UNIVERSITY OF CALIFORNIA SAN DIEGO

Modern Statistical Methods for Complex Survival Data

A dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy

in

Mathematics (with a specialization in Statistics)

by

Jue Hou

Committee in charge:

Professor Ronghui Xu, Chair
Professor Ery Arias-Castro
Professor Jelena Bradic
Professor Christina Chambers
Professor James Murphy

2019

The dissertation of Jue Hou is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

_____

_____

Chair

University of California San Diego

2019

DEDICATION

To future readers, the true witnesses to the value of my work.

EPIGRAPH

... the triumph of my art is in thoroughly examining whether the thought which the mind of the young man brings forth is a false idol or a noble and true birth.

—Socrates in Plato's *Theatetus*

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to first express my sincere appreciation and gratitude to my advisor Professor Ronghui Xu, who has been a tremendous mentor for me. Under your persistent guidance, I have learned from your model to work toward the excellence of scholarship. You have put a special attention in my train on finding the authentic motivation, as well as giving transparent and communicable presentation. All of these have not only prepared me for my future research, but also inspired me to see the ultimate worth of our discipline. You have instilled in me the patience for perfecting every detail, which I wholeheartedly believe will lead me to achievements with lasting impact. In the end, you got me a dream job better than you promised, a favor I can ever repay with words. Outside the academia, I appreciate your hospitality and approachability. I want to thank you for all the good memories.

My other special appreciation goes to Professor Jelena Bradic, who has invested so much in my training no less than a mentor. You have led me into the world of Big Data and contemplated the wonderful research topics on high-dimensional survival analysis. I am also deeply moved by your passion toward work, in all aspects of research, mentor, teaching and service. Serving under you as teaching assistant, I have witnessed how you gather the talents by sharing the beauty of Statistics and send them to top graduate programs. The memory about how you excel at the multiple fronts will become my secret weapon to survive along my career. Moreover, I would like to thank you for being so kind and caring in person.

I would like to thank Professor Ery Arias-Castro, Professor Christina Chambers and Professor James Murphy for serving as my committee members. I also want to thank Professor

Christina Chambers and Professor James Murphy for bringing your research and data to me and offering insights to my data analysis results. I would like to thank Professor Ruth Williams for your profound influence on me. You are the first faculty member I know in the department. I still remember your talk about the amazing application in Biology. Your class on Stochastic Process turned out to be the foundation of my research. I also want to thank you for setting a great example in teaching. You letter on my teaching for my job applications has been a great encouragement. I would like to thank Professor Dimitris Politis for the fascinating class on Bootstrap and speaking highly of me behind my back–that message gave me quite a boost of confidence.

I would like to thank the current and past staffs in Department Mathematics, Scott Rollans, Holly Proudfoot, Wilson Cheung, Terry Le, Debbie Shon, Derrick Hwa, Leslie Foley and Kelly Guerriero, for your service and support. I want to thank the friends I make in the departments, Zhehua Li, Hanbo Li, Yuchao Liu, Jiaqi Guo, Selene Xu, Xiao Pu, Jingwen Liang, Ruixuan Zhou, Andrew Ying, Rong Huang, Denise Rava, Davide Viviano and Xiao-ou Pan, for spending time together on academic activities and for fun.

I would like to thank Professor Lei Liu and Professor Ann Ragin for offering me the research opportunity at Northwestern University. You have introduced me to the application side of Biostatistics and helped me earning the first publications. To Professor Lei Liu, I also want to thank you for constant support in my various applications. I would like to thank Professor Zhiliang Ying for suggesting me the Ph. D. program here and offering the support in my application.

I would like to thank the brothers and sisters in my spiritual home, the Living Water Bible Church for the loving support in Christ and on earth. A special appreciation goes to my

godparents, Dr Dennis Cheng and Dr Qun Cheng, through whom I have learned about the gospel and seen the image of Christ.

Finally, I have my reserved thanks for my families including my wife, my parents, my parents in-law, my son in my wife's womb and my Pembroke Corgi. In your accompany, I have experienced so much in life beyond this thesis so that I have a clearer idea about the role of my work in my big picture about life. Related to the work, I want to thank my wife for being my first student and first reader. I want to thank my parents for passing to me your medical knowledge and experience, which has helped me greatly in my collaboration. I want to thank my parents and in-laws for supporting me and my wife financially when the stipend falls short to cover the skyrocketing living expense in San Diego.

Chapter 1, in full, is a reprint of the material as that appears in Lifetime Data Analysis. Hou, Jue; Chambers, Christina; Xu, Ronghui. A nonparametric maximum likelihood approach for survival data with observed cured subjects, left truncation and right-censoring, LIDA, **24**(4) 612-651, 2018. The dissertation/thesis author was the primary investigator and author of this paper.

Chapter 2, in full, has been accepted for publication of the material as it may appear in the Electronic Journal of Statistics. Hou, Jue; Bradic, Jelena; Xu, Ronghui. Inference under Fine-Gray competing risks model with high-dimensional covariates. The dissertation/thesis author was the primary investigator and author of this paper.

Chapter 3, in full, is currently being prepared for submission for publication of the material. Hou, Jue; Bradic, Jelena; Xu, Ronghui. Estimating treatment effect for time-to-event outcome with high-dimensional covariates in observational studies . The dissertation/thesis author is the

primary investigator and author of this material.

VITA

2011	B. S. in Mathematics, Fudan University
2013	M. S. in Statistics, University of Illinois at Urbana-Champaign
2013,2014,2015	Research Assistant, Northwestern University Feinberg School of Medicine
2014	Teaching Assistant, University of Illinois at Urbana-Champaign
2014-2019	Teaching Assistant, University of California San Diego
2016	C. Phil. in Mathematics, University of California San Diego
2016	Graduate Researcher, University of California San Diego Center for Better Beginnings
2019	Ph. D. in Mathematics with Specialization in Statistics, University of California San Diego

PUBLICATIONS

Hou Jue, Seneviratne C., Su X., Taylor J., Johnson B., Wang X.-Q., Zhang H., Kranzler H. R., Kang J. and Liu L.. "Subgroup identification in personalized treatment of alcohol dependence", *Alcoholism: Clinical and Experimental Research* **39**(7): 1253-1259, 2015.

Aanchal P., Hou Jue, Liu L., Gao Y., Kettering C. and Ragin A. B.. "Cognitive Function in Early HIV Infection", *Journal of NeuroVirology* **23**(3): 273-282, 2017.

Hou Jue, Chambers C. D. and Xu R.. "A nonparametric maximum likelihood approach for survival data with observed cured subjects, left truncation and right-censoring", *Lifetime Data Analysis* **24**(4): 612-651, 2017

Xu R., Hou Jue and Chambers C. D.. "The impact of confounder selection in propensity scores when applied to prospective cohort studies in pregnancy", *Reproductive Toxicology* **78**: 75-80, 2018.

Hou J., Paravati A., Hou Jue, Xu R. and Murphy J.. "High-dimensional variable selection and prediction under competing risks with application to SEER-Medicare linked data", *Statistics in Medicine* **37**:3486-3502, 2018.

Hou Jue, Bradic J. and Xu R.. "Inference under Fine-Gray Competing Risks Model with High-Dimensional Covariates", *to appear in Electronic Journal of Statistics*, 2019.

Hou Jue, Bradic J. and Xu R.. "Estimating treatment effect for time-to-event outcome with high-dimensional covariates in observational studies", *Manuscript in Preparation*, 2019.

ABSTRACT OF THE DISSERTATION

Modern Statistical Methods for Complex Survival Data

by

Jue Hou

Doctor of Philosophy in Mathematics (with a specialization in Statistics)

University of California San Diego, 2019

Professor Ronghui Xu, Chair

With the booming of big complex data, various Statistical methods and Data Science techniques have been developed to retrieve valuable information from them. The progress is slower with survival data due to the additional difficulty from censoring and truncation. Except for a few straightforward extensions, most modern learning methods have been absent in survival analysis for years since their invention. The theory on the survival version of those methods also falls further behind. There is a strong demand on computational efficient and theoretical reliable methods for big complex data with time-to-event outcomes in various Health related fields where

immense resource has been poured into.

This thesis is devoted to incorporating censoring and truncation to state-of-art Statistical methodology and theory, to promote the evolution of survival analysis and support Medical research with up-to-date tools. In Chapter 1, I study the mixture cure-rate model with left truncation and right-censoring. We propose a Nonparametric Maximum Likelihood Estimation (NPMLE) approach to effectively handle the truncation issue. We adopt an efficient and stable EM algorithm. We are able to give a closed form variance estimator giving rise to valid inference. In Chapter 2, I study the estimation and inference for the Fine-Gray competing risks model with high-dimensional covariates. We develop confidence intervals based on a one-step bias-correction to an initial regularized estimator. We lay down a methodological and theoretical framework for the one-step bias-corrected estimator with the partial likelihood. In Chapter 3, I study the inference on treatment effect with censored time-to-event outcome while adjusting for high-dimensional covariates. We propose an orthogonal score method to construct honest confidence intervals for the treatment effect. With a slight modification, we obtain a doubly robust estimator extremely tolerant to both estimation inconsistency and volatility. All the methods in aforementioned chapters are tested through extensive numerical experiments and applied on real data with authentic medical interests.

Chapter 1

A Nonparametric Maximum Likelihood Approach for Survival Data with Observed Cured Subjects, Left Truncation and Right-Censoring

1.1 Introduction

Our work was motivated by research carried out at the Organization of Teratology Information Specialists (OTIS), which is a North American network of university or hospital based teratology services that counsel between 70,000 and 100,000 pregnant women every year. Research subjects are enrolled from the Teratology Information Services and through other methods of recruitment, where the mothers and their babies are followed over time. Phone interviews

are conducted through the length of the pregnancy along with pregnancy diaries recorded by the mother. An outcome phone interview is conducted shortly after the pregnancy ends, and if it results in a live birth, a dysmorphology exam is done within six months and with further follow-ups at one year and possibly later dates. Recently it has been of interest to assess the effects of medication exposures on spontaneous abortion (SAB) [XC11, CJXJ11]. Here we examine the OTIS autoimmune disease in pregnancy database for risk factors as well as effects of medications on spontaneous abortion.

By definition SAB occurs within the first 20 weeks of gestation; any spontaneous pregnancy loss after that is called still birth. Ultimately we would like to know if an exposure modifies the risk of SAB for a woman, which may be increased or decreased. It is known that in the population for clinically recognized pregnancies the rate of SAB is about 12% [WWO$^+$88]. On the other hand, in our database the empirical SAB rate is consistently lower than 10%. This is due to the fact that women may enter a study any time before 20 weeks' gestation. Figure 1.1 left panel shows the histograms of study entry times up to 20 weeks of gestation from our autoimmune disease in pregnancy database. This way women who have early SAB events are less likely to be captured in our studies, and such selection bias is known as left truncation in survival analysis. Left truncation has been studied by many authors since the 1980s, and has attracted much recent attention in the context of length-biased data [AWZ06, QNLS11, among others]. In addition the women are subject to right-censoring due to loss to follow-up. Figure 1.1 right panel shows the left truncated as well as right-censored Kaplan-Meier curve for the SAB event.

As seen from the Kaplan-Meier curve the majority of the pregnant women are free of SAB; they are considered 'cured' in the time-to-event context. Cure models are used in

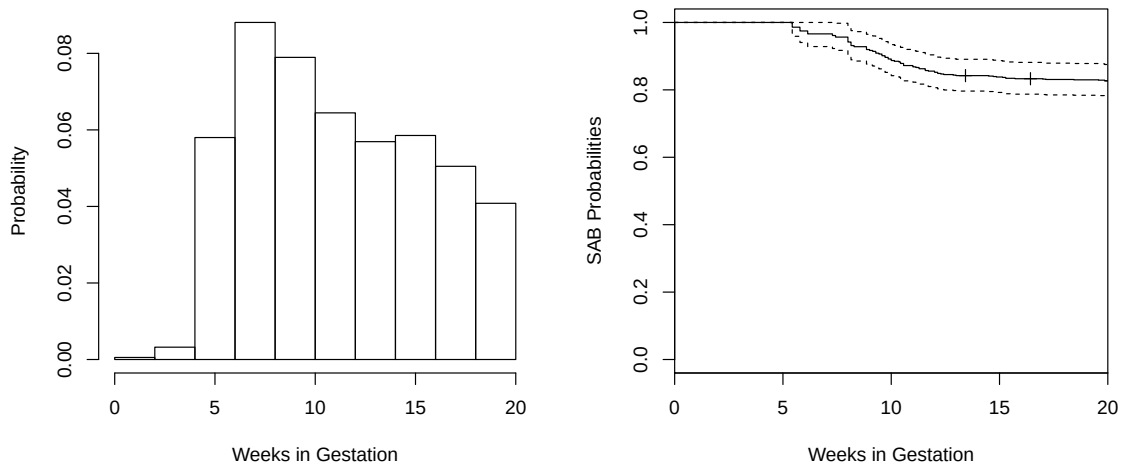


Figure 1.1: Study entry times for all individuals in the SAB data (left), and left truncated Kaplan-Meier curves (95% confidence intervals) for the SAB events (right).

various biomedical studies where data often include a substantial portion of 'long-term' survivors [Far82, Far86]. Our data, however, differ from classic cure data where the 'cured' subjects are always right-censored and never actually observed to be cured [ST00, LY04]. In our case, 'cured' is defined as surviving 20 weeks of gestation, and we observe over 80% of our subjects as cured from SAB.

Cure rate models are well studied in the literature for right-censored data. The models effectively analyze the survival distribution of those who are susceptible along with the probability of an individual being 'cured'. In the approaches using mixture models, logistic regression is often used to model the cured probability. For the dependency of the survival function on the covariates among the non-cured, various regression models have been considered: the Cox proportional hazards model [KC92, ST00], transformation models [LY04], and richly parametrized models

when the shape of the hazard function is of interest [HBJT03]. Cure rate models have also been developed along the lines of non-mixture models [CIS99, ZYI06]. In addition to right-censored data, cure-rate models have also been developed for interval-censored data [KJ08]. Recently, [CSWL17] studied left truncation under cure-rate transformation models. The data they considered was the 'classic' cure type mentioned above, i.e. all the cured subjects were right-censored. This led to an ease of handling in the likelihood function, where a maximum follow-up time was assumed before the 'cure' actually happened, resulting in a bounded parameter space for theoretical development. On the contrary our definition of SAB leads to a large portion of observed cured subjects, and forces the cumulative hazard function for the non-cured subjects to diverge to infinity as gestation timing approaches 20 weeks.

In the following we consider the mixture cure rate model. This choice has been made based on in-depth discussions with our scientific collaborators, because it is important to understand both the risk factors for SAB (yes/no) as well as the predictors of timing of SAB events among those who experience them. Different timing of SAB can reflect different underlying biological processes. In the next section we write out the likelihood function with many observed 'cured' women in our data. We discuss computational challenges with the likelihood, and adopt an EM algorithm using 'ghost copies' of the observed data. In Section 1.3, the resulting estimator is shown to be consistent and asymptotic normal, despite the fact that the cumulative baseline hazard function diverges at the finite time point before 'cure' is achieved. We illustrate the effectiveness of the method on finite samples via simulation experiments in Section 1.4. We analyze the SAB data from the OTIS database in Section 1.5, and conclude with some additional discussion.

1.2 Model and NPMLE

1.2.1 Model and data

Let $\tau < \infty$ be a strict upper bound of time for the event of interest, beyond which a subject is considered cured. In the pregnancy example above, this would be the 20 weeks of gestation. The whole population consists of two subpopulations: cured and non-cured. We note that the τ here is not defined as the longest possible follow-up time as in many other survival literatures. Let the binary random variable A indicate whether a subject belongs to the non-cured subpopulation; and let $T^* \in (0, \tau)$ be the failure time random variable for this subpopulation. The overall outcome time T is given by the mixture [LY04]: $T = AT^* + (1-A)\tau$. Let $\mathbf{Z_1}$ and $\mathbf{Z_2}$ be two covariate vectors; they may share common covariates depending on the application. We assume that A given $\mathbf{Z_1}$ follows the logistic regression model

$$\mathbb{P}(A = 1|\mathbf{Z_1}) = p(\mathbf{Z_1}) = \frac{e^{\boldsymbol{\alpha}^\top \mathbf{Z_1}}}{1 + e^{\boldsymbol{\alpha}^\top \mathbf{Z_1}}}, \tag{1.1}$$

and that T^* given $\mathbf{Z_2}$ follows the proportional hazard regression model with cumulative baseline hazard function $\lambda_0(t) = \int_0^t \lambda_0(u)du$:

$$\mathbb{P}(T^* \geq t|\mathbf{Z_2}) = S(t|\mathbf{Z_2}) = \exp\{-\Lambda_0(t)e^{\boldsymbol{\beta}^\top \mathbf{Z_2}}\}. \tag{1.2}$$

Note that $\Lambda_0(\tau) = +\infty$ so that $S(\tau|\mathbf{Z_2}) = 0$. The survival function for T is then

$$\mathbb{P}(T \geq t|\mathbf{Z_1}, \mathbf{Z_2}) = \{1 - p(\mathbf{Z_1})\} + p(\mathbf{Z_1})S(t|\mathbf{Z_2}). \tag{1.3}$$

Our data is subject to left truncation and right-censoring. Let Q be the left truncation time and C the right-censoring time, satisfying $0 \leq Q < C$; we also assume that they are independent

of (A, T^*) conditioning on $\mathbf{Z_1}$ and $\mathbf{Z_2}$. For subjecrts $i = 1, ..., n$, the observed data include $\mathbf{Z_1}_i$, $\mathbf{Z_2}_i$, Q_i, $X_i = T_i \wedge C_i$, $\delta_i^1 = A_i \cdot I(T_i \leq C_i)$, $\delta_i^0 = (1 - A_i) I(C_i \geq \tau)$ and $\delta_i^c = I(C_i < T_i \leq \tau)$. In other words δ_i^1 is the indicator that a subject has an observed event (non-cured), δ_i^0 is the indicator that a subject is observed to be cured, and δ_i^c is the indicator that a subject is censored before τ so that we do not know whether she is cured or not. All three indicators are necessary, since the subjects sojourn beyond time τ are observed as cured. This is different from the existing cure-rate model literature where the cured subjects are always marked as censored. Note also that the subject i is observed only if $T_i > Q_i$, hence left truncation is known to lead to a biased sample from the population. Because of right-censoring, A_i may not be observed; but we emphasize here that we do observe many $A_i = 0$ in our data.

Denote $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \boldsymbol{\beta}, \Lambda_0)$. For the purposes of nonparametric maximum likelihood estimation (NPMLE), it is necessary to discretize Λ_0 to be $\Lambda_0(t) = \sum_{k=1}^{K} \lambda_k I(t \geq t_k)$, where $0 < t_1 < \cdots < t_K < \infty$ are the unique failure times [Joh83, Mur94]. NPMLE under truncation and censoring was also discussed in [Tur76] and [LY91] for a distribution function, and under regression settings by, for example, [LMD88] and [GL96], and many other authors. We apply the likelihood approach conditional upon the left truncation time Q_i and the right-censoring time C_i, as no parametric distributional assumptions are made about these two random variables. Denote $p_i = e^{\boldsymbol{\alpha}^\top \mathbf{Z_1}_i} / (1 + e^{\boldsymbol{\alpha}^\top \mathbf{Z_1}_i})$, $\lambda_i(t) = \lambda_0(t) \exp(\boldsymbol{\beta}^\top \mathbf{Z_2}_i)$, $f_i(t) = \lambda_i(t) S_i(t)$, and $S_i(t) = \exp\{-\Lambda_0(t) e^{\boldsymbol{\beta}^\top \mathbf{Z_2}_i}\}$. The likelihood for our observed data is

$$
\begin{aligned}
L(\boldsymbol{\theta}) &= \prod_{i=1}^{n} L_i(\boldsymbol{\theta}; X_i, \delta_i^1, \delta_i^0, \delta_i^c | T_i > Q_i, \mathbf{Z_1}_i, \mathbf{Z_2}_i, Q_i, C_i) \\
&= \prod_{i=1}^{n} \frac{\left\{ p_i \lambda_i(X_i) S_i(X_i) \right\}^{\delta_i^1} (1 - p_i)^{\delta_i^0} \left\{ 1 - p_i + p_i S_i(X_i) \right\}^{\delta_i^c}}{1 - p_i + p_i S_i(Q_i)}, \qquad (1.4)
\end{aligned}
$$

where $1 - p_i + p_i S_i(X_i) = P(T_i > Q_i)$.

1.2.2 NPMLE through EM

Complete data likelihood

The complexity of observed likelihood (1.4) leads to the challenge of optimization. To reduce the problem we follow the approach of [QNLS11], who re-formulated the likelihood function of [Var85].

To augment the observed data, we first note that the group indicator A_i is latent whenever censoring occurs. In addition, we compensate for the left truncation through the "ghost copy" algorithm proposed in [QNLS11]. For each observed subject with the pair of covariates $(\mathbf{Z_1}_i, \mathbf{Z_2}_i)$ and entry time Q_i, there are M_i hypothetical "truncated samples" with latent event time $\widetilde{T}_{ij} < Q_i$, $j = 1, ..., M_i$. The resulting complete likelihood is

$$L^c(\boldsymbol{\theta}) = \prod_{i=1}^{n} \left\{ p_i \lambda_i(X_i) S_i(X_i) \right\}^{\delta_i^1} (1-p_i)^{\delta_i^0} \left\{ p_i S_i(X) \right\}^{A_i \delta_i^c} (1-p_i)^{(1-A_i)\delta_i^c}$$
$$\times p_i^{M_i} \prod_{j=1}^{M_i} \prod_{k: t_k \le Q_i} \left\{ \lambda_k e^{\boldsymbol{\beta}^\top \mathbf{Z_2}_i} S_i(t_k) \right\}^{I(\widetilde{T}_{ij} = t_k)} \tag{1.5}$$

In this way, the two sets of parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are separated in the complete data likelihood. All remaining product terms are those in the usual likelihoods for the logistic and the Cox regression model. Consequently, the M-step update is instantly available from existing solvers.

Given the observed data O, it can be seen that for subject i who is censored at X_i, the unobserved group indicator A_i follows Bernoulli distribution with $P(A_i = 1) = p_i S_i(X_i) / \{1 - p_i + p_i S_i(X_i)\}$. For a subject with truncation time Q_i and covariates $(\mathbf{Z_1}_i, \mathbf{Z_2}_i)$, it can be seen that the number of truncated "ghost" copies M_i follows the geometric distribution with probability

$P(T_i < Q_i) = p_i\{1 - S_i(Q_i)\}$. For the "ghost" event times let $\widetilde{T}_{ij}$ be one of the observed event times $t_k < Q_i$ with probability proportional to $f_i(t_k) = \lambda_k e^{\boldsymbol{\beta}^\top \mathbf{Z_{2}}_i} S_i(t_k)$:

$$\mathbb{P}(\widetilde{T}_{ij} = t_k | M_i, O) = \frac{I(t_k \leq Q_i)\lambda_k e^{\boldsymbol{\beta}^\top \mathbf{Z_{2}}_i} S_i(t_k)}{\sum_{k:t_k \leq Q} \lambda_k e^{\boldsymbol{\beta}^\top \mathbf{Z_{2}}_i} S_i(t_k)}. \tag{1.6}$$

By restricting the "ghost" event times to certain discrete times, we are able to exploit the convenience of directly applying the weighted Cox regression later. The price we pay for the discretization is a slight discrepancy between $\sum_{k:t_k \leq Q} \lambda_k e^{\boldsymbol{\beta}^\top \mathbf{Z_{2}}_i} S_i(t_k)$ and $1 - S_i(Q_i)$. Integrating out the latent variables in $L^c(\boldsymbol{\theta})$ does not give exactly the observed likelihood $L(\boldsymbol{\theta})$. However, we show later that this difference is asymptotically negligible so that the solution from the above EM is asymptotically equivalent to the true NPMLE.

The EM Algorithm

From (1.5) we can write the complete data log-likelihood $l^c = \log L^c$ as

$$\begin{aligned} l^c(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\lambda}) = \sum_{i=1}^{n} \Bigg[& \delta_i^1 A_i \sum_{k=1}^{K} I\{X_i = t_k\} \log f_i(t_k) + M_i \sum_{k:t_k < Q_i} I\{\widetilde{T}_i = t_k\} \log f_i(t_k) \\ & + (1 - \delta_i^1) A_i \log S_i(X_i) + (1 - A_i) \log(1 - p_i) + (A_i + M_i) \log(p_i) \Bigg], \end{aligned} \tag{1.7}$$

where $\boldsymbol{\lambda} = (\lambda_1, ..., \lambda_K)$.

Though the algorithm runs stably from any initial values of the parameters in the support, we recommend to fit a naïve logistic regression without censored subjects for $\boldsymbol{\alpha}^{(0)}$ and a naïve Cox regression for $\boldsymbol{\beta}^{(0)}$ and $\boldsymbol{\lambda}^{(0)}$ treating the observed cured subjects as censored at τ, to minimize the number of iterations until convergence.

E-step

At the $(l+1)$-th iteration ($l = 0, 1, ...$), let $\boldsymbol{\alpha}^{(l)}, \boldsymbol{\beta}^{(l)}, \boldsymbol{\lambda}^{(l)}$ be the parameter values at the current iteration upon which $p_i^{(l)}$, $f_i^{(l)}$ and $S_i^{(l)}$ are defined. The distributions of the latent variables conditioning on the observed data are given in the above, and their conditional expectations can be computed as

$$\mathbb{E}[I\{\widetilde{T}_{ij} = t_k\}|M_i, O; \boldsymbol{\alpha}^{(l)}, \boldsymbol{\beta}^{(l)}, \boldsymbol{\lambda}^{(l)}] = \frac{I(t_k < Q_i) f_i^{(l)}(t_k)}{\sum_{h:t_h<Q_i} f_i^{(l)}(t_h)}, \tag{1.8}$$

$$\mathbb{E}[M_i|O; \boldsymbol{\alpha}^{(l)}, \boldsymbol{\beta}^{(l)}, \boldsymbol{\lambda}^{(l)}] = \frac{p_i^{(l)} \sum_{k:t_k<Q_i} f_i^{(l)}(t_k)}{1 - p_i^{(l)} \sum_{k:t_k<Q_i} f_i^{(l)}(t_k)}, \tag{1.9}$$

$$\mathbb{E}[A_i|O; \boldsymbol{\alpha}^{(l)}, \boldsymbol{\beta}^{(l)}, \boldsymbol{\lambda}^{(l)}] = \delta_i^1 + \delta_i^c \frac{p_i^{(l)} S_i^{(l)}(X_i)}{1 - p_i^{(l)} + p_i^{(l)} S_i^{(l)}(X_i)}. \tag{1.10}$$

Since the latent variables all enter linearly into the complete data log-likelihood, the expected complete data log-likelihood is

$$E(l^c|O) = \sum_{i=1}^{n} \left\{ \sum_{k=1}^{K} w_{i,k}^f \log f_i(t_k) + w_i^S \log S_i(X_i) + w_{0,i}^p \log(1 - p_i) + w_{1,i}^p \log(p_i) \right\}, \tag{1.11}$$

where the weights are computed as

$$\begin{aligned}
w_{i,k}^f &= \delta_i^1 I\{X_i = t_k\} + \frac{p_i^{(l)} f_i^{(l)}(t_k)}{1 - p_i^{(l)} \sum_{h:t_h<Q_i} f_i^{(l)}(t_h)} I\{t_k < Q_i\}, \\
w_i^S &= \delta_i^c \frac{p_i^{(l)} S_i^{(l)}(X_i)}{1 - p_i^{(l)} + p_i^{(l)} S_i^{(l)}(X_i)}, \\
w_{0,i}^p &= \delta_i^0 + \delta_i^c \frac{1 - p_i^{(l)}}{1 - p_i^{(l)} + p_i^{(l)} S_i^{(l)}(X_i)}, \\
w_{1,i}^p &= \delta_i^1 A_i + \delta_i^c \frac{p_i^{(l)} S_i^{(l)}(X_i)}{1 - p_i^{(l)} + p_i^{(l)} S_i^{(l)}(X_i)} + \frac{p_i^{(l)} \sum_{k:t_k<Q_i} f_i^{(l)}(t_k)}{1 - p_i^{(l)} \sum_{k:t_k<Q_i} f_i^{(l)}(t_k)}.
\end{aligned}$$

M-step

From (1.11) the expected log-likelihood can be written as the sum of two parts, so that

the M-step can be achieved using a weighted logistic regression optimized over $\boldsymbol{\alpha}$:

$$l_{glm}(\boldsymbol{\alpha}) = \sum_{i=1}^{n} w_{0,i}^{p} \log(1-p_i) + w_{1,i}^{p} \log(p_i); \tag{1.12}$$

and a weighted Cox proportional hazard regression optimized over $\boldsymbol{\beta}$ and $\boldsymbol{\lambda}$:

$$l_{cox}(\boldsymbol{\beta}, \boldsymbol{\lambda}) = \sum_{i=1}^{n} \sum_{k=1}^{K} w_{i,k}^{f} \log f_i(t_k) + \sum_{i=1}^{n} w_i^{S} \log S_i(X_i). \tag{1.13}$$

Easily implemented solution is available from existing *glm* and *coxph* solvers in R, to obtain $\boldsymbol{\alpha}^{(l+1)}$, $\boldsymbol{\beta}^{(l+1)}$ and $\boldsymbol{\lambda}^{(l+1)}$, where $\boldsymbol{\lambda}^{(l+1)}$ is the Breslow-type baseline hazard estimator from the fitted *coxph* object with weights.

The application of the EM algorithm to NPMLE's under semiparametric models was discussed in [VX00], including convergence properties and variance estimation following the EM. In particular, once the baseline hazard function has been discretized to finite many mass points given the observed data, the convergence properties of the EM algorithm known for parametric models carry over. In this case, the convergence of the EM algorithm is guaranteed by the log-concaveness of the complete data likelihood $L^c(\boldsymbol{\theta})$, which in turn is guaranteed by the log-concaveness of the logistic likelihood and the Cox proportional hazards likelihood. We show in Section 1.3.1 that under mild conditions the EM solution converges to the NPMLE.

Variance Estimator

At convergence of the EM algorithm where $\widehat{\boldsymbol{\theta}}$ denotes the NPMLE, the [Lou82] formula can be used to give the observed Fisher information:

$$I_{obs}(\widehat{\boldsymbol{\theta}}) = \sum_{i=1}^{n} \mathbb{E}_{\widehat{\boldsymbol{\theta}}}[B_i|O] - \sum_{i=1}^{n} \mathbb{E}_{\widehat{\boldsymbol{\theta}}}[\mathbf{S}_i \mathbf{S}_i^{\top}|O] - 2\sum_{i<i'}^{n} \mathbb{E}_{\widehat{\boldsymbol{\theta}}}[\mathbf{S}_i|O]\mathbb{E}_{\widehat{\boldsymbol{\theta}}}[\mathbf{S}_i|O]^{\top}, \tag{1.14}$$

where $\mathbf{S}_i$ and B_i are the gradient ∇l_i^c and the negatives of Hessian $-\nabla^2 l_i^c$ of the complete data log-likelihood. The above is in closed form, and the details are given in Section 1.8. We show in the next section that (1.14) provides a consistent variance estimator for the NPMLE, and its use in association with the NPMLE has been advocated in the literature, in particular for its numerical stability [VX00, ZL07, GMX09].

1.3 Theory

Let $\boldsymbol{\theta}_0 = (\boldsymbol{\alpha}_0, \boldsymbol{\beta}_0, \Lambda_0(\cdot))$ denote the true parameter value. Following [ABGK93], we define the counting process $N_i(t) = \delta_i^1 I(X_i \leq t)$ and the at-risk process $Y_i(t) = I(Q_i \leq t \leq X_i)$. Their sums are denoted as $\bar{N}(t) = \sum_{i=1}^n N_i(t)$, and $\bar{Y}(t) = \sum_{i=1}^n Y_i(t)$. By Doob-Meyer decomposition, a martingale with respect to the filtration $\mathcal{F}_t = \sigma\{N_i(u), Y_i(u), \mathbf{Z_1}, \mathbf{Z_2}, u \leq t\}$ is

$$M_i(t) = N_i(t) - \int_0^t \phi_i^{\boldsymbol{\theta}_0}(u) Y_i(u) e^{\boldsymbol{\beta}_0^\top \mathbf{Z_{2}}_i} d\Lambda_0(u), \tag{1.15}$$

where

$$\phi_i^{\boldsymbol{\theta}}(t) = \frac{\exp\{\boldsymbol{\alpha}^\top \mathbf{Z_1}_i - \Lambda(t) e^{\boldsymbol{\beta}^\top \mathbf{Z_2}_i}\}}{1 + \exp\{\boldsymbol{\alpha}^\top \mathbf{Z_1}_i - \Lambda(t) e^{\boldsymbol{\beta}^\top \mathbf{Z_2}_i}\}} = \mathbb{P}_{\boldsymbol{\theta}}(A_i = 1 | X_i \geq t). \tag{1.16}$$

To make use of the martingale framework, we write the observed log-likelihood $l_n = \log L$, where $L(\boldsymbol{\theta})$ was given in (1.4), as

$$\begin{aligned} l_n = &\sum_{i=1}^n \int_0^\tau \log\Big(\phi_i^{\boldsymbol{\theta}}(u) e^{\boldsymbol{\beta}^\top \mathbf{Z_2}_i}\Big) dN_i(u) - \int_0^\tau Y_i(u) \phi_i^{\boldsymbol{\theta}}(u) e^{\boldsymbol{\beta}^\top \mathbf{Z_2}_i} d\Lambda(u) \\ &+ \int_0^\tau \log\Big(\Delta\Lambda(u)\Big) dN_i(u), \end{aligned}$$

where $\Delta\Lambda(u)$ is the size of jump of the baseline cumulative hazard at u [Mur94]. We establish the theory under the following assumptions. The vector norm throughout this paper is the uniform

norm, i.e. the largest absolute value among all elements.

Assumption 1. *The true finite-dimensional parameter* $(\boldsymbol{\alpha}_0, \boldsymbol{\beta}_0)$ *is an element of the interior of a compact set* $\{(\boldsymbol{\alpha}, \boldsymbol{\beta}) : \|\boldsymbol{\alpha}\| \vee \|\boldsymbol{\beta}\| \leq D_1\}$ *for some constant* D_1.

Assumption 2. *The covariates* $(\mathbf{Z_1}, \mathbf{Z_2})$ *follow distribution* $F_Z(\cdot,\cdot)$. *They are bounded a.s.: there exists* $D_2 > 0$, *such that* $\mathbb{P}(\max\{\|\mathbf{Z_1}\|, \|\mathbf{Z_2}\|\} \leq D_2) = 1$. *Also, their covariance matrices* $\mathrm{Var}(\mathbf{Z_1})$*(without intercept term) and* $\mathrm{Var}(\mathbf{Z_2})$ *are both positive-definite. Denote constant m such that*

$$0 < m^{-1} = e^{-D_1 D_2} \leq e^{\boldsymbol{\alpha}^\top \mathbf{Z_1}} \wedge e^{\boldsymbol{\beta}^\top \mathbf{Z_2}} \leq e^{\boldsymbol{\alpha}^\top \mathbf{Z_1}} \vee e^{\boldsymbol{\beta}^\top \mathbf{Z_2}} \leq e^{D_1 D_2} = m < \infty \quad a.s.. \tag{1.17}$$

Assumption 3. *The baseline cumulative hazard function* $\Lambda_0(t)$ *is a non-decreasing continuous function on* $[0, \tau)$. $\Lambda_0(0) = 0$ *and* $\Lambda_0(\tau-) = \infty$. *And*

$$\inf_{t \in [0,\tau]} \mathbb{E}[Y(t)|\mathbf{Z_1}, \mathbf{Z_2}] > \varepsilon > 0, \quad a.s.. \tag{1.18}$$

Assumption 4. *There exists* $\zeta \in (0, \tau)$ *such that* $\mathbb{P}(Q > \zeta) = 0$. $\Lambda_0(t)$ *is strictly increasing over* $[0, \zeta]$, *and* $\mathbb{E}[Y(t)|\mathbf{Z_1}, \mathbf{Z_2}]$ *is Lipschitz continuous w.r.t to* $\Lambda_0(t)$ *on* $[0, \zeta]$ *a.s.; that is, there is a constant*

$$\mathcal{L} \geq \sup_{0 \leq t < s \leq \zeta} \left\{ \frac{|\mathbb{E}[Y(t)|\mathbf{Z_1}, \mathbf{Z_2}] - \mathbb{E}[Y(s)|\mathbf{Z_1}, \mathbf{Z_2}]|}{|\Lambda_0(t) - \Lambda_0(s)|} \right\}, \quad a.s.. \tag{1.19}$$

Assumption 3 above is specifically made for cure rate models with an observed cured portion. This assumption enforces that the failure time must occur prior to a well-defined upper bound. Equation (1.18) requires that not all subjects (including cured) are censored prior to a maximal possible event time. Hence, all existing theoretical results for which the baseline cumulative hazard is assumed finite at a maximum follow-up time do not apply to our case.

Equation (1.18) also requires that certain proportion of subjects enter the study at time zero. While this may not always be the case for our pregnancy studies, time zero may be replaced by the earliest entry time into the study and the inference is conditional upon survival beyond that time, and all the results established in this section carry over. Assumption 4 gives the regularity conditions on truncation and censoring. The truncation times should be bounded away from time τ; this is required in order to establish Lemma 1 below. The truncation-censoring distribution also has to possess certain level of continuity with respect to the distribution of event time. For example, the continuity condition is satisfied when the distributions for Q, C and T given $\mathbf{Z_1}$ and $\mathbf{Z_2}$ all have densities that are bounded away from ∞ and 0 almost surely. This condition can be weakened to allow $\Lambda_0(t)$ to be constant over some open set and require only that $\mathbb{E}[Y(t)|\mathbf{Z_1},\mathbf{Z_2}]$ is Lipschitz continuous with respect to $\Lambda_0(t)$ on a open set $\Omega \subset [0,\zeta]$ consisting of finite many open intervals, on which $\int_\Omega d\Lambda_0 = \Lambda_0(\zeta)$. All theoretical results under this weakened condition can be achieved by repeatedly applying the steps in the current proof.

For the asymptotic normality we make the following assumption where τ' is defined later.

Assumption 3'. *The baseline cumulative hazard $\Lambda_0(t)$ is a non-decreasing continuous function on $[0,\tau']$. $\Lambda_0(0) = 0$, $\Lambda_0(\tau') < \infty$ and $\Lambda_0(\tau-) = \infty$. And*

$$\inf_{t\in[0,\tau']} \mathbb{E}[Y(t)|\mathbf{Z_1},\mathbf{Z_2}] > \varepsilon > 0, \quad a.s..$$

1.3.1 Existence of NPMLE and convergence of EM

First, we show the existence of the NPMLE.

Theorem 1. *Under Assumptions 1 and 2, if* $\sum_{i=1}^{n} N_i(\tau) > 0$*, then a maximizer of* $l_n(\boldsymbol{\theta})$*,* $\widehat{\boldsymbol{\theta}} = (\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}, \widehat{\Lambda}(\cdot))$ *exists and is finite.*

For the proof we use the same technique as in [Mur94]. All the proofs are in Section 1.7.

We now show that the previously described EM algorithm converges and its solution is asymptotically equivalent to the NPMLE.

Lemma 1. *Let* $\widehat{\boldsymbol{\theta}}$ *be the NPMLE for the observed likelihood* (1.4)*. Under Assumptions 1 - 4,*

1. *the EM algorithm with complete data likelihood* (1.5) *converges almost surely to* $\widetilde{\boldsymbol{\theta}}$*;*
2. $n^{-1}\{l_n(\widehat{\boldsymbol{\theta}}) - l_n(\widetilde{\boldsymbol{\theta}})\} = O_p(1/n)$.

Theorem 2. *Under Assumptions 1- 4,* $\|\widehat{\boldsymbol{\theta}} - \widetilde{\boldsymbol{\theta}}\| = o_p(1)$.

Theorem 2'. *Under Assumptions 1, 2, 4 and 3',* $\widehat{\boldsymbol{\theta}} - \widetilde{\boldsymbol{\theta}} = o_p(1/\sqrt{n})$.

1.3.2 Consistency of NPMLE

Next, we show the consistency of the NPMLE.

Theorem 3. *Under Assumptions 1 - 4, the NPMLE estimator for L in* (1.4)*,* $\widehat{\boldsymbol{\theta}} = (\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}, \widehat{\Lambda}(\cdot))$*, is consistent. That is*

$$\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0 \to 0, \quad \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 \to 0, \quad \sup_{t \in [0,\tau]} |e^{-\widehat{\Lambda}(t)} - e^{-\Lambda_0(t)}| \to 0 \quad a.s..$$

The proof follows the general framework in [Mur94]. The estimator for the baseline hazard satisfies the equation

$$\widehat{\Lambda}(t) = \int_0^t \left\{ \sum_{i=1}^{n} W_i^{\widehat{\boldsymbol{\theta}}}(u) e^{\widehat{\boldsymbol{\beta}}^\top \mathbf{Z}_{2i}} \right\}^{-1} d\bar{N}(u), \tag{1.20}$$

where

$$W_i^{\boldsymbol{\theta}}(t) = \left\{\delta_i^1 + \delta_i^c \phi_i^{\boldsymbol{\theta}}(X_i)\right\} I\{t \le X_i\} - \phi_i^{\boldsymbol{\theta}}(Q_i) I\{t \le Q_i\}, \tag{1.21}$$

and $\phi_i^{\boldsymbol{\theta}}(\cdot)$ is given in (1.16). A bridge between $\widehat{\Lambda}$ and Λ_0 is constructed as

$$\bar{\Lambda}(t) = \int_0^t \left\{ \sum_{i=1}^n \phi_i^{\boldsymbol{\theta}_0}(u) Y_i(u) e^{\boldsymbol{\beta}_0^\top \mathbf{Z_2}_i} \right\}^{-1} d\bar{N}(u). \tag{1.22}$$

The details of the proof deserve some extra comments here, as it achieves the a.s. convergence with a baseline hazard unbounded in its support using a few innovative steps. First, we apply Helly's selection theorem to the Càdlàg function sequence $e^{-\widehat{\Lambda}}$. Then, the upper bound for $\widehat{\Lambda}$ in any interval $[0, \tau^*] \subset (0, \tau)$ is established via the lower bound for $n^{-1}\sum_{i=1}^n W_i^{\widehat{\boldsymbol{\theta}}}(u) e^{\widehat{\boldsymbol{\beta}}^\top \mathbf{Z_2}_i}$. We manage to show that the ratio $\gamma(t) = d\widehat{\Lambda}(t)/d\bar{\Lambda}(t)$ is bounded between zero and infinity for all $t \in (0, \tau)$ despite the indefinite quotient at 0 and τ. Finally, we conclude the proof by showing that $\gamma(t) = 1$ using an identifiability argument.

For the purposes of the asymptotic normality below, we have a similar result:

Theorem 3'. *Under Assumptions 1, 2, 3' and 4, the NPMLE estimator for L^I defined later in* (1.24)*, $\widehat{\boldsymbol{\theta}} = (\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}, \widehat{\Lambda}(\cdot))$, is consistent. That is*

$$\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0 \to 0, \quad \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 \to 0, \quad \sup_{t \in [0, \tau']} |\widehat{\Lambda}(t) - \Lambda_0(t)| \to 0 \quad a.s.. \tag{1.23}$$

1.3.3 Asymptotic Normality of NPMLE

The divergence of the cumulative baseline hazard Λ_0 at τ eventually becomes an obstacle in the study of weak convergence. It is involved in all the second order terms including both the parametric parts and the nonparametric part. Existing techniques, mostly relying on a finite upper

bound of Λ_0, cannot deal with it. To proceed with the theoretical endeavor, we avoid the divergent tail by slightly modifying the likelihood. That is, we make an interval censoring window (τ', τ) close to the end of study, so that the failure indicator A is always observed for those at-risk at time τ', but their failure times are unknown if $A = 1$. We note that this is for technical reason only, so that the baseline cumulative hazard is always bounded at the observed failure times as $n \to \infty$. In practical applications this modification of the likelihood is unnecessary since the observed SAB events are recorded in dates, so that there is always at least one day gap between when a (possibly censored) SAB event can happen and when a woman is considered cured.

Let $\delta^\tau = A \cdot I(X > \tau')$ be the interval-censoring indicator in (τ', τ). Notice that $S(t) - S(\tau) = S(t)$ for any $t < \tau$. We have the resulting interval-censored data likelihood that is modified from (1.4):

$$L^I(\boldsymbol{\theta}) = \prod_{i=1}^{n} \frac{\left\{p\lambda_i(X_i)e^{\boldsymbol{\beta}^\top \mathbf{Z_2}_i} S_i(X_i)\right\}^{\delta_i^1} (1-p_i)^{\delta_i^0} \left\{1 - p_i + p_i S_i(X_i)\right\}^{\delta_i^c} \left\{p_i S_i(\tau')\right\}^{\delta_i^\tau}}{1 - p_i + p_i S_i(Q_i)}. \tag{1.24}$$

The corresponding log-likelihood $l_n^I = \log L^I$ is

$$\begin{aligned} l_n^I = \sum_{i=1}^{n} \int_0^{\tau'} \log\left(\phi_i^{\boldsymbol{\theta}}(u) e^{\boldsymbol{\beta}^\top \mathbf{Z_2}_i} \triangle\Lambda(u)\right) dN_i(u) - \int_0^{\tau'} Y_i(u)\phi_i^{\boldsymbol{\theta}}(u) e^{\boldsymbol{\beta}^\top \mathbf{Z_2}_i} d\Lambda(u) \\ + \{N_i(\tau) - N_i(\tau')\} \log \phi_i^{\boldsymbol{\theta}}(\tau') + Y_i(\tau) \log(1 - \phi_i^{\boldsymbol{\theta}}(\tau')). \end{aligned} \tag{1.25}$$

The proof then follows the framework in [Mur95] to verify the conditions of Theorem 3.3.1 from [VdVW96]. We shall describe the functional space in which weak convergence is established. Let H_∞ be the space containing elements in the form of $\mathbf{h} = (\mathbf{a}, \mathbf{b}, \eta)$, where the vectors $\mathbf{a}$ and $\mathbf{b}$ are of the same dimensions as $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, respectively, and the function $\eta(\cdot)$ is defined on $[0, \tau']$

with $\eta(0) = 0$ and is of bounded variation, i.e. the total variation of η over $[0, \tau']$,

$$V_0^{\tau'} \eta = \sup_{\substack{0=u_0<\cdots<u_s=\tau' \\ s=1,2,\ldots}} \sum_{j=1}^{s} |\eta(u_j) - \eta(u_{j-1})|$$

is finite. Define a norm $\|\cdot\|_H$ on H_∞:

$$\|(\mathbf{a}, \mathbf{b}, \eta)\|_H = \|\mathbf{a}\|_1 + \|\mathbf{b}\|_1 + V_0^{\tau'} |\eta|,$$

and spaces indexed by a positive real number p

$$H_p = \{\mathbf{h} : \|\mathbf{h}\|_H < p\}.$$

For each p, define $l^\infty(H_p)$ as the functional space of all uniformly bounded linear map $H_p \mapsto \mathbb{R}$, i.e.

$$\forall \Psi \in l^\infty(H_p), \ \sup_{\mathbf{h} \in H_p} |\Psi(\mathbf{h})| < \infty.$$

The parameter $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \boldsymbol{\beta}, \Lambda)$ as a function in $l^\infty(H_p)$ is defined as

$$\boldsymbol{\theta}(\mathbf{h}) = \mathbf{a}^\top \boldsymbol{\alpha} + \mathbf{b}^\top \boldsymbol{\beta} + \int_0^{\tau'} \eta(u) d\Lambda(u).$$

The induced functional norm is equivalent to the norm in (1.23) where consistency (Theorem 3') is established; we denote $\|\boldsymbol{\theta}\|$.

Theorem 4. *Let $\widehat{\boldsymbol{\theta}} = (\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}, \widehat{\Lambda}_0(\cdot))$ be the NPMLE for the log-likelihood l_n^I in* (1.25)*. Under Assumptions 1, 2, 3' and 4,*

$$\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \longrightarrow \mathcal{G}, \ \textit{in } l^\infty(H_p)$$

weakly for a tight Gaussian process $\mathcal{G}$ on $l^\infty(H_p)$ with covariance process

$$\mathrm{Cov}(\mathcal{G}(\mathbf{h}), \mathcal{G}(\mathbf{h}^*)) = \mathbf{a}^\top \boldsymbol{\sigma}_a^{-1}(\mathbf{h}^*) + \mathbf{b}^\top \boldsymbol{\sigma}_b^{-1}(\mathbf{h}^*) + \int_0^{\tau'} \eta(u) \sigma_\eta^{-1}(\mathbf{h}^*)(u) d\Lambda_0(u),$$

where $\mathbf{h} = (\mathbf{a}, \mathbf{b}, \eta)$, and $\sigma(\mathbf{h}) = \Big(\boldsymbol{\sigma}_a(\mathbf{h}), \boldsymbol{\sigma}_b(\mathbf{h}), \sigma_\eta(\mathbf{h})\Big)$ is given in the (1.48)*.*

Let $\widehat{\sigma}$ be a nature estimator for the operator σ by substituting the true parameter $\boldsymbol{\theta}_0$ and expectation with the estimator $\widehat{\boldsymbol{\theta}}$ and the sample average.

Theorem 5. *Under Assumptions 1, 2, 3' and 4,* $\widehat{\sigma}$ *is asymptotically equivalent to the information matrix in* (1.14)*. The solution to* $\mathbf{g} = \widehat{\sigma}^{-1}(\mathbf{h})$ *exists with probability going to* 1 *as n increases and*

$$\mathbf{a}^\top \widehat{\boldsymbol{\sigma}}_a^{-1}(\mathbf{h}^*) + \mathbf{b}^\top \widehat{\boldsymbol{\sigma}}_b^{-1}(\mathbf{h}^*) + \int_0^{\tau'} \eta(u)\widehat{\sigma}_\eta^{-1}(\mathbf{h}^*)(u)d\widehat{\Lambda}(u) \xrightarrow{\mathbb{P}} \mathrm{Cov}(\mathcal{G}(\mathbf{h}), \mathcal{G}(\mathbf{h}^*)).$$

1.4 Simulation study

1.4.1 Simulation setup

Here we detail our data simulation procedure for all of the simulation studies. Simulating cure-rate model data presents its own challenges. To be comparable with the spontaneous abortion data which we examine in the next section, we consider finite time τ, which is set to be 20 (weeks). The covariates are the same for the logistic and the Cox part of the regression models and, unless otherwise specified, consisting of $Z_1 \sim N(4,1)$, with corresponding parameters (α_1, β_1), and $Z_2 \sim$ Bernouilli $(p = 0.3)$, with corresponding parameters (α_2, β_2). The logistic regression part also includes an intercept α_0.

We begin by generating a larger sample than we desire to account for those who will be left out due to truncation. Values for $\boldsymbol{\alpha}$ are chosen to procure the desired percentage of cured individuals on average in the population, and we refer to this as the % of cured individuals in a simulation study. An individual is designated as either cured or not with the probability determined from the logistic model.

The baseline survival function for the Cox model is set as $S_0(t) = 20 - t$, the survival function of a Uniform $(0, 20)$ random variable. The baseline cumulative hazard is thus $\Lambda_0(t) = 20\{1 - e^{-t}\}$. For those not cured individuals we generate an event time $T = 20\{1 - U^{\exp(-\beta_1 Z_1 - \beta_2 Z_2)}\}$, where $U \sim$ Uniform $(0, 1)$.

Truncation times are generated from Uniform $(0, a)$ for some $a < 15$ chosen so that on average the desired percentage of uncured individuals are truncated out. We refer to this percentage as the % of truncation. Once the truncation times are generated, all individuals with event times less than their truncation times are removed, and we reduce the data set to the desired sample size by taking the first n individuals from those who remain. Finally, when there is censoring the censoring times are generated from Uniform $(15, b)$ for some $b > 20$ so that on average the desired percentage of the n individuals (including those who are cured) will have a censoring time less than $\min(T_i, 20)$. We refer to this percentage as the % of censoring. We ran all simulations with 500 trials below.

1.4.2 Simulation results

In Tables 1.1 we examine the performance of the NPMLE. We consider a smaller sample size n = 200 and a larger sample size n = 1000, and like in the pregnancy studies for SAB we assume that a majority 75% of the subjects are cured. We ran simulations over the combination of two truncation scenarios (10%, 20%) and two censoring scenarios (0%, 20%). In the tables we provide the average parameter estimates ("Estimate"), the sample standard deviation of these estimates over the 500 simulation trials ("Sample SD"), the mean over the 500 trials of the standard errors based on our variance estimation ("SE"), and the empirical coverage probabilities

Table 1.1: Simulation results using the EM algorithm for NPMLE.

	True	$n = 200$				$n = 1000$			
	Value	Estimate	Sample SD	SE	Coverage	Estimate	Sample SD	SE	Coverage
10% Truncation, 0% Censoring									
α_0	1.00	1.01	0.79	0.75	94.0 %	0.98	0.35	0.33	93.6 %
α_1	-0.63	-0.64	0.20	0.19	94.8 %	-0.63	0.09	0.08	94.3 %
α_2	1.00	1.00	0.36	0.37	95.6 %	1.01	0.16	0.16	94.8 %
β_1	-0.20	-0.23	0.20	0.17	92.2 %	-0.20	0.07	0.07	95.6 %
β_2	0.30	0.33	0.34	0.32	94.2 %	0.29	0.14	0.13	93.4 %
20% Truncation, 0% Censoring									
α_0	1.00	0.97	0.80	0.79	94.8 %	0.99	0.34	0.35	96.0 %
α_1	-0.63	-0.64	0.21	0.20	95.4 %	-0.63	0.09	0.09	96.2 %
α_2	1.00	0.98	0.40	0.39	95.4 %	0.99	0.17	0.17	95.2 %
β_1	-0.20	-0.20	0.20	0.18	94.6 %	-0.20	0.07	0.07	95.2 %
β_2	0.30	0.31	0.37	0.34	94.6 %	0.30	0.14	0.14	94.2 %
10% Truncation, 20% Censoring									
α_0	1.00	1.18	0.97	0.99	96.6 %	1.02	0.41	0.42	95.8 %
α_1	-0.63	-0.69	0.25	0.26	96.2 %	-0.64	0.11	0.11	96.2 %
α_2	1.00	1.01	0.50	0.49	95.4 %	1.00	0.20	0.21	96.0 %
β_1	-0.20	-0.21	0.30	0.26	91.6 %	-0.21	0.11	0.11	94.4 %
β_2	0.30	0.31	0.53	0.49	93.4 %	0.30	0.21	0.20	93.8 %
20% Truncation, 20% Censoring									
α_0	1.00	1.05	1.00	0.96	95.8 %	0.98	0.37	0.41	97.0 %
α_1	-0.63	-0.66	0.27	0.25	96.6 %	-0.63	0.10	0.11	96.6 %
α_2	1.00	1.05	0.49	0.47	95.6 %	1.01	0.20	0.20	94.6 %
β_1	-0.20	-0.19	0.30	0.26	90.4 %	-0.20	0.11	0.11	95.0 %
β_2	0.30	0.33	0.54	0.48	92.2 %	0.31	0.21	0.20	94.8 %

("Coverage") of the nominal 95% confidence intervals using the SE's.

According to the table, the performance of NPMLE is quite good. The average estimates of the parameters are generally close to their true values in all scenarios. This includes for the Cox part of the model under the smaller sample size n = 200, where only about 25% of the sample have events when there is no censoring, and even few in the presence of censoring. The variance estimator generally improves with larger sample size, especially for the Cox part of the model and with 20% censoring, which also reflects in the coverage probabilities of the nominal 95% confidence intervals. Note that with 500 simulation trials these empirical coverage probabilities have about $\pm 2\%$ margin of error.

At the suggestion of a reviewer, we also compared our EM algorithm with the numerical optimization algorithm of [CSWL17], and the results are summarized in Tables 1.2 and 1.3. We see that the performance of the two numerical algorithms were generally comparable. [CSWL17] reported occurrence of divergence (up to 0.3% of the times) in their simulation experiments, while we did not experience any such issues with the EM algorithm. This is consistent with our past experience with the EM algorithm under other semiparametric models [GMX09].

1.5 Analysis of spontaneous abortion data

The data we investigate come from the OTIS autoimmune disease in pregnancy database as mentioned earlier. Our sample includes pregnant women who entered a research study between 2005 and 2012. It consists of n = 911 women who entered the study before week 20 of their gestation, with complete covariate information. Among them 473 (52%) were pregnant women

Table 1.2: Comparison with Chen *et al.* (2017) simulation results, 20% censoring. In every two lines of the results, the first line is our approach, the second line is from Chen *et al.* (2017). True $\beta = -0.693$.

		$n = 200$			$n = 400$		
		β	γ_0	γ_1	β	γ_0	γ_1
$(\gamma_0, \gamma_1) =$	EST	-0.657	0.987	-0.535	-0.681	1.007	-0.534
		-0.657	1.015	-0.544	-0.657	1.004	-0.533
$(1, -0.5)$	SD	0.207	0.250	0.331	0.149	0.168	0.222
		0.254	0.242	0.340	0.170	0.168	0.245
	ASE	0.205	0.242	0.323	0.144	0.171	0.228
		0.235	0.247	0.343	0.167	0.173	0.242
	CP	0.952	0.937	0.937	0.935	0.953	0.952
		0.932	0.959	0.952	0.944	0.957	0.956
$(\gamma_0, \gamma_1) =$	EST	-0.661	0.975	-1.021	-0.672	0.992	-1.023
		-0.653	1.024	-1.032	-0.659	1.008	-1.015
$(1, -1)$	SD	0.233	0.257	0.331	0.159	0.175	0.222
		0.254	0.241	0.336	0.173	0.169	0.235
	ASE	0.220	0.242	0.322	0.154	0.171	0.226
		0.239	0.250	0.330	0.170	0.175	0.232
	CP	0.933	0.937	0.945	0.930	0.942	0.950
		0.931	0.956	0.943	0.928	0.963	0.949

Table 1.3: Comparison with Chen *et al.* (2017) simulation results, 40% censoring. in every two lines of the results, the first line is our approach, the second line is from Chen *et al.* (2017). True $\beta = -0.693$.

		$n = 200$			$n = 400$		
		β	γ_0	γ_1	β	γ_0	γ_1
$(\gamma_0, \gamma_1) =$	EST	-0.673	0.968	-0.509	-0.687	0.988	-0.521
		-0.719	0.999	-0.508	-0.707	1.003	-0.509
$(1, -0.5)$	SD	0.203	0.257	0.344	0.144	0.180	0.236
		0.210	0.263	0.336	0.143	0.182	0.239
	ASE	0.206	0.253	0.335	0.145	0.180	0.237
		0.204	0.266	0.344	0.142	0.187	0.242
	CP	0.956	0.946	0.946	0.950	0.944	0.952
		0.955	0.966	0.960	0.957	0.952	0.963
$(\gamma_0, \gamma_1) =$	EST	-0.664	0.967	-1.032	-0.672	0.989	-1.017
		-0.691	1.016	-1.018	-0.690	1.004	-1.001
$(1, -1)$	SD	0.229	0.254	0.342	0.161	0.179	0.228
		0.239	0.286	0.376	0.167	0.202	0.256
	ASE	0.223	0.256	0.338	0.155	0.182	0.239
		0.236	0.295	0.379	0.164	0.207	0.266
	CP	0.946	0.957	0.960	0.938	0.958	0.956
		0.955	0.958	0.955	0.939	0.956	0.952

with certain autoimmune diseases who were treated with medications under investigation, 261 (29%) were women with the same specific autoimmune diseases who were not treated with the medications under investigation, and the rest 177 (19%) were healthy pregnant women without autoimmune diseases who were not treated with the medications. [CBB$^+$01] discussed the importance of having a diseased control group, since some of the adverse outcomes in pregnancy may be due to the diseases instead of the medications. There were a total of 66 SAB events, and 2 women were lost to follow up before 20 weeks of gestation.

Since the data was collect through phone interviews roughly each trimester, despite the fact that we obtain medical records of all study subjects, there were 10 SAB events in the data set without exact event times, instead a window was available during which each event had occurred. These events were therefore interval censored. For the purposes of this analysis we applied the multiple imputation (MI) method of [Pan00]; a separate research project was carried out to develop specific methodology and theory to handle interval censoring in this setting (*https://arxiv.org/abs/1708.06838*). More specifically, we imputed the actual SAB event times from the uniform distribution over the interval censoring windows, and the imputation was repeated 10 times. The analysis results from each of the 10 imputed data sets were combined using standard MI methodology to obtain the final parameter estimates and their standard errors [Pan00, RL02].

There are a number of risk factors for spontaneous abortion that have been identified in the literature [CJX$^+$13, for example]. Alternatively, we can use a data driven selection method for risk factors in our cure rate model. For each baseline covariate, we use the Wald test with two degrees of freedom for both coefficients in the logistic and the Cox part of the model. We

first screen the covariates with a univariate cure rate model, with a p-value cutoff of 0.2 from the Wald test; this step also considers different possible codings of a variable, for example, some of the four categories of BMI (under-weight, normal, over-weight, obese) might be combined, if the p-value is lowered. We then run a backward selection, with a p-value cutoff of 0.1 from the Wald test. The selected variables are maternal age ≥ 34 years or not, body mass index (BMI) ≥ 25 or not, whether there was smoking (Y/N) or alchohol (Y/N) intake during early pregnancy. We fit our final cure model to the data with these covariates and exposure status, and the results using the NPMLE are given in Table 1.4 left columns.

From Table 1.4, we see that older maternal age significantly increases the probability of SAB in the logistic part of the model. The probability of SAB of either healthy control group or diseased control group is not significantly different from the medication exposed women. The Cox regression part of the model identified smoking and alcohol as significant factors for the hazard of SAB. In the cure model context since the Cox model is only used for those who eventually have events (observed or censored), this part of the model should be understood as impact of the covariates on the timing of SAB; that is, significantly later timing of SAB for those who smoked, and significantly earlier timing for those who had alcohol. The findings about BMI appears counter intuitive here. In discussion with our medical colleagues, it is possible that obese or overweight women have higher risk for early SAB, which were not captured in our data due to left truncation; it then might occur that for what we observe, they appear to be at lower risk for SAB. In addition in this data there was slightly more drinking in the BMI < 25 group, and the lower the BMI, the higher the blood alcohol level for a given alcohol dose, leading to possibly higher risk for SAB.

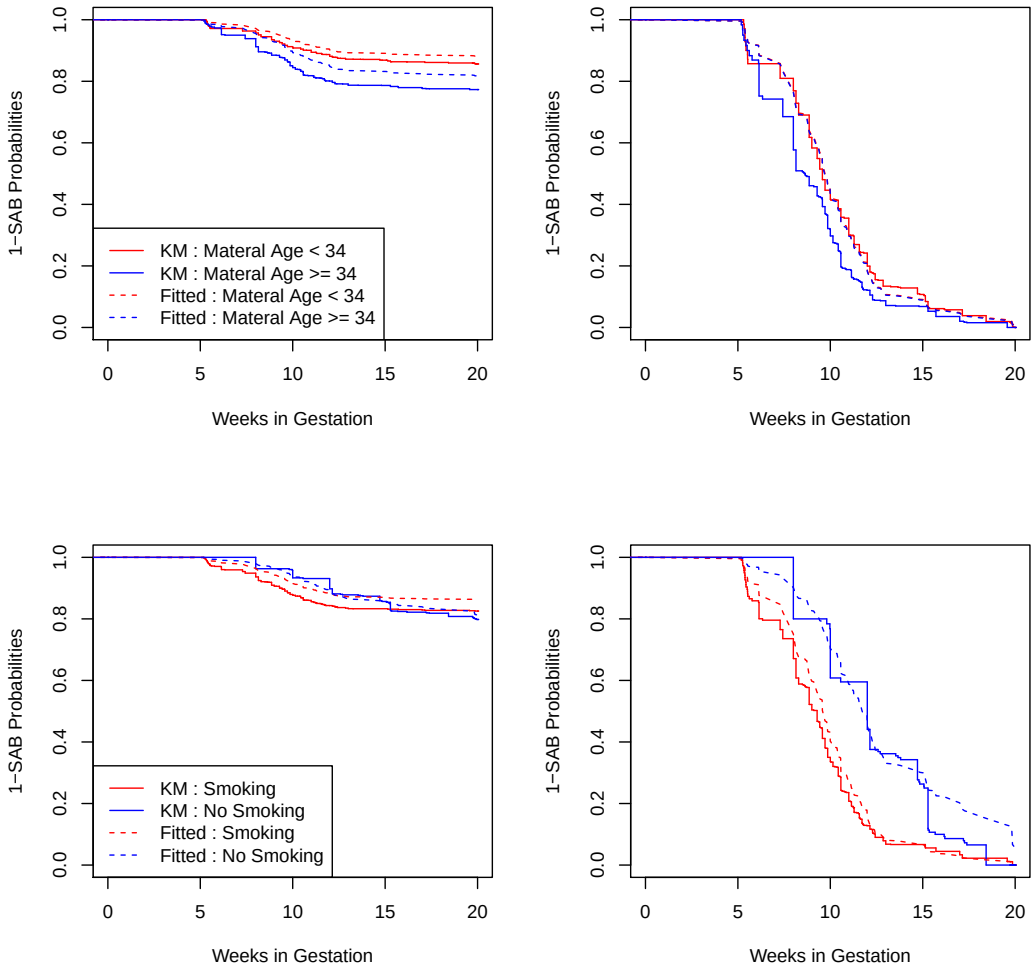


Figure 1.2: Left truncated Kaplan-Meier and fitted curves for SAB events according to maternal age (top) or smoking (bottom), among the full data set (left) and without the observed cured individuals (right). The fitted curves are averages of individual fitted curves in the group.

Figure 1.2 illustrates for maternal age and smoking the stratified Kaplan-Meier and the fitted curves under the cure model (the curves for BMI and alcohol were similar and not shown here). The fitted curves are averages of individual fitted curves in each group, such as among those with maternal age < 34, etc. It is seen that all curves drop to zero among the non-cured subjects (right panel of the figure), and that the effects in the Cox model part translate to timing of the SAB events. In the left panel of the figure the survival probabilities at 20 weeks of gestation reflect the cured portions in each group.

Accounting for the left truncation, classical survival analysis methods including the Cox proportional hazards regression model have been advocated in the literature [MS08, XC11]. As a comparison, Table 1.4 right columns (lower half) show the results of the classic Cox regression model fitted to the data by treating all the cured individuals as right-censored at 20 weeks of gestation, as is currently done in the practical analysis of SAB data [CJX$^+$13]. BMI and alcohol are no longer significant predictors of SAB. Note that under the proportional hazards assumption, nonsignificant effects of BMI or alcohol translates to no significant differences in the cumulative risks of SAB; that is, the impact on the timing of SAB is no longer distinguished from the impact on the overall cumulative risk of SAB (Y/N) by 20 weeks of gestation. In addition, as mentioned before, treating the majority of the women (who did not have SAB) as right-censored can lead to substantial loss of information.

Finally we also fit the 'naive' logistic regression model alone to the data, using whether a woman has SAB (Y/N) as the outcome, while excluding the two right-censored observations. The results are also given in the right columns (upper half) of Table 1.4. We note that this model does not properly handle left truncation, and the results should be not trusted.

1.6 Discussion and Conclusion

In this paper we have developed an NPMLE approach to fit the mixture type cure rate models to data with left truncation in addition to right-censoring. As illustrated in the data analysis, the cure rate model methodology developed here is able to make use of the information from both the women who had SAB and those who were observed not to have SAB, as well as to separate the differential regression effects of the covariates on both the cumulative risk of SAB as well as the timing of it among those who experience SAB. We anticipate this methodology to impact the practical analysis of pregnancy and other similar types of data. An 'alpha' version of a corresponding R package is currently being tested internally.

Different from the usual cure data literature where the long-term survivors are always right-censored, in our pregnancy studies we observe the majority of the 'cured' women. This greatly improves the practical identifiability of the cured portion (Sy and Taylor, 2000; Lu and Ying, 2004), as well as substantially increases the amount of information available for estimating the model parameters. Our inference procedures utilize the NPMLE, together with the "ghost copy" EM algorithm to produce estimators for the model parameters. The variances of the estimators can be obtained in closed form using the [Lou82] formula. In our simulations, the variance estimator leads to relatively accurate coverage of the 95% confidence intervals.

In our proof for consistency, we have worked through an unbounded cumulative baseline hazard, which has rarely been discussed in existing literatures. Ideally, we would like to show asymptotic normality without assuming the interval-censoring tail window. However, the weak convergence of nonparametric estimators often requires a stronger set of assumptions. As a

result, the unbounded Λ_0 in the log-likelihood causes trouble in the Fréchet differentiability and continuously invertibility steps. The "chop-off" argument applied in consistency does not work here as Λ_0 appears in both the parametric and the nonparametric part of the directional score.

Finally for left truncated data much work has been done recently under the length-biased assumption [AWZ06, NQS10, QNLS11, among others]. For enrollment into observational pregnancy studies like ours, we do not think that the uniform distributional assumption necessarily holds, as is evident in Figure 1.1. Other parametric assumptions might be explored that are more suitable for the entry times to pregnancy studies.

1.7 Proofs

1.7.1 The Existence of NPMLE

Proof of Theorem 1. Let θ_B be the maximizer on the compliment of compact set $\{\|\boldsymbol{\alpha}\| \vee \|\boldsymbol{\beta}\| \vee \|\boldsymbol{\lambda}\| \leq B\}$. We show that $l(\theta_B) \to -\infty$ when $B \to \infty$.

By Assumptions 1 and 2, we have the bound (1.17).

All terms in the log-likelihood are bounded except for

$$\sum_{i=1}^{n} \left\{ \delta_i^1 \log \lambda(X_i) - \delta_i^1 e^{\boldsymbol{\beta}^\top \mathbf{Z}_{\mathbf{2}i}} \Lambda(X_i) \right\}.$$

Let $\lambda_{\max}$ be the largest element in $\boldsymbol{\lambda}$. The expression above has the upper bound

$$\log(\lambda_{\max}/m) - \lambda_{\max}/m - K \log m,$$

which diverges to $-\infty$ when we set $B \to \infty$.

Then, the global maximizer must be in one of the compact set $\{\|\boldsymbol{\alpha}\| \vee \|\boldsymbol{\beta}\| \vee \|\boldsymbol{\lambda}\| \leq B^*\}$ for some $B^* > 0$. □

Let $W_i^{\boldsymbol{\theta}}(t)$ be defined as in (1.21). We define a generic inequality to be referenced later, for any $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \boldsymbol{\beta}, \Lambda)$ in the parameter space whose baseline cumulative hazard Λ is a step function jumping only at the observed event times, $t_1, \ldots, t_K$:

$$0 < d\Lambda(t_k) \leq \left(\sum_{j=1}^{n} W_j^{\boldsymbol{\theta}}(t_k) e^{\boldsymbol{\beta}^\top \mathbf{Z_2}_j} \right)^{-1} d\bar{N}(t_k), \quad k = 1, \ldots, K. \tag{1.26}$$

The conclusion of the following Lemma is used in the proofs of both Lemma 1 and Theorem 3.

Lemma 2. *Let $\boldsymbol{\theta}_{(n)} = \left(\boldsymbol{\alpha}_{(n)}, \boldsymbol{\beta}_{(n)}, \Lambda_{(n)}\right)$ be a sequence in the parameter space where $\Lambda_{(n)}$ is a non-decreasing step function with jumps only at the observed event times. Suppose that $\boldsymbol{\theta}_{(n)}$ satisfies* (1.26) *and has a subsequence $\boldsymbol{\theta}_{(n_k)}$ converging to a limiting point $\boldsymbol{\theta}^* = (\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*, \Lambda^*)$ a.s.:*

$$\boldsymbol{\alpha}_{(n_k)} - \boldsymbol{\alpha}^* \to 0, \quad \boldsymbol{\beta}_{(n_k)} - \boldsymbol{\beta}^* \to 0, \quad \sup_{t \in [0,\tau]} |e^{-\Lambda_{(n_k)}(t)} - e^{-\Lambda^*(t)}| \to 0, \quad a.s.. \tag{1.27}$$

Under Assumptions 1 - 4,

1. $\Lambda^*(t) < \infty$ *for all* $t < \tau$*;*

2. $\inf_{t \in [0,\zeta]} \mathbb{E}[W^{\boldsymbol{\theta}^*}(t) e^{\boldsymbol{\beta}^{*\top} \mathbf{Z_2}}] > C_w$, *for some* $C_w > 0$.

Proof of Lemma 2. By checking the uniform continuity of $W_i^{\boldsymbol{\theta}}(t) e^{\boldsymbol{\beta}^\top \mathbf{Z_2}_i}$ in $(\boldsymbol{\alpha}, \boldsymbol{\beta}, e^{-\Lambda(t)})$, we may establish

$$\sup_{t \in [0,\tau]} \left| W_i^{\boldsymbol{\theta}^*}(t) e^{\boldsymbol{\beta}^{*\top} \mathbf{Z_2}_i} - W_i^{\boldsymbol{\theta}_{(n_k)}}(t) e^{\boldsymbol{\beta}_{(n_k)}^\top \mathbf{Z_2}_i} \right| \to 0, \quad a.s..$$

$W_i^{\boldsymbol{\theta}}(t)$ as a function of observed random variables belongs to a Glivenko-Cantelli class of uniformly bounded functions with uniformly bounded variation. Thus, the pointwise convergence

can be strengthen to be uniform convergence,

$$\sup_{t\in[0,\tau]}\left|\frac{1}{n}\sum_{i=1}^{n_k} W_i^{\boldsymbol{\theta}_{(n_k)}}(t)e^{\boldsymbol{\beta}_{(n_k)}^\top \mathbf{Z_2}_i} - \mathbb{E}\left[W^{\boldsymbol{\theta}^*}(t)e^{\boldsymbol{\beta}^{*\top}\mathbf{Z_2}}\right]\right| \xrightarrow{a.s.} 0.$$

Note that $n^{-1}\sum_{i=1}^{n_k} W_i^{\boldsymbol{\theta}_{(n_k)}}(t)e^{\boldsymbol{\beta}_{(n_k)}^\top \mathbf{Z_2}_i}$ is càglàd, so its limit $\mathbb{E}[W^{\boldsymbol{\theta}^*}(t)e^{\boldsymbol{\beta}^{*\top}\mathbf{Z_2}}]$ must also be càglàd.

1. Let $\tau^* = \inf\{t \in [0,\zeta] : e^{-\Lambda^*(t)} = 0\}$. We shall prove that $\tau^* = \tau$.

   Suppose that τ^* is an interior point of $[0,\tau]$. From Assumption 4, $d\Lambda_0([s,t]) = \Lambda_0(t) - \Lambda_0(s) > 0$ for any $s < t$ in $[0,\tau]$. By the definition of τ^*, $\Lambda^*(t) = \infty$ and $\phi^{\boldsymbol{\theta}^*}(t) = 0$ for $t \in [\tau^*,\tau]$, so we have

   $$\mathbb{E}\left[W^{\boldsymbol{\theta}^*}(\tau^*)e^{\boldsymbol{\beta}^{*\top}\mathbf{Z_2}}\right] = \mathbb{E}\left[\int_{\tau^*_-}^{\tau} e^{\boldsymbol{\beta}^{*\top}\mathbf{Z_2}} dN(u)\right] > 0.$$

   By the left continuity of $W_i^{\boldsymbol{\theta}}(t)$, $\exists\, s < \tau^*$, s.t.

   $$\inf_{t\in[s,\tau^*]} \mathbb{E}\left[W^{\boldsymbol{\theta}^*}(t)e^{\boldsymbol{\beta}^{*\top}\mathbf{Z_2}}\right] \geq \frac{1}{2}\mathbb{E}\left[\int_{\tau^*_-}^{\tau} e^{\boldsymbol{\beta}^{*\top}\mathbf{Z_2}} dN(u)\right].$$

   The total increment of $\Lambda_{(n_k)}$ in $[s,\tau^*]$ must be bounded almost surely according to (1.26). By the definition of τ^*, $\Lambda^*(s) < \infty$. Putting these together, we reach the contradiction,

   $$\begin{aligned}\Lambda^*(\tau^*) \leq \varlimsup_{k\to\infty} \Lambda_{(n_k)}(\tau^*) \leq \varlimsup_{k\to\infty} \Lambda_{(n_k)}(s) + \int_{s_+}^{\tau^*} \frac{d\bar{N}(u)}{\sum_{i=1}^{n_k} W_i^{\boldsymbol{\theta}_{(n_k)}}(u)e^{\boldsymbol{\beta}_{(n_k)}^\top \mathbf{Z_2}_i}} \\ \leq \Lambda^*(s) + \frac{\tau^* - s}{\inf_{t\in[s,\tau^*]}\mathbb{E}[W^{\boldsymbol{\theta}^*}(t)e^{\boldsymbol{\beta}^{*\top}\mathbf{Z_2}}]} < \infty.\end{aligned}$$

   The other case is $\tau^* = 0$. Then, $\Lambda^*(t) = \infty$ and $\phi^{\boldsymbol{\theta}^*}(t) = 0$ for $t \in [0,\tau]$. The contradiction is easily established as

   $$\mathbb{E}\left[W^{\boldsymbol{\theta}^*}(0)e^{\boldsymbol{\beta}^{*\top}\mathbf{Z_2}}\right] = \mathbb{E}\left[\int_{0}^{\tau} e^{\boldsymbol{\beta}^{*\top}\mathbf{Z_2}} dN(u)\right] > 0.$$

2. Since $\mathbb{E}[W^{\boldsymbol{\theta}^*}(t)e^{\boldsymbol{\beta}^{*\top}\mathbf{Z_2}}]$ is càglàd, $\boldsymbol{\theta}_{(n_k)}$ satisfies (1.26) and converges uniformly to $\boldsymbol{\theta}^*$, it can be seen that $\mathbb{E}[W^{\boldsymbol{\theta}^*}(t)e^{\boldsymbol{\beta}^{*\top}\mathbf{Z_2}}] \geq 0$ over the interior of $[0,\zeta]$.

Write $n_k^{-1}\sum_{i=1}^{n_k} W_i^{\boldsymbol{\theta}}(t)e^{\boldsymbol{\beta}^\top \mathbf{Z_{2i}}}$ as

$$
\begin{aligned}
& n_k^{-1}\sum_{i=1}^{n_k}\int_{t-}^{\tau}\left\{1-\phi_i^{\boldsymbol{\theta}}(u)\right\}e^{\boldsymbol{\beta}^\top \mathbf{Z_{2i}}}dN_i(u)+\int_t^{\tau} Y_i(u)e^{\boldsymbol{\beta}^\top \mathbf{Z_{2i}}}d\phi_i^{\boldsymbol{\theta}}(u)+Y_i(t)\phi_i^{\boldsymbol{\theta}}(t)e^{\boldsymbol{\beta}^\top \mathbf{Z_{2i}}} \\
=& n_k^{-1}\sum_{i=1}^{n_k}\int_{t+}^{\tau}\left[1-\phi_i^{\boldsymbol{\theta}}(u)-\frac{\sum_{j=1}^{n_k} Y_j(u)\phi_j^{\boldsymbol{\theta}}(u)\left\{1-\phi_j^{\boldsymbol{\theta}}(u)\right\}e^{\boldsymbol{\beta}^\top \mathbf{Z_{2j}}}}{\sum_{j=1}^{n_k} W_j^{\boldsymbol{\theta}}(u)e^{\boldsymbol{\beta}^\top \mathbf{Z_{2j}}}}\right]e^{\boldsymbol{\beta}^\top \mathbf{Z_{2i}}}dN_i(u) \\
& +\left\{1-\phi_i^{\boldsymbol{\theta}}(t)\right\}e^{\boldsymbol{\beta}^\top \mathbf{Z_{2i}}}dN_i(t)+Y_i(t)\phi_i^{\boldsymbol{\theta}}(t)e^{\boldsymbol{\beta}^\top \mathbf{Z_{2i}}}. \qquad (1.28)
\end{aligned}
$$

By Assumption 4, all $Q_i < \zeta$ a.s.. Thus,

$$
\begin{aligned}
\mathbb{E}\left[W^{\boldsymbol{\theta}^*}(\zeta)e^{\boldsymbol{\beta}^{*\top}\mathbf{Z_2}}\right] =& \mathbb{E}\left[\left\{\delta^1+\delta^c\phi^{\boldsymbol{\theta}^*}(X)\right\}I\{\zeta\leq X\}e^{\boldsymbol{\beta}^{*\top}\mathbf{Z_2}}\right] \\
\geq& \mathbb{E}\left[\int_\zeta^{\tau} e^{\boldsymbol{\beta}^{*\top}\mathbf{Z_2}}dN(u)\right] > 0.
\end{aligned}
$$

For $t<\zeta$, the difference $\mathbb{E}[W^{\boldsymbol{\theta}^*}(t)e^{\boldsymbol{\beta}^{*\top}\mathbf{Z_2}}] - \mathbb{E}[W^{\boldsymbol{\theta}^*}(\zeta)e^{\boldsymbol{\beta}^{*\top}\mathbf{Z_2}}]$ is the limit of an integral like that in (1.28), where the integrand has $\sum_{j=1}^{n_k} W_j^{\boldsymbol{\theta}}(u)e^{\boldsymbol{\beta}^\top \mathbf{Z_{2j}}}$ in the denominator. So it has potential singularities at the zeros of $\mathbb{E}[W^{\boldsymbol{\theta}^*}(u)e^{\boldsymbol{\beta}^{*\top}\mathbf{Z_2}}]$ for $u\in[t,\zeta]$. We shall show that $\mathbb{E}[W^{\boldsymbol{\theta}^*}(u)e^{\boldsymbol{\beta}^{*\top}\mathbf{Z_2}}]$ is differentiable with respect to $d\Lambda_0(u)$ in $[0,\zeta]$, so that its zero u_0 leads to the divergent form $-\int_t^\zeta |u-u_0|^{-1}du$. We will then reach the contradiction that $\mathbb{E}[W^{\boldsymbol{\theta}^*}(t)e^{\boldsymbol{\beta}^{*\top}\mathbf{Z_2}}] = -\infty$, as seen below.

Denote R_0 the set of zeros and limiting zeros from right for $\mathbb{E}[W^{\boldsymbol{\theta}^*}(u)e^{\boldsymbol{\beta}^{*\top}\mathbf{Z_2}}]$. Let set $R_{\triangle u}$ be the $\triangle u$ neighborhood of R_0 and $\Omega^t_{\triangle u} = [t,\zeta]\setminus R_{\triangle u}$. $\mathbb{E}[W^{\boldsymbol{\theta}^*}(u)e^{\boldsymbol{\beta}^{*\top}\mathbf{Z_2}}]$ is bounded away from zero on $\Omega^t_{\triangle u}$. Through (1.28),

$$
\mathbb{E}\left[W^{\boldsymbol{\theta}^*}(t)e^{\boldsymbol{\beta}^{*\top}\mathbf{Z_2}}\right] - \mathbb{E}\left[W^{\boldsymbol{\theta}^*}(\zeta)e^{\boldsymbol{\beta}^{*\top}\mathbf{Z_2}}\right]
$$

$$\leq -\int_{\Omega^t_{\triangle u}} \frac{\mathbb{E}[Y(u)\phi^{\boldsymbol{\theta}^*}(u)\{1-\phi^{\boldsymbol{\theta}^*}(u)\}e^{\boldsymbol{\beta}^{*\top}\mathbf{Z_2}}]}{\mathbb{E}[W^{\boldsymbol{\theta}^*}(u)e^{\boldsymbol{\beta}^{*\top}\mathbf{Z_2}}]}\mathbb{E}\left[e^{\boldsymbol{\beta}^{*\top}\mathbf{Z_2}}dN(u)\right]$$
$$+\mathbb{E}\left[\int_{t+}^{\zeta}\{1-\phi^{\boldsymbol{\theta}^*}(u)\}dN(u)+\{1-\phi^{\boldsymbol{\theta}^*}(t)\}e^{\boldsymbol{\beta}^{*\top}\mathbf{Z_2}}dN(t)+Y(t)\phi^{\boldsymbol{\theta}^*}(t)e^{\boldsymbol{\beta}^{*\top}\mathbf{Z_2}}\right]. \quad (1.29)$$

From part 1, $e^{-\Lambda^*(\zeta)} > 0$. For any $u < \zeta$,

$$\phi_i^{\boldsymbol{\theta}^*}(u) \geq \phi_i^{\boldsymbol{\theta}^*}(\zeta) \geq \frac{m^{-1}e^{-m\Lambda^*(\zeta)}}{1+m^{-1}e^{-m\Lambda^*(\zeta)}} > 0.$$

So the limit of numerator term $\mathbb{E}[Y(u)\phi^{\boldsymbol{\theta}^*}(u)\{1-\phi^{\boldsymbol{\theta}^*}(u)\}e^{\boldsymbol{\beta}^{*\top}\mathbf{Z_2}}]$ is bounded away from zero. And $\forall u \in [0,\zeta]$,

$$\left|\frac{d\mathbb{E}W^{\boldsymbol{\theta}^*}(u)}{d\Lambda_0(u)}\right| = \left|\mathbb{E}\left[\{1-\phi^{\boldsymbol{\theta}^*}(u)\}Y(u)\phi^{\boldsymbol{\theta}_0}(u)e^{\boldsymbol{\beta}_0^\top\mathbf{Z_2}} - \phi^{\boldsymbol{\theta}^*}(u)\frac{d\mathbb{E}[Y(u)|\mathbf{Z_1},\mathbf{Z_2}]}{d\Lambda_0(u)}\right]\right|$$
$$\leq m+\mathcal{L} < \infty.$$

The first term in (1.29) diverges to $-\infty$ when $\triangle u \to 0$. The other terms are bounded, so this is the desired contradiction.

□

Proof of Lemma 1. 1. Define the marginal of the complete data likelihood

$$\widetilde{L}(\boldsymbol{\theta}) = \sum_{A_i=0,1}\sum_{M_i=0}^{\infty}\sum_{\widetilde{T}_{i1}=t_k:t_k\leq Q_i}\cdots\sum_{\widetilde{T}_{iM_i}=t_k:t_k\leq Q_i} L_i^c(\boldsymbol{\theta})$$
$$=\prod_{i=1}^{n}\frac{\{p_i\lambda_i(X_i)S_i(X_i)\}^{\delta_i^1}(1-p_i)^{\delta_i^0}\{p_iS_i(X_i)+1-p_i\}^{\delta_i^c}}{1-p_i\sum_{k:t_k\leq Q_i}\lambda_k e^{\boldsymbol{\beta}^\top\mathbf{Z_{2i}}}S_i(t_k)}.$$

From (1.5) it can be seen that the complete data likelihood $L^c(\boldsymbol{\theta})$ can be decomposed into the product of one logistic part with one Cox part. The Assumptions 1 - 3 contain the

regularity conditions of these two parts. The event rate $\mathbb{P}(A_i = 1)$ is bounded away from both zero and one,

$$0 < \frac{m^{-1}}{m^{-1}+1} \leq \mathbb{P}(A_i = 1) \leq \frac{m}{m+1} < 1.$$

The average at-risk process $\mathbb{E}[Y_i(t)]$ is bounded away from zero almost surely. The matrices $\mathbf{Z_1}$ and $\mathbf{Z_2}$ are almost surely of full rank, as $\text{Var}(\mathbf{Z_1})$ and $\text{Var}(\mathbf{Z_2})$ are positive definite. Under these conditions, both parts of the likelihood are concave in the associated sets of parameters, $\boldsymbol{\alpha}$ and $(\boldsymbol{\beta}, \boldsymbol{\lambda})$, respective. Thus, $L^c(\boldsymbol{\theta})$ is almost surely concave in $\boldsymbol{\theta}$. $\widetilde{L}(\boldsymbol{\theta})$ is also concave as the sum over concave functions. The almost sure convergence of the EM algorithm is guaranteed by the almost sure concaveness of the marginal of the complete data likelihood [DLR77].

2. To prove the second result, we take the following strategy. For any $\boldsymbol{\theta}$ denote $\lambda_{\max,\zeta} = \max\{\lambda_k : t_k \leq \zeta\}$, where ζ is the upper bound of truncation time defined in Assumption 4. Define a set in the parameter space:

$$\Theta = \left\{\boldsymbol{\theta} = (\boldsymbol{\alpha}, \boldsymbol{\beta}, \Lambda) | \lambda_{\max,\zeta} \leq n^{-1} 2/C_w\right\}, \tag{1.30}$$

with C_w defined in Lemma 2. We would like to show that

$$\lim_{n\to\infty} \mathbb{P}(\widehat{\boldsymbol{\theta}}, \widetilde{\boldsymbol{\theta}} \in \widehat{\Theta}) = 1. \tag{1.31}$$

This is done through applying Lemma 2, so we will need to verify condition (1.26) for $\widetilde{\boldsymbol{\theta}}$ and $\widehat{\boldsymbol{\theta}}$. The convergence of the EM algorithm is obtained in the first step.

First, we show that the EM finds the unique stationary point of $\widetilde{L}(\boldsymbol{\theta})$, which then must be the global maximizer since it is concave from the proof of part 1. Consider the conditional

expectation given the observed data as in (1.8) - (1.10). It can be verified directly (we skip the algebraic details here) that:

$$\nabla \log \widetilde{L}(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\theta}}[\nabla \log L^c(\boldsymbol{\theta}) | O].$$

The estimator $\widetilde{\boldsymbol{\theta}}$ is by definition the solution to the left-hand side of the above being zero, hence also the stationary point of $\widetilde{L}(\boldsymbol{\theta})$.

We write down the stationary equation $\boldsymbol{\theta}^{(l)} = \boldsymbol{\theta}^{(l+1)} = \widetilde{\boldsymbol{\theta}}$ for $\widetilde{\lambda}_k$'s at convergence,

$$\widetilde{\lambda}_k = \frac{1 + \widetilde{\lambda}_k \sum_{i=1}^n \frac{\widetilde{p}_i e^{\widetilde{\boldsymbol{\beta}}^\top \mathbf{Z_2}_i} \widetilde{S}_i(t_k) I(Q_i \geq t_k)}{1 - \widetilde{p}_i \sum_{h:h<Q_i} \widetilde{f}_i(t_h)}}{\sum_{i=1}^n \left\{ \delta_i^1 I(X_i \geq t_k) + \delta_i^c \phi_i^{\widetilde{\boldsymbol{\theta}}}(X_i) I(X_i \geq t_k) + \sum_{j \geq k} \frac{\widetilde{p}_i \widetilde{f}_i(t_j) I(Q_i \geq t_j)}{1 - \widetilde{p}_i \sum_{h:h<Q_i} \widetilde{f}_i(t_h)} \right\} e^{\widetilde{\boldsymbol{\beta}}^\top \mathbf{Z_2}_i}}, \quad (1.32)$$

where f_i was previously defined just above (1.6). Combining $\widetilde{\lambda}_k$ terms leads to

$$\widetilde{\lambda}_k^{-1} = \sum_{i=1}^n \Bigg\{ \delta_i^1 I(X_i \geq t_k) + \delta_i^c \phi_i^{\widetilde{\boldsymbol{\theta}}}(X_i) I(X_i \geq t_k) - \widetilde{p}_i \frac{\widetilde{S}_i(t_k) I(Q_i \geq t_k) - \sum_{j \geq k} \widetilde{f}_i(t_j) I(Q_i \geq t_j)}{1 - \widetilde{p}_i \sum_{h:h<Q_i} \widetilde{f}_i(t_h)} \Bigg\} e^{\widetilde{\boldsymbol{\beta}}^\top \mathbf{Z_2}_i}. \quad (1.33)$$

By the mean value theorem,

$$0 \leq e^{\lambda_k e^{\boldsymbol{\beta}^\top \mathbf{Z_2}_i}} - 1 - \lambda_k e^{\boldsymbol{\beta}^\top \mathbf{Z_2}_i} \leq \frac{1}{2} \left(\lambda_k e^{\boldsymbol{\beta}^\top \mathbf{Z_2}_i} \right)^2 e^{\lambda_k e^{\boldsymbol{\beta}^\top \mathbf{Z_2}_i}} \leq \frac{1}{2} m^2 \lambda_k^2 e^{\lambda_k m}. \quad (1.34)$$

where m is defined in (1.17). Applying (1.34) to the denominator in (1.33), we get

$$1 - \widetilde{p}_i \sum_{h:h<Q_i} \widetilde{f}_i(t_h) \geq 1 - \widetilde{p}_i \{1 - \widetilde{S}_i(Q_i)\}.$$

By a similar argument, we have almost surely

$$\widetilde{S}_i(t_k) I(Q_i \geq t_k) - \sum_{j \geq k} \widetilde{f}_i(t_j) I(Q_i \geq t_j)$$

$$
\begin{aligned}
&= \widetilde{S}_i(Q_i)I(Q_i \geq t_k) + \sum_{j \geq k} \left\{ 1 - e^{-\widetilde{\lambda}_j e^{\widetilde{\beta}^\top \mathbf{Z_{2}}_i}} - \widetilde{\lambda}_j e^{\widetilde{\beta}^\top \mathbf{Z_{2}}_i} \right\} \widetilde{S}_i(t_j) I(Q_i > t_j) \\
&\leq \widetilde{S}_i(Q_i)I(Q_i \geq t_k).
\end{aligned}
$$

Then, $\widetilde{\boldsymbol{\theta}}$ satisfies (1.26).

For $\widehat{\boldsymbol{\theta}}$, it must satisfy the score equation for λ_k's:

$$
\frac{\partial l(\boldsymbol{\theta})}{\partial \lambda_k} = \sum_{i=1}^{n} \left\{ \frac{dN_i(t_k)}{\lambda_k} - W_i^{\boldsymbol{\theta}}(t_k) e^{\boldsymbol{\beta}^\top \mathbf{Z_{2}}_i} \right\} = 0, \quad \forall k = 1, \ldots, K.
$$

This is the equation version of (1.26) after rearrangement.

Now let $\widehat{\lambda}_{\max,\zeta}$ and $\widetilde{\lambda}_{\max,\zeta}$ be the largest jump for $\widehat{\Lambda}$ and $\widetilde{\Lambda}$ on $[0, \zeta]$, correspondingly. By Lemma 2 part 2, we have

$$
\limsup_{n \to \infty} n\widehat{\lambda}_{\max,\zeta} \leq C_w^{-1}, \quad \limsup_{n \to \infty} n\widetilde{\lambda}_{\max,\zeta} \leq C_w^{-1}, a.s..
$$

Hence (1.31) is established.

In the set Θ, we evaluate the discrepancy between $\log \widetilde{L}(\boldsymbol{\theta})$ and $\log L(\boldsymbol{\theta})$, which can be bounded as following

$$
1 - S_i(Q_i) - \sum_{k: t_k < Q_i} \lambda_k e^{\boldsymbol{\beta}^\top \mathbf{Z_{2}}_i} S_i(t_k) = \sum_{k: t_k < Q_i} S_i(t_k) \left(e^{\lambda_k e^{\boldsymbol{\beta}^\top \mathbf{Z_{2}}_i}} - 1 - \lambda_k e^{\boldsymbol{\beta}^\top \mathbf{Z_{2}}_i} \right). \tag{1.35}
$$

Applying (1.34) to $|\log L(\boldsymbol{\theta}) - \log \widetilde{L}(\boldsymbol{\theta})|$, we have the bound

$$
\begin{aligned}
\left| \log L(\boldsymbol{\theta}) - \log \widetilde{L}(\boldsymbol{\theta}) \right| &\leq \sum_{i=1}^{n} \left| \log \{1 - p_i + p_i S_i(Q_i)\} - \log \left\{ 1 - p_i \sum_{k: t_k < Q_i} \lambda_k e^{\boldsymbol{\beta}^\top \mathbf{Z_{2}}_i} S_i(t_k) \right\} \right| \\
&\leq \sum_{i=1}^{n} \left| \frac{p_i}{1 - p_i} \frac{n}{2} m^2 \lambda_k^2 e^{\lambda_k m} \right| \leq \frac{1}{2} n^2 e^{m \lambda_{\max,\zeta}} m^3 \lambda_{\max,\zeta}^2.
\end{aligned}
$$

Using the upper bound for $\lambda_{\max,\zeta}$ in Θ, we can bound

$$
\sup_{\boldsymbol{\theta} \in \Theta} \left| \log L(\boldsymbol{\theta}) - \log \widetilde{L}(\boldsymbol{\theta}) \right| \leq e^{\frac{2m}{C_w}} \frac{2m^3}{C_w^2}. \tag{1.36}
$$

In summary whenever $\widehat{\boldsymbol{\theta}}, \widetilde{\boldsymbol{\theta}} \in \Theta$, we have

$$0 \leq \log L(\widehat{\boldsymbol{\theta}}) - \log L(\widetilde{\boldsymbol{\theta}}) \leq \log L(\widehat{\boldsymbol{\theta}}) - \log \widetilde{L}(\widehat{\boldsymbol{\theta}}) + \log \widetilde{L}(\widetilde{\boldsymbol{\theta}}) - \log L(\widetilde{\boldsymbol{\theta}}) < e^{\frac{2m}{C_w}} \frac{4m^3}{C_w^2}. \quad (1.37)$$

Combining (1.37) and (1.31) completes the proof.

□

Proof of Theorem 2 and 2'. From Lemma 1, we only need to establish the following two facts: 1) $\mathbb{E}[l_1(\boldsymbol{\theta})]$ exists with one unique maximal, and 2) it is locally invertible at the maximal. We will see that 1) is verified through the proof of Theorem 3, and 2) is verified through the proof of Theorem 4. □

1.7.2 Consistency of NPMLE

Proof of Theorem 3. The constants m, c, ε and $\mathcal{L}$ are defined in (1.17), (1.18) and (1.19).

First, we show that the "bridge" $\bar{\Lambda}$ defined in (1.22) converges to the true Λ_0 in the following sense:

$$\sup_{t \in [0,\tau]} \left| e^{-\bar{\Lambda}(t)} - e^{-\Lambda_0(t)} \right| \to 0, a.s. \quad (1.38)$$

as $n \to \infty$. We have the bound for $\forall t \in (0,\tau)$,

$$m \geq \frac{\mathbb{E}\left[Y(t) \phi^{\boldsymbol{\theta}_0}(t) e^{\boldsymbol{\beta}_0^\top \mathbf{Z_2}} \right]}{\mathbb{E}\left[\log\left\{ 1 + \exp\left(\boldsymbol{\alpha}_0^\top \mathbf{Z_1} - \Lambda_0(t) e^{\boldsymbol{\beta}_0^\top \mathbf{Z}_{2i}} \right) \right\} \right]} \geq \frac{\varepsilon}{m^2 + m}. \quad (1.39)$$

For any $\tau^* < \tau$ in $\mathbb{Q}$ the set of rational numbers, $\mathbb{E}[Y(t)\phi^{\boldsymbol{\theta}_0}(t)e^{\boldsymbol{\beta}_0^\top \mathbf{Z_2}}]$ is bounded away from zero over $[0,\tau^*]$. The uniform convergence of $\bar{\Lambda}$ to Λ_0 over any $[0,\tau^*]$ can be obtained in the way like [Mur94]. To extend the result to (1.38), we use a trick described in (1.40)-(1.43). By Assumption

3, Λ_0 is non-decreasing and diverges to ∞ at τ. Therefore,

$$\forall \varepsilon > 0, \exists \tau^* \in (0,\tau) \cap \mathbb{Q}, s.t. e^{-\Lambda_0(\tau^*)} < \varepsilon/3. \tag{1.40}$$

Through Rao's law of large number and Helly-Bray argument, we have

$$\sup_{t\in[0,\tau^*]} |\bar{\Lambda}(t) - \Lambda_0(t)| \to 0, \quad a.s.. \tag{1.41}$$

By continuity of the exponential function,

$$\exists N, \forall n > N, \sup_{t\in[0,\tau^*]} |e^{-\bar{\Lambda}(t)} - e^{-\Lambda_0(t)}| < \varepsilon/3. \tag{1.42}$$

Then,

$$\forall n > N, \sup_{t\in[\tau^*,\tau]} |e^{-\bar{\Lambda}(t)} - e^{-\Lambda_0(t)}| \leq 2e^{-\Lambda_0(\tau^*)} + |e^{-\bar{\Lambda}(\tau^*)} - e^{-\Lambda_0(\tau^*)}| < \varepsilon. \tag{1.43}$$

Therefore, we have proved (1.38).

Next, we evaluate the difference between the limits of $\widehat{\Lambda}$ and $\bar{\Lambda}$. According to Assumption 1 and $e^{-\widehat{\Lambda}(t)} \in [0,1]$, $(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}, e^{-\widehat{\Lambda}(t)})$ is bounded. $\widehat{\Lambda}(t)$ is Càdlàg, so is $e^{-\widehat{\Lambda}(t)}$. By Helly's Selection theorem, there is a subsequence converging uniformly almost surely to some $\boldsymbol{\theta}^* = (\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*, e^{-\Lambda^*})$. Lemma 2 part 2 gives the bound for $\mathbb{E}\{W^{\boldsymbol{\theta}}(t)e^{\boldsymbol{\beta}^\top \mathbf{Z_2}}\}$ over $[0,\zeta]$. We only need to find its bound on $[\zeta,\tau]$ in order to mimic the proof of Lemma 1 of [Mur94]. Note that

$$\begin{aligned}\mathbb{E}\left[W^{\boldsymbol{\theta}}(t)e^{\boldsymbol{\beta}^\top \mathbf{Z_2}}\right] =& \mathbb{E}\left[\int_{t-}^{\tau} \left\{1 - \phi^{\boldsymbol{\theta}}(u)\right\} e^{\boldsymbol{\beta}^\top \mathbf{Z_2}} dN(u)\right] \\ & - \mathbb{E}\left[\int_t^{\tau} \phi^{\boldsymbol{\theta}}(u) e^{\boldsymbol{\beta}^\top \mathbf{Z_2}} d\mathbb{E}[Y(u)|\mathbf{Z_1}, \mathbf{Z_2}]\right].\end{aligned}$$

By Assumption 4, $\mathbb{P}(Q_i \leq \zeta) = 1$, so $\mathbb{E}[Y(u)|\mathbf{Z_1}, \mathbf{Z_2}]$ is decreasing on $[\zeta,\tau]$. Along with the Lipschitz continuity, we have for $\forall t \in [\zeta,\tau)$

$$M\mathcal{L} \geq \frac{\mathbb{E}[W^{\boldsymbol{\theta}}(t)e^{\boldsymbol{\beta}^\top \mathbf{Z_2}}]}{\mathbb{E}\left[\log\left\{1 + \exp\left(\boldsymbol{\alpha}_0^\top \mathbf{Z_1} - \Lambda_0(t)e^{\boldsymbol{\beta}_0^\top \mathbf{Z}_{2i}}\right)\right\}\right]} \geq \frac{\varepsilon}{m^2 + m}.$$

Therefore, $\gamma(t) = \frac{\mathbb{E}\left[W^{\boldsymbol{\theta}_0}(t)e^{\boldsymbol{\beta}^\top \mathbf{Z_2}}\right]}{\mathbb{E}\left[W^{\boldsymbol{\theta}^*}(t)e^{\boldsymbol{\beta}^{*\top} \mathbf{Z_2}}\right]}$ is bounded away from both ∞ and zero, and

$$\sup_{t\in[0,\tau]}\left|\frac{d\widehat{\Lambda}}{d\bar{\Lambda}}(t)-\gamma(t)\right|\to 0 \text{ and } \sup_{t\in[0,\tau^*]}\left|\widehat{\Lambda}(t)-\int_0^t \gamma d\Lambda_0\right|\to 0\ a.s., \forall \tau^*<\tau \text{ in } \mathbb{Q}. \tag{1.44}$$

After all these preparation, we can use the semi-parametric Kullback-Leibler divergence argument from [Mur94]. We have

$$\begin{aligned}
0 \leq & \frac{1}{n}\left\{l_n(\widehat{\boldsymbol{\alpha}},\widehat{\boldsymbol{\beta}},\widehat{\Lambda})-l_n(\boldsymbol{\alpha}_0,\boldsymbol{\beta}_0,\bar{\Lambda})\right\} \\
= & \frac{1}{n}\sum_{i=1}^n\int_0^\tau \log\left\{\frac{\phi_i^{\widehat{\boldsymbol{\theta}}}(u)e^{\widehat{\boldsymbol{\beta}}^\top \mathbf{Z_{2i}}}d\widehat{\Lambda}(u)}{\phi_i^{\boldsymbol{\theta}_0}(u)e^{\boldsymbol{\beta}_0^\top \mathbf{Z_{2i}}}d\bar{\Lambda}(u)}\right\}\left\{dN_i(u)-\phi_i^{\boldsymbol{\theta}_0}(u)Y_i(u)e^{\boldsymbol{\beta}_0^\top \mathbf{Z_{2i}}}d\bar{\Lambda}(u)\right\} \\
& +\int_0^\tau\left[\log\left\{\frac{\phi_i^{\widehat{\boldsymbol{\theta}}}(u)e^{\widehat{\boldsymbol{\beta}}^\top \mathbf{Z_{2i}}}d\widehat{\Lambda}(u)}{\phi_i^{\boldsymbol{\theta}_0}(u)e^{\boldsymbol{\beta}_0^\top \mathbf{Z_{2i}}}d\bar{\Lambda}(u)}\right\}-\left\{\frac{\phi_i^{\widehat{\boldsymbol{\theta}}}(u)e^{\widehat{\boldsymbol{\beta}}^\top \mathbf{Z_{2i}}}d\widehat{\Lambda}(u)}{\phi_i^{\boldsymbol{\theta}_0}(u)e^{\boldsymbol{\beta}_0^\top \mathbf{Z_{2i}}}d\bar{\Lambda}(u)}-1\right\}\right] \\
& \times \phi_i^{\boldsymbol{\theta}_0}(u)e^{\boldsymbol{\beta}_0^\top \mathbf{Z_{2i}}}Y_i(u)d\bar{\Lambda}(u).
\end{aligned} \tag{1.45}$$

Denote the function in the logarithm above as $\psi_i(u)$. Using the definition of $\bar{\Lambda}$, we can rewrite the first term in (1.45) as

$$\begin{aligned}
& \frac{1}{n}\sum_{i=1}^n\left\{\int_0^\tau \log\left(\psi_i(u)\right)-\frac{\sum_{j=1}^n \log\left(\psi_j(u)\right)\phi_j^{\boldsymbol{\theta}_0}(u)Y_j(u)e^{\boldsymbol{\beta}_0^\top \mathbf{Z_{2j}}}}{\sum_{j=1}^n \phi_j^{\boldsymbol{\theta}_0}(u)Y_j(u)e^{\boldsymbol{\beta}_0^\top \mathbf{Z_{2j}}}}\right\}dN_i(u) \\
= & \frac{1}{n}\sum_{i=1}^n\left\{\int_0^\tau \log\left(\psi_i(u)\right)-\frac{\sum_{j=1}^n \log\left(\psi_j(u)\right)\phi_j^{\boldsymbol{\theta}_0}(u)Y_j(u)e^{\boldsymbol{\beta}_0^\top \mathbf{Z_{2j}}}}{\sum_{j=1}^n \phi_j^{\boldsymbol{\theta}_0}(u)Y_j(u)e^{\boldsymbol{\beta}_0^\top \mathbf{Z_{2j}}}}\right\}dM_i(u)
\end{aligned} \tag{1.46}$$

Inside $\psi_i(u)$, the ratio $d\widehat{\Lambda}/d\bar{\Lambda}$ is bounded away from 0 and ∞ according to (1.44). Denote the range of the ratio as $[1/R,R]$. The $\phi_i^{\boldsymbol{\theta}_0}(u)$ term and $\phi_i^{\widehat{\boldsymbol{\theta}}}(u)$ term in $\psi_i(u)$ creates potential singularity for (1.46) at τ, but its decay rate is bounded by $e^{-mR\Lambda_0(u)}$ by Assumptions 1 and 2. The integrands of martingale integral (1.46) are all bounded a.s., and the quadratic variation of (1.46) is bounded a.s. by

$$\frac{1}{n^2}\sum_{i=1}^n\int_0^\tau 4\left\{mR\Lambda_0(u)+\log(R)\right\}^2\phi_i^{\boldsymbol{\theta}_0}(u)Y_i(u)e^{\boldsymbol{\beta}_0^\top \mathbf{Z_{2i}}}d\Lambda_0(u).$$

It is of order $O_p(1/n)$, so the limit of (1.46) is zero almost surely.

The integrands in the second term of (1.45) is of the form $\log(x)-(x-1)\leq 0$. In order to satisfy the inequality in (1.45), we must have

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau}\left\{\log\left(\psi_i(u)\right)-\left(\psi_i(u)-1\right)\right\}\phi_i^{\boldsymbol{\theta}_0}(u)e^{\widehat{\boldsymbol{\beta}}^{\top}\mathbf{Z_{2}}_i}Y_i(u)d\bar{\Lambda}(u)=0.$$

Applying the same argument as in [Mur94], we get

$$\mathbb{E}\left(\int_0^{\tau}\left|\phi^{\boldsymbol{\theta}^*}(u)e^{\boldsymbol{\beta}^{*\top}\mathbf{Z_2}}\gamma(u)-\phi^{\boldsymbol{\theta}_0}(u)e^{\boldsymbol{\beta}_0^{\top}\mathbf{Z_2}}\right|Y(u)d\Lambda_0(u)\right)=0 \tag{1.47}$$

in the almost sure set. The identifiability of our model is verified in [LTS01] Theorem 2. Along with our regularity conditions in Assumptions 2 and 3, (1.47) leads to $\boldsymbol{\alpha}^*=\boldsymbol{\alpha}_0$, $\boldsymbol{\beta}^*=\boldsymbol{\beta}_0$ and $\gamma(t)=1$. This implies that

$$\sup_{t\in[0,\tau^*]}\left|\widehat{\Lambda}(t)-\Lambda_0(t)\right|\to 0\ a.s., \forall\tau^*<\tau \text{ in } \mathbb{Q}.$$

Repeating the trick in (1.40)-(1.43), we have

$$\sup_{t\in[0,\tau]}\left|e^{-\widehat{\Lambda}(t)}-e^{-\Lambda_0(t)}\right|\to 0\ a.s..$$

Finally, we summarize all usage of almost sure arguments to ensure that intersection of all almost sure sets still has probability one under σ-additivity. The steps (1.40)-(1.43) involves one almost sure argument for each choice of τ^*. We preserve the almost sure property by restricting τ^* to be in the countable set $\mathbb{Q}$. One almost sure argument is made for Helly's selection theorem. In Lemma 2, we use the Glivenko-Cantelli Theorem to avoid the dependence on the choice of $\boldsymbol{\theta}^*$, so the almost sure argument is only applied once. Two more almost sure arguments are used in calculating the limit of the terms in (1.45).

□

Proof of Theorem 3'. The proof is essentially the same as the proof of Theorem 3, so the details are omitted. In fact, it is less technical due to the boundedness of Λ_0 over $[0,\tau']$. □

1.7.3 Asymptotic Normality

First, we provide the definition of several quantities below. In Theorem 4 $\sigma(\mathbf{h}) = \Big(\boldsymbol{\sigma}_a(\mathbf{h}), \boldsymbol{\sigma}_b(\mathbf{h}), \sigma_\eta(\mathbf{h})\Big)$ is

$$\begin{aligned}
\boldsymbol{\sigma}_a(\mathbf{h}) &= \mathbb{E}\Bigg[\mathbf{Z_1}\Bigg\{-\int_0^{\tau'} K_1^{\boldsymbol{\theta}_0}(\mathbf{h})(u)Y(u)d\phi^{\boldsymbol{\theta}_0}(u) \\
&\qquad + K_2^{\boldsymbol{\theta}_0}(\mathbf{h})Y(\tau')\phi^{\boldsymbol{\theta}_0}(\tau')\Big(1-\phi^{\boldsymbol{\theta}_0}(\tau')\Big)\Bigg\}\Bigg], \\
\boldsymbol{\sigma}_b(\mathbf{h}) &= \mathbb{E}\Bigg[\mathbf{Z_2}\Bigg\{\int_0^{\tau'} K_1^{\boldsymbol{\theta}_0}(\mathbf{h})(u)Y(u)e^{\boldsymbol{\beta}_0^\top \mathbf{Z_2}}d\Big[\Lambda_0(u)\phi^{\boldsymbol{\theta}_0}(u)\Big] \\
&\qquad - K_2^{\boldsymbol{\theta}_0}(\mathbf{h})Y(\tau')e^{\boldsymbol{\beta}_0^\top \mathbf{Z_2}}\Lambda_0(\tau')\phi^{\boldsymbol{\theta}_0}(\tau')\Big(1-\phi^{\boldsymbol{\theta}_0}(\tau')\Big)\Bigg\}\Bigg], \\
\sigma_\eta(\mathbf{h}) &= \mathbb{E}\Bigg[e^{\boldsymbol{\beta}_0^\top \mathbf{Z_2}}\Bigg\{K_1^{\boldsymbol{\theta}_0}(\mathbf{h})(u)\phi^{\boldsymbol{\theta}_0}(u)Y(u) - K_2^{\boldsymbol{\theta}_0}(\mathbf{h})Y(\tau')\phi^{\boldsymbol{\theta}_0}(\tau')\Big(1-\phi^{\boldsymbol{\theta}_0}(\tau')\Big) \\
&\qquad - \int_u^{\tau'} K_1^{\boldsymbol{\theta}_0}(\mathbf{h})(s)\phi^{\boldsymbol{\theta}_0}(s)\Big(1-\phi^{\boldsymbol{\theta}_0}(s)\Big)Y(s)e^{\boldsymbol{\beta}_0^\top \mathbf{Z_2}}d\Lambda_0(s)\Bigg\}\Bigg], \qquad (1.48)
\end{aligned}$$

where

$$\begin{aligned}
K_1^{\boldsymbol{\theta}}(\mathbf{h})(u) =& \mathbf{a}^\top \mathbf{Z_1}\Big(1-\phi^{\boldsymbol{\theta}}(u)\Big) + \mathbf{b}^\top \mathbf{Z_2}\Big\{1-\Big(1-\phi^{\boldsymbol{\theta}}(u)\Big)\Lambda(u)e^{\boldsymbol{\beta}^\top \mathbf{Z_2}}\Big\} \\
&+ \eta(u) - \Big(1-\phi^{\boldsymbol{\theta}}(u)\Big)e^{\boldsymbol{\beta}^\top \mathbf{Z_2}}\int_0^u \eta d\Lambda, \\
K_2^{\boldsymbol{\theta}}(\mathbf{h}) =& \Big\{\mathbf{a}^\top \mathbf{Z_1} - \mathbf{b}^\top \mathbf{Z_2}\Lambda(\tau')e^{\boldsymbol{\beta}^\top \mathbf{Z_2}} - \int_0^{\tau'} \eta e^{\boldsymbol{\beta}^\top \mathbf{Z_2}}d\Lambda\Big\}. \qquad (1.49)
\end{aligned}$$

Let $\boldsymbol{\theta} + t\mathbf{h} = \Big(\boldsymbol{\alpha} + t\mathbf{a}, \boldsymbol{\beta} + t\mathbf{b}, \int_0^{\cdot}(1+t\eta)d\Lambda\Big)$. Define the directional derivatives

$$\lim_{t\to 0}\frac{l_n^I(\boldsymbol{\theta}+t\mathbf{h}) - l_n^I(\boldsymbol{\theta})}{t} = S_n^{\boldsymbol{\theta}} = S_{n,a}^{\boldsymbol{\theta}} + S_{n,b}^{\boldsymbol{\theta}} + S_{n,\eta}^{\boldsymbol{\theta}},$$

where

$$
\begin{aligned}
S^{\boldsymbol{\theta}}_{n,a} =& \frac{1}{n}\sum_{i=1}^{n} \mathbf{a}^\top \mathbf{Z_{1}}_i \Bigg\{ \int_0^{\tau'} \Big(1-\phi_i^{\boldsymbol{\theta}}(u)\Big) dN_i(u) - \int_0^{\tau'} Y_i(u)\phi_i^{\boldsymbol{\theta}}(u)\Big(1-\phi_i^{\boldsymbol{\theta}}(u)\Big) e^{\boldsymbol{\beta}^\top \mathbf{Z_{2}}_i} d\Lambda(u) \\
&+ \Big(N_i(\tau)-N_i(\tau')\Big)\Big(1-\phi_i^{\boldsymbol{\theta}}(\tau')\Big) - Y_i(\tau)\phi_i^{\boldsymbol{\theta}}(\tau') \Bigg\} \\
S^{\boldsymbol{\theta}}_{n,b} =& \frac{1}{n}\sum_{i=1}^{n} \mathbf{b}^\top \mathbf{Z_{2}}_i \Bigg[\int_0^{\tau'} \Big\{ 1-\Big(1-\phi_i^{\boldsymbol{\theta}}(u)\Big)\Lambda(u) e^{\boldsymbol{\beta}^\top \mathbf{Z_{2}}_i} \Big\} dN_i(u) \\
&+ \int_0^{\tau'} Y_i(u)\phi_i^{\boldsymbol{\theta}}(u) e^{\boldsymbol{\beta}^\top \mathbf{Z_{2}}_i} \Big\{ \Big(1-\phi_i^{\boldsymbol{\theta}}(u)\Big)\Lambda(u) e^{\boldsymbol{\beta}^\top \mathbf{Z_{2}}_i} - 1 \Big\} d\Lambda(u) \\
&- \Big(N_i(\tau)-N_i(\tau')\Big)\Big(1-\phi_i^{\boldsymbol{\theta}}(\tau')\Big)\Lambda(\tau') e^{\boldsymbol{\beta}^\top \mathbf{Z_{2}}_i} + Y_i(\tau)\phi_i^{\boldsymbol{\theta}}(\tau')\Lambda(\tau') e^{\boldsymbol{\beta}^\top \mathbf{Z_{2}}_i} \Bigg] \\
S^{\boldsymbol{\theta}}_{n,\eta} =& \frac{1}{n}\sum_{i=1}^{n} \int_0^{\tau'} \Bigg[\eta(u) - \Big\{1-\phi_i^{\boldsymbol{\theta}}(u)\Big\} e^{\boldsymbol{\beta}^\top \mathbf{Z_{2}}_i} \int_0^{u} \eta d\Lambda \Bigg] dN_i(u) \\
&+ \int_0^{\tau'} Y_i(u)\phi_i^{\boldsymbol{\theta}}(u) e^{\boldsymbol{\beta}^\top \mathbf{Z_{2}}_i} \Bigg[\Big\{1-\phi_i^{\boldsymbol{\theta}}(u)\Big\} e^{\boldsymbol{\beta}^\top \mathbf{Z_{2}}_i} \int_0^{u} \eta d\Lambda - \eta(u) \Bigg] d\Lambda(u) \\
&- \Big(N_i(\tau)-N_i(\tau')\Big)\Big(1-\phi_i^{\boldsymbol{\theta}}(\tau')\Big) \int_0^{\tau'} \eta d\Lambda e^{\boldsymbol{\beta}^\top \mathbf{Z_{2}}_i} + Y_i(\tau)\phi_i^{\boldsymbol{\theta}}(\tau') \int_0^{\tau'} \eta d\Lambda e^{\boldsymbol{\beta}^\top \mathbf{Z_{2}}_i}.
\end{aligned}
$$

Their expectations are denoted as

$$
S^{\boldsymbol{\theta}} = S^{\boldsymbol{\theta}}_a + S^{\boldsymbol{\theta}}_b + S^{\boldsymbol{\theta}}_\eta = \mathbb{E}\left(S^{\boldsymbol{\theta}}_{n,a}\right) + \mathbb{E}\left(S^{\boldsymbol{\theta}}_{n,b}\right) + \mathbb{E}\left(S^{\boldsymbol{\theta}}_{n,\eta}\right).
$$

Again let $\boldsymbol{\theta}_0$ be the true parameter and $\boldsymbol{\theta}$ another element in the paramter space. Define $\triangle\boldsymbol{\theta} = \boldsymbol{\theta} - \boldsymbol{\theta}_0$ with

$$
\triangle\boldsymbol{\alpha} = \boldsymbol{\alpha} - \boldsymbol{\alpha}_0,\ \triangle\boldsymbol{\beta} = \boldsymbol{\beta} - \boldsymbol{\beta}_0 \text{ and } \triangle\Lambda(\cdot) = \Big\{\Lambda(\cdot) - \Lambda_0(\cdot)\Big\}.
$$

Define $lin\Theta$ to be the linear space spanned by $\{\boldsymbol{\theta} - \boldsymbol{\theta}_0 : \boldsymbol{\theta} \text{ in parameter space}\}$. Let $\boldsymbol{\theta}_t = \boldsymbol{\theta}_0 + t\triangle\boldsymbol{\theta}$. The functional Hessian is a linear operator $lin\Theta \mapsto l^\infty(H_p)$ defined as

$$
\dot{S}^{\boldsymbol{\theta}_0}(\triangle\boldsymbol{\theta})(\mathbf{h}) = \lim_{t\to 0} \frac{S^{\boldsymbol{\theta}_t}(\mathbf{h}) - S^{\boldsymbol{\theta}_0}(\mathbf{h})}{t}
$$

$$= - \triangle\boldsymbol{\alpha}^\top \boldsymbol{\sigma}_a(\mathbf{h}) - \triangle\boldsymbol{\beta}^\top \boldsymbol{\sigma}_b(\mathbf{h}) - \int_0^{\tau'} \sigma_\eta(\mathbf{h})(u) d\triangle\Lambda(u) \tag{1.50}$$

with σ defined in (1.48).

The following Lemma 3 is used in the proofs of Theorems 4 and 5. It tells us about the property of σ, the essential element in the functional Hessian.

Lemma 3. *Let the operator* $\sigma : (\mathbf{a}, \mathbf{b}, \eta) \mapsto \Big(\boldsymbol{\sigma}_a(\mathbf{h}), \boldsymbol{\sigma}_b(\mathbf{h}), \sigma_\eta(\mathbf{h})\Big)$ *be defined as in* (1.48)*. Under the conditions of Theorem 4,* σ *is a continuously invertible bijection from* H_∞ *to* H_∞*.*

Proof of Lemma 3. First we prove that σ is injection by an identifiability argument. Define an inner-product between $\sigma(\mathbf{h})$ and $\mathbf{h}$ as

$$\begin{aligned}
\Big\langle \sigma(\mathbf{h}), \mathbf{h} \Big\rangle =& \mathbf{a}^\top \boldsymbol{\sigma}_a(\mathbf{h}) + \mathbf{b}^\top \boldsymbol{\sigma}_b(\mathbf{h}) + \int_0^{\tau'} \sigma_\eta(\mathbf{h})(u) \eta(u) d\Lambda_0(u) \\
=& \int_0^{\tau'} \mathbb{E}\left[\left\{ K_1^{\boldsymbol{\theta}_0}(\mathbf{h})(u) \right\}^2 Y(u) \phi^{\boldsymbol{\theta}_0}(u) e^{\boldsymbol{\beta}_0^\top \mathbf{Z_2}} \right] d\Lambda_0(u) \\
&+ \mathbb{E}\left[\left\{ K_2^{\boldsymbol{\theta}_0}(\mathbf{h}) \right\}^2 Y(\tau') \phi^{\boldsymbol{\theta}_0}(\tau') \Big(1 - \phi^{\boldsymbol{\theta}_0}(\tau') \Big) \right].
\end{aligned}$$

If $\Big\langle \sigma(\mathbf{h}), \mathbf{h} \Big\rangle = 0$, we have almost surely $K_2^{\boldsymbol{\theta}_0}(\mathbf{h}) = 0$ and $K_1^{\boldsymbol{\theta}_0}(\mathbf{h})(u) = 0$ a.e. $u \in [0, \tau']$. Therefore,

$$\int_0^t K_1^{\boldsymbol{\theta}_0}(\mathbf{h})(u) \phi^{\boldsymbol{\theta}_0}(u) e^{\boldsymbol{\beta}_0^\top \mathbf{Z_2}} d\Lambda_0(u) = 0, \forall t \in [0, \tau'], a.s..$$

Calculating the integral, we have for for any $t \in [0, \tau']$ a.s.

$$-\mathbf{a}^\top \mathbf{Z_1} \phi^{\boldsymbol{\theta}_0}(t) + \mathbf{b}^\top \mathbf{Z_2} \phi^{\boldsymbol{\theta}_0}(t) \Lambda_0(t) e^{\boldsymbol{\beta}_0^\top \mathbf{Z_2}} + \int_0^t \eta(u) d\Lambda_0(u) \phi^{\boldsymbol{\theta}_0}(t) e^{\boldsymbol{\beta}_0^\top \mathbf{Z_2}} = 0.$$

Setting $t = 0$, we have $-\mathbf{a}^\top \mathbf{Z_1} \phi^{\boldsymbol{\theta}_0}(0) = 0$, so $\mathbf{a}^\top \mathbf{Z_1} = 0$. By Assumption 2, $\mathbf{a} = 0$. Plugging $\mathbf{a} = 0$ into $K_2^{\boldsymbol{\theta}_0}$ yields

$$K_2^{\boldsymbol{\theta}_0}(\mathbf{h}) = e^{\boldsymbol{\beta}_0^\top \mathbf{Z_2}} \Big\{ \mathbf{b}^\top \mathbf{Z_2} \Lambda_0(\tau') - \int_0^{\tau'} \eta(u) d\Lambda_0(u) \Big\} = 0, a.s..$$

Again, $\mathbf{b}^\top \mathbf{Z_2} = \int_0^{\tau'} \eta(u) d\Lambda_0(u)/\Lambda_0(\tau')$ is deterministic, so $\mathbf{b} = 0$. This way η must also be constantly zero. As a result, $\sigma(\mathbf{h}) = \sigma(\mathbf{h}') \Rightarrow \left(\sigma(\mathbf{h}-\mathbf{h}'), \mathbf{h}-\mathbf{h}'\right) = 0 \Rightarrow \mathbf{h} = \mathbf{h}'$.

To show it is a bijection, we apply Theorem 3.11 in [Con90]. It suffices to decompose σ as the sum of one invertible operator and one compact operator. The invertible operator is defined as

$$\Sigma(\mathbf{h}) = \left(\mathbb{E}\left(\mathbf{Z_1}\mathbf{Z_1}^\top\right)\mathbf{a}, \mathbb{E}\left(\mathbf{Z_2}\mathbf{Z_2}^\top\right)\mathbf{b}, \eta(t)\mathbb{E}\left\{e^{\boldsymbol{\beta}_0^\top \mathbf{Z_2}}\phi^{\boldsymbol{\theta}_0}(t)Y(t)\right\}\right).$$

Since $\mathbb{E}\left(\mathbf{Z_1}\mathbf{Z_1}^\top\right)$, $\mathbb{E}\left(\mathbf{Z_2}\mathbf{Z_2}^\top\right)$ are both positive definite, and $\inf_{t\in[0,\tau']} \mathbb{E}e^{\boldsymbol{\beta}_0^\top \mathbf{Z_2}}\phi^{\boldsymbol{\theta}_0}(t)Y(t) > 0$, the inverse exists as

$$\Sigma^{-1}(\mathbf{h}) = \left(\left[\mathbb{E}\left\{\mathbf{Z_1}\mathbf{Z_1}^\top\right\}\right]^{-1}\mathbf{a}, \left[\mathbb{E}\left\{\mathbf{Z_2}\mathbf{Z_2}^\top\right\}\right]^{-1}\mathbf{b}, \eta(t)\left[\mathbb{E}\left\{e^{\boldsymbol{\beta}_0^\top \mathbf{Z_2}}\phi^{\boldsymbol{\theta}_0}(t)Y(t)\right\}\right]^{-1}\right).$$

For the compactness of $\sigma(\mathbf{h}) - \Sigma(\mathbf{h})$, classical Helly-selection plus dominated convergence method applies as all terms are conveniently bounded. □

The proof of Theorem 4 is the application of Theorem 3.3.1 from [VdVW96]. We shall verify all the required conditions for the Theorem.

Proof of Theorem 4. Since we work under a modified Assumption 3' now, the martingale representation in (1.15) needs to change accordingly beyond τ'. We still use $M_i(t)$ as the notation. Define the filtrations $\left\{\mathcal{F}_t : t \in [0,\tau]\right\}$. On $[0,\tau']$, $\mathcal{F}_t$ is the natural σ-algebra generated by $\{N_i(t), Y_i(t), \mathbf{Z_{1}}_i, \mathbf{Z_{2}}_i, i = 1,\ldots,n\}$. Since there is no extra information in the tail window (τ',τ), we set $\mathcal{F}_t = \mathcal{F}_{\tau'}$ for $t \in (\tau',\tau)$. $\mathcal{F}_\tau$ is the σ-algebra generated by $\{N_i(\tau) - N_i(\tau'), Y_i(\tau), \mathbf{Z_{1}}_i, \mathbf{Z_{2}}_i, i = 1,\ldots,n\}$, where $Y_i(\tau) = Y_i(\tau') - dN_i(\tau')$ is measurable in $\mathcal{F}_{\tau'}$. The filtrations on $[0,\tau']$ stay the same, so $M_i(t)$ defined in (1.15) is still a martingale up to time τ'. In the tail window (τ',τ), we

set $M_i(t)$ constantly equals $M_i(\tau')$. To extend its definition to time τ, we define

$$dM_i(\tau) = M_i(\tau) - M_i(\tau') = \big\{N_i(\tau) - N_i(\tau')\big\} - Y_i(\tau)\phi_i^{\boldsymbol{\theta}_0}(\tau'). \tag{1.51}$$

It is easy to verify that $\mathbb{E}[M_i(\tau)|\mathcal{F}_{\tau'}] = M_i(\tau')$, so $M_i(t)$ thus defined is a martingale with respect to the new filtrations $\big\{\mathcal{F}_t : t \in [0,\tau'] \cup \{\tau\}\big\}$. Analogously, we define the process $M_i^{\boldsymbol{\theta}}(\cdot)$ which replaces the true parameter $\boldsymbol{\theta}_0$ in $M_i(\cdot)$ by arbitrary $\boldsymbol{\theta}$ in the parameter space. Apparently, $M_i^{\boldsymbol{\theta}_0}(\cdot) = M_i(\cdot)$. From here, we establish the needed results based on the martingale theory.

First, we prove weak convergence of the empirical score

$$\sqrt{n}(S_n^{\boldsymbol{\theta}_0} - S^{\boldsymbol{\theta}_0}) \overset{l^\infty(H_p)}{\longrightarrow} \mathcal{W}. \tag{1.52}$$

Notice that $S_1^{\boldsymbol{\theta}_0} - S^{\boldsymbol{\theta}_0}$ is a martingale integral with respect to (1.51). The weak convergence follows from martingale central limit theorem. The covariance process is given by the expectation of its quadratic variation:

$$\begin{aligned}\mathrm{Cov}\big(\mathscr{G}(\mathbf{h}), \mathscr{G}(\mathbf{h}^*)\big) = \mathbb{E}\Big[\int_0^{\tau'} & K_1^{\boldsymbol{\theta}_0}(\mathbf{h}) K_1^{\boldsymbol{\theta}_0}(\mathbf{h}^*) Y(u)\phi_0(u) e^{\boldsymbol{\beta}_0^\top \mathbf{Z_2}} d\Lambda_0(u) \\ & + K_2^{\boldsymbol{\theta}_0}(\mathbf{h}) K_2^{\boldsymbol{\theta}_0}(\mathbf{h}^*)\phi_0(\tau')\big\{1 - \phi_0(\tau')\big\}\Big],\end{aligned}$$

where K_1 and K_2 are defined as in (1.49).

Next, we verify the approximation condition

$$\sqrt{n}\left(S_n^{\widehat{\boldsymbol{\theta}}} - S^{\widehat{\boldsymbol{\theta}}} - S_n^{\boldsymbol{\theta}_0} - S^{\boldsymbol{\theta}_0}\right) = o_p(1). \tag{1.53}$$

Consider the class $\{S_1^{\boldsymbol{\theta}}(\mathbf{h}) - S_1^{\boldsymbol{\theta}_0}(\mathbf{h}) : \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| \le \varepsilon, \mathbf{h} \in H_p\}$. All terms involved in this class are uniformly bounded with uniformly bounded variation, so it is a Donsker class for the set of

observable random variables. By checking that $\phi_i^{\boldsymbol{\theta}}$ is Lipschitz in $\boldsymbol{\theta}$ under the $l^\infty(H_p)$ norm, we have almost surely

$$\sup_{t,\mathbf{Z_2},\mathbf{Z_1}} |\phi_i^{\boldsymbol{\theta}}(t) - \phi_i^{\boldsymbol{\theta}_0}(t)| = O\left(\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|\right),$$

and similarly

$$\sup_{t,\mathbf{Z_2},\mathbf{Z_1}} |\phi_i^{\boldsymbol{\theta}}(t)\Lambda(t) - \phi_i^{\boldsymbol{\theta}_0}(t)\Lambda_0(t)| = O\left(\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|\right).$$

For a single summand in the score,

$$\sup_{h\in H_p} \mathbb{E}[S_1^{\boldsymbol{\theta}}(\mathbf{h}) - S_1^{\boldsymbol{\theta}_0}(\mathbf{h})]^2 = O\left(\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|^2\right).$$

We plug $\widehat{\boldsymbol{\theta}}$ into the expression above. Thus, the variance of the limiting process of (1.53) is $o(1)$ by the consistency of $\widehat{\boldsymbol{\theta}}$ from Theorem 3', so the process itself is $o_p(1)$.

We then show the Fréchet differentiability of expected score S at $\boldsymbol{\theta}_0$ in the direction of $\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0$,

$$S^{\widehat{\boldsymbol{\theta}}_t} - S^{\boldsymbol{\theta}_0} = t\dot{S}^{\boldsymbol{\theta}_0}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) + o_p(t\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|). \tag{1.54}$$

We use a shorthand notation for the expected score at $\boldsymbol{\theta}$:

$$S^{\boldsymbol{\theta}}(\mathbf{h}) = \mathbb{E}\left[\int_0^{\tau'} K_1^{\boldsymbol{\theta}}(\mathbf{h})(u)dM^{\boldsymbol{\theta}}(u) + K_2^{\boldsymbol{\theta}}(\mathbf{h})dM^{\boldsymbol{\theta}}(\tau)\right] = \mathbb{E}\left[\int_0^{\tau} V^{\boldsymbol{\theta}}(\mathbf{h})(u)dM^{\boldsymbol{\theta}}(u)\right],$$

by setting

$$V^{\boldsymbol{\theta}}(\mathbf{h})(t) = I(t \le \tau')K_1^{\boldsymbol{\theta}}(\mathbf{h})(t) + I(t = \tau)K_2^{\boldsymbol{\theta}}(\mathbf{h}).$$

By the Lipschitz continuity with respect to $\|\boldsymbol{\theta}\|$ for all terms involved, $K_1^{\boldsymbol{\theta}}(\mathbf{h})$, $K_2^{\boldsymbol{\theta}}(\mathbf{h})$ and $dM^{\boldsymbol{\theta}}$,

$$\begin{aligned} & S^{\boldsymbol{\theta}_t}(\mathbf{h}) - S^{\boldsymbol{\theta}}(\mathbf{h}) \\ = \;& \mathbb{E}\left[\int_0^{\tau'} V^{\boldsymbol{\theta}_t}(\mathbf{h})(u)dM^{\boldsymbol{\theta}_t}(u)\right] \end{aligned}$$

$$
\begin{aligned}
&= \mathbb{E}\left[\int_0^{\tau'} V^{\boldsymbol{\theta}_0}(\mathbf{h})(u)d\{M^{\boldsymbol{\theta}_t}(u)-M^{\boldsymbol{\theta}_0}(u)\}\right]+\mathbb{E}\left[\int_0^{\tau'} V^{\boldsymbol{\theta}_t}(\mathbf{h})(u)dM^{\boldsymbol{\theta}_0}(u)\right] \\
&\quad +\mathbb{E}\left[\int_0^{\tau'}\{V^{\boldsymbol{\theta}_t}(\mathbf{h})(u)-V^{\boldsymbol{\theta}_0}(\mathbf{h})(u)\}d\{M^{\boldsymbol{\theta}_t}(u)-M^{\boldsymbol{\theta}_0}(u)\}\right] \\
&= t\dot{S}^{\boldsymbol{\theta}_0}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)(\mathbf{h})+0+O_p(t^2\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|^2).
\end{aligned}
$$

Again, we plug-in $\widehat{\boldsymbol{\theta}}$ and use the consistency result to verify the condition (1.54).

Afterwards, we find the local inverse of the functional Hessian in (1.50). We have shown in Lemma 3 that the functional operator σ is a continuously invertible bijection from H_∞ to H_∞. The invertibility of $\dot{S}^{\boldsymbol{\theta}_0}$ in H_p follows from the following argument. By the continuous invertibility of σ, there is some q so that $\sigma^{-1}(H_q)\subseteq H_p$, and

$$
\begin{aligned}
&\inf_{\triangle\boldsymbol{\theta}\in lin\Theta}\frac{\sup_{\mathbf{h}\in H_p}|(\boldsymbol{\alpha}-\boldsymbol{\alpha}_0)^\top\boldsymbol{\sigma}_a(\mathbf{h})+(\boldsymbol{\beta}-\boldsymbol{\beta}_0)^\top\boldsymbol{\sigma}_b(\mathbf{h})+\int_0^{\tau'}\sigma_\eta(\mathbf{h})d(\Lambda-\Lambda_0)|}{\|\triangle\boldsymbol{\theta}\|_{l^\infty(H_p)}} \\
\geq &\inf_{\triangle\boldsymbol{\theta}\in lin\Theta}\frac{\sup_{\mathbf{h}\in\sigma^{-1}(H_q)}|(\boldsymbol{\alpha}-\boldsymbol{\alpha}_0)^\top\boldsymbol{\sigma}_a(\mathbf{h})+(\boldsymbol{\beta}-\boldsymbol{\beta}_0)^\top\boldsymbol{\sigma}_b(\mathbf{h})+\int_0^{\tau'}\sigma_\eta(\mathbf{h})d(\Lambda-\Lambda_0)|}{p\|\triangle\boldsymbol{\theta}\|} \\
= &\inf_{\triangle\boldsymbol{\theta}\in lin\Theta}\frac{\sup_{\mathbf{h}\in H_q}|\triangle\boldsymbol{\theta}(\mathbf{h})|}{p\|\triangle\boldsymbol{\theta}\|}>\frac{q}{2p}.
\end{aligned}
\tag{1.55}
$$

Finally, let us put everything together. The NPMLE $\widehat{\boldsymbol{\theta}}$ is shown to be consistent in Theorem 3', and (1.52), (1.53), (1.54) and (1.55) verify the conditions of Theorem 3.3.1 from [VdVW96]. □

Proof of Theorem 5. The proof for the continuous invertibility of $\widehat{\sigma}$ is similar to the proof of Lemma 3. The approximation error between the natural estimator $\widehat{\sigma}$ and Louis' formula variance estimator using (1.14) again comes from the "ghost copies" like the case in Lemma 1, so the same argument applies to show their asymptotic equivalence. □

1.8 Details on Variance Estimator

1.8.1 Derivatives of Log-likelihood

Let $l^c(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\lambda}) = \sum_{i=1}^n l_i^c(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\lambda})$ be the complete data log-likelihood,

$$\begin{aligned} l_i^c(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\lambda}) =& (A_i+M_i)\boldsymbol{\alpha}^\top \mathbf{Z_{1}}_i - (1+M_i)\log(1+e^{\boldsymbol{\alpha}^\top \mathbf{Z_{1}}_i}) \\ &+ \delta_i^1 A_i \sum_{k=1}^K I\{X_i = t_k\}(\log\lambda_k + \boldsymbol{\beta}^\top \mathbf{Z_{2}}_i) - A_i \sum_{k:t_k\le X_i} \lambda_k e^{\boldsymbol{\beta}^\top \mathbf{Z_{2}}_i} \\ &+ M_i \sum_{k:t_k<Q_i} I\{\kappa_i = k\}\Big(\log\lambda_k + \boldsymbol{\beta}^\top \mathbf{Z_{2}}_i - \sum_{h=1}^k \lambda_h e^{\boldsymbol{\beta}^\top \mathbf{Z_{2}}_i}\Big). \end{aligned}$$

Its gradient is given by

$$\nabla l_i^c = \left(\frac{\partial l_i^c}{\partial\boldsymbol{\alpha}}, \frac{\partial l_i^c}{\partial\boldsymbol{\beta}}, \frac{\partial l_i^c}{\partial\boldsymbol{\lambda}}\right)^\top,$$

where

$$\begin{aligned} \frac{\partial l_i^c}{\partial\boldsymbol{\alpha}} =& \mathbf{Z_{1}}_i\Big\{A_i + M_i - (1+M_i)\frac{e^{\boldsymbol{\alpha}^\top \mathbf{Z_{1}}_i}}{1+e^{\boldsymbol{\alpha}^\top \mathbf{Z_{1}}_i}}\Big\} = \mathbf{Z_{1}}_i\big\{A_i - p_i + M_i(1-p_i)\big\}, \\ \frac{\partial l_i^c}{\partial\boldsymbol{\beta}} =& \mathbf{Z_{2}}_i\Big\{A_i\delta_i^1 + M_i - \Big(A_i\sum_{k:t_k\le X_i}\lambda_k + M_i\sum_{k=1}^{\kappa_i}\lambda_k\Big)e^{\boldsymbol{\beta}^\top \mathbf{Z_{2}}_i}\Big\} \\ =& \mathbf{Z_{2}}_i\Big\{A_i\delta_i^1 + M_i - A_i\Lambda_i(X_i) - M_i\Lambda_i(\kappa_i)\Big\}, \\ \frac{\partial l_i^c}{\partial\lambda_k} =& \Big(A_i\delta_i^1 I\{X_i = t_k\} + M_i I\{\kappa_i = k\}\Big)\frac{1}{\lambda_k} - \Big(A_i I\{t_k\le X_i\} + M_i I\{\kappa_i\ge t_k\}\Big)e^{\boldsymbol{\beta}^\top \mathbf{Z_{2}}_i} \\ =& A_i\Big(\frac{\delta_i^1 I\{X_i = t_k\}}{\lambda_k} - I\{t_k\le X_i\}e^{\boldsymbol{\beta}^\top \mathbf{Z_{2}}_i}\Big) + M_i\Big(\frac{I\{\kappa_i = k\}}{\lambda_k} - I\{\kappa_i\ge t_k\}e^{\boldsymbol{\beta}^\top \mathbf{Z_{2}}_i}\Big). \end{aligned}$$

Its Hessian is given by

$$\nabla^2 l_i^c = \begin{pmatrix} \frac{\partial^2 l_i^c}{\partial\boldsymbol{\alpha}\partial\boldsymbol{\alpha}^\top} & 0 & 0 \\ 0 & \frac{\partial^2 l_i^c}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^\top} & \frac{\partial^2 l_i^c}{\partial\boldsymbol{\beta}\partial\boldsymbol{\lambda}^\top} \\ 0 & \left[\frac{\partial^2 l_i^c}{\partial\boldsymbol{\beta}\partial\boldsymbol{\lambda}^\top}\right]^\top & \mathrm{diag}(\frac{\partial^2 l_i^c}{\partial\lambda_k^2}) \end{pmatrix},$$

where

$$\begin{aligned}
\frac{\partial^2 l_i^c}{\partial\boldsymbol{\alpha}\partial\boldsymbol{\alpha}^\top} =& \mathbf{Z_1}_i\mathbf{Z_1}_i^\top\Big\{-(1+M_i)\frac{e^{\boldsymbol{\alpha}^\top\mathbf{Z_1}_i}}{(1+e^{\boldsymbol{\alpha}^\top\mathbf{Z_1}_i})^2}\Big\} = -\mathbf{Z_1}_i\mathbf{Z_1}_i^\top(1+M_i)p_i(1-p_i),\\
\frac{\partial^2 l_i^c}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^\top} =& \mathbf{Z_2}_i\mathbf{Z_2}_i^\top\Big\{-\Big(A_i\sum_{k:t_k\le X_i}\lambda_k + M_i\sum_{k=1}^{\kappa_i}\lambda_k\Big)e^{\boldsymbol{\beta}^\top\mathbf{Z_2}_i}\Big\},\\
\frac{\partial^2 l_i^c}{\partial\boldsymbol{\beta}\partial\lambda_k} =& \mathbf{Z_2}_i\Big\{-\Big(A_iI\{t_k\le X_i\} + M_iI\{t_k\le\kappa_i\}\Big)e^{\boldsymbol{\beta}^\top\mathbf{Z_2}_i}\Big\},\\
\frac{\partial^2 l_i^c}{\partial\lambda_k^2} =& -\Big(A_i\delta_i^1I\{X_i=t_k\} + M_iI\{\kappa_i=k\}\Big)\frac{1}{\lambda_k^2},\\
\frac{\partial^2 l_i^c}{\partial\boldsymbol{\alpha}\partial\boldsymbol{\beta}^\top} =& \frac{\partial^2 l_i^c}{\partial\boldsymbol{\alpha}\partial\boldsymbol{\lambda}^\top} = \frac{\partial^2 l_i^c}{\partial\lambda_k\partial\lambda_h} = 0, \quad k\ne h.
\end{aligned}$$

1.8.2 Conditional Expectations

By the conditional expectations (1.8) - (1.10), we are able to calculate the 'first order' conditional expectations, $\mathbb{E}[\nabla l_i^c|O]$ and $\mathbb{E}[\nabla^2 l_i^c|O]$:

$$\begin{aligned}
\mathbb{E}\left[\frac{\partial l_i^c}{\partial\boldsymbol{\alpha}}\right] =& \mathbf{Z_1}_i\Big\{\mathbb{E}(A_i) - p_i + \mathbb{E}(M_i)(1-p_i)\Big\},\\
\mathbb{E}\left[\frac{\partial l_i^c}{\partial\boldsymbol{\beta}}\right] =& \mathbf{Z_2}_i\Big[\mathbb{E}(A_i)\Big\{\delta_i^1 + \log S_i(X_i)\Big\} + \mathbb{E}(M_i)\Big\{1 + \sum_{k:t_k<Q_j}\mathbb{P}(\widetilde{T}_{ij}=t_k)\log S_i(t_k)\Big\}\Big],\\
\mathbb{E}\left[\frac{\partial l_i^c}{\partial\lambda_k}\right] =& \mathbb{E}(A_i)\Big\{\frac{\delta_i^1 I\{t_k=X_i\}}{\lambda_k} - I\{t_k\le X_i\}e^{\boldsymbol{\beta}^\top\mathbf{Z_2}_i}\Big\}\\
&+\mathbb{E}(M_i)\Big\{\frac{\mathbb{P}(\widetilde{T}_{ij}=t_k)}{\lambda_k} - \mathbb{P}(\widetilde{T}_{ij}\ge t_k)e^{\boldsymbol{\beta}^\top\mathbf{Z_2}_i}\Big\}.
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}\left[\frac{\partial^2 l_i^c}{\partial\boldsymbol{\alpha}\partial\boldsymbol{\alpha}^\top}\right] =& -\mathbf{Z_1}_i\mathbf{Z_1}_i^\top(1+\mathbb{E}(M_i))p_i(1-p_i),\\
\mathbb{E}\left[\frac{\partial^2 l_i^c}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^\top}\right] =& \mathbf{Z_2}_i\mathbf{Z_2}_i^\top\Big\{\mathbb{E}(A_i)\log S_i(X_i) + \mathbb{E}(M_i)\sum_{k:t_k<Q_i}\mathbb{P}(\widetilde{T}_{ij}=t_k)\log S_i(t_k)\Big\},\\
\mathbb{E}\left[\frac{\partial^2 l_i^c}{\partial\boldsymbol{\beta}\partial\lambda_k}\right] =& -\mathbf{Z_2}_i\Big\{\mathbb{E}(A_i)I\{t_k\le X_i\} + \mathbb{E}(M_i)\mathbb{P}(t_k\le\kappa_i)\Big\}e^{\boldsymbol{\beta}^\top\mathbf{Z_2}_i},
\end{aligned}$$

$$\mathbb{E}\left[\frac{\partial^2 l_i^c}{\partial \lambda_k^2}\right] = -\Big\{\mathbb{E}(A_i)\delta_i^1 I\{\widetilde{T}_{ij}=t_k\} + \mathbb{E}(M_i)\mathbb{P}(\widetilde{T}_{ij}=t_k)\Big\}\frac{1}{\lambda_k^2}.$$

To calculate 'second order' expectation $\mathbb{E}[\nabla l_i^c \nabla l_i^{c\top}|O]$, we first compute the conditional variances:

$$\mathrm{Var}[A_i|O] = \delta_i^c \frac{p_i(1-p_i)S_i(X_i)}{\big\{1-p_i+p_iS_i(X_i)\big\}^2},$$
$$\mathrm{Var}[M_i|O] = \frac{p_i\Big[1-S_i(Q_i)\big\}}{\big\{1-p_i+p_iS_i(Q_i)\big\}^2}.$$

Then,

$$\begin{aligned}
\mathbb{E}\left[\frac{\partial l_i^c}{\partial \boldsymbol{\alpha}}\frac{\partial l_i^c}{\partial \boldsymbol{\alpha}}^\top\right] =& \mathbb{E}\left[\frac{\partial l_i^c}{\partial \boldsymbol{\alpha}}\right]\mathbb{E}\left[\frac{\partial l_i^c}{\partial \boldsymbol{\alpha}}\right]^\top + \mathbf{Z_1}_i\mathbf{Z_1}_i^\top\big\{(1-p_i)^2\,\mathrm{Var}(M_i)+\mathrm{Var}(A_i)\big\},\\
\mathbb{E}\left[\frac{\partial l_i^c}{\partial \boldsymbol{\alpha}}\frac{\partial l_i^c}{\partial \boldsymbol{\beta}}^\top\right] =& \mathbb{E}\left[\frac{\partial l_i^c}{\partial \boldsymbol{\alpha}}\right]\mathbb{E}\left[\frac{\partial l_i^c}{\partial \boldsymbol{\beta}}\right]^\top + \mathbf{Z_1}_i\mathbf{Z_2}_i^\top\Bigg[\mathrm{Var}(A_i)\big\{\delta_i^1+\log S_i(X_i)\big\}\\
&+\mathrm{Var}(M_i)(1-p_i)\Big\{1+\sum_{k:t_k<Q_i}\mathbb{P}(\widetilde{T}_{ij}=t_k)\log S_i(t_k)\Big\}\Bigg],\\
\mathbb{E}\left[\frac{\partial l_i^c}{\partial \boldsymbol{\beta}}\frac{\partial l_i^c}{\partial \boldsymbol{\beta}}^\top\right] =& \mathbb{E}\left[\frac{\partial l_i^c}{\partial \boldsymbol{\beta}}\right]\mathbb{E}\left[\frac{\partial l_i^c}{\partial \boldsymbol{\beta}}\right]^\top + \mathbf{Z_2}_i\mathbf{Z_2}_i^\top\Bigg[\mathrm{Var}(A_i)\Big\{\delta_i^1+\log S_i(X_i)\Big\}^2\\
&+\mathrm{Var}(M_i)\Big\{1+\sum_{k:t_k<Q_i}\mathbb{P}(\widetilde{T}_{ij}=t_k)\log S_i(t_k)\Big\}^2\\
&+\mathbb{E}(M_i)\Big\{\sum_{k:t_k<Q_i}\mathbb{P}(\widetilde{T}_{ij}=t_k)\log S_i(t_k)^2-\Big(\sum_{k:t_k<Q_i}\mathbb{P}(\widetilde{T}_{ij}=t_k)\log S_i(t_k)\Big)^2\Big\}\Bigg],\\
\mathbb{E}\left[\frac{\partial l_i^c}{\partial \boldsymbol{\alpha}}\frac{\partial l_i^c}{\partial \lambda_k}\right] =& \mathbb{E}\left[\frac{\partial l_i^c}{\partial \boldsymbol{\alpha}}\right]\mathbb{E}\left[\frac{\partial l_i^c}{\partial \lambda_k}\right] + \mathbf{Z_1}_i\Bigg[\mathrm{Var}(A_i)\Big\{\frac{\delta_i^1 I\{t_k=X_i\}}{\lambda_k} - I\{t_k\le X_i\}e^{\boldsymbol{\beta}^\top\mathbf{Z_2}_i}\Big\}\\
&+\mathrm{Var}(M_i)(1-p_i)\Big\{\frac{\mathbb{P}(\widetilde{T}_{ij}=t_k)}{\lambda_k}-\mathbb{P}(\widetilde{T}_{ij}\ge t_k)e^{\boldsymbol{\beta}^\top\mathbf{Z_2}_i}\Big\}\Bigg],\\
\mathbb{E}\left[\frac{\partial l_i^c}{\partial \boldsymbol{\beta}}\frac{\partial l_i^c}{\partial \lambda_k}\right] =& \mathbb{E}\left[\frac{\partial l_i^c}{\partial \boldsymbol{\beta}}\right]\mathbb{E}\left[\frac{\partial l_i^c}{\partial \lambda_k}\right]\\
&+\mathbf{Z_2}_i\Bigg[\mathrm{Var}(A_i)\big\{\delta_i^1+\log S_i(X_i)\big\}\Big\{\frac{\delta_i^1 I\{t_k=X_i\}}{\lambda_k}-I\{t_k\le X_i\}e^{\boldsymbol{\beta}^\top\mathbf{Z_2}_i}\Big\}\\
&+\mathrm{Var}(M_i)\Big\{\frac{\mathbb{P}(\widetilde{T}_{ij}=t_k)}{\lambda_k}-\mathbb{P}(\widetilde{T}_{ij}\ge t_k)e^{\boldsymbol{\beta}^\top\mathbf{Z_2}_i}\Big\}\Big\{1+\sum_{h:t_h<Q_i}\mathbb{P}(\widetilde{T}_{ij}=t_h)\log S_i(t_h)\Big\}
\end{aligned}$$

$$
\begin{aligned}
&- \mathbb{E}(M_i)\Big\{ \sum_{h:t_h<Q_i} \mathbb{P}(\widetilde{T}_{ij}=t_h)\log S_i(t_h)\frac{\mathbb{P}(\widetilde{T}_{ij}=t_k)}{\lambda_k} - \frac{\mathbb{P}(\widetilde{T}_{ij}=t_k)\log S_i(t_k)}{\lambda_k} \\
&- \mathbb{P}\{\widetilde{T}_{ij}\geq t_k\}e^{\boldsymbol{\beta}^\top \mathbf{Z_{2}}_i} \sum_{h:t_h<Q_i} \mathbb{P}(\widetilde{T}_{ij}=t_h)\log S_i(t_h) \\
&+ e^{\boldsymbol{\beta}^\top \mathbf{Z_{2}}_i} \sum_{h=k}^{t_h<Q_i} \mathbb{P}(\widetilde{T}_{ij}=t_h)\log S_i(t_h)\Big\}\Big],
\end{aligned}
$$

$$
\begin{aligned}
&\mathbb{E}\left[\frac{\partial l_i^c}{\partial \lambda_k}\frac{\partial l_i^c}{\partial \lambda_h}\right] = \mathbb{E}A_i\left\{-\frac{\delta_i^1 I\{X_i=t_{k\vee h}\}}{\lambda_{k\vee h}}e^{\boldsymbol{\beta}^\top \mathbf{Z_{2}}_i} + I\{X_i\geq t_{k\vee h}\}e^{2\boldsymbol{\beta}^\top \mathbf{Z_{2}}_i}\right\} \\
&+ \mathbb{E}(A_i)\mathbb{E}(M_i)\left\{\frac{\delta_i^1 I\{\widetilde{T}_{ij}=t_k\}}{\lambda_k} - I\{X_i\geq t_k\}e^{\boldsymbol{\beta}^\top \mathbf{Z_{2}}_i}\right\}\left\{\frac{\mathbb{P}(\widetilde{T}_{ij}=t_h)}{\lambda_h} - \mathbb{P}(\widetilde{T}_{ij}\geq t_h)e^{\boldsymbol{\beta}^\top \mathbf{Z_{2}}_i}\right\} \\
&+ \mathbb{E}(A_i)\mathbb{E}(M_i)\left\{\frac{\delta_i^1 I\{X_i=t_h\}}{\lambda_h} - I\{X_i\geq t_h\}e^{\boldsymbol{\beta}^\top \mathbf{Z_{2}}_i}\right\}\left\{\frac{\mathbb{P}(\widetilde{T}_{ij}=t_k)}{\lambda_k} - \mathbb{P}(\widetilde{T}_{ij}\geq t_k)e^{\boldsymbol{\beta}^\top \mathbf{Z_{2}}_i}\right\} \\
&+ \mathbb{E}[M_i^2 - M_i]\left\{\frac{\mathbb{P}(\widetilde{T}_{ij}=t_k)}{\lambda_k} - \mathbb{P}(\widetilde{T}_{ij}\geq t_k)e^{\boldsymbol{\beta}^\top \mathbf{Z_{2}}_i}\right\}\left\{\frac{\mathbb{P}(\widetilde{T}_{ij}=t_h)}{\lambda_h} - \mathbb{P}(\widetilde{T}_{ij}\geq t_h)e^{\boldsymbol{\beta}^\top \mathbf{Z_{2}}_i}\right\} \\
&+ \mathbb{E}(M_i)\left\{-\frac{\mathbb{P}(\widetilde{T}_{ij}=t_{k\vee h})}{\lambda_{k\vee h}}e^{\boldsymbol{\beta}^\top \mathbf{Z_{2}}_i} + \mathbb{P}(\kappa_i\geq t_{k\vee h})e^{2\boldsymbol{\beta}^\top \mathbf{Z_{2}}_i}\right\}, \\
&\mathbb{E}\left[\frac{\partial l_i^c}{\partial \lambda_k}\frac{\partial l_i^c}{\partial \lambda_k}\right] = \mathbb{E}A_i\left\{\frac{\delta_i^1 I\{\widetilde{T}_{ij}=t_k\}}{\lambda_k} - I\{X_i\geq t_k\}e^{\boldsymbol{\beta}^\top \mathbf{Z_{2}}_i}\right\}^2 \\
&+ \mathbb{E}(A_i)\mathbb{E}(M_i)\left\{\frac{\delta_i^1 I\{\widetilde{T}_{ij}=t_k\}}{\lambda_k} - I\{X_i\geq t_k\}e^{\boldsymbol{\beta}^\top \mathbf{Z_{2}}_i}\right\}\left\{\frac{\mathbb{P}(\widetilde{T}_{ij}=t_k)}{\lambda_k} - \mathbb{P}(\widetilde{T}_{ij}\geq t_k)e^{\boldsymbol{\beta}^\top \mathbf{Z_{2}}_i}\right\} \\
&+ \mathbb{E}(A_i)\mathbb{E}(M_i)\left\{\frac{\delta_i^1 I\{\widetilde{T}_{ij}=t_k\}}{\lambda_k} - I\{X_i\geq t_k\}e^{\boldsymbol{\beta}^\top \mathbf{Z_{2}}_i}\right\}\left\{\frac{\mathbb{P}(\widetilde{T}_{ij}=t_k)}{\lambda_k} - \mathbb{P}(\widetilde{T}_{ij}\geq t_k)e^{\boldsymbol{\beta}^\top \mathbf{Z_{2}}_i}\right\} \\
&+ \mathbb{E}[M_i^2 - M_i]\left\{\frac{\mathbb{P}(\widetilde{T}_{ij}=t_k)}{\lambda_k} - \mathbb{P}(\widetilde{T}_{ij}\geq t_k)e^{\boldsymbol{\beta}^\top \mathbf{Z_{2}}_i}\right\}\left\{\frac{\mathbb{P}(\widetilde{T}_{ij}=t_k)}{\lambda_k} - \mathbb{P}(\widetilde{T}_{ij}\geq t_k)e^{\boldsymbol{\beta}^\top \mathbf{Z_{2}}_i}\right\} \\
&+ \mathbb{E}(M_i)\left\{\frac{\mathbb{P}(\widetilde{T}_{ij}=t_k)}{\lambda_k^2} - 2\frac{\mathbb{P}(\widetilde{T}_{ij}=t_k)}{\lambda_k}e^{\boldsymbol{\beta}^\top \mathbf{Z_{2}}_i} + \mathbb{P}(\widetilde{T}_{ij}\geq t_k)e^{2\boldsymbol{\beta}^\top \mathbf{Z_{2}}_i}\right\}.
\end{aligned}
$$

1.9 Acknowledgement

Chapter 1, in full, is a reprint of the material as it appears in Lifetime Data Analysis. Hou, Jue; Chambers, Christina; Xu, Ronghui. A nonparametric maximum likelihood approach for survival data with observed cured subjects, left truncation and right-censoring, LIDA, **24**(4) 612-651, 2018. The dissertation/thesis author was the primary investigator and author of this paper.

Table 1.4: Cure rate model versus separate model fits for SAB data.

	Cure model		Separate models	
	Estimate (SE)	P-value	Estimate (SE)	P-value
Logistic				
Intercept	-1.72 (0.28)	< 0.01	-2.54 (0.26)	< 0.01
Healthy	-0.60 (0.49)	0.22	-0.86 (0.45)	0.06
Diseased Control	0.15 (0.30)	0.63	0.01 (0.28)	0.98
Maternal Age ≥ 34	0.59 (0.28)	0.04	0.65 (0.26)	0.01
BMI $>$ normal	-0.62 (0.29)	0.04	-0.32 (0.28)	0.24
Smoking	0.51 (0.38)	0.17	0.80 (0.35)	0.02
Alcohol	-0.23 (0.29)	0.43	-0.34 (0.27)	0.20
Cox PH				
Healthy	-0.49 (0.44)	0.27	-0.35 (0.45)	0.43
Diseased Control	-0.30 (0.27)	0.26	0.29 (0.27)	0.28
Maternal Age ≥ 34	-0.04 (0.26)	0.86	0.55 (0.25)	0.03
BMI $>$ normal	-0.71 (0.29)	0.01	-0.39 (0.26)	0.14
Smoking	-1.18 (0.41)	< 0.01	0.78 (0.33)	0.02
Alcohol	0.78 (0.28)	0.01	-0.43 (0.26)	0.10

Chapter 2

Inference under Fine-Gray Competing Risks Model with High-Dimensional Covariates

2.1 Introduction

High-dimensional regression has attracted increasing interest in statistical analysis and has provided a useful tool in modern biomedical, ecological, astrophysical or economics data pertaining to the setting where the number of parameters is greater than the number of samples (see [BvdG11] for an overview). Regularized methods [Tib96, FL01] provide straightforward interpretation of resulting estimators while allowing the number of covariates to be exponentially larger than the sample size. While they can be consistent for prediction (i.e. estimating the underlying regression function), confidence intervals cannot be consistently formulated, as

firm guarantees of correct variable selection can only be established under a restrictive set of assumptions, including but not limited to the assumption of the minimal signal strength of the true parameter [WR09, FL10, MY09], which cannot be verified in practice. For practical purposes, it is of interest to develop inferential tools, most commonly confidence intervals and p-values, that do not depend on such assumptions and are yet able to provide theoretical guarantees of the quality of estimation and/or testing; and this is the goal of our work here.

For the purposes of constructing confidence intervals or testing significance of the effect from certain covariates, relying on a naive regularized estimation alone is not appropriate; notably, construction of confidence intervals for those coefficients that have been shrunk to zero is impossible. [ZZ14] and [vdGBRD14] proposed the one-step bias-correction estimator, which can be subsequently used to carry out proper statistical inference. Our work here was motivated by an illustration project of how information contained in patients' electronic medical records can be harvested for precision medicine. The data set linking the Surveillance, Epidemiology and End Results (SEER) Program database of the National Cancer Institute with the federal health insurance program Medicare database contained prostate cancer patients of age 65 or older. A total of 57,011 patients diagnosed between 2004 and 2009 had information available on 7 relevant clinical variables (age, PSA, Gleason score, AJCC stage, and AJCC stage T, N, M, respectively), 5 demographical variables (race, marital status, metro, registry and year of diagnosis), plus 9321 binary insurance claim codes. Among these patients 1,247 died due to cancer, and 5,221 had deaths unrelated to cancer by December 2013. An important goal of the project was to evaluate the impact of risk factors (clinical, demographical, and claim codes) on the non-cancer versus cancer mortality, with appropriate statistical inference. Cancer and non-cancer versus cancer

mortality are known as competing risks in survival analysis, and cannot be handled using linear or generalized linear regression models as considered in [ZZ14] and [vdGBRD14]. Instead, we consider the proportional subdistribution hazards regression model, often known as the Fine-Gray model [FG99]. Under classical low-dimensional setting, Fine and Gray derived the estimation and inference for the model coefficients via the partial likelihood principle, and handled right censoring by inverse probability censoring weighting (IPCW).

Considerable research effort has been devoted to developing regularized methods to handle various regression settings [RWL10, BC11, OWJ11, MB06, BM15, CF15], including those for right-censored time-to-event data [SLFL14, BFJ11, GG12a, Joh08, Lem16, BS15, HMX06, among others]. However, regression has been little studied for the competing risks setting, with random censoring and high-dimensional covariates. The purpose of this paper has two folds: 1) to study estimators under the Fine-Gray regression model for competing risks data with many more covariates than the number of events; 2) to develop statistical inference procedures in this setting. To our best knowledge, no published work exists on statistical inference for competing risks data that allows high-dimensional models; univariate testing was studied in Cox proportional hazards model – however, our construction allows for the testing of general linear hypothesis.

There are at least three significant challenges for addressing high-dimensional competing risks regression under the Fine-Gray model. The structure of the score function related to the partial likelihood causes a somewhat subtle issue with many of the unobserved factors preventing a simple martingale representation. Additionally, the structure, as well as, size of the sample information matrix renders both methodology and theoretical analysis based on the Hessian matrix problematic. Thirdly, random censoring presents non-trivial challenges in the presence of

competing risks and weighting is needed which further complicates the theoretical analysis. Also, although bootstrap has been considered for inference under the Fine-Gray regression model, this approach is no longer applicable given the known problems of the bootstrap in high-dimensional settings. Development of high-dimensional inferential methods for competing risks data and under the Fine-Gray model, in particular, may have been hampered by these considerations.

In this paper, we propose a natural and sensible formulation of inferential procedure for this high-dimensional competing risks regression. We first study a regularized estimator of the high-dimensional parameter of interest and derive its finite-sample properties where the interplay between the sparsity, ambient dimension and the sample size can be directly seen. We then propose a bias-correction procedure by formulating a new pragmatic estimator of the inverse of a large covariance matrix that allows broad dependence structures within the Fine-Gray model. We find that the bias-corrected estimator is effective at capturing strong as well as weak signals, and can be used for statistical inference. This combination leads to an efficient and simple-to-implement procedure under the Fine-Gray model with many covariates.

2.1.1 Model and notation

For subject $i = 1, ..., n$ in a study, let T_i be the event time, with the event type or cause ε_i; we use the two words interchangeably in the following. Under the Fine-Gray model that we consider below, we assume without loss of generality that the event type of interest is '1', and we code all the other event types as '2' without further specification. In the presence of a potential right-censoring time C_i, the observed time is $X_i = T_i \wedge C_i$. We denote the event indicator as $\delta_i = I(T_i \leq C_i)$. The type of the event ε_i is observed, if the event occurs before the censoring

time, i.e., when $\delta_i = 1$. Let $\mathbf{Z}_i(\cdot)$ be the vector of covariates that are possibly time-dependent. We focus on the situation that the dimension of $\mathbf{Z}_i(\cdot)$, p, is larger than the sample size n. Assume that the observed data $\{(X_i, \delta_i, \delta_i\varepsilon_i, \mathbf{Z}_i(\cdot))\}$ are independent and identically distributed (i.i.d.) for $i = 1, \ldots, n$.

Since the cumulative incidence function (CIF) is often the quantity of interest, [FG99] proposed a proportional subdistribution hazards model where the CIF

$$F_1(t|\mathbf{Z}_i(\cdot)) = \Pr(T_i \leq t, \varepsilon_i = 1|\mathbf{Z}_i(\cdot)) = 1 - \exp\left(-\int_0^t e^{\boldsymbol{\beta}^{o\top}\mathbf{Z}_i(u)} h_0^1(u)du\right), \tag{2.1}$$

the p-dimensional coefficient $\boldsymbol{\beta}^o$ is the unknown parameter of interest, and $h_0^1(t)$ is the baseline subdistribution hazard. Under the model (2.1) corresponding subdistribution hazard $h_1(t|\mathbf{Z}_i(\cdot)) = h_0^1(t)e^{\boldsymbol{\beta}^{o\top}\mathbf{Z}_i(t)}$. Throughout the paper, we assume that there exists a sparsity factor $s_o = |\text{supp}(\boldsymbol{\beta}^o)|$ for some $s_o \leq n$. Note that if we define an improper random variable $T_i^1 = T_i I(\varepsilon_i = 1) + \infty I(\varepsilon_i > 1)$, then the subdistribution hazard can be seen as the conditional hazard of T_i^1 given $\mathbf{Z}_i(\cdot)$.

We denote the counting process for type 1 event as $N_i^1(t) = I(T_i^1 \leq t)$ and its observed counterpart as $N_i^o(t) = I(\delta_i\varepsilon_i = 1)I(X_i \leq t)$. We also denote the counting process for the censoring time as $N_i^c(t) = I(C_i \leq t)$. Let $Y_i(t) = 1 - N_i^1(t-)$ (note that this is not the 'at risk' indicator like under the classic Cox model), and $r_i(t) = I(C_i \geq T_i \wedge t)$. Note that $r_i(t)Y_i(t) = I(t \leq X_i) + I(t > X_i)I(\delta_i\varepsilon_i > 1)$ is always observable, even though $Y_i(t)$ or $r_i(t)$ may not. Let $G(t) = \Pr(C_i \geq t)$ and let $\widehat{G}(\cdot)$ be the Kaplan-Meier estimator for $G(\cdot)$. Here we assume that C is independent of T, ε and $\mathbf{Z}$. Following the notation of Fine and Gray we call the IPW at-risk process:

$$\omega_i(t)Y_i(t) = r_i(t)Y_i(t)\frac{\widehat{G}(t)}{\widehat{G}(t \wedge X_i)}; \tag{2.2}$$

in other words, the weight for subject i is one if $t < X_i$, zero after being censored or failure due to

cause 1, and $\widehat{G}(t)/\widehat{G}(X_i)$ after failure due to other causes. The log pseudo likelihood (as recently named in [BKRF18]) that gives rise to the weighted score function in [FG99] for $\boldsymbol{\beta}$ is

$$m(\boldsymbol{\beta}) = n^{-1}\sum_{i=1}^{n}\int_0^{t^*}\left\{\boldsymbol{\beta}^\top \mathbf{Z}_i(t) - \log\left(\sum_{j=1}^{n}\omega_j(t)Y_j(t)e^{\boldsymbol{\beta}^\top \mathbf{Z}_j(t)}\right)\right\}dN_i^o(t). \tag{2.3}$$

where $t^* < \infty$ is the study end time.

In the following, for a vector $\mathbf{v}$, let $\mathbf{v}^{\otimes 0} = 1$, $\mathbf{v}^{\otimes 1} = \mathbf{v}$ and $\mathbf{v}^{\otimes 2} = \mathbf{v}\mathbf{v}^\top$. We define for $l = 0, 1, 2$

$$\mathbf{s}^{(l)}(t,\boldsymbol{\beta}) = \mathbb{E}\left\{G(t)/G(t\wedge X_i)r_i(t)Y_i(t)e^{\boldsymbol{\beta}^\top \mathbf{Z}_i(t)}\mathbf{Z}_i(t)^{\otimes l}\right\}, \quad \boldsymbol{\mu}(t) = \mathbf{s}^{(1)}(t,\boldsymbol{\beta}^o)/s^{(0)}(t,\boldsymbol{\beta}^o),$$
$$\mathbf{S}^{(l)}(t,\boldsymbol{\beta}) = n^{-1}\sum_{i=1}^{n}\omega_i(t)Y_i(t)e^{\boldsymbol{\beta}^\top \mathbf{Z}_i(t)}\mathbf{Z}_i(t)^{\otimes l}, \quad \bar{\mathbf{Z}}(t,\boldsymbol{\beta}) = \mathbf{S}^{(1)}(t,\boldsymbol{\beta})/S^{(0)}(t,\boldsymbol{\beta}). \tag{2.4}$$

We then have the score function, i.e. derivative of the log pseudo likelihood (2.3),

$$\dot{\mathbf{m}}(\boldsymbol{\beta}) = n^{-1}\sum_{i=1}^{n}\int_0^{t^*}\{\mathbf{Z}_i(t) - \bar{\mathbf{Z}}(t,\boldsymbol{\beta})\}dN_i^o(t).$$

Regarding notation, let us mention that all constants are assumed to be independent of n, p and s_o. We use K and ρ with corresponding enumerated subscripts to denote "big" and "small" constants. We use Q to denote intermediate terms used in the statements or the proofs of various results. We label the subscripts by the corresponding order of their appearance.

2.1.2 Organization of the paper

This paper is organized as follows. In Section 2.2, we provide the bias corrected estimator, Section 2.2.1, as well as the confidence interval construction, Section 2.2.2, for the high-dimensional Fine-Gray model. Construction of a new Hessian estimator, the cornerstone for our bias correction and variance estimation, is presented in Section 2.2.3. Section 2.3 presents

properties of the developed estimator. Additional notations for theoretical considerations are presented in Section 2.3.1. Bounds for the prediction error are presented in Section 2.3.2; Theorem 6 is the main result on estimation. Section 2.3.3 studies the sampling distribution of a newly develop test statistics while not requiring model selection consistency or minimal signal strength. Theorem 7 is the main result regarding asymptotic distribution. There we present a sequence of intermediate results as well. We examine our regularized estimator and the one-step bias-corrected estimator through simulation experiments in Section 2.4, and apply them to the SEER-Medicare data in Section 2.5.

2.2 Estimation and inference for competing risks with more regressors than events

2.2.1 One-step corrected estimator

A natural initial estimator to consider under the high dimensional setting is a l_1-regularized estimator, where the particular loss function of interest would be the log pseudo likelihood as defined in (2.3). We note that our results are easily generalizable to any sparsity-inducing and convex penalty functions, but due to the simplicity of presentation we present details only for the l_1 regularization. That is, we consider

$$\widehat{\boldsymbol{\beta}}(\lambda) = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}}\Big\{-m(\boldsymbol{\beta}) + \lambda\|\boldsymbol{\beta}\|_1\Big\} \tag{2.5}$$

for a suitable choice of the tuning parameter $\lambda > 0$. Whenever possible, we suppress λ in the notation above and use $\widehat{\boldsymbol{\beta}}$ to denote the l_1-regularized estimator. In the Section 2.3.2, we quantify

the non-asymptotic oracle risk bound and show that the estimator above, as a typical regularized estimator with $p \gg n$, converges at a rate slower than root-n. This implies that for inferential purposes the bias of the estimator cannot be ignored.

Inspired by the work of [ZZ14] and [vdGBRD14], we propose the one-step bias-correction estimator

$$\widehat{\mathbf{b}} := \widehat{\boldsymbol{\beta}} + \widehat{\boldsymbol{\Theta}}\dot{\mathbf{m}}(\widehat{\boldsymbol{\beta}}), \tag{2.6}$$

where $\widehat{\boldsymbol{\beta}}$ is defined in (2.5), $\widehat{\boldsymbol{\Theta}}$ is an estimator of the "asymptotic" precision matrix $\boldsymbol{\Theta}$ to be defined later. The above construction of the one-step estimator is inspired by the first order Taylor expansion of $\dot{\mathbf{m}}(\cdot)$,

$$\begin{aligned}\dot{\mathbf{m}}(\boldsymbol{\beta}^o) &\approx \dot{\mathbf{m}}(\widehat{\boldsymbol{\beta}}) - \ddot{\mathbf{m}}(\boldsymbol{\beta}^o)(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^o) \\ &\approx \ddot{\mathbf{m}}(\boldsymbol{\beta}^o)\left[\boldsymbol{\beta}^o - \{\widehat{\boldsymbol{\beta}} + \widehat{\boldsymbol{\Theta}}\dot{\mathbf{m}}(\widehat{\boldsymbol{\beta}})\}\right] = \ddot{\mathbf{m}}(\boldsymbol{\beta}^o)\{\boldsymbol{\beta}^o - \widehat{\mathbf{b}}\}.\end{aligned} \tag{2.7}$$

The notation "$\approx$" in the above indicates that the equivalence is approximate with the higher order error terms omitted. We aim to find a good candidate matrix $\widehat{\boldsymbol{\Theta}}$, such that $-\ddot{\mathbf{m}}(\boldsymbol{\beta}^o)\widehat{\boldsymbol{\Theta}} \approx \mathbb{I}_p$, with $\mathbb{I}_p$ denoting the $p \times p$ identity matrix. Note that when $p \leq n$ an inverse of the Hessian matrix above would naturally be a good candidate for $\widehat{\boldsymbol{\Theta}}$, but when $p \geq n$ such an inverse does not necessarily exist. We will elucidate the construction of $\widehat{\boldsymbol{\Theta}}$ towards the end of this section.

2.2.2 Confidence Intervals

To construct the confidence intervals for components of $\boldsymbol{\beta}^o$, we need the asymptotic distribution of $\widehat{\mathbf{b}}$. We will first establish the asymptotic distribution of the score $\dot{\mathbf{m}}(\boldsymbol{\beta}^o)$. With $p > n$, we have to restrict the space in which we want to establish the asymptotic distribution.

The asymptotic distribution for $\dot{\mathbf{m}}(\beta^o)$ is established in the following sense — for any $\mathbf{c} \in \mathbb{R}^p$ such that $\|\mathbf{c}\|_1 = 1$ we have

$$\sqrt{n}\mathbf{c}^\top \dot{\mathbf{m}}(\beta^o) \xrightarrow{d} N(0, \mathbf{c}^\top \mathcal{V}\mathbf{c}),$$

where $\mathcal{V}$ is the variance-covariance matrix for $\sqrt{n}\dot{\mathbf{m}}(\beta^o)$. We construct the following estimator for $\mathcal{V}$:

$$\widehat{\mathcal{V}} = n^{-1}\sum_{i=1}^{n}(\widehat{\boldsymbol{\eta}}_i + \widehat{\boldsymbol{\psi}}_i)^{\otimes 2}, \tag{2.8}$$

where $\widehat{\eta}_i$ and $\widehat{\psi}_i$ are defined as follows:

$$\widehat{\boldsymbol{\eta}}_i = \int_0^{t^*} \{\mathbf{Z}_i(t) - \bar{\mathbf{Z}}(t, \widehat{\boldsymbol{\beta}})\}\omega_i(t) d\widehat{M}_i^1(t), \tag{2.9}$$

$$\widehat{\boldsymbol{\psi}}_i = \int_0^{t^*} \frac{\widehat{\mathbf{q}}(t)}{\widehat{\pi}(t)} d\widehat{M}_i^c(t), \tag{2.10}$$

$$\widehat{\mathbf{q}}(t) = n^{-1}\sum_{i=1}^{n} I(t > X_i)\int_t^{t^*} \{\mathbf{Z}_i(u) - \bar{\mathbf{Z}}(u, \widehat{\boldsymbol{\beta}})\}\omega_i(u) d\widehat{M}_i^1(u), \tag{2.11}$$

$$\widehat{\pi}(t) = n^{-1}\sum_{i=1}^{n} I(X_i \geq t), \tag{2.12}$$

$$d\widehat{M}_i^1(t) = dN_i^o(t) - \frac{\omega_i(t)Y_i(t)e^{\widehat{\boldsymbol{\beta}}^\top \mathbf{Z}_i(t)}}{S^{(0)}(t, \widehat{\boldsymbol{\beta}})} n^{-1}\sum_{j=1}^{n} dN_j^o(t), \tag{2.13}$$

$$d\widehat{M}_i^c(t) = I(X_i \geq t)dN_i^c(t) - \frac{I(X_i \geq t)}{\widehat{\pi}(t)} n^{-1}\sum_{j=1}^{n} I(X_j \geq t)dN_j^c(t). \tag{2.14}$$

As illustrated in (2.7), we have $\sqrt{n}\mathbf{c}^\top(\widehat{\mathbf{b}} - \beta^o)$ to be asymptotically equivalent to

$$\sqrt{n}\mathbf{c}^\top \Theta \dot{\mathbf{m}}(\beta^o) \xrightarrow{d} N(0, \mathbf{c}^\top \Theta \mathcal{V} \Theta^\top \mathbf{c}).$$

We may now estimate the variance of $\sqrt{n}\mathbf{c}^\top(\widehat{\mathbf{b}} - \beta^o)$ using a "sandwich" estimator $\mathbf{c}^\top \widehat{\Theta}\widehat{\mathcal{V}}\widehat{\Theta}^\top \mathbf{c}$. Therefore a $(1-\alpha)100\%$ confidence interval for $\mathbf{c}^\top \beta^o$ is

$$\left[\mathbf{c}^\top \widehat{\mathbf{b}} - \mathcal{Z}_{1-\alpha/2}\sqrt{\mathbf{c}^\top \widehat{\Theta}\widehat{\mathcal{V}}\widehat{\Theta}^\top \mathbf{c}/n},\ \mathbf{c}^\top \widehat{\mathbf{b}} + \mathcal{Z}_{1-\alpha/2}\sqrt{\mathbf{c}^\top \widehat{\Theta}\widehat{\mathcal{V}}\widehat{\Theta}^\top \mathbf{c}/n}\right] \tag{2.15}$$

with standard normal quantile $\mathcal{Z}_{1-\alpha/2}$.

Our proposed approach addresses various practical questions as special cases. First, we can construct confidence interval for a chosen coordinate β_j^o in $\boldsymbol{\beta}^o$. To that end, one needs to consider $\mathbf{c} = \mathbf{e}_j$, the j-th natural basis for $\mathbb{R}^p$ and apply the result (2.15). Generally, we can construct a confidence interval for any linear contrasts $\mathbf{c}^\top \boldsymbol{\beta}^o$, potentially of any dimension. For example, we can have confidence intervals for the linear predictors $\mathbf{Z}^\top \boldsymbol{\beta}^o$ if the non-time-dependent covariate $\mathbf{Z}$ is also sparse so that we may assume $\|\mathbf{Z}\|_1$ to be bounded. As the dual problem, we may use the Wald test statistic

$$Z = \sqrt{n}(\mathbf{c}^\top \widehat{\mathbf{b}} - \theta_0) / \sqrt{\mathbf{c}^\top \widehat{\boldsymbol{\Theta}} \widehat{\mathcal{V}} \widehat{\boldsymbol{\Theta}}^\top \mathbf{c}} \tag{2.16}$$

to test the hypothesis with $H_0 : \mathbf{c}^\top \boldsymbol{\beta}^o = \theta_0$.

2.2.3 Construction of the inverse Hessian matrix

Although the early works under the linear model inspire the construction here, the specifics, as well as the theoretical analysis, the latter remains a challenge. We start by writing the negative Hessian of the log pseudolikelihood (2.3):

$$-\ddot{\mathbf{m}}(\boldsymbol{\beta}) = n^{-1} \sum_{i=1}^{n} \int_0^{t^*} \left\{ \frac{\mathbf{S}^{(2)}(t, \boldsymbol{\beta})}{S^{(0)}(t, \boldsymbol{\beta})} - \bar{\mathbf{Z}}(t, \boldsymbol{\beta})^{\otimes 2} \right\} dN_i^o(t). \tag{2.17}$$

We define

$$\boldsymbol{\Sigma} = \mathbb{E}\left[\int_0^{t^*} \{\mathbf{Z}_i(t) - \boldsymbol{\mu}(t)\}^{\otimes 2} dN_i^o(t) \right] = \mathbb{E}\left[\int_0^{t^*} \{\mathbf{Z}_i(t) - \boldsymbol{\mu}(t)\} dN_i^o(t) \right]^{\otimes 2}. \tag{2.18}$$

Under the regularity conditions, to be specified later, we have $\boldsymbol{\Sigma}$ as the "asymptotic negative Hessian" in the sense that the element-wise maximal norm $\|\boldsymbol{\Sigma} + \ddot{\mathbf{m}}(\boldsymbol{\beta}^o)\|_{\max}$ converges to zero in

probability. Our goal is to estimate its inverse $\boldsymbol{\Theta} = \boldsymbol{\Sigma}^{-1} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_p)^\top$, where $\boldsymbol{\theta}_j$'s are the rows of $\boldsymbol{\Theta}$.

By definition (2.18), the positive semi-definite matrix $\boldsymbol{\Sigma}$ is also the second moment of the random vector

$$\boldsymbol{\Xi}_i = \int_0^{t^*} \{\mathbf{Z}_i(t) - \boldsymbol{\mu}(t)\}\, dN_i^o(t) \tag{2.19}$$

with $\boldsymbol{\mu}(t)$ defined in (2.4). The expectation of $\boldsymbol{\Xi}_i$ is zero,

$$\mathbb{E}(\boldsymbol{\Xi}_i) = \mathbb{E}\left[\int_0^{t^*} \{\mathbf{Z}_i(t) - \boldsymbol{\mu}(t)\}\, Y_i(t) I(C_i \geq t) e^{\boldsymbol{\beta}^{t\top} \mathbf{Z}_i(o)} h_0^1(t) dt\right] = \mathbf{0}.$$

Hence, to estimate $\boldsymbol{\Theta}$, we may draw inspiration from the early work on inverting the high-dimensional variance-covariance matrix [ZRXB11]. Consider the minimizers of the expected loss functions

$$\boldsymbol{\gamma}_j^* = \underset{\boldsymbol{\gamma}_j \in \mathbb{R}^p}{\operatorname{argmin}} \mathbb{E}(\Xi_j - \boldsymbol{\Xi}_{-j}^\top \boldsymbol{\gamma}_j)^2, \quad \tau_j^2 = \mathbb{E}(\Xi_j - \boldsymbol{\Xi}_{-j}^\top \boldsymbol{\gamma}_j^*)^2, \tag{2.20}$$

where Ξ_j is the jth element of $\boldsymbol{\Xi}$, and $\boldsymbol{\Xi}_{-j}$ is a $p-1$ dimensional vector created by dropping the jth element from $\boldsymbol{\Xi}$. We show that the quantities $\boldsymbol{\gamma}_j^*$ and τ_j defined in (2.20) can be used to construct the inverse of $\boldsymbol{\Sigma}$. This is because τ_j^2 can also be alternatively written as

$$\mathbb{E}\{(\Xi_j - \boldsymbol{\Xi}_{-j}^\top \boldsymbol{\gamma}_j^*)\Xi_j\} - \boldsymbol{\gamma}_j^{*\top} \mathbb{E}\{(\Xi_j - \boldsymbol{\Xi}_{-j}^\top \boldsymbol{\gamma}_j^*)\boldsymbol{\Xi}_{-j}\}. \tag{2.21}$$

By the convexity of the target function $\mathbb{E}(\Xi_j - \boldsymbol{\Xi}_{-j}^\top \boldsymbol{\gamma}_j)^2$, $\boldsymbol{\gamma}_j^*$ must satisfy the first order Karush-Kuhn-Tucker conditions (KKT)

$$-\boldsymbol{\gamma}_j^{*\top} \mathbb{E}\{(\Xi_j - \boldsymbol{\Xi}_{-j}^\top \boldsymbol{\gamma}_j^*)\boldsymbol{\Xi}_{-j}\} = 0. \tag{2.22}$$

Applying (2.22) to (2.21), we have

$$\tau_j^2 = \mathbb{E}\{(\Xi_j - \boldsymbol{\Xi}_{-j}^\top \boldsymbol{\gamma}_j^*)\Xi_j\}.$$

We can then define a vector $\boldsymbol{\theta}_1 = (1, -\boldsymbol{\gamma}_1^{*\top})^\top / \tau_1^2$ that satisfies

$$\boldsymbol{\theta}_1^\top \boldsymbol{\Sigma} = \mathbb{E}\{(\Xi_1 - \boldsymbol{\Xi}_{-1}^\top \boldsymbol{\gamma}_1^{*\top})\boldsymbol{\Xi}\} / \mathbb{E}\{(\Xi_1 - \boldsymbol{\Xi}_{-1}^\top \boldsymbol{\gamma}_1^*)\Xi_1\} = (1, \mathbf{0}_{p-1}) = \mathbf{e}_1.$$

Without loss of generality, we may define $\boldsymbol{\theta}_j$ accordingly for $j = 2, \ldots, p$, satisfying $\boldsymbol{\theta}_j^\top \boldsymbol{\Sigma} = \mathbf{e}_j$. The matrix $\boldsymbol{\Theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_p)^\top$ satisfies

$$\boldsymbol{\Theta}\boldsymbol{\Sigma} = (\mathbf{e}_1, \ldots, \mathbf{e}_p) = \mathbb{I}_p,$$

therefore $\boldsymbol{\Theta}$ is the inverse of $\boldsymbol{\Sigma}$. We now utilize the sample form of $\boldsymbol{\Sigma}$, (2.18),

$$\widehat{\boldsymbol{\Sigma}} = n^{-1} \sum_{i=1}^n \int_0^{t^*} \{\mathbf{Z}_i(t) - \bar{\mathbf{Z}}(t, \widehat{\boldsymbol{\beta}})\}^{\otimes 2} dN_i^o(t). \tag{2.23}$$

In particular we observe that $\widehat{\boldsymbol{\Sigma}}$ is that it can be written as the sample second moment $\widehat{\boldsymbol{\Sigma}} = n^{-1} \sum_{i=1}^n \widehat{\boldsymbol{\Xi}}_i^{\otimes 2}$ where

$$\widehat{\boldsymbol{\Xi}}_i = \int_0^{t^*} \{\mathbf{Z}_i(t) - \bar{\mathbf{Z}}(t, \widehat{\boldsymbol{\beta}})\} dN_i^o(t). \tag{2.24}$$

This form allows us to define the inverse of $\boldsymbol{\Sigma}$ as a regression between the vectors $\widehat{\boldsymbol{\Xi}}_i$. For that purpose we define the least squares loss function as

$$\Gamma_j(\boldsymbol{\gamma}_j, \widehat{\boldsymbol{\beta}}) = n^{-1} \sum_{i=1}^n \left(\widehat{\Xi}_{i,j} - \widehat{\boldsymbol{\Xi}}_{i,-j}^\top \boldsymbol{\gamma}_j \right)^2, \ j = 1, \ldots, p, \tag{2.25}$$

where $\widehat{\Xi}_{i,j}$ is the jth element of $\widehat{\boldsymbol{\Xi}}_i$, and $\widehat{\boldsymbol{\Xi}}_{i,-j}$ is a $p-1$ dimensional vector obtained by dropping the jth element from $\widehat{\boldsymbol{\Xi}}_i$. We then define the nodewise LASSO in our context to be

$$\widehat{\boldsymbol{\gamma}}_j = \underset{\boldsymbol{\gamma}_j \in \mathbb{R}^{p-1}}{\operatorname{argmin}} \left\{ \Gamma_j(\boldsymbol{\gamma}_j, \widehat{\boldsymbol{\beta}}) + 2\lambda_j \|\boldsymbol{\gamma}_j\|_1 \right\}, \qquad \widehat{\tau}_j^2 = \Gamma_j(\widehat{\boldsymbol{\gamma}}_j, \widehat{\boldsymbol{\beta}}) + \lambda_j \|\widehat{\boldsymbol{\gamma}}_j\|_1. \tag{2.26}$$

Accordingly, we use $\widehat{\gamma}_j$ and $\widehat{\tau}_j^2$ to construct

$$\widehat{\Theta}_{jk} = \begin{cases} -\widehat{\gamma}_{j,k}/(\widehat{\tau}_j^2), & k < j; \\ 1/(\widehat{\tau}_j^2), & k = j; \\ \widehat{\gamma}_{j,k-1}/(\widehat{\tau}_j^2), & k > j. \end{cases} \tag{2.27}$$

By the first order KKT condition, we have $(\widehat{\Theta}\widehat{\Sigma})_{j,j} = 1$ and $|(\widehat{\Theta}\widehat{\Sigma})_{j,k}| \leq \lambda_j$ for $j \neq k$. Choosing $\lambda_{\max} = \max_{j=1,\ldots,p} \lambda_j = o_p(1)$, we achieve that $\|\widehat{\Theta}\widehat{\Sigma} - \mathbb{I}_p\|_{\max}$ goes to zero. The one-step estimator proposed in (2.6) with such $\widehat{\Theta}$ hence converges to the true coefficient β^o approximately at the rate equivalent to $\dot{\mathbf{m}}(\beta^o)$, as illustrated in (2.7).

Our proposed nodewise LASSO estimator is innovative in several aspects. Given the difficulty imposed by the model, we cannot make high-dimensional inference by simply inverting the $XX^\top$ for a design matrix X like in a linear or generalized linear model. The log pseudo likelihood (2.3) has dependent entries. The covariates $\mathbf{Z}_i(t)$ for $i = 1, \ldots, n$ are allowed to be time-dependent. Nevertheless, we identify for our model that the key element for the high-dimensional inference is each observation's contribution to the score, the Ξ_i's. Our solution generalizes high-dimensional matrix inversion in a non-trivial way to complex models with censoring, non-standard likelihoods and weighting.

2.3 Theoretical considerations

In this section, we present the theory for the estimators $\widehat{\beta}$, $\widehat{\mathbf{b}}$ and the confidence intervals described in the previous section. We will quantify the non-asymptotic oracle risk bound for the estimator above while allowing $p \gg n$ with a minimal set of assumptions. Theoretical study of this kind is novel, since in the context of competing risks, the martingale structures typically

utilized are unavailable and new techniques need to be developed. In particular, we show that the inverse probability weighting has a finite-sample effect that separates this model from the classical Cox model (see comments after Theorem 6). We will also establish that a certain tighter bound can be established whenever the hazard rate is bounded (Theorem 8).

Throughout our work we assume that $\{(T_i, C_i, \varepsilon_i, \mathbf{Z}_i(t)) : t \in [0,\infty)\}$ are i.i.d. with C_i independent of $(T_i, \varepsilon_i, \mathbf{Z}_i(\cdot))$. Moreover, for any $t \in [0, t^*]$, $G(t) = I(C_i \geq t)$ is differentiable, and its hazard function $h^c(t) = -G'(t)/G(t) \leq K_1$. We also assume that the baseline CIF $F_1(t;\mathbf{0})$ is differentiable. The baseline subdistribution hazard $h_0^1(t) = -d\log\{F_1(t;\mathbf{0})\}/dt \in [\rho_1, K_2]$ for all $t \in (0, t^*)$ and some $\rho_1 > 0$ and $K_2 < \infty$.

2.3.1 Additional notation

In the following, we introduce some additional notations. The counting process martingales

$$M_i^1(t) = N_i^1(t) - \int_0^t Y_i(u) e^{\boldsymbol{\beta}^{o\top} \mathbf{Z}_i(u)} h_0^1(u) du \tag{2.28}$$

are essentially helpful tools in high-dimensions for establishing theory with dependent partial likelihoods. Unfortunately, the uncensored counting processes $\{N_i^1(t), i = 1, \ldots, n\}$ are not always observable. The observable counterpart $N_i^o(t)$ has no known martingale related to it under the observed filtration $\mathcal{F}_t = \sigma\{N_i^o(u), I(X_i \geq u), r_i(u) : u \leq t, i = 1, \ldots, n\}$. The Doob-Meyer compensator for the submartingale $N_i^o(t)$ under the observed filtration involves the nuisance distribution of $T_i | \varepsilon_i > 1$. To utilize the martingale structure for our theory, we have to define the

"censoring complete" filtration

$$\mathcal{F}_t^* = \sigma\{N_i^o(u), I(C_i \geq u), \mathbf{Z}_i(\cdot) : u \leq t, i = 1, \ldots, n\}, \tag{2.29}$$

on which we have a martingale related to $N_i^o(t)$,

$$\int_0^t I(C_i \geq t) dM_i^1(u) = N_i^o(t) - \int_0^t I(C_i \geq u) Y_i(u) e^{\boldsymbol{\beta}^{o\top} \mathbf{Z}_i(u)} h_0^1(u) du. \tag{2.30}$$

To relate the martingale (2.30) with our log pseudo likelihood $m(\boldsymbol{\beta})$, we define its proxy with $\mathcal{F}_t^*$ measurable integrand

$$\widetilde{m}(\boldsymbol{\beta}) = n^{-1} \sum_{i=1}^n \int_0^{t^*} \boldsymbol{\beta}^\top \mathbf{Z}_i(t) - \log \left(\sum_{j=1}^n I(C_j \geq t) Y_j(t) e^{\boldsymbol{\beta}^\top \mathbf{Z}_j(t)} \right) dN_i^o(t). \tag{2.31}$$

We define processes related to $\widetilde{m}(\boldsymbol{\beta})$ and its derivatives as

$$\widetilde{\mathbf{S}}^{(l)}(t, \boldsymbol{\beta}) = n^{-1} \sum_{i=1}^n I(C_i \geq t) Y_i(t) e^{\boldsymbol{\beta}^\top \mathbf{Z}_i(t)} \mathbf{Z}_i(t)^{\otimes l}, \quad \widetilde{\mathbf{Z}}(t, \boldsymbol{\beta}) = \widetilde{\mathbf{S}}^{(1)}(t, \boldsymbol{\beta}) / \widetilde{S}^{(0)}(t, \boldsymbol{\beta}). \tag{2.32}$$

They can also be seen as proxies to the processes in (2.4). To see that, observe that by conditioning,

$$\begin{aligned} \mathbb{E}\left\{ \widetilde{\mathbf{S}}^{(l)}(t, \boldsymbol{\beta}) \right\} &= \mathbb{E}\left[\mathbb{E}\{ I(C_i \geq t) Y_i(t) | \mathcal{F}_t \} e^{\boldsymbol{\beta}^\top \mathbf{Z}_i(t)} \mathbf{Z}_i(t)^{\otimes l} \right] \\ &= \mathbb{E}\left\{ \widetilde{\omega}_i(t) Y_i(t) e^{\boldsymbol{\beta}^\top \mathbf{Z}_i(t)} \mathbf{Z}_i(t)^{\otimes 2} \right\}, \end{aligned}$$

where

$$\widetilde{\omega}_i(t) = r_i(t) G(t) / G(t \wedge X_i) \tag{2.33}$$

is the weight with the true censoring distribution $G(\cdot)$. We denote their expectations as

$$\mathbf{s}^{(l)}(t, \boldsymbol{\beta}) = \mathbb{E}\left\{ \widetilde{\mathbf{S}}^{(l)}(t, \boldsymbol{\beta}) \right\} = \mathbb{E}\left\{ \widetilde{\omega}_i(t) Y_i(t) e^{\boldsymbol{\beta}^\top \mathbf{Z}_i(t)} \mathbf{Z}_i(t)^{\otimes 2} \right\}. \tag{2.34}$$

Our proxies precisely target those weighted samples, as $\widetilde{\mathbf{S}}^{(l)}(t, \boldsymbol{\beta})$ differs from $\mathbf{S}^{(l)}(t, \boldsymbol{\beta})$ only at those summands with observed type-2 events.

Note that the Kaplan-Meier estimator for $G(t)$ can be written as

$$\widehat{G}(t) = \prod_{u \le t} \left(1 - \frac{dN_i^c(u)}{I(X_i \ge u)}\right).$$

To study the convergence of $\widehat{G}(t)$ to $G(t)$, we denote a martingale related to $N_i^c(t)$, the counting process of observed censoring, $M_i^c(t)$. Let the censoring hazard be defined as $h^c(t) = -d\log(G(t))/dt$. Under the "censoring" filtration

$$\mathcal{F}_t = \sigma\{N_i^c(u), T_i, \varepsilon_i, \mathbf{Z}_i(\cdot) : u \le t, i = 1, \ldots, n\}, \tag{2.35}$$

we have a martingale

$$M_i^c(t) = N_i^c(t) - \int_0^t I(C_i \ge u) h^c(u) du. \tag{2.36}$$

We use the integration-by-parts arguments [Mur94, the Helly-Bray argument on page 727] with random martingale measures, e.g. $dM_i^1(t)$, in our proof. The total variation of $M_i^1(t;w)$ is defined as

$$\bigvee_0^{t^*} M_i^1(t;w) = \sup_{k=1,2,\ldots} \sup_{0 \le t_1 < \cdots < t_k \le t^*} \sum_{j=2}^{n} |M_i^1(t_j;w) - M_i^1(t_{j-1};w)|. \tag{2.37}$$

Since $M_i^1(t;w)$ can be decomposed into a nondecreasing counting process $N_i^1(t)$ minus another nondecreasing compensator $\int_0^t Y_i(u) e^{\beta^{o\top} \mathbf{Z}_i(u)} h_0^1(u) du$, we have a bound for its total variation

$$\bigvee_0^{t^*} M_i^1(t;w) \le N_i^1(t^*) + \int_0^{t^*} Y_i(u) e^{\beta^{o\top} \mathbf{Z}_i(u)} h_0^1(u) du. \tag{2.38}$$

Similar conclusion also applies to $M_i^c(t)$, i.e. we have a bound for its total variation

$$\bigvee_0^{t^*} M_i^c(t;w) \le N_i^c(t^*) + \int_0^{t^*} I(C_i \ge t) h^c(u) du. \tag{2.39}$$

As a convention, from hereon we suppress the w in the notation to keep it simple.

2.3.2 Oracle inequality

We first establish oracle inequality for the initial estimation error $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^o\|_1$ based on the following set of conditions that are weaker than those in the next subsection.

(C1) (**Design**) With probability equal to one, the covariates satisfy

$$\sup_{i=1,\ldots,n} \sup_{t\in[0,t^*]} \|\mathbf{Z}_i(t)\|_\infty \leq K_3/2. \tag{2.40}$$

The expected at-risk process is bounded away from zero, i.e., for positive K_4 and ρ_2

$$\inf_{t\in[0,t^*]} \mathbb{E}\left[I(C_i \geq t^*) I(t^* < T_i^1 < \infty) \min\{K_4, e^{\boldsymbol{\beta}^{o\top}\mathbf{Z}_i(t)}\}\right] > \rho_2. \tag{2.41}$$

(C2) (**Covariance**) For K_4 in (2.41), the smallest eigenvalue of the matrix

$$\boldsymbol{\Sigma}(K_4) = \mathbb{E}\left\{\int_0^{t^*} \left(\mathbf{Z}(t) - \frac{\mathbb{E}\left[\mathbf{Z}(t)\{1-F_1(t;\mathbf{Z})\}\min\{K_4, e^{\boldsymbol{\beta}^{o\top}\mathbf{Z}(t)}\}\right]}{\mathbb{E}\left[\{1-F_1(t;\mathbf{Z})\}\min\{K_4, e^{\boldsymbol{\beta}^{o\top}\mathbf{Z}(t)}\}\right]}\right)^{\otimes 2} h_0^1(t)dt\right\}$$

is at least $\rho_3 > 0$.

(C3) (**Continuity**) $\mathbf{Z}_i(t)$ may have $K_{5,i}$ jumps at $t_{i,1} < t_{i,2} < \cdots < t_{i,K_{5,i}}$ with minimal gap between jumps bounded away from zero,

$$\min_{i=1,\ldots,n} \min_{1<k\leq K_{5,i}} t_{i,k} - t_{i,k+1} \geq \rho_4.$$

Between two consecutive jumps, $\mathbf{Z}_i(t)$ has at most K_6 elements Lipschitz continuous with Lipschitz constant K_7 while the rest of the elements are considered to be constant.

Remark 1. *Overall, the conditions above are minimal in the sense that they appear in results pertaining to the Cox model [HSY+13, see e.g, (3.9) on page 1149; (4.5) and Theorem 4.1 on page 1154].*

Remark 2. *We consider a finite interval* $[0,t^*]$. *Due to missing censoring times among those with observed type-2 events, we have to make the additional assumptions to control the weighting errors. Although the weighted at-risk processes* $\omega_i(t)$*'s are asymptotically unbiased, the approximation errors in the tail* $t \to \infty$ *are poor for any finite n. To avoid unnecessary complications, we set the* $[0,t^*]$ *such that we always have sufficient at-risk subjects; note that* (2.41) *implies that* $P(C > t^*) > 0$.

Remark 3. *We assume a finite maximal norm of* $\mathbf{Z}(t)$. *Condition* (2.40) *in (C1) is equivalent to the apparently weaker assumption (see for example [HSY+13] equation (3.9)):*

$$\sup_{1\le i<j\le n} \sup_{t\in[0,t^*]} \|\mathbf{Z}_i(t) - \mathbf{Z}_j(t)\|_\infty \le K_3. \tag{2.42}$$

This can be seen by noting that the Cox type partial likelihood for the proportional hazards model is invariant when subtracting $\mathbf{Z}_i(t)$ *by any deterministic* $\zeta(t)$.

Remark 4. *Condition (C1)* (2.41) *carries two facts. First, the at-risk rate for type 1 events is bounded away from zero. Second, relative-risks arbitrarily close to zero is truncated at a finite* K_4*; this is necessary in high-dimensions, in order to rule out the irregular cases where the non-zero expectation of the relative risk is dominated by a diminishing proportion of the excessively large relative risks. The same argument applies for (C2) in which a lower bound of the restricted eigenvalue of the negative Hessian [BRT09] is defined.*

Remark 5. *We assume the smoothness of the time-dependent covariates* $\mathbf{Z}(t)$. *Subjects with observed type 2 events, remain indefinitely in the risk sets for type 1 events. For time-dependent covariates, continuity is helpful in establishing a slow growing rate of the maximal relative risks*

among those subjects; something that is a fact for time independent covariates. Note that the coordinate wise continuity in $\mathbf{Z}_i(t)$ is insufficient as p grows to infinity. We propose (C3) taking into account likely practical scenarios, where the covariates are either constant, or change only at finitely many discrete time points.

Under the above assumptions, we are ready to present our estimation error result. Since the result holds in finite samples, we define a sequence of important constants first. For a $\varepsilon > 0$ and constants $K_1, \cdots, K_7$ as well as $\rho_1, \cdots, \rho_4$ (introduced in the conditions above)

$$Q_1(\varepsilon) = e^{K_6 K_7 \|\beta^o\|_\infty \rho_4} \log(n/\varepsilon)/\rho_4\rho_1, \tag{2.43}$$

$$Q_2^{(l)}(n,p,\varepsilon) = \frac{Q_1(\varepsilon)K_3^l}{2^l}\left\{\frac{4K_4^2(1+K_1t^*)}{\rho_2^2}\sqrt{\frac{4\log(2/\varepsilon)}{n}} + \frac{4K_4^2K_1t^*}{\rho_2^2 n} + \sqrt{\frac{2\log(2np^l/\varepsilon)}{n}} + \frac{1}{n}\right\}, \tag{2.44}$$

where $l = 0, 1$, and

$$Q_3(n,p,\varepsilon) = \left\{2Q_2^{(1)}(n,p,\varepsilon) + K_3 Q_2^{(0)}(n,p,\varepsilon)\right\}/\rho_2 + K_3\sqrt{2\log(2p/\varepsilon)/n}. \tag{2.45}$$

In high-dimensional models an additional constant, the so called compatibility factor, plays an important role. For a positive constant $\xi > 1$, the compatibility factor

$$\kappa\big(\xi, O; -\ddot{\mathbf{m}}(\beta^o)\big) = \sup_{0\neq\mathbf{b}\in\mathcal{C}(\xi,O)} \frac{\sqrt{s_o\mathbf{b}^\top\{-\ddot{\mathbf{m}}(\beta^o)\mathbf{b}\}}}{\|\mathbf{b}_O\|_1} \tag{2.46}$$

where $\mathcal{C}(\xi, O)$ denotes the cone set

$$\mathcal{C}(\xi, O) = \{\mathbf{b} \in \mathbb{R}^p : \|\mathbf{b}_{O_c}\|_1 \leq \xi\|\mathbf{b}_O\|_1\},$$

with O denoting the indices of non-zero elements β^o and O_c denoting its compliment.

Theorem 6. *For $\xi > 1$ and a $\varepsilon > 0$, let*

$$\lambda = Q_3(n, p, \varepsilon)(\xi - 1)/(\xi + 1)$$

with $Q_3(n, p, \varepsilon)$ defined in (2.45)*. When $n > -\log(\varepsilon/3)/(2\rho_2^2)$ with ρ_2 given in (C1), we have under regularity conditions (C1) and (C3) that*

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^o\|_1 < \frac{e^{\eta}(\xi + 1)s_o\lambda}{2Q_4^2}$$

occurs with probability no less than

$$\Pr\left(\kappa\left(\xi, O; -\ddot{\mathbf{m}}(\boldsymbol{\beta}^o)\right) > Q_4\right) - e^{-n\rho_2^2/(2K_4^2)} - ne^{-n(\rho_2 - 2/n)^2/(8K_4^2)} - 5\varepsilon,$$

where Q_4 is a positive constant satisfying

$$2K_3(\xi + 1)s_o\lambda/(2Q_4)^2 \leq 1/e$$

and η is the smaller solution of $\eta e^{-\eta} = 2K_3(\xi + 1)s_o\lambda/(2Q_4)^2$.

Our proof of Theorem 6 applies to the result with l_2-norm and general l_q-norm for $q \geq 1$. Namely, under the same conditions we have that

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^o\|_q < \frac{2e^{\eta}\xi s_o^{1/q}\lambda}{(\xi + 1)Q_4}$$

occurs with probability no less than

$$\Pr\left(F_q(\xi, O) > Q_4\right) - e^{-n\rho_2^2/(2K_4^2)} - ne^{-n(\rho_2 - 2/n)^2/(8K_4^2)} - 5\varepsilon,$$

with the weak cone invertibility condition defined as

$$F_q(\xi, O) = \sup_{0 \neq \mathbf{b} \in \mathcal{C}(\xi, O)} \frac{-s_o^{1/q}\mathbf{b}^{\top}\ddot{\mathbf{m}}(\boldsymbol{\beta}^o)\mathbf{b}}{\|\mathbf{b}_O\|_1\|\mathbf{b}\|_q}.$$

A few comments are in order. For a fixed ε, $Q_3(n,p,\varepsilon)$ is of order $\log(n)\sqrt{\log(p)/n}$. Thus, Theorem 6, together with Lemma 5 (see below), guarantee that for λ chosen to be of the order $\log(n)\sqrt{\log(p)/n}$

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^o\|_1 = O_p\left(s_o \log(n)\sqrt{\log(p)/n}\right).$$

The above estimation error rate to the error rate $\sqrt{\log(p)/n}$ of the simple Cox model [HSY$^+$13, YBS19], differing only by a factor of $\log(n)$. This factor is brought in by the error induced by the IPCW weights. Therefore, under the rate condition $s_o \log(n)\sqrt{\log(p)/n} = o(1)$, we obtain an asymptotically l_1-consistent regularized estimator $\widehat{\boldsymbol{\beta}}$.

The quantity $Q_1(\varepsilon)$ describes the error from IPCW weights through the measurable approximation to processes $\mathbf{S}^{(l)}$, $\mathbf{S}^{(l)}(t,\boldsymbol{\beta}^o) - \widetilde{\mathbf{S}}^{(l)}(t,\boldsymbol{\beta}^o)$. A naïve bound for the measurable approximation is proportional to the magnitude of the relative risks in $\mathbf{S}^{(l)}$, naturally of the order $e^{\|\boldsymbol{\beta}^o\|_1 K_3} \asymp e^{s_o}$, potentially growing in exponential rate of n if $s_o \asymp n^a$ for some $a > 0$. Such bound grows way too rapidly to deliver any meaningful result. Observing that the summands in $\mathbf{S}^{(l)}$ and $\widetilde{\mathbf{S}}^{(l)}$ at a particular index i differ from each other only when the i-th subject has type-2 event we are able to establish a significantly sharper bound. For that purpose, we develop ε-tail bound of the maximal relative risk among observed type 2 events (see Lemma 17). The quantity $Q_2^{(l)}(n,p,\varepsilon)$, involving $Q_1(\varepsilon)$ directly in the definition, gives the bound for the error from the measurable approximation to $\mathbf{S}^{(l)}$ (See in Section 2.7 Lemma 19).

For the rest of this section, we provide further details on the proof of Theorem 6, as well as the technical challenges involved. We highlight two results, Lemma 4 and 5. The first establishes properties of the score vector while the second one establishes the properties of the compatibility

factor (2.46).

Lemma 4. *Let $Q_3(n,p,\varepsilon)$ be defined as in* (2.45)*. Under Assumptions (C1) and (C3),*

$$\Pr\Big(\|\dot{\mathbf{m}}(\boldsymbol{\beta}^o)\|_\infty < Q_3(n,p,\varepsilon)\Big) \geq 1 - e^{-n\rho_2^2/(2K_4^2)} - ne^{-n(\rho_2-2/n)^2/(8K_4^2)} - 5\varepsilon.$$

Lemma 4 establishes that such event $\{\|\dot{\mathbf{m}}(\boldsymbol{\beta}^o)\|_\infty < \lambda(\xi-1)/(\xi+1)\}$ (of interest in Theorem 6) happens with high probability. This task is not straightforward in the presence of both competing risks and censoring. The greatest challenge is the lack of the martingale property in $\dot{\mathbf{m}}(\boldsymbol{\beta}^o)$. Even if we use its martingale proxy (an approach useful in low-dimensions) as the gradient of (2.31)

$$\dot{\tilde{\mathbf{m}}}(\boldsymbol{\beta}^o) = n^{-1}\sum_{i=1}^{n}\int_0^{t^*}\{\mathbf{Z}_i(t) - \widetilde{\mathbf{Z}}(t,\boldsymbol{\beta})\}dN_i^o(t) \tag{2.47}$$

with $\widetilde{\mathbf{Z}}(t,\boldsymbol{\beta})$ defined in (2.32), the approximation error between $\dot{\mathbf{m}}(\boldsymbol{\beta}^o)$ and $\dot{\tilde{\mathbf{m}}}(\boldsymbol{\beta}^o)$ is difficult to control because the error is determined by $\{\omega_i(t) - I(C_i \geq t)\}e^{\boldsymbol{\beta}^{o\top}\mathbf{Z}_i(t)}$ with $\omega_i(t)$ defined in (2.2), which can be significantly amplified when the relative risks grow with the dimension. To prove Lemma 4, we first show that the relative risks among subjects with observed type 2 events has sub-Gaussian tails. This is achieved through the argument that their CIF cannot be arbitrarily close to one; otherwise, these subjects would have probability close to one experiencing type 1 event. As the CIF is monotonically increasing with the relative risks, it is also unlikely to observe excessively large relative risks among the subjects with observed type 2 events. We then use Lemma 13(i) in the Section 2.7 to establish the concentration of $\mathbf{S}^{(l)}(t,\boldsymbol{\beta}^o) - \widetilde{\mathbf{S}}^{(l)}(t,\boldsymbol{\beta}^o)$ around zero across all observed type 1 event times.

Theorem 6 assumes that $\Pr\big(\kappa\big(\xi, O; -\ddot{\mathbf{m}}(\boldsymbol{\beta}^o)\big) > Q_4\big)$ converges to zero for a sequence of Q_4 bounded away from zero, as sample size n goes to infinity. In Lemma 5, we show that such

event happens with high probability. Using the connection between the compatibility factor and the restricted eigenvalue [vdGB09], we show that $\kappa\big(\xi, O; -\ddot{\mathbf{m}}(\beta^o)\big)$, the compatibility factor in the cone $\mathcal{C}(\xi, O)$, is bounded away from zero with probability tending to one.

Lemma 5. *Let* $Q_2^{(l)}(n,p,\varepsilon)$ *be defined as in* (2.44)*. Denote*

$$Q_5(n,p,\varepsilon) = \left\{2Q_2^{(2)}(n,p,\varepsilon) + 4K_3Q_2^{(1)}(n,p,\varepsilon) + (5/2)K_3^2Q_2^{(0)}(n,p,\varepsilon)\right\}/\rho_2 + K_3^2\left\{(1+t^*K_2)\sqrt{2\log\big(p(p+1)/\varepsilon\big)/n} + (2/\rho_2)t^*K_2Q_6(n,p,\varepsilon)^2\right\},$$

where $Q_6(n,p,\varepsilon)$ *is the solution of*

$$p(p+1)\exp\{-nQ_6(n,p,\varepsilon)^2/(2+2Q_6(n,p,\varepsilon)/3)\} = \varepsilon/2.221.$$

If $s_o\sqrt{\log(p)/n} = o(1)$*, we have under Assumptions (C1)- (C2) for n sufficiently large*

$$\Pr\left(\kappa\big(\xi, O; -\ddot{\mathbf{m}}(\beta^o)\big) \geq \sqrt{\rho_3 - s_o(\xi+1)Q_5(n,p,\varepsilon)}\right) \geq 1 - 6\varepsilon.$$

2.3.3 Asymptotic normality for one-step estimator and honest coverage of confidence intervals

Obtaining the asymptotic normality is technically challenging. The log-likelihood has dependent summands both through the initial lasso estimator as well as the Kaplan-Meier estimator. We establish the asymptotic normality for the one-step estimator $\widehat{\mathbf{b}}$ and coverage of the confidence intervals without requiring model-selection consistency of the initial estimator. To remove the small-sample bias of IPCW, we need slightly stronger conditions than in the previous section. In this section alone, we use K and ρ without subscript to denote the constants independent of n, p

and s_o; we have only one constant K_n that is allowed to grow with the dimension and is therefore denoted differently.

(D1) (**Design**) The true linear predictors are uniformly bounded with probability one

$$\sup_{i=1,\ldots,n} \sup_{t\in[0,t^*]} \left|\boldsymbol{\beta}^{o\top}\mathbf{Z}_i(t)\right| \leq K. \tag{2.48}$$

(D2) (**Hessian**) The smallest eigenvalue $\lambda_{\min}(\boldsymbol{\Sigma}) \geq \rho > 0$, where $\boldsymbol{\Sigma}$ is defined in (2.18).

(D3) (**Continuity**) Each $\mathbf{Z}_i(t)$ can be represented as

$$\mathbf{Z}_i(t) = \mathbf{Z}_i(0) + \int_0^t \mathbf{d}_i^z(u)du + \int_0^t \boldsymbol{\Delta}_i^z(u)dN_i^z(u).$$

for random processes $\mathbf{d}_i^z(t)$, $\boldsymbol{\Delta}_i^z(t)$ and the counting process $N_i^z(t)$ such that , $\boldsymbol{\beta}^{o\top}\mathbf{d}_i^z(t)$ is uniformly bounded between $\pm K$ and uniformly Lipschitz-K. Moreover, $N_i^z(t)$'s number of jumps $K_n = o\left(\sqrt{n/(\log(p)\log(n))}\right)$ and an intensity function $h^N(t) \leq K$.

(D4) (**Dimension**) The rows of the matrix $\boldsymbol{\Sigma}^{-1}$ are $\|\boldsymbol{\theta}_j/\Theta_{j,j}\|_1 \leq K$ and sparse with sparsities $s_1,\ldots,s_p \leq s_{\max}$. Lastly, $s_o(s_{\max}+s_o)\log(p)/\sqrt{n} = o(1)$.

We next present Theorem 7 that justifies all the proposed inference procedures in Section 2.2.2. For that purpose we denote the asymptotic variance of $\dot{\mathbf{m}}(\boldsymbol{\beta}^o)$ with

$$\mathcal{V} = \mathbb{E}\{\boldsymbol{\eta}_i + \boldsymbol{\psi}_i\}^{\otimes 2}, \tag{2.49}$$

where

$$\boldsymbol{\eta}_i = \int_0^{t^*} \{\mathbf{Z}_i(t) - \boldsymbol{\mu}(t)\}\widetilde{\omega}_i(t)dM_i^1(t), \tag{2.50}$$

$$\psi_i = \int_0^{t^*} \int_0^{t^*} \frac{\mathbf{q}(t)}{\pi(t)} I(X_i \geq t) dM_i^c(t), \tag{2.51}$$

$$\mathbf{q}(t) = \mathbb{E}\left[I(t > X_i) \int_t^{t^*} \{\mathbf{Z}_i(u) - \boldsymbol{\mu}(u)\} \widetilde{\omega}_i(u) dM_i^1(u) \right], \tag{2.52}$$

$$\pi(t) = \Pr(X_i \geq u), \tag{2.53}$$

with $M_i^1(t)$, $M_i^c(t)$ as defined in (2.28) and (2.36).

Theorem 7. *Let $\boldsymbol{\Theta}$ be defined as in Section 2.2.3. Let $\mathcal{V}$, $\widehat{\mathbf{b}}$, $\widehat{\boldsymbol{\Theta}}$ and $\widehat{\mathcal{V}}$ be defined as in* (2.49)*,* (2.6)*,* (2.27) *and* (2.8)*, respectively. Let $\mathbf{c} \in \mathbb{R}^p$ with $\|\mathbf{c}\|_1 = 1$ and $\mathbf{c}^\top \boldsymbol{\Theta} \mathcal{V} \boldsymbol{\Theta} \mathbf{c} \to \nu^2 \in (0, \infty)$. Then, whenever (C1) and (D1)-(D4) hold,*

$$\frac{\sqrt{n}\mathbf{c}^\top (\widehat{\mathbf{b}} - \boldsymbol{\beta}^o)}{\sqrt{\mathbf{c}^\top \widehat{\boldsymbol{\Theta}} \widehat{\mathcal{V}} \widehat{\boldsymbol{\Theta}}^\top \mathbf{c}}} \xrightarrow{d} N(0,1).$$

As a result of the stronger conditions required for Theorem 7, which we will explain in more details below, we are able to achieve an improved estimation error for the initial estimator as stated in the next theorem.

Theorem 8. *Under (C1) and (D1)-(D4), we can choose $\lambda \asymp \sqrt{\log(p)/n}$ and $Q_4 = \sqrt{\rho_3/2}$, such that*

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^o\|_1 = O_p\left(s_o \sqrt{\log(p)/n} \right) = o_p(1).$$

For the rest of this section, we explain the assumptions and theoretical results needed for Theorem 7 summarized in Lemmas 6-10. Condition (D1) is needed whenever the model departs significantly from the linear case [vdGBRD14, FNL17]. In our case, the asymptotic normality of $\sqrt{n}\dot{\mathbf{m}}(\boldsymbol{\beta}^o)$ depends fundamentally on the asymptotic tightness of $\sqrt{n}\dot{\widetilde{\mathbf{m}}}(\boldsymbol{\beta}^o)$. As a necessary

condition, the predictable quadratic variation under filtration $\mathcal{F}_t{}^*$ of the martingale $\sqrt{n}\dot{\widetilde{\mathbf{m}}}(\boldsymbol{\beta}^o)$

$$\langle\sqrt{n}\dot{\widetilde{\mathbf{m}}}(\boldsymbol{\beta}^o)\rangle_{t^*} = \int_0^{t^*} n^{-1}\sum_{i=1}^{n} I(C_i \geq t)Y_i(t)e^{\boldsymbol{\beta}^{o\top}\mathbf{Z}_i(t)}\{\mathbf{Z}_i(t) - \widetilde{\mathbf{Z}}(t,\boldsymbol{\beta}^o)\}^{\otimes 2}h_0^1(t)dt, \tag{2.54}$$

must have a finite bound independent of the dimension of the covariates. This requires that the magnitude of the summands in (2.54) either be bounded or have light tails. Hence, we cannot allow the relative risk $e^{\boldsymbol{\beta}^{o\top}\mathbf{Z}_i(t)}$ to grow arbitrarily large. We next observe that (D2) is a standard assumption for the validity of the nodewise penalized regressions (2.26). Finally, note that Theorem 7 utilizes Condition (D3); a condition stronger than (C3) needed for $\sqrt{n}$- approximation error between $\dot{\mathbf{m}}(\boldsymbol{\beta}^o)$ and $\dot{\widetilde{\mathbf{m}}}(\boldsymbol{\beta}^o)$.

If we define the population versions of the nodewise components defined in (2.24)-(2.26),

$$\begin{aligned}
&\boldsymbol{\Xi} = \int_0^{t^*}\{\mathbf{Z}(t) - \boldsymbol{\mu}(t)\}dN^o(t),\ \bar{\Gamma}_j(\boldsymbol{\gamma}) = \mathbb{E}\{\Xi_j - \boldsymbol{\Xi}_{i,-j}^{\top}\boldsymbol{\gamma}\}^2,\\
&\boldsymbol{\gamma}_j^* = \underset{\boldsymbol{\gamma}\in\mathbb{R}^{p-1}}{\operatorname{argmin}}\,\bar{\Gamma}_j(\boldsymbol{\gamma}),\ \tau_j^2 = \bar{\Gamma}_j(\boldsymbol{\gamma}_j^*),
\end{aligned} \tag{2.55}$$

then the true parameters $\{\boldsymbol{\gamma}_j^*, \tau_j^2 : j = 1,\ldots,p\}$ uniquely define the inverse negative Hessian $\boldsymbol{\Theta}$ as described in Section 2.2.3. We prove this statement in the following Lemma.

Lemma 6. *Under (D2),* $\Theta_{j,j} = 1/\tau_j^2$ *and* $\boldsymbol{\theta}_{j,-j}\tau_j^2 = \boldsymbol{\gamma}_j^*$. *Moreover,* $\|\boldsymbol{\gamma}_j^*\|_1 \leq K$, $\tau_j^2 \geq \rho$ *and* $\|\boldsymbol{\Theta}\|_1 \leq K/\rho$.

Next, we discuss the properties of estimands $\widehat{\boldsymbol{\gamma}}_j$, $\widehat{\tau}_j$ and $\widehat{\boldsymbol{\Theta}}$ – defining components of the variance estimate.

Lemma 7. *Under (C1) and (D1)-(D4), for* $\lambda_j \asymp s_o\sqrt{\log(p)/n}$, *we obtain*

$$\sup_j \|\widehat{\boldsymbol{\gamma}}_j - \boldsymbol{\gamma}_j^*\|_1 = O_p\left(s_o s_j\sqrt{\log(p)/n}\right)$$

and $\sup_j |\widehat{\tau}_j^2 - \tau_j^2| = O_p(s_o s_j \sqrt{\log(p)/n})$, *leading to* $\|\widehat{\Theta} - \Theta\|_1 = O_p\left(s_o s_{\max}\sqrt{\log(p)/n}\right)$.

The nodewise LASSO in (2.26), unlike [vdGB09] that has i.i.d. entries, has dependent $\widehat{\Xi}_i$'s through the common $\bar{\mathbf{Z}}(t,\widehat{\beta})$; see (2.24). Thus, our error rate takes the multiplicative form $s_o s_{\max}$, instead of the summation $s_o + s_{\max}$ that may be expected under the generalized linear models. In general, we consider our rate to be optimal under our model.

Using Lemma 7, we can establish the approximation condition for $\widehat{\mathbf{b}}$ proposed in (2.7).

Lemma 8. *Under (C1) and (D1)-(D4), the one-step estimator* $\widehat{\mathbf{b}}$ *satisfies the approximation condition*

$$\sqrt{n}\mathbf{c}^\top \left\{\Theta\dot{\mathbf{m}}(\beta^o) + \beta^o - \widehat{\mathbf{b}}\right\} = O_p\left(s_o(s_{\max}+s_o)\log(p)/\sqrt{n}\right) = o_p(1)$$

for any $\mathbf{c}$ *such that* $\|\mathbf{c}\|_1 = 1$.

Next, we show the asymptotic normality of $\dot{\mathbf{m}}(\beta^o)$.

Lemma 9. *Under conditions (C1) and (D1)-(D4), for directional vector* $\mathbf{c} \in \mathbb{R}^p$ *with* $\|\mathbf{c}\|_1 = 1$ *and* $\mathbf{c}^\top \Theta \mathcal{V} \Theta^\top \mathbf{c} \to \mathrm{v}^2 \in (0,\infty)$,

$$\sqrt{n}\mathbf{c}^\top \Theta \dot{\mathbf{m}}(\beta^o)/\sqrt{\mathbf{c}^\top \Theta \mathcal{V} \Theta^\top \mathbf{c}} \xrightarrow{d} N(0,1).$$

The proof uses the same approach as the initial low-dimensional result in [FG99]. We approximate $\dot{\mathbf{m}}(\beta^o)$ by the sample average of i.i.d. terms $\eta_i + \psi_i$ plus an $o_p\left(n^{-1/2}\right)$ term. We note that the same approach involves nontrivial techniques in order to be valid in high-dimensions. In particular, we discover and exploit the martingale property of the term $\{\omega_i(t) - I(C_i \geq t)\}/G(t)$.

The last piece of our proof for Theorem 7 is the element-wise convergence of the "meat" matrix (2.8) in the "sandwich" variance estimator.

Lemma 10. *Under conditions (C1) and (D1)-(D4),*

$$\sup_{i=1,\ldots,n} \|\widehat{\boldsymbol{\eta}}_i(\widehat{\boldsymbol{\beta}}) + \widehat{\boldsymbol{\psi}}_i(\widehat{\boldsymbol{\beta}}) - \boldsymbol{\eta}_i - \boldsymbol{\psi}_i\|_\infty = O_p\left(\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^o\|_1 + \sqrt{\log(p)/n}\right) = o_p(1).$$

Hence, $\|\widehat{\mathcal{V}} - \mathcal{V}\|_{\max} = o_p(1)$.

Putting Lemmas 9 and 10 together, we obtain the main result stated in the Theorem 7.

The details of the proofs are presented in the Section 2.7. Throughout the proof, we rely heavily on our concentration results for time-dependent processes, which we state in Section 2.7.1 and prove in Section 2.7.2.

2.4 Numerical Experiments

To assess the finite sample properties of our proposed methods, we conduct extensive simulation experiments with various dimensions and dependence structure among covariates.

2.4.1 Setup 1

Our first simulation setup follows closely the one of [FG99] but considers high-dimensional covariates. In particular, each $\mathbf{Z}_i$ is a vectors consisting of i.i.d. standard normal random variables. For cause 1, only $\beta_{1,1} = \beta_{1,2} = 0.5$ are non-zero. The cumulative incidence function is:

$$\Pr(T_i \leq t, \varepsilon_i = 1 | \mathbf{Z}_i) = 1 - [1 - p\{1 - \exp(-t)\}]^{\exp(\boldsymbol{\beta}_1^\top \mathbf{Z}_i)}.$$

For cause 2, $\beta_{2,1} = \beta_{2,3} = \cdots = \beta_{2,p-1} = -0.5$ and $\beta_{2,2} = \beta_{2,4} = \cdots = \beta_{2,p} = 0.5$, with

$$\Pr(T_i \leq t | \varepsilon_i = 2, \mathbf{Z}_i) = 1 - \exp\left(te^{\boldsymbol{\beta}_2^\top \mathbf{Z}_i}\right).$$

Table 2.1: Simulation results with independent covariates.

	True	Mean Est	SD	SE	SE corrected	Coverage	Level/Power
				n=200, p=300			
$\beta_{1,1}$	0.5	0.51	0.16	0.13	0.25	0.94	0.92
$\beta_{1,2}$	0.5	0.47	0.15	0.14	0.22	0.94	0.93
$\beta_{1,10}$	0	0.03	0.12	0.15	0.18	0.98	0.04
				n=200, p=500			
$\beta_{1,1}$	0.5	0.51	0.16	0.14	0.19	0.93	0.95
$\beta_{1,2}$	0.5	0.48	0.15	0.13	0.19	0.93	0.88
$\beta_{1,10}$	0	-0.01	0.10	0.14	0.16	1.00	0.01
				n=200, p=1000			
$\beta_{1,1}$	0.5	0.46	0.17	0.13	0.18	0.94	0.86
$\beta_{1,2}$	0.5	0.48	0.14	0.13	0.18	0.93	0.92
$\beta_{1,10}$	0	-0.00	0.11	0.14	0.17	0.99	0.06
				n=500, p=1000			
$\beta_{1,1}$	0.5	0.51	0.10	0.08	0.14	0.99	1.00
$\beta_{1,2}$	0.5	0.50	0.10	0.08	0.15	0.99	0.99
$\beta_{1,10}$	0	-0.00	0.07	0.08	0.14	1.00	0.03

We consider four different combinations: $n = 200$, $p = 300$; $n = 200$, $p = 500$; $n = 200$, $p = 1000$; and $n = 500$, $p = 1000$. Note that this setup considers sparsity for cause 1 but non-sparsity for cause 2 effects. As the Fine-Gray model does not require modeling cause 2 to make inference on cause 1, we expect that the non-sparsity in cause 2 effects should not affect the inference on cause 1.

The results are presented in Table 2.1. We focus on inference for the two non-zero

coefficients $\beta_{1,1}$ and $\beta_{1,2}$, as well as one arbitrarily chosen zero coefficient $\beta_{1,10}$. The mean estimates are the average of the one-step $\widehat{\mathbf{b}}$ over the 100 repetitions, reported together with other quantities described below. We can see from the average estimates column that the one-step $\widehat{\mathbf{b}}$ is bias-corrected and that the presence of many non-zero coefficients for causet 2 does not affect our inference on cause 1.

In practice the choice of the tuning parameters is particularly challenging; the optimal value is determined up to a constant. Moreover, the theoretical results are asymptotic. These together with the finite sample effects of $n \ll p$, lead to suboptimal performance of many proposed one-step correction estimators [vdGBRD14, FNL17]. Suboptimality is amplified for survival models, due to the nonlinearity of the loss function and the presence of censoring – both require more significant sample size (to observe asymptotic statements in the finite samples). In the following, we propose a finite-sample correction to the construction of confidence intervals and in particular the estimated standard error (SE).

Let $se(\widehat{b}_j;\widehat{\boldsymbol{\beta}})$ denote the asymptotic standard error as given in Section 2.2.2. As a finite-sample correction we propose to consider $se(\widehat{b}_j;\widehat{\mathbf{b}})$ in place of $se(\widehat{b}_j;\widehat{\boldsymbol{\beta}})$, where the variance estimation based on the initial LASSO estimate $\widehat{\boldsymbol{\beta}}$ is replaced by the one-step $\widehat{\mathbf{b}}$. This can be viewed as another iteration of the bias-correction formula. The resulting SE is therefore a "two-step" SE estimator. We report the coverage rate of the confidence intervals constructed with this finite-sample correction in Table 2.1 and we observe good coverage close to the nominal level of 95%. We note that with 100 simulation runs the margin of error for the simulated coverage probability is about 2.18%, if the true coverage is 95%; that is, the observed coverage can range between $95+/-4.36\%$. We note that the coverage is good for all three coefficients,

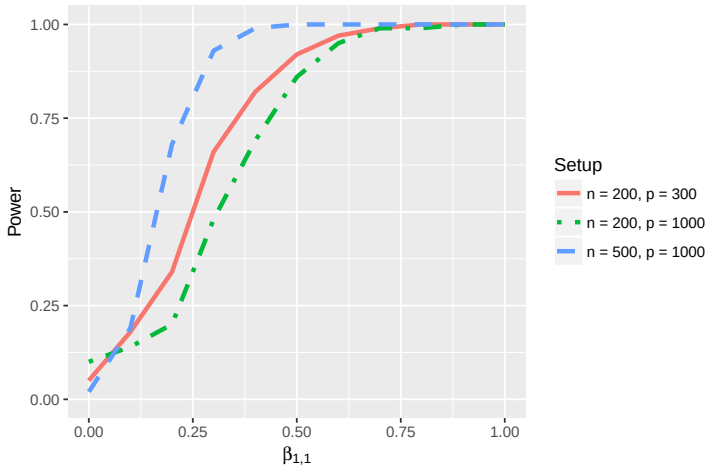


Figure 2.1: Power curve for testing $\beta_{1,1} = 0$ at nominal level 0.05.

where non-zero or zero. In contrast, results in the existing literature suffer under-coverage of the non-zero coefficients.

The last column 'level/power' in Table 2.1 refers to the empirical rejection rate of the null hypothesis that the coefficient is zero, by the two-sided Wald test $Z = (\widehat{b}_j - \beta_{1,j})/se(\widehat{b}_j; \widehat{\boldsymbol{\beta}})$ at a nominal 0.05 significance level. We see that although $se(\widehat{b}_j; \widehat{\boldsymbol{\beta}})$ is used, the nominal level is well preserved for the zero coefficient $\beta_{1,10}$, and the power is high for the non-zero coefficients $\beta_{1,1}$ and $\beta_{1,2}$ for the given sample sizes and signal strength.

We repeat the above simulations with different values for $\beta_{1,1}$ to investigate the power of the Wald test. The results are illustrated in Figure 2.1, where we see that the power increases with n and decreases with p as expected.

2.4.2 Setup 2

In the second setup we consider the case where the covariates are not all independent, which is more likely the case in practice for high dimensional data. We consider the block dependence structure also used in [BASB09]. We consider $n = 500$, $p = 1000$; $\beta_{1,1\sim8} = 0.5$, $\beta_{1,9\sim12} = -0.5$ and the rest are all zero. $\beta_{2,1\sim4} = \beta_{2,13\sim16} = 0.5$, $\beta_{2,5\sim8} = -0.5$ and the rest of β_1 are all zero. The covariates are grouped into four blocks of size 4, 4, 8 plus the rest, with the within-block correlations equal to 0.5, 0.35, 0.05 and 0. The four blocks are separated by the horizontal lines in Table 2.2.

Table 2.2 shows the inferential results for the non-zero coefficients $\beta_{1,1} \sim \beta_{1,12}$, as well as the zero coefficients $\beta_{1,13} \sim \beta_{1,16}$ from the third correlated block that also contains some of the non-zero coefficients, and plus arbitrarily chosen zero coefficient $\beta_{1,30}$. The initial LASSO estimator tended to select only one of every four non-zero coefficients of the correlated covariates (data not shown), as it is known that block dependence structure is particularly challenging for the Lasso type estimators. On the other hand, the one-step estimator performed remarkably well, capturing all of the non-zero coefficients.

Compared to the results in the last part of Table 2.1 with the same n and p, the block correlated covariates led to slightly more bias in $\widehat{\mathbf{b}}$, although the CI coverage remained high. The power also remained high, although in the third block with the mixed signal and noise variables the type I error rates appeared slightly high.

Table 2.2: Simulation results with block correlated covariates.

	True	Mean Est	SD	SE	SE corrected	Coverage	Level/Power
				n=500, p=1000			
$\beta_{1,1}$	0.5	0.47	0.10	0.07	0.12	0.97	1.00
$\beta_{1,2}$	0.5	0.48	0.10	0.07	0.12	0.94	0.98
$\beta_{1,3}$	0.5	0.47	0.10	0.07	0.12	0.98	1.00
$\beta_{1,4}$	0.5	0.47	0.10	0.07	0.12	0.94	1.00
$\beta_{1,5}$	0.5	0.48	0.10	0.06	0.11	0.93	1.00
$\beta_{1,6}$	0.5	0.46	0.10	0.06	0.11	0.94	1.00
$\beta_{1,7}$	0.5	0.47	0.09	0.06	0.11	0.95	1.00
$\beta_{1,8}$	0.5	0.47	0.08	0.06	0.11	0.98	1.00
$\beta_{1,9}$	-0.5	-0.44	0.08	0.06	0.11	0.93	1.00
$\beta_{1,10}$	-0.5	-0.42	0.08	0.06	0.11	0.92	1.00
$\beta_{1,11}$	-0.5	-0.41	0.08	0.06	0.11	0.91	1.00
$\beta_{1,12}$	-0.5	-0.43	0.07	0.05	0.11	0.94	1.00
$\beta_{1,13}$	0	-0.01	0.06	0.05	0.11	0.98	0.11
$\beta_{1,14}$	0	-0.02	0.05	0.05	0.11	1.00	0.06
$\beta_{1,15}$	0	-0.02	0.06	0.06	0.11	0.99	0.08
$\beta_{1,16}$	0	-0.02	0.06	0.05	0.11	1.00	0.05
$\beta_{1,30}$	0	-0.00	0.05	0.06	0.11	1.00	0.01

2.5 SEER-Medicare data example

The SEER-Medicare linked database contains clinical information and claims codes for 57011 patients diagnosed between 2004 and 2009. The clinical and demographic information were collected at diagnosis, and the insurance claim data were from the year prior to diagnosis. The clinical information contained PSA, Gleason Score, AJCC stage and year of diagnosis. Demographic information included age, race, and marital status. The same data set was considered in [HPH$^+$18a], where the emphasis was on variable selection and prediction error. Our focus is on testing and construction of confidence intervals.

In the following, we consider 2000 patients diagnosed during the year of 2004. The only cause for loss to follow-up was the administrative censoring at the end of the study which was year 2011. Consequently, the year of enrollment was the only factor affecting the censoring distribution. In our sample, all the subjects share the same year of enrollment 2004, so we may reasonably make the independent censoring assumption. Among them 76 died from the cancer and 337 had deaths unrelated to cancer. The process of identifying of the causes is detailed in [RTH$^+$19]. There were 9326 binary claims codes in the data. Here we would like to identify the risk factors for non-cancer mortality using the Fine-Gray model. We kept only the claims codes with at least 10 and at most 1990 occurrences. The resulting dataset had 1197 covariates. We center and standardize all the covariates before performing the analysis. To determine the penalty parameters λ and λ_j we used 10-fold cross-validation.

In Table 2.3, we present the result for 21 coefficients. Here, we focused on potential risk factors for non-cancer mortality, such as heart disease and colon cancer (different than prostate

cancer); the coefficients to be tested were chosen ahead of time following recommendations from the doctors. We also include the clinical markers associated with the prostate cancer in comparison. A descriptions of the variables is given in Table 2.4. For each coefficient, we report the initial estimate $\widehat{\beta}$, one-step estimate $\widehat{\mathbf{b}}$, corrected SE, the 95% CI constructed with the corrected SE and the Wald test p-value (2-sided) calculated using the uncorrected SE.

In Table 2.3, we see that the claims codes ICD-9 4280, CPT 93015, ICD-9 42731 are all related to the heart disease, and are all significant at 5% level Bonferonni correction for the 21 variables included in the table. However, a heart attack indicator variable, ICD-9 41189, shows up significant at 10% level although the naive regularized estimator was not able to select this variable as important; this indicates that our inference procedure is much more delicate (stable) at discovering significant variables. In support of that, an indicator of a possible cancer in the abdomen, CPT 74170, is reported as significant at 5% although the initial Lasso regularized method failed to include such variable. Similar result is seen for the indicator of a fall (CPT 72050) which for an elderly person can be fatal. An indicator of a colon cancer (CPT 45380) turns out to be significant at 10% although the Lasso method set it to zero initially. Therefore, our one-step method is able to recover important risk factors that would have been missed by the initial regularized estimator.

In contrast, non-life-threatening diseases, were not selected as significant predictors for the non-cancer mortality. These include Parkinson's (ICD-9 3320), Psychosis (ICD-9 2989), Bronchitis (ICD-9 49121) and Dementia (ICD-9 2948) in the table. It is worth noting that some of these were selected by the initial estimate but were then corrected by our test. We also note that the prostate cancer related variables, PSA, Gleason Score ahd AJCC all have large

Table 2.3: Inference for the SEER-Medicare linked data on non-cancer mortality among prostate cancer patients.

Variables	Initial estimate	One-step estimate and Inference			
	$\widehat{\beta}$	$\widehat{b}$	se($\widehat{b}$)	95% CI	p-value
Age	0.075	0.096	0.009	[0.078, 0.114]	2e-24*
Marital	0	0.218	0.147	[-0.071, 0.507]	0.042
Race.OvW	0	-0.213	0.224	[-0.652, 0.225]	0.317
Race.BvW	0.244	0.528	0.122	[0.288, 0.767]	1e-04*
PSA	0	0.005	0.003	[-0.000, 0.010]	0.041
GleasonScore	0	0.084	0.050	[-0.014, 0.182]	0.085
AJCC-T2	0	-0.130	0.146	[-0.418, 0.157]	0.218
ICD-9 51881	0.866	1.357	0.361	[0.650, 2.064]	4e-07*
ICD-9 4280	0.404	0.697	0.062	[0.576, 0.818]	2e-06*
CPT 93015	-0.061	-1.042	0.327	[-1.683, -0.401]	4e-05*
ICD-9 42731	0.135	0.459	0.191	[0.086, 0.833]	0.001*
CPT 72050	0	3.718	0.208	[3.310, 4.125]	4e-05*
ICD-9 6001	0	-2.454	0.577	[-3.585, -1.322]	0.000*
CPT 74170	0	-1.689	0.288	[-2.255, -1.124]	0.001*
ICD-9 2948	0.539	0.746	0.205	[0.343, 1.148]	0.009
ICD-9 49121	0.150	0.476	0.215	[0.055, 0.896]	0.015
ICD-9 2989	0.079	0.450	0.135	[0.184, 0.715]	0.062
ICD-9 79093	-0.056	-0.348	0.176	[-0.693, -0.002]	0.088
ICD-9 41189	0	1.332	0.434	[0.480, 2.184]	0.003**
CPT 45380	0	-2.250	0.544	[-3.318, -1.182]	0.003**
ICD-9 3320	0	0.378	0.373	[-0.353, 1.110]	0.327

* denotes 5% significance after Bonferroni correction for these 21 variables;

** denotes 10% significance after Bonferroni correction for these 21 variables

Table 2.4: Description of the variables in Table 2.3

Code	Description
Age	Age at diagnosis
Marital	marSt1: married vs other
Race.OvW	Race: Other vs White
Race.BvW	Race: Black with White
PSA	PSA
GleasonScore	Gleason Score
AJCC-T2	AJCC stage-T: T2 vs T1
ICD-9 51881	Acute respiratry failure (Acute respiratory failure)
ICD-9 4280	Congestive heart failure; nonhypertensive [108.]
CPT 93015	Global Cardiovascular Stress Test
ICD-9 42731	Cardiac dysrhythmias [106.]
CPT 72050	Diagnostic Radiology (Diagnostic Imaging) Procedures of the Spine and Pelvis
ICD-9 6001	Nodular prostate
CPT 74170	Diagnostic Radiology (Diagnostic Imaging) Procedures of the Abdomen
ICD-9 2948	Delirium dementia and amnestic and other cognitive disorders [653]
ICD-9 49121	Obstructive chronic bronchitis
ICD-9 2989	Unspecified psychosis
ICD-9 41189	acute and subacute forms of ischemic heart disease, other
CPT 45380	Under Endoscopy Procedures on the Rectum
ICD-9 3320	Parkinsons disease [79.]

p-values for non-cancer mortality. This is consistent with the results in [HPH$^+$18a], where under the competing risk models the predictors for a second cause only has secondary importance in predicting the events due to the first cause.

2.6 Discussion

This article focuses on estimation and inference under the Fine-Gray model with many more covariates than the number of events, which is well-known to be the effective sample size for survival data. The article studies the rate of convergence of a Lasso estimator and develops a new one-step estimator that can be utilized for asymptotically optimal inference: confidence intervals and testing. These results can be generalized to any sparsity-inducing and convex penalty functions including but not limited to one-step SCAD, adaptive LASSO, elastic net, to name a few. Moreover, it is worth noting that the variance estimation is novel in that it regresses a re-weighted score vector onto the score vector; in this way, the usual difficulty with asymptotic Hessian is avoided; it is worth pointing that the sandwich estimator or bootstrap carry biases in high-dimensions.

An often overlooked restriction on the time-dependent covariates $Z_i(t)$, $i = 1, \ldots, n$, under the Fine-Gray model is that $Z_i(t)$ must be observable even after the i-th subject experiences a type 2 event. In practice, $Z_i(t)$ should be either time independent or external [KP02]. In our case the continuity conditions (C3) and (D3) are easily satisfied if the majority of the elements in $Z_i(t)$ are time independent, which is most likely to be the case in practice. Our theory does not apply in studies involving longitudinal variables that are supposed to be truly measured continuously over

time.

We have illustrated that the method based on regularization only (without bias correction) might have severe disadvantages in many complex data situations – for example, it may potentially fail to identify relevant variables that are associated with the response. From the analysis of the SEER-medicare data, we see that variables like CPT 72050 (related to fall) or, CPT 74170 (related to diagnostic imaging of the abdomen, often for suspected malignancies) would not have been discovered as important risk factors for non-cancer mortality by regularization alone. In reality, both can be life-threatening events for an elderly patient. The one-step estimate, on the other hand, was able to detect these, therefore providing a valuable tool for practical applications. The one-step estimator is applicable as long as the model is sparse, and no minimum signal strength is required; this is another important aspect which makes the estimator more desirable for practical use than the LASSO type estimators.

2.7 Proof

We denote global quantities as Q and event sets as Ω with subscripts labelled by their order of appearance. Other quantities are all local, i.e. only defined for the current Lemma. We denote the ordered observed type-1 event times as $T^1_{(1)}, \ldots, T^1_{(K_T)}$.

2.7.1 Concentration Inequalities

Here we give the statements of the inequalities frequently used in our proofs. The notations in this section are all generic.

Classical Concentration Inequalities

Lemma 11. ***Hoeffding's Inequality*** *(Theorem 2 of [Hoe63] p.4) If $X_1,\ldots,X_n$ are independent and $a_i \le X_i \le b_i$ $(i = 1,2,\ldots,n)$, then for $t > 0$*

$$\Pr(\bar{X} - \mu \ge t) \le \exp\left(-\frac{2n^2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right).$$

Lemma 12. ***A version of Azuma's Inequality*** *(Theorem 1 and Remark 7 of [Sas13] p.3 and p.5) Let $\{X_k, \mathcal{F}_k\}_k = 0^\infty$ be a discrete-parameter real-valued martingale sequence such that for every k, the condition $|X_k - X_{k-1}| \le a_k$ holds almost surely for some non-negative constants $\{a_k\}_{k=1}^\infty$. Then*

$$\Pr\left(\max_{k\in 1,\ldots,n} |X_k - X_0| \ge t\right) \le 2\exp\left(-\frac{t^2}{2\sum_{k=1}^n a_k^2}\right)$$

Concentration Inequalities for Time-dependent Processes

Lemma 13. *Let $\{(\mathbf{S}_i(t), N_i(t)) \in \mathbb{R}^q \times \mathbb{N} : i = 1,\ldots,n, t \in [0,t^*]\}$ be i.i.d. pairs of random processes. Each $N_i(t)$ is a counting process bounded by K_N. Denote its jumps as $0 \le t_{i1} < \cdots < t_{iK_i} \le t^*$. Let $\bar{\mathbf{S}}(t) = n^{-1}\sum_{i=1}^n \mathbf{S}_i(t)$ and $\mathbb{E}\{\mathbf{S}_i(t)\} = \mathbf{s}(t)$. Suppose $\sup_{1\le i<j\le n} \sup_{t\in[0,t^*]} \|\mathbf{S}_i(t) - \mathbf{S}_j(t)\|_{\max} \le K_S$ almost surely. Then,*

(i) $\Pr\left(\sup_{i=1,\ldots,n} \sup_{j=1,\ldots,K_i} \left\|\bar{\mathbf{S}}(t_{ij}) - \mathbf{s}(t_{ij})\right\|_{\max} > K_S x + (K_S)/n\right) < 2nK_N q e^{-nx^2/2}.$

(ii) Assume in addition that each $\mathbf{S}_i(t)$ is càglàd generated by

$$\mathbf{S}_i(t) = \mathbf{S}_i(0) + \int_0^t \mathbf{d}_s(u)du + \int_0^t \mathbf{J}_s(u)dN_i(u)$$

for some $\mathbf{d}_s(t)$ and $\mathbf{J}_s(t)$ satisfying $\|\mathbf{d}_s(t)\|_{\max} < L_S$ and $\|\mathbf{J}_s(u)\|_{\max} < K_S$, and $\mathbb{E}\{N_i(t)\} =$

$\int_0^t h_i^N(u)du$ *for some* $h_i^N(t) \leq K$. *We have*

$$\sup_{i=1,\ldots,n} \sup_{t\in[0,t^*]} \left\|\bar{\mathbf{S}}(t) - \mathbf{s}(t)\right\|_{\max} = O_p(\sqrt{\log(nK_N q)/n}).$$

Lemma 14. *Let* $\{M_i(t) : t \in [0,t^*],\ i = 1,\ldots,n\}$ *be a* $\mathcal{F}_t$*-adapted counting process martingales* $M_i(t) = N_i(t) - \int_0^t Y_i(t)h_i(u)du$ *satisfying* $\sup_{i=1,\ldots,n} \sup_{t\in[0,t^*]} h_i(t) \leq K_h$. *Let* $\{\mathbf{\Phi}_i(t) : t \in [0,t^*],\ i = 1,\ldots,n\}$ *be the q dimensional* $\mathcal{F}_{t-}$*-measurable processes such that*

$$\sup_{i=1,\ldots,n} \sup_{t\in[0,t^*]} \|\mathbf{\Phi}_i(t)\|_{\max} \leq K_\Phi.$$

For $\mathbf{M}_\Phi(t) = n^{-1}\sum_{i=1}^n \int_0^t \mathbf{\Phi}_i(u)dM_i(u)$ *, we have*

(i) $\Pr\left(\sup_{t\in[0,t^*]} \|\mathbf{M}_\Phi(t)\|_{\max} \geq K_\Phi(1+K_h t^*)x + K_\Phi K_h t^*/n\right) \leq 2qe^{-nx^2/4}$.

(ii) Assume in addition $\sup_{i=1,\ldots,n} \sup_{t\in[0,t^*]} \|\mathbf{\Phi}_i(t)\|_{\max} = O_p(a_n)$ *and* $K_h t^* \asymp O(1)$. *Then,* $\sup_{t\in[0,t^*]} \|\mathbf{M}_\Phi(t)\|_{\max} = O_p(a_n\sqrt{\log(q)/n})$.

2.7.2 Proofs of Main Results

We shall present our proofs in the following order. First, we give the proofs to our theorems using the main Lemmas stated in Section 2.3. Second, we present the auxiliary lemmas necessary for the proofs of main Lemmas. Third, we present the proofs to the main Lemmas. Lastly, we present the proofs to the our concentration inequalities and auxiliary lemmas.

Proofs of Theorems

Proof of Theorem 6. Observe that the same techniques as those of [HSY$^+$13] apply (see for example Lemmas 3.1 and 3.2 therein). The structure of the partial likelihood is the same as that

of the Cox model modular the IPW weight functions $w_j(t)$. Following the same line of proof we can easily obtain on the event $\{\|\dot{\mathbf{m}}(\boldsymbol{\beta}^o)\|_\infty < \lambda(\xi-1)/(\xi+1)\}$, the estimation error of LASSO estimator $\widehat{\boldsymbol{\beta}}$ defined in (2.5) has the bound

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^o\|_1 \leq \frac{e^{\varsigma}(\xi+1)s_o\lambda}{2\kappa\big(\xi, O; -\ddot{\mathbf{m}}(\boldsymbol{\beta}^o)\big)^2}, \tag{2.56}$$

where ς is the smaller solution to

$$\varsigma e^{-\varsigma} = K_3(\xi+1)s_o\lambda/\{2\kappa\big(\xi, O; -\ddot{\mathbf{m}}(\boldsymbol{\beta}^o)\big)^2\}.$$

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^o\|_1 \leq \frac{e^{\varsigma}(\xi+1)s_o\lambda}{2\kappa\big(\xi, O; -\ddot{\mathbf{m}}(\boldsymbol{\beta}^o)\big)^2} \tag{2.57}$$

with $\varsigma_{\mathbf{b}} = \sup_{t\in[0,t^*]}\sup_{1\leq i<j\leq n}|\mathbf{b}^\top\{\mathbf{Z}_i(t) - \mathbf{Z}_j(t)\}|$ in the event $\|\dot{\mathbf{m}}(\boldsymbol{\beta}^o)\|_1 \leq \lambda(\xi-1)/(\xi+1)$. The proof is then completed by applying the conclusion of Lemma 4. □

Proof of Theorem 7. Be Lemmas 8 and 9, we have

$$\sqrt{n}\frac{\mathbf{c}^\top(\widehat{\mathbf{b}} - \boldsymbol{\beta}^o)}{\mathbf{c}^\top\boldsymbol{\Theta}\mathcal{V}\boldsymbol{\Theta}^\top\mathbf{c}} = \sqrt{n}\frac{\boldsymbol{\Theta}\dot{\mathbf{m}}(\boldsymbol{\beta}^o)}{\mathbf{c}^\top\boldsymbol{\Theta}\mathcal{V}\boldsymbol{\Theta}^\top\mathbf{c}} + o_p(1) \xrightarrow{d} N(0,1).$$

In Lemma 10, we have shown that $\|\mathcal{V}\|_{\max}$ is bounded by $K^2(1+Ke^Kt^*)^2\{1+2(1+K)e^K/\rho_2\}^2$ with probability tending to one. In Lemma 6, we have shown that $\|\boldsymbol{\Theta}\|_1$ is bounded by K/ρ. Then, we can apply Lemmas 7 and 10 to get

$$\begin{aligned}
|\mathbf{c}^\top\boldsymbol{\Theta}\mathcal{V}\boldsymbol{\Theta}^\top\mathbf{c} - \mathbf{c}^\top\widehat{\boldsymbol{\Theta}}\widehat{\mathcal{V}}\widehat{\boldsymbol{\Theta}}^\top\mathbf{c}| &\leq \|\mathbf{c}\|_1\|\boldsymbol{\Theta} - \widehat{\boldsymbol{\Theta}}\|_1\|\mathcal{V}\|_{\max}\|\boldsymbol{\Theta}\|_1\|\mathbf{c}\|_1 \\
&+ \|\mathbf{c}\|_1\{\|\boldsymbol{\Theta}\|_1 + \|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}\|_1\}\|\mathcal{V} - \widehat{\mathcal{V}}\|_{\max}\|\boldsymbol{\Theta}\|_1\|\mathbf{c}\|_1 \\
&+ \|\mathbf{c}\|_1\{\|\boldsymbol{\Theta}\|_1 + \|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}\|_1\}\{\|\widehat{\mathcal{V}} - \mathcal{V}\|_{\max} + \|\mathcal{V}\|_{\max}\}\|\boldsymbol{\Theta} - \widehat{\boldsymbol{\Theta}}\|_1\|\mathbf{c}\|_1
\end{aligned}$$

$$= 2O_p(\|\Theta - \widehat{\Theta}\|_1) + O_p(\|\mathcal{V} - \widehat{\mathcal{V}}\|_{\max}) = o_p(1).$$

Note that we use the following fact

$$\|\mathbf{c}^\top \Theta\|_1 = \sum_{j=1}^{p} |\sum_{i=1}^{p} c_i \Theta_{i,j}| \leq \sum_{i=1}^{p} |c_i| \sum_{j=1}^{p} |\Theta_{i,j}| \leq \|\mathbf{c}\|_1 \|\Theta\|_1.$$

□

Proof of Theorem 8. Since we assume (D1) now, the relative risks are bounded almost surely from above and below by constants $0 < e^{-K} \leq e^{\beta^{o\top} \mathbf{Z}_i(t)} \leq e^K < \infty$. We may set $K_4 = e^K$ to directly obtain (C2) from (D2). We can also improve the rate of estimation error in Theorem 6 by $\log(n)$ because we need not let $Q_1(\varepsilon)$ in Lemma 19 to grow with n. □

Auxiliary Lemmas

Lemma 15. *Let* $\{a_i(t) : t \in [0, t^*], i = 1, \ldots, n\}$ *be a set of nonnegative processes. Under* (2.40)*, where* K_3 *is defined,*

$$\left\| \frac{\sum_{i=1}^n a_i(t) \mathbf{Z}_i(t)^{\otimes l}}{\sum_{i=1}^n a_i(t)} \right\|_{\max} \leq (K_3/2)^l, \text{ and } \left\| \frac{\mathbb{E}\{a_i(t) \mathbf{Z}_i(t)^{\otimes l}\}}{\mathbb{E}\{a_i(t)\}} \right\|_{\max} \leq (K_3/2)^l.$$

As a result, the maximal norms defined in (2.4) *and* (2.32) *,*

$$\sup_{t \in [0,t^*]} \max \left\{ \left\| \frac{\mathbf{S}^{(l)}(t,\beta)}{S^{(0)}(t,\beta)} \right\|_\infty, \left\| \frac{\widetilde{\mathbf{S}}^{(l)}(t,\beta)}{\widetilde{S}^{(0)}(t,\beta)} \right\|_\infty, \left\| \frac{\mathbf{s}^{(l)}(t,\beta)}{s^{(0)}(t,\beta)} \right\|_\infty \right\} \leq (K_3/2)^l,$$

are all uniformly bounded.

Lemma 16. *Let* K_4 *and* ρ_2 *be defined as in* (2.41)*. Define*

$$\widetilde{S}^{(0)}(t; K_4) = n^{-1} \sum_{i=1}^{n} I(C_i \geq t^*) Y_i(t^*) \min\{K_4, e^{\beta^{o\top} \mathbf{Z}_i(t)}\}. \tag{2.58}$$

Let $T^1_{(1)},\ldots,T^1_{(K_T)}$ be the observed type-1 events. Under (C1), the event

$$\Omega_1 = \left\{ n^{-1}\sum_{i=1} I(X_i \geq t^*) \geq \rho_2/(2K_4), \sup_{k\in 1\ldots K_T} \widetilde{S}^{(0)}\left(T^1_{(k)}; K_4\right) \geq \rho_2/2 \right\} \tag{2.59}$$

occurs with probability at least $1-e^{-n\rho_2^2/(2K_4^2)}-ne^{-n(\rho_2-2/n)^2/(8K_4^2)}$.

On Ω_1, we have $\sup_{k\in 1\ldots K_T}\widetilde{S}^{(0)}(T^1_{(k)}) \geq \rho_2/2$.

Lemma 17. *Let $Q_1(\varepsilon) = e^{K_6K_7\|\boldsymbol{\beta}^o\|_\infty\rho_4}\log(n/\varepsilon)/(\rho_4\rho_1)$ be defined as in* (2.43)*. Under (C3), the event*

$$\Omega_2(\varepsilon) = \left\{ \sup_{i=1,\ldots,n}\sup_{t\in[0,t^*]} I(\delta_i\varepsilon_i > 1)e^{\boldsymbol{\beta}^{o\top}\mathbf{Z}_i(t)} < Q_1(\varepsilon) \right\} \tag{2.60}$$

occurs with probability at least $1-\varepsilon$.

Lemma 18. *Define the IPW weights with true $G(t)$, $\widetilde{\omega}_i(t) = r_i(t)G(t)/G(X_i\wedge t)$, as in* (2.33) *and*

$$Q_7(n,p,\varepsilon) = 4(K_4/\rho_2)^2\left\{(1+K_1t^*)\sqrt{4\log(2/\varepsilon)/n}+K_1t^*/n\right\}. \tag{2.61}$$

Under (C1),

$$\Omega_3(\varepsilon) = \left\{ \sup_{t\in[0,t^*]}\sup_{t\in[0,t^*]} |\omega_i(t)-\widetilde{\omega}_i(t)| \leq Q_7(n,p,\varepsilon) \right\} \tag{2.62}$$

occurs on event Ω_1 with probability at least $\Pr(\Omega_1)-\varepsilon$.

Lemma 19. *Define*

$$\boldsymbol{\Delta}^{(l)}(t) = \mathbf{S}^{(l)}(t,\boldsymbol{\beta}^o) - \widetilde{\mathbf{S}}^{(l)}(t,\boldsymbol{\beta}^o),$$

with $\mathbf{S}^{(l)}$ and $\widetilde{\mathbf{S}}^{(l)}$ defined in (2.4) *and* (2.32)*. Let $T^1_{(1)},\ldots,T^1_{(K_T)}$ be the observed type-1 events for some $K_T \leq n$. Denote $Q_1(\varepsilon) = e^{K_6K_7\|\boldsymbol{\beta}^o\|_\infty\rho_4}\log(n/\varepsilon)/(\rho_4\rho_1)$ and*

$$Q_2^{(l)}(n,p,\varepsilon) = \frac{Q_1(\varepsilon)K_3^l}{2^l}\left\{\frac{4K_4^2(1+K_1t^*)}{\rho_2^2}\sqrt{\frac{4\log(2/\varepsilon)}{n}}+\frac{4K_4^2K_1t^*}{\rho_2^2 n}+\sqrt{\frac{2\log(2np^l/\varepsilon)}{n}}+\frac{1}{n}\right\}$$

as in (2.43) *and* (2.44)*. Under (C1) and (C3),*

$$\Omega_4(\varepsilon) = \left\{ \max_{l=0,1,2} \sup_{k\in 1\ldots K_T} \left\| \boldsymbol{\Delta}^{(l)}\left(T^1_{(k)}\right)\right\|_{\max} \le Q_2^{(l)}(n,p,\varepsilon) \right\} \cap \Omega_1 \cap \Omega_2(\varepsilon) \cap \Omega_3(\varepsilon), \qquad (2.63)$$

with Ω_1, $\Omega_2(\varepsilon)$ *and* $\Omega_3(\varepsilon)$ *defined in Lemmas 16, 17 and 18, occurs with probability at least* $1 - e^{-n\rho_2^2/(2K_4^2)} - ne^{-n(\rho_2-2/n)^2/(8K_4^2)} - 5\varepsilon$.

On $\Omega_4(\varepsilon)$*, we have for* $l = 1,2$,

$$\sup_{k\in 1\ldots K_T} \left\| \frac{\mathbf{S}^{(l)}\left(T^1_{(k)},\boldsymbol{\beta}^o\right)}{S^{(0)}\left(T^1_{(k)},\boldsymbol{\beta}^o\right)} - \frac{\widetilde{\mathbf{S}}^{(l)}\left(T^1_{(k)},\boldsymbol{\beta}^o\right)}{\widetilde{S}^{(0)}\left(T^1_{(k)},\boldsymbol{\beta}^o\right)} \right\|_{\max} \le 2\{Q_2^{(l)}(n,p,\varepsilon) + (K_3/2)^l Q_2^{(0)}(n,p,\varepsilon)\}/\rho_2.$$

Lemma 20. *Denote* $\boldsymbol{\Delta}^{(l)}(t) = \mathbf{S}^{(l)}(t,\boldsymbol{\beta}^o) - \widetilde{\mathbf{S}}^{(l)}(t,\boldsymbol{\beta}^o)$ *as in Lemma 19, with* $\mathbf{S}^{(l)}(t,\boldsymbol{\beta}^o)$ *and* $\widetilde{\mathbf{S}}^{(l)}(t,\boldsymbol{\beta}^o)$ *defined in* (2.4) *and* (2.32)*, respectively. Under (C1), (D1) - (D3) and (D4),*

(i) $\sup_{t\in[0,t^*]} \|\Delta^{(0)}(t)\|_{\max} = O_p\left(\sqrt{\log(n)/n}\right)$;

$\sup_{l=1,2}\sup_{t\in[0,t^*]} \|\boldsymbol{\Delta}^{(l)}(t)\|_{\max}$, $\sup_{t\in[0,t^*]} \|\bar{\mathbf{Z}}(t,\boldsymbol{\beta}^o) - \widetilde{\mathbf{Z}}(t,\boldsymbol{\beta}^o)\|_\infty$,

$\sup_{t\in[0,t^*]} \|\widetilde{\mathbf{Z}}(t,\boldsymbol{\beta}^o) - \boldsymbol{\mu}(t)\|_\infty$ *and* $\sup_{t\in[0,t^*]} \|\bar{\mathbf{Z}}(t,\boldsymbol{\beta}^o) - \boldsymbol{\mu}(t)\|_\infty$ *are all* $O_p\left(\sqrt{\log(p)/n}\right)$;

(ii) Define

$$\Delta_i(t) = \{\omega_i(t) - I(C_i > t)\}Y_i(t). \qquad (2.64)$$

Let $\boldsymbol{\phi}(\mathbf{Z})$ *be a differentiable operator* $\mathbb{R}^p \mapsto \mathbb{R}^q$ *uniformly bounded by* $K_\phi \asymp 1$ *with* $\|\nabla\boldsymbol{\phi}(\mathbf{Z})\|_1 < L_h \asymp 1$*, and* $\mathbf{g}(t)$ *be a* $\mathcal{F}^*_{t^-}$ *adapted process in* $\mathbb{R}^{q'}$ *with bound in maximal norm uniformly in time* $\sup_{t\in[0,t^*]} \|\mathbf{g}(t)\|_{\max} \le K_g \asymp 1$*. Whenever* $qq' = p$*, we have*

$$\left\| n^{-1/2}\sum_{i=1}^n \int_0^{t^*} n^{-1}\sum_{j=1}^n \Delta_j(t)\boldsymbol{\phi}(\mathbf{Z}_j(t))\mathbf{g}(t)^\top I(C_i \ge t) dM_i^1(t) \right\|_{\max} = o_p(1); \qquad (2.65)$$

(iii) for any $\widetilde{\boldsymbol{\beta}} \in \mathbb{R}^p$, $\sup_{t\in[0,t^*]} \|\bar{\mathbf{Z}}(t,\boldsymbol{\beta}^o) - \bar{\mathbf{Z}}(t,\widetilde{\boldsymbol{\beta}})\|_\infty = O_p(\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^o\|_1)$*; if* $\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^o\|_1 = o_p(1)$,

$$\sup_{i=1,\ldots,n}\sup_{t\in[0,t^*]} \left| \frac{e^{\boldsymbol{\beta}^{o\top}\mathbf{Z}_i(t)}}{S^{(0)}(t,\boldsymbol{\beta}^o)} - \frac{e^{\widetilde{\boldsymbol{\beta}}^\top\mathbf{Z}_i(t)}}{S^{(0)}(t,\widetilde{\boldsymbol{\beta}})} \right| = O_p(\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^o\|_1).$$

Lemma 21. *Let $S^{(0)}$ and $\widetilde{S}^{(0)}$ be defined as in* (2.4) *and* (2.32)*, respectively. Under (C1) and (D1),* $\sup_{t\in[0,t^*]}|n/\{\sum_{i=1}^n I(X_i\geq t^*)\}|$, $\sup_{t\in[0,t^*]}|S^{(0)}(t,\boldsymbol{\beta}^o)^{-1}|$ *and* $\sup_{t\in[0,t^*]}|\widetilde{S}^{(0)}(t,\boldsymbol{\beta}^o)^{-1}|$ *are all* $O_p(1)$.

Lemma 22. *Let Γ_j, $\widehat{\boldsymbol{\beta}}$ and $\boldsymbol{\gamma}_j^*$ be defined as in* (2.25)*,* (2.5) *and* (2.20)*, respectively. On the event*

$$\Omega_5(\lambda,\xi_j):=\left\{\left\|\nabla_{\boldsymbol{\gamma}}\boldsymbol{\Gamma}_j\big(\boldsymbol{\gamma}_j^*,\widehat{\boldsymbol{\beta}}\big)\right\|_\infty\leq(\xi_j-1)\lambda_j/(\xi_j+1),\forall j=1,\ldots,p\right\},\tag{2.66}$$

we have under (D2)

(i) the estimation error $\widetilde{\boldsymbol{\gamma}}_j:=\widehat{\boldsymbol{\gamma}}_j-\boldsymbol{\gamma}_j^*$ belongs to the cone

$$\mathcal{C}_j(\xi_j,O_j):=\{\mathbf{v}\in\mathbb{R}^{p-1}:\|\mathbf{v}_{O_j^c}\|_1\leq\xi_j\|\mathbf{v}_{O_j}\|_1\}\tag{2.67}$$

(ii) and $\|\widehat{\boldsymbol{\gamma}}_j-\boldsymbol{\gamma}_j^*\|_1\leq\{s_j\lambda_j(\xi_j+1)\}/\{2\kappa_j(\xi_j,O_j)^2\}$, with compatibility factor

$$\kappa_j(\xi_j,O_j)=\sup_{0\neq\mathbf{g}\in\mathcal{C}_j(\xi_j,O_j)}\frac{\sqrt{s_j\mathbf{g}^\top\nabla^2_{\boldsymbol{\gamma}}\boldsymbol{\Gamma}(\boldsymbol{\gamma}^*,\widehat{\boldsymbol{\beta}})\mathbf{g}}}{\|\mathbf{g}_{O_j}\|_1}\tag{2.68}$$

for all $j=1,\ldots,p$.

Lemma 23. *Let Γ_j, $\widehat{\boldsymbol{\beta}}$ and $\boldsymbol{\gamma}_j^*$ be defined as in* (2.25)*,* (2.5) *and* (2.20)*, respectively. Under (C1) and (D1)-(D4),* $\max_{j=1,\ldots,p}\left\|\nabla_{\boldsymbol{\gamma}}\boldsymbol{\Gamma}_j\big(\boldsymbol{\gamma}_j^*,\widehat{\boldsymbol{\beta}}\big)\right\|_\infty=O_p\left(\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^o\|_1+\sqrt{\log(p)/n}\right)$.

Lemma 24. *Let $\widehat{\boldsymbol{\Sigma}}$, $\boldsymbol{\Sigma}$, $\ddot{\mathbf{m}}$ be defined as in* (2.23)*,* (2.18) *and* (2.17)*, respectively. Under (C1) and (D1)-(D4),*

(i) $\left\|\widehat{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma}\right\|_{\max}=O_p\left(s_o\sqrt{\log(p)/n}\right)$;

(ii) for any $\widetilde{\boldsymbol{\beta}}$ such that $\|\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta}^o\|_1=o_p(1)$,

$$\left\|-\ddot{\mathbf{m}}(\widetilde{\boldsymbol{\beta}})-\boldsymbol{\Sigma}\right\|_{\max}=O_p\left(\|\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta}^o\|_1+\sqrt{\log(p)/n}\right).$$

Lemma 25. *Let* $\kappa_j(\xi_j, O_j)$ *be define as in Lemma 22* (2.67)*. Under (C1) and (D1)-(D4), setting* $\xi_{\max} = \max_{j=1,\ldots,p} \xi_j \asymp 1$*, we have*

$$\Pr\left(\inf_j \kappa_j(\xi_j, O_j)^2 \geq \rho/2\right) \to 1.$$

Proof of Main Lemmas

Proof of Lemma 4. Let $T^1_{(1)}, \ldots, T^1_{(K_T)}$ be the observed type-1 events. We may decompose the score $\dot{\mathbf{m}}(\beta^o)$ as its martingale proxy plus an approximation error,

$$\dot{\mathbf{m}}(\beta^o) = \dot{\tilde{\mathbf{m}}}(\beta^o) + n^{-1} \sum_{k=1,\ldots,K_T} \left\{ \widetilde{\mathbf{Z}}\left(T^1_{(k)}, \beta^o\right) - \bar{\mathbf{Z}}\left(T^1_{(k)}, \beta^o\right) \right\},$$

with $\widetilde{\mathbf{Z}}$ and $\bar{\mathbf{Z}}$ defined in (2.4) and (2.32), respectively.

Recall that the counting process for observed type-1 event can be written as $N^o_i(t) = \int_0^t I(C_i \geq u) dN^1_i(t)$. Moreover, $\dot{\tilde{\mathbf{m}}}(\beta^o)$ takes the form of the Cox model score with counting process $\{N^o_i(t)\}$ and at-risk process $\{I(C_i \geq t) Y_i(t)\}$. The "censoring complete" filtration $\mathcal{F}_t^*$ can also be equivalently generated by $\{N^o_i(t), I(C_i \geq t) Y_i(t), \mathbf{Z}_i(t)\}$. Thus, we may apply Lemma 3.3 in [HSY$^+$13] under (2.40) from (C1),

$$\Pr(\|\dot{\tilde{\mathbf{m}}}(\beta^o)\|_\infty > K_3 x) \leq 2pe^{-nx^2/2}.$$

Notice that the inequality is sharper than that in Lemma 14(i) because the compensator part of $\dot{\tilde{\mathbf{m}}}(\beta^o)$ is zero.

The concentration result for approximation error

$$\widetilde{\mathbf{Z}}\left(T^1_{(k)}, \beta^o\right) - \bar{\mathbf{Z}}\left(T^1_{(k)}, \beta^o\right) = \frac{\mathbf{S}^{(1)}\left(T^1_{(k)}, \beta^o\right)}{S^{(0)}\left(T^1_{(k)}, \beta^o\right)} - \frac{\widetilde{\mathbf{S}}^{(1)}\left(T^1_{(k)}, \beta^o\right)}{\widetilde{S}^{(0)}\left(T^1_{(k)}, \beta^o\right)}$$

is established in Lemma 19 on $\Omega_4(\varepsilon)$. We obtain the concentration inequality for $\dot{\mathbf{m}}(\boldsymbol{\beta}^o)$ by adding the bounds and tail probabilities together. $\square$

Proof of Lemma 5. Our strategy here is the same as that for Lemma 4. We first show that $\kappa\big(\xi, O; -\ddot{\mathbf{m}}(\boldsymbol{\beta}^o)\big)$ is lower bounded by $\kappa(\xi, O; -\ddot{\tilde{\mathbf{m}}}(\boldsymbol{\beta}^o))$ plus a diminishing error. Since $\ddot{\tilde{\mathbf{m}}}(\boldsymbol{\beta}^o)$ takes the form of a Cox model Hessian, we then may apply the results from [HSY$^+$13].

By Lemma 4.1 in [HSY$^+$13] (for a similar result, see [vdGB09] Corollary 10.1),

$$\kappa^2(\xi, O; -\ddot{\mathbf{m}}(\boldsymbol{\beta}^o)) \geq \kappa^2(\xi, O; -\ddot{\tilde{\mathbf{m}}}(\boldsymbol{\beta}^o)) - s_o(\xi+1)^2 \|\ddot{\mathbf{m}}(\boldsymbol{\beta}^o) - \ddot{\tilde{\mathbf{m}}}(\boldsymbol{\beta}^o)\|_{\max}.$$

Let $T^1_{(1)}, \ldots, T^1_{(K_T)}$ be the observed type-1 events. We can write $\ddot{\mathbf{m}}(\boldsymbol{\beta}^o) - \ddot{\tilde{\mathbf{m}}}(\boldsymbol{\beta}^o)$ as

$$-n^{-1} \sum_{k=1}^{K_T} \left[\frac{\mathbf{S}^{(2)}\left(T^1_{(k)}, \boldsymbol{\beta}^o\right)}{S^{(0)}\left(T^1_{(k)}, \boldsymbol{\beta}^o\right)} - \frac{\widetilde{\mathbf{S}}^{(2)}\left(T^1_{(k)}, \boldsymbol{\beta}^o\right)}{\widetilde{S}^{(0)}\left(T^1_{(k)}, \boldsymbol{\beta}^o\right)} - \bar{\mathbf{Z}}\left(T^1_{(k)}, \boldsymbol{\beta}^o\right)^{\otimes 2} + \widetilde{\mathbf{Z}}\left(T^1_{(k)}, \boldsymbol{\beta}^o\right)^{\otimes 2} \right],$$

with $\mathbf{S}^{(l)}$, $\widetilde{\mathbf{S}}^{(l)}$, $\widetilde{\mathbf{Z}}$ and $\bar{\mathbf{Z}}$ defined in (2.4) and (2.32). By Lemma 15, $\sup_{t\in[0,t^*]} \|\bar{\mathbf{Z}}(t, \boldsymbol{\beta}^o)\|_\infty$ and $\sup_{t\in[0,t^*]} \|\widetilde{\mathbf{Z}}(t, \boldsymbol{\beta}^o)\|_\infty$ are both bounded by $K_3/2$. On the $\Omega_4(\varepsilon)$ as defined in Lemma 19, we apply Lemma 19 once with $l = 2$ and twice with $l = 1$ to get

$$\|\ddot{\mathbf{m}}(\boldsymbol{\beta}^o) - \ddot{\tilde{\mathbf{m}}}(\boldsymbol{\beta}^o)\|_{\max} \leq \left\{ 2Q_2^{(2)}(n, p, \varepsilon) + 4K_3 Q_2^{(1)}(n, p, \varepsilon) + (5/2)K_3^2 Q_2^{(0)}(n, p, \varepsilon) \right\} / \rho_2,$$

with $Q_2^{(l)}(n, p, \varepsilon)$ defined in (2.44).

Our (C1) and (C2) contains all the condition for Theorem 4.1 in [HSY$^+$13]. Hence, we may apply their result

$$\begin{aligned} \kappa^2(\xi, O; -\ddot{\tilde{\mathbf{m}}}(\boldsymbol{\beta}^o)) \geq{}& \kappa^2(\xi, O; \Sigma(K_4)) - s_o(\xi+1)^2 K_3^2 \\ & \times \left\{ (1 + t^* K_2) \sqrt{2 \log\left(p(p+1)/\varepsilon\right)/n} + (2/\rho_2) t^* K_2 Q_6(n, p, \varepsilon)^2 \right\} \end{aligned}$$

with probability at least $\Pr(\Omega_4(\varepsilon)) - 3\varepsilon$. We have bounded $\widetilde{S}^{(0)}(t;K_4)$ away from zero at all observed type-1 events in $\Omega_4(\varepsilon)$, so the $e^{-n\rho_2^2/(8K_4^2)}$ term is absorbed into $\Pr(\Omega_4(\varepsilon))$. □

Proof of Lemma 6. The notations in the proof are defined in Section 2.2.3. Denote

$$\boldsymbol{\Xi} = \int_0^{t^*} \{\mathbf{Z}(t) - \boldsymbol{\mu}(t)\} dN^o(t).$$

Without loss of generality, we set $j = 1$. Since we define $\boldsymbol{\gamma}_1^* = \operatorname{argmin}_{\boldsymbol{\gamma}} \bar{\Gamma}(\boldsymbol{\gamma})$ as the minimizer of a convex function, it must satisfy the first order condition

$$\nabla_{\boldsymbol{\gamma}} \bar{\Gamma}(\boldsymbol{\gamma}_1^*) = \mathbb{E}\left\{(\Xi_1 - \boldsymbol{\Xi}_{-1}^\top \boldsymbol{\gamma}_1^*)\boldsymbol{\Xi}_{-1}\right\} = \mathbf{0}_{p-1}.$$

Recall that $\tau_1^2 = \bar{\Gamma}(\boldsymbol{\gamma}_1^*)$. Applying the first order condition, we get

$$\tau_1^2 = \mathbb{E}\{\Xi_1 - \boldsymbol{\Xi}_{-1}^\top \boldsymbol{\gamma}_1^*\}^2 = \mathbb{E}\{(\Xi_1 - \boldsymbol{\Xi}_{-1}^\top \boldsymbol{\gamma}_1^*)\Xi_1\}.$$

We construct a vector $\boldsymbol{\theta}_1 = (1, -\boldsymbol{\gamma}_1^{*\top})^\top / \tau_1^2 \in \mathbb{R}^p$. Then, $\boldsymbol{\theta}_1$ satisfies

$$\boldsymbol{\theta}_1^\top \boldsymbol{\Sigma} = (1, -\boldsymbol{\gamma}_1^{*\top})\mathbb{E}\{\boldsymbol{\Xi}\boldsymbol{\Xi}^\top\}/\tau_1^2 = (1, \mathbf{0}_{p-1}^\top).$$

Hence, we have

$$(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_p)^\top = \boldsymbol{\Sigma}^{-1} = \boldsymbol{\Theta}.$$

We can directly bound

$$\|\boldsymbol{\gamma}_j^*\|_1 = \|\boldsymbol{\theta}_j / \Theta_{j,j}\|_1 - 1 \leq K - 1 < K.$$

By (D2), the minimal eigenvalue of $\boldsymbol{\Sigma}$ is at least ρ. We obtain through a spectral decomposition that the maximal eigenvalue of $\boldsymbol{\Theta} = \boldsymbol{\Sigma}^{-1}$ is at most ρ^{-1}. Hence, we have

$$\tau_j^2 = \left(\mathbf{e}_j^\top \boldsymbol{\Theta} \mathbf{e}_j\right)^{-1} \geq \rho$$

and

$$\|\Theta\|_1 \leq \max_{j=1,\ldots,p} \|\theta_j/\Theta j, j\| \max_{j=1,\ldots,p} |\Theta_{j,j}| \leq K/\rho.$$

□

Proof of Lemma 7. By Lemma 23, we may choose $\xi_1 = \cdots = \xi_p = 2$ and $\lambda_1 = \cdots = \lambda_p = \lambda_\varepsilon \asymp O_p(s_o\sqrt{\log(p)/n})$ such that $\Omega_5(\lambda, \xi_j)$ defined in Lemma 22 occurs with probability $1-\varepsilon$. Then, we establish the oracle inequality by Lemma 22,

$$\Pr\left(\max_{j=1,\ldots,p} \|\widehat{\gamma}_j - \gamma_j^*\|_1/s_j \leq \frac{2\lambda_\varepsilon}{\rho}\right) \geq \Pr\left(\min_{j=1,\ldots,p} \kappa_j(\xi_j, O_j)^2 \geq \rho/2\right) - \varepsilon.$$

We have shown that $\Pr\left(\min_{j=1,\ldots,p} \kappa_j(\xi_j, O_j)^2 \geq \rho/2\right)$ tends to one in Lemma 25. Hence, $\max_{j=1,\ldots,p} \|\widehat{\gamma}_j - \gamma_j^*\|_1 = O_p\left(s_o s_{\max}\sqrt{\log(p)/n}\right)$.

Define according to (2.55) $\Xi_i = \int_0^{t^*} \{\mathbf{Z}_i(t) - \boldsymbol{\mu}(t)\} dN_i^o(t)$. By Lemma 15, we have $\sup_{i=1,\ldots,n} \|\Xi_i\|_\infty \leq K$. We introduce

$$\widetilde{\Gamma}_j(\gamma) = n^{-1}\sum_{i=1}^n \{\Xi_j - \Xi_{i,-j}^\top \gamma_j^*\} = n^{-1}\sum_{i=1}^n \int_0^{t^*} \{Z_{ij}(t) - \mu_j(t) - \gamma^\top \mathbf{Z}_{i,-j}(t) + \gamma^\top \boldsymbol{\mu}_{-j}(t)\}^2 dN_i^o(t)$$

and decompose

$$\widehat{\tau}_j^2 - \tau_j^2 = \Gamma_j(\widehat{\gamma}_j, \widehat{\beta}) - \widetilde{\Gamma}_j(\gamma_j^*) + \widetilde{\Gamma}_j(\gamma_j^*) - \bar{\Gamma}_j(\gamma_j^*).$$

$\Gamma_j(\widehat{\gamma}_j, \widehat{\beta}) - \widetilde{\Gamma}_j(\gamma_j^*) = O_p\left(s_o s_j \sqrt{\log(p)/n}\right)$ by the results from Theorem 8, Lemma 20 and first part of this Lemma. Apparently, $\widetilde{\Gamma}_j(\gamma_j^*)$ is the average of i.i.d. terms. The expectation of the summands in $\widetilde{\Gamma}_j(\gamma_j^*)$ is defined as $\bar{\Gamma}_j(\gamma_j^*)$ in (2.55). Hence, we finish the proof by applying Lemma 11.

Along with Lemma 6, we can prove with the previous results in this Lemma, $\|\widehat{\Theta} - \Theta\|_1 = O_p\left(s_o s_{\max}\sqrt{\log(p)/n}\right)$. □

Proof of Lemma 8. We decompose

$$\sqrt{n}\mathbf{c}^\top \left\{ \Theta \dot{\mathbf{m}}(\beta^o) + \beta^o - \widehat{\mathbf{b}} \right\} \tag{2.69}$$

$$= \sqrt{n}\mathbf{c}^\top \{\Theta - \widehat{\Theta}\}\dot{\mathbf{m}}(\widehat{\beta}) + \sqrt{n}\mathbf{c}^\top \Theta\{\dot{\mathbf{m}}(\beta^o) - \dot{\mathbf{m}}(\widehat{\beta})\} + \sqrt{n}\mathbf{c}^\top(\beta^o - \widehat{\beta}). \tag{2.70}$$

By Lemma 7, $\|\Theta - \widehat{\Theta}\|_1 = O_p(s_o s_{\max}\sqrt{\log(p)/n})$. Each summand in $\dot{\mathbf{m}}(\widehat{\beta})$ is the integral of $\mathbf{Z}_i(t)$ minus a weighted average $\bar{\mathbf{Z}}(t,\widehat{\beta})$ over a counting measure $dN_i^o(t)$. By the KKT condition and Theorem 8, $\|\dot{\mathbf{m}}(\widehat{\beta})\|_\infty \asymp \lambda \asymp O(\sqrt{\log(p)/n})$. Putting these together, we obtain

$$\sqrt{n}|\mathbf{c}^\top \{\Theta - \widehat{\Theta}\}\dot{\mathbf{m}}(\widehat{\beta})| \le \sqrt{n}\|\mathbf{c}\|_1 \|\Theta - \widehat{\Theta}\|_1 \|\dot{\mathbf{m}}(\widehat{\beta})\|_\infty \tag{2.71}$$

$$= O_p\left(s_o s_{\max}\log(p)/\sqrt{n}\right) = o_p(1). \tag{2.72}$$

By the KKT condition and Theorem 6, $\|\dot{\mathbf{m}}(\widehat{\beta})\| \le \lambda \asymp n^{-(1/2-d)}$. Hence, the first term in (2.69) is $o_p(1)$. Like in the proof of Lemma 9, we have $\|\mathbf{c}^\top \Theta\|_1 \le \|\mathbf{c}\|_1 \|\Theta\|_1 \le Ke^K/\rho_2$ from Lemma 6.

Define $\beta_r = \beta^o + r(\widehat{\beta} - \beta^o)$. Applying mean value theorem to $h(r) = \mathbf{c}^\top \Theta \dot{\mathbf{m}}(\beta_r)$, we get

$$\mathbf{c}^\top \Theta \dot{\mathbf{m}}(\beta^o) - \mathbf{c}^\top \Theta \dot{\mathbf{m}}(\widehat{\beta}) = -h'(\widetilde{r}) = -\mathbf{c}^\top \Theta \ddot{\mathbf{m}}(\beta_{\widetilde{r}})(\widehat{\beta} - \beta^o)$$

for some $\widetilde{r} \in [0,1]$. By Theorem 8, we have

$$\|\beta_{\widetilde{r}} - \beta^o\|_1 = \widetilde{r}\|\widehat{\beta} - \beta^o\|_1 = O_p\left(s_o\sqrt{\log(p)/n}\right).$$

By Lemma 24(ii), $\|-\ddot{\mathbf{m}}(\beta_{\widetilde{r}}) - \Sigma\|_{\max} = O_p\left(s_o\sqrt{\log(p)/n}\right)$. Along with Theorem 8 and Lemma 6, we have

$$\sqrt{n}|\mathbf{c}^\top \Theta\{\dot{\mathbf{m}}(\beta^o) - \dot{\mathbf{m}}(\widehat{\beta})\} + \mathbf{c}^\top(\beta^o - \widehat{\beta})| = \sqrt{n}|\mathbf{c}^\top \Theta\{\Sigma + \ddot{\mathbf{m}}(\beta_{\widetilde{r}})\}(\beta^o - \widehat{\beta})|$$

$$\leq \sqrt{n}\|\mathbf{c}\|_1 \|\boldsymbol{\Theta}\|_1 \| - \ddot{\mathbf{m}}(\boldsymbol{\beta}_{\tilde{r}}) - \boldsymbol{\Sigma}\|_{\max} \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^o\|_1$$

$$= O_p\left(s_o^2 \log(p)/\sqrt{n}\right).$$

□

Proof of Lemma 9. Since $\omega_i(t)Y_i(t) \neq I(C_i \geq t)Y_i(t)$ implies $\varepsilon_i > 1$ thus $N_i^1(t^*) = 0$, we have the equivalence $dN_i^o(t) = \omega_i(t)dN_i^1(t) = I(C_i \geq t)dN_i^1(t)$. Recall for the following calculation that

$$\mathbf{S}^{(l)}(t,\boldsymbol{\beta}^o) = n^{-1}\sum_{i=1}^n \omega_i(t)Y_i(t)e^{\boldsymbol{\beta}^{o\top}\mathbf{Z}_i(t)}\mathbf{Z}_i(t)^{\otimes l},$$

$$\widetilde{\mathbf{S}}^{(l)}(t,\boldsymbol{\beta}^o) = n^{-1}\sum_{i=1}^n I(C_i \geq t)Y_i(t)e^{\boldsymbol{\beta}^{o\top}\mathbf{Z}_i(t)}\mathbf{Z}_i(t)^{\otimes l},$$

$$\boldsymbol{\Delta}^{(l)}(t) = \mathbf{S}^{(l)}(t,\boldsymbol{\beta}^o) - \widetilde{\mathbf{S}}^{(l)}(t,\boldsymbol{\beta}^o),$$

$$\mathbb{E}\{\mathbf{S}^{(l)}(t,\boldsymbol{\beta}^o)\} = \mathbb{E}\{\widetilde{\mathbf{S}}^{(l)}(t,\boldsymbol{\beta}^o)\} = \mathbf{s}^{(l)}(t,\boldsymbol{\beta}^o)$$

$$\bar{\mathbf{Z}}(t,\boldsymbol{\beta}^o) = \mathbf{S}^{(1)}(t,\boldsymbol{\beta}^o)/S^{(0)}(t,\boldsymbol{\beta}^o), \quad \widetilde{\mathbf{Z}}(t,\boldsymbol{\beta}^o) = \widetilde{\mathbf{S}}^{(1)}(t,\boldsymbol{\beta}^o)/\widetilde{S}^{(0)}(t,\boldsymbol{\beta}^o),$$

$$\boldsymbol{\mu}(t) = \mathbf{s}^{(1)}(t,\boldsymbol{\beta}^o)/s^{(0)}(t,\boldsymbol{\beta}^o), \quad Y_i(t) = 1 - N_i^1(t-)$$

and $M_i^1(t) = N_i^1(t) - \int_0^t Y_i(t)e^{\boldsymbol{\beta}^{o\top}\mathbf{Z}_i(u)}h_0^1(u)du.$

We decompose

$$\begin{aligned}
\sqrt{n}\dot{\mathbf{m}}(\boldsymbol{\beta}^o) =& n^{-1/2}\sum_{i=1}^n \int_0^{t^*} \left\{\mathbf{Z}_i(t) - \bar{\mathbf{Z}}(t,\boldsymbol{\beta}^o)\right\} dN_i^o(t) \\
=& n^{-1/2}\sum_{i=1}^n \int_0^{t^*} \left\{\mathbf{Z}_i(t) - \bar{\mathbf{Z}}(t,\boldsymbol{\beta}^o)\right\} \omega_i(t)dM_i^1(t) \\
=& n^{-1/2}\sum_{i=1}^n \int_0^{t^*} \left\{\boldsymbol{\mu}(t) - \widetilde{\mathbf{Z}}(t,\boldsymbol{\beta}^o)\right\} I(C_i \geq t)dM_i^1(t) \\
&+ n^{-1/2}\sum_{i=1}^n \int_0^{t^*} \left\{\widetilde{\mathbf{Z}}(t,\boldsymbol{\beta}^o) - \bar{\mathbf{Z}}(t,\boldsymbol{\beta}^o)\right\} I(C_i \geq t)dM_i^1(t)
\end{aligned}$$

$$
\begin{aligned}
&+n^{-1/2}\sum_{i=1}^{n}\int_0^{t^*}\left\{\bar{\mathbf{Z}}(t,\boldsymbol{\beta}^o)-\boldsymbol{\mu}(t)\right\}\Delta^{(0)}(t)h_0^1(t)dt\\
&+n^{-1/2}\sum_{i=1}^{n}\int_0^{t^*}\left\{\mathbf{Z}_i(t)-\boldsymbol{\mu}(t)\right\}\omega_i(t)dM_i^1(t)\\
\triangleq& I_1+I_2+I_3+I_4.
\end{aligned}
$$

Notice that I_1 is a $\mathcal{F}_t^*$ martingale. We have $\|\boldsymbol{\mu}(t)-\widetilde{\mathbf{Z}}(t,\boldsymbol{\beta}^o)\|_\infty = O_p(\sqrt{\log(p)/n})$ from Lemma 20(i). Hence, we can apply Lemma 14(ii) to get $\|I_1\|_\infty = \sqrt{n}O_p(\sqrt{\log(p)/n}^2) = o_p(1)$.

We further decompose I_2 into 3 terms

$$
\begin{aligned}
&-n^{-1/2}\sum_{i=1}^{n}\int_0^{t^*}\frac{\boldsymbol{\Delta}^{(1)}(t)}{\widetilde{S}^{(0)}(t,\boldsymbol{\beta}^o)}I(C_i\geq t)dM_i^1(t)-n^{-1/2}\sum_{i=1}^{n}\int_0^{t^*}\frac{\Delta^{(0)}(t)}{\widetilde{S}^{(0)}(t,\boldsymbol{\beta}^o)}\boldsymbol{\mu}(t)I(C_i\geq t)dM_i^1(t)\\
&+n^{-1/2}\sum_{i=1}^{n}\int_0^{t^*}\frac{\Delta^{(0)}(t)}{\widetilde{S}^{(0)}(t,\boldsymbol{\beta}^o)}\{\boldsymbol{\mu}(t)-\bar{\mathbf{Z}}(t,\boldsymbol{\beta}^o)\}I(C_i\geq t)dM_i^1(t)\\
\triangleq& I_2'+I_2''+I_2'''.
\end{aligned}
$$

By (D1) and (D3), each $M_i^1(t)$ has one jump at observed event time and $e^K K-$Lipschitz elsewhere. Since the $\{C_i, T_i^1 : i=1,\ldots,n\}$ is a set of independent continuous random variables, there is no tie among them with probability one. Hence, we may modify the integrand in I_2' and I_2'' at observed censoring times without changing the integral. Replacing $\boldsymbol{\Delta}^{(l)}(t)$ with $n^{-1}\sum_{j=1}^n \Delta_i(t)e^{\boldsymbol{\beta}^{o\top}\mathbf{Z}_i(t)}\mathbf{Z}_i(t)^{\otimes l}$, we can apply Lemma 20(ii) to get that $\|I_2'\|_\infty$ and $\|I_2''\|_\infty$ are both $o_p(1)$.

The total variation of $M_i^1(t)$ is at most $\max\{1, e^K K t^*\} \asymp 1$. By Lemma 20(i), we have $\|\boldsymbol{\Delta}^{(0)}(t)\{\boldsymbol{\mu}(t)-\bar{\mathbf{Z}}(t,\boldsymbol{\beta}^o)\}\|_\infty = O_p(\sqrt{\log(n)\log(p)}/n)$. Hence, we obtain the order for I_2''', $\|I_2'''\|_\infty = O_p(\sqrt{\log(n)\log(p)}/n) = o_p(1)$. Similarly, we obtain $\|I_3\|_\infty = O_p(\sqrt{\log(n)\log(p)}/n) = o_p(1)$.

Besides the one in Lemma 18, $\omega_i(t) - \widetilde{\omega}_i(t)$ has another martingale representation. Denote the Nelson-Aalen estimator

$$\widehat{H}^c(t) = \sum_{i=1}^{n} \int_0^t \frac{I(X_i \geq u)}{\sum_{j=1}^n I(X_j \geq u)} dN_i^c(u).$$

We have a $\mathcal{F}_t$ martingale

$$\overline{M^c}(t) = \widehat{H}^c(t) - \int_0^t h^c(u)du = \sum_{i=1}^{n} \int_0^t \frac{I(X_i \geq u)}{\sum_{j=1}^n I(X_j \geq u)} dM_i^c(u).$$

By Lemma 14(i), $\sup_{t\in[0,t^*]} |\overline{M^c}(t)| = O_p\left(n^{-1/2}\right)$ For $t > X_i$ and $\delta_i\varepsilon_i > 1$,

$$\omega_i(t) - \widetilde{\omega}_i(t) = -\widetilde{\omega}_i(t) \int_0^t I(u > X_i) d\overline{M^c}(u) + R_i(t)$$

with an error

$$R_i(t) = \frac{\widehat{G}(t)}{\widehat{G}(X_i)} - \exp\left\{\widehat{H}^c(X_i) - \widehat{H}^c(t)\right\} + \frac{G(t)}{G(X_i)} \left[e^{-\int_0^t I(u>X_i) d\overline{M^c}(u)} + \int_0^t I(u > X_i) d\overline{M^c}(u)\right].$$

It is the discrepancy between the Kaplan-Meier and the Nelson-Aalen plus a second order Tailer expansion remainder. We shall show that it is $O_p(1/n)$. Since

$$\left|\int_0^t I(u > X_i) d\overline{M^c}(u)\right| \leq 2 \sup_{t\in[0,t^*]} |\overline{M^c}(t)| = O_p\left(n^{-1/2}\right),$$

the second order remainder

$$\left|e^{-\int_0^t I(u>X_i) d\overline{M^c}(u)} + \int_0^t I(u > X_i) d\overline{M^c}(u)\right| = O_p(1/n).$$

Under (C1), $\{\sum_{i=1}^n I(X_i \geq t)\}^{-1} \leq \{\sum_{i=1}^n I(X_i \geq t^*)\}^{-1} = O_p(1/n)$. Let c_k be an observed censoring time. The increment in $-\log(\widehat{G}(t)) - \widehat{H}^c(t)$ at c_k is a second order remainder

$$\log\left(1 - \frac{1}{\sum_{i=1}^n I(X_i \geq c_k)}\right) - \frac{1}{\sum_{i=1}^n I(X_i \geq c_k)} = O_p\left(n^{-2}\right).$$

Hence, $\sup_{t\in[0,t^*]}|-\log(\widehat{G}(t))-\widehat{H}^c(t)|=O_p(1/n)$. Applying the Mean Value Theorem, we obtain $\sup_{t\in[0,t^*]}|\widehat{G}(t)-\exp\{-\widehat{H}^c(t)\}|=O_p(1/n)$. Under (C1), $G(t)\geq G(t^*)$ is bounded away from zero, and $-\log(G(t))\leq-\log(G(t^*))$ is bounded from above. We have shown that both $\widehat{G}(t)$ and $\widehat{H}^c(t)$ are uniformly $\sqrt{n}$ consistent. We obtain that $\widehat{G}(X_i)$ is bounded away from zero and $\widehat{H}^c(t)$ is bounded with probability tending to one. Putting these together, we obtain

$$\sup_{i=1,\ldots,n}\sup_{t\in[0,t^*]}|R_i(t)|=O_p(1/n).$$

Define

$$\widetilde{\mathbf{q}}(t)=n^{-1}\sum_{i=1}^{n}I(t\geq X_i)\int_t^{t^*}\{\mathbf{Z}_i(u)-\boldsymbol{\mu}(u)\}\,\widetilde{\omega}_i(u)dM_i^1(u),$$

$\widehat{\pi}(t)=n^{-1}\sum_{i=1}^{n}I(X_i\geq t)$ and $\mathbf{q}(t)=\mathbb{E}\{\widetilde{\mathbf{q}}(t)\}$, $\pi(t)=\mathbb{E}\{\widehat{\pi}(t)\}$. We write I_4 as i.i.d. sum plus error through integration by parts,

$$\begin{aligned}
&n^{-1/2}\sum_{i=1}^{n}\int_0^{t^*}\{\mathbf{Z}_i(t)-\boldsymbol{\mu}(t)\}\,\widetilde{\omega}_i(t)dM_i^1(t)\\
&\quad+n^{-1/2}\sum_{i=1}^{n}\int_0^{t^*}\{\mathbf{Z}_i(t)-\boldsymbol{\mu}(t)\}\,\{\omega_i(t)-\widetilde{\omega}_i(t)\}dM_i^1(t)\\
=&n^{-1/2}\sum_{i=1}^{n}\int_0^{t^*}\{\mathbf{Z}_i(t)-\boldsymbol{\mu}(t)\}\,\widetilde{\omega}_i(t)dM_i^1(t)+n^{-1/2}\sum_{i=1}^{n}\int_0^{t^*}\{\mathbf{Z}_i(t)-\boldsymbol{\mu}(t)\}\,R_i(t)dM_i^1(t)\\
&-n^{-1/2}\sum_{k=1}^{n}\int_0^{t^*}\frac{\mathbf{q}(t)}{\pi(t)}I(X_k\geq u)dM_k^c(t)\\
&\quad+n^{-1/2}\sum_{k=1}^{n}\int_0^{t^*}\frac{\mathbf{q}(t)}{\widehat{\pi}(t)\pi(t)}\{\widehat{\pi}(t)-\pi(t)\}I(X_k\geq u)dM_k^c(t)\\
&+n^{-1/2}\{\mathbf{q}(0)-\widetilde{\mathbf{q}}(0)\}\sum_{k=1}^{n}\int_0^{t^*}\frac{1}{\widehat{\pi}(t)}I(X_k\geq u)dM_k^c(t)\\
&-n^{-1/2}\sum_{k=1}^{n}\int_0^{t^*}\frac{\{\mathbf{q}(0)-\mathbf{q}(t)-\widetilde{\mathbf{q}}(0)+\widetilde{\mathbf{q}}(t)\}}{\widehat{\pi}(t)}I(X_k\geq u)dM_k^c(t)\\
\triangleq&I_4^{(1)}+I_4^{(2)}+I_4^{(3)}+I_4^{(4)}+I_4^{(5)}+I_4^{(6)}.
\end{aligned}$$

$I_4^{(1)} + I_4^{(3)}$ is already a sum of i.i.d.. We have shown that $\sup_{t\in[0,t^*]} |R_i(t)| = O_p(1/n)$. Hence, we have $\|I_4^{(2)}\|_\infty = O_p\left(n^{-1/2}\right) = o_p(1)$. $I(t \geq X_i) \int_t^{t^*} \{\mathbf{Z}_i(u) - \boldsymbol{\mu}(u)\} \widetilde{\omega}_i(u) dM_i^1(u)$ is uniformly bounded by $K(Kt^* + 1)$. It has at most one jump and is $KK-$Lipschitz elsewhere. Hence, we can apply Lemma 13(ii) to get $\sup_{t\in[0,t^*]} \|\mathbf{q}(t) - \widetilde{\mathbf{q}}(t)\|_\infty = O_p(\sqrt{\log(p)/n})$ and $\sup_{t\in[0,t^*]} |\pi(t) - \widehat{\pi}(t)| = O_p(\sqrt{\log(n)/n})$. Notice that $I_4^{(4)}$, $I_4^{(6)}$ and $n^{-1}\sum_{k=1}^n \int_0^{t^*} \widehat{\pi}(t)^{-1} I(X_k \geq u) dM_k^c(t)$ in $I_4^{(5)}$ are all $\mathcal{F}_t$ martingales. We may apply Lemmas 14(i) and 14(ii) to obtain $I_4^{(4)} = O_p(\sqrt{\log(n)\log p/n}) = o_p(1)$, $I_4^{(5)} = O_p(\sqrt{\log p/n}) = o_p(1)$ and $I_4^{(6)} = O_p(\log p/\sqrt{n}) = o_p(1)$.

By Lemma 6, we can bound the l_1 norm of $\mathbf{c}^\top \boldsymbol{\Theta}$ by

$$\|\mathbf{c}^\top \boldsymbol{\Theta}\|_1 = \sum_{i=1}^p \sum_{j=1}^p |c_i||\Theta_{ij}| \leq \sum_{i=1}^p |c_i| K/\rho = K/\rho.$$

Finally, we write $\mathbf{c}^\top \boldsymbol{\Theta} \dot{\mathbf{m}}(\boldsymbol{\beta}^o)$ as i.i.d. sum

$$\begin{aligned}
& n^{-1/2} \sum_{i=1}^n \mathbf{c}^\top \boldsymbol{\Theta} \left[\int_0^{t^*} \{\mathbf{Z}_i(t) - \boldsymbol{\mu}(t)\} \widetilde{\omega}_i(t) dM_i^1(t) - \int_0^{t^*} \frac{\mathbf{q}(t)}{\pi(t)} I(X_i \geq u) dM_i^c(t) \right] + o_p(1) \\
\triangleq & n^{-1/2} \sum_{i=1}^n \mathbf{c}^\top \boldsymbol{\Theta} \{\boldsymbol{\eta}_i - \boldsymbol{\psi}_i\} + o_p(1).
\end{aligned}$$

We have $\mathbb{E}\{\mathbf{c}^\top \boldsymbol{\Theta} \boldsymbol{\eta}_i\} = 0$ because of its martingale structure. We show $\mathbb{E}\{\mathbf{c}^\top \boldsymbol{\Theta} \boldsymbol{\psi}_i\} = 0$ again by introducing its martingale proxy

$$\begin{aligned}
\mathbb{E}\{\mathbf{c}^\top \boldsymbol{\Theta} \boldsymbol{\psi}_i\} = & \mathbb{E}\left[\int_0^{t^*} \mathbf{c}^\top \boldsymbol{\Theta} \{\mathbf{Z}_i(t) - \boldsymbol{\mu}(t)\} I(C_i \geq t) dM_i^1(t) \right] \\
& + \mathbb{E}\left[\int_0^{t^*} \mathbf{c}^\top \boldsymbol{\Theta} \{\mathbf{Z}_i(t) - \boldsymbol{\mu}(t)\} \mathbb{E}\{\widetilde{\omega}_i(t) - I(C_i \geq t) | T_i, \mathbf{Z}_i(\cdot)\} dM_i^1(t) \right].
\end{aligned}$$

The first term above is zero because of the martingale structure. The second term is zero because the IPW weights satisfy $\mathbb{E}\{\widetilde{\omega}_i(t) - I(C_i \geq t) | T_i, \mathbf{Z}_i(\cdot)\} = 0$. Each $\mathbf{c}^\top \boldsymbol{\Theta}\{\boldsymbol{\psi}_i - \boldsymbol{\eta}_i\}$ is mean zero and bounded by $K/\rho K(1 + Kt^*) + K/\rho K(1 + Kt^*)(1 + Kt^*)2e^K/\rho_2$ with probability

equaling one. The variance $\mathbf{c}^\top \Theta \mathcal{V} \Theta \mathbf{c}$ has a bounded and non-degenerating limit ν^2. Hence, $\{\mathbf{c}^\top \Theta(\psi_i - \eta_i) : i = 1, \ldots, n\}$ satisfies the Lindeberg condition.

By Lindeberg-Feller CLT,

$$\sqrt{n}\frac{\mathbf{c}^\top \Theta \dot{\mathbf{m}}(\beta^o)}{\sqrt{\mathbf{c}^\top \Theta \mathcal{V} \Theta \mathbf{c}}} = \frac{\mathbf{c}^\top \Theta \sum_{i=1}^n \{\eta_i - \psi_i\}}{\sqrt{n\mathbf{c}^\top \Theta \mathcal{V} \Theta \mathbf{c}}} + o_p(1) \xrightarrow{d} N(0,1).$$

We conclude the proof of the Lemma. □

Proof of Lemma 10. We define

$$\widetilde{\eta}_i = \int_0^{t^*} \{\mathbf{Z}_i(u) - \boldsymbol{\mu}(u)\}\widetilde{\omega}_i(u) d\widetilde{M}_i^1(u),$$

with

$$\widetilde{M}_i^1(t) = N_i^o(t) - n^{-1}\sum_{j=1}^n \int_0^t \frac{Y_i(u)e^{\beta^{o\top}\mathbf{Z}_i(u)}}{\widetilde{S}^{(0)}(u,\beta^o)} dN_j^o(u).$$

Under (D1) and (C1), the total variation of $\widetilde{M}_i^1(t)$ is at most $1 + 2e^{2K}/\rho_2$ with probability tending to one by Lemma 21. The difference between $\widetilde{\eta}_i$ and $\widehat{\eta}_i$ is

$$\begin{aligned}\widehat{\eta}_i - \widetilde{\eta}_i =& n^{-1}\sum_{j=1}^n \int_0^{t^*} \{\mathbf{Z}_i(u) - \bar{\mathbf{Z}}(u,\widehat{\beta})\}\omega_i(u)Y_i(u)\left\{\frac{e^{\beta^{o\top}\mathbf{Z}_i(u)}}{\widetilde{S}^{(0)}(u,\beta^o)} - \frac{e^{\widehat{\beta}^\top \mathbf{Z}_i(u)}}{S^{(0)}(u,\widehat{\beta})}\right\} dN_j^o(u) \\ &+ \int_0^{t^*} \{\boldsymbol{\mu}(u)\widetilde{\omega}_i(u) - \bar{\mathbf{Z}}(u,\widehat{\beta})\omega_i(u)\} d\widetilde{M}_i^1(u).\end{aligned}$$

By Lemmas 18, 20(i) and 20(iii), $\sup_{i=1,\ldots,n} \|\widehat{\eta}_i - \widetilde{\eta}_i\|_\infty = O_p\left(\|\widehat{\beta} - \beta^o\|_1 + \sqrt{\log(p)/n}\right)$.

Then, we study

$$\eta_i - \widetilde{\eta}_i = n^{-1}\sum_{j=1}^n \int_0^{t^*} \{\mathbf{Z}_i(u) - \boldsymbol{\mu}(u)\}\widetilde{\omega}_i(u) I(C_j \geq u) dM_j^1(u).$$

We have the bound $\|\mathbf{Z}_i(u) - \boldsymbol{\mu}(u)\|_\infty \leq K$ from Lemma 15. $\widetilde{\omega}_i(u)$ is not $\mathcal{F}_t^*$ measurable, but we can define a new filtration $\mathcal{F}_{i,t}^* = \sigma\{X_i, \delta_i, \varepsilon_i, \mathbf{Z}_i(\cdot), I(C_j \geq u), N_j^1(u), \mathbf{Z}_j(\cdot) : u \leq t, j \neq i\}$ for each

i, such that

$$n^{-1}\sum_{j\neq i}\int_0^{t^*}\{\mathbf{Z}_i(u)-\boldsymbol{\mu}(u)\}\widetilde{\omega}_i(u)I(C_j\geq u)dM_j^1(u)=\boldsymbol{\eta}_i-\widetilde{\boldsymbol{\eta}}_i+O_p(1/n)$$

is a $\mathcal{F}_{i,t}^*$ martingale. Hence, we can apply Lemma 14(i) to get

$$\Pr\left(\|\boldsymbol{\eta}_i-\widetilde{\boldsymbol{\eta}}_i\|_\infty\geq K(1+e^KKt^*)\sqrt{4\log(2np/\varepsilon)/n}+K(1+2e^KKt^*)/n\right)\leq\varepsilon/n.$$

Taking union bound, we get $\|\boldsymbol{\eta}_i-\widetilde{\boldsymbol{\eta}}_i\|_\infty=O_p(\sqrt{\log(p)/n})$. Hence, $\sup_{i=1,\ldots,n}\|\widehat{\boldsymbol{\eta}}_i-\boldsymbol{\eta}_i\|_\infty=O_p\left(\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^o\|_1+\sqrt{\log(p)/n}\right)$.

Recall that $\widehat{\mathbf{q}}(t)$ and $\mathbf{q}(t)$ also take a similar form. We can likewise define

$$\widetilde{\mathbf{q}}(t)=n^{-1}\sum_{i=1}^n I(t>X_i)\int_t^{t^*}\{\mathbf{Z}_i(u)-\boldsymbol{\mu}(u)\}\widetilde{\omega}_i(u)d\widetilde{M}_i^1(u)$$

and

$$\widetilde{\mathbf{q}}^*(t)=n^{-1}\sum_{i=1}^n I(t>X_i)\int_t^{t^*}\{\mathbf{Z}_i(u)-\boldsymbol{\mu}(u)\}\widetilde{\omega}_i(u)dM_i^1(u).$$

By Lemmas 18, 20(i) and 20(iii), we have

$$\sup_{i=1,\ldots,n}\sup_{t\in[0,t^*]}\|\widetilde{\mathbf{q}}(t)-\widehat{\mathbf{q}}(t)\|_\infty=O_p\left(\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^o\|_1+\sqrt{\log(p)/n}\right).$$

By Lemma 13(ii), $\sup_{t\in[0,t^*]}\|\widetilde{\mathbf{q}}^*(t)-\mathbf{q}(t)\|=O_p\left(\sqrt{\log(p)/n}\right)$. We only need to find the rate for

$$\widetilde{\mathbf{q}}^*(t)-\widetilde{\mathbf{q}}(t)=n^{-1}\sum_{i=1}^n I(t>X_i)n^{-1}\sum_{j=1}^n\int_t^{t^*}n^{-1}\sum_{i=1}^n\{\mathbf{Z}_i(u)-\boldsymbol{\mu}(u)\}\widetilde{\omega}_i(u)I(C_j\geq u)dM_j^1(u).$$

We repeat the trick for $\boldsymbol{\eta}_i-\widetilde{\boldsymbol{\eta}}_i$. Applying Lemma 14(ii) to the $\mathcal{F}_{i,t}^*$ martingale

$$\mathbf{M}_i^q(t)=n^{-1}\sum_{j\neq i}\int_0^t n^{-1}\sum_{i=1}^n\{\mathbf{Z}_i(u)-\boldsymbol{\mu}(u)\}\widetilde{\omega}_i(u)I(C_j\geq u)dM_j^1(u)$$

and obtain $\sup_{i=1,\ldots,n}\sup_{t\in[0,t^*]}\|\mathbf{M}_i^q(t)\|_\infty = O_p(\sqrt{\log(p)/n})$. Hence,

$$\sup_{t\in[0,t^*]}\|\widetilde{\mathbf{q}}^*(t)-\widetilde{\mathbf{q}}(t)\|_\infty \le 2\sup_{i=1,\ldots,n}\sup_{t\in[0,t^*]}\|\mathbf{M}_i^q(t)\|_\infty + O_p(1/n) = O_p(\sqrt{\log(p)/n}).$$

Putting the rates together, we have $\sup_{t\in[0,t^*]}\|\widehat{\mathbf{q}}(t)-\mathbf{q}(t)\|_\infty = O_p\left(\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^o\|_1+\sqrt{\log(p)/n}\right)$.

We can directly obtain $\sup_{t\in[0,t^*]}|\widehat{\pi}(t)-\pi(t)| = O_p\left(\sqrt{\log(n)/n}\right)$ from Lemma 13(ii).

Define

$$\widetilde{\boldsymbol{\psi}}_i = \int_0^{t^*}\frac{\mathbf{q}(t)}{\pi(t)}d\widehat{M}_i^c(t)$$

The total variation of $\widehat{M}_i^c(t)$ is at most $1+2e^K/\rho_2$ with probability tending to one by Lemma 21.

Using the results so far, we have

$$\sup_{i=1,\ldots,n}\|\widehat{\boldsymbol{\psi}}_i-\widetilde{\boldsymbol{\psi}}_i\|_\infty = O_p\left(\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^o\|_1+\sqrt{\log(p)/n}\right).$$

The remainder

$$\boldsymbol{\psi}_i-\widetilde{\boldsymbol{\psi}}_i = n^{-1}\sum_{j=1}^n\int_0^{t^*}\frac{\mathbf{q}(t)}{\pi(t)}I(C_i\ge t)I(X_j\ge t)dM_j^c(t)$$

is a $\mathcal{F}_t$ martingale. We can put the n martingales in $\mathbb{R}^p$ into a $\mathbb{R}^{np}$ vector and apply Lemma 14(i),

$$\sup_{i=1,\ldots,n}\|\boldsymbol{\psi}_i-\widetilde{\boldsymbol{\psi}}_i\|_\infty = O_p\left(\sqrt{\log(np)/n}\right) = O_p\left(\sqrt{\log(p)/n}\right).$$

Therefore, we get $\sup_{i=1,\ldots,n}\|\boldsymbol{\psi}_i-\widehat{\boldsymbol{\psi}}_i\|_\infty = O_p\left(\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^o\|_1+\sqrt{\log(p)/n}\right)$.

Finally, we decompose

$$\begin{aligned}\|\widehat{\mathcal{V}}-\mathcal{V}\|_{\max} \le & n^{-1}\sum_{i=1}^n\|\widehat{\boldsymbol{\eta}}_i+\widehat{\boldsymbol{\psi}}_i\|_\infty\|\widehat{\boldsymbol{\eta}}_i+\widehat{\boldsymbol{\psi}}_i-\boldsymbol{\eta}_i-\boldsymbol{\psi}_i\|_\infty \\ & + n^{-1}\sum_{i=1}^n\|\widehat{\boldsymbol{\eta}}_i+\widehat{\boldsymbol{\psi}}_i-\boldsymbol{\eta}_i-\boldsymbol{\psi}_i\|_\infty\|\boldsymbol{\eta}_i+\boldsymbol{\psi}_i\|_\infty \\ & + \left\|n^{-1}\sum_{i=1}^n(\boldsymbol{\eta}_i+\boldsymbol{\psi}_i)(\boldsymbol{\eta}_i+\boldsymbol{\psi}_i)^\top-\mathcal{V}\right\|_{\max}.\end{aligned}$$

We have shown that $\sup_{i=1,\ldots,n} \|\widehat{\boldsymbol{\eta}}_i + \widehat{\boldsymbol{\psi}}_i - \boldsymbol{\eta}_i - \boldsymbol{\psi}_i\|_\infty = o_p(1)$. Moreover, $\sup_{i=1,\ldots,n} \|\widehat{\boldsymbol{\eta}}_i + \widehat{\boldsymbol{\psi}}_i\|_\infty$ is $O_p(1)$ by Lemmas 15 and 21. In addition, we observe that $n^{-1}\sum_{i=1}^n (\boldsymbol{\eta}_i + \boldsymbol{\psi}_i)(\boldsymbol{\eta}_i + \boldsymbol{\psi}_i)^\top$ is an average of i.i.d. terms whose expectation is defined as $\mathcal{V}$. By Lemmas 15 and 21, we have the uniform maximal bound

$$\sup_{i=1,\ldots,n} \|(\boldsymbol{\eta}_i + \boldsymbol{\psi}_i)(\boldsymbol{\eta}_i + \boldsymbol{\psi}_i)^\top\|_{\max} = \sup_{i=1,\ldots,n} \|(\boldsymbol{\eta}_i + \boldsymbol{\psi}_i)\|_\infty^2$$

is also $O_p(1)$. We finish the proof by applying Lemma 11 to the last term in the decomposition above, $\left\|n^{-1}\sum_{i=1}^n (\boldsymbol{\eta}_i + \boldsymbol{\psi}_i)(\boldsymbol{\eta}_i + \boldsymbol{\psi}_i)^\top - \mathcal{V}\right\|_{\max}$.

□

Proofs of Auxiliary Lemmas

Proof of Lemma 13. (i) Without loss of generality, let t_{11} be the first jump time of $N_1(t)$. By the i.i.d. assumption, t_{11} is independent of all $\mathbf{S}_i(t)$ with $i \geq 2$. Thus, the sequence

$$\mathbf{L}_l = n^{-1}\sum_{i=2}^{l} \{\mathbf{S}_i(t_{11}) - \mathbf{s}(t_{11})\}$$

is a martingale with respect to filtration $\left\{\sigma\big(\mathbf{S}_i(t), i \leq l\big), l = 2, \ldots, n\right\}$. The increment is bounded as

$$n^{-1}\left\{\mathbf{S}_i(t_{11}) - \mathbf{s}(t_{11})\right\} = n^{-1}\mathbb{E}_{\mathbf{S}_j}\left\{\mathbf{S}_i(t_{11}) - \mathbf{S}_j(t_{11})\right\} \leq n^{-1}K_S.$$

Applying Lemma 12 to $\mathbf{L}_n$, we get $\Pr\left(\|\mathbf{L}_n\|_{\max} > K_S x\right) < 2qe^{-nx^2/2}$. Since the dropped first term is also bounded by K_S/n, we get

$$\Pr\left(\left\|\bar{\mathbf{S}}(t_{11}) - \mathbf{s}(t_{11})\right\|_{\max} > K_S x + K_S/n\right) < 2qe^{-nx^2/2}.$$

We use simple union bound to extend the result to all t_{ij}'s whose number is at most nK_N.

(ii) Define a deterministic set $\mathcal{T}_n = \{kt^*/n : k = 1,\dots,n\} \cup \mathcal{T}_z$. By the union bound of Hoeffding's inequality [Hoe63] , we have

$$\Pr\left(\sup_{t\in\mathcal{T}_n} \left\|\bar{\mathbf{S}}(t) - \mathbf{s}(t)\right\|_{\max} > K_S x\right) < 2(n + |\mathcal{T}_z|)qe^{-nx^2/2}.$$

Combining the result from Lemma 13(i), we obtain

$$\left\|\bar{\mathbf{S}}(t) - \mathbf{s}(t)\right\|_{\max} = O_p(\sqrt{\log(npq)/n})$$

over a grid containing $\mathcal{T}_n$ and jumps of $N_i(t)$. We only need to show that the variation of $\bar{\mathbf{S}}(t) - \mathbf{s}(t)$ is sufficiently small inside each bin created by the grid.

Let t' and t'' be consecutive elements by order in $\mathcal{T}_n$. By our construction, there is no jump of any of the counting processes $N_i(t)$ in the interval (t', t''). Otherwise, the jump time is another element in $\mathcal{T}_n$ between t' and t'' so that t' and t'' are not consecutive elements by order. Under the assumption of the lemma, elements of all $\mathbf{S}_i(t)$'s are L_S-Lipschitz in (t', t''). Moreover, $|t'' - t'| \le t^*/n$ because of the deterministic $\{kt^*/n : k = 1,\dots,n\} \subset \mathcal{T}_n$. Along with the càglàd property, we obtain a bound of variation of $\bar{\mathbf{S}}(t)$ in (t', t'')

$$\sup_{t\in(t',t'')} \|\bar{\mathbf{S}}(t) - \bar{\mathbf{S}}(t'')\|_{\max} \le \sup_{i=1,\dots,n} \sup_{t\in(t',t'')} \|\mathbf{S}_i(t) - \mathbf{S}_i(t'')\|_{\max} \le L_S|t'' - t'| \le L_S t^*/n.$$

For any $t \in (t', t'')$, we bound the variation of $\mathbf{s}(t)$ by

$$\|\mathbf{s}(t) - \mathbf{s}(t'')\|_{\max} \le \int_t^{t''} \mathbb{E}\|\mathbf{d}_s(u)\|_{\max} du + \int_t^{t''} \mathbb{E}\{\|\mathbf{J}_s(u)\|_{\max} h_i^N(u)\} du \le (L_S + K_S K)t^*/n.$$

For arbitrary $t \in [0, t^*]$, we find the the corresponding bin $(t', t'']$ contains t. Putting the results together, we have

$$\|\bar{\mathbf{S}}(t) - \mathbf{s}(t)\|_{\max} \le \|\bar{\mathbf{S}}(t) - \bar{\mathbf{S}}(t'')\|_{\max} + \|\mathbf{s}(t) - \mathbf{s}(t'')\|_{\max} + \|\bar{\mathbf{S}}(t'') - \mathbf{s}(t'')\|_{\max}$$

$$\leq O_p(\sqrt{\log(npq)/n}) + O(1/n).$$

□

Proof of Lemma 14. (i) The summands in $\mathbf{M}_\Phi(t)$ are the integrals of $\mathcal{F}_{t-}$-measurable processes over $\mathcal{F}_t$-adapted martingales, so $\mathbf{M}_\Phi(t)$ is a $\mathcal{F}_t$-adapted martingale [KP02, p.165].

Suppose $\{T_i : i = 1,\ldots,n\}$ are the jump times of $\{N_i(t)\}$. We artificially set $T_i = t^*$ if $N_i(t)$ has no jump in $[0,t^*]$. Define $0 \leq R_1 \leq \cdots \leq R_{2n}$ be the order statistics of $\{T_i : i = 1,\ldots,n\} \cup \{kt^*/n : k = 1,\ldots,n\}$. Hence, $\{R_k : k = 1,\ldots,2n\}$ is a set of ordered $\mathcal{F}_t$ stopping times. Applying optional stopping theorem, we get a discrete time martingale $\mathbf{M}_\Phi(R_k)$ adapted to $\mathcal{F}_{R_k}$.

The increment of $\mathbf{M}_\Phi(R_k)$ comes from either the counting part or the compensator part, which we can bound separately. By our construction of R_k's, each left-open right-closed bin $(R_k, R_{k+1}]$ satisfies two conditions. There is at most one jump from $\sum_{i=1}^n N_i(t)$ in the bin at R_{k+1}. The length of the bin is at most t^*/n. The increment of the martingale $\mathbf{M}_\Phi(t)$ over $(R_k, R_{k+1}]$ is decomposed into two coordinate-wise integrals, a jump minus a compensator,

$$\mathbf{M}_\Phi(t) = n^{-1}\sum_{i=1}^n \int_{R_k}^{R_{k+1}} \mathbf{\Phi}_i(u)dN_i(u) - n^{-1}\sum_{i=1}^n \int_{R_k}^{R_{k+1}} \mathbf{\Phi}_i(u)h_i(u)du.$$

With the assumed a.s. upper bound for $\sup_{t\in[0,t^*]} \|\mathbf{\Phi}_i(t)\|_{\max} \leq K_\Phi$, we have almost surely the jump of $\mathbf{M}_\Phi(t)$ in the bin be bounded by

$$\left\| n^{-1}\sum_{i=1}^n \int_{R_k}^{R_{k+1}} \mathbf{\Phi}_i(u)dN_i(u) \right\|_{\max} \leq K_\Phi/n.$$

Additionally with the assumed upper bound for $\sup_{i=1,\ldots,n}\sup_{t\in[0,t^*]} h_i(t) \le K_h$, we have the compensator of $\mathbf{M}_\Phi(t)$ increases over the bin by at most

$$\left\| \int_{R_k}^{R_{k+1}} n^{-1} \sum_{i=1}^{n} \mathbf{\Phi}_i(u) h_i(u) du \right\|_{\max} \le K_\Phi K_h (R_{k+1} - R_k) \le K_\Phi K_h t^*/n.$$

We obtain a uniform concentration inequality for $\mathbf{M}_\Phi(R_k)$ by Lemma 12

$$\Pr\left(\sup_{k=1,\ldots,2n} \|\mathbf{M}_\Phi(R_k)\|_{\max} \ge K_\Phi(1+K_h t^*)x \right) \le 2qe^{-nx^2/4}.$$

Remark that the uniform version of Lemma 12 is the application of Doob's maximal inequality [Dur10, Theorem 5.4.2, page 213]. For $t \in (R_k, R_{k+1})$, we use the bounded increment derived above

$$\|\mathbf{M}_\Phi(t) - \mathbf{M}_\Phi(R_k)\|_{\max} \le \left\| \int_{R_k+}^{R_{k+1}+} n^{-1} \sum_{i=1}^{n} \mathbf{\Phi}_i(u) h_i(u) du \right\|_{\max} \le K_\Phi K_h t^*/n.$$

(ii) Under the additional assumption $\sup_{i=1,\ldots,n}\sup_{t\in[0,t^*]} \|\mathbf{\Phi}_i(t)\|_{\max} = O_p(a_n)$, we can find $K_{\Phi,\varepsilon}$ for every $\varepsilon > 0$ such that

$$\Pr\left(\sup_{i=1,\ldots,n} \sup_{t\in[0,t^*]} \|\mathbf{\Phi}_i(t)\|_{\max} \le K_{\Phi,\varepsilon} a_n \right) \ge 1 - \varepsilon/2$$

for any n. We apply Lemma 14(i) to obtain that event

$$\left\{ \sup_{t\in[0,t^*]} \|\mathbf{M}_\Phi(t)\|_{\max} \le K_{\Phi,\varepsilon} a_n \{(1+K_h t^*)\sqrt{2\log(4q)/n} + K_h t^*/n\}, \right.$$
$$\left. \sup_{i=1,\ldots,n} \sup_{t\in[0,t^*]} \|\mathbf{\Phi}_i(t)\|_{\max} \le K_{\Phi,\varepsilon} a_n \right\}$$

occurs with probability no less than $1-\varepsilon$.

□

Proof of Lemma 15. Notice all $a_i(t)$'s are nonnegative. Hence, $\sum_{i=1}^n |a_i(t)| = \sum_{i=1}^n a_i(t)$. We apply Hölder's inequality for each coordinate

$$\left| \left\{ \frac{\sum_{i=1}^n a_i(t)\mathbf{Z}_i(t)^{\otimes l}}{\sum_{i=1}^n a_i(t)} \right\}_j \right| = \left| \sum_{i=1}^n \frac{a_i(t)}{\sum_{i=1}^n a_i(t)} \left\{ \mathbf{Z}_i(t)^{\otimes l} \right\}_j \right| \leq \frac{\sum_{i=1}^n |a_i(t)|}{|\sum_{i=1}^n a_i(t)|} \sup_{i=1,\ldots,n} \left| \left\{ \mathbf{Z}_i(t)^{\otimes l} \right\}_j \right|.$$

Hence, the maximal norm of $\sum_{i=1}^n a_i(t)\mathbf{Z}_i(t)^{\otimes l}$ is bounded by $(K_3/2)^l$ under (2.40). Similar result can be achieved with the sum replaced by the expectation.

To apply the result above to the processes $\mathbf{S}^{(l)}(t,\boldsymbol{\beta})/S^{(0)}(t,\boldsymbol{\beta})$, $\widetilde{\mathbf{S}}^{(l)}(t,\boldsymbol{\beta})/\widetilde{S}^{(0)}(t,\boldsymbol{\beta})$ and $\mathbf{s}^{(l)}(t,\boldsymbol{\beta})/s^{(0)}(t,\boldsymbol{\beta})$, we set $a_i(t)$ as $\omega_i(t)Y_i(t)e^{\boldsymbol{\beta}^\top \mathbf{Z}_i(t)}$ and $I(C_i \geq t)Y_i(t)e^{\boldsymbol{\beta}^\top \mathbf{Z}_i(t)}$. □

Proof of Lemma 16. Since $\{I(X_i \geq t^*), i = 1,\ldots,n\}$ are i.i.d. Bernoulli random variable, we may apply Lemma 11 for lower tail,

$$\Pr\left(n^{-1} \sum_{i=1}^n I(X_i \geq t^*) < \Pr(X_i \geq t^*) - x \right) \leq \exp(-2nx^2).$$

By (2.41), we can find lower bounds for the probability

$$\Pr(X_i \geq t^*) \geq \Pr(C_i \geq t^*, \infty > T_i^1 \geq t^*) = G(t^*)\mathbb{E}\{F_1(\infty;\mathbf{Z}_i) - F_1(t^*;\mathbf{Z}_i)\} \geq \rho_2/K_4.$$

We may relax the inequality at $x = \rho_2/(2K_4)$ to

$$\Pr\left(n^{-1} \sum_{i=1}^n I(X_i \geq t^*) < \rho_2/(2K_4) \right) \leq e^{-n\rho_2^2/(2K_4^2)}.$$

Because $I(C_i \geq t) \geq I(C_i \geq t^*)$ and $Y_i(t) \geq Y_i(t^*)$, $\widetilde{S}^{(0)}(t;K_4)$ is a lower bound for $\widetilde{S}^{(0)}(t)$. The summands in $\widetilde{S}^{(0)}(t;K_4)$ are i.i.d. uniformly bounded by K_4. Thus, we may apply Lemma 13(i) with one-sided version,

$$\Pr\left(\sup_{k \in 1\ldots K_T} \widetilde{S}^{(0)}(t;K_4) < \mathbb{E}\{\widetilde{S}^{(0)}(t;K_4)\} - K_4 x - K_4/n \right) < ne^{-nx^2/2}.$$

By (C3), the expectation has a lower bound

$$\mathbb{E}\{\widetilde{S}^{(0)}(t;K_4)\} = G(t^*)\mathbb{E}\left[\{1-F_1(t;\mathbf{Z}_i)\}\min\{K_4, e^{\boldsymbol{\beta}^{o\top}\mathbf{Z}_i(t)}\}\right] > \rho_2.$$

We relax the inequality at $x = (\rho_2/2 - 1/n)/K_4$,

$$\Pr\left(\sup_{k\in 1\ldots K_T} \widetilde{S}^{(0)}(t;K_4) < \rho_2\right) < ne^{-n(\rho_2-2/n)^2/(8K_4^2)}.$$

□

Proof of Lemma 17. Since $\varepsilon_i > 1$ implies $T_i^1 = \infty$, the probability of observing a type-2 event conditioning on $\mathbf{Z}_i(\cdot)$ has an upper bound

$$\begin{aligned}\Pr\left(\varepsilon_i > 1 | \mathbf{Z}_i(\cdot)\right) &= \exp\left\{-\int_0^\infty e^{\boldsymbol{\beta}^{o\top}\mathbf{Z}_i(u)} h_0^1(u)du\right\} \\ &\leq \exp\left\{-K_e x \int_0^\infty I\left(e^{\boldsymbol{\beta}^{o\top}\mathbf{Z}_i(u)} \geq K_e x\right) h_0^1(u)du\right\}.\end{aligned}$$

Hence, we may derive a bound for

$$\Pr\left(\delta_i\varepsilon_i > 1, \sup_{t\in[0,t^*]} e^{\boldsymbol{\beta}^{o\top}\mathbf{Z}_i(t)} > K_e\right) \leq \Pr\left(\varepsilon_i > 1 \middle| \sup_{t\in[0,t^*]} e^{\boldsymbol{\beta}^{o\top}\mathbf{Z}_i(t)} > K_e\right)$$

if we can bound $\int_0^\infty I\left(e^{\boldsymbol{\beta}^{o\top}\mathbf{Z}_i(u)} \geq K_e x\right) h_0^1(u)du$ away from zero with a certain x whenever $e^{\boldsymbol{\beta}^{o\top}\mathbf{Z}_i(t')} > K_e$ for some $t' \in [0,t^*]$.

Under (C3), there is an interval I' containing t' of length ρ_4 in which $\mathbf{Z}_i(\cdot)$ has no jumps. The variation of linear predictor is bounded

$$\sup_{t\in I'}\left|\boldsymbol{\beta}^{o\top}\mathbf{Z}_i(t) - \boldsymbol{\beta}^o\mathbf{Z}_i(t')\right| \leq K_6K_7\|\boldsymbol{\beta}^o\|_\infty\rho_4.$$

So, the relative risk $e^{\boldsymbol{\beta}^{o\top}\mathbf{Z}_i(t)}$ is greater than $K_e\exp\{-K_6K_7\|\boldsymbol{\beta}^o\|_\infty\rho_4\}$ over I'. Hence, we get a lower bound for

$$\int_0^\infty I\left(e^{\boldsymbol{\beta}^{o\top}\mathbf{Z}_i(u)} \geq K_e\exp\{-K_6K_7\|\boldsymbol{\beta}^o\|_\infty\rho_4\}\right) h_0^1(u)du \geq \rho_4\rho_1.$$

We finish the proof by taking a union bound over $i = 1, \ldots, n$. □

Proof of Lemma 18. Recall that $M_i^c(t) = I(C_i \le t) - \int_0^t I(C_i \ge u) h^c(u) du$ is a counting process martingale adapted to complete data filtration $\mathcal{F}_t$. The Kaplan-Meier estimator $\widehat{G}(t)$ has the martingale representation [KP02, p.170 (5.45)],

$$M^G(t) = \frac{\widehat{G}(t)}{G(t)} - 1 = n^{-1} \sum_{i=1}^{n} \int_0^t \frac{\widehat{G}(u-) I(X_i \ge u)}{G(u) n^{-1} \sum_{j=1}^{n} I(X_j \ge u)} dM_i^c(u).$$

For $\delta_i \varepsilon_i > 1$ and $t > X_i$,

$$\omega_i(t) - \widetilde{\omega}_i(t) = -\frac{\widehat{G}(t)}{\widehat{G}(X_i)} M^G(X_i) + \frac{G(t)}{G(X_i)} M^G(t),$$

so we will be able to establish a concentration result for the error from Kaplan-Meier

$$\left\| n^{-1} \sum_{i=1}^{n} \{\omega_i(t) - \widetilde{\omega}_i(t)\} Y_i(t) e^{\boldsymbol{\beta}^{o\top} \mathbf{Z}_i(t)} \mathbf{Z}_i(t)^{\otimes l} \right\|_{\max} \le 2 Q_1(\varepsilon) (K_3/2)^l \sup_{t \in [0, t^*]} |M^G(t)|$$

if we first obtain a concentration result for $\sup_{t \in [0,t^*]} |M^G(t)|$. On event $n^{-1} \sum_{j=1}^{n} I(X_j \ge u) \ge \rho_2/(2K_4)$, the integrated functions are $\mathcal{F}_{t-}$-adapted with uniform bound $2(K_4^{/} \rho_2)^2$. The hazard $h^c(t) \le K_1$ by (C1). Hence, we may apply Lemma 14(i) with $x = \sqrt{4 \log(2/\varepsilon)/n}$ to obtain

$$\Pr\left(\sup_{t \in [0,t^*]} |M^G(t)| < 2(K_4/\rho_2)^2 \left\{ (1 + K_1 t^*) \sqrt{4 \log(2/\varepsilon)/n} + K_1 t^*/n \right\} \right) \tag{2.73}$$

$$\le \Pr(\Omega_1 \cap \Omega_2(\varepsilon)) - \varepsilon. \tag{2.74}$$

□

Proof of Lemma 19. A sharper inequality is available if $\mathbf{Z}_i$'s are not time-dependent. We may exploit the martingale structure of $\boldsymbol{\Delta}^{(l)}(t)/G(t)$. With general time-dependent covariates, we would decompose the approximation error $\boldsymbol{\Delta}^{(l)}(t)$ into two parts, the error from Kaplan-Meier estimate $\widehat{G}(t)$ and the error from missingness in C_i's among the type-2 events.

Define the indicator $\upsilon_i(t) = I(t > X_i)I(\delta_i\varepsilon_i > 1)$. Since $\{\omega_i(t) - I(C_i \geq t)\}Y_i(t)$ is non-zero only when $\upsilon(t) = 1$, we may alternatively write

$$\boldsymbol{\Delta}^{(l)}(t) = n^{-1}\sum_{i=1}^{n}\{\omega_i(t) - I(C_i \geq t)\}\upsilon_i(t)e^{\boldsymbol{\beta}^{o\top}\mathbf{Z}_i(t)}\mathbf{Z}_i(t)^{\otimes l}.$$

We may use the upper bound $\sup_{i=1,\ldots,n}\sup_{t\in[0,t^*]}\left|\upsilon_i(t)e^{\boldsymbol{\beta}^{o\top}\mathbf{Z}_i(t)}\right| \leq Q_1(\varepsilon)$ on $\Omega_2(\varepsilon)$. By Lemma 18,

$$\left\|n^{-1}\sum_{i=1}^{n}\{\omega_i(t) - \widetilde{\omega}_i(t)\}\upsilon_i(t)e^{\boldsymbol{\beta}^{o\top}\mathbf{Z}_i(t)}\mathbf{Z}_i(t)^{\otimes l}\right\|_{\max} \leq Q_1(\varepsilon)(K_3/2)^l Q_7(n,p,\varepsilon)$$

on $\Omega_2(\varepsilon)\cap\Omega_3(\varepsilon)$.

Define the error from missingness in C_i's among the type-2 events as

$$\widetilde{\boldsymbol{\Delta}}^{(l)}(t) = n^{-1}\sum_{i=1}^{n}\{\widetilde{\omega}_i(t) - I(C_i \geq t)\}\upsilon_i(t)e^{\boldsymbol{\beta}^{o\top}\mathbf{Z}_i(t)}\mathbf{Z}_i(t)^{\otimes l}.$$

Since $\mathbb{E}\{r_i(t)|T_i\} = G(t\wedge T_i)$, [FG99] has shown that

$$\mathbb{E}\{\widetilde{\omega}_i(t)|T_i\} = \mathbb{E}\{I(C_i \geq t)|T_i\} = G(t).$$

Applying tower property, we have $\mathbb{E}\left\{\widetilde{\boldsymbol{\Delta}}^{(l)}(t)\right\} = \mathbf{0}$. Hence, we can apply Lemma 13(i) with $x = \sqrt{2\log(2np^l/\varepsilon)/n}$

$$\Pr\left(\sup_{k\in 1\ldots K_T}\left\|\widetilde{\boldsymbol{\Delta}}^{(l)}\left(T^1_{(k)}\right)\right\|_{\max} \leq Q_1(\varepsilon)(K_3/2)^l\left\{\sqrt{2\log(2np^l/\varepsilon)/n} + 1/n\right\}\right)$$

is at least $\Pr(\Omega_1\cap\Omega_2(\varepsilon)) - \varepsilon$. This finishes the proof of the first result.

We prove the other result by decomposing the differences into terms with $\boldsymbol{\Delta}^{(l)}(t)$,

$$\frac{\mathbf{S}^{(l)}(t,\boldsymbol{\beta}^o)}{S^{(0)}(t,\boldsymbol{\beta}^o)} - \frac{\widetilde{\mathbf{S}}^{(l)}(t,\boldsymbol{\beta}^o)}{\widetilde{S}^{(0)}(t,\boldsymbol{\beta}^o)} = \frac{1}{\widetilde{S}^{(0)}(t,\boldsymbol{\beta}^o)}\boldsymbol{\Delta}^{(l)}(t) - \frac{\mathbf{S}^{(l)}(t,\boldsymbol{\beta}^o)}{S^{(0)}(t,\boldsymbol{\beta}^o)\widetilde{S}^{(0)}(t,\boldsymbol{\beta}^o)}\Delta^{(0)}(t).$$

$\mathbf{S}^{(l)}(t,\beta^o)/S^{(0)}(t,\beta^o)$ is the weighted average of $\mathbf{Z}_i(t)^{\otimes l}$, so its maximal norm is bounded by $(K_3/2)^l$. On the event Ω_1,

$$\left\| \frac{\mathbf{S}^{(l)}(t,\beta^o)}{S^{(0)}(t,\beta^o)} - \frac{\widetilde{\mathbf{S}}^{(l)}(t,\beta^o)}{\widetilde{S}^{(0)}(t,\beta^o)} \right\|_\infty \leq \frac{2}{\rho_2} \|\boldsymbol{\Delta}^{(l)}(t)\|_\infty + \frac{K_3^l}{2^{l-1}\rho_2} |\Delta^{(0)}(t)|.$$

We can simply plug in the bounds and tail probabilities for $\Delta^{(0)}\left(T^1_{(k)}\right)$ and $\boldsymbol{\Delta}^{(1)}\left(T^1_{(k)}\right)$ in (2.63). □

Proof of Lemma 20. (i) By (C1) and (D1), we have $\left\| e^{\beta^{o\top}\mathbf{Z}_i(t)} \mathbf{Z}_i(t)^{\otimes l} \right\|_{\max} \leq (K_3/2)^l e^K \asymp 1$. Thus, all terms involved are bounded. Moreover, $e^{\beta^{o\top}\mathbf{Z}_i(t)}\mathbf{Z}_i(t)^{\otimes l}$ jumps only at the jumps of $N_i^z(t)$ by (D3). Define the outer product of arrays $\mathbf{u} \in \mathbb{R}^{p_1\times\cdots\times p_d}$ and $\mathbf{v} \in \mathbb{R}^{q_1\times\cdots\times q_{d'}}$ as

$$\mathbf{u}\otimes\mathbf{v} \in \mathbb{R}^{p_1\times\cdots\times p_d\times q_1\times\cdots\times q_{d'}}, \quad (\mathbf{u}\otimes\mathbf{v})_{i_1,\ldots,i_{d+d'}} = \mathbf{u}_{i_1,\ldots,i_d} \times \mathbf{v}_{i_{d+1},\ldots,i_{d+d'}}.$$

Between two consecutive jumps of $N_i^z(t)$,

$$\begin{aligned}
&\left\| \frac{d}{dt} e^{\beta^{o\top}\mathbf{Z}_i(t)} \mathbf{Z}_i(t)^{\otimes l} \right\|_{\max} \\
=&\left\| e^{\beta^{o\top}\mathbf{Z}_i(t)} \mathbf{Z}_i(t)^{\otimes l} \beta^{o\top} \mathbf{d}_i^z(t) + I(l>0) e^{\beta^{o\top}\mathbf{Z}_i(t)} l \mathbf{Z}_i(t)^{\otimes l-1} \otimes \mathbf{d}_i^z(t) \right\|_{\max} \\
\leq& e^K \{(K_3/2)^l K + I(l>1)(K_3/2)^{l-1} K\} \asymp 1.
\end{aligned}$$

Hence, $e^{\beta^{o\top}\mathbf{Z}_i(t)}\mathbf{Z}_i(t)^{\otimes l}$ satisfies the continuity condition for Lemma 13(ii).

Like in Lemma 19, we first replace $\omega_i(t)$ by $\widetilde{\omega}_i(t) = r_i(t)G(t)/G(X_i\wedge t)$. Denote $\widetilde{\boldsymbol{\Delta}}^{(l)}(t) = n^{-1}\sum_{i=1}^n \{\widetilde{\omega}_i(t) - I(C_i \geq t)\} Y_i(t) e^{\beta^{o\top}\mathbf{Z}_i(t)} \mathbf{Z}_i(t)^{\otimes l}$. By Lemma 18, $\sup_{t\in[0,t^*]} \|\boldsymbol{\Delta}^{(l)}(t) - \widetilde{\boldsymbol{\Delta}}^{(l)}(t)\|_{\max} = O_p\left(n^{-1/2}\right)$. Then, we apply Lemma 13(ii) to the i.i.d. mean zero process $\widetilde{\boldsymbol{\Delta}}^{(l)}(t)$,

$$\sup_{t\in[0,t^*]} \|\widetilde{\boldsymbol{\Delta}}^{(l)}(t)\|_{\max} = O_p\left(\sqrt{\log(np^lK_n)/n}\right).$$

Similarly,

$$\sup_{t\in[0,t^*]} \|\widetilde{\mathbf{S}}^{(l)}(t,\boldsymbol{\beta}^o) - \mathbf{s}^{(l)}(t,\boldsymbol{\beta}^o)\|_{\max} = O_p\left(\sqrt{\log(np^l K_n)/n}\right).$$

Finally, we extend to results to the quotients by decomposition

$$\frac{\mathbf{S}^{(1)}(t,\boldsymbol{\beta}^o)}{S^{(0)}(t,\boldsymbol{\beta}^o)} - \frac{\widetilde{\mathbf{S}}^{(1)}(t,\boldsymbol{\beta}^o)}{\widetilde{S}^{(0)}(t,\boldsymbol{\beta}^o)} = \frac{1}{\widetilde{S}^{(0)}(t,\boldsymbol{\beta}^o)}\boldsymbol{\Delta}^{(l)}(t) - \frac{\mathbf{S}^{(l)}(t,\boldsymbol{\beta}^o)}{S^{(0)}(t,\boldsymbol{\beta}^o)\widetilde{S}^{(0)}(t,\boldsymbol{\beta}^o)}\Delta^{(0)}(t).$$

The denominators are bounded away from zero by Lemma 21 by choosing $K_4 = e^K$.

(ii) First, we show that $\Delta_i(t)$ is related to the martingales

$$M_i^c(t) = N_i^c(t) - \int_0^t \{1 - N_i^c(u-)\}h^c(u)du.$$

$\Delta_i(t)$ is non-zero only after an observed type-2 event. To simplify notation, we define the indicator for non-zero $\Delta_i(t)$, $\upsilon_i(t) = r_i(t)Y_i(t)I(t > X_i) = I(\delta_i\varepsilon_i > 1)I(t > X_i)$.

Denote the Nelson-Aalen type estimator for censoring cumulative hazard as

$$\widehat{H}^c(t) = \sum_{i=1}^n \int_0^t \Big\{ \sum_{j=1}^n I(X_j \geq u)\Big\}^{-1} I(X_i \geq u)dN_i^c(u).$$

Define $R_i(t) = \widehat{G}(t)/\widehat{G}(X_i) - 1 + \int_{X_i}^t \widehat{G}(u-)d\widehat{H}^c(u)/\widehat{G}(X_i)$. Let c_k and c_{k+1} be two consecutive observed censoring times greater than X_i. The increment $R_i(c_{k+1}) - R_i(c_k)$ is in fact

$$\frac{\widehat{G}(c_k)}{\widehat{G}(X_i)}\left\{\frac{\sum_{j=1}^n I(X_j \geq c_{k+1}) - 1}{\sum_{j=1}^n I(X_j \geq c_{k+1})} - 1 + \frac{1}{\sum_{j=1}^n I(X_j \geq c_{k+1})}\right\} = 0.$$

For $t > X_i$, we have $R_i(t) = 0$. Thus,

$$\Delta_i(t) = \{\widehat{G}(t)/\widehat{G}(X_i) - 1 + N_i^c(t) - N_i^c(X_i) - R_i(t)\}\upsilon_i(t)$$

$$= \int_{X_i}^{t} \upsilon_i(u) dM_i^c(u) - \int_{X_i}^{t} \omega_i(u-)\upsilon_i(u) \frac{\sum_{j=1}^{n} I(X_j \geq u) dM_j^c(u)}{\sum_{j=1}^{n} I(X_j \geq u)} +$$

$$+ \int_{X_i}^{t} \{I(C_i \geq u) - \omega_i(u-)\}\upsilon_i(u) h^c(u) du. \tag{2.75}$$

Notice $\upsilon_i(t)$ does not change beyond X_i if $C_i > X_i$, i.e. an event is observed. Since $h^c(u) \leq K < \infty$, we may modify the integrand at countable many points without changing the integral

$$\int_{X_i}^{t} \{I(C_i \geq u) - \omega_i(u-)\}\upsilon_i(u) h^c(u) du = -\int_{X_i}^{t} \Delta_i(u) h^c(u) du.$$

Hence, (2.75) gives an first order linear integral equation for $\Delta_i(u)$. The general solution to the related homogeneous problem

$$\Delta_i(t) = -\int_{X_i}^{t} \Delta_i(u) h^c(u) du, \quad \Delta_i(X_i) = 0$$

has only one unique solution $\Delta_i(t) = 0$. Thus, we only need to find one specific solution to (2.75). Define an integral operator $I \circ f = \int_{X_i}^{t} f(u) h^c(u) du$. Then, the solution to $f(t) = g(t) - I \circ f(t)$ can be written as

$$f(t) = (1 - I + I^2 - I^3 + \ldots) \circ g(t) \triangleq e^{-I} \circ g(t).$$

By inductively using integration by parts, we are able to calculate

$$I^n \circ g(t) = \frac{1}{n!} \sum_{k=1}^{n} \binom{n}{k} (-1)^k H^c(t)^{n-k} \int_{X_i}^{t} H^c(u)^k dg(t).$$

Hence, the solution can be calculated as the series

$$f(t) = \sum_{n=1}^{\infty} (-1)^n I^n \circ g(t) = \sum_{n=1}^{\infty} \sum_{k=1}^{n} \frac{\{-H^c(t)\}^{n-k}}{(n-k)!} \int_{X_i}^{t} \frac{H^c(u)^k}{k!} dg(u)$$

$$= \sum_{k=1}^{\infty} \int_{X_i}^{t} \frac{H^c(u)^k}{k!} dg(u) \sum_{n=k}^{\infty} \frac{\{-H^c(t)\}^{n-k}}{(n-k)!} = G(t) \int_{X_i}^{t} G(u)^{-1} dg(u).$$

Applying to (2.75), we get

$$\Delta_i(t) = G(t) \int_{X_i}^{t} G(u)^{-1} dM_i^{\Delta}(u),$$

with a $\mathcal{F}_t-$ martingale

$$M_i^{\Delta}(t) = \int_0^t I(C_i \geq u) \upsilon_i(u) dM_i^c(u) - \int_0^t \omega_i(u-) \upsilon_i(u) \frac{\sum_{j=1}^n I(X_j \geq u) dM_j^c(u)}{\sum_{j=1}^n I(X_j \geq u)}.$$

Now, we use the martingale structure to prove the Lemma. Denote the $\mathcal{F}_t^*$ martingale

$$\mathbf{M}^g(t) = n^{-1} \sum_{i=1}^{n} \int_0^t G(u) e^{\boldsymbol{\beta}^{o\top} \mathbf{Z}_j(u)} \mathbf{g}(u) I(C_i \geq u) dM_i^1(u).$$

$\mathbf{M}^g(t)$ satisfies the condition for Lemma 14(i). Hence, we have $\sup_{t\in[0,t^*]} \|\mathbf{M}^g(t)\|_{\max} = O_p\left(\sqrt{\log(q')/n}\right)$. Also define

$$\widetilde{\Delta}_i(t) = \{\widetilde{\omega}_i(t) - I(C_i > t)\} Y_i(t).$$

By Lemma 18, $\sup_{i=1,\ldots,n} \sup_{t\in[0,t^*]} |\Delta_i(t) - \widetilde{\Delta}_i(t)| = O_p\left(n^{-1/2}\right)$. The total variation of each $\Delta_i(t)$ is at most 2. Hence, we can apply integration by parts to (2.65),

$$\begin{aligned}
& G^{-1}(t^*) \mathbf{M}^g(t^*-) \otimes n^{-1/2} \sum_{j=1}^{n} \Delta_j(t^*-) \boldsymbol{\phi}(\mathbf{Z}_j(t^*-)) \\
& - n^{-1/2} \sum_{j=1}^{n} \int_0^{t^*} \mathbf{M}^g(t) \otimes \boldsymbol{\phi}(\mathbf{Z}_j(t)) dM_j^{\Delta}(t) \\
& - n^{1/2} \int_0^{t^*} \mathbf{M}^g(t) \otimes G^{-1}(t) n^{-1} \sum_{j=1}^{n} \Delta_j(t) d\boldsymbol{\phi}(\mathbf{Z}_j(t)) \\
& \triangleq I_1 - I_2 - I_3.
\end{aligned}$$

We have shown that $|\mathbf{M}^g(t^*-)| = O_p\left(\sqrt{\log(q')/n}\right)$ and $\sup_{j=1,\ldots,n}|\Delta_j(t^*-) - \widetilde{\Delta}_j(t^*-)| = O_p\left(n^{-1/2}\right)$. By assumption, $\|\phi(\mathbf{Z}_j(t^*-))\|_{\max} \leq K_\phi \asymp 1$. As a result, we may replace the $\Delta_i(t)$ in I_1 by $\widetilde{\Delta}_i(t)$ with with an $O_p\left(\sqrt{\log(q')/n}\right)$ error. Since $\widetilde{\Delta}_j(t^*-)\phi(\mathbf{Z}_j(t^*-))$'s are i.i.d. mean zero random variables,

$$\|n^{-1}\sum_{j=1}^n \widetilde{\Delta}_j(t^*-)\phi(\mathbf{Z}_j(t^*-))\|_{\max} = O_p\left(\sqrt{\log(q)/n}\right)$$

by Lemma 11. Multiplying the rates together, we get $\|I_1\|_{\max} = O_p\left(\sqrt{\log(q)\log(q')/n}\right) = o_p(1)$.

I_2 can be expanded as

$$n^{-1/2}\sum_{j=1}^n \int_0^{t^*} G(t)^{-1}\mathbf{M}^g(t)\Big\{I(C_j \geq t)\upsilon_j(t)h(\mathbf{Z}_j(t)) - \frac{\sum_{k=1}^n \omega_k(t-)\upsilon_k(t)h(\mathbf{Z}_k(t))}{\sum_{k=1}^n I(X_k \geq t)} I(X_j \geq t)\Big\} dM_j^c(t)$$

By Lemma 21, $n\left\{\sum_{k=1}^n I(X_k \geq t)\right\}^{-1} = O_p(1)$. The integrand in I_2 is the product of $\mathbf{M}^g(t)$ and a $O_p(1)$ term. Hence, we can apply Lemma 14(ii) to get the order for I_2, $\|I_2\|_{\max} = O_p\left(\sqrt{\log(q')\log(qq')/n}\right) = o_p(1)$.

By (D3), we may further expand I_3 into

$$\begin{aligned} & n^{1/2}\int_0^{t^*} \mathbf{M}^g(t)\otimes G^{-1}(t)n^{-1}\sum_{j=1}^n \Delta_j(t)\nabla\phi(\mathbf{Z}_j(t))^\top \mathbf{d}_j^z(t)dt \\ & + n^{1/2}\int_0^{t^*} \mathbf{M}^g(t)\otimes G^{-1}(t)n^{-1}\sum_{j=1}^n \Delta_j(t)\triangle\phi(\mathbf{Z}_j(t))dN_j^z(t) \\ \triangleq & I_3' + I_3'', \end{aligned}$$

where $\triangle\phi(\mathbf{Z}_j(t)) = \phi(\mathbf{Z}_j(t)) - \phi(\mathbf{Z}_j(t-))$. By assumption on $h(\mathbf{Z})$ and (D3), we have $|\nabla\phi(\mathbf{Z}_j(t))^\top \mathbf{d}_j^z(t)|$ and $\triangle\phi(\mathbf{Z}_j(t))$ are bounded by L_hK and L_hK, respectively. With

$\sup_{t\in[0,t^*]}|\mathbf{M}^g(t)| = O_p\left(\sqrt{\log(q')/n}\right)$ and $N_j^z(t^*) < K_n = o(\sqrt{n/(\log(p)\log(n))})$, we may replace the $\Delta_j(t)$'s by $\widetilde{\Delta}_j(t)$'s with an $o_p(1)$ error. Each $\widetilde{\Delta}_j(t)\nabla\phi(\mathbf{Z}_j(t))^\top \mathbf{d}_j^z(t)$ has mean zero and at most K_n+1 jumps, and it is $(L_hK+K_\phi K)$-Lipschitz between two consecutive jumps under (D3) and conditions on $\phi(\mathbf{z})$. By applying Lemma 13(ii), we get

$$\sup_{t\in[0,t^*]}\left\|n^{-1}\sum_{j=1}^n \widetilde{\Delta}_j(t)\nabla\phi(\mathbf{Z}_j(t))^\top \mathbf{d}_j^z(t)\right\|_{\max} = O_p(\sqrt{\log(nq)/n}).$$

Hence, $\|I_3'\|_{\max} = O_p(\sqrt{\log(q')\log(nq)/n}) + o_p(1) = o_p(1)$. By applying Lemma 13(i) to

$$\{\widetilde{\Delta}_j(t)\triangle h(\mathbf{Z}_j(t)), N_j^z(t) : j = 1,\ldots,n\},$$

we get at the jumps of $N_i^z(t)$'s, at the t_{ik}, satisfy

$$\sup_{i=1,\ldots,n}\sup_{k\in 1\ldots K_T}\left|n^{-1}\sum_{j=1}^n \Delta_j(t_{ik})\triangle\phi(\mathbf{Z}_j(t_{ik}))\right|$$
$$= O_p(\sqrt{\log(nK_nq)/n}) = O_p(\sqrt{\log(nq)/n}).$$

Hence, $\|I_3''\|_{\max} = O_p\left(K_n\sqrt{\log(nq)\log(q')/n}\right) = o_p(1)$. This completes the proof.

(iii) Define $\boldsymbol{\beta}_r = \boldsymbol{\beta}^o + r\{\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^o\}$ and $h_j(r;t) = \bar{\mathbf{Z}}_j(t,\boldsymbol{\beta}_r)$. The subscript j means the j-th element of corespondent vector. By mean-value theorem, we have some $r\in(0,1)$ such that

$$h_j(1;t) - h_j(0;t) = \left(\{\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^o\}^\top \frac{\mathbf{S}^{(2)}(t,\boldsymbol{\beta}_r)S^{(0)}(t,\boldsymbol{\beta}_r) - \mathbf{S}^{(1)}(t,\boldsymbol{\beta}_r)^{\otimes 2}}{S^{(0)}(t,\boldsymbol{\beta}_r)^2}\right)_j$$
$$= \left(\{\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^o\}^\top \sum_{i=1}^n \frac{\omega_i(t)Y_i(t)e^{\boldsymbol{\beta}_r^\top \mathbf{Z}_i(t)}}{nS^{(0)}(t,\boldsymbol{\beta}_r)}\{\mathbf{Z}_i(t) - \bar{\mathbf{Z}}_j(t,\boldsymbol{\beta}_r)\}^{\otimes 2}\right)_j$$

Since each $\left\|\{\mathbf{Z}_i(t) - \bar{\mathbf{Z}}_j(t,\boldsymbol{\beta}_r)\}^{\otimes 2}\right\|_{\max} \le K_3^2$ under (C1), their weighted average

$$\left\|\sum_{i=1}^n \frac{\omega_i(t)Y_i(t)e^{\boldsymbol{\beta}_r^\top \mathbf{Z}_i(t)}}{nS^{(0)}(t,\boldsymbol{\beta}_r)}\{\mathbf{Z}_i(t) - \bar{\mathbf{Z}}_j(t,\boldsymbol{\beta}_r)\}^{\otimes 2}\right\|_{\max} \le K_3^2.$$

Hence, we have shown that

$$\sup_{t\in[0,t^*]} \|\bar{\mathbf{Z}}(t,\boldsymbol{\beta}^o) - \bar{\mathbf{Z}}(t,\widetilde{\boldsymbol{\beta}})\|_\infty \le \|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^o\|_1 K_3^2 = O_p(\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^o\|_1).$$

By a similar argument, we can show for some $r \in (0,1)$

$$\frac{e^{\boldsymbol{\beta}^{o\top}\mathbf{Z}_i(t)}}{S^{(0)}(t,\boldsymbol{\beta}^o)} - \frac{e^{\widetilde{\boldsymbol{\beta}}^\top\mathbf{Z}_i(t)}}{S^{(0)}(t,\widetilde{\boldsymbol{\beta}})} = \frac{e^{\boldsymbol{\beta}_r^\top\mathbf{Z}_i(t)}(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^o)^\top}{S^{(0)}(t,\boldsymbol{\beta}_r)} \sum_{j=1}^n \frac{\omega_j(t)Y_j(t)e^{\boldsymbol{\beta}_r^\top\mathbf{Z}_j(t)}}{nS^{(0)}(t,\boldsymbol{\beta}_r)}\{\mathbf{Z}_i(t) - \mathbf{Z}_j(t)\}.$$

On event $\{\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^o\|_1 \le K, n^{-1}\sum_{i=1}^n I(X_i \ge t^*) \ge e^{-K}\rho_2/2\}$, we have

$$\inf_{t\in[0,t^*]} S^{(0)}(t,\widetilde{\boldsymbol{\beta}}) > r^*/2 * e^{2K}, \quad \inf_{t\in[0,t^*]} S^{(0)}(t,\boldsymbol{\beta}^o) > r^*/2 * e^K.$$

Hence,

$$|e^{\boldsymbol{\beta}^{o\top}\mathbf{Z}_i(t)}/S^{(0)}(t,\boldsymbol{\beta}^o) - e^{\widetilde{\boldsymbol{\beta}}^\top\mathbf{Z}_i(t)}/S^{(0)}(t,\widetilde{\boldsymbol{\beta}})| \le \|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^o\|_1 2K_3 e^{4K}e^K/\rho_2 = O_p\left(\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^o\|_1\right).$$

The event occurs with probability tending to one because we have$\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^o\|_1 = o_p(1)$ from Theorem 8 and $\sup_{t\in[0,t^*]}|S^{(0)}(t,\boldsymbol{\beta}^o)^{-1}| = O_p(1)$ from Lemma 21.

□

Proof of Lemma 21. Consider the event

$$\Omega_1^* = \left\{ n^{-1}\sum_{i=1}^n I(X_i \ge t^*)I(\varepsilon_i = 1) \ge e^{-K}\rho_2/2 \right\}.$$

Each $I(X_i \ge t^*)I(\varepsilon_i = 1)$ is i.i.d. with expectation $G(t^*)\mathbb{E}[\{F_1(\infty;\mathbf{Z}) - F_1(t^*;\mathbf{Z})\}]$. Applying Lemma 11 under (2.41) and (2.48) from (C1) and (D1), we get that Ω_1 occurs with probability $1 - e^{-ne^{-2K}\rho_2^2}$.

Apparently, we have $I(X_i \ge t^*) \ge I(X_i \ge t^*)I(\varepsilon_i = 1)$. Morevoer, $S^{(0)}(t,\boldsymbol{\beta}^o)$ and $\widetilde{S}^{(0)}(t,\boldsymbol{\beta}^o)$ are both lower bounded by $n^{-1}\sum_{i=1}^n I(X_i \ge t^*)e^{-K}$.

On Ω_1, $\sup_{t\in[0,t^*]}|n/\{\sum_{i=1}^n I(X_i\geq t^*)\}|\leq 2e^K/\rho_2$ and

$$\max\left\{\sup_{t\in[0,t^*]}|S^{(0)}(t,\boldsymbol{\beta}^o)^{-1}|,\ \sup_{t\in[0,t^*]}|\widetilde{S}^{(0)}(t,\boldsymbol{\beta}^o)^{-1}|\right\}\leq 2e^K e^K/\rho_2.$$

□

Proof of Lemma 22. To simplify notation, wherever possible we will use $\widehat{\Gamma}_j(\boldsymbol{\gamma})=\Gamma_j(\boldsymbol{\gamma},\widehat{\boldsymbol{\beta}})$.

(i) We want to prove that for all $j=1,\ldots,p$, the differences $\widetilde{\boldsymbol{\gamma}}_j:=\widehat{\boldsymbol{\gamma}}_j-\boldsymbol{\gamma}_j^*$ belong to a certain convex cone.

It follows from the KKT conditions that, for $l=1,\ldots,p-1$,

$$\begin{cases}\frac{\partial\widehat{\Gamma}_j(\widehat{\boldsymbol{\gamma}}_j)}{\partial\boldsymbol{\gamma}_{j,l}}+\lambda_j\mathrm{sgn}(\widehat{\boldsymbol{\gamma}}_{j,l})=0 & \text{if } \widehat{\boldsymbol{\gamma}}_{j,l}\neq 0;\\ \left|\frac{\partial\widehat{\Gamma}_j(\widehat{\boldsymbol{\gamma}}_j)}{\partial\boldsymbol{\gamma}_{j,l}}\right|\leq\lambda_j & \text{if } \widehat{\boldsymbol{\gamma}}_{j,l}=0.\end{cases}$$

Denote $O_j:=\left\{l\in\{1,\ldots,p-1\}:\boldsymbol{\gamma}_{j,l}^*\neq 0\right\}$ and $O_j^c:=\{1,\ldots,p-1\}\setminus O_j$. For $\xi_j>1$, it follows from the KKT conditions above that on the event

$$\Omega_0:=\{\|\nabla_{\boldsymbol{\gamma}}\widehat{\boldsymbol{\Gamma}}_j(\boldsymbol{\gamma}_j^*)\|_\infty\leq(\xi_j-1)\lambda_j/(\xi_j+1)\},$$

with $\bar{\boldsymbol{\gamma}}_j=\alpha\widehat{\boldsymbol{\gamma}}_j+(1-\alpha)\boldsymbol{\gamma}_j^*$, $\alpha\in(0,1)$

$$\begin{aligned}0&\leq 2\widetilde{\boldsymbol{\gamma}}_j^\top\nabla_{\boldsymbol{\gamma}}^2\widehat{\boldsymbol{\Gamma}}_j(\bar{\boldsymbol{\gamma}}_j)\widetilde{\boldsymbol{\gamma}}_j\\&=\widetilde{\boldsymbol{\gamma}}_j^\top\left\{\nabla_{\boldsymbol{\gamma}}\widehat{\boldsymbol{\Gamma}}_j(\widehat{\boldsymbol{\gamma}}_j)-\nabla_{\boldsymbol{\gamma}}\widehat{\boldsymbol{\Gamma}}_j(\boldsymbol{\gamma}_j^*)\right\}\\&=\sum_{l\in O_j^c}\widetilde{\boldsymbol{\gamma}}_{j,l}\frac{\partial\widehat{\Gamma}_j(\widehat{\boldsymbol{\gamma}}_j)}{\partial\boldsymbol{\gamma}_{j,l}}+\sum_{l\in O_j}\widetilde{\boldsymbol{\gamma}}_{j,l}\frac{\partial\widehat{\Gamma}_j(\widehat{\boldsymbol{\gamma}}_j)}{\partial\boldsymbol{\gamma}_{j,l}}-\widetilde{\boldsymbol{\gamma}}_j^\top\nabla_{\boldsymbol{\gamma}}\widehat{\boldsymbol{\Gamma}}_j(\boldsymbol{\gamma}_j^*)\\&\leq-\lambda_j\sum_{l\in O_j^c}\widehat{\boldsymbol{\gamma}}_{j,l}\mathrm{sgn}(\widehat{\boldsymbol{\gamma}}_{j,l})+\lambda_j\sum_{l\in O_j}|\widetilde{\boldsymbol{\gamma}}_{j,l}|+\frac{(\xi_j-1)\lambda_j}{\xi_j+1}\|\widetilde{\boldsymbol{\gamma}}_{j,O_j}\|_1+\frac{(\xi_j-1)\lambda_j}{\xi_j+1}\|\widetilde{\boldsymbol{\gamma}}_{j,O_j^c}\|_1\\&=-\frac{2\lambda_j}{\xi_j+1}\|\widetilde{\boldsymbol{\gamma}}_{j,O_j^c}\|_1+\frac{2\xi_j\lambda_j}{\xi_j+1}\|\widetilde{\boldsymbol{\gamma}}_{j,O_j}\|_1.\end{aligned}$$

(ii) Let $\mathbf{v} = \widetilde{\gamma}/\|\widetilde{\gamma}\|_1$ be the l_1-standardized direction for $\widetilde{\gamma} = \widehat{\gamma} - \gamma^*$. By part (i) and convexity of Γ_j in γ_j, any $x \in (0, \|\widetilde{\gamma}\|_1]$ satisfies

$$\mathbf{v}^\top \left\{ \nabla_\gamma \widehat{\mathbf{\Gamma}}_j(\gamma^* + x\mathbf{v}) - \nabla_\gamma \widehat{\mathbf{\Gamma}}_j(\gamma^*) \right\} \leq -\frac{2\lambda_j}{\xi_j + 1} \|\mathbf{v}_{O_j^c}\|_1 + \frac{2\xi_j \lambda_j}{\xi_j + 1} \|\mathbf{v}_{O_j}\|_1.$$

We relax the inequality about x above to establish an upper bound for $\|\widetilde{\gamma}\|_1$. By the definition of κ_j, the left hand side can be bounded by

$$\mathbf{v}^\top \left\{ \nabla_\gamma \widehat{\mathbf{\Gamma}}_j(\gamma^* + x\mathbf{v}) - \nabla_\gamma \widehat{\mathbf{\Gamma}}_j(\gamma^*) \right\} = x\mathbf{v}^\top \nabla_\gamma^2 \widehat{\mathbf{\Gamma}}_j(\gamma^*)\mathbf{v} \geq \frac{x\|\mathbf{v}_{O_j}\|_1^2 \kappa_j(\xi_j, O_j)}{s_j}.$$

The right hand side can be bounded using the complete square $\{\|\mathbf{v}_{O_j}\|_1 - 2/(\xi_j + 1)\}^2$,

$$-\frac{2\lambda_j}{\xi_j + 1} \|\mathbf{v}_{O_j^c}\|_1 + \frac{2\xi_j \lambda_j}{\xi_j + 1} \|\mathbf{v}_{O_j}\|_1 = 2\lambda_j \|\mathbf{v}_{O_j}\|_1 - \frac{2\lambda_j}{\xi_j + 1} \leq \lambda_j(\xi_j + 1)\|\mathbf{v}_{O_j}\|_1^2.$$

Combining the bounds for both sides in the inequality, we get an upper bound for $\|\widetilde{\gamma}\|_1$.

□

Proof of Lemma 23. We define

$$\widetilde{\Gamma}_j(\gamma) = n^{-1} \sum_{i=1}^{n} \int_0^{t^*} \{Z_{ij}(t) - \mu_j(t) - \gamma^\top \mathbf{Z}_{i,-j}(t) + \gamma^\top \boldsymbol{\mu}_{-j}(t)\}^2 dN_i^o(t).$$

By Lemmas 20 and 6,

$$\max_{j=1,\ldots,p} \|\nabla_\gamma \widehat{\mathbf{\Gamma}}_j(\gamma_j^*, \widehat{\boldsymbol{\beta}}) - \nabla_\gamma \widetilde{\mathbf{\Gamma}}_j(\gamma_j^*)\|_\infty = O_p\left(\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^o\|_1 + \sqrt{\log(p)/n} \right).$$

$\nabla_\gamma \widetilde{\mathbf{\Gamma}}_j(\gamma_j^*)$ is the average of i.i.d. vectors with mean $\nabla_\gamma \bar{\mathbf{\Gamma}}_j(\gamma_j^*) = \mathbf{0}$ and maximal bound $K^2(1+K)$. We can apply Lemma 11 to the matrix $(\nabla_\gamma \widetilde{\mathbf{\Gamma}}_1(\gamma_1^*), \ldots, \nabla_\gamma \widetilde{\mathbf{\Gamma}}_p(\gamma_p^*))$ to get

$$\max_{j=1,\ldots,p} \|\nabla_\gamma \widetilde{\mathbf{\Gamma}}_j(\gamma_j^*)\|_\infty = \|(\nabla_\gamma \widetilde{\mathbf{\Gamma}}_1(\gamma_1^*), \ldots, \nabla_\gamma \widetilde{\mathbf{\Gamma}}_p(\gamma_p^*))\|_{\max}$$

$$= O_p(\sqrt{\log(p^2)/n}) = O_p(\sqrt{\log(p)/n}).$$

□

Proof of Lemma 24. (i) We define

$$\widetilde{\boldsymbol{\Sigma}} = n^{-1}\sum_{i=1}^{n}\int_0^{t^*}\{\mathbf{Z}_i(t) - \boldsymbol{\mu}(t)\}^{\otimes 2} dN_i^o(t).$$

The total variation of each $N_i^o(t)$ is at most 1. By Lemma 20, we have

$$\sup_{t\in[0,t^*]} \|\bar{\mathbf{Z}}(t,\widehat{\boldsymbol{\beta}}) - \boldsymbol{\mu}\|_\infty = O_p\left(\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^o\|_1 + \sqrt{\log(p)/n}\right).$$

Hence,

$$\|\widehat{\boldsymbol{\Sigma}} - \widetilde{\boldsymbol{\Sigma}}\|_{\max} \le 2KO_p\left(\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^o\|_1 + \sqrt{\log(p)/n}\right) = O_p\left(\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^o\|_1 + \sqrt{\log(p)/n}\right).$$

Now, $\widetilde{\boldsymbol{\Sigma}}$ is average of i.i.d. with mean $\boldsymbol{\Sigma}$ and bounded maximal norm K^2. We apply Lemma 11 with union bound,

$$\Pr\left(\|\widetilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_{\max} \ge K^2 x\right) \le 2p^2 e^{2nx^2}.$$

Choosing $x = \sqrt{\log(2p^2/\varepsilon)/(2n)}$, we have $\|\widetilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_{\max} = O_p(\sqrt{\log(p)/n})$.

(ii) We alternatively use the following form

$$\ddot{\mathbf{m}}(\boldsymbol{\beta}) = n^{-1}\sum_{i=1}^{n}\int_0^{t^*}\left\{n^{-1}\sum_{j=1}^{n}\frac{\omega_j(t)Y_j(t)e^{\boldsymbol{\beta}^\top\mathbf{Z}_j(t)}}{S^{(0)}(t,\boldsymbol{\beta})}\mathbf{Z}_i(t)^{\otimes 2} - \bar{\mathbf{Z}}(t,\boldsymbol{\beta})^{\otimes 2}\right\} dN_i^o(t).$$

By Lemma 20(iii), we have

$$\|\ddot{\mathbf{m}}(\widetilde{\boldsymbol{\beta}}) - \ddot{\mathbf{m}}(\boldsymbol{\beta}^o)\|_{\max} = O_p(\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^o\|_1).$$

We also have a similar form for

$$\ddot{\widetilde{\mathbf{m}}}(\boldsymbol{\beta}) = n^{-1}\sum_{i=1}^{n}\int_0^{t^*}\left\{n^{-1}\sum_{j=1}^{n}\frac{I(C_j\geq t)Y_j(t)e^{\boldsymbol{\beta}^\top\mathbf{Z}_j(t)}}{\widetilde{S}^{(0)}(t,\boldsymbol{\beta})}\mathbf{Z}_i(t)^{\otimes 2} - \widetilde{\mathbf{Z}}(t,\boldsymbol{\beta})^{\otimes 2}\right\}dN_i^o(t).$$

By Lemma 20(i), we have

$$\|\ddot{\mathbf{m}}(\boldsymbol{\beta}^o) - \ddot{\widetilde{\mathbf{m}}}(\boldsymbol{\beta}^o)\|_{\max} = O_p\left(\sqrt{\log(p)/n}\right).$$

Finally, we use the martingale property of

$$\begin{aligned}
\ddot{\widetilde{\mathbf{m}}}(\boldsymbol{\beta}^o) - \widetilde{\boldsymbol{\Sigma}} &= n^{-1}\sum_{i=1}^{n}\int_0^{t^*}\left\{\frac{\widetilde{\mathbf{S}}^{(2)}(t,\boldsymbol{\beta}^o)}{\widetilde{S}^{(0)}(t,\boldsymbol{\beta}^o)} - \widetilde{\mathbf{Z}}(t,\boldsymbol{\beta}^o)^{\otimes 2}\right\}I(C_i\geq t)dM_i^1(t)\\
&- n^{-1}\sum_{i=1}^{n}\int_0^{t^*}\{\mathbf{Z}_i(t) - \widetilde{\mathbf{Z}}(t,\boldsymbol{\beta}^o)\}^{\otimes 2}I(C_i\geq t)dM_i^1(t)\\
&+ n^{-1}\sum_{i=1}^{n}\int_0^{t^*}\left[\{\mathbf{Z}_i(t) - \widetilde{\mathbf{Z}}(t,\boldsymbol{\beta}^o)\}^{\otimes 2} - \{\mathbf{Z}_i(t) - \boldsymbol{\mu}(t)\}^{\otimes 2}\right]I(C_i\geq t)dN_i^o(t)
\end{aligned}$$

under filtration $\mathcal{F}_t^*$. The integrands in the first two martingale terms are bounded by K^2. Hence, we can apply Lemma 14(ii) to obtain that their maximal norms are both $O_p\left(\sqrt{\log(p)/n}\right)$. We apply Lemma 20(i) to the integrand of the third term, equivalently expressed as

$$\{\boldsymbol{\mu}(t) - \widetilde{\mathbf{Z}}(t,\boldsymbol{\beta}^o)\}\{\mathbf{Z}_i(t) - \widetilde{\mathbf{Z}}(t,\boldsymbol{\beta}^o)\}^\top + \{\mathbf{Z}_i(t) - \boldsymbol{\mu}(t)\}\{\boldsymbol{\mu}(t) - \widetilde{\mathbf{Z}}(t,\boldsymbol{\beta}^o)\}^\top.$$

Therefore, we obtain $\|\ddot{\widetilde{\mathbf{m}}}(\boldsymbol{\beta}^o) - \widetilde{\boldsymbol{\Sigma}}\|_{\max} = O_p\left(\sqrt{\log(p)/n}\right)$.

We put the rates together by the triangle inequality.

□

Proof of Lemma 25. The proof is similar to that of Lemma 5. Define the compatibility factor for $\mathcal{C}_j(\xi_j, O_j)$ and symmetric matrix $\boldsymbol{\Phi}$ as

$$\kappa_j(\xi_j, O_j; \boldsymbol{\Phi}) = \sup_{0\neq\mathbf{g}\in\mathcal{C}_j(\xi_j,O_j)}\frac{\sqrt{s_j\mathbf{g}^\top\boldsymbol{\Phi}\mathbf{g}}}{\|\mathbf{g}_{O_j}\|_1}.$$

Apparently, $\kappa_j(\xi_j, O_j) = \kappa_j\big(\xi_j, O_j; \nabla^2_{\gamma}\mathbf{\Gamma}(\gamma^*, \widehat{\boldsymbol{\beta}})\big)$. Notice that

$$\nabla^2_{\gamma}\mathbf{\Gamma}(\gamma^*, \widehat{\boldsymbol{\beta}}) = n^{-1}\sum_{i=1}^{n}\int_0^{t^*}\{\mathbf{Z}_{i,-j}(t) - \bar{\mathbf{Z}}_{-j}(t, \widehat{\boldsymbol{\beta}})\}^{\otimes 2} dN_i^o(t) = \widehat{\mathbf{\Sigma}}_{-j,-j},$$

where $\widehat{\mathbf{\Sigma}}_{-j,-j}$ is a $\widehat{\mathbf{\Sigma}}$ dropping its jth row and column. By Lemma 4.1 in [HSY$^+$13] (for a similar result, see [vdGB09] Corollary 10.1),

$$\kappa_j(\xi_j, O_j)^2 = \kappa_j^2\big(\xi_j, O_j; \widehat{\mathbf{\Sigma}}_{-j,-j}\big) \geq \kappa_j^2(\xi_j, O_j; \mathbf{\Sigma}_{-j,-j}) - s_j(\xi_j + 1)^2\|\mathbf{\Sigma}_{-j,-j} - \widehat{\mathbf{\Sigma}}_{-j,-j}\|_{\max}.$$

For any non-zero $\mathbf{g} \in \mathbb{R}^{p-1}$, let $\mathbf{g}^*$ be its embedding into $\mathbb{R}^p$ defined as

$$g_k^* = \begin{cases} g_k & k < j \\ 0 & k = j \\ g_{k-1} & k > j \end{cases}$$

Then, we may establish a lower bound for the smallest eigenvalue of $\mathbf{\Sigma}_{-j,-j}$ by (D2)

$$\inf_{0\neq\mathbf{g}\in\mathbb{R}^{p-1}} \mathbf{g}^\top \mathbf{\Sigma}_{-j,-j}\mathbf{g} = \inf_{0\neq\mathbf{g}\in\mathbb{R}^{p-1}} \mathbf{g}^{*\top}\mathbf{\Sigma}\mathbf{g}^* \geq \rho\|\mathbf{g}\|_2^2.$$

Hence, $\inf_{j=1,\ldots,p}\kappa_j^2(\xi_j, O_j; \mathbf{\Sigma}_{-j,-j}) \geq \rho$. Using the result in Lemma 24(i) under (D4), we have

$$\inf_{j=1,\ldots,p}\kappa_j(\xi_j, O_j)^2 \geq \rho - \|\mathbf{\Sigma} - \widehat{\mathbf{\Sigma}}\|_{\max} s_{\max} \max_{j=1,\ldots,p}(\xi_j + 1)^2 = \rho - o_p(1).$$

Therefor, if $\xi_{\max} \asymp 1$, we must have that $\big\{\inf_j \kappa_j(\xi_j, O_j)^2 \geq \rho/2\big\}$ occurs with probability tending to one. □

2.8 Acknowledgement

We would like to acknowledge our collaboration with Dr. James Murphy of the UC San Diego Department of Radiation Medicine and Applied Sciences on the linked Medicare-SEER

data analysis project that motivated this work. We would also like to thank his group for help in preparing the data set.

Chapter 2, in full, has been accepted for publication of the material as it may appear in the Electronic Journal for publication of the material. Hou, Jue; Bradic, Jelena; Xu, Ronghui. Inference under Fine-Gray competing risks model with high-dimensional covariates. The dissertation/thesis author was the primary investigator and author of this paper.

Chapter 3

Estimating Treatment Effect for Time-to-Event Outcome with High-dimensional Covariates in Observational Studies

3.1 INTRODUCTION

The proliferation of publicly accessible “big data" from Electronic Health Records (EHR) provides an abundant resource to study the effect of various treatments on the patients. This type of comparative effectiveness studies serve as the alternative or exploratory projects when a randomized trial is implausible or uneconomical [HYB$^+$10, SKD$^+$15]. With the availability of linked large databases the challenge in studying causal treatment effects, is to handle a large

"$p > n$" number of potential confounders.

Motived by studies in cancer, we consider the situation where the treatment effect of interest is on a survival outcome, while having to account for a large number of potential confounders. Despite scientific knowledge [HHWM02], in databases like these whether each covariate is a true confounder is often unknown. Machine learning methods had been considered for confounder selection in this type of high dimensional settings, but it is now known that directly applying such methods can lead to bias in the estimated treatment effect. This is well understood, for example, in the case of regularization, the control of estimation variance with diverging dimensions is achieved at the cost of estimation bias [vdGB11].

Orthogonal score is a familiar concept from the semiparametric statistics literature [New90, BKRW98]. It relates to the profile likelihood and the least favorable direction in likelihood inference, which includes nonparametric likelihood for semiparametric models [SW92, MvdV00]. The efficient score function generated by the profile likelihood is a special case of the orthogonal score function. It is known that an estimator obtained from an orthogonal score function should not be affected by the bias (or equivalently, slower than root-n convergence) in the estimation of the nuisance parameters, or misspecification of the nonparametric (i.e. nuisance parameter) part of the model [New90, BKRW98]. Recently the orthogonal score has been applied for debiasing purposes in estimating treatment effects [BCH13, Far15, CCD$^+$18], and was referred to as [Ney59] orthogonality.

There is a connection between orthogonal score and double robustness that has not always been made explicit in the literature. Doubly robust (DR) property refers to the context of causal inference and missing data problems, where there are at least two working models, one for the

outcome and another for the missing data mechanism (or equivalently, treatment assignment since the same subject cannot be observed under more than one treatment). An estimator is doubly robust if it is consistent as long as one of the two working models is correct [RR95, RR01, BR05]. When p is larger than n, [Far15] showed that the doubly robust estimator in [RR95] is still consistent. For time-to-event outcome subject to censoring, DR estimators have been studied by [ZS12], [ZZYK15], [KLZ18], [WLL$^+$17] and [JLS$^+$17] when the number of covaraites is fixed. A form of DR estimators was proposed in [RMN92] where the score function was the product of the error terms of the two working models [RRvdL00, VBC12, KLZ18]. By the error terms we mean the observed outcome or treatment assignment minus its expectation under the working model. Such a score function turns out to be very similar in its form as an orthogonal score under certain models, as will be seen in later works and our derivation below.

In this paper, we use the propensity score model for treatment assignment to construct an orthogonal score function for the estimation of the treatment effect. In addition, we also consider the cases where the propensity score model might be wrong, or the specified survival outcome model might be wrong, and in the high dimensional setting the sparsity assumption is violated. We study such double robustness properties of our estimator of the treatment effect. The organization of the rest of the paper is as follows. In Section 3.2, we propose inference method on treatment effect when both the Aalen additive hazards model and the logistic regression propensity model are correct. In Section 3.3, we develop doubly robust estimation with the extensions in closed form estimator, cross-fitting and further regularized estimators. Section 3.4 contains an extensive simulation study. In Section 3.5, we apply our method to the empirical study on the treatment effect of radical prostatectomy versus conservative management using SEER-Medicare

Linked Data. The conclusions and discussions are given in Section 3.6. The detail of theoretical derivations is given in the Section 3.7.

3.2 Treatment Effect with High-Dimensional Covariates

3.2.1 Model and Orthogonal Score

Let T be the event time of interest, C be the censoring time. We observe $X = \min(T,C)$ and $\delta = I(T \leq C)$, where $I(\cdot)$ is the indicator function. Let D be the treatment assignment, and $\mathbf{Z}$ be a $p \times 1$ vector of covariates. Let $\lambda(t;D,\mathbf{Z})$ be the conditional hazard function of T given D and $\mathbf{Z}$. Under the additive hazards model [CO84, Tho86, ED87, LY94],

$$\lambda(t;D,\mathbf{Z}) = \lambda_0(t) + D\theta + \beta^\top \mathbf{Z}, \tag{3.1}$$

where $\lambda_0(\cdot)$ is the baseline hazard function. The treatment effect under model (3.1) is θ, which is our parameter of interest. Our goal is to draw inference on θ while allowing the dimension of the covariates p to be much larger than the sample size n.

In this paper, we consider binary treatment assignments. For example, we may assume the logistic regression model for D given $\mathbf{Z}$:

$$\mathbb{P}(D = 1|\mathbf{Z}) = \frac{e^{\gamma^\top \mathbf{Z_1}}}{1 + e^{\gamma^\top \mathbf{Z_1}}}, \tag{3.2}$$

where $\mathbf{Z_1} = (1, Z_1, \ldots, Z_p)^\top$ represents the vector of covariates with the intercept term. The above conditional probability of treatment assignment is often called the propensity score in the literature.

In the following we will use $W = (X, \delta, D, \mathbf{Z})$ to denote a single copy of n independent and identically distributed observations. Let $\Lambda(\cdot) = \int_0^{\cdot} \lambda_0(u)du$ denote the baseline cumulative hazard function; and the subscript '0' in the following will be used to index the true parameter value under which the data are generated. Following the convention of [AG82], we denote the counting process and at-risk process for subjects $i = 1, \ldots, n$ as $N_i(t) = \delta_i I(X_i \leq t)$ and $Y_i(t) = I(X_i \geq t)$, respectively. By Doob-Meyer decomposition we define

$$M_i(t; \boldsymbol{\beta}, \Lambda) = N_i(t) - \int_0^t Y_i(u)\{(D_i\theta + \boldsymbol{\beta}^\top \mathbf{Z}_i)du + d\Lambda(u)\}, \tag{3.3}$$

which is a martingale with respect to the filtration $\mathcal{F}_{n,t} = \sigma\{N_i(u), Y_i(u), D_i, \mathbf{Z}_i : u \leq t, i = 1, \ldots, n\}$ when evaluated at the true parameter values.

In the presence of high dimensional covariates, directly fitting model (3.1) via regularization methods such as LASSO is known to lead to bias in the estimate of coefficient $(\theta, \boldsymbol{\beta}^\top)^\top$; this is also illustrated in our simulation results later (Table 3.1). Instead we consider the orthogonal score for estimating the treatment effect. Writing the nuisance parameter $\eta = (\boldsymbol{\beta}, \Lambda, \boldsymbol{\gamma})$, a score function ψ is an orthogonal score for θ if the Gâteaux derivative with respect to η

$$\frac{\partial}{\partial r}\mathbb{E}\{\psi(\theta_0; \eta_0 + r\triangle\eta)\}\Big|_{r=0} = 0, \tag{3.4}$$

where θ_0 and η_0 are the true values, respectively, and $\triangle\eta = \eta - \eta_0$. In other words, the orthogonality of a score function is defined as the local invariance of the score to a small perturbation in the nuisance parameter around the true parameters. Under orthogonality, the estimation to the treatment effect is not affected by the convergence rate of any consistent estimation to the nuisance parameter, as illustrated in Figure 3.1. We note that this orthogonality

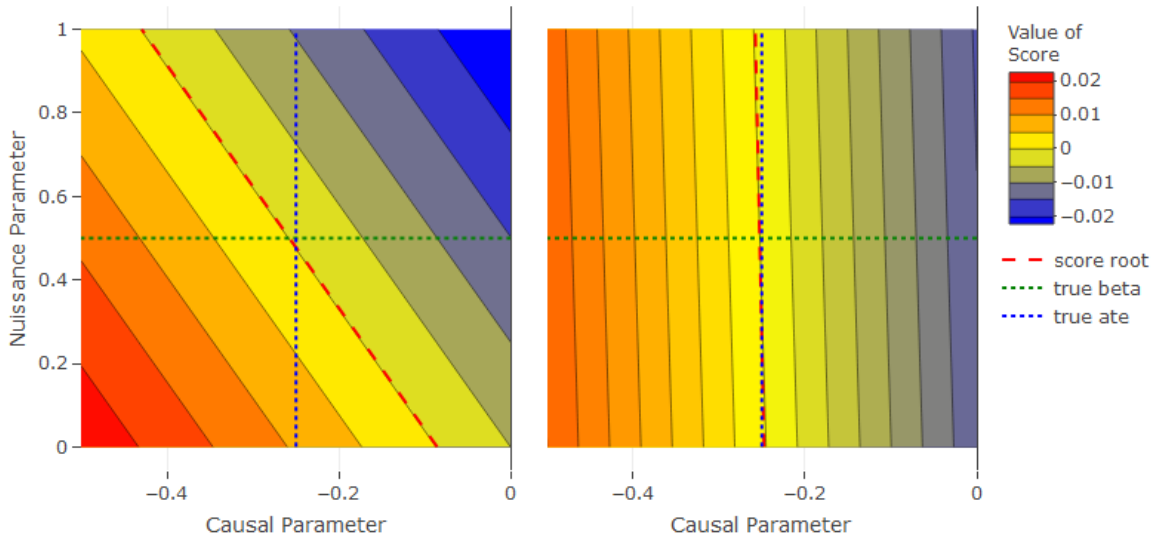


Figure 3.1: The contour for score functions with simulated data under additive hazards model and logistic regression models at sample size 5000. Left - the score without orthogonality [LY94]; Right - our orthogonal score with true propensity score. The orthogonal score is robust to error in nuisance parameter for estimation of the causal parameter.

was given in [New94] as the condition such that the limiting distribution of estimators of the parameters of interest is not affected by the estimation of the nuisance parameters.

In order to construct the orthogonal score for our parameter of interest, θ, we utilize the propensity model (3.2). We state our score and its orthogonality in the following Lemma.

Lemma 26. *Under models* (3.1) *and* (3.2) *and the assumption*

$$C \perp\!\!\!\perp (T, D) | \mathbf{Z}, \tag{3.5}$$

the score

$$\phi(\theta;\boldsymbol{\beta},\Lambda,\boldsymbol{\gamma}) = \frac{1}{n}\sum_{i=1}^{n}\left\{D_i - \text{expit}(\boldsymbol{\gamma}^\top \mathbf{Z}_{\mathbf{1}i})\right\}\int_0^\tau e^{D_i\theta t} dM_i(t;\boldsymbol{\beta},\Lambda), \tag{3.6}$$

where $\tau < \infty$ *is an upper limit of time, identifies the true parameters* $(\theta_0;\boldsymbol{\beta}_0,\Lambda_0,\boldsymbol{\gamma}_0)$ *and is an orthogonal score for* θ.

To utilize the orthogonal score to estimate θ, (3.6) is seen as an equation for θ only, and after solving for θ we plug in a consistent estimate of the nuisance parameter $(\boldsymbol{\beta},\Lambda,\boldsymbol{\gamma})$ to obtain $\widehat{\theta}$. Under proper regularity conditions regarding how the dimensions expand, we find that the asymptotic variance of $\widehat{\theta}$ may follow from the classic estimating equation theory, with $(\boldsymbol{\beta},\Lambda,\boldsymbol{\gamma})$ again replaced by its consistent estimator to obtain a consistent estimate of the asymptotic variance.

The condition (3.5) is stronger than the usual non-informative censoring assumption, where $C \perp\!\!\!\perp T|(D,\mathbf{Z})$. This condition can be relaxed if a consistent estimator of the conditional distribution of C given D and Z is available, and we discuss this more later.

3.2.2 Inference on θ

Before proceeding we extend the above description of how to use (3.6) by allowing an estimator of the baseline cumulative hazard function to depend θ, $\widehat{\Lambda}(\cdot;\theta)$. This is natural since under the additive hazards model (3.1), the usual estimator of $\Lambda(t)$ is a function of $\widehat{\theta}$ [LY94]. The extension is useful for the doubly robust estimator that we develop in the next section. Under conditions specified below, the property of the orthogonal score still holds with the 'estimator' of the nuisance parameter $(\widehat{\boldsymbol{\beta}},\widehat{\Lambda}(\cdot;\theta),\widehat{\boldsymbol{\gamma}})$.

For the purpose of guaranteeing firm guarantees on inference regarding θ, that is not dependent on exact model-selection consistency of regularized estimators, we make the following set of assumptions.

Assumption 5.

(i) *W is generated according to models* (3.1) *and* (3.2)*;*

(ii) $C \perp\!\!\!\perp (T,D)|\mathbf{Z}$*;*

(iii) $\mathbb{P}\left(\sup_{i=1,\ldots,n} \|\mathbf{Z}_i\|_\infty < K_Z\right) = 1$*;*

(iv) $\sup_{t\in[0,\tau]} \lambda_0(t) < K_\Lambda$*;*

(v) $\mathbb{E}\{\mathbb{E}(Y(\tau)|\mathbf{Z}; D=0)\,\mathrm{Var}(D|\mathbf{Z})\} \geq \varepsilon_Y > 0$*;*

(vi) $\mathbb{E}\{\mathbb{E}(N(\tau)|\mathbf{Z}; D=0)\,\mathrm{Var}(D|\mathbf{Z})\} \geq \varepsilon_N > 0$*;*

(vii) *the total variation of $\widehat{\Lambda}(\cdot;\theta)$ is bounded by K_v uniformly in θ with probability tending to one;*

(viii) *$\widehat{\Lambda}(t;\theta)$ is approximately linear in θ with respect to the total variation in the neighborhood of θ_0,*

$$\bigvee_{t=0}^{\tau} \left\{\widehat{\Lambda}(t;\theta) - \widehat{\Lambda}(t;\theta_0)\right\} = O_p(|\theta - \theta_0|); \tag{3.7}$$

(ix) *the rates of estimation errors follow*

$$\log(p)\|\widehat{\beta} - \beta_0\|_1 + \sup_{t\in[0,\tau]} |\widehat{\Lambda}(t;\theta_0) - \Lambda_0(t)| + \|\widehat{\gamma} - \gamma_0\|_1$$

$$+\sqrt{n}\|\widehat{\gamma}-\gamma_0\|_1\left(\|\widehat{\beta}-\beta_0\|_1+\sup_{t\in[0,\tau]}|\widehat{\Lambda}(t;\theta_0)-\Lambda_0(t)|\right)=o_p(1), \tag{3.8}$$

and the estimation error to the baseline hazard follows additionally

$$\int_0^\tau H(t)d\{\widehat{\Lambda}(t,\theta_0)-\Lambda_0(t)\}=o_p(1) \tag{3.9}$$

for any process $H(t)$ adapted to the filtration $\mathcal{F}_{n,t}=\sigma\{N_i(u),Y_i(u),D_i,\mathbf{Z}_i:u\le t,i=1,\ldots,n\}$ with tight uniform bound, $\sup_{t\in[0,\tau]}|H(t)|=O_p(1)$.

Then we have:

Theorem 9. *Under Assumption 5 , $\widehat{\theta}$ that solves $\phi(\theta;\widehat{\beta},\widehat{\Lambda}(\theta),\widehat{\gamma})=0$ satisfies*

$$\widehat{\sigma}^{-1}\sqrt{n}(\widehat{\theta}-\theta_0)\rightsquigarrow N(0,1), \tag{3.10}$$

where the variance estimator takes the closed form

$$\widehat{\sigma}^2=\frac{n^{-1}\sum_{i=1}^n\delta_i\{D_i-\operatorname{expit}(\widehat{\gamma}^\top\mathbf{Z}_{\mathbf{1}i})\}^2e^{2\widehat{\theta}D_iX_i}}{\left\{n^{-1}\sum_{i=1}^n(1-D_i)\operatorname{expit}(\widehat{\gamma}^\top\mathbf{Z}_{\mathbf{1}i})X_i\right\}^2}. \tag{3.11}$$

Remark 6. *Several penalization approaches are available to estimate the high-dimensional regression coefficients in the additive hazards model* (3.1) *under various assumptions [GG12b, LL13]. [LM07] proposed a Lasso estimator of the form*

$$\widehat{\beta}=\underset{\beta\in\mathbb{R}^p}{\operatorname{argmin}}\,\beta^\top H_n\beta-2\beta^\top\mathbf{h}_n+\lambda\|\beta\|_1, \tag{3.12}$$

where $H_n=\sum_{i=1}^n\int_0^\tau\{\mathbf{Z}_i-\bar{\mathbf{Z}}(t)\}^{\otimes2}Y_i(t)dt/n$, $\mathbf{h}_n=\sum_{i=1}^n\int_0^\tau\{\mathbf{Z}_i-\bar{\mathbf{Z}}(t)\}dN_i(t)/n$, with $\bar{\mathbf{Z}}(t)=\sum_{i=1}^n\mathbf{Z}_iY_i(t)/\sum_{i=1}^nY_i(t)$ and $a^{\otimes2}=aa^\top$ for a vector a. To estimate β in our model (3.1)*, we may use $\beta_*=(\theta,\beta)$ in* (3.12)*. The estimation of γ from model* (3.2) *is similar; we may use the LASSO estimator for the logistic regression model [SK03]:*

$$\widehat{\gamma}=\underset{\gamma\in\mathbb{R}^{p+1}}{\operatorname{argmin}}-\frac{1}{n}\sum_{i=1}^n\left\{\{D_i\gamma^\top\mathbf{Z}_{\mathbf{1}i}-\log(1+e^{\gamma^\top\mathbf{Z}_{\mathbf{1}i}})\right\}+\lambda\sum_{j=1}^p|\gamma_j|. \tag{3.13}$$

Remark 7. *For $\widehat{\Lambda}$ a natural choice is the Breslow type estimator*

$$\widehat{\Lambda}(t) = \int_0^t \frac{\sum_{i=1}^n \{dN_i(u) - Y_i(u)(\widehat{\beta}_*^\top \mathbf{Z}_{Di})du\}}{\sum_{i=1}^n Y_i(u)} \tag{3.14}$$

with $\mathbf{Z}_{Di} = (D_i, \mathbf{Z}_i^\top)^\top$. According to our Lemma 37 in Section 3.7.5, the total variation of (3.14) *is bounded with large probability as required by Assumption 5-vii. One may also use estimators dependent on θ under Assumption 5-viii. We expand on one special choice $\check{\Lambda}(\cdot, \theta)$ in Section 3.3.1.*

Remark 8. *For the LASSO estimators given in Remark 6, oracle inequality for l^1-estimation error have been established under either the restricted eigenvalue condition [BRT09] or the compatibility condition [vdG07, vdGB09]. By [GG12b], [ZSZH17] and [vdG08], the LASSO estimators with oracle penalty parameters follow*

$$\|\widehat{\beta} - \beta_0\|_1 = O_p\left(s_\beta \sqrt{\log(p)/n}\right), \|\widehat{\gamma} - \gamma_0\|_1 = O_p\left(s_\gamma \sqrt{\log(p)/n}\right) \tag{3.15}$$

under suitable regularity conditions, where s_β and s_γ are the sparsity of β_0 and γ_0, respectively. The estimator of the baseline cumulative hazard $\widehat{\Lambda}$ defined in (3.14) *has a rate of uniform convergence $O_p\left(\|\widehat{\beta} - \beta_0\|_1\right)$, and $\widehat{\Lambda}$ satisfies* (3.9) *whenever $\widehat{\beta}$ is consistent. If the dimensions and the sparsity levels satisfy $s_\beta s_\gamma = o(\sqrt{n}/\log(p))$, then Assumption 5-ix holds. This means that the dimension p can be of ultra-high exponential order when the sparsity levels grow slowly. When p is of polynomial order in n, the sparsity of both models may grow up to rate $n^{1/4}$. Or, the sparsity of one model is allowed to grow as rapid as $\sqrt{n}$ while the sparsity of the other model is constant.*

Remark 9. *Assumption 5-v has two implications. First, the treatment assignments cannot be deterministically decided by the covariates; we need a proportion of subjects with propensity*

strictly between zero and one. Second, there should be a proportion at-risk at the longest possible followup time among these subjects with propensity strictly between zero and one. The first part is a relaxed form of the positivity condition,also known as overlapping condition: the propensity score $\mathbb{P}(D=1|\mathbf{Z})$ *is bounded away from zero and one for every possible* $\mathbf{Z}$ *which applies to all subjects [Imb03, WC10]. In practice, the treatment assignment may violate the positivity condition when a subpopulation is guaranteed to receive the treatment or the control. As an example in our data, the distribution of estimated propensity of receiving received radical prostatectomy is not bounded away from one in Figure 3.4.*

Remark 10. *Assumption 5-iii is a common assumption for high-dimensional non-linear models [HSY*$^+$*13, vdGBRD14]. Assumption 5-iv is a standard assumption under additive hazards model [LY94, LL13]. We rely on the Assumptions 5-iii and 5-iv to obtain the concentration results of various empirical processes. Assumption 5-vi implies a positive event rate among subjects with positive chance of being assigned into either treatment arm.*

The proof of Theorem 9 has two parts. In the first part, we establish the asymptotic equivalence between the score with the estimated nuisance parameter and the score with the true nuisance parameter,

$$\phi(\theta;\widehat{\beta},\widehat{\Lambda}(\cdot,\theta),\widehat{\gamma}) = \phi(\theta;\beta_0,\Lambda_0,\gamma_0) + o_p\left(\sqrt{n}|\theta-\theta_0|+1\right), \tag{3.16}$$

by utilizing one consequence of the orthogonality that the score is insensitive to perturbation in nuisance parameter. In the second part, we use the identifiability from Lemma 26 to establish the asymptotic normality of $\widehat{\theta}$. Our proof allows the hazard contribution from the covariates, $\beta_0^\top \mathbf{Z}_i$, to grow arbitrarily large with the growing dimensions. Our asymptotic normality is the first result

established with unbounded hazards of a time-to-event outcome, which distinguished us from existing literatures on the topic [HBX17, YBS18].

Theorem 9 leads immediately to inference on the treatment effect θ. Using computationally efficient methods like LASSO for the nuisance parameters, our inference procedure adds no extra computational burden, and is therefore ready for practical uses.

3.3 Exploring the Doubly Robust Property

It is immediate from (3.6) that the estimator from solving the orthogonal score equation might be doubly robust (DR) when p is fixed, since it is of the form as the product of the error terms from the two models. While this may not be the most common way to construct a DR estimator in the literature, it was noted in [RMN92] , [VBC12] and [KLZ18]. In the classical literature on misspecified models, the estimators obtained under the misspecified model are known to converge to the so-called 'least false' parameters. Extension of DR approaches into high-dimensional settings so far mainly relies on the convergence of the estimators to the 'least false' parameters, which now need to be well-defined as p increases [Far15]. Few work has studied the asymptotic behavior of estimators under model misspecification under high-dimensional settings [Tan18]. In our empirical analysis, we find that the LASSO estimator with penalty parameter selected by cross-validation deviates substantially from the 'least false' parameters under dense or misspecified model (see Table 3.1). The observation suggests that the theory established on the convergence to the 'least false' parameter may be unsound for directing practice. Moreover, the 'least false' framework cannot handle a special type of 'misspecification' under high-dimensional

setting— the dense coefficient in a correct model cannot be estimated consistently. Given the limitation of 'least false' parameter framework, we propose the measurement on the magnitudes of estimators and develop our DR methodology and theory based on the magnitudes. As we shall show, our magnitudes framework provides both a weaker requirement in theory and a broader applicability in practice.

In this section, we first describe a special case of our estimators from Section 3.2, which has a closed-form expression that is computationally efficient and stable. In Section 3.3.2 we introduce the cross-fitting scheme, which generally leads to relaxed sparsity conditions. Finally in Section 3.3.3 we develop the the doubly robust estimator.

3.3.1 A closed-form estimator

As discussed in Remark 7 earlier, we may have different choices for the estimator of the cumulative baseline hazard Λ. A particular choice is the weighted Breslow estimator:

$$\check{\Lambda}(t,\theta;\boldsymbol{\beta},\boldsymbol{\gamma}) = \int_0^t \frac{\sum_{i=1}^n w_i^1(\boldsymbol{\gamma})\{dN_i(u) - Y_i(u)(D_i\theta + \boldsymbol{\beta}^\top \mathbf{Z}_i)du\}}{\sum_{i=1}^n w_i^1(\boldsymbol{\gamma})Y_i(u)}, \tag{3.17}$$

where $w_i^1(\boldsymbol{\gamma}) = D_i\{1 - \text{expit}(\boldsymbol{\gamma}^\top \mathbf{Z}_{\mathbf{1}i})\} = D_i P(D_i = 0|Z_i)$. Note that the w_i^1's are the weights among the treated subjects. With (3.17) the score (3.6) becomes a linear function in θ:

$$\begin{aligned} & \phi\big(\theta;\boldsymbol{\beta},\check{\Lambda}(\cdot,\theta;\boldsymbol{\beta},\boldsymbol{\gamma}),\boldsymbol{\gamma}\big) \\ = \; & -\frac{1}{n}\sum_{i=1}^n (1-D_i)\,\text{expit}(\boldsymbol{\gamma}^\top \mathbf{Z}_{\mathbf{1}i}) \int_0^\tau \Big(dN_i(u) - Y_i(u)\Big[\boldsymbol{\beta}^\top\{\mathbf{Z}_i - \widetilde{\mathbf{Z}}(u;\boldsymbol{\gamma})\}du + d\widetilde{N}(u;\boldsymbol{\gamma})\Big]\Big) \\ & -\frac{\theta}{n}\sum_{i=1}^n (1-D_i)\,\text{expit}(\boldsymbol{\gamma}^\top \mathbf{Z}_{\mathbf{1}i})X_i, \end{aligned} \tag{3.18}$$

where we denote the weighted processes $\widetilde{\mathbf{Z}}(t;\boldsymbol{\gamma}) = \sum_{i=1}^{n} \mathbf{Z}_i w_i^1(\boldsymbol{\gamma}) Y_i(t) / \sum_{i=1}^{n} w_i^1(\boldsymbol{\gamma}) Y_i(t)$, and $d\widetilde{N}(t;\boldsymbol{\gamma}) = \sum_{i=1}^{n} w_i^1(\boldsymbol{\gamma}) dN_i(t) / \sum_{i=1}^{n} w_i^1(\boldsymbol{\gamma}) Y_i(t)$ with the same weights as in (3.17). Therefore we have

$$\breve{\theta} = \frac{\sum_{i=1}^{n} w_i^0(\widehat{\boldsymbol{\gamma}}) \int_0^{\tau} \left(dN_i(u) - Y_i(u) \left[\widehat{\boldsymbol{\beta}}^{\top} \{ \mathbf{Z}_i - \widetilde{\mathbf{Z}}(u;\widehat{\boldsymbol{\gamma}}) \} du + d\widetilde{N}(u;\widehat{\boldsymbol{\gamma}}) \right] \right)}{-\sum_{i=1}^{n} w_i^0(\widehat{\boldsymbol{\gamma}}) X_i}, \tag{3.19}$$

where $w_i^0(\boldsymbol{\gamma}) = (1 - D_i) \operatorname{expit}(\boldsymbol{\gamma}^{\top} \mathbf{Z}_{\mathbf{1}i}) = (1 - D_i) P(D_i = 1 | Z_i)$ are the weights among the subjects in the control group.

In Section 3.7.1, we show that $\breve{\theta}$ can be seen as through directly estimating the difference between the two cumulative hazard functions of the treated and the control groups.

Remark 11. *Since* $\breve{\Lambda}$ *defined in* (3.17) *depends linearly on* θ*, it is easy to see that* $\breve{\Lambda}$ *satisfies Assumption 5-viii. The average weight in the risk set at any time* t *is bounded away from zero under Assumptions 5-ii and 5-v because*

$$\mathbb{E}\{w^1(\boldsymbol{\gamma}_0) Y(t)\} \geq \mathbb{E}\{w^1(\boldsymbol{\gamma}_0) Y(\tau)\} = \mathbb{E}\{\operatorname{Var}(D|\mathbf{Z}) \mathbb{E}(Y(\tau)|\mathbf{Z}; D = 0)\} e^{-\theta_0}.$$

Following that, one can verify that $\breve{\Lambda}$ *meets all conditions given in Assumption 5. Hence, the inference result in Theorem 9 applies for the closed-form estimator* $\breve{\theta}$ *in* (3.19)*. Note that according to Theorem 9, the asymptotic variance of* $\widehat{\theta}$ *does not depend on the specific estimator of the cumulative baseline hazard, so there is no loss of asymptotic efficiency by only using the treated subjects in* $\breve{\Lambda}$*.*

Remark 12. *In the construction of* $\breve{\theta}$*, we weight treated subjects by* $w_i^1(\boldsymbol{\gamma})$ *and controls by* $w_i^0(\boldsymbol{\gamma})$*, defined after* (3.17) *and* (3.19)*, respectively. Consider the standardized version of the weights*

with true parameter

$$\bar{w}_i^1 = w_i^1(\gamma_0)/\sum_{j=1}^n w_j^1(\gamma_0),\ \bar{w}_i^0 = w_i^0(\gamma_0)/\sum_{j=1}^n w_i^0(\gamma_0). \tag{3.20}$$

We discover that the weights in (3.20) *balance the covariates between treatment and control. Following the popular R package* twang *[RMM+17, RM07], we manifest the covariate balance after weighting by the weighted empirical cumulative distribution functions,*

$$F_{d,n}(\mathbf{z}) = \sum_{i=1}^n \bar{w}_i^d I(\mathbf{Z}_i \le \mathbf{z}) = \frac{n^{-1}\sum_{i=1}^n w_i^d(\gamma_0) I(\mathbf{Z}_i \le \mathbf{z})}{n^{-1}\sum_{i=1}^n w_i^d(\gamma_0)},\ d = 0,1, \tag{3.21}$$

where we denote the multivariate indicator $I(\mathbf{Z} \le \mathbf{z}) = \prod_{j=1}^p I(\mathbf{Z}^j \le \mathbf{z}^j)$ *with* $\mathbf{Z}^j$ *and* $\mathbf{z}^j$ *being the* j*-th coordinate of vectors* $\mathbf{Z}$ *and* $\mathbf{z}$*, respectively. Under logistic regression model* (3.2)*, the expectations of the weights given covarites both equal the conditional variance of treatment,*

$$\mathbb{E}\{w^d(\gamma_0)|\mathbf{Z}\} = \text{expit}(\gamma_0^\top \mathbf{Z_1})\{1 - \text{expit}(\gamma_0^\top \mathbf{Z_1})\} = \text{Var}(D|\mathbf{Z}). \tag{3.22}$$

Using (3.22)*, we heuristically obtain the limit of* (3.21)

$$F_{d,n}(\mathbf{z}) \rightsquigarrow \frac{\mathbb{E}\{w^d(\gamma_0) I(\mathbf{Z} \le \mathbf{z})\}}{\mathbb{E}\{w^d(\gamma_0)\}} = \frac{\mathbb{E}\{\text{Var}(D|\mathbf{Z}) I(\mathbf{Z} \le \mathbf{z})\}}{\mathbb{E}\{\text{Var}(D|\mathbf{Z})\}},\ d = 0,1. \tag{3.23}$$

Notice the right-hand side in (3.23) *does not depend on* d*, so the distributions of covariates in both treatment arms are roughly the same after weighting. The covariate balancing plays an important role in the robustness when the outcome model is misspecified. Following the tradition of [RR83], classical causal inference methods including matching, stratification and weighting [RR83, RR84, RR85, Ros87, RB00, VD14] have been recently revisited and improved with high-dimensional covariates when the estimation to the propensity score becomes challenging [IR14, vdL14, VV15, Tan17].*

3.3.2 A cross-fitted orthogonal score

Cross-fitting, also known as data-splitting [Cox75], allows relaxed conditions on ..., and is increasing being used in high-dimensional problems [CCD$^+$18, ATW19]. Algorithm 1 demonstrates a cross-fitted version of our orthogonal score method.

Data: split the data into k folds of equal size with the indices set $I_1, I_2, \ldots, I_k$

for *each fold indexed by* j **do**

1. estimate the nuisance parameters $\left(\widehat{\boldsymbol{\beta}}^{(j)}, \widehat{\Lambda}^{(j)}, \widehat{\boldsymbol{\gamma}}^{(j)}\right)$ using the out-of-fold samples indexed by $I_{-j} = \{1, \ldots, n\} \setminus I_j$;

2. construct the cross-fitted score using the in-fold samples

$$\phi^{(j)}\left(\theta; \widehat{\boldsymbol{\beta}}^{(j)}, \widehat{\Lambda}^{(j)}, \widehat{\boldsymbol{\gamma}}^{(j)}\right) = \frac{1}{|I_j|} \sum_{i \in I_j} \left[D_i - \text{expit}(\widehat{\boldsymbol{\gamma}}^{(j)\top} \mathbf{Z}_{\mathbf{1}i})\right]$$
$$\times \int_0^\tau e^{D_i \theta t} \left[dN_i(t) - Y_i(t) \left\{\left(D_i \theta + \widehat{\boldsymbol{\beta}}^{(j)\top} \mathbf{Z}_i\right) dt + d\widehat{\Lambda}^{(j)}(t; \theta)\right\}\right]. \quad (3.24)$$

end

Result: Obtain the estimated treatment effect $\widehat{\theta}_{cf}$ by solving

$$\frac{1}{k} \sum_{j=1}^{k} \phi^{(j)}\left(\theta; \widehat{\boldsymbol{\beta}}^{(j)}, \widehat{\Lambda}^{(j)}, \widehat{\boldsymbol{\gamma}}^{(j)}\right) = 0. \quad (3.25)$$

Algorithm 1: Estimation of the Treatment Effect via k-fold Cross-fitting

The cross-fitting algorithm described in the box induces independence between the score and the estimated nuisance parameters, further reducing the effect of the nuisance parameters on the estimation of the treatment effect in addition to the orthogonality of the score function. For our purposes, the cross-fitting has two main advantages. First, it simplifies our methodology and

theory for the doubly robust estimation in the next subsection, as we use survival information to further regularize the initial LASSO estimator; measurability issues in the martingale argument would otherwise arise without the independence induced by cross-fitting. Second, we are able to handle cases with less sparsity (see also Remark 13 below). This is achieved because it allows the convergence in Assumption 5 to be relaxed from uniform error (l^1 distance in the coefficients) to average model deviance below (often proportional to l^2 distance in the coefficients). To describe the estimation error of the out-of-fold estimators evaluated on the in-fold samples, we denote $(X_*, \delta_*, D_*, \mathbf{Z}_*)$ as an independent copy from the same distribution as the original data, for which the expectation $\mathbb{E}_*$ is taken. We define the *average model deviance* for the estimated model coefficients in (3.1) and (3.2) as

$$\begin{aligned}\mathcal{D}_\beta(\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta}_0) &= \sqrt{\mathbb{E}_* \left[\int_0^\tau \left\{ (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^\top \mathbf{Z}_* \right\}^2 Y_*(t) dt \right]}, \\ \mathcal{D}_\gamma(\widehat{\boldsymbol{\gamma}}, \boldsymbol{\gamma}_0) &= \sqrt{\mathbb{E}_* \left[\left\{ \operatorname{expit}\left(\widehat{\boldsymbol{\gamma}}^\top \mathbf{Z}_*\right) - \operatorname{expit}\left(\boldsymbol{\gamma}_0^\top \mathbf{Z}_*\right) \right\}^2 \right]}. \end{aligned} \tag{3.26}$$

Note that $\mathcal{D}_\beta(\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta}_0)$ is the same as the norm used in [GG12b] equation (15). Compared to the uniform error, the average model deviance has a convergence rate that grows slower when sparsity increases.

Now we state our relaxed conditions for inference.

Assumption 6. *Suppose conditions 5-i to 5-viii in Assumption 5 are satisfied with* $(\widehat{\boldsymbol{\beta}}, \widehat{\Lambda}, \widehat{\boldsymbol{\gamma}}) = \left(\widehat{\boldsymbol{\beta}}^{(j)}, \widehat{\Lambda}^{(j)}, \widehat{\boldsymbol{\gamma}}^{(j)}\right)$ *for all* $j = 1, \ldots, k$. *Assume additionally,*

(i) the rates of estimation errors follow

$$\mathcal{D}_\beta\left(\widehat{\boldsymbol{\beta}}^{(j)}, \boldsymbol{\beta}_0\right) + \sup_{t \in [0,\tau]} \left| \widehat{\Lambda}^{(j)}(t; \theta_0) - \Lambda_0(t) \right| + \mathcal{D}_\gamma\left(\widehat{\boldsymbol{\gamma}}^{(j)}, \boldsymbol{\gamma}_0\right)$$

$$+\sqrt{n}\mathcal{D}_{\gamma}\left(\widehat{\gamma}^{(j)},\gamma_0\right)\left(\mathcal{D}_{\beta}\left(\widehat{\beta}^{(j)},\beta_0\right)+\sup_{t\in[0,\tau]}\left|\widehat{\Lambda}^{(j)}(t;\theta_0)-\Lambda_0(t)\right|\right)=o_p(1). \tag{3.27}$$

We have the inference result for θ with the cross-fitted estimator $\widehat{\theta}_{cf}$ defined in (3.25).

Theorem 10. *Under Assumption 6, $\widehat{\theta}_{cf}$ obtained from Algorithm 1 satisfies*

$$\widehat{\sigma}_{cf}^{-1}\sqrt{n}(\widehat{\theta}_{cf}-\theta_0)\rightsquigarrow N(0,1), \tag{3.28}$$

with the closed-form variance estimator

$$\widehat{\sigma}_{cf}^2=\frac{n^{-1}\sum_{j=1}^{k}\sum_{i\in I_j}\delta_i\{D_i-\text{expit}(\widehat{\gamma}^{(j)\top}\mathbf{Z}_{\mathbf{1}i})\}^2e^{2\widehat{\theta}D_iX_i}}{\left\{n^{-1}\sum_{j=1}^{k}\sum_{i\in I_j}(1-D_i)\,\text{expit}(\widehat{\gamma}^{(j)\top}\mathbf{Z}_{\mathbf{1}i})X_i\right\}^2}. \tag{3.29}$$

The proof of Theorem 10 follows the same strategy as that of Theorem 9.

Remark 13. *The cross-fitted score* (3.25) *can handle a larger number of covariates, less sparse models and more flexible estimators of the baseline hazard. We explain these three advantages by comparing Assumption 6-i to Assumption 5-ix. First,* (3.27) *allows a larger dimension without the extra* $\log(p)$ *factor in* (3.8)*. Second, the average model deviance is less sensitive to the growth in sparsity than the uniform error. For LASSO estimators* (3.12) *and* (3.13) *with oracle penalty parameters in particular, the rates in terms of average model deviance has been established as [GG12b, ZSZH17, vdG08]*

$$\mathcal{D}_{\beta}(\widehat{\beta},\beta_0)=O_p\left(\sqrt{s_{\beta}\log(p)/n}\right),\ \mathcal{D}_{\gamma}(\widehat{\gamma},\gamma_0)=O_p\left(\sqrt{s_{\gamma}\log(p)/n}\right). \tag{3.30}$$

If $s_{\beta}s_{\gamma}=o\left(n/\{\log(p)\}^2\right)$, then (3.27) *holds. When p is of polynomial order in n, the sparsity of both models may grow up to rate $n^{1/2}$. Or, the sparsity of one model is allowed to grow as rapid as n while the sparsity of the other model is constant. The tolerance to sparsity of our method is*

comparable with results established under linear models without censoring [BCH13]. Third, the removal of condition (3.9) *allows various estimation methods of the baseline hazards besides the Breslow type estimators, e.g. parametric models or splines.*

3.3.3 A doubly robust estimator

With the preparation in Sections 3.3.1 and 3.3.2, we present our methodology on the doubly robust estimation. Here we consider the partially linear additive hazards model [YLZ08]

$$\lambda(t, D, \mathbf{Z}) = D\theta + g(t; \mathbf{Z}), \tag{3.31}$$

and the general propensity model

$$\mathbb{P}(D = 1|\mathbf{Z}) = \pi(\mathbf{Z}). \tag{3.32}$$

Under model (3.31), θ retains the treatment effect interpretation.

We apply the cross-fitting algorithm described in Section 3.3.2 to the score (3.18) in Section 3.3.1. Suppose that $(\widehat{\boldsymbol{\beta}}^{(j)}, \widehat{\boldsymbol{\gamma}}^{(j)})$ is the LASSO estimator of $(\boldsymbol{\beta}, \boldsymbol{\gamma})$ using the out-of-fold samples for fold j. The cross-fitted version of (3.19) takes the following form,

$$\breve{\theta}_{cf} = \frac{\sum_{j=1}^{k} \sum_{i \in I_j} w_i^0(\widehat{\boldsymbol{\gamma}}^{(j)}) \int_0^\tau \left(dN_i(u) - Y_i(u) \left[\widehat{\boldsymbol{\beta}}^{(j)\top} \{\mathbf{Z}_i - \widetilde{\mathbf{Z}}^{(j)}(u; \widehat{\boldsymbol{\gamma}}^{(j)})\} du + d\widetilde{N}^{(j)}(u; \widehat{\boldsymbol{\gamma}}^{(j)}) \right] \right)}{-\sum_{j=1}^{k} \sum_{i \in I_j} w_i^0(\widehat{\boldsymbol{\gamma}}^{(j)}) X_i}, \tag{3.33}$$

with the weighted processes under cross-fitting $\widetilde{\mathbf{Z}}^{(j)}(t; \boldsymbol{\gamma}) = \sum_{i \in I_j} \mathbf{Z}_i w_i^1(\boldsymbol{\gamma}) Y_i(t) / \sum_{i \in I_j} w_i^1(\boldsymbol{\gamma}) Y_i(t)$, and $d\widetilde{N}^{(j)}(t; \boldsymbol{\gamma}) = \sum_{i \in I_j} w_i^1(\boldsymbol{\gamma}) dN_i(t) / \sum_{i \in I_j} w_i^1(\boldsymbol{\gamma}) Y_i(t)$. The weights w^0 and w^1 defined in Section 3.3.1 for the weighted Breslow $\breve{\Lambda}$ and closed form estimator $\breve{\theta}$.

Define the average model deviance as in (3.26) for the correctly specified model. Similar to the definition of the average model deviance, we denote $(X_*, \delta_*, D_*, \mathbf{Z}_*)$ as an independent

copy from the same distribution as the original data, for which the expectation $\mathbb{E}_*$ is taken. We define the magnitude of the estimation for the misspecified model within each cross-fitting fold,

$$\mathcal{M}_\beta\left(\widehat{\boldsymbol{\beta}}\right) = \sqrt{\int_0^\tau \widehat{\boldsymbol{\beta}}^\top \mathbb{E}_*\left[\{\mathbf{Z}_* - \mu(t)\}^{\otimes 2} Y_*(t)\right]\widehat{\boldsymbol{\beta}} dt},$$
$$\mathcal{M}_\gamma(\widehat{\boldsymbol{\gamma}}) = \left[\mathbb{E}_*\left\{w_*^0(\widehat{\boldsymbol{\gamma}})X_*\right\}\right]^{-1} + \left[\mathbb{E}_*\left\{w_*^1(\widehat{\boldsymbol{\gamma}})Y_*(\tau)\right\}\right]^{-1} \tag{3.34}$$

with $\mu(t) = \mathbb{E}_*(\mathbf{Z}_*)/\mathbb{E}_*\{Y_*(t)\}$. Here we define the magnitudes with $(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\gamma}})$ for the brevity in notation of the ensuing discussion. Eventually, we shall establish our doubly robust estimation under the magnitude condition with the cross-fitted estimator $(\widehat{\boldsymbol{\beta}}^{(j)}, \widehat{\boldsymbol{\gamma}}^{(j)})$ and the samples in fold-j as the stared random variables.

Throughout this section we will make an assumption that the magnitudes above are bounded. That in turn, can be simply guaranteed by a bounded l^1-norm $\|\widehat{\boldsymbol{\beta}}\|_1$ or $\|\widehat{\boldsymbol{\gamma}}\|_1$. To see that, observe

$$\mathcal{M}_\beta(\widehat{\boldsymbol{\beta}}) \le \|\widehat{\boldsymbol{\beta}}\|_1 \|\mathbf{Z}_*\|_\infty \sqrt{\tau}, \quad \mathcal{M}_\gamma(\widehat{\boldsymbol{\gamma}}) \le \frac{1 + e^{\|\widehat{\boldsymbol{\gamma}}\|_1 K_Z}}{\tau \mathbb{E}_*\{(1-D_*)Y_*(\tau)\}} + \frac{1 + e^{\|\widehat{\boldsymbol{\gamma}}\|_1 K_Z}}{\mathbb{E}_*\{D_* Y_*(\tau)\}}.$$

Hence, with a finite $\|\mathbf{Z}_*\|_\infty$ and strictly positive $\mathbb{E}_*\{D_* Y_*(\tau)\}$ and $\mathbb{E}_*\{(1-D_*)Y_*(\tau)\}$, r.h.s. above is guaranteed to be finite as long as both l^1-norms are finite. Moreover, we discover that using cross-validation to select the penalty factor in LASSO is sufficient to control the magnitudes. Let $\{\widehat{\boldsymbol{\beta}}(\lambda) : \lambda > 0\}$ and $\{\widehat{\boldsymbol{\gamma}}(\lambda) : \lambda > 0\}$ be classes of LASSO estimators with different penalty factors λ under additive hazards model and logistic regression model, respectively. The sets are often called the LASSO regularization path [FHT10]. In practice, the most common way of deciding the penalty factors λ is k-fold cross-validation. Suppose the optimal penalty factors are selected

by the risk minimization

$$\widehat{\lambda}_{\beta} = \underset{\lambda>0}{\operatorname{argmin}}\, l^*_{\beta}(\widehat{\boldsymbol{\beta}}(\lambda)),\ \widehat{\lambda}_{\gamma} = \underset{\lambda>0}{\operatorname{argmin}}\, l^*_{\gamma}(\widehat{\boldsymbol{\gamma}}(\lambda)), \tag{3.35}$$

where the generalization losses for $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are defined as

$$\begin{aligned} l^*_{\beta}(\boldsymbol{\beta}) &= \int_0^{\tau} \mathbb{E}_* \left(\left[\boldsymbol{\beta}^{\top} \{\mathbf{Z}_* - \mu(t)\} \right]^2 Y_*(t) \right) dt - 2 \int_0^{\tau} \mathbb{E}_* \left[\boldsymbol{\beta}^{\top} \{\mathbf{Z}_* - \mu(t)\}\, dN_*(t) \right], \\ l^*_{\gamma}(\boldsymbol{\gamma}) &= -\mathbb{E}_*(D_* \boldsymbol{\gamma}^{\top} \mathbf{Z}_*) + \mathbb{E}_* \left\{ \log\left(1 + e^{\boldsymbol{\gamma}^{\top} \mathbf{Z}_*}\right) \right\}. \end{aligned} \tag{3.36}$$

$\mathcal{M}_{\beta}$ describes how large the average predicted contribution of covariates in the hazard is, so $\mathcal{M}_{\beta}\left(\widehat{\boldsymbol{\beta}}(\widehat{\lambda}_{\beta})\right)$ at the best estimator in the LASSO regularization path should not be excessively larger than the true contribution of the covariates in the hazard. Using the connection between $\mathcal{M}_{\beta}\left(\widehat{\boldsymbol{\beta}}\right)$ and the quadratic term in the $l^*_{\beta}(\boldsymbol{\beta})$ above, we establish a bound for the magnitude of the additive hazards LASSO estimator with optimal cross-validated penalty.

Lemma 27. *Under the partially linear additive hazards model* (3.31) *, we have*

$$\mathcal{M}_{\beta}\left(\widehat{\boldsymbol{\beta}}(\widehat{\lambda}_{\beta})\right)^2 \leq 4 \int_0^{\tau} \mathbb{E}_*\{g(t, \mathbf{Z}_*)^2 Y_*(t)\} dt. \tag{3.37}$$

We prove Lemma 27 by comparing the $l^*_{\beta}(\widehat{\boldsymbol{\beta}}(\widehat{\lambda}_{\beta}))$ to $l^*_{\beta}(0) = 0$. Since zero is always in the LASSO regularization path for a sufficiently large λ [GRS12], we must have the upper bound $l^*_{\beta}(\widehat{\boldsymbol{\beta}}(\widehat{\lambda}_{\beta})) \leq 0$. We use the Cauchy-Schwartz inequality to obtain a lower bound of $l^*_{\beta}(\widehat{\boldsymbol{\beta}}(\widehat{\lambda}_{\beta}))$. We reach (3.37) by the fact that the lower bound is always less than or equal to the upper bound for $l^*_{\beta}(\widehat{\boldsymbol{\beta}}(\widehat{\lambda}_{\beta}))$. $\mathcal{M}_{\gamma}$ describes how close the estimated propensity scores are to the actual treatment assignments in both treatment arms of the risk-set at $t = \tau$, so $\mathcal{M}_{\gamma}(\widehat{\boldsymbol{\gamma}}(\widehat{\lambda}_{\gamma}))$ should be bounded when the treatment has enough randomness in the risk-set at $t = \tau$, as required by the Assumption

5-v. We derive an equivalent characterization of the Assumption 5-v that there exist a set $\mathcal{Z}$ on which $\mathbb{P}(\mathbf{Z}_* \in \mathcal{Z})$, $\mathbb{E}_*(D_*|\mathcal{Z})$, $\mathbb{E}_*(1-D_*|\mathcal{Z})$ and $\mathbb{E}_*(Y_*(\tau)|\mathcal{Z})$ are all bounded away from zero. Focusing on the analysis of the set $\mathcal{Z}$, we establish a bound for the magnitude of the logistic regression LASSO estimator with optimal cross-validated penalty.

Lemma 28. *Under Assumption5-v, we have with probability tending to one*

$$\mathcal{M}_\gamma\left(\widehat{\gamma}(\widehat{\lambda}_\gamma)\right) \leq (1+\tau^{-1})8\varepsilon_Y^{-3}e^{-4\log(\varepsilon_Y)/\varepsilon_Y^2}. \tag{3.38}$$

We prove Lemma 28 by comparing the $l_\gamma^*(\widehat{\gamma}(\widehat{\lambda}_\gamma))$ to $l_\gamma^*(\widehat{\gamma}_0)$, where $\widehat{\gamma}_0$ is the intercept only estimator $(\log(1-n/\sum_{i=1}^n(D_i)),0,\ldots,0)$. The intercept only estimator $\widehat{\gamma}_0$ is also always in the LASSO regularization path for a sufficiently large λ [FHT10], so $l_\gamma^*(\widehat{\gamma}_0)$ is an upper bound for $l_\gamma^*(\widehat{\gamma}(\widehat{\lambda}_\gamma))$. Assumption 5-v implies the existence of a set $\mathcal{Z}$ in the covariate space with positive probability, at-risk rate at $t=\tau$ and true propensities bounded away from zero and one. Using the Jensen's inequality on the expectation taken over set $\mathcal{Z}$, we establish a lower bound for $l_\gamma^*(\widehat{\gamma}(\widehat{\lambda}_\gamma))$. The lower bound is closely connected with the magnitude $\mathcal{M}_\gamma\left(\widehat{\gamma}(\widehat{\lambda}_\gamma)\right)$. We obtain (3.38) by the fact that lower bound is less than or equal to the upper bound for $l_\gamma^*(\widehat{\gamma}(\widehat{\lambda}_\gamma))$.

Remark 14. *Lemmas 27 and 28 give surprisingly nice guarantees on the LASSO estimators with cross-validated penalty when the model assumption is wrong. Our results here has opened up a new direction of studying the properties of penalization methods under model misspecification. Unlike the common "least false parameter" argument, our bounds on the magnitudes require no quasi-model assumptions like sparsity. Consequently, our doubly robust estimation developed under the magnitude conditions is extremely sharp in theory and broadly valid in application.*

Remark 15. *When the model assumption is correct under suitable regularity conditions, the rate of consistency for LASSO with cross-validated penalty factor is always controlled by that for LASSO with oracle penalty factor. Quite surprisingly, this connection between the oracle penalty factor and the cross-validation is seldom manifested, though such fact is fundamental for any property established under oracle penalty factor to carry practical significance. Since the LASSO with oracle penalty factor is one element in the regularization path, any bound on its estimation error is also a bound for the estimation error of the "best" element along the regularization path. When the generalization loss possesses local convexity at the true coefficient, the cross-validated penalty attains such "best" element. Therefore, the LASSO with cross-validated penalty factor converges to the true coefficient in at least the same order of the LASSO with oracle penalty factor. Extension from the generalization loss to the sample cross-validated loss usually requires the same set of regularity conditions under which the oracle inequalities are established [GG12b, ZSZH17, vdG08].*

Our regularity conditions for the doubly robust estimation are stated with the estimation magnitudes below.

Assumption 7.

(a)- Suppose conditions 5-ii to 5-iv in Assumption 5 are satisfied.

(i) $\{W_i\}_{i=1}^n$ *is generated according to model* (3.1) *and* (3.31)*;*

(ii) $\sup_{j=1,\ldots,k} \mathcal{D}_\beta\left(\widehat{\beta}^{(j)}, \beta_0\right) = o_p(1)$ *and* $\sup_{j=1,\ldots,k} \mathcal{M}_\gamma\left(\widehat{\gamma}^{(j)}\right) \leq K_{\mathcal{M}_g} + o_p(1)$.

Or:

(b)- Suppose conditions 5-ii, 5-iii and 5-v in Assumption 5 are satisfied.

(i) $\{W_i\}_{i=1}^n$ *is generated according to model* (3.31) *and* (3.2)*;*

(ii) the hazard $g(t;\mathbf{Z})$ *satisfies* $\sqrt{\mathbb{E}\left\{\int_0^\tau g^2(t;\mathbf{Z}_i)Y_i(t)dt\right\}} = K_\Lambda = o(\sqrt{n})$*;*

(iii) the rate condition

$$\left\{K_\Lambda + \sup_{j=1,\ldots,k} \mathcal{M}_\beta\left(\widehat{\boldsymbol{\beta}}^{(j)}\right)\right\} \sup_{j=1,\ldots,k} \mathcal{D}_\gamma\left(\widehat{\boldsymbol{\gamma}}^{(j)}, \boldsymbol{\gamma}_0\right) = o_p(1). \tag{3.39}$$

Under either Assumption 7(a) or Assumption 7(b), the average model deviance for the correct model converges to zero. That is, $\mathcal{D}_\beta\left(\widehat{\boldsymbol{\beta}}^{(j)}, \beta_0\right) = o_p(1)$ under Assumption 7(a) and $\mathcal{D}_\gamma\left(\widehat{\boldsymbol{\gamma}}^{(j)}, \boldsymbol{\gamma}_0\right) = o_p(1)$ under Assumption 7(b). When the logistic regression model is wrong, we assume a bounded $\mathcal{M}_\gamma\left(\widehat{\boldsymbol{\gamma}}^{(j)}\right)$ so that the denominators in $\breve{\theta}_{cf}$ (3.33) are bounded away from zero. When the additive hazards model is wrong, we allow $\mathcal{M}_\beta\left(\widehat{\boldsymbol{\beta}}^{(j)}\right)$ and $\sqrt{\mathbb{E}\left\{\int_0^\tau g^2(t;\mathbf{Z}_i)Y_i(t)dt\right\}}$, the measure of the true average contribution of covariates in hazard in Lemma 27, to grow with sample size.

Theorem 11. *When either of Assumption 7(a) or Assumption 7(b) holds,* $\breve{\theta}_{cf}$*, defined in* (3.33)*, is consistent for* θ_0*, i.e.* $|\breve{\theta} - \theta_0| = o_p(1)$.

Our proof relies on the double robustness of our score at population level. When the additive hazards model (3.1) is correct, the true θ_0 solves the equation $\mathbb{E}\left\{\phi^{(j)}\left(\theta_0; \boldsymbol{\beta}_0, \Lambda_0, \widehat{\boldsymbol{\gamma}}^{(j)}\right)\right\} = 0$ for any $\widehat{\boldsymbol{\gamma}}^{(j)}$. When the logistic regression model (3.2) is correct, the true θ_0 solves the equation $\mathbb{E}\left\{\phi^{(j)}\left(\theta_0; \widehat{\beta}^{(j)}, \widehat{\Lambda}^{(j)}, \boldsymbol{\gamma}_0\right)\right\} = 0$ with any $\widehat{\beta}^{(j)}$ and $\widehat{\Lambda}^{(j)}$.

Theorem 11 implies an important corollary on the doubly robust estimation when the violation of the sparsity assumption on one model renders the consistent estimation of that model.

We state the immediate consequence of Lemmas 27 and 28 and Theorem 11 in the following corollary.

Corollary 1. *Suppose we use LASSO* (3.12) *and* (3.13) *to estimate* $(\widehat{\boldsymbol{\beta}}^{(j)}, \widehat{\boldsymbol{\gamma}}^{(j)})$ *in k-fold cross-fitting. The penalty factors are selected by generalized cross-validation. Under the conditions 5-i to 5-v in Assumption 5 and additionally:*

(i) $\int_0^\tau \mathbb{E}_*\{(\boldsymbol{\beta}_0^\top \mathbf{Z}_*)^2 Y_*(t)\}dt < K_\Lambda$,

(ii) the dimensions satisfy $\sqrt{(s_\beta \wedge s_\gamma)\log(p)/n} = o(1)$,

(iii) other regularity conditions from [GG12b] and [vdG08],

$\check{\theta}_{cf}$*, defined in* (3.33)*, is consistent for* θ_0*, i.e.* $|\check{\theta} - \theta_0| = o_p(1)$.

Corollary 1 demonstrates our unique contribution to the doubly robust estimation in high-dimensions in developing the magnitude condition. When the sparsity s exceeds sample size n, various concentration results on the estimators by penalization methods no longer hold. Regardless, we are able to show that our orthogonal score method produces consistent estimation with very common LASSO-cross-validation estimation procedure on the high-dimensional nuisance parameters when the sparsity of either model is small.

The other forms of our orthogonal score method, $\widehat{\theta}$ in Section 3.2 with Breslow estimator, the closed form $\check{\theta}$ in Section 3.3.1 and $\widehat{\theta}_{cf}$ in Section 3.3.2 with Breslow estimator, may also produce doubly robust estimation, as suggested by our simulation study later. The estimators constructed with Breslow estimator require obscure condition for the derivative of score with respect to θ being bounded away from zero. In practice, such requirement may trigger numerical

instability when the extra condition is violated. The estimators without cross-fitting generally require stronger conditions in dimensions or sparsities, similar to the inference case discussed in Section 3.3.2. Here we choose not to expand on those results to free the readers from unnecessary technicality for the suboptimal methods.

The orthogonality of our proposed score, as shown in Lemmas 26, no longer holds when one of the models is misspecified. As a result, the estimation error in causal parameter is dominated by the bias from the nuisance parameter estimator. With increasing dimensionality p, the bias from LASSO grows faster than $\sqrt{n}$-order. Therefore, the inference problem on θ through $\check{\theta}_{cf}$ under the double robustness setting is fundamentally different from the work with low-dimensional covariates [ZS12, ZZYK15, KLZ18, WLL$^+$17, JLS$^+$17]. The task requires a quite different approach, which goes beyond the scope of the current paper. Nevertheless, our consistency result provides a solid initial estimate for future pursuit of inference method.

3.4 Simulation

We assess the performance of the proposed estimators in the following simulation. In the simulation, we consider three pairs of dimensions $n = p = 300$ and $n = p = 1500$. The covariates $Z_1, \ldots, Z_p$ are independently generated from $N(0,1)$. The censoring time C is generated as the smaller between τ and $Unif(0, c_0)$. For each setup below, the parameters τ and c_0 are chosen such that $n/10$ treated subjects are expected to be at risk at $t = \tau$, and the censoring rate is around 30%. For each scenario with each sample size, we repeat the simulation 500 times.

To test the inference method, we generate the event time T from the additive hazards

model,

$$\lambda(t|D,\mathbf{Z}) = -0.25D + \boldsymbol{\beta}^\top \mathbf{Z} \tag{3.40}$$

and the treatment assignment D from the logistic regression propensity model

$$\mathbb{P}(D=1|\mathbf{Z}) = \text{logit}\left(\gamma_{\text{I}} + \boldsymbol{\gamma}^\top \mathbf{Z}\right). \tag{3.41}$$

Under model (3.40), the true treatment effect θ_0 is -0.25, and the baseline hazard λ_0 is set as 0. To ensure that the baseline hazard is non-negative, only samples satisfy $\beta^\top Z_i \geq 0.25$ are accepted in the data. The coefficients $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ contain two types of signals, the strong signal of size 1 and the weak signal of sizes 0.1 and 0.05 for $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, respectively. We consider the following 3 sparsity levels in :

very sparse ($s_\beta = 2$, $s_\gamma = 1$): $\boldsymbol{\beta} = (1, 0.1, \underbrace{0, \ldots, 0}_{p-2}), \boldsymbol{\gamma} = (1, \underbrace{0, \ldots, 0}_{p-1})$;

sparse ($s_\beta = 6$, $s_\gamma = 3$): $\boldsymbol{\beta} = (1, \underbrace{0.1, \ldots, 0.1}_{5}, \underbrace{0, \ldots, 0}_{p-6}), \boldsymbol{\gamma} = (1, 0.05, 0.05, \underbrace{0, \ldots, 0}_{p-3})$;

moderately sparse ($s_\beta = 15$, $s_\gamma = 10$): $\boldsymbol{\beta} = (1, \underbrace{0.1, \ldots, 0.1}_{13}, \underbrace{0, \ldots, 0}_{p-15})$,

$$\boldsymbol{\gamma} = (1, 1, \underbrace{0.05, \ldots, 0.05}_{8}, \underbrace{0, \ldots, 0}_{p-10}).$$

All the intercepts γ_I's in the propensity models are chosen such that $\mathbb{P}(D=1) = 0.5$, i.e. the treatment rate is around 0.5 marginally. We set the sparsities of the logistic regression model to be smaller than those of the additive hazards model because the LASSO of the former is empirically more sensitive to increase in sparsity. Four pairs of sparsities, $(s_\beta = 2, s_\gamma = 1)$, $(s_\beta = 2, s_\gamma = 10)$, $(s_\beta = 15, s_\gamma = 1)$ and $(s_\beta = 6, s_\gamma = 3)$, are studied in the simulation.

To test the doubly robust estimation method, we consider several setups under which the estimation to model is no longer consistent. First, we simulate from the models (3.40) and (3.41) with dense coefficients:

$$\text{Dense } (s_\beta = 30,\, s_\gamma = 20)\text{: } \boldsymbol{\beta} = (\underbrace{1,\ldots,1}_{4},\underbrace{0.1,\ldots,0.1}_{26},\underbrace{0,\ldots,0}_{p-30}),$$

$$\boldsymbol{\gamma} = (\underbrace{1,\ldots,1}_{4},\underbrace{0.05,\ldots,0.05}_{16},\underbrace{0,\ldots,0}_{p-20}).$$

Two pairs of very sparse - dense combinations, $(s_\beta = 2, s_\gamma = 20)$ and $(s_\beta = 30, s_\gamma = 1)$, are studied. Second, we consider the misspecified model for event time with exponential link

$$\lambda(t|D,\mathbf{Z}) = -0.25D + \exp\left(\boldsymbol{\beta}^\top \mathbf{Z}\right) + 0.25 \tag{3.42}$$

when the logistic regression propensity model is correct. The coefficients are set as $(s_\beta = 2, s_\gamma = 1)$. Third, we consider the misspecified propensity model with probit link

$$\mathbb{P}(D = 1|\mathbf{Z}) = \text{probit}\left(\gamma_I + \boldsymbol{\gamma}^\top \mathbf{Z}\right) \tag{3.43}$$

when the addive hazards model for event time is correct. The coefficients are alse set as $(s_\beta = 2, s_\gamma = 1)$. Finally, we consider another misspecified propensity model with deterministic treatment assignment

$$D|\mathbf{Z} = I\left(\boldsymbol{\gamma}^\top \mathbf{Z} > \mu\right) \tag{3.44}$$

with μ being the median of $\boldsymbol{\gamma}^\top \mathbf{Z}$. The additive hazards model for event time is correct, and the coefficients are also set as $(s_\beta = 2, s_\gamma = 1)$.

We estimate the coeffcient of the logistic regression with covariates $\mathbf{Z}$ by LASSO $\widehat{\gamma}$ (3.13) by the R-package *glmnet* and the coefficient of the additive hazards model with covariates $(D, \mathbf{Z})$

by LASSO $(\widehat{\theta}_{lasso}, \widehat{\boldsymbol{\beta}})$ (3.12) by the R-package *ahaz*. The penalty parameters are selected by 10-fold cross-validation. We estimate the baseline cumulative hazard by the Breslow estimator (3.14) $\widehat{\Lambda}(t; \widehat{\theta}_{lasso}, \widehat{\boldsymbol{\beta}})$. We set the number of folds in cross-fitting as 10, and estimate the coefficients for each fold by LASSO with 9-fold cross-validation. In Table 3.1, we present the estimation error of the nuisance parameters from LASSO with respect to the true parameters or the least false parameters. The uniform error panel contains the estimation errors from LASSO in l_1-norm and the Breslow estimator in l_∞-norm. The deviance panel contains the mean estimation errors from cross-fitted models in terms of the deviance (3.26). Here we use sample average $n^{-1}\sum_{j=1}^{k}\sum_{i\in I_k}$ to approximate expectation $\mathbb{E}_*$,

$$\begin{aligned}
\widehat{\mathcal{D}_\beta} &= \sqrt{n^{-1}\sum_{j=1}^{k}\sum_{i\in I_k}\left[\int_0^\tau \left\{(\widehat{\boldsymbol{\beta}}^{(j)} - \boldsymbol{\beta}_0)^\top \mathbf{Z}_i\right\}^2 Y_i(t)dt\right]}, \\
\widehat{\mathcal{D}_\gamma} &= \sqrt{n^{-1}\sum_{j=1}^{k}\sum_{i\in I_k}\left[\left\{\text{expit}\left(\widehat{\boldsymbol{\gamma}}^{(j)\top}\mathbf{Z}_i\right) - \text{expit}\left(\boldsymbol{\gamma}_0^\top \mathbf{Z}_i\right)\right\}^2\right]}.
\end{aligned} \tag{3.45}$$

The magnitude panel contains the median estimated magnitudes from cross-fitted models as defined in (3.34). Here we use sample average $k/n\sum_{i\in I_k}$ to approximate expectation $\mathbb{E}_*$, and take maximum across all folds,

$$\begin{aligned}
\widehat{\mathcal{M}_\beta} &= \max_{j=1,\ldots,k}\sqrt{\int_0^\tau \widehat{\boldsymbol{\beta}}^{(j)\top}\frac{k}{n}\sum_{i\in I_j}\left[\left\{\mathbf{Z}_i - \bar{\mathbf{Z}}^{(j)}(t)\right\}^{\otimes 2} Y_i(t)\right]\widehat{\boldsymbol{\beta}}^{(j)}dt}, \\
\widehat{\mathcal{M}_\gamma} &= \max_{j=1,\ldots,k}\left\{\frac{n/k}{\sum_{i\in I_j} w_i^0\left(\widehat{\boldsymbol{\beta}}^{(j)}\right)X_i} + \frac{n/k}{\sum_{i\in I_j} w_i^1\left(\widehat{\boldsymbol{\beta}}^{(j)}\right)Y_i(\tau)}\right\}.
\end{aligned} \tag{3.46}$$

When the Assumption 5-v holds, for all setups except the deterministic treatment assignment (3.44), the magnitudes are controlled quite well empirically by cross-validation, there is no evidence suggesting that magnitudes may blow up with larger dimensions.

Table 3.1: Estimation error of the nuisance parameters. Letters “E", “P" and “D" represent the misspecified models with exponential link, probit link and deterministic treatment assignment. Under dense case ($s_\beta = 30, s_\gamma = 20$), the LASSO estimators do not concentrate around the true coeffcients. The least false parameter is infinite under deterministic treatment assignment. The magnitude is properly controlled with cross-validation.

n & p	Sparsity/ Misspec.	Uniform Errors of LASSO and Breslow			Deviance from Cross-fitting		Magnitude from Cross-fitting	
	for β, γ	$\widehat{\beta}$ in l_1	$\widehat{\gamma}$ in l_1	$\widehat{\Lambda}$ in l_∞	$\widehat{\mathcal{D}_\beta}$	$\widehat{\mathcal{D}_\gamma}$	$\widehat{\mathcal{M}_\beta}$	$\widehat{\mathcal{M}_\gamma}$
300	2, 1	0.61	1.38	0.27	0.34	0.09	0.37	36.73
1500	2, 1	0.42	0.80	0.13	0.20	0.05	0.41	29.78
300	6, 3	1.13	1.75	0.31	0.46	0.10	0.35	29.84
1500	6, 3	0.90	1.13	0.17	0.27	0.05	0.40	23.49
300	15, 10	2.87	2.97	0.46	0.67	0.12	0.38	35.11
1500	15, 10	2.43	2.02	0.28	0.45	0.07	0.44	27.16
300	30, 20	6.22	5.01	0.60	0.96	0.15	0.42	48.72
1500	30, 20	5.19	3.75	0.36	0.63	0.09	0.48	34.97
300	E, P	–	–	–	0.71	0.09	0.58	29.19
1500	E, P	–	–	–	0.47	0.05	0.65	22.91
300	–, D	–	> 100	–	–	0.17	–	$> 100^{**}$
1500	–, D	–	> 100	–	–	0.13	–	$> 100^{**}$

* The dashed entries are not well-defined due to misspecification;

** The divergence of magnitude is expected because setup “D" violates Assumption 5-v.

We present the results of 4 proposed estimators for inference and doubly robust estimation: $\widehat{\theta}$ obtained from (3.6) with the Breslow estimator (3.14), the closed form estimator $\check{\theta}$ (3.19) and their cross-fitted counterparts $\widehat{\theta}_{cf}$ and $\check{\theta}_{cf}$ as described in Algorithm 1 and (3.33). As the benchmark, we also present the result of the estimation of θ from LASSO for the additive hazards model with covariates $(D, \mathbf{Z})$ which set the penalty for θ to be zero. Since $\widetilde{\theta}$ is not penalized, its estimation bias is entirely caused by the partially adjusted confounding due to estimation errors in the nuisance parameter.

We present the inference result in Table 3.2. The benchmark $\widetilde{\theta}_{\beta}$ has large bias/sd ratio, which confirms the difficulty of drawing inference in our design. Our devised orthogonality has corrected the bias in all four variations of our proposed method. The biases are at least reduced by half from LASSO, and the reduction rate grows rapidly with sample size. All of our four variations achieve the coverage rates of the 95 % confidence intervals very closed to the nominal level at the large sample size $n = p = 1500$. At the smaller sample size $n = p = 300$, $\check{\theta}_{cf}$ the closed from estimator with crossfitting also has a reasonably close coverage rate to the nominal level. has the best coverage. With a closer look especially at the smaller dimensions $n = p = 300$, we find the comparative advantage of the closed form estimator and the crossfitting scheme. $\widehat{\theta}_{cf}$ and $\check{\theta}_{cf}$ demonstrate the benefit of cross-fitting with even smaller Bias. The closed form $\check{\theta}_{cf}$ with enhanced numerical stability has smaller sd than $\widehat{\theta}_{cf}$ with Breslow, which leads to the improved CI coverage.

We present the estimation result with inconsistent estimation to one model in Table 3.3. In the dense scenarios, we observe from Table 3.1 that the LASSO estimators deviate substantially from the underlying true coefficients. When the treatment assignment is deterministic, the least

Table 3.2: Inference result for simulation under moderately sparse Scenarios. True ATE = -0.25. Censoring rate 30%.

Sparsity		Benchmark		ATE Inference Result							
		LASSO $\widetilde{\theta}_\beta$		$\widehat{\theta}$ with Breslow				Closed form $\check{\theta}$			
s_β	s_γ	Bias	sd	Bias	sd	se	Cover	Bias	sd	se	Cover
n=p=300											
2	1	0.054	0.097	0.029	0.097	0.091	92.4 %	0.032	0.094	0.091	93.0 %
6	3	0.071	0.094	0.050	0.095	0.093	92.0 %	0.052	0.092	0.093	93.0 %
15	1	0.088	0.135	0.051	0.123	0.128	93.6 %	0.049	0.122	0.128	94.0 %
2	10	0.099	0.094	0.050	0.099	0.094	89.6 %	0.052	0.096	0.094	89.8 %
n=p=1500											
2	1	0.031	0.040	0.009	0.041	0.041	94.2 %	0.011	0.041	0.041	94.0 %
6	3	0.033	0.042	0.015	0.043	0.042	93.0 %	0.017	0.042	0.042	93.6 %
15	1	0.047	0.063	0.019	0.064	0.058	91.4 %	0.020	0.063	0.058	91.2 %
2	10	0.077	0.041	0.019	0.043	0.043	91.6 %	0.022	0.043	0.043	91.6 %

Sparsity			ATE Inference Result with Cross-fitting							
			$\widehat{\theta}_{cf}$ with Breslow				Closed form $\check{\theta}_{cf}$			
s_β	s_γ		Bias	sd	se	Cover	Bias	sd	se	Cover
n=p=300										
2	1		0.012	0.100	0.090	91.0 %	0.011	0.093	0.090	93.4 %
6	3		0.027	0.100	0.092	92.4 %	0.028	0.089	0.092	94.6 %
15	1		0.018	0.134	0.127	93.8 %	0.013	0.123	0.127	95.8 %
2	10		0.032	0.106	0.094	89.4 %	0.032	0.097	0.094	93.2 %
n=p=1500										
2	1		0.006	0.042	0.041	94.8 %	0.009	0.040	0.041	95.4 %
6	3		0.010	0.044	0.041	92.8 %	0.014	0.041	0.042	94.2 %
15	1		0.006	0.064	0.058	92.4 %	0.012	0.061	0.058	93.0 %
2	10		0.017	0.044	0.043	92.4 %	0.019	0.042	0.043	92.8 %

Table 3.3: Doubly robust estimation with inconsistent nuisance estimator. True ATE = -0.25. Censoring rate 30%.

Sparsity		Benchmark			ATE Estimation Result					
		LASSO $\widetilde{\theta}_\beta$			$\widehat{\theta}$ with Breslow			Closed form $\breve{\theta}$		
s_β	s_γ	Bias	sd	$\sqrt{MSE}$	Bias	sd	$\sqrt{MSE}$	Bias	sd	$\sqrt{MSE}$
n=p=300										
30	1	0.141	0.202	0.247	0.080	0.169	0.187	0.074	0.169	0.185
2	20	0.078	0.090	0.119	0.051	0.099	0.111	0.052	0.096	0.109
E	1	0.233	0.397	0.461	0.117	0.375	0.393	0.106	0.366	0.381
2	P	0.095	0.102	0.139	0.041	0.101	0.109	0.043	0.097	0.106
2	D	-0.598	0.204	0.632	-0.103	0.217	0.240	-0.124	0.223	0.255
n=p=1500										
30	1	0.057	0.083	0.101	0.018	0.082	0.084	0.018	0.081	0.083
2	20	0.061	0.042	0.074	0.020	0.048	0.052	0.022	0.047	0.052
E	1	0.132	0.169	0.214	0.049	0.167	0.174	0.049	0.163	0.171
2	P	0.050	0.043	0.066	0.012	0.045	0.046	0.014	0.044	0.046
2	D	-0.308	0.092	0.321	-0.032	0.104	0.109	-0.040	0.107	0.114

Sparsity		ATE Estimation Result with Cross-fitting					
		$\widehat{\theta}_{cf}$ with Breslow			Closed form $\breve{\theta}_{cf}$		
s_β	s_γ	Bias	sd	$\sqrt{MSE}$	Bias	sd	$\sqrt{MSE}$
n=p=300							
30	1	0.049	0.197	0.203	0.026	0.177	0.179
2	20	0.036	0.106	0.112	0.034	0.099	0.104
E	1	0.054	0.396	0.400	0.001	0.364	0.364
2	P	0.021	0.105	0.107	0.019	0.097	0.099
2	D	-0.216 *	0.506 *	0.550*	-0.133	0.258	0.290
n=p=1500							
30	1	0.000	0.085	0.085	0.004	0.080	0.080
2	20	0.018	0.049	0.052	0.020	0.047	0.051
E	1	0.027	0.169	0.171	0.024	0.163	0.165
2	P	0.008	0.045	0.046	0.011	0.043	0.045
2	D	-0.031	0.107	0.111	-0.040	0.116	0.123

* One divergent iteration is removed from the summary.

false parameter does not exist. Under these situations, our magnitude condition may still hold or be enforced while the classical condition of concentration around the least false parameter is no longer valid. All four variations of our orthogonal score approach have reasonable estimation error decaying with larger sample size, showing evidence of consistency. Our estimators have smaller bias than LASSO $\widetilde{\theta}_{\beta}$. Most notably under the determinist treatment assignment scenario when the benchmark LASSO fails completely, our proposed methods still have quite accurate estimation. Similar pattern among our variations as in Table 3.2 is observed. $\widehat{\theta}_{cf}$ and $\check{\theta}_{cf}$ demonstrate the benefit of cross-fitting with even smaller Bias. Under most scenarios, the closed form $\check{\theta}_{cf}$ with enhanced numerical stability has smaller sd than $\widehat{\theta}_{cf}$ with Breslow at $n = p = 300$, which leads to the improved MSE. Again in the scenario "D" at $n = p = 300$, $\check{\theta}_{cf}$ demonstrates the clearest evidence of the enhanced numerical stability when it avoids the divergence experienced by $\widehat{\theta}_{cf}$ for exactly the same task.

3.5 Data Analysis

Typically clinical databases like the United States National Cancer Institute's Surveillance, Epidemiology, and End Results (SEER) contain disease specific variables, but only limited information on the subjects' health status such as comorbidities otherwise. In studying causal treatment effects, this leads to unobserved confounding [HYB$^+$10, YXM19]. On the other hand, the availability of information from claims databases could make up for some of these 'unobserved' confounders, as they have been shown to contain much information about these comorbidities [HPH$^+$18b, RTH$^+$19].

Table 3.4: Description of the SEER-Medicare Linked Data.

		12114 Conservative	5623 Surgery
Feature	Label	Count (%) or Mean (SD)	Count (%) or Mean (SD)
Age	66-69	4523 (37.3 %)	2443 (43.4 %)
	70-74	7591 (62.7 %)	3180 (56.6 %)
Marital status	Married	8743 (72.2 %)	4181 (74.4 %)
	Divorced	726 (6.0 %)	315 (5.6 %)
	Single	985 (8.1 %)	408 (7.3 %)
	Other	1660 (13.7 %)	719 (12.8 %)
Race	White	9855 (81.4 %)	4704 (83.7 %)
	Black	1560 (12.9 %)	543 (9.7 %)
	Asian	201 (1.7 %)	115 (2.0 %)
	Hispanic	143 (1.2 %)	86 (1.5 %)
	Other	355 (2.9 %)	175 (3.1 %)
Tumor stage	T1	8293 (68.5 %)	2426 (43.1 %)
	T2	3821 (31.5 %)	3197 (56.9 %)
Tumor grade	Well differentiated	6238 (51.5 %)	2963 (52.7 %)
	Moderately differentiated	71 (0.6 %)	58 (1.0 %)
	Poorly differentiated	5786 (47.8 %)	2594 (46.1 %)
	Undifferentiated	19 (0.2 %)	8 (0.1 %)
Prior Charlson	0	7932 (65.5 %)	3875 (68.9 %)
comorbidity score	≤ 1	2716 (22.4 %)	1174 (20.9 %)
	≥ 2	1466 (12.1 %)	574 (10.2 %)
Prostate-Specific-	< 10	9292 (76.7 %)	4519 (80.4 %)
-Antigen	≥ 10	2822 (23.3 %)	1104 (19.6 %)
Gleason score	< 7	6186 (51.1 %)	2979 (53.0 %)
	≥ 7	5928 (48.9 %)	2644 (47.0 %)
Year	2004	1783 (14.7 %)	953 (16.9 %)
	2005	1715 (14.2 %)	928 (16.5 %)
	2006	2153 (17.8 %)	1006 (17.9 %)
	2007	2393 (19.8 %)	1041 (18.5 %)
	2008	2260 (18.7 %)	954 (17.0 %)
	2009	1810 (14.9 %)	741 (13.2 %)
Claims codes	Ave count (SD)	57.7 (31.6)	60.9 (31.9)

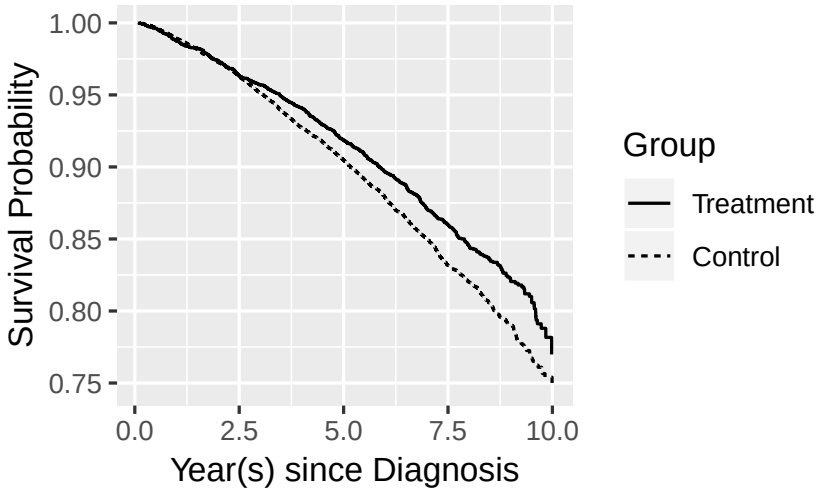
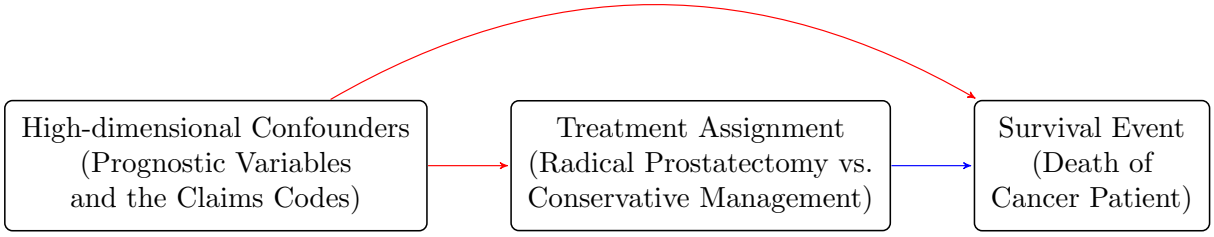


Figure 3.2: Kaplan-Meier curve for treatment (solid) vs control (dashed) across all years.

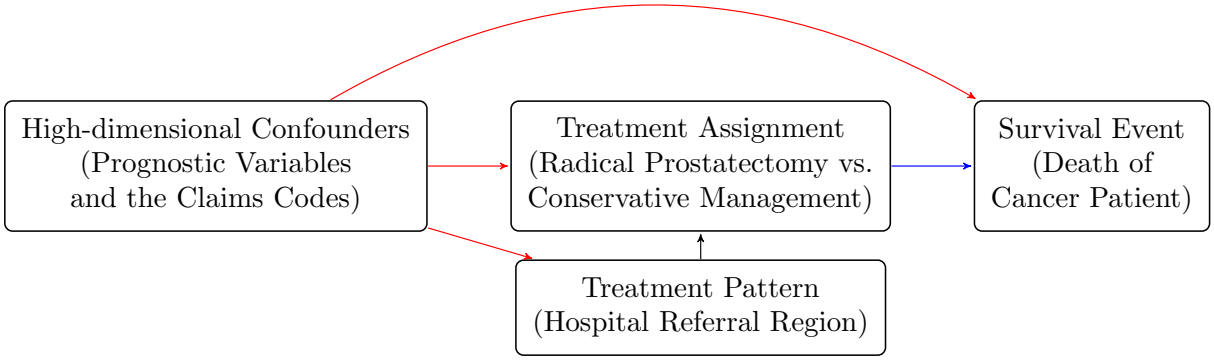
Motivated by our previous linked SEER-Medicare database projects, we consider 16854 prostate cancer patients diagnosed during 2004-2009 as recorded in the SEER-Medicare linked database. The data contains the survival of patients, the treatment information, demographic information, clinical markers and the insurance claims codes. We include in our analysis age, race, martital status, tumor stage, tumor grade, Prostate-Specific-Antigen (PSA), Gleason Score, Prior Charlson Comorbidity Score and 20675 claims codes possessed by at least 10 patients. Among all the patients, 1158 (6.87 %) deaths were observed while 15696 (93.13 %)were still alive by the end of year 2011. Our main focus is the treatment effect of surgery on the overall survival of the patient. In our sample, 5360 (31.80 %) patients received surgery while 11494 (68.20 %) patients received other types of treaments. The Kaplan-Meier curves for the treatment and control groups are presented in Figure 3.2. A summary of statistics of the clinical markers, demographical features and total number of claims codes are presented in Table 3.4. On average

at diagnosis, a patient receiving prostatectomy has 60.7 claims codes while a patient receiving conservative management has 57.7 claims codes.

With the improvement in diagnosis, treatment and management for the disease, the non-cancer causes become the dominant over cancer related causes for the death of patients over the years[LYSY04]. Such changes suggest a comprehensive consideration on both the cancer related and health maintenance related factors for medical decisions on initial treatment. Radical prostatectomy is quite effective reducing the cancer related death, but it comes with its own risk. With the progress in a combination of radiotherapy, chemotherapy and hormonal therapy, the benefit of conservative management can also be competitive for the disease with a rather slow rate of development. To account for the confounding issues, studying the comparative effect of radical prostatectomy on the overall survival of the patient using the observational data requires information from cancer related prognosis and general health factors including demographic information and the medical records, as reflected in the claims codes. Due to the lack of tool for handling high-dimensional claims code data, existing work on the topic either gives up the rich information on the patients' health status [SKD$^+$15, YXM19], or only uses very limited proportion of the information through some summary statistic [HYB$^+$10], and they reached different conclusions. Another issue reported in the study is the existence of heterogeneity in treatment pattern across geographic region [HPG$^+$01]. In our data, the information on the geographic region is described by the hospital referral region (HRR). Conventionally for low-dimensional data, covariates like HRR are excluded from the analysis if they do not associate with the outcome. For high-dimensional data, however, [BCH13] have shown with their proposed"double-selection" that adjusting for covariates associated with either the propensity or the outcome effectively

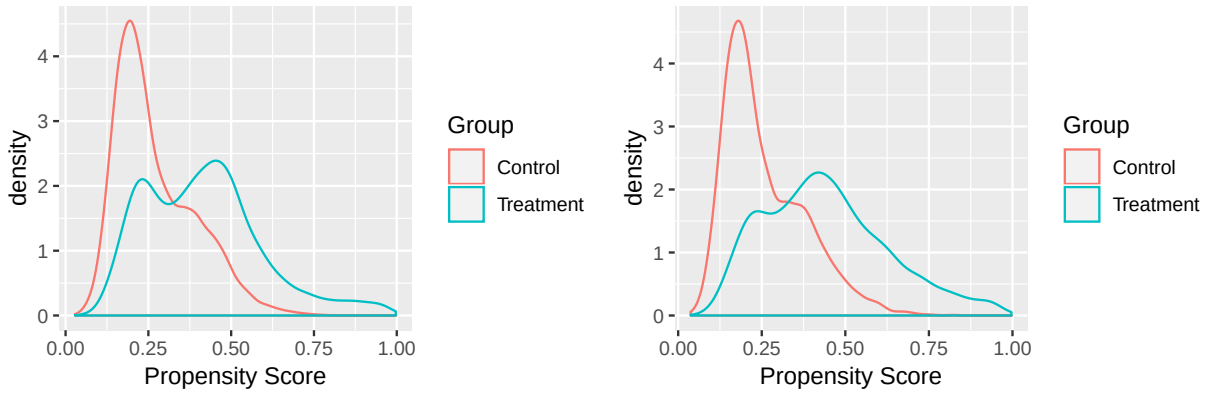


(a) Analysis I: adjust for confounding in clinical markers, demographic features and claims code.



(b) Analysis II: modelling the heterogeneity in treatment pattern across hospital referral regions (HRR) by interactions between HRR and patient features including clinical markers, demographic features and claims code in the propensity score model.

Figure 3.3: Causal diagram of our analyses.



(a) Analysis I: clinical markers, demographic features and claim codes.

(b) Analysis II: clinical markers, demographic features, claim codes and their interaction with Hospital Referral Region (HRR).

Figure 3.4: Distribution of Estimated Propensity Scores. In both analysis, some subjects in the treatment has propensity score close to one while others in control has propensity score close to zero.

corrects bias from the initial regularization and selection process. With the proposed method in Section 3.2-3.3, we study the comparative treatment effect of radical proctectomy in two analyses involving high-dimensional covariates. The causal diagrams of the analyses are illustrated in Figure 3.3. In our Analysis I, we adjust for the potential confounding effect from clinical markers, demographic features and high-dimensional claims code. In our Analysis II, we model the treatment heterogeneity across geographic regions by adding into the propensity score model the interaction between HRR and the covariates in Analysis I. The numbers of covariates considered in the analyses are 4056 and 20676, respectively, both exceeding the number of observed events 1158.

In both analyses, we apply the same methodology as described in the simulation Section 3.4 to estimate the propensity score model (PS) and the additive hazards model (OR) by LASSO. The penalty factors are selected by 10-fold cross-validation. In Figure 3.4, we present the kernel smoothed densities of the estimated propensity scores from both analyses. As mentioned earlier, we observe that some patients received deterministic treatment assignment with the propensity score very close to ether zero and one. The cross-fitted estimators for PS and OR are also obtained through LASSO. The penalty factors in cross-fitting are selected by 9-fold cross-validation. We estimate and draw inference on the treatment effect by $\widehat{\theta}$ with Breslow, closed form $\check{\theta}$ and their cross-fitted counterparts $\widehat{\theta}_{cf}$ and $\check{\theta}_{cf}$. We also provide the results from the marginal analysis without any adjustment, regression adjusted by low-dimensional clinical marks and demographical features and the IPW with the PS estimated by logistic regression on the low-dimensional clinical marks and demographical features. For Analyses I and II, we report the estimates by the additive hazard model LASSO estimate that does not penalize the treatment effect term and the IPW with PS estimated by LASSO. Since the estimators from these two methods are not regular, we do not give variance estimation or inference result.

We report the analysis results in Table 3.5. The point estimates for the treatment effect from all methods are negative, but they vary in the strength and significance of the effect. Both low-dimensional analyses, the covariate adjusted regression and the IPW, suggest that the treatment is not significant at 0.05 level. Once one of the PS model or OR model is employed to adjust for confounding explained by the claims codes through LASSO, the effect strength is magnified, but inference is not available. When both models are utilized through our orthogonal score method, three of the four variations of the method, $\widehat{\theta}$ with Breslow, closed form $\check{\theta}$ and closed

Table 3.5: Estimates of treatment effect ($\times 10^{-3}$) from the linked SEER-Medicare data. Crude analysis did not adjust for any covariates. $\widehat{\theta}$, $\breve{\theta}$, $\widehat{\theta}_{cf}$ and $\breve{\theta}_{cf}$ are the four variations of our orthogonal score approach, where the subscript 'cf' denotes the cross-fitted version; LASSO estimator $\widetilde{\theta}$ penalizes only the covariates effects β but not θ.

Approach	Estimate	SE	95 % CI	p value
Crude analysis	-3.910	0.911	[-5.695 , -2.125]	< 0.001
Adjusted by clinical and demographic features	-1.342	0.972	[-3.247 , 0.564]	0.168
IPW with PS estimated from clinical and demographic features	-0.437	0.627	[-1.666 , 0.792]	0.486
Analysis I: clinical, demographical and claims data				
LASSO $\widetilde{\theta}$ *	-2.674	–	–	–
IPW with PS estimated by LASSO *	-2.642	–	–	–
$\widehat{\theta}$ with Breslow	-2.112	0.969	[-4.012 , -0.212]	0.029
Closed form $\breve{\theta}$	-2.094	0.969	[-3.994 , -0.194]	0.031
$\widehat{\theta}_{cf}$ with Breslow and cross-fitting	-1.809	0.979	[-3.729 , 0.110]	0.065
Closed form $\breve{\theta}_{cf}$ with cross-fitting	-3.223	0.975	[-5.134 , -1.311]	0.001
Analysis II: clinical, demographical, claims data and their interaction with Hospital Referral Region				
LASSO $\widetilde{\theta}$ *	-2.674	–	–	–
IPW with PS estimated by LASSO *	-2.939	–	–	–
$\widehat{\theta}$ with Breslow	-2.050	0.973	[-3.957 , -0.142]	0.035
Closed form $\breve{\theta}$	-2.026	0.973	[-3.934 , -0.119]	0.037
$\widehat{\theta}_{cf}$ with Breslow and cross-fitting	-1.770	0.982	[-3.694 , 0.154]	0.071
Closed form $\breve{\theta}_{cf}$ with cross-fitting	-3.210	0.978	[-5.126 , -1.293]	0.001

* Inference is not directly available. Only the estimates are reported.

form $\check{\theta}_{cf}$ with cross-fitting, suggest that radical prostatectomy has significant benefit compared to conservative managements at level 0.05. According to our simulation study, we recommend to follow the conclusion of $\check{\theta}_{cf}$. The conclusion of our analysis differs from that of [HYB$^+$10], which suggests a potential change in treatment effectiveness between 1995-2003 and 2004-2009. Yet we infer from our analysis a message similar to that of [HYB$^+$10], that the low-dimensional covariates including clinical markers and the demographical features are inadequate to account for confounding. The detail information on the patients past medical records in the claims code, including but not limited to life threatening disease like the heart attack, is very likely the confounder omitted from the low-dimensional analysis. The finding is consistent with existing findings on treatment pattern [HPG$^+$01].

3.6 Discussion

In this paper we have devised the propensity score in a novel way so that the resulting estimate of the treatment effects with biased input from regularized regression is consistent and asymptotically normal (at root-*n* rate). In addition, we provide several refinements to our proposed method to achieve doubly robust estimation in the cases where the propensity score model might be wrong, or the specified survival model might be wrong, or the sparsity assumption is violated. With a delicate choice on the estimator for the nuisance parameter, we obtain a closed form estimator. We also improve our inference result with a relaxed model sparsity condition by incorporating cross-fitting (also known as data-splitting) to our method. We combine the closed form estimator and the cross-fitting together to achieve the doubly robust estimation. Our result

on double robustness extends the existing work by relaxing the assumption of convergence to the "least false" parameter to that of the boundedness of estimation magnitude. While the convergence assumption can hardly be verified in practice, we propose a further regularization step to enforce the proposed bound of the estimation magnitude.

Compared to existing literatures on the inference problem with high-dimensional data, most notably the work of [CCD$^+$18, ZZ14, vdGBRD14, JM14], our paper has its unique contribution. [CCD$^+$18] studied the inference on the treatment effect under partially linear conditional mean model. Their Double Machine Learning approach relies on the orthogonal score constructed with the cross-fitting scheme. The proposed inference method in the Section 4 of [CCD$^+$18] is very general, but it cannot be applied for survival data due to censoring, when the time to event response is not always observable. Moreover, typical models for survival outcome are conditional hazard models that cannot be directly treated as the conditional mean model. Our methodology makes significant progress in the analysis of censored time-to-event data based on the martingale technique under the conditional hazard models. By focusing on the LASSO, we are able to give clear theory for orthogonal score approach without cross-fitting in our Section 3.2, which shows a stronger performance in simulation than its cross-fitted counterpart. The one-step debiasing methods [ZZ14, vdGBRD14, JM14] address the inference problem for low-dimension projection of the coefficients from the high-dimensional regression, which can be applied to draw inference on the treatment effect. Unlike our orthogonal score approach that uses a single score for the treatment effect, the debiased estimation and inference for one coefficient of interest by the one-step debiasing involves the scores for all other nuisance coefficients. Since no covariate is identified as the treatment, the one-step debiasing methods do not involve any propensity

model. All methods of this class utilize a consistent estimation to the sparse precision matrix, i.e. the negative inverse Hessian, to reduce the bias of the initial regularized estimator. Compared to our sparsity assumption enforced upon the propensity model, the sparsity condition on the precision matrix is harder to interpret and verify in practice. As for the finite sample performance, the debiased LASSO is reported to have substantial under-coverage of confidence interval for non-zero coefficients [DBMM15], while our method has shown decent close-to-nominal coverage for the non-zero treatment effect in our simulation.

When the distribution of the censoring time given treatment and covariates can be consistently estimated, we may relax the independence between treatment and censoring in condition (3.5) for inference. Suppose the survival probability of censoring time follows

$$\mathbb{P}(C \geq t | D, \mathbf{Z}) = S(t; D, \mathbf{Z}). \tag{3.47}$$

We modify our score in (3.6) to include $\nu(t; D, \mathbf{Z}) = S(t; D = 0, \mathbf{Z})/S(t; D, \mathbf{Z})$ in the list of nuisance parameter,

$$\phi_c(\theta; \beta, \Lambda, \gamma, \nu) = \frac{1}{n} \sum_{i=1}^{n} \left\{ D_i - \text{expit}(\gamma^\top \mathbf{Z}_{\mathbf{1}i}) \right\} \int_0^\tau e^{D_i \theta t} \nu(t; D_i, \mathbf{Z}_i) dM_i(t; \beta, \Lambda). \tag{3.48}$$

The score (3.48) possesses orthogonality property, so we can use (3.48) to draw inference similar to Theorem 9 with suitable estimator to ν. The doubly robust estimation depends on the actual model for (3.47) and requires potentially delicate arrangements as in Sections 3.3.1-3.3.3.

It is natural to consider the extension of our inference method to the more popular Cox proportional hazards model,

$$\lambda(t; D, \mathbf{Z}) = e^{D\theta + \beta^\top \mathbf{Z}} \lambda_0(t). \tag{3.49}$$

However, such extension is not trivial as the usual score for the relative risk of the treatment is fully nonlinear,

$$\frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau} D_i\left\{dN_i(t)-Y_i(t)e^{D_i\theta+\beta^\top \mathbf{Z}_i}d\Lambda(t)\right\} \tag{3.50}$$

The orthogonalization takes a way more complicated form,

$$\frac{1}{n}\sum_{i=1}^{n}\left\{D_i-\text{expit}\left(\gamma^\top \mathbf{Z}_{\mathbf{1}i}\right)\right\}\int_0^{\tau}\frac{dN_i(t)-Y_i(t)e^{D_i\theta+\beta^\top \mathbf{Z}_i}d\Lambda(t)}{\exp\left\{D_i\theta+\int_0^t\left(e^{D_i\theta}-1\right)e^{\beta^\top \mathbf{Z}_i}d\Lambda(u)\right\}}. \tag{3.51}$$

We are working on the problem in a separate paper.

Another future direction is the inference on treatment affect based on the doubly robust estimation. We suggest three potential solutions to overcome the lack of orthogonality discussed at the end of Section 3.3.3. First, inference is possible when the distribution of the nuisance estimator either is known or can be approximated. Though the distribution of LASSO is largely unknown, some progress has been made under the linear models [TTLT16]. Second, replacing the LASSO by debiased LASSO [ZZ14, vdGBRD14, JM14] could be an attractive approach to supply information on the distribution of the nuisance estimator. However, several challenges exist for this approach. The performance of debiased LASSO under misspecification needs to be studied, and the modification of the debiased LASSO is needed to make consistent model predications. Third, alternative estimation methods for the nuisance parameter can be explored to minimize the bias passed to the estimation of the treatment effect. The idea has been studied for linear models under various names including Targeted estimation (TMLE)[vdL14], bias reduced estimation [VV15] and calibrated estimation [Tan18]. The extension to survival outcomes, however, can be quite technical.

3.7 Technical Details and Proofs

In this section we provide details of all of the theoretical results. We provide additional details on the closed form estimator in Section 3.7.1. We present the proofs of the Theorems and Lemmas stated in the main text in Section 3.7.2. The auxiliary results needed in the proofs, including classical and new concentration results, are stated and proved in Sections 3.7.3-3.7.6, whose proofs are given in Section 3.7.7. The results in Section 3.7.3 are technical preliminary steps in the proofs of the main results. We state and prove them separately to promote the conciseness and readability of the proofs of the main results. Section 3.7.4 contains the classical concentration equalities we use in our proofs. We establish some new concentration results in Section 3.7.5. We put some minor but frequently used results in Section 3.7.6. The notations with letter H are all generic and are replaced by suitable objects when we apply the results.

3.7.1 Details on the closed form estimator

Define

$$\begin{aligned}\check{\Lambda}^1(t;\boldsymbol{\beta},\boldsymbol{\gamma}) &= \int_0^t \frac{\sum_{i=1}^n w_i^1(\boldsymbol{\gamma})\{dN_i(u) - Y_i(u)\boldsymbol{\beta}^\top \mathbf{Z}_i du\}}{\sum_{i=1}^n w_i^1(\boldsymbol{\gamma}) Y_i(u)},\\ \check{\Lambda}^0(t;\boldsymbol{\beta},\boldsymbol{\gamma}) &= \int_0^t \frac{\sum_{i=1}^n w_i^0(\boldsymbol{\gamma})\{dN_i(u) - Y_i(u)\boldsymbol{\beta}^\top \mathbf{Z}_i du\}}{\sum_{i=1}^n w_i^0(\boldsymbol{\gamma}) Y_i(u)}.\end{aligned} \tag{3.52}$$

Under the additive hazards model (3.1), $\check{\Lambda}^1$ and $\check{\Lambda}^0$ can be seen to estimate $\Lambda_0(t)+\theta t$ and $\Lambda_0(t)$, respectively. It is then immediate that (3.19) is equivalent to

$$\check{\theta} = \frac{\sum_{i=1}^n \int_0^\tau w_i^0(\widehat{\boldsymbol{\gamma}}) Y_i(t) d\left\{\check{\Lambda}^1(t;\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\gamma}}) - \check{\Lambda}^0(t;\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\gamma}})\right\}}{\sum_{i=1}^n w_i^0(\widehat{\boldsymbol{\gamma}}) X_i}, \tag{3.53}$$

which estimates θ under the additive hazards model (3.1).

Remark 16. *As a generalization to* (3.53)*, we may draw inference on* θ *using the estimator*

$$\check{\theta}(H) = \frac{\int_0^\tau H(t) d\left\{\check{\Lambda}^1(t;\widehat{\beta},\widehat{\gamma}) - \check{\Lambda}^0(t;\widehat{\beta},\widehat{\gamma})\right\}}{\int_0^\tau H(t)dt} \tag{3.54}$$

for any adapted process $H(t)$ *such that* $\int_0^\tau H(t)dt$ *is bounded away from zero. It can be shown that all such* $\check{\theta}(H)$ *have the same asymptotic distribution under the conditions of Theorem 9.*

3.7.2 Proof of Main Results

Proof of Lemma 26. We first verify the identifiability of the true parameters by the score. At the true parameters $(\theta_0, \beta_0, \Lambda_0, \gamma_0)$, $M_i(t;\theta_0,\beta_0,\Lambda_0)$ is a martingale with respect to filtration $\mathcal{F}_{n,t} = \sigma\{N_i(u), Y_i(u), D_i, \mathbf{Z}_i : u \le t, i = 1,\ldots,n\}$. Since the other elements D_i and $\mathbf{Z}_i$ are all measurable with respect to $\mathcal{F}_{n,t}$, the martingale integral $\phi(\theta_0;\beta_0,\Lambda_0,\gamma_0)$ is also a $\mathcal{F}_{n,t}$-martingale. Therefore, $\mathbb{E}\{\phi(\theta_0;\beta_0,\Lambda_0,\gamma_0)\} = 0$.

To show the orthogonality, we define the directional perturbations

$$\beta_r = \beta_0 + r\triangle\beta,\ \Lambda_r(t) = \Lambda_0(t) + r\triangle\Lambda(t) \text{ and } \gamma_r = \gamma_0 + r\triangle\gamma.$$

We decompose the expected directional derivative in nuisance parameters at the true parameters into 2 terms,

$$\begin{aligned}
&\frac{\partial}{\partial r}\mathbb{E}\{\phi(\theta_0;\beta_r,\Lambda_r,\gamma_r)\}\Big|_{r=0} \\
=& -\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n \left\{D_i - \text{expit}(\gamma_0^\top \mathbf{Z}_{\mathbf{1}i})\right\} \int_0^\tau e^{D_i\theta_0 t} Y_i(t) \left\{\triangle\beta^\top \mathbf{Z}_i dt + d\triangle\Lambda(t)\right\}\right] \\
&-\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n \frac{e^{\gamma_0^\top \mathbf{Z}_{\mathbf{1}i}}}{\left(1+e^{\gamma_0^\top \mathbf{Z}_{\mathbf{1}i}}\right)^2}\triangle\gamma^\top \mathbf{Z}_{\mathbf{1}i} \int_0^\tau e^{D_i\theta_0 t} dM_i(t;\theta_0,\beta_0,\Lambda_0)\right].
\end{aligned}$$

The effect of treatment D_i on the conditional expectation of the at-risk process

$$\mathbb{E}\{Y_i(t)|D_i,\mathbf{Z}_i\} = \mathbb{P}(T_i \geq t|D_i,\mathbf{Z}_i)\mathbb{P}(C_i \geq t|D_i,\mathbf{Z}_i)$$

has two components, the effect on the event-time and that on the censoring time. Under the additive hazards model (3.1), the survival probability at time t is modified by a $e^{-D_i\theta_0 t}$ factor,

$$\mathbb{P}(T_i \geq t|D_i,\mathbf{Z}_i) = \mathbb{P}(T_i \geq t|D_i=0,\mathbf{Z}_i)e^{-\int_0^t D_i\theta_0 dt} = \mathbb{P}(T_i \geq t|D_i=0,\mathbf{Z}_i)e^{-D_i\theta_0 t}.$$

With the conditional independence between treatment and censoring (3.5), we have

$$\mathbb{P}(C_i \geq t|D_i,\mathbf{Z}_i) = \mathbb{P}(C_i \geq t|D_i=0,\mathbf{Z}_i) = \mathbb{P}(C_i \geq t|\mathbf{Z}_i).$$

We prove in Lemma 41, that $\mathbb{E}\{e^{D_i\theta_0 t}Y_i(t)|\mathbf{Z}_i,D_i\} = \mathbb{E}\{Y_i(t)|\mathbf{Z}_i,D_i=0\}$ is $\sigma\{\mathbf{Z}_i\}$-measurable under model (3.1) and condition (3.5). Using the fact $\mathbb{E}\left\{D_i - \text{expit}(\boldsymbol{\gamma}_0^\top \mathbf{Z}_{\mathbf{1}i})|\mathbf{Z}_i\right\} = 0$ under model (3.2), we apply the tower property of conditional expectation to calculate that the first term equals zero,

$$\int_0^\tau \mathbb{E}\left[\mathbb{E}\left\{D_i - \text{expit}(\boldsymbol{\gamma}_0^\top \mathbf{Z}_{\mathbf{1}i})|\mathbf{Z}_i\right\}\mathbb{E}\left\{e^{D_i\theta_0 t}Y_i(t)|D_i,\mathbf{Z}_i\right\}\left\{\triangle\boldsymbol{\beta}^\top \mathbf{Z}_i dt + d\triangle\Lambda(t)\right\}\right] = 0.$$

The second term is again a $\mathcal{F}_{n,t}$-martingale, so it also has mean zero. Therefore, the expected directional derivative in nuisance parameters at the true parameters is the sum of two zero terms, which is zero. By definition of orthogonality, the score ϕ is orthogonal. □

Proof of Theorem 9. We use the orthogonality of the score (3.6) to establish under Assumption 5

$$\widehat{\theta} - \theta_0 = \frac{\phi(\theta_0;\boldsymbol{\beta}_0,\Lambda_0,\boldsymbol{\gamma}_0) + o_p(|\widehat{\theta} - \theta_0|)}{n^{-1}\sum_{i=1}^n D_i\{1-\text{expit}(\boldsymbol{\gamma}_0^\top \mathbf{Z}_{\mathbf{1}i})\}(e^{\theta_0 X_i}-1)/\theta_0}. \tag{3.55}$$

The proof of (3.55) involves tedious calculation, so we present the proof separately in Lemma 29.

When the dimension of covariates $\mathbf{Z}$ is fixed, the representation (3.55) immediately leads to asymptotic normality through mere formality. However, the growing dimension of covariates in our high-dimensional setting may cause the violation of the classical boundedness assumptions on the summands of $\phi(\theta_0;\beta_0,\Lambda_0,\gamma_0)$. We go through the following technicalities to achieve the same asymptotic normality result in high-dimensions without any additional assumption.

The rest of our proof takes 4 steps. First, we show that $\widehat{\theta}$ is consistent for θ_0. Second, we establish the asymptotic normality of the score ϕ at true parameter. Third, we obtain the $\sqrt{n}$-tightness and the asymptotic distribution of $\widehat{\theta}-\theta_0$. Finally, we show that the variance estimator is consistent.

Step 1:

Under model (3.1),

$$\phi(\theta_0;\beta_0,\Lambda_0,\gamma_0)=\frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau}\{D_i-\text{expit}(\gamma_0^{\top}\mathbf{Z}_{\mathbf{1}i})\}e^{\theta_0 D_i t}dM_i(t) \tag{3.56}$$

is the final point of a martingale with respect to filtration $\mathcal{F}_{n,t}=\sigma\{N_i(u),Y_i(u),D_i,\mathbf{Z}_i : u\le t, i=1,\ldots,n\}$. Its expectation is thus zero,

$$\mathbb{E}\{\phi(\theta_0;\beta_0,\Lambda_0,\gamma_0)\}=0. \tag{3.57}$$

The true θ_0 is thus identified by the score ϕ. We apply the concentration result of Lemma 35 with (3.57), getting

$$\phi(\theta_0;\beta_0,\Lambda_0,\gamma_0,W_i)=o_p(1). \tag{3.58}$$

Under Assumption 5-ii, we use the martingale property of $M(t)$, defined by (3.3), and Lemma 41

to calculate the derivative with respect to θ at θ_0

$$\begin{aligned}
&\frac{\partial}{\partial\theta}\mathbb{E}\left\{\phi(\theta;\beta_0,\Lambda_0,\gamma_0)\right\}\Big|_{\theta=\theta_0} \\
= \;& \mathbb{E}\mathbb{E}\left(\{D-\text{expit}(\gamma_0^\top \mathbf{Z_1})\}D\mathbb{E}\left[\int_0^\tau e^{D\theta_0 t}\{tdM(t;D,\mathbf{Z})-Y_i(t)dt\}\Big|D,\mathbf{Z}\right]\Big|\mathbf{Z}\right) \\
= \;& -\int_0^\tau \mathbb{E}[\mathbb{E}\{Y(t)|\mathbf{Z};D=0\}\,\text{Var}(D|\mathbf{Z})]dt. \qquad (3.59)
\end{aligned}$$

Under Assumption 5-v, (3.59) is bounded away from zero

$$\int_0^\tau \mathbb{E}[\mathbb{E}\{Y(t)|\mathbf{Z};D=0\}\,\text{Var}(D|\mathbf{Z})]dt \geq \int_0^\tau \mathbb{E}[\mathbb{E}\{Y(\tau)|\mathbf{Z};D=0\}\,\text{Var}(D|\mathbf{Z})]dt \geq \tau\varepsilon_Y. \qquad (3.60)$$

Since the summands in the denominator of (3.55) has bound

$$|D_i\{1-\text{expit}(\gamma_0^\top \mathbf{Z_1}_i)\}(e^{\theta_0 X_i}-1)/\theta_0| \leq e^{\tau\theta_0}\tau, \qquad (3.61)$$

we can use the Hoeffding's inequality (as in Lemma 31) to establish a lower bound

$$\mathbb{P}\left(n^{-1}\sum_{i=1}^n D_i\{1-\text{expit}(\gamma_0^\top \mathbf{Z_1}_i)\}(e^{\theta_0 X_i}-1)/\theta_0 > \varepsilon_Y/2\right) > 1-e^{-\frac{n\varepsilon_Y^2}{8e^{2\tau\theta_0}\tau^2}}. \qquad (3.62)$$

Plugging the rate (3.58) and the lower bound (3.62) into (3.55), we conclude that $\widehat{\theta}-\theta_0 = o_p(1)$.

Step 2: Let $X_{(1)},\ldots,X_{(n)}$ be the order statistics of the observed times and

$$\begin{aligned}
M_k^1 &= \frac{1}{n}\sum_{i=1}^n \int_0^{X_{(k)}} D_i\{1-\text{expit}(\gamma_0^\top \mathbf{Z_1}_i)\}e^{\theta_0 t}dM_i(t), \\
M_k^0 &= \frac{1}{n}\sum_{i=1}^n \int_0^{X_{(k)}} (1-D_i)\,\text{expit}(\gamma_0^\top \mathbf{Z_1}_i)dM_i(t), \qquad (3.63)
\end{aligned}$$

for $k=0,\ldots,n$. We note that the score ϕ with true parameters can be alternatively expressed as

$$\phi(\theta_0;\beta_0,\Lambda_0,\gamma_0) = M_n^1 - M_n^0. \qquad (3.64)$$

Since both integrands in (3.63) are nonnegative and bounded by $\tau(1\vee e^{\theta_0\tau})$, we can apply the Lemma 36 to get that both M_k^1 and M_k^0, hence $M_k^1-M_k^0$, are martingales under filtration

$\mathcal{F}_k^M = \sigma\{N_i(u), Y_i(u+), D_i, \mathbf{Z}_i : u \in [0,t_k], i = 1,\ldots,n\}$ satisfying, most importantly,

$$\max\left\{\mathbb{E}\left\{(M_k^1 - M_{k-1}^1)^2 | \mathcal{F}_k^M\right\}, \mathbb{E}\left\{(M_k^0 - M_{k-1}^0)^2 | \mathcal{F}_k^M\right\}\right\} \leq 8\tau^2(1 \vee e^{\theta_0\tau})^2/n^2. \tag{3.65}$$

By the Cauchy-Schwartz inequality, we have

$$(M_k^1 - M_k^0 - M_{k-1}^1 + M_{k-1}^0)^2 \leq 2(M_k^1 - M_{k-1}^1)^2 + 2(M_k^0 - M_{k-1}^0)^2. \tag{3.66}$$

Hence, we establish an upper bound for the quadratic variation of $\psi(\theta_0; \beta_0, \Lambda_0, \gamma_0)$ from (3.65) and (3.66),

$$\mathbb{E}\left\{(M_k^1 - M_{k-1}^1 - M_k^0 + M_{k-1}^0)^2 | \mathcal{F}_k^M\right\} \leq 32\tau^2(1 \vee e^{\theta_0\tau})^2/n^2. \tag{3.67}$$

As a result, the variance

$$\sigma_\phi^2 = \mathrm{Var}\{\sqrt{n}\phi(\theta_0; \beta_0, \Lambda_0, \gamma_0)\} = n\mathbb{E}\left[\sum_{i=1}^n \mathbb{E}\left\{(M_k^1 - M_k^0 - M_{k-1}^1 + M_{k-1}^0)^2 | \mathcal{F}_k^M\right\}\right] \tag{3.68}$$

is finite, bounded by $32\tau^2(1 \vee e^{\theta_0\tau})^2$.

Now, we verify the Lindeberg condition for the martingale central limit theorem [Bro71]. The event

$$\sqrt{n}|M_k^1 - M_{k-1}^1 - M_k^0 + M_{k-1}^0| > \varepsilon \tag{3.69}$$

occurs only if one of

$$\sqrt{n}|M_k^1 - M_{k-1}^1| > \varepsilon/2 \text{ or } \sqrt{n}|M_k^0 - M_{k-1}^0| > \varepsilon/2 \tag{3.70}$$

occurs. Thus, we must have the following inequality

$$\begin{aligned} & I(\sqrt{n}|M_k^1 - M_{k-1}^1 - M_k^0 + M_{k-1}^0| > \varepsilon) \\ \leq\ & I(\sqrt{n}|M_k^1 - M_{k-1}^1| > \varepsilon/2) + I(\sqrt{n}|M_k^0 - M_{k-1}^0| > \varepsilon/2). \end{aligned} \tag{3.71}$$

Along with (3.66), we have

$$\begin{aligned} & n\sum_{i=1}^{n}\mathbb{E}\left\{(M_k^1-M_{k-1}^1-M_k^0+M_{k-1}^0)^2;\sqrt{n}|M_k^1-M_k^0-M_{k-1}^1+M_{k-1}^0|>\varepsilon\right\} \\ \leq\ & 2n\sum_{i=1}^{n}\mathbb{E}\left\{(M_k^1-M_{k-1}^1)^2;\sqrt{n}|M_k^1-M_{k-1}^1|>\varepsilon/2\right\} \\ & +2n\sum_{i=1}^{n}\mathbb{E}\left\{(M_k^0-M_{k-1}^0)^2;\sqrt{n}|M_k^0-M_{k-1}^0|>\varepsilon/2\right\}. \end{aligned} \tag{3.72}$$

By Lemma 36, the limit of the right hand side in (3.72) is zero. Hence, we can apply the martingale central limit theorem to

$$\sqrt{n}\sigma_\phi^{-1}\phi(\theta_0;\beta_0,\Lambda_0,\gamma_0,W_i)\rightsquigarrow N(0,1). \tag{3.73}$$

Step 3: We define the asymptotic standard deviation of $\sqrt{n}(\widehat{\theta}-\theta_0)$ as

$$\sigma=\sigma_\phi/\mathbb{E}[D\{1-\text{expit}(\gamma_0^\top\mathbf{Z}_1)\}(e^{\theta_0 X}-1)/\theta_0]. \tag{3.74}$$

Since $\widehat{\theta}$ solves $\phi(\theta;\widehat{\beta},\widehat{\Lambda}(\theta),\widehat{\gamma})=0$, we have along with Lemma 29

$$\begin{aligned} & \sqrt{n}\sigma^{-1}(\widehat{\theta}-\theta_0)-\frac{\sqrt{n}}{\sigma_\phi}\phi(\theta_0;\beta_0,\gamma_0,W_i) \\ =\ & \frac{\sqrt{n}(\widehat{\theta}-\theta_0)}{\sigma_\phi}\Bigg(\mathbb{E}[D\{1-\text{expit}(\gamma_0^\top\mathbf{Z}_1)\}(e^{\theta_0 X}-1)/\theta_0] \\ & -\frac{1}{n}\sum_{i=1}^{n}D_i\{1-\text{expit}(\gamma_0^\top\mathbf{Z}_{\mathbf{1}i})\}(e^{\theta_0 X_i}-1)/\theta_0\Bigg) \\ & +o_p(1+\sqrt{n}|\widehat{\theta}-\theta_0|). \end{aligned} \tag{3.75}$$

Again using the bound (3.61), we apply the Hoeffding's inequality (as in Lemma 31) to establish that

$$\mathbb{E}[D\{1-\text{expit}(\gamma_0^\top\mathbf{Z}_1)\}(e^{\theta_0 X}-1)/\theta_0]-\frac{1}{n}\sum_{i=1}^{n}D_i\{1-\text{expit}(\gamma_0^\top\mathbf{Z}_{\mathbf{1}i})\}(e^{\theta_0 X_i}-1)/\theta_0 \tag{3.76}$$

is of order $O_p(n^{-1/2})$. Hence, the right hand side of (3.75) is of order $o_p(1+\sqrt{n}|\widehat{\theta}-\theta_0|)$. Along with the normality (3.73), we establish the $\sqrt{n}$-tightness of the estimation error

$$|\widehat{\theta}-\theta_0| = O_p\big(n^{-1/2}\big). \tag{3.77}$$

Plugging in the rate of estimation error into the righthand side of (3.75), we obtain the asymptotic equivalence

$$\sqrt{n}\sigma^{-1}(\widehat{\theta}-\theta_0) - \sqrt{n}\sigma_\phi^{-1}\phi_n(\theta_0;\widehat{\beta},\widehat{\gamma}) = o_p(1). \tag{3.78}$$

Step 4: To show that $\widehat{\sigma}^{-1}$ defined in (3.11) is a consistent estimator for σ^{-1}, we decompose the numerator of $\widehat{\sigma}^2$ into

$$\begin{aligned}
& \frac{1}{n}\sum_{i=1}^n \delta_i\{D_i - \text{expit}(\widehat{\gamma}^\top \mathbf{Z_1}_i)\}^2 e^{2\widehat{\theta}D_iX_i} \\
= \; & \frac{1}{n}\sum_{i=1}^n \Big[\delta_i\{D_i - \text{expit}(\widehat{\gamma}^\top \mathbf{Z_1}_i)\}^2 e^{2\widehat{\theta}D_iX_i} - \{D_i - \text{expit}(\gamma_0^\top \mathbf{Z_1}_i)\}^2 e^{2\theta_0 D_iX_i}\Big] \\
& + \frac{1}{n}\sum_{i=1}^n \int_0^\tau [\{D_i - \text{expit}(\gamma_0^\top \mathbf{Z_1}_i)\}e^{\theta_0 D_i t}]^2 dN_i(t).
\end{aligned} \tag{3.79}$$

By mean value theorem, the first term in the righthand side of (3.79) can be written in terms of $\theta_\xi = (1-\xi)\theta_0 + \xi\widehat{\theta}$ and $\gamma_\xi = (1-\xi)\gamma_0 + \xi\widehat{\gamma}$ with some $\xi \in [0,1]$,

$$\begin{aligned}
& \frac{(\gamma_0-\widehat{\gamma})^\top}{n}\sum_{i=1}^n \frac{\delta_i\{D_i - \text{expit}(\gamma_\xi^\top \mathbf{Z_1}_i)\}e^{\gamma_\xi^\top \mathbf{Z_1}_i} e^{2\theta_\xi D_iX_i}}{\Big(1+e^{\gamma_\xi^\top \mathbf{Z_1}_i}\Big)^2} \\
& + \frac{(\widehat{\theta}-\theta_0)}{n}\sum_{i=1}^n 2\delta_i D_iX_i\{1 - \text{expit}(\gamma_\xi^\top \mathbf{Z_1}_i)\}^2 e^{2\theta_\xi D_iX_i} \\
= \; & O_p(\|\widehat{\gamma}-\gamma_0\|_1 + |\widehat{\theta}-\theta_0|).
\end{aligned} \tag{3.80}$$

The second term on the righthand side of (3.79) is the optional quadratic variation of $\phi(\theta_0;\beta_0,\gamma_0,\Lambda_0)$ bounded by $e^{2\theta_0\tau}$. By the Hoeffding's inequality (as in Lemma 31), we have the concentration of

the second term around the variance of $\phi(\theta_0;\beta_0,\gamma_0,\Lambda_0)$,

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau}[\{D_i-\text{expit}(\gamma_0^{\top}\mathbf{Z}_{\mathbf{1}i})\}e^{\theta_0 D_i t}]^2 dN_i(t)\\
=\;&\mathbb{E}\left(\int_0^{\tau}[\{D_i-\text{expit}(\gamma_0^{\top}\mathbf{Z}_{\mathbf{1}i})\}e^{\theta_0 D_i t}]^2 dN_i(t)\right)+O_p(n^{-1/2})\\
=\;&\sigma_{\phi}^2+o_p(1). \qquad (3.81)
\end{aligned}$$

Putting (3.80) and (3.81) together, we have the numerator of $\widehat{\sigma}^2$ (3.58) equals $\sigma_{\phi}^2+o_p(1)$. Similarly, we decompose the denominator of σ into

$$\frac{1}{n}\sum_{i=1}^{n}\left[D_i\{1-\text{expit}(\widehat{\gamma}^{\top}\mathbf{Z}_{\mathbf{1}i})\}\frac{e^{\widehat{\theta}X_i}-1}{\widehat{\theta}_0}-D_i\{1-\text{expit}(\gamma_0^{\top}\mathbf{Z}_{\mathbf{1}i})\}\frac{e^{\theta_0 X_i}-1}{\theta_0}\right] \qquad (3.82)$$

minus (3.76). Again, we have (3.82) is of order $O_p(|\widehat{\theta}-\theta_0|+\|\widehat{\gamma}-\gamma_0\|_1)=o_p(1)$ through mean value theorem. Under the additive hazards model (3.1), we must have

$$\beta_0^{\top}\mathbf{Z}+d\Lambda_0(t)\geq 0 \qquad (3.83)$$

for all $\mathbf{Z}$ such that $\Pr(D=0|\mathbf{Z})>0$. Under Assumptions 5-ii, 5-v and 5-vi, we can establish a lower bound for σ_{ϕ}

$$\begin{aligned}
\sigma_{\phi} &= \mathbb{E}\left[\int_0^{\tau}\{D-\text{expit}(\gamma_0^{\top}\mathbf{Z})\}^2 e^{2D\theta_0 t}Y(t)\{(D\theta_0+\beta_0^{\top}\mathbf{Z})dt+d\Lambda_0(t)\}\right]\\
&= \mathbb{E}\left[\int_0^{\tau}\{D-\text{expit}(\gamma_0^{\top}\mathbf{Z})\}^2 e^{D\theta_0 t}\mathbb{E}\{Y(t)|\mathbf{Z};D=0\}\{(D\theta_0+\beta_0^{\top}\mathbf{Z})dt+d\Lambda_0(t)\}\right]\\
&= \mathbb{E}\left[\int_0^{\tau}D\{1-\text{expit}(\gamma_0^{\top}\mathbf{Z})\}^2 e^{\theta_0 t}\mathbb{E}\{Y(\tau)|\mathbf{Z};D=0\}\theta_0 dt\right]\\
&\quad+\mathbb{E}\left[\int_0^{\tau}\{D-\text{expit}(\gamma_0^{\top}\mathbf{Z})\}^2 e^{D\theta_0 t}d\mathbb{E}\{N(t)|\mathbf{Z};D=0\}\right]\\
&\geq 0+e^{1\wedge\theta_0\tau}\mathbb{E}[\text{Var}(D|\mathbf{Z})\mathbb{E}\{N(\tau)|\mathbf{Z};D=0\}]\\
&\geq e^{1\wedge\theta_0\tau}\varepsilon_N. \qquad (3.84)
\end{aligned}$$

Hence, the limit is bounded by

$$\sigma^{-1} = \frac{\mathbb{E}\left[D\{1-\text{expit}(\gamma_0^\top \mathbf{Z}_1)\}\frac{e^{\theta_0 X}-1}{\theta_0}\right]}{\sqrt{\mathbb{E}[\delta\{D-\text{expit}(\gamma_0^\top \mathbf{Z}_1)\}^2 e^{2D\theta_0 X}]}} \leq \frac{\tau e^{\theta_0 \tau}}{\sqrt{e^{1\wedge\theta_0\tau}\varepsilon_N}}. \tag{3.85}$$

Therefore, we have

$$\widehat{\sigma}^{-1} = \sigma^{-1} + o_p(1) \tag{3.86}$$

by continuous mapping theorem.

Combining the results (3.73), (3.78) and (3.86), we obtain

$$\sqrt{n}\widehat{\sigma}^{-1}(\widehat{\theta}-\theta_0) \rightsquigarrow N(0,1). \tag{3.87}$$

This is the desired conclusion. We hence finish the proof. □

Proof of Theorem 10. We obtain from Lemma 30 the same representation as (3.55),

$$\widehat{\theta}_{cf} - \theta_0 = \frac{\phi(\theta_0;\beta_0,\Lambda_0,\gamma_0) + o_p(|\widehat{\theta}-\theta_0|)}{n^{-1}\sum_{i=1}^n D_i\{1-\text{expit}(\gamma_0^\top \mathbf{Z}_{\mathbf{1}i})\}(e^{\theta_0 X_i}-1)/\theta_0}. \tag{3.88}$$

The rest of the proof is identical to the Steps 1-4 in the proof of Theorem 9. □

Proof of Lemma 27. To see that zero is always in the LASSO regularization path, we shall spell out the associated penalty factor. The gradient of the loss $l_\beta(\boldsymbol{\beta}) = \boldsymbol{\beta}^\top H_n \boldsymbol{\beta} - 2\boldsymbol{\beta}^\top \mathbf{h}_n$ in the additive hazards model LASSO (3.12) at $\boldsymbol{\beta} = 0$ is

$$\nabla l_\beta(0) = -\mathbf{h}_n = -\frac{2}{n}\sum_{i=1}^n \int_0^\tau \left[\left\{\mathbf{Z}_i - \bar{\mathbf{Z}}(t)\right\} dN_i(t)\right].$$

Since we are studying a pure computational matter, we may use the computable vector $\nabla l_\beta(0)$ to set up $\lambda > \|\nabla l_\beta(0)\|_\infty$ [GRS12]. With the λ thus chosen, we have $\boldsymbol{\beta} = 0$ satisfy the LASSO KKT condition $\left\|\nabla l_\beta(0)\right\|_\infty < \lambda$. Therefore, zero is an element in the regularization path. By the optimality of $\widehat{\lambda}_\beta$ according to (3.35), $l_\beta^*(0) = 0$ must be an upper bound for $l_\beta^*(\widehat{\boldsymbol{\beta}}(\widehat{\lambda}_\beta))$.

Then, we derive the lower bound of $l^*_\beta(\widehat{\boldsymbol{\beta}}(\widehat{\lambda}_\beta))$ related to the magnitude $\mathcal{M}_\beta\left(\widehat{\boldsymbol{\beta}}\right)$. We apply the Cauchy-Schwartz inequality to the linear term in $l^*_\beta(\boldsymbol{\beta})$,

$$\begin{aligned}
l^*_\beta(\boldsymbol{\beta}) =& \mathcal{M}_\beta(\boldsymbol{\beta})^2 - 2\int_0^\tau \mathbb{E}_*\left[\boldsymbol{\beta}^\top\{\mathbf{Z}_* - \mu(t)\}\, dN_*(t)\right] \\
=& \mathcal{M}_\beta(\boldsymbol{\beta})^2 - 2\int_0^\tau \mathbb{E}_*\left[\boldsymbol{\beta}^\top\{\mathbf{Z}_* - \mu(t)\}\, Y_*(t) g(t,\mathbf{Z}_*) dt\right] \\
\geq& \mathcal{M}_\beta(\boldsymbol{\beta})\left(\mathcal{M}_\beta(\boldsymbol{\beta}) - 2\sqrt{\int_0^\tau \mathbb{E}_*[Y_*(t)g(t,\mathbf{Z}_*)^2]\, dt}\right).
\end{aligned}$$

Putting the upper bound and lower bound to gather, we must have

$$\mathcal{M}_\beta(\boldsymbol{\beta}) \leq 2\sqrt{\int_0^\tau \mathbb{E}_*[Y_*(t)g(t,\mathbf{Z}_*)^2]\, dt}.$$

This is the conclusion of the Lemma. □

Proof of Lemma 28. To see that the intercept only estimator $\widehat{\gamma}_0$ is always in the LASSO regularization path, we shall spell out the associated penalty factor. The intercept only estimator $\widehat{\gamma}_0$ makes constant predictions $\text{expit}(\widehat{\gamma}_0^\top \mathbf{z}) = \bar{D} = \sum_{i=1}^n D_i/n$. The gradient of the loss $\nabla l_\gamma(\boldsymbol{\gamma}) = -n^{-1}\sum_{i=1}^n\{D_i\boldsymbol{\gamma}^\top\mathbf{Z}_i - \log(1-e^{\boldsymbol{\gamma}^\top\mathbf{Z}_i})\}$ in the logistic regression LASSO (3.13) is

$$\nabla l_\gamma(\widehat{\gamma}_0) = -\frac{1}{n}\sum_{i=1}^n\left[\{D_i - \bar{D}\}\begin{pmatrix}1\\ \mathbf{Z}_i\end{pmatrix}\right] = \begin{pmatrix}0\\ -\frac{1}{n}\sum_{i=1}^n\{D_i-\bar{D}\}\mathbf{Z}_i\end{pmatrix}.$$

Since we are studying a pure computational matter, we may use the computable vector $\nabla l_\beta(0)$ to set up $\lambda > \|\frac{1}{n}\sum_{i=1}^n\{D_i - \bar{D}\}\mathbf{Z}_i\|_\infty$ [FHT10]. Notice that we follow [FHT10] in (3.13) by leaving the intercept term not penalized. With the λ thus chosen, we have the first coordinate in $|\nabla l_\gamma(\widehat{\gamma}_0)|$ being zero and the rest strictly smaller than λ. Therefore, $\widehat{\gamma}_0$ is an element in the regularization path. By the Markov inequality, $\widehat{\gamma}_0$ converges to $(\log(1-1/\mathbb{E}_*(D_*)), 0, \ldots, 0)$. Under Assumption 5-v, $\varepsilon_Y \leq \mathbb{E}_*(D_*) \leq 1-\varepsilon_Y$, so we have an upper bound for $l^*_\gamma(\widehat{\gamma}_0)$,

$$l^*_\gamma(\widehat{\gamma}_0) \leq -\mathbb{E}_*(D_*)\log(\mathbb{E}_*(D_*)) - \{1-\mathbb{E}_*(D_*)\}\log(1-\mathbb{E}_*(D_*)) + o_p(1) \leq -\log(\varepsilon_Y) + o_p(1).$$

By the optimality of $\widehat{\lambda}_\gamma$ according to (3.35), the upper bound of $l^*_\gamma(\widehat{\gamma}_0)$ must also be an upper bound for $l^*_\gamma(\widehat{\gamma}(\widehat{\lambda}_\gamma))$.

Define the set $\mathcal{Z} = \{\mathbf{z} : \mathbb{E}_*(D_*|\mathbf{Z}_* = \mathbf{z}) \geq \varepsilon_Y/2, \mathbb{E}_*(1-D_*|\mathbf{Z}_* = \mathbf{z}) \geq \varepsilon_Y/2$ and $\mathbb{E}_*(Y_*(\tau)|\mathbf{Z}_* = \mathbf{z}, D_* = 0) \geq \varepsilon_Y/2\}$. We decompose

$$
\begin{aligned}
&\mathbb{E}_*[\underset{*}{\mathrm{Var}}(D_*|\mathbf{Z}_*)\mathbb{E}_*\{Y_*(\tau)|\mathbf{Z}_*, D_* = 0\}] \\
=&\mathbb{E}_*[\mathbb{E}_*(D_*|\mathbf{Z}_*)\mathbb{E}_*(1-D_*|\mathbf{Z}_*)\mathbb{E}_*\{Y_*(\tau)|\mathbf{Z}_*, D_* = 0\}] \\
=&\mathbb{E}_*[\mathbb{E}_*(D_*|\mathbf{Z}_*)\mathbb{E}_*(1-D_*|\mathbf{Z}_*)\mathbb{E}_*\{Y_*(\tau)|\mathbf{Z}_*, D_* = 0\} I(\mathbf{Z}_* \in \mathcal{Z})] \\
&+\mathbb{E}_*[\mathbb{E}_*(D_*|\mathbf{Z}_*)\mathbb{E}_*(1-D_*|\mathbf{Z}_*)\mathbb{E}_*\{Y_*(\tau)|\mathbf{Z}_*, D_* = 0\} I(\mathbf{Z}_* \in \mathcal{Z}^c)] \\
\leq&\mathbb{P}_*(\mathbf{Z}_* \in \mathcal{Z}) + \varepsilon_Y/2.
\end{aligned}
$$

To satisfy Assumption 5-v, $\mathbb{P}_*(\mathbf{Z}_* \in \mathcal{Z})$ must be at least $\varepsilon_Y/2$. Then, we derive a lower bound of $l^*_\gamma(\gamma)$ by analyzing the expectation in set $\mathcal{Z}$

$$
\begin{aligned}
l^*_\gamma(\gamma) =& -\mathbb{E}_*[D_* \log\{\mathrm{expit}(\gamma^\top \mathbf{Z}_*)\} + (1-D_*)\log\{1-\mathrm{expit}(\gamma^\top \mathbf{Z}_*)\}] \\
\geq& -\mathbb{E}_*[D_* \log\{\mathrm{expit}(\gamma^\top \mathbf{Z}_*)\} + (1-D_*)\log\{1-\mathrm{expit}(\gamma^\top \mathbf{Z}_*)\}|\mathbf{Z}_* \in \mathcal{Z}]\mathbb{P}_*(\mathbf{Z}_* \in \mathcal{Z}) \\
\geq& -\varepsilon_Y^2/4\mathbb{E}_*[\log\{\mathrm{expit}(\gamma^\top \mathbf{Z}_*)\}|\mathbf{Z}_* \in \mathcal{Z}] - \varepsilon_Y^2/4\mathbb{E}_*[\log\{1-\mathrm{expit}(\gamma^\top \mathbf{Z}_*)\}|\mathbf{Z}_* \in \mathcal{Z}] \\
\geq& -\varepsilon_Y^2/4\log\left(\mathbb{E}_*\{\mathrm{expit}(\gamma^\top \mathbf{Z}_*)|\mathbf{Z}_* \in \mathcal{Z}\}\right) - \varepsilon_Y^2/4\log\left(\mathbb{E}_*\{1-\mathrm{expit}(\gamma^\top \mathbf{Z}_*)|\mathbf{Z}_* \in \mathcal{Z}\}\right).
\end{aligned}
$$

The last step above is the consequence of the Jensen's inequality.

Putting the upper bound and lower bound of $l^*_\gamma(\widehat{\gamma}(\widehat{\lambda}_\gamma))$ together, we have

$$
\begin{aligned}
&\mathbb{E}_*\{\mathrm{expit}(\widehat{\gamma}(\widehat{\lambda}_\gamma)^\top \mathbf{Z}_*)|\mathbf{Z}_* \in \mathcal{Z}\} \geq e^{-4\log(\varepsilon_Y)/\varepsilon_Y^2}, \\
&\mathbb{E}_*\{1-\mathrm{expit}(\widehat{\gamma}(\widehat{\lambda}_\gamma)^\top \mathbf{Z}_*)|\mathbf{Z}_* \in \mathcal{Z}\} \geq e^{-4\log(\varepsilon_Y)/\varepsilon_Y^2}.
\end{aligned}
$$

The bounds above are connected to $\mathcal{M}_\gamma\left(\widehat{\gamma}(\widehat{\lambda}_\gamma)\right)$ through

$$\mathbb{E}_*\{w_*^0(\widehat{\gamma}(\widehat{\lambda}_\gamma))X_*\} \geq \tau\varepsilon_Y^3/8\mathbb{E}_*\{\text{expit}(\widehat{\gamma}(\widehat{\lambda}_\gamma)^\top \mathbf{Z}_*)|\mathbf{Z}_* \in \mathcal{Z}\},$$

$$\mathbb{E}_*\{w_*^1(\widehat{\gamma}(\widehat{\lambda}_\gamma))Y_*(\tau)\} \geq \varepsilon_Y^3/8\mathbb{E}_*\{1-\text{expit}(\widehat{\gamma}(\widehat{\lambda}_\gamma)^\top \mathbf{Z}_*)|\mathbf{Z}_* \in \mathcal{Z}\}.$$

Therefore, we obtain the bound $\mathcal{M}_\gamma\left(\widehat{\gamma}(\widehat{\lambda}_\gamma)\right) \leq (1+\tau^{-1})8\varepsilon_Y^{-3}e^{-4\log(\varepsilon_Y)/\varepsilon_Y^2}$. □

Proof of Theorem 11. We prove the theorem under two setups given by Assumptions 7(a) and 7(b) separately. We denote the cross-fitted weighted Breslow estimator $\check{\Lambda}$ defined in (3.17) as

$$\check{\Lambda}^{(j)}(t,\theta;\boldsymbol{\beta},\boldsymbol{\gamma}) = \int_0^t \frac{\sum_{i\in I_j} w_i^1(\boldsymbol{\gamma})\{dN_i(u) - Y_i(u)(D_i\theta + \boldsymbol{\beta}^\top \mathbf{Z}_i)du\}}{\sum_{i\in I_j} w_i^1(\boldsymbol{\gamma})Y_i(u)}, \tag{3.89}$$

constructed with samples in fold-j. We denote the cross-fitted score associated with the closed form estimator $\check{\theta}_{cf}$ for fold-j as

$$\begin{aligned}
&\psi^{(j)}(\theta;\boldsymbol{\beta},\boldsymbol{\gamma}) \\
=&\phi^{(j)}\left(\theta;\boldsymbol{\beta},\check{\Lambda}^{(j)}(\cdot,\theta;\boldsymbol{\beta},\boldsymbol{\gamma}),\boldsymbol{\gamma}\right) \\
=&-\frac{1}{n}\sum_{i\in I_j} w_i^0(\widehat{\gamma}^{(j)})\int_0^\tau \left(dN_i(u) - Y_i(u)\left[\widehat{\boldsymbol{\beta}}^{(j)\top}\{\mathbf{Z}_i - \widetilde{\mathbf{Z}}^{(j)}(u;\widehat{\gamma}^{(j)})\}du + d\widetilde{N}^{(j)}(u;\widehat{\gamma}^{(j)})\right]\right) \\
&-\frac{\theta}{n}\sum_{i\in I_j}(1-D_i)\,\text{expit}(\boldsymbol{\gamma}^\top \mathbf{Z}_{\mathbf{1}i})X_i,
\end{aligned} \tag{3.90}$$

(a) First, we show that the true parameter is identified by the score. That is

$$\psi^{(j)}\left(\theta_0;\widehat{\boldsymbol{\beta}}^{(j)},\widehat{\boldsymbol{\gamma}}^{(j)}\right) = o_p(1). \tag{3.91}$$

We decompose

$$\psi^{(j)}\left(\theta_0;\widehat{\boldsymbol{\beta}}^{(j)},\widehat{\boldsymbol{\gamma}}^{(j)}\right)$$

$$
\begin{aligned}
&= -\frac{1}{n}\sum_{i\in I_j}\{D_i-\text{expit}(\widehat{\gamma}^{(j)\top}\mathbf{Z}_{\mathbf{1}i})\}\int_0^\tau e^{D_i\theta_0 t}Y_i(t)(\widehat{\boldsymbol{\beta}}^{(j)}-\boldsymbol{\beta}_0)^\top\mathbf{Z}_i dt \\
&\quad +\frac{1}{n}\sum_{i'\in I_j}\int_0^\tau \frac{\sum_{i\in I_j}\{D_i-\text{expit}(\widehat{\gamma}^{(j)\top}\mathbf{Z}_{\mathbf{1}i})\}e^{D_i\theta_0 t}Y_i(t)}{\sum_{i\in I_j}w_i^1(\widehat{\gamma}^{(j)})Y_i(t)}w_{i'}^1(\widehat{\gamma}^{(j)})Y_{i'}(t)(\widehat{\boldsymbol{\beta}}^{(j)}-\boldsymbol{\beta}_0)^\top\mathbf{Z}_{i'}dt \\
&\quad +\frac{1}{n}\sum_{i\in I_j}\{D_i-\text{expit}(\widehat{\gamma}^{(j)}\mathbf{Z}_{\mathbf{1}i})\}\int_0^\tau e^{D_i\theta_0 t}dM_i(t) \\
&\quad -\frac{1}{n}\sum_{i'\in I_j}\int_0^\tau \frac{\sum_{i\in I_j}\{D_i-\text{expit}(\widehat{\gamma}^{(j)\top}\mathbf{Z}_{\mathbf{1}i})\}e^{D_i\theta_0 t}Y_i(t)}{\sum_{i\in I_j}w_i^1(\widehat{\gamma}^{(j)})Y_i(t)}w_{i'}^1(\widehat{\gamma}^{(j)})dM_{i'}(t) \\
&= Q_1+Q_2+Q_3+Q_4. \qquad (3.92)
\end{aligned}
$$

We shall show that each term Q_1-Q_4 in (3.92) is negligible.

By applying twice the Cauchy-Schwartz inequality, first to the sum then to the integral, we have a bound for Q_1,

$$
|Q_1|\le\frac{1}{n}\sum_{i\in I_j}1e^{K_\theta\tau}\int_0^\tau Y_i(t)(\widehat{\boldsymbol{\beta}}^{(j)}-\boldsymbol{\beta}_0)^\top\mathbf{Z}_i dt\le\frac{1}{n}\sqrt{|I_j|e^{K_\theta\tau}}\sqrt{\sum_{i\in I_j}\{(\widehat{\boldsymbol{\beta}}^{(j)}-\boldsymbol{\beta}_0)^\top\mathbf{Z}_i\}^2X_i}. \qquad (3.93)
$$

Also from Assumption 7a-ii, the squared average model deviance $\mathbb{E}_*\{(\widehat{\boldsymbol{\beta}}^{(j)}-\boldsymbol{\beta}_0)^\top\mathbf{Z}_*\}^2X_*$ converges to zero. Applying the Markov inequality conditioning on the out-of-fold data, we have its asymptotic equivalence to the empirical counterpart

$$
\frac{1}{|I_j|}\sum_{i\in I_j}\{(\widehat{\boldsymbol{\beta}}^{(j)}-\boldsymbol{\beta}_0)^\top\mathbf{Z}_i\}^2X_i=\mathbb{E}_*\{(\widehat{\boldsymbol{\beta}}^{(j)}-\boldsymbol{\beta}_0)^\top\mathbf{Z}_*\}^2X_*+o_p(1)=o_p(1). \qquad (3.94)
$$

Plugging (3.94) to (3.93), we conclude that $Q_1=o_p(1)$.

Similarly for Q_2, we apply the Cauchy-Schwartz inequality twice,

$$
\begin{aligned}
|Q_2|\le&\frac{1}{n}\sqrt{\int_0^\tau\left[\frac{\sum_{i\in I_j}\{D_i-\text{expit}(\widehat{\gamma}^{(j)\top}\mathbf{Z}_{\mathbf{1}i})\}e^{D_i\theta_0 t}Y_i(t)}{\sum_{i\in I_j}w_i^1(\widehat{\gamma}^{(j)})Y_i(t)}\right]^2\sum_{i\in I_j}w_i^2(t;\theta_0,\widehat{\gamma}^{(j)})Y_i(t)dt} \\
&\times\sqrt{\sum_{i\in I_j}\{(\widehat{\boldsymbol{\beta}}^{(j)}-\boldsymbol{\beta}_0)^\top\mathbf{Z}_i\}^2X_i}. \qquad (3.95)
\end{aligned}
$$

From Assumption 7a-ii, we have a lower bound for $\mathbb{E}_*\{w_*^1(\widehat{\gamma}^{(j)})Y_*(\tau)\} \geq K_{\mathcal{M}_g}^{-1}$. Applying the Hoeffding's inequality to the empirical version of the process, we get $\frac{1}{|I_j|}\sum_{i\in I_j} w_i^1(\widehat{\gamma}^{(j)})Y_i(\tau) \geq K_{\mathcal{M}_g}^{-1}/2$ with probability tending to one. The denominator term in Q_2 is decreasing process in t thus achieves it minimal at $t=\tau$, so it has the lower bound $K_{\mathcal{M}_g}^{-1}/2$ with probability tending to one. Along with (3.94), we conclude from (3.95) with probability tending to one

$$Q_2 \leq 2e^{2K_\theta \tau}\tau K_{\mathcal{M}_g} O_p\left(\mathcal{D}_\beta\left(\widehat{\boldsymbol{\beta}}^{(j)}, \boldsymbol{\beta}_0\right)\right) = o_p(1). \tag{3.96}$$

Q_3 and Q_4 are martingale integrals with respect to filtration

$$\mathcal{F}_{I_j,t} = \sigma\left(\{(N_i(u), Y_i(u), D_i, \mathbf{Z}_i) : i \in I_j, u \leq t\} \cup \{(X_i, \delta_i, D_i, \mathbf{Z}_i) : i \in I_{-j}\}\right).$$

The integrands are bounded with probability tending to one under Assumption 7a-ii, so we obtain by Lemma 37 that $Q_3 = O_p\left(n^{-1/2}\right)$ and $Q_4 = O_p\left(n^{-1/2}\right)$.

We combine the results for $Q_1 - Q_4$ to establish the identifiability result (3.91).

By the Assumption 7a-ii, we have the denominator in $\breve{\theta}$ (3.33)

$$Q' = -\frac{1}{n}\sum_{j=1}^{k}\sum_{i\in I_j}(1-D_i)\operatorname{expit}(\widehat{\gamma}^{(j)\top}\mathbf{Z}_{\mathbf{1}j})X_i. \tag{3.97}$$

bounded from below by $2kK_{\mathcal{M}_g}^{-1}$ with probability tending to one.

Utilizing the linearity of ψ, we can write

$$(\breve{\theta} - \theta_0) = \frac{\frac{1}{n}\sum_{j=1}^{k}\psi^{(j)}(\theta_0; \widehat{\boldsymbol{\beta}}^{(j)}, \widehat{\gamma}^{(j)})}{Q'} = o_p(1). \tag{3.98}$$

We hence obtain the consistency of $\breve{\theta}$.

(b) Under model (3.31), we have for $i \in I_j$ the following martingale with respect to filtration

$$\mathcal{F}_{I_j,t} = \sigma\left(\{(N_i(u), Y_i(u), D_i, \mathbf{Z}_i) : i \in I_j, u \le t\} \cup \{(X_i, \delta_i, D_i, \mathbf{Z}_i) : i \in I_{-j}\}\right)$$

$$M_i(t) = N_i(t) - \int_0^\tau Y_i(t)\{D_i\theta_0 + g_0(t;\mathbf{Z}_i)\}dt. \tag{3.99}$$

First, we prove the identifiability result like (3.91). We decompose

$$\begin{aligned}
& \psi^{(j)}(\theta_0; \widehat{\boldsymbol{\beta}}^{(j)}, \widehat{\boldsymbol{\gamma}}^{(j)}) \\
= \; & \frac{1}{n}\sum_{i\in I_j}\{D_i - \text{expit}(\widehat{\boldsymbol{\gamma}}^{(j)\top}\mathbf{Z}_{\mathbf{1}i})\}\int_0^\tau e^{D_i\theta_0 t}dM_i(t) \\
& + \frac{1}{n}\sum_{i\in I_j}\int_0^\tau \{D_i - \text{expit}(\boldsymbol{\gamma}_0^\top\mathbf{Z}_{\mathbf{1}i})\}e^{D_i\theta_0 t}Y_i(t)g_0(t;\mathbf{Z}_i)dt \\
& - \int_0^\tau \frac{1}{n}\sum_{i\in I_j}\{D_i - \text{expit}(\boldsymbol{\gamma}_0^\top\mathbf{Z}_{\mathbf{1}i})\}e^{D_i\theta_0 t}Y_i(t)\widehat{\boldsymbol{\beta}}^{(j)\top}\{\mathbf{Z}_i - \mu(t)\}dt \\
& + \int_0^\tau \frac{1}{n}\sum_{i\in I_j}\{D_i - \text{expit}(\boldsymbol{\gamma}_0^\top\mathbf{Z}_{\mathbf{1}i})\}e^{D_i\theta_0 t}Y_i(t)\{\theta_0 dt - d\widetilde{N}^{(j)}(t;\widehat{\boldsymbol{\gamma}}^{(j)})\} \\
& + \int_0^\tau \frac{1}{n}\sum_{i\in I_j}\{\text{expit}(\boldsymbol{\gamma}_0^\top\mathbf{Z}_{\mathbf{1}i}) - \text{expit}(\widehat{\boldsymbol{\gamma}}^{(j)\top}\mathbf{Z}_{\mathbf{1}i})\}e^{D_i\theta_0 t}Y_i(t)g_0(t;\mathbf{Z}_i)dt \\
& - \int_0^\tau \frac{1}{n}\sum_{i\in I_j}\{\text{expit}(\boldsymbol{\gamma}_0^\top\mathbf{Z}_{\mathbf{1}i}) - \text{expit}(\widehat{\boldsymbol{\gamma}}^{(j)\top}\mathbf{Z}_{\mathbf{1}i})\}e^{D_i\theta_0 t}Y_i(t)\widehat{\boldsymbol{\beta}}^{(j)\top}\{\mathbf{Z}_i - \mu(t)\}dt \\
& + \int_0^\tau \frac{1}{n}\sum_{i\in I_j}\{\text{expit}(\boldsymbol{\gamma}_0^\top\mathbf{Z}_{\mathbf{1}i}) - \text{expit}(\widehat{\boldsymbol{\gamma}}^{(j)\top}\mathbf{Z}_{\mathbf{1}i})\}e^{D_i\theta_0 t}Y_i(t)\{\theta_0 dt - d\widetilde{N}^{(j)}(t;\widehat{\boldsymbol{\gamma}}^{(j)})\} \\
& - \int_0^\tau \frac{1}{n}\sum_{i\in I_j}\{D_i - \text{expit}(\boldsymbol{\gamma}_0^\top\mathbf{Z}_{\mathbf{1}i})\}e^{D_i\theta_0 t}Y_i(t)\widehat{\boldsymbol{\beta}}^{(j)\top}\{\mu(t) - \widetilde{\mathbf{Z}}^{(j)}(t;\widehat{\boldsymbol{\gamma}}^{(j)})\}dt \\
& - \int_0^\tau \frac{1}{n}\sum_{i\in I_j}\{\text{expit}(\boldsymbol{\gamma}_0^\top\mathbf{Z}_{\mathbf{1}i}) - \text{expit}(\widehat{\boldsymbol{\gamma}}^{(j)\top}\mathbf{Z}_{\mathbf{1}i})\}e^{D_i\theta_0 t}Y_i(t)\widehat{\boldsymbol{\beta}}^{(j)\top}\{\mu(t) - \widetilde{\mathbf{Z}}^{(j)}(t;\widehat{\boldsymbol{\gamma}}^{(j)})\}dt \\
= \; & Q_1 + Q_2 + Q_3 + Q_4 + Q_5 + Q_6 + Q_7 + Q_8 + Q_9.
\end{aligned} \tag{3.100}$$

Q_1 is the final element of the $\mathcal{F}_{I_j,t}$-martingale,

$$Q_{1,t} = \frac{1}{n}\sum_{i\in I_j}\{D_i - \text{expit}(\widehat{\boldsymbol{\gamma}}^{(j)\top}\mathbf{Z}_{\mathbf{1}i})\}\int_0^t e^{D_i\theta_0 u}dM_i(u). \tag{3.101}$$

The measurable quadratic variation of $Q_{1,t}$ is

$$\langle Q_{1,\cdot}\rangle_t = \frac{1}{n^2}\sum_{i\in I_j}\{D_i - \text{expit}(\widehat{\gamma}^{(j)\top}\mathbf{Z}_{\mathbf{1}i})\}^2\int_0^t e^{2D_i\theta_0 u}Y_i(u)g_0(t;\mathbf{Z}_i)du. \tag{3.102}$$

By the Cauchy-Schwartz's inequality, we have the upper bound for

$$\text{Var}(Q_1) = \mathbb{E}\langle Q_{1,\cdot}\rangle_\tau \leq \mathbb{E}\left\{\frac{1}{n^2}\sqrt{ne^{2K_\theta\tau}\tau}\sqrt{\sum_{i\in I_j}\int_0^\tau g_0(t;\mathbf{Z}_i)dt}\right\}.$$

Under Assumption 7(b), we have

$$\sqrt{\sum_{i\in I_j}\int_0^\tau g_0(t;\mathbf{Z}_i)dt} = O_p(nK_\Lambda)$$

by Markov's inequality. Thus, we have $\text{Var}(Q_1) = O(K_\Lambda/n) = o(1)$. By the Tchebychev's inequality, we obtain $Q_1 = o_p(1)$.

Using Lemma 41, we have for Q_2

$$\begin{aligned}
&\mathbb{E}[\{D_i - \text{expit}(\boldsymbol{\gamma}_0^\top\mathbf{Z}_{\mathbf{1}i})\}e^{D_i\theta_0 t}Y_i(t)g_0(t;\mathbf{Z}_i)]\\
=&\mathbb{E}(\mathbb{E}[\{D_i - \text{expit}(\boldsymbol{\gamma}_0^\top\mathbf{Z}_{\mathbf{1}i})\}e^{D_i\theta_0 t}Y_i(t)|\mathbf{Z}_i]g_0(t;\mathbf{Z}_i))\\
=&0.
\end{aligned}$$

The variance of Q_2 has bound

$$\begin{aligned}
\text{Var}(Q_2) =& \frac{1}{n}\mathbb{E}\left(\left[\int_0^\tau\{D_i - \text{expit}(\boldsymbol{\gamma}_0^\top\mathbf{Z}_{\mathbf{1}i})\}e^{D_i\theta_0 t}Y_i(t)g_0(t;\mathbf{Z}_i)dt\right]^2\right)\\
\leq& \frac{1}{n}e^{2K_\theta\tau}\mathbb{E}\left[\int_0^\tau Y_i(t)g_0^2(t;\mathbf{Z}_i)dt\right].
\end{aligned}$$

Under Assumption 7(b), we have $\text{Var}(Q_2) = O(K_\lambda/n) = o(1)$. By the Tchebychev's inequality, we obtain $Q_2 = o_p(1)$.

Similarly for Q_3, we obtain from Lemma 41 that $\mathbb{E}(Q_3) = 0$. Using the above fact, we give a bound for the variance of Q_3,

$$\begin{aligned}\text{Var}(Q_3) \leq & \frac{1}{n}\mathbb{E}\left(\left[\int_0^\tau \{D_i - \text{expit}(\gamma_0^\top \mathbf{Z}_{\mathbf{1}i})\} e^{D_i \theta_0 t} Y_i(t) \widehat{\boldsymbol{\beta}}^{(j)\top} \{\mathbf{Z}_i - \mu(t)\} dt\right]^2\right) \\ \leq & \frac{1}{n} e^{2K_\theta \tau} \mathbb{E}\left(\int_0^\tau \left[\widehat{\boldsymbol{\beta}}^{(j)\top} \{\mathbf{Z}_i - \mu(t)\}\right]^2 Y_i(t) dt\right). \end{aligned} \tag{3.103}$$

Under Assumption 7(b), we have $\text{Var}(Q_3) = O\left(\left\{\mathcal{M}_\beta\left(\widehat{\boldsymbol{\beta}}^{(j)}\right)\right\}^2 / n\right) = o(1)$. By the Tchebychev's inequality, we obtain $Q_3 = o_p(1)$.

For Q_4, we also have from Lemma 41

$$\sup_{t\in[0,\tau]} \left| \frac{1}{n} \sum_{i \in I_j} \{D_i - \text{expit}(\gamma_0^\top \mathbf{Z}_{\mathbf{1}i})\} e^{D_i \theta_0 t} Y_i(t) \right| = O_p\left(n^{-\frac{1}{2}}\right).$$

Again using the Cauchy-Schwartz inequality, we bound the total variation of the measure in Q_4,

$$\begin{aligned} & \int_0^\tau [\{\theta_0 + \widehat{\boldsymbol{\beta}}^{(j)\top} \widetilde{\mathbf{Z}}^{(j)}(t; \widehat{\gamma}^{(j)})\} dt + d\widetilde{N}^{(j)}(t; \widehat{\gamma}^{(j)})] dt \\ \leq & K_\theta \tau + 1 + \sqrt{\int_0^\tau \frac{n}{\left\{\sum_{i\in I_j} w_i^1(\widehat{\gamma}^{(j)}) Y_i(t)\right\}^2} dt} \sqrt{e^{2K_\theta \tau} \sum_{i \in I_j} X_i \left(\widehat{\boldsymbol{\beta}}^{(j)\top} \mathbf{Z}_i\right)^2}. \end{aligned}$$

Using Lemma 42 and the Markov inequality, we have the bound above is of order $O_p\left(\|\widehat{\boldsymbol{\beta}}^{(j)}\|_{I_j}\right)$. Therefore, we obtain under Assumption 7b-iii $Q_4 = O_p\left(\|\widehat{\boldsymbol{\beta}}^{(j)}\|_{I_j} n^{-\frac{1}{2}}\right) = o_p(1)$.

For terms Q_5, we use the Cauchy-Schwartz inequality

$$|Q_5| \leq \frac{1}{n} \sqrt{\tau \sum_{i \in I_j} \{\text{expit}(\gamma_0^\top \mathbf{Z}_{\mathbf{1}i}) - \text{expit}(\widehat{\gamma}^{(j)\top} \mathbf{Z}_{\mathbf{1}i})\}^2} \sqrt{e^{2K_\theta \tau} \sum_{i \in I_j} \int_0^\tau g_0^2(t; \mathbf{Z}_i) Y_i(t) dt}.$$

We apply the Markov's inequality under Assumptions 7(b) and 7b-iii to get

$$Q_5 = O_p\left(\mathcal{D}_\gamma\left(\widehat{\gamma}^{(j)}, \gamma_0\right) K_\Lambda\right) = o_p(1).$$

For terms Q_6, we use the Cauchy-Schwartz inequality

$$|Q_6| \leq \sqrt{\frac{\tau}{n}\sum_{i\in I_j}\{\text{expit}(\gamma_0^\top \mathbf{Z}_{\mathbf{1}i}) - \text{expit}(\widehat{\gamma}^{(j)\top}\mathbf{Z}_{\mathbf{1}i})\}^2 \frac{e^{2K_\theta \tau}}{n}\sum_{i\in I_j}\int_0^\tau [\widehat{\boldsymbol{\beta}}^{(j)\top}\{\mathbf{Z}_i - \mu(t)\}]^2 Y_i(t)dt}$$

We apply the Markov's inequality under Assumption 7b-iii to get

$$Q_6 = O_p\left(\mathcal{D}_\gamma\left(\widehat{\gamma}^{(j)}, \gamma_0\right) \mathcal{M}_\beta\left(\widehat{\boldsymbol{\beta}}^{(j)}\right)\right) = o_p(1).$$

In term Q_7, we establish a uniform bound

$$\begin{aligned}
&\sup_{t\in[0,\tau]}\left|\frac{1}{n}\sum_{i\in I_j}\{\text{expit}(\gamma_0^\top \mathbf{Z}_{\mathbf{1}i}) - \text{expit}(\widehat{\gamma}^{(j)\top}\mathbf{Z}_{\mathbf{1}i})\} e^{D_i\theta_0 t} Y_i(t)\right| \\
\leq &\frac{1}{n}\sqrt{\sum_{i\in I_j}\{\text{expit}(\gamma_0^\top \mathbf{Z}_{\mathbf{1}i}) - \text{expit}(\widehat{\gamma}^{(j)\top}\mathbf{Z}_{\mathbf{1}i})\}^2}\sqrt{|I_j| e^{2K_\theta\tau}}
\end{aligned}$$

by the Cauchy-Schwartz inequality. Hence, the process above has a bound of order $O_p\left(\mathcal{D}_\gamma\left(\widehat{\gamma}^{(j)}, \gamma_0\right)\right)$ uniformly in $t \in [0,\tau]$. We have the same upper bound for the total variation of the measure as that in Q_4, $O_p\left(\mathcal{M}_\beta\left(\widehat{\boldsymbol{\beta}}^{(j)}\right)\right)$. Thus, we establish the order $Q_7 = O_p\left(\mathcal{D}_\gamma\left(\widehat{\gamma}^{(j)}, \gamma_0\right) \mathcal{M}_\beta\left(\widehat{\boldsymbol{\beta}}^{(j)}\right)\right) = o_p(1)$.

For terms Q_8 and Q_9, we use the Cauchy-Schwartz inequality to bound the discrepancy between $\mu(t)$ in $\mathcal{M}_\beta\left(\widehat{\boldsymbol{\beta}}^{(j)}\right)$ and the empirical $\widetilde{\mathbf{Z}}^{(j)}(t, \widehat{\boldsymbol{\beta}}^{(j)})$,

$$\begin{aligned}
|\widehat{\boldsymbol{\beta}}^{(j)\top}\{\mu(t) - \widetilde{\mathbf{Z}}^{(j)}(t, \widehat{\gamma}^{(j)})\}|^2 &= \left|\sum_{i\in I_j}\frac{w_i^0(\widehat{\gamma}^{(j)})Y_i(t)}{\sum_{i'\in I_j} w_{i'}^0(\widehat{\gamma}^{(j)})Y_{i'}(t)}\widehat{\boldsymbol{\beta}}^{(j)\top}\{\mu(t) - \mathbf{Z}_i\}\right|^2 \\
&\leq \frac{\sum_{i'\in I_j}\{w_{i'}^0(\widehat{\gamma}^{(j)})Y_{i'}(t)\}^2}{\left\{\sum_{i'\in I_j} w_{i'}^0(\widehat{\gamma}^{(j)})Y_{i'}(t)\right\}^2}\sum_{i\in I_j}\left[\widehat{\boldsymbol{\beta}}^{(j)\top}\{\mathbf{Z}_i - \mu(t)\}\right]^2 Y_i(t)
\end{aligned}$$

$$\leq \frac{\sum_{i\in I_j}\left[\widehat{\boldsymbol{\beta}}^{(j)\top}\{\mathbf{Z}_i-\mu(t)\}\right]^2 Y_i(t)}{\sum_{i'\in I_j} w^0_{i'}(\widehat{\boldsymbol{\gamma}}^{(j)})Y_{i'}(t)}. \tag{3.104}$$

The last step above comes from the fact that $w^0_{i'}(\widehat{\boldsymbol{\gamma}}^{(j)})Y_{i'}(t)\in[0,1]$. Under Assumption (7b-iii), we obtain

$$\int_0^\tau |\widehat{\boldsymbol{\beta}}^{(j)\top}\{\mu(t)-\widetilde{\mathbf{Z}}^{(j)}(t,\widehat{\boldsymbol{\gamma}}^{(j)})\}|^2 dt = O_p\left(\mathcal{M}_\beta(\widehat{\boldsymbol{\beta}}^{(j)})\right) = o_p\left(\mathcal{D}_\gamma(\widehat{\boldsymbol{\gamma}}^{(j)})^{-1}\right).$$

Therefore, we follow the strategy of Q_3 and Q_6 to get $Q_8 = o_p(1)$ and $Q_9 = o_p(1)$.

Combining the results for Q_1-Q_9, we establish that $\psi^{(j)}(\theta_0;\widehat{\boldsymbol{\beta}}^{(j)},\widehat{\boldsymbol{\gamma}}^{(j)}) = o_p(1)$.

By the Lemma 42, we have the denominator in $\breve{\theta}$ (3.33)

$$Q' = -\frac{1}{n}\sum_{j=1}^{k}\sum_{i\in I_j}(1-D_i)\,\mathrm{expit}(\widehat{\boldsymbol{\gamma}}^{(j)\top}\mathbf{Z}_{\mathbf{1}j})X_i \tag{3.105}$$

is bounded from below by $k\varepsilon_Y/2$.

Along with the identifiability of θ_0 by ψ, we obtain the consistency for $\breve{\theta}$.

□

3.7.3 Preliminary Results

Lemma 29. *Under the Assumption 5, we have for* θ *in a compact neighborhood of* θ_0 *such that* $|\theta|\leq K_\theta$

$$\begin{aligned}
&\sqrt{n}\phi\left(\theta;\widehat{\boldsymbol{\beta}},\widehat{\Lambda}(\cdot,\theta),\widehat{\boldsymbol{\gamma}}\right)\\
=&\sqrt{n}\phi(\theta_0;\boldsymbol{\beta}_0,\Lambda_0,\boldsymbol{\gamma}_0) - \frac{1}{\sqrt{n}}(\theta-\theta_0)\sum_{i=1}^{n} D_i\{1-\mathrm{expit}(\boldsymbol{\gamma}_0^\top\mathbf{Z}_{\mathbf{1}i})\}(e^{\theta_0 X_i}-1)/\theta_0\\
&+o_p(1+\sqrt{n}|\theta-\theta_0|)+O_p(\sqrt{n}|\theta-\theta_0|^2)+O_p(\sqrt{n}|\theta-\theta_0|^3).
\end{aligned} \tag{3.106}$$

Lemma 30. *Suppose the $|I_j| \asymp n$. Under the Assumption 6, we have for θ in a compact neighborhood of θ_0 such that $|\theta| \leq K_\theta$*

$$\begin{aligned}
&\sqrt{n}\phi^{(j)}\left(\theta; \widehat{\beta}^{(j)}, \widehat{\Lambda}^{(j)}(\cdot,\theta), \widehat{\gamma}^{(j)}\right) \\
=&\sqrt{n}\phi^{(j)}(\theta_0;\beta_0,\Lambda_0,\gamma_0) - \frac{1}{\sqrt{n}}(\theta-\theta_0)\sum_{i\in I_j} D_i\{1-\text{expit}(\gamma_0^\top \mathbf{Z}_{\mathbf{1}i})\}(e^{\theta_0 X_i}-1)/\theta_0 \\
&+ o_p(1+\sqrt{n}|\theta-\theta_0|) + O_p(\sqrt{n}|\theta-\theta_0|^2) + O_p(\sqrt{n}|\theta-\theta_0|^3). \qquad (3.106)
\end{aligned}$$

3.7.4 Classical Concentration Inequalities

Lemma 31. ***Hoeffding's Inequality*** *Theorem 2 p.4 in [Hoe63]. If $X_1,\ldots,X_n$ are independent and $a_i \leq X_i \leq b_i$ $(i=1,2,\ldots,n)$, then for $t>0$*

$$\Pr(\bar{X}-\mu \geq t) \leq \exp\left(-\frac{2n^2t^2}{\sum_{i=1}^n (b_i-a_i)^2}\right).$$

Lemma 32. ***A version of Azuma's Inequality*** *Theorem 1 p.3 and Remark 7 p.5 in [Sas13]. Let $\{X_k, \mathcal{F}_k\}_k = 0^\infty$ be a discrete-parameter real-valued martingale sequence such that for every k, the condition $|X_k - X_{k-1}| \leq a_k$ holds almost surely for some non-negative constants $\{a_k\}_{k=1}^\infty$. Then*

$$\Pr\left(\max_{k\in 1,\ldots,n} |X_k - X_0| \geq t\right) \leq 2\exp\left(-\frac{t^2}{2\sum_{k=1}^n a_k^2}\right)$$

Lemma 33. ***Bernstein Inequality for Sub-exponential Random Variables***

a) For i.i.d. sample as in Chapter 2 Section 1.3 of [Wai19]:

Let X be a random variable with mean $\mathbb{E}(X)=\mu$. If X satisfies the Bernstein's condition with parameter b, i.e.

$$\left|\mathbb{E}\left\{(X-\mu)^k\right\}\right| \leq \frac{1}{2}k!b^k, \text{ for } k=2,3,\ldots,$$

the following concentration inequality holds for an i.i.d. sample $X_1, \ldots, X_n$

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| \geq t\right) \leq 2\exp\left\{-\frac{nt^2}{2(b^2+bt)}\right\}.$$

b) For martingale as in Chapter 2 Section 2.2 of [Wai19]: Let $M_1, \ldots, M_n$ *be a martingale series with respect to filtration* $\mathcal{F}_1 \subset \cdots \subset \mathcal{F}_n$*. If the martingale differences satisfies the Bernstein's condition with parameter b, i.e.*

$$\left|\mathbb{E}\left\{(M_{j+1} - M_j)^k | \mathcal{F}_j\right\}\right| \leq \frac{1}{2}k!b^k, \text{ for } j = 1, \ldots, n-1 \text{ and } k = 2, 3, \ldots,$$

the following concentration inequality holds

$$\mathbb{P}\left(\sup_{j=1,\ldots,n} |M_j| \geq t\right) \leq 2\exp\left\{-\frac{t^2}{2(nb^2+bt)}\right\}.$$

Lemma 34. ***Dvoretzky-Kiefer-Wolfowitz (DKW) Inequality*** *[DKW56, Mas90] Let* $X_1, \ldots, X_n$ *be i.i.d. samples from a distribution with c.d.f.* $F(x)$*. Define the empirical c.d.f. as* $F_n(x) = n^{-1}\sum_{i=1}^{n} I(X_i \leq x)$*. For any* $\varepsilon > 0$*,*

$$\Pr\left(\sup_{x\in\mathbb{R}} |F_n(x) - F(x)|\right) \leq 2e^{-2n\varepsilon^2}.$$

3.7.5 New Concentration Results

All the concentration results are adapted to the cross-fitting scheme. We repeated use the following two notations for index set and index set specific filtration.

Definition 1. We denote $I \subset \{1, \ldots, n\}$ be a index set independent of observed data $\{W_i, i = 1, \ldots, n\}$ whose cardinality satisfies $|I| \asymp n$.

Definition 2. We define the filtration for index set I as

$$\mathcal{F}_{I,t} = \sigma\left(\{N_i(u), Y_i(u+), D_i, Z_i : u \le t, i \in I\} \cup \{\delta_i, X_i, D_i, Z_i : i \in I^c\}\right).$$

Remark 17. *The difference between $\mathcal{F}_{I,t}$ with $Y_i(u+)$ and the usual filtration defined with $Y_i(u)$ is that the former contains information about independent out of fold samples and the censoring times at present time t so that the observed censoring times are stopping times with respect to $\mathcal{F}_{I,t}$. On the other hand, we still have the martingale property*

$$\mathbb{E}\{M_i(t)|\mathcal{F}_{I,t-}\} = \mathbb{E}\{M_i(t)|\mathcal{F}^*_{I,t-}\} = M_i(t-) \tag{3.107}$$

because the extra censoring information at t is not in $\mathcal{F}_{I,t-}$, and out of fold samples are independent of $M_i(t)$ for $i \in I$.

Lemma 35. *Define the filtration $F_t^{(i)} = \sigma\left(\{N_i(u), Y_i(u), D_i, \mathbf{Z}_i : u \le t\}\right)$. Let $H_i(t)$ be a $F_\tau^{(i)}$-measurable random process, satisfying $\mathbb{P}(\sup_{t\in[0,\tau]} |H_i(t)| < K_H) = 1$. Under the model* (3.1) *and the Assumption 5-iv,*

$$\mathbb{P}\left(\int_0^\tau H_i(t)Y_i(t)\beta_0^\top \mathbf{Z}_i dt > x\right) \tag{3.108}$$

Moreover, we have

$$\left|\int_0^\tau \mathbb{E}\{H_i(t)Y_i(t)\beta_0^\top \mathbf{Z}_i\}dt\right| < 2K_H^2(K_\Lambda + \theta_0 \vee 0)\tau + 4K_H \tag{3.109}$$

and the concentration result for all $\varepsilon \in \left[0, \sqrt{2}\right]$ and index set I defined as in Definition 1

$$\mathbb{P}\left(\left|\frac{1}{|I|}\sum_{i\in I}\int_0^\tau H_i(t)Y_i(t)\beta_0^\top \mathbf{Z}_i dt - \int_0^\tau \mathbb{E}\{H_i(t)Y_i(t)\beta_0^\top \mathbf{Z}_i\}dt\right| > K\varepsilon\right) < 4e^{-|I|\varepsilon^2/2}, \tag{3.110}$$

where $K = 2K_H(K_\Lambda + \theta_0 \vee 0)\tau + 2|\mu| + 4K_H$.

Lemma 36. *For an index set I defined as in Definition 1, we define the filtration $\mathcal{F}_{I,t}$ as in Definition 2. Let $M_i(t)$ be the martingale* (3.3) *under model* (3.1) *and $H_i(t)$ be a nonnegative $\mathcal{F}_{I,t}$-measurable random processes, satisfying $\mathbb{P}(\sup_{t\in[0,\tau]}|H_i(t)|<K_H)=1$. Denote the order statistics of observed times as $X_{(1)},\ldots,X_{(|I|)}$. Then,*

$$M_k^H=\frac{1}{|I|}\sum_{i\in I}\int_0^{X_{(k)}}H_i(t)dM_i(t),\,k=0,\ldots,|I| \tag{3.111}$$

is a martingale with respect to $\mathcal{F}_{I,t}$, and we have for $j\geq 2$

$$\left|\mathbb{E}\left\{(M_k^H-M_{k-1}^H)^j|\mathcal{F}_{k-1}^H\right\}\right|\leq j!(2K_H/|I|)^j. \tag{3.112}$$

Besides, for every $\varepsilon>K_H/\sqrt{|I|}$ we have

$$\begin{aligned}&|I|\sum_{i\in I}\mathbb{E}\left\{(M_k^H-M_{k-1}^H)^2;\sqrt{|I|}|M_k^H-M_{k-1}^H|>\varepsilon\right\}\\<&(\varepsilon^2|I|+2K_H\sqrt{|I|}+2K_H^2)e^{-\varepsilon\sqrt{|I|}/K_H}.\end{aligned} \tag{3.113}$$

Lemma 37. *For an index set I defined as in Definition 1, we define the filtration $\mathcal{F}_{I,t}$ as in Definition 2. Let $M_i(t)$ be the martingale* (3.3) *under model* (3.1) *and $H_i(t)$ be a $\mathcal{F}_{I,t}$-measurable random processes, satisfying $\mathbb{P}(\sup_{t\in[0,\tau]}|H_i(t)|<K_H)=1$. Denote $X_{(1)},\ldots,X_{(|I|)}$ be the order statistics of observed times. Under Assumption 5-iv, for any $\varepsilon<1$,*

$$\mathbb{P}\left(\left|\frac{1}{|I|}\sum_{i\in I}\int_0^{\tau}H_i(t)dM_i(t)\right|<8K_H\varepsilon\right)>1-4e^{-|I|\varepsilon^2/2}. \tag{3.114}$$

Moreover, we also have

$$\bigvee_{t=0}^{\tau}\left\{\frac{1}{|I|}\sum_{i\in I}\int_0^t H_i(u)dM_i(u)\right\}\leq\frac{1}{|I|}\sum_{i\in I}\bigvee_{t=0}^{\tau}\int_0^t H_i(u)dM_i(u)<4K_H+8K_H\varepsilon \tag{3.115}$$

where $\bigvee_{t=0}^{\tau} f(t)$ *is the total variation of function* $f(t)$ *over* $[0,\tau]$*, and*

$$\sup_{t\in[0,\tau]}\left|\frac{1}{|I|}\sum_{i\in I}\int_0^t H_i(u)dM_i(u)\right| < 8K_H\varepsilon + 2K_H/|I| \tag{3.116}$$

whenever the event in (3.114) *occurs.*

Lemma 38. *For an index set I defined as in Definition 1, we define the filtration* $\mathcal{F}_{I,t}$ *as in Definition 2. Let* $M_i(t)$ *be the martingale* (3.3) *under model* (3.1) *and* $H_i(t)$ *be a* $\mathcal{F}_{I,t}$*-measurable random processes with tight supremum norm* $\max_{i=1,\ldots,n}\sup_{t\in[0,\tau]}|H_i(t)| = O_p(1)$*. Under Assumption 5-iv, for any* $\varepsilon < 1$*,*

$$\left|\frac{1}{|I|}\sum_{i\in I}\int_0^{\tau} H_i(t)dM_i(t)\right| = O_p\left(n^{-\frac{1}{2}}\right). \tag{3.117}$$

Lemma 39. *Let* H_i *be a random variable, satisfying* $\mathbb{P}\left(\sup_{i=1,\ldots,n}|H_i| \le K_H\right) = 1$*. For an index set I defined as in Definition 1, we have the concentration result*

$$\mathbb{P}\left(\sup_{t\in[0,\tau]}\left|\frac{1}{|I|}\sum_{i\in I}H_iY_i(t) - \mathbb{E}\{H_iY_i(t)\}\right| > 5K_H\varepsilon\right) < 8e^{-|I|\varepsilon^2/2}. \tag{3.118}$$

Lemma 40. *For an index set I defined as in Definition 1, we define the filtration* $\mathcal{F}_{I,t}$ *as in Definition 2. Let* $M_i(t)$ *be the martingale* (3.3) *under model* (3.1) *and* $H_i(t)$ *be* $\mathcal{F}_{I,t}$*-measurable random processes, satisfying* $\mathbb{P}(\sup_{t\in[0,\tau]}|H_i(t)| < K_H) = 1$*. Let* $\mathcal{H}$ *be a set of functions, potentially not* $\mathcal{F}_{I,t}$*-measurable, but having a finite bound* $\mathbb{P}\left(\sup_{\widetilde{H}\in\mathcal{H}}\sup_{t\in[0,\tau]}|\widetilde{H}(t)| < K_V\right) = 1$ *and a finite total variation* $\mathbb{P}\left(\sup_{\widetilde{H}\in\mathcal{H}}\bigvee_0^{\tau}\widetilde{H}(t) < K_V\right) = 1$*, where* $\bigvee_0^{\tau}$ *is the total variation on* $[0,\tau]$*. Under Assumptions 5-iv and 5-v,*

$$\mathbb{P}\left(\sup_{\widetilde{H}\in\mathcal{H}}\left|\frac{1}{|I|}\sum_{i\in I}\int_0^{\tau}\widetilde{H}(t)H_i(t)dM_i(t)\right| > 16K_HK_V\varepsilon + 2K_HK_V/|I|\right) < 4e^{-|I|\varepsilon^2/2}. \tag{3.119}$$

3.7.6 Other Auxiliary Results

Lemma 41. *Under Assumption* (5-ii) *and models* (3.1)*, or more general partially linear additive risks model* (3.31)*, we have*

$$\mathbb{E}[e^{D_i\theta_0 t}Y_i(t)|D_i,\mathbf{Z}_i] = \mathbb{E}\{Y_i(t)|\mathbf{Z}_i, D_i = 0\}. \tag{3.120}$$

Under model (3.2)*,*

$$\mathbb{E}[\{D_i - \text{expit}(\boldsymbol{\gamma}_0^\top \mathbf{Z}_i)\}e^{D_i\theta_0 t}Y_i(t)] = 0 \text{ and } \mathbb{E}[\{D_i - \text{expit}(\boldsymbol{\gamma}_0^\top \mathbf{Z}_i)\}e^{D_i\theta_0 t}Y_i(t)\mathbf{Z}_i] = \mathbf{0}. \tag{3.121}$$

Moreover, we have for index set I defined as in Definition 1 under Assumption 5-iii,

$$\begin{aligned}
&\sup_{t\in[0,\tau]}\left|\frac{1}{|I|}\sum_{i\in I}\{D_i - \text{expit}(\boldsymbol{\gamma}_0^\top \mathbf{Z}_i)\}e^{D_i\theta_0 t}Y_i(t)\right| = O_p\left(n^{-\frac{1}{2}}\right) \text{ and} \\
&\sup_{t\in[0,\tau]}\left\|\frac{1}{|I|}\sum_{i\in I}\{D_i - \text{expit}(\boldsymbol{\gamma}_0^\top \mathbf{Z}_i)\}e^{D_i\theta_0 t}Y_i(t)\mathbf{Z}_i\right\| = O_p\left(\frac{\log(p)}{\sqrt{n}}\right).
\end{aligned} \tag{3.122}$$

Lemma 42. *Suppose model* (3.2) *is correct, and* $\widehat{\boldsymbol{\gamma}}$ *is consistent for* $\boldsymbol{\gamma}_0$*, i.e.* $\mathcal{D}_\gamma(\widehat{\boldsymbol{\gamma}}, \boldsymbol{\gamma}_0) = o_p(1)$*.*

For an index set I defined as in Definition 1, we have under Assumption 5-v

$$\lim_{n\to\infty}\mathbb{P}\left(\inf_{t\in[0,\tau]}\frac{1}{|I|}\sum_{i\in I}w_i^1(\widehat{\boldsymbol{\gamma}})Y_i(t) > e^{-K_\theta\tau}\varepsilon_Y/2\right) = 1 \tag{3.123}$$

$$\text{and } \lim_{n\to\infty}\mathbb{P}\left(\inf_{t\in[0,\tau]}\frac{1}{|I|}\sum_{i\in I}(1-D_i)\,\text{expit}(\widehat{\boldsymbol{\gamma}}^\top \mathbf{Z}_{\mathbf{1}i})Y_i(t) > \varepsilon_Y/2\right) = 1. \tag{3.124}$$

3.7.7 Proofs of the Auxiliary Results

Definition 3. By the Mean Value Theorem for $e^{\theta t} - e^{\theta_0 t}$, we have

$$e^{\theta t} - e^{\theta_0 t} = (\theta - \theta_0)te^{\theta_t t}, \text{ for } \theta_t = \xi_t\theta_0 + (1-\xi_t)\theta \text{ with } \xi_t \in [0,1]. \tag{3.125}$$

In a bounded set of θ such that $|\theta| < K_\theta$, we have the bound $\sup_{t\in[0,\tau]} e^{\theta_t t} \le e^{K_\theta \tau}$. Since θ_t depends on θ, potentially estimated with all information from the data, the process $e^{\theta_t t}$ is not necessarily $\mathcal{F}_{I_j,t}$-adapted, causing extra complication in our proof.

Proof of Lemma 29. We define the filtration as

$$\mathcal{F}_{n,t} = \sigma\left(\{N_i(u), Y_i(u+), D_i, Z_i : u \le t, i = 1,\ldots,n\}\right),$$

using $I = \{1,\ldots,n\}$ in Definition 2.

We prove the statement (3.106) by investigating each terms in the following expansion,

$$\begin{aligned}
& \sqrt{n}\phi(\theta; \widehat{\beta}, \widehat{\Lambda}(\cdot,\theta), \widehat{\gamma}) \\
= \;& \sqrt{n}\phi(\theta; \beta_0, \Lambda_0, \gamma_0) \\
& -n^{-\frac{1}{2}} \sum_{i=1}^{n} \{D_i - \text{expit}(\gamma_0^\top \mathbf{Z}_{1,i})\} \int_0^\tau e^{D_i\theta t} Y_i(t) (\widehat{\beta} - \beta_0)^\top \mathbf{Z}_i dt \\
& -n^{-\frac{1}{2}} \sum_{i=1}^{n} \{D_i - \text{expit}(\gamma_0^\top \mathbf{Z}_{1,i})\} \int_0^\tau e^{D_i\theta t} Y_i(t) \left\{ d\widehat{\Lambda}(t,\theta) - d\Lambda_0(t) \right\} \\
& -n^{-\frac{1}{2}} \sum_{i=1}^{n} \{\text{expit}(\widehat{\gamma}^\top \mathbf{Z}_{1,i}) - \text{expit}(\gamma_0^\top \mathbf{Z}_{1,i})\} \int_0^\tau e^{D_i\theta t} dM_i(t; \theta, \beta_0, \Lambda_0) \\
& +n^{-\frac{1}{2}} (\widehat{\beta} - \beta_0)^\top \sum_{i=1}^{n} \{\text{expit}(\widehat{\gamma}^\top \mathbf{Z}_{1,i}) - \text{expit}(\gamma_0^\top \mathbf{Z}_{1,i})\} \int_0^\tau e^{D_i\theta t} Y_i(t) \mathbf{Z}_i dt \\
& +n^{-\frac{1}{2}} \sum_{i=1}^{n} \{\text{expit}(\widehat{\gamma}^\top \mathbf{Z}_{1,i}) - \text{expit}(\gamma_0^\top \mathbf{Z}_{1,i})\} \int_0^\tau e^{D_i\theta t} Y_i(t) \left\{ d\widehat{\Lambda}(t,\theta) - d\Lambda_0(t) \right\} \\
= \;& Q_1 + Q_2 + Q_3 + Q_4 + Q_5 + Q_6. \qquad (3.126)
\end{aligned}$$

The first term Q_1 contains the leading terms. The rest $Q_2 - Q_6$ are the remainders.

We expand Q_1 with respect to θ at θ_0,

$$Q_1 \;=\; \sqrt{n}\phi(\theta_0; \beta_0, \Lambda_0, \gamma_0)$$

$$
\begin{aligned}
& -n^{-\frac{1}{2}}(\theta-\theta_0)\sum_{i=1}^{n}\{D_i-\text{expit}(\gamma_0^\top \mathbf{Z}_{1,i})\}\int_0^\tau e^{\theta_0 t}D_i Y_i(t)dt \\
& +\frac{1}{\sqrt{n}}(\theta-\theta_0)\sum_{i=1}^{n}\{D_i-\text{expit}(\gamma_0^\top \mathbf{Z}_{1,i})\}\int_0^\tau e^{\theta_0 t}D_i t dM_i(t) \\
& +\frac{1}{\sqrt{n}}(\theta-\theta_0)^2\sum_{i=1}^{n}\{D_i-\text{expit}(\gamma_0^\top \mathbf{Z}_{1,i})\}\int_0^\tau e^{\theta_t t}D_i\{t^2 dM_i(t)+tY_i(t)dt\} \\
= \;& Q_{1,1}+Q_{1,2}+Q_{1,3}+Q_{1,4}, \qquad (3.127)
\end{aligned}
$$

where $Q_{1,4}$ comes from the Mean Value Theorem for $e^{\theta t}-e^{\theta_0 t}$ (3.125). $Q_{1,1}$ is the leading term. Each summands in $Q_{1,2}$ is bounded by $e^{\theta_0\tau}$, so $Q_{1,2}$ is of order $O_p(\sqrt{n}|\theta-\theta_0|)$. Through an integral calculation, we have

$$
\int_0^\tau e^{D_i\theta_0 t}D_i Y_i(t)dt = D_i\int_0^{X_i} e^{\theta_0 t}dt = D_i(e^{\theta_0 X_i}-1)/\theta_0, \qquad (3.128)
$$

so we can write $Q_{1,2}$ as

$$
-\frac{1}{\sqrt{n}}(\theta-\theta_0)\sum_{i=1}^{n}D_i\{1-\text{expit}(\gamma_0^\top \mathbf{Z}_{1,i})\}(e^{\theta_0 X_i}-1)/\theta_0. \qquad (3.129)
$$

In $Q_{1,3}$, we have a $\mathcal{F}_{n,t}$-martingale

$$
\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau\{D_i-\text{expit}(\gamma_0^\top \mathbf{Z}_{1,i})\}e^{D_i\theta_0 t}D_i t dM_i(t), \qquad (3.130)
$$

whose integrand is bounded by $e^{\theta_0\tau}$. By Lemma 37, (3.130) is of order $O_p(n^{-1/2})$. Hence, $Q_{1,3}$ is of order $O_p(|\theta-\theta_0|)=o_p(\sqrt{n}|\theta-\theta_0|)$. Note that we need to prove our statement uniformly in θ, so we cannot directly utilize the martingale structure in $Q_{1,4}$

$$
\int_0^\tau e^{\theta_t t}\frac{1}{n}\sum_{i=1}^{n}\{D_i-\text{expit}(\gamma_0^\top \mathbf{Z}_{1,i})\}D_i t^2 dM_i(t). \qquad (3.131)
$$

Alternatively, we use Lemma 40 to establish the rate of (3.131) as $O_p(n^{-1/2})$. The other term in $Q_{1,4}$

$$
\frac{1}{n}\sum_{i=1}^{n}\{D_i-\text{expit}(\gamma_0^\top \mathbf{Z}_{1,i})\}\int_0^\tau e^{\theta_t t}D_i t Y_i(t)dt \qquad (3.132)
$$

is bounded by $e^{K_\theta \tau}\tau$. Then, $Q_{1,4}$ is of order $O_p(\sqrt{n}|\theta-\theta_0|^2)$. Therefore, we have term Q_1 equals

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\phi(\theta_0;\beta_0,\Lambda_0,\gamma_0)-\frac{1}{\sqrt{n}}(\theta-\theta_0)\sum_{i=1}^{n}D_i\{1-\text{expit}(\gamma_0^\top \mathbf{Z}_{1,i})\}(e^{\theta_0 X_i}-1)/\theta_0 \tag{3.133}$$

plus an $o_p(\sqrt{n}|\theta-\theta_0|)+O_p(\sqrt{n}|\theta-\theta_0|^2)$ error.

We expand Q_2 with respect to θ,

$$\begin{aligned} Q_2 &= -n^{-\frac{1}{2}}\sum_{i=1}^{n}\{D_i-\text{expit}(\gamma_0^\top \mathbf{Z}_{1,i})\}\int_0^\tau e^{D_i\theta_0 t}Y_i(t)(\widehat{\beta}-\beta_0)^\top \mathbf{Z}_i dt \\ &\quad -n^{-\frac{1}{2}}(\theta-\theta_0)\sum_{i=1}^{n}D_i\{1-\text{expit}(\gamma_0^\top \mathbf{Z}_{1,i})\}\int_0^\tau e^{\theta_t t}Y_i(t)(\widehat{\beta}-\beta_0)^\top \mathbf{Z}_i dt \\ &= Q_{2,1}+Q_{2,2}, \end{aligned} \tag{3.134}$$

where $Q_{2,2}$ comes from the Mean Value Theorem for $e^{\theta t}-e^{\theta_0 t}$ as in Definition 3. By the Hölder's inequality, we have an bound for $Q_{2,1}$,

$$|Q_{2,1}|\leq \sqrt{n}\tau\|\widehat{\beta}-\beta\|_1 \sup_{t\in[0,\tau]}\left\|\frac{1}{n}\sum_{i=1}^{n}\{D_i-\text{expit}(\gamma_0^\top \mathbf{Z}_{1,i})\}e^{D_i\theta_0 t}Y_i(t)\mathbf{Z}_i\right\|_\infty .$$

From Lemma 41, we have

$$\sup_{t\in[0,\tau]}\left\|\frac{1}{n}\sum_{i=1}^{n}\{D_i-\text{expit}(\gamma_0^\top \mathbf{Z}_{1,i})\}e^{D_i\theta_0 t}Y_i(t)\mathbf{Z}_i\right\|_\infty = O_p\left(\frac{\log(p)}{\sqrt{n}}\right).$$

Under Assumption 5-ix, we have $Q_{2,1}=O_p\left(\log(p)\|\widehat{\beta}-\beta\|_1\right)=o_p(1)$. We again apply the Hölder's inequality to find the upper bound for $Q_{2,2}$,

$$|Q_{2,2}|\leq \sqrt{n}|\theta-\theta_0|\tau\|\widehat{\beta}-\beta\|_1 \sup_{t\in[0,\tau]}\left\|\frac{1}{n}\sum_{i=1}^{n}D_i\{1-\text{expit}(\gamma_0^\top \mathbf{Z}_{1,i})\}e^{\theta_0 t}Y_i(t)\mathbf{Z}_i\right\|_\infty .$$

Under Assumptions 5-iii and 5-ix, we have $Q_{2,2}=O_p\left(\sqrt{n}|\theta-\theta_0|\|\widehat{\beta}-\beta\|_1\right)=o\left(\sqrt{n}|\theta-\theta_0|\right)$. Hence, term $Q_2=Q_{2,1}+Q_{2,2}$ is of order $o_p(\sqrt{n}|\theta-\theta_0|+1)$.

Very similar to our treatment of Q_2, we expand Q_3 with respect to θ,

$$\begin{aligned} Q_3 &= -\sqrt{n}\int_0^\tau \left[\frac{1}{n}\sum_{i=1}^n \{D_i - \text{expit}(\gamma_0^\top \mathbf{Z}_{1,i})\} e^{D_i\theta_0 t} Y_i(t)\right] \left\{d\widehat{\Lambda}(t,\theta) - d\widehat{\Lambda}(t,\theta_0)\right\} \\ &\quad -\sqrt{n}\int_0^\tau \left[\frac{1}{n}\sum_{i=1}^n \{D_i - \text{expit}(\gamma_0^\top \mathbf{Z}_{1,i})\} e^{D_i\theta_0 t} Y_i(t)\right] \left\{d\widehat{\Lambda}(t,\theta_0) - d\Lambda_0(t)\right\} \\ &\quad -n^{-\frac{1}{2}}(\theta-\theta_0)\sum_{i=1}^n D_i\{1-\text{expit}(\gamma_0^\top \mathbf{Z}_{1,i})\}\int_0^\tau t e^{\theta_t t} Y_i(t)\left\{d\widehat{\Lambda}(t,\theta) - d\Lambda_0(t)\right\} \\ &= Q_{3,1} + Q_{3,2} + Q_{3,3}, \end{aligned} \tag{3.135}$$

where $Q_{3,3}$ comes from the Mean Value Theorem for $e^{\theta t} - e^{\theta_0 t}$ as in Definition 3. From Lemma 41, we know that,

$$\sup_{t\in[0,\tau]}\left|\frac{1}{n}\sum_{i=1}^n \{D_i - \text{expit}(\gamma_0^\top \mathbf{Z}_{1,i})\} e^{D_i\theta_0 t} Y_i(t)\right| = O_p\left(n^{-\frac{1}{2}}\right).$$

Together with Assumption 5-viii, the integral $Q_{3,1}$ as an upper bound

$$\sqrt{n}\sup_{t\in[0,\tau]}\left|\frac{1}{n}\sum_{i=1}^n \{D_i - \text{expit}(\gamma_0^\top \mathbf{Z}_{1,i})\} e^{D_i\theta_0 t} Y_i(t)\right| \bigvee_{t=0}^{\tau}\left\{\widehat{\Lambda}(t,\theta) - \widehat{\Lambda}(t,\theta_0)\right\} = o_p(\sqrt{n}|\theta-\theta_0|).$$

We apply (3.9) in Assumption 5-ix to $Q_{3,2}$ and get $Q_{3,2} = o_p(1)$. By Helly-Bray argument [Mur94], we have a bound for $Q_{3,3}$

$$|Q_{3,3}| \le \sqrt{n}|\theta-\theta_0|\left\{\left|\widehat{\Lambda}(\tau,\theta) - \Lambda_0(\tau)\right| \tau e^{K_\theta \tau} + \int_0^\tau \left|\widehat{\Lambda}(t,\theta) - \Lambda_0(t)\right| dt e^{\theta_t t}\right\}.$$

Under Assumptions 5-viii and 5-ix, our bound gives the rate $Q_{3,3} = o_p(\sqrt{n}|\theta-\theta_0|) + O_p(\sqrt{n}|\theta-\theta_0|^2)$. Therefore, $Q_3 = Q_{3,1} + Q_{3,2} + Q_{3,3} = o_p(\sqrt{n}|\theta-\theta_0|+1) + O_p(\sqrt{n}|\theta-\theta_0|^2)$.

In terms $Q_4 - Q_6$, we have the model estimation error for the logistic regression. By a Mean Value Theorem argument, we have a uniform bound for the error

$$\left|\text{expit}(\widehat{\gamma}^\top \mathbf{Z}_{1,i}) - \text{expit}(\gamma_0^\top \mathbf{Z}_{1,i})\right| \le \|\widehat{\gamma}-\gamma\|_1 \sup_{i=1,\ldots,n}\|\mathbf{Z}_i\|_\infty \tag{3.136}$$

because the derivative of function expit$(\cdot)$ is uniformly bounded by one.

We expand Q_4 with respect to θ,

$$\begin{aligned}
Q_4 =& -n^{-\frac{1}{2}} \sum_{i=1}^{n} \int_0^{\tau} \{\text{expit}(\widehat{\gamma}^\top \mathbf{Z}_{1,i}) - \text{expit}(\gamma_0^\top \mathbf{Z}_{1,i})\} e^{D_i \theta_0 t} dM_i(t) \\
& - n^{-\frac{1}{2}} (\theta - \theta_0) \sum_{i=1}^{n} \int_0^{\tau} e^{\theta_t t} D_i \{\text{expit}(\widehat{\gamma}^\top \mathbf{Z}_{1,i}) - \text{expit}(\gamma_0^\top \mathbf{Z}_{1,i})\} t dM_i(t) \\
& + n^{-\frac{1}{2}} (\theta - \theta_0) \sum_{i=1}^{n} \{\text{expit}(\widehat{\gamma}^\top \mathbf{Z}_{1,i}) - \text{expit}(\gamma_0^\top \mathbf{Z}_{1,i})\} \int_0^{\tau} Y_i(t) D_i e^{\theta_t t} (t\theta_t - t\theta_0 + 1) dt \\
=& Q_{4,1} + Q_{4,2} + Q_{4,3}, \qquad (3.137)
\end{aligned}$$

where $Q_{4,3}$ comes from the Mean Value Theorem for $e^{\theta t} - e^{\theta_0 t}$ as in Definition 3. $\widehat{\gamma}$ is $\mathcal{F}_{n,t}$-measurable, so we can apply Lemma 38 to $Q_{4,1}$. According to (3.136) and Assumptions 5-iii and 5-ix, $Q_{4,1} = O_p(\|\widehat{\gamma} - \gamma\|_1) = o_p(1)$. For $Q_{4,2}$, we apply Lemma 40 with $\mathcal{H}$ be the set of $\{e^{\theta_t t} : |\theta_t| \leq K_\theta\}$ to get $Q_{4,2} = O_p(|\theta - \theta_0|)$. For $Q_{4,3}$, we use the uniform bound from (3.136)

$$|Q_{4,3}| \leq \sqrt{n}|\theta - \theta_0| \|\widehat{\gamma} - \gamma\|_1 \sup_{i=1,\ldots,n} \|\mathbf{Z}_i\|_\infty e^{K_\theta \tau} (2K_\theta \tau + 1).$$

Under Assumption 5-iii and 5-ix, $Q_{4,3} = o_p(\sqrt{n}|\theta - \theta_0|)$. Therefore, we obtain $Q_4 = o_p(\sqrt{n}|\theta - \theta_0| + 1)$.

We apply the Hölder's inequality and (3.122) to Q_5,

$$|Q_5| \leq \sqrt{n} \|\widehat{\gamma} - \gamma_0\|_1 \|\widehat{\beta} - \beta_0\|_1 e^{K_\theta \tau} \tau \left\{ \max_{i=1,\ldots,n} \|\mathbf{Z}_i\|_\infty \right\}^2.$$

Under Assumptions 5-iii and 5-ix, we have $Q_5 = o_p(1)$.

We expand Q_6 with respect to θ similar to Q_3,

$$Q_6 = \sqrt{n} \int_0^{\tau} \left[\frac{1}{n} \sum_{i=1}^{n} \{\text{expit}(\widehat{\gamma}^\top \mathbf{Z}_{1,i}) - \text{expit}(\gamma_0^\top \mathbf{Z}_{1,i})\} e^{D_i \theta_0 t} Y_i(t) \right] \left\{ d\widehat{\Lambda}(t, \theta) - d\widehat{\Lambda}(t, \theta_0) \right\}$$

$$+\sqrt{n}\frac{1}{n}\sum_{i=1}^{n}\{\text{expit}(\widehat{\gamma}^\top \mathbf{Z}_{1,i})-\text{expit}(\gamma_0^\top \mathbf{Z}_{1,i})\}\int_0^\tau e^{D_i\theta_0 t}Y_i(t)\left\{d\widehat{\Lambda}(t,\theta_0)-d\Lambda_0(t)\right\}$$

$$+n^{-\frac{1}{2}}(\theta-\theta_0)\sum_{i=1}^{n}\{\text{expit}(\widehat{\gamma}^\top \mathbf{Z}_{1,i})-\text{expit}(\gamma_0^\top \mathbf{Z}_{1,i})\}\int_0^\tau te^{\theta_t t}Y_i(t)\left\{d\widehat{\Lambda}(t,\theta)-d\Lambda_0(t)\right\}$$

$$=Q_{6,1}+Q_{6,2}+Q_{6,3},$$

where $Q_{6,3}$ comes from the Mean Value Theorem for $e^{\theta t}-e^{\theta_0 t}$ as in Definition 3. By (3.122) and Assumptions 5-viii, we have an upper bound for the integral $Q_{6,1}$

$$\sqrt{n}\sup_{t\in[0,\tau]}\|\widehat{\gamma}-\gamma\|_1\sup_{i=1,\ldots,n}\|\mathbf{Z}_i\|_\infty\bigvee_{t=0}^{\tau}\left\{\widehat{\Lambda}(t,\theta)-\widehat{\Lambda}(t,\theta_0)\right\}=o_p(\sqrt{n}|\theta-\theta_0|).$$

By Helly-Bray argument [Mur94], we have a bound for the integral

$$\left|\int_0^\tau e^{D_i\theta_0 t}Y_i(t)\left\{d\widehat{\Lambda}(t,\theta_0)-d\Lambda_0(t)\right\}\right|\leq\left\{\left|\widehat{\Lambda}(\tau,\theta_0)-\Lambda_0(\tau)\right|e^{K_\theta\tau}+\int_0^\tau\left|\widehat{\Lambda}(t,\theta_0)-\Lambda_0(t)\right|de^{\theta_t t}\right\}.$$

Apply the bound to $Q_{6,2}$ along with (3.122), we have

$$Q_{6,2}=O_p\left(\sqrt{n}\|\widehat{\gamma}-\gamma_0\|_1\max_{i=1,\ldots,n}\|\mathbf{Z}_i\|_\infty\sup_{t\in[0,\tau]}\left|\widehat{\Lambda}(t,\theta_0)-\Lambda_0(t)\right|\right).$$

Under Assumptions 5-iii and 5-ix, $Q_{6,2}=o_p(1)$. Similarly, we have $Q_{6,3}=O_p\left(\sqrt{n}|\theta-\theta_0|^2\right)+o_p(\sqrt{n}|\theta-\theta_0|)$. Hence, we obtain the rate $Q_6=Q_{6,1}+Q_{6,2}+Q_{6,3}=O_p\left(\sqrt{n}|\theta-\theta_0|^2\right)+o_p(\sqrt{n}|\theta-\theta_0|+1)$.

Combining the results for Q_1-Q_6, we finish the proof. □

Proof of Lemma 30. The proof of the lemma follows fundamentally the same strategy as that of Lemma 29. The main difference is that we use the Cauchy Schwartz inequality instead of the Hölder's inequality to derive MSE type of bounds.

We define the filtration for the j-th fold as

$$\mathcal{F}_{I_j,t}=\sigma\left(\{N_i(u),Y_i(u+),D_i,Z_i:u\leq t,i\in I_j\}\cup\{\delta_i,X_i,D_i,Z_i:i\in I_{-j}\}\right),$$

using $I = I_j$ in Definition 2.

We prove the statement (3.106) by investigating each terms in the following expansion,

$$
\begin{aligned}
& \sqrt{n}\phi^{(j)}(\theta;\widehat{\boldsymbol{\beta}}^{(j)},\widehat{\Lambda}^{(j)}(\cdot,\theta),\widehat{\boldsymbol{\gamma}}^{(j)}) \\
= \;& \sqrt{n}\phi^{(j)}(\theta;\boldsymbol{\beta}_0,\Lambda_0,\boldsymbol{\gamma}_0) \\
& -n^{-\frac{1}{2}}\sum_{i\in I_j}\{D_i-\mathrm{expit}(\boldsymbol{\gamma}_0^\top \mathbf{Z}_{1,i})\}\int_0^\tau e^{D_i\theta t}Y_i(t)(\widehat{\boldsymbol{\beta}}^{(j)}-\boldsymbol{\beta}_0)^\top \mathbf{Z}_i dt \\
& -n^{-\frac{1}{2}}\sum_{i\in I_j}\{D_i-\mathrm{expit}(\boldsymbol{\gamma}_0^\top \mathbf{Z}_{1,i})\}\int_0^\tau e^{D_i\theta t}Y_i(t)\left\{d\widehat{\Lambda}^{(j)}(t,\theta)-d\Lambda_0(t)\right\} \\
& -n^{-\frac{1}{2}}\sum_{i\in I_j}\{\mathrm{expit}(\widehat{\boldsymbol{\gamma}}^{(j)\top}\mathbf{Z}_{1,i})-\mathrm{expit}(\boldsymbol{\gamma}_0^\top \mathbf{Z}_{1,i})\}\int_0^\tau e^{D_i\theta t}dM_i(t;\theta,\boldsymbol{\beta}_0,\Lambda_0) \\
& +n^{-\frac{1}{2}}\sum_{i\in I_j}\{\mathrm{expit}(\widehat{\boldsymbol{\gamma}}^{(j)\top}\mathbf{Z}_{1,i})-\mathrm{expit}(\boldsymbol{\gamma}_0^\top \mathbf{Z}_{1,i})\}\int_0^\tau e^{D_i\theta t}Y_i(t)(\widehat{\boldsymbol{\beta}}^{(j)}-\boldsymbol{\beta}_0)^\top \mathbf{Z}_i dt \\
& +n^{-\frac{1}{2}}\sum_{i\in I_j}\{\mathrm{expit}(\widehat{\boldsymbol{\gamma}}^{(j)\top}\mathbf{Z}_{1,i})-\mathrm{expit}(\boldsymbol{\gamma}_0^\top \mathbf{Z}_{1,i})\}\int_0^\tau e^{D_i\theta t}Y_i(t)\left\{d\widehat{\Lambda}^{(j)}(t,\theta)-d\Lambda_0(t)\right\} \\
= \;& Q_1+Q_2+Q_3+Q_4+Q_5+Q_6. \qquad (3.138)
\end{aligned}
$$

The first term Q_1 contains the leading terms. The rest Q_2-Q_6 are the remainders.

Following exactly the same derivations in the proof of Lemma 29, we have term Q_1 equals

$$
\frac{1}{\sqrt{n}}\sum_{i\in I_j}\phi^{(j)}(\theta_0;\boldsymbol{\beta}_0,\Lambda_0,\boldsymbol{\gamma}_0)-\frac{1}{\sqrt{n}}(\theta-\theta_0)\sum_{i\in I_j}D_i\{1-\mathrm{expit}(\boldsymbol{\gamma}_0^\top \mathbf{Z}_{1,i})\}(e^{\theta_0 X_i}-1)/\theta_0 \qquad (3.139)
$$

plus an $o_p(\sqrt{n}|\theta-\theta_0|)+O_p(\sqrt{n}|\theta-\theta_0|^2)$ error.

We expand Q_2 with respect to θ,

$$
\begin{aligned}
Q_2 \;=\; & -n^{-\frac{1}{2}}\sum_{i\in I_j}\{D_i-\mathrm{expit}(\boldsymbol{\gamma}_0^\top \mathbf{Z}_{1,i})\}\int_0^\tau e^{D_i\theta_0 t}Y_i(t)(\widehat{\boldsymbol{\beta}}^{(j)}-\boldsymbol{\beta}_0)^\top \mathbf{Z}_i dt \\
& -n^{-\frac{1}{2}}(\theta-\theta_0)\sum_{i\in I_j}D_i\{1-\mathrm{expit}(\boldsymbol{\gamma}_0^\top \mathbf{Z}_{1,i})\}\int_0^\tau e^{\theta_t t}Y_i(t)(\widehat{\boldsymbol{\beta}}^{(j)}-\boldsymbol{\beta}_0)^\top \mathbf{Z}_i dt \\
= \;& Q_{2,1}+Q_{2,2}, \qquad (3.140)
\end{aligned}
$$

where $Q_{2,2}$ comes from the Mean Value Theorem for $e^{\theta t} - e^{\theta_0 t}$ as in Definition 3. Denote

$$Q_{2,1,i} = \{D_i - \text{expit}(\boldsymbol{\gamma}_0^\top \mathbf{Z}_{1,i})\} \int_0^\tau e^{D_i \theta_0 t} Y_i(t) (\widehat{\boldsymbol{\beta}}^{(j)} - \boldsymbol{\beta}_0)^\top \mathbf{Z}_i dt.$$

Using the independence across folds, we can calculate the expectation for $i \in I_j$

$$\begin{aligned}
&\mathbb{E}(Q_{2,1,i}) \\
&= \int_0^\tau \mathbb{E}(\widehat{\boldsymbol{\beta}}^{(j)} - \boldsymbol{\beta}_0)^\top \mathbb{E}[\{D_i - \text{expit}(\boldsymbol{\gamma}_0^\top \mathbf{Z}_{1,i})\} e^{D_i \theta_0 t} Y_i(t) \mathbf{Z}_i] dt \\
&= \int_0^\tau \mathbb{E}(\widehat{\boldsymbol{\beta}}^{(j)} - \boldsymbol{\beta}_0)^\top \mathbb{E}[\mathbb{E}\{D_i - \text{expit}(\boldsymbol{\gamma}_0^\top \mathbf{Z}_{1,i}) | \mathbf{Z}_i\} \mathbb{E}\{e^{D_i \theta_0 t} Y_i(t) | D_i, \mathbf{Z}_i\} \mathbf{Z}_i] dt, \qquad (3.141)
\end{aligned}$$

which equals zero by Lemma 41. Hence, $\mathbb{E}(Q_{2,1}) = 0$. We calculate the variance of $Q_{2,1}$

$$\text{Var}(Q_{2,1}) = n^{-1} \sum_{i \in I_j} \mathbb{E}(Q_{2,1,i}^2) + 2n^{-1} \sum_{i<j, \{i,j\} \subset I_j} \mathbb{E}(Q_{2,1,i} Q_{2,1,j}). \qquad (3.142)$$

Note that we have

$$\left| \int_0^\tau e^{D_i \theta t} Y_i(t) (\widehat{\boldsymbol{\beta}}^{(j)} - \boldsymbol{\beta}_0)^\top \mathbf{Z}_i dt \right| \leq e^{K_\theta \tau} X_i \left| (\widehat{\boldsymbol{\beta}}^{(j)} - \boldsymbol{\beta}_0)^\top \mathbf{Z}_i \right|. \qquad (3.143)$$

Under Assumption 6-i,

$$n^{-1} \sum_{i \in I_j} \mathbb{E}(Q_{2,1,i}^2) \leq \frac{|I_j|}{n} e^{2K_\theta \tau} \tau \left\{ \mathcal{D}_\beta \left(\widehat{\boldsymbol{\beta}}^{(j)}, \boldsymbol{\beta}_0 \right) \right\}^2 = O_p(r_n^{*2}) = o_p(1).$$

Using the independence across folds again, we have

$$\mathbb{E}(Q_{2,1,i} Q_{2,1,j}) = \mathbb{E}\{\mathbb{E}(Q_{2,1,i} | \widehat{\boldsymbol{\beta}}^{(j)}) \mathbb{E}(Q_{2,1,j} | \widehat{\boldsymbol{\beta}}^{(j)})\} = 0. \qquad (3.144)$$

Thus, we establish the rate $\text{Var}(Q_{2,1}) = o_p(1)$. By the Tchebychev's inequality, we have $Q_{2,1} = o_p(1)$. For $Q_{2,2}$, we denote

$$Q_{2,2,i} = D_i \{1 - \text{expit}(\boldsymbol{\gamma}_0^\top \mathbf{Z}_{1,i})\} \int_0^\tau e^{\theta_t t} Y_i(t) (\widehat{\boldsymbol{\beta}}^{(j)} - \boldsymbol{\beta}_0)^\top \mathbf{Z}_i dt \qquad (3.145)$$

apply Cauchy-Schwartz inequality to give an upper bound

$$|Q_{2,2}| \leq n^{-\frac{1}{2}}(\theta-\theta_0)\sqrt{n\sum_{i\in I_j} Q_{2,2,i}^2}. \tag{3.146}$$

Under Assumption 6-i, we have from bound (3.143)

$$\mathbb{E}\{Q_{2,2,i}^2\} \leq e^{2K_\theta\tau}\tau\left\{\mathcal{D}_\beta\left(\widehat{\boldsymbol{\beta}}^{(j)},\boldsymbol{\beta}_0\right)\right\}^2 = o_p(1).$$

Applying the Markov's inequality to $\sum_{i\in I_j} Q_{2,2,i}^2$, we have $Q_{2,2} = o_p(\sqrt{n}|\theta-\theta_0|)$. Hence, term Q_2 is of order $o_p(\sqrt{n}|\theta-\theta_0|+1)$.

Very similar to our treatment of Q_2, we expand Q_3 with respect to θ,

$$\begin{aligned} Q_3 &= -\sqrt{n}\int_0^\tau \left[\frac{1}{n}\sum_{i\in I_j}\{D_i - \text{expit}(\boldsymbol{\gamma}_0^\top \mathbf{Z}_{1,i})\}e^{D_i\theta_0 t}Y_i(t)\right]\left\{d\widehat{\Lambda}^{(j)}(t,\theta) - d\widehat{\Lambda}^{(j)}(t,\theta_0)\right\} \\ &\quad -n^{-\frac{1}{2}}\sum_{i\in I_j}\{D_i - \text{expit}(\boldsymbol{\gamma}_0^\top \mathbf{Z}_{1,i})\}\int_0^\tau e^{D_i\theta_0 t}Y_i(t)\left\{d\widehat{\Lambda}^{(j)}(t,\theta_0) - d\Lambda_0(t)\right\} \\ &\quad -n^{-\frac{1}{2}}(\theta-\theta_0)\sum_{i\in I_j} D_i\{1-\text{expit}(\boldsymbol{\gamma}_0^\top \mathbf{Z}_{1,i})\}\int_0^\tau t e^{D_i\theta_t t}Y_i(t)\left\{d\widehat{\Lambda}^{(j)}(t,\theta) - d\Lambda_0(t)\right\} \\ &= Q_{3,1}+Q_{3,2}+Q_{3,3}, \end{aligned} \tag{3.147}$$

where $Q_{3,3}$ comes from the Mean Value Theorem for $e^{\theta t} - e^{\theta_0 t}$ as in Definition 3. From Lemma 41, we have,

$$\sup_{t\in[0,\tau]}\left|\frac{1}{n}\sum_{i\in I_j}\{D_i - \text{expit}(\boldsymbol{\gamma}_0^\top \mathbf{Z}_{1,i})\}e^{D_i\theta_0 t}Y_i(t)\right| = O_p\left(n^{-\frac{1}{2}}\right).$$

Together with Assumption 5-viii, the integral $Q_{3,1}$ as an upper bound

$$\sqrt{n}\sup_{t\in[0,\tau]}\left|\frac{1}{n}\sum_{i\in I_j}\{D_i - \text{expit}(\boldsymbol{\gamma}_0^\top \mathbf{Z}_{1,i})\}e^{D_i\theta_0 t}Y_i(t)\right|\bigvee_{t=0}^{\tau}\left\{\widehat{\Lambda}^{(j)}(t,\theta) - \widehat{\Lambda}^{(j)}(t,\theta_0)\right\} = o_p(\sqrt{n}|\theta-\theta_0|).$$

Denote

$$Q_{3,2,i} = \{D_i - \text{expit}(\boldsymbol{\gamma}_0^\top \mathbf{Z}_{1,i})\}\int_0^\tau e^{D_i\theta_0 t}Y_i(t)\left\{d\widehat{\Lambda}^{(j)}(t,\theta_0) - d\Lambda_0(t)\right\}.$$

Using the independence across folds, we can calculate the expectation for $i \in I_j$ according to Lemma 41

$$\mathbb{E}(Q_{3,2,i}) = \int_0^\tau \mathbb{E}\left(\mathbb{E}[\{D_i - \text{expit}(\gamma_0^\top \mathbf{Z}_{1,i})\} e^{D_i \theta_0 t} Y_i(t) | \mathbf{Z}_i]\right) \left[d\mathbb{E}\left\{\widehat{\Lambda}^{(j)}(t, \theta_0)\right\} - d\Lambda_0(t)\right],$$

which equals zero by Lemma 41. Hence, $\mathbb{E}(Q_{3,2}) = 0$. Moreover, we have a diminishing bound for $Q_{3,2,i}$ by Helly-Bray argument [Mur94] under Assumption 6-i

$$\max_{i \in I_j} |Q_{3,2,i}| \leq \left|\widehat{\Lambda}^{(j)}(\tau, \theta_0) - \Lambda_0(\tau)\right| e^{K_\theta \tau} + \int_0^\tau \left|\widehat{\Lambda}^{(j)}(t, \theta_0) - \Lambda_0(t)\right| de^{\theta_0 t} = o_p(1).$$

We denote $M_{3,2,m} = \frac{1}{\sqrt{n}} \sum_{i \in I_j^{1:m}} Q_{3,2,i}$ as the partial sum of the first m terms in $Q_{3,2}$ whose indices are in $I_j^{1:m}$. It is a martingale with respect to filtration $\mathcal{F}_{3,2,m} = \sigma\left(\{W_i : i \in I_j^{1:m} \cup I_{-j}\}\right)$. By the Azuma's inequality (as in Lemma 32), we have $Q_{3,2} = M_{3,2,|I_j|} = o_p(1)$. Similarly, we apply Helly-Bray argument [Mur94] to show that

$$|Q_{3,3}| \leq \sqrt{n}|\theta - \theta_0| \left\{ \left|\widehat{\Lambda}^{(j)}(\tau, \theta) - \Lambda_0(\tau)\right| \tau e^{K_\theta \tau} + \int_0^\tau \left|\widehat{\Lambda}^{(j)}(t, \theta) - \Lambda_0(t)\right| de^{\theta_t t} \right\}.$$

Under Assumptions 5-viii and 6-i, we have $Q_{3,3} = o_p(\sqrt{n}|\theta - \theta_0|) + O_p(\sqrt{n}|\theta - \theta_0|^2)$. Therefore, $Q_3 = Q_{3,1} + Q_{3,2} + Q_{3,3} = o_p(\sqrt{n}|\theta - \theta_0| + 1) + O_p(\sqrt{n}|\theta - \theta_0|^2)$.

We expand Q_4 with respect to θ,

$$\begin{aligned}
Q_4 =& -n^{-\frac{1}{2}} \sum_{i \in I_j} \{\text{expit}(\widehat{\gamma}^{(j)\top} \mathbf{Z}_{1,i}) - \text{expit}(\gamma_0^\top \mathbf{Z}_{1,i})\} \int_0^\tau e^{D_i \theta_0 t} dM_i(t) \\
& - n^{-\frac{1}{2}} (\theta - \theta_0) \sum_{i \in I_j} \int_0^\tau e^{\theta_t t} D_i \{\text{expit}(\widehat{\gamma}^{(j)\top} \mathbf{Z}_{1,i}) - \text{expit}(\gamma_0^\top \mathbf{Z}_{1,i})\} t dM_i(t) \\
& + n^{-\frac{1}{2}} (\theta - \theta_0) \sum_{i \in I_j} \{\text{expit}(\widehat{\gamma}^{(j)\top} \mathbf{Z}_{1,i}) - \text{expit}(\gamma_0^\top \mathbf{Z}_{1,i})\} \int_0^\tau Y_i(t) D_i e^{\theta_t t} (t\theta_t - t\theta_0 + 1) dt \\
=& Q_{4,1} + Q_{4,2} + Q_{4,3}, \qquad (3.148)
\end{aligned}$$

where $Q_{4,3}$ comes from the Mean Value Theorem for $e^{\theta t} - e^{\theta_0 t}$ as in Definition 3. Denote

$$Q_{4,1,i}(t) = \{\text{expit}(\widehat{\gamma}^{(j)\top}\mathbf{Z}_{1,i}) - \text{expit}(\gamma_0^\top \mathbf{Z}_{1,i})\} \int_0^t e^{D_i\theta_0 t} dM_i(t).$$

Since $\{\text{expit}(\widehat{\gamma}^{(j)\top}\mathbf{Z}_{1,i}) - \text{expit}(\gamma_0^\top \mathbf{Z}_{1,i})\}e^{D_i\theta_0 t}$ is $\mathcal{F}_{I_j,t}$-adapted, each $Q_{4,1,i}(t)$ is $\mathcal{F}_{I_j,t}$-martingales. Then, $\mathbb{E}\{Q_{4,1}\} = 0$. The optional quadratic variation of $\sum_{i\in I_j} Q_{4,1,i}$ is

$$\begin{aligned}\left[\sum_{i\in I_j} Q_{4,1,i}\right]_t &= \sum_{i\in I_j} \{\text{expit}(\widehat{\gamma}^{(j)\top}\mathbf{Z}_{1,i}) - \text{expit}(\gamma_0^\top \mathbf{Z}_{1,i})\}^2 \int_0^t e^{2D_i\theta_0 t} dN_i(t) \\ &\leq \sum_{i\in I_j} \{\text{expit}(\widehat{\gamma}^{(j)\top}\mathbf{Z}_{1,i}) - \text{expit}(\gamma_0^\top \mathbf{Z}_{1,i})\}^2 e^{2K_\theta \tau}.\end{aligned}$$

Under Assumption 6-i, we have $\mathbb{E}\{\text{expit}(\widehat{\gamma}^{(j)\top}\mathbf{Z}_{1,i}) - \text{expit}(\gamma_0^\top \mathbf{Z}_{1,i})\}^2 = \left\{\mathcal{D}_\gamma\left(\widehat{\gamma}^{(j)}, \gamma_0\right)\right\}^2 = o_p(1)$. Hence,

$$\text{Var}(Q_{4,1}) = n^{-1} \sum_{i\in I_j} \mathbb{E}\left\{\left[\sum_{i\in I_j} Q_{4,1,i}\right]_\tau\right\} = o_p(1).$$

We obtain $Q_{4,1} = o_p(1)$ by the Tchebychev's inequality. For $Q_{4,2}$, we apply Lemma 40 with $\mathcal{H}$ be the set of $\{e^{\theta_t t} : |\theta_t| \leq K_\theta\}$ to get $Q_{4,2} = O_p(|\theta - \theta_0|)$. For $Q_{4,3}$, we apply the Cauchy-Schwartz inequality

$$|Q_{4,3}| \leq n^{-\frac{1}{2}}|\theta - \theta_0| \sqrt{\sum_{i\in I_j} \{\text{expit}(\widehat{\gamma}^{(j)\top}\mathbf{Z}_{1,i}) - \text{expit}(\gamma_0^\top \mathbf{Z}_{1,i})\}^2} \sqrt{ne^{2K_\theta\tau}(K_\theta\tau + \tau)^2}.$$

Again with $\mathbb{E}\{\text{expit}(\widehat{\gamma}^{(j)\top}\mathbf{Z}_{1,i}) - \text{expit}(\gamma_0^\top \mathbf{Z}_{1,i})\}^2 = O_p(q_n^*) = o_p(1)$, we obtain from the Markov's inequality that $\sum_{i\in I_j}\{\text{expit}(\widehat{\gamma}^{(j)\top}\mathbf{Z}_{1,i}) - \text{expit}(\gamma_0^\top \mathbf{Z}_{1,i})\}^2 = o_p(1)$. Hence, $Q_{4,3} = o_p(\sqrt{n}|\theta - \theta_0|)$. Therefore, we obtain $Q_4 = o_p(\sqrt{n}|\theta - \theta_0| + 1)$.

We apply the Cauchy-Schwartz inequality to Q_5,

$$|Q_5| \leq n^{-\frac{1}{2}} e^{K_\theta\tau} \sqrt{\sum_{i\in I_j} \{\text{expit}(\widehat{\gamma}^{(j)\top}\mathbf{Z}_{1,i}) - \text{expit}(\gamma_0^\top \mathbf{Z}_{1,i})\}^2} \sqrt{\sum_{i\in I_j} \left\{(\widehat{\boldsymbol{\beta}}^{(j)} - \boldsymbol{\beta}_0)^\top \mathbf{Z}_i\right\}^2 X_i^2}.$$

Using the independence across folds, we apply the Markov's inequality to get

$$Q_5 = O_p\left(\sqrt{n}\mathcal{D}_\gamma\left(\widehat{\boldsymbol{\gamma}}^{(j)},\boldsymbol{\gamma}_0\right)\mathcal{D}_\beta\left(\widehat{\boldsymbol{\beta}}^{(j)},\boldsymbol{\beta}_0\right)\right),$$

which is $o_p(1)$ under Assumption 6-i.

Similarly, we apply the Cauchy-Schwartz inequality to Q_6,

$$\begin{aligned}|Q_6| \leq & n^{-\frac{1}{2}} e^{K_\theta \tau} \sqrt{\sum_{i\in I_j}\{\text{expit}(\widehat{\boldsymbol{\gamma}}^{(j)\top}\mathbf{Z}_{1,i}) - \text{expit}(\boldsymbol{\gamma}_0^\top \mathbf{Z}_{1,i})\}^2} \\ & \times \sqrt{\sum_{i\in I_j}\left[\int_0^\tau e^{D_i\theta t} Y_i(t)\left\{d\widehat{\Lambda}^{(j)}(t,\theta) - d\Lambda_0(t)\right\}\right]^2}.\end{aligned}$$

Under Assumption 5-vii, we can apply the Helly-Bray argument [Mur94] to find the bound,

$$\begin{aligned}\left|\int_0^\tau e^{D_i\theta t} Y_i(t)\left\{d\widehat{\Lambda}^{(j)}(t,\theta) - d\Lambda_0(t)\right\}\right| \leq & \left|e^{D_i\theta X_i}\left\{\widehat{\Lambda}^{(j)}(X_i,\theta) - \Lambda_0(X_i)\right\}\right| \\ & + \left|\int_0^{X_i} D_i\theta e^{\theta t}\left\{\widehat{\Lambda}^{(j)}(t,\theta) - \Lambda_0(t)\right\} dt\right|.\end{aligned}$$

Hence, $Q_6 = O_p\left(\sqrt{n}\mathcal{D}_\gamma\left(\widehat{\boldsymbol{\gamma}}^{(j)},\boldsymbol{\gamma}_0\right)\sup_{t\in[0,\tau]}\left|\widehat{\Lambda}^{(j)}(t,\theta) - \Lambda_0\right|\right)$, which is $o_p(1+\sqrt{n}|\theta-\theta_0|)$ under Assumptions 5-viii and 6-i.

Combining the results for Q_1-Q_6, we finish the prove. □

Proof of Lemma 35. We prove the result for nonnegative $H_i(t)$. The general result can be obtained through decomposing $H_i(t)$ into the difference of two nonnegative processes

$$H_i(t) = H_i(t)\vee 0 - [-\{H_i(t)\wedge 0\}] \tag{3.149}$$

and use the union bound with the result for the nonnegative processes.

Under the model (3.1), μ satisfies $\mathbb{P}(D_i\theta_0 + \boldsymbol{\beta}_0^\top \mathbf{Z}_i \geq -d\Lambda_0(t)) = 1$. By the Assumption 5-iv, we have a lower bound $\mathbb{P}(\boldsymbol{\beta}_0^\top \mathbf{Z}_i > -K_\Lambda - \theta_0 \vee 0) = 1$. The $\boldsymbol{\beta}_0^\top \mathbf{Z}_i$ is potentially unbounded

from above, so we have to study the bound for the upper tail. For $x > K_H(K_\Lambda + \theta_0 \vee 0)\tau$,

$$\begin{aligned}
& \mathbb{P}\left(\int_0^\tau H_i(t)Y_i(t)\beta_0^\top \mathbf{Z}_i dt > x\right) \\
\leq\ & \mathbb{P}\left(K_H X_i \beta_0^\top \mathbf{Z}_i > x\right) \\
\leq\ & \mathbb{E}\left[I(\beta_0^\top \mathbf{Z}_i > K_\Lambda + \theta_0 \vee 0) I(C_i > x/K_H) \exp\left\{-\frac{x}{K_H}\frac{D_i\theta_0 + \beta_0^\top \mathbf{Z}_i}{\beta_0^\top \mathbf{Z}_i} - \Lambda_0\left(\frac{x/K_H}{\beta_0^\top \mathbf{Z}_i}\right)\right\}\right] \\
\leq\ & e^{-x/(2K_H)}. \qquad (3.150)
\end{aligned}$$

Denote $A_i = \int_0^\tau H_i(t)Y_i(t)\beta_0^\top \mathbf{Z}_i dt$, $\mu = \int_0^\tau \mathbb{E}\{H_i(t)Y_i(t)\beta_0^\top \mathbf{Z}_i\}dt$ and $K_A = K_H(K_\Lambda + \theta_0 \vee 0)\tau$. First, we can find a bound for the expectation

$$\begin{aligned}
|\mu| &= \left|\int_0^\tau \mathbb{E}\{H_i(t)Y_i(t)\beta_0^\top \mathbf{Z}_i\}dt\right| \\
&\leq \left|\int_0^\tau \mathbb{E}\{H_i(t)Y_i(t)\beta_0^\top \mathbf{Z}_i I(|\beta_0^\top \mathbf{Z}_i| < K_A)\}dt\right| \\
&\quad + \left|\mathbb{E}\left\{\int_0^\tau H_i(t)Y_i(t)\beta_0^\top \mathbf{Z}_i I(\beta_0^\top \mathbf{Z}_i \geq K_A)dt\right\}\right| \\
&\leq K_H K_A \tau + \int_0^\infty \mathbb{P}\left(\int_0^\tau H_i(t)Y_i(t)\beta_0^\top \mathbf{Z}_i I(\beta_0^\top \mathbf{Z}_i \geq K_A)dt > x\right)dx \\
&\leq K_H K_A + 2K_H. \qquad (3.151)
\end{aligned}$$

Then, we bound the centered moments for $k \geq 2$

$$\begin{aligned}
\mathbb{E}(A_i - \mu)^k &= \mathbb{E}\{(A_i - \mu)^k I(A_i < K_A + \mu \vee 0)\} + \mathbb{E}\{(A_i - \mu)^k I(A_i \geq K_A + \mu \vee 0)\} \\
&\leq (K_A + |\mu|)^k + \int_0^\infty \mathbb{P}\{(A_i - \mu)^k I(A_i \geq K_A + \mu \vee 0) > x\}dx \\
&\leq (K_A + |\mu|)^k + \int_0^{(K_A - \mu \wedge 0)^k} \mathbb{P}(A_i \geq K_A + \mu \vee 0)dx \\
&\quad + \int_{(K_A - \mu \wedge 0)^k}^\infty \mathbb{P}(A_i > x^{1/k} + \mu)dx \\
&\leq 2(K_A + |\mu|)^k + k!(2K_H)^k \\
&\leq k!(K_A + |\mu| + 2K_H)^k \qquad (3.152)
\end{aligned}$$

Thus, A_i is sub-exponential. By Bernstein inequality for sub-exponential random variables (as in Lemma 33), we have for any $\varepsilon \in [0, \sqrt{2}]$

$$\mathbb{P}\left(\left| \frac{1}{|I|} \sum_{i \in I} A_i - \mu \right| > \varepsilon (K_A + |\mu| + 2K_H) \right) < 2e^{-|I|\varepsilon^2/2}. \tag{3.153}$$

We thus complete the proof. □

Proof of Lemma 36. Let $X_{(1)}, \ldots, X_{(|I|)}$ be the order statistics of observed times. Under filtration $\mathcal{F}_{I,t}$, they are ordered stopping times (see Definition 2 and Remark 17). By optional stopping theorem [Dur10], we construct a discrete stopped martingale

$$M_k^H = \frac{1}{|I|} \sum_{i \in I} \int_0^{X_{(k)}} H_i(t) dM_i(t) \tag{3.154}$$

under filtration $\mathcal{F}_k^H = \sigma\{N_i(u), Y_i(u+), D_i, \mathbf{Z}_i, X_{(k)} : u \in [0, X_{(k)}], i \in I\}$. The increment of the discrete martingale has two components,

$$\begin{aligned} M_k^H - M_{k-1}^H &= \frac{1}{|I|} \sum_{i \in I} H_i(X_{(k)}) dN_i(X_{(k)}) \\ &\quad - \frac{1}{|I|} \sum_{i \in I} Y_i(X_{(k-1)}) \int_{X_{(k-1)}}^{X_{(k)}} H_i(t) [\{D_i \theta_0 + \beta_0^\top \mathbf{Z}_i\} dt + d\Lambda_0(t)], \end{aligned} \tag{3.155}$$

one from the jumps of $N_i(t)$ and the other from the compensator. Under Assumption 5-iv, there is almost surely no ties in the observed event times, so we have a bound

$$\mathbb{P}\left(\left| \frac{1}{|I|} \sum_{i \in I} H_i(X_{(k)}) dN_i(X_{(k)}) \right| \leq K_H/|I| \right) = \mathbb{P}\left(\frac{1}{|I|} \max_{i \in I} H_i(X_{(k)}) \leq K_H/|I| \right) = 1. \tag{3.156}$$

The compensator term in (3.155), second on the right hand side, is potentially unbounded. We have to study its tail distribution. Conditioning on $\mathcal{F}_{k-1}^H$, we calculate the distribution of $X_{(k)}$ as

$$\mathbb{P}(X_{(k)} \geq X_{(k-1)} + x | \mathcal{F}_{k-1}^H)$$

$$
\begin{aligned}
&= \prod_{i=1}^{|I|} \mathbb{P}(C_i \wedge T_i \geq X_{(k-1)} + x | C_i \wedge T_i \geq X_{(k-1)})^{Y_i(X_{(k-1)})} \\
&\leq \exp\left[-\sum_{i \in I} Y_i(X_{(k-1)}) \{ (D_i \theta_0 + \boldsymbol{\beta}_0^\top \mathbf{Z}_i) x + \Lambda_0(X_{(k-1)} + x) - \Lambda_0(X_{(k-1)}) \} \right]. \quad (3.157)
\end{aligned}
$$

We denote the function in the exponential index as

$$
A(x) = \sum_{i \in I} Y_i(X_{(k-1)}) \{ (D_i \theta_0 + \boldsymbol{\beta}_0^\top \mathbf{Z}_i) x + \Lambda_0(X_{(k-1)} + x) - \Lambda_0(X_{(k-1)}) \}. \quad (3.158)
$$

Note that $A(x)$ is nondecreasing, so its inverse $A^{-1}(x)$ is well defined. Next, we evaluate the tail distribution of the compensator term

$$
\begin{aligned}
&\mathbb{P}\left(\frac{1}{|I|} \sum_{i \in I} Y_i(X_{(k-1)}) \int_{X_{(k-1)}}^{X_{(k)}} H_i(t) [\{ D_i \theta_0 + \boldsymbol{\beta}_0^\top \mathbf{Z}_i \} dt + d\Lambda_0(t)] \geq x \middle| \mathcal{F}_{k-1}^H \right) \\
&\leq \mathbb{P}(K_H A(X_{(k)} - X_{(k-1)}) / |I| \geq x) \\
&= \mathbb{P}\{ X_{(k)} \geq X_{(k-1)} + A^{-1}(nx / K_H) \} \\
&\leq e^{-|I| x / K_H}. \quad (3.159)
\end{aligned}
$$

For $j \geq 2$, we calculate the moments

$$
\begin{aligned}
&\left| \mathbb{E}\left\{ (M_k^H - M_{k-1}^H)^j | \mathcal{F}_{k-1}^H \right\} \right| \\
&\leq \left[\mathbb{E}\left\{ \left| \frac{1}{|I|} \sum_{i \in I} H_i(X_{(k)}) dN_i(X_{(k)}) \right|^j \middle| \mathcal{F}_{k-1}^H \right\}^{\frac{1}{j}} \right. \\
&\quad \left. + \mathbb{E}\left\{ \left| \frac{1}{|I|} \sum_{i \in I} Y_i(X_{(k-1)}) \int_{X_{(k-1)}}^{X_{(k)}} H_i(t) [\{ D_i \theta_0 + \boldsymbol{\beta}_0^\top \mathbf{Z}_i \} dt + d\Lambda_0(t)] \right|^j \middle| \mathcal{F}_{k-1}^H \right\}^{\frac{1}{j}} \right]^j \\
&\leq \left[\frac{K_H}{|I|} + \left\{ \int_0^\infty e^{-|I| x^{\frac{1}{j}} / K_H} dx \right\}^{\frac{1}{j}} \right]^j \\
&= \left[\frac{K_H}{|I|} + \frac{K_H}{|I|} (j!)^{\frac{1}{j}} \right]^j \\
&\leq j! (2K_H / |I|)^j. \quad (3.160)
\end{aligned}
$$

This statement above proves (3.112), the first conclusion of the lemma.

For $\varepsilon > K_H/\sqrt{|I|}$, event

$$\sqrt{|I|}|M_k^H - M_{k-1}^H| > \varepsilon \tag{3.161}$$

occurs only if the following event occurs,

$$\begin{aligned} & \frac{1}{|I|}\sum_{i\in I} H_i(X_{(k)})dN_i(X_{(k)}) + \varepsilon/\sqrt{|I|} \\ < \; & \frac{1}{|I|}\sum_{i\in I} Y_i(X_{(k-1)}) \int_{X_{(k-1)}}^{X_{(k)}} H_i(t)[\{D_i\theta_0 + \beta_0^\top \mathbf{Z}_i\}dt + d\Lambda_0(t)]. \end{aligned} \tag{3.162}$$

We can bound

$$\begin{aligned} & \mathbb{E}\left\{(M_k^H - M_{k-1}^H)^2; \sqrt{|I|}|M_k^H - M_{k-1}^H| > \varepsilon\right\} \\ \leq \; & \mathbb{E}\Bigg\{ \left(\frac{1}{|I|}\sum_{i\in I} Y_i(X_{(k-1)}) \int_{X_{(k-1)}}^{X_{(k)}} H_i(t)[\{D_i\theta_0 + \beta_0^\top \mathbf{Z}_i\}dt + d\Lambda_0(t)]\right)^2 \\ & \quad \times I\left(\frac{1}{|I|}\sum_{i\in I} Y_i(X_{(k-1)}) \int_{X_{(k-1)}}^{X_{(k)}} H_i(t)[\{D_i\theta_0 + \beta_0^\top \mathbf{Z}_i\}dt + d\Lambda_0(t)] > \varepsilon/\sqrt{|I|}\right)\Bigg\} \\ \leq \; & \frac{\varepsilon^2}{|I|} e^{-\varepsilon\sqrt{|I|}/K_H} + \int_{\varepsilon^2/|I|}^{\infty} e^{-|I|\sqrt{x}/K_H} dx \\ = \; & \frac{\varepsilon^2|I| + 2K_H\sqrt{|I|} + 2K_H^2}{|I|^2} e^{-\varepsilon\sqrt{|I|}/K_H}. \end{aligned} \tag{3.163}$$

This proves (3.113), the other conclusion of the lemma. □

Proof of Lemma 37. Without loss of generality, we again prove the result for the nonnegative $H_i(t)$.

Let $X_{(1)}, \ldots, X_{(|I|)}$ be the order statistics of observed times. We define the sequence M_k^H, $k = 1, \ldots, n$, along with filtration $\mathcal{F}_k^* = \mathcal{F}_{I,X_{(k)}}$, as in Lemma 36. By Lemma 36, M_k^H is a $\mathcal{F}_k^*$-martingale satisfying (3.112), so we can apply the Bernstein's inequality for martingale

differences (as in Lemma 33). For $\varepsilon < 1$, we have

$$\mathbb{P}\left(\sup_{k=1,\ldots,|I|}\left|\frac{1}{|I|}\sum_{i\in I}\int_0^{X_{(i)}} H_i(t)dM_i(t)\right| > 4K_H\varepsilon\right) = \mathbb{P}\left(\sup_{k=1,\ldots,|I|}|M_k^H| > 4K_H\varepsilon\right) < 2e^{-|I|\varepsilon^2/2}. \tag{3.164}$$

This proves (3.114), the first result of the lemma.

The total variation of the integral with nonnegative $H_i(t)$'s can be written as

$$\begin{aligned}\bigvee_{t=0}^{\tau}\left\{\frac{1}{|I|}\sum_{i\in I}\int_0^t H_i(u)dM_i(u)\right\} &= \frac{1}{|I|}\sum_{i\in I}\bigvee_{t=0}^{\tau}\int_0^t H_i(u)dM_i(u)\\ &= \frac{2}{|I|}\sum_{i\in I}\int_0^{\tau} H_i(u)dN_i(u) - \frac{1}{|I|}\sum_{i\in I}\int_0^{\tau} H_i(u)dM_i(u)\end{aligned} \tag{3.165}$$

Hence, (3.115) the second result of the lemma follows directly from the first result (3.164).

To find the bound of variation between $X_{(k-1)}$ and $X_{(k)}$, simply consider that $H_i(t)$ is nonnegative while $dN_i(t)$ and $Y_i(t)\{(D_i\theta_0+\beta_0^\top\mathbf{Z}_i)dt+d\Lambda_0(t)\}$ are nonnegative measures. Hence, the extremal values in the intervals can be explicitly expressed as

$$\sup_{t\in[X_{(k-1)},X_{(k)})}\frac{1}{|I|}\sum_{i\in I}\int_0^t H_i(u)dM_i(u) = \frac{1}{|I|}\sum_{i\in I}\int_0^{X_{(k-1)}} H_i(u)dM_i(u) = M_{k-1}^H, \tag{3.166}$$

and

$$\inf_{t\in[X_{(k-1)},X_{(k)})}\frac{1}{|I|}\sum_{i\in I}\int_0^t H_i(u)dM_i(u) = \frac{1}{|I|}\sum_{i\in I}\int_0^{X_{(k)}-} H_i(u)dM_i(u) = M_k^H - \frac{H_{i_k}(X_{(k)})}{|I|}. \tag{3.167}$$

Therefore,

$$\sup_{t\in[0,\tau]}\left|\frac{1}{|I|}\sum_{i\in I}\int_0^t H_i(u)dM_i(u)\right| \le \sup_{k=1,\ldots,n}|M_k^H| + K_H/|I|. \tag{3.168}$$

For general $H_i(t)$, we simply decompose $H_i(t) = H_i^+(t) - H_i^-(t)$ and use the union bound. □

Proof of Lemma 38. The proof uses the conclusion of Lemma 37. For any $\varepsilon > 0$, we can find K_ε according to the tightness of $H_i(t)$ such that $\mathbb{P}\left(\max_{i=1,\ldots,n} \sup_{t\in[0,\tau]} |H_i(t)| > K_\varepsilon\right) < \varepsilon/2$. Define the truncated processes $H_{i,\varepsilon}(t) = (-K_\varepsilon) \vee \{H_i(t) \wedge K_\varepsilon\}$, which is still $\mathcal{F}_{I,t}$-adapted, as well as bounded by K_ε. By Lemma 37, we have

$$\mathbb{P}\left(\left|\frac{1}{|I|}\sum_{i\in I}\int_0^\tau H_{i,\varepsilon}(t)dM_i(t)\right| < 8K_\varepsilon\frac{\log(8/\varepsilon)}{\sqrt{|I|/2}}\right) > 1-\varepsilon/2.$$

Since $H_{i,\varepsilon}(t) = H_i(t)$ for all $i = 1,\ldots,n$ and $t \in [0,\tau]$ with probability at least $1-\varepsilon/2$, we have

$$\mathbb{P}\left(\left|\frac{1}{|I|}\sum_{i\in I}\int_0^\tau H_i(t)dM_i(t)\right| < 8K_\varepsilon\frac{\log(8/\varepsilon)}{\sqrt{|I|/2}}\right) > 1-\varepsilon.$$

The last equation defines the rate in (3.117). □

Proof of Lemma (39). Let B_i, $i \in I$, be independent Bernoulli random variables with rate $(H_i + K_H)/(2K_H)$. By a simple calculation, we have the following empirical distribution for B_iX_i

$$\frac{1}{|I|}\sum_{i\in I}B_iY_i(t) = \frac{1}{|I|}\sum_{i\in I}I(B_iX_i \geq t) \text{ and } \mathbb{E}\{B_iY_i(t)\} = \frac{1}{2K_H}\mathbb{E}\{H_iY_i(t)\} + \frac{1}{2}\mathbb{E}\{Y(t)\}. \quad (3.169)$$

We decompose

$$\begin{aligned}\frac{1}{|I|}\sum_{i\in I}H_iY_i(t) - \mathbb{E}\{H_iY_i(t)\} &= \frac{2K_H}{|I|}\sum_{i\in I}B_iY_i(t) - \mathbb{E}\{H_iY_i(t)\} - K_H\mathbb{E}\{Y(t)\} \\ &\quad - \frac{K_H}{|I|}\sum_{i\in I}Y_i(t) + K_H\mathbb{E}\{Y(t)\} \\ &\quad - \frac{2K_H}{|I|}\sum_{i\in I}\left(B_i - \frac{H_i+K_H}{2K_H}\right)Y_i(t). \quad (3.170)\end{aligned}$$

Applying the Dvoretzky-Kiefer-Wolfowitz (DKW) inequality (as in Lemma 34) to the first two terms in (3.170), we have

$$\mathbb{P}\left(\sup_{t\in[0,\tau]}\left|\frac{2K_H}{|I|}\sum_{i\in I}B_iY_i(t) - \mathbb{E}\{H_iY_i(t)\} - K_H\mathbb{E}\{Y(t)\}\right| > K_H\varepsilon\right) \leq 2e^{-|I|\varepsilon^2/2} \quad (3.171)$$

$$\text{and } \mathbb{P}\left(\sup_{t\in[0,\tau]}\left|\frac{K_H}{|I|}\sum_{i\in I}Y_i(t)-K_H\mathbb{E}\{Y(t)\}\right|>K_H\varepsilon\right)\leq 2e^{-|I|\varepsilon^2/2}. \tag{3.172}$$

Denote $X_{(i)}$, $i=1,\ldots,n$, as the order statistics of X_i's. We further decompose the third term in (3.170) as

$$\begin{aligned}\frac{2K_H}{|I|}\sum_{i\in I}\left(B_i-\frac{H_i+K_H}{2K_H}\right)Y_i(X_{(k)}) &= \frac{2K_H}{|I|}\sum_{i\in I}\left(B_i-\frac{H_i+K_H}{2K_H}\right)\\ &\quad -\frac{2K_H}{|I|}\sum_{i=1}^{k}\left(B_{(i)}-\frac{H_{(i)}+K_H}{2K_H}\right).\end{aligned} \tag{3.173}$$

By the Hoeffding's inequality (as in Lemma 31), we bound the first term in (3.173)

$$\mathbb{P}\left(\left|\frac{2K_H}{|I|}\sum_{i\in I}\left(B_i-\frac{H_i-K_H}{2K_H}\right)\right|>K_H\varepsilon\right)<2e^{-|I|\varepsilon^2/2}. \tag{3.174}$$

Let (i) be the i-th element in fold I. We define a filtration $\mathcal{F}_m^H=\sigma(\{(H_i,X_i):i\in I\}\cup\{B_{(i)}:i=1,\ldots,m\})$ under which we have the following martingale

$$M_m^H=\frac{2K_H}{|I|}\sum_{i=1}^{m}\left(B_{(i)}-\frac{H_{(i)}+K_H}{2K_H}\right). \tag{3.175}$$

By the Azuma's inequality (as in Lemma 32), we have

$$\mathbb{P}\left(\left|M_{|I|}^H\right|>2K_H\varepsilon\right)<2e^{-|I|\varepsilon^2/2}. \tag{3.176}$$

We finish the proof by putting the concentration inequalities together. □

Proof of Lemma 40. By Lemma 37, the probability that the event

$$\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau H_i(u)dM_i(u)<8K_H\varepsilon \tag{3.177}$$

is no less than $1-4e^{-n\varepsilon^2/2}$. We shall show that

$$\left|\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \widetilde{H}(t)H_i(t)dM_i(t)\right|<16K_HK_V\varepsilon+2K_HK_V/n \tag{3.178}$$

on such event.

By Lemma 37, the following function

$$\frac{1}{n}\sum_{i=1}^{n}\int_0^t H_i(u)dM_i(u) \tag{3.179}$$

has total variation bounded by $4K_H + 8K_H\varepsilon$ on event (3.178). As a result, we can apply the Helly-Bray integration by parts [Mur94]

$$\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \widetilde{H}(t)H_i(t)dM_i(t) = \frac{\widetilde{H}(\tau)}{n}\sum_{i=1}^{n}\int_0^\tau H_i(t)dM_i(t) - \int_0^\tau \left\{\frac{1}{n}\sum_{i=1}^{n}\int_0^t H_i(u)dM_i(u)\right\} d\widetilde{H}(t). \tag{3.180}$$

By Lemma 37, both terms have bound on event (3.178)

$$\left|\frac{\widetilde{H}(\tau)}{n}\sum_{i=1}^{n}\int_0^\tau H_i(t)dM_i(t)\right| \leq K_V \times 8K_H\varepsilon, \tag{3.181}$$

$$\left|\int_0^\tau \left\{\frac{1}{n}\sum_{i=1}^{n}\int_0^t H_i(u)dM_i(u)\right\} d\widetilde{H}(t)\right| \leq K_V \times (8K_H\varepsilon + 2K_H/n). \tag{3.182}$$

Plugging in the upper bounds to (3.180) finish the proof. □

Proof of Lemma 41. Since we assume that T_i and C_i are independent given D_i and $\mathbf{Z}_i$, we have

$$\mathbb{E}[Y_i(t)|D_i,\mathbf{Z}_i] = \mathbb{P}(T_i \wedge C_i \geq t|D_i,\mathbf{Z}_i) = \mathbb{P}(C_i \geq t|D_i,\mathbf{Z}_i)\mathbb{P}(T_i \geq t|D_i,\mathbf{Z}_i). \tag{3.183}$$

Under Assumption 5-ii, the censoring time is independent of treatment given covariates, so

$$\mathbb{P}(C_i \geq t|D_i,\mathbf{Z}_i) = \mathbb{P}(C_i \geq t|\mathbf{Z}_i) \tag{3.184}$$

is $\sigma\{\mathbf{Z}_i\}-$measurable. Under model (3.31),

$$\mathbb{P}(T_i \geq t|D_i,\mathbf{Z}_i) = e^{\int_0^t \lambda(t;D_i,\mathbf{Z}_i)dt} = e^{-D_i\theta_0 t - \int_0^t g_0(t;\mathbf{Z}_i)dt} = e^{-D_i\theta_0 t}\mathbb{P}(T_i \geq t|D_i=0,\mathbf{Z}_i). \tag{3.185}$$

Therefore, we have the following representation

$$\mathbb{E}[e^{D_i\theta_0 t}Y_i(t)|D_i,\mathbf{Z}_i] = \mathbb{P}(C_i \geq t|\mathbf{Z}_i)e^{-\int_0^t g_0(t;\mathbf{Z}_i)dt} = \mathbb{E}\{Y_i(t)|\mathbf{Z}_i, D_i = 0\}, \tag{3.186}$$

which is obviously $\sigma\{\mathbf{Z}_i\}-$measurable. By the tower property of conditional expectation, we can calculate the expectations for any $\sigma\{\mathbf{Z}_i\}$-measurable random variable U_i through

$$\begin{aligned}
& \mathbb{E}[\{D_i - \text{expit}(\boldsymbol{\gamma}_0^\top \mathbf{Z}_i)\}e^{D_i\theta_0 t}Y_i(t)U_i] \\
= \;& \mathbb{E}[\{D_i - \text{expit}(\boldsymbol{\gamma}_0^\top \mathbf{Z}_i)\}\mathbb{E}\{e^{D_i\theta_0 t}Y_i(t)|D_i,\mathbf{Z}_i\}U_i] \\
= \;& \mathbb{E}[\mathbb{E}\{D_i - \text{expit}(\boldsymbol{\gamma}_0^\top \mathbf{Z}_i)|\mathbf{Z}_i\}\mathbb{E}\{Y_i(t)|\mathbf{Z}_i, D_i = 0\}U_i] \\
= \;& 0.
\end{aligned} \tag{3.187}$$

We obtain the two equations in (3.121) by setting U_i above as 1 and $\mathbf{Z}_i$, respectively.

To deliver the concentration result (3.122), we decompose

$$\begin{aligned}
\frac{1}{|I|}\sum_{i\in I}\{D_i - \text{expit}(\boldsymbol{\gamma}_0^\top \mathbf{Z}_{\mathbf{1}i})\}e^{D_i\theta_0 t}Y_i(t) =& \frac{1}{|I|}\sum_{i\in I} D_i\{1 - \text{expit}(\boldsymbol{\gamma}_0^\top \mathbf{Z}_{\mathbf{1}i})\}Y_i(t) \\
& - \frac{1}{|I|}\sum_{i\in I}(1-D_i)\,\text{expit}(\boldsymbol{\gamma}_0^\top \mathbf{Z}_{\mathbf{1}i})\}Y_i(t).
\end{aligned}$$

Each coordinate of

$$\frac{1}{|I|}\sum_{i\in I} D_i\{1 - \text{expit}(\boldsymbol{\gamma}_0^\top \mathbf{Z}_{1,i})\}Y_i(t) \text{ and } \frac{1}{|I|}\sum_{i\in I}\text{expit}(\boldsymbol{\gamma}_0^\top \mathbf{Z}_{1,i})Y_i(t),$$

is bounded, so we can apply Lemma 39 to get

$$\begin{aligned}
&\sup_{t\in[0,\tau]}\left|e^{\theta_0 t}\frac{1}{|I|}\sum_{i\in I} D_i\{1 - \text{expit}(\boldsymbol{\gamma}_0^\top \mathbf{Z}_{1,i})\}Y_i(t) - e^{\theta_0 t}\mathbb{E}\left[D_i\{1 - \text{expit}(\boldsymbol{\gamma}_0^\top \mathbf{Z}_{1,i})\}Y_i(t)\right]\right| = O_p\left(n^{-\frac{1}{2}}\right), \\
&\sup_{t\in[0,\tau]}\left|\frac{1}{|I|}\sum_{i\in I}(1-D_i)\,\text{expit}(\boldsymbol{\gamma}_0^\top \mathbf{Z}_{\mathbf{1}i})\}Y_i(t) - \mathbb{E}\left[(1-D_i)\,\text{expit}(\boldsymbol{\gamma}_0^\top \mathbf{Z}_{\mathbf{1}i})\}Y_i(t)\right]\right| = O_p\left(n^{-\frac{1}{2}}\right).
\end{aligned} \tag{3.188}$$

From (3.121), we know that

$$e^{\theta_0 t}\mathbb{E}\left[D_i\{1-\text{expit}(\boldsymbol{\gamma}_0^\top \mathbf{Z}_{1,i})\}Y_i(t)\right] = \mathbb{E}\left[(1-D_i)\,\text{expit}(\boldsymbol{\gamma}_0^\top \mathbf{Z}_{1,i})Y_i(t)\right]. \tag{3.189}$$

Therefore, we have proved the first rate in (3.122) by combining (3.188) and (3.189). In the same way under Assumption 5-iii, we have a concentration result from Lemma 39 for each coordinate of $\frac{1}{|I|}\sum_{i\in I}\{D_i - \text{expit}(\boldsymbol{\gamma}_0^\top \mathbf{Z}_{\mathbf{1}i})\}e^{D_i\theta_0 t}Y_i(t)\mathbf{Z}_i$. We take the union bound to obtain the second rate in (3.122). □

Proof of Lemma 42. We provide the proof for the first result (3.123). The proof for the second result (3.124) is identical. Since the weights $w_i^1(\widehat{\boldsymbol{\gamma}})$ are nonnegative and $Y_i(t)$'s are non-increasing, we have lower bound

$$\frac{1}{|I|}\sum_{i\in I} w_i^1(\widehat{\boldsymbol{\gamma}})Y_i(t) \geq \frac{1}{|I|}\sum_{i\in I} D_i\{1-\text{expit}(\widehat{\boldsymbol{\gamma}}^\top \mathbf{Z}_{\mathbf{1}i})\}Y_i(\tau). \tag{3.190}$$

it is sufficient to show

$$\lim_{n\to\infty}\mathbb{P}\left(\frac{1}{|I|}\sum_{i\in I} w_i^1(\widehat{\boldsymbol{\gamma}})Y_i(\tau) > \varepsilon_Y/2\right) = 1. \tag{3.191}$$

We decompose

$$\begin{aligned}\frac{1}{|I|}\sum_{i\in I} w_i^1(\widehat{\boldsymbol{\gamma}})Y_i(\tau) &= \frac{1}{|I|}\sum_{i\in I} w_i^1(\boldsymbol{\gamma}_0)Y_i(\tau) \\ &\quad -\frac{1}{|I|}\sum_{i\in I} D_i\{\text{expit}(\widehat{\boldsymbol{\gamma}}^\top \mathbf{Z}_{\mathbf{1}i}) - \text{expit}(\boldsymbol{\gamma}_0^\top \mathbf{Z}_{\mathbf{1}i})\}Y_i(\tau).\end{aligned} \tag{3.192}$$

The first term in (3.192) has expectation bounded away from zero by Assumption 5-v

$$\mathbb{E}\{w_i^1(\boldsymbol{\gamma}_0)Y_i(\tau)\} = \mathbb{E}\{\text{Var}(D_i|\mathbf{Z}_{\mathbf{1}i})e^{\theta_0 t}\mathbb{E}\{Y_i(\tau)|\mathbf{Z}_{\mathbf{1}i}, D_i = 0\}\} \geq e^{-K_\theta \tau}\varepsilon_Y. \tag{3.193}$$

Since $w_i^1(\gamma_0)Y_i(\tau)$ are i.i.d. random variables in $[0,1]$, we have by Hoeffding's inequality (as in Lemma 31),

$$\frac{1}{|I|}\sum_{i\in I} w_i^1(\gamma_0)Y_i(\tau) = \mathbb{E}\{\mathrm{Var}(D_i|\mathbf{Z_1}_i)e^{\theta_0 t}\mathbb{E}\{Y_i(\tau)|\mathbf{Z_1}_i, D_i=0\}\} + O_p(n^{-1/2}) \geq e^{-K_\theta \tau}\varepsilon_Y + o_p(1). \tag{3.194}$$

By the Cauchy-Schwartz inequality, we have the bound for the second term in (3.192),

$$\begin{aligned} & \left|\frac{1}{|I|}\sum_{i\in I} D_i\{\mathrm{expit}(\widehat{\gamma}^\top \mathbf{Z_1}_i) - \mathrm{expit}(\gamma_0^\top \mathbf{Z_1}_i)\}Y_i(\tau)\right| \\ \leq & \sqrt{\frac{1}{|I|}\sum_{i\in I}\{\mathrm{expit}(\widehat{\gamma}^\top \mathbf{Z_1}_i) - \mathrm{expit}(\gamma_0^\top \mathbf{Z_1}_i)\}^2}. \end{aligned} \tag{3.195}$$

By the Markov's inequality, the bound above is of order $O_p\left(\mathcal{D}_\gamma(\widehat{\gamma},\gamma_0)\right) = o_p(1)$. Therefore, we have

$$\frac{1}{|I|}\sum_{i\in I} w_i^1(\widehat{\gamma})Y_i(\tau) + o_p(1) \geq \varepsilon_Y. \tag{3.196}$$

Hence, we obtain (3.191), a sufficient condition for (3.123). □

3.8 Acknowledgement

We would like to acknowledge our collaboration with Dr. James Murphy of the UC San Diego Department of Radiation Medicine and Applied Sciences on the linked Medicare-SEER data analysis project that motivated this work. We would also like to thank his group for help in preparing the data set.

Chapter 3, in full, is currently being prepared for submission for publication of the material. Hou, Jue; Bradic, Jelena; Xu, Ronghui. Estimating treatment effect for time-to-event outcome

with high-dimensional covariates in observational studies . The dissertation/thesis author is the primary investigator and author of this material.

Bibliography

[ABGK93] P. K. Andersen, O Borgan, R D Gill, and N Keiding. *Statistical Models Based on Counting Processes*. Springer, New York, USA, 1993.

[AG82] P. K. Andersen and R.D. Gill. Cox's regression model for counting processes: a large sample study. *The Annals of Statistics*, 10(4):1100–1120, 1982.

[ATW19] Susan Athey, Julie Tibshirani, and Stefan Wager. Generalized random forests. *Annals of Statistics*, 47(2):1148–1178, 2019.

[AWZ06] M. Asgharian, D. B. Wolfson, and X. Zhang. Checking stationarity of the incidence rate using prevalent cohort survival data. *Statistics in Medicine*, 25(10):1751–1767, 2006.

[BASB09] Harald Binder, Arthur Allignol, Martin Schumacher, and Jan Beyersmann. Boosting for high-dimensional time-to-event data with competing risks. *Bioinformatics*, 25(7):890–896, 2009.

[BC11] Alexandre Belloni and Victor Chernozhukov. l_1-penalized quantile regression in high-dimensional sparse models. *The Annals of Statistics*, 39(1):82–130, 2011.

[BCH13] Alexandre Belloni, Victor Chernozhukov, and Christian Hansen. Inference on Treatment Effects after Selection among High-Dimensional ControlsâĂă. *The Review of Economic Studies*, 81(2):608–650, 2013.

[BFJ11] Jelena Bradic, Jianqing Fan, and Jiancheng Jiang. Regularization for Cox's proportional hazards model with NP-dimensionality. *The Annals of Statistics*, 39(6):3092–3120, 2011.

[BKRF18] Anna Bellach, Michael R. Kosorok, Ludger Rüschendorf, and Jason P. Fine. Weighted NPMLE for the subdistribution of a competing risk. *Journal of the American Statistical Association*, page (online access), 2018.

[BKRW98] P Bickel, C A J Klaassen, Y Ritov, and J A. Wellner. *Efficient and Adaptive Estimation for Semiparametric Models*. Springer, New York, 1998.

[BM15] Sumanta Basu and George Michailidis. Regularized estimation in sparse high-dimensional time series models. *The Annals of Statistics*, 43(4):1535–1567, 2015.

[BR05] Heejung Bang and James M. Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61:962–972, 2005.

[Bro71] B.M. Brown. Martingale central limit theorems. *The Annals of Mathematical Statistics*, 42(1):59–66, 1971.

[BRT09] Peter J. Bickel, Ya'acov Ritov, and Alexandre B. Tsybakov. Simultaneous analysis of LASSO and Dantzig selector. *The Annals of Statistics*, 37(4):1705–1732, 2009.

[BS15] Jelena Bradic and Rui Song. Structured estimation for the nonparametric Cox model. *Electronic Journal of Statistics*, 9(1):492–534, 2015.

[BvdG11] Peter Bühlmann and Sara van de Geer. *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media, 2011.

[CBB+01] C. D. Chambers, S. R. Braddock, G. G. Briggs, A. Einarson, Y. R. Johnson, R. K. Miller, J. E. Polifka, L. K. Robinson, K. Stepanuk, and K. L. Jones. Postmarketing surveillance for human teratogenicity: a model approach. *Teratology*, 64:252–261, 2001.

[CCD+18] Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21:C1–C68, 2018.

[CF15] Haeran Cho and Piotr Fryzlewicz. Multiple-change-point detection for high dimensional time series via sparsified binary segmentation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 77(2):475–507, 2015.

[CIS99] M.-H. Chen, J. G. Ibrahim, and D. Sinha. A new bayesian model for survival data with a surviving fraction. *Journal of the American Statistical Association*, 94(447):909–919, 1999.

[CJX+13] C. D. Chambers, D. Johnson, R. Xu, Y. Luo, C. Louik, A. A. Mitchell, M. Schatz, and K. L. Jones. Risks and safety of pandemic h1n1 in uenza vaccine in pregnancy: Birth defects, spontaneous abortion, preterm delivery, and small for gestational age infants. *Teratology*, 31(44):5026–5032, 2013.

[CJXJ11] C. D. Chambers, D. Johnson, R. Xu, and K. L. Jones. Challenges and design of a prospective, observational cohort study to assess the risk of spontaneous abortion following administration of human papillomavirus (HPV) bivalent (types 16 and 18) recombinant vaccine. In *The 27th International Conference on Pharmacoepidemiology and Therapeutic Risk Management*, Chicago, IL, USA, 2011.

[CO84] D. R. Cox and D. Oakes. *Analysis of Survival Data*. Chapman & Hall, London, 1984.

[Con90] J. B. Conway. *A course in functional analysis*. Springer-Verlag, New York, second edition edition, 1990.

[Cox75] D. R. Cox. A note on data-splitting for the evaluation of significance levels. *Biometrika*, 62(2):441–444, 1975.

[CSWL17] Chyong-Mei Chen, Pao-Sheng Shen, James Cheng-Chung Wei, and Lichi Lin. A semiparametric mixture cure survival model for left-truncated and right-censored data. *Biometrical Journal*, 59:270–290, 2017.

[DBMM15] Ruben Dezeure, Peter BÃijhlmann, Lukas Meier, and Nicolai Meinshausen. High-dimensional inference: Confidence intervals, *p*-values and r-software hdi. *Statistical Science*, 30(4):533–558, 11 2015.

[DKW56] A. Dvoretzky, J. Kiefer, and J. Wolfowitz. Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *Annals of Mathematical Statistics*, 27(3):642–669, 1956.

[DLR77] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the American Statistical Association*, 39(1):1–38, 1977.

[Dur10] Rick Durrett. *Probability: Theory and Examples, 4th edition*. Cambridge University Press, 2010.

[ED87] Breslowm N. E. and N. E. Day. *Statistical Methods in Cancer Research*. IARC, Lyon, 1987.

[Far82] V. T. Farewell. The use of mixture models for the analysis of survival data with long-time survivors. *Biometrics*, 38:1041–1046, 1982.

[Far86] V. T. Farewell. Mixture models in survival analysis: Are they worth the risk? *Canadian Journal of Statistics*, 14(3):257–262, 1986.

[Far15] Max H. Farrell. Robust inference on average treatment effects with possibly more covariates than observations. *Journal of Econometrics*, 189:1–23, 2015.

[FG99] Jason P. Fine and Robert J. Gray. A proportional hazard model for the subdistribution of a competing risk. *Journal of the American Statistical Association*, 94:496–509, 1999.

[FHT10] Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software, Articles*, 33(1):1–22, 2010.

[FL01] Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360, 2001.

[FL10] Jianqing Fan and Jinchi Lv. A selective overview of variable selection in high dimensional feature space. *Statistica Sinica*, 20(1):101–148, 2010.

[FNL17] Ethan X Fang, Yang Ning, and Han Liu. Testing and confidence intervals for high dimensional proportional hazards models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79:1415–1437, 2017.

[GG12a] Stéphane Gaïffas and Agathe Guilloux. High-dimensional additive hazards models and the lasso. *Electronic Journal of Statistics*, 6:522–546, 2012.

[GG12b] Stéphane Gaïffas and Agathe Guilloux. High-dimensional additive hazards models and the lasso. *Electronic Journal of Statistics*, 6:522–546, 2012.

[GL96] S T Gross and T L Lai. Nonparametric estimation and regression analysis with left-truncated and right-censored data. *Journal of the American Statistical Association*, 91:1166–1180, 1996.

[GMX09] A Gamst, Donohue M, and R. Xu. Asymptotic properties and empirical evaluation of the *npmle* in the proportional hazards mixed-effects model. *Statistica Sinica*, 19:997–1011, 2009.

[GRS12] Anders Gorst-Rasmussen and Thomas Scheike. Coordinate descent methods for the penalized semiparametric additive hazards model. *Journal of Statistical Software, Articles*, 47(9):1–17, 2012.

[HBJT03] T. Hanson, E. J. Bedrick, W. O. Johnson, and M. C. Thurmond. A mixture model for bovine abortion and foetal survival. *Statistics in Medicine*, 22(10):1725–1739, 2003.

[HBX17] Jue Hou, Jelena Bradic, and Ronghui Xu. Fine-Gray competing risks model with high-dimensional covariates: estimation and Inference. *arXiv e-prints*, page arXiv:1707.09561, Jul 2017.

[HHWM02] M A Hernan, S Hernandez-Diaz, M M Werler, and A A Mitchell. Causal knowledge as a prerequisite for confounding evaluation: application to birth defects epidemiology. *American Journal of Epidemiology*, 155:176–184, 2002.

[HMX06] Jian Huang, Shuangge Ma, and Huiliang Xie. Regularized estimation in the accelerated failure time model with high-dimensional covariates. *Biometrics*, 62(3):813–820, 2006.

[Hoe63] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

[HPG+01] Linda C. Harlan, Arnold Potosky, Frank D. Gilliland, Richard Hoffman, Peter C. Albertsen, Ann S. Hamilton, J. W. Eley, Janet L. Stanford, and Robert A. Stephenson. Factors associated with initial therapy for clinically localized prostate cancer: Prostate cancer outcomes study. *Journal of the National Cancer Institute*, 93(24):1864–1871, 12 2001.

[HPH+18a] J. Hou, A. Paravati, J. Hou, R. Xu, and J. Murphy. High-Dimensional Variable Selection and Prediction under Competing Risks with Application to SEER-Medicare Linked Data. *Statistics in Medicine*, 37:3486–3502, 2018.

[HPH+18b] J Hou, A Paravati, J Hou, R Xu, and J Murphy. High-dimensional variable selection and prediction under competing risks with application to SEER-Medicare linked data. *Statistics in Medicine*, 37(4):3486–3502, 2018.

[HSY+13] Jian Huang, Tingni Sun, Zhiliang Ying, Yi Yu, and Cun-Hui Zhang. Oracle inequalities for the LASSO in the Cox model. *The Annals of Statistics*, 41(3):1142–1165, 2013.

[HYB+10] Jack Hadley, K. Robin Yabroff, Michael J. Barrett, David F. Penson, Christopher S. Saigal, and Arnold L. Potosky. Comparative effectiveness of prostate cancer treatments: Evaluating statistical adjustments for confounding in observational data. *Journal of the National Cancer Institute*, 103:1780–1793, 2010.

[Imb03] Guido W Imbens. Nonparametric estimation of average treatment effects under exogeneity: A review. Working Paper 294, National Bureau of Economic Research, 2003.

[IR14] Kosuke Imai and Marc Ratkovic. Covariate balancing propensity score. *Journal of the Royal Statistical Society, Serie B*, 76:243–263, 2014.

[JLS+17] Runchao Jiang, Wenbin Lu, Rui Song, Michael G. Hudgens, and Sonia Naprvavnik. Doubly robust estimation of optimal treatment regimes for survival dataâĂŤwith application to an hiv/aids study. *The Annals of Applied Statistics*, 11(3):1763–1786, 2017.

[JM14] Adel Javanmard and Andrea Montanari. Confidence intervals and hypothesis testing for high-dimensional regression. *Journal of Machine Learning Research*, 15:2869–2909, 2014.

[Joh83] S. Johansen. An extension of Cox's regression model. *International Statistics Review*, 51:165–174, 1983.

[Joh08] Brent A Johnson. Variable selection in semiparametric linear regression with censored data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(2):351–370, 2008.

[KC92] A. Y. Kuk and C.-H. Chen. A mixture model combining logistic regression with proportional hazards regression. *Biometrika*, 79(3):531–541, 1992.

[KJ08] Y.-J. Kim and M. Jhun. Cure rate model with interval censored data. *Statistics in Medicine*, 27(1):3–14, 2008.

[KLZ18] Suhyun Kang, Wenbin Lu, and Jiajia Zhang. On estimation of the optimal treatment regime with the additive hazards model. *Statistica Sinica*, 28(3):1539–1560, 2018.

[KP02] John D. Kalbfleisch and Ross L. Prentice. *The Statistical Analysis of Failure Time Data (2nd ed.)*. John Wiley & Sons, Inc., Hoboken, New Jersey, 2002.

[Lem16] Sarah Lemler. Oracle inequalities for the lasso in the high-dimensional multiplicative Aalen intensity model. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 52(2):981–1008, 2016.

[LL13] Wei Lin and Jinchi Lv. High-dimensional sparse additive hazards regression. *Journal of the American Statistical Association*, 108(501):247–264, 2013.

[LM07] Chenlei Leng and Shuangge Ma. Path consitent model selection in additive risk model via Lasso. *Statistics in Medicine*, 26:3753–3770, 2007.

[LMD88] S W Lagakos, L M Marraj, and V De Gruttola. Nonparametric analysis of truncated survival data, with application to aids. *Biometrika*, 75:515–523, 1988.

[Lou82] T. Louis. Finding the observed information matrix when using the em algorithm. *Journal of the Royal Statistical Society, Series B*, 44(2):226–233, 1982.

[LTS01] Chin-Shang Li, Jeremy MG Taylor, and Judy P Sy. Identifiability of cure models. *Statistics & Probability Letters*, 54(4):389–395, 2001.

[LY91] T L Lai and Z Ying. Estimating a distribution function with truncated and censored data. *Annals of Statistics*, 19:417–442, 1991.

[LY94] Dan Yu Lin and Zhiliang Ying. Semiparametric analysis of the additive risk model. *Biometrika*, 81(1):61–71, 1994.

[LY04] W. Lu and Z. Ying. On semiparametric transformation cure models. *Biometrika*, 91(2):331–343, 2004.

[LYSY04] Grace Lu-Yao, Therese A. Stukel, and Siu-Long Yao. Changing patterns in competing causes of death in men with prostate cancer: A population based study. *The Journal of Urology*, 171(6, Part 1):2285 – 2290, 2004. Part 1 of 2.

[Mas90] P. Massart. The tight constant in the dvoretzky-kiefer-wolfowitz inequality. *The Annals of Probability*, 18(3):1269–1283, 1990.

[MB06] Nicolai Meinshausen and Peter Bühlmann. High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, pages 1436–1462, 2006.

[MS08] R. Meister and C. Schaefer. Statistical methods for estimating the probability of spontaneous abortion in observational studies - analyzing pregnancies exposed to coumarin derivatives. *Reproductive Toxicology*, 26:31–35, 2008.

[Mur94] S. A. Murphy. Consistency in a proportional hazards model incorporating a random effect. *Annals of Statistics*, 22(2):712–731, 1994.

[Mur95] S. A. Murphy. Asymptotic theory for the frailty model. *Annals of Statistics*, 23(1):182–198, 1995.

[MvdV00] Susan Murphy and A van der Vaart. On profile likelihood. *Journal of the American Statistical Association*, 95:449–485, 2000.

[MY09] Nicolai Meinshausen and Bin Yu. Lasso-type recovery of sparse representations for high-dimensional data. *The Annals of Statistics*, pages 246–270, 2009.

[New90] Whitney K. Newey. Semiparametric efficient bounds. *Journal of Applied Econometrics*, 5(2):99–135, 1990.

[New94] Whitney K. Newey. The asymptotic variance of semiparametric estimators. *Econometrica*, 62(6):1349–1382, 1994.

[Ney59] Jerzy Neyman. Optimal asymptotic tests of composite statistical hypotheses. In Ulf Grenander, editor, *Probability and Statistics (The Harold Cramér Volume)*, pages 416–444. Almquist and Wiksells, Uppsala, Sweden, 1959.

[NQS10] J. Ning, J. Qin, and Y. Shen. Non-parametric tests for right-censored data with biased sampling. *Journal of the Royal Statistical Society, Series B*, 72:609–630, 2010.

[OWJ11] Guillaume Obozinski, Martin J. Wainwright, and Michael I. Jordan. Support union recovery in high-dimensional multivariate regression. *The Annals of Statistics*, 39(1):1–47, 2011.

[Pan00] Wei Pan. A multiple imputation approach to Cox regression with interval-censored data. *Biometrics*, 56(1):199–203, 2000.

[QNLS11] J. Qin, J. Ning, H. Liu, and Y. Shen. Maximum likelihood estimations and EM algorithms with length-biased data. *Journal of the American Statistical Association*, 106(496):1434–1449, 2011.

[RB00] Miguel Ángel Robins, James M.and Hernán and Babette Brumback. Marginal structural model and causal inference in epidemiology. *Epidemiology*, 11:550–560, 2000.

[RL02] D Rubin and R J A Little. *Statistical Analysis with Missing Data*. Wiley, New York, second edition edition, 2002.

[RM07] Greg Ridgeway and Daniel F. McCaffery. Comment: Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science*, 22(4):540–543, 2007.

[RMM+17] Greg Ridgeway, Daniel F. McCaffery, Andrew Morral, Morral Burgette, and Beth Ann Griffin. *Toolkit for Weighting and Analysis of Nonequivalent Groups: A tutorial for the* **twang** *package*. R Foundation for Statistical Computing, 2017.

[RMN92] James M. Robins, Steven D. Mark, and Whitney K. Newey. Estimating exposure effects by modelling the expectation of exposure conditional on confounders. *Biometrics*, 48:479–495, 1992.

[Ros87] Paul R. Rosenbaum. Model-based direct adjustment. *Journal of the American Statistical Association*, 82(398):387–394, 1987.

[RR83] Paul R. Rosenbaum and Donald B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

[RR84] Paul R. Rosenbaum and Donald B. Rubin. Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association*, 79(387):516–524, 1984.

[RR85] Paul R. Rosenbaum and Donald B. Rubin. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, 39(1):33–38, 1985.

[RR95] James M. Robins and Andrea Rotnitzky. Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, 90(429):122–129, 1995.

[RR01] James M. Robins and Andrea Rotnitzky. Comment on “inference for semiparametric models: Some questions and an answer". *Statistica Sinica*, 11(4):920–936, 2001.

[RRvdL00] James M. Robins, Andrea Rotnitzky, and Mark van der Laan. On profile likelihood: Comment. *Journal of the American Statistical Association*, 95(450):477–482, 2000.

[RTH+19] Paul Riviere, Christopher Tokeshi, Jiayi Hou, Vinit Nalawade, Reith Sarkar, Anthony J. Paravati, Melody Schiaffino, Brent Rose, Ronghui Xu, and James D. Murphy. Claims-based approach to predict cause-specific survival in men with prostate cancer. *JCO Clinical Cancer Informatics*, (3):1–7, 2019.

[RWL10] Pradeep Ravikumar, Martin J Wainwright, and John D Lafferty. High-dimensional ising model selection using l_1-regularized logistic regression. *The Annals of Statistics*, 38(3):1287–1319, 2010.

[Sas13] Igal Sason. On refined versions of the Azuma-Hoeffding inequality with applications in information theory. *ArXiv e-prints:1704.07989*, March 2013.

[SK03] S. K. Shevade and S. S. Keerthi. A simple and efficient algorithm for gene selection using sparse logistic regression. *Bioinformatics*, 19(17):2246–2253, 2003.

[SKD+15] Raj Satkunasivam, Andre E. Kim, Mihir Desai, Mike M. Nguyen, David I. Quinn, Leslie Ballas, Juan Pablo Lewinger, Mariana C. Stern, Ann S. Hamilton, Monish Aron, and Inderbir S. Gill. Radical prostatectomy or external beam radiation therapy vs no local therapy for survival benefit in metastatic prostate cancer: A seer-medicare analysis. *The Journal of Urology*, 194:378–385, 2015.

[SLFL14] Hokeun Sun, Wei Lin, Rui Feng, and Hongzhe Li. Network-regularized high-dimensional Cox regression for analysis of genomic data. *Statistica Sinica*, 24(3):1433–1459, 2014.

[ST00] J. P. Sy and J. M. Taylor. Estimation in a cox proportional hazards cure model. *Biometrika*, 56(1):227–236, 2000.

[SW92] Thomas A Severini and Wing Hung Wong. Profile likelihood and conditionally parametric models. *Annals of Statistics*, 20:1768–1802, 1992.

[Tan17] Zhiqiang Tan. Regularized calibrated estimation of propensity scores with model misspecification and high-dimensional data. *arXiv*, 2017.

[Tan18] Zhiqiang Tan. Model-assisted inference for treatment effects using regularized calibrated estimation with high-dimensional data. *arXiv*, 2018.

[Tho86] D. C. Thomas. Use of auxiliary information in fitting nonproportional hazards models. *Modern Statistical Methods in Chronic Disease Epidemiology*, pages 197–210, 1986.

[Tib96] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 58(1):267–288, 1996.

[TTLT16] Ryan J. Tibshirani, Jonathan Taylor, Richard Lockhart, and Robert Tibshirani. Exact post-selection inference for sequential regression procedures. *Journal of the American Statistical Association*, 111(514):600–620, 2016.

[Tur76] B. W. Turnbull. The empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society, Series B*, 38(3):290–295, 1976.

[Var85] Y. Vardi. Empirical distributions in selection bias models. *Annals of Statistics*, 13(1):178–203, 1985.

[VBC12] Stijn Vansteelandt, Maarten Bekaert, and Gerda Claeskens. On model selection and model misspecification in causal inference. *Statistical Methods in Medical Research*, 21:7–30, 2012.

[VD14] S. Vansteelandt and R.M. Daniel. On regression adjustment for the propensity score. *Statistics in Medicine*, 33(23):4053–4072, 2014.

[vdG07] Sara A. van de Geer. *The deterministic LASSO*. Technical Report 140. ETH Zürich, Switzerland, 2007.

[vdG08] Sara Anna van de Geer. High-dimensional generalized linear models and the lasso. *Annals of Statistics*, 36(2):614–645, 2008.

[vdGB09] Sara van de Geer and Peter Bühlmann. On the conditions used to prove oracle results for the Lasso. *Electronic Journal of Statistics*, 3:1360–1392, 2009.

[vdGB11] Sara van de Geer and Peter Bühlmann. *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer, 2011.

[vdGBRD14] Sara van de Geer, Peter Bühlmann, Ya'acov Ritov, and Ruben Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3):1166–1202, 2014.

[vdL14] Mark van der Laan. Targeted estimation of nuisance parameters to obtain valid statistical inference. *The International Journal of Biostatistics*, 10(1):29–57, 2014.

[VdVW96] A. W. Van der Vaart and J. A. Wellner. *Weak convergence and empirical processes*. Springer Series in Statistics. Springer-Verlag, New York, 1996.

[VV15] Karel Vermeulen and Stijn Vansteelandt. Bias-reduced doubly robust estimation. *Journal of the American Statistical Association*, 110(511):1024–1036, 2015.

[VX00] F. Vaida and R. Xu. Proportional hazards model with random effects. *Statistics in Medicine*, 19:3309–3324, 2000.

[Wai19] Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019.

[WC10] Daniel Westreich and Stephen R. Cole. Invited Commentary: Positivity in Practice. *American Journal of Epidemiology*, 171(6):674–677, 2010.

[WLL+17] Yan Wang, Mihye Lee, Pengfei Liu, Liuhua Shi, Zhi Yu, Yara Abu Awad, Antonella Zanobetti, and Joel D. Schwarts. Doubly robust additive hazards models to estimate effects of a continuous exposure on survival. *Epidemiology*, 28(6):771–779, 2017.

[WR09] Larry Wasserman and Kathryn Roeder. High dimensional variable selection. *The Annals of Statistics*, 37(5A):2178–2201, 2009.

[WWO+88] A. J. Wilcox, C. R. Weinberg, J. F. O'Connor, D. D. Baird, J. P. Schlatterer, R. E. Canfield, E. G. Armstrong, and B. C. Nisula. Incidence of early loss of pregnancy. *New England Journal of Medicine*, 319(4):189–194, 1988.

[XC11] R. Xu and C. Chambers. A sample size calculation for spontaneous abortion in observational studies. *Reproductive Toxicology*, 32(4):490–493, 2011.

[YBS18] Yi Yu, Jelena Bradic, and Richard J. Samworth. Confidence intervals for high-dimensional Cox models. *arXiv e-prints*, page arXiv:1803.01150, 2018.

[YBS19] Yi Yu, Jelena Bradic, and Richard J. Samworth. Confidence intervals for high-dimensional Cox models. *to appear in Statistica Sinica*, 2019.

[YLZ08] Guosheng Yin, Hui Li, and Donglin Zeng. Partially linear additive hazards regression with varying coefficients. *Journal of the American Statistical Association*, 103(483):1200–1213, 2008.

[YXM19] A. Ying, R. Xu, and J. Murphy. Two-stage residual inclusion for survival data and competing risks - an instrumental variable approach with application to SEER-Medicare linked data. *Statistics in Medicine*, 38(10):early view, 2019.

[ZL07] D Zeng and D Y Lin. Maximum likelihood estimation in semiparametric regression models with censored data. *Journal of the Royal Statistical Society, Series B*, 69:507–564, 2007.

[ZRXB11] Shuheng Zhou, Philipp Rütimann, Min Xu, and Peter Bühlmann. High-dimensional covariance estimation based on gaussian graphical models. *Journal of Machine Learning Research*, 12:2975–3026, 2011.

[ZS12] Min Zhang and Douglas E. Schaubel. Contrasting treatment-specific survival using double-robust estimators. *Statistics in Medicine*, 31(30):4255–4268, 2012.

[ZSZH17] Haixiang Zhang, Liuquan Sun, Yong Zhou, and Jian Huang. Oracle inequalities and selection consistency for weighted LASSO in high-dimensional additive hazards model. *Statistica Sinica*, 27:1903–1920, 2017.

[ZYI06] D. Zeng, G. Yin, and J. G. Ibrahim. Semiparametric transformation models for survival data with a cure fraction. *Journal of the American Statistical Association*, 101:670–684, 2006.

[ZZ14] Cun-Hui Zhang and Stephanie S. Zhang. Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):217–242, 2014.

[ZZYK15] Y. Q. Zhao, D. Zeng, M. Yuan, and M. R. Kosorok. Doubly robust learning for estimating individualized treatment with censored data. *Biometrika*, 102(1):151–168, 2015.