alpha, X)\, d\eta(t) \right\}, \end{aligned} \tag{3.6}$$

where τ is a truncation time at the residual scale defined in Condition III.4 in Subsection 3.6.1.

3.3 The pseudo-likelihood method

The log likelihood function (3.6) involves an unknown distribution function η and its derivative, hence a maximum likelihood estimation, if it exists, can be complicated. We propose a tractable two-stage pseudo-likelihood approach in which the nuisance parameters (ϕ, α, η) are estimated in stage 1 and the parameter of interest θ is then estimated by maximizing the data version of (3.6) in stage 2 with nuisance parameters replaced by their estimators obtained in stage 1 before maximization. Details are given below:

Stage 1. Nuisance parameter estimation. Dispersion parameter ϕ is estimated by the complete case analysis of the generalized linear model (3.2); the accelerated failure time model regression coefficient α is estimated by either the rank based methods, see Wei et al. (1990), Jin et al. (2003), Nan et al. (2009) or the sieve maximum likelihood method, see Ding and Nan (2011); and the accelerated failure time model error distribution η is estimated by the Kaplan-Meier estimator from the censored residuals.

Stage 2. Pseudo-likelihood estimation of θ. Replacing (ϕ, α, η) by their Stage 1 estimates $(\hat{\phi}_n, \hat{\alpha}_n, \hat{\eta}_{n,\hat{\alpha}_n})$ in the log likelihood function yields the following log pseudo-likelihood function for a random sample of n observations:

$$\begin{aligned} pl_n(\theta) \;=\; & \frac{1}{n}\sum_{i=1}^{n}\Bigg\{\Delta_i \log f_{\theta,\hat{\phi}_n}(Y_i|X_i,T_i) \\ & +(1-\Delta_i)\log \int_{C-X_i'\hat{\alpha}_n}^{\tau} f_{\theta,\hat{\phi}_n}(Y_i|X_i,t+X_i'\hat{\alpha}_n)\,d\hat{\eta}_{n,\hat{\alpha}_n}(t)\Bigg\}, \qquad (3.7) \end{aligned}$$

where

$$f_{\theta,\hat{\phi}_n}(Y_i|T_i,X_i) = \exp\left[\frac{Y_i\{D_i'(T_i)\theta\} - b\{D_i'(T_i)\theta\}}{a(\hat{\phi}_n)} + c(Y_i,\hat{\phi}_n)\right].$$

Note that the term $\Delta \log \dot{\eta}(T)$ in (3.6) is dropped because it does not involve θ. We maximize (3.7) by setting its derivative to be zero and then solving the equation for the pseudo-likelihood estimator $\hat{\theta}_n$.

Since $\hat{\theta}_n$ is obtained by solving an estimating equation, its asymptotic properties can be obtained from Z-estimation theory. It can be shown that all the estimates obtained in Stage 1 have desirable statistical properties for Stage 2. In particular,

$\hat{\phi}_n$ obtained from the complete case analysis is $n^{1/2}-$consistent by Little and Rubin (2002); $\hat{\alpha}_n$ is $n^{1/2}-$consistent by Nan et al. (2009) or Ding and Nan (2011); and $\hat{\eta}_{n,\hat{\alpha}_n}$ is also $n^{1/2}-$consistent in a finite interval, and its proof is provided in the Appendices.

3.4 Asymptotic properties

Define a random map as follows

$$\Psi_{\theta,n}(\phi,\alpha,\eta)=\frac{1}{n}\sum_{i=1}^{n}\psi_\theta(Y_i,X_i,V_i,\Delta_i;\phi,\alpha,\eta), \tag{3.8}$$

where

$$\begin{aligned}
&\psi_\theta(Y,X,V,\Delta;\phi,\alpha,\eta)\\
&\quad=\Delta\{Y-\dot{b}(D'(T)\theta)\}D(T)+(1-\Delta)\left\{\int_{C-X'\alpha}^{\tau} f_{\theta,\phi}(Y|t+X'\alpha,X)d\eta(t)\right\}^{-1}\\
&\qquad\times\int_{C-X'\alpha}^{\tau} f_{\theta,\phi}(Y|t+X'\alpha,X)\{Y-\dot{b}(D'(t+X'\alpha)\theta)\}D(t+X'\alpha)d\eta(t),
\end{aligned}$$

which is the derivative of (3.6) with respect to θ. Then with (ϕ,α,η) replaced by $(\hat{\phi}_n,\hat{\alpha}_n,\hat{\eta}_{n,\hat{\alpha}_n})$ in (3.8), $\Psi_{\theta,n}(\hat{\phi}_n,\hat{\alpha}_n,\hat{\eta}_{n,\hat{\alpha}_n})=0$ becomes the pseudo-likelihood estimating equation for θ, and its solution $\hat{\theta}_n$ is called the pseudo-likelihood estimator.

A set of regularity conditions is introduced in Subsection 3.6.1. Some conditions are commonly assumed for the accelerated failure time models, and other conditions are for the generalized linear models, which are easily verifiable for linear, logistic and Poisson regression models. We then have the following asymptotic results for $\hat{\theta}_n$.

Theorem III.1. *(Consistency and asymptotic normality.) Denote the true value of θ by θ_0. Suppose all the regularity conditions given in Subsection 3.6.1 hold. Then for the two-stage pseudo-likelihood estimator $\hat{\theta}_n$ satisfying $\Psi_{\hat{\theta}_n,n}(\hat{\phi}_n,\hat{\alpha}_n,\hat{\eta}_{n,\hat{\alpha}_n})=0$, we have: (i) $\hat{\theta}_n$ converges in outer probability to θ_0, and (ii) $n^{1/2}(\hat{\theta}_n-\theta_0)$ converges* *weakly to a mean zero normal random variable with variance $A^{-1}BA^{-1}$, where A and B are provided in Subsection 3.6.4.*

Because the asymptotic variance of $\hat{\theta}_n$ has a very complicated expression that prohibits the direct calculation of its estimate from observed data, we recommend using the bootstrap variance estimator.

The proof of Theorem III.1 is based on the general Z-estimation theory of Nan and Wellner (2013). Define a deterministic function

$$\Psi_\theta(\phi, \alpha, \eta) = E\Big\{\psi_\theta(Y, X, V, \Delta; \phi, \alpha, \eta)\Big\}, \tag{3.9}$$

and denote the true values of (ϕ, α, η) by $(\phi_0, \alpha_0, \eta_0)$. We can show that $\Psi_{\theta,n}(\hat{\phi}_n, \hat{\alpha}_n, \hat{\eta}_{n,\hat{\alpha}_n})$ converges uniformly to $\Psi_\theta(\phi_0, \alpha_0, \eta_0)$ as $n \to \infty$. Then the consistency is achieved given that θ_0 is the unique solution of $\Psi_\theta(\phi_0, \alpha_0, \eta_0) = 0$. The asymptotic normality is derived by showing the asymptotic linear representation of $n^{1/2}(\hat{\theta}_n - \theta_0)$. The detailed proofs rely heavily on empirical process theory and can be found in the Appendices, where we only provide the analytic form of the asymptotic variance for the Gehan weighted estimate of α. The analytic forms of the asymptotic variance for other rank based estimates and the sieve maximum likelihood estimate as well can be obtained similarly.

3.5 Numerical results

3.5.1 Simulations

We conduct simulations to investigate the finite sample performance of the proposed method. Simulation data sets are generated from the generalized linear model

$$g(E(Y)) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \gamma Z,$$

where $\beta_0 = -1$, $\beta_1 = 0.5$, $\beta_2 = -1$, $\gamma = 2$, and g is chosen to be the canonical link function for normal, bernoulli and poisson distributions, respectively. The normal error variance is chosen to be 1 for the linear regression model. The three covariates are: $X_1 \sim$ Bernoulli(0.5), X_2 is normal with mean 1 and standard deviation 1 truncated at ± 3, and $Z = \exp(-T)$ is generated from the following linear model

$$T = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \varsigma,$$

where $\alpha_0 = 0.25$, $\alpha_1 = 0.25$, $\alpha_2 = -0.5$, $\varsigma \sim 0.5N(0, 1/8^2) + 0.5N(0.5, 1/10^2)$, and T is subject to right-censoring. The limit of detection L for covariate Z is chosen to yield 30% censoring.

We simulate 1000 replications for each scenario, and compare the biases and variances of the proposed method with full data analysis, complete case analysis, and four different substitution methods. The full data analysis represents the case that all data are available, in other words, there is no limit of detection, which serves as a benchmark. For linear regression, we conduct simulations with three different sample sizes: 50, 200 and 400, where the sample size of 50 mimics the Michigan Bone Health and Metabolism anti-Mullerian hormone study. For logistic and Poisson regression models, we only consider sample sizes of 200 and 400. The four substitution methods for $Z < L$ are: (i) replacing Z by L, (ii) replacing Z by $L/\sqrt{2}$, (iii) replacing Z by zero, and (iv) replacing Z by $E(Z|Z < L)$. We only report biases for these substitution method. For the proposed two-stage method, we report the 90% coverage proportions for which the variances are obtained from 200 bootstrap samples. The results are presented in Table 3.1-3.3.

The results suggest that all the substitution methods yield biased estimates, including substituting Z by $E(Z|Z < L)$. The biases for the proposed two-stage method are minimal, which are comparable to both the full data analysis and the complete case analysis. Clearly, the proposed method is much more efficient than the complete case analysis, and the bootstrap method performs well in estimating the variance, which yields reasonable coverage rate of the confidence intervals for all considered sample sizes.

3.5.2 The hormone data analysis

We revisit the anti-Mullerian hormone data set analyzed by Sowers et al. (2008) and Sowers et al. (2010) as an illustrative example for the regression with a covariate subject to limit of detection. In particular, we focus on the effect of left-censored anti-Mullerian hormone on the time to final menstrual period.

The data set contains a subsample of 50 study participants of the Michigan Bone Health and Metabolism Study (Sowers et al., 2010). For each woman in this subsample, blood samples collected at six consecutive annual visits before her subsequent final menstrual period were assayed for hormone measures. The limit of detection was taken to be 0.05ng/ml for anti-Mullerian hormone. The percentage of subjects below this limit of detection increases over time and varies from 6% to 66%.

For illustration purpose, we focus on the 3rd visit where 18% subjects had anti-Mullerian hormone below limit of detection. Age, body mass index and follicle-stimulating hormone, all measured at visit 3, are used as covariates to fit the accelerated failure time model for $-\log(\text{AMH})$, here AMH stands for anti-Mullerian hormone. The final linear model for the time to final menstrual period only includes age and $\log(\text{AMH})$ as covariates. Table 3.4 shows the regression coefficient estimates, where we see that the proposed two-stage method yields similar point estimates with smaller variances comparing to the complete case analysis, indicating the efficiency gain of the proposed method. Figure 3.1 shows the plots of 50 realizations from the distributions of the score processes with dotted lines. The observed score processes are presented with solid lines which randomly fluctuated around zero. From Figure

Table 3.1: Simulation results for linear regression.

	Sample size		$\beta_0 = -1$	$\beta_1 = 0.5$	$\beta_2 = -1$	$\gamma = 2$
Full data	50	bias	0.020	-0.008	0.010	-0.020
		var	1.860	0.230	0.166	1.684
Two-stage		bias	0.024	-0.006	0.009	-0.023
		var	2.059	0.248	0.188	1.883
		bootstrap var	2.275	0.262	0.206	2.091
		90% coverage rate (%)	90.1	89.1	91.1	90.3
		95% coverage rate (%)	93.9	94.0	94.7	94.2
Complete case		bias	0.042	-0.013	0.014	-0.038
		var	2.571	0.398	0.325	2.487
L		bias	0.443	-0.232	0.236	-0.524
$L/\sqrt{2}$		bias	0.740	-0.149	0.160	-0.657
Zero		bias	1.840	-0.423	0.433	-1.689
$E(Z\|Z < L)$		bias	0.460	-0.146	0.154	-0.456
Full data	200	bias	-0.030	0.011	-0.006	0.024
		var	0.414	0.051	0.036	0.374
Two-stage		bias	-0.029	0.010	-0.005	0.022
		var	0.438	0.053	0.038	0.395
		bootstrap var	0.465	0.058	0.043	0.421
		90% coverage rate (%)	89.0	90.7	91.2	89.0
		95% coverage rate (%)	94.9	94.8	95.6	95.3
Complete case		bias	-0.018	0.004	0.000	0.010
		var	0.531	0.082	0.066	0.507
L		bias	0.399	-0.220	0.228	-0.491
$L/\sqrt{2}$		bias	0.701	-0.136	0.147	-0.620
Zero		bias	1.837	-0.417	0.427	-1.684
$E(Z\|Z < L)$		bias	0.415	-0.133	0.143	-0.418
Full data	400	bias	-0.019	0.007	-0.003	0.014
		var	0.212	0.028	0.019	0.192
Two-stage		bias	-0.019	0.008	-0.003	0.015
		var	0.225	0.029	0.020	0.204
		bootstrap var	0.226	0.028	0.021	0.205
		90% coverage rate (%)	89.4	89.2	90.4	89.9
		95% coverage rate (%)	95.0	93.8	95.0	95.0
Complete case		bias	-0.019	-0.001	0.004	0.008
		var	0.273	0.043	0.033	0.255
L		bias	0.404	-0.221	0.230	-0.495
$L/\sqrt{2}$		bias	0.724	-0.144	0.155	-0.642
Zero		bias	1.850	-0.426	0.436	-1.700
$E(Z\|Z < L)$		bias	0.433	-0.138	0.148	-0.434

Table 3.2: Simulation results for logistic regression.

	Sample size		$\beta_0 = -1$	$\beta_1 = 0.5$	$\beta_2 = -1$	$\gamma = 2$
Full data	200	bias	-0.030	0.013	-0.033	0.060
		var	2.157	0.268	0.216	2.033
Two-stage		bias	-0.041	0.016	-0.037	0.071
		var	2.313	0.278	0.230	2.191
		bootstrap var	2.424	0.299	0.235	2.260
		90% coverage rate (%)	91.4	91.7	90.7	91.0
		95% coverage rate (%)	96.2	96.3	95.9	96.2
Complete case		bias	-0.076	0.021	-0.045	0.106
		var	2.822	0.413	0.381	2.842
L		bias	0.309	-0.185	0.171	-0.368
$L/\sqrt{2}$		bias	0.690	-0.122	0.110	-0.570
Zero		bias	1.880	-0.453	0.441	-1.716
$E(Z\|Z < L)$		bias	0.350	-0.100	0.087	-0.313
Full data	400	bias	-0.033	0.007	-0.016	0.041
		var	0.930	0.123	0.096	0.881
Two-stage		bias	-0.043	0.011	-0.020	0.052
		var	1.013	0.129	0.104	0.964
		bootstrap var	1.101	0.138	0.107	1.022
		90% coverage rate (%)	90.8	91.2	90.5	90.6
		95% coverage rate (%)	95.8	96.3	95.8	95.2
Complete case		bias	-0.037	0.005	-0.018	0.048
		var	1.169	0.190	0.159	1.160
L		bias	0.319	-0.193	0.190	-0.398
$L/\sqrt{2}$		bias	0.651	-0.119	0.117	-0.553
Zero		bias	1.841	-0.442	0.440	-1.691
$E(Z\|Z < L)$		bias	0.332	-0.103	0.101	-0.317

Table 3.3: Simulation results for Poisson regression.

	Sample size		$\beta_0 = -1$	$\beta_1 = 0.5$	$\beta_2 = -1$	$\gamma = 2$
Full data	200	bias	0.022	-0.009	0.008	-0.026
		var	0.225	0.024	0.018	0.198
Two-stage		bias	0.034	-0.011	0.011	-0.037
		var	0.250	0.026	0.020	0.221
		bootstrap var	0.249	0.027	0.020	0.218
		90% coverage rate (%)	90.9	89.9	90.0	90.6
		95% coverage rate (%)	94.5	94.8	94.7	94.8
Complete case		bias	0.025	-0.011	0.010	-0.031
		var	0.351	0.053	0.041	0.325
L		bias	0.589	-0.288	0.286	-0.660
$L/\sqrt{2}$		bias	0.885	-0.200	0.210	-0.801
Zero		bias	1.867	-0.380	0.396	-1.691
$E(Z\|Z < L)$		bias	0.637	-0.213	0.217	-0.628
Full data	400	bias	0.018	-0.003	0.005	-0.020
		var	0.105	0.012	0.008	0.092
Two-stage		bias	0.019	-0.003	0.005	-0.021
		var	0.119	0.013	0.009	0.104
		bootstrap var	0.121	0.013	0.010	0.105
		90% coverage rate (%)	90.1	90.7	90.5	90.7
		95% coverage rate (%)	95.2	95.3	94.8	95.0
Complete case		bias	0.016	-0.004	0.007	-0.022
		var	0.175	0.027	0.022	0.163
L		bias	0.578	-0.281	0.283	-0.649
$L/\sqrt{2}$		bias	0.886	-0.196	0.208	-0.800
Zero		bias	1.870	-0.373	0.391	-1.689
$E(Z\|Z < L)$		bias	0.633	-0.208	0.215	-0.623

3.1 we see that the accelerated failure time model for anti-Mullerian hormone fits the data reasonably well, with respective goodness-of-fit empirical p-values 0.48, 0.664 and 0.602 for age, body mass index and follicle-stimulating hormone based on 500 simulated martingale residual score processes.

Table 3.4: Regression analysis results for the time to final menstrual period with covariate anti-Mullerian hormone subject to limit of detection: the Michigan Bone Health Metabolism Study.

		Intercept	Age	log(AMH)
Two-stage	estimate	19.59	-0.27	0.71
	bootstrap var	12.04	0.0064	0.032
Complete case	estimate	19.15	-0.25	0.74
	var	14.28	0.0073	0.078

AMH = anti-Mullerian hormone.

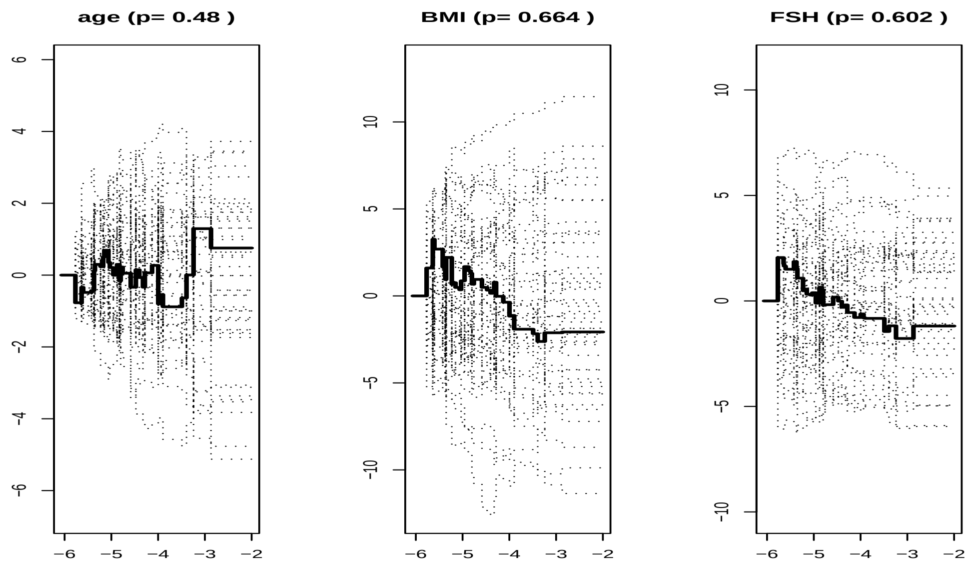

Figure 3.1: Accelerated failure time model for anti-Mullerian hormone: graphical checking for goodness of fit.

3.6 Appendix

3.6.1 Regularity Conditions

Denote the sample space of response variable Y by $\mathcal{Y}$, the sample space of covariate X by $\mathcal{X}$, the parameter space of θ by Θ, the parameter space of α by $\mathcal{A}$, and the parameter space of η by $\mathcal{H}$. In addition to the assumptions of bounded support for (X, Z) and compact parameter spaces Θ and $\mathcal{A}$, we provide a set of regularity conditions for Theorem III.1 in the following.

Condition III.1. $\Psi_\theta(\phi_0, \alpha_0, \eta_{0,\alpha_0})$ *has a unique root* θ_0.

Condition III.2. *For any constant* $U < \infty$, $\sup_{t\in[C,U]} |h(t)| \leq E_0 < \infty$ $\sup_{t\in[C,U]} |\dot{h}(t)| \leq E_1 < \infty$, *and* $\sup_{t\in[C,U]} |\ddot{h}(t)| \leq E_2 < \infty$, *where* $\dot{h}$ *and* $\ddot{h}$ *are the first and second derivatives of* h *respectively, and* E_0, E_1 *and* E_2 *are constants.*

Condition III.3. *Error* ς *has bounded density* $f = \dot{\eta}_{0,\alpha_0}$ *with bounded derivative* $\dot{f}$, *in other words,* $f \leq E_3 < \infty$, $|\dot{f}| \leq E_4 < \infty$ *for constants* E_3 *and* E_4, *and*

$$\int_{-\infty}^{\infty} (\dot{f}(t)/f(t))^2 f(t)dt < \infty.$$

Condition III.4. *There is a constant* $\tau < \infty$ *such that* $pr(V - X'\alpha \geq \tau) > \xi > 0$ *for all* $x \in \mathcal{X}$ *and* $\alpha \in \mathcal{A}$.

Condition III.5. $a(\phi)$ *is a monotone function satisfying* $|1/a(\phi)| \leq l < \infty$ *for a constant* l *with bounded derivatives* $\dot{a}(\cdot)$ *and* $\ddot{a}(\cdot)$.

Condition III.6. $\dot{b}(\cdot)$ *is a bounded monotone function.*

Condition III.7. $\ddot{b}(\cdot)$ *is a bounded Lipschitz function.*

Condition III.8. *There exist constants C_i, $i = 1, \ldots, 5$, such that for any constant $U < \infty$,*

$$\sup_{y\in\mathcal{Y},\theta\in\Theta,|1/a(\phi)|\le l,x\in\mathcal{X},t\in[C,U]} \left| f_{\theta,\phi}(y|t,x)\{y - \dot{b}(D'(t)\theta)\} \right| \le C_1 < \infty,$$

$$\sup_{y\in\mathcal{Y},\theta\in\Theta,|1/a(\phi)|\le l,x\in\mathcal{X},t\in[C,U]} \left| \frac{\partial f_{\theta,\phi}(y|t,x)}{\partial\phi}\{y - \dot{b}(D'(t)\theta)\} \right| \le C_2 < \infty,$$

$$\sup_{y\in\mathcal{Y},\theta\in\Theta,|1/a(\phi)|\le l,x\in\mathcal{X},t\in[C,U]} \left| \frac{\partial \left[f_{\theta,\phi}(y|t,x)\{y - \dot{b}(D'(t)\theta)\} \right]}{\partial t} \right| \le C_3 < \infty,$$

$$\sup_{y\in\mathcal{Y},\theta\in\Theta,|1/a(\phi)|\le l,x\in\mathcal{X},t\in[C,U]} \left| \frac{\partial f_{\theta,\phi}(y|t,x)}{\partial\phi} \right| \le C_4 < \infty,$$

$$\sup_{y\in\mathcal{Y},\theta\in\Theta,|1/a(\phi)|\le l,x\in\mathcal{X},t\in[C,U]} \left| \frac{\partial \left[f_{\theta,\phi}(y|t,x)\{y - \dot{b}(D'(t)\theta)\} \right]}{\partial\theta} \right| \le C_5 < \infty.$$

Condition III.9. *There exist constants $\delta_1 > 0$ and $\delta_2 > 0$, such that$\int_{C-X'\alpha}^{\tau} f_{\theta,\phi}(Y|t+X'\alpha, X)d\eta(t) \ge \delta_1$ with probability 1 for any $\theta \in \Theta$ and $|\phi-\phi_0|+|\alpha-\alpha_0|+\|\eta-\eta_0\| < \delta_2$.*

Remark: Condition III.1 is for the consistency, which may be unnecessarily strong for the proposed two-stage method. Direct calculation yields

$$\begin{aligned}
\dot{\Psi}_{\theta_0} &= \left.\frac{\partial\Psi_\theta(\phi_0,\alpha_0,\eta_0)}{\partial\theta}\right|_{\theta=\theta_0} \\
&= E\Bigg\{ -\Delta\ddot{b}\{D'(T)\theta_0\}D(T)^{\otimes 2} - (1-\Delta)\Bigg(\int_{C-X'\alpha_0}^{\tau} f_{\theta_0,\phi_0}(Y|t+X'\alpha_0,X)d\eta_0(t)\Bigg)^{-2} \\
&\Bigg(\int_{C-X'\alpha_0}^{\tau} f_{\theta_0,\phi_0}(Y|t+X'\alpha_0,X)[Y-\dot{b}\{D'(t+X'\alpha_0)\theta_0\}]D(t+X'\alpha_0)d\eta_0(t)\Bigg)^{\otimes 2}\Bigg\},
\end{aligned}$$

which is negative definite. Thus $\dot{\Psi}_\theta$, a continuous matrix with θ, is also negative definite in a neighborhood of θ_0, which guarantees that θ_0 is the unique solution of $\Psi_\theta(\phi_0,\alpha_0,\eta_0) = 0$ in a neighborhood of θ_0. The initial value we use in the Newton-Raphson algorithm for solving $\Psi_{\theta,n}(\hat{\phi}_n, \hat{\alpha}_n, \hat{\eta}_{n,\hat{\alpha}_n}) = 0$ is obtained from the complete case analysis, which is consistent, thus the solution of the proposed two-stage method should also be consistent.

Condition III.2 holds for many commonly used transformations, for example, $h(t) = \exp(-t)$ and polynomial functions. Condition III.3 and III.4 are usual assumptions for accelerated failure time models (Tsiatis, 1990; Nan et al., 2009). Conditions III.5-III.8 automatically hold for common generalized linear models, for example, linear, logistic or poisson regression.

Condition III.9 is mainly for technical convenience. One way to obtain Condition III.9 might be to truncate response variable Y such that $|Y| \leq M < \infty$ for a large constant M. In our simulations, however, we do not implement such truncations but still obtain satisfactory results.

3.6.2 General Z-estimation theory

The proof of Theorem III.1 in the main text is based on the general Z-estimation theory of Nan and Wellner (2013), which is provided in the following Lemmas III.1 and III.2 for our problem setting. Detailed discussion and proofs of these two lemmas can be found in Nan and Wellner (2013). Let $|\cdot|$ be the Euclidian norm and $\|\eta - \eta_0\| = \sup_t |\eta(t) - \eta_0(t)|$. Define $\rho\{(\phi, \alpha, \eta), (\phi_0, \alpha_0, \eta_0)\} = |\phi - \phi_0| + |\alpha - \alpha_0| + \|\eta - \eta_0\|$. We use P^* to denote outer probability, which is defined as $P^*(A) = \inf\{pr(B) : B \supset A, B \in \mathcal{B}\}$ for any subset A of Ω in a probability space $(\Omega, \mathcal{B}, P)$.

Lemma III.1. *(Consistency.) Suppose θ_0 is the unique solution to $\Psi_\theta(\phi_0, \alpha_0, \eta_0) = 0$ in the parameter space Θ and $(\hat{\phi}_n, \hat{\alpha}_n, \hat{\eta}_{n,\hat{\alpha}_n})$ are estimators of $(\phi_0, \alpha_0, \eta_0)$ such that $\rho\{(\hat{\phi}_n, \hat{\alpha}_n, \hat{\eta}_{n,\hat{\alpha}_n}), (\phi_0, \alpha_0, \eta_0)\} = o_{p^*}(1)$. If*

$$\sup_{\theta \in \Theta, \rho\{(\phi,\alpha,\eta),(\phi_0,\alpha_0,\eta_0)\} \leq \delta_n} \frac{|\Psi_{n,\theta}(\phi, \alpha, \eta) - \Psi_\theta(\phi_0, \alpha_0, \eta_0)|}{1 + |\Psi_{n,\theta}(\phi, \alpha, \eta)| + |\Psi_\theta(\phi_0, \alpha_0, \eta_0)|} = o_{p^*}(1) \qquad (3.10)$$

for every sequence $\{\delta_n \downarrow 0\}$, then $\hat{\theta}_n$ satisfying $\Psi_{n,\hat{\theta}_n}(\hat{\phi}_n, \hat{\alpha}_n, \hat{\eta}_{n,\hat{\alpha}_n}) = o_{p^*}(1)$ converges *in outer probability to* θ_0.

Lemma III.2. *(Rate of convergence and asymptotic representation.) Suppose that* $\hat{\theta}_n$ *satisfying* $\Psi_{n,\hat{\theta}_n}(\hat{\phi}_n, \hat{\alpha}_n, \hat{\eta}_{n,\hat{\alpha}_n}) = o_{p^*}(n^{-1/2})$ *is a consistent estimator of* θ_0 *that is a solution to* $\Psi_\theta(\phi_0, \alpha_0, \eta_0) = 0$ *in* Θ*, and that* $(\hat{\phi}_n, \hat{\alpha}_n, \hat{\eta}_{n,\hat{\alpha}_n})$ *is an estimator of* $(\phi_0, \alpha_0, \eta_0)$ *satisfying* $\rho\{(\hat{\phi}_n, \hat{\alpha}_n, \hat{\eta}_{n,\hat{\alpha}_n}),\ (\phi_0, \alpha_0, \eta_0)\} = O_{p^*}(n^{-1/2})$*. Suppose the following four conditions are satisfied:*

(i) (Stochastic equicontinuity.)

$$\frac{|n^{1/2}(\Psi_{n,\hat{\theta}_n} - \Psi_{\hat{\theta}_n})(\hat{\phi}_n, \hat{\alpha}_n, \hat{\eta}_{n,\hat{\alpha}_n}) - n^{1/2}(\Psi_{n,\theta_0} - \Psi_{\theta_0})(\phi_0, \alpha_0, \eta_0)|}{1 + n^{1/2}|\Psi_{n,\hat{\theta}_n}(\hat{\phi}_n, \hat{\alpha}_n, \hat{\eta}_{n,\hat{\alpha}_n})| + n^{1/2}|\Psi_{\hat{\theta}_n}(\hat{\phi}_n, \hat{\alpha}_n, \hat{\eta}_{n,\hat{\alpha}_n})|} = o_{p^*}(1).$$

(ii) $n^{1/2}\Psi_{n,\theta_0}(\phi_0, \alpha_0, \eta_0) = O_{p^*}(1)$.

(iii) (Smoothness.) There exist continuous matrices $\dot{\Psi}_{1,\theta_0}(\phi_0, \alpha_0, \eta_0)$, $\dot{\Psi}_{2,\theta_0}(\phi_0, \alpha_0, \eta_0)$, $\dot{\Psi}_{3,\theta_0}(\phi_0, \alpha_0, \eta_0)$*, and a continuous linear functional* $\dot{\Psi}_{4,\theta_0}(\phi_0, \alpha_0, \eta_0)$ *such that*

$$\begin{aligned} &|\Psi_{\hat{\theta}_n}(\hat{\phi}_n, \hat{\alpha}_n, \hat{\eta}_{n,\hat{\alpha}_n}) - \Psi_{\theta_0}(\phi_0, \alpha_0, \eta_0) \\ &\qquad -\dot{\Psi}_{1,\theta_0}(\phi_0, \alpha_0, \eta_0)(\hat{\theta}_n - \theta_0) - \dot{\Psi}_{2,\theta_0}(\phi_0, \alpha_0, \eta_0)(\hat{\phi}_n - \phi_0) \\ &\qquad -\dot{\Psi}_{3,\theta_0}(\phi_0, \alpha_0, \eta_0)(\hat{\alpha}_n - \alpha_0) - \dot{\Psi}_{4,\theta_0}(\alpha_0, \eta_0)(\hat{\eta}_{n,\hat{\alpha}_n} - \eta_0)| \\ &= o(|\hat{\theta}_n - \theta_0|) + o[\rho\{(\hat{\phi}_n, \hat{\alpha}_n, \hat{\eta}_{n,\hat{\alpha}_n}), (\phi_0, \alpha_0, \eta_0)\}]. \end{aligned} \tag{3.11}$$

Here the subscripts 1, 2, 3, and 4 correspond to θ, ϕ, α, *and* η *in* $\Psi_\theta(\phi, \alpha, \eta)$*, respectively, and we assume that the matrix* $\dot{\Psi}_{1,\theta_0}(\phi_0, \alpha_0, \eta_0)$ *is nonsingular.*

(iv) $n^{1/2}\dot{\Psi}_{2,\theta_0}(\phi_0, \alpha_0, \eta_0)(\hat{\phi}_n - \phi_0) = O_{p^*}(1)$, $n^{1/2}\dot{\Psi}_{3,\theta_0}(\phi_0, \alpha_0, \eta_0)(\hat{\alpha}_n - \alpha_0) = O_{p^*}(1)$, *and* $n^{1/2}\dot{\Psi}_{4,\theta_0}(\phi_0, \alpha_0, \eta_0)(\hat{\eta}_{n,\hat{\alpha}_n} - \eta_0) = O_{p^*}(1)$.

Then $\hat{\theta}_n$ is $n^{1/2}$ -consistent and further we have

$$\begin{aligned} n^{1/2}(\hat{\theta}_n - \theta_0) &= \{-\dot{\Psi}_{1,\theta_0}(\phi_0, \alpha_0, \eta_0)\}^{-1} n^{1/2}\{(\Psi_{n,\theta_0} - \Psi_{\theta_0})(\phi_0, \alpha_0, \eta_0) \\ &\quad + \dot{\Psi}_{2,\theta_0}(\phi_0, \alpha_0, \eta_0)(\hat{\phi}_n - \phi_0) + \dot{\Psi}_{3,\theta_0}(\phi_0, \alpha_0, \eta_0)(\hat{\alpha}_n - \alpha_0) \\ &\quad + \dot{\Psi}_{4,\theta_0}(\phi_0, \alpha_0, \eta_0)(\hat{\eta}_{n,\hat{\alpha}_n} - \eta_0)\} + o_{p^*}(1). \end{aligned} \quad (3.12)$$

3.6.3 Technical lemmas

Now we provide technical preparations for the proof of Theorem III.1, some of which are from Ying Ding's 2010 University of Michigan Ph.D. thesis. We adopt the empirical process notation of van der Vaart and Wellner (1996).

Let $\epsilon_\alpha = V - X'\alpha$ and $\epsilon_0 = V - X'\alpha_0$. Define

$$\begin{aligned} h^{(0)}(\alpha, s) &= P\{1(\epsilon_\alpha \le s, \Delta = 1)\}, \\ h^{(1)}(\alpha, s) &= P\{1(\epsilon_\alpha \ge s)\}, \\ h^{(2)}(\alpha, s) &= P\{1(\epsilon_\alpha \ge s)X\}, \end{aligned}$$

and

$$H_n^{(1)}(\alpha, s) = \mathbb{P}_n\{1(\epsilon_\alpha \ge s)\}.$$

The Kaplan-Meier estimator of the distribution function of $T - \alpha X$ is given by

$$\hat{\eta}_{n,\alpha}(t) = 1 - \prod_{i: V_i - X_i'\alpha \le t} \left\{ 1 - \frac{\Delta_i/n}{H_n^{(1)}(\alpha, V_i - X_i'\alpha)} \right\}.$$

Define

$$F(\alpha, t) = 1 - \exp\left\{ - \int_{u \le t} \frac{dh^{(0)}(\alpha, u)}{h^{(1)}(\alpha, u)} \right\},$$

and denote $\dot{F}_\alpha(\alpha, t) = \partial F(\alpha, t)/\partial\alpha$. For function c in the exponential family, denote

$\dot{c}_\phi(Y, \phi_0) = \partial c(Y, \phi)/\partial\phi|_{\phi=\phi_0}$.

Let $\Phi\{\alpha, h^{(1)}, h^{(2)}\} = P\left[\left\{h^{(1)}(\alpha, \epsilon_\alpha)X - h^{(2)}(\alpha, \epsilon_\alpha)\right\}\Delta\right]$, which corresponds to the limiting Gehan weighted estimating function, and define

$$m_1(\alpha_0, s; t) = -P\left\{\frac{\Delta 1(s \geq \epsilon_0)1(t \geq \epsilon_0)}{h^{(1)}(\alpha_0, \epsilon_0)^2}\right\},\ m_2(\alpha_0, s; t, \Delta) = \frac{\Delta 1(t \geq s)}{h^{(1)}(\alpha_0, s)}, \quad (3.13)$$

$$\begin{aligned} & m_3(\alpha_0, \epsilon_0; \Delta, X) \quad (3.14) \\ &= \left[-\dot{\Phi}_\alpha\{\alpha_0, h^{(1)}(\alpha_0, \cdot), h^{(2)}(\alpha_0, \cdot)\}\right]^{-1}\left[\left\{h^{(1)}(\alpha_0, \cdot)X - h^{(2)}(\alpha_0, \cdot)\right\}\Delta\right. \\ &\quad \left. - \int\{1(\epsilon_0 \geq t)X\}dP_{\epsilon_0,\Delta}(t, 1) + \int\{1(\epsilon_0 \geq t)\}x dP_{\epsilon_0,\Delta,X}(t, 1, x)\right]. \end{aligned}$$

Lemma III.3. *Suppose Conditions III.3-III.4 hold, and let $\hat{\alpha}_n$ be the Gehan weighted estimator for α_0, we have*

$$\sup_{t\in[C-E_5,\tau]} |\hat{\eta}_{n,\hat{\alpha}_n}(t) - \eta_0(t)| = O_{p^*}(n^{-1/2}),$$

where C is transformed L and $E_5 = \sup_{\alpha\in\mathcal{A},x\in\mathcal{X}} |x'\alpha| < \infty$.

Proof. From the proof of Ying Ding's Theorem 2.2.3 in her 2010 University of Michigan Ph.D. thesis, for t in a bounded interval, we have for $t \in [C - E_5, \tau]$,

$$\begin{aligned} \sup_t n^{1/2}\{\hat{\eta}_{n,\hat{\alpha}_n}(t) - \eta_0(t)\} &= \sup_t \mathbb{G}_n[\{1 - \eta_0(t)\}\{m_1(\alpha_0, \epsilon_0; t) + m_2(\alpha_0, \epsilon_0; t, \Delta)\} \\ &\quad + \dot{F}_\alpha(\alpha_0, t)m_3(\alpha_0, \epsilon_0; X, \Delta)] + o_p(1), \quad (3.15) \end{aligned}$$

where $m_1(\alpha_0, s; t)$, $m_2(\alpha_0, \epsilon_0; t, \Delta)$, $m_3(\alpha_0, \epsilon_0; X, \Delta)$ are defined in (3.13) and (3.14).

We first calculate the bracket numbers for $\mathcal{F}_1 = \{m_1(\alpha_0, \epsilon_0; t), t \in [C - E_5, \tau]\}$ and $\mathcal{F}_2 = \{m_2(\alpha_0, \epsilon_0; t, \Delta), t \in [C - E_5, \tau]\}$. For any nontrivial ε satisfying $1 > \varepsilon > 0$, let t_i be the i-th $\lceil 1/\varepsilon \rceil$ quantile of $\varsigma_0 = T - X'\alpha_0$, i.e.

$$pr(\varsigma_0 \leq t_i) = i\varepsilon,\ i = 1, \cdots, \lceil 1/\varepsilon \rceil - 1,$$

where $\lceil x \rceil$ is the smallest integer that is greater than or equal to x. Furthermore, denote $t_0 = 0$ and $t_{\lceil 1/\varepsilon \rceil} = +\infty$. For $i = 1, \cdots, \lceil 1/\varepsilon \rceil$, define brackets $[L_i, U_i]$ with

$$L_i(s) = -P\left\{\frac{\Delta 1(s \geq \epsilon_0)1(t_i \geq \epsilon_0)}{h^{(1)}(\alpha_0, \epsilon_0)^2}\right\},\ U_i(s) = -P\left\{\frac{\Delta 1(s \geq \epsilon_0)1(t_{i-1} \geq \epsilon_0)}{h^{(1)}(\alpha_0, \epsilon_0)^2}\right\}$$

such that $L_i(s) \leq -P\left\{\frac{\Delta 1(s \geq \epsilon_0)1(t \geq \epsilon_0)}{h^{(1)}(\alpha_0, \epsilon_0)^2}\right\} \leq U_i(s)$ when $t_{i-1} < t \leq t_i$. Since

$$E|U_i - L_i| \ \leq \ pr(t_{i-1} < \varsigma_0 \leq t_i)/\{h^{(1)}(\alpha_0, \tau)\}^2 = \varepsilon/\xi^2$$

from Condition III.4, we have $N_{[\ ]}(\varepsilon/\xi^2, \mathcal{F}_1, L_1) \leq 2/\varepsilon$ which yields

$$N_{[\ ]}(\varepsilon, \mathcal{F}_1, L_1) \leq K_1/\varepsilon,$$

where $K_1 = 2\xi^2$. Similarly, we have

$$N_{[\ ]}(\varepsilon, \mathcal{F}_2, L_1) \leq K_2/\varepsilon,$$

where $K_2 = 2\xi$. From Theorem 2.14.9 in van der Vaart and Wellner (1996), we have

$$P^*\left(\sup_{t \in [C-E_5, \tau]} |\mathbb{G}_n\{(1-\eta_0(t))m_1(\alpha_0, \epsilon_0; t)\}| > q\right)$$
$$\leq P^*\left(\sup_{t \in [C-E_5, \tau]} |\mathbb{G}_n\{m_1(\alpha_0, \epsilon_0; t)\}| > q\right) \leq D_1 q e^{-2q^2}, \quad (3.16)$$
$$P^*\left(\sup_{t \in [C-E_5, \tau]} |\mathbb{G}_n\{(1-\eta_0(t))m_2(\alpha_0, \epsilon_0; t, \Delta)\}| > q\right)$$
$$\leq P^*\left(\sup_{t \in [C-E_5, \tau]} |\mathbb{G}_n\{m_2(\alpha_0, \epsilon_0; t, \Delta)\}| > q\right) \leq D_2 q e^{-2q^2} \quad (3.17)$$

for some constant D_1 depends on K_1 and constant D_2 depends on K_2. We now show

$\sup_{t\in[C-E_5,\tau]}|\dot{F}_\alpha(\alpha_0,t)|$ is bounded. Direct calculation yields

$$\begin{aligned}
&\sup_{t\in[C-E_5,\tau]}|\dot{F}_\alpha(\alpha_0,t)| \\
&\quad= \sup_{t\in[C-E_5,\tau]} e^{-\int_{u\le t}\frac{dh^{(0)}(\alpha_0,u)}{h^{(1)}(\alpha_0,u)}}\left|\int_{u\le t}\frac{d\dot{h}_\alpha^{(0)}(\alpha_0,u)}{h^{(1)}(\alpha_0,u)} - \int_{u\le t}\frac{\dot{h}_\alpha^{(1)}(\alpha_0,u)dh^{(0)}(\alpha_0,u)}{\{h^{(1)}(\alpha_0,u)\}^2}\right| \\
&\quad\le \{h^{(1)}(\alpha_0,\tau)\}^{-1}\sup_{t\in[C-E_5,\tau]}|\dot{h}_\alpha^{(0)}(\alpha_0,t)| \\
&\qquad\quad +\{h^{(1)}(\alpha_0,\tau)\}^{-2}\sup_{u\in(-\infty,\infty)}\left|\dot{h}_u^{(0)}(\alpha_0,u)\right|\sup_{t\in[C-E_5,\tau]}\int_{u\le t}|\dot{h}_\alpha^{(1)}(\alpha_0,u)|du,
\end{aligned}$$

where $\dot{h}_\alpha^{(0)}(\alpha_0,t) = \frac{\partial}{\partial\alpha}h^{(0)}(\alpha,t)\big|_{\alpha=\alpha_0}$, $\dot{h}_\alpha^{(1)}(\alpha_0,t) = \frac{\partial}{\partial\alpha}h^{(1)}(\alpha,t)\big|_{\alpha=\alpha_0}$ and $\dot{h}_u^{(0)}(\alpha_0,u) = \frac{\partial}{\partial u}h^{(0)}(\alpha_0,u)$. Since

$$\begin{aligned}
h^{(0)}(\alpha,t) &= \int \eta_0(\min(t+x'\alpha-x'\alpha_0, C-x'\alpha_0))dF_X(x) \\
&= \int_{x'\alpha\ge C-t}\eta_0(C-x'\alpha_0)dF_X(x) + \int_{x'\alpha<C-t}\eta_0(t+x'\alpha-x'\alpha_0)dF_X(x), \\
h^{(1)}(\alpha,t) &= \int_{x'\alpha\le C-t}\{1-\eta_0(t+x'\alpha-x'\alpha_0)\}dF_X(x),
\end{aligned}$$

where $F_X(x)$ is the distribution function of X, from Condition III.3 we have

$$\begin{aligned}
&\sup_{t\in[C-E_5,\tau]}|\dot{h}_\alpha^{(0)}(\alpha_0,t)| = \sup_{t\in[C-E_5,\tau]}\left|\dot{\eta}_{0,\alpha_0}(t)\int_{t+x'\alpha_0<C}xdF_X(x)\right| \le E_3E|X| < \infty, \\
&\sup_{u\in(-\infty,\infty)}\left|\dot{h}_u^{(0)}(\alpha_0,u)\right| \le \sup_{u\in(-\infty,\infty)}|\dot{\eta}_{0,\alpha_0}(u)| \le E_3, \\
&\sup_{t\in[C-E_5,\tau]}\int_{u\le t}|\dot{h}_\alpha^{(1)}(\alpha_0,u)|du \\
&\quad\le \sup_{t\in[C-E_5,\tau]}\int_{u\le t}\left|\int_{t+x'\alpha_0\le C}xdF_X(x)\right|\dot{\eta}_{0,\alpha_0}(u)du + \int_{-\infty}^{\infty}|x|dF_X(x) \\
&\quad\le E|X|E_3 + E|X| < \infty.
\end{aligned}$$

Since it can be shown that $m_3(\alpha_0, \epsilon_0; X, \Delta)$ has finite second moment, we have $\sup_{t\in[C-E_5,\tau]} \mathbb{G}_n[\dot{F}_\alpha(\alpha_0, t)m_3(\alpha_0, \epsilon_0; X, \Delta)] = O_{p^*}(1)$, thus obtain the desired result. □

Lemma III.4. *Suppose Condition III.7 holds, we have that*

$$\Big\{\Delta\{Y - \dot{b}(D'(t)\theta)\}D(t), \theta \in \Theta, t \in \mathcal{T} \subset \mathbb{R}\Big\} \tag{3.18}$$

is Donsker.

Proof. From Condition III.7 we know that $\ddot{b}(\cdot)$ is bounded, hence $\dot{b}(\cdot)$ is a Lipschitiz function. From Theorem 2.10.6 in van der Vaart and Wellner (1996), we know that $D(t)$ and $\dot{b}(D'(t)\theta)$ are Donsker, hence (3.18) is Donsker. □

Lemma III.5. *Suppose $\mathcal{X}$ and $\mathcal{A}$ be the bounded covariate and parameter spaces. Let $\mathcal{H}$ be a collection of distribution functions satisfying Condition III.3. We have $\mathcal{F} = \{\eta(t - x'\alpha), t \in \mathcal{T} \subset \mathbb{R}, x \in \mathcal{X}, \alpha \in \mathcal{A}, \eta \in \mathcal{H}\}$ is Donsker.*

Proof. Let $\mathcal{F}_1 = \{\eta(t)\}$. From Theorem 2.7.5 in van der Vaart and Wellner (1996), the number of brackets $[L_i, U_i]$ such that $L_i(t) \le \eta(t) \le U_i(t)$ for any nontrivial ε with $1 > \varepsilon > 0$ and $\int |U_i(t) - L_i(t)| d\eta_0(t) \le \varepsilon$ satisfies $\log N_{[\ ]}(\varepsilon, \mathcal{F}_1, L_1(P)) \le K_1/\varepsilon$, where $K_1 < \infty$ is a constant.

For notational simplicity, we consider 1-dimensional $\mathcal{A}$. Because $\mathcal{A}$ is bounded, we partition $\mathcal{A}$ by a set of intervals $[l_k, u_k)$ such that $|u_k - l_k| \le \varepsilon$. Hence the number of such intervals is bounded by K_2/ε with a constant $K_2 < \infty$. Now we construct brackets for $\mathcal{F} \equiv \{\eta(t - x\alpha)\}$. Define

$$O_{ik}(t, x) = \min(L_i(t - xu_k), L_i(t - xl_k)),\ S_{ik}(t, x) = \max(U_i(t - xu_k), U_i(t - xl_k)).$$

We have

$$\begin{aligned} O_{ik}(t,x) &\leq \min(\eta(t-xl_k), \eta(t-xu_k)) \\ &\leq \eta(t-x\alpha) \\ &\leq \max(\eta(t-xl_k), \eta(t-xu_k)) \leq S_{ik}(t,x). \end{aligned}$$

Since

$$P \mid S_{ik} - O_{ik} \mid$$

$$\leq \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mid U_i(t-xu_k) - L_i(t-xu_k) \mid d\eta_0(t+x\alpha_0)dF_X(x) \tag{3.19}$$

$$+ \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mid U_i(t-xl_k) - L_i(t-xl_k) \mid d\eta_0(t+x\alpha_0)dF_X(x) \tag{3.20}$$

$$+ \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mid U_i(t-xl_k) - L_i(t-xu_k) \mid d\eta_0(t+x\alpha_0)dF_X(x) \tag{3.21}$$

$$+ \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mid U_i(t-xu_k) - L_i(t-xl_k) \mid d\eta_0(t+x\alpha_0)dF_X(x). \tag{3.22}$$

Since $[L_i, U_i]$ are brackets for $\mathcal{F}_1$, we have (3.19) $\leq \varepsilon$ and (3.20) $\leq \varepsilon$. Furthermore, by integration by parts and change of variables we obtain

$$\begin{aligned} (3.21) \;\leq\; & 2\varepsilon + \int_{0}^{\infty}\int_{-\infty}^{\infty} \{\eta(t-xl_k) - \eta(t-xu_k)\} d\eta_0(t+x\alpha_0)dF_X(x) \\ & + \int_{-\infty}^{0}\int_{-\infty}^{\infty} \{\eta(t-xu_k) - \eta(t-xl_k)\} d\eta_0(t+x\alpha_0)dF_X(x) \\ =\; & 2\varepsilon + \int_{0}^{\infty}\int_{-\infty}^{\infty} \{\eta_0(t+x\alpha_0+xu_k) - \eta_0(t+x\alpha_0+xl_k)\} d\eta(t)dF_X(x) \\ & + \int_{-\infty}^{0}\int_{-\infty}^{\infty} \{\eta_0(t+x\alpha_0+xl_k) - \eta_0(t+x\alpha_0+xu_k)\} d\eta(t)dF_X(x) \end{aligned}$$

$$\leq \ 2\varepsilon + \int_0^\infty \int_{-\infty}^\infty E_3 x(u_k - l_k) d\eta(t) dF_X(x) - \int_{-\infty}^0 \int_{-\infty}^\infty E_3 x(u_k - l_k) d\eta(t) dF_X(x)$$
$$\leq \ 2\varepsilon + E_3 E|X|\varepsilon = K_3\varepsilon,$$

where E_3 is defined in Condition III.3, and $K_3 = 2 + E_3 E|X| < \infty$. Similarly, we have (3.22) $\leq K_3\varepsilon$. Hence we have $N_{[\,]}((2+2K_3)\varepsilon, \mathcal{F}, L_1(P)) \leq \exp(K_1/\varepsilon)K_2/\varepsilon$, i.e. $N_{[\,]}(\varepsilon, \mathcal{F}, L_1(P)) \leq \exp(K_1(2+2K_3)/\varepsilon)K_2(2+2K_3)/\varepsilon \leq \exp((K_1+K_2)(2+2K_3)/\varepsilon)$. Hence, $\mathcal{F}$ is Donsker. □

Lemma III.6. *Suppose Conditions III.2, III.5-III.9 hold, we have*

$$\left\{ \frac{\int_{C-x'\alpha}^{\tau} f_{\theta,\phi}(y \mid t + x'\alpha, x)\{y - \dot{b}(D'(t+x'\alpha)\theta)\}D(t+x'\alpha)d\eta(t)}{\int_{C-x'\alpha}^{\tau} f_{\theta,\phi}(y \mid t + x'\alpha, x)d\eta(t)} : \right. \tag{3.23}$$
$$\left. \theta \in \Theta, |1/a(\phi)| < l, \alpha \in \mathcal{A}, \eta \in \mathcal{H}, \rho\{(\phi,\alpha,\eta),(\phi_0,\alpha_0,\eta_0)\} < \delta_2, x \in \mathcal{X}, y \in \mathcal{Y} \right\}$$

is Donsker.

Proof. From Condition III.9, we have $\{\int_{C-x'\alpha}^{\tau} f_{\theta,\phi}(y \mid t + x'\alpha, x)d\eta(t)\}$ bounded away from zero. From Section 2.10.2 of van der Vaart and Wellner (1996), we only need to show that both the numerator and denominator in (3.23) belong to Donsker classes. By integration by parts, we have

$$\begin{aligned}
&\int_{C-x'\alpha}^{\tau} f_{\theta,\phi}(y \mid t + x'\alpha, x)d\eta(t) \\
&\quad = f_{\theta,\phi}(y \mid \tau + x'\alpha, x)\eta(\tau) - f_{\theta,\phi}(y \mid C, x)\eta(C - x'\alpha) \\
&\qquad - \int_{C-E_5}^{\tau} 1(t \geq C - x'\alpha)\eta(t) f_{\theta,\phi}(y \mid t + x'\alpha, x) \\
&\qquad\qquad \gamma\{y - \dot{b}(D'(t+x'\alpha)\theta)\}\dot{h}(t+x'\alpha)/a(\phi)dt.
\end{aligned}$$

In the above, $\dot{h}(\cdot)$ is Lipschitz by Condition III.2 and $f_{\theta,\phi}(y \mid t + x'\alpha, x)$ is Lipschitz function for θ, ϕ and α by Conditions III.2, III.5 and III.8, thus both belong to

Donsker classes by Theorem 2.10.6 in van der Vaart and Wellner (1996). By Lemma III.5 we know that $\{\eta(C - x'\alpha)\}$ is Donsker. Since the class of indicator functions of half spaces is a VC-class, see e.g. Exercise 9 on page 151 and Exercise 14 on page 152 in van der Vaart and Wellner (1996), thus the set of functions $\{1(t \geq C - x'\alpha)\}$ is a Donsker class. By Theorem 2.10.3 in van der Vaart and Wellner (1996), the permanence of the Donsker property for the closure of the convex hull, we have $\left\{ \int_{C-E_5}^{\tau} 1(t \geq C - x'\alpha)\eta(t) f_{\theta,\phi}(y \mid t + x'\alpha, x)\gamma\{y - \dot{b}(D'(t+x'\alpha)\theta)\}/a(\phi)\dot{h}(t+x'\alpha)dt \right\}$ is Donsker. Hence the denominator of (3.23) belongs to a Donsker class.

Similarly, by integration by parts,

$$
\begin{aligned}
&\int_{C-x'\alpha}^{\tau} f_{\theta,\phi}(y \mid t + x'\alpha, x)\{y - \dot{b}(D'(t + x'\alpha)\theta)\}D(t + x'\alpha)d\eta(t) \\
&\quad = f_{\theta,\phi}(y \mid \tau + x'\alpha, x)\{y - \dot{b}(D'(\tau + x'\alpha)\theta)\}D(\tau + x'\alpha)\eta(\tau) \\
&\quad - f_{\theta,\phi}(y \mid C, x)\{y - \dot{b}(D'(C)\theta)\}D(C)\eta(C - x'\alpha) \\
&\quad - \int_{C-E_5}^{\tau} 1(t \geq C - x'\alpha)\eta(t) f_{\theta,\phi}(y \mid t + x'\alpha, x) \\
&\qquad (\gamma\{y - \dot{b}(D'(t + x'\alpha)\theta)\}^2 D(t + x'\alpha)/a(\phi) - \ddot{b}(D'(t + x'\alpha)\theta)\gamma D(t + x'\alpha) \\
&\qquad + \{y - \dot{b}(D'(t + x'\alpha)\theta)\}J_{p+2})\dot{h}(t + x'\alpha)dt,
\end{aligned}
$$

where $J_{p+2} = (0, \cdots, 0, 1)'_{1\times(p+2)}$. Similar to the denominator, we can show that the above function, which is the numerator of (3.23), belongs to a Donsker class provided that $\{\ddot{b}(D'(t + x'\alpha)\theta)\}$ is Donsker from Condition III.7. □

Lemma III.7. *Under Conditions III.5-III.9, when* $\theta \to \theta_0$ *and* $\rho\{(\phi, \alpha, \eta), (\phi_0, \alpha_0, \eta_0)\} \to 0$, *we have that* $E|\psi_\theta(\phi, \alpha, \eta) - \psi_{\theta_0}(\phi_0, \alpha_0, \eta_0)|^2 \to 0$.

Proof. The proof follows straightforward algebraic calculations based on the Mean Value Theorem. The details are thus omitted. □

Lemma III.8. *Suppose Conditions III.2, III.5-III.9 hold, we have* $E|\psi_{\theta_0}(\phi_0, \alpha_0, \eta_0)|^2 <$

∞.

Proof. Again, the proof is based on direct calculation. □

3.6.4 Proof of Theorem III.1

3.6.4.1 Proof of consistency

Proof. We prove consistency using Lemma III.1. Since $\hat{\phi}_n$ and $\hat{\alpha}_n$ are $n^{1/2}$ -consistent, see the last paragraph of Section 3.3, and $\hat{\eta}_{n,\hat{\alpha}_n}$ is also $n^{1/2}$ -consistent in a finite interval from Lemma III.3, we have

$$\rho\{(\hat{\phi}_n, \hat{\alpha}_n, \hat{\eta}_{n,\hat{\alpha}_n}), (\phi_0, \alpha_0, \eta_0)\} = o_{p^*}(1).$$

Given that θ_0 is the unique solution to $\Psi_\theta(\phi_0, \alpha_0, \eta_0) = 0$ from Condition III.1, we only need to show that

$$\sup_{\theta\in\Theta,\rho\{(\phi,\alpha,\eta),(\phi_0,\alpha_0,\eta_0)\}\le\delta_n} |\Psi_{\theta,n}(\phi,\alpha,\eta) - \Psi_\theta(\phi_0,\alpha_0,\eta_0)| = o_{p^*}(1) \tag{3.24}$$

for every sequence $\delta_n \downarrow 0$. Now

$$\begin{aligned}
&\sup_{\theta\in\Theta,\rho\{(\phi,\alpha,\eta),(\phi_0,\alpha_0,\eta_0)\}\le\delta_n} |\Psi_{\theta,n}(\phi,\alpha,\eta) - \Psi_\theta(\phi_0,\alpha_0,\eta_0)| \\
&\le \sup_{\theta\in\Theta} |(\mathbb{P}_n - P)[\Delta\{Y - \dot{b}(D(T)\theta)\}D(T)]| \qquad (3.25)\\
&\quad + \sup_{\theta\in\Theta,\rho\{(\phi,\alpha,\eta),(\phi_0,\alpha_0,\eta_0)\}\le\delta_n} \qquad (3.26)\\
&\quad P\left|\frac{\int_{C-X'\alpha}^{\tau} f_{\theta,\phi}(Y \mid t+X'\alpha, X)\{Y - \dot{b}(D'(t+X'\alpha)\theta)\}D(t+X'\alpha)d\eta(t)}{\int_{C-X'\alpha}^{\tau} f_{\theta,\phi}(Y \mid t+X'\alpha, X)d\eta(t)}\right. \\
&\quad \left. - \frac{\int_{C-X'\alpha_0}^{\tau} f_{\theta,\phi_0}(Y \mid t+X'\alpha_0, X)\{Y - \dot{b}(D'(t+X'\alpha_0)\theta)\}D(t+X'\alpha_0)d\eta_0(t)}{\int_{C-X'\alpha_0}^{\tau} f_{\theta,\phi_0}(Y \mid t+X_i'\alpha_0, X)d\eta_0(t)}\right|
\end{aligned}$$

$$+ \sup_{\theta\in\Theta,\rho\{(\phi,\alpha,\eta),(\phi_0,\alpha_0,\eta_0)\}\le\delta_n} \Bigg| (\mathbb{P}_n - P)(1-\Delta) \tag{3.27}$$

$$\frac{\int_{C-X'\alpha}^{\tau} f_{\theta,\phi}(Y \mid t+X'\alpha, X)\{Y - \dot{b}(D'(t+X'\alpha)\theta)\}D(t+X'\alpha)d\eta(t)}{\int_{C-X'\alpha}^{\tau} f_{\theta,\phi}(Y \mid t+X'\alpha, X)d\eta(t)}\Bigg| = o_{p^*}(1),$$

where (3.25) and (3.27) equal to $o_{p^*}(1)$ are from Lemma III.4 and Lemma III.6, respectively, and (3.26) equal to $o_{p^*}(1)$ follows a direct calculation similar to Lemma III.7 using the Mean Value Theorem. □

3.6.4.2 Proof of asymptotic normality

Proof. We now verify all the conditions in Lemma III.2. Condition (i) holds because $\{\psi_\theta(\phi,\alpha,\eta)\}$ is Donsker by Lemmas III.4 and III.6, together with the result in Lemma III.7. Condition (ii) holds by the classical central limit theorem for independent and identically distributed data with $E|\psi_{\theta_0}(\phi_0,\alpha_0,\eta_0(\alpha_0))|^2 < \infty$ from Lemma III.8.

For Condition (iii), given that $\rho\{(\hat{\phi}_n,\hat{\alpha}_n,\hat{\eta}_{n,\hat{\alpha}_n}),(\phi_0,\alpha_0,\eta_0)\} = O_{p^*}(n^{-1/2})$, taking the Taylor expansion for θ, ϕ and α we obtain

$$\begin{aligned}
&\Psi_{\hat{\theta}_n}(\hat{\phi}_n,\hat{\alpha}_n,\hat{\eta}_{n,\hat{\alpha}_n}) - \Psi_{\theta_0}(\phi_0,\alpha_0,\eta_0)\\
&\quad = \dot{\Psi}_{1,\tilde{\theta}}(\hat{\phi}_n,\hat{\alpha}_n,\hat{\eta}_{n,\hat{\alpha}_n})(\hat{\theta}_n-\theta_0) - \dot{\Psi}_{2,\theta_0}(\tilde{\phi},\hat{\alpha}_n,\hat{\eta}_{n,\hat{\alpha}_n})(\hat{\phi}_n-\phi_0)\\
&\qquad -\dot{\Psi}_{3,\theta_0}(\phi_0,\tilde{\alpha},\eta_0)(\hat{\alpha}_n-\alpha_0) - R(\theta_0,\phi_0,\hat{\alpha}_n,\hat{\eta}_{n,\hat{\alpha}_n},\eta_0),
\end{aligned}$$

where $\tilde{\theta}$ is between θ_0 and $\hat{\theta}_n$, $\tilde{\phi}$ is between ϕ_0 and $\hat{\phi}_n$, $\tilde{\alpha}$ is between α_0 and $\hat{\alpha}_n$, and the remainder has the following form

$$\begin{aligned}
&R(\theta_0,\phi_0,\alpha,\eta,\eta_0)\\
&\quad = P\left[(1-\Delta)\left\{\frac{\int_{C-X'\alpha}^{\tau} A(t,\theta_0,\phi_0,\alpha)d\eta(t)}{\int_{C-X'\alpha}^{\tau} B(t,\theta_0,\phi_0,\alpha)d\eta(t)} - \frac{\int_{C-X'\alpha}^{\tau} A(t,\theta_0,\phi_0,\alpha)d\eta_0(t)}{\int_{C-X'\alpha}^{\tau} B(t,\theta_0,\phi_0,\alpha)d\eta_0(t)}\right\}\right]
\end{aligned}$$

with

$$
\begin{aligned}
A(t,\theta_0,\phi_0,\alpha) &= f_{\theta_0,\phi_0}(Y \mid t+X'\alpha, X)\{Y-\dot{b}(D'(t+X'\alpha)\theta_0)\}D(t+X'\alpha),\\
B(t,\theta_0,\phi_0,\alpha) &= f_{\theta_0,\phi_0}(Y \mid t+X'\alpha, X).
\end{aligned}
$$

It can be show by direct calculation that $|\dot{\Psi}_{1,\tilde{\theta}}(\hat{\phi}_n,\hat{\alpha}_n,\hat{\eta}_{n,\hat{\alpha}_n}) - \dot{\Psi}_{1,\theta_0}(\phi_0,\alpha_0,\eta_0)| = o_{p^*}(1)$, $|\dot{\Psi}_{2,\theta_0}(\tilde{\phi},\hat{\alpha}_n,\hat{\eta}_{n,\hat{\alpha}_n}) - \dot{\Psi}_{2,\theta_0}(\phi_0,\alpha_0,\eta_0)| = o_{p^*}(1)$ and $|\dot{\Psi}_{3,\theta_0}(\phi_0,\tilde{\alpha},\eta_0) - \dot{\Psi}_{3,\theta_0}(\phi_0,\alpha_0,\eta_0)| = o_{p^*}(1)$.

Define

$$
\begin{aligned}
&\dot{\Psi}_{4,\theta_0}(\phi_0,\alpha_0,\eta_0)(\hat{\eta}_{n,\hat{\alpha}_n}-\eta_0) \qquad (3.28)\\
&= P\Bigg[(1-\Delta)\Bigg\{\frac{\int_{C-X'\alpha_0}^{\tau} A(t,\theta_0,\phi_0,\alpha_0)d[\hat{\eta}_{n,\hat{\alpha}_n}(t)-\eta_0(t)]}{\int_{C-X'\alpha_0}^{\tau} B(t,\theta_0,\phi_0,\alpha_0)d\eta_0(t)}\\
&\quad -\frac{\int_{C-X'\alpha_0}^{\tau} A(t,\theta_0,\phi_0,\alpha_0)d\eta_0(t)\int_{C-X'\alpha_0}^{\tau} B(t,\theta_0,\phi_0,\alpha_0)d[\hat{\eta}_{n,\hat{\alpha}_n}(t)-\eta_0(t)]}{\int_{C-X'\alpha_0}^{\tau} B(t,\theta_0,\phi_0,\alpha_0)d\eta_0(t)^2}\Bigg\}\Bigg]\\
&= P\Bigg[(1-\Delta)\Bigg\{\frac{\int_{C-X'\alpha_0}^{\tau} A(t,\theta_0,\phi_0,\alpha_0)d\hat{\eta}_{n,\hat{\alpha}_n}(t)}{\int_{C-X'\alpha_0}^{\tau} B(t,\theta_0,\phi_0,\alpha_0)d\eta_0(t)}\\
&\quad -\frac{\int_{C-X'\alpha_0}^{\tau} A(t,\theta_0,\phi_0,\alpha_0)d\eta_0(t)\int_{C-X'\alpha_0}^{\tau} B(t,\theta_0,\phi_0,\alpha_0)d\hat{\eta}_{n,\hat{\alpha}_n}(t)}{\int_{C-X'\alpha_0}^{\tau} B(t,\theta_0,\phi_0,\alpha_0)d\eta_0(t)^2}\Bigg\}\Bigg].
\end{aligned}
$$

Then we have

$$
\begin{aligned}
&|R(\theta_0,\phi_0,\hat{\alpha}_n,\hat{\eta}_{n,\hat{\alpha}_n},\eta_0) - \dot{\Psi}_{4,\theta_0}(\phi_0,\alpha_0,\eta_0)(\hat{\eta}_{n,\hat{\alpha}_n}-\eta_0)|\\
&\quad\le |R(\theta_0,\phi_0,\hat{\alpha}_n,\hat{\eta}_{n,\hat{\alpha}_n},\eta_0) - R(\theta_0,\phi_0,\alpha_0,\hat{\eta}_{n,\hat{\alpha}_n},\eta_0)|\\
&\qquad + |R(\theta_0,\phi_0,\alpha_0,\hat{\eta}_{n,\hat{\alpha}_n},\eta_0) - \dot{\Psi}_{4,\theta_0}(\phi_0,\alpha_0,\eta_0)(\hat{\eta}_{n,\hat{\alpha}_n}-\eta_0)|\\
&\quad = D_1 + D_2.
\end{aligned}
$$

Now $D_1 = o(|\hat{\alpha}_n - \alpha_0| + \|\hat{\eta}_{n,\hat{\alpha}_n} - \eta_0\|)$ can be shown by

$$\begin{aligned}
&\frac{A_1}{B_1} - \frac{A_2}{B_2} - \frac{A_3}{B_3} + \frac{A_4}{B_4} \\
&\quad = \frac{A_1}{B_1B_2}(B_2 - B_1 - B_4 + B_3) + \frac{A_1}{B_1B_2B_3B_4}(B_3B_4 - B_1B_2)(B_4 - B_3) \\
&\qquad + \frac{A_1 - A_3}{B_3B_4}(B_4 - B_3) + \frac{A_1 - A_2}{B_2B_4}(B_4 - B_2) + \frac{A_1 - A_2 - A_3 + A_4}{B_4},
\end{aligned}$$

and $D_2 = o(|\hat{\alpha}_n - \alpha_0| + \|\hat{\eta}_{n,\hat{\alpha}_n} - \eta_0\|)$ can be shown by

$$\frac{A_1}{B_1} - \frac{A_2}{B_2} - \frac{A_1}{B_2} + \frac{A_2B_1}{B_2^2} = \frac{1}{B_1B_2^2}\{A_1(B_1 - B_2)^2 - B_1(A_2 - A_1)(B_2 - B_1)\}.$$

Since $\hat{\phi}_n$, $\hat{\alpha}_n$ and $\hat{\eta}_n$ are all root-n consistent, under Conditions (i)-(iii), Condition (iv) holds automatically. Then by Lemma III.2 we have that $\hat{\theta}_n$ is $n^{1/2}$ -consistent and (3.12) holds with

$$\begin{aligned}
&\dot{\Psi}_{1,\theta_0}(\phi_0, \alpha_0, \eta_0) \\
&\quad = E[\Delta\ddot{b}\{D'(T)\theta_0\}D(T)D'(T)] \\
&\qquad - E\Bigg[(1 - \Delta)\Bigg\{\int_{C-X'\alpha_0}^{\tau} f_{\theta_0,\phi_0}(Y \mid t + X'\alpha_0, X)d\eta_0(t)\Bigg\}^{-2} \\
&\qquad\quad \Bigg(\int_{C-X'\alpha_0}^{\tau} f_{\theta_0,\phi_0}(Y \mid t + X'\alpha_0, X)\{Y - \dot{b}(D'(t + X'\alpha_0)\theta_0)\} \\
&\qquad\quad D(t + X'\alpha_0)d\eta_0(t)\Bigg)^{\otimes 2}\Bigg],
\end{aligned}$$

$$
\begin{aligned}
&\dot{\Psi}_{2,\theta_0}(\phi_0, \alpha_0, \eta_0) \\
&= -E\Bigg[(1-\Delta)\Bigg\{\int\limits_{C-X'\alpha_0}^{\tau} f_{\theta_0,\phi_0}(Y \mid t + X'\alpha_0, X) d\eta_0(t)\Bigg\}^{-1} \\
&\quad \Bigg\{\int\limits_{C-X'\alpha_0}^{\tau} f_{\theta_0,\phi_0}(Y \mid t + X'\alpha_0, X)\{Y - \dot{b}(D'(t + X'\alpha_0)\theta_0)\}D(t + X'\alpha_0) \\
&\quad \Big([Y\{D'(t + X'\alpha_0)\theta_0\} - b(D'(t + X'\alpha_0)\theta_0)]a'(\phi_0)/a(\phi_0)^2 - \dot{c}_\phi(Y, \phi_0)\Big) d\eta_0(t)\Bigg\} \\
&\quad -\Bigg\{\int\limits_{C-X'\alpha_0}^{\tau} f_{\theta_0,\phi_0}(Y \mid t + X'\alpha_0, X) d\eta_0(t)\Bigg\}^{-2}\Bigg\{\int\limits_{C-X'\alpha_0}^{\tau} f_{\theta_0,\phi_0}(Y \mid t + X'\alpha_0, X) \\
&\quad \Big([Y\{D'(t + X'\alpha_0)\theta_0\} - b(D'(t + X'\alpha_0)\theta_0)]a'(\phi_0)/a(\phi_0)^2 - \dot{c}_\phi(Y, \phi_0)\Big) d\eta_0(t) \\
&\quad \int\limits_{C-X'\alpha_0}^{\tau} f_{\theta_0,\phi_0}(Y \mid t + X'\alpha_0, X)\{Y - \dot{b}(D'(t + X'\alpha_0)\theta_0)\}D(t + X'\alpha_0) d\eta_0(t)\Bigg\}\Bigg],
\end{aligned}
$$

and

$$
\begin{aligned}
&\dot{\Psi}_{3,\theta_0}(\phi_0, \alpha_0, \eta_0) \\
&= -E\Bigg[(1-\Delta)\Bigg\{\int\limits_{C-X'\alpha_0}^{\tau} f_{\theta_0,\phi_0}(Y \mid t + X'\alpha_0, X) d\eta_0(t)\Bigg\}^{-2} \\
&\quad \int\limits_{C-X'\alpha_0}^{\tau} f_{\theta_0,\phi_0}(Y \mid t + X'\alpha_0, X)\{Y - \dot{b}(D'(t + X'\alpha_0)\theta_0)\}D(t + X'\alpha_0) d\eta_0(t) \\
&\quad \Bigg\{\int\limits_{C-X'\alpha_0}^{\tau} f_{\theta_0,\phi_0}(Y \mid t + X'\alpha_0, X)\{Y - \dot{b}(D'(t + X'\alpha_0)\theta_0)\} \\
&\quad \gamma_0 X'\dot{h}(t + X'\alpha_0)/a(\phi_0) d\eta_0(t) + f_{\theta_0,\phi_0}(Y \mid C, X)\dot{\eta}_0(C - X'\alpha_0)X'\Bigg\}\Bigg].
\end{aligned}
$$

Finally, we obtain

$$n^{1/2}\{(\Psi_{n,\theta_0}-\Psi_{\theta_0})(\phi_0,\alpha_0,\eta_0)\} = \mathbb{G}_n\Bigg(\Delta\{Y-\dot{b}(D'(T)\theta_0)\}D(T) \tag{3.29}$$

$$+(1-\Delta)\Bigg\{\int_{C-X'\alpha_0}^{\tau} f_{\theta_0,\phi_0}(Y\mid t+X'\alpha_0,X)d\eta_0(t)\Bigg\}^{-1}$$

$$\int_{C-X'\alpha_0}^{\tau} f_{\theta_0,\phi_0}(Y\mid t+X'\alpha_0,X)\{Y-\dot{b}(D'(t+X'\alpha_0)\theta_0)\}D(t+X'\alpha_0)d\eta_0(t)\Bigg),$$

$$=\mathbb{G}_n\{G_1(\theta_0,\phi_0,\alpha_0,\eta_0,\Delta,Y,X,V)\}$$

and

$$n^{1/2}\dot{\Psi}_{2,\theta_0}(\phi_0,\alpha_0,\eta_0)(\hat{\phi}_n-\phi_0)\} \tag{3.30}$$
$$=\mathbb{G}_n\Bigg\{\dot{\Psi}_{2,\theta_0}(\phi_0,\alpha_0,\eta_0)m_4(\theta_0,\Delta,Y,X,V)\Bigg\}+o_p(1),$$

where $n^{1/2}(\hat{\phi}_n-\phi_0) = \mathbb{G}_n m_4(\theta_0,\phi_0,Y,X)+o_p(1)$ with $m_4(\theta_0,\phi_0,Y,X) = \Delta\{Y - D'(T)\theta_0\}^2$ for linear regression and $m_4=0$ for the logistic and Poisson regressions. For Gehan weighted estimate $\hat{\alpha}_n$, we have

$$n^{1/2}\dot{\Psi}_{3,\theta_0}(\phi_0,\alpha_0,\eta_0)(\hat{\alpha}_n-\alpha_0)=\mathbb{G}_n\Bigg\{\dot{\Psi}_{3,\theta_0}(\phi_0,\alpha_0,\eta_0)m_3(\alpha_0,\epsilon_0;\Delta,X)\Bigg\}+o_p(1) \tag{3.31}$$

Furthermore, from (3.15) and (3.28) we obtain

$$n^{1/2}\dot{\Psi}_{4,\theta_0}(\phi_0,\alpha_0,\eta_0)(\hat{\eta}_{n,\hat{\alpha}_n}-\eta_0) \tag{3.32}$$
$$=\mathbb{G}_n\Bigg[-\int_{\mathcal{X}}\int_{-\infty}^{\infty}(1-\Delta)\Bigg\{f_{\theta_0,\phi_0}(y\mid\tau+x'\alpha_0,x)\Bigg(\{1-\eta_0(\tau)\}\{m_1(\alpha_0,\epsilon_0;\tau)$$
$$+m_2(\alpha_0,\epsilon_0;\tau,\Delta)\}+\dot{F}_\alpha(\alpha,\tau)m_3(\alpha_0,\epsilon_0;x,\Delta)\Bigg)$$
$$-f_{\theta_0,\phi_0}(y\mid C,x)\Bigg(\{1-\eta_0(C-x'\alpha_0)\}\{m_1(\alpha_0,\epsilon_0;C-x'\alpha_0)$$

$$
\begin{aligned}
&+m_2(\alpha_0,\epsilon_0;C-x'\alpha_0,\Delta)\}+\dot{F}_\alpha(\alpha,C-x'\alpha_0)m_3(\alpha_0,\epsilon_0;x,\Delta)\Big)\\
&-\int_{C-x'\alpha_0}^{\tau} f_{\theta_0,\phi_0}(y\mid t+x'\alpha_0,x)\gamma_0\dot{h}(t+x'\alpha_0)\{y-\dot{b}(D'(t+x'\alpha_0)\theta_0)\}/a(\phi_0)\\
&\Big(\{1-\eta_0(t)\}\{m_1(\alpha_0,\epsilon_0;t)+m_2(\alpha_0,\epsilon_0;t,\Delta)\}+\dot{F}_\alpha(\alpha,t)m_3(\alpha_0,\epsilon_0;x,\Delta)\Big)dt\Big\}\\
&\Big(\int_{C-x'\alpha_0}^{\tau} f_{\theta_0,\phi_0}(y\mid t+x'\alpha_0,x)\{y-\dot{b}(D'(t+x'\alpha_0)\theta_0)\}D(t+x'\alpha_0)d\eta_0(t)\Big)\\
&\Big(\int_{C-x'\alpha_0}^{\tau} f_{\theta_0,\phi_0}(y\mid t+x'\alpha_0,x)d\eta_0(t)\Big)^{-2}dydF_X(x)\Big]+o_p(1)\\
=\;&\mathbb{G}_n\{G_2(\theta_0,\phi_0,\alpha_0,\eta_0,\Delta,Y,X,V)\}
\end{aligned}
$$

Hence, $(\Psi_{n,\theta_0}-\Psi_{\theta_0})(\phi_0,\alpha_0,\eta_0)+\dot{\Psi}_{2,\theta_0}(\phi_0,\alpha_0,\eta_0)(\hat{\phi}_n-\phi_0)+\dot{\Psi}_{3,\theta_0}(\phi_0,\alpha_0,\eta_0)(\hat{\alpha}_n-\alpha_0)+\dot{\Psi}_{4,\theta_0}(\phi_0,\alpha_0,\eta_0)(\hat{\eta}_{n,\hat{\alpha}_n}-\eta_0)$ is the sum of independent and identically distributed terms and the classical central limit theorem applies. We have $\sqrt{n}(\hat{\theta}_n-\theta_0)$ converges weakly to a mean zero normal random variable with variance $A^{-1}BA^{-1}$, where

$$
\begin{aligned}
A &= -\dot{\Psi}_{1,\theta_0}(\phi_0,\alpha_0,\eta_0),\\
B &= \Big\{G_1(\theta_0,\phi_0,\alpha_0,\eta_0,\Delta,Y,X,V)+\dot{\Psi}_{2,\theta_0}(\phi_0,\alpha_0,\eta_0)m_4(\theta_0,\Delta,Y,X,V)\\
&\quad+\dot{\Psi}_{3,\theta_0}(\phi_0,\alpha_0,\eta_0)m_3(\alpha_0,\epsilon_0;\Delta,X)+G_2(\theta_0,\phi_0,\alpha_0,\eta_0,\Delta,Y,X,V)\Big\}^{\otimes 2}.
\end{aligned}
$$

Note that for other rank based estimates of α, m_3 in B is the corresponding influence function with different forms; For the sieve maximum likelihood estimates (Ding and Nan, 2011), m_3 is the efficient influence function (Ritov and Wellner, 1988). It is clearly seen that the analytic form of the asymptotic variance is too complicated to be useful for the asymptotic variance estimation. Hence in our numerical studies we use bootstrap to obtain the variance estimator. □

CHAPTER IV

Conditional Modeling of Longitudinal Data with Terminal Event

4.1 INTRODUCTION

In longitudinal studies, the collection of information can be stopped at the end of the study, or at the time of dropout of a study participant, or at the time of a terminal event. Death, the most common terminal event, often occurs in aging cohort studies and fatal disease follow-up studies, e.g., organ failure or cancer studies. Other types of terminal events also exist, for example, the final menstrual period is a terminal event for menstrual cycle data.

On primary focus of the current literature is how longitudinal measures affect survival time, and a popular method is the joint modeling approach using latent frailty, see e.g. Tsiatis and Davidian (2004) and Albert and Shih (2010). Such joint modeling strategy has been applied to longitudinal analysis with terminal event, see e.g. Huang and Wang (2004) and Ding and Wang (2008). Another widely used approach is the marginal estimating equation approach using inverse probability weighting, see e.g. Ghosh and Lin (2002). In either case, the relationship between the response variable and the covariates described by the fixed effects is the same no matter whether the terminal event occurs or not. In other words, an implicit assumption of such analyses is that the longitudinally measured response variable and covariates are stochastic processes holding the same relationship (i.e. fixed effects in the regression model) as would exist if not stopped by the terminal event. Thus the observed data terminated by the terminal event time is more or less treated as incomplete data.

These type of modeling strategies, however, are not reasonable for many longitudinal studies, where the explicit effect of terminal event time is of interest. For example, medical payments in dialysis patients (Liu et al., 2007) increase when patients approach death; functional limitations in an aging population (Sowers et al., 2007) become more severe when people are closer to the end of life; and menstrual cycles become longer and more variable when women approach the end of their reproductive life – menopause (Harlow et al., 2008). Consequently, the longitudinal data observed up to the terminal event time should be considered as complete data rather than incomplete data.

Therefore, we directly model event time as an additional covariate for repeated measures which provides much more intuitive and meaningful interpretation. The proposed model has the usual relationship of interest between the longitudinally measured response variable and covariates when the data collection time is far away from the terminal event time; whereas, the regression parameters become increasingly related to the terminal event time when the data collection time is closer to the terminal event. The parameter estimation in such models can be complicated when the terminal event times can be right censored. We propose a semiparametric likelihood based approach to a nonlinear regression model with a censored covariate. The tail distribution of the terminal event beyond each observed censoring time is estimated from the Cox regression model, conditional on all other covariates. Model checking is implemented by using martingale residuals. The proposed method is shown to be consistent and asymptotically normal, and outperforms complete case analysis in simulations, which simply eliminates subjects with censored terminal event times. The proofs of the asymptotic properties rely heavily on empirical process theory.

4.2 NONLINEAR REGRESSION MODEL WITH MIXED EFFECTS AND CENSORED COVARIATE

4.2.1 Complete data model with observed terminal event time

For a subject i, denote the terminal event time by S_i, the baseline covariates by a vector $\vec{X}_i$ with length p, the response by Y_{ij}, and the prespecified visit time by t_{ij}, where $i = 1, \cdots, n$, $j = 1, \cdots, n_i$. When S_i is observed, we can model Y_{ij} by the following nonlinear model with mixed effects for longitudinal data:

$$Y_{ij} = \vec{X}_i{}'\beta + \gamma e^{-(S_i - t_{ij} - \mu)^2 \xi} + \vec{Z}_i{}' b_i + U_i(t_{ij}) + \varepsilon_{ij}, \tag{4.1}$$

where β is a vector of regression coefficients with length p, b_i are independent random effects (vectors with length q_1) associated with covariates $\vec{Z}_i$, $U_i(t)$ are independent stochastic processes, and ε_{ij} are independent measurement errors.

We further assume that, for each subject i, (i) b_i follows a normal distribution $N(0, D(\varphi))$, where D is a positive definite matrix depending on a parameter vector φ with length q_2; (ii) $U_i(t)$ is a mean zero Gaussian process with covariance function $cov(U_i(t_1), U_i(t_2)) = \kappa(\nu, \rho; t_1, t_2)$, where $\kappa(\cdot)$ is a given function that depends on a parameter vector ν with length q_3 and a scalar ρ; for example, $U_i(t)$ can be the nonhomogeneous Ornstein-Uhlenbeck (NOU) process satisfying $var(U_i(t)) = \nu(t)$ with $\log(\nu(t)) = \nu_0 + \nu_1 t$ and $corr(U_i(t_1), U_i(t_2)) = \rho^{|t_1 - t_2|}$; (iii) ε_{ij} follows a normal distribution $N(0, \sigma^2)$; and (iv) b_i, $U_i(t)$, and ε_{ij} are mutually independent.

For a vector $b = (b_1, \cdots, b_m)$, denote $ab = (ab_1, \cdots, ab_m)$ for a scalar a, $b^2 = (b_1^2, \cdots, b_m^2)$ and $\exp(b) = (\exp(b_1), \cdots, \exp(b_m))$. Let $Y_i = (Y_{i1}, \cdots, Y_{in_i})'$, $t_i = (t_{i1}, \cdots, t_{in_i})$, $X_i = (\vec{X}_i, \cdots, \vec{X}_i)'_{p \times n_i}$ and $Z_i = (\vec{Z}_i, \cdots, \vec{Z}_i)'_{q_1 \times n_i}$. When S_i is observed, from (4.1) we have

$$f_{\theta,\phi}(Y_i|S_i, X_i) = \frac{1}{(2\pi)^{n_i/2}|\Sigma_i|^{1/2}} \tag{4.2}$$
$$\exp\Big\{ -\frac{1}{2}\big(Y_i - X_i\beta - \gamma e^{-(S_i 1_i - t_i - \mu 1_i)^2\xi}\big)'\Sigma_i^{-1}\big(Y_i - X_i\beta - \gamma e^{-(S_i 1_i - t_i - \mu 1_i)^2\xi}\big)\Big\},$$

where $1_i = (1, \cdots, 1)'$ with length n_i, $\theta = (\beta, \mu, \gamma, \xi)'$ with length $p + 3$, $\phi = (\varphi, \nu, \rho, \sigma^2)'$ with length $q = q_2 + q_3 + 2$, and $\Sigma_i = Z_i D Z_i' + \Gamma_i + \sigma^2 I_i$, where I_i is the $n_i \times n_i$ identity matrix and Γ_i is the covariance matrix of $(U(t_{i1}), \cdots, U(t_{in_i}))'$.

In model (4.1), the nonlinear predictor is the normal kernel, which is minimal when $S_i - t_{ij}$ is large; and the regression parameters become increasingly related to the terminal event when t_{ij} is close to the terminal event S_i. More general regression models can be considered, for example, time-dependent covariates and the semiparametric mixed effect model with a nonparametric smooth function of the terminal event time and the data collection time. We focus on the simpler model (4.1) to better illustrate the proposed methodology. The time-dependent covariate case involves predicting censored covariate process, which will be explored elsewhere.

4.2.2 Observed data model with potentially censored terminal event time

We denote the censoring time for ith subject by C_i. If $S_i \leq C_i \equiv t_{in_i}$, then S_i is observed; otherwise S_i is right-censored by C_i. We denote the observed time by $V_i = \min(S_i, C_i)$ and the censoring indicator by $\Delta_i = 1(S_i \leq C_i)$. Note that $t_{ij} \leq V_i$, for all $i = 1, \cdots, n$, $j = 1 \cdots, n_i$. Here, we assume that C_i and (S_i, Y_i) are conditionally independent given X_i.

For notational simplicity, assume that random effect Z is a sub-vector of X. For a single subject, we observe (V, Δ, Y, X). Denote the conditional cumulative distribution function of S given X by $F_1(s|X)$ with density $f_1(s|X)$. The likelihood function for the observed data (V, Δ, Y, X) can be factorized into

$$f(V, \Delta, Y, X) = f_2(V, \Delta|Y, X) f_3(Y|X) f_4(X),$$

where f denotes the joint density of (V, Δ, Y, X), f_2 denotes the conditional density of (V, Δ) given (Y, X), f_3 denotes the conditional density of Y given X, and f_4 denotes the marginal density of X. Since the conditional independence of C and (S, Y) given X implies that C and S are conditionally independent given (Y, X), we have

$$f_2(V, \Delta|Y, X) = \{f_S(S|Y, X)\bar{G}_C(S|Y, X)\}^{\Delta}\{\bar{F}_S(C|Y, X)g_C(C|Y, X)\}^{1-\Delta}, \quad (4.3)$$

where f_S denotes the conditional density of S given (Y, X), g_C denotes the conditional density of C given (Y, X), with $\bar{F}_S$ and $\bar{G}_C$ as the corresponding conditional survival function. Further assuming noninformative censoring, we can drop $g_C(C|Y, X)$ and $G_C(C|Y, X)$. Going through conditional arguments using the Bayes' rule and dropping $f_4(X)$, we obtain the likelihood function

$$L(V, \Delta, Y, X) = \{f_{\theta,\phi}(Y|S, X)f_1(S|X)\}^{\Delta}\left\{\int_C^{\infty} f_{\theta,\phi}(Y|s, X)dF_1(s|X)\right\}^{1-\Delta} \quad (4.4)$$

where only $f_{\theta,\phi}$ contains the parameter of interest θ and nuisance parameter ϕ, whereas f_1 is an additional nuisance parameter in addition to ϕ.

In (4.4), $\{f_{\theta,\phi}(Y|S, X)f_1(S|X)\}^{\Delta}$ is for a subject with observed terminal event time, which yields the fully observed data, and $\left\{\int_C^{\infty} f_{\theta,\phi}(Y|s, X)dF_1(s|X)\right\}^{1-\Delta}$ is for a subject with the terminal event time being censored. Later we show that the complete case analysis by dropping the second part in (4.4) yields consistent and asymptotically normally distributed estimator, but it loses efficiency. Making use of censored data can improve efficiency. We see from the second part of (4.4) that the amount efficiency gain depends on how well we can recover the right tail of the conditional distribution $F_1(s|X)$ beyond C. We consider a semiparametric approach that allows reliable extrapolation beyond C and is robust against any parametric assumption.

Due to the randomness of C, all the commonly used semiparametric models for right-censored data allow extrapolation beyond C. Here, we propose the most widely used Cox regression model (Cox, 1972). Other viable models include accelerated failure time model, additive hazard model, and transformation model (Kalbfleisch and Prentice, 2002). Suppose the hazard function of S given X has the following form:

$$\lambda(s|X) = \lambda(s)\exp(\alpha' X), \tag{4.5}$$

where α is the regression parameter with an unknown true value α_0, and $\lambda(\cdot)$ is the baseline hazard function. The conditional survival function is then given by

$$\eta(s; X) \equiv F_1(s|X) = 1 - \exp\{-\Lambda(s)\exp(\alpha' X)\},$$

where $\Lambda(s) = \int_0^s \lambda(u)du$ is the cumulative baseline hazard function with an unknown true value Λ_0. Note that X appears in both models (4.2) and (4.5), but it may refer to different forms of covariates in these models. For example, X_1 is a covariate in (4.2) whereas X_1^2 is a covariate in (4.5). We use the same X to denote all fully observed covariates for notational simplicity. The log-likelihood function then becomes

$$\begin{aligned}\log L = {} & \Delta \log f_{\theta,\phi}(Y|S,X) + \Delta \log \dot{\eta}(S;X) \\ & + (1-\Delta)\log \int_C^\tau f_{\theta,\phi}(Y|u,X)d\eta(u;X),\end{aligned} \tag{4.6}$$

where τ is the truncation time defined in Condition IV.5 provided in Section 4.4.

A similar idea has been used by Lu et al. (2010), but for a different problem. Lu et al. (2010) considered longitudinal data analysis with an event time, which does not terminate the observed data.

4.3 THE PSEUDO-LIKELIHOOD METHOD

The log likelihood function (4.6) involves an unknown distribution function η and the corresponding density function $\dot{\eta}$, hence a maximum likelihood estimation, if it exists, can be complicated. We propose a tractable two-stage pseudo-likelihood approach in which the nuisance parameters $(\phi, \eta(\alpha, \Lambda))$ are estimated in stage 1 and the parameter of interest θ is then estimated by maximizing the data version of (4.6) in stage 2, with nuisance parameters replaced by their estimators obtained in stage 1 before maximization. Details are given below:

Stage 1. Nuisance parameter estimation. The dispersion parameter ϕ is estimated by the complete case analysis of the nonlinear regression model (4.2); the Cox model regression coefficient α is estimated by maximizing the partial likelihood, denoted by $\tilde{\alpha}_n$; and the cumulative baseline hazard Λ is estimated by Breslow estimates $\tilde{\Lambda}_n$ (Breslow and Crowley, 1974). The c.d.f $\eta(s; X)$ is estimated by $\tilde{\eta}_n(s; X) = 1 - \exp\{-\tilde{\Lambda}_n(s) \exp(\tilde{\alpha}_n' X)\}$, which is asymptotically equivalent to the product integral expression. It can be shown that all the estimates obtained in Stage 1 have desirable statistical properties. In particular, $\tilde{\eta}_n$ is $n^{1/2}$ -consistent in a finite interval, see Lemma IV.3 in Subsection 4.6.2; $\tilde{\phi}_n$ obtained from the complete case analysis is $n^{1/2}$ -consistent, see Lemma IV.5 in Subsection 4.6.2.

Stage 2. Pseudo-likelihood estimation of θ. Replacing (ϕ, η) by their Stage 1 estimates $(\tilde{\phi}_n, \tilde{\eta}_n)$ in the log likelihood function yields the following log pseudo-likelihood function for a random sample of n subjects:

$$pl_n(\theta) = \frac{1}{n}\sum_{i=1}^{n}\Big\{\Delta_i \log f_{\theta,\tilde{\phi}_n}(Y_i|S_i,X_i) + (1-\Delta_i)\log\int_{C_i}^{\tau} f_{\theta,\tilde{\phi}_n}(Y_i|u,X_i)d\tilde{\eta}_n(u;X_i)\Big\}. \quad (4.7)$$

Note that the term $\Delta \log \dot{\eta}$ in (4.6) is dropped because it does not involve θ. However, if we want to maximize the log-likelihood directly without using the two-stage approach, then this term can not be omitted.

Let $\hat{\theta}_n$ denote the pseudo-likelihood estimator. Since it is obtained by maximizing the objective function (4.7), its asymptotic properties can be obtained from M-estimation theory, see van der vaart (2002), Wellner and Zhang (2007) and Li and Nan (2011).

The estimates $(\tilde{\eta}_n, \tilde{\Lambda}_n)$ are obtained using a standard package for the Cox regression model. The estimates $(\tilde{\theta}_n, \tilde{\phi}_n)$ from complete case analysis are obtained by maximizing $\frac{1}{n}\sum_{i=1}^{n}\left\{\Delta_i \log f_{\theta,\phi}(Y_i|S_i,X_i)\right\}$ using Newton-Raphson algorithm, where multiple initial values are tried. The two-stage estimator $\hat{\theta}_n$ is also obtained by Newton-Raphson algorithm with the initial value $\tilde{\theta}_n$ gained from complete case analysis.

4.4 ASYMPTOTIC PROPERTIES

Define

$$l_0(\theta,\phi;Y,X,\Delta,V) = \Delta \log f_{\theta,\phi}(Y|S,X), \quad (4.8)$$

which is the first part in the following log-likelihood for the observed data:

$$
\begin{aligned}
& l(\theta, \phi, \eta; Y, X, \Delta, V) \equiv l(\theta, \phi, \eta(\alpha, \Lambda); Y, X, \Delta, V) \\
& = \Delta \log f_{\theta,\phi}(Y|S, X) + (1-\Delta) \log \int_C^\tau f_{\theta,\phi}(Y|u, X) d\eta(u; X) \\
& = \Delta \log f_{\theta,\phi}(Y|S, X) \\
& \quad +(1-\Delta) \log \int_C^\tau f_{\theta,\phi}(Y|u, X) d[1 - \exp\{-\Lambda(u) \exp(\alpha' X)\}], \qquad (4.9)
\end{aligned}
$$

which is (4.6) with $\Delta \log \dot{\eta}$ dropped.

Denote the true value of θ by θ_0, the true value of ϕ by ϕ_0, the sample space of response variable Y by $\mathcal{Y}$, the sample space of covariate X by $\mathcal{X}$, the sample space of random effect Z by $\mathcal{Z} \subset \mathcal{X}$, the parameter space of θ by Θ, the parameter space of ϕ by Φ, and the parameter space of η by $\mathcal{F}$. In addition to the assumptions of bounded support for X, bounded parameter spaces Θ and Φ, and conditional independence between C and (S, Y) given X, we provide a set of regularity conditions in the following:

Condition IV.1. *(a)* $El_0(\theta, \phi; Y, X, \Delta, V)$ *has a unique maximizer* (θ_0, ϕ_0)*; (b)* $El(\theta, \phi_0, \eta_0; Y, X, \Delta, V)$ *has a unique maximizer* θ_0*.*

Condition IV.2. *The eigenvalues for* $\Sigma(\phi)$ *are bounded between* $[\lambda_1, \lambda_2]$*, where* $0 < \lambda_1 < \lambda_2 < \infty$ *for any* $\phi \in \Phi$ *and* $Z \in \mathcal{Z}$*.*

Condition IV.3. *The absolute values of all the elements in* $\frac{\partial \Sigma(\phi)}{\partial \phi_k}$ *and* $\frac{\partial^2 \Sigma(\phi)}{\partial \phi_j \partial \phi_k}$ *are bounded uniformly for all* $\phi \in \Phi$ *and* $Z \in \mathcal{Z}$*.*

Condition IV.4. *The absolute values of all the elements in* $\frac{\partial^3 \Sigma(\phi)}{\partial \phi_i \partial \phi_j \partial \phi_k}$ *are bounded uniformly for all* $\phi \in \Phi$ *and* $Z \in \mathcal{Z}$*.*

Condition IV.5. *The study stops at a finite time* $\tau > 0$ *such that* $\inf_{x \in \mathcal{X}} P(C \geq \tau | X = x) = \omega_1 > 0$ *and* $\inf_{x \in \mathcal{X}} P(S \geq \tau | X = x) = \omega_2 > 0$ *for constants* ω_1 *and* ω_2*.*

Condition IV.6. *The condition distribution of S given X possesses a continuous Lebesgue density.*

Condition IV.7. *The information matrix of the Cox regression model at the true parameter values is positive definite.*

Condition IV.8. *There exist constants $\delta_1 > 0$ and $\delta_2 > 0$, such that $\int_C^\tau f_{\theta,\phi}(Y|s,X)$ $d\eta(s) \geq \delta_1$ with probability 1 for any $\theta \in \Theta$ and $|\phi - \phi_0| + \|\eta - \eta_0\| < \delta_2$.*

REMARK: Condition IV.1(a) implies $\gamma_0\xi_0 \neq 0$; it holds by Theorem 2.1 of Lehmann (1998) provided model (4.1) is identifiable. Condition IV.1(b) is for the consistency of the proposed two-stage estimator $\hat{\theta}_n$, which may be unnecessarily strong and can be seen from the following. In the proof of Theorem IV.2 in Subsection 4.6.3.2, we can show $P\ddot{l}_{11}(\theta_0, \phi_0, \eta_0; Y, X, \Delta, V) = P\left\{ \frac{\partial^2 l(\theta,\phi_0,\eta_0;Y,X,\Delta,V)}{\partial\theta\partial\theta'} \Big|_{\theta=\theta_0} \right\}$ is negative definite by Condition IV.1(a). Thus $P\ddot{l}_{11}(\theta, \phi_0, \eta_0; Y, X, \Delta, V)$, a continuous matrix of θ, is also negative definite in a neighborhood of θ_0, which guarantees that θ_0 is a unique maximizer of $Pl(\theta, \phi_0, \eta_0; Y, X, \Delta, V)$ in a neighborhood of θ_0. The initial value we use in the algorithm for maximizing (4.7) is obtained from the complete case analysis, which is shown to be $n^{1/2}$ -consistent; thus, the solution of the proposed two-stage method is likely to be in the same neighborhood, and therefore also consistent without the uniqueness requirement in Condition IV.1(b).

Conditions IV.2-IV.4 automatically hold for model (4.1) with NOU process if $|\rho| \leq 1 - \delta$, and $t_{i,k+1} - t_{i,k} \geq \varepsilon$, $i = 1, \cdots, n$, $k = 1, \cdots, n_i - 1$, where $\delta > 0$ and $\varepsilon > 0$. And they are parallel to the conditions of bounded derivatives of the log likelihood in Theorem 1.1 and Theorem 2.3 of Lehmann (1998).

Conditions IV.5-IV.7 are usual assumptions for Cox regression models (Andersen and Gill, 1982; Nan and Wellner, 2013). Condition IV.8 is mainly for technical convenience. One way to obtain Condition IV.8 might be to truncate response variable Y such that $|Y| \leq M < \infty$ for a large constant M. In our simulations, however, we do not implement such truncations but still obtain satisfactory results.

We now have the following asymptotic results for $\hat{\theta}_n$:

Theorem IV.1. *(Consistency) Under Conditions IV.1-IV.3 and IV.5-IV.8, the two-stage pseudo-likelihood estimator* $\hat{\theta}_n$ *that maximizes (4.7) converges in outer probability to* θ_0*.*

The proof of Theorem IV.1 is similar to Li and Nan (2011) and van der vaart (2002). Details are provided in Subsection 4.6.3.1.

Theorem IV.2. *(Asymptotic normality) Under Conditions IV.1-IV.8,* $\sqrt{n}(\hat{\theta}_n - \theta_0)$ *converges weakly to a mean zero normal random variable with variance* $A^{-1}BA^{-1}$*, where* A *and* B *are provided in Subsection 4.6.3.2.*

The proof of Theorem IV.2 is based on the general M-estimation theory similar to Li and Nan (2011) and Wellner and Zhang (2007) which is given in Subsection 4.6.1. The detailed proof relies heavily on empirical process theory and is given in the Subsection 4.6.3.2.

Because the asymptotic variance of $\hat{\theta}_n$ has a very complicated expression that prohibits the direct calculation of its estimate from observed data, we use the bootstrap variance estimator.

4.5 SIMULATIONS

We conduct simulations to investigate the finite sample performance of the proposed method. Simulation data sets are generated from the nonlinear model with mixed effects:

$$Y_{ij} = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \gamma e^{-(S_i - T_{ij} - \mu)^2 \xi} + b_i + U_i(T_{ij}) + \varepsilon_{ij},$$

where $\beta_0 = 1$, $\beta_1 = 1$, $\beta_2 = -3$, $\mu = 1$, and $\gamma = 4$. The random effect $b_i \sim N(0, \exp(-0.5))$, the error term $\varepsilon_{ij} \sim N(0, \exp(-0.1))$, and $U_i(t)$ is an NOU process with $\nu_0 = 1$, $\nu_1 = -1$ and $\rho = \exp(-1)/(1 + \exp(-1))$. The two fully observed covariates are X_{1i} and X_{2i}, where $X_{1i} \sim$ Bernoulli(0.5) and X_{2i} follows $N(0,1)$ distribution truncated at ± 3. Terminal event time $S_i = 4 + S_{0i}$, where S_{0i} follows an exponential distribution with a conditional hazard function $\exp(-1 - 6X_{1i} + 4X_{2i})$. To generate the censoring time C_i, we first generate $C_{0i} = \kappa C^*_{0i}$, where C^*_{0i} follows an exponential distribution with a conditional hazard function $\exp(-3 - X_{1i} + X_{2i})$, then set $C_i = t_{ij}$, where j satisfies $t_{ij} \le C_i$ and $t_{ij+1} > C_i$ assuming $t_{in_i+1} = \infty$. The constant κ is chosen to yield 40% censoring. For each subject i, there are 10 scheduled visit times, and the first visit time t_{i1} is 0. There are two different settings to generate the subsequent visit times: (1) equally spaced time intervals with $t_{ij} = j-1$, $j = 2, \cdots, 10$; (2) non-equally spaced time intervals with the subsequent visit times generated recursively from $t_{ij} = t_{ij-1} + \min(4, W_i)$ for $j = 2, \cdots, 10$, where W_i follows an exponential distribution with a conditional hazard function $\exp(-3 - X_{1i} + X_{2i})$. In each setting, ξ takes two different values 1.2 and 0.2, corresponding to a flat and a sharp nonlinear predictor in the regression model, respectively.

We conduct simulations with sample size 300, which mimic a medical payment study in Liu et al. (2007), and simulate 500 replications for each scenario. The proposed method is compared to full data analysis and complete case analysis in terms of biases and variances. The full data analysis represents the case that all data are available; in other words, there is no censoring, which has more visits and serves as a benchmark. The complete case analysis simply eliminates subjects with censored terminal event time. For the proposed two-stage method, we report the 90% and 95% coverage proportions for which the variances estimators are obtained from 100 bootstrap samples. The results are presented in Tables 4.1-4.4.

The results suggest that the biases for the proposed two-stage method are minimal, which is comparable to both the full data analysis and the complete case analysis. From the tables, it can be clearly seen that the proposed method is much more

Table 4.1: Simulation results for equally spaced time interval with sharp nonlinear term. var_b=boostrap variance estimator; CR=coverage rate

		$\beta_0 = 1$	$\beta_1 = 1$	$\beta_2 = -3$	$\mu = 1$	$\gamma = 4$	$\xi = 1.2$
Full data	bias	-0.0064	0.0011	-0.0002	0.0001	-0.0032	0.0031
	var	0.0801	0.0115	0.0082	0.0002	0.0140	0.0032
Two-stage	bias	-0.0153	0.0045	0.0003	-0.0010	-0.0030	0.0039
	var	0.0973	0.0144	0.0098	0.0003	0.0166	0.0040
	var_b	0.1094	0.0161	0.0102	0.0003	0.0151	0.0043
	90% CR	0.904	0.876	0.898	0.918	0.896	0.912
	95% CR	0.966	0.944	0.960	0.972	0.952	0.946
Complete case	bias	-0.0092	0.0010	0.0041	-0.0008	-0.0024	0.0024
	var	0.1217	0.0208	0.0130	0.0003	0.0242	0.0053

Table 4.2: Simulation results for equally spaced time interval with flat nonlinear term. var_b=boostrap variance estimator; CR=coverage rate

		$\beta_0 = 1$	$\beta_1 = 1$	$\beta_2 = -3$	$\mu = 1$	$\gamma = 4$	$\xi = 0.2$
Full data	bias	0.0081	-0.0116	0.0209	-0.0007	0.0009	0.0001
	var	0.1279	0.0130	0.0455	0.0009	0.0121	0.0005
Two-stage	bias	0.0090	-0.0121	0.0229	0.0003	-0.0046	0.0010
	var	0.1599	0.0161	0.0561	0.0014	0.0152	0.0007
	var_b	0.1745	0.0166	0.0641	0.0015	0.0152	0.0007
	90% CR	0.902	0.890	0.912	0.904	0.910	0.902
	95% CR	0.958	0.948	0.946	0.950	0.964	0.952
Complete case	bias	0.0044	-0.0196	0.0463	-0.0004	-0.0062	-0.0004
	var	0.2248	0.0239	0.0821	0.0017	0.0236	0.0008

efficient than the complete case analysis, and the bootstrap method performs well in estimating the variance, which yields reasonable coverage rates of the confidence intervals for all the scenarios. Note that, in the case of non-equally spaced time intervals, the variance estimates for parameters in the flat nonlinear predictor is less accurate, see Table 4.4. However, the coverage rates of their confidence intervals become more accurate when sample size increases (additional simulation results not provided).

Table 4.3: Simulation results for non-equally spaced time interval with sharp nonlinear term var_b=boostrap variance estimator; CR=coverage rate

		$\beta_0 = 1$	$\beta_1 = 1$	$\beta_2 = -3$	$\mu = 1$	$\gamma = 4$	$\xi = 1.2$
Full data	bias	-0.0045	0.0020	0.0068	0.0016	-0.0051	0.0059
	var	0.1474	0.0203	0.0187	0.0005	0.0177	0.0084
Two-stage	bias	-0.0238	0.0089	0.0111	0.0015	-0.0046	0.0090
	var	0.1675	0.0235	0.0253	0.0007	0.0229	0.0113
	var_b	0.1550	0.0220	0.0276	0.0007	0.0213	0.0135
	90% CR	0.866	0.884	0.884	0.888	0.914	0.910
	95% CR	0.942	0.936	0.938	0.944	0.950	0.960
Complete case	bias	-0.0295	0.0093	0.0217	0.0004	0.0002	0.0120
	var	0.2480	0.0366	0.0353	0.0010	0.0340	0.0161

Table 4.4: Simulation results for non-equally spaced time interval with flat nonlinear term. var_b=boostrap variance estimator; CR=coverage rate

		$\beta_0 = 1$	$\beta_1 = 1$	$\beta_2 = -3$	$\mu = 1$	$\gamma = 4$	$\xi = 0.2$
Full data	bias	-0.0291	-0.0050	0.0459	-0.0048	-0.0083	0.0005
	var	0.2087	0.0198	0.0875	0.0038	0.0189	0.0014
Two-stage	bias	-0.0521	-0.0018	0.0642	-0.0052	-0.0076	0.0014
	var	0.2609	0.0253	0.1102	0.0055	0.0235	0.0019
	var_b	0.3059	0.0224	0.1416	0.0062	0.0216	0.0017
	90% CR	0.894	0.886	0.882	0.892	0.874	0.868
	95% CR	0.944	0.938	0.924	0.940	0.910	0.910
Complete case	bias	-0.0749	-0.0026	0.0984	-0.0025	-0.0104	0.0015
	var	0.3471	0.0363	0.1563	0.0064	0.0314	0.0024

4.6 Appendix

4.6.1 General M-theorems

Lemma IV.1 is a general M-estimation theory for parametric model , see van der Geer (2000); and it is a special case of Wellner and Zhang (2007), thus its proof is omitted. Lemma IV.2, which is used for the proof of Theorem IV.2, is similar to Theorem A.1 of Li and Nan (2011), but the former focuses on infinite-dimensional nuisance parameters; while the latter focuses on finite-dimensional nuisance parameters. Note that Lemma IV.2 reduces to Lemma IV.1 with the nuisance parameters fixed at true parameters. We provide Lemma IV.1 here for the ease of reference in the proofs for complete case analysis which is given in Lemma IV.5. Let $|\cdot|$ be the Euclidian norm and $\|\eta - \eta_0\| = \sup_{s,x} |\eta(s;x) - \eta_0(s;x)|$. We adopt the empirical process notation of van der Vaart and Wellner (1996).

Lemma IV.1. *(Asymptotic normality for M-estimation) Given i.i.d. observation $\boldsymbol{X}_i$, $i = 1, \cdots, n$. Suppose that the estimates $\tilde{\psi}_n$ of unknown parameters ψ are set to be maximizer of the objective function $\mathbb{P}_n m(\psi; \boldsymbol{X})$. Let $\dot{m}(\psi; \boldsymbol{X}) = \frac{\partial m(\psi;\boldsymbol{X})}{\partial \psi}$ and $\ddot{m}(\psi; \boldsymbol{X}) = \frac{\partial^2 m(\psi;\boldsymbol{X})}{\partial \psi \psi'}$. Consider the following conditions:*

A1. $|\tilde{\psi}_n - \psi_0| = o_p(1)$.

A2. $A = -P\{\ddot{m}(\psi; \boldsymbol{X})\}$ *is non-singular.*

A3. $P\dot{m}(\psi_0; \boldsymbol{X}) = 0$.

A4. The estimates $\tilde{\psi}_n$ satisfy $\mathbb{P}_n \dot{m}(\psi; \boldsymbol{X}) = o_p(n^{-1/2})$.

A5. For any $\delta_n > 0$, let $\Psi_n = \{\psi : |\psi - \psi_0| \le \delta_n\}$, we have $\sup_{\psi \in \Psi_n} |\mathbb{G}_n\{\dot{m}(\psi; \boldsymbol{X}) - \dot{m}(\psi_0; \boldsymbol{X})\}| = o_p(1)$.

A6. For $\psi \in \Psi_n$, $|P\{\dot{m}(\psi; \boldsymbol{X}) - \dot{m}(\psi_0; \boldsymbol{X}) - \ddot{m}(\psi_0; \boldsymbol{X})(\psi - \psi_0)\}| = o(|\psi - \psi_0|)$.

Suppose that Conditions A1-A6 hold, then we have

$$\sqrt{n}(\tilde{\psi}_n - \psi_0) = A^{-1}\sqrt{n}\mathbb{P}_n \dot{m}(\psi_0; \boldsymbol{X}) + o_{p^*}(1).$$

Lemma IV.2. *(Asymptotic normality for pseudo M-estimation) Given i.i.d. observation* $\boldsymbol{X}_i$, $i = 1, \cdots, n$. *Suppose that the estimates* $\hat{\theta}_n$ *of unknown parameters* θ *are set to be maximizer of the objective function* $\mathbb{P}_n m(\theta, \tilde{\phi}_n, \tilde{\eta}_n; \boldsymbol{X})$, *where* $\tilde{\phi}_n$ *is an estimator of true parameter* $\phi_0 \in \Phi \subset R^d$, *and* $\tilde{\eta}_n$ *is an estimator of the true parameter* $\eta_0 \in \mathcal{H}$, *which is an infinite dimensional Banach space. Suppose that* η_t *is a parametric submodel in* $\mathcal{F}$ *passing through* η, *that is,* $\eta_t \in \mathcal{F}$ *and* $\eta_{t=0} = \eta$. *Let* $H = \left\{ h : h = \left. \frac{\partial \eta_t}{\partial t} \right|_{t=0} \right\}$ *be the collection of all directions to approach* η. *Let* $\dot{m}_1(\theta, \phi, \eta; \boldsymbol{X}) = \frac{\partial m(\theta,\phi,\eta;\boldsymbol{X})}{\partial \theta}$, $\dot{m}_2(\theta, \phi, \eta; \boldsymbol{X}) = \frac{\partial m(\theta,\phi,\eta;\boldsymbol{X})}{\partial \phi}$, *and* $\dot{m}_3(\theta, \phi, \eta; \boldsymbol{X})[h] = \frac{\partial m(\theta,\phi,\eta;\boldsymbol{X})}{\partial t}$ *along the direction of* h. *Let* $\ddot{m}_{ij}$ *be the second order derivatives of* m *with respect to corresponding arguments defined in a similar way,* $i, j \in \{1, 2, 3\}$. *Consider the following conditions:*

B1. $|\tilde{\phi}_n - \phi_0| = o_p(1)$, $|\hat{\theta}_n - \theta_0| = o_p(1)$ *and* $\|\tilde{\eta}_n - \eta_0\| = O_p(n^{-\nu})$ *for some* $\nu > 0$ *and some norm* $\| \cdot \|$.

B2. $A = -P\{\ddot{m}_{11}(\theta_0, \phi_0, \eta_0; \boldsymbol{X})\}$ *is non-singular.*

B3. $P\dot{m}_1(\theta_0, \phi_0, \eta_0; \boldsymbol{X}) = 0$.

B4. The estimator $\hat{\theta}_n$ *satisfy* $\mathbb{P}_n \dot{m}_1(\hat{\theta}_n, \tilde{\phi}_n, \tilde{\eta}_n; \boldsymbol{X}) = o_p(n^{-1/2})$.

B5. For any $\delta_n \downarrow 0$ *and constant* $C > 0$, *let* $\Theta_n = \{(\theta, \phi, \eta) : |(\theta, \phi) - (\theta_0, \phi_0)| \leq \delta_n, \|\tilde{\eta}_n - \eta_0\|_2 \leq Cn^{-\nu}\}$, *we have* $\sup_{(\theta,\phi,\eta)\in\Theta_n} |\mathbb{G}_n\{\dot{m}_1(\theta, \phi, \eta; \boldsymbol{X}) - \dot{m}_1(\theta_0, \phi_0, \eta_0; \boldsymbol{X})\}| = o_p(1)$.

B6. For some $\varsigma > 1$ *satisfying* $\varsigma\nu > 1/2$, *and for* $(\theta, \phi, \eta) \in \Theta_n$,

$$
\begin{aligned}
&|P\{\dot{m}_1(\theta, \phi, \eta; \boldsymbol{X}) - \dot{m}_1(\theta_0, \phi_0, \eta_0; \boldsymbol{X}) \\
&\quad - \ddot{m}_{11}(\theta_0, \phi_0, \eta_0; \boldsymbol{X})(\theta - \theta_0)\}| - \ddot{m}_{12}(\theta_0, \phi_0, \eta_0; \boldsymbol{X})(\phi - \phi_0) \\
&\quad - \ddot{m}_{13}(\theta_0, \phi_0, \eta_0; \boldsymbol{X})[\eta - \eta_0]\}| \\
&= o(|\theta - \theta_0|) + o(|\phi - \phi_0|) + O(\|\eta - \eta_0\|^{\varsigma}).
\end{aligned}
$$

Suppose that Conditions B1-B6 hold, then we have

$$\begin{aligned}
&\sqrt{n}(\hat{\theta}_n - \theta_0) \\
&\quad = A^{-1}\sqrt{n}\mathbb{P}_n\dot{m}_1(\theta_0, \phi_0, \eta_0; \boldsymbol{X}) + A^{-1}\sqrt{n}P\{\ddot{m}_{12}(\theta_0, \phi_0, \eta_0; \boldsymbol{X})\}(\tilde{\phi}_n - \phi_0) \\
&\qquad + A^{-1}\sqrt{n}P\{\ddot{m}_{13}(\theta_0, \phi_0, \eta_0; \boldsymbol{X})[\tilde{\eta}_n - \eta_0]\} + o_p(1).
\end{aligned}$$

Proof. By B1, B3 and B5, we have

$$\mathbb{P}_n\dot{m}_1(\hat{\theta}_n, \tilde{\phi}_n, \tilde{\eta}_n; \mathbf{X}) - P\dot{m}_1(\hat{\theta}_n, \tilde{\phi}_n, \tilde{\eta}_n; \mathbf{X}) - \mathbb{P}_n\dot{m}_1(\theta_0, \phi_0, \eta_0; \mathbf{X}) = o_p(n^{-1/2})$$

In view of B4, this reduces to

$$P\dot{m}_1(\hat{\theta}_n, \tilde{\phi}_n, \tilde{\eta}_n; \mathbf{X}) + \mathbb{P}_n\dot{m}_1(\theta_0, \phi_0, \eta_0; \mathbf{X}) = o_p(n^{-1/2}).$$

Then by B6, it follows that

$$\begin{aligned}
&P\ddot{m}_{11}(\theta_0, \phi_0, \eta_0; \mathbf{X})(\hat{\theta}_n - \theta_0) + P\ddot{m}_{12}(\theta_0, \phi_0, \eta_0; \mathbf{X})(\tilde{\phi}_n - \phi_0) \\
&\quad + P\ddot{m}_{13}(\theta_0, \phi_0, \eta_0; \mathbf{X})[\tilde{\eta}_n - \eta_0] + \mathbb{P}_n\dot{m}_1(\theta_0, \phi_0, \eta_0; \mathbf{X}) \\
&\quad + o(|\hat{\theta}_n - \theta_0|) + o(|\tilde{\phi}_n - \phi_0|) + O(\|\tilde{\eta}_n - \eta_0\|^{\mu}) = o_p(n^{-1/2}).
\end{aligned}$$

Thus,

$$\begin{aligned}
&-(A + o_p(1))(\hat{\theta}_n - \theta_0) \\
&\quad = -P\{\ddot{m}_{12}(\theta_0, \phi_0, \eta_0; \mathbf{X})(\tilde{\phi}_n - \phi_0) + \ddot{m}_{13}(\theta_0, \phi_0, \eta_0; \mathbf{X})[\tilde{\eta}_n - \eta_0]\} \\
&\qquad - \mathbb{P}_n\dot{m}_1(\theta_0, \phi_0, \eta_0; \mathbf{X}) + o_p(n^{-1/2}).
\end{aligned}$$

□

4.6.2 Technical Lemmas

Now we provide technical preparations for the proofs of Theorem IV.1 and IV.2.

In order to obtain the influence function of the conditional survival function given X in the Cox regression model, we introduce the following notation:

$$
\begin{aligned}
&W_i(s) = 1(V_i \geq s), && N_i(s) = 1(V_i \leq s, \Delta_i = 1),\\
&dA_i(s;\alpha) = W_i(s)\exp(\alpha' X_i) d\Lambda_0(s), && dM_i(s;\alpha) = dN_i(s) - dA_i(s;\alpha)\\
&\bar{M}(s) = \sum_{i=1}^n M_i(s), && J(s) = 1\left(\sum_{i=1}^n W_i(s) > 0\right)\\
&S^{(0)}(u;\alpha) = \mathbb{P}_n\{W(u)\exp(\alpha' X)\}, && s^{(0)}(u;\alpha) = P\{W(u)\exp(\alpha' X)\},\\
&S^{(1)}(u;\alpha) = \mathbb{P}_n\{XW(u)\exp(\alpha' X)\}, && s^{(1)}(u;\alpha) = P\{XW(u)\exp(\alpha' X)\},\\
&S^{(2)}(u;\alpha) = \mathbb{P}_n\{X^{\otimes 2}W(u)\exp(\alpha' X)\}, && s^{(2)}(u;\alpha) = P\{X^{\otimes 2}W(u)\exp(\alpha' X)\},\\
&\zeta(u;\alpha) = s^{(1)}(u;\alpha)/s^{(0)}(u;\alpha).
\end{aligned}
$$

Lemma IV.3. *Under Conditions IV.5-IV.7, we have*

$$
\begin{aligned}
&\sqrt{n}(\tilde{\eta}_n(t;\tilde{X}) - \eta_0(t;\tilde{X}))\\
&\quad = [1 - \eta_0(t;\tilde{X})]\exp(\alpha_0'\tilde{X})\mathbb{G}_n\{A_1(\eta_0; t, \tilde{X}; X, \Delta, V)\} + o_p(1),
\end{aligned}
$$

where

$$
\begin{aligned}
&A_1(\eta_0; t, \tilde{X}; X, \Delta, V)\\
&\quad = \left[\tilde{X}' + h(t;\alpha_0)'\right] e(\alpha_0)^{-1}\left[-\int_0^\tau \{X - \zeta(u;\alpha_0)\}\exp(\alpha_0' X)W(u)d\Lambda_0(u)\right.\\
&\qquad \left. + \frac{1(V \leq t)\Delta}{s^{(0)}(V;\alpha_0)} + \{X - \zeta(V;\alpha_0)\}\Delta - \int_0^t \frac{1}{s^{(0)}(u;\alpha_0)}\exp(\alpha_0' X)W(u)d\Lambda_0(u)\right],
\end{aligned}
$$

with $h(t;\alpha_0) = -\int_0^t \zeta(u;\alpha_0)d\Lambda_0(u)$, *and* $e(\alpha) = E\left[\Delta\frac{s^{(2)}(V;\alpha)s^{(0)}(V;\alpha) - s^{(1)}(V;\alpha)^{\otimes 2}}{s^{(0)}(V;\alpha)^2}\right]$ *which* *is the Fisher information matrix for the Cox regression.*

Proof. From Nan and Wellner (2013) or Theorem 8.3.2 of Fleming and Harrington (2005), we have

$$\begin{aligned}&\sqrt{n}(\tilde{\alpha}_n - \alpha_0) \qquad (4.10)\\ &= e(\alpha_0)^{-1}\mathbb{G}_n\Bigg[\{X - \zeta(V;\alpha_0)\}\Delta - \int_0^\tau \{X - \zeta(u;\alpha_0)\}\exp(\alpha_0' X)W(u)d\Lambda_0(u)\Bigg]\\ &\quad + o_p(1).\end{aligned}$$

From the proof of Theorem 8.3.3 in Fleming and Harrington (2005), we have

$$\sqrt{n}(\tilde{\Lambda}_n(t) - \Lambda_0(t)) = h(t;\alpha_0)'\sqrt{n}(\tilde{\alpha}_n - \alpha_0) + n^{-1/2}\int_0^t \frac{J(u)d\bar{M}(u)}{S^{(0)}(u;\alpha_0)} + o_p(1). \quad (4.11)$$

The second term in the right hand side of equation (4.11) equals

$$\begin{aligned}n^{-1/2}&\int_0^t J(u)\Bigg[\frac{1}{S^{(0)}(u;\alpha_0)} - \frac{1}{s^{(0)}(u;\alpha_0)}\Bigg]d\bar{M}(u)\\ &+n^{-1/2}\int_0^t \frac{J(u)-1}{s^{(0)}(u;\alpha_0)}d\bar{M}(u) + n^{-1/2}\int_0^t \frac{d\bar{M}(u)}{s^{(0)}(u;\alpha_0)}.\end{aligned}$$

Define

$$\begin{aligned}A(t) &= n^{-1/2}\int_0^t J(u)\Bigg[\frac{1}{S^{(0)}(u;\alpha_0)} - \frac{1}{s^{(0)}(u;\alpha_0)}\Bigg]d\bar{M}(u),\\ B(t) &= n^{-1/2}\int_0^t \frac{J(u)-1}{s^{(0)}(u;\alpha_0)}d\bar{M}(u),\end{aligned}$$

and the martingales $A(t)$ and $B(t)$ have predictable variation process:

$$
\begin{aligned}
&< A(t), A(t) > \\
&= \frac{1}{n}\sum_{i=1}^{n}\int_{0}^{t}\left\{J(u)\left[\frac{1}{S^{(0)}(u;\alpha_0)} - \frac{1}{s^{(0)}(u;\alpha_0)}\right]\right\}^2 W_i(u)\exp(\alpha_0' X)d\Lambda_0(u) \\
&= \int_{0}^{t}\left\{J(u)\left[\frac{1}{S^{(0)}(u;\alpha_0)} - \frac{1}{s^{(0)}(u;\alpha_0)}\right]\right\}^2 S^{(0)}(u;\alpha_0)d\Lambda_0(u) \xrightarrow{p} 0, \\
&< B(t), B(t) > = \frac{1}{n}\sum_{i=1}^{n}\int_{0}^{t}\frac{[J(u)-1]^2}{s^{(0)}(u;\alpha_0)^2} W_i(u)\exp(\alpha_0' X_i)d\Lambda_0(u) \\
&\qquad = \int_{0}^{t}\frac{[J(u)-1]^2}{s^{(0)}(u;\alpha_0)^2} S^{(0)}(u;\alpha_0)d\Lambda_0(u) \xrightarrow{p} 0.
\end{aligned}
$$

Hence, $A(t) \to 0$ and $B(t) \to 0$ for any t, and

$$
\sqrt{n}(\tilde{\Lambda}_n(t) - \Lambda_0(t)) = h(t;\alpha_0)'\sqrt{n}(\tilde{\alpha}_n - \alpha_0) + \sum_{i=1}^{n}\int_{0}^{t}\frac{n^{-1/2}dM_i(u)}{s^{(0)}(u;\alpha_0)} + o_p(1).
$$

From Taylor expansion,

$$
\begin{aligned}
&\sqrt{n}(\tilde{\eta}_n(t;\tilde{X}) - \eta_0(t;\tilde{X})) \\
&= -\sqrt{n}\left[\exp\{-\tilde{\Lambda}_n(t)\exp(\tilde{\alpha}_n'\tilde{X})\} - \exp\{-\Lambda_0(t)\exp(\alpha_0'\tilde{X})\}\right] \\
&= -\sqrt{n}\Big[-\exp\{-\Lambda_0(t)\exp(\alpha_0'\tilde{X})\}\exp(\alpha_0'\tilde{X})(\tilde{\alpha}_n - \alpha_0)'\tilde{X} \\
&\qquad -\exp\{-\Lambda_0(t)\exp(\alpha_0'\tilde{X})\}\exp(\alpha_0'\tilde{X})[\tilde{\Lambda}_n(t) - \Lambda_0(t)] \\
&\qquad +o(|\tilde{\alpha}_n - \alpha_0|) + o(\|\tilde{\Lambda}_n - \Lambda_0\|)\Big].
\end{aligned}
$$

□

Now we want to derive the asymptotic properties for $\tilde{\phi}_n$ from the complete case analysis.

Lemma IV.4. *Under Conditions IV.1(a), IV.2, IV.3 and IV.5, the estimators* $(\tilde{\theta}_n, \tilde{\phi}_n)$ *that maximize* $\mathbb{P}_n l_0(\theta, \phi; Y, X, \Delta, V)$*, where* $l_0(\theta, \phi; Y, X, \Delta, V)$ *is defined in (4.8), converge in outer probability to* (θ_0, ϕ_0).

Proof. From Corollary 3.2.3 in van der Vaart and Wellner (1996), we need to show that (i)$El_0(\theta_0, \phi_0; Y, X, \Delta, V) > \sup_{(\theta,\phi)\notin G} El_0(\theta, \phi; Y, X, \Delta, V)$ for any open set G that contains (θ_0, ϕ_0); (ii) $\sup_{(\theta,\phi)} \|(\mathbb{P}_n - P)l_0(\theta, \phi; Y, X, \Delta, V)\| \to 0$. Condition (i) is satisfied from Condition IV.1(a) and non-informative censoring assumption. Condition (ii) is satisfied if the class of functions $\{-\frac{1}{2}\Delta(Y - X\beta - \gamma e^{-(S1-t-\mu 1)^2\xi})'\Sigma(\phi)^{-1}(Y - X\beta - \gamma e^{-(S1-t-\mu 1)^2\xi}) - \frac{1}{2}\log|\Sigma(\phi)| : \theta \in \Theta, \phi \in \Phi\}$ belongs to Glivenko-Cantelli. Under Conditions IV.2, IV.3 and IV.5, we have $e^{-(S1-t-\mu 1)^2\xi}$ is Lipschitz function for ξ and μ, $\log|\Sigma(\phi)|$ and all the elements in $\Sigma(\phi)$ are Lipschitz functions for ϕ, thus all belong to Donsker by Theorem 2.10.6 of van der Vaart and Wellner (1996); hence belong to Glivenko-Cantelli. □

Denote the element-wise product of two matrices A and B by $A * B$. Let

$$A_j(\phi) = \frac{\partial \Sigma(\phi)}{\partial \phi_j}, \quad A_{jk}(\phi) = \frac{\partial^2 \Sigma(\phi)}{\partial \phi_j \phi_k}, \quad r(\theta; V, Y, X) = Y - X\beta - \gamma e^{-(V1-t-\mu 1)^2\xi},$$

we obtain the influence function for $\tilde{\phi}_n$ as follows:

Lemma IV.5. *Under Conditions IV.1(a), and IV.2 -IV.5,*

$$\sqrt{n}(\tilde{\phi}_n - \phi_0) = D(\phi_0)^{-1}\sqrt{n}\mathbb{P}_n C(\theta_0, \phi_0; Y, X, \Delta, V),$$

where

$$D(\phi_0) = -\frac{1}{2}P\Big\{\Delta D_1(\phi_0; X)' D_1(\phi_0; X)\Big\}$$

with

$$D_1(\phi_0; X)' = \begin{pmatrix} vec\left(\Sigma(\phi_0)^{-1/2} A_1(\phi_0) \Sigma(\phi_0)^{-1/2}\right)' \\ \vdots \\ vec\left(\Sigma(\phi_0)^{-1/2} A_q(\phi_0) \Sigma(\phi_0)^{-1/2}\right)' \end{pmatrix},$$

and

$$C(\theta_0, \phi_0; Y, X, \Delta, V) = (C_1(\theta_0, \phi_0; Y, X, \Delta, V), \cdots, C_q(\theta_0, \phi_0; Y, X, \Delta, V))'$$

with

$$\begin{aligned} C_j(\theta_0, \phi_0; Y, X, \Delta, V) = & -\frac{1}{2}\Delta tr\left[\Sigma(\phi_0)^{-1} A_j(\phi_0)\right] \\ & +\frac{1}{2}\Delta r(\theta_0; V, Y, X)'\Sigma(\phi_0)^{-1} A_j(\phi_0)\Sigma(\phi_0)^{-1} r(\theta_0; V, Y, X). \end{aligned} \quad (4.12)$$

Proof. The proof follows Lemma IV.1 with $\psi = (\theta, \phi)$. Here

$$m(\theta, \phi; Y, X, \Delta, V) = l_0(\theta, \phi; Y, X, \Delta, V).$$

The first order derivative of $m(\theta, \phi; Y, X, \Delta, V)$ equals

$$\dot{m}(\theta, \phi; Y, X, \Delta, V) = \begin{pmatrix} \dot{m}_1(\theta, \phi; Y, X, \Delta, V) \\ \dot{m}_2(\theta, \phi; Y, X, \Delta, V) \end{pmatrix},$$

where

$$\dot{m}_1(\theta, \phi; Y, X, \Delta, V) = \frac{\partial m(\theta, \phi; Y, X, \Delta, V)}{\partial \theta} = \Delta D_2(\theta; V, X)'\Sigma(\phi)^{-1} r(\theta; V, Y, X)$$

with

$$D_2(\theta; V, X) = (D_{21}(\theta; V, X), D_{22}(\theta; V, X), D_{23}(\theta; V, X), D_{24}(\theta; V, X)), \quad (4.13)$$

where

$$\begin{aligned}
D_{21}(\theta; V, X) &= X \\
D_{22}(\theta; V, X) &= 2\gamma\xi(V1 - t - \mu 1) * e^{-(V1-t-\mu 1)^2\xi} \\
D_{23}(\theta; V, X) &= e^{-(V1-t-\mu 1)^2\xi} \\
D_{24}(\theta; V, X) &= -\gamma(V1 - t - \mu 1)^2 * e^{-(V1-t-\mu 1)^2\xi}
\end{aligned}$$

and

$$\begin{aligned}
\dot{m}_2(\theta, \phi; Y, X, \Delta, V) &= \frac{\partial m(\theta, \phi; Y, X, \Delta, V)}{\partial \phi} \\
&= (C_1(\theta, \phi; Y, X, \Delta, V), \cdots, C_q(\theta, \phi; Y, X, \Delta, V))'
\end{aligned}$$

with $C_j(\theta, \phi; Y, X, \Delta, V)$ defined in (4.12).

The second order derivative of $m(\theta, \phi; Y, X, \Delta, V)$ equals

$$\ddot{m}(\theta, \phi; Y, X, \Delta, V) = \begin{pmatrix} \ddot{m}_{11}(\theta, \phi; Y, X, \Delta, V) & \ddot{m}_{21}(\theta, \phi; Y, X, \Delta, V)' \\ \ddot{m}_{21}(\theta, \phi; Y, X, \Delta, V) & \ddot{m}_{22}(\theta, \phi; Y, X, \Delta, V) \end{pmatrix},$$

where

$$\begin{aligned}
\ddot{m}_{11}(\theta, \phi; Y, X, \Delta, V) &= \frac{\partial^2 m(\theta, \phi; Y, X, \Delta, V)}{\partial \theta \partial \theta'} \\
&= -\Delta D_2(\theta; V, X)' \Sigma^{-1}(\phi) D_2(\theta; V, X) + \Delta D_3(\theta, \phi; V, Y, X)
\end{aligned}$$

with

$$D_3(\theta,\phi;V,Y,X) \tag{4.14}$$
$$= \begin{pmatrix} 0_{p\times p} & 0_{p\times 1} & 0_{p\times 1} & 0_{p\times 1} \\ 0_{1\times p} & D_{311}(\theta,\phi;V,Y,X) & D_{312}(\theta,\phi;V,Y,X) & D_{313}(\theta,\phi;V,Y,X) \\ 0_{1\times p} & D_{312}(\theta,\phi;V,Y,X) & D_{322}(\theta,\phi;V,Y,X) & D_{323}(\theta,\phi;V,Y,X) \\ 0_{1\times p} & D_{313}(\theta,\phi;V,Y,X) & D_{323}(\theta,\phi;V,Y,X) & D_{333}(\theta,\phi;V,Y,X) \end{pmatrix},$$

$$D_{311}(\theta,\phi;V,Y,X) = 2\gamma\xi\left(\left[2\xi(V1-t-\mu 1)^2-1\right]*e^{-(V1-t-\mu 1)^2\xi}\right)' \Sigma(\phi)^{-1}r(\theta;V,Y,X),$$

$$D_{312}(\theta,\phi;V,Y,X) = 2\xi\left((V1-t-\mu 1)*e^{-(V1-t-\mu 1)^2\xi}\right)'\Sigma(\phi)^{-1}r(\theta;V,Y,X),$$

$$D_{313}(\theta,\phi;V,Y,X) = 2\gamma\left((V1-t-\mu 1)*\left[1-\xi(V1-t-\mu 1)^2\right]*e^{-(V1-t-\mu 1)^2\xi}\right)' \Sigma(\phi)^{-1}r(\theta;V,Y,X),$$

$$D_{322}(\theta,\phi;V,Y,X) = 0,$$

$$D_{323}(\theta,\phi;V,Y,X) = -((V1-t-\mu 1)^2*e^{-(V1-t-\mu 1)^2\xi})'\Sigma(\phi)^{-1}r(\theta;V,Y,X),$$

$$D_{333}(\theta,\phi;V,Y,X) = ((V1-t-\mu 1)^4*e^{-(V1-t-\mu 1)^2\xi})'\Sigma(\phi)^{-1}r(\theta;V,Y,X);$$

$$\ddot{m}_{21}(\theta,\phi;Y,X,\Delta,V) = \frac{\partial^2 m(\theta,\phi;Y,X,\Delta,V)}{\partial\phi\partial\theta'}$$
$$= (\ddot{m}_{211}(\theta,\phi;Y,X,\Delta,V),\cdots,\ddot{m}_{21q}(\theta,\phi;Y,X,\Delta,V))'$$

with

$$\ddot{m}_{21j}(\theta,\phi;Y,X,\Delta,V) = -\Delta D_2(\theta;V,X)'\Sigma(\phi)^{-1}A_j(\phi)\Sigma(\phi)^{-1}r(\theta;V,Y,X),$$

and

$$\ddot{m}_{22}(\theta,\phi;Y,X,\Delta,V) = \frac{\partial^2 m(\theta,\phi;Y,X,\Delta,V)}{\partial\phi\partial\phi'}$$
$$= \begin{pmatrix} \ddot{m}_{2211}(\theta,\phi;Y,X,\Delta,V) & \cdots & \ddot{m}_{221q}(\theta,\phi;Y,X,\Delta,V) \\ \vdots & \vdots & \vdots \\ \ddot{m}_{22q1}(\theta,\phi;Y,X,\Delta,V) & \cdots & \ddot{m}_{22qq}(\theta,\phi;Y,X,\Delta,V) \end{pmatrix}$$

with

$$\ddot{m}_{22jk}(\theta,\phi;Y,X,\Delta,V) = -\frac{1}{2}\Delta tr\left[-\Sigma(\phi)^{-1}A_j(\phi)\Sigma(\phi)^{-1}A_k(\phi) + \Sigma(\phi)^{-1}A_{jk}(\phi)\right]$$
$$-\frac{1}{2}\Delta r(\theta;V,Y,X)'\Sigma(\phi)^{-1}\left\{A_j(\phi)\Sigma(\phi)^{-1}A_k(\phi) - A_{jk}(\phi) + A_k(\phi)\Sigma(\phi)^{-1}A_j(\phi)\right\}$$
$$\Sigma(\phi)^{-1}r(\theta;V,Y,X).$$

A1 holds from Lemma IV.4. A2 holds since

$$\int_{-\infty}^{\infty} f_{\theta_0,\phi_0}(y|u,x)r(\theta_0;u,y,x)dy = 0, \tag{4.15}$$
$$\int_{-\infty}^{\infty} f_{\theta_0,\phi_0}(y|u,x)r(\theta_0;u,y,x)r(\theta_0;u,y,x)'dy = \Sigma(\phi_0). \tag{4.16}$$

We have

$$P\ddot{m}(\theta_0,\phi_0;Y,X,\Delta,V) = \begin{pmatrix} D_4(\theta_0,\phi_0) & 0 \\ 0 & D(\phi_0) \end{pmatrix},$$

where

$$D_4(\theta_0, \phi_0) = -P\{\Delta D_2(\theta_0; V, X)'\Sigma(\phi_0)^{-1}D_2(\theta_0; V, X)\},$$

$$D(\phi_0) = \begin{pmatrix} D_{11}(\phi_0) & \cdots & D_{1q}(\phi_0) \\ \vdots & \vdots & \vdots \\ D_{q1}(\phi_0) & \cdots & D_{qq}(\phi_0) \end{pmatrix}$$

with

$$D_{jk}(\phi_0) = -\frac{1}{2}P\Big\{\Delta tr\left[\Sigma(\phi_0)^{-1}A_k(\phi_0)\Sigma(\phi_0)^{-1}A_j(\phi_0)\right]\Big\}$$
$$= -\frac{1}{2}P\Big\{\Delta tr\left[\Sigma(\phi_0)^{-1/2}A_k(\phi_0)\Sigma(\phi_0)^{-1}A_j(\phi_0)\Sigma(\phi_0)^{-1/2}\right]\Big\}.$$

Hence,

$$D(\phi_0) = -\frac{1}{2}P\Big\{\Delta D_1(\phi_0; X)'D_1(\phi_0; X)\Big\}.$$

We have $P\ddot{m}(\theta_0, \phi_0; Y, X, \Delta, V)$ is negative definite from Condition IV.1(a).

From (4.15), we have Condition A3 holds. A4 holds automatically. A5 holds if the class of functions $\big\{-\frac{1}{2}\Delta tr\left[\Sigma(\phi)^{-1}A_j(\phi)\right]+\frac{1}{2}\Delta r(\theta; V, Y, X)'\Sigma(\phi)^{-1}A_j(\phi)\Sigma(\phi)^{-1}r(\theta; V, Y, X) : j = 1, \cdots, q, |\theta - \theta_0| < \delta, |\phi - \phi_0| < \delta\big\}$ is Donsker for some $\delta > 0$ and satisfies $E|\dot{m}(\theta, \phi; Y, X, \Delta, V) - \dot{m}(\theta_0, \phi_0; Y, X, \Delta, V)|^2 \to 0$ as $|(\theta, \phi) - (\theta_0, \phi_0)| \leq \delta_n \downarrow 0$. The two conditions hold from Conditions IV.2-IV.5, and Theorem 2.10.6 of van der Vaart and Wellner (1996). A6 holds from Taylor expansion and Conditions IV.2-IV.5. Hence,

$$\sqrt{n}((\tilde{\theta}_n, \tilde{\phi}_n) - (\theta_0, \phi_0))$$
$$= -[P\ddot{m}(\theta_0, \phi_0; Y, X, \Delta, V)]^{-1}\mathbb{P}_n\dot{m}(\theta_0, \phi_0; Y, X, \Delta, V) + o_{p^*}(1).$$

We focus on ϕ and obtain the influence function for $\tilde{\phi}_n$. □

Lemma IV.6. *Under Conditions IV.2, IV.3and IV.5, the class of functions* $\{l(\theta, \phi, \eta(\alpha, \Lambda); Y, X, \Delta, V) : \theta \in \Theta, \phi \in \Phi, \eta \in \mathcal{F}\}$ *defined in (4.9) belongs to Donsker class.*

Proof. In the proof of Lemma IV.4, we have shown that $\{\log f_{\theta,\phi}(Y|u, X) : \theta \in \Theta, \phi \in \Phi\}$ is Donsker; from Condition IV.2 and Theorem 2.10.6 of van der Vaart and Wellner (1996), we have $\{f_{\theta,\phi}(Y|u, X)\}$ is Donsker. From integration by parts,

$$\begin{aligned}
&\int_C^\tau f_{\theta,\phi}(Y|u, X)d[1 - \exp\{-\Lambda(u)\exp(\alpha' X)\}] \\
&= f_{\theta,\phi}(Y|\tau, X)[1 - \exp\{-\Lambda(\tau)\exp(\alpha' X)\}] \\
&\quad -f_{\theta,\phi}(Y|C, X)[1 - \exp\{-\Lambda(C)\exp(\alpha' X)\}] \\
&\quad + \int_C^\tau f_{\theta,\phi}(Y|u, X)(2\gamma\xi(u1 - t - \mu 1) * e^{-(u1-t-\mu 1)^2\xi})' \\
&\qquad\qquad \Sigma(\phi)^{-1}r(\theta)[1 - \exp\{-\Lambda(u)\exp(\alpha' X)\}]du.
\end{aligned}$$

In the above, $\exp\{-\Lambda(u)\exp(\alpha' X)\}$ is Lipschitz for Λ and α from Condition IV.5, and $e^{-(u1-t-\mu 1)^2\xi}$ is Lipschitz function for ξ and μ, thus belong to Donsker classes by Theorem 2.10.6 of van der Vaart and Wellner (1996). By Theorem 2.10.3 of van der Vaart and Wellner (1996), the permanence of the Donsker property for the closure of the convex hull, we have $\{\int_C^\tau f_{\theta,\phi}(Y|u, X)d[1 - \exp\{-\Lambda(u)\exp(\alpha' X)\}] : \theta \in \Theta, \phi \in \Phi, \eta \in \mathcal{F}\}$ is Donsker. By Condition IV.8, $\delta_1 \le \int_C^\tau f_{\theta,\phi}(Y|u, X)d[1 - \exp\{-\Lambda(u)\exp(\alpha' X)\}] \le \sup_{Y,u,X} f_{\theta,\phi}(Y|u, X)$, which is bounded from Condition IV.2. Hence, $\{\log \int_C^\tau f_{\theta,\phi}(Y|u, X)d[1 - \exp\{-\Lambda(u)\exp(\alpha' X)\}] : \theta \in \Theta, \phi \in \Phi, \eta \in \mathcal{F}\}$ is Donsker from Theorem 2.10.6 of van der Vaart and Wellner (1996). □

Lemma IV.7. *Under Conditions IV.1(b), IV.2-IV.3 and IV.8, we have*

$$\sup_{\theta\in\Theta} |Pl(\theta, \tilde{\phi}_n, \tilde{\eta}_n; Y, X, \Delta, V) - Pl(\theta, \phi_0, \eta_0; Y, X, \Delta, V)| = o_p(1).$$

Proof. From triangular inequality,

$$
\begin{aligned}
&\sup_{\theta\in\Theta}|Pl(\theta,\tilde{\phi}_n,\tilde{\eta}_n;Y,X,\Delta,V)-Pl(\theta,\phi_0,\eta_0;Y,X,\Delta,V)|\\
&\leq \frac{1}{2}|P\{\log|\Sigma(\tilde{\phi}_n)|-\log|\Sigma(\phi_0)|\\
&\quad+\frac{1}{2}\left|P\left\{r(\theta_0;V,Y,X)'\left[\Sigma(\tilde{\phi}_n)^{-1}-\Sigma(\phi_0)^{-1}\right]r(\theta_0;V,Y,X)\right\}\right|\\
&\quad+\sup_{\theta\in\Theta}\frac{1}{2}\left|P\left\{d(\theta;V,X)'\left[\Sigma(\tilde{\phi}_n)^{-1}-\Sigma(\phi_0)^{-1}\right]d(\theta;V,X)\right\}\right|\\
&\quad+\sup_{\theta\in\Theta}\left|P\left\{\log\int_C^\tau f_{\theta,\tilde{\phi}_n}(Y|u,X)d\tilde{\eta}_n(u;X)-\log\int_C^\tau f_{\theta,\phi_0}(Y|u,X)d\tilde{\eta}_n(u;X)\right\}\right|\\
&\quad+\sup_{\theta\in\Theta}\left|P\left\{\log\int_C^\tau f_{\theta,\phi_0}(Y|u,X)d\tilde{\eta}_n(u;X)-\log\int_C^\tau f_{\theta,\phi_0}(Y|u,X)d\eta_0(u;X)\right\}\right|\\
&\leq O(|\tilde{\phi}_n-\phi_0|)\\
&\quad+\sup_{\theta\in\Theta}\left|P\left\{\log\int_C^\tau f_{\theta,\phi_0}(Y|u,X)d\tilde{\eta}_n(u;X)-\log\int_C^\tau f_{\theta,\phi_0}(Y|u,X)d\eta_0(u;X)\right\}\right|,
\end{aligned}
$$

where $d(\theta;V,X)=X(\beta-\beta_0)+\left[\gamma e^{-(V1-t-\mu 1)^2\xi}-\gamma_0 e^{-(V1-t-\mu_0 1)^2\xi_0}\right]$, which is uniformly bounded for any (V,X) and $\theta\in\Theta$. The last inequality is obtained from the mean value theorem and Conditions IV.2-IV.3 and IV.8. Again, from the mean value theorem and Condition IV.8,

$$
\begin{aligned}
&\sup_{\theta\in\Theta}\left|P\left\{\log\int_C^\tau f_{\theta,\phi_0}(Y|u,X)d\tilde{\eta}_n(u;X)-\log\int_C^\tau f_{\theta,\phi_0}(Y|u,X)d\eta_0(u;X)\right|\right\}\\
&\leq \delta_1\sup_{\theta\in\Theta}P\left|\left\{\int_C^\tau f_{\theta,\phi_0}(Y|u,X)d[\tilde{\eta}_n(u;X)-\eta_0(u;X)]\right\}\right|.
\end{aligned}
$$

By integration by parts,

$$\sup_{\theta\in\Theta} P\left|\int_C^\tau f_{\theta,\phi_0}(Y|u,X)d\tilde{\eta}_n(u;X) - \int_C^\tau f_{\theta,\phi_0}(Y|u,X)d\eta_0(u;X)\right|$$
$$\leq \sup_{u\in[0,\tau]}\left(2+\tau P\left\{\left(2\gamma\xi|u1-t-\mu 1| * e^{-(u1-t-\mu 1)^2\xi}\right)'\Sigma(\phi_0)^{-1}\left|r(\theta;u,Y,X)\right|\right\}\right)$$
$$\times \sup_{\theta\in\Theta,Y\in\mathcal{Y},X\in\mathcal{X},u\in[0,\tau]} f_{\theta,\phi_0}(Y|u,X)\times\|\tilde{\eta}_n-\eta_0\| = O(\|\tilde{\eta}_n-\eta_0\|).$$

The last equality holds because all the elements in $\int_{-\infty}^{\infty}|y|f_{\theta_0,\phi_0}(y|u,x)dy$ are bounded uniformly for all $u\in[0,\tau]$ and $X\in\mathcal{X}$ from Kamart (1953). Hence,

$$\sup_{\theta\in\Theta}|Pl(\theta,\tilde{\phi}_n,\tilde{\eta}_n;Y,X,\Delta,V) - Pl(\theta,\phi_0,\eta_0;Y,X,\Delta,V)|$$
$$\leq O(|\tilde{\phi}_n-\phi_0|) + O(\|\tilde{\eta}_n-\eta_0\|) = o_p(1).$$

□

4.6.3 Proofs of Theorem IV.1 and IV.2

4.6.3.1 Proof of Theorem IV.1

Proof. From Condition IV.1, we have

$$\sup_{d(\theta,\theta_0)>\delta} Pl(\theta,\phi_0,\eta_0;Y,X,\Delta,V) < Pl(\theta_0,\phi_0,\eta_0;Y,X,\Delta,V) \tag{4.17}$$

holds for every $\delta>0$. By the definition of $\hat{\theta}_n$, we have

$$\begin{aligned}\mathbb{P}_n l(\hat{\theta}_n,\tilde{\phi}_n,\tilde{\eta}_n;Y,X,\Delta,V) &\geq \mathbb{P}_n l(\theta_0,\tilde{\phi}_n,\tilde{\eta}_n;Y,X,\Delta,V) \\ &= \mathbb{P}_n l(\theta_0,\phi_0,\eta_0;Y,X,\Delta,V)+o_p(1),\end{aligned} \tag{4.18}$$

where the equality is obtained by Lemma IV.6 and Lemma IV.7. The class of functions $\{l(\theta,\phi,\eta;Y,X,\Delta,V):\theta\in\Theta,\phi\in\Phi,\eta\in\mathcal{F}\}$ is Donsker from Lemma IV.6, hence is

Glivenko-Cantelli, we have

$$
\begin{aligned}
0 &\leq Pl(\theta_0, \phi_0, \eta_0; Y, X, \Delta, V) - Pl(\hat{\theta}_n, \phi_0, \eta_0; Y, X, \Delta, V) \\
&= \mathbb{P}_n l(\theta_0, \phi_0, \eta_0; Y, X, \Delta, V) - \mathbb{P}_n l(\hat{\theta}_n, \phi_0, \eta_0; Y, X, \Delta, V) + o_p(1) \\
&\leq \mathbb{P}_n l(\hat{\theta}_n, \tilde{\phi}_n, \tilde{\eta}_n; Y, X, \Delta, V) - \mathbb{P}_n l(\hat{\theta}_n, \phi_0, \eta_0; Y, X, \Delta, V) + o_p(1) \quad (4.19) \\
&= Pl(\hat{\theta}_n, \tilde{\phi}_n, \tilde{\eta}_n; Y, X, \Delta, V) - Pl(\hat{\theta}_n, \phi_0, \eta_0; Y, X, \Delta, V) + o_p(1) \\
&= o_p(1). \quad (4.20)
\end{aligned}
$$

where (4.19) is obtained from (4.18), and (4.20) is obtained by Lemma IV.7. By inequality (4.17), for every $\delta > 0$ we have

$$\{d(\hat{\theta}_n, \theta_0) \geq \delta\} \subset \{Pl(\hat{\theta}_n, \phi_0, \eta_0; Y, X, \Delta, V) < Pl(\theta_0, \phi_0, \eta_0; Y, X, \Delta, V)\}$$

with the sequence of the events on the right going to a null event in view of inequality (4.20), which yields the almost sure (thus in probability) convergence of $\hat{\theta}_n$. This argument is taken from the proof of Theorem 5.8 in van der vaart (2002) and the proof of Theorem 3 in Li and Nan (2011). □

4.6.3.2 Proof of Theorem IV.2

Proof. The proof follows Lemma IV.2. Here

$$m(\theta, \phi, \eta; Y, X, \Delta, V) = l(\theta, \phi, \eta; Y, X, \Delta, V).$$

Note that the function $m(\theta, \phi, \eta; Y, X, \Delta, V)$ is different from the function $m(\theta, \phi; Y, X, \Delta, V)$ in Lemma IV.5.

The partial derivative of $m(\theta,\phi,\eta;Y,X,\Delta,V)$ with respect to θ equals

$$
\begin{aligned}
&\dot{m}_1(\theta,\phi,\eta;Y,X,\Delta,V)\\
&=\Delta D_2(\theta;V,X)'\Sigma(\phi)^{-1}r(\theta;V,Y,X)+(1-\Delta)\Big[\int_C^\tau f_{\theta,\phi}(Y|u,X)d\eta(u;X)\Big]^{-1}\\
&\quad\Big[\int_C^\tau f_{\theta,\phi}(Y|u,X)D_2(\theta;u,X)'\Sigma(\phi)^{-1}r(\theta;u,Y,X)d\eta(u;X)\Big],
\end{aligned}
$$

where $D_2(\theta;u,X)$ is defined in (4.13).

The second order derivative of $m(\theta,\phi,\eta;Y,X,\Delta,V)$ with respect to θ equals

$$
\begin{aligned}
&\ddot{m}_{11}(\theta,\phi,\eta;Y,X,\Delta,V)\\
&=-\Delta D_2(\theta;V,X)'\Sigma(\phi)^{-1}D_2(\theta;V,X)+\Delta D_3(\theta,\phi;V,Y,X)+(1-\Delta)\\
&\times\Big\{\Big[\int_C^\tau f_{\theta,\phi}(Y|u,X)\Big\{-D_2(\theta;u,X)'\Sigma(\phi)^{-1}D_2(\theta;u,X)+D_3(\theta,\phi;u,Y,X)\\
&+\big[D_2(\theta;u,X)'\Sigma(\phi)^{-1}r(\theta;u,Y,X)\big]^{\otimes 2}\Big\}d\eta(u;X)\Big]\Big[\int_C^\tau f_{\theta,\phi}(Y|u,X)d\eta(u;X)\Big]^{-1}\\
&-\Big[\int_C^\tau f_{\theta,\phi}(Y|u,X)D_2(\theta;u,X)'\Sigma(\phi)^{-1}r(\theta;u,Y,X)d\eta(u;X)\Big]^{\otimes 2}\\
&\qquad\qquad\Big[\int_C^\tau f_{\theta,\phi}(Y|u,X)d\eta(u;X)\Big]^{-2}\Big\},
\end{aligned}
$$

where $D_3(\theta,\phi;V,Y,X)$ is defined in (4.14).

B1 holds from Lemma IV.3, Lemma IV.4 and Theorem IV.1. From (4.15) and

(4.16),

$$
\begin{aligned}
&P\ddot{m}_{11}(\theta_0, \phi_0, \eta_0; Y, X, \Delta, V) \qquad (4.21)\\
&= -P\Bigg\{\Delta D_2(\theta_0; V, X)'\Sigma^{-1}(\phi_0)D_2(\theta_0; V, X)\\
&\quad +(1-\Delta)\Bigg[\int_C^\tau f_{\theta_0,\phi_0}(Y|u, X)d\eta_0(u; X)\Bigg]^{-2}\\
&\quad \Bigg[\int_C^\tau f_{\theta_0,\phi_0}(Y|u, X)D_2(\theta_0; u, X)'\Sigma(\phi_0)^{-1}r(\theta_0; u, Y, X)d\eta_0(u; X)\Bigg]^{\otimes 2}\Bigg\},
\end{aligned}
$$

which is negative definite from Condition IV.1(a), thus B2 holds. From (4.15), we have B3 holds. And B4 holds automatically.

Since

$$\frac{A_1}{B_1} - \frac{A_2}{B_2} = \frac{A_1(B_2 - B_1)}{B_1 B_2} + \frac{A_1 - A_2}{B_2},$$

under Conditions IV.2-IV.5 and IV.8, we have

$$E|\dot{m}_1(\theta, \phi, \eta; Y, X, \Delta, V) - \dot{m}_1(\theta_0, \phi_0, \eta_0; Y, X, \Delta, V)|^2 \to 0$$

as $|(\theta, \phi) - (\theta_0, \phi_0)| \leq \delta_n \downarrow 0$, since the denominator of the censored subject part is bounded away from zero by Condition IV.8 and the numerator goes to zero from continuity. Similar to the proof of Lemma IV.6, we have the class of functions $\{\int_C^\tau f_{\theta,\phi}(Y|u, X)D_2(\theta; u, X)'\Sigma(\phi)^{-1}r(\theta; u, Y, X)d\eta(u; X) : \theta \in \Theta, \phi \in \Phi, \eta \in \mathcal{F}\}$ belongs to Donsker. Hence, $\{\dot{m}_1(\theta, \phi, \eta; Y, X, \Delta, V) : \theta \in \Theta, \phi \in \Phi\}$ is Donsker from Section 2.10.2 of van der Vaart and Wellner (1996) and Condition IV.8. Furthermore, from Corollary 2.3.12 of van der Vaart and Wellner (1996), we have B5 holds. Under Conditions IV.2-IV.5 and IV.8, similar to the proof of Theorem III.1 in Chapter III, we can show that B6 holds. Particularly in B6,

$$
\begin{aligned}
&P\ddot{m}_{12}(\theta_0, \phi_0, \eta_0; Y, X, \Delta, V) \\
&= (P\ddot{m}_{121}(\theta_0, \phi_0, \eta_0; Y, X, \Delta, V), \cdots, P\ddot{m}_{12q}(\theta_0, \phi_0, \eta_0; Y, X, \Delta, V))
\end{aligned}
$$

with

$$
\begin{aligned}
&P\ddot{m}_{12j}(\theta_0, \phi_0, \eta_0; Y, X, \Delta, V) \\
&= -\frac{1}{2}P\Bigg((1-\Delta)\Bigg[\int_C^\tau f_{\theta_0,\phi_0}(Y|u,X)D_2(\theta_0; u, X)'\Sigma(\phi_0)^{-1}r(\theta_0; u, Y, X)d\eta_0(u; X)\Bigg] \\
&\quad \Bigg[\int_C^\tau f_{\theta_0,\phi_0}(Y|u,X)\Bigg\{r(\theta_0; u, Y, X)'\Sigma(\phi_0)^{-1}A_j(\phi_0)\Sigma(\phi_0)^{-1}r(\theta_0; u, Y, X) \\
&\qquad -tr\Bigg[\Sigma(\phi_0)^{-1}A_j(\phi_0)\Bigg]\Bigg\}d\eta_0(u; X)\Bigg]\Bigg[\int_C^\tau f_{\theta_0,\phi_0}(Y|u,X)d\eta_0(u; X)\Bigg]^{-2}\Bigg),
\end{aligned}
$$

and

$$
\begin{aligned}
&P\ddot{m}_{13}(\theta_0, \phi_0, \eta_0; Y, X, \Delta, V)[\tilde{\eta}_n - \eta_0] \\
&= -P\Bigg((1-\Delta)\Bigg[\int_C^\tau f_{\theta_0,\phi_0}(Y|u,X)D_2(\theta_0; u, X)'\Sigma(\phi_0)^{-1}r(\theta_0; u, Y, X)d\eta_0(u; X)\Bigg] \\
&\quad \Bigg[\int_C^\tau f_{\theta_0,\phi_0}(Y|u,X)d\{\tilde{\eta}_n(u; X) - \eta_0(u; X)\}\Bigg]\Bigg[\int_C^\tau f_{\theta_0,\phi_0}(Y|u,X)d\eta_0(u; X)\Bigg]^{-2}\Bigg) \\
&= -P\Bigg((1-\Delta)\Bigg[\int_C^\tau f_{\theta_0,\phi_0}(Y|u,X)D_2(\theta_0; u, X)'\Sigma(\phi_0)^{-1}r(\theta_0; u, Y, X)d\eta_0(u; X)\Bigg] \\
&\quad \Bigg[f_{\theta_0,\phi_0}(Y|\tau,X)\{\tilde{\eta}_n(\tau; X) - \eta_0(\tau; X)\} - f_{\theta_0,\phi_0}(Y|C,X)\{\tilde{\eta}_n(C; X) - \eta_0(C; X)\} \\
&\quad + \int_C^\tau f_{\theta_0,\phi_0}(Y|u,X)(2\gamma_0\xi_0(u1 - t - \mu_0 1) * e^{-(u1-t-\mu_0 1)^2\xi_0})'\Sigma(\phi_0)^{-1}r(\theta_0; u, Y, X) \\
&\qquad \{\tilde{\eta}_n(u; X) - \eta_0(u; X)\}du\Bigg]\Bigg[\int_C^\tau f_{\theta_0,\phi_0}(Y|u,X)d\eta_0(u; X)\Bigg]^{-2}\Bigg)
\end{aligned}
$$

$$= -\mathbb{G}_n\{G(\theta_0, \phi_0, \eta_0; \tilde{X}, \tilde{\Delta}, \tilde{V})\} + o_p(1),$$

where

$$\begin{aligned}
&G(\theta_0, \phi_0, \eta_0; \tilde{X}, \tilde{\Delta}, \tilde{V}) \\
&= P\Big\{E_1(\theta_0, \phi_0, \eta_0; Y, X, \Delta, V)E_2(\theta_0, \phi_0, \eta_0; Y, X, \tau)A_1(\eta_0; \tau, X; \tilde{X}, \tilde{\Delta}, \tilde{V})\Big\} \\
&\quad -P\Big\{E_1(\theta_0, \phi_0, \eta_0; Y, X, \Delta, V)E_2(\theta_0, \phi_0, \eta_0; Y, X, C)A_1(\eta_0; C, X; \tilde{X}, \tilde{\Delta}, \tilde{V})\Big\} \\
&\quad +P\Big\{\int_C^\tau E_1(\theta_0, \phi_0, \eta_0; Y, X, \Delta, V)E_2(\theta_0, \phi_0, \eta_0; Y, X, u) \\
&\qquad\qquad E_3(\theta_0, \phi_0; Y, X, u)A_1(\eta_0; u, X; \tilde{X}, \tilde{\Delta}, \tilde{V})du\Big\}
\end{aligned}$$

with

$$\begin{aligned}
&E_1(\theta_0, \phi_0, \eta_0; Y, X, \Delta, V) = (1-\Delta)\Big[\int_C^\tau f_{\theta_0,\phi_0}(Y|u, X)d\eta_0(u; X)\Big]^{-2}, \\
&\quad \Big[\int_C^\tau f_{\theta_0,\phi_0}(Y|u, X)D_2(\theta_0; u, X)'\Sigma(\phi_0)^{-1}r(\theta_0; u, Y, X)d\eta_0(u; X)\Big], \\
&E_2(\theta_0, \phi_0, \eta_0; Y, X, u) = f_{\theta_0,\phi_0}(Y|u, X)[1-\eta_0(u; X)]\exp(\alpha_0'X), \\
&E_3(\theta_0, \phi_0; Y, X, u) = (2\gamma_0\xi_0(u1 - t - \mu_0 1) * e^{-(u1-t-\mu_0 1)^2\xi_0})'\Sigma(\phi_0)^{-1}r(\theta_0; u, Y, X),
\end{aligned}$$

and $A_1(\eta_0; u, X; \tilde{X}, \tilde{\Delta}, \tilde{V})$ is defined in Lemma IV.3.

Hence by Lemma IV.2 and the central limit theorem,

$$\begin{aligned}
&\sqrt{n}(\hat{\theta}_n - \theta_0) \\
&= A^{-1}\sqrt{n}\mathbb{P}_n\dot{m}_1(\theta_0, \phi_0, \eta_0; \mathbf{X}) + A^{-1}\sqrt{n}P\{\ddot{m}_{12}(\theta_0, \phi_0, \eta_0; \mathbf{X})\}(\tilde{\phi}_n - \phi_0) \\
&\quad +A^{-1}\sqrt{n}P\{\ddot{m}_{13}(\theta_0, \phi_0, \eta_0; \mathbf{X})[\tilde{\eta}_n - \eta_0]\} + o_p(1),
\end{aligned}$$

which converges weakly to a mean zero normal random variable with variance

$A^{-1}BA^{-1}$, where

$$
\begin{aligned}
A &= -P\{\ddot{m}_{11}(\theta_0, \phi_0, \eta_0; Y, X, \Delta, V)\}, \\
B &= P\Bigg[\dot{m}_1(\theta_0, \phi_0, \eta_0; \tilde{Y}, \tilde{X}, \tilde{\Delta}, \tilde{V}) + G(\theta_0, \phi_0, \eta_0; \tilde{X}, \tilde{\Delta}, \tilde{V}) \\
&\quad + P\left\{\ddot{m}_{12}(\theta_0, \phi_0, \eta_0; Y, X, \Delta, V)\right\} D(\phi_0)^{-1} C(\theta_0, \phi_0; \tilde{Y}, \tilde{X}, \tilde{\Delta}, \tilde{V})\Bigg]^{\otimes 2}
\end{aligned}
$$

with $D(\phi_0)$ and $C(\theta_0, \phi_0; \tilde{Y}, \tilde{X}, \tilde{\Delta}, \tilde{V})$ defined in Lemma IV.5. □

CHAPTER V

Future Work

In this dissertation, we derived the oracle inequalities for the high-dimensional Cox regression model via Lasso in Chapter II, and developed novel approaches to address censored covariate issues in Chapter III and Chapter IV.

The semiparametric likelihood-base estimates proposed in Chapter III allows more efficient and robust estimates of the regression parameters when a covariate is subject to limit of detection. This is an important issue, especially when the covariate subject to limit of detection is significantly associated with the response variable. The amount of efficiency gain of the proposed two-stage method depends on how far we can estimate $F(t|X)$ reasonably well beyond the limit of detection. We truncate the residuals with some finite value τ in this article. In practice, the upper limit of the integral in the pseudo-likelihood function can go as far as the largest observed residual in the fitted accelerated failure time model $\max_i(T_i - X_i'\hat{\alpha}_n)$, and theoretically, this upper limit is ∞ when the support of $X'\alpha_0$ is unbounded. In the latter case, it can be shown that $\hat{F}_n$ converges to F on the entire real line with a polynomial rate, e.g., $n^{-1/8}$, see Lai and Ying (1991) and the 2010 University of Michigan PhD thesis by Y. Ding. We may still be able to obtain consistent estimates for the parameters of interest. The asymptotic normality, however, will largely remain unknown.

The extrapolation of $F(t|X)$ beyond C depends on the semiparametric AFT model. New data with lower limit of detection can be used to check whether the AFT model is valid. For example, Study of Women's Health Across the Nation (SWAN) proposed a study to further investigate the the relationship between anti-Mullerian hormone and time to final menstrual period, where anti-Mullerian hormone is more accurately measured with lower limit of detection. We can predict the mean/medium of the anti-Mullerian hormone below the old limit of detection from the AFT model, see Yin Ding's 2010 University of Michigan Ph.D. thesis, and compare it to the measures in the new data. Furthermore, a AFT model can be fitted to the new data to see if it yields the same results. To evaluate the robustness of the proposed method, a sensitivity analysis can be considered where T is generated from a model that does not satisfy the AFT model assumptions.

We only consider the case with one covariate subject to limit of detection in this article for simplicity. Regression with multiple covariates subject to limits of detection may occur in practice. Parametric models have been considered for such problems (May et al., 2011; D'Angelo and Weissfeld, 2008). To achieve robust results, the proposed semiparametric approach can be generalized to tackle the problem with multiple covariates subject to limits of detection. The critical step is to provide an valid nonparametric estimate for the multivariate survival function, for which available methods include Dabrowska (1988), Prentice and Cai (1992), van der Laan (1996), and Prentice and Moodie (2004). The constant limit of detection assumption considered in this article, though commonly seen in practice, also can be relaxed to cases with random limit of detection.

Limit of detection issue can be viewed as a missing data problem. Multiple imputation (Little and Rubin, 2002) may be considered as an alternative method if the tail distribution of the covariate subject to limit of detection conditional on all other variables, including the response variable, can be estimated reasonably well.

More general regression models in Chapter IV could be considered, for example, the semiparametric mixed effect model with a nonparametric smooth function of terminal event time and the data collection time. Time-dependent covariates are common in longitudinal studies. The major challenge for handling time-dependent covariates is to provide a valid estimate of $F(s|\bar{X}(v))$ beyond censoring time, which needs valid extrapolation of covariate history. The proposed methodology may also apply to recurrent event data with terminal events.

In the medical payment cohort study, if the terminal event time is known, the prediction of medical payment is straightforward given the nonlinear regression model. However, when a new patient participated in the study, the prediction of the future medical payment given that the patient is still alive could be challenging and it is also of future interest.

BIBLIOGRAPHY

BIBLIOGRAPHY

Agresti, A. (2002). Categorical data analysis.*John Wiley*.

Albert, P.S. and Shih, J.H. (2010). An approach for jointly modeling multivariate longitudinal measurements and discrete time-to-event data .*The Annals of Applied Statistics* **4**, 1139-1620.

Andersen, P. K. and Gill, R. D. (1982). Cox's regression model for counting processes: a large sample study. *Ann. Statist.* **10**, 1100-1120.

Bach, F. (2010). Self-concordant analysis for logistic regression. *Electronic Journal of Statistics* **4**, 384–414.

Bickel, P., Ritov, Y., and Tsybakov, A. (2009). Simultaneous analysis of lasso and dantzig selector. *Ann. Statist.* **37**, 1705–1732.

Bradic, J., Fan, J., and Jiang, J. (2011). Regularization for Cox's proportional hazards model with NP-dimensionality. *Ann. Statist.* **39**, 3092–3120.

Breslow, N. and Crowley, J. (19874). A large sample study of the life table and product limit estimates under random censorship.*The Annals of Statistics* **2**, 415-614.

Box, G. E. P. and Cox, D. R. (1964). An analysis of transformations (with discussion). *Journal of the Royal Statistical Society, Series B.* **26**, 211-252.

Bühlmann, P. (2006). Boosting for high-dimensional linear models. *Ann. Statist.* **34**, 559–583.

Bunea, F. (2008). Honest variable selection in linear and logistic regression models via ℓ_1 and $\ell_1 + \ell_2$ penalization. *Electronic Journal of Statistics* **2**, 1153–1194.

Bunea, F., Tsybakov, A. B., and Wegkamp, M. H. (2007). Sparsity oracle inequalities for the Lasso. *Electronic Journal of Statistics* **1**, 169–194.

Cai, T., Tian, L. and Wei, L. J. (2005). Semiparametric Box-Cox power transformation models for censored survival observations. *Biometrika* **92**, 619-632.

Cox, D. R. (1972). Regression models and life tables (with discussion). *J. Roy. Statist. Soc. B* **34**, 187-220.

Crainiceanu, C.M., Guallar, E., Navas-Acien, A. and Tellez-Plaza, M. (2008). Cadmium exposure and hypertension in the 1999-2004 National Health and Nutrition Examination Survey (Nutrition Examination Survey).*Environmental Health Perspectives* **116**, 51-56.

Cole, S.R., Chu, H., Nie, L. and Schisterman, E.F. (2009). Estimating the odds ratio when exposure has a limit of detection.*International Journal of Epidemiology* **38**, 1674-1680.

Dabrowska, D.M. (1988). Kaplan-Meier estimate on the plane.*Annals of Statistics* **16**, 1475-1489.

D'Angelo, G. and Weissfeld, L. (2008). An index approach for the Cox model with left censored covariates.*Statistics in medicine* **72**, 4502-4514.

Ding, Y., and Nan, B. (2011). A sieve M-theorem for bundled parameters in semiparametric models, with application to the efficient estimation in a linear model for censored data. *Annals of Statistics* **39**, 3032-3061.

Ding, J. and Wang, J.L. (2008). Modeling longitudinal data with nonparametric multiplicative random effects jointly with survival data.*Biometrics.* **64**, 546-556.

Fleming, T.R. and Harrington, D.P. (2005). *Counting Processes and Survival Analysis*. Wiley.

Foster, A. M., Tian, L. and Wei, L. J. (2001). Estimation for the Box-Cox transformation model without assuming parametric error distribution. *Journal of the American Statistical Association* **96**, 1097-1101.

Gaiffas, S. and Guilloux, A. (2012). High-dimensional additive hazards models and the lasso. *Electronic Journal of Statistics* **6**, 522–546.

Ghosh, D. and Lin, D.Y. (2002). Marginal regression models for recurrent and terminal events.*Statistica Sinica.* **12**, 663-688.

Gui, J. and Li, H. (2005). Penalized Cox regression analysis in the high-dimensional and low-sample size settings, with applications to microarray gene expression data. *Bioinformatics* **21**, 3001–3008.

Harlow, Sioban D., E.S. Mitchell, S. Crawford, B. Nan, R. Little, and J. Taffe. (2008). The ReSTAGE Collaboration: defining optimal bleeding criteria for onset of early menopausal transition. *Fertility and Sterility* **89**. 129-140.

Helsel, D. R. (2005). Nondetects and Data Analysis: Statistics for Censored Environmental Data. *New York: Wiley* .

Helsel, D. R. (2006). Fabricating data: how substituting values for nondetects can ruin results, and what can be done about it. *Chemosphere* **56**, 2434-2439.

HORNUNG, R.W. AND REED, L.D. (1990). Estimation of average concentration in the presence of nondetectable values. *Applied Occupational and Environmental Hygiene* **4**, 46-51.

HUANG, C.Y. AND WANG, M.C. (2004). Joint modeling and estimation for recurrent event processes and failure time data.*Journal of the American Statistical Association.* **99**, 1153-1165.

JIN, Z. , LIN, D. Y. , WEI, L. J. AND YING, Z. (2003). Rank-based inference for the accelerated failure time model. *Biometrika* **90**, 341-353.

KALBFLEISCH, J. D. AND PRENTICE, R. L. (2002). *The Statistical Analysis of Failure Time Data, 2nd Ed.* John Wiley, New York.

KAMAT, A. R. (1953). Incomplete and Absolute Moments of the Multivariate Normal Distribution with Some Applications.*Biometrika.* **40**, 20-34.

KIM, C., KONG, S., LAUGHLIN, G. A., GOLDEN, S. H., MATHER, K. J., NAN, B., RANDOLPH, J. R., EDELSTEIN, S. L., LABRIE, F., BUSCHUR E., BARRETT-CONNOR, E. [a] (2012). Reductions in glucose among postmenopausal women who use and do not use estrogen therapy. *Menopause* **20**, 4.

LAI, T.L. AND YING, Z. (1991). Large sample theory of a modified Buckley-James estimator for regression analysis with censored data. *Annals of Statistics* **19**, 1370-1402.

Ledoux, M. and Talagrand, M. (1991). *Probability in Banach Spaces: Isoperimetry and Processes.* Springer, Berlin.

LEHMANN, E.L. (1998). *Theory of point estimation.* Springer.

LI, Z.G. AND NAN, B. (2011). Relative risk regression for current status data in case-cohort studies.*The Canadian Journal of Statistics.* **39**, 557-577.

LIN, D. Y., ROBINS, J.M. AND WEI, L.J. (1996). Comparing two failure time distributions in the presence of dependent censoring. *Biometrika* **83**, 381-393.

LIN, D. Y., WEI, L.J. AND YING, Z. (1993). Checking the Cox model with cumulative sums of martingale-based residuals. *Biometrika* **80**, 557-572.

LINDSTROM, M.J. AND BATES, D.M. (1988). Newton-Raphson and EM algorithms for linear mixed-effects models for repeated-measures data.*Journal of the American Statistical Association.* **83**, 1014-1022.

LITTLE, R. J, AND RUBIN, D. B. (2002). Statistical Analysis with Missing Data. *New York: Wiley.*

LIU, L., WOLFE, R.A., KALBFLEISCH, J.D. (2007). A shared random effects model for censored medical costs and mortality.*Statistics in Medicine* . **26**, 139-155.

LU, X., NAN, B., SONG P., SOWERS, M. (2010). Longitudinal data analysis with event time as a covariate. *Stat. Biosci.* **2**, 65-80.

LUBIN, J.H., COLT, J.S., CAMANN, D. (2004). Epidemiologic evaluation of measurement data in the presence of detection limits. *Environ Health Perspect* **112**, 1691-1696.

LYNN, H. (2001). Maximum likelihood inference for left-censored HIV RNA data. *Statistics in medicine* **20**, 33-45.

Martinussen, T. and Scheike, T. H. (2009). Covariate selection for the semiparametric additive risk model. *Scandinavian Journal of Statistics* **36**, 602–619.

Massart, P. (1990). The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality. *Ann. Probab.* **18**, 1269–1283.

MAY, R.C., IBRAHIM, J.G. AND CHU, H. (2011). Maximum likelihood estimation in generalized linear models with multiple covariates subject to detection limits. *Statistics in Medicine* **30**, 2551-2561.

MCCULLAGH, P. AND NELDER, J. (1989). Generalized Linear Models. *Chapman and Hall/CRC.*

MOULTON, L.H., CURRIERO, F.C.AND BARROSO, P.F. (2002). Mixture models for quantitative HIV RNA data. *Stat Meth Med Res.* **11**, 317-325.

NAN, B., KALBFLEISCH, J.D., AND YU, M. (2009). Asymptotic theory for the semiparametric accelerated failure time model with missing data. *Annals of Statistics* **37**, 2351-2376.

NAN, B. AND WELLNER J. A. (2013). A general semiparametric Z-estimation approach for case-cohort studies. *Statistica Sinica* **23**, 1155-1180.

NIE, L, CHU, H, LIU, C, COLE, SR, VEXLER, A, SCHISTERMAN, EF. (2010). Linear regression with an independent variable subject to a detection limit. *Epidemiology* **21**, S17- S24.

PENG, L. AND FINE, J. P. (2006). Rank estimation of accelerated lifetime models with dependent censoring. *Journal of the American Statistical Association* **101**, 1085-1093.

PRENTICE, R.L. AND CAI, J. (1992). Covariance and survivor function estimation usingcensored multivariate time data.*Biometrika* **79**, 495-512.

PRENTICE, R.L. AND MOODIE, F.Z. (2004).Hazard-based nonparametric survivor function estimation.*J.R.Statist. Soc. B.* **66**, 305-319.

RICHARDSON, D.B. AND CIAMPI, A. (2003). Effects of exposure measurement error when an exposure variable is constrained by a lower limit. *Am J Epidemiol.* **157**, 355-363.

Ritov, Y. and Wellner J. A. (1988). Censoring, martingale, and the Cox model, Contemporary Mathematics. *Statistical Inference for Stochastic processes, ed. N.H. Prabhu.* **80** 191-219.

Schisterman, E.F. and Little, R. J. (2010). Opening the black box of biomarker measurement error. *Epidemiology* **21**, S1- S3.

Schisterman, E.F. Vexler A., Whitcomb B.W. and Liu A. (2006). The limitations due to exposure detection limits for regression models.*Am J Epidemiol.* **163**, 374-383.

Sowers, M.R., Eyvazzadeh, A.D., McConnell, D., Yosef, M., Jannausch, M. L., Zhang, D., Harlow, S. and Randolph, J.F. Jr. (2008). Anti-mullerian hormone and inhibin B in the definition of ovarian aging and the menopause transition. *J Clin Endocrinol Metab.* **93**, 3478-3483.

Sowers, M.R., McConnell, D., Yosef, M., Jannausch, M. L., Harlow, S. and Randolph, J.F. Jr. (2010). Relating smoking, obseity, insulin resistance, and ovarian biomarker changes to the final menstrual period. *Ann. N.Y. Acad. Sci.* **1204**, 95-103.

Sowers, M., Tomey, K., Jannausch, M., Eyvazzdh, A., Crutchfield, M., Nan, B., Randolph, J. (2007). Physical functioning and menopause states. *Obstet Gynecol* **110**, 1290-1296.

Tarigan, B. and van de Geer, S. (2006). Classifiers of support vector machine type with l_1 complexity regularization. *Bernoulli* **12**, 1045–1076.

Thompsonand, M. and Nelson, K.P. (2003). Linear regression with type I interval- and left-censored response data. *Annals of Statistics.* **18**, 354- 372.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *J. R. Stat. Soc. B* **58**, 267–288.

Tibshirani, R. (1997). The Lasso method for variable selection in the Cox model. *Statistics in Medicine* **16**, 385–395.

Tsiatis, A A. (1990). Estimating regression parameters using linear rank tests for censored data. *Annals of Statistics.* **18**, 354- 372.

Tsiatis, A.A. and Davidian, M. (2004). Joint modeling of longitudinal and time-to-event data: an overview. *Statistica Sinica* **14**, 809-834.

van der Geer, S.A. (2000). *Empirical processes in M-esimation.* Cambridge University Press.

van de Geer, S. (2008). High-dimensional generalized linear models and the lasso. *Ann. Statist.* **36**, 614–645.

van der Laan, M.J. (1996). Efficient estimation in the bivariate censoring model and repairing NPMLE.*Annals of Statistics.* **24**, 596-627.

van der vaart, A. W. (2002). *Semiparametric Statistics. In Lectures on Probability and Statistics, Ecole dete de Saint-Flour XXIX1999, P. Bernard, editor.* Springer-Verlag, 330-457.

van der Vaart and Wellner, J. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics.* Wiley, New York.

Wellner, J. A. and Zhang, Y. (2007). Two likelihood-based semiparametric estimation methods for panel count data with covariates. *Annals of Statistics* **35**, 2106C2142.

Wei, L. J. , Ying, Z. and Lin, D. Y. (1990). Linear regression analysis of censored survival data based on rank tests.*Biometrika.* **77**, 845-851.

Yu, M. and Nan, B. (2009). Regression calibration in semiparametric accelerated failure time models. *Biometrics* **66**, 405- 414.