UNIVERSITY OF CALIFORNIA SAN DIEGO

Statistical Robustness - Distributed Linear Regression, Informative Censoring, Causal Inference, and Non-Proportional Hazards

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

in

Biostatistics

by

Jiyu Luo

Committee in charge:

Professor Ronghui Xu, Chair
Professor Ery Arias-Castro
Professor Loki Natarajan
Professor Siddharth Singh

2023

The Dissertation of Jiyu Luo is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California

2023

DEDICATION

To my wife, Mom, Dad, and my daughter.

EPIGRAPH

There are no routine statistical questions, only questionable statistical routines.

—D. R. Cox

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

Chapter 2, in full, is a reprint of the material as it may appear in Luo, Jiyu, Sun, Qiang; and Zhou, Wen-Xin. (2022). *Distributed adaptive Huber regression*. Computational Statistics and Data Analysis, 169, 107419. The dissertation author was the primary investigator and author of this paper.

Chapter 3, for the most part, is a reprint of the material as it may appear in Luo, Jiyu; Xu, Ronghui. (2022). *Doubly Robust Inference for Hazard Ratio under Informative Censoring with Machine Learning*. arXiv preprint arXiv:2206.02296. The dissertation author was the primary investigator and author of this paper.

Chapter 4 contains material being prepared for submission and is co-authored with Denise Rava, Jelena Bradic, and Ronghui Xu. The dissertation author is the primary investigator/author of this paper.

Chapter 5 contains material being prepared for submission and is co-authored with Ronghui Xu. The dissertation author is the primary investigator/author of this paper.

VITA

2013-2016	B. A. in Mathematics , University of Cambridge
2017-2019	M. S. in Statistics, University of Pennsylvania
2019-2023	Ph. D. in Biostatistics, University of California San Diego

PUBLICATIONS

Singh, S., Kim, J., **Luo, J.**, Paul, P., Rudrapatna, V., Park, S., Zheng, K., Syal, G., Ha, C., Fleshner, P., McGovern, D., Sauk, J.S., Limketkai, B., Dulai, P.S., Boland, B.S., Eisenstein, S., Ramamoorthy, S., Melmed, G., Mahadevan, U. and Sandborn, W.J. (2022). *Comparative Safety and Effectiveness of Biologic Therapy for Crohn's Disease: A CA-IBD Cohort Study*. Clinical Gastroenterology and Hepatology.

Gu, P., **Luo, J.**, Kim, J., Paul, P., Limketkai, B., Sauk, J.S., Park, S., Parekh, N., Zheng, K., Rudrapatna, V., Syal, G., Ha, C., McGovern, D.P., Melmed, G.Y., Fleshner, P., Eisenstein, S., Ramamoorthy, S., Dulai, P.S., Boland, B.S. and Grunvald, E. (2022). *Effect of Obesity on Risk of Hospitalization, Surgery, and Serious Infection in Biologic-Treated Patients With Inflammatory Bowel Diseases: A CA-IBD Cohort Study*. The American Journal of Gastroenterology, 117(10), pp.1639–1647.

Luo, J. and Xu, R. (2022). *Doubly Robust Inference for Hazard Ratio under Informative Censoring with Machine Learning*. arXiv:2206.02296

Nguyen, N.H., **Luo, J.**, Paul, P., Kim, J., Syal, G., Ha, C., Rudrapatna, V., Park, S., Parekh, N., Zheng, K., Sauk, J.S., Limketkai, B., Fleshner, P., Eisenstein, S., Ramamoorthy, S., Melmed, G., Dulai, P.S., Boland, B.S., Mahadevan, U. and Sandborn, W.J. (2022). *Effectiveness and Safety of Biologic Therapy in Hispanic Vs Non-Hispanic Patients With Inflammatory Bowel Diseases: A CA-IBD Cohort Study. Clinical Gastroenterology and Hepatology*.

Gu, P., **Luo, J.**, Paul, P., Limketkai, B.N., Sauk, J.S., Park, S., Parekh, N.K., Zheng, K., Rudrapatna, V.A., Syal, G., Ha, C., Mcgovern, D.P.B., Melmed, G., Fleshner, P., Eisenstein, S., Ramamoorthy, S., Dulai, P., Boland, B., Mahadevan, U. and Ohno-Machado, L. (2022). *454: Obesity Is Not Associated With Increased Risk of Adverse Treatment Outcomes or Serious Infection in Inflammatory Bowel Diseases Patients Starting New Biologic Therapy*. Gastroenterology, 162(7), p.S–S-102.

Venkat, P., Nguyen, N.H., **Luo, J.**, Qian, A. and Singh, S., 2022. Mo1476: *Impact of Recurrent Hospitalization for Clostridioides difficile on Longitudinal Outcomes in Patients With Inflammatory Bowel Diseases*. Gastroenterology, 162(7), pp.S-780.

Nguyen, N.H., **Luo, J.**, Paul, P., Kim, J., Syal, G., Ha, C., Rudrapatna, V.A., Park, S., Parekh, N.K., Zheng, K. and Sauk, J.S., 2022. *Su1018: Unplanned Healthcare Utilization and Safety of Biologic Therapy in Hispanic vs. Non-Hispanic Patients With IBD*. Gastroenterology, 162(7), pp.S-482.

Luo, J., Sun, Q. and Zhou, W.-X. (2022). *Distributed Adaptive Huber Hegression.* Computational Statistics and Data Analysis, 169, p.107419.

Singh, S., Qian, A.S., Nguyen, N.H., Ho, S.K.M., **Luo, J.**, Jairath, V., Sandborn, W.J. and Ma, C. (2021). *Trends in U.S. Health Care Spending on Inflammatory Bowel Diseases, 1996-2016.* Inflammatory Bowel Diseases, 28(3), pp.364–372.

Holmer, A.K., **Luo, J.**, Park, S., Yang, J.Y., Nguyen, V.Q., Sofia, M.A., Ertem, F., Dueker, J.M., Petrov, J.C., Al Bawardy, B.F., Llano, E.M., Fudman, D., Joseph, D., Jangi, S., Russ, K., Khakoo, N.S., Damas, O., Barnes, E.L., Hong, S.J. and Zenger, C. (2021). *S697 Comparative Safety of Biologic Agents in Patients With Inflammatory Bowel Disease with Active or Recent Malignancy: A Multi-Center Cohort Study.* Official journal of the American College of Gastroenterology, 116, p.S316.

Rozich, J.J., **Luo, J.**, Dulai, P.S., Collins, A.E., Pham, L., Boland, B.S., Sandborn, W.J. and Singh, S. (2020). *Disease- and Treatment-related Complications in Older Patients With Inflammatory Bowel Diseases: Comparison of Adult-onset vs Elderly-onset Disease.* Inflammatory Bowel Diseases, 27(8), pp.1215–1223.

Meserve, J., **Luo, J.**, Zhu, W., Veeravalli, N., Bandoli, G., Chambers, C.D., Singh, A.G., Boland, B.S., Sandborn, W.J., Mahadevan, U. and Singh, S. (2021). *Paternal Exposure to Immunosuppressive and/or Biologic Agents and Birth Outcomes in Patients With Immune-Mediated Inflammatory Diseases.* Gastroenterology, 161(1), pp.107-115.e3.

Nguyen, N.H., **Luo, J.**, Ohno-Machado, L., Sandborn, W.J. and Singh, S. (2020). *Burden and Outcomes of Fragmentation of Care in Hospitalized Patients With Inflammatory Bowel Diseases: A Nationally Representative Cohort.* Inflammatory Bowel Diseases.

Meserve, J., **Luo, J.**, Zhu, W., Bandoli, G., Chambers, C.D., Singh, A.G., Boland, B.S., Sandborn, W.J., Mahadevan, U. and Singh, S., 2021. *613 Paternal Exposure to Immunosuppressive and/or Biologic Agents and Birth Outcomes in Patients With Immune-Mediated Inflammatory Diseases.* Gastroenterology, 160(6), pp.S-121.

ABSTRACT OF THE DISSERTATION

Statistical Robustness - Distributed Linear Regression, Informative Censoring, Causal Inference, and Non-Proportional Hazards

by

Jiyu Luo

Doctor of Philosophy in Biostatistics

University of California San Diego, 2023

Professor Ronghui Xu, Chair

Robustness broadly refers to the property of the statistical method being valid even when some of the model assumptions are violated. We investigate 4 types of statistical robustness under 4 different problem setups. Firstly, we consider linear regression under the distributed setting where data are stored in separate machines. When errors are subject to heavy-tailed and/or asymmetric errors, we develop a tail-robust distributed estimator that achieves a sub-Gaussian-type deviation bound without pooling all the data together and without assuming Gaussian errors. Moreover, the algorithm only transfers gradient in each step and is hence communication efficient. Secondly, we explore the two-group Cox proportional hazards (PH) model in a randomized study. When the non-informative censoring assumption no longer holds, the inverse probability of censoring weighting (IPCW) estimator helps correct the censoring bias by modeling the nuisance function for conditional censoring survival. To protect against the misspecification of the nuisance function, we propose an augmented IPCW (AIPCW) estimator which also models conditional failure survival. The AIPCW estimator is model double robust (DR) in that the estimator will be consistent and asymptotically normal (CAN) even when one of the root-n nuisance estimators is wrong. The estimator is also CAN if both nuisance functions are consistently estimated with their product error rate being faster than root-n. This so-called rate DR property allows us to make use of machine learning (ML) methods, which directly address the non-collapsibility of the Cox model. Thirdly, we extend the problem to observational data with the two-group survival following the marginal structural Cox model. In addition to the missingness due to censoring, we also need to deal with missingness coming from partial observations of the potential outcomes. By extending the AIPCW estimator to include the nuisance propensity score function, we develop an augmented IPW (AIPW) estimator that is again DR with respect to the models for failure time and for missing mechanisms. Lastly, we consider the scenario when the PH assumption fails and propose a causal estimand that is a weighted average of the time-varying log hazards ratio. We show that this estimand enjoys several desirable properties and can be estimated using the same AIPW estimator we proposed for the marginal structural Cox model. A method for plotting the time-varying log hazard ratio under observational data is also proposed.

Chapter 1

Introduction

In this proposal, we outline the motivation and the statistical background of the dissertation topic. The proposal focuses on developing statistically robust methods in various problem domains. Robustness in statistics refers to the ability of a statistical method or model to produce reliable and stable results even under violation of underlying assumptions. In this proposal, we investigate 3 types of robustness: tail robustness, model and rate double robustness (DR), and robustness against violation of proportional hazards. The definition of these will be made clear later when we discuss each of them in detail.

To begin, we explore in Chapter 2 the linear regression under a distributed setting where data can only be stored at different locations due to either storage limitations or privacy concerns. We seek to develop a tail-robust distributed algorithm in that it can achieve a centralized sub-Gaussian type error bound even when the error is not Gaussian and is subject to heavy-tailed and/or asymmetric errors with finite second moments. The algorithm is also communication efficient since it only transfers gradient information during each update. We then extend our methods to high-dimensional settings and construct robust confidence intervals. Optimization schemes are proposed that achieve faster computational speed and these results are tested through numerical studies.

Next, we explore in Chapter 3 the two-group Cox proportional hazards (PH) model in a randomized trial where the random censoring assumption could be violated. We allow the censoring to be informative, where the censoring time and the failure time is only independent when conditional on additional covariates. Traditionally, inverse probability of censoring weighting (IPCW) estimator is used, where a conditional model for the censoring time given covariates needs to be correctly specified. By augmenting the IPCW estimator, we develop a DR estimator with respect to the censoring model and a conditional outcome model. Specifically, the augmented IPCW (AIPCW) estimator is model DR in that it is consistent and asymptotically normal (CAN) as long as one of the two nuisance functions is correctly estimated at the root-n rate. Moreover, with cross-fitting, the estimator is also rate DR, making it CAN if both nuisance functions are correctly estimated at an arbitrarily slow rate as long as their product error rate is faster than root-n. This allows us to apply slower than root-n ML methods to estimate both nuisance functions, which also meets the challenge of the non-collapsibility of the Cox model.

In Chapter 4, we extend our model from a study with random treatment assignment to observational data and assume that the potential failure times follow the marginal structural Cox model. In addition to dealing with informative censoring, we also simultaneously address the bias caused by partial observations of the potential outcomes. The common inverse probability of the treatment weighting approach would be biased if the propensity score model is misspecified. To protect against this misspecification, we develop a model and rate DR estimator by augmenting with respect to both treatment and censoring. Here, the augmented IPW (AIPW) estimator is double robust with respect to the conditional outcome model and the two models for censoring and treatment estimator. We apply our proposed methods to a study of Japanese men in Hawaii followed since the 1960s to examine the effect of mid-life drinking on overall survival.

In Chapter 5, we consider the implications for a potential violation of the PH assumption models presented in Chapter 4. Although the Cox PH model is one of the most widely used models, the PH assumption rarely, if ever strictly holds in practice. To this end, we propose a causal estimand that has the property of being the weighted average of the time-varying log hazard ratio. We also show that this estimand recovers the time-fixed log hazard ratio under the marginal structural Cox model, has a simple relationship with treatment effects defined through the marginal structural transformation models, and enjoys the same estimating functions as that used for the marginal structural Cox model. This property allows us to construct IPW, AIPW, and cross-fitted AIPW estimators for it under observational data with informative censoring. In particular, the cross-fitted AIPW estimator enjoys the desirable model and rate DR properties. The time-varying hazard ratio itself is non-parametric and hard to estimate, but methods for plotting it have been proposed by Therneau and Grambsch (2000) under the randomization with random censoring setting. We fill this gap by proposing an AIPW plot that generalizes it to observational data with informative censoring. Lastly, we apply our proposed method to the International Non-Hodgkin's Lymphoma Prognostic Factors Project dataset.

Chapter 2

Distributed Adaptive Huber Regression

2.1 Abstract

Distributed data naturally arise in scenarios involving multiple sources of observations, each stored at a different location. Directly pooling all the data together is often prohibited due to limited bandwidth and storage, or due to privacy protocols. This paper introduces a new robust distributed algorithm for fitting linear regressions when data are subject to heavy-tailed and/or asymmetric errors with finite second moments. The algorithm only communicates gradient information at each iteration, and therefore is communication-efficient. Statistically, the resulting estimator achieves the centralized nonasymptotic error bound as if all the data were pooled together and came from a distribution with sub-Gaussian tails. Under a finite $(2+\delta)$-th moment condition, we derive a Berry-Esseen bound for the distributed estimator, based on which we construct robust confidence intervals. Numerical studies further confirm that compared with extant distributed methods, the proposed methods achieve near-optimal accuracy with low variability and better coverage with tighter confidence width.

2.2 Introduction

In many applications, there are a massive number of individual agents/organizations collecting data independently. Multiple-site research has brought the possibility of studying rare outcome that require larger sample sizes, accelerating more generalizable findings, and bringing together investigators with different expertise from various backgrounds (Sidransky et al., 2009). Due to limited resources, such as bandwidth and storage, or privacy concerns, researchers across different sites are only allowed to share summary statistics without allowing collaborating parties to access raw data (Wu et al., 2012). Moreover, the collected data may often be contaminated by high level of noise, and thus of low quality. For example, in the context of gene expression data analysis, it has been observed that some gene expression levels have kurtosis values much larger than 3, despite of the normalization methods used (Wang et al., 2015). It is therefore important to develop robust and distributed learning algorithms with controlled communication cost and desirable statistical performance, measured by both efficiency and robustness.

Distributed learning algorithms have received considerable attention for multi-source studies in the past decade. Due to privacy concerns, data collected at each source, such as node, sensor or organization, must remain local. The goal is to develop efficient statistical learning methods that allow shared analyses or summary statistics without sharing individual level data. The classical divide-and-conquer principle is based on aggregating local estimators, that is, estimators computed separately on local machines, to form a final estimator; see for example, Chen and Xie (2014), Li et al. (2013), Zhang et al. (2015), Zhao et al. (2016), Rosenblatt and Nadler (2016), Lee et al. (2017), Battey et al. (2018) and Volgushev et al. (2019), among many others. We refer to Huo and Cao (2019) for a more complete literature review. One-step averaging takes one communication round, and therefore is convenient and has minimal communication cost. However, in order for the averaging estimator to achieve to same convergence rate as the centralized estimator, each local machine must have access to at least $\sqrt{N}$ samples, where N is the total sample size. This limits the number of machines allowed in the communication network.

To overcome this barrier of one-step averaging, multi-round procedures have been proposed for distributed data analysis with a large number of local agents (Shamir et al., 2014; Wang et al., 2017; Jordan et al., 2019; Wang et al., 2019). For linear and generalized linear models, Wang et al. (2017) and Jordan et al. (2019) proposed multi-round distributed (penalized) M-estimators that achieve optimal rates of convergence under very mild constraints on the number of machines. Chen et al. (2019) studied an iterative algorithm with proper smoothing for quantile regression under memory constraint, which may also apply under distributed computing platform. Alternatively, Dobriban and Sheng (2018) proposed an iterative weighted parameter averaging scheme for distributed linear regression when the dimension is comparable to the sample size.

For linear models under data parallelism, most of the existing distributed algorithms work with the least squares method, either by (weighted) averaging local least squares estimators or iteratively minimizing shifted (penalized) least squares loss functions. From a robustness viewpoint, distributed least squares based method inherits the sensitivity (non-robustness) of its centralized counterpart to the tails of the error distributions, hence increasing the variability of the estimator. In this paper, we propose a robust distributed algorithm for linear regression with heavy-tailed errors. Our setup includes the heteroscedastic linear model with asymmetric errors, to which the least absolute deviation (LAD) regression does not naturally apply. Following the terminology in Catoni (2012), the type of "robustness" considered in this paper is quantified by nonasymptotic exponential deviation of the estimator versus polynomial tail of the error distribution. The ensuing procedure does sacrifice a fair amount of robustness to adversarial contamination of the data. The motivation of this work is different from and should not be confused with the classical notion of robust statistics (Huber and Ronchetti, 2009).

The distributed method is built upon the iterative, multi-round algorithm proposed by Wang et al. (2017) and Jordan et al. (2019), which only communicates gradient information at each round and therefore is communication-efficient. By a delicate choice of local and global robustifications parameters, the proposed estimator satisfies exponential-type deviation bounds when the errors only have finite variance. Specifically, we show that the distributed estimator, obtained by a few rounds of communications, achieves the optimal centralized deviation bound as if the data were pooled together and subject to sub-Gaussian errors. The robustification parameters are also self-tuned, making the algorithm computationally convenient. We further derive a Berry-Esseen bound for the distributed estimator, based on which we construct robust confidence intervals. Finally, we propose a distributed penalized adaptive Huber regression estimator for high-dimensional sparse models, and establish its (near-)optimal theoretical guarantees.

NOTATION: For each integer $k \geq 1$, we use $\mathbb{R}^k$ to denote the the k-dimensional Euclidean space. The inner product of two vectors $u = (u_1, \ldots, u_k)^{\mathrm{T}}, v = (v_1, \ldots, v_k)^{\mathrm{T}} \in \mathbb{R}^k$ is defined by $u^{\mathrm{T}} v = \langle u, v \rangle = \sum_{i=1}^k u_i v_i$. We use $\| \cdot \|_p$ $(1 \leq p \leq \infty)$ to denote the ℓ_p-norm in $\mathbb{R}^k$: $\|u\|_p = (\sum_{i=1}^k |u_i|^p)^{1/p}$ and $\|u\|_\infty = \max_{1 \leq i \leq k} |u_i|$. For any $k \times k$ symmetric matrix $A \in \mathbb{R}^{k \times k}$, $\|A\|_2$ is the operator norm of A. For a positive semidefinite matrix $A \in \mathbb{R}^{k \times k}$, $\| \cdot \|_A$ denotes the norm induced by A, that is, $\|u\|_A = \|A^{1/2} u\|_2$, $u \in \mathbb{R}^k$. Moreover, we use $\mathbb{S}^{k-1} = \{u \in \mathbb{R}^k : \|u\|_2 = 1\}$ to denote the unit sphere in $\mathbb{R}^k$. For two sequences of non-negative numbers $\{a_n\}_{n \geq 1}$ and $\{b_n\}_{n \geq 1}$, $a_n \lesssim b_n$ indicates that there exists a constant $C > 0$ independent of n such that $a_n \leq C b_n$; $a_n \gtrsim b_n$ is equivalent to $b_n \lesssim a_n$; $a_n \asymp b_n$ is equivalent to $a_n \lesssim b_n$ and $b_n \lesssim a_n$.

2.3 Distributed Adaptive Huber Regression

2.3.1 Distributed Huber regression with adaptive robustification parameters

Consider a linear regression model

$$y_i = x_i^{\mathrm{T}}\beta^* + \varepsilon_i, \quad \mathbb{E}(\varepsilon_i | x_i) = 0, \;\; i = 1, \ldots, N, \tag{2.1}$$

where $x_i = (x_{i1}, \ldots, x_{ip})^{\mathrm{T}}$ with $x_{i1} \equiv 1$ is the covariate for the ith individual, and $\beta^* \in \mathbb{R}^p$ is the underlying coefficient vector. This setting allows conditional heteroscedastic models, where ε_i can depend on x_i. For example, in a local-scale model we have $\varepsilon_i = \sigma(x_i)e_i$, where $\sigma(x_i)$ is a function of x_i, and e_i is independent of x_i. In the absence of normality assumption on the (conditional) error distribution, Huber's M-estimator (Huber, 1973) is one of the most widely used robust alternative to the least squares estimator. Given some $\tau > 0$, referred to as the robustification parameter, Huber's regression M-estimator for estimating β^* is defined as

$$\widehat{\beta} = \widehat{\beta}_\tau \in \mathop{\arg\min}_{\beta \in \mathbb{R}^p} \; \widehat{\mathcal{L}}_\tau(\beta) := \frac{1}{N}\sum_{i=1}^{N} \ell_\tau(y_i - x_i^{\mathrm{T}}\beta),$$

where $\ell_\tau(u) = 0.5u^2 I(|u| \le \tau) + (\tau|u| - 0.5\tau^2)I(|u| > \tau)$ is the Huber loss. Traditionally, τ is often chosen to be 1.345σ with σ either determined by a robust scale estimate or simultaneously estimated by solving a system of equations, in order to achieve 95% asymptotic relative efficiency while gaining robustness when there are contaminated or heavy-tailed symmetric errors (Bickel, 1975; Western, 1995). In the presence of asymmetric heavy-tailed errors, Fan et al. (2017) and Sun et al. (2020) proposed (regularized) adaptive Huber regression estimators with τ scaling with the sample size and parametric dimension, and established exponential-type deviation bounds when ε_i's only have finite $(1+\delta)$-th moments for some $0 < \delta \le 1$.

In the linear model (2.1), we allow heteroscedastic errors that are of the form $\varepsilon_i = \sigma(x_i)e_i$, where $\sigma(\cdot)$ is an unknown function on $\mathbb{R}^p$ and e_i is independent of x_i. When the error variables ε_i are heavy-tailed, asymmetric and have finite variance σ^2, Sun et al. (2020) showed that Huber's estimator $\widehat{\beta}_\tau$ with $\tau \asymp \sigma\sqrt{N/(p+\log N)}$, referred to as the adaptive Huber regression (AHR) estimator, exhibits sharp finite-sample deviation properties (Catoni, 2012), while the least squares estimator is far less concentrated around β^*. We say ε_i is heavy-tailed if it has infinite k-th absolute moment for some $k > 2$.

In the distributed setting, assume that the overall dataset $\{(y_i, x_i)\}_{i=1}^N$ is stored on m node machines, one central machine and $m-1$ local machines that connected to the central. For $j = 1, \ldots, m$, the jth machine stores a subsample of n_j observations, denoted by $\{(y_i, x_i)\}_{i \in I_j}$, and I_j's are disjoint index sets that satisfy $\cup_{j=1}^m I_j = \{1, \ldots, N\}$ and $N = \sum_{j=1}^m |I_j| = \sum_{j=1}^m n_j$. Without loss of generality, we assume $n_1 = \cdots = n_j = n$ and $N = n \cdot m$ is divisible by m. We thus refer to n as the local sample size. When the entire dataset is available, the optimal τ scales with the total sample size N and dimension p for optimal bias and robustness tradeoff. With decentralized data, each local machine only has access to a subsample, so that the "locally optimal" τ depends on the local sample size. This, however, will lead to sub-optimal bounds for the aggregated estimator because τ is not large enough to offset the bias. To parallelize AHR in a distributed setting without compromising statistical optimality, we introduce two robustification parameters τ and κ, referred to as the global and local robustifiation parameters, and define the global and local Huber loss functions as $\widehat{\mathcal{L}}_\tau(\beta) = (1/N)\sum_{i=1}^N \ell_\tau(y_i - x_i^{\mathrm{T}}\beta)$ and $\widehat{\mathcal{L}}_{j,\kappa}(\beta) = (1/n)\sum_{i \in I_j} \ell_\kappa(y_i - x_i^{\mathrm{T}}\beta)$ for $j = 1, \ldots, m$. Using this adaptive robustification procedure, we then extend the approximate Newton-type method (Shamir et al., 2014; Jordan et al., 2019) to robust regression with heavy-tailed skewed errors.

Starting with an initial estimator $\widetilde{\beta}^{(0)}$ of β^*, we define the shifted adaptive Huber loss

$$
\begin{aligned}
\widetilde{\mathcal{L}}(\beta) &= \widehat{\mathcal{L}}_{1,\kappa}(\beta) - \big\langle \nabla\widehat{\mathcal{L}}_{1,\kappa}(\widetilde{\beta}^{(0)}) - \nabla\widehat{\mathcal{L}}_{\tau}(\widetilde{\beta}^{(0)}), \beta \big\rangle \\
&= \widehat{\mathcal{L}}_{1,\kappa}(\beta) - \Big\langle \nabla\widehat{\mathcal{L}}_{1,\kappa}(\widetilde{\beta}^{(0)}) - \frac{1}{m}\sum_{j=1}^{m} \widehat{\mathcal{L}}_{j,\tau}(\widetilde{\beta}^{(0)}), \beta \Big\rangle, \ \ \beta \in \mathbb{R}^p. \qquad (2.2)
\end{aligned}
$$

Implicitly the shifted loss $\widetilde{\mathcal{L}}(\cdot)$ depends on both local and global robustification parameters κ and τ. It uses data available only on the first machine, used as the central machine, along with p-dimensional gradient vectors $\widehat{\mathcal{L}}_{j,\kappa}(\widetilde{\beta}^{(0)})$ $(j = 2, \ldots, m)$ that were sent from the remaining local machines. The ensuing one-step estimator is given by

$$
\widetilde{\beta}^{(1)} = \widetilde{\beta}^{(1)}_{\kappa,\tau} \in \operatorname*{arg\,min}_{\beta \in \mathbb{R}^p} \widetilde{\mathcal{L}}(\beta). \qquad (2.3)
$$

This procedure requires one communication round of $O(pm)$ bits, and thus is communication-efficient. To investigate the statistical properties of $\widetilde{\beta}^{(1)}$, we impose the following moment condition on the data generating process.

(C1). The predictor $x \in \mathbb{R}^p$ is sub-Gaussian: there exists $\upsilon_1 \geq (2\log 2)^{-1/2}$ such that $\mathbb{P}(|z^{\mathsf{T}}u| \geq \upsilon_1 t) \leq 2e^{-t^2/2}$ for every unit vector $u \in \mathbb{S}^{p-1}$ and $t \geq 0$, where $z = \Sigma^{-1/2}x$ and $\Sigma = \mathbb{E}(xx^{\mathsf{T}})$ is positive definite. Moreover, the regression error ε satisfies $\mathbb{E}(\varepsilon|x) = 0$ and $\mathbb{E}(\varepsilon^2|x) \leq \sigma^2$ almost surely.

For prespecified parameters $r, r_* > 0$, define the events

$$
\mathcal{E}_0(r) = \big\{\widetilde{\beta}^{(0)} \in \Theta(r)\big\} \quad \text{and} \quad \mathcal{E}_*(r_*) = \big\{\|\nabla\widehat{\mathcal{L}}_{\tau}(\beta^*)\|_{\Omega} \leq r_*\big\}, \qquad (2.4)
$$

where $\Theta(r) := \{\beta \in \mathbb{R}^p : \|\beta - \beta^*\|_{\Sigma} \leq r\}$ and $\Omega := \Sigma^{-1}$. Here r quantifies the statistical accuracy of the initial estimator $\widetilde{\beta}^{(0)}$, and r^* determines the estimation error of the centralized AHR estimator which essentially depends on the score $\nabla\widehat{\mathcal{L}}_{\tau}(\beta^*)$ with the global robustification parameter.

Theorem 1. *Assume Condition (C1) holds. For any $u > 0$, let the robustification parameters* *satisfy* $\tau \geq \kappa \asymp \sigma\sqrt{n/(p+u)}$, *and suppose the local sample size satisfies* $n \gtrsim p+u$. *Then, conditioned on the event* $\mathcal{E}_0(r_0) \cap \mathcal{E}_*(r_*)$ *with* $8r_* \leq r_0 \leq \sigma$, *the one-step estimator* $\widetilde{\beta}^{(1)}$ *defined in* (2.3) *satisfies*

$$\|\widetilde{\beta}^{(1)} - \beta^*\|_\Sigma \lesssim \sqrt{\frac{p+u}{n}} \cdot r_0 + r_* \quad \textit{and} \tag{2.5}$$

$$\|\widetilde{\beta}^{(1)} - \beta^* + \Sigma^{-1}\nabla\widehat{\mathcal{L}}_\tau(\beta^*)\|_\Sigma \lesssim \sqrt{\frac{p+u}{n}} \cdot r_0 \tag{2.6}$$

with probability at least $1-3e^{-u}$.

In the above theorem, the bound (2.5) reflects the delicate dependence of the one-step error on the initial error r_0 as well as the centralized error rate r_*. If we take $\widetilde{\beta}^{(0)}$ to be a local estimator constructed on a single local machine that has access to only n observations, we may expect a sub-optimal convergence rate $r_0 \asymp \sigma\sqrt{p/n}$. Moreover, it can be shown that $\|\nabla\widehat{\mathcal{L}}_\tau(\beta^*)\|_\Omega \lesssim \sigma\sqrt{p/N} + \sigma^2/\tau + \tau p/N$ with high probability, up to logarithmic factors; see Lemma 6 in the Supplementary Material. Hence, the choice of r_* corresponds to the optimal rate of convergence when the entire dataset is available and $\tau \asymp \sigma\sqrt{N/p}$. Under the prescribed sample size scaling $n \gtrsim p$, the one-step estimator $\widetilde{\beta}^{(1)}$ refines the statistical accuracy of $\widetilde{\beta}^{(0)}$ by a factor of order $\sqrt{p/n}$, which is strictly less than 1. We thus expect the multi-step estimator, with sufficiently many communication rounds, will achieve the optimal convergence rate obtainable on the entire dataset.

The proposed multi-round procedure for adaptive Huber regression is iterative, starting at iteration 0 with an initial estimate $\widetilde{\beta}^{(0)} \in \mathbb{R}^p$. At iteration $t \geq 1$, it updates the estimate $\widetilde{\beta}^{(t)}$ by fitting a shifted adaptive Huber regression which leverages global first-order information, depending on τ, and local higher-order information, depending on κ. The procedure involves two steps.

1. COMMUNICATING GRADIENT INFORMATION. The central machine broadcasts $\widetilde{\beta}^{(t-1)}$ to every local machine. The jth machine, $1 \leq j \leq m$, computes the gradient $\nabla\mathcal{L}_{j,\tau}(\widetilde{\beta}^{(t-1)})$, and sends it back to the central machine. This step requires a communication of $2(m-1)p$ bits.

Algorithm 1 Communication-Efficient Adaptive Huber Regression.

```
Input: data batches $\{(y_i,x_i)\}_{i\in I_j}$, $j=1,\ldots,m$, stored on $m$ machines, robustification parameters $\tau \geq \kappa > 0$, initialization $\widetilde{\beta}^{(0)}$, number of iterations $T, g_0 = 1$.
for $t = 1,2\ldots,T$ do
    Broadcast $\widetilde{\beta}^{(t-1)}$ to all local machines;
    The $j$th ($1 \leq j \leq m$) machine computes $\nabla\widehat{\mathcal{L}}_{j,\tau}(\widetilde{\beta}^{(t-1)})$, and transmit it to the central machine;
    Compute $\nabla\widehat{\mathcal{L}}_{\tau}(\widetilde{\beta}^{(t-1)}) = (1/m)\sum_{j=1}^{m}\nabla\widehat{\mathcal{L}}_{j,\tau}(\widetilde{\beta}^{(t-1)})$, $\nabla\widehat{\mathcal{L}}_{1,\kappa}(\widetilde{\beta}^{(t-1)})$ and $g_t = \|\nabla\widehat{\mathcal{L}}_{\tau}(\widetilde{\beta}^{(t-1)})\|_\infty$ on the central machine;
    If $g_t \geq g_{t-1}$ or $g_t \leq 10^{-5}$ break ; otherwise, on the central machine, solve the shifted adaptive Huber regression problem in (2.7) to update the estimate $\widetilde{\beta}^{(t)}$;
end for
Output: $\widetilde{\beta}^{(T)}$.
```

2. Fitting local shifted AHR. The central machine computes the update $\widetilde{\beta}^{(t)}$, defined as a solution to the optimization problem

$$\min_{\beta\in\mathbb{R}^p} \widetilde{\mathcal{L}}^{(t)}(\beta) := \widehat{\mathcal{L}}_{1,\kappa}(\beta) - \left\langle \nabla\widehat{\mathcal{L}}_{1,\kappa}(\widetilde{\beta}^{(t-1)}) - \frac{1}{m}\sum_{j=1}^{m}\nabla\widehat{\mathcal{L}}_{j,\tau}(\widetilde{\beta}^{(t-1)}), \beta \right\rangle, \tag{2.7}$$

which can be solved by the method of iteratively reweighted least squares or quasi-Newton methods. Details are given in section 2.5.1. We summarize the procedure, with an early stopping criterion, in Algorithm 1.

Theorem 2. *Assume the same conditions in Theorem 1, and let* $8r_* \leq r_0 \leq \sigma$. *Conditioned on event* $\mathcal{E}_0(r_0)\cap\mathcal{E}_*(r_*)$, *the distributed AHR estimator* $\widetilde{\beta}^{(T)}$ *with* $T \gtrsim \lceil \log(r_0/r_*)/\log(n/(p+u))\rceil$ *satisfies the bounds*

$$\|\widetilde{\beta}^{(T)} - \beta^*\|_\Sigma \lesssim r_* \quad \textit{and} \quad \|\widetilde{\beta}^{(T)} - \beta^* + \Sigma^{-1}\nabla\widehat{\mathcal{L}}_\tau(\beta^*)\|_\Sigma \lesssim \sqrt{\frac{p+u}{n}}\cdot r_*$$

with probability at least $1-(2T+1)e^{-u}$.

The above result shows that, with proper choices of τ and κ as well as the number of iterations, the statistical error of the multi-step distributed AHR estimator matches that of the centralized AHR estimator on the entire dataset.For the initialization, we may take $\widetilde{\beta}^{(0)}$

to be a local AHR estimator computed on the central machine. With the above preparations, we are ready to explicitly describe the estimation error and Bahadur linearization error of the proposed distributed AHR estimator. The result is nonasymptotic, and carefully tracks the impact of the parametric dimension p, local sample size n and the number of machines m.

Theorem 3. *Assume Condition (C1) holds, and suppose the local sample size satisfies $n \gtrsim p + \log n + \log_2 m$, where $\log_2 m := \log(\log m)$ and $m = N/n$. Choose the robustification parameters $\tau \geq \kappa > 0$ as $\tau \asymp \sigma\sqrt{N/(p + \log n + \log_2 m)}$ and $\kappa \asymp \sigma\sqrt{n/(p + \log n + \log_2 m)}$. Then, starting at iteration 0 with a local AHR estimate $\widetilde{\beta}^{(0)}$, the distributed estimator $\widetilde{\beta} = \widetilde{\beta}^{(T)}$ with $T \asymp \lceil \frac{\log(m)}{\log(n/(p+\log n+\log_2 m))} \rceil$ satisfies*

$$\|\widetilde{\beta} - \beta^*\|_\Sigma \lesssim \sigma\sqrt{\frac{p + \log n + \log_2 m}{N}} \quad \textit{and} \tag{2.8}$$

$$\left\| \widetilde{\beta} - \beta^* - \Sigma^{-1}\frac{1}{N}\sum_{i=1}^{N} \psi_\tau(\varepsilon_i)x_i \right\|_\Sigma \lesssim \sigma \frac{p + \log n + \log_2 m}{(nN)^{1/2}} \tag{2.9}$$

with probability at least $1 - Cn^{-1}$, where $\psi_\tau(u) := \ell'_\tau(u) = \operatorname{sign}(u)\min(|u|, \tau)$.

The above theorem indicates that the multi-step distributed AHR estimator $\widetilde{\beta}$ achieves the optimal statistical rate of convergence by a delicate combination of the local robustification parameter, the global robustification parameter, and number of communication rounds. The second bound, (2.9), explicitly describes the error term of the Bahadur linearization. This allows to establish the asymptotic distribution of $\widetilde{\beta}$ when both p, n tend to infinity. Moreover, to achieve statistical optimality and communication efficiently simultaneously, the above results impose minimal conditions on the number of machines m. In summary, when data are heavy-tailed and collected on each machine remain local, the proposed procedure delivers a statistically optimal estimate by communicating as many as $O(pm\log(m))$ bits.

2.3.2 Distributed confidence estimation

In this section, we consider uncertainty quantification of the multi-step estimator in a distributed setting, with a particular focus on statistical confidence estimation. We first establish a Berry-Esseen bound for linear functionals of the distributed AHR estimator $\widetilde{\beta}$, which explicitly quantifies the normal approximation error.

Theorem 4. *In addition to the conditions in Theorem 3, assume* $\mathbb{E}(\varepsilon^2|x) = \sigma^2$ *and* $\mathbb{E}(|\varepsilon|^{2+\delta}|x) \le v_{2+\delta}$ *almost surely for some* $0 < \delta \le 1$. *Then, the distributed estimator* $\widetilde{\beta} = \widetilde{\beta}^{(T)}$ *satisfies*

$$\sup_{t \in \mathbb{R}, a \in \mathbb{R}^p} \left| \mathbb{P}\left[\frac{N^{1/2} a^{\mathrm{T}}(\widetilde{\beta} - \beta^*)}{\sqrt{\mathbb{E}\{\psi_\tau(\varepsilon) a^{\mathrm{T}} \Sigma^{-1} x\}^2}} \le t \right] - \Phi(t) \right|$$
$$\lesssim \frac{p + \log n + \log_2 m}{n^{1/2}} + \frac{v_{2+\delta}(p + \log n + \log_2 m)^{(1+\delta)/2}}{\sigma^{2+\delta} N^{\delta/2}},$$

where $\Phi(\cdot)$ *is the standard normal distribution function. In particular, assume* $\mathbb{E}(|\varepsilon|^3|x) \le v_3 < \infty$ *almost surely. Then, under the dimension constraint* $p + \log_2 m = o(n^{1/2})$,

$$\frac{N^{1/2} a^{\mathrm{T}}(\widetilde{\beta} - \beta^*)}{\sqrt{\mathbb{E}\{\psi_\tau(\varepsilon) a^{\mathrm{T}} \Sigma^{-1} x\}^2}} \xrightarrow{\mathrm{d}} \mathcal{N}(0,1) \quad \textit{and} \quad \frac{N^{1/2} a^{\mathrm{T}}(\widetilde{\beta} - \beta^*)}{\sigma (a^{\mathrm{T}} \Sigma^{-1} a)^{1/2}} \xrightarrow{\mathrm{d}} \mathcal{N}(0,1)$$

uniformly over $a \in \mathbb{R}^p$ *as* $n \to \infty$.

Let $\widetilde{\beta} = (\widetilde{\beta}_1, \ldots, \widetilde{\beta}_p)^{\mathrm{T}}$ be the distributed estimator described in the previous subsection. Theorem 4 implies that, for each $1 \le j \le p$, $N^{1/2}(\widetilde{\beta}_j - \beta_j^*)$ is asymptotically normal with zero mean and variance $(\Sigma^{-1}\mathbb{E}\{\psi_\tau(\varepsilon) x x^{\mathrm{T}}\}^2 \Sigma^{-1})_{jj}$. Let $\widehat{\Sigma} = (1/N)\sum_{i=1}^N x_i x_i^{\mathrm{T}}$ be the sample version of Σ, and $\widehat{\varepsilon}_i = y_i - x_i^{\mathrm{T}} \widetilde{\beta}$ be the fitted residuals. It can be shown that $(\widehat{\Sigma}^{-1} N^{-1} \sum_{i=1}^N \psi_\tau^2(\varepsilon_i) x_i x_i^{\mathrm{T}} \widehat{\Sigma}^{-1})_{jj}$ provides a consistent estimator of $(\Sigma^{-1}\mathbb{E}\{\psi_\tau(\varepsilon) x x^{\mathrm{T}}\}^2 \Sigma^{-1})_{jj}$. In a distributed setting, the computation of this variance estimator requires communicating $O(p^2 m)$ bits, thus incurring exorbitant communication costs when p is large.

To achieve a tradeoff between communication and statistical efficiencies, we propose averaging pointwise variance estimators, defined by $\widehat{\sigma}_j^2 := (1/m)\sum_{k=1}^m \widehat{\sigma}_{jm}^2$ for $j = 1, \ldots, p$,

where

$$\widehat{\sigma}_{jk}^2 = (\widehat{\Sigma}_k^{-1}\widehat{\Lambda}_k\widehat{\Sigma}_k^{-1})_{jj}, \quad \widehat{\Lambda}_k = \frac{1}{n}\sum_{i\in I_k}\psi_\tau^2(\widehat{\varepsilon}_i)x_ix_i^{\mathrm{T}} \text{ and } \widehat{\Sigma}_k = \frac{1}{n}\sum_{i\in I_k}x_ix_i^{\mathrm{T}}.$$

This approach takes one additional round of communication, and is robust against heteroscedastic errors that are of the form $\varepsilon_i = \sigma(x_i)e_i$. When ε_i is independent of x_i, the asymptotic variances reduce to $\mathbb{E}\{\psi_\tau^2(\varepsilon)\}(\Sigma^{-1})_{jj}$, and thus can be consistently estimated by $\widetilde{\sigma}_j^2 := (\widehat{\sigma}_\varepsilon^2/m)\sum_{k=1}^m(\widehat{\Sigma}_k^{-1})_{jj}$, where $\widehat{\sigma}_\varepsilon^2 = (N-p)^{-1}\sum_{i=1}^N\psi_\tau^2(\widehat{\varepsilon}_i)$. For $\alpha\in(0,1)$, the distributed $100(1-\alpha)\%$ normal-based confidence intervals for β_j^*, $j=1,\ldots,p$, are given by $[\widetilde{\beta}_j - z_{\alpha/2}\widehat{\sigma}_jN^{-1/2}, \widetilde{\beta}_j + z_{\alpha/2}\widehat{\sigma}_jN^{-1/2}]$ or $[\widetilde{\beta}_j - z_{\alpha/2}\widetilde{\sigma}_jN^{-1/2}, \widetilde{\beta}_j + z_{\alpha/2}\widetilde{\sigma}_jN^{-1/2}]$, where $z_{\alpha/2} = \Phi^{-1}(1-\alpha/2)$.

2.4 Distributed Regularized Adaptive Huber Regression

In this section, we consider high-dimensional linear models under sparsity. Specifically, we allow the parametric dimension p to be much larger than the local sample size n, and assume β^* is s-sparse, where $s = |\mathcal{S}|$ and $\mathcal{S} = \mathrm{supp}(\beta^*) = \{1 \le j \le p : \beta_j^* \neq 0\}$ denotes the true active set.

Given independent observations $\{(y_i,x_i)\}_{i=1}^N$ from the linear model (2.1), the centralized/global ℓ_1-penalized Huber regression estimator (ℓ_1-Huber) is defined as

$$\widehat{\beta} = \widehat{\beta}_\tau(\lambda) \in \mathop{\arg\min}_{\beta\in\mathbb{R}^p}\big\{\widehat{\mathcal{L}}_\tau(\beta) + \lambda\|\beta\|_1\big\}, \tag{2.10}$$

where $\lambda > 0$ is a regularization parameter. Statistical properties of ℓ_1-penalized Huber regression have been thoroughly studied by Lambert-Lacroix and Zwald (2011), Fan et al. (2017), Loh (2017) and Chinot et al. (2020) from different perspectives. To deal with asymmetric heavy-tailed errors, Fan et al. (2017) established high probability bounds for the ℓ_1-Huber estimator with $\tau \asymp \sigma\sqrt{N/\log(p)}$ in the high-dimensional regime $p \gg n \gtrsim s\log(p)$.

Remark 1. *In practice, it is natural to leave the intercept or a given subset of the parameters* *unpenalized in the penalized M-estimation framework* (2.10). *Denote by* $\mathcal{R} \subseteq \{1,\dots,p\}$ *the index set of unpenalized parameters, which is typically user-specified and contains at least index 1. A more flexible* ℓ_1*-Huber estimator can be obtained by solving* $\min_{\beta\in\mathbb{R}^p}\{\widehat{\mathcal{L}}_\tau(\beta)+\lambda\|\beta_{\mathcal{R}^c}\|_1\} = \min_{\beta\in\mathbb{R}^p}\{\widehat{\mathcal{L}}_\tau(\beta)+\lambda\sum_{j\in\mathcal{R}^c}|\beta_j|\}$. *Similar theoretical analysis can be carried out with slight modifications, and thus will be omitted.*

In a distributed setting, we integrate the ideas of Wang et al. (2017) and Jordan et al. (2019) with adaptive robustification to parallelize regularized Huber regression with controlled communication cost and optimal statistical guarantees. As before, let τ and κ be the global and local robustification parameters. Recall that $\widehat{\mathcal{L}}_{j,\kappa}(\cdot)$, $j=1,\dots,m$, denote local Huber loss functions. Commenced with a regularized estimator $\widetilde{\beta}^{(0)}$, let $\widetilde{\mathcal{L}}(\beta) = \widehat{\mathcal{L}}_{1,\kappa}(\beta) - \langle\nabla\widehat{\mathcal{L}}_{1,\kappa}(\widetilde{\beta}^{(0)}) - \nabla\widehat{\mathcal{L}}_\tau(\widetilde{\beta}^{(0)}),\beta\rangle$ be the shifted adaptive Huber loss as in (2.2). With slight abuse of notation, we define the one-step ℓ_1-penalized Huber regression estimator as

$$\widetilde{\beta}^{(1)} = \widetilde{\beta}^{(1)}_{\kappa,\tau}(\lambda) \in \operatorname*{arg\,min}_{\beta\in\mathbb{R}^p}\bigl\{\widetilde{\mathcal{L}}(\beta)+\lambda\|\beta\|_1\bigr\}. \tag{2.11}$$

To assess the statistical properties of the one-step estimator $\widetilde{\beta}^{(1)}$, we work under the the following moment condition on the random covariate vector in high dimensions.

(C2). The covariate vector $x=(x_1,\dots,x_p)^{\mathrm{T}}\in\mathbb{R}^p$ with $x_1\equiv 1$ has bounded components and uniformly bounded kurtosis. That is, $\max_{1\le j\le p}|x_j|\le B$ for some $B\ge 1$ and $\mu_4 = \sup_{u\in\mathbb{S}^{p-1}}\mathbb{E}(z^{\mathrm{T}}u)^4<\infty$, where $z=\Sigma^{-1/2}x$ and $\Sigma=(\sigma_{jk})_{1\le j,k\le p}=\mathbb{E}(xx^{\mathrm{T}})$. Write $\sigma_u = \max_{1\le j\le p}\sigma_{jj}^{1/2}$ and $\lambda_l=\lambda_{\min}(\Sigma)>0$. For simplicity, we assume $\lambda_l=1$. Moreover, the error variables ε_i satisfy $\mathbb{E}(\varepsilon_i|x_i)=0$ and $\mathbb{E}(\varepsilon_i^2|x_i)\le\sigma^2$ almost surely.

As before, we first examine the performance of $\widetilde{\beta}^{(1)}$ conditioned on certain "good" events in regard of the initialization and the centralized ℓ_1-Huber estimator. For $r_0,\lambda_*>0$,

define

$$\mathcal{E}_0(r_0) = \{\widetilde{\beta}^{(0)} \in \Theta(r_0) \cap \Lambda\} \quad \text{and} \quad \mathcal{E}_*(\lambda_*) = \{\|\nabla \widehat{\mathcal{L}}_\tau(\beta^*) - \nabla \mathcal{L}_\tau(\beta^*)\|_\infty \le \lambda_*\},$$

where $\Lambda := \{\beta \in \mathbb{R}^p : \|\beta - \beta^*\|_1 \le 4s^{1/2}\|\beta - \beta^*\|_\Sigma\}$ is an ℓ_1-cone.

Theorem 5. *Assume Condition (C2) holds. Given $\delta \in (0,1)$ and $0 < r_0, \lambda_* \lesssim \sigma$, let (τ, κ, λ) satisfy $\tau \ge \kappa \asymp \sigma\sqrt{n/\log(p/\delta)}$ and $\lambda = 2.5(\lambda_* + \rho)$ with*

$$\rho \asymp \max\left\{ r_0\sqrt{\frac{s\log(p/\delta)}{n}}, s^{-1/2}\sigma^2\tau^{-1} \right\}.$$

Moreover, suppose the local sample size satisfies $n \gtrsim s\log(p/\delta)$. Then, conditioned on the event $\mathcal{E}_0(r_0) \cap \mathcal{E}_*(\lambda_*)$, the one-step regularized estimator $\widetilde{\beta}^{(1)}$ defined in (2.11) satisfies $\widetilde{\beta}^{(1)} \in \Lambda$ and

$$\|\widetilde{\beta}^{(1)} - \beta^*\|_\Sigma \lesssim s\sqrt{\frac{\log(p/\delta)}{n}} \cdot r_0 + \sigma^2\tau^{-1} + s^{1/2}\lambda_* \tag{2.12}$$

with probability at least $1 - \delta$.

Theorem 5 indicates that the one-step procedure is able to reduce the statistical error of the initial estimator by a factor of $s\sqrt{\log(p)/n}$ when the local sample size satisfies $n \gtrsim s^2\log(p)$; see the first term on the right-hand of (2.12). The second term, $\sigma^2\tau^{-1} + s^{1/2}\lambda_*$, corresponds to the global error rate achievable on the entire dataset. In view of Theorem B.2 (with $\delta = 1$) in Sun et al. (2020), if we take $\lambda_* \asymp \sigma\sqrt{\log(p)/N}$ and $\tau \asymp \sigma\sqrt{N/\log(p)}$, the centralized ℓ_1-Huber estimator given in (2.10) satisfies $\|\widehat{\beta} - \beta^*\|_\Sigma \lesssim \sigma^2\tau^{-1} + s^{1/2}\lambda_* \asymp \sigma\sqrt{s\log(p)/N}$ with probability at least $1 - Cp^{-1}$.

Now we extend the iterative procedure in Section 2.3 to high-dimensional settings, starting at iteration 0 with an initial estimate $\widetilde{\beta}^{(0)} \in \mathbb{R}^p$. At iteration $t = 1, 2, \ldots$, it proceeds as follows:

Communicating gradient information. The jth ($2 \le j \le m$) machine receives $\widetilde{\beta}^{(t-1)}$ from the central machine, computes the local gradient $\nabla\widehat{\mathcal{L}}_{j,\tau}(\widetilde{\beta}^{(t-1)})$, and sends it back to the central.

Fitting local regularized AHR: On the central machine, solve $\min_{\beta\in\mathbb{R}^p}\{\widetilde{\mathcal{L}}^{(t)}(\beta)+\lambda_t\|\beta\|_1\}$ to obtain $\widetilde{\beta}^{(t)}$, where $\widetilde{\mathcal{L}}^{(t)}(\beta)=\widehat{\mathcal{L}}_{1,\kappa}(\beta)-\langle\nabla\widehat{\mathcal{L}}_{1,\kappa}(\widetilde{\beta}^{(t-1)})-(1/m)\sum_{j=1}^m\nabla\widehat{\mathcal{L}}_{j,\tau}(\widetilde{\beta}^{(t-1)}),\beta\rangle$ and $\lambda_t>0$ is a regularization parameter.

Computationally, we use a variant of the majorize-minimize algorithm (Lange et al., 2000), a proximal gradient descent type method, to solve the regularized optimization problem at each iteration. Details are provided in section 2.5.2. Theorem 6 below describes the statistical properties of the solution path $\{\widetilde{\beta}^{(t)}\}_{t\geq 1}$ conditioned on a prespecified level of accuracy of the initial estimator.

Theorem 6. *Assume Condition (C2) holds. Given $\delta\in(0,1)$ and $0<r_0,\lambda_*\lesssim\sigma$, let (τ,κ) satisfy $\tau\geq\kappa\asymp\sigma\sqrt{n/\log(p/\delta)}$. For $t=1,2,\ldots$, set $\lambda_t=2.5(\lambda_*+\rho_t)>0$ with $\rho_t\asymp s^{-1/2}\max\{\alpha^t r_0,\sigma^2\tau^{-1}\}$ and $\alpha\asymp s\sqrt{\log(p/\delta)/n}$. Suppose the local sample size satisfies $n\gtrsim s^2\log(p/\delta)$, and let $r_*\asymp\sigma^2\tau^{-1}+s^{1/2}\lambda_*$. Then, conditioned on event $\mathcal{E}_0(r_0)\cap\mathcal{E}_*(\lambda_*)$, the distributed regularized estimator $\widetilde{\beta}^{(T)}$ with $T\asymp\frac{\log(r_0/r_*)}{\log(1/\alpha)}$ satisfies $\widetilde{\beta}^{(T)}\in\Lambda$ and $\|\widetilde{\beta}^{(T)}-\beta^*\|_\Sigma\lesssim r_*$ with probability at least $1-T\delta$.*

With sufficiently many samples on the central machine—$n\gtrsim s^2\log(p)$, Theorems 5 and 6 ensure that the initial estimation error, albeit being sub-optimal, can be repeatedly refined by a factor of order $s\sqrt{\log(p)/n}$ until it reaches the optimal rate. For simplicity, we take $\widetilde{\beta}^{(0)}$ to be a local ℓ_1-penalized AHR estimator, that is, $\widetilde{\beta}^{(0)}\in\arg\min_{\beta\in\mathbb{R}^p}\{\mathcal{L}_{1,\kappa}(\beta)+\lambda_0\|\beta\|_1\}$.

Corollary 1. *Assume Condition (C2) holds, and the sample size per machine satisfies $n\gtrsim s^2\log p$. Choose the robustification and regularization parameters as $\tau\asymp\sigma\sqrt{N/\log(p)}$, $\kappa\asymp\sigma\sqrt{n/\log(p)}$ and*

$$\lambda_t\asymp\sigma\sqrt{\frac{\log p}{N}}+\sigma\left(\frac{s^2\log p}{n}\right)^{t/2}\sqrt{\frac{\log p}{n}},\ \ t=0,1,2,\ldots.$$

Starting at iteration 0 with a local ℓ_1-penalized AHR estimator, the multi-step estimator $\widetilde{\beta}^{(T)}$ after $T \asymp \lceil \log(m) \rceil$ rounds of communication satisfies the bounds

$$\|\widetilde{\beta}^{(T)} - \beta^*\|_\Sigma \lesssim \sigma\sqrt{\frac{s\log p}{N}} \quad \textit{and} \quad \|\widetilde{\beta}^{(T)} - \beta^*\|_1 \lesssim \sigma s\sqrt{\frac{\log p}{N}}$$

with probability at least $1 - C\log(m)/p$.

Corollary 1, along with the global error analysis in Fan et al. (2017) and Loh (2017), implies the optimality of distributed adaptive Huber regression in terms of the tradeoff between communication cost and statistical accuracy.

Remark 2. *Under light-tailed error distributions (e.g., sub-Gaussian errors), Lee et al. (2017) and Battey et al. (2018) studied a one-shot approach based on averaging debiased Lasso estimators (Zhang and Zhang, 2014; van de Geer et al., 2014). Theoretically, averaged debiased Lasso achieves the optimal error rate when the local size satisfies $n \gtrsim ms^2\log(p)$; and computationally, each local machine needs to estimate a $p \times p$ matrix for debiasing the Lasso. We may expect the same issues for the robust one-shot method that averages debiased ℓ_1-Huber estimators. The proposed distributed AHR method not only requires the minimum sample complexity but also is computationally efficient.*

2.5 Optimization Methods

2.5.1 Barzilai-Borwein gradient descent for distributed AHR

Let us first recall the multi-round distributed procedure for adaptive Huber regression. Starting with an initial estimator $\widetilde{\beta}^{(0)} \in \mathbb{R}^p$, and given robustification parameters τ and κ, for $t = 1, \ldots, T$, we update

$$\widetilde{\beta}^{(t)} \in \operatorname*{arg\,min}_{\beta \in \mathbb{R}^p} \widetilde{\mathcal{L}}^{(t)}(\beta) = \widehat{\mathcal{L}}_{1,\kappa}(\beta) - \big\langle \nabla\widehat{\mathcal{L}}_{1,\kappa}(\widetilde{\beta}^{(t-1)}) - \nabla\widehat{\mathcal{L}}_\tau(\widetilde{\beta}^{(t-1)}), \beta \big\rangle. \tag{2.13}$$

Since $\widetilde{\mathcal{L}}^{(t)}(\cdot)$ is convex, twice-differentiable and provably locally strongly convex, we propose to use the gradient descent method with a Barzilai-Borwein update step (GD-BB) (Barzilai and Borwein, 1988) to solve the optimization problem in (2.13). The Barzilai-Borwein method is motivated by quasi-Newton methods, which avoid calculating the inverse Hessian at each iteration. The latter is computationally expensive when p is large. To be specific, let us consider the optimization $\min_{\beta \in \mathbb{R}^p} \widetilde{\mathcal{L}}^{(t)}(\beta)$ for a fixed $t \geq 1$. Starting with the initialization $\widetilde{\beta}^{(t,0)} = \widetilde{\beta}^{(t-1)}$, at (inner) iteration $k = 1, 2, ...$, compute the update $\widetilde{\beta}^{(t,k+1)} = \widetilde{\beta}^{(t,k)} - \min\{\eta_k, 10\} \nabla \widetilde{\mathcal{L}}^{(t)}(\widetilde{\beta}^{(t,k)})$, where $\eta_1 = 1$ and for $k \geq 2$,

$$\eta_k = \frac{\langle \widetilde{\beta}^{(t,k)} - \widetilde{\beta}^{(t,k-1)}, \widetilde{\beta}^{(t,k)} - \widetilde{\beta}^{(t,k-1)} \rangle}{\langle \widetilde{\beta}^{(t,k)} - \widetilde{\beta}^{(t,k-1)}, \nabla \widetilde{\mathcal{L}}^{(t)}(\widetilde{\beta}^{(t,k)}) - \nabla \widetilde{\mathcal{L}}^{(t)}(\widetilde{\beta}^{(t,k-1)}) \rangle} \tag{2.14}$$

or

$$\eta_k = \frac{\langle \widetilde{\beta}^{(t,k)} - \widetilde{\beta}^{(t,k-1)}, \nabla \widetilde{\mathcal{L}}^{(t)}(\widetilde{\beta}^{(t,k)}) - \nabla \widetilde{\mathcal{L}}^{(t)}(\widetilde{\beta}^{(t,k-1)}) \rangle}{\| \nabla \widetilde{\mathcal{L}}^{(t)}(\widetilde{\beta}^{(t,k)}) - \nabla \widetilde{\mathcal{L}}^{(t)}(\widetilde{\beta}^{(t,k-1)}) \|_2^2}.$$

In practice, the step size computed in GD-BB may sometimes vibrate to some extent, and this may cause instability of the algorithm. Therefore, we set a upper bound for the step sizes by taking $\min\{\eta_k, 10\}$. This procedure is summarized in Algorithm 2.

2.5.2 Majorize-minimize algorithm for distributed penalized AHR

In the high-dimensional setting, we need to solve ℓ_1-penalized shifted Huber loss minimization problems. With slight abuse of notation, given an initial regularized estimator $\widetilde{\beta}^{(0)}$, at each iteration $t = 1, 2, \ldots, T$, define the update as

$$\begin{aligned} \widetilde{\beta}^{(t)} \in \operatorname*{arg\,min}_{\beta \in \mathbb{R}^p} \Big\{ & \widetilde{\mathcal{L}}^{(t)}(\beta) + \lambda \|\beta_-\|_1 = \widehat{\mathcal{L}}_{1,\kappa}(\beta) \\ & - \big\langle \nabla \widehat{\mathcal{L}}_{1,\kappa}(\widetilde{\beta}^{(t-1)}) - \nabla \widehat{\mathcal{L}}_{\tau}(\widetilde{\beta}^{(t-1)}), \beta \big\rangle + \lambda \|\beta_-\|_1 \Big\}. \end{aligned} \tag{2.15}$$

Algorithm 2 Gradient Descent with Barzilai-Borwein stepsize for solving (2.13)

```
Input: Local data vectors $\{(y_i, x_i)\}_{i \in I_1}$, initial estimator $\widehat{\beta}^0 = \widetilde{\beta}^{(t-1)}$, gradient $\nabla \widehat{\mathcal{L}}_{1,\kappa}(\widetilde{\beta}^{(t-1)})$ and $\nabla \widehat{\mathcal{L}}_{j,\tau}(\widetilde{\beta}^{(t-1)})$ for $j = 1, \ldots, m$, and gradient tolerance level $\delta = 10^{-4}$.
1: Compute $\widehat{\beta}^1 \leftarrow \widehat{\beta}^0 - \nabla \widetilde{\mathcal{L}}^{(t)}(\widehat{\beta}^0)$
2: for $k = 1, 2 \ldots$ do
3:     Compute $\eta_k$ as defined in (2.14).
4:     Update $\widehat{\beta}^{k+1} \leftarrow \widehat{\beta}^k - \min\{\eta_k, 10\} \nabla \widetilde{\mathcal{L}}^{(t)}(\widehat{\beta}^k)$;
5: end for when $\|\nabla \widetilde{\mathcal{L}}^{(t)}(\widehat{\beta}^k)\|_\infty \leq \delta$
```

Here we use $\beta_- \in \mathbb{R}^{p-1}$ to denote the subvector of β with its first component removed. To solve the optimization problem in (2.15), we employ the locally adaptive majorize-minimize (LAMM) principle Fan et al. (2018), which extends the classical MM algorithm (Hunter and Lange, 2000) to accommodate ℓ_1 penalty. This procedure minimizes a surrogate ℓ_1-penalized isotropic quadratic function at each iteration, thus permitting an analytical solution.

Let $\widetilde{\mathcal{L}}(\cdot)$ be the loss function of interest. For $k = 1, 2, \ldots$, define

$$g_k(\beta; \beta^{k-1}, \phi_k) = \widetilde{\mathcal{L}}(\beta^{k-1}) + \left\langle \nabla \widetilde{\mathcal{L}}(\beta^{k-1}), \beta - \beta^{k-1} \right\rangle + \frac{\phi_k}{2} \|\beta - \beta^{k-1}\|_2^2.$$

We say $g_k(\beta; \beta^{k-1}, \phi_k)$ majorizes $\widetilde{\mathcal{L}}(\beta)$ at β^{k-1} if

$$g_k(\beta; \beta^{k-1}, \phi_k) \geq \widetilde{\mathcal{L}}(\beta) \ \ \forall \beta \in \mathbb{R}^p \ \text{ and } \ g_k(\beta^{k-1}; \beta^{k-1}, \phi_k) = \widetilde{\mathcal{L}}(\beta^{k-1}). \tag{2.16}$$

By choosing ϕ_k large enough, $g_k(\cdot; \beta^{k-1}, \phi_k)$ is guaranteed to satisfy (2.16). To find the smallest such ϕ_k, we start with $\phi_0 = 0.0001$, and repeatedly inflate it by a constant factor, say 1.1, until (2.16) is satisfied. Finally, we update β^k by minimizing

$$g_k(\beta; \beta^{k-1}, \phi_k) + \lambda \|\beta_-\|_1. \tag{2.17}$$

Due to the isotropic quadratic term in $g_k(\beta; \beta^{k-1}, \phi_k)$, β^k can be obtained by a simple analytic

Algorithm 3 Local adaptive majorize-minimise (LAMM) algorithm for solving (2.11)

Input: Local data vectors $\{(y_i,x_i)\}_{i\in I_1}$, initial estimator $\widehat{\beta}^0 = \widetilde{\beta}^{(t-1)}$ gradient vectors $\nabla\widehat{\mathcal{L}}_{1,\kappa}(\widetilde{\beta}^{(t-1)})$ and $\nabla\widehat{\mathcal{L}}_{\tau}(\widetilde{\beta}^{(t-1)})$, regularisation parameter λ, initial isotropic parameter ϕ_0 and convergence tolerance δ

```
for $k = 1,2\dots$ do
    Set $\phi_k \leftarrow \max\{\phi_0, \phi_{k-1}/1.1\}$
    repeat
        Update $\widehat{\beta}_1^k \leftarrow \widehat{\beta}_1^{k-1} - \phi_k^{-1}\nabla_{\beta_1}\widetilde{\mathcal{L}}(\widehat{\beta}^{k-1})$
        Update $\widehat{\beta}_j^k \leftarrow S(\widehat{\beta}_j^{k-1} - \phi_k^{-1}\nabla_{\beta_j}\widetilde{\mathcal{L}}(\widehat{\beta}^{k-1}), \phi_k^{-1}\lambda)$ for $j = 2,\dots,p$
        If $g_k(\widehat{\beta}^k;\widehat{\beta}^{k-1},\phi_k) < \widetilde{\mathcal{L}}(\widehat{\beta}^k)$, set $\phi_k \leftarrow 1.1\phi_k$
    until $g_k(\widehat{\beta}^k;\widehat{\beta}^{k-1},\phi_k) \geq \widetilde{\mathcal{L}}(\widehat{\beta}^k)$
end for when $\|\widehat{\beta}^k - \widehat{\beta}^{k-1}\|_2 \leq \delta$
```

formula:

$$\begin{cases} \beta_1^k = \beta_1^{k-1} - \phi_k^{-1}(\nabla\widetilde{\mathcal{L}}(\beta^{k-1}))_1 \\ \beta_j^k = S(\beta_j^{k-1} - \phi_k^{-1}(\nabla\widetilde{\mathcal{L}}(\beta^{k-1}))_j, \phi_k^{-1}\lambda), \quad j = 2,\dots,p, \end{cases}$$

where $S(u,\lambda) = \operatorname{sign}(u)\max(|u|-\lambda, 0)$ denotes the soft-thresholding operator. This algorithm also guarantees a descent in the overall loss function at every iteration, which is a direct consequence of (2.16) and (2.17):

$$\begin{aligned} \widetilde{\mathcal{L}}(\beta^k) + \lambda\|\beta_-^k\|_1 &\leq g_k(\beta^k;\beta^{k-1},\phi_k) + \lambda\|\beta_-^k\|_1 \\ &\leq g_k(\beta^{k-1};\beta^{k-1},\phi_k) + \lambda\|\beta_-^{k-1}\|_1 = \widetilde{\mathcal{L}}(\beta^{k-1}) + \lambda\|\beta_-^{k-1}\|_1. \end{aligned}$$

Algorithm 3 summarizes the LAMM algorithm described above.

2.6 Numerical Studies

In this section, we compare the numerical performance of the proposed method with several state-of-the-art distributed regression methods in both low and high dimensions.

2.6.1 Distributed robust regression and inference

In the low-dimensional setting where $n \gg p$, we consider five distributed regression methods: (i) the global adaptive Huber regression (AHR) estimator (Sun et al., 2020) that uses all the available $N = mn$ observations; (ii) divide-and-conquer AHR (DC-AHR) estimator based on averaging m local AHR estimators; (iii) DC-OLS estimator that averages m local OLS estimators; (iv) distributed OLS estimator (Shamir et al., 2014); and (v) the proposed distributed AHR estimator with early stopping.

To implement methods (i) and (ii), we use the self-tuning principle proposed by Wang et al. (2021) which automatically selects the robustification parameter τ. The distributed procedures (iv) and (v) are iterative, and require a reasonably well initial estimator, say $\widetilde{\beta}^{(0)}$. In our simulations, we take $\widetilde{\beta}^{(0)}$ to be either the DC-AHR or the DC-OLS estimator, which only requires one communication round. When the error distribution is heavy-tailed and symmetric, DC-AHR often has better finite-sample performance than DC-OLS. However, it produces biased estimate when the error is asymmetric. In contrast, although the DC-OLS exhibits larger variability due to heavy-tailedness, it has smaller bias on average. Therefore, we use DC-OLS estimator as the initialization for both methods (iv) and (v). Recall that the distributed AHR estimator involves two robustification parameters κ and τ. The local parameter κ can be automatically obtained by the self-tuning procedure (Wang et al., 2021). Guided by theoretical orders of (κ, τ) stated in Theorem 3, we choose the global parameter τ to be $cm^{1/2}\kappa$, where $c \geq 1$ is a numerical constant that can be tuned by the validation set approach. We suggest to choose c from $\{1, 2, 3, 4, 5\}$, which suffices to achieve promising performance in a wide range of simulation settings.

We generate data vectors $\{(y_i, x_i)\}_{i=1}^N$ from a heteroscedastic model $y_i = x_i^{\mathrm{T}}\beta^* + c^{-1}(x_i^{\mathrm{T}}\beta^*)^2\varepsilon_i$, where $\beta^* = (1.5, \ldots, 1.5)^{\mathrm{T}} \in \mathbb{R}^p, x_i = (1, x_{i2}, \ldots, x_{ip})^{\mathrm{T}}$ with $x_{ij} \sim \mathcal{N}(0, 1)$ for $j = 2, \ldots, p$ and $c = \sqrt{3}\|\beta^*\|_2^2$ that makes $\mathbb{E}\{c^{-1}(x_i^{\mathrm{T}}\beta^*)^2\}^2 = 1$. The regression errors ε_i are generated from one of the following four distributions (centered if the mean is nonzero): (a) $\mathcal{N}(0, 1)$ (standard normal), (b) t_2 (t-distribution with 2 degrees of freedom), (c) Par$(4, 2)$–

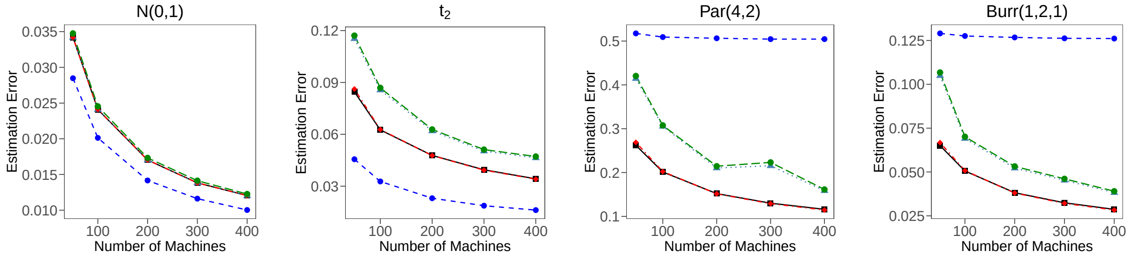
Pareto distribution with scale parameter 4 and shape parameter 2, and (d) Burr$(1,2,1)$–Burr distribution or the Singh-Maddala distribution (Singh and Maddala, 1976), which is commonly used to model household income. First, we fix $(n,p) = (400,20)$ and let the number of machines m increase from 10 to 500. Figure 2.1 plots the ℓ_2-error $\|\widehat{\beta} - \beta^*\|_2$ versus the number of machines, averaged over 500 replications, for all five methods. The global and distributed AHR estimators have almost identical performance, thus corroborating our theoretical results. The DC-AHR estimator only performs well under symmetric errors and suffers from non-negligible bias if the errors come from asymmetric distributions. This is largely expected because the robustification parameter for a local AHR estimator is tuned by a small subset of the data and results in a bias scaling with the local sample size. After averaging, this bias will not be offset when the number of machines increases. This points out a key drawback of the one-shot averaging approach when dealing with skewed data distributed across local machines. The DC-OLS method has decaying estimation error as m grows, but at a slower rate compared to the global and the distributed AHR estimators. The boxplots in Figure 2.2 show that the DC-OLS method often produces very poor estimates with high variability, while the distributed AHR method exhibits high degree of robustness.

Turning to uncertainty quantification, we construct approximate 95% confidence intervals for the slope coefficients based on distributed OLS and AHR methods. As before, we set $(n,p) = (400,20)$ and let m increase from 10 to 500. Table 2.1 shows the average coverage probabilities and widths, with standard errors in parentheses, across all slope coefficients based on 500 Monte Carlo simulations. Across all the settings, the AHR-based confidence intervals are consistently accurate with tight width and reliable with high coverage. In the presence of heavy-tailed errors, the OLS-based confidence intervals tend to be wider, and standard errors of the interval width are also larger than those of the AHR method by one order of magnitude.

2.6.2 Distributed regularized Huber regression

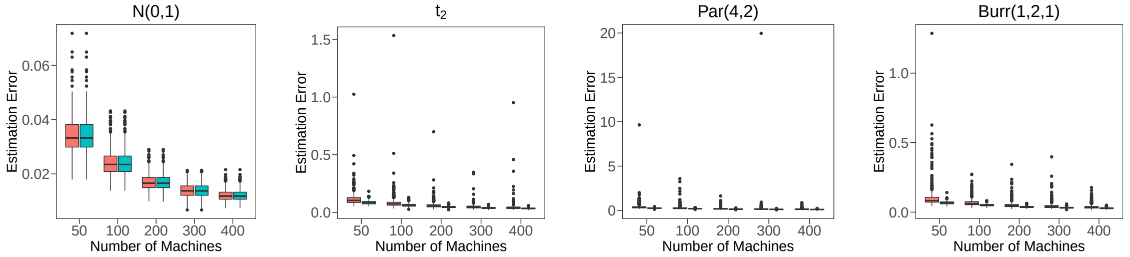
In the high-dimensional setting where the dimension p exceeds the sample size n, we compare four methods across a range of settings: (1) centralized ℓ_1-penalized AHR estimator; (2) DC ℓ_1-penalized AHR estimator; (3) centralized Lasso; and (4) distributed regularized AHR estimator with $T = \lfloor \log(m) \rfloor$ rounds of communication and with a local Lasso estimator as the initialization. All four methods involve a regularization parameter λ, which will be tuned by a held-out validation set of size $\lfloor 0.25N \rfloor$. The robustification parameter τ in methods (1), (2) and (4) is selected by the self-tuning principle proposed by Wang et al. (2021).

The simulated data $\{(y_i, x_i)\}_{i=1}^N$ is generated from a heteroscedastic model $y_i = x_i^{\mathrm{T}}\beta^* + c^{-1}(x_i^{\mathrm{T}}\beta^*)^2\varepsilon_i$, where $\beta^* = (1.5, 1.5, 1.5, 1.5, 1.5, 0, \ldots, 0)^{\mathrm{T}} \in \mathbb{R}^p$, $x_i = (1, x_{i2}, \ldots, x_{ip})^{\mathrm{T}}$ with $x_{ij} \sim \mathcal{N}(0,1)$ for $j = 2, \ldots, p$, and $c = \sqrt{3}\|\beta^*\|_2^2$. The regression errors ε_i are generated from one of the four distributions considered in Section 2.6.1, which are $\mathcal{N}(0,1)$, t_2 (heavy-tailed and symmetric), $\mathrm{Par}(4,2)$ and $\mathrm{Burr}(1,2,1)$ (heavy-tailed and skewed). We fix $(n, p) = (250, 1000)$ and let m increase from 10 to 50. Figure 2.3 plots the ℓ_2 error $\|\widehat{\beta} - \beta^*\|_2$ versus the number of machines m, averaged over 100 replications, for all four methods. The averaging ℓ_1-penalized AHR estimator has a nondecaying estimation error as m increases, which is expected because of its sub-optimal convergence rate that scales with the local sample size n. The distributed AHR estimator with $T = \lfloor \log(m) \rfloor$ rounds of communication performs as good as the centralized AHR on the entire data set, and has much smaller estimation errors than the centralized Lasso in heavy-tailed cases. Furthermore, from the boxplots displayed in Figure 2.4 we see that the distributed AHR improves upon centralized Lasso in terms of both average performance and variability.



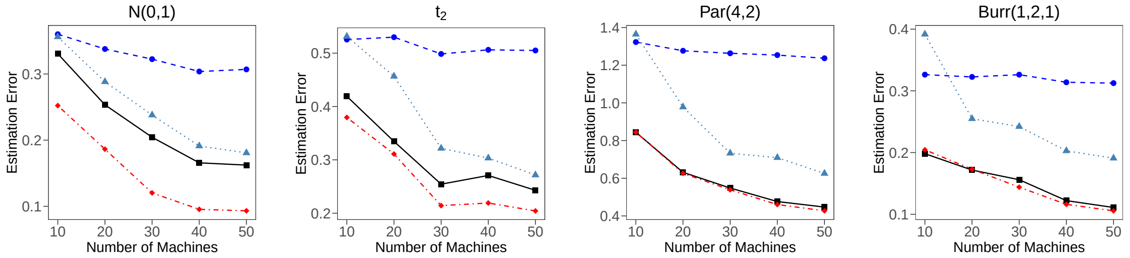
(a) (b) (c) (d)

Figure 2.1: Plots of estimation error (under ℓ_2-norm) versus number of machines when $(n,p) = (400,20)$, averaged over 500 replications. Five estimators are presented: global AHR estimator (━■━■━); DC-AHR estimator (-●- -●-); DC-OLS estimator (-●—●-); distributed OLS estimator (·▲··▲·); and distributed AHR estimator (-◆·-◆·).



(a) (b) (c) (d)

Figure 2.2: Boxplots of estimation error (under ℓ_2-norm) versus the number of machines when $(n,p) = (400,20)$ for distributed OLS estimator (🟥) and distributed AHR estimator (🟦), averaged over 500 replications.

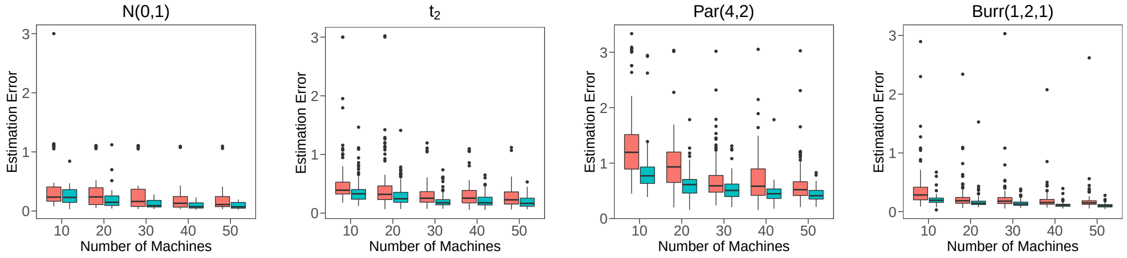


(a) (b) (c) (d)

Figure 2.3: Plots of estimation error (under ℓ_2-norm) versus the number of machines, over 100 replications, under a high-dimensional heteroscedastic model when $(n,p,s) = (250,1000,5)$. Four estimators are presented: centralized ℓ_1-penalized AHR estimator (━■━■━); DC ℓ_1-penalized AHR estimator (-●- -●-); centralized Lasso estimator (·▲··▲·); and proposed distributed regularized AHR estimator (-◆·-◆·)

Table 2.1: Coverage probabilities and widths (with standard errors in parentheses) of the normal-based CIs (averaged over all slope coefficients) for the distributed OLS and distributed AHR methods, based on 500 Monte Carlo simulations.

		$N(0,1)$		t_2		Par(4,2)		Burr(1,2,1)	
		Coverage	Width	Coverage	Width	Coverage	Width	Coverage	Width
		mean (sd)	mean (sd)	mean (sd)	mean (sd)	mean (sd)	mean (sd)	mean (sd)	mean (sd)
$m = 50$	Dist-OLS	0.93(0.011)	0.029(0.001)	0.93(0.011)	0.097(0.056)	0.93(0.012)	0.35(0.420)	0.94(0.011)	0.088(0.068)
	Dist-AHR	0.95(0.007)	0.031(0.001)	0.95(0.009)	0.077(0.007)	0.95(0.008)	0.23(0.025)	0.95(0.009)	0.058(0.006)
$m = 100$	Dist-OLS	0.93(0.012)	0.020(0.000)	0.94(0.010)	0.072(0.056)	0.93(0.012)	0.25(0.220)	0.93(0.008)	0.058(0.021)
	Dist-AHR	0.95(0.010)	0.022(0.001)	0.96(0.008)	0.058(0.005)	0.95(0.009)	0.18(0.017)	0.95(0.009)	0.044(0.004)
$m = 200$	Dist-OLS	0.93(0.011)	0.014(0.000)	0.93(0.013)	0.052(0.031)	0.93(0.010)	0.18(0.095)	0.94(0.015)	0.044(0.021)
	Dist-AHR	0.96(0.007)	0.015(0.000)	0.95(0.011)	0.043(0.003)	0.95(0.009)	0.13(0.012)	0.96(0.012)	0.034(0.003)
$m = 300$	Dist-OLS	0.93(0.013)	0.012(0.000)	0.94(0.011)	0.043(0.022)	0.94(0.011)	0.18(0.820)	0.93(0.008)	0.038(0.020)
	Dist-AHR	0.95(0.010)	0.013(0.000)	0.96(0.010)	0.036(0.003)	0.95(0.012)	0.11(0.009)	0.96(0.009)	0.028(0.002)
$m = 400$	Dist-OLS	0.93(0.010)	0.010(0.000)	0.94(0.011)	0.040(0.046)	0.93(0.008)	0.13(0.071)	0.94(0.010)	0.032(0.014)
	Dist-AHR	0.95(0.009)	0.011(0.000)	0.96(0.009)	0.031(0.002)	0.95(0.012)	0.10(0.008)	0.96(0.009)	0.025(0.002)



(a) (b) (c) (d)

Figure 2.4: Boxplots of estimation errors (under ℓ_2-norm) versus the number of machines, over 100 replications, for centralized Lasso (🟥) and distributed AHR (🟦) under a high-dimensional heteroscedastic model when $(n, p, s) = (250, 1000, 5)$.

2.7 Acknowledgement

Chapter 2, in full, is a reprint of the material as it may appear in Luo, Jiyu, Sun, Qiang; and Zhou, Wen-Xin. (2022). *Distributed adaptive Huber regression.* Computational Statistics and Data Analysis, 169, 107419. The dissertation author was the primary investigator and author of this paper.

Chapter 3

Doubly Robust Inference for Hazard Ratio under Informative Censoring with Machine Learning

3.1 Abstract

Randomized clinical trials with time-to-event outcomes have traditionally used the log-rank test followed by the Cox proportional hazards (PH) model to estimate the hazard ratio between the treatment groups. These are valid under the assumption that the right-censoring mechanism is non-informative, i.e. independent of the time-to-event of interest within each treatment group. More generally, the censoring time might depend on additional covariates, and inverse probability of censoring weighting (IPCW) can be used to correct for the bias resulting from the informative censoring. IPCW requires a correctly specified censoring time model conditional on the treatment and the covariates. Doubly robust inference in this setting has not been plausible previously due to the non-collapsibility of the Cox model. However, with the recent development of data-adaptive machine learning methods we derive an augmented IPCW (AIPCW) estimator that has the following doubly robust (DR) properties: it is model doubly robust, in that it is consistent and asymptotically

normal (CAN), as long as one of the two models, one for the failure time and one for the censoring time, is correctly specified; it is also rate doubly robust, in that it is CAN as long as the product of the estimation error rates under these two models is faster than root-n. We investigate the AIPCW estimator using extensive simulation in finite samples.

3.2 Introduction

In the analysis of time-to-event data, the Cox proportional hazards (PH) model (Cox, 1972) has been widely used to estimate the hazard ratio (HR) between two treatment groups in a randomized clinical trial, for example. The validity of the maximum partial likelihood estimator (MPLE) under the PH model relies on the non-informative censoring assumption (Fleming and Harrington, 1991); that is, the censoring time random variable is independent of the failure time random variable within each treatment group. In practice, this assumption can be violated which leads to informative censoring, and the censoring time may well depend on additional covariates. This issue was recently highlighted in Van Lancker et al. (2021), who aimed to develop procedures to select baseline covariates in order to be adjusted for in the Cox regression model. Such adjustment, however, changes the effect estimand, making it difficult to compare across different adjustment sets. Alternatively, the crude or marginal hazard ratio, as it is often referred to in the medical literature, between the two groups can still be consistently estimated using inverse probability of censoring weighting (IPCW) under the relaxed censoring assumption that the censoring time and the failure time are independent given the additional covariates.

IPCW was proposed in Robins and Finkelstein (2000) to correct for bias resulting from informative censoring of the log-rank test and, prior to that, in Robins (1993). Up until then, the main body of literature in both applied and theoretical survival analysis had assumed non-informative censoring, given the predictors in a regression model (Fleming and Harrington, 1991). A separate line of research where IPCW was called for, was under violation of the PH assumption, where it was recognized that the MPLE gave rise to a population quantity that involved the nuisance censoring distribution (Xu, 1996; Xu and O'Quigley, 2000). A series of work has since been done to correct for this bias using IPCW approaches, including Boyd et al. (2012); Hattori and Henmi (2012); Nguyen and Gillen (2017); Nuño and Gillen (2021). We note that the terminology 'IPCW' was not always mentioned in some of these works, which used the (conditional) survival distribution increments as weights in each risk set; but these are algebraically equivalent to the inverse probability of censoring weights.

The censoring distribution used in IPCW is often modeled parametrically or semi-parametrically, and the resulting IPCW estimator is consistent and asymptotically normal (CAN) if the model is correctly specified. Nguyen and Gillen (2017) proposed a survival tree approach to estimate the conditional censoring distribution given the covariates, but with no theoretical guarantee for inference. In fact, it is known that the resulting estimator is typically biased (Belloni et al., 2013).

Doubly robust (DR) approaches were developed when handling missing data (Robins, 1993; Robins et al., 1995; Scharfstein et al., 1999; Robins et al., 2000b; Robins and Rotnitzky, 2001; Van der Laan and Robins, 2003; Bang and Robins, 2005; Tsiatis, 2006). It is called doubly robust because two working models are involved, one for the outcome of interest, and one for the missing data mechanism, and the estimator is consistent as long as one of the two working models are correctly specified. When IPW is used to handle the missingness (referred to as coarsening), this usually comes down to augmentation with the coarsened data and the resulting DR estimator is an augmented IPW (AIPW) estimator (Tsiatis, 2006).

Since right censoring in survival data may be framed as a type of coarsening (Tsiatis, 2006), Rotnitzky and Robins (2005) developed an augmented IPCW (AIPCW) approach for censored survival data. For the PH model, however, this approach is not straightforward to apply to. As will be seen later, this is mainly due to the non-collapsibility of the Cox model (Martinussen and Vansteelandt, 2013; Tchetgen Tchetgen and Robins, 2012; Rava, 2021).

In this paper, we consider simultaneously the regression parameter and the nuisance baseline hazard function under the PH model. This naturally gives rise to full data estimating equations that are sums of independent and identically distributed (i.i.d.) martingales. The augmentation leads to working models for the failure time and the censoring time given the group indicator and the covariates. To specify a conditional failure time model that is compatible with the original (marginal) PH model given the group membership only, data adaptive machine learning (ML) or nonparametric methods are needed. With cross-fitting (Chernozhukov et al., 2018), the resulting AIPCW estimator has doubly robust properties not only in the classical sense, which is referred to as model doubly robust, but also rate doubly robust (Smucler et al., 2019; Hou et al., 2021). Here, rate double robustness refers to an estimator being CAN when the product of the estimation error rates under the two working models is faster than root-n, while either one of them is allowed to be arbitrarily slow.

The rest of the paper is organized as follows. In Section 3.2.1, we state the model and assumption about censoring. In Section 3.3, we take a missing data approach by constructing the AIPCW score from the full data score, and provide a detailed algorithm for the cross-fitted AIPCW estimator. Asymptotic properties of the AIPCW estimator are described in Section 3.4. In Section 3.5, we conduct simulations for the AIPCW estimator using different nuisance estimators, and also compare them with the IPCW estimators. Finally, we conclude with discussion in Section 3.6. Additional materials are provided in the Appendix.

3.2.1 Model and assumption

Let T and C be the failure time and the censoring time, respectively. Denote $X = \min(T, C)$, and $\Delta = I(T \leq C)$. Denote also $Y(t) = I(X \geq t)$ the at-risk process, and $N(t) = I(X \leq t, \Delta = 1)$ the failure event counting process. We consider the two-group survival setting where A is a binary group indicator. For a randomized trial, this can be the treatment groups. Let Z be a p-dimensional vector of baseline covariates. We assume that the data consist of n independent and identically distributed (i.i.d.) copies of the random vectors $O = (X, \Delta, A, Z)$.

Assumption 1. *(informative censoring) $C \perp T \mid (A, Z)$.*

We assume the PH model for the two-group survival:

$$\lambda(t|A) = \lambda_0(t)\exp(\beta A), \tag{3.1}$$

where $\lambda(t|A)$ denotes the group-specific hazard function of T, β is the log hazard ratio, and $\lambda_0(t)$ is the baseline hazard function.

3.3 Doubly robust inference

In this section following Tsiatis (2006) we treat right censoring as a coarsened data problem. We start with a set of full data score functions under the PH model, and show that when IPCW is applied to this set of full data score functions we obtain the familiar IPCW estimator under the Cox model (Boyd et al., 2012). We then mimic the approach of Rotnitzky and Robins (2005) to augment the IPCW score functions and arrive at a doubly robust AIPCW estimator. Finally, for inference purposes we introduce cross-fitting and describe the implementation of the cross-fitted AIPCW estimator.

3.3.1 Full data score functions

With the full data vector (T, A, Z), we follow the commonly used NPMLE approach for the semiparametric PH model and consider simultaneously the log hazard ratio β and the cumulative baseline hazard $\Lambda_0(t) = \int_0^t \lambda_0(u)du$, which is discretized to jumps at the observed event times only (Nielsen et al., 1992).

As in Fleming and Harrington (1991), we define the full data counting process

$N_T(t) = I(T \leq t)$ and the full data at-risk process $Y_T(t) = I(T \geq t)$. Let

$$M_T(t;\beta,\Lambda_0) = N_T(t) - \int_0^t Y_T(u)e^{\beta A}d\Lambda_0(u).$$

Then $M_T(t;\beta,\Lambda_0)$ is the full data martingale with respect to the full data filtration $\mathcal{F}_t^f = \{N_T(u), Y_T(u^+), A; 0 \leq u \leq t\}$ under model (3.1).

We have the following full data score functions for a single copy of the data:

$$\begin{aligned} D_1^f(\beta,\Lambda_0,t) &= dM_T(t;\beta,\Lambda_0), && (3.2)\\ D_2^f(\beta,\Lambda_0) &= \int_0^\tau A dM_T(t;\beta,\Lambda_0). && (3.3)\end{aligned}$$

where τ is the maximum follow-up time. Note that $D_1^f(\beta,\Lambda_0,t)$ is a martingale increment that is often used in survival analysis; see for example, Lu and Ying (2004). For each t, the true values of the parameters β and Λ_0 satisfy

$$E\{D_1^f(\beta,\Lambda_0,t)\} = 0 \text{ and } E\{D_2^f(\beta,\Lambda_0)\} = 0.$$

3.3.2 IPCW score functions

In survival analysis, it's common to consider the quantity

$$M(t) = N(t) - \int_0^t Y(u)e^{\beta A}d\Lambda_0(u),$$

even though it is not a martingale under informative censoring. We define $S_c(t|A,Z) = P(C \geq t|a,Z)$ the conditional survival function of C, $\widetilde{\Delta}(t) = I(\min(T,t) < C)$, and denote

$$\begin{aligned} dM^w(t;\beta,\Lambda_0,S_c) =& S_c(t|A,Z)^{-1}\widetilde{\Delta}(t)dM_T(t;\beta,\Lambda_0)\\ =& S_c(t|A,Z)^{-1}\left\{dN(t) - Y(t)e^{\beta A}d\Lambda_0(t)\right\}. && (3.4)\end{aligned}$$

The expression (3.4) then leads to the IPCW score functions:

$$D_1^w(\beta, \Lambda_0, t; S_c) = dM^w(t; \beta, \Lambda_0, S_c), \tag{3.5}$$

$$D_2^w(\beta, \Lambda_0; S_c) = \int_0^\tau A dM^w(t; \beta, \Lambda_0, S_c). \tag{3.6}$$

With n copies of i.i.d. data, this gives the following IPCW weighted estimating equations:

$$\frac{1}{n}\sum_{i=1}^{n} D_{1i}^w(\beta, \Lambda, t; S_c) = 0,$$

$$\frac{1}{n}\sum_{i=1}^{n} D_{2i}^w(\beta, \Lambda; S_c) = 0,$$

which after some algebra, can be combined to arrive at the IPCW estimating equation (Boyd et al., 2012):

$$\sum_{i=1}^{n}\int_0^\tau \widehat{S}_c(t|A_i, Z_i)^{-1}\left\{A_i - \frac{\widetilde{S}^{(1)}(\beta, t; \widehat{S}_c)}{\widetilde{S}^{(0)}(\beta, t; \widehat{S}_c)}\right\} dN_i(t) = 0,$$

where $\widetilde{S}^{(l)}(\beta, t; S_c) = \sum_{j=1}^{n} A_j^l S_c(t|A_j, Z_j)^{-1} Y_j(t) e^{\beta A_j}$ for $l = 0, 1$, and $\widehat{S}_c(t|A, Z)$ is some consistent estimator of $S_c(t|A, Z)$.

3.3.3 AIPCW score functions

The consistency of the IPCW estimator relies critically on $S_c(t|A, Z)$ being correctly specified. When it is misspecified, the IPCW estimator is biased. Rotnitzky and Robins (2005) provides an augmentation approach for an IPCW estimator in survival analysis, so that it has the doubly robust property to be detailed later. However, their approach cannot be directly applied because we have not only different weights for different individuals in the data set, but also different weights for each risk set. To this end, it is helpful to augment the martingale *increment* in (3.4) instead.

Denote $N_c(t) = I(X \leq t, \Delta = 0)$ the counting process for the censoring event, and $\Lambda_c(t|A,Z) = \int_0^t S_c(u|A,Z)^{-1} d\{1 - S_c(u|A,Z)\}$ the cumulative hazard function of C given A, Z. Then $M_c(t; S_c) = N_c(t) - \int_0^t Y(u) d\Lambda_c(u|A,Z)$ is the martingale corresponding to the censoring event counting process with respect to its natural history filtration. Also denote $S(t|A,Z) = P(T \geq t|A,Z)$. Define

$$
\begin{aligned}
& dM^{aug}(t; \beta, \Lambda_0, S, S_c) \\
=& dM^{w}(t; \beta, \Lambda_0, S_c) + \int_0^t E\{dM_T(t; \beta, \Lambda_0)|A,Z,T \geq u\} \frac{dM_c(u; S_c)}{S_c(u|A,Z)} \qquad (3.7) \\
=& \frac{dN(t) - Y(t) d\Lambda_0(t) e^{\beta A}}{S_c(t|A,Z)} - J(t; S, S_c) \left\{ dS(t|A,Z) + S(t|A,Z) e^{\beta A} d\Lambda_0(t) \right\}, \qquad (3.8)
\end{aligned}
$$

where $J(t; S, S_c) = \int_0^t S(u|A,Z)^{-1} S_c(u|A,Z)^{-1} dM_c(u; S_c)$. The last '=' above used the fact that, for $u \leq t$,

$$
\begin{aligned}
E\{N_T(t)|A,Z,T \geq u\} &= P(T \leq t|A,Z,T \geq u) = 1 - \frac{S(t|A,Z)}{S(u|A,Z)}, \qquad (3.9) \\
E\{Y_T(t)|A,Z,T \geq u\} &= P(T \geq t|A,Z,T \geq u) = \frac{S(t|A,Z)}{S(u|A,Z)}. \qquad (3.10)
\end{aligned}
$$

The above leads to the AIPCW score functions:

$$
\begin{aligned}
D_1(\beta, \Lambda_0, t; S, S_c) &= dM^{aug}(t; \beta, \Lambda_0, S, S_c), \qquad (3.11) \\
D_2(\beta, \Lambda_0; S, S_c) &= \int_0^\tau A \cdot dM^{aug}(t; \beta, \Lambda_0, S, S_c). \qquad (3.12)
\end{aligned}
$$

Assumption 2. *$S(\tau|a,z) > c$ for $a \in \{0,1\}, z \in \mathcal{Z}$ and some $c > 0$.*

Assumption 3. *$S_c(\tau|a,z) > c$ for $a \in \{0,1\}, z \in \mathcal{Z}$ and some $c > 0$.*

In Theorem 7 below, we will show that (3.11) and (3.12) are doubly robust score functions. We use superscript o to denote the truth; for example, $S^o(t|A,Z)$, $S_c^o(t|A,Z)$ and $\Lambda_c^o(t|A,Z)$ denote the true $S(t|A,Z)$, $S_c(t|A,Z)$ and $\Lambda_c(t|A,Z)$, respectively. Also let β^o and Λ_0^o denote the true values of the parameters of interest. We assume the following:

Theorem 7. *Under Assumptions 1-3, if either $S = S^o$ or $S_c = S_c^o$,*

$$E\{D_1(\beta^o, \Lambda_0^o, t; S, S_c)\} = E\{D_2(\beta^o, \Lambda_0^o; S, S_c)\} = 0.$$

The above theorem states that the scores (D_1, D_2) identifies the true parameters (β^o, Λ_0^o), as long as one of the two survival functions, $S(t|A,Z)$ and $S_c(t|A,Z)$, is true.

Given n i.i.d. data points, we estimate β^o, Λ^o by solving

$$\frac{1}{n}\sum_{i=1}^n D_{1i}(\beta, \Lambda_0, t; S, S_c) = 0, \tag{3.13}$$

$$\frac{1}{n}\sum_{i=1}^n D_{2i}(\beta, \Lambda_0; S, S_c) = 0. \tag{3.14}$$

Solving for (3.13) gives

$$\widetilde{\Lambda}_0(\beta, t; S, S_c) = \int_0^t \frac{\frac{1}{n}\sum_{i=1}^n S_c(u|A_i, Z_i)^{-1} dN_i(u) - J_i(u; S, S_c) dS(u|A_i, Z_i)}{\mathcal{S}^{(0)}(\beta, u; S, S_c)}, \tag{3.15}$$

where

$$\mathcal{S}^{(l)}(\beta, t; S, S_c) = \frac{1}{n}\sum_{i=1}^n A_i^l e^{\beta A_i}\{S_c(u|A_i, Z_i)^{-1} Y_i(t) + J_i(t; S, S_c) S(t|A_i, Z_i)\}$$

for $l = 0, 1$. Further define $\bar{A}(\beta, t; S, S_c) = \mathcal{S}^{(1)}(\beta, t; S, S_c)/\mathcal{S}^{(0)}(\beta, t; S, S_c)$. After plugging (3.15) into (3.14), we have:

$$\begin{aligned} & U(\beta; S, S_c) \\ =& \frac{1}{n}\sum_{i=1}^n \int_0^\tau \{S_c(t|A_i, Z_i)^{-1} dN_i(t) - J_i(u; S, S_c) dS(t|A_i, Z_i)\}\{A_i - \bar{A}(\beta, t; S, S_c)\} = 0 \end{aligned} \tag{3.16}$$

It's worth noting that like the partial likelihood score equation, (3.16) is not a sum of *i.i.d* terms due to $\bar{A}(\beta, t; S, S_c)$. As seen from the derivation leading to (3.8), the augmentation to the weighted martingale increment, which is linear in $N(t)$ and $Y(t)$, is the result of augmentation to the weighted $N(t)$ and $Y(t)$, respectively. It is apparent that $S_c(t|A_i,Z_i)^{-1}dN_i(t) - J_i(t;S,S_c)dS(t|A_i,Z_i)$ is the augmented weighted $dN_i(t)$, and the augmented weighted $Y_i(t)$'s give rise to the quantities $\mathcal{S}^{(l)}(\cdot)$ and $\bar{A}(\cdot)$, which are the analogies of similar quantities under the usual Cox model. For example, $\bar{A}(\beta,t;S,S_c)$ corresponds to the empirical mean of the treatment random variable A among subjects who fail at time t, which we may denote by $\rho(\beta,t)$.

The quantity $\rho(\beta,t)$ was implied in Rotnitzky and Robins (2005), as a nuisance parameter, based on the partial likelihood score function. It would, however, not be straightforward to construct compatible models for $\rho(\beta,t)$, which is defined on nested risk sets over time. The set of full data estimating functions we consider here, simultaneously for β and Λ_0, on the other hand, lead naturally to models for S and S_c.

3.3.4 Cross-fitted AIPCW estimator

In practice, both survival functions $S(t|A,Z)$ and $S_c(t|A,Z)$ are unknown and need to be estimated by some estimator $\widehat{S}(t|A,Z)$ and $\widehat{S}_c(t|A,Z)$. Parametric and semiparametric models, like the Cox model and the accelerated failure time (AFT) model, are often applied since their theoretical properties are well-studies and with little requirement on the computing power. However, these models can be misspecified, especially for $S(t|A,Z)$ due to the non-collapsibility of the Cox model. ML or nonparametric methods, like splines (Gray, 1992; Kooperberg et al., 1995a) and random survival forest (Ishwaran et al., 2008), offer a good alternative. ML or nonparametric estimators, however, do not have root-n convergence rate, which makes it difficult to conduct inference. We will show that the asymptotic normality can be established if we also apply cross-fitting, where the entire sample is first split into k folds, and for each fold, we estimate the nuisance functions using only the out-of-fold sample. Details of the cross-fitted AIPCW estimator $\widehat{\beta}$ are described in Algorithm 4. Heuristically, cross-fitting works by inducing independence between the nuisance parameter estimators and the rest of the quantities in the scores, thereby allowing

Algorithm 4 *k*-fold Cross-fitted AIPCW estimation of β

```
Input: A sample of $n$ observations that are split into $k$ folds of equal size with index sets $I_1, I_2, \ldots, I_k$.
for each fold indexed by $m$ do
    estimate nuisance functions $(\widehat{S}^{(-m)}, \widehat{S}_c^{(-m)})$ using the out-of-fold sample indexed by $I_{-m} := \{1, \ldots, n\} \setminus I_m$ to form the following estimating equations
        $\frac{1}{|I_m|} \sum_{i \in I_m}^{n} D_{1i}(\beta, \Lambda_0, t; \widehat{S}^{(-m)}, \widehat{S}_c^{(-m)}) = 0,$    (3.17)
        $\frac{1}{|I_m|} \sum_{i \in I_m}^{n} D_{2i}(\beta, \Lambda_0; \widehat{S}^{(-m)}, \widehat{S}_c^{(-m)}) = 0.$    (3.18)
    By first solving for $\Lambda_0(t)$ using (3.17) and plug into (3.18) we get the $m$-th fold estimating equation for $\beta$ as
        $U_m(\beta; \widehat{S}^{(-m)}, \widehat{S}_c^{(-m)})$
        $= \frac{1}{|I_m|} \sum_{i \in I_m} \int_0^{\tau} \left\{ \frac{dN_i(t)}{\widehat{S}_c^{(-m)}(t|A_i, Z_i)} - J_i(t; \widehat{S}^{(-m)}, \widehat{S}_c^{(-m)}) d\widehat{S}^{(-m)}(t|A_i, Z_i) \right\}$
        $\times \{A_i - \bar{A}_m(\beta, t; \widehat{S}^{(-m)}, \widehat{S}_c^{(-m)})\},$
    where $\bar{A}_m$ is $\bar{A}$ but evaluated using only data from fold $I_m$.
end for
Output: $\widehat{\beta}$, the solution to
        $\frac{1}{k} \sum_{m=1}^{k} U_m(\beta; \widehat{S}^{(-m)}, \widehat{S}_c^{(-m)}) = 0.$
```

asymptotic normality to be established (Smucler et al., 2019; Hou et al., 2021). Additional notations involving cross-fitted quantities are collected in Appendix B.1.

3.4 Asymptotic Properties

Let $O^{\dagger}$ denote a sample of n i.i.d. data vectors $\{(X_i^{\dagger}, \Delta_i^{\dagger}, A_i^{\dagger}, Z_i^{\dagger}), i = 1, \ldots, n\}$ used for estimating $\widehat{S}$ and $\widehat{S}_c$. Let (X, Δ, A, Z) be a data vector independent of $O^{\dagger}$ and drawn from the same distribution as $O^{\dagger}$. Define

$$\left\|\widehat{S}-S^{*}\right\| = E^{\dagger}\left(E\left[\left\{\sup_{t\in[0,\tau]}\left|\widehat{S}(t;A,Z)-S^{*}(t;A,Z)\right|\right\}^{2}\right]\right),$$
$$\left\|\widehat{S}_{c}-S_{c}^{*}\right\| = E^{\dagger}\left(E\left[\left\{\sup_{t\in[0,\tau]}\left|\widehat{S}_{c}(t;A,Z)-S_{c}^{*}(t;A,Z)\right|\right\}^{2}\right]\right),$$

where $E^{\dagger}$ denotes expectation taken with respect to $O^{\dagger}$, and E denotes expectation taken with respect to O conditional on $O^{\dagger}$.

Assumption 4 (Uniform Convergence)**.** *There exist $S^{*}(t;A,Z)$ and $S_{c}^{*}(t;A,Z)$ satisfying Assumptions 2-3 such that $\|\widehat{S}-S^{*}\| = o(1)$ and $\|\widehat{S}_{c}-S_{c}^{*}\| = o(1)$.*

Theorem 8. *Under Assumptions 1-4 and some regularity conditions, if either $S^{*}=S^{o}$ or $S_{c}^{*}=S_{c}^{o}$, then $\widehat{\beta}\xrightarrow{p}\beta^{o}$.*

Following the notation of Wang et al. (2022), we let O denote a sample of n i.i.d data vectors $\{(X_j,\Delta_j,A_j,Z_j), j=1,\dots,n\}$ that is independent from $O^{\dagger}$, and drawn from the same distribution as O, and define the cross-integral product as

$$\begin{aligned}
&\mathcal{D}^{\dagger}(\widehat{S},\widehat{S}_{c};S^{o},S_{c}^{o})\\
=&E^{\dagger}\left\{E\left[\left|\int_{0}^{\tau}\{A-\bar{A}(t;\beta^{o},S^{o},S_{c}^{o})\}\int_{0}^{t}\left\{\frac{d\widehat{S}(t;A,Z)}{\widehat{S}(u;A,Z)}-\frac{dS^{o}(t;A,Z)}{S^{o}(u;A,Z)}\right\}\right.\right.\right.\\
&\left.\left.\left.\times\left\{\frac{dM_{c}(u;A,Z,\widehat{S}_{c})}{\widehat{S}_{c}(t;A,Z)}-\frac{dM_{c}(u;A,Z,S_{c}^{o})}{S_{c}^{o}(t;A,Z)}\right\}\right|\right]\right\}\\
&+E^{\dagger}\left\{E\left[\max_{l\in\{0,1\}}\left|\int_{0}^{\tau}\{\bar{A}(t;\beta^{o},S^{o},\widehat{S}_{c})-\bar{A}(t;\beta^{o},S^{o},S_{c}^{o})\}\right.\right.\right.\\
&\left.\left.\left.\times J(t;A,Z,S^{o},\widehat{S}_{c})^{l}\{d\widehat{S}(t;A,Z)-dS^{o}(t;A,Z)\}\right|\right]\right\}.
\end{aligned}$$

where with a slight abuse of notation, we use E here to denote the expectation taken with respect to the sample O with n observations conditional on the sample $O^{\dagger}$.

Assumption 5 (Rate Condition)**.** $(S^*, S_c^*) = (S^o, S_c^o)$ *and*

$$\left\|\widehat{S} - S^o\right\| \left\|\widehat{S}_c - S_c^o\right\| + \mathcal{D}^\dagger(\widehat{S}, \widehat{S}_c; S^o, S_c^o) = o(n^{-1/2}).$$

The rate condition essentially requires that the product of the error rate of $\widehat{S}$ and $\widehat{S}_c$ is faster than root-n (Smucler et al., 2019). Due to the involvement of the time component in time-to-event analysis, an integral product of the errors like $\mathcal{D}^\dagger(\widehat{S}, \widehat{S}_c; S^o, S_c^o)$ is also required. Interested readers can refer to Ying (2023) for a thorough discussion on the role of this integral product term in survival analysis.

Theorem 9. *In addition to Assumptions 1-4 and some regularity conditions, if the rate condition Assumption 5 hold, we have*

$$\widehat{\sigma}^{-1}\sqrt{n}(\widehat{\beta} - \beta^o) \xrightarrow{d} N(0,1),$$

where the expressions for $\widehat{\sigma}^2 := \widehat{\sigma}^2(\widehat{\beta})$ *is provided in Appendix B.1.*

Theorem 9 establishes the rate DR property. Traditionally, the doubly robust inference is established assuming both working models are parametric or semiparametric with one of the nuisance estimators converging at the root-n rate, referred to as the model DR property. Although our estimator is also model DR, it is not helpful here due to the non-collapsibility of the Cox model. Non-collapsibility implies that any parametric or semiparametric conditional outcome models we specify will not be correct. To enable the possible use of ML/non-parametric models, we here establish a rate DR result which states that if all nuisance estimators converge to the truth and that their cross-product rate is faster than root-n, the proposed AIPCW estimator is CAN even if one of the nuisance estimators converges arbitrarily slowly.

Note that under the model DR case, if both nuisance function estimators are of root-n rate and only one of them is correctly specified, the AIPCW estimator is still CAN, but the asymptotic variance is rather complicated. In this case, resampling methods such as bootstrap (Efron, 1992) may be used to estimate the variance.

In Chapter 4, we will consider an extension of the two-group survival to observational data, which is more general than the current setting here. The regularity assumptions required and the proofs of Theorem 8 and 7 are also simplified versions of the assumptions and proofs we will present in Chapter 4, so we omit those details here.

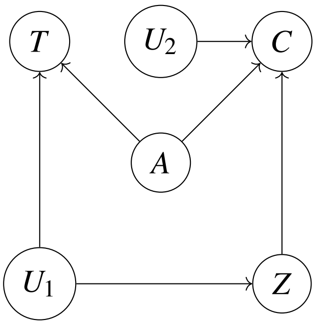
3.5 Simulation

In this section, we compare the performance of the cross-fitted AIPCW estimators $\widehat{\beta}$ using different working models, against different IPCW estimators and the MPLE. We consider sample sizes $n = 500$ and $n = 1000$, and 1000 data sets are simulated for each setting, which corresponds to margin of error of about $+/-1.35\%$ for the coverage probability of nominal 95% confidence intervals. Five-fold cross-fitting is used.

For data generation, we first follow the diagram in Figure 3.1(a) and generate $U_1 \sim$ Unif (-1, 1), $A \sim$ Bernoulli (0.5), $Z_1 \sim N(0.5U_1, 1)$, $Z_2 \sim N(U_1^2, 0.09)$, and $T = -\log(0.5U_1 + 0.5)e^A$. Here, T follows the PH model (3.1) with $\beta^o = -1$ and $\lambda_0^o(t) = 1$.

We consider two scenarios of data generating process for the censoring time C, as described in Figure 3.1(b). Both scenarios have around 25% samples administratively censored at $\tau = 1$, and 40% of the remaining samples censored during follow-up. Note that administrative censoring works in the same way for T and C, i.e. those events are consider as 'censored' for both the estimation of S and the estimation of S_c. It is obvious that Scenario 1 can be correctly modeled. Scenario 2 is designed such that most commonly used semiparametric models fail. As it turns out, under Scenario 2 $S_c(\tau|A, Z)$ can be very close to zero for some values of A and Z, leading to possible violation of Assumptions 2 and 3. This echoes the argument made in D'Amour et al. (2021) that the overlap assumption needed for DR estimates often fails in practice.

We consider three types of working models: PH model using the R package 'survival'; splines (Kooperberg et al., 1995a) using the R package 'polspline'; and random



(a)

Scenario	Data generating process for C
1: Cox PH	$\lambda_c(t) = \exp(-1+2Z_2)$
2: Mixture	$\log(U_2) \sim \mathrm{N}(0,1)$ $Z_1 > 0: \quad \log(C) = -0.2A - 2\sqrt{\lvert Z_2 \rvert} + 0.3U_2$ $Z_1 \leq 0: \quad \log(C) = 2.4 - 0.3A + 0.5\sqrt{\lvert Z_1 \rvert} + 0.5\sqrt{\lvert Z_2 \rvert} - U_2$

(b)

Figure 3.1: (a) Variable diagram. (b) Data generating process for C.

survival forest (RSF) (Ishwaran et al., 2008) using the R package ‘randomForestSRC’. We set splitrule = ’bs.gradient’ for RSF, while keeping all the others settings as default. We study 7 different combinations of working models for the proposed AIPCW estimator: Cox-Cox, Cox-spline, Cox-RSF, spline-Cox, RSF-Cox, spline-spline, and RSF-RSF, where the first part in the names denotes the model for S and the second part denotes the model for S_c. It is worth noting that due to the non-collapsibility of the Cox model, a semiparametric conditional model for S is almost always misspecified. Therefore the consistency of AIPCW-Cox-Cox, AIPCW-Cox-spline and AIPCW-Cox-RSF relies on the correct specification of the censoring model. We also note that the convergence rate of the spline and RSF is largely unknown, which depends on the choice of tuning parameters. See Discussion for more on this.

We also investigate the performance of MPLE and various IPCW estimators: IPCW-Cox, IPCW-spline, IPCW-RSF, IPCW-A and IPCW-1. More specifically, IPCW-A estimates S_c using the product-limit estimator for each group indicated by A, while IPCW-1 estimates S_c using the product-limit estimator on the entire sample. Robust variance estimator from Boyd et al. (2012) is used to estimate the model standard errors of the IPCW estimators. Standard errors for the cross-fitted AIPCW estimators are estimated using Theorem 9, which assumes both S and S_c models are correctly specified.

To avoid numerical problems, we impose a minimum on $\widehat{S}^{(-m)}(t|A,Z)$ and $\widehat{S}_c^{(-m)}(t|A,Z)$ in the above, so that values below 0.01 are trimmed to be 0.01. Finally, as a benchmark, we also fit model (3.1) to the full data without censoring.

The simulation results for Scenarios 1 and 2 are reported in Tables 3.1 and 3.2, respectively. It is immediate that under informative censoring, MPLE, IPCW-1 and IPCW-A have substantial bias leading to poor coverage of the confidence intervals (CI). Under Scenario 1 where the censoring model is correctly specified as Cox, the other three IPCW estimators (-Cox, -spline, -RSF) all appear to perform reasonably well. All seven AIPCW estimators also perform well under Scenario 1, with AIPCW-Cox-RSF having larger bias compared to the rest.

Under Scenario 2, IPCW-Cox appears more biased than IPCW-spline and IPCW-RSF, as expected. But even for the latter two estimators, their SE's severely under-estimate the SD's, leading to poor coverage of the CI's. This also points to the known fact that inference is not guaranteed when ML or nonparametric methods are used in IPCW, as discussed earlier. AIPCW-Cox-Cox also has large bias under Scenario 2, as expected. The rest six AIPCW's are less biased. For the larger sample size $n = 1000$, AIPCW using two ML or nonparametric methods appears to have the least bias, with close to nominal coverage probabilities. Finally we note that, under Scenario 2, spline-based AIPCWs tend to have larger variance. This might be explained by the fact that splines are less stable near the boundary τ, which under Scenario 2 has small $\widehat{S}_c(\tau|A,Z)$ for some values of A and Z as mentioned earlier.

Table 3.1: Simulation results under Scenario 1. Data are generated following Figures 3.1(a) and (b) with $\beta^o = -1$. Red indicates that the model or approach is invalid.

Sample Size	Estimators	Bias	SD	SE	CP
$n = 500$	AIPCW-Cox-Cox	0.002	0.196	0.191	0.94
	AIPCW-Cox-spline	-0.001	0.198	0.190	0.94
	AIPCW-Cox-RSF	0.023	0.197	0.207	0.96
	AIPCW-spline-Cox	0.005	0.185	0.177	0.94
	AIPCW-RSF-Cox	0.005	0.189	0.178	0.94
	AIPCW-spline-spline	0.002	0.185	0.177	0.94
	AIPCW-RSF-RSF	0.002	0.192	0.190	0.95
	IPCW-Cox	-0.006	0.186	0.179	0.94
	IPCW-spline	-0.005	0.188	0.179	0.94
	IPCW-RSF	0.008	0.190	0.177	0.93
	IPCW-A	-0.221	0.180	0.162	0.70
	IPCW-1	-0.221	0.179	0.162	0.70
	MPLE	-0.205	0.175	0.167	0.76
	Full data	0.002	0.103	0.099	0.93
$n = 1000$	AIPCW-Cox-Cox	-0.008	0.137	0.134	0.94
	AIPCW-Cox-spline	-0.010	0.138	0.133	0.94
	AIPCW-Cox-RSF	0.019	0.141	0.153	0.97
	AIPCW-spline-Cox	0.001	0.127	0.123	0.94
	AIPCW-RSF-Cox	0.002	0.130	0.125	0.94
	AIPCW-spline-spline	0.001	0.127	0.123	0.94
	AIPCW-RSF-RSF	-0.005	0.134	0.134	0.95
	IPCW-Cox	-0.009	0.130	0.128	0.94
	IPCW-spline	-0.007	0.135	0.128	0.94
	IPCW-RSF	0.011	0.134	0.128	0.95
	IPCW-A	-0.225	0.126	0.114	0.51
	IPCW-1	-0.224	0.126	0.114	0.51
	MPLE	-0.207	0.122	0.118	0.58
	Full data	-0.003	0.069	0.07	0.94

SD: standard deviation; SE: standard error; CP: coverage probability of nominal 95% CI

Table 3.2: Simulation results under Scenario 2. Data are generated following Figures 3.1(a) and (b) with $\beta^o = -1$. Red indicates that the model or approach is invalid.

Sample Size	Estimators	Bias	SD	SE	CP
$n = 500$	AIPCW-Cox-Cox	-0.129	0.285	0.276	0.93
	AIPCW-Cox-spline	-0.029	0.604	0.623	0.97
	AIPCW-Cox-RSF	-0.064	0.249	0.243	0.93
	AIPCW-spline-Cox	-0.068	0.282	0.256	0.93
	AIPCW-RSF-Cox	-0.034	0.275	0.250	0.93
	AIPCW-spline-spline	0.038	0.578	0.585	0.96
	AIPCW-RSF-RSF	-0.039	0.264	0.238	0.93
	IPCW-Cox	-0.114	0.266	0.174	0.77
	IPCW-spline	-0.046	0.452	0.192	0.68
	IPCW-RSF	-0.088	0.257	0.179	0.80
	IPCW-A	-0.227	0.184	0.170	0.74
	IPCW-1	-0.226	0.183	0.166	0.72
	MPLE	-0.216	0.179	0.174	0.77
	Full data	0.002	0.103	0.099	0.93
$n = 1000$	AIPCW-Cox-Cox	-0.127	0.195	0.192	0.90
	AIPCW-Cox-spline	-0.056	0.396	0.367	0.95
	AIPCW-Cox-RSF	-0.035	0.187	0.189	0.95
	AIPCW-spline-Cox	-0.056	0.191	0.180	0.93
	AIPCW-RSF-Cox	-0.021	0.185	0.178	0.92
	AIPCW-spline-spline	0.008	0.344	0.332	0.95
	AIPCW-RSF-RSF	-0.020	0.198	0.179	0.93
	IPCW-Cox	-0.103	0.204	0.126	0.71
	IPCW-spline	-0.045	0.377	0.146	0.63
	IPCW-RSF	-0.047	0.202	0.134	0.78
	IPCW-A	-0.220	0.127	0.120	0.56
	IPCW-1	-0.219	0.127	0.117	0.53
	MPLE	-0.211	0.123	0.123	0.61
	Full data	-0.003	0.069	0.07	0.94

SD: standard deviation; SE: standard error; CP: coverage probability of nominal 95% CI

3.6 Discussion

For the analysis of two-group survival, including for randomized clinical trials, non-informative censoring is assumed. When the simple PH model (3.1) is used with no covariates adjusted for, this requires the censoring distribution to be independent of any covariates. When this assumption is violated, the commonly used MPLE is biased and typically IPCW is used to correct that bias if the interest remains to estimate the marginal hazard ratio between the two groups. IPCW, on the other hand, requires modeling the censoring distribution, which can be wrong unless ML or nonparametric estimates are used. In this paper we have developed an AIPCW estimator that is both model DR and rate DR. Rate double robustness allows us to get around the non-collapsibility of the Cox regression model using more flexible ML or nonparametric methods for the conditional failure time model demanded by the DR construct, because almost any parametric or semiparametric would otherwise be invalid.

The theoretical results require certain rate condition of the estimates of the nuisance parameters. These are not always established for a given ML or nonparametric estimator. Cui et al. (2022) and Kooperberg et al. (1995b) demonstrated that under certain conditions, rate better than $n^{1/4}$ can be achieved for random survival forest and splines. Convergence rates were also studied for other ML methods. For example, a uniform rate for regression trees is shown in Wager and Walther (2015), while a root-mean-square rate is derived for neural networks (Chen and White, 1999). These results suggest that it is entirely possible to utilize even a slow converging ML method, so long as we use a fast converging ML method for the other nuisance function to achieve a better than $\sqrt{n}$ overall rate. The rates, of course, depend on the hyper-parameter values. In the simulations we used the default settings for the spline and the random survival forest. Investigation of other ML or nonparametric methods, as well as their tuning, in relationship with the performance of DR estimators, remains a topic of future work.

This work focused on two-group survival and a binary A. Generalization to continuous and/or multivariate A is conceptually straightforward although different algebra might be involved. In particular for continuous A, we would no longer have $A^2 = A$ and additional quantities like $\mathcal{S}^{(2)}$ need to be introduced.

Finally the models for S and S_c may include additional and different sets of covariates for these two models, so long as the failure time and the censoring time are independent given the common covariates Z.

The R codes for the cross-fitted AIPCW estimator as well as the simulation procedures investigated in this work are available online in
`http://github.com/charlesluo1002/DR-Cox`.

3.7 Acknowledgement

Chapter 3, with minor edits, is a reprint of the material as it may appear in Luo, Jiyu; Xu, Ronghui. (2022). *Doubly Robust Inference for Hazard Ratio under Informative Censoring with Machine Learning*. arXiv preprint arXiv:2206.02296. The dissertation author was the primary investigator and author of this paper.

Chapter 4

Doubly Robust Inference for Cox Marginal Structural Model with Informative Censoring

4.1 Abstract

The marginal structural Cox model has been widely used to draw causal inferences from observational studies with survival outcomes. The typical estimation approach under the marginal structural Cox model is inverse probability weighting, using a propensity score model for treatment assignment. Additionally censoring needs to be properly accounted for, especially when it depends on covariates. This is again typically handled using inverse probability weighting, with a censoring model given the treatment and covariates. Effort to protect against model misspecification involves augmentation, which has been a challenge in the past due to the non-collapsibility of the Cox regression model. In this work we develop an augmented inverse probability weighted estimator with doubly robust properties including rate doubly robust, that enables us to use machine learning and a large class of nonparametric methods, in order to overcome the non-collapsibility challenge. We study both the theoretical and empirical performance of the augmented inverse probability weighted estimator and

apply it to the data from a cohort of Japanese men in Hawaii followed since the 1960s in order to study the effect of mid-life alcohol exposure on late-life mortality.

4.2 Introduction

4.2.1 Background

Marginal structural Cox model (Hernán et al., 2001) has been widely used in observational studies with survival outcomes to estimate the causal hazard ratio; see for example, Cole et al. (2003); Feldman et al. (2004); Sterne et al. (2005); Hernán et al. (2006) and Buchanan et al. (2014), among many others. While the interpretation of the hazard function for causal inference has been under debate recently (Hernán, 2010; Martinussen et al., 2020), the Cox model formulation continues to be used broadly. More commonly agreed-upon interpretable quantities, such as survival probabilities, can also be obtained easily under the model.

For the definition of causal treatment effects in general, potential or counterfactual outcomes have often been considered (Neyman, 1923; Rubin, 1974), and marginal structural models (Robins et al., 2000a, MSM) are defined on the potential outcomes of interest, thereby providing a causal interpretation of their parameters. Under randomization, the absence of confounding of the relationship between the treatment and the outcome allows standard regression methods to consistently estimate the MSM parameters. Without randomization and in observational studies, under the assumption of no unmeasured confounding, inverse probability of treatment weighting has been used to estimate the causal parameters under the MSM (Robins, 1998; Robins et al., 2000a; Robins, 2000; Lunceford and Davidian, 2004; Hubbard et al., 2000; Hernán et al., 2001; Chen and Tsiatis, 2001; Zhang and Schaubel, 2011). It adjusts for the observed confounders by weighting each observation by the inverse of its propensity score, that is, the probability of receiving the treatment given the confounders.

Survival outcomes are often subject to right censoring. The non-informative censoring assumption is typically needed for the consistency of the (weighted) partial likelihood estimator under the Cox model (Fleming and Harrington, 1991); that is, the censoring time random variable should be independent of the failure time random variable within each treatment group. When the censoring time depends in addition on covariates, informative censoring occurs, and inverse probability of censoring weighting (IPCW) may be applied to consistently estimate the hazard ratio of interest. IPCW was proposed in Robins and Finkelstein (2000) to correct for bias resulting from informative censoring of the log-rank test and, prior to that, in Robins (1993). A separate line of research where IPCW was called for, was under violation of the proportional hazards assumption, where it was recognized that the partial likelihood estimator gave rise to an estimand that involved the nuisance censoring distribution (Xu, 1996; Xu and O'Quigley, 2000). A series of work has since been done to correct for this bias using IPCW approaches, including Boyd et al. (2012); Hattori and Henmi (2012); Nguyen and Gillen (2017); Nuño and Gillen (2021). We note that the terminology 'IPCW' was not always mentioned in some of these works, which used the (conditional) survival distribution increments as weights in each risk set; but these are algebraically equivalent to the inverse probability of censoring weights.

Propensity scores are unknown in practice, and similarly, the censoring probabilities given treatment and covariates. Both of them need to be estimated, and they are subject to misspecification if modeled parametrically or semiparametrically. Inconsistency in the estimation of either results in bias in the estimated causal effect of interest. For both the propensity score and the conditional censoring model, nonparametric or machine learning methods have also been proposed in the literature (Ridgeway et al., 2022). Nguyen and Gillen (2017) proposed a survival tree approach to estimate the conditional censoring distribution given the covariates. However, these approaches are without theoretical guarantees for statistical inference; in fact, it is known that the resulting estimator is typically biased (Belloni et al., 2013).

When the counterfactual outcomes and censoring are seen as missing data, augmented inverse probability weighting (AIPW) methods have been developed in order to protect against misspecification of the missing data mechanisms (Robins et al., 1995; Scharfstein et al., 1999; Robins et al., 2000a; Robins, 2000; Robins et al., 2000b; Robins and Rotnitzky, 2001; Van der Laan and Robins, 2003; Bang and Robins, 2005; Tsiatis, 2006). They possess doubly robust (DR) properties in the sense that the resulting estimator is consistent and asymptotically normal (CAN), as long as one of two sets of models is correctly specified: the missing data model(s) and a conditional outcome model.

For survival outcomes, Rotnitzky and Robins (2005) developed an augmented IPCW approach for censored survival data. Zhang and Schaubel (2012a), Bai et al. (2017) and Sjölander and Vansteelandt (2017) derived doubly robust estimators for the treatment effect defined as a contrast between the expected transformed potential failure times, i.e. the failure time that would be observed if a subject were treated or untreated, respectively. Yang et al. (2020) developed a doubly robust estimator for the structural accelerated failure time models. Petersen et al. (2014) and Zheng et al. (2016) derived targeted maximum likelihood estimators that are doubly robust after discretizing time and recasting the failure time as a binary outcome. Dukes et al. (2019) and Hou et al. (2021) proposed doubly robust estimators for the hazard difference under the additive hazards model in low and high dimensions, respectively, and Rava and Xu (2023) extended their approaches to competing risks setting.

In this paper, we consider the marginal structural Cox model. Robins (1998) derived a generic class of semiparametric estimators for the parameters of MSMs with a focus on efficiency, and without being robust against possible misspecification of the propensity score. A main challenge in developing doubly robust estimators under the Cox MSM is the non-collapsibility of the Cox regression model (Martinussen and Vansteelandt, 2013), i.e. the Cox model formulation including the proportional hazards assumption typically no longer holds when a covariate is integrated out from the model, a fact also well-known since the 1980s (Lancaster and Nickell, 1980; Gail et al., 1984; Ford et al., 1995; Xu, 1996). This gives rise to the difficulty of specifying a conditional survival outcome model that is compatible with the Cox MSM which defines the causal estimand, as illustrated in Tchetgen Tchetgen and Robins (2012) and also will become clear later in this paper.

4.2.2 Overview of the paper

In the following we derive an AIPW estimator under the Cox MSM, making use of contemporary machine learning methods that alleviate the compatibility problem described above. The approach considers simultaneously the log hazard ratio and the nuisance baseline hazard function under the Cox model. This gives rise to full data estimating equations that are sums of independent and identically distributed (i.i.d.) martingales. The augmentation leads to working models for the propensity score, the failure time, and the censoring time given the treatment and the covariates. To specify a conditional failure time model that is compatible with the original (marginal) Cox model given treatment only, data-adaptive machine learning or nonparametric methods can be used. With cross-fitting (Chernozhukov et al., 2018), the resulting AIPW estimator has doubly robust properties not only in the classical sense, which is referred to as model doubly robust, but also rate doubly robust (Rotnitzky et al., 2021; Hou et al., 2021). Here, rate double robustness refers to an estimator being CAN when the product of the estimation error rates under the two sets of working models is faster than root-n, while either one of them is allowed to be arbitrarily slow.

In the following after defining the notation, the model, and the assumptions in Section 4.3, we augment the Cox-IPW estimators in Section 4.4 of both structural parameters, the log hazard ratio, and the infinite-dimensional baseline hazard function. In Section 4.5, we establish the estimating equation for the log hazard ratio through cross-fitting. The asymptotic properties of the AIPW estimator are established in Section 4.6. Through simulations of Section 4.7, we show that our estimator outperforms the existing Cox-IPW estimator both in terms of finite sample bias and variance, for different combinations of parametric and nonparametric estimators under the propensity score and the conditional survival model. In Section 4.8 we apply our estimator to data from a cohort of Japanese men in Hawaii followed since the 1960s in order to study the effect of mid-life alcohol exposure on late-life mortality. We conclude with a discussion in Section 4.9. The proofs of all the theoretical results are given in the Supplementary Material.

4.3 Marginal Structural Cox Model

Let A be a binary treatment and let $T(0), T(1)$ be the potential failure time of a subject if s/he has been untreated or treated, respectively. Let $\lambda_{T(a)}(t)$ denote the hazard function of the potential failure time $T(a)$, $a \in \{0,1\}$. The marginal structural Cox model (Hernán et al., 2001, often referred to as the Cox MSM in the literature) postulates a model on the potential $T(a)$ by assuming

$$\lambda_{T(a)}(t) = \lambda_0(t)\exp(\beta a), \tag{4.1}$$

where $\lambda_0(t)$ is an unknown baseline hazard function, and $\beta = \log\{\lambda_{T(1)}(t)/\lambda_{T(0)}(t)\}$ is then the causal log hazard ratio, since it is a contrast between the distributions of the potential failure time outcomes under $a = 1$ versus $a = 0$. As previously mentioned in the Introduction, the Cox MSM (4.1) has been widely used in applications to estimate the causal hazard ratio, $\exp(\beta)$. The estimation of β under model (4.1) using inverse probability weighting was developed in Hernán et al. (2001).

As is typical for time-to-event outcomes, the event times of interest are subject to possible right censoring. The potential failure times $T(1), T(0)$ are therefore subject to potential right censoring at times $C(1), C(0)$, respectively. Let $\Delta(a) = \mathbf{1}\{T(a) \leq C(a)\}$ denote the potential event indicator and let $X(a) = \min\{T(a), C(a)\}$. We use T, C, X, Δ to indicate the failure, censoring, censored times and event indicator, respectively, once the treatment is received (as opposed to being counterfactual). We use $Z \in R^p$ to denote a vector of p-dimensional observed baseline covariates.

We make the following assumptions that are standard in causal inference (Hernán and Robins, 2020).

Assumption 6 (SUTVA)**.** *The potential outcomes of one subject are not affected by the treatment assignment of the other subjects, and there are no hidden versions of the treatments.*

Assumption 7 (Consistency)**.** $T = AT(1) + (1-A)T(0)$*, and* $C = AC(1) + (1-A)C(0)$*.*

Assumption 8 (Exchangeability)**.** $(T(a), C(a)) \perp A \mid Z$*, for* $a = 0, 1$*.*

Assumption 9 (Strict Positivity)**.** *There exists* $0 < \varepsilon < 1$ *such that* $\varepsilon < P(A=1|Z=z) < 1-\varepsilon$*,* $P(C > \tau|A=a, Z=z) > \varepsilon$*,* $P(T > \tau|A=a, Z=z) > \varepsilon$ *for all values of a and z, where* τ *is a maximum follow-up time.*

Traditionally survival analysis using regression models requires the censoring time to be conditionally independent of the failure time given the regressors, and this is referred to as non-informative censoring (Kalbfleisch and Prentice, 2011). Here our model under consideration is the Cox MSM (4.1), which does not involve the baseline covariates Z. Nonetheless, in this paper, we make the following informative censoring assumption that allows the censoring time to depend on the covariates.

Assumption 10 (Informative Censoring)**.** $T(a) \perp C(a) \mid Z$*, for* $a = 0, 1$*, where* $\perp$ *indicates statistical independence.*

4.4 Doubly Robust Inference

4.4.1 Full-data Estimating Function

When we consider both the counterfactual outcome and censoring as missing data, the full data, using the notion in Tsiatis (2006), is $(T(0), T(1), Z)$. From this, we can define the full data counting process $N_T^a(t) = I(T(a) \leq t)$, and the full data at-risk process $Y_T^a(t) = I(T(a) \geq t)$ for $a = 0, 1$, where $I(\cdot)$ is an indicator function. It can be shown that

$$M_T^a(t; \beta, \Lambda_0) = N_T^a(t) - \int_0^t Y_T^a(u) e^{\beta a} d\Lambda_0(u)$$

is a full data martingale with respect to the filtration $\mathcal{F}_t^a = \{N_T^a(u), Y_T^a(u^+) : 0 \leq u \leq t\}$ under model (4.1), for $a = 0, 1$, where $\Lambda_0(t) = \int_0^t \lambda_0(u) du$ (Fleming and Harrington, 1991).

We start by constructing a full data estimating function, i.e. an estimating function we would use if we were able to observe a single copy of the full data. Using the martingale property, we define full data estimating functions for $\Lambda_0(t)$ and β as follows:

$$D_1^f(t; \beta, \Lambda_0) = \sum_{a=0,1} dM_T^a(t; \beta, \Lambda_0), \tag{4.2}$$

$$D_2^f(\beta, \Lambda_0) = \sum_{a=0,1} \int_0^\tau a \cdot dM_T^a(t; \beta, \Lambda_0), \tag{4.3}$$

where τ is the maximum follow-up time defined before. We note that $D_1^f(\beta, \Lambda_0, t)$ is a martingale difference function that is often used in survival analysis; see for example, Lu and Ying (2004). For each t, the true values of β and $\Lambda_0(t)$ satisfy

$$E\{D_1^f(t; \beta, \Lambda_0)\} = 0 \text{ and } E\{D_2^f(\beta, \Lambda_0)\} = 0.$$

It can be readily verified that for a sample of size n of independent and identically distributed (i.i.d.) data, these would give the well-known Breslow's estimate of $\Lambda_0(t)$, as well as the partial likelihood score for β.

4.4.2 IPW Estimating Function

The full data are not observed. Instead, we have the observed counting process $N(t) = I(X \leq t, \Delta = 1)$ where $\Delta = I(T < C)$, and the observed at-risk process $Y(t) = I(X \geq t)$.

Define

$$M(t;\beta,\Lambda_0) = N(t) - \int_0^t Y(u)e^{\beta A} d\Lambda_0(u).$$

We note that $M(t;\beta,\Lambda_0)$ is generally not a martingale under model (4.1).

In order to bridge the gap between the full data estimating function above with the observed data, inverse probability weighting is often used. The idea is to weight an observation by its inverse probability of being sampled from the target population in general (Horvitz and Thompson, 1952), leading to a pseudo-random sample from the target population of interest. More specifically, since in observational studies treatment is typically not randomized, inverse probability of treatment weighting would give rise to a pseudo-random sample from a target population where the covariates are balanced, thereby giving a consistent estimate of the causal treatment effect when the observed data is fitted to a model with such weights (Hernán and Robins, 2020). Similarly, when informative censoring is present which depends on the covariates, the inverse probability of censoring weighting would lead to a consistent estimate of the parameter(s) of interest (Hernán et al., 2001).

Let $\pi(z) = P(A = 1|Z = z)$. The treatment weight is then

$$\frac{A}{\pi(Z)} + \frac{1-A}{\{1-\pi(Z)\}} = \frac{1}{\pi(Z)^A\{1-\pi(Z)\}^{1-A}}.$$

In addition, let $S(t;a,z) = P(T > t|A = a, Z = z)$ and $S_c(t;a,z) = P(C > t|A = a, Z = z)$ denote the conditional survival function of T and C, respectively.

We now have the IPW estimating function:

$$D_1^w(t;\beta,\Lambda_0,\pi,S_c) = \frac{dM(t;\beta,\Lambda_0)}{\pi(Z)^A\{1-\pi(Z)\}^{1-A}S_c(t;A,Z)}, \tag{4.4}$$

$$D_2^w(\beta,\Lambda_0,\pi,S_c) = \int_0^\tau \frac{A \cdot dM(t;\beta,\Lambda_0)}{\pi(Z)^A\{1-\pi(Z)\}^{1-A}S_c(t;A,Z)}. \tag{4.5}$$

It can also be readily verified that for a sample of i.i.d. observed data, $\sum_{i=1}^n D_{1i}^w(t;\beta,\Lambda_0,\pi,S_c) =$

0 gives a weighted Breslow's estimate of $\Lambda_0(t)$ and after profiling out $\Lambda_0(t)$, $\sum_{i=1}^{n} D_{2i}^{w}(\beta,\Lambda_0,\pi,S_c)$ gives the weighted partial likelihood score for β.

By Assumption 7, we have $S(t;a,Z) = P(T(a) > t|Z)$ and $S_c(t;a,Z) = P(C(a) > t|Z)$. For simplicity, we will omit writing out the explicit dependency of M and some future quantities on A and Z, unless they are needed for clarification.

Denote also $\Delta^a(t) = I\{\min(T(a),t) \leq C(a)\}$ for $a = 0,1$.

$$M(t;\beta,\Lambda_0) = A\Delta^1(t)M_T^1(t;\beta,\Lambda_0) + (1-A)\Delta^0(t)M_T^0(t;\beta,\Lambda_0), \tag{4.6}$$

where (4.6) is proved as Lemma 11 in Appendix.

4.4.3 Augmented IPW Estimating Function

The IPW estimating function is unbiased when the weights, or equivalently, $\pi(z)$ and $S_c(t;a,z)$ are known (Hernán et al., 2001). In practice, however, these quantities are often unknown and propensity score models as well as conditional censoring models are often used for the corresponding conditional distributions. When these models are misspecified, the resulting estimate of the causal hazard ratio is no longer consistent.

To protect against possible misspecification of the models, semiparametric theory has been developed to augment the IPW estimating function (Tsiatis, 2006), such that the resulting AIPW estimating function possesses the so-called doubly robust properties that will be described in detail later. In particular, Van der Laan and Robins (2003) augmented the inverse probability of treatment weighted estimating function for a binary treatment, and Rotnitzky and Robins (2005) augmented the inverse probability of censoring weighted estimating function for a survival parameter of interest. These approaches have been applied separately to the full data martingale increments similar to those here, when there is 1) confounding with non-informative censoring in Rava (2021), and 2) randomization with informative censoring in Luo and Xu (2022), respectively. In the following, we will combine the two augmentations and show that the resulting AIPW estimating function is doubly robust.

Denote the counting process for censoring events $N_c(t) = I(X \leq t, \Delta = 0)$, and $\Lambda_c(t;a,z) = \int_0^t S_c(u;a,z)^{-1} d\{1 - S_c(u;a,z)\}$ the cumulative hazard function of C given $A = a$ and $Z = z$. Define $M_c(t;a,z,S_c) = N_c(t) - \int_0^t Y(u) d\Lambda_c(u;a,z)$, then it is a martingale with respect to its natural history filtration if S_c is correctly modeled. Following Zhang and Schaubel (2012b) and Luo and Xu (2022), define also the censor-free failure process $N_T(t) = I(T \leq t)$, and the corresponding at-risk process $Y_T(t) = I(T \geq t)$. Let $M_T(t;\beta,\Lambda_0) = N_T(t) - \int_0^t Y_T(u) e^{\beta A} d\Lambda_0(u)$.

The augmented IPW estimating function based on (4.4) and (4.5) is:

$$
\begin{aligned}
D_1(t;\beta,\Lambda_0,\pi,S,S_c) =& \frac{dM(t;\beta,\Lambda_0)}{\pi(Z)^A\{1-\pi(Z)\}^{1-A} S_c(t;A,Z)} - \frac{E\{dM_T(t;\beta,\Lambda_0)|A,Z\}}{\pi(Z)^A\{1-\pi(Z)\}^{1-A}} \\
&+ \sum_{a=0,1} E\{dM_T(t;\beta,\Lambda_0)|A=a,Z\} \\
&+ \sum_{a=0,1} \frac{A^a(1-A)^{1-a}}{\pi(Z)^a\{1-\pi(Z)\}^{1-a}} \int_0^t \frac{dM_c(u;a,Z,S_c)}{S_c(u;a,Z)} \\
&\times E\{dM_T(t;\beta,\Lambda_0)|T \geq u, A=a,Z\}, \\
D_2(\beta,\Lambda_0,\pi,S,S_c) =& \int_0^\tau \left[\frac{A \cdot dM(t;\beta,\Lambda_0)}{\pi(Z) S_c(t;A,Z)} - \frac{A \cdot E\{dM_T(t;\beta,\Lambda_0)|A,Z\}}{\pi(Z)} \right. \\
&+ E\{dM_T(t;\beta,\Lambda_0)|A=1,Z\} \\
&\left. + \frac{A}{\pi(Z)} \int_0^t \frac{dM_c(u;1,Z,S_c)}{S_c(u;1,Z)} E\{dM_T(t;\beta,\Lambda_0)|T \geq u, A=1,Z\} \right].
\end{aligned}
$$

The above expectations can be written out explicitly. Using the fact that

$$
\begin{aligned}
E\{N_T(t)|A,Z\} =& P(T \leq t|A,Z) = 1 - S(t|A,Z), \\
E\{Y_T(t)|A,Z\} =& P(T \geq t|A,Z) = S(t|A,Z),
\end{aligned}
$$

and for $t \geq u$,

$$\begin{aligned}E\{N_T(t)|A,Z,T \geq u\} =& P(T \leq t|A,Z,T \geq u) = \frac{1 - S(t|A,Z)}{S(u|A,Z)},\\ E\{Y_T(t)|A,Z,T \geq u\} =& P(T \geq t|A,Z,T \geq u) = \frac{S(t|A,Z)}{S(u|A,Z)},\end{aligned}$$

we have

$$\begin{aligned}E\{dM_T(t;\beta,\Lambda_0)|A,Z\} =& -dS(t;A,Z) - S(t;A,Z)e^{\beta A}d\Lambda_0(t),\\ E\{dM_T(t;\beta,\Lambda_0)|T \geq u,A,Z\} =& -\frac{dS(t;A,Z) + S(t;A,Z)e^{\beta A}d\Lambda_0(t)}{S(u|A,Z)}. \qquad (4.7)\end{aligned}$$

This gives the final AIPW estimating function:

$$\begin{aligned}D_1(t;\beta,\Lambda_0,\pi,S,S_c) =& d\mathcal{N}^{(0)}(t;\pi,S,S_c) - \Gamma^{(0)}(t;\beta,\pi,S,S_c)d\Lambda_0(t), \qquad (4.8)\\ D_2(\beta,\Lambda_0,\pi,S,S_c) =& \int_0^\tau d\mathcal{N}^{(1)}(t;\pi,S,S_c) - \Gamma^{(1)}(t;\beta,\pi,S,S_c)d\Lambda_0(t), \qquad (4.9)\end{aligned}$$

where for $l = 0,1$,

$$\begin{aligned}d\mathcal{N}^{(l)}(t;\pi,S,S_c) =& \frac{A^l dN(t)}{\pi(Z)^A\{1-\pi(Z)\}^{1-A}S_c(t;A,Z)} + \frac{A^l dS(t;A,Z)}{\pi(Z)^A\{1-\pi(Z)\}^{1-A}}\\ & - \sum_{a=0,1} a^l \left\{1 + \frac{A^a(1-A)^{1-a}}{\pi(Z)^a\{1-\pi(Z)\}^{1-a}}J(t;a,S,S_c)\right\} dS(t;a,Z),\\ \Gamma^{(l)}(t;\beta,\pi,S,S_c) =& \frac{A^l Y(t)e^{\beta A}}{\pi(Z)^A\{1-\pi(Z)\}^{1-A}S_c(t;A,Z)} - \frac{A^l S(t;A,Z)e^{\beta A}}{\pi(Z)^A\{1-\pi(Z)\}^{1-A}}\\ & + \sum_{a=0,1} a^l \left\{1 + \frac{A^a(1-A)^{1-a}}{\pi(Z)^a\{1-\pi(Z)\}^{1-a}}J(t;a,S,S_c)\right\} S(t;a,Z)e^{\beta a},\end{aligned}$$

and $J(t;a,z,S,S_c) = \int_0^t dM_c(u;a,z,S_c)/\{S(u;a,z)S_c(u;a,z)\}$.Note that $d\mathcal{N}^{(l)}(t)$ can be seen as augmented weighted $A^l dN(t)$, and $\Gamma^{(l)}(t)$ can be seen as augmented weighted $A^l Y(t)e^{\beta A}$. In this sense the AIPW estimating function (4.8) - (4.9) parallels the original full data estimating function (4.2) - (4.3).

The following theorem gives the doubly robust property of the AIPW estimating function. In the following, the superscript '^{o}' denotes the true value of a parameter.

Theorem 10. *Under Assumptions 6-10, if either $S = S^o$, or both $S_c = S_c^o$ and $\pi = \pi^o$, then for all t, $E\{D_1(t;\beta^o,\Lambda_0^o,\pi,S,S_c)\} = E\{D_2(\beta^o,\Lambda_0^o,\pi,S,S_c)\} = 0$.*

4.5 Estimating Equations and Implementation

Given i.i.d. observations (X_i,Δ_i,A_i,Z_i), $i = 1,...,n$, we need to first estimate the nuisance functions $\widehat{\pi}(z),\widehat{S}(t;a,z)$ and $\widehat{S}_c(t;a,z)$. For example, common choices for the survival functions S include the Cox PH model or the accelerated failure time model. However, due to the non-collapsibility (Martinussen and Vansteelandt, 2013; Tchetgen Tchetgen and Robins, 2012) of the marginal structural Cox model, a conditional model for S would be misspecified. In fact, it's often not possible to come up with a parametric/semi-parametric conditional model for S that is compatible with (4.1). To this end, we need to resort to ML/non-parametric methods that are much more flexible for consistently estimating S. On the other hand, these methods usually converge at a slower than root-n rate, so while the resulting AIPW is still DR-consistent, they are no longer DR-asymptotically normal. The property of being doubly robust consistent and asymptotically normal (CAN) when all nuisance functions are estimated with root-n rate is often referred to as the model DR property. Since we have to work with slower than root-n rate methods, we make use of cross-fitting, which allows the AIPW estimator to achieve the so-called rate DR property (Smucler et al., 2019; Hou et al., 2021), in that the estimator is CAN if all nuisance functions are estimated consistently and that the product error rate between S and (π,S_c) is faster than root-n. This property allows each of the nuisance functions to converge at an arbitrarily slow rate as long as their product error rate is fast enough.

Suppose that the n observations are split into k folds of roughly equal size, with index sets $I_1, I_2, \dots, I_k$. For each fold $m = 1, ..., k$, let

$$\begin{aligned}\mathcal{S}_m^{(l)}(t;\beta,\pi,S,S_c) =& \frac{1}{|I_m|}\sum_{i\in I_m}\Gamma_i^{(l)}(t;\beta,\pi,S,S_c),\\ \bar{A}_m(t;\beta,\pi,S,S_c) =& \frac{\mathcal{S}_m^{(1)}(t;\beta,\pi,S,S_c)}{\mathcal{S}_m^{(0)}(t;\beta,\pi,S,S_c)}.\end{aligned}$$

Solving for $\sum_{i\in I_m} D_{1i}(t;\beta,\Lambda_0,\pi,S,S_c) = 0$, we first obtain

$$\widetilde{\Lambda}_{0,m}(t;\beta,\pi,S,S_c) = \frac{1}{|I_m|}\sum_{i\in I_m}\int_0^t \frac{d\mathcal{N}_i^{(0)}(u;\pi,S,S_c)}{\mathcal{S}_m^{(0)}(u;\beta,\pi,S,S_c)}.$$

We then plug it into $\sum_{i\in I_m} D_{2i}(\beta,\Lambda_0,\pi,S,S_c) = 0$ to obtain

$$U_m(\beta,\pi,S,S_c) = \frac{1}{|I_m|}\sum_{i\in I_m}\int_0^\tau \left\{d\mathcal{N}_i^{(1)}(t;\pi,S,S_c) - \bar{A}_m(t;\beta,\pi,S,S_c)d\mathcal{N}_i^{(0)}(t;\pi,S,S_c)\right\}.$$

The above can be seen as the ‘augmented’ partial likelihood score. Note that it is important to do the above ‘profiling’ of Λ_0 within the fold, in order to preserve the out-of-fold estimation of the nuisance parameters later.

The cross-fitted AIPW estimation algorithm is summarized in Algorithm 5.

Algorithm 5 *k*-fold Cross-fitted AIPW estimation of β

```
Input: A sample of $n$ observations that are split into $k$ folds with index sets $I_1, I_2, \dots, I_k$.
for each fold indexed by $m$ do
    Estimate nuisance functions $\widehat{\pi}^{(-m)}$, $\widehat{S}^{(-m)}$ and $\widehat{S}_c^{(-m)}$) using the out-of-fold sample indexed by $I_{-m} := \{1, \dots, n\} \setminus I_m$.
end for
Output: $\widehat{\beta}$, the solution to
    $\frac{1}{k}\sum_{m=1}^{k} U_m(\beta, \widehat{\pi}^{(-m)}, \widehat{S}^{(-m)}, \widehat{S}_c^{(-m)}) = 0.$
```

4.6 Asymptotic Properties

Let S^*, S_c^* and π^* be some nuisance functions satisfying Assumption 9. Let $O^\dagger$ denote a sample of n i.i.d. data vectors $\{(X_i^\dagger, \Delta_i^\dagger, A_i^\dagger, Z_i^\dagger), i = 1, \ldots, n\}$ used for estimating $\widehat{\pi}$, $\widehat{S}$ and $\widehat{S}_c$. Let (X, Δ, A, Z) be a data vector independent of $O^\dagger$ and drawn from the same distribution as $O^\dagger$. Define

$$\begin{aligned}
\|\widehat{\pi} - \pi^*\|^2 &= E^\dagger\left(E\left[\{\widehat{\pi}(Z) - \pi^*(Z)\}^2\right]\right), \\
\left\|\widehat{S} - S^*\right\|^2 &= E^\dagger\left(E\left[\left\{\sup_{t\in[0,\tau], a\in\{0,1\}}\left|\widehat{S}(t;a,Z) - S^*(t;a,Z)\right|\right\}^2\right]\right), \\
\left\|\widehat{S}_c - S_c^*\right\|^2 &= E^\dagger\left(E\left[\left\{\sup_{t\in[0,\tau], a\in\{0,1\}}\left|\widehat{S}_c(t;a,Z) - S_c^*(t;a,Z)\right|\right\}^2\right]\right),
\end{aligned}$$

where $E^\dagger$ denotes expectation taken with respect to $O^\dagger$, and E denotes expectation taken with respect to O conditional on $O^\dagger$.

Assumption 11 (Uniform Convergence)**.** *There exist $\pi^*(z), S^*(t;a,z)$ and $S_c^*(t;a,z)$ such that $\|\widehat{\pi} - \pi^*\| = o(1)$, $\left\|\widehat{S} - S^*\right\| = o(1)$ and $\left\|\widehat{S}_c - S_c^*\right\| = o(1)$.*

The uniform convergence Assumption 11 assumes that $\widehat{\pi}$, $\widehat{S}$ and $\widehat{S}_c$ converge to some limiting function π^*, S^* and S_c^* that are not necessarily the truth.

Given additional regularity Assumptions 21-23 listed in Supplementary Material C.3.1, the asymptotic properties of the cross-fitted AIPW estimator $\widehat{\beta}$ defined in Algorithm 5 can be summarized in Theorems 11 and 12 below.

Theorem 11. *Under Assumptions 6-11 and Assumptions 21-23, if either $S^* = S^o$, or $(\pi^*, S_c^*) = (\pi^o, S_c^o)$, then $\widehat{\beta} \xrightarrow{p} \beta^o$.*

For a sample of n i.i.d. observations, we define

$$\mathcal{S}^{(l)}(t;\beta,\pi,S,S_c) = \frac{1}{n}\sum_{i=1}^{n}\Gamma_i^{(l)}(t;\beta,\pi,S,S_c),$$

for $l=0,1$, and let $\bar{A}(t;\beta,\pi,S,S_c)=\mathcal{S}^{(1)}(t;\beta,\pi,S,S_c)/\mathcal{S}^{(0)}(t;\beta,\pi,S,S_c)$. This is a full sample version of $\mathcal{S}_m^{(l)}$ that is much more convenient to work with.

Following the notation of Wang et al. (2022), we let O denote a sample of n i.i.d data vectors $\{(X_j,\Delta_j,A_j,Z_j),j=1,\dots,n\}$ that is independent from $O^\dagger$, and drawn from the same distribution as O, and define the cross-integral product

$$\begin{aligned}
&\mathcal{D}^\dagger(\widehat{S},\widehat{S}_c;S^o,S_c^o)\\
=&E^\dagger\left\{E\left[\max_{a\in\{0,1\}}\left|\int_0^\tau\{a-\bar{A}(t;\beta^o,\pi^o,S^o,S_c^o)\}\right.\right.\right.\\
&\left.\left.\left.\times\int_0^t\left\{\frac{d\widehat{S}(t;a,Z)}{\widehat{S}(u;a,Z)}-\frac{dS^o(t;a,Z)}{S^o(u;a,Z)}\right\}\left\{\frac{dM_c(u;a,Z,\widehat{S}_c)}{\widehat{S}_c(u;a,Z)}-\frac{dM_c(u;a,Z,S_c^o)}{S_c^o(u;a,Z)}\right\}\right|\right]\right\}\\
&+E^\dagger\left\{E\left[\max_{a,l\in\{0,1\}}\left|\int_0^\tau\{\bar{A}(t;\beta^o,\pi^o,S^o,\widehat{S}_c)-\bar{A}(t;\beta^o,\pi^o,S^o,S_c^o)\}\right.\right.\right.\\
&\left.\left.\left.\times J(t;a,Z,S^o,\widehat{S}_c)^l\{d\widehat{S}(t;a,Z)-dS^o(t;a,Z)\}\right|\right]\right\},
\end{aligned}\tag{4.10}$$

where with a slight abuse of notation, we use E here to denote the expectation taken with respect to the sample O with n observations conditional on the sample $O^\dagger$.

Assumption 12 (Rate Condition). *$(\pi^*,S^*,S_c^*)=(\pi^o,S^o,S_c^o)$ and*

$$\left\|\widehat{S}-S^o\right\|\left(\|\widehat{\pi}-\pi^o\|+\left\|\widehat{S}_c-S_c^o\right\|\right)+\mathcal{D}^\dagger(\widehat{S},\widehat{S}_c;S^o,S_c^o)=o(n^{-1/2}).$$

Note that on top of the typical product rate condition that are products of error rates (Chernozhukov et al., 2018; Rotnitzky et al., 2021), we also include an integral cross-product term. This integral term is not needed for their previous work because at most one of their nuisance functions involves the time component, which allows them to make use of the mixed bias property (Rotnitzky et al., 2021) and simplifies the proof. Since we have two nuisance functions that involve time t, the mixed bias property no longer holds in our case, so an additional integral term is unavoidable. Similar integral terms in the rate condition can be found in Wang et al. (2022); Vansteelandt et al. (2022). Ying (2023) also provided a more general discussion of these integral remainders.

Theorem 12. *In addition to Assumptions 6-11 and Assumptions 21-23, if Assumption 12 holds, we have*

$$\widehat{\sigma}^{-1}\sqrt{n}(\widehat{\beta}-\beta^{o}) \xrightarrow{d} N(0,1),$$

where the expressions for the asymptotic variance estimator $\widehat{\sigma}^2 := \widehat{\sigma}^2(\widehat{\beta})$ *is provided in Supplementary Material C.1.*

Theorem 12 establishes the rate double robustness properties. Traditionally, DR inference only considers parametric/semi-parametric working models that converge at a root-n rate, which rules out the use of any ML/NP methods. On the other hand, it can be shown that all doubly robust estimating functions are also Neyman orthogonal scores (Neyman, 1959), which when combined with cross-fitting, gives root-n consistent estimators, as long as all nuisance parameters are estimated at $n^{-1/4}$ rate (Newey, 1994; Rotnitzky et al., 2021). This $n^{-1/4}$ rate requirement allows for the use of some ML/NP methods but still excludes many others. The rate double robustness result established here improves further upon these results. The estimator $\widehat{\beta}$ is CAN even if one of $\widehat{S}$ or $(\widehat{\pi},\widehat{S}_c)$ converge arbitrarily slow, as long as their product error rate is faster than root-n. So far, there have been very few results published on the convergence rate of ML methods for time-to-event data, we will empirically investigate the performance of some ML methods in the simulation section below.

We also note that due to the non-collapsibility of the Cox model that we discussed before, if we only use root-n nuisance function estimators, we would get at least one model wrong, which is why we do not consider it here. If someone insists on using root-n estimators and manages to get the other model correct, then the AIPW estimator would still be CAN. However, in this case, the variance estimator would be a lot more complicated due to the residual terms from the misspecified model, so we recommend the use of resampling methods such as bootstrap (Efron, 1992) for the estimation of the variance. In fact, resampling methods are valid under both model-DR and rate-DR, which we will investigate further in the simulation Section below.

4.7 Simulation

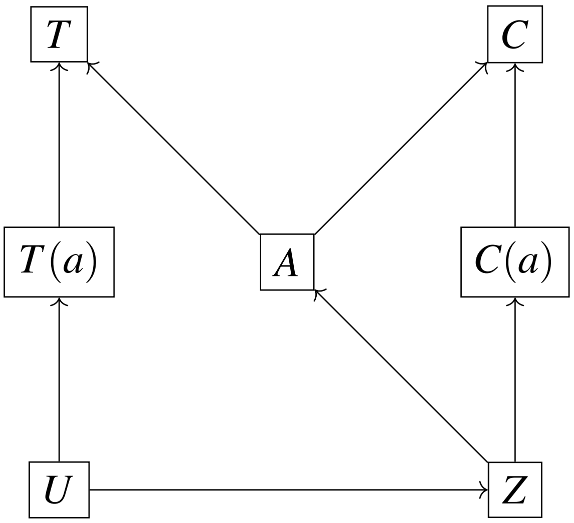


Figure 4.1: DAG for simulation

In this section, we investigate the performance of the AIPW estimator against the full data estimator that uses $(T(1), C(1), T(0), C(0))$, a naive Cox estimator that does not adjust for any covariates, and the IPW estimator. We set $n = 1000, p = 3$, and $\tau = 1$ for each dataset that we simulate, and we repeat this process 1000 times, which corresponds to a margin of errors around $+/-1.35\%$ for the coverage probability of nominal 95% confidence intervals. Five-fold cross-fitting is applied to all the AIPW estimators that we consider.

Data generation process is summarized in figure 4.1. We first simulate 3 i.i.d. latent variables $U_1 \sim \text{Unif}(-1,1), U_2 \sim \text{Unif}(-1,1)$, and $U_3 \sim \text{Unif}(-1,1)$, follwed by $Z_1 = 0.5U_1 + U_3$, $Z_2 = U_1 + 1.5U_1^2 - 0.5$, $Z_3 = U_1 + U_2$ and $T(a) = -\log(0.5U_1 + 0.5)\exp(a)$ for $a = 0, 1$. This allows $T(a)$ to follow the marginal structural Cox model as defined in (4.1) with $\lambda_0^o(t) = 1$ and $\beta^o = -1$. After that, we simulate $C(a)$ and A according to the 4 scenarios stated in Table 4.1, and set $T = AT(1) + (1-A)T(0)$, $C = AC(1) + (1-A)C(0)$.

As described in Table 4.1, we consider two scenarios for the conditional model of censoring time $C(a)$ given Z: Cox PH or a mixture of uniform and Cox PH, and two scenarios for the propensity score model $\pi(z)$: logistic or soft partition. This gives us $2 \times 2 = 4$ scenarios in total. We note that under the marginal structural Cox model (4.1), any parametric/semi-parametric conditional model of T, including the Cox PH model, is invalid due to the non-collapsibility of the Cox model that we discussed earlier. While for each of π and S_c, we consider one simple model that can be consistently estimated parametric/semi-parametrically, and one complex model that requires ML methods. All 4 Scenarios give around 50% being treated and around 40% being censored due to loss to follow-up.

For the proposed AIPW estimator, we investigate 2 methods for estimating the conditional T model: Cox and random survival forest (RSF) (Ishwaran et al., 2008), 2 for the conditional C model: Cox and RSF, as well as 2 for the propensity score model: logistic regression and the 'twang' package in R (Ridgeway et al., 2022), which makes up a total of 8 different AIPW estimators. Note that the conditional censoring model can be fit using the event time X and the event indicator $\Delta_c = (1-\Delta) \cdot I(X < \tau)$. We set the splitrule of RSF to 'bs.gradient' and kept all other hyperparameters of twang and RSF as the default. For each simulated data set, we calculate both a model-based standard error (SE) assuming the conditions for rate DR holds as well as a bootstrapped SE using 100 bootstrap runs for all the AIPW estimators. Bootstrap SE is more computationally expensive, but it does hold under the additional setting when all nuisance function estimators are parametric and one of them is misspecified. Coverage is calculated using normal-based 95% confidence intervals constructed from both the estimated model SEs and the bootstrap SEs.

We also consider 4 different IPW estimators, again with censoring weights estimated using Cox or RSF, and propensity score estimated using logistic regression or 'twang'. Bootstrap SE are calculated for these 4 estimators. As for the naive Cox estimator and the full data estimator, we calculate their SEs using the standard robust SE estimators.

To make sure the strict positivity assumption 9 is satisfied, we trim $\widehat{S}(t;a,z), \widehat{S}_c(t;a,z)$ so that all estimated values below 0.05 are set to 0.05, and we also set all estimated $\widehat{\pi}(z)$ values that are above 0.9 or below 0.1 to be 0.9 and 0.1.

Figures in 4.2 show the bias, standard deviation (SD), and bootstrap-based coverage probability of the 12 AIPW and IPW estimators under Scenarios 1-4 respectively. Tables 4.2 and 4.3 provide additional details for all 14 estimators that are considered. We see that IPW estimators can have small biases when both the censoring working model and the propensity score working model are correctly specified, but their coverage is often quite poor, especially in Scenarios 2-4 when all parametric/semi-parametric working models are invalid. This echoes the challenge of inference for IPW estimators when using ML methods that we discussed before. When all the working models are correct, AIPW estimators show very small biases, along with close to 0.95 coverage, across all 4 scenarios. We also note that the Cox/Cox-logit AIPW estimator, which exclusively uses parametric/semi-parametric nuisance function estimators, only attains good coverage under Scenario 1, which corresponds to the model-DR case. The AIPW estimator that uses a mix of parametric and ML methods, like the RSF/RSF-logit AIPW estimator, can perform well on multiple scenarios, but still fails when the parametric model fails. Lastly, the RSF/RSF-twang AIPW estimator that uses all ML methods performs well under all 4 settings. All these findings support the rate-DR result from Theorem 12.

4.8 Application

Here, we look at data collected from the Honolulu Health Program (HHP) and the subsequent linked Honolulu-Asia Aging Study (HAAS). HHP is an epidemiological study started in 1965, that concerns the rates and risk factors for heart disease and stroke in men of Japanese ancestry living in Oahu and born between 1900 and 1919. In 1991, HAAS is established as a continuation of the HHP study, which concerns instead brain aging,

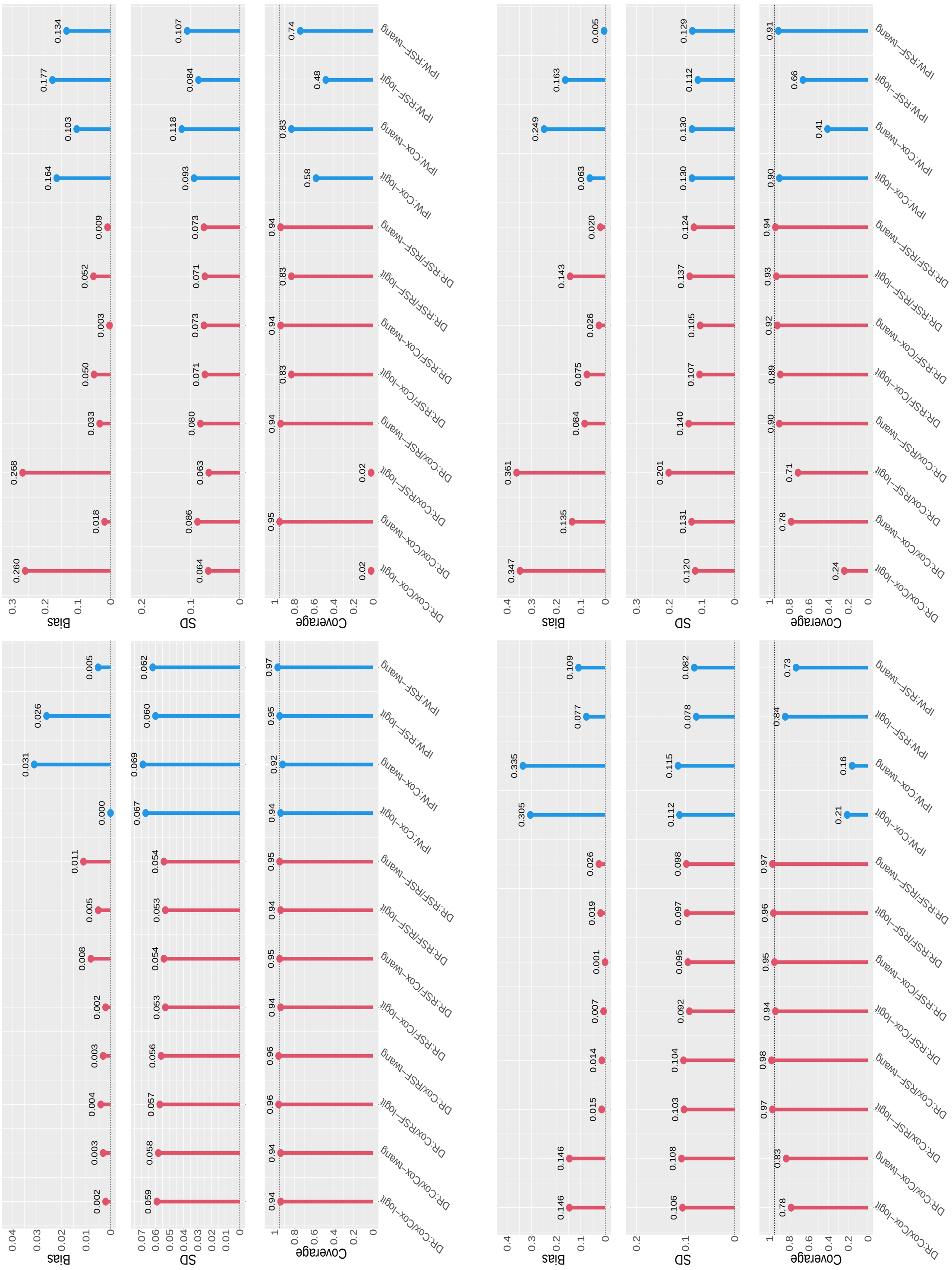


Figure 4.2: Plots of bias, SD, and bootstrap coverage for each of the 4 Scenarios considered in Simulation. Top-left, top-right, bottom-left, and bottom-right in landscape view correspond to Scenario 1 to Scenario 4, respectively.

Table 4.1: Scenarios 1-4 of the simulation. $\beta^o = -1$, $\lambda_0(t) = 1$.

Scenario	Data-generating mechanism
1	$\varepsilon \sim \mathrm{Unif}(0,1)$
Censoring: Cox	$C(a) = -\log(\varepsilon)\exp(0.5 + 0.5a - Z_2 + 0.5Z_3)$
PS: Logistic	$\mathrm{logit}\{\pi(Z)\} = 0.5Z_1 - 0.5Z_2 - 0.5Z_3$
2	$\varepsilon \sim \mathrm{Unif}(0,1)$
Censoring: Cox	$C(a) = -\log(\varepsilon)\exp(0.5 + 0.5a - Z_2 + 0.5Z_3)$
PS: Soft Partition	$\mathrm{logit}\{\pi(Z)\} = -3 \cdot \mathbf{1}\{Z_2 < -0.5\} + 3 \cdot \mathbf{1}\{-0.5 \leq Z_2 < 0.5\} - 3 \cdot \mathbf{1}\{Z_2 \geq 0.5\}$
3	$\varepsilon_a \sim \mathrm{Unif}(0,1)$ for $a = 0,1$
Censoring:	$C(0) = 1.05\varepsilon_0$
Uniform-Cox	$C(1) = -\log(\varepsilon_1)\exp(3.3 + 3.5Z_3)$
PS: Logistic	$\mathrm{logit}\{\pi(Z)\} = 0.5Z_1 - 0.5Z_2 - 0.5Z_3$
4	$\varepsilon_a \sim \mathrm{Unif}(0,1)$ for $a = 0,1$
Censoring:	$C(0) = 1.05\varepsilon_0$
Uniform-Cox	$C(1) = -\log(\varepsilon_1)\exp(3.3 + 3.5Z_3)$
PS: Soft Partition	$\mathrm{logit}\{\pi(Z)\} = -3 \cdot \mathbf{1}\{Z_2 < -0.5\} + 3 \cdot \mathbf{1}\{-0.5 \leq Z_2 < 0.5\} - 3 \cdot \mathbf{1}\{Z_2 \geq 0.5\}$

Alzheimer's disease, vascular dementia, other causes of cognitive and motor impairment, stroke, and the common chronic conditions of late life. In this section, we are particularly interested in the effect of mid-life alcohol exposures (captured in 1965-1973) on late-life overall survival (starting in 1991). Alcohol consumption was assessed through self-report and translated into units per month. Participants are categorized as heavy drinkers if they were heavy drinkers at some point during mid-life and non-heavy drinkers otherwise. After removing a few missing observations, we have a total of 2079 patients with 552 of them being heavy drinkers and 1527 non-heavy drinkers. Time to death is our primary endpoint, and in addition to mid-life alcohol exposure, we also include 4 covariates that were assessed at the start of the HHP: age, maximum years of education, systolic blood pressure, and heart rate in 30 seconds. It's worth noting that death certificates are available for many participants which allows us to retrospectively identify the death date, but we will not include it in our analysis so as to better understand the effect of informative censoring. We set a maximum follow-up time of 13 years, after which the participants' outcomes are artificially censored. This guarantees that the strict positivity assumption 9 is satisfied. The boxplots of $\widehat{\pi}(z), \widehat{S}(\tau; a, z)$ and $\widehat{S}_c(\tau; a, z)$ for all 2079 patients are provided in Figure C.1 of the supplementary material, which verifies that the strict positivity assumption is satisfied with

Table 4.2: Simulation based on 1000 data sets. $(n, p) = (1000, 3)$. $\beta^o = -1$. Red indicates that the model or approach is invalid.

Scenario	Estimator	T/C-PS Models	Bias	SD	SE	Coverage
					Model/Boot	Model/Boot
Scenario 1	AIPW	Cox/Cox-logit	0.002	0.059	0.060/0.059	0.95/0.94
		Cox/Cox-twang	0.003	0.058	0.061/0.060	0.96/0.94
		Cox/RSF-logit	0.004	0.057	0.060/0.059	0.96/0.96
		Cox/RSF-twang	0.003	0.056	0.062/0.060	0.97/0.96
		RSF/Cox-logit	0.002	0.053	0.053/0.053	0.95/0.94
		RSF/Cox-twang	0.008	0.054	0.054/0.055	0.95/0.95
		RSF/RSF-logit	0.005	0.053	0.053/0.054	0.95/0.94
		RSF/RSF-twang	0.011	0.054	0.055/0.056	0.95/0.95
	IPW	Cox-logit	0.000	0.067	- /0.066	- /0.94
		Cox-twang	0.031	0.069	- /0.069	- /0.92
		RSF-logit	0.026	0.060	- /0.064	- /0.95
		RSF-twang	0.005	0.062	- /0.068	- /0.97
	Naive Cox		0.496	0.100	0.100/0.101	0.00/0.00
	Full Data		0.001	0.029	0.028/ -	0.95/ -
Scenario 2	AIPW	Cox/Cox-logit	0.260	0.064	0.068/0.066	0.02/0.02
		Cox/Cox-twang	0.018	0.086	0.093/0.088	0.96/0.95
		Cox/RSF-logit	0.268	0.063	0.069/0.066	0.02/0.02
		Cox/RSF-twang	0.033	0.080	0.090/0.084	0.95/0.94
		RSF/Cox-logit	0.050	0.071	0.056/0.061	0.80/0.83
		RSF/Cox-twang	0.003	0.073	0.073/0.075	0.94/0.94
		RSF/RSF-logit	0.052	0.071	0.056/0.061	0.80/0.83
		RSF/RSF-twang	0.009	0.073	0.073/0.075	0.95/0.94
	IPW	Cox-logit	0.164	0.093	- /0.092	- /0.58
		Cox-twang	0.103	0.118	- /0.106	- /0.83
		RSF-logit	0.177	0.084	- /0.088	- /0.48
		RSF-twang	0.134	0.107	- /0.101	- /0.74
	Naive Cox		0.579	0.119	0.125/0.125	0.00/0.00
	Full Data		0.001	0.029	0.028/ -	0.95/ -

Table 4.3: Simulation based on 1000 data sets. $(n,p) = (1000,3)$. $\beta^o = -1$. Red indicates that the model or approach is invalid.

Scenario	Estimator	T/C-PS Models	Bias	SD	SE	Coverage
					Model/Boot	Model/Boot
Scenario 3	AIPW	Cox/Cox-logit	0.146	0.106	0.108/0.109	0.79/0.78
		Cox/Cox-twang	0.146	0.108	0.114/0.120	0.81/0.83
		Cox/RSF-logit	0.015	0.103	0.122/0.118	0.98/0.97
		Cox/RSF-twang	0.014	0.104	0.126/0.125	0.98/0.98
		RSF/Cox-logit	0.007	0.092	0.098/0.097	0.95/0.94
		RSF/Cox-twang	0.001	0.095	0.101/0.104	0.95/0.95
		RSF/RSF-logit	0.019	0.097	0.102/0.104	0.96/0.95
		RSF/RSF-twang	0.026	0.098	0.103/0.111	0.96/0.97
	IPW	Cox-logit	0.305	0.112	- /0.103	- /0.21
		Cox-twang	0.335	0.115	- /0.104	- /0.16
		RSF-logit	0.077	0.078	- /0.079	- /0.84
		RSF-twang	0.109	0.082	- /0.082	- /0.73
	Naive Cox		0.895	0.108	0.110/0.110	0.00/0.00
	Full Data		0.001	0.029	0.028/ -	0.95/ -
Scenario 4	AIPW	Cox/Cox-logit	0.347	0.120	0.138/0.143	0.14/0.24
		Cox/Cox-twang	0.135	0.131	0.116/0.129	0.72/0.78
		Cox/RSF-logit	0.361	0.201	0.343/0.224	0.50/0.71
		Cox/RSF-twang	0.084	0.140	0.149/0.157	0.88/0.90
		RSF/Cox-logit	0.075	0.107	0.101/0.105	0.89/0.89
		RSF/Cox-twang	0.026	0.105	0.096/0.106	0.90/0.92
		RSF/RSF-logit	0.143	0.137	0.174/0.155	0.88/0.93
		RSF/RSF-twang	0.020	0.124	0.119/0.132	0.94/0.94
	IPW	Cox-logit	0.063	0.130	- /0.124	- /0.90
		Cox-twang	0.249	0.130	- /0.117	- /0.41
		RSF-logit	0.163	0.112	- /0.104	- /0.66
		RSF-twang	0.005	0.129	- /0.112	- /0.91
	Naive Cox		0.562	0.120	0.132/0.132	0.00/0.00
	Full Data		0.001	0.028	0.028/ -	0.95/ -

our choice of τ. Overall, 47% of the participants died and 21% of the patients are censored due to loss to follow-up, and the remaining 32% of the participants are administratively censored at the 13-year mark.

We focus on estimating the log hazard ratio under the marginal structural Cox model (4.1). To determine whether there are informative censoring and confounding due to the 4 covariates, we look at the univariate association between each of the 4 covariates and the outcome T, the censoring C and the exposure A in Figure 4.4. From the p-values, we see that these 4 covariates lead to both informative censoring (covariate associated with both T and C) and confounding (covariate associated with both T and A). This suggests the need of considering exchangeability Assumption 8, the informative censoring Assumption 10 and the use of the proposed AIPW estimators.

We investigated here all 13 estimators that were considered in the simulation. Figure 4.3 shows the forest plot that provides a point estimate and a 95% bootstrap-based confidence interval for all 13 estimates and Table 4.5 provides additional details. We can see that the AIPW estimators, especially those that use ML methods, give visibly smaller estimates of log hazard ratio compared to the IPW estimates. The IPW estimates themselves are also smaller than the Naive Cox estimate. This could suggest that while all estimators agree on the negative impact of mid-life alcohol exposure on overall survival, the extent of this impact might not be as large as what the Naive Cox estimate suggests once we take better account of the biases resulting from both informative censoring and the treatment confounding.

To estimate the survival curves $P(T(a) > t) = \exp\{-\widehat{\Lambda}_0(t)\exp(\widehat{\beta}a)\}$, we construct a cross-fitted AIPW estimator for the cumulative baseline hazard function $\Lambda_0(t)$ at each time t as

$$\widehat{\Lambda}_0(t) = \frac{1}{k}\sum_{m=1}^{k} \widetilde{\Lambda}_{0,m}(t; \widehat{\beta}, \widehat{\pi}^{(-m)}, \widehat{S}^{(-m)}, \widehat{S}_c^{(-m)}).$$

$\widehat{\Lambda}_0(t)$ estimate is also doubly robust since each of the m-th fold estimates

$\widetilde{\Lambda}_{0,m}(t;\widehat{\beta},\widehat{\pi}^{(-m)},\widehat{S}^{(-m)},\widehat{S}_c^{(-m)})$ is DR. By estimating $\widehat{\Lambda}_0(t)$ and $\widehat{\beta}$ using the AIPW:RSF/RSF-twang method, we obtain average survival curves for the two treatment groups in Figure 4.4. Moreover, this allows us to estimate the risk difference $P(T(1) \leq t) - P(T(0) \leq t)$ and the risk ratio $P(T(1) \leq t)/P(T(0) \leq t)$. Table 4.6 presents the estimated risk difference, the estimated risk ratio, and their bootstrapped 95% normal-based confidence intervals for $t = 3, \ldots, 12$ years.

With informative censoring and treatment confounding, most classical methods of testing proportional hazards fail. Motivated by (Therneau and Grambsch, 2000), we present In Figure 4.5 a smoothed AIPW $\beta(t)$ plot using the AIPW:RSF/RSF-twang estimator. It seems to suggest that the effect of mid-life alcohol consumption on late-life mortality slightly decreases overtime. More details on this plot will be discussed in Section 5.7.2 of Chapter 5.

Table 4.4: The p-values for the univariate association of each of the 4 covariates in HAAS data with the outcome T, the censoring time C, and the exposure A. For T and C, the p-values are calculated using a univariate Cox PH model. For A, the p-value is from a univariate logistic regression.

Covariate	Association with T	Association with C	Association with A
Age	0.000	0.191	0.526
Education	0.001	0.001	0.000
SystolicBP	0.009	0.668	0.107
HeartRate	0.000	0.062	0.013

4.9 Discussion

We considered the estimation of the log hazard ratio under the marginal structural Cox model and developed an AIPW estimator under exchangeability and informative censoring. The proposed estimating functions are doubly robust with respect to possible misspecification of either the conditional outcome model or the missing data model. In particular, the AIPW estimator is rate DR, allowing for ML/NP methods for estimating nuisance functions, thereby meeting the challenge of the non-collapsibility that comes with

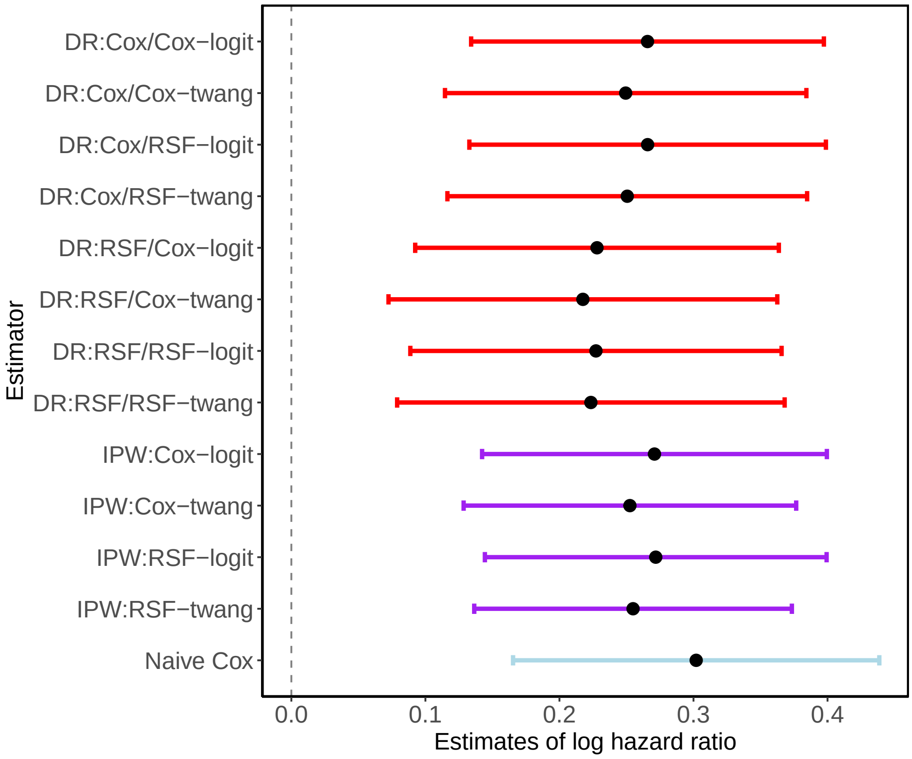


Figure 4.3: Forest plot of the log hazard ratio estimates examining the effect of drinking on overall survival for the HAAS dataset.

Table 4.5: Estimates of the log hazard ratio using 13 different estimators for the HAAS dataset, together with the bootstrapped standard errors and the 95% confidence interval constructed from it.

Estimator	Estimate	Boot SE	95% Boot CI
DR:Cox/Cox-logit	0.27	0.07	(0.13, 0.40)
DR:Cox/Cox-twang	0.25	0.07	(0.11, 0.38)
DR:Cox/RSF-logit	0.27	0.07	(0.13, 0.40)
DR:Cox/RSF-twang	0.25	0.07	(0.12, 0.39)
DR:RSF/Cox-logit	0.23	0.07	(0.09, 0.36)
DR:RSF/Cox-twang	0.22	0.07	(0.07, 0.36)
DR:RSF/RSF-logit	0.23	0.07	(0.09, 0.37)
DR:RSF/RSF-twang	0.22	0.07	(0.08, 0.37)
IPW:Cox-logit	0.27	0.07	(0.14, 0.40)
IPW:Cox-twang	0.25	0.06	(0.13, 0.38)
IPW:RSF-logit	0.27	0.07	(0.14, 0.40)
IPW:RSF-twang	0.25	0.06	(0.14, 0.37)
Naive Cox	0.30	0.07	(0.17, 0.44)

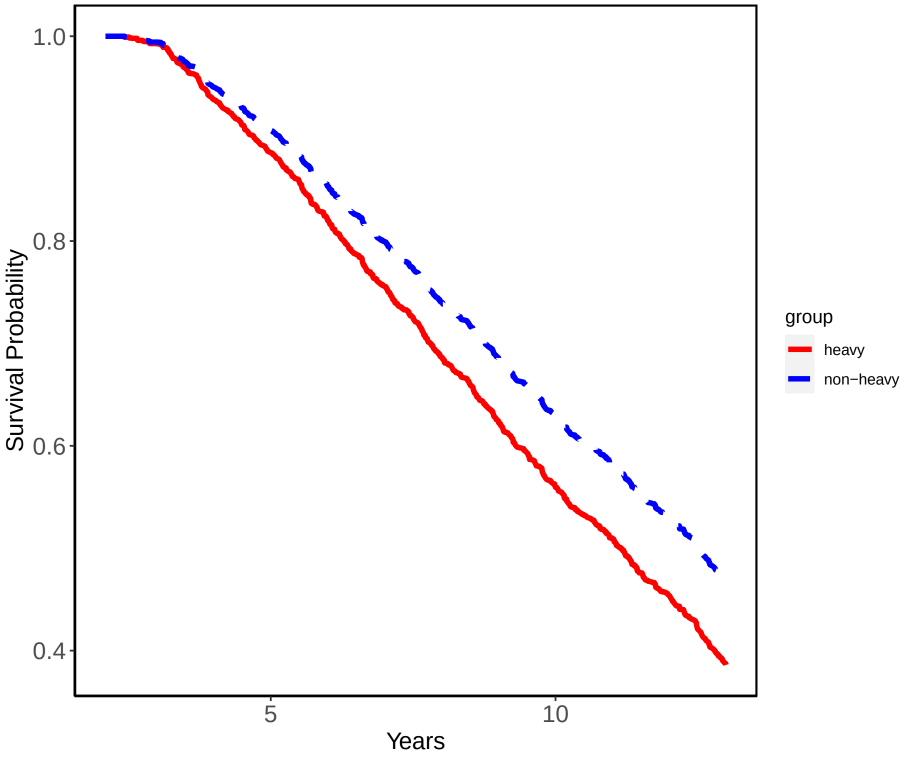


Figure 4.4: Estimated survival curves for heavy-drinkers vs non-heavy drinkers based on the cross-fitted AIPW estimator $\widehat{\Lambda}_0(t)$ for the cumulative baseline hazard.

Table 4.6: Estimated risk difference and risk ratio between the mid-life heavy-drinkers and the non-heavy drinkers along with the bootstrapped 95% confidence intervals.

Year	Risk Difference	Risk Ratio
3	0.001 (-0.0001,0.0030)	1.249 (1.0699,1.4291)
4	0.012 (0.0036,0.0207)	1.242 (1.0685,1.4164)
5	0.022 (0.0070,0.0364)	1.236 (1.0670,1.4043)
6	0.033 (0.0109,0.0554)	1.227 (1.0653,1.3879)
7	0.044 (0.0145,0.0727)	1.217 (1.0634,1.3711)
8	0.054 (0.0183,0.0892)	1.207 (1.0610,1.3523)
9	0.062 (0.0210,0.1030)	1.196 (1.0592,1.3337)
10	0.069 (0.0236,0.1144)	1.186 (1.0568,1.3146)
11	0.074 (0.0253,0.1221)	1.177 (1.0548,1.2984)
12	0.078 (0.0270,0.1286)	1.166 (1.0526,1.2795)

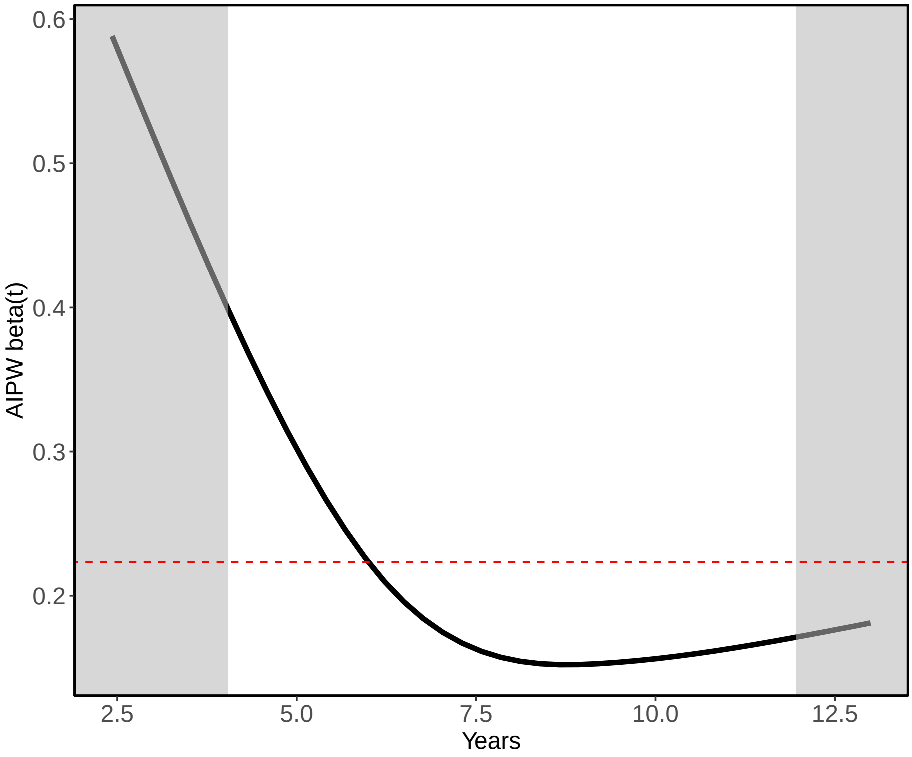


Figure 4.5: Smoothed AIPW $\beta(t)$ plot of the time-varying log hazard ratio for the effect of mid-life alcohol exposure on late-life mortality based on the AIPW:RSF/RSF-twang estimator. The red dotted line indicates the AIPW:RSF/RSF-twang estimate for β^*.

the marginal Cox model.

The marginal structural Cox model is a natural extension of the classical Cox model to causal inference. However, the use of hazard ratio under causal inference has brought some criticisms (Hernán, 2010; Martinussen et al., 2020) mainly due to the imbalance between risk sets post-treatment, while others defended its use (Prentice and Aragaki, 2022; Ying and Xu, 2023). Specifically, they argue that if there is a non-zero treatment effect, the two groups are no longer comparable at time $t > 0$ due to differential survival distributions between the two groups. On the other hand, the causal hazard ratio is the ratio of the log of potential survival functions between two groups at all times. Being a contrast of functionals of the two potential outcome distributions makes it a valid estimand that remains causally interpretable. Others have studied alternative estimands for time-to-event endpoints. Axelrod and Nevo (2022) looked at sensitivity analysis for the so-called 'causal HR' (Martinussen et al., 2020) that is based on patient groups who would have survived regardless of their treatment assignment, but this estimand is not even identifiable, which severely limits its practical use. Vansteelandt et al. (2022) proposed a model-free hazard ratio estimand which would simplify to the hazard ratio under the Cox model. Its estimation does not rely on the inverse probability of treatment weighting, which could bring more stability. However, their estimand explicitly depends on the specific propensity score model involved, and they require the restrictive assumption that the cumulative hazard is positive at all times. While researchers persist in their search for alternative causal estimands under time-to-event studies, we believe the significance of the causal hazard ratio should not be overlooked, owing to its inherent simplicity, interpretability, and widespread popularity.

In this paper, we implicitly assumed that the three nuisance functions depend on the same set of covariates. This need not be the case, and they may depend on different covariates, so long as the exchangeability and informative censoring assumptions are satisfied for the covariates that we choose to work with.

The proposed AIPW estimator was implemented in R and has been built into the R package 'CoxAIPW', and is available on the comprehensive R archive network (CRAN).

4.10 Acknowledgement

Chapter 4 contains material being prepared for submission and is co-authored with Denise Rava, Jelena Bradic, and Ronghui Xu. The dissertation author is the primary investigator/author of this paper.

We thank Ms.Yuyao Wang for suggesting a specific cross-fitting procedure where we first profile out within each fold before combining the estimating equations from each fold.

Chapter 5

Doubly Robust Inference under Non-Proportional Hazards Model

5.1 abstract

The marginal structural Cox model is commonly used in observational studies with time-to-event outcomes to draw causal inferences, but the proportional hazards (PH) assumption it makes often fails in practice. We propose an alternative causal estimand with clear causal interpretation under the non-PH model. It is a weighted average of the time-varying causal log hazard ratio, which recovers the time-fixed causal log hazard ratio under the marginal structural Cox model, and has a simple connection with treatment effects defined through transformation models. In observational data with time-to-event outcomes, both confounding and informative censoring occur frequently. They lead to potential biases in our estimation of the proposed causal estimand and can be accounted for through inverse probability weighting (IPW) to the full data estimating functions. By demonstrating the equivalence between its full data estimating functions and that of the marginal structural Cox model, we further construct an augmented IPW (AIPW) estimator that protects against misspecification of the working models through the doubly robust property. With cross-fitting, this AIPW estimator is not only model doubly-robust (DR)

but also rate DR, which allows for the use of machine learning methods in estimating the nuisance functions. Classical methods for plotting the time-varying log hazard ratio fail in the presence of informative censoring or treatment confounding, we propose an AIPW log hazard ratio plot that also works under observational data. Extensive simulations are conducted to study the properties of the proposed estimation methods empirically. Lastly, we apply our proposed method to the International Non-Hodgkin's Lymphoma Prognostic Factors Project dataset.

5.2 Introduction

In observational studies with time-to-event data, the marginal structural Cox model (Hernán et al., 2001) is one of the most commonly used models for drawing causal inference. When the proportional hazards (PH) assumption is satisfied, the causal hazard ratio is a scalar causal estimand for the treatment effect of a binary exposure and has been widely applied (Cole et al., 2003; Feldman et al., 2004; Sterne et al., 2005; Hernán et al., 2006; Buchanan et al., 2014). However, the PH assumption, which assumes that the ratio of hazards function between the two groups is constant at all times, is rarely satisfied in practice. This prompts us to consider the more general marginal structural non-PH model with a time-varying causal hazard ratio.

Under the marginal structural non-PH model, there have been efforts on developing a scalar causal estimand that summarizes the time-varying hazard ratio (Martinussen et al., 2020; Vansteelandt et al., 2022), but they all have their respective limitations, which we discuss in details later. Motivated by the work of Struthers and Kalbfleisch (1986); Xu and O'Quigley (2000), we propose a causal estimand β^* under the marginal structural non-PH model with the desirable property of being a weighted average of the time-varying log hazard ratio over time. Moreover, this causal estimand has a simple algebraic relationship with the treatment effects defined through the marginal structural transformation model. Under the marginal structural proportional odds model with log-link, which is a special case of the marginal structural transformation model, β^* reduces to exactly the average of time-varying log hazard ratio over time. Lastly, due to the desirable property of being a weighted average of time-varying log hazard ratio, it would recover the time-fixed causal log hazard ratio if the PH assumption is satisfied. This makes β^* a generalization of the causal log hazard ratio that carries causal interpretation even when the PH assumption fails.

In observational data that is without randomization, confounding exists that introduces bias. Assuming no unmeasured confounding, inverse probability of treatment weighting (IPTW) (Robins, 1998; Robins et al., 2000a; Robins, 2000; Lunceford and Davidian, 2004; Hubbard et al., 2000; Hernán et al., 2001; Chen and Tsiatis, 2001; Zhang and Schaubel, 2011) can be applied for unbiased estimation of our causal estimand. Time-to-event data is also commonly associated with right-censoring. If the censoring is informative, that is, the outcome event time and the censoring time are dependent unless we condition on additional covariates, then the inverse probability of censoring weighting (Robins, 1993; Robins and Finkelstein, 2000) can be similarly applied to correct for this bias.

Both the counterfactual outcomes and the censoring mechanisms can be seen as missing data. IPTW and IPCW require a known or estimated propensity score and censoring time model, respectively, and methods for protection against these models were developed. For counterfactual outcomes, augmented inverse probability weighting (AIPW) can be applied (Robins et al., 1995; Scharfstein et al., 1999; Robins et al., 2000a; Robins, 2000; Robins et al., 2000b; Robins and Rotnitzky, 2001; Van der Laan and Robins, 2003; Bang and Robins, 2005; Tsiatis, 2006), which protects against misspecification of the propensity score model. Rotnitzky and Robins (2005) developed the augmented IPCW approach that protects against misspecification of the censoring model.

In this paper, we first identify the full data estimating functions for our proposed causal estimand. By showing its equivalence with the full data estimating functions for the marginal structural Cox model proposed in Chapter 4, we proceed to develop the IPW and AIPW estimating functions in the same way. The AIPW estimating functions lead to the

AIPW estimator involving 3 working models, a conditional outcome model, a propensity score model, and a censoring time model, the latter two can be seen as the missing data model. This AIPW estimator is model DR (Rotnitzky et al., 2021) in that it is consistent and asymptotically normal (CAN) if any of the conditional outcome model or the missing data model is correctly specified. With the help of cross-fitting (Chernozhukov et al., 2018), the resulting cross-fitted AIPW estimator is also rate DR, which is CAN when the product of the estimation error rates under the two sets of working models is faster than root-*n*, even if any of them is arbitrarily slow.

We will also show that the proposed causal estimand is the limit of the partial likelihood (PL) estimator under full data with both potential outcomes and no censoring. Struthers and Kalbfleisch (1986) studied the limit of the PL estimator under randomization, Xu and O'Quigley (2000); Boyd et al. (2012); Hattori and Henmi (2012) then looked at the estimation of this limit under non-informative censoring, which assumes that the outcome and the censoring time are conditionally independent given treatment assignment. Our causal estimand can be thought of as an extension of this limit to observational studies and informative censoring, this motivates us to investigate the bias of the PL estimator when there is either informative censoring or treatment confounding.

Under randomization and random censoring, a plot of the time-varying log hazard ratio can be constructed following the approach of Therneau and Grambsch (2000). This is however biased under observational data without randomization or under informative censoring. Following our approach for constructing the AIPW estimator, we similarly propose an AIPW log hazard ratio plot. We then test it on a simulated dataset and compare it with the truth and the plot obtained using Therneau and Grambsch (2000)'s approach.

Causal inference has received wide interest due to the increasing scale and availability of observational data. However, we often remain interested in the associational or predictive effect of exposure on a time-to-event outcome through the likes of regression models. This leads us to study a non-causal estimand, which is equivalent to the average regression effect proposed in Struthers and Kalbfleisch (1986); Xu and O'Quigley (2000). We show that under randomization but with informative censoring, we can estimate the non-causal estimand using precisely the method we proposed in Chapter 3. We also discuss the properties of this estimand and its AIPCW estimator.

The rest of the paper is organized as follows. In Section 5.3, we state the non-PH model, the assumptions, and introduce a new causal estimand. In Section 5.4, we identify the full data estimating functions and construct the IPW and AIPW estimating functions. Implementation of the AIPW estimating equations is described in Section 5.5. Section 5.6 discusses the asymptotic properties of the non-cross-fitted and the cross-fitted AIPW estimator, including the model DR and rate DR properties. Some miscellaneous results, including the bias of the PL estimator, the AIPW plot for the time-varying log hazard ratio, as well as the non-causal estimand, are summarized in Section 5.7. In Section 5.8, we conduct simulations for our AIPW estimators under various non-PH models as well as various models for censoring and propensity scores. Lastly, the proposed estimator is applied to the Non-Hodgkin's Lymphoma dataset in Section 5.9. Additional materials are provided in the Supplementary Materials.

5.3 Non-Proportional Hazards Model and Estimand

5.3.1 Randomization

To motivate our estimand, we first consider a study under randomization. In a two-group study with time-to-event outcomes and without censoring, we observe the full data vector (T, A) where T is the failure time, and $A \in \{0,1\}$ is the exposure of interest. A saturated model here is non-PH, which can be expressed as

$$\lambda(t;A) = \lambda_0(t)\exp\{\beta(t)A\}, \tag{5.1}$$

where $\lambda(t;A)$ is the hazard function of T given exposure group A, $\lambda_0(t)$ is the baseline hazard function, and $\beta(t)$ is the time-varying log hazard ratio.

Under the non-PH model (5.1), challenges arise regarding both the inference as well as the interpretation of $\beta(t)$ due to its infinite-dimensionality. This motivates us to look for a scalar estimand that can be thought of as a weighted average of $\beta(t)$ of the form

$$\frac{\int_0^\tau \beta(t)w(t)dt}{\int_0^\tau w(t)dt},$$

where τ is the maximum follow-up time. Let $F(t)$ denote the cumulative density function of T. One choice of $w(t)$ with useful interpretations (Struthers and Kalbfleisch, 1986; Xu and O'Quigley, 2000) arises from the estimating equation

$$\int_0^\tau \left\{E_{\beta(t)}(A|T=t) - E_\beta(A|T=t)\right\} dF(t) = 0, \tag{5.2}$$

Here, $E_{\beta(t)}$ refers to the expectation taken over the time-varying model 5.1, while E_β sets $\beta(t)$ to be a constant β. We see that the solution β is the fixed log hazard ratio that would balance $E(A)$ under the non-PH model 5.1. This definition also provides us with other interesting interpretations which we will discuss later.

5.3.2 Marginal Structural Model

In practice, we are often interested in drawing causal inferences from observational data which involves confounding. To define a causal estimand, we adopt the counterfactual framework (Neyman, 1923; Rubin, 1974) and consider the marginal structural non-PH model, where with a slight abuse of notation, we let

$$\lambda_{T(a)}(t) = \lambda_0(t)e^{\beta(t)a}. \tag{5.3}$$

Here, $T(a)$ is the potential survival time under treatment $A = a$, $\lambda_{T(a)}(t)$ is the hazard function for $T(a)$, and $\lambda_0(t)$ is the baseline hazard function.

Like most studies of time-to-event outcomes, there is possible right censoring on the potential failure times $T(1), T(0)$. Denote $C(1), C(0)$ the potential censoring time for $T(1), T(0)$, respectively. Let $\Delta(a) = I\{T(a) \leq C(a)\}$ denote the potential event indicator and let $X(a) = \min\{T(a), C(a)\}$. We will use T, C, X, Δ to indicate the failure, censoring, censored times, and event indicator, respectively, once the treatment is received. In addition, we consider observed baseline covariates vector $Z \in R^p$.

Next, we make the following assumptions that are standard in causal inference (Hernán and Robins, 2020).

Assumption 13 (SUTVA). *The potential outcomes of one subject are not affected by the treatment assignment of the other subjects, and there are no hidden versions of the treatments.*

Assumption 14 (Consistency). $T = AT(1) + (1-A)T(0)$, *and* $C = AC(1) + (1-A)C(0)$.

Assumption 15 (Exchangeability). $(T(a), C(a)) \perp A \mid Z$, *for* $a = 0, 1$.

Assumption 16 (Strict Positivity). *There exists* $0 < \varepsilon < 1$ *such that* $\varepsilon < P(A = 1|Z = z) < 1 - \varepsilon$, $P(C > \tau|A = a, Z = z) > \varepsilon$, $P(T > \tau|A = a, Z = z) > \varepsilon$ *for all values of a and z, where* τ *is a maximum follow-up time.*

Non-informative censoring (Kalbfleisch and Prentice, 2011) is often assumed in traditional survival analysis, which requires the censoring time to be conditionally independent of the failure time given the regressors. Our model under consideration is the marginal structural non-PH model, which does not involve the baseline covariates Z. Nonetheless, we make the following informative censoring assumption that allows the censoring time to depend on the covariates.

Assumption 17 (Informative Censoring). $T(a) \perp C(a) \mid Z$, *for* $a = 0, 1$, *where* $\perp$ *indicates statistical independence.*

We now propose a new causal estimand β^* as the solution to (5.2) under 1:1 randomization. Under the marginal structural non-PH model (5.3), it can be equivalently defined using counterfactual quantities in the following Lemma.

Lemma 1. β^* *is the solution to*

$$h(\beta) = \int_0^\tau \left\{ \frac{\sum_{a=0,1} a e^{\beta(t)a} S_a(t)}{\sum_{a=0,1} e^{\beta(t)a} S_a(t)} - \frac{\sum_{a=0,1} a e^{\beta a} S_a(t)}{\sum_{a=0,1} e^{\beta a} S_a(t)} \right\} \sum_{a=0,1} dF_a(t) = 0, \qquad (5.4)$$

where $S_a(t) = P(T(a) > T)$ is the survival function for $T(a)$, $f_a(t), F_a(t)$ are the probability density function and the cumulative density function for $T(a)$ respectively.

Denote

$$v(\beta, t) = \frac{d}{d\beta} \left[\frac{\sum_{a=0,1} a e^{\beta a} S_a(t)}{\sum_{a=0,1} e^{\beta a} S_a(t)} \right] = \frac{\{\sum_{a=0,1} (1-a) e^{\beta a} S_a(t)\}\{\sum_{a=0,1} a e^{\beta a} S_a(t)\}}{\{\sum_{a=0,1} e^{\beta a} S_a(t)\}^2}.$$

Assume $f_a(t) > 0$ for $t \in [0, \tau], a = 0, 1$. We then have $v(\beta(t), t) > 0$ and $-dh(\beta)/d\beta = \sum_{a=0,1} \int_0^\tau v(\beta, t) dF_a(t) > 0$, so β^* is uniquely defined.

Using the mean-value theorem on the integrand of (5.4), we also have $\sum_{a=0,1} \int_0^\tau v(\widetilde{\beta}(t), t)\{\beta(t) - \beta^*\} dF_a(t) = 0$, where $\widetilde{\beta}(t) = \beta^* + s\{\beta(t) - \beta^*\}$ for some $s \in [0, 1]$. β^* is therefore a weighted average of $\beta(t)$ with weight $w(t) = v(\widetilde{\beta}(t), t) \sum_{a=0,1} f_a(t)$.

Lemma 2. *Suppose $T(a)$ follows the marginal structural transformation model*

$$g(T(a)) = \gamma a + \varepsilon,$$

for $a = 0, 1$, where $g(t)$ is any strictly increasing function. ε is a member of the G^ρ family ($\rho \geq 0$ is known) from Harrington and Fleming (1982), with survival function

$$P(\varepsilon > t) = \exp(-e^t) \qquad (\rho = 0),$$

$$P(\varepsilon > t) = (1 + \rho e^t)^{-1/\rho} \qquad (\rho > 0).$$

If τ is large so that $P(T(a) < \tau) = 1$ for $a = 0,1$, then

$$\beta^* = -\frac{\gamma}{\rho + 1}.$$

As a special case, if $T(a)$ follows the marginal structural proportional odds model $(\rho = 1)$ *with $g(t) = \log(t)$, we also have $\beta^* = \frac{1}{2}\sum_{a=0,1}\int_0^\tau \beta(t)dF_a(t)$.*

This Lemma generalizes the result in Section 2 of Xu and Harrington (2001) to the potential outcomes. This simple algebraic relationship above provides a simple estimator of the contrast γ in the marginal structural transformation model. Moreover, β^* is precisely the average of $\beta(t)$ over time between $a = 1$ and $a = 0$ when $T(a)$ follows the marginal structural proportional odds model with $g(t) = \log(t)$. Note that if $P(T(a) > \tau) > 0$, then the relationship may not be exact.

Lemma 3. *If $T(a)$ follows the marginal structural Cox model (Hernán et al., 2001)*

$$\lambda_{T(a)}(t) = \lambda_0(t)e^{\beta^o a},$$

for $a = 0,1$, then β^* equals the causal log hazard ratio β^o. In general, β^* is also the limit of the oracle estimator for β^o under the non-PH model (5.3).

The marginal structural Cox model is equivalent to setting $\beta(t)$ in the time-varying model (5.3) to β^o, so it is obvious that β^* would exactly recover the causal log hazard ratio under the marginal structural Cox model. The oracle log hazard ratio estimator for β^o is the PL estimator using the full data with both counterfactual outcomes, which is equivalent to the PL estimator under 1:1 randomization. The second statement of Lemma 3 is therefore a direct consequence of Struthers and Kalbfleisch (1986) since their result shows that β^* is the limit of the PL estimator under 1:1 randomization. This suggests that even under violation of the PH assumption, the oracle estimator converges to β^*, hence remains interpretable.

In the next section, we study the estimation of β^* under observational data, which could be subject to treatment confounding, as well as informative right censoring. Correcting for confounding and censoring would require us to build working models for the treatment assignment and the censoring process, so we will also explore doubly robust estimators that would be consistent and asymptotically normal even when some of the working models are misspecified or estimated at a slower rate.

5.4 Doubly Robust Estimating Functions

In this section, we treat both potential outcomes and censoring as missing data and apply the semiparametric theory (Bickel et al., 1993; Tsiatis, 2006). First, we show that β^* can be equivalently defined using a set of i.i.d. full data estimating functions. Next, we show that the full data estimating functions defined for the marginal structural Cox model in Chapter 4 is also the full data estimating functions for β^*. This allows us to obtain the inverse probability weighted (IPW) estimator and the Augmented IPW (AIPW) estimator for β^* that is the same as those we derived for the marginal structural Cox model earlier.

5.4.1 Full Data Estimating Function

β^* is the limit of the full data PL estimator for the hazard ratio from the marginal structural Cox model. To model a semiparametric marginal structural Cox model, a scalar parameter β^* is not sufficient, so it is natural for us to consider an additional parameter $\Lambda^*(t)$ that can be thought of as the least false baseline cumulative hazard function from the misspecified marginal structural Cox model. We require $\Lambda^*(t)$ to be right-continuous, non-decreasing with $\Lambda^*(0) = 0$ and $\Lambda^*(\tau) < \infty$. Next, using the counting process notations, we define the full data counting process $N_T^a(t) = I(T(a) \leq t)$, and the full data at-risk process $Y_T^a(t) = I(T(a) \geq t)$ for $a = 0, 1$, where $I(\cdot)$ is an indicator function. Consider

$$M_T^a(t; \beta, \Lambda) = N_T^a(t) - \int_0^t Y_T^a(u) e^{\beta a} d\Lambda(u).$$

Motivated by the estimating functions (4.2) and (4.3) for the Cox PH model proposed in Chapter 4, we consider the set of full data estimating functions

$$D_1^f(t;\beta,\Lambda) = \sum_{a=0,1} dM_T^a(t;\beta,\Lambda), \tag{5.5}$$

$$D_2^f(\beta,\Lambda) = \sum_{a=0,1} \int_0^\tau a \cdot dM_T^a(t;\beta,\Lambda). \tag{5.6}$$

Lemma 4. *β^* and $\Lambda^*(t)$ solves the set of equations $E\{D_1^f(t;\beta,\Lambda)\} = 0$ for each t, and $E\{D_2^f(\beta,\Lambda)\} = 0$, where β^* satisfies* (5.4)*,while*

$$\Lambda^*(t) = \int_0^t \frac{\sum_{a=0,1} dF_a(t)}{\sum_{a=0,1} S_a(t)e^{\beta^* a}} = -\int_0^t \frac{\sum_{a=0,1} dS_a(t)}{\sum_{a=0,1} S_a(t)e^{\beta^* a}}. \tag{5.7}$$

Lemma 4 shows that (5.5) and (5.6) indeed are the full data estimating functions for β^* and $\Lambda^*(t)$.

We note that with this formulation, β^* and $\Lambda^*(t)$ are also the least false log hazard ratio and cumulative baseline hazard function that maximizes the pseudo non-parametric likelihood for the marginal structural Cox model (White, 1982; Li and Duan, 1989; Lin and Wei, 1989). This can be seen from the fact the pseudo non-parametric likelihood for a single observation with a fixed hazard ratio β and a cumulative baseline hazard function $\Lambda_0(t)$ with jumps discretized only at the event times is

$$\begin{aligned} &\prod_{a=0,1} \left\{ e^{\beta a} \lambda_0\{T(a)\} \right\} \exp\left\{ -\int_0^{T(a)} e^{\beta a} d\Lambda_0(t) \right\} \\ =& \prod_{a=0,1} \left\{ e^{\beta a} \lambda_0\{t\} \right\}^{I\{t=T(a)\}} \exp\left\{ -\int_0^\tau Y_T^a(t) e^{\beta a} d\Lambda_0(t) \right\} \end{aligned}$$

and that the full data estimating functions (5.5) and (5.6) can be shown to be the scores of this likelihood with respect to $\Lambda_0(t)$ and β, respectively.

5.4.2 IPW Estimating Function

Full data is not available to us in practice, but instead, we have the observed counting process $N(t) = I(X \leq t, \Delta = 1)$ where $\Delta = I(T < C)$, and the observed at-risk process $Y(t) = I(X \geq t)$. Define

$$M(t;\beta,\Lambda) = N(t) - \int_0^t Y(u)e^{\beta A}d\Lambda(u).$$

To account for the bias caused by not observing part of the data, we may apply the inverse probability of censoring weighting. The general idea of weighting an observation by the inverse of its probability of being sampled dates back to Horvitz and Thompson (1952). In our case, we first apply the inverse probability of treatment weighting to obtain a pseudo population from the target population with balanced covariates (Hernán and Robins, 2020). Next, we apply the inverse probability of censoring weighting to similarly account for informative censoring bias (Hernán et al., 2001). Since the weightings are applied to the full data estimating functions (5.5) and (5.6), which are the same as those used in Chapter 4, we also arrive at the same IPW estimating functions

$$\begin{aligned} D_1^w(t;\beta,\Lambda,\pi,S_c) =& \frac{dM(t;\beta,\Lambda)}{\pi(Z)^A\{1-\pi(Z)\}^{1-A}S_c(t;A,Z)}, \\ D_2^w(\beta,\Lambda,\pi,S_c) =& \int_0^\tau \frac{A \cdot dM(t;\beta,\Lambda)}{\pi(Z)^A\{1-\pi(Z)\}^{1-A}S_c(t;A,Z)}, \end{aligned}$$

where $\pi(z) = P(A = 1|Z = z)$ is the propensity score, $S(t;a,z) = P(T > t|A = a, Z = z)$ and $S_c(t;a,z) = P(C > t|A = a, Z = z)$ are the conditional survival functions of T and C, respectively.

5.4.3 AIPW Estimating Function

For the IPW estimating functions to be unbiased, both $\pi(z)$ and $S_c(t;a,z)$ need to be known (Hernán et al., 2001). However, these quantities are often unknown in practice and working models for propensity score and conditional censoring time are used. When these models are misspecified, the resulting β^* estimate is no longer consistent.

To protect against possible misspecification of the models, we make use of semiparametric theory (Tsiatis, 2006) to augment the IPW estimating function, such that the resulting AIPW estimating function possesses the so-called doubly robust properties that will be described later. With identical full data estimating functions, we may augment the IPW estimating functions using the same approach as that of Chapter 4.

Denote the counting process for censoring events $N_c(t) = I(X \leq t, \Delta = 0)$, and $\Lambda_c(t;a,z) = \int_0^t S_c(u;a,z)^{-1} d\{1 - S_c(u;a,z)\}$ the cumulative hazard function of C given $A = a$ and $Z = z$. Define $M_c(t;a,z,S_c) = N_c(t) - \int_0^t Y(u) d\Lambda_c(u;a,z)$, which is a martingale with respect to its natural history filtration if S_c is correctly modeled. The AIPW estimating functions are

$$\begin{aligned}
D_1(t;\beta,\Lambda,\pi,S,S_c) &= d\mathcal{N}^{(0)}(t;\pi,S,S_c) - \Gamma^{(0)}(t;\beta,\pi,S,S_c)d\Lambda(t), \\
D_2(\beta,\Lambda,\pi,S,S_c) &= \int_0^\tau d\mathcal{N}^{(1)}(t;\pi,S,S_c) - \Gamma^{(1)}(t;\beta,\pi,S,S_c)d\Lambda(t),
\end{aligned}$$

where for $l = 0, 1$,

$$\begin{aligned}
d\mathcal{N}^{(l)}(t;\pi,S,S_c) =& \frac{A^l dN(t)}{\pi(Z)^A\{1-\pi(Z)\}^{1-A} S_c(t;A,Z)} + \frac{A^l dS(t;A,Z)}{\pi(Z)^A\{1-\pi(Z)\}^{1-A}} \\
& - \sum_{a=0,1} a^l \left\{1 + \frac{A^a(1-A)^{1-a}}{\pi(Z)^a\{1-\pi(Z)\}^{1-a}} J(t;a,S,S_c)\right\} dS(t;a,Z), \\
\Gamma^{(l)}(t;\beta,\pi,S,S_c) =& \frac{A^l Y(t) e^{\beta A}}{\pi(Z)^A\{1-\pi(Z)\}^{1-A} S_c(t;A,Z)} - \frac{A^l S(t;A,Z) e^{\beta A}}{\pi(Z)^A\{1-\pi(Z)\}^{1-A}} \\
& + \sum_{a=0,1} a^l \left\{1 + \frac{A^a(1-A)^{1-a}}{\pi(Z)^a\{1-\pi(Z)\}^{1-a}} J(t;a,S,S_c)\right\} S(t;a,Z) e^{\beta a},
\end{aligned}$$

and $J(t;a,z,S,S_c) = \int_0^t dM_c(u;a,z,S_c)/\{S(u;a,z)S_c(u;a,z)\}$.

Let superscript '^{o}' denote the true value of a parameter. The AIPW estimating functions are doubly robust in the sense that when Assumptions 13-17 hold,

$E\{D_1(t;\beta^*,\Lambda^*,S,S_c)\} = 0$ and $E\{D_2(\beta^*,\Lambda^*,S,S_c)\} = 0$ for $t \in [0,\tau]$ if either $S = S^o$ or $(\pi, S_c) = (\pi^o, S_c^o)$. The proof of this result is the same as that of Theorem 10.

Next, given n i.i.d. data points, we may estimate β^*, Λ^* by solving

$$\frac{1}{n}\sum_{i=1}^{n} D_{1i}(t;\beta,\Lambda,\pi,S,S_c) = 0, \tag{5.8}$$

$$\frac{1}{n}\sum_{i=1}^{n} D_{2i}(\beta,\Lambda,\pi,S,S_c) = 0. \tag{5.9}$$

Define

$$\mathcal{S}^{(l)}(t;\beta,\pi,S,S_c) = \frac{1}{n}\sum_{i=1}^{n}\Gamma_i^{(l)}(t;\beta,\pi,S,S_c)$$

for $l = 0,1$, and let $\bar{A}(t;\beta,\pi,S,S_c) = \mathcal{S}^{(1)}(t;\beta,\pi,S,S_c)/\mathcal{S}^{(0)}(t;\beta,\pi,S,S_c)$, we first profile out (5.8) to get

$$\widetilde{\Lambda}(t;\beta,\pi,S,S_c) = \frac{1}{n}\sum_{i=1}^{n}\int_0^t \frac{d\mathcal{N}_i^{(0)}(u;\pi,S,S_c)}{\mathcal{S}^{(0)}(u;\beta,\pi,S,S_c)},$$

which can be plugged into (5.9) and obtain

$$U(\beta,\pi,S,S_c) = \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau d\mathcal{N}_i^{(1)}(t;\pi,S,S_c) - \bar{A}(t;\beta,\pi,S,S_c)d\mathcal{N}_i^{(0)}(t;\pi,S,S_c) = 0,$$

5.5 Estimating Equation and Implementation

The three nuisance functions $\pi(z)$, $S(t;a,z)$ and $S_c(t;a,z)$ are often unknown, which requires us to construct estimators $\widehat{\pi}(z)$, $\widehat{S}(t;a,z)$ and $\widehat{S}_c(t;a,z)$ in practice. Parametric and semi-parametric models, like the logistic regression, Cox PH model, and the exponential model, were commonly used since they require low computing power and converge at root-n rate, which leads to the AIPW estimator also converging at root-n rate. This leads to the

AIPW estimator $\widehat{\beta}$, which solves the equation

$$U(\beta,\widehat{\pi},\widehat{S},\widehat{S}_c)=\frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau} d\mathcal{N}_i^{(1)}(t;\widehat{\pi},\widehat{S},\widehat{S}_c)-\bar{A}(t;\beta,\widehat{\pi},\widehat{S},\widehat{S}_c)d\mathcal{N}_i^{(0)}(t;\widehat{\pi},\widehat{S},\widehat{S}_c)=0.$$

However, the actual model is likely to be much more complex, this makes parametric/semi-parametric models prone to model misspecification. On the other hand, ML methods, like Splines (Gray, 1992; Kooperberg et al., 1995a) and Random Survival Forest (Ishwaran et al., 2008) are much more flexible and are often able to estimate the nuisance functions consistently even when they are complex. They were rarely used in practice mostly because ML estimators do not have root-n error rates, which forbids the asymptotic normality of $\widehat{\beta}$ to be established. To this end, we utilize the cross-fitting approach that helps to address this issue (Chernozhukov et al., 2018). Specifically, cross-fitting cycles the sample-splitting procedure over the entire samples, where for each sample-spitting procedure, we use one part of the data to estimate the nuisance functions, while the other part for constructing the estimating functions.

Suppose we have a sample of n observations that are split into k folds of equal size with index sets $I_1, I_2, \dots, I_k$, then for each fold m, we may define the m-th fold specific quantities as:

$$\mathcal{S}_m^{(l)}(t;\beta,\pi,S,S_c)=\frac{1}{|I_m|}\sum_{i\in I_m}\Gamma_i^{(l)}(t;\beta,\pi,S,S_c),$$

and $\bar{A}_m(t;\beta,\pi,S,S_c)=\mathcal{S}_m^{(1)}(t;\beta,\pi,S,S_c)/\mathcal{S}_m^{(0)}(t;\beta,\pi,S,S_c)$. A sample-splitting procedure for the m-th fold can be done by first solving for $\sum_{i\in I_m}^{n} D_{1i}(t;\beta,\Lambda,\pi,S,S_c)=0$ to get

$$\widetilde{\Lambda}_m(t;\beta,\pi,S,S_c)=\frac{1}{|I_m|}\sum_{i\in I_m}\int_0^t \frac{d\mathcal{N}_i^{(0)}(u;\pi,S,S_c)}{\mathcal{S}_m^{(0)}(u;\beta,\pi,S,S_c)},$$

which can then be plugged into $\sum_{i\in I_m}^{n} D_{2i}(\beta,\Lambda,\pi,S,S_c)=0$ to obtain the m-th fold estimating

Algorithm 6 Cross-fitted AIPW estimation for β^*

```
Input: A sample of $n$ observations that are split into $k$ folds with index sets $I_1, I_2, \dots, I_k$.
for each fold indexed by $m$ do
    Obtain estimated nuisance functions $\widehat{\pi}^{(-m)}$, $\widehat{S}^{(-m)}$ and $\widehat{S}_c^{(-m)}$) using the out-of-fold
    sample indexed by $I_{-m} := \{1, \dots, n\} \setminus I_m$.
end for
Output: $\widehat{\beta}_{cf}$, the solution to
    $$\frac{1}{k}\sum_{m=1}^{k} U_m(\beta, \widehat{\pi}^{(-m)}, \widehat{S}^{(-m)}, \widehat{S}_c^{(-m)}) = 0.$$
```

equation

$$U_m(\beta, \pi, S, S_c) = \frac{1}{|I_m|}\sum_{i \in I_m}\int_0^\tau \left\{ d\mathcal{N}_i^{(1)}(t; \pi, S, S_c) - \bar{A}_m(t; \beta, \pi, S, S_c) d\mathcal{N}_i^{(0)}(t; \pi, S, S_c) \right\}.$$

Cross-fitting repeats the above sample-splitting procedure for each of the k folds, and its details are summarized in Algorithm 6.

5.6 Asymptotic Properties

In this section, we discuss the asymptotic properties of the non-cross-fitted AIPW estimator $\widehat{\beta}$ and the cross-fitted AIPW estimator $\widehat{\beta}_{cf}$. Some of the technical assumptions will be omitted in this section to focus more on the discussion of the DR properties.

First, we look at the non-cross-fitted AIPW estimator $\widehat{\beta}$.

Assumption 18 (Uniform Convergence)**.** *There exist $\pi^*(z)$, $S^*(t; a, z)$ and $S_c^*(t; a, z)$ with*

$$E\{|\widehat{\pi}(Z) - \pi^*(Z)|^2\} = o(1),$$
$$E\left\{ \sup_{t \in [0,\tau], a \in \{0,1\}} |\widehat{S}(t; a, Z) - S^*(t; a, Z)|^2 \right\} = o(1),$$
$$E\left\{ \sup_{t \in [0,\tau], a \in \{0,1\}} |\widehat{S}_c(t; a, Z) - S_c^*(t; a, Z)|^2 \right\} = o(1).$$

Assumption 18 assumes that the nuisance estimators $\widehat{\pi}$, $\widehat{S}$ and $\widehat{S}_c$ converge to some limiting functions π^*, S^* and S_c^* that are not necessarily the truth. Under Assumptions 13-18 and some regularity assumptions, the non-cross-fitted AIPW estimator $\widehat{\beta}$ has the model double robustness property. That is, if either $S^* = S^o$ or $(\pi^*, S_c^*) = (\pi^o, S_c^o)$, then $\widehat{\beta} \xrightarrow{p} \beta^*$, and if $\widehat{\pi}(z)$, $\widehat{S}(t;a,z)$ and $\widehat{S}(t;a,z)$ are regular and asymptotically linear estimators of $\pi^*(z)$, $S^*(t;a,z)$ and $S_c^*(t;a,z)$, we then have $\sqrt{n}(\widehat{\beta} - \beta^*)/\sigma \xrightarrow{d} N(0,1)$ for some variance σ^2. When $(\pi^*, S^*, S_c^*) = (\pi^o, S^o, S_c^o)$, σ^2 can be consistently estimated by $\widehat{\sigma}^2 := \widehat{\sigma}^2(\widehat{\beta})$, and its expression is given in Appendix D.1.

Note that, the condition of regular and asymptotically linear estimators essentially restricts us to root-n consistent estimators, which include common parametric models like the exponential model, and semi-parametric models like the Cox PH model. This type of condition is also what most classical DR literature explicitly or implicitly assumes, and due to the root-n convergence rates of $\widehat{\pi}$, $\widehat{S}$ and $\widehat{S}_c$, we do not need cross-fitting to establish the asymptotic normality. The classical model DR result is widely studied for many survival problems, so we omit the proofs and refer readers to similar articles, like Wang et al. (2022), for details.

The asymptotic variance estimator is established when all nuisance estimators are root-n consistent. When one of the nuisance function estimators is not consistent, $\widehat{\beta}$ is still asymptotically normal, but the variance becomes too complicated to be estimated in close form. For this reason, we also suggest the use of resampling methods, such as bootstrap (Efron, 1992) for estimating the asymptotical variance, which works even when one of the working models is misspecified.

Next, we look at the cross-fitted AIPW estimator $\widehat{\beta}_{cf}$. Let $O^\dagger$ denote a sample of n i.i.d. data vectors $\{(X_i^\dagger, \Delta_i^\dagger, A_i^\dagger, Z_i^\dagger), i = 1, \ldots, n\}$ used for estimating $\widehat{\pi}$, $\widehat{S}$ and $\widehat{S}_c$. Let (X, Δ, A, Z) be a data vector independent of $O^\dagger$ and drawn from the same distribution as $O^\dagger$.

Define

$$
\begin{aligned}
\|\widehat{\pi}-\pi^*\|_{\dagger}^2 &= E^{\dagger}\left(E\left[\{\widehat{\pi}(Z)-\pi^*(Z)\}^2\right]\right),\\
\left\|\widehat{S}-S^*\right\|_{\dagger}^2 &= E^{\dagger}\left(E\left[\left\{\sup_{t\in[0,\tau],a\in\{0,1\}}\left|\widehat{S}(t;a,Z)-S^*(t;a,Z)\right|\right\}^2\right]\right),\\
\left\|\widehat{S}_c-S_c^*\right\|_{\dagger}^2 &= E^{\dagger}\left(E\left[\left\{\sup_{t\in[0,\tau],a\in\{0,1\}}\left|\widehat{S}_c(t;a,Z)-S_c^*(t;a,Z)\right|\right\}^2\right]\right),
\end{aligned}
$$

where $E^{\dagger}$ denotes expectation taken with respect to $(X^{\dagger},\Delta^{\dagger},A^{\dagger},Z^{\dagger})$, and E denotes expectation taken with respect to O conditional on $O^{\dagger}$.

Assumption 19 (Uniform Convergence)**.** *There exist $\pi^*(z)$, $S^*(t;a,z)$ and $S_c^*(t;a,z)$ such that $\|\widehat{\pi}-\pi^*\|_{\dagger}=o(1)$, $\|\widehat{S}-S^*\|_{\dagger}=o(1)$ and $\|\widehat{S}_c-S_c^*\|_{\dagger}=o(1)$.*

Assumption 20 (Rate Condition)**.** *$(\pi^*,S^*,S_c^*)=(\pi^o,S^o,S_c^o)$ and*

$$
\left\|\widehat{S}-S^o\right\|_{\dagger}\left(\|\widehat{\pi}-\pi^o\|_{\dagger}+\left\|\widehat{S}_c-S_c^o\right\|_{\dagger}\right)+\mathcal{D}^{\dagger}(\widehat{S},\widehat{S}_c;S^o,S_c^o)=o(n^{-1/2}),
$$

where $\mathcal{D}^{\dagger}(\widehat{S},\widehat{S}_c;S^o,S_c^o)$ is defined in (4.10).

The rate condition (Smucler et al., 2019) states that the product of the error rate between the conditional outcome model $\widehat{S}$ and the model for missing mechanism $(\widehat{\pi},\widehat{S}_c)$ is faster than root-n. Due to the involvement of the time component in time-to-event analysis, an integral product error term $\mathcal{D}^{\dagger}(\widehat{S},\widehat{S}_c;S^o,S_c^o)$ is also required, which we discussed in detail in Chapter 4. Under Assumptions 13-17, 19-20 and some regularity assumptions, $\widehat{\beta}_{cf}$ satisfies the rate double robustness property (Smucler et al., 2019). That is, $\widehat{\beta}_{cf}\xrightarrow{p}\beta^*$, and $\sqrt{n}(\widehat{\beta}_{cf}-\beta^*)/\widehat{\sigma}_{cf}\xrightarrow{d}N(0,1)$ for a cross-fitted variance estimator $\widehat{\sigma}_{cf}^2:=\widehat{\sigma}_{cf}^2(\widehat{\beta}_{cf})$ defined in (D.2). The proof of the rate DR property as well as some of the additional regularity conditions required can be found in the proofs of Theorem 11 and Theorem 12 in Chapter 4.

The rate condition is satisfied when we have consistent nuisance estimators $\widehat{S}$ and $(\widehat{\pi},\widehat{S}_c)$ with faster than root-n product error rate, even if one of them converges arbitrarily slowly. Therefore, in addition to the root-n consistent nuisance function estimators, the rate DR property also allows for the use of slower than root-n rate estimators. This condition is a relaxation over the root-n condition from model DR and makes it possible for us to apply many ML methods. To this day, few convergence rate results exist for ML methods in survival analysis, so we will empirically investigate the performance of some ML methods in the simulation section.

5.7 Miscellaneous Results

5.7.1 Bias of the PL Estimator

We showed in Section 5.3.2 that, as a result of (5.4), β^*, which is the limit of the PL estimator (sometimes called the Naive Cox estimator) under full data is a weighted average of $\beta(t)$. Following the derivations in Struthers and Kalbfleisch (1986), we can similarly show that under observed data, if we have randomization and non-informative censoring ($T(a) \perp C(a)|A$), then the limit of the PL estimator is the solution to the equation $\int_0^\tau h_{random}(t;\beta)dt = 0$ with

$$h_{random}(t;\beta) = \left[\frac{\sum_{a=0,1} ae^{\beta(t)a}S_c^a(t)p^a(1-p)^{1-a}}{\sum_{a=0,1} ae^{\beta(t)a}S_c^a(t)p^a(1-p)^{1-a}} - \frac{\sum_{a=0,1} ae^{\beta a}S_c^a(t)p^a(1-p)^{1-a}}{\sum_{a=0,1} ae^{\beta a}S_c^a(t)p^a(1-p)^{1-a}}\right] \times \left\{\sum_{a=0,1} f_a(t)S_c^a(t)p^a(1-p)^{1-a}\right\}.$$

where $S_c^a(t) = P(C(a) > t)$ and $p = (A = 1)$. Since $h_{random}(t;\beta(t)) = 0$, if we apply the mean-value Theorem on $h_{random}(t;\beta)$, we see that the limit of the PL estimator remains a weighted average of $\beta(t)$ under the observed data. This property forces the PL estimate to be within the range of $\beta(t)$, which restricts how large the bias can become.

On the other hand, under informative censoring and treatment confounding, the limit of the PL estimator is the solution to the equation $\int_0^\tau h_{general}(t;\beta)dt = 0$ with

$$\begin{aligned}
&h_{general}(t;\beta)\\
&=\int_0^\tau \left[\frac{\sum_{a=0,1} aE\left[f(t;a,Z)S(t;a,Z)S_c(t;a,Z)P(A=a|Z)\right]}{\sum_{a=0,1} E\left[f(t;a,Z)S(t;a,Z)S_c(t;a,Z)P(A=a|Z)\right]}\right.\\
&\qquad \left.-\frac{\sum_{a=0,1} ae^{\beta a}S(t;a,Z)S_c(t;a,Z)P(A=a|Z)}{\sum_{a=0,1} e^{\beta a}S(t;a,Z)S_c(t;a,Z)P(A=a|Z)}\right]\\
&\quad \times \sum_{a=0,1} E\left[f(t;a,Z)S_c(t;a,Z)P(A=a|Z)\right]dt. \qquad (5.10)
\end{aligned}$$

Unlike the special case of randomization and non-informative censoring where $h_{random}(t;\beta(t)) = 0$, $h_{general}(t;\beta(t))$ is non-zero in general, so applying mean-value Theorem on (5.10) leads to β^* being a weighted average of $\beta(t)$ plus a non-zero bias term. In fact, this non-zero bias term occurs if we have either informative censoring or treatment confounding. As we will demonstrate in the simulation section, the bias of the PL estimator can indeed be much larger under informative censoring or treatment confounding, with the PL estimate much outside the range of $\beta(t)$.

The dependency of the PL estimator on the censoring distribution and the propensity score is also a warning against naively reporting the hazard ratio estimate from the PL estimator, which is ubiquitous in practice. This leads to one of the important advantages of the proposed causal estimand β^*, which is independent of the study-specific censoring distribution and treatment assignment process. In particular, informative censoring can exist in many different situations, even under randomized trials, potentially leading to uninterpretable results (Nguyen and Gillen, 2017; Nuño and Gillen, 2021). The proposed causal estimand β^* allows us to replicate and compare results across studies that come with potentially different censoring distribution and treatment assignments.

5.7.2 AIPW Log Hazard Ratio Plot

Our estimand of interest β^* is a weighted average of $\beta(t)$. In this section, we look at a common way of plotting $\beta(t)$ under randomization and look to generalize it to observational data with informative censoring. Therneau and Grambsch (2000) showed that if the data is under randomization and randomly censored, then through Taylor's expansion, the PL estimator $\widehat{\beta}_{pl}$ satisfies the approximation

$$E\left\{\frac{s_{pl}(\widehat{\beta}_{pl},t)}{V_{pl}(\widehat{\beta}_{pl},t)}\right\}+\widehat{\beta}_{pl}\approx\beta(t),$$

where $s_{pl}(\widehat{\beta}_{pl},t)$ is the Schonfeld residual at event time t, while $V_{pl}(\widehat{\beta}_{pl},t)$ is the contribution of its variance at event time t, making the term inside the expectation a scaled Schoenfeld residual at an event time t. A smoothed plot of $s_{pl}(\widehat{\beta}_{pl},t_i)/V_{pl}(\widehat{\beta}_{pl},t_i)$ against event time t_i using splines would then lead to a good approximation of $\beta(t)$. This smoothed $\beta(t)$ plot is implemented in the 'cox.zph' function of the 'survival' package, which we used to construct a full data $\beta(t)$ plot for Scenario 3 and 4 as in Figure 5.3.

Since β^* is the limit of the PL estimator under full data, we have

$$E\left\{\frac{s_{pl}(\beta^*,t)}{V_{pl}(\beta^*,t)}\right\}+\beta^*\approx\beta(t),$$

under full data. It should not come as a surprise that, without randomization and with informative censoring, this approximation would fail under observational data. To correct for this bias, we apply AIPW to each of β^*, $s_{pl}(\beta^*,t)$ and $V_{pl}(\beta^*,t)$. In particular, we replace β^* and $s_{pl}(\beta^*,t)/V_{pl}(\beta^*,t)$ by the cross-fitted AIPW estimator and the cross-fitted AIPW scaled residual contribution to arrive at the AIPW approximation

$$E\left\{\frac{s_{aipw}(\widehat{\beta}_{cf},t)}{V_{aipw}(\widehat{\beta}_{cf},t)}\right\}+\widehat{\beta}_{cf}\approx\beta(t). \tag{5.11}$$

The expression for the cross-fitted scaled AIPW residual $s_{aipw}(\widehat{\beta}_{cf},t)/V_{aipw}(\widehat{\beta}_{cf},t)$ is given at the end of Appendix D.1.

Figure 5.1 presents the smoothed $\beta(t)$ plots using Therneau and Grambsch's method and our proposed AIPW $\beta(t)$ plot fitted using the observed data simulated according to Scenario 4 of the Simulation Section 5.8, as well as the true $\beta(t)$ plot generated using full data. The $\beta(t)$ from Therneau and Grambsch were generated using 1 million observed data points, while for the AIPW $\beta(t)$ plot, we can not train using 1 million observations in one go, so we follow our simulation approach to combine outputs from 1000 simulation runs, each with 1000 observations. We can see that under the observed data, Therneau and Grambsch's $\beta(t)$ plot shows significant bias compared to the truth, while the AIPW $\beta(t)$ plot closely follows the truth, except at the tails where the spline is known to be unstable.

5.7.3 A Non-Causal Estimand and Its Properties

In this paper, we primarily focused on the causal estimand β^* due to its broader application in observational studies. However, we are sometimes more interested in the associative relationship between the exposure and the outcome, which can be characterized by a regression model. For example, this could happen when our exposure of interest can not be applied or acted, which fails to satisfy the definition of treatment in causal inference (Hernán and Robins, 2020). This motivates us to consider the study of a non-causal β^*, which can be readily defined as the solution to the estimating equation (5.2). We note that this non-causal β^* is also equivalent to the average regression effect defined in Xu and O'Quigley (2000), and the estimation of this non-causal β^* are studied in Boyd et al. (2012); Hattori and Henmi (2012) where they assumed non-informative censoring ($T \perp C|A$). Assuming randomization and informative censoring $C \perp T|A,Z$, the non-causal β^* has similar properties as the causal β^* that we studied, and we can estimate it using AIPCW approaches proposed in Chapter 3. We now summarize the properties and the estimation of β^* without going into too much detail.

The non-causal β^* is a weighted average of the non-causal $\beta(t)$ where $\beta(t)$ is the log

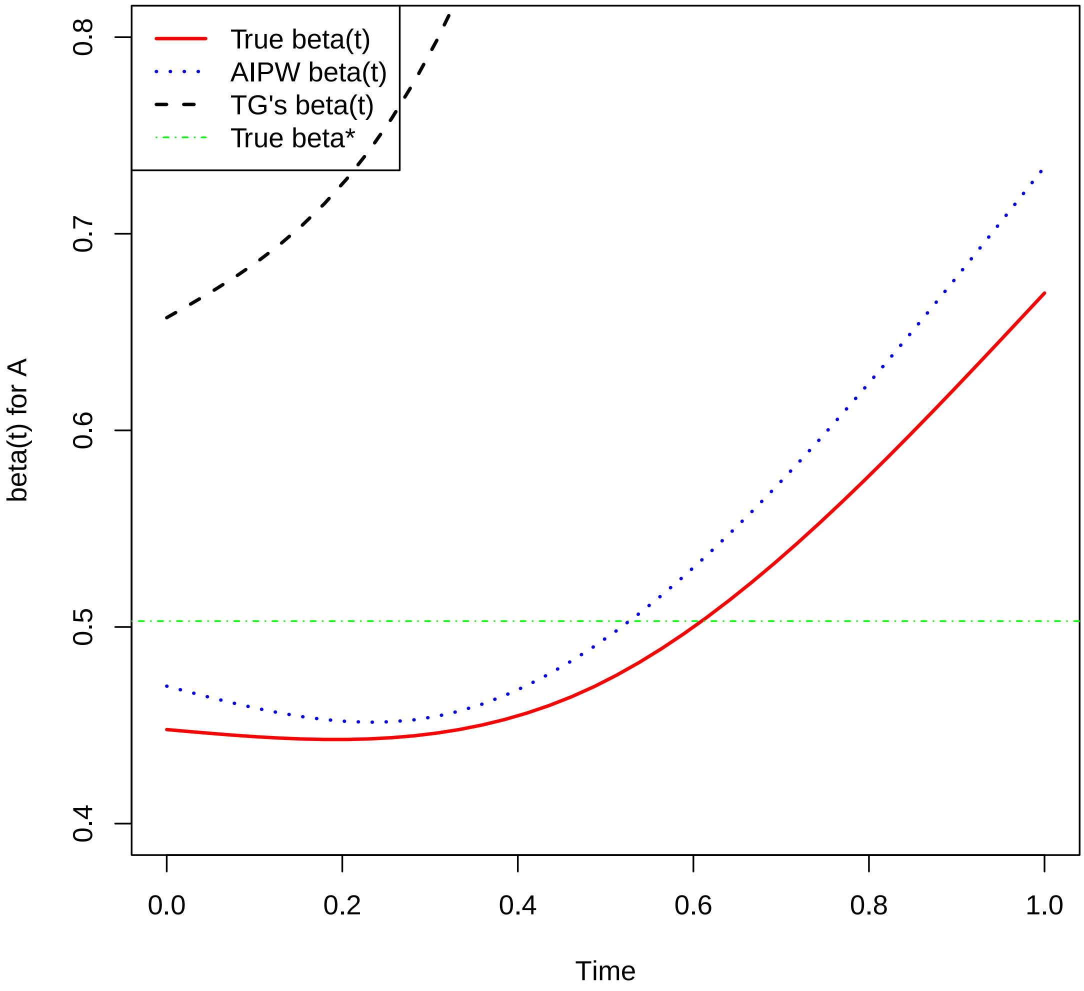


Figure 5.1: Smoothed $\beta(t)$ plot using 1 million observations from Scenario 4 in the Simulation Section 5.8.

hazard ratio from the non-causal non-PH model (5.1). It has a simple algebraic relationship with the parameters from the transformation model. It is also the limit of the PL estimator under the full data, where the full data here refers to observations with randomization without censoring. This last property again allows us to construct full data estimating function for it, which is the same as (3.2) and (3.3) we derived in Chapter 3. To correct for informative censoring, we similarly construct IPCW estimating functions (3.5) and (3.6). Then we augment the IPCW estimating functions to obtain AIPCW estimating functions (3.11) and (3.12). A Cross-fitted AIPCW estimating equation can be obtained following Algorithm 4.

The non-cross-fitted AIPCW estimator also has the model DR property in that it is CAN if one of the outcome and the censoring models is correctly specified. The cross-fitted AIPCW estimator would additionally have the rate DR property. Moreover, as we demonstrated, even if the treatment is randomized, if there is censoring, the bias of the PL estimator could potentially be large and outside the range of the $\beta(t)$. We can also construct AIPCW $\beta(t)$ plot for the non-causal $\beta(t)$.

Next, we conduct numerical studies on the estimation of the causal β^* where we compare our proposed AIPW estimator against other extant estimators. The performance of the AIPCW estimator under the non-causal setting is largely similar, which we omit.

5.8 Simulation

In this section, we compare the performance of the AIPW estimators $\widehat{\beta}$ and the cross-fitted AIPW estimators $\widehat{\beta}$ using different working models, against different IPW estimators, a Naive Cox estimator that does not adjust for any covariates, and a full data estimator that have access to full potential outcomes. For each simulated dataset, we set the number of observations to $n = 1000$. We also consider multiple scenarios for the underlying data-generating processes of T and π, C, and simulate 1000 datasets for each scenario, which corresponds to a margin of error of about $+/-1.35\%$ for the coverage probability of nominal 95% confidence intervals. Five-fold cross-fitting is used.

The data generation process is summarized in figure 5.2. We set $\tau = 1$ and simulate first the covariate $Z_1 \sim 0.5U_1 + 0.5U_2$, where $U_1, U_2 \sim \text{Unif}(-1, 1)$. Then we simulate $T(a)$ and $\{C(a), A\}$ conditional on the covariate Z_1 under 4 different scenarios according to Table 5.1. In half of the scenarios, $T(a)$ follows the Cox PH distribution that can be estimated using semi-parametric models, while in the other half $T(a)$ follows a mixture of distributions that would require ML methods to consistently estimate. This is the same for the model of missing mechanisms (π, S_c). After simulating the potential outcomes and A, we obtain the observed failure and censoring times by setting $T = AT(1) + (1 - A)T(0)$, $C = AC(1) + (1 - A)C(0)$.

All 4 scenarios have a $30 - 50\%$ event rate as well as a $30 - 50\%$ censoring rate and a $10 - 30\%$ administrate censoring rate, where administrative censoring happens when $X > \tau$. True β^* is 1.014 and 0.503 when T follows the Cox PH distribution and the Mixture distribution, respectively. Since there is no analytical solution for β^*, this is calculated from fitting a PL estimator on a simulated sample of one million full data observations. Figure 5.3 demonstrates a smoothed plot of how $\beta(t)$ changes across time for the mixture model of T which is highly non-PH. The plot is created using the 'cox.zph' function from the 'survival' package and is valid when applied to full data. More discussion on this smoothed $\beta(t)$ plot, as well as a proposed AIPW $\beta(t)$ plot applicable for observational data, can be found in Section 5.7.2 below.

For the AIPW estimator, we consider 2 types of working models for T and C: Cox PH model using the R package 'survival' and random survival forest (RSF) (Ishwaran et al., 2008) using the R package 'randomForestSRC' where we set splitrule = 'bs.gradient', nodesize = 50 and mtry = 2. For π, we consider the logistic regression model and the 'twang' package in R (Ridgeway et al., 2022). Mix and match 3 pairs of working models, we have a total of 8 AIPW estimators with different T/C-PS models: Cox/Cox-logit, Cox/Cox-twang, Cox/RSF-logit, Cox/RSF-twang, RSF/Cox-logit, RSF/Cox-twang, RSF/RSF-logit, RSF/RSF-twang. Note that the convergence rate of RSF and twang are largely unknown but slower than root-*n*, while the Cox PH model and the logistic regression converges at a root-*n* rate. This allows AIPW with Cox/Cox-logit to satisfy the conditions for model DR if either one of the working models for T and (C,π) is correctly specified. Therefore, we do not apply cross-fitting to AIPW with Cox/Cox-logit, while cross-fitting the remaining 7 AIPW estimators. Model-based standard errors (SE) for non-cross-fitted and cross-fitted AIPW estimators are estimated using (D.1) and (D.2) respectively, where we assume that all nuisance function estimators are consistent. We also consider 4 different IPW estimators, where the C-PS models are Cox-logit, Cox-twang, RSF-logit, RSF-twang, respectively. The bootstrap SE estimates for AIPW estimators and IPW estimators are reported where each dataset uses 100 bootstrap runs, and for the Naive Cox and the full data estimator, we report the robust variance estimator. Note that we also clustered each pair of potential outcomes when running the full data estimator to account for the dependence between potential outcomes. Coverage probability is then calculated using normal-based 95% confidence intervals constructed from both the estimated model SEs and the bootstrap SEs.

To avoid numerical problems, we trim $\widehat{S}(t;a,z),\widehat{S}_c(t;a,z)$ so that all estimated values below 0.05 are set to 0.05, and we also set all estimated $\widehat{\pi}(z)$ values that are above 0.9 or below 0.1 to be 0.9 and 0.1. From our experience, trimming with a small threshold is able to guarantee successful convergence while introducing little bias.

Figures 5.4 show the bias, standard deviation (SD), and bootstrap-based coverage probability of the AIPW and IPW estimators under Scenarios 1-4 respectively. Additional details for all 14 estimators are provided in Tables 5.2 and 5.3. IPW:Cox-logit performs well when *C-PS* models are correct but performs poorly when the working models are incorrect. IPW estimators with partly ML methods seem to perform slightly better than IPW:Cox-logit when *C-PS* are correct, but still fails hard when the working models are incorrect. This shows that although ML methods can be mostly consistent, if the parametric model is incorrect, the entire missing data model is still incorrect, causing the IPW estimator to be biased. The IPW:RSF-twang has smaller biases for all 4 scenarios, but due to its slower than root-n convergence rate, the coverage is quite poor. AIPW:Cox/Cox-logit's performance reflects the classical DR result, showing small bias, and great coverage when at least one working model is correct but fails badly when they are all wrong. The AIPW estimator that uses partly parametric and partly ML methods, like AIPW:RSF/Cox-logit, performs well when all 3 working models are correct but often shows larger biases or poor coverage when some of the working models are incorrect. AIPW:RSF/RSF-twang estimator that uses all ML methods has low biases under all 4 scenarios along with excellent model coverage and bootstrap coverage. Lastly, it's worth noting that in all 4 scenarios, the naive Cox estimates are way outside the range of the true $\beta(t)$ plot from Figure 5.3. This finding agrees with what we discussed in Section 5.7.1, and highlights the importance of controlling for treatment confounding and informative censoring.

The performance of the AIPCW estimator under the non-causal setting is largely similar, so we omit them here.

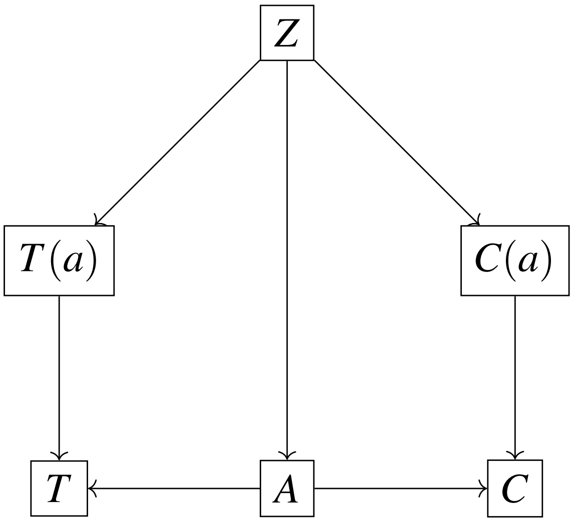


Figure 5.2: DAG for simulation

5.9 Application

We apply the proposed estimator to the International Non-Hodgkin's Lymphoma Prognostic Factors Project (Shipp, 1993). In this project, 5 risk factors deemed predictive of

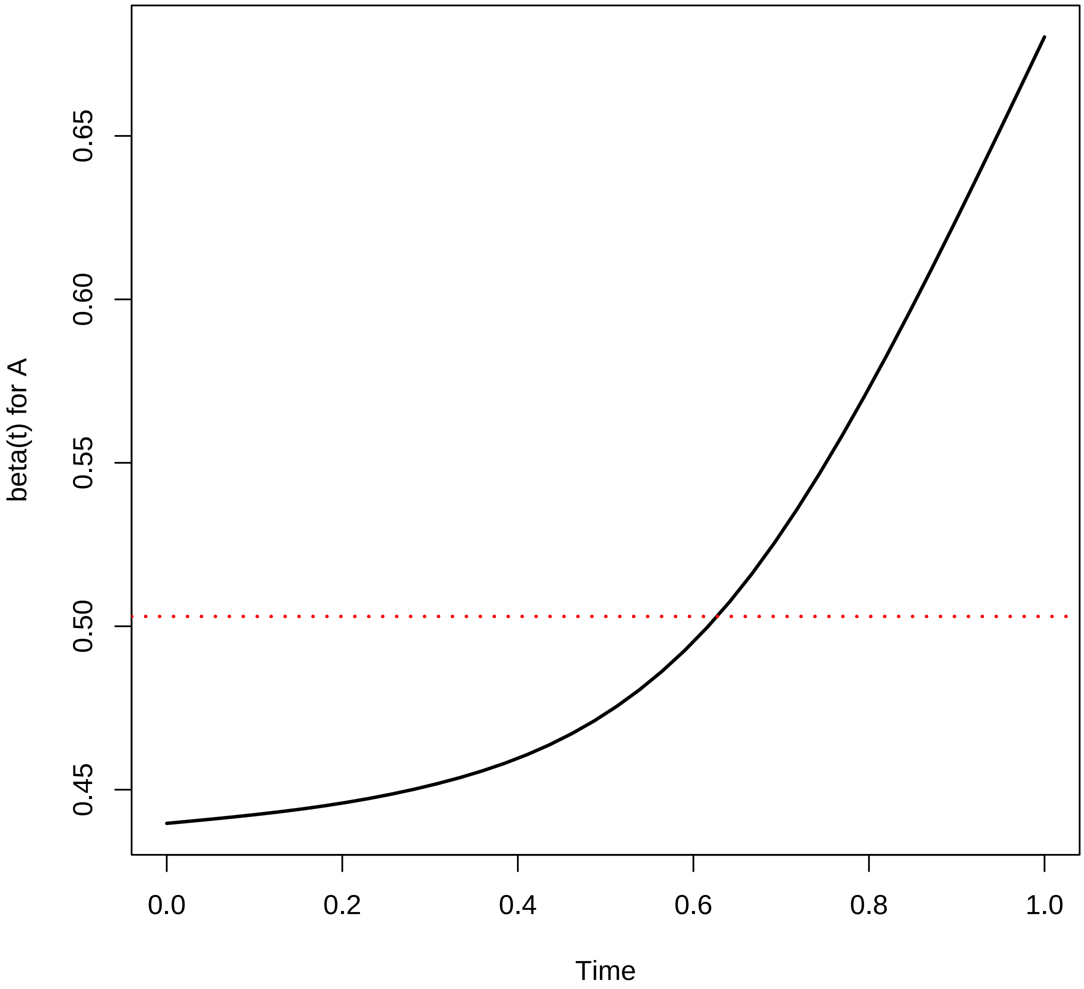


Figure 5.3: True $\beta(t)$ plot when T follows a Mixture model. True β^* is also shown in red.

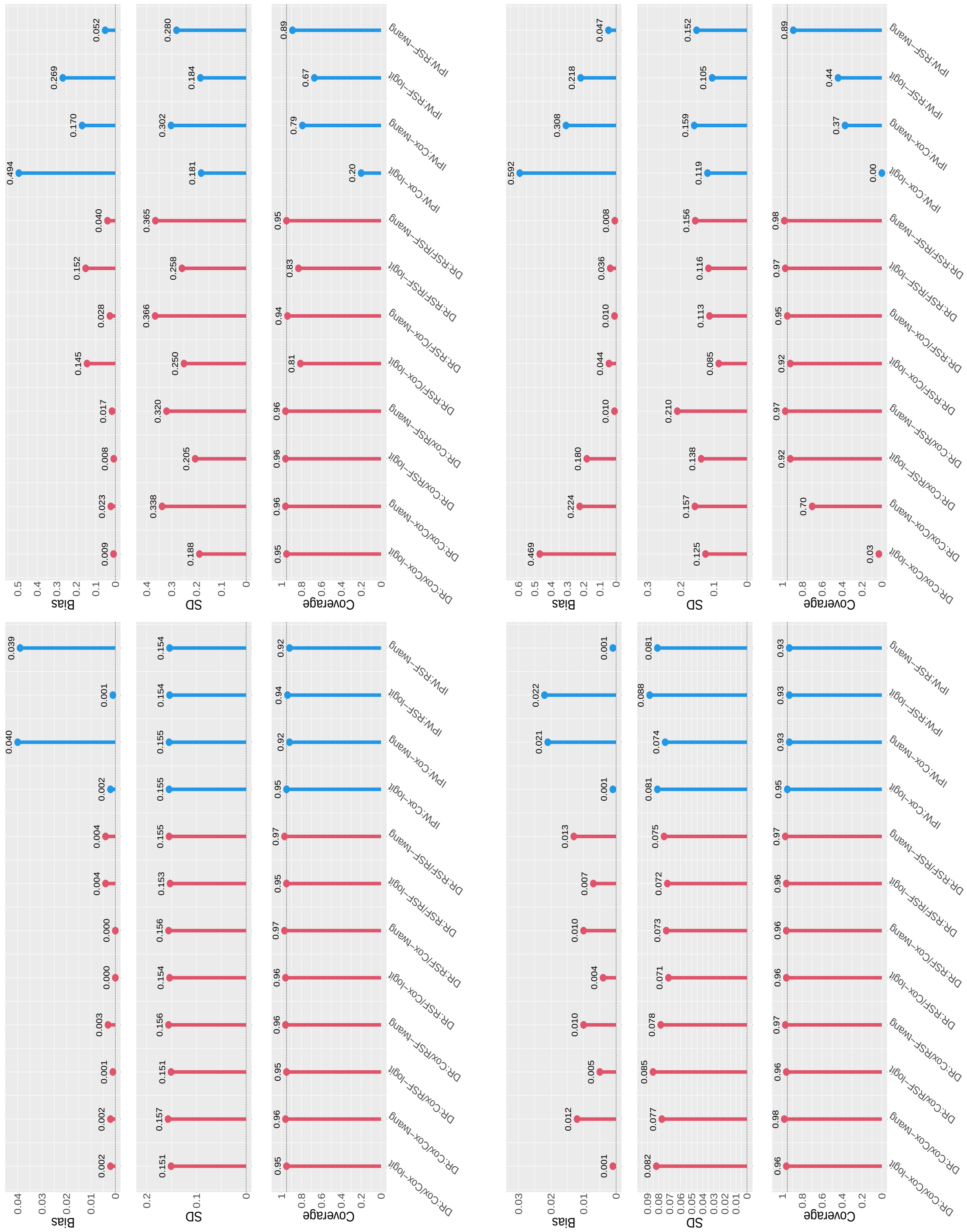


Figure 5.4: Plots of bias, bootstrap SD and bootstrap coverage for all 4 scenarios in simulation. Top-left, top-right, bottom-left, and bottom-right in the landscape view correspond to Scenario 1 to Scenario 4, respectively.

Table 5.1: Scenarios of data generating processes for T, C and π considered in the simulation.

Scenario	T/C Distributions	Data-generating mechanism
1	$T(a)$: Cox PH	$\lambda_{T(a)}(t;Z_1) = \exp(2 - 1.12a - 2Z_1)$.
	$C(a)$: Cox PH	$\lambda_{C(a)}(t;Z_1) = \exp(3.5 - 2a - 2.5Z_1)$.
	π: Logistic	$\text{logit}\{\pi(Z_1)\} = 2Z_1$
2	$T(a)$: Cox PH	$\lambda_{T(a)}(t;Z_1) = \exp(2 - 1.12a - 2Z_1)$.
	$C(a)$: Mixture	$Z_1 \leq 0$: $\lambda_{C(a)}(t;Z_1) = \exp(3.5 - 3a - 0.5Z_1)$, $Z_1 > 0$: $C(a) \sim \text{Unif}(0, 1.05)$.
	π: Soft Partition	$\text{logit}\{\pi(Z_1)\} = 2 \cdot \mathbf{1}\{Z_1 < -1/3\} - 2 \cdot \mathbf{1}\{-1/3 \leq Z_1 < 1/3\}$ $+ 2 \cdot \mathbf{1}\{Z_1 \geq 1/3\}$
3	$T(a)$: Mixture	$Z_1 \leq 0$: $\lambda_{T(a)}(t;Z_1) = \exp(5 - 3.4a + 2.5Z_1)$, $Z_1 > 0$: $T \sim \text{Unif}(0, 1.05)$.
	$C(a)$: Cox PH	$\lambda_{C(a)}(t;Z_1) = \exp(3.5 - 2a - 2.5Z_1)$.
	π: Logistic	$\text{logit}\{\pi(Z_1)\} = 2Z_1$
4	$T(a)$: Mixture	$Z_1 \leq 0$: $\lambda_{T(a)}(t;Z_1) = \exp(5 - 3.4a + 2.5Z_1)$, $Z_1 > 0$: $T(a) \sim \text{Unif}(0, 1.05)$.
	$C(a)$: Mixture	$Z_1 \leq 0$: $\lambda_{C(a)}(t;Z_1) = \exp(3.5 - 3a - 0.5Z_1)$, $Z_1 > 0$: $C(a) \sim \text{Unif}(0, 1.05)$.
	π: Soft Partition	$\text{logit}\{\pi(Z_1)\} = 2 \cdot \mathbf{1}\{Z_1 < -1/3\} - 2 \cdot \mathbf{1}\{-1/3 \leq Z_1 < 1/3\}$ $+ 2 \cdot \mathbf{1}\{Z_1 \geq 1/3\}$

overall survival were evaluated for patients with aggressive non-Hodgkin's lymphoma from 16 institutions and cooperative groups in the US, Europe, and Canada who were treated between 1982 and 1987 with combination-chemotherapy regimes containing doxorubicin. After removing observations with missing covariates, we obtained a total of 1968 patients with 628 (28.4%) events. Here we focus on the predictive effect of the Ann Arbor stage (STAGE) on overall survival, with the exposure groups defined as Ann Arbor stage III or IV (1251 patients) and the comparison group as Ann Arbor stage I or II (717 patients). 36.7% of the patients classified as STAGE III or IV experienced events during the follow-up, while 13.9% of the patients with STAGE I or II experienced events. We also include 4 binary covariates: Age > 60 (AGE), the number of extra-nodal disease sites ≥ 2 (XTRA), serum lactate dehydrogenase $> 1.5\times$ normal (LDH), and ECOG performance status > 1 (PS). To ensure the positivity assumption is satisfied, we apply administrative censoring at $\tau = 4$ years. Overall, 39.2% of the patients was censored due to lost to follow-up, and 32% was administratively censored. Figure 5.5 shows the plot of log negative log survival against log years for the 2 groups, where parallel lines would indicate PH. The figure indicates that

Table 5.2: Simulation based on 1000 data sets for Scenarios 1 and 2, each with 1000 observations. True $\beta^* = 1.014$. Red indicates that the working model or the approach is invalid.

Scenario	Estimator	T/C-PS Models	Bias	SD	SE	Coverage
					Model/Boot	Model/Boot
Scenario 1	AIPW	Cox/Cox-logit	0.002	0.151	0.151/0.151	0.95/0.95
		Cox/Cox-twang	0.002	0.157	0.164/0.168	0.95/0.96
		Cox/RSF-logit	0.001	0.151	0.154/0.157	0.95/0.95
		Cox/RSF-twang	0.003	0.156	0.165/0.167	0.96/0.96
		RSF/Cox-logit	0.000	0.154	0.155/0.159	0.95/0.96
		RSF/Cox-twang	0.000	0.156	0.166/0.175	0.96/0.97
		RSF/RSF-logit	0.004	0.153	0.156/0.159	0.95/0.95
		RSF/RSF-twang	0.004	0.155	0.167/0.175	0.96/0.97
	IPW	Cox-logit	0.002	0.155	- /0.153	- /0.95
		Cox-twang	0.040	0.155	- /0.144	- /0.92
		RSF-logit	0.001	0.154	- /0.153	- /0.94
		RSF-twang	0.039	0.154	- /0.144	- /0.92
	Naive Cox		0.470	0.152	0.151/0.151	0.11/0.11
	Full Data		0.001	0.061	0.063/0.063	0.96/0.96
Scenario 2	AIPW	Cox/Cox-logit	0.009	0.188	0.175/0.189	0.94/0.95
		Cox/Cox-twang	0.023	0.338	0.314/0.338	0.94/0.96
		Cox/RSF-logit	0.008	0.205	0.189/0.209	0.93/0.96
		Cox/RSF-twang	0.017	0.320	0.308/0.318	0.96/0.95
		RSF/Cox-logit	0.145	0.250	0.176/0.208	0.74/0.81
		RSF/Cox-twang	0.028	0.366	0.337/0.373	0.91/0.94
		RSF/RSF-logit	0.152	0.258	0.189/0.219	0.76/0.83
		RSF/RSF-twang	0.040	0.365	0.352/0.390	0.94/0.95
	IPW	Cox-logit	0.494	0.181	- /0.180	- /0.20
		Cox-twang	0.170	0.302	- /0.227	- /0.79
		RSF-logit	0.269	0.184	- /0.181	- /0.67
		RSF-twang	0.052	0.280	- /0.223	- /0.89
	Naive Cox		0.245	0.164	0.167/0.166	0.71/0.70
	Full Data		0.001	0.061	0.063/0.063	0.96/0.96

Table 5.3: Simulation based on 1000 data sets for Scenarios 3 and 4, each with 1000 observations. True $\beta^* = 0.503$. Red indicates that the working model or the approach is invalid.

Scenario	Estimator	T/C-PS Models	Bias	SD	SE	Coverage
					Model/Boot	Model/Boot
Scenario 3	AIPW	Cox/Cox-logit	0.001	0.082	0.086/0.083	0.96/0.96
		Cox/Cox-twang	0.012	0.077	0.094/0.088	0.98/0.98
		Cox/RSF-logit	0.005	0.085	0.089/0.091	0.96/0.96
		Cox/RSF-twang	0.010	0.078	0.095/0.089	0.98/0.97
		RSF/Cox-logit	0.004	0.071	0.072/0.074	0.95/0.95
		RSF/Cox-twang	0.010	0.073	0.077/0.081	0.96/0.96
		RSF/RSF-logit	0.007	0.072	0.074/0.076	0.95/0.96
		RSF/RSF-twang	0.013	0.075	0.079/0.083	0.96/0.97
	IPW	Cox-logit	0.001	0.081	- /0.082	- /0.95
		Cox-twang	0.021	0.074	- /0.070	- /0.93
		RSF-logit	0.022	0.088	- /0.085	- /0.93
		RSF-twang	0.001	0.081	- /0.073	- /0.93
	Naive Cox		0.518	0.097	0.099/0.099	0.00/0.00
	Full Data		0.001	0.035	0.034/0.034	0.95/0.94
Scenario 4	AIPW	Cox/Cox-logit	0.469	0.125	0.110/0.117	0.02/0.03
		Cox/Cox-twang	0.224	0.157	0.154/0.153	0.70/0.70
		Cox/RSF-logit	0.180	0.138	0.180/0.183	0.83/0.92
		Cox/RSF-twang	0.010	0.210	0.276/0.218	0.97/0.97
		RSF/Cox-logit	0.044	0.085	0.077/0.085	0.88/0.92
		RSF/Cox-twang	0.010	0.113	0.104/0.112	0.94/0.95
		RSF/RSF-logit	0.036	0.116	0.116/0.130	0.94/0.97
		RSF/RSF-twang	0.008	0.156	0.160/0.177	0.94/0.97
	IPW	Cox-logit	0.592	0.119	- /0.112	- /0.00
		Cox-twang	0.308	0.159	- /0.131	- /0.37
		RSF-logit	0.218	0.105	- /0.105	- /0.44
		RSF-twang	0.047	0.152	- /0.127	- /0.89
	Naive Cox		0.431	0.117	0.111/0.112	0.03/0.03
	Full Data		0.001	0.035	0.035/0.035	0.94/0.94

the 2 group survival does not seem to satisfy the PH assumption, which agrees with the findings in Xu and Adak (2002). Moreover, the STAGE variable is not a treatment that can be applied but a risk factor, so we consider the estimation of the non-causal β^*. Figure 5.6 shows the product-limit estimator of the survival functions for the censoring time for each of the 4 covariates, along with the p-values from the log-rank test. Note that the p-value from the log-rank test is not valid when the PH assumption is violated, like in the case of the covariate PS. We can see that the censoring time clearly differs among levels of the covariates. Xu and Adak (2002) also demonstrated that the failure time differs among levels of covariates, confirming the presence of informative censoring and the need to control for it.

We present in Figure 5.7 the β^* estimates for STAGE along with their 95% bootstrap-based confidence intervals. Since we are estimating the non-causal β^*, we considered the estimators that are studied in the Simulation Section of Chapter 3. A complete table of estimates, standard errors, and 95% bootstrap-based confidence intervals is given in Table 5.4. From Figure 5.7, we see that AIPCW-Cox-RSF, AIPCW-RSF-Cox, and AIPCW-RSF-RSF give much smaller β^* estimates than the PL estimator, which could suggest that the β^* estimate using the PL estimator is overestimating the true β^* due to informative censoring. IPCW-RSF involves a non-parametric nuisance estimator, which is typically biased Belloni et al. (2013), potentially resulting in its large estimate here. AIPCW-Cox-Cox also gives larger β^* estimates compared to the other AIPCW estimators, which could signal that none of the underlying distributions for T and C is PH. Therefore, we plot the standardized score residuals of the fitted conditional Cox PH Model for T against time in Figure 5.8, and for C in Figure 5.9. The standardized score converges to a Brownian bridge, so values outside of the blue lines indicate a violation of the PH assumption at the 5% significance level. This violation is observed in almost all the plots, confirming our hypothesis that neither T nor C follows a conditional Cox PH model. Lastly, Figure 5.10 presents the smoothed AIPCW $\beta(t)$ using the AIPCW:RSF-RSF estimator, which demonstrates that the effect of STAGE on survival is rapidly decaying towards zero as time increases. Admittedly, this data set is not collected from a randomized clinical trial, which limits its interpretation. Nevertheless, the comparison demonstrates the robustness of β^* when the PH assumption is violated. It also shows the importance of accounting for the biases resulting from informative censoring as well as the flexibility of the ML methods in practice.

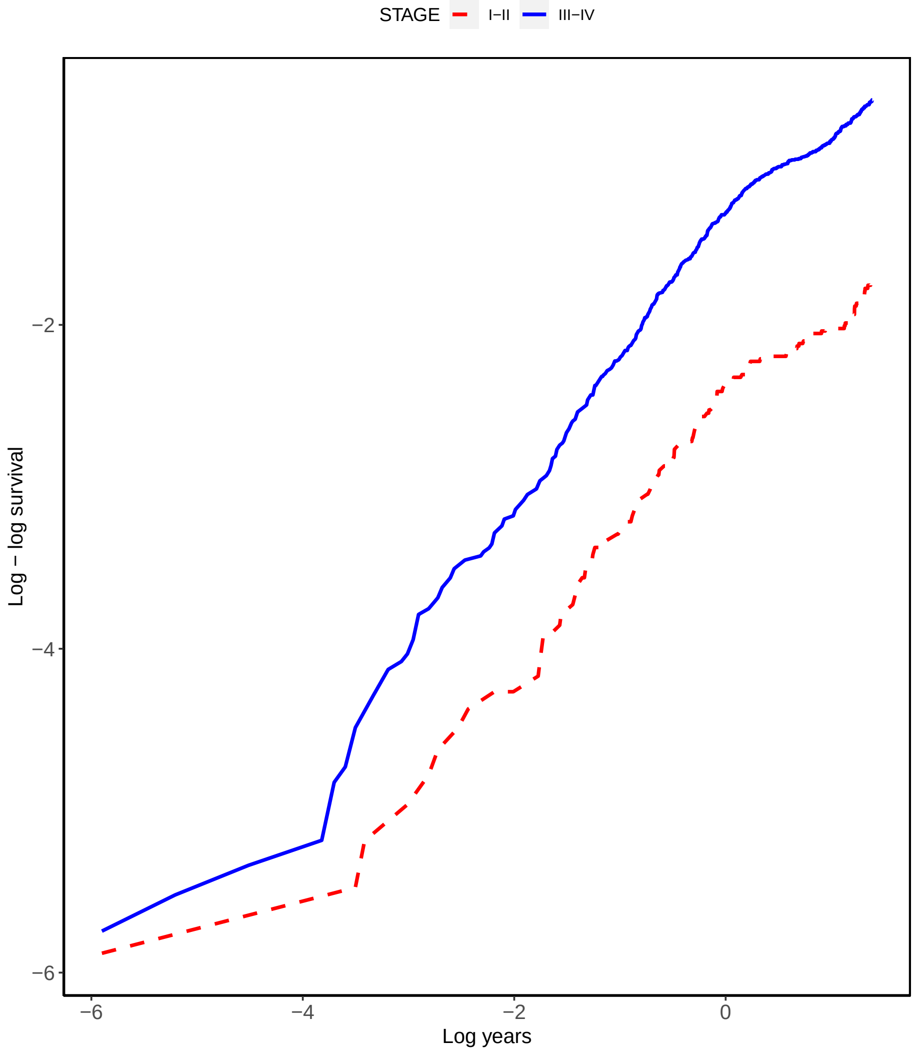


Figure 5.5: Log negative log of the survival curves against log years for STAGE III-IV vs STAGE I-II. Since parallel lines indicates PH, the figure shows that the two groups' survival does not seem to satisfy the PH assumption.

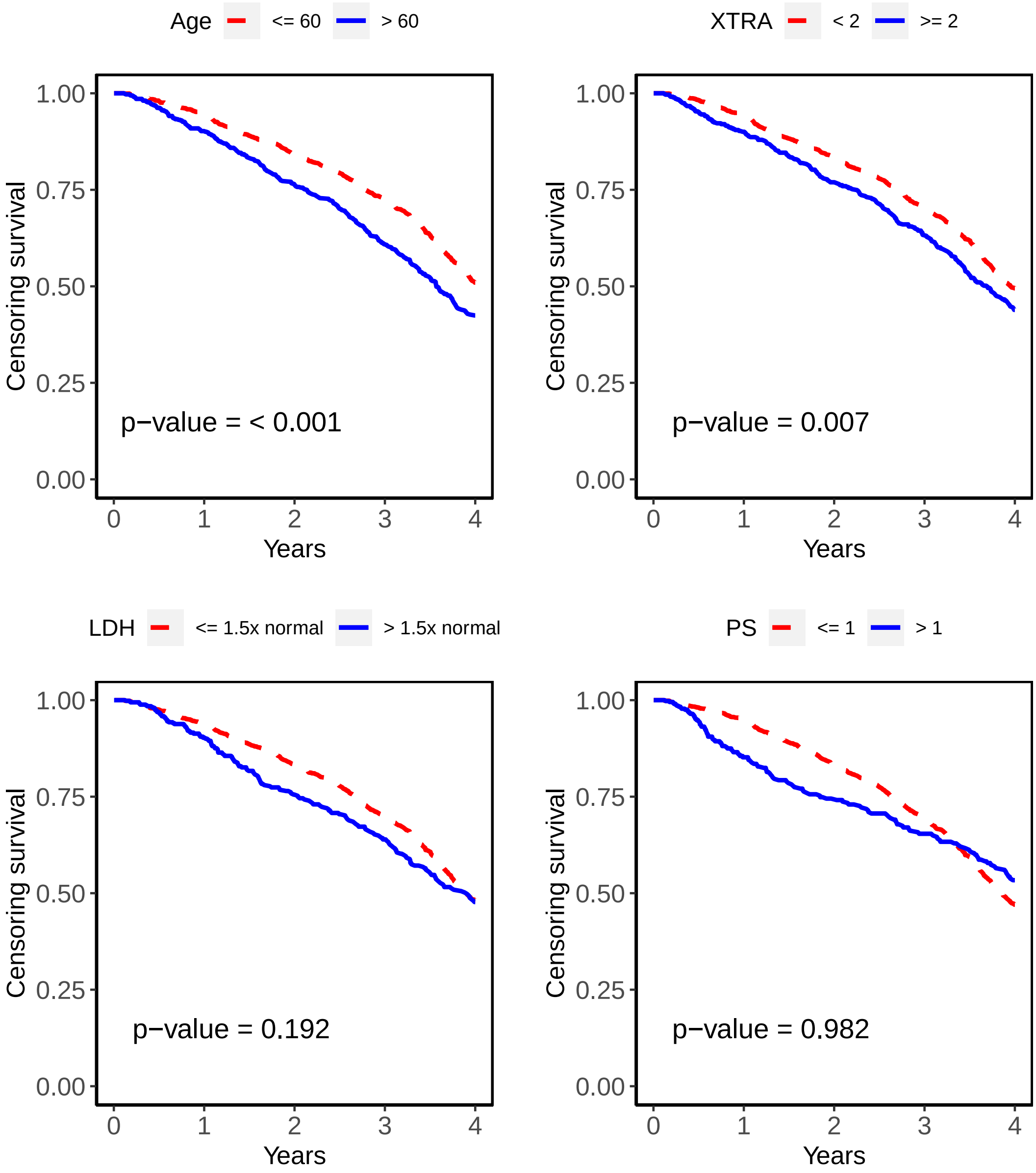


Figure 5.6: Product-limit estimators of the censoring survival curve for each of the 4 covariates. P-value is generated using the log-rank test, which is only valid under the PH assumption.

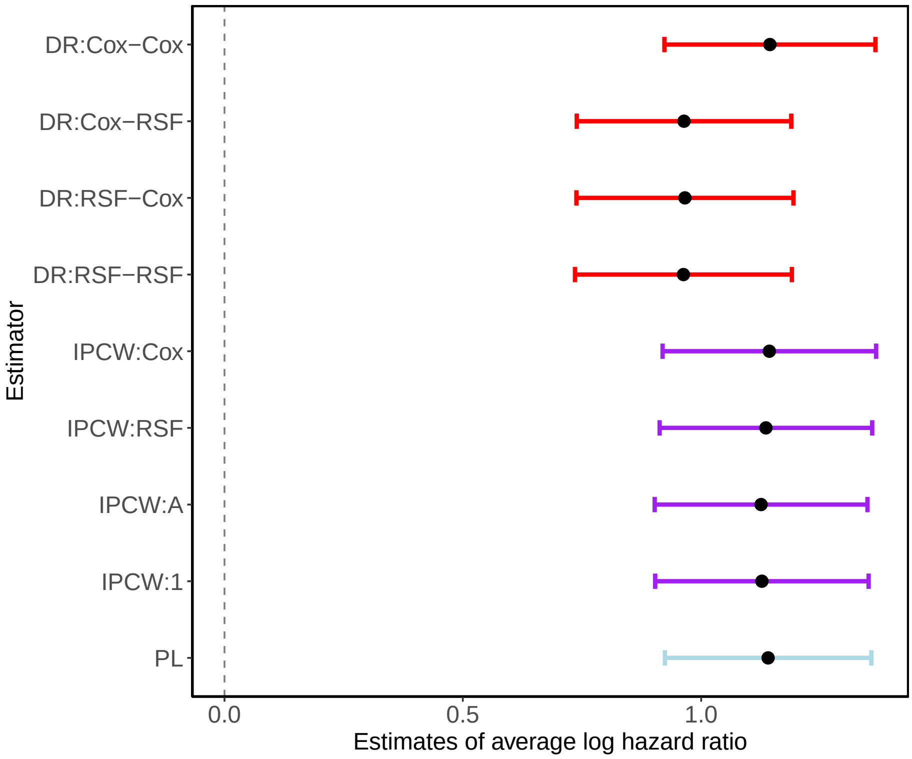


Figure 5.7: Forest plot of the β^* estimates examining the effect of STAGE on survival for patients with non-Hodgkin's lymphoma.

Table 5.4: β^* estimates for the effect of STAGE on overall survival in the Non-Hodgkin's Lymphoma dataset, together with the bootstrapped standard errors and the 95% constructed from it.

Estimator	Estimate	Boot SE	95% Boot CI
DR:Cox-Cox	1.14	0.11	(0.92, 1.37)
DR:Cox-RSF	0.96	0.11	(0.74, 1.19)
DR:RSF-Cox	0.97	0.12	(0.74, 1.19)
DR:RSF-RSF	0.96	0.12	(0.74, 1.19)
IPCW:Cox	1.14	0.11	(0.92, 1.37)
IPCW:RSF	1.14	0.11	(0.91, 1.36)
IPCW:A	1.13	0.11	(0.9, 1.35)
IPCW:1	1.13	0.11	(0.9, 1.35)
PL	1.14	0.11	(0.92, 1.36)

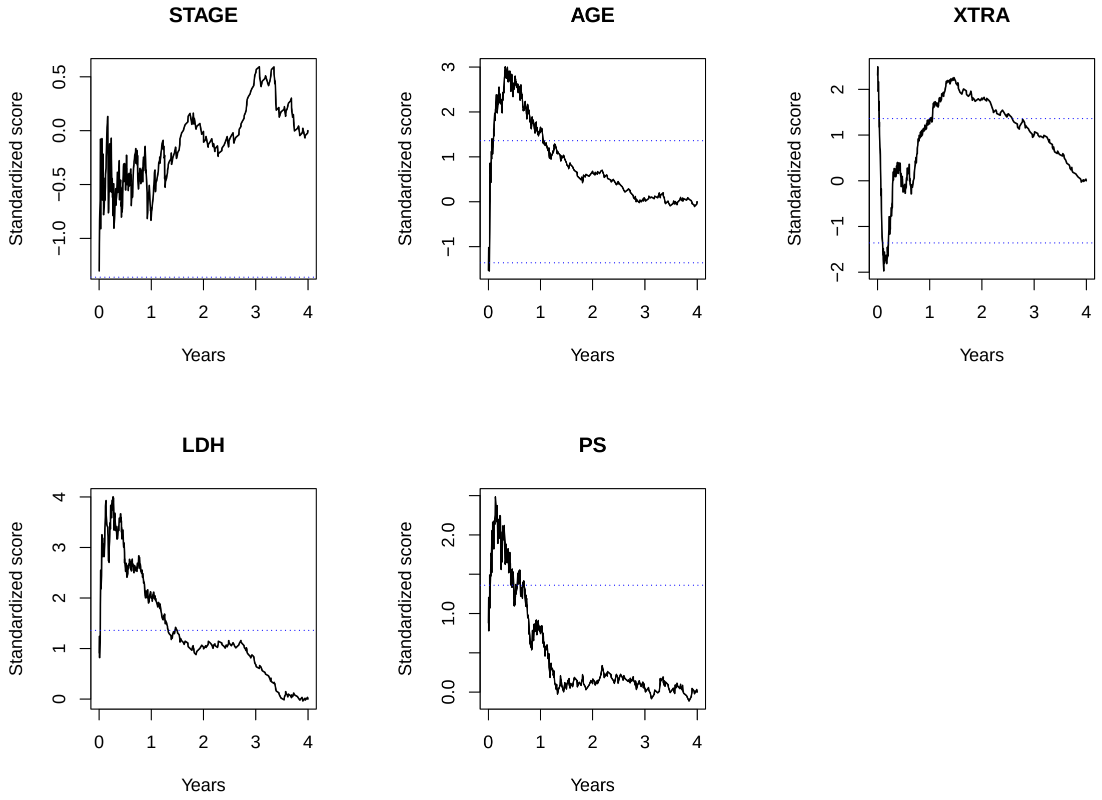


Figure 5.8: Standardized score residuals for the exposure and each of the covariates under the fitted conditional Cox PH model for T given exposure and covariates. Blue lines represent the significance level of the formal test for PH assumption.

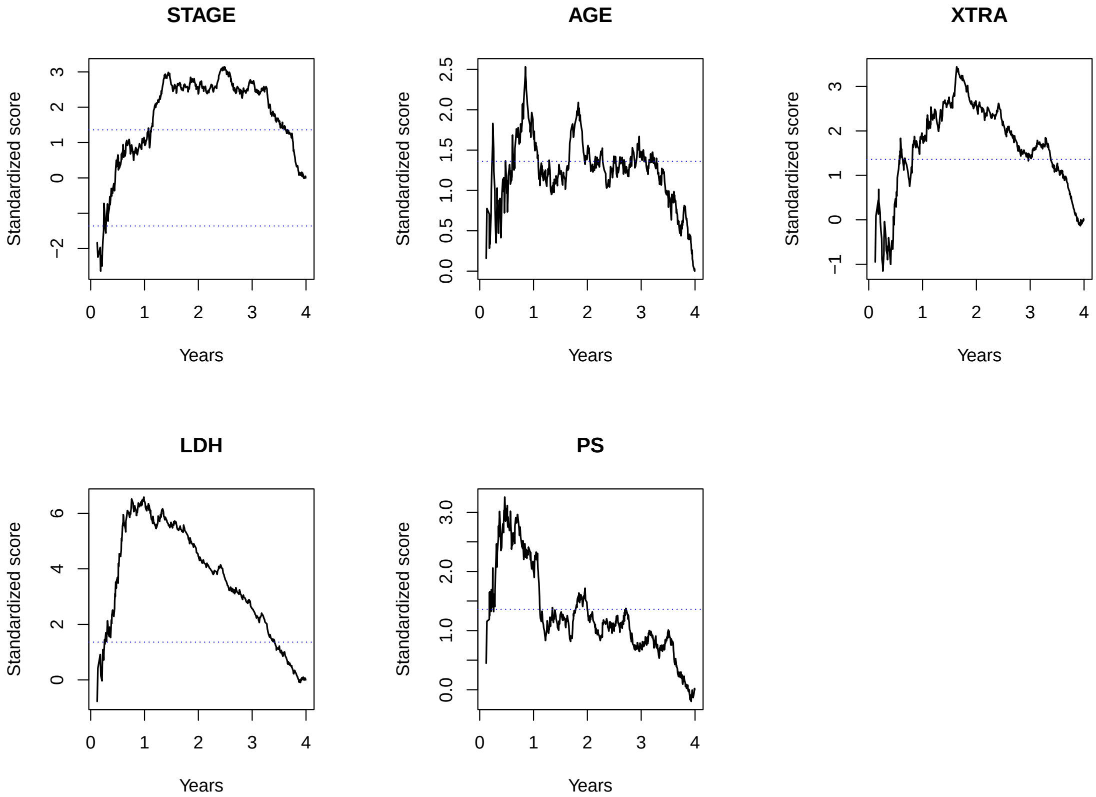


Figure 5.9: Standardized score residuals for the exposure and each of the covariates under the fitted conditional Cox PH model for *C* given exposure and covariates. Blue lines represent the significance level of the formal test for PH assumption.

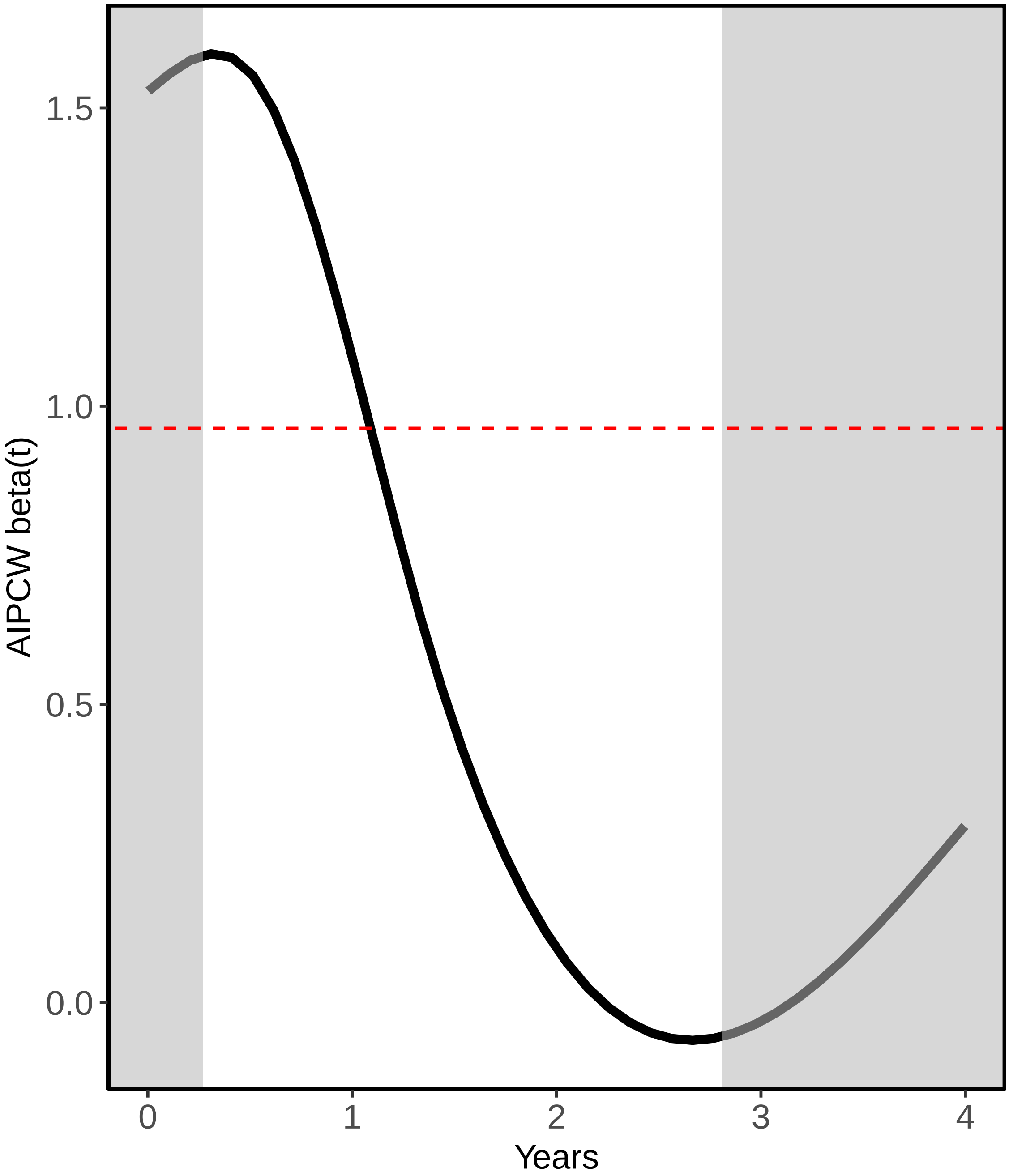


Figure 5.10: Smoothed AIPCW $\beta(t)$ plot for the effect of STAGE based on the AIPCW:RSF-RSF estimator in the Non-Hodgkin's Lymphoma dataset. The red dotted line indicates the AIPCW:RSF-RSF estimate for β^*.

5.10 Discussion

The marginal structural Cox PH model is most commonly used in the causal analysis of two-group survival, which assumes the PH assumption. When the PH assumption is violated, the saturated model is the marginal structural non-PH model (5.3). Under this model, we proposed and studied the estimation of a causal estimand β^* that is a weighted average of the causal log hazard ratio. β^* is also the limit of the oracle log hazard ratio estimator under full data.

By showing that β^* can be estimated using full data estimating functions that are identical to those for the marginal structural Cox model, we proceed to construct IPW, AIPW, and cross-fitted AIPW estimator for β^* using the results we already derived in Chapter 4. The cross-fitted AIPW estimator is again model and rate DR, which allows the use of more flexible ML methods for estimating the nuisance functions. Moreover, we demonstrated that without accounting for either treatment confounding or informative censoring, the naive PL estimator is no longer a weighted average of $\beta(t)$ and can result in bias of much larger magnitude. An AIPW $\beta(t)$ plot that works under observational data is also proposed and studied. The AIPW estimate for β^* converges to the causal log hazard ratio when the PH assumption holds, and by accounting for both treatment confounding and informative censoring, the estimate also carries causal interpretation when the PH assumption is violated. These developments suggest the potential for the causal estimand β^* to be more widely reported.

The causal hazard ratio used in our study has not been without criticism (Hernán, 2010; Martinussen et al., 2020), particularly regarding concerns about the potential imbalance between risk sets post-treatment. However, arguments made by the proponents also should not be overlooked (Prentice and Aragaki, 2022; Ying and Xu, 2023). They argue that although the comparability between the two groups is compromised at time $t > 0$ due to differential survival distributions, the causal hazard ratio still represents the ratio of the logarithm of potential survival functions between the two groups at all times.

This characteristic alone makes it a valid causal estimand with causal interpretability, as it contrasts the functionals of the two potential outcome distributions. Some researchers have explored alternative hazard ratio estimands for time-to-event endpoints. Axelrod and Nevo (2022) conducted sensitivity analysis for the so-called 'causal HR' (Martinussen et al., 2020), which is based on patient groups that would have survived regardless of their treatment assignment. However, the estimand itself is not identifiable, severely limiting its practical use. Another proposal by Vansteelandt et al. (2022) introduced a model-free hazard ratio estimand that simplifies to the hazard ratio under the Cox model. This approach does not rely on the inverse probability of treatment weighting, potentially leading to increased stability. However, it explicitly depends on the model for treatment assignment and requires the restrictive assumption that the cumulative hazard is positive at all times. Therefore, despite the ongoing quest for alternative causal estimands in time-to-event studies, we believe the causal hazard ratio should continue to play a central role due to its inherent simplicity, interpretability, and widespread popularity in the field.

5.11 Acknowledgement

Chapter 5 contains material being prepared for submission and is co-authored with Ronghui Xu. The dissertation author is the primary investigator/author of this paper.

Appendix A

Supplementary Materials for Chapter 2

A.1 Preliminaries

For any convex function $\psi : \mathbb{R}^k \to \mathbb{R}$, define the corresponding Bregman divergence $D_\psi(w', w) = \psi(w') - \psi(w) - \langle \nabla\psi(w), w' - w \rangle$ and its symmetrized version

$$\overline{D}_\psi(w, w') = D_\psi(w, w') + D_\psi(w', w) = \big\langle \nabla\psi(w) - \nabla\psi(w'), w - w' \big\rangle, \ \ w, w' \in \mathbb{R}^k.$$

Let $z = \Sigma^{-1/2} x \in \mathbb{R}^p$ be the standardized vector of covariates such that $\mathbb{E}(zz^{\mathrm{T}}) = I_p$, and define $\mu_k = \sup_{u \in \mathbb{S}^{p-1}} \mathbb{E}|z^{\mathrm{T}} u|^k$ for $k \geq 1$. In particular, $\mu_2 = 1$. For every $\delta \in (0, 1]$, define

$$\eta_\delta = \inf \left\{ \eta > 0 : \sup_{u \in \mathbb{S}^{p-1}} \mathbb{E}\big\{ (z^{\mathrm{T}} u)^2 \mathbb{1}(|z^{\mathrm{T}} u| > \eta) \big\} \leq \delta \right\}.$$

Under Condition (C1), η_δ depends only on δ and υ_1, and the map $\delta \mapsto \eta_\delta$ is non-increasing with $\eta_\delta \downarrow 0$ as $\delta \to 1$. A crude bound for η_δ, as a function of δ, is $\eta_\delta \leq (\mu_4/\delta)^{1/2}$.

In Lemmas 5 and 6 below, we provide a lower bound on the symmetrized Bregman divergence and an upper bound on the score, respectively. The former is a direct consequence of Lemmas C.3 and C.4 in Sun et al. (2020) with slight modifications, and the latter combines Lemmas C.5 and C.6 in Sun et al. (2020) with $\delta = 1$. For the shifted Huber loss $\widetilde{\mathcal{L}}(\cdot)$, note that

$$\bar{D}_{\widetilde{\mathcal{L}}}(\beta,\beta^*) = \big\langle \nabla\widehat{\mathcal{L}}_{1,\kappa}(\beta) - \nabla\widehat{\mathcal{L}}_{1,\kappa}(\beta^*), \beta-\beta^* \big\rangle.$$

Moreover, define the ℓ_1-cone

$$\Lambda = \big\{\beta \in \mathbb{R}^p : \|\beta-\beta^*\|_1 \le 4s^{1/2}\|\beta-\beta^*\|_\Sigma\big\}.$$

Lemma 5. *Let $\kappa, r > 0$ satisfy $\kappa \ge 4\max(\eta_{0.25}r, \sigma)$.*

(i) Condition (C1) ensures that, with probability at least $1-e^{-u}$,

$$\bar{D}_{\widetilde{\mathcal{L}}}(\beta,\beta^*) \ge \frac{1}{4}\|\beta-\beta^*\|_\Sigma^2 \ \textit{ holds uniformly over } \ \beta \in \Theta(r)$$

as long as $n \gtrsim (\kappa/r)^2(p+u)$.

(ii) Condition (C2) ensures that, with probability at least $1-e^{-u}$,

$$\bar{D}_{\widetilde{\mathcal{L}}}(\beta,\beta^*) \ge \frac{1}{4}\|\beta-\beta^*\|_\Sigma^2 \ \textit{ holds uniformly over } \ \beta \in \Theta(r)\cap\Lambda \tag{A.1}$$

as long as $n \gtrsim (\kappa/r)^2(s\log p+u)$.

Proof. Without loss of generality, assume $I_1 = \{1,\dots,n\}$. It suffices to prove (A.1) under Condition (C2). Following the proof of Lemma C.4 in Sun et al. (2020), the key is to upper bound the expected value of the maximum $\|(1/n)\sum_{i=1}^n e_i x_i\|_\infty$, where $e_1,\dots,e_n$ are independent Rademacher random variables. Let $\mathbb{E}_e$ be the expectation with respect to $e_1,\dots,e_n$ conditional on the remaining variables. By Hoeffding's moment inequality,

$$\mathbb{E}_e\left\|\frac{1}{n}\sum_{i=1}^n e_i x_i\right\|_\infty \le \max_{1\le j\le p}\left(\frac{1}{n}\sum_{i=1}^n x_{ij}^2\right)^{1/2}\sqrt{\frac{2\log(2p)}{n}} \le B\sqrt{\frac{2\log(2p)}{n}},$$

which in turns implies $\mathbb{E}\|(1/n)\sum_{i=1}^n e_i x_i\|_\infty \le B\sqrt{2\log(2p)/n}$. Keep the rest of the proof the same proves the claimed bound. □

Consider the gradient $\nabla\widehat{\mathcal{L}}_\tau(\cdot)$ evaluated at β^*, namely,

$$\nabla\widehat{\mathcal{L}}_\tau(\beta^*) = -\frac{1}{N}\sum_{i=1}^N \psi_\tau(\varepsilon_i)x_i,$$

where $\psi_\tau(u) = \ell_\tau'(u)$. The following lemma provides high probability bounds on both ℓ_2- and ℓ_∞-norms of $\nabla\widehat{\mathcal{L}}_\tau(\beta^*)$. Recall that $\Omega = \Sigma^{-1}$.

Lemma 6. *Let $u > 0$ and write $\mathcal{L}_\tau(\cdot) = \mathbb{E}\widehat{\mathcal{L}}_\tau(\cdot)$.*

(i) Condition (C1) ensures that, with probability at least $1 - e^{-u}$,

$$\|\nabla\widehat{\mathcal{L}}_\tau(\beta^*) - \nabla\mathcal{L}_\tau(\beta^*)\|_{\Sigma^{-1}} \le C_0\Big\{\sigma\sqrt{(p+u)/N} + \tau(p+u)/N\Big\}, \qquad (A.2)$$

where $C_0 > 0$ is a constant depending only on υ_1. Moreover, $\|\nabla\mathcal{L}_\tau(\beta^*)\|_\Omega \le \sigma^2/\tau$.

(ii) Condition (C2) ensures that, with probability at least $1 - e^{-u}$,

$$\|\nabla\widehat{\mathcal{L}}_\tau(\beta^*) - \nabla\mathcal{L}_\tau(\beta^*)\|_\infty \le \sigma\sigma_u\sqrt{\frac{2\{\log(2p)+u\}}{N}} + \frac{B\tau}{3}\frac{\log(2p)+u}{N}. \qquad (A.3)$$

Proof. The bound (A.2) is an immediate consequence of Lemma C.3 in Sun et al. (2020). It suffices to prove (A.3) under Condition (C2). Note that

$$\|\nabla\widehat{\mathcal{L}}_\tau(\beta^*) - \nabla\mathcal{L}_\tau(\beta^*)\|_\infty = \max_{1\le j\le p}\left|\frac{1}{N}\sum_{i=1}^N (1-\mathbb{E})\xi_i x_{ij}\right|,$$

where $\xi_i := \psi_\tau(\varepsilon_i)$ satisfy $|\xi_i| \le \tau$ and $\mathbb{E}(\xi_i^2|x_i) \le \mathbb{E}(\varepsilon_i^2|x_i) \le \sigma^2$. For any $1 \le j \le p$ and $z \ge 0$, applying Bernstein's inequality yields that with probability at least $1 - 2e^{-z}$,

$$\left|\frac{1}{N}\sum_{i=1}^N (1-\mathbb{E})\xi_i x_{ij}\right| \le \sigma_{jj}^{1/2}\sigma\sqrt{\frac{2z}{N}} + \frac{B\tau}{3}\frac{z}{N}.$$

Taking $z = \log(2p) + u$, the claimed bound (A.3) then follows from the union bound. □

A.2 Proof of Main Results

A.2.1 Proof of Theorem 1

PROOF OF (2.5). For simplicity, we write $\widetilde{\beta} = \widetilde{\beta}^{(1)}$, which minimizes the shifted Huber loss $\widetilde{\mathcal{L}}(\cdot)$ and thus satisfies the first-order condition $\nabla\widetilde{\mathcal{L}}(\widetilde{\beta}) = 0$. Throughout the proof we assume the event $\mathcal{E}_0(r_0) \cap \mathcal{E}_*(r_*)$ occurs. In view of Lemma 5, we consider a local region $\Theta(r_{\mathrm{loc}})$ with $r_{\mathrm{loc}} = \kappa/(4\eta_{0.25})$, and define an intermediate estimator $\widetilde{\beta}_c = (1-c)\beta^* + c\widetilde{\beta}$, where

$$c := \sup\left\{u \in [0,1] : (1-u)\beta^* + u\widetilde{\beta} \in \Theta(r_{\mathrm{loc}})\right\} \begin{cases} = 1 & \text{if } \widetilde{\beta} \in \Theta(r_{\mathrm{loc}}), \\ \in (0,1) & \text{otherwise.} \end{cases}$$

By construction, $\widetilde{\beta}_c \in \Theta(r_{\mathrm{loc}})$. In particular, if $\widetilde{\beta} \notin \Theta(r_{\mathrm{loc}})$, we must have $\widetilde{\beta}_c$ lying on the boundary of $\Theta(r_{\mathrm{loc}})$, i.e. $\|\widetilde{\beta}_c - \beta^*\|_\Sigma = r_{\mathrm{loc}}$.

Applying Lemma C.1 in Sun et al. (2020), we see that the three points $\widetilde{\beta}, \widetilde{\beta}_c$ and β^* satisfy $\bar{D}_{\widetilde{\mathcal{L}}}(\widetilde{\beta}_c, \beta^*) \le c\bar{D}_{\widetilde{\mathcal{L}}}(\widetilde{\beta}, \beta^*)$, where $\bar{D}_{\widetilde{\mathcal{L}}}(\beta, \beta^*) = \langle \nabla\widetilde{\mathcal{L}}(\beta) - \nabla\widetilde{\mathcal{L}}(\beta^*), \beta - \beta^* \rangle = \langle \nabla\widehat{\mathcal{L}}_{1,\kappa}(\beta) - \nabla\widehat{\mathcal{L}}_{1,\kappa}(\beta^*), \beta - \beta^* \rangle$. Together with the first-order condition $\nabla\widetilde{\mathcal{L}}(\widetilde{\beta}) = 0$, this implies

$$\bar{D}_{\widetilde{\mathcal{L}}}(\widetilde{\beta}_c, \beta^*) \le -c\left\langle \nabla\widetilde{\mathcal{L}}(\beta^*), \widetilde{\beta} - \beta^* \right\rangle \le \|\nabla\widetilde{\mathcal{L}}(\beta^*)\|_\Omega \cdot \|\widetilde{\beta}_c - \beta^*\|_\Sigma. \tag{A.4}$$

For the left-hand side of (A.4), applying Lemma 5 with $r = r_{\mathrm{loc}}$ and the fact $\widetilde{\beta}_c \in \Theta(r_{\mathrm{loc}})$ yields that with probability at least $1 - e^{-u}$,

$$\bar{D}_{\widetilde{\mathcal{L}}}(\widetilde{\beta}_c, \beta^*) \ge \frac{1}{4}\|\widetilde{\beta}_c - \beta^*\|_\Sigma^2 \tag{A.5}$$

as long as $n \gtrsim p + u$.

To bound the right-hand side of (A.4), we define vector-valued random processes

$$\begin{cases} \Delta_1(\beta) = \Sigma^{-1/2}\{\nabla\widehat{\mathcal{L}}_{1,\kappa}(\beta) - \nabla\widehat{\mathcal{L}}_{1,\kappa}(\beta^*)\} - \Sigma^{1/2}(\beta-\beta^*), \\ \Delta(\beta) = \Sigma^{-1/2}\{\nabla\widehat{\mathcal{L}}_{\tau}(\beta) - \nabla\widehat{\mathcal{L}}_{\tau}(\beta^*)\} - \Sigma^{1/2}(\beta-\beta^*), \end{cases} \tag{A.6}$$

Let $0 < r_0 \le \sigma$. Following the proof of Theorem B.1 in the supplement of Sun et al. (2020), it can be similarly shown that, with probability at least $1-2e^{-u}$,

$$\sup_{\beta\in\Theta(r_0)} \|\Delta_1(\beta)\|_2 \le C_1\left(\sqrt{\frac{p+u}{n}} + \frac{\sigma^2}{\kappa^2}\right) r_0 \text{ and } \sup_{\beta\in\Theta(r_0)} \|\Delta(\beta)\|_2 \le C_1\left(\sqrt{\frac{p+u}{N}} + \frac{\sigma^2}{\tau^2}\right) r_0 \tag{A.7}$$

as long as $n \gtrsim p+u$, where $C_1 > 0$ is a constant depending only on υ_1. Recall that $\tau \ge \kappa \asymp \sigma\sqrt{n/(p+u)}$. Conditioned on event $\mathcal{E}_0(r_0) \cap \mathcal{E}_*(r_*)$, it follows that

$$\begin{aligned} \|\nabla\widetilde{\mathcal{L}}(\beta^*)\|_\Omega &= \|\Delta(\widetilde{\beta}^{(0)}) - \Delta_1(\widetilde{\beta}^{(0)}) + \Sigma^{-1/2}\nabla\widehat{\mathcal{L}}_\tau(\beta^*)\|_2 \\ &\le \|\Delta(\widetilde{\beta}^{(0)}) - \Delta_1(\widetilde{\beta}^{(0)})\|_2 + \|\nabla\widehat{\mathcal{L}}_\tau(\beta^*)\|_\Omega \\ &\le C_2 r_0\sqrt{\frac{p+u}{n}} + r_*. \end{aligned} \tag{A.8}$$

Together, the bounds (A.4), (A.5) and (A.8) imply that, conditioning on $\mathcal{E}_0(r_0) \cap \mathcal{E}_*(r_*)$,

$$\|\widetilde{\beta}_c - \beta^*\|_\Sigma \le 4\|\nabla\widetilde{\mathcal{L}}(\beta^*)\|_\Omega \le 4\left(C_2 r_0\sqrt{\frac{p+u}{n}} + r^*\right) \tag{A.9}$$

with probability at least $1-3e^{-u}$. Provided that the sample size is sufficiently large—$n \gtrsim p+u$, the right-hand side of the above inequality is strictly less than $r_{\mathrm{loc}} = \kappa/(4\eta_{0.25})$ with $\kappa \asymp \sigma\sqrt{n/(p+u)}$. As a result, the intermediate estimator $\widetilde{\beta}_c$ falls into the interior of $\Theta(r_{\mathrm{loc}})$ with high probability conditioned on $\mathcal{E}_0(r_0) \cap \mathcal{E}_*(r_*)$. Via proof by contradiction, we must have $\widetilde{\beta} \in \Theta(r_{\mathrm{loc}})$ and hence $\widetilde{\beta} = \widetilde{\beta}_c$; otherwise if $\widetilde{\beta} \notin \Theta(r_{\mathrm{loc}})$, we have demonstrated that $\widetilde{\beta}_c$ must lie on the boundary of $\Theta(r_{\mathrm{loc}})$, which is a contradiction. Consequently, the bound (A.9) also applies to $\widetilde{\beta}$, as claimed.

PROOF OF (2.6). To establish the Bahadur representation, note that the random process $\Delta_1(\cdot)$ defined in (A.6) can be written as $\Delta_1(\beta) = \Sigma^{-1/2}\{\nabla\widetilde{\mathcal{L}}(\beta) - \nabla\widetilde{\mathcal{L}}(\beta^*)\} - \Sigma^{1/2}(\beta - \beta^*)$. Moreover, note that

$$\nabla\widetilde{\mathcal{L}}(\beta^*) = \nabla\widehat{\mathcal{L}}_{1,\kappa}(\beta^*) - \nabla\widehat{\mathcal{L}}_{1,\kappa}(\widetilde{\beta}^{(0)}) + \nabla\widehat{\mathcal{L}}_{\tau}(\widetilde{\beta}^{(0)}) - \nabla\widehat{\mathcal{L}}_{\tau}(\beta^*) + \nabla\widehat{\mathcal{L}}_{\tau}(\beta^*),$$

which in turn implies

$$\|\nabla\widetilde{\mathcal{L}}(\beta^*) - \nabla\widehat{\mathcal{L}}_{\tau}(\beta^*)\|_{\Omega} \le \|\Delta_1(\widetilde{\beta}^{(0)})\|_2 + \|\Delta(\widetilde{\beta}^{(0)})\|_2.$$

Recall that $\nabla\widetilde{\mathcal{L}}(\widetilde{\beta}) = 0$, and by (A.9), $\|\widetilde{\beta} - \beta^*\|_{\Sigma} \le r_1 := 4C_2 r_0\sqrt{(p+u)/n} + 4r_*$ with high probability conditioned on $\mathcal{E}_0(r_0) \cap \mathcal{E}_*(r_*)$. For $r_0 \ge 8r_*$, we have $r_1 \le r_0/2 + r_0/2 = r_0$ as long as $n \gtrsim p+u$, and hence $\widetilde{\beta} \in \Theta(r_0)$. Applying the bounds in (A.7) again, we obtain that conditioned on $\mathcal{E}_0(r_0) \cap \mathcal{E}_*(r_*)$,

$$\begin{aligned}
&\|\Sigma^{1/2}(\widetilde{\beta} - \beta^*) + \Sigma^{-1/2}\nabla\widehat{\mathcal{L}}_{\tau}(\beta^*)\|_2 \\
&= \|\Delta_1(\widetilde{\beta}) + \Sigma^{-1/2}\nabla\widetilde{\mathcal{L}}(\beta^*) - \Sigma^{-1/2}\nabla\widehat{\mathcal{L}}_{\tau}(\beta^*)\|_2 \\
&\le \|\Delta_1(\widetilde{\beta})\|_2 + \|\Delta_1(\widetilde{\beta}^{(0)})\|_2 + \|\Delta(\widetilde{\beta}^{(0)})\|_2 \\
&\le 2\sup_{\beta\in\Theta(r_0)} \|\Delta_1(\beta)\|_2 + \sup_{\beta\in\Theta(r_0)} \|\Delta(\beta)\|_2 \\
&\lesssim \sqrt{\frac{p+u}{n}} \cdot r_0
\end{aligned}$$

with probability at least $1 - 3e^{-u}$. This completes the proof. □

A.2.2 Proof of Theorem 2

Given a sequence of iterates $\{\widetilde{\beta}^{(t)}\}_{t=0,1,\ldots,T}$, we define "good" events

$$\mathcal{E}_t(r_t) = \{\widetilde{\beta}^{(t)} \in \Theta(r_t)\}, \ \ t = 0, \ldots, T,$$

for some sequence of radii $r_0 \geq r_1 \geq \cdots \geq r_T > 0$ to be determined. Examine the proof of Theorem 1, we see that the statistical properties of $\widetilde{\beta}^{(t)}$ depends on both first-order and second-order information of the loss function $\widetilde{\mathcal{L}}^{(t)}(\cdot)$, namely, the ℓ_2-norm of the gradient $\nabla\widetilde{\mathcal{L}}^{(t)}(\beta^*)$ and the (symmetrized) Bregman divergence of $\widetilde{\mathcal{L}}^{(t)}(\cdot)$. For the former, we have

$$\nabla\widetilde{\mathcal{L}}^{(t)}(\beta^*) = \nabla\widehat{\mathcal{L}}_{1,\kappa}(\beta^*) - \nabla\widehat{\mathcal{L}}_{1,\kappa}(\widetilde{\beta}^{(t-1)}) + \nabla\widehat{\mathcal{L}}_{\tau}(\widetilde{\beta}^{(t-1)}). \tag{A.10}$$

Let $\Delta_1(\cdot)$ and $\Delta(\cdot)$ be the random processes defined in (A.6), and observe that $\Sigma^{-1/2}\nabla\widetilde{\mathcal{L}}^{(t)}(\beta^*) = \Delta(\widetilde{\beta}^{(t-1)}) - \Delta_1(\widetilde{\beta}^{(t-1)}) + \Sigma^{-1/2}\nabla\widehat{\mathcal{L}}_{\tau}(\beta^*)$. By the triangle inequality,

$$\|\nabla\widetilde{\mathcal{L}}^{(t)}(\beta^*)\|_{\Omega} \leq \|\Delta(\widetilde{\beta}^{(t-1)})\|_2 + \|\Delta_1(\widetilde{\beta}^{(t-1)})\|_2 + \|\nabla\widehat{\mathcal{L}}_{\tau}(\beta^*)\|_{\Omega}. \tag{A.11}$$

On the other hand, note that the shifted Huber losses $\widetilde{\mathcal{L}}^{(t)}(\cdot)$ have the same Bregman divergence, denoted by

$$\bar{D}(\beta_1, \beta_2) = \langle \nabla\widetilde{\mathcal{L}}^{(t)}(\beta_1) - \nabla\widetilde{\mathcal{L}}^{(t)}(\beta_2), \beta_1 - \beta_2 \rangle = \langle \nabla\widetilde{\mathcal{L}}_{1,\kappa}(\beta_1) - \nabla\widetilde{\mathcal{L}}_{1,\kappa}(\beta_2), \beta_1 - \beta_2 \rangle.$$

Define the local radius $r_{\text{loc}} = \kappa/(4\eta_{0.25})$. Then, applying Lemma 5 with $r = r_{\text{loc}}$ yields that, with probability at least $1 - e^{-u}$,

$$\bar{D}(\beta, \beta^*) \geq \frac{1}{4}\|\beta - \beta^*\|_{\Sigma}^2 \tag{A.12}$$

holds uniformly over $\beta \in \Theta(r_{\text{loc}})$. Let $\mathcal{E}_{\text{lsc}}$ be the event that the local strong convexity (A.12) holds.

With the above preparations, we are ready to extend the argument in the proof of Theorem 1 to deal with $\widetilde{\beta}^{(t)}$ sequentially. At each iteration, we construct an intermediate estimator $\widetilde{\beta}^{(t)}_{\rm imd}$—a convex combination of $\widetilde{\beta}^{(t)}$ and β^*—which falls in $\Theta(r_{\rm loc})$ and satisfies

$$\bar{D}(\widetilde{\beta}^{(t)}_{\rm imd}, \beta^*) \leq \|\nabla \widetilde{\mathcal{L}}^{(t)}(\beta^*)\|_\Omega \cdot \|\widetilde{\beta}^{(t)}_{\rm imd} - \beta^*\|_\Sigma.$$

If event $\mathcal{E}_*(r_*) \cap \mathcal{E}_{\rm lsc}$ occurs, the bounds (A.11) and (A.12) imply

$$\|\widetilde{\beta}^{(t)}_{\rm imd} - \beta^*\|_\Sigma \leq 4\big\{\|\Delta_1(\widetilde{\beta}^{(t-1)})\|_2 + \|\Delta(\widetilde{\beta}^{(t-1)})\|_2\big\} + 4r_*. \tag{A.13}$$

Moreover, it follows from (A.10) and the first-order condition $\nabla \widetilde{\mathcal{L}}^{(t)}(\widetilde{\beta}^{(t)}) = 0$ that

$$\begin{aligned}
&\|\Sigma^{1/2}(\widetilde{\beta}^{(t)} - \beta^*) + \Sigma^{-1/2}\nabla\widehat{\mathcal{L}}_\tau(\beta^*)\|_2 \\
&= \|\Sigma^{-1/2}\{\nabla\widetilde{\mathcal{L}}^{(t)}(\widetilde{\beta}^{(t)}) - \nabla\widetilde{\mathcal{L}}^{(t)}(\beta^*)\} - \Sigma^{1/2}(\widetilde{\beta}^{(t)} - \beta^*) + \Sigma^{-1/2}\{\nabla\widetilde{\mathcal{L}}^{(t)}(\beta^*) - \nabla\widehat{\mathcal{L}}_\tau(\beta^*)\}\|_2 \\
&\leq \|\Delta_1(\widetilde{\beta}^{(t)})\|_2 + \|\Delta_1(\widetilde{\beta}^{(t-1)})\|_2 + \|\Delta(\widetilde{\beta}^{(t-1)})\|_2.
\end{aligned} \tag{A.14}$$

In view of the bounds in (A.7), for every $0 < r \leq \sigma$ we define the event

$$\mathcal{F}(r) = \left\{ \sup_{\beta \in \Theta(r)} \big\{\|\Delta_1(\beta)\|_2 + \|\Delta(\beta)\|_2\big\} \leq \gamma(u) \cdot r \right\} \tag{A.15}$$

with $\gamma(u) = C\sqrt{(p+u)/n}$ for some $C > 0$, which satisfies $\mathbb{P}\{\mathcal{F}(r)\} \geq 1 - 2e^{-u}$.

Let $8r^* \leq r_0 \leq \sigma$. In the following, we assume the event $\mathcal{E}_0(r_0) \cap \mathcal{E}_*(r_*) \cap \mathcal{E}_{\rm lsc}$ occurs, and deal with $\{(\widetilde{\beta}^{(t)}_{\rm imd}, \widetilde{\beta}^{(t)}), t = 1, 2, \ldots, T\}$ sequentially. At iteration 1, it follows from (A.13) that, conditioned on $\mathcal{F}(r_0)$,

$$\|\widetilde{\beta}^{(1)}_{\rm imd} - \beta^*\|_\Sigma \leq r_1 := 4\gamma(u) \cdot r_0 + 4r_*.$$

Provided that $n \gtrsim p + u$, we have $4\gamma(u) \leq 1/2 < 1$ and $r_1 \leq r_0 < r_{\rm loc} = \kappa/(4\eta_{0.25})$, so that

$\widetilde{\beta}^{(1)}_{\mathrm{imd}} \in \Theta(r_1) \subseteq \mathrm{int}(\Theta(r_{\mathrm{loc}}))$. Via proof by contradiction, we must have $\widetilde{\beta}^{(1)} = \widetilde{\beta}^{(1)}_{\mathrm{imd}} \in \Theta(r_{\mathrm{loc}})$, which in turns certifies event $\mathcal{E}(r_1)$. Combining this with (A.14), we see that conditioned on $\mathcal{F}(r_0)$, the event $\mathcal{E}_1(r_1)$ mush happen and hence

$$\begin{cases} \|\, \widetilde{\beta}^{(1)} - \beta^* \,\|_\Sigma \le r_1 = 4\gamma(u) \cdot r_0 + 4r_* \le r_0, \\ \|\, \widetilde{\beta}^{(1)} - \beta^* + \Sigma^{-1}\nabla \widehat{\mathcal{L}}_\tau(\beta^*) \,\|_\Sigma \le 2\gamma(u) \cdot r_0. \end{cases}$$

Now assume that for some $t \ge 1$, $\widetilde{\beta}^{(t)} \in \Theta(r_t)$ with $r_t = 4\gamma(u) \cdot r_{t-1} + 4r_* \le r_{t-1} < r_{\mathrm{loc}}$. At $(t+1)$-th iteration, applying (A.13) again yields that, conditioned on event $\mathcal{E}_t(r_t) \cap \mathcal{F}(r_t)$,

$$\|\widetilde{\beta}^{(t+1)}_{\mathrm{imd}} - \beta^*\|_\Sigma \le r_{t+1} := 4\gamma(u) \cdot r_t + 4r_*.$$

By induction, $r_t \le r_{t-1} < r_{\mathrm{loc}}$ so that $r_{t+1} \le 4\gamma(u) \cdot r_{t-1} + 4r_* = r_t < r_{\mathrm{loc}}$. This implies that $\widetilde{\beta}^{(t+1)}_{\mathrm{imd}}$ falls into the interior of $\Theta(r_{\mathrm{loc}})$, which enforces $\widetilde{\beta}^{(t+1)} = \widetilde{\beta}^{(t+1)}_{\mathrm{imd}} \in \Theta(r_{t+1})$ and thus certifies event $\mathcal{E}_{t+1}(r_{t+1})$. Combining this with the bound (A.14), we find that

$$\begin{cases} \|\, \widetilde{\beta}^{(t+1)} - \beta^* \,\|_\Sigma \le r_{t+1} = 4\gamma(u) \cdot r_t + 4r_* \le r_t, \\ \|\, \widetilde{\beta}^{(t+1)} - \beta^* + \Sigma^{-1}\nabla \widehat{\mathcal{L}}_\tau(\beta^*) \,\|_\Sigma \le 2\gamma(u) \cdot r_t. \end{cases}$$

Repeat the above argument until we obtain $\widetilde{\beta}^{(T)}$. We have shown that conditioned on $\mathcal{E}_*(r_*) \cap \mathcal{E}_{\mathrm{lsc}} \cap \mathcal{E}_{t-1}(r_{t-1}) \cap \mathcal{F}(r_{t-1})$ for every $0 \le t \le T-1$, the event $\mathcal{E}_t(r_t)$ must occur. Therefore, conditioned on $\mathcal{E}_*(r_*) \cap \mathcal{E}_{\mathrm{lsc}} \cap \mathcal{E}_0(r_0) \cap \{\cap_{t=0}^{T-1} \mathcal{F}(r_t)\}$, $\widetilde{\beta}^{(T)}$ satisfies the bounds

$$\begin{cases} \|\, \widetilde{\beta}^{(T)} - \beta^* \,\|_\Sigma \le r_T = 4\gamma(u) \cdot r_{T-1} + 4r_*, \\ \|\, \widetilde{\beta}^{(T)} - \beta^* + \Sigma^{-1}\nabla \widehat{\mathcal{L}}_\tau(\beta^*) \,\|_\Sigma \le 2\gamma(u) \cdot r_{T-1}. \end{cases} \tag{A.16}$$

Observe that $r_t = \{4\gamma(u)\}^t r_0 + \frac{1-\{4\gamma(u)\}^t}{1-4\gamma(u)} 4r_*$ for $t = 1, \ldots, T$. We choose T to be the smallest integer such that $\{4\gamma(u)\}^{T-1} r_0 \le r_*$, that is, $T = \lceil \log(r_0/r_*)/\log(1/\{4\gamma(u)\}) \rceil + 1$.

Consequently, the bounds in (A.16) become

$$\begin{cases} \|\widetilde{\beta}^{(T)} - \beta^* \|_\Sigma \le \{\gamma(u) + \frac{1}{1-4\gamma(u)}\} 4 r_* \le \{4\gamma(u) + 8\} r_*, \\ \|\widetilde{\beta}^{(T)} - \beta^* + \Sigma^{-1} \nabla \widehat{\mathcal{L}}_\tau(\beta^*) \|_\Sigma \le 18\gamma(u) \cdot r_*. \end{cases} \tag{A.17}$$

Finally, it suffices to show that the event $\mathcal{E}_{\rm lsc} \cap \{\cap_{t=0}^{T-1} \mathcal{F}(r_t)\}$ occurs with high probability. Recall from (A.12) and (A.15) that $\mathbb{P}(\mathcal{E}_{\rm lsc}) \ge 1 - e^{-u}$ and $\mathbb{P}\{\mathcal{F}(r_t)\} \ge 1 - 2e^{-u}$ for every $t = 0, 1, \ldots, T-1$. The claimed result then follows from (A.17) and the union bound. □

A.2.3 Proof of Theorem 3

Let $u > 0$. Applying Theorem B.1 in Sun et al. (2020) with a robustification parameter $\kappa \asymp \sigma\sqrt{n/(p+u)}$ yields that with probability at least $1 - 2e^{-u}$, $\|\widetilde{\beta}^{(0)} - \beta^*\|_\Sigma \le r_0 \asymp \sigma\sqrt{(p+u)/n}$ as long as $n \gtrsim p + u$. For event $\mathcal{E}^*(r^*)$ defined in (2.4), we take $r^* \asymp \sigma\sqrt{(p+u)/N} + \tau(p+u)/N + \sigma^2/\tau$ in Lemma 6 and obtain that $\mathbb{P}\{\mathcal{E}^*(r^*)\} \ge 1 - e^{-u}$. Putting together the pieces, we conclude that event $\mathcal{E}_0(r_0) \cap \mathcal{E}_*(r_*)$ occurs with probability at least $1 - 3e^{-u}$, provided that $n \gtrsim p + u$.

Set $u = \log n + \log_2 m$. Since $\tau \asymp \sigma\sqrt{N/(p + \log n + \log_2 m)}$, we see that

$$r_0 \asymp \sigma\sqrt{\frac{p + \log n + \log_2 m}{n}} \quad \text{and} \quad r_* \asymp \sigma\sqrt{\frac{p + \log n + \log_2 m}{N}},$$

and hence $r_0/r_* \asymp \sqrt{m}$. Finally, applying Theorem 2 yields the claimed bounds (2.8) and (2.9). □

A.2.4 Proof of Theorem 4

For simplicity, we write $q = p + \log n + \log_2 m$ throughout the proof. For every vector $a \in \mathbb{R}^p$, define $S_a = N^{-1/2} \sum_{i=1}^N \xi_i w_i$ and $S_a^0 = S_a - \mathbb{E} S_a$, where $\xi_i = \psi_\tau(\varepsilon_i)$ and $w_i = a^{\rm T} \Sigma^{-1} x_i$. Under the moment condition $\mathbb{E}(|\varepsilon|^{2+\delta} | x) \le v_{2+\delta}$, using Markov's inequality yields $|\mathbb{E}(\xi_i|x_i)| \le \tau^{-1-\delta}\mathbb{E}(|\varepsilon_i|^{1+\delta}|x_i) \le v_{2+\delta}\tau^{-1-\delta}$. Hence, $|\mathbb{E}(\xi_i w_i)| \le v_{2+\delta}\|a\|_\Omega \cdot \tau^{-1-\delta}$ and $|\mathbb{E}S_a| \le v_{2+\delta}\|a\|_\Omega \cdot N^{1/2}\tau^{-1-\delta}$.

With the above preparations, we are ready to prove the normal approximation for $\widetilde{\beta}$. Note that

$$\begin{aligned}
&|N^{1/2}a^{\mathrm{T}}(\widetilde{\beta}-\beta^*) - S_a^0| \\
&\le N^{1/2}\left|\left\langle \Sigma^{-1/2}a, \Sigma^{1/2}(\widetilde{\beta}-\beta^*) - \Sigma^{-1/2}\frac{1}{N}\sum_{i=1}^N \psi_\tau(\varepsilon_i)x_i \right\rangle\right| + |\mathbb{E}S_a| \\
&\le N^{1/2}\|a\|_\Omega \cdot \left\| \widetilde{\beta}-\beta^* - \Sigma^{-1}\frac{1}{N}\sum_{i=1}^N \psi_\tau(\varepsilon_i)x_i \right\|_\Sigma + v_{2+\delta}\|a\|_\Omega \cdot N^{1/2}\tau^{-1-\delta}.
\end{aligned}$$

Applying (2.9) in Theorem 3, we find that with probability at least $1-Cn^{-1}$,

$$|N^{1/2}a^{\mathrm{T}}(\widetilde{\beta}-\beta^*) - S_a^0| \le C_1\|a\|_\Omega \cdot \big(\sigma q n^{-1/2} + N^{1/2}v_{2+\delta}\tau^{-1-\delta}\big), \tag{A.18}$$

where $C_1 > 0$ is a constant independent of (N,n,p).

For the centered partial sum S_a^0, it follows from the Berry-Esseen inequality (see, e.g. Theorem 2.1 in Chen and Shao (2001)) that

$$\sup_{t\in\mathbb{R}} \big|\mathbb{P}\big\{S_a^0 \le \mathrm{var}(S_a^0)^{1/2}t\big\} - \Phi(t)\big| \le 4.1\frac{\mathbb{E}|\xi w - \mathbb{E}(\xi w)|^{2+\delta}}{\mathrm{var}(\xi w)^{1+\delta/2}N^{\delta/2}}, \tag{A.19}$$

where $\xi = \psi_\tau(\varepsilon)$ and $w = a^{\mathrm{T}}\Sigma^{-1}x$. Recall that $\tau \asymp \sigma\sqrt{N/q}$, and write $\sigma_{\tau,a}^2 = \mathbb{E}(\xi w)^2$. By Proposition A.2 in Zhou et al. (2018), $|\mathbb{E}(\xi^2|x) - \sigma^2| \le 2\delta^{-1}v_{2+\delta}\tau^{-\delta} \asymp \delta^{-1}v_{2+\delta}\sigma^{-\delta}(q/N)^{\delta/2}$, and hence

$$\big|\sigma_{\tau,a}^2/(\sigma\|a\|_\Omega)^2 - 1\big| \lesssim \frac{v_{2+\delta}}{\delta\sigma^{2+\delta}}\left(\frac{q}{N}\right)^{\delta/2}. \tag{A.20}$$

Moreover, $\mathbb{E}|\xi w|^{2+\delta} \le \mathbb{E}|\varepsilon w|^{2+\delta} \le \mu_{2+\delta}\|a\|_\Omega^{2+\delta}v_{2+\delta}$, where $\mu_{2+\delta} := \sup_{u\in\mathbb{S}^{p-1}}\mathbb{E}|z^{\mathrm{T}}u|^{2+\delta}$

depends only on υ_1 under Condition (C1). Substituting these bounds into (A.19) yields

$$\sup_{t\in\mathbb{R}} \big|\mathbb{P}\{S_a^0 \le \mathrm{var}(S_a^0)^{1/2} t\} - \Phi(t)\big| \le C_2 \frac{v_{2+\delta}}{\sigma^{2+\delta} N^{\delta/2}}, \tag{A.21}$$

provided that $N \gtrsim q$. For the variance term, the bound $|\mathbb{E}(\xi|x)| \le \sigma^2\tau^{-1}$ guarantees that

$$\mathbb{E}(\xi w)^2 \ge \mathrm{var}(S_a^0) = \mathbb{E}(\xi w)^2 - (\mathbb{E}\xi w)^2 \ge \mathbb{E}(\xi w)^2 - (\sigma \|a\|_\Omega)^2 \cdot \sigma^2 \tau^{-2}.$$

Combined with (A.20), this implies $|\mathrm{var}(S_a^0)/\sigma_{\tau,a}^2 - 1| \lesssim \sigma^2\tau^{-2}$, from which it follows that

$$\sup_{t\in\mathbb{R}} \big|\Phi(t/\mathrm{var}(S_a^0)^{1/2}) - \Phi(t/\sigma_{\tau,a})\big| \le C_3 \frac{\sigma^2}{\tau^2}. \tag{A.22}$$

Let $G \sim \mathcal{N}(0,1)$ and $t \in \mathbb{R}$. Combining the bounds (A.18), (A.21) and (A.22), we obtain

$$\begin{aligned}
&\mathbb{P}\{N^{1/2} a^{\mathrm{T}}(\widetilde{\beta} - \beta^*) \le t\} \\
&\le \mathbb{P}\{S_a^0 \le x + C_1 \|a\|_\Omega \cdot (\sigma q n^{-1/2} + N^{1/2} v_{2+\delta} \tau^{-1-\delta})\} + C n^{-1} \\
&\le \mathbb{P}\{\mathrm{var}(S_a^0)^{1/2} G \le t + C_1 \|a\|_\Omega \cdot (\sigma q n^{-1/2} + N^{1/2} v_{2+\delta} \tau^{-1-\delta})\} + C n^{-1} + C_2 \frac{v_{2+\delta}}{\sigma^{2+\delta} N^{\delta/2}} \\
&\le \mathbb{P}\{\sigma_{\tau,a} G \le t + C_1 \|a\|_\Omega \cdot (\sigma q n^{-1/2} + N^{1/2} v_{2+\delta} \tau^{-1-\delta})\} + C_2 \frac{v_{2+\delta}}{\sigma^{2+\delta} N^{\delta/2}} + C_3 \frac{\sigma^2}{\tau^2} \\
&\le \mathbb{P}(\sigma_{\tau,a} G \le t) + C n^{-1} + C_1 (2\pi)^{-1/2} (q n^{-1/2} + N^{1/2} v_{2+\delta} \sigma^{-1} \tau^{-1-\delta}) + C_2 \frac{v_{2+\delta}}{\sigma^{2+\delta} N^{\delta/2}} \\
&+ C_3 \frac{\sigma^2}{\tau^2}.
\end{aligned}$$

A similar argument leads to a series of reverse inequalities, and thus completes the proof. □

A.2.5 Proof of Theorem 5

As before, we assume without loss of generality that $I_1 = \{1, \ldots, n\}$. Write $\widetilde{\beta} = \widetilde{\beta}^{(1)}$ for simplicity, and let $g = \widetilde{\beta} - \beta^*$ be the error vector. By the first-order optimality condition, there exists a subgradient $g \in \partial\|\widetilde{\beta}\|_1$ such that $g^{\mathrm{T}}\widetilde{\beta} = \|\widetilde{\beta}\|_1$ and $\nabla\widetilde{\mathcal{L}}(\widetilde{\beta}) + \lambda \cdot g = 0$. Moreover, the convexity of $\widetilde{\mathcal{L}}(\cdot)$ implies

$$0 \le \bar{D}_{\widetilde{\mathcal{L}}}(\widetilde{\beta}, \beta^*) = g^{\mathrm{T}}\big\{\nabla\widetilde{\mathcal{L}}(\widetilde{\beta}) - \nabla\widetilde{\mathcal{L}}(\beta^*)\big\} = -\lambda \cdot g^{\mathrm{T}}g - g^{\mathrm{T}}\nabla\widetilde{\mathcal{L}}(\beta^*).$$

Recall the true active set $\mathcal{S} = \mathrm{supp}(\beta^*) \subseteq \{1, \ldots, p\}$, we have

$$-g^{\mathrm{T}}g \le \|\beta^*\|_1 - \|\widetilde{\beta}\|_1 = \|\beta^*_{\mathcal{S}}\|_1 - \|g_{\mathcal{S}^{\mathrm{c}}}\|_1 - \|g_{\mathcal{S}} + \beta_{\mathcal{S}}\|_1 \le \|g_{\mathcal{S}}\|_1 - \|g_{\mathcal{S}^{\mathrm{c}}}\|_1.$$

Together, the above two displays yield

$$0 \le \bar{D}_{\widetilde{\mathcal{L}}}(\widetilde{\beta}, \beta^*) \le \lambda\big(\|g_{\mathcal{S}}\|_1 - \|g_{\mathcal{S}^{\mathrm{c}}}\|_1\big) - g^{\mathrm{T}}\nabla\widetilde{\mathcal{L}}(\beta^*). \tag{A.23}$$

To deal with $\nabla\widetilde{\mathcal{L}}(\beta^*) = \nabla\widehat{\mathcal{L}}_{1,\kappa}(\beta^*) - \nabla\widehat{\mathcal{L}}_{1,\kappa}(\widetilde{\beta}^{(0)}) + \nabla\widehat{\mathcal{L}}_{\tau}(\widetilde{\beta}^{(0)})$, we define random processes

$$\widehat{D}_1(\beta) = \nabla\widehat{\mathcal{L}}_{1,\kappa}(\beta) - \nabla\widehat{\mathcal{L}}_{1,\kappa}(\beta^*), \quad \widehat{D}(\beta) = \nabla\widehat{\mathcal{L}}_{\tau}(\beta) - \nabla\widehat{\mathcal{L}}_{\tau}(\beta^*),$$

and write $D_1(\beta) = \mathbb{E}\widehat{D}_1(\beta)$ and $D(\beta) = \mathbb{E}\widehat{D}(\beta)$. The gradient $\nabla\widetilde{\mathcal{L}}(\beta^*)$ can thus be written as

$$\begin{aligned}\big\{\widehat{D}(\beta) - D(\beta)\big\}\Big|_{\beta=\widetilde{\beta}^{(0)}} + \big\{D_1(\beta) - \widehat{D}_1(\beta)\big\}\Big|_{\beta=\widetilde{\beta}^{(0)}} + \nabla\widehat{\mathcal{L}}_{\tau}(\beta^*) - \nabla\mathcal{L}_{\tau}(\beta^*)\\ + \big\{D(\beta) - D_1(\beta)\big\}\Big|_{\beta=\widetilde{\beta}^{(0)}} + \nabla\mathcal{L}_{\tau}(\beta^*).\end{aligned}$$

For any $r > 0$, define

$$\Delta_1(r) = \sup_{\beta\in\Theta(r)\cap\Lambda} \|\widehat{D}_1(\beta) - D_1(\beta)\|_\infty, \ \Delta(r) = \sup_{\beta\in\Theta(r)\cap\Lambda} \|\widehat{D}(\beta) - D(\beta)\|_\infty, \tag{A.24}$$

$$\delta(r) = \sup_{\beta\in\Theta(r)} \|D_1(\beta) - D(\beta)\|_\Omega \ \text{ and } \ b^* = \|\nabla\mathcal{L}_{\tau}(\beta^*)\|_\Omega. \tag{A.25}$$

The quantity b^* can be viewed as the robustification bias and by Lemma 6, $b^* \le \sigma^2\tau^{-1}$.

Back to the right-hand of (A.23), conditioning on the event $\mathcal{E}_0(r_0) \cap \mathcal{E}_*(\lambda_*)$, it follows from Hölder's inequality that

$$|g^{\mathrm{T}} \nabla \widetilde{\mathcal{L}}(\beta^*)| \leq \big\{\Delta(r_0) + \Delta_1(r_0) + \lambda_*\big\} \|g\|_1 + \big\{\delta(r_0) + b^*\big\} \|g\|_{\Sigma}. \tag{A.26}$$

Let $\lambda = 2.5(\lambda_* + \rho)$ for some $\rho > 0$. Provided that

$$\rho \geq \max\big[\Delta(r_0) + \Delta_1(r_0), s^{-1/2}\big\{\delta(r_0) + b^*\big\}\big], \tag{A.27}$$

we have $|g^{\mathrm{T}} \nabla \widetilde{\mathcal{L}}(\beta^*)| \leq 0.4\lambda \|g\|_1 + 0.4 s^{1/2} \lambda \|g\|_{\Sigma}$. Combined with (A.23), this yields $0 \leq 1.4\|g_{\mathcal{S}}\|_1 - 0.6\|g_{\mathcal{S}^c}\|_1 + 0.4 s^{1/2}\|g\|_{\Sigma}$. Consequently, $\|g\|_1 \leq (10/3)\|g_{\mathcal{S}}\|_1 + (2/3) s^{1/2} \|g\|_{\Sigma}$ $\leq 4 s^{1/2} \|g\|_{\Sigma}$, and hence $\widetilde{\beta} \in \Lambda$. Throughout the rest of the proof, we assume that the constraint (A.27) holds.

Next, we apply Lemma 5 to bound the left-hand side of (A.23) from below. As in the proof of Theorem 1, we set $r_{\mathrm{loc}} = \kappa/(4\eta_{0.25})$ and define $\widetilde{\beta}_c = (1-c)\beta^* + c\widetilde{\beta}$, where $c = \sup\{u \in [0,1] : (1-u)\beta^* + u\widetilde{\beta} \in \Theta(r_{\mathrm{loc}})\}$. The same argument therein implies $\bar{D}_{\widetilde{\mathcal{L}}}(\widetilde{\beta}_c, \beta^*) \leq c\bar{D}_{\widetilde{\mathcal{L}}}(\widetilde{\beta}, \beta^*)$. Recall that conditioned on $\mathcal{E}_0(r_0) \cap \mathcal{E}_*(\lambda_*)$, $\widetilde{\beta}$ falls in the ℓ_1-cone Λ and thus so does $\widetilde{\beta}_c$. Moreover, $\widetilde{\beta}_c \in \Theta(r_{\mathrm{loc}})$ by construction. Then it follows from Lemma 5 that, with probability at least $1 - e^{-u}$,

$$\bar{D}_{\widetilde{\mathcal{L}}}(\widetilde{\beta}_c, \beta^*) \geq \frac{1}{4}\|\widetilde{\beta}_c - \beta^*\|_{\Sigma}^2$$

as long as $n \gtrsim s \log p + u$. Combining this with (A.23), (A.26) and (A.27), we obtain that

$$\frac{1}{4}\|\widetilde{\beta}_c - \beta^*\|_{\Sigma}^2 \leq c\lambda\big(1.4\|g_{\mathcal{S}}\|_1 + 0.4 s^{1/2}\|g\|_{\Sigma}\big) \leq 1.8 s^{1/2} \lambda \|\widetilde{\beta}_c - \beta^*\|_{\Sigma}.$$

Canceling $\|\widetilde{\beta}_c - \beta^*\|_{\Sigma}$ on both sides yields

$$\|\widetilde{\beta}_c - \beta^*\|_{\Sigma} \leq 7.2 s^{1/2} \lambda. \tag{A.28}$$

Provided that $\kappa > 28.8\eta_{0.25}s^{1/2}\lambda$, the right-hand side is strictly less than $r_{\rm loc}$. Via proof by contradiction, we must have $\widetilde{\beta} = \widetilde{\beta}_c \in \Theta(r_{\rm loc})$, and hence the bound (A.28) also applies to $\widetilde{\beta}$.

It remains to choose ρ properly so that the constraint (A.27) holds with high probability. Recall from Lemma 6 that $b^* \le \sigma^2\tau^{-1}$. The following two lemmas provide upper bounds on the suprema $\Delta(r_0)$, $\Delta_1(r_0)$ and $\delta(r_0)$ defined in (A.24) and (A.25).

Lemma 7. *Assume Condition (C2) holds. Then, for any $r, u > 0$,*

$$\Delta(r) \le C_1 B^2 r \left\{ \sqrt{\frac{s\log(2p)}{N}} + s^{1/2}\frac{\log(2p)+u}{N} \right\} + C_2(\sigma_u \mu_4)^{1/2} r \sqrt{\frac{\log(2p)+u}{N}}$$

with probability at least $1-e^{-u}$, where $C_1, C_2 > 0$ are absolute constants. The same bound, with N replaced by n, holds for $\Delta_1(r)$.

Lemma 8. *Condition (C2) guarantees $\delta(r) \le \kappa^{-2} r(\sigma^2 + \mu_4 r^2/3)$ for any $r > 0$.*

Let $0 < r_0 \lesssim \sigma$ and set $\delta = 2e^{-u}$, so that $\log p + u \asymp \log(p/\delta)$. Suppose the sample size per machine satisfies $n \gtrsim s\log(p/\delta)$. Then, in view of Lemmas 7 and 8, a sufficiently large ρ, which is of order

$$\rho \asymp \max\left\{ r_0\sqrt{\frac{s\log(p/\delta)}{n}}, s^{-1/2}\sigma^2(\kappa^{-2}r_0 + \tau^{-1}) \right\},$$

guarantees that (A.27) holds with probability at least $1-\delta/2$. With this choice of ρ, we see that the right-hand of (A.28) is strictly less than $r_{\rm loc}$ as long as $\kappa \gtrsim s^{1/2}\{\lambda^* + r_0\sqrt{s\log(p/\delta)/n}\} + \sigma^2(\kappa^{-2}r_0 + \tau^{-1})$. Since $\kappa \asymp \sigma\sqrt{n/\log(p/\delta)}$, this holds trivially under the assumed sample size scaling, and thus completes the proof. □

We end this subsection with the proofs of Lemmas 7 and 8.

Proof of Lemma 7

For any r_1, r_2, define the ℓ_1/ℓ_2-ball $\mathbb{B}(r_1, r_2)$. Consider the change of variable $v = \beta - \beta^*$, so that $v \in \mathbb{B}(4s^{1/2}r, r)$ for $\beta \in \Theta(r) \cap \Lambda$. It follows that

$$\begin{aligned}
&\sup_{\beta \in \Theta(r) \cap \Lambda} \|\widehat{D}(\beta) - D(\beta)\|_\infty \\
&\leq \max_{1 \leq j \leq p} \sup_{v \in \mathbb{B}(4s^{1/2}r, r)} \left| \frac{1}{N} \sum_{i=1}^{N} (1 - \mathbb{E}) \underbrace{\{\psi_\tau(\varepsilon_i - x_i^{\mathrm{T}} v) - \psi_\tau(\varepsilon_i)\} x_{ij}}_{=:\phi_{ij}(v)} \right| = \max_{1 \leq j \leq p} \Phi_j,
\end{aligned}$$

where $\Phi_j := \sup_{v \in \mathbb{B}(4s^{1/2}r, r)} |(1/N) \sum_{i=1}^{N} (1 - \mathbb{E}) \phi_{ij}(v)|$ and $\psi_\tau(u) = \mathrm{sign}(u) \min(|u|, \tau)$. By the Lipschitz continuity of $\psi_\tau(\cdot)$, $\sup_{v \in \mathbb{B}(4s^{1/2}r, r)} |\phi_{ij}(v)| \leq \sup_{v \in \mathbb{B}(4s^{1/2}r, r)} |x_i^{\mathrm{T}} v| \cdot |x_{ij}| \leq 4B^2 s^{1/2} r$ and, for each $v \in \mathbb{B}(4s^{1/2}r, r)$,

$$\mathbb{E}\phi_{ij}^2(v) = \mathbb{E}\{x_{ij}^2 (x_i^{\mathrm{T}} v)^2\} \leq \left(\mathbb{E}x_{ij}^4\right)^{1/2} \left\{\mathbb{E}(x_i^{\mathrm{T}} v)^4\right\}^{1/2} \leq \sigma_{jj} \mu_4 \cdot r^2.$$

We then apply Bousquet's version of Talagrand's inequality (Bousquet, 2003) and obtain that, for any $z > 0$,

$$\begin{aligned}
\Phi_j &\leq \mathbb{E}\Phi_j + \sup_{v \in \mathbb{B}(4s^{1/2}r, r)} \left\{\mathbb{E}\phi_{ij}^2(v)\right\}^{1/2} \sqrt{\frac{2z}{N}} + 4\sqrt{\mathbb{E}\Phi_j \cdot B^2 s^{1/2} r \frac{z}{N}} + (4/3) B^2 s^{1/2} r \frac{z}{N} \qquad \text{(A.29)} \\
&\leq \mathbb{E}\Phi_j + (2\sigma_{jj}\mu_4)^{1/2} r \sqrt{\frac{z}{N}} + 4\sqrt{\mathbb{E}\Phi_j \cdot B^2 s^{1/2} r \frac{z}{N}} + (4/3) B^2 s^{1/2} r \frac{z}{N}
\end{aligned}$$

with probability at least $1 - 2e^{-z}$. For the expected value $\mathbb{E}\Phi_j$, by Rademacher symmetrization we have

$$\mathbb{E}\Phi_j \leq 2\mathbb{E} \sup_{v \in \mathbb{B}(4s^{1/2}r, r)} \left| \frac{1}{N} \sum_{i=1}^{N} e_i \phi_{ij}(v) \right| = 2\mathbb{E}\left\{ \mathbb{E}_e \sup_{v \in \mathbb{B}(4s^{1/2}r, r)} \left| \frac{1}{N} \sum_{i=1}^{N} e_i \phi_{ij}(v) \right| \right\},$$

where $e_1, \ldots, e_N$ are independent Rademacher random variables. For each i, write $\phi_{ij}(v) = \phi_j(x_i^{\mathrm{T}} v)$, where $\phi_j(\cdot)$ is such that $\phi_j(0) = 0$ and $|\phi_j(u) - \phi_j(v)| \leq |x_{ij}| \cdot |u - v| \leq B|u - v|$. It thus follows from Talagrand's contraction principle that

$$\mathbb{E}_e \sup_{v \in \mathbb{B}(4s^{1/2}r,r)} \left| \frac{1}{N} \sum_{i=1}^N e_i \phi_{ij}(v) \right| \leq 2B \cdot \mathbb{E}_e \sup_{v \in \mathbb{B}(4s^{1/2}r,r)} \left| \frac{1}{N} \sum_{i=1}^N e_i x_i^{\mathrm{T}} v \right| \leq 8Bs^{1/2} r \cdot \mathbb{E}_e \left\| \frac{1}{N} \sum_{i=1}^N e_i x_i \right\|_\infty .$$

Again, applying Hoeffding's moment inequality yields $\mathbb{E}_e \|(1/N) \sum_{i=1}^N e_i x_i\|_\infty \leq B\sqrt{2\log(2p)/N}$. Putting together the pieces, we conclude that, for $j = 1, \ldots, p$,

$$\mathbb{E}\Phi_j \leq 16B^2 r \sqrt{\frac{2s\log(2p)}{N}}.$$

Finally, taking $z = \log(2p) + u$ in (A.29), the claimed bound follows from the union bound. □

Proof of Lemma 8

Let $\mathcal{L}_\tau(\beta) = \mathbb{E}\widehat{\mathcal{L}}_\tau(\beta)$ be the population loss, so that

$$D_1(\beta) = \nabla \mathcal{L}_\kappa(\beta) - \nabla \mathcal{L}_\kappa(\beta^*) \quad \text{and} \quad D(\beta) = \nabla \mathcal{L}_\tau(\beta) - \nabla \mathcal{L}_\tau(\beta^*).$$

Starting with $D_1(\beta)$, consider the change of variable $v = \Sigma^{1/2}(\beta - \beta^*)$. Then, by the mean value theorem for vector-valued functions,

$$\begin{aligned} &\Sigma^{-1/2} D_1(\beta) - \Sigma^{1/2}(\beta - \beta^*) \\ &= \Sigma^{-1/2} \mathbb{E} \int_0^1 \nabla^2 \mathcal{L}_\tau\big((1-t)\beta^* + t\beta\big) \mathrm{d}t \, \Sigma^{-1/2} \cdot v - v \\ &= -\int_0^1 \mathbb{E}\big\{\mathbb{P}\big(|\varepsilon - tz^{\mathrm{T}} v| > \kappa | x\big) zz^{\mathrm{T}}\big\} \mathrm{d}t \cdot v. \end{aligned}$$

Similarly, it can be obtained that

$$\Sigma^{-1/2} D(\beta) - \Sigma^{1/2}(\beta - \beta^*) = -\int_0^1 \mathbb{E}\big\{\mathbb{P}\big(|\varepsilon - tz^{\mathrm{T}} v| > \tau | x\big) zz^{\mathrm{T}}\big\} \mathrm{d}t \cdot v.$$

Recall that $\tau \geq \kappa > 0$. We have

$$\Sigma^{-1/2}\{D_1(\beta) - D(\beta)\} = -\int_0^1 \mathbb{E}\{\mathbb{P}(\kappa < |\varepsilon - tz^{\mathrm{T}}v| \leq \tau|x)zz^{\mathrm{T}}\}\mathrm{d}t \cdot v$$

By Markov's inequality and the fact that $\mathbb{E}(\varepsilon|x) = 0$, $\mathbb{P}(|\varepsilon - tz^{\mathrm{T}}v| > \kappa|x) \leq \kappa^{-2}\{\mathbb{E}(\varepsilon^2|x) + t^2(z^{\mathrm{T}}v)^2\} \leq \kappa^{-2}\{\sigma^2 + t^2(z^{\mathrm{T}}v)^2\}$. Substituting this into the above bound yields

$$\sup_{\beta \in \Theta(r)} \|D_1(\beta) - D(\beta)\|_\Omega \leq \kappa^{-2}r\big(\sigma^2 + \mu_4 r^2/3\big),$$

as desired. □

A.2.6 Proof of Theorem 6

The proof will be carried out conditioning on the "good event" $\mathcal{E}_0(r_0) \cap \mathcal{E}_*(\lambda_*)$ for some predetermined $0 < r_0, \lambda_* \lesssim \sigma$. Given $\delta \in (0,1)$, let the robustification parameters satisfy $\tau \geq \kappa \asymp \sigma\sqrt{n/\log(p/\delta)}$. Theorem 5 implies that the first iterate $\widetilde{\beta}^{(1)} \in \arg\min_{\beta \in \mathbb{R}^p}\{\widetilde{\mathcal{L}}^{(1)}(\beta) + \lambda_1\|\beta\|_1\}$ with

$$\lambda_1 = 2.5(\lambda_* + \rho_1) \quad \text{and} \quad \rho_1 \asymp \max\left\{ r_0\sqrt{\frac{s\log(p/\delta)}{n}}, s^{-1/2}\sigma^2\tau^{-1} \right\}$$

satisfies the cone property $\widetilde{\beta}^{(1)} \in \Lambda$ and the error bound

$$\|\widetilde{\beta}^{(1)} - \beta^*\|_\Sigma \leq C_1 s\sqrt{\log(p/\delta)/n} \cdot r_0 + C_2(\sigma^2\tau^{-1} + s^{1/2}\lambda_*) =: r_1 \tag{A.30}$$

with probability at least $1-\delta$. In (A.30), we set $\alpha = \alpha(s,p,n,\delta) = C_1 s\sqrt{\log(p/\delta)/n}$ and $r_* = C_2(\sigma^2\tau^{-1} + s^{1/2}\lambda_*)$, so that $r_1 = \alpha r_0 + r_*$. Provided the sample size per machine is sufficiently large, namely, $n \gtrsim s^2\log(p/\delta)$, the contraction factor α is strictly less than 1, and hence the initial estimation error r_0 is reduced by a factor of α after one round of communication.

For $t = 2,3,\dots,T$, define the events $\mathcal{E}_t(r_t) = \{\widetilde{\beta}^{(t)} \in \Theta(r_t) \cap \Lambda\}$ and radius parameters

$$r_t = \alpha r_{t-1} + r_* = \alpha^2 r_{t-2} + (1+\alpha) r_* = \cdots = \alpha^t r_0 + \frac{1-\alpha^t}{1-\alpha} r_*.$$

In the t-th iteration, we choose the regularization parameter $\lambda_t = 2.5(\lambda_* + \rho_t)$ with

$$\rho_t \asymp \max\left\{ r_{t-1}\sqrt{\frac{s\log(p/\delta)}{n}}, s^{-1/2}\sigma^2\tau^{-1} \right\} \asymp s^{-1/2}\max\left\{\alpha^t r_0, \sigma^2\tau^{-1}\right\}.$$

Commenced with $\widetilde{\beta}^{(t-1)}$ at iteration $t \geq 2$, we apply Theorem 5 to obtain that conditioned on event $\mathcal{E}_{t-1}(r_{t-1}) \cap \mathcal{E}_*(\lambda_*)$,

$$\widetilde{\beta}^{(t)} \in \Lambda \quad \text{and} \quad \|\widetilde{\beta}^{(t)} - \beta^*\|_\Sigma \leq \alpha r_{t-1} + r_* = r_t \tag{A.31}$$

with probability at least $1-\delta$. In other words, event $\mathcal{E}_t(r_t)$ occurs with probability at least $1-\delta$ conditioned on $\mathcal{E}_{t-1}(r_{t-1}) \cap \mathcal{E}_*(\lambda_*)$.

Finally, we choose $T = \lceil \log(r_0/r_*)/\log(1/\alpha) \rceil$ so that $\alpha^T r_0 \leq r_*$. Then, applying (A.30), (A.31) and the union bound over $t = 1,\dots,T$ yields that, conditioned on $\mathcal{E}_0(r_0) \cap \mathcal{E}_*(r_*)$, the T-th iterate $\widetilde{\beta}^{(T)}$ falls into the cone Λ and satisfies the error bound

$$\|\widetilde{\beta}^{(T)} - \beta^*\|_\Sigma \leq r_T \asymp r_*$$

with probability at least $1 - T\delta$. This completes the proof of the theorem. □

Appendix B

Supplementary Materials for Chapter 3

B.1 Notation and Expressions

We define or repeat some of the important quantities that will be used in the proofs. For i in $1,\ldots,n$,

$$\begin{aligned}
&M_{ci}(t;S_c) = I(X_i \leq t, \Delta_i = 0) - \int_0^t I(X_i \geq u) d\Lambda_c(u;A_i,Z_i),\\
&J_i(t;S,S_c) = \int_0^t \frac{dM_c(u;S_c)}{S(u;A_i,Z_i)S_c(u;A_i,Z_i)},\\
&d\mathcal{N}_i(t;S,S_c) = \frac{dN_i(t)}{S_c(t;A_i,Z_i)} - J_i(t;S,S_c)dS(t;A_i,Z_i),\\
&dM_i(t;\beta,\Lambda,S,S_c) = d\mathcal{N}_i(t;S,S_c) - \left\{\frac{Y_i(t)}{S_c(t;A_i,Z_i)} + J_i(t;S,S_c)S(t;A_i,Z_i)\right\} e^{\beta A_i} d\Lambda(t),\\
&\mathcal{S}^{(l)}(t;\beta,S,S_c) = \frac{1}{n}\sum_{i=1}^{n} \Gamma_i^{(l)}(t;\beta,S,S_c),\\
&\bar{A}(t;\beta,S,S_c) = \frac{\mathcal{S}^{(1)}(t;\beta,S,S_c)}{\mathcal{S}^{(0)}(t;\beta,S,S_c)},\\
&V(t;\beta,S,S_c) = \bar{A}(t;\beta,S,S_c) - \bar{A}(t;\beta,S,S_c)^2,\\
&\widetilde{\Lambda}(t;\beta,S,S_c) = \frac{1}{n}\sum_{i=1}^{n}\int_0^t \frac{d\mathcal{N}_i(u;S,S_c)}{\mathcal{S}^{(0)}(u;\beta,S,S_c)},\\
&\widetilde{\psi}_i(\beta,\Lambda,S,S_c) = \int_0^\tau \{A_i - \bar{A}(t;\beta,S,S_c)\} dM_i(t;\beta,\Lambda,S,S_c),
\end{aligned}$$

where $\Lambda_c(t;A,Z)=\int_0^t \lambda_c(t;A,Z)du$ and $F(t;A,Z)=1-S(t;A,Z)$.

Next, for each fold m, we define the fold-specific quantities:

$$
\begin{aligned}
\mathcal{S}_m^{(l)}(t;\beta,S,S_c) &= \frac{1}{|I_m|}\sum_{i\in I_m}\Gamma_i^{(l)}(t;\beta,S,S_c),\\
\bar{A}_m(t;\beta,S,S_c) &= \frac{\mathcal{S}_m^{(1)}(t;\beta,S,S_c)}{\mathcal{S}_m^{(0)}(t;\beta,S,S_c)},\\
V_m(t;\beta,S,S_c) &= \bar{A}_m(t;\beta,S,S_c)-\bar{A}_m(t;\beta,S,S_c)^2,\\
\widetilde{\Lambda}_m(t;\beta,S,S_c) &= \frac{1}{|I_m|}\sum_{i\in I_m}\int_0^t \frac{d\mathcal{N}_i(u;S,S_c)}{\mathcal{S}_m^{(0)}(u;\beta,S,S_c)},\\
\widetilde{\psi}_{m,i}(\beta,\Lambda,S,S_c) &= \int_0^\tau \{A_i-\bar{A}_m(t;\beta,S,S_c)\}dM_i(t;\beta,\Lambda,S,S_c).
\end{aligned}
$$

The cross-fitted variance estimator of $\widehat{\beta}$ form Theorem 9 is defined as

$$
\widehat{\sigma}^2_{cf}(\beta)=\frac{\frac{1}{n}\sum_{m=1}^k\sum_{i\in I_m}\widetilde{\psi}_{m,i}(\beta,\widetilde{\Lambda}_m(\cdot;\beta,\widehat{S}^{(-m)},\widehat{S}_c^{(-m)}),\widehat{S}^{(-m)},\widehat{S}_c^{(-m)})^2}{\left\{\frac{1}{n}\sum_{m=1}^k\sum_{i\in I_m}\int_0^\tau V_m(t;\beta,\widehat{S}^{(-m)},\widehat{S}_c^{(-m)})d\mathcal{N}_i(t;\widehat{S}^{(-m)},\widehat{S}_c^{(-m)})\right\}^2},
$$

where $\widehat{S}^{(-m)}$ and $\widehat{S}_c^{(-m)}$ are estimated using the out of m-th fold data.

B.2 Proof of Double Robustness

Lemma 9. *For any $S_c(t|A,Z)$ with its corresponding censoring specific martingale $M_c(t;S_c)$,*

$$\int_0^t \frac{dM_c(u;S_c)}{S_c(u|A,Z)} = 1 - \frac{Y(t)}{S_c(t|A,Z)} - \frac{N(t-)}{S_c(X|A,Z)}, \tag{B.1}$$

where $N(t-) = I(X < t, T \le C)$.

Note, this can be seen as a continuous version of Lemma 10.4 in Tsiatis (2006).

Proof

First note that

$$\int_0^t \frac{dN_c(u)}{S_c(u|A,Z)} = \frac{N_c(t-)}{S_c(X|A,Z)}, \tag{B.2}$$

where $N_C(t-) = I(X < t, T > C)$. Next, since $S_c(u|A,Z) = \exp\{-\Lambda_c(u|A,Z)\}$,

$$\begin{aligned}
&\int_0^t \frac{-Y(u)d\Lambda_c(u|A,Z)}{S_c(u|A,Z)} \\
&= I(X \ge t)\int_0^t \frac{dS_c(u|A,Z)}{S_c(u|A,Z)^2} + I(X < t)\int_0^X \frac{dS_c(u|A,Z)}{S_c(u|A,Z)^2} \\
&= I(X \ge t)\{-S_c(u|A,Z)^{-1}\}|_{u=0}^{u=t} + I(X < t)\{-S_c(u|A,Z)^{-1}\}|_{u=0}^{u=X} \\
&= 1 - \frac{Y(t)}{S_c(t|A,Z)} - \frac{I(X < t)}{S_c(X|A,Z)}.
\end{aligned} \tag{B.3}$$

Since $I(X < t) = N(t-) + N_c(t-)$, (B.2) + (B.3) then gives the lemma. □

Proof of Theorem 7

Recall that

$$dM^{aug}(t;\beta,\Lambda_0,S,S_c) = dM^w(t;\beta,\Lambda_0,S_c) - J(t;S,S_c)\left\{dS(t|A,Z) + S(t|A,Z)e^{\beta A}d\Lambda_0(t)\right\},$$

where $J(t;S,S_c)$ is also included in Appendix B.1.

a) Assume $S_c = S_c^o$.

We first consider $dM^w(t;\beta^o,\Lambda_0^o,S_c^o)$. For $h(A) = 1$ or A,

$$
\begin{aligned}
&E\{h(A)dM^w(t;\beta^o,\Lambda_0^o,S_c^o)\}\\
=&E\left\{h(A)S_c^o(t|A,Z)^{-1}\left[dE\{I(T \le t)I(C \ge t)|T,A,Z\}\right.\right.\\
&\left.\left.-E\{I(T \ge t)I(C \ge t)|T,A,Z\} \cdot e^{\beta^o A}d\Lambda_0^o(t)\right]\right\}\\
=&E[h(A)S_c^o(t|A,Z)^{-1}\{dI(T \le t)P(C \ge t|A,Z)\\
&-I(T \ge t)P(C \ge t|A,Z) \cdot e^{\beta^o A}d\Lambda_0^o(t)\}]\\
=&E\{h(A)dM_T(t;\beta^o,\Lambda_0^o)\}\\
=&E[h(A)dE\{M_T(t;\beta^o,\Lambda_0^o)|A\}]\\
=&0,
\end{aligned}
$$

where the second '=' above uses the informative censoring Assumption 1, and the second last '=' above uses the martingale property of $M_T(t;\beta^o,\Lambda_0^o)$.

Next we consider $J(t;S,S_c)\{dS(t|A,Z)+S(t|A,Z)e^{\beta^o A}d\Lambda_0^o(t)\}$. Its expectation being zero follows immediately from the fact that $M_c(t;S_c^o)$ is a martingale.

b) Assume $S = S^o$.

Noting that $Y_T(t)N(t-) = N(t-)dN_T(t) = 0$ and $Y(t)dN_T(t) = dN(t)$, we multiply (B.1) by

$dM_T(t) = dN_T(t) - Y_T(t)e^{\beta^o A} d\Lambda_0(t)$ giving:

$$
\begin{aligned}
& dM_T(t;\beta^o,\Lambda_0^o)\int_0^t \frac{dM_c(u;S_c)}{S_c(u|A,Z)} \\
=& dN_T(t)\int_0^t \frac{dM_c(u;S_c)}{S_c(u|A,Z)} - Y_T(t)e^{\beta^o A} d\Lambda_0^o(t)\int_0^t \frac{dM_c(u;S_c)}{S_c(u|A,Z)} \\
=& dN_T(t) - \frac{dN_T(t)Y(t)}{S_c(t|A,Z)} - \frac{dN_T(t)N(t-)}{S_c(X|A,Z)} - Y_T(t)e^{\beta^o A} d\Lambda_0^o(t) + \\
& + \frac{Y(t)e^{\beta^o A} d\Lambda_0^o(t)}{S_c(t|A,Z)} + \frac{Y_T(t)N(t-)e^{\beta^o A} d\Lambda_0^o(t)}{S_c(X|A,Z)} \\
=& dM_T(t) - dM^w(t).
\end{aligned}
$$

Therefore

$$dM^w(t;\beta^o,\Lambda_0^o) = dM_T(t;\beta^o,\Lambda_0^o) - dM_T(t;\beta^o,\Lambda_0^o)\int_0^t \frac{dM_c(u;S_c)}{S_c(u|A,Z)}.$$

We note that (3.9) and (3.10) hold when $S = S^o$. From (3.7) then we have

$$
\begin{aligned}
& E\{dM^{aug}(t;\beta^o,\Lambda_0^o,S^o,S_c)\} \\
=& E\left[dM^w(t;\beta^o,\Lambda_0^o,S_c) + \int_0^t E\{dM_T(t;\beta^o,\Lambda_0^o)|A,Z,T \geq u\}\frac{dM_c(u;S_c)}{S_c(u|A,Z)}\right] \\
=& E\left[dM_T(t;\beta^o,\Lambda_0^o) - dM_T(t;\beta^o,\Lambda_0^o)\int_0^t \frac{dM_c(u;S_c)}{S_c(u|A,Z)}\right. \\
& \left. + \int_0^t E\{dM_T(t;\beta^o,\Lambda_0^o)|A,Z,T \geq u\}\frac{dM_c(u;S_c)}{S_c(u|A,Z)}\right] \\
=& E\left\{\int_0^t \left[E\{dM_T(t;\beta^o,\Lambda_0^o)|A,Z,T \geq u\} - dM_T(t;\beta^o,\Lambda_0^o)\right]\frac{dM_c(u;S_c)}{S_c(u|A,Z)}\right\} \\
=& E\left[E\left\{\int_0^t \left[E\{dM_T(t;\beta^o,\Lambda_0^o)|A,Z,T \geq u\} - dM_T(t;\beta^o,\Lambda_0^o)\right]\right.\right. \\
& \left.\left. \times\frac{dN_c(u)}{S_c(u|A,Z)}\right|A,Z,T \geq u, C = u\right\}\Bigg] \\
& - E\left[E\left\{\int_0^t \left[E\{dM_T(t;\beta^o,\Lambda_0^o)|A,Z,T \geq u\} - dM_T(t;\beta^o,\Lambda_0^o)\right]\right.\right. \\
& \left.\left. \times\frac{Y(u)d\Lambda_c(u)}{S_c(u|A,Z)}\right|A,Z,T \geq u, C \geq u\right\}\Bigg] \\
=& E\left\{\int_0^t \frac{dN_c(u)}{S_c(u|A,Z)}\left[E\{dM_T(t;\beta^o,\Lambda_0^o)|A,Z,T \geq u, C = u\}\right.\right. \\
& \left.\left. - E\{dM_T(t;\beta^o,\Lambda_0^o)|A,Z,T \geq u, C = u\}\right]\right\} \\
& - E\left\{\int_0^t \frac{Y(u)d\Lambda_c(u)}{S_c(u|A,Z)}\left[E\{dM_T(t;\beta^o,\Lambda_0^o)|A,Z,T \geq u, C \geq u\}\right.\right. \\
& \left.\left. - E\{dM_T(t;\beta^o,\Lambda_0^o)|A,Z,T \geq u, C \geq u\}\right]\right\} \\
=& 0,
\end{aligned}
$$

where in the 3rd line above $E\{dM_T(t;\beta^o,\Lambda_0^o)\} = 0$ because $M_T(t;\beta^o,\Lambda_0^o)$ is a martingale.

The above also gives

$$E\left\{\int_0^t AdM^{aug}(t;\beta^o,\Lambda_0^o,S^o,S_c)\right\} = 0.$$

□

Appendix C

Supplementary Materials for Chapter 4

C.1 Notation and Expressions

Throughout the Supplementary Material, we omit the notational dependencies of most quantities on A and Z, unless it requires clarification. For any random quantities a and b, we will use $a \lesssim b$ to denote a is less than or equal to b up to a constant factor.

Suppose we are given n i.i.d. data points, with $i \in \{1, \ldots, n\}$, that can be split into k equal-sized folds $I_1, \ldots, I_k$, we first collect notations that will be used repeatedly in the proofs. For $a = 0,1$, $l = 0,1$ and for each i,

$$
\begin{aligned}
&N_{Ti}(t) = I(T_i \le t), \qquad\qquad\qquad Y_{Ti}(t) = I(T_i \ge t),\\
&M_{Ti}(t;\beta,\Lambda_0) = N_{Ti}(t) - \int_0^t Y_{Ti}(u)e^{\beta A_i}d\Lambda_0(u),\\
&N_{ci}(t) = I(X_i \le t, \Delta_i = 0), \qquad\qquad Y_i(t) = I(X_i \ge t),\\
&M_{ci}(t;a,S_c) = N_{ci}(t) - \int_0^t Y_i(u)d\Lambda_c(u;a,Z_i),\\
&N^a(t) = I\{X(a) \le t, T(a) \le C(a)\}, \qquad Y^a(t) = I\{X(a) \ge t\}\\
&N_c^a(t) = I\{X(a) \le t, T(a) > C(a)\}, \qquad \Delta^a(t) = I\{\min(T(a),t) \le C(a)\},\\
&J_i(t;a,S,S_c) = \int_0^t \frac{dM_{ci}(u;a,S_c)}{S(u;a,Z_i)S_c(u;a,Z_i)},\\
&D_{1i}^w(t;\beta,\Lambda_0,\pi,S_c) = \frac{dM_i(t;\beta,\Lambda_0)}{\pi(Z_i)_i^A\{1-\pi(Z_i)\}^{1-A_i}S_c(t;A_i,Z_i)}\\
&d\mathcal{N}_i^{(l)}(t;\pi,S,S_c) = \frac{A_i^l dN_i(t)}{\pi(Z_i)^{A_i}\{1-\pi(Z_i)\}^{1-A_i}S_c(t;A_i,Z_i)} + \frac{A_i^l dS(t;A_i,Z_i)}{\pi(Z_i)^{A_i}\{1-\pi(Z_i)\}^{1-A_i}}\\
&\qquad\qquad - \sum_{a=0,1} a^l\left\{1 + \frac{A_i^a(1-A_i)^{1-a}}{\pi(Z_i)^a\{1-\pi(Z_i)\}^{1-a}}J_i(t;a,S,S_c)\right\}dS(t;a,Z_i),\\
&\Gamma_i^{(l)}(t;\beta,\pi,S,S_c) = \frac{A_i^l Y_i(t)e^{\beta A_i}}{\pi(Z_i)^{A_i}\{1-\pi(Z_i)\}^{1-A_i}S_c(t;A_i,Z_i)} - \frac{A_i^l S(t;A_i,Z_i)e^{\beta A_i}}{\pi(Z_i)^{A_i}\{1-\pi(Z_i)\}^{1-A_i}}\\
&\qquad\qquad + \sum_{a=0,1} a^l\left\{1 + \frac{A_i^a(1-A_i)^{1-a}}{\pi(Z_i)^a\{1-\pi(Z_i)\}^{1-a}}J_i(t;a,S,S_c)\right\}S(t;a,Z_i)e^{\beta a},\\
&D_{1i}(t;\beta,\Lambda_0,\pi,S,S_c) = d\mathcal{N}_i^{(0)}(t;\pi,S,S_c) - \Gamma_i^{(0)}(t;\beta,\pi,S,S_c)d\Lambda_0(t),\\
&D_{2i}(\beta,\Lambda_0,\pi,S,S_c) = \int_0^\tau d\mathcal{N}_i^{(1)}(t;\pi,S,S_c) - \Gamma_i^{(1)}(t;\beta,\pi,S,S_c)d\Lambda_0(t),\\
&s^{(1)}(t;\beta,\pi,S,S_c) = \frac{\partial}{\partial\beta}s^{(0)}(t;\beta,\pi,S,S_c) = \frac{\partial^2}{\partial\beta^2}s^{(0)}(t;\beta,\pi,S,S_c),\\
&\bar{\alpha}(t;\beta,\pi,S,S_c) = \frac{s^{(1)}(t;\beta,\pi,S,S_c)}{s^{(0)}(t;\beta,\pi,S,S_c)},\\
&v(t;\beta,\pi,S,S_c) = \bar{\alpha}(t;\beta,\pi,S,S_c) - \bar{\alpha}(t;\beta,\pi,S,S_c)^2,\\
&\nu(\beta,\pi,S,S_c) = \int_0^\tau v(t;\beta,\pi,S,S_c)s^{(0)}(t;\beta^o,\pi,S,S_c)d\Lambda_0^o(t)\\
&\mu(\beta,\pi,S,S_c) = \int_0^\tau \{\bar{\alpha}(t;\beta^o,\pi,S,S_c) - \bar{\alpha}(t;\beta,\pi,S,S_c)\}s^{(0)}(t;\beta^o,\pi,S,S_c)d\Lambda_0^o(t).
\end{aligned}
$$

Note that the quantities in the last 4 lines are defined in the Additional Assumptions Section C.3.1 for the Proof of Asymptotics Results later.

Next, we define quantities evaluated over the entire sample of n observations:

$$
\begin{aligned}
&\mathcal{S}^{(l)}(t;\beta,\pi,S,S_c) = \frac{1}{n}\sum_{i=1}^{n}\Gamma_i^{(l)}(t;\beta,\pi,S,S_c), \quad s^{(l)}(t;\beta,\pi,S,S_c) = E\{\mathcal{S}^{(l)}(t;\beta,\pi,S,S_c)\}\\
&\bar{A}(t;\beta,\pi,S,S_c) = \frac{\mathcal{S}^{(1)}(t;\beta,\pi,S,S_c)}{\mathcal{S}^{(0)}(t;\beta,\pi,S,S_c)},\\
&V(t;\beta,\pi,S,S_c) = \bar{A}(t;\beta,\pi,S,S_c) - \bar{A}(t;\beta,\pi,S,S_c)^2,\\
&\widetilde{\Lambda}_0(t;\beta,\pi,S,S_c) = \frac{1}{n}\sum_{i=1}^{n}\int_0^t \frac{d\mathcal{N}_i(u;\pi,S,S_c)}{\mathcal{S}^{(0)}(u;\beta,\pi,S,S_c)},\\
&U(\beta,\pi,S,S_c) = \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau d\mathcal{N}_i^{(1)}(t;\pi,S,S_c) - \bar{A}(t;\beta,\pi,S,S_c)d\mathcal{N}_i^{(0)}(t;\pi,S,S_c).
\end{aligned}
$$

Analogous to the quantities above, for each fold m, we define the fold-specific quantities $\mathcal{S}_m^{(l)}(t;\beta,\pi,S,S_c)$, $\bar{A}_m(t;\beta,\pi,S,S_c)$, $V_m(t;\beta,\pi,S,S_c)$, $\widetilde{\Lambda}_{0,m}(t;\beta,\pi,S,S_c)$, and $U_m(\beta,\pi,S,S_c)$. For example,

$$
\mathcal{S}_m^{(l)}(t;\beta,\pi,S,S_c) = \frac{1}{|I_m|}\sum_{i\in I_m}\Gamma_i^{(l)}(t;\beta,\pi,S,S_c).
$$

We will now denote the cross-fitted AIPW estimating equation as

$$
U_{cf}(\beta) = \frac{1}{k}\sum_{m=1}^{k} U_m(\beta,\widehat{\pi}^{(-m)},\widehat{S}^{(-m)},\widehat{S}_c^{(-m)}) = 0.
$$

The asymptotic variance of $\widehat{\beta}$ in Theorems 12 is defined as

$$
\sigma^2 = E\{\psi(\beta^o,\Lambda_0^o,\pi^o,S^o,S_c^o)^2\}/\nu^2(\beta^o,\pi^o,S^o,S_c^o),
$$

where

$$
\psi(\beta^o,\Lambda_0,\pi,S,S_c) = D_2(\beta,\Lambda_0,\pi,S,S_c) - \int_0^\tau \bar{\alpha}(t;\beta,\pi,S,S_c)D_1(t;\beta,\Lambda_0,\pi,S,S_c).
$$

The asymptotic variance σ^2 can be consistently estimated using

$$\widehat{\sigma}^2(\widehat{\beta}) = \frac{\frac{1}{n}\sum_{m=1}^{k}\sum_{i\in I_m}\widetilde{\psi}_{m,i}(\widehat{\beta},\widetilde{\Lambda}_{0,m}(\cdot;\widehat{\beta},\widehat{\pi}^{(-m)},\widehat{S}^{(-m)},\widehat{S}_c^{(-m)}),\widehat{\pi}^{(-m)},\widehat{S}^{(-m)},\widehat{S}_c^{(-m)})^2}{\left\{\frac{1}{n}\sum_{m=1}^{k}\sum_{i\in I_m}\int_0^\tau V_m(t;\widehat{\beta},\widehat{\pi}^{(-m)},\widehat{S}^{(-m)},\widehat{S}_c^{(-m)})d\mathcal{N}_i^{(0)}(t;\widehat{\pi}^{(-m)},\widehat{S}^{(-m)},\widehat{S}_c^{(-m)})\right\}^2},$$

where

$$\widetilde{\psi}_{m,i}(\beta^o,\Lambda_0,\pi,S,S_c) = D_{2i}(t;\beta,\Lambda_0,\pi,S,S_c) - \int_0^\tau \bar{A}_m(t;\beta,\pi,S,S_c)D_{1i}(t;\beta,\Lambda_0,\pi,S,S_c)$$

C.2 Proof of Double Robustness

We first state and prove some lemmas we will use in the proofs of the Double Robustness Theorem.

Lemma 10. *For any real-valued functions g and h, we have*

$$E\{g(A,Z)h(T,C,A,Z)\} = \sum_{a=0,1} E\left[g(a,Z)\pi^o(Z)^a\{(1-\pi^o(Z)\}^{1-a}E\{h(T,C,A,Z)|A=a,Z\}\right]$$

Proof

By the law of total expectation we have

$$\begin{aligned}
&E\{g(A,Z)h(T,C,A,Z)\}\\
=&E[E\{g(A,Z)h(T,C,A,Z)|A,Z\}]\\
=&E\left[\sum_{a=0,1} E\{g(A,Z)h(T,C,A,Z)|A=a,Z\}\pi^o(Z)^a\{1-\pi^o(Z)\}^{1-a}\right]\\
=&\sum_{a=0,1} E\left[g(a,Z)\pi^o(Z)^a\{(1-\pi^o(Z)\}^{1-a}E\{h(T,C,A,Z)|A=a,Z\}\right]
\end{aligned}$$

□

Lemma 11.

$$M(t;\beta,\Lambda_0) = A\Delta^1(t)M_T^1(t;\beta,\Lambda_0) + (1-A)\Delta^0(t)M_T^0(t;\beta,\Lambda_0).$$

Proof

We prove the result for $a=1$, the same arguments can be made for $a=0$.

The following is a potential outcome version of Lemma 1 from Luo and Xu (2022).

Note that

$$\int_0^t \frac{dN_c^1(u)}{S_c(u;1,Z)} = \frac{N_c^1(t-)}{S_c(X;1,Z)}. \tag{C.1}$$

Since $\Lambda_c(u;1,Z) = -\log\{S_c(u;1,Z)\}$,

$$\begin{aligned}
&\int_0^t \frac{-Y^1(u)d\Lambda_c(u;1,Z)}{S_c(u;1,Z)}\\
=&I\{X(1) \geq t\}\int_0^t \frac{dS_c(u;1,Z)}{S_c(u;1,Z)^2} + I\{X(1) < t\}\int_0^{X(1)} \frac{dS_c(u;1,Z)}{S_c(u;1,Z)^2}\\
=&I\{X(1) \geq t\}\{-S_c(u;1,Z)^{-1}\}|_{u=0}^{u=t} + I\{X(1) < t\}\{-S_c(u;1,Z)^{-1}\}|_{u=0}^{u=X(1)}\\
=&\frac{I\{X(1) \geq t\}}{S_c(0;1,Z)} + \frac{I\{X(1) < t\}}{S_c(0;1,Z)} - \frac{I\{X(1) \geq t\}}{S_c(t;1,Z)} - \frac{I(X(1) < t)}{S_c(X(1);1,Z)},\\
=&1 - \frac{Y^1(t)}{S_c(t;1,Z)} - \frac{I(X(1) < t)}{S_c(X(1);1,Z)}.
\end{aligned} \tag{C.2}$$

Since $I(X(1) < t) = N^1(t-) + N_c^1(t-)$, (C.1) + (C.2) gives

$$\int_0^t \frac{dM_c(u;1,S_c)}{S_c(u;1,Z)} = 1 - \frac{Y^1(t)}{S_c(t;1,Z)} - \frac{N^1(t-)}{S_c(X(1);1,Z)}, \tag{C.3}$$

The rest of the proof is analogous to part (b) of the proof of Theorem 1 from Luo and Xu (2022) . Noting that $Y^1(t)dN_T^1(t) = dN^1(t)$ and $dN_T^1(t)N^1(t-) = Y_T^1(t)N^1(t-) = 0$, we multiply (C.3) by $dM_T^1(t;\beta,\Lambda_0) = dN_T^1(t) - Y_T^1(t)e^{\beta}d\Lambda_0(t)$ giving:

$$
\begin{aligned}
& dM_T^1(t;\beta,\Lambda_0)\int_0^t \frac{dM_c(u;1,S_c)}{S_c(u;1,Z)} \\
=& dN_T^1(t)\int_0^t \frac{dM_c(u;1,S_c)}{S_c(u;1,Z)} - Y_T^1(t)e^{\beta}d\Lambda_0(t)\int_0^t \frac{dM_c(u;1,S_c)}{S_c(u;1,Z)} \\
=& dN_T^1(t) - \frac{dN_T^1(t)Y^1(t)}{S_c(t;1,Z)} - \frac{dN_T^1(t)N^1(t-)}{S_c(X(1);1,Z)} - Y_T^1(t)e^{\beta}d\Lambda_0(t) + \frac{Y^1(t)e^{\beta}d\Lambda_0(t)}{S_c(t;1,Z)} + \frac{Y_T^1(t)N^1(t-)e^{\beta^o}d\Lambda_0^o(t)}{S_c(X(1);1,Z)}. \\
=& dN_T^1(t) - Y_T^1(t)e^{\beta}d\Lambda_0(t) - \frac{dN^1(t)}{S_c(t;1,Z)} + \frac{Y^1(t)e^{\beta}d\Lambda_0(t)}{S_c(t;1,Z)} \\
=& dM_T^1(t;\beta,\Lambda_0) - \frac{dN^1(t) - Y^1(t)e^{\beta}d\Lambda_0(t)}{S_c(t;1,Z)} \\
=& dM_T^1(t;\beta,\Lambda_0) - \frac{\Delta^1(t)dN_T^1(t) - \Delta^1(t)Y_T^1(t)e^{\beta}d\Lambda_0(t)}{S_c(t;1,Z)} \\
=& dM_T^1(t;\beta,\Lambda_0) - \frac{\Delta^1(t)dM_T^1(t;\beta,\Lambda_0)}{S_c(t;1,Z)}.
\end{aligned}
$$

□

Lemma 12.

$$\frac{\Delta^a(t)dM_T^a(t;\beta,\Lambda_0)}{S_c(t;a,Z)} = dM_T^a(t;\beta,\Lambda_0) - dM_T^a(t;\beta,\Lambda_0)\int_0^t \frac{dM_c(u;a,S_c)}{S_c(u;a,Z)}$$

for $a = 0,1$.

Proof

By definition $N^a(t) = I\{T(a) \leq C(a)\}I\{T(a) \leq t\}$. Meanwhile $N_T^a(t)\Delta^a(t) = I\{T(a) \leq t\}I\{\min(T(a),t) \leq C(a)\} = I\{T(a) \leq t\}I\{T(a) \leq C(a)\}$. Therefore

$N^a(t) = N_T^a(t)\Delta^a(t)$.

In addition, $Y_T^a(t)\Delta^a(t) = I(T(a) \geq t)I\{\min(T(a),t) \leq C(a)\} = I(T(a) \geq t)I\{C(a) \geq t\}$

$= I(X(a) \geq t) = Y^a(t)$.

Then by the consistency Assumption 7, we have

$$\begin{aligned} N(t) &= AN^1(t) + (1-A)N^0(t) \\ &= AN_T^1(t)\Delta^1(t) + (1-A)N_T^0(t)\Delta^0(t). \end{aligned} \tag{C.4}$$

Similarly,

$$Y(t) = AY_T^1(t)\Delta^1(t) + (1-A)Y_T^0(t)\Delta^0(t). \tag{C.5}$$

We may then combine (C.4) and (C.5) to get

$$\begin{aligned} M(t;\beta,\Lambda_0) &= N(t) - \int_0^t Y(u)e^{\beta A}d\Lambda_0(u) \\ &= A\Delta^1(t)M_T^1(t;\beta,\Lambda_0) + (1-A)\Delta^0(t)M_T^0(t;\beta,\Lambda_0). \end{aligned}$$

□

Proof of Theorem 10

Note that

$$\begin{aligned} D_1(t;\beta^o,\Lambda_0^o,\pi,S,S_c) &= d\mathcal{N}_i^{(0)}(t;\pi,S,S_c) - \Gamma_i^{(0)}(t;\beta^o,\pi,S,S_c)d\Lambda_0^o(t), \\ D_2(\beta^o,\Lambda_0^o,\pi,S,S_c) &= \int_0^\tau d\mathcal{N}_i^{(1)}(t;\pi,S,S_c) - \Gamma_i^{(1)}(t;\beta^o,\pi,S,S_c)d\Lambda_0^o(t). \end{aligned}$$

.

By Assumptions 9-10, D_2 is absolutely integrable, so Fubini's theorem gives $ED_2 = E\{\int_0^\tau d\mathcal{N}_i^{(1)}(t) - \Gamma_i^{(1)}(t)d\Lambda_0^o(t)\} = \int_0^\tau E\{\int_0^\tau d\mathcal{N}_i^{(1)}(t) - \Gamma_i^{(1)}(t)d\Lambda_0^o(t)\}$. So it suffices to show that $E\{d\mathcal{N}_i^{(l)}(t) - \Gamma_i^{(l)}(t)d\Lambda_0^o(t)\} = 0$ for $l = 0,1$ and for any $t \in [0,\tau]$.

a) Assume $(\pi, S_c) = (\pi^o, S_c^o)$. For simplicity, we omit the dependency on all the arguments l and write $E\{d\mathcal{N}_i^{(l)}(t) - \Gamma_i^{(l)}(t)d\Lambda_0^o(t)\} = R_1 + R_2 - R_3$, where

$$R_1 = E\left[\frac{A^l dM(t;\beta^o,\Lambda_0^o)}{\pi^o(Z)^A\{1-\pi^o(Z)\}^{1-A}S_c^o(t;A,Z)}\right],$$

$$R_2 = E\left[\frac{A^l\{dS(t;A,Z)+S(t;A,Z)e^{\beta^o A}d\Lambda_0^o(t)\}}{\pi^o(Z)^A\{1-\pi^o(Z)\}^{1-A}} - \sum_{a=0,1} a^l\{dS(t;a,Z)+S(t;a,Z)e^{\beta^o a}d\Lambda_0^o(t)\}\right],$$

$$R_3 = E\left[\sum_{a=0,1} a^l \frac{A^a(1-A)^{1-a}}{\pi^o(Z)^a\{1-\pi^o(Z)\}^{1-a}} J(t;a,S,S_c^o)\{dS(t;a,Z)+S(t;a,Z)e^{\beta^o a}d\Lambda_0^o(t)\}\right].$$

Using Lemma 10 and Lemma 11, we then have

$$\begin{aligned}
R_1 &= \sum_{a=0,1} E\left[\frac{a^l}{S_c^o(t;a,Z)} E\left\{dM(t;\beta^o,\Lambda_0^o)|A=a,Z\right\}\right] \\
&= \sum_{a=0,1} E\left[\frac{a^l}{S_c^o(t;a,Z)} E\left\{\Delta^a(t)dM_T^a(t;\beta^o,\Lambda_0^o)|Z\right\}\right] \\
&= \sum_{a=0,1} E\left(\frac{a^l}{S_c^o(t;a,Z)} E\left[E\left\{\Delta^a(t)dM_T^a(t;\beta^o,\Lambda_0^o)|T(a)=t,Z\right\}|Z\right]\right) \\
&= \sum_{a=0,1} E\left\{\frac{a^l}{S_c^o(t;a,Z)} E\left(E\left[\{dN_T^a(t)I(C(a)\geq t) - Y_T^a(t)I(C(a)\geq t)e^{\beta^o a}d\Lambda_0^o(t)\}|T(a)=t,Z\right]\Big|Z\right)\right\} \\
&= \sum_{a=0,1} E\left\{\frac{a^l E\{I(C(a)\geq t)|Z\}}{S_c^o(t;a,Z)} E\left(E\left[\{dN_T^a(t) - Y_T^a(t)e^{\beta^o a}d\Lambda_0^o(t)\}|T(a)=t,Z\right]\Big|Z\right)\right\} \qquad \text{(C.6)} \\
&= \sum_{a=0,1} a^l dE\{M_T^a(t;\beta^o,\Lambda_0^o)\} \qquad \text{(C.7)} \\
&=0,
\end{aligned}$$

where (C.6) makes use of the informative censoring Assumption 10, and (C.7) uses the consistency Assumption 7 and the tower property.

Applying Lemma 10 to R_2, we have

$$R_2 = E\left[\sum_{a=0,1} a^l\{dS(t;a,Z)+S(t;a,Z)e^{\beta^o a}d\Lambda_0^o(t)\} - \sum_{a=0,1} a^l\{dS(t;a,Z)+S(t;a,Z)e^{\beta^o a}d\Lambda_0^o(t)\}\right]$$

$$=0.$$

Finally, again applying Lemma 10, we have

$$\begin{aligned} R_3 =& \sum_{a=0,1}\sum_{\alpha=0,1} a^l E\left[\frac{\alpha^a(1-\alpha)^{1-a}\pi^o(Z)^\alpha\{(1-\pi^o(Z)\}^{1-\alpha}}{\pi^o(Z)^a\{1-\pi^o(Z)\}^{1-a}}\{dS(t;a,Z)+S(t;a,Z)e^{\beta^o a}d\Lambda_0^o(t)\}\right. \\ & \left.\times E\{J(t;a,S,S_c^o)|A=\alpha,Z\}\right] \\ =& \sum_{a=0,1} a^l E\left[\{dS(t;a,Z)+S(t;a,Z)e^{\beta^o a}d\Lambda_0^o(t)\}E\{J(t;a,S,S_c^o)|A=a,Z\}\right] \qquad \text{(C.8)} \\ =& \sum_{a=0,1} a^l E\left[\{dS(t;a,Z)+S(t;a,Z)e^{\beta^o a}d\Lambda_0^o(t)\}\int_0^t \frac{dE\{M_c(u;a,S_c^o)|A=a,Z\}}{S(u;a,Z)S_c^o(u;a,Z)}\right] \\ =& 0, \qquad \text{(C.9)} \end{aligned}$$

where (C.8) comes from $\alpha^a(1-\alpha)^{1-a} = I(a=\alpha)$, and (C.9) uses the fact that for each $A=a$, $M_c(t;a,S_c^o)$ given Z is a martingale when $S_c = S_c^o$.

b) Assume $S = S^o$. We have $E\{d\mathcal{N}_i^{(l)}(t) - \Gamma_i^{(l)}(t)d\Lambda_0^o(t)\} = R_4 + R_5 + R_6$, where

$$\begin{aligned} R_4 =& E\left[\frac{A^l dM(t;\beta^o,\Lambda_0^o)}{\pi(Z)^A\{1-\pi(Z)\}^{1-A}S_c(t;A,Z)}\right. \\ & \left. - \sum_{a=0,1} a^l \frac{A^a(1-A)^{1-a}}{\pi(Z)^a\{1-\pi(Z)\}^{1-a}} J(t;a,S^o,S_c)\{dS^o(t;a,Z)+S^o(t;a,Z)e^{\beta^o a}d\Lambda_0^o(t)\}\right], \\ R_5 =& E\left[\frac{A^l\{dS^o(t;A,Z)+S^o(t;A,Z)e^{\beta^o A}d\Lambda_0^o(t)\}}{\pi(Z)^A\{1-\pi(Z)\}^{1-A}}\right], \\ R_6 =& -\sum_{a=0,1} a^l E\{dS^o(t;a,Z)+S^o(t;a,Z)e^{\beta^o a}d\Lambda_0^o(t)\}, \end{aligned}$$

We first make use of the fact that (4.7) holds under $S = S^o$ to get

$$R_4 = E\left[\frac{A^l dM(t;\beta^o,\Lambda_0^o)}{\pi(Z)^A\{1-\pi(Z)\}^{1-A}S_c(t;A,Z)}\right. \tag{C.10}$$

$$\left. + \sum_{a=0,1} a^l \frac{A^a(1-A)^{1-a}}{\pi(Z)^a\{1-\pi(Z)\}^{1-a}} \int_0^t \frac{dM_c(u;a,S_c)}{S_c(u;a,Z)} E\{dM_T(t;\beta^o,\Lambda_0^o)|T \geq u, A=a,Z\}\right]. \tag{C.11}$$

Applying Lemma 10 to both (C.10) and (C.11), we have

$$\begin{aligned}
R_4 = & \sum_{a=0,1} a^l E\left[\frac{\pi^o(Z)^a\{1-\pi^o(Z)\}^{1-a}}{\pi(Z)^a\{1-\pi(Z)\}^{1-a}} \frac{E\{dM(t;\beta^o,\Lambda_0^o)|A=a,Z\}}{S_c(t;a,Z)}\right] \\
& + \sum_{a=0,1}\sum_{\alpha=0,1} a^l E\left(\frac{\alpha^a(1-\alpha)^{1-a}\pi^o(Z)^\alpha\{1-\pi^o(Z)\}^{1-\alpha}}{\pi(Z)^a\{1-\pi(Z)\}^{1-a}}\right. \\
& \quad \left. \times E\left[\int_0^t \frac{dM_c(u;a,S_c)}{S_c(u;a,Z)} E\{dM_T(t;\beta^o,\Lambda_0^o)|T \geq u, A=a,Z\}\middle| A=\alpha,Z\right]\right)
\end{aligned}$$

$$= \sum_{a=0,1} a^l E\left[\frac{\pi^o(Z)^a\{1-\pi^o(Z)\}^{1-a}}{\pi(Z)^a\{1-\pi(Z)\}^{1-a}} \frac{\Delta^a(t) dM_T^a(t;\beta^o,\Lambda_0^o)}{S_c(t;a,Z)}\right] \tag{C.12}$$

$$\begin{aligned}
& + \sum_{a=0,1} a^l E\left(\frac{\pi^o(Z)^a\{1-\pi^o(Z)\}^{1-a}}{\pi(Z)^a\{1-\pi(Z)\}^{1-a}}\right. \\
& \quad \left. \times E\left[\int_0^t \frac{dM_c(u;a,S_c)}{S_c(u;a,Z)} E\{dM_T^a(t;\beta^o,\Lambda_0^o)|T(a) \geq u, A=a,Z\}\middle| A=a,Z\right]\right)
\end{aligned} \tag{C.13}$$

$$\begin{aligned}
= & \sum_{a=0,1} a^l E\left[\frac{\pi^o(Z)^a\{1-\pi^o(Z)\}^{1-a}}{\pi(Z)^a\{1-\pi(Z)\}^{1-a}}\left(\frac{\Delta^a(t) dM_T^a(t;\beta^o,\Lambda_0^o)}{S_c(t;a,Z)}\right.\right. \\
& \left.\left. + E\left[\int_0^t \frac{dM_c(u;a,S_c)}{S_c(u;a,Z)} E\{dM_T^a(t;\beta^o,\Lambda_0^o)|T(a) \geq u, A=a,Z\}\middle| A=a,Z\right]\right)\right]
\end{aligned}$$

$$= \sum_{a=0,1} a^l E\left[\frac{\pi^o(Z)^a\{1-\pi^o(Z)\}^{1-a}}{\pi(Z)^a\{1-\pi(Z)\}^{1-a}} dM_T^a(t;\beta^o,\Lambda_0^o)\right] \tag{C.14}$$

$$+ R_7,$$

where

$$
\begin{aligned}
&R_7\\
&= \sum_{a=0,1} a^l E\Bigg[\frac{\pi^o(Z)^a\{1-\pi^o(Z)\}^{1-a}}{\pi(Z)^a\{1-\pi(Z)\}^{1-a}}\\
&\quad \times \Bigg\{E\Big[\int_0^t \frac{dM_c(u;a,S_c)}{S_c(u;a,Z)} E\{dM_T^a(t;\beta^o,\Lambda_0^o)|T(a)\geq u, A=a,Z\}\Big| A=a,Z\Big]\\
&\qquad - \int_0^t \frac{dM_c(u;a,S_c)}{S_c(u;a,Z)} dM_T^a(t;\beta^o,\Lambda_0^o)\Bigg\}\Bigg]\\
&= \sum_{a=0,1} a^l E\Bigg[\frac{\pi^o(Z)^a\{1-\pi^o(Z)\}^{1-a}}{\pi(Z)^a\{1-\pi(Z)\}^{1-a}}\\
&\quad \times E\Bigg\{\int_0^t \frac{dM_c(u;a,S_c)}{S_c(u;a,Z)}[E\{dM_T^a(t;\beta^o,\Lambda_0^o)|T(a)\geq u, A=a,Z\} - dM_T^a(t;\beta^o,\Lambda_0^o)]\Bigg| A=a,Z\Bigg\}\Bigg],
\end{aligned}
$$

(C.12) uses Lemma 11 and the tower property, (C.13) makes use of $\alpha^a(1-\alpha)^{1-a} = I(a=\alpha)$ and the consistency Assumption 7, while (C.14) makes use of Lemma 12. Next, we show

$R_7 = 0$ by showing the inner conditional expectation is 0:

$$E\Big\{\int_0^t \frac{dM_c(u;a,S_c)}{S_c(u;a,Z)}[E\{dM_T^a(t;\beta^o,\Lambda_0^o)|T(a)\geq u,A=a,Z\} - dM_T^a(t;\beta^o,\Lambda_0^o)]\Big|A=a,Z\Big\}$$

$$=E\Big[E\Big\{\int_0^t \frac{dN_c(u)}{S_c(u;a,Z)}$$
$$\times [E\{dM_T^a(t;\beta^o,\Lambda_0^o)|T(a)\geq u,A=a,Z\} - dM_T^a(t;\beta^o,\Lambda_0^o)]\Big|A=a,Z,T\geq u,C=u\Big\}\Big|A=a,Z\Big]$$

$$-E\Big[E\Big\{\int_0^t \frac{Y(u)d\Lambda_c(u;a,Z)}{S_c(u;a,Z)}$$
$$\times [E\{dM_T^a(t;\beta^o,\Lambda_0^o)|T(a)\geq u,A=a,Z\} - dM_T^a(t;\beta^o,\Lambda_0^o)]\Big|A=a,Z,T\geq u,C=u\Big\}\Big|A=a,Z\Big]$$

$$=E\Big\{\int_0^t \frac{dN_c^a(u)}{S_c(u;a,Z)}$$
$$\times \Big[E\{dM_T^a(t;\beta^o,\Lambda_0^o)|T(a)\geq u,A=a,Z\} - E\{dM_T^a(t;\beta^o,\Lambda_0^o)|T(a)\geq u,A=a,Z\}\Big]\Big|A=a,Z\Big\} \tag{C.15}$$

$$-E\Big\{\int_0^t \frac{Y^a(u)d\Lambda_c(u;a,Z)}{S_c(u;a,Z)}$$
$$\times \Big[E\{dM_T^a(t;\beta^o,\Lambda_0^o)|T(a)\geq u,A=a,Z\} - E\{dM_T^a(t;\beta^o,\Lambda_0^o)|T(a)\geq u,A=a,Z\}\Big]\Big|A=a,Z\Big\} \tag{C.16}$$

$$=0,$$

where (C.15) and (C.16) uses consistency and informative censoring from Assumptions 7 and 10.

Next, using Lemma 10, we have

$$
\begin{aligned}
R_5 =& \sum_{a=0,1} a^l E\left[\frac{\pi^o(Z)^a\{1-\pi^o(Z)\}^{1-a}}{\pi(Z)^a\{1-\pi(Z)\}^{1-a}}\{dS^o(t;a,Z)+S^o(t;a,Z)e^{\beta^o a}d\Lambda_0^o(t)\}\right] \\
=& \sum_{a=0,1} a^l E\left[\frac{\pi^o(Z)^a\{1-\pi^o(Z)\}^{1-a}}{\pi(Z)^a\{1-\pi(Z)\}^{1-a}}E\{-dN_T^a(t)+Y_T^a(t)e^{\beta^o a}d\Lambda_0^o(t)|Z\}\right] \\
=& -\sum_{a=0,1} a^l E\left[E\left\{\frac{\pi^o(Z)^a\{1-\pi^o(Z)\}^{1-a}}{\pi(Z)^a\{1-\pi(Z)\}^{1-a}}dM_T^a(t;\beta^o,\Lambda_0^o)\Big|Z\right\}\right] \\
=& -\sum_{a=0,1} a^l E\left[\frac{\pi^o(Z)^a\{1-\pi^o(Z)\}^{1-a}}{\pi(Z)^a\{1-\pi(Z)\}^{1-a}}dM_T^a(t;\beta^o,\Lambda_0^o)\right].
\end{aligned}
$$

Lastly,

$$
\begin{aligned}
R_6 =& -\sum_{a=0,1} a^l E(E[\{dS^o(t;a,Z)+S^o(t;a,Z)e^{\beta^o a}d\Lambda_0^o(t)\}|Z]) \\
=& -\sum_{a=0,1} a^l E(E[\{-dN_T^a(t)+Y_T^a(t)e^{\beta^o a}d\Lambda_0^o(t)\}|Z]) \\
=& \sum_{a=0,1} a^l E\{dM_T^a(t;\beta^o,\Lambda_0^o)\} \\
=&0.
\end{aligned}
$$

The above gives $R_4+R_5+R_6=0$ as desired. □

C.3 Proof of Asymptotic Results

C.3.1 Additional Assumptions

Without loss of generality, we assume that the nuisance function estimates $\widehat{\pi},\widehat{S},\widehat{S}_c$ and their limits π^*,S^*,S_c^* only take value in $[0,1]$. Moreover, the survival functions $S^*(t;a,z)$ and $S_c^*(t;a,z)$ and their estimates are non-increasing in t.

Assumption 21. *There exists a neighbourhood $\mathcal{B}$ of β^o such that $\sup_{t\in[0,\tau],\beta\in\mathcal{B}}|\mathcal{S}^{(l)}(t;\beta,\pi^*,S^*,S_c^*)-s^{(l)}(t;\beta,\pi^*,S^*,S_c^*)|=o_p(1)$.*

Assumption 22. *For $l = 0,1$, $s^{(l)}(t;\beta,\pi^*,S^*,S_c^*)$ are continuous functions of $\beta \in \mathcal{B}$, uniformly in $t \in [0,\tau]$ and are bounded on $\mathcal{B} \times [0,\tau]$. $s^{(0)}(t;\beta,\pi^*,S^*,S_c^*)$ is bounded away from zero on $\mathcal{B} \times [0,\tau]$. For all $\beta \in \mathcal{B}$, $t \in [0,\tau]$:*

$$s^{(1)}(t;\beta,\pi,S,S_c) = \frac{\partial}{\partial\beta} s^{(0)}(t;\beta,\pi,S,S_c) = \frac{\partial^2}{\partial\beta^2} s^{(0)}(t;\beta,\pi,S,S_c).$$

In addition, let $\bar{\alpha} = s^{(1)}/s^{(0)}$ and $v = \bar{\alpha} - \bar{\alpha}^2$. We have

$$\nu(\beta^o,\pi^*,S^*,S_c^*) = \int_0^\tau v(t;\beta^o,\pi^*,S^*,S_c^*) s^{(0)}(t;\beta^o,\pi^*,S^*,S_c^*) d\Lambda_0^o(t) > 0.$$

Assumptions 21 and 22 are the typical regularity assumptions that are made under the Cox PH models (Andersen and Gill, 1982).

Assumption 23. *For $\pi = \widehat{\pi}$ or π^*, $S = \widehat{S}$ or S^*, and $S_c = \widehat{S}_c$ or S_c^*, where $\widehat{\pi}$, $\widehat{S}$ and $\widehat{S}_c$ are estimated using an independent sample, we have*

$$E\left\{\left[\sup_{t\in[0,\tau]}\left|s^{(l)}(t;\beta,\pi,S,S_c) - s^{(l)}(t;\beta,\pi^*,S^*,S_c^*)\right|\right]^2\right\} = o(1), \tag{C.17}$$

and

$$\sup_{t\in[0,\tau]} |\mathcal{S}^{(l)}(t;\beta,\pi,S,S_c) - s^{(l)}(t;\beta,\pi,S,S_c)| = O_p(n^{-1/2}), \tag{C.18}$$

for $\beta \in \mathcal{B}$ and $l = 0,1$. Moreover,

$$\int_0^\tau \{\bar{A}(t;\beta^o,\pi,S,S_c) - \bar{\alpha}(t;\beta^o,\pi,S,S_c)\} \cdot \frac{1}{\sqrt{n}} \sum_{i=1}^n D_{1i}(t;\beta^o,\Lambda_0^o,\pi^o,S^o,S_c^o) = o_p(1) \tag{C.19}$$

Assumption 23 is required due to the involvement of the time-dependent nuisance functions as well as the risk sets that are specific to the Cox MSM. Condition (C.17) simply states that the convergence of $\widehat{\pi},\widehat{S},\widehat{S}_c$ carries over to $s^{(l)}(t;\beta^o,\pi,S,S_c)$. Condition (C.18) should hold for most functions with simple structures even though the estimates of the nuisance function may converge at a slower than root-n rate. For example, if we have $G(t;h) = n^{-1}\sum_{i=1}^{n} A_i/h(t)$ and its limit $g(t;h) = E(A)/h(t)$, then

$$\sup_{t\in[0,\tau]} |G(t;\widehat{h}) - g(t;\widehat{h})| \le \left|\frac{1}{n}\sum_{i=1}^{n} A_i - E(A)\right| \cdot \sup_{t\in[0,\tau]} \left|\frac{1}{\widehat{h}(t)}\right| = O_p(n^{-1/2})$$

for any out-of-sample estimates $\widehat{h}(t)$ that are bounded away from zero. Condition (C.19) is required for the same reason the integral term $\mathcal{D}^{\dagger}(\widehat{S},\widehat{S}_c;S^o,S_c^o)$ in Assumption 12 is required. Although we have $\sqrt{n}\{\bar{A}(t;\beta^o,\pi,S,S_c) - \bar{\alpha}(t;\beta^o,\pi,S,S_c)\} = O_p(1)$ from (C.18), and $n^{-1}\sum_{i=1}^{n}\int_0^{\tau} D_{1i}(t;\beta^o,\Lambda_0^o,\pi^o,S^o,S_c^o) = o(1)$ from Theorem 10 and the law of large numbers, no existing tools allow us to generalize this product rate to increments within an integral, which is specific to our problem.

C.3.2 Proof of Main Results

We prove in this section the consistency and asymptotic normality of the cross-fitted AIPW estimator $\widehat{\beta}$. The proof of the main results is intentionally kept short and easy to follow, while the tedious details are put into the Lemmas 14 and 15. The proof of Lemma 14 involves standard convergence in probability arguments, regardless of whether we use cross-fitting or not. On the other hand, the proof of Lemma 15 makes use of the independence induced by cross-fitting and the rate condition Assumption 12, which we will elaborate on in more detail later.

Here, we first state Lemma 5.10 from Van der Vaart (2000), which will be used in the consistency proof.

Lemma 13. *Let Θ be a subset of the real line and let Ψ_n be random functions and Ψ a fixed function of θ such that $\Psi_n(\theta) \to \Psi(\theta)$ in probability for every θ. Assume that each map $\theta \to \Psi_n(\theta)$ is continuous and has exactly one zero $\widehat{\theta}_n$, or is non-decreasing with $\Psi_n(\widehat{\theta}_n) = o_p(1)$. Let θ_0 be a point such that $\Psi(\theta_0 - \varepsilon) < 0 < \Psi(\theta_0 + \varepsilon)$ for every $\varepsilon > 0$. Then $\widehat{\theta}_n \xrightarrow{p} \theta_0$.*

Lemma 14. *Under Assumptions 9, 11 and 21-23, if either $S^* = S^o$ or $(\pi^*, S_c^*) = (\pi^o, S_c^o)$, then for $\beta \in \mathcal{B}$,*

$$U_{cf}(\beta) \xrightarrow{p} \mu(\beta, \pi^*, S^*, S_c^*), \tag{C.20}$$

$$\frac{\partial}{\partial\beta} U_{cf}(\beta) \xrightarrow{p} -\nu(\beta, \pi^*, S^*, S_c^*), \tag{C.21}$$

where

$$\mu(\beta, \pi, S, S_c) = \int_0^\tau \{\bar{\alpha}(t; \beta^o, \pi, S, S_c) - \bar{\alpha}(t; \beta, \pi, S, S_c)\} s^{(0)}(t; \beta^o, \pi, S, S_c) d\Lambda_0^o(t),$$
$$\nu(\beta, \pi, S, S_c) = \int_0^\tau v(t; \beta, \pi, S, S_c) s^{(0)}(t; \beta^o, \pi, S, S_c) d\Lambda_0^o(t).$$

Lemma 15. *Under Assumptions 9, 11-12 and 21-23,*

$$\sqrt{n} U_{cf}(\beta^o) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \psi_i(\beta^o, \Lambda_0^o, \pi^o, S^o, S_c^o) + o_p(1).$$

Proof of Theorem 11

To show consistency, we make use of Lemma 13. Equation (C.20) of Lemma 14 states that

$$U_{cf}(\beta) \xrightarrow{p} \mu(\beta, \pi^*, S^*, S_c^*),$$

for β in a neighbourhood $\mathcal{B}$ of β^o.

Next, we assume that $\widehat{\beta}$ is a unique zero of $U_{cf}(\beta)$ and β^o is a unique zero of $\mu(\beta, \pi^*, S^*, S_c^*)$. Using Assumption 22, we have $\partial\mu(\beta, \pi^*, S^*, S_c^*)/\partial\beta|_{\beta=\beta^o} = -\nu(\beta^o, \pi^*, S^*, S_c^*) < 0$, and that $\mu(\beta, \pi^*, S^*, S_c^*)$ is continuous for $\beta \in \mathcal{B}$. These conditions together imply that

$$\mu(\beta - \varepsilon, \pi^*, S^*, S_c^*) > 0 > \mu(\beta + \varepsilon, \pi^*, S^*, S_c^*)$$

for any $\varepsilon > 0$.

Lastly, by noting that $U_{cf}(\beta)$ is also continuous in β, we have $\widehat{\beta} \xrightarrow{p} \beta^o$ from applying Lemma 13. □

Proof of Theorem 12

Applying the mean value theorem to $U_{cf}(\beta^o)$ around β^o, we have

$$\sqrt{n}(\widehat{\beta} - \beta^o) = \frac{-\sqrt{n}U_{cf}(\beta^o)}{\frac{\partial}{\partial\beta}U_{cf}(\widetilde{\beta})},$$

where $\widetilde{\beta}$ is some value between $\widehat{\beta}$ and β^o. To show the asymptotic linearity, we find the limit of $\partial U_{cf}(\widetilde{\beta})/\partial\beta$, and write the asymptotic expansions of $\sqrt{n}U_{cf}(\beta^o)$ as averages of i.i.d. terms.

By (C.21) of Lemma 14, we have $\partial U_{cf}(\beta^o)/\partial\beta \xrightarrow{p} -\nu(\beta^o, \pi^o, S^o, S_c^o)$. Using the same arguments as that used in Lemma 14, we also have $\partial U_{cf}(\widetilde{\beta})/\partial\beta - \partial U_{cf}(\beta^o)/\partial\beta = o_p(1)$, so

$$\frac{\partial}{\partial\beta}U_{cf}(\widetilde{\beta}) \xrightarrow{p} -\nu(\beta^o, \pi^o, S^o, S_c^o).$$

The asymptotic expansion of $\sqrt{n}U_{cf}(\beta^o)$ is derived in Lemma 15:

$$\sqrt{n}U_{cf}(\beta^o) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_i(\beta^o, \Lambda_0^o, \pi^o, S^o, S_c^o) + o_p(1).$$

By Assumptions 9 and 22, it's easy to see that $|\psi(\beta^o, \Lambda_0^o, \pi^o, S^o, S_c^o)|$ is bounded a.s., so by the central limit theorem,

$$\sqrt{n}U_{cf}(\beta^o) \xrightarrow{d} N(0, E\{\psi(\beta^o, \Lambda_0^o, \pi^o, S^o, S_c^o)^2\}).$$

Applying Slutsky's Theorem, we therefore have

$$\sqrt{n}(\widehat{\beta} - \beta^o) \xrightarrow{d} N(0, \sigma^2),$$

where $\sigma^2 = E\{\psi(\beta^o, \Lambda_0^o, \pi^o, S^o, S_c^o)^2\}/\nu^2(\beta^o, \pi^o, S^o, S_c^o)$.

Lastly, to show that $\widehat{\sigma}^2$ is a consistent estimator of σ^2, we show separately the convergence of its numerator and its denominator in probability:

$$\frac{1}{n}\sum_{m=1}^{k}\sum_{i\in I_k} \widetilde{\psi}_{m,i}(\widehat{\beta}, \widetilde{\Lambda}_{0,m}(\cdot; \widehat{\beta}, \widehat{\pi}^{(-m)}, \widehat{S}^{(-m)}, \widehat{S}_c^{(-m)}), \widehat{\pi}^{(-m)}, \widehat{S}^{(-m)}, \widehat{S}_c^{(-m)})^2 \xrightarrow{p} E\{\psi(\beta^o, \Lambda_0^o, \pi^o, S^o, S_c^o)^2\},$$
$$\left\{\frac{1}{n}\sum_{m=1}^{k}\sum_{i\in I_k} \int_0^\tau V_m(t; \widehat{\beta}, \widehat{\pi}^{(-m)}, \widehat{S}^{(-m)}, \widehat{S}_c^{(-m)}) d\mathcal{N}_i(t; \widehat{\pi}^{(-m)}, \widehat{S}^{(-m)}, \widehat{S}_c^{(-m)})\right\}^2 \xrightarrow{p} \nu^2(\beta^o, \pi^o, S^o, S_c^o).$$

These can be shown using the same arguments as used in Lemma 14, so we omit the proof here. Applying Slutsky's theorem again, we have

$$\widehat{\sigma}^{-1}\sqrt{n}(\widehat{\beta} - \beta^o) \xrightarrow{d} N(0, 1).$$

□

C.3.3 Proof of lemmas

Since the number of folds k is fixed as $n \to \infty$, to show that results in Lemma 14 hold for the cross-fitted estimating equations U_{cf}, it is sufficient to show that they hold for sample-splitting. Therefore, in the proof of Lemma 14 below, we will show

$$U(\beta, \widehat{\pi}, \widehat{S}, \widehat{S}_c) \xrightarrow{p} \mu(\beta, \pi^*, S^*, S_c^*),$$
$$\frac{\partial}{\partial\beta} U(\beta, \widehat{\pi}, \widehat{S}, \widehat{S}_c) \xrightarrow{p} -\nu(\beta, \pi^*, S^*, S_c^*),$$

where with a slight abuse of notation, we let $\widehat{\pi}, \widehat{S}, \widehat{S}_c$ denote nuisance functions estimated using a different set of data independent from but with the same distribution as the dataset that U is evaluated on. Similarly, in the proof of Lemma 15 below, we will show

$$\sqrt{n}U(\beta^o, \widehat{\pi}, \widehat{S}, \widehat{S}_c) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi_i(\beta^o, \Lambda_0^o, \pi^o, S^o, S_c^o) + o_p(1).$$

Before we begin the proof of the lemmas, note from the strict positivity Assumption 9 that $S^*(t; a, z)$ is bounded away from zero. By the uniform convergence Assumption 11, $\widehat{S}(t; a, z)$ converges to $S^*(t; a, z)$ in probability uniformly in t, so the probability that $\widehat{S}(t; a, z)$ is bounded away from zero goes to one. Same argument also applies to $\widehat{S}_c(t; a, z)$, $\widehat{\pi}(z)$, and $1 - \widehat{\pi}(z)$. We can also derive from (C.17) and (C.18) of Assumption 23 that for nuisance functions π, S, S_c that are either the estimates $\widehat{\pi}, \widehat{S}, \widehat{S}_c$ or their limits, and for $\beta \in \mathcal{B}$, $\mathcal{S}^{(l)}(t; \beta, \pi, S, S_c)$ converges to $s^{(l)}(t; \beta, \pi^*, S^*, S_c^*)$ in probability uniformly in t. Since assumption 22 states that $s^{(l)}(t; \beta, \pi^*, S^*, S_c^*)$ and $1/s^{(0)}(t; \beta, \pi^*, S^*, S_c^*)$ are bounded, so $\mathcal{S}^{(l)}(t; \beta, \pi, S, S_c)$ and $1/\mathcal{S}^{(0)}(t; \beta, \pi, S, S_c)$ are bounded with probability going to one. In the following to simplify the proofs, we will assume WLOG that the quantities are bounded almost surely, and this is due to the conditioning event argument below.

Both Lemmas 14 and 15 claim convergence in probability results. To prove them, we want to show that for some random quantity (i.e. remainder term) X_n and for any $\varepsilon > 0$, $P(|X_n| < \varepsilon) \to 1$ as $n \to \infty$. Let $\mathcal{G}_n$ denote the event that all those terms above are bounded. From Assumptions 9, 11, 22, and 23, we showed earlier that $P(\mathcal{G}_n) \to 1$ as $n \to \infty$. In our approach we first show that $E(|X_n| \,|\, \mathcal{G}_n) \to 0$ as $n \to \infty$, which by Markov' inequality implies that

$$P(|X_n| < \varepsilon \,|\, \mathcal{G}_n) > 1 - \frac{E(|X_n| \,|\, \mathcal{G}_n)}{\varepsilon} \to 1$$

as $n \to \infty$. This leads to

$$\begin{aligned} P(|X_n| < \varepsilon) =& P(|X_n| < \varepsilon \cap \mathcal{G}_n) + P(|X_n| < \varepsilon \cap \mathcal{G}_n^c) \geq P(|X_n| < \varepsilon \cap \mathcal{G}_n) \\ =& P(|X_n| < \varepsilon \,|\, \mathcal{G}_n) P(\mathcal{G}_n) \to 1 \end{aligned}$$

as $n \to \infty$.

Proof of Lemma 14

First, we have

$$U(\beta,\widehat{\pi},\widehat{S},\widehat{S}_c) = U(\beta,\pi^*,S^*,S_c^*) + Q_1 + Q_2 + Q_3,$$

where

$$\begin{aligned}
Q_1 =& U(\beta,\widehat{\pi},\widehat{S},\widehat{S}_c) - U(\beta,\pi^*,\widehat{S},\widehat{S}_c) \\
Q_2 =& U(\beta,\pi^*,\widehat{S},\widehat{S}_c) - U(\beta,\pi^*,S^*,\widehat{S}_c) \\
Q_3 =& U(\beta,\pi^*,S^*,\widehat{S}_c) - U(\beta,\pi^*,S^*,S_c^*).
\end{aligned}$$

We now show that Q_1, Q_2, and Q_3 are $o_p(1)$.

Consider Q_1. We write

$$Q_1 = Q_{11} - Q_{12} - Q_{13}$$

$$\begin{aligned}
Q_{11} =& \frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau} d\mathcal{N}_i^{(1)}(t;\widehat{\pi},\widehat{S},\widehat{S}_c) - d\mathcal{N}_i^{(1)}(t;\pi^*,\widehat{S},\widehat{S}_c) \\
Q_{12} =& \frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau} \left\{\bar{A}(t;\beta,\widehat{\pi},\widehat{S},\widehat{S}_c) - \bar{A}(t;\beta,\pi^*,\widehat{S},\widehat{S}_c)\right\} d\mathcal{N}_i^{(0)}(t;\widehat{\pi},\widehat{S},\widehat{S}_c) \\
Q_{13} =& \frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau} \bar{A}(t;\beta,\pi^*,\widehat{S},\widehat{S}_c)\left\{d\mathcal{N}_i^{(0)}(t;\widehat{\pi},\widehat{S},\widehat{S}_c) - d\mathcal{N}_i^{(0)}(t;\pi^*,\widehat{S},\widehat{S}_c)\right\}.
\end{aligned}$$

First, we note that $d\mathcal{N}_i^{(1)}(t;\pi,S,S_c)$ is a sum of several terms, each term is a product of a term that is bounded a.s. and an increment of a monotone function. Specifically, we have

$$
\begin{aligned}
Q_{11} =& \frac{1}{n}\sum_{i=1}^{n}\left[\frac{1}{\widehat{\pi}(Z_i)^{A_i}\{1-\widehat{\pi}(Z_i)\}^{1-A_i}} - \frac{1}{\pi^*(Z_i)^{A_i}\{1-\pi^*(Z_i)\}^{1-A_i}}\right]\int_0^\tau \frac{A_i}{\widehat{S}_c(t;A_i,Z_i)}dN_i(t) \\
&+\frac{1}{n}\sum_{i=1}^{n}\left[\frac{1}{\widehat{\pi}(Z_i)^{A_i}\{1-\widehat{\pi}(Z_i)\}^{1-A_i}} - \frac{1}{\pi^*(Z_i)^{A_i}\{1-\pi^*(Z_i)\}^{1-A_i}}\right]\int_0^\tau A_i d\widehat{S}(t;A_i,Z_i) \\
&-\frac{1}{n}\sum_{i=1}^{n}A_i\left\{\frac{1}{\widehat{\pi}(Z_i)}-\frac{1}{\pi^*(Z_i)}\right\}\int_0^\tau J_i(t;1,\widehat{S},\widehat{S}_c)d\widehat{S}(t;1,Z_i).
\end{aligned}
$$

This allows us to make use of the following property: for any function $f(t)$, and any monotone function $G(t)$ defined on $[a,b]$, we have

$$
\left|\int_a^b f(t)dG(t)\right| \le \sup_{t\in[a,b]} |f(t)|\cdot|G(b)-G(a)|. \tag{C.22}
$$

Since $N(t)$ and $\widehat{S}(t;a,z)$ are monotone in t, we apply (C.22) to each of the 3 terms in Q_{11}

above and have

$$
\begin{aligned}
&|Q_{11}| \\
\leq &\frac{1}{n}\sum_{i=1}^{n}\left|\frac{1}{\widehat{\pi}(Z_i)^{A_i}\{1-\widehat{\pi}(Z_i)\}^{1-A_i}}-\frac{1}{\pi^*(Z_i)^{A_i}\{1-\pi^*(Z_i)\}^{1-A_i}}\right|\cdot\sup_{t\in[0,\tau]}\left|\frac{A_i}{\widehat{S}_c(t;A_i,Z_i)}\right|\cdot|N_i(\tau)-N_i(0)| \\
&+\frac{1}{n}\sum_{i=1}^{n}\left|\frac{1}{\widehat{\pi}(Z_i)^{A_i}\{1-\widehat{\pi}(Z_i)\}^{1-A_i}}-\frac{1}{\pi^*(Z_i)^{A_i}\{1-\pi^*(Z_i)\}^{1-A_i}}\right|\cdot|A_i|\cdot\left|\widehat{S}(\tau;A_i,Z_i)-\widehat{S}(0;A_i,Z_i)\right| \\
&+\frac{1}{n}\sum_{i=1}^{n}|A_i|\left|\frac{1}{\widehat{\pi}(Z_i)}-\frac{1}{\pi^*(Z_i)}\right|\cdot\sup_{t\in[0,\tau]}\left|J_i(t;1,\widehat{S},\widehat{S}_c)\right|\cdot\left|\widehat{S}(\tau;1,Z_i)-\widehat{S}(0;1,Z_i)\right|. \\
\leq &\frac{1}{n}\sum_{i=1}^{n}\frac{|\widehat{\pi}(Z_i)-\pi^*(Z_i)|}{|\{\widehat{\pi}(Z_i)\pi^*(Z_i)\}^{A_i}[\{1-\widehat{\pi}(Z_i)\}\{1-\pi^*(Z_i)\}]^{1-A_i}|}\cdot\sup_{t\in[0,\tau]}\left|\frac{A_i}{\widehat{S}_c(t;A_i,Z_i)}\right|\cdot|N_i(\tau)-N_i(0)| \\
&+\frac{1}{n}\sum_{i=1}^{n}\frac{|\widehat{\pi}(Z_i)-\pi^*(Z_i)|}{|\{\widehat{\pi}(Z_i)\pi^*(Z_i)\}^{A_i}[\{1-\widehat{\pi}(Z_i)\}\{1-\pi^*(Z_i)\}]^{1-A_i}|}\cdot|A_i|\cdot\left|\widehat{S}(\tau;A_i,Z_i)-\widehat{S}(0;A_i,Z_i)\right| \\
&+\frac{1}{n}\sum_{i=1}^{n}|A_i|\frac{|\widehat{\pi}(Z_i)-\pi^*(Z_i)|}{|\widehat{\pi}(Z_i)\pi^*(Z_i)|}\cdot\sup_{t\in[0,\tau]}\left|J_i(t;1,\widehat{S},\widehat{S}_c)\right|\cdot\left|\widehat{S}(\tau;1,Z_i)-\widehat{S}(0;1,Z_i)\right|.
\end{aligned}
$$

Since $\widehat{S}(t;a,z)$ and $\widehat{S}_c(t;a,z)$ are bounded away from zero a.s., we can again apply (C.22) to

$$
J_i(t;1,\widehat{S},\widehat{S}_c)=\int_0^t\frac{dN_{ci}(u)+Y_i(u)d\log\{\widehat{S}_c(u;1,Z_i)\}}{\widehat{S}(u;1,Z_i)\widehat{S}_c(u;1,Z_i)},
$$

and have

$$
\begin{aligned}
\sup_{t\in[0,\tau]}\left|J_i(t;1,\widehat{S},\widehat{S}_c)\right|\leq &\sup_{t\in[0,\tau]}\left\{\sup_{u\in[0,t]}\left|\frac{1}{\widehat{S}(u;1,Z_i)\widehat{S}_c(u;1,Z_i)}\right|\cdot|N_{ci}(t)-N_{ci}(0)|\right\} \\
&+\sup_{t\in[0,\tau]}\left\{\sup_{u\in[0,t]}\left|\frac{Y_i(u)}{\widehat{S}(u;1,Z_i)\widehat{S}_c(u;1,Z_i)}\right|\cdot\left|\log\{\widehat{S}_c(t;1,Z_i)\}-\log\{\widehat{S}_c(0;1,Z_i)\}\right|\right\} \\
&\lesssim 1. \qquad\qquad (C.23)
\end{aligned}
$$

In addition, since $\widehat{\pi}(z)$ and $1-\widehat{\pi}(z)$ are bounded away from zero a.s., we have

$$|Q_{11}| \lesssim \frac{1}{n}\sum_{i=1}^{n} |\widehat{\pi}(Z_i) - \pi^*(Z_i)|.$$

As a reminder, $E^{\dagger}$ denotes expectations taken with respect to a sample $O^{\dagger}$ of n observations, and E denotes expectations taken with respect to an independent data O conditional on $O^{\dagger}$. O is used for constructing U, while $O^{\dagger}$ is used to estimate $(\widehat{\pi}, \widehat{S}, \widehat{S}_c)$. Using this notation, we have

$$E(|Q_{11}|) \lesssim E^{\dagger}\left[E\left\{|\widehat{\pi}(Z) - \pi^*(Z)|\right\}\right] \leq \|\widehat{\pi} - \pi^*\| = o(1),$$

where the last inequality follows from Jensen's inequality, and the last equality follows from Assumption 11. So we have $Q_{11} = o_p(1)$ from Markov's inequality.

Consider Q_{12}. We again break $d\mathcal{N}_i^{(0)}(t;\pi,S,S_c)$ into a sum of terms, each being a product of a term that is bounded a.s. and an increment of a monotone function.

$$\begin{aligned}
Q_{12} =& \frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau}\left\{\bar{A}(t;\beta,\widehat{\pi},\widehat{S},\widehat{S}_c) - \bar{A}(t;\beta,\pi^*,\widehat{S},\widehat{S}_c)\right\}\cdot\frac{1}{\widehat{\pi}(Z_i)^{A_i}\{1-\widehat{\pi}(Z_i)\}^{1-A_i}\widehat{S}_c(t;A_i,Z_i)}dN_i(t)\\
&+\frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau}\left\{\bar{A}(t;\beta,\widehat{\pi},\widehat{S},\widehat{S}_c) - \bar{A}(t;\beta,\pi^*,\widehat{S},\widehat{S}_c)\right\}\cdot\frac{1}{\widehat{\pi}(Z_i)^{A_i}\{1-\widehat{\pi}(Z_i)\}^{1-A_i}}d\widehat{S}(t;A_i,Z_i)\\
&-\frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau}\left\{\bar{A}(t;\beta,\widehat{\pi},\widehat{S},\widehat{S}_c) - \bar{A}(t;\beta,\pi^*,\widehat{S},\widehat{S}_c)\right\}\\
&\cdot\sum_{a=0,1}\left\{1+\frac{A_i^a(1-A_i)^{1-a}}{\widehat{\pi}(Z_i)^a\{1-\widehat{\pi}(Z_i)\}^{1-a}}J_i(t;a,\widehat{S},\widehat{S}_c)\right\}d\widehat{S}(t;a,Z_i).
\end{aligned}$$

Applying (C.22) and similar arguments as the above, also recall that $\mathcal{S}^{(0)}(t;\beta,\widehat{\pi},\widehat{S},\widehat{S}_c)$ and $\mathcal{S}^{(0)}(t;\beta,\pi^*,\widehat{S},\widehat{S}_c)$ are bounded away from zero a.s. and $\mathcal{S}^{(l)}(t;\beta,\widehat{\pi},\widehat{S},\widehat{S}_c)$ is bounded a.s., we have

$$
\begin{aligned}
|Q_{12}| &\lesssim \sup_{t\in[0,\tau]} \left|\bar{A}(t;\beta,\widehat{\pi},\widehat{S},\widehat{S}_c) - \bar{A}(t;\beta,\pi^*,\widehat{S},\widehat{S}_c)\right| \\
&= \sup_{t\in[0,\tau]} \left| \frac{\mathcal{S}^{(0)}(t;\beta,\widehat{\pi},\widehat{S},\widehat{S}_c)\mathcal{S}^{(1)}(t;\beta,\pi^*,\widehat{S},\widehat{S}_c) - \mathcal{S}^{(1)}(t;\beta,\widehat{\pi},\widehat{S},\widehat{S}_c)\mathcal{S}^{(0)}(t;\beta,\pi^*,\widehat{S},\widehat{S}_c)}{\mathcal{S}^{(0)}(t;\beta,\widehat{\pi},\widehat{S},\widehat{S}_c)\mathcal{S}^{(0)}(t;\beta,\pi^*,\widehat{S},\widehat{S}_c)} \right| \\
&\lesssim \sup_{t\in[0,\tau]} \left|\mathcal{S}^{(0)}(t;\beta,\widehat{\pi},\widehat{S},\widehat{S}_c)\mathcal{S}^{(1)}(t;\beta,\pi^*,\widehat{S},\widehat{S}_c) - \mathcal{S}^{(1)}(t;\beta,\widehat{\pi},\widehat{S},\widehat{S}_c)\mathcal{S}^{(0)}(t;\beta,\pi^*,\widehat{S},\widehat{S}_c)\right| \\
&\leq \sup_{t\in[0,\tau]} \left|\mathcal{S}^{(0)}(t;\beta,\widehat{\pi},\widehat{S},\widehat{S}_c)\{\mathcal{S}^{(1)}(t;\beta,\pi^*,\widehat{S},\widehat{S}_c) - \mathcal{S}^{(1)}(t;\beta,\widehat{\pi},\widehat{S},\widehat{S}_c)\}\right| \\
&\quad + \sup_{t\in[0,\tau]} \left|\mathcal{S}^{(1)}(t;\beta,\widehat{\pi},\widehat{S},\widehat{S}_c)\{\mathcal{S}^{(0)}(t;\beta,\pi^*,\widehat{S},\widehat{S}_c) - \mathcal{S}^{(0)}(t;\beta,\widehat{\pi},\widehat{S},\widehat{S}_c)\}\right| \\
&\lesssim \sum_{l=0,1} \sup_{t\in[0,\tau]} \left|\mathcal{S}^{(l)}(t;\beta,\widehat{\pi},\widehat{S},\widehat{S}_c) - \mathcal{S}^{(l)}(t;\beta,\pi^*,\widehat{S},\widehat{S}_c)\right| \\
&\leq \sum_{l=0,1} \cdot \frac{1}{n}\sum_{i=1}^{n} \sup_{t\in[0,\tau]} \left|\Gamma_i^{(l)}(t;\beta,\widehat{\pi},\widehat{S},\widehat{S}_c) - \Gamma_i^{(l)}(t;\beta,\pi^*,\widehat{S},\widehat{S}_c)\right| \\
&\lesssim \sum_{l=0,1} \cdot \frac{1}{n}\sum_{i=1}^{n} \left| \frac{1}{\widehat{\pi}(Z_i)^{A_i}\{1-\widehat{\pi}(Z_i)\}^{1-A_i}} - \frac{1}{\pi^*(Z_i)^{A_i}\{1-\pi^*(Z_i)\}^{1-A_i}} \right| \\
&\quad + \sum_{l=0,1} \cdot \frac{1}{n}\sum_{i=1}^{n}\sum_{a=0,1} a^l \left| \frac{1}{\widehat{\pi}(Z_i)^{a}\{1-\widehat{\pi}(Z_i)\}^{1-a}} - \frac{1}{\pi^*(Z_i)^{a}\{1-\pi^*(Z_i)\}^{1-a}} \right| \qquad \text{(C.24)} \\
&\lesssim \frac{1}{n}\sum_{i=1}^{n} |\widehat{\pi}(Z_i) - \pi^*(Z_i)|,
\end{aligned}
$$

where (C.24) follows since $\widehat{S}_c(t;A_i,Z_i)$ is bounded away from zero a.s. and $J_i(t;a,\widehat{S},\widehat{S}_c)$ is bounded a.s. following (C.23). Therefore, we again have $E(|Q_{12}|) = o(1)$ from Assumption 11, so $Q_{12} = o_p(1)$ by Markov's inequality.

$Q_{13} = o_p(1)$ can be shown using exactly the same arguments. We therefore have $Q_1 = o_p(1)$.

Next, we show $Q_2 = o_p(1)$. First, we write

$$Q_2 = Q_{21} - Q_{22} - Q_{23}$$

where

$$
\begin{aligned}
Q_{21} =& \frac{1}{n}\sum_{i=1}^{n}\int_{0}^{\tau} d\mathcal{N}_i^{(1)}(t;\pi^*,\widehat{S},\widehat{S}_c) - d\mathcal{N}_i^{(1)}(t;\pi^*,S^*,\widehat{S}_c) \\
Q_{22} =& \frac{1}{n}\sum_{i=1}^{n}\int_{0}^{\tau} \left\{\bar{A}(t;\beta,\pi^*,\widehat{S},\widehat{S}_c) - \bar{A}(t;\beta,\pi^*,S^*,\widehat{S}_c)\right\} d\mathcal{N}_i^{(0)}(t;\pi^*,\widehat{S},\widehat{S}_c) \\
Q_{23} =& \frac{1}{n}\sum_{i=1}^{n}\int_{0}^{\tau} \bar{A}(t;\beta,\pi^*,S^*,\widehat{S}_c)\left\{d\mathcal{N}_i^{(0)}(t;\pi^*,\widehat{S},\widehat{S}_c) - d\mathcal{N}_i^{(0)}(t;\pi^*,S^*,\widehat{S}_c)\right\}.
\end{aligned}
$$

Consider Q_{21}. We have

$$
Q_{21} = Q_{211} - Q_{212} - Q_{213},
$$

where

$$
\begin{aligned}
Q_{211} =& \frac{1}{n}\sum_{i=1}^{n}\frac{A_i\{\widehat{S}(\tau;A_i,Z_i) - S^*(\tau;A_i,Z_i)\}}{\pi^*(Z_i)^{A_i}\{1-\pi^*(Z_i)\}^{1-A_i}} + \frac{1}{n}\sum_{i=1}^{n}\sum_{a=0,1} a\{\widehat{S}(\tau;a,Z_i) - S^*(\tau;a,Z_i)\}, \\
Q_{212} =& \frac{1}{n}\sum_{i=1}^{n}\sum_{a=0,1}\frac{aA_i^a(1-A_i)^{1-a}}{\pi^*(Z_i)^a\{1-\pi^*(Z_i)\}^{1-a}}\int_0^\tau J_i(t;a,\widehat{S},\widehat{S}_c)\{d\widehat{S}(t;a,Z_i) - dS^*(t;a,Z_i)\}, \\
Q_{213} =& \frac{1}{n}\sum_{i=1}^{n}\sum_{a=0,1}\frac{aA_i^a(1-A_i)^{1-a}}{\pi^*(Z_i)^a\{1-\pi^*(Z_i)\}^{1-a}} \\
&\times\int_0^\tau\left[\int_0^t\left\{\frac{1}{\widehat{S}(u;a,Z_i)} - \frac{1}{S^*(u;a,Z_i)}\right\}\frac{dM_{ci}(u;a,\widehat{S}_c)}{\widehat{S}_c(u;a,Z_i)}\right] dS^*(t;a,Z_i).
\end{aligned}
$$

For Q_{211}, we can easily see that

$$
|Q_{211}| \lesssim \frac{1}{n}\sum_{i=1}^{n}\sum_{a=0,1}|\{\widehat{S}(\tau;a,Z_i) - S^*(\tau;a,Z_i)\}| \lesssim \frac{1}{n}\sum_{i=1}^{n}\sup_{t\in[0,\tau],a\in\{0,1\}}\left|\{\widehat{S}(t;a,Z_i) - S^*(t;a,Z_i)\}\right|,
$$

so $E(|Q_{211}|) = o(1)$ by Assumption 11 and $Q_{211} = o_p(1)$ by Markov's inequality.

Term Q_{212} involves a difference in increments $d\widehat{S}(t;a,Z_i) - dS^*(t;a,Z_i)$. Applying integration by parts to the integral term we have

$$\int_0^\tau J_i(t;a,\widehat{S},\widehat{S}_c)\{d\widehat{S}(t;a,Z_i)-dS^*(t;a,Z_i)\}$$
$$=\left[J_i(t;a,\widehat{S},\widehat{S}_c)\{\widehat{S}(t;a,Z_i)-S^*(t;a,Z_i)\}\right]\Big|_0^\tau-\int_0^\tau\frac{\{\widehat{S}(t;a,Z_i)-S^*(t;a,Z_i)\}dM_{ci}(t;a,\widehat{S}_c)}{\widehat{S}(t;a,Z_i)\widehat{S}_c(t;a,Z_i)}, \tag{C.25}$$

So

$$\begin{aligned}Q_{212}=&\frac{1}{n}\sum_{i=1}^n\sum_{a=0,1}\frac{aA_i^a(1-A_i)^{1-a}}{\pi^*(Z_i)^a\{1-\pi^*(Z_i)\}^{1-a}}\left[J_i(t;a,\widehat{S},\widehat{S}_c)\{\widehat{S}(t;a,Z_i)-S^*(t;a,Z_i)\}\right]\Big|_0^\tau\\&-\frac{1}{n}\sum_{i=1}^n\sum_{a=0,1}\frac{aA_i^a(1-A_i)^{1-a}}{\pi^*(Z_i)^a\{1-\pi^*(Z_i)\}^{1-a}}\int_0^\tau\frac{\{\widehat{S}(t;a,Z_i)-S^*(t;a,Z_i)\}dM_{ci}(t;a,\widehat{S}_c)}{\widehat{S}(t;a,Z_i)\widehat{S}_c(t;a,Z_i)}.\end{aligned}$$

Note that $dM_{ci}(t;a,\widehat{S}_c)=dN_{ci}(t)-Y_i(t)d\widehat{\Lambda}_c(t;a,Z_i)$. Since both $N_{ci}(t)$ and $\widehat{\Lambda}_c(t;a,Z_i)$ are monotone functions, we may again apply (C.22) on the second term above. The nuisance functions are bounded away from zero a.s., so we have

$$|Q_{212}|\lesssim\frac{1}{n}\sum_{i=1}^n\sup_{t\in[0,\tau],a\in\{0,1\}}\left|\{\widehat{S}(t;a,Z_i)-S^*(t;a,Z_i)\}\right|,$$

so $E(|Q_{212}|)=o(1)$ from Assumption 11 and $Q_{212}=o_p(1)$ by Markov's inequality.

By applying (C.22) twice on each of the double integrals in Q_{213}, we can show $Q_{213}=o_p(1)$ in exactly the same way.

Same approach used for Q_{21} also gives $Q_{22}=o_p(1)$ and $Q_{23}=o_p(1)$. We hence have $Q_2=o_p(1)$.

$Q_3=o_p(1)$ can again be shown using the same techniques we use for Q_2, so we omit the details.

Lastly, we show that $U(\beta,\pi^*,S^*,S_c^*)=\mu(\beta,\pi^*,S^*,S_c^*)+o_p(1)$ for $\beta\in\mathcal{B}$.

From the definition of the AIPW estimating functions $D_{1i}(t;\beta^o,\Lambda_0^o,\pi,S,S_c)$ and

$D_{2i}(\beta^o, \Lambda_0^o, \pi, S, S_c)$, we have

$$d\mathcal{N}_i^{(0)}(t;\pi,S,S_c) = D_{1i}(t;\beta^o,\Lambda_0^o,\pi,S,S_c) + \Gamma_i^{(0)}(t;\beta^o,\pi,S,S_c)d\Lambda_0^o(t),$$
$$\int_0^\tau d\mathcal{N}_i^{(1)}(t;\pi,S,S_c) = D_{2i}(\beta^o,\Lambda_0^o,\pi,S,S_c) + \int_0^\tau \Gamma_i^{(1)}(t;\beta^o,\pi,S,S_c)d\Lambda_0^o(t).$$

For $\beta \in \mathcal{B}$, we apply this to $U(\beta,\pi^*,S^*,S_c^*)$ and have

$$\begin{aligned}
&U(\beta,\pi^*,S^*,S_c^*)\\
=&\frac{1}{n}\sum_{i=1}^n \int_0^\tau d\mathcal{N}_i^{(1)}(t;\pi^*,S^*,S_c^*) - \bar{A}(t;\beta,\pi^*,S^*,S_c^*)d\mathcal{N}_i^{(0)}(t;\pi^*,S^*,S_c^*)\\
=&\frac{1}{n}\sum_{i=1}^n \Bigg[D_{2i}(\beta^o,\Lambda_0^o,\pi^*,S^*,S_c^*) + \int_0^\tau \Gamma_i^{(1)}(t;\beta^o,\pi^*,S^*,S_c^*)d\Lambda_0^o(t)\\
&-\int_0^\tau \bar{A}(t;\beta,\pi^*,S^*,S_c^*)\{\Gamma_i^{(0)}(t;\beta^o,\pi^*,S^*,S_c^*)d\Lambda_0^o(t) + D_{1i}(t;\beta^o,\Lambda_0^o,\pi^*,S^*,S_c^*)\}\Bigg]\\
=&\int_0^\tau \frac{1}{n}\sum_{i=1}^n\{\Gamma_i^{(1)}(t;\beta^o,\pi^*,S^*,S_c^*) - \bar{A}(t;\beta,\pi^*,S^*,S_c^*)\Gamma_i^{(0)}(t;\beta^o,\pi^*,S^*,S_c^*)\}d\Lambda_0^o(t)\\
&+\frac{1}{n}\sum_{i=1}^n D_{2i}(\beta^o,\Lambda_0^o,\pi^*,S^*,S_c^*)\\
&-\frac{1}{n}\sum_{i=1}^n\int_0^\tau \bar{A}(t;\beta,\pi^*,S^*,S_c^*)D_{1i}(t;\beta^o,\Lambda_0^o,\pi^*,S^*,S_c^*)\\
=&\int_0^\tau \{\mathcal{S}^{(1)}(t;\beta^o,\pi^*,S^*,S_c^*) - \bar{A}(t;\beta,\pi^*,S^*,S_c^*)\mathcal{S}^{(0)}(t;\beta^o,\pi^*,S^*,S_c^*)\}d\Lambda_0^o(t)\\
&+\frac{1}{n}\sum_{i=1}^n D_{2i}(\beta^o,\Lambda_0^o,\pi^*,S^*,S_c^*)\\
&-\frac{1}{n}\sum_{i=1}^n\int_0^\tau \bar{A}(t;\beta,\pi^*,S^*,S_c^*)D_{1i}(t;\beta^o,\Lambda_0^o,\pi^*,S^*,S_c^*)\\
=&\int_0^\tau \left\{\bar{A}(t;\beta^o,\pi^*,S^*,S_c^*) - \bar{A}(t;\beta,\pi^*,S^*,S_c^*)\right\}\mathcal{S}^{(0)}(t;\beta^o,\pi^*,S^*,S_c^*)d\Lambda_0^o(t)\\
&+\frac{1}{n}\sum_{i=1}^n D_{2i}(\beta^o,\Lambda_0^o,\pi^*,S^*,S_c^*)\\
&-\frac{1}{n}\sum_{i=1}^n\int_0^\tau \bar{A}(t;\beta,\pi^*,S^*,S_c^*)D_{1i}(t;\beta^o,\Lambda_0^o,\pi^*,S^*,S_c^*).
\end{aligned}$$

Next, for each of $\bar{A}$ and $\mathcal{S}^{(0)}$, we add and subtract its limits and have

$$
\begin{align}
& U(\beta,\pi^*,S^*,S_c^*) \notag\\
= \;& \mu(\beta^o,\pi^*,S^*,S_c^*) \notag\\
& +\int_0^\tau \{\bar{A}(t;\beta^o,\pi^*,S^*,S_c^*)-\bar{A}(t;\beta,\pi^*,S^*,S_c^*)-\bar{\alpha}(t;\beta^o,\pi^*,S^*,S_c^*)+\bar{\alpha}(t;\beta,\pi^*,S^*,S_c^*)\} \notag\\
& \quad \times \mathcal{S}^{(0)}(t;\beta^o,\pi^*,S^*,S_c^*)d\Lambda_0^o(t) \tag{C.26}\\
& +\int_0^\tau \{\bar{\alpha}(t;\beta^o,\pi^*,S^*,S_c^*)-\bar{\alpha}(t;\beta,\pi^*,S^*,S_c^*)\} \notag\\
& \quad \times \{\mathcal{S}^{(0)}(t;\beta^o,\pi^*,S^*,S_c^*)-s^{(0)}(t;\beta^o,\pi^*,S^*,S_c^*)\}d\Lambda_0^o(t) \tag{C.27}\\
& +\frac{1}{n}\sum_{i=1}^n D_{2i}(\beta^o,\Lambda_0^o,\pi^*,S^*,S_c^*) \tag{C.28}\\
& -\frac{1}{n}\sum_{i=1}^n \int_0^\tau \bar{\alpha}(t;\beta,\pi^*,S^*,S_c^*)D_{1i}(t;\beta^o,\Lambda_0^o,\pi^*,S^*,S_c^*) \tag{C.29}\\
& +\frac{1}{n}\sum_{i=1}^n \int_0^\tau \{\bar{\alpha}(t;\beta,\pi^*,S^*,S_c^*)-\bar{A}(t;\beta,\pi^*,S^*,S_c^*)\}D_{1i}(t;\beta^o,\Lambda_0^o,\pi^*,S^*,S_c^*). \tag{C.30}
\end{align}
$$

For (C.26), since $\Lambda_0^o(t)$ is an increasing function and $\mathcal{S}^{(0)}(t;\beta^o,\pi^*,S^*,S_c^*)$ is bounded a.s., we can apply (C.22) to it and have

$$
\begin{aligned}
& \left|\int_0^\tau \{\bar{A}(t;\beta^o,\pi^*,S^*,S_c^*)-\bar{A}(t;\beta,\pi^*,S^*,S_c^*)-\bar{\alpha}(t;\beta^o,\pi^*,S^*,S_c^*)+\bar{\alpha}(t;\beta,\pi^*,S^*,S_c^*)\}\right.\\
& \quad \left.\times \mathcal{S}^{(0)}(t;\beta^o,\pi^*,S^*,S_c^*)d\Lambda_0^o(t)\right|\\
\lesssim \;& \sup_{t\in[0,\tau]}\left|\bar{A}(t;\beta^o,\pi^*,S^*,S_c^*)-\bar{A}(t;\beta,\pi^*,S^*,S_c^*)-\bar{\alpha}(t;\beta^o,\pi^*,S^*,S_c^*)+\bar{\alpha}(t;\beta,\pi^*,S^*,S_c^*)\right|,
\end{aligned}
$$

which is $o_p(1)$ from Assumption 21. Similarly, (C.27) is $o_p(1)$.

Next, we note that the increments in $D_{1i}(t;\beta^o,\Lambda_0^o,\pi^*,S^*,S_c^*)$ are $dN_i(t)$, $dS^*(t;A_i,Z_i)$, $dS^*(t;a,Z_i)$ and $d\Lambda_0(t)$, all of which are increments of monotone functions. So similar to (C.26) and (C.27), we can apply (C.22), the strict positivity Assumptions 9 and Assumption 21 to show that (C.30) is $o_p(1)$.

Since we have $S^*=S^o$ or $(\pi^*,S_c^*)=(\pi^o,S_c^o)$, Theorem 10 gives that both (C.28) and (C.29) are sums of i.i.d. mean zero terms. The strict positivity Assumption 9 ensures that these i.i.d. mean zero terms are also bounded, hence having bounded variance. So (C.28) $= o_p(1)$ and (C.29) $= o_p(1)$ by the weak law of large numbers.

The second part of the Lemma,

$$\frac{\partial}{\partial\beta}U(\beta,\widehat{\pi},\widehat{S},\widehat{S}_c) \xrightarrow{p} -\nu(\beta,\pi^*,S^*,S_c^*),$$

can be shown using exactly the same arguments as how we proved $U(\beta,\widehat{\pi},\widehat{S},\widehat{S}_c) \xrightarrow{p} \mu(\beta,\pi^*,S^*,S_c^*)$ above, which completes the proof.

□

Proof of Lemma 15

First, write

$$\sqrt{n}U(\beta^o,\widehat{\pi},\widehat{S},\widehat{S}_c) = \sqrt{n}U(\beta^o,\pi^o,S^o,S_c^o) + Q_4 + Q_5 + Q_6,$$

where

$$\begin{aligned}
Q_4 =& \sqrt{n}\{U(\beta^o,\widehat{\pi},\widehat{S},\widehat{S}_c) - U(\beta^o,\pi^o,\widehat{S},S_c^o)\} - \sqrt{n}\{U(\beta^o,\widehat{\pi},S^o,\widehat{S}_c) - U(\beta^o,\pi^o S^o,S_c^o)\},\\
Q_5 =& \sqrt{n}\{U(\beta^o,\widehat{\pi},S^o,\widehat{S}_c) - U(\beta^o,\pi^o,S^o,S_c^o)\},\\
Q_6 =& \sqrt{n}\{U(\beta^o,\pi^o,\widehat{S},S_c^o) - U(\beta^o,\pi^o,S^o,S_c^o)\}.
\end{aligned}$$

The structure of the proof goes as follows: we first show using the rate condition Assumption 12 among other assumptions that Q_4, which is a difference in differences, is $o_p(1)$. Next, we show that Q_5 and Q_6 are $o_p(1)$, which uses, among other assumptions, the independence between in-fold and out-of-fold data induced by cross-fitting. Finally, we show that $\sqrt{n}U(\beta^o,\pi^o,S^o,S_c^o)$ is asymptotically equivalent to a sum of i.i.d. terms.

We first show that $Q_4 = o_p(1)$. For any fixed nuisance function S, we have

$$\begin{aligned}
&\sqrt{n}\{U_1(\beta^o,\widehat{\pi},S,\widehat{S}_c) - U_1(\beta^o,\pi^o,S,S_c^o)\}\\
=&\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\int_0^{\tau} d\mathcal{N}_i^{(1)}(t;\widehat{\pi},S,\widehat{S}_c) - d\mathcal{N}_i^{(1)}(t;\pi^o,S,S_c^o)\\
&-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\int_0^{\tau} \bar{A}(t;\beta^o,\pi^o,S,S_c^o)\{d\mathcal{N}_i^{(0)}(t;\widehat{\pi},S,\widehat{S}_c) - d\mathcal{N}_i^{(0)}(t;\pi^o,S,S_c^o)\}\\
&-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\int_0^{\tau} \{\bar{A}(t;\beta^o,\widehat{\pi},S,\widehat{S}_c) - \bar{A}(t;\beta^o,\pi^o,S,S_c^o)\}d\mathcal{N}_i^{(0)}(t;\widehat{\pi},S,\widehat{S}_c).
\end{aligned}$$

So we can write

$$Q_4 = Q_{41} - Q_{42} - Q_{43} - Q_{44},$$

where

$$\begin{aligned}
Q_{41} =&\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\int_0^{\tau} d\mathcal{N}_i^{(1)}(t;\widehat{\pi},\widehat{S},\widehat{S}_c) - d\mathcal{N}_i^{(1)}(t;\pi^o,\widehat{S},S_c^o) - d\mathcal{N}_i^{(1)}(t;\widehat{\pi},S^o,\widehat{S}_c) + d\mathcal{N}_i^{(1)}(t;\pi^o,S^o,S_c^o)\\
&-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\int_0^{\tau} \bar{A}(t;\beta^o,\pi^o,S^o,S_c^o)\\
&\times\{d\mathcal{N}_i^{(0)}(t;\widehat{\pi},\widehat{S},\widehat{S}_c) - d\mathcal{N}_i^{(0)}(t;\pi^o,\widehat{S},S_c^o) - d\mathcal{N}_i^{(0)}(t;\widehat{\pi},S^o,\widehat{S}_c) + d\mathcal{N}_i^{(0)}(t;\pi^o,S^o,S_c^o)\}
\end{aligned}$$

$$Q_{42} =\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\int_0^{\tau} \{\bar{A}(t;\beta^o,\widehat{\pi},S^o,\widehat{S}_c) - \bar{A}(t;\beta^o,\pi^o,S^o,S_c^o)\}\{d\mathcal{N}_i^{(0)}(t;\widehat{\pi},\widehat{S},\widehat{S}_c) - d\mathcal{N}_i^{(0)}(t;\widehat{\pi},S^o,\widehat{S}_c)\}$$

$$Q_{43} =\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\int_0^{\tau} \{\bar{A}(t;\beta^o,\pi^o,\widehat{S},S_c^o) - \bar{A}(t;\beta^o,\pi^o,S^o,S_c^o)\}\{d\mathcal{N}_i^{(0)}(t;\widehat{\pi},\widehat{S},\widehat{S}_c) - d\mathcal{N}_i^{(0)}(t;\pi^o,\widehat{S},S_c^o)\}$$

$$\begin{aligned}
Q_{44} =&\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\int_0^{\tau} d\mathcal{N}_i^{(0)}(t;\widehat{\pi},\widehat{S},\widehat{S}_c)\\
&\times\{\bar{A}(t;\beta^o,\widehat{\pi},\widehat{S},\widehat{S}_c) - \bar{A}(t;\beta^o,\pi^o,\widehat{S},S_c^o) - \bar{A}(t;\beta^o,\widehat{\pi},S^o,\widehat{S}_c) + \bar{A}(t;\beta^o,\pi^o,S^o,S_c^o)\}.
\end{aligned}$$

Consider Q_{41}, which can be written as $Q_{41} = -Q_{411} + Q_{412} - Q_{413} - Q_{414}$, where

$$\begin{aligned}
Q_{411} =& \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \sum_{a=0,1} \frac{A_i^a (1-A_i)^{1-a}}{\pi^o(Z_i)^a \{1-\pi^o(Z_i)\}^{1-a}} \int_0^\tau \left[\{a - \bar{A}(t;\beta^o,\pi^o,S^o,S_c^o)\} \right. \\
& \left. \times \int_0^t \left\{ \frac{d\widehat{S}(t;a,Z_i)}{\widehat{S}(u;a,Z_i)} - \frac{dS^o(t;a,Z_i)}{S^o(u;a,Z_i)} \right\} \left\{ \frac{dM_{ci}(u;a,\widehat{S}_c)}{\widehat{S}_c(u;a,Z_i)} - \frac{dM_{ci}(u;a,S_c^o)}{S_c^o(u;a,Z_i)} \right\} \right] \qquad \text{(C.31)} \\
Q_{412} =& \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{A_i - \bar{A}(\tau;\beta^o,\pi^o,S^o,S_c^o)\} \left\{ \frac{1}{\widehat{\pi}(Z_i)^{A_i}\{1-\widehat{\pi}(Z_i)\}^{1-A_i}} - \frac{1}{\pi^o(Z_i)^{A_i}\{1-\pi^o(Z_i)\}^{1-A_i}} \right\} \\
& \times \left\{ \widehat{S}(\tau;A_i,Z_i) - S^o(\tau;A_i,Z_i) \right\} \\
Q_{413} =& \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \sum_{a=0,1} A_i^a (1-A_i)^{1-a} \left\{ \frac{1}{\widehat{\pi}(Z_i)^a\{1-\widehat{\pi}(Z_i)\}^{1-a}} - \frac{1}{\pi^o(Z_i)^a\{1-\pi^o(Z_i)\}^{1-a}} \right\} \\
& \times \int_0^\tau dS^o(t;a,Z_i)\{a - \bar{A}(t;\beta^o,\pi^o,S^o,S_c^o)\} \int_0^t \left\{ \frac{1}{\widehat{S}(u;a,Z_i)} - \frac{1}{S^o(u;a,Z_i)} \right\} \frac{dM_{ci}(u;a,\widehat{S}_c)}{\widehat{S}_c(u;a,Z_i)} \\
Q_{414} =& \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \sum_{a=0,1} A_i^a (1-A_i)^{1-a} \left\{ \frac{1}{\widehat{\pi}(Z_i)^a\{1-\widehat{\pi}(Z_i)\}^{1-a}} - \frac{1}{\pi^o(Z_i)^a\{1-\pi^o(Z_i)\}^{1-a}} \right\} \\
& \times \int_0^\tau \{d\widehat{S}(t;a,Z_i) - dS^o(t;a,Z_i)\}\{a - \bar{A}(t;\beta^o,\pi^o,S^o,S_c^o)\} J_i(t;a,\widehat{S},\widehat{S}_c).
\end{aligned}$$

Recall that $\mathcal{D}^\dagger(\widehat{S},\widehat{S}_c;S^o,S_c^o)$ defined in Assumption 12 is made up of two terms, which we will denote as $\mathcal{D}^\dagger(\widehat{S},\widehat{S}_c;S^o,S_c^o) = \mathcal{D}_1^\dagger + \mathcal{D}_2^\dagger$, where

$$\begin{aligned}
\mathcal{D}_1^\dagger =& E^\dagger \left\{ E \left[\max_{a\in\{0,1\}} \left| \int_0^\tau \{a - \bar{A}(t;\beta^o,\pi^o,S^o,S_c^o)\} \right. \right. \right. \\
& \left. \left. \left. \times \int_0^t \left\{ \frac{d\widehat{S}(t;a,Z)}{\widehat{S}(u;a,Z)} - \frac{dS^o(t;a,Z)}{S^o(u;a,Z)} \right\} \left\{ \frac{dM_c(u;a,Z,\widehat{S}_c)}{\widehat{S}_c(u;a,Z)} - \frac{dM_c(u;a,Z,S_c^o)}{S_c^o(u;a,Z)} \right\} \right| \right] \right\} \\
\mathcal{D}_2^\dagger =& E^\dagger \left\{ E \left[\max_{a,l\in\{0,1\}} \left| \int_0^\tau \{\bar{A}(t;\beta^o,\pi^o,S^o,\widehat{S}_c) - \bar{A}(t;\beta^o,\pi^o,S^o,S_c^o)\} \right. \right. \right. \\
& \left. \left. \left. \times J(t;a,Z,S^o,\widehat{S}_c)^l \{d\widehat{S}(t;a,Z) - dS^o(t;a,Z)\} \right| \right] \right\}.
\end{aligned}$$

For Q_{411}, we first notice that by the strict positivity Assumption 9, $A_i^a(1 -$

$A_i)^{1-a}/\{\pi^o(Z_i)^a\{1-\pi^o(Z_i)\}^{1-a}\}$ is bounded a.s.. The expectation of the absolute value of the double integral in Q_{411} can be bounded directly using $\mathcal{D}_1^\dagger$, which leads to

$$E(|Q_{411}|) \lesssim \sqrt{n}\mathcal{D}_1^\dagger = o(1),$$

where the last equality follows from rate condition Assumption 12.

Integral remainders $\mathcal{D}_1^\dagger$ is specific to our case because both nuisance functions $S(t;a,z)$ and $S_c(t;a,z)$ are time-dependent, which can lead to a product between the differences $\widehat{S}_c - S_c$ and differences of increments $d\widehat{S} - dS^o$, like in (C.31). To the best of our knowledge, remainder terms like this can not be sufficiently controlled using existing tools, which requires us to make additional assumptions, such as $\mathcal{D}^\dagger(\widehat{S},\widehat{S}_c;S^o,S_c^o) = o(n^{-1/2})$ in the rate condition Assumption 12. More discussion and references on this can be found in Section 4.6

For Q_{412}, recall that $A_i - \bar{A}(\tau;\beta^o,\pi^o,S^o,S_c^o)$ is bounded a.s. and $\widehat{\pi}(Z_i)$, $\pi^o(Z_i)$, $1-\widehat{\pi}(Z_i)$ and $1-\pi^o(Z_i)$ are bounded away from zero a.s., so we have

$$\begin{aligned}|Q_{412}| \leq & \frac{1}{\sqrt{n}}\sum_{i=1}^n |A_i - \bar{A}(\tau;\beta^o,\pi^o,S^o,S_c^o)| \cdot \frac{|\widehat{\pi}(Z_i) - \pi^o(Z_i)| \cdot \left|\widehat{S}(\tau;A_i,Z_i) - S^o(\tau;A_i,Z_i)\right|}{|\widehat{\pi}(Z_i)^{A_i}\{1-\widehat{\pi}(Z_i)\}^{1-A_i}\pi^o(Z_i)^{A_i}\{1-\pi^o(Z_i)\}^{1-A_i}|} \\ \lesssim & \frac{1}{\sqrt{n}}\sum_{i=1}^n |\widehat{\pi}(Z_i) - \pi^o(Z_i)| \cdot \sup_{t\in[0,\tau],a\in\{0,1\}} \left|\widehat{S}(t;a,Z_i) - S^o(t;a,Z_i)\right|.\end{aligned}$$

Therefore

$$\begin{aligned}E(|Q_{412}|) \lesssim & \sqrt{n}E^\dagger\left\{E\left[|\widehat{\pi}(Z) - \pi^o(Z)| \cdot \sup_{t\in[0,\tau],a\in\{0,1\}} \left|\widehat{S}(t;a,Z) - S^o(t;a,Z)\right|\right]\right\} \\ \leq & \sqrt{n}\,\|\widehat{\pi} - \pi^o\| \cdot \left\|\widehat{S} - S^o\right\| \qquad \text{(C.32)} \\ = & o(1), \qquad \text{(C.33)}\end{aligned}$$

where (C.32) follows from the Cauchy-Schwartz inequality $|E(AB)|^2 \leq E(A^2)E(B^2)$, while

(C.33) uses the rate condition Assumption 12.

Q_{413} can be bounded similarly with the help of (C.22). First we note that $S^o(t;a,Z_i)$ is a monotone function by assumption. Recall that $dM_{ci}(u;a,\widehat{S}_c) = dN_{ci}(u) - Y_i(u)d\widehat{\Lambda}_c(u;a,Z_i)$ is also a sum of two terms, each being the product of a term bounded a.s. and an increment of a monotone function. We therefore apply (C.22) twice to each of the double integral in Q_{413} and have

$$\begin{aligned}
|Q_{413}| \lesssim & \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\sum_{a=0,1}\frac{|\widehat{\pi}(Z_i)-\pi^o(Z_i)|}{|\widehat{\pi}(Z_i)^a\{1-\widehat{\pi}(Z_i)\}^{1-a}\pi^o(Z_i)^a\{1-\pi^o(Z_i)\}^{1-a}|} \\
& \times \sup_{t\in[0,\tau]}\left|\{a-\bar{A}(t;\beta^o,\pi^o,S^o,S_c^o)\}\int_0^t\left\{\frac{1}{\widehat{S}(u;a,Z_i)}-\frac{1}{S^o(u;a,Z_i)}\right\}\frac{dM_{ci}(u;a,\widehat{S}_c)}{\widehat{S}_c(u;a,Z_i)}\right| \\
\lesssim & \frac{1}{\sqrt{n}}\sum_{i=1}^{n}|\widehat{\pi}(Z_i)-\pi^o(Z_i)| \\
& \times \sup_{t\in[0,\tau],a\in\{0,1\}}\left\{\sup_{u\in[0,t]}\left|\frac{\widehat{S}(u;a,Z_i)-S^o(u;a,Z_i)}{\widehat{S}(u;a,Z_i)S^o(u;a,Z_i)\widehat{S}_c(u;a,Z_i)}\right|\right\} \\
\lesssim & \frac{1}{\sqrt{n}}\sum_{i=1}^{n}|\widehat{\pi}(Z_i)-\pi^o(Z_i)|\cdot\sup_{t\in[0,\tau],a\in\{0,1\}}\left|\widehat{S}(t;a,Z_i)-S^o(t;a,Z_i)\right|.
\end{aligned}$$

So we again have

$$\begin{aligned}
E(|Q_{413}|) \lesssim & \sqrt{n}E^{\dagger}\left\{E\left[|\widehat{\pi}(Z)-\pi^o(Z)|\cdot\sup_{t\in[0,\tau],a\in\{0,1\}}\left|\widehat{S}(t;a,Z)-S^o(t;a,Z)\right|\right]\right\} \\
\leq & \sqrt{n}\,\|\widehat{\pi}-\pi^o\|\cdot\left\|\widehat{S}-S^o\right\| \\
= & o(1)
\end{aligned}$$

from the Cauchy-Schwartz inequality and the rate condition Assumption 12.

The integral in Q_{414} involves a difference in increments $d\widehat{S}(t;a,Z_i) - dS^o(t;a,Z_i)$, so we apply integration by parts and have

$$
\begin{aligned}
&\int_0^\tau \{d\widehat{S}(t;a,Z_i) - dS^o(t;a,Z_i)\}\{a - \bar{A}(t;\beta^o,\pi^o,S^o,S_c^o)\}J_i(t;a,\widehat{S},\widehat{S}_c) \\
=&\Big[\{\widehat{S}(t;a,Z_i) - S^o(t;a,Z_i)\}\{a - \bar{A}(t;\beta^o,\pi^o,S^o,S_c^o)\}J_i(t;a,\widehat{S},\widehat{S}_c)\Big]\Big|_0^\tau \\
&-\int_0^\tau \{\widehat{S}(t;a,Z_i) - S^o(t;a,Z_i)\}\frac{\partial}{\partial t}\Big[\{a - \bar{A}(t;\beta^o,\pi^o,S^o,S_c^o)\}J_i(t;a,\widehat{S},\widehat{S}_c)\Big] \\
=&\{\widehat{S}(\tau;a,Z_i) - S^o(\tau;a,Z_i)\}\{a - \bar{A}(\tau;\beta^o,\pi^o,S^o,S_c^o)\}J_i(\tau;a,\widehat{S},\widehat{S}_c) \\
&-\int_0^\tau \{\widehat{S}(t;a,Z_i) - S^o(t;a,Z_i)\}\{a - \bar{A}(t;\beta^o,\pi^o,S^o,S_c^o)\}\cdot\frac{dM_{ci}(t;a,\widehat{S}_c)}{\widehat{S}(t;a,Z_i)\widehat{S}_c(t;a,Z_i)} \\
&+\int_0^\tau \{\widehat{S}(t;a,Z_i) - S^o(t;a,Z_i)\}\frac{\partial}{\partial t}\left[\frac{\mathcal{S}^{(1)}(t;\beta^o,\pi^o,S^o,S_c^o)}{\mathcal{S}^{(0)}(t;\beta^o,\pi^o,S^o,S_c^o)}\right]\cdot J_i(t;a,\widehat{S},\widehat{S}_c) \\
=&\{\widehat{S}(\tau;a,Z_i) - S^o(\tau;a,Z_i)\}\{a - \bar{A}(\tau;\beta^o,\pi^o,S^o,S_c^o)\}J_i(\tau;a,\widehat{S},\widehat{S}_c) \\
&\int_0^\tau \{\widehat{S}(t;a,Z_i) - S^o(t;a,Z_i)\}\{a - \bar{A}(t;\beta^o,\pi^o,S^o,S_c^o)\}\cdot\frac{dM_{ci}(t;a,\widehat{S}_c)}{\widehat{S}(t;a,Z_i)\widehat{S}_c(t;a,Z_i)} \\
&+\int_0^\tau \{\widehat{S}(t;a,Z_i) - S^o(t;a,Z_i)\}J_i(t;a,\widehat{S},\widehat{S}_c)\frac{1}{\mathcal{S}^{(0)}(t;\beta^o,\pi^o,S^o,S_c^o)}\cdot\frac{1}{n}\sum_{j=1}^n d\Gamma_j^{(1)}(t;\beta^o,\pi^o,S^o,S_c^o) \\
&-\int_0^\tau \{\widehat{S}(t;a,Z_i) - S^o(t;a,Z_i)\}J_i(t;a,\widehat{S},\widehat{S}_c)\frac{\mathcal{S}^{(1)}(t;\beta^o,\pi^o,S^o,S_c^o)}{\mathcal{S}^{(0)}(t;\beta^o,\pi^o,S^o,S_c^o)^2}\cdot\frac{1}{n}\sum_{j=1}^n d\Gamma_j^{(0)}(t;\beta^o,\pi^o,S^o,S_c^o),
\end{aligned}
\tag{C.34}
$$

where the last two equalities follow from the product rule. For $l = 0,1$, we again apply the product rule and have

$$
\begin{aligned}
&d\Gamma_j^{(l)}(t;\beta^o,\pi^o,S^o,S_c^o) \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{(C.35)}\\
=&\frac{A_j^l e^{\beta^o A_j}}{\pi^o(Z_j)^{A_j}\{1-\pi^o(Z_j)\}^{1-A_j}S_c^o(t;A_j,Z_j)}dY_j(t)\\
&-\frac{A_j^l Y_j(t)e^{\beta^o A_j}}{\pi^o(Z_j)^{A_j}\{1-\pi^o(Z_j)\}^{1-A_j}S_c^o(t;A_j,Z_j)^2}dS_c^o(t;A_j,Z_j)\\
&-\frac{A_j^l e^{\beta^o A_j}}{\pi^o(Z_j)^{A_j}\{1-\pi^o(Z_j)\}^{1-A_j}}dS^o(t;A_j,Z_j)\\
&+\sum_{a=0,1}a^l\left\{1+\frac{A_j^a(1-A_j)^{1-a}}{\pi^o(Z_j)^a\{1-\pi^o(Z_j)\}^{1-a}}J_j(t;a,S^o,S_c^o)\right\}e^{\beta^o a}dS^o(t;a,Z_j)\\
&+\sum_{a=0,1}a^l\frac{A_j^a(1-A_j)^{1-a}}{\pi^o(Z_j)^a\{1-\pi^o(Z_j)\}^{1-a}}\frac{S^o(t;a,Z_j)e^{\beta^o a}}{S^o(t;a,Z_j)S_c^o(t;a,Z_j)}dM_{cj}(u;a,S_c^o).
\end{aligned}
$$

Since $dM_{cj}(u;a,S_c^o) = dN_{cj}(u) - Y_j(u)d\Lambda_c^o(u;a,Z_j)$, we can now see that $d\Gamma_j^{(l)}(t;\beta^o,\pi^o,S^o,S_c^o)$ is once again a sum of terms, each being a product between a term that is bounded a.s. and an increment of a monotone function. Therefore, applying (C.22), we have

$$
\begin{aligned}
&\left|\int_0^\tau\{d\widehat{S}(t;a,Z_i)-dS^o(t;a,Z_i)\}\{a-\bar{A}(t;\beta^o,\pi^o,S^o,S_c^o)\}J_i(t;a,\widehat{S},\widehat{S}_c)\right|\\
\lesssim&|\widehat{S}(\tau;a,Z_i)-S^o(\tau;a,Z_i)|\\
&+\sup_{t\in[0,\tau],a\in\{0,1\}}\left|\widehat{S}(t;a,Z_i)-S^o(t;a,Z_i)\right|\\
&+\frac{1}{n}\sum_{j=1}^n\sup_{t\in[0,\tau],a\in\{0,1\}}\left|\widehat{S}(t;a,Z_i)-S^o(t;a,Z_i)\right|\\
&+\frac{1}{n}\sum_{j=1}^n\sup_{t\in[0,\tau],a\in\{0,1\}}\left|\widehat{S}(t;a,Z_i)-S^o(t;a,Z_i)\right|\\
\lesssim&\sup_{t\in[0,\tau],a\in\{0,1\}}\left|\widehat{S}(t;a,Z_i)-S^o(t;a,Z_i)\right|.
\end{aligned}
$$

So

$$|Q_{414}| \lesssim \frac{1}{\sqrt{n}} \sum_{i=1}^{n} |\widehat{\pi}(Z_i) - \pi^o(Z_i)| \cdot \sup_{t\in[0,\tau], a\in\{0,1\}} \left|\widehat{S}(t;a,Z_i) - S^o(t;a,Z_i)\right|,$$

and we again have

$$\begin{aligned} E(|Q_{414}|) \lesssim & \sqrt{n} E^{\dagger}\left\{ E\left[|\widehat{\pi}(Z) - \pi^o(Z)| \cdot \sup_{t\in[0,\tau], a\in\{0,1\}} \left|\widehat{S}(t;a,Z) - S^o(t;a,Z)\right| \right] \right\} \\ \leq & \sqrt{n} \|\widehat{\pi} - \pi^o\| \cdot \left\|\widehat{S} - S^o\right\| \\ = & o(1), \end{aligned}$$

from the Cauchy-Schwartz inequality and the rate condition Assumption 12.

Therefore, we have

$$E(|Q_{41}|) \leq E(|Q_{411}|) + E(|Q_{412}|) + E(|Q_{413}|) + E(|Q_{414}|) \lesssim o(1),$$

so $Q_{41} = o_p(1)$ by Markov's inequality.

Next, we bound Q_{42}, which involves the use of $\mathcal{D}_2^{\dagger}$. First, we let $Q_{42} = Q_{421} + Q_{422}$, where

$$\begin{aligned} Q_{421} = & \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \int_0^{\tau} \{\bar{A}(t;\beta^o, \widehat{\pi}, S^o, \widehat{S}_c) - \bar{A}(t;\beta^o, \pi^o, S^o, \widehat{S}_c)\} \\ & \times \{d\mathcal{N}_i^{(0)}(t;\widehat{\pi}, \widehat{S}, \widehat{S}_c) - d\mathcal{N}_i^{(0)}(t;\widehat{\pi}, S^o, \widehat{S}_c)\} \\ Q_{422} = & \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \int_0^{\tau} \{\bar{A}(t;\beta^o, \pi^o, S^o, \widehat{S}_c) - \bar{A}(t;\beta^o, \pi^o, S^o, S_c^o)\} \\ & \times \{d\mathcal{N}_i^{(0)}(t;\widehat{\pi}, \widehat{S}, \widehat{S}_c) - d\mathcal{N}_i^{(0)}(t;\widehat{\pi}, S^o, \widehat{S}_c)\}. \end{aligned}$$

Note that like how we bounded $|Q_{12}|$ earlier, we also have

$$
\begin{aligned}
&\bar{A}(t;\beta^o,\widehat{\pi},S^o,\widehat{S}_c)-\bar{A}(t;\beta^o,\pi^o,S^o,\widehat{S}_c)\\
=&\frac{\mathcal{S}^{(1)}(t;\beta^o,\widehat{\pi},S^o,\widehat{S}_c)\mathcal{S}^{(0)}(t;\beta^o,\pi^o,S^o,\widehat{S}_c)-\mathcal{S}^{(1)}(t;\beta^o,\pi^o,S^o,\widehat{S}_c)\mathcal{S}^{(0)}(t;\beta^o,\widehat{\pi},S^o,\widehat{S}_c)}{\mathcal{S}^{(0)}(t;\beta^o,\widehat{\pi},S^o,\widehat{S}_c)\mathcal{S}^{(0)}(t;\beta^o,\pi^o,S^o,\widehat{S}_c)}\\
=&\frac{\{\mathcal{S}^{(1)}(t;\beta^o,\widehat{\pi},S^o,\widehat{S}_c)-\mathcal{S}^{(1)}(t;\beta^o,\pi^o,S^o,\widehat{S}_c)\}\mathcal{S}^{(0)}(t;\beta^o,\pi^o,S^o,\widehat{S}_c)}{\mathcal{S}^{(0)}(t;\beta^o,\widehat{\pi},S^o,\widehat{S}_c)\mathcal{S}^{(0)}(t;\beta^o,\pi^o,S^o,\widehat{S}_c)}\\
&-\frac{\mathcal{S}^{(1)}(t;\beta^o,\pi^o,S^o,\widehat{S}_c)\{\mathcal{S}^{(0)}(t;\beta^o,\widehat{\pi},S^o,\widehat{S}_c)-\mathcal{S}^{(0)}(t;\beta^o,\pi^o,S^o,\widehat{S}_c)\}}{\mathcal{S}^{(0)}(t;\beta^o,\widehat{\pi},S^o,\widehat{S}_c)\mathcal{S}^{(0)}(t;\beta^o,\pi^o,S^o,\widehat{S}_c)}\\
=&\frac{\mathcal{S}^{(0)}(t;\beta^o,\pi^o,S^o,\widehat{S}_c)}{\mathcal{S}^{(0)}(t;\beta^o,\widehat{\pi},S^o,\widehat{S}_c)\mathcal{S}^{(0)}(t;\beta^o,\pi^o,S^o,\widehat{S}_c)}\cdot\frac{1}{n}\sum_{j=1}^{n}\{\Gamma_j^{(1)}(t;\beta^o,\widehat{\pi},S^o,\widehat{S}_c)-\Gamma_j^{(1)}(t;\beta^o,\pi^o,S^o,\widehat{S}_c)\}\\
&-\frac{\mathcal{S}^{(1)}(t;\beta^o,\pi^o,S^o,\widehat{S}_c)}{\mathcal{S}^{(0)}(t;\beta^o,\widehat{\pi},S^o,\widehat{S}_c)\mathcal{S}^{(0)}(t;\beta^o,\pi^o,S^o,\widehat{S}_c)}\cdot\frac{1}{n}\sum_{j=1}^{n}\{\Gamma_j^{(0)}(t;\beta^o,\widehat{\pi},S^o,\widehat{S}_c)-\Gamma_j^{(0)}(t;\beta^o,\pi^o,S^o,\widehat{S}_c)\}\\
=&\frac{1}{n}\sum_{j=1}^{n}C_j(t)\{\widehat{\pi}(Z_j)-\pi^o(Z_j)\},
\end{aligned}
$$

where $C_j(t)$ are some functions bounded a.s.. Similarly, we have

$$
\{\bar{A}(t;\beta^o,\pi^o,S^o,\widehat{S}_c)-\bar{A}(t;\beta^o,\pi^o,S^o,S_c^o)\}=\frac{1}{n}\sum_{j=1}^{n}C'_j(t)\{\widehat{S}_c(t;a,Z_i)-S_c^o(t;a,Z_i)\},\tag{C.36}
$$

where $C'_j(t)$ are some other functions bounded a.s..

Next, let $d\mathcal{N}_i^{(0)}(t;\widehat{\pi},\widehat{S},\widehat{S}_c)-d\mathcal{N}_i^{(0)}(t;\widehat{\pi},S^o,\widehat{S}_c)=K_{1i}+K_{2i}$, where

$$
\begin{aligned}
K_{1i}=&\frac{d\widehat{S}(t;A_i,Z_i)-dS^o(t;A_i,Z_i)}{\widehat{\pi}(Z_i)^{A_i}\{1-\widehat{\pi}(Z_i)\}^{1-A_i}}\\
&-\sum_{a=0,1}\left\{1+\frac{A_i^a(1-A_i)^{1-a}}{\widehat{\pi}(Z_i)^a\{1-\widehat{\pi}(Z_i)\}^{1-a}}J_i(t;a,S^o,\widehat{S}_c)\right\}\{d\widehat{S}(t;a,Z_i)-dS^o(t;a,Z_i)\}\\
K_{2i}=&-\sum_{a=0,1}\frac{A_i^a(1-A_i)^{1-a}}{\widehat{\pi}(Z_i)^a\{1-\widehat{\pi}(Z_i)\}^{1-a}}\int_0^t\left\{\frac{1}{\widehat{S}(u;a,Z_i)}-\frac{1}{S^o(u;a,Z_i)}\right\}\frac{dM_{ci}(u;a,Z_i)}{\widehat{S}_c(u;a,Z_i)}\cdot d\widehat{S}(t;a,Z_i).
\end{aligned}
$$

We now have $Q_{421} = Q_{4211} + Q_{4212}$, where

$$Q_{4211} = \frac{1}{n^{3/2}} \sum_{i=1}^{n} \sum_{j=1}^{n} \{\widehat{\pi}(Z_j) - \pi^o(Z_j)\} \int_0^\tau C_j(t) K_{1i}$$
$$Q_{4212} = \frac{1}{n^{3/2}} \sum_{i=1}^{n} \sum_{j=1}^{n} \{\widehat{\pi}(Z_j) - \pi^o(Z_j)\} \int_0^\tau C_j(t) K_{2i}.$$

For Q_{4212}, we can apply (C.22), the rate Condition Assumption 12 and the boundedness of appropriate terms to show that

$$E(|Q_{4212}|) \lesssim \sqrt{n} \|\widehat{\pi} - \pi^o\| \cdot \left\|\widehat{S} - S^o\right\| = o(1).$$

$\int_0^\tau C_j(t) K_{1i}$ in Q_{4211} involves stochastic differences $d\widehat{S}(t;a,Z_i) - dS^o(t;a,Z_i)$, so like in (C.25) and (C.34) we first apply integration by parts to turn $d\widehat{S} - dS^o$ into $\widehat{S} - S^o$. Like (C.35), the $dC_j(t)$ term we have as a result of integration by parts can again be shown to be a sum of terms, each being a product between a term that is bounded a.s. and an increment of a monotone function. This allows us to apply (C.22), the rate Condition Assumption 12 and the boundedness of appropriate terms, which leads to $E(|Q_{4211}|) \lesssim \sqrt{n} \|\widehat{\pi} - \pi^o\| \cdot \left\|\widehat{S} - S^o\right\| = o(1)$. We therefore have $E(|Q_{421}|) = o(1)$ and $Q_{421} = o_p(1)$ by Markov's inequality.

For term Q_{422}, we make use of (C.36) and write $Q_{422} = Q_{4221} + Q_{4222}$, where

$$Q_{4221} = \frac{1}{n^{3/2}} \sum_{i=1}^{n} \sum_{j=1}^{n} \int_0^\tau C_j'(t) \{\widehat{S}_c(t;a,Z_i) - S_c^o(t;a,Z_i)\} K_{1i},$$
$$Q_{4222} = \frac{1}{n^{3/2}} \sum_{i=1}^{n} \sum_{j=1}^{n} \int_0^\tau C_j'(t) \{\widehat{S}_c(t;a,Z_i) - S_c^o(t;a,Z_i)\} K_{2i}.$$

By again applying (C.22), the rate condition Assumption 12 and the boundedness of appropriate terms to Q_{4222}, we have

$$E(|Q_{4222}|) \lesssim \sqrt{n} \left\|\widehat{S}_c - S_c^o\right\| \cdot \left\|\widehat{S} - S^o\right\| = o(1).$$

$E(|Q_{4221}|)$ involves a product between $d\widehat{S}(t;a,Z_i) - dS^o(t;a,Z_i)$ and $\widehat{S}_c(t;a,Z_i) -$

$S_c^o(t;a,Z_i)$, which can not be bounded using any existing tools we have. Therefore, we directly bound

$$Q_{4221} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \int_0^{\tau} \{\bar{A}(t;\beta^o,\pi^o,S^o,\widehat{S}_c) - \bar{A}(t;\beta^o,\pi^o,S^o,S_c^o)\} K_{1i}$$

using $\mathcal{D}_2^{\dagger}$ in Assumption 12, which gives

$$E(|Q_{4221}|) \lesssim \sqrt{n}\mathcal{D}_2^{\dagger} = o(1).$$

Therefore, $E(|Q_{422}|) = o(1)$ from rate condition Assumption 12.

Combining our results, we have

$$E(|Q_{42}|) \leq E(|Q_{421}|) + E(|Q_{422}|) = o(1).$$

Using the same techniques we used for Q_{41} and Q_{42} above, with the rate condition Assumption 12 and without using $\mathcal{D}^{\dagger}$, we can show that $E(|Q_{43}|) = o(1)$ and $E(|Q_{44}|) = o(1)$.

Hence we conclude that $E(|Q_4|) \leq E(|Q_{41}|) + E(|Q_{42}|) + E(|Q_{43}|) + E(|Q_{44}|) = o(1)$. Then by Markov's inequality, $Q_4 = o_p(1)$.

Next, we show that $Q_5 = o_p(1)$.

Using the definition of $D_{1i}(\beta,\Lambda_0,\pi,S,S_c), D_{2i}(\beta,\Lambda_0,\pi,S,S_c)$ defined in Supplementary Material C.1, it can be verified that

$$U(\beta^o,\pi,S,S_c) = \frac{1}{n}\sum_{i=1}^{n}\left[D_{2i}(\beta^o,\Lambda_0^o,\pi,S,S_c) - \int_0^{\tau} \bar{A}(t;\beta^o,\pi,S,S_c) D_{1i}(t;\beta^o,\Lambda_0^o,\pi,S,S_c)\right] \tag{C.37}$$

So we have $Q_5 = Q_{51} - Q_{52} - Q_{53} - Q_{54}$, where

$$\begin{aligned}
Q_{51} =& \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{D_{2i}(\beta^o, \Lambda_0^o, \widehat{\pi}, S^o, \widehat{S}_c) - D_{2i}(\beta^o, \Lambda_0^o, \pi^o, S^o, S_c^o)\} \\
Q_{52} =& \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \int_0^{\tau} \bar{\alpha}(t; \beta^o, \widehat{\pi}, S^o, \widehat{S}_c)\{D_{1i}(t; \beta^o, \Lambda_0^o, \widehat{\pi}, S^o, \widehat{S}_c) - D_{1i}(t; \beta^o, \Lambda_0^o, \pi^o, S^o, S_c^o)\} \\
Q_{53} =& \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \int_0^{\tau} \{\bar{A}(t; \beta^o, \widehat{\pi}, S^o, \widehat{S}_c) - \bar{\alpha}(t; \beta^o, \widehat{\pi}, S^o, \widehat{S}_c)\} \\
& \times \{D_{1i}(t; \beta^o, \Lambda_0^o, \widehat{\pi}, S^o, \widehat{S}_c) - D_{1i}(t; \beta^o, \Lambda_0^o, \pi^o, S^o, S_c^o)\} \\
Q_{54} =& \int_0^{\tau} \{\bar{A}(t; \beta^o, \widehat{\pi}, S^o, \widehat{S}_c) - \bar{A}(t; \beta^o, \pi^o, S^o, S_c^o)\} \cdot \frac{1}{\sqrt{n}} \sum_{i=1}^{n} D_{1i}(t; \beta^o, \Lambda_0^o, \pi^o, S^o, S_c^o).
\end{aligned}$$

First, consider Q_{51}. By the law of total variance, we have

$$\mathrm{Var}(Q_{51}) = \mathrm{Var}\{E(Q_{51}|O^\dagger)\} + E\{\mathrm{Var}(Q_{51}|O^\dagger)\}.$$

We note from Theorem 10 that $E\{D_{2i}(\beta^o, \Lambda_0^o, \widehat{\pi}, S^o, \widehat{S}_c) - D_{2i}(\beta^o, \Lambda_0^o, \pi^o, S^o, S_c^o)|O^\dagger\} = 0$ for each i, where $O^\dagger$ is the sample independent from O that is used for estimating the nuisance functions, so $E(Q_{51}|O^\dagger) = 0$. Moreover, when conditional on $O^\dagger$, Q_{51} is a sample average of mean-zero i.i.d terms, so we have

$$\mathrm{Var}(Q_{51}|O^\dagger) = \frac{n}{n} E\left[\{D_2(\beta^o, \Lambda_0^o, \widehat{\pi}, S^o, \widehat{S}_c) - D_2(\beta^o, \Lambda_0^o, \pi^o, S^o, S_c^o)\}^2 |O^\dagger\right].$$

Expand $D_2(\beta^o, \Lambda_0^o, \widehat{\pi}, S^o, \widehat{S}_c) - D_2(\beta^o, \Lambda_0^o, \pi^o, S^o, S_c^o)$, we have

$$\begin{aligned}
&D_2(\beta^o, \Lambda_0^o, \widehat{\pi}, S^o, \widehat{S}_c) - D_2(\beta^o, \Lambda_0^o, \pi^o, S^o, S_c^o) \qquad \text{(C.38)}\\
=&-\int_0^\tau \frac{A\{\widehat{\pi}(Z)-\pi^o(Z)\}}{\widehat{\pi}(Z)\pi^o(Z)} \cdot \{dS^o(t;A,Z) + S^o(t;A,Z)e^{\beta^o A} d\Lambda_0^o(t)\}\\
&+\int_0^\tau \frac{AS_c^o(t;A,Z)\{\pi^o(Z)-\widehat{\pi}(Z)\}}{\widehat{\pi}(Z)S_c^o(t;A,Z)\widehat{S}_c(t;A,Z)\pi^o(Z)^A\{1-\pi^o(Z)\}^{1-A}} \cdot \{dN(t) - Y(t)e^{\beta^o} d\Lambda_0^o(t)\}\\
&+\int_0^\tau \frac{A\widehat{\pi}(Z)\{S_c^o(t;A,Z)-\widehat{S}_c(t;A,Z)\}}{\widehat{\pi}(Z)S_c^o(t;A,Z)\widehat{S}_c(t;A,Z)\pi^o(Z)^A\{1-\pi^o(Z)\}^{1-A}} \cdot \{dN(t) - Y(t)e^{\beta^o} d\Lambda_0^o(t)\}\\
&-\int_0^\tau \frac{AJ(t;1,S^o,S_c^o)\{\widehat{\pi}(Z)-\pi^o(Z)\}\}}{\widehat{\pi}(Z)\pi^o(Z)} \cdot \{dS^o(t;1,Z) + S^o(t;1,Z)e^{\beta^o} d\Lambda_0^o(t)\}\\
&+\int_0^\tau \frac{A\{J(t;1,S^o,\widehat{S}_c)-J(t;1,S^o,S_c^o)\}}{\widehat{\pi}(Z)} \cdot \{dS^o(t;1,Z) + S^o(t;1,Z)e^{\beta^o} d\Lambda_0^o(t)\}.
\end{aligned}$$

We now see that $D_2(\beta^o, \Lambda_0^o, \widehat{\pi}, S^o, \widehat{S}_c) - D_2(\beta^o, \Lambda_0^o, \pi^o, S^o, S_c^o)$ consists of several terms, where each term is an integral of a difference in nuisance functions with respect to a monotone function. This allows us to apply (C.22) to each of the terms and have

$$\begin{aligned}
|D_2(\beta^o, \Lambda_0^o, \widehat{\pi}, S^o, \widehat{S}_c) - D_2(\beta^o, \Lambda_0^o, \pi^o, S^o, S_c^o)| \lesssim& |\widehat{\pi}(Z)-\pi^o(Z)|\\
&+ \sup_{t\in[0,\tau], a\in\{0,1\}} |S_c^o(t;a,Z) - \widehat{S}_c(t;a,Z)|.
\end{aligned}$$

From the inequality $(a+b)^2 \leq 2a^2 + 2b^2$, we have

$$\begin{aligned}
\mathrm{Var}(Q_{51}|O^\dagger) \lesssim& E\left[\left(|\widehat{\pi}(Z)-\pi^o(Z)| + \sup_{t\in[0,\tau], a\in\{0,1\}} |S_c^o(t;a,Z) - \widehat{S}_c(t;a,Z)|\right)^2 \middle| O^\dagger\right]\\
\leq& 2E[\{\widehat{\pi}(Z)-\pi^o(Z)\}^2|O^\dagger]\\
&+2E\left[\left\{\sup_{t\in[0,\tau], a\in\{0,1\}} |S_c^o(t;a,Z) - \widehat{S}_c(t;a,Z)|\right\}^2 \middle| O^\dagger\right].
\end{aligned}$$

So

$$
\begin{aligned}
&\operatorname{Var}(Q_{51})\\
=&\operatorname{Var}^{\dagger}\{E(Q_{51}|O^{\dagger})\}+E^{\dagger}\{\operatorname{Var}(Q_{51}|O^{\dagger})\}\\
\lesssim&0+E^{\dagger}(E[\{\widehat{\pi}(Z)-\pi^{o}(Z)\}^2|O^{\dagger}])\\
&+E^{\dagger}\left(E\left[\left\{\sup_{t\in[0,\tau],a\in\{0,1\}}|S_c^o(t;a,Z)-\widehat{S}_c(t;a,Z)|\right\}^2\middle|O^{\dagger}\right]\right)\\
=&\|\widehat{\pi}-\pi^o\|^2+\|\widehat{S}_c-S_c^o\|^2\\
=&o(1).
\end{aligned}
$$

Therefore, $Q_{51}=o_p(1)$ by Chebyshev's inequality.

Conditional on $O^{\dagger}$, we also have from Theorem 10 that $E\{D_{1i}(t;\beta^o,\Lambda_0^o,\widehat{\pi},S^o,\widehat{S}_c)-D_{1i}(t;\beta^o,\Lambda_0^o,\pi^o,S^o,S_c^o)|O^{\dagger}\}=0$ for each t and i, so Q_{52} is again a sample average of i.i.d. mean-zero terms when conditional on $O^{\dagger}$, and we can show $Q_{52}=o_p(1)$ in the same way as for Q_{51} above.

Consider Q_{53}. Just like the expansion of $D_2(\beta^o,\Lambda_0^o,\widehat{\pi},S^o,\widehat{S}_c)-D_2(\beta^o,\Lambda_0^o,\pi^o,S^o,S_c^o)$ in (C.38) above, we also have $D_{1i}(t;\beta^o,\Lambda_0^o,\widehat{\pi},S^o,\widehat{S}_c)-D_{1i}(t;\beta^o,\Lambda_0^o,\pi^o,S^o,S_c^o)$ as a sum of terms, where each term is a product between a difference in nuisance functions and an increment of a monotone function. So same as in Q_{51}, we apply (C.22) to each of the terms and have

$$
\begin{aligned}
|Q_{53}|\lesssim&\sqrt{n}\sup_{t\in[0,\tau]}\left|\bar{A}(t;\beta^o,\widehat{\pi},S^o,\widehat{S}_c)-\bar{\alpha}(t;\beta^o,\widehat{\pi},S^o,\widehat{S}_c)\right|\\
&\cdot\left\{\frac{1}{n}\sum_{i=1}^{n}|\widehat{\pi}(Z_i)-\pi^o(Z_i)|+\frac{1}{n}\sum_{i=1}^{n}\sup_{t\in[0,\tau],a\in\{0,1\}}|S_c^o(t;a,Z_i)-\widehat{S}_c(t;a,Z_i)|\right\}.
\end{aligned}
$$

From the uniform convergence Assumption 11 and the Markov's inequality, we have

$$
\frac{1}{n}\sum_{i=1}^{n}|\widehat{\pi}(Z_i)-\pi^o(Z_i)|+\frac{1}{n}\sum_{i=1}^{n}\sup_{t\in[0,\tau],a\in\{0,1\}}\left|S_c^o(t;a,Z_i)-\widehat{S}_c(t;a,Z_i)\right|=o_p(1).
$$

From (C.18) of Assumption 23, we have

$$\sqrt{n}\sup_{t\in[0,\tau]}\left|\bar{A}(t;\beta^o,\widehat{\pi},S^o,\widehat{S}_c)-\bar{\alpha}(t;\beta^o,\widehat{\pi},S^o,\widehat{S}_c)\right|=O_p(1).$$

We therefore have $Q_{53}=o_p(1)$.

For Q_{54}, we have $Q_{54}=Q_{541}-Q_{542}+Q_{543}$, where

$$\begin{aligned}
Q_{541}=&\int_0^\tau\{\bar{A}(t;\beta^o,\widehat{\pi},S^o,\widehat{S}_c)-\bar{\alpha}(t;\beta^o,\widehat{\pi},S^o,\widehat{S}_c)\}\cdot\frac{1}{\sqrt{n}}\sum_{i=1}^n D_{1i}(t;\beta^o,\Lambda_0^o,\pi^o,S^o,S_c^o),\\
Q_{542}=&\int_0^\tau\{\bar{A}(t;\beta^o,\pi^o,S^o,S_c^o)-\bar{\alpha}(t;\beta^o,\pi^o,S^o,S_c^o)\}\cdot\frac{1}{\sqrt{n}}\sum_{i=1}^n D_{1i}(t;\beta^o,\Lambda_0^o,\pi^o,S^o,S_c^o),\\
Q_{543}=&\frac{1}{\sqrt{n}}\sum_{i=1}^n\int_0^\tau\{\bar{\alpha}(t;\beta^o,\widehat{\pi},S^o,\widehat{S}_c)-\bar{\alpha}(t;\beta^o,\pi^o,S^o,S_c^o)\}D_{1i}(t;\beta^o,\Lambda_0^o,\pi^o,S^o,S_c^o)
\end{aligned}$$

By (C.19) of Assumption 23, we have $Q_{541}=o_p(1)$ and $Q_{542}=o_p(1)$. Q_{543} is again a sample average of i.i.d. terms when conditional on $O^\dagger$, and each of the increments in $D_{1i}(t;\beta^o,\Lambda_0^o,\pi^o,S^o,S_c^o)$ is an increment of a monotone function. So like Q_{51}, we apply (C.22), followed by the law of total variance and have

$$\mathrm{Var}(Q_{543})\lesssim 0+\frac{n}{n}E^\dagger\left(E\left[\left\{\sup_{t\in[0,\tau]}\left|\bar{\alpha}(t;\beta^o,\widehat{\pi},S^o,\widehat{S}_c)-\bar{\alpha}(t;\beta^o,\pi^o,S^o,S_c^o)\right|\right\}^2\right]\right)=o(1),$$

where $o(1)$ follows from (C.17) of Assumption 23. Therefore, $Q_{543}=o_p(1)$ by Chebyshev's inequality and $Q_{54}=o_p(1)$.

Combining our results on Q_{51} to Q_{54} ,we have $Q_5=o_p(1)$.

Same as how we dealt with Q_5, we can decompose Q_6 in a similar way and show that each of the terms is $o_p(1)$, so we omit the details here.

Lastly, we consider $\sqrt{n}U(\beta^o, \pi^o, S^o, S_c^o)$. Using (C.37), we have

$$\begin{aligned}
&\sqrt{n}U(\beta^o, \pi^o, S^o, S_c^o) \\
=&\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[D_{2i}(\beta^o, \Lambda_0^o, \pi^o, S^o, S_c^o) - \int_0^\tau \bar{A}(t; \beta^o, \pi^o, S^o, S_c^o) D_{1i}(t; \beta^o, \Lambda_0^o, \pi^o, S^o, S_c^o)\right] \\
=&\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_i(\beta^o, \Lambda_0^o, \pi^o, S^o, S_c^o) \\
&+\int_0^\tau \{\bar{\alpha}(t; \beta^o, \pi^o, S^o, S_c^o) - \bar{A}(t; \beta^o, \pi^o, S^o, S_c^o)\} \cdot \frac{1}{\sqrt{n}}\sum_{i=1}^{n} D_{1i}(t; \beta^o, \Lambda_0^o, \pi^o, S^o, S_c^o) \quad (C.39)
\end{aligned}$$

From (C.19) of Assumption 23, we have (C.39) $= o_p(1)$, so we have

$$\sqrt{n}U(\beta^o, \pi^o, S^o, S_c^o) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_i(\beta^o, \Lambda_0^o, \pi^o, S^o, S_c^o) + o_p(1).$$

□

C.4 Application

Here, we present the boxplot of the $\pi(z)$, $S(\tau; a, z)$ and $S_c(\tau; a, z)$ for each of the 2079 patients in HAAS study, where $\tau = 13$ is the maximum follow-up time that we set to ensure the strict positivity assumption is satisfied for S and S_c.

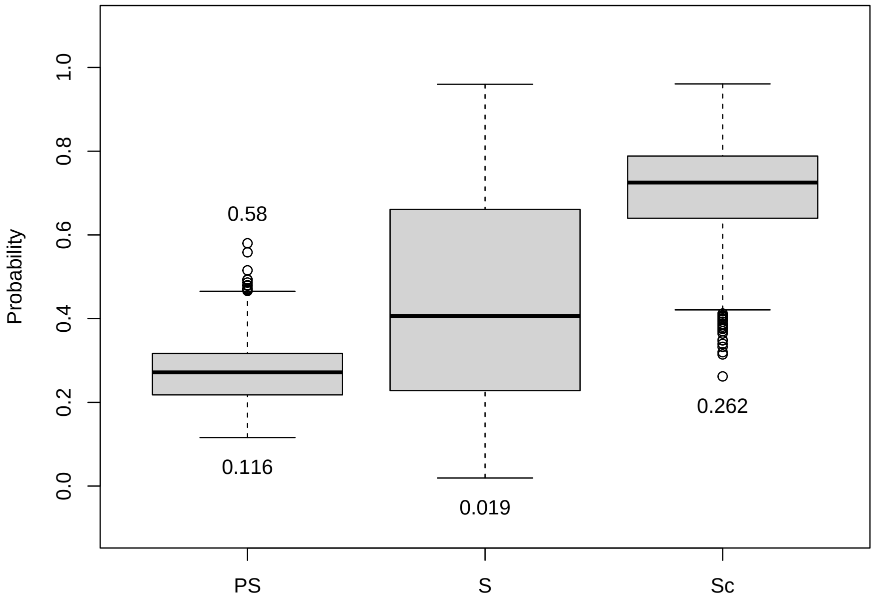


Figure C.1: Boxplot of the $\pi(z)$, $S(\tau;a,z)$ and $S_c(\tau;a,z)$ for all the patients in the HAAS dataset.

Appendix D

Supplementary Materials for Chapter 5

D.1 Quantities and Notations

We define or repeat some of the important quantities that will be used later. For i in $1,\ldots,n$,

$$
\begin{aligned}
M_{ci}(t;a,S_c) &= I(X_i \leq t, \Delta_i = 0) - \int_0^t I(X_i \geq u)\, d\Lambda_c(u;a,Z_i), \\
J_i(t;a,S,S_c) &= \int_0^t \frac{dM_{ci}(u;a,S_c)}{S(u;a,Z_i)S_c(u;a,Z_i)}, \\
d\mathcal{N}_i^{(l)}(t;\pi,S,S_c) &= \frac{A_i^l dN_i(t)}{\pi(Z_i)^{A_i}\{1-\pi(Z_i)\}^{1-A_i}S_c(t;A_i,Z_i)} + \frac{A_i^l dS(t;A_i,Z_i)}{\pi(Z_i)^{A_i}\{1-\pi(Z_i)\}^{1-A_i}} \\
&\quad - \sum_{a=0,1} a^l \left\{1 + \frac{A_i^a(1-A_i)^{1-a}}{\pi(Z_i)^a\{1-\pi(Z_i)\}^{1-a}} J_i(t;a,S,S_c)\right\} dS(t;a,Z_i), \\
\Gamma_i^{(l)}(t;\beta,\pi,S,S_c) &= \frac{A_i^l Y_i(t) e^{\beta A_i}}{\pi(Z_i)^{A_i}\{1-\pi(Z_i)\}^{1-A_i}S_c(t;A_i,Z_i)} - \frac{A_i^l S(t;A_i,Z_i) e^{\beta A_i}}{\pi(Z_i)^{A_i}\{1-\pi(Z_i)\}^{1-A_i}} \\
&\quad + \sum_{a=0,1} a^l \left\{1 + \frac{A_i^a(1-A_i)^{1-a}}{\pi(Z_i)^a\{1-\pi(Z_i)\}^{1-a}} J_i(t;a,S,S_c)\right\} S(t;a,Z_i)e^{\beta a}, \\
\mathcal{S}^{(l)}(t;\beta,\pi,S,S_c) &= \frac{1}{n}\sum_{i=1}^n \Gamma_i^{(l)}(t;\beta,\pi,S,S_c), \\
\bar{A}(t;\beta,\pi,S,S_c) &= \frac{\mathcal{S}^{(1)}(t;\beta,\pi,S,S_c)}{\mathcal{S}^{(0)}(t;\beta,\pi,S,S_c)}, \\
V(t;\beta,\pi,S,S_c) &= \bar{A}(t;\beta,\pi,S,S_c) - \bar{A}(t;\beta,\pi,S,S_c)^2, \\
\widetilde{\Lambda}(t;\beta,\pi,S,S_c) &= \frac{1}{n}\sum_{i=1}^n \int_0^t \frac{d\mathcal{N}_i(u;\pi,S,S_c)}{\mathcal{S}^{(0)}(u;\beta,\pi,S,S_c)}, \\
\widetilde{\psi}_i(\beta,\Lambda,\pi,S,S_c) &= D_{2i}(t;\beta,\Lambda,\pi,S,S_c) - \int_0^\tau \bar{A}(t;\beta,\pi,S,S_c) D_{1i}(t;\beta,\Lambda,\pi,S,S_c).
\end{aligned}
$$

The variance estimator of $\widehat{\beta}$ with model DR is defined as

$$
\widehat{\sigma}^2(\beta) = \frac{\frac{1}{n}\sum_{i=1}^n \widetilde{\psi}_i(\beta,\widetilde{\Lambda}(\cdot;\beta,\widehat{\pi},\widehat{S},\widehat{S}_c),\widehat{\pi},\widehat{S},\widehat{S}_c)^2}{\left\{\frac{1}{n}\sum_{i=1}^n \int_0^\tau V(t;\beta,\widehat{\pi},\widehat{S},\widehat{S}_c)\, d\mathcal{N}_i^{(0)}(t;\widehat{\pi},\widehat{S},\widehat{S}_c)\right\}^2}, \tag{D.1}
$$

where $\widehat{\pi}$, $\widehat{S}$ and $\widehat{S}_c$ are estimated using the same sample of n observations.

Next, for each fold m, we may define the fold-specific quantities:

$$
\begin{aligned}
\mathcal{S}_m^{(l)}(t;\beta,\pi,S,S_c) &= \frac{1}{|I_m|}\sum_{i\in I_m}\Gamma_i^{(l)}(t;\beta,\pi,S,S_c),\\
\bar{A}_m(t;\beta,\pi,S,S_c) &= \frac{\mathcal{S}_m^{(1)}(t;\beta,\pi,S,S_c)}{\mathcal{S}_m^{(0)}(t;\beta,\pi,S,S_c)},\\
V_m(t;\beta,\pi,S,S_c) &= \bar{A}_m(t;\beta,\pi,S,S_c) - \bar{A}_m(t;\beta,\pi,S,S_c)^2,\\
\widetilde{\Lambda}_m(t;\beta,\pi,S,S_c) &= \frac{1}{|I_m|}\sum_{i\in I_m}\int_0^t \frac{d\mathcal{N}_i^{(0)}(u;\pi,S,S_c)}{\mathcal{S}_m^{(0)}(u;\beta,\pi,S,S_c)},\\
\widetilde{\psi}_{m,i}(\beta^o,\Lambda_0,\pi,S,S_c) &= D_{2i}(t;\beta,\Lambda,\pi,S,S_c) - \int_0^\tau \bar{A}_m(t;\beta,\pi,S,S_c)D_{1i}(t;\beta,\Lambda,\pi,S,S_c),\\
\widetilde{\eta}_{m,i}(t;\beta,\Lambda,\pi,S,S_c) &= d\mathcal{N}_i^{(1)}(t;\pi,S,S_c) - \Gamma_i^{(1)}(t;\beta,\pi,S,S_c)d\Lambda(t)\\
&\qquad - \bar{A}_m(t;\beta,\pi,S,S_c)D_{1i}(t;\beta,\Lambda,\pi,S,S_c).
\end{aligned}
$$

The variance estimator $\widehat{\sigma}_{cf}^2$ for the cross-fitted AIPW estimator $\widehat{\beta}_{cf}$ is defined as

$$
\widehat{\sigma}_{cf}^2(\widehat{\beta}_{cf}) = \frac{\frac{1}{n}\sum_{m=1}^k\sum_{i\in I_m}\widetilde{\psi}_{m,i}(\widehat{\beta}_{cf},\widetilde{\Lambda}_m(\cdot;\widehat{\beta}_{cf},\widehat{\pi}^{(-m)},\widehat{S}^{(-m)},\widehat{S}_c^{(-m)}),\widehat{\pi}^{(-m)},\widehat{S}^{(-m)},\widehat{S}_c^{(-m)})^2}{\left\{\frac{1}{n}\sum_{m=1}^k\sum_{i\in I_m}\int_0^\tau V_m(t;\widehat{\beta}_{cf},\widehat{\pi}^{(-m)},\widehat{S}^{(-m)},\widehat{S}_c^{(-m)})d\mathcal{N}_i^{(0)}(t;\widehat{\pi}^{(-m)},\widehat{S}^{(-m)},\widehat{S}_c^{(-m)})\right\}^2}. \tag{D.2}
$$

The cross-fitted standardized AIPW score $s_{aipw}(\widehat{\beta}_{cf},t)/V_{aipw}(\widehat{\beta}_{cf},t)$ that is used in the AIPW $\beta(t)$ approximation (5.11) can be thought of as taking the differentials of the square root of both the numerator and denominator of (D.2), and is expressed as

$$
\frac{\frac{1}{n}\sum_{m=1}^k\sum_{i\in I_m}\widetilde{\eta}_{m,i}(t;\widehat{\beta}_{cf},\Lambda(\cdot;\widehat{\beta}_{cf},\widehat{\pi}^{(-m)},\widehat{S}^{(-m)},\widehat{S}_c^{(-m)}),\widehat{\pi}^{(-m)},\widehat{S}^{(-m)},\widehat{S}_c^{(-m)}))}{\frac{1}{n}\sum_{m=1}^k\sum_{i\in I_m}V_m(t;\widehat{\beta}_{cf},\widehat{\pi}^{(-m)},\widehat{S}^{(-m)},\widehat{S}_c^{(-m)})d\mathcal{N}_i^{(0)}(t;\widehat{\pi}^{(-m)},\widehat{S}^{(-m)},\widehat{S}_c^{(-m)})}.
$$

D.2 Proof of the Lemmas

We let $F(t)$ and $f(t)$ denote the marginal cumulative distribution and density of T, $F(t;a)$ and $f(t;a)$ denote the cumulative distribution and density of T conditional on $A = a$, while $F(t;z,z)$ and $f(t;a,z)$ denote the cumulative distribution and density of T conditional on $A = a$ and $Z = z$. We define $S(t;a)$ and $S(t;a,z)$ in similar fashion, where $S = 1 - F$.

D.2.1 Proof of Lemma 1

By conditional probability and consistency, we have

$$f(t) = P(A=1)f(t;A=1) + P(A=0)f(t;A=0) = \frac{1}{2}\{f_1(t) + f_0(t)\}.$$

Bayes' Theorem on two random variables X, Y states that

$$E(X|Y=y) = \frac{E\{Xf_{Y|X}(y|X)\}}{f_Y(y)}.$$

Applying this to $E_{\beta(t)}(A|T=t)$, where the expectation is taken assuming the true non-PH model 5.3, we have

$$\begin{aligned} E_{\beta(t)}(A|T=t) =& \frac{E\{Af(t;A)\}}{f(t)} \\ =& \frac{\frac{1}{2}\sum_{a=0,1} af_a(t)}{\frac{1}{2}\sum_{a=0,1} f_a(t)} \\ =& \frac{\sum_{a=0,1} a\lambda_{T(a)}(t)S_a(t)}{\sum_{a=0,1} \lambda_{T(a)}(t)S_a(t)} \\ =& \frac{\sum_{a=0,1} a\Lambda(t)e^{\beta(t)a}S_a(t)}{\sum_{a=0,1} \Lambda(t)e^{\beta(t)a}S_a(t)} \\ =& \frac{\sum_{a=0,1} ae^{\beta(t)a}S_a(t)}{\sum_{a=0,1} e^{\beta(t)a}S_a(t)} \end{aligned} \tag{D.3}$$

Similarly, replace $\beta(t)$ in $\lambda(t;A)$ with a fixed β while keeping everything else the same, we have

$$E_\beta(A|T=t) = \frac{\sum_{a=0,1} ae^{\beta a} S_a(t)}{\sum_{a=0,1} e^{\beta a} S_a(t)} \tag{D.4}$$

Substituting (D.3) and (D.4) into (5.2), we therefore have

$$\int_0^\tau \left\{ \frac{\sum_{a=0,1} ae^{\beta(t)a} S_a(t)}{\sum_{a=0,1} e^{\beta(t)a} S_a(t)} - \frac{\sum_{a=0,1} ae^{\beta a} S_a(t)}{\sum_{a=0,1} e^{\beta a} S_a(t)} \right\} \sum_{a=0,1} dF_a(t) = 0.$$

□

D.2.2 Proof of Lemma 4

First, we solve $E\{D_1^f(t;\beta,\Lambda)\} = 0$. The two limits of the integral from the expectation are constants, so by Leibniz integral rule, we may exchange the order of differentiation and integral and have

$$\begin{aligned} d\Lambda(t;\beta) =& \frac{\sum_{a=0,1} dE\{I\{T(a) < t\}\}}{\sum_{a=0,1} e^{\beta a} E\left[I\{T(a) \geq t\}\right]} \\ =& \frac{\sum_{a=0,1} dF_a(t)}{\sum_{a=0,1} e^{\beta a} S_a(t)} \\ =& -\frac{\sum_{a=0,1} dS_a(t)}{\sum_{a=0,1} e^{\beta a} S_a(t)} \end{aligned} \tag{D.5}$$

Plugging it into $E\{D_2^f(\beta,\Lambda)\} = 0$, we have

$$\begin{aligned}
0 &= \int_0^\tau \sum_{a=0,1} a \cdot dE[I\{T(a) < t\}] - \frac{\sum_{a=0,1} ae^{\beta a} E\left[I\{T(a) \geq t\}\right] \cdot \sum_{a=0,1} dF_a(t)}{\sum_{a=0,1} e^{\beta a} S_a(t)} \\
&= \int_0^\tau \sum_{a=0,1} a \cdot f_a(t)dt - \frac{\sum_{a=0,1} ae^{\beta a} S_a(t)}{\sum_{a=0,1} e^{\beta a} S_a(t)} \sum_{a=0,1} f_a(t)dt, \\
&= \int_0^\tau \left\{ \frac{\sum_{a=0,1} a \cdot f_a(t)}{\sum_{a=0,1} f_a(t)} - \frac{\sum_{a=0,1} ae^{\beta a} S_a(t)}{\sum_{a=0,1} e^{\beta a} S_a(t)} \right\} \sum_{a=0,1} f_a(t)dt \\
&= \int_0^\tau \left\{ \frac{\sum_{a=0,1} a \cdot \Lambda(t) e^{\beta(t)a} S_a(t)}{\sum_{a=0,1} \Lambda(t) e^{\beta(t)a} S_a(t)} - \frac{\sum_{a=0,1} ae^{\beta a} S_a(t)}{\sum_{a=0,1} e^{\beta a} S_a(t)} \right\} \sum_{a=0,1} f_a(t)dt \\
&= \int_0^\tau \left\{ \frac{\sum_{a=0,1} ae^{\beta(t)a} S_a(t)}{\sum_{a=0,1} e^{\beta(t)a} S_a(t)} - \frac{\sum_{a=0,1} ae^{\beta a} S_a(t)}{\sum_{a=0,1} e^{\beta a} S_a(t)} \right\} \sum_{a=0,1} f_a(t)dt,
\end{aligned}$$

which is equivalent to the definition of β^* defined in (5.4). Therefore β^* is the unique solution to β in the full data estimating functions. Plugging β^* into (D.5), we also see that $\Lambda^*(t)$ as defined in (5.7) is also the solution to $\Lambda(t)$ in the full data estimating functions. □

Bibliography

Andersen, P. K. and Gill, R. D. (1982). Cox's regression model for counting processes: a large sample study. The Annals of Statistics, 10:1100–1120.

Axelrod, R. and Nevo, D. (2022). A sensitivity analysis approach for the causal hazard ratio in randomized and observational studies. Biometrics.

Bai, X., Tsiatis, A. A., Lu, W., and Song, R. (2017). Optimal treatment regimes for survival endpoints using a locally-efficient doubly-robust estimator from a classification perspective. Lifetime Data Analysis, 23(4):585–604.

Bang, H. and Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. Biometrics, 61(4):962–973.

Barzilai, J. and Borwein, J. M. (1988). Two-point step size gradient methods. IMA journal of numerical analysis, 8(1):141–148.

Battey, H., Fan, J., Liu, H., Lu, J., and Zhu, Z. (2018). Distributed testing and estimation under sparse high dimensional models. The Annals of Statistics, 46(3):1352–1382.

Belloni, A., Chernozhukov, V., and Hansen, C. (2013). Inference on treatment effects after selection among high-dimensional controls†. The Review of Economic Studies, 81(2):608–650.

Bickel, P. J. (1975). One-step huber estimates in the linear model. Journal of the American Statistical Association, 70(350):428–434.

Bickel, P. J., Klaassen, C. A., Ritov, Y., and Wellner, J. A. (1993). Efficient and adaptive estimation for semiparametric models, volume 4. Johns Hopkins University Press, Baltimore.

Bousquet, O. (2003). Concentration inequalities for sub-additive functions using the entropy method. In Stochastic inequalities and applications, pages 213–247. Springer.

Boyd, A. P., Kittelson, J. M., and Gillen, D. L. (2012). Estimation of treatment effect under non-proportional hazards and conditionally independent censoring. Statistics in medicine, 31(28):3504–3515.

Buchanan, A. L., Hudgens, M. G., Cole, S. R., Lau, B., Adimora, A. A., and Women's Interagency HIV Study (2014). Worth the weight: using inverse probability weighted Cox models in AIDS research. AIDS research and human retroviruses, 30(12):1170–1177.

Catoni, O. (2012). Challenging the empirical mean and empirical variance: a deviation study. Annales de l'IHP Probabilités et statistiques, 48(4):1148–1185.

Chen, L. H. and Shao, Q.-M. (2001). A non-uniform berry-esseen bound via stein's method. Probability theory and related fields, 120:236–254.

Chen, P.-Y. and Tsiatis, A. A. (2001). Causal inference on the difference of the restricted mean lifetime between two groups. Biometrics, 57(4):1030–1038.

Chen, X., Liu, W., and Zhang, Y. (2019). Quantile regression under memory constraint. The Annals of Statistics, 47(6):3244–3273.

Chen, X. and White, H. (1999). Improved rates and asymptotic normality for nonparametric neural network estimators. IEEE Transactions on Information Theory, 45(2):682–691.

Chen, X. and Xie, M.-g. (2014). A split-and-conquer approach for analysis of extraordinarily large data. Statistica Sinica, pages 1655–1684.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. The Econometrics Journal, 21(1):C1–C68.

Chinot, G., Lecué, G., and Lerasle, M. (2020). Robust statistical learning with lipschitz and convex loss functions. Probability Theory and Related Fields, 176(3-4):897–940.

Cole, S. R., Hernán, M. A., Robins, J. M., Anastos, K., Chmiel, J., Detels, R., Ervin, C., Feldman, J., Greenblatt, R., Kingsley, L., et al. (2003). Effect of highly active antiretroviral therapy on time to acquired immunodeficiency syndrome or death using marginal structural models. American Journal of Epidemiology, 158(7):687–694.

Cox, D. R. (1972). Regression models and life-tables (with discussion). J. Roy. Statist. Soc. Ser. B, 34:187–220.

Cui, Y., Zhu, R., Zhou, M., and Kosorok, M. (2022). Consistency of survival tree and forest models: splitting bias and correction. Statistica Sinica, 32:1–23.

Dobriban, E. and Sheng, Y. (2018). Distributed linear regression by averaging. arXiv preprint arXiv:1810.00412.

Dukes, O., Martinussen, T., Tchetgen Tchetgen, E. J., and Vansteelandt, S. (2019). On doubly robust estimation of the hazard difference. Biometrics, 75(1):100–109.

D'Amour, A., Ding, P., Feller, A., Lei, L., and Sekhon, J. (2021). Overlap in observational studies with high-dimensional covariates. Journal of Econometrics, 221(2):644–654.

Efron, B. (1992). Bootstrap methods: another look at the jackknife. In Breakthroughs in statistics, pages 569–593. Springer.

Fan, J., Li, Q., and Wang, Y. (2017). Estimation of high dimensional mean regression in the absence of symmetry and light tail assumptions. Journal of the Royal Statistical Society. Series B, Statistical methodology, 79(1):247.

Fan, J., Liu, H., Sun, Q., and Zhang, T. (2018). I-lamm for sparse learning: Simultaneous control of algorithmic complexity and statistical error. The Annals of Statistics, 46(2):814–841.

Feldman, H. I., Joffe, M., Robinson, B., Knauss, J., Cizman, B., Guo, W., Franklin-Becker, E., and Faich, G. (2004). Administration of parenteral iron and mortality among hemodialysis patients. Journal of the American Society of Nephrology, 15(6):1623–1632.

Fleming, T. R. and Harrington, D. P. (1991). Counting processes and survival analysis. Wiley, New York.

Ford, I., Norrie, J., and Ahmadi, S. (1995). Model inconsistency, illustrated by the Cox proportional hazards model. Statistics in Medicine, 14:735–746.

Gail, M. H., Wieand, S., and Piantadosi, S. (1984). Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates. Biometrika, 71:431–444.

Gray, R. J. (1992). Flexible methods for analyzing survival data using splines, with applications to breast cancer prognosis. Journal of the American Statistical Association, 87(420):942–951.

Harrington, D. P. and Fleming, T. R. (1982). A class of rank test procedures for censored survival data. Biometrika, 69(3):553–566.

Hattori, S. and Henmi, M. (2012). Estimation of treatment effects based on possibly misspecified cox regression. Lifetime data analysis, 18(4):408–433.

Hernán, M. A. (2010). The hazards of hazard ratios. Epidemiology, 21(1):13–15.

Hernán, M. A., Brumback, B., and Robins, J. M. (2001). Marginal structural models to estimate the joint causal effect of nonrandomized treatments. Journal of the American Statistical Association, 96(454):440–448.

Hernán, M. A., Lanoy, E., Costagliola, D., and Robins, J. M. (2006). Comparison of dynamic treatment regimes via inverse probability weighting. Basic & Clinical Pharmacology & Toxicology, 98(3):237–242.

Hernán, M. A. and Robins, J. M. (2020). Causal Inference: What If. Chapman & Hall/CRC, Boca Raton.

Horvitz, D. G. and Thompson, D. J. (1952). Using the whole cohort in the analysis of case-cohort data. American Journal of Epidemiology, 169(11):1398–1405.

Hou, J., Bradic, J., and Xu, R. (2021). Treatment effect estimation under additive hazards models with high-dimensional confounding. Journal of the American Statistical Association, page https://doi.org/10.1080/01621459.2021.1930546.

Hubbard, A. E., Van Der Laan, M. J., and Robins, J. M. (2000). Nonparametric locally efficient estimation of the treatment specific survival distribution with right censored data and covariates in observational studies. In Statistical Models in Epidemiology, the Environment, and Clinical Trials, pages 135–177. Springer, New York.

Huber, P. and Ronchetti, E. (2009). Robust statistics. John Wiley & Sons Hoboken, NJ, USA.

Huber, P. J. (1973). Robust regression: asymptotics, conjectures and monte carlo. The annals of statistics, pages 799–821.

Hunter, D. R. and Lange, K. (2000). Quantile regression via an mm algorithm. Journal of Computational and Graphical Statistics, 9(1):60–77.

Huo, X. and Cao, S. (2019). Aggregated inference. Wiley Interdisciplinary Reviews: Computational Statistics, 11(1):e1451.

Ishwaran, H., Kogalur, U. B., Blackstone, E. H., Lauer, M. S., et al. (2008). Random survival forests. Annals of Applied Statistics, 2(3):841–860.

Jordan, M. I., Lee, J. D., and Yang, Y. (2019). Communication-efficient distributed statistical inference. Journal of the American Statistical Association, 114(526):668–681.

Kalbfleisch, J. D. and Prentice, R. L. (2011). The Statistical Analysis of Failure Time Data, 2nd Edition. John Wiley & Sons, New York.

Kooperberg, C., Stone, C. J., and Truong, Y. K. (1995a). Hazard regression. Journal of the American Statistical Association, 90(429):78–94.

Kooperberg, C., Stone, C. J., and Truong, Y. K. (1995b). The L2 rate of convergence for hazard regression. Scandinavian Journal of Statistics, pages 143–157.

Lambert-Lacroix, S. and Zwald, L. (2011). Robust regression through the huber's criterion and adaptive lasso penalty. Electronic Journal of Statistics, 5:1015–1053.

Lancaster, T. and Nickell, S. (1980). The analysis of re-employment probabili- ties for the unemployed. Journal of the Royal Statistical Society, Series A, 143(2):141–165.

Lange, K., Hunter, D. R., and Yang, I. (2000). Optimization transfer using surrogate objective functions. Journal of computational and graphical statistics, 9(1):1–20.

Lee, J. D., Liu, Q., Sun, Y., and Taylor, J. E. (2017). Communication-efficient sparse regression. The Journal of Machine Learning Research, 18(1):115–144.

Li, K.-C. and Duan, N. (1989). Regression analysis under link violation. The Annals of Statistics, 17(3):1009–1052.

Li, R., Lin, D. K., and Li, B. (2013). Statistical inference in massive data sets. Applied Stochastic Models in Business and Industry, 29(5):399–409.

Lin, D. Y. and Wei, L.-J. (1989). The robust inference for the cox proportional hazards model. Journal of the American statistical Association, 84(408):1074–1078.

Loh, P.-L. (2017). Statistical consistency and asymptotic normality for high-dimensional robust *m*-estimators. The Annals of Statistics, pages 866–896.

Lu, W. and Ying, Z. (2004). On semiparametric transformation cure models. Biometrika, 91(2):331–343.

Lunceford, J. K. and Davidian, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. Statistics in Medicine, 23(19):2937–2960.

Luo, J. and Xu, R. (2022). Doubly robust inference for hazard ratio under informative censoring with machine learning. arXiv preprint arXiv:2206.02296.

Martinussen, T. and Vansteelandt, S. (2013). On collapsibility and confounding bias in Cox and Aalen regression models. Lifetime Data Analysis, 19:279–296.

Martinussen, T., Vansteelandt, S., and Andersen, P. K. (2020). Subtleties in the interpretation of hazard contrasts. Lifetime Data Analysis, 26:833–855.

Newey, W. K. (1994). The asymptotic variance of semiparametric estimators. Econometrica: Journal of the Econometric Society, pages 1349–1382.

Neyman, J. (1923). Sur les applications de la théorie des probabilités aux experiences agricoles: Essai des principes. Roczniki Nauk Rolniczych, 10:1–51.

Neyman, J. (1959). Optimal asymptotic tests of composite statistical hypotheses. In Probability and Statistics, U. Grenander (Ed.), page 416–444.

Nguyen, V. Q. and Gillen, D. L. (2017). 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(2):406–430.

Nielsen, G. G., Gill, R. D., Andersen, P. K., and Sørensen, T. I. (1992). A counting process approach to maximum likelihood estimation in frailty models. Scandinavian journal of Statistics, pages 25–43.

Nuño, M. M. and Gillen, D. L. (2021). Censoring-robust time-dependent receiver operating characteristic curve estimators. Statistics in Medicine, 40(30):6885–6899.

Petersen, M., Schwab, J., Gruber, S., Blaser, N., Schomaker, M., and Van Der Laan, M. (2014). Targeted maximum likelihood estimation for dynamic and static longitudinal marginal structural working models. Journal of Causal Inference, 2(2):147–185.

Prentice, R. L. and Aragaki, A. K. (2022). Intention-to-treat comparisons in randomized trials. Statistical Science, 37(3):380–393.

Rava, D. (2021). Survival Analysis and Causal Inference: from Marginal Structural Cox to Additive Hazards Model and beyond. University of California, San Diego, Ph.D. Thesis.

Rava, D. and Xu, R. (2023). Doubly robust estimation of the hazard difference for competing risks data. Statistics in Medicine, 42(6):799–814.

Ridgeway, G., McCaffrey, D. F., Morral, A. R., Cefalu, M., Burgette, L. F., Pane, J. D., and Griffin, B. A. (2022). Toolkit for weighting and analysis of nonequivalent groups: a tutorial for the R TWANG package. Rand Santa Monica, Calif.

Robins, J. (1998). Marginal structural models. Proceedings of the American Statistical Association. Section on Bayesian Statistical Science, pages 1–10.

Robins, J. M. (1993). Information recovery and bias adjustment in proportional hazards regression analysis of randomized trials using surrogate markers. In Proceedings of the Biopharmaceutical Section, American Statistical Association, pages 24–33. San Francisco CA.

Robins, J. M. (2000). Marginal structural models versus structural nested models as tools for causal inference. In Statistical models in epidemiology, the environment, and clinical trials, pages 95–133. Springer, New York.

Robins, J. M. and Finkelstein, D. M. (2000). Correcting for noncompliance and dependent censoring in an aids clinical trial with inverse probability of censoring weighted (ipcw) log-rank tests. Biometrics, 56(3):779–788.

Robins, J. M., Hernan, M. A., and Brumback, B. (2000a). Marginal structural models and causal inference in epidemiology. Epidemiology, 11:550–560.

Robins, J. M. and Rotnitzky, A. (2001). Comment on “inference for semiparametric models: Some questions and an answer”. Statistical Science, 11(4):920–936.

Robins, J. M., Rotnitzky, A., and van der Laan, M. (2000b). On profile likelihood: comment. Journal of the American Statistical Association, 95(450):477–482.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. Journal of the American Statistical Association, 90(429):106–121.

Rosenblatt, J. D. and Nadler, B. (2016). On the optimality of averaging in distributed statistical learning. Information and Inference: A Journal of the IMA, 5(4):379–404.

Rotnitzky, A. and Robins, J. (2005). Inverse probability weighted estimation in survival analysis. Encyclopedia of Biostatistics, 4:2619–2625.

Rotnitzky, A., Smucler, E., and Robins, J. (2021). Characterization of parameters with a mixed bias property. Biometrika, 108:231–238.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of Educational Psychology, 66(5):688–701.

Scharfstein, D. O., Rotnitzky, A., and Robins, J. M. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models. Journal of the American Statistical Association, 94(448):1096–1120.

Shamir, O., Srebro, N., and Zhang, T. (2014). Communication-efficient distributed optimization using an approximate newton-type method. In International conference on machine learning, pages 1000–1008. PMLR.

Shipp, M. (1993). A predictive model for aggressive non-hodgkin's lymphoma. the international non-hodgkin's lymphoma prognostic factors project. N Engl j Med, 329:987–994.

Sidransky, E., Nalls, M. A., Aasly, J. O., Aharon-Peretz, J., Annesi, G., Barbosa, E. R., Bar-Shira, A., Berg, D., Bras, J., Brice, A., et al. (2009). Multicenter analysis of glucocerebrosidase mutations in parkinson's disease. New England Journal of Medicine, 361(17):1651–1661.

Singh, S. and Maddala, G. (1976). A function for size distribution of incomes. Econometrica, 44(5):963–970.

Sjölander, A. and Vansteelandt, S. (2017). Doubly robust estimation of attributable fractions in survival analysis. Statistical Methods in Medical Research, 26(2):948–969.

Smucler, E., Rotnitzky, A., and Robins, J. M. (2019). A unifying approach for doubly-robust L_1 regularized estimation of causal contrasts. arXiv preprint arXiv:1904.03737.

Sterne, J. A., Hernán, M. A., Ledergerber, B., Tilling, K., Weber, R., Sendi, P., Rickenbach, M., Robins, J. M., Egger, M., Study, S. H. C., et al. (2005). Long-term effectiveness of potent antiretroviral therapy in preventing aids and death: a prospective cohort study. The Lancet, 366(9483):378–384.

Struthers, C. A. and Kalbfleisch, J. D. (1986). Misspecified proportional hazard models. Biometrika, 73(2):363–369.

Sun, Q., Zhou, W.-X., and Fan, J. (2020). Adaptive huber regression. Journal of the American Statistical Association, 115(529):254–265.

Tchetgen Tchetgen, E. J. and Robins, J. (2012). On parametrization, robustness and sensitivity analysis in a marginal structural cox proportional hazards model for point exposure. Statistics & Probability Letters, 82(5):907–915.

Therneau, T. M. and Grambsch, P. M. (2000). Modeling Survival Data: Extending the Cox Model. Springer Science & Business Media.

Tsiatis, A. A. (2006). Semiparametric theory and missing data. Springer, New York.

van de Geer, S., Bühlmann, P., Ritov, Y., and Dezeure, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. The Annals of Statistics, 42(3):1166–1202.

Van der Laan, M. J. and Robins, J. M. (2003). Unified methods for censored longitudinal data and causality. Springer Science & Business Media.

Van der Vaart, A. W. (2000). Asymptotic statistics, volume 3. Cambridge university press.

Van Lancker, K., Dukes, O., and Vansteelandt, S. (2021). Principled selection of baseline covariates to account for censoring in randomized trials with a survival endpoint. Statistics in Medicine.

Vansteelandt, S., Dukes, O., Van Lancker, K., and Martinussen, T. (2022). Assumption-lean cox regression. Journal of the American Statistical Association, pages 1–10.

Volgushev, S., Chao, S.-K., and Cheng, G. (2019). Distributed inference for quantile regression processes. The Annals of Statistics, 47(3):1634–1662.

Wager, S. and Walther, G. (2015). Adaptive concentration of regression trees, with application to random forests. arXiv preprint arXiv:1503.06388.

Wang, J., Kolar, M., Srebro, N., and Zhang, T. (2017). Efficient distributed learning with sparsity. In International conference on machine learning, pages 3636–3645. PMLR.

Wang, L., Peng, B., and Li, R. (2015). A high-dimensional nonparametric multivariate test for mean vector. Journal of the American Statistical Association, 110(512):1658–1669.

Wang, L., Zheng, C., Zhou, W., and Zhou, W.-X. (2021). A new principle for tuning-free huber regression. Statistica Sinica, 31(4):2153–2177.

Wang, X., Yang, Z., Chen, X., and Liu, W. (2019). Distributed inference for linear support vector machine. The Journal of machine learning research, 20(113):1–41.

Wang, Y., Ying, A., and Xu, R. (2022). Doubly robust estimation under covariate-induced dependent left truncation. arXiv preprint arXiv:2208.06836.

Western, B. (1995). Concepts and suggestions for robust regression analysis. American Journal of Political Science, pages 786–817.

White, H. (1982). Maximum likelihood estimation of misspecified models. Econometrica: Journal of the econometric society, pages 1–25.

Wu, Y., Jiang, X., Kim, J., and Ohno-Machado, L. (2012). Grid binary logistic regression (glore): building shared models without sharing data. Journal of the American Medical Informatics Association, 19(5):758–764.

Xu, R. (1996). Inference for the Proportional Hazards Model. Ph.D. thesis, University of California, San Diego.

Xu, R. and Adak, S. (2002). Survival analysis with time-varying regression effects using a tree-based approach. Biometrics, 58(2):305–315.

Xu, R. and Harrington, D. P. (2001). A semiparametric estimate of treatment effects with censored data. Biometrics, 57(3):875–885.

Xu, R. and O'Quigley, J. (2000). Estimating average regression effect under non-proportional hazards. Biostatistics, 1(4):423–439.

Yang, S., Pieper, K., and Cools, F. (2020). Semiparametric estimation of structural failure time models in continuous-time processes. Biometrika, 107(1):123–136.

Ying, A. (2023). A cautionary note on doubly robust estimators involving continuous-time structure. arXiv preprint arXiv:2302.06739.

Ying, A. and Xu, R. (2023). On defense of the hazard ratio. arXiv preprint arXiv:2307.11971.

Zhang, C.-H. and Zhang, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. Journal of the Royal Statistical Society: Series B: Statistical Methodology, pages 217–242.

Zhang, M. and Schaubel, D. E. (2011). Estimating differences in restricted mean lifetime using observational data subject to dependent censoring. Biometrics, 67(3):740–749.

Zhang, M. and Schaubel, D. E. (2012a). Contrasting treatment-specific survival using double-robust estimators. Statistics in Medicine, 31(30):4255–4268.

Zhang, M. and Schaubel, D. E. (2012b). Double-robust semiparametric estimator for differences in restricted mean lifetimes in observational studies. Biometrics, 68(4):999–1009.

Zhang, Y., Duchi, J., and Wainwright, M. (2015). Divide and conquer kernel ridge regression: A distributed algorithm with minimax optimal rates. The Journal of Machine Learning Research, 16(1):3299–3340.

Zhao, T., Cheng, G., and Liu, H. (2016). A partially linear framework for massive heterogeneous data. The Annals of Statistics, 44(4):1400.

Zheng, W., Petersen, M., and Van Der Laan, M. J. (2016). Doubly robust and efficient estimation of marginal structural models for the hazard function. The International Journal of Biostatistics, 12(1):233–252.

Zhou, W.-X., Bose, K., Fan, J., and Liu, H. (2018). A new perspective on robust *m*-estimation: Finite sample theory and applications to dependence-adjusted multiple testing. The Annals of Statistics, 46(5):1904.