Marginal Screening on Survival Data

Tzu-Jung Huang

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

COLUMBIA UNIVERSITY

2017

ABSTRACT

Marginal Screening on Survival Data

Tzu-Jung Huang

This work develops a marginal screening test to detect the presence of significant predictors for a right-censored time-to-event outcome under a high-dimensional accelerated failure time (AFT) model. Establishing a rigorous screening test in this setting is challenging, not only because of the right censoring, but also due to the post-selection inference. The oracle property in such situations fails to ensure adequate control of the family-wise error rate, and this raises questions about the applicability of standard inferential methods. McKeague and Qian (2015) constructed an adaptive resampling test to circumvent this problem under ordinary linear regression. To accommodate right censoring, we develop a test statistic based on a maximally selected Koul–Susarla–Van Ryzin estimator from a marginal AFT model. A regularized bootstrap method is used to calibrate the test. Our test is more powerful and less conservative than the Bonferroni correction and other competing methods. This proposed method is evaluated in simulation studies and applied to two real data sets.

Table of Contents

List of Figures

Acknowledgements

While a completed dissertation bears the single name of a student, the whole process leading to its completion always accompanies the dedicated work of some other people. In this section, I wish to pay a sincere and public tribute to all the people who have ever shed glittering light on the path during this magical journey to my PhD degree.

Professor Ian Wray McKeague and Professor Min Qian, advisors of my dissertation, have been extremely helpful to me during the entire period of my work. Without their invaluable advice and lasting encouragement, this dissertation would not have been ever finished! I would like to express my full appreciation to them for being my trusty guides in academia, who can provide reliable guidance all the time but always yield abundant independence to me. Thanks to them for sustaining such an enlightening research environment, which is a moveable feast and will stay with me wherever I go for the rest of my life. Beyond research, more special thanks go to Professor Ian Wray McKeague for his phenomenal taste and sharing in art, wine, traveling and dog raising, which indeed added more spice to my academic life.

I would also like to thank my dissertation committee chair, Dr. Wei-Yann Tsai, and other members: Dr. Zhezhen Jin, Dr. Yang Feng and Dr. Bodhisattva Sen, for their helpful suggestions. Cordial acknowledgments are also due to my supervisors at

Columbia Center for Children's Environmental Health: Dr. Shuang Wang and Dr. Frederica P. Perera, and supervisors at Statistical Analysis Center: Dr. John L. P. (Seamus) Thompson and Dr. Bruce Levin. Thanks to them, I had funding resources to afford my student life and experienced impressive collaborations with experts from various disciplines.

With a special mention to Dr. Tsong-Cherng Lee from the department of mathematics in National Taiwan University, he was the instructor of my Calculus class and paved the way for my further studies in biostatistics. I would never have thought about seeking my PhD degree without his inspiration. I also have to thank Ms. I-Chen Lin, who has been one of my best friends and the eternal cheerleader since high school. Acknowledgments also go to Dr. Philippe Barbe from National Center for Scientific Research in France and Cox Media Group for his interesting and insightful comments on my dissertation.

Thanks must go to all my fellow doctoral students for their emotional supports, feedback, cooperation and of course friendship. I will never forget their screams of joy resounding through our doctoral room whenever a significantly momentous event was reached or whenever their unruly feeling just slipped the leash. I am also grateful to following department staff: Georgia A. Andre, Katy Hardy, Luminita Hellmann, Justine Herrera and Kevin Lee for their unfailing assistance with miscellaneous problems.

Last but certainly not least, I owe my unqualified gratitude to my family. My late maternal grandparents Ya Chen Kuo and Chi-Chen Kuo took care of me and taught me perseverance by their deed in my childhood. They founded my tough

attitude toward all the difficulties I have encountered in my research. I am also extremely grateful to my late paternal grandfather, Chun-Hsi Huang, and my paternal grandmother Chin-Tao Cheng-Huang for their generous moral supports. The deepest gratitude should be given to my dedicated parents, Yin-Wei Kuo and Chao-Cheng Huang, who raised me up by love and care. I could never have accomplished this dissertation without their supports and understanding along the way.

Every challenging work needs self efforts and guidance from the veterans.

My humble effort I dedicate to my loving and caring

Father, Mother and Grandparents,

whose sustaining affection and supports make me able to fulfill myself.

Introduction

The problem of detecting informative predictors of a survival outcome has received much attention over the past decade, especially since the advent of high-throughput genomic data. For example, a specific gene expression may influence a patient's survival time from diffuse large B-cell lymphoma (DLBCL), and how to discover such associations from massive collections of gene expression data still remains a challenging issue. Motivated by the DLBCL study (Rosenwald et al. (2002)), we consider the fundamental detection problem of whether there exists at least one predictor (or genetic feature) that is associated with the survival outcome in the presence of right-censoring.

To address this problem, we develop an adaptive resampling test for survival data (ARTS), related to the approach developed by McKeague and Qian (2015) (henceforth MQ) for uncensored outcomes. This test provides marginal screening of the predictors along with rigorous control of the family-wise error rate (FWER) resulting from the implicit multiple testing. Our testing procedure is further able to adjust for low-dimensional baseline clinical covariates that are not included in the systematic screening of the gene expression measurements. To identify the full set of active predictors, we further propose a forward-stepwise version of the ARTS

procedure that adjusts for previously-included predictors at each step, and continues until no further significant predictors are found.

We specify the link between the survival outcome and the predictors in terms of a general semiparametric accelerated failure time (AFT) model that does not make any distributional assumption on the error term. Our approach also applies when the error distribution is modeled parametrically (as in Kalbfleisch and Prentice (2002), Medeiros et al. (2014)) but we will focus on the semiparametric case. Let T be the (log-transformed) time-to-event outcome, and $\boldsymbol{U} = (U_1, \ldots, U_p)^T$ denote a p-dimensional vector of predictors. Here p can be large, although it is taken to be fixed for the purpose of developing the asymptotic theory. The AFT model is given by

$$T = \alpha_0 + \boldsymbol{U}^T \boldsymbol{\beta}_0 + \varepsilon, \tag{1}$$

where $\alpha_0 \in \mathbb{R}$ is an intercept, and $\boldsymbol{\beta}_0 \in \mathbb{R}^p$ is a vector of regression coefficients. We assume that the error term ε has zero mean, finite variance, and is uncorrelated with $\boldsymbol{U}$. The transformed survival outcome T is possibly right-censored by C, which is assumed independent of $(T, \boldsymbol{U})$ and bounded above by τ, the time to the end of the follow-up. We also make the standard assumption that $P(T \leq C) > 0$ to ensure that enough failure times can be observed over the follow-up period (asymptotically).

In the framework of semiparametric AFT models, Koul et al. (1981) (henceforth KSV) introduced the technique of inversely weighting the observed outcomes by the Kaplan–Meier estimate for the censoring, in order to apply standard least squares estimators from the uncensored linear model. Subsequently, two more sophisticated

methods were proposed to fit the semiparametric AFT model. The Buckley–James estimator replaces the censored survival outcome by the conditional expectation of T given the data (Buckley and James (1979), Ritov (1990)). The rank based method is an estimating equation approach formulated in terms of the partial likelihood score function (Tsiatis (1990), Lai and Ying (1991a), Lai and Ying (1991b), Ying (1993), Jin et al. (2003)). Our proposed marginal screening test will be based on the KSV estimator which has the advantage over the Buckley–James and rank-based methods that it preserves a direct link with the linear model; in particular it maintains the marginal correlations between the inversely weighted response and the predictors.

An especially attractive feature of the AFT model is that the marginal association between T and each predictor can be directly represented in terms of correlation. As we will see below, this allows a reduction of the high-dimensional screening problem to a single test of whether the most correlated predictor with T is significant. The most popular approach to the screening of predictors in the survival analysis setting is to use relative or excess conditional hazard function representations of association. However, the AFT approach has the advantage that the lack of any marginal correlation implies the absence of any correlation between T and $\boldsymbol{U}$; in the hazard rate setting, there is no such connection.

Another attractive feature of the AFT model is that it is relatively insensitive to unmeasured heterogeneity because the error term can act as a latent variable representing omitted confounders (Keiding et al. (1997)). In hazard rate approaches the inclusion of latent variables is typically handled using inflexible parametric frailty models that are not easily applied in practice. In general, the presence of unmeasured

heterogeneity causes the attenuation of parameter estimates, and this is especially pronounced in hazard rate approaches, such as the Cox model or additive risk models (Lin and Ying (1994), McKeague and Sasieni (1994)). On the other hand, such attenuation is much less problematic for the AFT model because the error term is only assumed to be uncorrelated with the predictors and requires no special distributional assumption.

Under the AFT model (1), we are interested in testing the null hypothesis $\boldsymbol{\beta}_0 = 0$, i.e., that no predictor is linearly associated with T, against the omnibus alternative. The data consist of iid copies $(X_i, \delta_i, \boldsymbol{U}_i), i = 1, \ldots, n$, of $(X, \delta, \boldsymbol{U})$, where $X = \min(T, C)$ and $\delta = 1(T \leq C)$. The idea of the ARTS marginal screening procedure is to fit a series of working AFT models only using one component of $\boldsymbol{U}$ at a time, and then select the marginal KSV regression parameter estimate $\hat{\theta}_n$ that has maximal absolute value. When the predictors are pre-standardized, the maximal regression parameter corresponds to the maximal correlation between T and any component of $\boldsymbol{U}$, motivating $\sqrt{n}\hat{\theta}_n$ as a suitable test statistic. The limiting distribution of this test statistic is non-regular (discontinuous at zero as a function of $\boldsymbol{\beta}_0$), causing difficulties in calibrating the test, as explained in the standard linear regression setting by MQ. Further, the presence of censoring introduces additional (discontinuous) dispersion in the limiting distribution of $\sqrt{n}\hat{\theta}_n$ that needs to be addressed.

The marginal KSV estimates stem from regressing the estimated synthetic response $Y = \delta X / \hat{G}_n(X-)$ on successive components of $\boldsymbol{U}$, where Y is regarded as an inverse probability weighted estimate; $\hat{G}_n$ is the standard Kaplan–Meier estimator of the survival function of C (denoted by G_0). Under independent censoring (as stated

earlier), the use of least squares estimators based on treating Y as a response variable is justified in view of the uniform consistency of $\hat{G}_n$ under mild conditions (e.g., when the distribution functions of T and C have no common jumps, see Stute and Wang (1993)). Independent censoring is a common assumption made in high-dimensional screening of predictors for survival outcomes (He et al. (2013), Song et al. (2014), Li et al. (2016)). It is much less restrictive, however, only to assume that T and C are conditionally independent given $\boldsymbol{U}$, in which case the conditional survival function $G_0(\cdot|\boldsymbol{U})$ of C given $\boldsymbol{U}$ can depend on the predictors. The estimation of $G_0(\cdot|\boldsymbol{U})$ is challenging unless there is prior knowledge that only a single predictor is involved, using a local Kaplan–Meier estimator (Dabrowska (1989)). For simplicity, however, we will assume independent censoring throughout.

Variable selection methods for right-censored survival data are widely available, although formal testing procedures are much less developed. For example, variants of regularized Cox regression have been studied by Tibshirani (1997), Fan and Li (2002), Bunea and McKeague (2005), Zhang and Lu (2007), Bøvelstad et al. (2009), Engler and Li (2009), Antoniadis et al. (2010), Binder et al. (2011), Wu (2012), and Sinnott and Cai (2016). Penalized AFT models have been considered by Huang et al. (2006), Datta et al. (2007), Johnson (2008), Johnson et al. (2008), Cai et al. (2009), Huang and Ma (2010), Bradic et al. (2011), Ma and Du (2012), and Li et al. (2014). These methods only ensure the consistency of variable selection (i.e., the oracle property) and do not address the issue of post-selection inference. Fang et al. (2016) have established asymptotically valid confidence intervals for a preconceived regression parameter in a high-dimensional Cox model after variable selection on

the remaining predictors, but this does not apply to marginal screening (where no regression parameter is singled-out a priori). Zhong et al. (2015) have considered the same problem for preconceived regression parameters within a high-dimensional additive risk model. Taylor and Tibshirani (2017) recently proposed a method of finding post-selection corrected p-values and confidence intervals for the Cox model based on conditional testing, but their method has not been explored theoretically (except in the linear regression setting with independent normal errors, see Lockhart et al. (2014)) as far as we know.

Statistical methods for variable selection based on marginal screening on survival data have been studied by Fan et al. (2010), who extended sure independence screening to survival outcomes based on the Cox model. Their method applies to the selection of components of ultra-high dimensional predictors, although no formal testing is available. Other relevant references include Zhao and Li (2012), Gorst-Rasmussen and Scheike (2013), He et al. (2013), Song et al. (2014), Zhao and Li (2014), Hong et al. (2016), Li et al. (2016), and Hong et al. (2017).

This work is organized as follows. In Chapter 1 we formulate the testing problem, and introduce the maximally selected KSV estimator. Then we propose a way of adapting to the non-regularity of its limiting distribution, which leads to the ARTS procedure. Consistency of the bootstrap used to calibrate ARTS is provided at the end of Chapter 1. In Chapter 2 we propose a variant of ARTS that adjusts for the effect of baseline clinical covariates. A forward-stepwise ARTS procedure is developed in Chapter 3, to successively identify all of the influential predictors while adjusting for previously-included predictors. Competing methods are discussed in Chapter 4.

Numerical results reported in Chapter 5 show that ARTS compares favorably with the competing methods. In Chapter 6 we present applications to gene expression data and primary biliary cirrhosis data. Concluding remarks are given in the last chapter. Proofs of all the results are provided in the supplementary materials.

Chapter 1

Adaptive resampling test for survival data

As mentioned in the Introduction, the ARTS procedure will be developed under the AFT model specified in (1) on the basis of iid right-censored data under that model. We assume that U_j has a finite fourth moment for $j = 1, \ldots, p$ and $|\mathrm{Corr}(U_j, U_k)| < 1$ for all $j \neq k$. The unknown survival function of the censoring, $G_0(t)$, is assumed to be continuous over $t \in \mathcal{T} = (-\infty, \tau]$, where τ is the end of follow-up. Further we assume that $G_0(\tau) > 0$ to prevent the Kaplan–Meier estimator $\hat{G}_n$ of G_0 from being too close to zero, to ensure the stable performance of the KSV estimator.

1 Maximally selected KSV estimator

Since T in model (1) may be censored, Koul et al. suggested replacing T by the synthetic response $\tilde{Y} = \delta X / G_0(X-)$ based on the equation below:

$$E[\tilde{Y}|\boldsymbol{U}] = E\left[\frac{\delta X}{G_0(X-)} \mid \boldsymbol{U}\right] = E\left[\frac{T}{G_0(T-)} E\left[\delta|T\right] \mid \boldsymbol{U}\right] = E[T|\boldsymbol{U}]. \tag{1.1}$$

This equation implies that T and $\tilde{Y}$ share the conditional mean on $\boldsymbol{U}$ under the assumption of independent censoring. Namely, we can see $\tilde{Y} = \alpha_0 + \boldsymbol{U}^T\boldsymbol{\beta}_0 + \tilde{\varepsilon}$, where the error term $\tilde{\varepsilon}$ is different from ε in model (1) but still has zero mean, finite variance, and is uncorrelated with $\boldsymbol{U}$ (see the supplementary materials for the proof). Using

similar arguments in (1.1), we can also show that $E[U_j\tilde{Y}] = E[U_jT]$ for $j = 1, \ldots, p$.

To get around the test on high-dimensional $\boldsymbol{\beta}_0$, we further tailor our hypotheses as follows. Define

$$j(\boldsymbol{b}) = \arg \max_{j=1,\ldots,p} |\mathrm{Corr}(U_j, \boldsymbol{U}^T\boldsymbol{b})| \text{ for any } \boldsymbol{b} \in \mathbb{R}^p. \tag{1.2}$$

Under model (1), it is natural to have $\mathrm{Corr}(U_j, T) = \mathrm{Corr}(U_j, \boldsymbol{U}^T\boldsymbol{\beta}_0)$, which indicates that $j(\boldsymbol{\beta}_0) = \arg \max_{j=1,\ldots,p} |\mathrm{Corr}(U_j, T)|$. We assume the uniqueness of $j(\boldsymbol{\beta}_0)$ when $\boldsymbol{\beta}_0 \neq \mathbf{0}$. Testing whether $\boldsymbol{\beta}_0 = \mathbf{0}$ is equivalent to a test of

$$H_0 : \theta_0 = 0 \quad \text{versus} \quad H_A : \theta_0 \neq 0,$$

where θ_0 denotes the marginal regression coefficient of $U_{j(\boldsymbol{\beta}_0)}$ (the most correlated predictor to T). The predictor $U_{j(\boldsymbol{\beta}_0)}$ also maximizes the marginal correlation with $\tilde{Y}$, because the aforementioned properties of $\tilde{Y}$ ensures that

$$\arg \max_{j=1,\ldots,p} |\mathrm{Corr}(U_j, T)| = \arg \max_{j=1,\ldots,p} |\mathrm{Corr}(U_j, \tilde{Y})|. \tag{1.3}$$

For notational simplicity, we denote the label $j(\boldsymbol{\beta}_0)$ by j_0 henceforth.

The further idea is to replace the "true but unknown" synthetic response $\tilde{Y}$ in (1.1) by the estimated synthetic response $Y = \delta X/\hat{G}_n(X-)$, and it gives the estimator of

j_0 by

$$\hat{j}_n = \arg\max_{j=1,\ldots,p} \left| \frac{\mathbb{P}_n(U_j - \mathbb{P}_n U_j)Y}{S_j S_Y} \right|, \tag{1.4}$$

where $\mathbb{P}_n$ is the empirical distribution; S_j and S_Y are the sample standard deviations of U_j and Y, respectively. Here, although we recommend that the components of $\boldsymbol{U}$ should be pre-standardized in ARTS, we find it more natural to present the procedure in terms of unstandardized variables.

Let $a_0 \in \mathbb{R}$ denote the intercept in a marginal linear model for T with predictor U_{j_0}, and we have

$$(a_0, \theta_0) = \left(ET - \theta_0 E U_{j_0}, \frac{\mathrm{Cov}(U_{j_0}, T)}{\mathrm{Var}(U_{j_0})} \right).$$

The maximally selected KSV estimator among all the marginal AFT models for Y gives the estimator of (a_0, θ_0) by

$$(\hat{a}_n, \hat{\theta}_n) = \left(\mathbb{P}_n Y - \hat{\theta}_n \mathbb{P}_n U_{\hat{j}_n}, \ \frac{1}{S^2_{\hat{j}_n}} \mathbb{P}_n (U_{\hat{j}_n} - \mathbb{P}_n U_{\hat{j}_n}) Y \right), \tag{1.5}$$

where $S^2_{\hat{j}_n}$ denotes the sample variance of $U_{\hat{j}_n}$.

2 Limiting distribution of $\hat{\theta}_n$ under a local model

The main test statistic is based on $\sqrt{n}\hat{\theta}_n$; however, the limiting distribution of $\sqrt{n}\hat{\theta}_n$ is discontinuous at $\boldsymbol{\beta}_0 = \mathbf{0}$. To examine the local asymptotic behavior of $\sqrt{n}\hat{\theta}_n$, we adopt a local model. Given sample size n, the model (1) can be re-specified as

$$T^{(n)} = \alpha_0 + \boldsymbol{U}^T \boldsymbol{\beta}_n + \varepsilon, \tag{1.6}$$

where $\boldsymbol{\beta}_n = \boldsymbol{\beta}_0 + \boldsymbol{b}_0/\sqrt{n}$ with a local parameter $\boldsymbol{b}_0$, $\boldsymbol{b}_0 \in \mathbb{R}^p$.

Under this local model (1.6), the observed time and the censoring status are redefined as $X^{(n)} = \min(T^{(n)}, C)$ and $\delta^{(n)} = 1(T^{(n)} \leq C)$, respectively. We can also create the synthetic response $\tilde{Y}^{(n)}$ and the estimated synthetic responses $Y^{(n)}$ in an analogous fashion, where

$$\tilde{Y}^{(n)} = \frac{\delta^{(n)} X^{(n)}}{G_0(X^{(n)}-)} \text{ and } Y^{(n)} = \frac{\delta^{(n)} X^{(n)}}{\hat{G}_n(X^{(n)}-)}.$$

For any fixed n, $\tilde{Y}^{(n)}$ has the same mean and the same covariance with $\boldsymbol{U}$ as $T^{(n)}$. The corresponding error term for $\tilde{Y}^{(n)}$ is expressed as $\tilde{\varepsilon}_n = \tilde{Y}^{(n)} - \alpha_0 - \boldsymbol{U}^T\boldsymbol{\beta}_n$, where $\tilde{\varepsilon}_n$ also has zero mean and is uncorrelated with $\boldsymbol{U}$ for any fixed n. Instead of j_0, the label of the most correlated predictor with $T^{(n)}$ is

$$j_n \equiv j(\boldsymbol{\beta}_n) = \arg \max_{j=1,\ldots,p} |\text{Corr}(U_j, T^{(n)})| = \arg \max_{j=1,\ldots,p} |\text{Corr}(U_j, \tilde{Y}^{(n)})|,$$

and our earlier hypotheses extend to

$$H_0 : \theta_n = 0 \quad \text{versus} \quad H_A : \theta_n \neq 0,$$

where

$$\theta_n = \frac{\text{Cov}(U_{j_n}, T^{(n)})}{\text{Var}(U_{j_n})}. \tag{1.7}$$

Note that $j_n = j(\boldsymbol{b}_0)$ when $\boldsymbol{\beta}_0 = \boldsymbol{0}$ but $\boldsymbol{b}_0 \neq \boldsymbol{0}$ and $j(\boldsymbol{b}_0)$ is assumed unique. Otherwise, j_n is not well-defined and the null hypothesis $\theta_n = 0$ holds when $\boldsymbol{\beta}_0 = \boldsymbol{0}$ and

$\boldsymbol{b}_0 = \mathbf{0}$. If j_0 is unique, then $j_n \to j_0$. Our estimators for θ_n and j_n are, respectively,

$$\hat{\theta}_n = \frac{1}{S^2_{\hat{j}_n}} \mathbb{P}_n (U_{\hat{j}_n} - \mathbb{P}_n U_{\hat{j}_n}) Y^{(n)} \quad \text{and} \quad \hat{j}_n = \arg\max_{j=1,\ldots,p} \left| \frac{\mathbb{P}_n (U_j - \mathbb{P}_n U_j) Y^{(n)}}{S_j S_{Y^{(n)}}} \right|, \tag{1.8}$$

where $S_{Y^{(n)}}$ is the sample standard deviation of $Y^{(n)}$.

Given the conditions stated above for $\boldsymbol{U}$ and G_0, the limiting distribution of $\sqrt{n}\hat{\theta}_n$ is provided in Theorem 1 below. The proof of Theorem 1 is developed based on functional delta method (van der Vaart (2000), Chap. 20) and functional central limit theorem (Pollard (1990), Sec. 10), and will be provided in the supplementary materials.

Theorem 1. *Suppose that $j_0 = j(\boldsymbol{\beta}_0)$ is unique when $\boldsymbol{\beta}_0 \neq \mathbf{0}$; that $j(\boldsymbol{b}_0)$ is unique when $\boldsymbol{\beta}_0 = \mathbf{0}$ and $\boldsymbol{b}_0 \neq \mathbf{0}$, and that some mild regularity conditions stated hold. Then, under the local model (1.6),*

$$\sqrt{n}(\hat{\theta}_n - \theta_n) \xrightarrow{d} \begin{cases} (M_{j_0} + \varphi_{j_0}(\mathbb{L}))/V_{j_0} & \text{if } \boldsymbol{\beta}_0 \neq \mathbf{0}, \\ (M_J + \varphi_J(\mathbb{L}))/V_J + (C_J/V_J - C_{j(\boldsymbol{b}_0)}/V_{j(\boldsymbol{b}_0)})^T \boldsymbol{b}_0 & \text{if } \boldsymbol{\beta}_0 = \mathbf{0}, \end{cases}$$

where $V_j = \mathrm{Var}(U_j)$; $C_j = \mathrm{Cov}(U_j, \boldsymbol{U})$; $J = \arg\max_{j=1,\ldots,p}\{M_j + \varphi_j(\mathbb{L}) + C_j^T \boldsymbol{b}_0\}^2/V_j$; $\boldsymbol{M} = \{M_j, j = 1, \ldots, p\}$ is a mean-zero normal random vector; $\mathbb{L}$ is a mean-zero Gaussian process, and $(\boldsymbol{M}, \mathbb{L})$ is also a mean-zero Gaussian process whose covariance is provided in the supplementary materials. The j-indexed functional $\varphi_j \colon \ell^\infty_\tau \to \mathbb{R}$ is defined by

$$\varphi_j(h) = E\left[\frac{(U_j - EU_j) T h(T-)}{G_0(T-)} \right],$$

where ℓ_τ^∞ denotes the space of bounded functions on $\mathcal{T}$ and any function $h \in \ell_\tau^\infty$.

Remark 1. *The Gaussian process $\mathbb{L}$ is the weak limit of the process $\sqrt{n}(\hat{G}_n - G_0)$. When there is no censoring, $\hat{G}_n(t) = G_0(t) = 1$ for all t so that $\mathbb{L}$ is a zero process. Then, $\varphi_j(\mathbb{L}) = 0$ for all j, and the limiting distribution reduces to that given by MQ. When there is censoring, $\mathbb{L}$ is a non-trivial Gaussian process and introduces further dispersion in our limiting distribution.*

Remark 2. *When there is censoring and $\boldsymbol{\beta}_0 \neq \mathbf{0}$, we have T and $\boldsymbol{U}$ correlated, leading to non-zero $\varphi_j(\mathbb{L})$ for all j. Along with the non-trivial process $\mathbb{L}$, the additional term $\varphi_{j_0}(\mathbb{L})$ will be present.*

Remark 3. *When there is censoring and $\boldsymbol{\beta}_0 = \mathbf{0}$, $\varphi_j(\mathbb{L})$ will vanish everywhere almost surely (a.s.) for all j, if ε and $\boldsymbol{U}$ are independent. This leads to the additional term $\varphi_J(\mathbb{L})$ disappearing. Given the independence between ε and $\boldsymbol{U}$, the limiting distribution simplifies to*

$$M_J/V_J + (C_J/V_J - C_{j(\boldsymbol{b}_0)}/V_{j(\boldsymbol{b}_0)})^T \boldsymbol{b}_0.$$

This less complex form of the limiting distribution can be easily estimated from the data, and it suggests not only the possibility of evaluating asymptotic power (discussed in Section 5), but also calibration via simulation from the estimated null limiting distribution of $\sqrt{n}\hat{\theta}_n$ (later introduced as "CEND" in Section 4). However, the validity of this approach relies on the highly restrictive assumption that ε and $\boldsymbol{U}$ are independent.

The discontinuity of the limiting distribution at $\boldsymbol{\beta}_0 = \mathbf{0}$ introduces difficulties for

designing a screening test based on $\hat{\theta}_n$. If $\boldsymbol{\beta}_0 \neq \mathbf{0}$, naive resampling methods can give consistent estimates of the limiting distribution of $\sqrt{n}(\hat{\theta}_n - \theta_n)$. If $\boldsymbol{\beta}_0 = \mathbf{0}$, resampling methods that fail to take the local behavior of $\sqrt{n}\hat{\theta}_n$ around $\boldsymbol{\beta}_0 = \mathbf{0}$ into account will give inconsistent estimates of the limiting distribution. To accommodate this non-uniform weak convergence at the point of non-regularity (i.e., $\boldsymbol{\beta}_0 = \mathbf{0}$), our proposed ARTS allows for the flexibility of using different bootstrap strategies to approximate the limiting distribution when $\boldsymbol{\beta}_0 \neq \mathbf{0}$ and when $\boldsymbol{\beta}_0 = \mathbf{0}$. Recall that S_j^2 is the sample variance of U_j for all j. We decompose $\sqrt{n}(\hat{\theta}_n - \theta_n)$ into

$$\sqrt{n}(\hat{\theta}_n - \theta_n)1(|\mathbb{T}_n| > \lambda_n \text{ or } \boldsymbol{\beta}_0 \neq \mathbf{0}) + \sqrt{n}(\hat{\theta}_n - \theta_n)1(|\mathbb{T}_n| \leq \lambda_n, \boldsymbol{\beta}_0 = \mathbf{0}), \tag{1.9}$$

where $\mathbb{T}_n = \sqrt{n}\hat{\theta}_n/\hat{\sigma}_n$ is the maximally selected studentized statistic for the pretest and

$$\hat{\sigma}_n^2 = \mathbb{P}_n(Y - \hat{a}_n - \hat{\theta}_n U_{\hat{j}_n})^2 / S_{\hat{j}_n}^2. \tag{1.10}$$

with $(\hat{a}_n, \hat{\theta}_n, \hat{j}_n)$ defined in (1.4) and (1.5). Since Koul et al. (1981) did not provide a computationally feasible estimator for the complex asymptotic variance of the KSV estimator, it is intractable to build our pretest statistic based on their result. Instead, we develop the studentized statistic above and its limiting distribution at the null based on Theorem 1. We show $\hat{\sigma}_n^2$ is asymptotically bounded away from zero and bounded above (the proof is provided in the supplementary materials). Together with results in Theorem 1, we then prove that $|\mathbb{T}_n| \xrightarrow{a.s.} \infty$ when $\boldsymbol{\beta}_0 \neq \mathbf{0}$ and $|\mathbb{T}_n| = O_p(1)$ when $\boldsymbol{\beta}_0 = \mathbf{0}$. The specification of λ_n will be presented in the next section.

We isolate the possibility of $\boldsymbol{\beta}_0 = \mathbf{0}$ by comparing $|\mathbb{T}_n|$ with some screening threshold λ_n. The first term in (1.9) can be consistently estimated by centered percentile bootstrap whenever $\lambda_n = o(\sqrt{n})$ and $\lambda_n \to \infty$, because we show $1(|\mathbb{T}_n| > \lambda_n) \xrightarrow{p} 1(\boldsymbol{\beta}_0 \neq \mathbf{0})$ (stated as Lemma 10 in the supplementary materials along with the proof thereof). For estimating the second term in (1.9), it entails more work. Recall that $\mathbb{P}_n$ is the empirical distribution; P is the distribution of $(X^{(n)}, \delta^{(n)}, \boldsymbol{U})$, and $\mathbb{G}_n = \sqrt{n}(\mathbb{P}_n - P)$. For $j = 1, \ldots, p$, we define

$$\mathbb{M}_{n,j} = \mathbb{G}_n \tilde{\varepsilon}_n (U_j - \mathbb{P}_n U_j) \text{ and } \mathbb{D}_{n,j} = \sqrt{n}\mathbb{P}_n (U_j - \mathbb{P}_n U_j)(Y^{(n)} - \tilde{Y}^{(n)}). \tag{1.11}$$

For $\boldsymbol{b} \in \mathbb{R}^p$, we define

$$J_n(\boldsymbol{b}) = \arg\max_{j=1,\ldots,p} (\mathbb{M}_{n,j} + \mathbb{D}_{n,j} + \mathbb{P}_n (U_j - \mathbb{P}_n U_j) U^T \boldsymbol{b})^2 / S_j^2, \tag{1.12}$$

and a $\boldsymbol{b}$-indexed process

$$\mathbb{Q}_n(\boldsymbol{b}) = (\mathbb{M}_{n,J_n(\boldsymbol{b})} + \mathbb{D}_{n,J_n(\boldsymbol{b})} + \mathbb{P}_n (U_{J_n(\boldsymbol{b})} - \mathbb{P}_n U_{J_n(\boldsymbol{b})}) U^T \boldsymbol{b}) / S_{J_n(\boldsymbol{b})}^2 - C_{j(\boldsymbol{b})}^T \boldsymbol{b} / V_{j(\boldsymbol{b})}. \tag{1.13}$$

Below we express the second term in (1.9) as a function $\mathbb{Q}_n(\boldsymbol{b}_0)$. When $\boldsymbol{\beta}_0 = \mathbf{0}$, it is easy to see

$$\begin{aligned}\sqrt{n}\hat{\theta}_j &= \sqrt{n}\mathbb{P}_n(U_j - \mathbb{P}_n U_j)\tilde{Y}^{(n)}/S_j^2 + \sqrt{n}\mathbb{P}_n(U_j - \mathbb{P}_n U_j)(Y^{(n)} - \tilde{Y}^{(n)})/S_j^2 \\ &= (\mathbb{G}_n\tilde{\varepsilon}_n(U_j - \mathbb{P}_n U_j) + \sqrt{n}\mathbb{P}_n(U_j - \mathbb{P}_n U_j)(Y^{(n)} - \tilde{Y}^{(n)}) + \mathbb{P}_n(U_j - \mathbb{P}_n U_j)\boldsymbol{U}^T\boldsymbol{b}_0)/S_j^2 \\ &= (\mathbb{M}_{n,j} + \mathbb{D}_{n,j} + \mathbb{P}_n(U_j - \mathbb{P}_n U_j)U^T\boldsymbol{b}_0)/S_j^2,\end{aligned} \tag{1.14}$$

for all j. Along with $\hat{j}_n = J_n(\boldsymbol{b}_0)$ and $j_n = j(\boldsymbol{b}_0)$ when $\boldsymbol{\beta}_0 = \mathbf{0}$, we have $\sqrt{n}\theta_n = C^T_{j(\boldsymbol{b}_0)}\boldsymbol{b}_0/V_{j(\boldsymbol{b}_0)}$, and therefore $\sqrt{n}(\hat{\theta}_n - \theta_n) = \mathbb{Q}_n(\boldsymbol{b}_0)$. Hence, the decomposition of $\sqrt{n}(\hat{\theta}_n - \theta_n)$ can be further expressed as

$$\sqrt{n}(\hat{\theta}_n - \theta_n) = \sqrt{n}(\hat{\theta}_n - \theta_n)1(|\mathbb{T}_n| > \lambda_n \text{ or } \boldsymbol{\beta}_0 \neq \mathbf{0}) + \mathbb{Q}_n(\boldsymbol{b}_0)1(|\mathbb{T}_n| \leq \lambda_n, \boldsymbol{\beta}_0 = \mathbf{0}). \tag{1.15}$$

In Theorem 2 below, we show $\mathbb{Q}_n(\boldsymbol{b})$ can be consistently bootstrapped for any given $\boldsymbol{b}$. Provided $\boldsymbol{b}_0$ known, we can directly bootstrap the expression in (1.15) to consistently estimate the limiting distribution of $\sqrt{n}(\hat{\theta}_n - \theta_n)$. Hereafter, a superscript $*$ is imposed to indicate the bootstrap version of all the estimators.

Theorem 2. *Suppose that all conditions for Theorem 1 hold, and the tuning parameter λ_n satisfies $\lambda_n = o(\sqrt{n})$ and $\lambda_n \to \infty$ as $n \to \infty$. Under the local model (1.6),*

$$\sqrt{n}(\hat{\theta}^*_n - \hat{\theta}_n)1(|\mathbb{T}^*_n| > \lambda_n \text{ or } |\mathbb{T}_n| > \lambda_n) + \mathbb{Q}^*_n(\boldsymbol{b}_0)1(|\mathbb{T}^*_n| \leq \lambda_n, |\mathbb{T}_n| \leq \lambda_n)$$

converges to the limiting distribution of $\sqrt{n}(\hat{\theta}_n - \theta_n)$ *conditionally (on the data) in probability.*

3 ARTS screening procedure

The ARTS screening procedure uses a bootstrap calibration for the test statistic $\sqrt{n}\hat{\theta}_n$ based on a special case of Theorem 2, specifically $\boldsymbol{b}_0 = \mathbf{0}$. To approximate the limiting distribution of $\sqrt{n}\hat{\theta}_n$ under the null, it suffices to bootstrap

$$B_n = \sqrt{n}(\hat{\theta}_n - \theta_n)1(|\mathbb{T}_n| > \lambda_n \text{ or } \boldsymbol{\beta}_0 \neq \mathbf{0}) + \mathbb{Q}_n(\mathbf{0})1(|\mathbb{T}_n| \leq \lambda_n, \boldsymbol{\beta}_0 = \mathbf{0}), \quad (1.16)$$

and the corresponding bootstrap version is

$$B_n^* = \sqrt{n}(\hat{\theta}_n^* - \hat{\theta}_n)1(|\mathbb{T}_n^*| > \lambda_n \text{ or } |\mathbb{T}_n| > \lambda_n) + \mathbb{Q}_n^*(\mathbf{0})1(|\mathbb{T}_n^*| \leq \lambda_n, |\mathbb{T}_n| \leq \lambda_n). \quad (1.17)$$

For some nominal level α, define the critical points c_l and c_u, respectively, by the lower and upper $100(\alpha/2)$-th percentiles of 1000 replications of B_n^*. We reject the null hypothesis and conclude that there is at least one significant predictor, if $\sqrt{n}\hat{\theta}_n$ falls outside the interval $[c_l, c_u]$.

Given the conditions that $\lambda_n = o(\sqrt{n})$ and $\lambda_n \to \infty$, the pretest demonstrates asymptotically negligible Type I error rate $P(|\mathbb{T}_n| > \lambda_n | \theta_n = 0) \to 0$ because we have shown that $P(|\mathbb{T}_n| > \lambda_n) \to 1(\boldsymbol{\beta}_0 \neq \mathbf{0})$ in Lemma 10 in the supplementary materials. Provided the independence between $\tilde{\varepsilon}$ and $\boldsymbol{U}$, a special case of Theorem 1 indicates that $\mathbb{T}_n \xrightarrow{d} \max_{j=1,\ldots,p} |Z_j|$ at the null, where $\{Z_j, j = 1, \ldots, p\}$ is a vector of standard normal random variables. Using similar arguments as in MQ's work, the

asymptotic Type I error rate of the pretest can be controlled below level α if we set $\lambda_n \geq \Phi^{-1}(1-\alpha/(2p))$, where Φ denotes the standard normal distribution function. To satisfy the conditions that $\lambda_n = o(\sqrt{n})$ and $\lambda_n \to \infty$, a reasonable selection of the threshold would be $\lambda_n = \max\{\sqrt{a \log n}, \Phi^{-1}(1-\alpha/(2p))\}$ for some constant $a > 0$.

To determine the value of the constant a in practice, the double bootstrap would be implemented. We produce 1000 bootstrap estimates $\hat{\theta}_n^*$, and apply ARTS on further generated 1000 nested double bootstrap samples to get the acceptance region $[c_l^*, c_u^*]$ for each $\hat{\theta}_n^*$. If the test statistic $\sqrt{n}(\hat{\theta}_n^* - \hat{\theta}_n)$ falls beyond $[c_l^*, c_u^*]$, we claim the ARTS procedure to reject. The constant a would be specified by the value that can have 5% of these 1000 ARTS procedures be rejected. This data-driven selection of a will be adopted in our numerical studies and applications to real data.

Chapter 2

ARTS adjusted for baseline covariates

When screening high-dimensional pedictors of survival outcomes, it is common practice to adjust for baseline demographic and clinical covariates. These baseline covariates include age, disease stage, tumor thickness, and lymph node status; in the DLBCL study, we have the International Prognostic Index (IPI). The IPI is a widely-used prognostic index developed by the combination of clinical covariates (cf. TIN (1993)). Such baseline covariates (with a moderate dimensionality) do not need to be screened, but need to be incorporated as covariates in the AFT model used in ARTS. In this section, we modify ARTS (as *adjusted ARTS*) in a way that accounts for the effect of these covariates.

Let $\tilde{\boldsymbol{U}} = (\tilde{U}_1, \ldots, \tilde{U}_q)^T$ be a vector of baseline covariates. With $\tilde{\boldsymbol{U}}$ included, the true AFT model (1) can be further expressed as

$$T = \alpha_0 + \boldsymbol{U}^T \boldsymbol{\beta}_0 + \tilde{\boldsymbol{U}}^T \boldsymbol{\gamma}_0 + \varepsilon, \tag{2.1}$$

where $\boldsymbol{\gamma}_0 \in \mathbb{R}^q$; $\tilde{\boldsymbol{U}}$ is assumed to have a finite fourth moment, and the error term ε is also uncorrelated with $\tilde{\boldsymbol{U}}$. Our interest is to test whether $\boldsymbol{\beta}_0 = \boldsymbol{0}$, which includes adjustment for $\tilde{\boldsymbol{U}}$. Projecting $\tilde{\boldsymbol{U}}$ on the space spanned by $\boldsymbol{U}$, we reformulate the AFT model (2.1) as

$$T = \alpha_0' + \boldsymbol{D}^T\boldsymbol{\beta}_0 + \varepsilon', \tag{2.2}$$

where $\boldsymbol{D} = (D_1, \ldots, D_p)^T$ with $D_j = U_j - \tilde{\alpha}_j - \tilde{\boldsymbol{U}}^T\tilde{\boldsymbol{\gamma}}_j$; meanwhile,

$$\begin{aligned}
&(\tilde{\alpha}_j, \tilde{\boldsymbol{\gamma}}_j^T) = (E[U_j] - E[\tilde{\boldsymbol{U}}^T\tilde{\boldsymbol{\gamma}}_j], (\Sigma_{\tilde{\boldsymbol{U}}}^{-1}\mathrm{Cov}(U_j, \tilde{\boldsymbol{U}}))^T)\,;\\
&\alpha_0' = \alpha_0 + (\tilde{\alpha}_1, \ldots, \tilde{\alpha}_p)\boldsymbol{\beta}_0 + E[\tilde{\boldsymbol{U}}^T((\tilde{\gamma}_1, \ldots, \tilde{\gamma}_p)\boldsymbol{\beta}_0 + \tilde{\gamma}_0)]\,;\\
&\varepsilon' = \tilde{\boldsymbol{U}}^T((\tilde{\gamma}_1, \ldots, \tilde{\gamma}_p)\boldsymbol{\beta}_0 + \tilde{\gamma}_0) - E[\tilde{\boldsymbol{U}}^T((\tilde{\gamma}_1, \ldots, \tilde{\gamma}_p)\boldsymbol{\beta}_0 + \tilde{\gamma}_0)] + \varepsilon,
\end{aligned}$$

and $\Sigma_{\tilde{\boldsymbol{U}}}$ is the covariance matrix of $\tilde{\boldsymbol{U}}$. Note that $\tilde{\alpha}_j + \tilde{\boldsymbol{U}}^T\tilde{\boldsymbol{\gamma}}_j$ is the best linear unbiased predictor of U_j based on $\tilde{\boldsymbol{U}}$. According to the definitions of $(\tilde{\alpha}_j, \tilde{\boldsymbol{\gamma}}_j)$, it is obvious that $E[D_j] = 0$ and $\mathrm{Cov}(D_j, \tilde{\boldsymbol{U}}^T\boldsymbol{\gamma}) = 0$, for all j and any vector $\boldsymbol{\gamma} \in \mathbb{R}^q$. The new error term ε' inherits the properties of ε and satisfies the moment conditions required for ARTS: $E[\varepsilon'] = 0$; $E[(\varepsilon')^2] < \infty$ and ε' is uncorrelated with $\boldsymbol{D}$. To test whether $\boldsymbol{\beta}_0 = \boldsymbol{0}$ under model (2.2), it suffices to test

$$H_0 : \theta_0' = 0 \quad \text{versus} \quad H_A : \theta_0' \neq 0,$$

where $\theta_0' = \mathrm{Cov}(D_{j'(\boldsymbol{\beta}_0)}, T)/\mathrm{Var}(D_{j'(\boldsymbol{\beta}_0)})$ and $j'(\boldsymbol{b}) = \arg\max_{j=1,\ldots,p} |\mathrm{Corr}(D_j, \boldsymbol{D}^T\boldsymbol{b})|$ for any $\boldsymbol{b} \in \mathbb{R}^p$, implying $j'(\boldsymbol{\beta}_0) = \arg\max_{j=1,\ldots,p} |\mathrm{Corr}(D_j, T)|$.

The idea of *adjusted ARTS* is to regress each screening predictor on baseline covariates and to apply ARTS with the corresponding residuals $\hat{\boldsymbol{D}} = (\hat{D}_1, \ldots, \hat{D}_p)^T$ as predictors. Since SLLN ensures that $\mathbb{P}_n D_j \xrightarrow{a.s.} 0$; $\mathbb{P}_n \hat{D}_j^2 = \mathbb{P}_n D_j^2$ a.s. and $\mathbb{P}_n \hat{D}_j Y^{(n)} = \mathbb{P}_n D_j Y^{(n)}$ a.s. for all j, we can straightforwardly justify the validity

of this adjustment by following arguments used for ARTS with $\boldsymbol{U}$ replaced by $\hat{\boldsymbol{D}}$. The bootstrap consistency can also be guaranteed, so we only need to resample residuals in the procedures of bootstrap and double bootstrap. This saves computation cost caused by implementing projections every time when we have bootstrap or double bootstrap samples. The saving is evident, especially when p is large. Through the idea suggested in this section, we tailor the adjustment of $\tilde{\boldsymbol{U}}$ to fit in the context of ARTS and avoid using a test statistic in matrix form that is inevitable if fitting a multivariate AFT model to adjust for $\tilde{\boldsymbol{U}}$. This idea is crucial in the sense that it has the advantage of extending theoretical results developed for ARTS to *adjusted ARTS*.

Chapter 3

Forward stepwise ARTS procedure

Given one significant predictor detected by ARTS, it is natural to continue searching for other potential predictors, conditional on the information provided by the found predictor. We implement the idea used in *adjusted ARTS* to fulfill this task in a forward and stepwise direction. Such a conditional screening will be continued until no more significance can be detected. We refer to this screening procedure as *forward stepwise ARTS* and carry it out in steps below.

(1) Given the predictor $U_{\hat{j}_n}$ detected by ARTS, obtain residuals from regressing U_j on $U_{\hat{j}_n}$ whenever $j \neq \hat{j}_n$. Treat the residuals as screened predictors and run *adjusted ARTS*. If no significant results returned, stop this procedure; otherwise, collect the newly-found significant predictor $U_{\tilde{j}_n}$.

(2) Use residuals from regressing U_j on $(U_{\hat{j}_n}, U_{\tilde{j}_n})$ as updated predictors, for all $j \notin (\hat{j}_n, \tilde{j}_n)$. Implement *adjusted ARTS* based on these updated predictors, with the aim to detect the next significant predictor.

(3) Keep this procedure proceeding forth and accumulate predictors until no more significant predictor can be detected.

Our *forward stepwise ARTS* procedure successively updates the predictors by using residuals from regressing on previously identified predictors. Compared with the

residual analysis suggested by MQ, our forward stepwise procedure allows the regression coefficients of all already-included predictors to be refit at each step. This implies that the detection of further significant predictors would be conducted, adjusting for those already-included predictors. Therefore, this procedure may have less susceptibility to inappropriate variable selections caused by high correlation between active and inactive predictors, since the effect of active and already-included predictors has been essentially removed from those unselected predictors.

Chapter 4

Competing methods

We compare the performance of ARTS with several procedures that are widely applied to detect the presence of significant predictors for the survival outcome. When considering the adjustment of baseline covariates, these procedures can be modified as alternatives to *adjusted ARTS.*

1 AFT model approaches

Marginal parametric AFT models with Bonferroni correction (BONF-AFT). Apart from the semiparametric AFT model adopted in ARTS, a marginal parametric AFT model is often used to predict T from each predictor. Through specifying a parametric form of the error distribution, we can derive the corresponding likelihood function, and obtain the maximum likelihood estimate of the marginal regression coefficient of U_j. A Z-test with Bonferroni correction is carried out for testing whether each marginal regression coefficient is zero or not. This method can be implemented in the `survreg` function from the `survival` package of R. To adjust for baseline covariates, we can treat the residual $\hat{D}_j$ as the predictor in a marginal parametric AFT model, $j = 1, \ldots, p$. In our finite sample simulations, we specify that the error term follows the standard normal distribution.

Centered percentile bootstrap with AFT model (CPB-AFT). In contrast with ARTS, this procedure works on the premise that there is at least one active predictor, and only bootstraps the first part of (1.15) to estimate the upper and lower $100(\alpha/2)$-th percentiles of the limiting distribution of $\sqrt{n}(\hat{\theta}_n - \theta_n)$. The estimated percentiles can be used to provide critical values for the test statistic $\sqrt{n}\hat{\theta}_n$ (Efron and Tibshirani (1993)). Note that this method gives a special case of ARTS with $\lambda_n = 0$. We are able to easily modify this method to adjust for baseline covariates via replacing θ_n and $\hat{\theta}_n$ by their counterparts in the framework of *adjusted ARTS*.

Calibration by simulation from the estimated null distribution (CEND). The asymptotic acceptance region is used to calibrate the test, and can be constructed in a special case that ε and $\boldsymbol{U}$ are independent. The idea is to simulate the limiting distribution of the scaled test statistic $\sqrt{n}\hat{\theta}_n/s$ under the null, where $s^2 = \mathbb{P}_n(Y_i^{(n)} - \hat{a}_n - \hat{\theta}_n U_{\hat{j}_n})^2$. At the null, we can show that $\sqrt{n}\hat{\theta}_n/s \xrightarrow{d} \tilde{M}_J/V_J$, where $\{\tilde{M}_j, j = 1, \ldots, p\} \sim \mathcal{N}_p(\mathbf{0}, \Sigma_{\boldsymbol{U}})$; $\Sigma_{\boldsymbol{U}}$ is the covariance matrix of $\boldsymbol{U}$, and $J = \arg\max_j \tilde{M}_j^2/V_j$ (the proof placed in the end of the supplementary materials). With $\Sigma_{\boldsymbol{U}}$ estimated by the sample covariance matrix of $\boldsymbol{U}$, we generate 1000 realizations from $\mathcal{N}_p(\mathbf{0}, \Sigma_{\boldsymbol{U}})$; use them to obtain 1000 random copies of $\sqrt{n}\hat{\theta}_n$, and take the corresponding percentiles to develop the acceptance region. Reject the null hypothesis if $\sqrt{n}\hat{\theta}_n$ falls beyond the region. The version to adjust for baseline covariates can be analogically developed by taking $\hat{D}$ as predictors.

2 Cox model approaches

The other popular approach for linking predictors to the survival outcome is Cox model, and the related statistical inference can be developed on the basis of partial likelihood (Cox (1972), Cox (1975)).

Partial likelihood ratio test (PLRT). This test is developed by the likelihood ratio test statistic Λ, the ratio of the partial likelihood from the full Cox model versus that from the reduced model at the null. Provided that $\Lambda \xrightarrow{d} \chi^2_p$ (chi-square distribution with p degrees of freedom), comparing Λ with a χ^2_p-distributed random variable gives the p-value to calibrates the test. However, PLRT is only feasible in the case of $n > p$, because it involves in a full linear model containing all of the predictors. To adjust for baseline covariates, we define the test statistic by the ratio of the partial likelihood from a Cox model containing $(\boldsymbol{U}, \tilde{\boldsymbol{U}})$ versus that from a Cox model only considering $\tilde{\boldsymbol{U}}$, and the statistic weakly converges to χ^2_p.

Marginal Cox models with Bonferroni correction (BONF-COX). This procedure is an analogy to BONF-AFT, but based on marginal Cox models for linking the survival outcome to each predictor U_j, $j = 1, \ldots, p$. Provided the asymptotic normality of the maximum partial likelihood estimator (MPLE) (Andersen and Gill (1982)), we conduct a Z-test with Bonferroni correction to investigate whether each marginal regression coefficient is zero or not. To adjust for baseline covariates, we can instead fit Cox models containing $(U_j, \tilde{\boldsymbol{U}})$ for all j and use the corresponding MPLE of the regression coefficient of U_j as the test statistic.

Centered percentile bootstrap with Cox model (CPB-COX). This procedure is similar to CPB-AFT in general, but the selected predictor is determined in a different fashion. The marginal p-values would be obtained from Z-tests based on separate marginal Cox models, and we select the predictor that marginally introduces the minimal p-value among others. We apply centered percentile bootstrap on the MPLE of the regression coefficient of this selected predictor (that is, the most significant predictor). To consider additional baseline covariates, we instead consider Cox models containing $(U_j, \tilde{\boldsymbol{U}})$ for all j, and bootstrap the MPLE of the regression coefficient of the most significant predictor among U_j's while adjusting for $\tilde{\boldsymbol{U}}$.

Global test based on Cox model (GLOBAL). A score test is proposed to investigate whether predictors $\boldsymbol{U}$ contribute to the hazard rate (Goeman et al. (2005)). The components of $\boldsymbol{\beta}_0$ are assumed random and independently follow a prior distribution with mean zero and common variance v, and it suffices to test whether $v = 0$ for investigating whether $\boldsymbol{\beta}_0 = \mathbf{0}$. Let $\mathbf{r} = (r_1, \ldots, r_n)^T$ with $r_i = \boldsymbol{U}_i^T \boldsymbol{\beta}_0$ for all i, and note that $\mathbf{r}$ is not observed because the unknown parameter vector $\boldsymbol{\beta}_0$ gets involved. By assumptions on $\boldsymbol{\beta}_0$, $\mathbf{r}$ has mean zero and covariance matrix $v\boldsymbol{U}\boldsymbol{U}^T$. Under non-informative censoring assumption, the marginal likelihood function of v is defined by

$$L(v) = E_{\mathbf{r}}\left[\exp\left(\sum_{i=1}^{n}[\delta_i(\ln(h_0(X_i)) + r_i) - \exp(r_i)H_0(X_i)]\right)\right], \tag{4.1}$$

where $H_0(t) = \int_0^t h_0(s)ds$ is the cumulative baseline hazards function up to time t. Second-order Taylor expanding the exponential term in (4.1) with respect to $\mathbf{r}$, we

can express $L(v)$ by the first and second moment of $\mathbf{r}$ (Le Cessie and van Houwelingen (1995)). This shows the sufficiency to establish the desired test statistic only based on the first and second moment of $\boldsymbol{\beta}_0$, without assuming a specific prior distribution. The test statistic can be constructed by the score function of v. There are two ways to calculate the p-value: by asymptotic theory and by permutation arguments. Both of them will be compared with ARTS in our numerical studies. This global test can be modified to adjust for baseline covariates, by simultaneously including $\boldsymbol{U}$ and $\tilde{\boldsymbol{U}}$ in the Cox model, and the test statistic will be constructed conditional on the MPLE of the regression coefficient of $\tilde{\boldsymbol{U}}$.

Chapter 5

Numerical studies

1 Finite sample simulations

The performance of ARTS is evaluated by numerical studies under different data generating scenarios. The underlying survival outcome can follow either an AFT model or a proportional hazards model. For the former, we consider three data generating models:

Model 1 $T = \varepsilon$;

Model 2 $T = U_1/4 + \varepsilon$;

Model 3 $T = \sum_{j=1}^{p} \beta_j U_j + \varepsilon$, where $\beta_1 = \ldots = \beta_5 = 0.15$, $\beta_6 = \ldots = \beta_{10} = -0.1$,

and $\beta_j = 0$ for $j \geq 11$,

where ε denotes the noise that follows a standard normal distribution and is independent of $\boldsymbol{U}$. In Model 1 there is no active predictor, while there is only a single active predictor in Model 2. In Model 3 we have ten active predictors and the most correlated predictor is not unique. The censoring time C is exponentially distributed with various rate parameters for light censoring (10% of subjects with censored survival outcomes), for moderate censoring (20%), and for heavy censoring (40%). The vector of predictors $\boldsymbol{U}$ follows a p-dimensional normal distribution with each compo-

nent $U_j \sim \mathcal{N}(0,1)$, and an exchangeable correlation structure $\text{Corr}(U_j, U_k) = 0.5$ for $j \neq k$.

We also generate the survival outcome based on the following proportional hazards models (Bender et al. (2005)):

Model 4 $h(t|\boldsymbol{U}) = 2\exp(t)$;

Model 5 $h(t|\boldsymbol{U}) = 2\exp(t)\exp(U_1/4)$;

Model 6 $h(t|\boldsymbol{U}) = 2\exp(t)\exp(\sum_{j=1}^{p}\beta_j U_j)$, where the value of β_j's as stated in **Model 3**.

To achieve designed censoring rates, we generate censoring times by an exponential random variable with different rate parameters. We use Model 1 and 4 to present null models, Model 2 and 5 to present alternative models with a sparse signal, and Model 3 and 6 to present alternative models with weak dense signals.

For each data generating scenario, we consider two sample sizes (n = 100 and 200), and five values for the dimension of predictors (p = 10, 50, 100, 150 and 200). A nominal significance level of 5% is used throughout. The number of bootstrap replications is set as 1000. To determine the threshold λ_n in ARTS, we apply ARTS to generate 1000 nested bootstrap replications of each bootstrap estimate $\hat{\theta}_n^*$, and construct the acceptance region based on the corresponding percentiles of these replications. If the test statistic $\sqrt{n}(\hat{\theta}_n^* - \hat{\theta}_n)$ falls beyond the acceptance region, we say that ARTS rejects the null hypothesis. The value of threshold λ_n will be selected to attain 5% rejection rate of ARTS among 1000 bootstrap estimates.

To provide a full comparison, we compare the performance of ARTS with several competing methods under these scenarios. These methods are commonly adopted

to detect significant predictors of the survival outcome, and have been introduced in Section 4. Empirical rejection rates based on 1000 Monte Carlo replications under various censoring rates are displayed in Figures 5.1-5.2. The panels for Model 1 and 4 give type I error rates, which we compare with the nominal level of 5%. Those panels for Model 2-6 indicate the power of each test.

In Figure 5.1, ARTS controls type I error rates (or equivalently, FWERs) around the nominal level, and demonstrates relatively high power throughout alternative models. The Bonf-AFT method gives more conservative type I error rates and lower power than ARTS, with the exception of achieving similar power to ARTS under alternative models with heavy censoring and $n = 200$. The Bonf-COX method and the global test based on asymptotic theory (Global-asymp) are highly conservative and lead to low power. Both CPB-AFT and CPB-COX are anti-conservative, with empirical type I error rates exceeding the nominal level at least by 5% under different sample sizes and various censoring rates (and thus going out of range at some points of dimension, see the left panels in Figure 5.1). The global test based on permutation arguments (Global-permut) takes a good control of type I error rates but claims a much lower power than ARTS, especially under light or moderate censoring. Both CEND and PLRT have poor performance: the former brings large type I error rates but low power, while the latter introduces extremely high type I error rates (PLRT results not shown here). The unsatisfying performance of CEND may result from small sample sizes used in simulations, in view of that CEND is developed based on the simplified form of the limiting distribution. The power of each approach can be observed higher as the sample size increases and the censoring rate decreases. The

comparison between results for Model 2 and 3 shows no adverse impact on the power of ARTS when the maximally correlated predictor is non-unique.

In Figure 5.2 where data are not generated from AFT models, ARTS retains a good control of type I error rate. On the other hand, ARTS suffers from an unstable performance of power when $n = 100$ or heavy censoring. Under light or moderate censoring, the power of ARTS under Model 5 and 6 deteriorates sharply when $n = 100$ and p increases, while ARTS retains stable power when $n = 200$. With a misspecified error distribution, Bonf-AFT surprisingly controls type I error rates well but leads to much worse power. In contrast, Bonf-COX gives relatively higher power when the underlying survival outcome is generated from the proportional hazards model, although it is still conservative at the null. Other competing methods present similar results as in Figure 5.1. Even though unstable in power due to model misspecification, ARTS still maintains a balanced performance between controlling type I error and achieving power than other methods, especially in the cases of light/moderate censoring and larger sample size. Comparing Figure 5.1 with Figure 5.2, we also observe that ARTS is less susceptible to model misspecification than competing methods. In the scenarios of AFT data generating models, ARTS apparently dominates Cox model approaches throughout; in the scenarios where data are generated from proportional hazards models, ARTS still exhibits better performance in FWER or power than Cox-model-relevant approaches when light/moderate censoring and $n = 200$.

2 Asymptotic power evaluation

In this section, we conduct a simulation study to evaluate the asymptotic FWER and power of ARTS, compared with that of Bonf-AFT. We will access the asymptotic FWER and power based on the limiting distribution shown in Theorem 1. This approach can be a computationally efficient alternative to the simulation method used in our finite-sample studies, because it evades the required double bootstrap (for threshold selection) that incurs a heavy computational expense of implementing ARTS.

Due to the complicated limiting distribution shown in Theorem 1, this approach is only feasible when $\varphi_j(\mathbb{L})$ can be reasonably negligible for all j. One possible situation is when $\boldsymbol{\beta}_0 = \mathbf{0}$ and the error term ε is independent of $\boldsymbol{U}$. This restriction on ε facilitates the evaluation of the asymptotic FWER at the null ($\boldsymbol{\beta}_0 = \mathbf{0}$, $\boldsymbol{b}_0 = \mathbf{0}$), and that of the asymptotic power at the local alternatives ($\boldsymbol{\beta}_0 = \mathbf{0}$, $\boldsymbol{b}_0 \neq \mathbf{0}$), saving computational costs at the price of being sensitive to model misspecification.

Consider a local model

$$T^{(n)} = (n^{-1/2} b_0) U_1 + \varepsilon, \tag{5.1}$$

where U_1 is the first element of $\boldsymbol{U}$. The predictors $\boldsymbol{U}$, the error term ε and the censoring time C are generated as in Section 5. We allow b_0 to vary over a grid in $[0, 5]$ by increments of 0.5. This local model reduces the complex limiting distribution

into a simpler form:

$$\sqrt{n}(\hat{\theta}_n - \theta_n) \xrightarrow{d} (M_J + b_0 \mathrm{Cov}(U_J, U_1))/\mathrm{Var}(U_J) - b_0, \tag{5.2}$$

where $J = \arg\max_j \{M_j + b_0 \mathrm{Cov}(U_j, U_1)\}^2/\mathrm{Var}(U_j)$, and $\boldsymbol{M} = \{M_j, j = 1, \ldots, p\}$ is a mean-zero normal random vector with the covariance matrix given by that of the random vector $\{\tilde{\varepsilon}(U_j - EU_j), j = 1, \ldots, p\}$. This evaluation procedure can be carried out as follows.

(1) For each value of b_0 on the grid, generate a large sample (with $n = 10,000$) from the local model (5.1) and compute the corresponding $Y^{(n)}$. With a fixed threshold λ_n, use ARTS to develop the acceptance region $[c_l, c_u]$ based on this sample.

(2) For each given b_0, take $10,000$ draws from the limiting distribution in (5.2), and then we can obtain $10,000$ realizations of $\sqrt{n}\hat{\theta}_n$.

(3) The asymptotic rejection rate of ARTS (for the given b_0) can be accessed by computing the proportion of falling beyond $[c_l, c_u]$ among $10,000$ realizations of $\sqrt{n}\hat{\theta}_n$.

To evaluate the asymptotic power of ARTS, we independently implement the above procedure 20 times and display these corresponding asymptotic rejection rates in a box plot, for each considered b_0. Note that the variation within each boxplot should result from randomness over 20 independent samples. For comparison, we also plot the asymptotic power of Bonf-AFT, which is approximated by the rejection rate from

1000 samples each of size $n = 10,000$.

To make the above evaluation practical for large p, say $p = 1000$, the threshold λ_n is fixed at $0, 4.3, 6.1, 7.4$ as the constant a takes corresponding values of $0, 2, 4, 6$. We present results under light censoring (Figure 5.3), moderate censoring (Figure 5.4) and heavy censoring (Figure 5.5). Since the plots appear similar between $a = 0$ and $a = 1$ and have no obvious difference when $a \geq 6$, we only present results at $a = 0, 2, 4, 6$ for conciseness. From these figures, we observe that smaller the value of a is, ARTS behaves more like CPB-AFT and gives more anti-conservative results as observed in previous numerical studies. When $a = 0$, in particular, ARTS reduces to CPB-AFT. We also perceive that the variation within each boxplot decreases as a grows larger.

Comparing the asymptotic power of Bonf-AFT (denoted by the circle) with the median of each boxplot, it indicates that ARTS has more satisfactory performance than Bonf-AFT in most cases. In terms of median power, ARTS can even provide an extra 20% power in some situations (e.g., at $b_0 = 3$ when $a = 4$ or $a = 6$ for all types of censoring). To control the asymptotic FWER, the reasonable choice should fall on $a = 4$ under light or moderate censoring, since the median FWER starts to touch the nominal level and the corresponding variation within the boxplot apparently diminishes. On the other hand, the selection of a should fall between 2 and 4 under heavy censoring because the median FWER remains higher than 5% at $a = 2$ but drops below 5% at $a = 4$.

3 Error dependent on predictors

In this section, we present the control on FWER of ARTS when the error term ε is still uncorrelated with but dependent on predictors $\boldsymbol{U}$. For simplicity, $\boldsymbol{U}$ follows a p-dimensional normal distribution with mean zero and identity covariance matrix, implying that predictors are independent of each other. To give a full comparison, the FWERs of AFT-model-relevant competing methods are also provided, especially for those require the independence between ε and $\boldsymbol{U}$ such as CEND.

To produce a dependent error structure on predictors, we generate the error term ε by random replications from a normal distribution with mean zero and standard deviation of $0.7(|U_1| + 0.7)$, and simulate the transformed time-to-event outcome under the null model $T = \varepsilon$. Note that ε is also non-identically distributed among subjects. Though not independent, we can see that ε still remains uncorrelated with $\boldsymbol{U}$ by $\text{Cov}(\varepsilon, U_1) = E[\varepsilon U_1] = E\{U_1 E[\varepsilon|U_1]\} = 0$ and $\text{Cov}(\varepsilon, U_j) = E\{U_j E[\varepsilon|U_1]\} = 0$ for $j \neq 1$. The censoring time C still follows the exponential distribution with varying rate parameters specified for different censoring rates. Figure 5.6 shows that only ART controls FWER around the nominal level in the case of dependent and non-identically distributed errors, except for providing slightly conservative FWERs when $p \geq 50$, heavy censoring and $n = 100$.

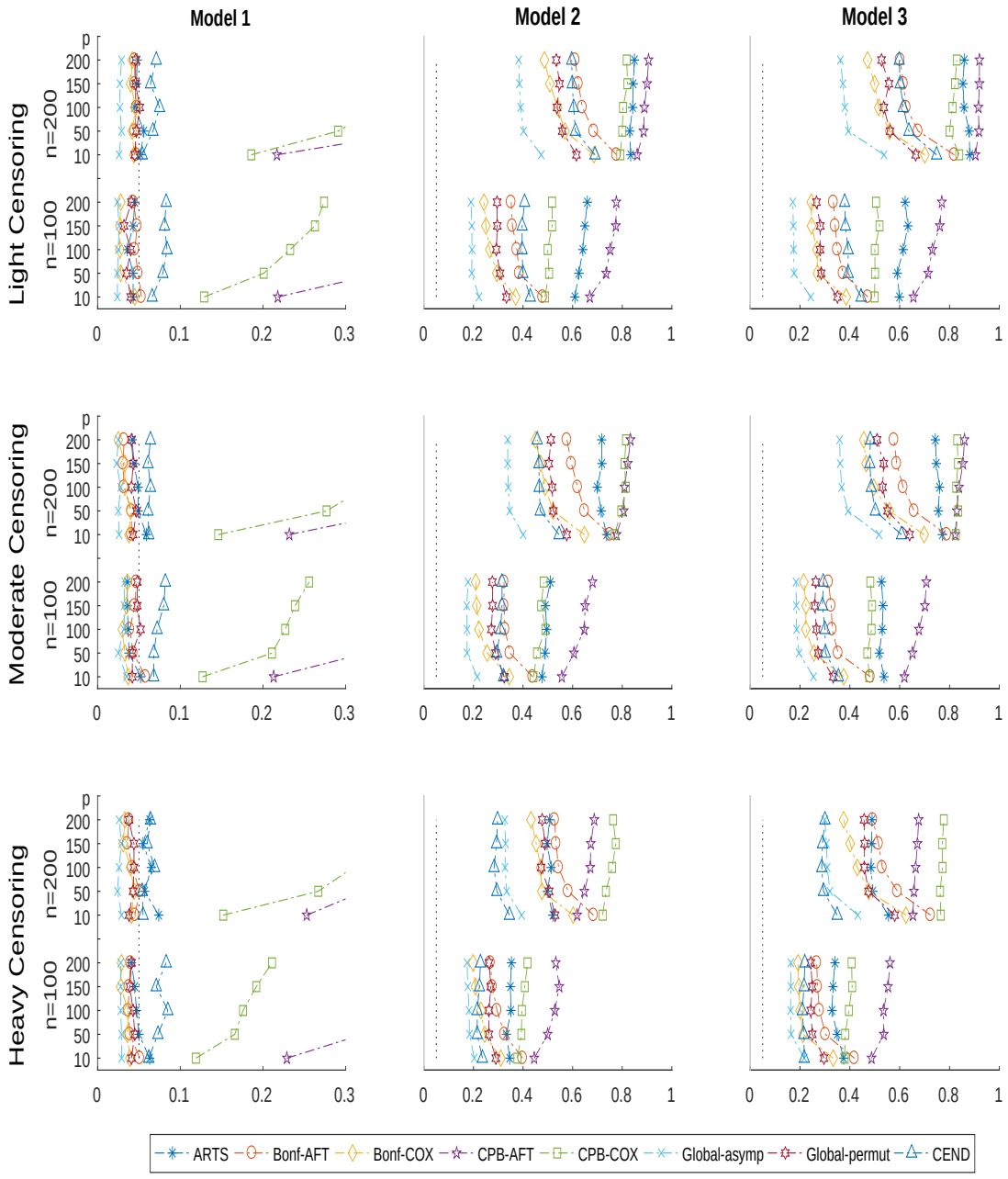

Figure 5.1: Empirical rejection rates based on 1000 samples generated from Model 1-3 with the dimension ranging from $p = 10$ to $p = 200$.

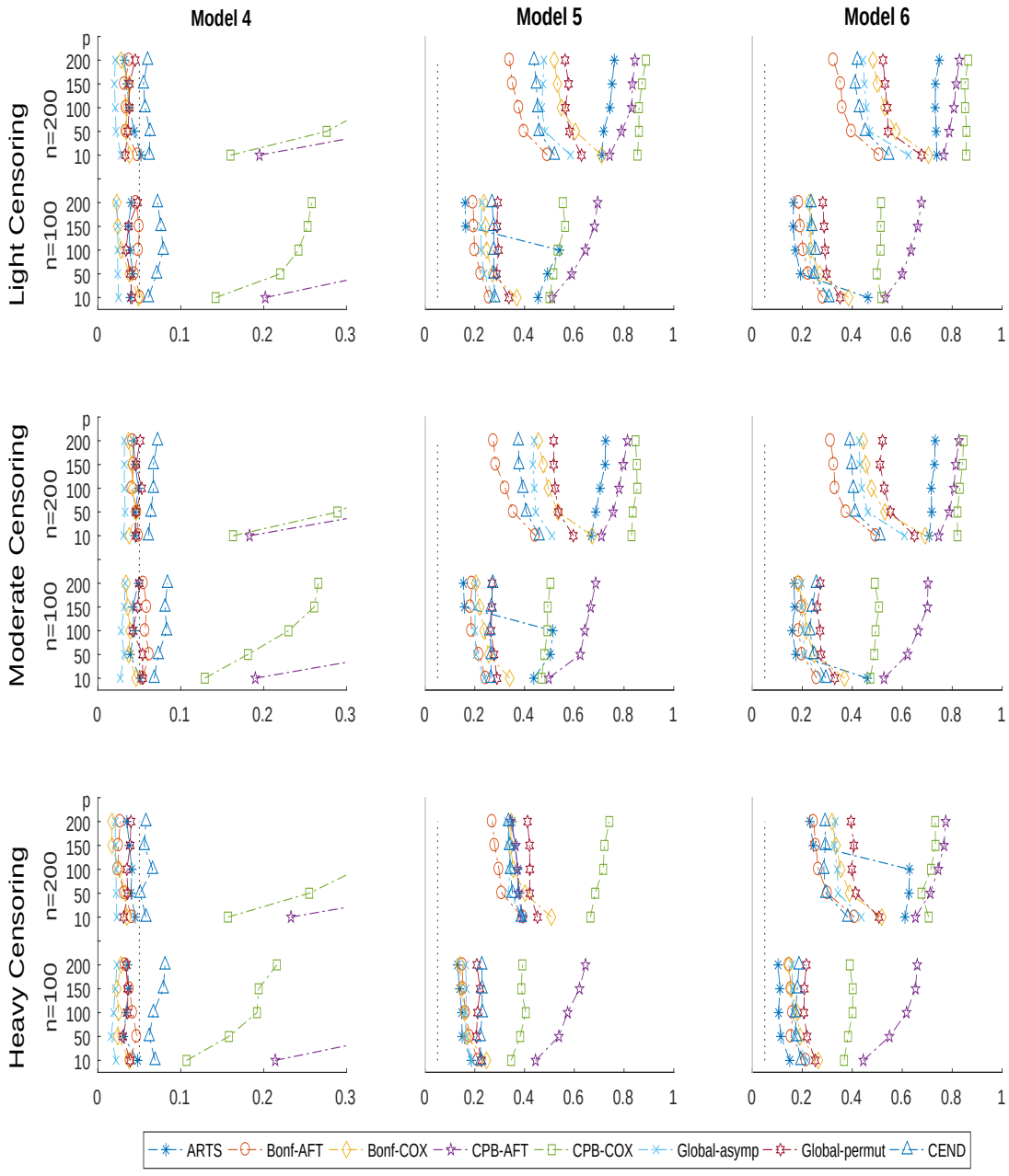

Figure 5.2: Empirical rejection rates based on 1000 samples generated from Model 4-6 with the dimension ranging from $p = 10$ to $p = 200$.

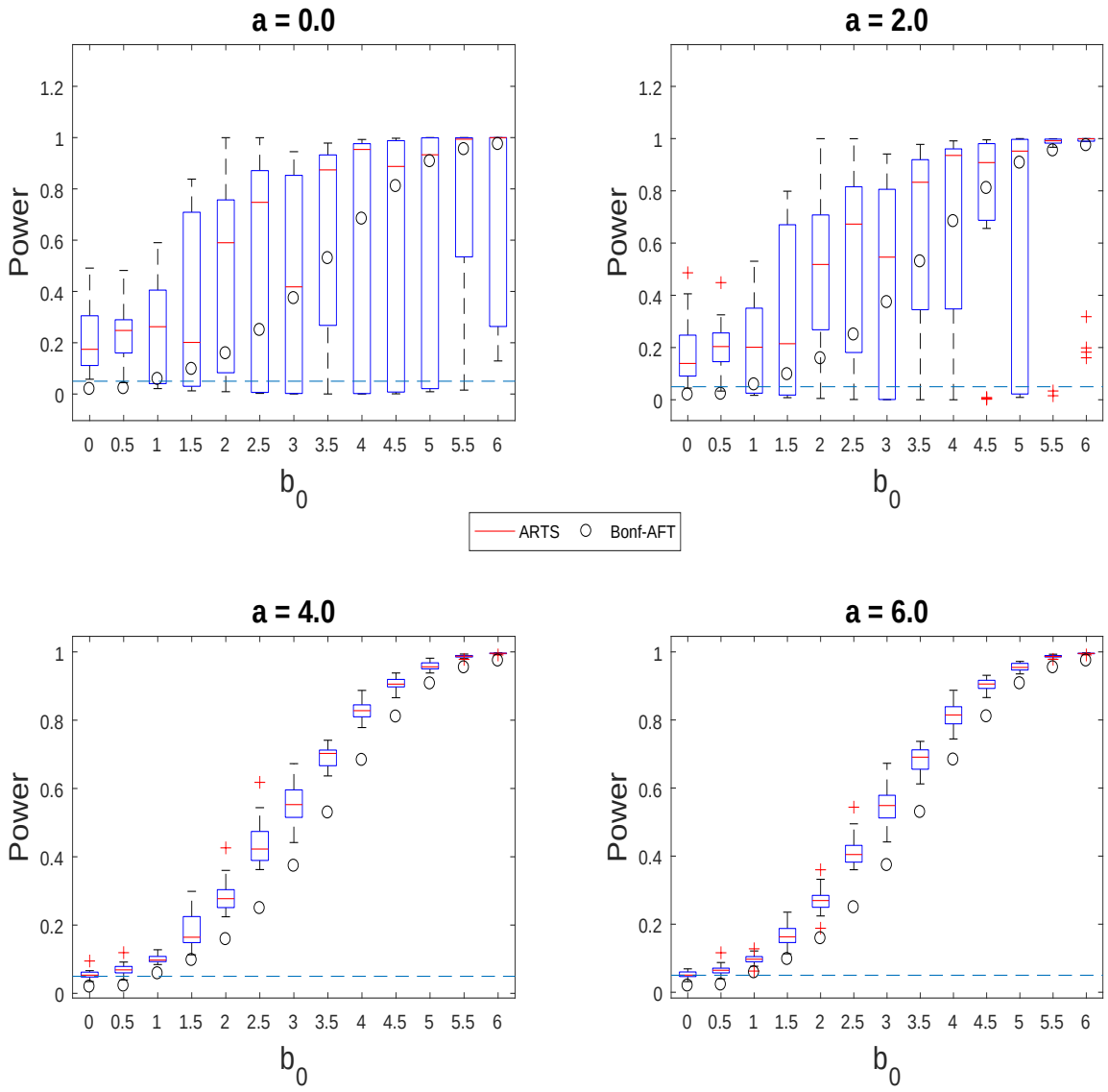

Figure 5.3: Asymptotic Type I error and power of ARTS compared with Bonferroni-AFT for $p = 1000$ under light censoring, where ARTS is implemented at fixed threshold λ_n specified by $a = \{0, 2, 4, 6\}$, and each boxplot is based on 20 independent replications with $n = 10,000$.

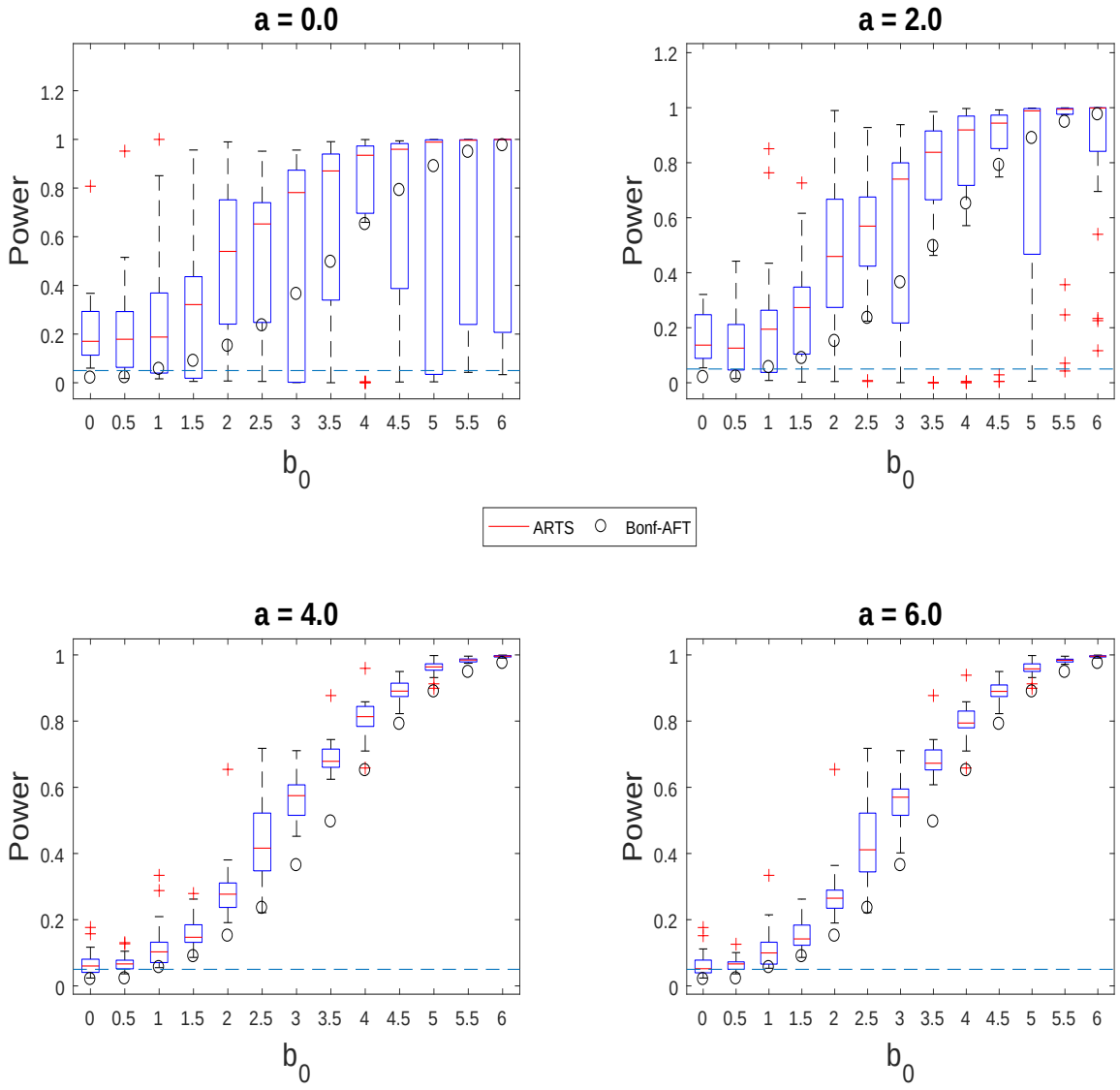

Figure 5.4: Asymptotic Type I error and power as in Figure 5.3 except under moderate censoring.

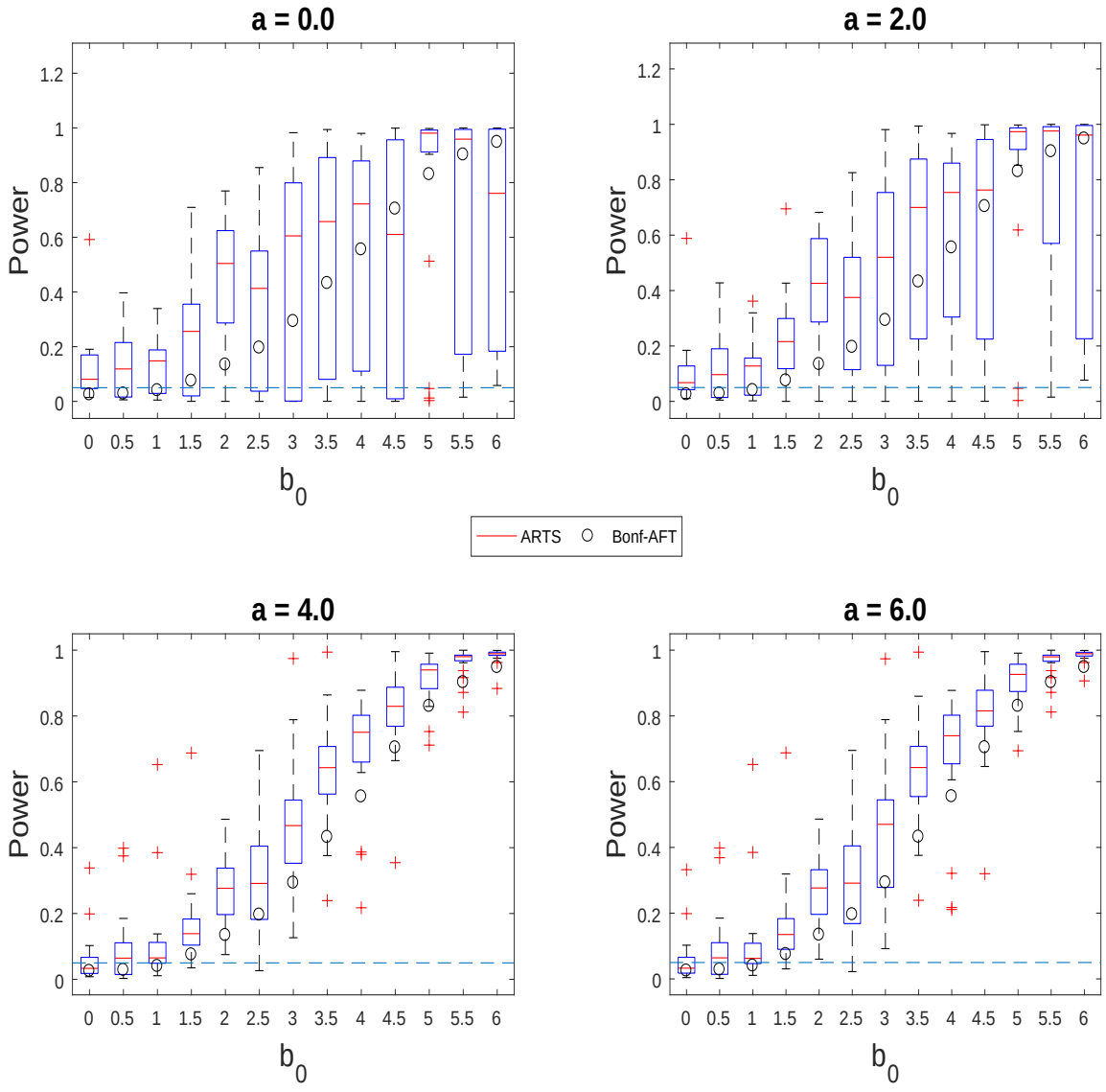

Figure 5.5: Asymptotic Type I error and power as in Figure 5.3 except under heavy censoring.

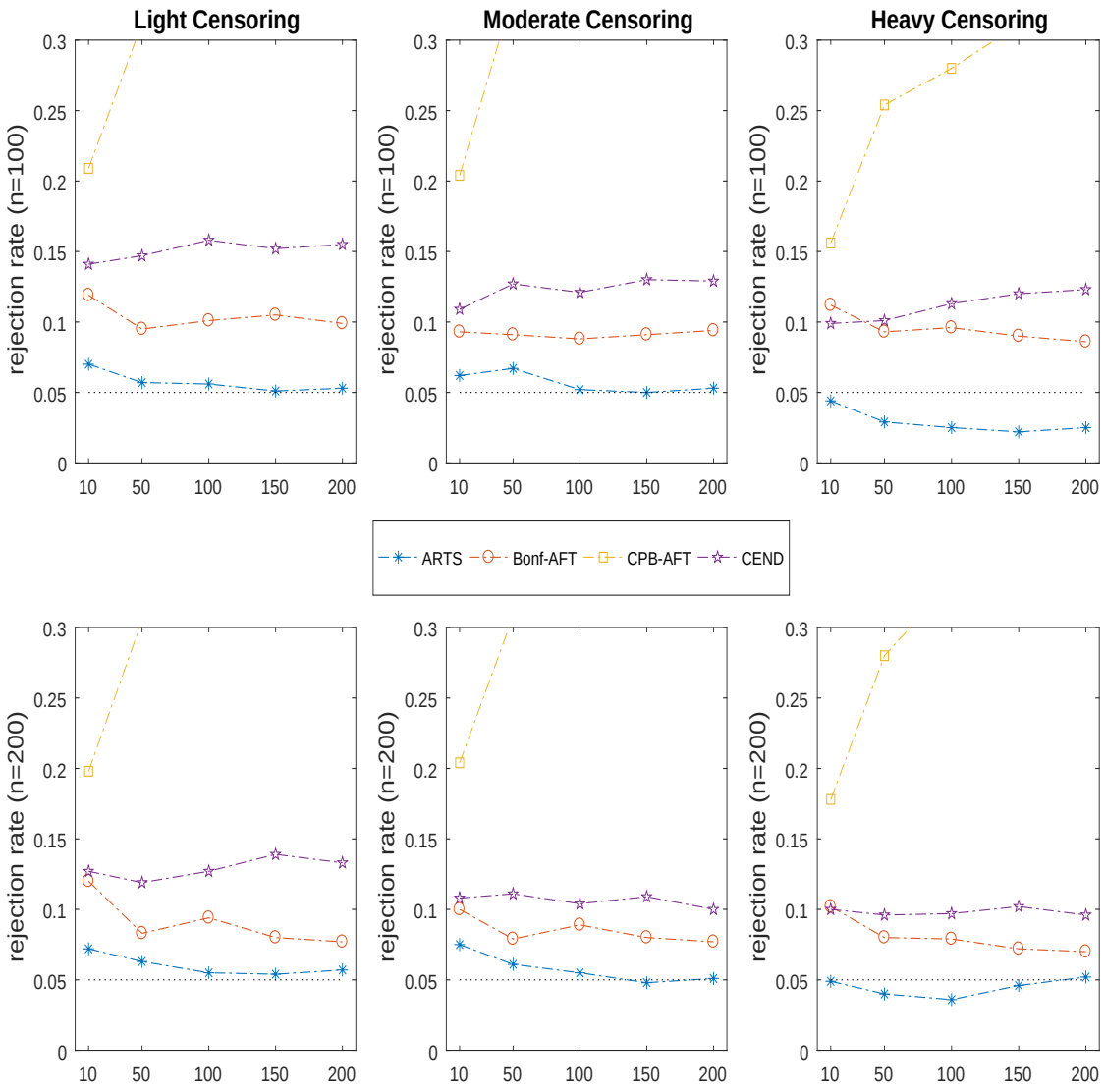

Figure 5.6: Empirical rejection rates of ARTS and CEND based on 1000 samples generated from the null model with dependent errors under various p and censoring rates.

Chapter 6

Applications to real data

1 DLBCL data

We revisit the DLBCL data introduced earlier (Rosenwald et al. (2002)). This data set contains the after-chemotherapy survival time from DLBCL diseases, the categorical IPI variable (with three levels low, medium and high) and 7399 genetic features of 222 patients ($n = 222$ and $p = 7399$), subject to a censoring rate of 43%. More details about the DLBCL data can be found in the literature (cf. Binder et al. (2011) and Bøvelstad et al. (2009)). To adjust for the prognostic information provided by IPI, we apply *adjusted ARTS* on this data set for detecting the presence of significant genetic features. To maintain the stability of the KSV estimator, the observed event time is restricted up to $\tau = 2.82$, which corresponds to the 98% empirical percentile of the observed event time. This excludes an observation whose estimated synthetic response value is 55.867 and severely distorts the estimation of marginal regression coefficients. In ARTS, We use the double bootstrap to select the constant a from 0 to 15 by increments of 0.5.

To give a fair comparison with ARTS, we also apply AFT-model-relevant competing methods: Bonf-AFT and CPB-AFT, with IPI information adjusted. The CEND method is not included, because it is challenging to verify its required assumption that

the error term is independent of predictors. The three implemented approaches yield very different p-values. The minimal Bonferroni corrected p-value from Bonf-AFT is 37.93%, indicating that no feature is significantly correlated with the survival time of patients suffering DLBCL. The CPB-AFT method gives a p-value of 4.0% and ARTS produces a p-value of 23.60%, from 1000 bootstrap samples. Figure 6.1 exhibits the sampling distribution of the test statistics used by ARTS and CPB-AFT based on these bootstrap samples, and it also illustrates how the corresponding p-values are obtained. In summary, ARTS detects no significant feature given the nominal level of 5.0%, while CPB-AFT gives a marginally significant result and Bonf-AFT behaves as conservative as expected.

2 Primary biliary cirrhosis data

Apart from massive gene expression data, ARTS could be useful in investigating interaction effects of clinical covariates, given that some covariates have been shown statistically or clinically significant. We demonstrate how ARTS works toward the aim of screening out potential interaction effects based on data from the Mayo Clinic trial in primary biliary cirrhosis (PBC) of the liver conducted between Year 1974 and Year 1984 (Fleming and Harrington (1991), Appendix D.1). A total of 424 PBC patients were referred to Mayo Clinic during that ten-year interval and met eligibility criteria for the randomized placebo controlled trial of the drug D-penicillamine. Survival times between registration and death (possibly right censored) are available for 312 patients; we only consider the 276 patients whose covariate information is complete and available at registration. Only five of the 17 risk factors were found statistically

significant under the setting of AFT model (Jin et al. (2003)), and they were also identified as active predictors of the survival time under Cox model (Bunea and McKeague (2005)). These significant risk factors are age (in years), presence of edema (0=no; 0.5=resolved; 1=unresolved with therapy), serum bilirubin (in mg/dl), albumin (in gm/dl), and protime (standardized blood clotting time). Of these risk factor, serum bilirubin, albumin and protime are log-transformed. Adjusting for these five risk factors, we apply *forward stepwise ARTS* on the subset of PBC data subject to a censoring rate of 60%, and have interest to successively screen out significant interaction terms of 17 clinical risk factors for the survival time ($n = 276$, $p = 136$).

Figure 6.2 displays the pattern of p-values for the newly entered interaction term at each step. When adjusting for the aforementioned risk factors, *Forward stepwise ARTS* concludes that there is no significant interaction term for the survival time. For comparison, we also present successive p-values given by CPB-AFT, which provides anti-conservative results as usual.

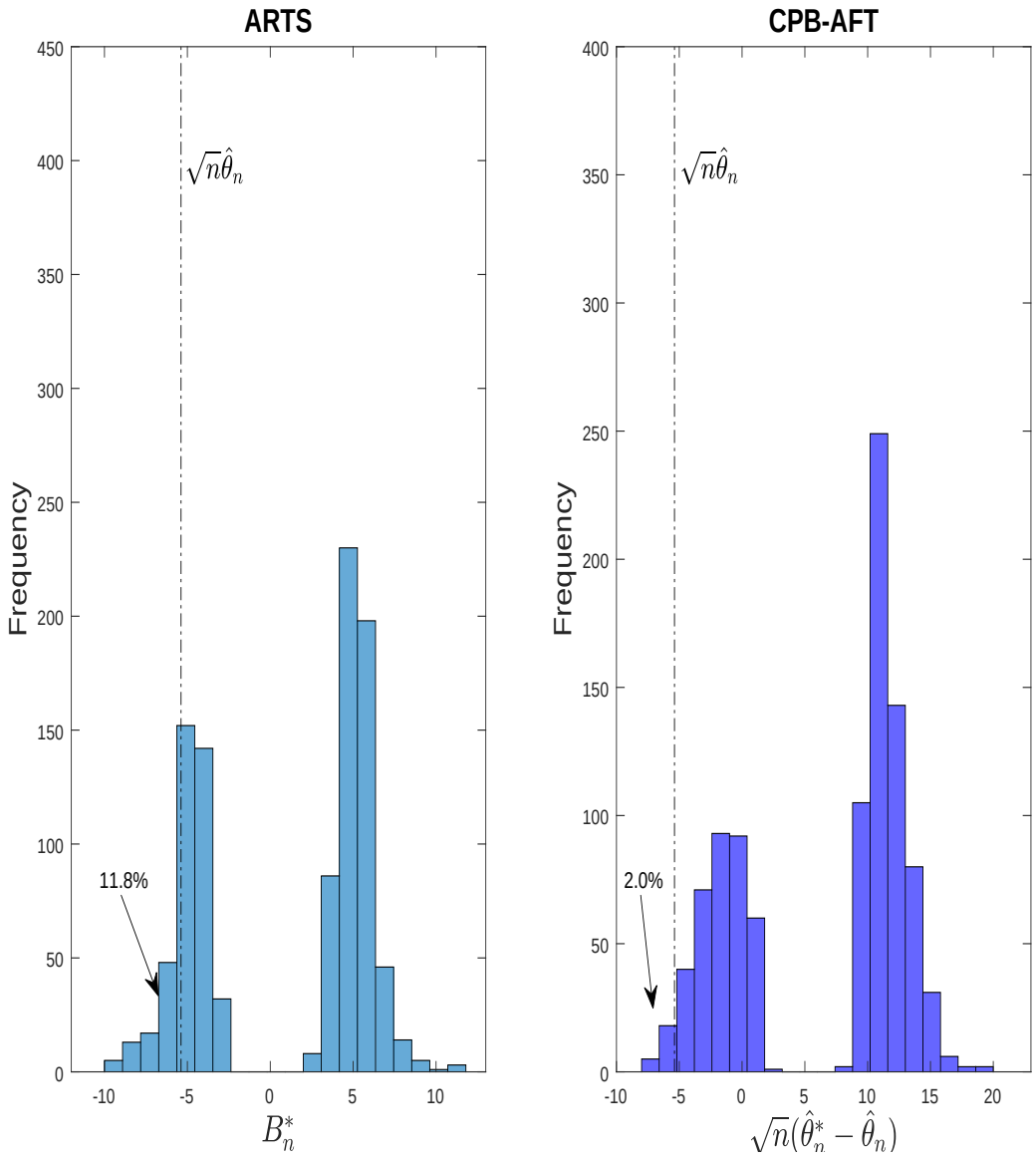

Figure 6.1: DLBCL example. Left panel: histogram of B_n^* giving the two-sided ARTS p-value 23.60%. Right panel: histogram of $\sqrt{n}(\hat{\theta}_n^* - \hat{\theta}_n)$ giving the two-sided CPB-AFT p-value 4.0%.

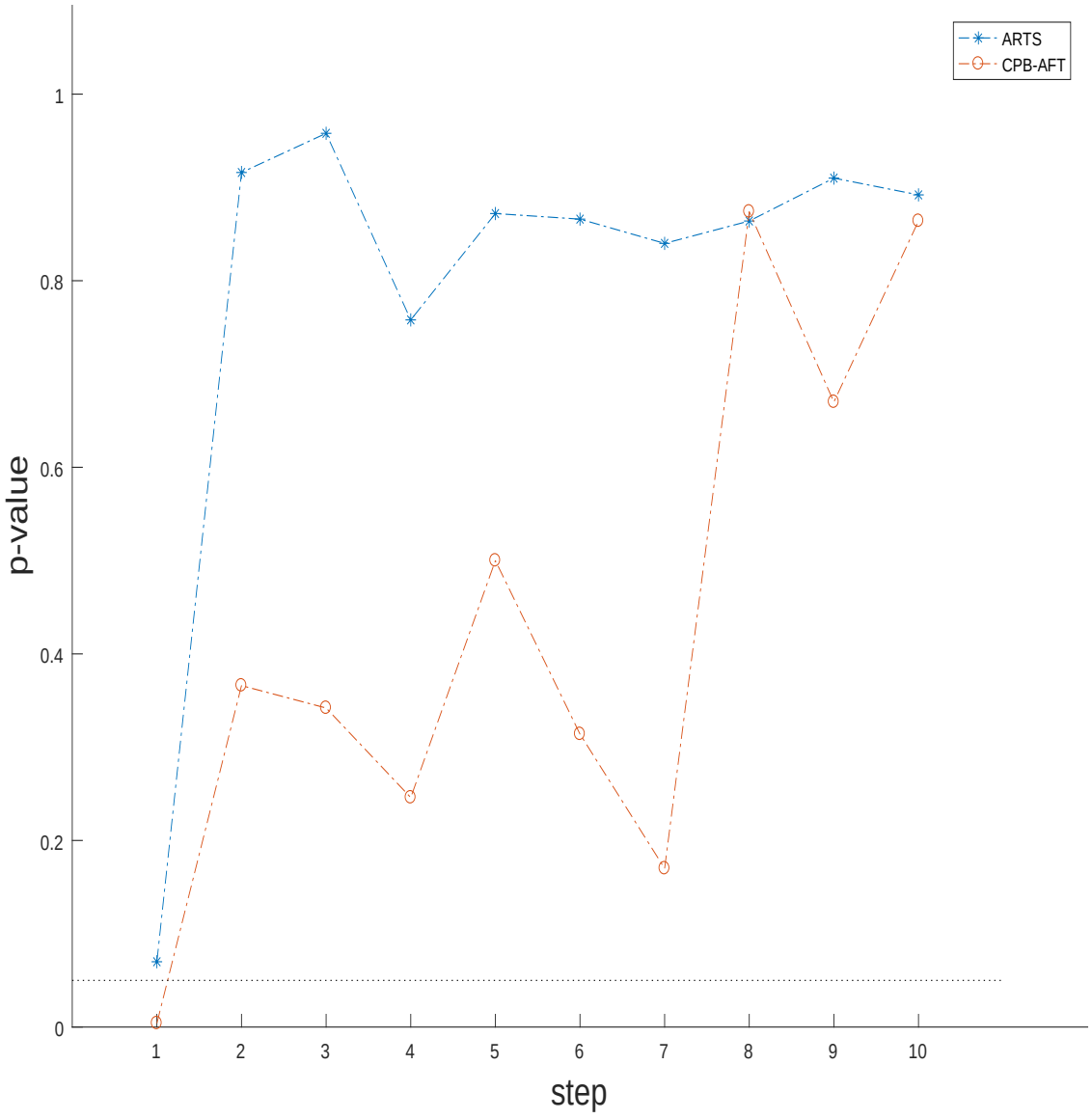

Figure 6.2: PBC example. The pattern of p-values for forward stepwise ARTS and CPB-AFT.

Conclusion

We have developed an adaptive resampling test for survival data (ARTS) to detect the presence of significant predictors for right-censored survival outcomes. We use marginal correlation screening to reduce the high-dimensional detection problem to a single test of whether $\theta_0 = 0$, where θ_0 is the marginal regression coefficient of the most correlated predictor to the survival outcome. In the setting of marginal screening for survival data the post-selection inference has been scarcely considered, and is challenging not only because of the non-regular asymptotic behavior of the test statistic at the null (i.e. $\theta_0 = 0$) but also owing to the presence of censoring. In this framework ARTS is designed to adapt to the non-regularity, while to deal with the discontinuous dispersion introduced by censoring that plays a crucial part asymptotically. The advantage of ARTS lies in it demonstrating a post selection corrected p-value without sacrificing the performance in power, while avoiding distributional assumptions, specific correlation structures between predictors, and preconceived information on regression parameters that are usually used for dimension reduction by some high-dimensional inference methods. The ARTS procedure is also versatile for practical use. Various extensions of ARTS are proposed, to adjust for additional baseline covariates of clinicians' interests and to successively identify all of the active

predictors.

We recognize that ARTS requires the independent censoring assumption that may be violated in some clinical contexts. One direction for future work is to develop rigorous theoretical results of ARTS under the assumption of conditionally independent censoring given predictors. To tackle this type of censoring mechanism, we could use the Cox model or the local Kaplan–Meier estimator for incorporating covariates into the estimation of the conditional survival function of censoring on predictors $G_0(\cdot|\boldsymbol{U})$. The generalization of the censoring mechanism still could be challenging in our framework, even given some proposals for estimating $G_0(\cdot|\boldsymbol{U})$ listed above. One challenge is how to determine the covariates to be included in the estimation of $G_0(\cdot|\boldsymbol{U})$ under the high-dimensional AFT model. The ensuing question is to ask whether the result of post selection inference would be affected as these included covariates may not be completely contained under a series of working AFT models only using one predictor at a time. As far as we know, this question has not been fully answered in the area of marginal screening based on survival data, and is worth further attention.

Although our simulation results show that ARTS performs well when $p \gg n$, we have only provided theoretical support assuming a fixed p. Formal testing procedures that can adjust to the non-regular behavior of $\hat{\theta}_n$ under diverging p appear challenging. One potentially fruitful alternative approach that might handle diverging p would be to extend the efficient influence function technique of Luedtke and van der Laan (2017) to the right censored setting in terms of a regularized version of the KSV estimator.

Bibliography

(1993). Project TINHLPF: A predictive model for aggressive non-Hodgkin's lymphoma. *New England Journal of Medicine*, 329(14):987–994.

Akritas, M. G. (1986). Bootstrapping the Kaplan–Meier estimator. *Journal of American Statistical Association*, 81(396):1032–1038.

Andersen, P. K. and Gill, R. D. (1982). Cox's regression model for counting processes: a large sample study. *The Annals of Statistics*, 10(4):1100–1120.

Antoniadis, A., Fryzlewicz, P., and Letué, F. (2010). The Dantzig selector in Cox's proportional hazards model. *Scandinavian Journal of Statistics*, 37(4):531–552.

Bender, R., Austin, T., and Blettner, M. (2005). Generating survival times to simulate Cox proportional hazards models. *Statistics in Medicine*, 24:1713–1723.

Binder, H., Porzelius, C., and Schumacher, M. (2011). An overview of techniques for linking high-dimensional molecular data to time-to-event endpoints by risk prediction models. *Biometrical Journal*, 53(2):170–189.

Bøvelstad, H. M., Nygård, S., and Borgan, Ø. (2009). Survival prediction from clinico-genomic models–A comparative study. *BMC bioinformatics*, 10:Article 413.

Bradic, J., Fan, J., and Jiang, J. (2011). Regularization for Cox's proportional hazards model with NP-dimensionality. *Annals of Statistics*, 39(6):3092–3120.

Buckley, J. and James, I. (1979). Linear regression with censored data. *Biometrika*, 66(3):429–436.

Bunea, F. and McKeague, I. W. (2005). Covariate selection for semiparametric hazard function regression models. *Journal of Multivariate Analysis*, 92(1):186–204.

Cai, T., Huang, J., and Tian, L. (2009). Regularized estimation for the accelerated failure time model. *Biometrics*, 65(1):394–404.

Cox, D. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society, Series B*, 34(2):187–202.

Cox, D. R. (1975). Partial likelihood. *Biometrika*, 62(2):269–276.

Dabrowska, D. (1989). Uniform consistency of the kernel conditional Kaplan–Meier estimate. *The Annals of Statistics*, 17(3):1157–1167.

Datta, S., Le-Rademacher, J., and Datta, S. (2007). Predicting patient survival from microarray data by accelerated failure time modeling using partial least squares and LASSO. *Biometrics*, 63(1):259–271.

Efron, B. (1981). Censored data and the bootstrap. *Journal of American Statistical Association*, 76(374):312–319.

Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap (Monographs on Statistics & Applied Probability)*. Chapman and Hall/CRC.

Engler, D. and Li, Y. (2009). Survival analysis with high-dimensional covariates: An application in microarray studies. *Statistical Applications in Genetics and Molecular Biology*, 8(1):Article 14.

Fan, J., Feng, Y., and Wu, Y. (2010). High-dimensional variable selection for Cox's proportional hazards model. *Borrowing Strength: Theory Powering Applications – A Festschrift for Lawrence D. Brown. Institute of Mathematical Statistics; Beachwood, OH*, 6:70–86.

Fan, J. and Li, R. (2002). Variable selection for Cox's proportional hazards model and frailty model. *Annals of Statistics*, 30(1):74–99.

Fang, E. X., Ning, Y., and Liu, H. (2016). Testing and confidence intervals for high dimensional proportional hazards models.

Fleming, T. R. and Harrington, D. P. (1991). *Counting Processes and Survival Analysis*. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., New York, NY.

Goeman, J. J., Oosting, J., Cleton-Jansen, A. M., Anninga, J. K., and van Houwelingen, H. C. (2005). Testing association of a pathway with survival using gene expression data. *Bioinformatics*, 21(9):1950–1957.

Gorst-Rasmussen, A. and Scheike, T. (2013). Independent screening for single-index hazard rate models with ultrahigh dimensional features. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 75:217–245.

He, X., Wang, L., and Hong, H. G. (2013). Quantile-adaptive model-free variable screening for high-dimensional heterogeneous data. *Annals of Statistics*, 41(1):342–369.

Hong, H. G., Chen, X., Christiani, D. C., and Li, Y. (2017). Integrated powered density: screening ultrahigh-dimensional covariates with survival outcomes.

Hong, H. G., Kang, J., and Li, Y. (2016). Conditional screening for ultra-high dimensional covariates with survival outcomes.

Huang, J. and Ma, S. (2010). Variable selection in the accelerated failure time model via the bridge method. *Lifetime Data Analysis*, 16(2):176–195.

Huang, J., Ma, S., and Xie, H. (2006). Regularized estimation in the accelerated failure time model with high-dimensional covariates. *Biometrics*, 62(3):813–820.

Jin, Z., Lin, D. Y., Wei, L. J., and Ying, Z. (2003). Rank-based inference for the accelerated failure time model. *Biometrika*, 90(2):341–353.

Johnson, B. A. (2008). Variable selection in semiparametric linear regression with censored data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(2):351–370.

Johnson, B. A., Lin, D. Y., and Zeng, D. (2008). Penalized estimating functions and variable selection in semiparametric regression models. *Journal of the American Statistical Association*, 103(482):672–680.

Kalbfleisch, J. D. and Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data*. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., Hoboken, NJ, USA.

Keiding, N., Andersen, P. K., and Klein, J. P. (1997). The role of frailty models and accelerated failure time models in describing heterogeneity due to omitted covariates. *Statistics in Medicine*, 16(2):215–224.

Kosorok, M. R. (2008). *Introduction to Empirical Processes and Semiparametric Inference*. Springer Series in Statistics. Springer.

Koul, H., Susarla, V., and Van Ryzin, J. (1981). Regression analysis with randomly right-censored data. *The Annals of Statistics*, 9(6):1276–1288.

Lai, T. L. and Ying, Z. (1991a). Large sample theory of a modified Buckley-James estimator for regression analysis with censored data. *The Annals of Statistics*, 19(3):1370–1402.

Lai, T. L. and Ying, Z. (1991b). Rank regression methods for left-truncated and right-censored data. *The Annals of Statistics*, 19(2):531–556.

Le Cessie, S. and van Houwelingen, H. C. (1995). Testing the fit of a regression model via score tests in random effects models. *Biometrics*, 51(2):600–614.

Li, J., Zheng, Q., Peng, L., and Huang, Z. (2016). Survival impact index and ultrahigh-dimensional model-free screening with survival outcomes. *Biometrics*, 72(4):1145–1154.

Li, Y., Dicker, L., and Zhao, S. D. (2014). The dantzig selector for censored linear regression models. *Statistica Sinica*, 24(1):251–268.

Lin, D. Y. and Ying, Z. (1994). Semiparametric analysis of the additive risk model. *Biometrika*, 81(1):61–71.

Lo, A. Y. (1993). A Bayesian bootstrap for censored data. *The Annals of Statistics*, 21(1):100–123.

Lockhart, R., Taylor, J., Tibshirani, R. J., and Tibshirani, R. (2014). A significance test for the lasso. *Annals of Statistics*, 42(2):413–468.

Luedtke, A. R. and van der Laan, M. J. (2017). Parametric-rate inference for one-sided differentiable parameters.

Ma, S. and Du, P. (2012). Variable selection in partly linear regression model with diverging dimensions for right censored data. *Statistica Sinica*, 22(3):1003–1020.

McKeague, I. W. and Qian, M. (2015). An adaptive resampling test for detecting the presence of significant predictors. *Journal of the American Statistical Association*, 110(512):1422–1433.

McKeague, I. W. and Sasieni, P. D. (1994). A partly parametric additive risk model. *Biometrika*, 81(3):501–514.

Medeiros, F. M., da Silva-Júnior, A. H., Valença, D. M., and Ferrari, S. L. (2014). Testing inference in Accelerated Failure Time models. *International Journal of Statistics and Probability*, 3(2):121–131.

Pollard, D. (1990). *Empirical Processes: Thepry and Applications*, volume 2. Institute of Mathematical Statistics.

Ritov, Y. (1990). Estimation in linear regression with censored data. *Annals of Statistics*, 18(1):303–328.

Rosenwald, A., Wright, G., Chan, W. C., et al. (2002). The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma. *The New England journal of medicine*, 346(25):1937–1947.

Sinnott, J. A. and Cai, T. (2016). Inference for survival prediction under the regularized Cox model. *Biostatistics*, page kxw016.

Song, R., Lu, W., Ma, S., and Jessie Jeng, X. (2014). Censored rank independence screening for high-dimensional survival data. *Biometrika*, 101(4):799–814.

Stute, W. (1995). The central limit theorem under random censorship. *The Annals of Statistics*, 23(2):422–439.

Stute, W. and Wang, J.-L. (1993). The strong law under random censorship. *The Annals of Statistics*, 21(3):1591–1607.

Taylor, J. and Tibshirani, R. (2017). Post-selection inference for ℓ_1-penalized likelihood models.

Tibshirani, R. (1997). The LASSO method for variable selection in the Cox model. *Statistics in Medicine*, 16(4):385–395.

Tsiatis, A. A. (1990). Estimating regression parameters using linear rank tests for censored data. *The Annals of Statistics*, 18(1):354–372.

van der Vaart, A. W. (2000). *Asymptotic Statistics*. Cambridge University Press.

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes with Applications to Statistics*. Springer, New York, NY, USA.

Wu, Y. (2012). Elastic net for Cox's proportional hazards model with a solution path algorithm. *Statistica Sinica*, 22:271–294.

Ying, Z. (1993). A large sample study of rank estimation for censored regression data. *The Annals of Statistics*, 21(1):76–99.

Zhang, H. H. and Lu, W. (2007). Adaptive LASSO for Cox's proportional hazards model. *Biometrika*, 94(3):691–703.

Zhao, S. D. and Li, Y. (2012). Principled sure independence screening for Cox models with ultra-high-dimensional covariates. *Journal of Multivariate Analysis*, 105(1):397–411.

Zhao, S. D. and Li, Y. (2014). Score test variable screening. *Biometrics*, 70(4):862–871.

Zhong, P.-S., Hu, T., and Li, J. (2015). Tests for coefficients in high-dimensional additive hazard models. *Scandinavian Journal of Statistics*, 42(3):649–664.

Supplementary Materials

Properties of $\tilde{\varepsilon}$

Given all conditions for Theorem 1, $\tilde{\varepsilon} = \tilde{Y} - \alpha_0 - \boldsymbol{U}^T\boldsymbol{\beta}_0$ has zero mean and finite variance (the square integrability of $\tilde{\varepsilon}$), and is uncorrelated with $\boldsymbol{U}$. The relevant proof goes as follows.

Proof. Because $E[\tilde{Y}|\boldsymbol{U}] = E[T|\boldsymbol{U}]$, it ensures that

$$E[\tilde{\varepsilon}|\boldsymbol{U}] = E[\tilde{Y}|\boldsymbol{U}] - \alpha_0 - \boldsymbol{U}^T\boldsymbol{\beta}_0 = E[T|\boldsymbol{U}] - \alpha_0 - \boldsymbol{U}^T\boldsymbol{\beta}_0 = 0, \tag{S.1}$$

which indicates the zero mean of $\tilde{\varepsilon}$ through taking expectation on both sides of (S.1). Suppose that $G_0(t)$ is bounded away from zero for all $t \in \mathcal{T} = (-\infty, \tau]$, where τ denotes the end of the follow-up. We show the boundedness of $E[\tilde{\varepsilon}^2|\boldsymbol{U}]$ as follows, which implies the finite variance of $\tilde{\varepsilon}$. By simple algebra, we have

$$\begin{aligned} &E[\tilde{\varepsilon}^2|\boldsymbol{U}] = E[(\tilde{Y} - \alpha_0 - \boldsymbol{U}^T\boldsymbol{\beta}_0)^2|\boldsymbol{U}] = E[\tilde{Y}^2|\boldsymbol{U}] - (\alpha_0 + \boldsymbol{U}^T\boldsymbol{\beta}_0)^2 \\ &= E[T^2 G_0(T-)^{-1}|\boldsymbol{U}] - (\alpha_0 + \boldsymbol{U}^T\boldsymbol{\beta}_0)^2 \\ &\leq G_0(\tau-)^{-1}E[\varepsilon^2|\boldsymbol{U}] + (G_0(\tau-)^{-1} - 1)(\alpha_0 + \boldsymbol{U}^T\boldsymbol{\beta}_0)^2. \end{aligned} \tag{S.2}$$

Taking expectation on both sides of (S.2) yields

$$E[\tilde{\varepsilon}^2] \leq G_0(\tau-)^{-1}E[\varepsilon^2] + (G_0(\tau-)^{-1} - 1)E[(\alpha_0 + \boldsymbol{U}^T\boldsymbol{\beta}_0)^2].$$

Because $G_0(\tau-)$ is non-zero and both ε as well as each covariate U_j have finite variance, it is easy to see $E[\tilde{\varepsilon}^2]$ is bounded above by a finite constant. Hence, we show that $\tilde{\varepsilon}$ has finite variance. Suppose that $\mathrm{Cov}(U_j, \varepsilon) = 0$ for all j. We can see $\mathrm{Cov}(U_j, \tilde{\varepsilon}) = 0$ because Proposition 1 gives $\mathrm{Cov}(U_j, \tilde{Y}) = \mathrm{Cov}(U_j, \bar{T})$ and $\mathrm{Cov}(U_j, \tilde{\varepsilon}) = \mathrm{Cov}(U_j, \tilde{Y}) - \mathrm{Cov}(U_j, \alpha_0 + \boldsymbol{U}^T\boldsymbol{\beta}_0) = \mathrm{Cov}(U_j, T) - \mathrm{Cov}(U_j, \alpha_0 + \boldsymbol{U}^T\boldsymbol{\beta}_0) = \mathrm{Cov}(U_j, \varepsilon) = 0$, for all j. □

Pollard's Functional Central Limit Theorem

We state Pollard's functional central limit theorem below for readers' convenience. Consider random processes developed from a triangular array $\{\tilde{f}_{ni}(t), i = 1, \ldots, m_n, t \in \mathcal{T}, n \in \mathbb{N}\}$, with the $\{\tilde{f}_{ni}\}$ independent within each row and $\mathcal{T}$ being the index set. In addition, we can define

$$\mathbb{W}_n(t) = \sum_{i \le m_n} (\tilde{f}_{ni}(t) - E\tilde{f}_{ni}(t))\,;\ \rho_n(s,t) = (\sum_{i \le m_n} E|\tilde{f}_{ni}(s) - \tilde{f}_{ni}(t)|^2)^{1/2}.$$

Let $UC(\mathcal{T}, \rho)$ denote the space of all bounded functions $\tilde{f}\colon \mathcal{T} \to \mathbb{R}$ which are uniformly ρ-continuous, that is, with any appropriately selected semimetric ρ,

$$\lim_{\tilde{\delta} \downarrow 0} \sup_{\rho(s,t) < \tilde{\delta}} |\tilde{f}(s) - \tilde{f}(t)| = 0.$$

Moreover, $\mathcal{T}$ is totally bounded by ρ (equivalently, $(\mathcal{T}, \rho)$ is totally bounded) if for every $\epsilon > 0$, there exists a finite collection $\mathcal{T}_m = \{t_1, \ldots, t_m\} \subset \mathcal{T}$ such that for all $t \in \mathcal{T}$, we have $\rho(s,t) \le \epsilon$ for some $s \in \mathcal{T}_m$. We would like to indicate that if the weak limit is a Gaussian process W, then the semimetric ρ can be selected as $\rho(s,t) = (E|W(s) - W(t)|^2)^{1/2}$.

Theorem (Pollard, (1990))**.** *Suppose the processes from the triangular array* $\{\tilde{f}_{ni}(t)\}$ *are independent within rows, and satisfies*

(A) the $\{\tilde{f}_{ni}\}$ *are manageable, with envelopes* $\{\tilde{F}_{ni}\}$ *which are also independent within rows;*

(B) $V(s,t) = \lim_{n\to\infty} E\mathbb{W}_n(s)\mathbb{W}_n(t)$ *exists for every* s, $t \in \mathcal{T}$*;*

(C) $\limsup_{n\to\infty} \sum_{i=1}^{m_n} E\tilde{F}_{ni}^2$ *is finite;*

(D) For each $\eta > 0$, $\lim_{n\to\infty} \sum_{i=1}^{m_n} E\tilde{F}_{ni}^2 1(\tilde{F}_{ni} > \eta) = 0$ *(an analogy to the Lindeberg condition);*

(E) For every $s, t \in \mathcal{T}$, $\rho(s,t) = \lim_{n\to\infty} \rho_n(s,t)$ *exists. And for all deterministic sequences* $\{s_n\}$ *and* $\{t_n\}$ *in* $\mathcal{T}$, $\rho_n(s_n,t_n) \to 0$ *if* $\rho(s_n,t_n) \to 0$.

Then, we have (i) $\mathcal{T}$ *is totally bounded under the pseudometric (semimetric)* ρ*;* (ii) *the finite dimensional distributions of* $\mathbb{W}_n$ *have Gaussian limits, with zero means and covariances given by* V*, which uniquely determine a Gaussian distribution concentrated on* $UC(\mathcal{T},\rho)$*;* (iii) $\mathbb{W}_n$ *converge weakly on* ℓ_τ^∞ *to a tight mean zero Gaussian process* W *concentrated on* $UC(\mathcal{T},\rho)$ *with* $V(s,t)$ *as covariance.*

Proof for Theorem 1

Theorem 1 follows from a series of lemmas below. To keep notational simplicity, we suppress the superscript "(n)" under the local model in this proof. Define the sample space by $\mathcal{X} = \mathcal{T} \times \{0,1\} \times \mathbb{R}^p$ (with σ-algebra $\mathcal{A}$), where we observe a random sample $\{X_i, \delta_i, \boldsymbol{U}_i\}_{i=1}^n$. In addition, $(X, \delta, \boldsymbol{U})$ follows a distribution P belonging to the set of Borel probability measure $\mathcal{P}$ on $(\mathcal{X}, \mathcal{A})$, and its empirical distribution is denoted by $\mathbb{P}_n$. In the following, we outline how all the lemmas play their roles to prove Theorem 1 and thereafter list these lemmas along with their corresponding proofs.

In Lemma 1, we take advantage of Stute's Theorem 1.1 (Stute (1995)), and express $\sqrt{n}[\hat{G}_n(t) - G_0(t)]$ as an i.i.d. sum, for any fixed t. Let $\mathbb{G}_n = \sqrt{n}(\mathbb{P}_n - P)$, and $\{\sqrt{n}[\hat{G}_n(t) - G_0(t)], t \in \mathcal{T}\}$ can be approximated by an empirical process (with probability approaching to one) $\mathbb{L}_n \colon \mathcal{X} \mapsto \ell_\tau^\infty$, which is

$$\{\mathbb{G}_n[\phi_t(X)\gamma_0(X)(1-\delta)+\gamma_1(X,t)\delta-\gamma_2(X,t)-G_0(t)], t\in\mathcal{T}\}$$

with ϕ_t, γ_0, γ_1 as well as γ_2 stated in Lemma 1. Moreover, we define a function $\Psi_j : \mathbb{R}\times\ell_\tau^\infty\times\mathcal{P}\to\mathbb{R}$, where

$$\Psi_j(m,h,Q)=m+Q\left[\frac{(U_j-EU_j)\tilde{Y}h(X-)}{G_0(X-)}\right], \tag{S.3}$$

and $\widetilde{\mathbb{M}}_n=\{\widetilde{\mathbb{M}}_{n,j}, j=1,\ldots,p\}$ with

$$\widetilde{\mathbb{M}}_{n,j}=\mathbb{G}_n\left(\tilde{\varepsilon}_n+(\boldsymbol{U}-\mathbb{P}_n\boldsymbol{U})^T\boldsymbol{\beta}_0-(U_j-\mathbb{P}_nU_j)C_j^T\boldsymbol{\beta}_0/V_j\right)(U_j-\mathbb{P}_nU_j),\ j=1,\ldots,p. \tag{S.4}$$

We further introduce Lemma 2 to indicate

$$\sqrt{n}(\hat{\theta}_n-\theta_n)S^2_{\hat{j}_n}=\Psi_{j_0}(\widetilde{\mathbb{M}}_{n,j_0},\mathbb{L}_n,\mathbb{P}_n).$$

To attain the limiting distribution of $\sqrt{n}\hat{\theta}_n$ when $\boldsymbol{\beta}_0\neq\mathbf{0}$, we need to derive the joint weak limit of $(\widetilde{\mathbb{M}}_n,\mathbb{L}_n)$. It suffices to show that the empirical process $\mathbb{W}_n=\{\mathbb{W}_n(t)=\mathbb{L}_n(t)+\sum_{j=1}^p a_j\widetilde{\mathbb{M}}_{n,j}, t\in\mathcal{T}\}$ converges weakly to a mean-zero Gaussian process $\mathbb{W}$ with covariance function σ_W, where for $(a_1,\ldots,a_p)\in\mathbb{R}^p$,

$$\begin{aligned}\mathbb{W}_n(t)&=\mathbb{L}_n(t)+\sum_{j=1}^p a_j\widetilde{\mathbb{M}}_{n,j}\\&=\mathbb{G}_n\{\phi_t(X)\gamma_0(X)(1-\delta)+\gamma_1(X,t)\delta-\gamma_2(X,t)-G_0(t-)\\&\qquad+\sum_{j=1}^p a_j(\tilde{\varepsilon}_n+(\boldsymbol{U}-\mathbb{P}_n\boldsymbol{U})^T\boldsymbol{\beta}_0-(U_j-\mathbb{P}_nU_j)C_j^T\boldsymbol{\beta}_0/V_j)(U_j-\mathbb{P}_nU_j)\}\end{aligned} \tag{S.5}$$

and

$$\sigma_W(s,t) = \sum_{j=1}^{p}\sum_{k=1}^{p} a_j a_k \sigma_M(j,k) + \sum_{j=1}^{p} a_j \sigma_{ML}(j,s) + \sum_{j=1}^{p} a_j \sigma_{ML}(j,t) + \sigma_L(s,t).$$

In Lemma 3, we obtain this desired result by checking some regularity conditions for Pollard's functional central limit theorem stated in Section 2. This result equivalently ensures the joint weak convergence of $(\widetilde{\mathbb{M}}_n, \mathbb{L}_n)$ to $(\boldsymbol{M}, \mathbb{L})$, where $(\boldsymbol{M}, \mathbb{L})$ is a mean-zero joint Gaussian process. Furthermore, multivariate central limit theorem implies that $\widetilde{\mathbb{M}}_n$ converges in distribution to a normal random vector $\boldsymbol{M}$, and the weak convergence of $\mathbb{L}_n$ to $\mathbb{L}$ can be developed by applying Pollard's functional central limit theorem again in a similar fashion as in Lemma 3.

In ensuing Lemma 4, we prove the continuity of Ψ_j on $\mathbb{R} \times \ell_\tau^\infty \times \mathcal{P}$ almost surely (a.s.) for all given j, and validate the application of continuous mapping theorem for empirical processes (Kosorok (2008), Chap. 7, Sec. 7.2.1). Based on Lemma 3 and 4, we develop the limiting distribution of $\sqrt{n}\hat{\theta}_n$ when $\boldsymbol{\beta}_0 \neq \mathbf{0}$ in Lemma 5. Moreover, we show the oracle property of $\hat{j}_n$ when $\boldsymbol{\beta}_0 \neq \mathbf{0}$ in Lemma 6. When $\boldsymbol{\beta}_0 = \mathbf{0}$, the joint limiting distribution of $\sqrt{n}\widehat{\boldsymbol{\theta}}$ and $n[S_Y^2 \mathbf{1}_p - \widehat{\boldsymbol{R}}]$ is constructed in Lemma 7. To establish the limiting distribution when $\boldsymbol{\beta}_0 = \mathbf{0}$, we use similar arguments to those in MQ's work (McKeague and Qian (2015)) and state one of their crucial lemmas as Lemma 8.

Let $\boldsymbol{Z} = (Z_1, \ldots, Z_p)^T$ in which the j-th component $Z_j = M_j + \varphi_j(\mathbb{L})$, and $\boldsymbol{Z}$ can be regarded as a function from $\mathbb{R}^{2p}$ to $\mathbb{R}^p$. Let $f(\boldsymbol{z}, \boldsymbol{b})_j = (z_j + C_j^T \boldsymbol{b})^2 / V_j$, for all j. Since $\boldsymbol{Z}$ is a random vector and $|\mathrm{Corr}(U_j, U_k)| < 1$ for $j \neq k$, it is indicated that $f(\boldsymbol{Z}, \boldsymbol{b}_0)_j \neq f(\boldsymbol{Z}, \boldsymbol{b}_0)_k$ for any $j \neq k$, a.s. Using Lemma 8, it further ensures that $J = \arg\max_{j=1,\ldots,p} f(\boldsymbol{Z}, \boldsymbol{b}_0)_j$ is uniquely determined a.s. Let $\boldsymbol{t}$ is a p-variate real

vector $(t_1, \ldots, t_p)^T$, and define

$$\boldsymbol{h}(\boldsymbol{t}) = (1(\arg\max_j t_j = 1), \ldots, 1(\arg\max_j t_j = p)).$$

We then show that $\sqrt{n}(\hat{\theta}_n - \theta_n)$ is a continuous function of $\sqrt{n}\widehat{\boldsymbol{\theta}}\boldsymbol{h}(n[S_Y^2\mathbf{1}_p - \widehat{\boldsymbol{R}}])$. Note that $\hat{j}_n$ is a unique maximizer to $n[S_Y^2\mathbf{1}_p - \widehat{\boldsymbol{R}}]$. Since both $\hat{j}_n$ and J are uniquely determined and $\boldsymbol{h}$ is continuous at $\boldsymbol{t}$ when $\arg\max_j t_j$ is unique, then in Lemma 9 we develop the desired limiting distribution of $\sqrt{n}\hat{\theta}_n$, applying continuous mapping theorem on the joint distribution of $\sqrt{n}\widehat{\boldsymbol{\theta}}$ and $n[S_Y^2\mathbf{1}_p - \widehat{\boldsymbol{R}}]$ obtained from Lemma 7. According to all the lemmas and results thereof (stated below in order), we finally complete this proof and show the desired theorem.

Lemma 1. *Suppose that all conditions for Theorem 1 hold and assume $\tau < \tau_H$, where $\tau_H = \inf\{x : H_0(x) = (F_0)(G_0)(x) = 0\} \leq \infty$ and F_0 is the survival function of T. For simplicity, we assume the continuity of F_0. For any fixed $t \leq \tau$,*

$$\sqrt{n}[\hat{G}_n(t) - G_0(t)] = -\mathbb{L}_n(t) + o_p(1),$$

where $\phi_t(\cdot) \equiv 1_{(-\infty,t]}(\cdot)$, and we define functions γ_0, γ_1 and γ_2 as follows.

Proof. For all $x \in \mathbb{R}$ and $t \leq \tau$, let

$$\begin{aligned}
\tilde{H}^0(x) &= P(X \leq x, \delta = 0) = -\int_{-\infty}^{x} F_0(y)G_0(dy);\\
\tilde{H}^1(x) &= P(X \leq x, \delta = 1) = -\int_{-\infty}^{x} G_0(y)F_0(dy);\\
\gamma_0(x) &= \exp\{\int_{-\infty}^{x} \frac{\tilde{H}^1(dy)}{H_0(y)}\};\\
\gamma_1(x,t) &= \frac{1}{H_0(x)}\int 1_{(x,\infty)}(w)\phi_t(w)\gamma_0(w)\tilde{H}^0(dw);\\
\gamma_2(x,t) &= \int\int \frac{1_{(-\infty,x)}(v)1_{(-\infty,w)}(v)\phi_t(w)\gamma_0(w)}{H_0(v)^2}\tilde{H}^1(dv)\tilde{H}^0(dw).
\end{aligned}$$

To apply Stute's Theorem 1.1 (Stute (1995)), we need to verify two conditions below:

$$\int \phi_t^2(x)\gamma_0^2(x)\tilde{H}^0(dx) = \int [\phi_t(X)\gamma_0(X)(1-\delta)]^2 dP_n < \infty;$$
$$-\int |\phi_t(x)|\Gamma^{1/2}(x)G_0(dx) < \infty, \text{ where } \Gamma(x) = \int_{-\infty}^{x} \frac{-F_0(dy)}{H_0(y)F_0(y)}.$$

Note that $\gamma_0(x) = F_0(x)^{-1}$ and the value of $1-\delta$ is either zero or one. The first condition then follows since

$$\int [\phi_t(X)\gamma_0(X)(1-\delta)]^2 dP_n \le \int_{-\infty}^{t} F_0(X)^{-2} dP_n < \frac{1}{F_0(\tau)^2} < \infty.$$

We wish to point out that the finiteness of $F_0(\tau)^{-2}$ in the above display is obvious because $\tau < \infty$ and $F_0(\infty) = 0$. Moreover for all $u \le \tau$, we have

$$\Gamma(u) \le \frac{-1}{H_0(\tau)} \left[\int_{-\infty}^{u} \frac{F_0(dy)}{F_0(y)} \right] = \frac{1}{H_0(\tau)F_0(\tau)} \left(1 - F_0(u)\right),$$

and then it implies that, for all $t \le \tau < \tau_H$,

$$\begin{aligned} -\int |\phi_t(x)|\Gamma^{1/2}(x)G_0(dx) &= -\int_{-\infty}^{t} \Gamma^{1/2}(x)G_0(dx) \le -\int_{-\infty}^{t} \Gamma^{1/2}(\tau)G_0(dx) \\ &\le \frac{(1-G_0(t))\sqrt{(1-F_0(\tau))}}{\sqrt{F_0(\tau)H_0(\tau)}} < \infty. \end{aligned}$$

Therefore, the second condition is shown satisfied. Since $P[\phi_t(X)\gamma_0(X)(1-\delta) + \gamma_1(X,t)\delta - \gamma_2(X,t) - G_0(t)] = 0$ for any fixed t, Stute's theorem implies Lemma 1 and further $\{\sqrt{n}[\hat{G}_n(t) - G_0(t)], t \in \mathcal{T}\}$ can be approximated by $-\mathbb{L}_n$ (with probability approaching to one). □

Lemma 2. *Suppose that all conditions for Theorem 1 hold and* $\boldsymbol{\beta}_0 \neq \mathbf{0}$.

$$\sqrt{n}(\hat{\theta}_n - \theta_n)S^2_{\hat{j}_n} = \Psi_{j_0}(\widetilde{\mathbb{M}}_{n,j_0}, \mathbb{L}_n, \mathbb{P}_n) + o_p(1),$$

where $\widetilde{\mathbb{M}}_{n,j}$, $\mathbb{L}_n$, *and* Ψ_j *are as previously defined.*

Proof. Since Proposition 1. implies $\mathrm{Cov}(U_j, T) = \mathrm{Cov}(U_j, \tilde{Y})$, for all j, then we can have

$$\begin{aligned}\sqrt{n}(\hat{\theta}_n - \theta_n)S^2_{\hat{j}_n} &= \sqrt{n}\left(\frac{1}{S^2_{\hat{j}_n}}\mathbb{P}_n(U_{\hat{j}_n} - \mathbb{P}_n U_{\hat{j}_n})Y - \frac{\mathrm{Cov}(U_{j_n}, T)}{V_{j_n}}\right)S^2_{\hat{j}_n} \\ &= \sqrt{n}\left(\mathbb{P}_n(U_{\hat{j}_n} - \mathbb{P}_n U_{\hat{j}_n})Y - \frac{\mathrm{Cov}(U_{j_n}, \tilde{Y})}{V_{j_n}}S^2_{\hat{j}_n}\right).\end{aligned} \tag{S.6}$$

Meanwhile, we can further observe that

$$\begin{aligned}&\sqrt{n}\mathbb{P}_n(U_{\hat{j}_n} - \mathbb{P}_n U_{\hat{j}_n})Y = \sqrt{n}\mathbb{P}_n(U_{\hat{j}_n} - \mathbb{P}_n U_{\hat{j}_n})\tilde{Y} + \sqrt{n}\mathbb{P}_n(U_{\hat{j}_n} - \mathbb{P}_n U_{\hat{j}_n})(Y - \tilde{Y}) \\ &= \sqrt{n}\mathbb{P}_n(U_{\hat{j}_n} - \mathbb{P}_n U_{\hat{j}_n})\tilde{Y} + \sqrt{n}\mathbb{P}_n(U_{\hat{j}_n} - \mathbb{P}_n U_{\hat{j}_n})\delta X\left[\frac{1}{\hat{G}_n(X-)} - \frac{1}{G_0(X-)}\right] \\ &= \sqrt{n}\mathbb{P}_n(U_{j_0} - \mathbb{P}_n U_{j_0})\tilde{Y} + \mathbb{P}_n\frac{(U_{j_0} - \mathbb{P}_n U_{j_0})\delta X\mathbb{L}_n(X-)}{G_0(X-)^2} + o_p(1),\end{aligned} \tag{S.7}$$

where the second term is contributed by the effect of estimating G_0 by the Kaplan-Meier estimator $\hat{G}_n$. The last equality in (S.7) can be ensured by $\hat{j}_n \xrightarrow{a.s.} j_0$ (shown in Lemma 4); the first order Taylor expansion around G_0, and Lemma 1.

Recall that $\tilde{Y} = \alpha_0 + \boldsymbol{U}^T\boldsymbol{\beta}_n + \tilde{\varepsilon}_n$, and then (S.6)-(S.7) further imply that

$$\begin{aligned}\sqrt{n}(\hat{\theta}_n - \theta_n)S^2_{\hat{j}_n} = \sqrt{n}\mathbb{P}_n\left(\tilde{\varepsilon}_n + (\boldsymbol{U} - \mathbb{P}_n\boldsymbol{U})^T\boldsymbol{\beta}_n - (U_{j_0} - \mathbb{P}_n U_{j_0})C^T_{j_0}\boldsymbol{\beta}_n/V_{j_0}\right)(U_{j_0} - \mathbb{P}_n U_{j_0}) \\ + \mathbb{P}_n\frac{(U_{j_0} - \mathbb{P}_n U_{j_0})\delta X\mathbb{L}_n(X-)}{G_0(X-)^2} + o_p(1).\end{aligned} \tag{S.8}$$

Because $\tilde{\varepsilon}_n$ and $\boldsymbol{U}$ are uncorrelated, it ensures that for any j,

$$P\left(\tilde{\varepsilon}_n + (\boldsymbol{U} - \mathbb{P}_n\boldsymbol{U})^T\boldsymbol{\beta}_n - (U_j - \mathbb{P}_n U_j)C^T_j\boldsymbol{\beta}_n/V_j\right)(U_j - \mathbb{P}_n U_j) = 0.$$

Since $\tilde{Y} = \delta X / G_0(X-)$, $\mathbb{P}_n U_{j_0} \xrightarrow{a.s.} EU_{j_0}$ along with (S.8), we can further have

$$\sqrt{n}(\hat{\theta}_n - \theta_n) S^2_{\hat{j}_n} = \mathbb{G}_n \left(\tilde{\varepsilon}_n + (\boldsymbol{U} - \mathbb{P}_n \boldsymbol{U})^T \boldsymbol{\beta}_n - (U_{j_0} - \mathbb{P}_n U_{j_0}) C_{j_0}^T \boldsymbol{\beta}_n / V_{j_0} \right) (U_{j_0} - \mathbb{P}_n U_{j_0}) + \mathbb{P}_n \frac{(U_{j_0} - EU_{j_0}) \tilde{Y} \mathbb{L}_n(X-)}{G_0(X-)} + o_p(1). \tag{S.9}$$

By $\boldsymbol{\beta}_n \to \boldsymbol{\beta}_0$ along with the definitions of $\widetilde{\mathbb{M}}_{n,j}$ and Ψ_j for any fixed j, (S.9) further gives the desired result. □

Lemma 3. *Suppose that all conditions for Theorem 1 hold. The empirical process* $\mathbb{W}_n$ *converges to a mean-zero Gaussian process* $\mathbb{W}$ *with covariance function* σ_W*, where for* $(a_1, \ldots, a_p) \in \mathbb{R}^p$,

$$\sigma_W(s,t) = \sum_{j=1}^{p} \sum_{k=1}^{p} a_j a_k \sigma_M(j,k) + \sum_{j=1}^{p} a_j \sigma_{ML}(j,s) + \sum_{j=1}^{p} a_j \sigma_{ML}(j,t) + \sigma_L(s,t)$$

with $\sigma_M(j,k)$, $\sigma_{ML}(j,t)$ *and* $\sigma_L(s,t)$ *provided in the proof, for any* j, k, s, t.

Proof. Recall that $\mathbb{W}_n = \{\mathbb{W}_n(t), t \in \mathcal{T}\}$, where

$$\begin{aligned}
\mathbb{W}_n(t) &= \mathbb{L}_n(t) + \sum_{j=1}^{p} a_j \widetilde{\mathbb{M}}_{n,j} \\
&= \mathbb{G}_n[\, \phi_t(X)\gamma_0(X)(1-\delta) + \gamma_1(X,t)\delta - \gamma_2(X,t) - G_0(t) \\
&\qquad + \sum_{j=1}^{p} a_j (\tilde{\varepsilon}_n + (\boldsymbol{U} - E\boldsymbol{U})^T \boldsymbol{\beta}_0 - (U_j - EU_j) C_j^T \boldsymbol{\beta}_0 / V_j)(U_j - EU_j)] + o_p(1).
\end{aligned}$$

Let $\boldsymbol{U}_{ij}$ denote the i-th subject's observation of U_j. The empirical process $\mathbb{W}_n$ can be approximated by triangular array:

$$\{h_{ni}(t) = \sum_{j=1}^{p} a_j f_{ni,j} + g_{ni}(t), i = 1, \ldots, n, t \in \mathcal{T}\},$$

where

$$f_{ni,j} = \frac{1}{\sqrt{n}} \left(\tilde{\varepsilon}_{ni} + (\boldsymbol{U}_i - E\boldsymbol{U})^T \boldsymbol{\beta}_0 - (\boldsymbol{U}_{ij} - EU_j) \frac{C_j^T \boldsymbol{\beta}_0}{V_j} \right) (\boldsymbol{U}_{ij} - EU_j)$$

and

$$g_{ni}(t) = \frac{1}{\sqrt{n}} [\phi_t(X_i)\gamma_0(X_i)(1-\delta_i) + \gamma_1(X_i,t)\delta_i - \gamma_2(X_i,t) - G_0(t)].$$

It is easy to see that $Ef_{ni,j} = 0$ for all i, j, and $Eg_{ni}(t) = 0$ for any $t \in \mathcal{T}$ (Stute (1995)). It implies that we can directly formulate $\mathbb{W}_n(t) = \sum_{i=1}^n h_{ni}(t) + o_p(1)$, and $\mathbb{L}_n(t) = \sum_{i=1}^n g_{ni}(t)$, respectively. Below, we check required conditions and apply Pollard's functional central limit theorem to establish the weak convergence of $\mathbb{W}_n$.

Condition (A) We start with verifying the manageability of triangular array. Let

$$\tilde{\mathcal{H}}_n = \{(h_{n1}(t), h_{n2}(t), \ldots, h_{nn}(t)) \in \mathbb{R}^n, t \in \mathcal{T}\}$$

whose envelope function is $\tilde{\boldsymbol{H}}_n = (H_{n1}, H_{n2}, \ldots, H_{nn}) \in \mathbb{R}^n$. For each i,

$$H_{ni} = \sum_{j=1}^{p} a_j F_{ni,j} + G_{ni} \quad \text{with} \quad F_{ni,j} = |f_{ni,j}| \quad \text{and} \quad G_{ni} = \sup_{t \in \mathcal{T}} |g_{ni}(t)|. \tag{S.10}$$

Let $\odot$ denote the operation of point-wise vector product. For any non-negative vector $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_n)^T \in \mathbb{R}^n$, we can create a class

$$\boldsymbol{\xi} \odot \tilde{\mathcal{H}}_n = \{(\xi_1 h_{n1}(t), \xi_2 h_{n2}(t), \ldots, \xi_n h_{nn}(t)) \in \mathbb{R}^n, t \in \mathcal{T}\}.$$

Let $\|\cdot\|$ denote L_2 norm, and $\|\cdot\|_{Q,2}$ denote $L_2(Q)$-norm, which is the norm of the class of square-integrable functions under a finitely discrete probability measure Q. Let $D(q, \mathcal{K})$ denote the packing number of class $\mathcal{K}$ (the maximal number of points that can fit in $\mathcal{K}$ while maintaining a distance greater than q (measured by a pre-specified

norm) between all points). Our triangular array of processes $\{h_{ni}(t), i = 1, \ldots, n, t \in \mathcal{T}\}$ is manageable (with respect to the envelopes $\tilde{\boldsymbol{H}}_n$) if we can find a deterministic function λ *(capacity bound)* such that

(1) $\int_0^1 \sqrt{\log \lambda(x)} dx < \infty$.

(2) $D(\zeta \| \boldsymbol{\xi} \odot \tilde{\boldsymbol{H}}_n \|, \boldsymbol{\xi} \odot \tilde{\mathcal{H}}_n) \leq \lambda(\zeta)$ for $0 < \zeta \leq 1$, $\boldsymbol{\xi} \in \mathbb{R}^n$ of non-negative weights, all $n \geq 1$.

Let $\boldsymbol{u}_j$ be the j-th element of $\boldsymbol{u} \in \mathbb{R}^p$. We define functions $f_{n,j} \colon \mathcal{X} \to \mathbb{R}$ and $g_{n,t} \colon \mathcal{T} \times \{0, 1\} \to \mathbb{R}$, where

$$f_{n,j}(x, d, \boldsymbol{u}) = \frac{1}{\sqrt{n}} \left[\left(\left(\frac{dx}{G_0(x-)} - \alpha_0 - \boldsymbol{u}^T \boldsymbol{\beta}_0 \right) + (\boldsymbol{u} - E\boldsymbol{U})^T \boldsymbol{\beta}_0 - (\boldsymbol{u}_j - EU_j) \frac{C_j^T \boldsymbol{\beta}_0}{V_j} \right) (\boldsymbol{u}_j - EU_j) \right],$$

and

$$g_{n,t}(x, d) = \frac{1}{\sqrt{n}} [\phi_t(x) \gamma_0(x)(1 - d) + \gamma_1(x, t) d - \gamma_2(x, t) - G_0(t-)].$$

We create another function class (changing with sample size n) $\mathcal{H}_n = \{h_{n,t}, t \in \mathcal{T}\}$, where the t-indexed function $h_{n,t} \colon \mathcal{X} \to \mathbb{R}$ is defined by

$$h_{n,t}(x, d, \boldsymbol{u}) = \sum_{j=1}^{p} a_j f_{n,j}(x, d, \boldsymbol{u}) + g_{n,t}(x, d) \text{ such that } h_{n,t}(X_i, \delta_i, \boldsymbol{U}_i) = h_{ni}(t).$$

Moreover, its envelope function $H_n : \mathcal{X} \to \mathbb{R}$, where $H_n(X_i, \delta_i, \boldsymbol{U}_i) = \sqrt{n} H_{ni}$. For any $\boldsymbol{\xi} \in \mathbb{R}^n$, it is easy to see that $\mathcal{H}_n \supseteq \boldsymbol{\xi} \odot \tilde{\mathcal{H}}_n$.

Let $N(q, \mathcal{K})$ denote the covering number of class $\mathcal{K}$ (the minimal number of closed balls of radius q (measured by a pre-specified norm) required to cover any class $\mathcal{K}$). Condition (1) for manageability could be fulfilled if we let

$$\lambda(x) = \limsup_{n \to \infty} \sup_{Q} N(x \| H_n \|_{Q,2} / 2, \mathcal{H}_n),$$

and if the class $\mathcal{H}_n$ satisfies the bounded uniform entropy integral (BUEI) condition

$$\limsup_{n\to\infty}\sup_{Q}\int_0^1\sqrt{\log N(x\|H_n\|_{Q,2}/2,\mathcal{H}_n)}dx<\infty, \tag{S.11}$$

where $\sup_Q$ means that the supremum is taken over all finitely discrete probability measures. To verify the BUEI condition in (S.11), it suffices to show $\mathcal{H}_n$ is a BUEI class, for all $n\geq 1$. Let $h^{\star}_{n,t}\colon \mathcal{X}\to\mathbb{R}$ and $h^{\divideontimes}_{n,t}\colon \mathcal{T}\times\{0,1\}\to\mathbb{R}$, where

$$h^{\star}_{n,t}(x,d,\boldsymbol{u})=\sum_{j=1}^{p}a_j f_{n,j}(x,d,\boldsymbol{u})+\frac{1}{\sqrt{n}}[\phi_t(x)\gamma_0(x)(1-d)+\gamma_1(x,t)d]$$

and

$$h^{\divideontimes}_{n,t}(x,d)=\frac{-1}{\sqrt{n}}[\gamma_2(x,t)+G_0(t-)],$$

such that we can further decompose $h_{n,t}=h^{\star}_{n,t}+h^{\divideontimes}_{n,t}$. Let $\mathcal{H}^{\star}_n=\{h^{\star}_{n,t},t\in\mathcal{T}\}$ and $\mathcal{H}^{\divideontimes}_n=\{h^{\divideontimes}_{n,t},t\in\mathcal{T}\}$. We can easily see for all $n\geq 1$, $\mathcal{H}^{\star}_n$ and $\mathcal{H}^{\divideontimes}_n$ are both VC classes because (1) the collection $\{(-\infty,t],t\in\mathcal{T}\}$ is a VC class (VC index=2), and (2) both $h^{\star}_{n,t}$ and $h^{\divideontimes}_{n,t}$ are monotone in t. Since VC class belongs to BUEI class, then $\mathcal{H}^{\star}_n$ and $\mathcal{H}^{\divideontimes}_n$ are both BUEI classes. The preservation property of BUEI class implies $\mathcal{H}_n$ is a BUEI class (Kosorok (2008)), such that for all $n\geq 1$,

$$\sup_{Q}\int_0^1\sqrt{\log N(x\|H_n\|_{Q,2}/2,\mathcal{H}_n)}dx<\infty.$$

Hence, the BUEI condition in (S.11) holds for $\mathcal{H}_n$. Subsequently, we verify Condition (2) for manageability as follows. For any $\boldsymbol{\xi}\in\mathbb{R}^n$, let $\|\cdot\|_{Q_{\boldsymbol{\xi}},2}$ denote $L_2(Q_{\boldsymbol{\xi}})$-norm, where $Q_{\boldsymbol{\xi}}$ is a finitely discrete probability measure:

$$Q_{\boldsymbol{\xi}}=(n\|\boldsymbol{\xi}\|)^{-1}\sum_{i=1}^{n}\xi_i^2 1(X_i,\delta_i,\boldsymbol{U}_i).$$

Thus for $0 < \zeta \leq 1$, $\boldsymbol{\xi} \in \mathbb{R}^n$ of non-negative weights and $n \geq 1$, we have

$$\zeta \|\boldsymbol{\xi} \odot \tilde{\boldsymbol{H}}_n\| = \zeta [\sum_{i=1}^{n} \xi_i^2 H_{ni}^2]^{1/2} = \zeta [\sum_{i=1}^{n} n^{-1} \xi_i^2 H_n^2(X_i, \delta_i, \boldsymbol{U}_i)]^{1/2}$$
$$\geq \zeta [\sum_{i=1}^{n} (n\|\boldsymbol{\xi}\|)^{-1} \xi_i^2 H_n^2(X_i, \delta_i, \boldsymbol{U}_i) 1(X_i, \delta_i, \boldsymbol{U}_i)]^{1/2} = \zeta \|H_n\|_{Q_{\boldsymbol{\xi}},2}.$$

Arguments used in Section 8.1.2 (Kosorok (2008), Chap. 8) indicate the relationship between packing number $D(q, \mathcal{K})$ and covering number $N(q, \mathcal{K})$ for each $q > 0$ and any class $\mathcal{K}$ with respect to the same norm :

$$N(q, \mathcal{K}) \leq D(q, \mathcal{K}) \leq N(q/2, \mathcal{K}).$$

If we let $q = \zeta \|\boldsymbol{\xi} \odot \tilde{\boldsymbol{H}}_n\|$, then this relationship implies for the class $\boldsymbol{\xi} \odot \tilde{\mathcal{H}}_n$,

$$D(\zeta \|\boldsymbol{\xi} \odot \tilde{\boldsymbol{H}}_n\|, \boldsymbol{\xi} \odot \tilde{\mathcal{H}}_n) \leq N(\zeta \|\boldsymbol{\xi} \odot \tilde{\boldsymbol{H}}_n\|/2, \boldsymbol{\xi} \odot \tilde{\mathcal{H}}_n).$$

Since we have perceived $\mathcal{H}_n \supseteq \boldsymbol{\xi} \odot \tilde{\mathcal{H}}_n$ and $\zeta \|\boldsymbol{\xi} \odot \tilde{\boldsymbol{H}}_n\| \geq \zeta \|H_n\|_{Q_{\boldsymbol{\xi}},2}$, it leads to

$$N(\zeta \|\boldsymbol{\xi} \odot \tilde{\boldsymbol{H}}_n\|/2, \boldsymbol{\xi} \odot \tilde{\mathcal{H}}_n) \leq N(\zeta \|H_n\|_{Q_{\boldsymbol{\xi}},2}/2, \mathcal{H}_n).$$

The above two equations further reveal that

$$D(\zeta \|\boldsymbol{\xi} \odot \tilde{\boldsymbol{H}}_n\|, \boldsymbol{\xi} \odot \tilde{\mathcal{H}}_n) \leq \sup_Q N(\zeta \|H_n\|_{Q,2}/2, \mathcal{H}_n). \tag{S.12}$$

Let $\lambda(\zeta) = \limsup_{n \to \infty} \sup_Q N(\zeta \|H_n\|_{Q,2}/2, \mathcal{H}_n)$. By (S.12), we can conclude

$$D(\zeta \|\boldsymbol{\xi} \odot \tilde{\boldsymbol{H}}_n\|, \boldsymbol{\xi} \odot \tilde{\mathcal{H}}_n) \leq \lambda(\zeta),$$

for $0 < \zeta \leq 1$, $\boldsymbol{\xi} \in \mathbb{R}^n$ of non-negative weights, and all $n \geq 1$. Note that λ does not depend on n.

Condition (B) Since $E\mathbb{W}_n(t) = 0$ for any t, we can obtain that for $s, t \in \mathcal{T}$,

$$\begin{aligned}\sigma_W(s,t) &= \lim_{n\to\infty} E\mathbb{W}_n(t)\mathbb{W}_n(s) = \lim_{n\to\infty}\sum_{i=1}^{n} Eh_{ni}(t)h_{ni}(s) \\ &= \sum_{j=1}^{p}\sum_{k=1}^{p}\tilde{a}_j\tilde{a}_k \lim_{n\to\infty}\sum_{i=1}^{n} Ef_{ni,j}f_{ni,k} + \sum_{j=1}^{p}\tilde{a}_j \lim_{n\to\infty}\sum_{i=1}^{n} Ef_{ni,j}g_{ni}(s) \\ &\quad + \sum_{j=1}^{p}\tilde{a}_j \lim_{n\to\infty}\sum_{i=1}^{n} Ef_{ni,j}g_{ni}(t) + \lim_{n\to\infty}\sum_{i=1}^{n} Eg_{ni}(s)g_{ni}(t) \\ &= \sum_{j=1}^{p}\sum_{k=1}^{p}\tilde{a}_j\tilde{a}_k\sigma_M(j,k) + \sum_{j=1}^{p}\tilde{a}_j\sigma_{ML}(j,s) + \sum_{j=1}^{p}\tilde{a}_j\sigma_{ML}(j,t) + \sigma_L(s,t).\end{aligned}$$

where $(\sigma_M(j,k))_{j,k=1,\ldots,p}$ is the covariance matrix of the mean-zero normal random vector $\boldsymbol{M}$, $\sigma_L(s,t)$ is the covariance function of the Gaussian process $\mathbb{L}$ at any s as well as t, and $\sigma_{ML}(j,t)$ is the covariance function of the joint Gaussian process $(\boldsymbol{M}, \mathbb{L})$ for any j, t.

Recall that $\tilde{Y} = \delta X/G_0(X-)$ and $\tilde{\varepsilon} = \tilde{Y} - \alpha_0 - \boldsymbol{U}^T\boldsymbol{\beta}_0$. Specifically, $(\sigma_M(j,k))_{j,k=1,\ldots,p}$ can be given by the covariance matrix of the random vector with components

$$(\tilde{\varepsilon} + (\boldsymbol{U} - E\boldsymbol{U})^T\boldsymbol{\beta}_0 - (U_j - EU_j)C_j^T\boldsymbol{\beta}_0/V_j)(U_j - EU_j), \tag{S.13}$$

for $j = 1, \ldots, p$. The dominated convergence theorem ensures that $\sigma_L(s,t)$ can be provided by the covariance function of a stochastic process at locations s and t, where the stochastic process is

$$\{\phi_t(X)\gamma_0(X)(1-\delta) + \gamma_1(X,t)\delta - \gamma_2(X,t) - G_0(t),\ t \in \mathcal{T}\}. \tag{S.14}$$

Moreover, we can obtain $\sigma_{ML}(j,t)$ by the cross covariance between the component $(\tilde{\varepsilon} + (\boldsymbol{U} - E\boldsymbol{U})^T\boldsymbol{\beta}_0 - (U_j - EU_j)C_j^T\boldsymbol{\beta}_0/V_j)(U_j - EU_j)$ and the process in (S.14) at location t. The square-integrability of $\tilde{\varepsilon}$ and the fourth moment condition of $\boldsymbol{U}$, along

with the results in Stute (1995), ensure the existence of $\sigma_W(s,t)$ for any $s,t \in \mathcal{T}$.

Condition (C) According to the definition of $\tilde{\boldsymbol{H}}_n$, we first express $\sum_{i=1}^n EH_{ni}^2$ as

$$\begin{aligned}
&\sum_{i=1}^n E(\sum_{j=1}^p a_j|f_{ni,j}| + \sup_{t\in\mathcal{T}}|g_{ni}(t)|)^2 \\
&= \sum_{i=1}^n \Big[\sum_{j,k=1}^p a_j a_k E|f_{ni,j}f_{ni,k}| + E(\sup_{t\in\mathcal{T}}|g_{ni}^2(t)|) + 2\sum_{j=1}^p a_j E(|f_{ni,j}|\sup_{t\in\mathcal{T}}|g_{ni}(t)|) \Big] \\
&\le \max_i\{ [\sum_{j,k=1}^p a_j a_k E(\tilde{\varepsilon}_{ni} + (\boldsymbol{U}_i - E\boldsymbol{U})^T\boldsymbol{\beta}_0 - (\boldsymbol{U}_{ij} - EU_j)\frac{C_j^T\boldsymbol{\beta}_0}{V_j})(\tilde{\varepsilon}_{ni} + (\boldsymbol{U}_i \\
&\qquad - E\boldsymbol{U})^T\boldsymbol{\beta}_0 - (\boldsymbol{U}_{ik} - EU_k)\frac{C_k^T\boldsymbol{\beta}_0}{V_k})(\boldsymbol{U}_{ij} - EU_j)(\boldsymbol{U}_{ik} - EU_k)\,] \\
&\qquad + E(\sup_{t\in\mathcal{T}}[\phi_t(X_i)\gamma_0(X_i)(1-\delta_i) + \gamma_1(X_i,t)\delta_i - \gamma_2(X_i,t) - G_0(t)]^2\,) \\
&\qquad + 2\sum_{j=1}^p a_j E(\,|(\tilde{\varepsilon}_{ni} + (\boldsymbol{U}_i - E\boldsymbol{U})^T\boldsymbol{\beta}_0 - (\boldsymbol{U}_{ij} - EU_j)\frac{C_j^T\boldsymbol{\beta}_0}{V_j})(\boldsymbol{U}_{ij} - EU_j)| \\
&\qquad \sup_{t\in\mathcal{T}}|\phi_t(X_i)\gamma_0(X_i)(1-\delta_i) + \gamma_1(X_i,t)\delta_i - \gamma_2(X_i,t) - G_0(t)|\,)\,\}.
\end{aligned} \tag{S.15}$$

We can further show that the first term in (S.15) is finite because both $\tilde{\varepsilon}$ and U_jU_k are square integrable, for all j,k. The restriction $X_i \le \tau < \tau_H$ leads to the uniform boundedness of the second term in (S.15) over $\mathcal{T}$, for all i. It is easy to see the third term is finite as well. Hence, we verify $\limsup_{n\to\infty}\sum_{i=1}^n EH_{ni}^2 < \infty$.

Condition (D) Recall that $H_n(X_i,\delta_i,\boldsymbol{U}_i) = \sqrt{n}H_{ni}$ and the definition of H_{ni} in (S.10). For each $\eta > 0$,

$$\sum_{i=1}^n EH_{ni}^2 1(H_{ni} > \eta) = n^{-1}\sum_{i=1}^n EH_n^2(X_i,\delta_i,\boldsymbol{U}_i)1(H_n(X_i,\delta_i,\boldsymbol{U}_i) > \eta\sqrt{n}), \tag{S.16}$$

where

$$H_n(X_i, \delta_i, \boldsymbol{U}_i) = \sqrt{n}[\sum_{j=1}^{p} a_j F_{ni,j} + G_{ni}]$$
$$= \sum_{j=1}^{p} a_j |(\tilde{\varepsilon}_{ni} + (\boldsymbol{U}_i - E\boldsymbol{U})^T \boldsymbol{\beta}_0 - (\boldsymbol{U}_{ij} - EU_j)\frac{C_j^T \boldsymbol{\beta}_0}{V_j})(\boldsymbol{U}_{ij} - EU_j)|$$
$$+ \sup_{t \in \mathcal{T}} |\phi_t(X_i)\gamma_0(X_i)(1-\delta_i) + \gamma_1(X_i, t)\delta_i - \gamma_2(X_i, t) - G_0(t)|.$$

Note that $\tilde{\varepsilon}$ and $U_j U_k$ are square-integrable for all j, k, and $\phi_t(X_i)\gamma_0(X_i)(1-\delta_i) + \gamma_1(X_i, t)\delta_i - \gamma_2(X_i, t) - G_0(t)$ is uniformly bounded over $\mathcal{T}$ for all i. Therefore, we have $H_n(X_i, \delta_i, \boldsymbol{U}_i)$ is bounded for all but finite many i for all $n \geq 1$. As $n \to \infty$, (S.16) tends to zero since the numerator is a finite sum but the denominator diverges. Hence, we show Condition (D) (the analogy of the Lindeberg condition) satisfied.

Condition (E) For every $s\ t \in \mathcal{T}$, $\rho_n(s,t) = (\sum_{i=1}^n E|h_{ni}(t) - h_{ni}(s)|^2)^{1/2}$, such that

$$\rho_n^2(s,t) = \sum_{i=1}^{n} E|h_{ni}(t) - h_{ni}(s)|^2 = \sum_{i=1}^{n} E|g_{ni}(t) - g_{ni}(s)|^2.$$

Without loss of generality, we assume $s < t$ and have $\sum_{i=1}^n E|g_{ni}(t) - g_{ni}(s)|^2$ as

$$\frac{1}{n}\sum_{i=1}^{n} E[\phi_{[s,t)}(X_i)\gamma_0(X_i)(1-\delta_i) + \tilde{\gamma}_1(X_i, s, t)\delta_i - \tilde{\gamma}_2(X_i, s, t) - G_0(t) + G_0(s)]^2,$$

where $\tilde{\gamma}_1(X, s, t) = \gamma_1(X, t) - \gamma_1(X, s)$ and $\tilde{\gamma}_2(X, s, t) = \gamma_2(X, t) - \gamma_2(X, s)$. The dominated convergence theorem implies that $\rho_n^2(s,t) \to \rho^2(s,t)$, where

$$\rho^2(s,t) = E[\phi_{[s,t)}(X)\gamma_0(X)(1-\delta) + \tilde{\gamma}_1(X, s, t)\delta - \tilde{\gamma}_2(X, s, t) - G_0(t) + G_0(s)]^2$$

and $\rho^2(s,t)$ can be easily shown finite. Since we obtain that $|\rho_n^2(s,t) - \rho^2(s,t)|$ converges to zero and $|\rho_n^2(s,t) - \rho^2(s,t)| = (\rho_n(s,t) + \rho(s,t))|\rho_n(s,t) - \rho(s,t)|$, then it further indicates the convergence of $|\rho_n(s,t) - \rho(s,t)|$ to zero by the fact that

$(\rho_n(s,t)+\rho(s,t))$ is definitely positive. We can also observe that, for any two deterministic sequences $\{s_n\}$ and $\{t_n\}$ in $\mathcal{T}$ and for all $n \geq 1$,

$$|\rho_n^2(s_n,t_n)| \leq |\rho_n^2(s_n,t_n)-\rho^2(s_n,t_n)| + |\rho^2(s_n,t_n)|. \tag{S.17}$$

If $\rho(s_n,t_n)$ converges to zero, then (S.17) would naturally imply $\rho_n^2(s_n,t_n) \to 0$ because we have

$$|\rho_n^2(s,t)-\rho^2(s,t)| \to 0, \text{ for arbitrary } s,t \in \mathcal{T}.$$

Hence, it leads to the convergence of $\rho_n(s_n,t_n)$ to zero for any two deterministic sequences $\{s_n\}$ and $\{t_n\}$ in $\mathcal{T}$, whenever $\rho(s_n,t_n)$ converges to zero. □

Lemma 4. *Suppose that all conditions for Theorem 1 hold and $\boldsymbol{\beta}_0 \neq \mathbf{0}$. The function Ψ_j is continuous on $\mathbb{R} \times \ell_\tau^\infty \times \mathcal{P}$, for all j.*

Proof. For $\alpha > 0$ and $A \in \mathcal{A}$, denote the Euclidean norm by $\|\cdot\|$ and we define the distance

$$d(\boldsymbol{x}, A) = \inf\{\|\boldsymbol{x}-\boldsymbol{a}\| : \boldsymbol{a} \in A\}$$

and $A_\alpha = \{\boldsymbol{x} : d(\boldsymbol{x}, A) \leq \alpha\}$ if $A \neq \emptyset$; otherwise, $A_\alpha = \emptyset$. For any probability measure $Q \in \mathcal{P}$, we can further define the Prokhorov metric between P and Q as

$$d_p(P,Q) = \inf\{\alpha > 0 : P(A) \leq Q(A_\alpha) + \alpha \text{ and } Q(A) \leq P(A_\alpha) + \alpha, \forall A \in \mathcal{A}\}.$$

For any given $\tilde{\epsilon} > 0$, suppose that there exists a probability measure $Q \in \mathcal{P}$ that satisfies $d_p(P,Q) < \tilde{\epsilon}$. Since $\tilde{\epsilon}$ can be arbitrarily small, it implies that there is a positive sequence $\alpha_n \downarrow 0$, such that $P(A) \leq Q(A_{\alpha_n}) + \alpha_n$ and $Q(A) \leq P(A_{\alpha_n}) + \alpha_n$, for all n. We can easily see A_α is closed, and therefore so is A_{α_n}. Let $\bar{A} = \cap_n A_{\alpha_n}$, where $\bar{A}$ is closed and $\bar{A}$ is exactly the closure of A. It follows that $P(A) \leq Q(\bar{A})$ and $Q(A) \leq P(\bar{A})$, which leads to $P(A) = Q(A)$ for all closed sets A. Hence, we can

conclude that $P = Q$ by inner regularity.

Recall that

$$\Psi_j(m, h, Q) = m + Q\left[\frac{(U_j - EU_j)\tilde{Y}h(X-)}{G_0(X-)}\right],$$

where we should point out that $Q[\cdot]$ is the expected value of a random variable with respect to the probability measure Q and EU_j denotes the expectation of U_j with respect to $P \in \mathcal{P}$. To show the continuity of Ψ_j on $\mathbb{R} \times \ell_\tau^\infty \times \mathcal{P}$, it suffices to prove that the second term of Ψ_j is continuous on $\ell_\tau^\infty \times \mathcal{P}$. For any $\epsilon > 0$, there exists $\tilde{\epsilon} > 0$ such that $\sup_{t\in\mathcal{T}} |\tilde{\mathbb{L}}(t) - \mathbb{L}(t)| < \tilde{\epsilon}$ and $d_p(\tilde{P}, P) < \tilde{\epsilon}$, where $\tilde{\mathbb{L}}, \mathbb{L} \in \ell_\tau^\infty$ and $\tilde{P}, P \in \mathcal{P}$. It follows that

$$\begin{aligned}
&\left|\tilde{P}\frac{(U_j - EU_j)\tilde{Y}\tilde{\mathbb{L}}(X-)}{G_0(X-)} - P\frac{(U_j - EU_j)\tilde{Y}\mathbb{L}(X-)}{G_0(X-)}\right| \\
&\quad\le \left|(\tilde{P} - P)\frac{(U_j - EU_j)\tilde{Y}\tilde{\mathbb{L}}(X-)}{G_0(X-)}\right| + \left|P\frac{(U_j - EU_j)\tilde{Y}(\tilde{\mathbb{L}}(X-) - \mathbb{L}(X-))}{G_0(X-)}\right| \\
&\quad\le \left|(\tilde{P} - P)\frac{(U_j - EU_j)\tilde{Y}\tilde{\mathbb{L}}(X-)}{G_0(X-)}\right| + P\left|\frac{(U_j - EU_j)\tilde{Y}(\tilde{\mathbb{L}}(X-) - \mathbb{L}(X-))}{G_0(X-)}\right| \\
&\quad\le \left|(\tilde{P} - P)\frac{(U_j - EU_j)\tilde{Y}\tilde{\mathbb{L}}(X-)}{G_0(X-)}\right| + \sup_{t\in\mathcal{T}}|\tilde{\mathbb{L}}(t) - \mathbb{L}(t)|P\left|\frac{(U_j - EU_j)\tilde{Y}}{G_0(X-)}\right|.
\end{aligned} \tag{S.18}$$

Because of $\tilde{P} = P$ by inner regularity, the first term in the last inequality from (S.18) disappears. Accompanying the square-integrability of $\tilde{\varepsilon}$ and U_jU_k for any j, k, and non-zero $G_0(t-)$ for all $t \in \mathcal{T}$, it leads to

$$P\left|\frac{(U_j - PU_j)\tilde{Y}}{G_0(X-)}\right| \le M,$$

where M is a constant. Hence, it implies that

$$\left|\tilde{P}\frac{(U_j - EU_j)\tilde{Y}\tilde{\mathbb{L}}(X-)}{G_0(X-)} - P\frac{(U_j - EU_j)\tilde{Y}\mathbb{L}(X-)}{G_0(X-)}\right| \le \tilde{\epsilon} \cdot M.$$

Let $\epsilon \geq \tilde{\epsilon} \cdot M$, and the proof of continuity is completed. □

Lemma 5. *Suppose that all conditions for Theorem 1 hold and* $\boldsymbol{\beta}_0 \neq \mathbf{0}$. *We have that*

$$\sqrt{n}(\hat{\theta}_n - \theta_n) \xrightarrow{d} \frac{\Psi_{j_0}(M_{j_0}, \mathbb{L}, P)}{V_{j_0}}.$$

Following notations in Theorem 1, it leads to

$$\sqrt{n}(\hat{\theta}_n - \theta_n) \xrightarrow{d} \frac{M_{j_0} + \varphi_{j_0}(\mathbb{L})}{V_{j_0}}, \text{ where } \varphi_{j_0}(\mathbb{L}) = E\left[\frac{(U_{j_0} - EU_{j_0})T\mathbb{L}(T-)}{G_0(T-)}\right].$$

Proof. Since Lemma 3 gives that $(\tilde{\mathbb{M}}_n, \mathbb{L}_n)$ converges weakly to $(\boldsymbol{M}, \mathbb{L})$ on $\mathbb{R}^p \times \ell_\tau^\infty$ a.s., and $\mathbb{P}_n$ converges a.s. to P, then we could have $(\tilde{\mathbb{M}}_n, \mathbb{L}_n, \mathbb{P}_n) \xrightarrow{d} (\boldsymbol{M}, \mathbb{L}, P)$ on $\mathbb{R}^p \times \ell_\tau^\infty \times \mathcal{P}$ a.s. It can further indicate that $(\mathbb{M}_{n,j_0}, \mathbb{L}_n, \mathbb{P}_n) \xrightarrow{d} (M_{j_0}, \mathbb{L}, P)$ on $\mathbb{R} \times \ell_\tau^\infty \times \mathcal{P}$ a.s. Recall that Lemma 2 gives

$$\sqrt{n}(\hat{\theta}_n - \theta_n)S^2_{\hat{j}_n} = \Psi_{j_0}(\mathbb{M}_{n,j_0}, \mathbb{L}_n, \mathbb{P}_n) + o_p(1).$$

Accompanying the continuity of Ψ_{j_0} shown in Lemma 4, therefore we can use continuous mapping theorem to develop that

$$\Psi_{j_0}(\mathbb{M}_{n,j_0}, \mathbb{L}_n, \mathbb{P}_n) \xrightarrow{d} \Psi_{j_0}(M_{j_0}, \mathbb{L}, P).$$

Along with the fact that $S^2_{\hat{j}_n}$ converges to V_{j_0} a.s. by $\hat{j}_n \xrightarrow{a.s.} j_0$ (shown in Lemma 6) and SLLN, Slutsky's lemma implies that,

$$\sqrt{n}(\hat{\theta}_n - \theta_n) \xrightarrow{d} \frac{\Psi_{j_0}(M_{j_0}, \mathbb{L}, P)}{V_{j_0}} = M_{j_0} + E\left[\frac{(U_{j_0} - EU_{j_0})T\mathbb{L}(T-)}{G_0(T-)}\right],$$

where the last equality follows from techniques of conditional expectation and the dominated convergence theorem. □

Lemma 6 (The oracle property)**.** *Suppose that all conditions for Theorem 1 hold and* $\boldsymbol{\beta}_0 \neq \mathbf{0}$*. We have* $\hat{j}_n$ *converges to* j_0 *a.s.*

Proof. Recall that $\boldsymbol{U}_{ij}$ denotes the i-th subject's U_j. Based on a marginal AFT model with respect to U_j, we can have mean squared errors $\hat{R}_j = \mathbb{P}_n(Y - \hat{\alpha}_j - \hat{\beta}_j U_j)^2$, where $(\hat{\alpha}_j, \hat{\beta}_j)$ denotes the KSV estimator of parameters in this marginal AFT model and can be written as $(\mathbb{P}_n Y - \hat{\beta}_j \mathbb{P}_n U_j,\ \mathbb{P}_n(U_j - \mathbb{P}_n U_j)Y/S_j^2)$. Therefore for all j, we have

$$\hat{R}_j = S_Y^2 - \mathbb{P}_n(U_j - \mathbb{P}_n U_j)Y/S_j^2,$$

and the above display indicates that the following two arguments are equivalent:

$$\arg\max_j \left| \frac{\mathbb{P}_n(U_j - \mathbb{P}_n U_j)Y}{S_Y S_j} \right| \text{ and } \arg\min_j \hat{R}_j. \tag{S.19}$$

Equation (S.19) reveals that

$$\hat{j}_n = \arg\max_j \left| \frac{\mathbb{P}_n(U_j - \mathbb{P}_n U_j)Y}{S_Y S_j} \right| = \arg\min_j \hat{R}_j.$$

We first need to prove: for all j,

$$\mathbb{P}_n U_j Y = \mathbb{P}_n U_j \tilde{Y} \text{ a.s., and } \mathbb{P}_n Y = \mathbb{P}_n \tilde{Y} \text{ a.s.} \tag{S.20}$$

To construct the first equality in (S.20), we re-express $\mathbb{P}_n U_j Y$ as

$$\mathbb{P}_n \left[\frac{U_j \delta X}{G_0(X-)} \right] - \mathbb{P}_n \left[\frac{U_j \delta X}{G_0(X-)} \left(\frac{\hat{G}_n(X-) - G_0(X-)}{\hat{G}_n(X-)} \right) \right],$$

which can be defined as $\mathbb{P}_n U_j \tilde{Y} - r_1$, and gives us $|\mathbb{P}_n U_j Y - \mathbb{P}_n U_j \tilde{Y}| = |r_1|$. It reveals

that the remainder term $|r_1|$ will be bounded by

$$\frac{\sup_{t\leq\tau}|\hat{G}_n(t)-G_0(t)|}{\hat{G}_n(\tau-)}\mathbb{P}_n\left|\frac{U_j\delta X}{G_0(X-)}\right|.$$

Note that this upper bound doesn't diverge since we assume non-zero $G_0(t-)$, for all $t\leq\tau$. Moreover, SLLN and the square-integrability of $\tilde{\varepsilon}$ as well as U_jU_k for all j,k imply that

$$\mathbb{P}_n\left|\frac{U_j\delta X}{G_0(X-)}\right|=\mathbb{P}_n|U_j\tilde{Y}|\overset{a.s.}{\to}E|U_jT|,$$

where we can easily see $E|U_jT|$ is a finite constant. Accompanying the strong uniform consistency of Kaplan-Meier estimator (Stute and Wang (1993)), it implies that the upper bound of $|r_1|$ converges to zero a.s., and so does $|r_1|$. Therefore, it leads to the first equality in (S.20). We can also ensure the second equality in (S.20) by similar arguments. Along with the square-integrability of $\tilde{\varepsilon}$ and U_jU_k for all j,k, we take advantage of SLLN and Proposition 1 to show that

$$\mathbb{P}_nU_j\tilde{Y}\overset{a.s.}{\to}EU_j\tilde{Y}=EU_jT \text{ and } \mathbb{P}_n\tilde{Y}\overset{a.s.}{\to}E\tilde{Y}=ET.$$

Combined with $\mathbb{P}_nU_j\overset{a.s.}{\to}EU_j$, the above display further indicates that

$$\mathbb{P}_n(U_j-\mathbb{P}_nU_j)\tilde{Y}\overset{a.s.}{\to}EU_jT-EU_jET=\mathrm{Cov}(U_j,T). \tag{S.21}$$

Because SLLN implies $\mathbb{P}_nU_j^2\overset{a.s.}{\to}EU_j^2$ and $\mathbb{P}_nU_j\overset{a.s.}{\to}EU_j$, it is also easy to see

$$S_j^2\overset{a.s.}{\to}V_j. \tag{S.22}$$

Applying continuous mapping theorem on (S.21) and (S.22), we can obtain that

$$\hat{\beta}_j = \frac{\mathbb{P}_n(U_j - \mathbb{P}_n U_j)\tilde{Y}}{S_j^2} \xrightarrow{a.s.} \frac{\mathrm{Cov}(U_j, T)}{V_j} \text{ for each } j,$$

so that

$$\widehat{\boldsymbol{\beta}} = (\hat{\beta}_1, \ldots, \hat{\beta}_p)^T \xrightarrow{a.s.} \left(\frac{\mathrm{Cov}(U_1, T)}{V_1}, \ldots, \frac{\mathrm{Cov}(U_p, T)}{V_p} \right)^T. \tag{S.23}$$

Let $\widehat{\boldsymbol{R}} = (\hat{R}_1, \ldots, \hat{R}_p)^T$ and $\mathbf{1}_p$ denote a p-variate vector $(1, \ldots, 1)^T$. When $\boldsymbol{\beta}_0 \neq \mathbf{0}$ such that $\mathrm{Var}(\boldsymbol{U}^T\boldsymbol{\beta}_0) > 0$, using continuous mapping theorem on (S.22) and (S.23) leads to

$$\begin{aligned} \frac{S_Y^2 \mathbf{1}_p - \widehat{\boldsymbol{R}}}{\mathrm{Var}(\boldsymbol{U}^T\boldsymbol{\beta}_0)} &= \left(\frac{\hat{\beta}_1^2 S_1^2}{\mathrm{Var}(\boldsymbol{U}^T\boldsymbol{\beta}_0)}, \ldots, \frac{\hat{\beta}_p^2 S_p^2}{\mathrm{Var}(\boldsymbol{U}^T\boldsymbol{\beta}_0)} \right) \xrightarrow{a.s.} \left(\frac{\mathrm{Cov}^2(U_1, T)}{V_1 \mathrm{Var}(\boldsymbol{U}^T\boldsymbol{\beta}_0)}, \ldots, \frac{\mathrm{Cov}^2(U_p, T)}{V_p \mathrm{Var}(\boldsymbol{U}^T\boldsymbol{\beta}_0)} \right)^T \\ &= (\mathrm{Corr}^2(U_1, T), \ldots, \mathrm{Corr}^2(U_p, T))^T. \end{aligned}$$

Note that $j_0 = \arg\max_j |\mathrm{Corr}(U_j, T)|$, which is equivalent to $j_0 = \arg\max_j \mathrm{Corr}^2(U_j, T)$. Since we have shown that $\hat{j}_n$ can also be the argument to maximize $(S_Y^2 - \hat{R}_j)/\mathrm{Var}(\boldsymbol{U}^T\boldsymbol{\beta}_0)$ among all j's, then $\hat{j}_n \xrightarrow{a.s.} j_0$, given that j_0 is unique. □

Lemma 7. *Suppose that all conditions for Theorem 1 hold and $\boldsymbol{\beta}_0 = \mathbf{0}$. The joint limiting distribution of $\sqrt{n}\widehat{\boldsymbol{\theta}}$ and $n(S_Y^2 \mathbf{1}_p - \widehat{\boldsymbol{R}})$ can be derived as*

$$\begin{pmatrix} (M_1 + \varphi_1(\mathbb{L}) + C_1^T \boldsymbol{b}_0)/V_1, & \ldots, & (M_p + \varphi_p(\mathbb{L}) + C_p^T \boldsymbol{b}_0)/V_p \\ (M_1 + \varphi_1(\mathbb{L}) + C_1^T \boldsymbol{b}_0)^2/V_1, & \ldots, & (M_p + \varphi_p(\mathbb{L}) + C_p^T \boldsymbol{b}_0)^2/V_p \end{pmatrix}^T,$$

where C_j as well as V_j are as previously defined, for any fixed j.

Proof. To prove this lemma, the first step is to derive the limiting distribution of $\sqrt{n}\widehat{\boldsymbol{\theta}} = (\sqrt{n}\hat{\theta}_1, \ldots, \sqrt{n}\hat{\theta}_p)^T$, where $\hat{\theta}_j$ is the KSV estimator of the regression coefficient

in a marginal AFT model with the predictor U_j and the outcome Y. The second step provides the joint limiting distribution of $\sqrt{n}\widehat{\boldsymbol{\theta}}$ and $n(S_Y^2\mathbf{1}_p - \widehat{\boldsymbol{R}}) = (n(S_Y^2 - \hat{R}_1), \ldots, n(S_Y^2 - \hat{R}_p))^T$, where $\hat{R}_j$ is defined as before, for all j. We begin with re-expressing $\sqrt{n}\hat{\theta}_j$ as $\sqrt{n}\mathbb{P}_n(U_j - \mathbb{P}_n U_j)Y/S_j^2$, which can be further written as, for all j,

$$\frac{\sqrt{n}}{S_j^2}\mathbb{P}_n(U_j - \mathbb{P}_n U_j)\tilde{Y} + \frac{1}{S_j^2}\mathbb{P}_n\left[\frac{(U_{j_0} - EU_{j_0})\tilde{Y}\mathbb{L}_n(X-)}{G_0(X-)}\right] + o_p(1). \tag{S.24}$$

Since $\tilde{\varepsilon}_n = \tilde{Y} - \alpha_0 - \boldsymbol{U}^T\boldsymbol{\beta}_n$, the linear property of sample covariance implies that

$$\mathbb{P}_n(U_j - \mathbb{P}_n U_j)\tilde{Y} = \mathbb{P}_n(U_j - \mathbb{P}_n U_j)\boldsymbol{U}^T\boldsymbol{\beta}_n + \mathbb{P}_n(U_j - \mathbb{P}_n U_j)\tilde{\varepsilon}_n, \tag{S.25}$$

where we can further have

$$\begin{aligned}\mathbb{P}_n(U_j - \mathbb{P}_n U_j)\boldsymbol{U}^T\boldsymbol{\beta}_n =& (\mathbb{P}_n - P)U_j\boldsymbol{U}^T\boldsymbol{\beta}_n + PU_j\boldsymbol{U}^T\boldsymbol{\beta}_n - (\mathbb{P}_n - P)U_j\mathbb{P}_n\boldsymbol{U}^T\boldsymbol{\beta}_n \\ & - PU_j(\mathbb{P}_n - P)\boldsymbol{U}^T\boldsymbol{\beta}_n - PU_jP\boldsymbol{U}^T\boldsymbol{\beta}_n,\end{aligned}$$

and

$$\mathbb{P}_n(U_j - \mathbb{P}_n U_j)\tilde{\varepsilon}_n = (\mathbb{P}_n - P)(\tilde{\varepsilon}_n(U_j - PU_j) - \mathbb{P}_n\tilde{\varepsilon}_n(\mathbb{P}_n - P)U_j. \tag{S.26}$$

Let $\mathbb{G}_n = \sqrt{n}(\mathbb{P}_n - P)$. Along with $C_j^T\boldsymbol{\beta}_n = PU_j\boldsymbol{U}^T\boldsymbol{\beta}_n - PU_jP\boldsymbol{U}^T\boldsymbol{\beta}_n$, (S.24)-(S.26) lead to

$$\begin{aligned}\sqrt{n}\hat{\theta}_j =& \frac{(\mathbb{G}_n U_j\boldsymbol{U}^T - PU_j\mathbb{G}_n\boldsymbol{U}^T - \mathbb{G}_n U_j\mathbb{P}_n\boldsymbol{U}^T)\boldsymbol{\beta}_n}{S_j^2} - \frac{\mathbb{P}_n\tilde{\varepsilon}_n\mathbb{G}_n U_j}{S_j^2} + \frac{\mathbb{G}_n\tilde{\varepsilon}_n(U_j - PU_j)}{S_j^2} \\ & + \frac{1}{S_j^2}\mathbb{P}_n\left[\frac{(U_{j_0} - EU_{j_0})\tilde{Y}\mathbb{L}_n(X-)}{G_0(X-)}\right] + \frac{\sqrt{n}C_j^T\boldsymbol{\beta}_n}{S_j^2} + o_p(1).\end{aligned}$$

When $\boldsymbol{\beta}_0 = \mathbf{0}$, then $\sqrt{n}\boldsymbol{\beta}_n = b_0$ such that

$$\sqrt{n}\hat{\theta}_j = \frac{(\mathbb{G}_n U_j \boldsymbol{U}^T - PU_j \mathbb{G}_n \boldsymbol{U}^T - \mathbb{G}_n U_j \mathbb{P}_n \boldsymbol{U}^T)\boldsymbol{b}_0}{\sqrt{n}\, S_j^2} - \frac{\mathbb{P}_n \tilde{\varepsilon}_n \mathbb{G}_n U_j}{S_j^2} + \frac{\mathbb{G}_n \tilde{\varepsilon}_n (U_j - PU_j)}{S_j^2} + \frac{1}{S_j^2}\mathbb{P}_n \left[\frac{(U_j - EU_j)\tilde{Y}\mathbb{L}_n(X-)}{G_0(X-)} \right] + \frac{C_j^T \boldsymbol{b}_0}{S_j^2} + o_p(1). \tag{S.27}$$

Since the first two terms in (S.27) are $o_p(1)$ by SLLN, along with the definition of $\mathbb{M}_{n,j}$, we can have

$$\begin{aligned}\sqrt{n}\hat{\theta}_j &= \frac{1}{S_j^2}\left\{ \mathbb{G}_n \tilde{\varepsilon}_n (U_j - PU_j) + \mathbb{P}_n \frac{(U_j - EU_j)\tilde{Y}\mathbb{L}_n(X-)}{G_0(X-)} \right\} + \frac{C_j^T \boldsymbol{b}_0}{S_j^2} + o_p(1) \\ &= \frac{1}{S_j^2}\left\{ \mathbb{M}_{n,j} + \mathbb{P}_n \frac{(U_j - EU_j)\tilde{Y}\mathbb{L}_n(X-)}{G_0(X-)} \right\} + \frac{C_j^T \boldsymbol{b}_0}{S_j^2} + o_p(1).\end{aligned}$$

Using previous arguments, it further leads to

$$\sqrt{n}\widehat{\boldsymbol{\theta}} \xrightarrow{d} \left(\frac{M_1 + \varphi_1(\mathbb{L}) + C_1^T \boldsymbol{b}_0}{V_1}, \ldots, \frac{M_p + \varphi_p(\mathbb{L}) + C_p^T \boldsymbol{b}_0}{V_p} \right)^T.$$

To complete the second step, we re-express $n(S_j^2 \mathbf{1}_p - \widehat{\boldsymbol{R}})$ as $(\sqrt{n}\widehat{\boldsymbol{\theta}}) \odot (\sqrt{n}\widehat{\boldsymbol{\theta}}) \odot (S_1^2, \ldots, S_p^2)^T$, where $\odot$ denotes the Hadamard product. Hence when $\boldsymbol{\beta}_0 = \mathbf{0}$, the joint distribution of $\sqrt{n}\widehat{\boldsymbol{\theta}}$ and $n(S_Y^2 \mathbf{1}_p - \widehat{\boldsymbol{R}})$ can be obtained as

$$\begin{pmatrix} \sqrt{n}\widehat{\boldsymbol{\theta}} \\ n(S_Y^2 \mathbf{1}_p - \widehat{\boldsymbol{R}}) \end{pmatrix} \xrightarrow{d} \begin{pmatrix} ((M_1 + \varphi_1(\mathbb{L}) + C_1^T \boldsymbol{b}_0)/V_1, & \ldots, & (M_p + \varphi_p(\mathbb{L}) + C_p^T \boldsymbol{b}_0)/V_p)^T \\ ((M_1 + \varphi_1(\mathbb{L}) + C_1^T \boldsymbol{b}_0)^2/V_1, & \ldots, & (M_p + \varphi_p(\mathbb{L}) + C_p^T \boldsymbol{b}_0)^2/V_p)^T \end{pmatrix}.$$

□

Lemma 8 (McKeague and Qian, (2015)). *Let $\boldsymbol{z}$ be a p-dimensional random vector and $f : \mathbb{R}^{2p} \to \mathbb{R}^p$ a function such that $f(\boldsymbol{z}, \cdot)$ is continuous for every $\boldsymbol{z} \in \mathbb{R}^p$, and $f(\boldsymbol{z}, \boldsymbol{b})_j \neq f(\boldsymbol{z}, \boldsymbol{b})_k$ a.s. for all $j \neq k$ and $\boldsymbol{b} \in \mathbb{R}^p$. Then, $J(\boldsymbol{b}) \equiv \arg\max_{j=1,\ldots,p} f(\boldsymbol{z}, \boldsymbol{b})_j$ is unique a.s. Also, if $\boldsymbol{b}_l \to \boldsymbol{b}_0$, then $J(\boldsymbol{b}_l) = J(\boldsymbol{b}_0)$ for l sufficiently large a.s.*

Lemma 9. *Suppose all conditions for Theorem 1 hold and* $\boldsymbol{\beta}_0 = \mathbf{0}$.

$$\sqrt{n}(\hat{\theta}_n - \theta_n) \xrightarrow{d} (M_J + \varphi_J(\mathbb{L}))/V_J + (C_J/V_J - C_{j(\boldsymbol{b}_0)}/V_{j(\boldsymbol{b}_0)})^T \boldsymbol{b}_0,$$

where J, $j(\boldsymbol{b}_0)$, C_j *and* V_j *are as defined in Theorem 1, for each* j.

Proof. It is easy to perceive $f(\boldsymbol{z}, \cdot)$ we defined is continuous with respect to $\boldsymbol{z}$. Also, $\boldsymbol{Z} = (Z_1, \ldots, Z_p)^T$ is a random vector and $|\mathrm{Corr}(U_j, U_k)| < 1$ for $j \neq k$, so it indicates that $f(\boldsymbol{Z}, \boldsymbol{b}_0)_j \neq f(\boldsymbol{Z}, \boldsymbol{b}_0)_k$ for any $j \neq k$ a.s., where $f(\boldsymbol{Z}, \boldsymbol{b}_0)_j = (Z_j + C_j^T \boldsymbol{b}_0)^2 / V_j$. Thus, we can point out that $J = J(\boldsymbol{b}_0) = \arg\max_{j=1,\ldots,p} f(\boldsymbol{Z}, \boldsymbol{b}_0)_j$ is unique a.s. Since $\hat{j}_n = \arg\min_j \hat{R}_j$ (equivalent to $\arg\max_j n(S_Y^2 - \hat{R}_j)$) and it is uniquely determined, then we can say that $\boldsymbol{h}(n(S_Y^2 \mathbf{1}_p - \widehat{\boldsymbol{R}}))$ is continuous. Moreover in the case of $\boldsymbol{\beta}_0 = \mathbf{0}$, we also see that

$$\sqrt{n}\hat{\theta}_n = \sqrt{n}\widehat{\boldsymbol{\theta}}\boldsymbol{h}(n(S_Y^2 \mathbf{1}_p - \widehat{\boldsymbol{R}})); \;\; \sqrt{n}\theta_n = \frac{\sqrt{n} C_{j(\boldsymbol{b}_0)}^T \boldsymbol{\beta}_n}{V_{j(\boldsymbol{b}_0)}} \equiv \frac{C_{j(\boldsymbol{b}_0)}^T \boldsymbol{b}_0}{V_{j(\boldsymbol{b}_0)}}.$$

Hence, the desired limiting distribution of $\sqrt{n}\hat{\theta}_n$ can be derived by applying continuous mapping theorem on the joint distribution of $\sqrt{n}\widehat{\boldsymbol{\theta}}$ and $n(S_Y^2 \mathbf{1}_p - \widehat{\boldsymbol{R}})$ derived in Lemma 7. □

Proof for Theorem 2

Before entering the core of the proof for Theorem 2, we clarify the large sample behavior of the maximally selected studentized statistic $\mathbb{T}_n$ in Lemma 10 below. Together with the conditions of the threshold λ_n, the results in this lemma would play a crucial role in designing adaptive resampling.

Lemma 10. *Suppose that the threshold* λ_n *satisfies* $\lambda_n = o(\sqrt{n})$ *and* $\lambda_n \to \infty$, *we have* $1(|\mathbb{T}_n| > \lambda_n) \xrightarrow{p} 1(\boldsymbol{\beta}_0 \neq \mathbf{0})$.

Proof. Recall that S_j^2 is the sample variance of U_j for all j and $\mathbb{T}_n = \sqrt{n}\hat{\theta}_n/\hat{\sigma}_n$, where $\hat{\sigma}_n^2 = \mathbb{P}_n(Y - \hat{\alpha}_n - \hat{\theta}_n U_{\hat{j}_n})^2/S_{\hat{j}_n}^2$. We start the proof with verifying that $\hat{\sigma}_n$ is asymptotically bounded above and below. Let $(\hat{\alpha}_j, \hat{\theta}_j)$ denote the estimated intercept and the estimated regression coefficient of U_j in the marginal AFT model that only contains one active predictor U_j for the outcome Y. By SLLN and the uniform consistency of Kaplan–Meier estimator, we can show $\hat{\theta}_j \xrightarrow{a.s.} \theta_j \equiv \text{Cov}(U_j, \boldsymbol{U})^T\boldsymbol{\beta}_0/\text{Var}(U_j)$ and $\hat{\alpha}_j \xrightarrow{a.s.} \alpha_0 + E\boldsymbol{U}^T\boldsymbol{\beta}_0 - \theta_j EU_j$, for all j. This further leads to

$$\mathbb{P}_n(Y - \hat{\alpha}_j - \hat{\theta}_n U_j)^2 \xrightarrow{a.s.} E(\tilde{Y} - \alpha_0 - E\boldsymbol{U}^T\boldsymbol{\beta}_0 - (U_j - EU_j)\theta_j)^2 = E(\tilde{\varepsilon} - (U_j - EU_j)\theta_j)^2.$$

Along with $S_j^2 \xrightarrow{a.s.} \text{Var}(U_j) > 0$ for all j, the continuous mapping theorem implies that

$$\frac{\mathbb{P}_n(Y - \hat{\alpha}_j - \hat{\theta}_j U_j)^2}{S_j^2} \xrightarrow{a.s.} \frac{E(\tilde{\varepsilon} - (U_j - EU_j)\theta_j)^2}{\text{Var}(U_j)}.$$

For all j, we ensure that $E(\tilde{\varepsilon} - (U_j - EU_j)\theta_j)^2 < \infty$ by $\text{Var}(U_j) > 0$ and the square-integrability of $\tilde{\varepsilon}$ and U_j. Therefore, $\max_{j=1,\ldots,p}\{\mathbb{P}_n(Y - \hat{\alpha}_j - \hat{\theta}_j U_j)^2/S_j^2\}$ converges to a finite constant. Since $\hat{\sigma}_n \leq [\max_{j=1,\ldots,p}\{\mathbb{P}_n(Y - \hat{\alpha}_j - \hat{\theta}_j U_j)^2/S_j^2\}]^{1/2}$, it implies that $\hat{\sigma}_n$ is asymptotically bounded above. Since it is obvious that

$$E(\tilde{\varepsilon} - (U_j - EU_j)\theta_j)^2/\text{Var}(U_j) > 0 \text{ for all } j,$$

then we see that $[\min_{j=1,\ldots,p}\{\mathbb{P}_n(Y - \hat{\alpha}_j - \hat{\theta}_j U_j)^2/S_j^2\}]^{1/2}$ converges to a non-zero finite constant. Because $\hat{\sigma}_n \geq [\min_{j=1,\ldots,p}\{\mathbb{P}_n(Y - \hat{\alpha}_j - \hat{\theta}_j U_j)^2/S_j^2\}]^{1/2}$, we therefore show that $\hat{\sigma}_n$ is asymptotically bounded below. Together with results in Theorem 1, we then prove that $|\mathbb{T}_n| \xrightarrow{a.s.} \infty$ when $\boldsymbol{\beta}_0 \neq \mathbf{0}$ and $|\mathbb{T}_n| = O_p(1)$ when $\boldsymbol{\beta}_0 = \mathbf{0}$.

To prove this lemma, it suffices to show that the probabilities in the follwing equation converge to zero:

$$
\begin{aligned}
E|1(|\mathbb{T}_n| > \lambda_n) - 1(\boldsymbol{\beta}_0 \neq \mathbf{0})| &= E|1(|\mathbb{T}_n| \leq \lambda_n) - 1(\boldsymbol{\beta}_0 = \mathbf{0})| \\
&= P(|\mathbb{T}_n| > \lambda_n, \boldsymbol{\beta}_0 = \mathbf{0}) + P(|\mathbb{T}_n| \leq \lambda_n, \boldsymbol{\beta}_0 \neq \mathbf{0}) \\
&= P(|\mathbb{T}_n| > \lambda_n | \boldsymbol{\beta}_0 = \mathbf{0}) 1(\boldsymbol{\beta}_0 = \mathbf{0}) + P(|\mathbb{T}_n| \leq \lambda_n | \boldsymbol{\beta}_0 \neq \mathbf{0}) 1(\boldsymbol{\beta}_0 \neq \mathbf{0}).
\end{aligned} \tag{S.28}
$$

We can see that the first probability in (S.28) converges to zero because $\lambda_n \to \infty$ along with $|\mathbb{T}_n| = O_p(1)$ when $\boldsymbol{\beta}_0 = \mathbf{0}$. Meanwhile, the second probability converges to zero because $\lambda_n = o(\sqrt{n})$ and $0 < |\mathbb{T}_n|/\sqrt{n} = O_p(1)$ when $\boldsymbol{\beta}_0 \neq \mathbf{0}$. □

More notations for the bootstrap version of estimators are introduced below. Let $\mathbb{P}_n^*$ be the nonparametric bootstrap of $\mathbb{P}_n$. Replacing P by $\mathbb{P}_n$ and $\mathbb{P}_n$ by $\mathbb{P}_n^*$, $\mathbb{G}_n^* = \sqrt{n}(\mathbb{P}_n^* - \mathbb{P}_n)$ is the bootstrapped empirical process, where $\mathbb{P}_n^*$, $\mathbb{P}_n$ and P only operate on functions defined on the sample space $\mathcal{X}$. The notation $\hat{\theta}_n^*$, $\hat{j}_n^*$ and $\hat{\theta}_j^*$ means that the bootstrap version of $\hat{\theta}_n$, $\hat{j}_n$ and $\hat{\theta}_j$, respectively. The bootstrapped Kaplan-Meier estimator is denote by $\hat{G}_n^*$. Note that under the operation of $\mathbb{P}_n^*$ or $\mathbb{G}_n^*$, we use $\hat{G}_n^*$ to replace $\hat{G}_n$ and $\hat{G}_n$ to replace G_0, respectively. All of the bootstrapped estimators are based on n i.i.d. observations taken from $\mathbb{P}_n$. Let E^* denote the expectation conditional on the data, and P^* be the corresponding probability measure.

To justify the claimed results, we first verify the following statements: (1) $1(|\mathbb{T}_n^*| > \lambda_n \text{ or } |\mathbb{T}_n| > \lambda_n) \xrightarrow{P^*} 1(\boldsymbol{\beta}_0 \neq \mathbf{0})$ and (2) $1(|\mathbb{T}_n^*| \leq \lambda_n \text{ and } |\mathbb{T}_n| \leq \lambda_n) \xrightarrow{P^*} 1(\boldsymbol{\beta}_0 = \mathbf{0})$ conditionally (on the data) in probability. Afterward, we prove Lemma 11 and 12, and obtain the desired results along with statements (1) and (2). To show statements (1) and (2), it suffices to give

$$
\begin{aligned}
&E^*|1(|\mathbb{T}_n^*| > \lambda_n) - 1(\boldsymbol{\beta}_0 \neq \mathbf{0})| = P^*(|\mathbb{T}_n^*| > \lambda_n, \boldsymbol{\beta}_0 = \mathbf{0}) + P^*(|\mathbb{T}_n^*| \leq \lambda_n, \boldsymbol{\beta}_0 \neq \mathbf{0}) \\
&= P^*(|\mathbb{T}_n^*| > \lambda_n | \boldsymbol{\beta}_0 = \mathbf{0}) 1(\boldsymbol{\beta}_0 = \mathbf{0}) + P^*(|\mathbb{T}_n^*| \leq \lambda_n | \boldsymbol{\beta}_0 \neq \mathbf{0}) 1(\boldsymbol{\beta}_0 \neq \mathbf{0}) \to 0
\end{aligned} \tag{S.29}
$$

in probability, implying that $1(|\mathbb{T}_n^*| > \lambda_n) \xrightarrow{p^*} 1(\boldsymbol{\beta}_0 \neq \mathbf{0})$ and $1(|\mathbb{T}_n^*| \leq \lambda_n) \xrightarrow{p^*} 1(\boldsymbol{\beta}_0 =$

$\mathbf{0}$) conditionally (on the data) in probability. The convergence in (S.29) follows from below arguments. Using Lemma 9 and the condition that $\lambda_n \to \infty$ as $n \to \infty$, we can have $P^*(|\mathbb{T}_n^*| > \lambda_n | \boldsymbol{\beta}_0 = \mathbf{0}) \to 0$ in probability. Besides, it is also easy to see $|\theta_n| \to |C_{j_0}^T \boldsymbol{\beta}_0|/V_{j_0}$ when $\boldsymbol{\beta}_0 \neq \mathbf{0}$ and j_0 is unique. Along with the condition that $\lambda_n = o(\sqrt{n})$ and that $\hat{\sigma}_n^*$ converges to a finite constant conditionally (on the data) in probability, we can use Lemma 5 and Lemma 11 (shown later) to prove

$$
\begin{aligned}
&P^*(|\mathbb{T}_n^*| \leq \lambda_n | \boldsymbol{\beta}_0 \neq \mathbf{0}) = P^*(\sqrt{n}|(\hat{\theta}_n^* - \hat{\theta}_n) + (\hat{\theta}_n - \theta_n) + \theta_n| \leq \lambda_n \hat{\sigma}_n^* | \boldsymbol{\beta}_0 \neq \mathbf{0}) \\
&\leq P^*(|\theta_n| \leq n^{-1/2}\lambda_n \hat{\sigma}_n^* + |\hat{\theta}_n^* - \hat{\theta}_n| + |\hat{\theta}_n - \theta_n| \,| \boldsymbol{\beta}_0 \neq \mathbf{0}) \to 0
\end{aligned}
$$

in probability. Since $1(|\mathbb{T}_n^*| > \lambda_n) \xrightarrow{p^*} 1(\boldsymbol{\beta}_0 \neq \mathbf{0})$ and $1(|\mathbb{T}_n^*| \leq \lambda_n) \xrightarrow{p^*} 1(\boldsymbol{\beta}_0 = \mathbf{0})$ conditionally (on the data) in probability, along with $1(|\mathbb{T}_n| > \lambda_n) \to 1(\boldsymbol{\beta}_0 \neq \mathbf{0})$ in probability, we can justify statements (1) and (2), using Slutsky's lemma.

Before stating necessary lemmas, we express the bootstrapped marginal regression coefficient as follows, which will appear in Lemma 11. For $j = 1, \ldots, p$,

$$
\begin{aligned}
\sqrt{n}\hat{\theta}_j^* &= \frac{\sqrt{n}[\mathbb{P}_n^* U_j Y - (\mathbb{P}_n^* U_j)(\mathbb{P}_n^* Y)]}{[\mathbb{P}_n^* U_j^2 - (\mathbb{P}_n^* U_j)^2]} \\
&= \frac{\mathbb{G}_n^* U_j Y - \mathbb{G}_n^* U_j \mathbb{P}_n^* Y - \mathbb{P}_n U_j \mathbb{G}_n^* Y + \sqrt{n}[\mathbb{P}_n U_j Y - \mathbb{P}_n U_j \mathbb{P}_n Y]}{[\mathbb{P}_n^* U_j^2 - (\mathbb{P}_n^* U_j)^2]} \\
&= \frac{\mathbb{G}_n^* U_j Y - \mathbb{G}_n^* U_j \mathbb{P}_n^* Y - \mathbb{P}_n U_j \mathbb{G}_n^* Y + \sqrt{n}\hat{\theta}_j[\mathbb{P}_n U_j^2 - (\mathbb{P}_n U_j)^2]}{[\mathbb{P}_n^* U_j^2 - (\mathbb{P}_n^* U_j)^2]}.
\end{aligned}
$$

Lemma 11. *Suppose the conditions for Theorem 1 hold and $\boldsymbol{\beta}_0 \neq \mathbf{0}$. We can have $\hat{j}_n^* \xrightarrow{p^*} j_0$ conditionally (on the data) a.s., and $\sqrt{n}(\hat{\theta}_n^* - \hat{\theta}_n) \xrightarrow{d} (M_{j_0} + \varphi_{j_0}(\mathbb{L}))/V_{j_0}$ conditionally (on the data) in probability.*

Proof. Let $S_Y^{*2} = \mathbb{P}_n^* Y^2 - (\mathbb{P}_n^* Y)^2$ and $S_j^{*2} = \mathbb{P}_n^* U_j^2 - (\mathbb{P}_n^* U_j)^2$. When $\boldsymbol{\beta}_0 \neq \mathbf{0}$, SLLN and Slutsky's lemma imply that,

$$S_j^{*2}\hat{\theta}_j^* = n^{-1/2}[\mathbb{G}_n^* U_j Y - \mathbb{G}_n^* U_j \mathbb{P}_n^* Y - \mathbb{P}_n U_j \mathbb{G}_n^* Y] + \hat{\theta}_j S_j^2 \xrightarrow{p^*} C_j^T \boldsymbol{\beta}_0 \text{ a.s.},$$

implying that $\hat{\theta}_j^* \xrightarrow{p^*} C_j^T \boldsymbol{\beta}_0 / V_j$ a.s., for $j = 1, \ldots, p$. Using a similar fashion to expressing the mean squared error, the corresponding bootstrap version can be written as $\hat{R}_j^* = S_Y^{*2} - \hat{\theta}_j^{*2} S_j^{*2}$, leading to that

$$\hat{j}_n^* = \arg\min_j \hat{R}_j^* = \arg\max_j \frac{S_Y^{*2} - \hat{R}_j^*}{\mathrm{Var}(\boldsymbol{U}^T \boldsymbol{\beta}_0)} = \arg\max_j \frac{\hat{\theta}_j^{*2} S_j^{*2}}{\mathrm{Var}(\boldsymbol{U}^T \boldsymbol{\beta}_0)}.$$

Moreover, Slutsky's lemma and SLLN indicate

$$\frac{\hat{\theta}_j^{*2} S_j^{*2}}{\mathrm{Var}(\boldsymbol{U}^T \boldsymbol{\beta}_0)} \xrightarrow{p^*} \mathrm{Corr}^2(U_j, \boldsymbol{U}^T \boldsymbol{\beta}_0) \text{ a.s., for } j = 1, \ldots, p.$$

Along with the condition that j_0 is unique when $\boldsymbol{\beta}_0 \neq \mathbf{0}$, it implies that

$$\begin{aligned} P^*(\hat{j}_n^* \neq j_0) &= P^* \left(\bigcup_{j: j \neq j_0} \left\{ \frac{\hat{\theta}_{j_0}^{*2} S_{j_0}^{*2}}{\mathrm{Var}(\boldsymbol{U}^T \boldsymbol{\beta}_0)} \leq \frac{\hat{\theta}_j^{*2} S_j^{*2}}{\mathrm{Var}(\boldsymbol{U}^T \boldsymbol{\beta}_0)} \right\} \right) \\ &\leq \sum_{j: j \neq j_0} P^* \left(\frac{\hat{\theta}_{j_0}^{*2} S_{j_0}^{*2}}{\mathrm{Var}(\boldsymbol{U}^T \boldsymbol{\beta}_0)} \leq \frac{\hat{\theta}_j^{*2} S_j^{*2}}{\mathrm{Var}(\boldsymbol{U}^T \boldsymbol{\beta}_0)} \right) \end{aligned}$$

converging to zero a.s. Let $\hat{\varepsilon}_n = Y - \hat{\alpha}_n - \hat{\theta}_n U_{\hat{j}_n}$. Recall that $\mathbb{P}_n \hat{\varepsilon}_n = 0$ and the definition of $\hat{\theta}_n^*$, we can have

$$\begin{aligned} \sqrt{n}(\hat{\theta}_n^* - \hat{\theta}_n) S_{\hat{j}_n^*}^{*2} &= \sqrt{n}[\mathbb{P}_n^* U_{\hat{j}_n^*} Y - \mathbb{P}_n^* U_{\hat{j}_n^*} \mathbb{P}_n^* Y - \hat{\theta}_n S_{\hat{j}_n^*}^{*2}] \\ &= \sqrt{n}[\mathbb{P}_n^* U_{\hat{j}_n^*} \hat{\varepsilon}_n - \mathbb{P}_n^* U_{\hat{j}_n^*} \mathbb{P}_n^* \hat{\varepsilon}_n - \hat{\theta}_n (\mathbb{P}_n^* U_{\hat{j}_n^*}^2 - (\mathbb{P}_n^* U_{\hat{j}_n^*})^2 - \mathbb{P}_n^* U_{\hat{j}_n^*} U_{\hat{j}_n} + \mathbb{P}_n^* U_{\hat{j}_n^*} \mathbb{P}_n^* U_{\hat{j}_n})] \\ &= \mathbb{G}_n^* U_{\hat{j}_n^*} \hat{\varepsilon}_n - \mathbb{G}_n^* \hat{\varepsilon}_n \mathbb{P}_n U_{\hat{j}_n^*} - \mathbb{G}_n^* U_{\hat{j}_n^*} \mathbb{P}_n^* \hat{\varepsilon}_n - \sqrt{n} \hat{\theta}_n [\mathbb{P}_n^* U_{\hat{j}_n^*}^2 - (\mathbb{P}_n^* U_{\hat{j}_n^*})^2 - \mathbb{P}_n^* U_{\hat{j}_n^*} U_{\hat{j}_n} \\ &\quad + \mathbb{P}_n^* U_{\hat{j}_n^*} \mathbb{P}_n^* U_{\hat{j}_n}] + o_{p^*}(1) \text{ a.s.} \\ &= \mathbb{G}_n^* \hat{\varepsilon}_n (U_{\hat{j}_n^*} - P U_{\hat{j}_n^*}) - \mathbb{G}_n^* \hat{\varepsilon}_n (\mathbb{P}_n - P) U_{\hat{j}_n^*} - \mathbb{G}_n^* U_{\hat{j}_n^*} (\mathbb{P}_n^* - \mathbb{P}_n) \hat{\varepsilon}_n \\ &\quad + \sqrt{n} \hat{\theta}_n [(\mathbb{P}_n^* U_{\hat{j}_n^*})^2 - \mathbb{P}_n^* U_{\hat{j}_n^*}^2 + \mathbb{P}_n^* U_{\hat{j}_n^*} U_{\hat{j}_n} - \mathbb{P}_n^* U_{\hat{j}_n^*} \mathbb{P}_n^* U_{\hat{j}_n}] + o_{p^*}(1) \text{ a.s.}, \end{aligned} \tag{S.30}$$

where the third equality follows from $\mathbb{P}_n U_{\hat{j}_n} \hat{\varepsilon}_n = 0$; $\hat{j}_n^* \xrightarrow{p^*} j_0$ a.s.; $\hat{j}_n \to j_0$ a.s., and the last equality follows from $\mathbb{P}_n \hat{\varepsilon}_n = 0$. In the last equality in (S.30), all the terms can be shown as $o_{p^*}(1)$ a.s. by similar arguments and SLLN, except for the first term. The next to show is the first term in (S.30) converges in distribution to some weak limit conditionally (on the data) in probability. According to Lemma 6, we can easily see that $\hat{\theta}_n \xrightarrow{p} \theta_0 \equiv C_{j_0}^T \boldsymbol{\beta}_0 / V_{j_0}$ and $\hat{\alpha}_n \xrightarrow{p} \alpha_0 + P\boldsymbol{U}^T \boldsymbol{\beta}_0 - \theta_0 P U_{j_0}$. Let $\bar{\varepsilon}_n = \tilde{\varepsilon}_n + (\boldsymbol{U} - P\boldsymbol{U})^T \boldsymbol{\beta}_0 - \theta_0 (U_{j_0} - P U_{j_0})$. The first term on the right-hand side (r.h.s.) of (S.30) can be decomposed as

$$\mathbb{G}_n^* \hat{\varepsilon}_n [(U_{\hat{j}_n^*} - P U_{\hat{j}_n^*}) - (U_{j_0} - P U_{j_0})] + \mathbb{G}_n^* (\hat{\varepsilon}_n - \bar{\varepsilon}_n)(U_{j_0} - P U_{j_0}) + \mathbb{G}_n^* \bar{\varepsilon}_n (U_{j_0} - P U_{j_0}). \quad \text{(S.31)}$$

In (S.31), the first term is $o_{p^*}(1)$ a.s. because for any $\epsilon > 0$,

$$P^*(\mathbb{G}_n^* \hat{\varepsilon}_n [(U_{\hat{j}_n^*} - P U_{\hat{j}_n^*}) - (U_{j_0} - P U_{j_0})] > \epsilon) \leq P^*(\hat{j}_n^* \neq j_0) \to 0 \text{ a.s.}$$

The second term in (S.31) can be reformatted as

$$\begin{aligned} &(\mathbb{P}_n^* - \mathbb{P}_n)[(U_{j_0} - P U_{j_0}) \boldsymbol{U}^T \boldsymbol{b}_0] - [\hat{\alpha}_n - (\alpha_0 + P\boldsymbol{U}^T \boldsymbol{\beta}_0 - \theta_0 P U_{j_0})] \mathbb{G}_n^* (U_{j_0} - P U_{j_0}) \\ &- (\hat{\theta}_n - \theta_0) \mathbb{G}_n^* U_{j_0} (U_{j_0} - P U_{j_0}) + \hat{\theta}_n \mathbb{G}_n^* [(U_{j_0} - U_{\hat{j}_n})(U_{j_0} - P U_{j_0})] \\ &+ \mathbb{G}_n^* (U_{j_0} - P U_{j_0})(Y - \tilde{Y}). \end{aligned} \quad \text{(S.32)}$$

Because $E^*[\hat{G}_n^*(t)] = \hat{G}_n(t)$ for all $t \in \mathcal{T}$ (Lo (1993)), along with first order Taylor expanding with respect to $\hat{G}_n$, the last term in (S.32) reduces to

$$\mathbb{P}_n^* \left[\frac{(U_{j_0} - P U_{j_0}) \tilde{Y} \mathbb{L}_n^*(X-)}{\hat{G}_n(X-)} \right] + o_{p^*}(1) \text{ a.s.},$$

where $\mathbb{L}_n^*\colon \mathcal{X} \mapsto \ell_\tau^\infty$ is a bootstrapped empirical process

$$\{\mathbb{G}_n^*[\phi_t(X)\gamma_0(X)(1-\delta) + \gamma_1(X,t)\delta - \gamma_2(X,t) - G_0(t)], t \in \mathcal{T}\}.$$

We use $\mathbb{L}_n^*$ to approximate $\{\sqrt{n}[\hat{G}_n^*(t) - \hat{G}_n(t)], t \in \mathcal{T}\}$ with ϕ_t, γ_0, γ_1 and γ_2 stated in Lemma 1. By the consistency of $(\hat{\alpha}_n, \hat{\theta}_n)$, bootstrap consistency of the sample mean and

$$P^*(\mathbb{G}_n^*[(U_{\hat{j}_n} - U_{j_0})(U_{j_0} - PU_{j_0})] > \epsilon) \leq 1(\hat{j}_n \neq j_0) \to 0 \text{ a.s.},$$

equation (S.32) reduces to

$$\mathbb{P}_n^* \left[\frac{(U_{j_0} - PU_{j_0})\tilde{Y}\mathbb{L}_n^*(X-)}{\hat{G}_n(X-)} \right] + o_{p^*}(1) \text{ in probability.} \tag{S.33}$$

Parallel to $\mathbb{M}_{n,j} = \mathbb{G}_n\tilde{\varepsilon}_n(U_j - \mathbb{P}_n U_j)$ for $j = 1, \ldots, p$, let

$$\mathbb{M}_{n,j}^* = \mathbb{G}_n^*\tilde{\varepsilon}_n(U_j - PU_j). \tag{S.34}$$

Since $\theta_0 = C_{j_0}^T \boldsymbol{\beta}_0 / V_{j_0}$ implying that $\bar{\varepsilon}_n = \tilde{\varepsilon}_n$, then we can express the remaining term in (S.31) $\mathbb{G}_n^*\bar{\varepsilon}_n(U_{j_0} - PU_{j_0})$ as $\mathbb{M}_{n,j_0}^*$. By the definition of Ψ_j in (S.3) and $EU_j = PU_j$ for all j along with the uniform consistency of $\hat{G}_n$, (S.30)-(S.33) lead to

$$\begin{aligned}
\sqrt{n}(\hat{\theta}_n^* - \hat{\theta}_n)S_{\hat{j}_n^*}^{*2} &= \mathbb{M}_{n,j_0}^* + \mathbb{P}_n^* \left[\frac{(U_{j_0} - EU_{j_0})\tilde{Y}\mathbb{L}_n^*(X-)}{G_0(X-)} \right] + o_{p^*}(1) \\
&= \Psi_{j_0}(\mathbb{M}_{n,j_0}^*, \mathbb{L}_n^*, \mathbb{P}_n^*) + o_{p^*}(1) \text{ in probability.}
\end{aligned}$$

Note that $S_{\hat{j}_n^*}^{*2} \xrightarrow{p^*} V_{j_0}$ in probability. Together with bootstrap consistency of Kaplan-Meier estimator based on Efron's resampling plan (Efron (1981), Akritas (1986)), we obtain the desired result, using similar arguments for the proofs of Lemmas 3-5

and Theorem 3.6.1 of van der Vaart and Wellner (van der Vaart and Wellner (1996), Chap. 3). □

Lemma 12. *Suppose that all conditions for Theorem 1 hold and* $\boldsymbol{\beta}_0 = \mathbf{0}$. *Then,* $\mathbb{Q}_n^*(\boldsymbol{b}_0)$ *converges to the limiting distribution of* $\sqrt{n}(\hat{\theta}_n - \theta_n)$ *conditionally (on the data) in probability.*

Proof. Following previous arguments, we can have

$$\sqrt{n}\hat{\theta}_j = (\mathbb{M}_{n,j} + \mathbb{D}_{n,j} + n^{-1}\sum_{i=1}^{n}(\boldsymbol{U}_{ij} - \bar{\boldsymbol{U}}_{\cdot j})\boldsymbol{U}_i^T\boldsymbol{b}_0)/S_j^2, \tag{S.35}$$

where

$$\mathbb{M}_{n,j} = \mathbb{G}_n\tilde{\varepsilon}_n(U_j - \mathbb{P}_nU_j)\,;$$
$$\mathbb{D}_{n,j} = \sqrt{n}\mathbb{P}_n(U_j - \mathbb{P}_nU_j)(Y - \tilde{Y}) = \mathbb{P}_n[(U_j - PU_j)\tilde{Y}\mathbb{L}_n(X-)/G_0(X-)] + o_p(1).$$

According to the definition of Ψ_j, (S.35) implies that

$$\sqrt{n}\hat{\theta}_j = \frac{\Psi_j(\mathbb{M}_{n,j}, \mathbb{L}_n, \mathbb{P}_n) + \widehat{\mathrm{Cov}}(U_j, \boldsymbol{U}^T\boldsymbol{b}_0)}{S_j^2}.$$

Let $\mathbb{M}_n$ be a p-dimensional vector with the j-th components given by $\mathbb{M}_{n,j}$. Let $\mathbb{J}_n(\boldsymbol{b})$ denote a p-dimensional vector with the j-th component defined by

$$\mathbb{J}_{n,j}(\boldsymbol{b}) = (\Psi_j(\mathbb{M}_{n,j}, \mathbb{L}_n, \mathbb{P}_n) + \widehat{\mathrm{Cov}}(U_j, \boldsymbol{U}^T\boldsymbol{b}))^2/S_j^2,$$

and $J(\boldsymbol{b})$ is a p-dimensional vector whose j-th component is $J_j(\boldsymbol{b}) = |\mathrm{Corr}(U_j, \boldsymbol{U}^T\boldsymbol{b})|$. Moreover, we define a $p \times p$ matrix $\mathbb{A}_n(\boldsymbol{b})$ whose (j,k)-th component is provided by

$$(\Psi_j(\mathbb{M}_{n,j}, \mathbb{L}_n, \mathbb{P}_n) + \widehat{\mathrm{Cov}}(U_j, \boldsymbol{U}^T\boldsymbol{b}))/S_j^2 - C_k/V_k.$$

In addition, let $\mathbb{H}_n(\boldsymbol{b})$ and $H(\boldsymbol{b})$ be p-dimensional vectors of zeros, except with a 1 at the entry that maximizes $\mathbb{J}_n(\boldsymbol{b})$ and $J(\boldsymbol{b})$, respectively. We can have that

$$\begin{aligned}\mathbb{Q}_n(\boldsymbol{b}) &= (\mathbb{M}_{n,J_n(\boldsymbol{b})} + \mathbb{D}_{n,J_n(\boldsymbol{b})} + \mathbb{P}_n(U_{J_n(\boldsymbol{b})} - \mathbb{P}_n U_{J_n(\boldsymbol{b})})\boldsymbol{U}^T\boldsymbol{b})/S^2_{J_n(\boldsymbol{b})} - C^T_{j(\boldsymbol{b})}\boldsymbol{b}/V_{j(\boldsymbol{b})} \\ &= \mathbb{H}_n(\boldsymbol{b})^T\mathbb{A}_n(\boldsymbol{b})H(\boldsymbol{b}).\end{aligned}$$

We define $\mathbb{J}(\boldsymbol{b})$, $\mathbb{A}(\boldsymbol{b})$ and $\mathbb{H}(\boldsymbol{b})$ as processes (not indexed by n) with the same form as $\mathbb{J}_n(\boldsymbol{b})$, $\mathbb{A}_n(\boldsymbol{b})$ and $\mathbb{H}_n(\boldsymbol{b})$, except with $\mathbb{M}_{n,j}$ replaced by M_j; $\mathbb{L}_n$ replaced by $\mathbb{L}$; $\mathbb{P}_n$ replaced by P, and the sample variance or covariances replaced by their population versions. According to Theorem 1, it implies that when $\boldsymbol{\beta}_0 = \mathbf{0}$,

$$\sqrt{n}(\hat{\theta}_n - \theta_n) = \mathbb{Q}_n(\boldsymbol{b}_0) = \mathbb{H}_n(\boldsymbol{b}_0)^T\mathbb{A}_n(\boldsymbol{b}_0)H(\boldsymbol{b}_0) \xrightarrow{d} \mathbb{H}(\boldsymbol{b}_0)^T\mathbb{A}(\boldsymbol{b}_0)H(\boldsymbol{b}_0). \tag{S.36}$$

Recall the bootstrap version of $\mathbb{M}_{n,j}$ defined in (S.34) Let $\mathbb{A}^*_n(\boldsymbol{b})$ and $\mathbb{J}^*_n(\boldsymbol{b})$ denote the bootstrap versions of $\mathbb{A}_n(\boldsymbol{b})$ and $\mathbb{J}_n(\boldsymbol{b})$, respectively, where the (j,k)-th component of $\mathbb{A}^*_n(\boldsymbol{b})$ is given by

$$\frac{\Psi^*_j(\mathbb{M}^*_{n,j}, \mathbb{L}^*_n, \mathbb{P}^*_n) + \widehat{\mathrm{Cov}}^*(U_j, \boldsymbol{U}^T\boldsymbol{b})}{S^{*2}_j} - \frac{\widehat{\mathrm{Cov}}(U_k, \boldsymbol{U}^T\boldsymbol{b})}{S^2_k},$$

and the j-th component of $\mathbb{J}^*_n(\boldsymbol{b})$ is provided by

$$\mathbb{J}^*_{n,j}(\boldsymbol{b}) = [\Psi^*_j(\mathbb{M}^*_{n,j}, \mathbb{L}^*_n, \mathbb{P}^*_n) + \widehat{\mathrm{Cov}}^*(U_j, \boldsymbol{U}^T\boldsymbol{b})]^2/S^{*2}_j.$$

The above display enables us to derive that, together with similar arguments used to close the proof of Lemma 11,

$$(\hat{H}_n(\boldsymbol{b}_0), \mathbb{A}^*_n(\boldsymbol{b}_0), \mathbb{J}^*_n(\boldsymbol{b}_0)) \xrightarrow{d} (H(\boldsymbol{b}_0), \mathbb{A}(\boldsymbol{b}_0), \mathbb{J}(\boldsymbol{b}_0))$$

conditionally (on the data) in probability, where $\hat{H}_n(\boldsymbol{b})$ denotes the sample version of $H(\boldsymbol{b})$. Moreover, we can observe that

$$\sqrt{n}\hat{\theta}_j^* = \frac{\Psi_j(\mathbb{M}_{n,j}^*, \mathbb{L}_n^*, \mathbb{P}_n^*) + \widehat{\mathrm{Cov}}^*(U_j, \boldsymbol{U}^T\boldsymbol{b}_0)}{S_j^{*2}} + o_{p^*}(1) \text{ a.s., for all } j,$$

Hence, parallel arguments to obtain (S.36) imply that

$$\sqrt{n}(\hat{\theta}_n^* - \hat{\theta}_n) = \mathbb{Q}_n^*(\boldsymbol{b}_0) = \mathbb{H}_n^*(\boldsymbol{b}_0)^T\mathbb{A}_n^*(\boldsymbol{b}_0)\hat{H}_n(\boldsymbol{b}_0) \xrightarrow{d} \mathbb{H}(\boldsymbol{b}_0)^T\mathbb{A}(\boldsymbol{b}_0)H(\boldsymbol{b}_0)$$

conditionally (on the data) in probability, where $\mathbb{H}_n^*(\boldsymbol{b})$ is a p-dimensional vector of zeros, except with a 1 at the entry that maximizes $\mathbb{J}_n^*(\boldsymbol{b})$. □