Appl. Math. J. Chinese Univ.
2017, 32(4): 379-396

Feature selection of ultrahigh-dimensional covariates with survival outcomes: a selective review

HONG Hyokyoung Grace[1] LI Yi[2]

Abstract. Many modern biomedical studies have yielded survival data with high-throughput predictors. The goals of scientific research often lie in identifying predictive biomarkers, understanding biological mechanisms and making accurate and precise predictions. Variable screening is a crucial first step in achieving these goals. This work conducts a selective review of feature screening procedures for survival data with ultrahigh dimensional covariates. We present the main methodologies, along with the key conditions that ensure sure screening properties. The practical utility of these methods is examined via extensive simulations. We conclude the review with some future opportunities in this field.

§1 Introduction

Modern biomedical studies have generated abundant survival data with high dimensional biomarkers for various scientific purposes. For instance, identifying genomic profiles that are associated with cancer patients' survival may help with understanding disease progression processes and designing more effective gene therapies. With the advent of new biotechnologies, the emergence of high-throughput data, such as gene expressions, SNPs, methylation and next-generation RNA sequencing, has pushed the dimensionality of data to a larger scale. In these cases, the dimensionality of covariates may grow exponentially with the sample size and such data has been commonly referred to as ultrahigh dimensional data ([5]).

When the number of covariates (p) is less than the sample size (n), the parametric regression, such as Weibull models, and the semiparametric regression, such as the Cox proportional hazards model and the Accelerated Failure Time (AFT) model, have been routinely used for modeling censored outcome data in many practical settings. When $p > n$, penalized likelihood methods have been proposed by various authors ([20], [4], [25], [29]) and the oracle properties and statistical error bounds of estimation have been established ([13], [15]). However,

Received: 2017-09-01. Revised: 2017-11-04.
MR Subject Classification: 97K80.
Keywords: survival analysis, ultrahigh dimensional predictors, variable screening, sure screening property.
Digital Object Identifier(DOI): https://doi.org/10.1007/s11766-017-3547-8.
Supported by the National Natural Science Foundation of China (11528102) and the National Institutes of Health (U01CA209414).

when $p \gg n$, computational issues inherent in these methods makes them non-applicable to ultrahigh-dimensional statistical learning problems because of serious challenges in "computational expediency, statistical accuracy, and algorithmic stability" ([6]). A recent work by [2] did establish the oracle properties of the regularized partial likelihood estimates under an ultrahigh dimensional setting. The results, however, required the optimizers to the penalized partial likelihood function to be unique and global, which is, in general, difficult to verify, especially when the dimension of covariates is exceedingly high.

A seminal paper by [5] has demonstrated a simple but useful way to deal with ultrahigh dimensional regression. First, a variable screening procedure is used as a fast and crude tool for reducing the dimensionality to a moderate size (usually below the sample size). In the second step, a more sophisticated technique, such as penalized likelihood methods, can be further applied to perform the final feature selection and parameter estimation simultaneously.

In the framework of linear regression with normal errors, [5] showed that sure independence screening (SIS), which recruits features with the largest marginal correlations with the response, has the desirable sure screening property. That is, with probability converging to 1, the screening procedure retains all of the important features in the model. While screening approaches have been actively pursued for completely observed outcome data, the development of ultrahigh dimensional screening tools with survival outcomes, however, has been less fruitful. Several ad-hoc solutions are available from [21] and [3], though detailed accounts of practical utility or theoretical support are still elusive. Recent years have seen a rapid surge in variable screening methods for survival data, but to our knowledge, no systematic reviews and comparisons are available. To fill the gap, we will review and compare several representative works in this field. Specifically, we first review model-motivated screening methods, including the principled sure screening by [27], the feature aberration at survival times screening by [8] and the conditional screening by [11]. We then review several model-free methods, including the quantile adaptive sure independence screening by [9], the censored rank independence screening procedure by [19], the survival impact index screening by [17] and the integrated powered density screening by [10]. We introduce the motivation of each method, describe the main conditions that lead to the sure screening property and numerically compare all the methods under the same settings.

Below we describe notation and terminologies that will be used throughout this paper. Suppose we have n observations with p covariates, where $p \gg n$. Denote by X_{ij} the jth covariate for subject i, and write $\mathbf{X}_i = (X_{i1}, \ldots, X_{ip})^{\mathrm{T}}$. Let T_i and C_i be the underlying survival and censoring times, respectively. We, however, only observe $Y_i = \min\{T_i, C_i\}$, and the event indicator $\delta_i = I(T_i \leq C_i)$, where $I(\cdot)$ is the indicator function. In general, we assume random censoring such that C_i and T_i are independent given $\mathbf{X}_i$. Let $N_i(t) = I(Y_i \leq t, \delta_i = 1)$ be the observed failure process and $\tilde{Y}_i(t) = I(Y_i \geq t)$ be the at-risk process. We assume $(Y_i, \delta_i, \mathbf{X}_i)$ are independently and identically distributed (i.i.d). In particular, we assume $(T_i, X_{ij}), i = 1, \ldots, n$, are i.i.d copies of (T, X_j), the random variables that underlie the survival time and covariates. Denote by $S(\cdot)$ the marginal survival function of T and by $S(t|\mathbf{X})$ the conditional survival function of T given $\mathbf{X}$, where $\mathbf{X} = (X_1, \ldots, X_p)$. Let $G(t) = \mathrm{pr}(C_i > t)$ be the survival function

of C_i and $\widehat{G}(t)$ be the Kaplan-Meier estimator of $G(t)$ based on $\{Y_i, \delta_i\}$, $i = 1, \ldots, n$. Suppose that $S(t|\mathbf{X})$ only depends on a subset of covariates, denoted by $\mathcal{M}$. The overarching goal of variable screening is to estimate $\mathcal{M}$, which we let be $\widehat{\mathcal{M}}$. We note, however, the specific definitions of $\mathcal{M}$ and $\widehat{\mathcal{M}}$ may differ in the reviewed papers and will be defined under the specific contexts.

§2 Screening methods

2.1 Principled Cox sure independence screening

[27] generalized the sure independence screening ([5]) to the Cox proportional hazards model, which stipulates that the hazard at time $t > 0$ for a subject i with the vector of covariates $\mathbf{X}_i$ is

$$\begin{aligned} h(t|\mathbf{X}_i) &= \lim_{\Delta t \to 0+} \frac{1}{\Delta t} P(t \le T \le t + \Delta t | T \ge t, \mathbf{X}_i) \\ &= h_0(t) \exp\left(\sum_{j=1}^{p} X_{ij} \beta_j \right), \end{aligned} \tag{1}$$

where $h_0(t)$ is an unspecified baseline hazard function.

Under (1), [27] defined $\mathcal{M} = \{j : \beta_j \neq 0\}$ and proposed to estimate it as follows. Assuming a (working and possibly misspecified) marginal Cox model on each X_j, namely, $h^*_{0,j}(t) \exp(X_{ij}\beta^*_j)$, they obtained the maximum partial likelihood estimate of β^*_j, denoted by $\hat{\beta}_j$. Then, the importance of X_j was measured by a Wald type statistic for testing $\beta^*_j = 0$. As a result, the estimated $\mathcal{M}$ was given by

$$\widehat{\mathcal{M}} = \{j : I_j(\hat{\beta}_j)^{\frac{1}{2}} |\hat{\beta}_j| \ge \lambda_n\},$$

where $\hat{\beta}_j$ solves the partial likelihood score equation

$$U_j(\beta) = \frac{1}{n} \sum_{i=1}^{n} \int_0^{\nu} \left\{ X_{ij} - \frac{\sum_{i=1}^{n} X_{ij} \exp(X_{ij}\beta) \tilde{Y}_i(t)}{\sum_{i=1}^{n} \exp(X_{ij}\beta) \tilde{Y}_i(t)} \right\} dN_i(t) = 0, \tag{2}$$

$I_j(\hat{\beta}_j) = -\frac{\partial U_j(\beta)}{\partial \beta}|_{\beta=\hat{\beta}_j}$ is the observed information at $\hat{\beta}_j$, and λ_n is a pre-specified cut-off that depends on n. Here, $\nu > 0$ in (2) is the study duration, which is assumed to be long enough to ensure that ample events are observed during the interval $[0, \nu]$.

As there were no principled ways of choosing λ_n, the work of [27] provided a new perspective of choosing λ_n such that one can achieve the sure screening property while controlling the false positive rate, or the proportion of unimportant covariates incorrectly included in $\widehat{\mathcal{M}}$. The rationale is as follows.

If the true model size $|\mathcal{M}| = s$, then the expected false positive rate can be written as

$$E\left(\frac{|\mathcal{M}^c \cap \widehat{\mathcal{M}}|}{|\mathcal{M}^c|} \right) = \frac{1}{p-s} \sum_{j \in \mathcal{M}^c} P\{I_j(\hat{\beta}_j)^{\frac{1}{2}} |\hat{\beta}_j| \ge \lambda_n\}.$$

Moreover, when $\beta_j = 0$ or $j \in \mathcal{M}^c$, $I_j(\hat{\beta}_j)^{\frac{1}{2}} \hat{\beta}_j$ converges in distribution to a standard normal variable. Hence, by setting $\lambda_n = \Phi^{-1}(1 - \mathtt{f}/(2p))$, the expected false positive rate is $2(1 - \Phi(\lambda_n)) = \mathtt{f}/p$, which is approximately equal to the desirable false positive rate, $\mathtt{f}/(p-s)$.

Here, $\Phi(\cdot)$ is the standard normal cumulative distribution function and `f` is the number of false positives that one is willing to tolerate. The screening procedure is commonly referred to as the principled Cox sure independence screening (PSIS).

To study the sure screening property, they first established the following "β-min" condition (that is, the true signals have enough marginal strengths): there exist constants $c_1 > 0$ and $0 < \kappa < 1/2$ such that $\min_{j\in\mathcal{M}} |\mathrm{cov}[X_{ij}, \mathrm{E}\{F_T(C_i|\mathbf{X}_i)|\mathbf{X}_i\}]| \geq c_1 n^{-\kappa}$, where $F_T(\cdot|\mathbf{X}_i)$ is the cumulative distribution function of T_i given $\mathbf{X}_i$. Then [27] proved that

$$\min_{j\in\mathcal{M}} |\beta_j| \geq c_2 n^{-\kappa}, \tag{3}$$

where c_2 is a positive constant. This result led to the sure screening property

$$\mathrm{pr}(\mathcal{M} \subset \widehat{\mathcal{M}}) \geq 1 - s\exp(-c_3 n^{1-2\kappa}), \tag{4}$$

where c_3 is a positive constant. As $n \to \infty$, $\mathrm{pr}(\mathcal{M} \subset \widehat{\mathcal{M}}) \to 1$ for the nonpolynomial (NP)-dimensionality problem $\log(p) = O(n^{1-2\kappa})$.

[27] demonstrated that, numerically, PSIS is stable and computationally efficient in selecting true signals. However, given that PSIS stems from a Wald test based on a Cox model, its performance is unclear when the underlying assumption of a Cox model fails.

2.2 Feature aberration at survival times screening

With the goal of making the screening procedure less model-centric, [8] proposed a 'feature aberration at survival times' (FAST) statistic that measures the aberration of each covariate relative to its at-risk average. Specifically, for covariate X_j, the FAST statistic is defined as

$$d_j = \frac{1}{n}\sum_{i=1}^{n}\int_0^{\nu}\left\{X_{ij} - \frac{\sum_{i=1}^{n} X_{ij}\tilde{Y}_i(t)}{\sum_{i=1}^{n}\tilde{Y}_i(t)}\right\} dN_i(t), \tag{5}$$

where $t \in [0, \nu]$.

The statistic can be justified under an administrative censoring scheme ($C_i \equiv \nu$). With standardized covariates such that $\mathrm{E}(X_j) = 0$ and $\mathrm{var}(X_j) = 1$ for $j = 1, \ldots, p$, first note that the population version of d_j is

$$\tilde{d}_j = \mathrm{E}(d_j) = \mathrm{cov}\{X_j, F_T(\nu|\mathbf{X}_i)\} + \int_0^{\nu} \mathrm{cov}\{X_j, F_T(t|\mathbf{X}_i)\}K(t)dt,$$

where $F_T(t|\mathbf{X}_i) = P(T_i \leq t|\mathbf{X}_i)$ and $K(\cdot)$ is a strictly positive function. Thus, $\tilde{d}_j$ is large if $\mathrm{cov}\{X_j, F_T(t|\mathbf{X}_i)\}$ has a constant sign throughout $t \in [0, \nu]$. Thus, it is reasonable to consider the magnitude of d_j as a marginal utility to rank the importance of X_j.

In addition, by taking β to be 0 in (2), we note that (2) reduces to (5). Therefore, FAST can also be viewed as a score test statistic based on a Cox model, which is a special case of the score-based screening proposed by [28].

To study the sure screening property, [8] assumed that the true hazard function is of the single-index form

$$h_i(t) = h(t, \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}), \quad i = 1, \ldots, n, \tag{6}$$

and required the resulting survival function $\exp\{\int_0^t h(s, \cdot)ds\}$ to be strictly monotonic for each $t \geq 0$. They argued that such an assumption holds true for a variety of models, including the

additive model ([18]), the Cox model and the AFT model. Under (6), [8] defined the true set as

$$\mathcal{M} = \{j : \beta_j \neq 0\},$$

and proposed to estimate it by

$$\widehat{\mathcal{M}} = \{j : |d_j| > \lambda_n\},$$

for a given λ_n.

Assuming various regularity conditions, [8] showed that there exists a threshold $\zeta_n > 0$ such that $\min_{j\in\mathcal{M}} |\tilde{d}_j| \geq \zeta_n$ and $\max_{j\notin\mathcal{M}} |\tilde{d}_j| = 0$. Thus, the signals $\tilde{d}_j$ when $j \in \mathcal{M}$ are stronger than those when $j \notin \mathcal{M}$. They further assumed that $|\mathrm{cov}(X_j, \mathbf{X}^{\mathrm{T}}\boldsymbol{\beta})| \geq c_1 n^{-\kappa},\ j \in \mathcal{M}$, for some $c_1 > 0$ and $\kappa < 1/2$. Then they showed that, by taking $\lambda_n = c_2 n^{-\kappa}$ for some constant $0 < c_2 \leq c_1/2$, the sure screening property holds even when p grows exponentially fast with n, or,

$$\mathrm{pr}(\mathcal{M} \subset \widehat{\mathcal{M}}) \to 1 \quad \text{when} \ \ n \to \infty.$$

As FAST is essentially a score test that requires fitting only the null model, its computation is simpler than the Wald test based screening methods, such as PSIS ([28]). Like the sure independence screening (SIS) of [5], FAST assumes that the covariates present in the true model $\mathcal{M}$ are independent of the irrelevant covariates. This assumption is often violated in practice. To account for possible correlations between variables, [5] proposed an iterative SIS procedure (ISIS): after applying the SIS procedure, the relevance of the unselected covariates is reassessed given the selected covariates and a small number of the most relevant features among them can be added to the selected set. These iterative steps are repeated until some stopping criterion is reached. [5] and [3] showed that an iterative screening procedure may perform better than a non-iterative procedure. A similar iterated FAST procedure was proposed by [8]; however, like the general ISIS procedure, its theoretical support is still an open problem.

2.3 Conditional screening

Intensive biomedical research has generated a large body of biological knowledge. Incorporating such knowledge may lead to improved accuracy in modeling. However, the marginal screening approaches, such as PSIS and FAST, were not designed to integrate prior knowledge into variable screening. Recent years have seen an advent of conditional sure independence screening methods that use *a priori* information; see [1], [12]. By including such important predictors as the conditioning variables, conditional screening ranks the marginal utility of each variable after adjusting for these conditioning variables.

[11] proposed a conditional screening method under the Cox proportional hazards model (1) by computing the marginal contribution of each covariate given priorly known information. This method is referred to as the Cox conditional screening (CoxCS).

Denote by $\mathcal{C}$ the indices of the set of covariates that are known *a priori* to be related to the number of covariates. Let $\mathcal{M}_{-\mathcal{C}} = \{j \notin \mathcal{C}, \beta_j \neq 0\}$, $q = |\mathcal{C}|$, and $\mathbf{X}_{\mathcal{C}} = (X_j, j \in \mathcal{C})^{\mathrm{T}}$. [11] proposed to fit the marginal Cox regression by including the known covariates in $\mathbf{X}_{\mathcal{C}}$.

Specifically, for each $X_j \notin \mathbf{X}_{\mathcal{C}}$, they considered the following Cox regression model

$$h_j(t, \mathbf{X}_i) = h_{j,0}(t) \exp(\boldsymbol{\beta}_{\mathcal{C}}^{\mathrm{T}} \mathbf{X}_{i\mathcal{C}} + \beta X_{ij}).$$

The maximum partial likelihood estimates $(\widehat{\boldsymbol{\beta}}_{\mathcal{C}}^{\mathrm{T}}, \widehat{\beta}_j)^{\mathrm{T}}$ can be obtained by solving the following equations:

$$\mathbf{V}_j(\boldsymbol{\beta}_{\mathcal{C}}, \beta) = [V_{j,k}(\boldsymbol{\beta}_{\mathcal{C}}, \beta)]^{\mathrm{T}} = \mathbf{0}_{q+1},$$

with

$$V_{j,k}(\boldsymbol{\beta}_{\mathcal{C}}, \beta) = \sum_{i=1}^{n} \int_0^{\nu} \left\{ X_{ik} - \frac{\sum_{i=1}^{n} X_{ik} \tilde{Y}_i(t) \exp(\boldsymbol{\beta}_{\mathcal{C}}^{\mathrm{T}} \mathbf{X}_{i\mathcal{C}} + \beta X_{ij})}{\sum_{i=1}^{n} \tilde{Y}_i(t) \exp(\boldsymbol{\beta}_{\mathcal{C}}^{\mathrm{T}} \mathbf{X}_{i\mathcal{C}} + \beta X_{ij})} \mathrm{d}N_i(t) \right\},$$

for $k \in \mathcal{C} \cup \{j\}$. The key is to recruit variables with large additional contributions given $\mathbf{X}_{\mathcal{C}}$. If the conditioning set $\mathbf{X}_{\mathcal{C}}$ is empty, the method reduces to PSIS. Using the magnitude of $\hat{\beta}_j$ as a marginal utility to rank the importance of X_j, the set of selected variables among $j \notin \mathcal{C}$ is given by

$$\widehat{\mathcal{M}}_{-\mathcal{C}} = \{j \notin \mathcal{C} : |\hat{\beta}_j| > \lambda_n\},$$

for a pre-defined λ_n.

To study the asymptotic behavior of the proposed procedure, [11] assumed the following "beta-min" condition: for constants $c_1 > 0$ and $0 < \kappa < 1/2$,

$$\min_{j \in \mathcal{M}_{-\mathcal{C}}} |\mathrm{E}[\mathrm{cov}^*\{X_j, \mathrm{pr}(\delta = 1|\mathbf{X})|\mathbf{X}_{\mathcal{C}}\}]| \geq c_1 n^{-\kappa}, \tag{7}$$

where $\mathrm{cov}^*(\zeta_1, \zeta_2|\boldsymbol{\xi})$ indicates the conditional linear covariance between ζ_1 and ζ_2 given $\boldsymbol{\xi}$. [11] introduced this new concept in order to approximate the ordinary conditional variance and facilitate the proof of the sure screening property.

Under (7) and some other conditions, [11] proved that the estimated $\hat{\beta}_j$ converges uniformly to β_j in probability. That is, for any given $\epsilon_1 > 0$ and $\epsilon_2 > 0$,

$$\mathrm{pr}\left(\max_{j \in \mathcal{M}_{-\mathcal{C}}} |\hat{\beta}_j - \beta_j| > \frac{c_2}{2}(n^{-\kappa} - \epsilon_1) \right) \leq 2s(q+1) \exp(-c_3 n^{1-2\kappa}) + \epsilon_2,$$

where c_2 and c_3 are some positive constants, q is the size of $\mathcal{C}$, and s is the size of $\mathcal{M}_{-\mathcal{C}}$. In addition, they established the following "β-min" condition:

$$\min_{j \in \mathcal{M}_{-\mathcal{C}}} |\beta_j| \geq c_2 n^{-\kappa},$$

where $0 < \kappa < 1/2$.

Taking $\lambda_n = c_4 n^{-\kappa}$ with $c_4 > 0$, this "β-min" condition leads to the following sure screening property:

$$\mathrm{pr}(\mathcal{M}_{-\mathcal{C}} \subset \widehat{\mathcal{M}}_{-\mathcal{C}}) \to 1, \text{ as } n \to \infty.$$

The conditional screening approach can identify "marginally unimportant but jointly important" variables, commonly referred to as hidden variables. Failing to include hidden variables in the screening stage may cause undesirable consequences. For example, important features may be missed in model selection or biased inference may occur in the subsequent analysis.

Since CoxCS requires the prior information to be known and informative, it remains statistically challenging to develop efficient screening methods in the absence of such information. In the context of generalized linear models, [12] proposed a data-driven alternative in the absence of prior knowledge. The question of how to extend the data-driven approach to a survival

setting, however, would require more intensive theoretical and empirical work.

2.4 Quantile adaptive nonparametric screening

The validity of model-based screening methods, such as PSIS and CoxCS, often hinges upon the assumptions of the underlying models. For example, when the proportional hazards assumption fails, the model-based approaches may incur a large number of false negatives and lead to an invalid result. To develop a model-free framework that can be applicable to a more general class of survival models, [9] proposed the quantile adaptive sure independence screening (QaSIS). This approach performs screening based on the disparity between unconditional and conditional quantiles given each covariate. They defined the true set at a given quantile level $\alpha \in (0,1)$ as

$$\mathcal{M}_\alpha = \{j : Q_\alpha(T|\mathbf{X}) \text{ functionally depends on } X_j\},$$

where $Q_\alpha(T|\mathbf{X})$ is the α-th conditional quantile of T given $\mathbf{X}$. That is, $Q_\alpha(T|\mathbf{X}) = \inf\{t : \mathrm{pr}(T \le t|\mathbf{X}) \ge \alpha\}$.

Now let $q_{\alpha j} = \mathrm{E}\{Q_\alpha(T|X_j) - Q_\alpha(T)\}^2$, where $Q_\alpha(T|X_j)$ is the α-th conditional quantile of T given X_j and $Q_\alpha(T)$ is the marginal α-th quantile of T. Then, $q_{\alpha j}$ gauges the magnitude of the association of T with X_j, and T and X_j are independent if and only if $q_{\alpha j} = 0$ for every $\alpha \in (0,1)$.

To estimate $Q_\alpha(T|X_j)$, [9] proposed a spline estimator, $\hat{Q}_\alpha(T|X_j) = \boldsymbol{\pi}(X_j)^T\hat{\boldsymbol{\beta}}_j$. Here, $\boldsymbol{\pi}(x) = \{B_1(x), \ldots, B_N(x)\}^T$ is the normalized B-spline basis functions and $\hat{\boldsymbol{\beta}}_j = (\beta_{j1}, \ldots, \beta_N)^{\mathrm{T}}$ is obtained via the inverse probability weighted marginal quantile regression estimator, i.e.,

$$\widehat{\boldsymbol{\beta}}_j = \underset{\boldsymbol{\pi}\in\mathbb{R}^N}{\mathrm{argmin}} \sum_{i=1}^{n} \frac{\delta_i}{\widehat{G}(Y_i)} \rho_\alpha(Y_i - \boldsymbol{\pi}(X_{ij})^{\mathrm{T}}\boldsymbol{\beta}),$$

where $\rho_\alpha(x) = x[\alpha - I(x < 0)]$ is the quantile loss function. On the other hand, the marginal quantile $Q_\alpha(T)$ can be simply estimated by the inverse function of the Kaplan-Meier estimate of the α-th conditional quantile of T, $F^{-1}_{\mathrm{KM},n}(\alpha)$. Hence, $q_{\alpha j}$ can empirically be estimated by $\hat{q}_{\alpha j} = n^{-1}\sum_{i=1}^{n}\{\boldsymbol{\pi}(X_{ij})^{\mathrm{T}}\hat{\boldsymbol{\beta}}_j - F^{-1}_{\mathrm{KM},n}(\alpha)\}^2$. Then, $\mathcal{M}_\alpha$ is estimated as

$$\widehat{\mathcal{M}}_\alpha = \{j : \hat{q}_{\alpha_j} \ge \lambda_n\}$$

for some $\lambda_n > 0$. The method is deemed a model-free approach as it does not resort to a specific model structure.

To guarantee sure screening and control the false selection rate, [9] assumed that $0 < \kappa < 1/4$ and $N^3 n^{2\kappa-1} = o(1)$, where N is the number of basis functions. By taking $\lambda_n = c^* n^{-\kappa}$ with $c^* \le c_1/16$ and $c_1 > 0$ the sure screening property holds. That is,

$$\mathrm{pr}(\mathcal{M}_\alpha \subset \widehat{\mathcal{M}}_\alpha) \ge 1 - s_\alpha\{17\exp(-c_3 n^{1-4\kappa}) + 12N^2\exp(-c_4 N^{-3} n^{1-2\kappa})\} \to 1, \text{ as } n \to \infty,$$

where c_3 and c_4 are some positive constants and s_α is the size of $\mathcal{M}_\alpha$.

Overall, QaSIS is flexible as it allows the set of active variables to vary across quantiles, which is appealing to the analysis of heterogeneous data. However, since not every quantile is estimable under censoring, its performance under heavy censoring is unclear. The question of how to choose the quantile level α in practice to optimize the procedure seems intriguing as

well.

2.5 Censored rank independence screening

In biomedical studies, variables are often transformed to achieve uniformity across different platforms, and from time to time, outliers are observed in predictors. It is desirable for screening tools to possess invariance properties under transformations of variables ($\mathbf{X}_i$ or T_i) and robustness against outliers. As Kendall's τ, a widely used measure of correlation, is robust against heavy tailed distributions and invariant under monotonic transformations, researchers proposed Kendall's τ based screening methods ([16]). To accommodate survival data, [19] considered the concordance between failure time T and covariate X_j in the presence of censoring.

Define $\tau_j = \mathrm{pr}(X_{ji} > X_{ji'}, T_i > T_{i'}) - 1/4, (j = 1, \ldots, p,\ i = 1, \ldots, n)$. [19] advocated using τ_j as the marginal utility measure for ranking predictors for several reasons. First, τ_j measures the association between T and X_j and τ_j is 0 if T and X_j are independent. Second, τ_j is invariant to any monotonic transformations on T and X_j. Third, this rank based measure is robust against outliers in the data.

Let $\psi_j = \delta_{i'} I(X_{ij} > X_{i'j}, Y_i > Y_{i'})/G^2(Y_{i'})$. It can be easily shown that $\mathrm{E}(\psi_j) = \mathrm{pr}(X_{ij} > X_{i'j}, T_i > T_{i'})$. Thus, a natural estimate of τ_j is:

$$\hat{\tau}_j = \binom{n}{2}^{-1} \sum_{i<i'} \frac{\delta_{i'}}{\widehat{G}^2(Y_{i'})} I(X_{ij} > X_{i'j}, Y_i > Y_{i'}) - 1/4,$$

where $\widehat{G}(\cdot)$ is the Kaplan–Meier estimator of $G(t) = \mathrm{pr}(C_i \geq t)$.

Define the true set as

$$\mathcal{M} = \{j : \mathrm{pr}(T > t|\mathbf{X}) \text{ functionally depends on } X_j\}.$$

Then, it is estimated by a set of important predictors with large $\hat{\tau}_j$:

$$\widehat{\mathcal{M}} = \{j : |\hat{\tau}_j| > \lambda_n\},$$

where λ_n is a predefined threshold value. This procedure is called the censored rank independence screening (CRIS). [19] showed that $\hat{\tau}_j$ is a consistent estimator for τ_j. That is,

$$\mathrm{pr}\left(\max_j |\hat{\tau}_j - \tau| < c_6 n^{-\kappa}\right) \leq p\{2.5n\exp(-c_1 n) + 4\exp(-c_4 n^{1-2\kappa}) + 2.5n\exp(-c_2 n^{1-2\kappa})\},$$

for some positive constants c_1, c_2, c_4, and c_6. Moreover, when $\min_{j\in\mathcal{M}} |\mathrm{pr}(X_{1j} > X_{2j}, T_1 > T_2) - 1/4| \geq c_0 n^{-\kappa}$ for some $0 < \kappa < 1/2$ and $c_0 > 0$, taking $\lambda_n = c_7 n^{-\kappa}$ with $c_7 \leq c_0/2$ leads to

$$\mathrm{pr}(\mathcal{M} \subset \widehat{\mathcal{M}}) \geq 1 - s\{2.5n\exp(-c_1 n) + 4\exp(-c_4 n^{1-2k}) + 2.5n\exp(-c_2 n^{1-2\kappa})\} \to 1, \text{ as } n \to \infty,$$

where s is the size of $\mathcal{M}$.

CRIS is a model-free approach that enjoys the sure screening property. Moreover, the screening statistic is a U-statistic with a bounded kernel function. The large sample results hold even without the tail probability conditions. However, the computation of $\hat{\tau}_j$ requires the comparison of all possible pairs of samples. This exceedingly heavy computational burden may hamper its applicability when the sample size is large.

2.6 Survival impact index screening

QaSIS and CRIS are model-free screeners, but both employ the inverse probability of censoring weighting (IPCW), which may be unstable, especially when evaluated at large observed survival times. In addition, they may not capture the full-range impact of covariates on the overall survival since QaSIS focuses on a specific quantile level and CRIS relies on a summarized value of association. To more fully capture the overall influence of a covariate on the outcome distribution, [17] proposed a new metric called the survival impact index (SII), which evaluates the absolute deviation of the covariate-stratified survival distribution from the unstratified survival distribution.

Specifically, for each X_j, $j = 1, \ldots, p$, SII is defined as

$$\xi_j = \int_{t\in\mathcal{T}, x\in\mathcal{X}} W_\xi(t,x)|S(t|X_j > x) - S(t)|dxdt,$$

where $W_\xi(t,x)$ is a pre-determined weight function introduced to capture the covariate impact on either early or late survival. The authors argued that if, for at least one t and one x, the survival function stratified on $X_j > x$ differs from the unstratified survival function at t, then ξ_j will be non-zero under mild conditions. On the other hand, if T and X_j are independent, then $\xi_j = 0$. Hence, ξ_j serves as a sensible index for characterizing the importance of X_j in influencing the distribution of T.

To estimate ξ_j, [17] proposed to use

$$\hat{\xi}_j = \int_{t\in\mathcal{T}, x\in\mathcal{X}} W_\xi(t,x)|\hat{S}(t|X_j > x) - \hat{S}(t)|dxdt, \tag{8}$$

where $\hat{S}(t|X_j > x)$ is the Kaplan-Meier estimator based on sub-sample $X_j > x$ and $\hat{S}(t)$ is the Kaplan-Meier estimator for the survival function of T. Note that (8) indicates SII does not need to adopt IPCW to handle the random censoring.

The set of important predictors with large $\hat{\xi}_j$ is defined by $\widehat{\mathcal{M}} = \{j : \hat{\xi}_j > \lambda_n\}$. Under some regularity conditions, given $24/(n\mu\gamma^4) < \epsilon < 1$, [17] proved that the estimated survival impact index $\hat{\xi}_j$ is uniformly consistent to ξ_j, or

$$\text{pr}\left(\max_j |\hat{\xi}_j - \xi_j| > \epsilon\right) \le c_3 \exp(-nc_4\epsilon^2 - c_5 \log \epsilon), \tag{9}$$

where c_3, c_4, c_5, μ, and γ are some positive constants and n is sufficiently large. With the true set being

$$\mathcal{M} = \{j : \xi_j > 0\},$$

if $p = O(\exp(n^c))$ for some $0 < c < 1$ and $\min_{j\in\mathcal{M}} \xi_j > c_0 n^{-\alpha}$ for some constant $c_0 > 0$, $0 \le \alpha < (1-c)/2$, [17] proved that

$$\text{pr}(\mathcal{M} \subset \widehat{\mathcal{M}}) \ge 1 - c_3 s \exp\{-nc_4(c_0 - b)^2 n^{-2\alpha} - c_5 \log((c_0 - b)n^{-\alpha})\},$$

where s is the size of $\mathcal{M}$. It follows that the sure screening property holds by taking $\lambda_n = bn^{-\alpha}$ with $b \le c_0/2$. That is, $\text{pr}(\mathcal{M} \subset \widehat{\mathcal{M}}) \to 1$ as $n \to \infty$.

One potential challenge with SII is that (8) needs to be computed over a range of values for both T and X_j, which is often associated with high computation cost. The problem of how to select the weight function may warrant more research.

2.7 Integrated powered density screening

In a survival setting, nonparametric variable screeners have focused on discerning how each candidate variable influences survival functions. One way of detecting such influence is by studying the variability of survival functions for the subpopulations or strata defined by each variable. The difference patterns, however, may vary across covariates. Specifically, the differences may occur either during the early or late period in the follow-up due to disease-related characteristics. For example, survival differences between chemotherapy versus both chemotherapy and radiation treatment groups among childhood cancer patients may be more pronounced right after treatment, while survival differences between the EGFR mutation status among non-small cell lung cancer patients may be more obvious long after the onset of cancer. Therefore, screening approaches that rely on a single screening criterion may not be able to capture the complex difference patterns and may lead to false non-discovery. [10] proposed to consider the following integrated powered density (IPOD):

$$\int_0^t f^\gamma(s)ds,$$

where a power index γ (>0) inflates either early ($\gamma>1$) or late differences ($\gamma<1$) during the life span and thus gives more flexibility in detecting distributional differences. IPOD resembles the cumulative distribution function (CDF) and satisfies the basic properties of CDFs, except that it does not necessarily approach one when $t\to\infty$.

To derive the screening criterion, first consider a discrete X_j with R_j categories. The unique property of IPOD motivates the following marginal utility to detect distributional differences:

$$\mathcal{I}_j^{(\gamma)}=\max_{r_1,r_2\in\{1,\ldots,R_j\}}\sup_{t\in[0,\nu]}\left|\int_0^t f_{T|X_j}^{\gamma}(s|X_j=r_1)ds-\int_0^t f_{T|X_j}^{\gamma}(s|X_j=r_2)ds\right|, \tag{10}$$

where $f_{T|X_j}(s|X_j=r)$ denotes the conditional density function of T given $X_j=r$. Since $\mathcal{I}_j^{(\gamma)}=0$ if and only if T and X_j are independent, it serves as a measure of marginal utility for each X_j. The framework of IPOD is general by accommodating different γ's. For example, when $\gamma=1$, (10) is simply the classical Kolmogorov difference: $\max_{r_1,r_2\in\{1,\ldots,R_j\}}\sup_{t\in[0,\nu]}|F_{T|X_j}(t|X_j=r_1)-F_{T|X_j}(t|X_j=r_2)|$.

Denote by $h_n>0$ the bandwidth of a kernel function $K(\cdot)$. Then, (10) can be estimated by

$$\hat{\mathcal{I}}_j^{(\gamma)}=\max_{r_1,r_2\in\{1,\ldots,R_j\}}\sup_{t\in[0,\nu]}\left|\int_0^t \hat{f}_{T|X_j}^{\gamma}(s|X_j=r_1)ds-\int_0^t \hat{f}_{T|X_j}^{\gamma}(s|X_j=r_2)ds\right|,$$

where $\hat{S}(t)$ is the Kaplan–Meier estimator for the survival function of T,

$$\hat{f}_T(t)=\sum_i K((t-t_i)/h_n)(\hat{S}_T(t_{i-1})-\hat{S}_T(t_i)),$$

and the conditional density estimator $\hat{f}_{T|X_j}(t|X_j=r)$ can be obtained similarly as $\hat{f}_T(t)$ by restricting samples to $X_j=r$.

[10] defined the true important feature set as

$$\mathcal{M}=\{j: S(t|\mathbf{X}) \text{ functionally depends on } X_j \text{ for some } t\in(0,\infty)\},$$

which is estimated by $\widehat{\mathcal{M}}_1=\left\{j:\hat{\mathcal{I}}_j^{(\gamma)}>\lambda_n\right\}$ where $\lambda_n>0$. This procedure is referred to as

the IPOD screening.

When a covariate X_j is continuous, it can be discretized into R_j slices by using the percentiles of the empirical distribution of X_j. Suppose there are N different ways of slicing a continuous covariate X_j, denoted by Λ_{ju}, $u = 1, \ldots, N$ with each slicing Λ_{ju} containing R_{ju} intervals. To ensure there are enough sample within each slice for all slicing schemes, one may take $R_{ju} = 3, \ldots, [\log(n)]$, which gives $N = [\log(n) - 2]$ slicing schemes. Denoting $\hat{\mathcal{I}}^{(\gamma)}_{j,\Lambda_{ju}}$ as the IPOD screening statistic corresponding to the slicing scheme of Λ_{ju}, they proposed the following fused IPOD screening statistic that collects all information from each slice:

$$\tilde{\mathcal{I}}^{(\gamma)}_j = \sum_{u=1}^{N} \hat{\mathcal{I}}^{(\gamma)}_{j,\Lambda_{ju}},$$

which leads to the following screening criterion:

$$\widehat{\mathcal{M}}_2 = \left\{ j : \tilde{\mathcal{I}}^{(\gamma)}_j > \lambda_n \right\},$$

where $\lambda_n > 0$ is a pre-specified constant.

For large sample results, [10] stipulated the following assumptions. Let Λ_{juo} be the partition based on the theoretical quantiles $q_{ju(r)}$, $r = 0, \ldots, R_{ju}$ of X_j and $\mathcal{I}^{(\gamma)}_{jo} = \sum_{u=1}^{N} \mathcal{I}^{(\gamma)}_{j,\Lambda_{juo}}$. Assume that there exist a $c > 0$ and $0 < v < 1/2$ such that $\min_{j \in \mathcal{M}} \mathcal{I}^{(\gamma)}_{jo} \geq 2cn^{-v}$ for a specific γ.

When covariates include both continuous and discrete values, under the above conditions, [10] proved that

$$\mathrm{pr}(\mathcal{M} \subset \widehat{\mathcal{M}}_2) \geq 1 - O(Np \exp(-b_2 n^{1-3\kappa-2v-2\mu-2\rho} + \kappa \log n)),$$

where b_2 is a positive constant. For $0 < \alpha < 1 - 3\kappa - 2v - 2\mu - 2\rho$, if $N = O(\log n)$ and $\log p = O(n^{\alpha})$, the fused IPOD has the sure screening property.

IPOD enjoys the invariance property like other nonparametric screeners such as SII and CRIS, but it is more computationally efficient with increasing n. The performance of the method depends on how well the distribution function can be estimated on each covariate-defined stratum. Hence, it may not work well for small sample sizes. Moreover, since smoothing techniques are used to evaluate the density functions, the choice of bandwidth affects the accuracy of results.

§3 Numerical comparisons

All the reviewed papers were exemplified with intensive simulations. As the simulation setup differs across different papers, to make fair comparisons and give general guidelines for practical utility, we considered various simulation settings, under which the methods reviewed in Section 2 are all compared. For each configuration, a total of 100 simulated datasets were generated. We considered $n = 100$ and $n = 300$ and explored how the performance of the methods improved with sample size.

Example 1 (Nonlinear covariate-response relationship): The survival time was generated from $\log(T) = g_1(X_1) + g_2(X_2) + g_3(X_3) + g_4(X_4) + \epsilon$, where $g_1(x) = 5x$, $g_2(x) = -4x(1 - x)$, $g_3(x) = 10[\exp\{-3(x-1)^2\} + \exp\{-4(x-3)^2\}] - 1.5$, and $g_4(x) = 4\sin(2\pi x)$. The vector of

covariates $\mathbf{X}$ was generated from a multivariate normal distribution with mean 0, variance 1, and a first order autoregressive correlation structure, i.e., $\mathrm{cor}(X_j, X_{j'}) = \rho^{|j-j'|}$ $(j,\ j' = 1, \ldots, p)$. The error $\epsilon \sim N(0,1)$ is independent of $\mathbf{X}$. The censoring time C was generated from a 3-component normal mixture distribution $N(0,4) - N(5,1) + 0.5N(25,1)$. This example was adopted from [17].

Example 2 (Rayleigh model): The survival time was generated from a hazard function $h(t|\mathbf{X}) = 2t(|X_1|+|X_2|)$. All covariates were generated from the multivariate normal distribution with mean 0, variance 1, and an exchangeable correlation structure with the true parameter value of ρ. In this case, the marginal correlation between each of the active variables, X_1 and X_2, and the survival time is 0. The censoring time C was generated from a uniform distribution on $(0,3)$. This example was adopted from [10].

Example 3 (Linear transformation models): The survival time was generated from the class of linear transformation models

$$\log\{0.5(e^{2T}-1)\} = -\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X} + \epsilon,$$

where $\boldsymbol{\beta} = (-1, -0.9, \mathbf{0}_6^{\mathrm{T}}, 0.8, 1.0, \mathbf{0}_{p-10}^{\mathrm{T}})^{\mathrm{T}}$ and $\mathbf{X}$ were generated from a multivariate normal distribution with mean 0, variance 1, and a first order autoregressive correlation structure, i.e., $\mathrm{cor}(X_j, X_{j'}) = \rho^{|j-j'|}$ $(j,\ j' = 1, \ldots, p)$. The error was generated from the standard normal distribution and the censoring time was generated from a uniform distribution on $(0,3)$. This example was adopted from [19].

Example 4 (Discrete covariates): The survival time was generated from the proportional hazards model,

$$h(t|\mathbf{X}) = 0.1 \exp\left\{\sum_{j=1}^{p} \beta_j I(X_j \in \{2,3\})\right\},$$

where $\beta = (-\mathbf{1}_5, \mathbf{0}_{p-5}^{\mathrm{T}})^{\mathrm{T}}$. The covariates $x^\star$ underlying these discrete variables were generated from a multivariate normal distribution with mean 0 and a covariance matrix $\Sigma = (\sigma_{jj'})_{p\times p}$, $\sigma_{jj} = 1$ and $\sigma_{jj'} = 0.5$ for $j \neq j'$. For each j, x_j^* was further quarterized by its quartile values: the obtained quarterly variable $X_j = 1$ if x_j^* is less than the lower quartile, 2 if between the lower quartile and the median, 3 if between the median and the upper quartile, and 4 if else. The censoring time was generated from a uniform distribution on $(0,40)$. This example was adopted and modified from [10].

Example 5 (Hidden variables): The survival time was generated from a Cox model:

$$h(t|\mathbf{X}) = \exp(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}),$$

where $\mathbf{X}$ were generated from a multivariate normal distribution with mean 0, variance 1, and an equal correlation of 0.5 and $\boldsymbol{\beta} = (\mathbf{1}_5^{\mathrm{T}}, -2.5, \mathbf{0}_{p-6}^{\mathrm{T}})$. In this case, X_6 has a lower marginal utility than all the noise variables. The censoring time was generated from a uniform distribution on $(0,10)$. This example was adopted from [11].

Following the literature, we used three metrics as the criteria for comparisons, the minimum model size (MMS; the minimum number of variables that need to be selected in order to include

Table 1: Average runtime (seconds) of nonparametric screeners for Example 1 on a CPU with 2.9 GHz Intel Core i5 and 8GB of memory

(n, p)	CRIS	SII	IPOD
$(100, 1000)$	10.873	53.307	12.030
$(100, 10000)$	83.384	524.126	119.233
$(300, 1000)$	77.244	659.949	18.139
$(300, 10000)$	625.341	5607.038	175.598

all active variables), the true positive rate (TPR; the proportion of active variables selected in the first $[n/\log n]$ variables), and the probability of inclusion of the true model (PIT; the probability of all active variables selected in the first $[n/\log n]$ variables). MMS was reported as the median and TPR and PIT as the averages over 100 repetitions in Tables 1 and 2. Three conditioning sets $\mathcal{C}_1 = \{X_1\}$, $\mathcal{C}_2 = \{X_1, X_2\}$, and $\mathcal{C}_3 = \{X_1, \text{an inactive variable}\}$ were used for the conditional screening method ([11]), whenever appropriate. Many screeners assumed that the active variables are independent of the inactive variables, which is difficult to satisfy in practice. To explore how sensitive the competing methods are toward this assumption, we set the correlation coefficient ρ to be 0, 0.5, and 0.8 in each example, except in Example 5 wherein $\rho = 0.5$ was carefully chosen to generate a hidden variable.

Table 1 documents a comparison of computation time of nonparametric screening methods with various n and p. It shows that the computation time of CRIS increases non-proportionally with the sample size. SII is also computationally expensive as it needs to exhaustively search the ranges of both T and X_j. On the other hand, IPOD seems to be less impacted by the increasing n and p.

Tables 2 and 3 demonstrate that when the covariates and the response have a nonlinear relationship (Example 1), IPOD, SII, and CoxCS outperformed the other methods. In Example 2 where the marginal correlation between each active variable and the survival time is 0, PSIS, FAST, and CRIS performed poorly. With the linear transformation model in Example 3, all methods easily identified active variables, especially when the sample size was large. When the covariates are discrete (Example 4), IPOD, which is designed to accommodate discrete variables, outperformed all the other methods. In Example 5, all marginal screeners had difficulties recruiting a "marginally unimportant but jointly important" variable X_6. Only the conditional screening approach, CoxCS, was able to detect the hidden variable. In addition, the performance of IPOD varied with different γ. The conditional approach might fail when the dependence among covariates is too strong, as shown in Example 3 with $\rho = 0.8$.

Table 2: Comparisons of competing methods when $(n, p) = (100, 1000)$

	MMS	**TPR**	**PIT**	**MMS**	**TPR**	**PIT**	**MMS**	**TPR**	**PIT**
Example 1		$\rho = 0$			$\rho = 0.5$			$\rho = 0.8$	
PSIS	660 (391)	0.46	0.00	744 (344)	0.27	0.01	580 (505)	0.22	0.01
FAST	601 (372)	0.18	0.00	765 (262)	0.17	0.00	672 (413)	0.08	0.00
CoxCS (x_1)	668 (361)	0.45	0.00	621 (518)	0.50	0.02	333 (618)	0.59	0.10
CoxCS (x_1, x_2)	677 (425)	0.54	0.00	488 (522)	0.57	0.02	390 (551)	0.59	0.06
CoxCS (x_1, an inactive)	694 (379)	0.44	0.00	604 (503)	0.51	0.02	391 (588)	0.60	0.11
QaSIS	702 (434)	0.22	0.00	618 (386)	0.22	0.00	381 (394)	0.31	0.00
CRIS	722 (344)	0.52	0.00	746 (399)	0.27	0.00	696 (424)	0.22	0.00
SII	509 (375)	0.48	0.00	560 (377)	0.41	0.00	231 (332)	0.56	0.08
IPOD ($\gamma = 0.8$)	388 (420)	0.63	0.02	441 (517)	0.56	0.00	160 (221)	0.62	0.08
IPOD ($\gamma = 1$)	402 (378)	0.61	0.03	457 (496)	0.52	0.01	176 (217)	0.59	0.06
IPOD ($\gamma = 1.2$)	456 (415)	0.57	0.02	495 (462)	0.45	0.00	216 (284)	0.52	0.05
Example 2		$\rho = 0$			$\rho = 0.5$			$\rho = 0.8$	
PSIS	764 (272)	0.01	0.00	733 (423)	0.02	0.00	699 (291)	0.01	0.00
FAST	757 (278)	0.01	0.00	719 (435)	0.02	0.00	694 (306)	0.01	0.00
CoxCS (x_1)	433 (539)	0.52	0.03	454 (492)	0.50	0.01	444 (505)	0.51	0.02
CoxCS (x_1, an inactive)	445 (491)	0.52	0.03	477 (540)	0.50	0.01	415 (531)	0.50	0.01
QaSIS	514 (372)	0.01	0.00	472 (489)	0.03	0.01	417 (450)	0.03	0.00
CRIS	705 (278)	0.01	0.00	710 (401)	0.00	0.00	658 (390)	0.02	0.00
SII	352 (307)	0.03	0.01	316 (315)	0.02	0.00	237 (239)	0.07	0.02
IPOD ($\gamma = 0.8$)	133 (265)	0.41	0.14	130 (237)	0.40	0.13	58 (172)	0.49	0.23
IPOD ($\gamma = 1$)	125 (279)	0.42	0.17	132 (238)	0.38	0.15	65 (142)	0.50	0.21
IPOD ($\gamma = 1.2$)	148 (270)	0.40	0.15	147 (232)	0.36	0.11	79 (148)	0.52	0.23
Example 3		$\rho = 0$			$\rho = 0.5$			$\rho = 0.8$	
PSIS	7 (21)	0.93	0.71	4 (0)	1.00	1.00	5 (2)	1.00	0.99
FAST	7 (28)	0.92	0.70	4 (0)	1.00	1.00	5 (2)	1.00	0.99
CoxCS (x_1)	5 (6)	0.97	0.88	4 (2)	0.98	0.92	27 (120)	0.87	0.48
CoxCS (x_1, x_2)	4 (1)	0.99	0.95	4 (0)	1.00	1.00	4 (0)	1.00	1.00
CoxCS (x_1, an inactive)	6 (6)	0.97	0.88	5 (2)	0.98	0.92	8 (11)	0.94	0.78
QaSIS	267 (261)	0.17	0.00	111 (139)	0.30	0.01	126 (110)	0.30	0.01
CRIS	7 (9)	0.95	0.79	4 (0)	1.00	1.00	5 (2)	1.00	0.99
SII	113 (220)	0.56	0.08	18 (33)	0.86	0.54	23 (50)	0.82	0.50
IPOD ($\gamma = 0.8$)	121 (194)	0.61	0.05	24 (58)	0.86	0.49	25 (52)	0.83	0.45
IPOD ($\gamma = 1$)	88 (136)	0.70	0.21	10 (20)	0.92	0.70	13 (17)	0.92	0.73
IPOD ($\gamma = 1.2$)	68 (112)	0.73	0.24	7 (9)	0.95	0.81	9 (11)	0.96	0.83
Example 4		$\rho = 0$			$\rho = 0.5$			$\rho = 0.8$	
PSIS	863 (194)	0.01	0.00	876 (156)	0.03	0.00	855 (209)	0.03	0.00
FAST	845 (211)	0.03	0.00	857 (166)	0.05	0.00	820 (234)	0.05	0.00
CoxCS (x_1)	851 (237)	0.22	0.00	833 (195)	0.23	0.00	783 (244)	0.23	0.00
CoxCS (x_1, x_2)	820 (244)	0.41	0.00	733 (303)	0.42	0.00	737 (316)	0.41	0.00
CoxCS (x_1, an inactive)	827 (245)	0.21	0.00	830 (240)	0.22	0.00	819 (244)	0.22	0.00
QaSIS	770 (330)	0.09	0.00	785 (304)	0.10	0.00	747 (297)	0.11	0.00
CRIS	872 (176)	0.01	0.00	863 (159)	0.03	0.00	879 (206)	0.01	0.00
SII	489 (393)	0.10	0.00	435 (292)	0.18	0.00	349 (258)	0.23	0.00
IPOD($\gamma = 0.8$)	549 (382)	0.15	0.00	457 (296)	0.20	0.00	401 (315)	0.21	0.00
IPOD ($\gamma = 1$)	411 (412)	0.27	0.00	363 (316)	0.32	0.00	286 (312)	0.33	0.00
IPOD ($\gamma = 1.2$)	398 (429)	0.33	0.01	294 (279)	0.35	0.00	242 (273)	0.41	0.00
Example 5		$\rho = 0$			$\rho = 0.5$			$\rho = 0.8$	
PSIS	-	-	-	1000 (0)	0.53	0.00	-	-	-
FAST	-	-	-	1000 (3)	0.25	0.00	-	-	-
CoxCS (x_1)	-	-	-	131 (199)	0.71	0.13	-	-	-
CoxCS (x_1, x_2)	-	-	-	93 (157)	0.80	0.20	-	-	-
CoxCS (x_1, an inactive)	-	-	-	108 (131)	0.74	0.14	-	-	-
QaSIS	-	-	-	895 (254)	0.08	0.00	-	-	-
CRIS	-	-	-	1000 (0)	0.54	0.00	-	-	-
SII	-	-	-	992 (42)	0.31	0.00	-	-	-
IPOD ($\gamma = 0.8$)	-	-	-	782 (351)	0.11	0.00	-	-	-
IPOD ($\gamma = 1$)	-	-	-	945 (122)	0.22	0.00	-	-	-
IPOD ($\gamma = 1.2$)	-	-	-	985 (57)	0.26	0.00	-	-	-

Table 3: Comparisons of competing methods when $(n, p) = (300, 1000)$

	MMS	**TPR**	**PIT**	**MMS**	**TPR**	**PIT**	**MMS**	**TPR**	**PIT**
Example 1		$\rho = 0$			$\rho = 0.5$			$\rho = 0.8$	
PSIS	571 (407)	0.61	0.01	745 (329)	0.35	0.00	448 (586)	0.46	0.11
FAST	538 (443)	0.56	0.04	803 (285)	0.29	0.00	569 (473)	0.32	0.00
CoxCS (x_1)	598 (508)	0.64	0.02	439 (525)	0.64	0.09	50 (187)	0.84	0.52
CoxCS (x_1, x_2)	488 (435)	0.67	0.03	454 (600)	0.66	0.09	350 (568)	0.67	0.12
CoxCS(x_1, an inactive)	602 (491)	0.64	0.02	487 (524)	0.63	0.10	71 (195)	0.81	0.45
QaSIS	645 (453)	0.50	0.00	590 (434)	0.52	0.02	290 (454)	0.66	0.10
CRIS	673 (374)	0.57	0.01	743 (282)	0.40	0.00	385 (599)	0.51	0.14
SII	423 (442)	0.68	0.07	333 (420)	0.65	0.01	36 (120)	0.87	0.56
IPOD ($\gamma = 0.8$)	76 (142)	0.85	0.39	42 (97)	0.90	0.59	5 (8)	0.99	0.95
IPOD ($\gamma = 1$)	115 (152)	0.81	0.25	57 (133)	0.87	0.47	9 (12)	0.97	0.89
IPOD ($\gamma = 1.2$)	161 (186)	0.80	0.20	82 (143)	0.84	0.37	12 (23)	0.96	0.84
Example 2		$\rho = 0$			$\rho = 0.5$			$\rho = 0.8$	
PSIS	739 (361)	0.01	0.00	719 (315)	0.04	0.00	737 (421)	0.06	0.01
FAST	725 (374)	0.03	0.00	705 (324)	0.06	0.00	731 (441)	0.07	0.01
CoxCS (x_1)	484 (548)	0.52	0.04	537 (504)	0.56	0.11	404 (466)	0.54	0.09
CoxCS (x_1, an inactive)	486 (557)	0.52	0.04	465 (428)	0.54	0.08	421 (499)	0.52	0.05
QaSIS	120 (163)	0.35	0.14	83 (134)	0.58	0.33	48 (102)	0.70	0.51
CRIS	734 (354)	0.02	0.00	712 (329)	0.05	0.00	701 (316)	0.04	0.00
SII	45 (53)	0.76	0.58	28 (37)	0.85	0.76	12 (34)	0.91	0.85
IPOD ($\gamma = 0.8$)	5 (9)	0.96	0.92	2 (2)	0.97	0.96	2 (4)	0.98	0.97
IPOD ($\gamma = 1$)	5 (12)	0.96	0.93	2 (2)	0.97	0.96	2 (3)	0.99	0.98
IPOD ($\gamma = 1.2$)	5 (12)	0.94	0.89	3 (3)	0.97	0.96	3 (4)	0.98	0.97
Example 3		$\rho = 0$			$\rho = 0.5$			$\rho = 0.8$	
PSIS	4 (0)	1.00	0.99	4 (0)	1.00	1.00	4 (0)	1.00	1.00
FAST	4 (0)	1.00	0.99	4 (0)	1.00	1.00	4 (1)	1.00	1.00
CoxCS (x_1)	4 (0)	1.00	0.99	4 (1)	1.00	1.00	12 (29)	0.95	0.81
CoxCS (x_1, x_2)	4 (0)	1.00	1.00	4 (0)	1.00	1.00	4 (0)	1.00	1.00
CoxCS (x_1, an inactive)	5 (0)	1.00	0.99	5 (1)	1.00	1.00	7 (1)	1.00	1.00
QaSIS	23 (21)	0.95	0.84	9 (13)	0.99	0.97	9 (10)	0.98	0.96
CRIS	4 (0)	1.00	1.00	4 (0)	1.00	1.00	4 (0)	1.00	1.00
SII	6 (6)	0.98	0.94	4 (0)	1.00	1.00	5 (2)	1.00	1.00
IPOD ($\gamma = 0.8$)	4 (1)	0.99	0.96	4 (0)	1.00	1.00	4 (1)	1.00	1.00
IPOD ($\gamma = 1$)	4 (1)	0.99	0.96	4 (0)	1.00	1.00	4 (1)	1.00	1.00
IPOD ($\gamma = 2.3$)	4 (0)	0.99	0.98	4 (0)	1.00	1.00	4 (1)	1.00	1.00
Example 4		$\rho = 0$			$\rho = 0.5$			$\rho = 0.8$	
PSIS	870 (163)	0.05	0.00	842 (212)	0.05	0.00	889 (209)	0.06	0.00
FAST	849 (191)	0.10	0.00	825 (241)	0.10	0.00	872 (236)	0.09	0.00
CoxCS (x_1)	831 (201)	0.27	0.00	807 (258)	0.26	0.00	812 (256)	0.24	0.00
CoxCS (x_1, x_2)	807 (228)	0.45	0.00	781 (346)	0.44	0.00	834 (261)	0.43	0.00
CoxCS (x_1, inactive)	836 (197)	0.27	0.00	799 (292)	0.25	0.00	842 (241)	0.24	0.00
QaSIS	636 (422)	0.36	0.00	488 (545)	0.40	0.02	531 (499)	0.41	0.00
CRIS	849 (192)	0.05	0.00	875 (144)	0.03	0.00	913 (109)	0.04	0.00
SII	103 (110)	0.73	0.17	42 (80)	0.90	0.57	36 (55)	0.91	0.64
IPOD ($\gamma = 0.8$)	94 (147)	0.77	0.26	81 (116)	0.83	0.34	50 (90)	0.89	0.56
IPOD ($\gamma = 1$)	39 (65)	0.90	0.60	28 (44)	0.94	0.72	22 (41)	0.95	0.74
IPOD ($\gamma = 1.2$)	28 (47)	0.94	0.71	18 (26)	0.96	0.84	19 (32)	0.96	0.81
Example 5		$\rho = 0$			$\rho = 0.5$			$\rho = 0.8$	
PSIS	-	-	-	1000 (0)	0.83	0.00	-	-	-
FAST	-	-	-	1000 (0)	0.71	0.00	-	-	-
CoxCS (x_1)	-	-	-	8 (11)	0.98	0.91	-	-	-
CoxCS (x_1, x_2)	-	-	-	6 (2)	1.00	0.98	-	-	-
CoxCS (x_1, inactive)	-	-	-	8 (3)	0.99	0.97	-	-	-
QaSIS	-	-	-	990 (49)	0.27	0.00	-	-	-
CRIS	-	-	-	1000 (0)	0.82	0.00	-	-	-
SII	-	-	-	1000 (0)	0.75	0.00	-	-	-
IPOD ($\gamma = 0.8$)	-	-	-	990 (59)	0.57	0.00	-	-	-
IPOD ($\gamma = 1$)	-	-	-	1000 (1)	0.69	0.00	-	-	-
IPOD ($\gamma = 1.2$)	-	-	-	1000 (0)	0.71	0.00	-	-	-

§4 Discussion

The aim of this paper is to provide a selective overview on feature screening for ultrahigh-dimensional survival data. We study the justifications of some commonly used methods, paying attention to the motivation and rationale of the screening statistics. Under the same simulation settings, we have numerically compared all the reviewed methods and commented on their practical utility.

With the advent of the biomedical big data era, variable screening for ultrahigh dimensional survival data has been rapidly evolving; see some recent works in [26], [22], [24], and [23]. Variable screening is becoming a standard and indispensable analytical tool for ultrahigh dimensional data analysis, especially when outcomes are subject to censoring.

Even with the discussed progress, from our perspectives, there are several open problems that may attract researchers' attention. First, as the majority of works in this field have focused on right-censored survival times, it would be of interest to extend the works to accommodate more complex censoring mechanisms, including left censoring, interval censoring and left/right truncation.

Second, most authors explicitly or implicitly assumed partial orthogonality ([7]). That is, they assumed the active variables and inactive variables are independent. This assumption is often violated in practice, as variables tend to be dependent, no matter whether they are active or inactive. It is worth investigating how to relax this condition in the theoretical development or to design a more efficient screening method for dependent predictors.

Third, as shown in the numerical studies, variable screening methods do not necessarily agree with each other as they tend to select different variables. Hence, a natural question is how to integrate the results from these different screening methods, while controlling false positives as well as false negatives.

Fourth, most screening methods developed so far focus on independent survival times. Generalizing the results to accommodate more complex survival settings, including multi-state models and competing risks, semi-competing risks, multivariate survival, etc., remains an open problem.

Finally, existing variable screening methods have often overlooked the useful information on covariates with similar functionality or spatial proximity. Partitioning biomarkers into smaller groups according to biological knowledge or other useful information may facilitate feature selection. Recently, a partition-based screening method, which accounts for such grouping, has been proposed in a generalized linear regression setting by [14]. It is worth investigating its extension to the survival setting.

Acknowledgement. We thank Dr. Jialiang Li for providing the code for the survival impact index screening and Ms. Martina Fu for proofreading the manuscript.

References

[1] E Barut, J Q Fan, A Verhasselt. *Conditional sure independence screening*, J Amer Statist Assoc, 2016, 111(515): 1266-1277.

[2] J Bradic, J Q Fan, J C Jiang. *Regularization for Cox's proportional hazards model with NP-dimensionality*, Ann Statist, 2011, 39(6): 3092-3120.

[3] J Q Fan, Y Feng, Y C Wu. *High-dimensional variable selection for Cox's proportional hazards model*, In: *IMS Collections 6, Borrowing Strength: Theory Powering Applications - A Festschrift for Lawrence D. Brown*, 2010, 70-86.

[4] J Q Fan, R Z Li. *Variable selection for Cox's proportional hazards model and frailty model*, Ann Statist, 2000, 30(1): 74-99.

[5] J Q Fan, J Lv. *Sure independence screening for ultrahigh dimensional feature space (with discussion)*, J Roy Statist Soc B, 2008, 70(5): 849-911.

[6] J Q Fan, R Samworth, Y C Wu. *Ultrahigh dimensional feature selection: beyond the linear model*, J Mach Learn Res, 2009, 10: 2013-2038.

[7] J Q Fan, R Song. *Sure independence screening in generalized linear models with NP-dimensionality*, Ann Statist, 2010, 38(6): 3567-3604.

[8] A Gorst-Rasmussen, T Scheike. *Independent screening for single-index hazard rate models with ultrahigh dimensional features*, J Roy Statist Soc B, 2013, 75(2): 217-245.

[9] X M He, L Wang, H G Hong. *Quantile-adaptive model-free variable screening for high-dimensional heterogeneous data*, Ann Statist, 2013, 41(1): 342-369.

[10] H G Hong, X R Chen, D C Christiani, Y Li. *Integrated powered density: screening ultrahigh dimensional covariates with survival outcomes*, Biometrics, in press.

[11] H G Hong, J Kang, Y Li. *Conditional screening for ultra-high dimensional covariates with survival outcomes*, Lifetime Data Anal, 2016, https://doi.org/10.1007/s10985-016-9387-7

[12] H G Hong, L Wang, X M He. *A data-driven approach to conditional screening of high dimensional variables*, Stat, 2016, 5(1): 200-212.

[13] J Huang, T N Sun, Z L Ying, Y Yu, C-H Zhang. *Oracle inequalities for the Lasso in the Cox model*, Ann Statist, 2013, 41(3): 1142-1165.

[14] J Kang, H G Hong, Y Li. *Partition-based ultrahigh-dimensional variable screening*, Biometrika, 2017, https://doi.org/10.1093/biomet/asx052

[15] S C Kong, B Nan. *Non-asymptotic oracle inequalities for the high-dimensional Cox regression via Lasso*, Statist Sinica, 2014, 24: 25-42.

[16] G R Li, H Peng, J Zhang, L X Zhu. *Robust rank correlation based screening*, Ann Statist, 2012, 40: 1846-1877.

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

[18] D Y Lin, Z L Ying. *Semiparametric analysis of the additive risk model*, Biometrika, 1994, 81(1): 61-71.

[19] R Song, W B Lu, S G Ma, X J Jeng. *Censored rank independence screening for high-dimensional survival data*, Biometrika, 2014, 101(4): 799-814.

[20] R J Tibshirani. *The lasso method for variable selection in the Cox model*, Stat Med, 1997, 16(4): 385-395.

[21] R J Tibshirani. *Univariate shrinkage in the Cox model for high dimensional data*, Stat Appl Genet Mol Biol, 2009, 8(1): 3498-3528.

[22] X D Yan, N S Tang, X Q Zhao. *The Spearman rank correlation screening for ultrahigh dimensional censored data*, eprint arXiv:1702.02708.

[23] G R Yang, Y Yu, R Z Li, A Buu. *Feature screening in ultrahigh dimensional Cox's model*, 2016, Statist Sinica, 26: 881-901.

[24] M Yue, J L Li. *Improvement screening for ultra-high dimensional data with censored survival outcomes and varying coefficients*, Int J Biostat, 2017, 13(1), https://doi.org/10.1515/ijb-2017-0024

[25] H H Zhang, W B Lu. *Adaptive Lasso for Cox's proportional hazards model*, Biometrika, 2007, 94(3): 691-703.

[26] J Zhang, G S Yin, Y Y Liu, Y S Wu. *Censored cumulative residual independent screening for ultrahigh-dimensional survival data*, Lifetime Data Anal, 2017, https://doi.org/10.1007/s10985-017-9395-2

[27] S D Zhao, Y Li. *Principled sure independence screening for Cox models with ultra-high-dimensional covariates*, J Multivariate Anal, 2012, 105(1): 397-411.

[28] S D Zhao, Y Li. *Score test variable screening*, Biometrics, 2014, 70(4): 862-871.

[29] H Zou. *A note on path-based variable selection in the penalized proportional hazards model*, Biometrika, 2008, 95: 241-247.

[1] Department of Statistics and Probability, Michigan State University, East Lansing, MI 48824, U.S.A.

[2] Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, U.S.A.
Email: yili@umich.edu