Annals of the Institute of Statistical Mathematics (2022) 74:379–397
https://doi.org/10.1007/s10463-021-00801-7

Surrogate-variable-based model-free feature screening for survival data under the general censoring mechanism

Jing Zhang[1] · Qihua Wang[2,3] · Xuan Wang[4]

Received: 20 May 2020 / Revised: 26 March 2021 / Accepted: 12 April 2021 /
Published online: 3 June 2021

Abstract

Feature screening has been seen as the first step in analyzing the ultrahigh-dimensional data with the censored survival time. In this article, we develop a surrogate-variable-based model-free feature screening approach for the censored data under the general censoring mechanism, where the censoring variable may depend on the survival variable and the covariates. This approach is developed by finding some observable variables whose active covariates contain the active covariates of the survival variable as a subset, respectively. Then, any existing model-free feature screening method with the sure screening property for full data can be applied to estimating the sets of the active covariates of the observable variables and hence the set of the active covariates of the survival variable. The sure screening property of the proposed approach is established, and its finite sample performances are demonstrated through some simulations. Further, we illustrate the proposed approach by analyzing two real datasets.

Keywords Feature screening · Model-free · Sure screening property · Survival data · Ultrahigh dimensionality

✉ Qihua Wang
qhwang@amss.ac.cn

Jing Zhang
jingzhang@amss.ac.cn

Xuan Wang
wangxuan209@hotmail.com

[1] School of Statistics and Mathematics, Shanghai Lixin University of Accounting and Finance, Shanghai 201209, China

[2] Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China

[3] School of Statistics and Mathematics, Zhejiang Gongshang University, Zhejiang 310018, China

[4] School of Mathematical Sciences, Zhejiang University, Zhejiang 310018, China

1 Introduction

In the analysis of ultrahigh-dimensional data, feature screening has become indispensable and received much attention in recent literatures. A crowd of model-based and model-free variable screening methods for fully observed outcomes have been proposed. For example, Fan and Lv (2008) proposed sure independence screening (SIS) methods based on the marginal Pearson correlation for the linear regression model, which was further extended to generalized linear models by Fan and Song (2010); Chang et al. (2013) developed a screening for generalized linear models based on the marginal empirical likelihood ratio; Zhu et al. (2011) proposed a sure independent ranking and screening (SIRS) approach for ultrahigh-dimensional multi-index models; Fan et al. (2011) presented a nonparametric independence screening (NIS) procedure based on the B-spline approximation for ultrahigh-dimensional additive models; He et al. (2013) gave a quantile-adaptive model-free screening means (QA-SIS) for the heteroscedastic model, which further improves the robustness of NIS; Li et al. (2012a) used Kendall τ correlation to replace the Pearson correlation in marginal correlation screening for the semiparametric single-index model with a monotone link function; Li et al. (2012b) similarly employed the distance correlation to replace Pearson correlation in marginal correlation screening, which is a model-free screening approach (DC-SIS); Mai and Zou (2015) presented a model-free variable screening method, called fused Kolmogorov filter basing on the Kolmogorov-Smirnov test statistic; Cui et al. (2015) developed a feature screening procedure based on the empirical conditional distribution function, which avoids the complex numerical optimization; Pan et al. (2019) gave a generic nonparametric sure independence screening procedure (BCor-SIS) on the basis of a universal dependence measure: Ball correlation, which can work for more scenarios under less restrictive assumptions.

These methods mentioned above have their own advantages. However, they mainly focus on the feature screening study for the fully observed data and may not be directly applied to the time-to-event survival data, which has widespread applications in biomedical research and other follow-up studies. For example, in an ultrahigh-dimensional gene expression cancer dataset, the primary interest is to find genes that are active and predictive for the survival time of patients. To handle this problem, some other screening methods for censored data have been developed. Gorst-Rasmussen and Scheike (2013) developed the feature aberration at survival times (FAST) screening procedure for the single-index hazard rate model, which requires the censoring mechanism to be partially random in the sense of depending on inactive covariates of the survival variable for obtaining the sure screening property. He et al. (2013) applied directly QA-SIS to the heterogeneous censored data by using the inverse probability weighting technique, and Song et al. (2014) suggested a censored rank independence screening (CRIS) based on an inverse probability-of-censoring weighted Kendall τ. However, these two methods both need to estimate the censoring probability under the complete random censoring (CRC) assumption, where the censoring variable is

independent of the survival variable and all covariates. Under the commonly used random censoring (RC) mechanism, where the censoring variable is independent of the survival variable given all covariates, more screening methods have also been proposed for the censored data problem. Fan et al. (2010) extended SIS to the Cox's proportional hazard model, but without yet the theoretical guarantee. Zhao and Li (2012) proposed the principled sure independence screening (PSIS) method based on the standardized marginal maximum partial likelihood estimators, and Yang et al. (2016) proposed the sure joint screening (SJS) method based on the joint likelihood of potential active predictors. Although both methods provided theoretical proofs, they was only suitable for the posited Cox model. Besides, He et al. (2013) further relaxed the CRC mechanism with the RC mechanism. Recently, Chen et al. (2018) presented two model-free screening approaches based on the robust distance correlation (Zhong et al. 2016). Nevertheless, both of them depend on the Kaplan-Meier estimator and hence more additional conditions are needed to make sure that the Kaplan-Meier estimator is well behaved. Since survival time cannot be observed completely under random censorship, these methods hence use the values of Kaplan-Meier estimator at every censored observation to replace their value at every survival time, which may be lead to bias. Li et al. (2016) proposed a survival impact index (SII) screener, and Liu et al. (2018) proposed a screener based on an appropriate Kolmogorov-Smirnov measure.

It is noted that all the existing feature screening approaches for censored survival data are developed under CRC or RC mechanism. In practice, however, it is hard to verify whether a censoring mechanism is CRC or RC. Actually, in many cases, the censoring may depend on the survival time variable and the covariates. This is a very general censoring (GC) mechanism, which include CRC and RC censoring mechanisms as special cases. Under the GC mechanism, the preceding feature screening methods may not be applicable. This is because these approaches depends on the assumed censoring mechanism heavily. For example, Kaplan-Meier estimator used in these approaches does not work well under the more general censoring mechanism. As Leung et al. (1997) pointed out: "the Kaplan-Meier estimator may overestimate the survival function if the survival time and the censoring time are positively correlated, and underestimate the survival function if the times are negatively correlated." This motivates us to develop the feature screening under the GC mechanism, which includes these censoring mechanisms considered in literature as special cases.

In this paper, based on the GC mechanism, we develop a model-free feature screening approach by proving some observable variables whose active covariates respectively contain the active covariates of the survival variable as a subset. Then, any available model-free feature screening methods for full data can be applied to estimating the sets of the active covariates of the observable variables and hence the set of the active covariates of the survival variable. The sure screening property can be kept as long as the used screening method for full data is of sure screening property. This method need to find at least a suitable observable variable, and is hence called surrogate-variable-based feature screening. The procedure enjoys several appealing merits, which are explained after Theorem 2.

The rest of this article is organized as follows. We present the results and the theoretical properties of the proposed feature screening approach in Section 2. Some simulation studies are conducted to evaluate the finite sample performances of the proposed approach in Section 3. In Section 4, two real data examples are analyzed using a mantle cell lymphoma data and a breast cancer data to illustrate the proposed approach. We also give a conclusion in Section 5. All of the technical proofs are provided in the Appendix.

2 Main results and methodologies

Let T and $\mathbf{X} = (X_1, \ldots, X_p)^T$ be the survival time variable and p-dimensional covariate vector, respectively. Suppose T is censored by the censoring variable C, and denote the observed response variable by $Y = \min(T, C)$ and the censoring indicator by $\delta = I(T \leq C)$, where $I(\cdot)$ is the indicator function.

Let τ denote the maximum follow-up time. Without any specified model, we define the index sets of the active covariates for T, δ given $T > t$ and C given $T > t$ by

$$\mathcal{A}(T|\mathbf{X}) = \{k : pr(T > t|\mathbf{X}) \text{ depends functionally on } X_k \text{ for some } t \in [0, \tau)\},$$
$$\mathcal{A}^*(\delta|\mathbf{X}) = \{k : pr(\delta = 1|T > t, \mathbf{X}) \text{ depends functionally on } X_k \text{ for some } t \in [0, \tau)\},$$
$$\mathcal{A}^*(C|\mathbf{X}) = \{k : pr(C > t|T > t, \mathbf{X}) \text{ depends functionally on } X_k \text{ for some } t \in [0, \tau)\},$$

respectively. Similarly, we can define $\mathcal{A}(C|\mathbf{X})$, $\mathcal{A}(Y|\mathbf{X})$ and $\mathcal{A}(\delta T|\mathbf{X})$, the index sets of the active covariates for C, Y and the product δT of T and δ, respectively. Our goal here is to recover the set of active variables $\mathcal{A}(T|\mathbf{X})$ on the basis of the sparsity assumption, which only a small number of covariates actually contribute to T.

Throughout this paper, we assume the GC mechanism. Under the GC mechanism, Y and δ are both likely to provide information which covariates are active to T. Hence, we expect to develop the feature screening approach based on observations of (Y, δ). Prior to this, the essential lemma deserves first attention. We first list the following conditions.

Condition 1. For any $k \in \mathcal{A}(T|\mathbf{X}) \cap \mathcal{A}^*(C|\mathbf{X})$, the product of $pr(T > t|\mathbf{X})$ and $pr(C > t|T > t, \mathbf{X})$ depends functionally on the covariate X_k for some $t \in [0, \tau)$.

Condition 2. For any $k \in \mathcal{A}(T|\mathbf{X}) \cap \mathcal{A}^*(\delta|\mathbf{X})$, the product of $pr(T > t|\mathbf{X})$ and $pr(\delta = 1|T > t, \mathbf{X})$ depends functionally on the covariate X_k for some $t \in [0, \tau)$.

Conditions 1 and *2* require that the product of the two conditional distribution functions does not cancel their mutual active covariates, respectively, which aims at ensuring that $pr(Y > t|\mathbf{X})$ and $pr(\delta T > t|\mathbf{X})$ depend functionally on these mutual active covariates. These are very weak conditions, which are widely satisfied in the practical application of censored data. It is easy to verify the two conditions in the case where $pr(T > t|\mathbf{X} = x)$, $pr(C > t|T > t, \mathbf{X} = x)$ and $pr(\delta = 1|T > t, \mathbf{X} = x)$ are differentiable on x.

Lemma 1 *Under Conditions 1 and 2, we have*

(i) $\mathcal{A}(Y|\mathbf{X}) = \mathcal{A}(T|\mathbf{X}) \cup \mathcal{A}^*(C|\mathbf{X})$,
(ii) $\mathcal{A}(\delta T|\mathbf{X}) = \mathcal{A}(T|\mathbf{X}) \cup \mathcal{A}^*(\delta|\mathbf{X})$.

Lemma 1 shows that $\mathcal{A}(T|\mathbf{X})$ is a subset of $\mathcal{A}(Y|\mathbf{X})$ and $\mathcal{A}(\delta T|\mathbf{X})$, respectively. Hence, we can directly apply any available model-free feature screening procedures for fully observed data to censored data based on fully observed variables $(Y, \mathbf{X})$ or $(\delta T, \mathbf{X})$. Denote the corresponding estimators $\hat{\mathcal{A}}_1(T|\mathbf{X}) = \hat{\mathcal{A}}(Y|\mathbf{X})$ and $\hat{\mathcal{A}}_2(T|\mathbf{X}) = \hat{\mathcal{A}}(\delta T|\mathbf{X})$, where $\hat{\mathcal{A}}(Y|\mathbf{X})$ and $\hat{\mathcal{A}}(\delta T|\mathbf{X})$ are respectively the estimator of $\mathcal{A}(Y|\mathbf{X})$ and $\mathcal{A}(\delta T|\mathbf{X})$ gained by some model-free feature screening method for full data. And the sure screening property can be kept as long as the model-free feature screening approach for full data is of sure screening property.

Theorem 1 (Sure Screening Property) *Assume the conditions of Lemma 1 and the corresponding conditions for the sure screening property of the used model-free feature screening method for full data, we then have*

(i) $\lim_{n\to\infty} pr\{\mathcal{A}(T|\mathbf{X}) \subseteq \hat{\mathcal{A}}_1(T|\mathbf{X})\} = 1$,
(ii) $\lim_{n\to\infty} pr\{\mathcal{A}(T|\mathbf{X}) \subseteq \hat{\mathcal{A}}_2(T|\mathbf{X})\} = 1$.

Up to now, we give two estimators for the active variable set $\mathcal{A}(T|\mathbf{X})$. Although the two estimators both possess the sure screening property, it can be seen from Lemma 1 that these estimators are somewhat conservative since this screening procedure uses the results $\mathcal{A}(T|\mathbf{X}) \subseteq \mathcal{A}(Y|\mathbf{X})$ and $\mathcal{A}(T|\mathbf{X}) \subseteq \mathcal{A}(\delta T|\mathbf{X})$ for the estimators, respectively. Thus, it deserves further research how to reduce the false positive number of elements in the estimated set. Here, a natural method is to take the intersection of these two estimators as the final estimator of $\mathcal{A}(T|\mathbf{X})$. That is, we define $\hat{\mathcal{A}}(T|\mathbf{X}) = \hat{\mathcal{A}}_1(T|\mathbf{X}) \cap \hat{\mathcal{A}}_2(T|\mathbf{X}) = \hat{\mathcal{A}}(Y|\mathbf{X}) \cap \hat{\mathcal{A}}(\delta T|\mathbf{X})$, which makes the false positive much less likely. What's more, the improved estimator is also of sure screening property.

Theorem 2 (Sure Screening Property) *Under the conditions of Theorem 1, we then have*

$$\lim_{n\to\infty} pr\{\mathcal{A}(T|\mathbf{X}) \subseteq \hat{\mathcal{A}}(T|\mathbf{X})\} = 1.$$

The proposed feature screening approach is easy to implement since it avoids complex operations needed in the screening methods with censored data in literature. It is widely applicable since the censoring mechanism is very general. It is flexible since it makes that any model free feature screening approaches for full data are applicable to the censored data. It can be extended to other survival data types, such as left censored data, interval censored data and truncated data.

It is noted that the CRC and RC mechanisms are two special cases of the GC mechanism. Hence, the proposed method can also be applied to the two cases.

1. Application to the CRC mechanism

The CRC mechanism, where C is assumed to be independent of T and $\mathbf{X}$, has been considered by many literatures, such as He et al. (2013), Song et al. (2014) and so on. Under this censoring mechanism, because C does not contain any information of $\mathbf{X}$, and hence Lemma 1 reduces to the following Lemma 2.

Lemma 2 *Under the CRC mechanism, we have* $\mathcal{A}(T|\mathbf{X}) = \mathcal{A}(Y|\mathbf{X}) = \mathcal{A}(\delta T|\mathbf{X})$.

Lemma 2 shows that the index set of interest is exactly the same as $\mathcal{A}(Y|\mathbf{X})$ and $\mathcal{A}(\delta T|\mathbf{X})$, which can be estimated by any model-free feature screening approach for full data, under the CRC mechanism. Compared with the estimators obtained in the GC case, the resulting estimator here may contain less false positive covariates, which is also verified by the simulation results.

2. Application to the RC mechanism

The RC mechanism, where C is independent of T given $\mathbf{X}$, is most commonly used in literatures (e.g., Li et al. 2016; Liu et al. 2018 and Chen et al. 2018). Due to $C \perp\!\!\!\perp T|\mathbf{X}$ for this case, then Lemma 1 reduces to the following Lemma 3.

Condition 3. For any $k \in \mathcal{A}(T|\mathbf{X}) \cap \mathcal{A}(C|\mathbf{X})$, the product of $pr(T > t|\mathbf{X})$ and $pr(C > t|\mathbf{X})$ depends functionally on the covariate X_k for some $t \in [0, \tau)$.

Lemma 3 *Under the RC mechanism, if Conditions 2 and 3 are satisfied, we then have*

(i) $\mathcal{A}(Y|\mathbf{X}) = \mathcal{A}(T|\mathbf{X}) \cup \mathcal{A}(C|\mathbf{X})$,
(ii) $\mathcal{A}(\delta T|\mathbf{X}) = \mathcal{A}(T|\mathbf{X}) \cup \mathcal{A}^*(\delta|\mathbf{X})$.

Under the RC mechanism, Lemma 3(i) replaces $\mathcal{A}^*(C|\mathbf{X})$ of Lemma 1(i) with $\mathcal{A}(C|\mathbf{X})$, where the latter is obviously a subset of the former. Hence, it is expected that the false positive number of the resulting estimator may be less in contrast to the estimators obtained in the general censoring case. In addition, the proposed screening method is also applicable to the case considered in Gorst-Rasmussen and Scheike (2013), where the censoring mechanism is assumed to be partially random in the sense of depending on inactive covariates of the survival variable.

3 Simulations

3.1 Simulation design

In this section, some simulation studies were conducted to investigate the finite sample performances of the proposed screening approach. Rather than comparing with all existing feature screening methods for survival data, we only compared with the latest screening procedures proposed by Chen et al. (2018) and Liu et al. (2018). This is because their works have already demonstrated that their proposed methods can outperform previous feature screening approaches in various survival cases.

Denote these two screening procedures from Chen et al. (2018) by Chen 1 and Chen 2, respectively. Denote the screening procedure from Liu et al. (2018) by Liu. To make it fair, we use the robust distance correlation

$$\omega_j = dcorr(F_j(X_j), F(Y)) = \frac{dcov(F_j(X_j), F(Y))}{\sqrt{dcov(F_j(X_j), F_j(X_j))dcov(F(Y), F(Y))}}$$

($docv$ is the distance covariance, $F(\cdot)$ is the distribution function)

to implement our approach (denoted by proposed-DC) for comparing with Chen 1 and Chen 2 since Chen 1 and Chen 2 use the above robust distance correlation based screening method. We use the fused Kolmogorov filter

$$\omega_j = \sum_{k=1}^{N_j} K_j^{\Lambda_{kj}} = \sum_{k=1}^{N_j} \max_{l_1, l_2} \sup_{0 \le t \le \tau} |S_j(t|I_{kj} = l_1) - S_j(t|I_{kj} = l_2)|$$

(Λ_{kj} is the kth partition of X_j, $S(\cdot|\cdot)$ is the conditional survival function)

to implement our approach (denoted by proposed-FKS) for comparing with Liu since Liu uses the fused Kolmogorov filter based screening method. In all the models, we set the sample size $n = 200$, the dimension of covariates $p = 2,000$, and repeated each experiment 500 times. To evaluate the proposed approach, we considered the following model settings.

Model 1. In the first example, we considered the common Cox proportional hazard model with the conditional hazard function given by

$$\lambda(t|\mathbf{X}) = \lambda_0(t) \cdot \exp(0.6X_1^3 + 1.3X_2 - 1.2\arccos(\pi X_3) - X_4 + 1.4X_5 + X_6),$$

where the baseline hazard function is set to be $\lambda_0(t) = 1$ and the covariate $\mathbf{X} \sim N(0, \Sigma)$ with $\Sigma = (0.8^{|k-m|})_{p\times p}$ for $k, m = 1, \cdots, p$. We considered the following four censoring mechanisms:

(a) $C = \tilde{C}$ with $\tilde{C} \sim U(0, c_0)$,
(b) $C = \exp(0.5\tan(0.5\pi X_2) + 6X_3^3 + c_0)$,
(c) $C = \exp(0.5\tan(0.5\pi X_2) + 3X_5^3 + 0.3X_{20} + c_0)$,
(d) $C = \exp(0.5\sin(T) + 0.5\tan(0.5\pi X_2) + 3X_5^5 + c_0)$,

where c_0 is chosen such that censoring rate is about 40% or 20% for these cases.

Model 2. In this example, we generated T from the following nonlinear accelerated failure time (AFT) model, which is another popular semi-parametric model in survival analysis.

$$\log(T) = X_1 + 0.4X_3 - \exp(-X_1 - 0.8X_2 - X_7) \cdot \epsilon,$$

where $\mathbf{X} \sim N(0, \Sigma)$ with $\Sigma = (0.8^{|k-m|})_{p\times p}$ for $k, m = 1, \cdots, p$, $\epsilon \sim N(0, 1)$ is independent of $\mathbf{X}$. And we considered the following four censoring mechanisms:

(a) $C = \tilde{C}$ with $\tilde{C} \sim U(0, c_0)$,

(b) $C = |0.5X_1^7 + c_0|$,
(c) $C = |-1.2X_1^5/(\mathbf{X}\beta_C) + c_0|$,
(d) $C = 0.3(\sin(T) - 0.2\sin(X_1^3) + c_0)^2$,

where c_0 is chosen such that censoring rate is about 40% or 20% for these cases, and $\beta_C = (2.5, 0.7, -3, -1.8, 0_{p-4})'$.

Model 3. Suppose that T took the following model, which represents various types of covariate functions with different degree of non-linearity, similarly adapted from Li et al. (2016).

$$\log(T) = g_1(X_1) + g_2(X_2) + g_3(X_3) + g_4(X_4) + \epsilon,$$

where $g_1(x) = 5\cos(2\pi x)$, $g_2(x) = 5\exp(1.2(x-1))$, $g_3(x) = -2.5x + 1$, $g_4(x) = 3\arctan(3x-2)$, $\mathbf{X} \sim N(0, \Sigma)$ with $\Sigma = (0.8^{|k-m|})_{p\times p}$ for $k, m = 1, \cdots, p$, $\epsilon \sim N(0,1)$ is independent of $\mathbf{X}$. And we considered the following four censoring mechanisms:

(a) $C \sim N(0,4) - N(5,1) + N(27,1)$,
(b) $C = \exp(0.6\tan(X_1) - X_3^3 + c_0)$,
(c) $C = \exp(0.6\tan(2\pi X_1) - X_3^3 + 2\sin(\mathbf{X}'\beta_C) + c_0)$,
(d) $C = \exp(0.6\tan(0.5\pi T) - X_3^3 + \sin(\mathbf{X}'\beta_C) + c_0)$,

where c_0 is chosen such that censoring rate is about 40% or 20% for these cases, and $\beta_C = (0, 0, 0, 0, -3.8, -4.2, 0, 0, 3, 5, -4, 0_{p-11})'$.

For all the above models, we consider four different censoring mechanisms. Case (a) means the CRC mechanism, where $\mathcal{A}(T|\mathbf{X}) = \mathcal{A}(Y|\mathbf{X}) = \mathcal{A}(\delta T|\mathbf{X})$. Cases (b) and (c) mean the RC mechanism, where $\mathcal{A}(T|\mathbf{X}) \subseteq \mathcal{A}(Y|\mathbf{X})$ and $\mathcal{A}(T|\mathbf{X}) \subseteq \mathcal{A}(\delta T|\mathbf{X})$. Case (d) means the GC mechanism. We assessed the performances of the screening approaches through the minimum model sizes (MMS), the minimum number of covariates needed to include all the active variables, like Fan and Lv (2008) and Mai and Zou (2015).

3.2 Simulation results and conclusions

Tables 1-2 present the simulation results, which contains the MMS's median of 500 replicates and the mean absolute deviation (parenthesis). In summary, we can find the following conclusions.

- The proposed feature screening approach works reasonably well and performs stably for all the models under four different censoring mechanisms considered here. In addition, our approach's finite sample performances are more robust according to these results of different censoring situations, whereas other screening methods have their ups and downs.
- For Case (a), the censoring variable completely independent of the survival variable and all covariates, the proposed approach is exactly comparable to other methods in terms of MMS, but its mean absolute deviation is gener-

Table 1 Simulation results for Models 1-3 under four censoring mechanisms with the censoring rate 40%

	Method	(a)	(b)	(c)	(d)
Model 1 (d=6)	proposed-DC	6(0.072)	6(0.216)	6(1.212)	6(1.802)
	Chen 1	6(0.294)	10(12.756)	42(73.080)	31(38.946)
	Chen 2	6(2.112)	7(13.532)	12(33.380)	7(10.990)
	proposed-FKS	6(0.274)	6(5.496)	8(22.552)	6(1.992)
	Liu	6(0.172)	33(94.672)	41(109.130)	11(31.194)
Model 2 (d=4)	proposed-DC	9(13.130)	7(7.024)	7(6.766)	9(13.202)
	Chen 1	25(65.104)	11(41.740)	17(49.232)	18(51.036)
	Chen 2	10(28.288)	13(35.600)	13(34.204)	15(42.854)
	proposed-FKS	8(22.406)	8(15.034)	8(14.910)	7(17.680)
	Liu	64(139.668)	262(392.178)	129(243.194)	56(129.486)
Model 3 (d=4)	proposed-DC	8(22.522)	10(37.470)	5(10.728)	5(5.352)
	Chen 1	10(31.618)	372(493.868)	515(612.042)	417(522.374)
	Chen 2	8(27.468)	17(31.950)	24(42.766)	18(31.312)
	proposed-FKS	5(20.432)	5(5.000)	4(1.234)	4(0.774)
	Liu	6(24.280)	7(45.252)	11(80.414)	7(50.102)

Table 2 Simulation results for Models 1-3 under four censoring mechanisms with the censoring rate 20%

	Method	(a)	(b)	(c)	(d)
Model 1 (d=6)	proposed-DC	6(0.008)	6(1.542)	6(0.990)	6(1.592)
	Chen 1	6(0.022)	8(12.612)	6(6.042)	24(27.054)
	Chen 2	6(0.028)	6(4.408)	6(1.316)	6(4.254)
	proposed-FKS	6(0.018)	6(4.120)	6(1.742)	6(5.048)
	Liu	6(0.036)	6(4.602)	6(2.016)	7(9.964)
Model 2 (d=4)	proposed-DC	6(3.336)	6(1.568)	6(1.908)	5(1.984)
	Chen 1	18(32.166)	23(49.640)	20(34.130)	19(30.134)
	Chen 2	7(12.818)	9(17.728)	7(13.756)	7(11.962)
	proposed-FKS	5(3.832)	5(1.206)	5(1.814)	5(2.232)
	Liu	23(60.800)	23(69.606)	17(42.526)	21(57.916)
Model 3 (d=4)	proposed-DC	5(4.790)	8(16.356)	9(16.802)	8(14.648)
	Chen 1	5(8.360)	312(438.976)	244(398.612)	264(404.038)
	Chen 2	5(6.290)	18(56.930)	17(56.968)	17(56.298)
	proposed-FKS	4(6.726)	5(3.228)	4(2.340)	4(2.910)
	Liu	5(7.702)	5(26.440)	5(21.828)	5(22.234)

ally smaller. However, Case (*a*) is rare in practice. In Cases (*b*) and (*c*) for all models, our approach outperforms the other methods in terms of MMS, and its mean absolute deviation is much smaller than others. The reason might be that our method uses fully observed data to estimate $\mathcal{A}(T|\mathbf{X})$ and does not use Kaplan-Meier estimator. In Case (*d*) for all the models, other screening meth-

ods perform poorly because they are only proposed under the RC mechanism. However, our proposed approach works well under this GC mechanism.

- As the censoring rate increases from 20% to 40%, the proposed screening approach always works well, but the performance of other screening methods become worse. This shows that the proposed method is robust for different censoring rates. The reason may be that the proposed method uses these observable variables more effectively.

In addition, in order to more clearly show the performance of our approach under the GC mechanism, we also present the bar graph of the MMS's median for all models' Case (d) (Figs. 1-2). It is easily to see that the MMSs of proposed-DC and proposed-FKS are much closer to the true model size in either censoring rate.

4 Applications

4.1 Mantle cell lymphoma data

Now, we applied the proposed screening approach to the mantle cell lymphoma (MCL) dataset. The dataset contains 8,810 expression genes for 92 patients diagnosed with MCL based on the morphologic and immunophenotypic criteria, which has been studied by Rosenwald et al. (2003) and Liu et al. (2018). This dataset is available at http://llmpp.nih.gov/MCL. During the follow-up, 64 patients were died of MCL and the rest of them were censored. That is, the overall censoring rate is

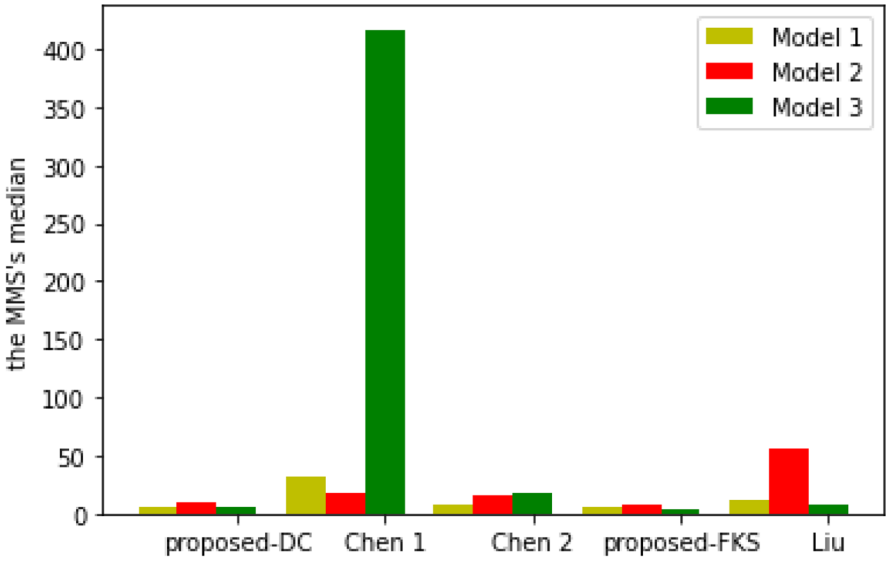

Fig. 1 Bar graph of the MMS's median for Case (d) of Models 1-3 with the censoring rate 40%

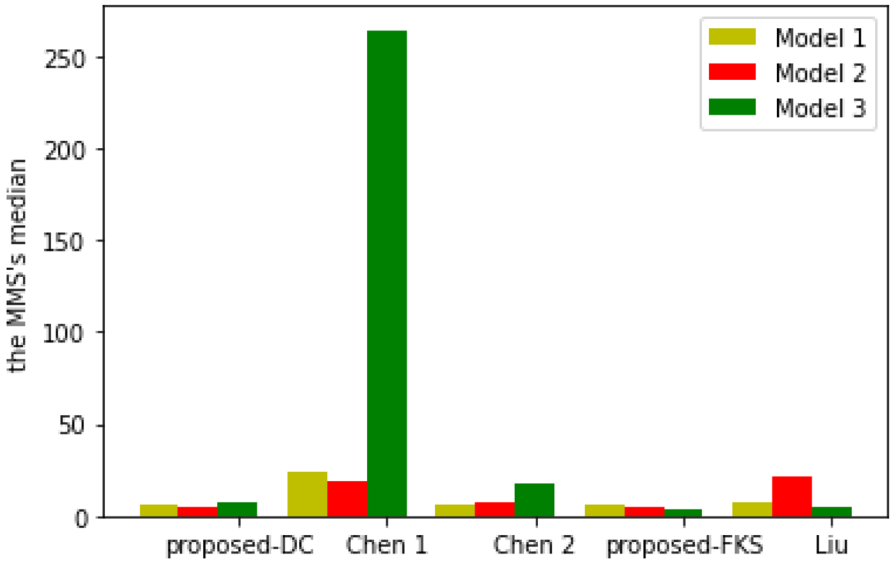

Fig. 2 Bar graph of the MMS's median for Case (d) of Models 1-3 with the censoring rate 20%

30.43%. Patient's survival time ranges from 0.02 to 14.05 years with the median observed survival time of 2.8 years. The primary goals of this research are to investigate which genes had expression patterns that correlated with life of patients, and to predict the length of survival of these patients. Because of such small sample size and huge number of covariates, a screening is necessary prior to any meaningful statistical analysis.

For comparisons, we also applied the other competing methods mentioned in Simulations to screen genes that may be relevant to the survival time. Instead of removing these genes with missing values, we firstly impute them by the K-nearest neighbor method with $K = 15$. In our subsequent analysis, we standardized each gene and screened the first $20 = [92/\log(92)]$ important genes with $[a]$ denoting the integer part of a. The gene unique identification (UNIQID) of these selected genes were displayed in Table 3. From this table, we can observe that seven genes with UNIQID 17198, 28346, 28990, 30334, 30898, 31420 and 34771 were selected by all the five feature screening approaches, indicating that they may be strongly associated with patients's survival time; five more genes with UNIQID 24656, 26950, 28534, 31443 and 32187 were only selected by our screening procedures. By consulting the literature, Rosenwald et al. (2003) had shown that the selected genes with UNIQID 28346, 28990, 30334, 34771 and 26950 had significant influences on the survival time. However, other screening methods could not select the important one with UNIQID 26950.

To evaluate the predictive accuracy of these methods, we randomly splitted the data into training dataset and test dataset, where the number of the training dataset accounts for about 2/3 of all samples (61 patients) and the censoring rate is roughly remained the same. The remainder of the observed dataset was

Table 3 The UNIQIDs of the top 20 selected genes for MCL data

proposed-DC	Chen 1	Chen 2	proposed-FKS	Liu
16787	27762	30898	17176	28990
17198	30898	30122	17198	30334
23826	17198	30334	17326	30898
24610	30334	28990	24656	31049
26944	27116	27762	26944	30157
27095	28640	31420	26950	17176
27116	24723	30157	27095	25234
27762	34771	34790	27116	17198
28346	23826	34771	28346	28346
28534	16787	17198	28872	24794
28872	34790	28346	28990	26944
28990	29897	29897	29897	31420
29897	28346	28872	30142	29357
30334	31420	27095	30282	30142
30898	30282	24610	30334	34771
31049	30122	17326	30898	27310
31420	24610	28640	31420	30378
32187	28990	25234	31443	17691
34771	27095	16787	32187	34790
34790	30949	26944	34771	24723

considered as the test part. We applied all the above feature screening approaches to the training dataset and selected $14 = [61/\log(61)]$ genes. Then, we fitted an AFT model with the lasso penalty based on selected genes on the training dataset and used this fitted model to make a prediction on the survival time in the test dataset. After repeating this procedure 200 times, we reported the average of the mean relative error (MRE, $\text{MRE}=\frac{1}{n}\sum|\frac{\textit{True value}-\textit{Prediction}}{\textit{True value}}|$) of different screening approaches, with a smaller value indicating better performance, in Table 4. And the box plots of the MREs are also given in Fig. 3. It can be seen that proposed-DC and proposed-FKS are the two top performers in terms of the average of MRE under the working AFT model. This indicates that our proposed screening procedure has better predictive performance for the MCL data, which again shows the reliability of our screening results in Table 3.

Table 4 The average of 200 MREs under a working AFT model for MCL data

	proposed-DC	Chen 1	Chen 2	proposed-FKS	Liu
MRE	1.6431	1.6485	1.6448	1.6340	1.7228

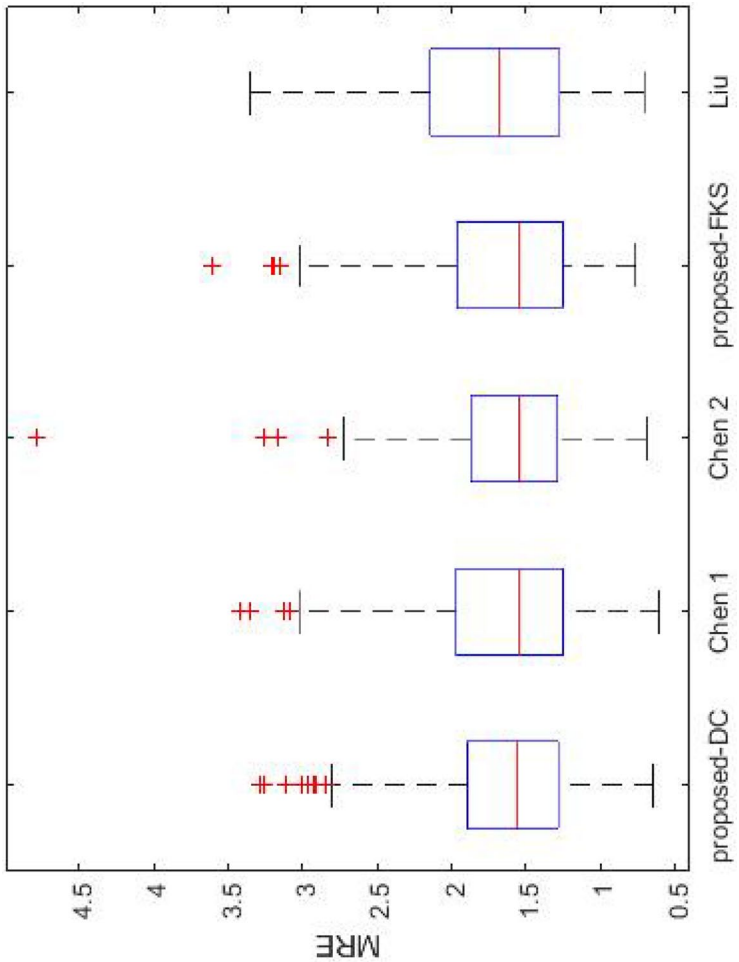

Fig. 3 Box plot of 200 MREs for different screening approaches in MCL data

4.2 Breast cancer data

Another real dataset from a breast cancer study (Van Houwelingen et al. 2006) contains 295 female patients with primary invasive breast carcinoma. For each patient, the expressions of 24,885 genes were profiled on cDNA arrays from all tumors. A set of 4,919 candidate genes were selected after initial screening by the Rosetta error model (Van't Veer et al. 2002). The median follow-up time is 7.2 years, and 73.2% of the observations (216 patients) were censored. The study aims to identify genes that are associated with the overall survival of breast cancer patients, and to predict patient's survival.

Similarly, we applied all the above feature screening methods to screen for active genes related to survival after the missing genes imputed by the K-nearest neighbor method with $K = 15$. All predictors were standardized with mean zero and variance one. We screened the first $51 = [295/\log(295)]$ genes and displayed these genes in Table 5. It can be easily seen that 19 genes were screened by all the five screening methods. One of them Contig38288.RC had been identified as an active predictive gene in Van't Veer et al. (2002). Besides, the gene NM.006623 was also selected by Song et al. (2014), but it was only selected by our screening approaches.

In addition, we examined and compared the predictive accuracy of these screened genes by different methods with 200 random partitions of the data. For each partition, about 2/3 of all samples (196 patients) were randomly selected as the training dataset with the roughly same censoring rate. The remainder of the observed dataset was considered as the test part. Like the above, we applied these screening approaches to the training dataset and selected $37 = [196/\log(196)]$ genes. Then, we fitted an AFT model with the lasso penalty based on selected genes on the training dataset and used this fitted model to make a prediction on the survival time in the test dataset. Table 6 reports the average of MRE of different screening approaches and Fig. 4 presents the box plots of the MREs. These results indicate that our procedure has better predictive performance for the breast cancer data in terms of the smallest average of MRE. Hence, our screening approach might give more reliable selection results than three others.

5 Conclusion

We propose a very easily implemented surrogate-variable-based model-free feature screening approach for the censored survival data under the GC mechanism, and demonstrate its superior performances by Monte Carlo simulations and the real applications. In this paper, for fairly comparing with the three existing screening methods, the proposed approach uses the corresponding dependence measures. Nevertheless, one can use any dependence measures from a wealth of available feature screening literatures, which is a rather appealing and distinct trait. In addition, compared with the existing screening methods, our approach is developed under a very general censoring mechanism. This shows that our approach is much more credible in reality due to the unknown censoring mechanism.

Table 5 The names of the top 51 selected genes for breast cancer data

proposed-DC	Chen 1	Chen 2	proposed-FKS	Liu
NM.002358	Contig38288.RC	NM.003981	NM.001605	NM.016359
NM.003158	NM.007057	NM.003600	NM.000926	NM.014176
NM.002497	NM.003981	D14678	NM.001673	NM.001168
NM.003258	Contig48328.RC	U74612	NM.003158	U96131
NM.001809	NM.003600	NM.004701	NM.002497	NM.016569
NM.002689	Contig31288.RC	NM.016359	NM.003258	D43950
AB024704	NM.003158	Contig57584.RC	NM.001809	NM.003035
AF279865	NM.001605	NM.007019	AF279865	NM.020974
Contig45816.RC	NM.005733	NM.003158	Contig45816.RC	NM.014321
NM.004217	Contig33814.RC	NM.004217	NM.004217	NM.004217
NM.003504	NM.018410	NM.018410	NM.003504	D38553
NM.012291	D14678	NM.014176	AB040926	NM.003600
NM.003600	NM.014585	AB024704	NM.003600	Contig38288.RC
NM.004336	NM.013277	NM.007057	NM.004336	Contig55725.RC
NM.003686	NM.004336	U96131	Contig31288.RC	NM.003504
Contig31288.RC	NM.006607	NM.013277	NM.004456	NM.005733
NM.012474	NM.001809	NM.001168	NM.013277	NM.014791
NM.004456	Contig51749.RC	NM.003258	NM.005375	NM.004456
NM.013277	Contig8818.RC	NM.003504	NM.020686	NM.012067
NM.006027	NM.004701	Contig48328.RC	NM.003981	NM.004358
NM.020675	NM.006845	Contig31288.RC	Contig35629.RC	NM.006845
NM.003981	NM.000270	Contig38288.RC	NM.014176	NM.006607
NM.014176	U74612	NM.018455	NM.004702	NM.004701
NM.004701	Contig34766.RC	NM.006607	U74612	NM.013277
NM.004702	AB040926	NM.001809	AL160131	NM.001122
Contig48328.RC	NM.014501	M96577	M96577	NM.006461
U74612	NM.001109	NM.004336	NM.007019	NM.003981
M96577	NM.003258	Contig45816.RC	NM.014321	NM.003579
NM.007019	NM.007019	NM.004456	NM.007057	Contig36879.RC
NM.007057	Contig38726.RC	NM.006027	NM.020974	Contig64688
D43950	Contig45816.RC	NM.005733	D43950	NM.001809
NM.014454	AL137566	NM.001333	NM.006461	NM.006500
NM.005733	NM.004217	NM.012291	NM.005733	AL137347
AF047002	NM.004456	Contig38901.RC	NM.014501	NM.012291
NM.006607	Contig55069.RC	D43950	NM.006607	NM.007019
Contig64688	Contig56843.RC	D38553	NM.006623	Contig31288.RC
NM.014791	NM.020974	NM.003686	Contig64688	Contig38901.RC
Contig38288.RC	Contig57584.RC	NM.006845	NM.014791	NM.001124
Contig38901.RC	NM.001333	NM.014791	Contig38288.RC	NM.001905
NM.006845	NM.001255	AL049265	NM.006845	NM.001333
NM.014875	NM.003686	NM.006461	AJ224741	NM.003258
D38553	NM.006027	NM.014501	Contig56843.RC	U74612

Table 5 (continued)

proposed-DC	Chen 1	Chen 2	proposed-FKS	Liu
NM.016359	NM.001168	NM.020675	D38553	NM.001071
NM.016448	NM.014176	NM.002497	AL049265	NM.018410
U96131	NM.020686	AF279865	NM.016359	Contig48270.RC
NM.018410	NM.006006	NM.004702	AL137566	NM.017522
NM.018455	NM.016359	Contig33814.RC	Contig38726.RC	NM.004336
D14678	NM.004648	NM.005804	U96131	NM.002106
NM.001168	AL160131	AB040926	NM.018410	Contig57584.RC
NM.001333	NM.007274	Contig34766.RC	D14678	NM.002497
NM.002106	NM.000926	NM.001605	NM.001168	NM.006027

Table 6 The average of 200 MREs under a working AFT model for breast cancer data

	proposed-DC	Chen 1	Chen 2	proposed-FKS	Liu
MRE	2.0486	2.1374	2.1469	2.1089	2.1384

As mentioned before, the active variable set estimated by the proposed screening approach may include some redundant covariates. Although we ulteriorly cut down the false positive number through taking a intersection of two estimated sets, it deserves further study how to obtain a more accurate estimator. We will consider the problem in a subsequent study.

Acknowledgements Wang's research was supported by the National Natural Science Foundation of China (General program 11871460 and program for Innovative Research Group in China 61621003), a grant from the Key Lab of Random Complex Structure and Data Science, CAS.

Appendix

Proof of Lemma 1 To facilitate the presentation, we write $\mathbf{X}_{\mathcal{A}} = \{X_k : k \in \mathcal{A}\}$ for any non-negative integer set $\mathcal{A}$. First, we prove Lemma 1 (i).

Under the GC mechanism, for any $t \in [0, \tau)$, we have

$$pr(Y > t|\mathbf{X}) = pr(\min(T, C) > t|\mathbf{X}) = pr(T > t|\mathbf{X}) \cdot pr(C > t|T > t, \mathbf{X}). \quad (1)$$

Recalling the definition of $\mathcal{A}(Y|\mathbf{X})$, it is easy to see $\mathcal{A}(Y|\mathbf{X}) \subseteq \mathcal{A}(T|\mathbf{X}) \cup \mathcal{A}^*(C|\mathbf{X})$. On the other hand, for any $X_j \in \mathbf{X}_{\mathcal{A}(T|\mathbf{X})\cup\mathcal{A}^*(C|\mathbf{X})}$, we have $X_j \in \mathbf{X}_{\mathcal{A}(T|\mathbf{X})}$ or $X_j \in \mathbf{X}_{\mathcal{A}^*(C|\mathbf{X})}$. That is, $pr(T > t|\mathbf{X})$ or $pr(C > t|T > t, \mathbf{X})$ depend functionally on X_j for some $t \in [0, \tau)$, and hence $pr(Y > t|\mathbf{X})$ depends functionally on X_j by the conditions of Lemma 1. This proves $X_j \in \mathbf{X}_{\mathcal{A}(Y|\mathbf{X})}$. Lemma 1 (i) is then proved.

Lemma 1 (ii) can be proved similar to Lemma 1(i) by noting

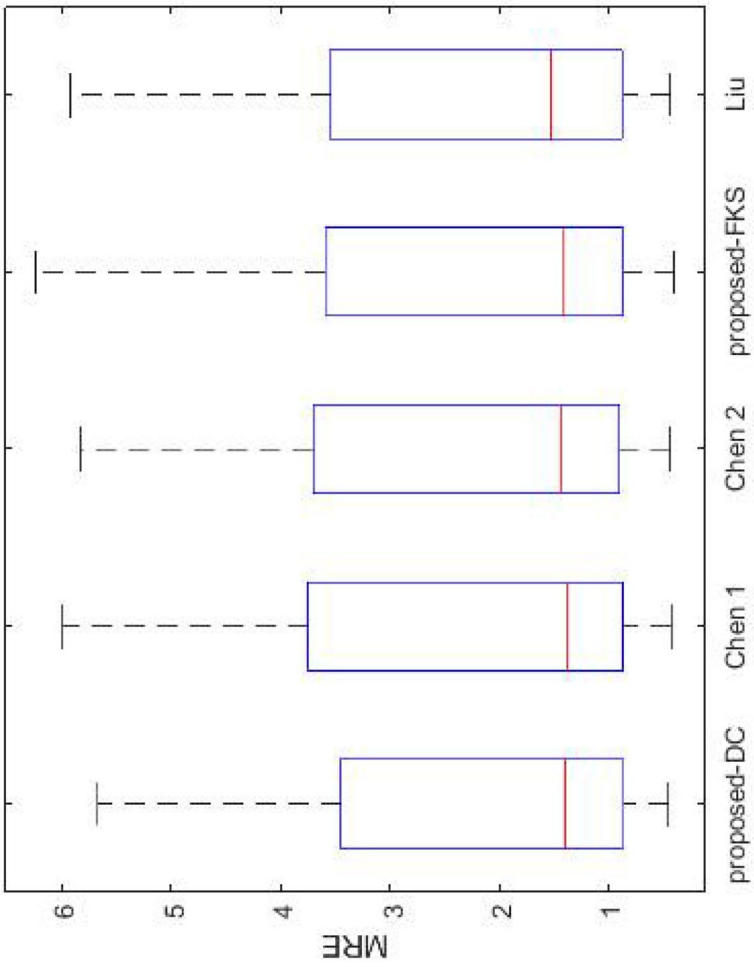

Fig. 4 Box plot of 200 MREs for different screening approaches in breast cancer data

$$\begin{aligned} pr(\delta T > t|\mathbf{X}) =& pr(\delta = 1, T > t|\mathbf{X}) \\ =& pr(T > t|\mathbf{X}) \cdot pr(\delta = 1|T > t, \mathbf{X}) \end{aligned}$$

for $t \in [0, \tau)$. □

Proofs of Theorems 1 and 2 The proofs are direct based on Lemma 1, and hence we omit it. □

Proof of Lemma 2 Under the CRC mechanism, namely, $C \perp\!\!\!\perp (T, \mathbf{X})$, we then have $\mathcal{A}^*(C|\mathbf{X})$ is an empty set and $\mathcal{A}^*(\delta|\mathbf{X})$ is a subset of $\mathcal{A}(T|\mathbf{X})$. This proves Lemma 2. □

Proof of Lemma 3 Under the RC mechanism, Lemma 3 is a direct result of Lemma 1 by noting $\mathcal{A}^*(C|\mathbf{X}) = \mathcal{A}(C|\mathbf{X})$. □

References

Chang, J., Tang, C. Y., Wu, Y. (2013). Marginal empirical likelihood and sure independence feature screening. *The Annals of Statistics, 41,* 2123–2148.

Chen, X., Chen, X., Wang, H. (2018). Robust feature screening for ultra-high dimensional right censored data via distance correlation. *Computational Statistics and Data Analysis, 119,* 118–138.

Cui, H., Li, R., Zhong, W. (2015). Model-free feature screening for ultrahigh dimensional discriminant analysis. *Journal of the American Statistical Association, 110,* 630–641.

Fan, J., Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society, Series B, 70,* 849–911.

Fan, J., Song, R. (2010). Sure independence screening in generalized linear models with NP-dimensionality. *The Annals of Statistics, 38,* 3567–3604.

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

Fan, J., Feng, Y., Song, R. (2011). Nonparametric independence screening in sparse ultra-high dimensional additive models. *Journal of the American Statistical Association, 106,* 544–557.

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

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

Leung, K. M., Elashoff, R. M., Afifi, A. A. (1997). Censoring issues in survival analysis. *Annual Review of Public Health*, *18*, 83–104.

Li, G., Peng, H., Zhang, J., Zhu, L. (2012a). Robust rank correlation based screening. *The Annals of Statistics, 40,* 1846–1877.

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

Li, R., Zhong, W., Zhu, L. (2012b). Feature screening via distance correlation learning. *Journal of American Statistical Association, 107,* 1129–1139.

Liu, Y., Zhang, J., Zhao, X. (2018). A new nonparametric screening method for ultrahigh-dimensional survival data. *Computational Statistics and Data Analysis, 119,* 74–85.

Mai, Q., Zou, H. (2015). The fused Kolmogorov filter: A nonparametric model-free screening method. *The Annals of Statistics, 43,* 1471–1497.

Pan, W. L., Wang, X. Q., Xiao, W. N., Zhu, H. T. (2019). A generic sure independence screening procedure. *Journal of American Statistical Association, 114,* 928–937.

Rosenwald, A., Wright, G., Wiestner, A., Chan, W. C., Connors, J. M., Campo, E., et al. (2003). The proliferation gene expression signature is a quantitative integrator of oncogenic events that predicts survival in mantle cell lymphoma. *Cancer Cell*, *3*, 185–197.

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

Van Houwelingen, H. C., Bruinsma, T., Hart, A. A., van't Veer, L. J., Wessels, L. F. (2006). Cross-validated Cox regression on microarray gene expression data. *Statistics in Medicine, 25,* 3201–3216.

Van't Veer, L., Dai, H., van de Vijver, M., He, Y., Hart, A., Mao, M., van der, H. P. K., Marton, M., Witteveen, A., Schreiber, G., Kerkhoven, R., Roberts, C., Linsley, P., Bernards, R., Firend, S. (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature, 415,* 530–536.

Yang, G., Yu, Y., Li, R., Buu, A. (2016). Feature screening in ultrahigh dimensional Cox's model. *Statistica Sinica, 26,* 881–901.

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

Zhong, W., Zhu, L., Li, R., Cui, H. (2016). Regularized quantile regression and robust feature screening for single index models. *Statistica Sinica, 26,* 69–95.

Zhu, L. P., Li, L., Li, R., Zhu, L. X. (2011). Model-free feature screening for ultrahigh-dimensional data. *Journal of American Statistical Association, 106,* 1464–1475.