Statistical Analysis of Complex Data in Survival and Event History Analysis

Hok Kan Ling

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
under the Executive Committee
of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2020

Abstract

Statistical Analysis of Complex Data in Survival and Event History Analysis

Hok Kan Ling

This thesis studies two aspects of the statistical analysis of complex data in survival and event history analysis. After a short introduction to survival and event history analysis in Chapter 1, we proposed a multivariate proportional intensity factor model for multivariate counting processes in Chapter 2. In an exploratory analysis on process data, a large number of possibly time-varying covariates maybe included. These covariates along with the high-dimensional counting processes often exhibit a low-dimensional structure that has meaningful interpretation. We explore such structure through specifying random coefficients in a low dimensional space through a factor model. For the estimation of the resulting model, we establish the asymptotic theory of the nonparametric maximum likelihood estimator (NPMLE). In particular, the NPMLE is consistent, asymptotically normal and asymptotically efficient with covariance matrix that can be consistently estimated by the inverse information matrix or the profile likelihood method under some suitable regularity conditions. Furthermore, to obtain a parsimonious model and to improve interpretation of parameters therein, variable selection and estimation for both fixed and random effects are developed by penalized likelihood. We illustrate the method using simulation studies as well as a real data application from The Programme for the International Assessment of Adult Competencies (PIAAC). Chapter 3 concerns rare events and sparse covariates in event history analysis. In large-scale longitudinal observational databases, the majority of subjects may not experience a particular event of interest. Furthermore, the associated covariate processes could also be zero for most of the subjects at any time. We formulate such setting of rare events and sparse covariates under the proportional intensity model and establish the validity of using the partial likelihood estimator and the observed information matrix for inference under this framework.

Table of Contents

List of Tables

List of Figures

Acknowledgements

Firstly, I would like to express my sincere gratitude to my advisor Prof. Zhiliang Ying for his continuous support of my Ph.D study and research. I could not forget his kindness, patience and immense knowledge. His guidance helped me in all the time of research and writing of this thesis. Besides my advisor, I would like to thank Professor Michael Sobel, Professor Jingchen Liu, Professor Hammou El Barmi and Professor Edward Hak-Sing Ip for serving on my dissertation committee and providing insightful comments on this thesis. I would also like to thank Qiwei He for hosting me in ETS for a summer and ETS for providing the PIAAC item portal and data that are used in this thesis.

I would also like to thank other faculty and staff in the statistics department. I am particularly thankful to Dood Kalicharan and Anthony Cruz for the assistance with various administrative works.

I am also deeply indebted to Professor Phillip Yam, Professor Gary Chan, Professor Tony Sit and Professor Chuan-Fa Tang for other collaborative works and leading me to other interesting research areas although the works do not constitute part of this thesis.

During the time in Columbia, I am fortunate to meet many good friends. Furthermore, I also benefit a lot from discussing with my friends and classmates.

Last, but not least, I would like to thank my parents, my brother for their unconditional support in encouraging me to achieve my goal in life, especially during the difficult and stressful times. It is their love and encouragement that has enabled me to complete this work.

Chapter 1: Introduction

Survival and event history analysis have been important tools in diverse disciplines, inlcuding biostatistics, reliability theory, insurance, business, and social sciences, where one is interested in the occurrence of events. Such events can be classified as survival events or recurrent events. Examples of survival events are the time from birth to death, the time from disease onset to death and the time from marriage to divorce. Examples of recurrent events are myocardial infarction, cancer tumors and birth of child. This chapter will discuss some of the important concepts and statistical models used in survival and event history analysis, where focus will be on the methods based on event counts. Methods that are based on waiting or gap times will not be discussed here. We also focus on the setting where events occur in continuous time.

In the rest of the following, we use a.s. to denote almost surely, $\xrightarrow{d}$ to denote convergence in distribution and $\xrightarrow{\mathbb{P}}$ to denote convergence in probability.

1.1 Survival Function and Hazard Function

We begin with survival data where we are interested in the survival time, which is the time from the initiating event to the event of interest. Note that we use the term “survival" even though the event of interest may not be death. Let T denote this survival time. Then, T is a nonnegative random variable and its survival function is defined by

$$S(t) := \mathbb{P}(T > t), \quad t \geq 0.$$

Assume now that T is a continuous random variable with density function f. The hazard function is defined as

$$\begin{aligned}\lambda(t) &:= \lim_{\Delta t \downarrow 0} \frac{1}{\Delta t}\mathbb{P}(t \le T < t + \Delta t | T \ge t) \\ &= -\frac{1}{S(t)}\frac{d}{dt}S(t) \\ &= \frac{f(t)}{S(t)}.\end{aligned}$$

The interpretation of λ is that $\lambda(t)\Delta t$ is the probability that the failure time occurs in the very small interval $[t, t + \Delta t)$ given that the event has not occurred by time t. The function $\Lambda(t) := \int_0^t \lambda(s)ds$ is called the cumulative hazard function for T.

1.2 Right-censored Survival Data

A common feature in survival data that is different from the usual data is that the survival time could be censored and/or truncated. The most common type of censoring is right-censoring as described below. Let U denote the censoring time variable. Instead of observing the survival time T, we only observe $(\tilde{T}, \delta)$, where $\tilde{T} := \min(T, U)$ and $\delta := I(T \le U)$. In the random (right) censorship model, U and $\tilde{T}$ are assumed to be independent or conditional independent given the covariates in the presence of covariates. Let S be the survival function of T and G be the distribution function of U. Suppose that T has density f that is parameterized by a parameter θ and U has density g that does not depend on θ (noninformative censoring). The likelihood function for θ from the data $(\tilde{T}_i, \delta_i)$, $i = 1, \ldots, n$, is

$$L(\theta) = \prod_{i=1}^{n} \{f(\tilde{T}_i|\theta)G(\tilde{T}_i)\}^{\delta_i}\{S(\tilde{T}_i)g(\tilde{T}_i)\}^{1-\delta_i} \propto \prod_{i=1}^{n} f(\tilde{T}_i|\theta)^{\delta_i} S(\tilde{T}_i)^{1-\delta_i}.$$

Now, define a process N by specifying $N(t) = 1$ if $\tilde{T} \leq t$ and $\delta = 1$ and $N(t) = 0$ otherwise. Then it can be shown that the process M given by

$$M(t) := N(t) - \int_0^t I(\tilde{T} \geq u)\lambda(u)du$$

is a martingale provided that

$$\lim_{\Delta t \downarrow 0} \frac{1}{\Delta t}\mathbb{P}(t \leq T < t + \Delta t | T \geq t) = \lim_{\Delta t \downarrow 0} \frac{1}{\Delta t}\mathbb{P}(t \leq T < t + \Delta t | T \geq t, U \geq t). \tag{1.1}$$

The condition in (1.1) is called independent censoring. The process N is an example of a counting process that will be described in the next section.

1.3 Counting Process and Martingales

Counting process is one of the notions that is central to survival and event history analysis. A counting process is a stochastic process $\{N(t) : t \geq 0\}$ adapted to a filtration $\{\mathcal{F}_t : t \geq 0\}$ with $N(0) = 0$ and $N(t) < \infty$ almost surely, and whose paths are with probability one right-continuous, piecewise constant, and have only jump discontinuities, with jumps of size +1. With this definition, $N(t) - N(s)$ is the number of events occurring in the interval $(s, t]$. If N is the counting process corresponding to the survival time T, i.e., $N(t) = 1$ if $t \geq T$ and $N(t) = 0$ otherwise, then the hazard function can be written as

$$\lambda(t) = \lim_{\Delta t \downarrow 0} \frac{1}{\Delta t}\mathbb{P}(N((t + \Delta t)-) - N(t-) = 1 | N(t-) = 0), \tag{1.2}$$

where $N(t-) := \lim_{u \uparrow t} N(u)$. This is because $\{N((t + \Delta t)-) - N(t-) = 1\} = \{t \leq T < t + \Delta t\}$ and $\{N(t-) = 0\} = \{T \geq t\}$. The form in (1.2) will give us the motivation for the definition of intensity function for general counting process as described in the next section.

One of the most important theorems in using counting process method for survival and event history analysis is the Doob-Meyer decomposition, where the following version is from Theorem

2.2.3 in [19].

Theorem 1. *Let $X = \{X(t) : t \geq 0\}$ be a nonnegative right-continuous $\mathcal{F}_t$-local submartingale with localizing sequence $\{\tau_n\}$, where $\{\mathcal{F}_t : t \geq 0\}$ is a right-continuous filtration. Then there exists a unique increasing right-continuous predictable process A such that $A(0) = 0$ a.s., $\mathbb{P}(A(t) < \infty) = 1$ for all $t > 0$, and $X - A$ is a right-continuous local martingale.*

As a result of Theorem 1, any counting process can be decomposed into the sum of a local martingale and a predictable increasing process. Alternatively, there exists a unique right-continuous predictable increasing process A such that $A(0) = 0$ a.s., $A(t) < \infty$ a.s., for any t, and the process $M = N - A$ is a local martingale. The process A is called a compensator. Furthermore, if $\mathbb{E}(A(t)) < \infty$ for all t, then $M = N - A$ is a martingale. In more complicated situation, we shall assume for a particular compensator A that $N - A$ is a martingale. This will be discussed in more details in Section 1.8.

Statistics for counting process data are often of the form

$$U_n = \sum_{i=1}^{n} \int H_i dM_i, \tag{1.3}$$

where $M_i = N_i - A_i$ are the compensated counting process. When H_i are locally bounded predictable processes, U_n is a local square integrable martingale. The following theorem establish the form of predictable variation and covariation processes for $\int H_i dM_i$, $i = 1, 2$.

Theorem 2 (Theorem 2.4.3 in [19])**.** *Assume that on a stochastic basis $(\Omega, \mathcal{F}, \{\mathcal{F}_t : t \geq 0\}, \mathbb{P})$:*

(1) H_i is a locally bounded $\mathcal{F}_t$-predictable process;

(2) N_i is a counting process.

Then for the local martingales $M_i = N_i - A_i$,

$$\left\langle \int H_1 dM_1, \int H_2 dM_2 \right\rangle = \int H_1 H_2 d\langle M_1, M_2 \rangle,$$

that is, the process

$$\int H_1 dM_1 \int H_2 dM_2 - \int H_1 H_2 d\langle M_1, M_2 \rangle$$

is a local martingale over $[0, \infty)$*, where* $\langle M_1, M_2 \rangle$ *is the predictable covariation process of* M_1 *and* M_2*.*

The results in the following lemma are special cases of Lenglart's inequality.

Lemma 1 (Lemma 8.2.1 in [19])**.** *Let* N *be a univariate counting process with continuous compensator* A*. Let* $M = N - A$*, and let* H *be a locally bounded, predictable process. Then for all* $\delta, \rho > 0$ *and any* $t \geq 0$*,*

(a)

$$\mathbb{P}(N(t) \geq \rho) \leq \frac{\delta}{\rho} + \mathbb{P}(A(t) \geq \delta).$$

(b)

$$\mathbb{P}\left(\sup_{0 \leq y \leq t} \left| \int_0^y H(x) dM(x) \right| \geq \rho \right) \leq \frac{\delta}{\rho^2} + \mathbb{P}\left(\int_0^t H^2(x) dA(x) \geq \delta \right).$$

The following theorem is a multivariate martingale central limit theorem that is useful in the development of the asymptotic distribution of the statistic of the form (1.3), for example, the score in the multiplicative intensity model.

Theorem 3 (Theorem 5.3.5 in [19])**.** *Let* $W_1^*, \ldots, W_r^*$ *be* r *dependent time-transformed Brownian motion processes. Specifically,* $(W_1^*, \ldots, W_r^*)$ *is an* r*-variate Gaussian process having components with independent increments,* $W_l^*(0) = 0$ *a.s. and, for all* $0 \leq s \leq t$*,* $\mathbb{E}(W_l^*(t)) = 0$ *and* $\mathbb{E}(W_l^*(s) W_{l'}^*(t)) = C_{ll'}(t)$*, where* $C_{ll'}$ *is a continuous function for all* $l, l' \in \{1, \ldots, r\}$*. Suppose* $\{N_i^{(n)} : i = 1, \ldots, n\}$ *satisfies* $\{N_{i,l}^{(n)} : i = 1, \ldots, n, l = 1, \ldots, r\}$ *is a multivariate counting process with stochastic basis* $(\Omega, \mathcal{F}, \mathcal{F}_t, P)$*, the compensator* $A_{i,l}^{(n)}$ *of* $N_{i,l}^{(n)}$ *is continuous, and* $H_{i,l}^{(n)}$ *is a locally bounded* $\mathcal{F}_t$*-predictable process. Consider the vector of local square integrable martingales*

$(U_1^{(n)}, \ldots, U_r^{(n)})$ *where, for* $l = 1, \ldots, r$ *and for* $t \geq 0$,

$$U_l^{(n)}(t) = \sum_{i=1}^{n} \int_0^t H_{i,l}^{(n)}(s) d\{N_i^{(n)}(s) - A_i^{(n)}(s)\}.$$

Suppose for each $l, l' \in \{1, \ldots, r\}$ *and for all* $t > 0$*:*

$$\begin{aligned} \langle U_l^{(n)}, U_{l'}^{(n)} \rangle(t) &= \sum_{i=1}^{n} \int_0^t H_{i,l}^{(n)}(s) H_{i,l'}^{(n)}(s) dA_i^{(n)}(s) && (1.4) \\ &\xrightarrow{\mathbb{P}} C_{ll'}(t), \text{ as } n \to \infty && (1.5) \end{aligned}$$

and

$$\begin{aligned} \langle U_{l,\varepsilon}^{(n)}, U_{l,\varepsilon}^{(n)} \rangle(t) &= \sum_{i=1}^{n} \int_0^t \{H_{i,l}^{(n)}(s)\}^2 I\{|H_{i,l}^{(n)}(s)| > \varepsilon\} dA_i^{(n)}(s) \\ &\xrightarrow{\mathbb{P}} 0 \text{ as } n \to \infty \text{ for any } \varepsilon > 0. \end{aligned}$$

Then

$$(U_1^{(n)}, \ldots, U_r^{(n)}) \xrightarrow{d} (W_1^*, \ldots, W_r^*) \text{ in } (D[0, \tau])^r \text{ as } n \to \infty.$$

1.4 Recurrent Event

As briefly mentioned in the introduction, we are often interested not only in a failure time but event that could happen more than once. Such an event is called recurrent event. For instance, in a carcinogenicty experiment on the times to the development of mammary tumors for rats ([21]), the rats can develop more than one tumor. The data is then a collection of the times at which the tumors are detected, denoted by $\{t_{ij} : i = 1, \ldots, j = 1, \ldots, n_i\}$ for a sample size of n, where n_i is the total number of tumors detected for rat i. More examples are given in [9] and the references therein. The information in t_{ij}'s can be summarized using the counting process notation by denoting $N_i(t)$ to be the number of events occur in $[0, t]$ for any t.

Let $\{\mathcal{F}_t\}_{t \geq 0}$ be a filtration where N is adapted to, where $\mathcal{F}_t$ can be understood as the history of

the process (together with other processes that could influence it) accumulate up to and including time t. As a generalization of hazard function defined in (1.2), the intensity function corresponding to a generally counting process N can be defined as

$$\lambda(t|\mathcal{F}_{t-}) = \lim_{\Delta t \downarrow 0} \frac{1}{\Delta t} \mathbb{P}(N((t+\Delta t)-) - N(t-) = 1|\mathcal{F}_{t-}).$$

In words, $\lambda(t|\mathcal{F}_{t-})$ is the instantaneous probability of an event occurring at t, conditional on the process history.

1.4.1 Likelihood for a Single Event Type

Given n events occur at times $t_1 < t_2 < \ldots < t_n$ over the time interval $[a, b]$, the probability density for a process with intensity function $\lambda(t|\mathcal{F}_{t-})$ conditional on $\mathcal{F}_a$ is

$$\prod_{j=1}^{n} \lambda(t_j|\mathcal{F}_{t_j-}) \cdot \exp\left\{ - \int_a^b \lambda(u|\mathcal{F}_{u-}) du \right\}. \tag{1.6}$$

To see this, consider a partition $a = u_0 < u_1 < \ldots < u_M = b$ of $[a, b]$, where each interval $[u_i, u_{i+1}]$ contains at most one event time and $\{t_1, \ldots, t_n\} \subset \{u_0, \ldots, u_M\}$. Then

$$\begin{aligned}
& \mathbb{P}(N(u_1) = n(u_1), \ldots, N(u_M) = n(u_M)|\mathcal{F}_a) \qquad (1.7)\\
= & \prod_{i=0}^{M} \mathbb{P}(N(u_i) = n(u_i)|\mathcal{F}_{u_i-}) \\
= & \prod_{i=0}^{M} \left\{ \mathbb{P}(\Delta N(u_i) = 1|\mathcal{F}_{u_i-})^{\Delta N(u_i)} \mathbb{P}(\Delta N(u_i) = 0|\mathcal{F}_{u_i-})^{1-\Delta N(u_i)} \right\}, \\
= & \prod_{i=1}^{n} \mathbb{P}(\Delta N(t_i) = 1|\mathcal{F}_{t_i-}) \prod_{i=0: u_i \notin \{t_1,\ldots,t_n\}}^{M} \mathbb{P}(\Delta N(u_i) = 0|\mathcal{F}_{u_i-})^{1-\Delta N(u_i)}.
\end{aligned}$$

where $\Delta N(u_i) = N(u_{i+1}-) - N(u_i-)$ is the number of events in $[u_i, u_{i+1})$. From the definition of intensity function and the assumption that events cannot occur simultaneously, we have

$$\begin{aligned}\mathbb{P}(\Delta N(u_i) = 0|\mathcal{F}_{u_i-}) &= 1 - \lambda(u_i|\mathcal{F}_{u_i-})\Delta u_i + o(\Delta u_i).\\ \mathbb{P}(\Delta N(u_i) = 1|\mathcal{F}_{u_i-}) &= \lambda(u_i|\mathcal{F}_{u_i-})\Delta u_i + o(\Delta u_i).\end{aligned}$$

Therefore, (1.7) equals

$$\prod_{i=0:u_i\in\{t_1,\ldots,t_n\}}^{M} \{\lambda(u_i|\mathcal{F}(u_i-))\Delta u_i + o(\Delta u_i)\} \prod_{i=0:u_i\notin\{t_1,\ldots,t_n\}}^{M} \{1 - \lambda(u_i|\mathcal{F}_{u_i-})\Delta u_i + o(\Delta u_i)\}.$$

Dividing the above expression by $\prod_{i=0:u_i\in\{t_1,\ldots,t_n\}}^{M} \Delta u_i$, we have

$$\prod_{i=1}^{n} \lambda(t_i|\mathcal{F}(t_i-)) \prod_{i=0}^{M} \{1 - \lambda(u_i|\mathcal{F}_{u_i-})\Delta u_i + o(\Delta u_i)\}\{1 + o(1))\}.$$

Letting $M \to \infty$ such that $\max_r \Delta u_r \to 0$, we have (1.6). Note that (1.6) holds for any $n \geq 0$. Therefore, in particular, the likelihood when there is no event in $[a, b]$ is

$$\exp\left\{ - \int_a^b \lambda(u|\mathcal{F}_{u-})du \right\}.$$

In the case when we have external covariate processes, we shall assume that σ-algebra $\mathcal{F}_a$ have already included them. See also [2] for a more general and rigorous derivation of the likelihood.

1.5 Multitype Recurrent Event

Now, suppose that we have more than one type of event, then each of them correspond to a counting process. When we consider them jointly, we have a multivariate counting process. Formally, a J-variate process $\{N_1, \ldots, N_J\}$ is called a multivariate counting process if each N_j, $j = 1, \ldots, J$, is a counting process and no two component processes jump at the same time. An important fact about a multivariate counting process with continuous compensators is that the

corresponding martingales M_i and M_j are orthogonal as shown in the following theorem.

Theorem 4 (Theorem 2.5.2 in [19]). *Let $\{N_1, \ldots, N_J\}$ be a multivariate counting process and for $j = 1, \ldots, J$, let A_j be the compensator of N_j. Assume that each A_j is a continuous process. Let $M_i = N_i - A_i$. Then $\langle M_j, M_j \rangle = A_j$ and $\langle M_i, M_j \rangle \equiv 0$ a.s. for $i \neq j$.*

1.5.1 Likelihood for Multitype Recurrent Event

Now, we consider the situation when we have J types of events. Let $N(t) = (N_1(t), \ldots, N_J(t))'$, where $N_j(t)$ denote the number of type j events occurring over the interval $[0, t]$ be a multivariate counting process adapted to the filtration $\{\mathcal{F}_t\}$. It should be noted that all the components of the counting process are adapted to the common filtration $\{\mathcal{F}_t\}$. The intensity function for N_j is defined as

$$\lambda_j(t|\mathcal{F}_{t-}) := \lim_{\Delta t \downarrow 0} \frac{1}{\Delta t} \mathbb{P}(N_j((t + \Delta t)-) - N_j(t-) = 1|\mathcal{F}_{t-}).$$

We also assume that at most one event can occur at any given time, with

$$\begin{aligned}
\mathbb{P}(\Delta N_j(t) = 1|\mathcal{F}_{t-}) &= \lambda_j(t|\mathcal{F}_{t-})\Delta t + o(\Delta t), \\
\mathbb{P}(\Delta N_\cdot(t) = 0|\mathcal{F}_{t-}) &= 1 - \sum_{j=1}^{J} \lambda_j(t|\mathcal{F}_{t-})\Delta t + o(\Delta t), \\
\mathbb{P}(\Delta N_\cdot(t) \geq 2|\mathcal{F}_{t-}) &= o(\Delta t),
\end{aligned}$$

where $\Delta N_\cdot(t) := \sum_{j=1}^{J} \Delta N_j(t)$. Let $t_{jk}, k = 1, \ldots, n_j$ denote the times of type j events over $[a, b]$ for $j = 1, \ldots, J$. To derive the likelihood of observing $\mathcal{T} := \{t_{jk} : k = 1, \ldots, n_j, j = 1, \ldots, J\}$, consider a partition $a = u_0 < u_1 < \ldots < u_M = b$ of $[a, b]$, where each interval $[u_i, u_{i+1}]$ contains

at most one event time and $\mathcal{T} \subset \{u_0, \ldots, u_M\}$. Then

$$
\begin{aligned}
& \mathbb{P}(N(u_1) = n(u_1), \ldots, N(u_M) = n(u_M)|\mathcal{F}_a) \qquad (1.8)\\
= & \prod_{i=0}^{M} \mathbb{P}(N(u_i) = n(u_i)|\mathcal{F}_{u_i-}) \\
= & \prod_{i=0}^{M} \prod_{j=1}^{J} \left\{ \mathbb{P}(\Delta N_j(u_i) = 1|\mathcal{F}_{u_i-})^{\Delta N_j(u_i)} \mathbb{P}(\Delta N_j(u_i) = 0|\mathcal{F}_{u_i-})^{1-\Delta N_j(u_i)} \right\}, \\
= & \prod_{j=1}^{J} \prod_{k=1}^{n_j} \mathbb{P}(\Delta N_j(t_{jk}) = 1|\mathcal{F}_{t_{jk}-}) \prod_{i=0:u_i \notin \mathcal{T}}^{M} \mathbb{P}(\Delta N_{j\cdot}(u_i) = 0|\mathcal{F}_{u_i-}). \\
= & \prod_{j=1}^{J} \prod_{k=1}^{n_j} \{\lambda_j(t_{jk}|\mathcal{F}_{t_{jk}-})\Delta t_{jk} + o(\Delta t_{jk})\} \prod_{i=0:u_i \notin \mathcal{T}}^{M} \left\{ 1 - \sum_{j=1}^{J} \lambda_{ij}(u_i|\mathcal{F}_{u_i-})\Delta u_i + o(\Delta u_i) \right\}.
\end{aligned}
$$

Now, dividing the last expression by $\prod_{j=1}^{J} \prod_{k=1}^{n_j} \Delta t_{jk}$ and taking the limit as $M \to \infty$ such that $\max_i \Delta u_i \to 0$, we obtain

$$
L = \prod_{j=1}^{J} \prod_{k=1}^{n_j} \lambda_j(t_{jk}|\mathcal{F}_{t_{jk}-}) \exp\left\{ - \sum_{j=1}^{J} \int_a^b \lambda_j(u|\mathcal{F}_{u-}) du \right\}.
$$

By writing

$$
L = \prod_{j=1}^{J} \left[\prod_{k=1}^{n_j} \lambda_j(t_{jk}|\mathcal{F}_{t_{jk}-}) \exp\left\{ - \int_a^b \lambda_j(u|\mathcal{F}_{u-}) du \right\} \right],
$$

we see that the type-specific intensity functions λ_j's are functionally independent. If the intensity functions do not share the same parameters, then estimation could be performed separately by maximum likelihood.

As a result of the discussion above, the event processes are mutually independent (conditional on the covariates). However, it is usually not the case that the covariates could explain all the association between different event types. Furthermore, in many cases, there could be unobserved covariate or latent effects in the processes. As discussed in [50], a routine use of random-effects model for multivariate failure time data is recommended. Let θ denote the random effect. We

enlarge our original filtration $\mathcal{F}_t$ by including θ in $\mathcal{F}_0$. In this way, the likelihood becomes

$$L = \int_\theta \prod_{j=1}^{J} \left[\prod_{k=1}^{n_j} \lambda_j(t_{jk}|\mathcal{F}_{t_{jk}-}) \exp\left\{ -\int_a^b \lambda_j(u|\mathcal{F}_{u-})du \right\} \right] \psi(\theta;\gamma)d\theta,$$

where ψ is the density of θ and γ is its variance component. Thus, the event processes are no longer independent conditional on the covariates and the type-specific intensity functions are no longer functionally independent in the likelihood.

1.6 Poisson Process

The Poisson process is considered canonical when modeling count data over time. It can be defined mathematically in various equivalent ways. Using intensity function, a Poisson process is the one which satisfies

$$\lambda(t|\mathcal{F}_{t-}) = \rho(t), \quad t > 0,$$

for some nonnegative integrable function ρ. When $\rho(t) \equiv \rho$, the process is called homogeneous; otherwise, it is called inhomogeneous. Let N be a counting process with intensity function $\rho(\cdot)$. Define $\mu(t) := \int_0^t \rho(u)du$ for $t > 0$. The Poisson process satisfies the following properties:

(i) $N(t) - N(s) \sim \text{Poisson}(\mu(t) - \mu(s))$, for $0 \le s < t$;

(ii) $N(t_1) - N(s_1)$ and $N(t_2) - N(s_2)$ are independent if $(s_1, t_1] \cap (s_2, t_2] = \emptyset$.

For a counting process, we define the mean function by $t \mapsto \mathbb{E}(N(t))$ and the rate function by $t \mapsto \frac{d}{ds}\mathbb{E}(N(s))|_{s=t}$. For a Poisson process, the mean function is $\mu(t)$ and the rate function is equal to the intensity function $\rho(t)$. In general, the rate function is not equal to the intensity function as $dN(t)$ is not independent of $\mathcal{F}_{t-}$.

A result that is particular useful for simulating an inhomogeneous Poisson process is the following. Let $\{N(t) : t \ge 0\}$ be a Poisson process with mean function $\mu(t)$. Define the time change $s = \mu(t)$ and define the process $\{N^*(s) : s \ge 0\}$ by $N^*(s) = N(\mu^{-1}(s))$ for $s > 0$. Then $\{N^*(s) : s \ge 0\}$ is a homogeneous Poisson process with rate function $\rho^*(s) = 1$.

1.7 Covariates

In survival and event history analysis, we are often interested in relating the process with covariates. In general, covariates have both fixed covariates and time-varying covariates. Examples of fixed covariates are age, gender and indicator representing the group. Examples of time-varying covariates are patient's measurements at different times. We shall use $X(\cdot)$ to denote a vector of covariate processes for fixed effects and $Z(\cdot)$ to denote a vector of covariate processes for random effects. An important distinction with a time-varying covariate is whether it is external or internal. We define a time-varying covariate to be external if it is independent of the recurrent event process under consideration. A time-varying covariate is internal if it is not external. Examples of internal covariate process are whether one experienced the event before time t and the total number of events experienced before time t. See also Chapter 6 in [27] for more details.

1.8 Regression Models

1.8.1 Cox Proportional Hazards Model

For right-censored survival data with covariates, we observe $(\tilde{T}, \delta, X)$, where $\tilde{T} = \min(T, U)$ is the minimum of a failure time and a censoring time, $\delta = I(T \leq U)$ is the indicator of the event that failure has been observed and X is a vector of covariates. Using counting process notation, the information contained in $(\tilde{T}, \delta)$ is the same as that contained in $N(t) = I(\tilde{T} \leq t, \delta = 1)$ and $Y(t) = I(\tilde{T} \geq t)$. As a result, there are two approaches to censored data regression models. The traditional approach, which is also used by [10], specifies that the conditional hazard function is

$$\lambda(t|X) = \lim_{\Delta t \downarrow 0} \frac{1}{\Delta t} \mathbb{P}(t \leq T < t + \Delta t | T \geq t, X).$$

Hence, for small values of Δt,

$$\lambda(t|Z)\Delta t \approx \mathbb{P}(t \leq T < t + \Delta t | T \geq t, X).$$

The interpretation is that $\lambda(t|X)\Delta t$ is approximately the conditional probability of observing a failure in $[t, t+\Delta t)$ given X and no failure before t. The Cox proportional hazards model ([10]) assumes that

$$\lambda(t|X) = \lambda_0(t)e^{\beta^T X}.$$

As $S(t|X) = e^{-\int_0^t \lambda(u|X)du}$, the proportional hazards model in terms of the conditional survival function is $S(t|X) = \{S_0(t)\}^{e^{\beta^T X}}$, where $S_0(t) = e^{-\int_0^t \lambda(u)du}$.

1.8.2 General Intensity-based Regression

A more general modeling approach is to based on the Doob-Meyer decomposition and by modeling the compensator as follows. Define the right-continuous filtration $\{\mathcal{F}_t\}$ by

$$\mathcal{F}_t := \sigma\{X(u), N(u), Y(u+) : 0 \le u \le t\}.$$

From the Doob-Meyer decomposition, there exists a unique predictable process A such that $N - A$ is a martingale. Therefore, heuristically, we have

$$\mathbb{E}(dN(t)|\mathcal{F}_{t-}) = \mathbb{E}(dA(t)|\mathcal{F}_{t-}) = dA(t),$$

because A is predictable. From this equation, we see that $dA(t)$ is the rate of change for N conditional on the information up to $t-$. If we assume that A is absolute continuous with

$$A(t) = \int_0^t l(s)ds$$

for some random function l, then we can model l instead. The process A and l are called the cumulative intensity function and intensity function for N respectively. The multiplicative intensity model or the proportional intensity model specifies that

$$A(t) = \int_0^t e^{\beta^T X(s)} Y(s)\lambda_0(s)ds,$$

where λ_0 is a so-called baseline intensity function. Note that as long as Y and X are predictable, A is predictable. In this case, the intensity function is therefore

$$\lambda(t|\mathcal{F}_{t-}) = Y(t)\lambda_0(t)e^{\beta^T X(t)}.$$

This is the formulation used in [3]. Under some regularity conditions, we also have

$$\lim_{\Delta t \downarrow 0} \frac{1}{\Delta t}\mathbb{P}(N(t+\Delta_t) - N(t)|\mathcal{F}_t) = \lambda(t+).$$

This provide an interpretation for the approach that models the compensator directly.

1.8.3 Semiparametric Models and Partial Likelihood

In the Cox proportional hazards model, the baseline function λ_0 is fully unspecified while the regression coefficients belong to the Euclidean space, therefore, it is a semiparametric model. Inference is usually based on the "partial likelihood". The partial likelihood could be derived as a profile likelihood from the full likelihood which is described below. First, we consider the case of right-censored survival data $\{(\tilde{T}_i, \delta_i, X_i) : i = 1, \ldots, n\}$ with hazard function $\lambda_i(t) = \lambda_i(t|\mathcal{F}_{t-}) = \lambda_0(t)e^{\beta^T X_i(t)}$ and denote its corresponding survival function by S_i. The likelihood is

$$\prod_{i=1}^{n} \lambda_i(\tilde{T}_i)^{\delta_i} S_i(\tilde{T}_i) = \prod_{i=1}^{n} \{\lambda_0(\tilde{T}_i)e^{\beta^T X_i(\tilde{T}_i)}\}^{\delta_i} e^{-\int_0^{\tilde{T}_i} e^{\beta^T X_i(t)} d\Lambda(t)}.$$

The maximum of this function does not exists if $\Lambda(\cdot)$ is restricted to be absolutely continuous. Thus we allow $\Lambda(\cdot)$ to be any increasing right-continuous function and replace $\lambda(t)$ with the jump size of Λ at time t, denoted by $\Lambda\{t\}$. We then maximize the modified log-likelihood function

$$\log L_n(\Lambda, \beta) := \sum_{i=1}^{n} \delta_i [\log \Lambda\{\tilde{T}_i\} + \beta^T X_i(\tilde{T}_i)] - \sum_{i=1}^{n} \int_0^{\tilde{T}_i} e^{\beta^T X_i(t)} d\Lambda(t),$$

over Λ and β. The maximizer of Λ can be seen to be a step function with jumps at X_i's. Hence,

$$\int_0^{\tilde{T}_i} e^{\beta^T X_i(t)} d\Lambda(t) = \sum_{j:\tilde{T}_j \le \tilde{T}_i} \Lambda\{\tilde{T}_j\} e^{\beta^T X_i(\tilde{T}_j)}.$$

Therefore,

$$\log L_n(\Lambda, \beta) = \sum_{i=1}^{n} \delta_i [\log \Lambda\{\tilde{T}_i\} + \beta^T X_i(\tilde{T}_i)] - \sum_{j:\tilde{T}_j \le \tilde{T}_i} \Lambda\{\tilde{T}_j\} e^{\beta^T X_i(\tilde{T}_j)}. \tag{1.9}$$

Assume there are no ties in the data $\{\tilde{T}_1, \ldots, \tilde{T}_n\}$. Denote $\Lambda_i := \Lambda\{X_i\}$. Then,

$$\frac{\partial \log L_n(\Lambda, \beta)}{\partial \Lambda_k} = \frac{\delta_k}{\Lambda_k} - \sum_{i=1}^{n} I(\tilde{T}_i \ge \tilde{T}_k) e^{\beta^T X_i(\tilde{T}_k)}.$$

Setting $\frac{\partial \log L(\Lambda,\beta)}{\partial \Lambda_k} = 0$, we hobtain

$$\Lambda_k = \frac{\delta_k}{\sum_{i=1}^{n} I(\tilde{T}_i \ge \tilde{T}_k) e^{\beta^T X_i(\tilde{T}_k)}}. \tag{1.10}$$

From (1.9) and (1.10), we have

$$\log L_n(\beta) := \sum_{i=1}^{n} \delta_i \left(\log \frac{\delta_i}{\sum_{k=1}^{n} I(\tilde{T}_k \ge \tilde{T}_i) e^{\beta^T X_k(\tilde{T}_i)}} + \beta^T X_i(\tilde{T}_i) \right) - \sum_{i=1}^{n} \sum_{j:\tilde{T}_j \le \tilde{T}_i} \frac{\delta_j e^{\beta^T X_i(\tilde{T}_j)}}{\sum_{k=1}^{n} I(\tilde{T}_k \ge \tilde{T}_j) e^{\beta^T X_k(\tilde{T}_j)}}.$$

The second term on the RHS of the the above equation equals

$$\sum_{i=1}^{n} \frac{\sum_{j=1}^{n} I(\tilde{T}_j \le \tilde{T}_i) \delta_j e^{\beta^T X_i(\tilde{T}_j)}}{\sum_{k=1}^{n} I(\tilde{T}_k \ge \tilde{T}_j) e^{\beta^T X_k(\tilde{T}_j)}} = \sum_{j=1}^{n} \delta_j \frac{\sum_{i=1}^{n} I(\tilde{T}_j \le \tilde{T}_i) e^{\beta^T X_i(\tilde{T}_j)}}{\sum_{k=1}^{n} I(\tilde{T}_k \ge \tilde{T}_j) e^{\beta^T X_k(\tilde{T}_j)}} = \sum_{j=1}^{n} \delta_j.$$

Therefore,

$$\begin{aligned} L_n(\beta) &= e^{-\sum_{j=1}^n \delta_j} \prod_{i=1}^n \left(\frac{e^{\beta^T X_i(\tilde{T}_i)}}{\sum_{k=1}^n I(\tilde{T}_k \geq \tilde{T}_i) e^{\beta^T X_k(\tilde{T}_i)}} \right)^{\delta_i} \\ &\propto \prod_{i=1}^n \left(\frac{e^{\beta^T X_i(\tilde{T}_i)}}{\sum_{k=1}^n I(\tilde{T}_k \geq \tilde{T}_i) e^{\beta^T X_k(\tilde{T}_i)}} \right)^{\delta_i}. \end{aligned}$$

The partial likelihood is

$$PL(\beta) := \prod_{i=1}^n \left(\frac{e^{\beta^T X_i(\tilde{T}_i)}}{\sum_{k=1}^n I(\tilde{T}_k \geq \tilde{T}_i) e^{\beta^T X_k(\tilde{T}_i)}} \right)^{\delta_i}.$$

Now, suppose we have recurrent event data $\{N_i(t), Y_i(t), X_i(t) : 0 \leq t \leq \tau_i, i = 1, \ldots, n\}$ with intensity function $\lambda_i(t|\mathcal{F}_{t-}) = Y_i(t)\lambda_0(t)e^{\beta^T X_i(t)}$. Let t_{ij}'s denote the event times. The likelihood is

$$L(\Lambda, \beta) = \prod_{i=1}^n \prod_{j=1}^{n_i} \lambda_i(t_{ij}) e^{-\int_0^{\tau_i} e^{\beta^T X_i(t)} Y_i(t) d\Lambda(t)}.$$

Similarly to the case for right-censored survival data, we only consider right-continuous step functions Λ with jumps at the event times. Then log-likelihood function is

$$\log L(\Lambda, \beta) = \sum_{i=1}^n \sum_{j=1}^{n_i} [\log \Lambda\{t_{ij}\} + \beta^T X_i(t_{ij})] - \sum_{i=1}^n \int_0^{\tau_i} e^{\beta^T X_i(t)} Y_i(t) d\Lambda(t). \tag{1.11}$$

Let $\{s_j : j = 1, \ldots, n^*\}$ be the set of all event times, where $n^* := \sum_{i=1}^n n_j$. Then $\int_0^{\tau_i} Y_i(t) e^{\beta^T X_i(t)} d\Lambda(t) = \sum_{j: s_j \leq \tau_i} Y_i(s_j) e^{\beta^T X_i(s_j)} \Lambda\{s_j\}$. Denote $\Lambda_k := \Lambda\{s_k\}$. Then,

$$\frac{\partial \log L(\Lambda, \beta)}{\partial \Lambda_k} = \frac{1}{\Lambda_k} - \sum_{i=1}^n Y_i(s_k) e^{\beta^T X_i(s_k)}.$$

Setting $\frac{\partial \log L(\Lambda, \beta)}{\partial \Lambda_k} = 0$, we obtain

$$\Lambda_k = \frac{1}{\sum_{l=1}^n Y_l(s_k) e^{\beta^T X_l(s_k)}}. \tag{1.12}$$

From (1.11) and (1.12), we have

$$\begin{aligned}\log L(\beta) &= \sum_{i=1}^{n}\sum_{j=1}^{n_i}\left\{\log\frac{1}{\sum_{l=1}^{n}Y_l(t_{ij})e^{\beta^T X_l(t_{ij})}}+\beta^T X_i(t_{ij})\right\}\\ &\quad -\sum_{i=1}^{n}\sum_{j:s_j\le\tau_i}\frac{Y_i(s_j)e^{\beta^T X_i(s_j)}}{\sum_{l=1}^{n}Y_l(s_j)e^{\beta^T X_l(s_j)}}.\end{aligned}$$

Note that

$$\sum_{i=1}^{n}\sum_{j:s_j\le\tau_i}\frac{Y_i(s_j)e^{\beta^T X_i(s_j)}}{\sum_{l=1}^{n}Y_l(s_j)e^{\beta^T X_l(s_j)}}=\sum_{j=1}^{n^*}\frac{\sum_{i=1}^{n}Y_i(s_j)e^{\beta^T X_i(s_j)}}{\sum_{l=1}^{n}e^{\beta^T X_l(s_j)}}=n^*.$$

Hence, we have the partial likelihood

$$PL(\beta):=\prod_{i=1}^{n}\prod_{j=1}^{n_i}\frac{e^{\beta^T X_i(t_{ij})}}{\sum_{l=1}^{n}Y_l(t_{ij})e^{\beta^T X_i(t_{ij})}}.$$

The log partial likelihood can be written as

$$\log \mathrm{PL}(\beta)=\sum_{i=1}^{n}\int_0^\infty\left[\beta^T X_i(t)-\log\left\{\sum_{k=1}^{n}Y_k(t)e^{\beta^T X_k(t)}\right\}\right]dN_i(t).$$

The score function is

$$\frac{\partial\log \mathrm{PL}(\beta)}{\partial\beta}=\sum_{i=1}^{n}\int\left\{X_i(t)-\frac{\sum_{k=1}^{n}Y_k(t)e^{\beta^T X_k(t)}X_k(t)}{\sum_{k=1}^{n}Y_k(t)e^{\beta^T X_k(t)}}\right\}dN_i(t),$$

which can be written as

$$\frac{\partial\log \mathrm{PL}(\beta)}{\partial\beta}=\sum_{i=1}^{n}\int\left\{X_i(t)-\frac{\sum_{k=1}^{n}Y_k(t)e^{\beta^T X_k(t)}X_k(t)}{\sum_{k=1}^{n}Y_k(t)e^{\beta^T X_k(t)}}\right\}dM_i(t), \tag{1.13}$$

where $M_i(t) := N_i(t) - A_i(t)$ and $A_i(t) := \int_0^t \lambda_0(s)e^{\beta^T X_i(s)}Y_i(s)ds$. Note that M_i is a local square-integrable martingale and the score process is a martingale transform. Therefore, martingale method is available for establishing the asymptotic theory for the proportional intensity

model. With these notations, the score function can be written as

$$\frac{\partial \log \mathrm{PL}(\beta)}{\partial \beta} = \sum_{i=1}^{n} \int \left\{ X_i(t) - \frac{S_n^{(1)}(\beta, t)}{S_0^{(n)}(\beta, t)} \right\} dN_i(t).$$

The observed Fisher information is

$$-\frac{\partial^2 \log \mathrm{PL}(\beta)}{\partial \beta \partial \beta^T} = \int \left[\frac{S_n^{(2)}(\beta, t)}{S^{(0)}(\beta, t)} - \left\{ \frac{S_n^{(1)}(\beta, t)}{S_n^{(0)}(\beta, t)} \right\}^2 \right] dN_i(t).$$

An important fact about the log partial likelihood is that it is a concave function in β. To see that, define

$$V_n(\beta, t) := \frac{S_n^{(2)}(\beta, t)}{S_n^{(0)}(\beta, t)} - \left\{ \frac{S_n^{(1)}(\beta, t)}{S_n^{(0)}(\beta, t)} \right\}^2$$

and

$$E_n(\beta, t) := \frac{S_n^{(1)}(\beta, t)}{S_n^{(0)}(\beta, t)}.$$

Note that V_n can be written as

$$V_n(\beta, t) = \frac{n^{-1} \sum_{i=1}^{n} \{X_i(t) - E_n(\beta, t)\}^{\otimes 2} Y_i(t) e^{\beta^T X_i(t)}}{S_n^{(0)}(\beta, t)}, \tag{1.14}$$

where $a^{\otimes 2} = aa^T$ for any vector a. Therefore, the Hessian matrix of the log partial likelihood is negative semidefinite and hence the log partial likelihood is concave.

1.8.4 Additive Models

An alternative to the Cox proportional hazards model or the multiplicative hazard model is the additive hazards model. In fact, both models belong to a family of hazard-based regression models where the conditional hazard function of the survival time T takes the form

$$\lambda(t|X) = L(\lambda_0(t), \beta^T X(t)),$$

where $\lambda_0(t)$ is a completely unspecified function and L is a known function to be specified. The choice of $L(x, y) = xe^y$ yields the Cox proportional hazards model and the choice $L(x, y) = x + y$ gives the following additive hazards model

$$\lambda(t|X) = \lambda_0(t) + \beta^T X(t).$$

The above model was studied by [4], [11], [43] and [32] among others. A related but fully non-parametric additive hazards model was originally proposed by [1], where the conditional hazard takes the form

$$\lambda(t|X) = \beta(t)^T X(t);$$

[1] and others have developed least squares estimation of the integrated regression coefficients

$$B(t) = \int_0^t \beta(s)ds.$$

A class of general additive-multiplicative hazards models is also studied in [33]:

$$\lambda(t|X, W) = g(\beta^T W(t)) + \lambda_0(t)h(\gamma^T X(t)),$$

where $(W^T, X^T)^T$ is a vector of covariates, $(\beta^T, \gamma^T)^T$ is a vector of unknown regression regression parameters, g and h are known link functions and λ_0 is again an unspecified "baseline hazard function" under $g \equiv 0$ and $h \equiv 1$.

1.8.5 Marginal Models

In the proportional intensity model, it is assumed that

(a) $\mathbb{E}(dN(t)|\mathcal{F}_{t-}) = \mathbb{E}(dN(t)|X(t))$;

(b) $\mathbb{E}(dN(t)|X(t)) = \exp\{\beta_0^T X(t)\}\lambda_0(t)Y(t)dt$.

The first assumption postulates that the influence of prior events on the future recurrence, if there is any, depends only on the covariate process X at t and the second assumption specifies how the covariate process affects the instantaneous rate of N. To relax Assumption (a), we can remove Assumption (a) and take only Assumption (b). Such an estimation method is called a robust method because it allows arbitrary dependence structure among recurrent events. For example, if the true intensity function is given by

$$\lambda(t|\theta) = \theta e^{\beta_0^T Z(t)} \lambda_0(t),$$

where θ is an unobserved positive random effect with mean 1 that is independent of Z. Then, the proportional intensity model does not hold but Assumption (b) is satisfied. [31] provided a rigorous justification of such robust procedures through empirical process theory. Another robust method for analyzing recurrent events is the method based on multivariate failure time data (see [48]).

1.8.6 Frailty Models and Random Effects Models

To accommodate heterogeneity across individuals, one may consider a model with random effects. For example, for a Poisson model with random effect, a conditional subject-specific intensity function is

$$\lambda_i(t|\mathcal{F}_{t-}, \theta_i) = \lim_{\Delta t \downarrow 0} \frac{1}{\Delta t} \mathbb{P}(N((t+\Delta t)-) - N(t-) = 1|\mathcal{F}_{t-}, \theta_i) = \theta_i \rho_i(t),$$

where θ_i is an unobserved random effect, assumed to have mean 1 and variance ϕ. Let $\mu_i(t) = \int_0^t \rho_i(u)du$. Then,

$$\begin{aligned} \mathbb{E}(N_i(t)) &= \mu_i(t) \\ \mathrm{Var}(N_i(t)) &= \mu_i(t) + \mu_i^2(t)\phi \\ \mathrm{Cov}(N_i(s_1,t_1), N_i(s_2,t_2)) &= \phi \mu_i(s_1,t_1)\mu_i(s_2,t_2). \end{aligned}$$

Therefore, the (unconditional) event process is not a Poisson process. If θ_i were observed, the likelihood of the data $(n_i, t_{i1}, \ldots, t_{in_i}, \theta_i, Y_i)$ for subject i is

$$\prod_{j=1}^{n_i} \{\theta_i \rho_i(t_{ij})\} \exp\left\{ - \int_0^\infty Y_i(s)\theta_i \rho_i(s) ds \right\}.$$

Because θ_i's are unobserved, the observed likelihood is

$$\int_0^\infty \prod_{j=1}^{n_i} \{\theta_i \rho_i(t_{ij})\} \exp\left\{ - \int_0^\infty Y_i(s)\theta_i \rho_i(s) ds \right\} dG(\theta_i; \phi).$$

More generally, if ϕ is a random vector with distribution function G and parameter ϕ and if

$$\lambda_i(t|\mathcal{F}_{t-}, \theta_i) = \lambda_0(t) e^{\beta^T X(t) + \theta_i^T Z(t)}, \tag{1.15}$$

where X and Z are covariates, then the observed likelihood for n subjects is

$$\prod_{i=1}^{n} \int_0^\infty \prod_{j=1}^{n_i} \{\lambda_0(t_{ij}) e\beta^T X(t_{ij}) + \theta_i^T Z(t_{ij})\}$$
$$\times \exp\left\{ - \int_0^\infty Y_i(s)\lambda_0(s) e^{\beta^T X(s) + \theta_i^T Z(s)} ds \right\} dG(\theta_i; \phi).$$

Clearly, if we maximize the above likelihood over any function λ_0, the MLE does not exist. A possible remedy for this is to consider only increasing right-continuous Λ and replace replace $\lambda(t)$ with the jump size of Λ at time t as in the argument for obtaining the partial likelihood by profile likelihood. In [51], a more general semiparametric transformation model with random effects is considered:

$$\Lambda(t|X, Z; \theta) = G\left(\int_0^t \lambda(s) e^{\beta^T X(s) + \theta^T Z(s)} ds \right),$$

where G is a three times continuously differentiable and strictly increasing transformation function with $G(0) = 0$ and $G(\infty) = \infty$. When $G(x) \equiv x$, we have (1.15).

In general, marginal models and random-effects models are two distinct approaches to model

event history data. However, random-effects models have several important advantages over marginal models. First, with random-effects models, we can predict future events based on individual's event history. Secondly, we can make use of the nonparametric maximum likelihood estimation (see [50] and [52]), which yields asymptotically efficient estimators. Thirdly, the dependence structure from the random effects and the random effects themselves could be of scientific interest, as illustrated in the multivariate proportional intensity factor model proposed in Chapter 2, where the factors could have interpretation.

Chapter 2: A Multivariate Proportional Intensity Factor Model for Multivariate Counting Processes

2.1 Introduction

This chapter is motivated by the need for statistical modeling and analysis of process data, which refer to sequences of events of different types and are commonly encountered in scientific studies when a subject undergoes a series of the same and different types of events. Analyzing such data is complex due to the dynamic nature of both the events of interest and the covariate processes. Furthermore, the data are often heterogeneous and contain a large number of different types of events and covariate processes. Our main goal of this article is to propose a model for the joint analysis of such data, motivated by the emerging computed-based assessment in education.

Computer-based assessments, such as simulation-based or scenario-based assessments, that involve interactive environments have become increasingly popular. For example, the Organization for Economic Cooperation and Development (OECD) has been administrating interactive and scenario-based questions in the Program for International Student Assessment (PISA) and the Programme for the International Assessment of Adult Competencies (PIAAC). In the US, the National Assessment of Educational Progress (NAEP) has been using interactive computer tasks in science and in technology and engineering literacy in recent years [5]. At the same time, technical advances now allow the action sequences together with the timestamps of solving a problem to be recorded in log-files. These process data could provide new insights on individual characteristics as traditional task analysis and scoring normally focus only on the final task outcomes. These may include, for example, test taker's motivation, engagement, persistence and planning. Because of the potential benefits and the additional information that could be obtained from analyzing process data, research related to it has received considerable attention recently [23, 25, 54, 39, 34]. For

instance, [29] used response times to filter for test taker motivation and [22] measured student engagement in collaboration using process data. However, few approaches have considered the joint statistical modeling of the process data that include all the events together with the timestamps.

Formally, a process data is of the form $\{(a_1, t_1), \ldots, (a_m, t_m)\}$, where $a_i \in \mathcal{J}$ is the ith event, $t_i \in \mathbb{R}_+$, with $t_i < t_{i+1}$, is the corresponding timestamp, m is the number of events and $\mathcal{J}$ is a discrete set. Without loss of generality, we can assume that $\mathcal{J} = \{1, \ldots, J\}$ with J equals the total number of possible event types. In this chapter, we are interested in the situation where we have independent observations from n subjects.

A natural way for representing process data is the use of multivariate counting processes. There are several methods available for analyzing multivariate event time data [46, 37, 48, 28, 30, 49, 45]. These methods mostly include the use of marginal models or frailty (or random effects) models. On the other hand, in educational and psychological measurement, applications often make use of factor analysis or multidimensional item response theory and find interpretation of the factors. To handle more general multivariate event time data and to explore a low-dimensional structure of the counting processes and covariates with the same spirit of factor models, we specify that the intensity function of the jth event type of the ith subject to be

$$\lambda_j(t|X_{ij}, Z_{ij}; \theta_i) = \lambda_{j0}(t) Y_{ij}^*(t) e^{\beta_j^T X_{ij}(t) + \theta_i^T A_j^T Z_{ij}(t)}, \tag{2.1}$$

where i indexes subject, j indexes event types, λ_{j0} is the event-specific baseline hazard function that is common to all subjects, Y_{ij}^* is an indicator process, X_{ij} and Z_{ij} are L_{1j}- and L_{2j}-dimensional covariate processes associated with the jth event type for the fixed and random effects respectively, β_j is a vector of regression coefficients for the event-specific fixed effects, θ_i is the subject-specific K-dimensional random effects, and A_j is an event-specific $L_{2j} \times K$ factor loading matrix. When $K < L_{2j}$, dimension reduction of the random coefficients is achieved. Here, we assume the random effects following a multivariate normal distribution: $\theta \sim N(0, \Sigma)$, with density $\phi(\theta; 0, \gamma)$ where γ is the vector of the variance components.

It is instructive to note that Model (2.1) contains many well-known models as special cases.

(i) When $L_2 = 0$ and $J = 1$, it reduces to the standard univariate proportional hazard model [10]

$$\lambda(t|X_i) = \lambda_0(t)e^{\beta^T X_i(t)}.$$

(ii) When $L_2 = 0$, it reduces to the multivariate proportional hazard model

$$\lambda_j(t|X_j) = \lambda_{j0}(t)e^{\beta_j^T X_{ij}(t)}.$$

(iii) When $K = L_2$, A is the identity matrix, it reduces to a proportional hazard model with random effects

$$\lambda_j(t|X_i, Z_i; \theta_i) = \lambda_{j0}(t)e^{\beta^T X_{ij}(t)+\theta_i^T Z_{ij}(t)}, \quad j = 1, \ldots, J.$$

In particular, when $\lambda_{j0}(t) \equiv \lambda_0(t)$, it is a model for clustered survival data [45].

(iv) When $L_2 = 1$, $Z_{il}(t) \equiv 1$, $K = 1$, $J = 1$, it reduces to the standard frailty model [46]

$$\lambda(t|X_i; \theta_i) = \lambda_0(t)e^{\beta^T X_i(t)+\theta_i} = \tilde{\theta}_i\lambda_0(t)e^{\beta^T X_i(t)},$$

where $\tilde{\theta}_i := e^{\theta_i}$.

(v) When $L_2 = 1, Z_{il}(t) \equiv 1$, $K = 1$, it reduces to a shared frailty model

$$\lambda_j(t|X_i; \theta_i) = \lambda_{j0}(t)e^{\beta_j^T X_i(t)+a_j\theta_i}, \quad j = 1, \ldots, J.$$

(vi) When $L_1 = 0$ and $L_2 = 1$ with $Z_i(t) \equiv 1$, it reduces to a factor model for multivariate counting processes

$$\lambda_j(t|X_i; \theta_i) = \lambda_{j0}(t)e^{a_j^T\theta_i}.$$

In particular, when the baseline functions are all constant, it becomes a Poisson factor model; see, for example, [47].

For inference of the parameters, we first discuss the method of nonparametric maximum likelihood estimator (NPMLE) in estimating the model parameters. We establish that the NPMLE is consistent, asymptotically normal and asymptotically efficient with covariance matrix that can be consistently estimated by the inverse information matrix or the profile likelihood method under some suitable regularity conditions.

In practice, we do not know which covariates should be included as the fixed effects and be loaded on which factors. Therefore, variable selection in both the fixed and random effects are usually necessary. For variable selection, the best subset selection procedure along with various information criteria, such as the Akaike information criterion and the Bayesian information criterion, become computationally infeasible with even moderate number of parameters. The least absolute shrinkage and selection operator (lasso) proposed by [44] has a much more tractable computation method [20] and has been applied to various models. In the same spirit of lasso, penalized likelihood with nonconcave penalty functions has been proposed to select significant variables [14, 15, 16, 53]. The resulting penalized estimators are shown to have the oracle properties. That is, the estimators perform as well as if the correct model were known. In this paper, we adopt the same approach by imposing a nonconcave penalty on the log-likelihood to select the significant variables for both the fixed and random effects. Note that the literature mentioned above focuses on variable selection of the fixed effects. For variable selection of the random effects, sparse estimation in the factor loadings have been studied in [8], [36] and [26] for the factor analysis models and [42] for the multidimensional item response theory models. The above papers study sparse factor loading which deal mainly with continuous and binary data type while our proposed method deals with process data, which is a much more complex data type.

Here is the outline of this chapter. In Section 2.2, we discuss nonparametric maximum likelihood estimation. In Section 2.3, we establish a variable selection procedure via penalized likelihood for parametric baseline intensity functions. Simulation studies are shown in Section 2.4. A

real data application is given in Section 2.5. In Section 2.6, we conclude with some remarks and extensions. All the technical proofs are relegated to the Appendix.

2.2 Nonparametric Maximum Likelihood Estimation

2.2.1 Setting

Recurrent event data are often subject to right-censoring. Therefore, we consider our model when the data are possibly right-censored. For a random sample of size n, the data consist of $\{N_i(t), Y_i(t), X_i(t), Z_i(t) : t \in [0,\tau], i = 1,\ldots,n\}$, where they are all vector-valued processes. For example, $N_i(t) = (N_{i1}(t),\ldots,N_{iJ}(t))^T$ and $Y_i(t) = (Y_{i1}(t),\ldots,Y_{iJ}(t))^T$. The process Y_{ij} is defined by $Y_{ij}(t) = I(C_i \geq t)Y^*_{ij}(t)$ and $N_{ij}(t) = N^*_{ij}(t \wedge C_{ij})$, where C_{ij} is the censoring time. Let τ be the duration of the study. We assume that the conditional probability of $C_{ij} > t$ given $\{X_{ij}(s), Z_{ij}(s), N^*_{ij}(s) : s \in [0,\tau]\}$ and θ depends only on $\{X_{ij}(s), Z_{ij}(s); s \leq t\}$ and is noninformative about $(\alpha, \mathcal{A})$, where α is the collection of all the finite dimensional parameters and $\mathcal{A} = (\Lambda_1,\ldots,\Lambda_J)$. In addition, we assume that the conditional distribution of $\{X(t), Z(t)\}$ given $\{X(s), Z(s), N(s), Y(s) : s < t\}$ is noninformative about $(\alpha, \mathcal{A})$ The first assumption is coarsening at random and the second assumption, which implies that no information on the parameters can be extracted from the covariate processes, is a standard assumption in any regression analysis.

Under the above two assumptions, the log-likelihood function for $(\alpha, \mathcal{A})$ is

$$\log \sum_{i=1}^{n} \int_{\theta} \prod_{j=1}^{J} \left[\prod_{t\leq\tau} \left\{ \lambda_{j0}(t) e^{\beta_j^T X_{ij}(t) + \theta_i^T A_j^T Z_{ij}(t)} \right\}^{dN_{ij}(t)} \right.$$
$$\left. \times \exp\left\{ -\int_0^{\tau} e^{\beta_j^T X_{ij}(t) + \theta_i^T A_j^T Z_{ij}(t)} Y_{ij}(t) d\Lambda_{j0}(t) \right\} \right] \phi_K(\theta; 0, \gamma) d\theta,$$

where $dN_{ij}(t) = N_{ij}(t) - N_{ij}(t-)$ denotes the jump of N_{ij} at t. The maximum of this function does not exist if we allow Λ to be absolute continuous. Thus, we replace $\lambda_{j0}(t)$ with the jump size of

Λ_{j0} at time t, denoted by $\Lambda_{j0}\{t\}$ and maximize the modified log-likelihood function

$$l_n(\alpha, \mathcal{A}) = \log \sum_{i=1}^{n} \int_{\theta} \prod_{j=1}^{J} \left[\prod_{t \le \tau} \left\{ \Lambda_{j0}\{t\} e^{\beta_j^T X_{ij}(t) + \theta_i^T A_j^T Z_{ij}(t)} \right\}^{dN_{ij}(t)} \right.$$
$$\left. \times \exp\left\{ - \int_0^{\tau} e^{\beta_j^T X_{ij}(t) + \theta_i^T A_j^T Z_{ij}(t)} Y_{ij}(t) d\Lambda_{j0}(t) \right\} \right] \phi_K(\theta; 0, \gamma) d\theta. \quad (2.2)$$

Clearly, the maximizer of $L_n(\alpha, \cdot)$ must be step functions Λ_j with jumps at the observed event times t_{ijm} ($i = 1, \ldots, n; j = 1, \ldots, J; m = 1, \ldots, n_{ij}$, where $n_{ij} = N_{ij}(\tau)$).

2.2.2 Theoretical Results

Denote $(\alpha_0, \mathcal{A}_0)$ to be the true value of $(\alpha, \mathcal{A})$. Let d be the length of α. We impose the following regularity conditions on the model and data structures.

(D1) α_0 lies in the interior of a compact set $\Theta \subset \mathbb{R}^d$ and $\Lambda'_{0j}(t) > 0$ for all $t \in [0, \tau]$, $j = 1, \ldots, J$.

(D2) With probability one, $X_{ijl}(\cdot)$ and $Z_{ijl}(\cdot)$ are of bounded variation in $[0, \tau]$ and are left-continuous with bounded left- and right-derivatives in $[0, \tau]$, for $j = 1, \ldots, J, l = 1, \ldots, L_j$.

(D3) With probability one, $\mathbb{P}(C_i \ge \tau | X_i, Z_i) > \delta_0 > 0$ for some constant δ_0.

(D4) For each $j = 1, \ldots, J$, if there exists a vector μ and a deterministic function $g(t)$ such that $g(t) + \mu^T X_{ij}(t) = 0$ with probability 1, then $\mu = 0$ and $g(t) = 0$; For each $j, l = 1, \ldots, J$, if there exists a matrix B such that $Z_{ij}^T(t) B Z_{il}(s) = 0$ with probability 1, then $B = 0$.

(D5) There exists K rows in $A := (A_1^T, \ldots, A_J^T)^T$ such that they form the $K \times K$ identity matrix.

Conditions (D1) and (D2) are standard assumptions in semiparametric regression models in survival analysis. Condition (D3) means that there is a positive probability for the events to be observed over the whole interval $[0, \tau]$. Conditions (D4) and (D5) ensure the model is identifiable and the information matrix is nonsingular. The first condition in (D4) is standard in regression analysis; the second condition in (D4) is due to the random coefficients. These are satisfied when

the covariates are linearly independent. Condition (D5) is a standard assumption in factor analysis when the covariance matrix in the random effects is unrestricted. This condition prohibit any orthogonal transformation between the factor loadings and the random effects.

The following theorem states the consistency of $\hat{\alpha}_n$ and $\hat{\Lambda}_j$, $j = 1, \ldots, J$.

Theorem 5. *Under Conditions (D1)-(D4),* $|\hat{\alpha}_n - \alpha_0| + \sum_{j=1}^{J} \sup_{t\in[0,\tau]} |\hat{\Lambda}_j(t) - \Lambda_{0j}(t)| \xrightarrow{a.s.} 0$.

To describe the asymptotic distribution, we need addition notations. For any set T, the space $l^\infty(T)$ is defined as the set of all uniformly bounded, real functions on T. Let $BV[0,\tau]$ denote the set of functions with bounded total variations on $[0,\tau]$. Define $\mathcal{V} = \{v \in \mathbb{R}^d : |v| \le 1\}$ and $\mathcal{Q} := \{h : ||h||_{V[0,\tau]} \le 1, h(0) = 0\}$, where $||h||_{V[0,\tau]}$ is the total variation of $h(\cdot)$ in $[0,\tau]$. Then $\hat{\Lambda}_{nj}$ can be considered as a bounded linear functional in $l^\infty(\mathcal{Q})$ by defining $\hat{\Lambda}_{nj}(h) = \int_0^\tau h(t)d\hat{\Lambda}_{nj}(t)$ for $h \in \mathcal{Q}$. We identify $(\hat{\alpha}_n - \alpha_0, \hat{\mathcal{A}}_n - \mathcal{A}_0)$ as a random element in $l^\infty(\mathcal{V}\times Q^J)$ through the definition $(\hat{\alpha}_n - \alpha_0)^T v + \sum_{j=1}^{J} \int_0^\tau h_j(s)d(\hat{\Lambda}_{j0} - \Lambda_{0j0})(s)$ for $v \in \mathcal{V}$ and $h_j \in Q$.

Theorem 6. *Under Conditions (D1)-(D5),* $\sqrt{n}(\hat{\alpha}_n - \alpha_0, \hat{\mathcal{A}}_n - \mathcal{A}_0) \xrightarrow{d} \mathcal{G}$ *in* $l^\infty(\mathcal{V}\times Q^J)$*, where* $\mathcal{G}$ *is a continuous zero-mean Gaussian process. Furthermore, the limiting covariance matrix of* $n^{1/2}(\hat{\alpha}_n - \alpha_0)$ *attains the semiparametric efficiency bound.*

By Theorem 6, we know that $\sqrt{n}(\alpha_n - \alpha_0)$ and $\sqrt{n}(\hat{\Lambda}_{j0} - \Lambda_{0j0})$ are asymptotically normal. To estimate their asymptotic variance, we can view (2.2) as a parametric log-likelihood with α and $\Lambda_{j0}\{t_{ijm}\}(j = 1, \ldots, J, i = 1, \ldots, n, m = 1, \ldots, n_{ij})$ the parameters. Then, the observed information matrix I_n is the negative of the Hessian matrix of (2.2) with respect to α and $\Lambda_{j0}\{t_{ijm}\}$'s evaluated at $\hat{\alpha}_n$ and $\hat{\Lambda}_{j0}\{t_{ijm}\}$'s. The asymptotic variance of $\sqrt{n}v^T(\hat{\alpha}_n - \alpha_0) + \sum_{j=1}^{J} \int_0^\tau h_j(s)d(\hat{\Lambda}_{j0} - \Lambda_{0j0})(s)$ equals that of $\sqrt{n}v^T(\hat{\alpha}_n - \alpha_0) + \sum_{j=1}^{J}\sum_{i=1}^{n}\sum_{m=1}^{n_{ij}} h_j(t_{ijm})\hat{\Lambda}_{j0}(t_{ijm})$, which can be consistently estimated by $n(v^T, \overline{h}_1^T, \ldots, \overline{h}_J^T)I_n^{-1}(v^T, \overline{h}_1^T, \ldots, \overline{h}_J^T)^T$, where $\overline{h}_j$ is the vector consisting of the values of $h_j(\cdot)$ at the observed event times, as shown in Theorem 7 below.

Theorem 7. *Under Conditions (D1)-(D5), I_n is invertible for all large enough n, and*

$$\sup_{v\in\mathcal{V},h_1,\ldots,h_J\in\mathcal{Q}}\left|n(v^T,\overline{h}_1^T,\ldots,\overline{h}_J^T)I_n^{-1}(v^T,\overline{h}_1^T,\ldots,\overline{h}_J^T)^T\right.$$
$$\left.-AVar\left[\sqrt{n}\left\{v^T(\hat{\alpha}_n-\alpha_0)+\sum_{j=1}^{J}\int h_j d(\hat{\Lambda}_j-\Lambda_{0j})\right\}\right]\right|\xrightarrow{\mathbb{P}}0,$$

where AVar denotes asymptotic variance.

Theorem 7 allows us to make inference about α and $\mathcal{A}$. An alternative approach to estimating the asymptotic covariance matrix of $\hat{\alpha}_n$ is to use the profile log-likelihood function with the negative second-order numerical difference of the profile log-likelihood function at $\hat{\alpha}_n$ as stated in Theorem 8.

Theorem 8. *Let $PL_n(\alpha)$ be the profile log-likelihood function for α. Under Conditions (D1)-(D5), for any $\varepsilon_n = O_P(n^{-1/2})$ and any vector v,*

$$-\frac{PL_n(\hat{\alpha}_n+\varepsilon_n v)-2PL_n(\hat{\alpha}_n)+PL_n(\hat{\alpha}_n-\varepsilon_n)}{n\varepsilon_n^2}\xrightarrow{\mathbb{P}}v^T\Sigma^{-1}v,$$

where Σ is the limiting covariance matrix of $\sqrt{n}(\hat{\alpha}_n-\alpha_0)$. Furthermore, $2\{PL_n(\alpha_n)-PL_n(\alpha_0)\}\xrightarrow{d}\chi_d^2$.

2.3 Variable Selection via Penalized Likelihood

2.3.1 Method

In some applications, we may want to include a large number of covariates in both the fixed and random coefficients parts of the model. It is then important and challenging to determine a subset of significant variables effectively and efficiently. Furthermore, a sparse factor loading matrix will usually provide a better interpretation of the factor. In such applications, to reduce computational burden, we consider parametric baseline function $\lambda_j(\cdot;\eta_j)$, where η_j is a finite

dimensional parameter, in this section. The log-likelihood is

$$\begin{aligned} l_n(\alpha) &= \log \sum_{i=1}^{n} \int_{\theta} \prod_{j=1}^{J} \left[\prod_{t} \left\{ \lambda_j(t;\eta_j) e^{\beta_j^T X_{ij}(t) + \theta_i^T A_j^T Z_{ij}(t)} \right\}^{dN_{ij}(t)} \right. \\ &\quad \left. \times \exp\left\{ - \int_0^{\tau} \lambda_j(t;\eta_j) e^{\beta_j^T X_{ij}(t) + \theta_i^T A_j^T Z_{ij}(t)} Y_{ij}(t) dt \right\} \right] \phi_K(\theta; 0, \gamma) d\theta, \end{aligned}$$

where α denotes all the parameters. We consider the following penalized likelihood

$$l_{n,p}(\alpha) = l_n(\alpha) - n\left\{ \sum_{j=1}^{J} \sum_{l=1}^{L} p_\gamma(\beta_{jl}) + \sum_{j=1}^{J} \sum_{l=1}^{L} \sum_{k=1}^{K} p_\gamma(a_{jlk}) \right\}, \tag{2.3}$$

where $p_\gamma(\cdot)$ is a penalty function and γ is the penalty parameter, for both variable selection and estimation. In both theory and practice, we can choose different penalty functions and parameters. Here, for simplicity, we have assumed that all the parameters share the same penalty function and penalty parameter γ. Note that we do not penalize the parameters in the baseline intensity function as well as the variance components in the random effects. The penalized estimator is defined as $\hat{\alpha}_n = \arg\max_\alpha l_{n,p}(\alpha)$. As discussed in [14], a good penalty function should result in an estimator with the properties of unbiasedness, sparsity and continuity. The smoothly clipped absolute deviation (scad) penalty ([14]) satisfies all the three requirements, which is defined by

$$p'_\gamma(x) = \gamma \left\{ I(x \le \gamma) + \frac{(a\gamma - x)_+}{(a-1)\gamma} I(x > \gamma) \right\}$$

for some $a > 2$ and $x > 0$. We adopt this penalty function in the application and choose $a = 3.7$ as suggested by [14].

2.3.2 Computation Algorithm

For a specific value of γ, to maximize (2.3), we could, in principle, apply the expectation-maximization algorithm [12] by treating θ_i, $i = 1, \ldots, n$, as the missing data. In the E-step, we compute the expectation of the complete data log-likelihood with respect to the conditional

distribution of the missing data given the observed data. In the present case, there is no closed form expression for this conditional expectation. Hence, numerical approximation of the E-step or stochastic versions of the expectation-maximization algorithm could be used instead. Here, we describe the estimating procedure using the stochastic expectation-maximization algorithm [7] with Metropolis algorithm [35] in the simulation step. In the stochastic E-step, we simulate θ_i from its conditional distribution given the observed data. In the M-step, the resulting complete data log-likelihood using the simulated θ_i is maximized. In this M-step, we apply the coordinate descent algorithm that is developed for the estimation for the generalized linear models with convex penalties [20]. The stochastic expectation-maximization algorithm iterates between the stochastic E-step and M-step until convergence. The details are in the appendix.

2.3.3 Theoretical Results

Since we assume a parametric baseline intensity function in this section, we have the following additional condition for the identifiability of the model, which simply says that the baseline intensity function is uniquely parameterized.

(D6) If $\lambda_j(t;\eta_j) = \lambda_j(t;\eta_{0j})$ for all t, then $\eta_j = \eta_{0j}$.

Write $\alpha_0 = (\alpha_{10}^T, \alpha_{20}^T)^T$ and $\hat{\alpha}_n = (\hat{\alpha}_1^T, \hat{\alpha}_2^T)$. Without loss of generality, assume that $\alpha_{20} = 0$. Let $I_1(\alpha_{10})$ be the Fisher information matrix knowing $\alpha_{20} = 0$. Let $a_n = \max\{p'_{\gamma_n}(|a_{j0}|) : a_{j0} \neq 0\}$. Theorem 9 shows that there exists a local maximizer of the penalized log-likelihood that converges at the rate $n^{-1/2} + a_n$. With the scad penalty functions, such a local maximizer is $\sqrt{n}$-consistent if $\gamma_n \to 0$.

Theorem 9. *Under Conditions (D1)- (D6), if* $\lim_{n\to\infty} \max\{|p''_{\gamma_n}(|\alpha_{j0}|) : \alpha_{j0} \neq 0\} = 0$*, then there exists a local maximizer* $\hat{\alpha}_n$ *of* $l_{n,p}(\alpha)$ *such that* $||\hat{\alpha}_n - \alpha_0|| = O_p(n^{-1/2} + a_n)$.

We need additional notations to describe the oracle property. Let s be the number of non-zero component in α_0. Denote

$$\tilde{\Sigma} = \text{diag}\{p''_{\gamma_n}(|\alpha_{10}|), \ldots, p''_{\gamma_n}(|\alpha_{s0}|)\}$$

and

$$b = (p'_{\gamma_n}(|\alpha_{10}|)\text{sgn}(\alpha_{10}), \ldots, p'_{\gamma_n}(|\alpha_{s0}|)\text{sgn}(\alpha_{s0}))^T.$$

Theorem 10. *Under Conditions (D1)- (D6), if the penalty function satisfies*

$$\liminf_{n\to\infty} \liminf_{x\to 0+} p'_{\gamma_n}(x)/\gamma_n > 0,$$

where $\gamma_n \to 0$ *and* $\sqrt{n}\gamma_n \to \infty$ *as* $n \to \infty$*, then the* $\sqrt{n}$ *consistent local maximizers* $\hat{\alpha}_n = (\hat{\alpha}_1, \hat{\alpha}_2)$ *in Theorem 9 satisfy* $\mathbb{P}(\hat{\alpha}_2 = 0) = 1$ *and*

$$\sqrt{n}(I_1(\alpha_{10}) + \tilde{\Sigma})(\hat{\alpha}_1 - \alpha_{10} + (I_1(\alpha_{10}) + \tilde{\Sigma})^{-1}b) \xrightarrow{d} N(0, I_1(\alpha_{10})).$$

As a result of Theorem 10, with the scad penalty functions, the $\sqrt{n}$-consistent local maximizer will satisfy $\hat{\alpha}_2 = 0$ with probability tending to 1 and $\sqrt{n}(\hat{\alpha}_1 - \alpha_{10})$ being asymptotically normal with covariance matrix $I_1^{-1}(\alpha_{10})$ if $\gamma_n \to 0$ and $\sqrt{n}\gamma_n \to \infty$.

2.3.4 Choice of regularization parameter

Let $x \in \mathbb{R}^d$. Define $\mathcal{S}(x)$ to be the binary vector $\mathcal{S}(x) = (I(x_1 \neq 0), \ldots, I(x_d \neq 0))$. For choosing the regularization parameters γ, we apply the Bayesian information criterion ([38]). Specifically, for each value of γ, we can obtain a penalized estimator $\hat{\alpha}_n^\gamma$ and and $\mathcal{S}(\hat{a}_n^\gamma)$. The Bayesian information criterion at this value of γ is computed as

$$\text{BIC}(\gamma) = \underset{\alpha:\mathcal{S}(\alpha)=\mathcal{S}(\hat{\alpha}_n^\gamma)}{\arg\max} \{-2l_{n,p}(\alpha) + \log(n)p\}, \tag{2.4}$$

where p is the number of parameters. In practice, $\text{BIC}(\gamma)$ is evaluated at a set of grid points, $\mathcal{G}$, that are uniformly spaced in log-scale. The proposed γ is chosen as $\gamma^* = \arg\min_{\gamma\in\mathcal{G}} \text{BIC}(\gamma)$.

2.4 Simulation Study

In this section, we perform simulation studies under a setting that is similar to the one in the real data example. Specifically, consider a test item where the test taker is required to evaluate information in some websites. A sample item similar to this one and the real data could be found on the PIAAC website of The Organisation for Economic Co-operation and Development (OECD). Two screenshots of the sample item are shown in Figure 2.1 and Figure 2.2 (Source: http://www.oecd.org/skills/piaac/Problem%20Solving%20in%20TRE%20Sample%20Items.pdf). Figure 2.1 shows the first page the test takers will see. They are required to access and evaluate information relating to job search in a simulated web environment that is similar to the one in the real world. In particular, they can click on the links and perform actions like going back and forward. If they click on the second link “Work Links", it will link to the page as shown in Figure 2.2. The test takers could then click the button “Learn More" to obtain further information.

In the simulation setting, for simplicity, assume that there are 3 such websites and in each website there is a further web link to another website that provide additional information about the website. In the item, one can click on these links, go back and forward in the browser. To answer the question, the test taker needs to click on a pull-down menu and select one of the 3 websites as the answer. The test taker can finish the item by clicking the “Next" button and confirm if he really wants to finish the item by answering “OK" or “Cancel". In total, there are 15 event types (see Table 2.4). An example of the process data from this setting is given in Table 2.1. The data is generated from our proposed model with covariate processes that include the information of the past two events. To be specific, for the ith subject, let $X_{il}(t) = 1$, for $l = 1, \ldots, 14$, if the last event happened before time t is the lth event type and $X_{il}(t) = 0$ otherwise. Also, let $X_{i,l+14}(t) = 1$, for $l = 1, \ldots, 3$, if the last event is Back and the second last event is Wl, for $l = 1, \ldots, 3$. The same covariate processes are used for the fixed effects, the random effects, and across different event types. That is, $X_{ijl}(\cdot) \equiv Z_{ijl}(\cdot) \equiv X_{il}(\cdot)$ for each $j = 1, \ldots, J, l = 1, \ldots, L$. For instance, using the example in Table 2.1, $X_{\mathrm{W2}}(t) = 1$ when $t \in [15, 25)$, $X_{\mathrm{Back}}(t) = 1$ when $t \in [25, 28) \cup [36, 42)$

and $X_{\text{W2, Back}}(t) = 1$ when $t \in [25, 28)$, where the subscripts i are suppressed and the names of the event type are used for clarity. In the simulation setting, there are 23 nonzero parameters for the fixed effects and there are 3 dimensions in the random coefficients, with 13 nonzero factor loadings.

The focus of the simulation study is to assess the performance of the penalized estimator obtained from the stochastic expectation-maximization algorithm together with choosing the tuning parameter using the Bayesian information criterion. We will first evaluate the recovery of the true structure by using the following criteria:

1. $C_0 = 1$ if there exists a tuning parameter γ such that the $\{j : \hat{\alpha}_j^\gamma \neq 0\} = \{j : \alpha_{j0} \neq 0\}$ and $\{j : \hat{\alpha}_j^\gamma = 0\} = \{j : \alpha_{j0} = 0\}$.

2. $C_1 = 1$ if the tunning parameter γ chosen using the Bayesian information criterion gives $\{j : \hat{\alpha}_j^\gamma \neq 0\} = \{j : \alpha_{j0} \neq 0\}$ and $\{j : \hat{\alpha}_j^\gamma = 0\} = \{j : \alpha_{j0} = 0\}$.

3. True positive rate:
$$\text{TPR} = \frac{|\{j : \hat{\alpha}_j \neq 0, \alpha_{j0} \neq 0\}|}{|\{j : \alpha_{j0} \neq 0\}|}.$$

4. False discovery rate:
$$\text{FDR} = \frac{|\{j : \hat{\alpha}_j \neq 0, \alpha_{j0} = 0\}|}{|\{j : \alpha_{j0} = 0\}|}.$$

For computing TPR and FDR, the estimate $\hat{\alpha}_j$ is the one with correspond to the minimum BIC. Table 2.2 shows the results of these criteria averaged across 100 independent simulations. It can be seen that when the sample size increases, the probability of BIC choosing the correct model increases. Also, when the true model is not selected, the nonzero parameters are always estimated nonzero and only very few parameters are estimated to be nonzero when they are actually zero.

We also evaluate the bias of the estimates, the accuracy of the standard error formula and the coverage probability. In computing the bias and the standard error, we only make use of the estimates that match the true structure. The results show that the bias is small except for a

few parameters. The standard error estimates are close to the sample standard deviation of the estimates and yield reasonable coverages except when the biases are relatively large.

Event types and their meanings in the simulation studies

Event Type (Simulation)	Meaning
Wi ($i = 1, \ldots, 3$)	Click the link of the ith webpage
Wi_M ($i = 1, \ldots, 3$)	Click the "More" link in the ith webpage
Next	Click the “Next" button
Next_Cancel	Click the “Cancel" button in the pop-up window that will appear after clicking the “Next" button
R_i ($i = 1, \ldots, 3$)	Choose the ith website as answer
R_Open	Click on the pull down menu for choosing an answer
R_Close	Close the pull down menu for choosing an answer without choosing an answer
Back	Click the back arrow in the toolbar
Forward	Click the forward arrow in the toolbar
Next_OK	Click the “OK" button in the pop-up window that will appear after clicking the “Next" button (the terminating event)

Event	W2	Back	W1	W1_M	Back	Back	W3	R_Open	R_3	Next	Next_OK
Time	15	25	28	34	36	38	42	45	50	52	53

Table 2.1: A hypothetical example of process data of a test taker in the simulation setting.

Evaluation critera

	C_0	C_1	TPR	FDR ($\times 10^{-2}$)
n = 500	0.85	0.60	1.00	0.13
n = 1000	0.96	0.78	1.00	0.05
n = 2000	0.99	0.83	1.00	0.03

Table 2.2: C_0 is the average of the number of times that there is a pair of tuning parameter that results in the true model. C_1 is the average of the number of times that the tuning parameter selected by BIC results in the true model. TPR and FDR denote the average of the true positive rates and false discovery rates from the models with tuning parameter selected by BIC respectively.

2.5 Application to PIAAC data

The Programme for the International Assessment of Adult Competencies (PIAAC) is a programme of assessment and analysis of adult skills. The survey measures adults' proficiency in key information-processing skills - literacy, numeracy and problem solving in technology rich environment (PSTRE) - and gathers information and data on how adults use their skills at home, at

work and in the wider community. The time-stamped action sequences data were logged during respondents' problem solving process. The proposed method is illustrated using one item in the PSTRE domain. The data used here consists of both response data and response process data of 3, 713 adults who answered all the items in the PSTRE domain from the United States, the England and Northern Ireland, Ireland, Japan and the Netherlands. This item shares the similar structure as the simulation settings, with a focus on evaluating respondents' skills in seeking key information through web pages

Table 2.3 summarizes the event types in the actual item and their corresponding meanings. Due to the nature of the item, the last two events will have a large impact on the next event to happen. Therefore, for the covariate processes, we include the information of the past two events. To be specific, for the ith subject, let $X_{il}(t) = 1$, for $l = 1, \ldots, 24$, if the last event happened before time t is the lth event type and $X_{il}(t) = 0$ otherwise. Also, let $X_{i,l+24}(t) = 1$, for $l = 1, \ldots, 5$, if the last event is Back and the second last event is Wl, for $l = 1, \ldots, 5$. The same covariate processes are used for the fixed effects, the random effects, and across different event types. That is, $X_{ijl}(\cdot) \equiv Z_{ijl}(\cdot) \equiv X_{il}(\cdot)$ for each $j = 1, \ldots, J, l = 1, \ldots, L$. See also the simulation section for an example. We shall use the notation $a \rightarrow b$ to represent the effect of the covariate processes a on the event type b.

We choose $K = 3$ for the dimension of the random effects. As discussed in the previous sections, to avoid possible rotation of the loading matrix A, we constrain three rows of A to be loading on only one dimension. These constraints are imposed on the effects W2 $\rightarrow$ W2_A, W2 $\rightarrow$ Back and W2, Back $\rightarrow$ W1. For example, the factor loading for W2 $\rightarrow$ W2_Author is not penalized in the first dimension and the factor loadings in the second and third dimensions are set to be 0. The first two constrains are set because these they represent different behaviors and are the most frequent patterns after the event W2. Furthermore, the second website is the correct answer. Hence, it will be of most interest to set the structure around the second website. The design for the third dimension is motivated by the results obtained from fitting the model without random effects, where the effects Wi, Back $\rightarrow$ Wj for different i and j have patterns that implies the test takers

tend to go to the next webpage instead of going back the previous page. By incorporating random effects and performing variable selection, we could see if these patterns are related across different webpages.

We apply the method with a sequence of pairs of penalty parameters (γ_1, γ_2), where γ_1 is for the fixed effects and γ_2 is for the random effects. The model with the smallest BIC occurs at $(0.000961, 0.00482)$, which are of different magnitude order. This suggests using two penalty parameters may provide a better exploration of different models with this data.

For the fixed effects, partial results are given below:

$$
\begin{aligned}
\lambda_{\mathrm{W1}}(t) &= \exp\{-3.83 + \ldots - 0.88\mathrm{W1, Back} - 0.87\mathrm{W2, Back} - 0.71\mathrm{W3, Back} - 0.5\mathrm{W4, Back} \\
&\quad +0.48\mathrm{W5, Back} + \ldots\}, \\
\lambda_{\mathrm{W2}}(t) &= \exp\{-5.75 + \ldots + 0.89\mathrm{W1, Back} - 2.12\mathrm{W2, Back} - 1.95\mathrm{W3, Back} - 1.42\mathrm{W4, Back} \\
&\quad -0.3\mathrm{W5, Back} + \ldots\}, \\
\lambda_{\mathrm{W3}}(t) &= \exp\{-6.94 + \ldots - 1.17\mathrm{W1, Back} + 1.62\mathrm{W2, Back} - 1.79\mathrm{W3, Back} - 2.16\mathrm{W4, Back} \\
&\quad -1.44\mathrm{W5, Back} + \ldots\}, \\
\lambda_{\mathrm{W4}}(t) &= \exp\{-6.59 + \ldots - 3.12\mathrm{W1, Back} - 0.34\mathrm{W2, Back} + 2.26\mathrm{W3, Back} - 1.35\mathrm{W4, Back} \\
&\quad -0.78\mathrm{W5, Back} + \ldots\}, \\
\lambda_{\mathrm{W5}}(t) &= \exp\{-7.43 + \ldots - 1.35\mathrm{W1, Back} - 0.83\mathrm{W2, Back} + 0\mathrm{W3, Back} + 3.34\mathrm{W4, Back} \\
&\quad -2.13\mathrm{W5, Back} + \ldots\}, \\
\lambda_{\mathrm{Back}}(t) &= \exp\{-9.6 + 6.12\mathrm{W1} + 7.34\mathrm{W1\_M} + 6.74\mathrm{W2} + 7.45\mathrm{W2\_A} + 6.9\mathrm{W3} + \\
&\quad 7.74\mathrm{W3\_A} + 6.55\mathrm{W3\_O1} + 6.69\mathrm{W3\_O2} + 6.79\mathrm{W4} + 6.83\mathrm{W5} + 6.78\mathrm{W5\_O} \\
&\quad + \ldots + 7.41\mathrm{Web}\}, \\
\lambda_{\mathrm{Next}}(t) &= \exp\{-6.454 + 4.78\mathrm{R1} + 5.29\mathrm{R2} + 5.23\mathrm{R3} + 5.28\mathrm{R4} + 4.97\mathrm{R5} + \ldots\}, \\
\lambda_{\mathrm{Web}}(t) &= \exp\{-8.3 + \ldots + 5.76\mathrm{Web} + \ldots\}.
\end{aligned}
$$

We see that the effects of clicking the links on the intensity of Back have large positive co-

efficients. This is because one has to go back to the main page in order to click on other links. In addition, we see that the coefficients for W1_M, W2_A and W3_A are slightly larger than the that for the other web links. This could be explained by the fact that the information contained in these three pages are much less than the other pages so that the test takers will finish reading and perform Back quicker. The coefficients of R_1 → Next, . . ., R_5 → Next are all positive and relatively large. This is because when the test takers have chosen an answer, they are much more likely to click Next to submit it. For the covariate process Web, its strongest effects are on Back and Web itself. This suggests that some test takers thought that Web will perform the action of going to the previous page and so they tended to try clicking Web one more time or realized that it will not work and performed Back instead. It is also interesting when we look at the coefficients of Wi, Back → Wj, where $i, j = 1, \ldots, 5$, where a sequential pattern of the browsing behavior is found. The coefficients of Wi, Back → Wj, where $i = 1, \ldots, 4$ and $j = i + 1$, are all positive and that when $i = 1, \ldots, 4$, $j \neq i + 1$, are all negative (with one zero). Also ,the coefficient of W5, Back → W1 is positive and those of W5, Back → Wj are all negative for $j = 2, \ldots, 5$.

For the random effects, partial results are given in Table 2.4. Recall that we constrain the effect of W2 → W2_A to be related to the first dimension only. It turns out that in the first dimension, many of the related relationships are of the same sign as the coefficient of W2 → W2_A. These include W1 → W1_M, W3 → W3_A, W3 → W3_O1 and W3_O1 → W3_O2. Also, the coefficients of these relationships are either 0 or very small in magnitude in the other two dimensions. Furthermore, we also see from the coefficients of R_Open → R_i, for $i = 1, \ldots, 5$, that the actions of checking for more details in the websites are positively related to choosing the correct answer. Another interesting finding is that the coefficient of the relationship Next → Next_Cancel is opposite to that of W2 → W2_A, suggesting people are more confident when they visit W1_M, W2_A, W3_A, W3_O1 and W3_O2. The second dimension is mainly related to the event Back. In particular, it can be seen that the coefficients of Wj → Back, for $j = 1, \ldots, 5$ are of similar magnitude and of the same sign. Finally, for the third dimension, it is mainly related to the sequential pattern observed in the fixed effects. We see that Wi, Back → Wj are positive when

$j = i + 1$ for $i = 1, \ldots, 4$ and are zero or negative when $j \neq i + 1$. Hence, the sequential patterns across different webpages are positively related.

Table 2.3: Event types and their meanings in real data

Event Type	Meaning
Wi ($i = 1, \ldots, 5$)	Click the link of the ith webpage
W1_More	Click the "More" link in the first webpage
Wi_A ($i = 2, 3$)	Click the "Author" link in the ith webpage
W3_Oi ($i = 1, 2$)	Click the ith order link in the third webpage
W5_O	Click the order link in the fifth webpage
Next	Click the “Next" button
Next_Cancel	Click the “Cancel" button in the pop-up window that will appear after clicking the “Next" button
R_i ($i = 1, \ldots, 5$)	Choose the ith website as answer
R_Open	Click on the pull down menu for choosing an answer
R_Close	Close the pull down menu for choosing an answer without choosing an answer
Back	Click the back arrow in the toolbar
Forward	Click the forward arrow in the toolbar
Home	Click the home button in the toolbar
Web	Click the Web environment icon
Next_OK	Click the “OK" button in the pop-up window that will appear after clicking the “Next" button (the terminating event)

2.6 Discussion

In this chapter, we proposed a multivariate proportional intensity factor model for multivariate event time data. We develop the theory of nonparametric maximum likelihood estimation as well as a variable selection and estimation method for the fixed effects and random effects simultaneously using parametric baseline intensity functions. From the simulation studies, we see that using the Bayesian information criterion provides a good choice of the tuning parameter and the whole procedure essentially recovers the true structure of the parameter with small bias and accurate standard errors. We further demonstrate the proposed method through a real data set from the Survey of Adult Skills in PIAAC. Our method finds meaningful relationships among different types of events that can help understanding both the task design and the behavior of subjects when solve a problem. Furthermore, the proposed method can be applied to both exploratory and confirmatory

analysis or a combination of them by controlling the number of constraints on the loading matrix.

Although we implicitly assume all the event types are recurrent, we can also allow some events to be survival times. For the distribution of the random effects, the multivariate normal distribution allows an unrestricted covariance structure between the random effects. However, other distributions can also be used and the theoretical results remain valid subject to some regularity conditions on the random effect distributions; see [52] for more details. The proposed model can also be easily extended to have a multilevel structure, where we could have, for example, a cluster level above the subject level with cluster-specific random effects.

While we illustrate the method using educational assessment data, the method is widely applicable. For example, in medical studies, for each person, we are often interested in several illnesses at the same time. When the number of random coefficients is moderate to large, the proposed model can achieve a parsimonious model.

2.7 Appendix

2.7.1 Sample Task

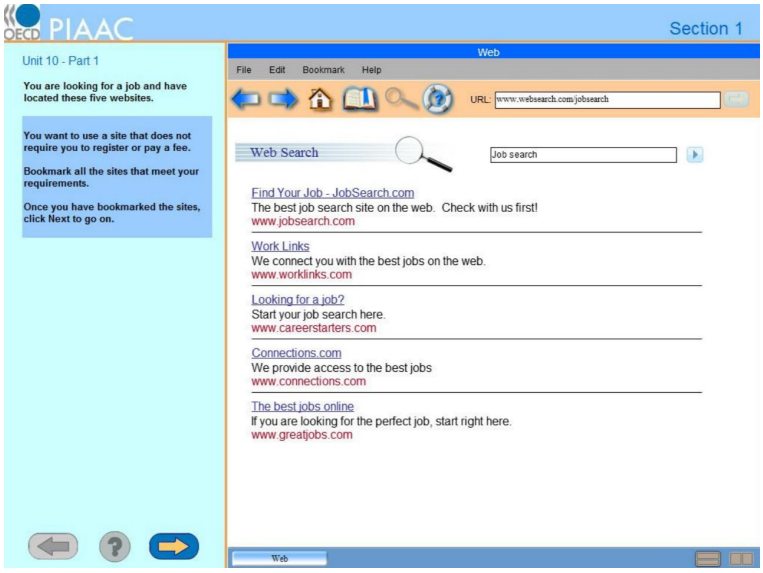

Figure 2.1: Screenshot of the sample item given in OECD website.

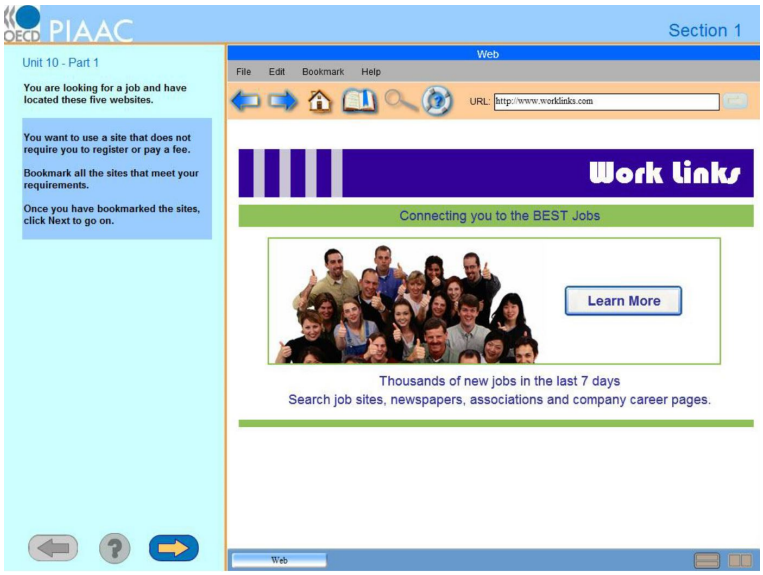

Figure 2.2: Screenshot of the sample item given in OECD website.

2.7.2 Estimation algorithm

In this section, we give the details of the computation algorithm described in Section 2.3.2. Let $(\eta^{(t)}, \beta^{(t)}, A^{(t)}, \gamma^{(t)})$ and $\theta^{(t)} = (\theta_1^{(t)}, \ldots, \theta_n^{(t)})$ denote the estimates and the simulated θ_i at the tth iteration respectively. At the $(t+1)$th iteration:

(1) Stochastic E-step via Metropolis Algorithm: for each $i = 1, \ldots, n$,

(i) Sample θ_i^* from the proposal distribution $N(\theta_i^{(t)}, \delta_i^2)$, where δ_i^2 is the proposal variance.

(ii) Compute the acceptance ratio

$$r_i = \frac{L_c(\alpha^{(t)}|N_i, X_i, Z_i, \theta_i^*)}{L_c(\alpha^{(t)}|N_i, X_i, Z_i, \theta_i^{(t)})},$$

where $L_c(\alpha|N_i, X_i, Z_i, \theta_i)$ denotes the complete data likelihood for the ith subject:

$$\begin{aligned} & L_c(\alpha|N_i, X_i, Z_i, \theta_i) \\ = & \prod_{j=1}^{J}\left[\prod_{m=1}^{n_{ij}} \lambda_j(t_{ijm};\eta_j) e^{\beta_j^T X_{ij}(t_{ijm})+\theta_i^T A_j^T Z_{ij}(t_{ijm})} \times \right. \\ & \left. \exp\left\{-\int_0^\tau \lambda_j(u;\eta_j) e^{\beta_j^T X_{ij}(u)+\theta_i^T A_j^T Z_{ij}(u)} Y_i(u)du\right\}\right] \times \phi(\theta_i;0,\gamma) \end{aligned}$$

(iii) Sample $U_i \sim U(0,1)$. Set $\theta_i^{(t+1)} = \theta_i^*$ if $U_i < r_i$ and $\theta_i^{(t+1)} = \theta^{(t)}$ otherwise.

(2) M-step via coordinate descent algorithm: maximize

$$\sum_{i=1}^{n} \log L_c(\alpha|N_i, X_i, Z_i, \theta_i^{(t+1)}) - n\left\{\sum_{j=1}^{J}\sum_{l=1}^{L} p_\gamma(\beta_{jl}) + \sum_{j=1}^{J}\sum_{l=1}^{L}\sum_{k=1}^{K} p_\gamma(a_{jlk})\right\}. \tag{2.5}$$

Denote

$$\begin{aligned} Q_j(\eta_j, \beta_j, A_j|\theta^{(t+1)}) = & \sum_{i=1}^{n}\left[\sum_{m=1}^{n_{ij}}\left\{\log\lambda_j(t_{ijm};\eta_j) + \beta_j^T X_{ij}(t_{ijm}) + (\theta_i^{(t+1)})^T A_j^T Z_{ij}(t_{ijm})\right\}\right. \\ & \left. - \int_0^\tau \lambda_j(u;\eta_j) e^{\beta_j^T X_{ij}(u)+\theta_i^T A_j^T Z_{ij}(u)} Y_i(u)du\right] \end{aligned}$$

It is clear that maximizing (2.5) is equivalent to maximizing the following terms separately:

$$Q_j(\eta_j, \beta_j, A_j|\theta^{(t+1)}) - n\left\{\sum_{l=1}^{L} p_\gamma(\beta_{jl}) + \sum_{l=1}^{L}\sum_{k=1}^{K} p_\gamma(a_{jlk})\right\}, \quad \text{for } j = 1, \ldots, J, \tag{2.6}$$

and

$$\sum_{i=1}^{n} \log\phi(\theta_i^{(t+1)};0,\gamma).$$

To maximize (2.6), we apply the coordinate descent algorithm to update each parameter. In each update, we form a quadratic approximation of Q_j with respect to that parameter at the current

value. In addition, we apply local linear approximation [55] to the scad penalty:

$$p_\gamma(|x|) \approx p_\gamma(|x_0|) + p'_\gamma(|x_0|)(|x| - |x_0|) \quad \text{for } x \approx x_0.$$

The resulting univariate maximization problem has a closed form solution. Specifically, we first update η_j by

$$\eta_j^{(t+1)} \leftarrow \eta_j^{(t)} - \frac{\partial_{\eta_j} Q_j(\eta_j^{(t)}, \beta_j^{(t)}, A_j^{(t)} | \theta^{(t+1)})}{\partial^2_{\eta_j} Q_j(\eta_j^{(t)}, \beta_j^{(t)}, A_j^{(t)} | \theta^{(t+1)})},$$

where ∂Q_j and $\partial^2 Q_j$ denote the first and second derivatives of Q with respect to the parameter η_j, β_{jl} or a_{jkl} as labeled by the subscripts. Denote $\beta_j^{(t,l)} = (\beta_{j1}^{(t+1)}, \ldots, \beta_{j,l-1}^{(t+1)}, \beta_{jl}^{(t)}, \ldots, \beta_{jL_j}^{(t)})$. To maximize β_{jl}, Update β_{jl} by

$$\beta_{jl}^{(t+1)} \leftarrow -\frac{S(\partial_{\beta_{jl}} Q_j(\eta_j^{(t+1)}, \beta_j^{(t,l)}, A_j^{(t)} | \theta^{(t+1)}) - \beta_{jl}^{(t)} \partial^2_{\beta_{jl}} Q_j(\eta_j^{(t+1)}, \beta_j^{(t,l)}, A_j^{(t)} | \theta^{(t+1)}), p'_\gamma(|\beta_j^{(t)}|))}{\partial^2_{\beta_{jl}} Q_j(\eta_j^{(t+1)}, \beta_j^{(t,l)}, A_j^{(t)} | \theta^{(t+1)})},$$

where S is the soft threshold operator ([13]) defined as $S(x, \gamma) = \text{sgn}(x)(|x| - \gamma)_+$. The updating procedure of α_{jlk} is similar to that of β_{jl} and is omitted.

2.7.3 Proofs for Theoretical Results

To prove Theorems 5-8, it suffices to verify Conditions (C1)-(C8) in [52] is satisfied under our regularity conditions (D1)-(D5). Our Theorems 5-8 then follow from Theorems 1-4 in [52]. Denote $B_j = (\beta_j, A_j)$. Denote

$$\Psi(O_i; \alpha, \mathcal{A}) = \int_\theta \prod_{j=1}^{J} \Omega_{ij}(\theta; B_j, \Lambda_j) \phi(\theta; \gamma) d\theta,$$

where

$$\begin{aligned}\Omega_{ij}(\theta; B_j, \Lambda_j) &= \prod_{t\le\tau}\left\{Y_{ij}(t)e^{\beta_j^T X_{ij}(t)+\theta^T A_j^T Z_{ij}(t)}\right\}^{dN^*_{ij}(t)} e^{-q_{ij}(\tau)},\\ q_{ij}(t) &= \int_0^t Y_{ij}(s)e^{\beta_j^T X_{ij}(s)+\theta^T A_j^T Z_{ij}(s)}d\Lambda_j(s).\end{aligned}$$

The likelihood can be written as

$$\prod_{i=1}^{n}\left\{\prod_{j=1}^{J}\prod_{t\le\tau}\Lambda_j\{t\}^{R_{ij}(t)dN^*_{ij}(t)}\right\}\Psi(O_i;\alpha,\mathcal{A}).$$

Let $\dot{\Psi}_\alpha$ denote the derivative of Ψ with respect to α. Let $\dot{\Psi}_j[H_j]$ denote the derivative of $\Psi(O_i;\alpha,\mathcal{A})$ along the path $(\Lambda_j+\varepsilon H_j)$, where H_j belongs to the set of functions in which $\Lambda_j+\varepsilon H_j$ is increasing with bounded total variation. That is,

$$\dot{\Psi}_j[H_j] = \lim_{\varepsilon\to 0}\frac{\Psi(O_i;\alpha,\mathcal{A}+(0,\ldots,\varepsilon H_j,\ldots,0))-\Psi(O_i;\alpha,\mathcal{A})}{\varepsilon}.$$

For easier reference, we list Conditions in (C1)-(C8) in [52] in terms of the current setting here.

(C1) The true value α_0 lies in the interior of a compact set Θ, and the true functions Λ_{0j} are continuously differentiable in $[0,\tau]$ with $\Lambda'_{0j}(t) > 0$, $j = 1,\ldots,J$.

(C2) With probability 1, $\mathbb{P}(\inf_{s\in[0,t]} Y_{ij}(s) \ge 1 | X_{ij}, Z_{ij}) > \delta_0 > 0$ for all $t \in [0,\tau]$.

(C3) There exist a constant $c_1 > 0$ and a random variable $r_1(O_i) > 0$ such that $\mathbb{E}(\log r_1(O_i)) < \infty$ and for any $\alpha \in \Theta$ and any finite $\Lambda_1,\ldots,\Lambda_J$,

$$\Psi(O_i;\alpha,\mathcal{A}) \le r_1(O_i)\prod_{j=1}^{J}\prod_{t\le\tau}\left\{1+\int_0^t Y_{ij}(t)d\Lambda_j(t)\right\}^{-dN^*_{ij}(t)}\left\{1+\int_0^\tau Y_{ij}(t)d\Lambda_j(t)\right\}^{-c_1}$$

almost surely. In addition, for any constant c_2,

$$\inf\{\Psi(O_i,\alpha,\mathcal{A}) : ||\Lambda_j||_{V[0,\tau]} \le c_2, j = 1,\ldots,J, \alpha\in\Theta\} > r_2(O_i) > 0,$$

where $r_2(O_i)$, which may depend on c_2, is a finite random variable with $\mathbb{E}(|\log r_2(O_i)|) < \infty$.

(C4) For any $(\alpha^{(1)}, \alpha^{(2)}) \in \Theta$, and $(\Lambda_1^{(1)}, \Lambda_1^{(2)}), \ldots, (\Lambda_J^{(1)}, \Lambda_J^{(2)}), (H_1^{(1)}, H_1^{(2)}), \ldots, (H_J^{(1)}, H_J^{(2)})$ with uniformly bounded total variations, there exists a random variable $\mathcal{F}(O_i) \in L_4(\mathbb{P})$ and J stochastic processes $\mu_{ij}(t; O_i) \in L_6(\mathbb{P})$, $j = 1, \ldots, J$ such that

$$
\begin{aligned}
&|\Psi(O_i; \alpha^{(1)}, \mathcal{A}^{(1)}) - \Psi(O_i; \alpha^{(2)}, \mathcal{A}^{(2)})| + |\dot{\Psi}_\alpha(O_i; \alpha^{(1)}, \mathcal{A}^{(1)}) - \dot{\Psi}_\alpha(O_i; \alpha^{(2)}, \mathcal{A}^{(2)})| \\
&\quad + \sum_{j=1}^{J} |\dot{\Psi}_j(O_i; \alpha^{(1)}, \mathcal{A}^{(1)})[H_k^{(1)}] - \dot{\Psi}_j(O_i; \alpha^{(2)}, \mathcal{A}^{(2)})[H_j^{(2)}]| \\
&\qquad + \sum_{j=1}^{J} \left| \frac{\dot{\Psi}_j(O_i; \alpha^{(1)}, \mathcal{A}^{(1)})[H_k^{(1)}]}{\Psi(O_i; \alpha^{(1)}, \mathcal{A}^{(1)})} - \frac{\dot{\Psi}_j(O_i; \alpha^{(2)}, \mathcal{A}^{(2)})[H_k^{(2)}]}{\Psi(O_i; \alpha^{(2)}, \mathcal{A}^{(2)})} \right| \\
\leq \;& \mathcal{F}(O_i) \Bigg[|\alpha^{(1)} - \alpha^{(2)}| + \sum_{j=1}^{J} \Bigg\{ \int_0^\tau |\Lambda_j^{(1)}(s) - \Lambda_j^{(2)}(s)| d\mu_{ij}(s; O_i) \\
&\quad + \int_0^\tau |H_j^{(1)}(s) - H_j^{(2)}(s)| d\mu_{ij}(s; O_i) \Bigg\} \Bigg].
\end{aligned}
$$

In addition, $\mu_{ij}(s; O_i)$ is non-decreasing, and $\mathbb{E}[\mathcal{F}(O_i)\mu_{ij}(s; O_i)]$ is bounded and left-continuous with uniformly bounded left- and right-derivatives for any $s \in [0, \tau]$.

(C5) The model is identifiable. That is, if

$$
\begin{aligned}
&\int_\theta \left[\prod_{j=1}^{J} \prod_{t \leq \tau} \{\lambda_j(t) e^{\beta_j^T X_{ij}(t) + \theta^T A_j^T Z_{ij}(t)}\}^{dN_{ij}(t)} e^{-\int_0^\tau e^{\beta_j^T X_{ij}(t) + \theta^T A_j^T Z_{ij}(t)} Y_{ij}(t) d\Lambda_j(t)} \right] \phi(\theta; \gamma) d\theta \\
= \;& \int_\theta \left[\prod_{j=1}^{J} \prod_{t \leq \tau} \{\lambda_{0j}(t) e^{\beta_{0j}^T X_{ij}(t) + \theta^T A_{0j}^T Z_{ij}(t)}\}^{dN_{ij}(t)} e^{-\int_0^\tau e^{\beta_{0j}^T X_{ij}(t) + \theta^T A_{0j}^T Z_{ij}(t)} Y_{ij}(t) d\Lambda_{0j}(t)} \right] \phi(\theta; \gamma_0) d\theta
\end{aligned}
$$

almost surely, then we have $(\alpha, \mathcal{A}) = (\alpha_0, \mathcal{A}_0)$.

Note that condition (C4) implies that the linear functional

$$
H_j \mapsto \mathbb{E}\left[\frac{\dot{\Psi}_j(O_i; \alpha, \mathcal{A})[H_j]}{\Psi(O_i; \alpha, \mathcal{A})} \right]
$$

is continuous from $BV[0,\tau]$ to $\mathbb{R}$. Thus, there exists a bounded function $\eta_{0j}(s;\alpha,\mathcal{A})$ such that

$$\mathbb{E}\left[\frac{\dot{\Psi}_j(O_i;\alpha,\mathcal{A})[H_j]}{\Psi(O_i;\alpha,\mathcal{A})}\right] = \int_0^\tau \eta_{0j}(s;\alpha,\mathcal{A})dH_j(s).$$

(C6) There exist functions $\zeta_{0j}(s;\alpha_0,\mathcal{A}_0) \in BV[0,\tau]$, $j = 1,\ldots,J$, and a matrix $\zeta_{0\alpha}(\alpha_0,\mathcal{A}_0)$ such that

$$\left|\mathbb{E}\left[\frac{\dot{\Psi}_\alpha(O_i;\alpha,\mathcal{A})}{\Psi(O_i;\alpha,\mathcal{A})} - \frac{\dot{\Psi}_\alpha(O_i;\alpha_0,\mathcal{A}_0)}{\Psi(O_i;\alpha_0,\mathcal{A}_0)}\right] - \zeta_{0\alpha}(\alpha_0,\mathcal{A}_0)(\alpha-\alpha_0) - \sum_{j=1}^J \int_0^\tau \zeta_{0j}(s;\theta_0,\mathcal{A}_0)d(\Lambda_j - \Lambda_{0j})\right| = o\left(|\alpha-\alpha_0| + \sum_{j=1}^J ||\Lambda_j - \Lambda_{0j}||_{V[0,\tau]}\right).$$

In addition, for $j = 1,\ldots,J$,

$$\sum_{j=1}^J \sup_{s\in[0,\tau]}\left|\{\eta_{0j}(s;\alpha,\mathcal{A}) - \eta_{0j}(s;\alpha_0,\mathcal{A}_0)\} - \eta_{0j\theta}(s;\alpha_0,\mathcal{A}_0)(\alpha-\alpha_0) - \int_0^\tau \sum_{j=1}^J \eta_{0jm}(s,t;\alpha_0,\mathcal{A}_0)d(\Lambda_m - \Lambda_{0m})(t)\right| = o\left(|\alpha-\alpha_0| + \sum_{j=1}^J ||\Lambda_j - \Lambda_{0j}||_{V[0,\tau]}\right),$$

where η_{0jm} is a bounded bivariate function and $\eta_{0j\alpha}$ is a d-dimensional bounded function. Furthermore, there exists a constant c_3 such that

$$|\eta_{0jm}(s,t_1;\alpha_0,\mathcal{A}_0) - \eta_{0jm}(s,t_2;\alpha_0,\mathcal{A}_0)| \le c_3|t_1 - t_2|$$

for any $s \in [0,\tau]$ and any $t_1, t_2 \in [0,\tau]$.

(C7) If for some $v \in \mathbb{R}^d$ and $h_j \in BV[0,\tau]$, $j = 1,\ldots,J$

$$\sum_{j=1}^J \int h_j(t)Y_{ij}(t)dN_{ij}^*(t) + \frac{\dot{\Psi}_\alpha(O_i;\alpha_0,\mathcal{A}_0)^T v + \sum_{j=1}^J \dot{\Psi}_j(O_i;\alpha_0,\mathcal{A}_0)[\int h_j d\Lambda_{0j}]}{\Psi(O_i;\alpha_0,\mathcal{A}_0)} = 0$$

almost surely, then $v = 0$ and $h_j = 0$ for $j = 1,\ldots,J$.

(C8) There exists a neighborhood of $(\alpha_0, \mathcal{A}_0)$ such that for $(\alpha, \mathcal{A})$ in this neighborhood, the first and second derivatives of $\log \Psi(O_i; \alpha, \mathcal{A})$ with respect to α and along the path $\Lambda_j + \varepsilon H_j$ with respect to ε satisfy the inequality in (C4).

Proposition 1. *The model is identifiable. That is, if*

$$\begin{aligned}
&\int_\theta \left[\prod_{j=1}^{J} \prod_{t \le \tau} \{\lambda_{j0}(t) e^{\beta_j^T X_{ij}(t) + \theta^T A_j^T Z_{ij}(t)}\}^{dN_{ij}(t)} e^{-\int_0^\tau e^{\beta_j^T X_{ij}(t)} + \theta^T A_j^T Z_{ij}(t) Y_{ij}(t) d\Lambda_j(t)} \right] \phi(\theta; \gamma) d\theta \\
= &\int_\theta \left[\prod_{j=1}^{J} \prod_{t \le \tau} \{\lambda_{0j0}(t) e^{\beta_{0j}^T X_{ij}(t) + \theta^T A_{0j}^T Z_{ij}(t)}\}^{dN_{ij}(t)} e^{-\int_0^\tau e^{\beta_{0j}^T X_{ij}(t)} + \theta^T A_{0j}^T Z_{ij}(t) Y_{ij}(t) d\Lambda_{0j}(t)} \right] \phi(\theta; \gamma_0) d\theta
\end{aligned}$$

almost surely, then we have $(\alpha, \mathcal{A}) = (\alpha_0, \mathcal{A}_0)$.

Remark 1. *λ_{j0} and λ_{0j0} are used to denote the general and true baseline intensity function for event type j respectively while λ_j is used to denote the intensity function for event type j.*

Proof of Proposition 1. Fix $k_{0j}, k_{1j} \in \mathbb{N}$. Consider $Y_j \equiv 1$ and event times $\{t_{j11}, \ldots, t_{j1k_1}\}$ and $\{t_{j1}, \ldots, t_{jk_0}\}$ for event type j, for $j = 1 \ldots, J$. Then,

$$\begin{aligned}
&\int_\theta \prod_{j=1}^{J} \left\{ \prod_{k=1}^{k_{1j}} \lambda_j(t_{j1k} | \alpha, \mathcal{A}, \theta) \prod_{k=1}^{k_{0j}} \lambda_j(t_{jk} | \alpha, \mathcal{A}, \theta) e^{-\int_0^\tau \lambda_j(t | \alpha, \mathcal{A}, \theta) dt} \right\} \phi(\theta, \gamma) d\theta \\
= &\int_\theta \prod_{j=1}^{J} \left\{ \prod_{k=1}^{k_{1j}} \lambda_j(t_{j1k} | \alpha_0, \mathcal{A}_0, \theta) \prod_{k=1}^{k_{0j}} \lambda_j(t_{jk} | \alpha_0, \mathcal{A}_0, \theta) e^{-\int_0^\tau \lambda_j(t | \alpha_0, \mathcal{A}_0, \theta) dt} \right\} \phi(\theta, \gamma_0) d\theta.
\end{aligned}$$

Integrating $t_{j11}, \ldots, t_{j1k_{1j}}$ from 0 to t_j for $j = 1, \ldots, J$ and integrating $t_{j11}, \ldots, t_{jk_0}$ from 0 to τ for $j = 1, \ldots, J$, we have

$$\begin{aligned}
&\int_\theta \prod_{j=1}^{J} \left[\left\{ \int_0^{t_j} \lambda_j(t | \alpha, \mathcal{A}, \theta) dt \right\}^{k_{1j}} \left\{ \int_0^{T} \lambda_j(t | \alpha, \mathcal{A}, \theta) dt \right\}^{k_{0j}} e^{-\int_0^\tau \lambda_j(t | \alpha, \mathcal{A}, \theta) dt} \right] \phi(\theta, \gamma) d\theta \\
= &\int_\theta \prod_{j=1}^{J} \left[\left\{ \int_0^{t_j} \lambda_j(t | \alpha_0, \mathcal{A}_0, \theta) dt \right\}^{k_{1j}} \left\{ \int_0^{T} \lambda_j(t | \alpha_0, \mathcal{A}_0, \theta) dt \right\}^{k_{0j}} e^{-\int_0^\tau \lambda_j(t | \alpha_0, \mathcal{A}_0, \theta) dt} \right] \phi(\theta, \gamma_0) d\theta.
\end{aligned}$$

Multiply both sides by $\prod_{j=1}^{J} [\frac{(\mathrm{i} s_j)^{k_{1j}!}}{k_{1j}!} \frac{1}{k_{0j}!}]$, where i is the imaginary number and s_j's are arbitrary

real numbers. Summing over $k_{1j} = 0, 1, 2, \ldots$ and $k_{0j} = 0, 1, 2, \ldots$, we get

$$\int_\theta \prod_{j=1}^{J} e^{\mathrm{i}s_j \int_0^{t_j} \lambda_j(t|\alpha,\mathcal{A},\theta)dt} \phi(\theta;\gamma)d\theta = \int_\theta \prod_{j=1}^{J} e^{\mathrm{i}s_j \int_0^{t_j} \lambda_j(t|\alpha_0,\mathcal{A}_0,\theta)dt} \phi(\theta;\gamma_0)d\theta.$$

Since this holds for any s_j, the distribution of $\{\int_0^{t_j} \lambda_j(t;\alpha,\mathcal{A},\theta)dt\}_{j=1,\ldots,J}$ and $\{\int_0^{t_j} \lambda_j(t;\alpha_0,\mathcal{A}_0,\theta_0)dt\}_{j=1,\ldots,J}$ are the same, where $\theta \sim N(0,\gamma)$ and $\theta_0 \sim N(0,\gamma_0)$. Therefore, $\{\log \lambda_j(t_j;\alpha,\mathcal{A},\theta)\}_{j=1,\ldots,J}$ and $\{\log \lambda_j(t_j;\alpha_0,\mathcal{A}_0,\theta_0)\}_{j=1,\ldots,J}$ have the same distribution. By considering the mean of $\log \lambda_j(t_j;\alpha,\mathcal{A},\theta)$ and $\log \lambda_j(t_j;\alpha_0,\mathcal{A}_0,\theta_0)$, we have $\lambda_{j0}(t_j)+\beta_j^T X_j(t_j) = \lambda_{0j0}(t_j)+\beta_{0j}{}^T X_j(t_j)$. By Condition (D4), we have $\lambda_{j0}(t) = \lambda_{0j0}(t)$ and $\beta_j = \beta_{0j}$ for all $j = 1, \ldots, J$. Then $\{\theta^T A_j^T Z_j(t_j)\}_{j=1,\ldots,J}$ has the same distribution as $\{\theta_0^T A_{0j}{}^T Z_j(t_j)\}_{j=1,\ldots,J}$. By considering the covariance matrices of these two random vectors, we have for each $j, l = 1, \ldots, J$,

$$Z_j^T(t_j) A_j \Sigma A_l^T Z_l(t_l) = Z_j^T(t_j) A_{0j} \Sigma_0 A_{0l}^T Z_l(t_l).$$

Let $B = A_j \Sigma A_l^T - A_{0j}\Sigma_0 A_{0l}^T$. We have $Z_j^T(t_j) B Z_l(t_l) = 0$. Condition (D4) then implies $B = 0$. Hence, we have $A_j \Sigma A_l^T = A_{0j}\Sigma_0 A_{0l}^T$ for all $j, l = 1, \ldots, J$. This is equivalent to $A\Sigma A^T = A_0 \Sigma A_0^T$. By Condition (D5), it is assumed that there is an $K \times K$ identity matrix in A. Without loss of generality, assume $A = (I, A_2^T)^T$ and $A_0 = (I, A_{02}^T)^T$. Hence, we have

$$I\Sigma I^T = I\Sigma_0 I^T \tag{2.7}$$

and

$$I\Sigma A_2^T = I\Sigma_0 A_{02}^T. \tag{2.8}$$

Therefore, $\Sigma = \Sigma_0$ and then $A_2 = A_{02}$, showing $A = A_0$. □

Proposition 2. *If for $v \in \mathbb{R}^d$ and $h_j \in BV[0,\tau]$, $j = 1, \ldots, J$, the score function*

$$\sum_{j=1}^{J} \int_0^\tau h_j(t) Y_{ij}(t) dN_{ij}^*(t) + \frac{v^T \dot{\Psi}_\alpha(O_i;\alpha_0,\mathcal{A}_0)}{\Psi(O_i;\alpha_0,\mathcal{A}_0)} + \sum_{j=1}^{J} \frac{\dot{\Psi}_j(O_i;\alpha_0,\mathcal{A}_0)[\int h_j d\Lambda_{0j}]}{\Psi(O_i;\alpha_0,\mathcal{A}_0)} = 0.$$

almost surely, then $v = 0$ *and* $h_j = 0$ *for* $j = 1, \ldots, J$.

Proof. We shall show that the model with $\lambda_j(t) = \lambda_{0j0}(t)e^{\beta_{0j}^T X_j(t)+\theta_j^T Z_j(t)}$, $\theta = (\theta_1^T, \ldots, \theta_J^T)^J \sim N(0, A_0\Sigma_0 A_0^T)$, and $A_0 = (A_{01}^T, \ldots, A_{0J}^T)^T$ satisfies the claim in this proposition. The result then follows by noting such model is simply a reparameterization of the proposed model. By an abuse of notation, we continue to denote the parameters in the distribution of θ_i by γ. Consider $Y_j = 1$ and observed event times $t_{j1}, \ldots, t_{jM}$ for the jth event type for $j = 1, \ldots, J$. From the score equation, by straightforward algebra, we have

$$\begin{aligned}
&\int_\theta \prod_{j=1}^{J}\prod_{m=1}^{M} e^{\beta_{0j}^T X_j(t_{jm})+\theta_j^T Z_j(t_{jm})} e^{-\int_0^\tau e^{\beta_{0j}^T X_j(s)+\theta_j^T Z_j(s)} d\Lambda_{0j0}(s)} \phi(\theta;\gamma_0) \\
&\quad \times \Bigg(\sum_{j=1}^{J}\sum_{m=1}^{M} \{h(t_{jm}) + X_j^T(t_{jm})v_{\beta_j}\} + \frac{\phi'(\theta;\gamma_0)^T v_\gamma}{\phi(\theta;\gamma_0)} \\
&\qquad - \sum_{j=1}^{J} \Bigg[\int_0^\tau \{h_j(s) + X_j^T(t)v_{\beta_j}\} e^{\beta_{0j}^T X_j(s)+\theta_j^T Z_j(s)} d\Lambda_{0j0}(s) \Bigg] d\theta = 0.
\end{aligned}$$

Now, we perform the following operations on both sides of the above equation. First, multiply both sides by $\prod_{j=1}^{J}\prod_{m=1}^{M} \lambda_{0j0}(t_{jm})$ and integrate t_{jm} from 0 to t_{jm} for $m = 1, \ldots, m_{0j}$ and from 0 to τ for $m = m_{0j} + 1, \ldots, M$. Then, divide the resulting equation by $(M - m_{0j})!$. After we sum over $M - m_{0j} = 0, 1, 2, \ldots$, we obtain

$$\int_\theta \prod_{j=1}^{J}\prod_{m=1}^{m_{0j}} H_{2j}(\theta, t_{jm})\phi(\theta;\gamma_0) \Bigg(\sum_{j=1}^{J}\sum_{m=1}^{m_{0j}} F_{2j}(\theta, t_{jm}) + \frac{\phi'(\theta;\gamma_0)^T v_\gamma}{\phi(\theta;\gamma_0)} \Bigg) d\theta = 0, \tag{2.9}$$

where

$$\begin{aligned}
H_{2j}(\theta, t) &:= \int_0^t e^{\beta_{0j}^T X_j(s)+\theta_j^T Z_j(t)} d\Lambda_{0j0}(s), \\
F_{2j}(\theta, t) &:= \frac{\int_0^t \{h_j(s) + X_j^T(s)v_{\beta_{0j}}\} e^{\beta_{0j}^T X_j(s)+\theta_j^T Z_j(s)} d\Lambda_{0j0}(s)}{\int_0^t e^{\beta_{0j}^T X_j(s)+\theta_j^T Z_j(s)} d\Lambda_{0j0}(s)}.
\end{aligned}$$

Letting t_{jm} has multiplicity k_{jm} and multiply both sides of the resulting equation by

$\prod_{j=1}^{J}\prod_{m=1}^{m_{0j}}(\mathrm{i}s_{jm})^{k_{jm}}/k_{jm}!$, we have

$$\int_{\theta}\prod_{j=1}^{J}\prod_{m=1}^{m_{0j}}\frac{(\mathrm{i}s_{jm})^{k_{jm}}}{k_{jm}!}H_{2j}(\theta,t_{jm})^{k_{jm}}\phi(\theta;\gamma_0)\left(\sum_{j=1}^{J}\sum_{m=1}^{m_{0j}}k_{jm}F_{2j}(\theta,t_{jm})+\frac{\phi'(\theta;\gamma_0)^T v_\gamma}{\phi(\theta;\gamma_0)}\right)d\theta=0.$$

Summing over $k_{jm}=1,2,\ldots,$ for $j=1,\ldots,J$, $m=1,\ldots,m_{0j}$, we have

$$\int_{\theta}\left\{\sum_{j=1}^{J}\sum_{m=1}^{m_{0j}}\mathrm{i}s_{jm}F_{2j}(\theta,t_{jm})H_2(\theta,t_{jm})+\frac{\phi'(\theta;\gamma_0)^T v_\gamma}{\phi(\theta;\gamma_0)}\right\}e^{\sum_{j=1}^{J}\sum_{m=1}^{m_{0j}}\mathrm{i}s_{jm}H_{2j}(\theta,t_{jm})}\phi(\theta;\gamma_0)d\theta=0,$$

for any s_{jm}. By making the variable transformation $\{y_{jm}: j=1,\ldots,J, m=1,\ldots,L_j\}=\{H_2(\theta,t_{jm}): j=1,\ldots,J, m=1,\ldots,L_j\}$ and using the relationship between the Fourier transform of a function and and the Fourier transform of its derivative, we see that

$$-\sum_{j=1}^{J}\sum_{m=1}^{m_{0j}}F_2(\theta,t_{jm})H_2(\theta,t_{jm})\phi(\theta;\gamma_0)+\phi'(\theta;\gamma_0)^T v_\gamma=0$$

almost everywhere. By letting t_{jm}'s go to 0, we obtain $\phi'(\theta;\gamma_0)^T v_\gamma=0$. By the identifiability of $\theta\sim N(0,A_0\Sigma_0A_0^T)$, we have $v_\gamma=0$. Then, (2.9) with only one t_{jm} is a homogeneous equation for $(h_j(t)+X_j(t)^T v_{\beta_j})$. It is easy to see that the equation has only a trivial solution. Therefore, $h_j(t)+X_j(t)^T v_{\beta_j}=0$. By Condition (D4), we have $h_j=0$ and $v_{\beta_j}=0$ for each $j=1,\ldots,J$. □

Now, we verify Conditions (C1)-(C8).

(i) Condition (C1) follows from Condition (D1)

(ii) Condition (C2) follows from Condition (D2)

(iii) To verify Condition (C3), note that

$$\Omega_{ij}(\theta;B_j,\Lambda_j)\le e^{O(1)(1+|\theta|)N_{ij}^*(\tau)}e^{-q_{ij}(\tau)}.$$

Note that there exist $\mu>0$ and $\kappa>0$ such that for any $m\in\mathbb{N}$ and $0<x_1\le\ldots\le x_m\le y$,

we have $\prod_{i=1}^{m}(1+x_i)e^{-y} \leq \mu^m(1+y)^{-\kappa}$. Thus,

$$\Omega_{ij}(\theta; B_j, \Lambda_j) \leq e^{O(1)(1+|\theta|)N^*_{ij}(\tau)}\mu^{N^*_{ij}(\tau)}\prod_{t\leq\tau}\{1+q_{ij}(t)\}^{-dN^*_{ij}(t)}\{1+q_{ij}(\tau)\}^{-\kappa}.$$

Note that $e^{\beta_j^T X_{ij}(s)+\theta^T A_j^T Z_{ij}(s)} \geq e^{-O(1)(1+|\theta|)}$. Hence,

$$1+q_{ij}(t) \geq e^{-m(1+|\theta|)}\left\{1+\int_0^t Y_{ij}(s)d\Lambda_j(s)\right\}$$

so that

$$\Omega_{ij}(\theta; B, \mathcal{A}) \leq e^{M(1+|\theta|)N^*_{ij}(\tau)}\mu^{N^*_{ij}(\tau)}\prod_{t\leq\tau}\left\{1+\int_0^t Y_{ij}(s)d\Lambda_j(s)\right\}^{-dN^*_{ij}(t)}\left\{1+\int_0^\tau Y_{ij}(s)d\Lambda_j(s)\right\}^{-\kappa}.$$

Therefore, condition (C1) holds with $r_1(O_i) = \mu^{\sum_{j=1}^J N^*_{ij}(\tau)}\int e^{M(1+|\theta|)\sum_{j=1}^J N^*_{ij}(\tau)}\phi(\theta;\gamma)d\theta$. Clearly, $\mathbb{E}(\log r_1(O_i)) < \infty$. To verify the second part of condition (C2), let $B_0 > 0$ be a fixed constant. Then, for $||\Lambda_j||_{V[0,\tau]} \leq c_2$, $j = 1, \ldots, J$,

$$\begin{aligned}
\Psi(O_i; \alpha, \mathcal{A}) &= \int_\theta \prod_{j=1}^J \Omega_{ij}(\theta; B, \Lambda_j)\phi(\theta;\gamma)d\theta \\
&\geq \int_{\{||\theta||\leq B_0\}} \prod_{j=1}^J \Omega_{ij}(\theta; B, \Lambda_j)\phi(\theta;\gamma)d\theta \\
&\geq \exp\{-O(1)N^*_{ij}(\tau)\}\mathbb{P}(|\theta| \leq B_0).
\end{aligned}$$

Therefore, the second part of condition (C2) is satisfied.

(iv) Note that for any B_j, $\Omega_{ij}(\theta; B_j, \Lambda_j) \leq e^{O(1)(1+|\theta|)N^*_{ij}(\tau)}$ and

$$\begin{aligned}
\left|\frac{\partial}{\partial\beta_j}\Omega_{ij}(\theta; B_j, \Lambda_j)\right| &= \left|\Omega_{ij}(\theta; \beta_j, \Lambda_j)\left\{\int Y_{ij}(t)X_{ij}(t)dN^*_{ij}(t)\right.\right. \\
&\quad \left.\left. - \int_0^\tau Y_{ij}(s)e^{\beta_j^T X_{ij}(s)+\theta^T A_j^T Z_{ij}(s)}X_{ij}(s)d\Lambda_j(s)\right\}\right| \\
&\leq e^{O(1)(1+|\theta|)(1+N^*_{ij}(\tau))}.
\end{aligned}$$

Similarly, we have

$$\left|\frac{\partial}{\partial A_j}\Omega_{ij}(\theta; B_j, \Lambda_j)\right| \leq e^{M(1+|\theta|)(1+N^*_{ij}(\tau))}.$$

Also,

$$\begin{aligned}\left|\frac{\partial}{\partial \Lambda_j}\Omega_{ij}(\theta; B_j, \Lambda_j)[H_j]\right| &= \left|-\Omega_{ij}(\theta; B_j, \Lambda_j)\int_0^\tau Y_{ij}(s)e^{\beta^T Z_{ij}(s)+\theta^T A_j(s)Z_{ij}(s)}dH_j(s)\right| \\ &\leq e^{O(1)(1+|\theta|)(1+N^*_{ij}(\tau))}.\end{aligned}$$

By the mean value theorem,

$$\begin{aligned}|\Omega_{ij}(\theta; B_j^{(1)}, \Lambda_j) - \Omega_{ij}(\theta; B_j^{(2)}, \Lambda_j)| &= \left|\frac{\partial}{\partial B_j}\Omega_{ij}(\theta; B_j^*, \Lambda_j)\right||B_j^{(1)} - B_j^{(2)}| \\ &\leq e^{O(1)(1+|\theta|)(1+N^*_{ij}(\tau))}|B_j^{(1)} - B_j^{(2)}|,\end{aligned}$$

and

$$\begin{aligned}&|\Omega_{ij}(\theta; B_j, \Lambda_j^{(1)}) - \Omega_{ij}(\theta; B_j, \Lambda_j^{(2)})| = \left|\frac{\partial}{\partial \Lambda_j}\Omega_{ij}(\theta; B_j, \Lambda_j^*)[\Lambda_j^{(1)} - \Lambda_j^{(2)}]\right| \\ \leq\ & e^{M(1+|\theta|)N^*_{ij}(\tau)}\left|\int_0^\tau Y_{ij}(s)e^{\beta^T X_{ij}(s)+\theta^T A_j(s)Z_{ij}(s)}d(\Lambda_j^{(1)} - \Lambda_j^{(2)})(s)\right| \\ \leq\ & e^{O(1)(1+|\theta|)(1+N^*_{ij}(\tau))}\int_0^\tau |\Lambda_j^{(1)}(s) - \Lambda_j^{(2)}(s)|ds,\end{aligned}$$

where the last inequality follows from integration by parts and the assumption that $X_{ij}(\cdot)$ and $Z_{ij}(\cdot)$ have bounded variations. Also,

$$\begin{aligned}&\left|\int_\theta \prod_{j=1}^J \Omega_{ij}(\theta; B_j; \Lambda_j)\phi(\theta; \gamma^{(1)})d\theta - \int_\theta \prod_{j=1}^J \Omega_{ij}(\theta; B_j; \Lambda_j)\phi(\theta; \gamma^{(2)})d\theta\right| \\ \leq\ & \left|\int e^{O(1)(1+|\theta|)\sum_{j=1}^J N^*_{ij}(\tau)}\frac{\partial\phi(\theta; \gamma^*)}{\partial\gamma}d\theta\right||\gamma^{(1)} - \gamma^{(2)}|.\end{aligned}$$

Using the same arguments, the other three terms in condition (C4) can be shown to satisfy the required bound.

(v) Condition (C5) is verified in Proposition 1.

(vi) To verify condition (C6), note that

$$\eta_{0j}(s;\alpha,\mathcal{A}) = -\mathbb{E}\Bigg[\int_\theta \frac{\prod_{m=1}^J \Omega_{im}(\theta;B_m,\Lambda_m)\phi(\theta;\gamma)}{\int_\theta \prod_{m=1}^J \Omega_{im}(\theta;B_m,\Lambda_m)\phi(\theta;\gamma)d\theta} Y_{ij}(s)e^{\beta_j^T X_{ij}(s)+\theta^T Z_{ij}(s)}d\theta\Bigg].$$

For $(\alpha,\mathcal{A})$ in a neighborhood of $(\alpha_0,\mathcal{A}_0)$,

$$\begin{aligned}&\left|\eta_{0j}(s;\alpha,\mathcal{A}) - \eta_{0j}(s;\alpha_0,\mathcal{A}_0) - \frac{\partial}{\partial\alpha}\eta_{0j}(s;\alpha_0,\mathcal{A}_0)^T(\alpha-\alpha_0) - \sum_{m=1}^J \frac{\partial\eta_{0j}}{\partial\Lambda_m}(s;\alpha_0,\mathcal{A}_0)[\Lambda_m-\Lambda_{0m}]\right| \\ = \;& o\left(|\alpha-\alpha_0| + \sum_{m=1}^J ||\Lambda_m-\Lambda_{0m}||_{V[0,\tau]}\right).\end{aligned}$$

Therefore, the second equation in (C6) will hold with $\eta_{0jm}(s,t;\theta_0,\mathcal{A}_0)$ being the derivative of η_{0j} with respect to Λ_m along the direction $\Lambda_m-\Lambda_{0m}$, and $\eta_{0j\alpha}$ begin the derivative of η_{0j} with respect to α. Straightforward calculation also yields the Lipschitz continuity of η_{0jm}. The verification of the first part in (C6) is similar.

(vii) Condition (C7) is verified in Proposition 2.

(viii) The verification of (C8) is similar to that of (C4) and is omitted.

The proofs of Theorems 9 and 10 follow the same argument as in [14] as we are assuming a parametric baseline intensity function in the penalized likelihood method. The only nonstandard ingredients that are model-dependent are the identifiability of the model and the invertibility of the information matrix, which can be verified as in Propositions 1 and 2 respectively.

2.7.4 Additional Simulation Results

Table 2.5 and Table 2.6 report the parameter setting, bias, average of the standard error estimates, estimated standard deviation of the parameters and the empirical coverage percentage of the 95% confidence interval.

Table 2.4: Partial results of the factor loading in the real data. The numbers outside the brackets are the estimated factor loadings and the numbers in the brackets are the estimated standard errors.

Covariate	Event	A1	A2	A3
Next	Next_C	0.81 (0.17)	-0.11 (0.19)	0.08 (0.23)
R_Open	R_1	0.71 (0.17)	0	0.74 (0.15)
R_Open	R_2	-0.79 (0.06)	0.39 (0.06)	0.16 (0.06)
R_Open	R_3	0	0.04 (0.12)	0
R_Open	R_4	0.51 (0.11)	0.57 (0.07)	0
R_Open	R_5	0.52 (0.23)	0	0
W1	Back	0	0.48 (0.06)	-0.1 (0.07)
W1	W1_M	-1.31 (0.12)	-0.03 (0.11)	0
W1, Back	W2	0.29 (0.09)	0.06 (0.07)	0.3 (0.09)
W1, Back	W3	0.75 (0.19)	0.03 (0.16)	-0.82 (0.24)
W1, Back	W4	0	0	-1.79 (0.35)
W2	Back	0	0.48 (0.04)	0
W2	W2_A	-2.12 (0.2)	0	0
W2, Back	W1	0	0	-1.02 (0.22)
W2, Back	W2	-0.04 (0.21)	0	-0.12 (0.17)
W2, Back	W3	0.97 (0.12)	0	0.41 (0.13)
W2, Back	W4	0.96 (0.17)	0.35 (0.14)	-0.33 (0.18)
W3	Back	-0.03 (0.07)	0.57 (0.05)	0
W3	W3_A	-2.55 (0.37)	-0.37 (0.25)	0
W3	W3_O1	-1.14 (0.31)	0	0
W3, Back	W2	-0.76 (0.22)	0	-1.31 (0.19)
W3, Back	W4	0.75 (0.14)	0.21 (0.1)	0.45 (0.13)
W3_O1	W3_O2	-0.97 (0.39)	0	0
W4	Back	-0.11 (0.06)	0.44 (0.05)	0.04 (0.07)
W4, Back	W2	-0.34 (0.14)	0	-0.64 (0.15)
W4, Back	W3	0	0	-1.46 (0.25)
W4, Back	W5	0.27 (0.15)	-0.05 (0.12)	1.14 (0.15)
W5	Back	-0.14 (0.06)	0.45 (0.06)	0.06 (0.06)
W5, Back	W1	0	0	0.17 (0.2)
W5, Back	W2	-0.34 (0.09)	-0.07 (0.08)	0
W5, Back	W3	0	0	-1.09 (0.23)
W5, Back	W4	0.53 (0.16)	0.34 (0.16)	-0.99 (0.2)

Table 2.5: Results of simulations. True, true value of the parameter; Bias, 100× {mean($\hat{\beta}$) - β_0}; SE, 100× average of standard error estimates, SD, 100× sample standard deviation; CP, empirical coverage percentage of the 95% confidence interval.

		n = 500				n = 1000				n = 2000			
	True	Bias	SE	SD	CP	Bias	SE	SD	CP	Bias	SE	SD	CP
1	-4	-1.87	5.45	5.39	0.95	-0.47	3.67	3.74	0.96	0.31	2.56	2.41	0.95
2	-5	-0.57	6.80	5.74	0.98	-1.33	4.53	4.12	0.97	-0.16	3.12	3.08	0.94
3	-5	-1.55	8.96	7.62	0.98	-0.48	6.12	6.22	0.96	-0.68	4.24	4.64	0.96
4	-5	-0.62	7.61	6.24	1.00	-0.32	5.06	4.37	0.99	-0.18	3.49	3.33	0.96
5	-5	2.18	8.87	8.14	0.95	-0.64	6.11	6.04	0.97	-0.64	4.26	4.85	0.93
6	-4	-0.01	5.45	4.71	0.98	0.09	3.65	3.70	0.96	0.39	2.51	2.86	0.90
7	-7	1.81	7.58	6.15	0.98	-0.24	5.12	4.87	0.96	0.33	3.56	3.19	0.94
8	-4	-2.29	13.39	13.41	0.98	-1.69	9.01	8.76	0.95	-1.81	6.23	6.03	0.96
9	-5	0.69	10.36	9.62	0.95	-0.62	6.97	7.12	0.95	-0.38	4.84	4.81	0.94
10	-5	-0.11	10.35	7.78	0.98	0.46	7.01	6.98	0.92	-0.10	4.83	4.08	0.98
11	-5	11.63	9.99	11.91	0.77	2.08	7.46	9.24	0.88	0.87	5.47	5.68	0.94
12	-6	0.35	4.80	4.51	0.97	-0.03	3.30	3.13	0.97	-0.17	2.28	2.26	0.94
13	-5	-0.25	10.63	8.37	1.00	-1.54	7.05	6.52	0.97	-0.48	4.87	5.08	0.93
14	-7	0.45	20.88	16.57	0.98	2.41	14.33	13.19	0.99	-0.14	10.05	9.71	0.96
15	-2	1.04	4.75	3.75	0.98	0.43	3.25	2.92	0.97	-0.01	2.27	2.09	0.98
16	1	2.50	7.07	7.32	0.93	-0.14	4.79	4.95	0.97	-0.48	3.34	3.18	0.93
17	-2	1.30	14.70	12.33	0.98	0.60	9.72	9.07	0.95	-0.51	6.67	6.70	0.95
18	-1	-0.56	10.35	9.47	0.98	1.53	6.97	6.75	0.96	1.09	4.68	4.87	0.95
19	1	-2.12	10.84	10.70	0.95	-0.90	7.46	7.26	0.95	0.21	5.15	5.50	0.94
20	2	1.57	9.37	9.39	0.90	2.26	6.26	6.51	0.92	-0.63	4.29	4.56	0.93
21	-2	32.79	25.34	22.82	0.72	13.33	17.21	19.37	0.83	2.99	11.73	13.99	0.89
22	1	-8.78	12.58	16.41	0.83	-3.74	8.48	10.17	0.86	-1.56	5.81	7.16	0.92
23	-2	25.51	23.61	31.84	0.75	11.25	15.89	20.37	0.79	3.53	11.02	12.63	0.88
24	2	9.20	11.37	14.14	0.82	5.51	7.53	9.58	0.82	1.63	5.17	5.18	0.95
25	-2	38.53	25.53	31.90	0.60	20.07	17.65	24.47	0.72	4.69	11.97	16.64	0.84
26	4	-0.48	17.05	14.40	0.98	1.11	11.35	10.78	0.96	0.87	7.89	7.85	0.93
27	5	2.96	13.16	11.38	0.98	-0.51	8.93	9.58	0.95	-0.80	6.24	5.95	0.96
28	5	-2.32	13.21	10.89	0.98	-2.07	8.98	9.58	0.92	-0.11	6.22	6.57	0.95
29	5	1.54	12.34	10.61	0.98	0.92	8.40	8.90	0.94	-0.46	5.88	6.14	0.95
30	2	-2.10	12.75	12.61	0.95	0.23	8.78	7.78	0.96	0.69	6.07	6.32	0.96
31	2	-0.18	10.54	9.90	0.98	0.68	6.97	6.23	0.97	0.64	4.81	4.40	0.96
32	3	0.90	22.12	18.85	0.95	0.27	15.09	13.89	0.95	-0.13	10.62	10.80	0.96
33	5	-1.45	22.39	19.31	0.98	-2.31	15.15	13.94	0.99	0.38	10.64	10.34	0.96
34	3	1.76	22.22	17.56	0.98	-0.19	15.22	13.07	0.99	0.24	10.73	9.74	0.98
35	5	0.41	22.52	19.99	0.98	-1.89	15.25	14.43	0.95	0.99	10.83	10.80	0.95
36	3	-1.14	22.67	19.30	0.98	-1.32	15.45	14.19	0.97	-0.09	10.87	10.66	0.96
37	5	-0.51	22.08	19.59	0.98	-1.67	15.20	14.18	0.97	0.43	10.63	10.69	0.98
38	3	-1.82	21.38	17.93	0.98	-1.95	14.53	14.07	0.99	0.24	10.21	10.20	0.96

Table 2.6: Results of simulations. True, true value of the parameter; Bias, 100× {mean($\hat{\beta}$) - β_0}; SE, 100× average of standard error estimates, SD, 100× sample standard deviation; CP, empirical coverage percentage of the 95% confidence interval.

		n = 500				n = 1000				n = 2000			
	True	Bias	SE	SD	CP	Bias	SE	SD	CP	Bias	SE	SD	CP
39	2	1.60	12.59	11.25	0.97	1.54	8.27	8.11	0.96	-0.05	5.60	5.66	0.96
40	2	-3.05	14.11	12.52	0.97	-0.11	9.36	10.54	0.90	-0.08	6.45	6.37	0.95
41	2	-5.08	13.47	11.90	0.97	0.85	8.82	9.60	0.90	0.16	5.97	6.76	0.90
42	1	-26.27	11.35	27.92	0.45	-0.75	8.13	14.75	0.86	-0.93	5.88	6.07	0.96
43	1	-0.21	6.69	6.09	0.98	1.32	4.46	4.86	0.94	0.40	3.02	3.02	0.96
44	1	1.91	8.08	6.74	0.98	1.23	5.33	4.68	0.96	1.82	3.66	3.61	0.94
45	1	1.69	6.96	6.61	0.93	1.06	4.59	4.67	0.92	0.61	3.15	2.82	0.96
46	1	3.25	9.23	9.03	0.97	1.89	6.04	5.37	0.99	2.03	4.17	3.51	0.98
47	1	2.01	7.79	8.09	0.95	0.94	5.14	5.28	0.94	1.00	3.51	3.73	0.90
48	1	-0.63	8.12	7.23	1.00	0.87	5.46	5.70	0.91	1.16	3.69	3.60	0.96
49	1	2.66	10.52	9.10	0.98	1.27	7.00	6.50	0.97	-0.08	4.69	5.24	0.94
50	1	2.82	7.56	6.56	0.95	0.83	4.94	4.86	0.91	0.05	3.35	3.89	0.90
51	1	3.70	8.14	7.02	0.97	0.91	5.31	5.46	0.96	0.87	3.57	3.96	0.93
52	-0.30	-0.17	6.43	5.56	0.97	0.79	4.25	3.67	0.96	0.60	2.95	3.19	0.94
53	0.30	1.09	8.65	9.65	0.92	0.56	5.82	6.71	0.90	0.11	3.99	4.12	0.94
54	-0.30	0.07	8.02	7.20	0.97	0.81	5.16	6.29	0.92	0.76	3.60	4.34	0.90

Table 2.7: Simulation setting for the fixed effects. Each row represents an event type. The columns are the corresponding covariate processes. The numbers are the regression coefficients. The dots represent the regression coefficient is 0.

	baseline	W1	W1_M	W2	W2_M	W3	W3_M	Next	Next_C	R_1
W1	-4	.	.	.	.	.	.	.	.	.
W1_M	-5	.	.	.	.	.	.	.	.	.
W2	-5	.	.	.	.	.	.	.	.	.
W2_M	-5	.	.	.	.	.	.	.	.	.
W3	-5	.	.	.	.	.	.	.	.	.
W3_M	-4	.	.	.	.	.	.	.	.	.
Next	-7	.	.	.	.	.	.	.	4	5
Next_C	-4	.	.	.	.	.	.	.	.	.
R_1	-5	.	.	.	.	.	.	.	.	.
R_2	-5	.	.	.	.	.	.	.	.	.
R_3	-5	.	.	.	.	.	.	.	.	.
R_Open	-6	.	.	.	.	.	2	.	.	.
R_Close	-5	.	.	.	.	.	.	.	.	.
Back	-7	3	5	3	5	3	5	.	.	.
Next_OK	-2	.	.	.	.	.	.	.	.	.

Table 2.8: Simulation setting for the fixed effects (continued)

	R_2	R_3	R_Open	R_Close	Back	W1..Back	W2..Back	W3..Back	Next_OK
W1	.	.	.	.	1	-2	-1	.	.
W1_M	.	.	.	.	.	.	.	.	.
W2	.	.	.	.	1	2	-2	.	.
W2_M	.	.	.	.	.	.	.	.	.
W3	.	.	.	.	1	-2	2	-2	.
W3_M	.	.	.	.	.	.	.	.	.
Next	5	5	.	.	.	.	.	.	.
Next_C	.	.	.	.	.	.	.	.	.
R_1	.	.	.	.	.	.	.	.	.
R_2	.	.	.	.	.	.	.	.	.
R_3	.	.	.	.	.	.	.	.	.
R_Open	.	.	.	.	.	.	.	2	.
R_Close	.	.	.	.	.	.	.	.	.
Back	.	.	.	.	3	.	.	.	.
Next_OK	.	.	.	.	.	.	.	.	.

Table 2.9: Simulation setting for the first dimension of the loading matrix

	W1	W1_M	W2	W2_M	W3	W3_M	Next	Next_C	R_1
W1	.	.	.	.	.	.	.	.	.
W1_M	2	.	.	.	.	.	.	.	.
W2	.	.	.	.	.	.	.	.	.
W2_M	.	.	2	.	.	.	.	.	.
W3	.	.	.	.	.	.	.	.	.
W3_M	.	.	.	.	2	.	.	.	.
Next	.	.	.	.	.	.	.	.	.
Next_C	.	.	.	.	.	.	.	.	.
R_1	.	.	.	.	.	.	.	.	.
R_2	.	.	.	.	.	.	.	.	.
R_3	.	.	.	.	.	.	.	.	.
R_Open	.	.	.	.	.	.	.	.	.
R_Close	.	.	.	.	.	.	.	.	.
Back	.	.	.	.	.	.	.	.	.
Next_OK	.	.	.	.	.	.	.	.	.

Table 2.10: Simulation setting for the first dimension of the loading matrix (continued)

	R_2	R_3	R_Open	R_Close	Back	W1..Back	W2..Back	W3..Back	Next_OK
W1	.	.	.	.	.	.	.	.	.
W1_M	.	.	.	.	.	.	.	.	.
W2	.	.	.	.	.	.	.	.	.
W2_M	.	.	.	.	.	.	.	.	.
W3	.	.	.	.	.	.	.	.	.
W3_M	.	.	.	.	.	.	.	.	.
Next	.	.	.	.	.	.	.	.	.
Next_C	.	.	.	.	.	.	.	.	.
R_1	.	.	.	.	.	.	.	.	.
R_2	.	.	.	.	.	.	.	.	.
R_3	.	.	1	.	.	.	.	.	.
R_Open	.	.	.	.	.	.	.	.	.
R_Close	.	.	.	.	.	.	.	.	.
Back	.	.	.	.	.	.	.	.	.
Next_OK	.	.	.	.	.	.	.	.	.

Table 2.11: Simulation setting for the second dimension of the loading matrix

	W1	W1_M	W2	W2_M	W3	W3_M	Next	Next_C	R_1
W1	.	.	.	.	.	.	.	.	.
W1_M	.	.	.	.	.	.	.	.	.
W2	.	.	.	.	.	.	.	.	.
W2_M	.	.	.	.	.	.	.	.	.
W3	.	.	.	.	.	.	.	.	.
W3_M	.	.	.	.	.	.	.	.	.
Next	.	.	.	.	.	.	.	.	.
Next_C	.	.	.	.	.	.	.	.	.
R_1	.	.	.	.	.	.	.	.	.
R_2	.	.	.	.	.	.	.	.	.
R_3	.	.	.	.	.	.	.	.	.
R_Open	.	.	.	.	.	.	.	.	.
R_Close	.	.	.	.	.	.	.	.	.
Back	1	1	1	1	1	1	.	.	.
Next_OK	.	.	.	.	.	.	.	.	.

Table 2.12: Simulation setting for the second dimension of the loading matrix (continued)

	R_2	R_3	R_Open	R_Close	Back	W1..Back	W2..Back	W3..Back	Next_OK
W1	.	.	.	.	.	.	.	.	.
W1_M	.	.	.	.	.	.	.	.	.
W2	.	.	.	.	.	.	.	.	.
W2_M	.	.	.	.	.	.	.	.	.
W3	.	.	.	.	.	.	.	.	.
W3_M	.	.	.	.	.	.	.	.	.
Next	.	.	.	.	.	.	.	.	.
Next_C	.	.	.	.	.	.	.	.	.
R_1	.	.	.	.	.	.	.	.	.
R_2	.	.	.	.	.	.	.	.	.
R_3	.	.	.	.	.	.	.	.	.
R_Open	.	.	.	.	.	.	.	.	.
R_Close	.	.	.	.	.	.	.	.	.
Back	.	.	.	.	.	.	.	.	.
Next_OK	.	.	.	.	.	.	.	.	.

Table 2.13: Simulation setting for the third dimension of the loading matrix

	W1	W1_M	W2	W2_M	W3	W3_M	Next	Next_C	R_1
W1	.	.	.	.	.	.	.	.	.
W1_M	.	.	.	.	.	.	.	.	.
W2	.	.	.	.	.	.	.	.	.
W2_M	.	.	.	.	.	.	.	.	.
W3	.	.	.	.	.	.	.	.	.
W3_M	.	.	.	.	.	.	.	.	.
Next	.	.	.	.	.	.	.	.	.
Next_C	.	.	.	.	.	.	.	.	.
R_1	.	.	.	.	.	.	.	.	.
R_2	.	.	.	.	.	.	.	.	.
R_3	.	.	.	.	.	.	.	.	.
R_Open	.	.	.	.	.	.	.	.	.
R_Close	.	.	.	.	.	.	.	.	.
Back	.	.	.	.	.	.	.	.	.
Next_OK	.	.	.	.	.	.	.	.	.

Table 2.14: Simulation setting for the third dimension of the loading matrix (continued)

	R_2	R_3	R_Open	R_Close	Back	W1..Back	W2..Back	W3..Back	Next_OK
W1	.	.	.	.	.	.	.	1	.
W1_M	.	.	.	.	.	.	.	.	.
W2	.	.	.	.	.	1	.	.	.
W2_M	.	.	.	.	.	.	.	.	.
W3	.	.	.	.	.	.	1	.	.
W3_M	.	.	.	.	.	.	.	.	.
Next	.	.	.	.	.	.	.	.	.
Next_C	.	.	.	.	.	.	.	.	.
R_1	.	.	.	.	.	.	.	.	.
R_2	.	.	.	.	.	.	.	.	.
R_3	.	.	.	.	.	.	.	.	.
R_Open	.	.	.	.	.	.	.	.	.
R_Close	.	.	.	.	.	.	.	.	.
Back	.	.	.	.	.	.	.	.	.
Next_OK	.	.	.	.	.	.	.	.	.

Chapter 3: Event History Analysis With Rare Events and Dynamic Sparse Covariates

3.1 Introduction

In recent years, event history analysis using large-scale longitudinal observational databases such as electronic health records and health insurance databases are becoming more popular. These include, for example, personalized treatment and identification of rare disease patients. A health insurance claims database typically contains millions of people with person-level prescription and medical diagnoses over a period of several years. An electronic health record of a person is a digital version of the patient's health information collected from all the clinicians involved in the patient's care. It includes a variety of information such as patient demographics, medications, diagnoses, vital signs, immunization, laboratory data, radiology images and allergies. It is built to share information across different health care providers. While electronic health records contain more variety of the information about a patient's medical history, claim databases may provide a more accurate prescription data as each time the patient gets a refill, it will be in the record when the patient files the claim. However, this is feasible only when the patient is insured. Therefore, combining both sources of information may provide a more comprehensive and accurate medical history of a patient. A common feature of these longitudinal observational data is that when our event of interest, for example, occurrence of certain diseases or prescription of certain drugs, is rare, a majority of the subjects do not experience any such event. Nevertheless, because of the size of the databases, studies on rare diseases and their relationship with different covariates or drug exposures are still feasible.

An example of analysis using the large-scale longitudinal observational databases is postmarketing drug safety surveillance, which is the continued monitoring of prescription drugs after they

have been approved in the market, see for example [40] and [41]. In such application, we are interested in the relationship between different (recurrent) adverse events related to health conditions and the drug exposures. Because of the scale of the data, to investigate the association between time-varying exposures and outcome events, one may want to consider only the cases (i.e. the subjects that experience the event of interest) to reduce computational complexity. One possible method is to use the self-controlled case series (SCCS) method, originally developed in [17]. In SCCS, the intensity function is assumed to be

$$\lambda_i(t|\mathcal{F}_{t-}) = e^{\phi_i + \beta^T X_i(t)},$$

where ϕ_i is a subject-specific parameter representing the person-level heterogeneity, $X(\cdot)$ is a vector of external time-varying covariates. Clearly, we cannot consistently estimate ϕ_i. However, it can be seen that the number of events n_i is a sufficient statistic for ϕ_i as the likelihood can be written as

$$e^{n_i\phi_i} \prod_{j=1}^{n_i} e^{\beta^T X_i(t_{ij})} \times \exp\left\{ - e^{\phi_i} \int_{a_i}^{b_i} e^{\beta^T X_i(u)} du \right\},$$

where t_{ij} are event times. Hence, conditioning on n_i will remove ϕ_i from the likelihood. The resulting conditional likelihood contribution from subject i is

$$\prod_{j=1}^{n_i} \frac{e^{\beta^T X(t_{ij})}}{\int_{a_i}^{b_i} e^{\beta^T X(t)} dt}, \tag{3.1}$$

and the inference of SCCS is based on this conditional likelihood. The benefit of using SCCS in analyzing large-scale data is that the conditional likelihood only depends on the cases. As a result, it greatly reduces the computational requirement. Another benefit of SCCS is that it can control for the multiplicative fixed individual covariates ϕ_i. However, a key assumption in SCCS is that event occurrence is conditional independent given the covariates. This is clearly violated in many important examples where the occurrence of the first event can alter the risk of having another one. For example, it is known that patients with a myocardial infarction is more likely to have a

subsequent one. To relax such assumption, [40] proposed a positive event dependence model for SCCS by specifying the intensity function as

$$\lambda_i(t|\mathcal{F}_{t-}) = \{e^{\phi_i} + \delta N_i(t-)\}e^{\beta^T X_i(t)}, \tag{3.2}$$

where $\delta \geq 0$. Clearly, when $\delta > 0$, the intensity function depends positively on the number of previous event and so event occurrence is no longer conditional independent given the covariates. Similarly to SCCS, n_i is still a sufficient statistic for ϕ_i for the model in (3.2). Hence, the computational benefits and self-control nature of SCCS using the conditional likelihood are retained while (3.2) allows a flexibility that the intensity depends on the number of previous event through the term $\delta N_i(t-)$ for $\delta > 0$.

For a more general intensity with external time-varying covariates $X(\cdot)$, the conditional likelihood for subject i on the number of events n_i is

$$\prod_{j=1}^{n_i} \frac{\lambda_i(t_{ij}|X_i)}{\left\{\int_{a_i}^{b_i} \lambda_i(t|X_i)dt\right\}^n}.$$

[18] showed that this case series method may be used with rare non-recurrent events. Let $\lambda_i(t|X_i)$ be the conditional hazard function for the rare non-recurrent event. The rare event setting is formulated by assuming $\lambda_i(t|X_i) = \varphi\nu_i(t|X_i)$ and then letting $\varphi \downarrow 0$. The conditional likelihood for subject i given X_i with an event at $t_i \in (a_i, b_i]$ is

$$\frac{S_i(t_i|X_i)\lambda_i(t_i|X_i)}{S(a_i|X_i) - S(b_i|X_i)}.$$

It is straightforward to see that $\lim_{\varphi\downarrow 0} S_i(t_i|X_i) = 1$ and

$$\lim_{\varphi\downarrow 0} \frac{S(a_i|X_i) - S(b_i|X_i)}{\int_{a_i}^{b_i} \lambda_i(t|X_i)dt} = 1.$$

Hence,

$$\lim_{\varphi \downarrow 0} \frac{S_i(t_i|X_i)\lambda_i(t_i|X_i)}{S(a_i|X_i) - S(b_i|X_i)} = \frac{\lambda_i(t_{ij}|X_i)}{\int_{a_i}^{b_i} \lambda_i(t|X_i)dt}$$

and in this sense, the case series method can be used with rare non-recurrent events.

Another setting of rare events is formulated in [6]. Motivated by the prostate, lung, colorectal, and ovarian (PLCO) cancer screening trial conducted by the National Cancer Institute, [6] studied a class of weighted log-rank tests for survival data when the event is rare. [6] showed that the popular G^ρ fmaily of weighted log-rank statistics essentially reduces to the special case of the unweighted log-rank statistics under the rare event setting. They proposed a simple modification to the G^ρ family and formulated a mathematical setting of rare event where the asymptotic properties of the statistics under both null and contiguous alternatives are studied. Since their setting is for comparing two distribution functions without covariates, the rare event setting is formulated in terms of the distribution functions as follow. First, the notion of rare event must be relative to the sample size. Hence, for each n, the underlying distribution function is $F_k^{(n)}$ and is assumed to be

$$F_k^{(n)}(t) = \frac{1}{m_n}\tilde{F}_k(t), \quad t \in [0, \tau],$$

where $m_n \to \infty$ and $\tilde{F}_k$'s are some increasing function for $k = 1, 2$. However, m_n cannot be arbitrarily large. Otherwise, we do not have enough number of events for consistency. Therefore, it is assumed that $n/m_n \to \infty$ so that the number of events also go to ∞. Informally, n/m_n is approximately proportional to the the information contained in the data.

Apart from having rare event, we could also have sparse covariates. Consider the case when the covariate is a process indicating whether a subject has taken certain drug in the last 30 days, that is, $X(t) = 1$ if the subject has taken that drug in the period $[t - 30, t)$. If a patient only takes the drug when he/she has certain rare diseases, then we expect that $X(t) = 0$ for the majority of the subjects at any time t. It is in this sense that the covariate is sparse. Another situation is that if we have a dynamic covariate that depends on the history of the event of interest, then such dynamic covariate could also be sparse when the event is rare. Note that this is different from a sparse

regression model in which case we have a lot of covariates but only a few of them have regression coefficients different from 0. Returning to the discussion of dynamic covariate, a popular model that captures a specific type of dynamic covariate is the Hawkes process, named after [24]. The defining characteristics of them is “self-excite", meaning that previous event increases the rate of having another future event. It is particularly suitable for modeling cluster events like earthquake and financial data. Mathematically, the intensity function for a Hawkes process is

$$\lambda(t) = \lambda + \int_0^t \mu(t-u)dN(u) = \lambda + \sum_{t_i<t} \mu(t-t_i),$$

where N is the underlying counting process, $\lambda > 0$ and $\mu : (0,\infty) \to [0,\infty)$ are called the background intensity and excitation function respectively. Since μ is positive, any previous event will increase the intensity and hence it is self-excited. A common choice of the excitation function is the exponential decay, where $\mu(t) = \alpha e^{-\beta t}$, for some $\alpha, \beta > 0$. For a J-variate counting process, the mutually exciting Hawkes process has intensity function

$$\lambda_j(t) = \lambda_j + \sum_{i=1}^{J} \int_0^t \mu_{ji}(t-\mu)dN_j(u), \quad j = 1,\ldots,J,$$

where $\lambda_i > 0$, $\mu_{ji} : (0,\infty) \to [0,\infty)$.

In Section 3.2, a mathematical formulation for the proportional intensity model with rare events is given. In Section 3.3, we establish the asymptotic theory of the maximum partial likelihood estimator under general conditions and illustrate the application to the cases when we have rare events and sparse covariates in Section 3.4. Simulation studies are given in Section 3.5. Some of the proofs are relegated to the Appendix

3.2 Setting and Notation

In this section, we first formalize the idea of rare events in a proportional intensity model. The data consists of $\{N_i^{(n)}(\cdot), X_i^{(n)}(\cdot), Y_i^{(n)}(\cdot) : i = 1,\ldots,n\}$. It is assumed that the counting process

$N_i^{(n)}(\cdot), i = 1, \ldots, n$ has intensity function

$$\lambda^{(n)}(t|\mathcal{F}_{t-}) = \lambda_0(t)e^{c_n+\beta^T X_i^{(n)}(t)}Y_i^{(n)}(t), \tag{3.3}$$

where $c_n \to -\infty$, $\lambda_0(\cdot)e^{c_n}$ is the baseline intensity function, β is the vector of regression parameters of interest, $X_i^{(n)}(\cdot)$ is a vector of predictable covariate processes and $Y_i^{(n)}(\cdot)$ is the at-risk indicator process. As the notion of rare event is relative to the sample size, we explicitly write down the dependence on n in these quantities. Let τ be the duration of study. Since $c_n \to -\infty$, the intensity function converges to 0 almost surely at any fixed t as $n \to \infty$. Hence, event occurrence becomes more rare as n increases. On the other hand, c_n cannot go to $-\infty$ arbitrarily fast. We require $ne^{c_n} \to \infty$ so that we have enough events for the consistency estimation β. Informally, ne^{c_n} is proportional to the observed Fisher information matrix and we require the information to go infinity for establishing consistency. If $\{c_n\}$ is a constant sequence, (3.3) becomes the usual proportional intensity model. It is straightforward to see that the same argument leading to the partial likelihood in the proportional intensity model also applies in the case under (3.3). Hence, the estimator of β is obtained by maximizing the log partial likelihood function:

$$\log \mathrm{PL}(\beta) = \sum_{i=1}^{n} \left[\int_0^\tau \beta^T X_i^{(n)}(t) - \log \left\{ \sum_{k=1}^{n} Y_k^{(n)}(t) e^{\beta^T X_k^{(n)}(t)} \right\} \right] dN_i^{(n)}(t).$$

That is, we define $\hat{\beta}_n := \arg\max_\beta \log \mathrm{PL}(\beta)$. Let β_0 denote the true value of β. To facilitate the proofs in this section, define

$$\begin{aligned}
S_n^{(0)}(\beta_0; s) &:= \frac{1}{n}\sum_{i=1}^{n} Y_i^{(n)}(s) e^{\beta_0^T X_i^{(n)}(s)}, \\
S_n^{(1)}(\beta_0; s) &:= \frac{1}{n}\sum_{i=1}^{n} Y_i^{(n)}(s) X_i^{(n)}(s) e^{\beta_0^T X_i^{(n)}(s)}, \\
S_n^{(2)}(\beta_0; s) &:= \frac{1}{n}\sum_{i=1}^{n} Y_i^{(n)}(s) X_i^{(n)}(s) X_i^{(n)}(s)^T e^{\beta_0^T X_i^{(n)}(s)}.
\end{aligned}$$

Let

$$V_n(\beta, s) := \frac{S_n^{(2)}(\beta; s)}{S_n^{(0)}(\beta; s)} - \left(\frac{S_n^{(1)}(\beta; s)}{S_n^{(0)}(\beta; s)}\right)^{\otimes 2}.$$

Define $M_i^{(n)} := N_i^{(n)} - A_i^{(n)}$, where $N_i^{(n)}$ is the compensator of $A_i^{(n)}$ given by

$$A_i^{(n)}(t) := \int_0^t \{\beta^T X_i^{(n)}(u) - \log S_n^{(0)}(\beta; u)\} Y_i^{(n)}(u) \lambda_0(u) e^{c_n} e^{\beta_0^T X_i^{(n)}(u)} du.$$

The score process is

$$U_n(\beta_0, t) := \int_0^t \left\{X_i^{(n)}(u) - \frac{S_n^{(1)}(\beta_0, u)}{S_n^{(0)}(\beta_0, u)}\right\} dN_i^{(n)}(u) = \int_0^t \left\{X_i^{(n)}(u) - \frac{S_n^{(1)}(\beta_0, u)}{S_n^{(0)}(\beta_0, u)}\right\} dM_i^{(n)}(u).$$

The observed Fisher information matrix is

$$I_n(\beta) := \int_0^\tau \left\{\frac{S_n^{(2)}(\beta; u)}{S_n^{(0)}(\beta; u)} - \left(\frac{S_n^{(1)}(\beta; u)}{S_n^{(0)}(\beta; u)}\right)^{\otimes 2}\right\} dN_i^{(n)}(u).$$

3.3 Asymptotic Theory under General Setting

In this section, we show the consistency and asymptotic normality of the partial likelihood estimator under general conditions on the score equation and the observed Fisher information. We shall verify the these conditions under various settings, in particular, including the setting when we have both rare events and sparse covariates.

Assumption 1. *(i) There exists a sequence of invertible deterministic matrices $\{D_n\}$ with $||D_n^{-1}|| \to 0$ as $n \to \infty$ such that $D_n^{-1} I_n(\beta_0)(D_n^{-1})^T \xrightarrow{\mathbb{P}} I$.*

(ii) $D_n^{-1} U_n(\beta_0) \xrightarrow{d} N(0, I)$.

(iii) Let $u \in \mathbb{R}^d$. Using multi-index notation, for $|\alpha| = 3$,

$$\sup_{\beta \in B(\beta_0, ||D_n^{-1} u||)} \left| \frac{\partial L_n(\beta)}{\partial \alpha} (D_n^{-1} u)^\alpha \right| = o_P(||u||^2),$$

where $B(\beta_0, \delta) := \{\beta : ||\beta - \beta_0|| \leq \delta\}$.

(iv) For any consistent estimator $\tilde{\beta}_n$ *of* β_0, $D_n^{-1} I_n(\tilde{\beta}_n)(D_n^{-1})^T \xrightarrow{\mathbb{P}} I$ *and* $I_n^{\frac{1}{2}}(\tilde{\beta}_n)(D_n^{-1})^T \xrightarrow{\mathbb{P}} I$.

Lemma 2. *Assumptions 1 ((i)) and ((iv)) are implied by the following conditions with* $D_n = \tilde{D}_n^{\frac{1}{2}}$*:*

(i') There exists a sequence of deterministic positive definite matrices $\{\tilde{D}_n\}$ *with* $||\tilde{D}_n^{-1}|| \to 0$ *such that* $\tilde{D}_n^{-1} I_n(\beta_0) \xrightarrow{\mathbb{P}} I$.

(iv') For any consistent estimator $\tilde{\beta}_n$ *of* β_0, $\tilde{D}_n^{-1} I_n(\tilde{\beta}_n) \xrightarrow{\mathbb{P}} I$.

Proof. The proof of the implication is seen by the two linear algebra results (see Lemmas 3 and 4) and the characterization of convergence in probability that $W_n \xrightarrow{\mathbb{P}} W$ if and only if every subsequence has a further subsequence along which $W_n \xrightarrow{a.s.} W$. □

Assumption 1 ((i)) is required so that the observed Fisher information suitably normalized will converge to a nondegenerate limit. Assumption 1 ((ii)) is usually a consequence of ((i)), which can be proved using the martingale central limit theorem under a suitable Linderberg condition. Assumption 1 ((iii)) is assumed to ensure the remainder terms in the Taylor's expansion of the log-likelihood is of smaller order and is easily satisfied. Assumption 1 ((iv)) is to ensure that inference based on the observed Fisher information is valid; see Theorems 12 and 13.

Theorem 11 (Consistency)**.** *Under Assumption 1 ((i)) - ((iii)), we have* $||\hat{\beta}_n - \beta_0|| = O_P(||D_n^{-1}||) = o_P(1)$.

Proof. By Taylor's theorem and Assumptions 1 ((i)) - ((iii)), we have

$$\begin{aligned} L_n(\beta_0 + D_n^{-1}u) - L_n(\beta_0) &= u^T D_n^{-1} U_n(\beta_0) - \frac{1}{2} u^T D_n^{-1} I_n(\beta_0) D_n^{-1} u + \sum_{|\alpha|=3} \frac{1}{\alpha!} \frac{\partial L_n(\beta_n^*)}{\partial \alpha} (D_n^{-1}u)^\alpha \\ &= u^T O_P(1) - \frac{1}{2}||u||^2(1 + o_P(1)) + o_P(||u||^2), \end{aligned}$$

where β_n^* lies on the segment between β_0 and $\beta_0 + D_n^{-1}u$. Hence, for all $\varepsilon > 0$, there exists $C_0 > 0$

such that for all large enough n,

$$\mathbb{P}\left(\sup_{u:||u||=C_0} \{L_n(\beta_0 + D_n^{-1}u) - L_n(\beta_0)\} < 0 \right) \geq 1 - \varepsilon.$$

This implies that there is a local maximizer in $\{\beta_0 + D_n^{-1}u : ||u|| \leq C_0\}$ with a probability at least $1 - \varepsilon$ for all large enough n. Since L_n is concave, the local maximizer is the global maximizer $\hat{\beta}_n$. Therefore, for all large enough n, $\mathbb{P}(||\hat{\beta}_n - \beta_0|| \leq ||D_n^{-1}||C_0) \geq 1 - \varepsilon$ and the claim follows. □

Theorem 12 (Asymptotic Normality)**.** *Under Assumptions 1 ((i)) - ((iv)), we have*

$$I_n^{\frac{1}{2}}(\hat{\beta}_n)(\hat{\beta}_n - \beta_0) \xrightarrow{d} N(0, I). \tag{3.4}$$

Proof. By Taylor's theorem, there exists a β_n^* on the line segment between β_0 and $\hat{\beta}_n$ such that

$$0 = U_n(\hat{\beta}_n) = U_n(\beta_0) - I_n(\beta_n^*)(\hat{\beta}_n - \beta_0).$$

Note that under Assumptions 1 ((i))-((iii)), $\hat{\beta}_n$ is consistent. Then, by Assumptions 1 ((ii)) and ((iv)), and Slutsky's theorem, we have

$$\begin{aligned} I_n^{\frac{1}{2}}(\hat{\beta}_n)(\hat{\beta}_n - \beta_0) &= I_n^{\frac{1}{2}}(\hat{\beta}_n) I_n^{-1}(\beta_n^*) U_n(\beta_0) \\ &= I_n^{\frac{1}{2}}(\hat{\beta}_n) D_n^{-1} \{D_n I_n^{-1}(\beta_n^*) D_n\} D_n^{-1} U_n(\beta_0) \\ &\xrightarrow{d} N(0, I). \end{aligned}$$

□

As a result of Theorem 12, we have $\hat{\beta}_n - \beta_0 \approx N(0, I_n^{-1}(\hat{\beta}_n))$. Therefore, inference can be based on the observed Fisher information matrix. Formally, we have the Wald test.

Theorem 13 (Wald Test)**.** *Under Assumptions 1 ((i)) - ((iv)), we have*

$$\frac{\hat{\beta}_{nj} - \beta_{0j}}{\sqrt{(I_n^{-1})_{jj}}} \xrightarrow{d} N(0,1), \tag{3.5}$$

where $(I_n^{-1})_{jj}$ *is the* (j,j)*-th element of* I_n^{-1}*. Let* $g : \mathbb{R}^p \to \mathbb{R}^r$*,* $r \leq p$*, be a continuously differentiable function where the Jacobian* J_g *has rank* r*. Then,*

$$\{g(\hat{\beta}_n) - g(\beta_0)\}^T \{J_g(\hat{\beta}_n) I_n^{-1}(\hat{\beta}_n) J_g^T(\hat{\beta}_n)\}^{-1} \{g(\hat{\beta}_n) - g(\beta_0)\} \xrightarrow{d} \chi^2(r), \tag{3.6}$$

where $\chi^2(r)$ *denotes the chi-square distribution with* r *degrees of freedom.*

Proof. We first prove (3.5). Let $e_j = (0, \ldots, 1, \ldots, 0)^T$, where the jth element is 1 and the other elements are 0. Then,

$$\frac{\hat{\beta}_{nj} - \beta_{0j}}{\sqrt{(I_n^{-1})_{jj}}} = \{e_j^T I_n^{-1}(\hat{\beta}_n) e_j\}^{-\frac{1}{2}} e_j^T (\hat{\beta}_n - \beta_0) = [\{e_j^T I_n^{-1}(\hat{\beta}_n) e_j\}^{-\frac{1}{2}} e_j^T I_n^{-\frac{1}{2}}(\hat{\beta}_n)] \{I_n^{\frac{1}{2}}(\hat{\beta}_n)(\hat{\beta}_n - \beta_0)\}.$$

Let $a_n^T = \{e_j^T I_n^{-\frac{1}{2}}(\hat{\beta}_n) e_j\}^{-1} e_j^T I_n^{-\frac{1}{2}}(\hat{\beta}_n)$. Then $a_n^T a_n = 1$. Hence, a_n is bounded and every subsequence $\{n_k\}$ has a further subsequence $\{n_{k_l}\}$ along which it converges to a, which is possibly random. By Skorokhod representation theorem, there is another probability space and random variables W_n such that $W_n \overset{d}{=} I_n^{\frac{1}{2}}(\hat{\beta}_n)(\hat{\beta}_n - \beta_0)$ and $W_n \xrightarrow{a.s.} N(0, I)$. By arguing along subsequences, we see that $a_n^T W_n \xrightarrow{a.s.} a^T N(0, I) = N(0, a^T a) = N(0,1)$. Therefore,

$$a_n^T \{I_n^{\frac{1}{2}}(\hat{\beta}_n)(\hat{\beta}_n - \beta_0)\} \xrightarrow{d} N(0,1). \tag{3.7}$$

To show (3.6), note that

$$\begin{aligned} & \{J_g(\hat{\beta}_n) I_n^{-1}(\hat{\beta}_n) J_g^T(\hat{\beta}_n)\}^{-\frac{1}{2}} \{g(\hat{\beta}_n) - g(\beta_0)\} \\ = \; & \{J_g(\hat{\beta}_n) I_n^{-1}(\hat{\beta}_n) J_g^T(\hat{\beta}_n)\}^{-\frac{1}{2}} J_g(\hat{\beta}_n^*) I_n^{-\frac{1}{2}}(\hat{\beta}_n) \{I_n^{\frac{1}{2}}(\hat{\beta}_n)(\hat{\beta}_n - \beta_0)\}, \end{aligned}$$

where β_n^* lies on the line segment between β_0 and $\hat{\beta}_n$. Let $A_n = \{J_g(\hat{\beta}_n)I_n^{-1}(\hat{\beta}_n)J_g^T(\hat{\beta}_n)\}^{-\frac{1}{2}}J_g(\hat{\beta}_n^*)I_n^{-\frac{1}{2}}(\hat{\beta}_n)$. By the consistency of $\hat{\beta}_n$ and the continuity of J_g, we see that $A_nA_n^T \xrightarrow{\mathbb{P}} I_r$, where I_r is an $r \times r$ identity matrix. Hence, every subsequence $\{n_k\}$ has a further subsequence $\{n_{k_l}\}$ along which $A_nA_n^T \to I_r$ almost surely. We may assume $A_{n_{k_l}}$ converges to A almost surely; otherwise argue along subsequences. Then $A_{n_{k_l}}W_{n_{k_l}} \xrightarrow{a.s.} AN(0, I_p) = N(0, AA^T) = N(0, I_r)$. Hence, $A_nW_n \xrightarrow{\mathbb{P}} N(0, I_r)$. As a result,

$$\{J_g(\hat{\beta}_n)I_n^{-1}(\hat{\beta}_n)J_g^T(\hat{\beta}_n)\}^{-\frac{1}{2}}\{g(\hat{\beta}_n) - g(\beta_0)\} \xrightarrow{d} N(0, I_r),$$

which implies (3.6). □

Let

$$T_n := g(\hat{\beta}_n)^T\{J_g(\hat{\beta}_n)I_n^{-1}(\hat{\beta}_n)J_g^T(\hat{\beta}_n)\}^{-1}g(\hat{\beta}_n) \xrightarrow{d} \chi^2(r).$$

Under the null hypothesis that $g(\theta_0) = 0$, we reject the null hypothesis when $T_n > F_r^{-1}(1 - \alpha)$, where F_r is the distribution function of $\chi^2(r)$ and α is the significance level. A confidence region for β_0 can be found inverting the Wald test.

Lemma 3. *Let $\{B_n\}$ be a sequence of $p \times p$ nonnegative definite matrices and $\{C_n\}$ be a sequence of $p \times p$ symmetric matrices, where $p \in \mathbb{N}$. Then $B_nC_n \to I$ as $n \to \infty$ implies that $B_n^{\frac{1}{2}}C_nB_n^{\frac{1}{2}} \to I$ as $n \to \infty$.*

Proof. Apply the eigendecomposition to B_n so that $B_n = E_n\Lambda_nE_n^T$, where E_n is orthornomal and Λ_n is a diagonal matrix. Since $B_nC_n \to I$ and E_n is bounded

$$\Lambda_nE_n^TC_nE_n - I = E_n^T(E_n\Lambda_nE_n^TC_n - I)E_n \to 0.$$

Let $\tilde{B}_n := \Lambda_n \geq 0$ and $\tilde{C}_n := E_n^TC_nE_n$. We have $\tilde{B}_n\tilde{C}_n \to I$. Write $\tilde{B}_n = \text{diag}(b_{n,1}, \ldots, b_{n,p})$ and $\tilde{C}_n = \{c_{n,ij}\}_{i,j=1,\ldots,p}$. Thus,

$$b_{n,j}c_{n,jj} \to 1, \quad \text{for all } j = 1, \ldots, p$$

and

$$b_{n,j}c_{n,ji} \to 0, \quad \text{for all } i, j = 1, \ldots, p.$$

Note that $\tilde{C}_n$ is symmetric as C_n is, so $c_{n,ji} = c_{n,ij}$ and thus

$$\max\{b_{n,j}c_{n,ji}, b_{n,i}c_{n,ij}\} = \max\{b_{n,j}, b_{n,i}\}c_{n,ij} \to 0.$$

Now, the diagonal elements in $\tilde{B}_n^{\frac{1}{2}}\tilde{C}_n\tilde{B}_n^{\frac{1}{2}}$ are simply $b_{n,j}c_{n,jj} \to 1$ and the off-diagonal elements are $\sqrt{b_{n,i}b_{b,j}}c_{n,ij} \leq \max\{b_{n,j}, b_{n,i}\}c_{n,ij} \to 0$. Therefore, we have shown that

$$\Lambda_n^{\frac{1}{2}}E_n^T C_n E_n \Lambda_n^{\frac{1}{2}} \to I.$$

This implies that

$$B_n^{\frac{1}{2}}C_n B_n^{\frac{1}{2}} - I = E_n(\Lambda_n^{\frac{1}{2}}E_n^T C_n E_n \Lambda_n^{\frac{1}{2}} - I)E_n^T \to 0.$$

□

Remark 2. *(i) The converse of Lemma 3 is not true, even when C_n is positive definite. For a counterexample, let*

$$B_n = \begin{pmatrix} \frac{1}{n} & 0 \\ 0 & \frac{1}{n^2} \end{pmatrix} \text{ and } C_n = \begin{pmatrix} n & n^{1.1} \\ n^{1.1} & n^2 \end{pmatrix}$$

Then C_n is positive definite, $B_n^{\frac{1}{2}}C_n B_n^{\frac{1}{2}} \to I$ but

$$B_n C_n \to \begin{pmatrix} 1 & \infty \\ \infty & 1 \end{pmatrix}.$$

Note that even when $c_{n,12} = n$, this is not true. A sufficient condition in this case is that $c_{n,ij} = o(\min(c_{n,ii}, c_{n,jj}))$.

(ii) In the proof of Lemma 3, we do not assume B_n or C_n are bounded and the result holds for general B_n and C_n. In fact, in our applications, the elements in the observed Fisher

information are indeed unbounded.

Remark 3. *If $\{B_n\}$ is a sequence of diagonal positive definite matrices and $\{C_n\}$ is a sequence of positive definite matrices, then B_nC_n does not necessarily converge to a matrix of the form BC where B is diagonal and C is nonnegative definite. For example,*

$$\begin{pmatrix} \frac{1}{n} & 0 \\ 0 & \frac{1}{n^2} \end{pmatrix}\begin{pmatrix} n & n \\ n & n^2 \end{pmatrix} \to \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}.$$

Lemma 4. *Let $\{B_n\}$ and $\{C_n\}$ be sequences of nonnegative definite matrices. Suppose that $A_n = B_nC_n$ for each n. Then $A_nA_n^T \to I$ as $n \to \infty$ if and only if $A_n \to I$ as $n \to \infty$.*

Proof. To show the forward implication, apply the eigendecomposition to B_n so that $B_n = E_n\Lambda_nE_n^T$, where E_n is orthornomal and Λ_n is a diagonal matrix. $B_n = E_n\Lambda_nE_n^T$. Then

$$\begin{aligned} A_nA_n^T &= E_n\Lambda_nE_n^TC_nC_nE_n\Lambda E_n^T = E_n\Lambda(E_n^TC_nE_n)(E_n^TC_nE_n)\Lambda E_n^T \\ &= E_n\tilde{B}_n\tilde{C}_n\tilde{C}_n^T\tilde{B}_nE_n^T, \end{aligned}$$

where $\tilde{B}_n := \Lambda_n \geq 0$ because $B_n \geq 0$ and $\tilde{C}_n := E_n^TC_nE_n \geq 0$ because $C_n \geq 0$. Since E_n is bounded, we have $E_n^T\{E_n\tilde{B}_n\tilde{C}_n\tilde{C}_n^T\tilde{B}_nE_n^T\}E_n \to I$ and so

$$\tilde{B}_n\tilde{C}_n\tilde{C}_n\tilde{B}_n \to I.$$

Let $\tilde{A}_n := \tilde{B}_n\tilde{C}_n$. We then have $\tilde{A}_n\tilde{A}_n^T \to I$, where $\tilde{B}_n$ and $\tilde{C}_n$ are both nonnegative definite. Now, because $\tilde{A}_n\tilde{A}_n^T \to I$, the elements in $\tilde{A}_n$ are bounded. Therefore, $\tilde{A}_n \to A$ for some A along a subsequence. We can assume the whole sequence converges; otherwise argue along subsequences. As $AA^T = I$, A is orthonormal. This result together with the facts that $\tilde{B}_n \geq 0$ is diagonal and $\tilde{C}_n \geq 0$ imply that $\tilde{B}_n\tilde{C}_n$ converges to BC, where B is again diagonal and C is nonnegative definite. Since A is orthonormal, the row vectors in C are orthogonal and hence $CC = D$, where $D > 0$ is a diagonal matrix. As C is nonnegative definite, C must be the unique nonnegative definite square

root matrix of D, i.e., $C = D^{\frac{1}{2}}$, implying C is also a diagonal matrix. As $A = BC$, which is a product of two nonnegative definite diagonal matrices, A is a nonnegative definite diagonal matrix. Then $AA = I$ implies $A = I$. Therefore, $\tilde{A}_n \to I$. Finally,

$$A_n - I = E_n \Lambda_n E_n^T C_n - I = E_n(\tilde{A}_n - I)E_n^T \to 0.$$

The other direction of implication is trivial. □

Remark 4. *(i) By setting $A_n = B_n^{\frac{1}{2}} C_n^{\frac{1}{2}}$ in Lemma 4, we have $B_n^{\frac{1}{2}} C_n B_n^{\frac{1}{2}} \to I$ if and only if $B_n^{\frac{1}{2}} C_n^{\frac{1}{2}} \to I$ provided $\{B_n\}$ and $\{C_n\}$ are sequences of nonnegative definite matrices.*

(ii) In the proof of Lemma 4, we do not assume B_n or C_n are bounded and the result holds for general B_n and C_n. In fact, in our applications, the elements in the observed Fisher information are indeed unbounded.

3.4 Applications

3.4.1 Proportional Intensity Model

In this section, we verify Assumption 1 under the proportional intensity model with setting following Chapter 8 in [19]:

Assumption 2. *(i) $\int_0^\tau \lambda_0(t)dt < \infty$.*

(ii) There exists a neighborhood $N(\beta_0)$ of β_0 and function $s^{(j)}$ such that, for $j = 0, 1, 2$,

$$\sup_{t\in[0,\tau],\beta\in N(\beta_0)} ||S_n^{(j)}(\beta, t) - s^{(j)}(\beta, t)|| \xrightarrow{\mathbb{P}} 0,$$

where for a scalar b, $||b|| := |b|$, for a vector $b = (b_1, \ldots, b_p)^T$, $||b|| := \max_{i=1,\ldots,p} b_i$ and for a matrix $\boldsymbol{B} = \{a_{ij}\}$, $||\boldsymbol{B}|| := \max_{i,j} |b_{ij}|$.

(iii) There exists a $\delta > 0$ such that

$$n^{-1/2} \sup_{1 \le i \le n, 0 \le t \le \tau} |X_i^{(n)}(t)| Y_i^{(n)}(t) I\{\beta_0^T X_i^{(n)}(t) > -\delta ||X_i^{(n)}(t)||\} \xrightarrow{\mathbb{P}} 0.$$

(iv) Let $e = s^{(1)}/s^{(0)}$ and $v = s^{(2)}/s^{(0)} - e^{\otimes 2}$. For all $\beta \in N(\beta_0)$ and $0 \le t \le \tau$,

$$\begin{aligned} \frac{\partial}{\partial \beta} s^{(0)}(\beta, t) &= s^{(1)}(\beta, t), \\ \frac{\partial^2}{\partial \beta^2} s^{(0)}(\beta, t) &= s^{(2)}(\beta, t). \end{aligned}$$

(v) The functions $s^{(j)}$ are bounded above and $s^{(0)}$ is bounded away from 0 on $N(\beta_0) \times [0, \tau]$; for $j = 0, 1, 2$, the family of functions $s^{(j)}(\cdot, t), 0 \le t \le \tau$, is an equicontinuous fmaily at β_0.

(vi) The matrix

$$\Sigma(\beta_0) = \int_0^\tau \left[\frac{s^{(2)}(\beta_0, t)}{s^{(0)}(\beta_0, t)} - \left\{ \frac{s^{(1)}(\beta_0, t)}{s^{(0)}(\beta_0, t)} \right\}^{\otimes 2} \right] s^{(0)}(\beta_0, t) \lambda_0(t) dt$$

is positive definite, where for a vector b, $b^{\otimes 2} = bb^T$ denotes the outer product of b with itself.

Take $D_n = \Sigma^{\frac{1}{2}} n^{\frac{1}{2}}$. Clearly, $||D_n^{-1}|| \to 0$. Under Assumption 2, by Theorem 8.2.1 in [19], $n^{-\frac{1}{2}} U_n(\beta_0) \xrightarrow{d} N(0, \Sigma(\beta_0))$ and for any consistent estimator $\tilde{\beta}_n$ of β_0, $n^{-1} I_n(\tilde{\beta}_n) \xrightarrow{\mathbb{P}} \Sigma$. It is then straightforward to see that Assumptions 1 (i), (ii) and (iv) are satisfied. Assumption 1 (iii) is also satisfied $\sup_{\beta \in N(\beta_0)} |\frac{1}{n} \frac{\partial L_n(\beta)}{\partial \alpha}|$ is easily seen to be $O_P(1)$.

3.4.2 Proportional Intensity Model with Rare Events

In this section, we verify Assumption 1 under the proportional intensity model with rare events as specified by (3.3) and Assumption 3.

Assumption 3. *(i)* $\int_0^\tau \lambda_0(t) dt < \infty$.

(ii) There exists a neighborhood $N(\beta_0)$ of β_0 and function $s^{(j)}$ such that, for $j = 0, 1, 2$,

$$\sup_{t\in[0,\tau],\beta\in N(\beta_0)} ||S_n^{(j)}(\beta, t) - s^{(j)}(\beta, t)|| \xrightarrow{\mathbb{P}} 0,$$

where for a scalar b, $||b|| := |b|$, for a vector $b = (b_1, \ldots, b_p)^T$, $||b|| := \max_{i=1,\ldots,p} b_i$ and for a matrix $\boldsymbol{B} = \{a_{ij}\}$, $||\boldsymbol{B}|| := \max_{i,j} |b_{ij}|$.

(iii) There exists a $\delta > 0$ such that

$$e^{-\frac{c_n}{2}} n^{-1/2} \sup_{1\le i\le n, 0\le t\le\tau} |X_i^{(n)}(t)|Y_i^{(n)}(t) I\{\beta_0^T X_i^{(n)}(t) > -\delta||X_i^{(n)}(t)||\} \xrightarrow{\mathbb{P}} 0.$$

(iv) Let $e = s^{(1)}/s^{(0)}$ and $v = s^{(2)}/s^{(0)} - e^{\otimes 2}$. For all $\beta \in N(\beta_0)$ and $0 \le t \le \tau$,

$$\begin{aligned} \frac{\partial}{\partial\beta} s^{(0)}(\beta, t) &= s^{(1)}(\beta, t), \\ \frac{\partial^2}{\partial\beta^2} s^{(0)}(\beta, t) &= s^{(2)}(\beta, t). \end{aligned}$$

(v) The functions $s^{(j)}$ are bounded above and $s^{(0)}$ is bounded away from 0 on $N(\beta_0) \times [0, \tau]$; for $j = 0, 1, 2$, the family of functions $s^{(j)}(\cdot, t), 0 \le t \le \tau$, is an equicontinuous fmaily at β_0.

(vi) The matrix

$$\Sigma(\beta_0) = \int_0^\tau \left[\frac{s^{(2)}(\beta_0, t)}{s^{(0)}(\beta_0, t)} - \left\{ \frac{s^{(1)}(\beta_0, t)}{s^{(0)}(\beta_0, t)} \right\}^{\otimes 2} \right] s^{(0)}(\beta_0, t)\lambda_0(t)dt$$

is positive definite, where for a vector b, $b^{\otimes 2} = bb^T$ denotes the outer product of b with itself.

(vii) $c_n \to -\infty$ and $ne^{c_n} \to \infty$.

Condition ((i))-((vi)) are regular conditions for establishing the asymptotic theory in the proportional intensity model (see Ch.8 in [19]). Condition ((vii)) is the setting of rare event where the condition $ne^{c_n} \to \infty$ is necessary so that we have enough data for the consistency of the estimator.

Lemma 5. *Under Assumption 3, we have*

(a)

$$e^{-\frac{c_n}{2}} n^{-\frac{1}{2}} U_n(\beta_0) \xrightarrow{d} N(0, \Sigma(\beta_0)); \tag{3.8}$$

(b)

$$e^{-c_n} n^{-1} I_n(\tilde{\beta}_n) \xrightarrow{\mathbb{P}} \Sigma(\beta_0), \tag{3.9}$$

for any consistent estimator $\tilde{\beta}_n$ of β_0.

Proof. See the appendix for details. □

Corollary 1. *Under Assumption 3, (3.4), (3.5) and (3.6) hold.*

Proof. It suffices to verify Assumption 1. Take $D_n = \Sigma^{\frac{1}{2}} n^{\frac{1}{2}} e^{\frac{c_n}{2}}$. By Assumption 3 ((vii)), $||D_n^{-1}|| \to 0$. By Lemma 5, we can see that Assumptions 1 (i), (ii) and (iv) are satisfied. Assumption 1 (iii) is also easily seen to be satisfied; we shall give the details in the more complicated case in Section 3.4.3. □

3.4.3 Proportional Intensity Model with Rare Events and Dynamic Sparse Covariates

In Section 3.4.2, we established the asymptotic theory of the maximum partial likelihood estimator in the proportional intensity model with rare events. In this section, in addition to rare events, we also consider the case we have sparse covariates. Informally, sparse covariates means that most of the covariates are 0. This arises when the covariates are for example indicating whether there was a certain rare event happen before. In particular, that event considered in the covariate could be the same as the event of interest that is being modeled, resulting a dynamic covaraite. Another example is that $X(t) = 1$ if the subject has taken a certain drug in the past 14 days before t. If the drug is very uncommon, then for most of the people, $X(t) = 0$.

The setting is similar to that in Section 3.4.2 except that we now expect

$$\sum_{i=1}^{n} Y_i^{(n)}(s) X_i^{(n)}(s) e^{\beta^T X_i^{(n)}(s)} = o_P\left(\sum_{i=1}^{n} Y_i^{(n)}(s) e^{\beta^T X_i^{(n)}(s)} \right)$$

and for some $l, l' \in \{1, \ldots, p\}$ with $l \neq l'$,

$$\sum_{i=1}^{n} Y_i^{(n)}(s) X_{il}^{(n)}(s) e^{\beta^T X_i^{(n)}(s)} = o_P\left(\sum_{i=1}^{n} Y_i^{(n)}(s) X_{il'}^{(n)} e^{\beta^T X_i^{(n)}(s)} \right),$$

corresponding to sparse covariates and highly unbalanced covariates. Again, we will estimate β using the maximum partial likelihood estimator. Assumption 4 below specifies the regularity conditions for the consistency and asymptotic normality of $\hat{\beta}_n$.

Assumption 4. *(i)* $\int_0^\tau \lambda_0(t)dt < \infty$.

(ii) There exists a neighborhood $N(\beta_0)$ *of* β_0 *and functions* $s^{(j)}$ *such that, for* $j = 0, 1, 2$,

$$\sup_{t\in[0,\tau],\beta\in N(\beta_0)} ||S_n^{(0)}(\beta, t) - s^{(0)}(\beta, t)|| \xrightarrow{\mathbb{P}} 0,$$
$$\sup_{t\in[0,\tau],\beta\in N(\beta_0)} ||diag(\gamma_{n1}, \ldots, \gamma_{np}) S_n^{(1)}(\beta, t) - s^{(1)}(\beta, t)|| \xrightarrow{\mathbb{P}} 0,$$
$$\sup_{t\in[0,\tau],\beta\in N(\beta_0)} ||diag(\gamma_{n1}, \ldots, \gamma_{np}) S_n^{(2)}(\beta, t) diag(\gamma_{n1}, \ldots, \gamma_{np}) - s^{(2)}(\beta, t)|| \xrightarrow{\mathbb{P}} 0,$$

where $\gamma_{nj} > 0$ *and* $diag(\gamma_{n1}, \ldots, \gamma_{np})$ *denotes the diagonal matrix with elements* $\gamma_{n1}, \ldots, \gamma_{np}$. *Let* $v = s^{(2)}/s^{(0)}$.

(iii) The covariate processes are bounded. That is, there exists a $M > 0$ *such that*

$$\sup_{i=1,\ldots,n,0\leq t\leq\tau} ||X_i^{(n)}(t)|| \leq M.$$

(iv) The functions $s^{(j)}$ *are bounded and* $s^{(0)}$ *is bounded away from* 0 *on* $N(\beta_0) \times [0, \tau]$*; for* $j = 0, 1, 2$, *the family of functions* $s^{(j)}(\cdot, t), 0 \leq t \leq \tau$, *is an equicontinuous fmaily at* β_0.

(v) The matrix

$$\Sigma(\beta_0) = \int_0^\tau s^{(2)}(\beta_0, t)\lambda_0(t)dt$$

is positive definite.

(vi) $ne^{c_n}(\max_{j=1,\ldots,p}\gamma_{nj})^{-1} \to \infty$ *as* $n \to \infty$.

Condition ((i))-((v)) are regular conditions for establishing asymptotic theory in the proportional intensity model (see Ch.8 in [19]). Condition ((vi)) is needed so that we have enough data for the consistency of the estimator. For simplicity, we shall also use $\mathrm{diag}(\gamma_n)$, $\mathrm{diag}(\sqrt{\gamma_n})$ and $\mathrm{diag}(\gamma_n^{-\frac{1}{2}})$ to denote $\mathrm{diag}(\gamma_{n1},\ldots,\gamma_{np})$, $\mathrm{diag}(\sqrt{\gamma_{n1}},\ldots,\sqrt{\gamma_{np}})$ and $\mathrm{diag}(\gamma_{n1}^{-\frac{1}{2}},\ldots,\gamma_{np}^{-\frac{1}{2}})$ respectively, where $\gamma_n = (\gamma_{n1},\ldots,\gamma_{np})^T$.

Lemma 6. *Under Assumption 4, we have*

(a)

$$e^{-c_n/2}n^{-1/2}diag(\sqrt{\gamma_n})U_n(\beta_0) \xrightarrow{d} N(0,\Sigma(\beta_0)); \tag{3.10}$$

(b) For any consistent estimator $\tilde{\beta}_n$ *of* β_0*,*

$$e^{-c_n}n^{-1}diag(\sqrt{\gamma_n})I_n(\tilde{\beta}_n)diag(\sqrt{\gamma_n}) \xrightarrow{\mathbb{P}} \Sigma(\beta_0). \tag{3.11}$$

Proof. See the appendix for details. □

Corollary 2. *Under Assumption 4, (3.4), (3.5) and (3.6) hold.*

Proof. Take $D_n = e^{\frac{c_n}{2}}n^{\frac{1}{2}}\mathrm{diag}(\gamma_n^{-\frac{1}{2}})\Sigma^{\frac{1}{2}}$. Then, $||D_n^{-1}|| \to 0$ as $n \to \infty$ by Assumption 4 ((vi)). By Lemma 6, for any consistent estimator $\tilde{\beta}_n$ of β_0, we see that $D_n^{-1}I_n(\tilde{\beta}_n)(D_n^{-1})^T \xrightarrow{\mathbb{P}} I$ and $D_n^{-1}U_n(\beta_0) \xrightarrow{d} N(0,I)$. To verify Assumption 1 (iii), let

$$S^{(3)}_{n,jkl}(\beta_0,u) := \frac{1}{n}\sum_{i=1}^{n} Y_i^{(n)}(u)X_{ij}^{(n)}(u)X_{ik}^{(n)}(u)X_{il}^{(n)}(u)e^{\beta_0^T X_i^{(n)}(u)}$$

and fix $\alpha = (j,k,l)$. Then

$$\begin{aligned}
\left|\frac{\partial^3 L_n(\beta)}{\partial\beta_j\partial\beta_k\partial\beta_l}\right| &= \left|\sum_{i=1}^n \int_0^\tau \left\{ -2\frac{S_{n,j}^{(1)}(u)S_{n,k}^{(1)}(u)S_{n,l}^{(1)}(u)}{\{S_n^{(0)}(u)\}^3} - \frac{S_{n,jkl}^{(3)}(u)}{S_n^{(0)}(u)} \right.\right. \\
&\qquad \left.\left. + \frac{S_{n,jl}^{(2)}(u)S_{n,k}^{(1)}(u) + S_{n,j}^{(1)}(u)S_{n,kl}^{(2)}(u) + S_{n,jk}^{(2)}(u)S_{n,l}^{(1)}(u)}{\{S_n^{(0)}(u)\}^2} \right\} dN_i^{(n)}(u) \right| \\
&\leq \sum_{i=1}^n N_i^{(n)}(\tau)\left[2\sup_u \left|\frac{S_{n,j}^{(1)}(u)S_{n,k}^{(1)}(u)S_{n,l}^{(1)}(u)}{\{S_n^{(0)}(u)\}^3}\right| + \sup_u \left|\frac{S_{n,jkl}^{(3)}(u)}{S_n^{(0)}(u)}\right| \right. \\
&\qquad \left. + \sup_u \left|\frac{S_{n,jl}^{(2)}(u)S_{n,k}^{(1)}(u) + S_{n,j}^{(1)}(u)S_{n,kl}^{(2)}(u) + S_{n,jk}^{(2)}(u)S_{n,l}^{(1)}(u)}{\{S_n^{(0)}(u)\}^2}\right| \right], \qquad (3.12)
\end{aligned}$$

where the argument β is suppressed in the integrand for simplicity. Note that

$$\begin{aligned}
\sup_u |\sqrt{\gamma_{nj}} S_{n,j}^{(1)}(u)| &\leq \gamma_{nj}^{-\frac{1}{2}} \sup_u \{|\gamma_{nj} S_{n,j}^{(1)}(u) - s_j^{(1)}(u)| + |s_j^{(1)}(u)|\} \xrightarrow{\mathbb{P}} 0, \\
\sup_u |\sqrt{\gamma_{nj}\gamma_{nk}} S_{n,jk}^{(2)}(u)| &\leq \gamma_{nj}^{-\frac{1}{2}}\gamma_{nk}^{-\frac{1}{2}} \sup_u \{|n^{\gamma_j} n^{\gamma_k} S_{n,jk}^{(2)}(u) - s_{jk}^{(2)}(u)| + |s_{jk}^{(2)}(u)|\} \xrightarrow{\mathbb{P}} 0, \\
\sup_u \left|\sqrt{\gamma_{nj}\gamma_{nk}} S_{n,jkl}^{(3)}(\beta_0,u)\right| &\leq M \sup_u |\sqrt{\gamma_{nj}\gamma_{nk}} S_{n,jk}^{(2)}(\beta_0,u)| \xrightarrow{\mathbb{P}} 0, \qquad (3.13)
\end{aligned}$$

by Assumption 4 ((ii)) and ((iii)). Note that

$$(D_n^{-1}u)^\alpha = O(n^{-\frac{1}{2}} e^{-\frac{c_n}{2}} \sqrt{\gamma_{nj}\gamma_{nk}\gamma_{nl}}||u||^3) = o(\sqrt{\gamma_{nj}\gamma_{nk}}||u||^3) \qquad (3.14)$$

Hence, by (3.12), (3.13), (3.14), the fact that $n^{-1}e^{-c_n}\sum_{i=1}^n N_i^{(n)}(\tau) = O_P(1)$, and Assumption 4 ((iv)), we have

$$\sum_{|\alpha|=3} \frac{1}{\alpha!} \frac{\partial L_n(\beta_n)}{\partial\alpha} (D_n^{-1}u)^\alpha = o_{\mathbb{P}}(||u||^3). \qquad (3.15)$$

Finally, we verify $I_n(\tilde{\beta}_n)^{\frac{1}{2}}(D_n^{-1})^T \xrightarrow{\mathbb{P}} I$. Let $B_n := e^{\frac{c_n}{2}} n^{\frac{1}{2}} \mathrm{diag}(\gamma_n^{-\frac{1}{2}})$ and $C_n := I_n(\tilde{\beta}_n)^{\frac{1}{2}}$. Let $A_n = B_n C_n$. Note that we have $A_n A_n^T \xrightarrow{\mathbb{P}} \Sigma$. It suffices to show that $A_n \xrightarrow{\mathbb{P}} \Sigma^{\frac{1}{2}}$. Now, every subsequence has a further subsequences along which $A_n A_n^T \to \Sigma$ almost surely. Assume that $A_n A_n^T \to \Sigma$; otherwise argue along subsequences. Let b_{nj}'s be the diagonal elements in B_n.

Without loss of generality, assume that $b_{n1} \geq \ldots \geq b_{np}$. It is not hard to see that the limit Σ must be a block diagonal matrix, where the elements corresponding to the covariates with the same order are grouped in the same block. Since $A_n A_n^T \to \Sigma$, A_n is bounded, we can assume that $A_n \to A$ for some A; otherwise argue along subsequences. Let a_j be the rows of A. The fact that Σ is a block diagonal matrix implies that $\langle a_i, a_j \rangle = 0$ if i and j are not in the same block. We only consider the case when b_{nj}'s do not all have the same order of going to ∞ as the case when they are of the same order is straightforward to show. Let $a_{j^*}, \ldots, a_p$ be the rows corresponding to the smallest order of b_n's. Note that $b_{ni} C_{n,ij} = b_{ni} C_{n,ji} = b_{ni} O(b_{nj}) = o(1)$ for $i = j^*, \ldots, p$ and $j = 1, \ldots, j^* - 1$. Therefore, we see that $(B_n)_{i,j=j^*,\ldots,p} (C_n)_{i,j=j^*,\ldots,p}$ converges to C^*, where C^* is positive definite. Using the orthogonality of the rows in A across different blocks, we must have $a_{ij} = 0$ for $i = 1, \ldots, j^* - 1, j = j^*, \ldots, p$. Repeating this argument, we see that A is a positive definite block diagonal matrix. Since Σ is positive definite, it has an unique nonnegative definite square root matrix. Thus, $A = \Sigma^{\frac{1}{2}}$. □

3.5 Simulation Studies

In this section, we perform simulations to illuminate the theoretical results. We consider the proportional intensity model with rare events and two sparse covariates by specifying the intensity function as:

$$\lambda_n(t|\mathcal{F}_{t-}) = e^{c_n + \beta_0 + \beta_1 X_1^{(n)} + \beta_2 X_2^{(n)}}.$$

That is, the intensity function has a constant baseline with $\lambda_0(t) = e^{c_n + \beta_0}$. The two covariates take value 0 or 1 with their joint probabilities specified in Table 3.5, where $\alpha_1, \alpha_2, \alpha_3 > 0$. Clearly, $\mathbb{P}(X_1^{(n)} = 0) \to 1$ and $\mathbb{P}(X_2^{(n)} = 0) \to 1$. The correlation between $X_1^{(n)}$ and $X_2^{(n)}$ is

$$\begin{aligned} \mathrm{Corr}(X_1^{(n)}, X_2^{(n)}) &= \frac{\frac{1}{n^{\alpha_3}} - (\frac{1}{n^{\alpha_1}} + \frac{1}{n^{\alpha_3}})(\frac{1}{n^{\alpha_2}} + \frac{1}{n^{\alpha_3}})}{\sqrt{(\frac{1}{n^{\alpha_1}} + \frac{1}{n^{\alpha_3}})(1 - \frac{1}{n^{\alpha_1}} - \frac{1}{n^{\alpha_3}})(\frac{1}{n^{\alpha_2}} + \frac{1}{n^{\alpha_3}})(1 - \frac{1}{n^{\alpha_2}} - \frac{1}{n^{\alpha_3}})}} \\ &\sim \frac{\frac{1}{n^{\alpha_3}} - (\frac{1}{n^{\alpha_1}} + \frac{1}{n^{\alpha_3}})(\frac{1}{n^{\alpha_2}} + \frac{1}{n^{\alpha_3}})}{\sqrt{(\frac{1}{n^{\alpha_1}} + \frac{1}{n^{\alpha_3}})(\frac{1}{n^{\alpha_2}} + \frac{1}{n^{\alpha_3}})}}. \end{aligned}$$

We consider the following two settings:

(1) $\alpha_1 = \alpha_3 < \alpha_2$: The correlation is asymptotically equivalent to

$$\frac{\frac{1}{n^{\alpha_3}}}{\sqrt{\frac{2}{n^{\alpha_3}}\frac{1}{n^{\alpha_3}}}} = \frac{\sqrt{2}}{2}.$$

To find the asymptotic variance, first note that

$$\begin{aligned}
\mathbb{E}(X_1^{(n)} e^{\beta_1 X_1^{(n)} + \beta_2 X_2^{(n)}}) =& \mathbb{P}(X_1^{(n)} = 1, X_2^{(n)} = 1) e^{\beta_1+\beta_2} + \mathbb{P}(X_1^{(n)} = 1, X_2^{(n)} = 0) e^{\beta_1} \\
=& \frac{10}{n^{\alpha_3}} e^{\beta_1+\beta_2} + \frac{10}{n^{\alpha_1}} e^{\beta_1} = \frac{10}{n^{\alpha_1}} (e^{\beta_1+\beta_2} + e^{\beta_1}), \\
\mathbb{E}(X_2^{(n)} e^{\beta_1 X_1^{(n)} + \beta_2 X_2^{(n)}}) =& \mathbb{P}(X_1^{(n)} = 1, X_2^{(n)} = 1) e^{\beta_1+\beta_2} + \mathbb{P}(X_1^{(n)} = 1, X_2^{(n)} = 0) e^{\beta_2} \\
=& \frac{10}{n^{\alpha_3}} e^{\beta_1+\beta_2} + \frac{10}{n^{\alpha_2}} e^{\beta_2} \sim \frac{10}{n^{\alpha_1}} e^{\beta_1+\beta_2}, \\
\mathbb{E}(X_1^{(n)} X_2^{(n)} e^{\beta_1 X_1^{(n)} + \beta_2 X_2^{(n)}}) =& \mathbb{P}(X_1^{(n)} = 1, X_2^{(n)} = 1) e^{\beta_1+\beta_2} = \frac{10}{n^{\alpha_1}} e^{\beta_1+\beta_2}.
\end{aligned}$$

Note that

(i)

$$S_n^{(0)}(\beta, s) \xrightarrow{\mathbb{P}} 1 =: s^{(0)}(\beta, s).$$

(ii)

$$\operatorname{diag}(n^{\alpha_1}, n^{\alpha_1}) S_n^{(1)}(\beta, s) \xrightarrow{\mathbb{P}} \begin{pmatrix} 10(e^{\beta_1+\beta_2} + e^{\beta_1}) \\ 10 e^{\beta_1+\beta_2} \end{pmatrix} =: s^{(1)}(\beta, s).$$

(iii)

$$\operatorname{diag}(n^{\alpha_1/2}, n^{\alpha_1/2}) S_n^{(2)}(\beta, s) \operatorname{diag}(n^{\alpha_1/2}, n^{\alpha_1/2}) \xrightarrow{\mathbb{P}} \begin{pmatrix} 10(e^{\beta_1+\beta_2} + e^{\beta_1}) & 10 e^{\beta_1+\beta_2} \\ 10 e^{\beta_1+\beta_2} & 10 e^{\beta_1+\beta_2}. \end{pmatrix} =: s^{(2)}(\beta, s).$$

The asymptotic variance of $n^{1/2-c_n/2}\text{diag}(n^{\alpha_1/2}, n^{\alpha_1/2})(\hat{\beta}_n - \beta_0)$ is

$$\left\{ \int_0^\tau s^{(2)}(\beta, s)\lambda_0(s)ds \right\}^{-1} = \left\{ e^{\beta_0}\tau \begin{pmatrix} 10(e^{\beta_1+\beta_2} + e^{\beta_1}) & 10e^{\beta_1+\beta_2} \\ 10e^{\beta_1+\beta_2} & 10e^{\beta_1+\beta_2}. \end{pmatrix} \right\}^{-1}$$

In this case, the off-diagonal elements of the asymptotic variance matrix are non-zero.

(2) $\alpha_2 < \alpha_1 = \alpha_3$: The correlation is asymptotically equivalent to

$$\frac{\frac{1}{n^{\alpha_3}}}{\sqrt{\frac{2}{n^{\alpha_3}}\frac{1}{n^{\alpha_2}}}} \to 0.$$

To find the asymptotic variance, first note that

$$\begin{aligned}
\mathbb{E}(X_1^{(n)} e^{\beta_1 X_1^{(n)} + \beta_2 X_2^{(n)}}) =& \mathbb{P}(X_1^{(n)} = 1, X_2^{(n)} = 1)e^{\beta_1+\beta_2} + \mathbb{P}(X_1^{(n)} = 1, X_2^{(n)} = 0)e^{\beta_1} \\
=& \frac{10}{n^{\alpha_3}}e^{\beta_1+\beta_2} + \frac{10}{n^{\alpha_1}}e^{\beta_1} = \frac{10}{n^{\alpha_1}}(e^{\beta_1+\beta_2} + e^{\beta_1}), \\
\mathbb{E}(X_2^{(n)} e^{\beta_1 X_1^{(n)} + \beta_2 X_2^{(n)}}) =& \mathbb{P}(X_1^{(n)} = 1, X_2^{(n)} = 1)e^{\beta_1+\beta_2} + \mathbb{P}(X_1^{(n)} = 1, X_2^{(n)} = 0)e^{\beta_2} \\
=& \frac{10}{n^{\alpha_3}}e^{\beta_1+\beta_2} + \frac{10}{n^{\alpha_2}}e^{\beta_2} \sim \frac{10}{n^{\alpha_2}}e^{\beta_2}, \\
\mathbb{E}(X_1^{(n)} X_2^{(n)} e^{\beta_1 X_1^{(n)} + \beta_2 X_2^{(n)}}) =& \mathbb{P}(X_1^{(n)} = 1, X_2^{(n)} = 1)e^{\beta_1+\beta_2} = \frac{10}{n^{\alpha_1}}e^{\beta_1+\beta_2}.
\end{aligned}$$

Note that

(i)

$$S_n^{(0)}(\beta, s) \xrightarrow{\mathbb{P}} 1 =: s^{(0)}(\beta, s).$$

(ii)

$$\text{diag}(n^{\alpha_1}, n^{\alpha_2})S_n^{(1)}(\beta, s) \xrightarrow{\mathbb{P}} \begin{pmatrix} 10(e^{\beta_1+\beta_2} + e^{\beta_1}) \\ 10e^{\beta_2} \end{pmatrix}.$$

(iii)

$$\text{diag}(n^{\alpha_1/2}, n^{\alpha_2/2}) S_n^{(2)}(\beta, s) \text{diag}(n^{\alpha_1/2}, n^{\alpha_2/2}) \xrightarrow{\mathbb{P}} \begin{pmatrix} 10(e^{\beta_1+\beta_2} + e^{\beta_1}) & 0 \\ 0 & 10e^{\beta_2}. \end{pmatrix} =: s^{(2)}(\beta, s).$$

The asymptotic variance of $n^{1/2-c_n/2}\text{diag}(n^{\alpha_1/2}, n^{\alpha_2/2})(\hat{\beta}_n - \beta_0)$ is

$$\left\{ \int_0^\tau s^{(2)}(\beta, s)\lambda_0(s)ds \right\}^{-1} = e^{-\beta_0}\tau^{-1} \begin{pmatrix} 10^{-1}(e^{-(\beta_1+\beta_2)} + e^{-\beta_1}) & 0 \\ 0 & 10^{-1}e^{-\beta_2} \end{pmatrix}.$$

In this case, the off-diagonal elements of the asymptotic variance matrix are zero and β_1 and β_2 are asymptotically independent.

In the simulation, we set $c_n = -0.1 \log n \to -\infty$, $\beta_0 = -6$, $\beta_1 = 1$, $\beta_2 = 1.5$ and consider two sets of α_j's as shown in the last two columns in Table 3.5. As a remark, when the covariates are sparse, they correlation cannot be negative asymptotically because there are too many 0's. We estimate the parameters with sample size 100×2^j, for $j = 0, \ldots, 20$. For each sample size, we simulate 500 independent datasets.

In Figure 3.1, we compare the standard errors from the normalized observed Fisher information (purple line) for $\hat{\beta}_1$ and the estimated standard deviations of the normalized $\hat{\beta}_1$ (black line) with the asymptotic standard deviation from theoretical calculation (red line) under Setting 1. Figure 3.2 shows the corresponding results for $\hat{\beta}_2$ under setting 1. The corresponding results for $\hat{\beta}_1$ and $\hat{\beta}_2$ under Setting 2 are shown in Figures 3.4 and 3.5. In Figure 3.3, the standard errors of $\hat{\beta}_1$ and $\hat{\beta}_2$, and the estimated $\text{Cov}(\hat{\beta}_1, \hat{\beta}_2)$ from the observed Fisher information are compared with the corresponding estimated true quantities. The corresponding results in Setting 2 are shown in 3.6.

Prob	Setting	Setting 1	Setting 2
$\mathbb{P}(X_1^{(n)} = 1, X_2^{(n)} = 0)$	$10n^{-\alpha_1}$	$10n^{-1/2}$	$10n^{-3/4}$
$\mathbb{P}(X_1^{(n)} = 0, X_2^{(n)} = 1)$	$10n^{-\alpha_2}$	$10n^{-3/4}$	$10n^{-1/2}$
$\mathbb{P}(X_1^{(n)} = 1, X_2^{(n)} = 1)$	$10n^{-\alpha_3}$	$10n^{-1/2}$	$10n^{-3/4}$

Table 3.1: Simulation Setting

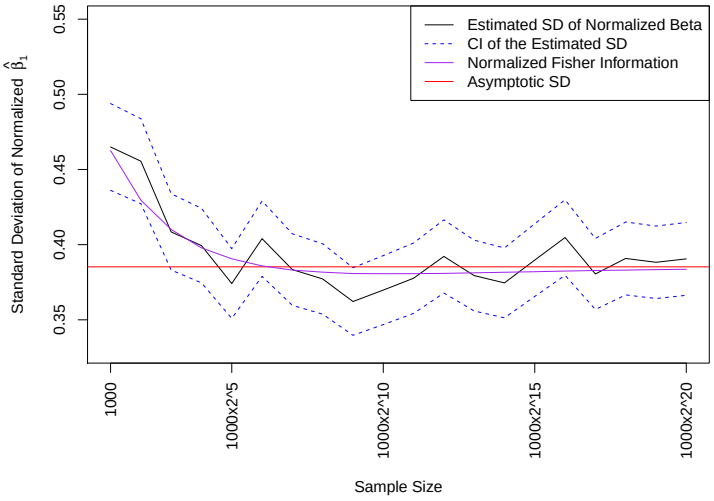

Figure 3.1: Comparison of the standard errors of the normalized β_1 in setting 1

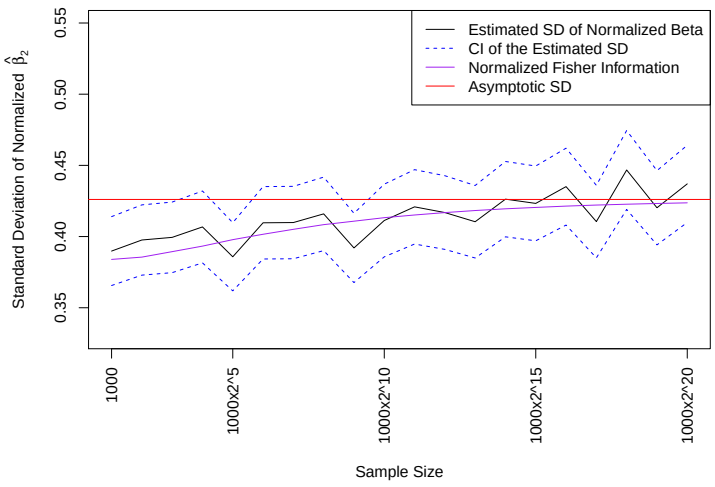

Figure 3.2: Comparison of the standard errors of the normalized β_2 in setting 1

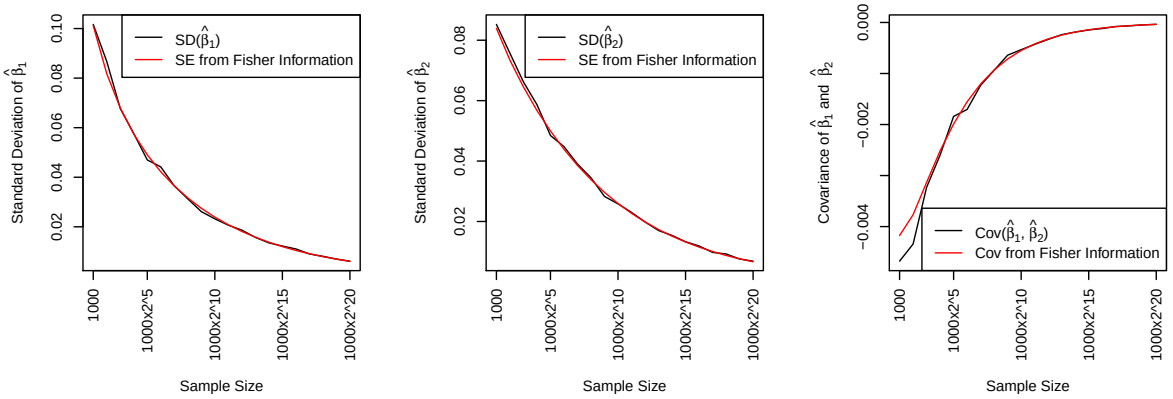

Figure 3.3: Comparison of the covariance matrix of $\hat{\beta}$ in setting 1

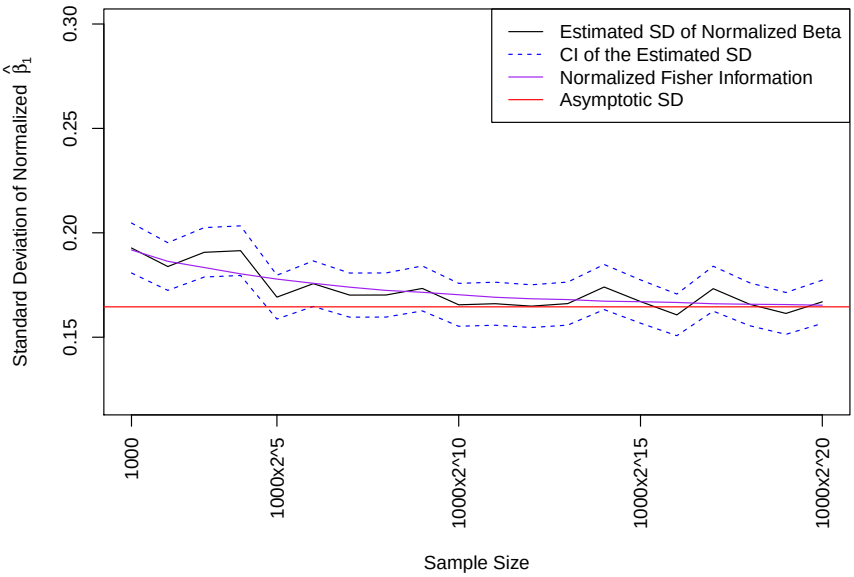

Figure 3.4: Comparison of the standard errors of the normalized β_1 in setting 2

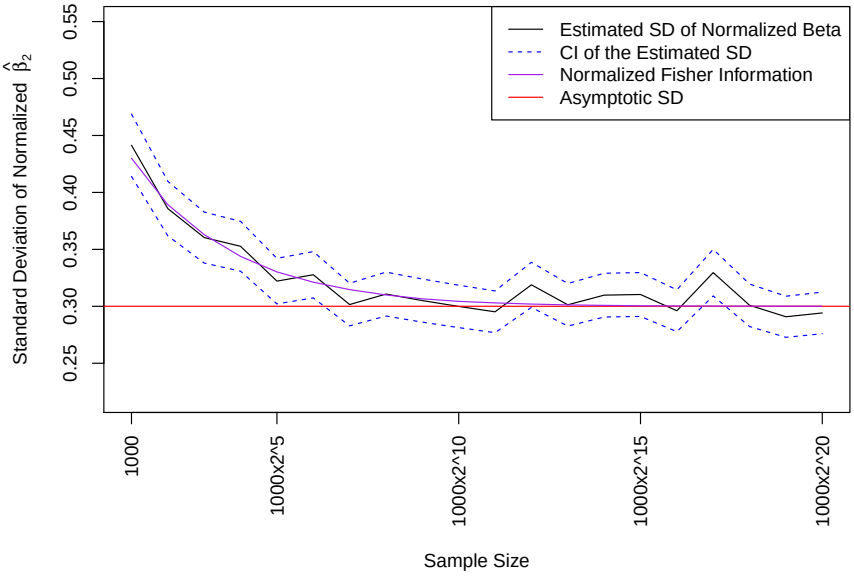

Figure 3.5: Comparison of the standard errors of the normalized β_2 in setting 2

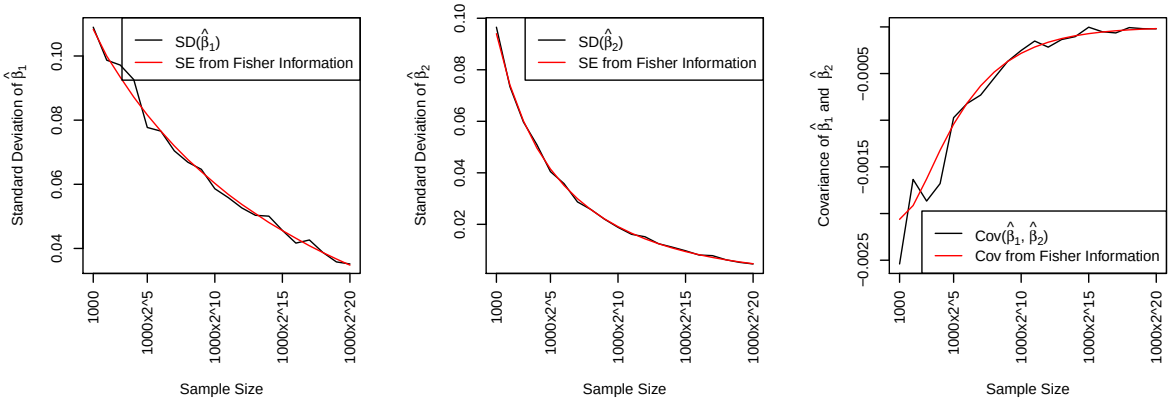

Figure 3.6: Comparison of the covariance matrix of $\hat{\beta}$ in setting 2

3.6 Appendix

Proof of Lemma 5. (a) We shall apply the martingale central limit theorem (see Theorem 5.3.5 in [19]). Let $U^{(n)}(t) := e^{-\frac{c_n}{2}} n^{-\frac{1}{2}} U_n(\beta_0, t)$ be the normalized score process. For any pair (l, l'), the predictable variation process

$$\begin{aligned}
& \langle U_l^{(n)}(\beta_0, \cdot), U_{l'}^{(n)}(\beta_0, \cdot) \rangle(t) \\
= \; & e^{-c_n} \int_0^t \frac{1}{n} \sum_{i=1}^n \left[\left\{ X_{il}(s) - \frac{S_{n,l}^{(1)}(\beta_0, s)}{S_n^{(0)}(\beta_0, s)} \right\} \left\{ X_{il'}^{(n)}(s) - \frac{S_{n,l'}^{(1)}(\beta_0, s)}{S_n^{(0)}(\beta_0, s)} \right\} \lambda_0(s) e^{c_n} Y_i^{(n)}(s) e^{\beta_0^T X_i^{(n)}(s)} \right] ds \\
= \; & \int_0^t V_{n,ll'}(\beta_0, s) S_n^{(0)}(\beta_0, s) \lambda_0(s) ds \\
\xrightarrow{\mathbb{P}} \; & \int_0^t v_{ll}(\beta_0, s) s^{(0)}(\beta_0, s) \lambda_0(s) ds,
\end{aligned}$$

where $S_{n,l}^{(1)}$ denotes the lth component of $S_n^{(1)}$, $V_{n,ll'}$ denotes the (l, l') element of V_n and the last convergence holds by Assumption 3 ((i)), ((ii)) and ((v)). Next, we verify the Lindeberg condition in the martingale central limit theorem. Define

$$H_{il}^{(n)}(s) := X_{il}^{(n)}(s) - \frac{S_{n,l}^{(1)}(\beta_0, s)}{S_n^{(0)}(\beta_0, s)}$$

and

$$\tilde{H}_{il}^{(n)}(s) \quad := \quad e^{-\frac{c_n}{2}} n^{-1/2} H_{il}^{(n)}(s).$$

For $\varepsilon > 0$, let

$$U_{l,\varepsilon}^{(n)}(\beta, t) := \sum_{i=1}^{n} \int_0^t \tilde{H}_{il}^{(n)}(s) I\{|\tilde{H}_{il}^{(n)}(s)| \geq \varepsilon\} dM_i^{(n)}(s),$$

Then,

$$\begin{aligned} & \langle U_{l,\varepsilon}^{(n)}, U_{l,\varepsilon}^{(n)} \rangle(t) \\ = & \sum_{i=1}^{n} \int_0^t \{\tilde{H}_{il}^{(n)}(s)\}^2 I\{|\tilde{H}_{il}^{(n)}(s) \geq \varepsilon\} Y_i^{(n)}(s) e^{\beta_0^T X_i^{(n)}(s)} \lambda_0(s) e^{c_n} ds \\ = & \frac{1}{n} \sum_{i=1}^{n} \int_0^t H_{il}^2(s) I\{e^{-\frac{c_n}{2}} n^{-1/2} |H_{il}(s)| \geq \varepsilon\} Y_i^{(n)}(s) e^{\beta_0^T X_i^{(n)}(s)} \lambda_0(s) ds. \end{aligned} \tag{3.16}$$

Let

$$E_n(\beta_0, t) := \frac{S_n^{(1)}(\beta_0, t)}{S_n^{(0)}(\beta_0, t)}.$$

Using the inequality

$$|a-b|^2 I\{|a-b| > \varepsilon\} \leq 4|a|^2 I\{|a| > \varepsilon/2\} + 4|b|^2 I\{|b| > \varepsilon/2\}, \quad \text{for any } a, b \in \mathbb{R},$$

(3.16) is bounded above by $T_{n1} + T_{n2}$, where

$$\begin{aligned} T_{n1} \quad := \quad & \frac{4}{n} \sum_{i=1}^{n} \int_0^\tau |X_{il}^{(n)}(u)|^2 I\{e^{-\frac{c_n}{2}} n^{-1/2} |X_{il}^{(n)}(u)| \geq \varepsilon/2\} Y_i^{(n)}(u) e^{\beta_0^T X_i^{(n)}(u)} \lambda_0(u) du \\ T_{n2} \quad := \quad & \frac{4}{n} \sum_{i=1}^{n} \int_0^\tau |E_{n,l}(\beta_0, u)|^2 I\{e^{-\frac{c_n}{2}} n^{-1/2} |E_{n,l}(\beta_0, u)| \geq \varepsilon/2\} Y_i^{(n)}(u) e^{\beta_0^T X_i^{(n)}(u)} \lambda_0(u) du. \end{aligned}$$

For T_{n1}, write

$$\{e^{-\frac{c_n}{2}} n^{-1/2} |X_{il}^{(n)}(t)| > \varepsilon/2\} = B_{1,i}^{n,l}(t) \bigcup B_{2,i}^{n,l}(t),$$

where

$$\begin{aligned} B_{1,i}^{n,l}(t) &:= \{e^{-\frac{c_n}{2}} n^{-1/2} |X_{il}^{(n)}(t)| > \varepsilon/2, \beta_0^T X_i^{(n)}(t) > -\delta ||X_i^{(n)}(t)||\}, \\ B_{2,i}^{n,l}(t) &:= \{e^{-\frac{c_n}{2}} n^{-1/2} |X_{il}^{(n)}(t)| > \varepsilon/2, \beta_0^T X_i^{(n)}(t) \leq -\delta ||X_i^{(n)}(t)||\}. \end{aligned}$$

Fix $\varepsilon' > 0$. By Condition ((iii)), with a probability more than $1 - \varepsilon'/2$, for all large n, we have $I\{B_{1,i}^{n,l}(u)\} Y_i^{(n)}(u) = 0$ for all $0 \leq u \leq \tau, 1 \leq i \leq n$. Hence,

$$\frac{1}{n} \int_0^\tau |X_{il}^{(n)}(u)|^2 I\{B_{1,i}^{n,l}(u)\} Y_i^{(n)}(u) e^{\beta_0^T X_i^{(n)}(u)} \lambda_0(u) du \xrightarrow{\mathbb{P}} 0.$$

Note that

$$\begin{aligned} & \frac{1}{n} \int_0^\tau |X_{il}^{(n)}(u)|^2 I\{B_{2,i}^{n,l}(u)\} Y_i^{(n)}(u) e^{\beta_0^T X_i^{(n)}(u)} \lambda_0(u) du \\ \leq \quad & \frac{1}{n} \int_0^\tau |X_{il}^{(n)}(u)|^2 I\{|X_{il}^{(n)}(u) > n^{1/2} e^{\frac{c_n}{2}} \varepsilon/2\} Y_i^{(n)}(u) e^{-\delta ||X_i^{(n)}(u)||} \lambda_0(u) du \\ \xrightarrow{\mathbb{P}} & 0, \end{aligned}$$

as $\sup_{i=1,\ldots,n} |X_{il}^{(n)}(u)|^2 I\{|X_{il}^{(n)}(u) > n^{1/2} e^{\frac{c_n}{2}} \varepsilon/2\} e^{-\delta ||X_i^{(n)}(u)||} \to 0$ as $n \to \infty$. For T_{n2}, by Condition ((ii)) and ((v)), with a probability more than $1-\varepsilon'/2$, for all large n, $I\{e^{-\frac{c_n}{2}} n^{-1/2} |E_{n,l}(\beta_0, u)| \geq \varepsilon/2\} = 0$. Hence, $T_{n2} \xrightarrow{\mathbb{P}} 0$. Therefore, we have shown that $\langle U_{l,\varepsilon}^{(n)}, U_{l,\varepsilon}^{(n)} \rangle(t) \xrightarrow{\mathbb{P}} 0$ for all $t \in [0, \tau]$. Hence, by the martingale central limit theorem, $U^{(n)}$ converges weakly in $D[0, \tau]^p$ to a mean zero p-variate Gaussian process such that each component process has independent increments and the covariance function at t for components l and l' is

$$\Sigma_{ll'}(\beta_0, t) = \int_0^t v_{ll'}(\beta_0, u) s^{(0)}(\beta_0, u) \lambda_0(u) du.$$

In particular, (3.8) holds.

(b) To show (3.9), note that for any consistent estimator $\tilde{\beta}_n$ of β_0,

$$\left\| e^{-c_n} n^{-1} I_n(\tilde{\beta}_n) - \Sigma(\beta_0) \right\| \leq E_{n1} + E_{n2} + E_{n3} + E_{n4},$$

where

$$\begin{aligned}
E_{n1} &:= \left\| \int_0^\tau \{V_n(\tilde{\beta}_n; t) - v(\tilde{\beta}_n; t)\} e^{-c_n} n^{-1} d\left(\sum_{i=1}^n N_i^{(n)}(t) \right) \right\| \\
E_{n2} &:= \left\| \int_0^\tau \{v(\tilde{\beta}_n; t) - v(\beta_0; t)\} e^{-c_n} n^{-1} d\left(\sum_{i=1}^n N_i^{(n)}(t) \right) \right\| \\
E_{n3} &:= \left\| \int_0^\tau v(\beta_0; t) d\left(e^{-c_n} n^{-1} \sum_{i=1}^n N_i^{(n)}(t) - n^{-1} \int_0^t \sum_{i=1}^n Y_i^{(n)}(s) e^{\beta_0^T X_i^{(n)}(s)} \lambda_0(s) ds \right) \right\| \\
E_{n4} &:= \left\| \int_0^\tau v(\beta_0; t) \{S_n^{(0)}(\beta_0; t) - s^{(0)}(\beta_0; t)\} \lambda_0(t) dt \right.
\end{aligned}$$

We shall show that these terms all converge to 0 in probability. We first observe that

$$e^{-c_n} n^{-1} \sum_{i=1}^n N_i^{(n)}(\tau) = O_P(1). \tag{3.17}$$

To see this, by Lemma 8.2.1 (1) in [19]

$$\begin{aligned}
\mathbb{P}\left(e^{-c_n} n^{-1} \sum_{i=1}^n N_i^{(n)}(t) > c \right) &= \mathbb{P}\left(\sum_{i=1}^n N_i^{(n)}(t) > e^{c_n} nc \right) \\
&\leq \frac{e^{c_n} n\delta}{e^{c_n} nc} + \mathbb{P}\left(\int_0^\tau \sum_{i=1}^n Y_i^{(n)}(u) e^{\beta_0^T X_i^{(n)}(u)} \lambda_0(u) e^{c_n} du > e^{c_n} n\delta \right) \\
&= \frac{\delta}{c} + \mathbb{P}\left(\int_0^\tau S_n^{(0)}(\beta_0; u) \lambda_0(u) du > \delta \right).
\end{aligned}$$

Hence, by Assumption 3 ((ii)),

$$\limsup_{n\to\infty} \mathbb{P}\left(e^{-c_n} n^{-1} \sum_{i=1}^n N_i^{(n)}(t) > c \right) \leq \frac{\delta}{c} + \mathbb{P}\left(\int_0^\tau s^{(0)}(\beta_0; u) \lambda_0(u) du > \delta \right).$$

Then, the claim in (3.17) follows by taking $\delta > \int_0^\tau s^{(0)}(\beta_0; u)\lambda_0(u)du$. Next, note that

$$E_{n1} \leq \sup_{0\leq u\leq\tau} ||V_n(\tilde{\beta}_n; u) - v(\tilde{\beta}_n; u)||e^{-c_n}n^{-1}\sum_{i=1}^{n} N_i^{(n)}(\tau) \xrightarrow{\mathbb{P}} 0,$$

by Assumption 3 ((ii)), ((v)) and (3.17). For E_{n2}, we have

$$E_{n2} \leq \sup_{0\leq u\leq\tau} ||v(\tilde{\beta}_n; u) - v(\beta_0; u)||e^{-c_n}n^{-1}\sum_{i=1}^{n} N_i^{(n)}(\tau) \xrightarrow{\mathbb{P}} 0,$$

by Assumption 3 ((v)), the consistency of $\tilde{\beta}_n$ and (3.17). To show E_{n3} converges to 0 in probability, note that for any pair j, k, by Lemma 8.2.1 (2) in [19],

$$\begin{aligned}
E_{n3,jk} &:= \mathbb{P}\left(\left|\int_0^\tau e^{-c_n}n^{-1}v(\beta_0; u)dM_i^{(n)}(u)\right| \geq \rho\right) \\
&\leq \mathbb{P}\left(\sup_{0\leq t\leq\tau}\left|\int_0^t e^{-c_n}n^{-1}v(\beta_0; u)dM_i^{(n)}(u)\right| \geq \rho\right) \\
&\leq \frac{\delta}{\rho^2} + \mathbb{P}\left(\int_0^\tau e^{-2c_n}n^{-2}v^2(\beta_0; u)\sum_{i=1}^{n} Y_i(t)e^{\beta_0^T X_i^{(n)}(t)}\lambda_0(t)e^{c_n}dt \geq \delta\right) \\
&= \frac{\delta}{\rho^2} + \mathbb{P}\left(e^{-c_n}n^{-1}\int_0^\tau v^2(\beta_0; u)S_n^{(0)}(\beta_0; u)du \geq \delta\right).
\end{aligned}$$

Hence,

$$\lim_{n\to\infty}\mathbb{P}\left(\left|\int_0^\tau e^{-c_n}n^{-1}v(\beta_0; u)dM_i^{(n)}(u)\right| \geq \rho\right) \leq \frac{\delta}{\rho^2}.$$

Since $\delta > 0$ is arbitrary, $E_{n3,jk} \to 0$. Therefore, $E_{n3} \xrightarrow{\mathbb{P}} 0$. Finally,

$$E_{n4} \leq \sup_{0\leq u\leq\tau} ||S_n^{(0)}(\beta_0; u) - s^{(0)}(\beta_0; u)|| \int_0^\tau v(\beta_0; u)\lambda_0(u)du \xrightarrow{\mathbb{P}} 0.$$

□

Proof of Lemma 6. The proof is similar to the proof of Lemma 5 and we only outline part of it.

(a) Let $U^{(n)}(t) := e^{-\frac{c_n}{2}}n^{-\frac{1}{2}}\text{diag}(\sqrt{\gamma_n})U_n(\beta_0, t)$ be the normalized score process. Its predictable

variation process is

$$\begin{aligned}
& \langle U_l^{(n)}(\beta_0,\cdot), U_{l'}^{(n)}(\beta_0,\cdot)\rangle(t) \\
= \quad & e^{-c_n}\int_0^t \frac{\sqrt{\gamma_{nl}\gamma_{nl'}}}{n}\sum_{i=1}^n \left\{X_{il}^{(n)}(s) - \frac{S_{n,l}^{(1)}(\beta_0,s)}{S_n^{(0)}(\beta_0,s)}\right\}\left\{X_{il'}^{(n)}(s) - \frac{S_{n,l'}^{(1)}(\beta_0,s)}{S_n^{(0)}(\beta_0,s)}\right\} \\
& \qquad \times Y_i^{(n)}(s) e^{\beta_0^T X_i^{(n)}(s)}\lambda_0(s)e^{c_n}ds \\
= \quad & \sqrt{\gamma_{nl}\gamma_{nl'}}\int_0^t V_{n,ll'}(\beta_0,s)S_n^{(0)}(\beta_0,s)\lambda_0(s)ds \\
\xrightarrow{\mathbb{P}} \quad & \int_0^t v_{ll'}(\beta_0,s)s^{(0)}(\beta_0,s)\lambda_0(s)ds,
\end{aligned}$$

where $S_{n,l}^{(1)}$ denotes the lth component of $S_n^{(1)}$ and the last convergence holds by Assumption 4 ((i)), ((ii)) and ((iv)). Next, we verify the Lindeberg condition in the martingale central limit theorem. Define

$$H_{il}^{(n)}(s) \quad := \quad X_{il}^{(n)}(s) - \frac{S_{n,l}^{(1)}(\beta_0,s)}{S_n^{(0)}(\beta_0,s)}$$

and

$$\tilde{H}_{il}^{(n)}(s) \quad := \quad e^{-\frac{c_n}{2}}n^{-1/2}\sqrt{\gamma_{nl}}H_{il}^{(n)}(s).$$

For $\varepsilon > 0$, let

$$U_{l,\varepsilon}^{(n)}(\beta,t) := \sum_{i=1}^n \int_0^t \tilde{H}_{il}^{(n)}(s) I\{|\tilde{H}_{il}^{(n)}(s) \geq \varepsilon\} dM_i^{(n)}(s),$$

The rest is essentially the same as in the proof of Lemma 5.

(b) To show (3.10), note that for any consistent estimator $\tilde{\beta}_n$ of β_0,

$$\left\| e^{-c_n}n^{-1}\mathrm{diag}(\sqrt{\gamma_n})I_n(\tilde{\beta}_n)\mathrm{diag}(\sqrt{\gamma_n}) - \Sigma(\beta_0)\right\| \leq E_{n1} + E_{n2} + E_{n3} + E_{n4},$$

where

$$
\begin{aligned}
E_{n1} &:= \left\| \int_0^\tau \{\mathrm{diag}(\sqrt{\gamma_n}) V_n(\tilde{\beta}_n; t) \mathrm{diag}(\sqrt{\gamma_n}) - v(\tilde{\beta}_n; t)\} e^{-c_n} n^{-1} d\left(\sum_{i=1}^n N_i^{(n)}(t) \right) \right\| \\
E_{n2} &:= \left\| \int_0^\tau \{v(\tilde{\beta}_n; t) - v(\beta_0; t)\} e^{-c_n} n^{-1} d\left(\sum_{i=1}^n N_i^{(n)}(t) \right) \right\| \\
E_{n3} &:= \left\| \int_0^\tau v(\beta_0; t) d\left(e^{-c_n} n^{-1} \sum_{i=1}^n N_i^{(n)}(t) - n^{-1} \int_0^t \sum_{i=1}^n Y_i^{(n)}(s) e^{\beta_0^T X_i^{(n)}(s)} \lambda_0(s) ds \right) \right\| \\
E_{n4} &:= \left\| \int_0^\tau v(\beta_0; t) \{S_n^{(0)}(\beta_0; t) - s^{(0)}(\beta_0; t)\} \lambda_0(t) dt \right\|.
\end{aligned}
$$

Each of these terms converges to 0 in probability as in the proof of Lemma 5.

□

Chapter 4: Discussion

In Chapter 2, we proposed a multivariate proportional intensity factor model for multivariate event time data. We develop the theory of nonparametric maximum likelihood estimation as well as a variable selection and estimation method for the fixed effects and random effects simultaneously using parametric baseline intensity functions. From the simulation studies, we see that using the Bayesian information criterion provides a good choice of the tuning parameter and the whole procedure essentially recovers the true structure of the parameter with small bias and accurate standard errors. We further demonstrate the proposed method through a real data set from the Survey of Adult Skills in PIAAC. Our method finds meaningful relationships among different types of events that can help understanding both the task design and the behavior of subjects when solve a problem. Furthermore, the proposed method can be applied to both exploratory and confirmatory analysis or a combination of them by controlling the number of constraints on the loading matrix.

Although we implicitly assume all the event types are recurrent, we can also allow some events to be survival times. For the distribution of the random effects, the multivariate normal distribution allows an unrestricted covariance structure between the random effects. However, other distributions can also be used and the theoretical results remain valid subject to some regularity conditions on the random effect distributions; see [52] for more details. The proposed model can also be easily extended to have a multilevel structure, where we could have, for example, a cluster level above the subject level with cluster-specific random effects.

While we illustrate the method using educational assessment data, the method is widely applicable. For example, in medical studies, for each person, we are often interested in several illnesses at the same time. When the number of random coefficients is moderate to large, the proposed model can achieve a parsimonious model.

In Chapter 3, we establish the consistency, asymptotic normality and the validity of the usual

inference procedure using the maximum partial likelihood estimator in the proportional intensity model under general conditions. We verify these conditions under setting with rare events and sparse covariates which are common in large-scale observational databases. A future direction is to study the corresponding results in a multivariate model with random effects, which is a more realistic setting because of the unobserved covariates that induces dependence between different events.

References

[1] O. Aalen, "A model for nonparametric regression analysis of counting processes," in *Mathematical statistics and probability theory*, Springer, 1980, pp. 1–25.

[2] P. K. Andersen, O. Borgan, R. D. Gill, and N. Keiding, *Statistical models based on counting processes*. Springer Science & Business Media, 1993.

[3] P. K. Andersen and R. D. Gill, "Cox's regression model for counting processes: A large sample study," *The annals of statistics*, pp. 1100–1120, 1982.

[4] F. J. Aranda-Ordaz, "An extension of the proportional-hazards model for grouped data," *Biometrics*, pp. 109–117, 1983.

[5] Y. Bergner and A. A. von Davier, "Process data in naep: Past, present, and future," *Journal of Educational and Behavioral Statistics*, p. 1 076 998 618 784 700, 2018.

[6] S. Buyske, R. Fagerstrom, and Z. Ying, "A class of weighted log-rank tests for survival data when the event is rare," *Journal of the American Statistical Association*, vol. 95, no. 449, pp. 249–258, 2000.

[7] G. Celeux and J. Diebolt, "The sem algorithm: A probabilistic teacher algorithm derived from the em algorithm for the mixture problem," *Computational statistics quarterly*, vol. 2, pp. 73–82, 1985.

[8] J. Choi, G. Oehlert, and H. Zou, "A penalized maximum likelihood approach to sparse factor analysis," *Statistics and its Interface*, vol. 3, no. 4, pp. 429–436, 2010.

[9] R. J. Cook and J. Lawless, *The statistical analysis of recurrent events*. Springer Science & Business Media, 2007.

[10] D. R. Cox, "Regression models and life-tables," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 34, no. 2, pp. 187–202, 1972.

[11] D. R. Cox and D. Oakes, *Analysis of survival data*. CRC Press, 1984, vol. 21.

[12] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 39, pp. 1–38, 1977.

[13] D. L. Donoho and J. M. Johnstone, "Ideal spatial adaptation by wavelet shrinkage," *Biometrika*, vol. 81, no. 3, pp. 425–455, 1994.

[14] J. Fan and R. Li, "Variable selection via nonconcave penalized likelihood and its oracle properties," *Journal of the American statistical Association*, vol. 96, no. 456, pp. 1348–1360, 2001.

[15] ——, "Variable selection for cox's proportional hazards model and frailty model," *Annals of Statistics*, vol. 30, pp. 74–99, 2002.

[16] J. Fan, H. Peng, *et al.*, "Nonconcave penalized likelihood with a diverging number of parameters," *The Annals of Statistics*, vol. 32, no. 3, pp. 928–961, 2004.

[17] C. Farrington, "Relative incidence estimation from case series for vaccine safety evaluation," *Biometrics*, pp. 228–235, 1995.

[18] C. Farrington and H. Whitaker, "Semiparametric analysis of case series data," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, vol. 55, no. 5, pp. 553–594, 2006.

[19] T. R. Fleming and D. P. Harrington, *Counting processes and survival analysis*. John Wiley & Sons, 2011, vol. 169.

[20] J. Friedman, T. Hastie, and R. Tibshirani, "Regularization paths for generalized linear models via coordinate descent," *Journal of Statistical Software*, vol. 33, no. 1, p. 1, 2010.

[21] M. Gail, T. Santner, and C. Brown, "An analysis of comparative carcinogenesis experiments based on multiple times to tumor," *Biometrics*, pp. 255–266, 1980.

[22] P. F. Halpin, A. A. von Davier, J. Hao, and L. Liu, "Measuring student engagement during collaboration," *Journal of Educational Measurement*, vol. 54, no. 1, pp. 70–84, 2017.

[23] J. Hao, Z. Shu, and A. von Davier, "Analyzing process data from game/scenario-based tasks: An edit distance approach," *Journal of Educational Data Mining*, vol. 7, no. 1, pp. 33–50, 2015.

[24] A. G. Hawkes, "Spectra of some self-exciting and mutually exciting point processes," *Biometrika*, vol. 58, no. 1, pp. 83–90, 1971.

[25] Q. He and M. von Davier, "Analyzing process data from problem-solving items with n-grams: Insights from a computer-based large-scale assessment," in *Handbook of research on technology tools for real-world skill development*, IGI Global, 2016, pp. 750–777.

[26] K. Hirose and M. Yamamoto, "Sparse estimation via nonconcave penalized likelihood in factor analysis model," *Statistics and Computing*, vol. 25, no. 5, pp. 863–875, 2015.

[27] J. D. Kalbfleisch and R. L. Prentice, *The statistical analysis of failure time data*. John Wiley & Sons, 2002, vol. 360.

[28] E. W. Lee, L. Wei, D. A. Amato, and S. Leurgans, "Cox-type regression analysis for large numbers of small groups of correlated failure time observations," in *Survival analysis: state of the art*, Springer, 1992, pp. 237–247.

[29] Y.-H. Lee and Y. Jia, "Using response time to investigate students' test-taking behaviors in a naep computer-based study," *Large-scale Assessments in Education*, vol. 2, no. 1, p. 8, 2014.

[30] K.-Y. Liang, S. G. Self, and Y.-C. Chang, "Modelling marginal hazards in multivariate failure time data," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 441–453, 1993.

[31] D. Y. Lin, L.-J. Wei, I Yang, and Z. Ying, "Semiparametric regression for the mean and rate functions of recurrent events," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 62, no. 4, pp. 711–730, 2000.

[32] D. Lin and Z. Ying, "Semiparametric analysis of the additive risk model," *Biometrika*, vol. 81, no. 1, pp. 61–71, 1994.

[33] D. Lin, Z. Ying, *et al.*, "Semiparametric analysis of general additive-multiplicative hazard models for counting processes," *The annals of Statistics*, vol. 23, no. 5, pp. 1712–1734, 1995.

[34] H. Liu, Y. Liu, and M. Li, "Analysis of process data of pisa 2012 computer-based problem solving: Application of the modified multilevel mixture irt model," *Frontiers in psychology*, vol. 9, 2018.

[35] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, "Equation of state calculations by fast computing machines," *The journal of chemical physics*, vol. 21, no. 6, pp. 1087–1092, 1953.

[36] L. Ning and T. T. Georgiou, "Sparse factor analysis via likelihood and $l1$-regularization," in *2011 50th IEEE Conference on Decision and Control and European Control Conference*, IEEE, 2011, pp. 5188–5192.

[37] R. L. Prentice, B. J. Williams, and A. V. Peterson, "On the regression analysis of multivariate failure time data," *Biometrika*, vol. 68, no. 2, pp. 373–379, 1981.

[38] G. Schwarz, "Estimating the dimension of a model," *The Annals of Statistics*, vol. 6, no. 2, pp. 461–464, 1978.

[39] Z. Shu, Y. Bergner, M. Zhu, J. Hao, and A. A. von Davier, "An item response theory analysis of problem-solving processes in scenario-based tasks," *Psychological Test and Assessment Modeling*, vol. 59, no. 1, p. 109, 2017.

[40] S. E. Simpson, "A positive event dependence model for self-controlled case series with applications in postmarketing surveillance," *Biometrics*, vol. 69, no. 1, pp. 128–136, 2013.

[41] S. E. Simpson, D. Madigan, I. Zorych, M. J. Schuemie, P. B. Ryan, and M. A. Suchard, "Multiple self-controlled case series for large-scale longitudinal observational databases," *Biometrics*, vol. 69, no. 4, pp. 893–902, 2013.

[42] J. Sun, Y. Chen, J. Liu, Z. Ying, and T. Xin, "Latent variable selection for multidimensional item response theory models via l_1 regularization," *Psychometrika*, vol. 81, no. 4, pp. 921–939, 2016.

[43] D. C. Thomas, "Use of auxiliary information in fitting nonproportional hazards models," *Modern statistical methods in chronic disease epidemiology*, vol. 197210, 1986.

[44] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996.

[45] F. Vaida and R. Xu, "Proportional hazards model with random effects," *Statistics in medicine*, vol. 19, no. 24, pp. 3309–3324, 2000.

[46] J. W. Vaupel, K. G. Manton, and E. Stallard, "The impact of heterogeneity in individual frailty on the dynamics of mortality," *Demography*, vol. 16, no. 3, pp. 439–454, 1979.

[47] M. Wedel, U. Böckenholt, and W. A. Kamakura, "Factor models for multivariate count data," *Journal of Multivariate Analysis*, vol. 87, no. 2, pp. 356–369, 2003.

[48] L.-J. Wei, D. Y. Lin, and L. Weissfeld, "Regression analysis of multivariate incomplete failure time data by modeling marginal distributions," *Journal of the American statistical association*, vol. 84, no. 408, pp. 1065–1073, 1989.

[49] A. I. Yashin, J. W. Vaupel, and I. A. Iachine, "Correlated individual frailty: An advantageous approach to survival analysis of bivariate data," *Mathematical population studies*, vol. 5, no. 2, pp. 145–159, 1995.

[50] D Zeng and D. Lin, "Maximum likelihood estimation in semiparametric regression models with censored data," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 69, no. 4, pp. 507–564, 2007.

[51] D. Zeng and D. Lin, "Semiparametric transformation models with random effects for recurrent events," *Journal of the American Statistical Association*, vol. 102, no. 477, pp. 167–180, 2007.

[52] ——, "A general asymptotic theory for maximum likelihood estimation in semiparametric regression models with censored data," *Statistica Sinica*, vol. 20, no. 2, p. 871, 2010.

[53] C.-H. Zhang *et al.*, "Nearly unbiased variable selection under minimax concave penalty," *The Annals of statistics*, vol. 38, no. 2, pp. 894–942, 2010.

[54] M. Zhu, Z. Shu, and A. A. von Davier, "Using networks to visualize and analyze process data for educational assessment," *Journal of Educational Measurement*, vol. 53, no. 2, pp. 190–211, 2016.

[55] H. Zou and R. Li, "One-step sparse estimates in nonconcave penalized likelihood models," *Annals of Statistics*, vol. 36, no. 4, p. 1509, 2008.