Statistica Sinica Preprint No: SS-2017-0298	
Title	Marginal screening for high-dimensional predictors of survival outcomes
Manuscript ID	SS-2017-0298
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202017.0298
Complete List of Authors	Tzu-Jung Huang Ian W. McKeague and Min Qian
Corresponding Author	Tzu-Jung Huang
E-mail	th2455@caa.columbia.edu
Notice: Accepted version subject to English editing.	

Marginal screening for high-dimensional predictors of survival outcomes

Tzu-Jung Huang, Ian W. McKeague, Min Qian

Department of Biostatistics, Columbia University

Abstract: This article 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 in the sense that the implicit variable selection step needs to be taken into account to avoid inflating the Type I error. McKeague and Qian (2015) constructed an adaptive resampling test to circumvent this problem under ordinary linear regression. To accommodate right censoring, we develop a new approach based on a maximally selected Koul–Susarla–Van Ryzin estimator from a marginal AFT working model. A regularized bootstrap method is used to calibrate the test. Our test is more powerful and less conservative than a Bonferroni correction of the marginal tests, and other competing methods. The proposed method is evaluated in simulation studies and applied to two real data sets.

Key words and phrases: Accelerated failure time model, Bootstrap, Family-wise error rate, Inverse probability weighting, Multiple testing, Post-selection inference

1 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 the 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).

The article is organized as follows. In Section 2 we formulate the testing problem, and introduce the proposed test statistic based on marginal KSV estimators. The adaptive bootstrap procedure used to calibrate the test is provided at the end of Section 2. In Section 3 we propose a variant of ARTS that adjusts for the effect of baseline clinical covariates. A forward-stepwise ARTS procedure is developed in Section 4. Various competing methods are discussed in Section 5. Numerical results reported in Section 6 show that ARTS has favorable performance compared with these competing methods. In Section 7 we present applications to gene expression data and primary biliary cirrhosis data. Concluding remarks are given in Section 8. Proofs of all the results are provided in the online supplementary materials.

2 ARTS procedure

2.1 Preliminaries

Koul et al. (1981)'s proposal for fitting the AFT model (1) is to replace T by the synthetic response $\tilde{Y} = \delta X/G_0(X)$, which is justified by the property

$$E[\tilde{Y}|\boldsymbol{U}] = E\left[\frac{\delta X}{G_0(X)} \mid \boldsymbol{U}\right] = E\left[\frac{T}{G_0(T)} E\left[\delta|T\right] \mid \boldsymbol{U}\right] = E[T|\boldsymbol{U}], \tag{2}$$

where G_0 is unknown but can be estimated by its Kaplan–Meier estimator. In other words, T and $\tilde{Y}$ have identical conditional means given $\boldsymbol{U}$ under the assumption of independent censoring. Therefore, we can recast the AFT model as $\tilde{Y} = \alpha_0 + \boldsymbol{U}^T\boldsymbol{\beta}_0 + \tilde{\varepsilon}$, using a new error term $\tilde{\varepsilon}$ that still has zero mean, finite variance, and is uncorrelated with $\boldsymbol{U}$ (see the supplementary materials for a detailed proof). Using similar arguments, it can also be shown that $E[\tilde{Y}^2] = E[T^2/G_0(T)] \geq E[T^2]$ and $E[U_j\tilde{Y}] = E[U_jT]$, for $j = 1, \ldots, p$. Hence this property further implies that the correlation between T and U_j is uniformly proportional to

the correlation between $\tilde{Y}$ and U_j over j, leading to the equality

$$\arg \max_{j=1,\ldots,p} |\mathrm{Corr}(U_j, T)| = \arg \max_{j=1,\ldots,p} |\mathrm{Corr}(U_j, \tilde{Y})|. \tag{3}$$

In the next section we will use (3) to reduce the screening problem to testing whether the most correlated predictor with T (or equivalently with $\tilde{Y}$) is significant. In practice we recommend pre-standardization of the predictors (as is common in variable selection) to provide scale-invariance, but we will develop the ARTS procedure in terms of the unstandardized predictors for simplicity of notation.

2.2 Maximally selected KSV estimator

To specify the most correlated predictor with T, we introduce the notation

$$j(\boldsymbol{b}) = \arg \max_{j=1,\ldots,p} |\mathrm{Corr}(U_j, \boldsymbol{U}^T\boldsymbol{b})| \text{ for any } \boldsymbol{b} \in \mathbb{R}^p. \tag{4}$$

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 therefore 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 (or equivalently to $\tilde{Y}$ by (3)). For notational simplicity, we denote the label $j(\boldsymbol{\beta}_0)$ by j_0 henceforth.

The synthetic response $\tilde{Y}$ is not observed, but it can be estimated by $Y = \delta X/\hat{G}_n(X)$, which leads to the sample version of j_0 given by

$$\hat{j}_n = \arg \max_{j=1,\ldots,p} \left| \frac{\mathbb{P}_n(U_j - \mathbb{P}_n U_j)Y}{S_j S_Y} \right|, \tag{5}$$

where $\mathbb{P}_n$ is the empirical distribution; S_j and S_Y are the sample standard deviations of U_j and Y, respectively. The best fitting marginal linear model for T with predictor U_{j_0} has intercept and slope

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

The maximally selected KSV estimator of (a_0, θ_0) is

$$(\hat{\alpha}_n, \hat{\theta}_n) = \left(\mathbb{P}_n Y - \hat{\theta}_n \mathbb{P}_n U_{\hat{j}_n}, \ \frac{1}{S^2_{\hat{j}_n}} \mathbb{P}_n (U_{\hat{j}_n} - \mathbb{P}_n U_{\hat{j}_n}) Y \right), \tag{6}$$

where $S^2_{\hat{j}_n}$ denotes the sample variance of $U_{\hat{j}_n}$. We reject H_0 in favor of H_A for extreme values of the test statistic $\sqrt{n}\hat{\theta}_n$.

2.3 Local behavior of $\hat{\theta}_n$

The challenge of calibrating a test based on $\sqrt{n}\hat{\theta}_n$ is to adapt to its non-regular limiting behavior at $\boldsymbol{\beta}_0 = \mathbf{0}$ (as shown in Theorem 1 below). To accurately capture the asymptotic behavior of $\hat{\theta}_n$ in $\sqrt{n}$-neighborhoods of $\boldsymbol{\beta}_0 = \mathbf{0}$, we consider the local linear model

$$T^{(n)} = \alpha_0 + \boldsymbol{U}^T \boldsymbol{\beta}_n + \varepsilon, \tag{7}$$

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

Under model (7), the observed time and the censoring status are denoted $X^{(n)} = \min(T^{(n)}, C)$ and $\delta^{(n)} = 1(T^{(n)} \leq C)$, respectively. We also define the synthetic response $\tilde{Y}^{(n)}$ and the estimated synthetic responses $Y^{(n)}$ in an analogous fashion:

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

For any fixed n, $\tilde{Y}^{(n)}$ has the same mean and the same covariance with $\boldsymbol{U}$ as $T^{(n)}$ does. The error term associated with $\tilde{Y}^{(n)}$ is $\tilde{\varepsilon}_n = \tilde{Y}^{(n)} - \alpha_0 - \boldsymbol{U}^T \boldsymbol{\beta}_n$, which also has zero mean and is

uncorrelated with $\boldsymbol{U}$. Instead of j_0, the label of the most correlated predictor with $T^{(n)}$ is

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

and our earlier hypotheses extend to

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

where

$$\theta_n = \frac{\mathrm{Cov}(U_{j_n}, T^{(n)})}{\mathrm{Var}(U_{j_n})}. \tag{8}$$

Note that $j_n = j(\boldsymbol{b}_0)$ when $\boldsymbol{\beta}_0 = \mathbf{0}$ but $\boldsymbol{b}_0 \neq \mathbf{0}$, and $j(\boldsymbol{b}_0)$ is assumed unique. Otherwise, j_n is not well-defined and the null hypothesis $\theta_n = 0$ holds when $\boldsymbol{\beta}_0 = \mathbf{0}$ and $\boldsymbol{b}_0 = \mathbf{0}$. If j_0 is unique, then $j_n \to j_0$. The estimators $\hat{j}_n$ and $\hat{\theta}_n$ are now defined by replacing Y by $Y^{(n)}$ in (5) and (6).

We develop the limiting distribution of $\sqrt{n}\hat{\theta}_n$ in the following theorem under assumptions (A.1)-(A.4) below. The proof is based on the functional delta method (van der Vaart (2000), Chap. 20) and a functional central limit theorem (Pollard (1990), Sec. 10), and is provided in supplementary materials.

(A.1) The predictors U_j, $j = 1, \ldots, p$, are bounded, and $|\mathrm{Corr}(U_j, U_k)| < 1$ for all $j \neq k$.

(A.2) The error term ε in (7) has zero mean, finite variance, and is uncorrelated with $\boldsymbol{U}$.

(A.3) The censoring time C is independent of $(T, \boldsymbol{U})$ and bounded above by τ (the time to the end of the follow-up).

(A.4) The marginal survival function of the censoring, G_0, is continuous on $\mathcal{T}$, and there exists a positive constant c_g such that $G_0(\tau) > c_g > 0$. Also, the marginal survival function of T, F_0, is continuous on $\mathcal{T}$, and there exists a positive constant c_f such that $F_0(\tau) > c_f > 0$.

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 suppose that regularity conditions (A.1)-(A.4) hold. Under the local model (7),*

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

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 supplementary materials. The j-indexed functional $\varphi_j \colon \ell_\tau^\infty \to \mathbb{R}$ is defined by

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

where ℓ_τ^∞ denotes the space of bounded functions on $\mathcal{T}$.

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 6), but also calibration via simulation from the estimated null limiting distribution of $\sqrt{n}\hat{\theta}_n$ *(later introduced as "CEND" in Section 5). 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

$$\sqrt{n}(\hat{\theta}_n - \theta_n)1(|\mathbb{T}_n| > \lambda_n \text{ or } \boldsymbol{\beta}_0 \neq \mathbf{0}) + \sqrt{n}(\hat{\theta}_n - \theta_n)1(|\mathbb{T}_n| \leq \lambda_n, \boldsymbol{\beta}_0 = \mathbf{0}), \tag{9}$$

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

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

with $(\hat{\alpha}_n, \hat{\theta}_n, \hat{j}_n)$ defined in (5) and (6). The statistic $\mathbb{T}_n$ serves for a pretest that is used to identify the non-regular situation in which we need a more accurate bootstrap strategy to capture the local asymptotic behavior of $\hat{\theta}_n$. Although the asymptotic variance of the KSV estimator in the fixed design case is known (Zhou (1992), Srinivasan and Zhou (1994)), in the present random design case it is simpler to avoid using such a complex standard error estimator. Instead we base the pretest on the relatively simple statistic $\mathbb{T}_n$. We show

$\hat{\sigma}_n^2$ is asymptotically bounded away from zero and bounded above (the proof is provided in supplementary materials). Together with results in Theorem 1, we further 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 (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 4.1 in supplementary materials along with a detailed proof). For estimating the second term in (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

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

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

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

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

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

Below we express the second term in (9) as a function $\mathbb{Q}_n(\boldsymbol{b}_0)$. When $\boldsymbol{\beta}_0 = \mathbf{0}$, it is easy to see

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

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 $\sqrt{n}\theta_n = C_{j(\boldsymbol{b}_0)}^T \boldsymbol{b}_0 / V_{j(\boldsymbol{b}_0)}$, 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

$$\sqrt{n}(\hat{\theta}_n - \theta_n) = \sqrt{n}(\hat{\theta}_n - \theta_n)1(|\mathbb{T}_n| > \lambda_n \text{ or } \boldsymbol{\beta}_0 \neq \mathbf{0}) + \mathbb{Q}_n(\boldsymbol{b}_0)1(|\mathbb{T}_n| \leq \lambda_n, \boldsymbol{\beta}_0 = \mathbf{0}). \quad (10)$$

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 (10) to consistently estimate the limiting distribution of $\sqrt{n}(\hat{\theta}_n - \theta_n)$. Hereafter, a superscript $*$ is used to indicate the bootstrap version of an estimator.

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 (7),*

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

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

2.4 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

$$B_n = \sqrt{n}(\hat{\theta}_n - \theta_n)1(|\mathbb{T}_n| > \lambda_n \text{ or } \boldsymbol{\beta}_0 \neq \mathbf{0}) + \mathbb{Q}_n(\mathbf{0})1(|\mathbb{T}_n| \leq \lambda_n, \boldsymbol{\beta}_0 = \mathbf{0}), \quad (11)$$

and the corresponding bootstrap version is

$$B_n^* = \sqrt{n}(\hat{\theta}_n^* - \hat{\theta}_n)1(|\mathbb{T}_n^*| > \lambda_n \text{ or } |\mathbb{T}_n| > \lambda_n) + \mathbb{Q}_n^*(\mathbf{0})1(|\mathbb{T}_n^*| \leq \lambda_n, |\mathbb{T}_n| \leq \lambda_n). \quad (12)$$

For some nominal level α, define the critical values 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 4.1 stated in 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 $\lambda_n \to \infty$, one 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, we use the double bootstrap. That is, 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 $[c_l^*, c_u^*]$, record as a rejection. The constant a is specified by the value that can have 5% of these 1000 ARTS procedures rejected. This data-driven selection of a will be adopted in our numerical studies and applications to real data. Note that in each bootstrap and nested double bootstrap sample, we set τ as the 90% empirical percentile of the observed time, and control the censoring rate around the same level as in the original data.

3 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. The International Non-Hodgkin's Lymphoma Prognostic Factors Project (1993)). Such baseline covariates (with moderate dimensionality)

do not need to be screened, but have to be incorporated as covariates in the AFT model. 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

$$T = \alpha_0 + \boldsymbol{U}^T\boldsymbol{\beta}_0 + \tilde{\boldsymbol{U}}^T\boldsymbol{\gamma}_0 + \varepsilon, \tag{13}$$

where $\boldsymbol{\gamma}_0 \in \mathbb{R}^q$; $\tilde{\boldsymbol{U}}$ is assumed bounded, 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 (13) as

$$T = \alpha_0' + \boldsymbol{D}^T\boldsymbol{\beta}_0 + \varepsilon', \tag{14}$$

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,

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

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 definition 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 (14), 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 $\hat{D}_j$ involves the least squares type estimate of $(\tilde{\alpha}_j, \tilde{\boldsymbol{\gamma}}_j)$ for $j = 1, \ldots, p$, we can use strong consistency of the estimates over all the j's (implied by SLLN and fixed p) to justify the replacement of $\boldsymbol{D}$ 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 considerable computation cost (caused by implementing projections every time when we have bootstrap or double bootstrap samples), especially when p is large. We tailor the adjustment of $\tilde{\boldsymbol{U}}$ to fit in the framework of ARTS, which avoids using a test statistic in matrix form that is inevitable when fitting a multi-variable 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.

4 Forward-stepwise ARTS

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 the adjusted ARTS procedure 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.

5 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 the adjusted ARTS procedure.

5.1 AFT model approaches

Marginal parametric AFT models with Bonferroni correction (BONF-AFT). A marginal parametric AFT model is often used to predict T from each predictor by specifying a parametric form of the error distribution, in which we can obtain the maximum likelihood estimate of the marginal regression coefficient of each predictor. 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 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 a standard normal distribution.

Marginal AFT models with higher criticism correction (HC). The higher criticism method is a test proposed by John Tukey for determining the overall significance of a collection of independent p-values. We use the statistic developed by Donoho and Jin, which is

expected to perform well if the predictors are nearly uncorrelated (Donoho and Jin (2004), Donoho and Jin (2015)).

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 (10) 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 Section 3.

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 $s^2 = \mathbb{P}_n(Y_i^{(n)} - \hat{\alpha}_n - \hat{\theta}_n U_{\hat{j}_n})^2$. At the null, Theorem 1 implies 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 $J = \arg\max_j \tilde{M}_j^2/V_j$. 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.

5.2 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 (namely, 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 $\boldsymbol{\beta}_0 = \boldsymbol{0}$.

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

$$L(v) = E_{\mathbf{r}}\left[\exp\left(\sum_{i=1}^{n}[\delta_i(\ln(h_0(X_i)) + r_i) - \exp(r_i)H_0(X_i)]\right)\right], \tag{15}$$

where $H_0(t) = \int_0^t h_0(s)ds$ is the cumulative baseline hazard function up to time t. Applying the second-order Taylor expansion on the exponential term in (15) with respect to $\mathbf{r}$, $L(v)$ can be expressed by the first and second moments of $\mathbf{r}$ (Le Cessie and van Houwelingen (1995)). This implies that the desired test statistic is able to be established in terms of the score function of v, which only involves the first and second moments of $\boldsymbol{\beta}_0$ without specifying the prior distribution. 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 coefficients of $\tilde{\boldsymbol{U}}$.

6 Numerical studies

6.1 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 $T = \sum_{j=1}^{p} \beta_j U_j + \varepsilon$ with $\beta_1 = \ldots = \beta_5 = 0.15$, $\beta_6 = \ldots = \beta_{10} = -0.1$, 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 component $U_j \sim \mathcal{N}(0,1)$, 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 $h(t|\boldsymbol{U}) = 2\exp(t)\exp(\sum_{j=1}^{p}\beta_j U_j)$ with the value of $(\beta_1, \ldots, \beta_p)$ as stated in **Model 3**.

To achieve designed censoring rates, we generate the censoring time 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. The selection of the threshold λ_n follows the steps stated in Section 2.4. To provide a full comparison, we compare the performance of ARTS with the competing methods that have been introduced in Section 5. Empirical rejection rates based on 1000 Monte Carlo replications under various censoring rates are displayed in Figures 1-2. 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 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 HC method is anti-conservative and fails to control Type I error; we suspect this is due to the relatively high correlation between predictors, for which HC is not designed. 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 considerably exceeding the nominal level under different sample sizes and various censoring rates (and thus going out of range somewhere in the left panels of Figure 1). The global test based on permutation arguments (GLOBAL-permut) takes good control of Type I error rates but claims 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 (the results of PLRT 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 the results of Model 2 and 3 shows no adverse impact on the power of ARTS when the maximally correlated predictor is non-unique.

In Figure 2 where data are not generated from AFT models, ARTS retains good control of Type I error rates. 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 maintains 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 1. Even though unstable in power due to model misspecification, ARTS still strikes more adequate balance between controlling Type I error and achieving power than other methods, especially in the cases of light or moderate censoring and large sample size. Comparing Figure 1 with Figure 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 and power than Cox-model-relevant approaches when the censoring is light or moderate and $n = 200$.

6.2 Screening performance of ARTS

We further assess the performance of ARTS as a full screening method (i.e. retaining all covariates whose marginal test statistics are beyond the critical values calculated for $\sqrt{n}\hat{\theta}_n$) in terms of false discovery rate (FDR), false negative rate (FNR) and false positive rate (FPR). Through a simulation study, we compare the screening performance of ARTS with the Benjamini–Hochberg procedure (BH, Benjamini and Hochberg (1995)) and the Holm–Bonferroni procedure (HB, Holm (1979)). Relevant results are presented in Section S5 of the supplementary materials.

The power (as given by the average values of (1 - FNR)) is seen to be slightly less for ARTS than for BH, which is expected because the acceptance region is constructed by the critical values of the maximum correlation statistic $\hat{\theta}_n$, leading to more conservative results. We expect, however, that forward-stepwise ARTS will have better performance than the ARTS screening procedure we just described because it re-calibrates at each step. In terms of FDR and FPR, the performance of ARTS and BH are comparable, although that of the Bonferroni method is more conservative as expected. The HB and Bonferroni methods have similar performance, with respect to all the measures.

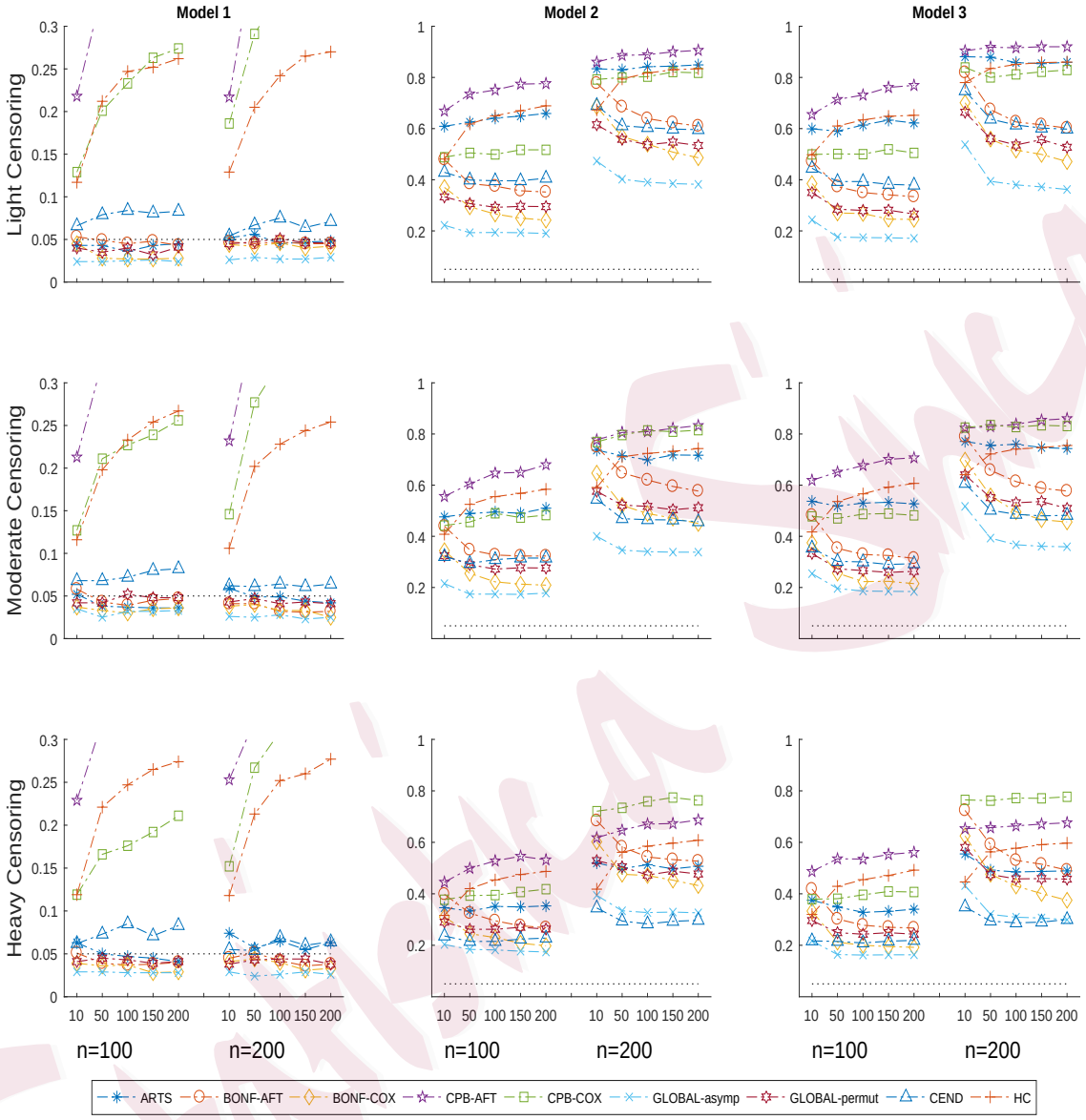

Figure 1: Empirical rejection rates based on 1000 samples generated from Model 1-3 with the dimension ranging from $p = 10$ to $p = 200$.

6.3 Asymptotic power evaluation

In this section, we conduct a simulation study to evaluate the asymptotic FWER and power of ARTS, compared with those of BONF-AFT. We will assess the asymptotic FWER and power based on the limiting distribution shown in Theorem 1. This approach can be a

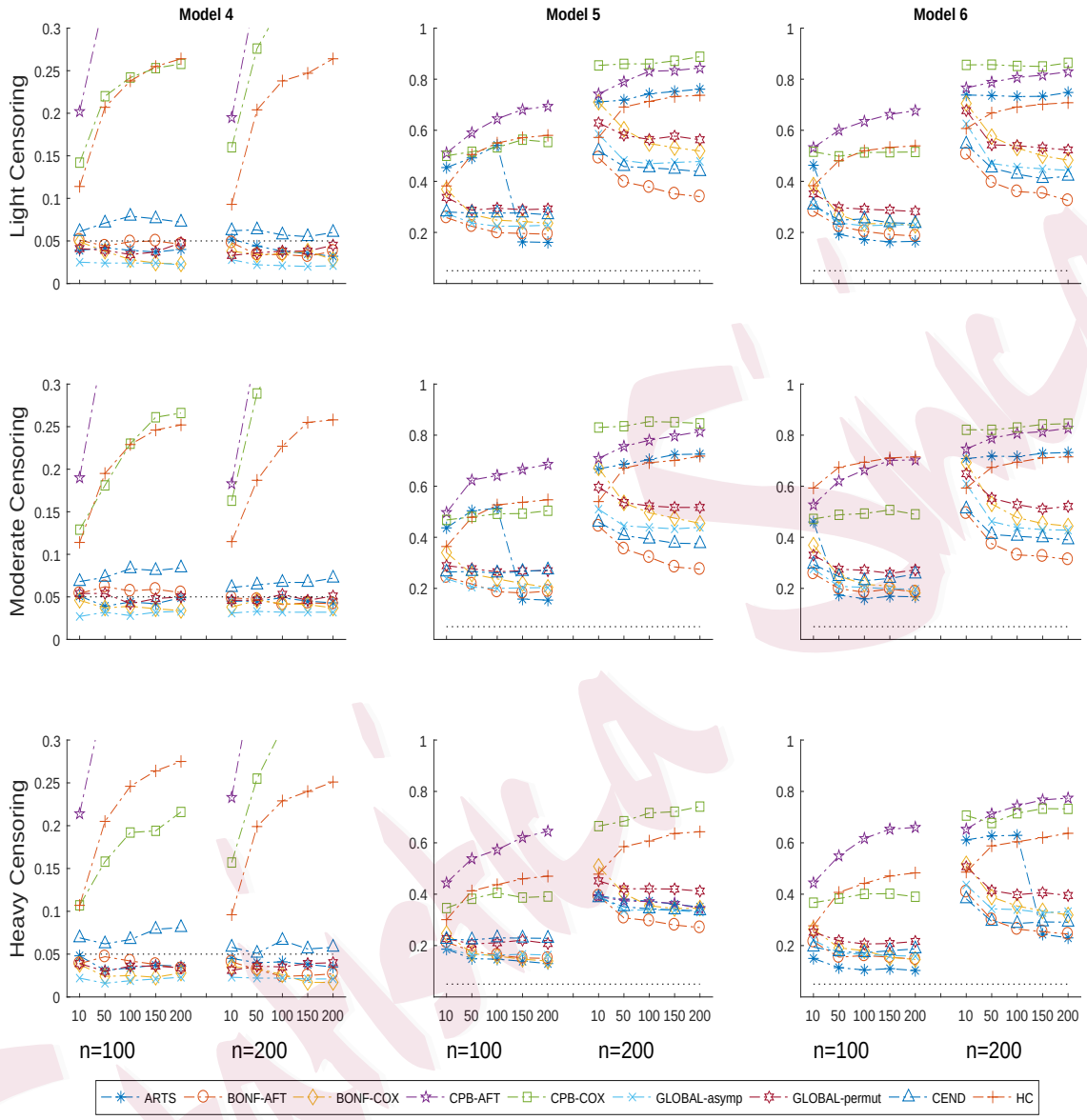

Figure 2: Empirical rejection rates based on 1000 samples generated from Model 4-6 with the dimension ranging from $p = 10$ to $p = 200$.

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 heavy computation to implement 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 the asymptotic power at 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

$$T^{(n)} = (n^{-1/2} b_0) U_1 + \varepsilon, \tag{16}$$

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 6.1. We allow b_0 to vary over a grid in $[0, 5]$ by increments of 0.5. Under this local model, the complex limiting distribution reduces to a simpler form:

$$\sqrt{n}(\hat{\theta}_n - \theta_n) \xrightarrow{d} (M_J + b_0 \mathrm{Cov}(U_J, U_1))/\mathrm{Var}(U_J) - b_0, \tag{17}$$

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 (16) 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 (17), 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 assessed by computing the proportion of falling beyond $[c_l, c_u]$ among $10,000$ realizations of $\sqrt{n}\hat{\theta}_n$.

To reflect the random variation of the asymptotic FWER and power over samples generated in Step 1, we independently implement the above procedure 20 times and display these corresponding asymptotic rejection rates in a box plot, for each considered b_0. 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 3), moderate censoring (Figure 4) and heavy censoring (Figure 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 gives more anti-conservative results as observed in previous numerical studies. When $a = 0$, in particular, ARTS reduces to CPB-AFT. On the other hand, ARTS behaves more stably and provides more accurate control of Type I error rates as a grows larger. We also perceive that the variation within each boxplot decreases when the value of a increases.

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% when $a = 2$ but drops below 5% at $a = 4$.

6.4 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. The FWERs of other AFT-model-relevant methods are also

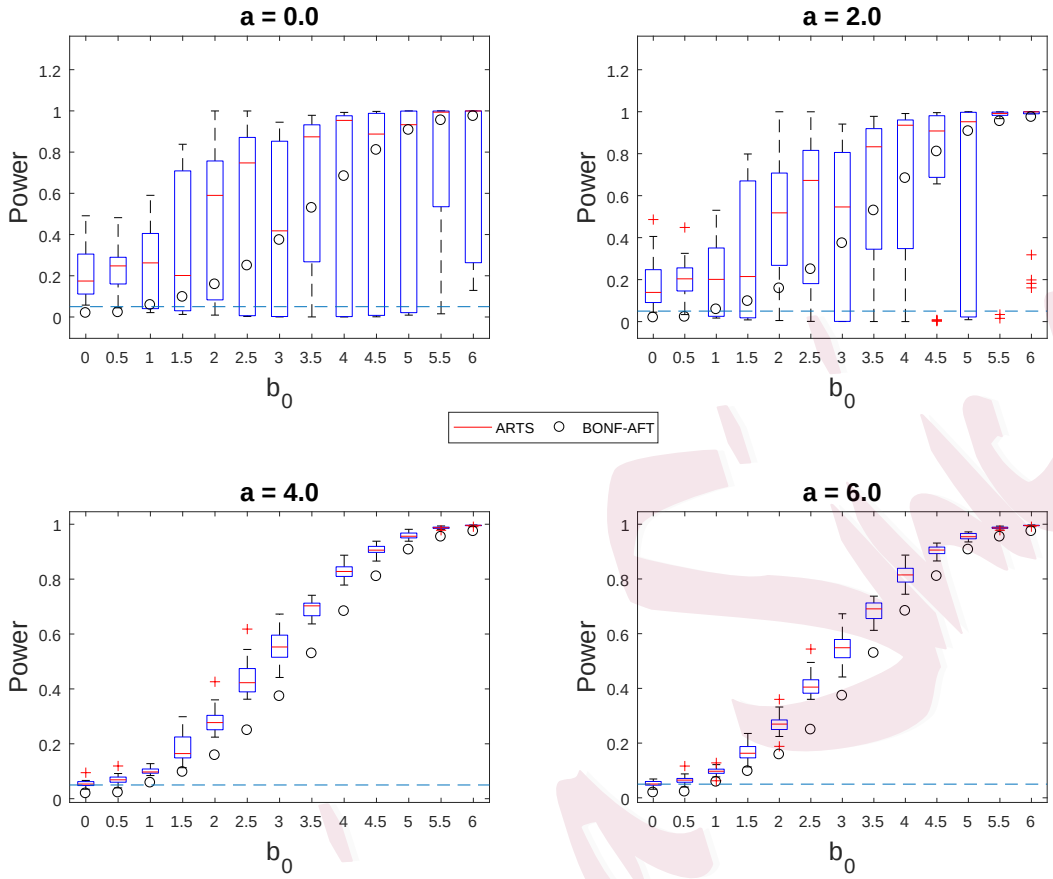

Figure 3: Asymptotic Type I error and power of ARTS compared with BONF-AFT for $p = 1000$ under light censoring, where ARTS is implemented with a fixed threshold λ_n specified by $a = \{0, 2, 4, 6\}$, and each boxplot is based on 20 independent replications with $n = 10,000$.

provided; here we omit the anti-conservative results of CPB-AFT for conciseness and pay more attention to CEND that requires the independence between ε and $\boldsymbol{U}$.

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 $T = \varepsilon$. 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 6 shows that only ARTS controls

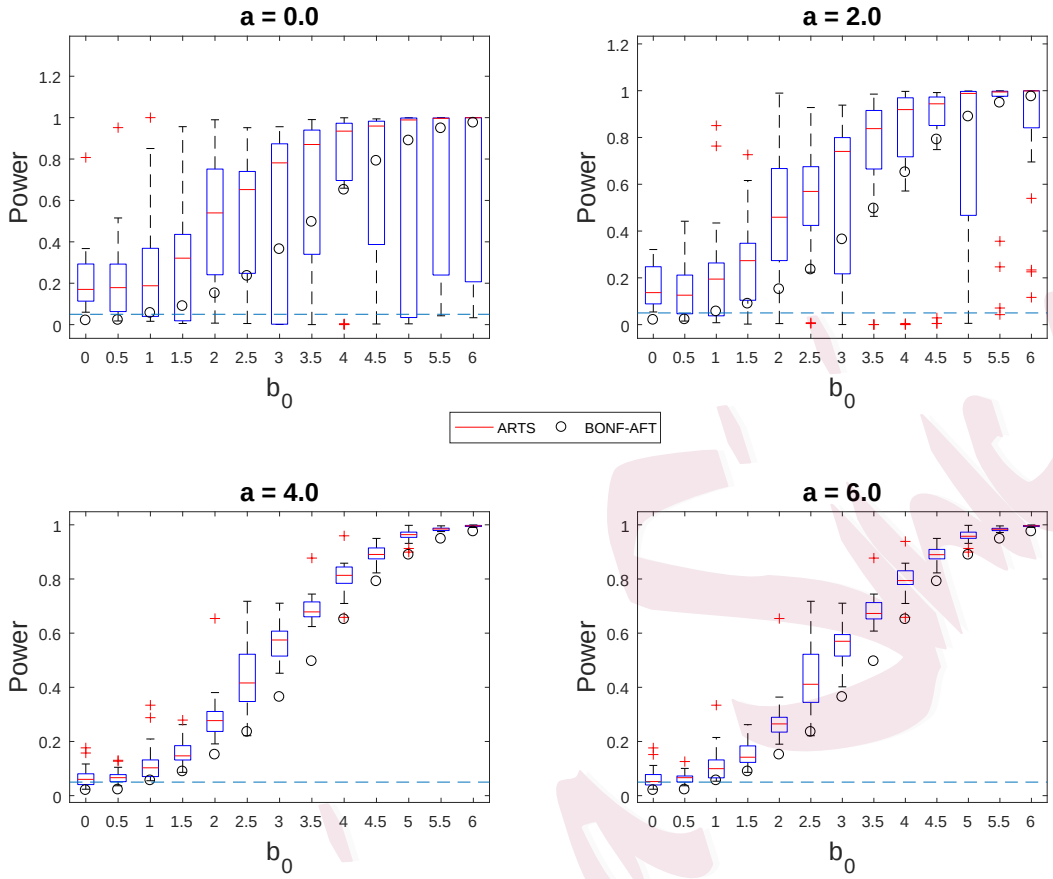

Figure 4: Asymptotic Type I error and power as in Figure 3 except under moderate censoring.

FWER around the nominal level in the case of dependent errors, except for giving slightly conservative FWERs when $p \geq 50$, heavy censoring and $n = 100$.

7 Applications to real data

7.1 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 with complete information on genetic predictors. The censoring rate is 43%. More details about the DLBCL data can be found in the literature (cf. Bøvelstad et al. (2009), Binder et al. (2011)). To adjust for the prognostic information provided by IPI, we apply adjusted ARTS

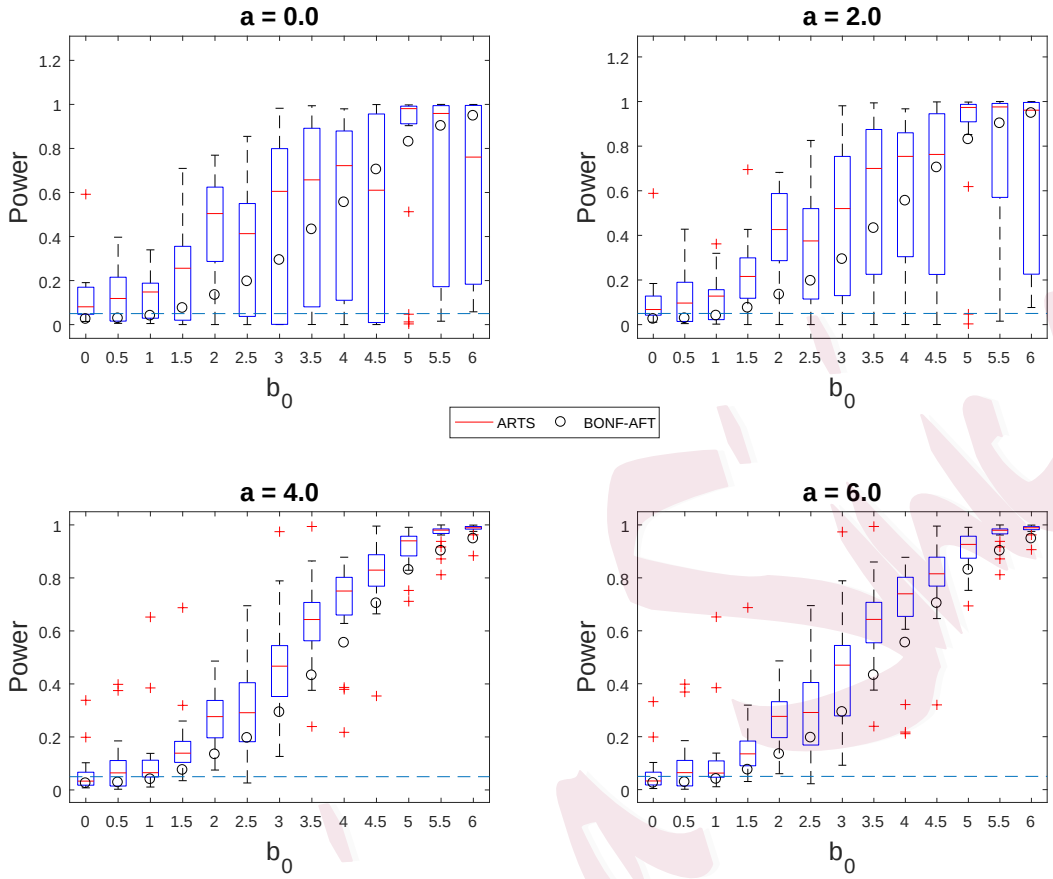

Figure 5: Asymptotic Type I error and power as in Figure 3 except under heavy censoring.

to this data set for detecting the presence of significant genetic features. To maintain the stability of the KSV estimator, the observed event times are restricted up to $\tau = 2.36$, which corresponds to the 90% empirical percentile of the observed event times. This excludes one observation whose value of the estimated synthetic response 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. Before implementing ARTS, one pre-processing step is taken to filter out the genes that lack significant differentiation between the censored group (patients still alive at the end of the follow-up) and the uncensored group (patients who died of DLBCL diseases within the follow-up). For each gene, a standard two-sample t-test is conducted to determine whether the gene expression measurement differentiates between these two groups. Comparing the corresponding p-values with

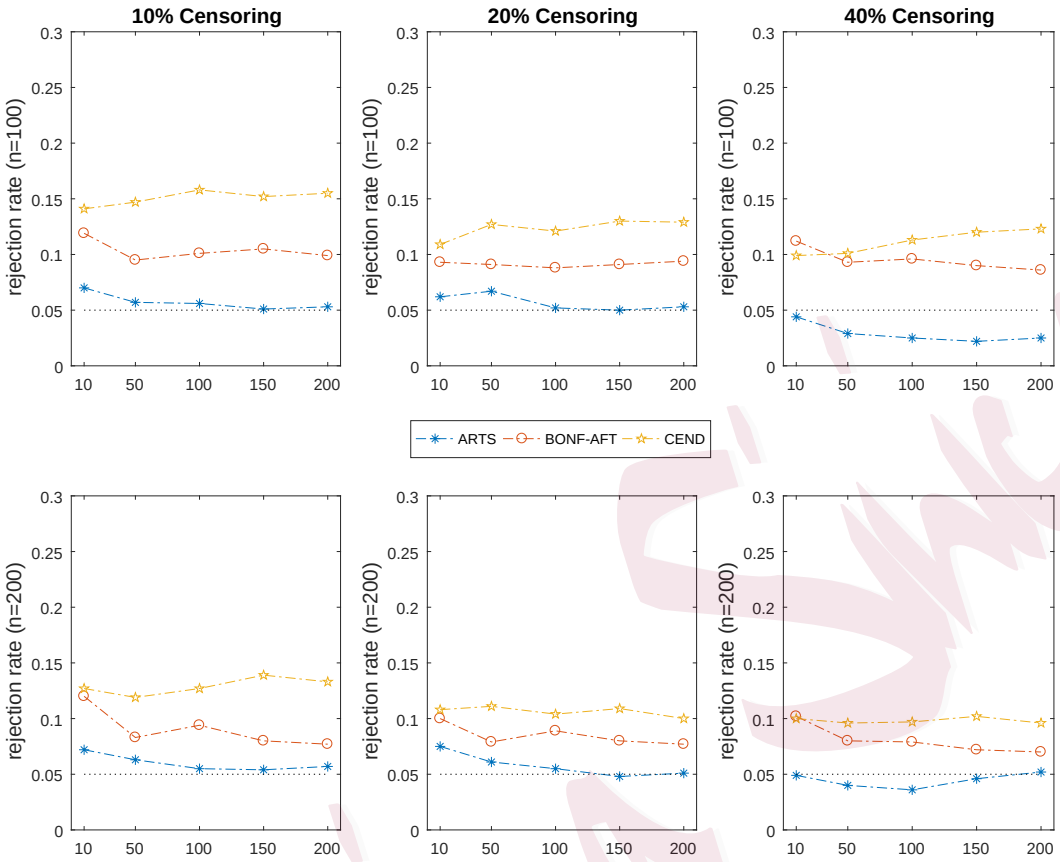

Figure 6: Empirical rejection rates based on 1000 samples generated from the null model with dependent errors under various p, sample sizes and censoring rates.

the nominal level of 5%, this pre-processing step reduces the number of screening genetic features to 1026 ($p = 1026$).

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 the required assumption of independence between the error and predictors. Also, the HC method is not considered since it is designed for the case of nearly uncorrelated predictors, which is unrealistic for gene expression data. The three implemented approaches yield similar p-values. The minimal Bonferroni corrected p-value from BONF-AFT is 4.39%. The ARTS procedure reduces to a special case with $\lambda_n = 0$ and gives the same p-value of 3.40% as CPB-AFT, from 1000 bootstrap samples.

Figure 7 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. Given the nominal level of 5%, these three approaches all indicate one significant gene for the survival time of patients. The ID of the detected gene is "27766" and it belongs to the group of the major histocompatibility class (MHC) II signature. This finding supports the notion that loss of MHC II expression correlates with a worse survival outcome, and responds to the results provided by Miller et al. (1988), Rosenwald et al. (2002), Rimsza et al. (2004), Roberts et al. (2006), and Higashi et al. (2016), among others.

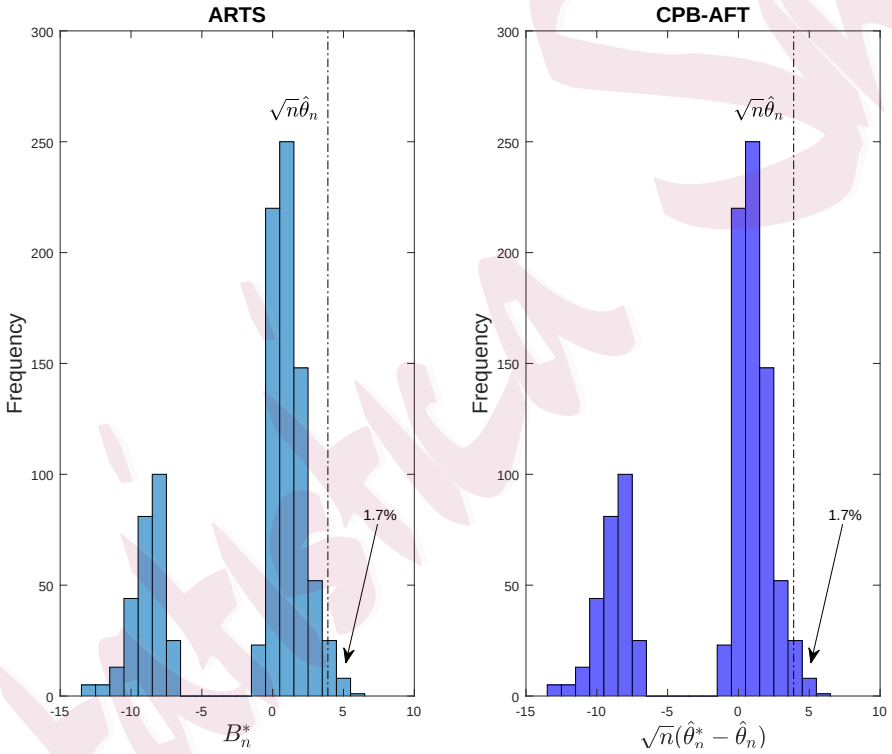

Figure 7: DLBCL example. Left panel: histogram of B_n^* giving the two-sided ARTS p-value 3.40%. Right panel: histogram of $\sqrt{n}(\hat{\theta}_n^* - \hat{\theta}_n)$ giving the two-sided CPB-AFT p-value 3.40%.

7.2 Primary biliary cirrhosis data

In this example, we demonstrate how to apply forward-stepwise ARTS to successively identify interaction effects, provided that the main effects of some covariates have been shown statistically or clinically significant. We use data from the Mayo Clinic trial in primary bil-

iary cirrhosis (PBC) of the liver conducted between 1974 and 1984 (Fleming and Harrington (1991), Appendix D.1). A total of 312 PBC patients participated in the randomized placebo controlled trial of the drug D-penicillamine; we restrict attention to the 276 patients who have complete covariate information in our data analysis. The censoring rate is 60%.

The survival outcome is the time from registration to death. Over the follow-up, there is no significant treatment effect (Fleming and Harrington (1991)). Only five of the 16 risk factors were found statistically significant under the setting of the Cox model (Dickson et al. (1989)) or under the AFT model (Jin et al. (2003)), and they were also identified as a subset of active predictors under the general 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, in seconds). Of these risk factors, serum bilirubin, albumin and protime are log-transformed. We are interested in successively locating significant pairwise interaction terms of 17 variables, adjusting for the five aforementioned risk factors. These 17 variables include the treatment indicator and 16 clinical risk factors for the survival time ($p = \binom{17}{2} = 136$).

Figure 8 displays the pattern of p-values for the newly entered interaction term at each step. The forward-stepwise ARTS procedure detects one significant interaction term, where the constant a and the end of the follow-up τ are selected in the same fashion as Section 7.1. This detected interaction is between platelet (platelets per cubic ml/1000) and alk.phos (alkaline phosphatase, in U/liter). For comparison, we also present successive p-values given by CPB-AFT. The conclusion remains the same, but the p-values of CPB-AFT are smaller, as expected.

To examine the effect of taking covariate-dependent censoring into account when applying ARTS in this example, we further run forward-stepwise ARTS as before, except replacing $\hat{G}_n$ by a Cox model based estimate conditional on selected covariates (alkaline phosphotase and log-transformed protime). In contrast to our earlier finding of one significant interaction

term, here we find none (results not shown). The CPB-AFT procedure (with the same Cox model estimate of G_0) also leads to the same conclusion.

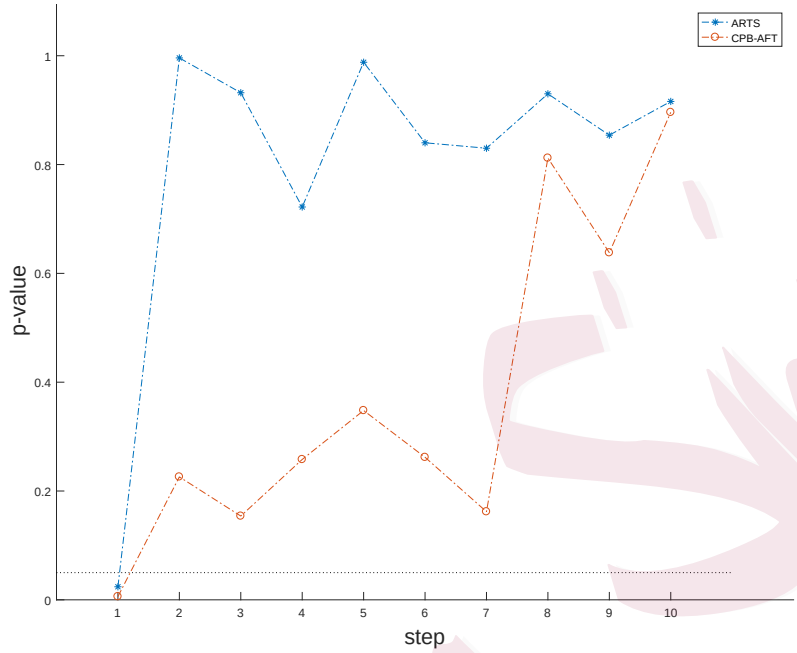

Figure 8: PBC example. The patterns of p-values for forward-stepwise ARTS and CPB-AFT.

8 Discussion

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 problem of 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 dealing with the increased dispersion introduced by the censoring. The advantage of ARTS lies in it providing a post-selection-corrected p-value without sacrificing power, while avoiding distributional assumptions, specific correlation structures between

predictors, and a preconceived choice of the regression parameters of interest. 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 further 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 for 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.

Supplementary Materials

The online supplementary materials include detailed proofs of theorems and additional simulation results.

Acknowledgements

This research was partially supported by NIH Grants R01GM095722; R21MH108999, and NSF Grant DMS-1307838. The authors thank the associate editor and reviewers for their helpful comments.

References

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

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

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **57**, 289–300.

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**, 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**, 3092–3120.

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

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

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

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

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

Dabrowska, D. (1989). Uniform consistency of the kernel conditional Kaplan–Meier estimate. *The Annals of Statistics*, **17**, 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**, 259–271.

Dickson, E. R., Grambsch, P. M., Fleming, T. R., Fisher, L. D., and Langworthy, A. (1989). Prognosis in primary biliary cirrhosis: model for decision making. *Hepatology*, **10**, 1–7.

Donoho, D. and Jin, J. (2004). Higher criticism for detecting sparse heterogeneous mixtures. *Annals of Statistics*, **32**, 962–994.

Donoho, D. and Jin, J. (2015). Higher criticism for large-scale inference, especially for rare and weak effects. *Statistical Science*, **30**, 1–25.

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**, 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**, 74–99.

Fang, E. X., Ning, Y., and Liu, H. (2016). Testing and confidence intervals for high dimensional proportional hazards models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) (in press)*.

Fleming, T. R. and Harrington, D. P. (1991). *Counting Processes and Survival Analysis*. John Wiley & Sons, Inc.

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**, 1950–1957.

Gorst-Rasmussen, A. and Scheike, T. (2013). Independent screening for single-index hazard rate models with ultra-high 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**, 342–369.

Higashi, M., Tokuhira, M., Fujino, S., Yamashita, T., Abe, K., Arai, E., Kizaki, M., and Tamaru, J.-I. (2016). Loss of HLA-DR expression is related to tumor microenvironment and predicts adverse outcome in diffuse large B-cell lymphoma. *Leukemia & Lymphoma*, **57**, 161–166.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, **6**, 65–70.

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

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

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

Huang, J., Ma, S., and Xie, H. (2006). Regularized estimation in the accelerated failure time model with high-dimensional covariates. *Biometrics*, **62**, 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**, 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**, 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**, 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**, 215–224.

Koul, H., Susarla, V., and Van Ryzin, J. (1981). Regression analysis with randomly right-censored data. *The Annals of Statistics*, **9**, 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**, 1370–1402.

Lai, T. L. and Ying, Z. (1991b). Rank regression methods for left-truncated and right-censored data. *The Annals of Statistics*, **19**, 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**, 600–614.

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

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

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

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

Luedtke, A. R. and van der Laan, M. J. (2017). Parametric-rate inference for one-sided differentiable parameters. *Journal of American Statistical Association (in press)*.

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

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

McKeague, I. W. and Sasieni, P. D. (1994). A partly parametric additive risk model. *Biometrika*, **81**, 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**, 121–131.

Miller, T. P., Lippman, S. M., Spier, C. M., Slymen, D. J., and Grogan, T. M. (1988). HLA-DR (Ia) immune phenotype predicts outcome for patients with diffuse large cell lymphoma. *The Journal of Clinical investigation*, **82**, 370–372.

Pollard, D. (1990). *Empirical Processes: Theory and Applications*. NSF-CBMS Regional Conference Series in Probability and Statistics: Vol 2. Institute of Mathematical Statistics.

Rimsza, L. M., Roberts, R. A., Miller, T. P., Unger, J. M., LeBlanc, M., Braziel, R. M., Weisenberger, D. D., Chan, W. C., Muller-Hermelink, H. K., Jaffe, E. S., Gascoyne, R. D., Campo, E., Fuchs, D. A., Spier, C. M., Fisher, R. I., Delabie, J., Rosenwald, A., Staudt, L. M., and Grogan, T. M. (2004). Loss of MHC class II gene and protein expression in diffuse large B-cell lymphoma is related to decreased tumor immunosurveillance and poor patient survival regardless of other prognostic factors: a follow-up study from the leukemia and lymphoma molecular profiling project. *Blood*, **103**, 4251–4258.

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

Roberts, R. A., Wright, G., Rosenwald, A. R., Jaramillo, M. A., Grogan, T. M., Miller, T. P., Frutiger, Y., Chan, W. C., Gascoyne, R. D., Ott, G., Muller-Hermelink, H. K.,

Staudt, L. M., and Rimsza, L. M. (2006). Loss of major histocompatibility class II gene and protein expression in primary mediastinal large B-cell lymphoma is highly coordinated and related to poor patient survival. *Blood*, **108**, 311–318.

Rosenwald, A., Wright, G., Chan, W. C., Connors, J. M., Campo, E., Fisher, R. I., Gascoyne, R. D., Muller-Hermelink, H. K., Smeland, E. B., Giltnane, J. M., Hurt, E. M., Zhao, H., Averett, L., Yang, L., Wilson, W. H., Jaffe, E. S., Simon, R., Klausner, R. D., Powell, J., Duffey, P. L., Longo, D. L., Greiner, T. C., Weisenburger, D. D., Sanger, W. G., Dave, B. J., Lynch, J. C., Vose, J., Armitage, J. O., Montserrat, E., Lpez-Guillermo, A., Grogan, T. M., Miller, T. P., LeBlanc, M., Ott, G., Kvaloy, S., Delabie, J., Holte, H., Krajci, P., Stokke, T., and Staudt, L. M. (2002). The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma. *The New England journal of medicine*, **346**, 1937–1947.

Sinnott, J. A. and Cai, T. (2016). Inference for survival prediction under the regularized Cox model. *Biostatistics*, **17**, 692–707.

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

Srinivasan, C. and Zhou, M. (1994). Linear regression with censoring. *Journal of Multivariate Analysis*, **49**, 179–201.

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

Taylor, J. and Tibshirani, R. (2017). Post-selection inference for ℓ_1-penalized likelihood models. *Canadian Journal of Statistics (in press)*.

The International Non-Hodgkin's Lymphoma Prognostic Factors Project (1993). A predictive model for aggressive non-Hodgkin's lymphoma. *New England Journal of Medicine*, **329**, 987–994.

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

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

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

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**, 76–99.

Zhang, H. H. and Lu, W. (2007). Adaptive Lasso for Cox's proportional hazards model. *Biometrika*, **94**, 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**, 397–411.

Zhao, S. D. and Li, Y. (2014). Score test variable screening. *Biometrics*, **70**, 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**, 649–664.

Zhou, M. (1992). Asymptotic normality of the 'synthetic data' regression estimator for censored survival data. *The Annals of Statistics*, **20**, 1002–1021.

Department of Biostatistics, Columbia University

E-mail: th2455@caa.columbia.edu

Phone: (212) 342-1242

Fax: (212) 305-9408

Department of Biostatistics, Columbia University

E-mail: im2131@cumc.columbia.edu

Department of Biostatistics, Columbia University

E-mail: mq2158@cumc.columbia.edu