Statistica Sinica Preprint No: SS-2017-0373	
Title	Joint variable screening in accelerated failure time model
Manuscript ID	SS-2017-0373
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202017.0373
Complete List of Authors	Yixin Fang and Jinfeng Xu
Corresponding Author	Jinfeng Xu
E-mail	xujf@hku.hk

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0373

JOINT VARIABLE SCREENING IN ACCELERATED FAILURE TIME MODELS

Yixin Fang[†] and Jinfeng Xu[‡,*]

[†] *AbbVie and* [‡] *The University of Hong Kong*

Abstract: Variable screening has become increasingly popular as a method for analyzing high-dimensional survival data. Most existing variable screening methods for such data assess the importance of their variables using marginal models that relate the time-to-event outcome to each variable separately. This implies that the relevance of one variable is examined while other variables are excluded. Therefore, these methods exclude variables that only manifest their influence jointly, and may retain irrelevant variables that are correlated with relevant ones. To circumvent these difficulties, we propose a new approach to evaluate the joint variable importance in censored accelerated failure time models. We establish the sure screening properties of the proposed approach and demonstrate its effectiveness through simulation studies and a real-data application. Furthermore, we propose a novel procedure using stability selection for tuning.

Key words and phrases: Inverse probability weighting, stability selection, survival data, ultrahigh dimensional covariates, variable importance.

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0373

1. Introduction

In many biomedical studies, such as high-throughput microarray or RNA-sequencing (RNA-seq) gene expression analyses, it is of practical interest to link gene expression profiles to censored survival phenotypes, such as the time to cancer recurrence or time to death. Because the number of genes greatly exceeds the sample size and the expression levels of some genes are often highly correlated, it is challenging to build a model that predicts the survival outcomes of future patients. In addition, the nuances of survival data such as right censoring and semiparametric modeling, make the task more challenging still. Therefore, it is important that we reduce the number of covariates when modeling survival data with ultrahigh-dimensional covariates.

Let T denote the survival time, $\boldsymbol{X} = (X_1, \cdots, X_{p_n})^{\mathsf{T}}$ be a covariate vector, and C be the censoring time. Define $\Delta = I(T \leqslant C)$, where $I(\cdot)$ is the indicator function. The observed data are independent and identically distributed (i.i.d.) copies of $(T \wedge C, \Delta, \boldsymbol{X})$, denoted by $(T_i \wedge C_i, \Delta_i, \boldsymbol{X}_i)$, for $i = 1, \cdots, n$. A semiparametric regression model relating the survival time T to the covariates $\boldsymbol{X}$ can be formulated as

$$T = g(\boldsymbol{\beta}_\star^{\mathsf{T}} \boldsymbol{X}, \varepsilon), \tag{1.1}$$

where ε is the mean-zero residual and $\boldsymbol{\beta}_\star = (\beta_{\star 1}, \cdots, \beta_{\star p_n})^{\mathsf{T}}$ is a vector of

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0373

the regression coefficients.

For specific choices of g or distributions of ε, model (1.1) leads to many useful survival models, such as the Cox's proportional hazards (PH) model (Cox, 1972) and the accelerated failure time (AFT) model (Jin et al., 2003; Kalbfleisch and Prentice, 2002, pp.218–219). When p_n is much larger than n, we assume that only a few of the covariates are truly relevant. Thus, we identify the following set of the active covariates:

$$\mathcal{M}_\star = \{j : \beta_{\star j} \neq 0,\ 1 \leqslant j \leqslant p_n\}\,. \tag{1.2}$$

To identify $\mathcal{M}_\star$, existing methods include the penalization methods for survival models, such as Cox's PH model or the AFT model (Tibshirani, 1997; Huang, Ma, and Xie, 2006; Zhang and Lu, 2007). Under a general design condition, lasso-type penalization methods are not selection-consistent (Zhao and Yu, 2006). Although adaptive lasso methods (Zou, 2006; Zhang and Lu, 2007) are selection-consistent, they require that the sample size be larger than the number of covariates. Therefore, under a ultrahigh-dimensional setting, where the number of covariates grows exponentially with the sample size, screening methods (Fan and Lv, 2008) are more appropriate and have emerged as an important tool.

There is a rich body of literature on screening methods for survival data with ultrahigh-dimensional covariates. However, most existing methods

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0373

evaluate the importance of their variables using separate marginal regression models. The partial likelihood ratio (PL) screening method, proposed by Fan, Feng, and Wu (2010), is based on a marginal Cox PH model. The feature aberration at survival times (FAST) screening method, proposed by Gorst-Rasmussen and Scheike (2013), is based on a marginal additive hazards model (Lin and Ying, 1994). The censored rank independence (CR) screening method, proposed by Song et al. (2014), is based on a marginal transformation model (Cheng, Wei, and Ying, 1995). These marginal screening methods examine the relevance of one variable, while excluding other variables. Thus, these methods fail to identify covariates that only manifest their influence jointly, and could retain irrelevant covariates that are correlated with relevant ones.

To overcome the drawbacks of existing marginal screening methods, Hong, Kang, and Li (2018) recently proposed a conditional screening approach for survival data with ultrahigh-dimensional covariates. However, their approach requires pre-selectioning a set of covariates. Furthermore, the computational burden of the approach is heavy because it requires fitting a multivariate survival model for each covariate that is not pre-selected. Therefore, we propose a joint screening approach for the AFT model that overcomes the drawbacks of current marginal screening approaches by mod-

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0373

eling all covariates jointly. In addition, it overcomes the drawbacks of the conditional screening approach by conducting only one model fitting.

Here, we consider the variable screening problem for the following AFT model:

$$\log(T_i) = \boldsymbol{X}_i^{\mathsf{T}}\boldsymbol{\beta}_\star + \varepsilon_i. \tag{1.3}$$

Let $Y_i = \log(T_i)$. Assume that $(Y_i, \boldsymbol{X}_i, \varepsilon_i)$, for $i = 1, \cdots, n$, are i.i.d. copies of $(Y, \boldsymbol{X}, \varepsilon)$. By relating the logarithm of the failure time linearly to the covariates, the AFT model provides an attractive alternative to the popular Cox's PH, owing to its direct physical interpretation and fast computation (Wei, 1992; Jin, Lin, and Ying, 2006).

The rest of the paper is organized as follows. In Section 2, we describe the proposed joint screening approach. In Section 3, we establish the sure screening properties of the approach under certain regularity conditions. In Section 4, we develop a novel stability selection-based bootstrap procedure for tuning the size of the estimated active set. We evaluate the proposed approach using simulation studies in Section 5 and by applying it to data on adult acute myeloid leukemia in Section 6. We conclude the paper in Section 7. All technical proofs are relegated to the online Supplementary Material.

2. The Joint Screening Method

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0373

The censoring time C_i is assumed to be independent of T_i, given the covariates $\boldsymbol{X}_i$. For ease of exposition, we assume that the censoring distribution is the same for all covariates; however, this assumption can be relaxed, as discussed in He, Wang, and Hong (2013, pp.349).

We consider the AFT model given in (1.3). Because the values of Y_i associated with $\Delta_i = 0$ are unknown, we use the inverse probability of censoring (IPC) weighting procedure (Ying, Jung, and Wei, 1995; Peng and Fine, 2009; Song et al., 2014) to impute Y_i; that is, $\widehat{Y}_i = \Delta_i Y_i / \widehat{G}(Y_i)$, where $G(t) = P(\log(C) \geqslant t)$ and $\widehat{G}(t)$ denotes the Kaplan–Meier estimator of $G(t)$. Then, the resulting least squares estimator of $\boldsymbol{\beta}_\star$ is obtained by minimizing

$$n^{-1} \sum_{i=1}^{n} \left(\widehat{Y}_i - \boldsymbol{\beta}^\mathsf{T} \boldsymbol{X}_i \right)^2 . \tag{2.1}$$

Let $\mathbf{X} = (\boldsymbol{X}_1^\mathsf{T}, \cdots, \boldsymbol{X}_n^\mathsf{T})^\mathsf{T}$ denote the design matrix and $\widehat{\boldsymbol{Y}} = (\widehat{Y}_1, \cdots, \widehat{Y}_n)^\mathsf{T}$. When the $p_n \times p_n$ matrix $\mathbf{X}^\mathsf{T}\mathbf{X}$ is invertible, the minimizer of (2.1) is

$$\breve{\boldsymbol{\beta}} = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\widehat{\boldsymbol{Y}}. \tag{2.2}$$

However, this is not applicable for screening problems in which p_n is much larger than n.

Motivated by the recent development of the high-dimensional ordinary least squares projection (HOLP) approach by Wang and Leng (2016), we

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0373

propose the following IPC-weighted projection estimator:

$$\widehat{\boldsymbol{\beta}} = \mathbf{X}^{\mathsf{T}}(\mathbf{X}\mathbf{X}^{\mathsf{T}})^{-1}\widehat{\boldsymbol{Y}}. \tag{2.3}$$

Note that the middle matrix in (2.3), $\mathbf{X}\mathbf{X}^{\mathsf{T}}$, is an invertible $n \times n$ matrix that can be computed easily for $p_n \gg n$. Letting $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_1, \cdots, \widehat{\beta}_{p_n})^{\mathsf{T}}$, we define the following two sets of variable indices:

$$\mathcal{M}_{\gamma_n} = \{j : |\widehat{\beta}_j| \geqslant \gamma_n, 1 \leqslant j \leqslant p_n\}, \tag{2.4}$$

where γ_n is the hard threshold to be determined, and

$$\mathcal{M}_{d_n} = \{j : |\widehat{\beta}_j| \text{ is among the largest } d_n \text{ of all } \widehat{\beta}_j\}, \tag{2.5}$$

where d_n is the size of the selected subset to be determined.

In the next section, we show that the magnitudes of $\min_{j\in\mathcal{M}_\star} |\widehat{\beta}_j|$ and $\max_{j\notin\mathcal{M}_\star} |\widehat{\beta}_j|$ can be separated. Thus, if the tuning parameter γ_n or d_n is selected appropriately, it is reasonable to use $\mathcal{M}_{\gamma_n}$ or $\mathcal{M}_{d_n}$, respectively, to estimate $\mathcal{M}_\star$. The proposed joint screening approach has computational complexity of $O(n^2 p_n)$, whereas that of the independence screening approach (Fan, Feng, and Wu, 2010) is $O(np_n)$. For the screening problem where $p_n \gg n$, the proposed joint screening approach is numerically fast, with only a slightly larger computational complexity than that of the independence screening approach.

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0373

The innovation of the proposed approach is threefold. First, it enables variable screening for survival data with ultrahigh-dimensional covariates based on the AFT model. There are existing screening approaches based on Cox's PH model, the additive hazard model, and the transformation model, but no screening approach has been proposed previously based on the AFT model. Second, it is the first joint screening approach for survival data with ultrahigh-dimensional covariates. There are existing marginal screening approaches and a conditional approach for survival data, but there are no existing joint screening approaches for survival data. Third, rather than being a simple application of the HOLP approach to a special case, the proposed projection estimator (2.3) is a smart application of the IPC procedure that makes it possible to utilize the HOLP concept. Note that including the IPC procedure makes the theoretical work of the sure consistency more challenging.

3. The Sure Screening Properties

Although the proposed IPC-weighted joint screening approach is numerically appealing, it must be able to separate relevant from irrelevant variables with probability tending to one, as the sample size increases. Therefore, in this section, we establish the screening properties of the proposed method.

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0373

Without loss of generality, assume that X_j, for $j = 1, \cdots, p_n$, have mean zero and variance one. Let $\mathrm{Cov}(\boldsymbol{X}) = \Sigma$. Define $\mathbf{Z}$ and $\boldsymbol{Z}$ as $\mathbf{Z} = \mathbf{X}\Sigma^{-1/2}$ and $\boldsymbol{Z} = \Sigma^{-1/2}\boldsymbol{X}$, respectively. Note that $\mathbf{X}$ and $\mathbf{Z}$ are $p_n \times p_n$ matrices, and that $\boldsymbol{X}$ and $\boldsymbol{Z}$ are p_n-dim vectors. The tail behavior of the random error ε has a significant impact on the screening performance. We present the following tail condition to characterize the tail behavior of different distribution families, as in Vershynin (2010).

Definition 1. *A zero-mean distribution F is said to have a q-exponential tail if any $K \geqslant 1$ independent random variables $\epsilon_i \sim F$ satisfy that, for any m constants a_i, with $\sum_{i=1}^{K} a_i^2 = 1$, the following inequality holds:*

$$P\left(\left|\sum_{i=1}^{K} a_i \epsilon_i\right| > t\right) \leqslant \exp(1 - q(t)),$$

for any $t > 0$ and some function $q(\cdot)$.

This characterization of the tail behavior is very general. As shown in Vershynin (2010), $q(t) = O(t^2/M^2)$ for some constant M depending on F if F is sub-Gaussian, including Gaussian, Bernoulli, and any bound random variables. In addition, we have $q(t) = O(\min\{t/M, t^2/M^2\})$ if F is sub-exponential, including the exponential, Poisson, and χ^2 distributions. Moreover, as shown in Zhao and Yu (2006), $q(t) = 2m \log t + O(1)$ if F has bounded $2m^{\text{th}}$ moments, for some positive integer m.

Throughout this paper, $\lambda_{\max}(\cdot)$ and $\lambda_{\min}(\cdot)$ denote the largest and s-

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0373

mallest eigenvalues, respectively, of a matrix and m, M, m_i, and M_i denote absolute constants independent of n and p_n. We make the following five theoretical assumptions.

A1. The transformed $\boldsymbol{Z}$ has a spherically symmetric distribution, and there exist some $M_1 > 0$ and $m_1 > 1$, such that

$$P\left(\lambda_{\max}(\mathbf{Z}\mathbf{Z}^{\mathsf{T}}/p_n) > m_1 \text{ or } \lambda_{\min}(\mathbf{Z}\mathbf{Z}^{\mathsf{T}}/p_n) < 1/m_1\right) \leqslant \exp(-M_1 n).$$

Assume $p_n > m_0 n$, for some $m_0 > 1$.

A2. Let $\epsilon = \Delta\varepsilon/G(Y)$, which has mean zero and standard deviation σ. Given $\boldsymbol{X} = \boldsymbol{x}$, the standardized error ϵ/σ has a q-exponential tail, with $q(t)$ independent of $\boldsymbol{x}$, as defined in Definition 1.

A3. For some $\kappa \geqslant 0$, $\nu \geqslant 0$, $\tau \geqslant 0$, $m_2 > 0$, $m_3 > 0$, and $m_4 > 0$,

$$\min_{j \in \mathcal{M}_\star} |\beta_{\star j}| \geqslant m_2/n^\kappa,\ s_n = |\mathcal{M}_\star| \leqslant m_3 n^\nu, \text{ and } \lambda_{\max}(\Sigma)/\lambda_{\min}(\Sigma) \leqslant m_4 n^\tau.$$

A4. There exists $C_{\max} > 0$, such that $\delta_1 = P(C = C_{\max}) > 0$ and $P(C > C_{\max}) = 0$. Assume that $0 < \delta_2 = P(T < C_{\max}) < 1$.

A5. Assume $\mathrm{Var}(Y_i) = O(1)$. Let M_2 be the sub-exponential norm of $\boldsymbol{\beta}_\star^{\mathsf{T}}\boldsymbol{X}$, and assume that $M_2 < \infty$.

Remark 1: Assumptions A1–A3 are similar to A1–A3 in Wang and Leng (2016). A4 is a technical condition adopted from Peng and Fine (2009) and

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0373

Song et al. (2014) to simplify the asymptotic arguments, ensuring that the IPC-weights, $\Delta_i/G(Y_i)$, are bounded by $1/\delta_1$. This is because $0 \leqslant \Delta_i/G(Y_i) \leqslant 1/G(Y_i \wedge C_i) \leqslant 1/G(C_{\max}) = 1/\delta_1$. A5 controls the tail behavior of the linear predictor. An example in which all assumptions are satisfied is $\boldsymbol{X} \sim N(\mathbf{0}, \sigma_x^2 \mathbf{I}_{p_n})$, $\varepsilon \sim N(0, \sigma_e^2)$, C is exponentially distributed, but truncated by a constant $C_{\max}$, and $s_n = s$.

Remark 2: Assumption A1 is similar to, but weaker than the concentration property in Fan and Lv (2008) and Fan, Feng, and Wu (2010). The latter require that all submatrices of $\mathbf{Z}$ consisting of more than $O(n)$ rows satisfy the eigenvalue concentration inequality, whereas A1 requires only that $\mathbf{Z}$ itself holds. As implied by the results in Section 5.4 of Vershynin (2010), the concentration inequality in A1 holds if $\boldsymbol{X}$ is sub-Gaussian.

Remark 3: The existence of the conditional q-exponential tail function $q(t)$ that is independent of $\boldsymbol{X} = \boldsymbol{x}$ in A2 is also ensured because $\Delta_i/G(Y_i)$ is bounded. For example, if ε is sub-Gaussian with norm $\|\varepsilon\|_{\psi_2} = M$, then, conditional on $\boldsymbol{X} = \boldsymbol{x}$, the sub-Gaussian norm of ϵ is $\|\Delta\varepsilon/G(Y)\|_{\psi_2} \leqslant M/\delta_1$. Therefore, we can consider $q(t) = O(\delta_1 t^2/M^2)$ that is independent of $\boldsymbol{x}$. Similarly, if ε is sub-exponential with norm $\|\varepsilon\|_{\psi_1} = M$, then we consider $q(t) = O(\min\{\delta_1 t/M, \delta_1^2 t^2/M^2\})$ that is independent of $\boldsymbol{x}$.

Remark 4: The assumption on $\min_{j \in \mathcal{M}_\star} |\beta_j|$ in A3 is the key differ-

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0373

ence between the proposed joint screening approach and existing marginal screening approaches. All existing approaches require some marginal correlation condition, which states that the covariates are active if and only if they are marginally relevant to the outcome; for example, see Condition 2 in Song et al. (2014). As pointed out by Fan and Lv (2008), this can be violated easily if the covariates are correlated.

We first show that, asymptotically, the screening based on the projection estimator retains all active covariates with probability tending to one.

Theorem 1. *Under Assumptions A1–A5, if* γ_n *is chosen such that* $p_n\gamma_n/n^{1-\tau-\kappa} \to 0$ *and* $p_n\gamma_n\sqrt{\log n}/n^{1-\tau-\kappa} \to \infty$, *then*

$$P\left(\mathcal{M}_\star \subset \mathcal{M}_{\gamma_n}\right) = 1 - O\left\{\exp\left(\frac{-Mn^{1-5\tau-2\kappa-\nu}}{\log n}\right) + \varpi(n)\right\},$$

where

$$\varpi(n) = 2s_n \exp\left\{1 - q\left(\frac{\sqrt{M}n^{1/2-2\tau-\kappa}}{\sqrt{\log n}}\right)\right\} + O\left\{\exp\left(-Mn^{1/4-\tau-\kappa/2}\right)\right\}.$$

In Theorem 1, we do not make any assumption on p_n, as long as $p_n > m_0 n$. With further mild conditions on p_n that still allow for the ultrahigh-dimensionality of p_n, we derive the following screening consistency property.

Theorem 2. *Under Assumptions A1–A5, if* p_n *satisfies*

$$\log p_n = o\left(\min\left\{\frac{n^{1-5\tau-2\kappa-\nu}}{\log n}, n^{1/4-\tau-\kappa/2}, q\left(\frac{\sqrt{M}n^{1/2-2\tau-\kappa}}{\sqrt{\log n}}\right)\right\}\right),$$

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0373

then, for the same γ_n as in Theorem 1, we have

$$P\left(\min_{j\in\mathcal{M}_\star}|\widehat{\beta}_j| > \gamma_n > \max_{j\notin\mathcal{M}_\star}|\widehat{\beta}_j|\right) = 1 - O\left\{\exp\left(-\frac{Mn^{1-5\tau-2\kappa-\nu}}{\log n}\right) + \varpi(n)\right\},$$

where $\varpi(n)$ is defined as in Theorem 1. Alternatively, we can choose a submodel $\mathcal{M}_{d_n}$ with $d_n \asymp n^\iota$, for some $\iota \in (\nu, 1]$, such that

$$P\left(\mathcal{M}_\star \subset \mathcal{M}_{d_n}\right) = 1 - O\left\{\exp\left(-\frac{Mn^{1-5\tau-2\kappa-\nu}}{\log n}\right) + \varpi(n)\right\}.$$

4. The Stability Selection Procedure for Tuning

The first part of Theorem 2 in Section 3 shows that, asymptotically, the magnitudes of the projection estimators for the important and unimportant covariates are separable with probability tending to one. This ensures the screening consistency. We can visualize such separability using the bootstrap method. First, we generate B bootstrap samples, $\mathcal{D}^{(b)}$, for $b = 1, \cdots, B$. From bootstrap sample $\mathcal{D}^{(b)}$, we obtain estimate $\widehat{\boldsymbol{\beta}}^{(b)}$ using (2.3). At each threshold $\gamma_n = \gamma$, we calculate the proportion of instances in which covariate j is selected as an important covariate from among B bootstrap samples,

$$\widehat{\Pi}_j(\gamma) = \frac{1}{B}\sum_{b=1}^{B} I\left\{|\widehat{\beta}_j^{(b)}| > \gamma\right\}. \tag{4.1}$$

We can display these curves, $\widehat{\Pi}_j(\gamma)$, for $j = 1, \cdots, p_n$, against γ on the same graph and identify the outstanding curves. We demonstrate this using a randomly selected simulated data set in the next section (see Figure 1).

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0373

The second part of Theorem 2 states that as long as we choose a submodel with a dimension larger than that of the true model, we are guaranteed to retain all active covariates with probability tending to one. If we choose $d_n = s_n$, then the proposed screening selects the true model with an overwhelming probability. In practice, it is important to determine d_n. As suggested by Fan and Lv (2008), with the high dimension reduced accurately to below the sample size, say $d_n = \lfloor n/\log n \rfloor$, where $\lfloor x \rfloor$ denotes the integer part of x, variable selection can be accomplished using a refined lower-dimensional method, such as the SCAD (Fan and Li, 2001), Dantzig selector (Candes and Tao, 2007), lasso Tibshirani (1996), or adaptive lasso Zou (2006). However, this is a two-stage strategy, with the first stage being the screening, and the second stage being the variable selection. Given the large volume of literature on variable selection and our focus on screening, we do not explore the performance of such two-stage approaches.

Instead, motivated by the stability selection of Bach (2008), who proposed the bootstrap-boosted lasso method, we propose a novel bootstrap procedure for tuning the size of the estimated active set. To this end, as suggested by Fan and Lv (2008), we also consider a submodel $\mathcal{M}_{d_n}$ of size $d_n = \lfloor n/\log n \rfloor$. Specifically, based on the bth bootstrap sample, we obtain estimate $\widehat{\boldsymbol{\beta}}^{(b)}$ using (2.3). Then we obtain the bth bootstrap submodel $\mathcal{M}_d^{(b)}$

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0373

with the same size $d_n = d$. Finally, we consider the intersection of these bootstrap submodels; that is,

$$\overline{\mathcal{M}}_d = \bigcap_{b=1}^{B} \mathcal{M}_d^{(b)}. \tag{4.2}$$

Based on Theorem 2, we have $\pi(n) = 1 - P\left(\mathcal{M}_\star \subset \mathcal{M}_{d_n}\right) \to 0$, if $s_n < d_n$. The following propositions show that, as long as $B \to \infty$ and $B\pi(n) \to 0$, we have $P(\overline{\mathcal{M}}_{d_n} = \mathcal{M}_\star) \to 1$.

Proposition 1. *Under the assumptions of Theorem 2, if $s_n < d$, then*

$$P(\mathcal{M}_\star \subset \overline{\mathcal{M}}_d) > 1 - B\pi(n),$$

where $\pi(n) = 1 - P\left(\mathcal{M}_\star \subset \mathcal{M}_d^{(b)}\right)$.

Proposition 2. *Denote the observed data set as $\mathcal{D}$. If $P(j \in \mathcal{M}_d^{(b)} | \mathcal{D}) = P(h \in \mathcal{M}_d^{(b)} | \mathcal{D})$, for any $j, h \notin \mathcal{M}_\star$, then*

$$P(\mathcal{M}_\star \subsetneq \overline{\mathcal{M}}_d) < (p_n - s_n)\left(\frac{d}{p_n - s_n}\right)^B.$$

We see that B is a hyperparameter and the choice of B is crucial. By Proposition 1, if B is too big, the procedure may miss relevant covariates in $\mathcal{M}_\star$. By Proposition 2, if B is too small, the procedure may include irrelevant covariates not in $\mathcal{M}_\star$. In our numerical studies, we use $B = 100$. In practice, we can use an outer-loop cross-validation to determine an appropriate value for B.

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0373

5. Simulation Studies

We conduct simulations to investigate the performance of the proposed inverse probability of the censoring-weighted projection screening procedure (JS). For comparison, we consider three alternative methods: FAST screening Gorst-Rasmussen and Scheike (2013), partial likelihood ratio screening (PL) Fan, Feng, and Wu (2010) based on marginal Cox PH models, and CR independence screening Song et al. (2014).

We consider three scenarios that accommodate a variety of correlation structures among the covariates and model parameters. We generate data from model (1.3), with $\varepsilon \sim N(0, \sigma^2)$. In all simulations, the censoring times are generated from a uniform distribution to yield a censoring proportion of 20% or 40%. For simplicity, denote p_n as p.

Scenario 1: $X_1, \cdots, X_p$ are multivariate normal, where $X_j \sim N(0,1)$ and $\text{cov}(X_j, X_h) = \rho$, for any j and h. Set $\mathcal{M}_\star = \{1,2,3,4,5\}$, with $\boldsymbol{\beta}_{\mathcal{M}_\star} = (5,5,5,5,-20\rho)^\mathsf{T}$. Set $\rho = 0.5$, $n = 200$ or 400, $p = 2000$, and $\sigma = 1$.

Scenario 2: $X_1, \cdots, X_p$ are multivariate normal, where $X_j \sim N(0,1)$ and $\text{cov}(X_j, X_h) = 0.1$, for any j and h. Set $\mathcal{M}_\star = \{1:15\}$, with $\boldsymbol{\beta}_{\mathcal{M}_\star} = (\mathbf{1}_{14}^\mathsf{T}, -1.4)^\mathsf{T}$.

Scenario 3: $X_1, \cdots, X_p$ are multivariate normal, where $X_j \sim N(0,1)$ and $\text{cov}(X_j, X_h) = 0.9^{|j-h|}$, for any j and h. Set $\mathcal{M}_\star = \{1,2,3,4,5\}$, with

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0373

$\boldsymbol{\beta}_{\mathcal{M}_\star} = (\mathbf{1}_4^\mathsf{T}, -3.09)^\mathsf{T}$.

In Scenario 1, a small number of nonzero regression coefficients have large effect sizes. In Scenarios 2, there are far more nonzero covariates, each with a relatively small effect size. In Scenarios 1 and 2, an equal-correlation structure for the covariates is adopted; in Scenario 3, a first-order autoregressive correlation structure is adopted.

To assess the performance of the screening procedures, we first examine the minimum model size, which is the smallest number of covariates such that all covariates in $\mathcal{M}_\star$ are included. We present the median and interquartile range of the minimum model size over 100 replications. The smaller the minimum model size, the better the procedure performs, because it results in a more parsimonious model. Second, we calculate the proportion of the 100 replications in which all covariates in $\mathcal{M}_\star$ are selected by submodel $\mathcal{M}_{d_n}$ of size $d_n = \lfloor n/\log n \rfloor$. We denote this proportion by $\mathcal{P}_{\text{All}}$. A screening procedure yielding $\mathcal{P}_{\text{All}}$ closer to one is considered more effective. In Table 1, we summarize the simulation results for different sample sizes (n) and censoring proportions (CP). Reported are the median and interquartile range (IQR) of the minimum model size needed to include all active covariates, along with the proportion $\mathcal{P}_{\text{All}}$ that all active predictors are selected by a submodel of size $\lfloor n/\log n \rfloor$.

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0373

Table 1: Simulation results for the AFT model.

Scenario	CP (%)	Method	Median	IQR	$\mathcal{P}_{\text{All}}$	Median	IQR	$\mathcal{P}_{\text{All}}$
			$n = 200$			$n = 400$		
S1	20	JS	13.0	35.75	0.73	5.0	0.00	0.99
		CR	2000.0	39.00	0.02	2000.0	0.25	0.00
		FAST	2000.0	0.00	0.00	2000.0	0.00	0.00
		PL	2000.0	0.00	0.00	2000.0	0.00	0.00
	40	JS	17.5	48.50	0.65	5.0	0.00	0.99
		CR	1193.5	1877.25	0.17	946.5	1821.25	0.22
		FAST	2000.0	0.00	0.00	2000.0	0.00	0.00
		PL	2000.0	0.00	0.00	2000.0	0.00	0.00
S2	20	JS	264	511.25	0.03	29.5	45.25	0.76
		CR	1981.5	52	0	2000	1	0
		FAST	1974.5	85.5	0	2000	1.25	0
		PL	1970	81.25	0	2000	2	0
	40	JS	450	698.25	0	53	84.5	0.62
		CR	763	874.5	0	317	610.25	0.12
		FAST	1982.5	63	0	2000	2	0
		PL	1982	61.75	0	2000	2	0
S3	20	JS	313.5	456.5	0.07	9	39.25	0.77
		CR	1134	1012	0	1180.5	1035.25	0.04
		FAST	998	1032.25	0.01	909	977	0.01
		PL	1034	1022.75	0.01	945	985.75	0.01
	40	JS	392.5	681	0	8	20.5	0.88
		CR	1707	521.75	0	1785	461.25	0
		FAST	1051.5	876	0	943.5	849.75	0.06
		PL	1041.5	867.5	0.02	920.5	870	0.06

In all three scenarios, the covariates are correlated with one another. Thus, some jointly irrelevant covariates may become marginally relevant

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0373

to the outcome. Consequently, marginal screening methods tend to yield erroneous results. Based on the results shown in Table 1, all three methods choose a substantial number of irrelevant covariates. In contrast, the proposed joint screening method tends to select much fewer irrelevant covariates. Moveover, for a given submodel size, the joint screening method has a much larger probability of selecting all truly relevant covariates. Therefore, the joint screening method outperforms the marginal screening methods.

To assess the robustness of the proposed approach toward a model misspecification, we generate data from the following semiparametric linear transformation model:

$$H(T_i) = \boldsymbol{X}_i^{\mathsf{T}} \boldsymbol{\beta}_\star + \varepsilon_i, \tag{5.1}$$

where $H(t) = \log\{0.5(e^{2t} - 1)\}$, and ε_i follows the standard extreme value distribution or the standard logistic distribution, corresponding to the proportional hazards model or the proportional odds model, respectively. The simulation settings for $\boldsymbol{X}_i$ and C_i are as before. The simulation results are summarized in Tables 2 and 3 for the proportional hazards model and proportional odds model, respectively. In both misspecified cases, the joint screening method performs very well in terms of both the screening consistency and the minimum model size. In contrast, the marginal screen-

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0373

ing methods are essentially unable to identify jointly relevant covariates. Therefore, the good performance of the joint screening method is quite robust toward model misspecifications.

Next, we illustrate the bootstrap-based tuning procedure proposed in Section 3.1. To do so, we use a randomly selected simulated data set from Scenario 1, with $n = 400$ and $p = 2000$. For the bth bootstrap sample and a given threshold $\gamma_n = \gamma$, let $\mathcal{M}_\lambda^{(b)}$ denote the indices of the covariates selected by the joint screening method. Figure 1 displays the proportion of each covariate being selected in 100 bootstrap samples against a fine grid of γ. The red curves refer to the five truly relevant covariates, and the blue curves refer to the remaining truly irrelevant covariates. We see that the red and blue curves are well separated from each other over a wide range of values of the threshold γ. Therefore, given a large proportion among 100 bootstrap samples, the joint screening method is able to identify all and only the truly relevant covariates for a wide range of γ. For example, when $\gamma = 0.32$, the joint screening method selects $\mathcal{M}_\lambda^{(b)} = \mathcal{M}_\star$ 87 out of 100 times.

To further examine the bootstrap-based tuning procedure, we repeat the above process by 100 times. In each repetition, 100 bootstrap samples are generated. The joint screening method is applied to each bootstrap

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0373

Table 2: Simulation results for the proportional hazards model.

Scenario	CP (%)	Method	Median	IQR	$\mathcal{P}_{\text{All}}$	Median	IQR	$\mathcal{P}_{\text{All}}$
			$n = 200$			$n = 400$		
S1	20	JS	53	127.75	0.4	5	2.25	0.98
		CR	2000	0	0	2000	0	0
		FAST	2000	0	0	2000	0	0
		PL	2000	0	0	2000	0	0
	40	JS	86.5	239	0.27	8	8	0.96
		CR	48	693.25	0.47	19	881	0.55
		FAST	2000	0	0	2000	0	0
		PL	2000	0	0	2000	0	0
S2	20	JS	521	530	0	77	130.5	0.45
		CR	1990.5	31	0	2000	2	0
		FAST	1983	77	0	2000	2	0
		PL	1983	69	0	2000	2	0
	40	JS	565	649.25	0	94.5	137.75	0.36
		CR	1965.5	112.75	0	1998	14	0
		FAST	1988	37	0	2000	2.25	0
		PL	1988.5	39.25	0	2000	2.25	0
S3	20	JS	187	329.75	0.09	21.5	108.75	0.7
		CR	925.5	848.25	0	989.5	1097.25	0.02
		FAST	1034.5	985.75	0	814.5	911.5	0.01
		PL	1051	984	0	778	878.25	0.02
	40	JS	372	737	0.05	13	20.75	0.87
		CR	1078.5	703.5	0	1112.5	835.5	0
		FAST	1103	890	0.02	1052	1070.25	0.03
		PL	1124.5	877	0.01	1048	1027.5	0.02

sample, resulting in $\mathcal{M}_d^{(b)}$, where $d = \lfloor n/\log n \rfloor$, for $b = 1, \cdots, 100$. Then, for the repetitions, we calculate the intersection of the submodels; that

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0373

Table 3: Simulation results for the proportional odds model.

Scenario	CP (%)	Method	Median	IQR	$\mathcal{P}_{\text{All}}$	Median	IQR	$\mathcal{P}_{\text{All}}$
			$n = 200$			$n = 400$		
S1	20	JS	32	72.75	0.56	5	2	0.96
		CR	2000	0	0	2000	0	0
		FAST	2000	0	0	2000	0	0
		PL	2000	0	0	2000	0	0
	40	JS	71.5	180	0.33	6	7.25	0.93
		CR	1076	1777.75	0.15	1008.5	1842	0.24
		FAST	2000	0	0	2000	0	0
		PL	2000	0	0	2000	0	0
S2	20	JS	459	632.5	0	65	84	0.51
		CR	1987	68.25	0	2000	1	0
		FAST	1984.5	66.5	0	2000	2	0
		PL	1986	68.75	0	2000	1	0
	40	JS	716	673.25	0	98.5	150	0.35
		CR	958	736.75	0	301.5	855.75	0.12
		FAST	1977.5	56.75	0	2000	1	0
		PL	1979	57.5	0	2000	1	0
S3	20	JS	177	550.75	0.09	15.5	62	0.72
		CR	897.5	984.5	0.01	987	787.25	0.02
		FAST	879	1053.75	0.03	1087	988.5	0.02
		PL	842	1087.5	0.02	1104.5	965.75	0.02
	40	JS	506.5	841	0.01	10	38	0.88
		CR	1391.5	806	0	1200.5	889.25	0
		FAST	1053	995.25	0.02	1029	891.25	0.05
		PL	1011	996.5	0.02	1025	952.5	0.06

is, $\overline{\mathcal{M}}_d = \bigcap_{b=1}^{100} \mathcal{M}_d^{(b)}$. Based on the sure screening properties in Theorem 2, we expect that, with large probability, the intersection $\overline{\mathcal{M}}_d$ is exactly

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0373

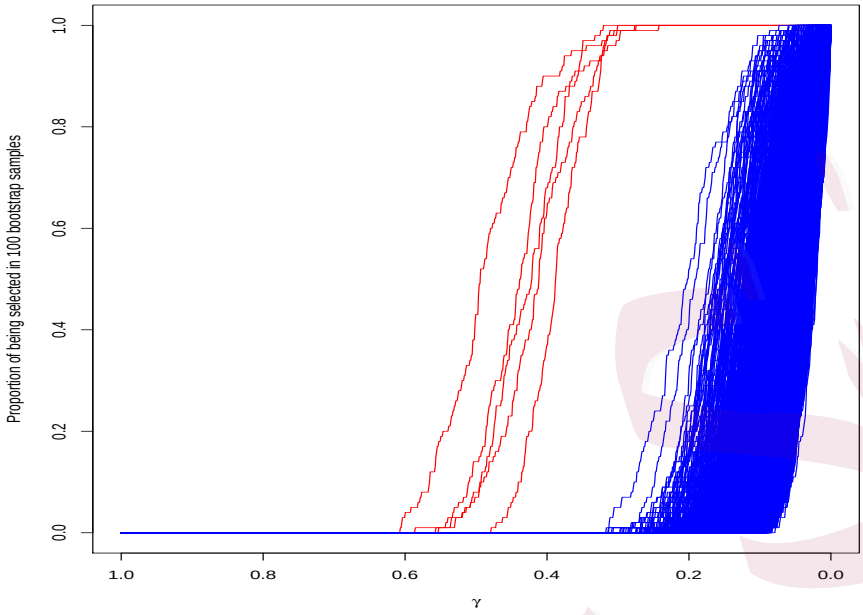

Figure 1: The selection result based on 100 bootstrap samples using a randomly selected simulated data set from Scenario 3, with $n = 400$ and $p = 2000$. The red curves correspond to the five truly relevant covaraites, and the blue curves correspond to the remaining irrelevant covariates.

$\mathcal{M}_\star = \{1, 2, 3, 4, 5\}$. Our simulation supports our expectation: of the 100 repetitions, 82 times $\overline{\mathcal{M}}_d = \mathcal{M}_\star$, 10 times $\overline{\mathcal{M}}_d$ consists of only four variables in $\mathcal{M}_\star$, and eight times $\overline{\mathcal{M}}_d$ misses two or more variables in $\mathcal{M}_\star$ or includes one irrelevant variable.

From this simulation study, we see that the bootstrap-based tuning

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0373

procedure performs very well. However, it seems crucial to determine the number of bootstrap samples; here, we subjectively use $B = 100$ samples. In practice, we should investigate different choices of B, with the help of proportion curves, such as those in Figure 1. If B is too large, we may miss some relevant covariates. However, if B is too small, we may include some irrelevant covariates.

6. Real-Data Application

We apply the proposed joint screening method and the three marginal screening methods to the adult acute myeloid leukemia (AML) data of Bullinger et al. (2004). Complementary-DNA microarrays were used to determine the levels of gene expression in peripheral-blood samples or bone marrow samples from 116 adults with AML (including 45 with a normal karyotype). The primary goal of the study was to identify relevant genes and devise a predictive model for the survival outcome of a patient using the genetic profile of a tumor. The data set contains data on $6,283$ genes and 116 patients. The median survival time was 1.09 years. During follow-up, 67 patients died of leukemia. The remaining 49 patients were censored, yielding a censoring rate of 42.2%.

The top 10 selected genes are reported in Table 4, showing that the joint approach yields quite different results to those of the marginal screening

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0373

Table 4: Top 10 genes selected by applying different screening methods to acute myeloid leukemia data.

Order	JS	CR	FAST	PL
1	117315	103875	112298	103875
2	112578	109607	111553	112298
3	101791	103308	119834	319580
4	112353	221677	330857	119821
5	247136	119133	109477	112283
6	112298	115614	117339	101364
7	117570	102345	119821	109607
8	103875	116402	313178	103308
9	223434	117549	221973	109541
10	225314	112298	117386	112105

methods. For the joint screening method, we visualize the effect paths of the top ten genes in Figure 2, showing the selection proportion curve for each gene against a fine grid of γ. Here, the selection curves are obtained using 100 bootstrap samples, with the red curves corresponding to the top 10 genes shown in the first column of Table 4. Figure 2 shows that the top ten genes are clearly identifiable over a wide range of γ.

To further compare the screening methods, we consider prediction errors (PE), as follows. First, we randomly split the data set into a training set comprising 58 observations and a test set containing the remaining 58

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0373

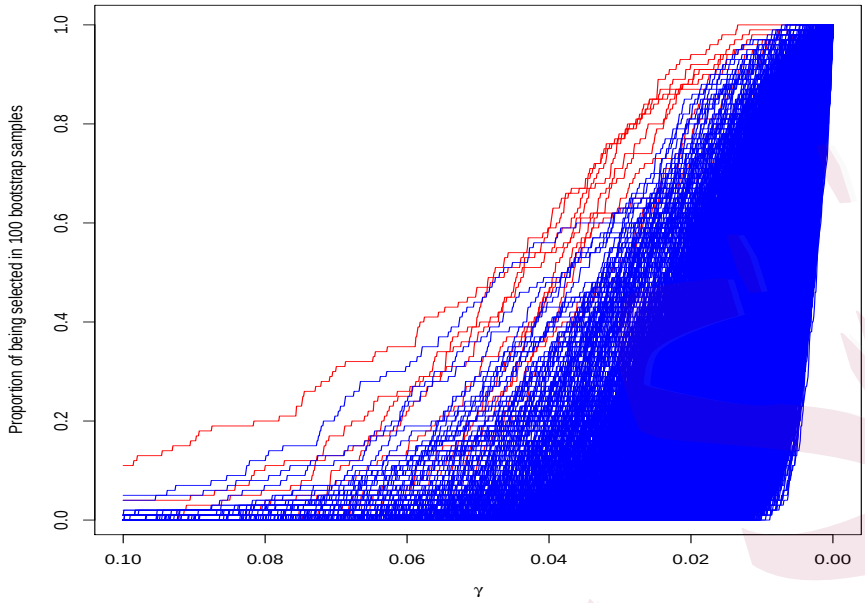

Figure 2: The selection-proportion curves based on 100 bootstrap samples when applying the joint screening method to acute myeloid leukemia data.

observations. We applied the method being evaluated to the training data set, obtaining estimators of the regression coefficients $\widehat{\boldsymbol{\beta}}_{\text{select}}$ for selected variables $\boldsymbol{X}^{\text{select}}$. Then, the prediction error was calculated as

$$\frac{\sum_{i=1}^{n} I(i \text{ in Test Set})\Delta_i(Y_i - \widehat{\boldsymbol{\beta}}_{\text{select}}^{\mathsf{T}}\boldsymbol{X}_i^{\text{select}})^2/\widehat{G}_{\text{test}}(Y_i)}{\sum_{i=1}^{n} I(i \text{ in Test Set})\Delta_i/\widehat{G}_{\text{test}}(Y_i)},$$

where $\widehat{G}_{\text{test}}$ denotes the Kaplan–Meier estimator of the survival function of the censoring time, computed based on the test data set. Here, a smaller

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0373

prediction error indicates better performance. The predictor errors for four screening methods with models of size from one to ten are summarized in Table 5. The average prediction error (APE) and median prediction error (MPE) based on 100 random splits of the data set are reported. We see that, overall, the proposed joint screening meothd outperforms the marginal screening methods, achieving the smallest APE and MPE. Furthermore, given that the selected model contains eight covariates, the joint screening method achieves the best prediction accuracy.

Table 5: Prediction errors for the four screening methods when applied to adult acute myeloid leukemia data.

Model size	JS	CR	FAST	PL	JS	CR	FAST	PL
	APE				MPE			
1	5.97	6.25	6.13	6.25	5.86	6.25	6.18	6.25
2	6.01	6.48	6.05	5.86	5.90	6.57	6.20	5.72
3	6.10	6.43	6.38	6.01	6.03	6.42	6.31	6.10
4	5.76	6.66	6.82	6.65	5.82	6.66	6.72	6.60
5	5.70	7.11	7.49	6.54	5.67	6.99	7.27	6.39
6	5.26	7.36	7.86	6.68	5.18	7.25	7.66	6.51
7	4.83	7.51	8.63	6.95	4.69	7.45	8.24	6.86
8	**4.66**	7.81	8.54	7.03	**4.41**	7.66	8.03	6.81
9	4.80	8.11	9.08	7.23	4.65	7.89	8.71	6.99
10	4.92	7.83	8.85	7.34	4.75	7.65	8.32	7.22

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0373

7. Discussion

In a high-dimensional survival analysis, where the number of covariates greatly exceeds the number of observations, a preliminary screening method reduces the data dimension effectively, simplifying the subsequent detailed data analysis. The effectiveness of a screening approach depends on whether important variables are retained when the data dimension is reduced. However, most existing approaches evaluate the relevance of variables, based only on marginal survival models. This prevents a joint regression analysis of survival data. This is less desirable because, in most high-dimensional situations, relevant variables exhibit significant effects in a joint manner, not marginally.

We develop a new screening scheme that employs the AFT model to directly evaluate the joint covariate-survival association in the presence of an ultrahigh-dimensional vector of covariates. The resulting screening procedure is easy to implement, enjoys easy and direct interpretation, and exhibits sure screening properties. For ease of exposition, we consider the situation where the censoring time is independent of both the covariates and the survival time. When the censoring time possibly depends on some covariates, our method can be extended by replacing the Kaplan–Meier estimator with the local Kaplan–Meier estimator using kernel smoothing

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0373

Gonzalez-Manteiga and Cadarso-Suarenz (1994).

Supplementary Material

The online Supplementary Material contains detailed proofs of Theorems 1–2 and Propositions 1–2.

Acknowledgements

The authors are grateful to the Editor, an associate editor, and the referees for many helpful comments. This work was supported in part by the University of Hong Kong Seed Fund for Translational and Applied Research (201711160015) and Basic Research (201811159052), the University of Hong Kong - Zhejiang Institute of Research and Innovation Seed Fund, and General Research Fund (17308018) of Hong Kong.

References

Bach, F. (2008). Bolasso: model consistent lasso estimation through the bootstrap. *Proceedings of the 25th international conference on Machine learning*, 33-40.

Bitouze, D., Laurent, B. and Massart, P. (1999). A Dvoretzky-Kiefer-Wolfowitz type inequality for the Kaplan-Meier estimator. *Annales de l'IHP Probabilites et Statistiques* **35**, 735-763.

Bullinger, L., Dohner, K., Bair, E., Frohling, S., Schlenk, R., Tibshirani, R., Dohner, H. and Pollack, J. (2004). Use of gene-expression profiling to identify prognostic subclasses in adult

acute myeloid leukemia. *N. Engl. J. Med.* **350**, 1605-1616.

Analysis of transformation models with censored data. *Biometrika* **82**, 835–45.

Chen, J. and Chen, Z. (2008). Extended bayesian information criteria for model selection with large model spaces. *Biometrika* **95**, 759-771.

Candes, E. and Tao, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics* **35**, 2313-2351.

Cox, D. R. (1972). Regression models and life-tables (with Discussion). *J. R. Statist. Soc. B* **34**, 187–220.

Fan, J., Feng, Y. and Wu, Y. (2010). Ultrahigh dimensional variable selection for Cox's proportional hazards model. *IMS Collections* **6**, 70-86.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* **96**, 1348-1360.

Fan, J. and Lv, J. (2008). Sure Independence Screening for ultrahigh dimensional feature space (with discussion). *J. R. Statist. Soc. B* **70**, 849-911.

Gail, M. H., Wieand, S. and Piantadosi, S. (1984). Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates. *Biometrika* **71**, 431-444.

Gonzalez-Manteiga, W. and Cadarso-Suarez, C. (1994). Asymptotic properties of a generalized Kaplan-Meier estimator with some applications. *J. Nonparam. Statist* **4**, 65-78.

Gorst-Rasmussen, A. and Scheike, T. (2013). Independent screening for single-index hazard rate models with ultrahigh dimensional features. *J. R. Statist. Soc. B* **75**, 217-245.

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

Hong, H. G., Kang, J. and Li, Y. (2018) Conditional screening for ultra-high dimensional covariates with survival outcomes. *Lifetime Data Analysis* **24**, 45-71.

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 models. *Biometrika* **90**, 341-353.

Jin, Z., Lin, D. Y. and Ying, Z. (2006). On least-squares regression with censored data, *Biometrika* **93**, 147-161.

Kalbfleisch, J. D. and Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data*, 2nd ed. New York: Wiley.

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

Peng, L. and Fine, J. (2009). Competing risks quantile regression. *Journal of the American Statistical Association* **104**, 1440-1453.

Song, R., Lu, W., Ma, S. and Jeng, X. J. (2014). Censored rank independence screening for

high-dimensional survival data. *Biometrika* **101**, 799-814.

Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society: Series B* **58**, 267-288.

Tibshirani, R. (1997). The lasso method for variable selection in the Cox model. *Statistics in Medicine* **16**, 385–95.

Vershynin, R. (2010). Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint* arXiv:1011.3027.

Wang, X. and Leng, C. (2016). High dimensional ordinary least squares projection for screening variables. *J. R. Statist. Soc. B* **78**, 589-611.

Wei, L. J. (1992). The accelerated failure time model: a useful alternative to the Cox regression model in survival analysis. *Statistics in medicine* **11**, 1871-1879.

Ying, Z., Jung, S. H. and Wei, L.J. (1995). Survival analysis with median regression models. *Journal of the American Statistical Association* **90**, 178-84.

Zhang, H. and Lu, W. (2007). Adaptive Lasso for Cox's proportional hazards model. *Biometrika* **94**, 691-703.

Zhao, P. and Yu, B. (2006). On model selection consistency of Lasso. *Journal of Machine Learning Research* **7**, 2541-2563.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association* **101**, 1418-1429.

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0373

Data and Statistical Science, AbbVie, North Chicago, IL, USA

E-mail: yf2113@gmail.com

Department of Statistics and Actuarial Science, The University of Hong Kong, Hong Kong

E-mail: xujf@hku.edu