Journal of Machine Learning Research 27 (2026) 1-63 Submitted 11/23; Revised 6/25; Published 1/26

Exploring Novel Uncertainty Quantification through Forward Intensity Function Modeling

Yudong Wang YUDONGW@U.NUS.EDU
Department of Industrial Systems Engineering and Management
National University of Singapore
1 Engineering Drive 2, Singapore 117576

Zhi-Sheng Ye* YEZ@NUS.EDU.SG
Department of Industrial Systems Engineering and Management
National University of Singapore
1 Engineering Drive 2, Singapore 117576

Cheng Yong Tang YONGTANG@TEMPLE.EDU
Department of Statistics, Operations, and Data Science
Temple University
1810 Liacouras Walk, Philadelphia, Pennsylvania 19122-6083, U.S.A.

Editor: Eric Laber

Abstract

Predicting future time-to-event outcomes is a foundational task in statistical learning. While various methods exist for generating point predictions, quantifying the associated uncertainties poses a more substantial challenge. In this study, we introduce an innovative approach specifically designed to address this challenge, accommodating dynamic predictors that may manifest as stochastic processes. Our investigation harnesses the forward intensity function in a novel way, providing a fresh perspective on this intricate problem. The framework we propose demonstrates remarkable computational efficiency, enabling efficient analyses of large-scale investigations. We validate its soundness with theoretical guarantees, and our in-depth analysis establishes the weak convergence of function-valued parameter estimations. We illustrate the effectiveness of our framework with two comprehensive real examples and extensive simulation studies.

Keywords: forward intensity, functional-valued parameter, interval predictions, time-to-event analysis, uncertainty quantification

1. Introduction

A fundamental task in learning time-to-event is the prediction of future event occurrences. Successfully accomplishing this task necessitates a comprehensive analysis of the predictive distribution of the time until a future event for an at-risk subject. This analysis takes into account available information such as the subject's event history and a set of covariates, which can either remain fixed or evolve as stochastic processes over time. In fields such as precision medicine, understanding the predictive distribution of event times for patients whose health conditions dynamically change over time is crucial for healthcare providers.

*. Corresponding author

This information empowers physicians to make personalized decisions for patient care. Essentially, the differences in predictive distributions among patients highlight the inherent heterogeneity that underpins personalized treatment choices.

Models for predicting event times have been extensively studied in the literature. Recently, there has been active development in learning, modeling, and predicting time-to-event data. This includes the "individual survival model" approach proposed by Haider et al. (2020), the model averaging approach introduced by He et al. (2020), the use of ordinary differential equations as applied by Tang et al. (2022a,b), and the utilization of neural network-based approaches as explored by Hu and Nan (2023). An established framework simultaneously models both the event time and the longitudinal covariate processes, as demonstrated in existing studies (Rizopoulos, 2012; Taylor et al., 2013; Rizopoulos et al., 2014; Elashoff et al., 2016). This "joint modeling" framework typically consists of a combination of a linear mixed-effects submodel for the longitudinal covariate process and a Cox-type regression submodel for survival times that shares the same random effects as the covariate process. The random effects are usually collectively modeled using a multivariate normal distribution, resulting in a likelihood function that necessitates multidimensional numerical integration (Wulfsohn and Tsiatis, 1997). This computational complexity can lead to high costs in joint modeling approaches, especially when dealing with a large dimensionality of time-varying covariates (Zhu et al., 2019). Another class of approaches focuses on modeling the future hazard rate function of an at-risk subject, conditioned on data information up to a specific time point. Such a time point is referred to as the "landmark time" in the literature; these approaches are often referred to as the "landmark analysis" (van Houwelingen, 2007; van Houwelingen and Putter, 2011; Zhu et al., 2019) or the "partly conditional model" (Zheng and Heagerty, 2005; Gong and Schaubel, 2017). In the literature on default predictions, in recognition of the forward-looking nature, the future hazard rate is also called the "forward intensity function" (Chen, 2007; Duan et al., 2012). As a striking feature, this class of approaches does not require explicitly modeling the covariate process; see Maziarz et al. (2017) for comparisons.

Remarkably, existing approaches are mostly used in developing point predictions. Meanwhile, constructing interval predictions—a task relevant to evaluating the associated uncertainties—is more meaningful and interesting, capable of revealing the trustworthiness of the discoveries. In the interim analysis of clinical trials, for example, prediction intervals of the survival times of enrolled patients can provide stronger decision support for the planning of subsequent duration and enrollment (Heitjan et al., 2015). In insurance systems, a prediction interval for the claim probability of a customer based on the claim history is helpful for customized pricing of an insurance product. Moreover, a prediction interval of the aggregate claims in the next fiscal year for a pool of customers is crucial for claim reserve management (Fredette and Lawless, 2007).

Notwithstanding the importance, computationally efficient procedures are less available for the interval predictions. In the joint modeling framework, interval estimations of the predicted survival probabilities can be constructed if the parameters are estimated using the Bayesian approach with the help of a Markov chain Monte Carlo (MCMC) algorithm (Wang and Taylor, 2001; Taylor et al., 2013; Wang et al., 2017). Clearly, the MCMC is computationally intensive, especially with large sample sizes and complicated random effects. When the dimensionality of the time-varying covariates is large, MCMC approaches

may encounter additional formidable challenges in both practical implementation and theoretical analysis. Similar computational challenges are also faced by another commonly used approach for interval estimations, the bootstrap. Recently, Yuan et al. (2018) constructed bootstrap-based confidence intervals for the predictive corporate default probability using a joint modeling approach. Nevertheless, their algorithm is computationally demanding. Moreover, no theoretical guarantee has been established for the resulting interval estimators using the bootstrap approach. In the aforementioned landmark analysis framework, interval estimation is also commonly done using the bootstrap (Ferrer et al., 2018; Wu et al., 2020), while its statistical properties, such as consistency, remain unknown.

In this study, we investigate a new framework for synthetically quantifying the uncertainties associated with the time-to-event predictions. As a foremost contribution, we approach the forward intensity function from a novel perspective. Specifically, we discover a new path that bridges the the forward intensity function and the probability distribution of the resulting predicted quantities. Our new framework is practically appealing; facilitated by the new path, quantifying the uncertainties of the point predictions becomes not only feasible but also computationally highly efficient. Unlike the MCMC and bootstrap approaches with tedious iterative loops, our new approaches essentially require no additional computational step besides those entailed in obtaining the point predictions. As a remark, the resampling method in Yuan et al. (2018) takes around 160 hours on a server with 80 CPUs. In contrast, it can be accomplished within one hour using a PC by applying our method on a data set with similar scale.

Furthermore, our theoretical investigation ensures the validity of our framework with a theoretical guarantee. Upon substantial and careful analysis that fully exploits the special structure of the forward intensity modeling device, we derive the asymptotic joint distribution of the estimators of the functional-valued model parameters. Remarkably, the asymptotic joint distribution of the estimators carries contributions from each time point prior to the landmark time, resulting in a complicated infinite-dimensional problem. We dedicatedly device the empirical process techniques in establishing the challenging weak convergence, as well as the important results that the asymptotic variances of the predictive quantities can be consistently estimated with closed forms.

The rest of the paper is organized as follows. Section 2 introduces the forward intensity model and discusses the challenges in uncertainty quantification. Section 3 investigates the estimation of the functional-valued parameters in the forward intensity function and establishes the weak convergence of the developed estimators; we provide detailed procedures to quantify the uncertainties associated with parameter estimation and future event prediction. Two real examples are analyzed in Section 4. A simulation study is conducted in Section 5 to examine the performance of the proposed method, and Section 6 concludes.

2. The Forward Intensity Model

2.1 Forward Intensity

We consider a time-to-event $T \geq 0$ and a p-dimensional covariate process $\{\boldsymbol{Z}(t) : t \geq 0\}$ that are defined on a probability space $(\Omega, \mathcal{F}, P)$. Given that a subject has survived up to t that is referred to as the landmark time, i.e., $T \geq t$, we are interested in the forward-looking predictive distribution of T. To this end, let $\{\mathcal{F}_t : t \geq 0\}$ denote the filtration that presents

the history of the subject up to time t. At time t, the forward intensity evaluated at a future time epoch $t+u$, denoted as $\lambda_{T,t}(u)$, is defined as

$$\lambda_{T,t}(u) \triangleq \lim_{\Delta u \downarrow 0} \frac{P(t+u \leq T < t+u+\Delta u | \mathcal{F}_t, T \geq t)}{\Delta u \; P(T \geq t+u | \mathcal{F}_t, T \geq t)}.$$

Based on $\lambda_{T,t}(u)$, the predictive survival probability at $t+\tau$ can be computed as

$$S_{T,t}(\tau) \triangleq P(T \geq t+\tau | \mathcal{F}_t, T \geq t) = \exp\left\{ - \int_0^\tau \lambda_{T,t}(u) du \right\}. \tag{1}$$

Note that if all covariates $\boldsymbol{Z}$ are time-invarying, then $\lambda_{T,t}(u)$ is the hazard rate of $[T|\boldsymbol{Z}]$ at time $t+u$. As a result, standard tools such as the classical Cox or varying-coefficient Cox models can be used to model the hazard rate of $[T|\boldsymbol{Z}]$. When the covariates are time-varying as a dynamic stochastic process, however, no standard model is available. In the context of the forward intensity function, one may directly impose a model on $\lambda_{T,t}(u)$. A plausible and simple model proposed by Duan et al. (2012) uses a log-linear model for forward intensity:

$$\lambda_{T,t}(u) = \exp\{\boldsymbol{\alpha}(u)^\top \boldsymbol{Z}(t)\}, \tag{2}$$

where $\boldsymbol{\alpha} : \mathbb{R}^+ \mapsto \mathbb{R}^p$ is the functional-valued parameter that captures the future dynamics of the unobserved covariates. Duan et al. (2012) further studied the point estimation of this model under a discrete time scale.

In survival analysis, it is common that the lifetime is subject to a random right-censoring time C, which can be lost follow-up, participant withdrawal, or competing risk events. We shall highlight that the censoring time C here does not include contributions from administrative censoring (or end-of-study censoring), because administrative censoring only affects parameter estimation, and it is irrelevant in the event-time prediction. Since we are interested in predicting the future time, a model for the right-censoring time is needed. We assume that for any landmark time $t > 0$, T and C are conditionally independent given $\mathcal{F}_t$ and $T \wedge C \geq t$. That is, the time-to-event T and the censoring time C for an at-risk subject are independent conditional on the data information up to the landmark time t. This assumption is similar to the noninformative censoring in survival analysis (Zeng and Lin, 2007). Further, we model the forward intensity of C as

$$\lambda_{C,t}(u) = \exp\{\boldsymbol{\beta}(u)^\top \boldsymbol{Z}(t)\}, \tag{3}$$

where $\boldsymbol{\beta} : \mathbb{R}^+ \mapsto \mathbb{R}^p$ is again a functional-valued regression parameter. Though the covariates incorporated in $\lambda_{T,t}(u)$ and $\lambda_{C,t}(u)$ can differ, for clarity in our presentation we use the same $\boldsymbol{Z}(t)$ for both T and C without loss of generality.

In the presence of the random censoring time C, the probability of observing the time-to-event T from t to $t+\tau$ can be evaluated by

$$F_t(\tau) \triangleq P(T < t+\tau, T < C | \mathcal{F}_t, T \wedge C \geq t) = \int_0^\tau S_{T,t}(u) S_{C,t}(u) \lambda_{T,t}(u) du, \tag{4}$$

where $S_{T,t}(u)$ and $S_{C,t}(u)$ are computed using (1) based on $\lambda_{T,t}$ and $\lambda_{C,t}$ in (2) and (3). The product $S_{T,t}(u)S_{C,t}(u)$ in the integrand of (4) reflects that no event or censoring occurs from the prediction origin t to time $t+u$.

2.2 A Promising Path to Uncertainty Quantification

We commence by exploring the inherent connection between the forward intensity function and predictive survival probabilities, opening up a promising avenue for quantifying uncertainty. To illustrate, we examine the forward-looking survival function, as described in (1). Suppose that we have $\hat{\lambda}_{T,t}(\cdot)$ as an estimator of the forward intensity function. Then by plugging it in (1), we have the estimator $\hat{S}_{T,t}(\tau)$ of $S_{T,t}(\tau)$. Since $-\log\{\hat{S}_{T,t}(\tau)\} = \int_0^\tau \hat{\lambda}_{T,t}(u)du$, we have that under mild conditions (i.e., existence of the second moments),

$$\mathrm{Var}[-\log\{\hat{S}_{T,t}(\tau)\}] = \int_0^\tau \int_0^\tau \mathrm{cov}\{\hat{\lambda}_{T,t}(u), \hat{\lambda}_{T,t}(s)\}duds. \tag{5}$$

This suggests that the key component in quantifying the variance associated with the predictions is the covariance function $\mathrm{cov}\{\hat{\lambda}_{T,t}(u), \hat{\lambda}_{T,t}(s)\}$. Once (5) is successfully evaluated, the uncertainties associated with the predictive survival probability can be quantified through the variance. If we further establish the weak convergence of $\hat{\lambda}_{T,t}(\cdot)$, the limit distribution of $\hat{S}_{T,t}(\tau)$ can be derived through the functional delta method, and uncertainty quantification with confidence intervals for $\hat{S}(\tau)$ becomes practically feasible.

While this approach holds promise, it faces notable challenges. Firstly, the interdependence between $\hat{\lambda}_{T,t}(u)$ and $\hat{\lambda}_{T,t}(s)$ $(0 < u, s < \tau)$ is expected to be intricate. Moreover, both u and s span a continuous range, reflecting that the ultimate prediction and its associated uncertainty stem from aggregating dynamic features at all time points before τ. Consequently, the estimation uncertainties in $\hat{\lambda}_{T,t}(u)$ for all $u \leq \tau$ contribute to the prediction's variance. Additionally, achieving weak convergence of $\hat{\lambda}_{T,t}(\cdot)$ necessitates uniform convergence to apply the functional delta method. Clearly, the covariance function and the weak convergence of $\hat{\lambda}_{T,t}(\cdot)$ depend on both the modeling and estimation approaches. Compared to pointwise properties for $\hat{\lambda}_{T,t}(u)$ with a given u, uniform weak convergence of $\hat{\lambda}_{T,t}(\cdot)$ is substantially more challenging and remains unexplored in the literature. In this study, we specifically investigate the log-linear model in (2), where the asymptotic properties of $\hat{\lambda}_{T,t}(\cdot)$ are determined by the estimated parameter $\hat{\boldsymbol{\alpha}}_n(\cdot)$.

3. Methodology

3.1 Data and Estimation

Additionally, alongside the event time T and the random censoring time C, there exists an administrative censoring time $\tilde{C}$, which represents the difference between the subject's study entry date and the end-of-study date when data collection concludes. The administrative censoring time $\tilde{C}$, known for all subjects, is a consequence of the finite data collection window, so it does not play a role in the predictive distribution. (4). Observable data for each subject include $X = \min(T, C, \tilde{C})$, $\delta_T = I\{T \leq X\}$ and $\delta_C = I\{C \leq X\}$, and measurements of the covariates $\boldsymbol{Z}(t)$ up to time X. We denote the observed data for a generic subject as $D \triangleq \{X, \delta_T, \delta_C, \boldsymbol{Z}(t); 0 \leq t \leq X\}$. It is worth noting that the complete likelihood function for $\boldsymbol{\alpha}(\cdot)$ and $\boldsymbol{\beta}(\cdot)$ based on D is generally unavailable due to the complex interdependence between likelihoods at different landmark times. As a convenient alternative, we employ a composite likelihood (Varin et al., 2011). In this context, we consider a forward-looking period of time with length u $(u < X)$ and treat $X - u$ as the landmark time. Upon fixing

u, the contribution from D to the log-likelihood of $\boldsymbol{\alpha}(\cdot)$ and $\boldsymbol{\beta}(\cdot)$ is

$$\begin{aligned}
&\ell\big(\alpha(\cdot), \boldsymbol{\beta}(\cdot); u, D\big) \\
=&\log\Big[S_{T,X-u}(u)\{\lambda_{T,X-u}(u)\}^{\delta_T} S_{C,X-u}(u)\{\lambda_{C,X-u}(u)\}^{\delta_C}\Big] \\
=&\delta_T \log\lambda_{T,X-u}(u) + \delta_C \log\lambda_{C,X-u}(u) - \int_0^u \{\lambda_{T,X-u}(s) + \lambda_{C,X-u}(s)\}\, ds.
\end{aligned}$$

Next, by aggregating the contribution from all u through integration, i.e., $\int_0^X \ell\big(\alpha(\cdot), \boldsymbol{\beta}(\cdot); u, D\big)du$, we obtain the composite log-likelihood:

$$\begin{aligned}
&\ell\big(\alpha(\cdot), \boldsymbol{\beta}(\cdot); D\big) = \int_0^X \ell\big(\alpha(\cdot), \boldsymbol{\beta}(\cdot); u, D\big)du \qquad (6)\\
=&\int_0^X \left[\delta_T \log\lambda_{T,X-u}(u) + \delta_C \log\lambda_{C,X-u}(u)\right] du - \int_0^X \int_0^u \{\lambda_{T,X-u}(s) + \lambda_{C,X-u}(s)\}\, dsdu.
\end{aligned}$$

By switching the order of integration and a change of variable $u = X + s - t$, we have

$$\int_0^X \int_0^u \lambda_{T,X-u}(s)dsdu = \int_0^X \int_s^X \lambda_{T,X-u}(s)duds = \int_0^X \int_s^X \lambda_{T,t-s}(s)dtds.$$

Let $N_T(t) = I\{T \le t\}$ and $N_C(t) = I\{C \le t\}$ be the counting processes associated with T and C, respectively. Denote $Y(t) = I\{t \le X\}$. Then, (6) can be written as

$$\begin{aligned}
&\ell\big(\alpha(\cdot), \boldsymbol{\beta}(\cdot); D\big) \\
=&\int_0^\infty Y(u)\Big[\delta_T \log\lambda_{T,X-u}(u) + \delta_C \log\lambda_{C,X-u}(u) - \int_u^X [\lambda_{T,t-u}(u) + \lambda_{C,t-u}(u)]dt\Big] du \qquad (7)\\
\triangleq&\int_0^\infty \left[\ell(\boldsymbol{\alpha}(u); D) + \ell(\boldsymbol{\beta}(u); D)\right] du,
\end{aligned}$$

where $\ell(\boldsymbol{\alpha}(u); D) = Y(u)\int_u^X \big[\log\lambda_{T,t-u}(u)dN_T(t) - \lambda_{T,t-u}(u)dt\big]$, and $\ell(\boldsymbol{\beta}(u); D)$ is defined similarly, replacing T with C. Here, (7) implies that composite log-likelihood can be decomposed into two parts: one involving $\boldsymbol{\alpha}(\cdot)$ only from the contributions of T and the other involving $\boldsymbol{\beta}(\cdot)$ only from the contributions of C.

Let $\mathbb{P}_n$ be the empirical measure based on n independent and identically distributed realizations of D. Then estimating $\boldsymbol{\alpha}(\cdot)$ and $\boldsymbol{\beta}(\cdot)$ can be done by maximizing $\mathbb{P}_n\ell\big(\alpha(\cdot), \boldsymbol{\beta}(\cdot); D\big)$. From (7), it is seen that the estimation procedure amounts to maximizing $\mathbb{P}_n\ell(\boldsymbol{\alpha}(u); D)$ w.r.t. $\alpha(u)$, and $\mathbb{P}_n\ell(\boldsymbol{\beta}(u); D)$ w.r.t. $\boldsymbol{\beta}(u)$ pointwisely for all $u \in [0, \tau]$, where $\tau > 0$ is determined by the range of the time-to-event data and/or the span of the future horizon of interest. Specifically, we have

$$\ell_n(\boldsymbol{\alpha}(u); \mathbf{D}_{\text{total}}) = n\mathbb{P}_n\ell(\boldsymbol{\alpha}(u); D) = \sum_{i=1}^n Y_i(u) \int_u^{X_i} \left[\log\lambda_{T_i,t-u}(u)dN_{T_i}(t) - \lambda_{T_i,t-u}(u)dt\right], \qquad (8)$$

where $\mathbf{D}_{\text{total}} = \cup_{1\le i\le n}\{X_i, \delta_{T_i}, \delta_{C_i}, \boldsymbol{Z}_i(t); 0 \le t \le X_i\}$ is the collection of data for all subjects and $Y_i(t) = I\{t \le X_i\}$, $i = 1, 2, \ldots, n$; $\ell_n(\boldsymbol{\beta}(u); \mathbf{D}_{\text{total}})$ is defined in the same way,

replacing T_i by C_i. For fixed $u \in [0,\tau]$, $\ell_n(\boldsymbol{\alpha}(u); \mathbf{D}_{\text{total}})$ and $\ell_n(\boldsymbol{\beta}(u); \mathbf{D}_{\text{total}})$ are strictly concave w.r.t. $\boldsymbol{\alpha}(u)$ and $\boldsymbol{\beta}(u)$, respectively. Hence the estimators of the functional-valued parameters $\boldsymbol{\alpha} : [0,\tau] \mapsto \mathbb{R}^p$ and $\boldsymbol{\beta} : [0,\tau] \mapsto \mathbb{R}^p$ exist and are unique. We denote these two estimators by $\hat{\boldsymbol{\alpha}}_n(\cdot)$ and $\hat{\boldsymbol{\beta}}_n(\cdot)$, respectively. That is, for $u \in [0,\tau]$, the pointwise estimator $\hat{\boldsymbol{\alpha}}_n(u)$ and $\hat{\boldsymbol{\beta}}_n(u)$ of the process-valued parameters $\boldsymbol{\alpha}(\cdot)$ and $\boldsymbol{\beta}(\cdot)$ are

$$\hat{\boldsymbol{\alpha}}_n(u) = \arg\max_{\boldsymbol{\alpha}(u)} \ell_n(\boldsymbol{\alpha}(u); \mathbf{D}_{\text{total}}), \quad \hat{\boldsymbol{\beta}}_n(u) = \arg\max_{\boldsymbol{\beta}(u)} \ell_n(\boldsymbol{\beta}(u); \mathbf{D}_{\text{total}}). \tag{9}$$

3.2 Weak Convergence

As the technical foundation, this section establishes the weak convergence, and develops the necessary limiting covariance functions in convenient analytical forms.

Let $\mathcal{X}$ be the support of D, and $\boldsymbol{\alpha}_0(\cdot)$ and $\boldsymbol{\beta}_0(\cdot)$ the truth. We denote the parameter space for the two functional-valued parameters $\boldsymbol{\alpha}(\cdot)$ and $\boldsymbol{\beta}(\cdot)$ by $\mathcal{F}$, and equip $\mathcal{F}$ with the uniform norm $\|\gamma(\cdot)\| = \sup_{0\le t\le\tau} |\gamma(t)|$ for $\gamma(\cdot) \in \mathcal{F}$, where $|a| = \max_{1\le j\le p} |a_j|$ for $a = (a_1, \cdots, a_p)^\top \in \mathbb{R}^p$. For notational convenience, we write the functional-valued parameters $\boldsymbol{\alpha}(\cdot)$, $\boldsymbol{\beta}(\cdot)$, $\boldsymbol{\alpha}_0(\cdot)$, $\boldsymbol{\beta}_0(\cdot)$ and the estimators $\hat{\boldsymbol{\alpha}}_n(\cdot)$, $\hat{\boldsymbol{\beta}}_n(\cdot)$ as $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, $\boldsymbol{\alpha}_0$, $\boldsymbol{\beta}_0$ and $\hat{\boldsymbol{\alpha}}_n$, $\hat{\boldsymbol{\beta}}_n$, respectively. Define the functions $\psi : \mathcal{F} \times \mathcal{X} \mapsto \left(\ell^\infty([0,\tau])\right)^p$ and $\vartheta : \mathcal{F} \times \mathcal{X} \mapsto \left(\ell^\infty([0,\tau])\right)^p$ as

$$[\psi(\boldsymbol{\alpha}; D)](u) = \int_u^\infty Y(t)\boldsymbol{Z}(t-u)dN_T(t) - \int_u^\infty Y(t)\boldsymbol{Z}(t-u)\exp\{\boldsymbol{\alpha}(u)^\top \boldsymbol{Z}(t-u)\}dt,$$
$$[\vartheta(\boldsymbol{\beta}; D)](u) = \int_u^\infty Y(t)\boldsymbol{Z}(t-u)dN_C(t) - \int_u^\infty Y(t)\boldsymbol{Z}(t-u)\exp\{\boldsymbol{\beta}(u)^\top \boldsymbol{Z}(t-u)\}dt.$$

Following the empirical process convention, we denote by Pg the mean of a function g: $Pg = \int g dP$. We assume the following conditions.

(A1) The parameter space $\mathcal{F}$ is uniformly bounded by a constant K_0.

(A2) The covariate process $\{\boldsymbol{Z}(t) : t \ge 0\}$ is bounded by a constant K_1. There exists K_2 and some $q \ge 1$ such that $E|\boldsymbol{Z}(t) - \boldsymbol{Z}(s)|^q \le K_2|t-s|^q$ for all $t, s \ge 0$. In addition, $P(X \ge \tau) > 0$, and $P\int_0^{(X-\tau)_+} \boldsymbol{Z}(t)\boldsymbol{Z}(t)^\top dt$ is positive definite with minimum eigenvalue $\lambda_{\min} > 0$.

(A3) The lifetime T and the censoring time C are absolute continuous, and the corresponding probability density functions f_T and f_C are bounded over $[0,\tau]$. The observed event time X is square integrable.

All three conditions are mild. Conditions (A1) and (A2) require that the parameter space and the covariates be uniformly bounded, which is a common regularity assumption in the literature of asymptotic analysis; see, for example, Wellner and Zhang (2007) and Zhao et al. (2017). The requirement $P(X \ge \tau) > 0$ in Condition (A2) is commonly used to ensure that the observed data contain sufficient information for estimating $\boldsymbol{\alpha}_0(u)$ and $\boldsymbol{\beta}_0(u)$, $u \in [0,\tau]$ (Ye et al., 2015; Chan and Wang, 2017). Continuity of the random variables T and C is also reasonable and frequently used (Berghaus and Bucher, 2018; Rousseau and Szabo, 2020).

We now present the foundational result of our framework in the following theorem on the weak convergence of the process $(\hat{\boldsymbol{\alpha}}_n, \hat{\boldsymbol{\beta}}_n)$.

Theorem 1 *Under Conditions (A1)–(A3), we have* $\|\hat{\alpha}_n - \alpha_0\| \to 0$ *and* $\|\hat{\beta}_n - \beta_0\| \to 0$ *almost surely. Moreover,*

$$\sqrt{n}\left(\begin{bmatrix}\hat{\alpha}_n \\ \hat{\beta}_n\end{bmatrix} - \begin{bmatrix}\boldsymbol{\alpha}_0 \\ \boldsymbol{\beta}_0\end{bmatrix}\right)$$

converges weakly to a Gaussian process with mean zero and covariance function

$$\boldsymbol{\rho}(u,s) = \begin{bmatrix}\boldsymbol{W}(u;\boldsymbol{\alpha}_0)^{-1}\boldsymbol{\varrho}(u,s,\psi)\boldsymbol{W}(s;\boldsymbol{\alpha}_0)^{-1} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{W}(u;\boldsymbol{\beta}_0)^{-1}\boldsymbol{\varrho}(u,s,\vartheta)\boldsymbol{W}(s;\boldsymbol{\beta}_0)^{-1}\end{bmatrix},$$

where

$$\begin{aligned}\boldsymbol{\varrho}(u,s,\psi) =& P[\psi(\boldsymbol{\alpha}_0;D)](u)[\psi(\boldsymbol{\alpha}_0;D)](s)^\top, \boldsymbol{\varrho}(u,s,\vartheta) = P[\vartheta(\boldsymbol{\beta}_0;D)](u)[\vartheta(\boldsymbol{\beta}_0;D)](s)^\top, \\ \boldsymbol{W}(u;\gamma) =& P\int_0^\infty I\{t>u\}Y(t)\boldsymbol{Z}(t-u)\boldsymbol{Z}(t-u)^\top\exp\{\gamma(u)^\top\boldsymbol{Z}(t-u)\}dt.\end{aligned}$$

The covariance function $\boldsymbol{\rho}(u,s)$ fully spells out the dependence between $\left(\hat{\boldsymbol{\alpha}}_n(t),\hat{\boldsymbol{\beta}}_n(t)\right)$ and $\left(\hat{\boldsymbol{\alpha}}_n(s),\hat{\boldsymbol{\beta}}_n(s)\right)$ for $0<t,s<\tau$. This result is the key enabling (5) for evaluating the relevant uncertainties. To apply it, a consistent estimator of the asymptotic covariance function is needed. A careful look at the sandwich form of the asymptotic covariance in Theorem 1 reveals that it is a function of $\boldsymbol{W}$ and $\boldsymbol{\varrho}$. We can estimate these two functions first, and a simple plug-in yields an estimator for the asymptotic covariance. Specifically, we define an estimator $\hat{\boldsymbol{W}}_n$ for $\boldsymbol{W}$ as

$$\hat{\boldsymbol{W}}_n(u;\gamma) = \mathbb{P}_n\int_u^X \boldsymbol{Z}(t-u)\boldsymbol{Z}(t-u)^\top\exp\{\gamma(u)^\top\boldsymbol{Z}(t-u)\}dt,$$

and an estimator $\hat{\boldsymbol{\varrho}}_n$ for $\boldsymbol{\varrho}$ as

$$\begin{aligned}&\hat{\boldsymbol{\varrho}}_n(u,s;\gamma) \\ =&\mathbb{P}_n\int_{u\vee s}^X \boldsymbol{Z}(t-u)\boldsymbol{Z}(t-s)^\top\exp\{\gamma(0)^\top\boldsymbol{Z}(t)\}dt \\ -&\mathbb{P}_n\int_s^X\int_u^{t_2}\boldsymbol{Z}(t_1-u)\boldsymbol{Z}(t_2-s)^\top\exp\{\gamma(u)^\top\boldsymbol{Z}(t_1-u)+\gamma(0)^\top\boldsymbol{Z}(t_2)\}dt_1dt_2 \\ -&\mathbb{P}_n\int_u^X\int_s^{t_1}\boldsymbol{Z}(t_1-u)\boldsymbol{Z}(t_2-s)^\top\exp\{\gamma(s)^\top\boldsymbol{Z}(t_2-s)+\gamma(0)^\top\boldsymbol{Z}(t_1)\}dt_2dt_1 \\ +&\mathbb{P}_n\int_s^X\int_u^X\boldsymbol{Z}(t_1-u)\boldsymbol{Z}(t_2-s)^\top\exp\{\gamma(u)^\top\boldsymbol{Z}(t_1-u)+\gamma(s)^\top\boldsymbol{Z}(t_2-s)\}dt_1dt_2.\end{aligned} \tag{10}$$

Based on the two estimators above, define an estimator $\hat{\boldsymbol{\rho}}_n$ for $\boldsymbol{\rho}$ as

$$\hat{\boldsymbol{\rho}}_n(u,s) = \begin{bmatrix}\hat{\boldsymbol{W}}_n(u;\hat{\boldsymbol{\alpha}}_n)^{-1}\hat{\boldsymbol{\varrho}}_n(u,s;\hat{\boldsymbol{\alpha}}_n)\hat{\boldsymbol{W}}_n(s;\hat{\boldsymbol{\alpha}}_n)^{-1} & \mathbf{0} \\ \mathbf{0} & \hat{\boldsymbol{W}}_n(u;\hat{\boldsymbol{\beta}}_n)^{-1}\hat{\boldsymbol{\varrho}}_n(u,s;\hat{\boldsymbol{\beta}}_n)\hat{\boldsymbol{W}}_n(s;\hat{\boldsymbol{\beta}}_n)^{-1}\end{bmatrix}. \tag{11}$$

The following proposition shows the consistency of $\hat{\boldsymbol{\rho}}_n$.

Proposition 2 *Suppose Conditions (A1)–(A3) hold. Then the estimator* $\hat{\boldsymbol{\rho}}_n(u,s)$ *defined in* (11) *converges to* $\boldsymbol{\rho}(u,s)$ *in probability.*

3.3 Interval Estimations for Predictive Probabilities

We focus on two predictive probabilities: the predictive survival probability $S_{T,t}(\tau)$ in (1) and the predictive probability $F_t(\tau)$ of observing an event in (4).

By plugging $\hat{\lambda}_{T,t}(u) = \exp\{\hat{\boldsymbol{\alpha}}_n(u)^\top \boldsymbol{Z}(t)\}$ and $\hat{\lambda}_{C,t}(u) = \exp\{\hat{\boldsymbol{\beta}}_n(u)^\top \boldsymbol{Z}(t)\}$ in $S_{T,t}(\tau)$ and $F_t(\tau)$ of (1) and (4), we have

$$\hat{S}_{T,t}(\tau) = \exp\left\{-\int_0^\tau \hat{\lambda}_{T,t}(u)du\right\},\ \hat{F}_t(\tau) = \int_0^\tau \hat{S}_{T,t}(u)\hat{S}_{C,t}(u)\hat{\lambda}_{T,t}(u)du,$$

where $\hat{S}_{C,t}(\tau) = \exp\{-\int_0^\tau \hat{\lambda}_{C,t}(u)du\}$. As functions of $(\hat{\boldsymbol{\alpha}}_n, \hat{\boldsymbol{\beta}}_n)$, the weak convergence of the two estimators $\hat{S}_{T,t}(\tau)$ and $\hat{F}_t(\tau)$ follows from Theorem 1 and the functional delta method. We summarize this result in the following proposition.

Proposition 3 *Suppose Conditions (A1)–(A3) hold. Then we have*

$$\sqrt{n}\big(\hat{S}_{T,t}(\tau) - S_{T,t}(\tau)\big) \rightsquigarrow N(0, \sigma^2),$$

where

$$\sigma^2 = S_{T,t}^2(\tau)\int_0^\tau\int_0^\tau \boldsymbol{Z}(t)^\top \boldsymbol{W}(u;\boldsymbol{\alpha}_0)^{-1}\boldsymbol{\varrho}(u,s,\psi)\boldsymbol{W}(s;\boldsymbol{\alpha}_0)^{-1}\boldsymbol{Z}(t)\lambda_{T,t}(u)\lambda_{T,t}(s)duds.$$

The asymptotic variance σ^2 can be consistently estimated by

$$\hat{\sigma}_n^2 = \hat{S}_{T,t}^2(\tau)\int_0^\tau\int_0^\tau \boldsymbol{Z}(t)^\top \hat{\boldsymbol{W}}_n(u;\hat{\boldsymbol{\alpha}}_n)^{-1}\hat{\boldsymbol{\varrho}}_n(u,s;\hat{\boldsymbol{\alpha}}_n)\hat{\boldsymbol{W}}_n(s;\hat{\boldsymbol{\alpha}}_n)^{-1}\boldsymbol{Z}(t)\hat{\lambda}_{T,t}(u)\hat{\lambda}_{T,t}(s)duds$$

with convergence rate $\sqrt{n}(\hat{\sigma}_n^2 - \sigma^2) = O_P(1)$. Similarly, $\sqrt{n}\big(\hat{F}_t(\tau) - F_t(\tau)\big) \rightsquigarrow N(0, \eta^2)$, where η^2 has an expression similar to σ^2. The expression of η^2 and its root-n consistent estimator is given in (41) *and* (42)*.*

Based on Proposition 3, the two-sided equal-tailed $100(1-\alpha)\%$ pointwise confidence intervals for $S_{T,t}(\tau)$ and $F_t(\tau)$ are

$$\hat{S}_{T,t}(\tau) \pm z_{1-\alpha/2}\hat{\sigma}_n/\sqrt{n} \ \text{ and } \ \hat{F}_t(\tau) \pm z_{1-\alpha/2}\hat{\eta}_n/\sqrt{n},$$

where $z_{1-\alpha/2}$ is the $(1-\alpha/2)$ upper quantile of the standard normal distribution.

In addition to $S_{T,t}(\tau)$ and $F_t(\tau)$, our approach can also be extended to construct confidence intervals for the restricted mean residual lifetime $E[\min(T-t,\tau) \mid \mathcal{F}_t, T \geq t]$ (Tian et al., 2020). The procedures are parallel to those for $S_{T,t}(\tau)$ and $F_t(\tau)$, and thus the details are deferred to Appendix A. Furthermore, the asymptotic normality of $\hat{S}_{T,t}(\tau)$ and $\hat{F}_t(\tau)$ for a fixed τ can be extended to the weak convergence of $\hat{S}_{T,t}(\cdot)$ and $\hat{F}_t(\cdot)$ to Gaussian processes over a bounded interval $[0,\bar{\tau}]$. This allows us to construct simultaneous confidence bands for functions of interest (e.g, $S_{T,t}(\tau)$) over $\tau \in [0,\bar{\tau}]$. Details of the simultaneous confidence bands can be found in Appendix B.

3.4 Interval Estimations for Total Events

Predicting the aggregate number of events within a pre-specified horizon is a practically important task. Methodologically, as a goodness-of-fit assessment, the point prediction and prediction intervals of the aggregate number of events can be compared with the true observed values. Such a check provides a diagnostic of the forward intensity model.

Given all the subjects that are still at risk until time t and the corresponding covariates observed at t, let us consider the point and interval predictions of the total number of events over $(t, t+\tau]$ – a random variable denoted by $N_{t,\tau}$. Let $R(t)$ be the set consisting of all individuals that are at risk at time t. The random variable $N_{t,\tau}$ can be written as

$$N_{t,\tau} = \sum_{i \in R(t)} I\{T_i < t+\tau, T_i < C_i | T_i \wedge C_i \geq t, \boldsymbol{Z}_i(t)\}.$$

The summands in $N_{t,\tau}$ are independent Bernoulli random variables with different success probabilities $F_{i,t}(\tau)$, $i \in R(t)$, where $F_{i,t}(\tau) = P(T_i < t+\tau, T_i < C_i | \mathcal{F}_t, T_i \wedge C_i \geq t)$ is the predictive probability of observing an event over $(t, t+\tau]$ for the ith subject, as given in (4). As a result, $N_{t,\tau}$ follows the Poisson binomial distribution with parameters $F_{i,t}(\tau)$, $i \in R(t)$; see also Yuan et al. (2018).

The point prediction for $N_{t,\tau}$ can be naturally obtained as $\sum_{i \in R(t)} \hat{F}_{i,t}(\tau)$, where $\hat{F}_{i,t}(\tau)$ is the subject-specific version of $\hat{F}_t(\tau)$ for the ith subject. On the other hand, the prediction intervals of $N_{t,\tau}$ are difficult to obtain due to the nature of $N_{t,\tau}$ as a count. There are two sources of uncertainties in the point prediction $\sum_{i \in R(t)} \hat{F}_{i,t}(\tau)$. One is the uncertainty from the estimator $(\hat{\boldsymbol{\alpha}}_n, \hat{\boldsymbol{\beta}}_n)$, and the other is the uncertainty from the Poisson binomial distribution of $N_{t,\tau}$. The uncertainties from $(\hat{\boldsymbol{\alpha}}_n, \hat{\boldsymbol{\beta}}_n)$ can be readily quantified through the weak convergence result in Theorem 1. To further account for the uncertainties in the Poisson binomial distribution, we adopt an efficient resampling method, as summarized in Algorithm 1.

Algorithm 1: Resampling method for prediction intervals

Input: Significance level α and estimators $\hat{\boldsymbol{\alpha}}_n, \hat{\boldsymbol{\beta}}_n, \hat{\boldsymbol{\rho}}_n(u,s)$.
Output: Prediction interval $(\underline{N}_{t,\tau}, \overline{N}_{t,\tau})$.

Initialization: $b \leftarrow 1$.
while $b \leq B$ **do**
  Generate $\left(\boldsymbol{\alpha}^{(b)}, \boldsymbol{\beta}^{(b)}\right)$ from the Gaussian process with mean $(\hat{\boldsymbol{\alpha}}_n, \hat{\boldsymbol{\beta}}_n)$ and covariance function $\hat{\boldsymbol{\rho}}_n(u,s)$.
  Compute $F_{i,t}^{(b)}(\tau)$ based on the parameters $\left(\boldsymbol{\alpha}^{(b)}, \boldsymbol{\beta}^{(b)}\right)$, $i \in R(t)$.
  Draw $N_{t,\tau}^{(b)}$ from the Poisson binomial distribution with parameters $F_{i,t}^{(b)}(\tau)$, $i \in R(t)$. $b \leftarrow b+1$.
end
Based on $\{N_{t,\tau}^{(b)}\}_{b=1}^{B}$, compute the $100(1-\alpha)\%$ prediction interval $(\underline{N}_{t,\tau}, \overline{N}_{t,\tau})$.
Return $(\underline{N}_{t,\tau}, \overline{N}_{t,\tau})$.

3.5 Working with Discrete-Time Data

In practical survival analysis, real-world problems often involve discrete-time data. This means that time-varying covariates are typically observed only at specific discrete time points, such as weekly or monthly intervals. In our numerical examples, we adhere to the standard practice in survival analysis by employing discrete-time approximations for any continuous-time integrations required in our methodology. Detailed information about these approximations can be found in Appendix C. Furthermore, for those who consider discrete-time survival models as the true representation, we establish corresponding asymptotic theory. A standalone asymptotic analysis in a discrete setting is also provided in Appendix C, which is of independent interest and leverages the forward intensity model in the discrete-time scale.

4. Real Data Examples

4.1 U.S. Corporate Default Data

We conducted an analysis using an illustrative dataset from Duan et al. (2012), comprising 2,000 publicly traded U.S. firms on the stock market, spanning from January 1991 to November 2011. Each of these 2,000 firms has monthly records containing time-varying financial covariates, which include: (i) the trailing 1-year return on the S&P 500 index (SP500); (ii) the 3-month U.S. Treasury bill rate (TreasRate); (iii) distance to default (Merton, 1974, DTD); (iv) the ratio of the sum of cash and short-term investments to total assets (CASH/TA); (v) the ratio of net income to total assets (NI/TA); (vi) logarithm of the ratio of a firm's market equity value to the average market equity value of S&P 500 firms (SIZE); (vii) the market-to-book asset ratio (M/B); and (viii) the 1-year idiosyncratic volatility (SIGMA). The first two covariates represent macroeconomic factors common to all firms, while the remaining six are firm-specific. If a firm exited the market before the end-of-study date, the exit type was categorized into one of two classes: defaults and other exit types, such as mergers and acquisitions. Out of the 2,000 firms, 168 exited the market due to default, while 1,334 firms exited for other reasons, leaving 498 firms at the study's conclusion.

Our primary interest is the default probability of a firm at some future time $t+\tau$. Hence the exit for non-default reasons is considered as random censoring. The basic time interval Δt is 1 month. Following Duan et al. (2012), we model the forward intensity of time T to default by

$$\begin{aligned}\lambda_{T,t_j}(t_k) = \exp\big\{&\alpha_0(t_k) + \alpha_1(t_k)\mathrm{SP500} + \alpha_2(t_k)\mathrm{TreasRate} + \alpha_3(t_k)\mathrm{DTD}_{\mathrm{level}}(t_j) + \\ &\alpha_4(t_k)\mathrm{DTD}_{\mathrm{trend}}(t_j) + \alpha_5(t_k)\mathrm{CASH/TA}_{\mathrm{level}}(t_j) + \alpha_6(t_k)\mathrm{CASH/TA}_{\mathrm{trend}}(t_j) + \\ &\alpha_7(t_k)\mathrm{NI/TA}_{\mathrm{level}}(t_j) + \alpha_8(t_k)\mathrm{NI/TA}_{\mathrm{trend}}(t_j) + \alpha_9(t_k)\mathrm{SIZE}_{\mathrm{level}}(t_j) + \\ &\alpha_{10}(t_k)\mathrm{SIZE}_{\mathrm{trend}}(t_j) + \alpha_{11}(t_k)\mathrm{M/B}(t_j) + \alpha_{12}(t_k)\mathrm{SIGMA}(t_j)\big\},\end{aligned}$$

where the subscript "level" denotes the average in the previous 12 months and "trend" denotes the difference between the current value and its previous 12-month average. The same covariates are used for modeling the random censoring time C; see Appendix D.1 for more detail.

To assess the default risk of each firm, we apply our method of Section 3.3 to make individual-level predictions of $F_t(\tau)$ and construct the associated 95% confidence intervals. For demonstrations, we consider three landmark times t near the outbreak of the financial crisis in 2008: Jan 2006, Jan 2007, and Jan 2008. We consider each month in the next year after the landmark time with $\tau \in \{\Delta t, 2\Delta t, \ldots, 12\Delta t\}$. For each of the three landmark times, one representative company that defaulted in that year is presented here for illustration. The three selected companies are from the sectors of apparel goods and notions, saving institutions, and pharmaceutical preparations. The company names are unknown, because they are not provided in the publicly available data set we have access to. Detailed default dates for the three selected companies are shown in Figure 1. The estimated predictive default probabilities $F_t(\tau)$ and the associated 95% CIs for the selected companies are shown in Figure 1. For comparison, we also present the average predictive default probabilities $F_t(\tau)$ and 95% CIs of all companies that are at risk at landmark time t (excluding the selected company). When estimating $F_t(\tau)$ for a selected company, we follow the principle of leave-one-out cross-validation. That is, we first exclude the selected company and use the remaining ones as the training set for parameter estimation, then use the selected company for validation.

From Figure 1, we observe that these companies indeed exhibit significantly higher predicted default probabilities compared to the benchmarking average levels. Notably, the predicted probabilities show rapid growth, starting as early as a few months before their defaults, with varying rates for the three firms. This provides valuable insights when assessing individual default risk. Furthermore, it is worth noting that the associated confidence intervals (CIs) for these selected companies are substantially wider than the average ones. Our discovery emphasizes that the width of the CIs adds a new dimension to evaluating a company's default risk. It is reasonable to infer that the level of uncertainty also reflects the vulnerability of the company in the near future. As a proactive measure, companies with larger and/or expanding CIs of the predicted $F_t(\tau)$ should be subject to closer monitoring. This approach can aid in identifying companies that may be at higher risk of default in the future.

Next, we assess the default risk at the aggregate level, by applying the proposed method in Section 3.4 to produce both point and interval predictions for the total events $N_{t,\tau}$ over $(t, t+\tau]$. The prediction is performed every month after Dec 1995; i.e., t ranges from Jan 1996 to Nov 2011. We set $\tau \in \{3\Delta t, 6\Delta t, 12\Delta t\}$ to assess the risk levels of short-, mid-, and long-terms, respectively. More specifically, at the beginning of each month (time t), we first identify the risk set $R(t)$ consisting of all active firms, then estimate the unknown model parameters using the data observed up to time t, and finally calculate point and interval predictions for $N_{t,\tau}$.

The point predictions and the associated 95% PIs of $N_{t,\tau}$ are presented in Figure 2. In order to check the goodness-of-fit, the observed values of $N_{t,\tau}$ are shown and compared with the predicted values in Figure 2. From Figure 2, we see that the width of the intervals are reasonable with satisfactory coverages: most of the observed data points fall within the 95% PI for all short-, mid-, and long-term predictions. This demonstrates that the forward intensity modeling approach works satisfactorily. In general, goodness-of-fit assessment using confidence intervals is more reliable, upon incorporating information from the uncertainty assessments.

An interesting observation from Figure 2 is how informative the results are in reflecting the progression of the overall market risk level. For instance, both the prediction for $N_{t,\tau}$ and the width of the associated 95% PIs grew drastically overall all three terms near the time of July 2007 to July 2009, which is known near when the 2008 financial crisis occurred. This suggests that at an early stage of catastrophic credit events, such results are helpful for market risk monitoring and systemic risk management.

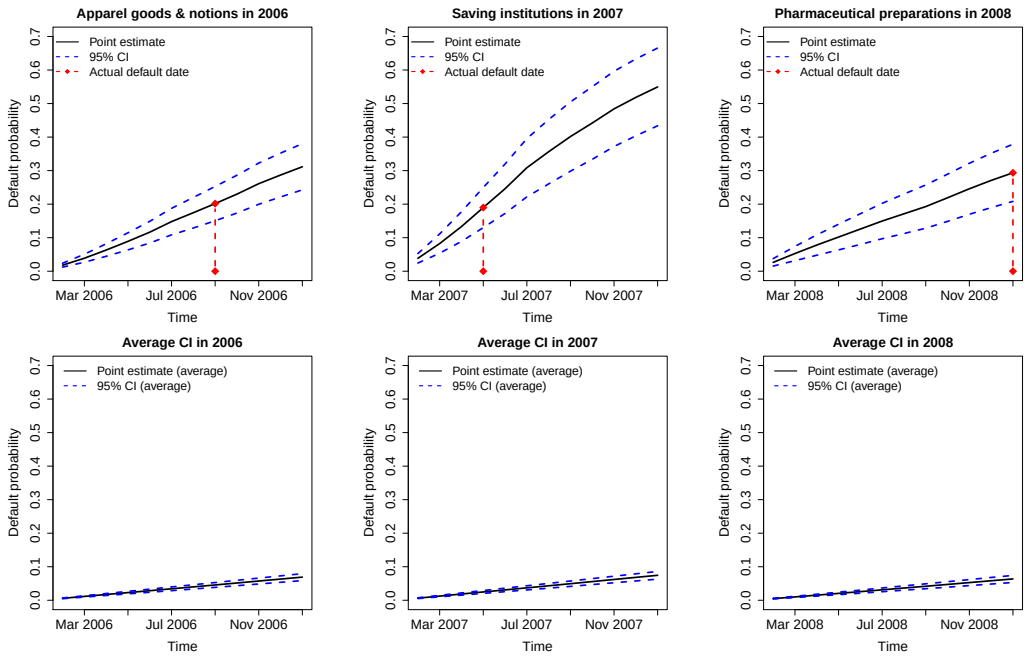

Figure 1: Estimated predictive default probabilities and the associated 95% CIs for selected companies, together with the average default probabilities and 95% CIs for all companies at risk. The landmark time t in the left, middle, and right panels is fixed at Jan 2006, Jan 2007, and Jan 2008, respectively.

4.2 Primary Biliary Cirrhosis Data

We apply the proposed method to the primary biliary cirrhosis (PBC) data reported in Fleming and Harrington (2005). This data set was collected from a clinical trial conducted by the Mayo Clinic between January 1974 and July 1986. In the trial, 312 patients diagnosed with PBC of the liver were randomly allocated to a treatment group with D-penicillamine or a placebo group. At the time of enrollment, certain patient demographic information, such as age and gender, was recorded. Afterward, each patient was regularly followed (approximately every 6 months) to measure some time-dependent clinical covariates until the date of death, the date of drop-out, or the end-of-study date (July 1986), whichever came first. Specifically, 15 covariates were collected in total, and detailed descriptions of these covariates are listed in Table 1 in Appendix D.2. By the end-of-study date, 125 of the

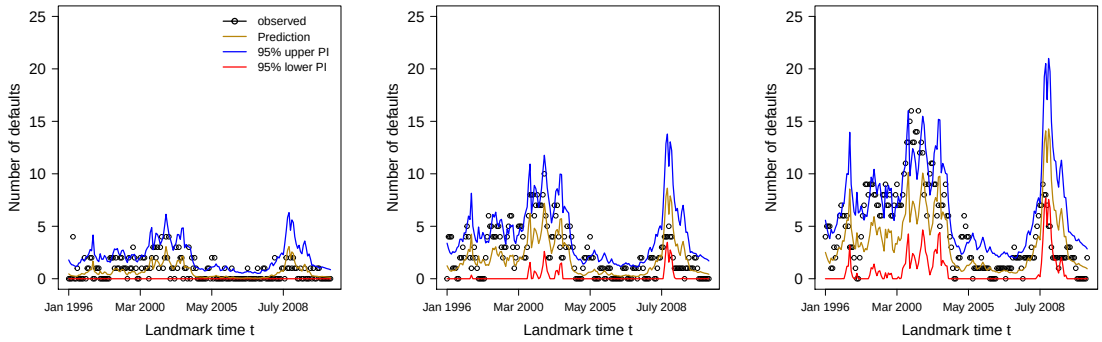

Figure 2: Observed values, point predictions and 95% prediction intervals for the aggregate number $N_{t,\tau}$ of defaults within $(t, t+\tau]$. The left, middle, and right panels correspond to the cases in which $\tau = 3\Delta t$, $6\Delta t$, and $12\Delta t$, respectively.

312 patients had died and 27 had dropped out during the trial. The remaining 160 were still alive at the end-of-study date.

In clinical practice, the main focus is the survival time T from the date of enrollment to the date of death. We treat patient drop-outs as random censoring, and thus random censoring time C is the duration between enrollment and drop-out. The basic time interval Δt is 6 months. We model the forward intensity of T as

$$\begin{aligned}\lambda_{T,t_j}(t_k) = \exp\{&\alpha_0(t_k) + \alpha_1(t_k)\mathrm{Trt} + \alpha_2(t_k)\mathrm{Gen} + \alpha_3(t_k)\mathrm{Age}(t_j) + \alpha_4(t_k)\mathrm{Ascts}(t_j) + \\ &\alpha_5(t_k)\mathrm{Hepato}(t_j) + \alpha_6(t_k)\mathrm{Spiders}(t_j) + \alpha_7(t_k)\mathrm{Edema}(t_j) + \alpha_8(t_k)\mathrm{Bili}(t_j) + \\ &\alpha_9(t_k)\mathrm{Chol}(t_j) + \alpha_{10}(t_k)\mathrm{Albumin}(t_j) + \alpha_{11}(t_k)\mathrm{Alkphos}(t_j) + \alpha_{12}(t_k)\mathrm{Ast}(t_j) + \\ &\alpha_{13}(t_k)\mathrm{Platelet}(t_j) + \alpha_{14}(t_k)\mathrm{Protime}(t_j) + \alpha_{15}(t_k)\mathrm{Stage}(t_j)\}.\end{aligned}$$

The same covariates are used for the forward intensity of C, and the detailed expression of $\lambda_{C,t_j}(t_k)$ is given in Appendix D.2.

To evaluate the risk of death at patient level, we apply the method proposed in Section 3.3 to make point and interval estimation for $F_t(\tau)$ for patients at risk at landmark time t. Because risk assessment at the early stage of a trial is of special interest in clinical practice (Beever and Swaby, 2019), we set landmark time t as the enrollment time (Jan 1974). Following Murtaugh et al. (1994), we let the prediction horizon $t+\tau$ range within the 2-year period after the landmark time, i.e., $\tau \in \{\Delta t, \ldots, 4\Delta t\}$.

To demonstrate the difference in the prediction between patients who survive beyond or die within the prediction horizon, we randomly select two patients, labeled as 2 and 97; Patient 2 was still alive at the end-of-study date (July 1986) and Patient 97 died in Jan 1976. As with the preceding example, we follow the principle of leave-one-out cross-validation when estimating $F_{i,t}(\tau)$ for a selected patient. Results for the two selected patients are shown in Figure 3. As can be seen, the estimated predictive death probability for Patient 97 is significantly larger than that for Patient 2. Moreover, the CI for Patient 97 is also much wider than that for Patient 2. This indicates that the interval estimate of $F_{i,t}(\tau)$

is informative in detecting future events. This result is helpful for clinicians seeking to distinguish high-risk patients for targeted interventions (Bansal and Heagerty, 2018).

To evaluate the risk of death at the aggregate level, the proposed method in Section 3.4 is applied to produce both point and interval predictions for $N_{t,\tau}$. The prediction is conducted at each monitor time over the whole follow-up period—i.e., landmark time t ranges from Jan 1974 to Jul 1986. In addition, we set forward-looking time $\tau \in \{2\Delta t, 4\Delta t\}$ to assess the short- and mid-term aggregate-level risks, respectively. To make the prediction fair, we randomly split the data into a training set and a testing set with the same number of patients. The training set is used for parameter estimation, and the testing set for checking prediction accuracy. Results for point prediction and the associated 95% PIs of the aggregate number of deaths are displayed in Figure 4. For checking the model goodness-of-fit, we also present the observed values of $N_{t,\tau}$ in Figure 4. As can be seen, there is good agreement between the predicted and true observed values of $N_{t,\tau}$, and most of the observed data points fall in the 95% PI for both short- and mid-term predictions. This indicates satisfactory adequacy of the forward intensity approach on this data set from a clinical trial. Moreover, it can be seen that the variability of the width of PIs at different landmark times is small. This implies that our uncertainty on this cohort of patients does not change a lot as the trial proceeds. That is, the trial was under good control throughout the follow-up period.

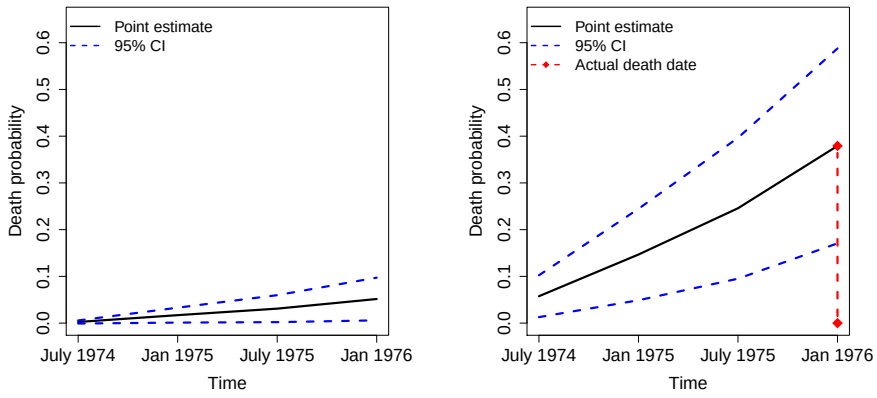

Figure 3: Estimated predictive failure probabilities $F_t(\tau)$ and the associated 95% CIs for Patients 2 (left panel) and 97 (right panel). Landmark time t is fixed at the time of enrollment (Jan 1974).

5. Simulation Study

5.1 Simulation Settings

We conduct extensive simulation studies to examine the performance of the proposed methods. To mimic the real cases, we consider a discrete observational scheme and set the basic time interval Δt as 1 month. Inspired by Zhu et al. (2019), we generate the data $(T, \boldsymbol{Z})$ from

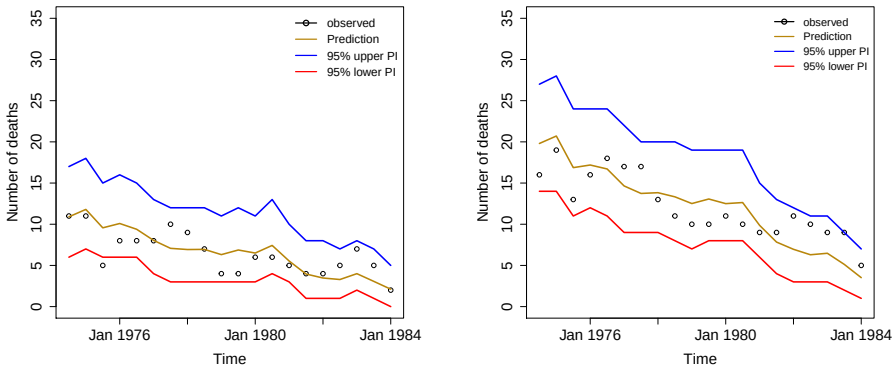

Figure 4: Observed values, point predictions and 95% prediction intervals for the aggregate number $N_{t,\tau}$ of deaths within time interval $(t, t+\tau]$. The left and right panels correspond to cases in which $\tau = 2\Delta t$ (1 year) and $4\Delta t$ (2 years), respectively.

a designated joint distribution for which the proposed log-linear forward intensity model holds. Specifically, we consider a log-linear forward intensity model with four time-varying covariates:

$$\lambda_{T,t}(u) = \exp\{\alpha_0(u)Z_0(t) + \alpha_1(u)Z_1(t) + \alpha_2(u)Z_2(t) + \alpha_3(u)Z_3(t)\}. \tag{12}$$

We carefully choose the covariate processes so that they mimic the predictors that are commonly seen in real applications. In particular, $Z_1(t)$ mimics the predictors with a linear growing trend, and it follows a linear random-effects model $Z_1(t) = \theta_1 + \theta_2 t + \epsilon_t$, where $\epsilon_t \sim N(0,1)$ and $(\theta_1, \theta_2)^\top$ follows a bivariate normal distribution with mean 1, variance 0.5, and covariance 0.2. The second covariate $Z_2(t)$ follows the change-point model $Z_2(t) = f(t) + \epsilon_t$, where

$$f(t) = \begin{cases} t, & \text{if } t \le 5; \\ 3t - 10, & \text{if } t > 5, \end{cases} \quad \text{and } \epsilon_t \sim N(0,1).$$

As for the third variable $Z_3(t)$, it follows a linear model with autoregressive random errors: $Z_3(t) = t + \nu_t$, where ν_t follows the AR(1) model with the autoregressive coefficient equal to 0.2. Finally, the variable $Z_0(t)$ is an adjusting covariate with a specially designed distribution to ensure the validity of the forward intensity model; see also Zhu et al. (2019). Detailed descriptions about $Z_0(t)$ and the data-generation procedure are provided in Appendix D.3. Without loss of generality, we set the true value of $\boldsymbol{\alpha}(u) = (\alpha_0(u), \alpha_1(u), \alpha_2(u), \alpha_3(u))^\top$ as $(1, 1, 0.5, 2)^\top$ for all $u \in [0, \tau]$. Using the same procedures as those for generating T, we generate random censoring time C under the same settings of $\boldsymbol{Z}(t)$. For simplicity, the true value of $\boldsymbol{\beta}(u) = (\beta_0(u), \beta_1(u), \beta_2(u), \beta_3(u))^\top$ is set equal to $\boldsymbol{\alpha}(u)$, i.e., $\boldsymbol{\beta}(u) = (1, 1, 0.5, 2)^\top$ for $u \in [0, \tau]$. The administrative censoring time is fixed at $\tilde{C} = 100$ for all subjects. In our simulation setting, the percentage of subjects whose event times are observed, randomly censored, and right-censored by the end-of-study date (administrative censoring) are around 45%, 45%, and 10%, respectively.

We examine different sample sizes $n \in \{50, 100, 200, 500\}$. For each n, we generate 1,000 synthetic datasets. For each synthetic dataset, we apply the proposed approach to perform three tasks:

(i) Construct 95% confidence intervals for $F_t(\tau) = P(T < t + \tau, T < C \mid \boldsymbol{Z}(t), T \wedge C \geq t)$, in which we set the time-dependent covariates $(Z_1(t), Z_2(t), Z_3(t))$ equal to their expected values, and the detailed setting of $Z_0(t)$ is given in Appendix D.3.

(ii) Construct 95% confidence intervals for $S_{T,t}(\tau) = P(T \geq t + \tau \mid \boldsymbol{Z}(t), T \wedge C \geq t)$.

(iii) Apply Algorithm 4 in Appendix C.4 to construct 95% prediction intervals for $N_{t,\tau}$, in which we set the number B of replications as 5,000. To assess the out-of-sample predictive performance of our approach, we randomly split each dataset into two of equal size. One is the training set used for parameter estimation and the other is the testing set used to check prediction accuracy.

In all the three tasks, we consider different values of landmark times $t \in \{20, 25, 30, 35, 40\}$ and forward-looking times $\tau \in \{1\Delta t, 2\Delta t, 5\Delta t, 10\Delta t\}$.

5.2 Interval Estimation for $F_t(\tau)$

First, we examine the performance of the proposed 95% confidence intervals for $F_t(\tau)$. The simulation results for $S_{T,t}(\tau)$ are similar and thus deferred to Appendix D.4.

Empirical coverage probabilities and the average 95% confidence limits based on the 1,000 synthetic datasets are summarized in Figures 5 and 6. As can be seen, the coverage probabilities are close to the nominal level, especially when the sample sizes are large. In addition, for fixed sample size n and landmark time t, the average length of CIs grows with forward-looking time τ. This aligns with the common sense and our theory that uncertainties are aggregating over future time.

5.3 Interval Prediction for $N_{t,\tau}$

In this subsection, we examine the performance of the proposed prediction interval for the number $N_{t,\tau}$ of failures over $(t, t + \tau]$.

Empirical coverage probabilities and average 95% prediction limits are summarized in Figures 7 and 8. As shown in Figure 7, the coverage probabilities of the proposed prediction interval are generally close to the nominal level. An interesting finding is that the coverage probability for small τ (e.g., $\tau = 1, 2$) is slightly larger than that for large τ (e.g., $\tau = 5, 10$). This is because the lower prediction limits of the observed number $N_{t,\tau}$ of failures are generally zero when τ is small, as shown in Figure 8. As a result, the observed value of $N_{t,\tau}$, which is always positive by definition, would never fall below the lower prediction limits when τ is small. This makes the coverage probability slightly higher than our expectation.

6. Discussion

We view the forward intensity function as a novel tool for quantifying uncertainties linked to model-based predictions. Our results underscore the effectiveness and computational efficiency of this framework, showcasing its wide range of potential applications.

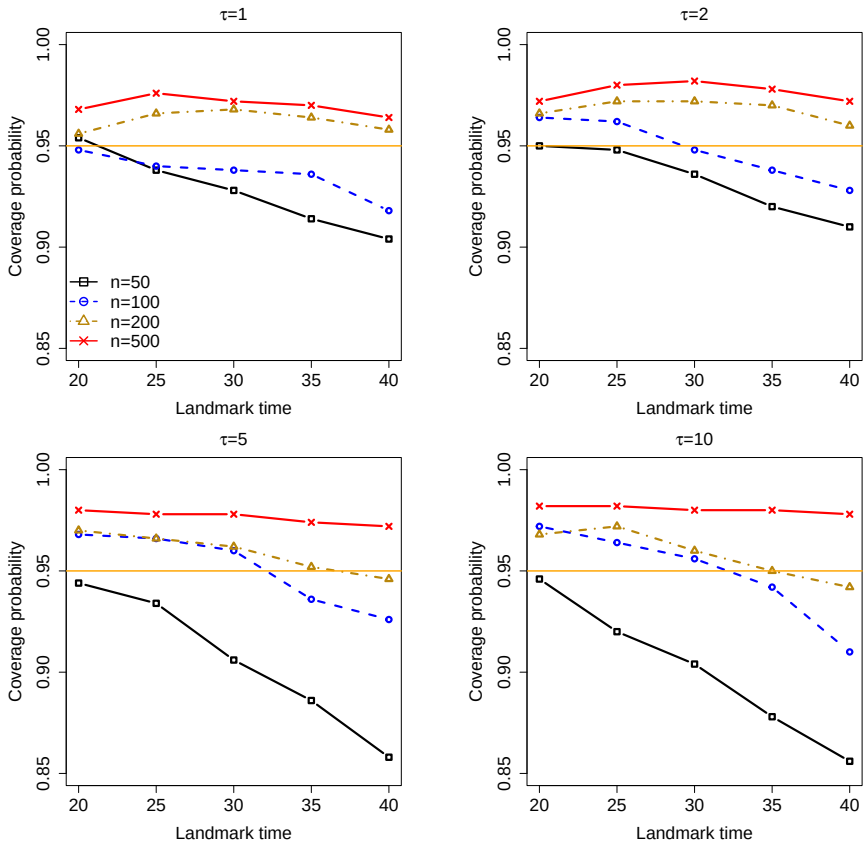

Figure 5: Empirical coverage probability of the 95% confidence interval for $F_t(\tau)$ under different values of $n \in \{50, 100, 200, 500\}$, $t \in \{20, 25, 30, 35, 40\}$, and $\tau \in \{1, 2, 5, 10\}$.

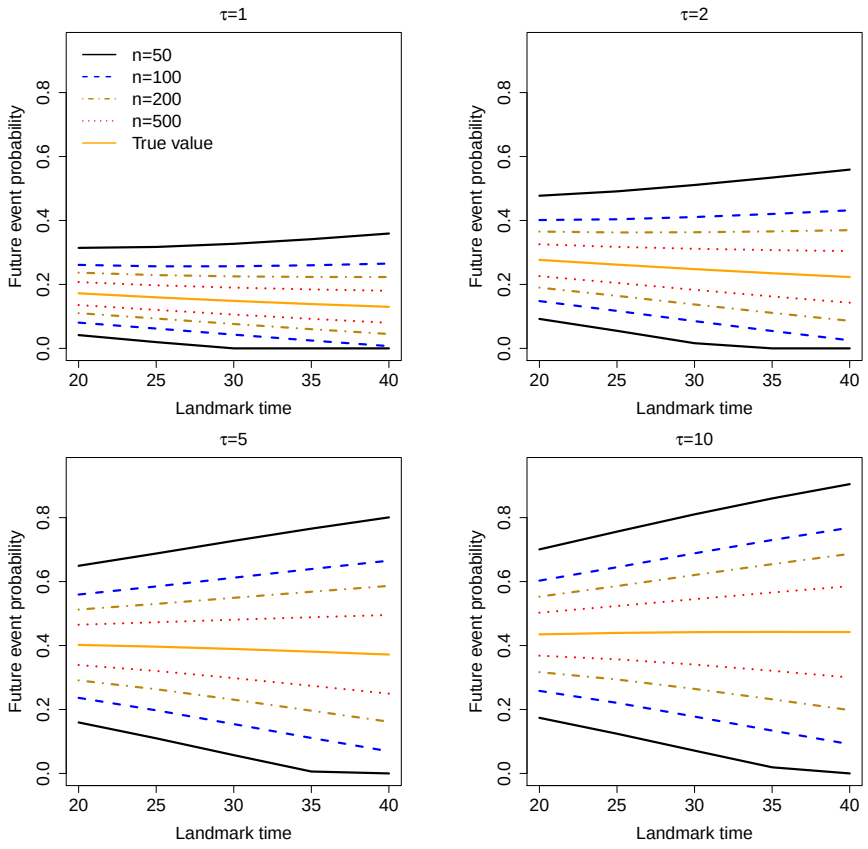

Figure 6: Average upper and lower limits of the 95% confidence interval for $F_t(\tau)$, together with the true value of $F_t(\tau)$ under different values of $n \in \{50, 100, 200, 500\}$, $t \in \{20, 25, 30, 35, 40\}$, and $\tau \in \{1, 2, 5, 10\}$.

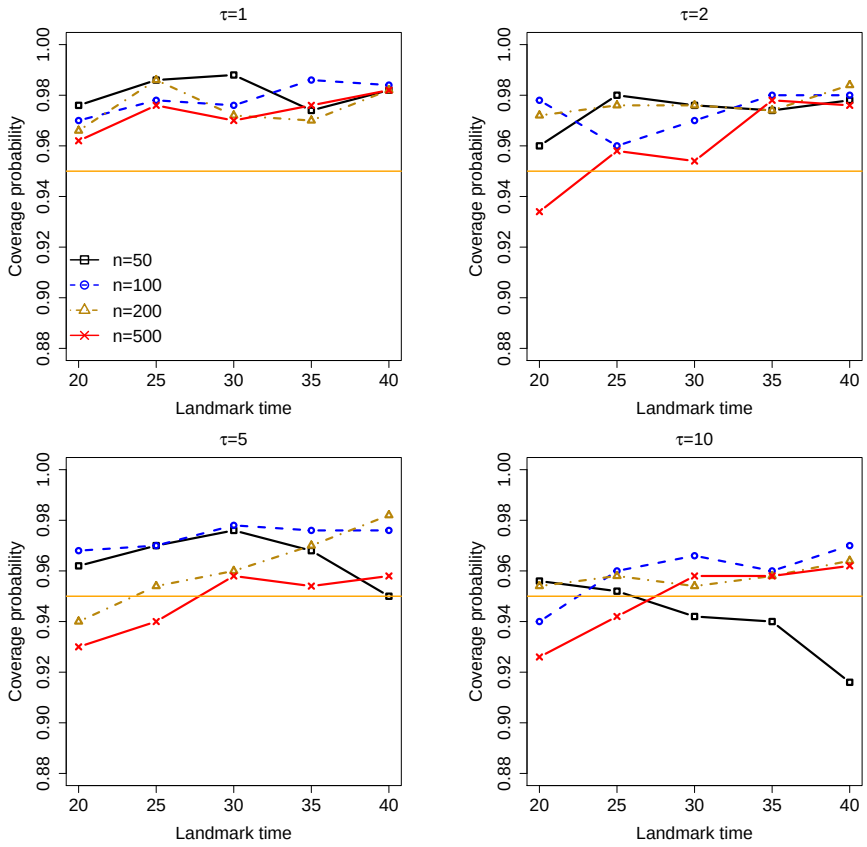

Figure 7: Empirical coverage probability of the proposed 95% prediction intervals for the number $N_{t,\tau}$ of failures within $(t, t+\tau]$ under different values of $n \in \{50, 100, 200, 500\}$, $t \in \{20, 25, 30, 35, 40\}$, and $\tau \in \{1, 2, 5, 10\}$.

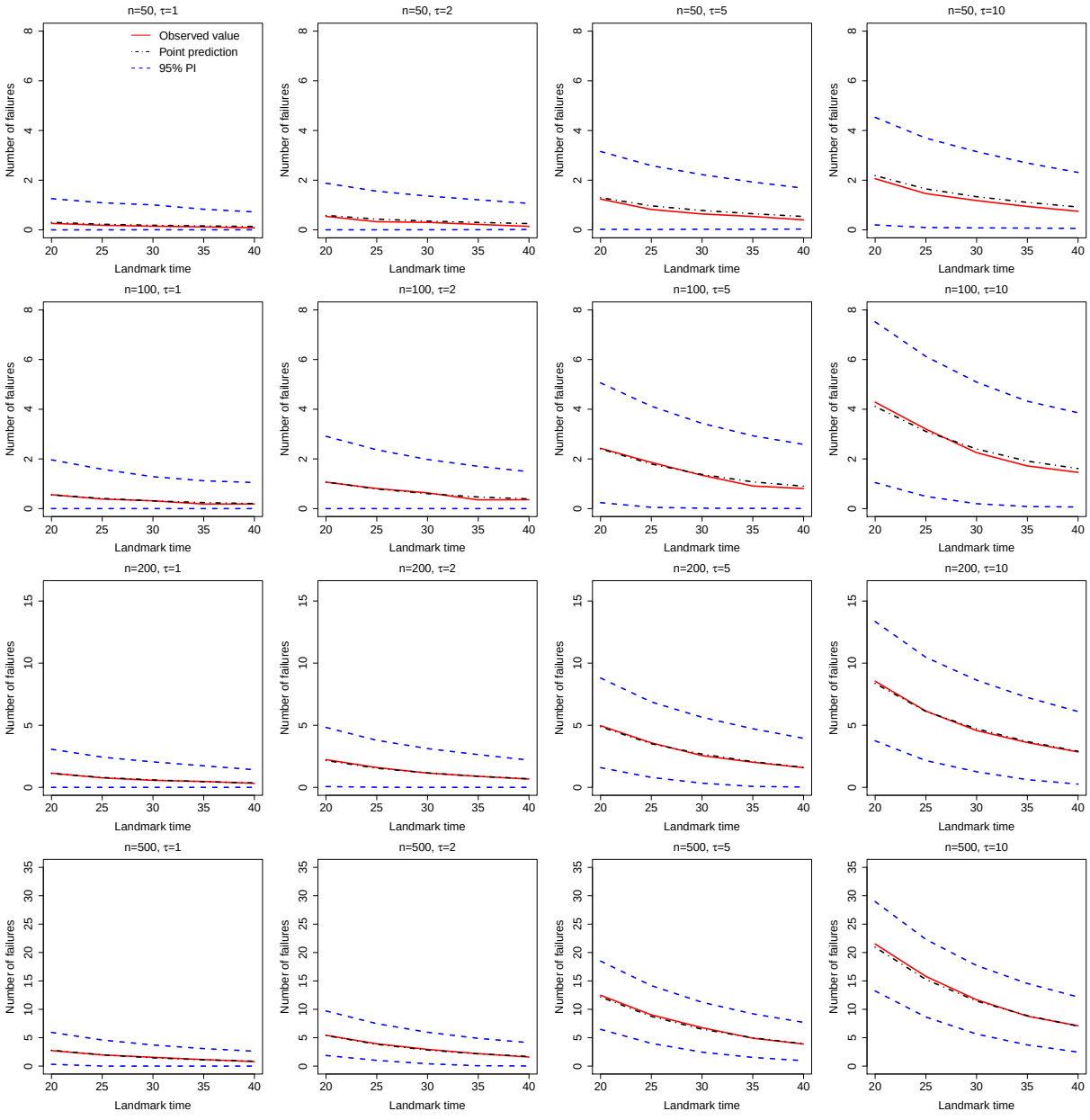

Figure 8: Average upper and lower limits of the proposed 95% prediction intervals for the number $N_{t,\tau}$ of failures within $(t, t+\tau]$, together with the average of the observed values of $N_{t,\tau}$ under different values of $n \in \{50, 100, 200, 500\}$, $t \in \{20, 25, 30, 35, 40\}$, and $\tau \in \{1, 2, 5, 10\}$.

As elucidated in our theoretical and methodological discussions, the extent of uncertainty quantification crucially hinges on the chosen modeling approach. In our study, we utilize the log-linear model in tandem with a composite likelihood. However, alternative modeling approaches, such as the semi-parametric framework commonly applied in landmark analysis, are also applicable but more challenging. Specifically, the nonparametric component introduces an additional infinite-dimensional parameter for estimation. A successful investigation in this context necessitates addressing weak convergence issues of the nonparametric component and understanding the dependencies between the nonparametric and parametric components. We plan to explore and address these challenges in a future project.

Acknowledgments

We sincerely thank the Editor, Eric Laber, for his thoughtful feedback throughout the review process. We are also grateful to the anonymous reviewers for their constructive comments and insightful suggestions, which have substantially improved the quality of this work. Wang and Ye were supported in part by Singapore MOE AcRF Tier 2 grant (A-8001052-00-00, A-8002472-00-00). Tang was supported in part by subaward of NIH R01GM140476 and NSF DMS-2210687.

Appendix A. Interval Estimations for Restricted Mean Residual Lifetime

Our approach in Section 3.3 can be extended to construct confidence intervals for the restricted mean residual lifetime $E[\min(T-t,\tau) \mid \mathcal{F}_t, T \geq t]$ (Tian et al., 2020). Specifically, it is defined as

$$\mathrm{RMRL}_t(\tau) = E[\min(T-t,\tau) \mid \mathcal{F}_t, T \geq t] = \int_0^\tau S_{T,t}(u)du, \tag{13}$$

which can be considered as the expected residual lifetime up to the pre-determined time horizon τ.

The point estimator for $\mathrm{RMRL}_t(\tau)$ can be directly obtained as

$$\hat{\mathrm{RMRL}}_t(\tau) = \int_0^\tau \hat{S}_{T,t}(u)du = \int_0^\tau \exp\left\{-\int_0^u \exp\left\{\hat{\boldsymbol{\alpha}}_n(u)^\top \boldsymbol{Z}(t)\right\}\right\} du,$$

where $\hat{\boldsymbol{\alpha}}_n(u)$ is the proposed estimator for the functional-valued parameter $\boldsymbol{\alpha} : [0,\tau] \to \mathbb{R}^p$. To quantify the uncertainties associated with the point estimator $\hat{\mathrm{RMRL}}_t(\tau)$, we need to further establish the asymptotic normality of $\hat{\mathrm{RMRL}}_t(\tau)$. This can be established using the weak convergence of the functional-valued estimator $\hat{\boldsymbol{\alpha}}_n(u)$ (see Theorem 1), together with the functional delta method (van der Vaart, 1998, Theorem 20.8). The following proposition summarizes the asymptotic normality of $\hat{\mathrm{RMRL}}_t(\tau)$.

Proposition 4 *Suppose Conditions (A1)–(A3) hold. Then we have*

$$\sqrt{n}\big(\hat{RMRL}_t(\tau) - RMRL_t(\tau)\big) \rightsquigarrow N(0,\zeta^2),$$

where

$$\zeta^2 = \int_0^\tau \int_0^\tau S_{T,t}(u)S_{T,t}(v)Q(u,v)dudv,$$

with

$$Q(u,v) = \int_0^u \int_0^v \lambda_{T,t}(s_1)\lambda_{T,t}(s_2)\boldsymbol{Z}(t)^\top \boldsymbol{W}(s_1;\boldsymbol{\alpha}_0)^{-1}\boldsymbol{\varrho}(s_1,s_2,\psi)\boldsymbol{W}(s_2;\boldsymbol{\alpha}_0)^{-1}\boldsymbol{Z}(t)ds_1ds_2.$$

The asymptotic variance ζ^2 can be consistently estimated by

$$\hat{\zeta}_n^2 = \int_0^\tau \int_0^\tau \hat{S}_{T,t}(u)\hat{S}_{T,t}(v)\hat{Q}(u,v)dudv,$$

where

$$\hat{Q}(u,v) = \int_0^u \int_0^v \hat{\lambda}_{T,t}(s_1)\hat{\lambda}_{T,t}(s_2)\boldsymbol{Z}(t)^\top \hat{\boldsymbol{W}}_n(s_1;\hat{\boldsymbol{\alpha}}_n)^{-1}\hat{\boldsymbol{\varrho}}_n(s_1,s_2;\hat{\boldsymbol{\alpha}}_n)\hat{\boldsymbol{W}}_n(s_2;\hat{\boldsymbol{\alpha}}_n)^{-1}\boldsymbol{Z}(t)ds_1ds_2.$$

Based on this proposition, the two-sided equal-tailed $100(1-\alpha)\%$ confidence interval for $\mathrm{RMRL}_t(\tau)$ can be constructed as

$$\hat{\mathrm{RMRL}}_t(\tau) \pm z_{1-\alpha/2}\hat{\zeta}_n/\sqrt{n},$$

where $z_{1-\alpha/2}$ is the $(1-\alpha/2)$ upper quantile of the standard normal distribution.

Appendix B. Simultaneous Confidence Bands

In this subsection, we extend our approach to construct confidence bands for $S_{T,t}(\tau)$ and $F_t(\tau)$ over a bounded interval $[0, \bar{\tau}]$, where $\bar{\tau}$ is the largest forward-looking time of interest.

One important building block for the construction of simultaneous confidence bands is to show the weak convergence of $\hat{S}_{T,t} : [0, \bar{\tau}] \to \mathbb{R}$ and $\hat{F}_t : [0, \bar{\tau}] \to \mathbb{R}$. This can be established based on the weak convergence of the functional-valued estimators $\hat{\boldsymbol{\alpha}}_n$ and $\hat{\boldsymbol{\beta}}_n$ (see Theorem 1 in the main text), together with the functional delta method (van der Vaart, 1998, Theorem 20.8). Detailed weak convergence results are summarized in the following proposition.

Proposition 5 *Suppose Conditions (A1)–(A3) hold. Then $\sqrt{n}(\hat{S}_{T,t} - S_{T,t})$ weakly converges to a Gaussian process with mean zero and covariance function $\rho_S(u, v)$, $u, v \in [0, \bar{\tau}]$, given by*

$$\rho_S(u,v) = S_{T,t}(u)S_{T,t}(v) \int_0^u \int_0^v \lambda_{T,t}(s_1)\lambda_{T,t}(s_2)\boldsymbol{Z}(t)^\top \boldsymbol{\varrho}_T(s_1,s_2)\boldsymbol{Z}(t)ds_1ds_2,$$

where $\boldsymbol{\varrho}_T(s_1, s_2) = \boldsymbol{W}(s_1; \boldsymbol{\alpha}_0)^{-1}\boldsymbol{\varrho}(s_1, s_2, \psi)\boldsymbol{W}(s_2; \boldsymbol{\alpha}_0)^{-1}$. The covariance function can be consistently estimated by

$$\hat{\rho}_S(u,v) = \hat{S}_{T,t}(u)\hat{S}_{T,t}(v) \int_0^u \int_0^v \hat{\lambda}_{T,t}(s_1)\hat{\lambda}_{T,t}(s_2)\boldsymbol{Z}(t)^\top \hat{\boldsymbol{\varrho}}_T(s_1,s_2)\boldsymbol{Z}(t)ds_1ds_2,$$

where $\hat{\boldsymbol{\varrho}}_T(s_1, s_2) = \hat{\boldsymbol{W}}_n(s_1; \hat{\boldsymbol{\alpha}}_n)^{-1}\hat{\boldsymbol{\varrho}}_n(s_1, s_2; \hat{\boldsymbol{\alpha}}_n)\hat{\boldsymbol{W}}_n(s_2; \hat{\boldsymbol{\alpha}}_n)^{-1}$. Similarly, $\sqrt{n}(\hat{F}_t - F_t)$ also weakly converges to a Gaussian process with mean zero and covariance function $\rho_F(u, v) = \rho_F^{(1)}(u, v) - \rho_F^{(2)}(u, v) - \rho_F^{(3)}(u, v) + \rho_F^{(4)}(u, v)$, where $u, v \in [0, \bar{\tau}]$ and

$$\begin{aligned}
\rho_F^{(1)}(u,v) &= \int_0^u \int_0^v R(s_1,s_2)\boldsymbol{Z}(t)^\top \boldsymbol{\varrho}_T(s_1,s_2)\boldsymbol{Z}(t)ds_2ds_1,\\
\rho_F^{(2)}(u,v) &= \int_0^u \int_0^v R(s_1,s_2) \int_0^{s_2} \lambda_{T,t}(w)\boldsymbol{Z}(t)^\top \boldsymbol{\varrho}_T(s_1,w)\boldsymbol{Z}(t)dwds_2ds_1,\\
\rho_F^{(3)}(u,v) &= \int_0^u \int_0^v R(s_1,s_2) \int_0^{s_1} \lambda_{T,t}(w)\boldsymbol{Z}(t)^\top \boldsymbol{\varrho}_T(s_2,w)\boldsymbol{Z}(t)dwds_2ds_1,\\
\rho_F^{(4)}(u,v) &= \int_0^u \int_0^v R(s_1,s_2) \int_0^{s_1} \int_0^{s_2} \Big[\lambda_{T,t}(w_1)\lambda_{T,t}(w_2)\boldsymbol{Z}(t)^\top \boldsymbol{\varrho}_T(w_1,w_2)\boldsymbol{Z}(t) +\\
&\qquad \lambda_{C,t}(w_1)\lambda_{C,t}(w_2)\boldsymbol{Z}(t)^\top \boldsymbol{\varrho}_C(w_1,w_2)\boldsymbol{Z}(t)\Big]dw_2dw_1ds_2ds_1,
\end{aligned}$$

with $R(s_1, s_2) = S_{T,t}(s_1)S_{C,t}(s_1)\lambda_{T,t}(s_1)S_{T,t}(s_2)S_{C,t}(s_2)\lambda_{T,t}(s_2)$ and $\boldsymbol{\varrho}_C(s_1, s_2) = \boldsymbol{W}(s_1; \boldsymbol{\beta}_0)^{-1}\boldsymbol{\varrho}(s_1, s_2, \vartheta)\boldsymbol{W}(s_2; \boldsymbol{\beta}_0)^{-1}$. The covariance function can be consistently estimated

by $\hat{\rho}_F(u,v) = \hat{\rho}_F^{(1)}(u,v) - \hat{\rho}_F^{(2)}(u,v) - \hat{\rho}_F^{(3)}(u,v) + \hat{\rho}_F^{(4)}(u,v)$*, where* $u, v \in [0, \bar{\tau}]$ *and*

$$\begin{aligned}
\hat{\rho}_F^{(1)}(u,v) &= \int_0^u \int_0^v \hat{R}(s_1,s_2) \boldsymbol{Z}(t)^\top \hat{\boldsymbol{\varrho}}_T(s_1,s_2) \boldsymbol{Z}(t) ds_2 ds_1, \\
\hat{\rho}_F^{(2)}(u,v) &= \int_0^u \int_0^v \hat{R}(s_1,s_2) \int_0^{s_2} \hat{\lambda}_{T,t}(w) \boldsymbol{Z}(t)^\top \hat{\boldsymbol{\varrho}}_T(s_1,w) \boldsymbol{Z}(t) dw ds_2 ds_1, \\
\hat{\rho}_F^{(3)}(u,v) &= \int_0^u \int_0^v \hat{R}(s_1,s_2) \int_0^{s_1} \hat{\lambda}_{T,t}(w) \boldsymbol{Z}(t)^\top \hat{\boldsymbol{\varrho}}_T(s_2,w) \boldsymbol{Z}(t) dw ds_2 ds_1, \\
\hat{\rho}_F^{(4)}(u,v) &= \int_0^u \int_0^v \hat{R}(s_1,s_2) \int_0^{s_1} \int_0^{s_2} \Big[\hat{\lambda}_{T,t}(w_1) \hat{\lambda}_{T,t}(w_2) \boldsymbol{Z}(t)^\top \hat{\boldsymbol{\varrho}}_T(w_1,w_2) \boldsymbol{Z}(t) + \\
&\qquad \hat{\lambda}_{C,t}(w_1) \hat{\lambda}_{C,t}(w_2) \boldsymbol{Z}(t)^\top \hat{\boldsymbol{\varrho}}_C(w_1,w_2) \boldsymbol{Z}(t) \Big] dw_2 dw_1 ds_2 ds_1,
\end{aligned}$$

with $\hat{R}(s_1,s_2) = \hat{S}_{T,t}(s_1)\hat{S}_{C,t}(s_1)\hat{\lambda}_{T,t}(s_1)\hat{S}_{T,t}(s_2)\hat{S}_{C,t}(s_2)\hat{\lambda}_{T,t}(s_2)$ *and* $\hat{\boldsymbol{\varrho}}_C(s_1,s_2) = \hat{\boldsymbol{W}}_n(s_1;\hat{\boldsymbol{\beta}}_n)^{-1} \hat{\boldsymbol{\varrho}}_n(s_1,s_2;\hat{\boldsymbol{\beta}}_n) \hat{\boldsymbol{W}}_n(s_2;\hat{\boldsymbol{\beta}}_n)^{-1}$.

Based on the weak convergence results in Proposition 5, we can construct simultaneous confidence bands for $S_{T,t}$ and F_t following the approach in Section 3 of Li et al. (2021). Detailed procedures are summarized in Algorithm 2, in which we leverage the limiting Gaussian process for $\sqrt{n}(\hat{S}_{T,t} - S_{T,t})$ to account for the uncertainties in $\hat{S}_{T,t}$ over the entire interval $[0, \bar{\tau}]$. The procedure for F_t is similar and thus omitted.

Algorithm 2: Construction of simultaneous confidence bands for $S_{T,t}$

Input: Pointwise estimator $\hat{S}_{T,t}$ for $S_{T,t}$, variance estimator $\hat{\rho}_S$, number B of resampling, and significance level α

Output: $100(1-\alpha)\%$ confidence bands for $S_{T,t}(u)$ over $u \in [0, \bar{\tau}]$

$b \leftarrow 1$.

while $b \le B$ **do**

  Generate $S_{T,t}^{(b)} : [0, \bar{\tau}] \to [0,1]$ from the Gaussian process with mean $\hat{S}_{T,t}$ and covariance function $n^{-1}\hat{\rho}_S(u,v)$.

  Calculate $z^{(b)} = \sup_{u \in [0,\bar{\tau}]} \left| \sqrt{n}\big(S_{T,t}^{(b)}(u) - \hat{S}_{T,t}(u)\big) / \sqrt{\hat{\rho}_S(u,u)} \right|$.

end

Calculate the $(1-\alpha/2)$-th sample quantile $\tilde{z}_{\alpha/2}$ based on $\{z^{(b)}\}_{b=1}^B$.

Construct the $100(1-\alpha)\%$ confidence bands for $S_{T,t} : [0, \bar{\tau}] \to [0,1]$ as

$$\hat{S}_{T,t}(u) \pm n^{-1/2} \tilde{z}_{\alpha/2} \sqrt{\hat{\rho}_S(u,u)}, \quad \forall u \in [0, \bar{\tau}].$$

return $\hat{S}_{T,t}(u) \pm n^{-1/2} \tilde{z}_{\alpha/2} \sqrt{\hat{\rho}_S(u,u)}, \quad \forall u \in [0, \bar{\tau}]$.

Appendix C. Discrete-time Analysis

C.1 Discrete-time Approximation

In real data collection, the time-varying covariates are usually observed on a regular basis, e.g., every month or every half year. In view of this common data-collection scheme, this section further discusses inference for the forward intensity model under discrete observations. As with Duan et al. (2012), we consider an equally-spaced observational scheme with Δt being the basic time interval. In corporate default analysis, Δt is usually chosen as one month. Point estimation of the forward intensity models in (2) and (3) has been explored in Duan et al. (2012) under the discrete setting. Our main focus in this section is on the uncertainty quantification.

Consider a generic subject. Measurements of the subject's covariates are taken at $t_k = k\Delta t$, $k = 1, 2, \cdots, M$, where M is random and $M\Delta t = X$ is the maximum follow-up time. Let $\tilde{\delta}_T(t_k) = N_T(t_k) - N_T(t_{k-1})$ and $\tilde{\delta}_C(t_k) = N_C(t_k) - N_C(t_{k-1})$ be the event indicators at the kth observation time, $k = 1, \cdots, M$. The observed data for the subject are $\mathbf{D} = \{M, \delta_T(t_j), \delta_C(t_j), \boldsymbol{Z}(t_j); j = 1, \ldots, M\}$. Under this observational scheme, the event times are only know to lie within certain intervals, and only one point of $\boldsymbol{Z}(t)$ is observed over the interval $(t_{j-1}, t_j]$, $j \in \mathbb{N}$. As a result, the continuous-time composite log-likelihood in (8) can only be evaluated approximately in this case. A natural and feasible approach is to approximate the two functions $\boldsymbol{\alpha} : [0, \tau] \mapsto \mathbb{R}^p$ and $\boldsymbol{\beta} : [0, \tau] \mapsto \mathbb{R}^p$ with piecewise constants over the intervals between successive observation times:

$$\boldsymbol{\alpha}_*(u) = \sum_{k=0}^{\tau/\Delta t-1} \boldsymbol{\alpha}(t_k) I\{t_k \le u < t_{k+1}\}, \quad \boldsymbol{\beta}_*(u) = \sum_{k=0}^{\tau/\Delta t-1} \boldsymbol{\beta}(t_k) I\{t_k \le u < t_{k+1}\}. \tag{14}$$

Suppose we are interested in evaluating the survival probability of an subject $\tau \triangleq \kappa\Delta t$ units of time ahead of the current time. Then the parameters to estimate under the discrete setting are $\underline{\boldsymbol{\alpha}}^\top = \big(\boldsymbol{\alpha}(t_0)^\top, \boldsymbol{\alpha}(t_1)^\top, \ldots, \boldsymbol{\alpha}(t_{\kappa-1})^\top\big)$ and $\underline{\boldsymbol{\beta}}^\top = \big(\boldsymbol{\beta}(t_0)^\top, \boldsymbol{\beta}(t_1)^\top, \ldots, \boldsymbol{\beta}(t_{\kappa-1})^\top\big)$. As with Duan et al. (2012), the composite log-likelihood of $(\underline{\boldsymbol{\alpha}}, \underline{\boldsymbol{\beta}})$ can be obtained based on the discrete-time approximation in (14):

$$\begin{aligned}
\tilde{\ell}_n\left(\underline{\boldsymbol{\alpha}}; \mathbf{D}_{\text{disct}}\right) &= \frac{1}{n} \sum_{k=0}^{\kappa-1} \sum_{i=1}^{n} \sum_{j=1}^{M_i-k-1} \Big\{ -\lambda_{T_i,t_j}(t_k)\Delta t + \tilde{\delta}_{T_i}(t_j + t_{k+1}) \log \frac{1 - \exp\{-\lambda_{T_i,t_j}(t_k)\Delta t\}}{\exp\{-\lambda_{T_i,t_j}(t_k)\Delta t\}} \Big\}, \\
\tilde{\ell}_n\left(\underline{\boldsymbol{\beta}}; \mathbf{D}_{\text{disct}}\right) &= \frac{1}{n} \sum_{k=0}^{\kappa-1} \sum_{i=1}^{n} \sum_{j=1}^{M_i-k-1} \Big\{ -\lambda_{C_i,t_j}(t_k)\Delta t + \tilde{\delta}_{C_i}(t_j + t_{k+1}) \log \frac{1 - \exp\{-\lambda_{C_i,t_j}(t_k)\Delta t\}}{\exp\{-\lambda_{C_i,t_j}(t_k)\Delta t\}} \Big\},
\end{aligned} \tag{15}$$

where $\mathbf{D}_{\text{disct}} = \{M_i, \tilde{\delta}_{T_i}(t_j), \tilde{\delta}_{C_i}(t_j), \boldsymbol{Z}_i(t_j); j = 1, 2, \ldots, M_i\}_{i=1}^n$ denotes the collection of the observed data for all subjects with M_i being the total number of observations of the ith subject. As $\Delta t \to 0$, the above two composite log-likelihoods converge to the continuous-time composite log-likelihoods in (8). Thus (15) is essentially a discrete-time approximation of the composite log-likelihood under the continuous setting.

Similar to the continuous case, $\tilde{\ell}_n\left(\underline{\boldsymbol{\alpha}}; \mathbf{D}_{\text{disct}}\right)$ and $\tilde{\ell}_n(\underline{\boldsymbol{\beta}}; \mathbf{D}_{\text{disct}})$ are strictly concave w.r.t. $\underline{\boldsymbol{\alpha}}$ and $\underline{\boldsymbol{\beta}}$, respectively. Thus the maximum composite likelihood estimators exist and are

unique. We denote these two estimators by $\hat{\underline{\boldsymbol{\alpha}}}_n^\top = \big(\hat{\boldsymbol{\alpha}}_n(t_0)^\top, \hat{\boldsymbol{\alpha}}_n(t_1)^\top, \ldots, \hat{\boldsymbol{\alpha}}_n(t_{\kappa-1})^\top\big)$ and $\hat{\underline{\boldsymbol{\beta}}}_n^\top = \big(\hat{\boldsymbol{\beta}}_n(t_0)^\top, \hat{\boldsymbol{\beta}}_n(t_1)^\top, \ldots, \hat{\boldsymbol{\beta}}_n(t_{\kappa-1})^\top\big)$, respectively.

C.2 Discrete-time Asymptotic Properties

Suppose the true functional parameters $\boldsymbol{\alpha}_0 : [0, \tau] \mapsto \mathbb{R}^p$ and $\boldsymbol{\beta}_0 : [0, \tau] \mapsto \mathbb{R}^p$ are piecewise constants over the intervals between successive observation times. That is, the following equality holds:

$$\boldsymbol{\alpha}_0(u) = \sum_{k=0}^{\kappa-1} \boldsymbol{\alpha}_0(t_k) I\{t_k \le u < t_{k+1}\}, \quad \boldsymbol{\beta}_0(u) = \sum_{k=0}^{\kappa-1} \boldsymbol{\beta}(t_k) I\{t_k \le u < t_{k+1}\}.$$

Then the composite log-likelihoods in (15) are equal to the continuous-time composite log-likelihoods in (8). In this case, the uncertainties associated with the point predictions can be quantified through establishing the asymptotic properties of the estimators $(\hat{\underline{\boldsymbol{\alpha}}}_n, \hat{\underline{\boldsymbol{\beta}}}_n)$, as detailed below.

In the discrete setting, both $\hat{\underline{\boldsymbol{\alpha}}}_n$ and $\hat{\underline{\boldsymbol{\beta}}}_n$ are κp-dimensional random vectors because $\kappa \triangleq \tau/\Delta t$ is finite. Hence the asymptotic normality of $(\hat{\underline{\boldsymbol{\alpha}}}_n, \hat{\underline{\boldsymbol{\beta}}}_n)$ follows from the standard M-estimation theory (Duan et al., 2012). We first summarize the results on the asymptotic normality of $(\hat{\underline{\boldsymbol{\alpha}}}_n, \hat{\underline{\boldsymbol{\beta}}}_n)$ together with its consistent variance estimator, based on which the asymptotic normality of the estimated predictive probabilities can then be established using the delta method.

Denote the parameter space for the two parameters $\underline{\boldsymbol{\alpha}}$ and $\underline{\boldsymbol{\beta}}$ as $\mathcal{F}_0$. Denote $\underline{\boldsymbol{\alpha}}_0^\top = \big(\boldsymbol{\alpha}_0(t_0)^\top, \boldsymbol{\alpha}_0(t_1)^\top, \ldots, \boldsymbol{\alpha}_0(t_{\kappa-1})^\top\big)$ and $\underline{\boldsymbol{\beta}}_0^\top = \big(\boldsymbol{\beta}_0(t_0)^\top, \boldsymbol{\beta}_0(t_1)^\top, \ldots, \boldsymbol{\beta}_0(t_{\kappa-1})^\top\big)$ as the truth. For the asymptotic normality of $(\hat{\underline{\boldsymbol{\alpha}}}_n, \hat{\underline{\boldsymbol{\beta}}}_n)$, the following regularity conditions are needed.

(B1) The parameter space $\mathcal{F}_0$ is compact.

(B2) The covariate process $\{\boldsymbol{Z}(t_k) : k \in \mathbb{N}^+\}$ is bounded, and $E\boldsymbol{Z}(t_k)\boldsymbol{Z}(t_k)^\top$ is positive definite for $k \in \mathbb{N}^+$.

(B3) $E[X^2] < \infty$ and $P(X \ge \tau) > 0$, where $X = \min(T, C, \tilde{C})$.

All above conditions are mild. Similar to Conditions (A1) and (A2) under the continuous setting, Conditions (B1) and (B2) require the parameter space and the discrete covariate process to be bounded, and it is a common regularity condition in the literature (Wellner and Zhang, 2007; Zhao et al., 2017). In Condition (B3), the requirement on the distribution of the maximum follow-up time X is commonly assumed in survival analysis (Ye et al., 2015; Chan and Wang, 2017).

Theorem 6 *Suppose Conditions (B1)–(B3) hold. Then* $(\hat{\underline{\boldsymbol{\alpha}}}_n, \hat{\underline{\boldsymbol{\beta}}}_n)$ *converges to* $(\underline{\boldsymbol{\alpha}}_0, \underline{\boldsymbol{\beta}}_0)$ *in probability, and we have*

$$\sqrt{n}\left(\begin{bmatrix}\hat{\underline{\boldsymbol{\alpha}}}_n \\ \hat{\underline{\boldsymbol{\beta}}}_n\end{bmatrix} - \begin{bmatrix}\underline{\boldsymbol{\alpha}}_0 \\ \underline{\boldsymbol{\beta}}_0\end{bmatrix}\right) \rightsquigarrow N(\mathbf{0}, \boldsymbol{\Sigma}),$$

where $\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_1 & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_2 \end{bmatrix}$ *with*

$$\begin{aligned}\boldsymbol{\Sigma}_1 &= (E[\nabla^2 \tilde{\ell}_n(\underline{\boldsymbol{\alpha}}_0)])^{-1}\mathrm{Var}[\sqrt{n}\nabla\tilde{\ell}_n(\underline{\boldsymbol{\alpha}}_0)](E[\nabla^2\tilde{\ell}_n(\underline{\boldsymbol{\alpha}}_0)])^{-1}\\ \boldsymbol{\Sigma}_2 &= (E[\nabla^2 \tilde{\ell}_n(\underline{\boldsymbol{\beta}}_0)])^{-1}\mathrm{Var}[\sqrt{n}\nabla\tilde{\ell}_n(\underline{\boldsymbol{\beta}}_0)](E[\nabla^2\tilde{\ell}_n(\underline{\boldsymbol{\beta}}_0)])^{-1}.\end{aligned}$$

Furthermore, the asymptotic variance $\boldsymbol{\Sigma}$ *can be consistently estimated by* $\hat{\boldsymbol{\Sigma}}_n = \begin{bmatrix} \hat{\boldsymbol{\Sigma}}_{1,n} & \mathbf{0} \\ \mathbf{0} & \hat{\boldsymbol{\Sigma}}_{2,n} \end{bmatrix}$, *where the detailed expression of* $\hat{\boldsymbol{\Sigma}}_n$ *is shown in* (50).

In order to construct confidence intervals for the predictive probabilities $S_{T,t}(\tau)$ and $F_t(\tau)$ and the restricted mean residual lifetime $\mathrm{RMRL}_t(\tau)$ under the discrete setting, we need the weak convergence of the estimators of $S_{T,t}(\tau)$, $F_t(\tau)$, and $\mathrm{RMRL}_t(\tau)$. In contrast to the integral forms of $S_{T,t}(\tau)$, $F_t(\tau)$, and $\mathrm{RMRL}_t(\tau)$ in (1), (4), and (13), these quantities take the forms of sums under the discrete setting:

$$\begin{aligned} S_{T,t}(\tau) &= P(T \geq t+\tau|\mathcal{F}_t, T\geq t) = \exp\big\{-\sum_{k=0}^{\kappa-1}\lambda_{T,t}(t_k)\Delta t\big\},\\ F_t(\tau) &= \sum_{h=0}^{\kappa-1} P(t+h\Delta t \leq T\wedge C = T < t+(h+1)\Delta t|\mathcal{F}_t, T\wedge C\geq t) \qquad (16)\\ &= \sum_{h=1}^{\kappa-1}\exp\{-\sum_{k=0}^{h-1}\lambda_{T,t}(t_k)\Delta t - \sum_{k=0}^{h}\lambda_{C,t}(t_k)\Delta t\}\left[1-\exp\{-\lambda_{T,t}(t_h)\Delta t\}\right]\\ &\qquad + \exp\{-\lambda_{C,t}(t_0)\Delta t\}\left[1-\exp\{-\lambda_{T,t}(t_0)\Delta t\}\right],\\ \mathrm{RMRL}_t(\tau) &= \frac{\Delta t}{2}\Big[1+2\sum_{k=1}^{\kappa-1}\exp\big\{-\sum_{l=0}^{k-1}\lambda_{T,t}(t_l)\Delta t\big\} + \exp\big\{-\sum_{l=0}^{\kappa-1}\lambda_{T,t}(t_l)\Delta t\big\}\Big]. \end{aligned}$$

The point estimators $\hat{S}_{T,t}(\tau)$, $\hat{F}_t(\tau)$, $\hat{\mathrm{RMRL}}_t(\tau)$ of $S_{T,t}(\tau)$, $F_t(\tau)$, and $\mathrm{RMRL}_t(\tau)$ can be readily obtained by replacing $\lambda_{T,t}(t_k)$ and $\lambda_{C,t}(t_k)$ in (16) with the corresponding plug-in estimators $\hat{\lambda}_{T,t}(t_k) = \exp\{\hat{\underline{\boldsymbol{\alpha}}}_n(t_k)^\top \boldsymbol{Z}(t)\}$ and $\hat{\lambda}_{C,t}(t_k) = \exp\{\hat{\underline{\boldsymbol{\beta}}}_n(t_k)^\top \boldsymbol{Z}(t)\}$, $k = 0,1,2,\ldots,\kappa-1$ (Duan et al., 2012). Similar to the continuous-time setting, the point estimators $\hat{S}_{T,t}(\tau)$, $\hat{F}_t(\tau)$, and $\hat{\mathrm{RMRL}}_t(\tau)$ under the discrete-time setting are functions of $(\hat{\underline{\boldsymbol{\alpha}}}_n, \hat{\underline{\boldsymbol{\beta}}}_n)$, and thus the weak convergence of the three point estimators follows from Theorem 6 and the delta method, as summarized in the following proposition.

Proposition 7 *Suppose Conditions (B1)–(B3) hold. Then we have*

$$\begin{aligned}&\sqrt{n}\big(\hat{S}_{T,t}(\tau) - S_{T,t}(\tau)\big) \rightsquigarrow N(0,\sigma_0^2),\\ &\sqrt{n}\big(\hat{F}_t(\tau) - F_t(\tau)\big) \rightsquigarrow N(0,\eta_0^2),\\ &\sqrt{n}\big(\hat{RMRL}_t(\tau) - RMRL_t(\tau)\big) \rightsquigarrow N(0,\varsigma_0^2),\end{aligned}$$

where $\sigma_0^2 = \nabla_{\underline{\boldsymbol{\alpha}}} S_{T,t}(\tau)\boldsymbol{\Sigma}_1\nabla_{\underline{\boldsymbol{\alpha}}}S_{T,t}(\tau)^\top$, $\eta_0^2 = \nabla_{\underline{\boldsymbol{\alpha}}}F_t(\tau)\boldsymbol{\Sigma}_1\nabla_{\underline{\boldsymbol{\alpha}}}F_t(\tau)^\top + \nabla_{\underline{\boldsymbol{\beta}}}F_t(\tau)\boldsymbol{\Sigma}_2\nabla_{\underline{\boldsymbol{\beta}}}F_t(\tau)^\top$, *and* $\varsigma_0^2 = \nabla_{\underline{\boldsymbol{\alpha}}}RMRL_t(\tau)\boldsymbol{\Sigma}_1\nabla_{\underline{\boldsymbol{\alpha}}}RMRL_t(\tau)^\top$. *In addition, the asymptotic variances* σ_0^2 *and*

η_0^2 *can be consistently estimated by*

$$\hat{\sigma}_0^2 = \nabla_{\underline{\boldsymbol{\alpha}}} S_{T,t}(\tau)|_{\underline{\boldsymbol{\alpha}}=\hat{\underline{\boldsymbol{\alpha}}}_n} \hat{\boldsymbol{\Sigma}}_{1,n} \nabla_{\underline{\boldsymbol{\alpha}}} S_{T,t}(\tau)^\top|_{\underline{\boldsymbol{\alpha}}=\hat{\underline{\boldsymbol{\alpha}}}_n},$$
$$\hat{\eta}_0^2 = \nabla_{\underline{\boldsymbol{\alpha}}} F_t(\tau)|_{\underline{\boldsymbol{\alpha}}=\hat{\underline{\boldsymbol{\alpha}}}_n} \hat{\boldsymbol{\Sigma}}_{1,n} \nabla_{\underline{\boldsymbol{\alpha}}} F_t(\tau)^\top|_{\underline{\boldsymbol{\alpha}}=\hat{\underline{\boldsymbol{\alpha}}}_n} + \nabla_{\underline{\boldsymbol{\beta}}} F_t(\tau)|_{\underline{\boldsymbol{\beta}}=\hat{\underline{\boldsymbol{\beta}}}_n} \hat{\boldsymbol{\Sigma}}_{2,n} \nabla_{\underline{\boldsymbol{\beta}}} F_t(\tau)^\top|_{\underline{\boldsymbol{\beta}}=\hat{\underline{\boldsymbol{\beta}}}_n},$$
$$\hat{\varsigma}_0^2 = \nabla_{\underline{\boldsymbol{\alpha}}} RMRL_t(\tau)|_{\underline{\boldsymbol{\alpha}}=\hat{\underline{\boldsymbol{\alpha}}}_n} \hat{\boldsymbol{\Sigma}}_{1,n} \nabla_{\underline{\boldsymbol{\alpha}}} RMRL_t(\tau)^\top|_{\underline{\boldsymbol{\alpha}}=\hat{\underline{\boldsymbol{\alpha}}}_n}.$$

Based on the above asymptotic result, the uncertainties associated with $\hat{S}_{T,t}(\tau)$, $\hat{F}_t(\tau)$, and $\hat{\text{RMRL}}_t(\tau)$ can be quantified by constructing confidence intervals for $S_{T,t}(\tau)$, $F_t(\tau)$, and $\text{RMRL}_t(\tau)$ through the large sample normal approximation. In detail, the two-sided equal-tailed $100(1-\alpha)\%$ confidence intervals of $S_{T,t}(\tau)$, $F_t(\tau)$, and $\text{RMRL}_t(\tau)$ can be constructed as

$$\hat{S}_{T,t}(\tau) \pm z_{1-\alpha/2}\hat{\sigma}_0/\sqrt{n}, \quad \hat{F}_t(\tau) \pm z_{1-\alpha/2}\hat{\eta}_0/\sqrt{n}, \quad \text{and} \quad \hat{\text{RMRL}}_t(\tau) \pm z_{1-\alpha/2}\hat{\varsigma}_0/\sqrt{n}.$$

C.3 Simultaneous Confidence Intervals

In parallel with Appendix B, this subsection considers the simultaneous confidence intervals for $\underline{\boldsymbol{S}}_t \triangleq (S_{T,t}(t_1), S_{T,t}(t_2), \ldots, S_{T,t}(t_\kappa))^\top$ and $\underline{\boldsymbol{F}}_t \triangleq (F_t(t_1), F_t(t_2), \ldots, F_t(t_\kappa))^\top$ under the discrete-time setting. To this end, we need to establish the asymptotic normality of $(\hat{S}_{T,t}(t_1), \hat{S}_{T,t}(t_2), \ldots, \hat{S}_{T,t}(t_\kappa))$ and $(\hat{F}_t(t_1), \hat{F}_t(t_2), \ldots, \hat{F}_t(t_\kappa))$. This follows from the asymptotic normality of $(\hat{\underline{\boldsymbol{\alpha}}}_n, \hat{\underline{\boldsymbol{\beta}}}_n)$ in Theorem 6 and the delta method, as summarized below.

Proposition 8 *Suppose Conditions (B1)–(B3) hold. Then we have*

$$\sqrt{n}\left(\begin{bmatrix}\hat{S}_{T,t}(t_1)\\ \hat{S}_{T,t}(t_2)\\ \vdots \\ \hat{S}_{T,t}(t_\kappa)\end{bmatrix} - \begin{bmatrix}S_{T,t}(t_1)\\ S_{T,t}(t_2)\\ \vdots \\ S_{T,t}(t_\kappa)\end{bmatrix}\right) \rightsquigarrow N(\mathbf{0}, \boldsymbol{\Sigma}_S),$$
$$\sqrt{n}\left(\begin{bmatrix}\hat{F}_t(t_1)\\ \hat{F}_t(t_2)\\ \vdots \\ \hat{F}_t(t_\kappa)\end{bmatrix} - \begin{bmatrix}F_t(t_1)\\ F_t(t_2)\\ \vdots \\ F_t(t_\kappa)\end{bmatrix}\right) \rightsquigarrow N(\mathbf{0}, \boldsymbol{\Sigma}_F),$$

where

$$\boldsymbol{\Sigma}_S = (\nabla_{\underline{\boldsymbol{\alpha}}} \underline{\boldsymbol{S}}_t)\boldsymbol{\Sigma}_1(\nabla_{\underline{\boldsymbol{\alpha}}} \underline{\boldsymbol{S}}_t^\top),$$
$$\boldsymbol{\Sigma}_F = (\nabla_{\underline{\boldsymbol{\alpha}}} \underline{\boldsymbol{F}}_t)\boldsymbol{\Sigma}_1(\nabla_{\underline{\boldsymbol{\alpha}}} \underline{\boldsymbol{F}}_t^\top) + (\nabla_{\underline{\boldsymbol{\beta}}} \underline{\boldsymbol{F}}_t)\boldsymbol{\Sigma}_2(\nabla_{\underline{\boldsymbol{\beta}}} \underline{\boldsymbol{F}}_t^\top).$$

The asymptotic variance can be consistently estimated by

$$\hat{\boldsymbol{\Sigma}}_S = (\nabla_{\underline{\boldsymbol{\alpha}}} \underline{\boldsymbol{S}}_t|_{\underline{\boldsymbol{\alpha}}=\hat{\underline{\boldsymbol{\alpha}}}_n})\hat{\boldsymbol{\Sigma}}_{1,n}(\nabla_{\underline{\boldsymbol{\alpha}}} \underline{\boldsymbol{S}}_t^\top|_{\underline{\boldsymbol{\alpha}}=\hat{\underline{\boldsymbol{\alpha}}}_n}),$$
$$\hat{\boldsymbol{\Sigma}}_F = (\nabla_{\underline{\boldsymbol{\alpha}}} \underline{\boldsymbol{F}}_t|_{\underline{\boldsymbol{\alpha}}=\hat{\underline{\boldsymbol{\alpha}}}_n})\hat{\boldsymbol{\Sigma}}_{1,n}(\nabla_{\underline{\boldsymbol{\alpha}}} \underline{\boldsymbol{F}}_t^\top|_{\underline{\boldsymbol{\alpha}}=\hat{\underline{\boldsymbol{\alpha}}}_n}) + (\nabla_{\underline{\boldsymbol{\beta}}} \underline{\boldsymbol{F}}_t|_{\underline{\boldsymbol{\beta}}=\hat{\underline{\boldsymbol{\beta}}}_n})\hat{\boldsymbol{\Sigma}}_{2,n}(\nabla_{\underline{\boldsymbol{\beta}}} \underline{\boldsymbol{F}}_t^\top|_{\underline{\boldsymbol{\beta}}=\hat{\underline{\boldsymbol{\beta}}}_n}).$$

Based on the weak convergence results in Proposition 8, we can construct simultaneous confidence bands for $\underline{\boldsymbol{S}}_t$ and $\underline{\boldsymbol{F}}_t$ using a similar approach as in Appendix B. Detailed procedures for $\underline{\boldsymbol{S}}_t$ are summarized in Algorithm 3. The procedures for $\underline{\boldsymbol{F}}_t$ are similar and thus omitted.

Algorithm 3: Construction of simultaneous confidence bands for $\underline{\boldsymbol{S}}_t$

Input: Point estimator $(\hat{S}_{T,t}(t_1), \hat{S}_{T,t}(t_2), \ldots, \hat{S}_{T,t}(t_\kappa))$ for $\underline{\boldsymbol{S}}_t$, variance estimator $\hat{\boldsymbol{\Sigma}}_S$, number B of resampling, and significance level α

Output: $100(1-\alpha)\%$ confidence bands for $\underline{\boldsymbol{S}}_t$

$b \leftarrow 1$.

while $b \leq B$ **do**

  Generate $\big(S^{(b)}_{T,t}(t_1), S^{(b)}_{T,t}(t_2), \ldots, S^{(b)}_{T,t}(t_\kappa)\big)$ from the normal distribution with mean $(\hat{S}_{T,t}(t_1), \hat{S}_{T,t}(t_2), \ldots, \hat{S}_{T,t}(t_\kappa))$ and variance $n^{-1}\hat{\boldsymbol{\Sigma}}_S$.

  Calculate $z^{(b)} = \max_{k\in\{1,2,\ldots,\kappa\}} \left|\sqrt{n}\big(S^{(b)}_{T,t}(t_k) - \hat{S}_{T,t}(t_k)\big)/\sqrt{\hat{\boldsymbol{\Sigma}}_S(k)}\right|$, where $\hat{\boldsymbol{\Sigma}}_S(k)$ is the kth diagonal element of $\hat{\boldsymbol{\Sigma}}_S$.

end

Calculate the $(1-\alpha/2)$-th sample quantile $\tilde{z}_{\alpha/2}$ based on $\{z^{(b)}\}_{b=1}^B$.

Construct the $100(1-\alpha)\%$ confidence bands for $\underline{\boldsymbol{S}}_t$ as

$$\hat{S}_{T,t}(t_k) \pm n^{-1/2}\tilde{z}_{\alpha/2}\sqrt{\hat{\boldsymbol{\Sigma}}_S(k)}, \quad \forall k \in \{1,2,\ldots,\kappa\}.$$

return $\hat{S}_{T,t}(t_k) \pm n^{-1/2}\tilde{z}_{\alpha/2}\sqrt{\hat{\boldsymbol{\Sigma}}_S(k)},\ \forall k \in \{1,2,\ldots,\kappa\}$.

C.4 Aggregate Number of Future Events

Similar to the continuous case, the prediction interval for $N_{t,\tau}$ under the discrete setting can be constructed using the resampling method. The resampling method under the discrete setting is almost the same as that in Section 3.4. The detailed procedure is summarized in Algorithm 4.

Algorithm 4: Resampling method for prediction intervals under discrete settings

Input: Significance level α, repetition B, and estimators $\hat{\underline{\boldsymbol{\alpha}}}_n, \hat{\underline{\boldsymbol{\beta}}}_n, \hat{\boldsymbol{\Sigma}}_{1,n}, \hat{\boldsymbol{\Sigma}}_{2,n}$.

Output: Prediction interval $(\underline{N}_{t,\tau}, \overline{N}_{t,\tau})$.

Initialization: $b \leftarrow 1$.

while $b \leq B$ **do**

  Draw $\underline{\boldsymbol{\alpha}}^{(b)}$ from $N(\hat{\underline{\boldsymbol{\alpha}}}_n, \hat{\boldsymbol{\Sigma}}_{1,n}/n)$. Similarly, draw $\underline{\boldsymbol{\beta}}^{(b)}$ from $N(\hat{\underline{\boldsymbol{\beta}}}_n, \hat{\boldsymbol{\Sigma}}_{2,n}/n)$.

  Compute $F^{(b)}_{i,t}(\tau)$ according to the sampled parameters $(\underline{\boldsymbol{\alpha}}^{(b)}, \underline{\boldsymbol{\beta}}^{(b)})$, $i \in R(t)$.

  Draw $N^{(b)}_{t,\tau}$ from the Poisson binomial distribution with parameters $F^{(b)}_{i,t}(\tau)$, $i \in R(t)$. $b \leftarrow b+1$.

end

Based on $\{N^{(b)}_{t,\tau}\}_{b=1}^B$, compute the $100(1-\alpha)\%$ PI $(\underline{N}_{t,\tau}, \overline{N}_{t,\tau})$.

Return $(\underline{N}_{t,\tau}, \overline{N}_{t,\tau})$.

Appendix D. More Details about the Real Examples and Simulations

D.1 U.S. Corporate Default Data

In this example, the detailed expression of the forward intensity $\lambda_{C,t_j}(t_k)$ of the random censoring time C is

$$\begin{aligned}\lambda_{C,t_j}(t_k) = \exp\big\{&\beta_0(t_k) + \beta_1(t_k)\text{SP500} + \beta_2(t_k)\text{TreasRate} + \beta_3(t_k)\text{DTD}_{\text{level}}(t_j) + \\ &\beta_4(t_k)\text{DTD}_{\text{trend}}(t_j) + \beta_5(t_k)\text{CASH/TA}_{\text{level}}(t_j) + \beta_6(t_k)\text{CASH/TA}_{\text{trend}}(t_j) + \\ &\beta_7(t_k)\text{NI/TA}_{\text{level}}(t_j) + \beta_8(t_k)\text{NI/TA}_{\text{trend}}(t_j) + \beta_9(t_k)\text{SIZE}_{\text{level}}(t_j) + \\ &\beta_{10}(t_k)\text{SIZE}_{\text{trend}}(t_j) + \beta_{11}(t_k)\text{M/B}(t_j) + \beta_{12}(t_k)\text{SIGMA}(t_j)\big\}.\end{aligned}$$

D.2 Primary Biliary Cirrhosis Data

This data set was collected from a clinical trial conducted by the Mayo Clinic between January 1974 and July 1986. In the trial, 312 patients diagnosed with PBC of the liver were randomly allocated to a treatment group with D-penicillamine or a placebo group. At the time of enrollment, certain patient demographic information, such as age and gender, was recorded. Afterward, each patient was regularly followed (approximately every 6 months) to measure some time-dependent clinical covariates until the date of death, the date of drop-out, or the end-of-study date (July 1986), whichever came first. Specifically, 15 covariates were collected in total, and detailed descriptions of these covariates are listed in Table 1.

Table 1: Description of the covariates in the example of primary biliary cirrhosis

Covariates	Details
Trt	binary, equals one if the patient is in the treatment group
Gen	gender of the patient
Age(t)	age of the patient at time t
Ascts(t)	binary, equals one if the ascites have occurred by t
Hepato(t)	binary, equals one if the hepatomegaly or enlarged liver has occurred by t
Spiders(t)	blood vessel malformations in the skin at time t
Edema(t)	0 if no edema at t; 0.5 if untreated or successfully treated at t; 1 if edema despite diuretic therapy at t
Bili(t)	serum bilirunbin (mg/dl) at t
Chol(t)	serum cholesterol (mg/dl) at t
Albumin(t)	serum albumin (mg/dl) at t
Alkphos(t)	alkaline phosphotase (U/liter) at t
Ast(t)	aspartate aminotransferase (U/ml) at t
Platelet(t)	platelet count at t
Protime(t)	standardised blood clotting time at t
Stage(t)	histologic stage of disease (needs biopsy) at t

When analyzing this data set, we model the forward intensity $\lambda_{C,t_j}(t_k)$ of the random censoring time C by

$$\begin{aligned}\lambda_{C,t_j}(t_k) = \exp\{&\beta_0(t_k) + \beta_1(t_k)\text{Trt} + \beta_2(t_k)\text{Gen} + \beta_3(t_k)\text{Age}(t_j) + \beta_4(t_k)\text{Ascts}(t_j) + \\ &\beta_5(t_k)\text{Hepato}(t_j) + \beta_6(t_k)\text{Spiders}(t_j) + \beta_7(t_k)\text{Edema}(t_j) + \beta_8(t_k)\text{Bili}(t_j) + \\ &\beta_9(t_k)\text{Chol}(t_j) + \beta_{10}(t_k)\text{Albumin}(t_j) + \beta_{11}(t_k)\text{Alkphos}(t_j) + \beta_{12}(t_k)\text{Ast}(t_j) + \\ &\beta_{13}(t_k)\text{Platelet}(t_j) + \beta_{14}(t_k)\text{Protime}(t_j) + \beta_{15}(t_k)\text{Stage}(t_j)\}.\end{aligned}$$

D.3 Data Generation in Simulation Studies

In this section, we give a detailed description about the data-generation procedure in the simulation. Our procedure for generating $(T, \boldsymbol{Z})$ mainly consists of three steps. In the first step, we generate the survival time T based on a marginal model. The second step identifies the constraint on the distribution of $\boldsymbol{Z}(t)$ that ensures the forward intensity model holds. In the third step, we generate the covariates $\boldsymbol{Z}(t)$ from the posterior distribution $\boldsymbol{Z}(t)|T$ under the constraint identified in the second step.

Step One. First, the predictive distribution of T at landmark time t is

$$P(T \geq t + t_k | T \geq t, \boldsymbol{Z}(t)) = \exp\big\{ - \Delta t \sum_{h=0}^{k-1} \exp\{\boldsymbol{\alpha}(t_h)^\top \boldsymbol{Z}(t)\}\big\}.$$

Under the assumption that $\boldsymbol{\alpha}(t_{h_1}) = \boldsymbol{\alpha}(t_{h_2})$ for any $h_1, h_2 \in \mathbb{N}$, the above predictive distribution can be written as $P(T \geq t + t_k | T \geq t, \boldsymbol{Z}(t)) = \exp\{-t_k\theta(t)\}$, where $\theta(t) = \exp\{\boldsymbol{\alpha}(0)^\top \boldsymbol{Z}(t)\}$. We assume that $\theta(t)$ follows a gamma distribution with shape parameter η_0 and rate parameter $\gamma(t)$. As a result, the marginal predictive distribution of T at landmark time t is

$$P(T \geq t + t_k | T \geq t) = \left[\frac{\gamma(t)}{\gamma(t) + t_k}\right]^{\eta_0}.$$

We fix the landmark time t as zero, and generate the survival time from the marginal distribution:

$$P(T \geq t_k) = \big[\frac{\gamma(0)}{\gamma(0) + t_k}\big]^{\eta_0}. \tag{17}$$

This completes the first step.

Step Two. In this step, we identify the constraint on the distribution of $\boldsymbol{Z}(t)$ that ensures the forward intensity model holds. Suppose the forward intensity model holds, then the following equality should hold:

$$P(T \geq t + t_k | T \geq t) = \frac{P(T \geq t + t_k | T \geq 0)}{P(T \geq t | T \geq 0)}.$$

The above constraint can be simplified as

$$\big[\frac{\gamma(t)}{\gamma(t) + t_k}\big]^{\eta_0} = \big[\frac{\gamma(0) + t}{\gamma(0 + t + t_k)}\big]^{\eta_0},$$

which implies $\gamma(t) = \gamma(0) + t$. Hence the rate parameter $\gamma(t)$ of $\theta(t) = \exp\{\boldsymbol{\alpha}(0)^\top \boldsymbol{Z}(t)\}$ should satisfy the constraint $\gamma(t) = \gamma(0) + t$. This completes the second step.

Step Three. In this step, we generate $\boldsymbol{Z}(t) = (Z_0(t), Z_1(t), \ldots, Z_k(t))$ by first generating $\theta(t) = \exp\{\boldsymbol{\alpha}(0)^\top \boldsymbol{Z}(t)\}$ from the posterior distribution $p\big(\theta(t)|T\big)$ under the constraint $\gamma(t) = \gamma(0) + t$, and then generating $Z_\ell(t)$, $\ell = 0, 1, \ldots, k$, individually. Specifically, we generate $\theta(t)$ from the following posterior distribution:

$$\begin{aligned} p(\theta(t)|T = s) &\propto p\big(T = s|\theta(t)\big)p(\theta(t)) = p\big(T = s|T \geq t, \theta(t)\big)p(\theta(t)) \\ &\propto [\exp\{-(s-t)\theta(t)\} - \exp\{-(s-t+1)\theta(t)\}]\theta(t)^{\eta_0 - 1} \exp\{-\big(\gamma(0) + t\big)\theta(t)\}. \end{aligned} \tag{18}$$

Given $\theta(t) = \exp\{\boldsymbol{\alpha}(0)^\top \boldsymbol{Z}(t)\}$, we then generate $Z_\ell(t)$, $\ell = 0, 1, \ldots, k$, individually. We first generate $Z_\ell(t)$, $\ell = 1, \ldots, k$ from certain arbitrary distributions. To ensure that $\theta(t) = \exp\{\boldsymbol{\alpha}(0)^\top \boldsymbol{Z}(t)\}$ follows the gamma distribution with shape parameter η_0 and rate parameter $\gamma(0) + t$, we introduce an adjusting covariate $Z_0(t)$, and the value of $Z_0(t)$ is determined by

$$Z_0(t) = \frac{\ln \theta(t) - \alpha_1(0)Z_1(t) - \alpha_2(0)Z_2(t) - \cdots - \alpha_k(0)Z_k(t)}{\alpha_0(0)}.$$

This completes the whole data-generation procedure. We summarize the detailed steps for generating $(T, \boldsymbol{Z})$ in Algorithm 5.

Algorithm 5: Procedure for generating $(T, \boldsymbol{Z})$

Initialization: $t \leftarrow 0$. Generate survival time T according to the marginal distribution in (17).
while $t \leq T$ **do**
  Generate $\theta(t)$ according to the posterior distribution in (18).
  Generate $Z_1(t), Z_2(t), Z_3(t)$ according to the distributions specified in Section 5.
  Solve the adjusting variable $Z_0(t)$ using

$$Z_0(t) = \frac{\ln \theta(t) - \alpha_1(0)Z_1(t) - \alpha_2(0)Z_2(t) - \alpha_3(0)Z_3(t)}{\alpha_0(0)}. \tag{19}$$

  $t \leftarrow t + 1$.
end
Return $(T, \boldsymbol{Z})$.

In our simulation study, we set $\eta_0 = 1$ and $\gamma(0) = 20$. When calculating the predictive probability $F_t(\tau) = P(T < t + \tau, T < C|\boldsymbol{Z}(t), T \wedge C \geq t)$ in the simulation section, we let $\big(Z_1(t), Z_2(t), Z_3(t)\big)$ be its expected values, fix $\theta(t)$ as $10/(\gamma(0) + t)$, and solve $Z_0(t)$ using (19).

D.4 Interval Estimation for $S_{T,t}(\tau)$

This section summarizes the simulation results for the proposed 95% confidence intervals for the predictive survival probability $S_{T,t}(\tau)$.

Figures 9 and 10 show the empirical coverage probabilities and the average 95% confidence limits based on the 1,000 simulation replications under different values of $n \in \{50, 100, 200, 500\}$, $t \in \{20, 25, 30, 35, 40\}$, and $\tau \in \{1, 2, 5, 10\}$. As can be seen, the results for the confidence intervals of $S_{T,t}(\tau)$ are generally similar to those for $F_t(\tau)$—the coverage probabilities are close to the nominal level, especially when the sample sizes are large. However, an interesting observation, which differs from the results for $F_t(\tau)$, is that the average length of the confidence intervals does not consistently increase with the forward-looking time τ when the sample size n and landmark time t are fixed. This behavior can be explained by the specific form of the asymptotic variance for $\hat{S}_{T,t}(\tau)$

$$\sigma^2 = S^2_{T,t}(\tau) \int_0^\tau \int_0^\tau \boldsymbol{Z}(t)^\top \boldsymbol{W}(u; \boldsymbol{\alpha}_0)^{-1} \boldsymbol{\varrho}(u, s, \psi) \boldsymbol{W}(s; \boldsymbol{\alpha}_0)^{-1} \boldsymbol{Z}(t) \lambda_{T,t}(u) \lambda_{T,t}(s) du ds.$$

It can be seen that the second term on the right (i.e., the double integral) increases with the forward-looking τ, while the first term $S^2_{T,t}(\tau)$ decreases with τ. Consequently, the decreasing first term may offset the increasing second term, leading to a smaller variance of $\hat{S}_{T,t}(\tau)$ for longer forward-looking horizons. This phenomenon explains why the average width of the confidence intervals slightly decreases when τ increases from 5 to 10, as shown in the bottom panel of Figure 10.

D.5 Additional Simulation on Confidence Bands

In this section, we conduct an additional simulation study to validate the performance of the proposed simultaneous confidence bands. The simulation setting follows from that in Section 5.1 of the main text. Specifically, we generate the event time T and the random censoring time C based on the log-linear forward intensity model (Duan et al., 2012), in which we include four covariates and details about the generation of the covariates are given in Appendix D.3. We examine different sample sizes $n \in \{50, 100, 200, 500\}$ and landmark times $t \in \{20, 25, 30, 35, 40\}$. For each combination of (n, t), we apply Algorithm 2 to construct 95% simultaneous confidence bands for $S_{T,t}(\tau)$ and $F_t(\tau)$ over $\tau \in \{1, 2, \ldots, 10\}$, in which the number B of resampling is chosen as 1,000. This simulation procedure is repeated 1,000 times.

Figures 11 and 12 summarize the empirical coverage probabilities and average confidence bounds of the 95% simultaneous confidence bands for $F_t(\tau)$ based on the 1,000 Monte Carlo replications, where the empirical coverage probability is calculated as the proportion of replications in which the true value of $F_t(\tau)$ falls in the confidence bands for all $\tau = 1, 2, \ldots, 10$. Similarly, Figures 13 and 14 show the respective results for $S_{T,t}(\tau)$.

As can be seen, the coverage probabilities for both F_t and $S_{T,t}$ are close to the nominal level under all values of landmark times t, especially when sample size n is large. In addition, the average width of the confidence bands gradually shrinks as sample size grows. This demonstrates satisfactory performance of the proposed confidence bands.

Appendix E. Technical Proofs

E.1 Proof of Theorem 1

We first give two useful lemmas and then prove Theorem 1.

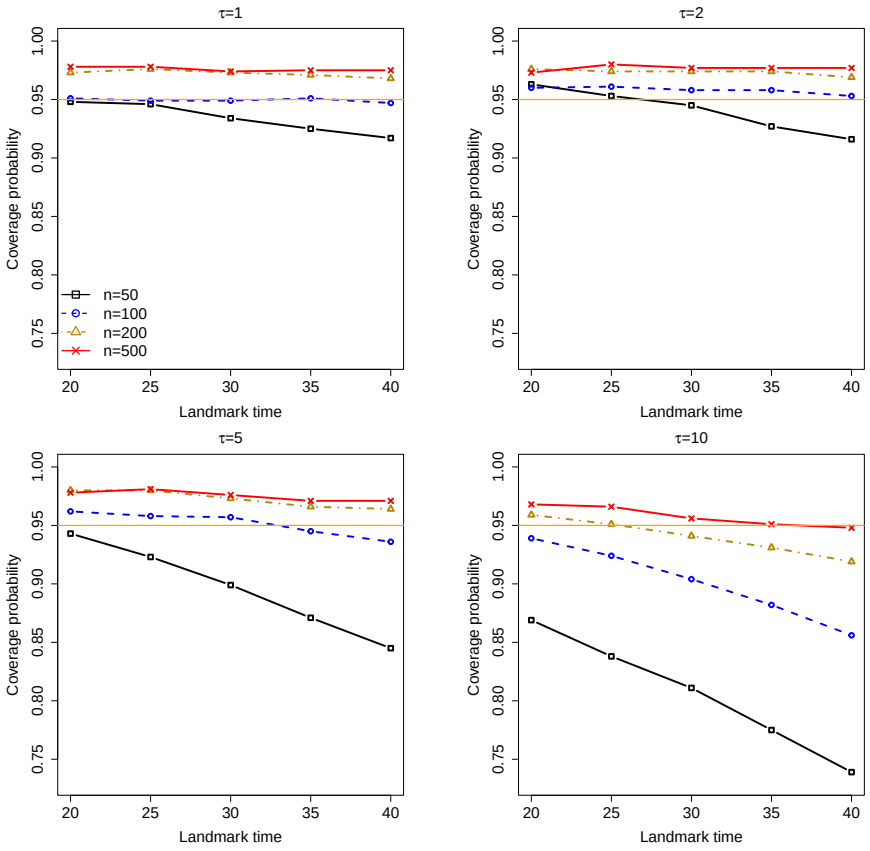

Figure 9: Empirical coverage probability of the 95% confidence interval for $S_{T,t}(\tau)$ under different values of $n \in \{50, 100, 200, 500\}$, $t \in \{20, 25, 30, 35, 40\}$, and $\tau \in \{1, 2, 5, 10\}$.

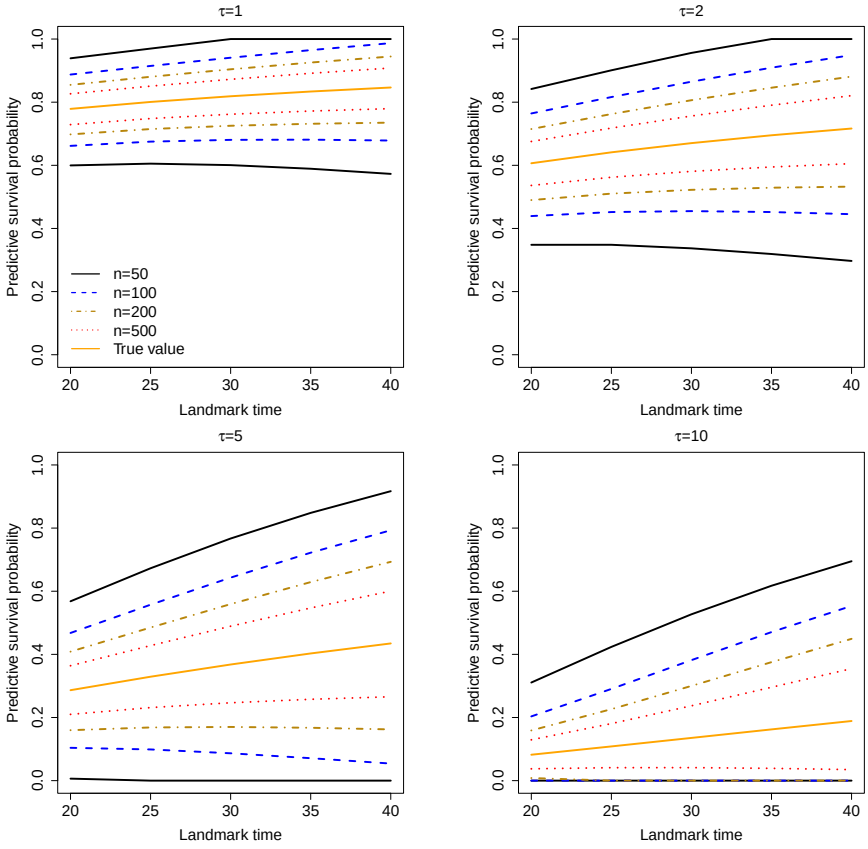

Figure 10: Average upper and lower limits of the 95% confidence interval for $S_{T,t}(\tau)$, together with the true value of $F_t(\tau)$ under different values of $n \in \{50, 100, 200, 500\}$, $t \in \{20, 25, 30, 35, 40\}$, and $\tau \in \{1, 2, 5, 10\}$.

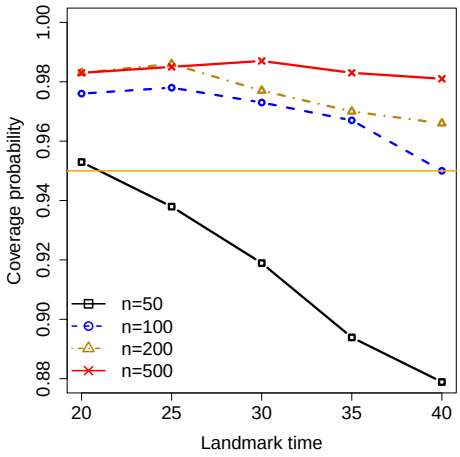

Figure 11: Empirical coverage probability of the proposed 95% simultaneous confidence bands for $F_t(\tau)$ over $\tau \in \{1, 2, \ldots, 10\}$ under different values of sample size $n \in \{50, 100, 200, 500\}$ and landmark time $t \in \{20, 25, 30, 35, 40\}$.

Lemma 9 *Define the class of functions* $\mathfrak{F} = \{\varphi(\beta, u, \mathbf{D}) : \beta \in \mathbb{R}^p, |\beta| \leq 2K_0, u \in [0, \tau]\}$ *where*

$$\varphi(\beta, u; \mathbf{D}) = \int_0^\infty I\{t > u\} Y(t) \boldsymbol{Z}(t-u) dN_T(t) - \int_0^\infty I\{t > u\} Y(t) \boldsymbol{Z}(t-u) \exp\{\beta^\top \boldsymbol{Z}(t-u)\} dt.$$

Then $\mathfrak{F}$ *is a P-Donsker class.*

Proof First, consider the class of functions $\mathfrak{F}_1 = \{\varphi_1(u; \mathbf{D}) : u \in [0, \tau]\}$, where

$$\varphi_1(u; \mathbf{D}) \triangleq \int_0^\infty I\{t > u\} Y(t) \boldsymbol{Z}(t-u) dN_T(t) = I\{X \geq u, T \leq C\} \boldsymbol{Z}(X-u).$$

Observe that $\varphi_1(u; \mathbf{D})$ is bounded by K_1. In addition, we have

$$\begin{aligned}
&E|\varphi_1(u_1; \mathbf{D}) - \varphi_1(u_2; \mathbf{D})| \\
=&E|\varphi_1(u_1; \mathbf{D}) - I\{X \geq u_2, T \leq C\} \boldsymbol{Z}(X-u_1) + I\{X \geq u_2, T \leq C\} \boldsymbol{Z}(X-u_1) - \varphi_1(u_2; \mathbf{D})| \\
\leq&E\left|\left[I\{X \geq u_1, T \leq C\} - I\{X \geq u_2, T \leq C\}\right] \boldsymbol{Z}(X-u_1)\right| \\
&\quad + E\left|I\{X \geq u_2, T \leq C\} \left[\boldsymbol{Z}(X-u_1) - \boldsymbol{Z}(X-u_2)\right]\right| \\
\leq&K_1 E\left|I\{X \geq u_1, T \leq C\} - I\{X \geq u_2, T \leq C\}\right| + E\left|\boldsymbol{Z}(X-u_1) - \boldsymbol{Z}(X-u_2)\right|. \qquad (20)
\end{aligned}$$

The first term in (20) is bounded by $K_1 \|f_T\| |u_1 - u_2|$ because

$$\begin{aligned}
&K_1 E\left|I\{X \geq u_1, T \leq C\} - I\{X \geq u_2, T \leq C\}\right| \\
\leq&K_1 E|I\{T \geq u_1\} - I\{T \geq u_2\}| \qquad (21) \\
=&K_1 P(u_1 \wedge u_2 \leq T \leq u_1 \vee u_2) \leq K_1 \|f_T\| |u_1 - u_2|.
\end{aligned}$$

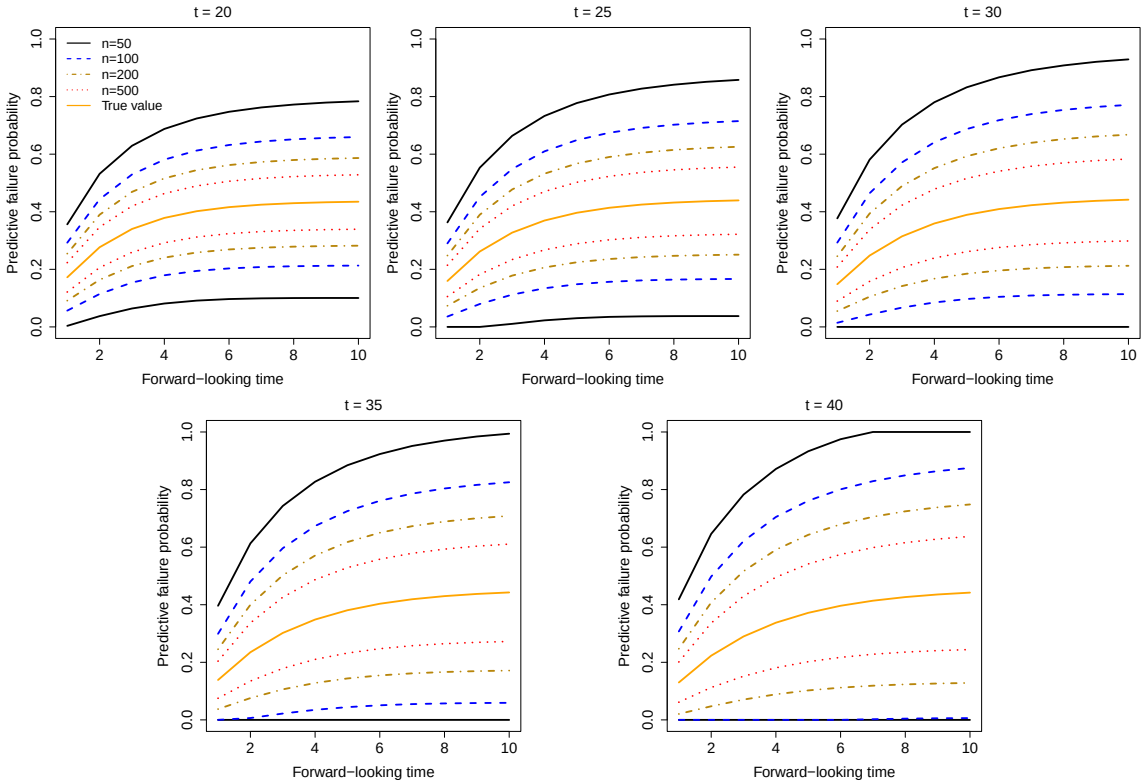

Figure 12: Average confidence bounds of the 95% simultaneous confidence bands for $F_t(\tau)$ over $\tau \in \{1, 2, \ldots, 10\}$ under different values of sample size $n \in \{50, 100, 200, 500\}$ and landmark time $t \in \{20, 25, 30, 35, 40\}$.

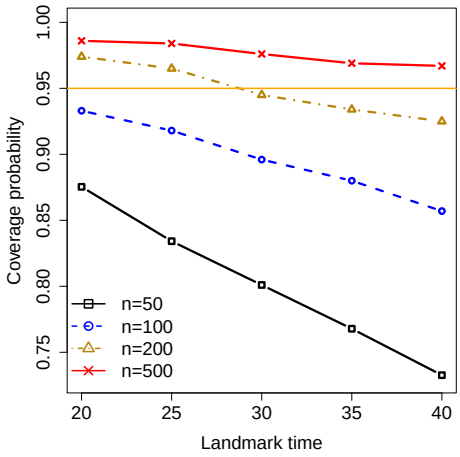

Figure 13: Empirical coverage probability of the proposed 95% simultaneous confidence bands for $S_{T,t}(\tau)$ over $\tau \in \{1, 2, \ldots, 10\}$ under different values of sample size $n \in \{50, 100, 200, 500\}$ and landmark time $t \in \{20, 25, 30, 35, 40\}$.

The second is bounded by $K_2^{1/q}|u_1 - u_2|$ because

$$(E\,|\boldsymbol{Z}(X-u_1) - \boldsymbol{Z}(X-u_2)|)^q \le E\,|\boldsymbol{Z}(X-u_1) - \boldsymbol{Z}(X-u_2)|^q \le K_2|u_1-u_2|^q, \qquad (22)$$

where the first inequality follows from Jensen's inequality and the second inequality follows from Condition (A2). Combining (20)–(22) yields

$$E|\varphi_1(u_1;\mathbf{D}) - \varphi_1(u_2;\mathbf{D})| \le (K_1\|f_T\| + K_2^{1/q})|u_1-u_2|.$$

By Example 2.11.14 in van der Vaart and Wellner (1996), $\mathfrak{F}_1$ is P-Donsker.

Then, consider the class of functions $\mathfrak{F}_2 = \{\varphi_2(\beta, u; \mathbf{D}) : \beta \in \mathbb{R}^p, |\beta| \le 2K_0, u \in [0,\tau]\}$, where

$$\varphi_2(\beta, u; \mathbf{D}) \triangleq \int_u^\infty Y(t)\boldsymbol{Z}(t-u)\exp\{\beta^\top \boldsymbol{Z}(t-u)\}dt = \int_0^{(X-u)_+} \boldsymbol{Z}(t)\exp\{\beta^\top \boldsymbol{Z}(t)\}dt.$$

Observe that

$$|\varphi_2(\beta_1,u_1;\mathbf{D}) - \varphi_2(\beta_2,u_2;\mathbf{D})| \le |\varphi_2(\beta_1,u_1;\mathbf{D}) - \varphi_2(\beta_2,u_1;\mathbf{D})| + |\varphi_2(\beta_2,u_1;\mathbf{D}) - \varphi_2(\beta_2,u_2;\mathbf{D})|\,. \qquad (23)$$

Consider the first term on the right hand side of (23):

$$\begin{aligned}
|\varphi_2(\beta_1,u_1;\mathbf{D}) - \varphi_2(\beta_2,u_1;\mathbf{D})| &= \left| \int_0^{(X-u_1)_+} \boldsymbol{Z}(t)\left[\exp\{\beta_1^\top \boldsymbol{Z}(t)\}dt - \exp\{\beta_2^\top \boldsymbol{Z}(t)\}\right] dt \right| \\
&\le p(X-u_1)_+ K_1^2 \exp\{2pK_0K_1\}|\beta_1-\beta_2| \\
&\le p\tau K_1^2 \exp\{2pK_0K_1\}|\beta_1-\beta_2|.
\end{aligned}$$

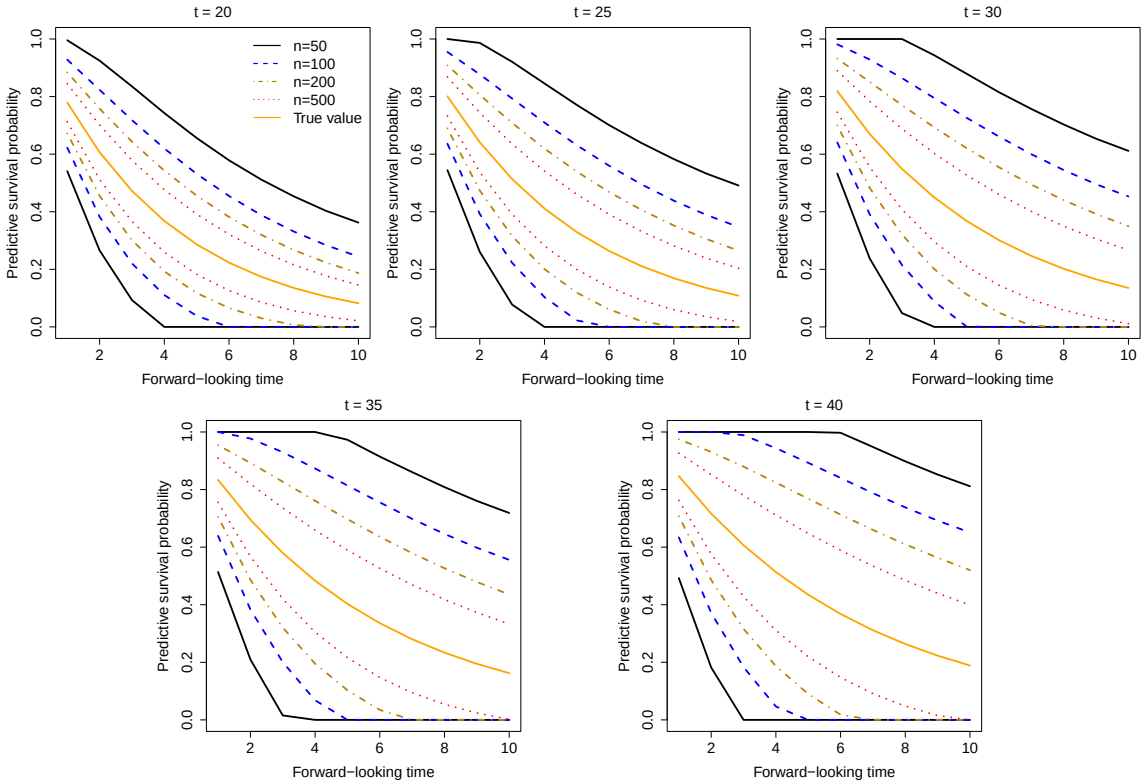

Figure 14: Average confidence bounds of the 95% simultaneous confidence bands for $S_{T,t}(\tau)$ over $\tau \in \{1, 2, \ldots, 10\}$ under different values of sample size $n \in \{50, 100, 200, 500\}$ and landmark time $t \in \{20, 25, 30, 35, 40\}$.

Consider the second term on the right hand side of (23):

$$|\varphi_2(\beta_2,u_1;\mathbf{D})-\varphi_2(\beta_2,u_2;\mathbf{D})| = \left|\int_{(X-u_1)_+}^{(X-u_2)_+} \boldsymbol{Z}(t)\exp\{\beta_2^\top \boldsymbol{Z}(t)\}dt\right| \le K_1\exp\{2pK_0K_1\}|u_1-u_2|.$$

The above shows that $(\beta,u)\mapsto\varphi_2(\beta,u;\mathbf{D})$ is Lipschitz continuous. Therefore, $\mathfrak{F}_2$ is P-Donsker by Theorem 2.7.11 in van der Vaart and Wellner (1996).

According to Example 19.20 in van der Vaart (1998), the sum of two P-Donsker classes is a P-Donsker class. Since the class of functions $\mathfrak{F}$ is the sum of two P-Donsker classes $\mathfrak{F}_1$ and $\mathfrak{F}_2$, it is a P-Donsker class. ■

Lemma 10 *Suppose Conditions (A1)–(A3) hold. Then we have* $P[\psi(\boldsymbol{\alpha}_0)](u)[\vartheta(\boldsymbol{\beta}_0)](s)^\top = \mathbf{0}$ *for any* $u,s\in[0,\tau]$.

Proof By Fubini's theorem, we have

$$\begin{aligned}
&P[\psi(\boldsymbol{\alpha}_0)](u)[\vartheta(\boldsymbol{\beta}_0)](s)^\top\\
=&E\Big[\int_u^\infty Y(t)\boldsymbol{Z}(t-u)[dN_T(t)-\exp\{\boldsymbol{\alpha}_0(u)\boldsymbol{Z}(t-u)\}dt]\times\\
&\qquad\int_s^\infty Y(t)\boldsymbol{Z}(t-s)^\top[dN_C(t)-\exp\{\boldsymbol{\beta}_0(s)\boldsymbol{Z}(t-s)\}dt]\Big] \qquad (24)\\
=&\int_u^\infty\int_s^\infty E\Big[\big\{Y(t_1)\boldsymbol{Z}(t_1-u)[dN_T(t_1)-\exp\{\boldsymbol{\alpha}_0(u)\boldsymbol{Z}(t_1-u)\}dt_1]\big\}\times\\
&\qquad\qquad\big\{Y(t_2)\boldsymbol{Z}(t_2-s)^\top[dN_C(t_2)-\exp\{\boldsymbol{\beta}_0(s)\boldsymbol{Z}(t_2-s)\}dt_2]\big\}\Big].
\end{aligned}$$

In the following, we will show the integrand in (24) is zero. First, if $t_1-u>t_2-s$, we have

$$\begin{aligned}
&E\Big[\big\{Y(t_1)\boldsymbol{Z}(t_1-u)[dN_T(t_1)-\exp\{\boldsymbol{\alpha}_0(u)\boldsymbol{Z}(t_1-u)\}dt_1]\big\}\times\\
&\quad\big\{Y(t_2)\boldsymbol{Z}(t_2-s)^\top[dN_C(t_2)-\exp\{\boldsymbol{\beta}_0(s)\boldsymbol{Z}(t_2-s)\}dt_2]\big\}\Big]\\
=&E\Big[E\Big[\big\{Y(t_1)\boldsymbol{Z}(t_1-u)[dN_T(t_1)-\exp\{\boldsymbol{\alpha}_0(u)\boldsymbol{Z}(t_1-u)\}dt_1]\big\}\times\\
&\qquad\big\{Y(t_2)\boldsymbol{Z}(t_2-s)^\top[dN_C(t_2)-\exp\{\boldsymbol{\beta}_0(s)\boldsymbol{Z}(t_2-s)\}dt_2]\big\}|\mathcal{F}_{t_1-u}\Big]\Big]\\
=&E\Big[E\Big[\big\{Y(t_1)\boldsymbol{Z}(t_1-u)[dN_T(t_1)-\exp\{\boldsymbol{\alpha}_0(u)\boldsymbol{Z}(t_1-u)\}dt_1]\big\}|\mathcal{F}_{t_1-u}\Big]\times\\
&\quad E\Big[\big\{Y(t_2)\boldsymbol{Z}(t_2-s)^\top[dN_C(t_2)-\exp\{\boldsymbol{\beta}_0(s)\boldsymbol{Z}(t_2-s)\}dt_2]\big\}|\mathcal{F}_{t_1-u}\Big]\Big]\\
=&E\Big[\mathbf{0}\times E\Big[\big\{Y(t_2)\boldsymbol{Z}(t_2-s)^\top[dN_C(t_2)-\exp\{\boldsymbol{\beta}_0(s)\boldsymbol{Z}(t_2-s)\}dt_2]\big\}|\mathcal{F}_{t_1-u}\Big]\Big]=\mathbf{0},
\end{aligned}$$

where the first equality follows from the law of iterated expectation; the second equality follows from the conditional independence between T and C; the third equality follows from

the fact that $E[dN_T(t_1)|\mathcal{F}_{t_1-u}] = P(T \geq t_1|\mathcal{F}_{t_1-u})\exp\{\boldsymbol{\alpha}_0(u)\boldsymbol{Z}(t_1-u)\}dt_1$. Similarly, if $t_1 - u \leq t_2 - s$, we have

$$\begin{aligned}
&E\Big[\big\{Y(t_1)\boldsymbol{Z}(t_1-u)[dN_T(t_1) - \exp\{\boldsymbol{\alpha}_0(u)\boldsymbol{Z}(t_1-u)\}dt_1]\big\}\times \\
&\quad \big\{Y(t_2)\boldsymbol{Z}(t_2-s)^\top[dN_C(t_2) - \exp\{\boldsymbol{\beta}_0(s)\boldsymbol{Z}(t_2-s)\}dt_2]\big\}\Big] \\
=&E\Big[E\Big[\big\{Y(t_1)\boldsymbol{Z}(t_1-u)[dN_T(t_1) - \exp\{\boldsymbol{\alpha}_0(u)\boldsymbol{Z}(t_1-u)\}dt_1]\big\}\times \\
&\quad \big\{Y(t_2)\boldsymbol{Z}(t_2-s)^\top[dN_C(t_2) - \exp\{\boldsymbol{\beta}_0(s)\boldsymbol{Z}(t_2-s)\}dt_2]\big\}|\mathcal{F}_{t_2-s}\Big]\Big] \\
=&E\Big[E\Big[\big\{Y(t_1)\boldsymbol{Z}(t_1-u)[dN_T(t_1) - \exp\{\boldsymbol{\alpha}_0(u)\boldsymbol{Z}(t_1-u)\}dt_1]\big\}|\mathcal{F}_{t_2-s}\Big]\times \\
&\quad E\Big[\big\{Y(t_2)\boldsymbol{Z}(t_2-s)^\top[dN_C(t_2) - \exp\{\boldsymbol{\beta}_0(s)\boldsymbol{Z}(t_2-s)\}dt_2]\big\}|\mathcal{F}_{t_2-s}\Big]\Big] \\
=&E\Big[E\Big[\big\{Y(t_1)\boldsymbol{Z}(t_1-u)[dN_T(t_1) - \exp\{\boldsymbol{\alpha}_0(u)\boldsymbol{Z}(t_1-u)\}dt_1]\big\}|\mathcal{F}_{t_2-s}\Big]\times \mathbf{0}\Big] = \mathbf{0}.
\end{aligned}$$

Therefore, the integrand in (24) is always zero, and thus $P[\psi(\boldsymbol{\alpha}_0)](u)[\vartheta(\boldsymbol{\beta}_0)](s)^\top = \mathbf{0}$. ■

Proof of Theorem 1: Note that $\|\mathbb{P}_n\psi(\hat{\boldsymbol{\alpha}}_n)\| = 0$ because of the definition of $\hat{\boldsymbol{\alpha}}_n$. Furthermore,

$$\begin{aligned}
\sup_{\boldsymbol{\alpha}\in\mathcal{F}}\|\mathbb{P}_n\psi(\boldsymbol{\alpha}) - P\psi(\boldsymbol{\alpha})\| &= \sup_{\boldsymbol{\alpha}\in\mathcal{F}}\sup_u|\mathbb{P}_n\varphi(\boldsymbol{\alpha}(u),u) - P\varphi(\boldsymbol{\alpha}(u),u)| \\
&\leq \sup_{\varphi\in\mathfrak{F}}|\mathbb{P}_n\varphi - P\varphi| \overset{a.s.}{\to} 0.
\end{aligned}$$

The inequality holds because $\mathfrak{F}$ is a larger class. The almost sure convergence holds because $\mathfrak{F}$ is P-Donsker, and thus P-Glivenko-Cantelli. If we can show that

$$\inf_{\boldsymbol{\alpha}\in\mathcal{F},\|\boldsymbol{\alpha}-\boldsymbol{\alpha}_0\|>\epsilon}\|P\psi(\boldsymbol{\alpha})\| > 0 \tag{25}$$

for any $\epsilon > 0$, then we have

$$\|\mathbb{P}_n\psi(\boldsymbol{\alpha})\| \geq \|P\psi(\boldsymbol{\alpha})\| - \|(\mathbb{P}_n - P)\psi(\boldsymbol{\alpha})\| > 0$$

when $\|\boldsymbol{\alpha} - \boldsymbol{\alpha}_0\| > \epsilon$ and n is large enough. This means $\|\hat{\boldsymbol{\alpha}}_n - \boldsymbol{\alpha}_0\| < \epsilon$ for sufficiently large n. Hence it suffices to verify (25). For $\boldsymbol{\alpha} \in \mathcal{F}$, if $\|\boldsymbol{\alpha} - \boldsymbol{\alpha}_0\| > 2\epsilon$, then there exists $u \in [0, \tau]$

such that $|\boldsymbol{\alpha}(u) - \boldsymbol{\alpha}_0(u)| > \epsilon$. Let $\delta = \boldsymbol{\alpha}(u) - \boldsymbol{\alpha}_0(u)$. Since $P\varphi(\boldsymbol{\alpha}_0, u) = 0$, we have

$$\begin{aligned}
|P\varphi(\boldsymbol{\alpha}, u)| &= \left| P \int_u^\infty Y(t)\boldsymbol{Z}(t-u) \left[\exp\{\boldsymbol{\alpha}(u)^\top \boldsymbol{Z}(t-u)\} - \exp\{\boldsymbol{\alpha}_0(u)^\top \boldsymbol{Z}(t-u)\right] dt \right| \\
&= \left| P \int_u^\infty \int_0^1 Y(t)\boldsymbol{Z}(t-u)\boldsymbol{Z}(t-u)^\top \exp\left\{ \left(\boldsymbol{\alpha}_0(u) + v\delta\right)^\top \boldsymbol{Z}(t-u) \right\} \delta dv dt \right| \\
&\geq \exp(-pK_0K_1) \times \left| P \int_u^\infty Y(t)\boldsymbol{Z}(t-u)\boldsymbol{Z}(t-u)^\top dt \cdot \delta \right| \\
&\geq \exp(-pK_0K_1) \times \left| P \int_0^{(X-u)_+} \boldsymbol{Z}(t)\boldsymbol{Z}(t)^\top dt \cdot \delta \right| \\
&\geq \exp(-pK_0K_1) \times \left| P \int_0^{(X-\tau)_+} \boldsymbol{Z}(t)\boldsymbol{Z}(t)^\top dt \cdot \delta \right| \\
&\geq \exp(-pK_0K_1)\lambda_{\min}|\delta| \\
&\geq \exp(-pK_0K_1)\lambda_{\min}\epsilon
\end{aligned}$$

where $\lambda_{\min}$ is the smallest eigenvalue of the positive definite matrix $P\int_0^{(X-\tau)_+} \boldsymbol{Z}(t)\boldsymbol{Z}(t)^\top dt$. The second equality is by the mean value theorem for multivariate functions, and the third inequality is based on the fact that $|(\mathbf{A}+\mathbf{B})\delta| \geq |\mathbf{A}\delta|$ when $\mathbf{A} \in \mathbb{R}^{p\times p}$ is positive definite, and $\mathbf{B} \in \mathbb{R}^{p\times p}$ is nonnegative definite. This establishes (25) and completes the proof of the strong consistency of $\hat{\boldsymbol{\alpha}}_n$. Similarly, the strong consistency of $\hat{\boldsymbol{\beta}}_n$ can also be established.

Next, we show the weak convergence of the estimator. The estimator $\hat{\boldsymbol{\alpha}}_n$ can be regarded as an Z-estimator with infinitely many estimating equations indexed by $u \in [0,\tau]$. Since $\hat{\boldsymbol{\alpha}}_n$ is consistent, we shall consider the class of functions $\{[\psi(\boldsymbol{\alpha};\mathbf{D})](u) : \|\boldsymbol{\alpha} - \boldsymbol{\alpha}_0\| < \epsilon, u \in [0,\tau]\}$ for some small $\epsilon > 0$. This class of functions is P-Donsker because it is a subset of $\mathfrak{F}$ defined above. By a continuous mapping argument, the asymptotic continuity condition $\mathbb{G}_n\psi(\hat{\boldsymbol{\alpha}}_n;\mathbf{D}) - \mathbb{G}_n\psi(\boldsymbol{\alpha}_0;\mathbf{D}) = o_P(1)$ will hold if $\left\| P\left|[\psi(\boldsymbol{\alpha})](u) - [\psi(\boldsymbol{\alpha}_0)](u)\right|^2 \right\| \to 0$ as $\boldsymbol{\alpha} \to \boldsymbol{\alpha}_0$. To show this, let $\delta = \boldsymbol{\alpha}(u) - \boldsymbol{\alpha}_0(u)$. Then

$$\begin{aligned}
&P\left|[\psi(\boldsymbol{\alpha})](u) - [\psi(\boldsymbol{\alpha}_0)](u)\right|^2 \\
&= P\left| \int_0^{(X-u)_+} \boldsymbol{Z}(t) \left[\exp\{\boldsymbol{\alpha}(u)^\top \boldsymbol{Z}(t)\} - \exp\{\boldsymbol{\alpha}_0(u)^\top \boldsymbol{Z}(t)\} \right] dt \right|^2 \\
&= P\left| \int_0^{(X-u)_+} \int_0^1 \boldsymbol{Z}(t)\boldsymbol{Z}(t)^\top \exp\left\{ \left(\boldsymbol{\alpha}_0(u) + v\delta\right)^\top \boldsymbol{Z}(t) \right\} \delta dv dt \right|^2 \\
&\leq P\left((X-u)_+ pK_1^2 \exp\{pK_0K_1\} \cdot |\delta| \right)^2 \leq P\left(XpK_1^2 \exp\{pK_0K_1\} \cdot |\delta| \right)^2
\end{aligned}$$

When $\boldsymbol{\alpha} \to \boldsymbol{\alpha}_0$, the above converges to zero, uniformly in u. The asymptotic continuity condition implies that

$$\mathbb{G}_n\psi(\boldsymbol{\alpha}_0) = \sqrt{n}\left[P\psi(\hat{\boldsymbol{\alpha}}_n) - P\psi(\boldsymbol{\alpha}_0)\right] + o_P(1). \tag{26}$$

Next, we show that $P\psi(\boldsymbol{\alpha})$ is Frechet differentiable and the derivative has a continuous inverse. Note that

$$\begin{aligned} P[\psi(\boldsymbol{\alpha})](u) &= P\int_0^\infty I\{t>u\}\boldsymbol{Z}(t-u)dN_T(t) - P\int_0^\infty I\{t>u\}Y(t)\boldsymbol{Z}(t-u)\exp\{\boldsymbol{\alpha}(u)^\top\boldsymbol{Z}(t-u)\}dt \\ &= P\int_0^\infty I\{t>u\}Y(t)\boldsymbol{Z}(t-u)\left[\exp\{\boldsymbol{\alpha}_0(u)^\top\boldsymbol{Z}(t-u)\} - \exp\{\boldsymbol{\alpha}(u)^\top\boldsymbol{Z}(t-u)\}\right]dt. \end{aligned}$$

Consider a vector-valued function $h:[0,\tau]\mapsto\mathbb{R}^p$ such that $\boldsymbol{\alpha}_0+h\in\mathcal{F}$. By Taylor's expansion, $P[\psi(\boldsymbol{\alpha}_0+h)](u)-P[\psi(\boldsymbol{\alpha}_0)](u)$ can be written as

$$\begin{aligned} &-P\int_0^\infty I\{t>u\}Y(t)\boldsymbol{Z}(t-u)\boldsymbol{Z}(t-u)^\top\exp\{\boldsymbol{\alpha}_0(u)^\top\boldsymbol{Z}(t-u)\}dth(u) \\ &-h^\top(u)\int_0^1\int_0^1 P\int_u^\infty Y(t)\boldsymbol{Z}(t-u)\boldsymbol{Z}(t-u)\boldsymbol{Z}(t-u)^\top\exp\{[\boldsymbol{\alpha}_0(u)+rsh(u)]^\top\boldsymbol{Z}(t-u)\}dtdrdsh(u). \end{aligned}$$

By the boundness of $\boldsymbol{Z}$, the second line is $O(\|h\|^2)$ as $h\to 0$, and thus is $o(\|h\|)$. Therefore, $P\psi(\boldsymbol{\alpha})$ is Frechet differentiable at $\boldsymbol{\alpha}=\boldsymbol{\alpha}_0$ with derivative

$$[V(h)](u) = -P\int_0^\infty I\{t>u\}Y(t)\boldsymbol{Z}(t-u)\boldsymbol{Z}(t-u)^\top\exp\{\boldsymbol{\alpha}_0(u)^\top\boldsymbol{Z}(t-u)\}dth(u).$$

It is readily seen that the inverse of $V(h)$ is given by $[V^{-1}(g)](u) = -(\boldsymbol{W}(u;\boldsymbol{\alpha}_0))^{-1}g(u)$, where

$$\boldsymbol{W}(u;\boldsymbol{\alpha}_0) = P\int_0^\infty I\{t>u\}Y(t)\boldsymbol{Z}(t-u)\boldsymbol{Z}(t-u)^\top\exp\{\boldsymbol{\alpha}_0(u)^\top\boldsymbol{Z}(t-u)\}dt.$$

Checking the continuity of the inverse amounts to verifying that there exists a positive number ϵ such that $\sup_u|[V(h)](u)|\geq\epsilon\|h\|$. This is obvious by letting ϵ be the infimum of the minimum eigenvalues of the following collection of positive definite matrices

$$\left\{P\int_0^\infty I\{t>u\}Y(t)\boldsymbol{Z}(t-u)\boldsymbol{Z}(t-u)^\top\exp\{\boldsymbol{\alpha}_0(u)^\top\boldsymbol{Z}(t-u)\}dt : u\in[0,\tau]\right\},$$

which is equal to

$$\left\{P\int_0^{(X-u)_+}\boldsymbol{Z}(t)\boldsymbol{Z}(t)^\top\exp\{\boldsymbol{\alpha}_0(u)^\top\boldsymbol{Z}(t)\}dt : u\in[0,\tau]\right\}$$

The ϵ chosen above is positive because it is larger than $\exp\{-pK_0K_1\}\lambda_{\min}$. Therefore, (26) can be expanded as

$$\mathbb{G}_n\psi(\boldsymbol{\alpha}_0) = \sqrt{n}V(\hat{\boldsymbol{\alpha}}_n-\boldsymbol{\alpha}_0) + o_P\left(\sqrt{n}(\hat{\boldsymbol{\alpha}}_n-\boldsymbol{\alpha}_0)\right) + o_P(1).$$

By the continuous invertibility of V, we have the second term on the right hand side is $o_P(1)$, and thus $\sqrt{n}(\hat{\boldsymbol{\alpha}}_n-\boldsymbol{\alpha}_0) = V^{-1}\left(\mathbb{G}_n\psi(\boldsymbol{\alpha}_0)\right)+o_P(1)$. Since $\mathbb{G}_n\psi(\boldsymbol{\alpha}_0)$ converges weakly to a Gaussian process with covariance $\boldsymbol{\varrho}(s,u,\psi) = P[\psi(\boldsymbol{\alpha}_0)](s)[\psi(\boldsymbol{\alpha}_0)](u)^\top$, a continuous

mapping argument yields that $\sqrt{n}(\hat{\boldsymbol{\alpha}}_n - \boldsymbol{\alpha}_0)$ converges weakly to a Gaussian process with covariance function given by

$$\mathrm{Cov}\big(\hat{\boldsymbol{\alpha}}_n(s), \hat{\boldsymbol{\alpha}}_n(u)\big) = \left(\boldsymbol{W}(s;\boldsymbol{\alpha}_0)\right)^{-1} \boldsymbol{\varrho}(s,u,\psi) \left(\boldsymbol{W}(u;\boldsymbol{\alpha}_0)\right)^{-1}.$$

This completes the proof for the weak convergence of $\hat{\boldsymbol{\alpha}}_n$. The weak convergence of $(\hat{\boldsymbol{\alpha}}_n, \hat{\boldsymbol{\beta}}_n)$ can be similarly established by following the same arguments as above, except that we need to replace $[\psi(\boldsymbol{\alpha})](u)$ with

$$[\Psi(\boldsymbol{\alpha},\boldsymbol{\beta})](u) = \begin{pmatrix} [\psi(\boldsymbol{\alpha})](u) \\ [\vartheta(\boldsymbol{\beta})](u) \end{pmatrix},$$

which is the derivative of $\big(\ell(\boldsymbol{\alpha}(u);\mathbf{D}) + \ell(\boldsymbol{\beta}(u);\mathbf{D})\big)$ w.r.t. $\big(\boldsymbol{\alpha}(u), \boldsymbol{\beta}(u)\big)$. To avoid redundancy, we omit the details and directly give the asymptotic distribution of $(\hat{\boldsymbol{\alpha}}_n, \hat{\boldsymbol{\beta}}_n)$—$\sqrt{n}\left(\begin{bmatrix}\hat{\boldsymbol{\alpha}}_n \\ \hat{\boldsymbol{\beta}}_n\end{bmatrix} - \begin{bmatrix}\boldsymbol{\alpha}_0 \\ \boldsymbol{\beta}_0\end{bmatrix}\right)$ weakly converges to a Gaussian process with mean zero and covariance function

$$\boldsymbol{\rho}(s,u) = \big(\check{\boldsymbol{W}}(s;\boldsymbol{\alpha}_0,\boldsymbol{\beta}_0)\big)^{-1} \check{\boldsymbol{\varrho}}(s,u,\Psi) \big(\check{\boldsymbol{W}}(u;\boldsymbol{\alpha}_0,\boldsymbol{\beta}_0)\big)^{-1},$$

where

$$\check{\boldsymbol{W}}(u;\boldsymbol{\alpha}_0,\boldsymbol{\beta}_0) = P\nabla\Psi(\boldsymbol{\alpha}_0,\boldsymbol{\beta}_0)[u] = \begin{bmatrix} \boldsymbol{W}(u;\boldsymbol{\alpha}_0) & \mathbf{0} \\ \mathbf{0} & \boldsymbol{W}(u;\boldsymbol{\beta}_0) \end{bmatrix},$$

$$\check{\boldsymbol{\varrho}}(s,u,\Psi) = P[\Psi(\boldsymbol{\alpha}_0,\boldsymbol{\beta}_0)](s)[\Psi(\boldsymbol{\alpha}_0,\boldsymbol{\beta}_0)](u)^\top = \begin{bmatrix} P[\psi(\boldsymbol{\alpha}_0)](s)[\psi(\boldsymbol{\alpha}_0)](u)^\top & P[\psi(\boldsymbol{\alpha}_0)](s)[\vartheta(\boldsymbol{\beta}_0)](u)^\top \\ P[\vartheta(\boldsymbol{\beta}_0)](s)[\psi(\boldsymbol{\alpha}_0)](u)^\top & P[\vartheta(\boldsymbol{\beta}_0)](s)[\vartheta(\boldsymbol{\beta}_0)](u)^\top \end{bmatrix}.$$

By Lemma 10, we have $P[\psi(\boldsymbol{\alpha}_0)](s)[\vartheta(\boldsymbol{\beta}_0)](u)^\top = P[\vartheta(\boldsymbol{\beta}_0)](s)[\psi(\boldsymbol{\alpha}_0)](u)^\top = \mathbf{0}$. Thus the covariance function $\boldsymbol{\rho}(s,u)$ can be expressed as

$$\boldsymbol{\rho}(s,u) = \begin{bmatrix} \boldsymbol{W}(s;\boldsymbol{\alpha}_0)^{-1}\boldsymbol{\varrho}(s,u,\psi)\boldsymbol{W}(u;\boldsymbol{\alpha}_0)^{-1} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{W}(s;\boldsymbol{\beta}_0)^{-1}\boldsymbol{\varrho}(s,u,\vartheta)\boldsymbol{W}(u;\boldsymbol{\beta}_0)^{-1} \end{bmatrix}.$$

This completes the proof.

E.2 Proof of Proposition 2

In this proof, we first show $\hat{\boldsymbol{\varrho}}_n(u,s;\hat{\boldsymbol{\alpha}}_n) \xrightarrow{p} \boldsymbol{\varrho}(u,s,\psi)$ using the Glivenko-Cantelli theorem. Define

$$\begin{aligned}
G(\boldsymbol{Z}_i, X_i; \boldsymbol{\alpha}, u, s) \triangleq & \int_{u\vee s}^{X_i} \boldsymbol{Z}_i(t-u)\boldsymbol{Z}_i(t-s)^\top \exp\{\boldsymbol{\alpha}(0)^\top \boldsymbol{Z}_i(t)\} dt \\
& - \int_s^{X_i}\int_u^{t_2} \boldsymbol{Z}_i(t_1-u)\exp\{\boldsymbol{\alpha}(u)^\top \boldsymbol{Z}_i(t_1)\}\boldsymbol{Z}_i(t_2-s)^\top \exp\{\boldsymbol{\alpha}(0)^\top \boldsymbol{Z}_i(t_2)\} dt_1 dt_2 \\
& - \int_u^{X_i}\int_s^{t_1} \boldsymbol{Z}_i(t_1-u)\boldsymbol{Z}_i(t_2-s)^\top \exp\{\boldsymbol{\alpha}(s)^\top \boldsymbol{Z}_i(t_2)\}\exp\{\boldsymbol{\alpha}(0)^\top \boldsymbol{Z}_i(t_1)\} dt_2 dt_1 \qquad (27) \\
& + \int_s^{X_i}\int_u^{X_i} \boldsymbol{Z}_i(t_1-u)\exp\{\boldsymbol{\alpha}(u)^\top \boldsymbol{Z}_i(t_1)\}\boldsymbol{Z}_i(t_2-s)^\top \exp\{\boldsymbol{\alpha}(s)^\top \boldsymbol{Z}_i(t_2)\} dt_1 dt_2.
\end{aligned}$$

Based on the above definition of $G(\boldsymbol{Z}_i, X_i; \boldsymbol{\alpha}, u, s)$ and the expression of $\hat{\boldsymbol{\varrho}}_n(u, s; \boldsymbol{\alpha})$ in (10), we have

$$\hat{\boldsymbol{\varrho}}_n(u, s; \boldsymbol{\alpha}) = \frac{1}{n} \sum_{i=1}^{n} G(\boldsymbol{Z}_i, X_i; \boldsymbol{\alpha}, u, s) \tag{28}$$

Consider the class of functions $\mathfrak{F}_3 = \{G_{j_1,j_2}(\boldsymbol{Z}_i, X_i; \boldsymbol{\alpha}, u, s) : \boldsymbol{\alpha} \in \mathcal{F}, u, s \in [0, \tau]\}$, where $G_{j_1,j_2}(\boldsymbol{Z}_i, X_i; \boldsymbol{\alpha}, u, s)$ is the (j_1, j_2)th entry of $G(\boldsymbol{Z}_i, X_i; \boldsymbol{\alpha}, u, s)$, $1 \leq j_1, j_2 \leq p$. We then show this class $\mathfrak{F}_3$ of functions is Glivenko-Cantelli according to Example 19.8 in van der Vaart (1998). First, Condition (A1) guarantees the compactness of the parameter space $\mathcal{F}$ of $\boldsymbol{\alpha}$. Second, according to the boundness of $\mathcal{F}$ and $\boldsymbol{Z}(t)$ in Conditions (A1) and (A2), we have $|G_{j_1,j_2}(\boldsymbol{Z}_i, X_i; \boldsymbol{\alpha}, u, s)| \leq K_1^2 \exp\{2pK_0K_1\}X^2$. Based on the assumption that $E[X^2] < \infty$, we know that $|G_{j_1,j_2}(\boldsymbol{Z}_i, X_i; \boldsymbol{\alpha}, u, s)|$ is bounded above by an integrable function. Third, it is straightforward to check that $\boldsymbol{\alpha} \mapsto G_{j_1,j_2}(\boldsymbol{Z}_i, X_i; \boldsymbol{\alpha}, u, s)$ is continuous for any given observed data $\boldsymbol{Z}_i$ and X_i using the dominated convergence theorem. Therefore, according to Example 19.8 in van der Vaart (1998), $\mathfrak{F}_3$ is Glivenko-Cantelli and we have

$$\sup_{\boldsymbol{\alpha} \in \mathcal{F}} \left\| \frac{1}{n} \sum_{i=1}^{n} G_{j_1,j_2}(\boldsymbol{Z}_i, X_i; \boldsymbol{\alpha}, u, s) - E[G_{j_1,j_2}(\boldsymbol{Z}_i, X_i; \boldsymbol{\alpha}, u, s)] \right\| \xrightarrow{a.s.} 0. \tag{29}$$

Therefore, for any $\epsilon > 0$, we have

$$\begin{aligned}
&P\Big(\Big\| \frac{1}{n} \sum_{i=1}^{n} G_{j_1,j_2}(\hat{\boldsymbol{\alpha}}_n; \boldsymbol{Z}_i, X_i, u, s) - E[G_{j_1,j_2}(\boldsymbol{Z}_i, X_i; \boldsymbol{\alpha}_0, u, s)] \Big\| > \epsilon \Big) \\
\leq & P\Big(\Big\| \frac{1}{n} \sum_{i=1}^{n} G_{j_1,j_2}(\hat{\boldsymbol{\alpha}}_n; \boldsymbol{Z}_i, X_i, u, s) - E[G_{j_1,j_2}(\boldsymbol{Z}_i, X_i; \hat{\boldsymbol{\alpha}}_n, u, s)] \Big\| > \frac{\epsilon}{2} \Big) \\
&+ P\Big(\big\| E[G_{j_1,j_2}(\boldsymbol{Z}_i, X_i; \hat{\boldsymbol{\alpha}}_n, u, s)] - E[G_{j_1,j_2}(\boldsymbol{Z}_i, X_i; \boldsymbol{\alpha}_0, u, s)] \big\| > \frac{\epsilon}{2} \Big) \\
\leq & P\Big(\sup_{\boldsymbol{\alpha} \in \mathcal{F}} \Big\| \frac{1}{n} \sum_{i=1}^{n} G_{j_1,j_2}(\boldsymbol{Z}_i, X_i; \boldsymbol{\alpha}, u, s) - E[G_{j_1,j_2}(\boldsymbol{Z}_i, X_i; \boldsymbol{\alpha}, u, s)] \Big\| > \frac{\epsilon}{2} \Big) \\
&+ P\Big(\big\| E[G_{j_1,j_2}(\boldsymbol{Z}_i, X_i; \hat{\boldsymbol{\alpha}}_n, u, s)] - E[G_{j_1,j_2}(\boldsymbol{Z}_i, X_i; \boldsymbol{\alpha}_0, u, s)] \big\| > \frac{\epsilon}{2} \Big).
\end{aligned} \tag{30}$$

The term in the second last row of (30) converges to zero because of (29). Recall $\hat{\boldsymbol{\alpha}}_n \xrightarrow{p} \boldsymbol{\alpha}_0$, and it is straightforward to check that $\boldsymbol{\alpha} \mapsto E[G_{j_1,j_2}(\boldsymbol{Z}_i, X_i; \boldsymbol{\alpha}, u, s)]$ is a continuous function using the dominated convergence theorem. By the continuous mapping theorem, the term in the last row of (30) also converges to zero. Therefore, we have

$$\frac{1}{n} \sum_{i=1}^{n} G_{j_1,j_2}(\hat{\boldsymbol{\alpha}}_n; \boldsymbol{Z}_i, X_i, u, s) \xrightarrow{p} E[G_{j_1,j_2}(\boldsymbol{Z}_i, X_i; \boldsymbol{\alpha}_0, u, s)], \quad 1 \leq j_1, j_2 \leq p. \tag{31}$$

Combining (28) and (31), we know that the (j_1, j_2)th entry of $\hat{\boldsymbol{\varrho}}_n(u, s; \boldsymbol{\alpha})$ converges to $E[G_{j_1,j_2}(\boldsymbol{Z}_i, X_i; \boldsymbol{\alpha}_0, u, s)]$ in probability, $1 \leq j_1, j_2 \leq p$. This implies $\hat{\boldsymbol{\varrho}}_n(u, s; \boldsymbol{\alpha}) \xrightarrow{p} E[G(\boldsymbol{Z}_i, X_i; \boldsymbol{\alpha}_0, u, s)]$. To achieve our goal $\hat{\boldsymbol{\varrho}}_n(u, s; \hat{\boldsymbol{\alpha}}_n) \xrightarrow{p} \boldsymbol{\varrho}(u, s, \psi)$, it suffices to show

$\boldsymbol{\varrho}(u,s,\psi) = E[G(\boldsymbol{Z}_i, X_i; \boldsymbol{\alpha}_0, u, s)]$. To this end, we expand $\boldsymbol{\varrho}(u,s,\psi)$ as follows.

$$
\begin{aligned}
&\boldsymbol{\varrho}(u,s,\psi) = P[\psi(\boldsymbol{\alpha}_0;\mathbf{D})](u)[\psi(\boldsymbol{\alpha}_0;\mathbf{D})](s)^\top \\
=&E\Bigg[\left(\int_0^\infty I\{t_1>u\}Y(t_1)\boldsymbol{Z}(t_1-u)dN_T(t_1) - \int_0^\infty I\{t_1>u\}Y(t_1)\boldsymbol{Z}(t_1-u)\lambda_{T,t_1-u}(u)dt_1\right) \\
&\times\left(\int_0^\infty I\{t_2>s\}Y(t_2)\boldsymbol{Z}(t_2-s)^\top dN_T(t_2) - \int_0^\infty I\{t_2>s\}Y(t_2)\boldsymbol{Z}(t_2-s)^\top\lambda_{T,t_2-s}(s)dt_2\right)\Bigg] \\
=&E\int_0^\infty\int_0^\infty I\{t_1>u\}Y(t_1)\boldsymbol{Z}(t_1-u)I\{t_2>s\}Y(t_2)\boldsymbol{Z}(t_2-s)^\top dN_T(t_1)dN_T(t_2) \qquad (32)\\
&- E\int_0^\infty\int_0^\infty I\{t_1>u\}Y(t_1)\boldsymbol{Z}(t_1-u)I\{t_2>s\}Y(t_2)\boldsymbol{Z}(t_2-s)^\top\lambda_{T,t_2-s}(s)dN_T(t_1)dt_2 \qquad (33)\\
&- E\int_0^\infty\int_0^\infty I\{t_1>u\}Y(t_1)\boldsymbol{Z}(t_1-u)\lambda_{T,t_1-u}(u)I\{t_2>s\}Y(t_2)\boldsymbol{Z}(t_2-s)^\top dt_1dN_T(t_2) \qquad (34)\\
&+ E\int_0^\infty\int_0^\infty I\{t_1>u\}Y(t_1)\boldsymbol{Z}(t_1-u)\lambda_{T,t_1-u}(u)I\{t_2>s\}Y(t_2)\boldsymbol{Z}(t_2-s)^\top\lambda_{T,t_2-s}(s)dt_1dt_2. \qquad (35)
\end{aligned}
$$

Then, we derive the above four expectations in (32)–(35) one by one. The expectation in (32) is

$$
\begin{aligned}
&E\int_0^\infty\int_0^\infty I\{t_1>u\}Y(t_1)\boldsymbol{Z}(t_1-u)I\{t_2>s\}Y(t_2)\boldsymbol{Z}(t_2-s)^\top dN_T(t_1)dN_T(t_2)\\
=&E\int_0^\infty I\{t>u\}I\{t>s\}\boldsymbol{Z}(t-u)\boldsymbol{Z}(t-s)^\top Y(t)dN_T(t)\\
=&E\left[E\left[\int_0^\infty I\{t>u\}I\{t>s\}\boldsymbol{Z}(t-u)\boldsymbol{Z}(t-s)^\top Y(t)dN_T(t)\Big|\mathcal{F}_t\right]\right]\\
=&E\left[\int_0^\infty E\left[I\{t>u\}I\{t>s\}\boldsymbol{Z}(t-u)\boldsymbol{Z}(t-s)^\top Y(t)dN_T(t)\Big|\mathcal{F}_t\right]\right] \qquad (36)\\
=&E\int_{u\vee s}^X \boldsymbol{Z}(t-u)\boldsymbol{Z}(t-s)^\top\lambda_{T,t}(0)dt,
\end{aligned}
$$

where the first equality follows because the integral is zero when $t_1 \neq t_2$; the second equality follows from law of iterated expectation; the third equality follows from Fubini's theorem;

the last equality follows from the definition of forward intensity. The expectation in (34) is

$$
\begin{aligned}
&E\int_0^\infty\int_0^\infty I\{t_1>u\}Y(t_1)\boldsymbol{Z}(t_1-u)\lambda_{T,t_1-u}(u)I\{t_2>s\}Y(t_2)\boldsymbol{Z}(t_2-s)^\top dt_1 dN_T(t_2)\\
=&E\int_0^\infty\int_0^{t_2} I\{t_1>u\}Y(t_1)\boldsymbol{Z}(t_1-u)\lambda_{T,t_1-u}(u)I\{t_2>s\}Y(t_2)\boldsymbol{Z}(t_2-s)^\top dt_1 dN_T(t_2)\\
=&E\left[E\left[\int_0^\infty\int_0^{t_2} I\{t_1>u\}Y(t_1)\boldsymbol{Z}(t_1-u)\lambda_{T,t_1-u}(u)I\{t_2>s\}Y(t_2)\boldsymbol{Z}(t_2-s)^\top dt_1 dN_T(t_2)\Big|\mathcal{F}_{t_2}\right]\right]\\
=&E\left[\int_0^\infty\int_0^{t_2} E\left[I\{t_1>u\}Y(t_1)\boldsymbol{Z}(t_1-u)\lambda_{T,t_1-u}(u)I\{t_2>s\}Y(t_2)\boldsymbol{Z}(t_2-s)^\top dt_1 dN_T(t_2)\Big|\mathcal{F}_{t_2}\right]\right]\\
=&E\int_s^X\int_u^{t_2}\boldsymbol{Z}(t_1-u)\lambda_{T,t_1-u}(u)\boldsymbol{Z}(t_2-s)^\top\lambda_{T,t_2}(0)dt_1dt_2, \qquad (37)
\end{aligned}
$$

where the first equality follows because the integrand is nonzero iff $t_1 \le t_2$. Similarly, the expectation in (33) is

$$
\begin{aligned}
&E\int_0^\infty\int_0^\infty I\{t_1>u\}Y(t_1)\boldsymbol{Z}(t_1-u)I\{t_2>s\}Y(t_2)\boldsymbol{Z}(t_2-s)^\top\lambda_{T,t_2}(s)dN_T(t_1)dt_2\\
=&E\int_u^X\int_s^{t_1}\boldsymbol{Z}(t_1-u)\boldsymbol{Z}(t_2-s)^\top\lambda_{T,t_2-s}(s)\lambda_{T,t_1}(0)dt_2dt_1. \qquad (38)
\end{aligned}
$$

The expectation in (35) is

$$
\begin{aligned}
&E\int_0^\infty\int_0^\infty I\{t_1>u\}Y(t_1)\boldsymbol{Z}(t_1-u)\lambda_{T,t_1}(u)I\{t_2>s\}Y(t_2)\boldsymbol{Z}(t_2-s)^\top\lambda_{T,t_2}(s)dt_1dt_2\\
=&E\int_s^X\int_u^X\boldsymbol{Z}(t_1-u)\lambda_{T,t_1-u}(u)\boldsymbol{Z}(t_2-s)^\top\lambda_{T,t_2-s}(s)dt_1dt_2. \qquad (39)
\end{aligned}
$$

Combining (36)–(39) and the definition (27) of $G(\boldsymbol{Z}_i, X_i; \boldsymbol{\alpha}, u, s)$, it can be readily seen that $E[G(\boldsymbol{Z}_i, X_i; \boldsymbol{\alpha}_0, u, s)] = \boldsymbol{\varrho}(u, s, \psi)$. This completes the proof of $\hat{\boldsymbol{\varrho}}_n(u, s; \hat{\boldsymbol{\alpha}}_n) \xrightarrow{p} \boldsymbol{\varrho}(u, s, \psi)$.

In like manner, we can show that $\boldsymbol{W}_n(u; \hat{\boldsymbol{\alpha}}_n) \xrightarrow{p} \boldsymbol{W}(u; \boldsymbol{\alpha}_0)$. Then, by the continuous mapping theorem, we have

$$
\hat{\boldsymbol{W}}_n(u; \hat{\boldsymbol{\alpha}}_n)^{-1}\hat{\boldsymbol{\varrho}}_n(u, s; \hat{\boldsymbol{\alpha}}_n)\hat{\boldsymbol{W}}_n(s; \hat{\boldsymbol{\alpha}}_n)^{-1} \xrightarrow{p} \boldsymbol{W}(u; \boldsymbol{\alpha}_0)^{-1}\boldsymbol{\varrho}(u, s, \psi)\boldsymbol{W}(s; \boldsymbol{\alpha}_0)^{-1}.
$$

In exactly the same way, we can show

$$
\hat{\boldsymbol{W}}_n(u; \hat{\boldsymbol{\beta}}_n)^{-1}\hat{\boldsymbol{\varrho}}_n(u, s; \hat{\boldsymbol{\beta}}_n)\hat{\boldsymbol{W}}_n(s; \hat{\boldsymbol{\beta}}_n)^{-1} \xrightarrow{p} \boldsymbol{W}(u; \boldsymbol{\beta}_0)^{-1}\boldsymbol{\varrho}(u, s, \vartheta)\boldsymbol{W}(s; \boldsymbol{\beta}_0)^{-1}.
$$

This implies $\hat{\boldsymbol{\rho}}_n(u, s) \xrightarrow{p} \boldsymbol{\rho}(u, s)$.

E.3 Proof of Proposition 3

In this proof, we focus on the predictive probability $F_t(\tau)$ and its estimator $\hat{F}_t(\tau)$. The results regarding $S_{T,t}(\tau)$ and $\hat{S}_{T,t}(\tau)$ can be proved in a similar way, and thus we omit the details.

Recall that $F_t(\tau)$ is a function of $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, and $\boldsymbol{Z}(t)$. To highlight its dependence on model parameters $(\boldsymbol{\alpha}, \boldsymbol{\beta})$, we write this predictive probability as $F_t(\tau; \boldsymbol{\alpha}, \boldsymbol{\beta})$. Our target is to show the weak convergence of $F_t(\tau; \hat{\boldsymbol{\alpha}}_n, \hat{\boldsymbol{\beta}}_n)$. Since $F_t(\tau; \hat{\boldsymbol{\alpha}}_n, \hat{\boldsymbol{\beta}}_n)$ is a function of $(\hat{\boldsymbol{\alpha}}_n, \hat{\boldsymbol{\beta}}_n)$ and the weak convergence of $(\hat{\boldsymbol{\alpha}}_n, \hat{\boldsymbol{\beta}}_n)$ has already been established in Theorem 1, the weak convergence of $F_t(\tau; \hat{\boldsymbol{\alpha}}_n, \hat{\boldsymbol{\beta}}_n)$ can be readily established by showing $(\boldsymbol{\alpha}, \boldsymbol{\beta}) \mapsto F_t(\tau; \boldsymbol{\alpha}, \boldsymbol{\beta})$ is Hadamard differentiable and applying the functional delta method (van der Vaart, 1998, Theorem 20.8). We show that $(\boldsymbol{\alpha}, \boldsymbol{\beta}) \mapsto F_t(\tau; \boldsymbol{\alpha}, \boldsymbol{\beta})$ is Hadamard differentiable using the chain rule. First, consider $F_t(\tau)$ as a function of $\big(\lambda_{T,t}, \lambda_{C,t}\big)$ and write $F_t(\tau) \triangleq \tilde{F}_t(\tau; \lambda_{T,t}, \lambda_{C,t})$. Let $\mathcal{E}$ be the parameter space for $\lambda_{T,t}(u) = \exp\{\boldsymbol{\alpha}(u)^\top \boldsymbol{Z}(t)\}$, where $\boldsymbol{\alpha} \in \mathcal{F}$. For any $(h_1, h_2) \in \mathcal{E} \times \mathcal{E}$, let $h_{1,\delta} \to h_1$ and $h_{2,\delta} \to h_2$ as $\delta \to 0^+$. In addition, define $\phi_1'(h_1, h_2) = \int_0^\tau S_{T,t}(u) S_{C,t}(u) h_1(u) du - \int_0^\tau S_{T,t}(u) S_{C,t}(u) \lambda_{T,t}(u) \int_0^u [h_1(s) + h_2(s)] ds du$. Consider

$$
\begin{aligned}
&\Big| \frac{\tilde{F}_t(\tau; \lambda_{T,t} + \delta h_{1,\delta}, \lambda_{C,t} + \delta h_{2,\delta}) - \tilde{F}_t(\tau; \lambda_{T,t}, \lambda_{C,t})}{\delta} - \phi_1'(h_1, h_2) \Big| \qquad (40) \\
\leq &\Big| \int_0^\tau S_{T,t}(s) S_{C,t}(u) h_{1,\delta}(u) \big[\exp\{-\delta \int_0^u [h_{1,\delta}(s) + h_{2,\delta}(s)] ds\} - 1 \big] du \Big| \\
+ &\Big| \int_0^\tau S_{T,t}(s) S_{C,t}(u) \lambda_{T,t}(u) \Big[\frac{\exp\{-\delta \int_0^u [h_{1,\delta}(s) + h_{2,\delta}(s)] ds\} - 1}{\delta} + \int_0^u [h_1(s) + h_2(s)] ds \Big] du \Big| \\
\leq &\exp\{pK_0K_1\} \delta \int_0^\tau e^{\xi^*} \int_0^u [h_{1,\delta}(s) + h_{2,\delta}(s)] ds du \\
+ &\exp\{pK_0K_1\} \int_0^\tau \Big| \int_0^u [h_1(s) + h_2(s) - h_{1,\delta}(s) - h_{2,\delta}(s)] ds + e^{\xi_*} \delta \Big\{ \int_0^u [h_{1,\delta}(s) + h_{2,\delta}(s)] ds \Big\}^2 \Big| du \\
\leq &\exp\{pK_0K_1\} \delta \int_0^\tau \int_0^u [h_{1,\delta}(s) + h_{2,\delta}(s)] ds du \\
+ &\exp\{pK_0K_1\} \left[\int_0^\tau \int_0^u |h_1(s) + h_2(s) - h_{1,\delta}(s) - h_{2,\delta}(s)| ds du + \delta \int_0^\tau \Big\{ \int_0^u [h_{1,\delta}(s) + h_{2,\delta}(s)] ds \Big\}^2 du \right] \\
\leq &2\delta\tau^2 \exp\{2pK_0K_1\} + \exp\{pK_0K_1\} \tau \int_0^\tau |h_1(s) + h_2(s) - h_{1,\delta}(s) - h_{2,\delta}(s)| ds + 4\delta\tau^3 \exp\{3pK_0K_1\},
\end{aligned}
$$

where ξ_* lies between zero and $-\delta \int_0^u [h_{1,\delta}(s) + h_{2,\delta}(s)] ds$. In the above display, the first inequality is obvious by plugging in the detailed expressions of $\tilde{F}_t(\tau; \lambda_{T,t} + \delta h_1, \lambda_{C,t} + \delta h_2)$ and $\tilde{F}_t(\tau; \lambda_{T,t}, \lambda_{C,t})$; the second inequality follows from the Taylor expansion and Conditions (A1) and (A2). By dominated convergence theorem, we can see that the last row in the above display uniformly converge to zero as $\delta \to 0$. Hence $(\lambda_{T,t}, \lambda_{C,t}) \mapsto \tilde{F}_t(\tau; \lambda_{T,t}, \lambda_{C,t})$ is Hadamard differetiable, and the Hadamard derivative is $\phi_1' : \mathcal{E} \times \mathcal{E} \mapsto \mathbb{R}$. In a similar way, we can prove that the function $(\boldsymbol{\alpha}, \boldsymbol{\beta}) \mapsto (\lambda_{T,t}, \lambda_{C,t})$ is also Hadamard differentiable for any $(\boldsymbol{\alpha}, \boldsymbol{\beta}) \in \mathcal{F} \times \mathcal{F}$, and the Hadamard derivative is

$$
\phi_2'(h_3, h_4) = \big(\lambda_{T,t} h_3^\top \boldsymbol{Z}(t), \lambda_{C,t} h_4^\top \boldsymbol{Z}(t) \big)^\top,
$$

where $(h_3, h_4) \in \mathcal{F} \times \mathcal{F}$. Based on the chain rule (van der Vaart, 1998, Theorem 20.9), the function $(\boldsymbol{\alpha}, \boldsymbol{\beta}) \mapsto F_t(\tau; \boldsymbol{\alpha}, \boldsymbol{\beta})$ is Hadamard differentiable at $(\boldsymbol{\alpha}_0, \boldsymbol{\beta}_0)$, and the Hadamard

derivative $\phi' : \mathcal{F} \times \mathcal{F} \mapsto \mathbb{R}$ at $(\boldsymbol{\alpha}_0, \boldsymbol{\beta}_0)$ is

$$\begin{aligned}\phi'(h_1, h_2) = \phi_1' \circ \phi_2' = &\int_0^\tau S_{T,t}(u) S_{C,t}(u) \lambda_{T,t}(u) h_1(u)^\top \boldsymbol{Z}(t) du \\ &- \int_0^\tau S_{T,t}(u) S_{C,t}(u) \lambda_{T,t}(u) \int_0^u [\lambda_{T,t}(s) h_1(s)^\top \boldsymbol{Z}(t) + \lambda_{C,t}(s) h_2(s)^\top \boldsymbol{Z}(t)] ds du.\end{aligned}$$

Therefore, based on Theorem 1 and the functional delta method, we have

$$\sqrt{n}\big(F_t(\tau; \hat{\boldsymbol{\alpha}}_n, \hat{\boldsymbol{\beta}}_n) - F_t(\tau; \boldsymbol{\alpha}_0, \boldsymbol{\beta}_0)\big) \rightsquigarrow N(0, \eta^2),$$

where

$$\begin{aligned}\eta^2 = \int_0^\tau \int_0^\tau &S_{T,t}(u) S_{C,t}(u) \lambda_{T,t}(u) S_{T,t}(s) S_{C,t}(s) \lambda_{T,t}(s) \\ &[Q_1(u,s) - Q_2(u,s) - Q_3(u,s) + Q_4(u,s)] du ds, \end{aligned} \tag{41}$$

with

$$\begin{aligned}Q_1(u,s) &= \boldsymbol{Z}(t)^\top \boldsymbol{W}(u; \boldsymbol{\alpha}_0)^{-1} \boldsymbol{\varrho}(u, s, \psi) \boldsymbol{W}(s; \boldsymbol{\alpha}_0)^{-1} \boldsymbol{Z}(t), \\ Q_2(u,s) &= \int_0^s \lambda_{T,t}(\nu) Q_1(u, \nu) d\nu, \\ Q_3(u,s) &= \int_0^u \lambda_{T,t}(\nu) Q_1(s, \nu) d\nu, \\ Q_4(u,s) &= \boldsymbol{Z}(t)^\top \int_0^u \int_0^s \lambda_{C,t}(\nu_1) \lambda_{C,t}(\nu_2) \boldsymbol{W}(\nu_1; \boldsymbol{\beta}_0)^{-1} \boldsymbol{\varrho}(\nu_1, \nu_2, \vartheta) \boldsymbol{W}(\nu_2; \boldsymbol{\beta}_0)^{-1} d\nu_1 d\nu_2 \boldsymbol{Z}(t) \\ &\quad + \int_0^u \int_0^s \lambda_{T,t}(\nu_1) \lambda_{T,t}(\nu_2) Q_1(\nu_1, \nu_2) d\nu_1 d\nu_2.\end{aligned}$$

The asymptotic variance η^2 can be estimated by

$$\begin{aligned}\hat{\eta}_n^2 = \int_0^\tau \int_0^\tau &\hat{S}_{T,t}(u) \hat{S}_{C,t}(u) \hat{\lambda}_{T,t}(u) \hat{S}_{T,t}(s) \hat{S}_{C,t}(s) \hat{\lambda}_{T,t}(s) \\ &[\hat{Q}_1(u,s) - \hat{Q}_2(u,s) - \hat{Q}_3(u,s) + \hat{Q}_4(u,s)] du ds, \end{aligned} \tag{42}$$

where

$$\begin{aligned}\hat{Q}_1(u,s) &= \boldsymbol{Z}(t)^\top \hat{\boldsymbol{W}}_n(u; \hat{\boldsymbol{\alpha}}_n)^{-1} \hat{\boldsymbol{\varrho}}_n(u, s; \hat{\boldsymbol{\alpha}}_n) \hat{\boldsymbol{W}}_n(s; \hat{\boldsymbol{\alpha}}_n)^{-1} \boldsymbol{Z}(t), \\ \hat{Q}_2(u,s) &= \int_0^s \hat{\lambda}_{T,t}(\nu) \hat{Q}_1(u, \nu) d\nu, \\ \hat{Q}_3(u,s) &= \int_0^u \hat{\lambda}_{T,t}(\nu) \hat{Q}_1(s, \nu) d\nu, \\ \hat{Q}_4(u,s) &= \boldsymbol{Z}(t)^\top \int_0^u \int_0^s \hat{\lambda}_{C,t}(\nu_1) \hat{\lambda}_{C,t}(\nu_2) \hat{\boldsymbol{W}}_n(\nu_1; \hat{\boldsymbol{\beta}}_n)^{-1} \hat{\boldsymbol{\varrho}}_n(\nu_1, \nu_2; \hat{\boldsymbol{\beta}}_n) \hat{\boldsymbol{W}}_n(\nu_2; \hat{\boldsymbol{\beta}}_n)^{-1} d\nu_1 d\nu_2 \boldsymbol{Z}(t) \\ &\quad + \int_0^u \int_0^s \hat{\lambda}_{T,t}(\nu_1) \hat{\lambda}_{T,t}(\nu_2) \hat{Q}_1(\nu_1, \nu_2) d\nu_1 d\nu_2.\end{aligned}$$

Conditional on $\boldsymbol{Z}(t)$, the asymptotic variance η^2 is a function of $\boldsymbol{\alpha}_0$ and $\boldsymbol{\beta}_0$. To highlight its dependence on $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, we write $\eta^2 = \eta^2(\boldsymbol{\alpha}_0, \boldsymbol{\beta}_0)$. Then, the estimator for η^2 is essentially $\hat{\eta}_n^2 \triangleq \eta^2(\hat{\boldsymbol{\alpha}}_n, \hat{\boldsymbol{\beta}}_n)$. The weak convergence of $(\hat{\boldsymbol{\alpha}}_n, \hat{\boldsymbol{\beta}}_n)$ to $(\boldsymbol{\alpha}_0, \boldsymbol{\beta}_0)$ has been established in Theorem 1. In addition, we can follow the derivation in (40) to show that $(\boldsymbol{\alpha}, \boldsymbol{\beta}) \mapsto \eta^2(\boldsymbol{\alpha}, \boldsymbol{\beta})$ is Hadamard differentiable. Therefore, we can apply the functional delta method to show that $\sqrt{n}(\hat{\eta}_n^2 - \eta^2) \rightsquigarrow [\eta^2]'(\boldsymbol{\alpha}_0, \boldsymbol{\beta}_0)$, where $(\boldsymbol{\alpha}, \boldsymbol{\beta}) \mapsto [\eta^2]'(\boldsymbol{\alpha}, \boldsymbol{\beta})$ is the Hadamard derivative of $(\boldsymbol{\alpha}, \boldsymbol{\beta}) \mapsto \eta^2(\boldsymbol{\alpha}, \boldsymbol{\beta})$. This implies $\sqrt{n}(\hat{\eta}_n^2 - \eta^2) = O_P(1)$.

E.4 Proof of Proposition 4

Proposition 4 can be proved in the same manner as Proposition 3. The argument proceeds analogously. Hence, we omit the details.

E.5 Proof of Proposition 5

Define $\mathcal{S} : \mathcal{F} \mapsto (\ell^\infty([0, \bar{\tau}]))^p$ as following:

$$[\mathcal{S}(\boldsymbol{\alpha})](\tau) = \exp\left\{ -\int_0^\tau \exp\{\boldsymbol{Z}(t)^\top \boldsymbol{\alpha}(u)\} du \right\}.$$

Then we have $\hat{S}_{T,t}(\tau) = [\mathcal{S}(\hat{\boldsymbol{\alpha}}_n)](\tau)$. The weak convergence of $\hat{\boldsymbol{\alpha}}_n$ has been established in Theorem 1. Next, we will establish the weak convergence of $[\mathcal{S}(\hat{\boldsymbol{\alpha}}_n)](\tau)$ using the functional delta method.

First, we show that $\boldsymbol{\alpha} \mapsto [\mathcal{S}(\boldsymbol{\alpha})](\tau)$ is Hadamard differentiable. Define $[\lambda_{T,t}(\boldsymbol{\alpha})](u) = \exp\{\boldsymbol{Z}(t)^\top \boldsymbol{\alpha}(u)\}$ and $[\mathcal{S}_1(\lambda_T)](\tau) = \exp\{-\int_0^\tau \lambda_{T,t}(u) du\}$. Then, we can write $\mathcal{S} = \mathcal{S}_1 \circ \lambda_{T,t}$, and we will use the chain rule to show that $\mathcal{S}$ is Hadamard differentiable. We begin with $\mathcal{S}_1$. Let $\mathcal{E}$ be the parameter space for $\lambda_{T,t}(\cdot)$. For any h, let $h_\delta \to h$ as $\delta \to 0^+$, and define $\phi_3'(h) = -\mathcal{S}_1(\lambda_{T,t}) \int_0^\tau h(u) du$. Then, we have

$$\begin{aligned}
&\left\| \frac{\mathcal{S}_1(\lambda_{T,t} + \delta h_\delta) - \mathcal{S}_1(\lambda_{T,t})}{\delta} - \phi_3'(h) \right\| \\
&= \left\| \mathcal{S}_1(\lambda_{T,t})[\tau] \left[\frac{\exp\left\{-\int_0^\tau \delta h_\delta(u) du\right\} - 1}{\delta} + \int_0^\tau h(u) du \right] \right\| \\
&\le \left\| \frac{\exp\left\{-\int_0^\tau \delta h_\delta(u) du\right\} - \exp\{-\int_0^\tau \delta h(u) du\}}{\delta} + \frac{\exp\left\{-\int_0^\tau \delta h(u) du\right\} - 1}{\delta} + \int_0^\tau h(u) du \right\| \\
&\le \left\| \frac{\exp\left\{-\int_0^\tau \delta h_\delta(u) du\right\} - \exp\{-\int_0^\tau \delta h(u) du\}}{\delta} \right\| + \left\| \frac{\exp\left\{-\int_0^\tau \delta h(u) du\right\} - 1}{\delta} + \int_0^\tau h(u) du \right\| \\
&\le \left\| \int_0^\tau [h_\delta(u) - h(u)] du \right\| + \left\| \frac{\exp\left\{-\int_0^\tau \delta h(u) du\right\} - 1}{\delta} + \int_0^\tau h(u) du \right\| \\
&\le \int_0^{\bar{\tau}} |h_\delta(u) - h(u)| du + \left\| 1 - \exp\{-\tilde{\delta} \int_0^\tau h(u) du\} \right\| \left\| \int_0^\tau h(u) du \right\|, \qquad (43)
\end{aligned}$$

where $\tilde{\delta}$ lies between 0 and δ. As $\delta \to 0$, (43) goes to zero. Thus, $\mathcal{S}_1 : \mathcal{E} \mapsto (\ell^\infty([0, \bar{\tau}]))^p$ is Hadamard differentiable and the Hadamard derivative is $\phi_3'(h) = -\mathcal{S}_1(\lambda_{T,t}) \int_0^\tau h(u) du$. In

addition, as shown in the proof of Proposition 3, $\lambda_{T,t} : \mathcal{F} \mapsto \mathcal{E}$ is Hadamard differentiable and the Hadamard derivative is $\phi_4'(h) = \lambda_{T,t}(u)h(u)^\top \boldsymbol{Z}(t)$ for $h \in \mathcal{F}$. Based on the chain rule (van der Vaart, 1998, Theorem 20.9), we can conclude that $\mathcal{S} : \mathcal{F} \mapsto (\ell^\infty([0,\bar{\tau}]))^p$ is Hadamard differentiable and its Hadamard derivative is

$$\phi'(h) = \phi_3'(\phi_4'(h)) = -\exp\left\{-\int_0^\tau \lambda_{T,t}(u)du\right\}\int_0^\tau \lambda_{T,t}(u)h(u)^\top \boldsymbol{Z}(t)du,$$

for $h \in \mathcal{F}$. Then, based on Theorem 1 and the functional delta method, $\sqrt{n}(\hat{S}_{T,t} - S_{T,t})$ weakly converges to a Gaussian process with mean zero and covariance function $\rho_S(u,v)$, $u, v \in [0,\bar{\tau}]$, given by

$$\rho_S(u,v) = S_{T,t}(u)S_{T,t}(v)\int_0^u\int_0^v \lambda_{T,t}(s_1)\lambda_{T,t}(s_2)\boldsymbol{Z}(t)^\top \boldsymbol{\varrho}_T(s_1,s_2)\boldsymbol{Z}(t)ds_1ds_2,$$

where $\boldsymbol{\varrho}_T(s_1,s_2) = \boldsymbol{W}(s_1;\boldsymbol{\alpha}_0)^{-1}\boldsymbol{\varrho}(s_1,s_2,\psi)\boldsymbol{W}(s_2;\boldsymbol{\alpha}_0)^{-1}$. The covariance function can be consistently estimated by

$$\hat{\rho}_S(u,v) = \hat{S}_{T,t}(u)\hat{S}_{T,t}(v)\int_0^u\int_0^v \hat{\lambda}_{T,t}(s_1)\hat{\lambda}_{T,t}(s_2)\boldsymbol{Z}(t)^\top \hat{\boldsymbol{\varrho}}_T(s_1,s_2)\boldsymbol{Z}(t)ds_1ds_2,$$

where $\hat{\boldsymbol{\varrho}}_T(s_1,s_2) = \hat{\boldsymbol{W}}_n(s_1;\hat{\boldsymbol{\alpha}}_n)^{-1}\hat{\boldsymbol{\varrho}}_n(s_1,s_2;\hat{\boldsymbol{\alpha}}_n)\hat{\boldsymbol{W}}_n(s_2;\hat{\boldsymbol{\alpha}}_n)^{-1}$. The proof for the weak convergence of $\sqrt{n}(\hat{F}_t - F_t)$ is similar and thus the details are omitted.

E.6 Proof of Theorem 6

We first give a useful lemma and then prove Theorem 6.

Lemma 11 *Suppose Conditions (B1)–(B3) hold. Define* $\check{\ell}_n(\underline{\boldsymbol{\alpha}},\underline{\boldsymbol{\beta}};\mathbf{D}_{\text{disct}}) \triangleq \tilde{\ell}_n(\underline{\boldsymbol{\alpha}};\mathbf{D}_{\text{disct}}) + \tilde{\ell}_n(\underline{\boldsymbol{\beta}};\mathbf{D}_{\text{disct}})$. *Then we have*

$$\text{Var}[\sqrt{n}\nabla\check{\ell}_n(\underline{\boldsymbol{\alpha}}_0,\underline{\boldsymbol{\beta}}_0)] = \begin{bmatrix}\text{Var}[\sqrt{n}\nabla\tilde{\ell}_n(\underline{\boldsymbol{\alpha}}_0)] & \mathbf{0} \\ \mathbf{0} & \text{Var}[\sqrt{n}\nabla\tilde{\ell}_n(\underline{\boldsymbol{\beta}}_0)]\end{bmatrix}.$$

Proof Denote $\boldsymbol{Z}_i^{(k)}(t) = (\mathbf{0}_p^\top,\ldots,\mathbf{0}_p^\top,\boldsymbol{Z}_i(t)^\top,\mathbf{0}_p^\top,\ldots,\mathbf{0}_p^\top)^\top$, where $\boldsymbol{Z}_i(t)$ locates at the kth coordinate, $k = 1,2,\ldots,\kappa$. Then $\nabla\tilde{\ell}_n(\underline{\boldsymbol{\alpha}};\mathbf{D}_{\text{disct}})$ can be expressed as

$$\begin{aligned}&\nabla\tilde{\ell}_n(\underline{\boldsymbol{\alpha}};\mathbf{D}_{\text{disct}}) \\ =&\frac{1}{n}\sum_{i=1}^n\sum_{k=0}^{\kappa-1}\sum_{h=1}^{\infty}\boldsymbol{Z}_i^{(k)}(t_h)\left[-\Delta t\lambda_{T_i,t_h}(t_k) + \delta_{T_i}(t_h+t_{k+1})\frac{\Delta t\lambda_{T_i,t_h}(t_k)}{1-\exp\{-\Delta t\lambda_{T_i,t_h}(t_k)\}}\right]\mathbf{1}_{(t_h+t_k,\infty)}(X_i).\end{aligned}$$

Based on the law of iterated expectation and the fact that $E[\delta_{T_i}(t_h+t_{k+1})|\mathcal{F}_{t_h}] = [1-\exp\{-\Delta t\lambda_{T_i,t_h}(t_k)\}]P(T_i \geq t_h+t_k|\mathcal{F}_{t_h})$, we have $E[\nabla\tilde{\ell}_n(\underline{\boldsymbol{\alpha}}_0;\mathbf{D}_{\text{disct}})] = \mathbf{0}$. Similarly, we can also show $E[\nabla\tilde{\ell}_n(\underline{\boldsymbol{\beta}}_0;\mathbf{D}_{\text{disct}})] = \mathbf{0}$. Thus $E[\nabla\check{\ell}_n(\underline{\boldsymbol{\alpha}}_0,\underline{\boldsymbol{\beta}}_0;\mathbf{D}_{\text{disct}})] = \mathbf{0}$ and

$\mathrm{Var}[\sqrt{n}\nabla\check{\ell}_n(\underline{\boldsymbol{\alpha}}_0,\underline{\boldsymbol{\beta}}_0)]$ can be expressed as $= nE\nabla\check{\ell}_n(\underline{\boldsymbol{\alpha}}_0,\underline{\boldsymbol{\beta}}_0)\big(\nabla\check{\ell}_n(\underline{\boldsymbol{\alpha}}_0,\underline{\boldsymbol{\beta}}_0)\big)^\top$.

$$\begin{aligned}\mathrm{Var}[\sqrt{n}\nabla\check{\ell}_n(\underline{\boldsymbol{\alpha}}_0,\underline{\boldsymbol{\beta}}_0)] &= nE\nabla\check{\ell}_n(\underline{\boldsymbol{\alpha}}_0,\underline{\boldsymbol{\beta}}_0)\big(\nabla\check{\ell}_n(\underline{\boldsymbol{\alpha}}_0,\underline{\boldsymbol{\beta}}_0)\big)^\top\\ &= \begin{bmatrix} nE\nabla\tilde{\ell}_n(\underline{\boldsymbol{\alpha}}_0)\big(\nabla\tilde{\ell}_n(\underline{\boldsymbol{\alpha}}_0)\big)^\top & nE\nabla\tilde{\ell}_n(\underline{\boldsymbol{\alpha}}_0)\big(\nabla\tilde{\ell}_n(\underline{\boldsymbol{\beta}}_0)\big)^\top\\ nE\nabla\tilde{\ell}_n(\underline{\boldsymbol{\beta}}_0)\big(\nabla\tilde{\ell}_n(\boldsymbol{\alpha}_0)\big)^\top & nE\nabla\tilde{\ell}_n(\underline{\boldsymbol{\beta}}_0)\big(\nabla\tilde{\ell}_n(\underline{\boldsymbol{\beta}}_0)\big)^\top\end{bmatrix}\\ &= \begin{bmatrix} \mathrm{Var}[\sqrt{n}\nabla\tilde{\ell}_n(\underline{\boldsymbol{\alpha}}_0)] & nE\nabla\tilde{\ell}_n(\underline{\boldsymbol{\alpha}}_0)\big(\nabla\tilde{\ell}_n(\underline{\boldsymbol{\beta}}_0)\big)^\top\\ nE\nabla\tilde{\ell}_n(\underline{\boldsymbol{\beta}}_0)\big(\nabla\tilde{\ell}_n(\boldsymbol{\alpha}_0)\big)^\top & \mathrm{Var}[\sqrt{n}\nabla\tilde{\ell}_n(\underline{\boldsymbol{\beta}}_0)]\end{bmatrix}.\end{aligned}$$

We then show the off-diagonal elements are zeros. First, note that

$$\begin{aligned}&nE\nabla\tilde{\ell}_n(\underline{\boldsymbol{\alpha}}_0)\big(\nabla\tilde{\ell}_n(\underline{\boldsymbol{\beta}}_0)\big)^\top\\ =&\sum_{k_1=0}^{\kappa-1}\sum_{k_2=0}^{\kappa-1}\sum_{h_1=1}^{\infty}\sum_{h_2=1}^{\infty}E\Big\{\boldsymbol{Z}_i^{(k)}(t_{h_1})\Big[\frac{\delta_{T_i}(t_{h_1}+t_{k_1+1})\Delta t\lambda_{T_i,t_{h_1}}(t_{k_1})}{1-\exp\{-\Delta t\lambda_{T_i,t_{h_1}}(t_{k_1})\}}-\Delta t\lambda_{T_i,t_{h_1}}(t_{k_1})\Big]\mathbf{1}_{(t_{h_1}+t_{k_1},\infty)}(X_i)\times\\ &\qquad\boldsymbol{Z}_i^{(k)}(t_{h_2})^\top\Big[\frac{\delta_{C_i}(t_{h_2}+t_{k_2+1})\Delta t\lambda_{C_i,t_{h_2}}(t_{k_2})}{1-\exp\{-\Delta t\lambda_{C_i,t_{h_2}}(t_{k_2})\}}-\Delta t\lambda_{C_i,t_{h_2}}(t_{k_2})\Big]\mathbf{1}_{(t_{h_2}+t_{k_2},\infty)}(X_i)\Big\}.\end{aligned}\tag{44}$$

Next, we show $nE\nabla\tilde{\ell}_n(\underline{\boldsymbol{\alpha}}_0)\big(\nabla\tilde{\ell}_n(\underline{\boldsymbol{\beta}}_0)\big)^\top=\mathbf{0}$ by showing the terms in the summation in the above display are all zeros. If $t_{h_1}>t_{h_2}$, we have

$$\begin{aligned}&E\Big\{\boldsymbol{Z}_i^{(k)}(t_{h_1})\Big[\frac{\delta_{T_i}(t_{h_1}+t_{k_1+1})\Delta t\lambda_{T_i,t_{h_1}}(t_{k_1})}{1-\exp\{-\Delta t\lambda_{T_i,t_{h_1}}(t_{k_1})\}}-\Delta t\lambda_{T_i,t_{h_1}}(t_{k_1})\Big]\mathbf{1}_{(t_{h_1}+t_{k_1},\infty)}(X_i)\times\\ &\quad\boldsymbol{Z}_i^{(k)}(t_{h_2})^\top\Big[\frac{\delta_{C_i}(t_{h_2}+t_{k_2+1})\Delta t\lambda_{C_i,t_{h_2}}(t_{k_2})}{1-\exp\{-\Delta t\lambda_{C_i,t_{h_2}}(t_{k_2})\}}-\Delta t\lambda_{C_i,t_{h_2}}(t_{k_2})\Big]\mathbf{1}_{(t_{h_2}+t_{k_2},\infty)}(X_i)\Big\}\\ =&E\Big\{E\Big[\boldsymbol{Z}_i^{(k)}(t_{h_1})\Big[\frac{\delta_{T_i}(t_{h_1}+t_{k_1+1})\Delta t\lambda_{T_i,t_{h_1}}(t_{k_1})}{1-\exp\{-\Delta t\lambda_{T_i,t_{h_1}}(t_{k_1})\}}-\Delta t\lambda_{T_i,t_{h_1}}(t_{k_1})\Big]\mathbf{1}_{(t_{h_1}+t_{k_1},\infty)}(X_i)\times\\ &\quad\boldsymbol{Z}_i^{(k)}(t_{h_2})^\top\Big[\frac{\delta_{C_i}(t_{h_2}+t_{k_2+1})\Delta t\lambda_{C_i,t_{h_2}}(t_{k_2})}{1-\exp\{-\Delta t\lambda_{C_i,t_{h_2}}(t_{k_2})\}}-\Delta t\lambda_{C_i,t_{h_2}}(t_{k_2})\Big]\mathbf{1}_{(t_{h_2}+t_{k_2},\infty)}(X_i)\Big|\mathcal{F}_{t_{h_1}}\Big]\Big\}\\ =&E\Big\{E\Big[\boldsymbol{Z}_i^{(k)}(t_{h_1})\Big[\frac{\delta_{T_i}(t_{h_1}+t_{k_1+1})\Delta t\lambda_{T_i,t_{h_1}}(t_{k_1})}{1-\exp\{-\Delta t\lambda_{T_i,t_{h_1}}(t_{k_1})\}}-\Delta t\lambda_{T_i,t_{h_1}}(t_{k_1})\Big]\mathbf{1}_{(t_{h_1}+t_{k_1},\infty)}(X_i)\Big|\mathcal{F}_{t_{h_1}}\Big]\\ &\quad E\Big[\boldsymbol{Z}_i^{(k)}(t_{h_2})^\top\Big[\frac{\delta_{C_i}(t_{h_2}+t_{k_2+1})\Delta t\lambda_{C_i,t_{h_2}}(t_{k_2})}{1-\exp\{-\Delta t\lambda_{C_i,t_{h_2}}(t_{k_2})\}}-\Delta t\lambda_{C_i,t_{h_2}}(t_{k_2})\Big]\mathbf{1}_{(t_{h_2}+t_{k_2},\infty)}(X_i)\Big|\mathcal{F}_{t_{h_1}}\Big]\Big\}\\ =&E\Big\{\mathbf{0}\times E\Big[\boldsymbol{Z}_i^{(k)}(t_{h_2})^\top\Big[\frac{\delta_{C_i}(t_{h_2}+t_{k_2+1})\Delta t\lambda_{C_i,t_{h_2}}(t_{k_2})}{1-\exp\{-\Delta t\lambda_{C_i,t_{h_2}}(t_{k_2})\}}-\Delta t\lambda_{C_i,t_{h_2}}(t_{k_2})\Big]\mathbf{1}_{(t_{h_2}+t_{k_2},\infty)}(X_i)\Big|\mathcal{F}_{t_{h_1}}\Big]\Big\}=\mathbf{0},\end{aligned}$$

where the first equality follows from the law of iterated expectation; the second equality follows from the conditional independence between T and C; the third equality follows from the fact that $E[\delta_{T_i}(t_{h_1}+t_{k_1+1})|\mathcal{F}_{t_{h_1}}]=[1-\exp\{-\Delta t\lambda_{T_i,t_{h_1}}(t_{k_1})\}]P(T_i\geq t_{h_1}+t_{k_1}|\mathcal{F}_{t_{h_1}})$.

Similarly, if $t_{h_1} \leq t_{h_2}$, we have

$$
\begin{aligned}
&E\Big\{ \boldsymbol{Z}_i^{(k)}(t_{h_1}) \Big[\frac{\delta_{T_i}(t_{h_1}+t_{k_1+1})\Delta t \lambda_{T_i,t_{h_1}}(t_{k_1})}{1-\exp\{-\Delta t \lambda_{T_i,t_{h_1}}(t_{k_1})\}} - \Delta t \lambda_{T_i,t_{h_1}}(t_{k_1}) \Big] \mathbf{1}_{(t_{h_1}+t_{k_1},\infty)}(X_i) \times \\
&\quad \boldsymbol{Z}_i^{(k)}(t_{h_2})^\top \Big[\frac{\delta_{C_i}(t_{h_2}+t_{k_2+1})\Delta t \lambda_{C_i,t_{h_2}}(t_{k_2})}{1-\exp\{-\Delta t \lambda_{C_i,t_{h_2}}(t_{k_2})\}} - \Delta t \lambda_{C_i,t_{h_2}}(t_{k_2}) \Big] \mathbf{1}_{(t_{h_2}+t_{k_2},\infty)}(X_i) \Big\} \\
=&E\Bigg\{ E\Bigg[\boldsymbol{Z}_i^{(k)}(t_{h_1}) \Big[\frac{\delta_{T_i}(t_{h_1}+t_{k_1+1})\Delta t \lambda_{T_i,t_{h_1}}(t_{k_1})}{1-\exp\{-\Delta t \lambda_{T_i,t_{h_1}}(t_{k_1})\}} - \Delta t \lambda_{T_i,t_{h_1}}(t_{k_1}) \Big] \mathbf{1}_{(t_{h_1}+t_{k_1},\infty)}(X_i) \times \\
&\quad \boldsymbol{Z}_i^{(k)}(t_{h_2})^\top \Big[\frac{\delta_{C_i}(t_{h_2}+t_{k_2+1})\Delta t \lambda_{C_i,t_{h_2}}(t_{k_2})}{1-\exp\{-\Delta t \lambda_{C_i,t_{h_2}}(t_{k_2})\}} - \Delta t \lambda_{C_i,t_{h_2}}(t_{k_2}) \Big] \mathbf{1}_{(t_{h_2}+t_{k_2},\infty)}(X_i) \Big| \mathcal{F}_{t_{h_2}} \Bigg] \Bigg\} \\
=&E\Bigg\{ E\Bigg[\boldsymbol{Z}_i^{(k)}(t_{h_1}) \Big[\frac{\delta_{T_i}(t_{h_1}+t_{k_1+1})\Delta t \lambda_{T_i,t_{h_1}}(t_{k_1})}{1-\exp\{-\Delta t \lambda_{T_i,t_{h_1}}(t_{k_1})\}} - \Delta t \lambda_{T_i,t_{h_1}}(t_{k_1}) \Big] \mathbf{1}_{(t_{h_1}+t_{k_1},\infty)}(X_i) \Big| \mathcal{F}_{t_{h_2}} \Bigg] \\
&\quad E\Bigg[\boldsymbol{Z}_i^{(k)}(t_{h_2})^\top \Big[\frac{\delta_{C_i}(t_{h_2}+t_{k_2+1})\Delta t \lambda_{C_i,t_{h_2}}(t_{k_2})}{1-\exp\{-\Delta t \lambda_{C_i,t_{h_2}}(t_{k_2})\}} - \Delta t \lambda_{C_i,t_{h_2}}(t_{k_2}) \Big] \mathbf{1}_{(t_{h_2}+t_{k_2},\infty)}(X_i) \Big| \mathcal{F}_{t_{h_2}} \Bigg] \Bigg\} \\
=&E\Bigg\{ E\Bigg[\boldsymbol{Z}_i^{(k)}(t_{h_1}) \Big[\frac{\delta_{T_i}(t_{h_1}+t_{k_1+1})\Delta t \lambda_{T_i,t_{h_1}}(t_{k_1})}{1-\exp\{-\Delta t \lambda_{T_i,t_{h_1}}(t_{k_1})\}} - \Delta t \lambda_{T_i,t_{h_1}}(t_{k_1}) \Big] \mathbf{1}_{(t_{h_1}+t_{k_1},\infty)}(X_i) \Big| \mathcal{F}_{t_{h_2}} \Bigg] \times \mathbf{0} \Bigg\} = \mathbf{0}.
\end{aligned}
$$

The terms in the summation in (44) are all zeros, and thus we have $nE\nabla\tilde{\ell}_n(\underline{\boldsymbol{\alpha}}_0)\big(\nabla\tilde{\ell}_n(\underline{\boldsymbol{\beta}}_0)\big)^\top = \mathbf{0}$ and

$$
\mathrm{Var}[\sqrt{n}\nabla\check{\ell}_n(\underline{\boldsymbol{\alpha}}_0, \underline{\boldsymbol{\beta}}_0)] = \begin{bmatrix} \mathrm{Var}[\sqrt{n}\nabla\tilde{\ell}_n(\underline{\boldsymbol{\alpha}}_0)] & \mathbf{0} \\ \mathbf{0} & \mathrm{Var}[\sqrt{n}\nabla\tilde{\ell}_n(\underline{\boldsymbol{\beta}}_0)] \end{bmatrix}.
$$

■

Proof of Theorem 6: First, we prove $\hat{\underline{\boldsymbol{\alpha}}}_n \xrightarrow{p} \underline{\boldsymbol{\alpha}}_0$. Let $Q(\underline{\boldsymbol{\alpha}}) = E[-\nabla^2\tilde{\ell}_n(\underline{\boldsymbol{\alpha}}; \mathbf{D}_{\text{disct}})]$. According to Condition (B2), $Q(\underline{\boldsymbol{\alpha}})$ exists and it is positive definite. Define $B_n(c) = \{\underline{\boldsymbol{\alpha}} : \|[Q(\underline{\boldsymbol{\alpha}}_0)]^{1/2}\sqrt{n}(\underline{\boldsymbol{\alpha}} - \underline{\boldsymbol{\alpha}}_0)\| \leq c\}$, which shrinks to $\underline{\boldsymbol{\alpha}}_0$ as $n \to \infty$. Following the proof of Theorem 4.17 in Shao (2003), we show the weak consistency of $\hat{\underline{\boldsymbol{\alpha}}}_n$ by showing that for any $\epsilon > 0$, there exists $c > 0$ and $n_0 > 0$ such that

$$
P\big(n\tilde{\ell}_n(\underline{\boldsymbol{\gamma}}; \mathbf{D}_{\text{disct}}) - n\tilde{\ell}_n(\underline{\boldsymbol{\alpha}}_0; \mathbf{D}_{\text{disct}}) < 0 \text{ for all } \underline{\boldsymbol{\gamma}} \in \partial B_n(c)\big) > 1-\epsilon, \ \forall n > n_0, \tag{45}
$$

where $\partial B_n(c)$ is the boundary of $B_n(c)$. For $\underline{\boldsymbol{\gamma}} \in \partial B_n(c)$, expand $n\tilde{\ell}_n(\underline{\boldsymbol{\gamma}}; \mathbf{D}_{\text{disct}})$ at the truth $\underline{\boldsymbol{\alpha}}_0$ yields

$$
\begin{aligned}
&n\tilde{\ell}_n(\underline{\boldsymbol{\gamma}}; \mathbf{D}_{\text{disct}}) - n\tilde{\ell}_n(\underline{\boldsymbol{\alpha}}_0; \mathbf{D}_{\text{disct}}) \\
=&c\boldsymbol{\lambda}^\top[nQ(\underline{\boldsymbol{\alpha}}_0)]^{-1/2} n\nabla\tilde{\ell}_n(\underline{\boldsymbol{\alpha}}_0; \mathbf{D}_{\text{disct}}) + \frac{c^2}{2}\boldsymbol{\lambda}^\top[Q(\underline{\boldsymbol{\alpha}}_0)]^{-1/2}\nabla^2\tilde{\ell}_n(\underline{\boldsymbol{\gamma}}^*; \mathbf{D}_{\text{disct}})[Q(\underline{\boldsymbol{\alpha}}_0)]^{-1/2}\boldsymbol{\lambda},
\end{aligned} \tag{46}
$$

where $\boldsymbol{\lambda} = [nQ(\underline{\boldsymbol{\alpha}}_0)]^{1/2}(\underline{\boldsymbol{\gamma}} - \underline{\boldsymbol{\alpha}}_0)/c$ satisfying $\|\boldsymbol{\lambda}\| = 1$, and $\underline{\boldsymbol{\gamma}}^*$ lies between $\underline{\boldsymbol{\gamma}}$ and $\underline{\boldsymbol{\alpha}}_0$. Write $\tilde{\ell}_n(\underline{\boldsymbol{\alpha}}; \mathbf{D}_{\text{disct}}) = \frac{1}{n}\sum_{i=1}^n \tilde{\ell}(\underline{\boldsymbol{\alpha}}; \mathbf{D}_{i,\text{disct}})$, where $\mathbf{D}_{i,\text{disct}}$ denotes the observed data for the ith

subject and

$$\tilde{\ell}\left(\underline{\boldsymbol{\alpha}};\mathbf{D}_{i,\text{disct}}\right)=\sum_{k=0}^{\kappa-1}\sum_{j=1}^{M_i-k-1}\Big\{-\lambda_{T_i,t_j}(t_k)\Delta t+\tilde{\delta}_{T_i}(t_j+t_{k+1})\log\frac{1-\exp\{-\lambda_{T_i,t_j}(t_k)\Delta t\}}{\exp\{-\lambda_{T_i,t_j}(t_k)\Delta t\}}\Big\}.$$

Then we have the following inequality

$$\begin{aligned}E\|\nabla^2\tilde{\ell}_n\left(\boldsymbol{\gamma}^*;\mathbf{D}_{\text{disct}}\right)-\nabla^2\tilde{\ell}_n\left(\underline{\boldsymbol{\alpha}}_0;\mathbf{D}_{\text{disct}}\right)\|&\le E\max_{\boldsymbol{\gamma}\in B_n(c)}\|\nabla^2\tilde{\ell}_n\left(\boldsymbol{\gamma};\mathbf{D}_{\text{disct}}\right)-\nabla^2\tilde{\ell}_n\left(\underline{\boldsymbol{\alpha}}_0;\mathbf{D}_{\text{disct}}\right)\|\\&\le E\max_{\boldsymbol{\gamma}\in B_n(c)}\|\nabla^2\tilde{\ell}\left(\boldsymbol{\gamma};\mathbf{D}_{1,\text{disct}}\right)-\nabla^2\tilde{\ell}\left(\underline{\boldsymbol{\alpha}}_0;\mathbf{D}_{1,\text{disct}}\right)\|.\end{aligned}$$

The function $\nabla^2\tilde{\ell}\left(\underline{\boldsymbol{\alpha}};\mathbf{D}_{1,\text{disct}}\right)$ is continuous and $B_n(c)$ shrinks to $\underline{\boldsymbol{\alpha}}_0$, so

$$\max_{\boldsymbol{\gamma}\in B_n(c)}\|\nabla^2\tilde{\ell}\left(\boldsymbol{\gamma};\mathbf{D}_{1,\text{disct}}\right)-\nabla^2\tilde{\ell}\left(\underline{\boldsymbol{\alpha}}_0;\mathbf{D}_{1,\text{disct}}\right)\|\to 0$$

as $n\to\infty$ for any given observed data $\mathbf{D}_{1,\text{disct}}$. According to Condition (B2), the covariate process $\{\boldsymbol{Z}(k\Delta t):k\in\mathbb{N}\}$ is bounded, and thus $\max_{\boldsymbol{\gamma}\in B_n(c)}\|\nabla^2\tilde{\ell}\left(\boldsymbol{\gamma};\mathbf{D}_{1,\text{disct}}\right)-\nabla^2\tilde{\ell}\left(\underline{\boldsymbol{\alpha}}_0;\mathbf{D}_{1,\text{disct}}\right)\|$ is bounded above by a constant. By the dominated convergence theorem, we have

$$\begin{aligned}&E\|\nabla^2\tilde{\ell}_n\left(\boldsymbol{\gamma}^*;\mathbf{D}_{\text{disct}}\right)-\nabla^2\tilde{\ell}_n\left(\underline{\boldsymbol{\alpha}}_0;\mathbf{D}_{\text{disct}}\right)\|\\\le&E\max_{\boldsymbol{\gamma}\in B_n(c)}\|\nabla^2\tilde{\ell}\left(\boldsymbol{\gamma};\mathbf{D}_{1,\text{disct}}\right)-\nabla^2\tilde{\ell}\left(\underline{\boldsymbol{\alpha}}_0;\mathbf{D}_{1,\text{disct}}\right)\|\to 0,\end{aligned}$$

which implies $\nabla^2\tilde{\ell}_n\left(\boldsymbol{\gamma}^*;\mathbf{D}_{\text{disct}}\right)-\nabla^2\tilde{\ell}_n\left(\underline{\boldsymbol{\alpha}}_0;\mathbf{D}_{\text{disct}}\right)=o_P(1)$ (by Markov's inequality). In addition, by WLLN, we have $\nabla^2\tilde{\ell}_n\left(\underline{\boldsymbol{\alpha}}_0;\mathbf{D}_{\text{disct}}\right)\xrightarrow{p}-Q(\underline{\boldsymbol{\alpha}}_0)$ as $n\to\infty$. Then the Taylor expansion in (46) can be rewritten as

$$n\tilde{\ell}_n\left(\boldsymbol{\gamma};\mathbf{D}_{\text{disct}}\right)-n\tilde{\ell}_n\left(\underline{\boldsymbol{\alpha}}_0;\mathbf{D}_{\text{disct}}\right)=c\boldsymbol{\lambda}^\top[nQ(\underline{\boldsymbol{\alpha}}_0)]^{-1/2}n\nabla\tilde{\ell}_n\left(\underline{\boldsymbol{\alpha}}_0;\mathbf{D}_{\text{disct}}\right)-\frac{c^2}{2}+o_P(1).\quad(47)$$

Note that $\max_{\{\boldsymbol{\lambda}:\|\boldsymbol{\lambda}\|=1\}}\boldsymbol{\lambda}^\top[nQ(\underline{\boldsymbol{\alpha}}_0)]^{-1/2}n\nabla\tilde{\ell}_n\left(\underline{\boldsymbol{\alpha}}_0;\mathbf{D}_{\text{disct}}\right)\le\|[nQ(\underline{\boldsymbol{\alpha}}_0)]^{-1/2}n\nabla\tilde{\ell}_n\left(\underline{\boldsymbol{\alpha}}_0;\mathbf{D}_{\text{disct}}\right)\|$ for sufficiently large n. Therefore, we have

$$\begin{aligned}&P\Big(\boldsymbol{\lambda}^\top[nQ(\underline{\boldsymbol{\alpha}}_0)]^{-1/2}n\nabla\tilde{\ell}_n\left(\underline{\boldsymbol{\alpha}}_0;\mathbf{D}_{\text{disct}}\right)<c/4\Big)\\\ge&P\Big(\max_{\{\boldsymbol{\lambda}:\|\boldsymbol{\lambda}\|=1\}}\boldsymbol{\lambda}^\top[nQ(\underline{\boldsymbol{\alpha}}_0)]^{-1/2}n\nabla\tilde{\ell}_n\left(\underline{\boldsymbol{\alpha}}_0;\mathbf{D}_{\text{disct}}\right)<c/4\Big)\\\ge&P\Big(\|[nQ(\underline{\boldsymbol{\alpha}}_0)]^{-1/2}n\nabla\tilde{\ell}_n\left(\underline{\boldsymbol{\alpha}}_0;\mathbf{D}_{\text{disct}}\right)\|<c/4\Big)\\\ge&1-(4/c)^2E\|[nQ(\underline{\boldsymbol{\alpha}}_0)]^{-1/2}n\nabla\tilde{\ell}_n\left(\underline{\boldsymbol{\alpha}}_0;\mathbf{D}_{\text{disct}}\right)\|^2=1-(4/c)^2C_0\ge 1-\epsilon.\end{aligned}\quad(48)$$

In (48), $C_0\triangleq E\|[nQ(\underline{\boldsymbol{\alpha}}_0)]^{-1/2}n\nabla\tilde{\ell}_n\left(\underline{\boldsymbol{\alpha}}_0;\mathbf{D}_{\text{disct}}\right)\|^2$ is a constant, and the last inequality holds by choosing a large c. Combining (47) and (48) yields (45). This completes the proof of weak consistency of $\hat{\underline{\boldsymbol{\alpha}}}_n$. Similarly, the weak consistency of $\hat{\boldsymbol{\beta}}_n$ can also be established.

Then, we show the weak convergence of $\hat{\underline{\boldsymbol{\alpha}}}_n$. Expand $\nabla\tilde{\ell}_n\left(\hat{\underline{\boldsymbol{\alpha}}}_n;\mathbf{D}_{\text{disct}}\right)$ at $\underline{\boldsymbol{\alpha}}_0$ gives

$$\nabla\tilde{\ell}_n\left(\hat{\underline{\boldsymbol{\alpha}}}_n;\mathbf{D}_{\text{disct}}\right)=\nabla\tilde{\ell}_n\left(\underline{\boldsymbol{\alpha}}_0;\mathbf{D}_{\text{disct}}\right)+\nabla^2\tilde{\ell}_n\left(\underline{\boldsymbol{\alpha}}_n^*;\mathbf{D}_{\text{disct}}\right)(\hat{\underline{\boldsymbol{\alpha}}}_n-\underline{\boldsymbol{\alpha}}_0),$$

where $\underline{\boldsymbol{\alpha}}_n^*$ lies between $\hat{\underline{\boldsymbol{\alpha}}}_n$ and $\underline{\boldsymbol{\alpha}}_0$. In addition, it is known $\nabla\tilde{\ell}_n\left(\hat{\underline{\boldsymbol{\alpha}}}_n;\mathbf{D}_{\text{disct}}\right)=0$. Hence we have

$$\sqrt{n}(\hat{\underline{\boldsymbol{\alpha}}}_n-\underline{\boldsymbol{\alpha}}_0)=-[\nabla^2\tilde{\ell}_n\left(\underline{\boldsymbol{\alpha}}_n^*;\mathbf{D}_{\text{disct}}\right)]^{-1}\sqrt{n}\nabla\tilde{\ell}_n\left(\underline{\boldsymbol{\alpha}}_0;\mathbf{D}_{\text{disct}}\right).$$

It is straightforward to check $E[\nabla\tilde{\ell}\left(\underline{\boldsymbol{\alpha}}_0;\mathbf{D}_{i,\text{disct}}\right)]=\mathbf{0}$ using the law of iterated expectation and the definition of the forward intensity. Therefore, by Lindeberg-Levy CLT, we have

$$\sqrt{n}\nabla\tilde{\ell}_n\left(\underline{\boldsymbol{\alpha}}_0;\mathbf{D}_{\text{disct}}\right)\rightsquigarrow N\big(\mathbf{0},\text{Var}[\nabla\tilde{\ell}\left(\underline{\boldsymbol{\alpha}}_0;\mathbf{D}_{1,\text{disct}}\right)]\big).$$

Then it suffices to check the convergence of $\nabla^2\tilde{\ell}_n\left(\underline{\boldsymbol{\alpha}}_n^*;\mathbf{D}_{\text{disct}}\right)$. Note that

$$E\|\nabla^2\tilde{\ell}_n\left(\underline{\boldsymbol{\alpha}}_n^*;\mathbf{D}_{\text{disct}}\right)-\nabla^2\tilde{\ell}_n\left(\underline{\boldsymbol{\alpha}}_0;\mathbf{D}_{\text{disct}}\right)\|\leq E\|\nabla^2\tilde{\ell}\left(\underline{\boldsymbol{\alpha}}_n^*;\mathbf{D}_{1,\text{disct}}\right)-\nabla^2\tilde{\ell}\left(\underline{\boldsymbol{\alpha}}_0;\mathbf{D}_{1,\text{disct}}\right)\|.$$

Because $\hat{\underline{\boldsymbol{\alpha}}}_n$ is consistent and $\underline{\boldsymbol{\alpha}}_n^*$ lies between $\hat{\underline{\boldsymbol{\alpha}}}_n$ and $\underline{\boldsymbol{\alpha}}_0$, $\|\nabla^2\tilde{\ell}\left(\underline{\boldsymbol{\alpha}}_n^*;\mathbf{D}_{1,\text{disct}}\right)-\nabla^2\tilde{\ell}\left(\underline{\boldsymbol{\alpha}}_0;\mathbf{D}_{1,\text{disct}}\right)\|$ converges to zero for any given observed data $\mathbf{D}_{1,\text{disct}}$. In addition, $\|\nabla^2\tilde{\ell}\left(\underline{\boldsymbol{\alpha}}_n^*;\mathbf{D}_{1,\text{disct}}\right)-\nabla^2\tilde{\ell}\left(\underline{\boldsymbol{\alpha}}_0;\mathbf{D}_{1,\text{disct}}\right)\|$ is bounded above by a constant because the covariate process $\{\boldsymbol{Z}(k\Delta t):k\in\mathbb{N}\}$ is bounded. By the dominated convergence theorem, we have

$$E\|\nabla^2\tilde{\ell}_n\left(\underline{\boldsymbol{\alpha}}_n^*;\mathbf{D}_{\text{disct}}\right)-\nabla^2\tilde{\ell}_n\left(\underline{\boldsymbol{\alpha}}_0;\mathbf{D}_{\text{disct}}\right)\|\leq E\|\nabla^2\tilde{\ell}\left(\underline{\boldsymbol{\alpha}}_n^*;\mathbf{D}_{1,\text{disct}}\right)-\nabla^2\tilde{\ell}\left(\underline{\boldsymbol{\alpha}}_0;\mathbf{D}_{1,\text{disct}}\right)\|\to 0,$$

which further implies $\nabla^2\tilde{\ell}_n\left(\underline{\boldsymbol{\alpha}}_n^*;\mathbf{D}_{\text{disct}}\right)-\nabla^2\tilde{\ell}_n\left(\underline{\boldsymbol{\alpha}}_0;\mathbf{D}_{\text{disct}}\right)=o_P(1)$ by Markov's inequality. Besides, it is easy to see $\nabla^2\tilde{\ell}_n\left(\underline{\boldsymbol{\alpha}}_0;\mathbf{D}_{\text{disct}}\right)\xrightarrow{p}E[\nabla^2\tilde{\ell}\left(\underline{\boldsymbol{\alpha}}_0;\mathbf{D}_{1,\text{disct}}\right)]$ by WLLN. Hence we have $\nabla^2\tilde{\ell}_n\left(\underline{\boldsymbol{\alpha}}_n^*;\mathbf{D}_{\text{disct}}\right)\xrightarrow{p}E[\nabla^2\tilde{\ell}\left(\underline{\boldsymbol{\alpha}}_0;\mathbf{D}_{1,\text{disct}}\right)]$. By Slutsky's lemma, we have $\sqrt{n}(\hat{\underline{\boldsymbol{\alpha}}}_n-\underline{\boldsymbol{\alpha}}_0)\rightsquigarrow N\big(\mathbf{0},\boldsymbol{\Sigma}_1\big)$ with

$$\boldsymbol{\Sigma}_1=\big[E[\nabla^2\tilde{\ell}\left(\underline{\boldsymbol{\alpha}}_0;\mathbf{D}_{1,\text{disct}}\right)]\big]^{-1}\text{Var}[\nabla\tilde{\ell}\left(\underline{\boldsymbol{\alpha}}_0;\mathbf{D}_{1,\text{disct}}\right)]\big[E[\nabla^2\tilde{\ell}\left(\underline{\boldsymbol{\alpha}}_0;\mathbf{D}_{1,\text{disct}}\right)]\big]^{-1}.$$

Besides, it is easy to see $E[\nabla^2\tilde{\ell}\left(\underline{\boldsymbol{\alpha}}_0;\mathbf{D}_{1,\text{disct}}\right)]=E[\nabla^2\tilde{\ell}_n\left(\underline{\boldsymbol{\alpha}}_0;\mathbf{D}_{\text{disct}}\right)]$ and $\text{Var}[\nabla\tilde{\ell}\left(\underline{\boldsymbol{\alpha}}_0;\mathbf{D}_{1,\text{disct}}\right)]=\text{Var}[\sqrt{n}\nabla\tilde{\ell}_n\left(\underline{\boldsymbol{\alpha}}_0;\mathbf{D}_{\text{disct}}\right)]$, so the asymptotic variance $\boldsymbol{\Sigma}_1$ can also be written as

$$\boldsymbol{\Sigma}_1=(E[\nabla^2\tilde{\ell}_n(\underline{\boldsymbol{\alpha}}_0)])^{-1}\text{Var}[\sqrt{n}\nabla\tilde{\ell}_n(\underline{\boldsymbol{\alpha}}_0)](E[\nabla^2\tilde{\ell}_n(\underline{\boldsymbol{\alpha}}_0)])^{-1}.$$

This completes the proof of the asymptotic normality of $\hat{\underline{\boldsymbol{\alpha}}}_n$. The asymptotic normality of $(\hat{\underline{\boldsymbol{\alpha}}}_n,\hat{\underline{\boldsymbol{\beta}}}_n)$ can be similarly established by following the same arguments as above, except that we need to replace $\nabla\tilde{\ell}_n\left(\underline{\boldsymbol{\alpha}};\mathbf{D}_{\text{disct}}\right)$ by $\nabla\check{\ell}_n(\underline{\boldsymbol{\alpha}},\underline{\boldsymbol{\beta}};\mathbf{D}_{\text{disct}})$, where

$$\check{\ell}_n(\underline{\boldsymbol{\alpha}},\underline{\boldsymbol{\beta}};\mathbf{D}_{\text{disct}})=\tilde{\ell}_n\left(\underline{\boldsymbol{\alpha}};\mathbf{D}_{\text{disct}}\right)+\tilde{\ell}_n\left(\underline{\boldsymbol{\beta}};\mathbf{D}_{\text{disct}}\right).$$

To avoid redundancy, we omit the details and directly give the asymptotic distribution of $(\hat{\underline{\boldsymbol{\alpha}}}_n,\hat{\underline{\boldsymbol{\beta}}}_n)$. In particular, $\sqrt{n}\left(\begin{bmatrix}\hat{\underline{\boldsymbol{\alpha}}}_n\\ \hat{\underline{\boldsymbol{\beta}}}_n\end{bmatrix}-\begin{bmatrix}\underline{\boldsymbol{\alpha}}_0\\ \underline{\boldsymbol{\beta}}_0\end{bmatrix}\right)$ converges to a normal distribution with mean zero and covariance matrix

$$(E[\nabla^2\check{\ell}_n(\underline{\boldsymbol{\alpha}}_0,\underline{\boldsymbol{\beta}}_0)])^{-1}\text{Var}[\sqrt{n}\nabla\check{\ell}_n(\underline{\boldsymbol{\alpha}}_0,\underline{\boldsymbol{\beta}}_0)](E[\nabla^2\check{\ell}_n(\underline{\boldsymbol{\alpha}}_0,\underline{\boldsymbol{\beta}}_0)])^{-1}.$$

It is straightforward to check

$$E[\nabla^2\check{\ell}_n(\underline{\boldsymbol{\alpha}}_0,\underline{\boldsymbol{\beta}}_0)]=\begin{bmatrix}E[\nabla^2\tilde{\ell}_n(\underline{\boldsymbol{\alpha}}_0)] & \mathbf{0}\\ \mathbf{0} & E[\nabla^2\tilde{\ell}_n(\underline{\boldsymbol{\beta}}_0)]\end{bmatrix}.$$

In addition, by Lemma 11, we have

$$\mathrm{Var}[\sqrt{n}\nabla\check{\ell}_n(\underline{\boldsymbol{\alpha}}_0,\underline{\boldsymbol{\beta}}_0)] = \begin{bmatrix}\mathrm{Var}[\sqrt{n}\nabla\tilde{\ell}_n(\underline{\boldsymbol{\alpha}}_0)] & \mathbf{0}\\ \mathbf{0} & \mathrm{Var}[\sqrt{n}\nabla\tilde{\ell}_n(\underline{\boldsymbol{\beta}}_0)]\end{bmatrix}.$$

Thus the covariance matrix can be written as $\begin{bmatrix}\boldsymbol{\Sigma}_1 & \mathbf{0}\\ \mathbf{0} & \boldsymbol{\Sigma}_2\end{bmatrix}$ with

$$\begin{aligned}\boldsymbol{\Sigma}_1 &= (E[\nabla^2\tilde{\ell}_n(\underline{\boldsymbol{\alpha}}_0)])^{-1}\mathrm{Var}[\sqrt{n}\nabla\tilde{\ell}_n(\underline{\boldsymbol{\alpha}}_0)](E[\nabla^2\tilde{\ell}_n(\underline{\boldsymbol{\alpha}}_0)])^{-1},\\ \boldsymbol{\Sigma}_2 &= (E[\nabla^2\tilde{\ell}_n(\underline{\boldsymbol{\beta}}_0)])^{-1}\mathrm{Var}[\sqrt{n}\nabla\tilde{\ell}_n(\underline{\boldsymbol{\beta}}_0)](E[\nabla^2\tilde{\ell}_n(\underline{\boldsymbol{\beta}}_0)])^{-1}.\end{aligned}$$

This completes the proof of the asymptotic normality.

Afterward, we develop the estimator of the asymptotic variance $\boldsymbol{\Sigma}_1$ of $\hat{\underline{\boldsymbol{\alpha}}}_n$ and show its consistency. Denote $\boldsymbol{Z}_i^{(k)}(t) = (\mathbf{0}_p^\top,\dots,\mathbf{0}_p^\top,\boldsymbol{Z}_i(t)^\top,\mathbf{0}_p^\top,\dots,\mathbf{0}_p^\top)^\top$, where $\boldsymbol{Z}_i(t)$ locates at the kth coordinate, $k=1,2,\dots,\kappa$. For $i\in\{1,2,\dots,n\}$ and $j,k\in\mathbb{N}$, we write $\lambda_{T_i,t_j}(t_k;\underline{\boldsymbol{\alpha}}) = \exp\{\boldsymbol{\alpha}(t_k)^\top\boldsymbol{Z}_i(t_j)\}$ to highlight that $\lambda_{T_i,t_j}(t_k)$ depends on the parameter $\underline{\boldsymbol{\alpha}}$. For $0\le k_1,k_2<\kappa$ and $u,s\in\{t_0,t_1,\dots\}$, denote

$$\begin{aligned}&\mathrm{Cov}^{(i)}_{u,s,k_1,k_2}(\underline{\boldsymbol{\alpha}})\\ &= \begin{cases}(\Delta t)^2\lambda_{T_i,u}(t_{k_1};\underline{\boldsymbol{\alpha}})\Big[1-(1-\exp\{-\lambda_{T_i,u+t_{k_1}}(0;\underline{\boldsymbol{\alpha}})\Delta t\})\times & \\ \dfrac{\exp\{-\lambda_{T_i,u}(t_{k_1};\underline{\boldsymbol{\alpha}})\Delta t\}+\exp\{-\lambda_{T_i,s}(t_{k_2};\underline{\boldsymbol{\alpha}})\Delta t\}-1}{(1-\exp\{-\lambda_{T_i,u}(t_{k_1};\underline{\boldsymbol{\alpha}})\Delta t\})(1-\exp\{-\lambda_{T_i,s}(t_{k_2};\underline{\boldsymbol{\alpha}})\Delta t\})}\Big]\lambda_{T_i,s}(t_{k_2};\underline{\boldsymbol{\alpha}}), & \text{if } u+t_{k_1}=s+t_{k_2};\\ (\Delta t)^2\lambda_{T_i,u}(t_{k_1};\underline{\boldsymbol{\alpha}})\lambda_{T_i,s}(t_{k_2};\underline{\boldsymbol{\alpha}})\Big[1-\dfrac{1-\exp\{-\lambda_{T_i,s+t_{k_2}}(0;\underline{\boldsymbol{\alpha}})\Delta t\}}{1-\exp\{-\lambda_{T_i,s}(t_{k_2};\underline{\boldsymbol{\alpha}})\Delta t\}}\Big], & \text{if } u+t_{k_1}<s+t_{k_2};\\ (\Delta t)^2\lambda_{T_i,u}(t_{k_1};\underline{\boldsymbol{\alpha}})\lambda_{T_i,s}(t_{k_2};\underline{\boldsymbol{\alpha}})\Big[1-\dfrac{1-\exp\{-\lambda_{T_i,u+t_{k_1}}(0;\underline{\boldsymbol{\alpha}})\Delta t\}}{1-\exp\{-\lambda_{T_i,u}(t_{k_1};\underline{\boldsymbol{\alpha}})\Delta t\}}\Big], & \text{if } u+t_{k_1}>s+t_{k_2}.\end{cases}\end{aligned} \tag{49}$$

Then the estimator $\hat{\boldsymbol{\Sigma}}_{1,n}$ of $\boldsymbol{\Sigma}_1$ is

$$\hat{\boldsymbol{\Sigma}}_{1,n} = \boldsymbol{\mathcal{W}}_n^{-1}(\hat{\underline{\boldsymbol{\alpha}}}_n)\tilde{\boldsymbol{\varrho}}_n(\hat{\underline{\boldsymbol{\alpha}}}_n)\boldsymbol{\mathcal{W}}_n^{-1}(\hat{\underline{\boldsymbol{\alpha}}}_n), \tag{50}$$

where

$$\begin{aligned}\boldsymbol{\mathcal{W}}_n(\underline{\boldsymbol{\alpha}}) = \frac{1}{n}\sum_{i=1}^{n}\sum_{k=0}^{\kappa-1}\sum_{h=1}^{M_i-k-1}\boldsymbol{Z}_i^{(k)}(t_h)\boldsymbol{Z}_i^{(k)}(t_h)^\top\Big[&\lambda_{T_i,t_h}(t_k;\underline{\boldsymbol{\alpha}})\Delta t-(1-\exp\{-\lambda_{T_i,t_h+t_k}(0;\underline{\boldsymbol{\alpha}})\Delta t\})\\ &\times\frac{\lambda_{T_i,t_h}(t_k;\underline{\boldsymbol{\alpha}})\Delta t\,(1-\exp\{-\lambda_{T_i,t_h}(t_k;\underline{\boldsymbol{\alpha}})\Delta t\})-(\Delta t)^2\lambda^2_{T_i,t_h}(t_k;\underline{\boldsymbol{\alpha}})\exp\{-\lambda_{T_i,t_h}(t_k;\underline{\boldsymbol{\alpha}})\Delta t\}}{(1-\exp\{-\lambda_{T_i,t_h}(t_k;\underline{\boldsymbol{\alpha}})\Delta t\})^2}\Big]\end{aligned}$$

and

$$\tilde{\boldsymbol{\varrho}}_n(\underline{\boldsymbol{\alpha}}) = \frac{1}{n}\sum_{i=1}^{n}\sum_{k_1=0}^{\kappa-1}\sum_{k_2=0}^{\kappa-1}\sum_{h_1=1}^{M_i-k_1-1}\sum_{h_2=1}^{M_i-k_2-1}\boldsymbol{Z}_i^{(k_1)}(t_{h_1})\boldsymbol{Z}_i^{(k_2)}(t_{h_2})^\top\mathrm{Cov}^{(i)}_{t_{h_1},t_{h_2},k_1,k_2}(\underline{\boldsymbol{\alpha}}).$$

Define $\tilde{\boldsymbol{\varrho}}(\underline{\boldsymbol{\alpha}}) = \mathrm{Var}[\nabla\tilde{\ell}(\underline{\boldsymbol{\alpha}}; \mathbf{D}_{1,\mathrm{disct}})]$. In the following, we prove $\tilde{\boldsymbol{\varrho}}_n(\underline{\hat{\boldsymbol{\alpha}}}_n) \xrightarrow{p} \tilde{\boldsymbol{\varrho}}(\underline{\boldsymbol{\alpha}}_0)$ using the Glivenko-Cantelli theorem. To this end, for any $1 \le j_1, j_2 \le p$, define

$$U_{j_1,j_2}(\boldsymbol{Z}_i, X_i, \underline{\boldsymbol{\alpha}}) = \sum_{k_1=0}^{\kappa-1}\sum_{k_2=0}^{\kappa-1}\sum_{h_1=1}^{\infty}\sum_{h_2=1}^{\infty} Z_{i,j_1}(t_{h_1})Z_{i,j_2}(t_{h_2})\mathrm{Cov}^{(i)}_{t_{h_1},t_{h_2},k_1,k_2}(\underline{\boldsymbol{\alpha}})\times I\{X_i > t_{h_1} + k_1\Delta t\}I\{X_i > t_{h_2} + k_2\Delta t\},$$

where $Z_{i,j}(t)$ is the jth coordinate of the p-dimensional vector $\boldsymbol{Z}_i(t)$. Using this definition, the (j_1, j_2)th element of $\tilde{\boldsymbol{\varrho}}_n(\underline{\boldsymbol{\alpha}})$ can be written as $\frac{1}{n}\sum_{i=1}^n U_{j_1,j_2}(\boldsymbol{Z}_i, X_i, \underline{\boldsymbol{\alpha}})$. In order to apply the G-C theorem, we need show: (i) the parameter space $\mathcal{F}_0$ of $\underline{\boldsymbol{\alpha}}$ is compact; (ii) the function $\underline{\boldsymbol{\alpha}} \mapsto U_{j_1,j_2}(\boldsymbol{Z}_i, X_i, \underline{\boldsymbol{\alpha}})$ is continuous for any given X_i and $\boldsymbol{Z}_i$; (iii) $U_{j_1,j_2}(\boldsymbol{Z}_i, X_i, \boldsymbol{\alpha})$ is upper bounded by an integrable function. The compactness of $\mathcal{F}_0$ follows from Condition (B1). Based on expression (49), it is obvious that $\underline{\boldsymbol{\alpha}} \mapsto \mathrm{Cov}^{(i)}_{t_{h_1},t_{h_2},k_1,k_2}(\underline{\boldsymbol{\alpha}})$ is continuous. Hence the continuity of $\underline{\boldsymbol{\alpha}} \mapsto U_{j_1,j_2}(\boldsymbol{Z}_i, X_i, \underline{\boldsymbol{\alpha}})$ holds for any given X_i and $\boldsymbol{Z}_i$. To check the third condition, note that

$$\begin{aligned}
&|U_{j_1,j_2}(\boldsymbol{Z}_i, X_i, \underline{\boldsymbol{\alpha}})| \\
\le& \exp\{4pK_3K_4\}\sum_{k_1=0}^{\kappa-1}\sum_{k_2=0}^{\kappa-1}\sum_{h_1=1}^{\infty}\sum_{h_2=1}^{\infty}|Z_{i,j_1}(t_1)Z_{i,j_2}(t_2)|I\{X_i > t_{h_1} + k_1\Delta t\}I\{X_i > t_{h_2} + k_2\Delta t\} \\
\le& K_4^2\exp\{4pK_3K_4\}\sum_{k_1=0}^{\kappa-1}\sum_{k_2=0}^{\kappa-1}\sum_{h_1=1}^{\infty}\sum_{h_2=1}^{\infty} I\{X_i > t_{h_1} + k_1\Delta t\}I\{X_i > t_{h_2} + k_2\Delta t\} \\
\le& K_4^2\kappa^2\exp\{4pK_3K_4\}\big[\sum_{h=1}^{\infty} I\{X_i > t_h\}\big]^2 \le K_4^2\kappa^2\exp\{4pK_3K_4\}X_i^2,
\end{aligned}$$

where K_3 and K_4 are the constant bounding $\mathcal{F}_0$ and $\{\boldsymbol{Z}(k\Delta t); t \in \mathbb{N}\}$, respectively. In the above display, the first inequality holds because $|\mathrm{Cov}^{(i)}_{t_{h_1},t_{h_2},k_1,k_2}(\underline{\boldsymbol{\alpha}})| \le \exp\{4pK_3K_4\}$. Since X is assumed to be square integrable, $U_{j_1,j_2}(\boldsymbol{Z}_i, X_i, \boldsymbol{\alpha})$ is upper bounded by an integrable function. Therefore, we can apply the G-C theorem and this theorem tells us that (i) the function $\underline{\boldsymbol{\alpha}} \mapsto E[U_{j_1,j_2}(\boldsymbol{Z}_i, X_i, \underline{\boldsymbol{\alpha}})]$ is continuous; (ii) the following result holds:

$$\sup_{\boldsymbol{\alpha}\in\mathcal{F}_0}\Big|\frac{1}{n}\sum_{i=1}^n U_{j_1,j_2}(\boldsymbol{Z}_i, X_i, \underline{\boldsymbol{\alpha}}) - E[U_{j_1,j_2}(\boldsymbol{Z}_i, X_i, \underline{\boldsymbol{\alpha}})]\Big| \xrightarrow{a.s.} 0. \tag{51}$$

Therefore, for any $\epsilon > 0$,

$$\begin{aligned}
&P\Big(\Big|\frac{1}{n}\sum_{i=1}^{n} U_{j_1,j_2}(\boldsymbol{Z}_i, X_i, \hat{\underline{\boldsymbol{\alpha}}}_n) - E[U_{j_1,j_2}(\boldsymbol{Z}_i, X_i, \underline{\boldsymbol{\alpha}}_0)]\Big| > \epsilon\Big) \\
\leq & P\Big(\Big|\frac{1}{n}\sum_{i=1}^{n} U_{j_1,j_2}(\boldsymbol{Z}_i, X_i, \hat{\underline{\boldsymbol{\alpha}}}_n) - E[U_{j_1,j_2}(\boldsymbol{Z}_i, X_i, \hat{\underline{\boldsymbol{\alpha}}}_n)]\Big| > \frac{\epsilon}{2}\Big) \\
&+ P\Big(\Big|E[U_{j_1,j_2}(\boldsymbol{Z}_i, X_i, \hat{\underline{\boldsymbol{\alpha}}}_n)] - E[U_{j_1,j_2}(\boldsymbol{Z}_i, X_i, \underline{\boldsymbol{\alpha}}_0)]\Big| > \frac{\epsilon}{2}\Big) \qquad (52) \\
\leq & P\Big(\sup_{\underline{\boldsymbol{\alpha}} \in \mathcal{F}_0} \Big|\frac{1}{n}\sum_{i=1}^{n} U_{j_1,j_2}(\boldsymbol{Z}_i, X_i, \underline{\boldsymbol{\alpha}}) - E[U_{j_1,j_2}(\boldsymbol{Z}_i, X_i, \underline{\boldsymbol{\alpha}})]\Big| > \frac{\epsilon}{2}\Big) \\
&+ P\Big(\Big|E[U_{j_1,j_2}(\boldsymbol{Z}_i, X_i, \hat{\underline{\boldsymbol{\alpha}}}_n)] - E[U_{j_1,j_2}(\boldsymbol{Z}_i, X_i, \underline{\boldsymbol{\alpha}}_0)]\Big| > \frac{\epsilon}{2}\Big).
\end{aligned}$$

The term in the second last row of (52) converges to zero because of (51). Recall $\hat{\underline{\boldsymbol{\alpha}}}_n \xrightarrow{p} \underline{\boldsymbol{\alpha}}_0$, and $\underline{\boldsymbol{\alpha}} \mapsto E[U_{j_1,j_2}(\boldsymbol{Z}_i, X_i, \underline{\boldsymbol{\alpha}})]$ is a continuous function. By the continuous mapping theorem, the term in the last row of (52) also converges to zero. Therefore, we have

$$\frac{1}{n}\sum_{i=1}^{n} U_{j_1,j_2}(\boldsymbol{Z}_i, X_i, \hat{\underline{\boldsymbol{\alpha}}}_n) \xrightarrow{p} E[U_{j_1,j_2}(\boldsymbol{Z}_i, X_i, \underline{\boldsymbol{\alpha}}_0)], \quad 1 \leq j_1, j_2 \leq p. \qquad (53)$$

Furthermore, using the law of iterated expectation and the definition of the forward intensity function, we have

$$E\big[U_{j_1,j_2}(\boldsymbol{Z}_i, X_i, \underline{\boldsymbol{\alpha}})\big] = \tilde{\varrho}_{i,j}(\underline{\boldsymbol{\alpha}}), \qquad (54)$$

where $\tilde{\varrho}_{i,j}(\underline{\boldsymbol{\alpha}})$ denotes the (i,j)th element of $\tilde{\boldsymbol{\varrho}}(\underline{\boldsymbol{\alpha}})$, $1 \leq j_1, j_2 \leq p$. Combining (53) and (54) yields

$$\frac{1}{n}\sum_{i=1}^{n} U_{j_1,j_2}(\boldsymbol{Z}_i, X_i, \hat{\underline{\boldsymbol{\alpha}}}_n) \xrightarrow{p} \tilde{\varrho}_{i,j}(\underline{\boldsymbol{\alpha}}_0), \quad 1 \leq j_1, j_2 \leq p.$$

In addition, recall that $\frac{1}{n}\sum_{i=1}^{n} U_{j_1,j_2}(\boldsymbol{Z}_i, X_i, \underline{\boldsymbol{\alpha}})$ is equal to the (j_1,j_2)th element of $\tilde{\boldsymbol{\varrho}}_n(\underline{\boldsymbol{\alpha}})$, $1 \leq j_1, j_2 \leq p$. As a result, every entry of $\tilde{\boldsymbol{\varrho}}_n(\hat{\underline{\boldsymbol{\alpha}}}_n)$ converges to the corresponding entry of $\tilde{\boldsymbol{\varrho}}(\underline{\boldsymbol{\alpha}}_0)$ in probability. Thus $\tilde{\boldsymbol{\varrho}}_n(\hat{\underline{\boldsymbol{\alpha}}}_n) \xrightarrow{p} \tilde{\boldsymbol{\varrho}}(\underline{\boldsymbol{\alpha}}_0)$. In the same way, we can show that $\boldsymbol{\mathcal{W}}_n(\hat{\underline{\boldsymbol{\alpha}}}_n) \xrightarrow{p} E[\nabla^2 \tilde{\ell}\,(\underline{\boldsymbol{\alpha}}_0; \mathbf{D}_{1,\text{disct}})]$. Then, by the continuous mapping theorem, $\hat{\boldsymbol{\Sigma}}_{1,n} = \boldsymbol{\mathcal{W}}_n^{-1}(\hat{\underline{\boldsymbol{\alpha}}}_n)\tilde{\boldsymbol{\varrho}}_n(\hat{\underline{\boldsymbol{\alpha}}}_n)\boldsymbol{\mathcal{W}}_n^{-1}(\hat{\underline{\boldsymbol{\alpha}}}_n)$ converges in probability to $\boldsymbol{\Sigma}_1$. The estimator $\hat{\boldsymbol{\Sigma}}_{2,n}$ of the asymptotic variance $\boldsymbol{\Sigma}_2$ of $\hat{\underline{\boldsymbol{\beta}}}_n$ and its consistency can be developed in a similar way, and thus the details are omitted.

E.7 Proof of Proposition 7

The asymptotic normality of $\hat{S}_{T,t}(\tau)$, $\hat{F}_t(\tau)$, and $\hat{\text{RMRL}}_t(\tau)$ follows from the asymptotic normality of $(\hat{\underline{\boldsymbol{\alpha}}}_n, \hat{\underline{\boldsymbol{\beta}}}_n)$ in Theorem 6 and the delta method. The consistency of $\hat{\sigma}_0^2$, $\hat{\eta}_0^2$, and $\hat{\varsigma}_0^2$ follows from the consistency of $(\hat{\underline{\boldsymbol{\alpha}}}_n, \hat{\underline{\boldsymbol{\beta}}}_n)$, $\hat{\boldsymbol{\Sigma}}_{1,n}$, $\hat{\boldsymbol{\Sigma}}_{2,n}$ in Theorem 6 and the continuous mapping theorem.

E.8 Proof of Proposition 8

The asymptotic normality of $(\hat{S}_{T,t}(t_1), \hat{S}_{T,t}(t_2), \ldots, \hat{S}_{T,t}(t_\kappa))$ and $(\hat{F}_t(t_1), \hat{F}_t(t_2), \ldots, \hat{F}_t(t_\kappa))$ follows from the asymptotic normality of $(\underline{\hat{\boldsymbol{\alpha}}}_n, \underline{\hat{\boldsymbol{\beta}}}_n)$ in Theorem 6 and the delta method. The consistency of the variance estimators follows from the consistency of $(\underline{\hat{\boldsymbol{\alpha}}}_n, \underline{\hat{\boldsymbol{\beta}}}_n)$, $\hat{\boldsymbol{\Sigma}}_{1,n}$, $\hat{\boldsymbol{\Sigma}}_{2,n}$ in Theorem 6 and the continuous mapping theorem.

References

A. Bansal and P. J. Heagerty. A tutorial on evaluating the time-varying discrimination accuracy of survival models used in dynamic decision making. *Medical Decision Making*, 38:904–916, 2018.

D. Beever and L. Swaby. An evaluation of risk-based monitoring in pragmatic trials in uk clinical trials units. *Trials*, 20:1–7, 2019.

B. Berghaus and A. Bucher. Weak convergence of a pseudo maximum likelihood estimator for the extremal index. *Annals of Statistics*, 46:2307–2335, 2018.

K. C. G. Chan and M.-C. Wang. Semiparametric modeling and estimation of the terminal behavior of recurrent marker processes before failure events. *Journal of the American Statistical Association*, 112:351–362, 2017.

C.-J. Chen. The instantaneous and forward default intensity of structural models. Technical report, University of Alberta, 2007.

J.-C. Duan, J. Sun, and T. Wang. Multiperiod corporate default prediction—a forward intensity approach. *Journal of Econometrics*, 170:191–209, 2012.

L. Elashoff, G. Li, and N. Li. *Joint Modeling of Longitudinal and Time-to-Event Data*. Chapman and Hall/CRC, 2016. ISBN 978-1439807828.

L. Ferrer, H. Putter, and C. Proust-Lima. Individual dynamic predictions using landmarking and joint modelling: Validation of estimators and robustness assessment. *Statistical Methods in Medical Research*, 28:3649–3666, 2018.

T. R. Fleming and D. P. Harrington. *Counting Process and Survival Analysis*. Wiely, 2005.

M. Fredette and J. F. Lawless. Finite-horizon prediction of recurrent events, with application to forecasts of warranty claims. *Technometrics*, 49:66–80, 2007.

Q. Gong and D. E. Schaubel. Estimating the average treatment effect on survival based on observational data and using partly conditional modeling. *Biometrics*, 73:134–144, 2017.

H. Haider, B. Hoehn, S. Davis, and R. Greiner. Effective ways to build and evaluate individual survival distributions. *Journal of Machine Learning Research*, 21(85):1–63, 2020.

B. He, Y. Liu, Y. Wu, G. Yin, and X. Zhao. Functional martingale residual process for high-dimensional cox regression with model averaging. *Journal of Machine Learning Research*, 21(207):1–37, 2020.

D. F. Heitjan, Z. Ge, and G.-s. Ying. Real-time prediction of clinical trial enrollment and event counts: A review. *Contemporary Clinical Trials*, 45:26–33, 2015.

B. Hu and B. Nan. Conditional distribution function estimation using neural networks for censored and uncensored data. *Journal of Machine Learning Research*, 24(223):1–26, 2023.

Xiao-Yang Li, Zhi-Sheng Ye, and Cheng Yong Tang. Estimating the inter-occurrence time distribution from superposed renewal processes. *Bernoulli*, 27(4):2804–2826, 2021.

M. Maziarz, P. Heagerty, T. Cai, and Y. Zheng. On longitudinal prediction with time-to-event outcome: comparison of modeling options. *Biometrics*, 73:83–93, 2017.

C.Robert Merton. On the pricing of corporate debt: the risk structure of interest rates. *The Journal of Finance*, 29:449–470, 1974.

P. A. Murtaugh, E. R. Dickson, G. M. Van Dam, M. Malinchoc, P. M. Grambsch, A. L. Langworthy, and C. H. Gips. Primary biliary cirrhosis: prediction of short-term survival based on repeated patient visits. *Hepatology (Baltimore, Md.)*, 20(1 Pt 1):126—134, July 1994.

D. Rizopoulos. *Joint Models for Longitudinal and Time-to-Event Data: With Applications in R*. Chapman and Hall/CRC, 2012.

D. Rizopoulos, L. A. Hatfield, B. P. Carlin, and J. J. M. Takkenberg. Combining dynamic predictions from joint models for longitudinal and time-to-event data using bayesian model averaging. *Journal of the American Statistical Association*, 109:1385–1397, 2014.

J. Rousseau and B. Szabo. Asymptotic frequentist coverage properties of Bayesian credible sets for sieve priors. *Annals of Statistics*, 48:2155–2179, 2020.

J. Shao. *Mathematical Statistics*. Springer, 2003.

W. Tang, K. He, G. Xu, and J. Zhu. Survival analysis via ordinary differential equations. *Journal of the American Statistical Association*, pages 1–16, apr 2022a. doi: 10.1080/01621459.2022.2051519.

W. Tang, J. Ma, Q. Mei, and J. Zhu. Soden: A scalable continuous-time survival model through ordinary differential equation networks. *Journal of Machine Learning Research*, 23(34):1–29, 2022b.

J. M. G. Taylor, Y. Park, D. P. Ankerst, C. Proust-Lima, S. Williams, L. Kestin, K. Bae, T. Pickles, and H. Sandler. Real-time individual predictions of prostate cancer recurrence using joint models. *Biometrics*, 69:206–213, 2013.

Lu Tian, Hua Jin, Hajime Uno, Ying Lu, Bo Huang, Keaven M Anderson, and LJ Wei. On the empirical choice of the time window for restricted mean survival time. *Biometrics*, 76(4):1157–1166, 2020.

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

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

H. C. van Houwelingen. Dynamic prediction by landmarking in event history analysis. *Scandinavian Journal of Statistics*, 34:70–85, 2007.

H. C. van Houwelingen and H. Putter. *Dynamic Prediction in Clinical Survival Analysis*. CRC Press, 2011.

C. Varin, N. Reid, and D. Firth. An overview of composite likelihood methods. *Statistica Sinica*, 21:5–42, 2011.

J. Wang, S. Luo, and L. Li. Dynamic prediction for multiple repeated measures and event time data: An application to parkinson's disease. *The Annals of Applied Statistics*, 11: 1787–1809, 2017.

Y. Wang and J. M. G. Taylor. Jointly modeling longitudinal and event time data with application to acquired immunodeficiency syndrome. *Journal of the American Statistical Association*, 96:895–905, 2001.

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

L. Wu, L. Li, and R. Li. Dynamic prediction of competing risk events using landmark sub-distribution hazard model with multiple longitudinal biomarkers. *Statistical Methods in Medical Research*, 29(11):3179–3191, 2020.

M. S. Wulfsohn and A. A. Tsiatis. A joint model for survival and longitudinal data measured with error. *Biometrics*, 53:330–339, 1997.

P. Ye, X. Zhao, L. Sun, and W. Xu. A semiparametric additive rates model for multivariate recurrent events with missing event categories. *Computational Statistics & Data Analysis*, 89:39–50, 2015.

M. Yuan, C. Y. Tang, Y. Hong, and J. Yang. Disentangling and assessing uncertainties in multiperiod corporate default risk predictions. *Annals of Applied Statistics*, 12:2587–2617, 2018.

D. Zeng and D. Lin. Maximum likelihood estimation in semiparametric regression models with censored data. *Journal of the Royal Statistical Society: Series B*, 69(4):507–564, 2007.

X. Zhao, Y. Wu, and G. Yin. Sieve maximum likelihood estimation for a general class of accelerated hazards models with bundled parameters. *Bernoulli*, 23(4B):3385–3411, 2017.

Y. Zheng and P. J. Heagerty. Partly conditional survival models for longitudinal data. *Biometrics*, 61:379–391, 2005.

Y. Zhu, L. Li, and X. Huang. Landmark linear transformation model for dynamic prediction with application to a longitudinal cohort study of chronic disease. *Journal of the Royal Statistical Society: Series C*, 68:771–791, 2019.