UNIVERSITY OF CALIFORNIA SAN DIEGO

Statistical Inference: Global Testing, Multiple Testing and Causal Inference in Survival Analysis

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

in

Mathematics

by

Andrew Ying

Committee in charge:

Professor Ery Arias-Castro, Chair
Professor Ronghui Xu, Co-Chair
Professor Loki Natarajan
Professor Yixiao Sun
Professor Wen-Xin Zhou

2020

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

____

____

____

____

Co-Chair

____

Chair

University of California San Diego

2020

DEDICATION

To my parents, who are always there for me.

EPIGRAPH

One thing only I know, and that is that I know nothing.

— Socrates

Cogito, ergo sum

— René Descartes

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to first express my special appreciation and thanks to my advisor Professor Ery Arias-Castro and my coadvisor Professor Ronghui (Lily) Xu. Professor Ery Arias-Castro was the first person who guided me into both advanced statistics and research. Without his patience with my constant failures in the early stage of researching, I would not stick to the end. Professor Ronghui (Lily) Xu not only showed me the mysterious world of survival analysis but also encouraged me to explore causal inference. She has been a role model of me for being rigorous and careful in researching. I cannot thank her enough for bringing me to different conferences and helping me to network.

I would like to thank Professor Ian Abramson, the very first teacher who brought me into the beautiful world of statistics. Without his stimulating and appealing courses in mathematical statistics, I would not switch my research interest into statistics. I would like to thank Professor Wen-Xin Zhou for his patience as my collaborator. I would like to thank Professor Jelena Bradic and Professor Dimitris Politis for showing me the beauty of other pieces of statistics including high dimensional statistics and time series. To the Mathematics Department, I appreciate all the resources and support provided. Thanks to Wilson Cheung, Holly Proudfoot, Scott Rollans, and Debbie Shon for all your patience and assistance. I would also like to thank Professor Qi-Man Shao for introducing me asymptotic theory in probability and statistics and having me as a visiting researcher in CUHK.

I am immensely grateful to all my friends at UCSD. I would to thank Fred Li, Yinzhuo Ding, Jiaqi Guo, Hanbo Li, Jingwen Liang, Marquis Hou, Yuchao Liu, Yaqing Wang, Ran Pan, Xiaolong Li, Dun Qiu, Nan Zou, Bowen Wang, Xiaoou Pan, Zonglin Han, Shiyun Chen, Rong Huang, Lin Zheng, Toni Gui, Selene Xu and Ruixuan Zhou. I would like to express my sincere gratitude to Yibo Wu, Qiqi Liu, Ka Li, Tian Gao, Yang Sun and Minwen Liu, my friends outside UCSD, who greatly enriched and fulfilled my leisure activities.

My sincere thanks also go to Dennis Sun, Yebin Tao, Yifei Yuan, Muxuan Liang, Xi

Chen, Nikos Ignatiadis, Qimeng Bao, who brought an amazing internship experience in Google. I would also thank Tian Liang, Jianjun Xie, who introduced me into industry world in CoreLogic for my first time in my life. I learned a lot about how top companies handle modern data using state-of-the-art techniques.

Last but not least, my sincere thanks go to my whole family. Thanks for my parents for bringing me into the world and nurturing me up. I thank my mom for devoting her lifetime in caring me and educating me. I thank my dad for being my first math teacher and constantly showing the beauty of mathematics.

Chapter 1, in full, is a reprint of the material as it appears in Electronic Journal of Statistics. Arias-Castro, Ery; Ying, Andrew. Detection of Sparse Mixtures: Higher Criticism and Scan Statistic, 13(1): 208-230, 2019. The dissertation author was the primary investigator and author of this paper.

Chapter 2, in full, has been submitted for publication of the material as it may appear in Bernoulli. Ying, Andrew; Zhou, Wen-Xin. On the Asymptotic Distribution of the Scan Statistic for Point Clouds. The dissertation author was the primary investigator and author of this paper.

Chapter 3, in full, is a reprint of the material as it appears in Statistics in Medicine. Ying, Andrew; Xu, Ronghui; Murphy, James.Two-Stage Residual Inclusion for Survival Data and Competing Risks - An Instrumental Variable Approach with Application to SEER-Medicare Linked Data, 38(10): 1775-1801, 2019. The dissertation author was the primary investigator and author of this paper.

Chapter 4, in full, has been submitted for publication of the material as it may appear in Journal of the American Statistical Association. Ying, Andrew; Xu, Ronghui; Chambers, Christina; Jones, Kenneth. Causal Effects on Birth Defects with Missing by Terathanasia. The dissertation author was the primary investigator and author of this paper.

VITA

2015	B. S. in Mathematics, Zhejiang University
2015-2020	Graduate Teaching Assistant, University of California San Diego
2018	C. Phil. in Mathematics, University of California San Diego
2020	Ph. D. in Mathematics, University of California San Diego

PUBLICATIONS

Arias-Castro, Ery, and Ying, Andrew. “Detection of Sparse Mixtures: Higher Criticism and Scan Statistic”, *Electronic Journal of Statistics*, 13(1): 208-230, 2019.

Ying, Andrew, Xu, Ronghui, and Murphy, James. “Two-Stage Residual Inclusion for Survival Data and Com- peting Risks - An Instrumental Variable Approach with Application to SEER-Medicare Linked Data”, *Statistics in Medicine*, 38(10): 1775-1801, 2019.

Chen, Shiyun, Ying, Andrew, and Arias-Castro, Ery. “A Scan Procedure for Multiple Testing”, *Journal of Statistical Planning and Inference*, Under Review, 2020.

Ying, Andrew, and Wen-Xin Zhou. “On the Asymptotic Distribution of the Scan Statistic for Point Clouds”, *Bernoulli*, Under Review, 2020.

Ying, Andrew, Xu, Ronghui, Chambers, Christina, and Jones, Kenneth. “Causal Effects on Birth Defects with Missing by Terathanasia”, *Journal of the American Statistical Association*, Under Review, 2020.

ABSTRACT OF THE DISSERTATION

Statistical Inference: Global Testing, Multiple Testing and Causal Inference in Survival Analysis

by

Andrew Ying

Doctor of Philosophy in Mathematics

University of California San Diego, 2020

Professor Ery Arias-Castro, Chair
Professor Ronghui Xu, Co-Chair

In Chapter 1, we consider the problem of detecting a sparse mixture as studied by Ingster (1997) and Donoho and Jin (2004). We consider a wide array of base distributions. In particular, we study the situation when the base distribution has polynomial tails, a situation that has not received much attention in the literature. Perhaps surprisingly, we find that in the context of such a power-law distribution, the higher criticism does not achieve the detection boundary. However, the scan statistic does. In Chapter 2, we derive the large-sample distribution of several variants of the scan statistic applied to a point process on an interval, which can be applied to

detect the presence of an anomalous interval with any length. The main ingredients in the proof are Kolmogorov's theorem, a Poisson approximation, and recent technical results by [KW14]. In Chapter 3, we consider causal inference in survival analysis in the presence of unmeasured confounders. Instrumental variable is an essential tool for addressing unmeasured confounding in observational studies. Two stage predictor substitution (2SPS) estimator and two stage residual inclusion(2SRI) are two commonly used approaches in applying instrumental variables. Recently 2SPS was studied under the additive hazards model in the presence of competing risks of time-to-events data, where linearity was assumed for the relationship between the treatment and the instrument variable. This assumption may not be the most appropriate when we have binary treatments. We consider the 2SRI estimator under the additive hazards model for general survival data and in the presence of competing risks, which allows generalized linear models for the relation between the treatment and the instrumental variable. We derive the asymptotic properties including a closed-form asymptotic variance estimate for the 2SRI estimator. We carry out numerical studies in finite samples, and apply our methodology to the linked Surveillance, Epidemiology and End Results (SEER) - Medicare database comparing radical prostatectomy versus conservative treatment in early-stage prostate cancer patients. In Chapter 4, we investigate the causal effects of etanercept (trade name Enbrel) on birth defects, a pharmaceutical that treats autoimmune diseases and recently went through the US FDA revised labeling for use in pregnancy, as the proportion of liveborn infants with major birth defects was higher for women exposed to etanercept compared to diseased etanercept unexposed women. An outstanding problem, which was not addressed in the data analysis leading up to the FDA relabeling, is the missing birth defect outcomes due to spontaneous abortion since, in accepted standard practice an infant or a fetus is assumed not to be malformed unless a defect is found. This led to likely bias (and missing not at random) because, according to the theory of "terathanasia", a defected fetus is more likely to be spontaneously aborted. In addition, the previous analysis stratified on live birth against spontaneous abortion, which was itself a post-exposure variable showing higher rate of

spontaneous abortion in the unexposed women, hence did not lead to causal interpretation of the stratified results. We aim to estimate and provide inference for the causal parameters of scientific interest, including the principal effects, making use of the missing data mechanism informed by terathanasia. During the process we also deal with complications in the data including left truncation, observational nature, and rare events. We report our findings which not only provide a more in-depth analysis than previously done on etanercept, but also shed light on how similar studies on causal effects of medication (or vaccine, other substances etc.) during pregnancy may be analyzed.

Chapter 1

Detection of Sparse Mixtures: Higher Criticism and Scan Statistic

1.1 Introduction

We consider the problem of detecting a sparse mixture. A simple variant of the problem can be formulated as follows. Let F be a continuous distribution function on the real line, and $\varepsilon \in (0, 1/2]$ and $\mu > 0$. The problem is to test

$$\mathcal{H}_0^n : X_1, \dots, X_n \overset{\text{iid}}{\sim} F, \tag{1.1}$$

versus

$$\mathcal{H}_1^n : X_1, \dots, X_n \overset{\text{iid}}{\sim} (1-\varepsilon)F(\cdot) + \varepsilon F(\cdot - \mu). \tag{1.2}$$

Mixtures models such as in (1.2) have been considered for quite some time, particularly in the context of robust statistics, where they are known as contamination models [HR09, Eq 1.22].

Rather, our contribution is in line with the testing problems studied by [Ing97] in the context of the normal sequence model, where F above corresponds to the standard normal distribution. In that setting, Ingster considered the following parameterization

$$\varepsilon = \varepsilon_n = n^{-\beta}, \qquad \mu = \mu_n = \sqrt{2r\log n}, \tag{1.3}$$

for some $\beta > 0$ and $r > 0$. The advantage of this parameterization is that, holding β and r fixed, the situation admits a relatively simple description. Indeed, since both the null and the alternative hypotheses are simple, by the Neyman-Pearson Lemma, the likelihood ratio test (set at level α) is most powerful. Ingster studied the large-sample behavior of this test procedure and discovered that, in the case where $\beta > 1/2$, when $r < \rho(\beta)$, the test is powerless in the sense of achieving power α, while when $r > \rho(\beta)$, the test was fully powerful in the sense of achieving power 1, where the function ρ is given by

$$\rho(\beta) := \begin{cases} \beta - 1/2, & 1/2 < \beta \leq 3/4, \\ (1 - \sqrt{1-\beta})^2, & 3/4 < \beta < 1. \end{cases} \tag{1.4}$$

Thus the existence of a detection boundary in the (β, r) plane given by $r = \rho(\beta)$. In such a situation, we will say that a test procedure 'achieves the detection boundary', or is 'first-order optimal' (or simply 'optimal'), if it is fully powerful when $r > \rho(\beta)$.

Such detection boundaries where derived for other models, for example, in [CW14, CJJ11, DJ04]. We also mention that the situation where $\beta \leq 1/2$ is also well-understood, but quite different, and will not be considered here. Most of the literature has focused on the more interesting setting where $\beta > 1/2$ and we do the same here.

1.1.1 Threshold tests

After determining what one can hope for, it becomes of interest to understand what one can achieve with less information. Indeed, the likelihood ratio test requires knowledge of all the quantities and objects defining the testing problem, in this case (F, ε, μ), and even in the present stylized setting we might want to know what can be done when some of this information is missing, in particular what defines the alternative, namely (ε, μ). (The case where F is also unknown has attracted much less attention. We discuss it in Section 1.5.)

When F is known, the problem is that of goodness-of-fit testing, albeit with alternatives of the form (1.2) in mind. [DJ04] opened this investigation with the analysis of various tests, including the max test based on $\max_i X_i$ and a variant of the Anderson-Darling test [AD52]. Seeing as a problem of multiple testing based on p-values defined as $U_i = 1 - F(X_i)$, the max test coincides with the Tippett-Šidák test combination test, while the Anderson-Darling test coincides with a proposal by Tukey called the higher criticism (HC). More recently, [MNS16] analyzed a goodness-of-fit (BJ) test proposed by [BJ79] in the same setting. For $t \in \mathbb{R}$, define

$$N_n(t) = \#\{i \in [n] : X_i \geq t\}. \tag{1.5}$$

We note that, under the null hypothesis, $N_n(t)$ is binomial with parameters $(n, 1 - F(t))$, which motivates the test that rejects for large values of

$$\sup_{t : F(t) \geq 1/2} \frac{N_n(t) - n(1 - F(t))}{\sqrt{nF(t)(1 - F(t)) + 1}}. \tag{1.6}$$

This is one of many possible variants of HC.[1]

Let $U_{(1)} \leq \cdots \leq U_{(n)}$ denote the ordered U_i's. We note that, under the null hypothesis, $U_{(i)}$ has the beta distribution with parameters $(i, n - i + 1)$, which motivates the definition of BJ, rejecting for small values of

$$\min_{i \in [n]} P_i, \tag{1.7}$$

[1] The constraint '$F(t) \leq 1/2$ can be replaced by $F(t) \leq \gamma$, where γ can be taken to be smaller, say $\gamma = 0.05$. The '+1' in the denominator is roughly equivalent to adding the constraint $F(t) \geq 1/n$, which [DJ04] recommend for reasons of stability. In any case, this variant performs as well (to first order) as any other variant of HC considered in the literature, at least in all the regimes commonly considered.

where $P_i := \mathsf{B}(U_{(i)}; i, n-i+1)$ and $\mathsf{B}(\cdot; a, b)$ denotes the distribution function of the beta distribution with parameters (a,b).

The verdict is the following. In the normal setting, HC and BJ achieve the detection boundary in the full range $\beta > 1/2$, while the max test is only able to achieve the detection in the upper half of the range $\beta > 3/4$. The same extends to other models, in particular to generalized Gaussian models where F has density proportional to $\exp(-|x|^a/a)$ for some $a > 1$. (The case $a \le 1$ is qualitatively different. HC and BJ are still first-order optimal while the max test is suboptimal everywhere.)

These tests are all threshold tests, where we define a threshold test as any test with a rejection region of the form $\bigcup_{t \in \mathcal{T}} \{N_n(t) \ge c_t\}$, for some subset $\mathcal{T} \subset \mathbb{R}$ and some critical values $c_t > 0$. More broadly, any combination test that we know of that is discussed in the multiple-testing literature is a threshold test. (This includes the tests proposed by Fisher, Lipták-Stouffer, Tippett-Šidák, Simes, and more.) Thus it might be of interest to understand what can be achieved with a threshold test. In this regard, it is useful to examine how one would optimize such an approach if one had perfect knowledge of the model. Let ϕ_t denote the test with rejection region $\{N_n(t) \ge c_t\}$, where

$$c_t := \min\big\{c \ge 0 : \mathbb{P}_0(N_n(t) \ge c) \le \alpha\big\}. \tag{1.8}$$

We define the oracle threshold test as the test ϕ_{t*}, where

$$t_* := \operatorname*{arg\,max}_{t \in \mathbb{R}} \mathbb{P}_1(N_n(t) \ge c_t), \tag{1.9}$$

with $\mathbb{P}_0$ denoting the distribution under the null (1.1) and $\mathbb{P}_1$ that under the alternative (1.2). (Here and elsewhere, α denotes the desired significance level.) Note that computing c_t only requires knowledge of F, while computing t_* requires knowledge of the entire model, namely (F, ε, μ). Thus the construction of the test ϕ_{t*} relies on the oracle knowledge of (ε, μ).

1.1.2 Scan tests

Detection problems arise in a variety of contexts and in very many applications. An important example is in spatial statistics (itself a rather wide area), where the detection of ‘hot spots’, meaning areas of unusually high concentration, has been considered for quite some time [Kul97]. An early contribution to this literature is that of [Nau65], who considered the distribution of the maximum number of points in an interval of given length (say ℓ) when the points are drawn iid from the uniform distribution on $[0,1]$. This would nowadays be referred to as the scan statistic and arises when testing the null that the points are uniformly distributed in $[0,1]$ against the (composite) alternative that there is an sub-interval of length ℓ with higher intensity. Settings where sub-interval length is unknown have been considered [ACDH05].

For $s \le t$, define $N_n[s,t] = \#\{i \in [n] : X_i \in [s,t]\}$ and $F[s,t] = F(t) - F(s)$. We note that, under the null hypothesis, $N_n[s,t]$ is binomial with parameters $(n, F[s,t])$, which motivates the test that rejects for large values of

$$\sup_{(s,t): F[s,t] \le 1/2} \frac{N_n[s,t] - nF[s,t]}{\sqrt{nF[s,t](1 - F[s,t]) + 1}}. \tag{1.10}$$

Although there are many possible variants, this is the one we will be working with.

We note that, under the null hypothesis, for any pair of indices $i < j$, $U_{(j)} - U_{(i)}$ has the beta distribution with parameters $(j - i, n - j + i + 1)$ — see [GC11, Th 11.1]. This motivates the definition of the scan test which rejects for small values of

$$\min_{0 \le i < j \le n+1} P_{i,j}, \tag{1.11}$$

where $P_{i,j} := \mathsf{B}(U_{(j)} - U_{(i)}; j - i, n - j + i + 1)$, $U_{(0)} := 0$, $U_{(n+1)} := 1$, $P_{0,n+1} := 1$.

In general, we define a scan test as any test with region rejection of the form

$$\bigcup_{(s,t)\in\mathcal{K}} \{N_n[s,t] \geq c_{s,t}\}, \tag{1.12}$$

where $\mathcal{K}$ is a subset of $\{(s,t) : s < t\}$ and $c_{s,t} \geq 0$ are critical values. Let $\phi_{s,t}$ denote the test with rejection region $\{N_n[s,t] \geq c_{s,t}\}$, where

$$c_{s,t} := \min\big\{c \geq 0 : \mathbb{P}_0(N_n[s,t] \geq c) \leq \alpha\big\}. \tag{1.13}$$

We define the oracle scan test as the test $\phi_{s_\bullet,t_\bullet}$, where

$$(s_\bullet,t_\bullet) := \arg\max_{s<t} \mathbb{P}_1(N_n[s,t] \geq c_{s,t}). \tag{1.14}$$

Indeed, $\phi_{s_\bullet,t_\bullet}$ relies on oracle knowledge of (ε,μ).

To the best of our knowledge, this is the first time that such tests are considered in the line of work that concerns us here with roots in the work of [Ing97] and [DJ04] — although a similar procedure is used in [CJL07] to estimate the contamination proportion ε. The main reason for considering these tests in the present context is that they happen to be first-order optimal, not only in the models considered in the literature (such as generalized Gaussian), but also in power-law models where F has fat tails (e.g., t distribution, Cauchy or Pareto), whereas threshold tests fail are suboptimal for such models. We observe that power-law models are mostly absent from this literature, although they are mentioned in [JSD$^+$05] in the context of an application in cosmology.

1.1.3 Content

For simplicity and the sake of clarity, we will focus on oracle-type, rather than likelihood ratio, performance bounds. The former are indeed more transparent and can be obtained under more generality and with simpler arguments. Also our main intention here is to compare what can be achieved with threshold tests compared to the more general scan tests, defined next, and comparing the corresponding oracle tests seems more appropriate.

In Section 1.2, we study the oracle threshold test and the oracle scan test. We then consider a number of models. In Section 1.3, we consider the two scan tests described above and compare them to the oracle scan test. In Section 1.4, we present the result of some numerical experiments that illustrate our theory. We briefly discuss the performance of the likelihood ratio test and that of nonparametric approaches in Section 1.5.

1.2 Oracle threshold test and oracle scan test

In this section we state and prove some basic results for the oracle threshold and oracle scan tests.

1.2.1 Power monotonicity

It is natural to guess that the testing (1.1) versus (1.2) becomes easier as the shift μ increases. This is indeed the case, at least from the point of view of both oracle tests.

Proposition 1.2.1.1. *The oracle threshold test has monotonic power in the shift.*

Proof. We assume that $\varepsilon > 0$ is fixed and let $\mathbb{P}_\mu$ denote the data distribution under the alternative (1.2). Take $\mu_1 \leq \mu_2$ and let t_k denote the oracle threshold (1.9) for μ_k, so that the oracle test for μ_k, meaning ϕ_{t_k}, has rejection region $\{N_n(t_k) \geq c_{t_k}\}$ and power $\pi_k := \mathbb{P}_{\mu_k}(N_n(t_k) \geq c_{t_k})$. Thus we need to show that $\pi_1 \leq \pi_2$. This is so because of the fact that, for any t, $N_n(t)$ is stochastically non-decreasing in μ, leading to

$$\pi_1 = \mathbb{P}_{\mu_1}(N_n(t_1) \geq c_{t_1}) \leq \mathbb{P}_{\mu_2}(N_n(t_1) \geq c_{t_1}) \leq \mathbb{P}_{\mu_2}(N_n(t_2) \geq c_{t_2}) = \pi_2, \tag{1.15}$$

where the last inequality is by construction of t_2 and c_2. □

Clearly, the oracle scan test has at least as much power as the oracle threshold test. Interestingly, it does not have monotonic power in general, although it does under some natural assumptions on the base distribution.

Proposition 1.2.1.2. *Assume that F, as a distribution, is unimodal. Then the oracle scan test has monotonic power in the shift.*

Proof. We stay with the setting and notation introduced in the proof of Proposition 1.2.1.1. Let $d \geq 0$ be smallest such that

$$F[s_1 + d, t_1 + \mu_2 - \mu_1] = F[s_1, t_1]. \tag{1.16}$$

The fact that F, as a distribution, is unimodal implies that $d \leq \mu_2 - \mu_1$. Now, under the null, by construction,

$$\mathbb{P}_0(N_n[s_1 + d, t_1 + \mu_2 - \mu_1] \geq c_{s_1,t_1}) = \mathbb{P}_0(N_n[s_1, t_1] \geq c_{s_1,t_1}) \leq \alpha. \tag{1.17}$$

On the other hand, under $\mathbb{P}_{\mu_1}$, $N_n[s_1, t_1]$ is binomial with parameters n and $q_1 := (1 - \varepsilon)F[s_1, t_1] + \varepsilon F[s_1 - \mu_1, t_1 - \mu_1]$, while under $\mathbb{P}_{\mu_2}$, $N_n[s_1 + d, t_1 + \mu_2 - \mu_1]$ is binomial with parameters n and

$$\begin{aligned} q_2 &:= (1 - \varepsilon)F[s_1 + d, t_1 + \mu_2 - \mu_1] + \varepsilon F[s_1 + d - \mu_2, t_1 + \mu_2 - \mu_1 - \mu_2] \\ &= (1 - \varepsilon)F[s_1, t_1] + \varepsilon F[s_1 + d - \mu_2, t_1 - \mu_1] \\ &\geq q_1, \end{aligned}$$

using the fact that $d \leq \mu_2 - \mu_1$. This explains the first inequality in the following derivation

$$\begin{aligned}
\pi_1 &= \mathbb{P}_{\mu_1}(N_n[s_1,t_1] \geq c_{s_1,t_1}) \\
&\leq \mathbb{P}_{\mu_2}(N_n[s_1+d,t_1+\mu_2-\mu_1] \geq c_{s_1,t_1}) \\
&\leq \mathbb{P}_{\mu_2}(N_n[s_2,t_2] \geq c_{s_2,t_2}) = \pi_2,
\end{aligned}$$

and the second inequality is by definition of (s_2,t_2). □

1.2.2 Performance bounds

We now provide necessary and sufficient conditions for the the oracle threshold test and the oracle scan test to be fully powerful in the large-sample limit ($n \to \infty$). We focus on the case where

$$n\varepsilon_n \to \infty, \qquad \sqrt{n}\varepsilon_n \to 0, \tag{1.18}$$

where the first condition implies that, under the alternative, the sample is indeed contaminated with probability tending to 1, while the second condition puts us in the regime corresponding to $\beta > 1/2$ under Ingster's parameterization (1.3).

Our analysis below is based on the following simple result, which is an immediate consequence of Chebyshev's inequality and the central limit theorem.

Lemma 1.2.2.1. *Suppose that we are testing $N \sim Bin(n, p_n)$ versus $N \sim Bin(n, q_n)$ where $p_n \leq 1/2$ and $p_n \leq q_n$, and consider the test at level α that rejects for large values of N — which is the most powerful test. It is asymptotically powerful if $n(q_n - p_n)^2/q_n \to \infty$, while it is asymptotically powerless if $n(q_n - p_n)^2/p_n \to 0$.*

Using Lemma 1.2.2.1, we easily obtain performance guarantees for the oracle threshold test and the oracle scan test.

Proposition 1.2.2.1. *The oracle threshold test is powerful if there is a sequence of thresholds (t_n) such that*

$$\begin{aligned} &n\varepsilon_n \bar{F}(t_n - \mu_n) \to \infty, \quad \textit{and} \\ &n\varepsilon_n^2 \bar{F}(t_n - \mu_n)^2 / \bar{F}(t_n) \to \infty. \end{aligned} \tag{1.19}$$

It is powerless if for any sequence of thresholds (t_n)

$$n\varepsilon_n^2 \bar{F}(t_n - \mu_n)^2 / \bar{F}(t_n) \to 0. \tag{1.20}$$

Proof. Let (t_n) denote a sequence of thresholds satisfying (1.19), and define $p_n = \bar{F}(t_n)$ and $q_n = (1-\varepsilon_n)\bar{F}(t_n) + \varepsilon_n \bar{F}(t_n - \mu_n)$. We know that $N_n(t_n) \sim \text{Bin}(n, p_n)$ under the null and $N_n(t_n) \sim \text{Bin}(n, q_n)$ under the alternative, with

$$n(q_n - p_n)^2 / q_n = \frac{n\varepsilon_n^2 (\bar{F}(t_n - \mu_n) - \bar{F}(t_n))^2}{(1-\varepsilon_n)\bar{F}(t_n) + \varepsilon_n \bar{F}(t_n - \mu_n)}.$$

If the second part of (1.19) holds, then necessarily $\bar{F}(t_n - \mu_n) \gg \bar{F}(t_n)$, since

$$\begin{aligned} n\varepsilon_n^2 \bar{F}(t_n - \mu_n)^2 / \bar{F}(t_n) &= \left[n\varepsilon_n^2 \bar{F}(t_n)\right] \left[\bar{F}(t_n - \mu_n)/\bar{F}(t_n)\right]^2 \\ &\leq (n\varepsilon_n^2) \left[\bar{F}(t_n - \mu_n)/\bar{F}(t_n)\right]^2, \end{aligned}$$

with $n\varepsilon_n^2 = o(1)$ by assumption. Hence,

$$\begin{aligned} n(q_n - p_n)^2 / q_n &\sim \frac{n\varepsilon_n^2 \bar{F}(t_n - \mu_n)^2}{(1-\varepsilon_n)\bar{F}(t_n) + \varepsilon_n \bar{F}(t_n - \mu_n)} \\ &\asymp n\varepsilon_n \bar{F}(t_n - \mu_n) \bigwedge n\varepsilon_n^2 \bar{F}(t_n - \mu_n)^2 / \bar{F}(t_n). \end{aligned}$$

Therefore, by Lemma 1.2.2.1, the sequence of tests (ϕ_{t_n}) has full power in the limit when (1.19) holds.

Now let (t_n) be any sequence of thresholds and consider the sequence of tests (ϕ_{t_n}). By

Lemma 1.2.2.1, it has power α in the limit since

$$n(q_n - p_n)^2/p_n \leq n\varepsilon_n^2 \bar{F}(t_n - \mu_n)^2/(1-\varepsilon_n)\bar{F}(t_n) \to 0, \tag{1.21}$$

where the convergence to 0 comes from (1.20). □

Remark 1.2.2.1. Note that the first part of (1.19) may be replaced by

$$n\bar{F}(t_n) \to \infty. \tag{1.22}$$

This is because this and $n\varepsilon_n^2 \bar{F}(t_n - \mu_n)^2/\bar{F}(t_n) \to \infty$ implies $n\varepsilon_n \bar{F}(t_n - \mu_n) \to \infty$.

Proposition 1.2.2.2. *The oracle scan test is powerful if there is a sequence of intervals* $([s_n, t_n])$ *such that*

$$\begin{gathered} n\varepsilon_n F[s_n - \mu_n, t_n - \mu_n] \to \infty, \quad \textit{and} \\ n\varepsilon_n^2 F[s_n - \mu_n, t_n - \mu_n]^2/F[s_n, t_n] \to \infty. \end{gathered} \tag{1.23}$$

It is powerless if for any sequence of intervals $([s_n, t_n])$

$$n\varepsilon_n^2 F[s_n - \mu_n, t_n - \mu_n]^2/F[s_n, t_n] \to 0. \tag{1.24}$$

The proof is completely parallel to that of Proposition 1.2.2.1 and is omitted.

1.2.3 Examples: generalized Gaussian models and more

We look at a number of models and in each case derive the performance of the oracle threshold and oracle scan tests, and compare that with the performance of the likelihood ratio test.

To place the results in line with the literature on the topic, we adopt Ingster's parameterization (1.3) for ε_n, in fact a softer version of that

$$\varepsilon = \varepsilon_n \sim n^{-\beta}, \tag{1.25}$$

for some fixed β. The parameterization of $\mu = \mu_n$ will depend on on the model.

To further simplify matters, we assume throughout that

$$\log \bar{F}(x) \sim -\varphi(x), \tag{1.26}$$

where $\varphi(x)$ is continuous and strictly increasing for x large enough. In that case, in view of Remark 1.2.2.1, we note that (1.19) is satisfied when

$$\begin{gathered} \log n - \varphi(t_n) \to \infty, \\ (1-2\beta)\log n + \varphi(t_n) - 2\varphi(t_n - \mu_n) \to \infty. \end{gathered} \tag{1.27}$$

Extended generalized Gaussian

This class of models is defined by the property that φ satisfies[2]

$$\varphi(ut)/\varphi(t) \to u^a, \quad t \to \infty, \quad \forall u \geq 0. \tag{1.28}$$

Here $a > 0$ parameterizes this class of models. This covers the generalized Gaussian models, which are often used as benchmarks in this line of work. It also covers the case where $\varphi(t) \sim t^a (\log t)^b$ where $b \in \mathbb{R}$ is arbitrary.

[2] It is tempting to consider a more general condition where there is a function ω on $\mathbb{R}_+$ such that $\lim_{t\to\infty} \varphi(ut)/\varphi(t) \to \omega(u)$ for all $u \geq 0$. However, as long as ω is not constant (equal to zero in that case), it can easily be shown that $\omega(u) = u^a$ for some $a > 0$.

For $a > 1$, define

$$\rho_a(\beta) = \begin{cases} (2^{1/(a-1)} - 1)^{a-1}(\beta - 1/2), & 1/2 < \beta < 1 - 2^{-a/(a-1)}, \\ (1 - (1-\beta)^{1/a})^a, & 1 - 2^{-a/(a-1)} \leq \beta < 1. \end{cases} \tag{1.29}$$

For $a \leq 1$, define

$$\rho_a(\beta) = 2(\beta - 1/2). \tag{1.30}$$

In addition to (1.25), assume that

$$\mu = \mu_n \text{ satisfies } \varphi(\mu_n) \sim r \log n, \text{ with } r \geq 0 \text{ fixed.} \tag{1.31}$$

Proposition 1.2.3.1. *The curve $r = \rho_a(\beta)$ in the (β, r) plane is the detection boundary that the oracle threshold test achieves.*

Proof. We focus on proving that the oracle threshold test achieves that boundary. A simple inspection of the arguments reveal that they are tight, so that this is the precise detection boundary that the test achieves. (See the proof of Proposition 1.2.4.2 for an example.)

We divide the proof into several cases.

Case 1: $a > 1$. Define $b = 2^{-1/(a-1)}$ and note that $0 < b < 1$.

Case 1.1: $1/2 < \beta < 1 - b^a$ *and* $r > \rho_a(\beta)$. Under these conditions, $\beta < 1/2 + r(1/b-1)^{-(a-1)}$, and in particular there is $\eta > 0$ such that

$$1 - 2\beta \geq -2r(1/b-1)^{-(a-1)} + \eta. \tag{1.32}$$

Setting $t_n = (1-b)^{-1}\mu_n$, by (1.28) and (1.31), we have the following

$$\varphi(t_n - \mu_n) = \big(rb^a/(1-b)^a + o(1)\big)\log n, \tag{1.33}$$

$$\varphi(t_n) = \big(r/(1-b)^a + o(1)\big)\log n. \tag{1.34}$$

By Proposition 1.2.1.1 we may focus on r small enough that $r/(1-b)^a < 1$. This is possible because $\rho_a(\beta) < (1-b)^a$ when $\beta < 1-b^a$, which we assume here. (This can be easily verified using the definition of b.) Assuming that r is as such, the first part of (1.27) is satisfied. For the second part, with (1.32), we have

$$\begin{aligned}
&(1-2\beta)\log n - 2\varphi(t_n-\mu_n) + \varphi(t_n)\\
&\geq \big[-2r(1/b-1)^{-(a-1)} + \eta - 2rb^a/(1-b)^a + r/(1-b)^a + o(1)\big]\log n\\
&= [\eta + o(1)]\log n \to \infty,
\end{aligned}$$

using the definition of b and simplifying. Thus the second part of (1.27) is also satisfied and the oracle threshold test is powerful.

Case 1.2: $1-b^a \leq \beta < 1$ *and* $r > \rho_a(\beta)$. Under these conditions, we have $1-\beta > (1-r^{1/a})^a$, and in particular there is $\eta > 0$ such that

$$1-\beta-\eta \geq (1-r^{1/a})^a \geq ((1-\eta)^{1/a} - r^{1/a})^a. \tag{1.35}$$

Set $t_n = (\frac{1}{r}(1-\eta))^{1/a}\mu_n$, we have the following

$$\varphi(t_n-\mu_n) = \big((1-\eta)^{1/a} - r^{1/a})^a + o(1)\big)\log n, \tag{1.36}$$

$$\varphi(t_n) = (1-\eta + o(1))\log n. \tag{1.37}$$

By looking at the speed of $\varphi(t_n)$, the first part of (1.27) is satisfied immediately. For the second part, with (1.32), we have

$$
\begin{aligned}
&(1-2\beta)\log n - 2\varphi(t_n-\mu_n)+\varphi(t_n)\\
&= (1-2\beta)\log n - 2\big((1-\eta)^{1/a}-r^{1/a})^a+o(1)\big)\log n + (1-\eta+o(1))\log n\\
&= 2\big[1-\beta-\eta/2-((1-\eta)^{1/a}-r^{1/a})^a+o(1)\big]\log n \to \infty.
\end{aligned}
$$

Thus the second part of (1.27) is also satisfied and the oracle threshold test is powerful.

Case 2: $a \leq 1$. By Proposition 1.2.1.1 we may restrict attention to the case where $2\beta - 1 < r < 1$. Here we set $t_n = \mu_n$. Then the first part in (1.27) is clearly satisfied. For the second part, notice that

$$
\begin{aligned}
&(1-2\beta)\log n - 2\varphi(t_n-\mu_n)+\varphi(t_n)\\
&= (1-2\beta)\log n + (r+o(1))\log n\\
&= [1-2\beta+r+o(1)]\log n \to \infty.
\end{aligned}
$$

This completes the proof. □

Thus, although the conditions are much more general here, the detection boundary is the same as in the corresponding generalized Gaussian model and, moreover, the oracle threshold test achieves that boundary.

Remark 1.2.3.1 (max test). In this class of models, it can be shown that the max test achieves the detection boundary over the upper range, meaning when $\beta \geq 1-2^{-a/(a-1)}$. In fact, $\rho^{\max}(\beta) := (1-(1-\beta)^{1/a})^a$ defines the detection boundary for the max test.

Other models

In the next few classes of models, φ satisfies

$$\frac{\varphi^{-1}(t)-\varphi^{-1}(vt)}{\lambda(t)} \to \omega(v), \quad t\to\infty, \quad \forall v\in(0,1]. \tag{1.38}$$

for some functions λ and ω, with the latter being non-increasing, continuous, and such that $\omega(1)=0$. This is actually also the case when $\varphi(t)\sim t^a(\log t)^b$ with $a>0$ and $b\in\mathbb{R}$, with $\lambda(t)=t^{1/a}(\log t)^{-b/a}$ and $\omega(v)=(1-v^{1/a})/a^{b/a}$.

Define

$$\rho(\beta)=\inf_{0<h<1-\beta}\big[\omega(h)-\omega(2\beta-1+2h)\big]. \tag{1.39}$$

In addition to (1.25), assume that

$$\mu=\mu_n\sim r\lambda(\log n), \quad r\geq 0 \text{ fixed}. \tag{1.40}$$

Proposition 1.2.3.2. *The curve $r=\rho(\beta)$ in the (β,r) plane is the detection boundary that the oracle threshold test achieves.*

Proof. We focus on proving that the oracle threshold test achieves that boundary.

Since $\omega(v)$ is continuous, we may define

$$h^*=\operatorname*{arg\,min}_{0\leq h\leq 1-\beta}\big[\omega(h)-\omega(2\beta-1+2h)\big]. \tag{1.41}$$

We focus on the case where $h^*<1-\beta$. In the case where $h^*=1-\beta$, the max test is powerful (Remark 1.2.3.2), and therefore so is the oracle threshold test. By Proposition 1.2.1.1 we may focus on the case where $r<\omega(h^*)$. With these assumptions and the fact that $\omega(h^*)-\omega(2\beta-1+$

$2h^*) = \rho(\beta) < r$, there is $\eta > 0$ be such that

$$2\beta - 1 + 2h^* + 2\eta < 1, \tag{1.42}$$

and

$$\omega(h^*) - \omega(2\beta - 1 + 2h^* + \eta) < r < \omega(h^*) - \omega(2\beta - 1 + 2h^* + 2\eta). \tag{1.43}$$

Define $t_n := \mu_n + \varphi^{-1}(h^* \log n)$. Using (1.38) multiple times, for n sufficiently large, we have the following

$$\begin{aligned}
\mu_n &= (r + o(1))\lambda(\log n) \\
&\le [\omega(h^*) - \omega(2\beta - 1 + 2h^* + 2\eta)]\lambda(\log n) \\
&= \varphi^{-1}(\log n) - \varphi^{-1}(h^* \log n) - \varphi^{-1}(\log n) + \varphi^{-1}((2\beta - 1 + 2h^* + 2\eta)\log n) \\
&= \varphi^{-1}((2\beta - 1 + 2h^* + 2\eta)\log n) - \varphi^{-1}(h^* \log n).
\end{aligned}$$

Hence, eventually, $t_n \le \varphi^{-1}((2\beta - 1 + 2h^* + 2\eta)\log n)$, implying that

$$\begin{aligned}
\log n - \varphi(t_n) &= \log n - (2\beta - 1 + 2h^* + 2\eta)\log n \\
&= [1 - (2\beta - 2 + 2h^* + 2\eta)]\log n \to \infty,
\end{aligned}$$

using (1.42). Thus the first part of (1.27) is satisfied.

Similarly, for n sufficiently large,

$$\begin{aligned}
\mu_n &= (r + o(1))\lambda(\log n) \\
&\ge [\omega(h^*) - \omega(2\beta - 1 + 2h^* + \eta)]\lambda(\log n) \\
&= \varphi^{-1}((2\beta - 1 + 2h^* + \eta)\log n) - \varphi^{-1}(h^* \log n),
\end{aligned}$$

so that, eventually, $t_n \geq \varphi^{-1}((2\beta - 1 + 2h^* + \eta)\log n)$, implying that

$$\begin{aligned}
&(1-2\beta)\log n - 2\varphi(t_n - \mu_n) + \varphi(t_n) \\
&\geq (1-2\beta)\log n - 2h^* \log n + (2\beta - 1 + 2h^* + \eta)\log n \\
&= \eta \log n \to \infty.
\end{aligned}$$

Thus the second part of (1.27) is satisfied. □

Remark 1.2.3.2 (max test). In the present situation, it can be shown that $\rho^{\max}(\beta) := \omega(1-\beta)$ defines the detection boundary for the max test.

Extended generalized Gumbel

This class of models is defined by $\varphi(t) = \exp(t^a)$ for some $a > 0$, which satisfies (1.38) with $\lambda(t) = \frac{1}{a}(\log t)^{1/a-1}$ and $\omega(v) = \log(1/v)$. In this case,

$$\mu = \mu_n \sim \frac{r}{a}(\log\log n)^{1/a-1}, \tag{1.44}$$

and the detection boundary is given by $r = -\log(1-\beta)$. Note that, at the detection boundary, $\mu_n \to \infty$ when $a > 1$; that $\mu_n \asymp 1$ when $a = 1$; and $\mu_n \to 0$ when $a < 1$.

Extended generalized Gumbel

This class of models is defined by $\varphi(t) = \exp((\log t)^a)$ for some $a > 1$, which satisfies (1.38) with $\lambda(t) = \frac{1}{a}(\log t)^{1/a-1}\exp((\log t)^{1/a})$ and $\omega(v) = \log(1/v)$. In this case,

$$\mu = \mu_n \sim \frac{r}{a}(\log\log n)^{1/a-1}\exp((\log\log n)^{1/a}), \tag{1.45}$$

and the detection boundary is given by $r = -\log(1-\beta)$ as in the previous class of models (since ω is the same).

Remark 1.2.3.3 (max test). Based on Remark 1.2.3.2, in the last two classes of models, the max test achieves the detection boundary over the whole β range. The same is true, more generally, when the infimum in (1.39) is at $h = 1-\beta$.

1.2.4 Examples: power-law models and more

In the next few classes of models, F satisfies

$$\log(F(t+v) - F(t)) \sim -\lambda(t), \quad t \to \infty, \quad \forall v \geq 0, \tag{1.46}$$

for some function λ which is increasing eventually and such that $\lambda(t) \to \infty$ as $t \to \infty$. This includes models where

$$\bar{F}(t) \propto t^{-a}(\log t)^b(1+o(1/t)), \quad t \to \infty, \tag{1.47}$$

with $a > 0$ and $b \in \mathbb{R}$, in which case (1.46) holds with $\lambda(t) = (a+1)\log t$. It also includes models where $\bar{F}(t) \propto (\log t)^{-a}(1+o(1/t\log t))$, with $a > 0$, in which case (1.46) holds with $\lambda(t) = \log t$, as well as other distribution with even slower decay.

In addition to (1.25), assume that

$$\mu = \mu_n \quad \text{satisfies} \quad \lambda(\mu_n) \sim r\log n, \quad r \geq 0 \text{ fixed.} \tag{1.48}$$

Proposition 1.2.4.1. *The curve $r = \rho(\beta) := 2\beta - 1$ in the (β, r) plane is the detection boundary that the oracle scan test achieves.*

Proof. We focus on proving that the oracle scan test achieves that boundary.

Fix r such that $r > 2\beta - 1$. Consider the interval $[s_n, t_n]$ with $s_n := \mu_n$ and $t_n := \mu_n + v$, where $v > 0$ is such that $F[0, v] > 0$. We need to verify that (1.23) holds. On the one hand, we have

$$n\varepsilon_n F[s_n - \mu_n, t_n - \mu_n] = n^{1-\beta} F[0, v] \to \infty, \tag{1.49}$$

because $\beta < 1$ by assumption. So the first part of (1.23) holds. On the other hand,

$$\begin{aligned} n\varepsilon_n^2 F[s_n - \mu_n, t_n - \mu_n]^2 / F[s_n, t_n] &= n^{1-2\beta} F[0, v]^2 / n^{r+o(1)} \\ &= n^{r+1-2\beta+o(1)} \to \infty, \end{aligned}$$

since $r > 2\beta - 1$. So the second part of (1.23) holds. □

We now show that threshold tests are suboptimal in the main class of models satisfying (1.46), namely (1.47). (The same happens to be true in other models with fat tails satisfying (1.46).) This is the main motivation for considering scan tests.

Proposition 1.2.4.2. *In a model satisfying* (1.47)*, and with the same parameterization* (1.48)*, the curve* $r = (1 + 1/a)(2\beta - 1)$ *in the* (β, r) *plane is the detection boundary that the oracle threshold test achieves.*

Proof. We first prove that the oracle threshold test achieves this detection boundary. By Proposition 1.2.1.1 we may assume that $r < 1 + 1/a$. Therefore, fix r such that $(1 + 1/a)(2\beta - 1) < r < 1 + 1/a$. Set the threshold $t_n = \mu_n + v$, where v is such that $\bar{F}(v) > 0$. We need to verify that (1.19) holds, and we do so via Remark 1.2.2.1. Note that $t_n \sim \mu_n = n^{r/(a+1)+o(1)}$. In particular,

$$n\bar{F}(t_n) \sim n\mu_n^{-a} (\log \mu_n)^b = n^{1-ar/(a+1)+o(1)} \to \infty, \tag{1.50}$$

and, by the same token,

$$n\varepsilon_n^2 \bar{F}(t_n - \mu_n)^2 / \bar{F}(t_n) \sim n^{1-2\beta} n^{-ar/(a+1)+o(1)} = n^{1-2\beta-ar/(a+1)+o(1)} \to \infty. \tag{1.51}$$

We now turn to proving that this is the statement boundary is the best that the oracle threshold test can hope for. For this, fix $r < (1+1/a)(2\beta-1)$. We need to verify (1.20). Suppose for contradiction that there is a sequence of thresholds, (t_n), such that (1.20) does not hold. By extracting a subsequence if needed, we may assume that

$$n\varepsilon_n^2 \bar{F}(t_n - \mu_n)^2/\bar{F}(t_n) \to \lambda \in (0,\infty]. \tag{1.52}$$

First, suppose that $\liminf t_n/\mu_n < \infty$. Extracting a subsequence if needed, we may assume that $t_n = O(\mu_n)$. In that case, we have

$$\begin{aligned} n\varepsilon_n^2 \bar{F}(t_n - \mu_n)^2/\bar{F}(t_n) &\le n\varepsilon_n^2/\bar{F}(t_n) \\ &\le n^{1-2\beta+o(1)} \mu_n^{-a+o(1)} \\ &= n^{1-2\beta-ar/(a+1)+o(1)} \to 0. \end{aligned}$$

Since this contradicts (1.52), we must have $\liminf t_n/\mu_n = \infty$, meaning that $t_n \gg \mu_n$. In that case, we have $\bar{F}(t_n - \mu_n) \sim \bar{F}(t_n)$, implying that

$$n\varepsilon_n^2 \bar{F}(t_n - \mu_n)^2/\bar{F}(t_n) \sim n\varepsilon_n^2 \bar{F}(t_n) \le n\varepsilon_n^2 \to 0. \tag{1.53}$$

This also contradicts (1.52). Since there is no other option, it must be that (1.52) cannot hold. We conclude that, indeed, (1.20) holds for any sequence of thresholds. □

1.3 Scan tests

In this section, we study the scan tests (1.10) and (1.11), and show that both of them do as well as the oracle scan test, at least to first-order in the asymptote where $n \to \infty$ and under the various parameterizations used in the previous section. We refer to (1.10) as the Stouffer scan test, as it is constructed as Stouffer's combination test [SSD$^+$49]; while we refer to (1.11) as the Tippett scan test, for similar reasons [Tip31].

1.3.1 Stouffer scan test

We study the Stouffer scan test (1.10). The main work goes into controlling this statistic under the null hypothesis. The limiting distribution of higher criticism can be derived from [Jae79] and the limiting distributions of some variants of the scan statistic are known under other models [Kab11, SAC16]. We will not pursue such a fine result here, but contend ourselves with a relatively rough upper bound.

Lemma 1.3.1.1. *Given observations $x_1, \dots, x_n$, the maximum in* (1.10) *is attained at some $(s,t) = (x_i, x_j)$.*

Proof. Define

$$R_n(s,t) := \frac{N_n[s,t] - nF[s,t]}{\sqrt{nF[s,t](1-F[s,t])+1}}. \tag{1.54}$$

Let $x_{(1)} \leq \cdots \leq x_{(n)}$ denote the ordered observations, and set $x_{(0)} = -\infty$ and $x_{(n+1)} = \infty$. It suffices to show that, for any $1 \leq i \leq j \leq n$ and any (s,t) such that $x_{(i-1)} < s \leq x_{(i)}$ and $x_{(j)} \leq t < x_{(j+1)}$, in addition to $F[s,t] \leq 1/2$, we have $R_n(x_{(i)}, x_{(j)}) \geq R_n(s,t)$. The crucial observation is that $N_n[s,t] = N_n[x_{(i)}, x_{(j)}]$ while $F[x_{(i)}, x_{(j)}] \leq F[s,t]$.

It is thus enough to show that the function $p \mapsto (a-p)/(p(1-p)+b)^{1/2}$ is decreasing over $[0,1/2]$ for any $a,b \geq 0$. This is so since this function has derivative $-(a(1-2p)+2b+p)/(p(1-p)+b)^{3/2}$. □

Theorem 1.3.1.1. *With S_n defined as the statistic* (1.10)*, we have*

$$\mathbb{P}_0(S_n \geq 3\log n) \to 0. \tag{1.55}$$

Proof. We place ourselves under the null hypothesis. Recall the definition of R_n in (1.54). By

Lemma 1.3.1.1 and the fact that $R_n(X_i,X_i)=1$ for all i, if $S_n \geq 3\log n$ necessarily $S_n = S_n^* := \max_{i\neq j} R_n(X_i,X_j)$. For any $i \neq j$, we have

$$R_n(X_i,X_j) \leq 2 + S_{i,j}, \tag{1.56}$$

with

$$\begin{aligned} S_{i,j} &:= \frac{N_{i,j} - 2 - (n-2)p_{i,j}}{\sqrt{(n-2)p_{i,j}(1-p_{i,j})+1}}, \\ N_{i,j} &:= N_n[X_i,X_j], \quad p_{i,j} := F[X_i,X_j]. \end{aligned} \tag{1.57}$$

The point of this reorganizing is that, given (X_i,X_j), $N_{i,j}-2 \sim \mathrm{Bin}(n-2,p_{i,j})$, and an application of Bernstein's inequality gives

$$\begin{aligned} \mathbb{P}_0(S_{i,j} \geq s \mid X_i,X_j) &\leq \exp\left(-\frac{s^2 b_{i,j}^2/2}{b_{i,j}^2 + s b_{i,j}/3}\right) \\ &\leq \exp\left(-\frac{s^2/2}{1+s/3}\right) \\ &\leq \exp(-s), \quad \forall s \geq 6, \end{aligned}$$

because $b_{i,j} := \sqrt{(n-2)p_{i,j}(1-p_{i,j})+1} \geq 1$. Thus, with the union bound, as $n \to \infty$, we have

$$\begin{aligned} \mathbb{P}_0(S_n \geq 3\log n) &= \mathbb{P}_0(\exists i \neq j : S_{i,j} + 2 \geq 3\log n) \\ &\leq \sum_{i<j} \mathbb{P}_0(S_{i,j} \geq 3\log n - 2) \leq n^2 \exp(-3\log n + 2) \to 0, \end{aligned}$$

which proves the statement. □

With Theorem 1.3.1.1, one obtains the following performance bound for the Stouffer scan test.

Corollary 1.3.1.1. *The Stouffer scan test is powerful if there is a sequence of intervals* $([s_n,t_n])$

such that

$$\begin{aligned} n\varepsilon_n F[s_n - \mu_n, t_n - \mu_n] &\gg \log n, \quad \textit{and} \\ n\varepsilon_n^2 F[s_n - \mu_n, t_n - \mu_n]^2 / F[s_n, t_n] &\gg (\log n)^2. \end{aligned} \tag{1.58}$$

Proof. By Theorem 1.3.1.1, the Stouffer scan test at level α is at least as powerful as the test $\{S_n \geq 3\log n\}$, eventually. Now, under the alternative, this test is powerful if we can prove that $p_n := F[s_n, t_n] \leq 1/2$ and $R_n(s_n, t_n) \geq 3\log n$. Define $p_n' := F[s_n - \mu_n, t_n - \mu_n]$ and $q_n := (1-\varepsilon_n)p_n + \varepsilon_n p_n'$, so that (1.58) can be expressed as

$$n\varepsilon_n p_n' \gg \log n \to \infty, \quad \text{and} \quad n\varepsilon_n^2 {p_n'}^2 / p_n \gg (\log n)^2 \to \infty. \tag{1.59}$$

That $p_n \leq 1/2$ is true, eventually, comes from the fact that

$$\infty \leftarrow n\varepsilon_n^2 {p_n'}^2 / p_n \leq n\varepsilon_n^2 / p_n, \tag{1.60}$$

with $n\varepsilon_n^2 \to 0$ by assumption, so that necessarily $p_n \to 0$. Note that this implies that $q_n \to 0$ also.

Given that $N_n[s_n, t_n]$ is a binomial distribution with parameters n and q_n, with $nq_n \geq np_n' \to \infty$ by the first part of (1.58), we have $N_n[s_n, t_n] = nq_n + O_P(\sqrt{nq_n(1-q_n)})$, and so

$$R_n(s_n, t_n) = \frac{n\varepsilon_n(p_n' - p_n) + O_P(\sqrt{nq_n(1-q_n)})}{\sqrt{np_n(1-p_n)+1}} \sim \frac{n\varepsilon_n p_n' + O_P(\sqrt{nq_n})}{\sqrt{np_n+1}}, \tag{1.61}$$

since $p_n' \gg p_n$, by the fact that

$$\infty \leftarrow n\varepsilon_n^2 {p_n'}^2 / p_n = n\varepsilon_n^2 p_n (p_n'/p_n)^2 = o(p_n'/p_n)^2. \tag{1.62}$$

In addition, the same conditions imply

$$\frac{n\varepsilon_n p_n'}{\sqrt{nq_n}} \asymp \sqrt{n\varepsilon_n^2 {p_n'}^2 / p_n} \bigvee n\varepsilon_n p_n' \to \infty, \tag{1.63}$$

so that

$$R_n(s_n,t_n) \sim_P n\varepsilon_n p_n'/\sqrt{np_n+1} \asymp_P \sqrt{n\varepsilon_n^2 {p_n'}^2/p_n} \bigvee n\varepsilon_n p_n' \gg \log n. \tag{1.64}$$

We conclude that $R_n(s_n,t_n) \geq 3\log n$ holds with probability tending to 1. □

With this performance bound, it is straightforward to verify that the Stouffer scan test performs as well as the oracle scan test to first order, at least in the context of the parameterization used in the models studied in Section 1.2.3 and Section 1.2.4. This comes from the fact that, in context of these sections, the quantity appearing in (1.58) increases as a (fixed) positive power of n under the alternative. We formalize this into the following statement, left without formal proof.

Corollary 1.3.1.2. *The Stouffer scan test achieves the oracle scan detection boundary in all the settings considered in Section 1.2.3 and Section 1.2.4.*

1.3.2 Tippett scan test

We study the Tippett scan test (1.11), which we denote by T_n. We control this statistic under the null hypothesis by a simple application of the union bound. A more refined control seems possible in view of [MNS16], where the limiting distribution of (1.7) is obtained.

Proposition 1.3.2.1. *With T_n defined as the statistic* (1.11)*, we have*

$$\mathbb{P}_0(T_n \leq 1/n^3) \to 0. \tag{1.65}$$

Proof. Under the null, each $P_{i,j}$ is uniformly distributed in $[0,1]$. Thus the union bound gives

$$\mathbb{P}_0(T_n \leq 1/n^3) \leq n^2\, \mathbb{P}_0(P_{i,j} \leq 1/n^3) = n^2/n^3 = 1/n \to 0, \tag{1.66}$$

which concludes the proof. □

Thus most of the work goes into controlling the statistic under the alternative. We do so by bounding the Tippett scan statistic by an expression that resembles that of the Stouffer scan statistic. We make use of the following simple concentration bound.[3]

Lemma 1.3.2.1. *For $k \in [n]$,*

$$\mathsf{B}(u;k,n-k+1) \le \exp\left(-\frac{(k-nu)^2/2}{nu(1-u)+(k-nu)/3}\right), \quad 0 \le u \le k/n. \tag{1.67}$$

Proof. Let $U_{k:n}$ denote the k-th order statistic of an iid sample of size n from the uniform distribution on $[0,1]$. For $u \in [0,1]$ such that $nu \le k$, we have

$$\mathsf{B}(u;k,n-k+1) = \mathbb{P}(U_{k:n} \le u) = \mathbb{P}(\mathrm{Bin}(n,u) \ge k),$$

and we conclude with an application of Bernstein's inequality. □

Proposition 1.3.2.2. *The Tippett scan test is powerful if there is a sequence of intervals $([s_n,t_n])$ such that*

$$\begin{aligned} n\varepsilon_n F[s_n-\mu_n, t_n-\mu_n] &\gg \sqrt{\log n}, \\ n\varepsilon_n^2 F[s_n-\mu_n, t_n-\mu_n]^2/F[s_n,t_n] &\gg \log n. \end{aligned} \tag{1.68}$$

Proof. Recall that $T_n = \min_{i<j} P_{i,j}$ and the expression of $P_{i,j}$. Thus applying Lemma 1.3.2.1 gives

$$T_n \le 1/n^3 \;\Leftrightarrow\; \max_{i<j} \frac{(j-i-V_{i,j})_+^2/2}{nV_{i,j}(1-V_{i,j})+(j-i-V_{i,j})_+/3} \ge 3\log n, \tag{1.69}$$

where $V_{i,j} := U_{(j)} - U_{(i)}$, after taking a logarithm.

Moreover, $V_{i,j} = F[X_{(n-j+1)}, X_{(n-i+1)}]$ and $j-i = N_n[X_{(n-j+1)}, X_{(n-i+1)}] - 1$, yielding

$$T_n \le 1/n^3 \;\Leftrightarrow\; \max_{i \ne j} \frac{(N_{i,j}-1-np_{i,j})_+^2}{np_{i,j}(1-p_{i,j})+(N_{i,j}-1-np_{i,j})_+} \ge 6\log n, \tag{1.70}$$

[3] Many things are known about the beta distribution and order statistics in general, but we could not immediately find such a simple bound.

with the notation of (1.57). The latter inequality holds when there is $i \neq j$ such that

$$N_{i,j} - 1 - np_{i,j} \geq 12 \log n \quad \text{and} \quad \frac{N_{i,j} - 1 - np_{i,j}}{\sqrt{np_{i,j}(1 - p_{i,j})}} \geq \sqrt{12 \log n}, \tag{1.71}$$

which is the case when

$$np_{i,j} \geq \sqrt{12 \log n} \quad \text{and} \quad \frac{N_{i,j} - 1 - np_{i,j}}{\sqrt{np_{i,j}(1 - p_{i,j})}} \geq \sqrt{12 \log n}. \tag{1.72}$$

Let (s_n, t_n) be as in the statement and let (i, j) be such that $U_{(i)} \leq s < U_{(i+1)}$ and $U_{(j-1)} < t \leq U_{(j)}$. By construction, $p_{i,j} \geq F[s_n, t_n]$, so that the first part of (1.68) implies that the first part of (1.72) holds eventually. We also have $N_{i,j} \geq N_n[s_n, t_n] - 2$, so that

$$\frac{N_{i,j} - 1 - np_{i,j}}{\sqrt{np_{i,j}(1 - p_{i,j})}} \geq \frac{N_n[s_n, t_n] - 3 - nF[s_n, t_n]}{\sqrt{nF[s_n, t_n](1 - F[s_n, t_n])}}, \tag{1.73}$$

and the quantity on the RHS is controlled using the second part (1.68) exactly as in the proof of Proposition 1.2.2.2. □

Here too, these results make it straightforward to verify that the Tippett scan test performs as well as the oracle scan test (to first order) in the models and regimes seen earlier, leading us to state the following (left without a formal proof).

Corollary 1.3.2.1. *The Tippett scan test achieves the oracle scan detection boundary in all the settings considered in Section 1.2.3 and Section 1.2.4.*

1.4 Numerical experiments

We performed small-scale numerical experiments to probe our theory. We generated Student t-distributions with varying numbers of degrees of freedom, df $= 0.5, 1, 2, 5\}$. Recall that the Student t-distribution with k degrees of freedom has density $\propto (1 + x/k)^{-(k+1)/2}$. We considered three different scenarios with varying sparsity exponents, $\beta = 0.6, 0.7, 0.8$. The sample size was set to $n = 30,000$. We compared the higher criticism test, the Berk-Jones test, the Stouffer scan test, and the Tippett scan test in each of these settings. We repeat each setting 200 times. See Figure 1.1, Figure 1.2, and Figure 1.3.

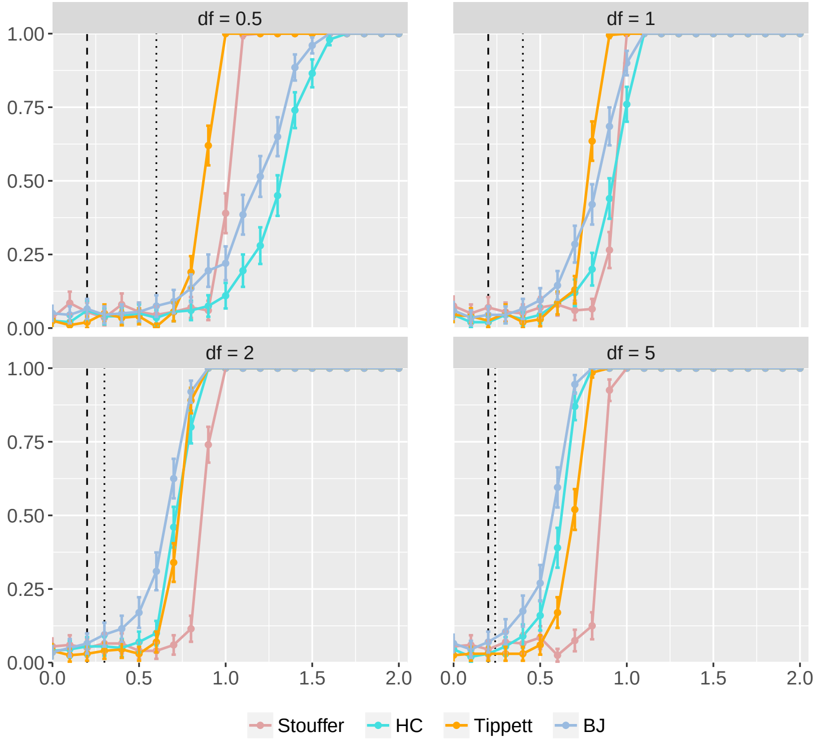

Figure 1.1: Here $\beta = 0.6$, the x-axis represents r in the parameterization (1.48), y-axis the power of the tests identified in the legend. Each subfigure corresponds to a Student t-distribution with the specified number of degrees of freedom. The black dashed vertical line corresponds to the oracle scan detection boundary established in Proposition 1.2.4.1, while the dotted line corresponds to the oracle threshold detection boundary established in Proposition 1.2.4.2.

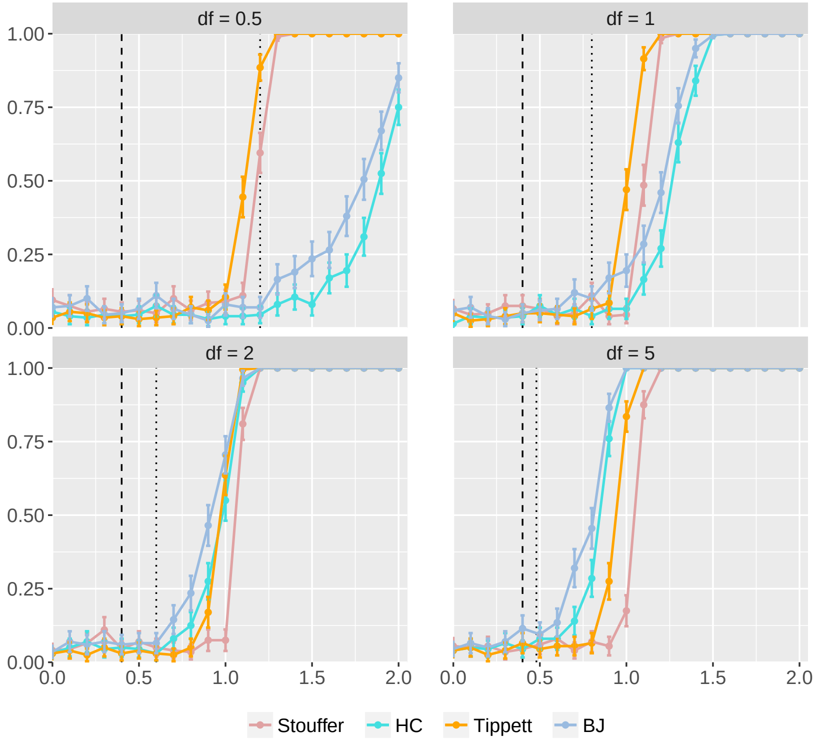

Figure 1.2: Here $\beta = 0.7$, otherwise, see Figure 1.1 for more details.

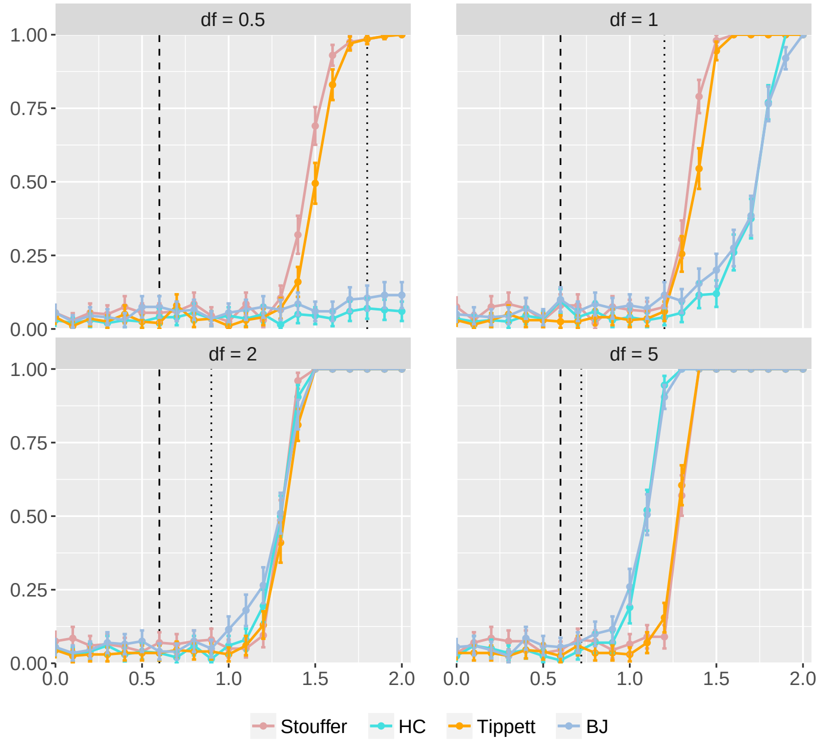

Figure 1.3: Here $\beta = 0.8$, otherwise, see Figure 1.1 for more details.

As the theory predicts, We can check that when the number of degrees of freedom is smaller, implying that the base distribution has fatter tails, the scan procedures dominate the threshold procedure. The threshold procedures become dominant as the tails become lighter. This is so at this particular sample size as, in principle, our theory indicates that with a larger sample size, the scan procedures would still dominate. The transition from powerless to powerful takes place at a larger effect size than predicted by the theory, which is also explain by the limited sample size.[4]

1.5 Discussion

While scan tests are commonly used in a number of detection problems, threshold tests are almost exclusively used in multiple testing situations. The main purpose of our work here

[4] The scan tests have computational complexity of order $O(n^2)$, which has limited the scale of our experiments.

was to reveal that scan tests can improve on threshold tests in somewhat standard multiple testing settings, particularly when the null distribution (F in the paper) has heavy tails.

Likelihood ratio performance bounds

Given our main objective, it was more natural to consider oracle-type performance bounds rather than using the likelihood ratio performance as benchmark. We can say nonetheless that, for representative models, the oracle threshold boundaries stated in Proposition 1.2.3.1 and Proposition 1.2.3.2 match those of the likelihood ratio test — for example, this is true of generalized Gaussian models where F has density of the form $f(t) \propto \exp(-|t|^a)$ for some $a > 0$. The same is true of the oracle scan boundary stated in Proposition 1.2.4.1 — for example, this is true of power law models where F has density of the form $f(t) \propto (1+|t|^a)^{-1}$ for some $a > 0$.

Nonparametric approaches

[ACCTW17] consider the situation where the null distribution, F, is symmetric about 0 but otherwise unknown. They suggest two tests for symmetry: the CUSUM sign test and the tail-run test, which are meant to be the nonparametric equivalent of the higher criticism test and the tail-run sign test, respectively. Back-of-the-envelope calculations seem to indicate that these nonparametric tests achieve the same detection boundaries as their parametric counterparts in all the settings considered here.

Multiple testing

In separate work [CYAC18], we uncover a similar phenomenon in the context of multiple testing, where the goal is maximizing the number of rejections while controlling the false discovery rate (FDR). Indeed, in a similar mixture model, standard in that literature at least since the work of [GW02, GW$^+$04], we find that with heavy tail distributions, scanning can improve on thresholding (what the procedure of [BH95] does). This is established in the context of the asymptotic framework of [GW02, GW$^+$04], which is different than the one considered here in that the mixture proportion, ε, does not converge to zero with the sample size. However, we expect this to extend to the present asymptotic model.[5]

[5] The present asymptotic model has been considered in the context of multiple testing, in particular in some of our own recent work [ACC17, CAC17].

1.6 Acknowledgement

Chapter 1, in full, is a reprint of the material as it appears in Electronic Journal of Statistics. Arias-Castro, Ery; Ying, Andrew. Detection of Sparse Mixtures: Higher Criticism and Scan Statistic, 13(1): 208-230, 2019. The dissertation author was the primary investigator and author of this paper.

Chapter 2

On the Asymptotic Distribution of the Scan Statistic for Point Clouds

2.1 Introduction

The study of the scan statistic dates back[1] to [Nau65], who derived the probability that an interval of a certain length contains a certain fraction of independent and identically distributed (iid) samples from the uniform distribution on $[0,1]$. Specifically, let $U_1,\dots,U_n$ be iid random variables from $\mathrm{Unif}(0,1)$ with empirical distribution function denoted by F_n, and let h be the length of the underlying interval of interest, Naus studied the distribution of

$$\sup_{0\le a\le 1} F_n(a+h)-F_n(a). \tag{2.1}$$

Knowing this distribution is essential to calibrating the scan statistic in the context of detecting, in a uniform background, the presence of an interval of a certain length with an unusually high density of points. This is considered today a quintessential detection problem, with applications in the detection of disease clusters [BN91] and syndromic surveillance [HMD$^+$04], among many

[1] Naus himself cites even earlier work in the 1940's by [Sil45], [Ber45], and [Mac48].

others [GNW$^+$01, GPW09, GB12, GK18].

In practice, even in the simplest case where only a single anomalous interval may be present, the length of that interval is almost always unknown. In that case, it is natural to consider intervals of various lengths, but standardize the counts, leading to

$$\sup_{0\le a\le 1}\sup_{h_-\le h\le h_+}\frac{\sqrt{n}(F_n(a+h)-F_n(a)-h)}{\sqrt{h(1-h)}}. \tag{2.2}$$

This can be seen to approximate the likelihood ratio test [Kul97]. The parameters h_- and h_+ limit the search to intervals that are neither too short and nor too large. The main goal of this paper is to derive the asymptotic (as $n\to\infty$) distribution of (2.2) along with its studentized counterpart

$$\sup_{0\le a\le 1}\sup_{h_-\le F_n(a+h)-F_n(a)\le h_+}\frac{\sqrt{n}(F_n(a+h)-F_n(a)-h)}{\sqrt{(F_n(a+h)-F_n(a))(1-F_n(a+h)+F_n(a))}}. \tag{2.3}$$

2.1.1 Related work: point processes

In one of the most celebrated results in what is now the empirical process literature, [Kol33] derived the limiting distribution of $\sqrt{n}\sup_{0\le a\le 1}(F_n(a)-a)$. This is the Kolmogorov-Smirnov statistic, and it can be seen as scanning over intervals of the form $[0,a]$, $0\le a\le 1$.

For similar reasons that motivated the introduction of the normalized scan statistic (2.2) as an improvement over the unnormalized one (2.1), [AD52] introduced and studied normalized variants of the Kolmogorov-Smirnov statistic, some of them of the form $\sqrt{n}\sup_a(F_n(a)-a)\sqrt{\psi(a)}$, where ψ is a given weight function. The choice $\psi(a)=[a(1-a)]^{-1}$ is particularly compelling, leading to the statistic

$$\sup_{0\le a\le 1}\frac{\sqrt{n}(F_n(a)-a)}{\sqrt{a(1-a)}}. \tag{2.4}$$

[Eic79] and [Jae79] obtained the limiting distributions of this statistic, its variants of the form

$$V_n = \sup_{\varepsilon_n \le a \le \delta_n} \frac{\sqrt{n}(F_n(a) - a)}{\sqrt{a(1-a)}}, \tag{2.5}$$

and its Studentized counterpart

$$\hat{V}_n = \sup_{\varepsilon_n \le a \le \delta_n} \frac{\sqrt{n}(F_n(a) - a)}{\sqrt{F_n(a)(1-F_n(a))}}, \tag{2.6}$$

for some given $0 \le \varepsilon_n \le \delta_n \le 1$. We note that these statistics can be directly expressed in terms of the order statistics, $U_{(1)} \le \cdots \le U_{(n)}$, which when $\varepsilon_n = 0$ and $\delta_n = 1$, is as follows

$$\max_{1 \le i \le n} \frac{i - nU_{(i)}}{\sqrt{nU_{(i)}(1-U_{(i)})}}, \tag{2.7}$$

and

$$\max_{1 \le i < n} \frac{i - nU_{(i)}}{\sqrt{i(1-\frac{i}{n})}}, \tag{2.8}$$

respectively.

[BJ79] proposed to directly look at each order statistic individually, combining the resulting tests using Tippett's method, leading to

$$\min_{1 \le i \le n} B(U_{(i)}; i, n-i+1), \tag{2.9}$$

with $B(\cdot; a, b)$ denoting the distribution function of the Beta(a,b) distribution. [MNS16] and [GF17] derived the asymptotic distribution of this statistic.

We note that the two-sided version of the above-mentioned tests have been considered and studied.

2.1.2 Related work: signals

Closely related to the work above is the setting where, instead of observing a point cloud, one observes a signal. The simplest situation is that of a one-dimensional signal defined on a regular lattice, that is, of the form $X_1, \dots, X_n$. The null situation is when these are iid from some underlying distribution on the real line, for example, the standard normal distribution. When the goal is to detect an interval where the observations are unusually large, and the length of the (discrete) interval is unknown, it becomes of interest to study the following scan statistic

$$Z_n = \max_{1 \le i < j \le n} \frac{S_j - S_i}{\sqrt{j-i}}, \tag{2.10}$$

where $S_k = \sum_{i=1}^k X_i$.

The study of such statistics dates back to the work of [DE56], who derived the limiting distribution of

$$\max_{1 \le j \le n} \frac{S_j}{\sqrt{j}}, \tag{2.11}$$

which can be seen as scanning intervals of the form $\{1, \dots, j\}$.

[SV95] provided the limiting distribution of the statistic (2.10) under the assumption that the X_i's are iid normal. This study was extended by [MR10] to the case where the underlying distribution is heavy-tailed, and by [KW14] when the underlying distribution has finite moment generating function in a neighborhood of the origin. [Kab11] generalized the result to the multivariate setting where the variables are indexed by a multi-dimensional lattice; see also [SAC16, KMW18]. [PWM18] studied more general scanning procedures motivated within the framework of inverse problems.

There is a parallel literature for continuous processes, where one observes instead $X_t, t \in [0,1]$ (in dimension 1). See, for example, [Ald13, QW73] and [CL06].

2.1.3 Content

The rest of the paper is organized as follows. We state our main results in Section 2.2, where we provides the asymptotic distributions of some scan statistics and their variants. The proofs are provided in Section 2.3.

2.2 Main results

Recall that $U_1, \ldots, U_n$ are iid from the uniform distribution on $[0,1]$, and that $U_{(1)} \le \cdots \le U_{(n)}$ denote the order statistics. (Whenever needed, we write $U_{(0)} \equiv 0$ and $U_{(n+1)} \equiv 1$.)

2.2.1 Studentized scan statistics

We derive the asymptotics for (2.3) before (2.2) for convenience of the proof. As we did earlier, we may rewrite (2.3) directly in terms of the order statistics, in the form of

$$M_n^+(k,l) = \max_{0 \le i < j \le n:\, k \le j-i < l} M_{i,j}, \tag{2.12}$$

where

$$M_{i,j} = \frac{j - i - n(U_{(j)} - U_{(i)})}{\sqrt{(j-i)(1 - \frac{j-i}{n})}}. \tag{2.13}$$

We will be particularly interested in the following special case

$$M_n^+ := M_n^+(1,n), \tag{2.14}$$

which is the analog of (2.8). Not surprisingly, the limiting distribution is an extreme value distribution, specifically, a Gumbel distribution. Indeed, we have the following.

Theorem 2.2.1.1. *For any* $\tau \in \mathbb{R}$,

$$\lim_{n\to\infty} \mathbb{P}\left\{ M_n^+ \leq \sqrt{2\log n} - \frac{3\log\log n}{2\sqrt{2\log n}} + \frac{\tau}{\sqrt{2\log n}} \right\} = \exp\big(-c\exp(-\tau)\big), \tag{2.15}$$

where $c = \frac{8}{9\sqrt{\pi}}$.

Similarly, define the opposite one-sided statistics

$$M_n^-(k,l) = -\min_{0\leq i<j\leq n:\, k\leq j-i\leq l} M_{i,j}, \tag{2.16}$$

and

$$M_n^- := M_n^-(1,n). \tag{2.17}$$

Finally, define the two-sided statistics

$$M_n(k,l) = \max\{M_n^+(k,l), M_n^-(k,l)\} = \max_{0\leq i<j\leq n:\, k\leq j-i<l} |M_{i,j}|, \tag{2.18}$$

and

$$M_n := M_n(1,n) = \max\{M_n^+, M_n^-\}. \tag{2.19}$$

For these statistics too, the limiting distribution is a Gumbel distribution, but what is surprising here is that these statistics do not behave the same way as M_n^+. In particular, $M_n^- = (1+o_P(1))\log n$, and therefore dominates M_n^+ in the large-sample limit, implying that $M_n = M_n^-$ with probability tending to 1. Indeed, we have the following.

Theorem 2.2.1.2. *For any* $\tau \in \mathbb{R}$,

$$\lim_{n\to\infty} \mathbb{P}\left\{ M_n^- \leq \log n + \tau \right\} = \exp(-\exp(1-\tau)). \tag{2.20}$$

Moreover,

$$\lim_{n\to\infty} \mathbb{P}\left\{M_n = M_n^-\right\} = 1. \tag{2.21}$$

2.2.2 Standardized scan statistics

We also examine the large-sample behavior of standardized scan statistics (2.2). Following the same way as rewriting (2.3) before. Define

$$\tilde{M}_n^+(k,l) := \max_{0\le i<j\le n: k\le j-i\le l} \tilde{M}_{i,j}, \tag{2.22}$$

where

$$\tilde{M}_{i,j} := \frac{j-i-n(U_{(j)}-U_{(i)})}{\sqrt{n(U_{(j)}-U_{(i)})(1-U_{(j)}+U_{(i)})}}. \tag{2.23}$$

Note that

$$\tilde{M}_n^+ := \tilde{M}_n^+(1,n), \tag{2.24}$$

is the analog of (2.7).

The behavior of $\tilde{M}_n^+$ turns out to be very different from that of its studentized analog M_n^+. However, we recover a similar behavior if we appropriately bound the length of the scanning interval from below.

Theorem 2.2.2.1. *For any* $\tau \in \mathbb{R}$*,*

$$\lim_{n\to\infty} \mathbb{P}\left\{\tilde{M}_n^+ \le \sqrt{\frac{n}{\tau}}\right\} = \exp(-\tau). \tag{2.25}$$

Moreover, for any $A > 0$*, defining* $k_n = \lceil A(\log n)^3 \rceil$*,*

$$\lim_{n\to\infty} \mathbb{P}\left\{\tilde{M}_n^+(k_n,n) \le \sqrt{2\log n} - \frac{3\log\log n}{2\sqrt{2\log n}} + \frac{\tau}{\sqrt{2\log n}}\right\} = \exp(-c_A \exp(-\tau)), \tag{2.26}$$

where $c_A = \int_A^\infty \Lambda_1(a)da$ *with* $\Lambda_1(a) = \frac{1}{2\sqrt{\pi}a^2}\exp\left(\frac{\sqrt{2}}{3\sqrt{a}}\right)$.

Remark 2.2.2.1. Here we choose $k_n \propto (\log n)^3$ because we want to examine the behavior of $\tilde{M}^+(K,L)$, compared to its counterpart $M^+(K,L)$ at the most contributed part, which is reflected in the proof of Theorem 2.2.1.1. For readers who are curious about other choices of k_n, we note that $\tilde{M}_{i,j}$ behaves like subgaussian, or named as "sublogarithmic" in [KW14]. Roughly speaking,$\tilde{M}_n^+(k_n,n)$ will likely to take its maximum around the indices i, j with small length, that is, when $j-i$ is close to k_n.

Define the standardized analog of (2.17)

$$\tilde{M}_n^-(k,l) = -\min_{0\le i<j\le n:\, k\le j-i\le l} \tilde{M}_{i,j}, \tag{2.27}$$

with

$$\tilde{M}_n^- := \tilde{M}_n^-(1,n), \tag{2.28}$$

as well as the analog of (2.19)

$$\tilde{M}_n(k,l) = \max\{\tilde{M}_n^+(k,l), \tilde{M}_n^-(k,l)\}, \tag{2.29}$$

with

$$\tilde{M}_n := \tilde{M}_n(1,n) = \max\{\tilde{M}_n^+, \tilde{M}_n^-\}. \tag{2.30}$$

Theorem 2.2.2.2. *We have*

$$\lim_{n\to\infty} \mathbb{P}\left(\tilde{M}_n = \tilde{M}_n^+\right) = 1. \tag{2.31}$$

Thus for any $\tau \in \mathbb{R}$,

$$\lim_{n\to\infty} \mathbb{P}\left(\tilde{M}_n \le \sqrt{\frac{n}{\tau}}\right) = \exp(-\tau). \tag{2.32}$$

Remark 2.2.2.2. While the behavior of the Studentized statistic M_n^+ is driven by the smallest intervals, this is not as much the case for the standardized statistic $\tilde{M}_n^+$. Indeed, a large value of

M_n^+ comes from some $n(U_{(j)} - U_{(i)})$ being large compared to $j-i$, however, $n(U_{(j)} - U_{(i)})$ being in the denominator defining $\tilde{M}_n^+$, its impact is lessened.

2.3 Proofs of Main Results

Our proof arguments are based on standard moderate and large deviation results, Kolmogorov's theorem, a Poisson approximation [AGG89], as well as some technical results developed by [KW14] in their study of the limiting distribution of the scan statistic in the form of (2.10).

2.3.1 Preliminaries

Throughout the paper, we assume that $\{X_k, k \in \mathbb{Z}\}$ are iid distributed with the density,

$$f(x) = \mathbb{1}(x \leq 1) \exp(x-1), \tag{2.33}$$

noting that $-X_1 + 1$ follows standard exponential distribution. This distribution has zero mean and unit variance. Define the two-sided partial sums,

$$S_k^+ = \sum_{i=1}^{k} X_i, \quad S_0^+ = 0, \quad S_{-k}^+ = -\sum_{i=1}^{k} X_{-i}, \ \ k \in \mathbb{N} \tag{2.34}$$

and

$$S_k^- := -S_k^+. \tag{2.35}$$

They will play a central role in what follows. Define the normalized increments

$$Z_{i,j}^{\pm} = \frac{S_j^{\pm} - S_i^{\pm}}{\sqrt{j-i}}, \tag{2.36}$$

$$Z_n^{\pm}(k,l) := \max_{1\le i<j\le n: k\le j-i\le l} Z_{i,j}^{\pm}, \qquad Z_n^{\pm} := Z_n^{\pm}(1,n). \tag{2.37}$$

Let $\varphi^{\pm}(t)$ be the cumulant generating functions of $\pm X_1$ respectively. We have

$$\varphi^{+}(t) = t - \log(1+t), \quad \text{if } t \ge 0. \tag{2.38}$$

$$\varphi^{-}(t) = \begin{cases} -t - \log(1-t), & \text{if } 0 \le t \le 1, \\ \infty, & \text{if } t \ge 1, \end{cases} \tag{2.39}$$

Also, define $I^{+}(s)$ and $I^{-}(s)$ as the respective Legendre-Fenchel transforms (a.k.a., rate functions). We have

$$I^{+}(s) = \begin{cases} -s - \log(1-s), & \text{if } 0 \le s \le 1, \\ \infty, & \text{if } s \ge 1, \end{cases} \tag{2.40}$$

and

$$I^{-}(s) = s - \log(1+s), \tag{2.41}$$

with respective Taylor expansions at 0 (as $s \to 0$)

$$I^{+}(s) = s^2/2 + s^3/3 + o(s^3),$$
$$I^{-}(s) = s^2/2 - s^3/3 + o(s^3).$$

We also prepare several usefull lemmas. The first two lemmas are well-known moderate and large deviations results [Cra38, BR60].

Lemma 2.3.1.1. *Let (x_k) be a sequence satisfying $x_k \to \infty$ and $x_k = o(\sqrt{k})$ as $k \to \infty$. Then, as $k \to \infty$,*

$$\mathbb{P}\left(\frac{S_k^{\pm}}{\sqrt{k}} \ge x_k\right) \sim \frac{1}{\sqrt{2\pi} x_k} \exp\left\{ -kI^{\pm}\left(\frac{x_k}{\sqrt{k}}\right)\right\}. \tag{2.42}$$

Lemma 2.3.1.2. *For every $k \in \mathbb{N}$ and $x > 0$, we have*

$$\mathbb{P}\left(\frac{S_k^{\pm}}{\sqrt{k}} \geq x\right) \leq \exp\left\{-kI^{\pm}\left(\frac{x}{\sqrt{k}}\right)\right\}. \tag{2.43}$$

Moreover, for every $A \leq s_\infty$, where $s_\infty = \sup\{s \in \mathbb{R} : \mathbb{P}(X_1 \leq s) \leq 1\}$, there is $C_A > 0$ such that, for all $k \in \mathbb{N}$ and $x \in (0, A\sqrt{k})$,

$$\mathbb{P}\left(\frac{S_k^{\pm}}{\sqrt{k}} \geq x\right) \leq \frac{C_A}{x}\exp\left\{-kI^{\pm}\left(\frac{x}{\sqrt{k}}\right)\right\}, \tag{2.44}$$

The following result is obtained from a simple application of Theorem 2.4 in [Pet95], which provides an upper bound of the tail distribution of $\max_{1\leq k\leq n} S_k^{\pm}$ by that of $S_n^{\pm}$.

Lemma 2.3.1.3. *We have*

$$\mathbb{P}\left\{\max_{1\leq k\leq n} S_k^{\pm} \geq x\right\} \leq 2\,\mathbb{P}\left\{S_n^{\pm} \geq x - \sqrt{2(n-1)}\right\}. \tag{2.45}$$

For completeness, we include Lemma 4.4 and 4.5 from [KW14] below. For integers $r > 0$ and $x < y$, define

$$\mathbb{T}_r(x,y) := \big\{(i,j) \in \mathbb{I} : x - r \leq i \leq x \text{ and } y \leq j \leq y + r\big\}. \tag{2.46}$$

Lemma 2.3.1.4. *Fix constants $B_1, B_2 > 0$. Then for all $x \in \mathbb{Z}$, $l, r \in \mathbb{N}$ and all $u > 0$ such that $B_1 l > u^2$ and $r \leq B_2 l u^{-2}$, we have*

$$Q(l,r,u) := \mathbb{P}\left(\max_{i,j\in\mathbb{T}_r(x,x+l)} \frac{S_j^+ - S_i^+}{\sqrt{l}} \geq u\right) \leq \frac{C}{u}\exp\left(-\frac{u^2}{2} - \frac{cu^3}{\sqrt{l}}\right), \tag{2.47}$$

where the constants c and C depend on B_1 and B_2 but do not depend on x, l, r, u.

Lemma 2.3.1.5. *Let ν, ν_n, $n \in \mathbb{N}$, be measures on $[0, \infty)$ which are finite on compact intervals.*

Let G, G_n, $n \in \mathbb{N}$, be measurable functions on $[0,\infty)$ which are uniformly bounded on compact intervals. Assume that

1. *ν_n converges to ν weakly on every interval $[0,t]$, $t \geq 0$;*
2. *for ν-a.e. $s \geq 0$, we have $\lim_{n\to\infty} G_n(s_n) = G(s)$, for every sequence $s_n \to s$;*
3. *$\lim_{T\to\infty} \int_T^\infty |G_n| d\nu_n = 0$ uniformly when $n \geq N$ for some $N \in \mathbb{N}$.*

Then, $\lim_{n\to\infty} \int_0^T G_n d\nu_n = \int_0^T G d\nu$.

We also provide an upper bound of the tail distribution $\max_{i,j\in\mathbb{T}_r(x,x+l)}(S_j^- - S_i^-)/\sqrt{l}$ also, which is cruder than its counterpart for S_k^+ in Lemma 2.3.1.4 but shall suffice for our purposes.

Lemma 2.3.1.6. *For all $x \in \mathbb{Z}$, $l, r \in \mathbb{N}^+$ and all $u > 40$ such that $l > u^2 r$ and $r > 10u^2$, we have*

$$Q(l,r,u) := \mathbb{P}\left(\max_{i,j\in\mathbb{T}_r(x,x+l)} \frac{S_j^- - S_i^-}{\sqrt{l}} \geq u\right) \leq C\exp\left(-\frac{u^2}{3}\right), \tag{2.48}$$

where the constant C does not depend on x, l, r, u.

Proof. Before we proceed into the proof, one fact about $I^-(s)$ is

$$I^-(s) \geq \frac{1.01s^2}{3}, \quad 0 \leq s \leq 0.5, \tag{2.49}$$

which can be easily checked. Define $V_{l,u} := u^2 - uS_l^-/\sqrt{l}$, $S_{k_1}^{(1)-}$ and $S_{k_2}^{(2)-}$ to be two partial sums of $-X_i$ independent of each other and S_l^-. With translation invariance, we bound $Q(l,r,u)$ as follows,

$$\begin{aligned}
Q(l,r,u) &= \mathbb{P}\left(\max_{i,j\in\mathbb{T}_r(0,0+l)} \frac{S_j^- - S_i^-}{\sqrt{l}} \geq u\right) \\
&= \mathbb{P}\left(\max_{0\leq k_1,k_2\leq r} \frac{S_{k_1}^{(1)-} + S_{k_2}^{(2)-}}{\sqrt{l}} + \frac{S_l^-}{\sqrt{l}} \geq u\right) \\
&= \mathbb{P}\left(\max_{0\leq k_1,k_2\leq r} \frac{S_{k_1}^{(1)-} + S_{k_2}^{(2)-}}{\sqrt{l}} \geq \frac{V_{l,u}}{u}\right) \\
&\leq \mathbb{P}\left(\max_{0\leq k_1,k_2\leq r} \frac{S_{k_1}^{(1)-} + S_{k_2}^{(2)-}}{\sqrt{l}} \geq \frac{V_{l,u}}{u}, V_{l,u} \leq u^2\sqrt{\frac{r}{l}}\right) \\
&+ \mathbb{P}\left(\max_{0\leq k_1,k_2\leq r} \frac{S_{k_1}^{(1)-} + S_{k_2}^{(2)-}}{\sqrt{l}} \geq \frac{V_{l,u}}{u}, V_{l,u} > u^2\sqrt{\frac{r}{l}}\right) \\
&\leq \mathbb{P}(V_{l,u} \leq u^2\sqrt{r/l}) + \mathbb{P}\left(\max_{0\leq k_1,k_2\leq r} \frac{S_{k_1}^{(1)-} + S_{k_2}^{(2)-}}{\sqrt{l}} > u\sqrt{\frac{r}{l}}\right),
\end{aligned}$$

where we bound these two terms individually. By the assumptions on u,l,r, we have $u(1-\sqrt{r/l})/\sqrt{l} \leq 0.5$. Thus with (2.43) and (2.49), we have

$$\mathbb{P}\left(V_{l,u} \leq u^2\sqrt{\frac{r}{l}}\right) = \mathbb{P}\left(\frac{S_l^-}{\sqrt{l}} \geq u - u\sqrt{\frac{r}{l}}\right) \leq \exp\left[-lI^+\left\{\frac{u(1-\sqrt{r/l})}{\sqrt{l}}\right\}\right] \leq \exp\left(-\frac{u^2}{3}\right). \tag{2.50}$$

Now we switch to the second item, with Lemma 2.3.1.3, (2.43) and assumption that $r > 10u^2$, $u > 40$,

$$\begin{aligned}
\mathbb{P}\left(\max_{0\leq k_1,k_2\leq r} \frac{S_{k_1}^{(1)-} + S_{k_2}^{(2)-}}{\sqrt{l}} \geq u\sqrt{\frac{r}{l}}\right) &\leq 2\,\mathbb{P}\left(\max_{0\leq k\leq r} \frac{S_k^-}{\sqrt{r}} \geq \frac{u}{2}\right) \\
&\leq 4\,\mathbb{P}\left(\frac{S_r^-}{\sqrt{r}} \geq \frac{u}{2} - \sqrt{2}\right) \\
&\leq C\exp\left\{-rI^-\left(\frac{u-2\sqrt{2}}{2\sqrt{r}}\right)\right\} \\
&\leq C\exp\left(-\frac{u^2}{3}\right).
\end{aligned}$$

Putting the two terms together, we get the stated bound. □

We now adjust the Lemma 2.3.1.4 to suit for proving Theorem 2.2.2.1, in which we need to deal with

$$\tilde{Z}^+_{i,j} := \frac{S^+_j - S^+_i}{\sqrt{j-i-(S^+_j - S^+_i)}}. \tag{2.51}$$

Define a function

$$\phi(x) = \frac{x}{\sqrt{1-x}}, \;\; x < 1, \tag{2.52}$$

and thus we have

$$\frac{\tilde{Z}^+_{i,j}}{\sqrt{j-i}} = \phi\left(\frac{Z^+_{i,j}}{\sqrt{j-i}}\right). \tag{2.53}$$

Since $\phi(x)$ is strictly increasing on $(-\infty, 1)$ with range $\mathbb{R}$, we write its inverse function as

$$g^+(x) := \frac{1}{2}(x\sqrt{x^2+4} - x^2), \;\; x \in \mathbb{R}, \tag{2.54}$$

which is also strictly increasing. Therefore, $\tilde{Z}^+_{i,j} \geq u$ if and only if

$$Z^+_{i,j} \geq \sqrt{j-i} \cdot g^+\left(\frac{a}{\sqrt{j-i}}\right). \tag{2.55}$$

This is an important transformation which enables us to deal with $Z^+_{i,j}$ instead. We compute the Taylor expansion of $I^+(g^+(s))$ at $s = 0$,

$$I^+(g^+(s)) = \frac{s^2}{2} - \frac{s^3}{6} + O(s^4). \tag{2.56}$$

We have

Lemma 2.3.1.7. *Fix constants B_1, $B_2 > 0$. Then for all $x \in \mathbb{Z}$, $l, r \in \mathbb{N}$ and all $u > 0$ such that*

$B_1 l > u^2$ and $r < B_2 l u^{-2}$, we have

$$Q(l,r,u) := \mathbb{P}\left(\max_{(i,j)\in\mathbb{T}_r(x,x+l)} \tilde{Z}^+_{i,j} \geq u\right) \leq \frac{C}{u}\exp\left(-\frac{u^2}{2}+\frac{cu^3}{\sqrt{l}}\right), \tag{2.57}$$

where the constants $c, C > 0$ depend on B_1 and B_2 but do not depend on x, l, r, u.

Proof. By the transformation (2.55), translation invariance and the fact that $g^+(x)/x^2$ is strictly decreasing,

$$\begin{aligned}
Q(l,r,u) &= \mathbb{P}\left(\max_{(i,j)\in\mathbb{T}_r(0,l)} \tilde{Z}^+_{i,j} \geq u\right) && (2.58)\\
&= \mathbb{P}\left[\max_{0\leq k_1,k_2\leq r}\left\{S^{(1)+}_{k_1}+S^{(2)+}_{k_2}-(l+k_1+k_2)\cdot g^+\left(\frac{u}{\sqrt{l+k_1+k_2}}\right)\right\}+S^+_l \geq 0\right] && (2.59)\\
&\leq \mathbb{P}\left[\max_{0\leq k_1,k_2\leq r}\left\{S^{(1)+}_{k_1}+S^{(2)+}_{k_2}\right\}-l\cdot g^+\left(\frac{u}{\sqrt{l}}\right)+S^+_l \geq 0\right], && (2.60)
\end{aligned}$$

where $S^{(1)+}_{k_1}$, $S^{(2)+}_{k_2}$ are two partial sums of X_i independent of each other and S^+_l. Define

$$V_{l,u} = u\left(u - \frac{S^+_l}{\sqrt{l-S^+_l}}\right). \tag{2.61}$$

Thus

$$\frac{S^+_l}{\sqrt{l-S^+_l}} = \frac{l\cdot S^+_l/l}{\sqrt{l}\sqrt{1-S^+_l/l}} = \sqrt{l}\cdot\phi\left(\frac{S^+_l}{l}\right) = u - \frac{V_{l,u}}{u}, \tag{2.62}$$

which gives

$$S^+_l = l\cdot g^+\left(\frac{u-V_{l,u}/u}{\sqrt{l}}\right). \tag{2.63}$$

Therefore,

$$
\begin{aligned}
&Q(l,r,u) && (2.64)\\
&\leq \mathbb{P}\left[\max_{0\leq k_1,k_2\leq r}\left\{S_{k_1}^{(1)+}+S_{k_2}^{(2)+}\right\}-l\cdot g^{+}\left(\frac{u}{\sqrt{l}}\right)+l\cdot g^{+}\left(\frac{u-V_{l,u}/u}{\sqrt{l}}\right)\geq 0, V_{l,u}\leq 0\right] && (2.65)\\
&+\mathbb{P}\left[\max_{0\leq k_1,k_2\leq r}\left\{S_{k_1}^{(1)+}+S_{k_2}^{(2)+}\right\}-l\cdot g^{+}\left(\frac{u}{\sqrt{l}}\right)+l\cdot g^{+}\left(\frac{u-V_{l,u}/u}{\sqrt{l}}\right)\geq 0, V_{l,u}> 0\right] && (2.66)\\
&=\mathbb{P}(V_{l,u}\leq 0) && (2.67)\\
&+\mathbb{P}\left[\max_{0\leq k_1,k_2\leq r}\left\{S_{k_1}^{(1)+}+S_{k_2}^{(2)+}\right\}-l\cdot g^{+}\left(\frac{u}{\sqrt{l}}\right)+l\cdot g^{+}\left(\frac{u-V_{l,u}/u}{\sqrt{l}}\right)\geq 0, V_{l,u}> 0\right] && (2.68)\\
&=F_{l,u}(0)+\int_0^\infty G_{l,r,u}(s)dF_{l,u}(s), && (2.69)
\end{aligned}
$$

where the last equality is obtained by conditioning on $V_{l,u}=s$, which is independent of $S_{k_1}^{(1)+}$, $S_{k_2}^{(2)+}$. $F_{l,u}$ therein is the probability distribution of $V_{l,u}$ and

$$
G_{l,r,u}(s):=\mathbb{P}\left[\max_{0\leq k_1,k_2\leq r}\left\{S_{k_1}^{(1)+}+S_{k_2}^{(2)+}\right\}-l\cdot g^{+}\left(\frac{u}{\sqrt{l}}\right)+l\cdot g^{+}\left(\frac{u-s/u}{\sqrt{l}}\right)\geq 0\right],
$$

which is decreasing. To obtain an upper bound for $Q(l,r,u)$, first we bound $F_{l,u}(s)$ for $s\in[0,\frac{3}{4}u^2]$ so that $u-s/u\in[u/4,u]$. Applying (2.44),

$$
\begin{aligned}
F_{l,u}(s)&=\mathbb{P}\left(\frac{S_l^{+}}{\sqrt{l-S_l^{+}}}\geq u-\frac{s}{u}\right)\\
&=\mathbb{P}\left\{\frac{S_l^{+}}{\sqrt{l}}\geq\sqrt{l}\cdot g^{+}\left(\frac{u-s/u}{\sqrt{l}}\right)\right\}\\
&\leq C\left\{\sqrt{l}\cdot g^{+}\left(\frac{u-s/u}{\sqrt{l}}\right)\right\}^{-1}\exp\left[-l\cdot I^{+}\left\{g^{+}\left(\frac{u-s/u}{\sqrt{l}}\right)\right\}\right]\\
&\leq\frac{C}{u}\exp\left[-l\cdot I^{+}\left\{g^{+}\left(\frac{u-s/u}{\sqrt{l}}\right)\right\}\right],
\end{aligned}
$$

where the last inequality follows from the fact that when $0 < x < 1$,

$$xg^+\left(\frac{1}{x}\right) > \frac{1}{2}. \tag{2.70}$$

By Taylor expansion of $I^+(g^+(s))$, we have

$$\begin{aligned} F_{l,u}(s) &\le \frac{C}{u}\exp\left\{-\frac{1}{2}\left(u-\frac{s}{u}\right)^2 + \frac{c}{2\sqrt{l}}\left(u-\frac{s}{u}\right)^3\right\} \\ &\le \frac{Ce^s}{u}\exp\left(-\frac{u^2}{2}+\frac{cu^3}{\sqrt{l}}\right). \end{aligned} \tag{2.71}$$

It is however easy to see that this inequality continues to hold for $s \ge \frac{3}{4}u^2$. Indeed, if c is sufficiently small, then the assumption $B_1 l > u^2$ implies that $cu^3/\sqrt{l} \le u^2/8$. Hence, when $s \ge \frac{3}{4}u^2$, the above inequality becomes

$$F_{l,u}(s) \le \frac{C}{u}\exp\left(\frac{3u^2}{8}\right). \tag{2.72}$$

If C is sufficiently large, the right-hand side of previous inequality is greater than 1 and hence the inequality trivially holds. We bound $G_{l,r,u}(s)$ for $s \ge 0$,

$$\begin{aligned} G_{l,r,u}(s) &\le \mathbb{P}\left\{\max_{0\le k_1,k_2<r} S_{k_1}^{(1)+} + S_{k_2}^{(2)+} > \frac{s}{2u}\sqrt{\left(u-\frac{s}{u}\right)^2+4l} + \frac{s^2}{2u^2} - s\right\} \\ &\le 2\,\mathbb{P}\left\{\max_{0\le k<r} S_k^+ > \frac{s}{4u}\sqrt{\left(u-\frac{s}{u}\right)^2+4l} - \frac{s}{2}\right\} \\ &\le 2\,\mathbb{P}\left\{\max_{0\le k<r} S_k^+ > \frac{s}{2u}\sqrt{l} - \frac{s}{2}\right\}. \end{aligned}$$

Applying the Lemma 2.3.1.3 to the above equation we obtain

$$\begin{aligned} G_{l,r,u}(s) &\leq 4\,\mathbb{P}\left(S_r^+ > \frac{s}{2u}\sqrt{l} - \frac{s}{2} - \sqrt{2r}\right) \\ &\leq 4\,\mathbb{P}\left(\frac{S_r^+}{\sqrt{r}} > \frac{s}{2u\sqrt{r}}\sqrt{l} - \frac{s}{2\sqrt{r}} - \sqrt{2}\right) \\ &\leq 4\exp\left\{-rI^+\left(\frac{cs-\sqrt{2}}{\sqrt{r}}\right)\right\}. \end{aligned}$$

In the second inequality, we used the assumption $r < B_2 l u^{-2}$. By noticing the fact that $I^+(s) \geq s^2/2$, we have

$$G_{l,r,u}(s) \leq Ce^{-cs^2}. \tag{2.73}$$

Strictly speaking, this is valid only as long as $cs \geq \sqrt{2}$, however, we can choose the constant C so large that (2.73) continues to hold in the case $cs < \sqrt{2}$. To obtain (2.57), by (2.69), (2.71), (2.73), it is clear that

$$\begin{aligned} Q(l,r,u) &\leq F_{l,u}(0) + \sum_{k=0}^{\infty} G_{l,r,u}(k) F_{l,u}(k+1) \\ &\leq \frac{C}{u}\left(1 + \sum_{k=0}^{\infty} e^{-ck^2} e^k\right) \exp\left(-\frac{u^2}{2} + \frac{cu^3}{\sqrt{l}}\right) \\ &\leq \frac{C}{u} \exp\left(-\frac{u^2}{2} + \frac{cu^3}{\sqrt{l}}\right). \end{aligned}$$

□

2.3.2 Proof of Theorem 2.2.1.1 and Theorem 2.2.1.2

The roadmap of our proof. We know that $(U_{(1)}, U_{(2)}, \ldots, U_{(n)})$ has the same distribution as

$$\left(\frac{Y_1}{\sum_{i=1}^{n+1} Y_i}, \frac{Y_1+Y_2}{\sum_{i=1}^{n+1} Y_i}, \ldots, \frac{\sum_{i=1}^{n} Y_i}{\sum_{i=1}^{n+1} Y_i}\right), \quad \text{where } Y_1, \ldots, Y_{n+1} \text{ are iid exponential.} \tag{2.74}$$

In particular, Y_i can be set as $1-X_i$. We use this fact, together with a comparison of $\sum_{i=1}^{n+1} Y_i$ with its mean using a central limit theorem, to deal with the dependency among order statistics above, effectively reducing the problem to partial sums of iid random variables. We then divide the intervals into smaller intervals, which end up contributing the most to the maximum, and larger intervals, whose contribution we show to be negligible. Although $U_{(i)}$ and Y_i may be defined on different probability spaces with different probability measure, we may switch between them when there is no confusion. Because we only prove convergence in distribution, from now on, we put $U_{(j)} = \sum_{i=1}^{j} Y_i / \sum_{i=1}^{n+1} Y_i$ throughout the proof.

Proof of (2.15)

We study the asymptotic behavior of the statistic based on different regions of $j-i$. For $b > 0$, define the event

$$\begin{aligned} A_{i,j}^{n+}(b) &= \left\{ \frac{j-i-n(U_{(j)}-U_{(i)})}{\sqrt{(j-i)(1-\frac{j-i}{n})}} \leq b \right\} \\ &= \left\{ U_{(j)} - U_{(i)} \geq \frac{j-i}{n} - \frac{b}{\sqrt{n}} w_{i,j}^{n} \right\}, \end{aligned}$$

where

$$w_{i,j}^{n} := \sqrt{\frac{j-i}{n}\left(1-\frac{j-i}{n}\right)}. \tag{2.75}$$

Under this notation, we have

$$\left\{ M_n^+ \leq b \right\} = \bigcap_{0 \leq i < j \leq n} A_{i,j}^{n+}(b). \tag{2.76}$$

Define

$$u_n(\tau) = \left(1 + \frac{-3\log\log n + 2\tau}{4\log n}\right)\sqrt{2\log n}. \tag{2.77}$$

Throughout the proof, we abbreviate $u_n(\tau)$ as u_n with τ fixed. With this choice, we have

$u_n \sim \sqrt{2\log n}$.

Step 1: Upper bound

For the upper bound, it suffices to focus on the optimal range so that the maximum is achieved. This turns out to be at $j-i \propto (\log n)^3$, as discussed below.

Define the events

$$\Omega_n = \big\{|S_{n+1}^{+}| \le (\log\log n)\sqrt{n}\big\}. \tag{2.78}$$

By the central limit theorem,

$$\mathbb{P}(\Omega_n) \to 1 \;\text{ as }\; n \to \infty. \tag{2.79}$$

When $j-i \le \frac{n}{\log n \log\log n}$,

$$\begin{aligned}
&A_{i,j}^{n+}(u_n)\\
&\subseteq \Omega_n^{\mathsf{c}} \bigcup \{\Omega_n \bigcap A_{i,j}^{n+}(u_n)\}\\
&= \Omega_n^{\mathsf{c}} \bigcup \left(\Omega_n \bigcap \left\{\frac{j-i-S_j^{+}+S_i^{+}}{n+1-S_{n+1}^{+}} \ge \frac{j-i}{n} - \frac{u_n}{\sqrt{n}} w_{i,j}^{n}\right\}\right)\\
&\subseteq \Omega_n^{\mathsf{c}} \bigcup \left(\Omega_n \bigcap \left\{S_j^{+} - S_i^{+} \le (j-i)\frac{-1+S_{n+1}^{+}}{n} + \frac{u_n}{\sqrt{n}}(n+1-S_{n+1}^{+}) w_{i,j}^{n}\right\}\right)\\
&\subseteq \Omega_n^{\mathsf{c}} \bigcup \left\{S_j^{+} - S_i^{+} \le (j-i)\frac{\log\log n}{\sqrt{n}} + \frac{u_n}{\sqrt{n}}(n+1+(\log\log n)\sqrt{n}) w_{i,j}^{n}\right\}\\
&= \Omega_n^{\mathsf{c}} \bigcup \left\{Z_{i,j}^{+} \le \frac{(\log\log n)}{\sqrt{n}}\sqrt{j-i} + u_n \cdot \left(1 + \frac{(\log\log n)\sqrt{n}+1}{n}\right)\sqrt{1-\frac{j-i}{n}}\right\}\\
&\subseteq \Omega_n^{\mathsf{c}} \bigcup \left\{Z_{i,j}^{+} \le \sqrt{\frac{\log\log n}{\log n}} + u_n \cdot \left(1 + \frac{(\log\log n)\sqrt{n}+1}{n}\right)\right\}\\
&\subseteq \Omega_n^{\mathsf{c}} \bigcup \{Z_{i,j}^{+} \le u_n(\tau+\varepsilon)\},
\end{aligned}$$

for any fixed $\varepsilon > 0$ provided that n is large enough. To deal with the standardized sums $Z_{i,j}^{+}$, we need Theorem 1.1 and Theorem 1.2 in [KW14]. Because $X_1 \le 1$, it belongs to the superlogarithm family defined in [KW14]. Applying Theorem 1.1 and Theorem 1.2 in [KW14], we obtain

$$\lim_{n\to\infty} \mathbb{P}\{Z_n^+ \leq u_n\} = \exp\left\{-\frac{8}{9\sqrt{\pi}}e^{-\tau}\right\}, \tag{2.80}$$

and

$$\lim_{A\to\infty} \liminf_{n\to\infty} \mathbb{P}\{Z_n^+ = Z_n^+(A^{-1}(\log n)^3, A(\log n)^3)\} = 1. \tag{2.81}$$

By (2.79), (2.80) and the fact that $(\log n)^3 \ll \frac{n}{\log n(\log\log n)}$,

$$\begin{aligned}
&\limsup_{n\to\infty} \mathbb{P}(M_n^+ \leq u_n)\\
&= \limsup_{n\to\infty} \mathbb{P}\left\{\bigcap_{0\leq i<j\leq n} A_{i,j}^{n+}(u_n)\right\}\\
&\leq \limsup_{n\to\infty} \mathbb{P}\left\{\bigcap_{0\leq i<j\leq n: j-i\leq \frac{n}{\log n\log\log n}} A_{i,j}^{n+}(u_n(\tau+\varepsilon))\right\} + \limsup_{n\to\infty} \mathbb{P}(\Omega_n^{\mathrm{c}})\\
&\leq \limsup_{n\to\infty} \mathbb{P}\left\{Z_n^+\left(1, \frac{n}{\log n\log\log n}\right) \leq u_n(\tau+\varepsilon)\right\} + \limsup_{n\to\infty} \mathbb{P}(\Omega_n^{\mathrm{c}})\\
&= \exp\left\{-\frac{8}{9\sqrt{\pi}}e^{-\tau-\varepsilon}\right\}.
\end{aligned}$$

As $\varepsilon > 0$ is arbitrary we get

$$\limsup_{n\to\infty} \mathbb{P}(M_n^+ \leq u_n) \leq \lim_{\varepsilon\to 0} \exp\left\{-\frac{8}{9\sqrt{\pi}}e^{-\tau-\varepsilon}\right\} = \exp\left\{-\frac{8}{9\sqrt{\pi}}e^{-\tau}\right\}. \tag{2.82}$$

Step 2: Lower bound

Define

$$k_n = \frac{n}{\log n(\log\log n)}, \quad K_n = \frac{n\log\log n}{\log n}. \tag{2.83}$$

We establish the lower bound by dividing the range of $j-i$ into five regions:

$$R_1 = [1, u_n^2), \qquad R_2 = [u_n^2, k_n),$$
$$R_3 = [k_n, K_n), \qquad R_4 = [K_n, n-K_n),$$
$$R_5 = [n-K_n, n).$$

• For R_1, note that

$$\frac{j-i}{n} - \frac{u_n}{\sqrt{n}} w_{i,j}^n \leq 0, \tag{2.84}$$

is equivalent to

$$j-i \leq \frac{u_n^2}{1+u_n^2/n}. \tag{2.85}$$

Since $u_n^4 \ll n$, i, j only take value in integers, it is further equivalent to $j-i \leq u_n^2$ when n is large enough, which is exactly R_1. Therefore, when n is large enough,

$$A_{i,j}^{n+}(u_n) = \Omega, \tag{2.86}$$

for any (i,j) satisfying $j-i \in R_1$ so that

$$\bigcap_{0\leq i<j\leq n: j-i \in R_1} A_{i,j}^{n+}(u_n) = \Omega. \tag{2.87}$$

• For R_2, following the same argument that was used to prove the upper bound, it can be shown that

$$\liminf_{n\to\infty} \mathbb{P}\left\{ \bigcap_{0\leq i<j\leq n: j-i\in R_2} A_{i,j}^{n+}(u_n) \right\} \geq \exp\left\{ -\frac{8}{9\sqrt{\pi}} e^{-\tau} \right\}. \tag{2.88}$$

• Turning to R_3, we shall show that

$$\mathbb{P}\left(\max_{0\leq i\leq n-k_n} \frac{S_{i+k_n}^+ - S_i^+}{\sqrt{k_n}} \leq \log\log n \right) \to 1, \tag{2.89}$$

and then use this fact to prove that the maximum of $M^+_{i,j}$ over R_3 is ignorable. First we bound $\max_{0\le i\le n-k_n}(S^+_{i+k_n}-S^+_i)$. Define

$$q_n = \frac{k_n}{(\log\log n)^2} \ll k_n, \tag{2.90}$$

and introduce a positive sequence ε_n such that $q_n \ll \varepsilon_n \ll k_n$. Consider the following two-dimensional grid with mesh size q_n:

$$\mathcal{J}_n = \{(x,y) \in q_n\mathbb{Z}^2 : x \in [-\varepsilon_n, n+\varepsilon_n], y-x \in [0.9k_n-\varepsilon_n, 1.1k_n+\varepsilon_n]\}. \tag{2.91}$$

By the union bound,

$$\mathbb{P}\left\{Z^+_n(0.9k_n, 1.1k_n) > \log\log n\right\} \le \sum_{(x,y)\in\mathcal{J}_n} \mathbb{P}\left\{\max_{(i,j)\in\mathbb{T}_{q_n}(x,y)} Z^+_{i,j} \ge \log\log n\right\}. \tag{2.92}$$

Note that the cardinality of $\mathcal{J}_n$ satisfies

$$|\mathcal{J}_n| \sim \frac{(1.1-0.9)nk_n}{(q_n)^2} = 0.2(\log\log n)^5\log n. \tag{2.93}$$

By the translation invariance property of $\mathbb{T}_{q_n}(x,y)$ and Lemma 2.3.1.4, taking $l=y-x$, $r=q_n$ and $u=\log\log n$ for large enough n (and thus satisfying the conditions in Lemma 2.3.1.4) temporarily, we have

$$\mathbb{P}\left\{Z^+_n(0.9k_n, 1.1k_n) \ge \log\log n\right\} \le C|\mathcal{J}_n|\exp\left\{-\frac{(\log\log n)^2}{2}\right\} \to 0,$$

where $C>0$ is a constant. Since

$$\max_{0\le i\le n-k_n} \frac{S^+_{i+k_n}-S^+_i}{\sqrt{k_n}} \le Z^+_n(0.9k_n, 1.1k_n), \tag{2.94}$$

it follows that

$$\limsup_{n\to\infty} \mathbb{P}\left(\max_{0\le i\le n-k_n} \frac{S_{i+k_n}^+ - S_i^+}{\sqrt{k_n}} \ge \log\log n\right) = 0. \tag{2.95}$$

We may now prove the ignorability of maximum of $M_{i,j}^+$ when taking values on R_3. Define

$$\Omega_{1n} := \Omega_n \bigcap \left\{\max_{0\le i\le n-k_n} \frac{S_{i+k_n}^+ - S_i^+}{\sqrt{k_n}} \le \log\log n\right\}. \tag{2.96}$$

By (2.95), $\mathbb{P}(\Omega_{1n}) \to 1$ as $n\to\infty$. For $j-i \in R_3$,

$$\begin{aligned}
&A_{i,j}^{n+}(u_n)\\
&\supseteq \Omega_{1n} \bigcap \left\{S_j^+ - S_i^+ \le (j-i)\frac{S_{n+1}^+ - 1}{n} + \frac{u_n}{\sqrt{n}}(n+1-S_{n+1})w_{i,j}^n\right\}\\
&= \Omega_{1n} \bigcap \left\{S_j^+ - S_{i+k_n}^+ \le (j-i)\frac{S_{n+1}^+ - 1}{n} - S_{i+k_n}^+ + S_i^+ + \frac{u_n}{\sqrt{n}}(n+1-S_{n+1}^+)w_{i,j}^n\right\}\\
&\supseteq \Omega_{1n} \bigcap \left\{S_j^+ - S_{i+k_n}^+ \le -(j-i)\frac{\log\log n}{\sqrt{n}} - \sqrt{k_n}\log\log n + \frac{u_n}{\sqrt{n}}(n+1-\log\log n\sqrt{n})w_{i,j}^n\right\}\\
&\supseteq \Omega_{1n} \bigcap \left\{\frac{S_j^+ - S_{i+k_n}^+}{\sqrt{j-i-k_n}} \le \sqrt{\frac{j-i}{j-i-k_n}}\left[u_n\cdot\left(1-\frac{\log\log n}{\sqrt{n}}\right) - \sqrt{\frac{(\log\log n)^3}{\log n}} - \log\log n\right]\right\}\\
&\supseteq \Omega_{1n} \bigcap \left\{\frac{S_j^+ - S_{i+k_n}^+}{\sqrt{j-i-k_n}} \le \sqrt{1+\frac{k_n}{K_n}}\left[u_n\cdot\left(1-\frac{\log\log n}{\sqrt{n}}\right) - \sqrt{\frac{(\log\log n)^3}{\log n}} - \log\log n\right]\right\}\\
&\supseteq \Omega_{1n} \bigcap \left\{\frac{S_j^+ - S_{i+k_n}^+}{\sqrt{j-i-k_n}} \le u_n(\log\log n)\right\},
\end{aligned}$$

where the last line follows by noting that $k_n/K_n = 1/(\log\log n)^2$. Thus

$$\begin{aligned}
&\bigcap_{0\le i<j\le n:\ k_n+1\le j-i\le K_n} A_{i,j}^{n+}(u_n)\\
&\supset \Omega_{1n} \bigcap \left\{\max_{0\le i<j\le n:\ k_n+1\le j-i\le K_n} \frac{S_j^+ - S_{i+k_n}^+}{\sqrt{j-i-k_n}} \le u_n(\log\log n)\right\}\\
&\supset \Omega_{1n} \bigcap \left\{\max_{0\le i<j\le n:\ j-i\le K_n} \frac{S_j^+ - S_i^+}{\sqrt{j-i}} \le u_n(\log\log n)\right\},
\end{aligned}$$

and recall that $u_n(\cdot)$ is a function. Since $(\log n)^3 \ll K_n$, (2.80) and (2.81) together imply that

$$\begin{aligned}
&\liminf_{n\to\infty} \mathbb{P}\left\{M_n^+(k_n+1, K_n) \le u_n(\tau)\right\} \\
&\ge \liminf_{n\to\infty} \mathbb{P}\left[\Omega_{1n} \bigcap \{Z_n^+(1, K_n) \le u_n(\log\log n)\}\right] \\
&\ge \liminf_{n\to\infty} \mathbb{P}\left[\Omega_{1n} \bigcap \{Z_n^+(1, K_n) \le u_n(\tau')\}\right] = \exp\left(-\frac{8}{9\sqrt{\pi}} e^{-\tau'}\right),
\end{aligned}$$

for any τ, τ'. We now take $\tau' \to \infty$, yielding

$$\liminf_{n\to\infty} \mathbb{P}\left\{M_n^+(k_n+1, K_n) \le u_n(\tau)\right\} = \liminf_{\tau'\to\infty} \exp\left(-\frac{8}{9\sqrt{\pi}} e^{-\tau'}\right) = 1. \tag{2.97}$$

• Next we apply the Kolmogorov's Theorem to deal with R_4. Define the centered order statistics

$$\bar{U}_{(i)} = U_{(i)} - \frac{i}{n+1}. \tag{2.98}$$

Note that when n is large enough,

$$\begin{aligned}
A_{i,j}^{n+}(u_n) &= \left\{\bar{U}_{(j)} - \bar{U}_{(i)} \ge \frac{j-i}{n(n+1)} - \frac{u_n}{\sqrt{n}} w_{i,j}^n\right\} \\
&= \left\{\sqrt{n}(\bar{U}_{(j)} - \bar{U}_{(i)}) \ge \frac{j-i}{\sqrt{n}(n+1)} - u_n w_{i,j}^n\right\} \\
&\supseteq \left\{\sqrt{n}(\bar{U}_{(j)} - \bar{U}_{(i)}) \ge -0.9 u_n w_{i,j}^n\right\} \\
&\supseteq \left\{0.9 u_n w_{i,j}^n \ge \sqrt{n}(\bar{U}_{(j)} - \bar{U}_{(i)}) \ge -0.9 u_n w_{i,j}^n\right\} \\
&\supseteq \left\{2\sqrt{n}\max\{|\bar{U}_{(i)}|, |\bar{U}_{(j)}|\} \le 0.9 u_n w_{i,j}^n\right\}.
\end{aligned}$$

For (i,j) such that $j-i \in R_4$, $w_{i,j}^n$ is minimized at either $j-i = \frac{n\log\log n}{\log n}$ or $n - \frac{n\log\log n}{\log n}$. Consequently,

$$\bigcap_{0\le i<j\le n: j-i\in R_4} A^n_{i,j}(u_n) \supseteq \left\{ \sqrt{n} \max_{1\le i\le n}\{|\bar{U}_{(i)}|\} \le \frac{0.9u_n}{2} \min_{0\le i<j\le n: j-i\in R_4} w^n_{i,j} \right\}$$
$$= \left\{ \sqrt{n} \max_{1\le i\le n}\{|\bar{U}_{(i)}|\} \le \frac{0.9u_n}{2} \sqrt{\frac{\log\log n}{\log n}\left(1-\frac{\log\log n}{\log n}\right)} \right\}.$$

The Kolmogorov's Theorem states that for any $y \ge 0$,

$$\lim_{n\to\infty} \mathbb{P}\left(\sqrt{n} \max_{1\le i\le n} |\bar{U}_{(i)}| \le y \right) = K(y) := 1 - 2e^{-2y^2} + 2e^{-8y^2} - \cdots . \tag{2.99}$$

In particular, $(\sqrt{n}\max_{1\le i\le n}|\bar{U}_{(i)}|)$ is tight. Therefore, by the fact that

$$\frac{0.9u_n}{2}\sqrt{\frac{\log\log n}{\log n}\left(1-\frac{\log\log n}{\log n}\right)} \asymp \sqrt{\log\log n} \to \infty, \tag{2.100}$$

we obtain

$$\lim_{n\to\infty} \mathbb{P}\left\{ \bigcap_{0\le i<j\le n: j-i\in R_4} A^{n+}_{i,j}(u_n) \right\} = 1. \tag{2.101}$$

- For R_5, define $j' = n-j$ and $U'_{(j'+1)} = 1 - U_{(n+1-j'-1)} = 1 - U_{(j)}$. A simple change of indices gives

$$\begin{aligned} M^+_n(n-K_n, n) &= \max_{\substack{0\le i<j\le n\\ n-K_n\le j-i<n}} \frac{j-i-n(U_{(j)}-U_{(i)})}{\sqrt{(j-i)(1-\frac{j-i}{n})}} \\ &\le \max_{\substack{i,j'\ge 0\\ i+j'<K_n}} \frac{nU'_{(j'+1)}-(j'+1)+nU_{(i)}-i}{\sqrt{(i+j')(1-\frac{i+j'}{n})}} \\ &\le 1.01 \max_{\substack{i,j\ge 0\\ 1\le i+j<K_n}} \frac{nU'_{(j)}-j+nU_{(i)}-i}{\sqrt{i+j}} + 1.01 \end{aligned}$$

where the last inequality holds when n is large enough since $K_n \ll n$. Now, by the above statements, to prove

$$\limsup_{n\to\infty} \mathbb{P}(M_n^+(n-K_n,n) \geq u_n) = 0, \tag{2.102}$$

it suffices to prove

$$\limsup_{n\to\infty} \mathbb{P}\left(\max_{\substack{i,j\geq 0 \\ 1\leq i+j<K_n}} \frac{nU_{(i)} - i + nU'_{(j)} - j}{\sqrt{i+j}} \geq \sqrt{1.9\log n} \right) = 0. \tag{2.103}$$

Assuming $0/0 = 0$, observe that

$$\begin{aligned}
&\mathbb{P}\left(\max_{\substack{i,j\geq 0 \\ 1\leq i+j<K_n}} \frac{nU_{(i)} - i + nU'_{(j)} - j}{\sqrt{i+j}} \geq \sqrt{1.9\log n} \right) \\
&= \mathbb{P}\left\{ \max_{\substack{i,j\geq 0 \\ 1\leq i+j\leq K_n}} \left(\frac{nU_{(i)} - i}{\sqrt{i+j}} + \frac{nU'_{(j)} - j}{\sqrt{i+j}} \right) \geq \sqrt{1.9\log n} \right\} \\
&\leq \mathbb{P}\left\{ \max_{\substack{i,j\geq 0 \\ 1\leq i+j\leq K_n}} \left(\frac{nU_{(i)} - i}{\sqrt{i}} + \frac{nU'_{(j)} - j}{\sqrt{j}} \right) \geq \sqrt{1.9\log n} \right\} \\
&\leq \mathbb{P}\left(\max_{0\leq i\leq n} \frac{nU_{(i)} - i}{\sqrt{i}} + \max_{0\leq j\leq n} \frac{nU'_{(j)} - j}{\sqrt{j}} \geq \sqrt{1.9\log n} \right) \\
&\leq 2\,\mathbb{P}\left(\max_{0\leq i\leq n} \frac{nU_{(i)} - i}{\sqrt{i}} \geq \frac{\sqrt{1.9\log n}}{2} \right) \\
&\leq 2\,\mathbb{P}\left(\max_{0\leq i\leq n} \frac{nU_{(i)} - i}{\sqrt{i(1-i/n)}} \geq \frac{\sqrt{1.9\log n}}{2} \right).
\end{aligned}$$

However, [Eic79] showed that

$$\max_{0\leq i\leq n} \frac{nU_{(i)} - i}{\sqrt{i(1-i/n)}} \sim \sqrt{2\log\log n}, \tag{2.104}$$

which finishes the proof for R_5.

• Now combining all the results gives the lower bound, which, together with the upper bound, establishes the proof of Theorem 2.2.1.1. □

Proof of (2.20)

In what follows, we let

$$u_n = u_n(\tau) := \log n + \tau, \tag{2.105}$$

with τ fixed. Define

$$A_{i,j}^{n-}(u_n) = \left\{ \frac{n(U_{(j)} - U_{(i)}) - (j-i)}{\sqrt{(j-i)(1-\frac{j-i}{n})}} \le u_n \right\} = \left\{ U_{(j)} - U_{(i)} \le \frac{j-i}{n} + \frac{u_n}{\sqrt{n}} w_{i,j}^n \right\},$$

where $w_{i,j}^n$ is defined in (2.75), and note that

$$\left\{ M_n^- \le u_n \right\} = \bigcap_{0 \le i < j \le n} A_{i,j}^{n-}(u_n). \tag{2.106}$$

Step 1: Upper bound

For the upper bound, again, we only consider a particular order of magnitude for the length, the one that contributes the most to the maximum. When $j - i \le \frac{n \log\log n}{(\log n)^2}$,

$$\begin{aligned} A_{i,j}^{n-}(u_n) &\subset \Omega_n^{\mathsf{c}} \bigcup \{\Omega_n \bigcap A_{i,j}^{n-}(u_n)\} \\ &\subset \Omega_n^{\mathsf{c}} \bigcup \left\{ \frac{S_j^- - S_i^-}{\sqrt{j-i}} \le (\log\log n) \sqrt{\frac{j-i}{n}} + u_n \cdot \left(1 + \frac{\log\log n}{\sqrt{n}} \right) \right\} \\ &\subset \Omega_n^{\mathsf{c}} \bigcup \left\{ \frac{S_j^- - S_i^-}{\sqrt{j-i}} \le u_n(\tau + \varepsilon) \right\}, \end{aligned}$$

for any $\varepsilon > 0$, where Ω_n is given in (2.78). By (2.79), it suffices to consider the second event on the RHS. Applying Theorem 1.7 in [KW14], the limiting distribution of Z_n^- is the same as that of $\max_{1 \le i \le n}(-X_i)$. By the independence of $\{X_i\}$, we obtain

$$\lim_{n\to\infty} \mathbb{P}(Z_n^- \le u_n) = \lim_{n\to\infty} \mathbb{P}\{ \max_{1 \le i \le n} (-X_i) \le u_n \} = \exp\{-\exp(1-\tau)\}. \tag{2.107}$$

Therefore, taking $\varepsilon \to 0$,

$$\begin{aligned}
\limsup_{n\to\infty} \mathbb{P}(M_n^- \le u_n) &= \limsup_{n\to\infty} \mathbb{P}\left\{ \bigcap_{0\le i<j\le n} A_{i,j}^{n-}(u_n) \right\} \\
&\le \limsup_{n\to\infty} \mathbb{P}\left\{ \bigcap_{0\le i<j\le n: j-i\le \frac{n\log\log n}{(\log n)^2}} A_{i,j}^{n-}(u_n) \right\} + \mathbb{P}(\Omega_n^{\mathsf{c}}) \\
&\le \limsup_{\varepsilon\to 0} \exp\{-\exp(1-\tau-\varepsilon)\} \\
&= \exp\{-\exp(1-\tau)\}.
\end{aligned}$$

Step 2: Lower bound

As in the proof of (2.2.1.1), we divide the range of $j-i$ into several subintervals. Similar to the upper bound case,

$$\lim_{n\to\infty} \mathbb{P}\left\{ M_n^-\left(1, \frac{n\log\log n}{(\log n)^2}\right) \le u_n \right\} = \exp\{-\exp(1-\tau)\}. \tag{2.108}$$

With the same argument that was used to prove (2.101), we obtain

$$\lim_{n\to\infty} \mathbb{P}\left\{ \bigcap_{0\le i<j\le n: \frac{n\log\log n}{(\log n)^2} \le j-i\le n-\frac{n\log\log n}{(\log n)^2}} A_{i,j}^{n-}(u_n) \right\} = 1. \tag{2.109}$$

The case where $j-i \ge n - \frac{n\log\log n}{(\log n)^2}$ can be treated similarly to proving the region R_5 in the proof of Theorem 2.2.1.1, even easier since now $u_n \sim \log n$ (and details are omitted).

Proof of (2.21)

This follows directly from (2.15), where we learn that $M_n^+ \asymp_P \sqrt{\log n}$, and (2.20), which states that $M_n^- \asymp_P \log n$, which when combined imply that $M_n^- \gg_P M_n^+$, and therefore $M_n = \max(M_n^-, M_n^+) = M_n^-$ with probability tending to 1 as n increases.

2.3.3 Proof of Theorem 2.2.2.1

Proof of (2.25)

We first derive the asymptotic distribution of

$$\tilde{M}_n^+(1,2) = \max_{0 \le i \le n-1} \frac{1 - n(U_{(i+1)} - U_{(i)})}{\sqrt{n(U_{(i+1)} - U_{(i)})(1 - U_{(i+1)} + U_{(i)})}}, \tag{2.110}$$

which is exactly the same as that of (2.25) and then show that $\tilde{M}_n^+(2,n) \ll_P \sqrt{n}$. These together imply (2.25). To get the asymptotic distribution of $\tilde{M}_n^+(1,2)$, note that

$$\tilde{M}_n^+(1,2) \le \max_{0 \le i \le n-1} \frac{1}{\sqrt{n(U_{(i+1)} - U_{(i)})[1 - (U_{(i+1)} - U_{(i)})]}} \tag{2.111}$$

$$\text{and } \tilde{M}_n^+(1,2) \ge \max_{0 \le i \le n-1} \frac{1 - n(U_{(i+1)} - U_{(i)})}{\sqrt{n(U_{(i+1)} - U_{(i)})}}, \tag{2.112}$$

where both upper and lower bounds are functions of

$$T := \min_{0 \le i \le n-1} (U_{(i+1)} - U_{(i)}). \tag{2.113}$$

Therefore it suffices to work on T instead. It is easy to see that $T \le 1/n$. By symmetry,

$$\mathbb{P}(T \ge t) = n!\,\mathbb{P}(T \ge t, U_1 \le U_2 \le \cdots \le U_n). \tag{2.114}$$

Define the subset

$$A_t = \{(u_1, \ldots, u_n) \in [0,1]^n : u_i + t \le u_{i+1}, i = 0, 1, \ldots, n-1\}, \tag{2.115}$$

where $u_0 = 0$. Then,

$$\{(U_1, \cdots, U_n) \in A_t\} = \{T \geq t, U_1 \leq U_2 \leq \cdots \leq U_n\}, \tag{2.116}$$

and hence

$$\mathbb{P}(T \geq t, U_1 \leq U_2 \leq \cdots \leq U_n) = \lambda_n(A_t), \tag{2.117}$$

where λ_n is the Lebesgue measure on $\mathbb{R}^n$. Define a mapping

$$h: \quad A_t \longrightarrow Q \subset [0, 1-nt]^n, \quad h(u_1, u_2, \cdots, u_n) = (u_1 - t, u_2 - 2t, u_n - nt), \tag{2.118}$$

where

$$Q := \{(y_1, \ldots, y_n) : y_i \leq y_{i+1}, \forall\ 1 \leq i \leq n-1\} \cap [0, 1-nt]^n. \tag{2.119}$$

It is easy to verify that h is a volume-preserving bijection. Hence

$$\mathbb{P}(T \geq t, U_1 \leq U_2 \leq \cdots \leq U_n) = \lambda_n(A_t) = \lambda_n(Q) = \frac{(1-nt)^n}{n!} \tag{2.120}$$

Therefore, we have

$$\mathbb{P}(T \geq t) = \frac{n!(1-nt)^n}{n!} = (1-nt)^n, \tag{2.121}$$

for $0 \leq t \leq 1/n$. For any $0 \leq t \leq 1/n$,

$$\mathbb{P}\left\{ \min_{0 \leq i \leq n-1} (U_{(i+1)} - U_{(i)}) \geq t \right\} = (1-nt)^n, \tag{2.122}$$

which implies

$$\lim_{n \to \infty} \mathbb{P}\left\{ \min_{0 \leq i \leq n-1} (U_{(i+1)} - U_{(i)}) \geq \frac{\tau}{n^2} \right\} = \exp(-\tau). \tag{2.123}$$

This, together with (2.111) and (2.112), implies

$$\lim_{n\to\infty} \mathbb{P}\left(\tilde{M}_n^+(1,1) \le \sqrt{\frac{n}{\tau}}\right) = \exp(-\tau). \tag{2.124}$$

It remains to show that $\tilde{M}_n^+(2,n) \ll_P \sqrt{n}$. We will divide the region $(2,n)$ into

$$(2,\ (\log n)^2),\ ((\log n)^2,\ n-(\log n)^2)\ \mathit{and}\ (n-(\log n)^2,\ n). \tag{2.125}$$

When $2 \le j-i \le (\log n)^2$, note that

$$\begin{aligned} 1-(U_{(j)}-U_{(i)}) &= 1-\frac{j-i}{n+1}-(\bar{U}_{(j)}-\bar{U}_{(i)}) && (2.126)\\ &\ge 1-\frac{(\log n)^2}{n+1}-2\max_{1\le i\le n}|\bar{U}_{(i)}| \\ &= 1+O_P(1/\sqrt{n}) \\ &\ge 0.5, && (2.127) \end{aligned}$$

where the last inequality holds on a sequence of events with probability tending to one, by Kolmogorov's Theorem mentioned in the proof of Theorem 2.2.1.1 when n is large enough. Meanwhile,

$$\begin{aligned} \frac{j-i-n(U_{(j)}-U_{(i)})}{\sqrt{n(U_{(j)}-U_{(i)})}} &= \frac{j-i-\frac{n}{n+1-S_{n+1}^+}(j-i-S_j^+ + S_i^+)}{\sqrt{n\frac{n}{n+1-S_{n+1}^+}(j-i-S_j^+ + S_i^+)}} \\ &= (1+O_P(1/\sqrt{n}))\tilde{Z}_{i,j}+O_P(1/\sqrt{n}) \\ &\le 1.01\tilde{Z}_{i,j}+0.01, && (2.128) \end{aligned}$$

on the sequence of events Ω_n defined in (2.79). With these results, the union bound, (2.43) and the fact that $I^+(s) = -s - \log(1-s)$ on $[0,1)$, for any $\varepsilon > 0$,

$$\begin{aligned}
&\mathbb{P}(\tilde{M}_n^+(2,(\log n)^2) \geq \varepsilon\sqrt{n}) \\
&\leq \mathbb{P}(\tilde{Z}_n^+(2,(\log n)^2) \geq 0.9\varepsilon\sqrt{n}) + \mathbb{P}(\Omega_n^c) \\
&\leq \sum_{0\leq i<j\leq n: 2\leq j-i\leq(\log n)^2} \mathbb{P}(\tilde{Z}_{i,j}^+ \geq 0.9\varepsilon\sqrt{n}) + \mathbb{P}(\Omega_n^c) \\
&\leq n \sum_{2\leq k\leq(\log n)^2} \mathbb{P}\left(\frac{S_k^+}{\sqrt{k-S_k^+}} \geq 0.9\varepsilon\sqrt{n}\right) + \mathbb{P}(\Omega_n^c) \\
&\leq n \sum_{2\leq k\leq(\log n)^2} \exp\left[-kI^+\left\{g^+\left(\frac{0.9\varepsilon\sqrt{n}}{\sqrt{k}}\right\}\right]\right\} + \mathbb{P}(\Omega_n^c) \\
&\leq n \sum_{2\leq k\leq(\log n)^2} \exp\left[kg^+\left(\frac{0.9\varepsilon\sqrt{n}}{\sqrt{k}}\right) + k\log\left\{1-g^+\left(\frac{0.9\varepsilon\sqrt{n}}{\sqrt{k}}\right)\right\}\right] + \mathbb{P}(\Omega_n^c).
\end{aligned}$$

As $a \to \infty$, $0.9\varepsilon\sqrt{n}/\sqrt{k} \to \infty$ and $g^+(a) \uparrow 1$. In addition,

$$1-g^+(a) = 1 - \frac{a(\sqrt{a^2+4}-a)}{2} = 1 - \frac{2a}{\sqrt{a^2+4}+a} = \frac{\sqrt{a^2+4}-a}{\sqrt{a^2+4}+a} = \frac{4}{(\sqrt{a^2+4}+a)^2}. \quad (2.129)$$

Note that

$$\frac{0.9}{a^2} \leq \frac{4}{(\sqrt{a^2+4}+a)^2} \leq \frac{1}{a^2}, \quad (2.130)$$

when a is large enough. Therefore, when n is sufficiently large,

$$\begin{aligned}
\mathbb{P}(\tilde{M}_n^+(2,(\log n)^2) \geq \varepsilon\sqrt{n}) &\leq n \sum_{2\leq k\leq(\log n)^2} \exp\left\{k - k\log\left(\frac{0.9\varepsilon n}{k}\right)\right\} \\
&\leq n \sum_{2\leq k\leq(\log n)^2} \exp(-0.9k\log n) \\
&\leq n \sum_{2\leq k\leq(\log n)^2} \exp(-1.8\log n) \to 0,
\end{aligned}$$

where the last inequality uses that $k \geq 2$.

When $(\log n)^2 \leq j-i \leq n-(\log n)^2$, by Theorem 2.2.1.1 and Theorem 2.2.1.2, we have

$$U_{(j)} - U_{(i)} \leq \frac{j-i}{n} + \frac{1.01\log n}{\sqrt{n}} w_{i,j}^n, \tag{2.131}$$

$$1-(U_{(j)} - U_{(i)}) \geq 1 - \frac{j-i}{n} - \frac{1.01\log n}{\sqrt{n}} w_{i,j}^n, \tag{2.132}$$

$$U_{(j)} - U_{(i)} \geq \frac{j-i}{n} - \frac{1.01\log n}{\sqrt{n}} w_{i,j}^n, \tag{2.133}$$

and

$$1-(U_{(j)} - U_{(i)}) \leq 1 - \frac{j-i}{n} + \frac{1.01\log n}{\sqrt{n}} w_{i,j}^n, \tag{2.134}$$

with probability tending to one. Together, (2.131) and (2.133) lead to

$$\left| \frac{n(U_{(j)} - U_{(i)})}{j-i} \right| = O_P(1), \tag{2.135}$$

uniformly in (i,j) satisfying $j-i \geq (\log n)^2$. (2.132) and (2.134) imply

$$\left| \frac{1-(U_{(j)} - U_{(i)})}{1-(j-i)/n} \right| = O_P(1). \tag{2.136}$$

These, combined with the definitions of M_n^+ and $\tilde{M}_n^+$, imply

$$\tilde{M}_n^+\{(\log n)^2, n-(\log n)^2\} \asymp_P M_n^+\{(\log n)^2, n-(\log n)^2\}. \tag{2.137}$$

By Theorem 2.2.1.1, it follows that for any $\varepsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}[\tilde{M}_n^+\{(\log n)^2, n-(\log n)^2\} \geq \varepsilon\sqrt{n}] = 0. \tag{2.138}$$

Finally, when $n-(\log n)^2 \leq j-i \leq n$, define $j' = n-j$ and thus $U'_{(j'+1)} = 1-U_{(n+1-j'-1)} =$

$1 - U_{(j)}$. A simple change of indices gives

$$\begin{aligned}
&\tilde{M}_n^+(n-(\log n)^2, n) \\
&= \max_{\substack{0\le i<j\le n \\ n-(\log n)^2\le j-i\le n}} \frac{j-i-n(U_{(j)}-U_{(i)})}{\sqrt{n(U_{(j)}-U_{(i)})(1-(U_{(j)}-U_{(i)}))}} \\
&= \max_{\substack{i,j'\ge 0 \\ i+j'\le(\log n)^2}} \frac{nU'_{(j'+1)}-(j'+1)+nU_{(i)}-i}{\sqrt{n(U_{(i)}+U'_{(j'+1)})(1-U_{(i)}-U'_{(j'+1)})}} \\
&= \max_{\substack{i,j\ge 0 \\ 1\le i+j\le(\log n)^2}} \frac{nU_{(i)}-i+nU'_{(j)}-j}{\sqrt{n(U_{(i)}+U'_{(j)})(1-U_{(i)}-U'_{(j)})}} + O_P(1).
\end{aligned}$$

Notice that when $i, j \ge 0$ and $1 \le i + j \le (\log n)^2$,

$$1-U_{(i)}-U'_{(j)} > 1-2\max_{0\le i\le(\log n)^2} U_{(i)} > 0.5, \tag{2.139}$$

with probability tending to one, which can be seen by a simple application of Kolmogorov's Theorem. By a similar speech when proving R_5 in the proof of Theorem 2.2.1.1,

$$\begin{aligned}
&\mathbb{P}\left(\max_{\substack{i,j\ge 0 \\ 1\le i+j\le(\log n)^2}} \frac{nU_{(i)}-i+nU'_{(j)}-j}{\sqrt{n(U_{(i)}+U'_{(j)})(1-U_{(i)}-U'_{(j)})}} \ge \varepsilon\sqrt{n}\right) && (2.140) \\
&\le \mathbb{P}\left(\max_{\substack{i,j\ge 0 \\ 1\le i+j\le(\log n)^2}} \frac{nU_{(i)}-i+nU'_{(j)}-j}{\sqrt{n(U_{(i)}+U'_{(j)})}} \ge 0.5\varepsilon\sqrt{n}\right) && (2.141) \\
&\le 2\,\mathbb{P}\left(\max_{0\le i\le(\log n)^2} \frac{nU_{(i)}-i}{\sqrt{nU_{(i)}}} \ge 0.25\varepsilon\sqrt{n}\right) && (2.142) \\
&\le 2\,\mathbb{P}\left(\max_{0\le i\le(\log n)^2} \frac{nU_{(i)}-i}{\sqrt{nU_{(i)}(1-U_{(i)})}} \ge 0.25\varepsilon\sqrt{n}\right) && (2.143) \\
&\to 0, && (2.144)
\end{aligned}$$

where the last line again follows from [Eic79]. These eventually establish the proof of (2.25).

Proof of (2.26)

The roadmap of our proof.

To derive the asymptotic distribution, we first focus on the most contributed part, i.e., those with length $j-i=l_n \sim a\log^3 n$ for $a>0$. Define

$$u_n = u_n(\tau) := \sqrt{2\log n}\left(1+\frac{-3\log\log n+2\tau}{4\log n}\right). \tag{2.145}$$

For any two constants $0<A_1<A_2<\infty$, define $l_n^- = A_1\log^3 n$ and $l_n^+ = A_2\log^3 n$. We prove

$$\lim_{n\to\infty}\mathbb{P}\{\tilde{M}_n^+(l_n^-,l_n^+)\le u_n\} = \exp\left\{-e^{-\tau}\int_{A_1}^{A_2}\Lambda_1(a)da\right\}. \tag{2.146}$$

It turns out that to prove (2.146), within that region, it suffices to focus on

$$\tilde{Z}_{i,j}^+ := \frac{S_j^+ - S_i^+}{\sqrt{j-i-(S_j^+ - S_i^+)}}, \tag{2.147}$$

instead, up to restricting on subset Ω_n defined in (2.79). Write

$$\tilde{Z}_n^+(k,l) = \max_{0\le i<j\le n: k\le j-i\le l} \tilde{Z}_{i,j}^+, \tag{2.148}$$

and

$$\tilde{Z}_n^+ = \tilde{Z}_n^+(1,n). \tag{2.149}$$

We will use Lemma 2.3.1.5 to show that

$$Q_n := \mathbb{P}\left(\max_{(i,j)\in\mathbb{T}_{Bq_n}(x,x+l_n)} \tilde{Z}_{i,j}^+ \ge u_n\right) \sim P_n(0)\left\{1+H^2\left(\frac{B}{a}\right)\right\}, \tag{2.150}$$

where $B\ge 1$ is an integer and the quantities $P_n(0)$, $H(x)$, q_n will be specified later. Next, with a domain $\mathbb{J}_n(z)$ (to be specified) larger than $\mathbb{T}_{Bq_n}$, we will show that

$$\mathbb{P}\left(\max_{(i,j)\in\mathbb{J}_n(z)} \tilde{Z}^+_{i,j} \geq u_n\right) \sim e^{-\tau}\frac{w_n}{n}\int_{A_1}^{A_2}\Lambda_1(a)da, \tag{2.151}$$

which no longer depends on B, with $\Lambda_1(a)$ defined in the theorem part. This enables us to apply Poisson limit theorem in [AGG89] to get

$$\lim_{n\to\infty}\mathbb{P}\{\tilde{Z}^+_n(l^-_n,l^+_n) \leq u_n\} = \exp\left\{-e^{-\tau}\int_{A_1}^{A_2}\Lambda_1(a)da\right\}. \tag{2.152}$$

The final step will be showing that the region beyond $A_2(\log n)^3$ is negligible, that is,

$$\limsup_{A_2\to\infty}\limsup_{n\to\infty}\mathbb{P}\{\tilde{M}^+_n(l^+_n,n) \geq u_n\} = 0. \tag{2.153}$$

Therefore setting $A_1 = A$ and letting $A_2 \to \infty$ yield (2.26).

We first argue why we can focus on (2.51) instead when $j-i \asymp \log^3 n$. Note that (2.127) and (2.128) continue to hold when $j-i \asymp (\log n)^3$. Hence,

$$\tilde{M}^+_n(l^-_n,l^+_n) = \{1+O_P(1/\sqrt{n})\}\tilde{Z}^+_n(l^-_n,l^+_n) + O_P(1/\sqrt{n}), \tag{2.154}$$

which implies

$$\mathbb{P}\{\tilde{Z}^+_n(l^-_n,l^+_n) \leq u_n(\tau-\varepsilon)\} \leq \mathbb{P}\{\tilde{M}^+_n(l^-_n,l^+_n) \leq u_n(\tau)\} \leq \mathbb{P}\{\tilde{Z}^+_n(l^-_n,l^+_n) \leq u_n(\tau+\varepsilon)\},$$

for any $\varepsilon > 0$. If we had established (2.152), taking $\varepsilon \to 0$ would yield (2.146). Now we turn to the mainstream of the proof.

PROOF OF (2.150). We will prove this following a similar strategy as in [KW14]. Necessary adjustments are still needed since [KW14] focused on $Z^+_{i,j}$ while we are dealing with $\tilde{Z}^+_{i,j}$. We will present the parts that need to be adjusted and refer to their results when nothing needs to be changed.

First we work on Q_n. For any $\tau \in \mathbb{R}$ and $a \geq 0$, let $l_n = a(\log n)^3$ and define

$$P_n(s) = \mathbb{P}\left(\frac{S_{l_n}^+}{\sqrt{l_n - S_{l_n}^+}} \geq u_n - \frac{s}{u_n} \right). \tag{2.155}$$

Define

$$b_n := \frac{u_n - s/u_n}{\sqrt{l_n}}, \tag{2.156}$$

for ease of notation. Since $u_n^3 \propto \sqrt{l_n}$ and $b_n \sim \sqrt{2/a}/\log n \to 0$, for fixed $s > 0$ with sufficiently large n, with the transformation (2.55), Lemma 2.3.1.1 and Taylor's expansion

$$\begin{aligned} P_n(s) &= \mathbb{P}\left\{ \frac{S_{l_n}^+}{\sqrt{l_n}} \geq \sqrt{l_n} g^+(b_n) \right\} \\ &\sim \frac{1}{\sqrt{2\pi} u_n} \exp\left\{ -\frac{(u_n - s/u_n)^2}{2} \frac{2I^+(g^+(b_n))}{b_n^2} \right\} \\ &= \frac{1}{\sqrt{2\pi} u_n} \exp\left\{ -\frac{(u_n - s/u_n)^2}{2} \left(1 - \frac{1}{3} b_n \right) + o(1) \right\} \\ &\sim \frac{1}{2\sqrt{\pi}} e^{s + \frac{\sqrt{2}}{3} a^{-1/2}} \frac{e^{-\tau} \log n}{n}. \end{aligned} \tag{2.157}$$

Recall that $\mathbb{T}_r(x,y)$ is defined in (2.46). Define $q_n = (\log n)^2$. By the same techniques in the proof of Lemma 2.3.1.7 we have

$$\begin{aligned} Q_n &= \mathbb{P}\left(\max_{(i,j) \in \mathbb{T}_{Bq_n}(x, x+l_n)} \tilde{Z}_{i,j}^+ \geq u_n \right) \\ &= \mathbb{P}\left[\max_{(i,j) \in \mathbb{T}_{Bq_n}(x, x+l_n)} \left\{ S_j^+ - S_i^+ - (j-i) g^+\left(\frac{u_n}{j-i} \right) \right\} \geq 0 \right] \\ &= \mathbb{P}\left[\max_{0 \leq k_1, k_2 \leq Bq_n} \left\{ S_{k_1}^{(1)+} + S_{k_2}^{(2)+} - (l_n + k_1 + k_2) g^+\left(\frac{u_n}{l_n + k_1 + k_2} \right) \right\} + S_{l_n}^+ \geq 0 \right] \\ &= P_n(0) \left\{ 1 + \int_0^\infty G_n(s) d\nu_n(s) \right\}, \end{aligned}$$

where $P_n(s)$ defined in (2.155) is actually the probability distribution of V_{l_n, u_n}, defined in (2.61).

Therein

$$
\begin{aligned}
G_n(s) := \mathbb{P}\Bigg[\max_{0 \le k_1, k_2 \le Bq_n} \Bigg\{ S_{k_1}^{(1)+} + S_{k_2}^{(2)+} - (l_n + k_1 + k_2) g^+ \left(\frac{u_n}{\sqrt{l_n + k_1 + k_2}} \right) \Bigg\} \\
+ l_n \cdot g^+ \left(\frac{u_n - s/u_n}{\sqrt{l_n}} \right) \ge 0 \Bigg],
\end{aligned}
$$

and

$$
\nu_n(\cdot) := P_n(\cdot)/P_n(0). \tag{2.158}
$$

It is immediate that the first and second conditions in Lemma 2.3.1.5 hold by directly mimicking the details in the proof of Lemma 4.3 in [KW14], that is, for any fixed $s > 0$ and any sequence $s_n \to s$,

$$
\lim_{n \to \infty} G_n(s_n) = \mathbb{P}(M_1 + M_2 \ge s), \tag{2.159}
$$

and

$$
\lim_{n \to \infty} \nu_n([0, s)) = \lim_{n \to \infty} \frac{P_n(s)}{P_n(0)} = e^s. \tag{2.160}
$$

M_1 and M_2 are independent copies with the same distribution as

$$
M = \sup_{t \in [0, a^{-1}B]} \{ \sqrt{2} W(t) - t \}, \tag{2.161}
$$

where $W(t)$ is a standard Brownian motion (similar but more detailed arguments can be found in the proof of lemma 4.3 in [Kab11]). To verify the third condition in Lemma 2.3.1.5, we need to bound the integral $\int_0^\infty G_n(s) d\nu_n(s)$ from above. This can be immediately completed by using Lemma 2.3.1.7. Hence applying Lemma 2.3.1.5 completes the proof of (2.150), where

$$
H(x) := \mathbb{E}\{ \sup_{t \in [0,x]} e^{\sqrt{2}W(t) - t} \}, x > 0, \tag{2.162}
$$

therein.

PROOF OF (2.151). Define $w_n = (\log n)^3$. For $z \in \mathbb{Z}$, define

$$\mathbb{J}_n(z) = \{(i,j) \in \mathbb{I} : z \leq i < z + w_n, j - i \in [l_n^-, l_n^+]\}. \tag{2.163}$$

To derive the rate of $\mathbb{P}(\max_{(i,j)\in\mathbb{J}_n(z)} \tilde{Z}_{i,j}^+ \geq u_n)$, by translation invariance we may take $z = 0$. Let δ_n be a real sequence satisfying $\delta_n = o(w_n)$ and $q_n = o(\delta_n)$, e.g. $\delta_n = (\log n)^{2.5}$. For $B \in \mathbb{N}$, we introduce the following two-dimensional discrete grids with mesh size Bq_n:

$$\mathcal{J}_n(B) = \{(x,y) \in Bq_n\mathbb{Z} \times Bq_n\mathbb{Z} : x \in [-\delta_n, w_n + \delta_n], y - x \in [l_n^- - \delta_n, l_n^+ + \delta_n]\}, \tag{2.164}$$

$$\mathcal{J}_n'(B) = \{(x,y) \in Bq_n\mathbb{Z} \times Bq_n\mathbb{Z} : x \in [\delta_n, w_n - \delta_n], y - x \in [l_n^- + \delta_n, l_n^+ - \delta_n]\}. \tag{2.165}$$

By Bonferroni inequality,

$$S_n'(B) - S_n''(B) \leq \mathbb{P}\left(\max_{(i,j)\in\mathbb{J}_n(0)} \tilde{Z}_{i,j}^+ \geq u_n\right) \leq S_n(B), \tag{2.166}$$

where

$$S_n(B) = \sum_{(x,y)\in\mathcal{J}_n(B)} \mathbb{P}\left(\max_{(i,j)\in\mathbb{T}_{Bq_n}(x,y)} \tilde{Z}_{i,j}^+ \geq u_n\right), \tag{2.167}$$

$$S_n'(B) = \sum_{(x,y)\in\mathcal{J}_n'(B)} \mathbb{P}\left(\max_{(i,j)\in\mathbb{T}_{Bq_n}(x,y)} \tilde{Z}_{i,j}^+ \geq u_n\right), \tag{2.168}$$

and

$$S_n''(B) = \sum_{(x_1,y_1),(x_2,y_2)} \mathbb{P}\left(\max_{(i,j)\in\mathbb{T}_{Bq_n}(x_1,y_1)} \tilde{Z}_{i,j}^+ \geq u_n, \max_{(i,j)\in\mathbb{T}_{Bq_n}(x_2,y_2)} \tilde{Z}_{i,j}^+ \geq u_n\right), \tag{2.169}$$

where the summation is taken over $(x_1,y_1) \neq (x_2,y_2) \in \mathcal{J}_n'(B)$. As long as we can show

$$\lim_{B\to\infty} \limsup_{n\to\infty} n w_n^{-1} S_n(B) \leq e^{-\tau} \int_{A_1}^{A_2} \Lambda_1(a) da, \tag{2.170}$$

$$\lim_{B\to\infty}\liminf_{n\to\infty} nw_n^{-1}S_n'(B) \geq e^{-\tau}\int_{A_1}^{A_2}\Lambda_1(a)da, \tag{2.171}$$

and

$$\lim_{B\to\infty}\limsup_{n\to\infty} nw_n^{-1}S_n''(B) = 0, \tag{2.172}$$

(2.151) will follow immediately. The proof of (2.171) is almost identical to that of (2.170), so we only focus on proving (2.170) based on the dominated convergence theorem. Define

$$\mathcal{L}_n(B) = Bq_n\mathbb{Z}\cap[l_n^- - \delta_n, l_n^+ + \delta_n], \tag{2.173}$$

such that $|\mathcal{L}_n(B)| \sim (A_2 - A_1)(\log n)/B$. Since the probability on the right-hand side of (2.167) depends only on $l := y - x$, by translation invariance we have

$$S_n(B) \leq \frac{w_n + \delta_n}{Bq_n}\sum_{l\in\mathcal{L}_n(B)}\mathbb{P}\left(\max_{(i,j)\in T_{Bq_n}(0,l)}\tilde{Z}_{i,j}^+ \geq u_n\right). \tag{2.174}$$

Next we apply (2.150) to bound each probability with l fixed and replace the summation $(Bq_n)^{-1}\sum_{l\in\mathcal{L}_n(B)}$ by an integral as $n\to\infty$. By (2.150) and (2.157),

$$\lambda_{n,B}(a) := \frac{n}{\log n}\mathbb{P}\left(\max_{(i,j)\in T_{Bq_n}(0,l_{n,B}(a))}\tilde{Z}_{i,j}^+ \geq u_n\right) \to \frac{1}{2\sqrt{\pi}}e^{\frac{\sqrt{2}}{3}a^{-1/2}-\tau}\left\{1 + H^2\left(\frac{B}{a}\right)\right\}, \tag{2.175}$$

as $n\to\infty$, where

$$l_{n,B}(a) = \max\{l\in Bq_n\mathbb{Z} : l \leq aw_n\}. \tag{2.176}$$

The function $\lambda_{n,B}(a)$ takes constant values on sub-intervals with widths $Bq_n/w_n = B/\log n$. It follows that

$$S_n(B) \leq \frac{w_n+\delta_n}{B^2 n}\sum_{l\in\mathcal{L}_n(B)}\frac{B\lambda_{n,B}(a)}{\log n} = \frac{w_n+\delta_n}{B^2 n}\int_{A_1-\frac{2\delta_n}{w_n}}^{A_2+\frac{2\delta_n}{w_n}}\lambda_{n,B}(a)da. \tag{2.177}$$

From Lemma 2.3.1.7, we can upper bound the integrand $\lambda_{n,B}(a)$ by an integrable function that is independent of n. Therefore, applying Fatou's lemma on lim sup gives

$$\limsup_{n\to\infty} n w_n^{-1} S_n(B) \le e^{-\tau} \int_{A_1}^{A_2} \frac{a^2 \Lambda_1(a)}{B^2} \left\{ 1 + H^2\left(\frac{B}{a}\right) \right\} da. \tag{2.178}$$

This result holds for any $B \in \mathbb{N}$. Note that $\lim_{B\to\infty} H(B)/B = 1$. Letting $B \to \infty$, we arrive at (2.170).

To prove (2.172), we bound $S_n''(B)$ by similar quantities of $Z_{i,j}^+$, which allows us to use results in [KW14] immediately. For any interval (x,y) define the event

$$E_n(x,y) = \left\{ \max_{(i,j)\in\mathbb{T}_{Bq_n}(x,y)} \tilde{Z}_{i,j}^+ \ge u_n \right\}. \tag{2.179}$$

Note that

$$\frac{g^+(x)}{x} = \frac{1}{2}(\sqrt{x^2+4} - x) \ge 1 - \frac{x}{2}, \text{ when } x \to 0. \tag{2.180}$$

When $y - x \propto (\log n)^3$, $u_n/(y-x) \propto 1/(\log n)$,

$$\begin{aligned} E_n(x,y) &= \left\{ \max_{0\le l_1,l_2\le Bq_n} \left\{ S_{y+l_2}^+ - S_{x-l_1}^+ - (y-x+l_1+l_2) g^+\left(\frac{u_n}{\sqrt{y-x+l_1+l_2}} \right) \right\} \ge 0 \right\} \\ &\subset \left\{ \max_{0\le l_1,l_2\le Bq_n} \frac{S_{y+l_2}^+ - S_{x-l_1}^+}{\sqrt{y-x+l_1+l_2}} \ge \sqrt{y-x+l_1+l_2} g^+\left(\frac{u_n}{\sqrt{y-x+l_1+l_2}} \right) \right\} \\ &\subset \left\{ \max_{(i,j)\in\mathbb{T}_{Bq_n}(x,y)} Z_{i,j}^+ \ge u_n(\tau)\left(1 - \frac{u_n}{2\sqrt{y-x+l_1+l_2}} \right) \right\} \\ &\subset \left\{ \max_{(i,j)\in\mathbb{T}_{Bq_n}(x,y)} Z_{i,j}^+ \ge u_n(\tau - 0.1) \right\}. \end{aligned}$$

Therefore,

$$\begin{aligned} &\mathbb{P}\{E_n(i_1,j_1) \cap E_n(i_2,j_2)\} \\ &\le \mathbb{P}\left[\left\{ \max_{(i,j)\in\mathbb{T}_{Bq_n}(i_1,j_1)} Z_{i,j}^+ \ge u_n(\tau-0.1) \right\} \bigcap \left\{ \max_{(i,j)\in\mathbb{T}_{Bq_n}(i_2,j_2)} Z_{i,j}^+ \ge u_n(\tau-0.1) \right\} \right]. \end{aligned}$$

This allows us to work on $Z^{+}_{i,j}$ instead. Directly applying Lemma 4.12, Lemma 4.14, Lemma 4.15 and Lemma 4.16 in [KW14] yields (2.172).

PROOF OF (2.152). We will temporarily adopt the notations in [AGG89]. Define

$$I = \{\alpha \in \mathbb{N} : \alpha w_n \leq n\}, \tag{2.181}$$

which implies $|I| \leq n/w_n$. For any $\alpha \in I$, define

$$X_\alpha = \mathbb{1}\{\max_{(i,j)\in \mathbb{J}_n(\alpha w_n)} \tilde{Z}^{+}_{i,j} \geq u_n\}, \tag{2.182}$$

$$p_\alpha = \mathbb{P}(X_\alpha), \tag{2.183}$$

and

$$B_\alpha = \{\beta \in I : |(\beta - \alpha) w_n| \leq l^{+}_n + w_n\}. \tag{2.184}$$

Hence $|B_\alpha| \leq A_2 + 1$. To apply Theorem 1 in [AGG89], we need to show that

$$b_1 := \sum_{\alpha \in I} \sum_{\beta \in B_\alpha} p_\alpha p_\beta, \tag{2.185}$$

$$b_2 := \sum_{\alpha \in I} \sum_{\alpha \neq \beta \in B_\alpha} p_{\alpha\beta}, \text{ where } p_{\alpha\beta} := \mathbb{E}(X_\alpha X_\beta), \tag{2.186}$$

and

$$b'_3 := \sum_{\alpha \in I} s'_\alpha \tag{2.187}$$

therein vanish as $n \to \infty$, where

$$s'_\alpha := \mathbb{E}\left|\mathbb{E}\left(X_\alpha - p_\alpha \middle| \sum_{\beta \in I - B_\alpha} X_\beta\right)\right| \tag{2.188}$$

By the definition of B_α, $X_\alpha - p_\alpha$ and $\sum_{\beta \in I - B_\alpha} X_\beta$ are independent. Hence $s'_\alpha = 0$, so is b'_3. It follows from (2.151) that

$$b_1 \sim |I||B_\alpha| p_\alpha p_\beta \to 0. \tag{2.189}$$

With slight modification on (2.151),

$$\mathbb{P}\left(\max_{(i,j)\in \mathbb{J}_n(\alpha w_n)\cup \mathbb{J}_n(\beta w_n)} \tilde{Z}^+_{i,j} \geq u_n \right) \sim e^{-\tau} \frac{2w_n}{n} \int_{A_1}^{A_2} \Lambda_1(a)da. \tag{2.190}$$

This, together with (2.151), implies

$$p_{\alpha\beta} = \mathbb{P}\left(\max_{(i,j)\in \mathbb{J}_n(\alpha w_n)} \tilde{Z}^+_{i,j} \geq u_n, \max_{(i,j)\in \mathbb{J}_n(\beta w_n)} \tilde{Z}^+_{i,j} \geq u_n \right) = o\left(\frac{w_n}{n}\right). \tag{2.191}$$

Thus,

$$b_2 \leq |I||B_\alpha| \max_{\alpha\neq\beta} p_{\alpha\beta} \to 0. \tag{2.192}$$

Now, by Theorem 1 in [AGG89],

$$\lim_{n\to\infty} \mathbb{P}\{\tilde{Z}^+_n(l^-_n, l^+_n) \leq u_n\} = \lim_{n\to\infty} \mathbb{P}\left(\sum_{\alpha\in I} X_\alpha = 0 \right) = e^{-\lambda}, \tag{2.193}$$

where

$$\lambda = \sum_{\alpha\in I} p_\alpha \to e^{-\tau} \int_{A_1}^{A_2} \Lambda_1(a)da. \tag{2.194}$$

Therefore,

$$\lim_{n\to\infty} \mathbb{P}\{\tilde{M}^+_n(l^-_n, l^+_n) \leq u_n\} = \exp\left(-e^{-\tau} \int_{A_1}^{A_2} \Lambda_1(a)da \right), \tag{2.195}$$

by the statement in the beginning of our proof.

PROOF OF (2.153). Divide $(l^+_n, n]$ into

$$(l^+_n,\ (\log n)^4],\ ((\log n)^4,\ n-(\log n)^4]\ \mathit{and}\ (n-(\log n)^4,\ n]. \tag{2.196}$$

Within the first region, for any $k \in \mathbb{N}$, any pair (i, j) with length

$$2^k(\log n)^3 \leq j-i \leq 2^{k+1}(\log n)^3 \tag{2.197}$$

can be covered by the union of at most $2^{-k}n/\log n$ disjoint discrete squares of the form

$$\mathbb{T}_{2^k(\log n)^2}(x, x+j-i). \tag{2.198}$$

By (2.132),

$$1-(U_{(j)}-U_{(i)}) \geq 1-1.1(\log n)^4/n, \tag{2.199}$$

with probability tending to one. With these facts, by the union bound and Lemma 2.3.1.7,

$$\begin{aligned}
&\mathbb{P}\{\tilde{M}_n^+(l_n^+, (\log n)^4) \geq u_n\} \\
&\leq \mathbb{P}\left\{\max_{k:\log_2 A_2 \leq k \leq \log_2(\log n)} \tilde{M}_n^+(2^k(\log n)^3, 2^{k+1}(\log n)^3) \geq u_n\right\} \\
&\leq \mathbb{P}\left\{\max_{k:\log_2 A_2 \leq k \leq \log_2(\log n)} \tilde{Z}_n^+(2^k(\log n)^3, 2^{k+1}(\log n)^3) \geq u_n(\tau-0.1)\right\} \\
&\leq \sum_{k \geq \log_2 A_2} 2^{-k}\frac{n}{\log n}\mathbb{P}\left\{\max_{(i,j)\in T_{2^k(\log n)^2}(0,2^{k+1}(\log n)^3)} \tilde{Z}_{i,j}^+ \geq u_n(\tau-0.1)\right\} + \mathbb{P}(\Omega_n^c) \\
&\leq C\sum_{k \geq \log_2 A_2} 2^{-k} + \mathbb{P}(\Omega_n^c).
\end{aligned}$$

Taking $\limsup_{n\to\infty}$ and letting $A_2 \to \infty$ gives the desired result.

In the meantime, on $((\log n)^4, n-(\log n)^4]$, a finer examination of (2.131) and (2.133) yields

$$\left|\frac{n(U_{(j)}-U_{(i)})}{j-i}-1\right| = O_p\left(\frac{1}{\log n}\right). \tag{2.200}$$

(2.132) and (2.134) imply

$$\left|\frac{1-(U_{(j)}-U_{(i)})}{1-(j-i)/n}-1\right| = O_p\left(\frac{1}{\log n}\right). \tag{2.201}$$

Therefore,

$$\mathbb{P}\{\tilde{M}_n^+((\log n)^4, n-(\log n)^4) \geq u_n\} \leq \mathbb{P}\{M_n^+(l_n^+, (\log n)^4) \geq u_n(\tau-0.1)\} \to 0,$$

by Theorem 2.2.1.1.

The proof of the region $(n-(\log n)^4, n]$ is immediate by following the proof for (2.144), which we omit here. □

2.3.4 Proof of Theorem 2.2.2.2

Define

$$\tilde{Z}_{i,j}^- := \frac{S_j^- - S_i^-}{\sqrt{j-i+S_j^- - S_i^-}}, \tag{2.202}$$

and

$$g^-(a) := \frac{1}{2}(a\sqrt{a^2+4}+a^2). \tag{2.203}$$

$$I^-(g^-(s)) \geq s^2/2. \tag{2.204}$$

The theorem follows immediately after showing that

$$\limsup_{n\to\infty} \mathbb{P}(\tilde{M}_n^- \geq \varepsilon\sqrt{n}) = 0, \tag{2.205}$$

for any $\varepsilon > 0$. This can be proved similarly by dividing the regions, transforming the statistic $\tilde{M}_{i,j}^-$ into $\tilde{Z}_{i,j}^-$, combined with (2.204). We omit the detail here.

2.4 Acknowledgement

Chapter 2, in full, has been submitted for publication of the material as it may appear in Bernoulli. Ying, Andrew; Zhou, Wen-Xin. On the Asymptotic Distribution of the Scan Statistic for Point Clouds. The dissertation author was the primary investigator and author of this paper.

Chapter 3

Two-Stage Residual Inclusion for Survival Data and Competing Risks - An Instrumental Variable Approach with Application to SEER-Medicare Linked Data

3.1 Introduction

Interpreting the causal meaning of a treatment or exposure is straightforward under the randomized trials, because the randomization guarantees that there are no confounders for the exposure or treatment of interest. However, randomized experiment is not always feasible in practice. In observational studies issues such as endogeneity or potential confounding will arise. Instrumental variable (IV) is a useful method in some of these settings [AIR96], including when we may have unmeasured confounders. It finds increasing application in research on health care practices (see for example [HL04, SFW^{+}07]) including comparative effectiveness studies, and also in genetic studies where certain genes are used as IV for Mendelian randomization [GTTR12, LHS^{+}08, ZBL^{+}15].

Figure 3.1 illustrates a typical setting where IV methods can be applied. It is called a causal directed acyclic graph, where the nodes represent random variables, arrows represent direct causal effect, in such a way so that the common cause of any two nodes is included in the graph. In Figure 3.1 X_e is the exposure we are interested in testing the causal effect on the survival time outcome T, and we use a dashed line to represent the uncertainty of causation. In addition, X_o is the observed confounder, X_u the unobserved confounder. X_I is the instrument variable which has to satisfy the following three conditions: 1) X_I is associated with X_e, 2) X_I doesn't affect T except through its potential effect on X_e, and 3) X_I and X_u do not share causes.

There are two commonly used IV approaches: two-stage predictor substitution (2SPS), and two-stage residual inclusion (2SRI). For survival outcomes [TTWV^{+}15] considered these two methods under the additive hazards model, and gave conditions under which the causal parameters of interest can be correctly estimated. We note that while the Cox proportional hazards model has been more widely used in practice for survival data, an important appeal of additive hazards models is that unlike proportional hazards, a hazards difference is a collapsible effect measure.[TTWV^{+}15, MV13] This means that when there is an unobserved (exogenous) covariate for the survival outcome, integrating out this covariate under the additive hazards model still gives a model satisfying the additive hazards assumption. But this has long been known not to be the case for the proportional hazards model.[LN80, GWP84, SK86, BHC88, AF95, FNA95, XO00] In 2SPS under the additive hazards model the exposure is substituted by its prediction given the IV, and included in the model as a covariate, together with possibly other observed confounders. [LFB15] investigated its large sample properties under the additive hazards model. [ZDHZ17] considered 2SPS under competing risks.

The 2SRI is different from 2SPS in that instead of the predicted exposure, the residual from regressing the exposure on the IV(i.e. first stage) is included as an extra term in the additive hazards model, in addition to the exposure and possibly other observed confounders. This was considered more suitable for binary or discrete exposures [TBR08, TTWV$^+$15], when generalized linear models (GLM) for example are used in the first stage. Note that linear regression is used in the first stage of 2SPS, and therefore it is more suitable for continuous exposures. Very recently [JLF18] studied the 2SRI estimator with linear regression in the first stage for survival data. Our goal in this paper is to study the 2SRI estimator with GLM in the first stage, and also to develop the methodology for competing risks data.

This work was motivated by the desire to conduct comparative effectiveness research in large observational databases. In the field of oncology we lack gold standard randomized clinical trials in many clinical scenarios to optimally inform clinical decision making. With the lack of randomized trials investigators turn to comparative effectiveness with large observational data sets such as the linked (Surveillance, Epidemiology, and End Results) SEER-Medicare database. These databases include information on the specifics of cancer, staging, treatment, patient comorbidity, as well as information on long-term outcomes including toxicity and survival. Despite this wealth of information, these databases do not contain information on unmeasured confounding factors such as patient weight, smoking status, diet, exercise, patient compliance with treatment, and patient performance status. These unmeasured confounders can substantially influence outcomes (in particular survival), adding bias to comparative effectiveness research using observational data. Recently [HYB$^+$10] compared aggressive (radical prostatectomy) versus conservative treatments of prostate cancer with SEER-Medicare linked data. The question of radical prostatectomy versus conservative treatment has been addressed in randomized clinical trials which demonstrate no clear survival advantage for either treatment approach. [BAHR$^+$05, HDL$^+$16]Hadley found IV to be a useful technique as compared to for example propensity scores in adjustment for confounding in such data. Proportional hazards model was used in their analysis which, as pointed out by [LFB15] as well as explained above, due to the noncollapsibility is not suitable for the two-stage approaches. We would like to instead consider the additive hazards model for the reasons given earlier. Since the treatment choices are binary, we would like to use the 2SRI estimator, both for overall survival, and for cancer specific mortality. For these purposes we need to develop the inference procedure under the model for both general survival data and under competing risks.

The rest of the paper is organized as follows. In section 3.2 we describe the assumptions needed for the 2SRI approach under the additive hazards model for general survival data with right censoring, and we study the asymptotic behavior including consistency and asymptotic normality of the 2SRI estimator. Following that, we extend the results to competing risks data in Section 3.3 under subdistribution hazard modeling. For both settings we provide a closed-form variance estimate of the 2SRI estimator. Section 3.4 contains finite sample simulation results, and Section 3.5 the analysis of the SEER-Medicare data. Section 3.6 contains some further discussion. All technical details are provided in the Appendix.

3.2 Additive hazards model for survival data

In the presence of possible right censoring, let T and C be the failure time and the censoring time random variables, respectively. We can only observe $T^* = \min(T,C)$ and $\delta = \mathbb{1}\{T \leq C\}$. Similar to the setting in Figure 3.1, denote X_e as the exposure variable, whose causal effect is of primary interest, X_I as the IV, and X_o as the (vector of) observed confounders of dimension p. Our observed data for each individual is $\{T_i^*, \delta_i, X_{ei}, X_{oi}, X_{Ii}\}$ $(i = 1, ..., n)$, which we assume are independent and identically distributed. In this section, we will assume that T and C are independent conditional on X_e, X_I and X_o. Under the additive hazards model [Aal80, Aal89, LY94,], the hazard function of T given X_e, X_I, X_o and the unobserved confounders is assumed to be in the form

$$\lambda(t|X_e, X_I, X_o, X_u) = \lambda_0(t) + \beta_e X_e + \beta_o{}^\top X_o + X_u, \tag{3.1}$$

where X_u is a function of the unobserved confounders. We assume that X_u is independent of X_o and X_I. Denote

$$\Delta = X_e - \mathbb{E}(X_e|X_I, X_o). \tag{3.2}$$

Following [TTWV$^+$15] we put a key assumption on X_u:

$$X_u = \rho_0 \Delta + \varepsilon. \tag{3.3}$$

where ε is an error term independent of X_e, X_I and X_o. Proposition 3.7.2.1 in the Appendix shows that integrating out X_u we have

$$\lambda(t|X_e, X_I, X_o) = \bar{\lambda}_0(t) + \beta_e X_e + \beta_o{}^\top X_o + \rho_0 \Delta. \tag{3.4}$$

Note that the same coefficient β_e (and β_o) from (3.1) is remained in (3.4).

The error term Δ in (3.4) is not readily available from the data. Nonetheless we can 'estimate' Δ and use this estimate as a substitute. For this we need to impose an assumption on the form of $\mathbb{E}(X_e|X_I, X_o)$, for example,

$$g(\mathbb{E}(X_e|X_I, X_o)) = \alpha_c + \alpha_I X_I + \alpha_o{}^\top X_o, \tag{3.5}$$

where $g(\cdot)$ is a link function.

The two stage residual inclusion (2SRI) estimator is then defined as follows: in the first stage, we fit model (3.5) and obtain

$$\hat{\Delta} = X_e - \hat{X}_e = X_e - \hat{\mathbb{E}}(X_e | X_I, X_o). \tag{3.6}$$

Then in the second stage, we fit (3.4) with Δ replaced by $\hat{\Delta}$.

Denote $Z_i = [X_{ei}, X_{oi}^\top, \hat{\Delta}_i]^\top$ the regressors in (3.4) with Δ replaced by $\hat{\Delta}$, $\tilde{X}_i = [1, X_{Ii}, X_{oi}{}^\top]^\top$ the regressors in (3.5). Let $N_i(t) = \mathbb{1}\{T_i^* \leq t, \delta_i = 1\}$ be the counting process, and $Y_i(t) = \mathbb{1}\{T_i^* \geq t\}$ the at-risk process. Define the filtration $\mathcal{F}_t = \sigma\{N_i(u), Y_i(u), X_{Ii}, X_{oi}, X_{ei}, u \leq t, i = 1, .., n\}$. By the usual counting process theory, $M_i(t) = N_i(t) - \int_0^t Y_i(u)\lambda_i(u)du$ is a mean zero martingale with respect to the filtration $\mathcal{F}_t$. Under the additive hazards model the estimating equation for $\beta = (\beta_e, \beta_o^\top, \rho_0)^\top$ in (3.4) is

$$U(\beta) = \frac{1}{n}\sum_{i=1}^{n}\int_0^1 (Z_i - \bar{Z}(t))(dN_i(t) - Y_i(t)\beta^\top Z_i dt), \tag{3.7}$$

where $\bar{Z}(t) = \sum_{l=1}^{n} Z_l Y_l(t) / \sum_{l=1}^{n} Y_l(t)$. This gives our estimator

$$\hat{\beta} = \Big\{\sum_{i=1}^{n}\int_0^1 Y_i(t)(Z_i - \bar{Z}(t))^{\otimes 2} dt\Big\}^{-1}\Big\{\sum_{i=1}^{n}\int_0^1 (Z_i - \bar{Z}(t))dN_i(t)\Big\}. \tag{3.8}$$

In the following we show that $\hat{\beta}$ is consistent for the true β and therefore $\hat{\beta_e}$ is consistent for the causal parameter β_e. The estimator is also asymptotically normal and we provide a closed form expression for its asymptotic variance.

In addition to β, the cumulative baseline hazard function $\Lambda_0(t) = \int_0^t \bar{\lambda}_0(s)ds$ can be estimated by

$$\hat{\Lambda}_0(t) = \sum_{i=1}^{n}\int_0^t \frac{1}{\sum_{j=1}^{n} Y_j(u)} dN_i(u) - \hat{\beta}^\top \int_0^t \bar{Z}(u)du. \tag{3.9}$$

Using this we can also estimate the conditional survival function given the observed variables

$x = (x_e, x_o^\top, x_I)^\top$, the value of the variables of a future patient whose survival we are interested in predicting:

$$\hat{S}(t|x) = \exp(-\hat{\Lambda}_0(t) - \hat{\beta}^\top zt), \tag{3.10}$$

where $z = (x_e, x_o^\top, x_e - \hat{\mathbb{E}}(X_e|x_I, x_o))^\top$. Note that under the additive hazards model $\hat{\Lambda}_0(t)$ can be negative, or the estimated survival function $\hat{S}(t|x)$ not decreasing. Therefore we follow the approach of [LY94] and use a modified $\hat{\Lambda}_0^*(t) = \max_{0\le s\le t} \hat{\Lambda}_0(s)$, and $\hat{S}^*(t|x) = \min_{0\le s\le t} \hat{S}(s|x)$. Under regularity condition, the modified version is asymptotically equivalent to the original version. Now we state our main results below.

Theorem 3.2.0.1. *Under* (3.1)*,* (3.3)*,* (3.5) *and Condition 3.7.2.1, Condition 3.7.2.4, Condition 3.7.2.5 given in the Appendix, the two stage residual inclusion estimator* $\hat{\beta}$ *is consistent for the true value of* β *in* (3.4)*, denoted by* β_T*, i.e.* $\hat{\beta} \to \beta_T$ *in probability as* $n \to \infty$*.*

Theorem 3.2.0.2. *Under* (3.1)*,* (3.3)*,* (3.5) *and Condition 3.7.2.1, Condition 3.7.2.4, Condition 3.7.2.5,* $\sqrt{n}(\hat{\beta} - \beta_T)$ *is asymptotically normally distributed with asymptotic covariance matrix that can be consistently estimated by* $\hat{\Omega}^{-1}(\hat{\Sigma}_1 + \hat{\Sigma}_2)\hat{\Omega}^{-1}$*, where*

$$\begin{aligned}
\hat{\Omega} &= \frac{1}{n}\sum_{i=1}^{n}\int_0^1 Y_i(t)(Z_i - \bar{Z}(t))^{\otimes 2} dt, && (3.11)\\
\hat{\Sigma}_1 &= \frac{1}{n}\sum_{i=1}^{n}\int_0^1 (Z_i - \bar{Z}(t))^{\otimes 2} dN_i(t), && (3.12)\\
\hat{\Sigma}_2 &= \hat{\Psi}\hat{\Theta}\hat{\Psi}^\top, && (3.13)\\
\hat{\Psi} &= \frac{\hat{\rho}_0}{n}\sum_{i=1}^{n}\Big\{\int_0^1 Y_i(t)(Z_i - \bar{Z}(t))dt\Big\}\tilde{X}_i^\top (g^{-1})'(\tilde{X}_i^\top \hat{\alpha}), && (3.14)
\end{aligned}$$

$\alpha = (\alpha_c, \alpha_I, \alpha_o{}^\top)^\top$, $\hat{\Theta}$ *is the estimated covariance matrix of* $\sqrt{n}(\hat{\alpha} - \alpha_T)$ *from the first stage, and* $(g^{-1})'$ *is the derivative of the inverse function of* g*.*

Remark 3.2.0.1. Note that $\hat{\Theta}$ can typically be obtained when using software for fitting linear or generalized linear regression models in the first stage. For linear regression of X_e on X_I and X_o in (3.5), $g(y) = y$, so $(g^{-1})' \equiv 1$. Note this special case was investigated in [JLF18]. For logistic regression $(g^{-1})'(y) = \exp(y)/\{1+\exp(y)\}^2$.

Theorem 3.2.0.3. *For a new observation x, Under* (3.1)*,* (3.3)*,* (3.5) *and Condition 3.7.2.1, Condition 3.7.2.4, Condition 3.7.2.5, the estimated survival function in* (3.10) *converges to* $S(t|x)$ *uniformly and the process* $\sqrt{n}\{\hat{S}(\cdot|x) - S(\cdot|x)\}$ *converges weakly to a zero-mean Gaussian process whose covariance function at* (t,s)*, where* $0 \le s \le t$*, can be consistently estimated by*

$$\begin{aligned}&\hat{S}(t|x)\hat{S}(s|x)\Big\{n\sum_{i=1}^{n}\int_0^s \frac{1}{(\sum_{j=1}^{n} Y_j(u))^2} dN_i(u) + \hat{G}^\top(t)\hat{\Omega}^{-1}(\hat{\Sigma}_1 + \hat{\Sigma}_2)\hat{\Omega}^{-1}\hat{G}(s)\\ &+\hat{E}^\top(t)\hat{\Theta}\hat{E}(s) + \hat{G}^\top(t)\hat{\Omega}^{-1}\hat{D}(s) + \hat{G}^\top(s)\hat{\Omega}^{-1}\hat{D}(t)\Big\},\end{aligned} \tag{3.15}$$

where

$$\hat{D}(t) = \sum_{i=1}^{n}\int_0^t \frac{Z_i - \bar{Z}(u)}{\sum_{j=1}^{n} Y_j(u)} dN_i(u), \tag{3.16}$$

$$\hat{E}(t) = \hat{\rho}_0 \sum_{i=1}^{n} \tilde{X}_i (g^{-1})'(\tilde{X}_i^\top \hat{\alpha}) \int_0^t \frac{Y_i(u)}{\sum_{j=1}^{n} Y_j(u)} du, \tag{3.17}$$

$$\hat{G}(t) = \int_0^t (z - \bar{Z}(u)) du. \tag{3.18}$$

Remark 3.2.0.2. In forming the confidence interval (CI) for $S(t|x)$, we can take the log-log transformation of $\hat{S}(t|x)$ and use Delta method to obtain the confidence interval of $\log\Lambda(t|x)$. This way the transformed-back confidence interval of $\hat{S}(t|x)$ is guaranteed to be within the range of $[0,1]$.

The asymptotic results for the cumulative baseline hazard estimator is in the appendix.

3.3 Competing risks

We now consider competing risks data. As before let T and C be the failure time and the censoring time, respectively. In addition let $J \in \{1,...,K\}$ to be the indicator for cause of failure, and $J = 1$ will be our cause of interest. Denote $X = (X_e, X_I, X_o, X_u)$, $F_1(t|X) = P(T \leq t, J = 1|X)$, and let $\lambda_1(t|X) = -d\log\{1 - F_1(t|X)\}/dt$ be the subdistribution hazard. In principle we may assume that C and T are independent conditional on all the observed covariates, but for the estimation approach below we will make use of the marginal Kaplan-Meier estimate of the distribution of C. Therefore we will make the stronger assumption that C and T are independent; we will discuss the relaxation of this assumption later.

Similar to Section 3.2 we assume that

$$\lambda_1(t|X_e, X_I, X_o, X_u) = \lambda_{10}(t) + \beta_e X_e + {\beta_o}^\top X_o + X_u. \tag{3.19}$$

This is the additive subdistribution hazards model. Keeping the same notation as in (3.2) and assumption (3.3), we have according to Proposition 3.7.3.1 in the Appendix,

$$\lambda_1(t|X_e, X_I, X_o) = \bar{\lambda}_{10}(t) + \beta_e X_e + \bar{\beta}_o{}^\top X_o + \rho_0 \Delta. \tag{3.20}$$

Note that although the derivation of Proposition 3.7.3.1 is similar to that of Proposition 3.7.2.1 in the Appendix, this is a new result to our best knowledge and the 2SRI approach has not been previously considered under competing risks in the literature.

In the 2SRI approach for competing risks data here, the first stage is the same as that in Section 3.2, and we replace Δ by $\hat{\Delta}$ to fit (3.20). Our observed data are $\{T_i^*, \delta_i, \delta_i J_i, X_{ei}, X_{oi}, X_{Ii}\}_{1\leq i\leq n}$. The following are common quantities used in the regression modeling and inference of the subdistribution hazard function. With a slight abuse of notation in this section, define the event time process as $N_i(t) = \mathbb{1}\{T_i \leq t, \delta_i J_i = 1\}$, and the at-risk process as $Y_i(t) = 1 - N_i(t-)$ (note that these have different meanings from Section 3.2).

Let $r_i(t) = \mathbb{1}\{C_i \geq T_i \wedge t\}$ denote an individual not yet censored, so that both $r_i(t)N_i(t)$ and $r_i(t)Y_i(t)$ are computable from the observed data at any time t. In particular,

$$r_i(t)N_i(t) = \mathbb{1}\{T_i^* \leq t, \delta_i J_i = 1\}, \tag{3.21}$$

$$r_i(t)Y_i(t) = \mathbb{1}\{T_i^* \geq t\} + \mathbb{1}\{T_i^* < t, \delta_i = 1, \delta_i J_i \neq 1\}. \tag{3.22}$$

Define

$$w_i(t) = r_i(t)G(t)/G(T_i^* \wedge t), \tag{3.23}$$

and

$$\hat{w}_i(t) = r_i(t)\hat{G}(t)/\hat{G}(T_i^* \wedge t), \tag{3.24}$$

where $G(t) = \mathbb{P}(C \geq t)$ and $\hat{G}(t)$ is the Kaplan-Meier estimate for $G(t)$ using $\{T_i^*, 1-\delta_i\}_{1 \leq i \leq n}$. The $\hat{w}_i(t)$'s are the weights that will be used in the estimating equation below.

Multiple filtrations and martingales are needed under the subdistribution hazard modeling of competing risks. We will use $M_i^1(t)$ to denote the martingale for the i-th object with respect to the complete-data filtration, that is, $\mathcal{F}^1(t) = \sigma\{N_i(u), Y_i(u), X_{ei}, X_{Ii}, X_{oi}, u \leq t, \forall\ 1 \leq i \leq n\}$. We will also use $M_i^c(t)$ to denote the martingale for the censoring related process of the i-th subject, $M_i^c(t) = N_i^c(t) - \int_0^t \mathbb{1}\{T_i^* \geq u\} d\Lambda^c(u)$, where $N_i^c(t) = \mathbb{1}\{T_i^* \leq t, \delta_i = 0\}$ is the censoring counting process, $\Lambda^c(t)$ is the cumulative hazard function of the censoring distribution. The censoring filtration is $\mathcal{F}^c(t) = \sigma\{\mathbb{1}\{T_i^* \geq u\}, \mathbb{1}\{T_i^* \leq u, \delta_i = 0\}, X_{ei}, X_{Ii}, X_{oi}, u \leq t, \forall\ 1 \leq i \leq n\}$.

The estimating function for $\beta = (\beta_e, \beta_o^\top, \rho_0)^\top$ can be written as [ZDHZ17, LXL17]

$$U(\beta) = \frac{1}{n}\sum_{i=1}^{n}\int_0^1 (Z_i - \bar{Z}(t))\hat{w}_i(t)(dN_i(t) - Y_i(t)\beta^\top Z_i dt), \tag{3.25}$$

where Z_i is the same as defined in Section 3.2, $\bar{Z}(t) = \frac{\sum_{l=1}^{n} Z_l \hat{w}_l(t) Y_l(t)}{\sum_{l=1}^{n} \hat{w}_l(t) Y_l(t)}$. Therefore,

$$\hat{\beta} = \Big\{\sum_{i=1}^{n}\int_0^1 \hat{w}_i(t)Y_i(t)(Z_i - \bar{Z}(t))^{\otimes 2} dt\Big\}^{-1}\Big\{\sum_{i=1}^{n}\int_0^1 (Z_i - \bar{Z}(t))\hat{w}_i(t) dN_i(t)\Big\}. \tag{3.26}$$

The baseline cumulative hazard function $\Lambda_{10} = \int_0^{\cdot} \lambda_{10}$ is then estimated by

$$\hat{\Lambda}_{10}(t) = \sum_{i=1}^{n}\int_0^t \frac{\hat{w}_i(u)}{\sum_{j=1}^{n} \hat{w}_j(u) Y_j(u)} dN_i(u) - \hat{\beta}^\top \int_0^t \bar{Z}(u) du. \tag{3.27}$$

Therefore the estimated cumulative incidence function (CIF) is

$$\hat{F}_1(t|x) = 1 - \hat{S}_1(t|x) = 1 - \exp(-\hat{\Lambda}_{10}(t) - \hat{\beta}^\top z t), \tag{3.28}$$

where $z = (x_e, x_o^\top, x_e - \hat{\mathbb{E}}(x_e|x_I, x_o))^\top$, and (x_e, x_o, x_I) is the value of the variables of a future patient whose CIF we are interested in predicting. We use the same modified version $\hat{\Lambda}_{10}^*(t) = \max_{0\le s\le t} \hat{\Lambda}_{10}(s)$ and $\hat{F}_1^*(t|x) = \max_{0\le s\le t} \hat{F}_1(s|x)$ as in Section 3.2 to ensure that the estimated hazard is non-negative. Now we state our main results below.

Theorem 3.3.0.1. *Under* (3.19), (3.3), (3.5) *and Condition 3.7.2.1, Condition 3.7.3.1, Condition 3.7.3.2, the two stage residual inclusion estimator $\hat{\beta}$ is consistent for the true value of β in* (3.4), *denoted by β_T, i.e. $\hat{\beta} \to \beta_T$ in probability as $n \to \infty$.*

Theorem 3.3.0.2. *Under* (3.19), (3.3), (3.5) *and Condition 3.7.2.1, Condition 3.7.3.1, Condition 3.7.3.2, $\sqrt{n}(\hat{\beta} - \beta_T)$ is asymptotically normally distributed with asymptotic covariance matrix* *that can be consistently estimated by* $\hat{\Omega}^{-1}(\hat{\Sigma}_1+\hat{\Sigma}_2+\hat{\Sigma}_3)\hat{\Omega}^{-1}$*, where*

$$
\begin{aligned}
\hat{\Omega} &= \frac{1}{n}\sum_{i=1}^{n}\int_0^1 \hat{w}_i(t)Y_i(t)(Z_i-\bar{Z}(t))^{\otimes 2}dt, &(3.29)\\
\hat{\Sigma}_1 &= \frac{1}{n}\sum_{i=1}^{n}\int_0^1 (Z_i-\bar{Z}(t))^{\otimes 2}\hat{w}_i(t)dN_i(t), &(3.30)\\
\hat{\Sigma}_2 &= \hat{\Psi}\hat{\Theta}\hat{\Psi}^{\top}, &(3.31)\\
\hat{\Sigma}_3 &= \frac{1}{n}\sum_{i=1}^{n}\int_0^1 \Big(\frac{\hat{q}(t)}{\hat{\pi}(t)}\Big)^{\otimes 2}dN_i^c(t), &(3.32)\\
\hat{\Psi} &= \frac{\hat{\rho}_0}{n}\sum_{i=1}^{n}\Big\{\int_0^1 \hat{w}_i(t)Y_i(t)(Z_i-\bar{Z}(t))dt\Big\}\tilde{X}_i^{\top}(g^{-1})'(\tilde{X}_i^{\top}\hat{\alpha}), &(3.33)\\
\hat{q}(t) &= -\frac{1}{n}\sum_{i=1}^{n}\int_0^1 \mathbb{1}\{T_i^* < t \le u\}\hat{w}_i(u)\big(Z_i-\bar{Z}(u)\big)d\hat{M}_i(u), &(3.34)\\
\hat{\pi}(t) &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}\{T_i^* \ge t\}, &(3.35)\\
\hat{M}_i(t) &= N_i(t)-\int_0^t Y_i(u)(d\hat{\Lambda}_{10}(u)+\hat{\beta}^{\top}Z_i du), &(3.36)
\end{aligned}
$$

and $\hat{\Theta}$ *is the estimated variance-covariance matrix of* $\sqrt{n}(\hat{\alpha}-\alpha_T)$ *from the first stage.*

Theorem 3.3.0.3. *For a new observation x, under* (3.19), (3.3), (3.5) *and Condition 3.7.2.1, Condition 3.7.3.1, Condition 3.7.3.2, the estimated CIF in* (3.28) *converges to* $F_1(t|x)$ *uniformly and the process* $\sqrt{n}\{\hat{F}_1(\cdot|x)-F_1(\cdot|x)\}$ *converges weakly to a zero-mean Gaussian process whose covariance function at* (t,s)*, where* $0\le s\le t$*, can be consistently estimated by*

$$
\begin{aligned}
&(1-\hat{F}_1(t|x))(1-\hat{F}_1(s|x))\Big\{\int_0^s \frac{n\sum_{i=1}^n \hat{w}_i^2(u)dN_i(u)}{(\sum_{j=1}^n \hat{w}_j(u)Y_j(u))^2}+\hat{G}^{\top}(t)\hat{\Omega}^{-1}(\hat{\Sigma}_1+\hat{\Sigma}_2+\hat{\Sigma}_3)\hat{\Omega}^{-1}\hat{G}(s)\\
&+n\sum_{i=1}^{n}\int_0^1 \frac{\hat{q}_t(u)\hat{q}_s(u)}{\hat{\pi}^2(u)}dN_i^c(u)+\hat{E}^{\top}(t)\hat{\Theta}\hat{E}(s)+\hat{G}^{\top}(t)\hat{\Omega}^{-1}\hat{D}(s)+\hat{G}^{\top}(s)\hat{\Omega}^{-1}\hat{D}(t)\Big\}, \quad (3.37)
\end{aligned}
$$

where

$$
\begin{aligned}
\hat{D}(t) &= \sum_{i=1}^{n}\int_0^t \frac{(Z_i-\bar{Z}(u))\hat{w}_i(u)}{\sum_{j=1}^{n}\hat{w}_j(u)Y_j(u)}dN_i(u), & (3.38)\\
\hat{E}(t) &= \hat{\rho}_0\sum_{i=1}^{n}\tilde{X}_i(g^{-1})'(\tilde{X}_i^\top\hat{\alpha})\int_0^t \frac{\hat{w}_i(u)Y_i(u)}{\sum_{j=1}^{n}\hat{w}_j(u)Y_j(u)}du, & (3.39)\\
\hat{G}(t) &= \int_0^t (z-\bar{Z}(u))du, & (3.40)\\
\hat{q}_t(u) &= \frac{1}{n}\sum_{i=1}^{n}\int_0^t \frac{\mathbb{1}\{T_i^* < u \le v\}\hat{w}_i(v)}{\sum_{j=1}^{n}\hat{w}_j(v)Y_j(v)}d\hat{M}_i(v). & (3.41)
\end{aligned}
$$

Remark 3.3.0.1. As in Section 2 we can take the log-log transformation of $1-\hat{F}_1(t|x)$, and use Delta method to obtain the confidence interval of $\log\Lambda_1(t|x)$. This way the transformed-back confidence interval of $1-\hat{F}_1(t|x)$ is guaranteed to be within the range of $[0,1]$.

The asymptotic results for the cumulative baseline hazard estimator is in the appendix.

3.4 Simulation

We are in the process of completing an R package for our estimators. The following numerical results were obtained using the program which is the core of the package.

To study the performance of our estimators under both survival and competing risks settings, we carried out simulation studies with sample size $100, 200, 400, 800, 1200$, and repeated 1000 times for each sample size. We provided in the tables the bias of the estimator, the empirical variance of the estimator from the 1000 repeats, the mean of the variance estimate, and the coverage rate of the nominal 95% confidence intervals.

3.4.1 Regular survival model

For this part without competing risks, we considered the following three scenarios.

Scenario I: We sampled X_I, X_o from independent standard normal distributions, set $\alpha =$

$[1,1,0.5]^\top$, and generated $X_e = [1, X_I, X_o]\alpha^\top + \Delta$, where $\Delta \sim N(0, 0.2)$ was independent of X_I, X_o. We simulated X_u according to (3.3) with independent $\varepsilon_i \sim N(0, 0.1)$, and $\rho_0 = 1$. We set $\beta = [1, 0.5, 1.5]^\top$ and the baseline hazard $\lambda_0(t) \equiv 10.5$ in (3.1) to generate the survival time T. There was no censoring in this case.

Scenario II: Similar to I above, but with $\alpha = [0.25, 0.3, 0.2]^\top$, $\beta = [0.5, 0.2, 0.3]^\top$, and $\lambda_0(t) \equiv 5t + 5$. The censoring distribution followed exponential with rate of 2 so that the censoring rate is around 40%.

Scenario III: We generated binary exposure as follows. First we sampled X_I from Bernoulli distribution with $P(X_I = 1) = 0.5$, and independent X_o from standard normal distribution. We generated X_e from

$$X_{ei}|X_{Ii}, X_{oi} \sim \text{Bern}\Big(\frac{1}{1+\exp(-(\alpha_0 + \alpha_i X_{Ii} + \alpha_o X_{oi}))}\Big).$$

with $\alpha = [1, 0.5, 1]^\top$. We then simulated

$$X_{ui} = \rho_0\{X_{ei} - E(X_{ei}|X_{Ii}, X_{oi})\} + \varepsilon_i,$$

with $\rho_0 = 1$ and independent $\varepsilon_i \sim N(0, 0.1)$. Finally we generate the survival time by setting $\beta = [1, 0.5, 1.5]^\top$ and the baseline hazard $\lambda_0(t) \equiv 10.5$. The censoring distribution was exponential with rate of 5 so that the censoring rate is around 30%.

3.4.2 Competing risks model

For the competing risks model, let

$$\mathbb{P}(J = 1|X) = 1 - \exp\Big(-\int_0^{t_0} \lambda_{10}(u)du - (\beta_{e1}X_e + \beta_{o1}^\top X_o + \beta_{u1}X_u)t_0\Big), \tag{3.42}$$

and $\mathbb{P}(J=2|X) = 1-\mathbb{P}(J=1|X)$. Here t_0 is a maximum follow-up time. We then simulated the event time data from the following subdistributions:

$$F(t|J=1,X) = \frac{1-\exp(-\int_0^{\min(t,t_0)} \lambda_{10}(u)du - (\beta_{e1}X_e + \beta_{o1}^\top X_o + \beta_{u1}X_u)\min(t,t_0))}{1-\exp(-\int_0^{t_0} \lambda_{10}(u)du - (\beta_{e1}X_e + \beta_{o1}^\top X_o + \beta_{u1}X_u)t_0)}, \quad (3.43)$$

$$F(t|J=2,X) = \frac{1-\exp(-(\lambda_{20} + \beta_{e2}X_e + \beta_{o2}^\top X_o + \beta_{u2}X_u)\min(t,t_0))}{1-\exp(-(\lambda_{20} + \beta_{e2}X_e + \beta_{o2}^\top X_o + \beta_{u2}X_u)t_0)}. \quad (3.44)$$

Note that

$$F(t,J=1|X) = 1-\exp(-\int_0^{\min(t,t_0)} \lambda_{10}(u)du - (\beta_{e1}X_e + \beta_{o1}^\top X_o + \beta_{u1}X_u)\min(t,t_0)). \quad (3.45)$$

Note that the maximum follow-up time t_0 is necessary because under the additive hazards model $\Lambda_1(t|\bar{Z}) = \Lambda_{10}(t) + \beta^\top \bar{Z}t$ goes to infinity as $t \to \infty$, implying that $\lim_{t\to\infty} F_1(t|\bar{Z}) = 1$, which is no longer a subdistribution function.

Scenario I: Similar to Scenario I under the regular survival model but with the parameters $\alpha = [1.5, 1, 0.7]^\top$, $\beta_1 = [1, 0.5, 0.75]^\top$ and $\lambda_{10}(t) \equiv 11$ for cause 1, and $\beta_2 = [1.2, 1, 1.3]^\top$ and $\lambda_{20} = 15$ for cause 2. We set $t_0 = 0.095$. There was no censoring, and the cause 1 event rate was around 60%.

Scenario II: similar to I above, but with $\alpha = [1, 1, 0.5]^\top$, $\beta_1 = [1, 0.5, 0.75]^\top$ and $\lambda_{10}(t) = 5t + 10$ for cause 1, and the same as in Scenario I for cause 2. We set $t_0 = 0.06$. The censoring distribution was exponential with rate of 1. The cause 1 event rate was around 28% and censoring rate around 38%.

Scenario III: Similar to Scenario III under the regular survival model, but with $\alpha = [-1, 2, 1]^\top$, $\beta_1 = [1, 0.5, 0.75]^\top$ and $\lambda_{10}(t) \equiv 10$ for cause 1, and the same as in Scenario I, II for cause 2. We set $t_0 = 0.06$. The censoring distribution was exponential with rate of 25. The cause 1 event rate was around 26% and censoring rate aroung 44%.

3.4.3 Simulation results

The results are summarized in Table 3.1 and 3.2, respectively. From the tables we can see that as the sample size increases the bias goes down, and the variance estimate also becomes closer to the empirical variance. The coverage rate is quite close to nominal 95% level in all cases. The variance of the estimator becomes larger in Scenario III both with or without competing risks, probably because we have a binary treatment, which leads to a error term with large variance, compared to the error terms in other scenarios. Nonetheless the variance estimate still works well. In general our estimator behaves well under all the finite-sample settings considered.

3.5 SEER-Medicare data analysis

For this analysis we consider prostate cancer patients with localized non-metastatic disease identified from the linked SEER-Medicare database diagnosed between 2000-2011 and followed up through 12/31/2013. The variables included were age, race/ethnicity, marital status, tumor stage, tumor grade, Prior Charlson comorbidity score measured during the year prior to diagnosis, year of diagnosis, and hospital referral regions. The hospital referral region was an important variable for us to construct the instrumental variable. Hospital referral regions represents a set of contiguous zip codes around a major hospital. Following [HYB$^+$10] we restricted the analysis to early stage (T1 and T2) patients, aged 66 to 74 years, as well as eliminated patients in geographic areas with fewer than 50 patients over the entire observation period. This led to an overall sample size of $n = 29806$. Among them 493 (1.65%) patients died due to cancer, 2066 (6.93%) died due to other causes, and the remaining 27247 (91.4%) were alive at the end of the follow-up. There were four types of treatments: surgery, radiation, chemotherapy and hormonal therapy; 10977 people received surgery, 21357 radiation, 9577 chemotherapy, and 9527 hormonal therapy. Note that some patients received more than one treatment. Following [HYB$^+$10] we will label patients who received surgery as “radical prostatectomy” and the remaining “conservative management”.

We will then compare the effects of these two treatments on the time to death for all causes and due to cancer, respectively. A summary of the patient characteristics is presented in Table 3.3. It can be seen that patients who received surgery tended to be younger, married, non-black, have T2 stage, well differentiated tumor grade, comorbidity score 0, and diagnosed no later than 2005. [HYB$^+$10] showed that the treatment pattern varied by hospital referral regions, beyond what was captured by the patient characteristics in Table 3.3.

For comparison purposes we first fitted the additive hazards model including the variables in Table 3.3 and all their pairwise interactions, but without using any IV. It turned out that the treatment had a significant effect with p-value of 0.001 for all causes of death. On the other hand the treatment effect was not significantly different from zero for our data with p-value of 0.17 for cancer specific survival. We note that censoring in this data set was administrative only, i.e. at the time of data export, and the only covariate that was correlated with censoring was year of diagnosis. In fitting the subdistribution hazards model we let the weights be conditional on this categorical variable. In [HYB$^+$10] the treatment had a significant effect on both overall survival and cancer specific survival; our data was a later export than those used by [HYB$^+$10] (diagnosed between 1995 and 2003) from the linked database. In addition, [HYB$^+$10] used the Cox multiplicative hazards model as opposed to our additive hazards model.

We now consider the 2SRI approach. We used the same instrumental variable as in [HYB$^+$10]. Specifically, we constructed the IV as follows. We first applied logistic regression to obtain the predicted probability for conservative management given covariates including age, race/ethnicity, marital status, tumor stage, tumor grade description, Prior Charlson comorbidity score, year of diagnosis and all the two-way interactions. Then for each hospital referral region and each year, we calculated the difference between the proportion of patients receiving conservative management and the average predicted probability of conservative management. Clearly, a larger difference indicated that the corresponding hospital referral region favored the conservative management more than those with a smaller difference. Therefore this difference was likely correlated with the treatment a patient received and, on the other hand, this difference was unlikely to directly influence the survival of an individual patient beyond the treatment assignment. For use as an IV we lagged this difference for one year for the patients coming from the same hospital referral region. Therefore, the data that we used to analysis survival were patients diagnosed from 2001 to 2011.

We then performed the first step of the IV analysis, using logistic regression of treatment on the IV obtained above, together with the other observed confounders including age, race or ethnicity, marital status, tumor stage, grade description and Prior Charlson comorbidity score, year of diagnosis and all two-way interactions. We then subtracted the predicted probability of treatment from the observed treatment to obtained the residuals. In the second step, we included this residual term together with the treatment and all the confounders and their pairwise interactions to fit the survival models. The results for overall survival are shown in Table 3.4, and for cancer specific survival in Table 3.5.

From the tables we see that the causal effect of treatment remained significant for overall survival, although the p-value increased from 0.001 to 0.042. The p-value for the causal effect of treatment on cancer specific survival also increased from 0.17 to 0.83. The differences between the IV analysis results and the initial analysis results earlier indicate that there were likely unobserved confounders for the treatment effect on both overall and cancer specific survival, beyond those captured in Table 3.3 (and their interactions). At the request of a reviewer, we also compared our results with the 2SRI approach of [JLF18], where linear regression was used in the first stage. For this data set the results were similar: for overall survival the treatment effect was estimated to be -0.0013 with a p-value of 0.021, and for cancer specific survival the treatment effect was 4.3×10^{-5} with a p-value of 0.86.

Finally, Figure 3.2 illustrates the predicted overall survival as well as cancer specific cumulative incidence function for a patient who received radical prostatectomy, was diagnosed in 2001, aged 71, white, with ‘other’ marital status, tumor stage T2, moderately differentiated tumor, Charlson comorbidity score 2, and instrument value 0.0.3429. In reality the patient survived for 95 months, and died of other causes.

3.6 Discussion

In this paper we have developed statistical inference procedures for the 2SRI IV estimator under GLM in the first stage and additive hazards model in the second stage for survival data that was conceptually described in [TTWV$^+$15]. As mentioned earlier 2SRI was considered more suitable for binary or discrete exposures than 2SPS, as GLM may be used in the first stage to model the exposure. On the other hand, assumption such as (3.3) is needed for the 2SRI to work, although when allowing for general covariates a very strong linearity condition was imposed on certain function in the proof of [TTWV$^+$15] Result 1 for 2SPS. We have also extended the approach to competing risks data under the additive subdistribution hazards model. More practical experience is needed to compare 2SRI and 2SPS for complex outcomes such as survival with competing risks, etc. An R package is being completed that computes these estimators and their closed-form estimated asymptotic variances, as well as prediction under these models given the observed covariates. Our simulation results show the satisfactory performance of the procedures, and the SEER-Medicare analysis shows the usefulness of the approaches.

The causal effect of interest β_e that we have considered in this work is conditional on the unobserved confounder X_u, although from (3.4) and (3.20) we may also understand it as conditional on the observed variables. This seems reasonable in our comparative effectiveness settings, where a relatively large number of observed confounders are typically considered. [CSH11] considered the setting for compliance in randomized clinical trials with binary outcomes, and pointed out that the 2SPS and 2SRI approaches may not estimate the causal odds ratio among compliers under the principal stratification framework.[AIR96, FR02a] Future work may consider similar analysis under the additive hazards model.

Under competing risks and using subdistribution hazards modeling, weights based on the estimated censoring distribution are needed in order to consistently estimate the regression coefficients. Using the marginal Kaplan-Meier estimate for the weights requires independent censoring. Alternatively one may estimate the conditional distribution of censoring given covariates using, in our case, Kaplan-Meier estimate give each category of year of diagnosis. For continuous covariates semiparametric survival models have been used in the literature. Recently [NG17] proposed to estimate this conditional distribution nonparametrically using a survival tree approach. When any of these estimates are used, the assumption on censoring distribution can then be relaxed to be conditionally independent of failure time (and type/cause) given the covariates.

All our technical proofs are compatible with time-dependent covariates. However, the causal inference problem is more complex with time-varying confounders and time-varying treatments, especially if the later confounders are affected by the earlier treatments.[HBR01] To our best knowledge time-varying instrument variable method has not been developed in the literature.

Acknowledgement

The authors would like to thank Mr. Vinit Nalawade for help with preparing the SEER-Medicare data set.

3.7 Appendix

3.7.1 Preparations

The following results are either from or easy consequences of [FH11]. The predictable variation process $\langle\rangle$ and quadratic variation process $[]$ are also defined in [FH11].

Lemma 3.7.1.1. *Assume that for each $t \geq 0$, given $\mathcal{F}_{t-}$, $\{dN_1(t), ..., dN_n(t)\}$ are independent $0, 1$ random variables, set $M_j = N_j - A_j$, where A_j is the compensator for N_j. Then for any $i \neq j$ and $t \geq 0$,*

$$\langle M_i, M_j \rangle(t) = 0, \quad a.s. \tag{3.46}$$

Lemma 3.7.1.2. *Let $M(t)$ be a martingale with respect to the filtration $\mathcal{F}(t)$, $H(t)$ be a predictable process, then the predictable variation process of the martingale integral $\int_0^t H(s)dM(s)$ is*

$$\langle \int_0^t H(s)dM(s) \rangle = \int_0^t H^2(s)d\langle M \rangle(s). \tag{3.47}$$

Lemma 3.7.1.3. *Let $M_1(t)$, $M_2(t)$ be martingales with respect to the filtration $\mathcal{F}(t)$, $H_1(t)$, $H_2(t)$ be predictable processes, then the predictable covariation process of the martingale integral $\int_0^t H_1(s)dM_1(s)$ and $\int_0^t H_2(s)dM_2(s)$ is*

$$\langle \int_0^t H_1(s)dM_1(s), \int_0^t H_2(s)dM_2(s) \rangle = \int_0^t H_1(s)H_2(s)d\langle M_1, M_2 \rangle(s). \tag{3.48}$$

Lemma 3.7.1.4. *For independent and identically distributed sequences of martingale $\{M_i\}$ and predictable process $H_i(t)$, $1 \leq i \leq n$,*

$$\mathrm{Var}\Big(\int_0^t H_1(u)dM_1(u)\Big) = \mathbb{E}\Big(\langle \int_0^t H_1(u)dM_1(u) \rangle\Big) = \mathbb{E}\Big(\int_0^t H_1^2(u)d\langle M_1 \rangle(u)\Big), \tag{3.49}$$

can be estimated by

$$\frac{1}{n}\sum_{i=1}^{n}\int_0^t H_i^2(u)d[M_i](u). \tag{3.50}$$

Lemma 3.7.1.5 (Lenglart's inequality)**.** *Let W be a local square integrable martingale. Then for all* $\delta, \eta > 0$*,*

$$\mathbb{P}(\sup_{t\in[0,1]}|W(t)| > \eta) \leq \frac{\delta}{\eta^2} + \mathbb{P}(\langle W, W\rangle(1) > \delta). \tag{3.51}$$

Basically the Lenglart's inequality tells us that the convergence in probability of the supreme of a martingale can be infered from its endpoint.

Lemma 3.7.1.6. *Let* $M_i(t)$ *be independent and identically distributed martingales with respect to the filtration* $\mathcal{F}(t)$*,* $i = 1, ..., n$*,* $H_i(t)$ *also be i.i.d. predictable processes, then*

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\int_0^t H_i(s)dM_i(s) \to_p 0 \quad \textit{uniformly in } t \in [0,1], \tag{3.52}$$

if each $\sup_{0\leq t\leq 1}|H_i(t)| = o_p(1)$*.*

Proof. By Lemma 3.7.1.1 and 3.7.1.3,

$$\begin{aligned}\langle\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\int_0^t H_i(s)dM_i(s)\rangle &= \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\int_0^t H_i(s)H_j(s)d\langle M_i, M_j\rangle(s) \\ &= \frac{1}{n}\sum_{i=1}^{n}\int_0^t H_i^2(s)d\langle M_i\rangle(s).\end{aligned}$$

Therefore,

$$\mathbb{P}\Big(\sup_{t\in[0,1]}|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\int_0^t H_i(s)dM_i(s)| > \eta\Big) \leq \frac{\delta}{\eta^2} + \mathbb{P}\Big(\frac{1}{n}\sum_{i=1}^{n}\int_0^1 H_i^2(s)d\langle M_i\rangle(s) > \delta\Big).$$

Since $\langle M_i \rangle$ is increasing,

$$\frac{1}{n}\sum_{i=1}^{n}\int_0^1 H_i^2(s)d\langle M_i\rangle(s) \le \frac{1}{n}\sum_{i=1}^{n}[\sup_{0\le s\le t} H_i^2(s)]\langle M_i\rangle(t), \tag{3.53}$$

if $\sup_{0\le t\le 1}|H_i(t)| = o_p(1)$, so is $\frac{1}{n}\sum_{i=1}^{n}[\sup_{0\le s\le t} H_i^2(s)]\langle M_i\rangle(t)$ and $\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\int_0^t H_i(s)dM_i(s)$, the latter by Lemma 3.7.1.5. □

The following proposition can be easily proved by mimicking the proof of Glivenko-Cantelli theorem.

Proposition 3.7.1.1. *Assume that $(X_1,Y_1),\ldots,(X_n,Y_n)$ are independent and identically distributed pairs of random variables with distribution function $F(x,y)$. Also, X_1 has finite first moment. Define*

$$G_n(t) = \frac{1}{n}\sum_{i=1}^{n} X_i \mathbb{1}\{[Y_i,\infty)\}(t), \tag{3.54}$$

and

$$G(t) = \mathbb{E}[X_1\mathbb{1}\{[Y_1,\infty)\}], \tag{3.55}$$

then we have a similar result to Glivenko-Cantelli theorem, that is,

$$||G_n - G||_\infty = \sup_{t\in\mathbb{R}}|G_n(t) - G(t)| \longrightarrow 0. \tag{3.56}$$

3.7.2 Additive hazards model for survival data

Proposition 3.7.2.1. *Assuming* (3.1) *and* (3.3)*, we have*

$$\bar{\lambda}(t|X_e,X_I,X_o) = \bar{\lambda}_0(t) + \beta_e X_e + \bar{\beta_o}^\top X_o + \rho_0\Delta, \tag{3.57}$$

where $\bar{\lambda}_0(t) = \lambda_0(t) - \frac{\partial}{\partial t}\log\Big[\mathbb{E}\big\{\exp(-\varepsilon t)\big\}\Big]$.

Proof. By the assumption (3.1), we have

$$S(t|X_e,X_I,X_o,X_u) = \exp\Big\{-\int_0^t \big[\lambda_0(s)+\beta_e X_e+{\beta_o}^\top X_o+X_u\big]ds\Big\}.$$

Therefore, integrating out X_u we have,

$$\begin{aligned}
& S(t|X_e,X_I,X_o) \\
= \;& \mathbb{E}[S(t|X_e,X_I,X_o,X_u)|X_e,X_I,X_o] \\
= \;& \mathbb{E}\big[\exp\big\{-\int_0^t(\lambda_0(s)+\beta_e X_e+{\beta_o}^\top X_o+\mathbb{E}(X_u|X_e,X_I,X_o))\big\}ds|X_e,X_I,X_o\big] \\
= \;& \mathbb{E}\big[\exp\big\{-\int_0^t(\lambda_0(s)+\beta_e X_e+{\beta_o}^\top X_o+\rho_0\Delta+\varepsilon)ds\big\}|X_e,X_I,X_o\big] \\
= \;& \exp\big\{-\int_0^t(\lambda_0(s)+\beta_e X_e+{\beta_o}^\top X_o+\rho_0\Delta\big\}\times\mathbb{E}\big[\exp\{-\varepsilon t\}|X_e,X_I,X_o\big] \\
= \;& \exp\big\{-\int_0^t(\lambda_0(s)+\beta_e X_e+{\beta_o}^\top X_o+\rho_0\Delta)ds\big\}\times\mathbb{E}\big[\exp\{-\varepsilon t\}\big].
\end{aligned}$$

Now the hazard function becomes,

$$\begin{aligned}
\lambda(t|X_e,X_I,X_o) &= -\frac{\partial}{\partial t}\log S(t|X_e,X_I,X_o) \\
&= \lambda_0(t)-\frac{\partial}{\partial t}\log\Big[\mathbb{E}\big\{\exp(-\varepsilon t)\big\}\Big]+\beta_e X_e+{\beta_o}^\top X_o+\rho_0\Delta,
\end{aligned}$$

where $\bar{\lambda}(t|X_e,X_I,X_o) = \bar{\lambda}_0(t)+\beta_e X_e+\bar{\beta_o}^\top X_o+\rho_0\Delta$ is our new baseline hazard function. □

In the following we give the regularity conditions and addition notation for the results of the Theorems in section 3.2.

Condition 3.7.2.1. *The covariates* $\{X_{ei},X_{Ii},X_{oi}\}$ *are bounded.*

Note that

$$\begin{aligned}\hat{\Delta}-\Delta &= -\{\hat{\mathbb{E}}(X_e|X_I,X_o)-\mathbb{E}(X_e|X_I,X_o)\} \\ &= -\{g^{-1}(\tilde{X}^{\top}\hat{\alpha})-g^{-1}(\tilde{X}^{\top}\alpha_T)\}\end{aligned}$$

Condition 3.7.2.2. *We assume that $\hat{\alpha}$ is consistent for α_T, which implies that $|\hat{\Delta}-\Delta| \to_p 0$ with Condition 3.7.2.1. We also assume that*

$$\sqrt{n}(\hat{\alpha}-\alpha_T) = I^{-1}\cdot(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}U_i)+o_p(1);$$

in other words, $\sqrt{n}(\hat{\alpha}-\alpha_T)$ can be written into a sum of i.i.d. terms with finite variance plus one $o_p(1)$ term.

Remark 3.7.2.1. This Condition is usually fulfilled by the first step estimator, such as the maximum likelihood estimator under the GLM.

Condition 3.7.2.3.

$$\int_0^1 \bar{\lambda}_0(t)dt < \infty. \tag{3.58}$$

Define $S^{(j)}(t)=\sum_{i=1}^{n}Y_i(t)Z_i^{\otimes j}/n$ for $j=0$ and 1. Then by Gilvenko-Cantelli theorem there exists a scalar and vector function $s^{(0)}(t)$ and $s^{(1)}(t)$ defined on $[0,1]$ such that

$$\sup_{t\in[0,1]} ||S^{(j)}(t)-s^{(j)}(t)|| \xrightarrow{\mathcal{P}} 0. \tag{3.59}$$

Condition 3.7.2.4 (Asymptotic regularity conditions)**.** *$s^{(0)}(t)$ is bounded away from zero.*

Condition 3.7.2.5. *There exists a positive definite matrix* Ω *such that*

$$\frac{1}{n}\sum_{i=1}^{n}\int_{0}^{1} Y_i(t)\Big(\bar{Z}_{0i}-\frac{s^{(1)}(t)}{s^{(0)}(t)}\Big)^{\otimes 2}dt \xrightarrow{a.s.} \Omega. \tag{3.60}$$

Let $\bar{Z}_{0i}=[X_{ei},{X_{oi}}^{\top},\Delta_i]$, $\bar{Z}_0(t)=\sum_{l=1}^{n}\bar{Z}_{0l}Y_l(t)/\sum_{l=1}^{n}Y_l(t)$, define

$$\mathbb{A}_1 = \sum_{i=1}^{n}\int_{0}^{1}(Z_i-\bar{Z}_{0i})dM_i(t), \tag{3.61}$$

$$\mathbb{A}_2 = \sum_{i=1}^{n}\int_{0}^{1}(\bar{Z}(t)-\bar{Z}_0(t))dM_i(t), \tag{3.62}$$

$$\mathbb{A}_3 = \rho_0\sum_{i=1}^{n}\Big\{\int_{0}^{1}(Z_i-\bar{Z}_{0i})Y_i(t)dt\Big\}(\Delta_i-\hat{\Delta}_i), \tag{3.63}$$

$$\mathbb{A}_4 = \rho_0\sum_{i=1}^{n}\Big\{\int_{0}^{1}(\bar{Z}(t)-\bar{Z}_0(t))Y_i(t)dt\Big\}(\Delta_i-\hat{\Delta}_i). \tag{3.64}$$

Lemma 3.7.2.1. $\mathbb{A}_1$*,* $\mathbb{A}_2$*,* $\mathbb{A}_3$*,* $\mathbb{A}_4$ *are bounded in probability.*

Proof. Notice that

$$Z_i-\bar{Z}_{0i}=[0,0_p,\hat{\Delta}_i-\Delta_i]^{\top},$$

where $\hat{\Delta}_i-\Delta_i=g^{-1}(\tilde{X}_i^{\top}\alpha_T)-g^{-1}(\tilde{X}_i^{\top}\hat{\alpha})=(g^{-1})'(\tilde{X}_i^{\top}\alpha_T)+o_p(1)$. Therefore it suffices to check the $\hat{\Delta}_i-\Delta_i$ components. Then the only nonzero entry of $\mathbb{A}_1$ is

$$\sum_{i=1}^{n}\int_{0}^{1}\{\hat{\Delta}_i-\Delta_i\}dM_i(t)=\Big\{-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\int_{0}^{1}\tilde{X}_i^{\top}(g^{-1})'(\tilde{X}_i^{\top}\alpha_T)dM_i(t)\Big\}\sqrt{n}(\hat{\alpha}-\alpha_T)+o_p(1),$$

which is bounded in probability by central limit theorem. Next, we check the nonzero entry of

$\mathbb{A}_2$,

$$\begin{aligned}
&\sum_{i=1}^{n}\int_0^1 \frac{\sum_{l=1}^n(\hat{\Delta}_l-\Delta_l)Y_l(t)}{\sum_{l=1}^n Y_l(t)}dM_i(t)\\
=&\ \Big\{-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\int_0^1 \frac{\sum_{l=1}^n \tilde{X}_l^\top(g^{-1})'(\tilde{X}_l^\top\alpha_T)Y_l(t)}{\sum_{l=1}^n Y_l(t)}dM_i(t)\Big\}\sqrt{n}(\hat{\alpha}-\alpha_T)+o_p(1),
\end{aligned}$$

where $\frac{\sum_{l=1}^n \tilde{X}_l^\top(g^{-1})'(\tilde{X}_l^\top\alpha_T)Y_l(t)}{\sum_{l=1}^n Y_l(t)}$ will have a sup norm limit by Proposition 3.7.1.1. Name that limit $\mathcal{K}(t)$, we will get

$$\begin{aligned}
&\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\int_0^1 \frac{\sum_{l=1}^n \tilde{X}_l^\top(g^{-1})'(\tilde{X}_l^\top\alpha_T)Y_l(t)}{\sum_{l=1}^n Y_l(t)}dM_i(t)\\
=&\ \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\int_0^1 \Big\{\frac{\sum_{l=1}^n \tilde{X}_l^\top(g^{-1})'(\tilde{X}_l^\top\alpha_T)Y_l(t)}{\sum_{l=1}^n Y_l(t)}-\mathcal{K}(t)\Big\}dM_i(t)+\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\int_0^1\mathcal{K}(t)dM_i(t).
\end{aligned}$$

From Lemma 3.7.1.6 and Proposition 3.7.1.1, the first term will be $o_p(1)$ and second will converge weakly. Finally, rearranging the nonzero term in $\mathbb{A}_3$ gives

$$\begin{aligned}
&\rho_0\sum_{i=1}^{n}(\hat{\Delta}_i-\Delta_i)(\int_0^1 Y_i(t)dt)(\hat{\Delta}_i-\Delta_i)\\
=&\ \sqrt{n}(\hat{\alpha}-\alpha_T)^\top\Big\{\frac{\rho_0}{n}\sum_{i=1}^{n}(\int_0^1 Y_i(t)dt)\tilde{X}_i\tilde{X}_i^\top(g^{-1})'^2(\tilde{X}_i^\top\alpha_T)\Big\}\sqrt{n}(\hat{\alpha}-\alpha_T)+o_p(1),
\end{aligned}$$

together with the nonzero entry in $\mathbb{A}_4$,

$$\sqrt{n}(\hat{\alpha}-\alpha_T)^\top\Big\{\frac{\rho_0}{n}\sum_{i=1}^{n}\int_0^1\frac{\sum_{l=1}^n\tilde{X}_lY_l(t)}{\sum_{l=1}^n Y_l(t)}Y_i(t)\tilde{X}_i^\top(g^{-1})'^2(\tilde{X}_i^\top\alpha_T)dt\Big\}\sqrt{n}(\hat{\alpha}-\alpha_T)+o(1).$$

The boundedness of these two terms in probability follows from similar statement. □

Proof of Theorem 3.2.0.1. With simple algebra, we have

$$\begin{aligned} U(\beta) &= \frac{1}{n}\sum_{i=1}^{n}\int_0^1 (Z_i-\bar{Z}(t))(dN_i(t)-Y_i(t)\beta^\top Z_i dt) \\ &= \frac{1}{n}\sum_{i=1}^{n}\int_0^1 (Z_i-\bar{Z}(t))(dM_i(t)+Y_i(t)d\Lambda_0(t)+Y_i(t)\beta_T^\top \bar{Z}_{0i}dt-Y_i(t)\beta^\top Z_i dt) \\ &= \frac{1}{n}\sum_{i=1}^{n}\int_0^1 (Z_i-\bar{Z}(t))(dM_i(t)+Y_i(t)\beta_T^\top \bar{Z}_{0i}dt-Y_i(t)\beta^\top Z_i dt). \end{aligned}$$

Plugging in the true parameter β_T, we can decompose it into

$$\begin{aligned} U(\beta_T) &= \frac{1}{n}\sum_{i=1}^{n}\int_0^1 (Z_i-\bar{Z}(t))dM_i(t)+\frac{\rho_0}{n}\sum_{i=1}^{n}\Big\{\int_0^1 (Z_i-\bar{Z}(t))Y_i(t)dt\Big\}(\Delta_i-\hat{\Delta}_i) \\ &= \frac{1}{n}\sum_{i=1}^{n}\int_0^1 \Big(\bar{Z}_{0i}-\frac{s^{(1)}(t)}{s^{(0)}(t)}\Big)dM_i(t) \\ &\quad +\frac{\rho_0}{n}\Big[\sum_{i=1}^{n}\Big\{\int_0^1 \Big(\bar{Z}_{0i}-\frac{s^{(1)}(t)}{s^{(0)}(t)}\Big)Y_i(t)dt\Big\}\tilde{X}_i^\top (g^{-1})'(\tilde{X}_i^\top \alpha_T)\Big](\hat{\alpha}-\alpha_T) \\ &\quad +\frac{1}{n}\mathbb{A}_1+\frac{1}{n}\mathbb{A}_2+\frac{1}{n}\mathbb{A}_3+\frac{1}{n}\mathbb{A}_4+o_p(1). \end{aligned}$$

The first term is a sample mean of n i.i.d. martingale integrals of predictable functions, thus by Law of Large Numbers is $o_p(1)$. Following from Condition 3.7.2.2 and Law of Large Numbers again, the second term will tend to zero in probability. The remaining terms are all $o_p(1)$ by Lemma 3.7.2.1. Hence $U(\beta_T)=o_p(1)$.

Meanwhile, we can also express

$$U(\beta_T)=U(\beta_T)-U(\hat{\beta})=\Big\{\frac{1}{n}\sum_{i=1}^{n}\int_0^1 (Z_i-\bar{Z}(t))Y_i(t)Z_i dt\Big\}(\hat{\beta}-\beta_T).$$

Solving for $\hat{\beta}-\beta_T$ we get,

$$\begin{aligned}\hat{\beta}-\beta_T &= \Big\{\frac{1}{n}\sum_{i=1}^{n}\int_0^1 Y_i(t)(Z_i-\bar{Z}(t))^{\otimes 2}dt\Big\}^{-1}U(\beta_T)\\ &= \Big\{\frac{1}{n}\sum_{i=1}^{n}\int_0^1 Y_i(t)\Big(\bar{Z}_{0i}-\frac{s^{(1)}(t)}{s^{(0)}(t)}\Big)^{\otimes 2}dt+o_p(1)\Big\}^{-1}U(\beta_T).\end{aligned}$$

Condition 3.7.2.5 and $U(\beta_T)=o_p(1)$ imply the consistency of $\hat{\beta}$ from Slutsky's theorem. □

Proof of Theorem 3.2.0.2. Observe that

$$\begin{aligned}&\sqrt{n}U(\beta_T)\\ =\;& \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\int_0^1\Big(\bar{Z}_{0i}-\frac{s^{(1)}(t)}{s^{(0)}(t)}\Big)dM_i(t)\\ &+\frac{\rho_0}{n}\Big[\sum_{i=1}^{n}\Big\{\int_0^1\Big(\bar{Z}_{0i}-\frac{s^{(1)}(t)}{s^{(0)}(t)}\Big)Y_i(t)dt\Big\}\tilde{X}_i^{\top}(g^{-1})'(\tilde{X}_i^{\top}\alpha_T)\Big]\sqrt{n}(\hat{\alpha}-\alpha_T)\\ &+\frac{1}{\sqrt{n}}\mathbb{A}_1+\frac{1}{\sqrt{n}}\mathbb{A}_2+\frac{1}{\sqrt{n}}\mathbb{A}_3+\frac{1}{\sqrt{n}}\mathbb{A}_4+o_p(1)\\ =\;& \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\int_0^1\Big(\bar{Z}_{0i}-\frac{s^{(1)}(t)}{s^{(0)}(t)}\Big)dM_i(t)+\Psi\sqrt{n}(\hat{\alpha}-\alpha_T)+o_p(1),\end{aligned}$$

where

$$\Psi=\rho_0\mathbb{E}\Big\{\int_0^1\Big(\bar{Z}_{01}-\frac{s^{(1)}(t)}{s^{(0)}(t)}\Big)Y_1(t)\tilde{X}_1^{\top}(g^{-1})'(\tilde{X}_1^{\top}\alpha_T)dt\Big\}. \tag{3.65}$$

With Condition 3.7.2.1, 3.7.2.2 and 3.7.2.3, $\sqrt{n}U(\beta_T)$ can be written into a sum of i.i.d. random variables with mean zero and finite second moments. Thus the Multivariate Central Limit Theorem together with the Slutsky's Theorem proves that our estimator is asymptotically normally distributed with mean zero.

Lastly, we can compute the covariance matrix of this asymptotic normal distribution.

Notice that the asymptotic covariance matrix of $\sqrt{n}U(\beta_T)$ is by Condtion 3.7.2.2,

$$\begin{aligned}
& E\Big\{\Big(\int_0^1 \Big(\bar{Z}_{01} - \frac{s^{(1)}(t)}{s^{(0)}(t)}\Big) dM_1(t) + \Psi I^{-1} U_1\Big)^{\otimes 2}\Big\} \\
= \ & \mathbb{E}\Big\{\int_0^1 \Big(\bar{Z}_{01} - \frac{s^{(1)}(t)}{s^{(0)}(t)}\Big) dM_1(t)\Big\}^{\otimes 2} \\
& + \Psi I^{-1} \mathbb{E}\{U_1^{\otimes 2}\}(I^{-1})^\top \Psi^\top \\
& + 2\mathbb{E}\Big\{\int_0^1 \Big(\bar{Z}_{01} - \frac{s^{(1)}(t)}{s^{(0)}(t)}\Big) dM_1(t) U_1^\top\Big\}\Big] (I^{-1})^\top \Psi^\top \\
= \ & \Sigma_1 + \Sigma_2 + 2 \lim_{n\to\infty} \mathbb{E}\Big\{\sum_{i=1}^n \int_0^1 \Big(\bar{Z}_{0i} - \frac{s^{(1)}(t)}{s^{(0)}(t)}\Big) dM_i(t)(\hat{\alpha} - \alpha_T)^\top\Big\}\Psi^\top \\
= \ & \mathbb{E}\Big\{\int_0^1 \Big(\bar{Z}_{01} - \frac{s^{(1)}(t)}{s^{(0)}(t)}\Big)^{\otimes 2} d[M_1(t), M_1(t)]\Big\} \\
& + \Psi I^{-1} \mathbb{E}(U_1^{\otimes 2})(I^{-1})^\top \Psi^\top \\
& + 2\mathbb{E}\Big\{\int_0^1 \Big(\bar{Z}_{01} - \frac{s^{(1)}(t)}{s^{(0)}(t)}\Big) dM_1(t) U_1^\top\Big\}\Big] (I^{-1})^\top \Psi^\top \\
= \ & \Sigma_1 + \Sigma_2 + 2\mathbb{E}\Big\{\int_0^1 \Big(\bar{Z}_{01} - \frac{s^{(1)}(t)}{s^{(0)}(t)}\Big) dM_1(t) U_1^\top\Big\}\Big] (I^{-1})^\top \Psi^\top, \qquad (3.66)
\end{aligned}$$

where

$$\begin{aligned}
\Sigma_1 &= \mathbb{E}\Big\{\int_0^1 \Big(\bar{Z}_{01} - \frac{s^{(1)}(t)}{s^{(0)}(t)}\Big)^2 d\langle M_1, M_1\rangle(t)\Big\}, && (3.67) \\
\Sigma_2 &= \Psi\Theta\Psi^\top, && (3.68)
\end{aligned}$$

and Θ is the variance-covariance matrix of the first step estimator. Note that the last term in (3.66) is zero since it is still a martingale integral. Thus the asymptotic variance of our estimator $\sqrt{n}(\hat{\beta} - \beta) = \Omega^{-1}(\sqrt{n}U(\beta_T) + o_p(1))$, where Ω is given in Condtion 3.7.2.5, is clearly $\Omega^{-1}(\Sigma_1 + \Sigma_2)\Omega^{-1}$. To consistently estimate the variance, we just use their corresponding empirical parts. Specifically, Σ_1 can be instead estimated in the form of (3.13) by Lemma 3.7.1.4. □

Proposition 3.7.2.2. *Under* (3.1)*,* (3.3)*,* (3.5) *and Conditions 3.7.2.1,3.7.2.4,3.7.2.5, the cumulative baseline hazard function estimator defined in* (3.9) *converges in probability to the true value*

Λ_{0T} *of* $\Lambda_0(\cdot) = \int_0^{\cdot} \bar{\lambda}_0(t)dt$ *uniformly in* $t \in [0,1]$*, where* $\bar{\lambda}_0(t)$ *is the baseline hazard in equation* (3.4)*, and the process* $\sqrt{n}\{\hat{\Lambda}_0(\cdot) - \Lambda_{0T}(\cdot)\}$ *converges weakly to a zero-mean Gaussian process whose covariance function at* (t,s)*, where* $0 \leq s \leq t$*, can be consistently estimated by*

$$\begin{aligned} & n\sum_{i=1}^{n}\int_0^s \frac{1}{(\sum_{j=1}^{n} Y_j(u))^2} dN_i(u) + \hat{C}^\top(t)\hat{\Omega}^{-1}(\hat{\Sigma}_1 + \hat{\Sigma}_2)\hat{\Omega}^{-1}\hat{C}(s) + \hat{E}^\top(t)\hat{\Theta}\hat{E}(s) \\ & -\hat{C}^\top(t)\hat{\Omega}^{-1}\hat{D}(s) - \hat{C}^\top(s)\hat{\Omega}^{-1}\hat{D}(t), \qquad (3.69) \end{aligned}$$

where

$$\begin{aligned} \hat{C}(t) &= \int_0^t \bar{Z}(u)du, & (3.70) \\ \hat{D}(t) &= \sum_{i=1}^{n}\int_0^t \frac{Z_i - \bar{Z}(u)}{\sum_{j=1}^{n} Y_j(u)} dN_i(u), & (3.71) \\ \hat{E}(t) &= \hat{\rho}_0 \sum_{i=1}^{n} \tilde{X}_i (g^{-1})'(\tilde{X}_i^\top \hat{\alpha}) \int_0^t \frac{Y_i(u)}{\sum_{j=1}^{n} Y_j(u)} du. & (3.72) \end{aligned}$$

Proof of Proposition 3.7.2.2.

$$U_1(\Lambda_0(t),\beta,t) = \frac{1}{n}\sum_{i=1}^{n} M_i(\Lambda_0(t),\beta,t) = \frac{1}{n}\sum_{i=1}^{n}\int_0^t (dN_i(u) - Y_i(u)d\Lambda_0(u) - Y_i(t)\beta^\top Z_i du),$$

note that $U_1(\hat{\Lambda}_0(t),\hat{\beta},t) \equiv 0$. Thus we have

$$U_1(\Lambda_{0T}(t),\hat{\beta},t) = U_1(\Lambda_{0T}(t),\hat{\beta},t) - U_1(\hat{\Lambda}_0(t),\hat{\beta},t) = \frac{1}{n}\int_0^t \sum_{i=1}^{n} Y_i(u)d(\hat{\Lambda}_0(u) - \Lambda_{0T}(u)).$$

Meanwhile, observe that

$$\begin{aligned}
& U_1(\Lambda_{0T}(t), \hat{\beta}, t) \\
= \ & \frac{1}{n}\sum_{i=1}^{n}\int_0^t (dN_i(u) - Y_i(u)d\Lambda_{0T}(u) - Y_i(t)\hat{\beta}^\top Z_i du) \\
= \ & \frac{1}{n}\sum_{i=1}^{n}\int_0^t \{dM_i(u) - Y_i(u)(\hat{\beta}^\top Z_i - \beta_T^\top \bar{Z}_{0i})du\} \\
= \ & \frac{1}{n}\sum_{i=1}^{n}\int_0^t \big\{dM_i(u) - Y_i(u)(\hat{\beta} - \beta_T)^\top \bar{Z}_{0i}du - Y_i(u)\hat{\rho}_0(\hat{\Delta}_i - \Delta_i)du\big\} \\
= \ & \frac{1}{n}\sum_{i=1}^{n}\int_0^t \big\{dM_i(u) - Y_i(u)(\hat{\beta} - \beta_T)^\top \bar{Z}_{0i}du \\
& + Y_i(u)\hat{\rho}_0\tilde{X}_i^\top (g^{-1})'(\tilde{X}_i^\top \alpha_T)(\hat{\alpha} - \alpha_T)du\big\} + o_p(1).
\end{aligned}$$

where we shall emphasize that the $o_p(1)$ here is uniformly over $t \in [0,1]$, which can be easily proved with the help of Lemma 3.7.1.6. These together gives us that

$$\begin{aligned}
& \hat{\Lambda}_0(t) - \Lambda_{0T}(t) \\
= \ & \sum_{i=1}^{n}\int_0^t \frac{1}{\sum_{j=1}^n Y_j(u)} dM_i(u) - \Big\{\sum_{i=1}^{n}\int_0^t \frac{Y_i(u)Z_i^\top}{\sum_{j=1}^n Y_j(u)} du\Big\}(\hat{\beta} - \beta_T) \\
& - \Big\{\hat{\rho}_0 \sum_{i=1}^{n}\int_0^t \frac{Y_i(u)\tilde{X}_i^\top (g^{-1})'(\tilde{X}_i^\top \alpha_T)}{\sum_{j=1}^n Y_j(u)} du\Big\}(\hat{\alpha} - \alpha_T).
\end{aligned}$$

It is easy to establish the uniform convergence in time t by checking for each term. For the weak convergence, observe that

$$\begin{aligned}
& \sqrt{n}(\hat{\Lambda}_0(t) - \Lambda_{0T}(t)) \\
= \ & \sqrt{n}\sum_{i=1}^{n}\int_0^t \frac{dM_i(u) - Y_i(u)(\hat{\beta} - \beta_T)^\top Z_i du - Y_i(u)\hat{\rho}_0\tilde{X}_i^\top (g^{-1})'(\tilde{X}_i^\top \alpha_T)(\hat{\alpha} - \alpha_T)du}{\sum_{j=1}^n Y_j(u)} \\
= \ & \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\int_0^t \frac{1}{s^{(0)}(u)} dM_i(u) - \Big(\int_0^t \frac{s^{(1)}(u)}{s^{(0)}(u)} du\Big)^\top \sqrt{n}(\hat{\beta} - \beta_T) \\
& - \Big(\int_0^t \frac{\gamma(u)}{s^{(0)}(u)} du\Big)^\top \sqrt{n}(\hat{\alpha} - \alpha_T) + o_p(1),
\end{aligned}$$

where $\gamma(u) = \lim_{n\to\infty} \frac{\hat{\rho}_0}{n}\sum_{i=1}^{n} Y_i(u)\tilde{X}_i^\top (g^{-1})'(\tilde{X}_i^\top \alpha_T)$. With Condition 3.7.2.4, Martingale Central Limit Theorem 5.1.1 in [FH11] can be employed to show that $\sqrt{n}(\hat{\Lambda}_0(t) - \Lambda_{0T}(t))$ converges a mean zero Gaussian process with respect to the Skorohod topology with covariance function at $0 \leq s \leq t$ obtained with the help of pointwise Multivariate Central Limit Theorem,

$$
\begin{aligned}
&\mathbb{E}\Big[\Big\{\int_0^t \frac{1}{s^{(0)}(u)} dM_1(u) - \Big(\int_0^t \frac{s^{(1)}(u)}{s^{(0)}(u)} du\Big)^\top \Omega^{-1} \int_0^1 \Big(\bar{Z}_{01} - \frac{s^{(1)}(u)}{s^{(0)}(u)}\Big) dM_1(u) \\
&-\Big(\int_0^t \frac{\gamma(u)}{s^{(0)}(u)} du\Big)^\top \Psi I^{-1} U_1\Big\}\Big\{\int_0^s \frac{1}{s^{(0)}(u)} dM_1(u) \\
&-\Big(\int_0^s \frac{s^{(1)}(u)}{s^{(0)}(u)} du\Big)^\top \Omega^{-1} \int_0^1 \Big(\bar{Z}_{01} - \frac{s^{(1)}(t)}{s^{(0)}(t)}\Big) dM_1(t) - \Big(\int_0^s \frac{\gamma(u)}{s^{(0)}(u)} du\Big)^\top \Psi I^{-1} U_1\Big\}\Big] \\
= \;& \mathbb{E}\Big\{\int_0^s \frac{1}{(s^{(0)}(u))^2} d\langle M_1, M_1\rangle(u)\Big\} \\
&+\Big(\int_0^t \frac{s^{(1)}(u)}{s^{(0)}(u)} du\Big)^\top \Omega^{-1} \mathbb{E}\Big\{\int_0^1 \Big(\bar{Z}_{01} - \frac{s^{(1)}(u)}{s^{(0)}(u)}\Big)^{\otimes 2} d\langle M_1, M_1\rangle(u)\Big\} \Omega^{-1} \Big(\int_0^s \frac{s^{(1)}(u)}{s^{(0)}(u)} du\Big) \\
&+\Big(\int_0^t \frac{\gamma(u)}{s^{(0)}(u)} du\Big)^\top \Psi \mathbb{E}(U_1^{\otimes 2}) \Psi^\top \Big(\int_0^s \frac{\gamma(u)}{s^{(0)}(u)} du\Big) \\
&-\mathbb{E}\Big\{\Big(\int_0^t \frac{s^{(1)}(u)}{s^{(0)}(u)} du\Big)^\top \Omega^{-1} \int_0^t \frac{\bar{Z}_{01} - \frac{s^{(1)}(u)}{s^{(0)}(u)}}{s^{(0)}(u)} d\langle M_1, M_1\rangle(u)\Big\} \\
&-\mathbb{E}\Big\{\Big(\int_0^s \frac{s^{(1)}(u)}{s^{(0)}(u)} du\Big)^\top \Omega^{-1} \int_0^t \frac{\bar{Z}_{01} - \frac{s^{(1)}(u)}{s^{(0)}(u)}}{s^{(0)}(u)} d\langle M_1, M_1\rangle(u)\Big\} \\
&-2\mathbb{E}\Big\{\int_0^t \frac{1}{s^{(0)}(u)} dM_1(u) U_1^\top\Big\} (I^{-1})^\top \Psi^\top \int_0^s \frac{\gamma(u)}{s^{(0)}(u)} du \\
&+2\mathbb{E}\Big\{\Big(\int_0^t \frac{s^{(1)}(u)}{s^{(0)}(u)} du\Big)^\top \Omega^{-1} \int_0^1 \Big(\bar{Z}_{01} - \frac{s^{(1)}(u)}{s^{(0)}(u)}\Big) dM_1(u) U_1^\top\Big\} (I^{-1})^\top \Psi^\top \int_0^s \frac{\gamma(u)}{s^{(0)}(u)} du
\end{aligned}
$$

where the last two terms can be proved to be zero by the similar approach as in the proof of Theorem 3.2.0.2. The remaining terms can be estimated by their empirical parts with the help of Lemma 3.7.1.4. □

Proof of Theorem 3.2.0.3. The proof is similarly to the proof of Theorem 3.3.0.3, so omitted. □

3.7.3 Competing risks

Proposition 3.7.3.1. *Assuming* (3.19) *and* (3.3)*, we have*

$$\bar{\lambda}_{10}(t|X_e, X_I, X_o) = \bar{\lambda}_{10}(t) + \beta_e X_e + \bar{\beta_o}^\top X_o + \rho_0 \Delta. \tag{3.73}$$

where $\bar{\lambda}_{10}(t) = \lambda_{10}(t) - \frac{\partial}{\partial t} \log \big[\mathbb{E} \big\{ \exp \big(-\varepsilon t \big) \big\} \big]$.

Proof. The proof is completely parallel to that of Proposition 3.7.2.1 with $1 - F_1$ in place of S. □

In the following we give the regularity conditions and additional notation for the Theorems in Section 3.3.

Define $S^{(j)}(t) = \sum_{i=1}^n Y_i(t) \hat{w}_i(t) Z_i^{\otimes j}/n$ for $j = 0$ and 1. Then there exists a scalar and vector function $s^{(0)}(t)$ and $s^{(1)}(t)$, respectively, defined on $[0,1]$ such that for $j = 0, 1$,

$$\sup_{t \in [0,1]} \big| |S^{(j)}(t) - s^{(j)}(t)| \big| \xrightarrow{\mathcal{P}} 0. \tag{3.74}$$

The above is guaranteed by Gilvenko-Cantelli theorem, Condition 3.7.2.2 and consistency of the Kaplan-Meier estimator.

Condition 3.7.3.1 (Asymptotic regularity conditions)**.** $s^{(0)}(t)$ *is bounded away from zero.*

Condition 3.7.3.2. *There exists a positive definite matrix* Ω *such that*

$$\frac{1}{n} \sum_{i=1}^n \int_0^1 \tilde{w}_i(t) Y_i(t) \Big(\bar{Z}_{0i} - \frac{s^{(1)}(t)}{s^{(0)}(t)} \Big)^{\otimes 2} dt \xrightarrow{a.s.} \Omega. \tag{3.75}$$

Although $w(t)$ is not adaptable to $\mathcal{F}_{t-}^1$, we still have the following result:

Lemma 3.7.3.1. *For any* $H(T,t)$ *adaptable to* $\mathcal{F}_{t-}$*, we have*

$$\mathbb{E} \Big\{ \int_0^1 H(T,t) w(t) dM^1(t) \Big\} = 0. \tag{3.76}$$

Proof. The martingale $M^1(t)$ associated with the counting process $N(t)$ has bounded variation. So we can define the integral pathwisely, i.e.,

$$\int_0^1 H(T,t)w(t)dM^1(t) = \lim_{\max\{t_i - t_{i-1}\} \to 0} \sum_{i=1}^n H(T,t_i)w(t_i)(M^1(t_i) - M^1(t_{i-1})) \tag{3.77}$$

where $\{0 = t_0 < t_1 < ... < t_{n-1} < t_n = 1\}$ is a partition on $[0,1]$. Notice that

$$\mathbb{E}\left\{H(T,t_i)w(t_i)(M^1(t_i) - M^1(t_{i-1}))\right\} = \mathbb{E}\left\{H(T,t_i)\mathbb{E}\left(w(t_i)(M^1(t_i) - M^1(t_{i-1}))\big|\mathcal{F}^1_{t_{i-1}}\right)\right\},$$

where

$$\begin{aligned}
& \mathbb{E}\left\{w(t_i)(M^1(t_i) - M^1(t_{i-1}))\big|\mathcal{F}^1_{t_{i-1}}\right\} \\
= \; & \mathbb{E}\left\{(M^1(t_i) - M^1(t_{i-1}))\mathbb{E}\left(w(t_i)\big|\mathcal{F}^1_{t_i}\right)\big|\mathcal{F}^1_{t_{i-1}}\right\} \\
= \; & \mathbb{E}\left\{(M^1(t_i) - M^1(t_{i-1}))\mathbb{E}\left(\mathbb{1}\{C \geq T \wedge t_i\}\frac{G(t_i)}{G(X \wedge t_i)}\big|\mathcal{F}_{t_i}\right)\big|\mathcal{F}^1_{t_{i-1}}\right\} \\
= \; & \mathbb{E}\left\{(M^1(t_i) - M^1(t_{i-1}))G(t_i)\big|\mathcal{F}^1_{t_{i-1}}\right\} = 0.
\end{aligned}$$

Thus it suffices to prove the result by dominated convergence theorem,

$$\begin{aligned}
\mathbb{E}\left\{\int_0^1 H(T,t)w(t)dM^1(t)\right\} &= \mathbb{E}\left\{\lim_{\max\{t_i - t_{i-1}\} \to 0} \sum_{i=1}^n H(T,t_i)w(t_i)(M^1(t_i) - M^1(t_{i-1}))\right\} \\
&= \lim_{\max\{t_i - t_{i-1}\} \to 0} \mathbb{E}\left\{\sum_{i=1}^n H(T,t_i)w(t_i)(M^1(t_i) - M^1(t_{i-1}))\right\} \\
&= 0.
\end{aligned}$$

□

Define

$$
\begin{aligned}
\mathbb{A}_1 &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\int_0^1 (Z_i-\bar{Z}_{0i})(\hat{w}_i(t)-w_i(t))dM_i^1(t)), && (3.78)\\
\mathbb{A}_2 &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\int_0^1 (Z_i-\bar{Z}_{0i})w_i(t)dM_i^1(t), && (3.79)\\
\mathbb{A}_3 &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\int_0^1 (\bar{Z}(t)-\bar{Z}_0(t))(\hat{w}_i(t)-w_i(t))dM_i^1(t), && (3.80)\\
\mathbb{A}_4 &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\int_0^1 (\bar{Z}(t)-\bar{Z}_0(t))w_i(t)dM_i^1(t), && (3.81)\\
\mathbb{A}_5 &= \frac{\rho_0}{\sqrt{n}}\sum_{i=1}^{n}\int_0^1 \left\{\Big(Z_i-\bar{Z}(t)\Big)-\Big(\bar{Z}_{0i}-\frac{s^{(1)}(t)}{s^{(0)}(t)}\Big)\right\}\hat{w}_i(t)Y_i(t)(\Delta_i-\hat{\Delta}_i)dt, && (3.82)
\end{aligned}
$$

Lemma 3.7.3.2. $\mathbb{A}_1$, $\mathbb{A}_2$, $\mathbb{A}_3$, $\mathbb{A}_4$, $\mathbb{A}_5$ *all converge to 0 in probability as $n \to \infty$.*

Proof. For the first term $\mathbb{A}_1$, by the same trick as in [FG99] on the martingale expression of Kaplan-Meier estimator we have,

$$
\begin{aligned}
& \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\int_0^1 (Z_i-\bar{Z}_{0i})(\hat{w}_i(t)-w_i(t))dM_i^1(t)\\
= {} & \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\int_0^1 (Z_i-\bar{Z}_{0i})r_i(t)\frac{G(t)\mathbb{1}\{T_i^*<t\}}{G(X_i\wedge t)}\sum_{j=1}^{n}\int_{T_i^*}^{t}\frac{1}{\sum_{k=1}^{n}\mathbb{1}\{T_k^*>u\}}dM_j^c(u)dM_i^1(t)+o_p(1)\\
= {} & \frac{1}{\sqrt{n}}\sum_{j=1}^{n}\int_0^1 \frac{\sum_{i=1}^{n}(Z_i-\bar{Z}_{0i})w_i(t)dM_i^1(t)\mathbb{1}\{T_i^*\le u\le t\}}{\sum_{k=1}^{n}\mathbb{1}\{T_k^*>u\}}dM_j^c(u)+o_p(1),
\end{aligned}
$$

where $M_j^c(u)$ is a martingale with respect to the censoring filtration

$$
\mathcal{F}^c(u)=\{\mathbb{1}\{X_i\ge t\},\mathbb{1}\{X_i\le t,\delta_i=0\},X_{Ii},X_{oi},J_i,X_{ei},t\le u,i=1,...,n\}. \qquad (3.83)
$$

Because the integrands are adaptable to this filtration after a careful analysis, the first term now becomes a sum of martingale integrals, which by Lemma 3.7.1.6 is $o_p(1)$.

Next, the second term $\mathbb{A}_2$ becomes,

$$
\begin{aligned}
& \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \int_0^1 \{\hat{\Delta}_i - \Delta_i\} \hat{w}_i(t) dM_i^1(t) \\
= \quad & \Big\{ \frac{1}{n} \sum_{i=1}^{n} \int_0^1 \tilde{X}_i^\top (g^{-1})'(\tilde{X}_i^\top \alpha_T) w_i(t) dM_i^1(t) \Big\} \sqrt{n}(\hat{\alpha} - \alpha_T) + o_p(1),
\end{aligned}
$$

by Lemma 3.7.3.1 and law of large numbers, it is $o_p(1)$.

Proofs for $\mathbb{A}_3$ and $\mathbb{A}_4$ are exactly parallel to the first two but more laborious, which we omit here.

For $\mathbb{A}_5$, we have

$$
\begin{aligned}
& \frac{\rho_0}{\sqrt{n}} \sum_{i=1}^{n} \int_0^1 \Big\{ \Big(Z_i - \bar{Z}(t) \Big) \hat{w}_i(t) - \Big(\bar{Z}_{0i} - \frac{s^{(1)}(t)}{s^{(0)}(t)} \Big) \hat{w}_i(t) \Big\} Y_i(t)(\Delta_i - \hat{\Delta}_i) dt \\
= \quad & \Big[\frac{\rho_0}{n} \sum_{i=1}^{n} \int_0^1 \Big\{ \Big(Z_i - \bar{Z}(t) \Big) \hat{w}_i(t) - \Big(\bar{Z}_{0i} - \frac{s^{(1)}(t)}{s^{(0)}(t)} \Big) \hat{w}_i(t) \Big\} \\
& Y_i(t) \tilde{X}_i^\top (g^{-1})'(\tilde{X}_i^\top \alpha_T) dt \Big] \sqrt{n}(\hat{\alpha} - \alpha_T),
\end{aligned}
$$

which will tend to zero in probability by law of large numbers and Gilvenko-Cantelli theorem. □

Proof of Theorem 3.3.0.1. By lemma 3.7.3.2, we can rewrite $U(\beta_T)$ as

$$
\begin{aligned}
& \frac{1}{n} \sum_{i=1}^{n} \int_0^1 \Big(\bar{Z}_{0i} - \frac{s^{(1)}(t)}{s^{(0)}(t)} \Big) w_i(t) dM_i^1(t) + \frac{1}{n} \sum_{i=1}^{n} \int_0^1 \Big(\bar{Z}_{0i} - \frac{s^{(1)}(t)}{s^{(0)}(t)} \Big) (\hat{w}_i(t) - w_i(t)) dM_i^1(t) \\
& + \frac{\rho_0}{n} \sum_{i=1}^{n} \int_0^1 \Big(\bar{Z}_{0i} - \frac{s^{(1)}(t)}{s^{(0)}(t)} \Big) w_i(t) Y_i(t)(\Delta_i - \hat{\Delta}_i) dt + o_p(1).
\end{aligned}
$$

Notice that the first term is not a martingale integral since $w_i(t)$ is not adapted to the complete data filtration $\mathcal{F}^1(t)$. So partition it into

$$\begin{aligned}
& \frac{1}{n}\sum_{i=1}^{n}\int_0^1 \Big(\bar{Z}_{0i} - \frac{s^{(1)}(t)}{s^{(0)}(t)}\Big) w_i(t) dM_i^1(t) \\
= \; & \frac{1}{n}\sum_{i=1}^{n}\int_0^1 \Big(\bar{Z}_{0i} - \frac{s^{(1)}(t)}{s^{(0)}(t)}\Big) \mathbb{1}\{C_i \geq t\} dM_i^1(t) \\
& + \frac{1}{n}\sum_{i=1}^{n}\int_0^1 \Big(\bar{Z}_{0i} - \frac{s^{(1)}(t)}{s^{(0)}(t)}\Big) \big(w_i(t) - \mathbb{1}\{C_i \geq u\}\big) dM_i^1(t).
\end{aligned}$$

The first term tends to zero from the conclusion in censoring complete situation. For the second term, by Fubini-Toneli Theorem and the property of conditional expectation, we have

$$\begin{aligned}
& \mathbb{E}\Big\{\int_0^1 \Big(\bar{Z}_{0i} - \frac{s^{(1)}(t)}{s^{(0)}(t)}\Big) \big(w_i(t) - \mathbb{1}\{C_i \geq t\}\big) dM_i^1(t)\Big\} \\
= \; & \mathbb{E}\Big[\mathbb{E}\Big\{\int_0^1 \Big(\bar{Z}_{0i} - \frac{s^{(1)}(t)}{s^{(0)}(t)}\Big) \big(w_i(t) - \mathbb{1}\{C_i \geq t\}\big) dM_i^1(t) | Z_i\Big\}\Big] \\
= \; & \mathbb{E}\Big[\int_0^1 \Big(\bar{Z}_{0i} - \frac{s^{(1)}(t)}{s^{(0)}(t)}\Big) \mathbb{E}\Big\{\big(w_i(t) - \mathbb{1}\{C_i \geq t\}\big) dM_i^1(t) | Z_i\Big\}\Big],
\end{aligned}$$

where

$$\begin{aligned}
\mathbb{E}\big\{w_1(t) dM_1^1(t) | \bar{Z}_1\big\} &= \mathbb{E}\big\{\mathbb{1}\{C_1 \geq T_1 \wedge t\} G(t)/G(T_1^* \wedge t) dM_1^1(t) | \bar{Z}_1\big\} \\
&= \mathbb{E}\big\{G(t) dM_1^1(t) \mathbb{E}\Big[\mathbb{1}\{C_1 \geq T_1 \wedge t\}/G(T_1^* \wedge t) | \bar{Z}_{0\cdot}, T_\cdot, J_i\Big] | \bar{Z}_1\big\} \\
&= \mathbb{E}\big\{G(t) dM_1^1(t) | \bar{Z}_1\big\} \\
&= \mathbb{E}\big\{\mathbb{E}\Big[\mathbb{1}\{C_1 \geq t\} | \bar{Z}_{01}, T_1, J_1\Big] dM_1^1(t) | \bar{Z}_1\big\} \\
&= \mathbb{E}\big\{\mathbb{1}\{C_1 \geq t\} dM_1^1(t) | \bar{Z}_{01}\big\}.
\end{aligned}$$

Hence the second term also has zero expectation. This established the consistency of the first term by Strong Law of Large Numbers. The remaining two terms can be handled by a similar way as we do in the Lemma 3.7.3.2. □

Proof of Theorem 3.3.0.2. Observe that

$$
\begin{aligned}
\sqrt{n}U(\beta_T) &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\int_0^1 \Big(\bar{Z}_{0i}-\frac{s^{(1)}(t)}{s^{(0)}(t)}\Big)w_i(t)dM_i^1(t) \\
&\quad +\Big\{\frac{\rho_0}{n}\sum_{i=1}^{n}\int_0^1 \Big(\bar{Z}_{0i}-\frac{s^{(1)}(t)}{s^{(0)}(t)}\Big)w_i(t)Y_i(t)\tilde{X}_i^\top(g^{-1})'(\tilde{X}_i^\top\alpha_T)dt\Big\}\sqrt{n}(\hat{\alpha}-\alpha_T) \\
&\quad +\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\int_0^1 \Big(\bar{Z}_{0i}-\frac{s^{(1)}(t)}{s^{(0)}(t)}\Big)(\hat{w}_i(t)-w_i(t))dM_i^1(t)+o_p(1) \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\int_0^1 \Big(\bar{Z}_{0i}-\frac{s^{(1)}(t)}{s^{(0)}(t)}\Big)w_i(t)dM_i^1(t)-\Psi\sqrt{n}(\hat{\alpha}-\alpha_T) \\
&\quad +\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\int_0^1 \frac{q(t)}{\pi(t)}dM_i^c(t)+o_p(1),
\end{aligned}
$$

where

$$
\begin{aligned}
\Psi &= \rho_0\mathbb{E}\Big\{\int_0^1 \Big(\bar{Z}_{01}-\frac{s^{(1)}(t)}{s^{(0)}(t)}\Big)w_1(t)Y_1(t)\tilde{X}_1^\top(g^{-1})'(\tilde{X}_1^\top\alpha_T)dt\Big\}, \\
q(t) &= -\mathbb{E}\Big\{\int_0^1 \mathbb{1}\{T_1^*<t\le u\}w_1(u)\Big(\bar{Z}_{01}-\frac{s^{(1)}(u)}{s^{(0)}(u)}\Big)dM_1^1(u)\Big\}, \\
\pi(t) &= \mathbb{P}(T_1^*\ge t).
\end{aligned}
$$

Therefore now $\sqrt{n}U(\beta_T)$ has been written into a sum of i.i.d. random variables with mean zero and finite second moments. By the Multivariate Central Limit Theorem and Slutsky's theorem, our estimator is asymptotically normal with mean zero and covariance matrix $\Omega^{-1}(\Sigma_1+\Sigma_2+\Sigma_3)\Omega^{-1}$, where the cross terms are zero by similar arguments as in the proof of Theorem 3.2.0.2; here the cross terms are between the event martingale and the censoring martingale. Note that since we assume that the event distribution and the censoring distribution are marginally independent, then

$$
\begin{aligned}
&\mathbb{E}\Big\{\int_0^1 \Big(\bar{Z}_{01}-\frac{s^{(1)}(t)}{s^{(0)}(t)}\Big)w_1(t)dM_1^1(t)\int_0^1\frac{q(t)}{\pi(t)}dM_1^c(t)\Big\} \\
= \;&\mathbb{E}\Big\{\int_0^1 \Big(\bar{Z}_{01}-\frac{s^{(1)}(t)}{s^{(0)}(t)}\Big)w_1(t)dM_1^1(t)\Big\}\mathbb{E}\Big\{\int_0^1\frac{q(t)}{\pi(t)}dM_1^c(t)\Big\}=0.
\end{aligned}
$$

□

Proposition 3.7.3.2. *Under* (3.19)*,* (3.3)*,* (3.5) *on the subdistribution hazard function and assumptions 3.7.2.1,3.7.3.1,3.7.3.2, the baseline hazard function estimator defined in* (3.27) *converges to the baseline hazard function uniformly after introducing the residual term and the process* $\sqrt{n}\{\hat{\Lambda}_{10}(\cdot) - \Lambda_{10T}(\cdot)\}$ *converges weakly to a zero-mean Gaussian process whose covariance function at* (t,s)*, where* $0 \leq s \leq t$*, can be consistently estimated by*

$$\begin{aligned}&\int_0^s \frac{n\sum_{i=1}^n \hat{w}_i(u)dN_i(u)}{(\sum_{j=1}^n \hat{w}_j(u)Y_j(u))^2} + \hat{C}^\top(t)\hat{\Omega}^{-1}(\hat{\Sigma}_1 + \hat{\Sigma}_2 + \hat{\Sigma}_3)\hat{\Omega}^{-1}\hat{C}(s) + \hat{E}^\top(t)\hat{\Theta}\hat{E}(s) \\ &+n\sum_{i=1}^n \int_0^1 \frac{\hat{q}_t(u)\hat{q}_s(u)}{\hat{\pi}^2(u)} dN_i^c(u) - \hat{C}^\top(t)\hat{\Omega}^{-1}\hat{D}(s) - \hat{C}^\top(s)\hat{\Omega}^{-1}\hat{D}(t), \qquad (3.84)\end{aligned}$$

where

$$\hat{C}(t) = \int_0^t \bar{Z}(u)du. \qquad (3.85)$$

Proof of Proposition 3.7.3.2. Similarly in the proof of Proposition 3.7.2.2 in the appendix, we will turn to another score function,

$$U_1(\Lambda_0(t), \beta, t) = \frac{1}{n}\sum_{i=1}^n \int_0^t \hat{w}_i(u)\{dN_i(u) - Y_i(u)d\Lambda_0(u) - Y_i(t)\beta^\top Z_i du\},$$

note that $U_1(\hat{\Lambda}_0(t), \hat{\beta}, t) \equiv 0$. Thus we have

$$\begin{aligned}U_1(\Lambda_{0T}(t), \hat{\beta}, t) &= U_1(\Lambda_{0T}(t), \hat{\beta}, t) - U_1(\hat{\Lambda}_0(t), \hat{\beta}, t) \qquad (3.86)\\ &= \frac{1}{n}\int_0^t \sum_{i=1}^n \hat{w}_i(u)Y_i(u)d(\hat{\Lambda}_0(u) - \Lambda_{0T}(u)). \qquad (3.87)\end{aligned}$$

In the meantime, observe that

$$\begin{aligned}
& U_1(\Lambda_{0T}(t),\hat{\beta},t) \\
= & \frac{1}{n}\sum_{i=1}^{n}\int_0^t \hat{w}_i(u)\{dN_i(u)-Y_i(u)d\Lambda_{0T}(u)-Y_i(t)\hat{\beta}^\top Z_i du\} \\
= & \frac{1}{n}\sum_{i=1}^{n}\int_0^t \hat{w}_i(u)\{dM_i(u)-Y_i(u)(\hat{\beta}^\top Z_i-\beta_T^\top \bar{Z}_{0i})du\} \\
= & \frac{1}{n}\sum_{i=1}^{n}\int_0^t \hat{w}_i(u)\{dM_i(u)-Y_i(u)(\hat{\beta}-\beta_T)^\top Z_i du-Y_i(u)\hat{\rho}_0(\hat{\Delta}_i-\Delta_i)du\} \\
= & \frac{1}{n}\sum_{i=1}^{n}\int_0^t \hat{w}_i(u)\{dM_i(u)-Y_i(u)(\hat{\beta}-\beta_T)^\top \bar{Z}_{0i} du-Y_i(u)\hat{\rho}_0\tilde{X}_i^\top(g^{-1})'(\tilde{X}_i^\top\alpha_T)(\hat{\alpha}-\alpha_T)du\}.
\end{aligned}$$

These together gives us that

$$\begin{aligned}
& \hat{\Lambda}_0(t)-\Lambda_{0T}(t) \\
= & \sum_{i=1}^{n}\int_0^t \frac{\hat{w}_i(u)}{\sum_{j=1}^n \hat{w}_j(u)Y_j(u)}dM_i^1(u)-\Big\{\sum_{i=1}^{n}\int_0^t \frac{\hat{w}_i(u)Y_i(u)\bar{Z}_{0i}^\top}{\sum_{j=1}^n \hat{w}_j(u)Y_j(u)}du\Big\}(\hat{\beta}-\beta_T) \\
& -\Big\{\hat{\rho}_0\sum_{i=1}^{n}\int_0^t \frac{\hat{w}_i(u)Y_i(u)\tilde{X}_i^\top(g^{-1})'(\tilde{X}_i^\top\alpha_T)}{\sum_{j=1}^n \hat{w}_j(u)Y_j(u)}du\Big\}(\hat{\alpha}-\alpha_T).
\end{aligned}$$

Hence the uniform convergence can be established by an application of Gilvenko-Cantelli theorem. For the weak convergence, we just need to check the variance-covariance function of $\sqrt{n}(\hat{\Lambda}_0(t)-$

$\Lambda_{0T}(t))$. First note that

$$\begin{aligned}
&\sqrt{n}(\hat{\Lambda}_0(t) - \Lambda_{0T}(t)) \\
= \;& \sqrt{n}\sum_{i=1}^{n}\int_0^t \frac{w_i(u)}{\sum_{j=1}^{n}\hat{w}_j(u)Y_j(u)} dM_i^1(u) + \sqrt{n}\sum_{i=1}^{n}\int_0^t \frac{\hat{w}_i(u) - w_i(u)}{\sum_{j=1}^{n}\hat{w}_j(u)Y_j(u)} dM_i^1(u) \\
& -\sqrt{n}\sum_{i=1}^{n}\int_0^t \frac{\hat{w}_i(u)Y_i(u)(\hat{\beta} - \beta_T)^\top \bar{Z}_{0i}}{\sum_{j=1}^{n}\hat{w}_j(u)Y_j(u)} du \\
& -\sqrt{n}\sum_{i=1}^{n}\int_0^t \frac{\hat{w}_i(u)Y_i(u)\hat{\rho}_0\tilde{X}_i^\top (g^{-1})'(\tilde{X}_i^\top \alpha_T)(\hat{\alpha} - \alpha_T)}{\sum_{j=1}^{n}\hat{w}_j(u)Y_j(u)} du \\
= \;& \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\int_0^t \frac{w_i(u)}{s^{(0)}(u)} dM_i^1(u) - \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\int_0^t \frac{w_i(u)\sum_{j=1}^{n}\int_{T_i^*}^{u}\frac{1}{\sum_{k=1}^{n}\mathbb{1}\{T_k^* > v\}} dM_j^c(v)}{s^{(0)}(u)} dM_i^1(u) \\
& -\sqrt{n}\Big(\int_0^t \frac{s^{(1)}(u)}{s^{(0)}(u)} du\Big)^\top (\hat{\beta} - \beta_T) - \sqrt{n}\Big(\int_0^t \frac{\gamma(u)}{s^{(0)}(u)} du\Big)^\top (\hat{\alpha} - \alpha_T) + o_p(1) \\
= \;& \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\int_0^t \frac{w_i(u)}{s^{(0)}(u)} dM_i^1(u) \\
& + \frac{1}{\sqrt{n}}\sum_{j=1}^{n}\int_0^1 \frac{1}{\sum_{k=1}^{n}\mathbb{1}\{T_k^* > v\}} \int_0^t \frac{\sum_{i=1}^{n} w_i(u) dM_i^1(u)\mathbb{1}\{T_i^* < v \leq u\}}{s^{(0)}(u)} dM_j^c(v) \\
& -\Big(\int_0^t \frac{s^{(1)}(u)}{s^{(0)}(u)} du\Big)^\top \sqrt{n}(\hat{\beta} - \beta_T) - \Big(\int_0^t \frac{\gamma(u)}{s^{(0)}(u)} du\Big)^\top \sqrt{n}(\hat{\alpha} - \alpha_T) + o_p(1),
\end{aligned}$$

where

$$\gamma(t) = \rho_0 \mathbb{E}\left\{\hat{w}_1(t)Y_i(t)\tilde{X}_1(g^{-1})'(\tilde{X}_1^\top \alpha_T)\right\} \tag{3.88}$$

in the above. With Condition 3.7.3.1, Martingale Central Limit Theorem 5.1.1 in [FH11] can again be employed to show that $\sqrt{n}(\hat{\Lambda}_0(t) - \Lambda_{0T}(t))$ converges in the sense of Skorohod topology to a mean zero Gaussian process with covariance function to be computed by similar way as used in Theorem 3.2.0.2 and Proposition 3.7.2.2. □

Proof of Theorem 3.3.0.3. The consistency simply follows from the results of Theorem 3.3.0.1, Proposition 3.7.3.2 in the appendix and that $\exp(-x)$ is continuously differentiable. For the asymptotic process, we just resort to delta method and Donsker theorem. □

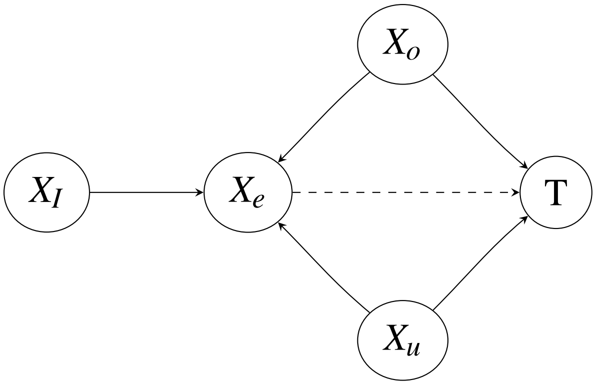

Figure 3.1: A causal directed acyclic graph (DAG) describing the causal relation between variables

Table 3.1: Simulation results for general survival data without competing risks.

Scenario	Sample size	Bias	Emp. Var	Est. Var	Coverage
I	100	-0.15	6.10	6.80	95.8%
I	200	0.03	2.70	2.80	94.9%
I	400	-0.04	0.95	0.92	96.1%
I	800	-0.01	0.43	0.43	94.7%
I	1200	0.00	0.28	0.28	95.3%
II	100	0.06	2.10	2.00	94.6%
II	200	-0.05	0.93	0.89	94.4%
II	400	-0.05	0.45	0.42	96.0%
II	800	0.01	0.21	0.20	95.2%
II	1200	0.00	0.14	0.13	94.4%
III	100	0.12	62.50	64.44	95.7%
III	200	0.07	27.50	26.85	96.0%
III	400	0.07	13.74	12.53	94.9%
III	800	0.08	6.04	6.02	95.5%
III	1200	-0.04	4.09	3.96	95.5%

Table 3.2: Simulation results under the competing risks model.

Scenario	Sample size	Bias	Emp. Var	Est. Var	Coverage
I	100	0.07	3.60	3.40	94.8%
I	200	0.04	1.60	1.50	94.5%
I	400	0.04	0.77	0.74	94.0%
I	800	0.02	0.36	0.36	95.0%
I	1200	0.01	0.24	0.24	94.5%
II	100	0.13	4.10	4.30	95.3%
II	200	0.04	2.00	2.00	94.6%
II	400	-0.04	0.85	0.95	95.7%
II	800	0.01	0.43	0.46	95.8%
II	1200	-0.01	0.31	0.31	95.2%
III	100	0.23	134.80	128.90	95.7%
III	200	-0.16	53.38	58.34	95.0%
III	400	-0.12	26.17	27.14	95.9%
III	800	0.10	13.56	13.87	94.5%
III	1200	0.05	8.85	8.12	95.5%

Table 3.3: Patient characteristics of the SEER-Medicare data set.

	Radical prostatectomy $n = 10977$	Conservative management $n = 18829$
Age		
66-69	4852 (45.3%)	6925 (37.2%)
70-74	5859 (54.7%)	11694 (62.8%)
Marital Status		
Married	7815 (73.0%)	12889 (69.2%)
Divorced	536 (5.0%)	1068 (5.7%)
Single	786 (7.3%)	1450 (7.9%)
Other	1574 (14.7%)	3212 (17.3%)
Race or Ethnity		
Asian	206 (1.9%)	302 (1.6%)
Black	1022 (9.5%)	2495 (13.4%)
Hispanic	184 (1.7%)	222 (1.2%)
White	8973 (83.8%)	15047 (80.8%)
Other	326 (3.0%)	553 (3.0%)
Tumor Stage		
T1	4132 (38.6%)	12059 (64.8%)
T2	6579 (61.4%)	6560 (35.2%)
Tumor Grade		
Well differentiated	168 (1.6%)	140 (0.8%)
Moderately differentiated	5537 (51.7%)	9070 (48.7%)
Poorly differentiated	4793 (44.7%)	9153 (49.2%)
Undifferentiated	17 (0.2%)	26 (0.1%)
Cell type not determined	196 (1.8%)	230 (1.2%)
Prior Charlson comorbidity score		
0	7217 (67.4%)	11868 (63.7%)
1	2301 (21.5%)	4260 (22.9%)
≥ 2	1193 (11.1%)	2491 (13.4%)
Diagnosis year		
2001	345 (3.2%)	241 (1.3%)
2002	311 (2.9%)	268 (1.4%)
2003	277 (2.6%)	207 (1.1%)
2004	1284 (12.0%)	1908 (10.2%)
2005	1217 (11.4%)	1838 (9.9%)
2006	1334 (12.5%)	2252 (12.1%)
2007	1351 (12.6%)	2486 (13.4%)
2008	1291 (12.1%)	2372 (12.7%)
2009	1215 (11.3%)	2381 (12.8%)
2010	1020 (9.5%)	2263 (12.2%)
2011	1066 (10.0%)	2403 (12.9%)

Table 3.4: Results of two stage residual inclusion IV analysis on overall survival with all two-way interactions.

Variable Label	Hazard difference	Standard Errors	Two sided P-value
Radical prostatectomy vs Conservative management	-0.0012	0.00057	0.042
Residual term	0.0010	0.0006	0.083
Age 70-74 vs 66-69	0.0006	0.0005	0.24
Stage T2 vs T1	0.0005	0.0005	0.31
Married vs Other	-0.0008	0.0007	0.23
Divorced vs Other	0.0011	0.0016	0.51
Single vs Other	-0.0006	0.001	0.53
Asian vs Other	-0.0004	0.0019	0.85
Black vs Other	-0.0004	0.0013	0.75
Hispanic vs Other	0.0026	0.0022	0.24
White vs Other	-0.0004	0.0011	0.7
Grade moderately differentiated vs Well differentiated	0.0007	0.0004	0.052
Grade poorly differentiated vs Well differentiated	0.0019	0.0006	0.0009
Grade undifferentiated vs Well differentiated	0.0023	0.001	0.021
Grade cell type not determined vs Well differentiated	0.0035	0.0011	0.0014
Prior Charlson comorbidity score 0 vs ≥ 2	-0.004	0.0010	< 0.0001
Prior Charlson comorbidity score 1 vs ≥ 2	-0.0026	0.0005	< 0.0001
2002 vs 2001	-0.00058	0.0003	0.022
2003 vs 2001	-0.00033	0.0003	0.33
2004 vs 2001	-0.0008	0.0004	0.032
2005 vs 2001	-0.001	0.0005	0.022
2006 vs 2001	-0.0013	0.0005	0.016
2007 vs 2001	-0.0013	0.0006	0.043
2008 vs 2001	-0.0014	0.0007	0.041
2009 vs 2001	-0.0017	0.0008	0.036
2010 vs 2001	-0.0018	0.0009	0.039
2011 vs 2001	-0.0018	0.0010	0.061

Table 3.5: Results of two stage residual inclusion IV analysis on cancer specific survival with all two-way interactions.

Variable Label	Hazard difference	Standard Errors	Two sided P-value
Radical prostatectomy vs Conservative management	4.8×10^{-5}	0.0002	0.83
Residual term	-8.1×10^{-5}	0.0002	0.72
Age 70-74 vs 66-69	0.0005	0.0002	0.03
Stage T2 vs T1	0.0002	0.0002	0.32
Married vs Other	-0.0008	0.0004	0.026
Divorced vs Other	0.0004	0.0009	0.67
Single vs Other	-0.0002	0.0006	0.76
Asian vs Other	-0.0011	0.0011	0.32
Black vs Other	0.0002	0.0007	0.83
Hispanic vs Other	-0.0003	0.0009	0.69
White vs Other	-0.0003	0.0006	0.67
Grade moderately differentiated vs Well differentiated	0.0008	0.0002	0.0003
Grade poorly differentiated vs Well differentiated	0.0018	0.0004	< 0.0001
Grade undifferentiated vs Well differentiated	0.0031	0.0008	0.0001
Grade cell type not determined vs Well differentiated	0.0033	0.0007	< 0.0001
Prior Charlson comorbidity score 0 vs ≥ 2	0.0009	0.0003	0.0031
Prior Charlson comorbidity score 1 vs ≥ 2	0.0004	0.0002	0.0093
2002 vs 2001	-6.1×10^{-5}	0.0001	0.6
2003 vs 2001	-0.0002	0.0002	0.16
2004 vs 2001	-0.0005	0.0002	0.0076
2005 vs 2001	-0.0006	0.0002	0.0046
2006 vs 2001	-0.0007	0.0003	0.0082
2007 vs 2001	-0.0008	0.0003	0.011
2008 vs 2001	-0.0009	0.0004	0.009
2009 vs 2001	-0.001	0.0004	0.0086
2010 vs 2001	-0.0012	0.0005	0.0062
2011 vs 2001	-0.0012	0.0005	0.0082

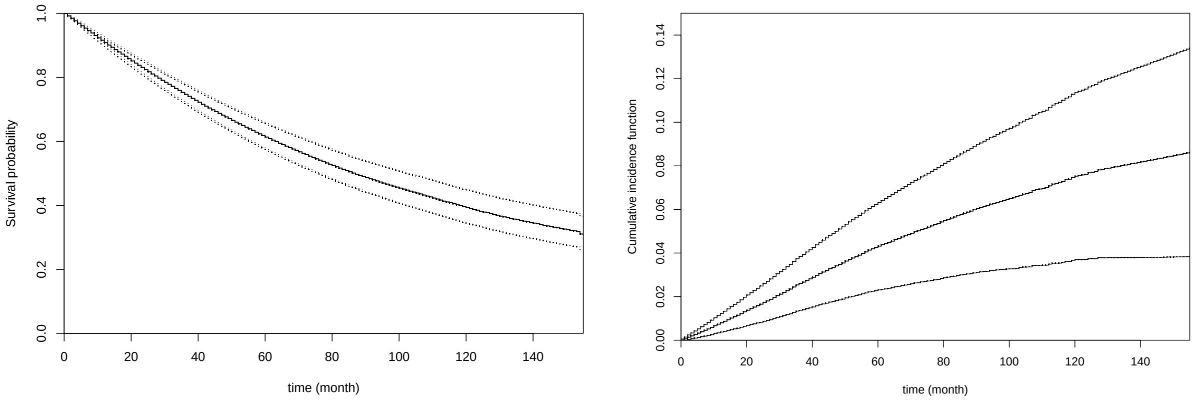

Figure 3.2: Predicted overall survival (left) and cancer specific cumulative incidence (right) function for a patient with pointwise 95% confidence intervals.

3.8 Acknowledgement

Chapter 3, in full, is a reprint of the material as it appears in Statistics in Medicine. Ying, Andrew; Xu, Ronghui; Murphy, James.Two-Stage Residual Inclusion for Survival Data and Competing Risks - An Instrumental Variable Approach with Application to SEER-Medicare Linked Data, 38(10): 1775-1801, 2019. The dissertation author was the primary investigator and author of this paper.

Chapter 4

Causal Effects on Birth Defects with Missing by Terathanasia

4.1 Introduction

Our work was motivated by a recent observational study carried out by the North American Organization of Teratology Information Specialists (OTIS), on the use of etanercept (trade name Enbrel) during pregnancy. Etanercept is a tumor necrosis factor-alpha (TNFα) inhibitor that treats autoimmune diseases such as rheumatoid arthritis and psoriasis. The study found that the proportion of liveborn infants with major birth defects was higher for women exposed to etanercept compared to diseased etanercept unexposed women (US Food & Drug Administration Prescribing Information, revised 9/2017).[1] The elevated birth defect rate was also independently replicated in a Scandinavian study; see the same Food & Drug Administration (FDA) document above. The biological mechanism, if any, behind these elevated birth defect rates was not understood, due to the lack of pattern of major birth defects as traditionally seen in other teratogens such as thalidomide. This led to the FDA statement that "available studies with use of etanercept during

[1] https://www.accessdata.fda.gov/drugsatfda_docs/label/2017/103795s5556lbl.pdf

pregnancy do not reliably support an association between etanercept and major birth defects".

Table 4.1 shows the five types of pregnancy outcomes for a total 494 pregnant women in the OTIS data set for the Etanercept study: live birth, spontaneous abortion (SAB), therapeutic abortion (TAB), stillbirth, and loss to follow up (LTFU). The distinction between SAB and stillbirth is a technical one, if the loss of pregnancy occurs before week 20 of gestation it is SAB, otherwise it is stillbirth. Of these, 336 women were exposed to etanercept during their first trimester, which was the exposure window of interest for major birth defects, and the rest 158 women were unexposed any time during their pregnancy. From Table 4.1 we see that there was a total of 40 observed major birth defects. In addition, there were 27 missing birth defect outcomes, with 25 of these came from SAB. In fact, of the 26 total SAB pregnancy outcomes, only one was observed to have a major birth defect, the rest 25 were missing major birth defect outcomes. As mentioned earlier, SAB is known to be associated with a higher risk of major birth defects, therefore the fact that most SAB's are missing this outcome falls under the mechanism of missing not at random (MNAR).

In the following we study the causal effect of etanercept on major birth defects, making use of the established terathanasia mechanism in dysmorphology to handle the missing data due to SAB. We adopt the potential outcomes framework to define the causal effects, also referred to as the Rubin Causal Model [Ney23, Rub74, Hol86, RCM] In addition, careful analysis of SAB, which was subject to left truncation since most women enrolled in the study after recognition of their pregnancies, showed that the rate was almost twice as high in the unexposed as in the exposed women. This is not surprising since SAB as a type of pregnancy outcome, is clearly a post-exposure variable. It has been recognized in the literature [FR02b] that stratification based on the observed values of such a variable invalidates any causal interpretation. In this case, the analysis of major birth defects within the liveborn stratum did not take into consideration that the pregnancy outcome of live birth versus SAB, for example, might be affected by the exposure to etanercept. Instead, the concept of principal stratification was proposed to address such post-exposure effects within the potential outcomes framework [FR02b].

The rest of the paper proceeds as follows. Due to the multiple pregnancy outcomes and complexity of the data structure, we devote the next section to describe the multiple challenges, detailed notation, as well as the general assumptions. Section 4.3 presents the approach we use to estimate the average treatment effect on major birth defect. Section 4.4 presents the definition of the principal strata as defined by the potential pregnancy outcomes, the causal estimands that we are interested in, together with the estimation and inference approach. Section 4.5 provides the detailed data analysis using the approaches developed in this paper. Section 4.6 contains further discussion. The technical details are provided in the online supplement.

4.2 Challenges, Notation and Assumptions

4.2.1 Main challenges

In the following we provide a roadmap of the main challenges in analyzing the birth defects data for causal inference, and outline our solutions to them.

Missing not at random

As shown in Table 4.1, the major birth defect outcome is missing for most SAB cases. Meanwhile the established terathanasia theory in dysmorphology tells us that the SAB cases are at higher risk of having major birth defects than the pregnancies that end in live births. This results in missing not at random [LR19, MNAR]. Fortunately, the terathanasia theory also informs us how to model such a missing data mechanism, i.e. using the so-called ‘selection model’, which models the marginal distribution of the complete data and the conditional distribution of missingness given the complete data. Here the complete data corresponds to assuming that all major birth defect outcomes are observed, and the missing data mechanism can be modeled as the conditional distribution of SAB/stillbirth given the major birth defect outcomes.

Left Truncation and Right Censoring

As mentioned earlier, left truncation has been known to exist for our SAB data, because women typically enroll in pregnancy studies at OTIS after clinical recognition of their pregnancies [XC11]. This leads to selection bias as women who have early SAB events tend not to be captured in our studies. In addition, loss to follow-up, for example, leads to right censoring. Finally, unlike death, SAB or stillbirth does not happen to all pregnancies; when a pregnancy ends in live birth, we consider it censored for SAB/stillbirth at that time (i.e. gestational age). An alternative consideration is that the pregnancy is 'cured' from SAB/stillbirth if it ends in live birth; we will further discuss this later.

Survival analysis methods have been well studied for left truncated and right-censored (LTRC) time-to-event data. In particular for length-biased data [QNLS11] further developed an early framework of [Var89], by augmenting "ghost copies" for each observed individual in the data set, i.e. through recovery of those similar individuals who have been left truncated out of the observed data set. These approaches lead to an EM algorithm, which was further adopted in [HCX18] for data with an observed cured portion and applied to SAB data analysis. The "ghost copies" approach will also be used in this paper, as it integrates nicely with the missing data above so that EM type algorithms may be applied.

Observational nature and moderate sample size

The prospective cohort studies in pregnancy carried out by OTIS are observational in nature. As such, the presence of confounders is inevitably an important issue to address. Many approaches exist in the literature; for pregnancy studies with birth defect outcomes in particular, various ways to select confounders and use propensity scores were discussed in [XHSC19]. Given the numbers of events in Table 4.1, we will consider parsimonious modeling approaches in the next sections, together with inverse probability weighting (IPW) using propensity scores [RR83, DJ98, WMLC12, WFW14] that were identified in the original analysis that was submitted to the FDA. This results in the minimal number of parameters that need to be estimated.

Principal effects and rare events

Principal stratification is a widely used framework for addressing post-randomization complications. The principal effects [FR02b] provide finer causal effects within principal strata, with exact definitions provided in Section 4.4, at the expense of more unknown parameters. More specifically, we divide the whole population into three sub-populations by monotonicity, with one of them being the main target and the rest two being nuisance. The two nuisance subpopulations have more than 90 percent missing outcomes for at least one arm (treatment or control). Writing down the likelihood for the target subpopulation inevitably introduces nuisance due to the nature of principal stratification, that is, hidden membership. The extremely high missing rates prevent the usage of common strategies like EM algorithm for missing outcomes adopted in Section 4.3 or multiple imputation [Rub96, Rub04]. To draw valid information for the target subpopulation, we offset the parameters introduced by the nuisance and conduct sensitivity analysis. More details can be found in Section 4.5.2.

4.2.2 Outcomes and notation

Among the five types of pregnancy outcomes listed in Table 4.1, we combine SAB and Stillbirth into one outcome SAB/Stillbirth for the purposes of this paper; the early versus late timing during gestation of pregnancy loss is further considered later. As mentioned earlier, left truncation exists in our data for this variable, therefore time to SAB/Stillbirth event will be considered and survival analysis methods will be applied in order to properly handle this selection bias [XC11, HCX18]. In addition, TAB in pregnancy studies should be considered as a competing risk of SAB [MS08], but due to the extremely low number of events in our data, it will be treated as non-informative right censoring. LTFU is the usual right censoring. Finally, live birth informs us that the pregnancy is no longer at risk of SAB, and it will be treated as right censoring for SAB/Stillbirth and further discussion can be found later. For all of these survival random variables, the time scale is gestational age in weeks, and time zero is the start of gestation which is defined as the the first day of the last menstrual period of a pregnant woman.

The data for subjects $i = 1, ..., n$ are treated as independent and identically distributed. Define

- $D_i = 1$ if subject i is treated or exposed, 0 otherwise;
- $Y_i = 1$ if subject i has a major birth defect, 0 otherwise; note that some Y_i's are missing;
- $O_i = 1$ if Y_i is observed, 0 otherwise;
- Q_i the gestational age (in weeks) of subject i at study enrollment;
- T_i the time to SAB/Stillbirth, in gestational weeks;
- C_i the right censoring time;
- $X_i = \min(T_i, C_i)$;
- $\Delta_i = I(T_i \leq C_i)$;
- $M_i = 1$ if subject i has an event of of SAB/Stillbirth, 0 otherwise; note that M_i is missing if subject i is right censored;
- V_i the vector of covariates;
- $t_1 < t_2 < \cdots < t_K$ the K distinct observed SAB/Stillbirth event times.

In addition, we define the following potential outcomes:

- $(Y_i(1), Y_i(0))$ the potential major birth defect outcome under exposure or not, respectively;
- $(M_i(1), M_i(0))$ the potential SAB/Stillbirth outcome under exposure or not, respectively;
- $(T_i(1), T_i(0))$ the potential time to SAB/Stillbirth under exposure or not, respectively.

4.2.3 Causal framework and assumptions

The causal relationship of the variables defined above can be depicted in a graphical display as in Figure 4.1 and Figure 4.2. They give a causal causal directed acyclic graph (DAG) with its single world intervention graph (SWIG), which illustrates the possible causal relationship of all the variables in this study. The dashed lines between D and Y, (M,T) are the causal relations that we want to identify. The covariates V are confounders of (D,Y,T,M). The arrow from Y to (M,T) illustrates the effect of "terathanisia". One can read Assumption 4.3 from the DAG due to the absence of direct arrows between (D,Y,V) and O. One can read Assumption 4.2.3.3 from the SWIG by d-separation rule.

In order to proceed, we assume the following throughout the paper. The first four assumptions are commonly used in causal inference.

Assumption 4.2.3.1 (Stable unit treatment value assumption)**.** *The potential outcomes for one subject are unaffected by the treatment assignments of other subjects, and for each subject there are no hidden versions of treatment or control being considered.*

Assumption 4.2.3.2 (Consistency)**.** *We observe one of the potential outcomes at a time, that is,* $Y = D\cdot Y(1)+(1-D)Y(0)$, $M = D\cdot M(1)+(1-D)M(0)$ *and* $T = D\cdot T(1)+(1-D)T(0)$.

Assumption 4.2.3.3 (Conditionall ignorability)**.** *The treatment assignment is randomized, once given the covariates; that is,*

$$(Y(d),M(d),T(d)) \perp D \mid V, \tag{4.1}$$

where '$\perp$' denotes statistically independent.

Assumption 4.2.3.4 (Positivity)**.** *The propensity scores are bounded away from* 0 *or* 1 *given any covariates; that is, there exists* $\varepsilon > 0$ *such that*

$$\varepsilon \leq \mathbb{P}(D=1|V=v) \leq 1-\varepsilon, \text{ for any } v. \tag{4.2}$$

The next two assumptions are commonly used for survival data, commonly known as non-informative censoring and truncation; see for example [QNLS11, KPJ$^+$17, HCX18]. Note that in the selection model to be specified in the next section, the birth defect outcome Y is a predictor of the SAB outcome T.

Assumption 4.2.3.5 (Conditional independent censoring)**.** *C is independent of (T,Q) given (Y,D,V). There exists a finite number $\tau > 0$ such that $\mathbb{P}(C > \tau) = 0$ and $\mathbb{P}(T > \tau) > 0$.*

Assumption 4.2.3.6 (Conditional quasi-independent truncation)**.** *The truncation time Q and the event time T are independent given (Y,D,V) on the nontruncated region. There exists $X_I \in (0,\tau)$ such that $\mathbb{P}(Q > X_I) = 0$.*

Finally, the following assumption is needed for the selection model; it states that once the information on SAB is included in the data, the major birth defect outcome is missing at random.

Assumption 4.2.3.7 (Missing at random given SAB)**.**

$$\mathbb{P}(O = 1|D,Y,V,T,M) = \mathbb{P}(O = 1|T,M). \tag{4.3}$$

4.3 Average Treatment Effect

4.3.1 Models and weighted likelihood

In this section we focus on the average treatment effect (ATE) of etanercept on major birth defects. We consider the following models for the potential outcomes:

$$\mathbb{P}(Y(d) = 1) = \frac{\exp(\alpha_c + \alpha_D \cdot d)}{1 + \exp(\alpha_c + \alpha_D \cdot d)}, \tag{4.4}$$

and

$$\mathbb{P}(T(d) > t|Y(d) = y) = \exp\{-\Lambda(t)\exp(\beta_D \cdot d + \beta_Y \cdot y)\}, \tag{4.5}$$

where $\Lambda(t)$ is the cumulative baseline hazard function for the conditional distribution of $T(d)$ given $Y(d)$. Then $\exp(\alpha_D)$ is the ATE, which is the causal odds ratio

$$\frac{P(Y(1)=1)/P(Y(1)=0)}{P(Y(0)=1)/P(Y(0)=0)}. \tag{4.6}$$

The parameters β_D and β_Y represent the effect of the treatment and birth defect on SAB/Stillbirth. In particular, β_Y plays the role of quantifying "terathanasia". Higher β_Y implies a stronger effect of "terathanasia", leading to earlier occurrence of SAB/Stillbirth, thus not censored by "livebirth" and so forth.

Remark 4.3.1.1. Note that model (4.4) is in fact saturated. While we might attempt to include the interaction term between d and y in model (4.5), it turns out that the estimation algorithm failed to converge due to too few (7) observed birth defects ($Y = 1$) in the control group ($D = 0$), as seen from Table 4.1.

The counterfactual outcome, by definition, is not observed. In order to estimate the parameters in models (4.4) and (4.5), we use the inverse probability (IP) weights to create a pseudo-randomized sample. This enables us to write down a weighted likelihood based on the observed variables in order to estimate the parameters in these two models, as discussed in [BW07, Page 4]. Specifically, define the stabilized IP weights as

$$w_i^{IP} = \frac{\hat{\mathbb{P}}(D_i=1)D_i}{\hat{\mathbb{P}}(D_i=1|V_i)} + \frac{\hat{\mathbb{P}}(D_i=0)(1-D_i)}{\hat{\mathbb{P}}(D_i=0|V_i)}, \tag{4.7}$$

where $\hat{\mathbb{P}}(D_i = d|V_i)$ is the estimated propensity score using, for example, the package 'twang' [RMM$^+$17] in R; more discussion will be given later.

Before we write down the weighted likelihood based on the complete data, here we introduce additional notation for the "ghost copies" of subject i in order to properly account from left truncation. The idea of "ghost copies" is that for an observed subject i, there are A_i "ghost copies" with the same value of (Y_i, D_i, V_i) that have been truncated out and not observed in the data. It can be seen that A_i follows a geometric distribution with probability $\mathbb{P}(T_i > Q_i|Y_i, D_i, Q_i)$ in the pseudo randomized population. Under the the nonparametric likelihood framework for semiparametric models, the "ghost copy" event times $T_{ij} < Q_i$ ($j = 1, ..., A_i$) are discrete random variables taking values among $t_1, \cdots, t_K$ with probabilities $\mathbb{P}(T_{ij} = t_k|T_{ij} < Q_i, Q_i)$. The fact that the T_{ij}'s are assumed to be discrete is closely related the fact that in a nonparametric likelihood, the baseline hazard is understood as discretized to point masses $\lambda_1, ..., \lambda_K$ at the observed event times $t_1, ..., t_K$ [Joh93, Mur94, Mur95]. See [HCX18] for more details and discussion on this.

Write $\alpha = (\alpha_c, \alpha_D)^\top$, $\beta = (\beta_D, \beta_Y)^\top$, and $\theta = (\alpha^\top, \beta^\top, \lambda_1, ..., \lambda_K)^\top$. The weighted likelihood based on the complete data, including the augmented Y_i's if they are missing, can then be written:

$$L^c_w(\theta) = \prod_{i=1}^{n} \left[\pi_i^{Y_i}(1-\pi_i)^{1-Y_i} \lambda_i(X_i)^{\Delta_i} S_i(X_i) \prod_{j=1}^{A_i} \left\{ \lambda_i(T_{ij}) S_i(T_{ij}) \right\} \right]^{w_i^{IP}},$$

where $\pi_i = \exp(\alpha_c + \alpha_D D_i)/(1 + \exp(\alpha_c + \alpha_D D_i))$, $\lambda_i(t) = \lambda_0(t)\exp(\beta_D D_i + \beta_Y Y_i)$ with $\lambda_0(\cdot)$ equal to the corresponding λ_k, and $S_i(t) = \exp\{-\Lambda(t)\exp(\beta_D D_i + \beta_Y Y_i)\}$ with $\Lambda(\cdot)$ equal to the corresponding cumulative sum of the λ_k's. This yields a weighted complete data log-likelihood:

$$\begin{aligned} l^c_w(\theta) &= \log L^c_w(\theta) \\ &= \sum_{i=1}^{n} w_i^{IP} \Big[Y_i \log \pi_i + (1-Y_i)\log(1-\pi_i) + \Delta_i \log \lambda_i(X_i) + \log S_i(X_i) \\ &\quad + \sum_{j=1}^{A_i} \left\{ \log \lambda_i(T_{ij}) + \log S_i(T_{ij}) \right\} \Big]. \end{aligned} \tag{4.8}$$

We note that (4.8) is a weighted nonparametric likelihood for semiparametric models discussed in [BW07].

4.3.2 ES algorithm

We use the ES algorithm to compute the estimates. The ES algorithm computationally can be achieved by an equivalent weighted EM algorithm, which we present here.

INITIALIZATION

We use "twang" package [RMM$^+$17] in R to compute w_i^{IP} in (4.7) for $1 \leq i \leq n$. By treating all missing outcomes Y as 0 and ignoring left truncation, we run "glm" and "coxph" (in "survival" package [TG00]) functions weighted by w_i^{IP} to initialize. Write the parameter as $\theta^{(0)}$.

E-STEP

By setting $Q(\theta|\theta^{(t)}) = \mathbb{E}_{\theta^{(t)}}[l_w^c(\theta)|O]$, where we use O to represent all the observed variables, we obtain

$$\begin{aligned}Q(\theta|\theta^{(t)}) = \sum_{i=1}^{n} \Big[& w_{1,i}^{\pi} \log \pi_i + w_{0,i}^{\pi} \log(1-\pi_i) \\ & + \sum_{k=1}^{K} w_{i,k,1}^{f} \log f_i(t_k|Y_i=1) + \sum_{k=1}^{K} w_{i,k,0}^{f} \log f_i(t_k|Y_i=0) \\ & + w_{i,1}^{S} \log S_i(X_i|Y_i=1) + w_{i,0}^{S} \log S_i(X_i|Y_i=0) \Big], \end{aligned} \tag{4.9}$$

where

$$w_{1,i}^{\pi} = w_i^{IP}\{O_iY_i + (1-O_i)\mathbb{P}_i^{(t)}(Y_i^{mis}=1)\}, \tag{4.10}$$

$$w_{0,i}^{\pi} = w_i^{IP}\{O_i(1-Y_i) + (1-O_i)\mathbb{P}_i^{(t)}(Y_i^{mis}=0)\}, \tag{4.11}$$

$$\begin{aligned} w_{i,k,1}^{f} \quad = \quad & w_i^{IP}(O_iY_i + (1-O_i)\mathbb{P}_i^{(t)}(Y_i^{mis}=1)) \\ & \cdot[\Delta_i \mathbb{1}(X_i=t_k) + \mathbb{E}_i^{(t)}(A_i|Y_i=1)\mathbb{P}_i^{(t)}(T_{i1}=t_k|Y_i=1)], \end{aligned} \tag{4.12}$$

$$
\begin{aligned}
w^{f}_{i,k,0} &= w^{IP}_i(O_i(1-Y_i)+(1-O_i)\,\mathbb{P}^{(t)}_i(Y^{mis}_i=0)) \\
&\quad \cdot[\Delta_i \mathbb{1}(X_i=t_k)+\mathbb{E}^{(t)}_i(A_i|Y_i=0)\,\mathbb{P}^{(t)}_i(T_{i1}=t_k|Y_i=0)], \qquad (4.13)
\end{aligned}
$$

$$
w^{S}_{i,1}=w^{IP}_i(1-\Delta_i)[O_iY_i+(1-O_i)\,\mathbb{P}^{(t)}_i(Y^{mis}_i=1)]. \qquad (4.14)
$$

$$
w^{S}_{i,0}=w^{IP}_i(1-\Delta_i)[O_i(1-Y_i)+(1-O_i)\,\mathbb{P}^{(t)}_i(Y^{mis}_i=0)]. \qquad (4.15)
$$

The forms of the E-functions appeared in the equations above including $\mathbb{P}^{(t)}_i(Y^{mis}_i=y)$, $\mathbb{E}^{(t)}_i(A_i|Y_i=y)$ and $\mathbb{P}^{(t)}_i(T_{i1}=t_k|Y_i=y)$ are given in the online supplement.

S-STEP

The S-step is achieved by finding the maximizer of $Q(\theta|\theta^{(t)})$. It is easy to see that the Q function $Q(\theta|\theta^{(t)})$ is a sum of two parts with parameters separated. Indeed,

$$
Q(\theta|\theta^{(t)})=l_{glm}(\alpha)+l_{cox}(\beta,\lambda_1,\cdots,\lambda_K), \qquad (4.16)
$$

where

$$
l_{glm}(\alpha)=\sum_{i=1}^{n}\Big[w^{\pi}_{1,i}\log\pi_i+w^{\pi}_{0,i}\log(1-\pi_i)\Big], \qquad (4.17)
$$

and

$$
\begin{aligned}
l_{cox}(\beta,\lambda_1,\cdots,\lambda_K) &= \sum_{i=1}^{n}\Big\{\sum_{k=1}^{K}\Big[w^{f}_{i,k,1}\log f_i(t_k|Y_i=1)+\sum_{k=1}^{K}w^{f}_{i,k,0}\log f_i(t_k|Y_i=0)\Big] \\
&\quad +w^{S}_{i,1}\log S_i(X_i|Y_i=1)+w^{S}_{i,0}\log S_i(X_i|Y_i=0)\Big\}, \qquad (4.18)
\end{aligned}
$$

Consequently, we can solve two weighted regressions separately, which can be done by plugging the corresponding weights into the “weights” arguments in “glm” and “coxph” (in “survival” package [TG00]) functions.

The E-step and S-step are alternately repeated until the overall change of parameters (e.g. L^2 norm) is below a prespecified threshold (0.00001 in our paper). The ES algorithm returns its convergence point, The expectations of the weighted first derivatives and the weighted second derivatives that appeared in the variance estimate are estimated using a Monte Carlo simulation. The details are given in the online supplement A.

4.3.3 Asymptotic properites

At the convergence of the ES algorithm, the solution $\hat{\theta}$ satisfies

Theorem 4.3.3.1. *Under assumptions given in the online supplement A,* $(\hat{\alpha}, \hat{\beta}, \hat{\Lambda})$ *is consistent for* $(\alpha_0, \beta_0, \Lambda_0)$*, that is,*

$$\hat{\alpha} - \alpha_0 \to 0, \quad \hat{\beta} - \beta_0 \to 0, \quad \sup_{t \in [0,\tau]} |\hat{\Lambda}(t) - \Lambda_0(t)| \to 0 \quad a.s.. \tag{4.19}$$

Theorem 4.3.3.2. *Under assumptions given in the online supplement A, the standardized term* $\sqrt{n}\{(\hat{\alpha}, \hat{\beta}, \hat{\Lambda}) - (\alpha_0, \beta_0, \Lambda_0)\}$ *converges weakly to a Gaussian process in* $\tilde{\theta}$*, more specifically,*

$$\sqrt{n}\{(\hat{\alpha}, \hat{\beta}, \hat{\Lambda}) - (\alpha_0, \beta_0, \Lambda_0)\} \to_d -\dot{\Psi}^{-1}(\alpha_0, \beta_0, \Lambda_0)\mathbb{G}(\psi_{\alpha_0, \beta_0, \Lambda_0}), \tag{4.20}$$

where $\dot{\Psi}(\cdot)$ *and* $\psi_{\cdot}$ *are given in the online supplement A,* $\mathbb{G}$ *is a Brownian Bridge.*

The proofs of Theorem 4.3.3.1 and 4.3.3.2 are given in the online supplement A. In particular, we have

$$\sqrt{n}(\hat{\theta} - \theta_0) \longrightarrow_d \mathcal{N}(0, \Omega^{-1}\Sigma\Omega^{-1}), \tag{4.21}$$

where we write the estimate $\hat{\theta} = (\hat{\alpha}, \hat{\beta}, \hat{\lambda}_1, \cdots \hat{\lambda}_K)$, the true value $\theta_0 = (\alpha_0, \beta_0, \lambda_{01}, \cdots \lambda_{0K})$, the estimated jumps $\hat{\lambda}_k = \hat{\Lambda}(t_k) - \hat{\Lambda}(t_{k-1})$ and the true differences $\lambda_{0k} = \Lambda_0(t_k) - \Lambda_0(t_{k-1})$. Denote $S_i(\theta)$ and $H_i(\theta)$ as the score and negative Hessian of the complete data log likelihood contributed by subject i, with respect to θ. By writing

$$U_i(\theta) = w_i^{IP} \mathbb{E}_\theta \Big(\frac{\partial l^c_{w_i}(\theta)}{\partial \theta} \Big| O \Big), \tag{4.22}$$

we can estimate Σ and Ω by

$$\frac{1}{n} \sum_{i=1}^{n} U_i(\theta) U_i(\theta)^\top \Big|_{\theta=\hat{\theta}} = \frac{1}{n} \sum_{i=1}^{n} (\hat{w}_i^{IP})^2 \hat{\mathbb{E}}_\theta [S_i(\theta)|O] \hat{\mathbb{E}}_\theta [S_i(\theta)|O]^\top \Big|_{\theta=\hat{\theta}}, \tag{4.23}$$

and

$$\frac{1}{n} \sum_{i=1}^{n} \frac{\partial U_i(\theta)}{\partial \theta} \Big|_{\theta=\hat{\theta}} \tag{4.24}$$

$$= \frac{1}{n} \sum_{i=1}^{n} \hat{w}_i^{IP} \Big\{ \hat{\mathbb{E}}_\theta [H_i(\theta)|O] - \hat{\mathbb{E}}_\theta [S_i(\theta) S_i(\theta)^\top |O] + \hat{\mathbb{E}}_\theta [S_i(\theta)|O] \hat{\mathbb{E}}_\theta [S_i(\theta)|O]^\top \Big\} \Big|_{\theta=\hat{\theta}} \tag{4.25}$$

We put $\hat{\mathbb{E}}$ here in the variance estimates because we use the Monte Carlo simulation to estimate the asymptotic variance.

We provide the results including the estimates, the confidence intervals, and P-values in Section 4.5.1. Due to the insufficient number of events, we conduct a sensitivity analysis by varying β_Y in model (4.5). This allows us to gradually see the shift of conclusions of ATE based on the strength of terathanasia effect, changing with β_Y.

4.4 Principal Effects

As mentioned earlier SAB is a post exposure variable, therefore as we defined in the notation section for subject i it can take on two potential values $M_i(0)$ and $M_i(1)$. As explained in

[FR02b] a stratified comparison of the Y_i's based on the observed values of the M_i's, is equivalent to comparing $\mathbb{P}(Y_i(1) = 1|M_i(1) = m)$ versus $\mathbb{P}(Y_i(0) = 1|M_i(0) = m)$. Such a comparison is problematic because the set of subjects $\{i : M_i(1) = m\}$ is not the same set of subjects $\{i : M_i(0) = m\}$, as long as the exposure has non-zero effect on SAB. Unfortunately this is the case for etanercept, leading to so-called posttreatment selection bias in the estimated exposure effect [Ros84, RG92, FR02b].

In this section we consider principal stratification which is the stratification with respect to the joint potential values of M, and use $(M(0), M(1))$ to stratify the whole population. The whole population is then divided into:

1. SS: always-survivors, $(M(0), M(1)) = (0, 0)$, are those who will not experience SAB no matter whether treated or not;
2. NS: treatment-survivors, $(M(0), M(1)) = (1, 0)$, are those who will experience SAB only when not treated;
3. SN: control-survivors, $(M(0), M(1)) = (0, 1)$, are those who will experience SAB only when treated;
4. NN: never-survivors, $(M(0), M(1)) = (1, 1)$, are those who will experience SAB no matter treated or not.

Table 4.2 shows the division of the whole population into the above four principal strata.

Due to the very limited number of events in our data, in the following we further make a monotonicity assumption that eliminates the 'control-survivor' stratum, in order to reduce the number of parameters that need to be estimated later. This is the supported by the fact that the estimated β_D under model (4.5) is negative (see Section 4.5); that is, etanercept reduces the risk of SAB. We assume

Assumption 4.4.0.1 (Monotonicity). *$M(1) \leq M(0)$ with probability one.*

It is unknown which principal stratum a subject belongs to. However, certain relationship can be derived between the latent principal strata and the observed group of subjects defined according to (D, M^{obs}), where $M^{obs} = M$ if observed, and $M^{obs} = \ ?$ otherwise. Table 4.5 summarizes the correspondence between the observed groups and the latent strata. For example, those with $(D_i, M_i^{obs}) = (0,0)$ i.e. no SAB events under no treatment, can only belong to the always-survivor stratum SS due to the monotonicity assumption. On the other had, those with $(D_i, M_i^{obs}) = (0,1)$ i.e. having had SAB events under no treatment, can belong to either NS (treatment-survivors) or NN (never-survivors). Missing M_i leads to possibilities of all three strata, etc. Table 4.5 also gives the number of subjects (group size) and the number of birth defects in each observed group for the OTIS etanercept data.

Following [FMPR12] we define G as the latent indicator for the principal strata, which takes values in $\{\text{SS}, \text{NS}, \text{NN}\}$. We assume a multinomial distribution for the principal strata membership:

$$\mathbb{P}(G = g) = \frac{\exp(\gamma_g)}{\sum_{g'} \exp(\gamma_{g'})}, \tag{4.26}$$

where $g \in \{\text{SS}, \text{NS}, \text{NN}\}$, and we treat the group SS as reference, i.e. $\gamma_{\text{SS}} = 0$. Parallel to model (4.4) for the ATE in Section 4.3, the causal estimands are now the principal effects $\alpha_{D,g}$ in each stratum:

$$\mathbb{P}(Y(d) = 1 | G = g) = \frac{\exp(\alpha_{0,g} + \alpha_{D,g} \cdot d)}{1 + \exp(\alpha_{0,g} + \alpha_{D,g} \cdot d)}. \tag{4.27}$$

Though as parsimonious as we intend to be, we allow (4.27) here to be saturated for each stratum since treatment effect for birth defect is of main interest. The parameter $\alpha_{D,\text{SS}}$ in (4.27) is of scientific interest whereas $(\alpha_{0,\text{NS}}, \alpha_{D,\text{NS}}, \alpha_{0,\text{NN}}, \alpha_{D,\text{NN}})$ are treated as nuisance. The reason is twofold. Firstly, we are only interested in the treatment effect of etanercept for the always-survivors since only these individuals' fetus can "survive" through both treatment and control, compared to treatment-survivors and never-survivors. Women and practitioners will be less interested in the treatment effect on malformation if not using etanercept leads to SAB/Stillbirth. Low incidence rate of birth defects implies always-survivors as the main constitution in the whole population. Secondly, the data also supports this setup. By examining Table 4.5, most subjects within always-survivors (SS) are in $O(0,0)$ and $O(1,0)$, who are both complete. Main information about treatment-survivors (NS) is stored in the observed subgroup $O(0,1)$ and $O(1,1)$, where the control arm ($O(0,1)$) is almost all missing. On the other hand, the subgroups $O(0,1)$ and $O(1,1)$ that contain information for never-survivors (NN) are both closely all missing. $O(0,?)$ and $O(1,?)$ are mixtures of three strata and basically do not provide any useful information. These high missing rates prevent us from valid inference for treatment-survivors (NN) and never-survivors (NN). In fact, (4.27) has its outcomes and predictors almost missing among those subjects from $O(0,1)$ and $O(1,1)$, which intuitively cannot converge in practice. To put in another way, we expect that any parameters $(\alpha_{0,\mathrm{NS}}, \alpha_{D,\mathrm{NS}}, \alpha_{0,\mathrm{NN}}, \alpha_{D,\mathrm{NN}})$ shall fit into this data. What we can hope is isolate the influence from the strata treatment-survivors and never-survivors to gain valid information for always-surviros. This isolation is advantageous to the ATE in that we roughly separate observed part and missing part, in the modeling process.

Finally by definition of the principal strata, the potential time to SAB/Stillbirth $T(d)$ is infinity in SS and NS if $d = 1$. That is,

$$\begin{aligned} & \mathbb{P}(T(1) > t|Y(1) = y, G = \mathrm{NS}) = \mathbb{P}(T(0) > t|Y(0) = y, G = \mathrm{NN}) \\ = \ & \mathbb{P}(T(1) > t|Y(1) = y, G = \mathrm{NN}) = 1, \end{aligned}$$

for any $t > 0$. On the other hand, $T(d)$ is finite in NN and NS if $d = 0$. For these latter cases where $T(d) < \infty$ we assume:

$$\begin{aligned} & \mathbb{P}(T(d) > t | Y(d) = y, G = g) \\ = \ & \exp\left\{-\Lambda(t)\exp(\beta_{0,\text{NS}} \cdot (1-d) \cdot \mathbb{1}(g = \text{NS}) + \beta_{D,\text{NN}} \cdot d \cdot \mathbb{1}(g = \text{NN}) + \beta_Y \cdot y)\right\}. \end{aligned} \tag{4.28}$$

Note that only three cases have finite time to SAB: $(d = 0, G = \text{NS})$, $(d = 0, G = \text{NN})$, $(d = 1, G = \text{NN})$. The intercept of $G = \text{NN}$ is in the baseline cumulative hazards $\Lambda(t)$ and thus we only need $\beta_{0,\text{NS}}$ and $\beta_{D,\text{NN}}$. That is,

$$\mathbb{P}(T(0) > t | Y(0) = y, G = \text{NS}) = \exp\left\{-\Lambda(t)\exp(\beta_{0,\text{NS}} + \beta_Y \cdot y)\right\}, \tag{4.29}$$

$$\mathbb{P}(T(0) > t | Y(0) = y, G = \text{NN}) = \exp\{-\Lambda(t)\exp(\beta_Y \cdot y)\}, \tag{4.30}$$

$$\mathbb{P}(T(1) > t | Y(1) = y, G = \text{NN}) = \exp\{-\Lambda(t)\exp(\beta_{D,\text{NN}} + \beta_Y \cdot y)\}. \tag{4.31}$$

Note that the effect of treatment on the occurrence of SAB is now reflected in the definition of the principal strata and (4.26), and no longer modeled in (4.28). This is also part of the attempt to be parsimonious again due to the limited number of events in the data. The parameters involved in (4.28) only influence the time to SAB/Stillbirth not whether it will happen or not. In particular, β_Y does not quantify "terathanasia" as its counterpart in 4.3 does. The model (4.28) is mainly needed to deal with left truncation.

Remark 4.4.0.1. The model for $T(d)$ given $Y(d)$ and G is similar to a cure model [Far82, Far86, KC92, ST00, LY04, HCX18]. The cure indicator now is $\mathbb{1}(g = \text{SS}) + \mathbb{1}(g = \text{NS}) \cdot (1 - d)$, hence the potential time to SAB/Stillbirth is a mixture.

Remark 4.4.0.2. We would like to discuss the closeness and difference between our model to topic known as “truncation by death”. The primary outcome cannot be measured and is not well defined once the subject dies, hence being truncated by death. In our case, whether an embryo is malformed is well defined before birth but can only be measured after birth (“death” here means SAB/Stillbirth). Though it is easily missing when baby is aborted, the probability of observation is not completely zero.

Denote $p_g = \frac{\exp(\gamma_g)}{\sum_{g'} \exp(\gamma_{g'})}$, $\pi_i = \frac{\exp(\alpha_{0,G_i} + \alpha_{D,G_i} D_i)}{1+\exp(\alpha_{0,G_i} + \alpha_{D,G_i} D_i)}$,

$\lambda_i(t) = \lambda(t)\exp(\beta_{0,\mathrm{NS}}(1-D_i)\mathbb{1}(G_i = \mathrm{NS}) + \beta_{D,\mathrm{NN}} D_i \mathbb{1}(G_i = \mathrm{NN}) + \beta_Y Y_i)$, and $S_i(t) = \exp(-\Lambda(t)\exp(\beta_{0,\mathrm{NS}}(1-D_i)\mathbb{1}(G_i = \mathrm{NS}) + \beta_{D,\mathrm{NN}} D_i \mathbb{1}(G_i = \mathrm{NN}) + \beta_Y Y_i))$.

Note that $p_{\mathrm{SS}} + p_{\mathrm{NS}} + p_{\mathrm{NN}} = 1$. Treating G_i, $(1-O_i)Y_i$ and ghost copies as observed, the weighted complete data likelihood is

$$
\begin{aligned}
& L_w^c(\theta) \\
= & \prod_{i \in O(0,0)} \Big[p_{\mathrm{SS}} \pi_i^{Y_i} (1-\pi_i)^{1-Y_i} \Big]^{w_i^{IP} \mathbb{1}(G_i=\mathrm{SS})} \\
& \prod_{i \in O(0,1)} \Big[p_{\mathrm{NS}} \pi_i^{Y_i} (1-\pi_i)^{1-Y_i} \lambda_i(X_i) S_i(X_i) \prod_{j=1}^{A_i} \Big\{ \lambda_i(T_{ij}) S_i(T_{ij}) \Big\} \Big]^{w_i^{IP} \mathbb{1}(G_i=\mathrm{NS})} \\
& \prod_{i \in O(0,1)} \Big[p_{\mathrm{NN}} \pi_i^{Y_i} (1-\pi_i)^{1-Y_i} \lambda_i(X_i) S_i(X_i) \prod_{j=1}^{A_i} \Big\{ \lambda_i(T_{ij}) S_i(T_{ij}) \Big\} \Big]^{w_i^{IP} \mathbb{1}(G_i=\mathrm{NN})} \\
& \prod_{i \in O(0,?)} \Big[p_{\mathrm{SS}} \pi_i^{Y_i} (1-\pi_i)^{1-Y_i} \Big]^{w_i^{IP} \mathbb{1}(G_i=\mathrm{SS})} \\
& \prod_{i \in O(0,?)} \Big[p_{\mathrm{NS}} \pi_i^{Y_i} (1-\pi_i)^{1-Y_i} S_i(X_i) \prod_{j=1}^{A_i} \Big\{ \lambda_i(T_{ij}) S_i(T_{ij}) \Big\} \Big]^{w_i^{IP} \mathbb{1}(G_i=\mathrm{NS})} \\
& \prod_{i \in O(0,?)} \Big[p_{\mathrm{NN}} \pi_i^{Y_i} (1-\pi_i)^{1-Y_i} S_i(X_i) \prod_{j=1}^{A_i} \Big\{ \lambda_i(T_{ij}) S_i(T_{ij}) \Big\} \Big]^{w_i^{IP} \mathbb{1}(G_i=\mathrm{NN})} \\
& \prod_{i \in O(1,0)} \Big[p_{\mathrm{SS}} \pi_i^{Y_i} (1-\pi_i)^{1-Y_i} \Big]^{w_i^{IP} \mathbb{1}(G_i=\mathrm{SS})} \\
& \prod_{i \in O(1,0)} \Big[p_{\mathrm{NS}} \pi_i^{Y_i} (1-\pi_i)^{1-Y_i} \Big]^{w_i^{IP} \mathbb{1}(G_i=\mathrm{NS})} \\
& \prod_{i \in O(1,1)} \Big[p_i \pi_i^{Y_i} (1-\pi_i)^{1-Y_i} \lambda_i(X_i) S_i(X_i) \prod_{j=1}^{A_i} \Big\{ \lambda_i(T_{ij}) S_i(T_{ij}) \Big\} \Big]^{w_i^{IP} \mathbb{1}(G_i=\mathrm{NN})} \\
& \prod_{i \in O(1,?)} \Big[p_{\mathrm{SS}} \pi_i^{Y_i} (1-\pi_i)^{1-Y_i} \Big]^{w_i^{IP} \mathbb{1}(G_i=\mathrm{SS})} \\
& \prod_{i \in O(1,?)} \Big[p_{\mathrm{NS}} \pi_i^{Y_i} (1-\pi_i)^{1-Y_i} S_i(X_i) \prod_{j=1}^{A_i} \Big\{ \lambda_i(T_{ij}) S_i(T_{ij}) \Big\} \Big]^{w_i^{IP} \mathbb{1}(G_i=\mathrm{NS})} \\
& \prod_{i \in O(1,?)} \Big[p_{\mathrm{NN}} \pi_i^{Y_i} (1-\pi_i)^{1-Y_i} S_i(X_i) \prod_{j=1}^{A_i} \Big\{ \lambda_i(T_{ij}) S_i(T_{ij}) \Big\} \Big]^{w_i^{IP} \mathbb{1}(G_i=\mathrm{NN})},
\end{aligned}
$$

which yields the weighted complete data log-likelihood

$$
\begin{aligned}
l_w^c(\theta) &= \log L_w^c(\theta) \\
&= \sum_{i \in O(0,0)} w_i^{IP} \mathbb{1}(G_i = \text{SS}) \Big[\log p_{\text{SS}} + Y_i \log \pi_i + (1 - Y_i) \log(1 - \pi_i) \Big] \\
&+ \sum_{i \in O(0,1)} w_i^{IP} \mathbb{1}(G_i = \text{NS}) \Big[\log p_{\text{NS}} + Y_i \log \pi_i + (1 - Y_i) \log(1 - \pi_i) + \log \lambda_i(X_i) \\
& \quad + \log S_i(X_i) + \sum_{j=1}^{A_i} \Big\{ \log \lambda_i(T_{ij}) + \log S_i(T_{ij}) \Big\} \Big] \\
&+ \sum_{i \in O(0,1)} w_i^{IP} \mathbb{1}(G_i = \text{NN}) \Big[\log p_{\text{NN}} + Y_i \log \pi_i + (1 - Y_i) \log(1 - \pi_i) + \log \lambda_i(X_i) \\
& \quad + \log S_i(X_i) + \sum_{j=1}^{A_i} \Big\{ \log \lambda_i(T_{ij}) + \log S_i(T_{ij}) \Big\} \Big] \\
&+ \sum_{i \in O(0,?)} w_i^{IP} \mathbb{1}(G_i = \text{SS}) \Big[\log p_{\text{SS}} + Y_i \log \pi_i + (1 - Y_i) \log(1 - \pi_i) \Big] \\
&+ \sum_{i \in O(0,?)} w_i^{IP} \mathbb{1}(G_i = \text{NS}) \Big[\log p_{\text{NS}} + Y_i \log \pi_i + (1 - Y_i) \log(1 - \pi_i) + \log S_i(X_i) \\
& \quad + \sum_{j=1}^{A_i} \Big\{ \log \lambda_i(T_{ij}) + \log S_i(T_{ij}) \Big\} \Big] \\
&+ \sum_{i \in O(0,?)} w_i^{IP} \mathbb{1}(G_i = \text{NN}) \Big[\log p_{\text{NN}} + Y_i \log \pi_i + (1 - Y_i) \log(1 - \pi_i) + \log S_i(X_i) \\
& \quad + \sum_{j=1}^{A_i} \Big\{ \log \lambda_i(T_{ij}) + \log S_i(T_{ij}) \Big\} \Big]
\end{aligned}
$$

$$
\begin{aligned}
&+ \sum_{i\in O(1,0)} w_i^{IP}\mathbb{1}(G_i=\mathrm{SS})\Big[\log p_{\mathrm{SS}} + Y_i\log\pi_i + (1-Y_i)\log(1-\pi_i)\Big]\\
&+ \sum_{i\in O(1,0)} w_i^{IP}\mathbb{1}(G_i=\mathrm{NS})\Big[\log p_{\mathrm{NS}} + Y_i\log\pi_i + (1-Y_i)\log(1-\pi_i)\Big]\\
&+ \sum_{i\in O(1,1)} w_i^{IP}\mathbb{1}(G_i=\mathrm{NN})\Big[\log p_{\mathrm{NN}} + Y_i\log\pi_i + (1-Y_i)\log(1-\pi_i) + \log\lambda_i(X_i)\\
&\quad +\log S_i(X_i) + \sum_{j=1}^{A_i}\Big\{\log\lambda_i(T_{ij}) + \log S_i(T_{ij})\Big\}\Big]\\
&+ \sum_{i\in O(1,?)} w_i^{IP}\mathbb{1}(G_i=\mathrm{SS})\Big[\log p_{\mathrm{SS}} + Y_i\log\pi_i + (1-Y_i)\log(1-\pi_i)\Big]\\
&+ \sum_{i\in O(1,?)} w_i^{IP}\mathbb{1}(G_i=\mathrm{NS})\Big[\log p_{\mathrm{NS}} + Y_i\log\pi_i + (1-Y_i)\log(1-\pi_i) + \log S_i(X_i)\\
&\quad +\sum_{j=1}^{A_i}\Big\{\log\lambda_i(T_{ij}) + \log S_i(T_{ij})\Big\}\Big]\\
&+ \sum_{i\in O(1,?)} w_i^{IP}\mathbb{1}(G_i=\mathrm{NN})\Big[\log p_{\mathrm{NN}} + Y_i\log\pi_i + (1-Y_i)\log(1-\pi_i) + \log S_i(X_i)\\
&\quad +\sum_{j=1}^{A_i}\Big\{\log\lambda_i(T_{ij}) + \log S_i(T_{ij})\Big\}\Big].
\end{aligned}
$$

Therefore we obtain the Q-function

$$
\begin{aligned}
& Q(\theta|\theta^{(t)})\\
= & \sum_{i=1}^{N}\sum_{g}\Big[w^{p}_{g,i}\log p_i^{g} + w^{\pi}_{1,g,i}\log\pi_i(G_i=g) + w^{\pi}_{0,g,i}\log(1-\pi_i(G_i=g))\\
& +\sum_{k=1}^{K} w^{f}_{g,1,i,k}\log f_i(t_k|G_i=g,Y_i=1) + \sum_{k=1}^{K} w^{f}_{g,0,i,k}\log f_i(t_k|G_i=g,Y_i=0)\\
& +w^{S}_{g,i,1}\log S_i(X_i|G_i=g,Y_i=1) + w^{S}_{g,0,i}\log S_i(X_i|G_i=g,Y_i=0)\Big], \qquad (4.32)
\end{aligned}
$$

where

$$
w^{p}_{g,i} = w_i^{IP}\,\mathbb{P}_i^{(t)}(G_i=g), \qquad (4.33)
$$

$$w^{\pi}_{1,g,i} = w_i^{IP}\,\mathbb{P}_i^{(t)}(G_i = g)\{O_iY_i + (1-O_i)\,\mathbb{P}_i^{(t)}(Y_i = 1|G_i = g)\}, \tag{4.34}$$

$$w^{\pi}_{0,g,i} = w_i^{IP}\,\mathbb{P}_i^{(t)}(G_i = g)\{O_i(1-Y_i) + (1-O_i)\,\mathbb{P}_i^{(t)}(Y_i = 0|G_i = g)\}, \tag{4.35}$$

$$\begin{aligned} w^{f}_{g,1,i,k} &= w_i^{IP}\,\mathbb{P}_i^{(t)}(G_i = g)[O_iY_i + (1-O_i)\,\mathbb{P}_i^{(t)}(Y_i = 1|G_i = g)][\Delta_i \mathbb{1}(X_i = t_k) \\ &\quad + \mathbb{E}_i^{(t)}(A_i|G_i = g, Y_i = 1)\,\mathbb{P}_i^{(t)}(T_{i1} = t_k|G_i = g, Y_i = 1)], \end{aligned} \tag{4.36}$$

$$\begin{aligned} w^{f}_{g,0,i,k} &= w_i^{IP}\,\mathbb{P}_i^{(t)}(G_i = g)[O_i(1-Y_i) + (1-O_i)\,\mathbb{P}_i^{(t)}(Y_i = 0|G_i = g)][\Delta_i \mathbb{1}(X_i = t_k) \\ &\quad + \mathbb{E}_i^{(t)}(A_i|G_i = g, Y_i = 0)\,\mathbb{P}_i^{(t)}(T_{i1} = t_k|G_i = g, Y_i = 0)], \end{aligned} \tag{4.37}$$

$$w^{S}_{g,1,i} = w_i^{IP}\,\mathbb{1}(M_i = \text{“?''})\,\mathbb{P}_i^{(t)}(G_i = g)[O_iY_i + (1-O_i)\,\mathbb{P}_i^{(t)}(Y_i = 1|G_i = g)]. \tag{4.38}$$

$$w^{S}_{g,0,i} = w_i^{IP}\,\mathbb{1}(M_i = \text{“?''})\,\mathbb{P}_i^{(t)}(G_i = g)[O_i(1-Y_i) + (1-O_i)\,\mathbb{P}_i^{(t)}(Y_i = 0|G_i = g)]. \tag{4.39}$$

S-STEP

The S-step is achieved by finding the maximizer of $Q(\theta|\theta^{(t)})$. It is easy to check that the Q function $Q(\theta|\theta^{(t)})$ is a sum of three parts with parameterseparated. Indeed,

$$Q(\theta|\theta^{(t)}) = l_{multi}(\gamma) + l_{glm}(\alpha) + l_{cox}(\beta, \lambda_1, \cdots, \lambda_K), \tag{4.40}$$

where

$$l_{multi}(\gamma) = \sum_{i=1}^{n} \sum_{g} w_{g,i}^{p} \log p_i^g, \tag{4.41}$$

$$l_{glm}(\alpha) = \sum_{i=1}^{n} \sum_{g} \Big[w_{1,g,i}^{\pi} \log \pi_i(G_i = g) + w_{0,g,i}^{\pi} \log(1 - \pi_i(G_i = g)) \Big], \tag{4.42}$$

and

$$\begin{aligned} & l_{cox}(\beta, \lambda_1, \cdots, \lambda_K) & (4.43) \\ = \; & \sum_{i=1}^{n} \sum_{g} \Big[\sum_{k=1}^{K} w_{g,1,i,k}^{f} \log f_i(t_k | G_i = g, Y_i = 1) & \\ & + \sum_{k=1}^{K} w_{g,0,i,k}^{f} \log f_i(t_k | G_i = g, Y_i = 0) & \\ & + w_{g,i,1}^{S} \log S_i(X_i | G_i = g, Y_i = 1) + w_{g,0,i}^{S} \log S_i(X_i | G_i = g, Y_i = 0) \Big]. & (4.44) \end{aligned}$$

Therefore, the S-step can be achieved by solving three weighted regression problems, which can be easily done with the existing software.

The E-step and S-step are alternately repeated until the overall change of parameters (e.g. L^2 norm) is below a prespecified threshold 0.0001. We use the final result as our estimator.

Analogous to Section 4.3.3, we have

Theorem 4.4.0.1. *Under assumptions given in the online supplement A,* $(\hat{\gamma}, \hat{\alpha}, \hat{\beta}, \hat{\Lambda}(t))$ *is consistent for* $(\gamma_0, \alpha_0, \beta_0, \Lambda_0(t))$ *uniformly, that is,*

$$\hat{\gamma} - \gamma_0 \to 0, \quad \hat{\alpha} - \alpha_0 \to 0, \quad \hat{\beta} - \beta_0 \to 0, \quad \sup_{t \in [0,\tau]} |\hat{\Lambda}(t) - \Lambda_0(t)| \to 0 \quad a.s.. \tag{4.45}$$

Theorem 4.4.0.2. *Under assumptions given in the online supplement A, the standardized term*

$\sqrt{n}(\hat{\theta} - \theta_0)$ *converges weakly to a Gaussian process in* $\tilde{\Theta}$*, more specifically,*

$$\sqrt{n}(\hat{\theta} - \theta_0) \to_d -\dot{\Psi}^{-1}(\theta_0)\mathbb{G}(\psi_{\theta_0}), \tag{4.46}$$

where $\dot{\Psi}$ *and* ψ_{θ_0} *are given in the online supplement A,* $\mathbb{G}$ *is a Brownian Bridge.*

The proofs of Theorem 4.4.0.1 and 4.4.0.2 are given in the online supplement A. By writing $\hat{\theta} = (\hat{\gamma}, \hat{\alpha}, \hat{\beta}, \hat{\lambda}_1, \cdots \hat{\lambda}_K)$, the true value $\theta_0 = (\gamma_0, \alpha_0, \beta_0, \lambda_{01}, \cdots \lambda_{0K})$, where the estimated jumps $\hat{\lambda}_k = \hat{\Lambda}(t_k) - \hat{\Lambda}(t_{k-1})$ and the true differences $\lambda_{0k} = \Lambda_0(t_k) - \Lambda_0(t_{k-1})$, we have

$$\sqrt{n}(\hat{\theta} - \theta_0) \longrightarrow_d \mathcal{N}(0, \Omega^{-1}\Sigma\Omega^{-1}). \tag{4.47}$$

The variance estimates are the same as (4.23) and (4.24) except changes in the score $S_i(\theta)$ and negative Hessian $H_i(\theta)$, which are given in the online supplement A.

4.5 Birth Defect Data Analysis

In this analysis, we consider the confounders identified in the original analysis that was submitted the US FDA, which included asthma (yes/no), maternal height and referral source. A key step in identifying the confounders was to examine whether or not the relationship between the exposure and the outcome was altered by including or excluding the covariate in question; more investigation on confounder selection for birth defect studies was carried out in [XHC18] The purpose of the first part of the analysis here is to investigate the robustness of the original analysis results with respect to the handling of missing major birth defect outcomes. The distribution of the covariates are summarized in Table 4.3.

Figure 4.3 shows the distribution of Q, i.e. gestational age at enrollment for women in the etanercept study, as well as the Kaplan-Meier (KM) estimate of the time to SAB/stillbirth distribution accounting for left truncation.

4.5.1 Average treatment effect

We used the R package "twang" [RMM+17] to estimate the propensity score $\hat{\mathbb{P}}(D_i = 1|V_i)$ with the above identified confounders, which was then used to form the stabilized weights as in (4.7). The convergence of the ES algorithm was achieved when the overall change (in L^2 norm) of the parameters values between two consecutive steps was less than 0.00001. The estimates and their standard errors etc. are presented in Table 4.4.

From the table we see that the estimated $\beta_Y > 0$ is consistent with the known "terathanasia" theory mentioned before. However, it has a very wide 95% confidence interval (CI), which is perhaps not surprising as we only had two observed major birth defect outcome Y_i's among the 27 SAB/stillbirth events. The results otherwise show that etanercept has a significant effect in increasing major birth defects, with an causal odds ratio close to 3 which is consistent with the original analysis. In addition, etanercept has a negative albeit not significant effect on SAB/Stillbirth, reducing the hazard to less than half of the unexposed.

Due to the little confidence we have in the estimated β_Y as reflected in its wide CI, we further conduct a sensitivity analysis to examine the robustness of our conclusion about the ATE α_D with respect to the value of β_Y, which affects the probability of major birth defects among those with missing values. Following the theory of terathanasia, we restrict β_Y to be non-negative, and vary it on the interval $[0,5]$; note that 5 would be considered an extremely large log hazard ratio. Figure 4.4 shows the posterior probabilities of the 27 subjects (ordered by their SAB event times) with missing major birth defect, which increase as β_Y becomes larger.

Figure 4.5 shows the sensitivity analysis results. Note that as β_Y increases past 1.4, α_D becomes non-significantly different from zero at 0.05 level two-sided. This makes sense because as more missing major birth defect outcomes become 'yes', the rates of major birth defect between the exposed and unexposed groups become less differentiated. Meanwhile, β_D becomes significantly less than zero, implying that exposure to etanercept reduces the risk of spontaneous abortion. Such a mechanism allows more malformed fetus in the exposed group, which would have had a high chance of being spontaneously aborted had the women not been exposed to etanercept, to develop into live born infants.

4.5.2 Principal effects

Under the principal strata models (4.26) - (4.28), more parameters need to be estimated than that when estimating the ATE. The ES algorithm failed to converge with so few events and the missing major birth defects.

As we state in Section 4.4, most missing subjects come from subgroups $O(0,1)$ and $O(1,1)$, which consist of (NS, NN) and NN, respectively. Consequently, both outcomes and predictors (Y and G) are almost missing within those subgroups. Regressing on NS and NN is impossible. Owing to the fact that the missing outcomes in $O(0,1)$ and $O(1,1)$ are mainly controlled by $(\alpha_{0,\text{NS}}, \alpha_{0,\text{NN}}, \alpha_{D,\text{NN}}$, it seems natural to offset them and investigate how $\hat{\alpha}_{D,\text{SS}}$ reacts. To this end, we allow $\alpha_{0,\text{NS}}$, $\alpha_{0,\text{NN}}$, $\alpha_{D,\text{NN}})$ to range in $\{-2,-1,0\}$, $\{-2,-1,0\}$, $\{-2,-1,0,1,2\}$, and examine how $\hat{\alpha}_{D,\text{SS}}$ changes accordingly. The non-positive ranges of $(\alpha_{0,\text{NS}}, \alpha_{0,\text{NN}})$ are chosen to respect the birth defect rate around 10% in the whole population. The sensitivity here does not only serve for the purpose of offsetting some unknown parameters for us to estimate parameters of interest, but also plays a similar role as in the sensitivity analysis in Section 4.5.1. Indeed, though directly tuning the effect of "terathanasia" in Section 4.5.1, this eventually leads to our belief in the constitution of real birth defects among those missing subjects, see Figure 4.4. Under principal stratification, we cannot directly tune "terathanasia", as explained in Section 4.4. Nonetheless, we are able to vary probabilities of birth defects among missing subjects, which is achieved by varying $(\alpha_{0,\text{NS}}, \alpha_{0,\text{NN}}, \alpha_{D,\text{NN}})$ also. By varying those parameters, we mainly target the birth defects probabilities of all missing subjects in $O(0,1)$, $O(1,1)$ and $O(1,?)$. More discussion on tuning "terathanasia" in the context of principal stratification can be found in Section 4.6.

The resulted $\hat{\alpha}_{D,\text{SS}}$ and corresponding P-values with respect to different offset values are given in Figure 4.6.

From Figure 4.6, it is seen that the principal effect of etanercept on major birth defect among the always-survivors remains significant with an estimated OR over 3. This implies that if a pregnant diseased woman will not experience SAB regardless of the exposure status, etanercept exposure increases her risk of major birth defects.

As a secondary interest, we estimate the treatment effect of etanercept on SAB/Stillbirth. In particular, we report the estimated log of odds ratio

$$\log\left\{\frac{\hat{\mathbb{P}}(M(1)=1)/\hat{\mathbb{P}}(M(1)=0)}{\hat{\mathbb{P}}(M(0)=1)/\hat{\mathbb{P}}(M(0)=0)}\right\}. \tag{4.48}$$

Since $\mathbb{P}(M(1)=1, M(0)=0)=0$ by monotonicity assumption, the odds ratio in (4.48) can be rewritten as,

$$\begin{aligned} & \frac{\hat{\mathbb{P}}(M(1)=1)/\hat{\mathbb{P}}(M(1)=0)}{\hat{\mathbb{P}}(M(0)=1)/\hat{\mathbb{P}}(M(0)=0)} && (4.49) \\ = & \frac{\hat{\mathbb{P}}(M(1)=1,M(0)=1)\big/[\hat{\mathbb{P}}(M(1)=0,M(0)=1)+\hat{\mathbb{P}}(M(1)=0,M(0)=0)]}{[\hat{\mathbb{P}}(M(1)=1,M(0)=1)+\hat{\mathbb{P}}(M(1)=0,M(0)=1)]\big/\hat{\mathbb{P}}(M(1)=0,M(0)=0)} \\ = & \frac{\exp(\hat{\gamma}_{\mathrm{NN}})}{[\exp(\hat{\gamma}_{\mathrm{NS}})+\exp(\hat{\gamma}_{\mathrm{NN}})][1+\exp(\hat{\gamma}_{\mathrm{NS}})]}. && (4.50) \end{aligned}$$

We apply the delta method to get a variance estimate of (4.48). This enables us to draw inference without refitting another model. The result is presented in 4.7.

As one can tell, in most cases, etanercept has a non-significant negative treatment effect on SAB/Stillbirth. Especially, the significance is explained away as $\alpha_{0,\mathrm{NN}}$ becomes smaller, $\alpha_{D,\mathrm{NN}}$ becomes larger and $\alpha_{0,\mathrm{NS}}$ becomes larger. This can be explained by Figure 4.8 given below. One can see as $\alpha_{0,\mathrm{NN}}$ becomes smaller, $\alpha_{D,\mathrm{NN}}$ becomes larger and $\alpha_{0,\mathrm{NS}}$ becomes larger, $\hat{\mathbb{P}}(G=\mathrm{NS})$ decreases and $\hat{\mathbb{P}}(G=\mathrm{NN})$ increases, resulting in an increase of the treatment effect seen by (4.50). One explanation is, the amount of belief in the proportion of birth defect in $O(0,1)$ is explained more by $\alpha_{0,\mathrm{NS}}$ but stays invariant in the sensitivity analysis, therefore pushing the stratum NS to shrink.

We are also curious about the estimated principal strata membership for those subjects in $O(0,1)$ (above) and $O(0,1)$, as in Table 4.5. We choose two extreme cases and one middle case ($(\alpha_{0,\text{NS}}, \alpha_{0,\text{NN}}, \alpha_{D,\text{NN}})$ are equal to $(-2,-2,-2)$, $(-1,-1,0)$ and $(0,0,2)$) to present. It suffices to just give $\hat{\mathbb{P}}(G_i = \text{NS}|O)$ in 4.9 in $O(0,1)$ (above) and $O(0,1)$ since they are just mixtures of two strata. The rest 6 subjects in $O(0,?)$ and $O(1,?)$ is presented in the online supplement A, for simplicity.

As the figures reflect, the pattern of $\hat{\mathbb{P}}(G_i = \text{NS}|O)$ stays unchanged with respect to the sensitivity parameters. We basically believe the $O(0,1)$ is equally divided into NS and NN, with 5% of $O(1,0)$ originating from NS.

Finally, we also present $\hat{\mathbb{P}}(Y = 1|O)$ in Figure 4.10, compared with Figure 4.4. We also show the result based on two extreme cases and one middle case ($(\alpha_{0,\text{NS}}, \alpha_{0,\text{NN}}, \alpha_{D,\text{NN}})$ are equal to $(-2,-2,-2)$, $(-1,-1,0)$ and $(0,0,2)$). The pattern looks clearly different but is due to difference in modeling assumptions. $\hat{\mathbb{P}}(Y = 1|O)$ goes higher for each subject as $(\alpha_{0,\text{NS}}, \alpha_{0,\text{NN}}, \alpha_{D,\text{NN}})$ increases. The average $\hat{\mathbb{P}}(Y = 1|O)$ for the control arm is around 10% as we compute, to respect the population birth defect rate, which validates the choice of the sensitivity range. As we expected, the posterior probability increase as three parameters all increase. We see that subjects within the treatment arm has a higher probability in blue lines, since $\alpha_{D,\text{NN}} = 2$.

Note that there is a difference between the sensitivity analysis here and that in Section 4.5.1. We both examine how the treatment effect of etanercept changes with the change in the sensitivity parameter. However, in Section 4.5.1, because of the high variability of $\hat{\beta}_Y$ (β_Y in (4.5)), we offset β_Y, which also reflects the effect of terathanasia in theory. Here, we simply offset the other parameters because of data shortage. More discussion on how to adjust the effect of terathanasia under principal stratification can be found in Section 4.6. One simple reason here is, otherwise, we have to offset too many parameters.

4.6 Discussion

In this paper we have considered prospective pregnancy cohort studies where spontaneous abortion often results in unknown major birth defect outcomes. By convention of coding in the database a pregnancy is recorded as no birth defects unless one is found. In the meanwhile established terathanasia theory tells us that a malformed fetus have an increased chance of being aborted early(?). By modeling the missing major birth defect mechanism after the terathanasia theory, we are able to turn the MNAR problem of major birth defect into an MAR setting by including information on the spontaneous abort outcome.

Missing outcome is the leading problem in this study. Other possible techniques exist for handling other than EM algorithm that we adopt for ATE [LR19], for instance, multiple imputation [Rub96, Rub04], which also enables a complete data analysis. However, we note that this still fails in the case of estimating principal effects. A frequentist multiple imputation [WR98, RW00] requires an initial estimate based on the observed data with a variance estimate, which we do not have. Apart from the real data issue, multiple imputation shall only be used with missing rate lower than 40% (90% for some PS in our case), otherwise leading to a severe bias, as pointed out in [WR98, RW00].

Another feature of these prospective pregnancy cohort studies is left truncation, which leads to selection bias as early aborted pregnancies tend not to be captures in the study. In the etanercept data set, the earliest gestational age at study enrollment is 3.9 weeks, therefore all conclusions only apply to the population of pregnancies that have lasted at least 3.9 weeks. Given this condition, survival analysis method is able to properly handle any remaining bias that might otherwise result from the differential gestational ages at enrollment. This applies to both major birth defects as well as time to spontaneous abortion as both outcomes are analyzed together.

Left truncation, however, does reduce the number of observed SAB events in the data. Major birth defects, on the other hand, is known to be a rare outcome in the population.[2] Both

[2] https://www.cdc.gov/ncbddd/birthdefects/macdp.html

lead to very limited number of events, given the moderate sample sizes of this type of prospective pregnancy cohort studies. In alleviating the problem we have carried out sensitivity analysis with respect to the missing data model parameter β_Y, which can also be seen as a special case of Bayesian analysis.

A second objective of this paper is to properly handle the post exposure variable live birth versus SAB. We found principal strata to be a useful framework for this.

We notice that there are two ways of principally stratifying the population: one adopted by us in Section 4.4 and the other by further introducing $(M(d,y), T(d,y)$, where we will have $M(d) = Y(d)M(d,1) + (1-Y(d))M(d,0)$, $T(d) = Y(d)T(d,1) + (1-Y(d))T(d,0)$. Then we can investigate the principal effects within the 16 subpopulations stratified by $M(d,y)$. Introducing more strata enables us to explore more quantities of scientific interest at the expense of exponentially increased number of parameters. Note that stratifying by $M(d,y)$ enables us to directly tune the effect of "terathanasia" under principal stratification, for instance, by tuning the relative size of the stratum $(M(1,1), M(1,0), M(0,1), M(0,0)) = (1,0,1,0)$ compared to the stratum $(M(1,1), M(1,0), M(0,1), M(0,0)) = (0,0,0,0)$. We choose our framework for the sake of parsimony.

Although we do not explore on this it in this project, one should note that our procedure is ready for it. If we put assumptions on the regression model, we could get an estimate for the heterogeneous treatment effect conditional on any level v (or a subset L of V).

To handle observational nature, we adopt a propensity score-based method. As discussed about heterogeneous treatment effect, the regression-based method can serve as an alternative, so are doubly robust estimators for causal odds ratio [Che07, TTRR09, ZVVS19], taking account into the missingness and left truncation.

The principal effects can be seen as lying between ATE and CATE (strictly speaking, this holds if we include G into V). The effect of the treatment D on (M,T) can be decomposed into a direct effect and an indirect effect through Y, see Figure 4.1 and Figure 4.2. As noted by [Pea01], the controlled direct and the controlled indirect effect are particularly relevant for the policy-making whereas the natural direct and the natural indirect effect are more useful for understanding the underlying mechanism by which the exposure operates. We write“natural principal effect” or “controlled principal effect” here when the population is stratified by $M(d)$ or $M(d,y)$, respectively. People can instead estimate the principal effects when the population is stratified by $M(d,y)=0$ for $d,y=0,1$. One can have a better interpretation of the mechanism of both d and y on M. However, we should note that the number of strata increases exponentially with the number of distinct complications.

In our case, SAB and TAB can act as competing risks. Removing the selection bias induced by the left truncation requires knowledge on the overall survival function. This, however, requires a good estimate of the competing event distribution. Our data is inadequate to deal with this since we only observe 4 TAB in total. To fix this problem, we assume TAB is non-informative, which can be treated as independent censoring.

Table 4.1: Missing major birth defects by pregnancy outcomes

	Exposed ($n = 336$)			Unexposed ($n = 158$)		
Birth Defect	Yes	No	Missing	Yes	No	Missing
Live Birth	30	287		5	139	
SAB			13	1		12
Stillbirth		1				
TAB	3			1		
LTFU			2			

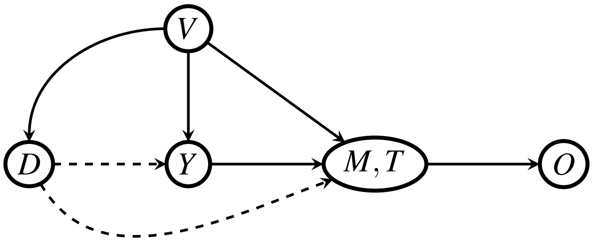

Figure 4.1: The causal directed acyclic graph

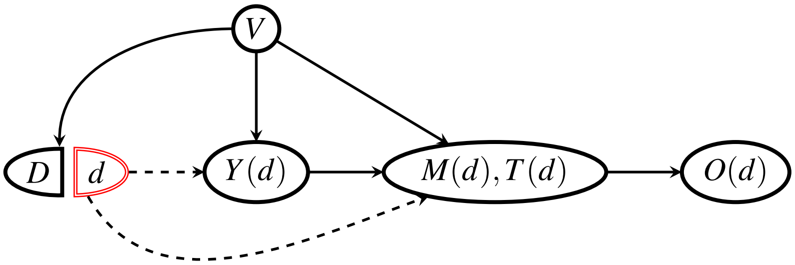

Figure 4.2: The single world intervention graph

Table 4.2: Division of the whole population into four principal strata

		M(1)	
		0	1
M(0)	0	Always-survivors (SS)	Control-survivors (SN)
	1	Treatment-survivors (NS)	Never-survivors (NN)

Table 4.4: Estimated parameters including the ATE from the etanercept data

	Estimate	Standard Error	exp(Estimate)	95% CI of OR/HR	P-value
α_c	-3.336	0.443	0.035	(0.014, 0.084)	<0.001
α_D	1.093	0.489	2.983	(1.144, 7.779)	0.025
β_D	-0.801	0.498	0.448	(0.169, 1.191)	0.107
β_Y	0.485	1.935	1.624	(0.036, 72.067)	0.802

Table 4.3: Distribution of the identified confounders in etanercept study: mean (SD) or *n* (%).

Confounders	Exposed ($n = 336$)	Unexposed ($n = 158$)
Asthma	45 (13.4%)	32 (20.3%)
Maternal Height (cm)	165 (6.98)	167 (7.01)
Referral Source:*		
Type I	26 (7.7%)	65 (41.1%)
Type II	199 (59.2%)	52 (32.9%)
Type III	111 (33.1%)	41 (26.0%)

*I: TIS; II: Pharmaceutical Company/Sponsor, Healthcare Professional; III: Patient Support Group, Internet, or Other.

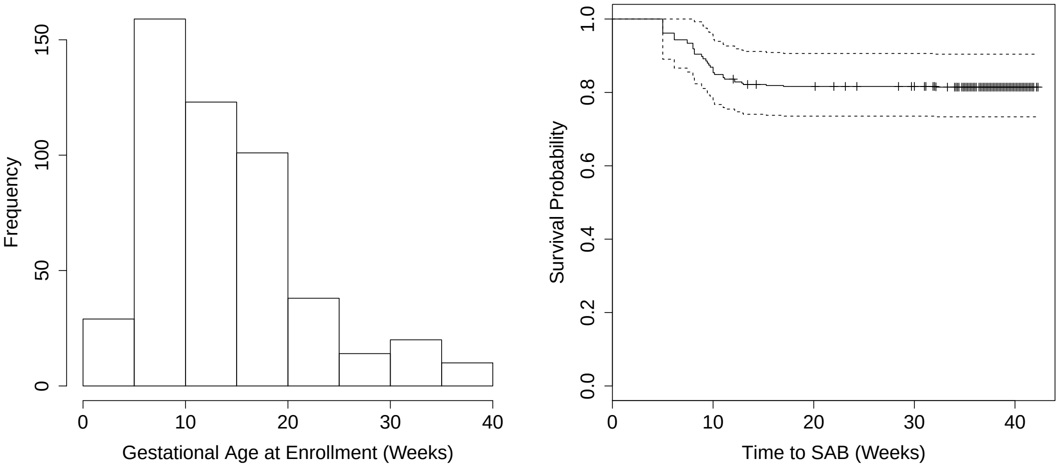

Figure 4.3: Distribution of gestational age at study enrollment (left) and KM estimate for time to SAB/stillbirth (right)

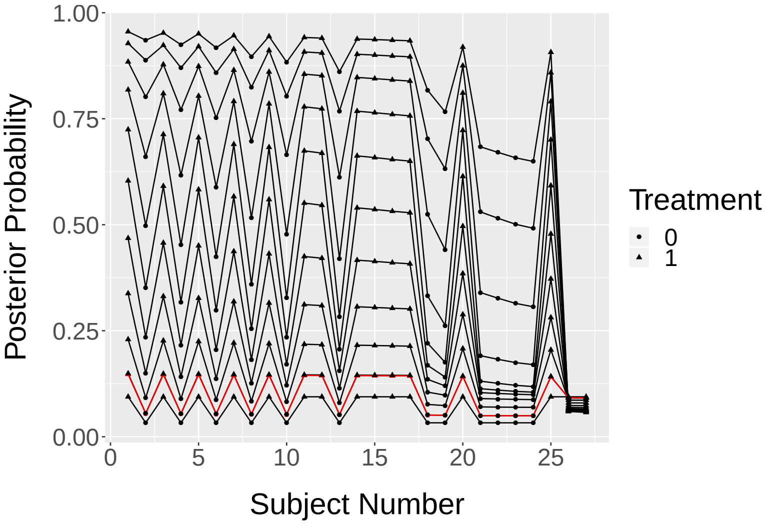

Figure 4.4: Posterior probabilities of major birth defect as a function of β_Y for subjects with missing major birth defect outcomes, ordered by censored event time; red line is when β_Y is set at the estimated β_Y value, black lines from bottom to top correspond to β_Y set from 0 to 5 with increment of 0.5. Last two subjects are censored among 27 missing subjects.

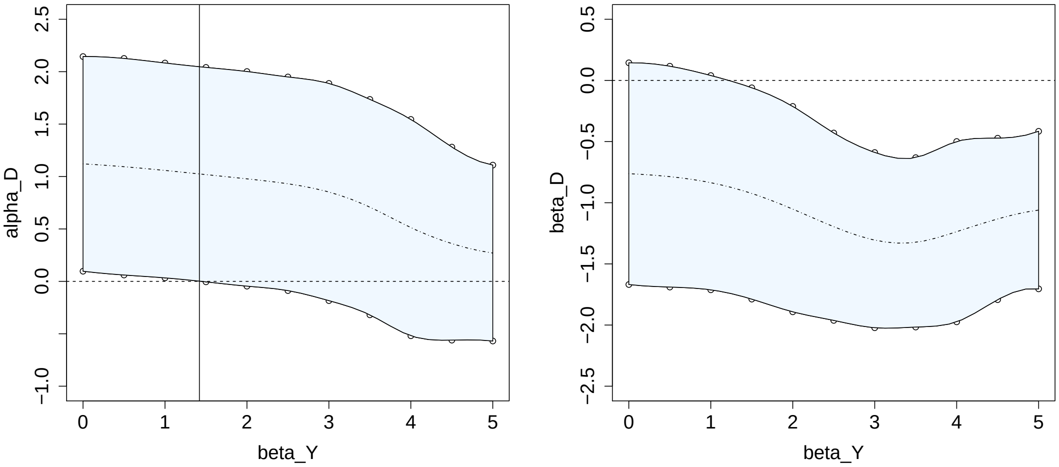

Figure 4.5: 95% confidence intervals of ATE (left) and β_D (right) for given values of β_Y; the dashed lines denote point estimates.

Table 4.5: Correspondence between the observed $O(D, M^{obs})$ groups and the latent principal strata

$O(D, M^{obs})$	Size	Birth Defects	Missing Defects	Principal Strata
$O(0,0)$	144	5	0	SS
$O(0,1)$	13	1	12	NS, NN
$O(0,?)$	1	1	0	SS, NS, NN
$O(1,0)$	317	30	0	SS, NS
$O(1,1)$	14	0	13	SS
$O(1,?)$	5	3	2	SS, NS, NN

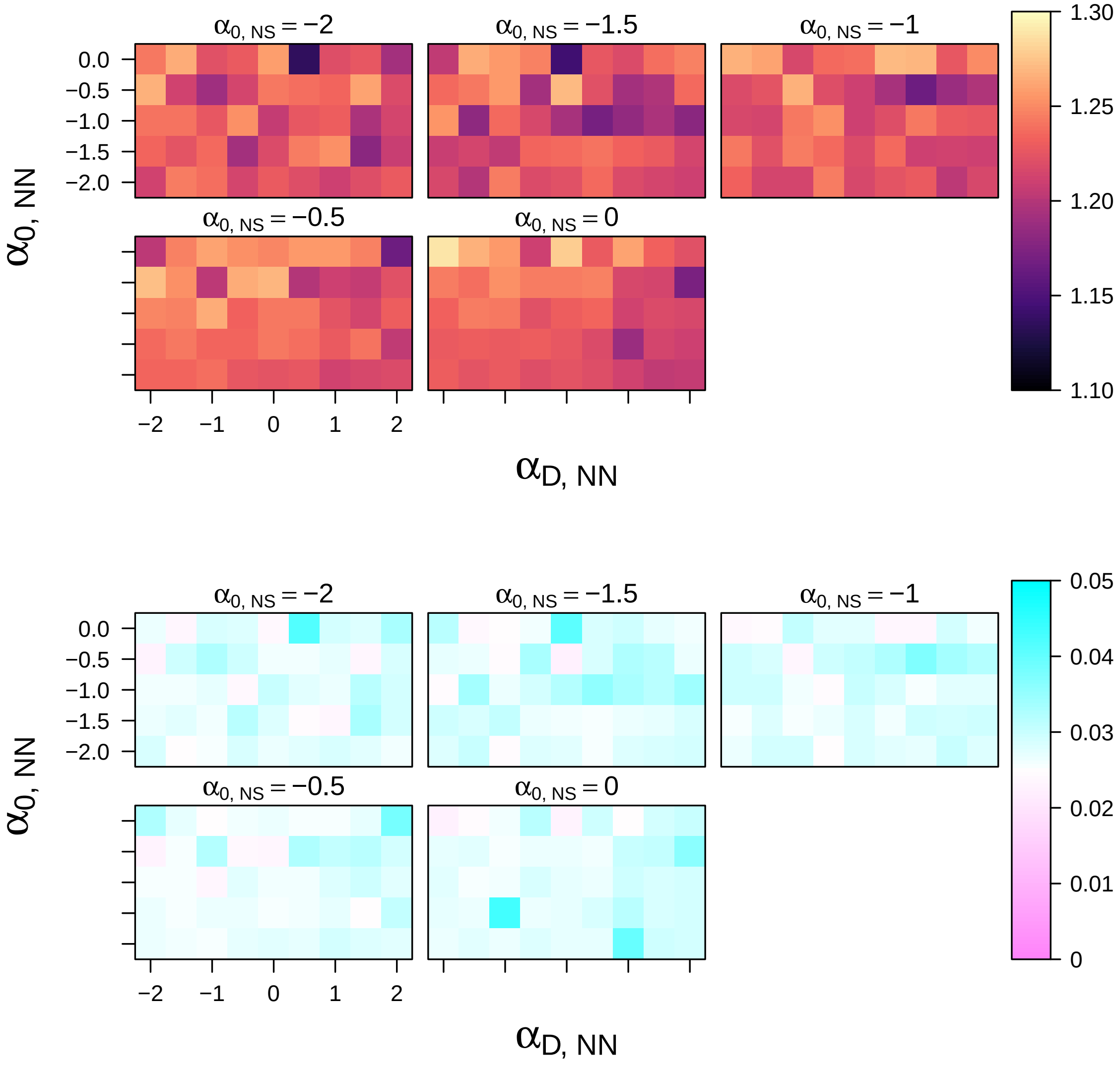

Figure 4.6: Levelplots of estimated $\hat{\alpha}_{D,\mathrm{SS}}$ (above) and P-values (below), with the corresponding offset $\alpha_{0,\mathrm{NS}}$ on the subtitles. The offset parameters $\alpha_{0,\mathrm{NN}}$ and $\alpha_{D,\mathrm{NN}}$ range in $\{-2,-1,0\}$, $\{-2,-1,0,1,2\}$. The treatment effect within always-survivors stays significantly positive at significance level 0.05.

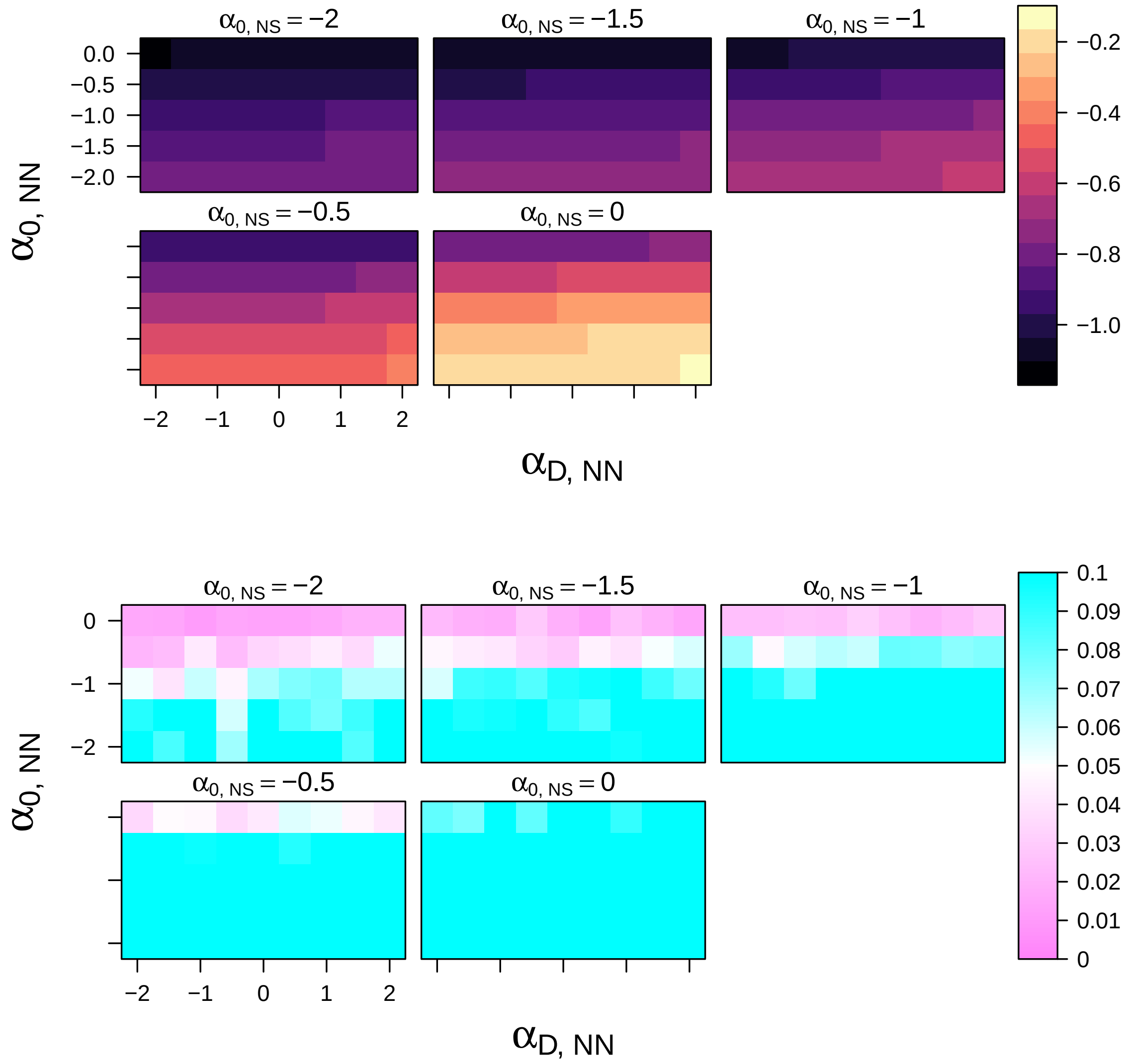

Figure 4.7: Levelplots of estimated causal effect of etanercept on SAB/Stillbirth (above) and its P-values (below), with the corresponding offset $\alpha_{0,\text{NS}}$ on the subtitles. The offset parameters $\alpha_{0,\text{NN}}$ and $\alpha_{D,\text{NN}}$ range in $\{-2,-1,0\}$, $\{-2,-1,0,1,2\}$.

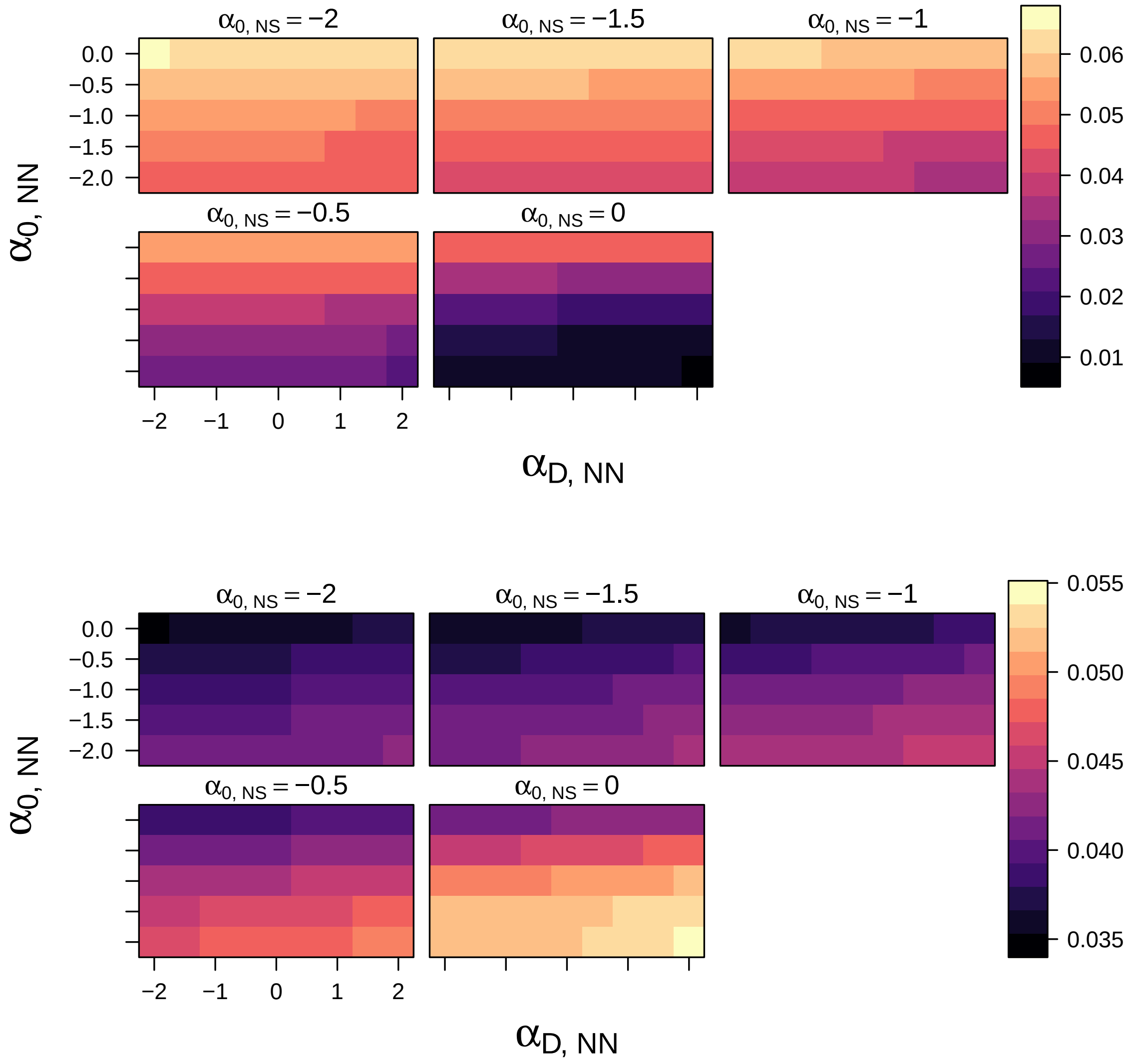

Figure 4.8: Levelplots of estimated $\hat{\mathbb{P}}(G = \mathrm{NS})$ (above) and $\hat{\mathbb{P}}(G = \mathrm{NN})$ (below), with the corresponding offset $\alpha_{0,\mathrm{NS}}$ on the subtitles. The offset parameters $\alpha_{0,\mathrm{NN}}$ and $\alpha_{D,\mathrm{NN}}$ range in $\{-2,-1,0\}$, $\{-2,-1,0,1,2\}$.

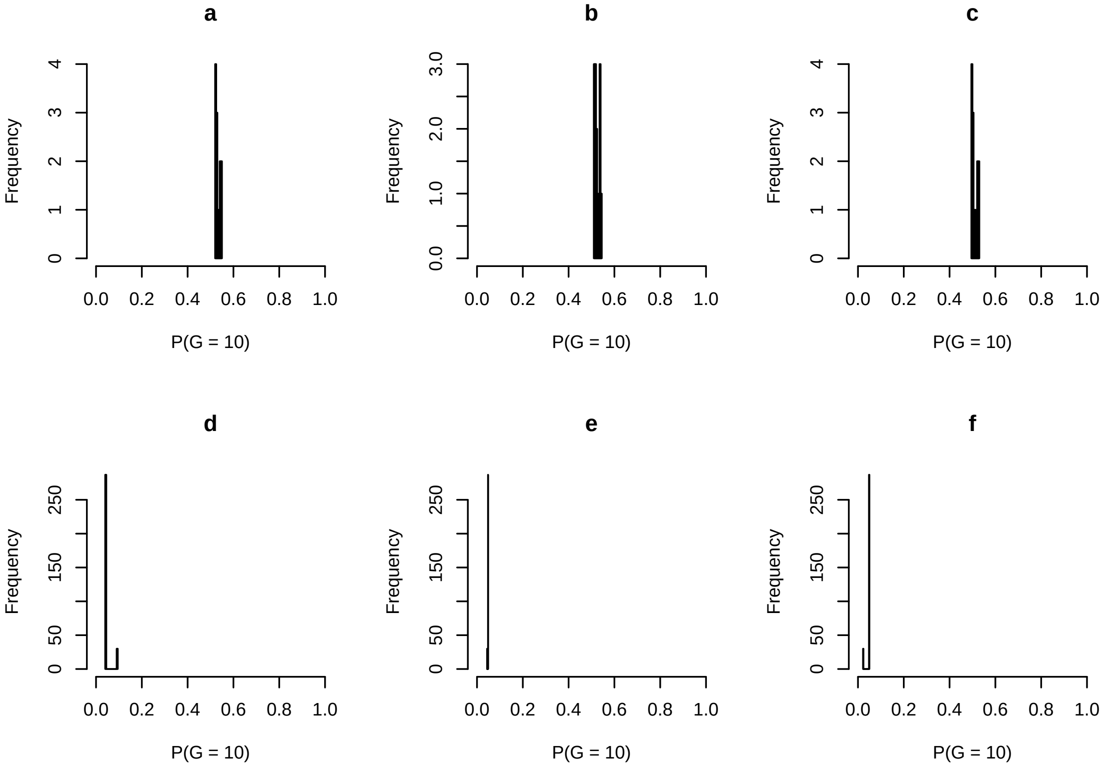

Figure 4.9: Histograms of $\hat{\mathbb{P}}(G_i = \text{NS}|O)$ in $O(0,1)$ (a, b, c) and $O(0,1)$ (d, e, f), with the corresponding offset $(\alpha_{0,\text{NS}}, \alpha_{0,\text{NN}}, \alpha_{D,\text{NN}})$ equal to $(-2,-2,-2)$ (a, d), $(-1,-1,0)$ (b, e) and $(0,0,2)$ (c, f) from left to right.

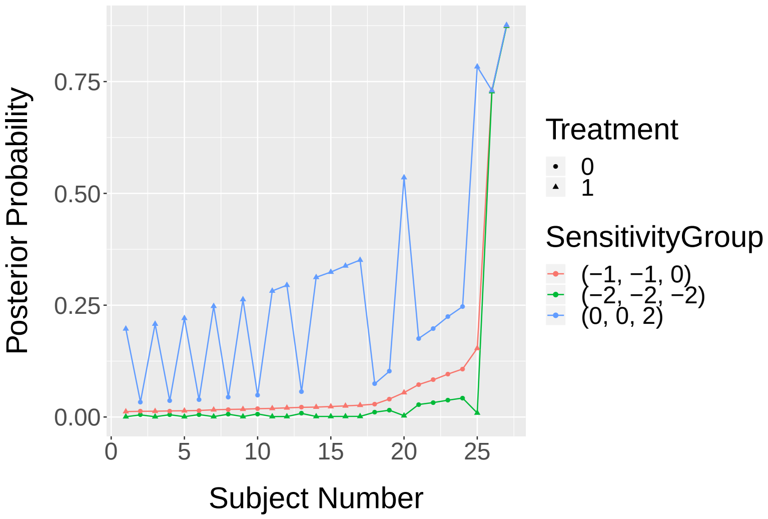

Figure 4.10: The estimated $\hat{\mathbb{P}}(Y=1|O)$ for the 27 missing subjects, with the corresponding offset $(\alpha_{0,\text{NS}},\alpha_{0,\text{NN}},\alpha_{D,\text{NN}})$ equal to $(-2,-2,-2)$, $(-1,-1,0)$ and $(0,0,2)$ in black, red, blue lines, respectively.

4.7 Appendix

4.7.1 Proofs of Main Results and the Weighted EM Steps in Section 4.3

Denote the parameter set $\Theta = \{\vartheta\}$ as a subset of

$$\mathbb{R}^2 \times \mathbb{R}^2 \times \{\text{cumulative hazards on } [0,\tau]\},$$

equipped with the norm

$$||\cdot||_\infty + ||\cdot||_\infty + ||\cdot||_{TV}, \tag{4.51}$$

where $||\cdot||_\infty$ is the sup norm and $||\cdot||_{BV}$ is the total variation norm. Define $\mathcal{H}_0$ as the unit ball in the space of functions on $[0,\tau]$ with bounded variations, equipped with the total variation norm. Define

$$\mathcal{H} = [-1,1]^2 \times [-1,1]^2 \times \mathcal{H}_0.$$

Note that $\mathcal{H}$ is uniformly bounded with an envelope, that is, an upper bound $(1,1,\mathbb{1}[0,\tau]) \in \mathcal{H}$. We use $l^\infty(\mathcal{H})$ to represent the set of all the uniformly bounded, real-valued functions on $\mathcal{H}$, that is, all functions $z:\ \mathcal{H} \longrightarrow \mathbb{R}$ such that

$$||z||_{l^\infty(\mathcal{H})} = \sup_{h\in\mathcal{H}} |z(h)| < \infty.$$

Write the score operator induced by the complete data as $A^c_\vartheta : \mathbb{H}_\vartheta \longrightarrow L_2(\mathbb{P}_\vartheta)$, where $\vartheta \in \Theta$. When A^c_ϑ acts on $h = (h_1,h_2,h_3) \in [-1,1]^2 \times [-1,1]^2 \times \mathcal{H}_0$, the score takes the form,

$$A^c_\vartheta(h) = h_1^\top \frac{\partial l^c(\vartheta)}{\partial\alpha} + h_2^\top \frac{\partial l^c(\vartheta)}{\partial\beta} + B_\vartheta(h_3), \tag{4.52}$$

where, by setting $z_1 = (1,d)^\top$ and $z_2 = (d,y)^\top$, we have

$$\frac{\partial l^c(\vartheta)}{\partial \alpha} = z_1\left(y - \frac{e^{\alpha^\top z_1}}{1+e^{\alpha^\top z_1}}\right), \tag{4.53}$$

$$\frac{\partial l^c(\vartheta)}{\partial \beta} = z_2\Big[\delta + a - \Big(\Lambda_0(x) + \sum_{j=1}^{a} \Lambda_0(t_{\cdot j})\Big)\exp(\beta^\top z_2)\Big], \tag{4.54}$$

$$B_\vartheta(h_3) = \delta h_3(x) + \sum_{j=1}^{a} h_3(t_{\cdot j}) - \exp(\beta^\top z_2)\int\Big[\mathbb{1}(0,x) + \sum_{j=1}^{a}\mathbb{1}(0,t_{\cdot j})\Big]h_3 d\Lambda. \tag{4.55}$$

Define another score operator ψ_ϑ as

$$\psi_\vartheta(h) = w^{IP}\,\mathbb{E}_\vartheta(A^c_\vartheta(h)|O). \tag{4.56}$$

Note that ϑ enters into both A^c_ϑ and the conditional expectation.

Write the empirical measure as $\mathbb{P}_n$ and the underlying probability measure as $\mathbb{P}_{\vartheta_0}$. Define a random map $\Psi_n : \Theta \longrightarrow l^\infty(\mathcal{H})$ by

$$\Psi_n(\vartheta)(h) = \mathbb{P}_n(\psi_\vartheta h) - \mathbb{P}_\vartheta(\psi_\vartheta h). \tag{4.57}$$

Define a deterministic map $\Psi :\ \Theta \longrightarrow l^\infty(\mathcal{H})$ as

$$\Psi(\vartheta)(h) = \mathbb{P}_{\vartheta_0}(\psi_\vartheta(h)) - \mathbb{P}_\vartheta(\psi_\vartheta(h)). \tag{4.58}$$

The estimator $\hat{\vartheta}$ and the true parameter ϑ_0 are the solutions to $\Psi_n(\vartheta) = 0$ and $\Psi(\vartheta) = 0$.

We prepare a remark that will be useful for the proof of Theorem 4.3.3.2.

Remark 4.7.1.1. By definition, A^c_ϑ is the score operator induced by the complete data density. We can repeat the same procedure and obtain the score operator A^o_ϑ for the observed data density. By inserting all one dimensional submodels through ϑ_0 and computing their scores, we find that

$$A^o_\vartheta(h) = \mathbb{E}_\vartheta(A^c_\vartheta(h)|O). \tag{4.59}$$

Hence ψ_ϑ can be seen as the score operator of the weighted observed data density. That is,

$$\psi_\vartheta = w^{IP} A^o_\vartheta(h). \tag{4.60}$$

Our proofs mainly rely on the Z-estimation theory developed in [VDVW96, Chapter 3.3], [VdV00, Chapter 25.12], and [Kos08, Chapter 13]. The proofs are quite straightforward based on our assumptions. More thoughts about relaxing the assumptions can be considered but are beyond our scope. More information about Fréchet derivatives can be found in [BKB+93, Appendix A.5].

In addition to assumptions in the main paper, we assume the followings for both Theorem 4.3.3.1 and Theorem 4.3.3.2.

Assumption 4.7.1.1. *The true parameter* $(\alpha_0, \beta_0, \Lambda_0)$ *is within the interior of a compact set* $\{(\alpha, \beta, \Lambda) : ||\alpha||_\infty \vee ||\beta||_\infty \vee ||\Lambda||_{TV} \leq D\}$ *for some constant D.*

Assumption 4.7.1.2. *The map* Ψ *is one-to-one.*

Proof of Theorem 4.3.3.1

We apply Theorem 2.10 in [Kos08]. Accordingly, it suffices to verify:

1. $||\Psi_n(\hat{\vartheta})||_{l^\infty(\mathcal{H})} \to_P 0$;

2. $\{\psi_\vartheta(h) : ||\vartheta - \vartheta_0|| < \delta, h \in \mathcal{H}\}$ is $\mathbb{P}_{\vartheta_0}$-Gilvenko-Cantelli;

3. $||\Psi(\vartheta_n)||_{l^\infty(\mathcal{H})} \to_P 0$ implies $\vartheta_n \to \vartheta_0$ for any sequence $\{\vartheta_n\} \in \Theta$.

Condition 1 is immediate by the definition of $\hat{\vartheta}$.

The index set $\{\psi_{\vartheta}(h) : ||\vartheta - \vartheta_0|| < \delta, h \in \mathcal{H}\}$ will be shown to be $\mathbb{P}_{\vartheta_0}$-Donsker in the proof of Theorem 4.3.3.2, hence being $\mathbb{P}_{\vartheta_0}$-Gilvenko-Cantelli.

To verify Condition 3, it suffices to prove [Kos08, Section 13.1] that $\Psi(\vartheta_0) = 0$, which is automatic, and that $\Psi : \Theta \longrightarrow l^{\infty}(\mathcal{H})$ is one-to-one, assumed by Assumption 4.7.1.2.

The consistency of $\hat{\vartheta}$ for ϑ_0 thus follows.

□

We further assume the following for Theorem 4.3.3.2.

Assumption 4.7.1.3. *The Fréchet derivative $\dot{\Psi}(\vartheta_0) : \Theta \longrightarrow l^{\infty}(\mathcal{H})$ of Ψ at ϑ_0 is continuously invertible.*

Proof of Theorem 4.3.3.2

To prove asymptotic Gaussianity, we adopt Theorem 3.3.1, together with Lemma 3.3.5, in [VDVW96] (or similarly, as pointed out in the Section 3 in [BW07] for weighted complete data likelihood). The theorem states that when the following conditions (referred later as condition 1-4) are satisfied,

1. The set $\{\psi_{\vartheta}(h) : ||\vartheta - \vartheta_0|| < \delta, h \in \mathcal{H}\}$ for some $\delta > 0$ is $\mathbb{P}_{\vartheta_0}$-Donsker;
2. $\sup_{h \in \mathcal{H}} \mathbb{P}_{\vartheta_0}[(\psi_{\vartheta} h - \psi_{\vartheta_0} h)^2] \to 0$ whenever $\vartheta \to \vartheta_0$;
3. The map Ψ has a Fréchet derivative $\dot{\Psi}$ at ϑ_0 that is continuously invertible on its range;
4. $\hat{\vartheta}$ is consistent for ϑ_0 and satisfies $\Psi_n(\hat{\vartheta}) = 0$;

we have

$$\sqrt{n}(\hat{\vartheta} - \vartheta_0) = -\dot{\Psi}^{-1}(\vartheta_0)\mathbb{G}_n(\psi_{\vartheta_0}) + o_P(1) \to_d -\dot{\Psi}^{-1}(\vartheta_0)\mathbb{G}(\psi_{\vartheta_0}), \tag{4.61}$$

where $\mathbb{G}_n$ is the empirical process $\sqrt{n}(\mathbb{P}_n - \mathbb{P}_{\vartheta_0})$ and $\mathbb{G}$ is the $\mathbb{P}_{\vartheta_0}$-Brownian bridge.

To prove condition 1, we recall some basic properties of Donsker classes:

- All functions that are of variation bounded by one form a Donsker class [VdV00, Example 19.11].
- The sum of two Donsker classes with an integrable envelope function is Donsker [VDVW96, Example 2.10.7].
- A Donsker class multiplied by a uniformly bounded, measurable function remains Donsker [VDVW96, Example 2.10.10].
- Closures and convex hulls of Donsker classes remain Donsker [VDVW96, Theorem 2.10.2, Theorem 2.10.3]. Therefore, the conditional expectation, which can be seen as the L_2 limit of a sequence of convex combinations by its definition, preserves Donsker property.

We separate the complete data scores (4.53), (4.54) and (4.57) into two parts $A_{1,\vartheta}$ and $A_{2,\vartheta}$: the one without and the other one involving the ghost copies.

$$\frac{\partial l^{c,1}(\vartheta)}{\partial \alpha} = z_1\left(y - \frac{e^{\alpha^\top z_1}}{1+e^{\alpha^\top z_1}}\right), \tag{4.62}$$

$$\frac{\partial l^{c,1}(\vartheta)}{\partial \beta} = z_2\left[\delta - \Lambda_0(x)\exp(\beta^\top z_2)\right], \tag{4.63}$$

$$B^c_{1,\vartheta}(h_3) = \delta h_3(x) - \exp(\beta^\top z_2)\int \mathbb{1}_{(0,x)}(u)h_3(u)d\Lambda(u), \tag{4.64}$$

and

$$\frac{\partial l^{c,2}(\vartheta)}{\partial \alpha} = 0, \tag{4.65}$$

$$\frac{\partial l^{c,2}(\vartheta)}{\partial \beta} = z_2\Big[a - \sum_{j=1}^{a} \Lambda_0(t_{\cdot j})\exp(\beta^\top z_2)\Big], \tag{4.66}$$

$$B^c_{2,\vartheta}(h_3) = \sum_{j=1}^{a} h_3(t_{\cdot j}) - \exp(\beta^\top z_2)\int \sum_{j=1}^{a} \mathbb{1}_{(0,t_{\cdot j})}(u)h_3(u)d\Lambda(u). \tag{4.67}$$

We show that, after conditional expectation and multiplication with IP weights, both of them are Donsker, so is their sum.

The variables y, δ, z_1 and z_2 that enter into the complete data scores (4.62), (4.63) and (4.64) are bounded since they are binary. Parameters α, β and Λ are bounded by Assumption 4.7.1.1. The index set $\mathcal{H}$ is Donsker by its definition. Therefore, (4.62), (4.63) and (4.64) are Donsker, so are them after conditional expectation. IP weights, by Assumption 4.2.3.4, are uniformly bounded, measurable functions. We conclude that $\{w^{IP}\mathbb{E}(A_{1,\vartheta}|O) : ||\vartheta - \vartheta_0|| < \delta, h \in \mathcal{H}\}$ is Donsker.

After taking conditional expectation on $A_{2,\vartheta}$, (4.65), (4.66) and (4.67) become

$$\mathbb{E}_\vartheta\left[\frac{\partial l^{c,2}(\vartheta)}{\partial \alpha}\middle| O\right] = 0, \tag{4.68}$$

$$\mathbb{E}_\vartheta\left[\frac{\partial l^{c,2}(\vartheta)}{\partial \beta}\middle| O\right] = z_2\,\mathbb{E}_\vartheta(A|z_2)\Big[1 - \mathbb{E}_\vartheta(\Lambda_0(T_{\cdot 1})\exp(\beta^\top z_2)|A, z_2\Big], \tag{4.69}$$

and

$$\begin{aligned} & \mathbb{E}_\vartheta[B^c_{2,\vartheta}(h_3)|O] & (4.70)\\ = \;& \mathbb{E}_\vartheta(A|z_2)\left[\mathbb{E}(h_3(T_{\cdot 1}|A,z_2) - \exp(\beta^\top z_2)\,\mathbb{E}\left(\int \mathbb{1}_{(0,T_{\cdot 1})}(u)h_3(u)d\Lambda(u)\middle| A, z_2\right)\right]. & (4.71)\end{aligned}$$

The term $\mathbb{E}(A|z_2) = 1/P(T > Q|z_2) - 1$ is bounded by Assumption 4.2.3.6. Other parts go through similar to those in the first part, by which we reach to the conclusion that $\{w^{IP}\mathbb{E}(A_{2,\vartheta}|O) : ||\vartheta-\vartheta_0|| < \delta, h \in \mathcal{H}\}$ is Donsker. Condition 1 is verified.

We obtained an envelope function for $\{\psi_\vartheta(h) : \vartheta \in \Theta, h \in \mathcal{H}\}$ when proving condition 1. By Dominated Convergence Theorem, condition 2 simplifies to showing pointwise convergence of $\psi_\vartheta(h)$ to $\psi_{\vartheta_0}(h)$, uniformly in $h \in \mathcal{H}$. This follows from the facts that $A^c_\vartheta(h)$ converges to $A^c_{\vartheta_0}(h)$ pointwisely, uniformly in $h \in \mathcal{H}$ whenever $\vartheta \to \vartheta_0$, and that the conditional expectation is a continuous operator.

For condition 3, heuristically, Fubini's Theorem implies Fréchet differentiability of Ψ at ϑ_0. By computing the Gâteaux derivative, we find that the Fréchet derivative $\dot{\Psi}$ (the Fréchet derivative coincides with the Gâteaux derivative when an operator is Fréchet differentiable) at ϑ_0 is $-\mathbb{P}_{\vartheta_0}(w^{IP}A^{o*}_\vartheta A^o_\vartheta)$, where A^{o*}_ϑ is the adjoint operator of A^o_ϑ, see [VdV00, Equation (25.91)]. The continuous invertibility of $\dot{\Psi}(\vartheta_0)$ in Condition 3 is assumed in Assumption 4.7.1.3.

Condition 4 is immediate by Theorem 4.3.3.1.

These finish the proof. □

Variance Estimation

The ES algorithm presented in the main text is a specific case of [ER04] by taking $U^{(1)}(Y^{obs}, S(Y^c), \vartheta) = S(Y^c)$ and $S(Y^c)$, following their notation, as the derivatives $\frac{\partial l^c_w(\vartheta)}{\partial \vartheta}$ of the weighted complete data likelihood $l^c_w(\vartheta)$ in (4.8) with respect to $(\alpha, \beta, \lambda_1, \cdots, \lambda_K)$.

The Louis' formula in [ER04, Page 5] proved to provide a valid variance estimate for estimator derived from a finite number of estimating equations. We show this formula is still valid in our case with an infinite number of estimating equations.

We first depict the formula. Within our context, an estimate of the asymptotic variance of $\hat{\theta} = (\hat{\alpha}, \hat{\beta}, \hat{\lambda}_1, \cdots, \hat{\lambda}_K)$ is given by

$$\left(\frac{1}{n}\sum_{i=1}^n \frac{\partial}{\partial\theta}U_i(\theta)\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^n U_i(\theta)U_i(\theta)^\top\right)\left(\frac{1}{n}\sum_{i=1}^n \frac{\partial}{\partial\theta}U_i(\theta)^\top\right)^{-1}\Bigg|_{\theta=\hat{\theta}}, \tag{4.72}$$

where

$$U_i(\theta) = w_i^{IP}\, \mathbb{E}_\theta\Big(\frac{\partial l^c_{w,i}(\theta)}{\partial\theta}\Big| O\Big), \tag{4.73}$$

and $l^c_{w,i}(\theta)$ is the contribution from the i-th subject in (4.8). This in turn, can provide a variance estimate of $\hat{\vartheta}$ as an element in Θ. To prove so, it suffices to show that this variance estimate is valid for any finite choices of $h_1, \cdots, h_L$.

An initial variance estimate might be, informally writing, based on approximating the variance of $-\dot{\Psi}_n^{-1}(\hat{\vartheta})\mathbb{G}(\psi_{\vartheta_0})$. However, $\dot{\Psi}_n(\hat{\vartheta})$ may not be invertible as an operator on Θ.

Nonetheless, since we are interested in providing a variance estimate for a finite dimensional estimator, we may project $\sqrt{n}(\hat{\vartheta} - \vartheta_0)$ into the finite dimensional set Θ_1, where $\Theta_1 \subset \Theta$ is defined as subset of

$$\mathbb{R}^2 \times \mathbb{R}^2 \times \{\text{cumulative hazards on } [0,\tau] \text{ with nonnegative jumps only at } t_1 < \cdots < t_K\},$$

which can be identified as a subset of $\mathbb{R}^{2+2+K}$ equipped with the sup norm. It is easy to show that $\dot{\Psi}_n(\hat{\vartheta})$ stay invariant under this projection, which by (4.61) yields

$$\sqrt{n}(\hat{\vartheta} - \vartheta_0) \to N(0, (\dot{\Psi}_n(\hat{\vartheta}) + o_P(1))^{-1}\, \mathbb{E}(\psi_{\vartheta_0}\psi_{\vartheta_0}^\top)(\dot{\Psi}_n(\hat{\vartheta}) + o_P(1))^{-1})^\top. \tag{4.74}$$

Within this set, the operator $\dot{\Psi}_n(\hat{\vartheta})$ is indeed invertible, thus giving us a variance estimate as

$$\dot{\Psi}_n^{-1}(\hat{\vartheta})\, \mathbb{E}_n(\psi_{\hat{\vartheta}}\psi_{\hat{\vartheta}}^\top)(\dot{\Psi}_n^{-1}(\hat{\vartheta}))^\top \tag{4.75}$$

Implementation of the weighted EM Algorithm

In this section, we present E-functions, the scores and the negative Hessians.

E-FUNCTIONS

We use $\theta^{(t)}$ and superscript (t) to represent t-th iteration ES parameter and conditioning on $\theta^{(t)}$. At the $(t+1)$-th iteration $(t = 0, 1, \cdots)$, we have

$$\mathbb{P}^{(t)}(Y_i^{mis} = 1 | D_i, T_i > X_i, Q_i) = \frac{\pi_i^{(t)} S_i^{(t)}(X_i | Y_i = 1)}{\pi_i^{(t)} S_i^{(t)}(X_i | Y_i = 1) + (1 - \pi_i^{(t)}) S_i^{(t)}(X_i | Y_i = 0)}, \tag{4.76}$$

$$\mathbb{P}^{(t)}(Y_i^{mis} = 1 | D_i, T_i = X_i, Q_i) = \frac{\pi_i^{(t)} f_i^{(t)}(X_i | Y_i = 1)}{\pi_i^{(t)} f_i^{(t)}(X_i | Y_i = 1) + (1 - \pi_i^{(t)}) f_i^{(t)}(X_i | Y_i = 0)}, \tag{4.77}$$

$$\mathbb{E}^{(t)}(A_i | D_i, Y_i = 1, Q_i) = \frac{1 - S_i^{(t)}(Q_i | Y_i = 1)}{S_i^{(t)}(Q_i | Y_i = 1)}, \tag{4.78}$$

$$\mathbb{E}^{(t)}(A_i | D_i, Y_i = 0, Q_i) = \frac{1 - S_i^{(t)}(Q_i | Y_i = 0)}{S_i^{(t)}(Q_i | Y_i = 0)}, \tag{4.79}$$

$$\mathbb{P}^{(t)}(T_{i1} = t_k | D_i, Y_i = 1, Q_i) = \frac{\mathbb{1}(t_k < Q_i) f_i^{(t)}(t_k | Y_i = 1)}{1 - S_i^{(t)}(Q_i | Y_i = 1)}, \tag{4.80}$$

$$\mathbb{P}^{(t)}(T_{i1} = t_k | D_i, Y_i = 0, Q_i) = \frac{\mathbb{1}(t_k < Q_i) f_i^{(t)}(t_k | Y_i = 0)}{1 - S_i^{(t)}(Q_i | Y_i = 0)}, \tag{4.81}$$

$$\begin{aligned} & \mathbb{E}^{(t)}\left[Y_i \sum_{j=1}^{A_i} \mathbb{1}(T_{ij} = t_k) \Big| D_i, T_i > X_i, Q_i \right] \\ = \ & \mathbb{P}^{(t)}(Y_i = 1 | D_i, T_i > X_i, Q_i) \, \mathbb{E}^{(t)}(A_i | D_i, Y_i = 1, Q_i) \\ & \cdot \mathbb{P}^{(t)}(T_{i1} = t_k | D_i, Y_i = 1, Q_i), \end{aligned} \tag{4.82}$$

$$\begin{aligned}
&\mathbb{E}^{(t)}\left[(1-Y_i)\sum_{j=1}^{A_i}\mathbb{1}(T_{ij}=t_k)\Big|D_i,T_i>X_i,Q_i\right]\\
=\;&(1-\mathbb{P}^{(t)}(Y_i=1|D_i,T_i>X_i,Q_i))\,\mathbb{E}^{(t)}(A_i|D_i,Y_i=0,Q_i)\\
&\cdot\mathbb{P}^{(t)}(T_{i1}=t_k|D_i,Y_i=0,Q_i),
\end{aligned}\tag{4.83}$$

$$\begin{aligned}
&\mathbb{E}^{(t)}\left[Y_i\sum_{j=1}^{A_i}\mathbb{1}(T_{ij}=t_k)\Big|D_i,T_i=X_i,Q_i\right]\\
=\;&\mathbb{P}^{(t)}(Y_i=1|D_i,T_i=X_i,Q_i)\,\mathbb{E}^{(t)}(A_i|D_i,Y_i=1,Q_i)\\
&\cdot\mathbb{P}^{(t)}(T_{i1}=t_k|D_i,Y_i=1,Q_i),
\end{aligned}\tag{4.84}$$

$$\begin{aligned}
&\mathbb{E}^{(t)}\left[(1-Y_i)\sum_{j=1}^{A_i}\mathbb{1}(T_{ij}=t_k)\Big|D_i,T_i=X_i,Q_i\right]\\
=\;&(1-\mathbb{P}^{(t)}(Y_i=1|D_i,T_i=X_i,Q_i))\,\mathbb{E}^{(t)}(A_i|D_i,Y_i=0,Q_i)\\
&\cdot\mathbb{P}^{(t)}(T_{i1}=t_k|D_i,Y_i=0,Q_i),
\end{aligned}\tag{4.85}$$

where the first two E-functions serve for missing outcomes and the remaining for left truncation.

Remark 4.7.1.2. Since T_{ij}, A_i are conditionally independent of T_i, we have

$$\mathbb{E}^{(t)}(A_i|D_i,Y_i,T_i=X_i,Q_i)=\mathbb{E}^{(t)}(A_i|D_i,Y_i,T_i>X_i,Q_i)=\mathbb{E}^{(t)}(A_i|D_i,Y_i,Q_i),\tag{4.86}$$

$$\mathbb{P}^{(t)}(T_{i1}=t_k|A_i,D_i,Y_i,T_i=X_i,Q_i)=\mathbb{P}^{(t)}(T_{i1}=t_k|D_i,Y_i,Q_i),\tag{4.87}$$

and

$$\mathbb{E}^{(t)}\Big[\sum_{j=1}^{A_i}\mathbb{1}(T_{ij}=t_k)\Big|D_i,Y_i,Q_i\Big]=\mathbb{E}^{(t)}(A_i|D_i,Y_i,Q_i)\,\mathbb{P}^{(t)}(T_{i1}=t_k|D_i,Y_i,Q_i),\tag{4.88}$$

hence the last four E-functions.

Thus the Q function, that is, the expectation of the complete data log-likelihood given the observed variables at the $(l+1)$-iteration, becomes

$$
\begin{aligned}
& Q(\theta|\theta^{(t)}) \\
= \quad & \sum_{\Delta_i=0,O_i=1} w_i^{IP}\Big\{Y_i \log\pi_i + (1-Y_i)\log(1-\pi_i) + \log S_i(X_i)\Big\} \\
+ \quad & \sum_{\Delta_i=0,O_i=0} w_i^{IP}\Big\{\,\mathbb{P}_i^{(t)}(Y_i^{mis}=1)\log\pi_i + \mathbb{P}_i^{(t)}(Y_i^{mis}=0)\log(1-\pi_i) \\
& +\mathbb{P}_i^{(t)}(Y_i^{mis}=1)\log S_i(X_i|Y_i=1) + \mathbb{P}_i^{(t)}(Y_i^{mis}=0)\log S_i(X_i|Y_i=0)\Big\} \\
+ \quad & \sum_{\delta_i=1,O_i=1} w_i^{IP}\Big\{Y_i \log\pi_i + (1-Y_i)\log(1-\pi_i) + \log f_i(X_i)\Big\} \\
+ \quad & \sum_{\Delta_i=1,O_i=0} w_i^{IP}\Big\{\,\mathbb{P}_i^{(t)}(Y_i^{mis}=1)\log\pi_i + \mathbb{P}_i^{(t)}(Y_i^{mis}=0)\log(1-\pi_i) \\
& +\mathbb{P}_i^{(t)}(Y_i^{mis}=1)\log f_i(X_i|Y_i=1) + \mathbb{P}_i^{(t)}(Y_i^{mis}=0)\log f_i(X_i|Y_i=0)\Big\} \\
+ \quad & \sum_{i=1}^{n} w_i^{IP}\Big\{\,\mathbb{E}_i^{(t)}\Big[Y_i\sum_{j=1}^{A_i}\log f_i(T_{ij}|Y_i=1) + (1-Y_i)\sum_{j=1}^{A_i}\log f_i(T_{ij}|Y_i=0))\Big]\Big\}.
\end{aligned}
$$

Inference

To conduct inference, by setting $Z_{1,i} = (1, D_i)^\top$ and $Z_{2,i} = (D_i, Y_i)^\top$, we compute the complete data scores and negative Hessians $H_i(\theta)$ for each subject i,

$$
\frac{\partial l_i^c(\theta)}{\partial\alpha} = Z_{1,i}\bigg(Y_i - \frac{e^{\alpha^\top Z_{1,i}}}{1+e^{\alpha^\top Z_{1,i}}}\bigg), \tag{4.89}
$$

$$
\frac{\partial l_i^c(\theta)}{\partial\beta} = Z_{2,i}\Big[\Delta_i + A_i - \Big(\Lambda_0(X_i) + \sum_{j=1}^{A_i}\Lambda_0(T_{ij})\Big)\exp(\beta^\top Z_{2,i})\Big], \tag{4.90}
$$

$$
\begin{aligned}
\frac{\partial l_i^c(\theta)}{\partial \lambda_k} &= \frac{\Delta_i \mathbb{1}(X_i = t_k) + \sum_{j=1}^{A_i} \mathbb{1}(T_{ij} = t_k)}{\lambda_k} && (4.91)\\
&\quad - \exp(\beta^\top Z_{2,i}) \Big[\mathbb{1}(X_i \geq t_k) + \sum_{j=1}^{A_i} \mathbb{1}(T_{ij} \geq t_k) \Big], && (4.92)
\end{aligned}
$$

$$
-\frac{\partial^2}{\partial^2 \alpha} l_i^c(\theta) = Z_{1,i}^{\otimes 2} \frac{e^{\alpha^\top Z_i}}{(1 + e^{\alpha^\top Z_i})^2} = Z_{1,i}^{\otimes 2} \mathbb{P}(Y_i = 1 | Z_{1,i}) \mathbb{P}(Y_i = 0 | Z_{1,i}), \tag{4.93}
$$

$$
-\frac{\partial^2}{\partial^2 \beta} l_i^c(\theta) = Z_{2,i}^{\otimes 2} \Big[\Lambda_{0i}(X_i) + \sum_{j=1}^{A_i} \Lambda(T_{ij}) \Big] \exp(\beta^\top Z_{2,i}), \tag{4.94}
$$

$$
-\frac{\partial^2}{\partial^2 \lambda_k} l_i^c(\theta) = \frac{\Delta_i \mathbb{1}(X_i = t_k) + \sum_{j=1}^{A_i} \mathbb{1}(T_{ij} = t_k)}{\lambda_k^2}, \tag{4.95}
$$

$$
-\frac{\partial^2}{\partial \beta \partial \lambda_k} l_i^c(\theta) = Z_{2,i} \exp(\beta^\top Z_{2,i}) \Big[\mathbb{1}(X_i \geq t_k) + \sum_{j=1}^{A_i} \mathbb{1}(T_{ij} \geq t_k) \Big]. \tag{4.96}
$$

All the other off-diagonal terms are zero. With the observed data, we can simulate the complete data at $\hat{\theta}$, compute the complete data score S_i and negative Hessians H_i, repeat 2000 times for example, and average to get the variance estimate.

Sensitivity Analysis

When conducting the sensitivity analysis in Section 4.5.1, we offset β_Y at some prechosen value $\beta_{Y,0}$. The dimension of parameters decreases by one. To obtain $\hat{\theta}$, it suffices to set $\beta_Y^{(t)} = \beta_{Y,0}$ for any non-negative integer t in the ES algorithm. Here we adjust the scores and negative Hessians appeared in the variance estimate,

$$\frac{\partial l_i^c(\theta)}{\partial \alpha} = Z_{1,i}\left(Y_i - \frac{e^{\alpha^\top Z_{1,i}}}{1+e^{\alpha^\top Z_{1,i}}}\right), \tag{4.97}$$

$$\frac{\partial l_i^c(\theta)}{\partial \beta_D} = D_i\Big[\Delta_i + A_i - \Big(\Lambda_0(X_i) + \sum_{j=1}^{A_i}\Lambda_0(T_{ij})\Big)\exp(\beta_D D_i + \beta_{Y,0}Y_i)\Big], \tag{4.98}$$

$$\begin{aligned} \frac{\partial l_i^c(\theta)}{\partial \lambda_k} &= \frac{\Delta_i \mathbb{1}(X_i = t_k) + \sum_{j=1}^{A_i}\mathbb{1}(T_{ij} = t_k)}{\lambda_k} && (4.99)\\ &\quad - \exp(\beta_D D_i + \beta_{Y,0}Y_i)\Big[\mathbb{1}(X_i \geq t_k) + \sum_{j=1}^{A_i}\mathbb{1}(T_{ij} \geq t_k)\Big], && (4.100) \end{aligned}$$

$$-\frac{\partial^2}{\partial^2\alpha} l_i^c(\theta) = Z_{1,i}^{\otimes 2}\frac{e^{\alpha^\top Z_i}}{(1+e^{\alpha^\top Z_i})^2} = Z_{1,i}^{\otimes 2}\,\mathbb{P}(Y_i = 1|Z_{1,i})\,\mathbb{P}(Y_i = 0|Z_{1,i}), \tag{4.101}$$

$$-\frac{\partial^2}{\partial^2\beta_D} l_i^c(\theta) = D_i^2\Big[\Lambda_{0i}(X_i) + \sum_{j=1}^{A_i}\Lambda(T_{ij})\Big]\exp(\beta_D D_i + \beta_{Y,0}Y_i), \tag{4.102}$$

$$-\frac{\partial^2}{\partial^2\lambda_k} l_i^c(\theta) = \frac{\Delta_i \mathbb{1}(X_i = t_k) + \sum_{j=1}^{A_i}\mathbb{1}(T_{ij} = t_k)}{\lambda_k^2}, \tag{4.103}$$

$$-\frac{\partial^2}{\partial\beta_D\partial\lambda_k} l_i^c(\theta) = D_i\exp(\beta_D D_i + \beta_{Y,0}Y_i)\Big[\mathbb{1}(X_i \geq t_k) + \sum_{j=1}^{A_i}\mathbb{1}(T_{ij} \geq t_k)\Big]. \tag{4.104}$$

4.7.2 Proofs of the Main Results and ES steps in Section 4.4

Proofs of Theorem 4.4.0.1 and Theorem 4.4.0.2

The proofs are similar to those of of Theorem 4.3.3.1 and Theorem 4.3.3.2, thus we omit.

ES Steps for the Principal Effects

We still provide the E-functions of missing outcomes for completeness although we use multiple imputation in the study.

E-STEP

We use $\theta^{(t)}$ and superscript (t) to represent t-th iteration ES parameter and conditioning on $\theta^{(t)}$. At the $(t+1)$-th iteration $(t = 0, 1, \cdots)$, we present E-functions corresponding to different observed groups, see Table 4.5,

• for $i \in$ O(0, 0),

$$\mathbb{P}^{(t)}(G_i = g | D_i = 0, M_i = 0, T_i > X_i, Q_i) = 1, \tag{4.105}$$

$$\mathbb{P}^{(t)}(G_i = g | D_i = 0, M_i = 0, Y_i = y, T_i > X_i, Q_i) = 1, \tag{4.106}$$

when $g = \text{SS}$,

$$\mathbb{P}^{(t)}(G_i = g | D_i = 0, T_i > X_i, Q_i) = \mathbb{P}^{(t)}(G_i = g | D_i = 0, Y_i = y, T_i > X_i, Q_i) = 0, \tag{4.107}$$

when $g = \text{NS}, \text{NN}$,

$$\mathbb{P}^{(t)}(Y_i = 1 | G_i = g, D_i = 1, T_i > X_i, Q_i) = \pi_i^{(t)}(G_i = g), \tag{4.108}$$

$$\mathbb{E}^{(t)}(A_i|G_i = g, D_i = 0, M_i = 0, Y_i, Q_i) = 0, \tag{4.109}$$

$$\mathbb{P}^{(t)}(T_{i1} = t_k|G_i = g, D_i, Y_i = 1, Q_i) = 0, \tag{4.110}$$

for all g.

- for $i \in$ O(0, 1),

$$\mathbb{P}^{(t)}(G_i = g|D_i = 0, T_i = X_i, Q_i) = \mathbb{P}^{(t)}(G_i = g|D_i = 0, Y_i = y, T_i = X_i, Q_i) = 0, \tag{4.111}$$

$$\mathbb{P}^{(t)}(Y_i = 1|G_i = g, D_i = 0, M_i = 1, T_i = X_i, Q_i) = 0, \tag{4.112}$$

$$\mathbb{E}^{(t)}(A_i|G_i = g, D_i = 1, M_i = 0, Y_i, Q_i) = 0, \tag{4.113}$$

$$\mathbb{P}^{(t)}(T_{i1} = t_k|G_i = g, D_i, Y_i = 1, Q_i) = 0, \tag{4.114}$$

when $g = \text{SS}$,

$$\begin{aligned} &\mathbb{P}^{(t)}(G_i = g|D_i = 0, T_i = X_i, Q_i) \\ = \quad &\frac{\sum_y p_i^{(t)}(g)\pi_i^{(t)}(G_i = g) f_i^{(t)}(X_i|G_i = g, Y_i = y)}{\sum_{g=\text{NS,NN}} \sum_y p_i^{(t)}(g)\, \mathbb{P}^{(t)}(Y_i = y|G_i = g) f_i^{(t)}(X_i|G_i = g, Y_i = y)}, \end{aligned} \tag{4.115}$$

$$\begin{aligned} &\mathbb{P}^{(t)}(G_i = g|D_i = 0, Y_i = y, T_i = X_i, Q_i) \\ = \quad &\frac{p_i^{(t)}(g)\, \mathbb{P}^{(t)}(Y_i = y|G_i = g) f_i^{(t)}(X_i|G_i = g, Y_i = y)}{\sum_{g=\text{NS,NN}} p_i^{(t)}(g)\, \mathbb{P}^{(t)}(Y_i = y|G_i = g) f_i^{(t)}(X_i|G_i = g, Y_i = y)}, \end{aligned} \tag{4.116}$$

$$\begin{aligned}&\mathbb{P}^{(t)}(Y_i = 1|G_i = g, D_i = 0, M_i = 0, T_i = X_i, Q_i)\\ = \quad &\frac{\pi_i^{(t)}(G_i = g) f_i^{(t)}(X_i|G_i = g, Y_i = 1)}{\sum_y \pi_i^{(t)}(G_i = g) f_i^{(t)}(X_i|G_i = g, Y_i = y)},\end{aligned} \tag{4.117}$$

$$\mathbb{E}^{(t)}(A_i|G_i = g, D_i, Y_i = 1, Q_i) = \frac{1 - S_i^{(t)}(Q_i|G_i = g, Y_i = 1)}{S_i^{(t)}(Q_i|G_i = g, Y_i = 1)}, \tag{4.118}$$

$$\mathbb{P}^{(t)}(T_{i1} = t_k|G_i = g, D_i, Y_i = 1, Q_i) = \frac{\mathbb{1}(t_k < Q_i) f_i^{(t)}(t_k|G_i = g, Y_i = 1)}{\sum_{h:t_h<Q_i} f_i^{(t)}(t_h|G_i = g, Y_i = 1)}, \tag{4.119}$$

when $g = \text{NS}, \text{NN}$.

- for $i \in$ O(0, ?),

$$\begin{aligned}&\mathbb{P}^{(t)}(G_i = g|D_i = 0, T_i > X_i, Q_i)\\ = \quad &\frac{\sum_y p_i^{(t)}(g)\, \mathbb{P}^{(t)}(Y_i = y|G_i = g) S_i^{(t)}(X_i|G_i = g, Y_i = y)}{\sum_g \sum_y p_i^{(t)}(g)\, \mathbb{P}^{(t)}(Y_i = y|G_i = g) S_i^{(t)}(X_i|G_i = g, Y_i = y)},\end{aligned} \tag{4.120}$$

$$\begin{aligned}&\mathbb{P}^{(t)}(G_i = g|D_i = 0, Y_i = y, T_i > X_i, Q_i)\\ = \quad &\frac{p_i^{(t)}(g)\, \mathbb{P}^{(t)}(Y_i = y|G_i = g) S_i^{(t)}(X_i|G_i = g, Y_i = y)}{\sum_g p_i^{(t)}(g)\, \mathbb{P}^{(t)}(Y_i = y|G_i = g) S_i^{(t)}(X_i|G_i = g, Y_i = y)},\end{aligned} \tag{4.121}$$

when $g = \text{SS}, \text{NS}, \text{NN}$,

$$\mathbb{P}^{(t)}(Y_i = 1|G_i = g, D_i = 1, M_i =?, T_i > X_i, Q_i) = 0, \tag{4.122}$$

$$\mathbb{E}^{(t)}(A_i|G_i = g, D_i = 0, M_i =?, Y_i, Q_i) = 0, \tag{4.123}$$

$$\mathbb{P}^{(t)}(T_{i1} = t_k | G_i = g, D_i, M_i = ?, Y_i = 1, Q_i) = 0, \tag{4.124}$$

when $g = \text{SS}$,

$$\begin{aligned} & \mathbb{P}^{(t)}(Y_i = 1 | G_i = g, D_i = 1, M_i = ?, T_i > X_i, Q_i) \\ = & \frac{\pi_i^{(t)}(G_i = g) S_i^{(t)}(X_i | G_i = g, Y_i = 1)}{\sum_y \pi_i^{(t)}(G_i = g) S_i^{(t)}(X_i | G_i = g, Y_i = y)}, \end{aligned} \tag{4.125}$$

$$\mathbb{E}^{(t)}(A_i | G_i = g, D_i, Y_i = 1, Q_i) = \frac{1 - S_i^{(t)}(Q_i | G_i = g, Y_i = 1)}{S_i^{(t)}(Q_i | G_i = g, Y_i = 1)}, \tag{4.126}$$

$$\mathbb{P}^{(t)}(T_{i1} = t_k | G_i = g, D_i, Y_i = 1, Q_i) = \frac{\mathbb{1}(t_k < Q_i) f_i^{(t)}(t_k | G_i = g, Y_i = 1)}{\sum_{h: t_h < Q_i} f_i^{(t)}(t_h | G_i = g, Y_i = 1)}, \tag{4.127}$$

when $g = \text{NS}, \text{NN}$.

- for $i \in$ O(1, 0),

$$\mathbb{P}^{(t)}(G_i = g | D_i = 1, T_i > X_i, Q_i) = \frac{\sum_y p_i^{(t)}(g)}{\sum_{g = \text{SS,NS}} p_i^{(t)}(g)}, \tag{4.128}$$

$$\begin{aligned} & \mathbb{P}^{(t)}(G_i = g | D_i = 1, Y_i = y, T_i > X_i, Q_i) \\ = & \frac{p_i^{(t)}(g) \mathbb{P}^{(t)}(Y_i = y | G_i = g)}{\sum_{g = \text{SS,NS}} p_i^{(t)}(g) \mathbb{P}^{(t)}(Y_i = y | G_i = g)}, \end{aligned} \tag{4.129}$$

$$\mathbb{P}^{(t)}(Y_i = 1 | G_i = g, D_i = 1, T_i > X_i, Q_i) = \pi_i^{(t)}(G_i = g) \tag{4.130}$$

when $g = \text{SS}, \text{NS}$,

$$\mathbb{P}^{(t)}(G_i = g | D_i = 1, T_i > X_i, Q_i) = \mathbb{P}^{(t)}(G_i = g | D_i = 1, Y_i = y, T_i > X_i, Q_i) = 0, \tag{4.131}$$

$$\mathbb{P}^{(t)}(Y_i = 1 | G_i = g, D_i = 1, T_i > X_i, Q_i) = 0, \tag{4.132}$$

when $g = \text{SS}$,

$$\mathbb{E}^{(t)}(A_i | G_i = g, D_i = 1, M_i = 0, Y_i, Q_i) = 0, \tag{4.133}$$

$$\mathbb{P}^{(t)}(T_{i1} = t_k | G_i = g, D_i, Y_i = 1, Q_i) = 0, \tag{4.134}$$

for all g.

- for $i \in$ O(1, 1),

$$\mathbb{P}^{(t)}(G_i = g | D_i = 1, T_i = X_i, Q_i) = \mathbb{P}^{(t)}(G_i = g | D_i = 1, Y_i = y, T_i = X_i, Q_i) = 0, \tag{4.135}$$

$$\mathbb{P}^{(t)}(Y_i = 1 | G_i = g, D_i = 1, M_i = 1, T_i = X_i, Q_i) = 0, \tag{4.136}$$

$$\mathbb{P}^{(t)}(Y_i = 1 | G_i = g, D_i = 1, T_i > X_i, Q_i) = 0, \tag{4.137}$$

$$\mathbb{E}^{(t)}(A_i | G_i = g, D_i = 1, M_i = 0, Y_i, Q_i) = 0, \tag{4.138}$$

$$\mathbb{P}^{(t)}(T_{i1} = t_k | G_i = g, D_i, Y_i = 1, Q_i) = 0, \tag{4.139}$$

when $g = \text{SS}, \text{NS}$,

$$\mathbb{P}^{(t)}(G_i = g | D_i = 1, T_i = X_i, Q_i) = 1, \tag{4.140}$$

$$\mathbb{P}^{(t)}(G_i = g | D_i = 1, Y_i = y, T_i = X_i, Q_i) = 1, \tag{4.141}$$

$$\begin{aligned} & \mathbb{P}^{(t)}(Y_i = 1 | G_i = g, D_i = 1, M_i = 1, T_i = X_i, Q_i) \\ = \; & \frac{\pi_i^{(t)}(G_i = g) f_i^{(t)}(X_i | G_i = g, Y_i = 1)}{\sum_y \pi_i^{(t)}(G_i = g) f_i^{(t)}(X_i | G_i = g, Y_i = y)}, \end{aligned} \tag{4.142}$$

$$\mathbb{E}^{(t)}(A_i | G_i = g, D_i, Y_i = 1, Q_i) = \frac{1 - S_i^{(t)}(Q_i | G_i = g, Y_i = 1)}{S_i^{(t)}(Q_i | G_i = g, Y_i = 1)}, \tag{4.143}$$

$$\mathbb{P}^{(t)}(T_{i1} = t_k | G_i = g, D_i, Y_i = 1, Q_i) = \frac{\mathbb{1}(t_k < Q_i) f_i^{(t)}(t_k | G_i = g, Y_i = 1)}{\sum_{h: t_h < Q_i} f_i^{(t)}(t_h | G_i = g, Y_i = 1)}, \tag{4.144}$$

when $g = \text{NN}$.

- for $i \in$ O(1, ?),

$$\begin{aligned} & \mathbb{P}^{(t)}(G_i = g | D_i = 1, T_i > X_i, Q_i) \\ = \; & \frac{\sum_y p_i^{(t)}(g) \, \mathbb{P}^{(t)}(Y_i = y | G_i = g) S_i^{(t)}(X_i | G_i = g, Y_i = y)}{\sum_g \sum_y p_i^{(t)}(g) \, \mathbb{P}^{(t)}(Y_i = y | G_i = g) S_i^{(t)}(X_i | G_i = g, Y_i = y)}, \end{aligned} \tag{4.145}$$

$$\begin{aligned} & \mathbb{P}^{(t)}(G_i = g|D_i = 1, Y_i = y, T_i > X_i, Q_i) \\ = \quad & \frac{p_i^{(t)}(g)\,\mathbb{P}^{(t)}(Y_i = y|G_i = g)S_i^{(t)}(X_i|G_i = g, Y_i = y)}{\sum_g p_i^{(t)}(g)\,\mathbb{P}^{(t)}(Y_i = y|G_i = g)S_i^{(t)}(X_i|G_i = g, Y_i = y)}, \end{aligned} \tag{4.146}$$

for any g,

$$\mathbb{P}^{(t)}(Y_i = 1|G_i = g, D_i = 1, T_i > X_i, Q_i) = 0, \tag{4.147}$$

$$\mathbb{E}^{(t)}(A_i|G_i = g, D_i = 1, M_i = 0, Y_i, Q_i) = 0, \tag{4.148}$$

$$\mathbb{P}^{(t)}(T_{i1} = t_k|G_i = g, D_i, Y_i = 1, Q_i) = 0, \tag{4.149}$$

when $g = \mathrm{SS}, \mathrm{NS}$,

$$\begin{aligned} & \mathbb{P}^{(t)}(Y_i = 1|G_i = g, D_i = 1, T_i > X_i, Q_i) \\ = \quad & \frac{\pi_i^{(t)}(G_i = g)S_i^{(t)}(X_i|G_i = g, Y_i = 1)}{\sum_y \pi_i^{(t)}(G_i = g)S_i^{(t)}(X_i|G_i = g, Y_i = y)}, \end{aligned} \tag{4.150}$$

$$\mathbb{E}^{(t)}(A_i|G_i = g, D_i, Y_i = 1, Q_i) = \frac{1 - S_i^{(t)}(Q_i|G_i = g, Y_i = 1)}{S_i^{(t)}(Q_i|G_i = g, Y_i = 1)}, \tag{4.151}$$

$$\mathbb{P}^{(t)}(T_{i1} = t_k|G_i = g, D_i, Y_i = 1, Q_i) = \frac{\mathbb{1}(t_k < Q_i) f_i^{(t)}(t_k|G_i = g, Y_i = 1)}{\sum_{h: t_h < Q_i} f_i^{(t)}(t_h|G_i = g, Y_i = 1)}, \tag{4.152}$$

when $g = \mathrm{NN}$.

$$\begin{aligned}
& \mathbb{E}^{(t)}\left[Y_i \sum_{j=1}^{A_i} \mathbb{1}(T_{ij}=t_k)\middle| G_i=g, D_i, T_i=X_i, Q_i\right] \\
= \ & \mathbb{P}^{(t)}(Y_i=1|G_i=g, D_i, T_i=X_i, Q_i)\,\mathbb{E}^{(t)}(A_i|G_i=g, D_i, Y_i=1, Q_i) \\
& \cdot \mathbb{P}^{(t)}(T_{i1}=t_k|G_i=g, D_i, Y_i=1, Q_i),
\end{aligned} \tag{4.153}$$

$$\begin{aligned}
& \mathbb{E}^{(t)}\left[(1-Y_i) \sum_{j=1}^{A_i} \mathbb{1}(T_{ij}=t_k)\middle| G_i=g, D_i, T_i=X_i, Q_i\right] \\
= \ & (1-\mathbb{P}^{(t)}(Y_i=1|G_i=g, D_i, T_i=X_i, Q_i))\,\mathbb{E}^{(t)}(A_i|G_i=g, D_i, Y_i=0, Q_i) \\
& \cdot \mathbb{P}^{(t)}(T_{i1}=t_k|G_i=g, D_i, Y_i=0, Q_i).
\end{aligned} \tag{4.154}$$

$$\begin{aligned}
& \mathbb{E}^{(t)}\left[Y_i \sum_{j=1}^{A_i} \mathbb{1}(T_{ij}=t_k)\middle| G_i=g, D_i, T_i>X_i, Q_i\right] \\
= \ & \mathbb{P}^{(t)}(Y_i=1|G_i=g, D_i, T_i>X_i, Q_i)\,\mathbb{E}^{(t)}(A_i|G_i=g, D_i, Y_i=1, Q_i) \\
& \cdot \mathbb{P}^{(t)}(T_{i1}=t_k|G_i=g, D_i, Y_i=1, Q_i),
\end{aligned} \tag{4.155}$$

$$\begin{aligned}
& \mathbb{E}^{(t)}\left[(1-Y_i) \sum_{j=1}^{A_i} \mathbb{1}(T_{ij}=t_k)\middle| G_i=g, D_i, T_i>X_i, Q_i\right] \\
= \ & (1-\mathbb{P}^{(t)}(Y_i=1|G_i=g, D_i, T_i>X_i, Q_i))\,\mathbb{E}^{(t)}(A_i|G_i=g, D_i, Y_i=0, Q_i) \\
& \cdot \mathbb{P}^{(t)}(T_{i1}=t_k|G_i=g, D_i, Y_i=0, Q_i).
\end{aligned} \tag{4.156}$$

The Q function becomes

$$
\begin{aligned}
& Q(\theta|\theta^{(t)}) \\
= & \sum_{M_i=0,O_i=1} w_i^{IP} \sum_g \Big\{ \mathbb{P}_i^{(t)}(G_i=g)\log p_i^g + \mathbb{P}_i^{(t)}(G_i=g) Y_i \log \pi_i(G_i=g) \\
& + \mathbb{P}_i^{(t)}(G_i=g)(1-Y_i)\log(1-\pi_i(G_i=g)) \Big\} \\
+ & \sum_{M_i=0,O_i=0} w_i^{IP} \sum_g \Big\{ \mathbb{P}_i^{(t)}(G_i=g)\log p_i^g + \mathbb{P}_i^{(t)}(G_i=g,Y_i=1)\log \pi_i(G_i=g) \\
& + \mathbb{P}_i^{(t)}(G_i=g,Y_i=0)\log(1-\pi_i(G_i=g)) \Big\} \\
+ & \sum_{M_i=1,O_i=1} w_i^{IP} \sum_g \Big\{ \mathbb{P}_i^{(t)}(G_i=g)\log p_i^g + \mathbb{P}_i^{(t)}(G_i=g) Y_i \log \pi_i^{(t)}(G_i=g) \\
& + \mathbb{P}_i^{(t)}(G_i=g)(1-Y_i)\log(1-\pi_i(G_i=g)) + \mathbb{P}_i^{(t)}(G_i=g)\log f_i(X_i|G_i=g) \Big\} \\
+ & \sum_{M_i=1,O_i=0} w_i^{IP} \sum_g \Big\{ \mathbb{P}_i^{(t)}(G_i=g)\log p_i^g + \mathbb{P}_i^{(t)}(G_i=g,Y_i=1)\log \pi_i(G_i=g) \\
& + \mathbb{P}_i^{(t)}(G_i=g,Y_i=0)\log(1-\pi_i(G_i=g)) \\
& + \mathbb{P}_i^{(t)}(G_i=g,Y_i=1)\log f_i(X_i|Y_i=1,G_i=g) \\
& + \mathbb{P}_i^{(t)}(G_i=g,Y_i=0)\log f_i(X_i|Y_i=0,G_i=g) \Big\} \\
+ & \sum_{M_i=?,O_i=1} w_i^{IP} \sum_g \Big\{ \mathbb{P}_i^{(t)}(G_i=g)\log p_i^g + \mathbb{P}_i^{(t)}(G_i=g) Y_i \log \pi_i(G_i=g) \\
& + \mathbb{P}_i^{(t)}(G_i=g)(1-Y_i)\log(1-\pi_i(G_i=g)) + \mathbb{P}_i^{(t)}(G_i=g)\log S_i(X_i|G_i=g) \Big\} \\
+ & \sum_{M_i=?,O_i=0} w_i^{IP} \sum_g \Big\{ \mathbb{P}_i^{(t)}(G_i=g)\log p_i^g + \mathbb{P}_i^{(t)}(G_i=g,Y_i=1)\log \pi_i(G_i=g) \\
& + \mathbb{P}_i^{(t)}(G_i=g,Y_i=0)\log(1-\pi_i(G_i=g)) \\
& + \mathbb{P}_i^{(t)}(G_i=g,Y_i=1)\log S_i(X_i|Y_i=1,G_i=g) \\
& + \mathbb{P}_i^{(t)}(G_i=g,Y_i=0)\log S_i(X_i|Y_i=0,G_i=g) \Big\} \\
+ & \sum_{i=1}^{n} w_i^{IP} \Big\{ \sum_g \mathbb{E}_i^{(t)} \Big[\sum_{j=1}^{A_i} Y_i \mathbb{1}(G_i^* = g, T_{i,j}=t_k) \Big] \log f_i(t_k|Y=y,G_i=g) \\
& + \sum_g \mathbb{E}_i^{(t)} \Big[\sum_{j=1}^{A_i} (1-Y_i) \mathbb{1}(G_i^* = g, T_{i,j}=t_k) \Big] \log f_i(t_k|Y=y,G_i=g) \Big\}.
\end{aligned}
$$

Inference

To conduct inference, we resort to Louis' formula. We compute the complete data scores and complete data negative hessians.

$$\frac{\partial}{\partial \gamma_{\text{NS}}} f_i = \mathbb{1}(G_i = \text{NS}) - \frac{e^{\gamma_{\text{NS}}}}{1 + e^{\gamma_{\text{NS}}} + e^{\gamma_{\text{NN}}}} = \mathbb{1}(G_i = \text{NS}) - \mathbb{P}(G_i = \text{NS}), \tag{4.157}$$

$$\frac{\partial}{\partial \gamma_{\text{NN}}} f_i = \mathbb{1}(G_i = \text{NN}) - \frac{e^{\gamma_{\text{NN}}}}{1 + e^{\gamma_{\text{NS}}} + e^{\gamma_{\text{NN}}}} = \mathbb{1}(G_i = \text{NS}) - \mathbb{P}(G_i = \text{NN}), \tag{4.158}$$

$$\begin{aligned} \triangledown \alpha f_i &= \sum_g \mathbb{1}(G_i = g) \Big[Z_i^g \Big(Y_i - \frac{e^{\alpha^\top Z_i^g}}{1 + e^{\alpha^\top Z_i^g}} \Big) \Big] && (4.159) \\ &= \sum_g \mathbb{1}(G_i = g) \Big[Z_i^g \Big(Y_i - \mathbb{P}(Y_i = 1 | Z_i^g) \Big) \Big], && (4.160) \end{aligned}$$

$$\begin{aligned} &\triangledown_\beta f_i && (4.161) \\ = \;& \mathbb{1}(G_i = \text{NS}) \Big\{ (1 - D_i) Z_i^{10} \Big[\Delta_i + A_i - \Big(\Lambda_{0i}(X_i) + \sum_{j=1}^{A_i} \Lambda(T_{ij}) \Big) \exp(\beta^\top Z_i^{10}) \Big] \Big\} && (4.162) \\ & + \mathbb{1}(G_i = \text{NN}) \Big\{ Z_i^{11} \Big[\Delta_i + A_i - \Big(\Lambda_{0i}(X_i) + \sum_{j=1}^{A_i} \Lambda(T_{ij}) \Big) \exp(\beta^\top Z_i^{11}) \Big] \Big\}, && (4.163) \end{aligned}$$

$$\begin{aligned} \frac{\partial}{\partial \lambda_k} f_i &= \mathbb{1}(G_i = \text{NS})(1 - D_i) \Big\{ \frac{\Delta_i \mathbb{1}(X_i = t_k) + \sum_{j=1}^{A_i} \mathbb{1}(T_{ij} = t_k)}{\lambda_k} \\ &\qquad - \exp(\beta^\top Z_i^{10}) \Big(\mathbb{1}(X_i \geq t_k) + \sum_{j=1}^{A_i} \mathbb{1}(T_{ij} \geq t_k) \Big) \Big] \Big\} && (4.164) \\ &\quad + \mathbb{1}(G_i = \text{NN}) \Big\{ \frac{\Delta_i \mathbb{1}(X_i = t_k) + \sum_{j=1}^{A_i} \mathbb{1}(T_{ij} = t_k)}{\lambda_k} \\ &\qquad - \exp(\beta^\top Z_i^{11}) \Big(\mathbb{1}(X_i \geq t_k) + \sum_{j=1}^{A_i} \mathbb{1}(T_{ij} \geq t_k) \Big) \Big] \Big\}, && (4.165) \end{aligned}$$

$$\frac{\partial^2}{\partial^2\gamma_{\mathrm{NS}}} f_i = -\frac{e^{\gamma_{\mathrm{NS}}}(1+e^{\gamma_{\mathrm{NN}}})}{(1+e^{\gamma_{\mathrm{NS}}}+e^{\gamma_{\mathrm{NN}}})^2} = -\mathbb{P}(G_i=\mathrm{NS})(1-\mathbb{P}(G_i=\mathrm{NS})), \tag{4.166}$$

$$\frac{\partial^2}{\partial^2\gamma_{\mathrm{NN}}} f_i = -\frac{e^{\gamma_{\mathrm{NN}}}(1+e^{\gamma_{\mathrm{NS}}})}{(1+e^{\gamma_{\mathrm{NS}}}+e^{\gamma_{\mathrm{NN}}})^2} = -\mathbb{P}(G_i=\mathrm{NN})(1-\mathbb{P}(G_i=\mathrm{NN})), \tag{4.167}$$

$$\frac{\partial^2}{\partial\gamma_{\mathrm{NS}}\partial\gamma_{\mathrm{NN}}} f_i = \frac{e^{\gamma_{\mathrm{NS}}+\gamma_{\mathrm{NN}}}}{(1+e^{\gamma_{\mathrm{NS}}}+e^{\gamma_{\mathrm{NN}}})^2} = \mathbb{P}(G_i=\mathrm{NS})\,\mathbb{P}(G_i=\mathrm{NN}), \tag{4.168}$$

$$\begin{aligned}
\bigtriangledown^2_{\alpha} f_i &= -\sum_g \mathbb{1}(G_i=g)\Big[Z_i^{g\otimes 2}\frac{e^{\alpha^\top Z_i^g}}{(1+e^{\alpha^\top Z_i^g})^2}\Big] && (4.169)\\
&= -\sum_g \mathbb{1}(G_i=g)\Big[Z_i^{g\otimes 2}\,\mathbb{P}(Y_i=1|Z_i^g)\,\mathbb{P}(Y_i=0|Z_i^g)\Big], && (4.170)
\end{aligned}$$

$$\begin{aligned}
\bigtriangledown^2_{\beta} f_i = &-\mathbb{1}(G_i=\mathrm{NS})\Big\{(1-D_i)Z_i^{10\otimes 2}\Big[\Big(\Lambda_{0i}(X_i)+\sum_{j=1}^{A_i}\Lambda(T_{ij})\Big)\exp(\beta^\top Z_i^{10})\Big]\Big\} && (4.171)\\
&-\mathbb{1}(G_i=\mathrm{NN})\Big\{Z_i^{11\otimes 2}\Big[\Big(\Lambda_{0i}(X_i)+\sum_{j=1}^{A_i}\Lambda(T_{ij})\Big)\exp(\beta^\top Z_i^{11})\Big]\Big\}, && (4.172)
\end{aligned}$$

$$\begin{aligned}
\frac{\partial^2}{\partial^2\lambda_k} f_i &= -\mathbb{1}(G_i=\mathrm{NS})(1-D_i)\Big\{\frac{\Delta_i\mathbb{1}(X_i=t_k)+\sum_{j=1}^{A_i}\mathbb{1}(T_{ij}=t_k)}{\lambda_k^2}\Big\} && (4.173)\\
&\quad -\mathbb{1}(G_i=\mathrm{NN})\Big\{\frac{\Delta_i\mathbb{1}(X_i=t_k)+\sum_{j=1}^{A_i}\mathbb{1}(T_{ij}=t_k)}{\lambda_k^2}\Big\}, && (4.174)
\end{aligned}$$

$$\begin{align}
& \frac{\partial^2}{\partial\beta\partial\lambda_k} f_i \tag{4.175}\\
= \quad & -\mathbb{1}(G_i = \mathrm{NS})(1 - D_i)\Big\{ Z_i^{10} \exp(\beta^\top Z_i^{10}) \Big(\mathbb{1}(X_i \geq t_k) + \sum_{j=1}^{A_i} \mathbb{1}(T_{ij} \geq t_k) \Big) \Big] \Big\} \tag{4.176}\\
& -\mathbb{1}(G_i = \mathrm{NN}) \Big\{ Z_i^{11} \exp(\beta^\top Z_i^{11}) \Big(\mathbb{1}(X_i \geq t_k) + \sum_{j=1}^{A_i} \mathbb{1}(T_{ij} \geq t_k) \Big) \Big] \Big\}. \tag{4.177}
\end{align}$$

All the other off-diagonal terms are zero.

4.8 Acknowledgement

Chapter 4, in full, has been submitted for publication of the material as it may appear in Journal of the American Statistical Association. Ying, Andrew; Xu, Ronghui; Chambers, Christina; Jones, Kenneth. Causal Effects on Birth Defects with Missing by Terathanasia. The dissertation author was the primary investigator and author of this paper.

Bibliography

[Aal80] Odd Aalen. A model for nonparametric regression analysis of counting processes. In *Mathematical statistics and probability theory*, pages 1–25. Springer, 1980.

[Aal89] Odd O Aalen. A linear regression model for the analysis of life times. *Statistics in medicine*, 8(8):907–925, 1989.

[ACC17] Ery Arias-Castro and Shiyun Chen. Distribution-free multiple testing. *Electronic Journal of Statistics*, 11(1):1983–2001, 2017.

[ACCTW17] Ery Arias-Castro, Rui M Castro, Ervin Tánczos, and Meng Wang. Distribution-free detection of structured anomalies: Permutation and rank-based scans. *Journal of the American Statistical Association*, (just-accepted), 2017.

[ACDH05] Ery Arias-Castro, David L Donoho, and Xiaoming Huo. Near-optimal detection of geometric objects by fast multiscale methods. *IEEE Transactions on Information Theory*, 51(7):2402–2425, 2005.

[AD52] Theodore W Anderson and Donald A Darling. Asymptotic theory of certain "goodness of fit" criteria based on stochastic processes. *The Annals of Mathematical Statistics*, 23(2):193–212, 1952.

[AF95] G L Anderson and T R Fleming. Model misspecification in proportional hazards regression. *Biometrika*, 82:527–541, 1995.

[AGG89] Richard Arratia, Larry Goldstein, and Louis Gordon. Two moments suffice for poisson approximations: the chen-stein method. *The Annals of Probability*, 17(1):9–25, 1989.

[AIR96] Joshua D Angrist, Guido W Imbens, and Donald B Rubin. Identification of causal effects using instrumental variables. *Journal of the American statistical Association*, 91(434):444–455, 1996.

[Ald13] David Aldous. *Probability Approximations via the Poisson Clumping Heuristic*, volume 77. Springer Science & Business Media, 2013.

[BAHR$^+$05] Anna Bill-Axelson, Lars Holmberg, Mirja Ruutu, Michael Häggman, Swen-Olof Andersson, Stefan Bratell, Anders Spångberg, Christer Busch, Stig Nordling, Hans Garmo, Juni Palmgren, Hans-Olov Adami, Bo Johan Norlén, and Jan-Erik Johansson. Radical prostatectomy versus watchful waiting in early prostate cancer. *New England journal of medicine*, 352(19):1977–1984, 2005.

[Ber45] WF Berg. Aggregates in one-and two-dimensional random distributions. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 36(256):337–346, 1945.

[BH95] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995.

[BHC88] J Bretagnolle and C Huber-Carol. Effects of omitting covariates in Cox's model for survival data. *Scandinavian Journal of Statistics*, 15:125–138, 1988.

[BJ79] Robert H Berk and Douglas H Jones. Goodness-of-fit test statistics that dominate the kolmogorov statistics. *Probability Theory and Related Fields*, 47(1):47–59, 1979.

[BKB$^+$93] Peter J Bickel, Chris AJ Klaassen, Peter J Bickel, Ya'acov Ritov, J Klaassen, Jon A Wellner, and YA'Acov Ritov. *Efficient and adaptive estimation for semiparametric models*, volume 4. Johns Hopkins University Press Baltimore, 1993.

[BN91] Julian Besag and James Newell. The detection of clusters in rare diseases. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 154(1):143–155, 1991.

[BR60] Raghu Raj Bahadur and R Ranga Rao. On deviations of the sample mean. *Ann. Math. Statist*, 31(4):1015–1027, 1960.

[BW07] Norman E Breslow and Jon A Wellner. Weighted likelihood for semiparametric models and two-phase stratified samples, with application to Cox regression. *Scandinavian Journal of Statistics*, 34(1):86–102, 2007.

[CAC17] Shiyun Chen and Ery Arias-Castro. Sequential multiple testing. *arXiv preprint arXiv:1705.10190*, 2017.

[Che07] Hua Yun Chen. A semiparametric odds ratio model for measuring association. *Biometrics*, 63(2):413–421, 2007.

[CJJ11] Tony T Cai, X Jessie Jeng, and Jiashun Jin. Optimal detection of heterogeneous and heteroscedastic mixtures. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(5):629–662, 2011.

[CJL07] T Tony Cai, Jiashun Jin, and Mark G Low. Estimation and confidence sets for sparse normal mixtures. *The Annals of Statistics*, 35(6):2421–2449, 2007.

[CL06] Hock Peng Chan and Tze Leung Lai. Maxima of asymptotically gaussian random fields and moderate deviation approximations to boundary crossing probabilities of sums of random variables with multidimensional indices. *The Annals of Probability*, 34(1):80–121, 2006.

[Cra38] Harald Cramér. *Les sommes et les fonctions de variables aléatoires*, volume 736. Hermann, 1938.

[CSH11] Bing Cai, Dylan S Small, and Thomas R Ten Have. Two-stage instrumental variable methods for estimating the causal odds ratio: Analysis of bias. *Statistics in medicine*, 30(15):1809–1824, 2011.

[CW14] T Tony Cai and Yihong Wu. Optimal detection of sparse mixtures against a given null distribution. *IEEE Transactions on Information Theory*, 60(4):2217–2232, 2014.

[CYAC18] Shiyun Chen, Andrew Ying, and Ery Arias-Castro. A scan procedure for multiple testing. *arXiv preprint arXiv:1808.00631*, 2018.

[DE56] DA Darling and P Erdös. A limit theorem for the maximum of normalized sums of independent random variables. *Duke Mathematical Journal*, 23(1):143–155, 1956.

[DJ98] Ralph B D'Agostino Jr. Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Statistics in medicine*, 17(19):2265–2281, 1998.

[DJ04] David Donoho and Jiashun Jin. Higher criticism for detecting sparse heterogeneous mixtures. *The Annals of Statistics*, 32(3):962–994, 2004.

[Eic79] F Eicker. The asymptotic distribution of the suprema of the standardized empirical processes. *The Annals of Statistics*, 7(1):116–138, 1979.

[ER04] Michael Elashoff and Louise Ryan. An em algorithm for estimating equations. *Journal of Computational and Graphical Statistics*, 13(1):48–65, 2004.

[Far82] V T Farewell. The use of mixture models for the analysis of survival data with long-term survivors. *Biometrics*, 38:1041–1046, 1982.

[Far86] V T Farewell. Mixture models in survival analysis: Are they worth the risk? 14:257–262, 1986.

[FG99] Jason P Fine and Robert J Gray. A proportional hazards model for the subdistribution of a competing risk. *Journal of the American statistical association*, 94(446):496–509, 1999.

[FH11] Thomas R Fleming and David P Harrington. *Counting Processes and Survival Analysis*, volume 169. John Wiley & Sons, 2011.

[FMPR12] Paolo Frumento, Fabrizia Mealli, Barbara Pacini, and Donald B Rubin. Evaluating the effect of training on wages in the presence of noncompliance, nonemployment, and missing outcome data. *Journal of the American Statistical Association*, 107(498):450–466, 2012.

[FNA95] I Ford, J. Norrie, and S Ahmadi. Model inconsistency, illustrated by the Cox proportional hazards model. *Statistics in Medicine*, 14:735–746, 1995.

[FR02a] C E Frangakis and Donald B Rubin. Principal stratification in causal inference. *Biometrics*, 58:21–29, 2002.

[FR02b] Constantine E Frangakis and Donald B Rubin. Principal stratification in causal inference. *Biometrics*, 58(1):21–29, 2002.

[GB12] Joseph Glaz and Narayanaswamy Balakrishnan. *Scan Statistics and Applications*. Springer Science & Business Media, 2012.

[GC11] Jean Dickinson Gibbons and Subhabrata Chakraborti. *Nonparametric Statistical Inference*. Springer, 2011.

[GF17] Veronika Gontscharuk and Helmut Finner. Asymptotics of goodness-of-fit tests based on minimum p-value statistics. *Communications in Statistics-Theory and Methods*, 46(5):2332–2342, 2017.

[GK18] Joseph Glaz and Markos V. Koutras, editors. *Handbook of Scan Statistics*. Springer, New York, 2018+.

[GNW$^+$01] Joseph Glaz, Joseph I Naus, Sylvan Wallenstein, Sylvan Wallenstein, and Joseph I Naus. *Scan Statistics*. Springer, 2001.

[GPW09] Joseph Glaz, Vladimir Pozdnyakov, and Sylvan Wallenstein. *Scan Statistics: Methods and Applications*. Springer Science & Business Media, 2009.

[GTTR12] M Maria Glymour, Eric J Tchetgen Tchetgen, and James M Robins. Credible mendelian randomization studies: approaches for evaluating the instrumental variable assumptions. *American journal of epidemiology*, 175(4):332–339, 2012.

[GW02] Christopher Genovese and Larry Wasserman. Operating characteristics and extensions of the false discovery rate procedure. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3):499–517, 2002.

[GW$^+$04] Christopher Genovese, Larry Wasserman, et al. A stochastic process approach to false discovery control. *The Annals of Statistics*, 32(3):1035–1061, 2004.

[GWP84] M H Gail, S Wieand, and S Piantadosi. Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates. *Biometrika*, 71:431–444, 1984.

[HBR01] M A Hernan, B Brumback, and J M Robins. Marginal structural models to estimate the joint causal effect of nonrandomized treatments. *Journal of the American Statistical Association*, 96:440–448, 2001.

[HCX18] Jue Hou, Christina D Chambers, and Ronghui Xu. A nonparametric maximum likelihood approach for survival data with observed cured subjects, left truncation and right-censoring. *Lifetime Data Analysis*, 24(4):612–651, 2018.

[HDL$^+$16] Freddie C Hamdy, Jenny L Donovan, J Athene Lane, Malcolm Mason, Chris Metcalfe, Peter Holding, Michael Davis, Tim J Peters, Emma L Turner, Richard M Martin, Jon Oxley, Mary Robinson, John M.B. Staffurth, Eleanor Walsh, Prasad Bollina, James Catto, Andrew Doble, Alan Doherty, David Gillatt, Roger Kockelbergh, Howard Kynaston, Alan Paul, Philip Powell, Stephen Prescott, Derek J. Rosario, Edward Rowe, and David E. Neal. 10-year outcomes after monitoring, surgery, or radiotherapy for localized prostate cancer. *New England Journal of Medicine*, 375(15):1415–1424, 2016.

[HL04] J W Hogan and T Lancaster. Instrumental variables and inverse probability weighting for causal inference from longitudinal observational studies. *Statistical Methods in Medical Research*, 13:17–48, 2004.

[HMD$^+$04] R. Heffernan, F. Mostashari, D. Das, A. Karpati, M. Kulldorff, and D. Weiss. Syndromic surveillance in public health practice, New York City. *Emerging Infectious Diseases*, 10(5):858–864, 2004.

[Hol86] Paul W Holland. Statistics and causal inference. *Journal of the American statistical Association*, 81(396):945–960, 1986.

[HR09] Peter J Huber and Elvezio M Ronchetti. *Robust Statistics*. John Wiley & Sons, 2009.

[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*, 102(23):1780–1793, 2010.

[Ing97] Yuri I Ingster. Some problems of hypothesis testing leading to infinitely divisible distributions. *Mathematical Methods of Statistics*, 6(1):47–69, 1997.

[Jae79] D Jaeschke. The asymptotic distribution of the supremum of the standardized empirical distribution function on subintervals. *The Annals of Statistics*, 7(1):108–115, 1979.

[JLF18] B Jiang, Jialiang Li, and Jason Fine. On two-step residual inclusion estimator for instrumental variable additive hazards model. *Biostatistics and Epidemiology*, 2(1):47–60, 2018.

[Joh93] Søren Johansen. An extension of Cox's regression model. *International Statistical Review*, 51:258–262, 1993.

[JSD$^+$05] Jiashun Jin, J-L Starck, David L Donoho, Nabila Aghanim, and Olivier Forni. Cosmological non-Gaussian signature detection: Comparing performance of different statistical tests. *EURASIP Journal on Advances in Signal Processing*, 2005(15):297184, 2005.

[Kab11] Zakhar Kabluchko. Extremes of the standardized gaussian noise. *Stochastic Processes and their Applications*, 121(3):515–533, 2011.

[KC92] Anthony YC Kuk and Chen-Hsin Chen. A mixture model combining logistic regression with proportional hazards regression. *Biometrika*, 79(3):531–541, 1992.

[KMW18] Claudia König, Axel Munk, and Frank Werner. Multidimensional multiscale scanning in exponential families: Limit theory and statistical consequences. *arXiv preprint arXiv:1802.07995*, 2018.

[Kol33] Andrey Kolmogorov. Sulla determinazione empirica di una legge di distribuzione. *Giornale dell'Istituto Italiano degli Attuari*, 4:89–91, 1933.

[Kos08] M. R. Kosorok. *Introduction to Empirical Processes and Semiparametric Inference*. Springer, 2008.

[KPJ$^+$17] Minjin Kim, Myunghee Cho Paik, Jiyeong Jang, Ying K Cheung, Joshua Willey, Mitchell SV Elkind, and Ralph L Sacco. Cox proportional hazards models with left truncation and time-varying coefficient: Application of age at event as outcome in cohort studies. *Biometrical Journal*, 59(3):405–419, 2017.

[Kul97] Martin Kulldorff. A spatial scan statistic. *Communications in Statistics: Theory and Methods*, 26(6):1481–1496, 1997.

[KW14] Zakhar Kabluchko and Yizao Wang. Limiting distribution for the maximal standardized increment of a random walk. *Stochastic Processes and their Applications*, 124(9):2824–2867, 2014.

[LFB15] Jialiang Li, Jason Fine, and Alan Brookhart. Instrumental variable additive hazards models. *Biometrics*, 71(1):122–130, 2015.

[LHS$^+$08] Debbie A Lawlor, Roger M Harbord, Jonathan AC Sterne, Nic Timpson, and George Davey Smith. Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. *Statistics in medicine*, 27(8):1133–1163, 2008.

[LN80] T Lancaster and S Nickell. The analysis of re-employment probabilities for the unemployed. *Journal of the Royal Statistical Society, Series A*, 143:141–165, 1980.

[LR19] Roderick JA Little and Donald B Rubin. *Statistical Analysis with Missing Data*. John Wiley & Sons, 2019.

[LXL17] Wanxing Li, Xiaoming Xue, and Yonghong Long. An additive subdistribution hazards model for competing risks data. *Communications in Statistics-Theory and Methods*, 46:11667–11687, 2017.

[LY94] DY 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*, 91:331–343, 2004.

[Mac48] C Mack. An exact formula for $q_k(n)$, the probable number of k-aggregates in a random distribution of n points. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 39(297):778–790, 1948.

[MNS16] Amit Moscovich, Boaz Nadler, and Clifford Spiegelman. On the exact Berk-Jones statistics and their p-value calculation. *Electronic Journal of Statistics*, 10(2):2329–2354, 2016.

[MR10] Thomas Mikosch and Alfredas Račkauskas. The limit distribution of the maximum increment of a random walk with regularly varying jump size distribution. *Bernoulli*, 16(4):1016–1038, 2010.

[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.

[MV13] T Martinussen and S. Vansteelandt. On collapsibility and confounding bias in Cox and Aalen regression models. *Lifetime Data Analysis*, 19:279–296, 2013.

[Nau65] Joseph I Naus. The distribution of the size of the maximum cluster of points on a line. *Journal of the American Statistical Association*, 60(310):532–538, 1965.

[Ney23] Jersey Neyman. Sur les applications de la théorie des probabilités aux experiences agricoles: Essai des principes. *Roczniki Nauk Rolniczych*, 10:1–51, 1923.

[NG17] V Q Nguyen and D L Gillen. Censoring-robust estimation in observational survival studies: Assessing the relative effectiveness of vascular access type on patency among end-stage renal disease patients. *Statistics in Biosciences*, 9:406–430, 2017.

[Pea01] Judea Pearl. Direct and indirect effects. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, pages 411–420. Morgan Kaufmann Publishers Inc., 2001.

[Pet95] Valentin V Petrov. Limit theorems of probability theory: sequences of independent random variables. Technical report, Oxford, New York, 1995.

[PWM18] Katharina Proksch, Frank Werner, and Axel Munk. Multiscale scanning in inverse problems. *The Annals of Statistics*, 46(6B):3569–3602, 2018.

[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*, 106:1434–1449, 2011.

[QW73] Clifford Qualls and Hisao Watanabe. Asymptotic properties of gaussian random fields. *Transactions of the American Mathematical Society*, 177:155–171, 1973.

[RG92] James M Robins and Sander Greenland. Identifiability and exchangeability for direct and indirect effects. *Epidemiology*, 3(2):143–155, 1992.

[RMM$^+$17] Greg Ridgeway, Dan McCaffrey, Andrew Morral, Beth Ann Griffin, and Lane Burgette. *twang: Toolkit for Weighting and Analysis of Nonequivalent Groups*, 2017. R package version 1.5.

[Ros84] Paul R Rosenbaum. The consequences of adjustment for a concomitant variable that has been affected by the treatment. *Journal of the Royal Statistical Society: Series A (General)*, 147(5):656–666, 1984.

[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.

[Rub74] Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701, 1974.

[Rub96] Donald B Rubin. Multiple imputation after 18+ years. *Journal of the American statistical Association*, 91(434):473–489, 1996.

[Rub04] Donald B Rubin. *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons, 2004.

[RW00] James M Robins and Naisyin Wang. Inference for imputation estimators. *Biometrika*, 87(1):113–124, 2000.

[SAC16] James Sharpnack and Ery Arias-Castro. Exact asymptotics for the scan statistic and fast alternatives. *Electronic Journal of Statistics*, 10(2):2641–2684, 2016.

[SFW⁺07] Thérèse A Stukel, Elliott S Fisher, David E Wennberg, David A Alter, Daniel J Gottlieb, and Marian J Vermeulen. Analysis of observational studies in the presence of treatment selection bias: effects of invasive cardiac management on AMI survival using propensity score and instrumental variable methods. *Journal of the American Medical Association*, 297(3):278–285, 2007.

[Sil45] Ludwik Silberstein. The probable number of aggregates in distributions of points. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 36(256):319–336, 1945.

[SK86] C. A. Struthers and J. D. Kalbfleisch. Misspecified proportional hazards model. *Biometrika*, 73:363–369, 1986.

[SSD⁺49] Samuel A Stouffer, Edward A Suchman, Leland C DeVinney, Shirley A Star, and Robin M Williams Jr. *The American Soldier, Vol 1: Adjustment During Army Life*. Princeton University Press, 1949.

[ST00] Judy P Sy and Jeremy MG Taylor. Estimation in a cox proportional hazards cure model. *Biometrics*, 56(1):227–236, 2000.

[SV95] David Siegmund and ES Venkatraman. Using the generalized likelihood ratio statistic for sequential detection of a change-point. *The Annals of Statistics*, 23(1):255–271, 1995.

[TBR08] Joseph V Terza, Anirban Basu, and Paul J Rathouz. Two-stage residual inclusion estimation: addressing endogeneity in health econometric modeling. *Journal of health economics*, 27(3):531–543, 2008.

[TG00] T.M. Therneau and P.M. Grambsch. *Modeling Survival Data: Extending the Cox Model*. Springer Verlag, New York, USA, 2000.

[Tip31] Leonard Henry Caleb Tippett. *Methods of Statistics*. Williams Norgate: London, 1931.

[TTRR09] Eric J Tchetgen Tchetgen, James M Robins, and Andrea Rotnitzky. On doubly robust estimation in a semiparametric odds ratio model. *Biometrika*, 97(1):171–180, 2009.

[TTWV⁺15] Eric J Tchetgen Tchetgen, Stefan Walter, Stijn Vansteelandt, Torben Martinussen, and Maria Glymour. Instrumental variable estimation in a survival context. *Epidemiology (Cambridge, Mass.)*, 26(3):402, 2015.

[Var89] Y Vardi. Multiplicative censoring, renewal processes, deconvolution and decreasing density: nonparametric estimation. *Biometrika*, 76:751–761, 1989.

[VdV00] Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

[VDVW96] Aad W Van Der Vaart and Jon A Wellner. Weak convergence. In *Weak convergence and empirical processes*, pages 16–28. Springer, 1996.

[WFW14] Elizabeth J Williamson, Andrew Forbes, and Ian R White. Variance reduction in randomised trials by inverse probability weighting using the propensity score. *Statistics in medicine*, 33(5):721–737, 2014.

[WMLC12] Elizabeth Williamson, Ruth Morley, Alan Lucas, and James Carpenter. Propensity scores: from naive enthusiasm to intuitive understanding. *Statistical methods in medical research*, 21(3):273–293, 2012.

[WR98] Naisyin Wang and James M Robins. Large-sample theory for parametric multiple imputation procedures. *Biometrika*, 85(4):935–948, 1998.

[XC11] R Xu and C Chambers. A sample size calculation for spontaneous abortion in observational studies. *Reproductive Toxicology*, 32:490–493, 2011.

[XHC18] Ronghui Xu, Jue Hou, and Christina D Chambers. The impact of confounder selection in propensity scores when applied to prospective cohort studies in pregnancy. *Reproductive Toxicology*, 78:75–80, 2018.

[XHSC19] Ronghui Xu, Gordon Honerkamp-Smith, and Christina D Chambers. Statistical sensitivity analysis for the estimation of fetal alcohol spectrum disorders prevalence. *Reproductive Toxicology*, 86:62–67, 2019.

[XO00] R. Xu and J O'Quigley. Estimating average regression effect under non-proportional hazards. *Biostatistics*, 1:423–439, 2000.

[ZBL$^+$15] Ge Zhang, Jonas Bacelis, Candice Lengyel, Kari Teramo, Mikko Hallman, Øyvind Helgeland, Stefan Johansson, Ronny Myhre, Verena Sengpiel, Pål Rasmus Njølstad, Bo Jacobsson, and Louis Muglia. Assessing the causal relationship of maternal height on birth size and gestational age at birth: a mendelian randomization analysis. *PLoS Med*, 12(8):e1001865, 2015.

[ZDHZ17] Cheng Zheng, Ran Dai, Parameswaran N Hari, and Mei-Jie Zhang. Instrumental variable with competing risk model. *Statistics in Medicine*, 36:1240–1255, 2017.

[ZVVS19] Johan Zetterqvist, Karel Vermeulen, Stijn Vansteelandt, and Arvid Sjölander. Doubly robust conditional logistic regression. *Statistics in Medicine*, 38(23):4749–4760, 2019.