Advances in Survival Analysis: Accurate Partial Likelihood Computation by Poisson-Binomial Distributions and Nonparametric Competing Risk Cox Model

Youngjin Cho

Dissertation submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Statistics

Pang Du, Chair
Yili Hong, Co-chair
Inyoung Kim
Hongxiao Zhu

August 22, 2025
Blacksburg, Virginia

Keywords: Cox Model, Breslow Estimator, Efron Estimator, Competing Risk, Cause-Specific Hazard, Smoothing Spline ANOVA

Advances in Survival Analysis: Accurate Partial Likelihood Computation by Poisson-Binomial Distributions and Nonparametric Competing Risk Cox Model

Youngjin Cho

(ABSTRACT)

Two novel contributions to survival analysis are presented. The first project revisits the partial likelihood in the Cox model, which traditionally approximates conditional probabilities using risk score ratios under a continuous-time assumption. We propose a new accurate partial likelihood computation method based on the Poisson-binomial distribution. Although ties are common in real studies, existing Cox model theory largely overlooks tied data. In contrast, our approach accommodates both grouped data with ties and continuous data without ties, offering a unified theoretical framework for accurate partial likelihood computation regardless of data type. Simulations and real data analyses show that the method reduces bias and mean squared error while improving confidence interval coverage rates, particularly when ties are frequent or risk score variability is high. The second project develops a nonparametric regression model for competing risks survival data by combining the proportional cause-specific hazards framework with a smoothing spline ANOVA approach. We establish estimation procedures and theoretical convergence rates. Simulation studies demonstrate the method's effectiveness, and application to a multiple myeloma dataset reveals that for each gene expression covariate, at least one cause-specific effect is nonlinear and differs from the others. The proposed model fills a gap in the existing literature, where competing risks are often overlooked or covariate effects are assumed to follow parametric forms, by providing a flexible and practical framework for data analysis.

Advances in Survival Analysis: Accurate Partial Likelihood Computation by Poisson-Binomial Distributions and Nonparametric Competing Risk Cox Model

Youngjin Cho

(GENERAL AUDIENCE ABSTRACT)

Two new methodological developments in survival analysis are introduced. For the first project, the Cox model is a widely used tool for analyzing the relationship between event times and covariates, which uses partial likelihood to estimate regression coefficients. Because the original partial likelihood is computationally complex, approximations are often used in practice. We revisit the original formulation and develop a more accurate computation method by leveraging fast techniques for the Poisson binomial distribution, which underlies the partial likelihood structure. Unlike existing theory, which typically excludes ties (simultaneous events), our method accommodates them and also applies to data without ties, providing a unified framework. Simulation studies and real data applications demonstrate the improved performance of our approach, especially in settings with many ties. The second project develops a flexible regression model for competing risks, where individuals may experience one of several types of events, such as death from different causes. By combining the proportional cause-specific hazards framework with a smoothing spline ANOVA approach, for each risk factor, we enable the estimation of various nonlinear effects of covariates in a nonparametric manner. Simulation studies confirm the method's accuracy, and application to a multiple myeloma dataset reveals that for each gene expression covariate, at least one cause-specific effect is nonlinear and distinct.

Dedication

To my family — my father, my mother, and my sister — for their endless love and support.

Acknowledgments

First of all, I am deeply grateful to my advisor, Dr. Pang Du, for his tremendous and steadfast support throughout my Ph.D. studies. In September 2020, shortly after joining the graduate program in the Department of Statistics at Virginia Tech, I approached Dr. Du because I was intrigued by his research on smoothing splines. I asked him to be my advisor, and he graciously accepted. This decision is one of the best choices I have made in my life. Under Dr. Du's patient, enthusiastic, and generous guidance, I was able to build my own understanding of the methodology and theories for the smoothing splines, which now represent the primary focus of my research endeavors. He also instilled in me a profound motivation for the significance of statistical research, emphasizing how our findings can be applied by researchers across various fields, where both theoretical and applied researches have unique ways to contribute to the field. This has become the most compelling driving force behind my work in Statistics. For my dissertation, I cannot express enough gratitude to him for providing invaluable motivation in making critical theoretical assumptions, guiding me when I got stuck in technical proofs, meticulously proofreading my theoretical derivations, and assisting in writing the manuscript. Working with him on my dissertation not only improved the quality of the dissertation but also provided me with invaluable experience that will greatly benefit my future research endeavors.

I am also profoundly thankful to my co-advisor, Dr. Yili Hong, for his invaluable support throughout my Ph.D. journey. Our collaboration began in October 2021 when I enrolled in his survival analysis course, where he suggested exploring a interesting topic on the Cox model for my final project in the course, which eventually laid the groundwork for my first project in dissertation, "An Accurate Computational Approach for Partial Likelihood Using

Poisson-Binomial Distributions". With Dr. Hong's motivating, meticulous, and sincere guidance, I expanded my knowledge of methodologies in lifetime data analysis, which has become another primary focus of my research efforts. Talking with Dr. Hong, I was impressed by his passion for research and its real-world impact. This has become another motivating factor for my research in Statistics. Also, I am deeply grateful to him for providing invaluable input and guidance in formulating the main idea and constructing the model for the method proposed in the dissertation. His insights throughout the lifetime data analysis, emphasizing the unique contributions of our new method compared to existing ones, were invaluable and significantly influenced the direction of the dissertation.

In addition, I extend my deep gratitude to my committee members, Dr. Inyoung Kim and Dr. Hongxiao Zhu. I invited them to join my committee based on my profound admiration for them during their classes and discussions, which have been valuable sources for me in constructing my own foundations in statistical modeling, along with the guidance and classes provided by my advisor and co-advisor. While my current research may not directly align with the subjects covered in their classes, the insights and motivation I gained from them will undoubtedly impact my future research endeavors. The probing questions they posed, the motivational comments they offered, and the invaluable feedback they provided have been instrumental in significantly enhancing and refining the quality of my dissertation. I sincerely express my heartfelt gratitude towards them.

Furthermore, I wish to convey my gratitude to the faculty members and staffs of the Department of Statistics at Virginia Tech for their support during my Ph.D. studies. Engaging in meaningful discussions with faculty members, whether during their classes or elsewhere, has provided me with ample motivation for research endeavors. I am also appreciative of my friends in the same program not only for their collaborative spirit in discussing research topics and ideas, but also for the friendships and cherished memories formed while being together in Blacksburg. Despite it being a new environment initially, their presence has made Blacksburg feel like a second home to me.

I wish to extend my heartfelt appreciation to two professors who played pivotal roles during my time studying Statistics at Sungkyunkwan University in South Korea prior to enrolling at Virginia Tech, whether it was during my time as an undergraduate or pursuing my master's degree. Firstly, I would like to express my heartfelt gratitude to my former advisor, Dr. Seyoung Park, who has since moved to the Department of Applied Statistics at Yonsei University. Secondly, I would like to express my sincere thanks to Dr. Eun Ryung Lee. They have been paramount role models for me, fostering strong motivations for my journey to become a statistician by demonstrating excellence as a researcher and mentoring me as an experienced senior researcher. Their thorough and attentive guidance, coupled with their insightful perspective on statistics, left a lasting impact on my journey in Statistics.

Finally, I want to extend my heartfelt gratitude to my parents and my sister, as well as to my friends in South Korea and generous people who helped me adjust to life in the United States. In particular, I want to extend my deepest appreciation to my parents for their unwavering support not only throughout my Ph.D. journey but also throughout my entire life. Their support has been both unconditional and limitless, and I am truly grateful for everything they have done for me.

Attribution

This dissertation comprises two projects, both of which include coauthored material. The individual contributions are acknowledged as follows:

- *An Accurate Computational Approach for Partial Likelihood Using Poisson-Binomial Distributions*, presented in Chapter 2, is a published work [6], coauthored with Dr. Yili Hong and Dr. Pang Du, both from the Department of Statistics at Virginia Tech. The methodological and theoretical development, computational implementation, simulations and real application, and writing were primarily conducted by the author, with guidance and contributions from the coauthors.
- *Competing Risk Model with A Nonparametric Form of Spline-Estimated Relative Risks*, presented in Chapter 3, is a collaborative work with Dr. Pang Du from the Department of Statistics at Virginia Tech. The author developed the methodology and theory, conducted simulations and real data analysis, wrote the manuscript, and received critical guidance and feedback from the coauthor throughout the research.

Contents

List of Figures

Chapter 1

Introduction

This dissertation comprises two projects: the first, titled *An Accurate Computational Approach for Partial Likelihood Using Poisson-Binomial Distributions*, and the second, *Competing Risk Model with A Nonparametric Form of Spline-Estimated Relative Risks.*

1.1 An Accurate Computational Approach for Partial Likelihood Using Poisson-Binomial Distributions

1.1.1 Motivation

The Cox proportional hazard model [8], alternatively referred to as the Cox model or proportional hazard model, is undoubtedly one of the most widely used methods for analyzing survival data. It is employed to discern the relationship between covariates and the hazard of the time to event response.

In a typical survival data analysis scenario, each observation is represented by $\{t_i, \delta_i, \boldsymbol{x}_i\}$, where t_i is the observed time (event or censoring), δ_i is the event indicator (1 for event, 0 for censored), and $\boldsymbol{x}_i$ is the covariate vector with d covariates. The Cox model links covariates to the hazard function $\lambda(t; \boldsymbol{x}_i)$ as $\lambda(t; \boldsymbol{x}_i) = \lambda_0(t) \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta})$, where $\lambda_0(t)$ is the baseline hazard function and $\boldsymbol{\beta}$ is the vector of regression coefficients. This model is semiparametric because it does not specify $\lambda_0(t)$. The regression coefficients $\boldsymbol{\beta}$ are estimated using the partial likelihood (PL). The baseline cumulative hazard function $\Lambda_0(t)$ is estimated as a step function, and the cumulative hazard function $\Lambda(t; \boldsymbol{x}_i) = \int_0^t \lambda(s; \boldsymbol{x}_i) ds$ for a specific $\boldsymbol{x}_i$ can also be estimated.

To introduce PL, let $t_{(j)}$ represent the jth ordered distinct event times in the dataset $\{t_i, \delta_i, \boldsymbol{x}_i\}_{i=1}^n$, where j ranges from 1 to k. Define $\mathcal{R}(t_{(j)})$ as the at-risk set, containing subjects who survived upto $t_{(j)}$, and let $n_j \equiv |\mathcal{R}(t_{(j)})|$. Following the original idea of PL as described in Cox [9], when there are no ties, PL is constructed as $L(\boldsymbol{\beta}) = \prod_{j=1}^k L_j(\boldsymbol{\beta})$. where

$$\begin{aligned} L_j(\boldsymbol{\beta}) &= \frac{\Pr(\text{unit } j_1 \text{ had the event at } t_{(j)} \big| \, n_j \text{ units survived up to } t_{(j)})}{\Pr(1 \text{ out of } n_j \text{ units had the event at } t_{(j)} \big| \, n_j \text{ units survived up to } t_{(j)})} \\ &= \frac{d\Lambda(t_{(j)}; \boldsymbol{x}_{j_1}) \prod_{i \in \mathcal{R}(t_{(j)}) \setminus \{j_1\}} (1 - d\Lambda(t_{(j)}; \boldsymbol{x}_i))}{\sum_{i \in \mathcal{R}(t_{(j)})} \left\{ d\Lambda(t_{(j)}; \boldsymbol{x}_i) \prod_{l \in \mathcal{R}(t_{(j)}) \setminus \{i\}} (1 - d\Lambda(t_{(j)}; \boldsymbol{x}_l)) \right\}} \, . \end{aligned} \tag{1.1}$$

Here, j_1 indicates the index of the sole failed subject at time $t_{(j)}$. PL in (1.1) follows the original approach of the partial likelihood described in Cox [9], providing an accurate calculation for the contribution from observations at time $t_{(j)}$. However, this form is not commonly used in practice due to computational challenges, especially with tied data where multiple failures can occur simultaneously.

Neglecting higher-order terms assuming a continuous time model, (1.1) can be approximated by

$$L_j(\boldsymbol{\beta}) = \frac{\exp(\boldsymbol{x}_{j_1}^\top \boldsymbol{\beta})}{\sum_{i \in \mathcal{R}(t_{(j)})} \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta})}. \tag{1.2}$$

Note that continuous time model eliminates the occurrence of ties. The formula in (1.2) has long been used in practice and incorporated into major software. Moreover, it is often presented simply as the PL, burying the original idea of an accurate PL in (1.1).

Another challenge in calculating the accurate PL arises from tied event times, complicating the process even more compared to the accurate PL calculations without ties in (1.1). Let d_j represent the number of units that have failed at $t_{(j)}$. To address ties in Cox models, various approximating methods have been proposed, including the Cox correction [8], the Kalbfleisch-Prentice correction [26], the Breslow correction [4], and the Efron correction [13], where all of these methods are direct extensions of (1.2) to scenarios with ties, involving the incorporation of the averaged contribution of d_j units that experienced an event at $t_{(j)}$ to calculate the tie-corrected PL. These methods are widely accepted for handling ties in Cox models within statistical software. However, they may not be ideal because they are based on (1.2), which is already an approximation to (1.1) assuming continuous times without ties. Also, their averaging approaches oversimplify the contributions of tied event times, neglecting distributional differences among ties. These factors lead to deviations from the original accurate PL for tied data. Additionally, to our knowledge, no rigorous studies have examined the asymptotic behavior of these correction methods with respect to the extent of permissible order of ties in the model.

In summary, the conventional utilization of PL in practice, whether for data without ties or data with ties, represents an approximated PL. Furthermore, for tied data cases, owing to the reasons mentioned earlier, relying on existing tie correction methods may not be ideal.

1.1.2 Literature Review

Before we introduce our new challenges in Section 1.1.3, which are motivated by the limitations of approximated PL methods widely utilized in practice as discussed in Section 1.1.1, we provide a review of relevant literature.

Tie Correction Methods for Tied Survival Data

Here, we outline tie correction methods referenced in Section 1.1.1 for data with ties. Note that the accurate PL for tied data entails a complex structure and requires extensive explanations regarding its underlying settings for tie generation scenario, i.e., grouped data scenario. Given the intricacies involved, this accurate PL under ties generated by grouped data scenario, which is denoted by

$$L(\boldsymbol{\beta}, \boldsymbol{\Lambda}) = \prod_{j=1}^{k} L_j(\boldsymbol{\beta}, \lambda_j) = \prod_{j=1}^{k} \frac{A_j(\boldsymbol{\beta}, \lambda_j)}{B_j(\boldsymbol{\beta}, \lambda_j)}, \tag{1.3}$$

is not detailed here but is thoroughly discussed in Chapter 2.

Let $\mathcal{D}(t_{(j)})$ denote an event set at time $t_{(j)}$, where $d_j \equiv |\mathcal{D}(t_{(j)})|$ and $j_1, \ldots, j_{d_j}$ represent the d_j individuals in $\mathcal{D}(t_{(j)})$. Assuming grouped data scenario, where ties are allowed, thus d_j may be larger than 1. The Cox correction [8] employs the following PL term,

$$L_j(\boldsymbol{\beta}, \lambda_j) \approx \frac{\exp\left(\sum_{i \in \mathcal{D}(t_{(j)})} \boldsymbol{x}_i^\top \boldsymbol{\beta}\right)}{\sum_{\mathcal{A}_{d_j} \in \mathcal{F}_{d_j}} \exp\left(\sum_{i \in \mathcal{A}_{d_j}} \boldsymbol{x}_i^\top \boldsymbol{\beta}\right)},$$

where $\mathcal{F}_{d_j}$ represents the set of all subsets of d_j individuals that can be selected from $\mathcal{R}(t_{(j)})$. To clarify, each set $\mathcal{A}_{d_j} \in \mathcal{F}_{d_j}$ contains d_j elements, and $|\mathcal{F}_{d_j}| = \binom{n_j}{d_j}$. Let Q_j denote the set of $d_j!$ permutations of the subjects $j_1, \ldots, j_{d_j}$. Let $S_l^j = \{s_{l1}^j, \ldots, s_{ld_j}^j\}$ represent the lth element in Q_j. We can express Q_j as $Q_j = \{S_1^j, \ldots, S_{d_j!}^j\}$ and $\mathcal{R}(t_{(j)}, S_l^j, m) = \mathcal{R}(t_{(j)}) \setminus \{s_{l1}^j, \ldots, s_{l,m-1}^j\}$. The Kalbfleisch-Prentice correction [26] utilizes the following PL term,

$$L_j(\boldsymbol{\beta}, \lambda_j) \approx \frac{1}{d_j!} \sum_{l=1}^{d_j!} \frac{\exp\left(\sum_{i \in \mathcal{D}(t_{(j)})} \boldsymbol{x}_i^\top \boldsymbol{\beta}\right)}{\prod_{m=1}^{d_j} \left\{\sum_{i \in \mathcal{R}(t_{(j)}, S_l^j, m)} \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta})\right\}}.$$

The Breslow correction [4] employs the following PL term,

$$L_j(\boldsymbol{\beta}, \lambda_j) \propto \frac{\exp\left(\sum_{i\in\mathcal{D}(t_{(j)})} \boldsymbol{x}_i^\top \boldsymbol{\beta}\right)}{\left\{\sum_{i\in\mathcal{R}(t_{(j)})} \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta})\right\}^{d_j}}.$$

The Efron correction [13], which can be regarded as a centralized version of the Breslow correction, utilizes the following PL term,

$$L_j(\boldsymbol{\beta}, \lambda_j) \propto \frac{\exp\left(\sum_{i\in\mathcal{D}(t_{(j)})} \boldsymbol{x}_i^\top \boldsymbol{\beta}\right)}{\prod_{\ell=0}^{d_j-1}\left\{\sum_{i\in\mathcal{R}(t_{(j)})} \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta}) - \ell\overline{A}\left(\boldsymbol{\beta}, t_{(j)}\right)\right\}},$$

where $\overline{A}(\boldsymbol{\beta}, t_{(j)}) = \sum_{i\in\mathcal{D}(t_{(j)})} \exp\left(\boldsymbol{x}_i^\top \boldsymbol{\beta}\right)/d_j$. Note that the Cox correction, the Kalbfleisch-Prentice correction, the Breslow correction, and the Efron correction are all the same as (1.2) when $d_j = 1$.

Poisson Binomial Distribution

Here, we introduce the Poisson binomial (PB) distribution, which holds a significant role in our proposed accurate PL method discussed in Section 1.1.3 and detailed in Chapter 2.

Let $\mathcal{I}_1, \cdots, \mathcal{I}_m$ are independent random indicators, where each $\mathcal{I}_i$ follows Bernoulli distribution with probability $\mathcal{P}_i$. Define $\mathcal{I} = \sum_{i=1}^m \mathcal{I}_i$. For any $\ell \in \{0, 1, \cdots, m\}$, the probability mass function of $\mathcal{I}$ at ℓ can be calculated as

$$\Pr(\mathcal{I} = \ell) = \sum_{\mathcal{A}_\ell \in \mathcal{F}_\ell} \left\{ \prod_{i\in\mathcal{A}_\ell} \mathcal{P}_i \prod_{i\in\{1,\cdots,m\}\setminus\mathcal{A}_\ell} (1 - \mathcal{P}_i) \right\}, \tag{1.4}$$

where $\mathcal{F}_\ell$ represents the set of all subsets of ℓ individuals that can be selected from $\{1, \cdots, m\}$. In other words, each set $\mathcal{A}_\ell \in \mathcal{F}_\ell$ contains ℓ elements, and $|\mathcal{F}_\ell| = \binom{m}{\ell}$. We call the distribution of $\mathcal{J}$ the PB distribution with parameter $\{\mathcal{P}_i\}_{i=1}^m$ [22].

Because of its intricate structure, exact calculations for the probability mass function of the PB distribution have rarely been employed in practice. Instead, the following two approximation methods for the probability mass function of PB distribution have been commonly utilized. Firstly, the Poisson approximation to (1.4) is,

$$\Pr(\mathcal{J} = \ell) \approx \frac{\mathcal{M}^\ell \exp(-\mathcal{M})}{\ell!},$$

where $\mathcal{M} = \sum_{i=1}^m \mathcal{P}_i$ is the mean of $\mathcal{J}$. Secondly, the Normal approximation to (1.4) is,

$$\Pr(\mathcal{J} = \ell) \approx \frac{1}{\sqrt{2\pi\mathcal{V}^2}} \exp\left(-\frac{1}{2\mathcal{V}^2}\left(\ell - \mathcal{M}\right)^2\right),$$

where $\mathcal{V}^2 = \sum_{i=1}^m \mathcal{P}_i(1 - \mathcal{P}_i)$ is the variance of $\mathcal{J}$.

However, recently, the discrete Fourier transform of the characteristic function (DFT-CF) method, proposed by Hong [21], has emerged as an efficient approach for computing the probability mass function of the PB distribution using the fast Fourier transform. This method is conveniently accessible through the R package “`poibin`” developed by Hong [22]. We leverage the DFT-CF method to compute the probability mass function of PB distribution in our proposed method.

It is acknowledged that for continuous data without ties, the denominator of accurate PL (1.1) is exactly the probability mass function of a PB distribution. Similarly, for grouped data with ties, the denominator of accurate PL (1.3), which is meticulously explained in Chapter 2, is also exactly the probability mass function of a PB distribution. Notably, regardless of the presence of ties, accurate PL inherently involves PB distribution.

1.1.3 New Challenges

As detailed in Section 1.1.1, despite many studies on Cox models, the original idea of accurate PL has received limited attention due to the computational challenges associated with its denominator, particularly exacerbated for the grouped data with ties. Here, we revisit the original idea of accurate PL and introduce a novel method for computing accurate PL based on the PB distribution advancements presented in Hong [21]. This method effectively addresses the computational complexities associated with the denominator and accommodates tied data. The most important idea of our approach lies in recognizing that the denominator of the accurate PL is the probability mass function of a PB distribution, irrespective of the presence of ties. Consequently, efficient computational techniques for PB distribution introduced by Hong [21] can be leveraged when computing the accurate PL. In Chapter 2, we elaborate on our proposed method, derive its theoretical properties, and demonstrate its advantages over existing approaches through comprehensive numerical studies.

1.2 Competing Risk Model with A Nonparametric Form of Spline-Estimated Relative Risks

1.2.1 Motivation

In survival analysis, it is frequently encountered that each subject is at risk of experiencing one of several distinct and mutually exclusive failure types, a situation commonly referred to as competing risks. When covariates are available, various modeling strategies have been proposed to accommodate the structure imposed by competing risks. A widely used method is the cause-specific hazards model under the proportional hazards assumption, often referred to as the competing risk Cox model [27, 39]. Subsequently, alternative and extended methodologies have been developed. For example, the proportional subdistribution hazards model was introduced to directly quantify the effect of covariates on the cumulative incidence function [15]. Other approaches include data augmentation techniques for modeling all cause-specific hazards jointly through standard Cox regression [33], and copula-based frameworks that describe the joint distribution of latent failure times [31]. A comprehensive review on competing risk survival analysis can be found in Monterrubio-Gómez et al. [37].

Nonparametric modeling becomes essential when the effect of covariates on the hazard function is nonlinear. Among available tools, smoothing spline ANOVA offers both flexibility and interpretability, enabling decomposition of a multivariate function into additive components and interactions, providing insight into the individual and joint effects of covariates. For a thorough review, see Gu [18]. This modeling approach has been developed for both hazard-based and proportional hazards settings [18]. Theoretical properties, such as convergence rates for nonparametric estimation, have been established in both the general hazard model [18] and the proportional hazards context [38]. In the setting of correlated survival data, smoothing spline ANOVA frailty models have been proposed [10, 12].

Despite parallel advancements in the competing risks literature and in nonparametric survival modeling, the integration of these two directions remains uncommon. For discrete-time competing risks data, Luo et al. [34] introduced a Cox–logistic model where baseline hazards are estimated with spline expansions, but covariate effects remain linear. A joint distribution method using kernel was proposed by Fermanian [14], decomposing the survival function into a cumulative distribution function, cause-specific time functions, and cause-specific covariate functions. While fully nonparametric, this approach is complex, difficult to implement, and yields estimates that are often hard to interpret.

In summary, both competing risk survival analysis and nonparametric modeling for survival data have been extensively developed within their respective domains, but their integration remains relatively underexplored despite the long-standing history of each.

1.2.2 Literature Review

Prior to presenting the new challenges in Section 1.2.3, which are inspired by the scarcity of studies integrating competing risk survival analysis with smoothing spline ANOVA as highlighted in Section 1.2.1, we give a review of the related literature.

Proportional Cause-Specific Hazards Model

Here, we provide a thorough review of the proportional cause-specific hazards model [27, 39] in Section 1.2.1.

Let $\tilde{T} \in (0, \infty)$ denote the true event time, assumed to be continuously distributed. The continuous non-informative right-censoring time is denoted by $C \in (0, \tau]$, where $\tau \in (0, \infty)$ represents the end of the study period. Let $X = (X_{[1]}, \dots, X_{[d]})^\top \in \mathcal{X} = \prod_{j=1}^{d} \mathcal{X}_{[j]}$ be a vector of covariates. We consider the setting of K mutually exclusive event types (risks), indexed by the set $\{1, \dots, K\}$. The underlying cause of failure is denoted by $\tilde{G} = \sum_{k=1}^{K} k\mathbb{1}(\tilde{T} \text{ is due to the } k\text{-th risk}) \in \{1, \dots, K\}$. Each individual is assumed to be susceptible to only one of the K risks. We assume that $\tilde{T}$ and C are conditionally independent given X.

The observed time is defined as $T = \min\{\tilde{T}, C\} \in (0, \tau]$. The event indicator is defined as

$$\mathcal{I} = \mathbb{1}(\tilde{T} \leq C) = \begin{cases} 1, & \text{if the event is observed,} \\ 0, & \text{if the observation is censored.} \end{cases}$$

The observed event type is then $G = \tilde{G} \times \mathcal{I} \in \{0, 1, \dots, K\}$, where $G = 0$ indicates that the event was censored, and the true cause was unobserved.

Let $\{W_i\}_{i=1}^n = \{\tilde{T}_i, C_i, X_i, \tilde{G}_i\}_{i=1}^n$ be an independent and identically distributed (i.i.d.) sample drawn from the distribution of $W = \{\tilde{T}, C, X, \tilde{G}\}$. The observed data consist of $\{T_i, X_i, \mathcal{I}_i, G_i\}_{i=1}^n$, calculated analogously to $\{T, X, \mathcal{I}, G\}$ for each individual i.

In the competing risks framework, for $t \in [0, \infty)$, the conditional distributional functions of $\tilde{T}$ given X are defined. The hazard function is

$$\begin{aligned} dH(t;X) &= h(t;X)dt = \mathbb{P}(\tilde{T} \in [t, t+dt) \mid \tilde{T} \geq t, X) \\ &= \sum_{k=1}^{K} \mathbb{P}(\tilde{T} \in [t, t+dt), \tilde{G} = k \mid \tilde{T} \geq t, X) = \sum_{k=1}^{K} dH_k(t;X) = \sum_{k=1}^{K} h_k(t;X)dt, \end{aligned}$$

where $dH_k(t;X)$ is the cause-specific hazard function corresponding to the k-th risk.

The survival function is $\mathsf{S}(t;X) = \mathbb{P}(\tilde{T} > t \mid X) = \exp(-H(t;X))$, and the cumulative distribution function is $\mathsf{F}(t;X) = \mathbb{P}(\tilde{T} \leq t \mid X) = 1 - \mathsf{S}(t;X)$. The overall density function is $\mathsf{f}(t;X) = \partial\mathsf{F}(t;X)/(\partial t) = h(t;X)\mathsf{S}(t;X) = \sum_{k=1}^{K} \mathsf{f}_k(t;X)$, where $\mathsf{f}_k(t;X) = h_k(t;X)\mathsf{S}(t;X)$. The cumulative distribution function can also be decomposed as $\mathsf{F}(t;X) = \sum_{k=1}^{K} \mathsf{F}_k(t;X)$, where $\mathsf{F}_k(t;X) = \mathbb{P}(\tilde{T} \leq t, \tilde{G} = k \mid X) = \int_0^t \mathsf{f}_k(s;X)ds$ is the cumulative incidence function for the k-th cause. Let $\mathsf{P}(k;X) = \mathbb{P}(\tilde{G} = k \mid X) = \lim_{t\to\infty} \mathsf{F}_k(t;X)$, where $\sum_{k=1}^{K} \mathsf{P}(k;X) = 1$.

For each k, we adopt the proportional cause-specific hazards model, under which the cause-specific hazard function for the k-th cause is expressed as

$$dH_k(t;X) = dH_{0k}(t)\exp(X^\top \boldsymbol{\beta}_k^*),$$

where $\boldsymbol{\beta}_k^* = (\beta_{k,1}^*, \dots, \beta_{k,d}^*)^\top$ is the linear covariate effect for the k-th cause and $dH_{0k}(t) = h_{0k}(t)dt$ represents the baseline hazard for the k-th cause, with the convention that $h_{0k}(0) =$

0.

Define the counting process and at-risk process used in the construction of the partial likelihood. Let $t \in [0, \tau]$. For each risk k, the counting process for the k-th cause is given by

$$N_k(t) = \mathbb{1}(T \leq t, G = k, \mathfrak{I} = 1) = \mathbb{1}(\tilde{T} \leq t, \tilde{T} \leq C, \tilde{G} = k),$$

and the at-risk process is defined as

$$Y(t) = \mathbb{1}(T \geq t) = \mathbb{1}(\tilde{T} \geq t \text{ and } C \geq t).$$

For each individual $i = 1, \ldots, n$, the individual-level counting and at-risk processes are denoted by $N_{ik}(t)$ and $Y_i(t)$, respectively, defined analogously to $N_k(t)$ and $Y(t)$.

Given $\boldsymbol{\beta} \in \mathbb{R}^d$, the negative log-partial likelihood function corresponding to the k-th cause is defined as

$$\mathcal{L}_{k(n)}(\boldsymbol{\beta}) = -\int_0^\tau \frac{1}{n} \sum_{i=1}^{n} \left\{ X_i^\top \boldsymbol{\beta} - \ln \left(\frac{1}{n} \sum_{i'=1}^{n} Y_{i'}(t) \exp(X_{i'}^\top \boldsymbol{\beta}) \right) \right\} dN_{ik}(t)$$

and the estimation of $\boldsymbol{\beta}_k^*$ is done by

$$\hat{\boldsymbol{\beta}}_k^* = \underset{\boldsymbol{\beta} \in \mathbb{R}^d}{\operatorname{argmin}} \; \mathcal{L}_{k(n)}(\boldsymbol{\beta}).$$

For large sample properties of this model, refer to Kalbfleisch and Prentice [27].

Smoothing Spline ANOVA Proportional Hazards Model

We review smoothing spline ANOVA [18] in Section 1.2.1 for proportional hazards model, without considering competing risks at this stage. Other variants of smoothing spline

ANOVA—such as those used for exponential family response regression, density estimation, and hazard modeling—are formulated in a similar manner. For further details, see Gu [18]. For simplicity, we assume that all d covariates are continuous, although in general each of the d covariates can be either continuous or categorical [18].

Since competing risks are not considered here, we disregard $\tilde{G}$ and G—that is, we set $K = 1$. Accordingly, $\{\tilde{T}_i, C_i, X_i\}_{i=1}^n$ are n i.i.d. copies of $\{\tilde{T}, C, X\}$ and the observed data $\{T_i, X_i, \mathcal{I}_i\}_{i=1}^n$ is structured analogously to $\{T, X, \mathcal{I}\}$ for each individual i. We model the hazard function under the proportional hazards assumption with a nonparametric covariate effect:

$$dH(t; X) = dH_0(t)\exp(\eta^*(X)),$$

where $\eta^* : \mathcal{X} \to \mathbb{R}$ denotes the nonparametric covariate effect, and $dH_0(t) = h_0(t)\,dt$ represents the baseline hazard function, with the convention that $h_0(0) = 0$. We assume $\eta^* \in \mathcal{H}$ for some reproducing kernel hilbert space (RKHS) $\mathcal{H}$.

Define RKHS $\mathcal{H}$, which captures the structural characteristics of the model under consideration, using the smoothing spline ANOVA framework introduced by Gu [18]. Let $m \in \mathbb{N}$. The Sobolev space of order m associated with the jth axis, $\mathcal{X}_{[j]}$, is given by

$$\mathcal{H}_{[j]} = \left\{ f \in \mathcal{L}_2(\mathcal{X}_{[j]}) : f, f^{(1)}, \ldots, f^{(m-1)} \text{ are absolutely continuous, } f^{(m)} \in \mathcal{L}_2(\mathcal{X}_{[j]}) \right\},$$

where $\mathcal{L}_2(\mathcal{X}_{[j]}) = \left\{ f : \mathcal{X}_{[j]} \to \mathbb{R}, \int_{\mathcal{X}_{[j]}} |f(x_{[j]})|^2 dx_{[j]} < \infty \right\}$ denotes the square-integrable function space on $\mathcal{X}_{[j]}$.

Let $\mathcal{A}_{[j]}$ denote the averaging operator along the jth axis, defined as

$$\mathcal{A}_{[j]}f(x) = \mathcal{A}_{[j]}f(x_{[1]}, \ldots, x_{[d]}) = \int_{\mathcal{X}_{[j]}} f(x_{[1]}, \ldots, x_{[d]})dx_{[j]}.$$

Using this operator, the space $\mathcal{H}_{[j]}$ can be decomposed as

$$\mathcal{H}_{[j]} = \mathcal{H}_{\emptyset[j]} \oplus \mathcal{H}_{\{j\}},$$

where $\mathcal{H}_{\emptyset[j]} = \left\{c_j \mathcal{A}_{[j]} f_{[j]} : c_j \in \mathbb{R}, f_{[j]} \in \mathcal{H}_{[j]}\right\} = \text{Span}\{1\}$, and $\mathcal{H}_{\{j\}} = \{f_{[j]} \in \mathcal{H}_{[j]} : \mathcal{A}_{[j]} f_{[j]} = 0\}$. Any function $f_{[j]} \in \mathcal{H}_{[j]}$ admits a decomposition of the form $f_{[j]} = f_{\emptyset[j]} + f_{\{j\}}$, where $f_{\emptyset[j]} = \mathcal{A}_{[j]} f_{[j]} \in \mathcal{H}_{\emptyset[j]}$, $f_{\{j\}} = (id - \mathcal{A}_{[j]}) f_{[j]} \in \mathcal{H}_{\{j\}}$, and id denotes the identity operator.

The tensor product Sobolev space over the full domain $\mathcal{X}$, formed via tensor products over all coordinate directions, is

$$\begin{aligned}\otimes_{j=1}^{d} \mathcal{H}_{[j]} &= \otimes_{j=1}^{d} \left\{\mathcal{H}_{\emptyset[j]} \oplus \mathcal{H}_{\{j\}}\right\} \\ &= \oplus_{S \in \mathcal{P}_d} \mathcal{H}_S = \mathcal{H}_\emptyset \oplus \left\{\oplus_{j=1}^{d} \mathcal{H}_{\{j\}}\right\} \oplus \left\{\oplus_{j<j'} \mathcal{H}_{\{j,j'\}}\right\} \oplus \ldots \oplus \mathcal{H}_{\{1,\ldots,d\}},\end{aligned}$$

where $\mathcal{P}_d$ denotes the power set of $\{1, \ldots, d\}$, and $|\mathcal{P}_d| = 2^d$. As an illustration, when $d = 3$, we have $\mathcal{P}_d = \{\emptyset, \{1\}, \{2\}, \{3\}, \{1,2\}, \{1,3\}, \{2,3\}, \{1,2,3\}\}$. For each $S \in \mathcal{P}_d$, the corresponding subspace is defined as

$$\mathcal{H}_S = \left\{\otimes_{j \in S} \mathcal{H}_{\{j\}}\right\} \otimes \left\{\otimes_{j \in \{1,\ldots,d\} \setminus S} \mathcal{H}_{\emptyset[j]}\right\} = \left\{\otimes_{j \in S} \mathcal{H}_{\{j\}}\right\} \otimes \text{Span}\{1\} = \otimes_{j \in S} \mathcal{H}_{\{j\}}.$$

The domain of $\mathcal{H}_S$ is $\mathcal{X}_S = \prod_{j \in S} \mathcal{X}_{[j]}$, and the subspace $\mathcal{H}_S$ represents the function components associated with effect S. The constant function space is given by $\mathcal{H}_\emptyset = \otimes_{j=1}^{d} \mathcal{H}_{\emptyset[j]} = \text{Span}\{1\}$. Every function $f \in \otimes_{j=1}^{d} \mathcal{H}_{[j]}$ admits a decomposition of the form $f = \sum_{S \in \mathcal{P}_d} f_S$, where each component f_S is given by

$$f_S = \left\{ \prod_{j \in S} (id - \mathcal{A}_{[j]}) \prod_{j \in \{1,\ldots,d\} \setminus S} \mathcal{A}_{[j]} \right\} f \in \mathcal{H}_S.$$

Let $\mathcal{H} = \oplus_{S \in \mathbb{S}} \mathcal{H}_S$ for some subset $\emptyset \notin \mathbb{S} \subseteq \mathcal{P}_d$. Note that $\mathcal{H}_\emptyset$ is excluded from $\mathcal{H}$, as the Cox model does not incorporate an intercept term. The structure of $\mathcal{H}$ is determined by the choice of $\mathbb{S}$, allowing one to control the complexity of the model. For example, taking $\mathbb{S} = \{\{1\}, \ldots, \{d\}\}$ yields an additive model involving only main effects. In contrast, setting $\mathbb{S} = \{\{1\}, \ldots, \{d\}, \{1,2\}, \ldots, \{d-1, d\}\}$ includes all two-way interactions along with main effects.

Given a semi-inner product $J(\cdot, \cdot)$ defined on $\mathcal{H}$, the space $\mathcal{H}$ can be decomposed as

$$\mathcal{H} = \mathcal{N}_J \oplus \mathcal{H}_J,$$

where $\mathcal{N}_J = \{f \in \mathcal{H} : J(f) = J(f,f) = 0\} = \text{Span}\{\phi_1, \ldots, \phi_p\}$ is the null space of the penalty and corresponds to a finite-dimensional parametric component. The complement $\mathcal{H}_J$ is an infinite-dimensional nonparametric subspace. On $\mathcal{H}_J$, the semi-inner product $J(\cdot, \cdot)$ becomes an inner product and induces a reproducing kernel $\mathcal{K}_J(\cdot, \cdot)$.

Let $t \in [0, \tau]$. The counting process is defined by

$$N(t) = \mathbb{1}(T \leq t, \mathfrak{I} = 1) = \mathbb{1}(\tilde{T} \leq t, \tilde{T} \leq C),$$

and for each subject i, $N_i(\cdot)$ is defined analogously.

For $f \in \mathcal{H}$, the negative log-partial likelihood is given by

$$\mathcal{L}_{(n)}(f) = -\int_0^\tau \frac{1}{n}\sum_{i=1}^n \left\{ f(X_i) - \ln\left(\frac{1}{n}\sum_{i'=1}^n Y_{i'}(t)\exp(f(X_{i'})) \right) \right\} dN_i(t).$$

A subset of q covariate values is randomly selected from the n subjects, denoted as $\{Z_\ell\}_{\ell=1}^q \subseteq \{X_i\}_{i=1}^n$. Define the space $\mathcal{H}_J^{(q)} = \text{Span}\{\mathcal{K}_J(Z_1,\cdot),\ldots,\mathcal{K}_J(Z_q,\cdot)\}$, which forms a finite-dimensional subspace of $\mathcal{H}_J$. We then set $\mathcal{H}^{(q)} = \mathcal{N}_J \oplus \mathcal{H}_J^{(q)}$, a finite-dimensional subspace of $\mathcal{H}$. Applying the efficient approximation from Gu [18], the estimator of η^* is obtained by

$$\hat{\eta} = \underset{f\in\mathcal{H}^{(q)}}{\text{argmin}}\ \mathcal{L}_{(n)}(f) + \frac{\lambda}{2}J(f), \tag{1.5}$$

where $J(f) = J(f,f)$ serves as the roughness penalty and $\lambda \in (0,\infty)$ is the smoothing parameter. In the special case where $q = n$, the set $\{Z_\ell\}_{\ell=1}^q$ becomes identical to $\{X_i\}_{i=1}^n$, and by the representer theorem, we have $\hat{\eta} = \hat{\eta}^* \equiv \underset{f\in\mathcal{H}}{\text{argmin}}\ \mathcal{L}_{(n)}(f) + \lambda J(f)/2$.

Solution to the optimization problem in (1.5) can then be expressed as

$$\hat{\eta} = \sum_{l=1}^p \hat{d}_l\phi_l + \sum_{\ell=1}^q \hat{c}_\ell \mathcal{K}_J(Z_\ell,\cdot),$$

where

$$(\hat{\boldsymbol{d}}^\top, \hat{\boldsymbol{c}}^\top)^\top = \underset{\boldsymbol{d}\in\mathbb{R}^p,\boldsymbol{c}\in\mathbb{R}^q}{\text{argmin}}\ \mathcal{L}_{(n)}\left(\sum_{l=1}^p d_l\phi_l + \sum_{\ell=1}^q c_\ell\mathcal{K}_J(Z_\ell,\cdot) \right) + \frac{\lambda}{2}\boldsymbol{c}^\top\boldsymbol{\mathcal{K}}_J\boldsymbol{c}$$

with $\boldsymbol{\mathcal{K}}_J = \{\mathcal{K}_J(Z_\ell, Z_{\ell'})\}_{\ell,\ell'\in\{1,\ldots,q\}}$.

1.2.3 New Challenges

As described in Section 1.2.1, although the proportional cause-specific hazards model is a classical and widely adopted approach for analyzing competing risks data, it has not been previously integrated with nonparametric modeling of covariate effects. Such integration can be achieved by the smoothing spline ANOVA framework to accommodate flexible, nonlinear covariate effects. This dissertation introduces a new modeling framework that unites the cause-specific Cox model with smoothing spline ANOVA, enabling flexible nonparametric covariate effects estimation under the proportional cause-specific hazards assumption in competing risks data. We refer to this method as the smoothing spline competing risk Cox model, which is developed in detail in Chapter 3, along with its theoretical properties and supporting numerical studies.

Chapter 2

An Accurate Computational Approach for Partial Likelihood Using Poisson-Binomial Distributions

2.1 Introduction

This project is a published work [6], coauthored with Dr. Yili Hong and Dr. Pang Du, both from the Department of Statistics at Virginia Tech.

Ever since its birth half a century ago, the Cox model [8] has been arguably the most widely used method to analyze survival data with covariates. Since the literature for Cox models is too vast to give a comprehensive review here, we only name a few widely-used monographs here, such as Fleming and Harrington [16], Therneau and Grambsch [47], Kalbfleisch and Prentice [27], and Klein and Moeschberger [29].

Consider a typical survival data setting where observations are denoted by $\{t_i, \delta_i, \boldsymbol{x}_i\}$, $i = 1, \ldots, n$. Here, n is the number of subjects in the study, t_i is the observed event time or censoring time, and δ_i is the event indicator that equals to 1 if the observed time is from an event and 0 if the event time is censored. The covariate vector is denoted by $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{id})^\top$, where d is the number of covariates. The Cox model [9] incorporates the covariate effects into the hazard function as $\lambda(t;\boldsymbol{x}_i) = \lambda_0(t)\exp(\boldsymbol{x}_i^\top\boldsymbol{\beta})$, where $\lambda_0(t)$ is the unknown baseline hazard function and $\boldsymbol{\beta} = (\beta_1,\ldots,\beta_d)^\top$ is the vector of regression coefficients. Let $\Lambda_0(t) = \int_0^t \lambda_0(s)ds$ be the baseline cumulative hazard function, which is often estimated as a step function. Denote the cumulative hazard function for a specific $\boldsymbol{x}_i$ by $\Lambda(t;\boldsymbol{x}_i) = \int_0^t \lambda(s;\boldsymbol{x}_i)ds$. The regression coefficients $\boldsymbol{\beta}$ are typically estimated by the partial likelihood.

To revisit the idea of partial likelihood, let $t_{(j)}$, $j = 1,\ldots,k$, be the k ordered *distinct* events times in data $\{t_i,\delta_i,\boldsymbol{x}_i\}_{i=1}^n$. To clarify, when referring to something up to t, this includes all times from 0 to t^- but explicitly excludes t itself. Let $\mathcal{R}(t_{(j)})$ be the at-risk set containing all the subjects that survived up to time $t_{(j)}$. Let $n_j \equiv |\mathcal{R}(t_{(j)})|$. When there are no ties, Cox [9] constructs the partial likelihood as $L(\boldsymbol{\beta}) = \prod_{j=1}^k L_j(\boldsymbol{\beta})$, where

$$\begin{aligned} L_j(\boldsymbol{\beta}) &= \frac{\Pr(\text{unit } j_1 \text{ failed at } t_{(j)} \big| \, n_j \text{ units survived up to } t_{(j)})}{\Pr(1 \text{ out of } n_j \text{ units failed at } t_{(j)} \big| \, n_j \text{ units survived up to } t_{(j)})} \\ &= \frac{d\Lambda(t_{(j)};\boldsymbol{x}_{j_1})\prod_{i\in\mathcal{R}(t_{(j)})\setminus\{j_1\}}(1-d\Lambda(t_{(j)};\boldsymbol{x}_i))}{\sum_{i\in\mathcal{R}(t_{(j)})}\left\{d\Lambda(t_{(j)};\boldsymbol{x}_i)\prod_{l\in\mathcal{R}(t_{(j)})\setminus\{i\}}(1-d\Lambda(t_{(j)};\boldsymbol{x}_l))\right\}} \end{aligned} \tag{2.1}$$

is the partial likelihood contribution from observations at time $t_{(j)}$, and j_1 is the index for the only failed subject at time $t_{(j)}$. The expression (2.1) from Cox [9] represents the accurate likelihood contribution from the observation at time $t_{(j)}$. We refer to (2.1) as the *accurate partial likelihood* (APL). Despite its accuracy, (2.1) is not widely used in practice due to its computational difficulty in the denominator.

Ignoring higher order terms under a continuous failure time model when no ties are present, (2.1) can be approximated by

$$L_j(\boldsymbol{\beta}) = \frac{\exp(\boldsymbol{x}_{j_1}^\top\boldsymbol{\beta})}{\sum_{i\in\mathcal{R}(t_{(j)})}\exp(\boldsymbol{x}_i^\top\boldsymbol{\beta})}; \tag{2.2}$$

see, e.g., Page 140 of Fleming and Harrington [16]. For decades, (2.2) has been used in practice and implemented in major software packages. In many textbooks and research papers, (2.2) is often directly introduced as the partial likelihood and the original APL idea gets buried. The goal of this project is to re-visit the idea of APL with some new development from another area, by realizing that the denominator of (2.1) is in the form of the probability mass function of a Poisson-binomial (PB) distribution. The PB distribution describes the sum of independent but non-identically distributed random indicators [3, 21].

When ties are present, the calculation of the APL gets even harder and more time consuming. To incorporate ties into Cox models, several approximating approaches have been proposed. In essence, they are all direct extensions of (2.2) to the with-tie scenario, with some additional ad hoc approximations to include the contributions from all the event times tied at a time point. Suppose d_j is the number of failures at time $t_{(j)}$. The Cox correction [8] averages the corresponding partial likelihoods over all the size-d_j subsets of the at risk set at time $t_{(j)}$. The Kalbfleisch-Prentice correction [26] averages on all the possible permutations of the d_j underlying event times. The Breslow correction [4] is a simplified version of the Kalbfleisch-Prentice correction assuming an equal contribution from each permutation. The Efron correction [13] can be viewed as a centerized version of the Breslow correction. Therneau and Grambsch [47] provided an overview of tie corrections in partial likelihood and also introduced various Cox regression based models. Hertz-Picciotto and Rockhill [19] and Scheike and Sun [43] used numerical studies to compare performances of different tie correction methods, and similar simulation settings are adopted in this project. These methods are now standards for tie corrections in Cox models and well-implemented in all the major statistical software.

However, they may not be the ideal approaches for tie corrections. Firstly, they are all based on (2.2), which is already an approximation to the APL and derived actually for continuous failure times without ties. Secondly, the various averaging approaches are really crude ways to count the contributions of all the tied event times without fully incorporating the distributional differences among these ties. This creates further deviations from the APL. Furthermore, as far as we know, there are no formal investigations on the asymptotic properties of these correction methods under consideration for how large the permissible ties can be in the model.

In this project, we propose a new computationally efficient method to calculate the APL based on the PB distribution development in Hong [21]. The key idea is that the denominator of the APL is exactly in the form of the probability mass function of a PB distribution, regardless the presence of ties or not. We use the method in Hong [21] to compute the PB probability mass function, which is based on the discrete Fourier transformation of the characteristic function. Alternatively, one can also use the convolution-based method in Biscarri et al. [3] based on the direct convolution or the divide-and-conquer fast Fourier transform tree convolution. As far as we know, the idea of a direct and exact computation of the APL for the Cox model in this project is completely new to the literature.

We consider the common scenario where ties are caused by rounding or grouping of underlying continuous failure times, which we refer to as the grouped continuous failure time model. Our first result shows that all the aforementioned common methods, namely, the expression (2.2), the Cox correction, the Kalbfleisch-Prentice correction, the Breslow correction, and the Efron correction, are all approximations to the PB distribution approach based on Poisson approximation or approximation by enumeration averaging. Under the grouped continuous failure time model, we first establish the consistency and asymptotic normality for the Breslow estimator, which can be easily extended to the estimators from the other existing correction methods. For our PB distribution approach, we establish the consistency and asymptotic normality for its estimator under both the grouped continuous failure time model (ties present) and the ungrouped one (no ties). Note that although Prentice and Kalbfleisch [40] provided theoretical results for the Breslow estimator in the presence of ties, they did not address how large the order of ties can be for the model, whereas we provide practical insights into the permissible order of ties. In simulations, we compare the performance of our PB distribution approach with those of the existing approaches such as the Breslow correction and the Efron correction. Our result shows that our approach has much lower biases and mean squared errors, as well as higher confidence interval coverage rates, than the existing methods for survival data with heavy ties or high-variation covariates. Our real data examples further confirm our findings in the theory and simulation.

In summary, our new PB distribution based approach to compute the partial likelihood accurately has the following distinct contributions: (1) we show that all the existing tie correction methods are approximations to our exact approach; (2) we derive the consistency and asymptotic normality of these methods under this unified framework, suggesting allowable order of ties, which has been lacking in the literature despite their popularity and long history; (3) we show that the proposed approach possesses the same asymptotic properties and demonstrate its clear numerical advantage in terms of reduced bias and mean squared error, along with enhanced confidence interval coverages.

The rest of the project is organized as follows. In Section 2.2, we introduce the Cox model and the PB distribution. In Section 2.3, we obtain the accurate partial likelihood for the Cox model using the PB distribution, and propose a new method to estimate the coefficients and the baseline hazard function for the Cox model. In Section 2.4, we develop statistical theory for the existing methods and our new method. In Section 2.5, we evaluate the performance of our method in simulated data and compare its performance with others. In Section 2.6, we analyze survival datasets with our new method and the existing ones. Lastly, in Section 2.7, we conclude the project with recommendations and some areas for future research.

2.2 The Cox Model

2.2.1 The Underlying Continuous Failure Time Model

In a continuous failure time model, the underlying event times $\tilde{T}_i \in (0, \infty)$, $i = 1, \ldots, n$, are from a continuous distribution. Here, $\tilde{T}_1, \ldots, \tilde{T}_n$ are assumed to be conditionally independent given the covariates $\{\boldsymbol{x}_i\}_{i=1}^n$. Recall the general form of the Cox model: for $i = 1, \ldots, n$,

$$d\Lambda(t; \boldsymbol{x}_i) = \Pr(\tilde{T}_i \in [t, t + dt) | \boldsymbol{x}_i, \tilde{T}_i \geq t) = \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta}) d\Lambda_0(t). \tag{2.3}$$

The cumulative hazard and survival functions are respectively $\Lambda(t; \boldsymbol{x}_i) = \int_0^t d\Lambda(s; \boldsymbol{x}_i) = \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta}) \Lambda_0(t)$ and $S(t; \boldsymbol{x}_i) = \Pr(\tilde{T}_i > t | \boldsymbol{x}_i) = \exp(-\Lambda(t; \boldsymbol{x}_i))$. Under the continuous failure time model, the baseline hazard function can be written as $d\Lambda_0(t) = \lambda_0(t) dt$. Let $\zeta \in (0, \infty)$ be the ending time of the study and $C_i \in (0, \zeta], i = 1, \ldots, n$, be non-informative right censoring times, which are conditionally independent from each other and from $\{\tilde{T}_i\}_{i=1}^n$ given the covariates $\{\boldsymbol{x}_i\}_{i=1}^n$. Assume that the covariates are deterministic, allowing the conditioning argument for $\{\boldsymbol{x}_i\}_{i=1}^n$ to be omitted for simplicity.

2.2.2 The Grouped Continuous Failure Time Model

In some scenarios, it is possible that the event times are rounded or grouped, leading to ties in the event times. Let $\lceil a \rceil$ be the smallest integer which is not smaller than a. Let $\tau \in (0, \infty)$ be a grouping parameter such that event times and censoring times are discretized as

$$\tilde{T}_i^* = \tau \lceil \tilde{T}_i / \tau \rceil \in \Omega_{\mathrm{G}} = \{\tau, 2\tau, \ldots\} \text{ and } C_i^* = \tau \lceil C_i / \tau \rceil \in \Omega = \{\tau, 2\tau, \ldots, \zeta\},$$

where we assume, without loss of generality, that the ending time ζ is a multiple of τ. The observed time is $T_i^* = \min(\tilde{T}_i^*, C_i^*) \in \Omega$, where $\delta_i = \mathbb{1}(\tilde{T}_i^* \leq C_i^*)$ is the indicator of an event. That is, when $\delta_i = 1$, the ith unit had the event and when $\delta_i = 0$, the ith unit was censored. As Ω_{G} is the support for grouped times, Ω is the support for the observed times, which is a finite subset of Ω_{G}. In this project, we assume ties are generated by the grouping of an underlying continuous random variable. From the discretized distribution, for $t \in \Omega_{\mathrm{G}}$, we have

$$\Pr(\tilde{T}_i^* \geq t | \boldsymbol{x}_i) = \Pr(\tilde{T}_i > t - \tau | \boldsymbol{x}_i) = S(t - \tau; \boldsymbol{x}_i), \tag{2.4}$$

$$\Pr(\tilde{T}_i^* \in [t, t + dt) | \boldsymbol{x}_i) = \Pr(\tilde{T}_i \in (t - \tau, t] | \boldsymbol{x}_i) = S(t - \tau; \boldsymbol{x}_i) - S(t; \boldsymbol{x}_i), \tag{2.5}$$

for $i = 1, \ldots, n$. From (2.4) and (2.5), the hazard function of $\tilde{T}_i^*$ is

$$\begin{aligned} d\Lambda^*(t; \boldsymbol{x}_i) &= \Pr(\tilde{T}_i^* \in [t, t + dt) | \boldsymbol{x}_i, \tilde{T}_i^* \geq t) = \frac{\Pr(\tilde{T}_i^* \in [t, t + dt) | \boldsymbol{x}_i)}{\Pr(\tilde{T}_i^* \geq t | \boldsymbol{x}_i)} \\ &= 1 - \exp\left(-\exp(\boldsymbol{x}_i^\top \boldsymbol{\beta}) d\Lambda_0^*(t)\right), \quad i = 1, \ldots, n, \end{aligned} \tag{2.6}$$

where the baseline hazard function is $d\Lambda_0^*(t) = \Lambda_0(t) - \Lambda_0(t - \tau)$ for $t \in \Omega_{\mathrm{G}}$, $d\Lambda_0^*(t) = 0$ for $t \notin \Omega_{\mathrm{G}}$, and the baseline cumulative hazard function is

$$\Lambda_0^*(t) = \int_0^t d\Lambda_0^*(s) = \sum_{s \leq t, s \in \Omega_{\mathrm{G}}} d\Lambda_0^*(s). \tag{2.7}$$

As $\Lambda_0(0) = 0$, we have $\Lambda_0^*(t) = \Lambda_0(t)$ for $t \in \Omega_{\mathrm{G}} \cup \{0\}$. The cumulative hazard function and survival function of $\tilde{T}_i^*$ are respectively $\Lambda^*(t; \boldsymbol{x}_i) = \int_0^t d\Lambda^*(s; \boldsymbol{x}_i) = \sum_{s \leq t, s \in \Omega_{\mathrm{G}}} d\Lambda^*(s; \boldsymbol{x}_i)$ and $S^*(t; \boldsymbol{x}_i) = \Pr(\tilde{T}_i^* > t | \boldsymbol{x}_i) = \prod_{s \leq t, s \in \Omega_{\mathrm{G}}} (1 - d\Lambda^*(s; \boldsymbol{x}_i)) = \exp(-\exp(\boldsymbol{x}_i^\top \boldsymbol{\beta})\ \Lambda_0^*(t))$. One can see that $S^*(t; \boldsymbol{x}_i) = S(t; \boldsymbol{x}_i)$ for $t \in \Omega_{\mathrm{G}} \cup \{0\}$.

2.2.3 The Fitted Model

In a typical semiparametric inference setting, the baseline cumulative hazard function is estimated as a step function, regardless of whether data are with grouping or not. Hence, the fitted model is essentially a discretized version of the continuous model in (2.3).

Let $t_{(j)}$, $j = 1, \ldots, k$, be the ordered distinct events times as defined in Section 2.1 and let $\Omega^{\mathcal{K}} = \{t : t = t_{(j)}, j = 1, \ldots, k\} \subseteq (0, \infty)$ be the set of distinct event times. It is worth reiterating no matter how data are generated (i.e., from a continuous model with or without grouping), the fitted Cox model is always a discretized model, because the estimated cumulative hazard function is a step function which has jumps at those time points in $\Omega^{\mathcal{K}}$. Under the grouped continuous failure time model, we have $\Omega^{\mathcal{K}} \subseteq \Omega \subseteq \Omega_{\mathrm{G}} \subseteq (0, \infty)$.

2.2.4 Poisson-Binomial Distribution

Recall that n_j is the number of subjects in the at-risk set $\mathcal{R}(t_{(j)})$ at time $t_{(j)}$, and d_j is the number of subjects in the event set $\mathcal{D}(t_{(j)})$ at time $t_{(j)}$. Here d_j can be larger than one so that ties are possible. For each $i \in \mathcal{R}(t_{(j)})$, given the history up to time $t_{(j)}$, the probability p_{ij} for subject i to have an event at time $t_{(j)}$ based on the fitted model is defined as $p_{ij} = p_{ij}(\boldsymbol{\beta}, \lambda_j) \equiv d\Lambda^*(t_{(j)}; \boldsymbol{x}_i)$ by (2.6), where λ_j is defined as $\lambda_j \equiv d\Lambda_0^*(t_{(j)})$ by (2.7). From now until the end of this section, all distributional and independence derivations are conditional on the history up to $t_{(j)}$, but for simplicity, we omit this conditioning and use abused notation. We can define a random indicator $I_{ij} \sim \text{Bernoulli}(p_{ij})$, where $\text{Bernoulli}(p_{ij})$ indicates a Bernoulli distribution with parameter p_{ij}. In this case, the number of events I_j at time $t_{(j)}$ can be written as $I_j = \sum_{i \in \mathcal{R}(t_{(j)})} I_{ij}$. Because p_{ij}'s are different, the distribution of I_j is not necessarily a binomial distribution. In general, I_j follows a PB distribution. The probability mass function of a PB distribution can be computed using enumeration, the method of discrete Fourier transform of characteristic function, an approximation method such as the Poisson approximation [21], or the convolution-based method [3].

The enumeration method computes the probability mass function of I_j as follows,

$$\Pr(I_j = d_j) = \sum_{\mathcal{A}_{d_j} \in \mathcal{F}_{d_j}} \left\{ \prod_{i \in \mathcal{A}_{d_j}} p_{ij} \right\} \left\{ \prod_{i \in \mathcal{R}(t_{(j)}) \setminus \mathcal{A}_{d_j}} (1 - p_{ij}) \right\}, \tag{2.8}$$

where $\mathcal{F}_{d_j}$ is the set of all subsets of d_j individuals that can be selected from $\mathcal{R}(t_{(j)})$. That is, each set $\mathcal{A}_{d_j} \in \mathcal{F}_{d_j}$ has d_j number of elements and $|\mathcal{F}_{d_j}| = \binom{n_j}{d_j}$. The method of discrete transform of characteristic function computes the probability mass function of I_j as $\Pr(I_j = d_j) = (n_j + 1)^{-1} \sum_{l=0}^{n_j} \exp(-\boldsymbol{i}\omega_j l d_j) z_{lj}$, where $\omega_j = 2\pi/(n_j + 1)$, $\boldsymbol{i} = \sqrt{-1}$, and $z_{lj} = \prod_{i \in \mathcal{R}(t_{(j)})} \{1 - p_{ij} + p_{ij} \exp(\boldsymbol{i}\omega_j l)\}$. The method can be done efficiently using the fast Fourier transform, and it is available in the R package "`poibin`" by Hong [22]. The Poisson approximation to the probability mass function of a PB distribution is,

$$\Pr(I_j = d_j) \approx \frac{\mu_j^{d_j} \exp(-\mu_j)}{d_j!}, \tag{2.9}$$

where $\mu_j = \sum_{i \in \mathcal{R}(t_{(j)})} p_{ij}$ is the mean of I_j.

2.3 Partial Likelihood and Parameter Estimation

2.3.1 The Original Idea of Partial Likelihood

Consider the general situation where ties could be present or not. The APL function is $L(\boldsymbol{\beta}, \boldsymbol{\Lambda}) = \prod_{j=1}^{k} L_j(\boldsymbol{\beta}, \lambda_j) = \prod_{j=1}^{k} A_j(\boldsymbol{\beta}, \lambda_j)/B_j(\boldsymbol{\beta}, \lambda_j)$, where $\boldsymbol{\Lambda} = (\lambda_1, \dots, \lambda_k)^\top$ and

$$L_j(\boldsymbol{\beta}, \lambda_j) = \frac{\Pr(\text{units } j_1, \dots, j_{d_j} \text{ had event at } t_{(j)} \big| \, n_j \text{ units survived up to } t_{(j)})}{\Pr(d_j \text{ out of } n_j \text{ units had event at } t_{(j)} \big| \, n_j \text{ units survived up to } t_{(j)})}. \tag{2.10}$$

Here $j_1, \dots, j_{d_j}$ are the d_j individuals in $\mathcal{D}(t_{(j)})$. The accurate calculation for $A_j(\boldsymbol{\beta}, \lambda_j)$ is

$$A_j(\boldsymbol{\beta}, \lambda_j) = \left\{ \prod_{i \in \mathcal{D}(t_{(j)})} p_{ij}(\boldsymbol{\beta}, \lambda_j) \right\} \left\{ \prod_{i \in \mathcal{R}(t_{(j)}) \setminus \mathcal{D}(t_{(j)})} (1 - p_{ij}(\boldsymbol{\beta}, \lambda_j)) \right\}. \tag{2.11}$$

The accurate calculation for $B_j(\boldsymbol{\beta}, \lambda_j)$ is

$$B_j(\boldsymbol{\beta}, \lambda_j) = \Pr(I_j = d_j), \tag{2.12}$$

which is the probability mass function of the PB distribution.

2.3.2 An Estimation Procedure Based on PB distributions

When there are no ties ($d_j = 1$), the expression (2.2) is used to approximate the APL [8]. In the presence of ties, various corrections can be applied, including the Cox correction [8], the Kalbfleisch-Prentice correction [26], the Breslow correction [4], and the Efron correction [13]. More details about these methods are in Section A.1. It is important to note that all of these existing methods rely on some approximations to both $A_j(\boldsymbol{\beta}, \lambda_j)$ and $B_j(\boldsymbol{\beta}, \lambda_j)$.

Once the probability mass function of the PB distribution is calculated, we can compute the accurate probability terms (2.10) in the partial likelihood function. However, the probability p_{ij} depends on $\mathbf{\Lambda}$. Let $\widehat{\mathbf{\Lambda}} = (\widehat{\lambda}_1, \ldots, \widehat{\lambda}_k)^\top$ be any estimate of $\mathbf{\Lambda}$, which satisfies a mild condition in Assumption A.3.9. We use $\widehat{\mathbf{\Lambda}}$ to substitute $\mathbf{\Lambda}$. Our proposed partial likelihood to be calculated based on PB distribution is

$$L(\boldsymbol{\beta}, \widehat{\mathbf{\Lambda}}) = \prod_{j=1}^{k} \frac{A_j(\boldsymbol{\beta}, \widehat{\lambda}_j)}{B_j(\boldsymbol{\beta}, \widehat{\lambda}_j)}, \tag{2.13}$$

where A_j and B_j are given in (2.11) and (2.12), respectively. Then the PB distribution estimate of $\boldsymbol{\beta}$ is

$$\widehat{\boldsymbol{\beta}}_{\text{pb}} = \arg\max_{\tilde{\boldsymbol{\beta}}} L(\tilde{\boldsymbol{\beta}}, \widehat{\mathbf{\Lambda}}), \tag{2.14}$$

which can be computed optimization algorithms such as the BFGS algorithm (e.g., Fletcher [17]). The new method uses the exact $A_j(\tilde{\boldsymbol{\beta}}, \widehat{\lambda}_j)$ and $B_j(\tilde{\boldsymbol{\beta}}, \widehat{\lambda}_j)$ for a given $\tilde{\boldsymbol{\beta}}$. Hence, $\widehat{\boldsymbol{\beta}}_{\text{pb}}$ is the maximizer of a more accurate partial likelihood.

The optimization of (2.14) requires initial estimates $\widehat{\mathbf{\Lambda}}$ and $\widehat{\boldsymbol{\beta}}$ whose choices are flexible. For example, one can use the Efron baseline hazard function estimate in (A.5), $\widehat{\mathbf{\Lambda}}_{\text{e}} = (\widehat{\lambda}_{\text{e}1}, \ldots, \widehat{\lambda}_{\text{e}k})^\top$, the Breslow baseline hazard function estimate in (A.4), $\widehat{\mathbf{\Lambda}}_{\text{b}} = (\widehat{\lambda}_{\text{b}1}, \ldots, \widehat{\lambda}_{\text{b}k})^\top$, or even the Nelson-Aalen baseline hazard function estimate, $\widehat{\mathbf{\Lambda}}_{\text{na}} = (\widehat{\lambda}_{\text{na},1}, \ldots, \widehat{\lambda}_{\text{na},k})^\top$ as $\widehat{\mathbf{\Lambda}}$, where $\widehat{\lambda}_{\text{na},j} = d_j/n_j$. Our numerical example in Section A.6.2 suggests choosing $\widehat{\mathbf{\Lambda}}_{\text{e}}$ yields a smaller bias comparing to the other methods. So we use $\widehat{\mathbf{\Lambda}}_{\text{e}}$ as $\widehat{\mathbf{\Lambda}}$ in our numerical examples in Section 2.5 and 2.6. For $\widehat{\boldsymbol{\beta}}$, one can use Efron estimator $\widehat{\boldsymbol{\beta}}_{\text{e}}$ in (A.5) or Breslow estimator $\widehat{\boldsymbol{\beta}}_{\text{b}}$ in (A.4). We use $\widehat{\boldsymbol{\beta}}_{\text{e}}$ in our numerical examples in Sections 2.5 and 2.6.

Next, we estimate the baseline hazard function. The likelihood for the baseline hazard function is $\prod_{j=1}^{k} A_j(\boldsymbol{\beta}, \lambda_j)$ as shown in page 115 of Kalbfleisch and Prentice [27]. So, by the

BFGS algorithm with the initial values $\widehat{\lambda}_j$, the baseline hazard function can be updated as $\widehat{\lambda}_{\mathrm{pb},j} = \arg\max_{\tilde{\lambda}_j} A_j(\widehat{\boldsymbol{\beta}}_{\mathrm{pb}}, \tilde{\lambda}_j)$, $j = 1, \ldots, k$, where the fitted cumulative hazard function is $\widehat{\Lambda}_{\mathrm{pb}}(t) = \sum_{j=1}^{k} \widehat{\lambda}_{\mathrm{pb},j} \mathbb{1}(t_{(j)} \leq t)$.

2.4 Statistical Properties

2.4.1 Connections to PB Distribution

We first draw some connections between the PB distribution probability and existing partial likelihood approximation methods. Interestingly, all the existing methods are connected to the PB distribution approach, and our results shed theoretical insights on when the existing computing methods tend to work well.

Theorem 2.1. *(i) the approximate partial likelihood (A.1) is based on the Poisson approximation of* $\Pr(I_j = 1)$ *to* $B_j(\boldsymbol{\beta}, \lambda_j)$*, (ii) the Breslow correction in (A.4) based on the Poisson approximation of* $\Pr(I_j = d_j)$ *to* $B_j(\boldsymbol{\beta}, \lambda_j)$*, (iii) the Cox correction in (A.2) is based on the enumeration in (2.8) to compute the* $B_j(\boldsymbol{\beta}, \lambda_j)$*, and (iv) the Kalbfleisch-Prentice correction in (A.3) is based on the enumeration to compute* $A_j(\boldsymbol{\beta}, \lambda_j)/B_j(\boldsymbol{\beta}, \lambda_j)$*.*

The proof of Theorem 2.1 is in Section A.4.1. As a note, the Efron correction in (A.5) is of a similar form to the Breslow correction but with some further adjustments. Thus, the Efron correction can be viewed as a more refined Poisson approximation to the PB distribution probability than the Breslow correction. We provide some insights into the accuracy of their Poisson approximations in the following remark.

Remark 2.2. The error bound of the Poisson approximation to the PB distribution can be obtained by the Le Cam theorem [5]. The average error is bounded as

$$\frac{1}{n_j}\sum_{d_j=0}^{n_j}\left|\Pr(I_j=d_j)-\frac{\mu_j^{d_j}\exp(-\mu_j)}{d_j!}\right| \le \frac{2}{n_j}\sum_{i\in\mathcal{R}(t_{(j)})} p_{ij}^2 = \frac{2\sum_{i\in\mathcal{R}(t_{(j)})}(p_{ij}-\bar{p}_j)^2}{n_j} + 2\bar{p}_j^2.$$

Here $\bar{p}_j = \sum_{i\in\mathcal{R}(t_{(j)})} p_{ij}/n_j$ and μ_j is defined in (2.9). Thus, the performance of the Breslow and Efron estimators depends on the average $\bar{p}_j$ and the variation $\sum_{i\in\mathcal{R}(t_{(j)})}(p_{ij}-\bar{p}_j)^2/n_j$ of the risk scores $r_i = \exp(\boldsymbol{x}_i^\top\boldsymbol{\beta})$. The average and variation of risk scores depend on the values of $\boldsymbol{\beta}$ and the distribution of the covariates $\boldsymbol{x}_i$. Thus both the scale of $\boldsymbol{\beta}$ and the variation in $\boldsymbol{x}_i$'s can affect the approximation accuracy.

2.4.2 Asymptotics under the Grouped Continuous Failure Time Model

Now we study the large sample properties of the estimators of $\boldsymbol{\beta}$ under the grouped continuous failure time model where event times can be tied. Among the estimators from the existing methods, we select the Breslow estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{b}}$ to show its consistency and asymptotic normality, with the notion that the properties for the other estimators can be derived similarly, given our result in Theorem 2.1. Then we establish similar large sample properties for the PB distribution estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{pb}}$ based on (2.14).

In this section, we restrict $t \in [0,\zeta]$. We denote by $\mathcal{D}(t) \equiv \{i : T_i^* = t \text{ and } \delta_i = 1\}$ and $\mathcal{R}(t) \equiv \{i : T_i^* \ge t\}$ respectively the event and at-risk sets at time t. Let $d(t) \equiv |\mathcal{D}(t)|$ and $n(t) \equiv |\mathcal{R}(t)|$. Let $\Omega_t = \Omega \cap [0,t]$ for some $t \in [0,\zeta]$. We define $\Delta H(t) = H(t) - H(t^-)$ for a function $H(\cdot)$. Let the PB distribution partial likelihood component at t be

$$L_t(\boldsymbol{\beta}) = \frac{A_t(\boldsymbol{\beta})}{B_t(\boldsymbol{\beta})} = \frac{\left\{\prod_{i\in\mathcal{D}(t)} p_i(\boldsymbol{\beta},t)\right\}\left\{\prod_{i\in\mathcal{R}(t)\setminus\mathcal{D}(t)}(1-p_i(\boldsymbol{\beta},t))\right\}}{\sum_{\mathcal{A}_{d(t)}\in\mathcal{F}_{d(t)}}\left\{\prod_{i\in\mathcal{A}_{d(t)}} p_i(\boldsymbol{\beta},t)\right\}\left\{\prod_{i\in\mathcal{R}(t)\setminus\mathcal{A}_{d(t)}}(1-p_i(\boldsymbol{\beta},t))\right\}},$$

where $p_i(\boldsymbol{\beta}, t) = 1 - \exp(-\exp(\boldsymbol{x}_i^\top \boldsymbol{\beta})\Delta\widehat{\Lambda}(t))$. Let the Breslow partial likelihood component at t be

$$L_t^{\mathrm{b}}(\boldsymbol{\beta}) = \frac{\exp\left(\sum_{i\in\mathcal{D}(t)} \boldsymbol{x}_i^\top \boldsymbol{\beta}\right)}{\left\{\sum_{i\in\mathcal{R}(t)} \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta})\right\}^{d(t)} / d(t)!}.$$

Here $\widehat{\Lambda}(t) = \sum_{j=1}^{k} \widehat{\lambda}_j \mathbb{1}(t_{(j)} \le t) = \int_0^t d\widehat{\Lambda}(s) = \sum_{s\in\Omega_t} \Delta\widehat{\Lambda}(s)$ stands for any estimator of $\Lambda_0^*(t)$, which corresponds to $\widehat{\boldsymbol{\Lambda}}$ in (2.13). We write the PB distribution partial likelihood in (2.13) at $t \in [0, \zeta]$ as $L(\boldsymbol{\beta}, \widehat{\boldsymbol{\Lambda}}, t) = \prod_{s\in\Omega_t} L_s(\boldsymbol{\beta})$ and the Breslow partial likelihood in (A.4) at $t \in [0, \zeta]$ as $L^{\mathrm{b}}(\boldsymbol{\beta}, t) = \prod_{s\in\Omega_t} L_s^{\mathrm{b}}(\boldsymbol{\beta})$, where if $t \notin \Omega^{\mathcal{K}}$ (i.e., $d(t) = 0$), we have $\Delta\widehat{\Lambda}(t) = 0$, $p_i(\boldsymbol{\beta}, t) = 0$, and $L_t^{\mathrm{b}}(\boldsymbol{\beta}) = L_t(\boldsymbol{\beta}) = 1$.

Let $\mathcal{B} \subseteq \mathbb{R}^d$ be an open neighborhood of $\boldsymbol{\beta}$. Without loss of generality, we assume $\mathcal{B}$ is sufficiently large and $\widehat{\boldsymbol{\beta}}_{\mathrm{b}}, \widehat{\boldsymbol{\beta}}_{\mathrm{pb}} \in \mathcal{B}$, resulting in $\widehat{\boldsymbol{\beta}}_{\mathrm{b}} \equiv \arg\max_{\tilde{\boldsymbol{\beta}}\in\mathbb{R}^d} L^{\mathrm{b}}(\tilde{\boldsymbol{\beta}}, \zeta) = \arg\max_{\tilde{\boldsymbol{\beta}}\in\mathcal{B}} L^{\mathrm{b}}(\tilde{\boldsymbol{\beta}}, \zeta)$ and $\widehat{\boldsymbol{\beta}}_{\mathrm{pb}} \equiv \arg\max_{\tilde{\boldsymbol{\beta}}\in\mathbb{R}^d} L(\tilde{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Lambda}}, \zeta) = \arg\max_{\tilde{\boldsymbol{\beta}}\in\mathcal{B}} L(\tilde{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Lambda}}, \zeta)$.

For the counting processes, let $\tilde{N}_i(t) = \mathbb{1}(\tilde{T}_i^* \le t)$, $N_i(t) = \mathbb{1}(T_i^* \le t, \delta_i = 1) = \mathbb{1}(\tilde{T}_i^* \le t, \tilde{T}_i^* \le C_i^*)$, and $Y_i(t) = \mathbb{1}(T_i^* \ge t) = \mathbb{1}(\tilde{T}_i^* \ge t, C_i^* \ge t)$ be the underlying counting process, the observed counting process, and the at-risk process for the grouped continuous failure time model, respectively. Let the history (filtration) for the grouped continuous failure time model be

$$\mathcal{F}_t = \sigma\left\{N_i(u), Y_i(u^+), \boldsymbol{x}_i; i = 1, \dots, n; 0 \le u \le t\right\} \text{ and}$$
$$\mathcal{F}_{t^-} = \sigma\left\{N_i(u), Y_i(u), \boldsymbol{x}_i; i = 1, \dots, n; 0 \le u < t\right\}.$$

Note that $dN_i(t) = Y_i(t)d\tilde{N}_i(t)$. Thus, one can show that $\mathrm{E}(dN_i(t)|\mathcal{F}_{t^-}) = \Pr(dN_i(t) = 1|\mathcal{F}_{t^-}) = Y_i(t)d\Lambda^*(t; \boldsymbol{x}_i)$ due to our assumption that $\{\tilde{T}_i^*\}_{i=1}^n$ and $\{C_i^*\}_{i=1}^n$ are conditionally independent given $\{\boldsymbol{x}_i\}_{i=1}^n$. We refer to equation (5.7) from the Kalbfleisch and Prentice [27] for more details. So by the Doob-Meyer decomposition, $N_i(t) = A_i(t) + M_i(t)$, where $A_i(t) =$

$\int_0^t Y_i(s)d\Lambda^*(s;\boldsymbol{x}_i) = \sum_{s\in\Omega_t} Y_i(s)\Delta\Lambda^*(s;\boldsymbol{x}_i)$ is the compensator and $M_i(t) = N_i(t) - A_i(t)$ is a zero-mean martingale. Let $N(t) = \sum_{i=1}^n N_i(t)$, $A(t) = \sum_{i=1}^n A_i(t)$, and $M(t) = \sum_{i=1}^n M_i(t)$. Again, we refer to Sections 5.2 and 5.3 of Kalbfleisch and Prentice [27] for more details on the construction of counting processes and martingales.

Although the time points in Ω all depend on the grouping parameter τ, for the simplicity of notation, we don't include τ into the subscripts. Every process or statement depending on a time point in Ω also depends on τ implicitly. Furthermore, one can see that $\tilde{T}_i^*$, C_i^*, T_i^*, and all their related functions including $\tilde{N}_i(t)$, $N_i(t)$, and $Y_i(t)$ depend on τ. All our asymptotic notations and technical assumptions are collected in Sections A.2 and A.3, respectively.

For $t \in [0, \zeta]$, the logarithm of the Breslow partial likelihood is

$$\begin{aligned}\log\left(L^{\mathrm{b}}(\boldsymbol{\beta},t)\right) &= \sum_{i=1}^n \int_0^t \left\{\boldsymbol{x}_i^\top\boldsymbol{\beta} - \log\left(\sum_{l=1}^n Y_l(s)\exp(\boldsymbol{x}_l^\top\boldsymbol{\beta})\right)\right\} dN_i(s) + \sum_{s\in\Omega_t}\log(\Delta N(s)!)\\ &= \sum_{i=1}^n \sum_{s\in\Omega_t} \left\{\boldsymbol{x}_i^\top\boldsymbol{\beta} - \log\left(\sum_{l=1}^n Y_l(s)\exp(\boldsymbol{x}_l^\top\boldsymbol{\beta})\right)\right\} \Delta N_i(s) + \sum_{s\in\Omega_t}\log(\Delta N(s)!).\end{aligned}$$

The score function is

$$U_{\mathrm{b}}(\boldsymbol{\beta},t) = \frac{\partial}{\partial\boldsymbol{\beta}}\log(L^{\mathrm{b}}(\boldsymbol{\beta},t)) = \sum_{i=1}^n \int_0^t (\boldsymbol{x}_i - \epsilon(\boldsymbol{\beta},s))\, dN_i(s) = \sum_{i=1}^n \sum_{s\in\Omega_t} (\boldsymbol{x}_i - \epsilon(\boldsymbol{\beta},s))\,\Delta N_i(s),$$

where $\epsilon(\boldsymbol{\beta},t) = \sum_{i=1}^n q_i(\boldsymbol{\beta},t)\boldsymbol{x}_i$ with $q_i(\boldsymbol{\beta},t) = Y_i(t)\exp(\boldsymbol{x}_i^\top\boldsymbol{\beta})/(\sum_{l=1}^n Y_l(t)\exp(\boldsymbol{x}_l^\top\boldsymbol{\beta}))$, which satisfies $\sum_{i=1}^n q_i(\boldsymbol{\beta},t) = 1$. The information matrix is

$$I_{\mathrm{b}}(\boldsymbol{\beta},t) = -\frac{\partial^2}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^\top}\log(L^{\mathrm{b}}(\boldsymbol{\beta},t)) = \int_0^t \mathcal{V}(\boldsymbol{\beta},s)dN(s) = \sum_{s\in\Omega_t}\mathcal{V}(\boldsymbol{\beta},s)\Delta N(s),$$

where $\mathcal{V}(\boldsymbol{\beta},t) = \sum_{i=1}^n q_i(\boldsymbol{\beta},t)(\boldsymbol{x}_i - \epsilon(\boldsymbol{\beta},t))(\boldsymbol{x}_i - \epsilon(\boldsymbol{\beta},t))^\top$. Let $\Sigma(\boldsymbol{\beta},\zeta)$ be the matrix defined in Assumption A.3.7. Our asymptotic results for $\widehat{\boldsymbol{\beta}}_{\mathrm{b}}$ are

Theorem 2.3. *Under Assumptions A.3.1 – A.3.7,* $\widehat{\boldsymbol{\beta}}_b$ *converges in probability to* $\boldsymbol{\beta}$ *and* $I_b(\widehat{\boldsymbol{\beta}}_b, \zeta)/n$ *converges in probability to* $\Sigma(\boldsymbol{\beta}, \zeta)$ *when* $\tau \to 0$ *as* $n \to \infty$.

Theorem 2.4. *Under Assumptions A.3.1 – A.3.7,* $n^{-1/2} U_b(\boldsymbol{\beta}, \zeta)$ *converges in distribution to* $N(\mathbf{0}, \Sigma(\boldsymbol{\beta}, \zeta))$ *and* $n^{1/2}(\widehat{\boldsymbol{\beta}}_b - \boldsymbol{\beta})$ *converges in distribution to* $N(\mathbf{0}, \Sigma(\boldsymbol{\beta}, \zeta)^{-1})$ *when* $n^{1/2}\tau \to 0$ *as* $n \to \infty$.

The proofs of Theorems 2.3 and 2.4 are respectively in Sections A.4.6 and A.4.7. As we have $\sup_{t \in \Omega} d(t) = \mathcal{O}_{\mathrm{P}}(n\tau)$ for $t \in \Omega$ under Assumption A.3.5, if the order of τ is no smaller than $1/n$ and $\tau \to 0$, the Breslow method can achieve consistency allowing ties. If the order of τ is no smaller than $1/n$ and $n^{1/2}\tau \to 0$, the Breslow estimator have the asymptotic normality allowing ties.

Now we derive the asymptotic properties of the PB distribution estimator with three more assumptions, A.3.8 – A.3.10. In particular, Assumption A.3.8 requires τ to be of the order $1/n$. The following theorem shows the asymptotic equivalence between $L(\tilde{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Lambda}}, t)$ and $L^{\mathrm{b}}(\tilde{\boldsymbol{\beta}}, t)$ uniformly for $\tilde{\boldsymbol{\beta}} \in \mathcal{B}$.

Theorem 2.5. *When Assumptions A.3.1 – A.3.5 and A.3.8 – A.3.9 are satisfied, for all* $t \in [0, \zeta]$,

$$\sup_{\tilde{\boldsymbol{\beta}} \in \mathcal{B}} \left| \log\left(L(\tilde{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Lambda}}, t) \right) - \log\left(L^b(\tilde{\boldsymbol{\beta}}, t) \right) \right| = \mathcal{O}_P(1).$$

Let $X_{\mathrm{pb}}(\tilde{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Lambda}}, t) = \{\log(L(\tilde{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Lambda}}, t)) - \log(L(\boldsymbol{\beta}, \widehat{\boldsymbol{\Lambda}}, t))\}/n$ and $X(\tilde{\boldsymbol{\beta}}, t) = \{\log(L^{\mathrm{b}}(\tilde{\boldsymbol{\beta}}, t)) - \log(L^{\mathrm{b}}(\boldsymbol{\beta}, t))\}/n$. By Theorem 2.5, when Assumptions A.3.1 – A.3.5 and A.3.8 – A.3.9 are satisfied, we have $\sup_{\tilde{\boldsymbol{\beta}} \in \mathcal{B}} |X_{\mathrm{pb}}(\tilde{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Lambda}}, t) - X(\tilde{\boldsymbol{\beta}}, t)| = \mathcal{O}_{\mathrm{P}}(1/n)$ for all $t \in [0, \zeta]$, where $\widehat{\boldsymbol{\beta}}_{\mathrm{pb}} = \arg\max_{\tilde{\boldsymbol{\beta}} \in \mathcal{B}} X_{\mathrm{pb}}(\tilde{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Lambda}}, \zeta)$ and $\widehat{\boldsymbol{\beta}}_{\mathrm{b}} = \arg\max_{\tilde{\boldsymbol{\beta}} \in \mathcal{B}} X(\tilde{\boldsymbol{\beta}}, \zeta)$. Based on closeness between PB distribution and Breslow partial likelihoods by Theorem 2.5, the asymptotic results for

$\widehat{\boldsymbol{\beta}}_{\mathrm{pb}}$ are

Theorem 2.6. *Under Assumptions A.3.1 – A.3.9,* $\widehat{\boldsymbol{\beta}}_{pb}$ *converges in probability to* $\boldsymbol{\beta}$ *and* $I_b(\widehat{\boldsymbol{\beta}}_{pb}, \zeta)/n$ *converges in probability to* $\Sigma(\boldsymbol{\beta}, \zeta)$ *as* $n \to \infty$.

Theorem 2.7. *Under Assumptions A.3.1 – A.3.10, the quantity* $n^{1/2}(\widehat{\boldsymbol{\beta}}_{pb} - \boldsymbol{\beta})$ *converges in distribution to* $N(\mathbf{0}, \Sigma(\boldsymbol{\beta}, \zeta)^{-1})$ *as* $n \to \infty$.

The proofs of Theorems 2.6 and 2.7 are respectively in Sections A.4.10 and A.4.11. Even though the estimators are all asymptotically unbiased, the score function based on the Breslow and Efron methods do not have a zero mean due to the Poisson approximation under small samples. However, the accurate likelihood calculation based on the PB distribution provides the exact probability under any sample size. Thus, the score function always has a zero mean using the PB distribution calculation if the true $\boldsymbol{\Lambda}$ is used. That is, although the new method uses $\widehat{\boldsymbol{\Lambda}}$ instead of $\boldsymbol{\Lambda}$, it tends to have less bias even under small samples. We indeed observe a smaller bias for the new estimator in small samples from the simulation studies in Section 2.5. Because of the variation caused by $\widehat{\boldsymbol{\Lambda}}$, the PB distribution estimator tends to have a slightly larger variance in practice. We also observe this in the simulation studies. However, the difference is minimal.

2.4.3 Asymptotics Under Continuous Model

In this section we show that the PB distribution estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{pb}}$ based on (2.14) still possesses excellent asymptotic properties even when observations are generated from a continuous model and contain no ties. Let $L_j^{\mathrm{c}}(\tilde{\boldsymbol{\beta}}) = \exp(\boldsymbol{x}_{j1}^\top \tilde{\boldsymbol{\beta}}) \,/\{\sum_{i \in \mathcal{R}(t_{(j)})} \exp(\boldsymbol{x}_i^\top \tilde{\boldsymbol{\beta}})\}$, $\mathcal{L}_{\mathrm{c}}(\tilde{\boldsymbol{\beta}}) = \sum_{j=1}^{k} \log(L_j^{\mathrm{c}}(\tilde{\boldsymbol{\beta}}))$, $U_{\mathrm{c}}(\tilde{\boldsymbol{\beta}}, \zeta) = \partial \mathcal{L}_{\mathrm{c}}(\tilde{\boldsymbol{\beta}})/(\partial \tilde{\boldsymbol{\beta}})$, and $I_{\mathrm{c}}(\tilde{\boldsymbol{\beta}}, \zeta) = -\partial^2 \mathcal{L}_{\mathrm{c}}(\tilde{\boldsymbol{\beta}})/(\partial \tilde{\boldsymbol{\beta}} \partial \tilde{\boldsymbol{\beta}}^\top)$ be respectively the log partial likelihood, the score function, and the information matrix for the approximate partial likelihood in (2.2). Let $\widehat{\boldsymbol{\beta}}_{\mathrm{c}} \equiv \arg\max_{\tilde{\boldsymbol{\beta}} \in \mathbb{R}^d} \mathcal{L}_{\mathrm{c}}(\tilde{\boldsymbol{\beta}})$ the estimator based on

(2.2). Without loss of generality, assume $\widehat{\boldsymbol{\beta}}_{\mathrm{c}} \in \mathcal{B}$. So we can write $\widehat{\boldsymbol{\beta}}_{\mathrm{c}}$ as $\arg\max_{\tilde{\boldsymbol{\beta}}\in\mathcal{B}} \mathcal{L}_{\mathrm{c}}(\tilde{\boldsymbol{\beta}})$. Let $\mathcal{L}(\tilde{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Lambda}}) = \log(L(\tilde{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Lambda}}))$, $\chi_{\mathrm{pb}}(\tilde{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Lambda}}) = \{\mathcal{L}(\tilde{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Lambda}}) - \mathcal{L}(\boldsymbol{\beta}, \widehat{\boldsymbol{\Lambda}})\}/n$, and $\chi(\tilde{\boldsymbol{\beta}}) = \{\mathcal{L}_{\mathrm{c}}(\tilde{\boldsymbol{\beta}}) - \mathcal{L}_{\mathrm{c}}(\boldsymbol{\beta})\}/n$. One can see that the estimators satisfy $\widehat{\boldsymbol{\beta}}_{\mathrm{pb}} = \arg\max_{\tilde{\boldsymbol{\beta}}\in\mathcal{B}} \chi_{\mathrm{pb}}(\tilde{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Lambda}})$ and $\widehat{\boldsymbol{\beta}}_{\mathrm{c}} = \arg\max_{\tilde{\boldsymbol{\beta}}\in\mathcal{B}} \chi(\tilde{\boldsymbol{\beta}})$. Let $\Sigma_{\mathrm{c}}(\boldsymbol{\beta}, \zeta)$ be the matrix defined in Assumption A.3.11. The asymptotic results for the PB distribution estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{pb}}$ under the continuous model without ties are

Theorem 2.8. *Under Assumptions A.3.2 and A.3.11, $\widehat{\boldsymbol{\beta}}_{pb}$ converges in probability to $\boldsymbol{\beta}$ and $I_c(\widehat{\boldsymbol{\beta}}_{pb}, \zeta)/n$ converges in probability to $\Sigma_c(\boldsymbol{\beta}, \zeta)$ as $n \to \infty$.*

Theorem 2.9. *Under Assumptions A.3.2 and A.3.11 – A.3.12, $n^{1/2}(\widehat{\boldsymbol{\beta}}_{pb} - \boldsymbol{\beta})$ converges in distribution to $N(\mathbf{0}, \Sigma_c(\boldsymbol{\beta}, \zeta)^{-1})$ as $n \to \infty$.*

The proofs of Theorems 2.8 and 2.9, similar to those of Theorems 2.6 and 2.7, are in Section A.4.12.

2.5 Simulation Studies

Our simulation studies are done in settings with a single covariate x_i, using scalar notation for simplicity, but it can be readily extended to multiple covariates. Overall, we follow the settings for the underlying continuous model described in Section 2.2.1. For the covariate, we generate n i.i.d. x_i's from $\mathrm{N}(0, \sigma_x^2)$. For event times, we use the Weibull distribution with the baseline hazard function $\lambda_0(t) = \gamma t^{\gamma-1}/\eta^{\gamma}$ such that $\lambda(t; x_i) = \gamma\eta^{-\gamma}t^{\gamma-1}\exp(x_i\beta) = \gamma\eta_i^{-\gamma}t^{\gamma-1}$, where $\eta_i = \eta\exp(-x_i\beta/\gamma)$. In particular, we generate $\tilde{T}_i$ as $\tilde{T}_i = \exp(\mu_i + \sigma W_i)$, where $\sigma = 1/\gamma$, $\mu_i = \log(\eta_i)$, and W_i's are n i.i.d. following the standard smallest extreme value distribution. For censoring times, we first generate n i.i.d. $\tilde{C}_i$'s from the Weibull distribution with hazard function $\lambda_c(t) = \gamma_c t^{\gamma_c-1}/\eta_c^{\gamma_c}$ and then set the censoring times as $C_i = \min\{\tilde{C}_i, \zeta\}$. To generate ties by grouping, we use the grouped continuous failure time model rule described in Section 2.2.2 with several τ values. We fix $\zeta = 1$, $\eta = \eta_c = 1.31$ and $\gamma = \gamma_c = 1.5$, and repeat the simulations $B = 10,000$ times for all the simulation cases. We vary $\beta \in \{1, 1.5\}$, $\sigma_x \in \{1.5, 2\}$, $\tau \in \{0.01, 0.1, 0.2\}$, and $n \in \{50, 100, 200, 500, 1000\}$.

For estimation performance metrics, we use the root mean square error (RMSE) and the absolute bias (|Bias|), each scaled by the true β, to compare our PB distribution method with the Breslow method and the Efron method. To evaluate inference performance, we compare the empirical coverage rates of confidence intervals and the average standard errors for the three methods with $n \in \{50, 100, 200\}$. The standard errors are calculated using respectively $I_{\mathrm{b}}^{-1/2}(\widehat{\beta}_{\mathrm{b}}, \zeta)$ for the Breslow method, $I_{\mathrm{e}}^{-1/2}(\widehat{\beta}_{\mathrm{e}}, \zeta)$ for the Efron method, and $I_{\mathrm{b}}^{-1/2}(\widehat{\beta}_{\mathrm{pb}}, \zeta)$ for the PB distribution method.

Figure 2.1 illustrate that for many cases, $\widehat{\beta}_{\mathrm{pb}}$ exhibits notably lower |Bias| and smaller RMSE compared to existing methods. On the other hand, the standard deviations of the estimators from the three methods are similar, as shown in Figure A.1. This indicates that $\widehat{\beta}_{\mathrm{pb}}$'s better RMSE performance primarily arises from its significant reduction in |Bias|. To be more specific, when β, σ_x^2, and τ are all small, all three methods give comparable performance. When at least one of these model parameters gets bigger, the Breslow method's performance starts to deteriorate. The Efron method can maintain a performance competitive to the PB distribution method until at least two of these model parameters get bigger. Besides its dominance in the cases of larger τ, σ_x^2, or β, the PB distribution estimators can deliver a competitive performance in all the other cases too.

Table 2.1 provides empirical coverage rates of confidence intervals and average standard errors. Clearly, the three methods deliver comparable standard errors. In terms of the empirical coverage, the message is similar to that of the esimation performance. When all the three model parameters, τ, β, and covariate variation, are small, the three method all perform well with coverage rates close to the nominal. When at least one of the three model parameters gets larger, the coverage rates for the PB distribution method dominate those for the other methods.

Overall, our method outperforms the Breslow and Efron methods in both parameter estimation and confidence interval coverages when at least one of the three model parameters, τ, β, or covariate variation, gets larger. Intuitively, an increase in σ_x^2 or β leads to an increase in the variation among p_{ij}'s. As discussed in Remark 2.2, an increased level of variation among p_{ij}'s leads to a larger value of $\sum_{i \in \mathcal{R}(t_{(j)})} p_{ij}^2/n_j$, which can also be enlarged by increment in τ considering the fact that $\sup_{ij} p_{ij} = \mathcal{O}_{\mathrm{P}}(\tau)$. This makes the Poisson approximation in the Breslow partial likelihood (A.4) and the Efron partial likelihood (A.5) less precise, and thus reduces the accuracy of $\widehat{\beta}_{\mathrm{b}}$ and $\widehat{\beta}_{\mathrm{e}}$. Additionally, note that Theorem 2.5 requires Assumption A.3.8, which is $\tau \asymp 1/n$. Or more simply, note that $\sup_{ij} |p_{ij} - \exp(x_i\beta)\lambda_j| = \mathcal{O}_{\mathrm{P}}(\tau^2)$. Thus, an increment in τ makes (A.4) and (A.5) less precise, and thus less accurate $\widehat{\beta}_{\mathrm{b}}$ and $\widehat{\beta}_{\mathrm{e}}$.

As a side note, the average computing time for our method with $\beta = 1.5$, $\sigma_x = 1.5$, $\tau = 0.2$, and $n = 200$ is approximately 0.028 seconds, on a MacBook with an 8-core Apple M1 chip and 17.2 GB RAM, echoing its computational efficiency demonstrated in Hong [21].

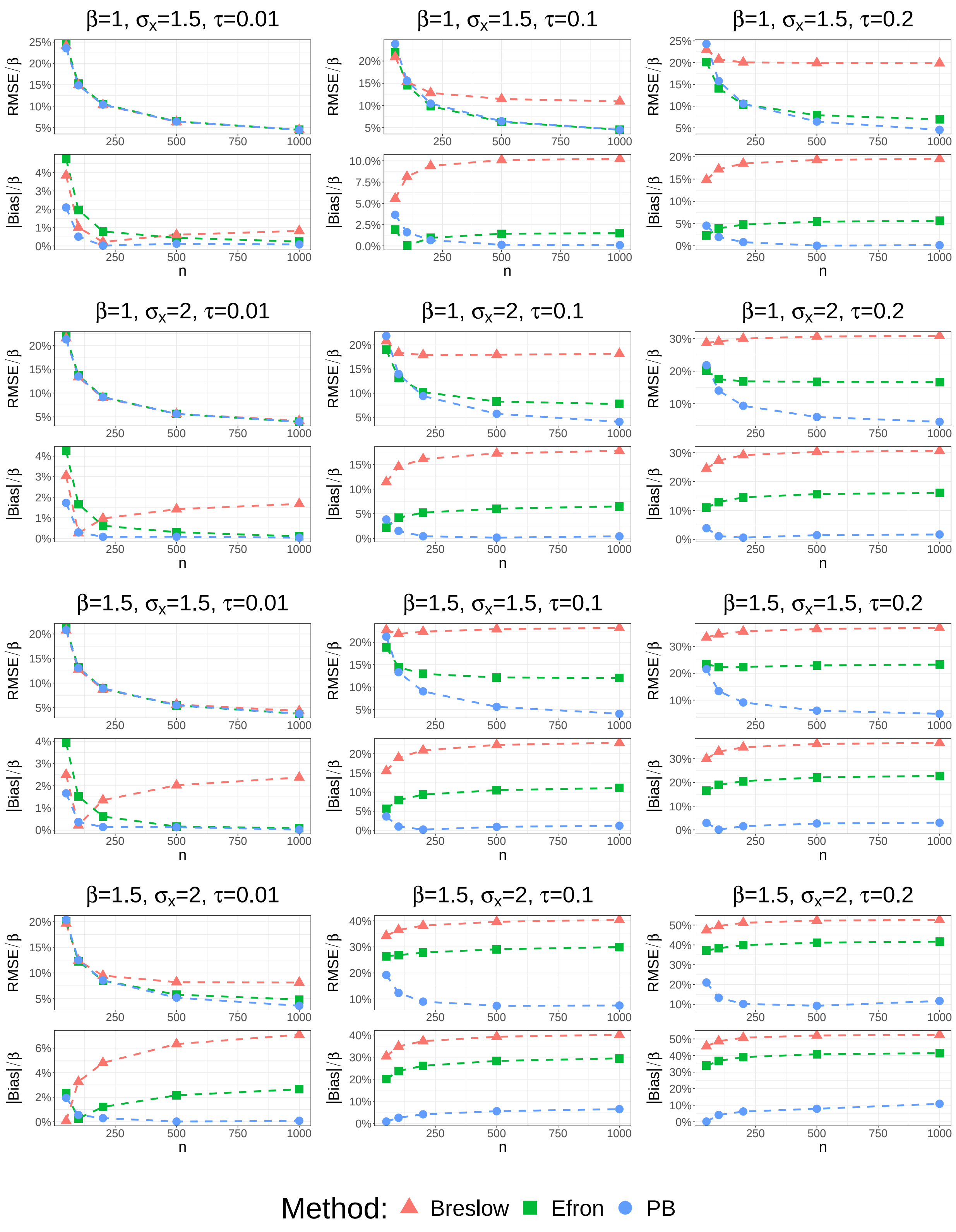

Figure 2.1: Simulation results for RMSE and |Bias|.

Table 2.1: Simulation results for inference. Each cell shows the empirical coverage rate of the confidence interval and the average estimated standard error (in parentheses).

β	σ_x	τ	Method	n		
				50	100	200
1	1.5	0.01	Breslow	0.959(0.226)	0.956(0.150)	0.952(0.102)
			Efron	0.958(0.226)	0.954(0.150)	0.952(0.103)
			PB	0.957(0.224)	0.956(0.149)	0.952(0.103)
		0.1	Breslow	0.939(0.211)	0.903(0.140)	0.828(0.096)
			Efron	0.959(0.217)	0.951(0.145)	0.950(0.100)
			PB	0.953(0.219)	0.939(0.145)	0.942(0.099)
		0.2	Breslow	0.857(0.196)	0.713(0.131)	0.441(0.090)
			Efron	0.949(0.205)	0.934(0.138)	0.908(0.095)
			PB	0.941(0.213)	0.932(0.140)	0.922(0.094)
	2	0.01	Breslow	0.958(0.202)	0.952(0.132)	0.948(0.090)
			Efron	0.958(0.203)	0.953(0.133)	0.954(0.091)
			PB	0.954(0.200)	0.951(0.132)	0.950(0.090)
		0.1	Breslow	0.856(0.179)	0.715(0.117)	0.460(0.079)
			Efron	0.936(0.185)	0.910(0.123)	0.864(0.084)
			PB	0.950(0.197)	0.937(0.128)	0.931(0.086)
		0.2	Breslow	0.592(0.158)	0.282(0.104)	0.041(0.071)
			Efron	0.835(0.166)	0.712(0.110)	0.498(0.075)
			PB	0.942(0.190)	0.920(0.121)	0.907(0.080)
1.5	1.5	0.01	Breslow	0.955(0.293)	0.953(0.191)	0.940(0.130)
			Efron	0.957(0.294)	0.954(0.193)	0.950(0.131)
			PB	0.951(0.291)	0.949(0.192)	0.945(0.131)
		0.1	Breslow	0.768(0.249)	0.546(0.162)	0.234(0.109)
			Efron	0.896(0.259)	0.832(0.170)	0.708(0.115)
			PB	0.947(0.288)	0.939(0.184)	0.923(0.122)
		0.2	Breslow	0.418(0.215)	0.128(0.141)	0.008(0.095)
			Efron	0.700(0.226)	0.489(0.148)	0.232(0.100)
			PB	0.932(0.276)	0.909(0.174)	0.885(0.114)
	2	0.01	Breslow	0.943(0.276)	0.921(0.177)	0.878(0.119)
			Efron	0.951(0.278)	0.943(0.179)	0.931(0.121)
			PB	0.949(0.281)	0.949(0.183)	0.944(0.124)
		0.1	Breslow	0.379(0.199)	0.099(0.126)	0.004(0.083)
			Efron	0.585(0.206)	0.323(0.131)	0.099(0.086)
			PB	0.937(0.273)	0.909(0.170)	0.860(0.111)
		0.2	Breslow	0.098(0.163)	0.004(0.105)	0.000(0.070)
			Efron	0.256(0.168)	0.059(0.108)	0.002(0.072)
			PB	0.910(0.265)	0.854(0.161)	0.755(0.102)

2.6 Real Applications

In Section 2.5, we demonstrated that for many cases, $\widehat{\boldsymbol{\beta}}_{\text{pb}}$ surpasses $\widehat{\boldsymbol{\beta}}_{\text{b}}$ and $\widehat{\boldsymbol{\beta}}_{\text{e}}$ in terms of reduced |Bias| and RMSE, and better confidence interval coverages. This is particularly evident when τ is large or when there is considerable variation in the p_{ij} values. This finding depicts an important insight: there are positive relationships between τ, $\sum_{i\in\mathcal{R}(t_{(j)})} p_{ij}^2/n_j$, and the inaccuracies of $\widehat{\boldsymbol{\beta}}_{\text{b}}$ and $\widehat{\boldsymbol{\beta}}_{\text{e}}$ in estimating $\boldsymbol{\beta}$. Here we confirm these relationships using real datasets, anticipating that a large τ or a high $\sum_{i\in\mathcal{R}(t_{(j)})} p_{ij}^2/n_j$ leads to an increase in the RMSE for $\widehat{\boldsymbol{\beta}}_{\text{b}}$ and $\widehat{\boldsymbol{\beta}}_{\text{e}}$, demonstrating the advantage of $\widehat{\boldsymbol{\beta}}_{\text{pb}}$ over these estimators.

For the ease of comparison, we first transformed the variables in all the datasets as follows. All the covariates, except for binary ones, are standardized to have mean 0 and standard deviation 1. The observed times $\{t_i\}_{i=1}^n$ are first scaled to all lie within the interval $[0,1]$. Then to create ties at different levels, we group the scaled times $\{t_i\}_{i=1}^n$ with different choices of τ. Specifically, let $\tau \in \{0, 0.01, 0.02, \ldots, 0.25\}$. When $\tau = 0$, the original observed times are used directly, i.e., $t_i^* = t_i$. For other values of τ, t_i^* is defined as $t_i^* = \lceil t_i/\tau \rceil \tau$.

For each τ, we fit $\widehat{\boldsymbol{\beta}}_{\text{b}}$, $\widehat{\boldsymbol{\beta}}_{\text{e}}$, and $\widehat{\boldsymbol{\beta}}_{\text{pb}}$ using the data $\{t_i^*, \delta_i, \boldsymbol{x}_i\}_{i=1}^n$. Due to lack of truth in real applications, we consider an estimation accuracy metric defined with respect to $\widehat{\boldsymbol{\beta}}_{\text{pb}}$, considering its theoretical accuracy and consistently strong performance in simulations. In particular, we define the estimation discrepancy (ED) of an estimator $\widehat{\boldsymbol{\beta}}$ from $\widehat{\boldsymbol{\beta}}_{\text{pb}}$ as

$$\max_{l\in\{1,\cdots,d\}} \left\{ \exp\left(|\widehat{\beta}_l - \widehat{\beta}_{\text{pb},l}| \right) - 1 \right\},$$

where $\widehat{\boldsymbol{\beta}}$ can be either $\widehat{\boldsymbol{\beta}}_{\text{b}}$ or $\widehat{\boldsymbol{\beta}}_{\text{e}}$. To gain more insights on the estimation accuracy performance, we also record the sum of squared hazards (SSH), defined as $\frac{1}{k}\sum_{j=1}^k \frac{1}{n_j} \sum_{i\in\mathcal{R}(t_{(j)})} \widehat{p}_{ij}^2$, which is the upper bound in Remark 2.2 with $\widehat{p}_{ij}$ being the evaluations of p_{ij} at $\widehat{\boldsymbol{\beta}}_{\text{pb}}$ and $\widehat{\lambda}_{\text{pb},j}$. Based on our findings in theory and simulation, we expect to observe a positive relationship of ED versus τ or SSH.

Besides the measures introduced above, it is also desirable to evaluation the performance of the methods via goodness-of-fit metrics. One such metric is the APL $L(\cdot, \boldsymbol{\Lambda}, \zeta)$ with $\boldsymbol{\Lambda}$ estimated by $\widehat{\boldsymbol{\Lambda}}_{\mathrm{pb}}$. The evaluations of the APL at the corresponding estimates of $\boldsymbol{\beta}$ yield $L_{\mathrm{b}} = L(\widehat{\boldsymbol{\beta}}_{\mathrm{b}}, \widehat{\boldsymbol{\Lambda}}_{\mathrm{pb}}, \zeta)$, $L_{\mathrm{e}} = L(\widehat{\boldsymbol{\beta}}_{\mathrm{e}}, \widehat{\boldsymbol{\Lambda}}_{\mathrm{pb}}, \zeta)$, and $L_{\mathrm{pb}} = L(\widehat{\boldsymbol{\beta}}_{\mathrm{pb}}, \widehat{\boldsymbol{\Lambda}}_{\mathrm{pb}}, \zeta)$, respectively for the three methods.

2.6.1 Male Laryngeal Cancer Patients Study

The original data are available as larynx dataset from R package “`KMsurv`”. Conducted at a Dutch hospital from 1970 to 1978, this study was reported by Kardaun [28]. It includes data on 90 male patients diagnosed with larynx cancer, documenting the time (in years) to death or censoring after their initial treatment. The dataset also records the covariates such as the age at diagnosis, the year of diagnosis, and the disease stage (four stages in total). More information can be found in Klein and Moeschberger [29]. For our analysis, we selected age and indicators for Stages 3 and 4 as covariates.

As shown in Figure 2.2, the study exhibits the desired positive relationships between the recorded values. In this example, the deviation of the Breslow estimator from the PB distribution estimator increases at an linear rate against τ or SSH throughout the whole range, while the deviation of the Efron estimator from the PB distribution estimator increases after $\tau \geq 0.125$ or SSH ≥ 0.375. The PB distribution estimator shows its advantage against the others in the domain of higher τ or higher SSH again. For the goodness-of-fit, our method delivers slightly higher APLs than the Breslow and Efron methods when τ is small, and has a clear advantage when τ increases to higher values.

Table A.1 indicates that the estimated β_j values are similar across all the methods for small

τ. However, as τ increases, Breslow's estimates diverge significantly from those of our PB distribution based estimates. Although the Efron estimates show closer alignment with our estimates, their discrepancy with our estimates also grows as τ increases when viewed alongside ED. On the other hand, the standard error estimates are similar across all the methods, regardless of the τ values.

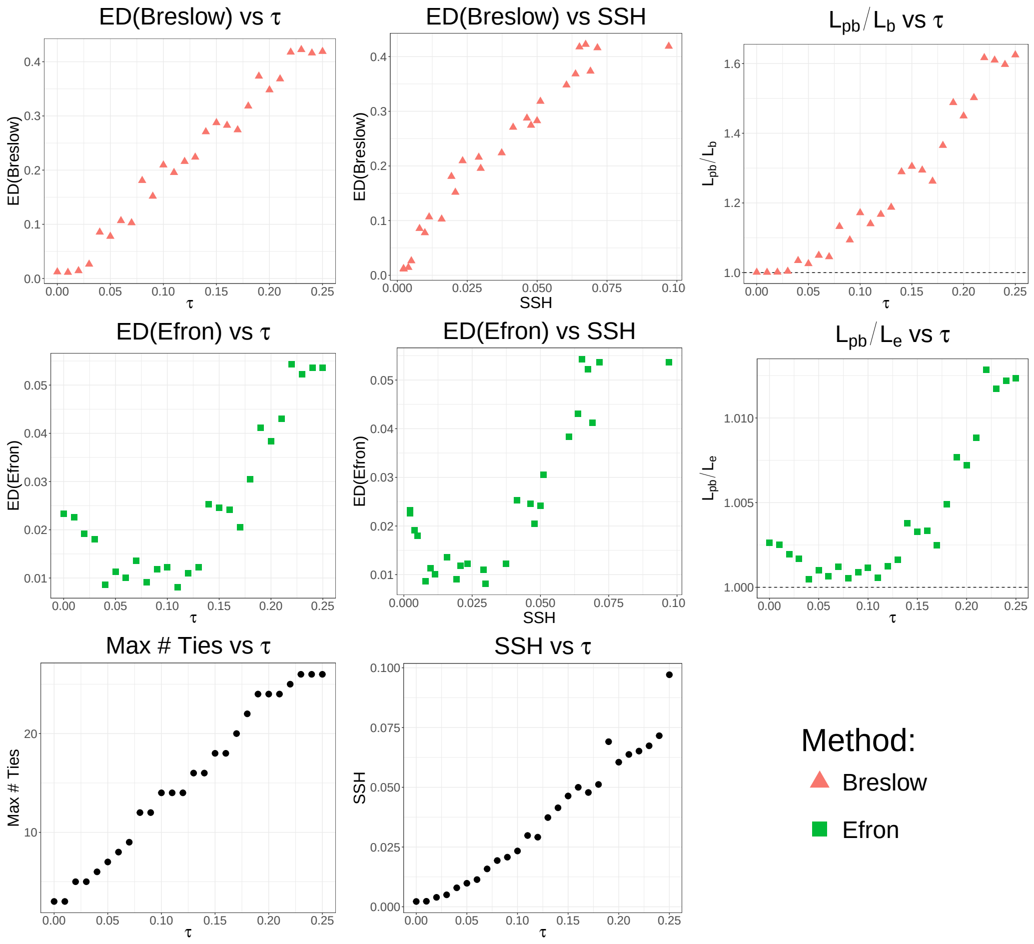

Figure 2.2: Results for the male laryngeal cancer patients study.

2.6.2 North Central Cancer Treatment Group Lung Cancer Study

The lung dataset, also available in the R package “`survival`”, originates from a study conducted by the north central cancer treatment group. This dataset captures survival data for patients with advanced lung cancer, noting the time (in days) to death or censoring for 228 patients. The covariates we consider here are sex, ECOG performance score as assessed by a physician, Karnofsky performance scores (both patient-rated and physician-rated ones), and weight loss over the last six months (in pounds). For more details, see Loprinzi et al. [32].

Figure 2.3 illustrates the desired positive relationships among the recorded values in this study. The trends are similar to what we observe in the laryngeal cancer study except that the extra drastic increases of the Efron estimator's ED happens respectively at $\tau = 0.2$ or $\text{SSH} = 0.175$. The goodness-of-fit comparison and estimation results in Table A.2 are similar to those in the male laryngeal cancer patients study. However, larger ratios of the APLs using the PB distribution versus the others are observed for large τ values in this study compared to the laryngeal cancer patients study.

General Finding: Through the examples with varying tie levels, we have further confirmed our finding in theory and simulation: $\widehat{\boldsymbol{\beta}}_{\text{pb}}$ could estimate $\boldsymbol{\beta}$ more accurately than existing methods in datasets with heavy ties or datasets with large variability among covariates. The PB distribution method also outperforms the existing methods in goodness-of-fit when there are heavy ties.

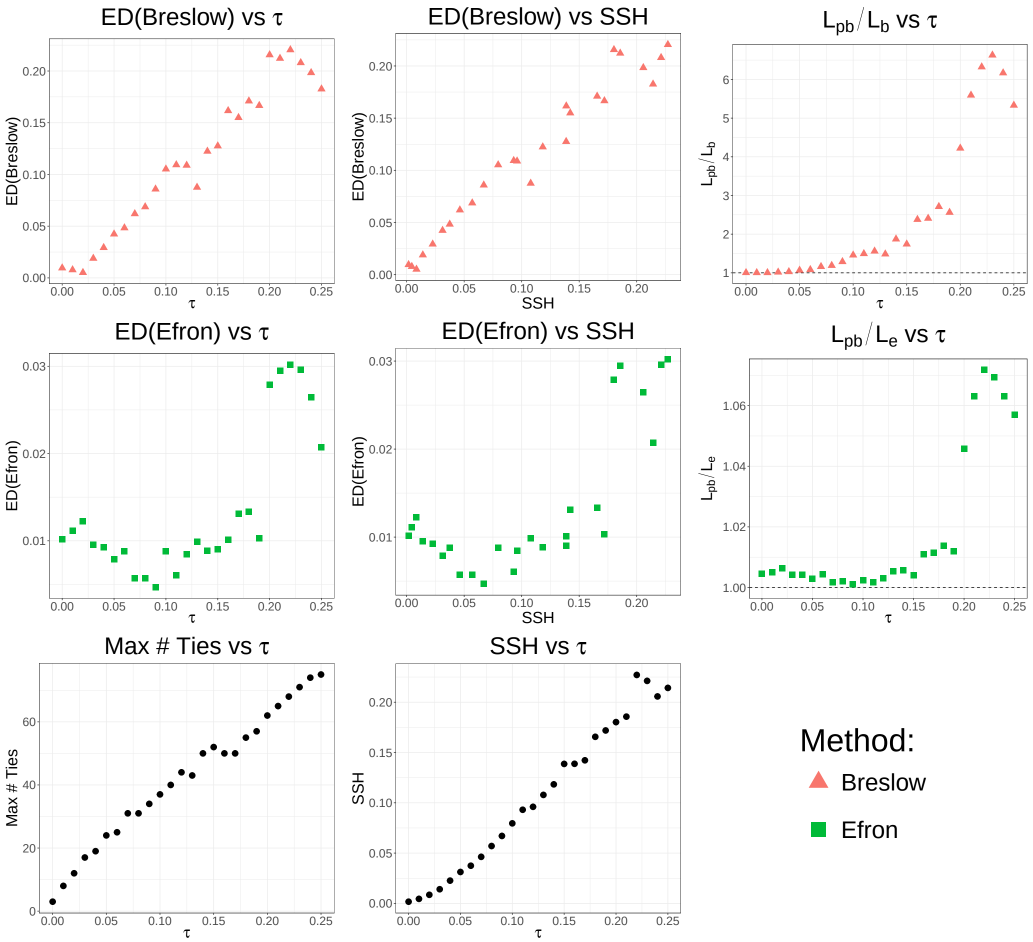

Figure 2.3: Results for the north central cancer treatment group lung cancer study.

2.7 Conclusion

In this project, we propose a new approach for accurately computing the partial likelihood for the Cox model, which integrates the original idea of APL and the recent development in efficient calculation of PB distribution. We also thoroughly study the properties of the new computing method. That is, our consistency and asymptotic normality for the new approach can cover not only grouped data with ties, but also continuous data without ties. Our numerical results show that due to reduction in bias, the new approach outperforms existing methods in reducing RMSE and improving coverage of confidence interval for the estimated coefficients, especially when data have many ties or the variation among the risk scores is high. Therefore, we recommend using the new method rather than the existing methods for these cases. One can see that the choice of $\widehat{\mathbf{\Lambda}}$, which is used as substitution of $\mathbf{\Lambda}$ in APL, is very flexible, only requiring $\widehat{\mathbf{\Lambda}}$ to satisfy a mild condition in Assumption A.3.9.

Without loss of generality and for notation simplicity, we only consider non-time-varying covariates in this project. The results can be easily extended to time-varying covariates. For our future theory, we might try to relax order condition for τ needed in the asymptotic theory of $\widehat{\boldsymbol{\beta}}_{\mathrm{pb}}$. Under the grouped continuous failure time model, to have consistency and asymptotic normality, the PB distribution estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{pb}}$ requires $\tau \asymp 1/n$. Meanwhile, the Breslow estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{b}}$ requires $\tau \to 0$ for consistency and $n^{1/2}\tau \to 0$ for asymptotic normality, which are much relaxed conditions than $\tau \asymp 1/n$. However, in our numerical example, $\widehat{\boldsymbol{\beta}}_{\mathrm{pb}}$ performs much better than $\widehat{\boldsymbol{\beta}}_{\mathrm{b}}$ under a large τ, i.e., lots of ties. The current theory for $\widehat{\boldsymbol{\beta}}_{\mathrm{pb}}$ only allows a non-diverging number of ties, i.e., $\sup_{t\in\Omega} d(t) = \mathcal{O}_{\mathrm{P}}(n\tau) = \mathcal{O}_{\mathrm{P}}(n \times 1/n) = \mathcal{O}_{\mathrm{P}}(1)$, but we might be able to allow a diverging number of ties for $\widehat{\boldsymbol{\beta}}_{\mathrm{pb}}$ in the theory considering its high performances in numerical examples with a large τ and, correspondingly, lots of ties.

Another direction of future research is to extend this idea to more complex survival models such as the competing risk Cox model, calculating the exact partial likelihood using a Poisson multinomial distribution (PMD). As Lin et al. [30] have proposed an efficient calculation for the PMD, we can use it for calculating the exact partial likelihood. Similar statistical theories in this project can also be considered.

As demonstrated in our numerical studies, the PB distribution method not only dominates the existing methods in the case of heavy ties, but also delivers competitive performance in the case of no or fewer ties. In addition, the dominance can also happen when covariate effects or covariate variances are large. However, the exact thresholds of these model parameters for such dominance can be hard to determine and most likely vary in different settings. Therefore, a practical guideline we would like to provide is to always fit a PB distribution method alongside the existing methods whenever there are ties in the data. When discrepancies are observed among the three methods, the PB distribution estimate, whose robustness is demonstrated here, is the one we would recommend.

Here we treat ties as grouped data from underlying continuous times and focus on proportional hazards modeling of the underlying continuous time hazard function. An alternative approach based on such grouped-continuous-times assumption, called discrete-time models that include parametric and machine learning classification methods, is discussed in Suresh et al. [46]. However, their approach directly models the discretized hazards instead of the underlying continuous hazard function. The latter form of hazard, as our focus here, is often of more interest in most survival analysis studies.

Chapter 3

Competing Risk Model with A Nonparametric Form of Spline-Estimated Relative Risks

3.1 Introduction

In survival analysis, it is common to observe time-to-event data subject to multiple, mutually exclusive causes of failure, giving rise to competing risks. When covariates are present, various statistical models have been developed to account for the structure imposed by competing risks. Among these, the cause-specific hazards model based on the Cox proportional hazards framework, or the competing risk Cox model, is one of the most widely adopted approaches and is treated in detail in Prentice et al. [39] and Chapter 8 of Kalbfleisch and Prentice [27]. Several alternative or extended approaches of this method have since been proposed. For example, Fine and Gray [15] proposed a proportional subdistribution hazards model to directly assess the effect of each risk factor on the cumulative incidence function. A data augmentation method to model multiple cause-specific hazards simultaneously using standard Cox regression was proposed by Lunn and Mcneil [33]. Modeling of correlated latent cause-specific event times using a copula is presented in Lo and Wilke [31]. For a comprehensive review of methodological developments in competing risks analysis, see

Monterrubio-Gómez et al. [37].

Meanwhile, nonparametric smooth modeling in survival analysis becomes essential when parametric assumptions become doubtful in practice. For example, kernel-based survival methods were considered in López-Cheda et al. [35, 36] for cure rate data and Akritas and Keilegom [1] for bivariate survival data. Penalty-based regression spline methods were studied in Joly et al. [24], Joly et al. [25] and Commenges et al. [7] for single event survival data with various censoring/truncation schemes, Rondeau et al. [42] for correlated survival data, and Joly and Commenges [23] for three-state data. Among nonparametric methods, smoothing spline survival models [18] provide an especially appealing approach due to its rigorous mathematical foundation in the reproducing kernel Hilbert space framework, straightforward extension to the multi-dimensional setting through tensor product spaces and the functional ANOVA structure, and reliable performance even in estimating multivariate functions. Besides the single event survival models [18], smoothing splines based methods have been studied in various other survival settings, such as high dimensional survival data [11], recurrent event data [12], and cure rate data [48].

Despite the rich literature in their respective areas, an integration of nonparametric smoothing into competing risk models have been rarely studied, with a few exceptions described below. For discrete-time competing risks survival data, Luo et al. [34] proposed a Cox–logistic regression model of the cause-specific hazards, where only the baseline hazards are estimated nonparametrically by simple quadratic or cubic spline expansions but the covariate effects are modeled linearly. A joint distribution approach based on kernel methods was proposed by Fermanian [14]. In particular, the joint survival function of the competing failure times is assumed to have a form defined by three components, an unknown cumulative distribution function, a set of cause-specific time functions, and a set of cause-specific covariate functions. Various constraints on the component functions, such as boundary conditions, monotonicity, and differenitability are necessary to make the model identifiable. Under these constraints, these three components are estimated separately by kernel methods. Despite its nonparametric nature, this method can be cumbersome to implement and its estimates are hard to interpret in practice due to its complex structure.

In this paper, we consider a nonparametric competing risk model based on the smoothing spline ANOVA framework. Instead of linear forms of covariate effects adopted in most parametric competing risk models, we consider cause-specific hazards with the true covariate effect functions assumed to lie in a tensor-product Sobolev space. This nonparametric formulation allows for flexible inclusion and decomposition of covariate effects. The estimation is through the optimization of a penalized partial likelihood consisting of the negative log partial likelihood representing the goodness-of-fit, a roughness penalty enforcing the smoothness constraint, and a smoothing parameter balancing the tradeoff. For the inference purpose, we consider the Bayesian confidence intervals based on the penalized partial likelihood, which provide point-wise coverage for the function parameters. A smoothing parameter selection criterion based on the Kullback-Leibler distance is introduced. For theoretical properties, we establish the rates of convergence for the function parameters in the proposed model. The model is evaluated through simulations, and then applied to a multiple myeloma dataset studying the effect of genes on cause-specific hazards. The analysis reveals nonlinear and distinct effects across risk factors, demonstrating the advantage of the proposed model over existing methods that often neglect competing risks or rely on linear covariate effects.

For the remainder of the paper, Section 3.2 presents the model formulation, function space structure, estimation procedure, and theoretical result. Section 3.3 provides simulation studies for the proposed model. Section 3.4 presents a real data application. Finally, Section 3.5 concludes the paper with a discussion of future research directions.

3.2 Smoothing Spline Competing Risk Cox Model

3.2.1 Model Structure

Let $\tilde{T} \in (0, \infty)$ and $C \in (0, \tau]$ be respectively the underlying event time and non-informative right-censoring time, with $\tau \in (0, \infty)$ representing the ending time of the study. Let $X = (X_{[1]}, \ldots, X_{[d]})^\top \in \mathcal{X} = \prod_{j=1}^{d} \mathcal{X}_{[j]}$ be a vector of d covariates. Without loss of generality, we assume $\mathcal{X}_{[j]} = [0, 1]$. Suppose there are K mutually exclusive risks, denoted by $\{1, \ldots, K\}$, that can be the cause of an event. Let $\tilde{G} = \sum_{k=1}^{K} k\mathbb{1}(\tilde{T} \text{ is due to the } k\text{-th risk}) \in \{1, \ldots, K\}$ represent the underlying risk associated with the event. We assume that a single subject can experience at most one event, and that $\tilde{T}$ and C are conditionally independent given X. Let $T = \min\{\tilde{T}, C\} \in (0, \tau]$ be the observed time and $\mathcal{I} = \mathbb{1}(\tilde{T} \leq C)$ be the censoring indicator. Let $G = \tilde{G} \cdot \mathcal{I} \in \{0, 1, \ldots, K\}$ be the observed risk indicator. Note that $G = 0$ when the subject is censored and is the risk type otherwise.

Let $\{W_i\}_{i=1}^{n} = \{\tilde{T}_i, C_i, X_i, \tilde{G}_i\}_{i=1}^{n}$ be the underlying data, which are n independent and identically distributed copies from $W = \{\tilde{T}, C, X, \tilde{G}\}$. The observations $\{T_i, X_i, \mathcal{I}_i, G_i\}_{i=1}^{n}$ are then copies from $\{T, X, \mathcal{I}, G\}$. In competing risk models, for $t \in [0, \infty)$, the hazard function of $\tilde{T}$ given X is

$$\begin{aligned} dH(t; X) &= h(t; X)dt = \mathbb{P}(\tilde{T} \in [t, t+dt) | \tilde{T} \geq t, X) \\ &= \sum_{k=1}^{K} \mathbb{P}(\tilde{T} \in [t, t+dt), \tilde{G} = k | \tilde{T} \geq t, X) = \sum_{k=1}^{K} dH_k(t; X) = \sum_{k=1}^{K} h_k(t; X)dt, \end{aligned}$$

where $dH_k(t; X)$ is the cause-specific hazard function of the k-th risk factor. The survival function is given by $\mathsf{S}(t; X) = \mathbb{P}(\tilde{T} > t | X) = \exp(-H(t; X))$ and the distribution function is then calculated as $\mathsf{F}(t; X) = \mathbb{P}(\tilde{T} \leq t | X) = 1 - \mathsf{S}(t; X)$. The density function is $\mathsf{f}(t; X) = \partial \mathsf{F}(t; X)/(\partial t) = h(t; X)\mathsf{S}(t; X) = \sum_{k=1}^{K} \mathsf{f}_k(t; X)$, where $\mathsf{f}_k(t; X) = h_k(t; X)\mathsf{S}(t; X)$. The distribution function can also be expressed as $\mathsf{F}(t;X) = \sum_{k=1}^{K} \mathsf{F}_k(t;X)$, where $\mathsf{F}_k(t;X) = \mathbb{P}(\tilde{T} \leq t, \tilde{G} = k|X) = \int_0^t \mathsf{f}_k(s;X)ds$ is the cumulative incidence function of the k-th risk factor. Define $\mathsf{P}(k;X) = \mathbb{P}(\tilde{G} = k|X) = \lim_{t\to\infty} \mathsf{F}_k(t;X)$, where $\sum_{k=1}^{K} \mathsf{P}(k;X) = 1$.

For each k, let $\eta_k^* : \mathcal{X} \to \mathbb{R}$ denote the true covariate effect function for the k-th risk. We assume the proportional cause-specific hazards model, that is, the cause-specific hazard functions follow the Cox model

$$dH_k(t;X) = dH_{0k}(t)\exp(\eta_k^*(X)),$$

where $dH_{0k}(t) = h_{0k}(t)dt$ is the baseline cause-specific hazard function, and $h_{0k}(0) = 0$.

3.2.2 Reproducing Kernel Hilbert Spaces

Here we introduce reproducing kernel Hilbert spaced (RKHS) following the smoothing spline ANOVA framework described in Gu [18]. The Sobolev space of order $m \in \mathbb{N}$ corresponding to the jth axis $\mathcal{X}_{[j]}$ is given by

$$\mathcal{H}_{[j]} = \left\{ f : f, f^{(1)}, \ldots, f^{(m-1)} \text{ are absolutely continuous, } \int_{\mathcal{X}_{[j]}} |f^{(m)}(x_{[j]})|^2 dx_{[j]} < \infty \right\}.$$

Let $\mathcal{A}_{[j]}$ be the averaging operator of the jth axis, defined by

$$\mathcal{A}_{[j]}f(x) = \mathcal{A}_{[j]}f(x_{[1]}, \ldots, x_{[d]}) = \int_{\mathcal{X}_{[j]}} f(x_{[1]}, \ldots, x_{[d]})dx_{[j]}.$$

Then $\mathcal{H}_{[j]}$ can be decomposed as $\mathcal{H}_{[j]} = \mathcal{H}_{\emptyset[j]} \oplus \mathcal{H}_{\{j\}}$, where $\mathcal{H}_{\emptyset[j]} = \left\{ c_j \mathcal{A}_{[j]} f_{[j]} : c_j \in \mathbb{R}, f_{[j]} \in \mathcal{H}_{[j]} \right\} = \text{Span}\{1\}$, and $\mathcal{H}_{\{j\}} = \{f_{[j]} \in \mathcal{H}_{[j]} : \mathcal{A}_{[j]} f_{[j]} = 0\}$.

The tensor product Sobolev space on $\mathcal{X} = \prod_{j=1}^{d} \mathcal{X}_{[j]}$ can be written as

$$\begin{aligned}\otimes_{j=1}^{d}\mathcal{H}_{[j]} &= \otimes_{j=1}^{d}\left\{\mathcal{H}_{\emptyset[j]} \oplus \mathcal{H}_{\{j\}}\right\} \\ &= \oplus_{S\in\mathcal{P}_d}\mathcal{H}_S = \mathcal{H}_\emptyset \oplus \left\{\oplus_{j=1}^{d}\mathcal{H}_{\{j\}}\right\} \oplus \left\{\oplus_{j<j'}\mathcal{H}_{\{j,j'\}}\right\} \oplus \ldots \oplus \mathcal{H}_{\{1,\ldots,d\}},\end{aligned}$$

where $\mathcal{P}_d$ denotes the power set of $\{1,\ldots,d\}$. For each $S \in \mathcal{P}_d$, the associated space is

$$\mathcal{H}_S = \{\otimes_{j\in S}\mathcal{H}_{\{j\}}\} \otimes \{\otimes_{j\in\{1,\ldots,d\}\setminus S}\mathcal{H}_{\emptyset[j]}\} = \{\otimes_{j\in S}\mathcal{H}_{\{j\}}\} \otimes \mathrm{Span}\{1\} = \otimes_{j\in S}\mathcal{H}_{\{j\}}.$$

Note that $\mathcal{H}_S$ is defined over $\mathcal{X}_S = \prod_{j\in S}\mathcal{X}_{[j]}$ and represents the space for effect S. The constant part, i.e., the intercept space, is $\mathcal{H}_\emptyset = \otimes_{j=1}^{d}\mathcal{H}_{\emptyset[j]} = \mathrm{Span}\{1\}$.

Let

$$\mathcal{H} = \oplus_{S\in\mathbb{S}}\mathcal{H}_S$$

for some collection $\mathbb{S}$ satisfying $\emptyset \notin \mathbb{S}$ and $\mathbb{S} \subseteq \mathcal{P}_d$. Note that, since the Cox model does not include an intercept, $\mathcal{H}_\emptyset$ is excluded from $\mathcal{H}$. As a result, $\int_{\mathcal{X}} f(x)\,dx = 0$ for all $f \in \mathcal{H}$. The selection of $\mathbb{S}$ specifies the structure of $\mathcal{H}$. For instance, if $\mathbb{S} = \{\{1\},\ldots,\{d\}\}$, then $\mathcal{H}$ adopts an additive form that involves only main effects. On the other hand, choosing $\mathbb{S} = \{\{1\},\ldots,\{d\},\{1,2\},\ldots,\{d-1,d\}\}$ allows $\mathcal{H}$ to incorporate all main effects together with two-factor interactions. For each k, we assume $\eta_k^* = \sum_{S\in\mathbb{S}}\eta_{k,S}^* \in \mathcal{H}$, where $\eta_{k,S}^* \in \mathcal{H}_S$ for every $S \in \mathbb{S}$.

Given a semi-inner product $J(\cdot,\cdot)$ on $\mathcal{H}$, $\mathcal{H}$ can be decomposed as

$$\mathcal{H} = \mathcal{N}_J \oplus \mathcal{H}_J,$$

where $\mathcal{N}_J = \{f \in \mathcal{H} : J(f) = J(f,f) = 0\}$ forms a finite-dimensional parametric subspace of

$\mathcal{H}$, and $\mathcal{H}_J$ forms an infinite-dimensional nonparametric subspace of $\mathcal{H}$. Furthermore, $J(\cdot,\cdot)$ is an inner product on $\mathcal{H}_J$, with its associated reproducing kernel (RK) $\mathcal{K}_J(\cdot,\cdot)$ on $\mathcal{H}_J$.

In practice, for each risk k, when estimating η_k^*, we use J_k and $\mathcal{K}_{J_k}$ instead of J and $\mathcal{K}_J$. Here, J_k is an inner product on $\mathcal{H}_J$ whose induced norm is equivalent to that of J, and $\mathcal{K}_{J_k}$ denotes the reproducing kernel associated with J_k on $\mathcal{H}_J$. Further details on the RKHS, along with the calculation for J, $\mathcal{K}_J$, J_k, and $\mathcal{K}_{J_k}$, are provided in Section B.6.

3.2.3 Estimation

We provide estimation procedure for the proposed model. First, define counting process for each k and at-risk process. Let $t \in [0,\tau]$. For each risk k, the counting process for the k-th risk is defined as

$$N_k(t) = \mathbb{1}(T \le t, G = k, \mathfrak{I} = 1) = \mathbb{1}(\tilde{T} \le t, \tilde{T} \le C, \tilde{G} = k),$$

and the at-risk process is defined as

$$Y(t) = \mathbb{1}(T \ge t) = \mathbb{1}(\tilde{T} \ge t \text{ and } C \ge t).$$

For each subject i, $N_{ik}(\cdot)$ and $Y_i(\cdot)$ are defined in the same manner. For details on the martingale framework in the theoretical derivation, see Section B.1.

For $f \in \mathcal{H}$, define minus log-partial likelihood for k-th risk as

$$\mathcal{L}_{k(n)}(f) = -\int_0^\tau \frac{1}{n}\sum_{i=1}^n \left\{ f(X_i) - \ln\left(\frac{1}{n}\sum_{i'=1}^n Y_{i'}(t)\exp(f(X_{i'}))\right)\right\} dN_{ik}(t).$$

For covariates, for each k, we randomly select q_k elements from n subjects, i.e., $\{Z_{\ell(k)}\}_{\ell=1}^{q_k}$

$\subseteq \{X_i\}_{i=1}^n$. Then, $\mathcal{H}_J^{(k)} = \text{Span}\{\mathcal{K}_{J_k}(Z_{1(k)}, \cdot), \ldots, \mathcal{K}_{J_k}(Z_{q_k(k)}, \cdot)\}$ forms a finite-dimensional subspace of $\mathcal{H}_J$, which further defines $\mathcal{H}^{(k)} = \mathcal{N}_J \oplus \mathcal{H}_J^{(k)}$, a finite-dimensional subspace of $\mathcal{H}$. Note that q_k and the selected q_k elements may be different for each k. For each risk k, using the efficient approximation [18], we get estimator of η_k^* by

$$\hat{\eta}_k = \underset{f \in \mathcal{H}^{(k)}}{\text{argmin}}\ \mathcal{L}_{k(n)}(f) + \frac{\lambda_k}{2} J_k(f), \tag{3.1}$$

where $J_k(f) = J_k(f, f)$ is a roughness penalty and $\lambda_k \in (0, \infty)$ is a smoothing parameter. Note that when $q_k = n$, we have $\{Z_{\ell(k)}\}_{\ell=1}^{q_k} = \{X_i\}_{i=1}^n$, and then, by the representer theorem, $\hat{\eta}_k = \hat{\eta}_k^* \equiv \underset{f \in \mathcal{H}}{\text{argmin}}\ \mathcal{L}_{k(n)}(f) + \lambda_k J_k(f)/2$.

One can represent $\mathcal{N}_J = \text{Span}\{\phi_1, \ldots, \phi_p\}$ for known basis functions $\phi_1, \ldots, \phi_p$. See Section B.6 for more details. Solution of the optimization problem in (3.1) is equivalent to $\hat{\eta}_k = \sum_{l=1}^p \hat{d}_{k,l}\phi_l + \sum_{\ell=1}^{q_k} \hat{c}_{k,\ell}\mathcal{K}_{J_k}(Z_{\ell(k)}, \cdot)$, where

$$(\hat{\boldsymbol{d}}_k^\top, \hat{\boldsymbol{c}}_k^\top)^\top = \underset{\boldsymbol{d}_k \in \mathbb{R}^p, \boldsymbol{c}_k \in \mathbb{R}^{q_k}}{\text{argmin}}\ \mathcal{L}_{k(n)}\left(\sum_{l=1}^p d_{k,l}\phi_l + \sum_{\ell=1}^{q_k} c_{k,\ell}\mathcal{K}_{J_k}(Z_{\ell(k)}, \cdot)\right) + \frac{\lambda_k}{2}\boldsymbol{c}_k^\top \boldsymbol{\mathcal{K}}_{J_k}\boldsymbol{c}_k \tag{3.2}$$

with $\boldsymbol{\mathcal{K}}_{J_k} = \{\mathcal{K}_{J_k}(Z_{\ell(k)}, Z_{\ell'(k)})\}_{\ell,\ell' \in \{1,\ldots,q_k\}}$.

3.2.4 Bayesian Confidence Interval

We can compute point-wise Bayesian confidence intervals for the fitted functions. Similar to Gu [18], for each k, $\boldsymbol{\varphi}_k(\cdot) = [\phi_1(\cdot), \ldots, \phi_p(\cdot), \mathcal{K}_{J_k}(Z_{1(k)}, \cdot), \ldots, \mathcal{K}_{J_k}(Z_{q_k(k)}, \cdot)]^\top$ and $\boldsymbol{V}_{k(n)}(\hat{\eta}_k)$ is the hessian matrix for the optimization in (3.2) evaluated at $(\hat{\boldsymbol{d}}_k^\top, \hat{\boldsymbol{c}}_k^\top)^\top$, which can be expressed as

$$\boldsymbol{V}_{k(n)}(\hat{\eta}_k) = \int_0^\tau \left\{ \frac{S_{k(n)}^{(2)}(\hat{\eta}_k, t)}{S_{k(n)}^{(0)}(\hat{\eta}_k, t)} - \frac{S_{k(n)}^{(1)}(\hat{\eta}_k, t)}{S_{k(n)}^{(0)}(\hat{\eta}_k, t)} \frac{S_{k(n)}^{(1)}(\hat{\eta}_k, t)^\top}{S_{k(n)}^{(0)}(\hat{\eta}_k, t)} \right\} d\overline{N}_k(t) + \text{Diag}\left(\boldsymbol{O}_p, \lambda_k \boldsymbol{\mathcal{K}}_{J_k}\right),$$

where $S^{(0)}_{k(n)}(\hat{\eta}_k, t) = \sum_{i=1}^n Y_i(t)\exp(\hat{\eta}_k(X_i))/n$, $S^{(1)}_{k(n)}(\hat{\eta}_k, t) = \sum_{i=1}^n Y_i(t)\boldsymbol{\varphi}_k(X_i)\exp(\hat{\eta}_k(X_i))/n$, $S^{(2)}_{k(n)}(\hat{\eta}_k, t) = \sum_{i=1}^n Y_i(t)\boldsymbol{\varphi}_k(X_i)\boldsymbol{\varphi}_k(X_i)^\top \exp(\hat{\eta}_k(X_i))/n$, $\overline{N}_k(\cdot) = \sum_{i=1}^n N_{ik}(\cdot)/n$, and $\boldsymbol{O}_p$ is a $p \times p$ zero matrix.

The point-wise $100(1-\alpha)\%$ confidence interval for $\eta_k^*(x)$ at a fixed point $x \in \mathcal{X}$ is given by

$$\hat{\eta}_k(x) \pm z_{1-\alpha/2}\sqrt{\frac{1}{n}\boldsymbol{\varphi}_k(x)^\top \boldsymbol{V}^{-1}_{k(n)}(\hat{\eta}_k)\boldsymbol{\varphi}_k(x)},$$

where $z_{1-\alpha/2}$ is the $(1-\alpha/2)$-quantile of the standard normal distribution. The point-wise Bayesian confidence intervals for each effect function can be computed in a similar manner.

3.2.5 Smoothing Parameter Selection

We propose a rule for selecting the smoothing parameter for each k. For each k, let $n_k = \sum_{i=1}^n N_{ik}(\tau)$ represent the number of subjects who failed due to the k-th risk factor. Let $\{X_{\ell(k)}\}_{\ell=1}^{n_k} \subseteq \{X_i\}_{i=1}^n$ be the subset of n_k covariates out of n that satisfy $N_{ik}(\tau) = 1$.

Using the relative Kullback-Leibler distance (RKL) proxy for leave-one-out cross validation, as outlined in Du et al. [11], we select the smoothing parameter λ_k that minimizes

$$\mathrm{RKL}_k(\lambda_k) = \frac{n}{n_k}\mathcal{L}_{k(n)}(\hat{\eta}_k) + \frac{\alpha_k}{n_k(n_k-1)}\mathrm{tr}\left(\boldsymbol{P}_k^\top \boldsymbol{Q}_k \boldsymbol{V}^{-1}_{k(n)}(\hat{\eta}_k)\boldsymbol{Q}_k^\top \boldsymbol{P}_k\right),$$

where $\boldsymbol{P}_k = \boldsymbol{I}_{n_k} - \mathbf{1}_{n_k}\mathbf{1}_{n_k}^\top/n_k$, $\boldsymbol{I}_{n_k}$ is $n_k \times n_k$ identity matrix, $\mathbf{1}_{n_k}$ is length n_k one vector, and $\boldsymbol{Q}_k = [\boldsymbol{\varphi}_k(X_{1(k)}), \ldots, \boldsymbol{\varphi}_k(X_{n_k(k)})]^\top$. We set $\alpha_k = 1.4$, following the recommendation in Gu [18].

3.2.6 Rate of Convergence

Here, we present the rate of convergence result for the estimator in (3.1). The asymptotic notations and the assumptions are listed in Sections B.2 and B.3, respectively.

For each k, define a non-negative definite bivariate functional $\mathcal{V}_k(\cdot,\cdot)$ on $\mathcal{H}$ by

$$\mathcal{V}_k(f,g)=\int_0^\tau\left\{\frac{S(fg;\eta_k^*,t)}{S(\eta_k^*,t)}-\frac{S(f;\eta_k^*,t)}{S(\eta_k^*,t)}\frac{S(g;\eta_k^*,t)}{S(\eta_k^*,t)}\right\}S(\eta_k^*,t)dH_{0k}(t)$$

for $f,g\in\mathcal{H}$, where $S(\eta,t)\equiv\mathbb{E}_W[Y(t)\exp(\eta(X))]$, $S(f;\eta,t)\equiv\mathbb{E}_W[Y(t)f(X)\exp(\eta(X))]$, and $S(fg;\eta,t)\equiv\mathbb{E}_W[Y(t)f(X)g(X)\exp(\eta(X))]$ for $f,g,\eta\in\mathcal{H}$ and $t\in[0,\tau]$. Let $\mathcal{V}_k(f)=\mathcal{V}_k(f,f)$ for $f\in\mathcal{H}$.

Theorem 3.1. *Under Assumption B.3.1, for each k, as $n\to\infty$, we have*

$$\sqrt{(\mathcal{V}_k+\lambda_k J_k)(\hat{\eta}_k-\eta_k^*)}=\mathcal{O}_{\mathbb{P}}\left(n^{-1/2}\lambda_k^{-1/(2r)}+\lambda_k^{b_k/2}\right)=o_{\mathbb{P}}(1).$$

The proof of Theorem 3.1 is provided in Section B.4. One can show that Theorem 3.1 implies $\|\hat{\eta}_k-\eta_k^*\|_{\mathcal{L}_2}=\mathcal{O}_{\mathbb{P}}\left(n^{-1/2}\lambda_k^{-1/(2r)}+\lambda_k^{b_k/2}\right)$, where $\|\cdot\|_{\mathcal{L}_2}$ represents the $\mathcal{L}_2$ norm. Note that $r>1$ is required, where $r=2m$ when $d=1$, or when $d>1$ with $\mathcal{H}$ having an additive structure; and $r=2m-\delta$ for any $\delta\in(0,2m-1)$ when $d>1$ with $\mathcal{H}$ including at least one interaction. See Gu [18] for additional details. The parameter $b_k\in[1,2]$ represents the smoothness of the true function η_k^*, with η_k^* being referred to as supersmooth when $b_k>1$. To achieve the optimal rate of convergence, if $\lambda_k\asymp n^{-r/(rb_k+1)}$, which solves $n^{-1/2}\lambda_k^{-1/(2r)}\asymp\lambda_k^{b_k/2}$, we obtain the optimal rate of convergence $\mathcal{O}_{\mathbb{P}}(n^{-rb_k/(2rb_k+2)})$.

3.3 Simulation Study

We perform a simulation study for the proposed method. We set $K = 2$ and $d = 3$, and impose an additive structure on $\mathcal{H}$; that is, for each k, we assume $\eta_k^* = \sum_{j=1}^d \eta_{k,\{j\}}^*$, where $\eta_{k,\{j\}}^* = \tilde{\eta}_{k,\{j\}}^* - \int_{\mathcal{X}_{[j]}} \tilde{\eta}_{k,\{j\}}^*(x_{[j]})dx_{[j]} \in \mathcal{H}_{\{j\}}$ and

$$\begin{aligned}
&\tilde{\eta}_{1,\{1\}}^*(x_{[1]}) = 2\sin(1.2\pi x_{[1]}) + 1.6x_{[1]}^2,\ \ \tilde{\eta}_{1,\{2\}}^*(x_{[2]}) = 1.4\sin(2\pi x_{[2]}) + 2.8x_{[2]},\\
&\tilde{\eta}_{1,\{3\}}^*(x_{[3]}) = 1.5\cos(2\pi x_{[3]}) - 0.5x_{[3]},\\
&\tilde{\eta}_{2,\{1\}}^*(x_{[1]}) = 2.4x_{[1]}^3 - 2.4x_{[1]}^2 - 0.9x_{[1]} - 1.8\sin\left(\frac{\pi x_{[1]}}{1.2}\right),\ \ \tilde{\eta}_{2,\{2\}}^*(x_{[2]}) = -2.25\sin\left(\frac{\pi x_{[2]}}{1.1}\right),\\
&\tilde{\eta}_{2,\{3\}}^*(x_{[3]}) = 1.6\sin(2\pi x_{[3]}) + 0.5x_{[3]}^2.
\end{aligned}$$

We assume a Weibull baseline cause-specific hazard functions; that is, for each k, $h_{0k}(t) = \gamma_k t^{\gamma_k - 1}/\beta_k^{\gamma_k}$. We set $\gamma_1 = \gamma_2 = 1.4$ and $\beta_1 = \beta_2 = 1.7$. Similarly, the hazard function for the underlying censoring time is also assumed to be Weibull; that is, for a random variable $\tilde{C} \in (0, \infty)$, we have $C = \min\{\tilde{C}, \tau\}$, where $h_c(t) = \gamma_c t^{\gamma_c - 1}/\beta_c^{\gamma_c}$ and $dH_c(t) = h_c(t)dt = \mathbb{P}(\tilde{C} \in [t, t+dt) \mid \tilde{C} \geq t)$. We set $\tau = 1$.

Let $\tilde{X} = (\tilde{X}_{[1]}, \tilde{X}_{[2]}, \tilde{X}_{[3]})^\top$ be a multivariate normal random variable with mean vector $(0,0,0)^\top$ and covariance matrix $\{0.5^{|j-j'|}\}_{1\leq j,j'\leq 3}$. We have $X = (X_{[1]}, X_{[2]}, X_{[3]})^\top$ $= (\Phi(\tilde{X}_{[1]}), \Phi(\tilde{X}_{[2]}), \Phi(\tilde{X}_{[3]}))^\top$, where Φ denotes the distribution function of the standard normal distribution. Each $X_{[j]}$ follows a uniform distribution on $[0,1]$ and $X_{[1]}, X_{[2]}, X_{[3]}$ are dependent.

Considering that $\mathbb{P}(\tilde{G} = k | X, \tilde{T} \in [t, t+dt)) = h_k(t; X)/h(t; X)$ for each k, we generate simulation data following the approach outlined in Beyersmann et al. [2]. Specifically, for each subject i, (1) generate C_i; (2) generate X_i; (3) generate $\tilde{T}_i$ from $\mathsf{F}(t; X_i)$; (4) generate $\tilde{G}_i$ from $\text{Multinomial}\big(1, h_1(\tilde{T}_i; X_i)/h(\tilde{T}_i; X_i), \ldots, h_K(\tilde{T}_i; X_i)/h(\tilde{T}_i; X_i)\big)$; (5) compute T_i, $\mathfrak{I}_i$, and G_i.

For the simulation setting, we vary the censoring levels by $(\gamma_c, \beta_c) \in \{(0.4, 0.3), (1.2, 1.1), (2.8, 2.7)\}$ and the sample size $n \in \{250, 500\}$, where smaller γ_c or smaller β_c induces a heavier censoring rate. We repeat the simulation over $1,000$ replicates, where, for each replicate, we compute $\hat{\eta}_k$ for each k. For each simulation setting, we evaluate the performance of the proposed method across all replicates, focusing on the rate of convergence and the Bayesian confidence interval.

Table 3.1 presents the data summary of the simulation settings. One can observe that events are equally likely to occur for both risk factors. With smaller values of γ_c or β_c—more observations are censored, and fewer subjects remain in the at-risk set until the end of the study.

Estimation performances are reported in Table 3.2, while both estimation performances and the Bayesian confidence interval performances are shown in Figures 3.1 – 3.3. As expected, lower censoring levels and larger sample sizes yield improved estimation accuracy, narrower confidence intervals, and generally better coverage rates. The proposed Bayesian confidence intervals align closely with empirical quantiles across replicates, demonstrating their reliability.

n	h_c		Data Summary			
	γ_c	β_c	$\sum_{i=1}^{n} \mathbb{1}(T_i = \tau)$	$\sum_{i=1}^{n} \mathbb{1}(G_i = 0)$	$\sum_{i=1}^{n} \mathbb{1}(G_i = 1)$	$\sum_{i=1}^{n} \mathbb{1}(G_i = 2)$
250	0.4	0.3	15.46(3.79)	184.60(6.98)	31.77(5.17)	33.63(5.43)
	1.2	1.1	31.92(5.26)	118.31(7.74)	65.46(6.75)	66.23(6.81)
	2.8	2.7	72.80(7.06)	79.12(7.11)	86.41(7.21)	84.47(7.35)
500	0.4	0.3	30.88(5.39)	369.26(9.90)	64.07(7.51)	66.66(7.41)
	1.2	1.1	64.07(7.50)	237.84(11.13)	130.44(9.74)	131.72(9.40)
	2.8	2.7	146.85(9.99)	159.48(10.28)	171.98(10.41)	168.54(10.61)

Table 3.1: Summary of data by simulation settings. Each cell shows the mean (standard deviation) over 1,000 replicates.

n	h_c		$\|\hat{\eta}_{k,\{j\}} - \eta^*_{k,\{j\}}\|_{\mathcal{L}_2}$					
	γ_c	β_c	$j = 1$		$j = 2$		$j = 3$	
			$k = 1$	$k = 2$	$k = 1$	$k = 2$	$k = 1$	$k = 2$
250	0.4	0.3	0.42(0.21)	0.39(0.22)	0.47(0.18)	0.41(0.26)	0.41(0.20)	0.60(0.34)
	1.2	1.1	0.29(0.13)	0.26(0.12)	0.31(0.12)	0.26(0.14)	0.29(0.10)	0.36(0.18)
	2.8	2.7	0.26(0.12)	0.21(0.10)	0.27(0.10)	0.22(0.11)	0.25(0.08)	0.30(0.14)
500	0.4	0.3	0.29(0.13)	0.25(0.13)	0.31(0.12)	0.26(0.14)	0.30(0.10)	0.37(0.18)
	1.2	1.1	0.21(0.09)	0.17(0.07)	0.21(0.08)	0.18(0.08)	0.21(0.07)	0.25(0.10)
	2.8	2.7	0.18(0.07)	0.15(0.07)	0.19(0.07)	0.15(0.07)	0.19(0.06)	0.21(0.08)

Table 3.2: Summary of $\mathcal{L}_2$ loss by simulation settings. Each cell shows the mean (standard deviation) over 1,000 replicates.

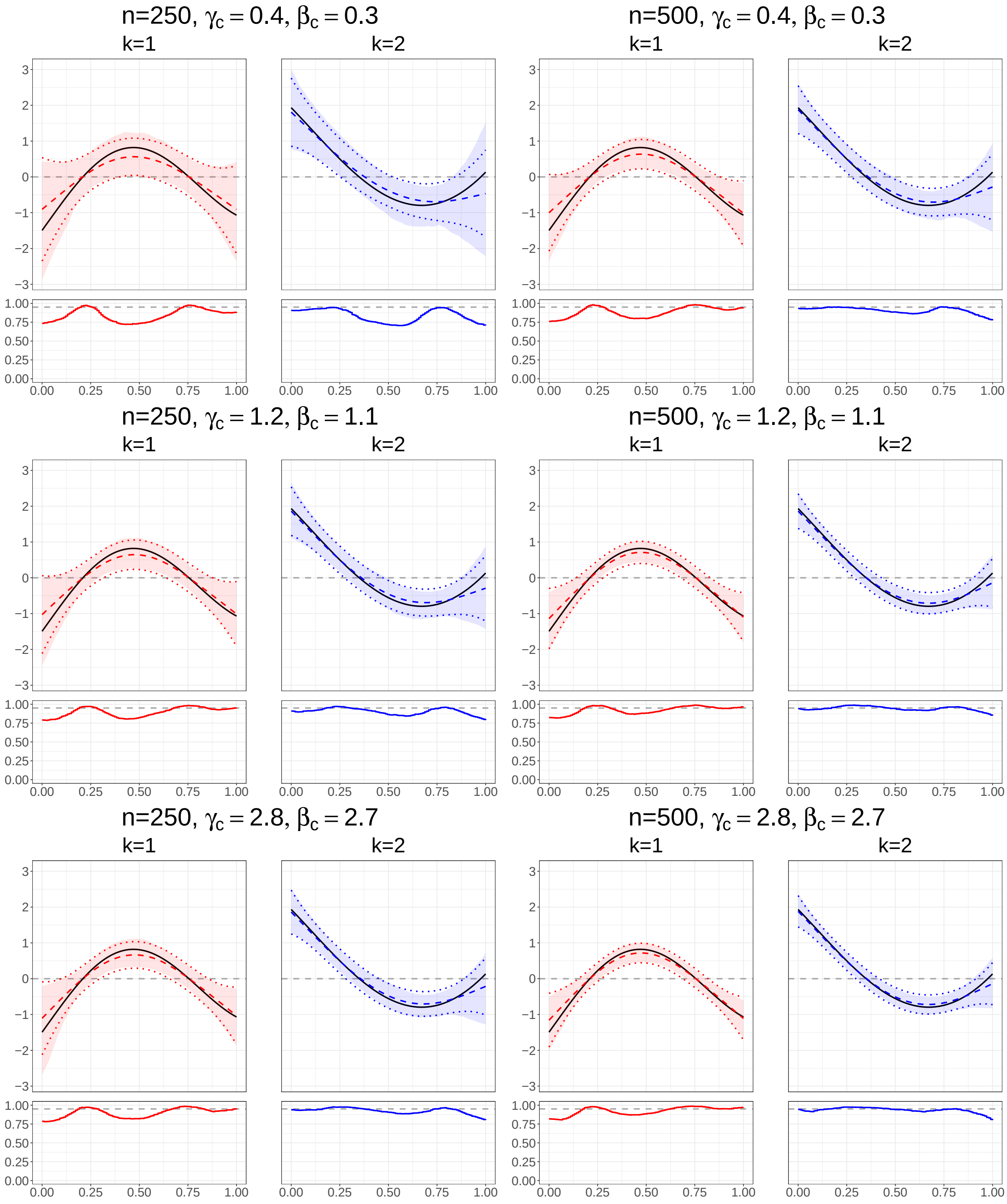

Figure 3.1: Simulation results for $\mathcal{X}_{[1]}$. For each simulation setting and each risk factor, the top panel shows $\eta^*_{k,\{1\}}$ (solid line), average of $\hat{\eta}_{k,\{1\}}$ (dashed line), average of the 95% Bayesian confidence interval for $\eta^*_{k,\{1\}}$ (dotted lines), and 2.5% and 97.5% empirical quantiles of $\hat{\eta}_{k,\{1\}}$ (shaded area). The bottom panel displays the empirical coverage rate of the 95% Bayesian confidence interval for $\eta^*_{k,\{1\}}$. Note that all averages and empirical quantiles are computed point-wise across all replicates.

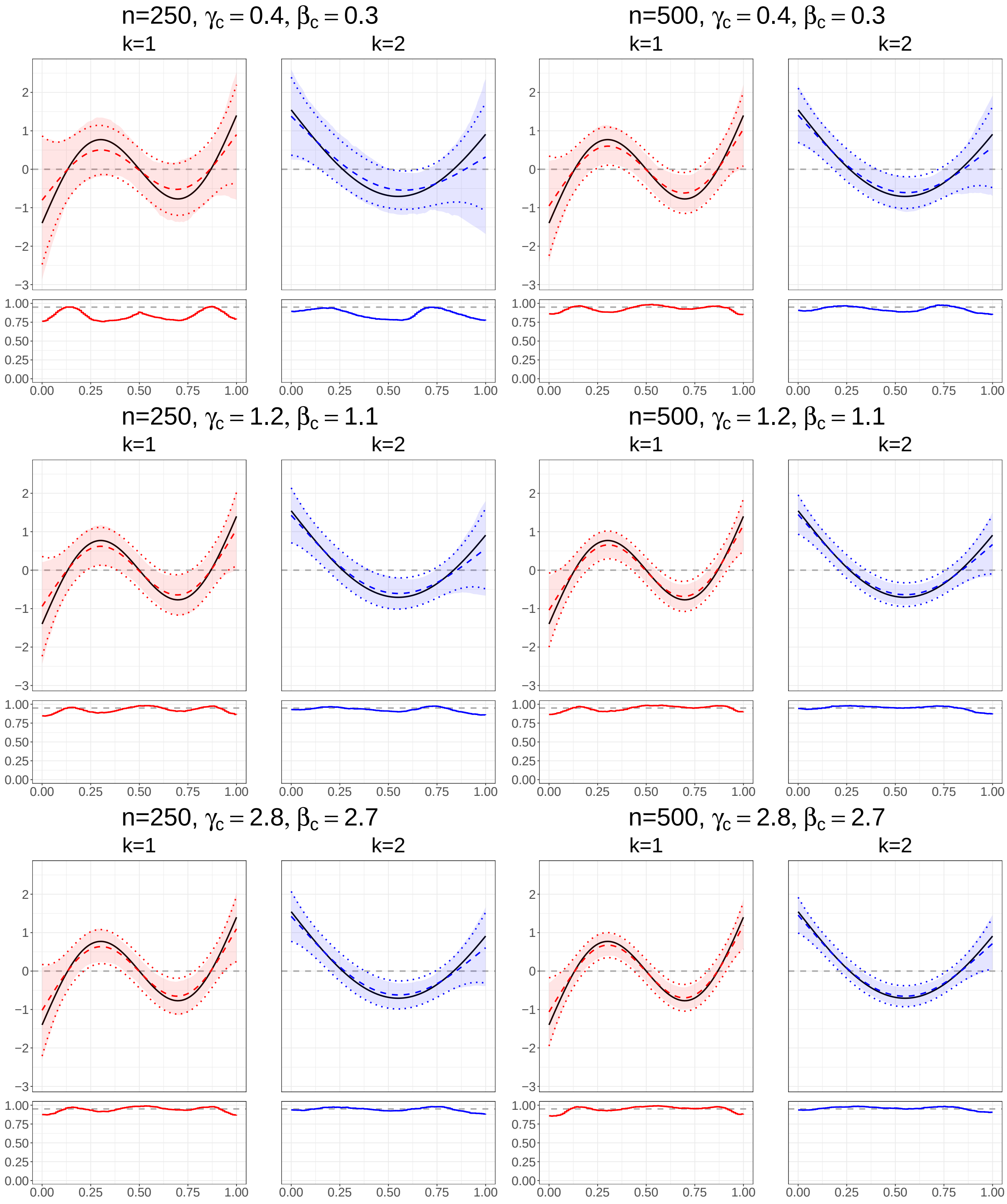

Figure 3.2: Simulation results for $\mathcal{X}_{[2]}$. For each simulation setting and each risk factor, the top panel shows $\eta^*_{k,\{2\}}$ (solid line), average of $\hat{\eta}_{k,\{2\}}$ (dashed line), average of the 95% Bayesian confidence interval for $\eta^*_{k,\{2\}}$ (dotted lines), and 2.5% and 97.5% empirical quantiles of $\hat{\eta}_{k,\{2\}}$ (shaded area). The bottom panel displays the empirical coverage rate of the 95% Bayesian confidence interval for $\eta^*_{k,\{2\}}$. Note that all averages and empirical quantiles are computed point-wise across all replicates.

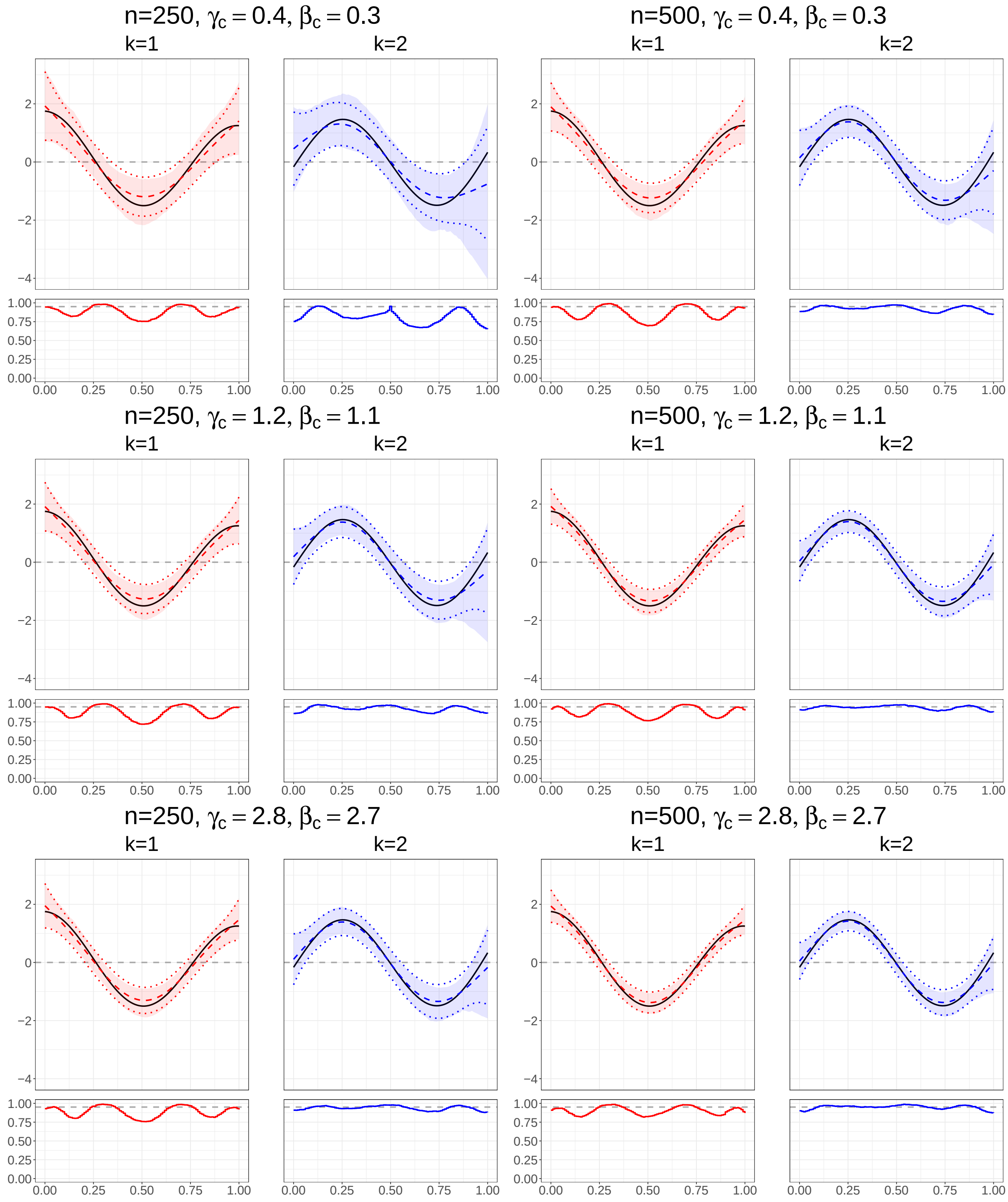

Figure 3.3: Simulation results for $\mathcal{X}_{[3]}$. For each simulation setting and each risk factor, the top panel shows $\eta^*_{k,\{3\}}$ (solid line), average of $\hat{\eta}_{k,\{3\}}$ (dashed line), average of the 95% Bayesian confidence interval for $\eta^*_{k,\{3\}}$ (dotted lines), and 2.5% and 97.5% empirical quantiles of $\hat{\eta}_{k,\{3\}}$ (shaded area). The bottom panel displays the empirical coverage rate of the 95% Bayesian confidence interval for $\eta^*_{k,\{3\}}$. Note that all averages and empirical quantiles are computed point-wise across all replicates.

3.4 Real Application

In this section, we conduct a real data analysis using the Multiple Myeloma Research Foundation (MMRF) CoMMpass study, which provides clinical trial data and gene expression data, publicly available through the Genomic Data Commons (GDC) portal. Multiple myeloma is a type of blood cancer that originates in plasma cells and affects the bone marrow and immune function. The dataset includes, for each subject, time-to-event or censoring, a censoring indicator, cause of death (categorized as multiple myeloma-related or non-cancer-related), age, and gene expression profiles for various genes.

Previous studies on time-to-event modeling in multiple myeloma have either used linear survival models [45, 49], which focus on variable selection or interpretation but cannot accommodate nonlinear effects, or machine learning approaches [41, 44], which aim for predictive performance but lack interpretability, especially with respect to how each covariate nonlinearly influences the event time.

The study by Zhao et al. [50], which focuses on survival modeling with omics covariates using The Cancer Genome Atlas (TCGA) data—also available through the GDC portal and formatted similarly to the MMRF CoMMpass data—mentions competing risks modeling as a potential approach for analyzing multiple causes of death, such as cancer-related and other-cause mortality. This suggestion motivates our use of competing risks modeling in the MMRF CoMMpass data analysis.

Despite the richness of the MMRF CoMMpass study, existing literature lacks smoothing spline modeling frameworks that can flexibly capture and interpret nonlinear covariate effects, as well as methods for jointly modeling multiple causes of death. Our proposed method addresses both limitations to analyze the MMRF CoMMpass data.

Under a competing risks framework, following previous studies on multiple myeloma [44, 45, 49], we focus on the effect of gene expression as covariates. Note that age is also an important factor [50]. In the MMRF CoMMpass dataset, using modified R code for data extraction as described in Zhao et al. [50], after excluding subjects with missing age information, the number of observations in the censored group and each risk group is as follows: $\sum_{i=1}^{n} \mathbb{1}(\mathfrak{I}_i = 0) = 577$, $\sum_{i=1}^{n} \mathbb{1}(G_i = 1) = 83$, and $\sum_{i=1}^{n} \mathbb{1}(G_i = 2) = 66$, where $k = 1$ corresponds to multiple myeloma-related death and $k = 2$ to non-multiple myeloma-related death, with $K = 2$ total causes. The high censoring rate in this dataset poses challenges for estimation and inference. To mitigate this, we apply matching [20] based on age to reduce the censoring rate and simultaneously control for age imbalance between censored and uncensored subjects.

After matching, the dataset is reduced to $\sum_{i=1}^{n} \mathbb{1}(\mathfrak{I}_i = 0) = 74$, $\sum_{i=1}^{n} \mathbb{1}(G_i = 1) = 83$, and $\sum_{i=1}^{n} \mathbb{1}(G_i = 2) = 66$. The original gene expression data contains measurements on 60,660 genes. We preprocess the gene expression data by (1) applying a logarithmic transformation and (2) scaling each variable to lie in the interval $[0, 1]$. The latter step alone resulted in covariate distributions that were too extreme.

To reduce dimensionality, we first select the top 25 genes with the highest standard deviation among those with less than 10% missing values. From this subset, we further select $d = 3$ covariates based on $\sum_{k=1}^{2} \|\hat{\eta}_k\|_{\mathcal{L}_2}$, computed via univariate modeling using the proposed method. The selected genes—IGKC ($j = 1$), UCHL1 ($j = 2$), and SHROOM3 ($j = 3$)—are listed in decreasing order of the screening criterion.

The final dataset, after excluding subjects with missing values for any of the three selected genes, consists of $\sum_{i=1}^{n} \mathbb{1}(\mathfrak{I}_i = 0) = 69$, $\sum_{i=1}^{n} \mathbb{1}(G_i = 1) = 82$, and $\sum_{i=1}^{n} \mathbb{1}(G_i = 2) = 62$, with $d = 3$ selected gene expression covariates as described above. We fit our proposed model to this dataset, assuming an additive structure on $\mathcal{H}$.

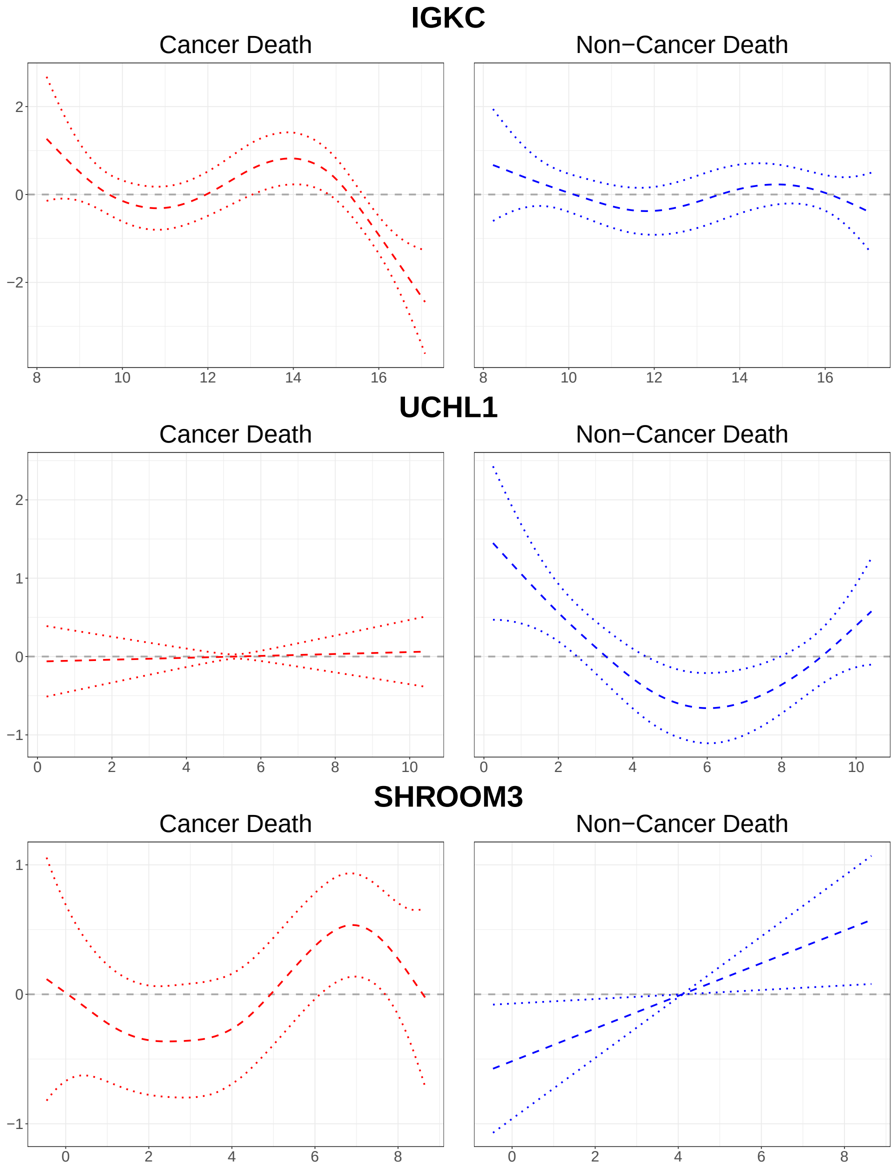

Figure 3.4: MMRF data analysis results. For each line, the title represents the gene and the subtitles correspond to the associated risk factors. For each plot, the dashed line represents $\hat{\eta}_{k,\{j\}}$ and the dotted lines indicate the 95% Bayesian confidence interval for $\eta^*_{k,\{j\}}$. Note that the domains are rescaled back to the logarithmic of gene expression for better interpretability.

In Figure 3.4, we observe that for each gene, the fitted curves $\hat{\eta}_{k,\{j\}}$ for multiple myeloma-related death and non-multiple myeloma-related death exhibit notably different shapes. This suggests the necessity of competing risk modeling for these risk factors. Additionally, for at least one of the risk factors, either multiple myeloma-related death or non-multiple myeloma-related death, nonlinearity is present in the fitted curves. For example, the cancer death group in IGKC and the non-cancer death group in UCHL1 exhibit pronounced nonlinearity in the fitted curve, where the Bayesian confidence interval does not include zero for a substantial portion of the domain. Although the Bayesian confidence interval for cancer death group in SHROOM3 is relatively wide, its fitted curve still displays nonlinearity. These findings are not reported in the existing literature.

Overall, the analysis demonstrates that for each gene expression, at least one cause-specific effect is nonlinear and varies across different causes of death. This highlights that the proposed model fills a gap in the existing literature, where competing risks are often overlooked or effects are assumed to be linear, providing a useful approach for analyzing MMRF CoMMpass dataset.

3.5 Conclusion

In this study, we introduced the smoothing spline competing risk Cox model, which combines the cause-specific Cox model with smoothing spline ANOVA to capture nonparametric covariate effects in competing risks. The proposed model bridges a gap in the literature by estimating nonlinear and distinct cause-specific covariate effects under the proportional cause-specific hazards assumption. For each risk factor, the estimation procedures are proposed, and the rate of convergence results are established. Our simulations and application to a multiple myeloma dataset demonstrate that the proposed model can estimate distinct effect functions across risk factors, in the presence of nonlinearity.

For future work, we plan to apply the smoothing spline ANOVA framework to various competing risk survival models. The proposed method in this paper combines the classical and fundamental proportional cause-specific hazards model with smoothing spline ANOVA. However, we aim to extend this framework by integrating alternative approaches to proportional cause-specific hazards modeling with smoothing spline ANOVA. For instance, we could combine the proportional subdistribution hazards model [15] with smoothing spline ANOVA or integrate correlated latent cause-specific event times model by copula structure [31] with smoothing spline ANOVA.

Bibliography

[1] Michael G. Akritas and Ingrid Van Keilegom. Estimation of bivariate and marginal distributions with censored data. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 65(2):457–471, 2003.

[2] Jan Beyersmann, Aurélien Latouche, Anika Buchholz, and Martin Schumacher. Simulating competing risks data in survival analysis. *Statistics in medicine*, 28:956–71, 03 2009.

[3] William Biscarri, Sihai Dave Zhao, and Robert J. Brunner. A simple and fast method for computing the Poisson binomial distribution function. *Computational Statistics & Data Analysis*, 122:92–100, 2018.

[4] N. Breslow. Covariance analysis of censored survival data. *Biometrics*, 30:89–99, 1974.

[5] Lucien Le Cam. An approximation theorem for the Poisson binomial distribution. *Pacific Journal of Mathematics*, 10:1181 – 1197, 1960.

[6] Youngjin Cho, Yili Hong, and Pang Du. An accurate computational approach for partial likelihood using poisson-binomial distributions. *Computational Statistics & Data Analysis*, 208:108161, 02 2025.

[7] Daniel Commenges, Jérémie Bureau, and Hein Putter. Inference with penalized likelihood, 2014.

[8] D. R. Cox. Regression models and life tables (with discussion). *Journal of the Royal Statistical Society, Series B*, 34:187–220, 1972.

[9] D. R. Cox. Partial likelihood. *Biometrika*, 62:269–276, 1975.

[10] Pang Du and Shuangge Ma. Frailty model with spline estimated nonparametric hazard function. *Statistica Sinica*, 20(2):561–580, 2010.

[11] Pang Du, Shuangge Ma, and Hua Liang. Penalized variable selection procedure for Cox models with semiparametric relative risk. *The Annals of Statistics*, 38(4):2092 – 2117, 2010.

[12] Pang Du, Yihua Jiang, and Yuedong Wang. Smoothing spline anova frailty model for recurrent event data. *Biometrics*, 67(4):1330–1339, 2011.

[13] Bradley Efron. The efficiency of Cox's likelihood function for censored data. *Journal of the American Statistical Association*, 72:557–565, 1977.

[14] Jean-David Fermanian. Nonparametric estimation of competing risks models with covariates. *Journal of Multivariate Analysis*, 85(1):156–191, 2003.

[15] Jason P. Fine and Robert J. Gray. A proportional hazards model for the subdistribution of a competing risk. *Journal of the American Statistical Association*, 94(446):496–509, 1999.

[16] T. R. Fleming and D. P. Harrington. *Counting Processes and Survival Analysis*. John Wiley & Sons, Inc, Hoboken, New Jersey, 1991.

[17] Roger Fletcher. *Practical Methods of Optimization*. John Wiley & Sons, New York, NY, USA, second edition, 1987.

[18] Chong Gu. *Smoothing Spline ANOVA Models*. Springer Series in Statistics. Springer, New York, 2nd edition, 2013.

[19] Irva Hertz-Picciotto and Beverly Rockhill. Validity and efficiency of approximation methods for tied survival times in Cox regression. *Biometrics*, 53:1151–1156, 1997.

[20] Daniel E. Ho, Kosuke Imai, Gary King, and Elizabeth A. Stuart. Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis*, 15(3):199–236, 2007.

[21] Y. Hong. On computing the distribution function for the Poisson bionomial distribution. *Computational Statistics and Data Analysis*, 59:41–51, 2013.

[22] Yili Hong. *poibin: The Poisson Binomial Distribution*. Virginia Tech, 2020. R package version 1.5.

[23] Pierre Joly and Daniel Commenges. A penalized likelihood approach for a progressive three-state model with censored and truncated data: Application to aids. *Biometrics*, 55(3):887–890, 1999.

[24] Pierre Joly, Daniel Commenges, and Luc Letenneur. A penalized likelihood approach for arbitrarily censored and truncated data: Application to age-specific incidence of dementia. *Biometrics*, 54(1):185–194, 1998.

[25] Pierre Joly, Daniel Commenges, Catherine Helmer, and Luc Letenneur. A penalized likelihood approach for an illness-death model with interval-censored data: Application to age-specific incidence of dementia. *Biostatistics*, 3:433–43, 2002.

[26] J. D. Kalbfleisch and R. L. Prentice. Marginal likelihoods based on Cox's regression and life model. *Biometrika*, 60:267–278, 1973.

[27] John D. Kalbfleisch and Ross L. Prentice. *The Statistical Analysis of Failure Time Data*. John Wiley and Sons Inc, 2nd edition, 2002.

[28] O Kardaun. Statistical survival analysis of male larynx-cancer patients - a case study. *Statistica Neerlandica*, 37:103–125, 1983.

[29] John P. Klein and Melvin L. Moeschberger. *Survival Analysis: Techniques for Censored and Truncated Data*. Springer, New York, second edition, 2003.

[30] Zhengzhi Lin, Yueyao Wang, and Yili Hong. The computing of the Poisson multinomial distribution and applications in ecological inference and machine learning. *Computational Statistics*, 38:1851–1877, 2023.

[31] Simon M. S. Lo and Ralf A. Wilke. A copula model for dependent competing risks. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 59(2):359–376, 2010.

[32] CL Loprinzi, JA Laurie, HS Wieand, JE Krook, PJ Novotny, JW Kugler, J Bartel, M Law, M Bateman, and NE Klatt. Prospective evaluation of prognostic variables from patient-completed questionnaires. north central cancer treatment group. *Journal of Clinical Oncology: official journal of the American Society of Clinical Oncology*, 12 (3):601–607, 1994.

[33] Mary Lunn and Don Mcneil. Applying Cox regression to competing risks. *Biometrics*, 51:524–32, 1995.

[34] Sirong Luo, Xiao Kong, and Tingting Nie. Spline based survival model for credit risk modeling. *European Journal of Operational Research*, 253(3):869–879, 2016.

[35] Ana López-Cheda, Ricardo Cao, M. Amalia Jácome, and Ingrid Van Keilegom. Nonparametric incidence estimation and bootstrap bandwidth selection in mixture cure models. *Computational Statistics & Data Analysis*, 105:144–165, 2017.

[36] Ana López-Cheda, Yingwei Peng, and M. Jácome. Nonparametric estimation in mixture cure models with covariates. *TEST*, 32, 2023.

[37] Karla Monterrubio-Gómez, Nathan Constantine-Cooke, and Catalina A. Vallejos. A review on statistical and machine learning competing risks methods. *Biometrical Journal*, 66(2):2300060, 2024.

[38] Finbarr O'Sullivan. Nonparametric Estimation in the Cox Model. *The Annals of Statistics*, 21(1):124 – 145, 1993.

[39] R. L. Prentice, J. D. Kalbfleisch, A. V. Peterson, Jr., N. Flournoy, V. T. Farewell, and N. E. Breslow. The analysis of failure times in the presence of competing risks. *Biometrics*, 34:541–554, 1978.

[40] Ross L. Prentice and John D. Kalbfleisch. Mixed discrete and continuous cox regression model. *Lifetime Data Analysis*, 9:195–210, 2003.

[41] Liang Ren, Bei Xu, Jiadai Xu, Jing Li, Jifeng Jiang, Yuhong Ren, and Peng Liu. A machine learning model to predict survival and therapeutic responses in multiple myeloma. *International Journal of Molecular Sciences*, 24(7), 2023.

[42] Virginie Rondeau, Daniel Commenges, and Pierre Joly. Maximum penalized likelihood estimation in a gamma-frailty model. *Lifetime data analysis*, 9:139–53, 2003.

[43] Thomas H. Scheike and Yanqing Sun. Maximum likelihood estimation for tied survival data under Cox's regression model via the EM algorithm. *Lifetime Data Analysis*, 13: 339–420, 2007.

[44] Krish U. Shah, Kaina A. Millan, Anna E. Pula, Tadeusz F. Kubicki, Joseph Cannova, Sulin Wu, Manoj Bhagwat, Quincy C. Guenther, Jennifer Cooperrider, Gregory Roloff, Aarti Venkat, Benjamin A. Derman, Andrzej J. Jakubowiak, and Michael W. Drazer. Rnaseq-based machine learning models for prognostication of multiple myeloma. *medRxiv*, 2025.

[45] Anish Simhal, Kylee Maclachlan, Rena Elkin, Jiening Zhu, Larry Norton, Joseph Deasy, Jung Hun Oh, Saad Usmani, and Allen Tannenbaum. Gene interaction network analysis in multiple myeloma detects complex immune dysregulation associated with shorter survival. *Blood Cancer Journal*, 13, 11 2023.

[46] Krithika Suresh, Cameron Severn, and Debashis Ghosh. Survival prediction models: an introduction to discrete-time modeling. *BMC Medical Research Methodology*, 22, 07 2022.

[47] Terry M. Therneau and Patricia M. Grambsch. *Modeling Survival Data: Extending the Cox Model*. Springer, New York, 2000.

[48] Lu Wang, Pang Du, and Hua Liang. Two-component mixture cure rate model with spline estimated nonparametric components. *Biometrics*, 68(3):726–735, 2011.

[49] Cuiling Zhang, Sungui Wu, and Bing Chen. A novel prognostic model based on pyroptosis-related genes for multiple myeloma. *BMC Medical Genomics*, 16, 02 2023.

[50] Zhi Zhao, John Zobolas, Manuela Zucknick, and Tero Aittokallio. Tutorial on survival modeling with applications to omics data. *Bioinformatics*, 40(3):btae132, 03 2024.

Appendices

Appendix A

Details for "An Accurate Computational Approach for Partial Likelihood Using Poisson-Binomial Distributions"

A.1 Existing APL Approximation Methods

When there are no ties (i.e., d_j=1), the APL is

$$L_j(\boldsymbol{\beta}, \lambda_j) = \frac{A_j(\boldsymbol{\beta}, \lambda_j)}{B_j(\boldsymbol{\beta}, \lambda_j)} = \frac{p_{j_1,j}(\boldsymbol{\beta}, \lambda_j) \prod_{i \in \mathcal{R}(t_{(j)}) \setminus \{j_1\}} (1 - p_{ij}(\boldsymbol{\beta}, \lambda_j))}{\sum_{i \in \mathcal{R}(t_{(j)})} \left\{ p_{ij} \prod_{\ell \in \mathcal{R}(t_{(j)}) \setminus \{i\}} (1 - p_{\ell j}) \right\}}.$$

When λ_j is small enough, say $\sup_j \lambda_j = \mathcal{O}_{\mathrm{P}}(\tau)$, $p_{ij} \approx \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta}) \lambda_j$ by the Taylor expansion. Then $\sup_{ij} p_{ij} = \mathcal{O}_{\mathrm{P}}(\tau)$, $1 - p_{ij} \approx 1$, and

$$L_j(\boldsymbol{\beta}, \lambda_j) \approx \frac{\exp(\boldsymbol{x}_{j_1}^\top \boldsymbol{\beta}) \lambda_j}{\sum_{i \in \mathcal{R}(t_{(j)})} \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta}) \lambda_j} = \frac{\exp(\boldsymbol{x}_{j_1}^\top \boldsymbol{\beta})}{\sum_{i \in \mathcal{R}(t_{(j)})} \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta})}, \tag{A.1}$$

which is the approximate partial likelihood (2.2) used in literature [8].

When there are ties, the Cox correction [8] uses the following partial likelihood term,

$$L_j(\boldsymbol{\beta}, \lambda_j) \approx \frac{\exp\left(\sum_{i \in \mathcal{D}(t_{(j)})} \boldsymbol{x}_i^\top \boldsymbol{\beta}\right)}{\sum_{\mathcal{A}_{d_j} \in \mathcal{F}_{d_j}} \exp\left(\sum_{i \in \mathcal{A}_{d_j}} \boldsymbol{x}_i^\top \boldsymbol{\beta}\right)}. \tag{A.2}$$

We have d_j ties at $t_{(j)}$, and subject $j_1, \ldots, j_{d_j}$ failed at $t_{(j)}$. Let Q_j be the set of $d_j!$ permutations of the subjects $j_1, \ldots, j_{d_j}$. Let $S_l^j = \{s_{l1}^j, \ldots, s_{ld_j}^j\}$ be the lth element in Q_j. We have $Q_j = \{S_1^j, \ldots, S_{d_j!}^j\}$ and $\mathcal{R}(t_{(j)}, S_l^j, m) = \mathcal{R}(t_{(j)}) \setminus \{s_{l1}^j, \ldots, s_{l,m-1}^j\}$. The Kalbfleisch-Prentice correction [26] uses the following partial likelihood term,

$$L_j(\boldsymbol{\beta}, \lambda_j) \approx \frac{1}{d_j!} \sum_{l=1}^{d_j!} \frac{\exp\left(\sum_{i \in \mathcal{D}(t_{(j)})} \boldsymbol{x}_i^\top \boldsymbol{\beta}\right)}{\prod_{m=1}^{d_j} \left\{\sum_{i \in \mathcal{R}(t_{(j)}, S_l^j, m)} \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta})\right\}}, \tag{A.3}$$

which is based on the average partial likelihood contribution at $t_{(j)}$.

The Breslow correction [4] uses the following partial likelihood term,

$$L_j(\boldsymbol{\beta}, \lambda_j) \propto \frac{\exp\left(\sum_{i \in \mathcal{D}(t_{(j)})} \boldsymbol{x}_i^\top \boldsymbol{\beta}\right)}{\left\{\sum_{i \in \mathcal{R}(t_{(j)})} \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta})\right\}^{d_j}}. \tag{A.4}$$

We denote by $\widehat{\boldsymbol{\beta}}_{\mathrm{b}}$ and $\widehat{\lambda}_{\mathrm{b}j} = d_j / \{\sum_{i \in \mathcal{R}(t_{(j)})} \exp(\boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}}_{\mathrm{b}})\}$ the Breslow estimates for $\boldsymbol{\beta}$ and λ_j, respectively, where $\widehat{\Lambda}_{\mathrm{b}}(t) = \sum_{j=1}^k \widehat{\lambda}_{\mathrm{b}j} \mathbb{1}(t_{(j)} \leq t)$.

The Efron correction [13] uses the following partial likelihood term,

$$L_j(\boldsymbol{\beta}, \lambda_j) \propto \frac{\exp\left(\sum_{i \in \mathcal{D}(t_{(j)})} \boldsymbol{x}_i^\top \boldsymbol{\beta}\right)}{\prod_{\ell=0}^{d_j-1} \left\{\sum_{i \in \mathcal{R}(t_{(j)})} \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta}) - \ell \overline{A}\left(\boldsymbol{\beta}, t_{(j)}\right)\right\}}, \tag{A.5}$$

where $\overline{A}(\boldsymbol{\beta}, t_{(j)}) = \sum_{i \in \mathcal{D}(t_{(j)})} \exp\left(\boldsymbol{x}_i^\top \boldsymbol{\beta}\right) / d_j$. We denote the Efron estimates for $\boldsymbol{\beta}$ and λ_j by $\widehat{\boldsymbol{\beta}}_{\mathrm{e}}$ and $\widehat{\lambda}_{\mathrm{e}j}$, respectively. In particular, $\widehat{\lambda}_{\mathrm{e}j} = \sum_{\gamma=0}^{d_j-1} \left\{ \sum_{i \in \mathcal{R}(t_{(j)})} \exp(\boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}}_{\mathrm{e}}) - \gamma \overline{A}(\widehat{\boldsymbol{\beta}}_{\mathrm{e}}, t_{(j)}) \right\}^{-1}$, where $\widehat{\Lambda}_{\mathrm{e}}(t) = \sum_{j=1}^{k} \widehat{\lambda}_{\mathrm{e}j} \mathbb{1}(t_{(j)} \leq t)$. The Cox correction, the Kalbfleisch-Prentice correction, the Breslow correction, and the Efron correction are all the same as the approximate partial likelihood (2.2) when $d_j = 1$.

A.2 Asymptotic Notations

Here we introduce some asymptotic notations. Without loss of generality, C or $\mathcal{C}$, including versions with subscripts, represent constants in $(0, \infty)$. To avoid confusion with censoring values denoted by C_i, we refrain from using numerical subscripts for C when representing constants. We denote by $\|A\|_\infty$ the infinite norm of a scalar, vector, or matrix. Let X_n and Y_n be random scalar, vector, or matrix. We write $X_n = \mathcal{O}_{\mathrm{P}}(Y_n)$ if $\Pr(\|X_n\|_\infty > C\|Y_n\|_\infty) \to 0$ as $n \to \infty$ for some $C > 0$ and $X_n = o_{\mathrm{P}}(Y_n)$ when $\Pr(\|X_n\|_\infty > \epsilon\|Y_n\|_\infty) \to 0$ as $n \to \infty$ for all $\epsilon > 0$. One can see that $\|\cdot\|_\infty$ in definitions of $\mathcal{O}_{\mathrm{P}}(\cdot)$ or $o_{\mathrm{P}}(\cdot)$ can be substituted by $\|\cdot\|_2$ for vectors with finite, fixed, and therefore non-diverging dimensions as $n \to \infty$, i.e., $\|a\|_\infty \leq \|a\|_2 \leq \sqrt{m}\|a\|_\infty$ for $a \in \mathbb{R}^m$. Let $\xrightarrow{\mathrm{D}}$, $\xrightarrow{\mathrm{P}}$, and $\xrightarrow{\mathrm{a.s.}}$ stand for convergence in distribution, convergence in probability, and almost sure convergence, respectively. Note that summation over empty set is set to be 0 and product over empty set is set to be 1.

A.3 Assumptions

Recall that ζ is the ending time of the failure time study, which we assume, for the sake of notational simplicity, to be a multiple of the grouping parameter τ.

Assumption A.3.1. *Recall that $n(t)$ and $d(t)$ are respectively the numbers of at-risk subjects and failed subjects at time t. We assume $n(\zeta) - d(\zeta) \geq 1$.*

Note that Assumption A.3.1 is basically saying that there is at least one censored subjects in the end of the study, which holds for most datasets. This assumption basically allows $n(\zeta)$ or $n(\zeta) - d(\zeta)$ appearing as the denominators of fractions.

Assumption A.3.2. *The covariates* $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$ *are deterministic and are elements of some compact set* $\mathcal{M}$.

Assumption A.3.2 implies $\sup_i \|\boldsymbol{x}_i\|_\infty \le \rho$ for some $\rho > 0$, or $\sup_i \|\boldsymbol{x}_i\|_\infty = \mathcal{O}(1)$, that is, there are no extreme outliers in the model. This is commonly assumed in the survival analysis literature.

Assumption A.3.3. *We assume* $\sup_{t\in\Omega} d\Lambda_0^*(t) = \mathcal{O}(\tau)$, *where* $\Lambda_0^*(\cdot)$ *is the discretized baseline cumulative hazard function defined in* (2.7).

One can see that Assumption A.3.3 can be generally satisfied for common cumulative hazard functions. One example is the cumulative hazard function of the Weibull distribution ($\Lambda_0(t) = gt^m$) with the parameters $g > 0$ and $m \ge 1$. To check this, one can see that $d\Lambda_0^*(t)/\tau$ is non-decreasing function on Ω and

$$\frac{d\Lambda_0^*(\zeta)}{\tau} = \frac{g\zeta^m - g(\zeta-\tau)^m}{\tau} = \frac{g(\zeta-\tau)^m - g\zeta^m}{-\tau} \to \frac{\partial}{\partial\zeta} g\zeta^m = gm\zeta^{m-1}$$

as $\tau \to 0$ as $n \to \infty$. So, $0 \le d\Lambda_0^*(\tau)/\tau \le d\Lambda_0^*(2\tau)/\tau \le \cdots \le d\Lambda_0^*(\zeta)/\tau = \{\sup_{t\in\Omega} d\Lambda_0^*(t)\}/\tau = \mathcal{O}(1)$, directly implying $\sup_{t\in\Omega} d\Lambda_0^*(t) = \mathcal{O}(\tau)$. To see point-wise boundedness for some individual $t \in \Omega$, as $d\Lambda_0^*(\zeta)/\tau \to gm\zeta^{m-1}$ as $\tau \to 0$ by $n \to \infty$, $d\Lambda_0^*(\zeta) \asymp \tau$, which is also $\mathcal{O}(\tau)$, and $d\Lambda_0^*(\tau) = g\tau^m$ is $o(\tau)$ for $m > 1$ and $\asymp \tau$ for m=1, which are also $\mathcal{O}(\tau)$.

Assumption A.3.4. *We assume* $\sup_{t\in[0,\zeta]} |n(t)/n - y(t)| \xrightarrow{a.s.} 0$ *as* $n \to \infty$ *with* $y(t) \in (0, 1]$ *for all* $t \in [0, \zeta]$.

Assumption A.3.4 means that there are always a nonzero fraction of subjects at risk even throughout the study. This is a reasonable assumption since most survival data contain a good portion of censored subjects at the end of the study.

Assumption A.3.5. *We assume* $\sup_{t\in\Omega} dN(t) = \mathcal{O}_P\left(\sup_{t\in\Omega}\{\mathrm{E}(dN(t)|\mathcal{F}_{t^-})\}\right).$

Assumption A.3.5, together with Assumptions A.3.1 – A.3.4, indicates that

$$\sup_{t\in\Omega}\{\mathrm{E}(dN(t)|\mathcal{F}_{t^-})\} = \sup_{t\in\Omega} dA(t) = \sup_{t\in\Omega}\left\{\sum_{i=1}^{n} Y_i(t)d\Lambda^*(t;\boldsymbol{x}_i)\right\} = \mathcal{O}_{\mathrm{P}}(n\tau),$$

implying that $\sup_{t\in\Omega} d(t) = \sup_{t\in\Omega} dN(t) = \mathcal{O}_{\mathrm{P}}(n\tau)$. Considering the fact that the number of elements in Ω is ζ/τ and the number of total events is $\mathcal{O}_{\mathrm{P}}(n)$, it is natural that $\sup_{t\in\Omega} dN(t) = \mathcal{O}_{\mathrm{P}}(n/(\zeta/\tau)) = \mathcal{O}_{\mathrm{P}}(n\tau)$, which means the number of ties are balanced for all $t \in \Omega$, i.e., no extreme number of ties for some $t \in \Omega$.

For martingales $G(\cdot)$ and $H(\cdot)$, we define predictable variation process as $\langle G\rangle(t) = \int_0^t d\langle G\rangle(s)$, where $d\langle G\rangle(t) = \mathrm{Var}(dG(t)|\mathcal{F}_{t^-})$ and the predictable covariation process as $\langle G, H\rangle(t) = \int_0^t d\langle G, H\rangle(s)$, where $d\langle G, H\rangle(t) = \mathrm{Cov}(dG(t),\, dH(t)|\mathcal{F}_{t^-})$.

Assumption A.3.6. *We assume*

$$\mathrm{E}(dN_i(t)dN_l(t)|\mathcal{F}_{t^-}) = \mathrm{E}(dN_i(t)|\mathcal{F}_{t^-})\,\mathrm{E}(dN_l(t)|\mathcal{F}_{t^-}),$$

when $i \neq l$ *for all* $t \in [0, \zeta]$.

Assumption A.3.6 is common in the survival analysis literature. One can see that it implies $d\langle M_i, M_l\rangle(t) = 0$, when $i \neq l$ for all $t \in [0, \zeta]$.

Define the processes

$$
\begin{aligned}
S^{(0)}(\boldsymbol{\beta}, t) &= \sum_{i=1}^{n} Y_i(t) \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta}), \\
S^{(1)}(\boldsymbol{\beta}, t) &= \frac{\partial}{\partial \boldsymbol{\beta}} S^{(0)}(\boldsymbol{\beta}, t) = \sum_{i=1}^{n} Y_i(t) \boldsymbol{x}_i \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta}), \text{ and} \\
S^{(2)}(\boldsymbol{\beta}, t) &= \frac{\partial^2}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\top} S^{(0)}(\boldsymbol{\beta}, t) = \sum_{i=1}^{n} Y_i(t) \boldsymbol{x}_i \boldsymbol{x}_i^\top \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta}).
\end{aligned}
$$

Assumption A.3.7. *There exists an open convex neighborhood $\mathcal{B}$ of $\boldsymbol{\beta}$ and functions $s^{(j)}(\cdot, \cdot)$, $j = 0, 1, 2$ defined on $\mathcal{B} \times [0, \zeta]$, that satisfy the following four conditions.*

(i) $\int_0^\zeta d\Lambda_0(t) < \infty$.

(ii) $\sup_{\tilde{\boldsymbol{\beta}} \in \mathcal{B}, t \in [0,\zeta]} \|S^{(j)}(\tilde{\boldsymbol{\beta}}, t)/n - s^{(j)}(\tilde{\boldsymbol{\beta}}, t)\|_\infty \xrightarrow{P} 0$ *as* $n \to \infty$ *for* $j = 0, 1, 2$,

(iii) $s^{(0)}(\tilde{\boldsymbol{\beta}}, t) > m_s > 0$ *for all* $\tilde{\boldsymbol{\beta}} \in \mathcal{B}$, $t \in [0, \zeta]$ *and* $\sup_{\tilde{\boldsymbol{\beta}} \in \mathcal{B}, t \in [0,\zeta]} \|s^{(j)}(\tilde{\boldsymbol{\beta}}, t)\|_\infty \leq C_S$ *for* $j = 0, 1, 2$ *for some* $C_S > 0$,

(iv) For $j = 0, 1, 2$, $s^{(j)}(\tilde{\boldsymbol{\beta}}, t)$ *are continuous functions of* $\tilde{\boldsymbol{\beta}} \in \mathcal{B}$ *uniformly in* $t \in [0, \zeta]$, *and* $s^{(1)}(\tilde{\boldsymbol{\beta}}, t) = \partial s^{(0)}(\tilde{\boldsymbol{\beta}}, t)/(\partial \tilde{\boldsymbol{\beta}})$, $s^{(2)}(\tilde{\boldsymbol{\beta}}, t) = \partial^2 s^{(0)}(\tilde{\boldsymbol{\beta}}, t)/\ (\partial \tilde{\boldsymbol{\beta}} \partial \tilde{\boldsymbol{\beta}}^\top)$ *for all* $\tilde{\boldsymbol{\beta}} \in \mathcal{B}, t \in [0, \zeta]$, *and*

(v) $\Sigma(\tilde{\boldsymbol{\beta}}, \zeta) = \int_0^\zeta v(\tilde{\boldsymbol{\beta}}, t) s^{(0)}(\tilde{\boldsymbol{\beta}}, t) d\Lambda_0(t)$ *is positive definite for all* $\tilde{\boldsymbol{\beta}} \in \mathcal{B}$, *where* $e(\tilde{\boldsymbol{\beta}}, t) = s^{(1)}(\tilde{\boldsymbol{\beta}}, t)/s^{(0)}(\tilde{\boldsymbol{\beta}}, t)$ *and* $v(\tilde{\boldsymbol{\beta}}, t) = s^{(2)}(\tilde{\boldsymbol{\beta}}, t)/s^{(0)}(\tilde{\boldsymbol{\beta}}, t) - e(\tilde{\boldsymbol{\beta}}, t) e(\tilde{\boldsymbol{\beta}}, t)^\top$ *for all* $\tilde{\boldsymbol{\beta}} \in \mathcal{B}, t \in [0, \zeta]$.

Assumption A.3.7 is common in the survival analysis literature. See, for example, Section 8 of Fleming and Harrington [16]. Without loss of generality, there exists $C_{\mathcal{B}} > 0$ such that $\|\tilde{\boldsymbol{\beta}}\|_2 \leq C_{\mathcal{B}}$ for all $\tilde{\boldsymbol{\beta}} \in \mathcal{B}$.

Assumption A.3.8. *Let* $\tau \asymp 1/n$, *which means* $\tau = \mathcal{O}(1/n)$ *and* $1/n = \mathcal{O}(\tau)$.

Assumption A.3.8, together with Assumptions A.3.1 – A.3.5, indicates $\sup_{t\in\Omega} d(t) = \mathcal{O}_{\mathrm{P}}(1)$, which means the supremum number of ties does not diverge with n.

Assumption A.3.9. *For all* $t \in \Omega$, $d\widehat{\Lambda}(t) = d(t)/n^*(t)$ *for some non-negative and non-increasing function* $n^*(t)$ *over* $t \in [0, \zeta]$, *where* $\sup_{t\in\Omega} n^*(t) = \mathcal{O}_P(n)$ *and* $\sup_{t\in\Omega}\{1/n^*(t)\} = \mathcal{O}_P(1/n)$.

Under Assumptions A.3.1 – A.3.5 and A.3.9, one can show that $\sup_{t\in\Omega} d\widehat{\Lambda}(t) = \mathcal{O}_{\mathrm{P}}(\tau)$. One can see that $d\widehat{\Lambda}_{\mathrm{b}}(\cdot)$ and $d\widehat{\Lambda}_{\mathrm{na}}(\cdot)$ satisfy Assumption A.3.9, where $\widehat{\Lambda}_{\mathrm{na}}(t) = \sum_{j=1}^{k} \widehat{\lambda}_{\mathrm{na},j}\mathbb{1}(t_{(j)} \le t)$. See Section A.4.8. Although there is no rigorous proof that $d\widehat{\Lambda}_{\mathrm{e}}(\cdot)$ satisfies Assumption A.3.9, as the Efron correction is a modification of the Breslow correction and performance of $\widehat{\boldsymbol{\Lambda}}_{\mathrm{e}}$ in Figure A.2 is better than the other baseline hazard function estimators in reducing bias of $\widehat{\boldsymbol{\beta}}_{\mathrm{pb}}$, we use $d\widehat{\Lambda}_{\mathrm{e}}(\cdot)$ as $d\widehat{\Lambda}(\cdot)$ in our numerical examples in Sections 2.5 and 2.6.

Assumption A.3.10. $X_{pb}(\cdot, \widehat{\boldsymbol{\Lambda}}, \zeta)$ *and* $X(\cdot, \zeta)$ *are strictly concave on* $\mathcal{B}$ *and there exist* $a \in [0, 1/2)$, $C_{pb} > 0$, *and* $C_b > 0$ *satisfying the following as* $n \to \infty$*:*

(i) $\Pr\left(\|\widehat{\boldsymbol{\beta}}_{pb} - \widehat{\boldsymbol{\beta}}_b\|_2 \le C_{pb} n^a \left\{X_{pb}(\widehat{\boldsymbol{\beta}}_{pb}, \widehat{\boldsymbol{\Lambda}}, \zeta) - X_{pb}(\widehat{\boldsymbol{\beta}}_b, \widehat{\boldsymbol{\Lambda}}, \zeta)\right\}\right) \to 1.$

(ii) $\Pr\left(\|\widehat{\boldsymbol{\beta}}_b - \widehat{\boldsymbol{\beta}}_{pb}\|_2 \le C_b n^a \left\{X(\widehat{\boldsymbol{\beta}}_b, \zeta) - X(\widehat{\boldsymbol{\beta}}_{pb}, \zeta)\right\}\right) \to 1.$

One can see that for any $\tilde{\boldsymbol{\beta}} \in \mathcal{B}$, we have $X_{\mathrm{pb}}(\widehat{\boldsymbol{\beta}}_{\mathrm{pb}}, \widehat{\boldsymbol{\Lambda}}, \zeta) - X_{\mathrm{pb}}(\tilde{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Lambda}}, \zeta) \ge 0$ and $X(\widehat{\boldsymbol{\beta}}_{\mathrm{b}}, \zeta) - X(\tilde{\boldsymbol{\beta}}, \zeta) \ge 0$ by the definition of $\widehat{\boldsymbol{\beta}}_{\mathrm{pb}}$ and $\widehat{\boldsymbol{\beta}}_{\mathrm{b}}$. Due to strict concavity, $X_{\mathrm{pb}}(\widehat{\boldsymbol{\beta}}_{\mathrm{pb}}, \widehat{\boldsymbol{\Lambda}}, \zeta) - X_{\mathrm{pb}}(\tilde{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Lambda}}, \zeta) = 0$ only when $\tilde{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}_{\mathrm{pb}}$, which also makes $\|\widehat{\boldsymbol{\beta}}_{\mathrm{pb}} - \tilde{\boldsymbol{\beta}}\|_2 = 0$. Similarly, $X(\widehat{\boldsymbol{\beta}}_{\mathrm{b}}, \zeta) - X(\tilde{\boldsymbol{\beta}}, \zeta) = 0$ only when $\tilde{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}_{\mathrm{b}}$, which also makes $\|\widehat{\boldsymbol{\beta}}_{\mathrm{b}} - \tilde{\boldsymbol{\beta}}\|_2 = 0$. Assumption A.3.10 essentially requires that the difference between $\widehat{\boldsymbol{\beta}}_{\mathrm{pb}}$ and $\widehat{\boldsymbol{\beta}}_{\mathrm{b}}$ is bounded by both the difference between their values in $n^a X_{\mathrm{pb}}(\cdot, \widehat{\boldsymbol{\Lambda}}, \zeta)$ and the difference between their values in $n^a X(\cdot, \zeta)$. Considering the fact that $\widehat{\boldsymbol{\beta}}_{\mathrm{pb}}, \widehat{\boldsymbol{\beta}}_{\mathrm{b}} \in \mathcal{B}$, we have $\|\widehat{\boldsymbol{\beta}}_{\mathrm{pb}} - \widehat{\boldsymbol{\beta}}_{\mathrm{b}}\|_2 = \mathcal{O}_{\mathrm{P}}(1)$. Also, one can show that $n^a\{X_{\mathrm{pb}}(\widehat{\boldsymbol{\beta}}_{\mathrm{pb}}, \widehat{\boldsymbol{\Lambda}}, \zeta) - X_{\mathrm{pb}}(\widehat{\boldsymbol{\beta}}_{\mathrm{b}}, \widehat{\boldsymbol{\Lambda}}, \zeta)\} = \mathcal{O}_{\mathrm{P}}(n^a)$ and $n^a\{X(\widehat{\boldsymbol{\beta}}_{\mathrm{b}}, \zeta) - X(\widehat{\boldsymbol{\beta}}_{\mathrm{pb}}, \zeta)\} = \mathcal{O}_{\mathrm{P}}(n^a)$. So this assumption is expected to be satisfied in many scenarios.

For $j = 0, 1, 2$, define the processes $\mathcal{S}^{(j)}(\boldsymbol{\beta}, t)$ similar to $S^{(j)}(\boldsymbol{\beta}, t)$ with the at-risk process $Y_i(t)$ replaced by $\mathcal{Y}_i(t) = \mathbb{1}(T_i \geq t)$ for the continuous model, where $T_i = \min(\tilde{T}_i, C_i)$.

Assumption A.3.11. *Let* $\sum_{i=1}^n \mathcal{Y}_i(\zeta) - 1 \geq 1$ *and* $\sup_{t\in[0,\zeta]} |\sum_{i=1}^n \mathcal{Y}_i(t)/n - \mathscr{Y}(t)| \xrightarrow{a.s.} 0$ *as* $n \to \infty$ *with* $\mathscr{Y}(t) \in (0, 1]$ *for all* $t \in [0, \zeta]$. *There exists an open convex neighborhood* $\mathcal{B}$ *of* $\boldsymbol{\beta}$ *and functions* $\mathscr{S}^{(j)}(\cdot, \cdot)$, $j = 0, 1, 2$ *defined on* $\mathcal{B} \times [0, \zeta]$, *that satisfy the following four conditions.*

(i) $\int_0^\zeta d\Lambda_0(t) < \infty$.

(ii) $\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B}, t\in[0,\zeta]} \|\mathcal{S}^{(j)}(\tilde{\boldsymbol{\beta}}, t)/n - \mathscr{S}^{(j)}(\tilde{\boldsymbol{\beta}}, t)\|_\infty \xrightarrow{P} 0$ *as* $n \to \infty$ *for* $j = 0, 1, 2$,

(iii) $\mathscr{S}^{(0)}(\tilde{\boldsymbol{\beta}}, t) > m_{\mathscr{S}} > 0$ *for all* $\tilde{\boldsymbol{\beta}} \in \mathcal{B}$, $t \in [0, \zeta]$ *and* $\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B}, t\in[0,\zeta]} \|\mathscr{S}^{(j)}(\tilde{\boldsymbol{\beta}}, t)\|_\infty \leq C_{\mathscr{S}}$ *for* $j = 0, 1, 2$ *for some* $C_{\mathscr{S}} > 0$,

(iv) For $j = 0, 1, 2$, $\mathscr{S}^{(j)}(\tilde{\boldsymbol{\beta}}, t)$ *are continuous functions of* $\tilde{\boldsymbol{\beta}} \in \mathcal{B}$ *uniformly in* $t \in [0, \zeta]$, *and* $\mathscr{S}^{(1)}(\tilde{\boldsymbol{\beta}}, t) = \partial\mathscr{S}^{(0)}(\tilde{\boldsymbol{\beta}}, t)/(\partial\tilde{\boldsymbol{\beta}})$, $\mathscr{S}^{(2)}(\tilde{\boldsymbol{\beta}}, t) = \partial^2\mathscr{S}^{(0)}(\tilde{\boldsymbol{\beta}}, t)/\ (\partial\tilde{\boldsymbol{\beta}}\partial\tilde{\boldsymbol{\beta}}^\top)$ *for all* $\tilde{\boldsymbol{\beta}} \in \mathcal{B}, t \in [0, \zeta]$, *and*

(v) $\Sigma_c(\tilde{\boldsymbol{\beta}}, \zeta) = \int_0^\zeta v_c(\tilde{\boldsymbol{\beta}}, t)\mathscr{S}^{(0)}(\tilde{\boldsymbol{\beta}}, t)d\Lambda_0(t)$ *is positive definite for all* $\tilde{\boldsymbol{\beta}} \in \mathcal{B}$, *where* $e_c(\tilde{\boldsymbol{\beta}}, t) = \mathscr{S}^{(1)}(\tilde{\boldsymbol{\beta}}, t)/\mathscr{S}^{(0)}(\tilde{\boldsymbol{\beta}}, t)$ *and* $v_c(\tilde{\boldsymbol{\beta}}, t) = \mathscr{S}^{(2)}(\tilde{\boldsymbol{\beta}}, t)/\mathscr{S}^{(0)}(\tilde{\boldsymbol{\beta}}, t) - e_c(\tilde{\boldsymbol{\beta}}, t)e_c(\tilde{\boldsymbol{\beta}}, t)^\top$ *for all* $\tilde{\boldsymbol{\beta}} \in \mathcal{B}, t \in [0, \zeta]$.

As there is at most a single tie in continuous model, Assumption A.3.11 is saying that at least one subject is censored at the end of the study.

Assumption A.3.12. $\chi_{pb}(\cdot, \widehat{\boldsymbol{\Lambda}})$ *and* $\chi(\cdot)$ *are strictly concave on* $\mathcal{B}$ *and there exist* $\alpha \in [0, 1/2)$, $\mathcal{C}_{pb} > 0$, *and* $\mathcal{C}_c > 0$ *satisfying the following as* $n \to \infty$:

(i) $\Pr\left(\|\widehat{\boldsymbol{\beta}}_{pb}-\widehat{\boldsymbol{\beta}}_{c}\|_2 \le \mathcal{C}_{pb} n^{\alpha}\left\{\chi_{pb}(\widehat{\boldsymbol{\beta}}_{pb},\widehat{\boldsymbol{\Lambda}})-\chi_{pb}(\widehat{\boldsymbol{\beta}}_{c},\widehat{\boldsymbol{\Lambda}})\right\}\right)\to 1.$

(ii) $\Pr\left(\|\widehat{\boldsymbol{\beta}}_{c}-\widehat{\boldsymbol{\beta}}_{pb}\|_2 \le \mathcal{C}_{c} n^{\alpha}\left\{\chi(\widehat{\boldsymbol{\beta}}_{c})-\chi(\widehat{\boldsymbol{\beta}}_{pb})\right\}\right)\to 1.$

A.4 Lemmas and Proofs

A.4.1 Proof of Theorem 2.1

We begin with the proof for the approximate partial likelihood. First,

$$A_j(\boldsymbol{\beta},\lambda_j)=p_{j_1j}\prod_{i\in\mathcal{R}(t_{(j)})\setminus\{j_1\}}(1-p_{ij})$$

is approximated by $p_{j_1j}=\lambda_j\exp(\boldsymbol{x}_{j_1}^{\top}\boldsymbol{\beta})(1+\mathcal{O}_{\mathrm{P}}(\tau))\approx\lambda_j\exp(\boldsymbol{x}_{j_1}^{\top}\boldsymbol{\beta})$ and we have

$$\prod_{i\in\mathcal{R}(t_{(j)})\setminus\{j_1\}}(1-p_{ij})=\exp\left(-\sum_{i\in\mathcal{R}(t_{(j)})\setminus\{j_1\}}\lambda_j\exp(\boldsymbol{x}_i^{\top}\boldsymbol{\beta})\right).$$

For $B_j(\boldsymbol{\beta},\lambda_j)=\Pr(I_j=1)$, we have

$$B_j(\boldsymbol{\beta},\lambda_j)\approx\mu_j\exp(-\mu_j)\approx\left\{\sum_{i\in\mathcal{R}(t_{(j)})}\lambda_j\exp(\boldsymbol{x}_i^{\top}\boldsymbol{\beta})\right\}\exp\left(-\sum_{i\in\mathcal{R}(t_{(j)})}\lambda_j\exp(\boldsymbol{x}_i^{\top}\boldsymbol{\beta})\right)$$

using Poisson approximation. So, one can approximate (2.10) by

$$\frac{A_j(\boldsymbol{\beta},\lambda_j)}{B_j(\boldsymbol{\beta},\lambda_j)}\approx\frac{\lambda_j\exp\left(\boldsymbol{x}_{j_1}^{\top}\boldsymbol{\beta}\right)\exp\left(-\sum_{i\in\mathcal{R}(t_{(j)})\setminus\{j_1\}}\lambda_j\exp(\boldsymbol{x}_i^{\top}\boldsymbol{\beta})\right)}{\lambda_j\left\{\sum_{i\in\mathcal{R}(t_{(j)})}\exp(\boldsymbol{x}_i^{\top}\boldsymbol{\beta})\right\}\exp\left(-\sum_{i\in\mathcal{R}(t_{(j)})}\lambda_j\exp(\boldsymbol{x}_i^{\top}\boldsymbol{\beta})\right)}\approx\frac{\exp(\boldsymbol{x}_{j_1}^{\top}\boldsymbol{\beta})}{\sum_{i\in\mathcal{R}(t_{(j)})}\exp(\boldsymbol{x}_i^{\top}\boldsymbol{\beta})},$$

which corresponds to the approximation in (A.1).

Now we turn to our proof for the Breslow correction. First,

$$A_j(\boldsymbol{\beta},\lambda_j) = \prod_{i\in\mathcal{D}(t_{(j)})} p_{ij} \prod_{i\in\mathcal{R}(t_{(j)})\setminus\mathcal{D}(t_{(j)})} (1-p_{ij})$$

is approximated by $\prod_{i\in\mathcal{D}(t_{(j)})} p_{ij} \approx \prod_{i\in\mathcal{D}(t_{(j)})} \lambda_j \exp(\boldsymbol{x}_i^\top\boldsymbol{\beta})$ and we have

$$\prod_{i\in\mathcal{R}(t_{(j)})\setminus\mathcal{D}(t_{(j)})} (1-p_{ij}) = \exp\left(-\sum_{i\in\mathcal{R}(t_{(j)})\setminus\mathcal{D}(t_{(j)})} \lambda_j \exp(\boldsymbol{x}_i^\top\boldsymbol{\beta})\right).$$

For $B_j(\boldsymbol{\beta},\lambda_j)$, it is approximated by

$$B_j(\boldsymbol{\beta},\lambda_j) \approx \frac{\mu_j^{d_j}\exp(-\mu_j)}{d_j!} \approx \frac{\left\{\sum_{i\in\mathcal{R}(t_{(j)})} \lambda_j \exp(\boldsymbol{x}_i^\top\boldsymbol{\beta})\right\}^{d_j} \exp\left(-\sum_{i\in\mathcal{R}(t_{(j)})} \lambda_j \exp(\boldsymbol{x}_i^\top\boldsymbol{\beta})\right)}{d_j!}$$

using Poisson approximation. So, one can approximate (2.10) by

$$\begin{aligned}\frac{A_j(\boldsymbol{\beta},\lambda_j)}{B_j(\boldsymbol{\beta},\lambda_j)} &\approx \frac{\left\{\prod_{i\in\mathcal{D}(t_{(j)})} \lambda_j \exp(\boldsymbol{x}_i^\top\boldsymbol{\beta})\right\} \exp\left(-\sum_{i\in\mathcal{R}(t_{(j)})\setminus\mathcal{D}(t_{(j)})} \lambda_j \exp(\boldsymbol{x}_i^\top\boldsymbol{\beta})\right)}{\left\{\sum_{i\in\mathcal{R}(t_{(j)})} \lambda_j \exp(\boldsymbol{x}_i^\top\boldsymbol{\beta})\right\}^{d_j} \exp\left(-\sum_{i\in\mathcal{R}(t_{(j)})} \lambda_j \exp(\boldsymbol{x}_i^\top\boldsymbol{\beta})\right)/d_j!} \\ &\approx \frac{\exp\left(\sum_{i\in\mathcal{D}(t_{(j)})} \boldsymbol{x}_i^\top\boldsymbol{\beta}\right)}{\left\{\sum_{i\in\mathcal{R}(t_{(j)})} \exp(\boldsymbol{x}_i^\top\boldsymbol{\beta})\right\}^{d_j}/d_j!} \propto \frac{\exp\left(\sum_{i\in\mathcal{D}(t_{(j)})} \boldsymbol{x}_i^\top\boldsymbol{\beta}\right)}{\left\{\sum_{i\in\mathcal{R}(t_{(j)})} \exp(\boldsymbol{x}_i^\top\boldsymbol{\beta})\right\}^{d_j}},\end{aligned}$$

which corresponds to the approximation in (A.4).

Now we turn to our proof for the Cox correction. One can approximate (2.10) by

$$\frac{A_j(\boldsymbol{\beta},\lambda_j)}{B_j(\boldsymbol{\beta},\lambda_j)} = \frac{\prod_{i\in\mathcal{D}(t_{(j)})} p_{ij} \prod_{i\in\mathcal{R}(t_{(j)})\setminus\mathcal{D}(t_{(j)})} (1-p_{ij})}{\sum_{\mathcal{A}_{d_j}\in\mathcal{F}_{d_j}} \prod_{i\in\mathcal{A}_{d_j}} p_{ij} \prod_{i\in\mathcal{R}(t_{(j)})\setminus\mathcal{A}_{d_j}} (1-p_{ij})} \approx \frac{\prod_{i\in\mathcal{D}(t_{(j)})} p_{ij}}{\sum_{\mathcal{A}_{d_j}\in\mathcal{F}_{d_j}} \prod_{i\in\mathcal{A}_{d_j}} p_{ij}}$$

$$\approx \frac{\lambda_j^{d_j}\exp\left(\sum_{i\in\mathcal{D}(t_{(j)})}\boldsymbol{x}_i^\top\boldsymbol{\beta}\right)}{\lambda_j^{d_j}\sum_{\mathcal{A}_{d_j}\in\mathcal{F}_{d_j}}\exp\left(\sum_{i\in\mathcal{A}_{d_j}}\boldsymbol{x}_i^\top\boldsymbol{\beta}\right)} = \frac{\exp\left(\sum_{i\in\mathcal{D}(t_{(j)})}\boldsymbol{x}_i^\top\boldsymbol{\beta}\right)}{\sum_{\mathcal{A}_{d_j}\in\mathcal{F}_{d_j}}\exp\left(\sum_{i\in\mathcal{A}_{d_j}}\boldsymbol{x}_i^\top\boldsymbol{\beta}\right)},$$

where is a direct approximation result based on the enumeration in (2.8) to compute the B_j and corresponds to the approximation in (A.2).

Then we turn to our proof for the Kalbfleisch-Prentice correction. Let t_{jm}^l be the latent event time for s_{lm}^j, where $t_{j1}^l < \cdots < t_{jd_j}^l$. Denote by n_{jm}^l the number of subjects in $\mathcal{R}(t_{(j)}, S_l^j, m)$. One can approximate (2.10) by

$$\begin{aligned}
\frac{A_j(\boldsymbol{\beta},\lambda_j)}{B_j(\boldsymbol{\beta},\lambda_j)} &= \frac{\Pr(\text{item } j_1,\ldots,j_{d_j} \text{ failed at } t_{(j)} \mid n_j \text{ units survived up to } t_{(j)})}{\Pr(d_j \text{ out of } n_j \text{ units failed at } t_{(j)} \mid n_j \text{ units survived up to } t_{(j)})} \\
&\approx \frac{1}{d_j!}\sum_{l=1}^{d_j!}\prod_{m=1}^{d_j}\frac{\Pr(\text{item } s_{lm}^j \text{ failed at } t_{jm}^l \mid n_{jm}^l \text{ units survived up to } t_{jm}^l)}{\Pr(1 \text{ out of } n_{jm}^l \text{ units failed at } t_{jm}^l \mid n_{jm}^l \text{ units survived up to } t_{jm}^l)} \\
&\approx \frac{1}{d_j!}\sum_{l=1}^{d_j!}\prod_{m=1}^{d_j}\frac{d\Lambda^*(t_{jm}^l;\boldsymbol{x}_{s_{lm}^j})\prod_{i\in\mathcal{R}(t_{(j)},S_l^j,m)\setminus\{s_{lm}^j\}}\left(1-d\Lambda^*(t_{jm}^l;\boldsymbol{x}_i)\right)}{\sum_{i\in\mathcal{R}(t_{(j)},S_l^j,m)}d\Lambda^*(t_{jm}^l;\boldsymbol{x}_i)\prod_{i'\in\mathcal{R}(t_{(j)},S_l^j,m)\setminus\{i\}}\left(1-d\Lambda^*(t_{jm}^l;\boldsymbol{x}_{i'})\right)} \\
&\approx \frac{1}{d_j!}\sum_{l=1}^{d_j!}\prod_{m=1}^{d_j}\frac{d\Lambda^*(t_{jm}^l;\boldsymbol{x}_{s_{lm}^j})}{\sum_{i\in\mathcal{R}(t_{(j)},S_l^j,m)}d\Lambda^*(t_{jm}^l;\boldsymbol{x}_i)} \approx \frac{1}{d_j!}\sum_{l=1}^{d_j!}\prod_{m=1}^{d_j}\frac{d\Lambda_0^*(t_{jm}^l)\exp(\boldsymbol{x}_{s_{lm}^j}^\top\boldsymbol{\beta})}{\sum_{i\in\mathcal{R}(t_{(j)},S_l^j,m)}d\Lambda_0^*(t_{jm}^l)\exp(\boldsymbol{x}_i^\top\boldsymbol{\beta})} \\
&= \frac{1}{d_j!}\sum_{l=1}^{d_j!}\frac{\exp\left(\sum_{i\in\mathcal{D}(t_{(j)})}\boldsymbol{x}_i^\top\boldsymbol{\beta}\right)}{\prod_{m=1}^{d_j}\left\{\sum_{i\in\mathcal{R}(t_{(j)},S_l^j,m)}\exp(\boldsymbol{x}_i^\top\boldsymbol{\beta})\right\}},
\end{aligned}$$

which corresponds to the approximation in (A.3). The first approximation in above is based on Kalbfleisch and Prentice [26]. ■

A.4.2 Basic Asymptotic Results for the Breslow Estimator and the PB Distribution Estimator under the Grouped Continuous Failure Time Model

Here we derive some asymptotic orders that are used in the proofs for $\widehat{\boldsymbol{\beta}}_{\mathrm{b}}$ and $\widehat{\boldsymbol{\beta}}_{\mathrm{pb}}$.

Lemma A.1. *Let* $X_1, \cdots, X_n \in \mathbb{R}$ *be random variables and* $a_n > 0$ *be a constant depending on* n*. If* $\sup_i |X_i| = \mathcal{O}_P(a_n)$*, then* $\sum_{i=1}^n X_i = \mathcal{O}_P(na_n)$*. Similarly, if* $\sup_i |X_i| = o_P(a_n)$*, then* $\sum_{i=1}^n X_i = o_P(na_n)$*.*

Proof.

$$0 \le \left| \frac{\sum_{i=1}^n X_i}{na_n} \right| \le \sum_{i=1}^n \frac{|X_i|}{na_n} \le \frac{\sup_i |X_i|}{a_n}.$$

If we assume $\sup_i |X_i| = \mathcal{O}_{\mathrm{P}}(a_n)$, there exists $C > 0$ such that

$$\Pr\left(\left| \frac{\sum_{i=1}^n X_i}{na_n} \right| > C \right) \le \Pr\left(\frac{\sup_i |X_i|}{a_n} > C \right) \to 0 \text{ as } n \to \infty,$$

directly implying $\sum_{i=1}^n X_i = \mathcal{O}_{\mathrm{P}}(na_n)$. If we assume $\sup_i |X_i| = o_{\mathrm{P}}(a_n)$, for all $\epsilon > 0$,

$$\Pr\left(\left| \frac{\sum_{i=1}^n X_i}{na_n} \right| > \epsilon \right) \le \Pr\left(\frac{\sup_i |X_i|}{a_n} > \epsilon \right) \to 0 \text{ as } n \to \infty,$$

directly implying $\sum_{i=1}^n X_i = o_{\mathrm{P}}(na_n)$. ■

Here we give a lemma showing the convergence of Stieltjes integrals with respect to the baseline hazard, which is used in the proofs for the asymptotic theory of the grouped continuous failure time model.

Lemma A.2. *Let $A_n(\cdot)$ be a stochastic process that can be scalar, vector, or matrix, where $\sup_{t\in\Omega}\|A_n(t)-a(t)\|_\infty \xrightarrow{P} 0$ for some deterministic process $a(\cdot)$. Under Assumption A.3.3, we have*

$$\int_0^\zeta A_n(t)d\Lambda_0^*(t)=\sum_{t\in\Omega}A_n(t)\Delta\Lambda_0^*(t)\xrightarrow{P}\int_0^\zeta a(t)d\Lambda_0(t)\quad \textit{as } n\to\infty.$$

Proof. Let $A_n(\cdot)$ and $a(\cdot)$ be scalars. First, by the definition of the Stieltjes integral, as $n\to\infty$,

$$\int_0^\zeta a(t)d\Lambda_0^*(t)=\sum_{t\in\Omega}a(t)\Delta\Lambda_0^*(t)=\sum_{t\in\Omega}a(t)\left(\Lambda_0(t)-\Lambda_0(t-\tau)\right)\to\int_0^\zeta a(t)d\Lambda_0(t),$$

considering the fact that as $n\to\infty$, Ω gets larger and the bin size τ for the Stieltjes integral goes to 0. Second, by the given conditions, we have

$$0\le\sup_{t\in\Omega}\left|\left(A_n(t)-a(t)\right)\Delta\Lambda_0^*(t)\right|\le\sup_{t\in\Omega}\left|\left(A_n(t)-a(t)\right)\right|\sup_{t\in\Omega}\left|\Delta\Lambda_0^*(t)\right|=o_{\mathrm{P}}(1)\mathcal{O}(\tau)=o_{\mathrm{P}}(\tau),$$

implying that $\sup_{t\in\Omega}\left|\left(A_n(t)-a(t)\right)\Delta\Lambda_0^*(t)\right|=o_{\mathrm{P}}(\tau)$. By Lemma A.1, we have

$$\int_0^\zeta\left(A_n(t)-a(t)\right)d\Lambda_0^*(t)=\sum_{t\in\Omega}\left(A_n(t)-a(t)\right)\Delta\Lambda_0^*(t)=o_{\mathrm{P}}\left(\frac{\zeta}{\tau}\tau\right)=o_{\mathrm{P}}(1),$$

which completes the proof. The result can be generalized to the cases when $A_n(\cdot)$ and $a(\cdot)$ are vectors or matrices. ■

Lemma A.3. *Under Assumption A.3.2,*

$$0<\exp\left(-\sqrt{d}\rho C_{\mathcal{B}}\right)\le\inf_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},i}\exp\left(\boldsymbol{x}_i^\top\tilde{\boldsymbol{\beta}}\right)\le\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},i}\exp\left(\boldsymbol{x}_i^\top\tilde{\boldsymbol{\beta}}\right)\le\exp\left(\sqrt{d}\rho C_{\mathcal{B}}\right)<\infty.$$

Proof.

$$\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},i} \exp\left(\boldsymbol{x}_i^\top\tilde{\boldsymbol{\beta}}\right) = \exp\left(\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},i} \boldsymbol{x}_i^\top\tilde{\boldsymbol{\beta}}\right) \le \exp\left(\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},i} |\boldsymbol{x}_i^\top\tilde{\boldsymbol{\beta}}|\right) \le \exp\left(\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},i} \|\boldsymbol{x}_i\|_2\|\tilde{\boldsymbol{\beta}}\|_2\right) \le \exp\left(\sqrt{d}\rho C_{\mathcal{B}}\right).$$

Similarly,

$$\inf_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},i} \exp\left(\boldsymbol{x}_i^\top\tilde{\boldsymbol{\beta}}\right) = \exp\left(\inf_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},i} \boldsymbol{x}_i^\top\tilde{\boldsymbol{\beta}}\right) = \exp\left(-\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},i} \{-\boldsymbol{x}_i^\top\tilde{\boldsymbol{\beta}}\}\right) \ge \exp\left(-\sqrt{d}\rho C_{\mathcal{B}}\right),$$

where the last inequality is due to $\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},i}\{-\boldsymbol{x}_i^\top\tilde{\boldsymbol{\beta}}\} \le \sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},i} |\boldsymbol{x}_i^\top\tilde{\boldsymbol{\beta}}| \le \sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},i} \|\boldsymbol{x}_i\|_2\|\tilde{\boldsymbol{\beta}}\|_2 \le \sqrt{d}\rho C_{\mathcal{B}}$. ■

Now we show the asymptotic order for the discretized hazard. By the Taylor series expansion, for all $i \in \{1, \cdots, n\}$ and $t \in \Omega$,

$$d\Lambda^*(t;\boldsymbol{x}_i) = \exp(\boldsymbol{x}_i^\top\boldsymbol{\beta})d\Lambda_0^*(t)(1 + a_i(\boldsymbol{\beta}, t)) = \exp(\boldsymbol{x}_i^\top\boldsymbol{\beta})d\Lambda_0^*(t) + b_i(\boldsymbol{\beta}, t),$$

where

$$a_i(\boldsymbol{\beta}, t) = -\frac{\exp(-c_i(\boldsymbol{\beta}, t))}{2}\exp(\boldsymbol{x}_i^\top\boldsymbol{\beta})d\Lambda_0^*(t)$$

with some $c_i(\boldsymbol{\beta}, t) \in [0, \exp(\boldsymbol{x}_i^\top\boldsymbol{\beta})d\Lambda_0^*(t)]$.

Lemma A.4. *Under Assumptions A.3.2 and A.3.3, we have* $\sup_{t\in\Omega,i}\exp(\boldsymbol{x}_i^\top\boldsymbol{\beta})d\Lambda_0^*(t) = \mathcal{O}(\tau)$ *and* $\sup_{t\in\Omega,i}|a_i(\boldsymbol{\beta}, t)| = \mathcal{O}(\tau)$*, resulting in* $\sup_{t\in\Omega,i}|b_i(\boldsymbol{\beta}, t)| = \mathcal{O}(\tau^2)$ *and* $\sup_{t\in\Omega,i} d\Lambda^*(t;\boldsymbol{x}_i) = \mathcal{O}(\tau)$.

Proof. First, we have

$$\sup_{t\in\Omega,i}\exp(\boldsymbol{x}_i^\top\boldsymbol{\beta})d\Lambda_0^*(t) \le \sup_i \exp(\boldsymbol{x}_i^\top\boldsymbol{\beta})\sup_{t\in\Omega} d\Lambda_0^*(t) \le \exp\left(\sqrt{d}\rho C_{\mathcal{B}}\right)\mathcal{O}(\tau) = \mathcal{O}(\tau),$$

implying that $\sup_{t\in\Omega,i}\exp(\boldsymbol{x}_i^\top\boldsymbol{\beta})d\Lambda_0^*(t)=\mathcal{O}(\tau)$. Second, we have

$$\sup_{t\in\Omega,i}|a_i(\boldsymbol{\beta},t)|\le\sup_{t\in\Omega,i}\left|\frac{\exp(-c_i(\boldsymbol{\beta},t))}{2}\right|\sup_{t\in\Omega,i}\exp(\boldsymbol{x}_i^\top\boldsymbol{\beta})d\Lambda_0^*(t)\le\frac{1}{2}\sup_{t\in\Omega,i}\exp(\boldsymbol{x}_i^\top\boldsymbol{\beta})d\Lambda_0^*(t)=\mathcal{O}(\tau),$$

implying that $\sup_{t\in\Omega,i}|a_i(\boldsymbol{\beta},t)|=\mathcal{O}(\tau)$. Using these results, we have

$$\sup_{t\in\Omega,i}|b_i(\boldsymbol{\beta},t)|\le\sup_{t\in\Omega,i}|a_i(\boldsymbol{\beta},t)|\sup_{t\in\Omega,i}\exp(\boldsymbol{x}_i^\top\boldsymbol{\beta})d\Lambda_0^*(t)=\mathcal{O}(\tau)\mathcal{O}(\tau)=\mathcal{O}(\tau^2)\text{ and}$$

$$\sup_{t\in\Omega,i}d\Lambda^*(t;\boldsymbol{x}_i)\le\sup_{t\in\Omega,i}\exp(\boldsymbol{x}_i^\top\boldsymbol{\beta})d\Lambda_0^*(t)+\sup_{t\in\Omega,i}|b_i(\boldsymbol{\beta},t)|=\mathcal{O}(\tau)+\mathcal{O}(\tau^2)=\mathcal{O}(\tau).$$

■

Now we present some basic convergence and boundedness results used in the proofs of the asymptotic theory for $\widehat{\boldsymbol{\beta}}_{\mathrm{b}}$.

Lemma A.5. *Under Assumptions A.3.1 – A.3.5,* $\sup_{t\in\Omega}n(t)=\mathcal{O}_P(n)$, $\sup_{t\in\Omega}\{1/n(t)\}=1/\{\inf_{t\in\Omega}n(t)\}=1/n(\zeta)=\mathcal{O}_P(1/n)$, *and we have*

$$\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega}S^{(0)}(\tilde{\boldsymbol{\beta}},t)=\mathcal{O}_P(n),\quad\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega}\frac{1}{S^{(0)}(\tilde{\boldsymbol{\beta}},t)}=\frac{1}{\inf_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega}S^{(0)}(\tilde{\boldsymbol{\beta}},t)}=\frac{1}{\inf_{\tilde{\boldsymbol{\beta}}\in\mathcal{B}}S^{(0)}(\tilde{\boldsymbol{\beta}},\zeta)}=\mathcal{O}_P(1/n),$$

$$\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega}\left|\log\left\{S^{(0)}(\tilde{\boldsymbol{\beta}},t)/n\right\}\right|=\mathcal{O}_P(1),\quad\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega}\|S^{(1)}(\tilde{\boldsymbol{\beta}},t)\|_\infty=\mathcal{O}_P(n),\quad\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega}\|S^{(2)}(\tilde{\boldsymbol{\beta}},t)\|_\infty=\mathcal{O}_P(n),$$

$$\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega}\|\epsilon(\tilde{\boldsymbol{\beta}},t)\|_\infty=\mathcal{O}_P(1),\ \textit{and}\ \sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega}\|\mathcal{V}(\tilde{\boldsymbol{\beta}},t)\|_\infty=\mathcal{O}_P(1).$$

If Assumption A.3.7 is further assumed, then

$$\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega} \|\epsilon(\tilde{\boldsymbol{\beta}},t)-e(\tilde{\boldsymbol{\beta}},t)\|_\infty = o_P(1),$$

$$\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega} \|\mathcal{V}(\tilde{\boldsymbol{\beta}},t)-v(\tilde{\boldsymbol{\beta}},t)\|_\infty = o_P(1),$$

and $\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega}\left|\log\left\{S^{(0)}(\tilde{\boldsymbol{\beta}},t)/n\right\}-\log\left\{s^{(0)}(\tilde{\boldsymbol{\beta}},t)\right\}\right| = o_P(1)$.

Proof. The proof of Lemma A.5 is in Section A.5.1. ■

As $\boldsymbol{\beta}\in\mathcal{B}$, one can use results in Lemma A.5 for $\boldsymbol{\beta}$.

A.4.3 Lemma A.6 and Its Proof

Lemma A.6. *Under Assumptions A.3.1 – A.3.5, for all* $t\in[0,\zeta]$, $U_b(\boldsymbol{\beta},t) = U(\boldsymbol{\beta},t) + \mathcal{O}_P(n\tau)$, *where* $U(\boldsymbol{\beta},t) = \sum_{i=1}^n\int_0^t(\boldsymbol{x}_i-\epsilon(\boldsymbol{\beta},s))\,dM_i(s) = \sum_{i=1}^n\sum_{s\in\Omega_t}(\boldsymbol{x}_i-\epsilon(\boldsymbol{\beta},s))\,\Delta M_i(s)$ *is a martingale.*

Proof. For $t\in[0,\zeta]$, one has

$$\begin{aligned}
&\sum_{i=1}^n\int_0^t(\boldsymbol{x}_i-\epsilon(\boldsymbol{\beta},s))\,dA_i(s) = \sum_{i=1}^n\sum_{s\in\Omega_t}(\boldsymbol{x}_i-\epsilon(\boldsymbol{\beta},s))\,\Delta A_i(s)\\
&=\sum_{i=1}^n\sum_{s\in\Omega_t}(\boldsymbol{x}_i-\epsilon(\boldsymbol{\beta},s))\,Y_i(s)\exp(\boldsymbol{x}_i^\top\boldsymbol{\beta})\Delta\Lambda_0^*(s)\,(1+a_i(\boldsymbol{\beta},s))\\
&=\left\{\sum_{s\in\Omega_t}\sum_{i=1}^n(\boldsymbol{x}_i-\epsilon(\boldsymbol{\beta},s))\,Y_i(s)\exp(\boldsymbol{x}_i^\top\boldsymbol{\beta})\Delta\Lambda_0^*(s)\right\}+\mathcal{O}_{\mathrm{P}}(n\tau)=\mathcal{O}_{\mathrm{P}}(n\tau),
\end{aligned}$$

where for $t\in[0,\zeta]$, $\sum_{s\in\Omega_t}\sum_{i=1}^n(\boldsymbol{x}_i-\epsilon(\boldsymbol{\beta},s))\,Y_i(s)\exp(\boldsymbol{x}_i^\top\boldsymbol{\beta})\Delta\Lambda_0^*(s)=\mathbf{0}$ by the summation structure and $\sum_{s\in\Omega_t}\sum_{i=1}^n(\boldsymbol{x}_i-\epsilon(\boldsymbol{\beta},s))\,Y_i(s)\exp(\boldsymbol{x}_i^\top\boldsymbol{\beta})\Delta\Lambda_0^*(s)a_i(\boldsymbol{\beta},s) = \mathcal{O}_{\mathrm{P}}(n\tau)$. To see this, for all $\ell \in \{1, \cdots, d\}$, let $\epsilon_\ell(\boldsymbol{\beta}, t)$ be the ℓth element of $\epsilon(\boldsymbol{\beta}, t)$. Then

$$\begin{aligned}
&\left| \sum_{s \in \Omega_t} \sum_{i=1}^{n} \left(x_{i\ell} - \epsilon_\ell(\boldsymbol{\beta}, s)\right) Y_i(s) \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta}) \Delta\Lambda_0^*(s) a_i(\boldsymbol{\beta}, s) \right| \\
&\quad \leq \sum_{t \in \Omega} \sum_{i=1}^{n} \left| \left(x_{i\ell} - \epsilon_\ell(\boldsymbol{\beta}, t)\right) Y_i(t) \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta}) \Delta\Lambda_0^*(t) a_i(\boldsymbol{\beta}, t) \right| \\
&\quad \leq \frac{\zeta n}{\tau} \sup_{t \in \Omega, i} \left| \left(x_{i\ell} - \epsilon_\ell(\boldsymbol{\beta}, t)\right) Y_i(t) \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta}) \Delta\Lambda_0^*(t) a_i(\boldsymbol{\beta}, t) \right| = \frac{\zeta n}{\tau} \mathcal{O}_{\mathrm{P}}(\tau^2) = \mathcal{O}_{\mathrm{P}}(n\tau),
\end{aligned}$$

where

$$\begin{aligned}
&\sup_{t \in \Omega, i} \left| \left(x_{i\ell} - \epsilon_\ell(\boldsymbol{\beta}, t)\right) Y_i(t) \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta}) \Delta\Lambda_0^*(t) a_i(\boldsymbol{\beta}, t) \right| \\
&\qquad \leq \left(\sup_i |x_{il}| + \sup_{t \in \Omega} |\epsilon_\ell(\boldsymbol{\beta}, t)| \right) \sup_i \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta}) \sup_{t \in \Omega} \Delta\Lambda_0^*(t) \sup_{t \in \Omega, i} |a_i(\boldsymbol{\beta}, t)| \\
&\qquad = \left(\mathcal{O}(1) + \mathcal{O}_{\mathrm{P}}(1)\right) \mathcal{O}(1) \mathcal{O}(\tau) \mathcal{O}(\tau) = \mathcal{O}_{\mathrm{P}}(\tau^2).
\end{aligned}$$

So, for $t \in [0, \zeta]$, one can represent the score function as

$$U_{\mathrm{b}}(\boldsymbol{\beta}, t) = \sum_{i=1}^{n} \int_0^t \left(\boldsymbol{x}_i - \epsilon(\boldsymbol{\beta}, s)\right) dM_i(s) + \mathcal{O}_{\mathrm{P}}(n\tau) = \sum_{i=1}^{n} \sum_{s \in \Omega_t} \left(\boldsymbol{x}_i - \epsilon(\boldsymbol{\beta}, s)\right) \Delta M_i(s) + \mathcal{O}_{\mathrm{P}}(n\tau).$$

■

A.4.4 Lemma A.7 and Its Proof

Lemma A.7. *Under Assumptions A.3.1 – A.3.6, for $t \in [0, \zeta]$, we have*

$$\langle U(\boldsymbol{\beta}) \rangle(t) = \left\{ \sum_{i=1}^{n} \int_0^t n G_i(s) G_i(s)^\top Y_i(s) d\Lambda^*(s; \boldsymbol{x}_i) \right\} + \mathcal{O}_P(n\tau)$$

$$= \left\{ \sum_{i=1}^{n} \sum_{s \in \Omega_t} n G_i(s) G_i(s)^\top Y_i(s) \Delta\Lambda^*(s; \boldsymbol{x}_i) \right\} + \mathcal{O}_P(n\tau),$$

where $G_i(t) = n^{-1/2}(\boldsymbol{x}_i - \epsilon(\boldsymbol{\beta}, t))$.

Proof. First, for $t \in \Omega$, we have

$$\begin{aligned} d\langle M_i \rangle(t) &= \mathrm{Var}(dM_i(t) | \mathcal{F}_{t^-}) = \mathrm{Var}(dN_i(t) - dA_i(t) | \mathcal{F}_{t^-}) \\ &= \mathrm{Var}(dN_i(t) | \mathcal{F}_{t^-}) = \Pr(dN_i(t) = 1 | \mathcal{F}_{t^-}) \left(1 - \Pr(dN_i(t) = 1 | \mathcal{F}_{t^-})\right) \\ &= Y_i(t) d\Lambda^*(t; \boldsymbol{x}_i) \left(1 - \Delta\Lambda^*(t; \boldsymbol{x}_i)\right), \end{aligned}$$

where $\sup_{t \in \Omega, i} d\langle M_i \rangle(t) \leq \sup_{t \in \Omega, i} d\Lambda^*(t; \boldsymbol{x}_i) = \mathcal{O}(\tau)$, which leads to $\sup_{t \in \Omega, i} d\langle M_i \rangle(t) = \mathcal{O}_{\mathrm{P}}(\tau)$. For $t \in [0, \zeta]$, the predictable variation process for the score function can be derived as

$$\begin{aligned} \langle U(\boldsymbol{\beta}) \rangle(t) =& \sum_{i=1}^{n} \int_0^t (\boldsymbol{x}_i - \epsilon(\boldsymbol{\beta}, s))(\boldsymbol{x}_i - \epsilon(\boldsymbol{\beta}, s))^\top d\langle M_i \rangle(s) \\ =& \sum_{i=1}^{n} \sum_{s \in \Omega_t} (\boldsymbol{x}_i - \epsilon(\boldsymbol{\beta}, s))(\boldsymbol{x}_i - \epsilon(\boldsymbol{\beta}, s))^\top Y_i(s) \Delta\Lambda^*(s; \boldsymbol{x}_i) \left(1 - \Delta\Lambda^*(s; \boldsymbol{x}_i)\right) \\ =& \left\{ \sum_{i=1}^{n} \sum_{s \in \Omega_t} (\boldsymbol{x}_i - \epsilon(\boldsymbol{\beta}, s))(\boldsymbol{x}_i - \epsilon(\boldsymbol{\beta}, s))^\top Y_i(s) \Delta\Lambda^*(s; \boldsymbol{x}_i) \right\} + \mathcal{O}_{\mathrm{P}}(n\tau) \\ =& \left\{ \sum_{i=1}^{n} \int_0^t n G_i(s) G_i(s)^\top Y_i(s) d\Lambda^*(s; \boldsymbol{x}_i) \right\} + \mathcal{O}_{\mathrm{P}}(n\tau), \end{aligned}$$

where for all $\ell, \ell' \in \{1, \cdots, d\}$, we have

$$\begin{aligned} & \sup_{t \in \Omega, i} \left| (x_{i\ell} - \epsilon_\ell(\boldsymbol{\beta}, t))(x_{i\ell'} - \epsilon_{\ell'}(\boldsymbol{\beta}, t)) Y_i(t) \{\Delta\Lambda^*(t; \boldsymbol{x}_i)\}^2 \right| \\ & \qquad \leq \sup_{t \in \Omega, i} \left| (x_{i\ell} - \epsilon_\ell(\boldsymbol{\beta}, t)) \right| \sup_{t \in \Omega, i} \left| (x_{i\ell'} - \epsilon_{\ell'}(\boldsymbol{\beta}, t)) \right| \sup_{t \in \Omega, i} \{\Delta\Lambda^*(t; \boldsymbol{x}_i)\}^2 = \mathcal{O}_{\mathrm{P}}(\tau^2), \end{aligned}$$

resulting in

$$\begin{aligned}&\left|\sum_{i=1}^{n}\sum_{s\in\Omega_t}(x_{i\ell}-\epsilon_\ell(\boldsymbol{\beta},s))(x_{i\ell'}-\epsilon_{\ell'}(\boldsymbol{\beta},s))Y_i(s)\{\Delta\Lambda^*(s;\boldsymbol{x}_i)\}^2\right|\\ &\quad\le\sum_{i=1}^{n}\sum_{t\in\Omega}\left|(x_{i\ell}-\epsilon_\ell(\boldsymbol{\beta},t))(x_{i\ell'}-\epsilon_{\ell'}(\boldsymbol{\beta},t))Y_i(t)\{\Delta\Lambda^*(t;\boldsymbol{x}_i)\}^2\right|\\ &\quad\le\frac{n\zeta}{\tau}\sup_{t\in\Omega,i}\left|(x_{i\ell}-\epsilon_\ell(\boldsymbol{\beta},t))(x_{i\ell'}-\epsilon_{\ell'}(\boldsymbol{\beta},t))Y_i(t)\{\Delta\Lambda^*(t;\boldsymbol{x}_i)\}^2\right|=\frac{n\zeta}{\tau}\mathcal{O}_{\mathrm{P}}(\tau^2)=\mathcal{O}_{\mathrm{P}}(n\tau).\end{aligned}$$

■

A.4.5 Lemma A.8 and Its Proof

Lemma A.8. *Under Assumptions A.3.1 – A.3.7, when* $\tau\to0$ *as* $n\to\infty$*, we have* $I_b(\boldsymbol{\beta},\zeta)/n\xrightarrow{P}\Sigma(\boldsymbol{\beta},\zeta)$ *and* $I_b(\widehat{\boldsymbol{\beta}},\zeta)/n\xrightarrow{P}\Sigma(\boldsymbol{\beta},\zeta)$ *for any random variable* $\widehat{\boldsymbol{\beta}}\in\mathcal{B}$ *satisfying* $\widehat{\boldsymbol{\beta}}\xrightarrow{P}\boldsymbol{\beta}$.

Proof. We show some details for proving $I_{\mathrm{b}}(\boldsymbol{\beta},\zeta)/n\xrightarrow{\mathrm{P}}\Sigma(\boldsymbol{\beta},\zeta)$. First, we show that

$$\int_0^\zeta\mathcal{V}(\boldsymbol{\beta},t)dM(t)/n=o_{\mathrm{P}}(1),$$

which is equivalent to show

$$\int_0^\zeta\mathcal{V}_{\ell,\ell'}(\boldsymbol{\beta},t)dM(t)/n=o_{\mathrm{P}}(1)$$

for all $\ell,\ell'\in\{1,\dots,d\}$, where $\mathcal{V}_{\ell,\ell'}(\boldsymbol{\beta},t)$ is the element of $\mathcal{V}(\boldsymbol{\beta},t)$ in the ℓth row and ℓ'th column. For $t\in[0,\zeta]$, one can derive the predictable variation process of martingale

$\int_0^t \mathcal{V}_{\ell,\ell'}(\boldsymbol{\beta}, s)dM(s)/n$ as

$$\sum_{i=1}^n \int_0^t \mathcal{V}^2_{\ell,\ell'}(\boldsymbol{\beta}, s)\frac{d\langle M_i\rangle(s)}{n^2},$$

which is non-decreasing in $t \in [0, \zeta]$. So,

$$\sup_{t\in[0,\zeta]}\left|\sum_{i=1}^n \int_0^t \mathcal{V}^2_{\ell,\ell'}(\boldsymbol{\beta}, s)\frac{d\langle M_i\rangle(s)}{n^2}\right| = \sum_{i=1}^n \int_0^\zeta \mathcal{V}^2_{\ell,\ell'}(\boldsymbol{\beta}, t)\frac{d\langle M_i\rangle(t)}{n^2} = \mathcal{O}_{\mathrm{P}}\left(\frac{n\zeta}{\tau}\frac{\tau}{n^2}\right) = \mathcal{O}_{\mathrm{P}}\left(\frac{1}{n}\right) = o_{\mathrm{P}}(1),$$

where we used Lemma A.1 and the fact that

$$\sup_{t\in\Omega,i} \mathcal{V}^2_{\ell,\ell'}(\boldsymbol{\beta}, t)\frac{d\langle M_i\rangle(t)}{n^2} \le \frac{1}{n^2}\left\{\sup_{t\in\Omega}|\mathcal{V}_{\ell,\ell'}(\boldsymbol{\beta}, t)|\right\}^2 \sup_{t\in\Omega,i} d\langle M_i\rangle(t) = \mathcal{O}_{\mathrm{P}}\left(\frac{\tau}{n^2}\right).$$

This result leads to $\sup_{t\in[0,\zeta]}|\int_0^t \mathcal{V}_{\ell,\ell'}(\boldsymbol{\beta}, s)dM(s)/n| = o_{\mathrm{P}}(1)$ by Lenglart’s Inequality on page 167 of Kalbfleisch and Prentice [27], which allows triangular array structure. Note that filtration and every processes depend on τ, changing as $n \to \infty$, that is, they have triangular array structure. So, we have $\int_0^\zeta \mathcal{V}(\boldsymbol{\beta}, t)dM(t)/n = o_{\mathrm{P}}(1)$. Then

$$\begin{aligned}
\frac{1}{n}I_{\mathrm{b}}(\boldsymbol{\beta}, \zeta) &= \int_0^\zeta \mathcal{V}(\boldsymbol{\beta}, t)\frac{dN(t)}{n} = \int_0^\zeta \mathcal{V}(\boldsymbol{\beta}, t)\frac{dA(t)}{n} + \int_0^\zeta \mathcal{V}(\boldsymbol{\beta}, t)\frac{dM(t)}{n} \\
&= \int_0^\zeta \mathcal{V}(\boldsymbol{\beta}, t)\frac{dA(t)}{n} + o_{\mathrm{P}}(1) = \int_0^\zeta \mathcal{V}(\boldsymbol{\beta}, t)\frac{\sum_{i=1}^n Y_i(t)d\Lambda^*(t;\boldsymbol{x}_i)}{n} + o_{\mathrm{P}}(1) \\
&= \sum_{t\in\Omega} \mathcal{V}(\boldsymbol{\beta}, t)\frac{\sum_{i=1}^n Y_i(t)\Delta\Lambda^*(t;\boldsymbol{x}_i)}{n} + o_{\mathrm{P}}(1) = \sum_{t\in\Omega} \mathcal{V}(\boldsymbol{\beta}, t)\frac{S^{(0)}(\boldsymbol{\beta}, t)}{n}\Delta\Lambda_0^*(t) + \mathcal{O}_{\mathrm{P}}(\tau) + o_{\mathrm{P}}(1) \\
&\xrightarrow{\mathrm{P}} \int_0^\zeta v(\boldsymbol{\beta}, t)s^{(0)}(\boldsymbol{\beta}, t)d\Lambda_0(t) = \Sigma(\boldsymbol{\beta}, \zeta).
\end{aligned}$$

Hence we use the fact that $\Delta\Lambda^*(t;\boldsymbol{x}_i) = \exp(\boldsymbol{x}_i^\top\boldsymbol{\beta})\Delta\Lambda_0^*(t)\left(1 + a_i(\boldsymbol{\beta},t)\right)$ for all $t \in \Omega$, and

$$\sum_{t\in\Omega}\mathcal{V}(\boldsymbol{\beta},t)\frac{\sum_{i=1}^n Y_i(t)\exp(\boldsymbol{x}_i^\top\boldsymbol{\beta})\Delta\Lambda_0^*(t)a_i(\boldsymbol{\beta},t)}{n} = \mathcal{O}_{\mathrm{P}}\left(\frac{n\zeta}{\tau}\frac{\tau^2}{n}\right) = \mathcal{O}_{\mathrm{P}}(\tau)$$

by

$$\begin{aligned}&\sup_{t\in\Omega,i}\left|\mathcal{V}_{\ell,\ell'}(\boldsymbol{\beta},t)Y_i(t)\exp(\boldsymbol{x}_i^\top\boldsymbol{\beta})\Delta\Lambda_0^*(t)a_i(\boldsymbol{\beta},t)\right|\\&\qquad\le\sup_{t\in\Omega}\left|\mathcal{V}_{\ell,\ell'}(\boldsymbol{\beta},t)\right|\sup_i\exp(\boldsymbol{x}_i^\top\boldsymbol{\beta})\sup_{t\in\Omega}\Delta\Lambda_0^*(t)\sup_{t\in\Omega,i}\left|a_i(\boldsymbol{\beta},t)\right| = \mathcal{O}_{\mathrm{P}}(\tau^2)\end{aligned}$$

for all $\ell,\ell' \in \{1,\cdots,d\}$. Also, using the fact that

$$\begin{aligned}&\sup_{t\in\Omega}\left|\mathcal{V}_{\ell,\ell'}(\boldsymbol{\beta},t)\frac{S^{(0)}(\boldsymbol{\beta},t)}{n} - v_{\ell,\ell'}(\boldsymbol{\beta},t)s^{(0)}(\boldsymbol{\beta},t)\right|\\&\quad\le\sup_{t\in\Omega}\left|\mathcal{V}_{\ell,\ell'}(\boldsymbol{\beta},t)\left\{\frac{S^{(0)}(\boldsymbol{\beta},t)}{n} - s^{(0)}(\boldsymbol{\beta},t)\right\}\right| + \sup_{t\in\Omega}\left|s^{(0)}(\boldsymbol{\beta},t)\left\{\mathcal{V}_{\ell,\ell'}(\boldsymbol{\beta},t) - v_{\ell,\ell'}(\boldsymbol{\beta},t)\right\}\right|\\&\quad\le\sup_{t\in\Omega}\left|\mathcal{V}_{\ell,\ell'}(\boldsymbol{\beta},t)\right|\sup_{t\in\Omega}\left|\frac{S^{(0)}(\boldsymbol{\beta},t)}{n} - s^{(0)}(\boldsymbol{\beta},t)\right| + \sup_{t\in\Omega}\left|s^{(0)}(\boldsymbol{\beta},t)\right|\sup_{t\in\Omega}\left|\left\{\mathcal{V}_{\ell,\ell'}(\boldsymbol{\beta},t) - v_{\ell,\ell'}(\boldsymbol{\beta},t)\right\}\right|\\&\quad= \mathcal{O}_{\mathrm{P}}(1)o_{\mathrm{P}}(1) + \mathcal{O}(1)o_{\mathrm{P}}(1) = o_{\mathrm{P}}(1)\end{aligned}$$

for all $\ell,\ell' \in \{1,\cdots,d\}$, by Lemma A.2, $\xrightarrow{\mathrm{P}}$ in the last line holds.

Let $\widehat{\boldsymbol{\beta}} \in \mathcal{B}$ satisfies $\widehat{\boldsymbol{\beta}} \xrightarrow{\mathrm{P}} \boldsymbol{\beta}$. We have

$$\begin{aligned}\frac{1}{n}I_{\mathrm{b}}(\widehat{\boldsymbol{\beta}},\zeta) - \Sigma(\boldsymbol{\beta},\zeta) &= \frac{1}{n}I_{\mathrm{b}}(\widehat{\boldsymbol{\beta}},\zeta) - \frac{1}{n}I_{\mathrm{b}}(\boldsymbol{\beta},\zeta) + \frac{1}{n}I_{\mathrm{b}}(\boldsymbol{\beta},\zeta) - \Sigma(\boldsymbol{\beta},\zeta)\\&= \int_0^\zeta\left\{\mathcal{V}(\widehat{\boldsymbol{\beta}},t) - \mathcal{V}(\boldsymbol{\beta},t)\right\}\frac{dN(t)}{n} + o_{\mathrm{P}}(1).\end{aligned}$$

So we need to show $\int_0^\zeta \left\{ \mathcal{V}(\widehat{\boldsymbol{\beta}}, t) - \mathcal{V}(\boldsymbol{\beta}, t) \right\} dN(t)/n \xrightarrow{\mathrm{P}} 0$. For all $\ell, \ell' \in \{1, \cdots, d\}$,

$$\begin{aligned}
\left| \int_0^\zeta \left\{ \mathcal{V}_{\ell,\ell'}(\widehat{\boldsymbol{\beta}}, t) - \mathcal{V}_{\ell,\ell'}(\boldsymbol{\beta}, t) \right\} \frac{dN(t)}{n} \right| \leq & \left| \int_0^\zeta \left\{ \mathcal{V}_{\ell,\ell'}(\widehat{\boldsymbol{\beta}}, t) - v_{\ell,\ell'}(\widehat{\boldsymbol{\beta}}, t) \right\} \frac{dN(t)}{n} \right| \\
& + \left| \int_0^\zeta \left\{ \mathcal{V}_{\ell,\ell'}(\boldsymbol{\beta}, t) - v_{\ell,\ell'}(\boldsymbol{\beta}, t) \right\} \frac{dN(t)}{n} \right| \\
& + \left| \int_0^\zeta \left\{ v_{\ell,\ell'}(\widehat{\boldsymbol{\beta}}, t) - v_{\ell,\ell'}(\boldsymbol{\beta}, t) \right\} \frac{dN(t)}{n} \right|.
\end{aligned} \tag{A.1}$$

For the first part of the right-hand side of (A.1), we have

$$\begin{aligned}
\left| \int_0^\zeta \left\{ \mathcal{V}_{\ell,\ell'}(\widehat{\boldsymbol{\beta}}, t) - v_{\ell,\ell'}(\widehat{\boldsymbol{\beta}}, t) \right\} \frac{dN(t)}{n} \right| &= \left| \sum_{t \in \Omega} \left\{ \mathcal{V}_{\ell,\ell'}(\widehat{\boldsymbol{\beta}}, t) - v_{\ell,\ell'}(\widehat{\boldsymbol{\beta}}, t) \right\} \frac{\Delta N(t)}{n} \right| \\
&\leq \sum_{t \in \Omega} \left| \mathcal{V}_{\ell,\ell'}(\widehat{\boldsymbol{\beta}}, t) - v_{\ell,\ell'}(\widehat{\boldsymbol{\beta}}, t) \right| \frac{\Delta N(t)}{n} \\
&\leq \sup_{t \in \Omega} \left| \mathcal{V}_{\ell,\ell'}(\widehat{\boldsymbol{\beta}}, t) - v_{\ell,\ell'}(\widehat{\boldsymbol{\beta}}, t) \right| \sum_{t \in \Omega} \frac{\Delta N(t)}{n} \\
&\leq \sup_{\tilde{\boldsymbol{\beta}} \in \mathcal{B}} \sup_{t \in \Omega} \left| \mathcal{V}_{\ell,\ell'}(\tilde{\boldsymbol{\beta}}, t) - v_{\ell,\ell'}(\tilde{\boldsymbol{\beta}}, t) \right| \sum_{t \in \Omega} \frac{\Delta N(t)}{n} \\
&= \sup_{\tilde{\boldsymbol{\beta}} \in \mathcal{B}, t \in \Omega} \left| \mathcal{V}_{\ell,\ell'}(\tilde{\boldsymbol{\beta}}, t) - v_{\ell,\ell'}(\tilde{\boldsymbol{\beta}}, t) \right| \sum_{t \in \Omega} \frac{\Delta N(t)}{n} \\
&\leq \sup_{\tilde{\boldsymbol{\beta}} \in \mathcal{B}, t \in \Omega} \left| \mathcal{V}_{\ell,\ell'}(\tilde{\boldsymbol{\beta}}, t) - v_{\ell,\ell'}(\tilde{\boldsymbol{\beta}}, t) \right| = o_{\mathrm{P}}(1),
\end{aligned}$$

where $\sum_{t \in \Omega} \Delta N(t)/n = \int_0^\zeta dN(t)/n = N(\delta)/n \leq 1$. As $\boldsymbol{\beta} \in \mathcal{B}$, one can show that the second part of the right-hand side of (A.1) is $o_{\mathrm{P}}(1)$ similarly. For the third part of the (A.1), we have

$$\begin{aligned}
\left| \int_0^\zeta \left\{ v_{\ell,\ell'}(\widehat{\boldsymbol{\beta}}, t) - v_{\ell,\ell'}(\boldsymbol{\beta}, t) \right\} \frac{dN(t)}{n} \right| &= \left| \sum_{t \in \Omega} \left\{ v_{\ell,\ell'}(\widehat{\boldsymbol{\beta}}, t) - v_{\ell,\ell'}(\boldsymbol{\beta}, t) \right\} \frac{\Delta N(t)}{n} \right| \\
&\leq \sum_{t \in \Omega} \left| v_{\ell,\ell'}(\widehat{\boldsymbol{\beta}}, t) - v_{\ell,\ell'}(\boldsymbol{\beta}, t) \right| \frac{\Delta N(t)}{n}
\end{aligned}$$

$$\leq \sup_{t\in\Omega}\left|v_{\ell,\ell'}(\widehat{\boldsymbol{\beta}},t)-v_{\ell,\ell'}(\boldsymbol{\beta},t)\right|\sum_{t\in\Omega}\frac{\Delta N(t)}{n}$$
$$\leq \sup_{t\in[0,\zeta]}\left|v_{\ell,\ell'}(\widehat{\boldsymbol{\beta}},t)-v_{\ell,\ell'}(\boldsymbol{\beta},t)\right| = o_{\mathrm{P}}(1).$$

Note that continuity of functions in Assumption A.3.7 with respect to $\tilde{\boldsymbol{\beta}} \in \mathcal{B}$ uniformly holds over $t \in [0,\zeta]$. So $g(\cdot) \equiv \sup_{t\in[0,\zeta]}|v_{\ell,\ell'}(\cdot,t)-v_{\ell,\ell'}(\boldsymbol{\beta},t)| : \mathcal{B} \to [0,\infty)$ is continuous at $\boldsymbol{\beta}$. Combined with $\widehat{\boldsymbol{\beta}} \xrightarrow{\mathrm{P}} \boldsymbol{\beta}$ and $\widehat{\boldsymbol{\beta}} \in \mathcal{B}$, by the continuous mapping principle, $g(\widehat{\boldsymbol{\beta}}) \xrightarrow{\mathrm{P}} g(\boldsymbol{\beta}) = 0$, which makes $o_{\mathrm{P}}(1)$ in the last line hold.

■

A.4.6 Proof of Theorem 2.3

The following proof is based on Kalbfleisch and Prentice [27] and Fleming and Harrington [16]. One can see that $\widehat{\boldsymbol{\beta}}_{\mathrm{b}}$ maximizes

$$\log\left(L^{\mathrm{b}}(\tilde{\boldsymbol{\beta}},\zeta)\right) = \sum_{i=1}^{n}\int_0^{\zeta}\left\{\boldsymbol{x}_i^{\top}\tilde{\boldsymbol{\beta}} - \log\left(S^{(0)}(\tilde{\boldsymbol{\beta}},t)\right)\right\}dN_i(t) + \sum_{t\in\Omega}\log(\Delta N(t)!).$$

One can see that $\widehat{\boldsymbol{\beta}}_{\mathrm{b}} = \arg\max_{\tilde{\boldsymbol{\beta}}\in\mathcal{B}} X(\tilde{\boldsymbol{\beta}},\zeta)$, where for $t \in [0,\zeta]$ and $\tilde{\boldsymbol{\beta}} \in \mathcal{B}$,

$$X(\tilde{\boldsymbol{\beta}},t) = \frac{1}{n}\{\log(L^{\mathrm{b}}(\tilde{\boldsymbol{\beta}},t)) - \log(L^{\mathrm{b}}(\boldsymbol{\beta},t))\} = \frac{1}{n}\sum_{i=1}^{n}\int_0^{t}\left\{\boldsymbol{x}_i^{\top}(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta}) - \log\left(\frac{S^{(0)}(\tilde{\boldsymbol{\beta}},s)}{S^{(0)}(\boldsymbol{\beta},s)}\right)\right\}dN_i(s).$$

Here, for $t \in [0,\zeta]$ and $\tilde{\boldsymbol{\beta}} \in \mathcal{B}$, let

$$\tilde{X}(\tilde{\boldsymbol{\beta}},t) = \frac{1}{n}\sum_{i=1}^{n}\int_0^{t}\left\{\boldsymbol{x}_i^{\top}(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta}) - \log\left(\frac{S^{(0)}(\tilde{\boldsymbol{\beta}},s)}{S^{(0)}(\boldsymbol{\beta},s)}\right)\right\}dA_i(s)$$

$$\begin{aligned}
&= \frac{1}{n}\sum_{i=1}^{n}\sum_{s\in\Omega_t}\left\{\boldsymbol{x}_i^\top(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta}) - \log\left(\frac{S^{(0)}(\tilde{\boldsymbol{\beta}},s)}{S^{(0)}(\boldsymbol{\beta},s)}\right)\right\} Y_i(s)\Delta\Lambda^*(s;\boldsymbol{x}_i)\\
&= \frac{1}{n}\sum_{i=1}^{n}\sum_{s\in\Omega_t}\left\{\boldsymbol{x}_i^\top(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta}) - \log\left(\frac{S^{(0)}(\tilde{\boldsymbol{\beta}},s)}{S^{(0)}(\boldsymbol{\beta},s)}\right)\right\} Y_i(s)\exp(\boldsymbol{x}_i^\top\boldsymbol{\beta})\Delta\Lambda_0^*(s)(1+a_i(\boldsymbol{\beta},s))\\
&= \frac{1}{n}\sum_{i=1}^{n}\sum_{s\in\Omega_t}\left\{\boldsymbol{x}_i^\top(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta}) - \log\left(\frac{S^{(0)}(\tilde{\boldsymbol{\beta}},s)}{S^{(0)}(\boldsymbol{\beta},s)}\right)\right\} Y_i(s)\exp(\boldsymbol{x}_i^\top\boldsymbol{\beta})\Delta\Lambda_0^*(s) + \mathcal{O}_{\mathrm{P}}(\tau)\\
&= \frac{1}{n}\sum_{s\in\Omega_t}\left\{S^{(1)}(\boldsymbol{\beta},s)^\top(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta}) - \log\left(\frac{S^{(0)}(\tilde{\boldsymbol{\beta}},s)}{S^{(0)}(\boldsymbol{\beta},s)}\right)S^{(0)}(\boldsymbol{\beta},s)\right\}\Delta\Lambda_0^*(s) + \mathcal{O}_{\mathrm{P}}(\tau),
\end{aligned}$$

where we use the fact that $\mathrm{Rem}_{\tilde{X}}(\tilde{\boldsymbol{\beta}},t) \equiv$

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{s\in\Omega_t}\left\{\boldsymbol{x}_i^\top(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta}) - \log\left(\frac{S^{(0)}(\tilde{\boldsymbol{\beta}},s)}{S^{(0)}(\boldsymbol{\beta},s)}\right)\right\} Y_i(s)\exp(\boldsymbol{x}_i^\top\boldsymbol{\beta})\Delta\Lambda_0^*(s)a_i(\boldsymbol{\beta},s) = \mathcal{O}_{\mathrm{P}}(\tau)$$

for $t\in[0,\zeta]$ and $\tilde{\boldsymbol{\beta}}\in\mathcal{B}$. To see this,

$$\left|\mathrm{Rem}_{\tilde{X}}(\tilde{\boldsymbol{\beta}},t)\right| \leq \sum_{i=1}^{n}\sum_{t\in\Omega}\left|\frac{1}{n}\left\{\boldsymbol{x}_i^\top(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta}) - \log\left(\frac{S^{(0)}(\tilde{\boldsymbol{\beta}},t)}{S^{(0)}(\boldsymbol{\beta},t)}\right)\right\} Y_i(t)\exp(\boldsymbol{x}_i^\top\boldsymbol{\beta})\Delta\Lambda_0^*(t)a_i(\boldsymbol{\beta},t)\right|$$

$$\leq \frac{n\zeta}{\tau}\sup_{t\in\Omega,i}\left|\frac{1}{n}\left\{\boldsymbol{x}_i^\top(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta}) - \log\left(\frac{S^{(0)}(\tilde{\boldsymbol{\beta}},t)}{S^{(0)}(\boldsymbol{\beta},t)}\right)\right\} Y_i(t)\exp(\boldsymbol{x}_i^\top\boldsymbol{\beta})\Delta\Lambda_0^*(t)a_i(\boldsymbol{\beta},t)\right| = \mathcal{O}_{\mathrm{P}}\left(\frac{n\zeta}{\tau}\frac{\tau^2}{n}\right) = \mathcal{O}_{\mathrm{P}}(\tau)$$

by

$$\begin{aligned}
&\sup_{t\in\Omega,i}\left|\frac{1}{n}\left\{\boldsymbol{x}_i^\top(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta}) - \log\left(\frac{S^{(0)}(\tilde{\boldsymbol{\beta}},t)/n}{S^{(0)}(\boldsymbol{\beta},t)/n}\right)\right\} Y_i(t)\exp(\boldsymbol{x}_i^\top\boldsymbol{\beta})\Delta\Lambda_0^*(t)a_i(\boldsymbol{\beta},t)\right|\\
&\quad\leq \frac{1}{n}\sup_{t\in\Omega,i}\left|\boldsymbol{x}_i^\top(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta}) - \log\left(\frac{S^{(0)}(\tilde{\boldsymbol{\beta}},t)/n}{S^{(0)}(\boldsymbol{\beta},t)/n}\right)\right|\sup_i\exp(\boldsymbol{x}_i^\top\boldsymbol{\beta})\sup_{t\in\Omega}\Delta\Lambda_0^*(t)\sup_{t\in\Omega,i}|a_i(\boldsymbol{\beta},t)|\\
&\quad= \frac{1}{n}\mathcal{O}_{\mathrm{P}}(1)\mathcal{O}_{\mathrm{P}}(1)\mathcal{O}_{\mathrm{P}}(\tau)\mathcal{O}_{\mathrm{P}}(\tau) = \mathcal{O}_{\mathrm{P}}\left(\frac{\tau^2}{n}\right),
\end{aligned}$$

where we used the fact that

$$\sup_{t\in\Omega,i}\left|\boldsymbol{x}_i^\top(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta})-\log\left(\frac{S^{(0)}(\tilde{\boldsymbol{\beta}},t)/n}{S^{(0)}(\boldsymbol{\beta},t)/n}\right)\right|$$
$$\leq \sup_i|\boldsymbol{x}_i^\top(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta})|+\sup_{t\in\Omega}|\log\{S^{(0)}(\tilde{\boldsymbol{\beta}},t)/n\}|+\sup_{t\in\Omega}|\log\{S^{(0)}(\boldsymbol{\beta},t)/n\}|=\mathcal{O}_{\mathrm{P}}(1),$$

where $\sup_i|\boldsymbol{x}_i^\top(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta})|\leq\sup_i\|\boldsymbol{x}_i\|_2\|\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta}\|_2\leq\sqrt{d}\rho(\|\tilde{\boldsymbol{\beta}}\|_2+\|\boldsymbol{\beta}\|_2)\leq 2\sqrt{d}\rho C_{\mathcal{B}}=\mathcal{O}(1)$.

We can show the convergence of $\tilde{X}(\tilde{\boldsymbol{\beta}},\zeta)$ for $\tilde{\boldsymbol{\beta}}\in\mathcal{B}$ as below

$$\begin{aligned}\tilde{X}(\tilde{\boldsymbol{\beta}},\zeta)&=\frac{1}{n}\int_0^\zeta\left\{S^{(1)}(\boldsymbol{\beta},t)^\top(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta})-\log\left(\frac{S^{(0)}(\tilde{\boldsymbol{\beta}},t)}{S^{(0)}(\boldsymbol{\beta},t)}\right)S^{(0)}(\boldsymbol{\beta},t)\right\}d\Lambda_0^*(t)+\mathcal{O}_{\mathrm{P}}(\tau)\\&=\int_0^\zeta\left\{\frac{S^{(1)}(\boldsymbol{\beta},t)^\top}{n}(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta})-\log\left(\frac{S^{(0)}(\tilde{\boldsymbol{\beta}},t)/n}{S^{(0)}(\boldsymbol{\beta},t)/n}\right)\frac{S^{(0)}(\boldsymbol{\beta},t)}{n}\right\}d\Lambda_0^*(t)+\mathcal{O}_{\mathrm{P}}(\tau)\\&\overset{\mathrm{P}}{\to}\int_0^\zeta\left\{s^{(1)}(\boldsymbol{\beta},t)^\top(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta})-\log\left(\frac{s^{(0)}(\tilde{\boldsymbol{\beta}},t)}{s^{(0)}(\boldsymbol{\beta},t)}\right)s^{(0)}(\boldsymbol{\beta},t)\right\}d\Lambda_0(t)=f(\tilde{\boldsymbol{\beta}}),\end{aligned}$$

where the $\overset{\mathrm{P}}{\to}$ in the last line is by the following. Let $A_n(t)=S^{(1)}(\boldsymbol{\beta},t)^\top(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta})/n$, $a(t)=s^{(1)}(\boldsymbol{\beta},t)^\top(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta})$, $B_n(t)=\log\left(\{S^{(0)}(\tilde{\boldsymbol{\beta}},t)/n\}/\{S^{(0)}(\boldsymbol{\beta},t)/n\}\right)S^{(0)}(\boldsymbol{\beta},t)/n$, and $b(t)=\log\left(s^{(0)}(\tilde{\boldsymbol{\beta}},t)/s^{(0)}(\boldsymbol{\beta},t)\right)s^{(0)}(\boldsymbol{\beta},t)$. We have

$$\begin{aligned}\sup_{t\in\Omega}|A_n(t)-a(t)|&\leq\sup_{t\in\Omega}\|S^{(1)}(\boldsymbol{\beta},t)/n-s^{(1)}(\boldsymbol{\beta},t)\|_2\|\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta}\|_2\\&\leq 2C_{\mathcal{B}}\sup_{t\in\Omega}\|S^{(1)}(\boldsymbol{\beta},t)/n-s^{(1)}(\boldsymbol{\beta},t)\|_\infty=o_{\mathrm{P}}(1),\end{aligned}$$

$$\begin{aligned}\sup_{t\in\Omega}|B_n(t)-b(t)|\leq&\sup_{t\in\Omega}\left|\log\left(\frac{S^{(0)}(\tilde{\boldsymbol{\beta}},t)/n}{S^{(0)}(\boldsymbol{\beta},t)/n}\right)\left\{\frac{S^{(0)}(\boldsymbol{\beta},t)}{n}-s^{(0)}(\boldsymbol{\beta},t)\right\}\right|\\&+\sup_{t\in\Omega}\left|s^{(0)}(\boldsymbol{\beta},t)\left\{\log\left(\frac{S^{(0)}(\tilde{\boldsymbol{\beta}},t)/n}{S^{(0)}(\boldsymbol{\beta},t)/n}\right)-\log\left(\frac{s^{(0)}(\tilde{\boldsymbol{\beta}},t)}{s^{(0)}(\boldsymbol{\beta},t)}\right)\right\}\right|\end{aligned}$$

$$\leq \left\{ \sup_{t\in\Omega} |\log(S^{(0)}(\tilde{\boldsymbol{\beta}},t)/n)| + \sup_{t\in\Omega} |\log(S^{(0)}(\boldsymbol{\beta},t)/n)| \right\} \sup_{t\in\Omega} \left| \frac{S^{(0)}(\boldsymbol{\beta},t)}{n} - s^{(0)}(\boldsymbol{\beta},t) \right|$$

$$+ \sup_{t\in\Omega} s^{(0)}(\boldsymbol{\beta},t) \left\{ \sup_{t\in\Omega} \left| \log\left\{ \frac{S^{(0)}(\tilde{\boldsymbol{\beta}},t)/n}{s^{(0)}(\tilde{\boldsymbol{\beta}},t)} \right\} \right| + \sup_{t\in\Omega} \left| \log\left\{ \frac{S^{(0)}(\boldsymbol{\beta},t)/n}{s^{(0)}(\boldsymbol{\beta},t)} \right\} \right| \right\}$$

$$= (\mathcal{O}_{\mathrm{P}}(1) + \mathcal{O}_{\mathrm{P}}(1))\, o_{\mathrm{P}}(1) + \mathcal{O}(1)\,(o_{\mathrm{P}}(1) + o_{\mathrm{P}}(1)) = o_{\mathrm{P}}(1), \text{ and}$$

$\sup_{t\in\Omega} |(A_n(t) - B_n(t)) - (a(t) - b(t))| \leq \sup_{t\in\Omega} |A_n(t) - a(t)| + \sup_{t\in\Omega} |B_n(t) - b(t)| = o_{\mathrm{P}}(1)$. One can see that $\partial f(\tilde{\boldsymbol{\beta}})/(\partial\tilde{\boldsymbol{\beta}})\big|_{\tilde{\boldsymbol{\beta}}=\boldsymbol{\beta}} = \mathbf{0}$ and $-\partial^2 f(\tilde{\boldsymbol{\beta}})/(\partial\tilde{\boldsymbol{\beta}}\partial\tilde{\boldsymbol{\beta}}^\top)\big|_{\tilde{\boldsymbol{\beta}}=\boldsymbol{\beta}} = \Sigma(\boldsymbol{\beta},\zeta)$ is positive definite by Assumption A.3.7, indicating $\boldsymbol{\beta} = \arg\max_{\tilde{\boldsymbol{\beta}}\in\mathcal{B}} f(\tilde{\boldsymbol{\beta}})$.

For $t \in [0,\zeta]$ and $\tilde{\boldsymbol{\beta}} \in \mathcal{B}$, the predictable variation process for the martingale $X(\tilde{\boldsymbol{\beta}},t) - \tilde{X}(\tilde{\boldsymbol{\beta}},t)$ is

$$\langle X(\tilde{\boldsymbol{\beta}}) - \tilde{X}(\tilde{\boldsymbol{\beta}})\rangle(t) = \sum_{i=1}^{n} \int_0^t \frac{1}{n^2} \left\{ \boldsymbol{x}_i^\top(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}) - \log\left(\frac{S^{(0)}(\tilde{\boldsymbol{\beta}},s)}{S^{(0)}(\boldsymbol{\beta},s)} \right) \right\}^2 d\langle M_i\rangle(s),$$

which is non-decreasing in $t \in [0,\zeta]$. So, for $\tilde{\boldsymbol{\beta}} \in \mathcal{B}$,

$$\sup_{t\in[0,\zeta]} \left| \langle X(\tilde{\boldsymbol{\beta}}) - \tilde{X}(\tilde{\boldsymbol{\beta}})\rangle(t) \right| = \langle X(\tilde{\boldsymbol{\beta}}) - \tilde{X}(\tilde{\boldsymbol{\beta}})\rangle(\zeta) = \mathcal{O}_{\mathrm{P}}\left(\frac{n\zeta}{\tau} \frac{\tau}{n^2} \right) = \mathcal{O}_{\mathrm{P}}\left(\frac{1}{n} \right) \xrightarrow{\mathrm{P}} 0$$

by Lemma A.1 where

$$\sup_{t\in\Omega,i} \left[\frac{1}{n^2} \left\{ \boldsymbol{x}_i^\top(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}) - \log\left(\frac{S^{(0)}(\tilde{\boldsymbol{\beta}},t)/n}{S^{(0)}(\boldsymbol{\beta},t)/n} \right) \right\}^2 d\langle M_i\rangle(t) \right]$$

$$\leq \frac{1}{n^2} \left\{ \sup_i |\boldsymbol{x}_i^\top(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})| + \sup_{t\in\Omega} |\log(S^{(0)}(\tilde{\boldsymbol{\beta}},t)/n)| + \sup_{t\in\Omega} |\log(S^{(0)}(\boldsymbol{\beta},t)/n)| \right\}^2 \sup_{t\in\Omega,i} d\langle M_i\rangle(t)$$

$$= \frac{1}{n^2} \mathcal{O}_{\mathrm{P}}(1)^2 \mathcal{O}_{\mathrm{P}}(\tau) = \mathcal{O}_{\mathrm{P}}\left(\frac{\tau}{n^2} \right).$$

By Lenglart's Inequality, which allows triangular array structure in filtration and every processes, for any $\delta > 0$, $\eta > 0$, and $\tilde{\boldsymbol{\beta}} \in \mathcal{B}$,

$$\Pr\left(\sup_{t\in[0,\zeta]} |X(\tilde{\boldsymbol{\beta}}, t) - \tilde{X}(\tilde{\boldsymbol{\beta}}, t)| > \sqrt{\eta}\right) \leq \frac{\delta}{\eta} + \Pr\left(\langle X(\tilde{\boldsymbol{\beta}}) - \tilde{X}(\tilde{\boldsymbol{\beta}})\rangle(\zeta) > \delta\right),$$

which means that $\sup_{t\in[0,\zeta]} |X(\tilde{\boldsymbol{\beta}}, t) - \tilde{X}(\tilde{\boldsymbol{\beta}}, t)| \xrightarrow{\mathrm{P}} 0$ for any $\tilde{\boldsymbol{\beta}} \in \mathcal{B}$.

Now one can see that $X(\tilde{\boldsymbol{\beta}}, \zeta) \xrightarrow{\mathrm{P}} f(\tilde{\boldsymbol{\beta}})$ for any $\tilde{\boldsymbol{\beta}} \in \mathcal{B}$. Considering the fact that $\widehat{\boldsymbol{\beta}}_{\mathrm{b}} = \arg\max_{\tilde{\boldsymbol{\beta}}\in\mathcal{B}} X(\tilde{\boldsymbol{\beta}}, \zeta)$ and $\boldsymbol{\beta} = \arg\max_{\tilde{\boldsymbol{\beta}}\in\mathcal{B}} f(\tilde{\boldsymbol{\beta}})$, where $X(\cdot, \zeta)$ and $f(\cdot)$ are concave functions on $\mathcal{B}$, we finally have $\widehat{\boldsymbol{\beta}}_{\mathrm{b}} \xrightarrow{\mathrm{P}} \boldsymbol{\beta}$. See Lemma 8.3.1 in Fleming and Harrington [16].

As $\widehat{\boldsymbol{\beta}}_{\mathrm{b}} \xrightarrow{\mathrm{P}} \boldsymbol{\beta}$, $\widehat{\boldsymbol{\beta}}_{\mathrm{b}} \in \mathcal{B}$, and $n^{-1}I_{\mathrm{b}}(\boldsymbol{\beta}, \zeta) \xrightarrow{\mathrm{P}} \Sigma(\boldsymbol{\beta}, \zeta)$, by Lemma A.8, $n^{-1}I_{\mathrm{b}}(\widehat{\boldsymbol{\beta}}_{\mathrm{b}}, \zeta) \xrightarrow{\mathrm{P}} \Sigma(\boldsymbol{\beta}, \zeta)$ is directly obtained. ■

A.4.7 Proof of Theorem 2.4

The following proof is based on Kalbfleisch and Prentice [27] and Fleming and Harrington [16]. For $t \in \Omega$, one can represent the predictable variation process as

$$d\langle M_i\rangle(t) = Y_i(t)d\Lambda^*(t; \boldsymbol{x}_i)\left(1 - \Delta\Lambda^*(t; \boldsymbol{x}_i)\right) = Y_i(t)\exp(\boldsymbol{x}_i^\top\boldsymbol{\beta})\Delta\Lambda_0^*(t)(1 + a_i(\boldsymbol{\beta}, t))\left(1 - \Delta\Lambda^*(t; \boldsymbol{x}_i)\right)$$

$$= Y_i(t)\exp(\boldsymbol{x}_i^\top\boldsymbol{\beta})\Delta\Lambda_0^*(t)(1 + d_i(\boldsymbol{\beta}, t)),$$

where $d_i(\boldsymbol{\beta}, t) = a_i(\boldsymbol{\beta}, t) - \Delta\Lambda^*(t; \boldsymbol{x}_i) - a_i(\boldsymbol{\beta}, t)\Delta\Lambda^*(t; \boldsymbol{x}_i)$. Note that

$$\sup_{t\in\Omega,i} |d_i(\boldsymbol{\beta}, t)| \leq \sup_{t\in\Omega,i} |a_i(\boldsymbol{\beta}, t)| + \sup_{t\in\Omega,i} \Delta\Lambda^*(t; \boldsymbol{x}_i) + \sup_{t\in\Omega,i} |a_i(\boldsymbol{\beta}, t)| \sup_{t\in\Omega,i} \Delta\Lambda^*(t; \boldsymbol{x}_i)$$
$$= \mathcal{O}(\tau) + \mathcal{O}(\tau) + \mathcal{O}(\tau)\mathcal{O}(\tau) = \mathcal{O}(\tau).$$

We have $n^{-1/2}U(\boldsymbol{\beta},\zeta) = \sum_{i=1}^n \int_0^\zeta G_i(t)dM_i(t)$. Then, we can further derive its predictable variation process as below

$$\langle n^{-1/2}U(\boldsymbol{\beta})\rangle(\zeta) = \sum_{i=1}^n \int_0^\zeta G_i(t)G_i(t)^\top d\langle M_i\rangle(t) = \sum_{i=1}^n \sum_{t\in\Omega} G_i(t)G_i(t)^\top Y_i(t)\Delta\Lambda^*(t;\boldsymbol{x}_i)\left(1-\Delta\Lambda^*(t;\boldsymbol{x}_i)\right)$$

$$\begin{aligned}
&= \sum_{i=1}^n \sum_{t\in\Omega} G_i(t)G_i(t)^\top Y_i(t)\exp(\boldsymbol{x}_i^\top\boldsymbol{\beta})\Delta\Lambda_0^*(t)\left(1+d_i(\boldsymbol{\beta},t)\right)\\
&= \sum_{i=1}^n \sum_{t\in\Omega} G_i(t)G_i(t)^\top Y_i(t)\exp(\boldsymbol{x}_i^\top\boldsymbol{\beta})\Delta\Lambda_0^*(t) + \mathcal{O}_{\mathrm{P}}(\tau)\\
&= \sum_{t\in\Omega} \mathcal{V}(\boldsymbol{\beta},t)\frac{S^{(0)}(\boldsymbol{\beta},t)}{n}\Delta\Lambda_0^*(t) + \mathcal{O}_{\mathrm{P}}(\tau) \xrightarrow{\mathrm{P}} \int_0^\zeta v(\boldsymbol{\beta},t)s^{(0)}(\boldsymbol{\beta},t)d\Lambda_0(t) = \Sigma(\boldsymbol{\beta},\zeta),
\end{aligned}$$

where we use the fact that

$$\sum_{i=1}^n \sum_{t\in\Omega} G_i(t)G_i(t)^\top Y_i(t)\exp(\boldsymbol{x}_i^\top\boldsymbol{\beta})\Delta\Lambda_0^*(t)d_i(\boldsymbol{\beta},t) = \mathcal{O}_{\mathrm{P}}\left(\frac{n\zeta}{\tau}\frac{\tau^2}{n}\right) = \mathcal{O}_{\mathrm{P}}(\tau).$$

To see this, let $G_i(t) = (G_{i1}(t),\ldots,G_{id}(t))^\top$. For all $\ell,\ell' \in \{1,\cdots,d\}$, we have

$$\begin{aligned}
&\sup_{t\in\Omega,i}\left|G_{i\ell}(t)G_{i\ell'}(t)Y_i(t)\exp(\boldsymbol{x}_i^\top\boldsymbol{\beta})\Delta\Lambda_0^*(t)d_i(\boldsymbol{\beta},t)\right|\\
&\quad\le \frac{1}{n}\sup_{t\in\Omega,i}\left|(x_{i\ell}-\epsilon_\ell(\boldsymbol{\beta},t))\right|\sup_{t\in\Omega,i}\left|(x_{i\ell'}-\epsilon_{\ell'}(\boldsymbol{\beta},t))\right|\sup_i\exp(\boldsymbol{x}_i^\top\boldsymbol{\beta})\sup_{t\in\Omega}\Delta\Lambda_0^*(t)\sup_{t\in\Omega,i}|d_i(\boldsymbol{\beta},t)| = \mathcal{O}_{\mathrm{P}}\left(\frac{\tau^2}{n}\right).
\end{aligned}$$

For $\ell = 1,\ldots,d$, define

$$n^{-1/2}U_{\epsilon,\ell}(\boldsymbol{\beta},\zeta) = \sum_{i=1}^n \int_0^\zeta G_{i\ell}(t)\mathbb{1}(|G_{i\ell}(t)|>\epsilon)dM_i(t).$$

For $\ell = 1, \ldots, d$ and for all $\epsilon > 0$, we have

$$
\begin{aligned}
&0 \leq \langle n^{-1/2} U_{\epsilon,\ell}(\boldsymbol{\beta}) \rangle(\zeta) = \sum_{i=1}^{n} \int_0^{\zeta} G_{i\ell}^2(t) \mathbb{1}(|G_{i\ell}(t)| > \epsilon) d\langle M_i \rangle(t) \\
&\leq 4 \sum_{i=1}^{n} \int_0^{\zeta} \frac{1}{n} x_{i\ell}^2 \mathbb{1}\left(\left| \frac{1}{\sqrt{n}} x_{i\ell} \right| > \frac{\epsilon}{2} \right) d\langle M_i \rangle(t) + 4 \sum_{i=1}^{n} \int_0^{\zeta} \frac{1}{n} \epsilon_\ell(\boldsymbol{\beta}, t)^2 \mathbb{1}\left(\left| \frac{1}{\sqrt{n}} \epsilon_\ell(\boldsymbol{\beta}, t) \right| > \frac{\epsilon}{2} \right) d\langle M_i \rangle(t) \\
&\leq 4 \mathbb{1}\left(\sup_i \left| \frac{1}{\sqrt{n}} x_{i\ell} \right| > \frac{\epsilon}{2} \right) \sum_{i=1}^{n} \int_0^{\zeta} \frac{1}{n} x_{i\ell}^2 d\langle M_i \rangle(t) + 4 \mathbb{1}\left(\sup_{t \in \Omega} \left| \frac{1}{\sqrt{n}} \epsilon_\ell(\boldsymbol{\beta}, t) \right| > \frac{\epsilon}{2} \right) \sum_{i=1}^{n} \int_0^{\zeta} \frac{1}{n} \epsilon_\ell(\boldsymbol{\beta}, t)^2 d\langle M_i \rangle(t) \\
&= o(1) \mathcal{O}_{\mathrm{P}}\left(\frac{n\zeta}{\tau} \frac{\tau}{n} \right) + o_{\mathrm{P}}(1) \mathcal{O}_{\mathrm{P}}\left(\frac{n\zeta}{\tau} \frac{\tau}{n} \right) = o_{\mathrm{P}}(1),
\end{aligned}
$$

where $0 \leq n^{-1/2} \sup_i |x_{i\ell}| \leq n^{-1/2} \rho = o(1)$, $\epsilon_\ell(\boldsymbol{\beta}, t) = \sum_{i=1}^{n} q_i(\boldsymbol{\beta}, t) x_{i\ell}$ with

$$
0 \leq \sup_{t \in \Omega} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} q_i(\boldsymbol{\beta}, t) x_{i\ell} \right| \leq \frac{1}{\sqrt{n}} \sup_{t \in \Omega} \sum_{i=1}^{n} q_i(\boldsymbol{\beta}, t) |x_{i\ell}| \leq \frac{1}{\sqrt{n}} \rho = o(1),
$$

$$
\sup_{t \in \Omega, i} \frac{1}{n} x_{i\ell}^2 d\langle M_i \rangle(t) \leq \frac{1}{n} \sup_i x_{i\ell}^2 \sup_{t \in \Omega, i} d\langle M_i \rangle(t) = \mathcal{O}_{\mathrm{P}}\left(\frac{\tau}{n} \right),
$$

$$
\sup_{t \in \Omega, i} \frac{1}{n} \epsilon_\ell(\boldsymbol{\beta}, t)^2 d\langle M_i \rangle(t) \leq \frac{1}{n} \left\{ \sup_{t \in \Omega} |\epsilon_\ell(\boldsymbol{\beta}, t)| \right\}^2 \sup_{t \in \Omega, i} d\langle M_i \rangle(t) = \mathcal{O}_{\mathrm{P}}\left(\frac{\tau}{n} \right),
$$

and we use the fact that

$$
|a - b|^2 \mathbb{1}(|a - b| > \epsilon) \leq 4|a|^2 \mathbb{1}(|a| > \epsilon/2) + 4|b|^2 \mathbb{1}(|b| > \epsilon/2).
$$

Then, by the martingale central limit theorem (Rebolledo's Theorem) in page 166 of Kalbfleisch and Prentice [27], which allows triangular array structure in filtration and every processes, we have $n^{-1/2} U(\boldsymbol{\beta}, \zeta) \xrightarrow{\mathrm{D}} \mathrm{N}(\mathbf{0}, \Sigma(\boldsymbol{\beta}, \zeta))$. As $U_{\mathrm{b}}(\boldsymbol{\beta}, t) = U(\boldsymbol{\beta}, t) + \mathcal{O}_{\mathrm{P}}(n\tau)$ for $t \in [0, \zeta]$ and since $n^{1/2}\tau \to 0$, we have $n^{-1/2} U_{\mathrm{b}}(\boldsymbol{\beta}, \zeta) = n^{-1/2} U(\boldsymbol{\beta}, \zeta) + o_{\mathrm{P}}(1) \xrightarrow{\mathrm{D}} \mathrm{N}(\mathbf{0}, \Sigma(\boldsymbol{\beta}, \zeta))$.

Lastly, we derive the limiting distribution for $\widehat{\boldsymbol{\beta}}_{\mathrm{b}}$. As $U_{\mathrm{b}}(\widehat{\boldsymbol{\beta}}_{\mathrm{b}}, \zeta) = \mathbf{0}$, by the Taylor series expansion, we have the following:

$$\begin{aligned}\mathbf{0} = n^{-1/2}U_{\mathrm{b}}(\widehat{\boldsymbol{\beta}}_{\mathrm{b}}, \zeta) &= n^{-1/2}U_{\mathrm{b}}(\boldsymbol{\beta}, \zeta) + \left\{ n^{-1/2}\frac{\partial}{\partial \tilde{\boldsymbol{\beta}}^{\top}} U_{\mathrm{b}}(\tilde{\boldsymbol{\beta}}, \zeta) \right\} (\widehat{\boldsymbol{\beta}}_{\mathrm{b}} - \boldsymbol{\beta}) \\ &= n^{-1/2}U_{\mathrm{b}}(\boldsymbol{\beta}, \zeta) - n^{-1}I_{\mathrm{b}}(\tilde{\boldsymbol{\beta}}, \zeta) n^{1/2}(\widehat{\boldsymbol{\beta}}_{\mathrm{b}} - \boldsymbol{\beta}),\end{aligned}$$

where $\tilde{\boldsymbol{\beta}}$ is between $\boldsymbol{\beta}$ and $\widehat{\boldsymbol{\beta}}_{\mathrm{b}}$, i.e., $\tilde{\boldsymbol{\beta}} = \theta\boldsymbol{\beta} + (1-\theta)\widehat{\boldsymbol{\beta}}_{\mathrm{b}}$ for some random variable $\theta \in [0,1]$. Then

$$n^{-1}I_{\mathrm{b}}(\tilde{\boldsymbol{\beta}}, \zeta) n^{1/2}(\widehat{\boldsymbol{\beta}}_{\mathrm{b}} - \boldsymbol{\beta}) = n^{-1/2}U_{\mathrm{b}}(\boldsymbol{\beta}, \zeta) \xrightarrow{\mathrm{D}} \mathrm{N}\left(\mathbf{0}, \Sigma(\boldsymbol{\beta}, \zeta)\right).$$

As $\widehat{\boldsymbol{\beta}}_{\mathrm{b}} \xrightarrow{\mathrm{P}} \boldsymbol{\beta}$ by Theorem 2.3 and $\tilde{\boldsymbol{\beta}}$ is between $\boldsymbol{\beta}$ and $\widehat{\boldsymbol{\beta}}_{\mathrm{b}}$, we have $\|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2 = \|(1-\theta)(\widehat{\boldsymbol{\beta}}_{\mathrm{b}} - \boldsymbol{\beta})\|_2 = |1-\theta| \|\widehat{\boldsymbol{\beta}}_{\mathrm{b}} - \boldsymbol{\beta}\|_2 \le \|\widehat{\boldsymbol{\beta}}_{\mathrm{b}} - \boldsymbol{\beta}\|_2 = o_{\mathrm{P}}(1)$, leading to $\tilde{\boldsymbol{\beta}} \xrightarrow{\mathrm{P}} \boldsymbol{\beta}$. As $\mathcal{B}$ is a convex set and $\boldsymbol{\beta}, \widehat{\boldsymbol{\beta}}_{\mathrm{b}} \in \mathcal{B}$, we have $\tilde{\boldsymbol{\beta}} \in \mathcal{B}$. Combined with $n^{-1}I_{\mathrm{b}}(\boldsymbol{\beta}, \zeta) \xrightarrow{\mathrm{P}} \Sigma(\boldsymbol{\beta}, \zeta)$, by Lemma A.8, we have $n^{-1}I_{\mathrm{b}}(\tilde{\boldsymbol{\beta}}, \zeta) \xrightarrow{\mathrm{P}} \Sigma(\boldsymbol{\beta}, \zeta)$. So using the Slutsky's theorem and continuous mapping principle, we finally have

$$\begin{aligned}&n^{1/2}(\widehat{\boldsymbol{\beta}}_{\mathrm{b}} - \boldsymbol{\beta}) = \left\{ n^{-1}I_{\mathrm{b}}(\tilde{\boldsymbol{\beta}}, \zeta) \right\}^{-1} n^{-1/2}U_{\mathrm{b}}(\boldsymbol{\beta}, \zeta) \xrightarrow{\mathrm{D}} \Sigma(\boldsymbol{\beta}, \zeta)^{-1}\mathrm{N}\left(\mathbf{0}, \Sigma(\boldsymbol{\beta}, \zeta)\right) \\ &= \mathrm{N}\left(\mathbf{0}, \Sigma(\boldsymbol{\beta}, \zeta)^{-1}\Sigma(\boldsymbol{\beta}, \zeta)\Sigma(\boldsymbol{\beta}, \zeta)^{-1}\right) = \mathrm{N}\left(\mathbf{0}, \Sigma(\boldsymbol{\beta}, \zeta)^{-1}\right).\end{aligned}$$

■

A.4.8 Basic Asymptotic Results for the PB Distribution Estimator under the Grouped Continuous Failure Time Model

Here we derive some asymptotic orders that are used in proofs for $\widehat{\boldsymbol{\beta}}_{\mathrm{pb}}$.

Lemma A.9. *Under Assumptions A.3.1 – A.3.5 and A.3.9,* $\sup_{t\in\Omega} d\widehat{\Lambda}(t) = \mathcal{O}_P(\tau)$.

Proof. We have $\sup_{t\in\Omega} d\widehat{\Lambda}(t) \le \{\sup_{t\in\Omega} d(t)\}\{\sup_{t\in\Omega} 1/n^*(t)\} = \mathcal{O}_{\mathrm{P}}(n\tau)\mathcal{O}_{\mathrm{P}}(1/n) = \mathcal{O}_{\mathrm{P}}(\tau)$. Thus it shows that $\sup_{t\in\Omega} d\widehat{\Lambda}(t) = \mathcal{O}_{\mathrm{P}}(\tau)$. ■

Now we check $d\widehat{\Lambda}_{\mathrm{na}}(\cdot)$ and $d\widehat{\Lambda}_{\mathrm{b}}(\cdot)$ satisfy Assumption A.3.9, which is a requirement for proper baseline hazard. For $t \in [0, \zeta]$,

$$\widehat{\Lambda}_{\mathrm{na}}(t) = \sum_{j=1}^{k} \widehat{\lambda}_{\mathrm{na},j} \mathbb{1}(t_{(j)} \le t) = \int_0^t d\widehat{\Lambda}_{\mathrm{na}}(s) = \sum_{s\in\Omega_t} \Delta\widehat{\Lambda}_{\mathrm{na}}(s),$$

where for all $t \in \Omega$, $d\widehat{\Lambda}_{\mathrm{na}}(t) = d(t)/n(t) = dN(t)/\sum_{i=1}^{n} Y_i(t)$. Similarly, for $t \in [0, \zeta]$,

$$\widehat{\Lambda}_{\mathrm{b}}(t) = \sum_{j=1}^{k} \widehat{\lambda}_{\mathrm{b}j} \mathbb{1}(t_{(j)} \le t) = \int_0^t d\widehat{\Lambda}_{\mathrm{b}}(s) = \sum_{s\in\Omega_t} \Delta\widehat{\Lambda}_{\mathrm{b}}(s),$$

where for all $t \in \Omega$, $d\widehat{\Lambda}_{\mathrm{b}}(t) = d(t)/\left\{\sum_{i\in\mathcal{R}(t)} \exp(\boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}}_{\mathrm{b}})\right\} = dN(t)/\left\{\sum_{i=1}^{n} Y_i(t) \exp(\boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}}_{\mathrm{b}})\right\}$. When $d(t) = dN(t) = \Delta N(t)$ is observed to be 0 ($t \notin \Omega^{\mathcal{K}}$) for some $t \in \Omega$, we also observe $d\widehat{\Lambda}_{\mathrm{na}}(t) = d\widehat{\Lambda}_{\mathrm{b}}(t) = 0$.

Lemma A.10. *Under Assumptions A.3.1 – A.3.5, $d\widehat{\Lambda}_{na}(\cdot)$ and $d\widehat{\Lambda}_{b}(\cdot)$ satisfy Assumption A.3.9.*

Proof. When $d\widehat{\Lambda}(\cdot) = d\widehat{\Lambda}_{\mathrm{na}}(\cdot)$, $n^*(\cdot) = n(\cdot)$. We know that $\sup_{t\in\Omega} n(t) = \mathcal{O}_{\mathrm{P}}(n)$ and $\sup_{t\in\Omega}\{1/n(t)\} = 1/\{\inf_{t\in\Omega} n(t)\} = 1/n(\zeta) = \mathcal{O}_{\mathrm{P}}(1/n)$.

When $d\widehat{\Lambda}(\cdot) = d\widehat{\Lambda}_{\mathrm{b}}(\cdot)$, $n^*(\cdot) = S^{(0)}(\widehat{\boldsymbol{\beta}}_{\mathrm{b}}, \cdot)$. As $\widehat{\boldsymbol{\beta}}_{\mathrm{b}} \in \mathcal{B}$, $0 \le \sup_{t\in\Omega} S^{(0)}(\widehat{\boldsymbol{\beta}}_{\mathrm{b}}, t) \le \sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B}} \sup_{t\in\Omega} S^{(0)}(\tilde{\boldsymbol{\beta}}, t)$ $= \sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B}, t\in\Omega} S^{(0)}(\tilde{\boldsymbol{\beta}}, t) = \mathcal{O}_{\mathrm{P}}(n)$, $0 \le \sup_{t\in\Omega} 1/S^{(0)}(\widehat{\boldsymbol{\beta}}_{\mathrm{b}}, t) \le \sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B}} \sup_{t\in\Omega} 1/S^{(0)}(\tilde{\boldsymbol{\beta}}, t) = \sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B}, t\in\Omega} 1/S^{(0)}(\tilde{\boldsymbol{\beta}}, t) = \mathcal{O}_{\mathrm{P}}(1/n)$, where we have $\sup_{t\in\Omega} 1/S^{(0)}(\widehat{\boldsymbol{\beta}}_{\mathrm{b}}, t) = 1/\inf_{t\in\Omega} S^{(0)}(\widehat{\boldsymbol{\beta}}_{\mathrm{b}}, t) = 1/S^{(0)}(\widehat{\boldsymbol{\beta}}_{\mathrm{b}}, \zeta)$. ■

Now we show asymptotic order for the PB distribution hazard. By the Taylor series expansion, for all $i \in \{1, \cdots, n\}$, $t \in \Omega$, and $\tilde{\boldsymbol{\beta}} \in \mathcal{B}$,

$$p_i(\tilde{\boldsymbol{\beta}}, t) = \exp(\boldsymbol{x}_i^\top \tilde{\boldsymbol{\beta}}) d\widehat{\Lambda}(t)(1 + \tilde{a}_i(\tilde{\boldsymbol{\beta}}, t)) = \exp(\boldsymbol{x}_i^\top \tilde{\boldsymbol{\beta}}) d\widehat{\Lambda}(t) + \tilde{b}_i(\tilde{\boldsymbol{\beta}}, t),$$

where

$$\tilde{a}_i(\tilde{\boldsymbol{\beta}}, t) = -\frac{\exp(-\tilde{c}_i(\tilde{\boldsymbol{\beta}}, t))}{2} \exp(\boldsymbol{x}_i^\top \tilde{\boldsymbol{\beta}}) d\widehat{\Lambda}(t)$$

with some $\tilde{c}_i(\tilde{\boldsymbol{\beta}}, t) \in [0, \exp(\boldsymbol{x}_i^\top \tilde{\boldsymbol{\beta}}) d\widehat{\Lambda}(t)]$. Note that when $d(t) = 0$, we have $d\widehat{\Lambda}(t) = p_i(\tilde{\boldsymbol{\beta}}, t) = \tilde{a}_i(\tilde{\boldsymbol{\beta}}, t) = \tilde{b}_i(\tilde{\boldsymbol{\beta}}, t) = 0$.

Lemma A.11. *Under Assumptions A.3.1 – A.3.5 and A.3.9, we have* $\sup_{\tilde{\boldsymbol{\beta}} \in \mathcal{B}, t \in \Omega, i} \exp(\boldsymbol{x}_i^\top \tilde{\boldsymbol{\beta}}) d\widehat{\Lambda}(t) = \mathcal{O}_P(\tau)$ *and*

$\sup_{\tilde{\boldsymbol{\beta}} \in \mathcal{B}, t \in \Omega, i} |\tilde{a}_i(\tilde{\boldsymbol{\beta}}, t)| = \mathcal{O}_P(\tau)$*, resulting in* $\sup_{\tilde{\boldsymbol{\beta}} \in \mathcal{B}, t \in \Omega, i} |\tilde{b}_i(\tilde{\boldsymbol{\beta}}, t)| = \mathcal{O}_P(\tau^2)$ *and* $\sup_{\tilde{\boldsymbol{\beta}} \in \mathcal{B}, t \in \Omega, i} p_i(\tilde{\boldsymbol{\beta}}, t) = \mathcal{O}_P(\tau)$.

Proof. First, we have

$$\sup_{\tilde{\boldsymbol{\beta}} \in \mathcal{B}, t \in \Omega, i} \exp(\boldsymbol{x}_i^\top \tilde{\boldsymbol{\beta}}) d\widehat{\Lambda}(t) \leq \sup_{\tilde{\boldsymbol{\beta}} \in \mathcal{B}, i} \exp(\boldsymbol{x}_i^\top \tilde{\boldsymbol{\beta}}) \sup_{t \in \Omega} d\widehat{\Lambda}(t) \leq \exp\left(\sqrt{d} \rho C_{\mathcal{B}}\right) \mathcal{O}_{\mathrm{P}}(\tau) = \mathcal{O}_{\mathrm{P}}(\tau),$$

implying that $\sup_{\tilde{\boldsymbol{\beta}} \in \mathcal{B}, t \in \Omega, i} \exp(\boldsymbol{x}_i^\top \tilde{\boldsymbol{\beta}}) d\widehat{\Lambda}(t) = \mathcal{O}_{\mathrm{P}}(\tau)$. Second, we have

$$\begin{aligned} \sup_{\tilde{\boldsymbol{\beta}} \in \mathcal{B}, t \in \Omega, i} |\tilde{a}_i(\tilde{\boldsymbol{\beta}}, t)| &\leq \sup_{\tilde{\boldsymbol{\beta}} \in \mathcal{B}, t \in \Omega, i} \left| \frac{\exp(-\tilde{c}_i(\tilde{\boldsymbol{\beta}}, t))}{2} \right| \sup_{\tilde{\boldsymbol{\beta}} \in \mathcal{B}, t \in \Omega, i} \exp(\boldsymbol{x}_i^\top \tilde{\boldsymbol{\beta}}) d\widehat{\Lambda}(t) \\ &\leq \frac{1}{2} \sup_{\tilde{\boldsymbol{\beta}} \in \mathcal{B}, t \in \Omega, i} \exp(\boldsymbol{x}_i^\top \tilde{\boldsymbol{\beta}}) d\widehat{\Lambda}(t) = \mathcal{O}_{\mathrm{P}}(\tau), \end{aligned}$$

implying that $\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega,i}|\tilde{a}_i(\tilde{\boldsymbol{\beta}},t)| = \mathcal{O}_{\mathrm{P}}(\tau)$. Using these results, we have

$$\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega,i}|\tilde{b}_i(\tilde{\boldsymbol{\beta}},t)| \leq \sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega,i}|\tilde{a}_i(\tilde{\boldsymbol{\beta}},t)| \sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega,i}\exp(\boldsymbol{x}_i^\top\tilde{\boldsymbol{\beta}})d\widehat{\Lambda}(t) = \mathcal{O}_{\mathrm{P}}(\tau)\mathcal{O}_{\mathrm{P}}(\tau) = \mathcal{O}_{\mathrm{P}}(\tau^2) \text{ and}$$

$$\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega,i} p_i(\tilde{\boldsymbol{\beta}},t) \leq \sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega,i}\exp(\boldsymbol{x}_i^\top\tilde{\boldsymbol{\beta}})d\widehat{\Lambda}(t) + \sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega,i}|\tilde{b}_i(\tilde{\boldsymbol{\beta}},t)| = \mathcal{O}_{\mathrm{P}}(\tau) + \mathcal{O}_{\mathrm{P}}(\tau^2) = \mathcal{O}_{\mathrm{P}}(\tau).$$

■

The following convergence results are directly used in proofs for $\widehat{\boldsymbol{\beta}}_{\mathrm{pb}}$.

Lemma A.12. *When Assumptions A.3.1 – A.3.5 and A.3.8 – A.3.9 are satisfied,*

$$\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega}\left|\log\left(B_t(\tilde{\boldsymbol{\beta}})\right) - \log\left(\frac{\mu(\tilde{\boldsymbol{\beta}},t)^{d(t)}\exp(-\mu(\tilde{\boldsymbol{\beta}},t))}{d(t)!}\right)\right| = \mathcal{O}_P\left(\frac{1}{n}\right),$$

$$\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega}\left|\log\left(\prod_{i\in\mathcal{R}(t)\setminus\mathcal{D}(t)}\left(1-p_i(\tilde{\boldsymbol{\beta}},t)\right)\right) - \log\left(\exp\left(-\mu(\tilde{\boldsymbol{\beta}},t)+\tilde{\mu}(\tilde{\boldsymbol{\beta}},t)\right)\right)\right| = \mathcal{O}_P\left(\frac{1}{n}\right),$$

$$\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega}\left|\log\left(\frac{\prod_{i\in\mathcal{D}(t)}p_i(\tilde{\boldsymbol{\beta}},t)}{\prod_{i\in\mathcal{D}(t)}\left\{\exp(\boldsymbol{x}_i^\top\tilde{\boldsymbol{\beta}})d\widehat{\Lambda}(t)\right\}}\right)\right| = \mathcal{O}_P\left(\frac{1}{n}\right), \text{ and}$$

$$\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega}\left|\log\left(\frac{\mu(\tilde{\boldsymbol{\beta}},t)^{d(t)}}{\left\{\sum_{i\in\mathcal{R}(t)}\exp(\boldsymbol{x}_i^\top\tilde{\boldsymbol{\beta}})d\widehat{\Lambda}(t)\right\}^{d(t)}}\right)\right| = \mathcal{O}_P\left(\frac{1}{n}\right), \text{ where}$$

$\mu(\tilde{\boldsymbol{\beta}},t) = \sum_{i\in\mathcal{R}(t)}p_i(\tilde{\boldsymbol{\beta}},t)$ *and* $\tilde{\mu}(\tilde{\boldsymbol{\beta}},t) = \sum_{i\in\mathcal{D}(t)}p_i(\tilde{\boldsymbol{\beta}},t)$.

Note that the first two results are based on Poisson approximation and the other results are based on approximating $p_i(\tilde{\boldsymbol{\beta}},t)$ by $\exp(\boldsymbol{x}_i^\top\tilde{\boldsymbol{\beta}})d\widehat{\Lambda}(t)$. The proof of Lemma A.12 is in Section A.5.1.

A.4.9 Proof of Theorem 2.5

Recall that for $t \in \Omega$ and $\tilde{\boldsymbol{\beta}} \in \mathcal{B}$,

$$L_t(\tilde{\boldsymbol{\beta}}) = \frac{A_t(\tilde{\boldsymbol{\beta}})}{B_t(\tilde{\boldsymbol{\beta}})} = \frac{\left\{\prod_{i\in\mathcal{D}(t)} p_i(\tilde{\boldsymbol{\beta}},t)\right\}\left\{\prod_{i\in\mathcal{R}(t)\setminus\mathcal{D}(t)}(1-p_i(\tilde{\boldsymbol{\beta}},t))\right\}}{\sum_{\mathcal{A}_{d(t)}\in\mathcal{F}_{d(t)}}\left\{\prod_{i\in\mathcal{A}_{d(t)}} p_i(\tilde{\boldsymbol{\beta}},t)\right\}\left\{\prod_{i\in\mathcal{R}(t)\setminus\mathcal{A}_{d(t)}}(1-p_i(\tilde{\boldsymbol{\beta}},t))\right\}},$$

$$L_t^{\mathrm{b}}(\tilde{\boldsymbol{\beta}}) = \frac{\exp\left(\sum_{i\in\mathcal{D}(t)} \boldsymbol{x}_i^\top \tilde{\boldsymbol{\beta}}\right)}{\left\{\sum_{i\in\mathcal{R}(t)} \exp(\boldsymbol{x}_i^\top \tilde{\boldsymbol{\beta}})\right\}^{d(t)}/d(t)!}.$$

For $t \in \Omega$ and $\tilde{\boldsymbol{\beta}} \in \mathcal{B}$,

$$\frac{L_t(\tilde{\boldsymbol{\beta}})}{L_t^{\mathrm{b}}(\tilde{\boldsymbol{\beta}})} = \frac{L_t(\tilde{\boldsymbol{\beta}})}{L_t^1(\tilde{\boldsymbol{\beta}})}\frac{L_t^1(\tilde{\boldsymbol{\beta}})}{L_t^2(\tilde{\boldsymbol{\beta}})}\frac{L_t^2(\tilde{\boldsymbol{\beta}})}{L_t^3(\tilde{\boldsymbol{\beta}})}\frac{L_t^3(\tilde{\boldsymbol{\beta}})}{L_t^{\mathrm{b}}(\tilde{\boldsymbol{\beta}})},$$

where

$$L_t^1(\tilde{\boldsymbol{\beta}}) = \frac{\left\{\prod_{i\in\mathcal{D}(t)} p_i(\tilde{\boldsymbol{\beta}},t)\right\}\exp\left(-\mu(\tilde{\boldsymbol{\beta}},t)+\tilde{\mu}(\tilde{\boldsymbol{\beta}},t)\right)}{\mu(\tilde{\boldsymbol{\beta}},t)^{d(t)}\exp(-\mu(\tilde{\boldsymbol{\beta}},t))/d(t)!},$$

$$L_t^2(\tilde{\boldsymbol{\beta}}) = \frac{\left\{\prod_{i\in\mathcal{D}(t)}\left\{\exp(\boldsymbol{x}_i^\top\tilde{\boldsymbol{\beta}})d\widehat{\Lambda}(t)\right\}\right\}\exp\left(-\mu(\tilde{\boldsymbol{\beta}},t)+\tilde{\mu}(\tilde{\boldsymbol{\beta}},t)\right)}{\mu(\tilde{\boldsymbol{\beta}},t)^{d(t)}\exp(-\mu(\tilde{\boldsymbol{\beta}},t))/d(t)!},$$

$$L_t^3(\tilde{\boldsymbol{\beta}}) = \frac{\left\{\prod_{i\in\mathcal{D}(t)}\left\{\exp(\boldsymbol{x}_i^\top\tilde{\boldsymbol{\beta}})d\widehat{\Lambda}(t)\right\}\right\}\exp\left(-\mu(\tilde{\boldsymbol{\beta}},t)+\tilde{\mu}(\tilde{\boldsymbol{\beta}},t)\right)}{\left\{\sum_{i\in\mathcal{R}(t)}\exp(\boldsymbol{x}_i^\top\tilde{\boldsymbol{\beta}})d\widehat{\Lambda}(t)\right\}^{d(t)}\exp(-\mu(\tilde{\boldsymbol{\beta}},t))/d(t)!}.$$

We have

$$\begin{aligned}\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega}\left|\log\left(\frac{L_t(\tilde{\boldsymbol{\beta}})}{L_t^1(\tilde{\boldsymbol{\beta}})}\right)\right| &\le \sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega}\left|\log\left(\frac{\prod_{i\in\mathcal{R}(t)\setminus\mathcal{D}(t)}\left(1-p_i(\tilde{\boldsymbol{\beta}},t)\right)}{\exp\left(-\mu(\tilde{\boldsymbol{\beta}},t)+\tilde{\mu}(\tilde{\boldsymbol{\beta}},t)\right)}\right)\right| \\ &\quad + \sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega}\left|\log\left(\frac{B_t(\tilde{\boldsymbol{\beta}})}{\mu(\tilde{\boldsymbol{\beta}},t)^{d(t)}\exp(-\mu(\tilde{\boldsymbol{\beta}},t))/d(t)!}\right)\right| \\ &= \mathcal{O}_{\mathrm{P}}\left(\frac{1}{n}\right) + \mathcal{O}_{\mathrm{P}}\left(\frac{1}{n}\right) = \mathcal{O}_{\mathrm{P}}\left(\frac{1}{n}\right),\end{aligned}$$

$$\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega}\left|\log\left(\frac{L_t^1(\tilde{\boldsymbol{\beta}})}{L_t^2(\tilde{\boldsymbol{\beta}})}\right)\right| = \sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega}\left|\log\left(\frac{\prod_{i\in\mathcal{D}(t)} p_i(\tilde{\boldsymbol{\beta}},t)}{\prod_{i\in\mathcal{D}(t)}\left\{\exp(\boldsymbol{x}_i^\top\tilde{\boldsymbol{\beta}})d\widehat{\Lambda}(t)\right\}}\right)\right| = \mathcal{O}_{\mathrm{P}}\left(\frac{1}{n}\right),$$

$$\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega}\left|\log\left(\frac{L_t^2(\tilde{\boldsymbol{\beta}})}{L_t^3(\tilde{\boldsymbol{\beta}})}\right)\right| = \sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega}\left|\log\left(\frac{\mu(\tilde{\boldsymbol{\beta}},t)^{d(t)}}{\left\{\sum_{i\in\mathcal{R}(t)}\exp(\boldsymbol{x}_i^\top\tilde{\boldsymbol{\beta}})d\widehat{\Lambda}(t)\right\}^{d(t)}}\right)\right| = \mathcal{O}_{\mathrm{P}}\left(\frac{1}{n}\right),$$

$$\begin{aligned}\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega}\left|\log\left(\frac{L_t^3(\tilde{\boldsymbol{\beta}})}{L_t^{\mathrm{b}}(\tilde{\boldsymbol{\beta}})}\right)\right| &= \sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega}\left|\tilde{\mu}(\tilde{\boldsymbol{\beta}},t)\right| = \sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega}\sum_{i\in\mathcal{D}(t)} p_i(\tilde{\boldsymbol{\beta}},t)\\ &\le \sup_{t\in\Omega} d(t)\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega,i} p_i(\tilde{\boldsymbol{\beta}},t) = \mathcal{O}_{\mathrm{P}}(n\tau)\mathcal{O}_{\mathrm{P}}(\tau) = \mathcal{O}_{\mathrm{P}}(n\tau^2) = \mathcal{O}_{\mathrm{P}}\left(\frac{1}{n}\right),\end{aligned}$$

where the last equality is due to $\tau \asymp 1/n$ by Assumption A.3.8. So, for $t \in \Omega$,

$$\begin{aligned}\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega}\left|\log\left(\frac{L_t(\tilde{\boldsymbol{\beta}})}{L_t^{\mathrm{b}}(\tilde{\boldsymbol{\beta}})}\right)\right| &\le \sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega}\left|\log\left(\frac{L_t(\tilde{\boldsymbol{\beta}})}{L_t^1(\tilde{\boldsymbol{\beta}})}\right)\right| + \sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega}\left|\log\left(\frac{L_t^1(\tilde{\boldsymbol{\beta}})}{L_t^2(\tilde{\boldsymbol{\beta}})}\right)\right|\\ &+ \sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega}\left|\log\left(\frac{L_t^2(\tilde{\boldsymbol{\beta}})}{L_t^3(\tilde{\boldsymbol{\beta}})}\right)\right| + \sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega}\left|\log\left(\frac{L_t^3(\tilde{\boldsymbol{\beta}})}{L_t^{\mathrm{b}}(\tilde{\boldsymbol{\beta}})}\right)\right| = \mathcal{O}_{\mathrm{P}}\left(\frac{1}{n}\right) = \mathcal{O}_{\mathrm{P}}(\tau).\end{aligned}$$

Note that when $d(t) = 0$ is observed for $t \in \Omega$, for any $\tilde{\boldsymbol{\beta}} \in \mathcal{B}$, we also observe $\log L_t(\tilde{\boldsymbol{\beta}}) = \log L_t^1(\tilde{\boldsymbol{\beta}}) = \log L_t^2(\tilde{\boldsymbol{\beta}}) = \log L_t^3(\tilde{\boldsymbol{\beta}}) = \log L_t^{\mathrm{b}}(\tilde{\boldsymbol{\beta}}) = 0$.

For $t \in [0, \zeta]$,

$$\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B}}\left|\log\left(L(\tilde{\boldsymbol{\beta}},\widehat{\boldsymbol{\Lambda}},t)\right) - \log\left(L^{\mathrm{b}}(\tilde{\boldsymbol{\beta}},t)\right)\right| = \sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B}}\left|\sum_{s\in\Omega_t}\log\left(\frac{L_s(\tilde{\boldsymbol{\beta}})}{L_s^{\mathrm{b}}(\tilde{\boldsymbol{\beta}})}\right)\right| \le \sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B}}\sum_{t\in\Omega}\left|\log\left(\frac{L_t(\tilde{\boldsymbol{\beta}})}{L_t^{\mathrm{b}}(\tilde{\boldsymbol{\beta}})}\right)\right|$$

$$\leq \frac{\zeta}{\tau} \sup_{\tilde{\boldsymbol{\beta}} \in \mathcal{B}, t \in \Omega} \left| \log \left(\frac{L_t(\tilde{\boldsymbol{\beta}})}{L_t^{\mathrm{b}}(\tilde{\boldsymbol{\beta}})} \right) \right| = \frac{\zeta}{\tau} \mathcal{O}_{\mathrm{P}}(\tau) = \mathcal{O}_{\mathrm{P}}(1).$$

■

A.4.10 Proof of Theorem 2.6

When Assumptions A.3.1 – A.3.9 are satisfied, by Theorems 2.3 and 2.5, we have $\sup_{\tilde{\boldsymbol{\beta}} \in \mathcal{B}} |X_{\mathrm{pb}}(\tilde{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Lambda}}, \zeta) - X(\tilde{\boldsymbol{\beta}}, \zeta)| = \mathcal{O}_{\mathrm{P}}(1/n) = o_{\mathrm{P}}(1)$, leading to $X_{\mathrm{pb}}(\tilde{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Lambda}}, \zeta) - X(\tilde{\boldsymbol{\beta}}, \zeta) = \mathcal{O}_{\mathrm{P}}(1/n) = o_{\mathrm{P}}(1)$ for any $\tilde{\boldsymbol{\beta}} \in \mathcal{B}$. We already have $X(\tilde{\boldsymbol{\beta}}, \zeta) \xrightarrow{\mathrm{P}} f(\tilde{\boldsymbol{\beta}})$ for any $\tilde{\boldsymbol{\beta}} \in \mathcal{B}$. So, we have $X_{\mathrm{pb}}(\tilde{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Lambda}}, \zeta) \xrightarrow{\mathrm{P}} f(\tilde{\boldsymbol{\beta}})$ for any $\tilde{\boldsymbol{\beta}} \in \mathcal{B}$. Considering the fact that $\widehat{\boldsymbol{\beta}}_{\mathrm{pb}} = \arg\max_{\tilde{\boldsymbol{\beta}} \in \mathcal{B}} X_{\mathrm{pb}}(\tilde{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Lambda}}, \zeta)$ and $\boldsymbol{\beta} = \arg\max_{\tilde{\boldsymbol{\beta}} \in \mathcal{B}} f(\tilde{\boldsymbol{\beta}})$, where $X_{\mathrm{pb}}(\cdot, \widehat{\boldsymbol{\Lambda}}, \zeta)$ and $f(\cdot)$ are concave functions on $\mathcal{B}$, we have $\widehat{\boldsymbol{\beta}}_{\mathrm{pb}} \xrightarrow{\mathrm{P}} \boldsymbol{\beta}$ by Lemma 8.3.1 in Fleming and Harrington [16].

As $\widehat{\boldsymbol{\beta}}_{\mathrm{pb}} \xrightarrow{\mathrm{P}} \boldsymbol{\beta}$, $\widehat{\boldsymbol{\beta}}_{\mathrm{pb}} \in \mathcal{B}$, and $n^{-1} I_{\mathrm{b}}(\boldsymbol{\beta}, \zeta) \xrightarrow{\mathrm{P}} \Sigma(\boldsymbol{\beta}, \zeta)$, by Lemma A.8, $n^{-1} I_{\mathrm{b}}(\widehat{\boldsymbol{\beta}}_{\mathrm{pb}}, \zeta) \xrightarrow{\mathrm{P}} \Sigma(\boldsymbol{\beta}, \zeta)$ is directly derived. ■

A.4.11 Proof of Theorem 2.7

Let $\tilde{C} = \max(C_{\mathrm{pb}}, C_{\mathrm{b}})$. Then, we have

$$\Pr\left(\|\widehat{\boldsymbol{\beta}}_{\mathrm{pb}} - \widehat{\boldsymbol{\beta}}_{\mathrm{b}}\|_2 \leq \tilde{C} n^{a} \left\{ X_{\mathrm{pb}}(\widehat{\boldsymbol{\beta}}_{\mathrm{pb}}, \widehat{\boldsymbol{\Lambda}}, \zeta) - X_{\mathrm{pb}}(\widehat{\boldsymbol{\beta}}_{\mathrm{b}}, \widehat{\boldsymbol{\Lambda}}, \zeta) \right\} \right) \to 1$$

and

$$\Pr\left(\|\widehat{\boldsymbol{\beta}}_{\mathrm{b}} - \widehat{\boldsymbol{\beta}}_{\mathrm{pb}}\|_2 \leq \tilde{C} n^{a} \left\{ X(\widehat{\boldsymbol{\beta}}_{\mathrm{b}}, \zeta) - X(\widehat{\boldsymbol{\beta}}_{\mathrm{pb}}, \zeta) \right\} \right) \to 1$$

by Assumption A.3.10. Consequently,

$$\mathrm{Pr}\left(\|\widehat{\boldsymbol{\beta}}_{\mathrm{pb}} - \widehat{\boldsymbol{\beta}}_{\mathrm{b}}\|_2 \le Cn^{a}\left\{X_{\mathrm{pb}}(\widehat{\boldsymbol{\beta}}_{\mathrm{pb}}, \widehat{\boldsymbol{\Lambda}}, \zeta) - X_{\mathrm{pb}}(\widehat{\boldsymbol{\beta}}_{\mathrm{b}}, \widehat{\boldsymbol{\Lambda}}, \zeta) + X(\widehat{\boldsymbol{\beta}}_{\mathrm{b}}, \zeta) - X(\widehat{\boldsymbol{\beta}}_{\mathrm{pb}}, \zeta)\right\}\right) \to 1,$$

where $C = \tilde{C}/2$. The above equation is equivalent to

$$\mathrm{Pr}\left(n^{1/2}\|\widehat{\boldsymbol{\beta}}_{\mathrm{pb}} - \widehat{\boldsymbol{\beta}}_{\mathrm{b}}\|_2 \le Cn^{a+1/2}\left\{X_{\mathrm{pb}}(\widehat{\boldsymbol{\beta}}_{\mathrm{pb}}, \widehat{\boldsymbol{\Lambda}}, \zeta) - X(\widehat{\boldsymbol{\beta}}_{\mathrm{pb}}, \zeta) + X(\widehat{\boldsymbol{\beta}}_{\mathrm{b}}, \zeta) - X_{\mathrm{pb}}(\widehat{\boldsymbol{\beta}}_{\mathrm{b}}, \widehat{\boldsymbol{\Lambda}}, \zeta)\right\}\right) \to 1, \tag{A.2}$$

where $a + 1/2 \in [1/2, 1)$. Let $\epsilon > 0$. Then the following holds:

$$\begin{aligned}
&\mathrm{Pr}\left(Cn^{a+1/2}\left\{X_{\mathrm{pb}}(\widehat{\boldsymbol{\beta}}_{\mathrm{pb}}, \widehat{\boldsymbol{\Lambda}}, \zeta) - X(\widehat{\boldsymbol{\beta}}_{\mathrm{pb}}, \zeta) + X(\widehat{\boldsymbol{\beta}}_{\mathrm{b}}, \zeta) - X_{\mathrm{pb}}(\widehat{\boldsymbol{\beta}}_{\mathrm{b}}, \widehat{\boldsymbol{\Lambda}}, \zeta)\right\} \le \epsilon\right)\\
&\ge \mathrm{Pr}\left(\left|Cn^{a+1/2}\left(X_{\mathrm{pb}}(\widehat{\boldsymbol{\beta}}_{\mathrm{pb}}, \widehat{\boldsymbol{\Lambda}}, \zeta) - X(\widehat{\boldsymbol{\beta}}_{\mathrm{pb}}, \zeta)\right)\right| + \left|Cn^{a+1/2}\left(X(\widehat{\boldsymbol{\beta}}_{\mathrm{b}}, \zeta) - X_{\mathrm{pb}}(\widehat{\boldsymbol{\beta}}_{\mathrm{b}}, \widehat{\boldsymbol{\Lambda}}, \zeta)\right)\right| \le \epsilon\right)\\
&\ge \mathrm{Pr}\left(\left|Cn^{a+1/2}\left(X_{\mathrm{pb}}(\widehat{\boldsymbol{\beta}}_{\mathrm{pb}}, \widehat{\boldsymbol{\Lambda}}, \zeta) - X(\widehat{\boldsymbol{\beta}}_{\mathrm{pb}}, \zeta)\right)\right| \le \frac{\epsilon}{2} \text{ and } \left|Cn^{a+1/2}\left(X(\widehat{\boldsymbol{\beta}}_{\mathrm{b}}, \zeta) - X_{\mathrm{pb}}(\widehat{\boldsymbol{\beta}}_{\mathrm{b}}, \widehat{\boldsymbol{\Lambda}}, \zeta)\right)\right| \le \frac{\epsilon}{2}\right)\\
&= 1 - \mathrm{Pr}\left(\left|Cn^{a+1/2}\left(X_{\mathrm{pb}}(\widehat{\boldsymbol{\beta}}_{\mathrm{pb}}, \widehat{\boldsymbol{\Lambda}}, \zeta) - X(\widehat{\boldsymbol{\beta}}_{\mathrm{pb}}, \zeta)\right)\right| > \frac{\epsilon}{2} \text{ or } \left|Cn^{a+1/2}\left(X(\widehat{\boldsymbol{\beta}}_{\mathrm{b}}, \zeta) - X_{\mathrm{pb}}(\widehat{\boldsymbol{\beta}}_{\mathrm{b}}, \widehat{\boldsymbol{\Lambda}}, \zeta)\right)\right| > \frac{\epsilon}{2}\right)\\
&\ge 1 - \mathrm{Pr}\left(\left|Cn^{a+1/2}\left(X_{\mathrm{pb}}(\widehat{\boldsymbol{\beta}}_{\mathrm{pb}}, \widehat{\boldsymbol{\Lambda}}, \zeta) - X(\widehat{\boldsymbol{\beta}}_{\mathrm{pb}}, \zeta)\right)\right| > \frac{\epsilon}{2}\right) - \mathrm{Pr}\left(\left|Cn^{a+1/2}\left(X(\widehat{\boldsymbol{\beta}}_{\mathrm{b}}, \zeta) - X_{\mathrm{pb}}(\widehat{\boldsymbol{\beta}}_{\mathrm{b}}, \widehat{\boldsymbol{\Lambda}}, \zeta)\right)\right| > \frac{\epsilon}{2}\right)\\
&\to 1,
\end{aligned}$$

where the $\to$ 1 in the last line is because $X_{\mathrm{pb}}(\widehat{\boldsymbol{\beta}}_{\mathrm{pb}}, \widehat{\boldsymbol{\Lambda}}, \zeta) - X(\widehat{\boldsymbol{\beta}}_{\mathrm{pb}}, \zeta) = \mathcal{O}_{\mathrm{P}}(1/n)$ and $X_{\mathrm{pb}}(\widehat{\boldsymbol{\beta}}_{\mathrm{b}}, \widehat{\boldsymbol{\Lambda}}, \zeta) - X(\widehat{\boldsymbol{\beta}}_{\mathrm{b}}, \zeta) = \mathcal{O}_{\mathrm{P}}(1/n)$. So, we have

$$\mathrm{Pr}\left(Cn^{a+1/2}\left\{X_{\mathrm{pb}}(\widehat{\boldsymbol{\beta}}_{\mathrm{pb}}, \widehat{\boldsymbol{\Lambda}}, \zeta) - X(\widehat{\boldsymbol{\beta}}_{\mathrm{pb}}, \zeta) + X(\widehat{\boldsymbol{\beta}}_{\mathrm{b}}, \zeta) - X_{\mathrm{pb}}(\widehat{\boldsymbol{\beta}}_{\mathrm{b}}, \widehat{\boldsymbol{\Lambda}}, \zeta)\right\} \le \epsilon\right) \to 1. \tag{A.3}$$

Considering (A.2) and (A.3), we have $\mathrm{Pr}\left(n^{1/2}\|\widehat{\boldsymbol{\beta}}_{\mathrm{pb}} - \widehat{\boldsymbol{\beta}}_{\mathrm{b}}\|_2 \le \epsilon\right) \to 1$, which means $n^{1/2}(\widehat{\boldsymbol{\beta}}_{\mathrm{pb}} - \widehat{\boldsymbol{\beta}}_{\mathrm{b}}) = o_{\mathrm{P}}(1)$. Then we have $n^{1/2}(\widehat{\boldsymbol{\beta}}_{\mathrm{pb}} - \boldsymbol{\beta}) = n^{1/2}(\widehat{\boldsymbol{\beta}}_{\mathrm{pb}} - \widehat{\boldsymbol{\beta}}_{\mathrm{b}}) + n^{1/2}(\widehat{\boldsymbol{\beta}}_{\mathrm{b}} - \boldsymbol{\beta}) = n^{1/2}(\widehat{\boldsymbol{\beta}}_{\mathrm{b}} - \boldsymbol{\beta}) + o_{\mathrm{P}}(1) \xrightarrow{\mathrm{D}} \mathrm{N}\left(\mathbf{0}, \Sigma(\boldsymbol{\beta}, \zeta)^{-1}\right)$ since $n^{1/2}(\widehat{\boldsymbol{\beta}}_{\mathrm{b}} - \boldsymbol{\beta}) \xrightarrow{\mathrm{D}} \mathrm{N}\left(\mathbf{0}, \Sigma(\boldsymbol{\beta}, \zeta)^{-1}\right)$. ■

A.4.12 Proof of Theorems 2.8 and 2.9

As no ties are allowed in a continuous model, $d_j = 1$ for all $j = 1, \ldots, k$. We use $\widehat{\boldsymbol{\Lambda}}_{\mathrm{e}}$ as $\widehat{\boldsymbol{\Lambda}}$. One can also use $\widehat{\boldsymbol{\Lambda}}_{\mathrm{na}}$ as $\widehat{\boldsymbol{\Lambda}}$, considering $\widehat{\lambda}_{\mathrm{na},j} = 1/n_j$. As there are no ties, the approximate partial likelihood estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{c}}$, the Breslow estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{b}}$, and the Efron estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{e}}$ are the same, and

$$d\widehat{\Lambda}_{\mathrm{e}}(t) = d\widehat{\Lambda}_{\mathrm{b}}(t) = \frac{1}{\sum_{i \in \mathcal{R}(t)} \exp(\boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}}_{\mathrm{e}})}$$

for $t = t_{(j)} \in \Omega^{\mathcal{K}}$ and $d\widehat{\Lambda}_{\mathrm{e}}(t) = d\widehat{\Lambda}_{\mathrm{b}}(t) = 0$ for $t \notin \Omega^{\mathcal{K}}$. We have $\widehat{\lambda}_{\mathrm{e}j} = d\widehat{\Lambda}_{\mathrm{e}}(t_{(j)})$ and $\widehat{\Lambda}_{\mathrm{e}}(t) = \sum_{s \le t, s \in \Omega^{\mathcal{K}}} d\widehat{\Lambda}_{\mathrm{e}}(s)$. Using similar techniques in the proof of Theorem 2.5, one can show that

$$\sup_{\tilde{\boldsymbol{\beta}} \in \mathcal{B}} \left| \mathcal{L}(\tilde{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Lambda}}_{\mathrm{e}}) - \mathcal{L}_{\mathrm{c}}(\tilde{\boldsymbol{\beta}}) \right| = \mathcal{O}_{\mathrm{P}}(1),$$

where $\mathcal{L}(\tilde{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Lambda}}_{\mathrm{e}}) = \log\left(L(\tilde{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Lambda}}_{\mathrm{e}})\right)$ and $\mathcal{L}_{\mathrm{c}}(\tilde{\boldsymbol{\beta}}) = \sum_{j=1}^{k} \left\{ \boldsymbol{x}_{j_1}^\top \tilde{\boldsymbol{\beta}} - \log\left(\sum_{i \in \mathcal{R}(t_{(j)})} \exp(\boldsymbol{x}_i^\top \tilde{\boldsymbol{\beta}}) \right) \right\}$ is the logarithm of the approximate partial likelihood (2.2). Under the continuous model, $\widehat{\boldsymbol{\beta}}_{\mathrm{c}} = \arg\max_{\tilde{\boldsymbol{\beta}} \in \mathcal{B}} \mathcal{L}_{\mathrm{c}}(\tilde{\boldsymbol{\beta}})$ satisfies the following under Assumptions A.3.2 and A.3.11 by Kalbfleisch and Prentice [27]:

(i) $n^{-1/2} U_{\mathrm{c}}(\boldsymbol{\beta}, \zeta) \xrightarrow{\mathrm{D}} \mathrm{N}\left(\mathbf{0}, \Sigma_{\mathrm{c}}(\boldsymbol{\beta}, \zeta)\right)$,

(ii) $n^{1/2}\left(\widehat{\boldsymbol{\beta}}_{\mathrm{c}} - \boldsymbol{\beta}\right) \xrightarrow{\mathrm{D}} \mathrm{N}\left(\mathbf{0}, \Sigma_{\mathrm{c}}(\boldsymbol{\beta}, \zeta)^{-1}\right)$, and

(iii) $n^{-1} I_{\mathrm{c}}(\widehat{\boldsymbol{\beta}}_{\mathrm{c}}, \zeta) \xrightarrow{\mathrm{P}} \Sigma_{\mathrm{c}}(\boldsymbol{\beta}, \zeta)$.

Showing the consistency and limiting covariance estimation of $\widehat{\boldsymbol{\beta}}_{\mathrm{pb}}$ is similar to the proof of Theorem 2.6 and showing the asymptotic normality of $\widehat{\boldsymbol{\beta}}_{\mathrm{pb}}$ is similar to the proof of Theorem 2.7. ■

A.5 Asymptotic Orders

A.5.1 Asymptotic Orders for the Grouped Continuous Failure Time Model

In this section, we derive the asymptotic orders results used in the proofs for Section 2.4.2 under the grouped continuous failure time model. Here we first derive some basic asymptotic order calculations.

Lemma A.13. *Let* $X_n \in \mathbb{R}$ *be a random variable and* $a_n > 0$ *satisfies* $a_n \to 0$ *as* $n \to \infty$. *If* $X_n = \mathcal{O}_P(a_n)$, *then* $\log(1 + X_n) = \mathcal{O}_P(a_n)$ *and* $1/(1 + X_n) = 1 + \mathcal{O}_P(a_n)$. *Similarly, if* $X_n = o_P(1)$, *then* $\log(1 + X_n) = o_P(1)$ *and* $1/(1 + X_n) = 1 + o_P(1)$.

Proof. We only prove the case when $X_n = \mathcal{O}_{\mathrm{P}}(a_n)$. By the Taylor series expansion,

$$\begin{aligned}\log(1 + X_n) &= X_n - \frac{1}{2(1+\tilde{X}_n)^2}X_n^2 = \mathcal{O}_{\mathrm{P}}(a_n) - \frac{1}{2(1+\mathcal{O}_{\mathrm{P}}(a_n))^2}\mathcal{O}_{\mathrm{P}}(a_n)^2 \\ &= \mathcal{O}_{\mathrm{P}}(a_n) + \mathcal{O}_{\mathrm{P}}(1)\mathcal{O}_{\mathrm{P}}(a_n)^2 = \mathcal{O}_{\mathrm{P}}(a_n)\end{aligned}$$

for some $\tilde{X}_n$ between 0 and X_n, where $\tilde{X}_n \in [0, X_n]$ when $X_n \geq 0$ and $\tilde{X}_n \in [X_n, 0]$ when $X_n < 0$. Similarly, the Taylor series expansion,

$$\begin{aligned}\frac{1}{1+X_n} &= 1 - X_n + \frac{1}{(1+X_n^*)^3}X_n^2 = 1 + \mathcal{O}_{\mathrm{P}}(a_n) + \frac{1}{(1+\mathcal{O}_{\mathrm{P}}(a_n))^3}\mathcal{O}_{\mathrm{P}}(a_n)^2 \\ &= 1 + \mathcal{O}_{\mathrm{P}}(a_n) + \mathcal{O}_{\mathrm{P}}(1)\mathcal{O}_{\mathrm{P}}(a_n)^2 = 1 + \mathcal{O}_{\mathrm{P}}(a_n)\end{aligned}$$

for some X_n^* between 0 and X_n, where $X_n^* \in [0, X_n]$ when $X_n \geq 0$ and $X_n^* \in [X_n, 0]$ when $X_n < 0$. ■

Lemma A.14. *Let stochastic processes* $A_n(\tilde{\boldsymbol{\beta}}, t), B_n(\tilde{\boldsymbol{\beta}}, t) > 0$ *for* $\tilde{\boldsymbol{\beta}} \in \mathcal{B}$ *and* $t \in \Omega$*, where* $\sup_{\tilde{\boldsymbol{\beta}} \in \mathcal{B}, t \in \Omega} A_n(\tilde{\boldsymbol{\beta}}, t) = \mathcal{O}_P(1)$*,* $\sup_{\tilde{\boldsymbol{\beta}} \in \mathcal{B}, t \in \Omega} \{1/A_n(\tilde{\boldsymbol{\beta}}, t)\} = \mathcal{O}_P(1)$*,* $\sup_{\tilde{\boldsymbol{\beta}} \in \mathcal{B}, t \in \Omega} B_n(\tilde{\boldsymbol{\beta}}, t) = \mathcal{O}_P(1)$*, and* $\sup_{\tilde{\boldsymbol{\beta}} \in \mathcal{B}, t \in \Omega} \{1/B_n(\tilde{\boldsymbol{\beta}}, t)\} = \mathcal{O}_P(1)$*. If* $\sup_{\tilde{\boldsymbol{\beta}} \in \mathcal{B}, t \in \Omega} |A_n(\tilde{\boldsymbol{\beta}}, t) - B_n(\tilde{\boldsymbol{\beta}}, t)| = \mathcal{O}_P(a_n)$ *for some* $a_n > 0$ *satisfying* $a_n \to 0$ *as* $n \to \infty$*, then*

$$\sup_{\tilde{\boldsymbol{\beta}} \in \mathcal{B}, t \in \Omega} \{A_n(\tilde{\boldsymbol{\beta}}, t)/B_n(\tilde{\boldsymbol{\beta}}, t)\} = 1 + \mathcal{O}_P(a_n) \text{ and } \sup_{\tilde{\boldsymbol{\beta}} \in \mathcal{B}, t \in \Omega} \{B_n(\tilde{\boldsymbol{\beta}}, t)/A_n(\tilde{\boldsymbol{\beta}}, t)\} = 1 + \mathcal{O}_P(a_n).$$

If $\sup_{\tilde{\boldsymbol{\beta}} \in \mathcal{B}, t \in \Omega} |A_n(\tilde{\boldsymbol{\beta}}, t) - B_n(\tilde{\boldsymbol{\beta}}, t)| = o_P(1)$*, then*

$$\sup_{\tilde{\boldsymbol{\beta}} \in \mathcal{B}, t \in \Omega} \{A_n(\tilde{\boldsymbol{\beta}}, t)/B_n(\tilde{\boldsymbol{\beta}}, t)\} = 1 + o_P(1) \text{ and } \sup_{\tilde{\boldsymbol{\beta}} \in \mathcal{B}, t \in \Omega} \{B_n(\tilde{\boldsymbol{\beta}}, t)/A_n(\tilde{\boldsymbol{\beta}}, t)\} = 1 + o_P(1).$$

Proof. It is sufficient to prove the first case. For $\tilde{\boldsymbol{\beta}} \in \mathcal{B}$ and $t \in \Omega$, $A_n(\tilde{\boldsymbol{\beta}}, t)/B_n(\tilde{\boldsymbol{\beta}}, t) > 0$ and $B_n(\tilde{\boldsymbol{\beta}}, t)/A_n(\tilde{\boldsymbol{\beta}}, t) > 0$ can be defined by the given condition. We have

$$\begin{aligned}
&\sup_{\tilde{\boldsymbol{\beta}} \in \mathcal{B}, t \in \Omega} \left| \frac{A_n(\tilde{\boldsymbol{\beta}}, t)}{B_n(\tilde{\boldsymbol{\beta}}, t)} \right| = \sup_{\tilde{\boldsymbol{\beta}} \in \mathcal{B}, t \in \Omega} \left| 1 + \frac{A_n(\tilde{\boldsymbol{\beta}}, t)}{B_n(\tilde{\boldsymbol{\beta}}, t)} - 1 \right| \leq 1 + \sup_{\tilde{\boldsymbol{\beta}} \in \mathcal{B}, t \in \Omega} \left| \frac{A_n(\tilde{\boldsymbol{\beta}}, t)}{B_n(\tilde{\boldsymbol{\beta}}, t)} - 1 \right|, \\
&1 = \sup_{\tilde{\boldsymbol{\beta}} \in \mathcal{B}, t \in \Omega} \left| \frac{A_n(\tilde{\boldsymbol{\beta}}, t)}{B_n(\tilde{\boldsymbol{\beta}}, t)} + 1 - \frac{A_n(\tilde{\boldsymbol{\beta}}, t)}{B_n(\tilde{\boldsymbol{\beta}}, t)} \right| \leq \sup_{\tilde{\boldsymbol{\beta}} \in \mathcal{B}, t \in \Omega} \left| \frac{A_n(\tilde{\boldsymbol{\beta}}, t)}{B_n(\tilde{\boldsymbol{\beta}}, t)} \right| + \sup_{\tilde{\boldsymbol{\beta}} \in \mathcal{B}, t \in \Omega} \left| \frac{A_n(\tilde{\boldsymbol{\beta}}, t)}{B_n(\tilde{\boldsymbol{\beta}}, t)} - 1 \right|, \\
&- \sup_{\tilde{\boldsymbol{\beta}} \in \mathcal{B}, t \in \Omega} \left| \frac{A_n(\tilde{\boldsymbol{\beta}}, t)}{B_n(\tilde{\boldsymbol{\beta}}, t)} - 1 \right| \leq \left\{ \sup_{\tilde{\boldsymbol{\beta}} \in \mathcal{B}, t \in \Omega} \left| \frac{A_n(\tilde{\boldsymbol{\beta}}, t)}{B_n(\tilde{\boldsymbol{\beta}}, t)} \right| \right\} - 1 = \left\{ \sup_{\tilde{\boldsymbol{\beta}} \in \mathcal{B}, t \in \Omega} \frac{A_n(\tilde{\boldsymbol{\beta}}, t)}{B_n(\tilde{\boldsymbol{\beta}}, t)} \right\} - 1 \\
&\leq \sup_{\tilde{\boldsymbol{\beta}} \in \mathcal{B}, t \in \Omega} \left| \frac{A_n(\tilde{\boldsymbol{\beta}}, t)}{B_n(\tilde{\boldsymbol{\beta}}, t)} - 1 \right|,
\end{aligned}$$

$$\begin{aligned}
\text{and } \left| \left\{ \sup_{\tilde{\boldsymbol{\beta}} \in \mathcal{B}, t \in \Omega} \frac{A_n(\tilde{\boldsymbol{\beta}}, t)}{B_n(\tilde{\boldsymbol{\beta}}, t)} \right\} - 1 \right| &\leq \sup_{\tilde{\boldsymbol{\beta}} \in \mathcal{B}, t \in \Omega} \left| \frac{A_n(\tilde{\boldsymbol{\beta}}, t)}{B_n(\tilde{\boldsymbol{\beta}}, t)} - 1 \right| = \sup_{\tilde{\boldsymbol{\beta}} \in \mathcal{B}, t \in \Omega} \left| \frac{A_n(\tilde{\boldsymbol{\beta}}, t) - B_n(\tilde{\boldsymbol{\beta}}, t)}{B_n(\tilde{\boldsymbol{\beta}}, t)} \right| \\
&\leq \sup_{\tilde{\boldsymbol{\beta}} \in \mathcal{B}, t \in \Omega} \left| \frac{1}{B_n(\tilde{\boldsymbol{\beta}}, t)} \right| \sup_{\tilde{\boldsymbol{\beta}} \in \mathcal{B}, t \in \Omega} \left| A_n(\tilde{\boldsymbol{\beta}}, t) - B_n(\tilde{\boldsymbol{\beta}}, t) \right| \\
&= \mathcal{O}_{\mathrm{P}}(1) \mathcal{O}_{\mathrm{P}}(a_n) = \mathcal{O}_{\mathrm{P}}(a_n),
\end{aligned}$$

which leads to $\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega}\left\{A_n(\tilde{\boldsymbol{\beta}},t)/B_n(\tilde{\boldsymbol{\beta}},t)\right\}=1+\mathcal{O}_{\mathrm{P}}(a_n)$. $\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega}\left\{B_n(\tilde{\boldsymbol{\beta}},t)/A_n(\tilde{\boldsymbol{\beta}},t)\right\}=1+\mathcal{O}_{\mathrm{P}}(a_n)$ can be shown similarly. ■

Now we derive order of ties.

Lemma A.15. *Under Assumptions A.3.1 – A.3.5,*

$$\sup_{t\in\Omega}\left\{\mathrm{E}(dN(t)|\mathcal{F}_{t^-})\right\}=\mathcal{O}_P(n\tau),$$

implying that $\sup_{t\in\Omega}dN(t)=\mathcal{O}_P(n\tau)$.

Proof.

$$\begin{aligned}0\le\sup_{t\in\Omega}\left\{\mathrm{E}(dN(t)|\mathcal{F}_{t^-})\right\}&=\sup_{t\in\Omega}\left\{\sum_{i=1}^n Y_i(t)d\Lambda^*(t;\boldsymbol{x}_i)\right\}\le\sup_{t\in\Omega}\left\{\sum_{i=1}^n d\Lambda^*(t;\boldsymbol{x}_i)\right\}\\&\le n\sup_{t\in\Omega,i}d\Lambda^*(t;\boldsymbol{x}_i)=n\mathcal{O}(\tau)=\mathcal{O}(n\tau),\end{aligned}$$

directly implying $\sup_{t\in\Omega}dN(t)=\mathcal{O}_{\mathrm{P}}(n\tau)$ by Assumption A.3.5. ■

Proof of the Lemma A.5

First, we show some properties for $Y_i(t)$. For all i and all $t\in\Omega$, we have $Y_i(t)\ge Y_i(\zeta)$. So, for all $t\in\Omega$, we have $n(t)=\sum_{i=1}^n Y_i(t)\ge\sum_{i=1}^n Y_i(\zeta)=n(\zeta)$, resulting in $n(\zeta)=\inf_{t\in\Omega}n(t)$. Similarly, for any $\tilde{\boldsymbol{\beta}}\in\mathcal{B}$, for all $t\in\Omega$, $S^{(0)}(\tilde{\boldsymbol{\beta}},t)=\sum_{i=1}^n Y_i(t)\exp(\boldsymbol{x}_i^\top\tilde{\boldsymbol{\beta}})\ge\sum_{i=1}^n Y_i(\zeta)\exp(\boldsymbol{x}_i^\top\tilde{\boldsymbol{\beta}})=S^{(0)}(\tilde{\boldsymbol{\beta}},\zeta)$, resulting in $S^{(0)}(\tilde{\boldsymbol{\beta}},\zeta)=\inf_{t\in\Omega}S^{(0)}(\tilde{\boldsymbol{\beta}},t)$.

As $\sup_{t\in\Omega}n(t)\le n$, $\sup_{t\in\Omega}n(t)=\mathcal{O}_{\mathrm{P}}(n)$. By Assumption A.3.4, we have $|n(\zeta)/n-y(\zeta)|\le\sup_{t\in\Omega}|n(t)/n-y(t)|\le\sup_{t\in[0,\zeta]}|n(t)/n-y(t)|\xrightarrow{\text{a.s.}}0$, leading to $|n(\zeta)/n-y(\zeta)|\xrightarrow{\mathrm{P}}0$. For all $\epsilon > 0$,

$$\Pr(n(y(\zeta) - \epsilon) \le n(\zeta)) \ge \Pr\left(-\epsilon \le n(\zeta)/n - y(\zeta) \le \epsilon\right) \to 1.$$

Set small enough $\tilde{\epsilon} \in (0, y(\zeta))$ to make $C_\zeta^{-1} = y(\zeta) - \tilde{\epsilon} > 0$. Note that $C_\zeta > 1$ due to ranges of $y(\zeta)$ and $\tilde{\epsilon}$. Then $\Pr(1/n(\zeta) \le C_\zeta/n) \to 1$. Note that

$$\sup_{t\in\Omega} \frac{1}{n(t)} = \frac{1}{\inf_{t\in\Omega} n(t)} = \frac{1}{n(\zeta)}, \text{ which leads to } \sup_{t\in\Omega} \frac{1}{n(t)} = \mathcal{O}_{\mathrm{P}}\left(\frac{1}{n}\right).$$

We have

$$\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega} S^{(0)}(\tilde{\boldsymbol{\beta}}, t) = \sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega} \sum_{i=1}^{n} Y_i(t)\exp(\boldsymbol{x}_i^\top \tilde{\boldsymbol{\beta}}) \le \sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B}} \sum_{i=1}^{n} \exp(\boldsymbol{x}_i^\top \tilde{\boldsymbol{\beta}}) \le n \sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},i} \exp(\boldsymbol{x}_i^\top \tilde{\boldsymbol{\beta}}) \le n\exp(\sqrt{d}\rho C_{\mathcal{B}}),$$

directly proving $\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega} S^{(0)}(\tilde{\boldsymbol{\beta}}, t) = \mathcal{O}_{\mathrm{P}}(n)$. Similarly,

$$\begin{aligned}\inf_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega} S^{(0)}(\tilde{\boldsymbol{\beta}}, t) &= \inf_{\tilde{\boldsymbol{\beta}}\in\mathcal{B}} \inf_{t\in\Omega} S^{(0)}(\tilde{\boldsymbol{\beta}}, t) = \inf_{\tilde{\boldsymbol{\beta}}\in\mathcal{B}} S^{(0)}(\tilde{\boldsymbol{\beta}}, \zeta) \ge \inf_{\tilde{\boldsymbol{\beta}}\in\mathcal{B}} n(\zeta) \inf_i \exp(\boldsymbol{x}_i^\top \tilde{\boldsymbol{\beta}}) \\ &= n(\zeta) \inf_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},i} \exp(\boldsymbol{x}_i^\top \tilde{\boldsymbol{\beta}}) \ge n(\zeta)\exp(-\sqrt{d}\rho C_{\mathcal{B}}),\end{aligned}$$

where we used the fact that $S^{(0)}(\tilde{\boldsymbol{\beta}}, \zeta) \ge n(\zeta)\inf_i \exp(\boldsymbol{x}_i^\top \tilde{\boldsymbol{\beta}})$ for any $\tilde{\boldsymbol{\beta}} \in \mathcal{B}$, resulting in $\inf_{\tilde{\boldsymbol{\beta}}\in\mathcal{B}} S^{(0)}(\tilde{\boldsymbol{\beta}}, \zeta) = S^{(0)}(\boldsymbol{\beta}^*, \zeta) \ge n(\zeta)\inf_i \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta}^*) \ge \inf_{\tilde{\boldsymbol{\beta}}\in\mathcal{B}} n(\zeta)\inf_i \exp(\boldsymbol{x}_i^\top \tilde{\boldsymbol{\beta}})$ for some $\boldsymbol{\beta}^* \in \mathcal{B}$.

As $n(\zeta) \ge 1$ by Assumption A.3.1, we have

$$\Pr\left(\frac{1}{\inf_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega} S^{(0)}(\tilde{\boldsymbol{\beta}}, t)} \le C_\zeta \exp(\sqrt{d}\rho C_{\mathcal{B}})\frac{1}{n}\right) \to 1.$$

Note that

$$\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega}\frac{1}{S^{(0)}(\tilde{\boldsymbol{\beta}},t)}=\frac{1}{\inf_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega}S^{(0)}(\tilde{\boldsymbol{\beta}},t)}=\frac{1}{\inf_{\tilde{\boldsymbol{\beta}}\in\mathcal{B}}S^{(0)}(\tilde{\boldsymbol{\beta}},\zeta)},$$

which leads to

$$\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega}\frac{1}{S^{(0)}(\tilde{\boldsymbol{\beta}},t)}=\mathcal{O}_{\mathrm{P}}\left(\frac{1}{n}\right).$$

For any $\ell,\ell'\in\{1,\cdots,d\}$,

$$\begin{aligned}\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega}\left|S_{\ell}^{(1)}(\tilde{\boldsymbol{\beta}},t)\right|&=\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega}\left|\sum_{i=1}^{n}Y_i(t)x_{i\ell}\exp(\boldsymbol{x}_i^{\top}\tilde{\boldsymbol{\beta}})\right|\le\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B}}\sum_{i=1}^{n}|x_{i\ell}|\exp(\boldsymbol{x}_i^{\top}\tilde{\boldsymbol{\beta}})\\&\le n\sup_{i}|x_{i\ell}|\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},i}\exp(\boldsymbol{x}_i^{\top}\tilde{\boldsymbol{\beta}})\le n\rho\exp(\sqrt{d}\rho C_{\mathcal{B}}),\end{aligned}$$

$$\begin{aligned}\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega}\left|S_{\ell,\ell'}^{(2)}(\tilde{\boldsymbol{\beta}},t)\right|&=\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega}\left|\sum_{i=1}^{n}Y_i(t)x_{i\ell}x_{i\ell'}\exp(\boldsymbol{x}_i^{\top}\tilde{\boldsymbol{\beta}})\right|\le\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},}\sum_{i=1}^{n}|x_{i\ell}|\,|x_{i\ell'}|\exp(\boldsymbol{x}_i^{\top}\tilde{\boldsymbol{\beta}})\\&\le n\sup_{i}|x_{i\ell}|\sup_{i}|x_{i\ell'}|\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},i}\exp(\boldsymbol{x}_i^{\top}\tilde{\boldsymbol{\beta}})\le n\rho^2\exp(\sqrt{d}\rho C_{\mathcal{B}}),\end{aligned}$$

implying that $\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega}\|S^{(1)}(\tilde{\boldsymbol{\beta}},t)\|_{\infty}=\mathcal{O}_{\mathrm{P}}(n)$ and $\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega}\|S^{(2)}(\tilde{\boldsymbol{\beta}},t)\|_{\infty}=\mathcal{O}_{\mathrm{P}}(n)$.

One can represent $\epsilon(\tilde{\boldsymbol{\beta}},t)=S^{(1)}(\tilde{\boldsymbol{\beta}},t)/S^{(0)}(\tilde{\boldsymbol{\beta}},t)$ and $\mathcal{V}(\tilde{\boldsymbol{\beta}},t)=S^{(2)}(\tilde{\boldsymbol{\beta}},t)/S^{(0)}(\tilde{\boldsymbol{\beta}},t)-\epsilon(\tilde{\boldsymbol{\beta}},t)\epsilon(\tilde{\boldsymbol{\beta}},t)^{\top}$. For all $\ell,\ell'\in\{1,\cdots,d\}$,

$$\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega}\left|\epsilon_{\ell}(\tilde{\boldsymbol{\beta}},t)\right|=\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega}\left|\frac{S_{\ell}^{(1)}(\tilde{\boldsymbol{\beta}},t)}{S^{(0)}(\tilde{\boldsymbol{\beta}},t)}\right|\le\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega}\left|S_{\ell}^{(1)}(\tilde{\boldsymbol{\beta}},t)\right|\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega}\frac{1}{S^{(0)}(\tilde{\boldsymbol{\beta}},t)}=\mathcal{O}_{\mathrm{P}}(n)\mathcal{O}_{\mathrm{P}}\left(\frac{1}{n}\right)=\mathcal{O}_{\mathrm{P}}(1),$$

$$\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega}\left|\mathcal{V}_{\ell,\ell'}(\tilde{\boldsymbol{\beta}},t)\right|=\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega}\left|\frac{S_{\ell,\ell'}^{(2)}(\tilde{\boldsymbol{\beta}},t)}{S^{(0)}(\tilde{\boldsymbol{\beta}},t)}-\epsilon_{\ell}(\tilde{\boldsymbol{\beta}},t)\epsilon_{\ell'}(\tilde{\boldsymbol{\beta}},t)\right|$$

$$\leq \sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega} \left|S^{(2)}_{\ell,\ell'}(\tilde{\boldsymbol{\beta}},t)\right| \sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega} \frac{1}{S^{(0)}(\tilde{\boldsymbol{\beta}},t)} + \sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega} \left|\epsilon_{\ell}(\tilde{\boldsymbol{\beta}},t)\right| \sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega} \left|\epsilon_{\ell'}(\tilde{\boldsymbol{\beta}},t)\right|$$

$$= \mathcal{O}_{\mathrm{P}}(n)\mathcal{O}_{\mathrm{P}}\left(\frac{1}{n}\right) + \mathcal{O}_{\mathrm{P}}(1)\mathcal{O}_{\mathrm{P}}(1) = \mathcal{O}_{\mathrm{P}}(1),$$

which shows $\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega} \|\epsilon(\tilde{\boldsymbol{\beta}},t)\|_{\infty} = \mathcal{O}_{\mathrm{P}}(1)$ and $\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega} \|\mathcal{V}(\tilde{\boldsymbol{\beta}},t)\|_{\infty} = \mathcal{O}_{\mathrm{P}}(1)$.

Now we show orders related to log. We have

$$\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega} \left|\log\left\{\frac{1}{n}S^{(0)}(\tilde{\boldsymbol{\beta}},t)\right\}\right| = \max\left\{\left|\log\left\{\frac{1}{n}\inf_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega} S^{(0)}(\tilde{\boldsymbol{\beta}},t)\right\}\right|, \left|\log\left\{\frac{1}{n}\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega} S^{(0)}(\tilde{\boldsymbol{\beta}},t)\right\}\right|\right\}. \tag{A.1}$$

From previous derivations, we have

$$\Pr\left(C_{\zeta}^{-1}\exp(-\sqrt{d}\rho C_{\mathcal{B}}) \leq \frac{1}{n}\inf_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega} S^{(0)}(\tilde{\boldsymbol{\beta}},t) \leq \frac{1}{n}\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega} S^{(0)}(\tilde{\boldsymbol{\beta}},t) \leq \exp(\sqrt{d}\rho C_{\mathcal{B}})\right) \to 1,$$

As $C_{\zeta} > 1$, We have $\mathcal{C}_1 = C_{\zeta}^{-1}\exp(-\sqrt{d}\rho C_{\mathcal{B}}) \in (0,1)$ and $\mathcal{C}_2 = \exp(\sqrt{d}\rho C_{\mathcal{B}}) \in (1,\infty)$. Then

$$C = \max\left\{|\log \mathcal{C}_1|, |\log \mathcal{C}_2|\right\} = \max\left\{\log C_{\zeta} + \sqrt{d}\rho C_{\mathcal{B}}, \sqrt{d}\rho C_{\mathcal{B}}\right\} = \log C_{\zeta} + \sqrt{d}\rho C_{\mathcal{B}} \in (0,\infty).$$

Considering (A.1),

$$\Pr\left(\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega} \left|\log\left\{\frac{1}{n}S^{(0)}(\tilde{\boldsymbol{\beta}},t)\right\}\right| \leq C\right) \to 1,$$

which proves $\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega} \left|\log\left\{S^{(0)}(\tilde{\boldsymbol{\beta}},t)/n\right\}\right| = \mathcal{O}_{\mathrm{P}}(1)$.

Now we show some convergence results under Assumption A.3.7, where for $j = 0,1,2$,

$$\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega} \|S^{(j)}(\tilde{\boldsymbol{\beta}},t)/n - s^{(j)}(\tilde{\boldsymbol{\beta}},t)\|_{\infty} \leq \sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in[0,\zeta]} \|S^{(j)}(\tilde{\boldsymbol{\beta}},t)/n - s^{(j)}(\tilde{\boldsymbol{\beta}},t)\|_{\infty} \xrightarrow{\mathrm{P}} 0,$$

implying that $\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega}\|S^{(j)}(\tilde{\boldsymbol{\beta}},t)/n-s^{(j)}(\tilde{\boldsymbol{\beta}},t)\|_\infty=o_{\mathrm{P}}(1)$. For all $\ell\in\{1,\cdots,d\}$,

$$\begin{aligned}\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega}\left|\frac{S^{(1)}_\ell(\tilde{\boldsymbol{\beta}},t)/n}{S^{(0)}(\tilde{\boldsymbol{\beta}},t)/n}-\frac{s^{(1)}_\ell(\tilde{\boldsymbol{\beta}},t)}{S^{(0)}(\tilde{\boldsymbol{\beta}},t)/n}\right| &\le \sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega}\frac{n}{S^{(0)}(\tilde{\boldsymbol{\beta}},t)}\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega}\left|S^{(1)}_\ell(\tilde{\boldsymbol{\beta}},t)/n-s^{(1)}_\ell(\tilde{\boldsymbol{\beta}},t)\right|\\ &= n\mathcal{O}_{\mathrm{P}}\left(\frac{1}{n}\right)o_{\mathrm{P}}(1)=o_{\mathrm{P}}(1),\end{aligned}$$

$$\begin{aligned}\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega}\left|\frac{s^{(1)}_\ell(\tilde{\boldsymbol{\beta}},t)}{S^{(0)}(\tilde{\boldsymbol{\beta}},t)/n}-\frac{s^{(1)}_\ell(\tilde{\boldsymbol{\beta}},t)}{s^{(0)}(\tilde{\boldsymbol{\beta}},t)}\right| &\le \sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega}\left|s^{(1)}_\ell(\tilde{\boldsymbol{\beta}},t)\right|\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega}\left|\frac{1}{S^{(0)}(\tilde{\boldsymbol{\beta}},t)/n}-\frac{1}{s^{(0)}(\tilde{\boldsymbol{\beta}},t)}\right|\\ &= \sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega}\left|s^{(1)}_\ell(\tilde{\boldsymbol{\beta}},t)\right|\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega}\left|\frac{s^{(0)}(\tilde{\boldsymbol{\beta}},t)-S^{(0)}(\tilde{\boldsymbol{\beta}},t)/n}{s^{(0)}(\tilde{\boldsymbol{\beta}},t)S^{(0)}(\tilde{\boldsymbol{\beta}},t)/n}\right|\\ &\le \sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega}\left|s^{(1)}_\ell(\tilde{\boldsymbol{\beta}},t)\right|\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega}\frac{1}{s^{(0)}(\tilde{\boldsymbol{\beta}},t)}\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega}\frac{n}{S^{(0)}(\tilde{\boldsymbol{\beta}},t)}\\ &\quad\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega}\left|S^{(0)}(\tilde{\boldsymbol{\beta}},t)/n-s^{(0)}(\tilde{\boldsymbol{\beta}},t)\right|\\ &= \mathcal{O}(1)\mathcal{O}(1)n\mathcal{O}_{\mathrm{P}}\left(\frac{1}{n}\right)o_{\mathrm{P}}(1)=o_{\mathrm{P}}(1),\end{aligned}$$

$$\begin{aligned}\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega}\left|\epsilon_\ell(\tilde{\boldsymbol{\beta}},t)-e_\ell(\tilde{\boldsymbol{\beta}},t)\right| &= \sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega}\left|\frac{S^{(1)}_\ell(\tilde{\boldsymbol{\beta}},t)/n}{S^{(0)}(\tilde{\boldsymbol{\beta}},t)/n}-\frac{s^{(1)}_\ell(\tilde{\boldsymbol{\beta}},t)}{s^{(0)}(\tilde{\boldsymbol{\beta}},t)}\right|\\ \le \sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega}\left|\frac{S^{(1)}_\ell(\tilde{\boldsymbol{\beta}},t)/n}{S^{(0)}(\tilde{\boldsymbol{\beta}},t)/n}-\frac{s^{(1)}_\ell(\tilde{\boldsymbol{\beta}},t)}{S^{(0)}(\tilde{\boldsymbol{\beta}},t)/n}\right| &+ \sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega}\left|\frac{s^{(1)}_\ell(\tilde{\boldsymbol{\beta}},t)}{S^{(0)}(\tilde{\boldsymbol{\beta}},t)/n}-\frac{s^{(1)}_\ell(\tilde{\boldsymbol{\beta}},t)}{s^{(0)}(\tilde{\boldsymbol{\beta}},t)}\right|=o_{\mathrm{P}}(1),\end{aligned}$$

which leads to $\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega}\|\epsilon(\tilde{\boldsymbol{\beta}},t)-e(\tilde{\boldsymbol{\beta}},t)\|_\infty=o_{\mathrm{P}}(1)$.

For all $\ell,\ell'\in\{1,\cdots,d\}$,

$$\begin{aligned}\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega}\left|\frac{S^{(2)}_{\ell,\ell'}(\tilde{\boldsymbol{\beta}},t)/n}{S^{(0)}(\tilde{\boldsymbol{\beta}},t)/n}-\frac{s^{(2)}_{\ell,\ell'}(\tilde{\boldsymbol{\beta}},t)}{S^{(0)}(\tilde{\boldsymbol{\beta}},t)/n}\right| &\le \sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega}\frac{n}{S^{(0)}(\tilde{\boldsymbol{\beta}},t)}\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega}\left|S^{(2)}_{\ell,\ell}(\tilde{\boldsymbol{\beta}},t)/n-s^{(2)}_{\ell,\ell'}(\tilde{\boldsymbol{\beta}},t)\right|\\ &= n\mathcal{O}_{\mathrm{P}}\left(\frac{1}{n}\right)o_{\mathrm{P}}(1)=o_{\mathrm{P}}(1),\end{aligned}$$

$$
\begin{aligned}
\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega}\left|\frac{s^{(2)}_{\ell,\ell'}(\tilde{\boldsymbol{\beta}},t)}{S^{(0)}(\tilde{\boldsymbol{\beta}},t)/n}-\frac{s^{(2)}_{\ell,\ell'}(\tilde{\boldsymbol{\beta}},t)}{s^{(0)}(\tilde{\boldsymbol{\beta}},t)}\right| &\leq \sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega}\left|s^{(2)}_{\ell,\ell'}(\tilde{\boldsymbol{\beta}},t)\right|\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega}\left|\frac{1}{S^{(0)}(\tilde{\boldsymbol{\beta}},t)/n}-\frac{1}{s^{(0)}(\tilde{\boldsymbol{\beta}},t)}\right| \\
&= \sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega}\left|s^{(2)}_{\ell,\ell'}(\tilde{\boldsymbol{\beta}},t)\right|\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega}\left|\frac{s^{(0)}(\tilde{\boldsymbol{\beta}},t)-S^{(0)}(\tilde{\boldsymbol{\beta}},t)/n}{s^{(0)}(\tilde{\boldsymbol{\beta}},t)S^{(0)}(\tilde{\boldsymbol{\beta}},t)/n}\right| \\
&\leq \sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega}\left|s^{(2)}_{\ell,\ell'}(\tilde{\boldsymbol{\beta}},t)\right|\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega}\frac{1}{s^{(0)}(\tilde{\boldsymbol{\beta}},t)}\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega}\frac{n}{S^{(0)}(\tilde{\boldsymbol{\beta}},t)} \\
&\quad \sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega}\left|S^{(0)}(\tilde{\boldsymbol{\beta}},t)/n-s^{(0)}(\tilde{\boldsymbol{\beta}},t)\right| \\
&= \mathcal{O}(1)\mathcal{O}(1)n\mathcal{O}_{\mathrm{P}}\left(\frac{1}{n}\right)o_{\mathrm{P}}(1)=o_{\mathrm{P}}(1),
\end{aligned}
$$

$$
\begin{aligned}
\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega}\left|\frac{S^{(2)}_{\ell,\ell'}(\tilde{\boldsymbol{\beta}},t)/n}{S^{(0)}(\tilde{\boldsymbol{\beta}},t)/n}-\frac{s^{(2)}_{\ell,\ell'}(\tilde{\boldsymbol{\beta}},t)}{s^{(0)}(\tilde{\boldsymbol{\beta}},t)}\right| &\leq \sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega}\left|\frac{S^{(2)}_{\ell,\ell'}(\tilde{\boldsymbol{\beta}},t)/n}{S^{(0)}(\tilde{\boldsymbol{\beta}},t)/n}-\frac{s^{(2)}_{\ell,\ell'}(\tilde{\boldsymbol{\beta}},t)}{S^{(0)}(\tilde{\boldsymbol{\beta}},t)/n}\right| \\
&\quad+\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega}\left|\frac{s^{(2)}_{\ell,\ell'}(\tilde{\boldsymbol{\beta}},t)}{S^{(0)}(\tilde{\boldsymbol{\beta}},t)/n}-\frac{s^{(2)}_{\ell,\ell'}(\tilde{\boldsymbol{\beta}},t)}{s^{(0)}(\tilde{\boldsymbol{\beta}},t)}\right| \\
&= o_{\mathrm{P}}(1),
\end{aligned}
$$

$$
\begin{aligned}
&\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega}\left|\mathcal{V}_{\ell,\ell'}(\tilde{\boldsymbol{\beta}},t)-v_{\ell,\ell'}(\tilde{\boldsymbol{\beta}},t)\right| \\
&\leq \sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega}\left|\frac{S^{(2)}_{\ell,\ell'}(\tilde{\boldsymbol{\beta}},t)/n}{S^{(0)}(\tilde{\boldsymbol{\beta}},t)/n}-\frac{s^{(2)}_{\ell,\ell'}(\tilde{\boldsymbol{\beta}},t)}{s^{(0)}(\tilde{\boldsymbol{\beta}},t)}\right|+\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega}\left|\epsilon_{\ell}(\tilde{\boldsymbol{\beta}},t)\epsilon_{\ell'}(\tilde{\boldsymbol{\beta}},t)-e_{\ell}(\tilde{\boldsymbol{\beta}},t)e_{\ell'}(\tilde{\boldsymbol{\beta}},t)\right| \\
&\leq \sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega}\left|\frac{S^{(2)}_{\ell,\ell'}(\tilde{\boldsymbol{\beta}},t)/n}{S^{(0)}(\tilde{\boldsymbol{\beta}},t)/n}-\frac{s^{(2)}_{\ell,\ell'}(\tilde{\boldsymbol{\beta}},t)}{s^{(0)}(\tilde{\boldsymbol{\beta}},t)}\right|+\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega}\left|\epsilon_{\ell}(\tilde{\boldsymbol{\beta}},t)\epsilon_{\ell'}(\tilde{\boldsymbol{\beta}},t)-\epsilon_{\ell}(\tilde{\boldsymbol{\beta}},t)e_{\ell'}(\tilde{\boldsymbol{\beta}},t)\right| \\
&+\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega}\left|\epsilon_{\ell}(\tilde{\boldsymbol{\beta}},t)e_{\ell'}(\tilde{\boldsymbol{\beta}},t)-e_{\ell}(\tilde{\boldsymbol{\beta}},t)e_{\ell'}(\tilde{\boldsymbol{\beta}},t)\right| \\
&\leq \sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega}\left|\frac{S^{(2)}_{\ell,\ell'}(\tilde{\boldsymbol{\beta}},t)/n}{S^{(0)}(\tilde{\boldsymbol{\beta}},t)/n}-\frac{s^{(2)}_{\ell,\ell'}(\tilde{\boldsymbol{\beta}},t)}{s^{(0)}(\tilde{\boldsymbol{\beta}},t)}\right|+\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega}\left|\epsilon_{\ell}(\tilde{\boldsymbol{\beta}},t)\right|\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega}\left|\epsilon_{\ell'}(\tilde{\boldsymbol{\beta}},t)-e_{\ell'}(\tilde{\boldsymbol{\beta}},t)\right| \\
&+\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega}\left|e_{\ell'}(\tilde{\boldsymbol{\beta}},t)\right|\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega}\left|\epsilon_{\ell}(\tilde{\boldsymbol{\beta}},t)-e_{\ell}(\tilde{\boldsymbol{\beta}},t)\right|=o_{\mathrm{P}}(1)+O_{\mathrm{P}}(1)o_{\mathrm{P}}(1)+\mathcal{O}(1)o_{\mathrm{P}}(1)=o_{\mathrm{P}}(1),
\end{aligned}
$$

which leads to $\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega}\|\mathcal{V}(\tilde{\boldsymbol{\beta}},t)-v(\tilde{\boldsymbol{\beta}},t)\|_\infty=o_{\mathrm{P}}(1)$.

Lastly, we show convergence result for log. Considering the fact that $s^{(0)}(\tilde{\boldsymbol{\beta}},t)>0$ and $S^{(0)}(\tilde{\boldsymbol{\beta}},t)/n>0$ for all $\tilde{\boldsymbol{\beta}}\in\mathcal{B}$ and all $t\in\Omega$ ($n(\zeta)\geq 1$), $\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega}\{S^{(0)}(\tilde{\boldsymbol{\beta}},t)/n\}=\mathcal{O}_{\mathrm{P}}(1)$, $\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega}\{s^{(0)}(\tilde{\boldsymbol{\beta}},t)\}=\mathcal{O}(1)$, $\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega}\{n/S^{(0)}(\tilde{\boldsymbol{\beta}},t)\}=\mathcal{O}_{\mathrm{P}}(1)$, $\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega}\{1/s^{(0)}(\tilde{\boldsymbol{\beta}},t)\}=\mathcal{O}(1)$, and

$$\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega}\left|S^{(0)}(\tilde{\boldsymbol{\beta}},t)/n-s^{(0)}(\tilde{\boldsymbol{\beta}},t)\right|=o_{\mathrm{P}}(1),\text{ by Lemma A.14, we have}$$

$$\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega}\frac{S^{(0)}(\tilde{\boldsymbol{\beta}},t)/n}{s^{(0)}(\tilde{\boldsymbol{\beta}},t)}=1+o_{\mathrm{P}}(1)\text{ and }\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega}\frac{s^{(0)}(\tilde{\boldsymbol{\beta}},t)}{S^{(0)}(\tilde{\boldsymbol{\beta}},t)/n}=1+o_{\mathrm{P}}(1).$$

Note that

$$\inf_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega}\frac{S^{(0)}(\tilde{\boldsymbol{\beta}},t)/n}{s^{(0)}(\tilde{\boldsymbol{\beta}},t)}=\inf_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega}\frac{1}{s^{(0)}(\tilde{\boldsymbol{\beta}},t)/(S^{(0)}(\tilde{\boldsymbol{\beta}},t)/n)}=\frac{1}{\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega}\left\{s^{(0)}(\tilde{\boldsymbol{\beta}},t)/(S^{(0)}(\tilde{\boldsymbol{\beta}},t)/n)\right\}}$$

$$=\frac{1}{1+o_{\mathrm{P}}(1)}=1+o_{\mathrm{P}}(1).$$

$$\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega}\left|\log\left\{\frac{S^{(0)}(\tilde{\boldsymbol{\beta}},t)/n}{s^{(0)}(\tilde{\boldsymbol{\beta}},t)}\right\}\right|=\max\left\{\left|\log\left\{\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega}\frac{S^{(0)}(\tilde{\boldsymbol{\beta}},t)/n}{s^{(0)}(\tilde{\boldsymbol{\beta}},t)}\right\}\right|,\left|\log\left\{\inf_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega}\frac{S^{(0)}(\tilde{\boldsymbol{\beta}},t)/n}{s^{(0)}(\tilde{\boldsymbol{\beta}},t)}\right\}\right|\right\}$$

$$=o_{\mathrm{P}}(1)\text{ because}$$

$$\log\left\{\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega}\frac{S^{(0)}(\tilde{\boldsymbol{\beta}},t)/n}{s^{(0)}(\tilde{\boldsymbol{\beta}},t)}\right\}=\log(1+o_{\mathrm{P}}(1))=o_{\mathrm{P}}(1),\log\left\{\inf_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega}\frac{S^{(0)}(\tilde{\boldsymbol{\beta}},t)/n}{s^{(0)}(\tilde{\boldsymbol{\beta}},t)}\right\}=\log(1+o_{\mathrm{P}}(1))=o_{\mathrm{P}}(1).$$

■

Now we derive some asymptotic orders for covariance estimation.

Lemma A.16. *Under Assumptions A.3.1 – A.3.6, for* $t \in [0, \zeta]$, $\langle n^{-1/2} U(\boldsymbol{\beta}) \rangle(t) = \mathcal{O}_P(1)$ *and* $\sup_{\tilde{\boldsymbol{\beta}} \in \mathcal{B}} \| I_b(\tilde{\boldsymbol{\beta}}, t)/n \|_\infty = \mathcal{O}_P(1)$.

Proof. In this proof, we restrict $t \in [0, \zeta]$. For all $\ell, \ell' \in \{1, \ldots, d\}$,

$$\begin{aligned}
\left| \left\{ \langle n^{-1/2} U(\boldsymbol{\beta}) \rangle (t) \right\}_{\ell, \ell'} \right| &\leq \int_0^t \sum_{i=1}^n \frac{1}{n} \Big| (x_{i\ell} - \epsilon_\ell(\boldsymbol{\beta}, s))(x_{i\ell'} - \epsilon_{\ell'}(\boldsymbol{\beta}, s)) \Big| d\langle M_i \rangle(s) \\
&= \sum_{s \in \Omega_t} \sum_{i=1}^n \frac{1}{n} \Big| (x_{i\ell} - \epsilon_\ell(\boldsymbol{\beta}, s))(x_{i\ell'} - \epsilon_{\ell'}(\boldsymbol{\beta}, s)) \Big| \Delta\langle M_i \rangle(s) \\
&\leq \sum_{t \in \Omega} \sum_{i=1}^n \frac{1}{n} \Big| (x_{i\ell} - \epsilon_\ell(\boldsymbol{\beta}, t))(x_{i\ell'} - \epsilon_{\ell'}(\boldsymbol{\beta}, t)) \Big| \Delta\langle M_i \rangle(t) \\
&\leq \frac{n\zeta}{\tau} \sup_{t \in \Omega, i} \frac{1}{n} \Big| (x_{i\ell} - \epsilon_\ell(\boldsymbol{\beta}, t))(x_{i\ell'} - \epsilon_{\ell'}(\boldsymbol{\beta}, t)) \Big| \Delta\langle M_i \rangle(t) \\
&= \frac{n\zeta}{\tau} \mathcal{O}_{\mathrm{P}}\left(\frac{\tau}{n} \right) = \mathcal{O}_{\mathrm{P}}(1), \text{ where}
\end{aligned}$$

$$\begin{aligned}
&\sup_{t \in \Omega, i} \frac{1}{n} \Big| (x_{i\ell} - \epsilon_\ell(\boldsymbol{\beta}, t))(x_{i\ell'} - \epsilon_{\ell'}(\boldsymbol{\beta}, t)) \Big| \Delta\langle M_i \rangle(t) \\
&\quad \leq \frac{1}{n} \sup_{t \in \Omega, i} |(x_{i\ell} - \epsilon_\ell(\boldsymbol{\beta}, t))| \sup_{t \in \Omega, i} |(x_{i\ell'} - \epsilon_{\ell'}(\boldsymbol{\beta}, t))| \sup_{t \in \Omega, i} \Delta\langle M_i \rangle(t) \\
&\quad = \frac{1}{n} \mathcal{O}_{\mathrm{P}}(1) \mathcal{O}_{\mathrm{P}}(1) \mathcal{O}_{\mathrm{P}}(\tau) = \mathcal{O}_{\mathrm{P}}\left(\frac{\tau}{n} \right), \text{ implying that } \langle n^{-1/2} U(\boldsymbol{\beta}) \rangle(t) = \mathcal{O}_{\mathrm{P}}(1).
\end{aligned}$$

For all $\ell, \ell' \in \{1, \ldots, d\}$,

$$\begin{aligned}
\sup_{\tilde{\boldsymbol{\beta}} \in \mathcal{B}} \left| \left\{ \frac{1}{n} I_{\mathrm{b}}(\tilde{\boldsymbol{\beta}}, t) \right\}_{\ell, \ell'} \right| &\leq \sup_{\tilde{\boldsymbol{\beta}} \in \mathcal{B}} \int_0^t \frac{1}{n} \left| \mathcal{V}_{\ell, \ell'}(\tilde{\boldsymbol{\beta}}, s) \right| dN(s) = \sup_{\tilde{\boldsymbol{\beta}} \in \mathcal{B}} \sum_{s \in \Omega_t} \frac{1}{n} \left| \mathcal{V}_{\ell, \ell'}(\tilde{\boldsymbol{\beta}}, s) \right| \Delta N(s) \\
&\leq \sup_{\tilde{\boldsymbol{\beta}} \in \mathcal{B}} \sum_{t \in \Omega} \frac{1}{n} \left| \mathcal{V}_{\ell, \ell'}(\tilde{\boldsymbol{\beta}}, t) \right| \Delta N(t) \leq \sup_{\tilde{\boldsymbol{\beta}} \in \mathcal{B}} \frac{\zeta}{\tau} \sup_{t \in \Omega} \frac{1}{n} \left| \mathcal{V}_{\ell, \ell'}(\tilde{\boldsymbol{\beta}}, t) \right| \Delta N(t) \\
&= \frac{\zeta}{\tau} \sup_{\tilde{\boldsymbol{\beta}} \in \mathcal{B}, t \in \Omega} \frac{1}{n} \left| \mathcal{V}_{\ell, \ell'}(\tilde{\boldsymbol{\beta}}, t) \right| \Delta N(t) = \frac{\zeta}{\tau} \mathcal{O}_{\mathrm{P}}(\tau) = \mathcal{O}_{\mathrm{P}}(1), \text{ where}
\end{aligned}$$

$$\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega}\frac{1}{n}\left|\mathcal{V}_{\ell,\ell'}(\tilde{\boldsymbol{\beta}},t)\right|\Delta N(t)\leq\frac{1}{n}\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega}\left|\mathcal{V}_{\ell,\ell'}(\tilde{\boldsymbol{\beta}},t)\right|\sup_{t\in\Omega}\Delta N(t)=\frac{1}{n}\mathcal{O}_{\mathrm{P}}(1)\mathcal{O}_{\mathrm{P}}(n\tau)=\mathcal{O}_{\mathrm{P}}(\tau),$$

implying that $\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B}}\|I_{\mathrm{b}}(\tilde{\boldsymbol{\beta}},t)/n\|_\infty=\mathcal{O}_{\mathrm{P}}(1)$. ■

Note that under Assumptions A.3.1 – A.3.6, for all $t\in[0,\zeta]$, as $\|I_{\mathrm{b}}(\widehat{\boldsymbol{\beta}}_{\mathrm{b}},t)/n\|_\infty\leq\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B}}\|I_{\mathrm{b}}(\tilde{\boldsymbol{\beta}},t)/n\|_\infty$ $=\mathcal{O}_{\mathrm{P}}(1)$ and $\|I_{\mathrm{b}}(\widehat{\boldsymbol{\beta}}_{\mathrm{pb}},t)/n\|_\infty\leq\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B}}\|I_{\mathrm{b}}(\tilde{\boldsymbol{\beta}},t)/n\|_\infty=\mathcal{O}_{\mathrm{P}}(1)$, covariance estimators satisfy $I_{\mathrm{b}}(\widehat{\boldsymbol{\beta}}_{\mathrm{b}},t)/n=\mathcal{O}_{\mathrm{P}}(1)$ and $I_{\mathrm{b}}(\widehat{\boldsymbol{\beta}}_{\mathrm{pb}},t)/n=\mathcal{O}_{\mathrm{P}}(1)$.

Now we derive asymptotic order for $X_{\mathrm{pb}}(\tilde{\boldsymbol{\beta}},\widehat{\boldsymbol{\Lambda}},t)-X(\tilde{\boldsymbol{\beta}},t)$, $X(\tilde{\boldsymbol{\beta}},t)$, and $X_{\mathrm{pb}}(\tilde{\boldsymbol{\beta}},\widehat{\boldsymbol{\Lambda}},t)$.

Lemma A.17. *Under Assumptions A.3.1 – A.3.5 and A.3.8 – A.3.9, for* $t\in[0,\zeta]$,

$$\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B}}|X_{pb}(\tilde{\boldsymbol{\beta}},\widehat{\boldsymbol{\Lambda}},t)-X(\tilde{\boldsymbol{\beta}},t)|=\mathcal{O}_P(1/n).$$

Proof. For $t\in[0,\zeta]$,

$$\begin{aligned}
&\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B}}|X_{\mathrm{pb}}(\tilde{\boldsymbol{\beta}},\widehat{\boldsymbol{\Lambda}},t)-X(\tilde{\boldsymbol{\beta}},t)|\\
&=\frac{1}{n}\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B}}\left|\left\{\log\left(L(\tilde{\boldsymbol{\beta}},\widehat{\boldsymbol{\Lambda}},t)\right)-\log\left(L(\boldsymbol{\beta},\widehat{\boldsymbol{\Lambda}},t)\right)\right\}-\left\{\log\left(L^{\mathrm{b}}(\tilde{\boldsymbol{\beta}},t)\right)-\log\left(L^{\mathrm{b}}(\boldsymbol{\beta},t)\right)\right\}\right|\\
&\leq\frac{1}{n}\left|\log\left(L(\boldsymbol{\beta},\widehat{\boldsymbol{\Lambda}},t)\right)-\log\left(L^{\mathrm{b}}(\boldsymbol{\beta},t)\right)\right|+\frac{1}{n}\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B}}\left|\log\left(L(\tilde{\boldsymbol{\beta}},\widehat{\boldsymbol{\Lambda}},t)\right)-\log\left(L^{\mathrm{b}}(\tilde{\boldsymbol{\beta}},t)\right)\right|\\
&\leq2\frac{1}{n}\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B}}\left|\log\left(L(\tilde{\boldsymbol{\beta}},\widehat{\boldsymbol{\Lambda}},t)\right)-\log\left(L^{\mathrm{b}}(\tilde{\boldsymbol{\beta}},t)\right)\right|=\mathcal{O}_{\mathrm{P}}(1/n).
\end{aligned}$$

■

Lemma A.18. *Under Assumptions A.3.1 – A.3.5, for* $t\in[0,\zeta]$, $\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B}}|X(\tilde{\boldsymbol{\beta}},t)|=\mathcal{O}_P(1)$.

Proof. For $t\in[0,\zeta]$,

$$\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B}}|X(\tilde{\boldsymbol{\beta}},t)|\leq\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B}}\int_0^t\frac{1}{n}\sum_{i=1}^n\left|\boldsymbol{x}_i^\top(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta})-\log\left(\frac{S^{(0)}(\tilde{\boldsymbol{\beta}},s)/n}{S^{(0)}(\boldsymbol{\beta},s)/n}\right)\right|dN_i(s)$$

$$
\begin{aligned}
&= \sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B}} \sum_{s\in\Omega_t} \frac{1}{n} \sum_{i=1}^{n} \left| \boldsymbol{x}_i^\top(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta}) - \log\left(\frac{S^{(0)}(\tilde{\boldsymbol{\beta}},s)/n}{S^{(0)}(\boldsymbol{\beta},s)/n}\right)\right| \Delta N_i(s) \\
&\le \sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B}} \sum_{t\in\Omega} \frac{1}{n} \sum_{i=1}^{n} \left| \boldsymbol{x}_i^\top(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta}) - \log\left(\frac{S^{(0)}(\tilde{\boldsymbol{\beta}},t)/n}{S^{(0)}(\boldsymbol{\beta},t)/n}\right)\right| \Delta N_i(t) \\
&\le \sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B}} \sum_{t\in\Omega} \sup_i \left\{ \left| \boldsymbol{x}_i^\top(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta}) - \log\left(\frac{S^{(0)}(\tilde{\boldsymbol{\beta}},t)/n}{S^{(0)}(\boldsymbol{\beta},t)/n}\right)\right| \right\} \frac{\Delta N(t)}{n} \\
&\le \frac{\zeta}{\tau} \sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega,i} \left\{ \left| \boldsymbol{x}_i^\top(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta}) - \log\left(\frac{S^{(0)}(\tilde{\boldsymbol{\beta}},t)/n}{S^{(0)}(\boldsymbol{\beta},t)/n}\right)\right| \right\} \frac{1}{n} \sup_{t\in\Omega} \Delta N(t) \\
&= \frac{\zeta}{\tau} \mathcal{O}_{\mathrm{P}}(1) \frac{1}{n} \mathcal{O}_{\mathrm{P}}(n\tau) = \mathcal{O}_{\mathrm{P}}(1),
\end{aligned}
$$

implying that $\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B}} |X(\tilde{\boldsymbol{\beta}},t)| = \mathcal{O}_{\mathrm{P}}(1)$. Note that we used the fact that

$$
\begin{aligned}
&\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega,i} \left| \boldsymbol{x}_i^\top(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta}) - \log\left(\frac{S^{(0)}(\tilde{\boldsymbol{\beta}},t)/n}{S^{(0)}(\boldsymbol{\beta},t)/n}\right)\right| \\
&\qquad \le \sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},i} |\boldsymbol{x}_i^\top(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta})| + \sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega} |\log\{S^{(0)}(\tilde{\boldsymbol{\beta}},t)/n\}| + \sup_{t\in\Omega} |\log\{S^{(0)}(\boldsymbol{\beta},t)/n\}| = \mathcal{O}_{\mathrm{P}}(1),
\end{aligned}
$$

where $\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},i} |\boldsymbol{x}_i^\top(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta})| \le \sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},i} \|\boldsymbol{x}_i\|_2 \|\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta}\|_2 \le \sup_i \|\boldsymbol{x}_i\|_2 \sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B}} \|\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta}\|_2 \le \sqrt{d}\rho(\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B}} \|\tilde{\boldsymbol{\beta}}\|_2 + \|\boldsymbol{\beta}\|_2) \le 2\sqrt{d}\rho C_{\mathcal{B}} = \mathcal{O}(1)$. ■

Lemma A.19. *Under Assumptions A.3.1 – A.3.5 and A.3.8 – A.3.9, for $t \in [0,\zeta]$,*

$$\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B}} |X_{pb}(\tilde{\boldsymbol{\beta}},\widehat{\boldsymbol{\Lambda}},t)| = \mathcal{O}_P(1).$$

Proof. For $t \in [0,\zeta]$,

$$\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B}} |X_{\mathrm{pb}}(\tilde{\boldsymbol{\beta}},\widehat{\boldsymbol{\Lambda}},t)| \le \sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B}} |X_{\mathrm{pb}}(\tilde{\boldsymbol{\beta}},\widehat{\boldsymbol{\Lambda}},t) - X(\tilde{\boldsymbol{\beta}},t)| + \sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B}} |X(\tilde{\boldsymbol{\beta}},t)| = \mathcal{O}_{\mathrm{P}}(1/n) + \mathcal{O}_{\mathrm{P}}(1) = \mathcal{O}_{\mathrm{P}}(1).$$

■

Considering Lemmas A.17 – A.19, under Assumptions A.3.1 – A.3.5 and A.3.8 – A.3.9, for any $t \in [0, \zeta]$, we have the following results. As $|X(\widehat{\boldsymbol{\beta}}_{\mathrm{b}}, t)| \leq \sup_{\tilde{\boldsymbol{\beta}} \in \mathcal{B}} |X(\tilde{\boldsymbol{\beta}}, t)| = \mathcal{O}_{\mathrm{P}}(1)$ and $|X(\widehat{\boldsymbol{\beta}}_{\mathrm{pb}}, t)| \leq \sup_{\tilde{\boldsymbol{\beta}} \in \mathcal{B}} |X(\tilde{\boldsymbol{\beta}}, t)| = \mathcal{O}_{\mathrm{P}}(1)$, we have $X(\widehat{\boldsymbol{\beta}}_{\mathrm{b}}, t) = \mathcal{O}_{\mathrm{P}}(1)$ and $X(\widehat{\boldsymbol{\beta}}_{\mathrm{pb}}, t) = \mathcal{O}_{\mathrm{P}}(1)$. As $|X_{\mathrm{pb}}(\widehat{\boldsymbol{\beta}}_{\mathrm{b}}, \widehat{\boldsymbol{\Lambda}}, t)| \leq \sup_{\tilde{\boldsymbol{\beta}} \in \mathcal{B}} |X_{\mathrm{pb}}(\tilde{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Lambda}}, t)| = \mathcal{O}_{\mathrm{P}}(1)$ and $|X_{\mathrm{pb}}(\widehat{\boldsymbol{\beta}}_{\mathrm{pb}}, \widehat{\boldsymbol{\Lambda}}, t)| \leq \sup_{\tilde{\boldsymbol{\beta}} \in \mathcal{B}} |X_{\mathrm{pb}}(\tilde{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Lambda}}, t)| = \mathcal{O}_{\mathrm{P}}(1)$, we have $X_{\mathrm{pb}}(\widehat{\boldsymbol{\beta}}_{\mathrm{b}}, \widehat{\boldsymbol{\Lambda}}, t) = \mathcal{O}_{\mathrm{P}}(1)$ and $X_{\mathrm{pb}}(\widehat{\boldsymbol{\beta}}_{\mathrm{pb}}, \widehat{\boldsymbol{\Lambda}}, t) = \mathcal{O}_{\mathrm{P}}(1)$. Similarly, as $|X_{\mathrm{pb}}(\widehat{\boldsymbol{\beta}}_{\mathrm{b}}, \widehat{\boldsymbol{\Lambda}}, t) - X(\widehat{\boldsymbol{\beta}}_{\mathrm{b}}, t)| \leq \sup_{\tilde{\boldsymbol{\beta}} \in \mathcal{B}} |X_{\mathrm{pb}}(\tilde{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Lambda}}, t) - X(\tilde{\boldsymbol{\beta}}, t)| = \mathcal{O}_{\mathrm{P}}(1/n)$ and $|X_{\mathrm{pb}}(\widehat{\boldsymbol{\beta}}_{\mathrm{pb}}, \widehat{\boldsymbol{\Lambda}}, t) - X(\widehat{\boldsymbol{\beta}}_{\mathrm{pb}}, t)| \leq \sup_{\tilde{\boldsymbol{\beta}} \in \mathcal{B}} |X_{\mathrm{pb}}(\tilde{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Lambda}}, t) - X(\tilde{\boldsymbol{\beta}}, t)| = \mathcal{O}_{\mathrm{P}}(1/n)$, we have $X_{\mathrm{pb}}(\widehat{\boldsymbol{\beta}}_{\mathrm{pb}}, \widehat{\boldsymbol{\Lambda}}, t) - X(\widehat{\boldsymbol{\beta}}_{\mathrm{pb}}, t) = \mathcal{O}_{\mathrm{P}}(1/n)$ and $X_{\mathrm{pb}}(\widehat{\boldsymbol{\beta}}_{\mathrm{b}}, \widehat{\boldsymbol{\Lambda}}, t) - X(\widehat{\boldsymbol{\beta}}_{\mathrm{b}}, t) = \mathcal{O}_{\mathrm{P}}(1/n)$.

Proof of the Lemma A.12

We start with proof for the first line. For $t \in \Omega$ and $\tilde{\boldsymbol{\beta}} \in \mathcal{B}$, let $\tilde{B}_t(\tilde{\boldsymbol{\beta}}) = \mu(\tilde{\boldsymbol{\beta}}, t)^{d(t)} \exp(-\mu(\tilde{\boldsymbol{\beta}}, t))/d(t)!$, where $B_t(\tilde{\boldsymbol{\beta}})$ has the form of probability mass function of PB distribution and $\tilde{B}_t(\tilde{\boldsymbol{\beta}})$ is Poisson approximation for $B_t(\tilde{\boldsymbol{\beta}})$. By Remark 2.2, for all $t \in \Omega$ and $\tilde{\boldsymbol{\beta}} \in \mathcal{B}$, $\left|B_t(\tilde{\boldsymbol{\beta}}) - \tilde{B}_t(\tilde{\boldsymbol{\beta}})\right| \leq 2\sum_{i \in \mathcal{R}(t)} p_i(\tilde{\boldsymbol{\beta}}, t)^2$, which leads to

$$
\begin{aligned}
\sup_{\tilde{\boldsymbol{\beta}} \in \mathcal{B}, t \in \Omega} \left|B_t(\tilde{\boldsymbol{\beta}}) - \tilde{B}_t(\tilde{\boldsymbol{\beta}})\right| &\leq \sup_{\tilde{\boldsymbol{\beta}} \in \mathcal{B}, t \in \Omega} \left\{ 2 \sum_{i \in \mathcal{R}(t)} p_i(\tilde{\boldsymbol{\beta}}, t)^2 \right\} \leq \sup_{\tilde{\boldsymbol{\beta}} \in \mathcal{B}, t \in \Omega} \left\{ 2 \sum_{i=1}^{n} p_i(\tilde{\boldsymbol{\beta}}, t)^2 \right\} \\
&\leq 2n \sup_{\tilde{\boldsymbol{\beta}} \in \mathcal{B}, t \in \Omega, i} p_i(\tilde{\boldsymbol{\beta}}, t)^2 \leq 2n \left\{ \sup_{\tilde{\boldsymbol{\beta}} \in \mathcal{B}, t \in \Omega, i} p_i(\tilde{\boldsymbol{\beta}}, t) \right\}^2 = 2n\mathcal{O}_{\mathrm{P}}\left(\frac{1}{n}\right)^2 = \mathcal{O}_{\mathrm{P}}\left(\frac{1}{n}\right).
\end{aligned}
$$

Note that $B_t(\tilde{\boldsymbol{\beta}}), \tilde{B}_t(\tilde{\boldsymbol{\beta}}) \in [0, 1]$ for $t \in \Omega$ and $\tilde{\boldsymbol{\beta}} \in \mathcal{B}$ since they are in the form of probability mass function, leading to $\sup_{\tilde{\boldsymbol{\beta}} \in \mathcal{B}, t \in \Omega} B_t(\tilde{\boldsymbol{\beta}}) = \mathcal{O}_{\mathrm{P}}(1)$ and $\sup_{\tilde{\boldsymbol{\beta}} \in \mathcal{B}, t \in \Omega} \tilde{B}_t(\tilde{\boldsymbol{\beta}}) = \mathcal{O}_{\mathrm{P}}(1)$.

Note that by Lemma A.20,

$$
\sup_{\tilde{\boldsymbol{\beta}} \in \mathcal{B}, t \in \Omega, d(t) \neq 0} \left| \frac{d(t)}{\mu(\tilde{\boldsymbol{\beta}}, t)} \right| = \mathcal{O}_{\mathrm{P}}(1) \text{ and } \sup_{\tilde{\boldsymbol{\beta}} \in \mathcal{B}, t \in \Omega, d(t) \neq 0} \left| \frac{\mu(\tilde{\boldsymbol{\beta}}, t)}{d(t)} \right| = \mathcal{O}_{\mathrm{P}}(1),
$$

which means that the number of event and mean in the poisson probability mass function have the same order. Also, for $t \in \Omega$ and $\tilde{\boldsymbol{\beta}} \in \mathcal{B}$, $d(t) = 0$ implies $d\widehat{\Lambda}(t) = 0$, $p_i(\tilde{\boldsymbol{\beta}}, t) = 0$ for all i, $\mu(\tilde{\boldsymbol{\beta}}, t) = 0$, and $B_t(\tilde{\boldsymbol{\beta}}) = \tilde{B}_t(\tilde{\boldsymbol{\beta}}) = 1$, whereas $d(t) \to \infty$ implies $B_t(\tilde{\boldsymbol{\beta}})$ and $\tilde{B}_t(\tilde{\boldsymbol{\beta}})$ going to 0. Considering aforementioned and the fact that $\Pr(\sup_{t\in\Omega} d(t) \leq C) \to 1$ for some $C > 0$, i.e., $\sup_{t\in\Omega} d(t) = \mathcal{O}_{\mathrm{P}}(1)$, we have $\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega}\{1/B_t(\tilde{\boldsymbol{\beta}})\} = \mathcal{O}_{\mathrm{P}}(1)$ and $\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega}\{1/\tilde{B}_t(\tilde{\boldsymbol{\beta}})\} = \mathcal{O}_{\mathrm{P}}(1)$. So, by Lemma A.14,

$$\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega}\{B_t(\tilde{\boldsymbol{\beta}})/\tilde{B}_t(\tilde{\boldsymbol{\beta}})\} = 1 + \mathcal{O}_{\mathrm{P}}(1/n) \text{ and } \sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega}\{\tilde{B}_t(\tilde{\boldsymbol{\beta}})/B_t(\tilde{\boldsymbol{\beta}})\} = 1 + \mathcal{O}_{\mathrm{P}}(1/n).$$

Note that

$$\inf_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega}\frac{B_t(\tilde{\boldsymbol{\beta}})}{\tilde{B}_t(\tilde{\boldsymbol{\beta}})} = \inf_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega}\frac{1}{\tilde{B}_t(\tilde{\boldsymbol{\beta}})/B_t(\tilde{\boldsymbol{\beta}})} = \frac{1}{\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega}\left\{\tilde{B}_t(\tilde{\boldsymbol{\beta}})/B_t(\tilde{\boldsymbol{\beta}})\right\}} = \frac{1}{1+\mathcal{O}_{\mathrm{P}}(1/n)} = 1 + \mathcal{O}_{\mathrm{P}}(1/n).$$

$$\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega}\left|\log\left\{\frac{B_t(\tilde{\boldsymbol{\beta}})}{\tilde{B}_t(\tilde{\boldsymbol{\beta}})}\right\}\right| = \max\left\{\left|\log\left\{\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega}\frac{B_t(\tilde{\boldsymbol{\beta}})}{\tilde{B}_t(\tilde{\boldsymbol{\beta}})}\right\}\right|, \left|\log\left\{\inf_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega}\frac{B_t(\tilde{\boldsymbol{\beta}})}{\tilde{B}_t(\tilde{\boldsymbol{\beta}})}\right\}\right|\right\} = \mathcal{O}_{\mathrm{P}}(1/n) \text{ because}$$

$$\log\left\{\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega}\frac{B_t(\tilde{\boldsymbol{\beta}})}{\tilde{B}_t(\tilde{\boldsymbol{\beta}})}\right\} = \log(1+\mathcal{O}_{\mathrm{P}}(1/n)) = \mathcal{O}_{\mathrm{P}}(1/n),$$
$$\log\left\{\inf_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega}\frac{B_t(\tilde{\boldsymbol{\beta}})}{\tilde{B}_t(\tilde{\boldsymbol{\beta}})}\right\} = \log(1+\mathcal{O}_{\mathrm{P}}(1/n)) = \mathcal{O}_{\mathrm{P}}(1/n).$$

Now we prove second line. For $t \in \Omega$ and $\tilde{\boldsymbol{\beta}} \in \mathcal{B}$, let $E_t(\tilde{\boldsymbol{\beta}}) = \prod_{i\in\mathcal{R}(t)\setminus\mathcal{D}(t)}\left(1 - p_i(\tilde{\boldsymbol{\beta}}, t)\right)$ and $\tilde{E}_t(\tilde{\boldsymbol{\beta}}) = \exp\left(-\mu(\tilde{\boldsymbol{\beta}}, t) + \tilde{\mu}(\tilde{\boldsymbol{\beta}}, t)\right)$, where $E_t(\tilde{\boldsymbol{\beta}})$ has the form of probability mass function of PB distribution and $\tilde{E}_t(\tilde{\boldsymbol{\beta}})$ is Poisson approximation for $E_t(\tilde{\boldsymbol{\beta}})$. By Remark 2.2, for all $t \in \Omega$

and $\tilde{\boldsymbol{\beta}} \in \mathcal{B}$,

$$\left| E_t(\tilde{\boldsymbol{\beta}}) - \tilde{E}_t(\tilde{\boldsymbol{\beta}}) \right| \leq 2 \sum_{i \in \mathcal{R}(t) \setminus \mathcal{D}(t)} p_i(\tilde{\boldsymbol{\beta}}, t)^2, \text{ which leads to}$$

$$\sup_{\tilde{\boldsymbol{\beta}} \in \mathcal{B}, t \in \Omega} \left| E_t(\tilde{\boldsymbol{\beta}}) - \tilde{E}_t(\tilde{\boldsymbol{\beta}}) \right| \leq \sup_{\tilde{\boldsymbol{\beta}} \in \mathcal{B}, t \in \Omega} \left\{ 2 \sum_{i \in \mathcal{R}(t) \setminus \mathcal{D}(t)} p_i(\tilde{\boldsymbol{\beta}}, t)^2 \right\} \leq \sup_{\tilde{\boldsymbol{\beta}} \in \mathcal{B}, t \in \Omega} \left\{ 2 \sum_{i=1}^{n} p_i(\tilde{\boldsymbol{\beta}}, t)^2 \right\}$$
$$\leq 2n \sup_{\tilde{\boldsymbol{\beta}} \in \mathcal{B}, t \in \Omega, i} p_i(\tilde{\boldsymbol{\beta}}, t)^2 \leq 2n \left\{ \sup_{\tilde{\boldsymbol{\beta}} \in \mathcal{B}, t \in \Omega, i} p_i(\tilde{\boldsymbol{\beta}}, t) \right\}^2 = 2n \mathcal{O}_{\mathrm{P}} \left(\frac{1}{n} \right)^2 = \mathcal{O}_{\mathrm{P}} \left(\frac{1}{n} \right).$$

Note that $E_t(\tilde{\boldsymbol{\beta}}), \tilde{E}_t(\tilde{\boldsymbol{\beta}}) \in [0, 1]$ for $t \in \Omega$ and $\tilde{\boldsymbol{\beta}} \in \mathcal{B}$ since they are in the form of probability mass function, leading to $\sup_{\tilde{\boldsymbol{\beta}} \in \mathcal{B}, t \in \Omega} E_t(\tilde{\boldsymbol{\beta}}) = \mathcal{O}_{\mathrm{P}}(1)$ and $\sup_{\tilde{\boldsymbol{\beta}} \in \mathcal{B}, t \in \Omega} \tilde{E}_t(\tilde{\boldsymbol{\beta}}) = \mathcal{O}_{\mathrm{P}}(1)$.

Note that by Lemma A.20,

$$\sup_{\tilde{\boldsymbol{\beta}} \in \mathcal{B}, t \in \Omega, d(t) \neq 0} \left| \frac{d(t)}{\mu(\tilde{\boldsymbol{\beta}}, t) - \tilde{\mu}(\tilde{\boldsymbol{\beta}}, t)} \right| = \mathcal{O}_{\mathrm{P}}(1) \text{ and } \sup_{\tilde{\boldsymbol{\beta}} \in \mathcal{B}, t \in \Omega, d(t) \neq 0} \left| \frac{\mu(\tilde{\boldsymbol{\beta}}, t) - \tilde{\mu}(\tilde{\boldsymbol{\beta}}, t)}{d(t)} \right| = \mathcal{O}_{\mathrm{P}}(1),$$

which means that the number of event and mean in the poisson probability mass function have the same order. Also, for $t \in \Omega$ and $\tilde{\boldsymbol{\beta}} \in \mathcal{B}$, $d(t) = 0$ implies $d\widehat{\Lambda}(t) = 0$, $p_i(\tilde{\boldsymbol{\beta}}, t) = 0$ for all i, $\mu(\tilde{\boldsymbol{\beta}}, t) - \tilde{\mu}(\tilde{\boldsymbol{\beta}}, t) = 0$, and $E_t(\tilde{\boldsymbol{\beta}}) = \tilde{E}_t(\tilde{\boldsymbol{\beta}}) = 1$, whereas $d(t) \to \infty$ implies $E_t(\tilde{\boldsymbol{\beta}})$ and $\tilde{E}_t(\tilde{\boldsymbol{\beta}})$ going to 0. Considering aforementioned and the fact that $\Pr(\sup_{t \in \Omega} d(t) \leq C) \to 1$ for some $C > 0$, i.e., $\sup_{t \in \Omega} d(t) = \mathcal{O}_{\mathrm{P}}(1)$, we have $\sup_{\tilde{\boldsymbol{\beta}} \in \mathcal{B}, t \in \Omega} \{1/E_t(\tilde{\boldsymbol{\beta}})\} = \mathcal{O}_{\mathrm{P}}(1)$ and $\sup_{\tilde{\boldsymbol{\beta}} \in \mathcal{B}, t \in \Omega} \{1/\tilde{E}_t(\tilde{\boldsymbol{\beta}})\} = \mathcal{O}_{\mathrm{P}}(1)$. So, by Lemma A.14,

$$\sup_{\tilde{\boldsymbol{\beta}} \in \mathcal{B}, t \in \Omega} \{E_t(\tilde{\boldsymbol{\beta}}) / \tilde{E}_t(\tilde{\boldsymbol{\beta}})\} = 1 + \mathcal{O}_{\mathrm{P}}(1/n) \text{ and } \sup_{\tilde{\boldsymbol{\beta}} \in \mathcal{B}, t \in \Omega} \{\tilde{E}_t(\tilde{\boldsymbol{\beta}}) / E_t(\tilde{\boldsymbol{\beta}})\} = 1 + \mathcal{O}_{\mathrm{P}}(1/n).$$

Note that

$$\inf_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega}\frac{E_t(\tilde{\boldsymbol{\beta}})}{\tilde{E}_t(\tilde{\boldsymbol{\beta}})}=\inf_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega}\frac{1}{\tilde{E}_t(\tilde{\boldsymbol{\beta}})/E_t(\tilde{\boldsymbol{\beta}})}=\frac{1}{\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega}\left\{\tilde{E}_t(\tilde{\boldsymbol{\beta}})/E_t(\tilde{\boldsymbol{\beta}})\right\}}=\frac{1}{1+\mathcal{O}_{\mathrm{P}}(1/n)}=1+\mathcal{O}_{\mathrm{P}}(1/n).$$

$$\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega}\left|\log\left\{\frac{E_t(\tilde{\boldsymbol{\beta}})}{\tilde{E}_t(\tilde{\boldsymbol{\beta}})}\right\}\right|=\max\left\{\left|\log\left\{\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega}\frac{E_t(\tilde{\boldsymbol{\beta}})}{\tilde{E}_t(\tilde{\boldsymbol{\beta}})}\right\}\right|,\left|\log\left\{\inf_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega}\frac{E_t(\tilde{\boldsymbol{\beta}})}{\tilde{E}_t(\tilde{\boldsymbol{\beta}})}\right\}\right|\right\}=\mathcal{O}_{\mathrm{P}}(1/n)\text{ because}$$

$$\begin{aligned}\log\left\{\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega}\frac{E_t(\tilde{\boldsymbol{\beta}})}{\tilde{E}_t(\tilde{\boldsymbol{\beta}})}\right\}&=\log(1+\mathcal{O}_{\mathrm{P}}(1/n))=\mathcal{O}_{\mathrm{P}}(1/n),\\ \log\left\{\inf_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega}\frac{E_t(\tilde{\boldsymbol{\beta}})}{\tilde{E}_t(\tilde{\boldsymbol{\beta}})}\right\}&=\log(1+\mathcal{O}_{\mathrm{P}}(1/n))=\mathcal{O}_{\mathrm{P}}(1/n).\end{aligned}$$

Now we prove third line. Note that when $d(t)=0$, we have

$$\frac{\prod_{i\in\mathcal{D}(t)}p_i(\tilde{\boldsymbol{\beta}},t)}{\prod_{i\in\mathcal{D}(t)}\left\{\exp(\boldsymbol{x}_i^\top\tilde{\boldsymbol{\beta}})d\widehat{\Lambda}(t)\right\}}=1\text{ for all }\tilde{\boldsymbol{\beta}}\in\mathcal{B}.$$

$$\begin{aligned}&\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega}\left|\log\left(\frac{\prod_{i\in\mathcal{D}(t)}p_i(\tilde{\boldsymbol{\beta}},t)}{\prod_{i\in\mathcal{D}(t)}\left\{\exp(\boldsymbol{x}_i^\top\tilde{\boldsymbol{\beta}})d\widehat{\Lambda}(t)\right\}}\right)\right|=\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega}\left|\log\left(\prod_{i\in\mathcal{D}(t)}(1+\tilde{a}_i(\tilde{\boldsymbol{\beta}},t))\right)\right|\\ &=\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega}\left|\sum_{i\in\mathcal{D}(t)}\log\left(1+\tilde{a}_i(\tilde{\boldsymbol{\beta}},t)\right)\right|\le\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega}\sum_{i\in\mathcal{D}(t)}\left|\log\left(1+\tilde{a}_i(\tilde{\boldsymbol{\beta}},t)\right)\right|\\ &\le\sup_{t\in\Omega}d(t)\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega,i}\left|\log\left(1+\tilde{a}_i(\tilde{\boldsymbol{\beta}},t)\right)\right|=\mathcal{O}_{\mathrm{P}}(1)\mathcal{O}_{\mathrm{P}}(1/n)=\mathcal{O}_{\mathrm{P}}(1/n).\end{aligned}$$

Note that for $t\in\Omega$ and $\tilde{\boldsymbol{\beta}}\in\mathcal{B}$, by the Taylor series expansion, for $\tilde{\alpha}_i(\tilde{\boldsymbol{\beta}},t)$ between $\tilde{a}_i(\tilde{\boldsymbol{\beta}},t)$

and 0,

$$\log(1+\tilde{a}_i(\tilde{\boldsymbol{\beta}},t)) = \tilde{a}_i(\tilde{\boldsymbol{\beta}},t) - \frac{1}{2(1+\tilde{\alpha}_i(\tilde{\boldsymbol{\beta}},t))^2}\tilde{a}_i(\tilde{\boldsymbol{\beta}},t)^2.$$

$$\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega,i}\left|\log\Big(1+\tilde{a}_i(\tilde{\boldsymbol{\beta}},t)\Big)\right| \leq \sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega,i}\left|\tilde{a}_i(\tilde{\boldsymbol{\beta}},t)\right| + \sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega,i}\left|\frac{1}{2(1+\tilde{\alpha}_i(\tilde{\boldsymbol{\beta}},t))^2}\right| \sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega,i}\tilde{a}_i(\tilde{\boldsymbol{\beta}},t)^2$$

$$= \mathcal{O}_{\mathrm{P}}(\tau) + \mathcal{O}_{\mathrm{P}}(1)\mathcal{O}_{\mathrm{P}}(\tau^2) = \mathcal{O}_{\mathrm{P}}(\tau) = \mathcal{O}_{\mathrm{P}}(1/n).$$

Finally, we prove the last line. Note that when $d(t)=0$, we have

$$\frac{\mu(\tilde{\boldsymbol{\beta}},t)^{d(t)}}{\left\{\sum_{i\in\mathcal{R}(t)}\exp(\boldsymbol{x}_i^\top\tilde{\boldsymbol{\beta}})d\widehat{\Lambda}(t)\right\}^{d(t)}} = 1 \text{ for all } \tilde{\boldsymbol{\beta}}\in\mathcal{B}.$$

$$\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega}\left|\log\left(\frac{\mu(\tilde{\boldsymbol{\beta}},t)^{d(t)}}{\left\{\sum_{i\in\mathcal{R}(t)}\exp(\boldsymbol{x}_i^\top\tilde{\boldsymbol{\beta}})d\widehat{\Lambda}(t)\right\}^{d(t)}}\right)\right| = \sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega}\left|d(t)\log\Big(1+b(\tilde{\boldsymbol{\beta}},t)\Big)\right|,$$

where

$$b(\tilde{\boldsymbol{\beta}},t) = -\frac{\sum_{i\in\mathcal{R}(t)}\exp(-\tilde{c}_i(\tilde{\boldsymbol{\beta}},t))\exp(2\boldsymbol{x}_i^\top\tilde{\boldsymbol{\beta}})d\widehat{\Lambda}(t)}{2\sum_{i\in\mathcal{R}(t)}\exp(\boldsymbol{x}_i^\top\tilde{\boldsymbol{\beta}})}.$$ We have

$$\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega}|b(\tilde{\boldsymbol{\beta}},t)| \leq \frac{1}{2}\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega}\frac{1}{\sum_{i\in\mathcal{R}(t)}\exp(\boldsymbol{x}_i^\top\tilde{\boldsymbol{\beta}})}\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega}\sum_{i\in\mathcal{R}(t)}\exp(-\tilde{c}_i(\tilde{\boldsymbol{\beta}},t))\exp(2\boldsymbol{x}_i^\top\tilde{\boldsymbol{\beta}})d\widehat{\Lambda}(t)$$

$$= \mathcal{O}_{\mathrm{P}}(1/n)\mathcal{O}_{\mathrm{P}}(1) = \mathcal{O}_{\mathrm{P}}(1/n), \text{ where}$$

$$\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega}\sum_{i\in\mathcal{R}(t)}\exp(-\tilde{c}_i(\tilde{\boldsymbol{\beta}},t))\exp(2\boldsymbol{x}_i^\top\tilde{\boldsymbol{\beta}})d\widehat{\Lambda}(t) \leq \sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega}\sum_{i\in\mathcal{R}(t)}\exp(2\boldsymbol{x}_i^\top\tilde{\boldsymbol{\beta}})d\widehat{\Lambda}(t)$$

$$\leq \sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega}\sum_{i=1}^{n}\exp(2\boldsymbol{x}_i^{\top}\tilde{\boldsymbol{\beta}})d\widehat{\Lambda}(t) \leq n \sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},i}\exp(2\boldsymbol{x}_i^{\top}\tilde{\boldsymbol{\beta}})\sup_{t\in\Omega}d\widehat{\Lambda}(t) = n\exp\left(2\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},i}\boldsymbol{x}_i^{\top}\tilde{\boldsymbol{\beta}}\right)\sup_{t\in\Omega}d\widehat{\Lambda}(t)$$

$$\leq n\exp\left(2\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},i}|\boldsymbol{x}_i^{\top}\tilde{\boldsymbol{\beta}}|\right)\sup_{t\in\Omega}d\widehat{\Lambda}(t) \leq n\exp(2\sqrt{d}\rho C_{\mathcal{B}})\sup_{t\in\Omega}d\widehat{\Lambda}(t) = n\mathcal{O}_{\mathrm{P}}(1/n) = \mathcal{O}_{\mathrm{P}}(1).$$

For $t\in\Omega$ and $\tilde{\boldsymbol{\beta}}\in\mathcal{B}$, by the Taylor series expansion, for $\tilde{b}(\tilde{\boldsymbol{\beta}},t)$ between $b(\tilde{\boldsymbol{\beta}},t)$ and 0,

$$\log(1+b(\tilde{\boldsymbol{\beta}},t)) = b(\tilde{\boldsymbol{\beta}},t) - \frac{1}{2(1+\tilde{b}(\tilde{\boldsymbol{\beta}},t))^2}b(\tilde{\boldsymbol{\beta}},t)^2.$$

We have

$$\begin{aligned}\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},\in\Omega}\left|\log\left(1+b(\tilde{\boldsymbol{\beta}},t)\right)\right| &\leq \sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega}\left|b(\tilde{\boldsymbol{\beta}},t)\right| + \sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega}\left|\frac{1}{2(1+\tilde{b}(\tilde{\boldsymbol{\beta}},t))^2}\right|\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega}b(\tilde{\boldsymbol{\beta}},t)^2\\ &= \mathcal{O}_{\mathrm{P}}(1/n) + \mathcal{O}_{\mathrm{P}}(1)\mathcal{O}_{\mathrm{P}}\left((1/n)^2\right) = \mathcal{O}_{\mathrm{P}}(1/n) \text{ and}\end{aligned}$$

$$\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega}\left|d(t)\log\left(1+b(\tilde{\boldsymbol{\beta}},t)\right)\right| \leq \sup_{t\in\Omega}d(t)\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega}\left|\log\left(1+b(\tilde{\boldsymbol{\beta}},t)\right)\right| = \mathcal{O}_{\mathrm{P}}(1)\mathcal{O}_{\mathrm{P}}(1/n) = \mathcal{O}_{\mathrm{P}}(1/n),$$

which competes the proof.

■

Lemma A.20. *When Assumptions A.3.1 – A.3.5 and A.3.8 – A.3.9 are satisfied,*

$$\begin{aligned}&\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega,d(t)\neq0}\left|\frac{d(t)}{\mu(\tilde{\boldsymbol{\beta}},t)}\right| = \mathcal{O}_P(1), \quad \sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega,d(t)\neq0}\left|\frac{\mu(\tilde{\boldsymbol{\beta}},t)}{d(t)}\right| = \mathcal{O}_P(1),\\ &\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega,d(t)\neq0}\left|\frac{d(t)}{\mu(\tilde{\boldsymbol{\beta}},t)-\tilde{\mu}(\tilde{\boldsymbol{\beta}},t)}\right| = \mathcal{O}_P(1), \text{ and } \sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega,d(t)\neq0}\left|\frac{\mu(\tilde{\boldsymbol{\beta}},t)-\tilde{\mu}(\tilde{\boldsymbol{\beta}},t)}{d(t)}\right| = \mathcal{O}_P(1).\end{aligned}$$

Proof. For $t \in \Omega$, $\tilde{\boldsymbol{\beta}} \in \mathcal{B}$, and $d(t) \neq 0$, one can represent $d(t)/\mu(\tilde{\boldsymbol{\beta}}, t)$ as

$$\frac{d(t)}{\mu(\tilde{\boldsymbol{\beta}}, t)} = \frac{d(t)}{\sum_{i=1}^{n} Y_i(t) \exp(\boldsymbol{x}_i^\top \tilde{\boldsymbol{\beta}}) \frac{d(t)}{n^*(t)} \left(1 - \frac{\exp(-\tilde{c}_i(\tilde{\boldsymbol{\beta}}, t))}{2} \exp(\boldsymbol{x}_i^\top \tilde{\boldsymbol{\beta}}) d\widehat{\Lambda}(t)\right)}$$
$$= \frac{n^*(t)}{\sum_{i=1}^{n} Y_i(t) \exp(\boldsymbol{x}_i^\top \tilde{\boldsymbol{\beta}}) \left(1 - \frac{\exp(-\tilde{c}_i(\tilde{\boldsymbol{\beta}}, t))}{2} \exp(\boldsymbol{x}_i^\top \tilde{\boldsymbol{\beta}}) d\widehat{\Lambda}(t)\right)}$$

where $1 - \exp(-\tilde{c}_i(\tilde{\boldsymbol{\beta}}, t)) \exp(\boldsymbol{x}_i^\top \tilde{\boldsymbol{\beta}}) d\widehat{\Lambda}(t)/2 \geq 0$ by the fact that $p_i(\tilde{\boldsymbol{\beta}}, t) \in [0, 1]$. Also, it is trivial that $1 - \exp(-\tilde{c}_i(\tilde{\boldsymbol{\beta}}, t)) \exp(\boldsymbol{x}_i^\top \tilde{\boldsymbol{\beta}}) d\widehat{\Lambda}(t)/2 \leq 1$, resulting in it to be reside in $[0, 1]$. We finish proving first result by the following:

$$\sup_{\tilde{\boldsymbol{\beta}} \in \mathcal{B}, t \in \Omega, d(t) \neq 0} \left| \frac{\mu(\tilde{\boldsymbol{\beta}}, t)}{d(t)} \right| = \sup_{\tilde{\boldsymbol{\beta}} \in \mathcal{B}, t \in \Omega, d(t) \neq 0} \left| \frac{\sum_{i=1}^{n} Y_i(t) \exp(\boldsymbol{x}_i^\top \tilde{\boldsymbol{\beta}}) \left(1 - \frac{\exp(-\tilde{c}_i(\tilde{\boldsymbol{\beta}}, t))}{2} \exp(\boldsymbol{x}_i^\top \tilde{\boldsymbol{\beta}}) d\widehat{\Lambda}(t)\right)}{n^*(t)} \right|$$

$$\leq \sup_{t \in \Omega, d(t) \neq 0} \frac{1}{n^*(t)} \sup_{\tilde{\boldsymbol{\beta}} \in \mathcal{B}, t \in \Omega, d(t) \neq 0} \sum_{i=1}^{n} Y_i(t) \exp(\boldsymbol{x}_i^\top \tilde{\boldsymbol{\beta}}) \left(1 - \frac{\exp(-\tilde{c}_i(\tilde{\boldsymbol{\beta}}, t))}{2} \exp(\boldsymbol{x}_i^\top \tilde{\boldsymbol{\beta}}) d\widehat{\Lambda}(t)\right)$$
$$\leq \sup_{t \in \Omega, d(t) \neq 0} \frac{1}{n^*(t)} \sup_{\tilde{\boldsymbol{\beta}} \in \mathcal{B}, t \in \Omega, d(t) \neq 0} \sum_{i=1}^{n} Y_i(t) \exp(\boldsymbol{x}_i^\top \tilde{\boldsymbol{\beta}}) = \mathcal{O}_{\mathrm{P}}(1/n) \mathcal{O}_{\mathrm{P}}(n) = \mathcal{O}_{\mathrm{P}}(1).$$

Now we prove second result. We have

$$\inf_{\tilde{\boldsymbol{\beta}} \in \mathcal{B}, t \in \Omega, d(t) \neq 0} \sum_{i=1}^{n} Y_i(t) \exp(\boldsymbol{x}_i^\top \tilde{\boldsymbol{\beta}}) \left(1 - \frac{\exp(-\tilde{c}_i(\tilde{\boldsymbol{\beta}}, t))}{2} \exp(\boldsymbol{x}_i^\top \tilde{\boldsymbol{\beta}}) d\widehat{\Lambda}(t)\right) \tag{A.2}$$
$$= \inf_{\tilde{\boldsymbol{\beta}} \in \mathcal{B}, t \in \Omega, d(t) \neq 0} \left[\sum_{i=1}^{n} Y_i(t) \exp(\boldsymbol{x}_i^\top \tilde{\boldsymbol{\beta}}) - \sum_{i=1}^{n} Y_i(t) \frac{\exp(-\tilde{c}_i(\tilde{\boldsymbol{\beta}}, t))}{2} \exp(2\boldsymbol{x}_i^\top \tilde{\boldsymbol{\beta}}) d\widehat{\Lambda}(t)\right]$$
$$\geq \inf_{\tilde{\boldsymbol{\beta}} \in \mathcal{B}, t \in \Omega, d(t) \neq 0} \left[\sum_{i=1}^{n} Y_i(t) \exp(\boldsymbol{x}_i^\top \tilde{\boldsymbol{\beta}})\right] + \inf_{\tilde{\boldsymbol{\beta}} \in \mathcal{B}, t \in \Omega, d(t) \neq 0} \left[- \sum_{i=1}^{n} Y_i(t) \frac{\exp(-\tilde{c}_i(\tilde{\boldsymbol{\beta}}, t))}{2} \exp(2\boldsymbol{x}_i^\top \tilde{\boldsymbol{\beta}}) d\widehat{\Lambda}(t)\right]$$

$$= \inf_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega,d(t)\neq 0}\left[\sum_{i=1}^{n} Y_i(t)\exp(\boldsymbol{x}_i^\top\tilde{\boldsymbol{\beta}})\right] - \sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega,d(t)\neq 0}\left[\sum_{i=1}^{n} Y_i(t)\frac{\exp(-\tilde{c}_i(\tilde{\boldsymbol{\beta}},t))}{2}\exp(2\boldsymbol{x}_i^\top\tilde{\boldsymbol{\beta}})d\widehat{\Lambda}(t)\right].$$

We have $\inf_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega,d(t)\neq 0}\sum_{i=1}^{n} Y_i(t)\exp(\boldsymbol{x}_i^\top\tilde{\boldsymbol{\beta}}) \geq \inf_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega}\sum_{i=1}^{n} Y_i(t)\exp(\boldsymbol{x}_i^\top\tilde{\boldsymbol{\beta}})$ and

$$\Pr\left(\inf_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega}\sum_{i=1}^{n} Y_i(t)\exp(\boldsymbol{x}_i^\top\tilde{\boldsymbol{\beta}}) \geq Cn\right) \to 1$$

for some $C > 0$. So, we have

$$\Pr\left(\inf_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega,d(t)\neq 0}\sum_{i=1}^{n} Y_i(t)\exp(\boldsymbol{x}_i^\top\tilde{\boldsymbol{\beta}}) \geq Cn\right) \to 1. \tag{A.3}$$

We have

$$\begin{aligned}\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega,d(t)\neq 0}\sum_{i=1}^{n} Y_i(t)\frac{\exp(-\tilde{c}_i(\tilde{\boldsymbol{\beta}},t))}{2}\exp(2\boldsymbol{x}_i^\top\tilde{\boldsymbol{\beta}})d\widehat{\Lambda}(t) &\leq \sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega,d(t)\neq 0}\sum_{i=1}^{n}\exp(2\boldsymbol{x}_i^\top\tilde{\boldsymbol{\beta}})d\widehat{\Lambda}(t)\\ &\leq n\exp\left(2\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},i}\boldsymbol{x}_i^\top\tilde{\boldsymbol{\beta}}\right)\sup_{t\in\Omega,d(t)\neq 0} d\widehat{\Lambda}(t) = n\mathcal{O}(1)\mathcal{O}_{\mathrm{P}}(1/n) = \mathcal{O}_{\mathrm{P}}(1),\end{aligned}$$

which means there exists $\mathcal{C} > 0$ such that

$$\begin{aligned}&\Pr\left(\inf_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega,d(t)\neq 0}\left[-\sum_{i=1}^{n} Y_i(t)\frac{\exp(-\tilde{c}_i(\tilde{\boldsymbol{\beta}},t))}{2}\exp(2\boldsymbol{x}_i^\top\tilde{\boldsymbol{\beta}})d\widehat{\Lambda}(t)\right] \geq -\mathcal{C}\right)\\ &= \Pr\left(\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega,d(t)\neq 0}\sum_{i=1}^{n} Y_i(t)\frac{\exp(-\tilde{c}_i(\tilde{\boldsymbol{\beta}},t))}{2}\exp(2\boldsymbol{x}_i^\top\tilde{\boldsymbol{\beta}})d\widehat{\Lambda}(t) \leq \mathcal{C}\right) \to 1.\end{aligned} \tag{A.4}$$

Considering (A.2), (A.3), and (A.4), there exists $\tilde{C} > 0$ such that

$$\Pr\left(\inf_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega,d(t)\neq 0}\sum_{i=1}^{n} Y_i(t)\exp(\boldsymbol{x}_i^\top\tilde{\boldsymbol{\beta}})\left(1 - \frac{\exp(-\tilde{c}_i(\tilde{\boldsymbol{\beta}},t))}{2}\exp(\boldsymbol{x}_i^\top\tilde{\boldsymbol{\beta}})d\widehat{\Lambda}(t)\right) \geq \tilde{C}n\right) \to 1.$$

We finish proving second result by the following:

$$\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega,d(t)\neq 0}\left|\frac{d(t)}{\mu(\tilde{\boldsymbol{\beta}},t)}\right| = \sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega,d(t)\neq 0}\left|\frac{n^*(t)}{\sum_{i=1}^n Y_i(t)\exp(\boldsymbol{x}_i^\top\tilde{\boldsymbol{\beta}})\left(1-\frac{\exp(-\tilde{c}_i(\tilde{\boldsymbol{\beta}},t))}{2}\exp(\boldsymbol{x}_i^\top\tilde{\boldsymbol{\beta}})d\widehat{\Lambda}(t)\right)}\right|$$

$$\leq \sup_{t\in\Omega,d(t)\neq 0} n^*(t) \sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega,d(t)\neq 0}\frac{1}{\sum_{i=1}^n Y_i(t)\exp(\boldsymbol{x}_i^\top\tilde{\boldsymbol{\beta}})\left(1-\frac{\exp(-\tilde{c}_i(\tilde{\boldsymbol{\beta}},t))}{2}\exp(\boldsymbol{x}_i^\top\tilde{\boldsymbol{\beta}})d\widehat{\Lambda}(t)\right)}$$

$$= \left\{\sup_{t\in\Omega,d(t)\neq 0} n^*(t)\right\}\frac{1}{\inf_{\tilde{\boldsymbol{\beta}}\in\mathcal{B},t\in\Omega,d(t)\neq 0}\sum_{i=1}^n Y_i(t)\exp(\boldsymbol{x}_i^\top\tilde{\boldsymbol{\beta}})\left(1-\frac{\exp(-\tilde{c}_i(\tilde{\boldsymbol{\beta}},t))}{2}\exp(\boldsymbol{x}_i^\top\tilde{\boldsymbol{\beta}})d\widehat{\Lambda}(t)\right)}$$

$$= \mathcal{O}_{\mathrm{P}}(n)\mathcal{O}_{\mathrm{P}}(1/n) = \mathcal{O}_{\mathrm{P}}(1).$$

We have $\sup_{t\in\Omega}\{n(t)-d(t)\}\leq n$, resulting in $\sup_{t\in\Omega}\{n(t)-d(t)\}=\mathcal{O}_{\mathrm{P}}(n)$. Also,

$$\inf_{t\in\Omega}\{n(t)-d(t)\}\geq\inf_{t\in\Omega}n(t)+\inf_{t\in\Omega}\{-d(t)\}=\inf_{t\in\Omega}n(t)-\sup_{t\in\Omega}\{d(t)\},$$

where $\Pr(\inf_{t\in\Omega}n(t)\geq\mathcal{C}_1 n)\to 1$ and $\Pr(\inf_{t\in\Omega}\{-d(t)\}\geq-\mathcal{C}_2)=\Pr(\sup_{t\in\Omega}d(t)\leq\mathcal{C}_2)\to 1$ together imply $\Pr(\inf_{t\in\Omega}\{n(t)-d(t)\}\geq\mathcal{C}_3 n)\to 1$ for some $\mathcal{C}_1,\mathcal{C}_2,\mathcal{C}_3>0$. So,

$$\sup_{t\in\Omega}\frac{1}{n(t)-d(t)}=\frac{1}{\inf_{t\in\Omega}\{n(t)-d(t)\}}=\mathcal{O}_{\mathrm{P}}(1/n).$$

Note that for all $t\in\Omega$, $n(t)=n(t-\tau)-d(t-\tau)-c(t-\tau)\leq n(t-\tau)$, where $c(t-\tau)$ is the number of subjects censored at $t-\tau$, so $n(t)$ is non-increasing over $t\in\Omega$. Surely, we have $n(0)=n$ and $d(0)=c(0)=0$. For $t\in\Omega$,

$$n(t-\tau)-d(t-\tau)=n(t)+c(t-\tau)\geq n(t)\geq n(t)-d(t),$$

so $n(t) - d(t)$ is non-increasing over $t \in \Omega$, resulting in $\inf_{t\in\Omega}\{n(t) - d(t)\} = n(\zeta) - d(\zeta) \geq 1$. As we can see that $n(\cdot) - d(\cdot)$ has the same asymptotic orders with $n(\cdot)$, the rest of the proof can be completed in a similar manner to what has already been done. ■

A.5.2 Example of ζ and τ Satisfying $\zeta = \tau\lceil\zeta/\tau\rceil$

Here we give example of ζ and τ satisfying $\zeta = \tau\lceil\zeta/\tau\rceil$, where $\tau \asymp 1/n^a \to 0$ as $n \to \infty$, e.g., $a = 1$ or $a = 3/4$. Let $\zeta \in \{1, 2, \dots\}$. For all $M \in \{0, 1, 2, \dots\}$, when $10^M \leq n^a < 10^{M+1}$, choose $\tau = 1/10^M$, which implies $\tau/(1/n^a) = \tau n^a \in [1, 10)$ for any n, directly showing $\tau \asymp 1/n^a$.

A.6 Additional Numerical Results

A.6.1 Standard Deviation Results in Simulation Studies

In this section, we compare the standard deviations (SD) for $\widehat{\beta}_{\mathrm{b}}$, $\widehat{\beta}_{\mathrm{e}}$, and $\widehat{\beta}_{\mathrm{pb}}$ from the simulation results in Sections 2.5. For scaling purposes, we present SD/β. Figure A.1 show the simulation results for SD considering all simulation cases. It is evident that for all cases, the SDs for $\widehat{\beta}_{\mathrm{b}}$, $\widehat{\beta}_{\mathrm{e}}$, and $\widehat{\beta}_{\mathrm{pb}}$ are all comparable, showing minimal differences even when τ is large or there is large variation among p_{ij}s. In these scenarios, as described in Section 2.5, $\widehat{\beta}_{\mathrm{pb}}$ outperforms the others in terms of RMSE and |Bias|.

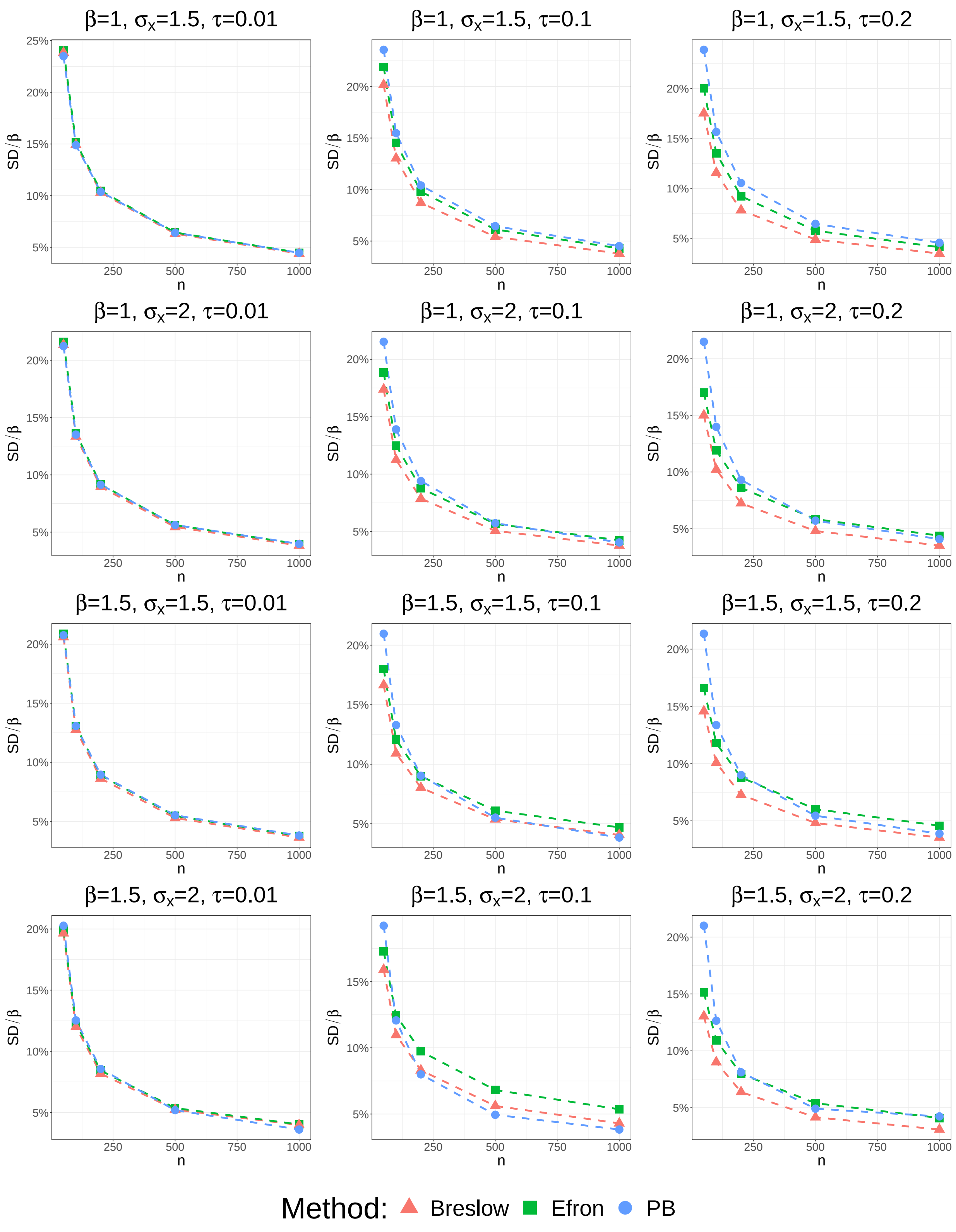

Figure A.1: Simulation results for SD.

A.6.2 Simulation for Choosing Baseline Hazard in the PB distribution Method

Here we do simulation to compare performances of using $\widehat{\boldsymbol{\Lambda}}_{\mathrm{b}}$, $\widehat{\boldsymbol{\Lambda}}_{\mathrm{e}}$, or $\widehat{\boldsymbol{\Lambda}}_{\mathrm{na}}$ as $\widehat{\boldsymbol{\Lambda}}$ in (2.14). The simulation setting in here is based on the Section 2.5, where $\beta = 1$, $\zeta = 1$, $\eta = \eta_c = 1.31$, $\gamma = \gamma_c = 1.5$, and $B = 10,000$. We generate x_i i.i.d. from $\mathrm{N}(0, 1)$. We vary $n \in \{50, 100, 200, 500, 1000\}$ and $\tau \in \{0.01, 0.1, 0.2\}$.

In Figure A.2, one can see that RMSE is very similar for different choices of $\widehat{\boldsymbol{\Lambda}}$ in (2.14), but choosing $\widehat{\boldsymbol{\Lambda}}_{\mathrm{e}}$ as $\widehat{\boldsymbol{\Lambda}}$ in (2.14) has less bias than the others for many cases.

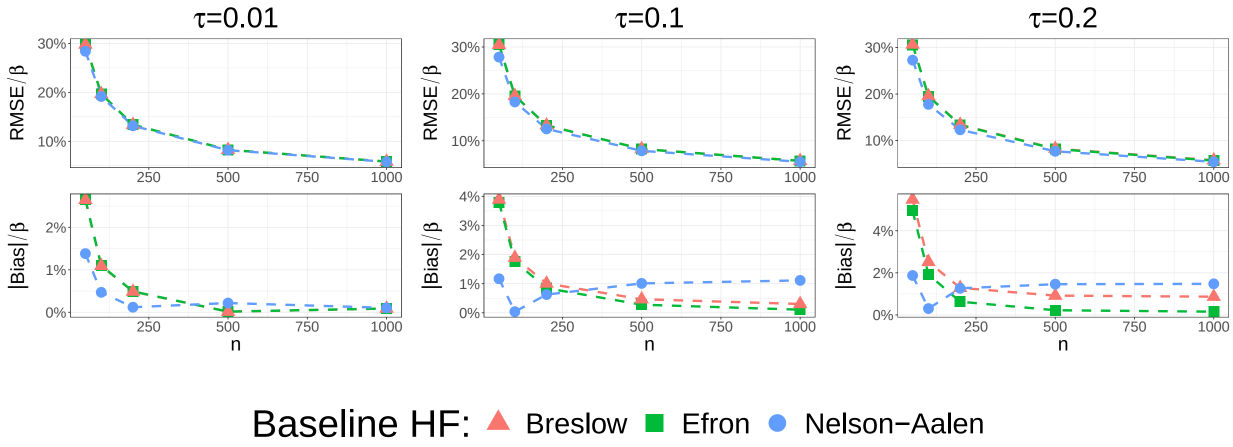

Figure A.2: Simulations for $\widehat{\beta}_{\mathrm{pb}}$ using Breslow, Efron, or Nelson-Aalen baseline hazard function as $\widehat{\boldsymbol{\Lambda}}$ in (2.14), respectively.

A.6.3 Simulation for Assumption A.3.10

Here we do simulation to check the important result from Assumption A.3.10, which is $n^{1/2}(\widehat{\boldsymbol{\beta}}_{\mathrm{pb}} - \widehat{\boldsymbol{\beta}}_{\mathrm{b}}) = o_{\mathrm{P}}(1)$. The simulation setting in here is based on the Section 2.5, where $\beta = 1$, $\tau = 1/n$, $\zeta = 1$, $\eta = \eta_c = 1.31$, $\gamma = \gamma_c = 1.5$, and $B = 10,000$. As $\tau = 1/n$, Assumption A.3.8 is satisfied allowing some ties. We generate x_i i.i.d. from $\mathrm{N}(0, 1)$. We vary $n \in \{50, 100, 200, 500, 1000\}$ to check $\widehat{\beta}_{\rm pb}$ and $\widehat{\beta}_{\rm b}$ get similar enough as n grows.

In Figure A.3, one can see that as n grows, one can see that $n^{1/2}(\widehat{\beta}_{\rm pb} - \widehat{\beta}_{\rm b})$ goes to 0 fast as n grows, indicating asymptotic normality of $n^{1/2}(\widehat{\beta}_{\rm pb} - \beta)$.

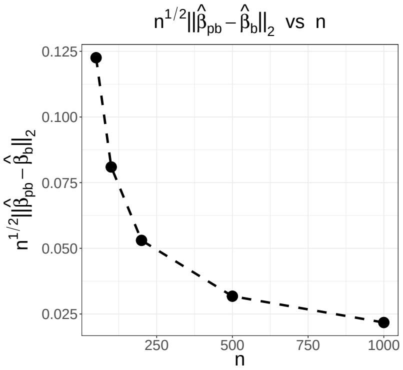

Figure A.3: Simulation for $n^{1/2}(\widehat{\boldsymbol{\beta}}_{\rm pb} - \widehat{\boldsymbol{\beta}}_{\rm b}) = o_{\rm P}(1)$.

A.6.4 Estimation Results in Real Applications

This section presents the estimates and standard errors for each β_j from real applications in Section 2.6. Table A.1 summarizes results from the laryngeal cancer patients study, while Table A.2 provides results from the north central cancer treatment group lung cancer study. The content in each cell of Tables A.1 and A.2 includes:

- (Breslow) the coefficient estimate $\widehat{\beta}_{{\rm b},j}$ and its standard error $\sqrt{\{I_{\rm b}(\widehat{\boldsymbol{\beta}}_{\rm b}, \zeta)^{-1}\}_{j,j}}$.
- (Efron) the coefficient estimate $\widehat{\beta}_{{\rm e},j}$ and its standard error $\sqrt{\{I_{\rm e}(\widehat{\boldsymbol{\beta}}_{\rm e}, \zeta)^{-1}\}_{j,j}}$.
- (PB distribution) the coefficient estimate $\widehat{\beta}_{{\rm pb},j}$ and its standard error $\sqrt{\{I_{\rm b}(\widehat{\boldsymbol{\beta}}_{\rm pb}, \zeta)^{-1}\}_{j,j}}$.

From Tables A.1 and A.2, it is evident that for estimating β_j, the results from the three methods are similar when τ is small. However, as τ increases, the Breslow method diverges significantly from PB distribution method. The Efron method shows less deviation from PB distribution method compared to Breslow. In contrast, the standard error estimates for β_j are consistent across all the methods.

Table A.1: Estimation result for the male laryngeal cancer patients study.

Covariate	Method	τ					
		0	0.05	0.1	0.15	0.2	0.25
Age	Breslow	0.20(0.15)	0.19(0.15)	0.19(0.15)	0.18(0.15)	0.22(0.15)	0.17(0.15)
	Efron	0.20(0.15)	0.20(0.15)	0.22(0.15)	0.20(0.15)	0.25(0.15)	0.20(0.15)
	PB	0.20(0.15)	0.20(0.15)	0.22(0.15)	0.21(0.15)	0.26(0.15)	0.20(0.15)
Stage 3	Breslow	0.60(0.32)	0.62(0.32)	0.62(0.32)	0.59(0.32)	0.57(0.32)	0.59(0.33)
	Efron	0.60(0.32)	0.64(0.32)	0.65(0.32)	0.64(0.32)	0.65(0.32)	0.68(0.32)
	PB	0.58(0.32)	0.63(0.33)	0.64(0.33)	0.63(0.33)	0.64(0.33)	0.68(0.34)
Stage 4	Breslow	1.65(0.40)	1.60(0.40)	1.49(0.39)	1.44(0.39)	1.23(0.39)	1.23(0.39)
	Efron	1.66(0.40)	1.68(0.40)	1.67(0.39)	1.67(0.40)	1.49(0.39)	1.53(0.39)
	PB	1.64(0.40)	1.67(0.39)	1.68(0.38)	1.69(0.38)	1.53(0.37)	1.58(0.38)

Table A.2: Estimation result for the north central cancer treatment group lung cancer study.

Covariate	Method	τ					
		0	0.05	0.1	0.15	0.2	0.25
Sex	Breslow	0.61(0.18)	0.58(0.18)	0.58(0.18)	0.55(0.18)	0.52(0.18)	0.49(0.18)
	Efron	0.62(0.18)	0.61(0.18)	0.65(0.18)	0.63(0.18)	0.64(0.18)	0.62(0.18)
	PB	0.61(0.18)	0.60(0.18)	0.65(0.18)	0.64(0.18)	0.66(0.18)	0.64(0.18)
ECOG (Physician)	Breslow	0.68(0.20)	0.63(0.20)	0.56(0.20)	0.54(0.20)	0.50(0.20)	0.48(0.20)
	Efron	0.68(0.20)	0.67(0.20)	0.66(0.20)	0.66(0.20)	0.67(0.20)	0.62(0.20)
	PB	0.68(0.20)	0.67(0.20)	0.66(0.20)	0.66(0.20)	0.70(0.20)	0.64(0.20)
Karnofsky (Patient)	Breslow	-0.22(0.11)	-0.20(0.10)	-0.18(0.10)	-0.13(0.10)	-0.16(0.10)	-0.18(0.10)
	Efron	-0.22(0.11)	-0.21(0.11)	-0.21(0.10)	-0.15(0.10)	-0.19(0.10)	-0.22(0.10)
	PB	-0.22(0.11)	-0.21(0.10)	-0.21(0.10)	-0.16(0.10)	-0.20(0.10)	-0.24(0.10)
Karnofsky (Physician)	Breslow	0.23(0.13)	0.20(0.13)	0.17(0.13)	0.16(0.13)	0.15(0.13)	0.17(0.13)
	Efron	0.23(0.13)	0.22(0.13)	0.20(0.13)	0.20(0.13)	0.20(0.13)	0.21(0.13)
	PB	0.23(0.13)	0.22(0.13)	0.21(0.13)	0.20(0.13)	0.21(0.13)	0.21(0.13)
Weight Loss	Breslow	-0.17(0.09)	-0.15(0.09)	-0.14(0.09)	-0.11(0.09)	-0.14(0.09)	-0.11(0.09)
	Efron	-0.17(0.09)	-0.16(0.09)	-0.16(0.09)	-0.14(0.09)	-0.18(0.09)	-0.16(0.09)
	PB	-0.17(0.09)	-0.16(0.09)	-0.16(0.09)	-0.15(0.09)	-0.19(0.09)	-0.18(0.09)

Appendix B

Details for "Competing Risk Model with A Nonparametric Form of Spline-Estimated Relative Risks"

B.1 Martingale Structure

Here, we derive the martingale structure used in the statistical theory for our model, based on Fleming and Harrington [16] and Kalbfleisch and Prentice [27]. Let $t \in [0, \tau]$. The filtration at t and t^- are each defined as

$$\mathcal{F}_t = \sigma\left(\left\{\{N_{ik}(s), Y_i(s^+), X_i\}_{i=1}^n; k \in \{1, \ldots, K\}; 0 \le s \le t\right\}\right),$$

$$\mathcal{F}_{t^-} = \sigma\left(\{\{N_{ik}(s), Y_i(s), X_i\}_{i=1}^n; k \in \{1, \ldots, K\}; 0 \le s < t\}\right),$$

where the sigma-algebra generated by $\cdot$ is denoted by $\sigma(\cdot)$.

As $\{\tilde{T}_i\}_{i=1}^n$ and $\{C_i\}_{i=1}^n$ are conditionally independent given $\{X_i\}_{i=1}^n$, we have

$$\mathbb{E}(dN_{ik}(t)|\mathcal{F}_{t^-}) = \mathbb{P}(dN_{ik}(t) = 1|\mathcal{F}_{t^-}) = Y_i(t)dH_k(t; X_i),$$

where $dN_{ik}(t) = N_{ik}(t^- + dt) - N_{ik}(t^-)$ and $dt > 0$. By the Doob-Meyer decomposition, we have

$$N_{ik}(t) = A_{ik}(t) + M_{ik}(t),$$

where $M_{ik}(t)$ is a zero-mean martingale and $A_{ik}(t) = \int_0^t dA_{ik}(s) = \int_0^t Y_i(s)dH_k(s; X_i)$ is the compensator. The predictable variation process and covariation process are each calculated as $\langle M_{ik}\rangle(t) = \int_0^t d\langle M_{ik}\rangle(s)$ and $\langle M_{ik}, M_{i'k}\rangle(t) = \int_0^t d\langle M_{ik}, M_{i'k}\rangle(s)$, where

$$\begin{aligned} d\langle M_{ik}\rangle(t) &\equiv \mathrm{Var}(dM_{ik}(t)|\mathcal{F}_{t^-}) = \mathrm{Var}(dN_{ik}(t)|\mathcal{F}_{t^-}) \\ &= \mathbb{E}(dN_{ik}(t)|\mathcal{F}_{t^-})(1 - \mathbb{E}(dN_{ik}(t)|\mathcal{F}_{t^-})) = Y_i(t)dH_k(t; X_i), \end{aligned}$$

where we regard $dt \times dt$ as 0. We have

$$\begin{aligned} d\langle M_{ik}, M_{i'k'}\rangle(t) &\equiv \mathrm{Cov}(dM_{ik}(t)dM_{i'k'}(t)|\mathcal{F}_{t^-}) = \mathrm{Cov}(dN_{ik}(t)dN_{i'k'}(t)|\mathcal{F}_{t^-}) \\ &= \mathbb{E}(dN_{ik}(t)dN_{i'k'}(t)|\mathcal{F}_{t^-}) - \mathbb{E}(dN_{ik}(t)|\mathcal{F}_{t^-})\mathbb{E}(dN_{i'k'}(t)|\mathcal{F}_{t^-}) \\ &= \mathbb{P}(dN_{ik}(t) = 1, dN_{i'k'}(t) = 1|\mathcal{F}_{t^-}) = 0 \end{aligned}$$

for $\{i, k\} \neq \{i', k'\}$ due to the continuity of the random variables (at most one jump at t among all n subjects) and the fact that a single subject cannot experience more than two risks.

B.2 Notations

We introduce the notations used throughout our theory. We write $\mathrm{P}_W = \mathbb{P}(W^{-1}(\cdot))$ for the distribution of W, and $\mathrm{P}_{W^n} = \mathbb{P}(\{W_1, \ldots, W_n\}^{-1}(\cdot))$ for the distribution of $\{W_i\}_{i=1}^n$. We use the operator $\mathbb{E}_W$ to indicate the Lebesgue integral taken with respect to the distribution of either W or $\{W_i\}_{i=1}^n$. That is,

$$\mathbb{E}_W[g(W)] = \int_{\mathcal{W}} g \, d\mathrm{P}_W \text{ and } \mathbb{E}_W[g(\{W_i\}_{i=1}^n)] = \int_{\mathcal{W}^n} g \, d\mathrm{P}_{W^n}$$

for any function g defined on $\mathcal{W}$ or $\mathcal{W}^n$, where $\mathcal{W} = (0,\infty) \times (0,\tau] \times \mathcal{X} \times \{1,\ldots,K\}$. The function g may be deterministic or a stochastic process. If g is deterministic,

$$\mathbb{E}_W[g(W)] = \mathbb{E}[g(W)] \text{ and } \mathbb{E}_W[g(\{W_i\}_{i=1}^n)] = \mathbb{E}[g(\{W_i\}_{i=1}^n)].$$

For deterministic sequences $a_n, b_n \in \mathbb{R}$ for $n \in \mathbb{N}$, we set $a_n = \mathcal{O}(b_n)$ ($|a_n| \lesssim |b_n|$) if there exists a constant $\mathcal{C} \in (0,\infty)$ and $N \in \mathbb{N}$ such that $|a_n| \leq \mathcal{C}|b_n|$ for all $n \geq N$. Likewise, $a_n = o(b_n)$ ($|a_n| \ll |b_n|$) is satisfied if for every constant $\epsilon \in (0,\infty)$, there exists $N_\epsilon \in \mathbb{N}$ such that $|a_n| \leq \epsilon |b_n|$ for all $n \geq N_\epsilon$. We adopt $a_n \asymp b_n$ to mean that $a_n = \mathcal{O}(b_n)$ and $b_n = \mathcal{O}(a_n)$.

For sequences $X_n, Y_n \in \mathbb{R}$ for $n \in \mathbb{N}$, with at least one being random, we refer to $X_n = \mathcal{O}_{\mathbb{P}}(Y_n)$ ($|X_n| \lesssim |Y_n|$) if there is a constant $\mathcal{C} \in (0,\infty)$ such that $\mathbb{P}(|X_n| \leq \mathcal{C}|Y_n|) \to 1$ as $n \to \infty$. In addition, $X_n = o_{\mathbb{P}}(Y_n)$ ($|X_n| \ll |Y_n|$) if $\mathbb{P}(|X_n| \leq \epsilon|Y_n|) \to 1$ as $n \to \infty$ for any constant $\epsilon \in (0,\infty)$.

Here, if a constant does not have a subscript k, it is independent of k.

Let $\mathcal{U}_1$ and $\mathcal{U}_2$ be vector spaces that are subspaces of a vector space $\mathcal{U}$. We refer to $\mathcal{U} = \mathcal{U}_1 \oplus \mathcal{U}_2$, which can also be expressed as $\mathcal{U}_1 = \mathcal{U} \ominus \mathcal{U}_2$ or $\mathcal{U}_2 = \mathcal{U} \ominus \mathcal{U}_1$, if the followings are satisfied:

(i) $\mathcal{U}_1 \cap \mathcal{U}_2 = \{0\}$,

(ii) For every $f \in \mathcal{U}$, there exist unique $f_1 \in \mathcal{U}_1$ and $f_2 \in \mathcal{U}_2$ such that $f = f_1 + f_2$.

For each $j = 1,\ldots,d$, suppose that $\{\varphi_{j,v}\}_{v \in \mathcal{I}_j}$ forms a basis of Hilbert spaces $\mathcal{V}_j$, where index set $\mathcal{I}_j$ is either finite (when $\mathcal{V}_j$ is finite-dimensional, $|\mathcal{I}_j| \in \mathbb{N}$) or countably infinite (when $\mathcal{V}_j$ is infinite-dimensional, $\mathcal{I}_j = \mathbb{N}$). We denote tensor product space by $\otimes_{j=1}^{d}\mathcal{V}_j$, which is the completion of Span $\left\{\prod_{j=1}^{d} \varphi_{j,v_j}\right\}_{v_j\in\mathcal{I}_j, j=1,\ldots,d}$ with respect to some inner product $\langle\cdot,\cdot\rangle_{\otimes_{j=1}^{d}\mathcal{V}_j}$ meeting $\langle\prod_{j=1}^{d} f_j, \prod_{j=1}^{d} g_j\rangle_{\otimes_{j=1}^{d}\mathcal{V}_j} = \prod_{j=1}^{d}\langle f_j, g_j\rangle_{\mathcal{V}_j}$ for all $f_j, g_j \in \mathcal{V}_j$, $j = 1,\ldots,d$, where for each $j = 1,\ldots,d$, $\langle\cdot,\cdot\rangle_{\mathcal{V}_j}$ is some inner product on $\mathcal{V}_j$.

B.3 Assumptions

In this section, we provide assumptions needed in the rate of convergence for the proposed model. Let $\mathsf{f}_{\mathcal{X}}(\cdot)$ be the density function of X. For $t \in [0,\tau]$ and $x \in \mathcal{X}$, let $\mathsf{f}_{\mathcal{Y}|\mathcal{X}}(t;x) = \mathbb{E}[Y(t)|X=x] = \mathbb{P}(Y(t)=1|X=x)$ and $\mathsf{f}_{\mathcal{Y},\mathcal{X}}(t,x) = \mathsf{f}_{\mathcal{Y}|\mathcal{X}}(t;x)\mathsf{f}_{\mathcal{X}}(x)$.

Assumption B.3.1. *Assume the followings:*

- *(i) There exist some constants $\mathcal{C}_\tau \in (0,1)$ and $\mathcal{C}_S \in (0,\infty)$ such that $\mathbb{P}(T \geq \tau) \geq \mathcal{C}_\tau$ and $\inf_{t\in[0,\tau]} S(\eta_k^*, t) = S(\eta_k^*, \tau) > \mathcal{C}_S$ for each k.*
- *(ii) There exists some constant $\mathcal{C}_R \in (0,1)$ such that for all k, $\mathbb{P}(\tilde{G} = k) \geq \mathcal{C}_R$.*
- *(iii) There exist some constants $0 < \mathcal{C}_{h,1} < \mathcal{C}_{h,2} < \infty$ such that for all k, $\mathcal{C}_{h,1} \leq h_{0k}(t) \leq \mathcal{C}_{h,2}$ for all $t \in [0,\tau]$.*
- *(iv) There exist some constants $0 < \mathcal{C}_{\mathsf{f},1} < \mathcal{C}_{\mathsf{f},2} < \infty$ such that for all k, $\mathcal{C}_{\mathsf{f},1} < \mathsf{f}_{\mathcal{Y},\mathcal{X}}(t,x) < \mathcal{C}_{\mathsf{f},2}$ for all $t \in [0,\tau]$ and all $x \in \mathcal{X}$.*
- *(v) There exists some constant $\mathcal{C}_* \in (0,\infty)$ such that for all k, $-\mathcal{C}_* \leq \eta_k^*(x) \leq \mathcal{C}_*$ for all $x \in \mathcal{X}$. Furthermore, for each k, there exists deterministic big convex neighborhood of η_k^*, denoted by $\mathcal{B}_k = \{f \in \mathcal{H} : \mathcal{V}_k(f - \eta_k^*) \leq \rho_k\}$ with some constant $\rho_k \in (0,\infty)$, which*

does not depend on n, such that there exist some constants $0 < \mathcal{C}_{k,1} < 1 < \mathcal{C}_{k,2} < \infty$ such that for all $\eta \in \mathcal{B}_k$, $\mathcal{C}_{k,1} \exp(\eta_k^*(x)) \leq \exp(\eta(x)) \leq \mathcal{C}_{k,2} \exp(\eta_k^*(x))$ for all $x \in \mathcal{X}$.

(vi) *For each k, $\mathcal{V}_k$ is completely continuous with respect to J_k on $\mathcal{H}$, implying that there exists an eigensystem $\{\rho_{k,v}, \psi_{k,v}\}_{v\in\mathbb{N}}$ such that for all $v \in \mathbb{N}$, the eigenfunction $\psi_{k,v} \in \mathcal{H}$ and the eigenvalue $\rho_{k,v} \geq 0$ satisfy $\rho_{k,v} \to \infty$ as $v \to \infty$, and $\mathcal{V}_k(\psi_{k,v}, \psi_{k,v'}) = \mathbb{1}_{v,v'} \equiv \mathbb{1}(v = v')$, while $J_k(\psi_{k,v}, \psi_{k,v'}) = \rho_{k,v}\mathbb{1}_{v,v'}$ for all $v, v' \in \mathbb{N}$. Furthermore, for any $f \in \mathcal{H}$, we have $f = \sum_{v\in\mathbb{N}} \mathcal{V}_k(f, \psi_{k,v})\psi_{k,v}$, since for all $f \in \mathcal{H}$, $J(f) < \infty$, which is equivalent to $J_k(f) < \infty$.*

There exist some constants $\mathcal{C}_\psi, \mathcal{C}_\rho \in (0,\infty)$ such that for each k, $\sup_{v\in\mathbb{N}} \|\psi_{k,v}\|_{\sup} \leq \mathcal{C}_\psi$ and $\rho_{k,v} > \mathcal{C}_\rho v^r$ as $v \to \infty$, where $r > 1$. Note that $r = 2m$ when $d = 1$, or when $d > 1$ and $\mathcal{H}$ has an additive structure; and $r = 2m - \delta$ for any $\delta \in (0, 2m-1)$ when $d > 1$ and $\mathcal{H}$ contains at least one interaction.

(vii) *For each k, there exists some constant $b_k \in [1,2]$ such that $\sum_{v\in\mathbb{N}} \rho_{k,v}^{b_k} \mathcal{V}_k^2(\eta_k^*, \psi_{k,v}) < \infty$.*

(viii) *For each k, there exists some constant $\zeta_k \in (0,\infty)$ such that as $n \to \infty$, $q_k \asymp n^{\zeta_k + 2/(rb_k+1)}$, $\lambda_k \to 0$, and $q_k \lambda_k^{2/r} \to \infty$.*

Let $n(\cdot) = \sum_{i=1}^n Y_i(\cdot)$. By the strong law of large numbers and (1), which is common in the survival analysis literature, we have $n(\tau)/n \xrightarrow{\text{a.s.}} \mathbb{P}(T \geq \tau) > \mathcal{C}_\tau$ as $n \to \infty$. Given that $n(\cdot)$ is non-increasing on $[0,\tau]$, there are sufficient number of units at risk for any $t \in [0,\tau]$. Therefore, for simplicity, we assume $n(\tau) \geq 1$, which is reasonable based on the assumption above.

Note that (2) indicates that the probability of each risk factor is above a certain threshold. This implies that, when n is sufficiently large, for each risk factor, we have a sufficiently large number of subjects with observed risk corresponding to that risk factor.

The uniformly boundedness assumptions in (3) and (4) are borrowed from Du et al. [11]. Specifically, (3) leads to $\mathcal{C}_{h,1}\tau \leq \inf_{t\in[0,\tau]} h_{0k}(t)\times\tau \leq \int_0^\tau dH_{0k}(t) = H_{0k}(\tau) \leq \sup_{t\in[0,\tau]} h_{0k}(t)\times\tau \leq \mathcal{C}_{h,2}\tau$. Observe that (4) is a uniformly boundedness condition used to show that Theorem 3.1 implies rate of convergence with respect to the $\mathcal{L}_2$ norm.

The set $\mathcal{B}_k$ in (5) is large enough to ensure that all estimators of η_k^* in the technical proofs, including $\hat{\eta}_k$ and $\hat{\eta}_k^*$, are elements of $\mathcal{B}_k$. This is a common assumption in the smoothing spline literature [11, 18].

Constructing the eigensystem on $\mathcal{H}$ for each k, as described in (6), is a standard approach in the smoothing spline literature [11, 18].

The smoothness level of the true function η_k^* is described by the parameter b_k. When $b_k = 1$, we have $\sum_{v\in\mathbb{N}} \rho_{k,v}^{b_k} \mathcal{V}_k^2(\eta_k^*, \psi_{k,v}) = J_k(\eta_k^*)$, ensuring that condition (7) is automatically satisfied by $\eta_k^* \in \mathcal{H}$.

The order of the number of knots required for efficient approximation, as well as the order of the smoothing parameter in relation to the number of knots, are discussed in (8). Both are common assumptions in the smoothing spline literature [11, 18]. When $\lambda_k \asymp n^{-r/(rb_k+1)} \to 0$ as $n \to \infty$, which guarantees the optimal convergence rate, $q_k\lambda_k^{2/r} \asymp n^{\zeta_k} \to \infty$ as $n \to \infty$ is also satisfied, thereby ensuring that (8) holds.

B.4 Proof of Theorem 3.1

Here, we employ quadratic approximation techniques as outlined by Gu [18] to prove Theorem 3.1, with proofs closely following those in Du et al. [11]. Without loss of generality, any arguments here are for any k. First, we introduce some notations that are used in the proofs. For $\eta, f, g \in \mathcal{H}$, let $\mathfrak{U}(f;\eta,t) = S(f;\eta,t)/S(\eta,t)$, $\mathfrak{U}_{(n)}(f;\eta,t) = S_{(n)}(f;\eta,t)/S_{(n)}(\eta,t)$,

$\mathcal{V}(f, g; \eta, t) = S(fg; \eta, t)/S(\eta, t) - \mathcal{U}(f; \eta, t)\mathcal{U}(g; \eta, t)$, and

$$\mathcal{V}_{(n)}(f, g; \eta, t) = S_{(n)}(fg; \eta, t)/S_{(n)}(\eta, t) - \mathcal{U}_{(n)}(f; \eta, t)\mathcal{U}_{(n)}(g; \eta, t),$$

where $S_{(n)}(\eta, t) \equiv \sum_{i=1}^{n} Y_i(t)\exp(\eta(X_i))/n$, $S_{(n)}(f; \eta, t) \equiv \sum_{i=1}^{n} Y_i(t)f(X_i)\exp(\eta(X_i))/n$, and $S_{(n)}(fg; \eta, t) \equiv \sum_{i=1}^{n} Y_i(t)f(X_i)g(X_i)\exp(\eta(X_i))/n$ for $f, g, \eta \in \mathcal{H}$ and $t \in [0, \tau]$.

Then, we can express $\mathcal{V}_k(\cdot, \cdot)$ as

$$\mathcal{V}_k(f, g) = \int_0^{\tau} \mathcal{V}(f, g; \eta_k^*, t)S(\eta_k^*, t)dH_{0k}(t)$$

for $f, g \in \mathcal{H}$. For $\eta \in \mathcal{H}$, we define a non-negative definite bivariate functional $\mathcal{V}_{\eta,k}^{(n)}(\cdot, \cdot)$ on $\mathcal{H}$ as follows. For $f, g \in \mathcal{H}$, let

$$\mathcal{V}_{\eta,k}^{(n)}(f, g) = \int_0^{\tau} \mathcal{V}_{(n)}(f, g; \eta, t)d\overline{N}_k(t).$$

Let $\mathcal{V}_{\eta,k}^{(n)}(f) = \mathcal{V}_{\eta,k}^{(n)}(f, f)$ for $\eta, f \in \mathcal{H}$. Define $\mathcal{V}_k^{(n)}(\cdot, \cdot) = \mathcal{V}_{\eta_k^*,k}^{(n)}(\cdot, \cdot)$. For $\eta, f \in \mathcal{H}$, let

$$\mathcal{U}_{\eta,k}^{(n)}(f) = \int_0^{\tau} \mathcal{U}_{(n)}(f; \eta, t)d\overline{N}_k(t), \text{ where } \mathcal{U}_k^{(n)}(\cdot) \equiv \mathcal{U}_{\eta_k^*,k}^{(n)}(\cdot).$$

We begin by introducing the remarks and lemmas that are instrumental in our main proof. Initially, we require an order bound for the eigenvalues and λ_k, as described in Remark B.1.

Remark B.1. Under Assumption B.3.1, for each k,

$$\sum_{v\in\mathbb{N}} \frac{\lambda_k \rho_{k,v}}{(1 + \lambda_k \rho_{k,v})^2}, \sum_{v\in\mathbb{N}} \frac{1}{(1 + \lambda_k \rho_{k,v})^2}, \text{ and } \sum_{v\in\mathbb{N}} \frac{1}{1 + \lambda_k \rho_{k,v}}$$

are $\mathcal{O}_{\lambda_k}(\lambda_k^{-1/r})$, where $\mathcal{O}_{\lambda_k}(\cdot)$ is $\mathcal{O}(\cdot)$ with respect to $\lambda_k \to 0$.

The proof of Remark B.1 is provided in Gu [18], where the condition $r > 1$ is required. Since $\lambda_k \to 0$ as $n \to \infty$, $\mathcal{O}_{\lambda_k}(\cdot)$ in Remark B.1 is replaced with $\mathcal{O}(\cdot)$ as $n \to \infty$.

We demonstrate that $\mathcal{V}_k$ and $\|\cdot\|^2_{\mathcal{L}_2}$ are equivalent in $\mathcal{H}$, as established by Lemma B.2.

Lemma B.2. *Under Assumption B.3.1, there exist some constants* $0 < \mathcal{C}_{\mathcal{L}_2,1} < \mathcal{C}_{\mathcal{L}_2,2} < \infty$ *such that for each* k*, for all* $f \in \mathcal{H}$, $\mathcal{C}_{\mathcal{L}_2,1}\|f\|^2_{\mathcal{L}_2} \leq \mathcal{V}_k(f) \leq \mathcal{C}_{\mathcal{L}_2,2}\|f\|^2_{\mathcal{L}_2}$.

The proof of Lemma B.2 is detailed in Section B.5. According to Lemma B.2, the convergence rate between $\hat{\eta}_k$ and η^*_k with respect to $\|\cdot\|_{\mathcal{L}_2}$ is equivalent to that with respect to $(\mathcal{V}_k + \lambda_k J_k)^{1/2}$ as stated in Theorem 3.1.

We present a series of important results related to convergence and asymptotics in Lemmas B.3 – B.5, with detailed proofs provided in Section B.5. Additionally, we establish results concerning boundedness and orthogonality in Lemmas B.6 and B.7, also included in Section B.5.

Lemma B.3. *Under Assumption B.3.1, for each* k*, for all* $f, g \in \mathcal{H}$*, as* $n \to \infty$*,*

$$\begin{aligned}\mathcal{V}^{(n)}_k(f,g) &= \mathcal{V}_k(f,g) + \sqrt{(\mathcal{V}_k + \lambda_k J_k)(f)(\mathcal{V}_k + \lambda_k J_k)(g)}\,\mathcal{O}_{\mathbb{P}}\left(n^{-1/2}\lambda_k^{-1/r}\right)\\ &= \mathcal{V}_k(f,g) + o_{\mathbb{P}}\left(\sqrt{(\mathcal{V}_k + \lambda_k J_k)(f)(\mathcal{V}_k + \lambda_k J_k)(g)}\right).\end{aligned}$$

Lemma B.4. *Under Assumption B.3.1, for each* k*, for all* $f, g \in \mathcal{H}$*, there exists some random variable* $\alpha \in (0,1)$*, such that*

$$\mathcal{U}^{(n)}_{f,k}(f-g) - \mathcal{U}^{(n)}_{g,k}(f-g) = \mathcal{V}^{(n)}_{g+\alpha(f-g),k}(f-g).$$

Lemma B.5. *Under Assumption B.3.1, for each* k*, let a stochastic process* $f_k \in \mathcal{H}$ *satisfies*

$\mathcal{V}_k^{(n)}(f_k - \eta_k^*) = o_{\mathbb{P}}(1)$ *as* $n \to \infty$. *Then for all* $h \in \mathcal{H}$, *as* $n \to \infty$,

$$\mathcal{U}_{f_k,k}^{(n)}(h) - \mathcal{U}_k^{(n)}(h) = \mathcal{V}_k^{(n)}(f_k - \eta_k^*, h)\left(1 + o_{\mathbb{P}}(1)\right).$$

Furthermore, if another stochastic process $g_k \in \mathcal{H}$ *satisfies* $\mathcal{V}_k^{(n)}(g_k - \eta_k^*) = o_{\mathbb{P}}(1)$ *as* $n \to \infty$. *Then for all* $h \in \mathcal{H}$, *as* $n \to \infty$,

$$\mathcal{U}_{f_k,k}^{(n)}(h) - \mathcal{U}_{g_k,k}^{(n)}(h) = \mathcal{V}_k^{(n)}(f_k - g_k, h)\left(1 + o_{\mathbb{P}}(1)\right).$$

The proof for the main result in Theorem 3.1 is presented in the following three steps.

Step 1: Rate for the Quadratic Approximation Optimizer

For $\eta, f \in \mathcal{H}$, define

$$\mathcal{S}_{\eta,k}^{(n)}(f) = \int_0^\tau \frac{1}{n}\sum_{i=1}^n \left\{f(X_i) - \mathfrak{U}_{(n)}(f;\eta,t)\right\} dN_{ik}(t),$$

which represents the Fréchet derivative of $-\mathcal{L}_{k(n)}(\cdot)$ at η in the direction of f.

The quadratic approximation optimizer $\tilde{\eta}_k^*$ is defined by

$$\tilde{\eta}_k^* = \underset{f \in \mathcal{H}}{\operatorname{argmin}} \; -\mathcal{S}_{\eta_k^*,k}^{(n)}(f) + \frac{1}{2}\mathcal{V}_k(f - \eta_k^*) + \frac{\lambda_k}{2} J_k(f), \tag{B.1}$$

where assume $\tilde{\eta}_k^* \in \mathcal{B}_k$. The loss function in (B.1) is a quadratic approximation of the loss function in (3.1) at η_k^*, which can be represented as

$$\sum_{v \in \mathbb{N}} \left\{ -\mathcal{V}_k(f, \psi_{k,v})\mathcal{S}_{\eta_k^*,k}^{(n)}(\psi_{k,v}) + \frac{1}{2}\left(\mathcal{V}_k(f,\psi_{k,v}) - \mathcal{V}_k(\eta_k^*, \psi_{k,v})\right)^2 + \frac{\lambda_k}{2}\rho_{k,v}\mathcal{V}_k^2(f,\psi_{k,v}) \right\}. \tag{B.2}$$

Setting the derivative of (B.2) at $\mathcal{V}_k(f, \psi_{k,v})$ to be 0, one gets $\tilde{\eta}_k^* = \sum_{v \in \mathbb{N}} \mathcal{V}_k(\tilde{\eta}_k^*, \psi_{k,v})\psi_{k,v}$, where

$$\mathcal{V}_k(\tilde{\eta}_k^*, \psi_{k,v}) = \frac{1}{1 + \lambda_k \rho_{k,v}} \left\{ \mathcal{V}_k(\eta_k^*, \psi_{k,v}) + \mathcal{S}^{(n)}_{\eta_k^*,k}(\psi_{k,v}) \right\}.$$

One can represent $\mathcal{S}^{(n)}_{\eta_k^*,k}(\psi_{k,v})$ by

$$\mathcal{S}^{(n)}_{\eta_k^*,k}(\psi_{k,v}) = \frac{1}{n}\sum_{i=1}^{n} \int_0^\tau \left\{ \psi_{k,v}(X_i) - \mathcal{U}_{(n)}(\psi_{k,v}; \eta_k^*, t) \right\} dM_{ik}(t) = \frac{1}{n}\sum_{i=1}^{n} \int_0^\tau B^{(v)}_{i,k}(t) dM_{ik}(t),$$

where

$$\begin{aligned}
\mathbb{E}\left[\mathcal{S}^{(n)}_{\eta_k^*,k}(\psi_{k,v})\right] &= \frac{1}{n}\sum_{i=1}^{n} \int_0^\tau \mathbb{E}\left[B^{(v)}_{i,k}(t) dM_{ik}(t)\right] = \frac{1}{n}\sum_{i=1}^{n} \int_0^\tau \mathbb{E}\left[\mathbb{E}\left[B^{(v)}_{i,k}(t) dM_{ik}(t) \middle| \mathcal{F}_{t^-}\right]\right] \\
&= \frac{1}{n}\sum_{i=1}^{n} \int_0^\tau \mathbb{E}\left[B^{(v)}_{i,k}(t) \mathbb{E}\left[dM_{ik}(t) | \mathcal{F}_{t^-}\right]\right] = 0
\end{aligned}$$

and

$$\mathbb{E}\left[\left\{\mathcal{S}^{(n)}_{\eta_k^*,k}(\psi_{k,v})\right\}^2\right] = \frac{1}{n}\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} \int_0^\tau \left\{B^{(v)}_{i,k}(t)\right\}^2 d\langle M_{ik}\rangle(t)\right] \le \frac{2\exp(\mathcal{C}_*)\mathcal{C}_\psi^2 \mathcal{C}_{h,2}\tau}{n}. \quad \text{(B.3)}$$

In (B.3), the first equality is derived by

$$\begin{aligned}
\mathbb{E}\left[\int_0^\tau B^{(v)}_{i,k}(t) dM_{ik}(t) \int_0^\tau B^{(v)}_{i',k}(t) dM_{i'k}(t)\right] &= \mathbb{E}\left[\int_0^\tau B^{(v)}_{i,k}(t) B^{(v)}_{i',k}(t) d\langle M_{ik}, M_{i'k}\rangle(t)\right] \\
&= \mathbb{E}\left[\int_0^\tau \left\{B^{(v)}_{i,k}(t)\right\}^2 d\langle M_{ik}\rangle(t)\right] \mathbb{1}_{i,i'},
\end{aligned}$$

which follows from Theorem 2.4.4. in Fleming and Harrington [16], and the second equality is derived by

$$\frac{1}{n}\sum_{i=1}^{n} \int_0^\tau \left\{B^{(v)}_{i,k}(t)\right\}^2 d\langle M_{ik}\rangle(t) = \int_0^\tau \mathcal{V}_{(n)}(\psi_{k,v}, \psi_{k,v}; \eta_k^*, t) S_{(n)}(\eta_k^*, t) dH_{0k}(t)$$

$$\leq \left\{ \sup_{v,v'\in\mathbb{N},t\in[0,\tau]} \left|\mathcal{V}_{(n)}(\psi_{k,v}, \psi_{k,v'}; \eta_k^*, t)\right| \right\} \left\{ \sup_{t\in[0,\tau]} S_{(n)}(\eta_k^*, t) \right\} H_{0k}(\tau)$$

$$\leq 2\exp(\mathcal{C}_*)\mathcal{C}_\psi^2 \mathcal{C}_{h,2}\tau,$$

where the last inequality follows from Lemma B.6.

Simple calculation yields

$$\left(\mathcal{V}_k + \lambda_k J_k\right)\left(\tilde{\eta}_k^* - \eta_k^*\right) = \sum_{v\in\mathbb{N}} \frac{\left\{\mathcal{S}^{(n)}_{\eta_k^*,k}(\psi_{k,v})\right\}^2 - 2\lambda_k\rho_{k,v}\mathcal{V}_k(\eta_k^*, \psi_{k,v})\mathcal{S}^{(n)}_{\eta_k^*,k}(\psi_{k,v}) + \lambda_k^2\rho_{k,v}^2\mathcal{V}_k^2(\eta_k^*, \psi_{k,v})}{1 + \lambda_k\rho_{k,v}}$$

and

$$\mathbb{E}\left[\left(\mathcal{V}_k + \lambda_k J_k\right)\left(\tilde{\eta}_k^* - \eta_k^*\right)\right] \leq \frac{2\exp(\mathcal{C}_*)\mathcal{C}_\psi^2\mathcal{C}_{h,2}\tau}{n} \sum_{v\in\mathbb{N}} \frac{1}{1 + \lambda_k\rho_{k,v}} + \lambda_k^{b_k} \sum_{v\in\mathbb{N}} \frac{\lambda_k^{2-b_k}\rho_{k,v}^{2-b_k}}{1 + \lambda_k\rho_{k,v}} \rho_{k,v}^{b_k}\mathcal{V}_k^2(\eta_k^*, \psi_{k,v})$$

$$\begin{aligned} &\leq \frac{2\exp(\mathcal{C}_*)\mathcal{C}_\psi^2\mathcal{C}_{h,2}\tau}{n} \sum_{v\in\mathbb{N}} \frac{1}{1 + \lambda_k\rho_{k,v}} + \lambda_k^{b_k} \sum_{v\in\mathbb{N}} \rho_{k,v}^{b_k}\mathcal{V}_k^2(\eta_k^*, \psi_{k,v}) \\ &= \mathcal{O}\left(n^{-1}\lambda_k^{-1/r} + \lambda_k^{b_k}\right), \end{aligned}$$

directly implying $\sqrt{\left(\mathcal{V}_k + \lambda_k J_k\right)\left(\tilde{\eta}_k^* - \eta_k^*\right)} = \mathcal{O}_{\mathbb{P}}\left(n^{-1/2}\lambda_k^{-1/(2r)} + \lambda_k^{b_k/2}\right)$.

Step 2: Rate for the Global Optimizer

For $\eta, f \in \mathcal{H}$, let $A(\eta, f)$ and $B(\eta, f)$ denote the Fréchet derivatives of the loss functions in (3.1) and (B.1), respectively, at η in the direction of f. These derivatives are expressed as

$$A(\eta, f) = -\mathcal{S}^{(n)}_{\eta,k}(f) + \lambda_k J_k(\eta, f), \tag{B.4}$$

$$B(\eta, f) = -\mathcal{S}^{(n)}_{\eta^*_k,k}(f) + \mathcal{V}_k(\eta - \eta^*_k, f) + \lambda_k J_k(\eta, f), \tag{B.5}$$

where, by definition, $A(\hat{\eta}^*_k, f) = 0$ and $B(\tilde{\eta}^*_k, f) = 0$ for any $f \in \mathcal{H}$. Set $\eta = \hat{\eta}^*_k$ and $f = \hat{\eta}^*_k - \tilde{\eta}^*_k$ in (B.4) and set $\eta = \tilde{\eta}^*_k$ and $f = \hat{\eta}^*_k - \tilde{\eta}^*_k$ in (B.5). Subtracting the resulting equations gives

$$\begin{aligned}
&\mathcal{U}^{(n)}_{\hat{\eta}^*_k,k}(\hat{\eta}^*_k - \tilde{\eta}^*_k) - \mathcal{U}^{(n)}_{\tilde{\eta}^*_k,k}(\hat{\eta}^*_k - \tilde{\eta}^*_k) + \lambda_k J_k(\hat{\eta}^*_k - \tilde{\eta}^*_k) \\
&\quad = \mathcal{V}_k(\tilde{\eta}^*_k - \eta^*_k, \hat{\eta}^*_k - \tilde{\eta}^*_k) + \mathcal{U}^{(n)}_k(\hat{\eta}^*_k - \tilde{\eta}^*_k) - \mathcal{U}^{(n)}_{\tilde{\eta}^*_k,k}(\hat{\eta}^*_k - \tilde{\eta}^*_k).
\end{aligned} \tag{B.6}$$

By Lemma B.4, for some random variable $\alpha_1 \in (0, 1)$, we have

$$\begin{aligned}
0 \leq \quad & \mathcal{U}^{(n)}_{\hat{\eta}^*_k,k}(\hat{\eta}^*_k - \tilde{\eta}^*_k) - \mathcal{U}^{(n)}_{\tilde{\eta}^*_k,k}(\hat{\eta}^*_k - \tilde{\eta}^*_k) + \lambda_k J_k(\hat{\eta}^*_k - \tilde{\eta}^*_k) \\
& = \mathcal{V}^{(n)}_{\alpha_1\hat{\eta}^*_k + (1-\alpha_1)\tilde{\eta}^*_k,k}(\hat{\eta}^*_k - \tilde{\eta}^*_k) + \lambda_k J_k(\hat{\eta}^*_k - \tilde{\eta}^*_k) \\
& \geq \mathcal{C}_{\mathcal{B}_k}(\mathcal{V}^{(n)}_k + \lambda_k J_k)(\hat{\eta}^*_k - \tilde{\eta}^*_k) \\
& = \mathcal{C}_{\mathcal{B}_k}(\mathcal{V}_k + \lambda_k J_k)(\hat{\eta}^*_k - \tilde{\eta}^*_k) + o_{\mathbb{P}}\left((\mathcal{V}_k + \lambda_k J_k)(\hat{\eta}^*_k - \tilde{\eta}^*_k)\right),
\end{aligned} \tag{B.7}$$

where $\alpha_1\hat{\eta}^*_k + (1-\alpha_1)\tilde{\eta}^*_k \in \mathcal{B}_k$, and considering the boundedness of η^*_k and the closeness of $\exp(\eta)$ to $\exp(\eta^*_k)$ for $\eta \in \mathcal{B}_k$, as described in (6) of Assumption B.3.1, we assume, without loss of generality, that for some constant $\mathcal{C}_{\mathcal{B}_k} \in (0, 1)$, the inequality $\mathcal{V}^{(n)}_{\eta,k}(f) \geq \mathcal{C}_{\mathcal{B}_k}\mathcal{V}^{(n)}_k(f)$ holds for all $\eta \in \mathcal{B}_k$ and $f \in \mathcal{H}$.

By Lemma B.3, we have $\mathcal{V}^{(n)}_k(\tilde{\eta}^*_k - \eta^*_k) = \mathcal{V}_k(\tilde{\eta}^*_k - \eta^*_k) + o_{\mathbb{P}}\left((\mathcal{V}_k + \lambda_k J_k)(\tilde{\eta}^*_k - \eta^*_k)\right) = o_{\mathbb{P}}(1)$. So, by Lemmas B.5 and B.3, we get

$$\begin{aligned} 0 \leq\ & \mathcal{V}_k(\tilde{\eta}_k^* - \eta_k^*, \hat{\eta}_k^* - \tilde{\eta}_k^*) + \mathcal{U}_k^{(n)}(\hat{\eta}_k^* - \tilde{\eta}_k^*) - \mathcal{U}_{\tilde{\eta}_k^*,k}^{(n)}(\hat{\eta}_k^* - \tilde{\eta}_k^*) \qquad \text{(B.8)} \\ &= \mathcal{V}_k(\tilde{\eta}_k^* - \eta_k^*, \hat{\eta}_k^* - \tilde{\eta}_k^*) - \mathcal{V}_k^{(n)}(\tilde{\eta}_k^* - \eta_k^*, \hat{\eta}_k^* - \tilde{\eta}_k^*)\,(1 + o_{\mathbb{P}}(1)) \\ &= o_{\mathbb{P}}\Big(\mathcal{V}_k(\tilde{\eta}_k^* - \eta_k^*, \hat{\eta}_k^* - \tilde{\eta}_k^*)\Big) + o_{\mathbb{P}}\left(\sqrt{(\mathcal{V}_k + \lambda_k J_k)(\tilde{\eta}_k^* - \eta_k^*)(\mathcal{V}_k + \lambda_k J_k)(\hat{\eta}_k^* - \tilde{\eta}_k^*)}\right) \\ &= o_{\mathbb{P}}\left(\sqrt{(\mathcal{V}_k + \lambda_k J_k)(\tilde{\eta}_k^* - \eta_k^*)(\mathcal{V}_k + \lambda_k J_k)(\hat{\eta}_k^* - \tilde{\eta}_k^*)}\right), \end{aligned}$$

where the last equality is by $|\mathcal{V}_k(\tilde{\eta}_k^* - \eta_k^*, \hat{\eta}_k^* - \tilde{\eta}_k^*)| \leq \sqrt{\mathcal{V}_k(\tilde{\eta}_k^* - \eta_k^*)\mathcal{V}_k(\hat{\eta}_k^* - \tilde{\eta}_k^*)}$.

Combining (B.6), (B.7), and (B.8), we have

$$\begin{aligned} (\mathcal{V}_k + \lambda_k J_k)(\hat{\eta}_k^* - \tilde{\eta}_k^*) &= o_{\mathbb{P}}\left(\sqrt{(\mathcal{V}_k + \lambda_k J_k)(\tilde{\eta}_k^* - \eta_k^*)(\mathcal{V}_k + \lambda_k J_k)(\hat{\eta}_k^* - \tilde{\eta}_k^*)}\right), \\ \sqrt{(\mathcal{V}_k + \lambda_k J_k)(\hat{\eta}_k^* - \tilde{\eta}_k^*)} &= o_{\mathbb{P}}\left(\sqrt{(\mathcal{V}_k + \lambda_k J_k)(\tilde{\eta}_k^* - \eta_k^*)}\right) = o_{\mathbb{P}}\left(n^{-1/2}\lambda_k^{-1/(2r)} + \lambda_k^{b_k/2}\right), \end{aligned}$$

where $\sqrt{(\mathcal{V}_k + \lambda_k J_k)(\hat{\eta}_k^* - \eta_k^*)} \leq \sqrt{(\mathcal{V}_k + \lambda_k J_k)(\hat{\eta}_k^* - \tilde{\eta}_k^*)} + \sqrt{(\mathcal{V}_k + \lambda_k J_k)(\tilde{\eta}_k^* - \eta_k^*)}$ leads to $\sqrt{(\mathcal{V}_k + \lambda_k J_k)(\hat{\eta}_k^* - \eta_k^*)} = \mathcal{O}_{\mathbb{P}}\left(n^{-1/2}\lambda_k^{-1/(2r)} + \lambda_k^{b_k/2}\right)$.

Step 3: Rate for the Efficient Approximation Optimizer

Define $\mathcal{V}_k^{(q_k)}$ analogously to $\mathcal{V}_k^{(n)}$, with the distinction that $\mathcal{V}_k^{(q_k)}$ incorporates only q_k subjects corresponding to $\{Z_{\ell(k)}\}_{\ell=1}^{q_k}$ in its calculation, specifically for both counting process and predictable processes. Denote $\mathcal{H}^{(k)\perp} = \{f \in \mathcal{H} : \langle f, g\rangle_{\mathcal{H},k} = 0 \text{ for all } g \in \mathcal{H}^{(k)}\}$ by the orthogonal complement of $\mathcal{H}^{(k)}$ in $\mathcal{H}$ with respect to $\langle \cdot, \cdot\rangle_{\mathcal{H},k} = \langle \cdot, \cdot\rangle_{\mathcal{N}_J} + J_k(\cdot, \cdot)$, which is an inner product on $\mathcal{H}$. Here, $\langle \cdot, \cdot\rangle_{\mathcal{N}_J}$ is an inner product on $\mathcal{N}_J$, which can be found in Section B.6. Similarly, denote $\mathcal{H}_J^{(k)\perp} = \{f \in \mathcal{H}_J : J_k(f, g) = 0 \text{ for all } g \in \mathcal{H}_J^{(k)}\}$ by the orthogonal complement of $\mathcal{H}_J^{(k)}$ in $\mathcal{H}_J$ with respect to $J_k(\cdot, \cdot)$.

One can see that

$$\begin{aligned}\mathcal{H}^{(k)\perp} &= \mathcal{H} \ominus \mathcal{H}^{(k)} \\ &= \mathcal{H}_J^{(k)\perp} = \mathcal{H}_J \ominus \mathcal{H}_J^{(k)}\end{aligned}$$

and any $f \in \mathcal{H} \ominus \mathcal{H}^{(k)}$ satisfies $f(Z_{\ell(k)}) = J_k\left(f, \mathcal{K}_{J_k}(Z_{\ell(k)}, \cdot)\right) = 0$ for $\ell = 1, \ldots, q_k$. Considering this, for any $f \in \mathcal{H} \ominus \mathcal{H}^{(k)}$, a modification of Lemma B.3 with respect to $\mathcal{V}_k^{(q_k)}$ leads to

$$\begin{aligned}\mathcal{V}_k(f) = \left|\mathcal{V}_k^{(q_k)}(f) - \mathcal{V}_k(f)\right| &= (\mathcal{V}_k + \lambda_k J_k)(f)\mathcal{O}_{\mathbb{P}}\left(q_k^{-1/2}\lambda_k^{-1/r}\right) \\ &= o_{\mathbb{P}}\left((\mathcal{V}_k + \lambda_k J_k)(f)\right) = o_{\mathbb{P}}\left(\lambda_k J_k(f)\right).\end{aligned} \tag{B.9}$$

Let $\eta_k = \operatorname{Proj}(\hat{\eta}_k^* \mid \mathcal{H}^{(k)}) = \operatorname{arginf}_{\eta \in \mathcal{H}^{(k)}} \langle \hat{\eta}_k^* - \eta, \hat{\eta}_k^* - \eta \rangle_{\mathcal{H},k}$ denote the projection of $\hat{\eta}_k^*$ onto $\mathcal{H}^{(k)}$ with respect to $\langle \cdot, \cdot \rangle_{\mathcal{H},k}$, with the assumption that $\eta_k \in \mathcal{B}_k$. One can see that $\hat{\eta}_k^* - \eta_k \in \mathcal{H} \ominus \mathcal{H}^{(k)}$. Set $\eta = \hat{\eta}_k^*$ and $f = \hat{\eta}_k^* - \eta_k$ in (B.4). Considering $J_k(\eta_k, \hat{\eta}_k^* - \eta_k) = 0$ by Lemma B.7, we have

$$\lambda_k J_k(\hat{\eta}_k^* - \eta_k) = \mathcal{S}_{\eta_k^*,k}^{(n)}(\hat{\eta}_k^* - \eta_k) + \mathcal{U}_k^{(n)}(\hat{\eta}_k^* - \eta_k) - \mathcal{U}_{\hat{\eta}_k^*,k}^{(n)}(\hat{\eta}_k^* - \eta_k). \tag{B.10}$$

The order of the first term on the right-hand side of (B.10) can be calculated as

$$\begin{aligned}\left|\mathcal{S}_{\eta_k^*,k}^{(n)}(\hat{\eta}_k^* - \eta_k)\right| &= \left|\sum_{v\in\mathbb{N}} \left(\mathcal{V}_k(\hat{\eta}_k^*, \psi_{k,v}) - \mathcal{V}_k(\eta_k, \psi_{k,v})\right) \mathcal{S}_{\eta_k^*,k}^{(n)}(\psi_{k,v})\right| \\ &\leq \left\{\sum_{v\in\mathbb{N}} (1 + \lambda_k \rho_{k,v}) \left(\mathcal{V}_k(\hat{\eta}_k^*, \psi_{k,v}) - \mathcal{V}_k(\eta_k, \psi_{k,v})\right)^2\right\}^{1/2} \\ &\times \left\{\sum_{v\in\mathbb{N}} \frac{1}{1 + \lambda_k \rho_{k,v}} \left\{\mathcal{S}_{\eta_k^*,k}^{(n)}(\psi_{k,v})\right\}^2\right\}^{1/2}\end{aligned} \tag{B.11}$$

$$
\begin{aligned}
&= \sqrt{(\mathcal{V}_k + \lambda_k J_k)(\hat{\eta}_k^* - \eta_k)}\,\mathcal{O}_{\mathbb{P}}\left(n^{-1/2}\lambda_k^{-1/(2r)}\right) \\
&= \sqrt{\lambda_k J_k(\hat{\eta}_k^* - \eta_k)}\,\mathcal{O}_{\mathbb{P}}\left(n^{-1/2}\lambda_k^{-1/(2r)} + \lambda_k^{b_k/2}\right)(1 + o_{\mathbb{P}}(1)),
\end{aligned}
$$

where (B.9) is used, and the following is utilized:

$$
\mathbb{E}\left[\sum_{v\in\mathbb{N}} \frac{1}{1+\lambda_k\rho_{k,v}}\left\{\mathcal{S}^{(n)}_{\eta_k^*,k}(\psi_{k,v})\right\}^2\right] \le \frac{2\exp(\mathcal{C}_*)\mathcal{C}_\psi^2\mathcal{C}_{h,2}\tau}{n}\sum_{v\in\mathbb{N}}\frac{1}{1+\lambda_k\rho_{k,v}} = \mathcal{O}\left(n^{-1}\lambda_k^{-1/r}\right).
$$

By Lemma B.3, we have $\mathcal{V}_k^{(n)}(\hat{\eta}_k^* - \eta_k^*) = \mathcal{V}_k(\hat{\eta}_k^* - \eta_k^*) + o_{\mathbb{P}}\left((\mathcal{V}_k + \lambda_k J_k)(\hat{\eta}_k^* - \eta_k^*)\right) = o_{\mathbb{P}}(1)$. So, by (B.9) and Lemmas B.5 and B.3, the order of the remaining terms on the right-hand side of (B.10) can be calculated as

$$
\begin{aligned}
&\left|\mathcal{U}_k^{(n)}(\hat{\eta}_k^* - \eta_k) - \mathcal{U}^{(n)}_{\hat{\eta}_k^*,k}(\hat{\eta}_k^* - \eta_k)\right| = \left|\mathcal{V}_k^{(n)}(\hat{\eta}_k^* - \eta_k^*, \hat{\eta}_k^* - \eta_k)\,(1 + o_{\mathbb{P}}(1))\right| \\
&\quad= \left|\left\{\mathcal{V}_k(\hat{\eta}_k^* - \eta_k^*, \hat{\eta}_k^* - \eta_k) + o_{\mathbb{P}}\left(\sqrt{(\mathcal{V}_k + \lambda_k J_k)(\hat{\eta}_k^* - \eta_k^*)\lambda_k J_k(\hat{\eta}_k^* - \eta_k)}\right)\right\}(1 + o_{\mathbb{P}}(1))\right| \\
&\quad= o_{\mathbb{P}}\left(\sqrt{(\mathcal{V}_k + \lambda_k J_k)(\hat{\eta}_k^* - \eta_k^*)\lambda_k J_k(\hat{\eta}_k^* - \eta_k)}\right) = \sqrt{\lambda_k J_k(\hat{\eta}_k^* - \eta_k)}\,o_{\mathbb{P}}\left(n^{-1/2}\lambda_k^{-1/(2r)} + \lambda_k^{b_k/2}\right),
\end{aligned}
\tag{B.12}
$$

where

$$
\begin{aligned}
|\mathcal{V}_k(\hat{\eta}_k^* - \eta_k^*, \hat{\eta}_k^* - \eta_k)| &\le \sqrt{(\mathcal{V}_k + \lambda_k J_k)(\hat{\eta}_k^* - \eta_k^*)\mathcal{V}_k(\hat{\eta}_k^* - \eta_k)} \\
&= o_{\mathbb{P}}\left(\sqrt{(\mathcal{V}_k + \lambda_k J_k)(\hat{\eta}_k^* - \eta_k^*)\lambda_k J_k(\hat{\eta}_k^* - \eta_k)}\right).
\end{aligned}
$$

Combining (B.10), (B.11), and (B.12) and considering (B.9), we get

$$
\begin{aligned}
\lambda_k J_k(\hat{\eta}_k^* - \eta_k) &= \sqrt{\lambda_k J_k(\hat{\eta}_k^* - \eta_k)}\,\mathcal{O}_{\mathbb{P}}\left(n^{-1/2}\lambda_k^{-1/(2r)} + \lambda_k^{b_k/2}\right), \\
(\mathcal{V}_k + \lambda_k J_k)(\hat{\eta}_k^* - \eta_k) &= \mathcal{O}_{\mathbb{P}}\left(\left\{n^{-1/2}\lambda_k^{-1/(2r)} + \lambda_k^{b_k/2}\right\}^2\right),
\end{aligned}
$$

where $\sqrt{(\mathcal{V}_k+\lambda_k J_k)(\eta_k-\eta_k^*)} \le \sqrt{(\mathcal{V}_k+\lambda_k J_k)(\eta_k-\hat{\eta}_k^*)} + \sqrt{(\mathcal{V}_k+\lambda_k J_k)(\hat{\eta}_k^*-\eta_k^*)}$ leads to $\sqrt{(\mathcal{V}_k+\lambda_k J_k)(\eta_k-\eta_k^*)} = \mathcal{O}_{\mathbb{P}}\left(n^{-1/2}\lambda_k^{-1/(2r)}+\lambda_k^{b_k/2}\right)$.

As $J_k(\eta_k,\hat{\eta}_k^*-\eta_k)=J_k(\hat{\eta}_k,\hat{\eta}_k^*-\eta_k)=0$ by Lemma B.7, we get

$$\begin{aligned} J_k(\hat{\eta}_k^*,\hat{\eta}_k^*-\hat{\eta}_k) &= J_k(\hat{\eta}_k^*-\eta_k+\eta_k,\hat{\eta}_k^*-\eta_k+\eta_k-\hat{\eta}_k) \\ &= J_k(\hat{\eta}_k^*-\eta_k)+J_k(\eta_k,\eta_k-\hat{\eta}_k). \end{aligned} \tag{B.13}$$

By definition, $A(\hat{\eta}_k,f)=0$ for any $f\in\mathcal{H}^{(k)}$. Set $\eta=\hat{\eta}_k$ and $f=\hat{\eta}_k-\eta_k$ in (B.4) and set $\eta=\hat{\eta}_k^*$ and $f=\hat{\eta}_k^*-\hat{\eta}_k$ in (B.4). Using (B.13) and (B.10), summing the resulting equations gives

$$\begin{aligned} &\mathcal{U}^{(n)}_{\hat{\eta}_k,k}(\hat{\eta}_k-\eta_k)-\mathcal{U}^{(n)}_{\hat{\eta}_k^*,k}(\hat{\eta}_k-\eta_k)+\lambda_k J_k(\hat{\eta}_k-\eta_k) \\ &\quad = \mathcal{S}^{(n)}_{\eta_k^*,k}(\hat{\eta}_k^*-\eta_k)-\lambda_k J_k(\hat{\eta}_k^*-\eta_k)+\mathcal{U}^{(n)}_k(\hat{\eta}_k^*-\eta_k)-\mathcal{U}^{(n)}_{\hat{\eta}_k^*,k}(\hat{\eta}_k^*-\eta_k)=0, \end{aligned}$$

which can be represented as

$$\mathcal{U}^{(n)}_{\hat{\eta}_k,k}(\hat{\eta}_k-\eta_k)-\mathcal{U}^{(n)}_{\eta_k,k}(\hat{\eta}_k-\eta_k)+\lambda_k J_k(\hat{\eta}_k-\eta_k)=\mathcal{U}^{(n)}_{\hat{\eta}_k^*,k}(\hat{\eta}_k-\eta_k)-\mathcal{U}^{(n)}_{\eta_k,k}(\hat{\eta}_k-\eta_k). \tag{B.14}$$

By Lemma B.4, for some random variable $\alpha_2\in(0,1)$, we have

$$\begin{aligned} 0 \le\ & \mathcal{U}^{(n)}_{\hat{\eta}_k,k}(\hat{\eta}_k-\eta_k)-\mathcal{U}^{(n)}_{\eta_k,k}(\hat{\eta}_k-\eta_k)+\lambda_k J_k(\hat{\eta}_k-\eta_k) \\ &= \mathcal{V}^{(n)}_{\alpha_2\hat{\eta}_k+(1-\alpha_2)\eta_k,k}(\hat{\eta}_k-\eta_k)+\lambda_k J_k(\hat{\eta}_k-\eta_k) \\ &\ge \mathcal{C}_{\mathcal{B}_k}(\mathcal{V}^{(n)}_k+\lambda_k J_k)(\hat{\eta}_k-\eta_k) \\ &= \mathcal{C}_{\mathcal{B}_k}(\mathcal{V}_k+\lambda_k J_k)(\hat{\eta}_k-\eta_k)+o_{\mathbb{P}}\left((\mathcal{V}_k+\lambda_k J_k)(\hat{\eta}_k-\eta_k)\right), \end{aligned} \tag{B.15}$$

where $\alpha_2\hat{\eta}_k + (1-\alpha_2)\eta_k \in \mathcal{B}_k$.

We have $\mathcal{V}_k^{(n)}(\hat{\eta}_k^* - \eta_k^*) = o_{\mathbb{P}}(1)$ and by Lemma B.3, we have $\mathcal{V}_k^{(n)}(\eta_k - \eta_k^*) = \mathcal{V}_k(\eta_k - \eta_k^*) + o_{\mathbb{P}}\left((\mathcal{V}_k + \lambda_k J_k)(\eta_k - \eta_k^*)\right) = o_{\mathbb{P}}(1)$. So, by Lemmas B.5 and B.3, we get

$$\begin{aligned} 0 \leq\ & \mathcal{U}^{(n)}_{\hat{\eta}_k^*,k}(\hat{\eta}_k - \eta_k) - \mathcal{U}^{(n)}_{\eta_k,k}(\hat{\eta}_k - \eta_k) = \mathcal{V}_k^{(n)}(\hat{\eta}_k^* - \eta_k, \hat{\eta}_k - \eta_k)\left(1 + o_{\mathbb{P}}(1)\right) \\ & = \mathcal{V}_k(\hat{\eta}_k^* - \eta_k, \hat{\eta}_k - \eta_k) + o_{\mathbb{P}}\left(\mathcal{V}_k(\hat{\eta}_k^* - \eta_k, \hat{\eta}_k - \eta_k)\right) \\ & + o_{\mathbb{P}}\left(\sqrt{(\mathcal{V}_k + \lambda_k J_k)(\hat{\eta}_k^* - \eta_k)(\mathcal{V}_k + \lambda_k J_k)(\hat{\eta}_k - \eta_k)}\right) \\ & = \mathcal{O}_{\mathbb{P}}\left(\sqrt{(\mathcal{V}_k + \lambda_k J_k)(\hat{\eta}_k^* - \eta_k)(\mathcal{V}_k + \lambda_k J_k)(\hat{\eta}_k - \eta_k)}\right), \end{aligned} \tag{B.16}$$

where $|\mathcal{V}_k(\hat{\eta}_k^* - \eta_k, \hat{\eta}_k - \eta_k)| \leq \sqrt{(\mathcal{V}_k + \lambda_k J_k)(\hat{\eta}_k^* - \eta_k)(\mathcal{V}_k + \lambda_k J_k)(\hat{\eta}_k - \eta_k)}$.

Combining (B.14), (B.15), and (B.16), we have

$$\begin{aligned} & (\mathcal{V}_k + \lambda_k J_k)(\hat{\eta}_k - \eta_k) = \mathcal{O}_{\mathbb{P}}\left(\sqrt{(\mathcal{V}_k + \lambda_k J_k)(\hat{\eta}_k^* - \eta_k)(\mathcal{V}_k + \lambda_k J_k)(\hat{\eta}_k - \eta_k)}\right), \\ & \sqrt{(\mathcal{V}_k + \lambda_k J_k)(\hat{\eta}_k - \eta_k)} = \mathcal{O}_{\mathbb{P}}\left(\sqrt{(\mathcal{V}_k + \lambda_k J_k)(\hat{\eta}_k^* - \eta_k)}\right) = \mathcal{O}_{\mathbb{P}}\left(n^{-1/2}\lambda_k^{-1/(2r)} + \lambda_k^{b_k/2}\right), \end{aligned}$$

where $\sqrt{(\mathcal{V}_k + \lambda_k J_k)(\hat{\eta}_k - \eta_k^*)} \leq \sqrt{(\mathcal{V}_k + \lambda_k J_k)(\hat{\eta}_k - \eta_k)} + \sqrt{(\mathcal{V}_k + \lambda_k J_k)(\eta_k - \eta_k^*)}$ leads to $\sqrt{(\mathcal{V}_k + \lambda_k J_k)(\hat{\eta}_k - \eta_k^*)} = \mathcal{O}_{\mathbb{P}}\left(n^{-1/2}\lambda_k^{-1/(2r)} + \lambda_k^{b_k/2}\right)$.

Note that $n^{-1/2}\lambda_k^{-1/(2r)} \ll n^{-1/4}\lambda_k^{-1/(2r)} \ll q_k^{-1/4}\lambda_k^{-1/(2r)} = o(1)$ implies

$$\mathcal{O}_{\mathbb{P}}\left(n^{-1/2}\lambda_k^{-1/(2r)} + \lambda_k^{b_k/2}\right) = o_{\mathbb{P}}(1),$$

which completes the proof. ■

B.5 Technical Details for Theorem 3.1

Proof of Lemma B.2. Here, we consider $t \in [0, \tau]$ and $f \in \mathcal{H}$, where f may represent either a deterministic function or a stochastic process. Let $\mathsf{f}_k(t,x) \equiv \exp(\eta_k^*(x))\mathsf{f}_{\mathcal{Y},\mathcal{X}}(t,x)/S(\eta_k^*,t)$. Then, we have

$$\begin{aligned}\int_{\mathcal{X}} \mathsf{f}_k(t,x)dx &= \frac{\int_{\mathcal{X}} \mathbb{E}[Y(t)|X=x]\exp(\eta_k^*(x))\mathsf{f}_{\mathcal{X}}(x)dx}{S(\eta_k^*,t)} \\ &= \frac{\mathbb{E}\left[\mathbb{E}\left[Y(t)\exp(\eta_k^*(X))|X\right]\right]}{S(\eta_k^*,t)} = \frac{S(\eta_k^*,t)}{S(\eta_k^*,t)} = 1\end{aligned}$$

and

$$\begin{aligned}\mu_k(f)(t) \equiv \frac{S(f;\eta_k^*,t)}{S(\eta_k^*,t)} &= \frac{\mathbb{E}_W\left[\mathbb{E}_W\left[Y(t)|X\right]f(X)\exp(\eta_k^*(X))\right]}{S(\eta_k^*,t)} \\ &= \frac{\mathbb{E}_W\left[\mathbb{E}\left[Y(t)|X\right]f(X)\exp(\eta_k^*(X))\right]}{S(\eta_k^*,t)} \\ &= \frac{\int_{\mathcal{X}} \mathsf{f}_{\mathcal{Y}|\mathcal{X}}(t;x)f(x)\exp(\eta_k^*(x))\mathsf{f}_{\mathcal{X}}(x)dx}{S(\eta_k^*,t)} = \int_{\mathcal{X}} f(x)\mathsf{f}_k(t,x)dx.\end{aligned}$$

Using above, we have

$$\begin{aligned}\mathcal{V}_k(f) &= \int_0^\tau \left\{\int_{\mathcal{X}} \Big(f(x)-\mu_k(f)(t)\Big)^2 \mathsf{f}_k(t,x)dx\right\} S(\eta_k^*,t)dH_{0k}(t) \qquad \text{(B.17)} \\ &= \int_0^\tau \left\{\int_{\mathcal{X}} \Big(f(x)-\mu_k(f)(t)\Big)^2 \exp(\eta_k^*(x))\mathsf{f}_{\mathcal{Y},\mathcal{X}}(t,x)dx\right\} dH_{0k}(t) \\ &\in \Bigg[\exp(-\mathcal{C}_*)\mathcal{C}_{\mathsf{f},1}\int_0^\tau \left\{\int_{\mathcal{X}} \Big(f(x)-\mu_k(f)(t)\Big)^2 dx\right\} dH_{0k}(t), \\ &\qquad \exp(\mathcal{C}_*)\mathcal{C}_{\mathsf{f},2}\int_0^\tau \left\{\int_{\mathcal{X}} \Big(f(x)-\mu_k(f)(t)\Big)^2 dx\right\} dH_{0k}(t)\Bigg].\end{aligned}$$

Let $\mathfrak{m}(\cdot)$ be the Lebesgue measure on $\mathcal{X}$, Then,

$$\int_0^\tau \left\{ \int_{\mathcal{X}} \Big(f(x) - \mu_k(f)(t) \Big)^2 dx \right\} dH_{0k}(t) = H_{0k}(\tau) \|f\|_{\mathcal{L}_2}^2 + \int_0^\tau \Big(\mu_k(f)(t) \Big)^2 dH_{0k}(t),$$

where we used the fact that $\int_{\mathcal{X}} f(x)dx = 0$ due to $f \in \mathcal{H}$ and $\mathfrak{m}(\mathcal{X}) = \mathfrak{m}([0,1]^d) = 1$. By the Cauchy-Schwarz inequality, we have

$$0 \leq \Big(\mu_k(f)(t) \Big)^2 \leq \int_{\mathcal{X}} (f(x))^2 \, dx \int_{\mathcal{X}} (\mathsf{f}_k(t,x))^2 \, dx = \|f\|_{\mathcal{L}_2}^2 \int_{\mathcal{X}} (\mathsf{f}_k(t,x))^2 \, dx,$$

which leads to

$$\begin{aligned} \int_0^\tau \Big(\mu_k(f)(t) \Big)^2 dH_{0k}(t) &\leq \|f\|_{\mathcal{L}_2}^2 \int_0^\tau \left\{ \int_{\mathcal{X}} (\mathsf{f}_k(t,x))^2 \, dx \right\} dH_{0k}(t) \\ &\leq \frac{\exp(2\mathcal{C}_*)\mathcal{C}_{\mathsf{f},2}^2 \mathcal{C}_{h,2}}{\mathcal{C}_S^2} \tau \|f\|_{\mathcal{L}_2}^2 = \tilde{\mathcal{C}} \tau \|f\|_{\mathcal{L}_2}^2 \end{aligned}$$

and

$$\mathcal{C}_{h,1}\tau\|f\|_{\mathcal{L}_2}^2 \leq H_{0k}(\tau)\|f\|_{\mathcal{L}_2}^2 \leq \int_0^\tau \left\{ \int_{\mathcal{X}} \Big(f(x) - \mu_k(f)(t) \Big)^2 dx \right\} dH_{0k}(t) \leq (\mathcal{C}_{h,2} + \tilde{\mathcal{C}})\tau\|f\|_{\mathcal{L}_2}^2. \tag{B.18}$$

Using (B.17) and (B.18), we have

$$\mathcal{C}_{\mathcal{L}_2,1}\|f\|_{\mathcal{L}_2}^2 \leq \mathcal{V}_k(f) \leq \mathcal{C}_{\mathcal{L}_2,2}\|f\|_{\mathcal{L}_2}^2,$$

where $\mathcal{C}_{\mathcal{L}_2,1} = \exp(-\mathcal{C}_*)\mathcal{C}_{\mathsf{f},1}\mathcal{C}_{h,1}\tau$ and $\mathcal{C}_{\mathcal{L}_2,2} = \exp(\mathcal{C}_*)\mathcal{C}_{\mathsf{f},2}(\mathcal{C}_{h,2} + \tilde{\mathcal{C}})\tau$. ■

Lemma B.6. *Under Assumption B.3.1, for each k, the following results on boundedness* *hold:*

$$\sup_{t\in[0,\tau]} S(\eta_k^*,t) \le \exp(\mathcal{C}_*), \quad \sup_{t\in[0,\tau]} S_{(n)}(\eta_k^*,t) \le \exp(\mathcal{C}_*),$$
$$\sup_{v\in\mathbb{N},t\in[0,\tau]} |S(\psi_{k,v};\eta_k^*,t)| \le \exp(\mathcal{C}_*)\mathcal{C}_\psi, \quad \sup_{v\in\mathbb{N},t\in[0,\tau]} \left|S_{(n)}(\psi_{k,v};\eta_k^*,t)\right| \le \exp(\mathcal{C}_*)\mathcal{C}_\psi,$$
$$\sup_{v,v'\in\mathbb{N},t\in[0,\tau]} |S(\psi_{k,v}\psi_{k,v'};\eta_k^*,t)| \le \exp(\mathcal{C}_*)\mathcal{C}_\psi^2, \quad \sup_{v,v'\in\mathbb{N},t\in[0,\tau]} \left|S_{(n)}(\psi_{k,v}\psi_{k,v'};\eta_k^*,t)\right| \le \exp(\mathcal{C}_*)\mathcal{C}_\psi^2,$$

$$\sup_{v\in\mathbb{N},t\in[0,\tau]} |\mathcal{U}(\psi_{k,v};\eta_k^*,t)| \le \mathcal{C}_\psi, \quad \sup_{v\in\mathbb{N},t\in[0,\tau]} \left|\mathcal{U}_{(n)}(\psi_{k,v};\eta_k^*,t)\right| \le \mathcal{C}_\psi,$$
$$\sup_{v,v'\in\mathbb{N},t\in[0,\tau]} |\mathcal{V}(\psi_{k,v},\psi_{k,v'};\eta_k^*,t)| \le 2\mathcal{C}_\psi^2, \quad \sup_{v,v'\in\mathbb{N},t\in[0,\tau]} \left|\mathcal{V}_{(n)}(\psi_{k,v},\psi_{k,v'};\eta_k^*,t)\right| \le 2\mathcal{C}_\psi^2.$$

Furthermore, for any $v, v' \in \mathbb{N}$, the following results hold:

$$\begin{aligned}
&\mathbb{E}\left[\int_0^\tau \frac{1}{S_{(n)}(\eta_k^*,t)} d\overline{N}_k(t)\right] \le \mathcal{C}_{h,2}\tau,\\
&\mathbb{E}\left[\int_0^\tau \frac{\left(S_{(n)}(\eta_k^*,t) - S(\eta_k^*,t)\right)^2}{S_{(n)}(\eta_k^*,t)} d\overline{N}_k(t)\right] \le \frac{\exp(2\mathcal{C}_*)\mathcal{C}_{h,2}\tau}{n},\\
&\mathbb{E}\left[\int_0^\tau \frac{\left(S_{(n)}(\psi_{k,v};\eta_k^*,t) - S(\psi_{k,v};\eta_k^*,t)\right)^2}{S_{(n)}(\eta_k^*,t)} d\overline{N}_k(t)\right] \le \frac{\exp(2\mathcal{C}_*)\mathcal{C}_\psi^2\mathcal{C}_{h,2}\tau}{n},\\
&\mathbb{E}\left[\int_0^\tau \frac{\left(S_{(n)}(\psi_{k,v}\psi_{k,v'};\eta_k^*,t) - S(\psi_{k,v}\psi_{k,v'};\eta_k^*,t)\right)^2}{S_{(n)}(\eta_k^*,t)} d\overline{N}_k(t)\right] \le \frac{\exp(2\mathcal{C}_*)\mathcal{C}_\psi^4\mathcal{C}_{h,2}\tau}{n}.
\end{aligned}$$

Proof of Lemma B.6. Considering $Y(\cdot), Y_i(\cdot) \le 1$ and boundedness of η_k^*, we have

$$\sup_{t\in[0,\tau]} S(\eta_k^*,t) = \sup_{t\in[0,\tau]} \mathbb{E}\left[Y(t)\exp(\eta_k^*(X))\right] \le \left\{\sup_{t\in[0,\tau]} \mathbb{E}\left[Y(t)\right]\right\}\exp(\mathcal{C}_*) \le \exp(\mathcal{C}_*),$$
$$\sup_{t\in[0,\tau]} S_{(n)}(\eta_k^*,t) = \sup_{t\in[0,\tau]} \frac{1}{n}\sum_{i=1}^n Y_i(t)\exp(\eta_k^*(X_i)) \le \left\{\sup_{t\in[0,\tau]} \frac{1}{n}\sum_{i=1}^n Y_i(t)\right\}\exp(\mathcal{C}_*) \le \exp(\mathcal{C}_*).$$

Regarding boundedness of eigenfunctions, we have

$$\begin{aligned}
\sup_{v,v'\in\mathbb{N},t\in[0,\tau]} |S(\psi_{k,v}\psi_{k,v'};\eta_k^*,t)| &= \sup_{v,v'\in\mathbb{N},t\in[0,\tau]} \left|\mathbb{E}\big[Y(t)\psi_{k,v}(X)\psi_{k,v'}(X)\exp(\eta_k^*(X))\big]\right| \\
&\le \sup_{v,v'\in\mathbb{N},t\in[0,\tau]} \mathbb{E}\Big[Y(t)\big|\psi_{k,v}(X)\big|\big|\psi_{k,v'}(X)\big|\exp(\eta_k^*(X))\Big] \\
&\le \exp(\mathcal{C}_*)\mathcal{C}_\psi^2, \\
\sup_{v,v'\in\mathbb{N},t\in[0,\tau]} \big|S_{(n)}(\psi_{k,v}\psi_{k,v'};\eta_k^*,t)\big| &= \sup_{v,v'\in\mathbb{N},t\in[0,\tau]} \left|\frac{1}{n}\sum_{i=1}^n Y_i(t)\psi_{k,v}(X_i)\psi_{k,v'}(X_i)\exp(\eta_k^*(X_i))\right| \\
&\le \sup_{v,v'\in\mathbb{N},t\in[0,\tau]} \frac{1}{n}\sum_{i=1}^n Y_i(t)\big|\psi_{k,v}(X_i)\big|\big|\psi_{k,v'}(X_i)\big|\exp(\eta_k^*(X_i)) \\
&\le \exp(\mathcal{C}_*)\mathcal{C}_\psi^2,
\end{aligned}$$

where boundedness for $\sup_{v\in\mathbb{N},t\in[0,\tau]} |S(\psi_{k,v};\eta_k^*,t)|$ and $\sup_{v\in\mathbb{N},t\in[0,\tau]} \big|S_{(n)}(\psi_{k,v};\eta_k^*,t)\big|$ can be similarly proved.

$$\begin{aligned}
\sup_{v,v'\in\mathbb{N},t\in[0,\tau]} \left|\frac{S(\psi_{k,v}\psi_{k,v'};\eta_k^*,t)}{S(\eta_k^*,t)}\right| &\le \sup_{v,v'\in\mathbb{N},t\in[0,\tau]} \frac{\mathbb{E}\big[Y(t)\big|\psi_{k,v}(X)\big|\big|\psi_{k,v'}(X)\big|\exp(\eta_k^*(X))\big]}{\mathbb{E}[Y(t)\exp(\eta_k^*(X))]} \\
&\le \left\{\sup_{v,v'\in\mathbb{N},t\in[0,\tau]} \frac{\mathbb{E}[Y(t)\exp(\eta_k^*(X))]}{\mathbb{E}[Y(t)\exp(\eta_k^*(X))]}\right\}\mathcal{C}_\psi^2 = \mathcal{C}_\psi^2, \\
\sup_{v,v'\in\mathbb{N},t\in[0,\tau]} \left|\frac{S_{(n)}(\psi_{k,v}\psi_{k,v'};\eta_k^*,t)}{S_{(n)}(\eta_k^*,t)}\right| &\le \sup_{v,v'\in\mathbb{N},t\in[0,\tau]} \frac{\frac{1}{n}\sum_{i=1}^n Y_i(t)\big|\psi_{k,v}(X_i)\big|\big|\psi_{k,v'}(X_i)\big|\exp(\eta_k^*(X_i))}{\frac{1}{n}\sum_{i=1}^n Y_i(t)\exp(\eta_k^*(X_i))} \\
&\le \left\{\sup_{v,v'\in\mathbb{N},t\in[0,\tau]} \frac{\frac{1}{n}\sum_{i=1}^n Y_i(t)\exp(\eta_k^*(X_i))}{\frac{1}{n}\sum_{i=1}^n Y_i(t)\exp(\eta_k^*(X_i))}\right\}\mathcal{C}_\psi^2 = \mathcal{C}_\psi^2,
\end{aligned}$$

where boundedness for $\sup_{v\in\mathbb{N},t\in[0,\tau]} |\mathcal{U}(\psi_{k,v};\eta_k^*,t)|$ and $\sup_{v\in\mathbb{N},t\in[0,\tau]} \big|\mathcal{U}_{(n)}(\psi_{k,v};\eta_k^*,t)\big|$ can be similarly proved. Combining above results,

$$\begin{aligned}
\sup_{v,v'\in\mathbb{N},t\in[0,\tau]}|\mathcal{V}(\psi_{k,v},\psi_{k,v'};\eta_k^*,t)| &\leq \sup_{v,v'\in\mathbb{N},t\in[0,\tau]}\left|\frac{S(\psi_{k,v}\psi_{k,v'};\eta_k^*,t)}{S(\eta_k^*,t)}\right| \\
&\quad + \left\{\sup_{v\in\mathbb{N},t\in[0,\tau]}|\mathcal{U}(\psi_{k,v};\eta_k^*,t)|\right\}^2 \leq 2\mathcal{C}_\psi^2, \\
\sup_{v,v'\in\mathbb{N},t\in[0,\tau]}\left|\mathcal{V}_{(n)}(\psi_{k,v},\psi_{k,v'};\eta_k^*,t)\right| &\leq \sup_{v,v'\in\mathbb{N},t\in[0,\tau]}\left|\frac{S_{(n)}(\psi_{k,v}\psi_{k,v'};\eta_k^*,t)}{S_{(n)}(\eta_k^*,t)}\right| \\
&\quad + \left\{\sup_{v\in\mathbb{N},t\in[0,\tau]}\left|\mathcal{U}_{(n)}(\psi_{k,v};\eta_k^*,t)\right|\right\}^2 \leq 2\mathcal{C}_\psi^2,
\end{aligned}$$

We show boundedness of moment calculations, which are

$$\begin{aligned}
\mathbb{E}\left[\int_0^\tau \frac{1}{S_{(n)}(\eta_k^*,t)}d\overline{N}_k(t)\right] &= \int_0^\tau \mathbb{E}\left[\frac{1}{S_{(n)}(\eta_k^*,t)}d\overline{N}_k(t)\right] \\
&= \int_0^\tau \mathbb{E}\left[\mathbb{E}\left[\frac{1}{S_{(n)}(\eta_k^*,t)}d\overline{N}_k(t)\middle|\mathcal{F}_{t^-}\right]\right] \\
&= \int_0^\tau \mathbb{E}\left[\frac{1}{S_{(n)}(\eta_k^*,t)}\mathbb{E}\left[d\overline{N}_k(t)|\mathcal{F}_{t^-}\right]\right] \\
&= \int_0^\tau \mathbb{E}\left[\frac{S_{(n)}(\eta_k^*,t)}{S_{(n)}(\eta_k^*,t)}dH_{0k}(t)\right] = H_{0k}(\tau) \leq \mathcal{C}_{h,2}\tau,
\end{aligned}$$

$$\begin{aligned}
&\mathbb{E}\left[\int_0^\tau \frac{\left(S_{(n)}(\psi_{k,v}\psi_{k,v'};\eta_k^*,t) - S(\psi_{k,v}\psi_{k,v'};\eta_k^*,t)\right)^2}{S_{(n)}(\eta_k^*,t)}d\overline{N}_k(t)\right] \\
&\quad = \int_0^\tau \mathbb{E}\left[\mathbb{E}\left[\frac{\left(S_{(n)}(\psi_{k,v}\psi_{k,v'};\eta_k^*,t) - S(\psi_{k,v}\psi_{k,v'};\eta_k^*,t)\right)^2}{S_{(n)}(\eta_k^*,t)}d\overline{N}_k(t)\middle|\mathcal{F}_{t^-}\right]\right] \\
&\quad = \int_0^\tau \mathbb{E}\left[\frac{\left(S_{(n)}(\psi_{k,v}\psi_{k,v'};\eta_k^*,t) - S(\psi_{k,v}\psi_{k,v'};\eta_k^*,t)\right)^2}{S_{(n)}(\eta_k^*,t)}\mathbb{E}\left[d\overline{N}_k(t)|\mathcal{F}_{t^-}\right]\right] \\
&\quad = \int_0^\tau \mathbb{E}\left[\left(S_{(n)}(\psi_{k,v}\psi_{k,v'};\eta_k^*,t) - S(\psi_{k,v}\psi_{k,v'};\eta_k^*,t)\right)^2\right]dH_{0k}(t) \\
&\quad = \frac{1}{n}\int_0^\tau \mathrm{Var}\Big[Y(t)\psi_{k,v}(X)\psi_{k,v'}(X)\exp(\eta_k^*(X))\Big]dH_{0k}(t)
\end{aligned}$$

$$
\begin{aligned}
&\leq \frac{1}{n}\int_0^\tau \mathbb{E}\Big[\big(Y(t)\psi_{k,v}(X)\psi_{k,v'}(X)\exp(\eta_k^*(X))\big)^2\Big]dH_{0k}(t)\\
&\leq \frac{1}{n}\mathcal{C}_\psi^2\mathcal{C}_\psi^2\exp(2\mathcal{C}_*)H_{0k}(\tau) \leq \frac{\exp(2\mathcal{C}_*)\mathcal{C}_\psi^4\mathcal{C}_{h,2}\tau}{n},
\end{aligned}
$$

where $\mathbb{E}_W\left[Y(t)\psi_{k,v}(X)\psi_{k,v'}(X)\exp(\eta_k^*(X))\right] = \mathbb{E}\left[Y(t)\psi_{k,v}(X)\psi_{k,v'}(X)\exp(\eta_k^*(X))\right]$. Similar derivation yields

$$
\begin{aligned}
\mathbb{E}\left[\int_0^\tau \frac{\big(S_{(n)}(\psi_{k,v};\eta_k^*,t) - S(\psi_{k,v};\eta_k^*,t)\big)^2}{S_{(n)}(\eta_k^*,t)}d\overline{N}_k(t)\right] &= \frac{1}{n}\int_0^\tau \mathrm{Var}\Big[Y(t)\psi_{k,v}(X)\exp(\eta_k^*(X))\Big]dH_{0k}(t)\\
&\leq \frac{1}{n}\int_0^\tau \mathbb{E}\Big[\big(Y(t)\psi_{k,v}(X)\exp(\eta_k^*(X))\big)^2\Big]dH_{0k}(t)\\
&\leq \frac{\exp(2\mathcal{C}_*)\mathcal{C}_\psi^2\mathcal{C}_{h,2}\tau}{n},
\end{aligned}
$$

$$
\begin{aligned}
\mathbb{E}\left[\int_0^\tau \frac{\big(S_{(n)}(\eta_k^*,t) - S(\eta_k^*,t)\big)^2}{S_{(n)}(\eta_k^*,t)}d\overline{N}_k(t)\right] &= \frac{1}{n}\int_0^\tau \mathrm{Var}\Big[Y(t)\exp(\eta_k^*(X))\Big]dH_{0k}(t)\\
&\leq \frac{1}{n}\int_0^\tau \mathbb{E}\Big[\big(Y(t)\exp(\eta_k^*(X))\big)^2\Big]dH_{0k}(t) \leq \frac{\exp(2\mathcal{C}_*)\mathcal{C}_{h,2}\tau}{n}.
\end{aligned}
$$

■

Proof of Lemma B.3. Here, each $f \in \mathcal{H}$ or $g \in \mathcal{H}$ can represent either a deterministic function or a stochastic process. We use the fact that $\mathbb{E}(|X_n|^b) = \mathcal{O}(a_n^b)$ implies $X_n = \mathcal{O}_{\mathbb{P}}(a_n)$, and $\mathbb{E}(|X_n|^b) = o(a_n^b)$ implies $X_n = o_{\mathbb{P}}(a_n)$, where $b \in (0,\infty)$ is a constant and for $n \in \mathbb{N}$, $X_n \in \mathbb{R}$ is a random sequence and $a_n \in (0,\infty)$ is a deterministic sequence. Note that Lemma B.3 and its proof are similar to the result in O'Sullivan [38].

The upper bound of the difference between $\mathcal{V}_k^{(n)}(f,g)$ and $\mathcal{V}_k(f,g)$ is given by

$$\left|\mathcal{V}_k^{(n)}(f,g)-\mathcal{V}_k(f,g)\right| \leq \left|\int_0^\tau \left\{\mathcal{V}_{(n)}(f,g;\eta_k^*,t)-\mathcal{V}(f,g;\eta_k^*,t)\right\} d\overline{N}_k(t)\right| + \left|\int_0^\tau \mathcal{V}(f,g;\eta_k^*,t)d\overline{N}_k(t)-\mathcal{V}_k(f,g)\right|, \tag{B.19}$$

where we demonstrate the asymptotic order bounds for the first and second components on the right-hand side as follows.

For the second component in (B.19), we have

$$\begin{aligned}
&\left|\int_0^\tau \mathcal{V}(f,g;\eta_k^*,t)d\overline{N}_k(t)-\mathcal{V}_k(f,g)\right| \\
&=\left|\sum_{v\in\mathbb{N}}\sum_{v'\in\mathbb{N}}\mathcal{V}_k(f,\psi_{k,v})\mathcal{V}_k(g,\psi_{k,v'})\left\{\int_0^\tau \mathcal{V}(\psi_{k,v},\psi_{k,v'};\eta_k^*,t)d\overline{N}_k(t)-\mathcal{V}_k(\psi_{k,v},\psi_{k,v'})\right\}\right| \\
&\leq\left\{\sum_{v\in\mathbb{N}}\sum_{v'\in\mathbb{N}}(1+\lambda_k\rho_{k,v})(1+\lambda_k\rho_{k,v'})\mathcal{V}_k^2(f,\psi_{k,v})\mathcal{V}_k^2(g,\psi_{k,v'})\right\}^{1/2} \\
&\times\left\{\sum_{v\in\mathbb{N}}\sum_{v'\in\mathbb{N}}\frac{1}{1+\lambda_k\rho_{k,v}}\frac{1}{1+\lambda_k\rho_{k,v'}}\left\{\int_0^\tau \mathcal{V}(\psi_{k,v},\psi_{k,v'};\eta_k^*,t)d\overline{N}_k(t)-\mathcal{V}_k(\psi_{k,v},\psi_{k,v'})\right\}^2\right\}^{1/2} \\
&=\sqrt{(\mathcal{V}_k+\lambda_kJ_k)(f)(\mathcal{V}_k+\lambda_kJ_k)(g)}\,\mathcal{O}_{\mathbb{P}}\left(n^{-1/2}\lambda_k^{-1/r}\right),
\end{aligned} \tag{B.20}$$

where the last equality is obtained by

$$\begin{aligned}
&\mathbb{E}\left|\sum_{v\in\mathbb{N}}\sum_{v'\in\mathbb{N}}\frac{1}{1+\lambda_k\rho_{k,v}}\frac{1}{1+\lambda_k\rho_{k,v'}}\left\{\int_0^\tau \mathcal{V}(\psi_{k,v},\psi_{k,v'};\eta_k^*,t)d\overline{N}_k(t)-\mathcal{V}_k(\psi_{k,v},\psi_{k,v'})\right\}^2\right| \\
&=\mathbb{E}\left[\sum_{v\in\mathbb{N}}\sum_{v'\in\mathbb{N}}\frac{1}{1+\lambda_k\rho_{k,v}}\frac{1}{1+\lambda_k\rho_{k,v'}}\left\{\int_0^\tau \mathcal{V}(\psi_{k,v},\psi_{k,v'};\eta_k^*,t)d\overline{N}_k(t)-\mathcal{V}_k(\psi_{k,v},\psi_{k,v'})\right\}^2\right] \\
&\leq\left\{\sum_{v\in\mathbb{N}}\sum_{v'\in\mathbb{N}}\frac{1}{1+\lambda_k\rho_{k,v}}\frac{1}{1+\lambda_k\rho_{k,v'}}\right\}\frac{4\mathcal{C}_\psi^4}{n}=\mathcal{O}\left(\lambda_k^{-1/r}\right)\mathcal{O}\left(\lambda_k^{-1/r}\right)\frac{4\mathcal{C}_\psi^4}{n}=\mathcal{O}\left(n^{-1}\lambda_k^{-2/r}\right).
\end{aligned}$$

The last inequality above is obtained by

$$\mathbb{E}\left[\left\{\frac{1}{n}\sum_{i=1}^{n}\left\{\int_0^\tau \mathcal{V}(\psi_{k,v},\psi_{k,v'};\eta_k^*,t)dN_{ik}(t)-\mathcal{V}_k(\psi_{k,v},\psi_{k,v'})\right\}\right\}^2\right]$$
$$=\frac{1}{n}\mathrm{Var}\left[\int_0^\tau \mathcal{V}(\psi_{k,v},\psi_{k,v'};\eta_k^*,t)dN_k(t)\right]\le\frac{1}{n}\mathbb{E}\left[\left\{\int_0^\tau \mathcal{V}(\psi_{k,v},\psi_{k,v'};\eta_k^*,t)dN_k(t)\right\}^2\right]\le\frac{4\mathcal{C}_\psi^4}{n},$$

where

$$\begin{aligned}\mathbb{E}\left[\int_0^\tau \mathcal{V}(\psi_{k,v},\psi_{k,v'};\eta_k^*,t)dN_{ik}(t)\right]&=\int_0^\tau \mathcal{V}(\psi_{k,v},\psi_{k,v'};\eta_k^*,t)\mathbb{E}\big[\mathbb{E}\left[dN_{ik}(t)|\mathcal{F}_{t-}\right]\big]\\&=\int_0^\tau \mathcal{V}(\psi_{k,v},\psi_{k,v'};\eta_k^*,t)\mathbb{E}\left[Y_i(t)\exp(\eta_k^*(X_i))\right]dH_{0k}(t)\\&=\int_0^\tau \mathcal{V}(\psi_{k,v},\psi_{k,v'};\eta_k^*,t)\mathbb{E}_W\left[Y_i(t)\exp(\eta_k^*(X_i))\right]dH_{0k}(t)\\&=\int_0^\tau \mathcal{V}(\psi_{k,v},\psi_{k,v'};\eta_k^*,t)S(\eta_k^*,t)dH_{0k}(t)=\mathcal{V}_k(\psi_{k,v},\psi_{k,v'}),\\\left|\int_0^\tau \mathcal{V}(\psi_{k,v},\psi_{k,v'};\eta_k^*,t)dN_k(t)\right|&\le\int_0^\tau\left|\mathcal{V}(\psi_{k,v},\psi_{k,v'};\eta_k^*,t)\right|dN_k(t)\\&\le\left\{\sup_{t\in[0,\tau]}\left|\mathcal{V}(\psi_{k,v},\psi_{k,v'};\eta_k^*,t)\right|\right\}N_k(\tau)\le 2\mathcal{C}_\psi^2.\end{aligned}$$

For the first component in (B.19), simple calculation yields

$$\begin{aligned}&\left|\int_0^\tau\left\{\mathcal{V}_{(n)}(f,g;\eta_k^*,t)-\mathcal{V}(f,g;\eta_k^*,t)\right\}d\overline{N}_k(t)\right|\\&\quad\le\left|\int_0^\tau\frac{S_{(n)}(fg;\eta_k^*,t)-S(fg;\eta_k^*,t)}{S_{(n)}(\eta_k^*,t)}d\overline{N}_k(t)\right|\\&\quad+\left|\int_0^\tau\frac{S(fg;\eta_k^*,t)}{S(\eta_k^*,t)}\frac{S(\eta_k^*,t)-S_{(n)}(\eta_k^*,t)}{S_{(n)}(\eta_k^*,t)}d\overline{N}_k(t)\right|\\&\quad+\left|\int_0^\tau\frac{S_{(n)}(f;\eta_k^*,t)}{S_{(n)}(\eta_k^*,t)}\frac{S_{(n)}(g;\eta_k^*,t)-S(g;\eta_k^*,t)}{S_{(n)}(\eta_k^*,t)}d\overline{N}_k(t)\right|\end{aligned}\tag{B.21}$$

$$+\left|\int_0^\tau \frac{S_{(n)}(f;\eta_k^*,t)}{S_{(n)}(\eta_k^*,t)}\frac{S(g;\eta_k^*,t)}{S(\eta_k^*,t)}\frac{S(\eta_k^*,t)-S_{(n)}(\eta_k^*,t)}{S_{(n)}(\eta_k^*,t)}d\overline{N}_k(t)\right|$$
$$+\left|\int_0^\tau \frac{S(g;\eta_k^*,t)}{S(\eta_k^*,t)}\frac{S_{(n)}(f;\eta_k^*,t)-S(f;\eta_k^*,t)}{S_{(n)}(\eta_k^*,t)}d\overline{N}_k(t)\right|$$
$$+\left|\int_0^\tau \frac{S(g;\eta_k^*,t)}{S(\eta_k^*,t)}\frac{S(f;\eta_k^*,t)}{S(\eta_k^*,t)}\frac{S(\eta_k^*,t)-S_{(n)}(\eta_k^*,t)}{S_{(n)}(\eta_k^*,t)}d\overline{N}_k(t)\right|.$$

For the first component of (B.21), we have

$$\begin{aligned}
&\left|\int_0^\tau \frac{S_{(n)}(fg;\eta_k^*,t)-S(fg;\eta_k^*,t)}{S_{(n)}(\eta_k^*,t)}d\overline{N}_k(t)\right| \\
&\quad=\left|\sum_{v\in\mathbb{N}}\sum_{v'\in\mathbb{N}}\mathcal{V}_k(f,\psi_{k,v})\mathcal{V}_k(g,\psi_{k,v'})\int_0^\tau \frac{S_{(n)}(\psi_{k,v}\psi_{k,v'};\eta_k^*,t)-S(\psi_{k,v}\psi_{k,v'};\eta_k^*,t)}{S_{(n)}(\eta_k^*,t)}d\overline{N}_k(t)\right| \\
&\quad\le\left\{\sum_{v\in\mathbb{N}}\sum_{v'\in\mathbb{N}}(1+\lambda_k\rho_{k,v})(1+\lambda_k\rho_{k,v'})\mathcal{V}_k^2(f,\psi_{k,v})\mathcal{V}_k^2(g,\psi_{k,v'})\right\}^{1/2} \\
&\quad\times\left\{\sum_{v\in\mathbb{N}}\sum_{v'\in\mathbb{N}}\frac{1}{1+\lambda_k\rho_{k,v}}\frac{1}{1+\lambda_k\rho_{k,v'}}\left\{\int_0^\tau \frac{S_{(n)}(\psi_{k,v}\psi_{k,v'};\eta_k^*,t)-S(\psi_{k,v}\psi_{k,v'};\eta_k^*,t)}{S_{(n)}(\eta_k^*,t)}d\overline{N}_k(t)\right\}^2\right\}^{1/2} \\
&\quad=\sqrt{(\mathcal{V}_k+\lambda_kJ_k)(f)(\mathcal{V}_k+\lambda_kJ_k)(g)}\,\mathcal{O}_{\mathbb{P}}\left(n^{-1/2}\lambda_k^{-1/r}\right).
\end{aligned}\tag{B.22}$$

The last equality in (B.22) is obtained by

$$\begin{aligned}
0&\le\sum_{v\in\mathbb{N}}\sum_{v'\in\mathbb{N}}\frac{1}{1+\lambda_k\rho_{k,v}}\frac{1}{1+\lambda_k\rho_{k,v'}}\left\{\int_0^\tau \frac{S_{(n)}(\psi_{k,v}\psi_{k,v'};\eta_k^*,t)-S(\psi_{k,v}\psi_{k,v'};\eta_k^*,t)}{S_{(n)}^{1/2}(\eta_k^*,t)S_{(n)}^{1/2}(\eta_k^*,t)}d\overline{N}_k(t)\right\}^2 \\
&\le\left\{\sum_{v\in\mathbb{N}}\sum_{v'\in\mathbb{N}}\frac{1}{1+\lambda_k\rho_{k,v}}\frac{1}{1+\lambda_k\rho_{k,v'}}\int_0^\tau \frac{\left(S_{(n)}(\psi_{k,v}\psi_{k,v'};\eta_k^*,t)-S(\psi_{k,v}\psi_{k,v'};\eta_k^*,t)\right)^2}{S_{(n)}(\eta_k^*,t)}d\overline{N}_k(t)\right\} \\
&\times\int_0^\tau \frac{1}{S_{(n)}(\eta_k^*,t)}d\overline{N}_k(t)=Z_{1,1}\times Z_{1,2}=\mathcal{O}_{\mathbb{P}}\left(n^{-1}\lambda_k^{-2/r}\right)\mathcal{O}_{\mathbb{P}}(1)=\mathcal{O}_{\mathbb{P}}\left(n^{-1}\lambda_k^{-2/r}\right),
\end{aligned}$$

where

$$\mathbb{E}|Z_{1,1}| = \mathbb{E}[Z_{1,1}] \leq \left\{ \sum_{v\in\mathbb{N}} \sum_{v'\in\mathbb{N}} \frac{1}{1+\lambda_k \rho_{k,v}} \frac{1}{1+\lambda_k \rho_{k,v'}} \right\} \frac{\exp(2\mathcal{C}_*)\mathcal{C}_\psi^4 \mathcal{C}_{h,2}\tau}{n} = \mathcal{O}\left(n^{-1}\lambda_k^{-2/r}\right)$$

and $\mathbb{E}|Z_{1,2}| = \mathbb{E}[Z_{1,2}] \leq \mathcal{C}_{h,2}\tau = \mathcal{O}(1)$.

For the second component of (B.21), we have

$$\begin{aligned}
&\left| \int_0^\tau \frac{S(fg;\eta_k^*,t)}{S(\eta_k^*,t)} \frac{S(\eta_k^*,t) - S_{(n)}(\eta_k^*,t)}{S_{(n)}(\eta_k^*,t)} d\overline{N}_k(t) \right| \\
&= \left| \sum_{v\in\mathbb{N}} \sum_{v'\in\mathbb{N}} \mathcal{V}_k(f,\psi_{k,v})\mathcal{V}_k(g,\psi_{k,v'}) \int_0^\tau \frac{S(\psi_{k,v}\psi_{k,v'};\eta_k^*,t)}{S(\eta_k^*,t)} \frac{S(\eta_k^*,t) - S_{(n)}(\eta_k^*,t)}{S_{(n)}(\eta_k^*,t)} d\overline{N}_k(t) \right| \\
&\leq \left\{ \sum_{v\in\mathbb{N}} \sum_{v'\in\mathbb{N}} (1+\lambda_k\rho_{k,v})(1+\lambda_k\rho_{k,v'})\mathcal{V}_k^2(f,\psi_{k,v})\mathcal{V}_k^2(g,\psi_{k,v'}) \right\}^{1/2} \\
&\times \left\{ \sum_{v\in\mathbb{N}} \sum_{v'\in\mathbb{N}} \frac{1}{1+\lambda_k \rho_{k,v}} \frac{1}{1+\lambda_k \rho_{k,v'}} \left\{ \int_0^\tau \frac{S(\psi_{k,v}\psi_{k,v'};\eta_k^*,t)}{S(\eta_k^*,t)} \frac{S(\eta_k^*,t) - S_{(n)}(\eta_k^*,t)}{S_{(n)}(\eta_k^*,t)} d\overline{N}_k(t) \right\}^2 \right\}^{1/2} \\
&= \sqrt{(\mathcal{V}_k + \lambda_k J_k)(f)(\mathcal{V}_k + \lambda_k J_k)(g)}\,\mathcal{O}_\mathbb{P}\left(n^{-1/2}\lambda_k^{-1/r}\right). \qquad \text{(B.23)}
\end{aligned}$$

The last equality in (B.23) is obtained by

$$\begin{aligned}
0 &\leq \sum_{v\in\mathbb{N}} \sum_{v'\in\mathbb{N}} \frac{1}{1+\lambda_k \rho_{k,v}} \frac{1}{1+\lambda_k \rho_{k,v'}} \left\{ \int_0^\tau \frac{S(\psi_{k,v}\psi_{k,v'};\eta_k^*,t)}{S(\eta_k^*,t)} \frac{S(\eta_k^*,t) - S_{(n)}(\eta_k^*,t)}{S_{(n)}(\eta_k^*,t)} d\overline{N}_k(t) \right\}^2 \\
&\leq \left\{ \sum_{v\in\mathbb{N}} \sum_{v'\in\mathbb{N}} \frac{1}{1+\lambda_k \rho_{k,v}} \frac{1}{1+\lambda_k \rho_{k,v'}} \int_0^\tau \frac{\left(S_{(n)}(\eta_k^*,t) - S(\eta_k^*,t)\right)^2}{S_{(n)}(\eta_k^*,t)} d\overline{N}_k(t) \right\} \\
&\times \mathcal{C}_\psi^4 \int_0^\tau \frac{1}{S_{(n)}(\eta_k^*,t)} d\overline{N}_k(t) = Z_{2,1} \times Z_{2,2} = \mathcal{O}_\mathbb{P}\left(n^{-1}\lambda_k^{-2/r}\right)\mathcal{O}_\mathbb{P}(1) = \mathcal{O}_\mathbb{P}\left(n^{-1}\lambda_k^{-2/r}\right),
\end{aligned}$$

where

$$\begin{aligned}
&\left|\int_0^\tau \frac{S(\psi_{k,v}\psi_{k,v'};\eta_k^*,t)}{S(\eta_k^*,t)}\frac{S(\eta_k^*,t)-S_{(n)}(\eta_k^*,t)}{S_{(n)}(\eta_k^*,t)}d\overline{N}_k(t)\right| \\
&\quad\le \int_0^\tau \left|\frac{S(\psi_{k,v}\psi_{k,v'};\eta_k^*,t)}{S(\eta_k^*,t)}\frac{S(\eta_k^*,t)-S_{(n)}(\eta_k^*,t)}{S_{(n)}(\eta_k^*,t)}\right| d\overline{N}_k(t) \\
&\quad\le \left\{\sup_{t\in[0,\tau]}\left|\frac{S(\psi_{k,v}\psi_{k,v'};\eta_k^*,t)}{S(\eta_k^*,t)}\right|\right\}\times\left|\int_0^\tau \frac{\left|S_{(n)}(\eta_k^*,t)-S(\eta_k^*,t)\right|}{S_{(n)}^{1/2}(\eta_k^*,t)S_{(n)}^{1/2}(\eta_k^*,t)}d\overline{N}_k(t)\right| \\
&\quad\le \mathcal{C}_\psi^2\left\{\int_0^\tau \frac{\left|S_{(n)}(\eta_k^*,t)-S(\eta_k^*,t)\right|^2}{S_{(n)}(\eta_k^*,t)}d\overline{N}_k(t)\right\}^{1/2}\left\{\int_0^\tau \frac{1}{S_{(n)}(\eta_k^*,t)}d\overline{N}_k(t)\right\}^{1/2},
\end{aligned}$$

$$\mathbb{E}|Z_{2,1}| = \mathbb{E}[Z_{2,1}] \le \left\{\sum_{v\in\mathbb{N}}\sum_{v'\in\mathbb{N}}\frac{1}{1+\lambda_k\rho_{k,v}}\frac{1}{1+\lambda_k\rho_{k,v'}}\right\}\frac{\exp(2\mathcal{C}_*)\mathcal{C}_{h,2}\tau}{n} = \mathcal{O}\left(n^{-1}\lambda_k^{-2/r}\right),$$

and $\mathbb{E}|Z_{2,2}| = \mathbb{E}[Z_{2,2}] \le \mathcal{C}_\psi^4\mathcal{C}_{h,2}\tau = \mathcal{O}(1)$.

For the third component of (B.21), we have

$$\begin{aligned}
&\left|\int_0^\tau \frac{S_{(n)}(f;\eta_k^*,t)}{S_{(n)}(\eta_k^*,t)}\frac{S_{(n)}(g;\eta_k^*,t)-S(g;\eta_k^*,t)}{S_{(n)}(\eta_k^*,t)}d\overline{N}_k(t)\right| \\
&\quad= \left|\sum_{v\in\mathbb{N}}\sum_{v'\in\mathbb{N}}\mathcal{V}_k(f,\psi_{k,v})\mathcal{V}_k(g,\psi_{k,v'})\int_0^\tau \frac{S_{(n)}(\psi_{k,v};\eta_k^*,t)}{S_{(n)}(\eta_k^*,t)}\frac{S_{(n)}(\psi_{k,v'};\eta_k^*,t)-S(\psi_{k,v'};\eta_k^*,t)}{S_{(n)}(\eta_k^*,t)}d\overline{N}_k(t)\right| \\
&\quad\le \left\{\sum_{v\in\mathbb{N}}\sum_{v'\in\mathbb{N}}(1+\lambda_k\rho_{k,v})(1+\lambda_k\rho_{k,v'})\mathcal{V}_k^2(f,\psi_{k,v})\mathcal{V}_k^2(g,\psi_{k,v'})\right\}^{1/2} \\
&\quad\times\left\{\sum_{v\in\mathbb{N}}\sum_{v'\in\mathbb{N}}\frac{1}{1+\lambda_k\rho_{k,v}}\frac{1}{1+\lambda_k\rho_{k,v'}}\left\{\int_0^\tau \frac{S_{(n)}(\psi_{k,v};\eta_k^*,t)}{S_{(n)}(\eta_k^*,t)}\frac{S_{(n)}(\psi_{k,v'};\eta_k^*,t)-S(\psi_{k,v'};\eta_k^*,t)}{S_{(n)}(\eta_k^*,t)}d\overline{N}_k(t)\right\}^2\right\}^{1/2} \\
&\quad= \sqrt{(\mathcal{V}_k+\lambda_kJ_k)(f)(\mathcal{V}_k+\lambda_kJ_k)(g)}\,\mathcal{O}_{\mathbb{P}}\left(n^{-1/2}\lambda_k^{-1/r}\right). \qquad \text{(B.24)}
\end{aligned}$$

The last equality in (B.24) is obtained by

$$0 \leq \sum_{v \in \mathbb{N}} \sum_{v' \in \mathbb{N}} \frac{1}{1+\lambda_k \rho_{k,v}} \frac{1}{1+\lambda_k \rho_{k,v'}} \left\{ \int_0^\tau \frac{S_{(n)}(\psi_{k,v};\eta_k^*,t)}{S_{(n)}(\eta_k^*,t)} \frac{S_{(n)}(\psi_{k,v'};\eta_k^*,t) - S(\psi_{k,v'};\eta_k^*,t)}{S_{(n)}(\eta_k^*,t)} d\overline{N}_k(t) \right\}^2$$

$$\leq \left\{ \sum_{v \in \mathbb{N}} \sum_{v' \in \mathbb{N}} \frac{1}{1+\lambda_k \rho_{k,v}} \frac{1}{1+\lambda_k \rho_{k,v'}} \int_0^\tau \frac{\left(S_{(n)}(\psi_{k,v'};\eta_k^*,t) - S(\psi_{k,v'};\eta_k^*,t) \right)^2}{S_{(n)}(\eta_k^*,t)} d\overline{N}_k(t) \right\}$$
$$\times \mathcal{C}_\psi^2 \int_0^\tau \frac{1}{S_{(n)}(\eta_k^*,t)} d\overline{N}_k(t) = Z_{3,1} \times Z_{3,2} = \mathcal{O}_{\mathbb{P}}\left(n^{-1}\lambda_k^{-2/r}\right) \mathcal{O}_{\mathbb{P}}(1) = \mathcal{O}_{\mathbb{P}}\left(n^{-1}\lambda_k^{-2/r}\right),$$

where

$$\left| \int_0^\tau \frac{S_{(n)}(\psi_{k,v};\eta_k^*,t)}{S_{(n)}(\eta_k^*,t)} \frac{S_{(n)}(\psi_{k,v'};\eta_k^*,t) - S(\psi_{k,v'};\eta_k^*,t)}{S_{(n)}(\eta_k^*,t)} d\overline{N}_k(t) \right|$$
$$\leq \int_0^\tau \left| \frac{S_{(n)}(\psi_{k,v};\eta_k^*,t)}{S_{(n)}(\eta_k^*,t)} \frac{S_{(n)}(\psi_{k,v'};\eta_k^*,t) - S(\psi_{k,v'};\eta_k^*,t)}{S_{(n)}(\eta_k^*,t)} \right| d\overline{N}_k(t)$$
$$\leq \left\{ \sup_{t \in [0,\tau]} \left| \frac{S_{(n)}(\psi_{k,v};\eta_k^*,t)}{S_{(n)}(\eta_k^*,t)} \right| \right\} \times \left| \int_0^\tau \frac{\left| S_{(n)}(\psi_{k,v'};\eta_k^*,t) - S(\psi_{k,v'};\eta_k^*,t) \right|}{S_{(n)}^{1/2}(\eta_k^*,t) S_{(n)}^{1/2}(\eta_k^*,t)} d\overline{N}_k(t) \right|$$
$$\leq \mathcal{C}_\psi \left\{ \int_0^\tau \frac{\left| S_{(n)}(\psi_{k,v'};\eta_k^*,t) - S(\psi_{k,v'};\eta_k^*,t) \right|^2}{S_{(n)}(\eta_k^*,t)} d\overline{N}_k(t) \right\}^{1/2} \left\{ \int_0^\tau \frac{1}{S_{(n)}(\eta_k^*,t)} d\overline{N}_k(t) \right\}^{1/2},$$

$$\mathbb{E}|Z_{3,1}| = \mathbb{E}[Z_{3,1}] \leq \left\{ \sum_{v \in \mathbb{N}} \sum_{v' \in \mathbb{N}} \frac{1}{1+\lambda_k \rho_{k,v}} \frac{1}{1+\lambda_k \rho_{k,v'}} \right\} \frac{\exp(2\mathcal{C}_*)\mathcal{C}_\psi^2 \mathcal{C}_{h,2}\tau}{n} = \mathcal{O}\left(n^{-1}\lambda_k^{-2/r}\right),$$

and $\mathbb{E}|Z_{3,2}| = \mathbb{E}[Z_{3,2}] \leq \mathcal{C}_\psi^2 \mathcal{C}_{h,2}\tau = \mathcal{O}(1)$.

For the fourth component of (B.21), we have

$$\left| \int_0^\tau \frac{S_{(n)}(f;\eta_k^*,t)}{S_{(n)}(\eta_k^*,t)} \frac{S(g;\eta_k^*,t)}{S(\eta_k^*,t)} \frac{S(\eta_k^*,t) - S_{(n)}(\eta_k^*,t)}{S_{(n)}(\eta_k^*,t)} d\overline{N}_k(t) \right|$$

$$
\begin{aligned}
&= \left| \sum_{v\in\mathbb{N}}\sum_{v'\in\mathbb{N}} \mathcal{V}_k(f,\psi_{k,v})\mathcal{V}_k(g,\psi_{k,v'}) \int_0^\tau \frac{S_{(n)}(\psi_{k,v};\eta_k^*,t)}{S_{(n)}(\eta_k^*,t)}\frac{S(\psi_{k,v'};\eta_k^*,t)}{S(\eta_k^*,t)}\frac{S(\eta_k^*,t)-S_{(n)}(\eta_k^*,t)}{S_{(n)}(\eta_k^*,t)}d\overline{N}_k(t)\right| \\
&\le \left\{ \sum_{v\in\mathbb{N}}\sum_{v'\in\mathbb{N}} (1+\lambda_k\rho_{k,v})(1+\lambda_k\rho_{k,v'})\mathcal{V}_k^2(f,\psi_{k,v})\mathcal{V}_k^2(g,\psi_{k,v'}) \right\}^{1/2} \\
&\times \left\{ \sum_{v\in\mathbb{N}}\sum_{v'\in\mathbb{N}} \frac{1}{1+\lambda_k\rho_{k,v}}\frac{1}{1+\lambda_k\rho_{k,v'}} \left\{ \int_0^\tau \frac{S_{(n)}(\psi_{k,v};\eta_k^*,t)}{S_{(n)}(\eta_k^*,t)}\frac{S(\psi_{k,v'};\eta_k^*,t)}{S(\eta_k^*,t)}\frac{S(\eta_k^*,t)-S_{(n)}(\eta_k^*,t)}{S_{(n)}(\eta_k^*,t)}d\overline{N}_k(t)\right\}^2 \right\}^{1/2} \\
&= \sqrt{(\mathcal{V}_k+\lambda_k J_k)(f)(\mathcal{V}_k+\lambda_k J_k)(g)}\,\mathcal{O}_{\mathbb{P}}\left(n^{-1/2}\lambda_k^{-1/r}\right). \qquad \text{(B.25)}
\end{aligned}
$$

The last equality in (B.25) is obtained by

$$
\begin{aligned}
0 &\le \sum_{v\in\mathbb{N}}\sum_{v'\in\mathbb{N}} \frac{1}{1+\lambda_k\rho_{k,v}}\frac{1}{1+\lambda_k\rho_{k,v'}} \left\{ \int_0^\tau \frac{S_{(n)}(\psi_{k,v};\eta_k^*,t)}{S_{(n)}(\eta_k^*,t)}\frac{S(\psi_{k,v'};\eta_k^*,t)}{S(\eta_k^*,t)}\frac{S(\eta_k^*,t)-S_{(n)}(\eta_k^*,t)}{S_{(n)}(\eta_k^*,t)}d\overline{N}_k(t)\right\}^2 \\
&\le \left\{ \sum_{v\in\mathbb{N}}\sum_{v'\in\mathbb{N}} \frac{1}{1+\lambda_k\rho_{k,v}}\frac{1}{1+\lambda_k\rho_{k,v'}} \int_0^\tau \frac{\left(S_{(n)}(\eta_k^*,t)-S(\eta_k^*,t)\right)^2}{S_{(n)}(\eta_k^*,t)}d\overline{N}_k(t) \right\} \\
&\times \mathcal{C}_\psi^4 \int_0^\tau \frac{1}{S_{(n)}(\eta_k^*,t)}d\overline{N}_k(t) = Z_{4,1}\times Z_{4,2} = \mathcal{O}_{\mathbb{P}}\left(n^{-1}\lambda_k^{-2/r}\right)\mathcal{O}_{\mathbb{P}}(1) = \mathcal{O}_{\mathbb{P}}\left(n^{-1}\lambda_k^{-2/r}\right),
\end{aligned}
$$

where

$$
\begin{aligned}
&\left| \int_0^\tau \frac{S_{(n)}(\psi_{k,v};\eta_k^*,t)}{S_{(n)}(\eta_k^*,t)}\frac{S(\psi_{k,v'};\eta_k^*,t)}{S(\eta_k^*,t)}\frac{S(\eta_k^*,t)-S_{(n)}(\eta_k^*,t)}{S_{(n)}(\eta_k^*,t)}d\overline{N}_k(t)\right| \\
&\quad \le \int_0^\tau \left| \frac{S_{(n)}(\psi_{k,v};\eta_k^*,t)}{S_{(n)}(\eta_k^*,t)}\frac{S(\psi_{k,v'};\eta_k^*,t)}{S(\eta_k^*,t)}\frac{S(\eta_k^*,t)-S_{(n)}(\eta_k^*,t)}{S_{(n)}(\eta_k^*,t)}\right| d\overline{N}_k(t) \\
&\quad \le \left\{ \sup_{t\in[0,\tau]} \left| \frac{S_{(n)}(\psi_{k,v};\eta_k^*,t)}{S_{(n)}(\eta_k^*,t)} \right| \right\} \left\{ \sup_{t\in[0,\tau]} \left| \frac{S(\psi_{k,v'};\eta_k^*,t)}{S(\eta_k^*,t)} \right| \right\} \times \left| \int_0^\tau \frac{\left|S_{(n)}(\eta_k^*,t)-S(\eta_k^*,t)\right|}{S_{(n)}^{1/2}(\eta_k^*,t)S_{(n)}^{1/2}(\eta_k^*,t)}d\overline{N}_k(t) \right|
\end{aligned}
$$

$$\leq \mathcal{C}_\psi^2 \left\{ \int_0^\tau \frac{\left|S_{(n)}(\eta_k^*, t) - S(\eta_k^*, t)\right|^2}{S_{(n)}(\eta_k^*, t)} d\overline{N}_k(t) \right\}^{1/2} \left\{ \int_0^\tau \frac{1}{S_{(n)}(\eta_k^*, t)} d\overline{N}_k(t) \right\}^{1/2},$$

$$\mathbb{E}|Z_{4,1}| = \mathbb{E}[Z_{4,1}] \leq \left\{ \sum_{v\in\mathbb{N}} \sum_{v'\in\mathbb{N}} \frac{1}{1+\lambda_k\rho_{k,v}} \frac{1}{1+\lambda_k\rho_{k,v'}} \right\} \frac{\exp(2\mathcal{C}_*)\mathcal{C}_{h,2}\tau}{n} = \mathcal{O}\left(n^{-1}\lambda_k^{-2/r}\right),$$

and $\mathbb{E}|Z_{4,2}| = \mathbb{E}[Z_{4,2}] \leq \mathcal{C}_\psi^4 \mathcal{C}_{h,2}\tau = \mathcal{O}(1)$.

For the fifth component of (B.21), we have

$$\begin{aligned}
&\left| \int_0^\tau \frac{S(g;\eta_k^*, t)}{S(\eta_k^*, t)} \frac{S_{(n)}(f;\eta_k^*, t) - S(f;\eta_k^*, t)}{S_{(n)}(\eta_k^*, t)} d\overline{N}_k(t) \right| \\
&= \left| \sum_{v\in\mathbb{N}} \sum_{v'\in\mathbb{N}} \mathcal{V}_k(f, \psi_{k,v}) \mathcal{V}_k(g, \psi_{k,v'}) \int_0^\tau \frac{S(\psi_{k,v'};\eta_k^*, t)}{S(\eta_k^*, t)} \frac{S_{(n)}(\psi_{k,v};\eta_k^*, t) - S(\psi_{k,v};\eta_k^*, t)}{S_{(n)}(\eta_k^*, t)} d\overline{N}_k(t) \right| \\
&\leq \left\{ \sum_{v\in\mathbb{N}} \sum_{v'\in\mathbb{N}} (1+\lambda_k\rho_{k,v})(1+\lambda_k\rho_{k,v'}) \mathcal{V}_k^2(f, \psi_{k,v}) \mathcal{V}_k^2(g, \psi_{k,v'}) \right\}^{1/2} \\
&\times \left\{ \sum_{v\in\mathbb{N}} \sum_{v'\in\mathbb{N}} \frac{1}{1+\lambda_k\rho_{k,v}} \frac{1}{1+\lambda_k\rho_{k,v'}} \left\{ \int_0^\tau \frac{S(\psi_{k,v'};\eta_k^*, t)}{S(\eta_k^*, t)} \frac{S_{(n)}(\psi_{k,v};\eta_k^*, t) - S(\psi_{k,v};\eta_k^*, t)}{S_{(n)}(\eta_k^*, t)} d\overline{N}_k(t) \right\}^2 \right\}^{1/2} \\
&= \sqrt{(\mathcal{V}_k + \lambda_k J_k)(f)(\mathcal{V}_k + \lambda_k J_k)(g)} \mathcal{O}_{\mathbb{P}}\left(n^{-1/2}\lambda_k^{-1/r}\right). \qquad \text{(B.26)}
\end{aligned}$$

The last equality in (B.26) is obtained by

$$\begin{aligned}
0 &\leq \sum_{v\in\mathbb{N}} \sum_{v'\in\mathbb{N}} \frac{1}{1+\lambda_k\rho_{k,v}} \frac{1}{1+\lambda_k\rho_{k,v'}} \left\{ \int_0^\tau \frac{S(\psi_{k,v'};\eta_k^*, t)}{S(\eta_k^*, t)} \frac{S_{(n)}(\psi_{k,v};\eta_k^*, t) - S(\psi_{k,v};\eta_k^*, t)}{S_{(n)}(\eta_k^*, t)} d\overline{N}_k(t) \right\}^2 \\
&\leq \left\{ \sum_{v\in\mathbb{N}} \sum_{v'\in\mathbb{N}} \frac{1}{1+\lambda_k\rho_{k,v}} \frac{1}{1+\lambda_k\rho_{k,v'}} \int_0^\tau \frac{\left(S_{(n)}(\psi_{k,v};\eta_k^*, t) - S(\psi_{k,v};\eta_k^*, t)\right)^2}{S_{(n)}(\eta_k^*, t)} d\overline{N}_k(t) \right\}
\end{aligned}$$

$$\times \mathcal{C}_\psi^2 \int_0^\tau \frac{1}{S_{(n)}(\eta_k^*, t)} d\overline{N}_k(t) = Z_{5,1} \times Z_{5,2} = \mathcal{O}_{\mathbb{P}}\left(n^{-1}\lambda_k^{-2/r}\right) \mathcal{O}_{\mathbb{P}}(1) = \mathcal{O}_{\mathbb{P}}\left(n^{-1}\lambda_k^{-2/r}\right),$$

where

$$\begin{aligned}
&\left|\int_0^\tau \frac{S(\psi_{k,v'}; \eta_k^*, t)}{S(\eta_k^*, t)} \frac{S_{(n)}(\psi_{k,v}; \eta_k^*, t) - S(\psi_{k,v}; \eta_k^*, t)}{S_{(n)}(\eta_k^*, t)} d\overline{N}_k(t)\right| \\
&\quad \le \int_0^\tau \left|\frac{S(\psi_{k,v'}; \eta_k^*, t)}{S(\eta_k^*, t)} \frac{S_{(n)}(\psi_{k,v}; \eta_k^*, t) - S(\psi_{k,v}; \eta_k^*, t)}{S_{(n)}(\eta_k^*, t)}\right| d\overline{N}_k(t) \\
&\quad \le \left\{\sup_{t\in[0,\tau]} \left|\frac{S(\psi_{k,v'}; \eta_k^*, t)}{S(\eta_k^*, t)}\right|\right\} \times \left|\int_0^\tau \frac{\left|S_{(n)}(\psi_{k,v}; \eta_k^*, t) - S(\psi_{k,v}; \eta_k^*, t)\right|}{S_{(n)}^{1/2}(\eta_k^*, t) S_{(n)}^{1/2}(\eta_k^*, t)} d\overline{N}_k(t)\right| \\
&\quad \le \mathcal{C}_\psi \left\{\int_0^\tau \frac{\left|S_{(n)}(\psi_{k,v}; \eta_k^*, t) - S(\psi_{k,v}; \eta_k^*, t)\right|^2}{S_{(n)}(\eta_k^*, t)} d\overline{N}_k(t)\right\}^{1/2} \left\{\int_0^\tau \frac{1}{S_{(n)}(\eta_k^*, t)} d\overline{N}_k(t)\right\}^{1/2},
\end{aligned}$$

$$\mathbb{E}|Z_{5,1}| = \mathbb{E}[Z_{5,1}] \le \left\{\sum_{v\in\mathbb{N}} \sum_{v'\in\mathbb{N}} \frac{1}{1+\lambda_k\rho_{k,v}} \frac{1}{1+\lambda_k\rho_{k,v'}}\right\} \frac{\exp(2\mathcal{C}_*)\mathcal{C}_\psi^2 \mathcal{C}_{h,2}\tau}{n} = \mathcal{O}\left(n^{-1}\lambda_k^{-2/r}\right),$$

and $\mathbb{E}|Z_{5,2}| = \mathbb{E}[Z_{5,2}] \le \mathcal{C}_\psi^2 \mathcal{C}_{h,2}\tau = \mathcal{O}(1)$.

For the sixth component of (B.21), we have

$$\begin{aligned}
&\left|\int_0^\tau \frac{S(g; \eta_k^*, t)}{S(\eta_k^*, t)} \frac{S(f; \eta_k^*, t)}{S(\eta_k^*, t)} \frac{S(\eta_k^*, t) - S_{(n)}(\eta_k^*, t)}{S_{(n)}(\eta_k^*, t)} d\overline{N}_k(t)\right| \\
&\quad = \left|\sum_{v\in\mathbb{N}} \sum_{v'\in\mathbb{N}} \mathcal{V}_k(f, \psi_{k,v}) \mathcal{V}_k(g, \psi_{k,v'}) \int_0^\tau \frac{S(\psi_{k,v'}; \eta_k^*, t)}{S(\eta_k^*, t)} \frac{S(\psi_{k,v}; \eta_k^*, t)}{S(\eta_k^*, t)} \frac{S(\eta_k^*, t) - S_{(n)}(\eta_k^*, t)}{S_{(n)}(\eta_k^*, t)} d\overline{N}_k(t)\right|
\end{aligned}$$

$$\begin{aligned}
&\le \left\{\sum_{v\in\mathbb{N}} \sum_{v'\in\mathbb{N}} (1+\lambda_k\rho_{k,v})(1+\lambda_k\rho_{k,v'}) \mathcal{V}_k^2(f, \psi_{k,v}) \mathcal{V}_k^2(g, \psi_{k,v'})\right\}^{1/2} \\
&\times \left\{\sum_{v\in\mathbb{N}} \sum_{v'\in\mathbb{N}} \frac{1}{1+\lambda_k\rho_{k,v}} \frac{1}{1+\lambda_k\rho_{k,v'}} \left\{\int_0^\tau \frac{S(\psi_{k,v'}; \eta_k^*, t)}{S(\eta_k^*, t)} \frac{S(\psi_{k,v}; \eta_k^*, t)}{S(\eta_k^*, t)} \frac{S(\eta_k^*, t) - S_{(n)}(\eta_k^*, t)}{S_{(n)}(\eta_k^*, t)} d\overline{N}_k(t)\right\}^2\right\}^{1/2}
\end{aligned}$$

$$= \sqrt{(\mathcal{V}_k + \lambda_k J_k)(f)(\mathcal{V}_k + \lambda_k J_k)(g)}\mathcal{O}_{\mathbb{P}}\left(n^{-1/2}\lambda_k^{-1/r}\right). \tag{B.27}$$

The last equality in (B.27) is obtained by

$$0 \leq \sum_{v\in\mathbb{N}}\sum_{v'\in\mathbb{N}} \frac{1}{1+\lambda_k\rho_{k,v}}\frac{1}{1+\lambda_k\rho_{k,v'}}\left\{\int_0^\tau \frac{S(\psi_{k,v'};\eta_k^*,t)}{S(\eta_k^*,t)}\frac{S(\psi_{k,v};\eta_k^*,t)}{S(\eta_k^*,t)}\frac{S(\eta_k^*,t)-S_{(n)}(\eta_k^*,t)}{S_{(n)}(\eta_k^*,t)}d\overline{N}_k(t)\right\}^2$$

$$\leq \left\{\sum_{v\in\mathbb{N}}\sum_{v'\in\mathbb{N}} \frac{1}{1+\lambda_k\rho_{k,v}}\frac{1}{1+\lambda_k\rho_{k,v'}}\int_0^\tau \frac{\left(S_{(n)}(\eta_k^*,t)-S(\eta_k^*,t)\right)^2}{S_{(n)}(\eta_k^*,t)}d\overline{N}_k(t)\right\}$$
$$\times\, \mathcal{C}_\psi^4\int_0^\tau \frac{1}{S_{(n)}(\eta_k^*,t)}d\overline{N}_k(t) = Z_{6,1}\times Z_{6,2} = \mathcal{O}_{\mathbb{P}}\left(n^{-1}\lambda_k^{-2/r}\right)\mathcal{O}_{\mathbb{P}}(1) = \mathcal{O}_{\mathbb{P}}\left(n^{-1}\lambda_k^{-2/r}\right),$$

where

$$\left|\int_0^\tau \frac{S(\psi_{k,v'};\eta_k^*,t)}{S(\eta_k^*,t)}\frac{S(\psi_{k,v};\eta_k^*,t)}{S(\eta_k^*,t)}\frac{S(\eta_k^*,t)-S_{(n)}(\eta_k^*,t)}{S_{(n)}(\eta_k^*,t)}d\overline{N}_k(t)\right|$$
$$\leq \int_0^\tau \left|\frac{S(\psi_{k,v'};\eta_k^*,t)}{S(\eta_k^*,t)}\frac{S(\psi_{k,v};\eta_k^*,t)}{S(\eta_k^*,t)}\frac{S(\eta_k^*,t)-S_{(n)}(\eta_k^*,t)}{S_{(n)}(\eta_k^*,t)}\right|d\overline{N}_k(t)$$
$$\leq \left\{\sup_{t\in[0,\tau]}\left|\frac{S(\psi_{k,v'};\eta_k^*,t)}{S(\eta_k^*,t)}\right|\right\}\left\{\sup_{t\in[0,\tau]}\left|\frac{S(\psi_{k,v};\eta_k^*,t)}{S(\eta_k^*,t)}\right|\right\}\times\left|\int_0^\tau \frac{\left|S_{(n)}(\eta_k^*,t)-S(\eta_k^*,t)\right|}{S_{(n)}^{1/2}(\eta_k^*,t)S_{(n)}^{1/2}(\eta_k^*,t)}d\overline{N}_k(t)\right|$$

$$\leq \mathcal{C}_\psi^2\left\{\int_0^\tau \frac{\left|S_{(n)}(\eta_k^*,t)-S(\eta_k^*,t)\right|^2}{S_{(n)}(\eta_k^*,t)}d\overline{N}_k(t)\right\}^{1/2}\left\{\int_0^\tau \frac{1}{S_{(n)}(\eta_k^*,t)}d\overline{N}_k(t)\right\}^{1/2},$$

$$\mathbb{E}|Z_{6,1}| = \mathbb{E}[Z_{6,1}] \leq \left\{\sum_{v\in\mathbb{N}}\sum_{v'\in\mathbb{N}} \frac{1}{1+\lambda_k\rho_{k,v}}\frac{1}{1+\lambda_k\rho_{k,v'}}\right\}\frac{\exp(2\mathcal{C}_*)\mathcal{C}_{h,2}\tau}{n} = \mathcal{O}\left(n^{-1}\lambda_k^{-2/r}\right),$$

and $\mathbb{E}|Z_{6,2}| = \mathbb{E}[Z_{6,2}] \leq \mathcal{C}_\psi^4\mathcal{C}_{h,2}\tau = \mathcal{O}(1)$.

Considering (B.21), (B.22), (B.23), (B.24), (B.25), (B.26), and (B.27), we get

$$\begin{aligned}&\left|\int_0^\tau \left\{\mathcal{V}_{(n)}(f,g;\eta_k^*,t)-\mathcal{V}(f,g;\eta_k^*,t)\right\} d\overline{N}_k(t)\right| \\ &\quad = \sqrt{(\mathcal{V}_k+\lambda_k J_k)(f)(\mathcal{V}_k+\lambda_k J_k)(g)}\,\mathcal{O}_{\mathbb{P}}\left(n^{-1/2}\lambda_k^{-1/r}\right).\end{aligned} \tag{B.28}$$

Combining (B.19), (B.20), and (B.28), we obtain

$$\left|\mathcal{V}_k^{(n)}(f,g)-\mathcal{V}_k(f,g)\right| = \sqrt{(\mathcal{V}_k+\lambda_k J_k)(f)(\mathcal{V}_k+\lambda_k J_k)(g)}\,\mathcal{O}_{\mathbb{P}}\left(n^{-1/2}\lambda_k^{-1/r}\right),$$

where $n^{-1/2}\lambda_k^{-1/r} \ll q_k^{-1/2}\lambda_k^{-1/r} = o(1)$ completes the proof. ■

Proof of Lemma B.4. Here, each $f \in \mathcal{H}$ or $g \in \mathcal{H}$ can represent either a deterministic function or a stochastic process. For $\tilde{\alpha} \in [0,1]$, let $D_k(\tilde{\alpha}) = \mathcal{U}^{(n)}_{g+\tilde{\alpha}(f-g),k}(f-g)$. By the mean value theorem, for some random variable $\alpha \in (0,1)$, we have

$$\mathcal{U}^{(n)}_{f,k}(f-g) - \mathcal{U}^{(n)}_{g,k}(f-g) = \frac{D_k(1)-D_k(0)}{1-0} = \frac{\partial}{\partial\alpha}D_k(\alpha) = \mathcal{V}^{(n)}_{g+\alpha(f-g),k}(f-g).$$

■

Proof of Lemma B.5. Here, $h \in \mathcal{H}$ can represent either a deterministic function or a stochastic process. Let $\sigma_1 = \{\mathcal{V}_k^{(n)}(f_k-\eta_k^*)\}^{1/2}$ and $C_1(\alpha_1) = \mathcal{U}^{(n)}_{\eta_k^*+\alpha_1(f_k-\eta_k^*)/\sigma_1,k}(h) - \mathcal{U}_k^{(n)}(h)$. For $\alpha_1 \approx 0$, by the Taylor series expansion of $C_1(\alpha_1)$ at 0, we get

$$\begin{aligned}C_1(\alpha_1) &= C_1(0) + C_1'(0)\alpha_1 + o_{\alpha_1}\left(C_1'(0)\alpha_1\right) = \alpha_1 C_1'(0)\left(1+o_{\alpha_1}(1)\right) \\ &= \frac{\alpha_1}{\sigma_1}\mathcal{V}_k^{(n)}(f_k-\eta_k^*,h)\left(1+o_{\alpha_1}(1)\right),\end{aligned}$$

where $C_1'(\alpha_1) = \partial C_1(\alpha_1)/(\partial\alpha_1)$ and $o_{\alpha_1}(\cdot)$ is $o(\cdot)$ with respect to $\alpha_1 \to 0$. As $\sigma_1 \approx 0$, we have

$$\mathcal{U}^{(n)}_{f_k,k}(h) - \mathcal{U}^{(n)}_k(h) = C_1(\alpha_1 = \sigma_1) = \mathcal{V}^{(n)}_k(f_k - \eta^*_k, h)\,(1 + o_{\mathbb{P}}(1))\,,$$

where the fact that $\sigma_1 = o_{\mathbb{P}}(1)$ as $n \to \infty$ leads to the substitution of $o_{\alpha_1=\sigma_1}(\cdot)$ with $o_{\mathbb{P}}(\cdot)$ as $n \to \infty$.

Let

$$\sigma_2 = \max\left\{\{\mathcal{V}^{(n)}_k(f_k - \eta^*_k)\}^{1/2}, \{\mathcal{V}^{(n)}_k(g_k - \eta^*_k)\}^{1/2}\right\}$$

and $C_2(\alpha_2) = \mathcal{U}^{(n)}_{\eta^*_k+\alpha_2(f_k-\eta^*_k)/\sigma_2,k}(h) - \mathcal{U}^{(n)}_{\eta^*_k+\alpha_2(g_k-\eta^*_k)/\sigma_2,k}(h)$. For $\alpha_2 \approx 0$, by the Taylor series expansion of $C_2(\alpha_2)$ at 0, we get

$$\begin{aligned} C_2(\alpha_2) &= C_2(0) + C_2'(0)\alpha_2 + o_{\alpha_2}\left(C_2'(0)\alpha_2\right) = \alpha_2 C_2'(0)\,(1 + o_{\alpha_2}(1)) \\ &= \frac{\alpha_2}{\sigma_2}\left(\mathcal{V}^{(n)}_k(f_k - \eta^*_k, h) - \mathcal{V}^{(n)}_k(g_k - \eta^*_k, h)\right)(1 + o_{\alpha_2}(1)) = \frac{\alpha_2}{\sigma_2}\mathcal{V}^{(n)}_k(f_k - g_k, h)\,(1 + o_{\alpha_2}(1))\,, \end{aligned}$$

where $C_2'(\alpha_2) = \partial C_2(\alpha_2)/(\partial\alpha_2)$ and $o_{\alpha_2}(\cdot)$ is $o(\cdot)$ with respect to $\alpha_2 \to 0$. As $\sigma_2 \approx 0$, we have

$$\mathcal{U}^{(n)}_{f_k,k}(h) - \mathcal{U}^{(n)}_{g_k,k}(h) = C_2(\alpha_2 = \sigma_2) = \mathcal{V}^{(n)}_k(f_k - g_k, h)\,(1 + o_{\mathbb{P}}(1))\,,$$

where the fact that $\sigma_2 = o_{\mathbb{P}}(1)$ as $n \to \infty$ leads to the substitution of $o_{\alpha_2=\sigma_2}(\cdot)$ with $o_{\mathbb{P}}(\cdot)$ as $n \to \infty$. ■

Lemma B.7. *Under Assumption B.3.1, for each* k, $J_k(\eta_k, \hat{\eta}^*_k - \eta_k) = 0$ *and* $J_k(\hat{\eta}_k, \hat{\eta}^*_k - \eta_k) = 0$.

Proof of Lemma B.7. Considering $\eta_k = \operatorname{arginf}_{\eta \in \mathcal{H}^{(k)}} \langle \hat{\eta}^*_k - \eta, \hat{\eta}^*_k - \eta \rangle_{\mathcal{H},k} \in \mathcal{H}^{(k)}$, we have $\langle f, \hat{\eta}^*_k - \eta_k \rangle_{\mathcal{H},k} = 0$ for all $f \in \mathcal{H}^{(k)}$. By the representer theorem, $\hat{\eta}^*_k \in \mathcal{H}^{(k)}_{(n)}$, where $\mathcal{H}^{(k)} \subseteq$

$\mathcal{H}^{(k)}_{(n)} \equiv \mathcal{N}_J \oplus \mathcal{H}^{(k)}_{J(n)} \subseteq \mathcal{H}$ with $\mathcal{H}^{(k)}_J \subseteq \mathcal{H}^{(k)}_{J(n)} \equiv \mathrm{Span}\{\mathcal{K}_{J_k}(X_1,\cdot),\ldots,\mathcal{K}_{J_k}(X_n,\cdot)\} \subseteq \mathcal{H}_J$. By projecting with respect to $\langle\cdot,\cdot\rangle_{\mathcal{H},k}$ onto subspaces, one has $\hat{\eta}^*_k = \hat{\eta}^*_{k,N} + \hat{\eta}^*_{k,J}$, $\eta_k = \eta_{k,N} + \eta_{k,J}$, and $\hat{\eta}_k = \hat{\eta}_{k,N} + \hat{\eta}_{k,J}$, where

$$\hat{\eta}^*_{k,N}, \eta_{k,N}, \hat{\eta}_{k,N} \in \mathcal{N}_J,\ \hat{\eta}^*_{k,J} \in \mathcal{H}^{(k)}_{J(n)},\ \text{and } \eta_{k,J}, \hat{\eta}_{k,J} \in \mathcal{H}^{(k)}_J.$$

Note that

$$\langle \hat{\eta}^*_k - \eta_k, \hat{\eta}^*_k - \eta_k\rangle_{\mathcal{H},k} = \langle \hat{\eta}^*_{k,N} - \eta_{k,N}, \hat{\eta}^*_{k,N} - \eta_{k,N}\rangle_{\mathcal{N}_J} + J_k(\hat{\eta}^*_{k,J} - \eta_{k,J}, \hat{\eta}^*_{k,J} - \eta_{k,J}),$$

where, by the definition of η_k, we have $\eta_{k,N} = \hat{\eta}^*_{k,N} \in \mathcal{N}_J$, which minimizes $\langle\cdot,\cdot\rangle_{\mathcal{N}_J}$ term in above to zero. As a result, $\hat{\eta}^*_k - \eta_k = \hat{\eta}^*_{k,J} - \eta_{k,J}$. Considering $\eta_{k,J}, \hat{\eta}_{k,J} \in \mathcal{H}^{(k)}_J \subseteq \mathcal{H}^{(k)}$, we have $0 = \langle \eta_{k,J}, \hat{\eta}^*_k - \eta_k\rangle_{\mathcal{H},k} = \langle \eta_{k,J}, \hat{\eta}^*_{k,J} - \eta_{k,J}\rangle_{\mathcal{H},k} = J_k(\eta_{k,J}, \hat{\eta}^*_{k,J} - \eta_{k,J})$ and $0 = \langle \hat{\eta}_{k,J}, \hat{\eta}^*_k - \eta_k\rangle_{\mathcal{H},k} = \langle \hat{\eta}_{k,J}, \hat{\eta}^*_{k,J} - \eta_{k,J}\rangle_{\mathcal{H},k} = J_k(\hat{\eta}_{k,J}, \hat{\eta}^*_{k,J} - \eta_{k,J})$.

Finally, we get $J_k(\eta_k, \hat{\eta}^*_k - \eta_k) = J_k(\eta_{k,N}, \hat{\eta}^*_{k,J} - \eta_{k,J}) + J_k(\eta_{k,J}, \hat{\eta}^*_{k,J} - \eta_{k,J}) = 0$ and $J_k(\hat{\eta}_k, \hat{\eta}^*_k - \eta_k) = J_k(\hat{\eta}_{k,N}, \hat{\eta}^*_{k,J} - \eta_{k,J}) + J_k(\hat{\eta}_{k,J}, \hat{\eta}^*_{k,J} - \eta_{k,J}) = 0$. ■

B.6 Details for Reproducing Kernel Hilbert Space

The RKHS considered here is constructed following the smoothing spline ANOVA from Gu [18]. The Sobolev space of order $m \in \mathbb{N}$ for $\mathcal{X}_{[j]}$ is given by

$$\mathcal{H}_{[j]} = \left\{ f : f, f^{(1)}, \ldots, f^{(m-1)} \text{ are absolutely continuous, and } \int_{\mathcal{X}_{[j]}} |f^{(m)}(x_{[j]})|^2 dx_{[j]} < \infty \right\}.$$

For $\mathcal{X}_{[j]}$, define $\kappa_{l[j]}(x) = B_l(x)/l!$ as the l-th scaled Bernoulli polynomial. For instance, $\kappa_{0[j]}(x) = 1$, $\kappa_{1[j]}(x) = x - 0.5$, $\kappa_{2[j]} = (\kappa_{1[j]}^2 - 1/12)/2$, and $\kappa_{4[j]} = (\kappa_{1[j]}^4 - \kappa_{1[j]}^2/2 + 7/240)/24$.

The space $\mathcal{H}_{[j]}$ is expressed as

$$\mathcal{H}_{[j]} = \oplus_{l=0}^{m-1} \mathcal{H}_{0l[j]} \oplus \mathcal{H}_{1[j]} = \mathcal{H}_{0[j]} \oplus \mathcal{H}_{1[j]} = \mathcal{H}_{\emptyset[j]} \oplus \mathcal{H}_{\{j\}},$$

where $\mathcal{H}_{0l[j]} = \text{Span}\{\kappa_{l[j]}\}$ is equipped with RK $\mathcal{K}_{0l[j]}(x_{[j]}, x'_{[j]}) = \kappa_{l[j]}(x_{[j]})\kappa_{l[j]}(x'_{[j]})$ and its paired inner product $\langle f, g\rangle_{\mathcal{H}_{0l[j]}} = \int_{\mathcal{X}_{[j]}} f^{(l)}(x_{[j]})dx_{[j]} \int_{\mathcal{X}_{[j]}} g^{(l)}(x_{[j]})dx_{[j]}$. In particular, $\mathcal{H}_{\emptyset[j]} = \mathcal{H}_{00[j]}$, and

$$\mathcal{H}_{0[j]} = \text{Span}\{\kappa_{0[j]}, \kappa_{1[j]}, \ldots, \kappa_{m-1[j]}\}$$

has RK $\mathcal{K}_{0[j]} = \sum_{l=0}^{m-1} \mathcal{K}_{0l[j]}$ and paired inner product $\langle \cdot, \cdot\rangle_{\mathcal{H}_{0[j]}} = \sum_{l=0}^{m-1} \langle \cdot, \cdot\rangle_{\mathcal{H}_{0l[j]}}$. The space $\mathcal{H}_{1[j]}$ is

$$\mathcal{H}_{1[j]} = \left\{ f \in \mathcal{H}_{[j]} : \int_{\mathcal{X}_{[j]}} f^{(l)}(x_{[j]})dx_{[j]} = 0 \text{ for } l = 0, 1, \ldots, m-1 \right\},$$

with RK $\mathcal{K}_{1[j]}(x_{[j]}, x'_{[j]}) = \kappa_{m[j]}(x_{[j]})\kappa_{m[j]}(x'_{[j]}) + (-1)^{m-1}\kappa_{2m[j]}(|x_{[j]} - x'_{[j]}|)$ and paired inner product $\langle f, g\rangle_{\mathcal{H}_{1[j]}} = \int_{\mathcal{X}_{[j]}} f^{(m)}(x_{[j]})g^{(m)}(x_{[j]})dx_{[j]}$. Therefore, $\mathcal{H}_{[j]}$ itself has RK $\mathcal{K}_{[j]} = \mathcal{K}_{0[j]} + \mathcal{K}_{1[j]}$ and corresponding inner product $\langle \cdot, \cdot\rangle_{\mathcal{H}_{[j]}} = \langle \cdot, \cdot\rangle_{\mathcal{H}_{0[j]}} + \langle \cdot, \cdot\rangle_{\mathcal{H}_{1[j]}}$. One can see that $\mathcal{H}_{0[j]}$ serves as the parametric space on the jth axis (polynomials of degree up to $m-1$), while $\mathcal{H}_{1[j]}$ forms the nonparametric space on the jth axis.

In addition, $\mathcal{H}_{\emptyset[j]} = \text{Span}\{1\}$ is the constant space of axis j, and

$$\mathcal{H}_{\{j\}} = \oplus_{l=1}^{m-1} \mathcal{H}_{0l[j]} \oplus \mathcal{H}_{1[j]}$$

is the effect space of axis j with its RK $\mathcal{K}_{\{j\}} = \sum_{l=1}^{m-1} \mathcal{K}_{0l[j]} + \mathcal{K}_{1[j]}$ and corresponding inner product $\langle \cdot, \cdot\rangle_{\mathcal{H}_{\{j\}}} = \sum_{l=1}^{m-1} \langle \cdot, \cdot\rangle_{\mathcal{H}_{0l[j]}} + \langle \cdot, \cdot\rangle_{\mathcal{H}_{1[j]}}$.

Let $\mathcal{A}_{[j]}$ be the averaging operator of the jth axis, defined by

$$\mathcal{A}_{[j]}f(x) = \mathcal{A}_{[j]}f(x_{[1]}, \ldots, x_{[d]}) = \int_{\mathcal{X}_{[j]}} f(x_{[1]}, \ldots, x_{[d]})dx_{[j]}.$$

Any $f_{[j]} \in \mathcal{H}_{[j]}$ can be decomposed as $f_{[j]} = f_{\emptyset[j]} + f_{\{j\}}$, with $f_{\emptyset[j]} = \mathcal{A}_{[j]}f_{[j]} \in \mathcal{H}_{\emptyset[j]}$ and $f_{\{j\}} = (id - \mathcal{A}_{[j]})f_{[j]} \in \mathcal{H}_{\{j\}}$, where id is the identity operator.

The tensor product Sobolev space across all d axes is

$$\begin{aligned}\otimes_{j=1}^{d}\mathcal{H}_{[j]} &= \otimes_{j=1}^{d}\left\{\mathcal{H}_{\emptyset[j]} \oplus \mathcal{H}_{\{j\}}\right\} \\ &= \oplus_{S\in\mathcal{P}_d}\mathcal{H}_S = \mathcal{H}_\emptyset \oplus \left\{\oplus_{j=1}^{d}\mathcal{H}_{\{j\}}\right\} \oplus \left\{\oplus_{j<j'}\mathcal{H}_{\{j,j'\}}\right\} \oplus \ldots \oplus \mathcal{H}_{\{1,\ldots,d\}},\end{aligned}$$

where $\mathcal{P}_d$ denotes the power set of $\{1, \ldots, d\}$, whose size is $|\mathcal{P}_d| = 2^d$. As an example, if $d = 3$, $\mathcal{P}_d = \{\emptyset, \{1\}, \{2\}, \{3\}, \{1, 2\}, \{1, 3\}, \{2, 3\}, \{1, 2, 3\}\}$. For each $S \in \mathcal{P}_d$, the associated space is

$$\mathcal{H}_S = \{\otimes_{j\in S}\mathcal{H}_{\{j\}}\} \otimes \{\otimes_{j\in\{1,\ldots,d\}\setminus S}\mathcal{H}_{\emptyset[j]}\} = \{\otimes_{j\in S}\mathcal{H}_{\{j\}}\} \otimes \text{Span}\{1\} = \otimes_{j\in S}\mathcal{H}_{\{j\}}.$$

Note that $\mathcal{H}_S$ is defined over $\mathcal{X}_S = \prod_{j\in S}\mathcal{X}_{[j]}$ and represents the space for effect S. The constant part, i.e., the intercept space, is $\mathcal{H}_\emptyset = \otimes_{j=1}^{d}\mathcal{H}_{\emptyset[j]} = \text{Span}\{1\}$. Any $f \in \otimes_{j=1}^{d}\mathcal{H}_{[j]}$ is decomposed by

$$f = \sum_{S\in\mathcal{P}_d} f_S, \text{ where } f_S = \left\{\prod_{j\in S}(id - \mathcal{A}_{[j]}) \prod_{j\in\{1,\ldots,d\}\setminus S}\mathcal{A}_{[j]}\right\} f \in \mathcal{H}_S.$$

For each $S \in \mathcal{P}_d \setminus \{\emptyset\}$, $\mathcal{K}_S = \prod_{j\in S}\mathcal{K}_{\{j\}}$ is the RK for $\mathcal{H}_S$, with paired inner product $\langle\cdot,\cdot\rangle_{\mathcal{H}_S}$.

The subspace

$$\mathcal{N}_{J,S} = \otimes_{j\in S}\left\{\oplus_{l_j=1}^{m-1}\mathcal{H}_{0l_j[j]}\right\} = \mathrm{Span}\left\{\left\{\prod_{j\in S}\kappa_{l_j[j]} : l_j \in \{1,\ldots,m-1\} \text{ for } j \in S\right\}\right\}$$

of $\mathcal{H}_S$ has its RK $\mathcal{K}_{N,S} = \prod_{j\in S}\left\{\sum_{l_j=1}^{m-1}\mathcal{K}_{0l_j[j]}\right\}$ and paired inner product $\langle\cdot,\cdot\rangle_{\mathcal{N}_{J,S}}$, while

$$\mathcal{H}_{J,S} = \mathcal{H}_S \ominus \mathcal{N}_{J,S}$$

has its RK $\mathcal{K}_{J,S} = \mathcal{K}_S - \mathcal{K}_{N,S}$ and corresponding inner product $J_S(\cdot,\cdot) = \langle\cdot,\cdot\rangle_{\mathcal{H}_{J,S}}$, so $\langle\cdot,\cdot\rangle_{\mathcal{H}_S} = \langle\cdot,\cdot\rangle_{\mathcal{N}_{J,S}} + \langle\cdot,\cdot\rangle_{\mathcal{H}_{J,S}}$. Intuitively, $\mathcal{N}_{J,S}$ collects the subspaces where all axes are parametric, while $\mathcal{H}_{J,S}$ contains those with at least one nonparametric axis. The constant space $\mathcal{H}_\emptyset$ has RK $\mathcal{K}_\emptyset = 1$ and its paired inner product $\langle\cdot,\cdot\rangle_{\mathcal{H}_\emptyset}$.

Let

$$\mathcal{H} = \oplus_{S\in\mathbb{S}}\mathcal{H}_S$$

for some collection $\mathbb{S}$ satisfying $\emptyset \notin \mathbb{S}$ and $\mathbb{S} \subseteq \mathcal{P}_d$. As a result, $\int_{\mathcal{X}} f(x)\,dx = 0$ for all $f \in \mathcal{H}$. The selection of $\mathbb{S}$ specifies the structure of $\mathcal{H}$. For instance, if $\mathbb{S} = \{\{1\},\ldots,\{d\}\}$, then $\mathcal{H}$ adopts an additive form that involves only main effects. On the other hand, choosing $\mathbb{S} = \{\{1\},\ldots,\{d\},\{1,2\},\ldots,\{d-1,d\}\}$ allows $\mathcal{H}$ to incorporate all main effects together with two-factor interactions.

We can write

$$\mathcal{H} = \mathcal{N}_J \oplus \mathcal{H}_J,$$

where the parametric space

$$\mathcal{N}_J = \oplus_{S\in\mathbb{S}}\mathcal{N}_{J,S} = \mathrm{Span}\{\phi_1,\ldots,\phi_p\}$$

with RK $\mathcal{K}_N = \sum_{S \in \mathbb{S}} \mathcal{K}_{N,S}$ and its paired inner product $\langle \cdot, \cdot \rangle_{\mathcal{N}_J} = \sum_{S \in \mathbb{S}} \langle \cdot, \cdot \rangle_{\mathcal{N}_{J,S}}$; and the nonparametric space

$$\mathcal{H}_J = \oplus_{S \in \mathbb{S}} \mathcal{H}_{J,S}$$

with RK $\mathcal{K}_J = \sum_{S \in \mathbb{S}} \mathcal{K}_{J,S}$ and its paired inner product $J(\cdot, \cdot) = \langle \cdot, \cdot \rangle_{\mathcal{H}_J} = \sum_{S \in \mathbb{S}} \langle \cdot, \cdot \rangle_{\mathcal{H}_{J,S}}$. Consequently, the RK and its corresponding inner product on $\mathcal{H}$ are $\mathcal{K} = \mathcal{K}_N + \mathcal{K}_J = \sum_{S \in \mathbb{S}} \mathcal{K}_S$ and $\langle \cdot, \cdot \rangle_{\mathcal{H}} = \langle \cdot, \cdot \rangle_{\mathcal{N}_J} + \langle \cdot, \cdot \rangle_{\mathcal{H}_J} = \sum_{S \in \mathbb{S}} \langle \cdot, \cdot \rangle_{\mathcal{H}_S}$, respectively.

The following is an example when $d = 2$, $m = 2$, and an interaction effect is considered. For $l, l' \in \{00, 01, 1\}$, define

$$\mathcal{H}_{l,l'} = \mathcal{H}_{l[1]} \otimes \mathcal{H}_{l'[2]}$$

with its RK and paired inner product $\mathcal{K}_{l,l'} = \mathcal{K}_{l[1]} \times \mathcal{K}_{l'[2]}$ and $\langle \cdot, \cdot \rangle_{\mathcal{H}_{l,l'}}$. Note that $\mathcal{H}_{l,l'} = \mathrm{Span}\{\phi_{l,l'}\}$ for $l, l' \in \{00, 01\}$, where $\phi_{0l,0l'} = \kappa_{l[1]} \kappa_{l'[2]}$ for $l, l' \in \{0, 1\}$.

We have

$$\otimes_{j=1}^{2} \{\mathcal{H}_{00[j]} \oplus \mathcal{H}_{01[j]} \oplus \mathcal{H}_{1[j]}\} = \mathcal{H}_{\emptyset} \oplus \mathcal{H}_{\{1\}} \oplus \mathcal{H}_{\{2\}} \oplus \mathcal{H}_{\{1,2\}} = \mathcal{H}_{\emptyset} \oplus \mathcal{H} = \mathcal{H}_{\emptyset} \oplus \mathcal{N}_J \oplus \mathcal{H}_J,$$

where $\mathcal{H}_{\emptyset} = \mathcal{H}_{00,00}$ is the intercept space, $\mathcal{H}_{\{1\}} = \mathcal{H}_{01,00} \oplus \mathcal{H}_{1,00}$ is the main effect space for the 1st axis, $\mathcal{H}_{\{2\}} = \mathcal{H}_{00,01} \oplus \mathcal{H}_{00,1}$ is the main effect space for the 2nd axis, $\mathcal{H}_{\{1,2\}} = \mathcal{H}_{01,01} \oplus \mathcal{H}_{01,1} \oplus \mathcal{H}_{1,01} \oplus \mathcal{H}_{1,1}$ is the interaction space between the 1st and 2nd axes,

$$\mathcal{H}_{00,00} \oplus \mathcal{H}_{00,01} \oplus \mathcal{H}_{01,00} \oplus \mathcal{H}_{01,01} = \mathrm{Span}\{\phi_{00,00}, \phi_{00,01}, \phi_{01,00}, \phi_{01,01}\}$$

is the parametric space,

$$\mathcal{N}_J = \mathcal{H}_{00,01} \oplus \mathcal{H}_{01,00} \oplus \mathcal{H}_{01,01} = \mathrm{Span}\{\phi_{00,01}, \phi_{01,00}, \phi_{01,01}\}$$

is the parametric space without intercept, and

$$\mathcal{H}_J = \mathcal{H}_{00,1} \oplus \mathcal{H}_{01,1} \oplus \mathcal{H}_{1,00} \oplus \mathcal{H}_{1,01} \oplus \mathcal{H}_{1,1}$$

is the nonparametric space.

One can show that

$$\begin{aligned}
\langle f, g \rangle_{\mathcal{H}_{00,00}} &= \int_{\mathcal{X}} f(x)dx \int_{\mathcal{X}} g(x)dx, \\
\langle f, g \rangle_{\mathcal{H}_{01,00}} &= \int_{\mathcal{X}} f^{(1)}_{\langle 1 \rangle}(x)dx \int_{\mathcal{X}} g^{(1)}_{\langle 1 \rangle}(x)dx, \\
\langle f, g \rangle_{\mathcal{H}_{01,01}} &= \int_{\mathcal{X}} f^{(2)}_{\langle 12 \rangle}(x)dx \int_{\mathcal{X}} g^{(2)}_{\langle 12 \rangle}(x)dx, \\
\langle f, g \rangle_{\mathcal{H}_{1,00}} &= \int_{\mathcal{X}_{[1]}} \{\int_{\mathcal{X}_{[2]}} f^{(2)}_{\langle 11 \rangle}(x)dx_{[2]} \int_{\mathcal{X}_{[2]}} g^{(2)}_{\langle 11 \rangle}(x)dx_{[2]}\}dx_{[1]}, \\
\langle f, g \rangle_{\mathcal{H}_{1,01}} &= \int_{\mathcal{X}_{[1]}} \{\int_{\mathcal{X}_{[2]}} f^{(3)}_{\langle 112 \rangle}(x)dx_{[2]} \int_{\mathcal{X}_{[2]}} g^{(3)}_{\langle 112 \rangle}(x)dx_{[2]}\}dx_{[1]}, \\
\langle f, g \rangle_{\mathcal{H}_{1,1}} &= \int_{\mathcal{X}} f^{(4)}_{\langle 1122 \rangle}(x) g^{(4)}_{\langle 1122 \rangle}(x)dx,
\end{aligned}$$

where $f^{(1)}_{\langle 1 \rangle}(x) = \partial f(x)/(\partial x_{[1]})$, $f^{(2)}_{\langle 12 \rangle}(x) = \partial^2 f(x)/(\partial x_{[1]} \partial x_{[2]})$, $f^{(2)}_{\langle 11 \rangle}(x) = \partial^2 f(x)/(\partial x^2_{[1]})$, $f^{(3)}_{\langle 112 \rangle}(x) = \partial^3 f(x)/(\partial x^2_{[1]} \partial x_{[2]})$, and $f^{(4)}_{\langle 1122 \rangle}(x) = \partial^4 f(x)/(\partial x^2_{[1]} \partial x^2_{[2]})$.

In practice, for each k, for some $\theta_{l,l'(k)} \in (0, \infty)$, $\{\langle \cdot, \cdot \rangle_{\mathcal{H}_{l,l'}}/\theta_{l,l'(k)}, \theta_{l,l'(k)} \mathcal{K}_{l,l'}\}$ forms an equivalent pair of inner product and RK with $\{\langle \cdot, \cdot \rangle_{\mathcal{H}_{l,l'}}, \mathcal{K}_{l,l'}\}$ on $\mathcal{H}_{l,l'}$. One can regard $\theta_{l,l'(k)}$ as weight parameter for $\mathcal{H}_{l,l'}$. We set $\theta_{l,l'(k)} = 1$ for $l, l' \in \{00, 01\}$.

Then, we have RK and paired inner product for $\mathcal{N}_J$ as

$$\mathcal{K}_N = \mathcal{K}_{00,01} + \mathcal{K}_{01,00} + \mathcal{K}_{01,01}$$

and

$$\langle\cdot,\cdot\rangle_{\mathcal{N}_J} = \langle\cdot,\cdot\rangle_{\mathcal{H}_{00,01}} + \langle\cdot,\cdot\rangle_{\mathcal{H}_{01,00}} + \langle\cdot,\cdot\rangle_{\mathcal{H}_{01,01}}.$$

Similarly, for each k, RK and paired inner product for $\mathcal{H}_J$ are

$$\mathcal{K}_{J_k} = \theta_{00,1(k)}\mathcal{K}_{00,1} + \theta_{01,1(k)}\mathcal{K}_{01,1} + \theta_{1,00(k)}\mathcal{K}_{1,00} + \theta_{1,01(k)}\mathcal{K}_{1,01} + \theta_{1,1(k)}\mathcal{K}_{1,1}$$

and

$$J_k(\cdot,\cdot) = \frac{\langle\cdot,\cdot\rangle_{\mathcal{H}_{00,1}}}{\theta_{00,1(k)}} + \frac{\langle\cdot,\cdot\rangle_{\mathcal{H}_{01,1}}}{\theta_{01,1(k)}} + \frac{\langle\cdot,\cdot\rangle_{\mathcal{H}_{1,00}}}{\theta_{1,00(k)}} + \frac{\langle\cdot,\cdot\rangle_{\mathcal{H}_{1,01}}}{\theta_{1,01(k)}} + \frac{\langle\cdot,\cdot\rangle_{\mathcal{H}_{1,1}}}{\theta_{1,1(k)}},$$

which leads to RK and paired inner product for $\mathcal{H}$ being $\mathcal{K}_k = \mathcal{K}_N + \mathcal{K}_{J_k}$ and $\langle\cdot,\cdot\rangle_{\mathcal{H},k} = \langle\cdot,\cdot\rangle_{\mathcal{N}_J} + J_k(\cdot,\cdot)$. When we set $\theta_{l,l'(k)} = 1$ for $l, l' \in \{00, 01, 1\}$, we have $\mathcal{K}_{J_k} = \mathcal{K}_J$, $J_k = J$, $\mathcal{K}_k = \mathcal{K}$, and $\langle\cdot,\cdot\rangle_{\mathcal{H},k} = \langle\cdot,\cdot\rangle_{\mathcal{H}}$.

The true function $\eta_k^* \in \mathcal{H}$ is decomposed as

$$\eta_k^* = \eta_{k,\{1\}}^* + \eta_{k,\{2\}}^* + \eta_{k,\{1,2\}}^* = \eta_{k,N}^* + \eta_{k,J}^*$$

with the 1st axis main effect $\eta_{k,\{1\}}^* = \eta_{k,01,00}^* + \eta_{k,1,00}^* \in \mathcal{H}_{\{1\}}$, 2nd axis main effect $\eta_{k,\{2\}}^* = \eta_{k,00,01}^* + \eta_{k,00,1}^* \in \mathcal{H}_{\{2\}}$, interaction between the 1st and 2nd axes $\eta_{k,\{1,2\}}^* = \eta_{k,01,01}^* + \eta_{k,01,1}^* + \eta_{k,1,01}^* + \eta_{k,1,1}^* \in \mathcal{H}_{\{1,2\}}$, parametric effect

$$\eta_{k,N}^* = \eta_{k,00,01}^* + \eta_{k,01,00}^* + \eta_{k,01,01}^* \in \mathcal{N}_J,$$

and nonparametric effect

$$\eta_{k,J}^* = \eta_{k,00,1}^* + \eta_{k,01,1}^* + \eta_{k,1,00}^* + \eta_{k,1,01}^* + \eta_{k,1,1}^* \in \mathcal{H}_J,$$

where $\eta^*_{k,l,l'} \in \mathcal{H}_{l,l'}$ for $l, l' \in \{00, 01, 1\}$.

Similarly, the estimator $\hat{\eta}_k$ having the following form

$$\hat{\eta}_k = \hat{d}_{k,00,01}\phi_{00,01} + \hat{d}_{k,01,00}\phi_{01,00} + \hat{d}_{k,01,01}\phi_{01,01} + \sum_{\ell=1}^{q_k} \hat{c}_{k,\ell}\mathcal{K}_{J_k}(Z_{\ell(k)}, \cdot) \in \mathcal{N}_J \oplus \mathcal{H}_J^{(k)} = \mathcal{H}^{(k)} \subseteq \mathcal{H}$$

is decomposed as

$$\hat{\eta}_k = \hat{\eta}_{k,\{1\}} + \hat{\eta}_{k,\{2\}} + \hat{\eta}_{k,\{1,2\}} = \hat{\eta}_{k,N} + \hat{\eta}_{k,J}$$

with the 1st axis main effect $\hat{\eta}_{k,\{1\}} = \hat{\eta}_{k,01,00} + \hat{\eta}_{k,1,00} \in \mathcal{H}_{\{1\}}$, 2nd axis main effect $\hat{\eta}_{k,\{2\}} = \hat{\eta}_{k,00,01} + \hat{\eta}_{k,00,1} \in \mathcal{H}_{\{2\}}$, interaction between the 1st and 2nd axes $\hat{\eta}_{k,\{1,2\}} = \hat{\eta}_{k,01,01} + \hat{\eta}_{k,01,1} + \hat{\eta}_{k,1,01} + \hat{\eta}_{k,1,1} \in \mathcal{H}_{\{1,2\}}$, parametric effect

$$\hat{\eta}_{k,N} = \hat{\eta}_{k,00,01} + \hat{\eta}_{k,01,00} + \hat{\eta}_{k,01,01} \in \mathcal{N}_J,$$

and nonparametric effect

$$\hat{\eta}_{k,J} = \hat{\eta}_{k,00,1} + \hat{\eta}_{k,01,1} + \hat{\eta}_{k,1,00} + \hat{\eta}_{k,1,01} + \hat{\eta}_{k,1,1} \in \mathcal{H}_J^{(k)} \subseteq \mathcal{H}_J,$$

where $\hat{\eta}_{k,l,l'} = \hat{d}_{k,l,l'}\phi_{l,l'} \in \mathcal{H}_{l,l'}$ for $l, l' \in \{00, 01\}$ and

$$\hat{\eta}_{k,l,l'} = \sum_{\ell=1}^{q_k} \hat{c}_{k,\ell}\theta_{l,l'(k)}\mathcal{K}_{l,l'}(Z_{\ell(k)}, \cdot) \in \text{Span}\left\{\mathcal{K}_{l,l'}(Z_{1(k)}, \cdot), \ldots, \mathcal{K}_{l,l'}(Z_{q_k(k)}, \cdot)\right\} \subseteq \mathcal{H}_{l,l'}$$

for $(l, l') \in \{(00, 1), (01, 1), (1, 00), (1, 01), (1, 1)\}$.

For other d or other m, or any other structures of $\mathcal{H}$, for example, additive structure, one can do similarly like above.