Statistica Sinica Preprint No: SS-2023-0309	
Title	New Feature Screening Methods for Massive Interval-censored Failure Time Data
Manuscript ID	SS-2023-0309
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202023.0309
Complete List of Authors	Huiqiong Li, Zhimiao Cao, Jianguo Sun and Niansheng Tang
Corresponding Authors	Jianguo Sun
E-mails	sunj@missouri.edu
Notice: Accepted author version.	

Statistica Sinica

New Feature Screening Methods for Massive Interval-censored Failure Time Data

Huiqiong Li[a], Zhimiao Cao[a], Jianguo Sun[b] and Niansheng Tang[a]

[a] *Yunnan Key Laboratory of Statistical Modeling and Data Analysis, Yunnan University,* Kunming, 650091, China

[b] *Department of Statistics, University of Missouri,Columbia, MO, 65211, U.S.A.*

Abstract: Screening important features has become one of the important tasks in statistical analysis and correspondingly, various screening procedures have been proposed for various types of studies or data including both complete and incomplete data. However, these methods would be computationally costly or even infeasible when one faces massive health databases with both high dimensionality and huge sample size, which have become increasingly popular for comparative effectiveness and safety studies of medical products. In this paper, we consider such a type of incomplete data, interval-censored failure time data, that have not be discussed before and propose two procedures with the use of distance correlation and orthogonal sampling as well as the the jackknife debiased average technique. The proposed approaches can be easily implemented and their sure screening and rank consistency properties are established. Simulation studies demonstrate that the proposed methods work well for practical situations and they are applied to the SEER breast cancer data.

Key words and phrases: Distance correlation, Jackknife debiased average, Orthogonal subsampling, Rank consistency, Sure screening.

1. Introduction

This paper considers the feature screening for massive interval-censored failure time data. By being massive, we assume that the data have both large numbers of features (p) and huge sample sizes (N), while by interval censoring, we mean that the failure time of interest is only observed to belong to an interval instead of being observed exactly (Sun, 2006). It is easy to see that such data naturally occur in many studies such as epidemiological or medical follow-up studies, in particular clinical trials. Two specific examples of them are given by the medicare data in Wang et al. (2021b) and the LEGEND-HTN data in Yang et al. (2023). For the problem, two methods will be developed.

Screening important features has become one of the important tasks in statistical analysis and correspondingly, various model-based or model-free screening procedures have been proposed for various types of studies or data including both complete and incomplete data. Among them, one important contribution was given by Fan and Lv (2008) who proposed a sure independence screening (SIS) procedure under the framework of linear

models. Following Fan and Lv (2008), many researchers have generalized SIS to other models, including the generalized linear model (Fan and Song, 2010), the additive model (Fan et al., 2011), and the multi-index model (Zhu et al., 2011). Model-free methods include the robust-ranking-correlation-based screening (Li et al., 2012a), the distance-correlation-based SIS (DC-SIS) (Li et al., 2012b; Zhong et al., 2016), the conditional quantile SIS (Wu and Yin, 2015), the projection-correlation-based SIS (Zhu et al., 2017), and the jackknife debiased R-squared screening (Zhu et al., 2022). Feature screening for interval-valued data are also studied(Zhong et al., 2023; Zhang and Feng, 2024; Dai et al., 2020; Shu et al., 2024; Peng and Zhang, 2021).

Massive datasets often occur in many modern scientific fields, including genetic studies, signal processing, and the internet. It is well-known that the handling or analysis of such data present many challenges including data storage, communication, computational speed, statistical accuracy and algorithmic stability (Fan et al., 2020). Also the usual screening methods such as these described above would be computationally costly or even infeasible, and the development of computationally convenient methods is pivotal to overcome these challenges. For this, some methods have been proposed, including the divide-and-conquer strategy (Zhao et al., 2016; Battey et al., 2018; Shi et al., 2018), the online-updating approach (Schifano et al.,

2016; Wang et al., 2018a; Kong and Xia, 2019; Luo and Song, 2020), and the subsampling method(Ma et al., 2015; Wang et al., 2018b; Wang, 2019). However, most of these approaches apply only to completely observed data.

Several authors have considered the screening problem for massive incomplete or failure time data. For example, Kawaguchi et al. (2020) developed a scalable sparse Cox regression method for the problem, while Wang et al. (2021b) proposed an efficient divide-and-conquer algorithm for fitting the sparse Cox regression. Also Xue et al. (2019) proposed an online updating approach for testing the proportional hazards assumption, and Zuo et al. (2021) developed a subsampling algorithm to efficiently approximate the estimators of regression parameters in the additive hazards model. Furthermore, Wu et al. (2021) proposed some online-updating estimators for both the regression coefficient and the baseline hazard function. A major limitation of the methods above is that they were developed for right-censored data, a special case of interval-censored data (Kalbfleisch and Prentice, 2002). In other words, they cannot be applied to interval-censored data (Sun, 2006) since the analysis of the latter is much more complicated and difficult than that of right-censored data due to their more complex structures. Corresponding to this, we will propose two new subsampling-based approaches.

The subsampling approach has recently attracted a good deal of attention for the analysis of the data with either large dimensionality (p) or sample size (N). For example, the available methods include the information-based optimal subdata selection method given in Wang et al. (2019), the subsampling winner algorithm developed by Fan and Sun (2021), the sequential addressing subsampling method introduced by Pan et al. (2023), and the orthogonal subsampling (OSS) approach proposed by Wang et al. (2021a). Note that the subsampling approach is typically designed for the data with fixed dimensions and no one has considered its application to the analysis of interval-censored data. In the following, we will discuss the feature screening for massive interval-censored failure time data with both $N \gg p$ and p being large and develop two methods with the use of the DC-SIS and the sampling technique.

The proposed methods can be regarded as generalizations of the DC-SIS screening method for regular or non-massive interval-censored data given in Zhang et al. (2023), who demonstrated its excellent screening performance. However, their method cannot be applied to massive interval-censored data because the needed computation of the DC-SIS is in the degree of $O(N^3)$ and thus it cannot run on ordinary machines for large data. Also for the massive data, one needs to consider whether it is necessary to use all avail-

able information because the characteristics of the massive data can cause information overlap. To address these and other challenges, we will first develop a simple average distance correlation screening procedure of orthogonal subsampling (SDC-OSS) and then a jackknife debiased distance correlation screening procedure of orthogonal subsampling (JDC-OSS), which incorporate the concepts of orthogonal subsampling and jackknife debiased average screening.

The proposed methods have several features or advantages. First, they are model-free or do not rely on any specific regression model, allowing for flexible analysis of the relationship between response and predictor variables. Second, they efficiently manage computation complexity by limiting the original $O(N^3)$ calculations to $O(Bn^3)$, where B is the number of segments and n is the number of subsamples in each segment to be defined below. Third, they have the sure screening and rank consistency properties. Note that the main difference between the method given by Zhang et al. (2023) and that proposed in this paper is that the former makes use of full data and thus may be infeasible or fail for the situation discussed here. In contrast, the proposed method only utilizes a subset of the full samples through the subsampling and incorporates the concepts of aggregation and jackknife debiasing. These new features make the establishment of the

theoretical properties of the proposed methods much more challenging and difficult. More discussion on this are given below.

The remainder of the paper is organized as follows. We will first introduce the notation and the set-up of the problem and then describe the idea of the proposed methods in Section 2. In Section 3, two screening procedures, SDC-OSS and JDC-OSS, will be developed, and their asymptotic properties, the sure screening and rank consistency properties, will be established in Section 4. An extensive simulation study will be conducted in Section 5 to investigate the finite sample performance of the proposed methods and the results suggest that they work well for practical situations. In Section 6, an application to the SEER breast cancer data is provided and Section 7 gives some discussion and concluding remarks.

2. Notation and Set-up

Consider a failure time study consisting of N independent subjects and let T denote the failure time of interest. Suppose that for each subject, there exists a p-dimensional vector of covariates denoted by $\mathbf{X} = (X_1, X_2, \ldots, X_p)^\top$ and let $S(t|\mathbf{X}) = P(T > t|\mathbf{X})$ denote the survival function for a subject with the covariate $\mathbf{X}$. Furthermore, define

$$\mathcal{M} = \{k : S(t|\mathbf{X}) \text{ functionally depends on } X_k \text{ for } t \geq 0, k = 1, \ldots, p\},$$

representing the index set of the active covariates or covariates that have an effect on T. The main goal here is to perform the feature screening to identify the active covariates or estimate the index set $\mathcal{M}$ when one faces massive interval-censored data. For this, as others, we will make the sparse assumption that only a handful or small number of covariates or features are relevant to the failure time of interest T.

For the observed data, we will assume that for each subject, there exist two monitoring or observation times denoted by U and V with $U < V$ and T is only known to be in one of three situations: T is between U and V or interval-censored, T is greater than V or right-censored, and T is less than U or left-censored. That is, we have case II interval-censored data (Sun, 2006) given by $\mathcal{D} = \{(U_i, V_i, \delta_{i1}, \delta_{i2}, \delta_{i3}, \mathbf{X}_i), i = 1, 2, \ldots, N\}$, the N i.i.d. copies of $(U, V, \delta_1, \delta_2, \delta_3, \mathbf{X})$, where $\delta_1 = I(T < U)$, $\delta_2 = I(U \leq T < V)$, $\delta_3 = 1-\delta_1-\delta_2$, and $\mathbf{X}_i = (X_{i1}, X_{i2}, \ldots, X_{ip})^\top$. In the following, we will assume that the censoring mechanism is independent or non-informative (Sun, 2006), meaning that given $\mathbf{X}$, T is independent of U and V, denoted as $T \perp (U, V)|\mathbf{X}$. Also similar to Zhu et al. (2022), it will be assumed that the sample size N is extremely large (i.e., N in thousands to hundreds of millions) and the covariate dimension p is ultrahigh dimensional (i.e., p in thousands or tens of thousands), meaning that $\log(p) = O(N^\zeta)$ with

$\zeta \in \left(0, (\alpha+\iota)\max\{1-2\gamma-2\kappa, \gamma\}\right)$ (Li and Xu, 2023). Here α, ι, γ and κ will be detailed in Section 4.

Let $\omega_j \geq 0$ denote a marginal correlation measure or measure of the correlation strength between T and X_j. In general, a commonly used strategy for feature screening is first to estimate ω_j, say, by a centralized estimator $\hat{\omega}_j$ based on the observed data $\mathcal{D}$ and then estimate $\mathcal{M}$ by

$$\widehat{\mathcal{M}} = \{j : \hat{\omega}_j \geq \gamma, j = 1, \ldots, p\} \tag{2.1}$$

for a pre-specified threshold $\gamma > 0$. Similarly, one can select a positive integer d_0 and define the estimated active set as

$$\widehat{\mathcal{M}} = \{j : \hat{\omega}_j \text{ is amongst the first } \ d_0 \text{ largest of all } \hat{\omega}_j (j = 1, 2, \ldots, p)\}.$$

In the next section, we will develop two methods for obtaining the estimators $\hat{\omega}_j$'s, which lead to two feature screening procedures, a simple average distance correlation screening method and a jackknife debiased average distance correlation screening method based on orthogonal sampling. For the threshold value d_0, by following Fan and Lv (2008), one simple choice is $d_0 = [N/\log(N)]$, where $[a]$ represents the integer part of a real value a.

Before presenting the proposed estimators $\hat{\omega}_j$'s in the next section, we need to discuss the transformation of the observed data. For this, first note that an ideal method for estimating the ω_j's would be to use their

empirical estimators if one observes the exact value of T. However, this is impossible for the situation here due to interval censoring. To deal with this, by following Zhang et al. (2023), we consider the transformed data or variables

$$L = \delta_1 \cdot U + \delta_2 \cdot (V - U) + \delta_3 \cdot (\eta - V),$$

$$H = \delta_1 \cdot U + \delta_2 \cdot V + \delta_3 \cdot V,$$

which represent the length and endpoint of the time interval within which the event time lies, respectively. Here η can be any large constant and $\eta = 10^6$ will be used in the numerical studies below.

For the transformed data, we have that $H = V$ if T is either between U and V (interval-censored) or greater than V (right-censored) and $H = U$ if T is less than U (left-censored). In other words, H represents either the left or right endpoint of the observed interval. Furthermore, define the standardized L and H as $L^* = [L - E(L)]/\sqrt{Var(L)}$ and $H^* = [H - E(H)]/\sqrt{Var(H)}$, where $E(\cdot)$ and $Var(\cdot)$ denote the corresponding expectations and variances, respectively. Both Székely, Rizzo, and Bakirov (2007) and Zhang et al. (2023) have showed that the zero distance correlation between (L^*, H^*) and X implies the zero distance correlation between the observed data on T and X. Also it can be shown that if T depends on X, then the observed data depend on X and vice versa. Thus

it is natural and reasonable to develop the estimators of the ω_j's by using the transformed data $\{ (L_i^*, H_i^*, \mathbf{X}_i);\ i = 1, ..., N \}$.

3. Feature Screening for Massive Interval-censored Data

In this section, as mentioned above, we will propose two estimators $\hat{\omega}_j$'s for the marginal correlation measure ω_j and then give two feature screening procedures or methods to estimate the index set of the active covariates $\mathcal{M}$. It is well-known that in the case of massive data or when the sample size N may be tens of thousands or more, computing the correlation between each covariate and the response variable can be computationally challenging. Also running such computations may result in memory issues and unacceptable computing times, and the information overlap can become a significant issue too. To address these issues, we propose first perform the orthogonal subsampling and then randomly divide the entire subsample into B segments. On each data segment, we will computer or estimate the distance correlation for each covariate to improve the calculation speed while ensuring that the samples overlap as little as possible. The estimates are then aggregated based on each subsample segment. In the following, we will first describe the orthogonal subsampling and distance correlation estimators and then the proposed feature screening methods.

3.1 Orthogonal subsampling

As mentioned above, the orthogonal subsampling (OSS) was initially proposed by Wang et al. (2021a) with the aim of selecting data points that are 'dissimilar' to approximate combinatorial optimization. This is because large datasets often contain overlapping information between data points and to ensure that a subsample covers diverse information, it is desirable to avoid including a data point and its 'similar' points. By 'similar', we mean that they share overlap information.

Assume that all covariates are scaled to the range of $[-1, 1]$, which can be accomplished by dividing each covariate by its largest absolute value observed. To avoid the confusion or for the generality, we will refer the resulting design matrix as to $\mathbf{Z}$. The optimality of orthogonal arrays is based on two features of their arrangements of row points. Firstly, the extreme values points that have large distances from the center are located at the corners of the data domain $[-1, 1]^p$, the space between -1 and 1 in the p-dimensional space. Secondly, the combinatorial orthogonality points (or more precisely, their signs) are as dissimilar as possible. The OSS approach then minimizes a discrepancy function that measures the distortion of a data point while maintaining the two features simultaneously.

More specifically, let $\mathbf{Z}_{i\cdot}$ denote the ith data point in the design matrix

$\mathbf{Z}$ and Z_{ij} the (i, j) element of $\mathbf{Z}$. The discrepancy function contains two parts targeting the two features. For the feature of extreme values, the function can include $p - \|\mathbf{Z}_{i\cdot}\|$, where $\|\cdot\|$ denotes the Euclidean norm. The dimension p is to ensure that the term $p - \|\mathbf{Z}_{i\cdot}\|$ is positive. For the feature of combinatorial orthogonality, define $\xi(\mathbf{Z}_{i\cdot}, \mathbf{Z}_{j\cdot}) = \sum_{l=1}^{p} \xi_1(Z_{il}, Z_{jl})$, where $\xi_1(x, y)$ is equal to 1 if both x and y have the same sign and 0 otherwise. It is easy to see that $\xi(\mathbf{Z}_{i\cdot}, \mathbf{Z}_{j\cdot})$ gives the number of the components in $\mathbf{Z}_{i\cdot}$ and $\mathbf{Z}_{j\cdot}$ that have the same signs. Combining the two parts, the discrepancy function is given by

$$L(\mathbf{Z}_{A\cdot}) = \sum_{\{i<j\}\in A} \left[p - \|\mathbf{Z}_{i\cdot}\|^2/2 - \|\mathbf{Z}_{j\cdot}\|^2/2 + \xi(\mathbf{Z}_{i\cdot}, \mathbf{Z}_{j\cdot})\right]^2 ,$$

where A denotes a subset of $\{1, \ldots, N\}$ with the cardinality K and $\mathbf{Z}_{A\cdot}$ a matrix with K rows and p columns consisting of data points $\mathbf{Z}_{i\cdot}, i \in A$. The indicator set for the subsample $\widetilde{A}$ is then obtained by solving the optimization problem

$$\mathbf{Z}_{\widetilde{A}\cdot} = \arg\min_{A\subseteq\{1,\ldots,N\}} L(\mathbf{Z}_{A\cdot}) \qquad s.t.\ \mathbf{Z}_{\widetilde{A}\cdot} \text{ contains } K \text{ points.} \tag{3.2}$$

The summary of the description above is given in Algorithm 1 of Supplementary Material.

Note for the computational complexity, one is usually interested in the situation when N is large. In this case, $\mathbf{Z}$ consists of $k_i = N/i$ points at

each iteration and the time for finding $\mathbf{z}$ becomes $O(Np/i)$. Therefore, the complexity for selecting n data points is $O(Np/1) + \ldots + O(Np/(K)) = O(Np\log(K))$. Note that n can be any integer less than N in the OSS approach, not restricted to be a multiple of 4.

3.2 Distance correlation measure based on subsamples

In this subsection, we discuss the computation and estimation of the correlation measure for each covariate based on the subsamples given by the orthogonal sampling on each segment. Without loss of generality, assume that the entire subsample $\mathcal{D}$ is randomly divided into B manageable segments $\{\mathcal{D}_b\}_{b=1}^{B}$, where $B = O(N^{\alpha}), \alpha \in (\frac{1}{5}, \frac{3}{5})$. So each segment contains $[K/B]$ samples, where $[\cdot]$ represents the rounding to the integer. As discussed above and depending on the computational environment, these segments can be either distributively stored and processed by multiple computers or sequentially processed by a single computer. Then the orthogonal subsampling is performed on each segment to yield n samples with $n = O(N^{\iota}), \iota \in (0, \frac{1}{5})$.

For the simplicity of notation, we will write the random vector $(L^*, H^*)^{\top}$ as $\mathbf{Y}$ in the following. Let $\phi_{X_j}(w)$ and $\phi_{\mathbf{Y}}(\boldsymbol{r})$ denote the characteristic functions of X_j and $\mathbf{Y}$, respectively, and $\phi_{X_j,\mathbf{Y}}(w, \boldsymbol{r})$ their joint characteristic function. Then by the definition, the nonnegative distance covariance be-

tween X_j and $\mathbf{Y}$ is given by

$$dcov^2(X_j, \mathbf{Y}) = \int_{R^{d_x+d_y}} \frac{\left\|\phi_{X_j,\mathbf{Y}}(w, \boldsymbol{r}) - \phi_{X_j}(w)\phi_{\mathbf{Y}}(\boldsymbol{r})\right\|^2}{c_{d_x}c_{d_y}\|w\|_{d_x}^{1+d_x}\|\boldsymbol{r}\|_{d_y}^{1+d_y}} dwd\boldsymbol{r},$$

where d_x and d_y represent the dimensions of X_j and $\mathbf{Y}$, respectively, $c_d = \pi^{(1+d)/2}/\Gamma((1+d)/2)$, and $\|\mathbf{a}\|_d$ denotes the Euclidean norm of $\mathbf{a} \in R^d$. Correspondingly, a correlation measure, the distance correlation (DC), can be computed as

$$\omega_j^{DC} = \frac{dcov^2(X_j, \mathbf{Y})}{\sqrt{dcov^2(X_j, X_j)dcov^2(\mathbf{Y}, \mathbf{Y})}},$$

which is equal to 0 if and only if X_j and $\mathbf{Y}$ are independent (Székely, Rizzo, and Bakirov, 2007; Zhang et al., 2023).

To estimate the ω_j's, note that it follows from Székely, Rizzo, and Bakirov (2007) that $dcov^2(X_j, \mathbf{Y})$ can be partitioned as $dcov^2(X_j, \mathbf{Y}) = S_{j1} + S_{j2}S_{j3} - 2S_{j4}$, where

$$S_{j1} = E(\|X_j - \tilde{X}_j\|_1\|\mathbf{Y} - \tilde{\mathbf{Y}}\|_2),$$

$$S_{j2} = E(\|X_j - \tilde{X}_j\|_1), \quad S_{j3} = E(\|\mathbf{Y} - \tilde{\mathbf{Y}}\|_2),$$

$$S_{j4} = E\{E(\|X_j - \tilde{X}_j\|_1|X_j)E(\|\mathbf{Y} - \tilde{\mathbf{Y}}\|_2|\mathbf{Y})\}$$

with $(\tilde{X}_j, \tilde{\mathbf{Y}})$ denoting an independent copy of $(X_j, \mathbf{Y})$. These suggest that we can obtain the unbiased estimators of S_{j1}, S_{j2}, S_{j3}, and S_{j4} given by the following U-statistics

$$\hat{S}_{j1} = \binom{n}{2}^{-1} \sum_{i,k}^{n} \frac{1}{2!} \sum_{\Omega\{i,k\}} \|X_{ji} - X_{jk}\|_1\|\mathbf{Y}_i - \mathbf{Y}_k\|_2$$

$$= \frac{1}{n(n-1)} \sum_{i \neq k}^{n} \|X_{ji} - X_{jk}\|_1 \|\mathbf{Y}_i - \mathbf{Y}_k\|_2 \,,$$

$$\hat{S}_{j2} = \binom{n}{2}^{-1} \sum_{i,k}^{n} \frac{1}{2!} \sum_{\Omega\{i,k\}} \|X_{ji} - X_{jk}\|_1 = \frac{1}{n(n-1)} \sum_{i \neq k}^{n} \|X_{ji} - X_{jk}\|_1 \,,$$

$$\hat{S}_{j3} = \binom{n}{2}^{-1} \sum_{i,k}^{n} \frac{1}{2!} \sum_{\Omega\{i,k\}} \|\mathbf{Y}_i - \mathbf{Y}_k\|_2 = \frac{1}{n(n-1)} \sum_{i \neq k}^{n} \|\mathbf{Y}_i - \mathbf{Y}_k\|_2 \,,$$

$$\hat{S}_{j4} = \binom{n}{3}^{-1} \sum_{i,k}^{n} \frac{1}{3!} \sum_{\Omega\{i,k,l\}} \|X_{ji} - X_{jl}\|_1 \|\mathbf{Y}_k - \mathbf{Y}_l\|_2$$

$$= \frac{1}{n(n-1)(n-2)} \sum_{i \neq k \neq l}^{n} \|X_{ji} - X_{jl}\|_1 \|\mathbf{Y}_k - \mathbf{Y}_l\|_2 \,.$$

In the above, $\Omega\{i,k\}$ and $\Omega\{i,k,l\}$ represent the set of all possible permutations $\{i,k\}$, and $\{i,k,l\}$, respectively. Similarly, we can obtain the unbiased estimators of $dcov^2(X_j, X_j)$ and $dcov^2(\mathbf{Y}, \mathbf{Y})$ and consequently an unbiased estimator of ω_j given by

$$\hat{\omega}_j^{DC} = \frac{\widehat{dcov^2}(X_j, \mathbf{Y})}{\sqrt{\widehat{dcov^2}(X_j, X_j)\widehat{dcov^2}(\mathbf{Y}, \mathbf{Y})}}. \tag{3.3}$$

The estimator $\hat{\omega}_j^{DC}$ is expected to fluctuate around zero if X_j is an inactive covariate and be away from zero otherwise.

3.3 Average distance correlation procedures of OSS

Now we are ready to present the two proposed screening methods. Let

$$\hat{\omega}_{(b),j}^{DC} = \frac{\widehat{dcov^2_{(b)}}(X_j, \mathbf{Y})}{\sqrt{\widehat{dcov^2_{(b)}}(X_j, X_j)\widehat{dcov^2_{(b)}}(\mathbf{Y}, \mathbf{Y})}}$$

denote the estimator given in (3.3) based on the data segment $\mathcal{D}_b$. Then it is natural to estimate the marginal correlation between X_j and $\mathbf{Y}$ based on the whole dataset by the simple average

$$\hat{\omega}_j^{SDC} = \frac{1}{B}\sum_{b=1}^{B} \hat{\omega}_{(b),j}^{DC}$$

over all data segments, $j = 1, ..., p$. The process or procedure is summarized in Algorithm 2 of Supplementary Material. It is apparent that the main advantage of the method based on $\hat{\omega}_j^{SDC}$ or the SDC-OSS is that it is straightforward and can be easily implemented. On the other hand, the measure $\hat{\omega}_j^{SDC}$ could be biased due to the accumulated bias inherited from the local estimators as seen in the numerical study below. To address this issue while maintaining a compact subsampling size, we propose the second method, JDC-OSS, or the jackknife debiased version of the first method.

Let $\mathcal{D}_b$ denote the subsample indicator set of the b-th segment, $\mathbf{X}_{(b,-m)} = (\mathbf{X}_i : i \in \mathcal{D}_b, j \neq m)^\top$, and $\mathbf{Y}_{(b,-m)} = (\mathbf{Y}_i : i \in \mathcal{D}_b, j \neq m)^\top$ represent the b-th subsample with the m-th subject removed. Based on $\mathbf{X}_{(b,-m)}$ and $\mathbf{Y}_{(b,-m)}$, we can obtain the leave-one-out estimator $\widehat{dcov^2}_{(b,-m)}(X_j, \mathbf{Y})$ of the distance covariance $dcov^2_{(b,-m)}(X_j, \mathbf{Y})$. Then we can estimate the bias by

$$\widehat{\triangle}_{(b)}(X_j, \mathbf{Y}) = \frac{(n-1)}{n}\sum_{m=1}^{n} \widehat{dcov^2}_{(b,-m)}(X_j, \mathbf{Y}) - (n-1)\widehat{dcov^2}_{(b)}(X_j, \mathbf{Y}), \quad (3.4)$$

and thus obtain the jackknife debiased simple average distance covariance estimator between X_j and $\mathbf{Y}$ given by

$$\widehat{dcov^2}^{JDC}(X_j, \mathbf{Y}) = \frac{1}{B}\sum_{b=1}^{B}\{\widehat{dcov^2}_{(b)}(X_j, \mathbf{Y}) - \widehat{\triangle}_{(b)}(X_j, \mathbf{Y})\}.$$

Similarly, we can obtain jackknife debiased simple average distance covariance estimators of $dcov^2(X_j, X_j)$ and $dcov^2(\mathbf{Y}, \mathbf{Y})$. These naturally give the jackknife debiased simple average distance correlation estimator

$$\hat{\omega}_j^{JDC} = \frac{\widehat{dcov^2}^{JDC}(X_j, \mathbf{Y})}{\sqrt{\widehat{dcov^2}^{JDC}(X_j, X_j)\widehat{dcov^2}^{JDC}(\mathbf{Y}, \mathbf{Y})}},$$

and thus the second proposed screening procedure, the JDC-OSS, which is summarized in Algorithm 3 of Supplementary Material. In the next section, we will establish the asymptotic properties of the two proposed methods that replace the $\hat{\omega}_j$'s in (2.1) by the $\hat{\omega}_j^{SDC}$'s and $\hat{\omega}_j^{JDC}$'s, respectively.

4. Asymptotic Properties

To establish the asymptotic properties of the two proposed screening procedures, SDC-OSS and JDC-OSS, we need the following regularity conditions.

(C1) Assume that there exists $\eta > 0$ such that $P(V - U \geq \eta) = 1$. The union of the supports of U and V is contained in the interval $[\sigma, \tau]$, where $0 < \sigma < \tau < \infty$.

(C2) There exists a positive constant s_0 such that for all $0 < s \leq 2s_0$, we have $\sup_p \max_{1\leq k\leq p} E\{\exp(s\|X_k\|_1^2)\} < \infty$, and $E\{\exp(s\|\max\{U, V\}\|_1^2)\} < \infty$.

(C3) The minimum distance correlation of active predictors satisfies $\min_{k\in\mathcal{M}} \omega_k \geq 2cN^{-(\alpha+\iota)\kappa}$ for some constants $c > 0$ and $\kappa \in [0, 1/2]$.

Note that Condition (C1) is commonly used in the studies of interval-censored data (Zhang et al., 2010; Zhou et al., 2017) and usually satisfied in practice. Condition (C2) is a common assumption required by most of the existing screening procedures (Zhang et al., 2023). Condition (C3) requires that the values of marginal utilities between each active variable and response are not too small. It is also a standard assumption in the feature screening literature and similar to Condition 3 of Fan and Lv (2008), Condition 2 of Li et al. (2012b), and Conditions 2 and 5 of Wu and Cook (2015) among others.

In the following, we will first establish the asymptotic order of the variances of the two proposed estimators $\widehat{\omega}_j^{SDC}$ and $\hat{\omega}_j^{JDC}$. Note that both $\widehat{\omega}_j^{SDC}$ and $\hat{\omega}_j^{JDC}$ are unbiased and thus their estimation accuracies are determined by their variances. Then in Theorem 2 and 3, we will give the sure screening property and rank consistency property of the proposed method, respectively. For simplicity, we use the $\hat{\omega}_j$'s to denote both $\hat{\omega}_j^{SDC}$'s and $\hat{\omega}_j^{JDC}$'s in Theorems 2 and 3 with the proof for all results sketched in the Supplementary Material.

Theorem 1. *Assume that* $B = O(N^{\alpha})$ *with* $\alpha \in (\frac{1}{5}, \frac{3}{5})$ *and* $n = O(N^{\iota})$ *with* $\iota \in (0, \frac{1}{5})$. *Then under Condition (C2), we have that*

$$\max_{j=1,\ldots,p} Var(\widehat{\omega}_j^{SDC}) = O\Big(\frac{1}{N^{\alpha+2\iota}}\Big) + O\Big(\frac{1}{N^{\alpha+5/2\iota}}\Big) + O\Big(\frac{1}{N^{\alpha+3\iota}}\Big),$$

and

$$\max_{j=1,\ldots,p} Var(\widehat{\omega}_j^{JDC}) = O\Big(\frac{1}{N^{\alpha}}\Big) + O\Big(\frac{1}{N^{\alpha+1/2\iota}}\Big) + O\Big(\frac{1}{N^{\alpha+\iota}}\Big).$$

The results above tell us that the largest variances of the estimators $\widehat{\omega}_j^{SDC}$ and $\widehat{\omega}_j^{JDC}$ have the orders $O\Big(\frac{1}{N^{\alpha+2\iota}}\Big)$ and $O\Big(\frac{1}{N^{\alpha}}\Big)$, respectively. That is, both variances converge to zero under massive data.

Theorem 2. *[The sure screening property of SDC-OSS and JDC-OSS]. Assume that* $B = O(N^{\alpha})$ *with* $\alpha \in (\frac{1}{5}, \frac{3}{5})$, *and* $n = O(N^{\iota})$ *with* $\iota \in (0, \frac{1}{5})$. *Also assume that Conditions (C1) and (C2) hold. Then for any* $0 < \gamma < 1/2 - \kappa$ *with* $\kappa \in [0, 1/2]$, *there exist positive constants* $c, c_1,$ *and* c_2 *such that*

$$\Pr\left(\max_{1\le j\le p} |\hat{\omega}_j - \omega_j| \ge cN^{-(\alpha+\iota)\kappa}\right) \le O\left(p\exp(-c_1 N^{(\alpha+\iota)(1-2\gamma-2\kappa)}) + N^{\alpha+\iota} p\exp(-c_2 N^{(\alpha+\iota)\gamma})\right).$$

In addition, under Conditions (C1)-(C3), we have that

$$\Pr(\mathcal{M} \subseteq \widehat{\mathcal{M}}) \ge 1 - O\left(|\mathcal{M}|\left[\exp(-c_1 N^{(\alpha+\iota)(1-2\gamma-2\kappa)}) + N^{\alpha+\iota}\exp(-c_2 N^{(\alpha+\iota)\gamma})\right]\right),$$

where $|\mathcal{M}|$ *denotes the cardinality of* $\mathcal{M}$.

Remark 1: Theorem 2 is usually called the sure screening property and it establishes the connection between the estimated and true correlation measures between covariates and a response variable. Also it provides a bound on the probability that the genuinely significant set of variables is contained within the estimated important set. The theorem tells us that the probability of the event 'there exists a j such that the distance between $\hat{\omega}_j$ and ω_j is greater than $cN^{-(\alpha+\iota)\kappa}$,

converges to 0 at the exponential rate $O(p\exp(-cN^{(\alpha+\iota)\max\{1-2\gamma-2\kappa,\gamma\}}))$. Similarly, the probability of the event '$\mathcal{M} \subseteq \widehat{\mathcal{M}}$' converges to 1 at the same rate.

Theorem 3. *[The rank consistency property of SDC-OSS and JDC-OSS]. Assume that* $B = O(N^{\alpha})$ *with* $\alpha \in (\frac{1}{5}, \frac{3}{5})$, *and* $n = O(N^{\iota})$ *with* $\iota \in (0, \frac{1}{5})$. *Also suppose that* $\omega_j = 0$ *for* $j \notin \mathcal{M}$ *and Conditions (C1)-(C3) given above hold. Then there exist positive constants* c, c_1 *and* c_2 *such that*

$$Pr\Big(\max_{j\notin\mathcal{M}} \hat{\omega}_j \leq \min_{j\in\mathcal{M}} \widehat{\omega}_j\Big) \geq 1 - O\left(p\left[\exp(-c_1 N^{(\alpha+\iota)(1-2\gamma-2\kappa)}) + N^{\alpha+\iota}\exp(-c_2 N^{(\alpha+\iota)\gamma})\right]\right).$$

Remark 2: Theorem 3 is commonly referred to as the rank consistency property. It states that the probability of the estimated correlation measure for irrelevant variables being lesser than the estimated correlation measure for significant variables converges to 1 at the exponential speed $O(p\exp(-cN^{(\alpha+\iota)\max\{1-2\gamma-2\kappa,\gamma\}}))$.

Remark 3: Note that in Theorems 2 and 3 there are only the conditions for the number of segments B and the subsample size n within each segment. Other existing sampling methods (e.g., simple random subsampling, random addressing sampling(Zhu et al., 2022), sequential addressing subsampling(Pan et al., 2023), etc.) can also set the number of segments B and the subsample size n within each segment to satisfy the conditions in the above theorems. Therefore, the theoretical framework given in the manuscript remains broadly applicable across various sampling methods, not just OSS.

Remark 4: On the comparison of the proposed method to that given in

Zhang et al. (2023), note that the latter is effective only for hundreds of samples and cannot be implemented for massive sample sizes due to its computational complexity of $O(pN^3)$. In contrast, the total computational complexity of the proposed method is $O(Nplog(nB)) + O(Bpn^3) = O(p((\alpha+\iota)N + N^{\alpha+3\iota})) = O(pN^{\max\{\alpha+3\iota,1\}})$. Thus the proposed method effectively solves the computational challenges under massive data and significantly reduces the computational complexity from $O(pN^3)$ to $O(pN^{\max\{\alpha+3\iota,1\}})$. Another difference between the two methods is that unlike or beyond Zhang et al. (2023), we showed in Theorem 1 that the $\hat{\omega}_j$'s are effective and efficient in estimating the ω_j's. This further ensures the screening performance of the proposed method and serves as a theoretical foundation of the proposed method. In addition, the methodology used in the proof of Theorem 2 is significantly different from that of Zhang et al. (2023).

5. A Simulation Study

In this section, we present some results obtained from an extensive simulation study conducted to investigate the finite sample performance of the two screening procedures, SDC-OSS and JDC-OSS, proposed in the previous sections. For comparison, we also considered the divide-and-conquer (DC) method, which will be referred to as DC-DC, and the simple random subsampling combined with JDC and SDC, which will be abbreviated as JDC-RSS and SDC-RSS, respectively. Note that because of the large sample size, it is not possible to conduct the

DC screening based on the complete data using standard computers. To generate the true failure times, we considered the following three set-ups or cases.

Model 1. We first generated the covariate vector $\mathbf{X}$ from the normal distribution $N_p(\mathbf{0}, \mathbf{\Sigma})$ with $\mathbf{\Sigma}_{i,j} = (0.3^{|i-j|})$ for $i, j = 1, \ldots, p$. The true failure time of interest T_i was then generated under the Cox Proportional Hazards model with the baseline hazard function $\lambda(t) = (t - 0.5)^2$ and the true parameter $\boldsymbol{\beta} = (\mathbf{1}_{10}, \mathbf{0}_{p-10})^\top$, where $\mathbf{1}_{10}$ represents the 10-dimensional vector with all elements being 1. That is, there exist ten important or relevant variables.

Model 2. In this setup, the covariate vector $\mathbf{X}$ was generated in the same way as above but with $\mathbf{\Sigma}_{i,j} = (0.5^{|i-j|})$ for $i, j = 1, \ldots, p$, and the true failure time of interest was generated under the transformation model $H(T) = -\mathbf{X}^\top \boldsymbol{\beta} + \epsilon$. For the selection of the pre-specified function $H(\cdot)$, the true parameter $\boldsymbol{\beta}$ and the distribution of ϵ, we considered the following three scenarios.

(a) $H(t) = \log(0.5(e^{2t} - 1))$, ϵ follows the Student t distribution with 3 degrees of freedom, and $\boldsymbol{\beta} = (1, 0.7, \mathbf{0}_6, 0.8, 1, \mathbf{0}_{p-10})^\top$.

(b) Both the function $H(\cdot)$ and the distribution of ϵ were set to be the same as in (a). For the regression coefficients, we took the first ten components to be $\boldsymbol{\beta}_0 = (\beta_1, \ldots, \beta_{10})^\top = \alpha_0(-1)^{U_1}U_2$ and the remaining to be 0. Here U_1 was generated from the Bernoulli distribution $B(0.6)$, U_2 generated from the uniform distribution $U(1, 2)$, and $\alpha_0 = 1$.

(c) For this scenario, we took $H(t) = \log(t)$ and assumed that ϵ follows the

standard normal distribution. For the regression coefficients, we took the first six components as non-zero coefficients, and their values are generated in the manner described in (b). All of the other coefficients were set to be zero.

Model 3. For this model, all set-ups are the same as Model 1 except that we generated the failure time of interest from the log-linear model $\log(T) = \alpha + \mathbf{X}^\top \boldsymbol{\beta} + \sigma\epsilon$, where $\alpha = 0.4$, $\sigma = 2$ and ϵ follows the standard normal distribution.

For the generation of the observed interval-censored data, for each subject, we first randomly generated m values $0 < t_1 < \ldots < t_m < \tau$ from the uniform distribution $U(0, \tau)$ to yield $m + 1$ intervals $[t_0, t_1), [t_1, t_2), \ldots, [t_m, t_{m+1})$, where $t_0 = 0$, $t_{m+1} = \infty$, τ was chosen to give the required right-censored percentage. The observation interval $[U_i, V_i)$ was then taken to be the interval $[t_j, t_{j+1})$ that includes the true failure time T_i. In the following, we set $m = 20$ and $\eta = 10^6$ and considered the right-censored rates of low (20%), moderate (40%) and high (60%). The screening results given below are based on $N = 10^5$ and $p = 1000$ with 200 replications.

To evaluate the performance of the proposed procedures, we calculated the following four metrics.

(1) time (sec): the execution time of the respective method in seconds across 200 replications.

(2) PA: the proportion of all active predictors selected for a given model size $d_0 = \lceil 5(nB)^{(1/5-1/500)} \rceil$ in 200 replications.

(3) S: the 5%, 25%, 50%, 75%, and 95% quantiles of the minimum model size to include all active predictors over 200 replications.

(4) AUS: the rank consistency index given by

$$AUS = 1 - \Big(\frac{1}{|\mathcal{M}||\mathcal{M}^c|}\sum_{i\in\mathcal{M}}\sum_{j\in\mathcal{M}^c}\big[I\{\hat{\omega}_i < \hat{\omega}_j\} + 0.5I\{\hat{\omega}_i = \hat{\omega}_j\}\big]\Big),$$

where $|\mathcal{M}|$ and $|\mathcal{M}^c|$ denote the sizes of the sets of true important and non-important variables, respectively.

First we present the time used by the orthogonal subsampling across various total sample sizes and sub-sample sizes, with the covariate dimension of $p = 1000$. It shows that the execution time of orthogonal subsampling aligns with the theoretical expectation of $O(Np\log(nB))$, as described in Section 3.1. Table 2 gives the outcomes of the two proposed feature screening methods, SDC-OSS and JDC-OSS, in comparison with DC-DC, JDC-RSS, and SDC-OSS under **Model 1** with $n = 10$ or 15 and $B = 100$ or 150. One can see from Table 2 that both methods seem to perform well for the considered situations and this is especially the case with larger total subsample sizes (nB). That is, a sufficiently large total subsample size (nB) can lead to optimal performance, where the important variables can be selected before the unimportant ones and the probability of correctly selecting all of the essential variables is close to 1. More specifically, the two proposed methods gave similar performance, especially in terms of the AUC and PA measures or for large total subsample size (nB), with the JDC-

Table 1: The times (sec) required by OSS under different settings

N \ nB	100	1000	1500	3000	5000
10^5	46.67971	241.6879	442.1897	1354.251	3211.391
10^6	592.0568	921.4064	1410.246	3418.268	7295.892

OSS slightly better than the SDC-OSS as expected. In other words, the main difference between the two procedures occurred at the higher quantiles of the S measure and smaller total subsample sizes.

In terms of comparison, one can see from Table 2 that the DC-DC method can provide effective screening but it can be slower in hundred times than the proposed method. Also the proposed method attains satisfactory screening outcomes when the total subsample size (nB) is sufficiently large. That is, a sufficiently large total subsample size (nB) can lead to optimal performance, meaning that the important variables can be selected before the unimportant ones and the probability of correctly selecting all of the essential variables is close to 1. Furthermore, the screening method combining JDC (and SDC) with simple random sampling (RSS) exhibited inferior screening performance compared to JDC (and SDC) with orthogonal subsampling (OSS).

To save the space, the results obtained under **Models 2** and **3** are given in Tables 1 - 4 of Supplementary Material. Here we have that $n = 10$, 15 or 20 and $B = 25$, 40, 60, 80, 100, 150, 200 or 300 for **Model 2**, and for **Model 3**, $n = 8$

Table 2: Simulation results under Model 1.

n	$B(d_0)$	$Method$	time(sec)	$AUC(\%)$	PA	S 5%	25%	50%	75%	95%
Right-censored rate = 20%										
100	N/B(20)	DC-DC	15629.24	100.000	1.00	10	10	10	10	10
	10(20)	JDC-OSS	132.74	99.471	0.48	10	12.75	22.5	54.75	228.1
		SDC-OSS	103.14	99.450	0.48	10	13	23	57	242.9
		JDC-RSS	34.26	99.277	0.31	11	17	35	78.5	241.45
		SDC-RSS	5.44	99.297	0.30	11	17.75	34.5	85	243
	15(21)	JDC-OSS	503.44	99.975	0.95	10	10	10	12	18.3
		SDC-OSS	425.83	99.975	0.95	10	10	10	12	22
		JDC-RSS	91.87	99.927	0.93	10	10	10	12	31.05
		SDC-RSS	15.77	99.908	0.91	10	10	10	13	38.1
150	N/B(18)	DC-DC	10500.55	100.000	1.00	10	10	10	10	10
	10(21)	JDC-OSS	325.50	99.835	0.74	10	10	11	22	84.45
		SDC-OSS	273.94	99.814	0.78	10	10	12	19	98.45
		JDC-RSS	52.52	99.786	0.71	10	10	13	25.25	125.05
		SDC-RSS	10.92	99.725	0.65	10	11	14	31.25	131.3
	15(23)	JDC-OSS	474.57	99.991	0.99	10	10	10	10	12
		SDC-OSS	414.56	99.988	0.99	10	10	10	10	12
		JDC-RSS	66.88	99.994	0.99	10	10	10	10	13
		SDC-RSS	9.92	99.995	1.00	10	10	10	10	13
Right-censored rate = 40%										
100	N/B(20)	DC-DC	15611.10	100.000	1.00	10	10	10	10	10
	10(20)	JDC-OSS	135.47	99.706	0.68	10	11	13	25.25	150.2
		SDC-OSS	104.66	99.687	0.67	10	11	14	28.25	180.05
		JDC-RSS	35.83	99.615	0.57	10	11	17	46.25	193
		SDC-RSS	5.53	99.584	0.54	10	12	17	46.25	164.1
	15(21)	JDC-OSS	503.78	99.995	0.99	10	10	10	10	13.05
		SDC-OSS	421.29	99.993	1.00	10	10	10	10	12
		JDC-RSS	96.10	99.989	0.98	10	10	10	10	14
		SDC-RSS	15.73	99.988	0.98	10	10	10	10	13.05
150	N/B(18)	DC-DC	10557.73	100.000	1.00	10	10	10	10	10
	10(21)	JDC-OSS	362.10	99.946	0.91	10	10	10	13	33.45
		SDC-OSS	302.08	99.936	0.90	10	10	10	14	41.05
		JDC-RSS	58.27	99.886	0.89	10	10	10	13	53.15
		SDC-RSS	11.83	99.855	0.84	10	10	10.5	14	88
	15(23)	JDC-OSS	453.07	100.000	1.00	10	10	10	10	10
		SDC-OSS	395.47	100.000	1.00	10	10	10	10	10
		JDC-RSS	65.91	99.998	1.00	10	10	10	10	10
		SDC-RSS	9.55	99.999	1.00	10	10	10	10	10
Right-censored rate = 60%										
100	N/B(20)	DC-DC	15568.42	100.000	1.00	10	10	10	10	10
	10(20)	JDC-OSS	128.70	99.749	0.66	10	11	15	29.5	119.85
		SDC-OSS	99.50	99.697	0.67	10	11	15	27.25	188.65
		JDC-RSS	33.81	99.623	0.52	10	12	19	38	207.45
		SDC-RSS	5.47	99.595	0.53	10	12.75	19	48.75	183.2
	15(21)	JDC-OSS	497.38	99.989	0.99	10	10	10	10	13
		SDC-OSS	417.26	99.988	0.98	10	10	10	10	13.05
		JDC-RSS	94.92	99.985	0.98	10	10	10	10	15
		SDC-RSS	15.84	99.975	0.97	10	10	10	10	17.1
150	N/B(18)	DC-DC	10525.82	100.000	1.00	10	10	10	10	10
	10(21)	JDC-OSS	367.93	99.948	0.91	10	10	11	14	29.15
		SDC-OSS	304.56	99.947	0.90	10	10	11	15	35.2
		JDC-RSS	57.70	99.934	0.89	10	10	10	14	30.15
		SDC-RSS	11.98	99.917	0.84	10	10	11	15	55.05
	15(23)	JDC-OSS	502.84	100.000	1.00	10	10	10	10	10
		SDC-OSS	441.43	99.999	1.00	10	10	10	10	10
		JDC-RSS	68.41	99.999	1.00	10	10	10	10	10
		SDC-RSS	10.08	100.000	1.00	10	10	10	10	10

or 10 and $B = 80$, 100, 150 or 200. The results are similar to those given above and again suggest that the two proposed approaches performed well in general, which is especially the case when the larger total sub-sample size (nB) was used. In particular, they indicate that as expected from the asymptotic properties, the proposed methods can achieve the optimal performance to identify all important variables before the unimportant ones with a high probability of correctly selecting all important variables with a sufficiently large total subsample size. Furthermore, the performance seems to be robust with respect to the underlying model used to generate the true failure time of interest. Also again as seen before, the main difference between the two proposed methods occurred when the total subsample size is small.

In addition, we also carried out the assessment of the two proposed procedures in terms of the False Discovery Rate (FDR) control and the results obtained under **Model 1** are provided in Supplementary Material. In particular, they indicated that the two propsoed methods gave good and consistent performances. Also they are similar in terms of FDR and power.

6. An Application

In this section, we apply the two feature screening procedures proposed in the previous sections to the data on the survival time of breast cancer patients in the Surveillance, Epidemiology, and End Results (SEER) program, a commonly used

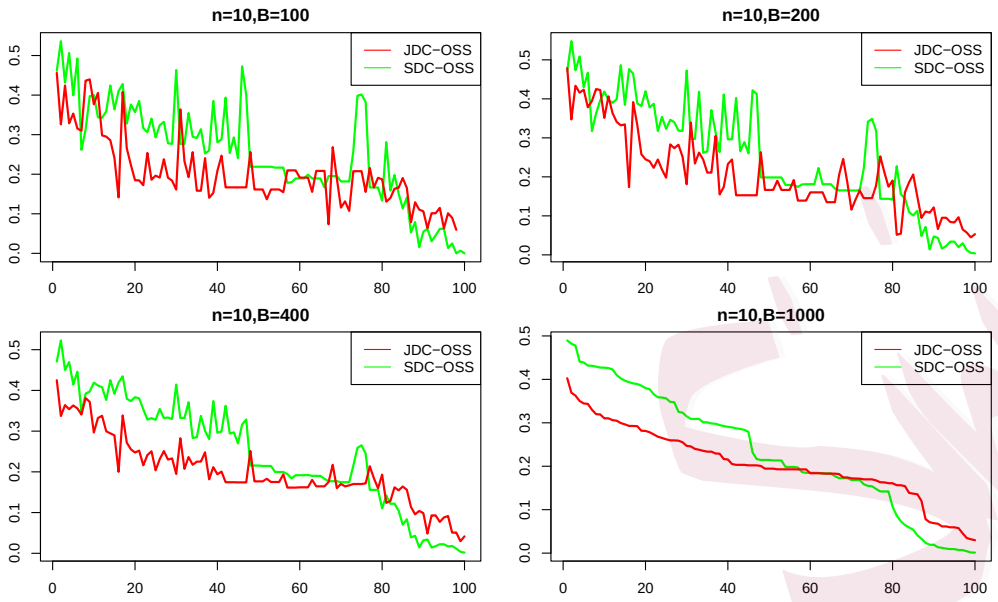

Figure 1: The comparison of JDC-OSS and SDC-OSS in different B with $n = 10$.

data source in cancer research (https://seer.cancer.gov). The SEER program was established as one of the first steps in the War on Cancer declared by President Nixon's Administration and began collecting information on January 1, 1973 in some of US states with other areas added to the SEER database over the years. After the year 2000, the SEER captured approximately 25% of all cancers diagnosed in the United States each year.

In the following, we consider the data released in April 2021 and based on the November 2020 submission. The dataset includes the information on 157,905 breast cancer patients diagnosed between 2010 and 2015 and 191 covariates, including sex, age, race, year of diagnosis, marital status, and the information

on initial treatment as well as the survival time if the death happened. For the analysis here, we are mainly interested in the death time, the failure time of interest, of the cancer patient, on which only interval-censored data are available due to the periodic collection nature of the data.

Figure 1 presents the line plots of the estimated correlation measures under different settings for the two methods proposed in the previous sections. The plots are based on the top 100 covariate rankings from the highest segment for fixed per-segment sample size $n = 10$ and four different segment sizes $B = 100, 200, 400$ and 1000. It is apparent that for both procedures, as seen in the simulation study, a larger total subsample size yielded the improved screening performance and the corresponding measure value became more stable or less fluctuated. It can be observed that under a fixed n and as the number of segments increases, the decreasing trend tends to flatten. Meanwhile, for a smaller number of segments, the fluctuation of the latter part of the SDC-OSS method is relatively large. In other words, the screening stability of the JDC-OSS method is higher than that of the SDC-OSS method.

Furthermore, similar to Figure 1, Figure 2 in S2 of the Supplementary Material provides the line plots of the correlation measures for the two methods JDC-OSS and SDC-OSS based on the rankings of the top 100 covariates under different settings of fixed segments $B = 300$ and varying per-segment sample size $n = 5, 10, 20$ and 50. From the figure, it can be observed that the fluctuation am-

Table 3: The screening results for the SEER data with all 191 covariates.

Covariates	DC (B=1000) (n=N/n)	JDC-OSS (B=1000) (n=8)	SDC-OSS (B=1000) (n=8)	JDC-RSS (B=1000) (n=8)	SDC-RSS (B=1000) (n=8)	DC (B=200) (n=N/n)	JDC-OSS (B=200) (n=8)	SDC-OSS (B=200) (n=8)	JDC-RSS (B=200) (n=8)	SDC-RSS (B=200) (n=8)
Age		√	√	√	√		√	√	√	√
Year.of.diagnosis		√	√	√	√	√	√	√	√	√
CS.version.input.original	√	√	√		√	√	√	√		√
CS.version.input.current		√	√	√	√	√	√	√	√	√
Regional.nodes.examined		√	√	√	√		√	√	√	√
CS.tumor.size		√	√		√		√	√		√
RX.Summ.Surg.Prim.Site		√	√	√	√		√	√	√	√
Breast.T	√	√	√	√	√		√	√	√	√
Breast.Stage		√	√	√	√		√	√	√	√
Primary.Site		√	√	√	√		√	√		√
Marital.status.at.diagnosis		√	√	√	√		√	√	√	√
CS.lymph.nodes		√		√	√		√			√
Radiation.recode	√	√	√	√	√		√	√		√
Breast.N		√		√	√	√	√	√	√	√
Regional.nodes.positive		√					√			√
AYA.site.recode.2020.Revision	√	√	√							
Histologic.Type.ICD.O.3	√	√	√	√			√	√		
Histology.ICD.O.2	√	√	√				√	√		

plitude of the SDC-OSS method is significantly higher than that of the JDC-OSS method. Similarly, it can be inferred that JDC-OSS performs better in terms of screening stability.

Table 3 presents the screening results given by the DC-DC method, which divided the full sample into B segments for divide-and-conquer analysis, along with the JDC-OSS, SDC-OSS, JDC-RSS, and SDC-RSS methods. The settings are the same as described in the simulation section, where B represents the number of segments and n is the sample size within each segment. To maintain the consistency in the number of screenings, we artificially set the threshold value d_0 to 20 across all methods. It can be observed from Table 3 that different number of segments in the full samples corresponds to a small overlap in important

variables screened by the DC-DC method. This indicates that the method is less stable. The stability of JDC and SDC based on two different sampling methods is relatively better, and the stability of the two methods based on OSS sampling is better than that of the two methods based on RSS method.

On the identified factors,, the covariates 'Radiation.recode' represent methods of radiation therapy, 'Primary.Site' indicate tumor location information, 'CS.tumor.size' denote tumor size information, 'RX.Summ.Surg.Prim.Site' describe surgical conditions at the primary site, 'Breast.Stage' 'Breast.N' and 'Breast.T' respectively represent different stages of breast tumors, while 'CS.lymph.nodes' signify expanded tumor information. Clearly, from a practical perspective, all of the aforementioned variables significantly impact the survival time or time of death. Moreover, the aforementioned screening methods effectively identify these variables as important variables.

To further see the effectiveness of the proposed methods, we apply the methods above to a set of randomly selected 50 variables along with their interactions, resulting in a total of 1275 variables, and the results are given in Table 5 in S2 of the Supplementary Material. Here as above, B represents the number of segments and n represents the sample size within each segment. From the table, it is apparent that there is a high degree of overlaps in the important variables selected by the five methods. In particular, the methods based on orthogonal subsampling seem to be more stable compared to these based on random sub-

sampling.

7. Discussion and Concluding Remark

This paper discussed feature screening for massive interval-censored failure time data where both the sample size and the number of covariates or factors are huge. For the problem, two distance correlation and orthogonal sampling-based screening methods were developed with the use of the jackknife debiasing technique in the second method. Both methods are model-free and thus offer a broader range of applications without model constraints. In other words, the proposed methods can efficiently handle large-scale interval-censored datasets under different models by utilizing orthogonal sampling to extract a small number of samples comprising the maximum amount of information for segmented massive data. Furthermore, the proposed procedures have been shown to possess the sure screening and rank consistency properties, and the numerical studies demonstrated that they performed well in different settings.

It is apparent that the first proposed method, SDC-OSS, can be relatively easily implemented compared to the second proposed method, JDC-OSS, but the latter may be more stable and perform better than the former when the total subsample size is small. Nevertheless, the two should give similar performances if a larger total subsample size can be used, which could be realized if more individual computers are available. A possible shortcoming of the two procedures

is that both are marginal screening methods and thus cannot take into account the relationship between covariates or factors.

In the previous sections, we only considered case II interval-censored data and the ideas discussed above should be applicable to more general situations such as case K interval-censored data and truncated interval-censored data. However, more work would be needed to investigate, for example, the effectiveness of the generalized methods as well as their theoretical properties. Another direction for future research is to establish the theoretical properties of the FDR control for the two screening methods proposed above.

Supplementary Material

The online Supplementary Material includes the three algorithms mentioned above, some additional simulation results, and the proofs of all the theorems.

Acknowledgements

The authors wish to thank the Co-Editor, Dr. Huixia Wang, the Associate Editor and two reviewers for their many helpful and insightful comments and suggestions that greatly improved the paper.The research was partially supported by a grant from the National Key R&D Program of China(Grant Number 2022YFA1003701), a grant from the Natural Science Foundation of China [Grant Number 12261102], and the grants from Yunnan Fundamental Research Project,

China [Grant Numbers 202201BF070001-004, 202301AS070044,202401AS070152].

References

Battey, H., Fan, J., Liu, H., Lu, J., and Zhu, Z. (2018). Distributed testing and estimation under sparse high dimensional models. *Annals of Statistics* **46(3)**, 1352–1382.

Dai, J., Liu, Y., Chen, J. and Liu, X. (2020). Fast feature selection for interval-valued data through kernel density estimation entropy. International Journal of Machine Learning and Cybernetics **11**, 2607–2624.

Fan, J., Feng, Y., and Song, R. (2011). Nonparametric independence screening in sparse ultra-high-dimensional additive models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **70(5)**, 849–911.

Fan, J., Li, R., Zhang, C., and Zou, H. (2020). *Statistical Foundations of Data Science*. Boca Raton, FL: CRC press.

Fan, J., and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **70(5)**, 849–911.

Fan, J. , and Song, R. (2010). Sure independence screening in generalized linear models with np-dimensionality. *Annals of Statistics* **38(6)**, 3567–3604.

Fan, Y., and Sun, J. (2021). Subsampling from features in large regression to find "winning features". *Statistical Analysis and Data Mining* **14**, 168–184.

Kawaguchi, E., Suchard, M., Liu, Z., and Li, G. (2020). A surrogate l_0 sparse Cox's regres-

sion with applications to sparse high-dimensional massive sample size time-to-event data. *Statistics in Medicine* **39(6)**, 675–686.

Kalbfleisch, J., and Prentice, R. (2002). *The statistical analysis of failure time data, 2nd edition.* John Wiley and Sons, New York.

Kong, E., and Xia, Y. (2019). On the efficiency of online approach to nonparametric smoothing of big data. *Statistica Sinica* **29**, 185–201.

Li, G., Peng, H., Zhang, J., and Zhu, L. (2012a). Robust rank correlation based screening. *The Annals of Statistics* **40(3)**, 1846–1877.

Li, R., Zhong, W., and Zhu, L. (2012b). Feature screening via distance correlation learning. *Journal of the American Statistical Association* **107(499)**, 1129–1139.

Li, X., and Xu, C. (2023). Feature screening with conditional rank utility for big-data classification. *Journal of the American Statistical Association* **119(546)**, 1385–1395.

Luo, L., and Song, P. (2020). Renewable estimation and incremental inference in generalized linear models with streaming data sets. *Journal of the Royal Statistical Society, Series B* **82**, 69–97.

Ma, P., Mahoney, M., and Yu, B. (2015). A statistical perspective on algorithmic leveraging. *Journal of Machine Learning Research* **16(1)**, 861–911.

Pan, R., Zhu, Y., Guo, B., Zhu, X., and Wang, H. (2023). A sequential addressing subsampling method for massive data analysis under memory constraint. *IEEE Transactions on Knowledge and Data Engineering* **35(9)**, 9502-9513.

Peng, Y., and Zhang, Q. (2021). Feature Selection for Interval-Valued Data Based on D-S Evidence Theory. *IEEE Access* **9**, 122754–122765.

Schifano, E., Wu, J., Wang, C., Yan, J., and Chen, M. (2016). Online updating of statistical inference in the big data setting. *Technometrics* **58**, 393–403.

Shi, C., Lu, W., and Song, R. (2018). A massive data framework for M-estimators with cubic-rate. *Journal of the American Statistical Association* **113(524)**, 1698–1709.

Shu, W., Chen, T., Cao, D. and Qian, Q. (2024). Incremental feature selection based on uncertainty measure for dynamic interval-valued data. International Journal of Machine Learning and Cybernetics **15**, 1453–1472.

Sun, J. (2006). *The Statistical Analysis of Interval-Censored Failure Time Data*. New York, NY: Springer.

Székely, G., Rizzo, M., and Bakirov, N. (2007). Measuring and testing dependence by correlation of distances. *The Annals of Statistics* **35**, 2769–2794.

Wang, H. (2019). More efficient estimation for logistic regression with optimal subsample. *Journal of Machine Learning Research* **20**, 1–59.

Wang, C., Chen, M., Wu, J., Yan, J., Zhang, Y., and Schifano, E. (2018a). Online updating method with new variables for big data streams. *Canadian Journal of Statistics* **46**, 123–146.

Wang, L., Elmstedt, J., Wong, W., and Xu, H. (2021a). Orthogonal subsampling for big data linear regression. *The Annals of Applied Statistics* **15(3)**, 1273–1290. DOI: 10.1214/21-

AOAS1462

Wang, Y., Hong, C., Palmer, N., Di, Q., Di, Q., Schwartz, J., Kohane, I., and Cai, T. (2021b). A fast divide-and-conquer sparse Cox regression. *Biostatistics* **22(2)**, 381-401.

Wang, H., Yang, M., and Stufken, J. (2019). Information-based optimal subdata selection for big data linear regression. *Journal of the American Statistical Association* **114(525)**, 393–405.

Wang, H., Zhu, R., and Ma, P. (2018b). Optimal subsampling for large sample logistic regression. *Journal of the American Statistical Association* **113(522)**, 829–844.

Wu, Y., and Cook, R. (2015). Penalized regression for interval-censored times of disease progression: selection of HLA markers in psoriatic arthritis. *Biometrics* **71**, 782–791.

Wu, J., Chen, M., Schifano, E., and Yan, J. (2021). Online updating of survival analysis. *Journal of Computational and Graphical Statistics* **30(4)**, 1209–1223. doi: 10.1080/10618600.2020.1870481.

Wu, Y., and Yin, G. (2015). Conditional quantile screening in ultrahigh-dimensional heterogeneous data. *Biometrika* **102(1)**, 65-76.

Xue, Y., Wang, H., Yan, J., and Schifano, E. (2019). An online updating approach for testing the proportional hazards assumption with streams of survival data. *Biometrics* **76(1)**, 171–182.

Yang, J., Schuemie, M., Ji, X., and Suchard, M. (2023). Massive Parallelization of Massive Sample-Size Survival Analysis. *Journal of Computational and Graphical Statistics* **0(0)**, 1–14.

Zhang, J., Du, M., Liu, Y., and Sun, J. (2023). A new model-free feature screening procedure for ultrahigh-dimensional interval-censored failure time data. *Statistica Sinica* **33**, 1809-1830.

Zhang, X., and Feng, Z. (2024). Feature selection based on contradictory state sequence for multi-scale interval valued decision table. Information Sciences **677**, 120926.

Zhang, Y., Hua, L., and Huang, J. (2010). A spline-based semiparametric maximum likelihood estimation method for the cox model with interval-censored data. *Scandinavian Journal of Statistics* **37**, 338–354.

Zhao, T., Cheng, G., and Liu, H. (2016). A partially linear framework for massive heterogeneous data. *Annals of Statistics* **44(4)**, 1400–1437.

Zhong, W., Zhu, L., Li, R., and Cui, H. (2016). Regularized quantile regression and robust feature screening for single index models. Statistica Sinica **26(1)**, 69–95.

Zhong, W., Qian, C., Liu, W., Zhu, L., and Li, R. (2023). Feature screening for interval-valued response with application to study association between posted salary and required skills. Journal of the American Statistical Association **118(542)**, 805–817.

Zhou, Q., Zhou, H., and Cai, J. (2017). case-cohort studies with interval-censored failure time data. *Biometrika* **104**, 17–29.

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

Zhu, X., Pan, R., Wu, S., and Wang, H. (2022). Feature screening for massive data analysis by subsampling. *Journal of Business & Economic Statistics* **40(4)**, 1892–1903.

REFERENCES

Zhu, L., Xu, K., Li, R., and Zhong, W. (2017). Projection correlation between two random vectors. *Biometrika* **104**, 829—843.

Zuo, L., Zhang, H., Wang, H., and Liu, L. (2021). Sampling-based estimation for massive survival data with additive hazards model. *Statistics in Medicine* **40(2)**, 441–450.

Yunnan Key Laboratory of Statistical Modeling and Data Analysis, Yunnan University, 2 Cuihu North Road, Kunming City, Yunnan Province, Kunming, 650091, China

E-mail: lihuiqiong@ynu.edu.cn

Yunnan Key Laboratory of Statistical Modeling and Data Analysis, Yunnan University, 2 Cuihu North Road, Kunming City, Yunnan Province, Kunming, 650091, China

E-mail: caozmyx@163.com

Department of Statistics, University of Missouri, 134J Middlebush Hall, Columbia, MO, 65211, U.S.A.

E-mail: sunj@missouri.edu

Yunnan Key Laboratory of Statistical Modeling and Data Analysis, Yunnan University, 2 Cuihu North Road, Kunming City, Yunnan Province, Kunming, 650091, China

E-mail: nstang@ynu.edu.cn