https://doi.org/10.1016/j.csda.2020.106942

Joint Model-free Feature Screening for Ultra-high Dimensional Semi-competing Risks Data

Shuiyun Lu[a], Xiaolin Chen[a,1], Sheng Xu[b], Chunling Liu[b]

[a] *School of Statistics, Qufu Normal University, Qufu, 273165, China*
[b] *Department of Applied Mathematics, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong*

Abstract

High-dimensional semi-competing risks data consisting of two probably correlated events, namely terminal event and non-terminal event, arise commonly in many biomedical studies. However, the corresponding statistical analysis is rarely investigated. A joint model-free feature screening procedure for both terminal and non-terminal events is proposed, which could allow the associated covariates to be in an ultra-high dimensional feature space. The joint screening utility is constructed from distance correlation between each predictor's survival function and joint survival function of terminal and non-terminal events. Under rather mild technical assumptions, it is demonstrated that the proposed joint feature screening procedure enjoys sure screening and consistency in ranking properties. An adaptive threshold rule is further proposed to simultaneously identify important covariates and determine number of these covariates. Extensive numerical studies are conducted to examine the finite-sample performance of the proposed methods. Lastly, the suggested joint feature screening procedure is illustrated through a real example.

Keywords: Clayton copula; Distance correlation; Feature screening; Semi-competing risks data; Ultra-high dimensionality.

1. Introduction

In various biomedical fields, researchers frequently collect semi-competing risks data (Fine et al., 2001), which are significantly different from the traditional survival data with only one type of failure and typical competing risks data including several mutually exclusive failures. Under the semi-competing risks context, multiple potential event times, one terminal event and some non-terminal events, could be observed. The terminal event censors the non-terminal ones, but non-terminal events could not hinder the occurrence and hence observation of the terminal event. For instance, in a clinical trial, one patient may drop out the study before the end of follow-up and time to the interested failure. Then interested failure and drop out could be regarded as the terminal and non-terminal events respectively in this scenario.

[1] Corresponding author. E-mail address: xlchen@amss.ac.cn

There is a broad literature on statistical methods for semi-competing risks failure time data. See Fine et al. (2001), Peng and Fine (2007), Lakhal et al. (2008), Lin et al. (2014), Li and Peng (2015) and the references therein. As for all we know, substantially all of the existing work has focused on situations without or only with low-dimensional covariates, and little literature has paid attention to the statistical analysis of high-dimensional semi-competing risks data so far. Although there already exist plenty of statistical procedures for complete data, missing data, traditional survival data and competing risks data with high-dimensional predictors, such as Zhang and Lu (2007), Zhao and Li (2012), Fu et al. (2017), Lai et al. (2017), Hong et al. (2018), Chen et al. (2018), Yan et al. (2018), Chen et al. (2019c) to cite a few, they could not be naively applied to semi-competing risks data. New approaches tailored for high-dimensional semi-competing risks data should be developed.

In high-dimensional data analysis, it is commonly assumed that only a few of all covariates are truly predictive of the response, which is called sparsity assumption in the literature (Zhu et al., 2011). Under this assumption, regularization-based variable selection methods have been well developed for varieties of types of data with moderate-high dimensional covariates. However, for the ultra-high dimensional feature space, they will come across the challenges of computational expediency, statistical accuracy and algorithmic stability simultaneously (Fan et al., 2009). As a feasible alternative before the more sophisticated penalization-based approaches could be used, marginal independence screening procedures pioneered by Fan and Lv (2008) for complete data under linear regression model have been substantially explored in recent years. The purpose of this article is to put forward a new feature screening method for ultra-high semi-competing risks data.

As mentioned above, there are two types of event times in semi-competing risks data, in which the terminal event censors the non-terminal events, but not vice versa. Without loss of generality, we assume that there exists only one non-terminal event subsequently. Methods of feature screening could be developed for terminal event and non-terminal event separately. However, we believe that the joint feature screening for both events is necessary in practice for two reasons. Firstly, almost all the existing survival feature screening methods at least require the assumption of independent censoring. Nevertheless, while conducting feature screening for non-terminal event, the terminal event and censoring times in semi-competing risks constitute the hypothetical censoring time for non-terminal event. It is obviously that non-terminal event and hypothetical censoring are dependent, thus the existing survival feature screening procedures are not applicable. Secondly, in spite of the feasibility of performing feature screening for terminal event only, the correlation information between terminal and non-terminal events is ignored completely during the individual feature screening. Therefore, executing feature screening solely for terminal event is not the most ideal method. In fact, in some cases, the marginal feature screening approach could completely fail, however our joint feature screening procedure behaves quite satisfactory; see Section 3 for more details.

In this paper, we come up with a new joint feature screening method for ultra-high dimensional semi-competing risks data. To the best of our knowledge, this problem has only been considered by Peng (2019) in the literature, where Pearson correlations between the covariates and the joint survival distribution of both terminal and non-terminal events are used to construct marginal screening utility. Here, we carry out the joint feature screening via distance correlation (Székely et al., 2007) of survival function of each predictor and the joint survival function of both terminal and non-terminal events. Compared with Pearson correlation, the distance correlation has a remarkable property that the distance correlation of two random vectors equals zero if and only if these two random vectors are independent. This makes our screening method significantly better than that of Peng (2019). In addition, by transforming each predictor through its survival function, we not only avoid the subexponential tail probability assumption for covariates, but also establish a better convergence rate than that in Peng (2019). Moreover, robustness is obtained numerically when some features contain outliers or follow heavy-tailed distributions. Besides these, we also develop an adaptive threshold rule for our suggested joint feature screening method to pick out the important covariates and determine the number of important covariates simultaneously. Last but not the least, we want to emphasis that our suggested approach is model-free, and thus do not need to specify a specific regression structure.

The rest of the article is organized as follows. Section 2 describes our methodology of joint model-free feature screening procedure and presents the corresponding theoretical properties. In Section 3, we provide an adaptive feature screening algorithm to automatically determine the threshold of number of active features, and present extensive simulation studies to evaluate the finite-sample performance of suggested methods. A real data example is illustrated in Section 4, while a brief summary and discussion is given in Section 5. All the technical proofs are relegated to the Appendix.

2. Methodology

2.1. Joint Model-free Feature Screening

Let's begin this section with some notations. Denote the times to non-terminal and terminal events by T_1 and T_2, respectively. Let $\mathbf{x} = (X_1, \cdots, X_p)^T$ be a p-dimensional vector of covariates. Both of T_1 and T_2 are subject to right censoring, the time to which is written as C. It is assumed, throughout this paper, that C is independent of T_1, T_2 and $\mathbf{x}$. Define $Y = \min\{T_1, T_2, C\}$, $\delta_1 = I(T_1 \leq T_2 \wedge C)$, $Z = \min\{T_2, C\}$ and $\delta_2 = I(T_2 \leq C)$, where $I(\cdot)$ is the indicator function and $\wedge$ is the minimum operator. Assume that one observes n independent and identically distributed copies of $\{Y, \delta_1, Z, \delta_2, \mathbf{x}\}$, expressed as $\{Y_i, \delta_{1i}, Z_i, \delta_{2i}, \mathbf{x}_i\}_{i=1}^n$. In the ultra-high dimensional circumstances considered here, p is on a large or huge scale, and much larger than n, which means that p could increase at an exponential rate of n technically.

To pick up the small number of predictors, which have influences on only one of T_1 and T_2 or both, we first define the active and inactive predictors without specifying the correlation structure of non-terminal and terminal events and specific regression model. Denote by $S(t_1, t_2|\mathbf{x}) = \Pr(T_1 > t_1, T_2 > t_2|\mathbf{x})$ the joint survival function of T_1 and T_2 conditional on $\mathbf{x}$. Then we define the index set of jointly active predictors as

$$\mathcal{A}=\{k : S(t_1, t_2|\mathbf{x}) \text{ functionally depends on } X_k \text{ for some } k = 1, \cdots, p\}.$$

If $k \in \mathcal{A}$, X_k is regarded as an active feature; otherwise, an inactive one.

Before presenting our joint screening utility, we will review the distance correlation and its estimation (Székely et al., 2007) briefly. Assume that U and V are two random vectors, the dimensionality of which are d_U and d_V respectively. The distance covariance is defined as the nonnegative number of $\mathrm{dcov}(U,V)$ given by $\mathrm{dcov}^2(U,V) = S_1 + S_2 - 2S_3$, where $S_1 = E(\|U - \tilde{U}\|_{d_U}\|V - \tilde{V}\|_{d_V})$, $S_2 = E(\|U - \tilde{U}\|_{d_U})E(\|V - \tilde{V}\|_{d_V})$, $S_3 = E\{E(\|U - \tilde{U}\|_{d_U}|U)E(\|V - \tilde{V}\|_{d_V}|V)\}$, $(\tilde{U}, \tilde{V})$ is an independent copy of (U,V) and $\|a\|_d$ represents the Euclidean norm for a d-dimensional vector a. Then distance correlation between U and V is defined as

$$\mathrm{dcorr}(U,V) = \mathrm{dcov}(U,V)/\sqrt{\mathrm{dcov}(U,U)\mathrm{dcov}(V,V)}. \tag{1}$$

Given an independent and identically distributed sample $\{U_i, V_i\}_{i=1}^n$, we could estimate $\mathrm{dcov}(U,V)$ by

$$\widehat{\mathrm{dcov}}(U,V) = \hat{S}_1 + \hat{S}_2 - 2\hat{S}_3 \tag{2}$$

with

$$\hat{S}_1 = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\|U_i - U_j\|_{d_U}\|V_i - V_j\|_{d_V}, \tag{3}$$

$$\hat{S}_2 = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\|U_i - U_j\|_{d_U}\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\|V_i - V_j\|_{d_V}, \tag{4}$$

and

$$\hat{S}_3 = \frac{1}{n^3}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\|U_i - U_l\|_{d_U}\|V_j - V_l\|_{d_V}. \tag{5}$$

$\widehat{\mathrm{dcov}}(U,U)$ and $\widehat{\mathrm{dcov}}(V,V)$ could be obtained in the similar way. Thus, from a sample, $\mathrm{dcorr}(U,V)$ could be estimated by $\widehat{\mathrm{dcorr}}(U,V) = \widehat{\mathrm{dcov}}(U,V)/\sqrt{\widehat{\mathrm{dcov}}(U,U)\widehat{\mathrm{dcov}}(V,V)}$. As a measure of correlation, the distance correlation has an attractive property, i.e. $\mathrm{dcorr}(U,V) = 0$ if and only if U and V are independent. This makes distance correlation particularly suitable for variable screening of ultra-high dimensional data. Li et al. (2012), Zhong and Zhu (2014), Zhong et al. (2016), Chen et al. (2019a), and Chen et al. (2018) have investigated distance correlation-based screening in the complete data and traditional survival data setting.

Just like the scenario of distance correlation-based screening for traditional survival data (Chen et al., 2018), the original distance correlation (1) could not be directly applied to our current circumstances to measure the correlation of each predictor and $(T_1, T_2)^T$. In the same sprite of Chen et al. (2018), we propose a modified distance correlation for the semi-competing risks context as

$$\rho_k = \frac{\mathrm{dcov}\{S_k(X_k), S(T_1,T_2)\}}{\sqrt{\mathrm{dcov}\{S_k(X_k), S_k(X_k)\}}\sqrt{\mathrm{dcov}\{S(T_1,T_2), S(T_1,T_2)\}}}, \tag{6}$$

where $S_k(x_k) = \Pr(X_k > x_k)$ is the survival function of X_k and $S(t_1,t_2) = \Pr(T_1 > t_1, T_2 > t_2)$ is the joint survival function of $(T_1,T_2)^T$. It is noted that the distance correlation could evaluate the correlation between two random vectors. Thus the correlation between X_k

and $(T_1, T_2)^T$ could be measured directly by the distance correlation $\text{dcorr}\{X_k, (T_1, T_2)^T\}$. However, due to the right censoring, it is not easy to estimate $\text{dcorr}\{X_k, (T_1, T_2)^T\}$ based on an independent and identically distributed sample. But, as explained below, our suggested modified distance correlation could be estimated without difficulties. Based on the modified distance correlation, the joint screening utility is defined by $\omega_k = \rho_k^2$ for $k = 1, \cdots, p$.

To conduct feature screening, we need to derive the sample version of ω_k based on $\{Y_i, \delta_{1i}, Z_i, \delta_{2i}, \mathbf{x}_i\}_{i=1}^n$. As for $S_k(x)$, it could be easily estimated by the empirical survival function, i.e. $\hat{S}_k(x_k) = n^{-1}\sum_{i=1}^n I(X_{ki} > x_k)$. A number of estimators of the joint survival function have been suggested in the literature; see the references in Lin and Ying (1993). Lin and Ying's estimator is the simplest, and enjoys many desirable properties, such as the weak convergence. It should be emphasized that the joint survival function could only be estimated on the upper wedge of entire plane without other additional information. For $0 < t_1 < t_2$, Lin and Ying's estimator takes the following form

$$\hat{S}(t_1, t_2) = \frac{n^{-1}\sum_{i=1}^n I(Y_i \geq t_1, Z_i \geq t_2)}{\hat{G}(t_2)}, \tag{7}$$

where $\hat{G}(\cdot)$ is the Kaplan-Meier estimator of survival function of C based on $\{Z_i, 1 - \delta_{2i}\}_{i=1}^n$, and can be expressed as

$$\hat{G}(t_2) = \prod_{i=1}^n \left(1 - \frac{1}{\sum_{j=1}^n I(Z_j \geq Z_i)}\right)^{(1-\delta_{2i})I(Z_i \leq t_2)}. \tag{8}$$

The feasibility of (7) and (8) is due to the assumption that T_1 and T_2 are rightly censored by C independently. From Equations (2), (7) and (8), we could estimate $\text{dcov}\{S_k(X_k), S(T_1, T_2)\}$ by

$$\widehat{\text{dcov}}\{S_k(X_k), S(T_1, T_2)\} = \hat{S}_{1k} + \hat{S}_{2k} - 2\hat{S}_{3k}, \tag{9}$$

where $\hat{S}_{1k}$, $\hat{S}_{2k}$ and $\hat{S}_{3k}$ are defined by replacing U_i, U_j, V_i, V_j and V_l with $\hat{S}_k(X_{ki})$, $\hat{S}_k(X_{kj})$, $\hat{S}(Y_i, Z_i)$, $\hat{S}(Y_j, Z_j)$ and $\hat{S}(Y_l, Z_l)$. In the similar way, we could get $\widehat{\text{dcov}}\{S_k(X_k), S_k(X_k)\}$ and $\widehat{\text{dcov}}\{S(T_1, T_2), S(T_1, T_2)\}$. Then the sample version of ω_k could be achieved as $\hat{\omega}_k = \hat{\rho}_k^2$ for $k = 1, \cdots, p$, where $\hat{\rho}_k$ is the estimated modified distance correlation of (6).

From the values of ω_ks, we identify the indexes of important covariates by selecting out the covariates with large ω_ks. Specifically, the index set of active predictors are estimated by

$$\hat{\mathcal{A}} = \{k : \hat{\omega}_k \geq \gamma_n,\ k = 1, \cdots, p\},$$

where γ_n is a threshold sequence given in advance and varies with sample size n. For practical use, we can find a pre-determined size d_0 (may change with sample size) and pick out the covariates with corresponding ω_ks among the first d_0 largest of all. In the sequel, the joint model-free distance correlation-based sure independence screening procedure is referred to as JMDC-SIS for short.

2.2. Theoretical Properties

Ahead of the presentation of theoretical properties, let's introduce the necessary technical assumptions as follows:

(A1): There exists a positive constant η such that $\Pr(t_2 \leq T_2 \leq C) \geq \eta$, where $t_2 \in (0, \tau]$ with τ being the maximum follow-up time; Furthermore, $\sup\{t_2 : \Pr(T_2 > t_2) > 0\} \geq \sup\{t_2 : \Pr(C > t_2) > 0\}$; $G(\cdot)$ has uniformly bounded first order derivative.

(A2): It holds that $\min_{k \in \mathcal{A}} \omega_k \geq 2c_1 n^{-\kappa}$ for some constants $c_1 > 0$ and $\kappa \in [0, 1/2)$.

Assumption (A1) is common in survival analysis literature to guarantee the well performance of Kaplan-Meier estimator, and has been imposed in much survival feature screening literature, such as He et al. (2013), Zhou and Zhu (2017), Chen et al. (2019b) and so on. Assumption (A2) is standard in feature screening investigation to make sure that the minimal signal does not degenerate too fast, and then guarantee the sure screening property. It should be noted that, in Assumption (A1), we only make assumptions on survival time of terminal event and censoring time, but not on non-terminal event. This is different from assumptions in Peng (2019), in which assumptions are made on joint distribution of terminal and non-terminal events besides that on censoring time. Due to the fact that terminal event will censor the non-terminal event, but not vice versa, we believe that our assumption is more flexible than that in Peng (2019). Variances of all the covariates are required to be uniformly bounded in Peng (2019), while our suggested procedure avoids this restrictive assumption through transformation.

The sure screening property is stated in the following Theorem 1.

Theorem 1. *(Sure Screening Property) Under Assumptions (A1) and (A2), there exist positive constants d_1 and d_2 such that*

$$\Pr\left(\max_{1 \leq k \leq p} |\hat{\omega}_k - \omega_k| \geq c_1 n^{-\kappa}\right) \leq d_1 p \exp\{-d_2 n^{1-2\kappa}\}.$$

Furthermore, if taking $\gamma_n = c_1 n^{-\kappa}$, we have

$$\Pr(\mathcal{A} \subseteq \hat{\mathcal{A}}) \geq 1 - d_1 q \exp\{-d_2 n^{1-2\kappa}\},$$

where q is the number of truly important covariates.

From the result in Theorem 1, it can be seen that our JMDC-SIS could manage covariates with dimensionality $\log(p) = o(n^{1-2\kappa})$, which is better than that of Peng (2019). The subsequent corollary provides a result about the size of selected predictors by JMDC-SIS with $\gamma_n = c_1 n^{-\kappa}$.

Corollary 1. *(False Discovery Control) Under Assumptions (A1) and (A2), we have that, for $\gamma_n = c_1 n^{-\kappa}$, there exist positive constants d_3 and d_4 such that*

$$\Pr\left(|\hat{\mathcal{A}}| \leq 2c_1 n^{\kappa} \sum_{1 \leq k \leq p} |\omega_k|\right) \geq 1 - d_3 p \exp\{-d_4 n^{1-2\kappa}\},$$

where $|\hat{\mathcal{A}}|$ represents the number of elements in $\hat{\mathcal{A}}$.

Let $\mathbf{x}_{\mathcal{A}}$ denote the subvector of $\mathbf{x}$ consisting of all X_k with $k \in \mathcal{A}$. Define $\mathbf{x}_{\mathcal{A}^c}$ in the same way. Theorem 2 below affords JMDC-SIS's ranking consistency under some additional assumptions.

Theorem 2. *Assuming that (i) $(T_1, T_2)^T$ and $\mathbf{x}_{\mathcal{A}^c}$ are conditionally independent given $\mathbf{x}_{\mathcal{A}}$; (ii) $\mathbf{x}_{\mathcal{A}}$ is independent of $\mathbf{x}_{\mathcal{A}^c}$. Under Assumption (A2), we have that*

$$\max_{k\in\mathcal{A}^c} \omega_k < \min_{k\in\mathcal{A}} \omega_k,$$

and $\omega_k = 0$ if and only if $k \in \mathcal{A}^c$. Furthermore, it holds that

$$\Pr\left(\min_{k\in\mathcal{A}} \hat{\omega}_k > \max_{k\in\mathcal{A}^c} \hat{\omega}_k\right) \geq 1 - 2d_1 p \exp\{-d_2 n^{1-2\kappa}\}.$$

3. Numerical Studies

3.1. JMDC-SIS with adaptive threshold rule

To perform our JMDC-SIS, one needs to specify a threshold sequence in advance, which is hard in practice. Instead, we could assign an integer sequence d_0 and keep the covariates with the estimated screening utilities being among the first d_0 largest ones. The most frequently used reference sequences are $d_0 = [n/\log(n)]$ or $n-1$ (Fan and Lv, 2008), where $[a]$ denotes the integer part of a. The choice of threshold sequence will substantially influence the performance of feature screening on one hand, and on the other hand, is hard to interpret in both of theory and practice. Several methods have been advocated to address this problem, such as those based on combination of soft and hard thresholding rules (Zhu et al., 2011), false positive rate control (Zhao and Li, 2012) and p-values of multiple testing (Wen et al., 2017).

In this subsection, we propose a data-driven means for JMDC-SIS to conduct feature screening without pre-specification of threshold value based on an inequality of distance covariance (Kong et al., 2015). Suppose that U and V are two random vectors with the same dimensionality, i.e. $d_U = d_V$, and W is an additional d_W-dimensional random vector. It has been proved, in Kong et al. (2015), that $\mathrm{dcov}(U+V, W) \leq \mathrm{dcov}(U, W)$ on the condition of V being independent of (U, W). Besides that, Theorem 4 of Székely et al. (2007) declares that $\mathrm{dcov}(U+V, U+V) \leq \mathrm{dcov}(U, U) + \mathrm{dcov}(V, V)$. Therefore, we can conclude that

$$\mathrm{dcorr}(U+V, W) \leq \mathrm{dcorr}(U, W) \tag{10}$$

according to the definition of distance correlation (1). The inequality (10) motivates us to identify features in a sequential and model-free fashion just as the well-known forward variable selection, and terminate the process once the decrease of distance correlation occurs. A similar procedure based on distance covariance has been applied to feature screening for complete data (Kong et al., 2015). Here, we tailor our JMDC-SIS based on (10) to determine the number of retained covariates in an adaptive way, and name the corresponding procedure as adaptive JMDC-SIS (aJMDC-SIS for short), the steps of which are summarized below:

Step 1: Obtain $\hat{S}(t_1, t_2)$, and $\hat{S}_k(x_k)$ for $k = 1, \cdots, p$. Then compute $\hat{\omega}_k$ for each X_k, $k = 1, \cdots, p$.

Step 2: Sort $\hat{\omega}_k$, $k = 1, \cdots, p$, in decreasing order. Reorder the covariates according to the sorted $\hat{\omega}_k$s, and denoted them by $X_{(1)}, X_{(2)}, \cdots, X_{(p)}$. Initialise by setting $m = 1$ and $\tilde{\mathcal{A}}^{(1)} = \{k : X_k = X_{(1)}\}$.

Step 3: Update $\tilde{\mathcal{A}}^{(m+1)} = \tilde{\mathcal{A}}^{(m)} \bigcup \{k : X_k = X_{(m+1)}\}$. If $\widehat{\mathrm{dcorr}}\left(\sum_{k \in \tilde{\mathcal{A}}^{(m+1)}} \hat{S}_k(X_k), \hat{S}(T_1, T_2)\right) \leq \widehat{\mathrm{dcorr}}\left(\sum_{k \in \tilde{\mathcal{A}}^{(m)}} \hat{S}_k(X_k), \hat{S}(T_1, T_2)\right)$, stop and set $\hat{\mathcal{A}} = \tilde{\mathcal{A}}^{(m)}$; otherwise, set $m = m + 1$ and continue this process.

3.2. Simulation Settings and Results

In this subsection, we assess via simulation studies the finite sample performance of the proposed JMDC-SIS and aJMDC-SIS, also compare them with joint correlation rank screening (JCR) of Peng (2019) and robust censored distance correlation screening (RCDCS) of Chen et al. (2018) for terminal event only. To make easy comparisons between joint feature screening for both terminal and non-terminal events and marginal feature screening RCDCS for terminal event only, the truly predictive covariates are designed to be the same for the two events in Examples 1 to 5.

In all the following examples, we generate the censoring times from the uniform distribution on interval $(0, 6)$, which causes censoring rates from 30% to 78% for terminal events. The dimensionality p of covariates is set to be 2000. We specify the threshold value for JCR, RCDCS and JMDC-SIS to be $d_0 = [n/\log(n)]$. Based on 500 simulation runs, we evaluate the performance of these screening procedures using the following criteria: selection proportions for each truly predictive covariate for $(T_1, T_2)^T$, selection proportions for all active covariates, and quantiles of the minimum model size to include all active covariates.

Example 1. In this example, we consider the log-linear model for times of non-terminal and terminal events. Specifically, non-terminal and terminal event times are generated from the following models:

$$\log(T_1) = \mathbf{x}^T \beta - 0.5 e_1$$

and

$$\log(T_2) = \mathbf{x}^T \alpha + e_2,$$

where $\beta = (1.0, 0.5, 1.0, 0, \cdots, 0)^T$, $\alpha = (0.2, -0.45, 0.25, 0, \cdots, 0)^T$, $\mathbf{x} = (X_1, X_2, \cdots, X_p)^T$ follows the multivariate normal distribution $N(\mathbf{0}_{p\times 1}, \Sigma = (\rho^{|i-j|})_{p\times p})$ with $\rho = 0.6$ and 0.9, and the joint distribution of error vector $(e_1, e_2)^T$ satisfies the Clayton copula (Nelsen, 2007), that is, $\Pr(e_1 > t_1, e_2 > t_2) = [S^*(t_1)^{-\theta} + S^*(t_2)^{-\theta} - 1]^{-1/\theta}$ with $\Pr(e_1 > t) = \Pr(e_2 > t) = S^*(t) = \exp\{-\exp(t)\}$. Here we choose the parameter $\theta = 0.5$, 2 and 8, which reflect different associations between the two events through Kendall's τ (Nelsen, 2007). The censoring rates for T_1 and T_2 are approximately 32% and 30%, respectively.

Insert Figure 1 about here

We firstly make use of this relative simple example to examine the feasibility of adaptive threshold rule for JMDC-SIS, i.e. the aJMDC-SIS. Under several different sample sizes, the threshold values selected by aJMDC-SIS are recorded for chosen settings. We repeat this operation 100 times for each combination of different sample sizes and settings. Figure 1 exhibits the line charts of adaptive threshold value versus sample size for aJMDC-SIS under different settings. It is easy to see that the adaptive threshold value gets closer and closer to the number of truly important features as the sample size increases. When the sample size is equal to 400, the threshold values identified by aJMDC-SIS are quite close to the true number 3.

Insert Tables 1 and 2 about here

The results for the three criteria based on 500 data repetitions are listed in Tables 1 and 2, from which we could easily see that our suggested aJMDC-SIS and JMDC-SIS outperform the JCR under each setting. The correlations among covariates play greater impacts on the performance of JCR than those of aJMDC-SIS and JMDC-SIS, that is, JCR is more sensitive to the change of correlations among covariates. It is so amazing to observe that RCDCS, which performs feature screening only for terminal event, fails completely. This observation makes the use of joint feature screening for semi-competing risks data more necessary. Besides, these results also show that different choices of levels of association between the two events have little influences on the behavior of the proposed screening methods. As we anticipate, the performance of various methods become better as the sample size increases.

Example 2. We consider in this example complex varying-coefficient nonlinear models for both of non-terminal and terminal events' times. To be concrete, non-terminal and terminal event times are generated from the following models:

$$\log(T_1) = \beta_1(U)\sin(X_1) + \beta_2(U)X_2^2 + \beta_3(U)X_3 + e_1$$

and

$$\log(T_2) = \alpha_1(U)X_1^2 + \alpha_2(U)X_2X_3 + \alpha_3(U)|X_3| + e_2,$$

where $\beta_1(U) = 1+U$, $\beta_2(U) = 2\cos(2\pi U)$, $\beta_3(U) = 2U$, $\alpha_1(U) = 1+U$, $\alpha_2(U) = 2\sin(2\pi U)$, $\alpha_3(U) = U^2$, and U is a standard uniform random variable. The other elements are the same as those in Example 1. The censoring rates for T_1 and T_2 are approximately 42% and 66%, respectively.

Insert Tables 3 and 4 about here

Tables 3 and 4 present the simulation results for Example 2, and the similar phenomena as those in Example 1 could be observed. However, in this example, the advantages of JMDC-SIS and aJMDC-SIS over JCR and RCDCS are more significant than those in Example 1. The threshold value determined by aJMDC-SIS decreases as the sample size increases, and is far smaller than that specified by convention under the cases with $n = 300$ and $\rho = 0.9$.

Example 3. This example presents general nonlinear models for non-terminal and terminal events' times. T_1 and T_2 satisfy the beneath regression models:

$$\log(T_1) = 0.5(X_1 + 0.5X_2 + 0.5|X_3|)^2 + \sin(X^T\beta) + e_1$$

and

$$\log(T_2) = 0.5(X^T\beta)^2 + \cos(X_1X_2 + 0.5\sin(X_2) + 0.5|X_3|)^2 + e_2,$$

where $\beta = (1, 0.5, 0.5, 0, \cdots, 0)^T$ and the other elements are the same as those in Example 1. The censoring rates for T_1 and T_2 are also about 42% and 66%, respectively.

Insert Tables 5 and 6 about here

Tables 5 and 6 summarize the simulation results with similar observations as those in Examples 1 and 2.

Example 4. We consider other general nonlinear regression models for non-terminal and terminal events' times here. The regression models for T_1 and T_2 are

$$\log(T_1) = 0.5(X_1X_2 + 2X_2 + X_3)^3 + 0.5\exp(|X_1|^2 + 2\tan(X_2) + X_3^2) + e_1$$

and

$$\log(T_2) = 0.5(X^T\alpha)^2 + 0.5\tan(X^T\alpha) + e_2,$$

where $\alpha = (2, 1, 1, 0, \cdots, 0)^T$ and the other settings are the same as those in Example 1. The censoring rates for T_1 and T_2 are around 63% and 78%, respectively. Compared with former examples, the censoring rates are rather high in this example.

Insert Tables 7 and 8 about here

The simulation results are presented in Tables 7 and 8. In addition to the phenomena similar to those in former examples, we could see that JCR behaves badly when the correlations among covariates are low, even when the sample size is relatively large.

Example 5. In this example, we change the autoregressive type correlation structure to the simple independent and identically distributed case. The regression models for T_1 and T_2 are

$$\log(T_1) = 0.15X_1X_2 + 0.15X_2 + 0.15\sin(X_3) + 0.15e_1$$

and

$$\log(T_2) = 0.15X_1 + 0.15X_2^2 + 0.35|X_3| + 0.15e_2,$$

where $\mathbf{x} = (X_1, X_2, \cdots, X_p)^T$ follows the multivariate normal distribution $N(\mathbf{0}_{p\times 1}, I_{p\times p})$ with $I_{p\times p}$ being an identity matrix of size p. The other settings are the same as those in former examples. The censoring rates for T_1 and T_2 are approximately 20% and 39%, respectively.

Insert Tables 9 and 10 about here

Simulation results are exhibited in Tables 9 and 10. Different from previous examples, JCR fails completely in this example. The improvement of JCR is very limited as the sample size increases from 200 to 300, even the RCDCS obtains satisfactory results. Although RCDCS, which performs feature screening only for terminal event, achieves acceptable performance, the gaps between RCDCS and JMDC-SIS or aJMDC-SIS are still fairly wide.

Example 6. As suggested by one reviewer, it is interested to examine the situation with different sets of the truly important convariates for non-terminal and terminal events. Thus in this example, we consider the following regression models for T_1 and T_2:

$$\log(T_1) = 0.5(X_1 + 0.5X_2 + 0.5|X_3|)^2 + \sin(X^T\beta) + e_1$$

and

$$\log(T_2) = 0.5(X^T\alpha)^2 + \cos(X_2X_3 + 0.5\sin(X_3) + 0.5|X_4|)^2 + e_2,$$

where $\beta = (1, 0.5, 0.5, 0, \cdots, 0)^T$, $\alpha = (0, 1, 0.5, 0.5, 0, \cdots, 0)^T$) and the other elements are the same as those in Example 1. The censoring rates for T_1 and T_2 are approximately 57% and 83%, respectively. It is easy to see that the predictors X_2 and X_3 are commonly important covariates for T_1 and T_2. However, X_1 is only truly predictive for T_1, while X_4 is only important for T_2.

Insert Tables 11 and 12 about here

The simulation results are displayed in Tables 11 and 12, from which similar phenomena as that in the former examples are observed. In addition, it is worth noting that X_1 and X_4, which are truly important for T_1 and T_2 respectively, have significantly different inclusion frequencies. The inclusion frequency of X_4 is dramatically lower than that of X_1. There may be two reasons for this. Firstly, the signal for X_1 is stronger than that of X_4. This could be seen by comparing the coefficients of X_1 and X_4 in the regression models of T_1 and T_2. Secondly, the censoring rate for T_2 is very high, which leads that the data contain less information of T_2 than that of T_1. Furthermore, the high censoring rate results in that the information of X_4 is less than that of X_1.

4. Real Data Analysis

As an illustration, in this section, we apply JMDC-SIS and aJMDC-SIS, along with JCR and RCDCS, to a breast cancer dataset (van De Vijver et al., 2002). In this dataset, there are 295 patients from the Netherlands Cancer Institute. In addition to the main interested event, death from breast cancer, data about time to distant metastasis were also collected. Thus, in this application, distant metastasis is the non-terminal event, while death from breast cancer is the terminal event. Among all 295 patients, 101 persons experienced distant metastasis and 79 experienced death, corresponding to around 65% censoring rate for non-terminal event and 75% censoring rate for terminal event. Besides the clinical data, this dataset contains data for 24,881 genes for each patient. Our goal is to identify the genes which are predictive for distant metastasis or death from breast cancer. The data for our analysis could be obtained from the R package ”cancerdata” at https://www.bioconductor.org/packages/release/data/experiment/html/cancerdata.html.

To conduct JMDC-SIS, JCR and RCDCS, we specify the threshold value as $[295/\log 295] = 51$. As described in Section 3.1, aJMDC-SIS will determine the model size in an adaptive way, and does not need to appoint one. It should be noted that aJMDC-SIS, JMDC-SIS and JCR aim to identify the genes which are predictive for distant metastasis or death from breast cancer, while RCDCS is used to screen predictive genes for death from breast cancer only.

The names of selected genes by various procedures are listed in Table 13. It is easy to see that aJMDC-SIS confirms that 25 genes are predictive for distant metastasis or death from breast cancer, which is significantly smaller than the threshold 51 used for the other three methods. This shows that our aJMDC-SIS is effective to determine the predictive genes and number of them simultaneously for this dataset. Among the 25 genes, Contig48328_RC and Contig38288_RC have been confirmed to be related to at least death from breast cancer (van't Veer et al., 2002). However, JCR could not select out Contig48328_RC even with the threshold value 51.

There are 21 genes picked out by both of JMDC-SIS and JCR. It is more likely that these 21 genes are truly important for either distant metastasis or death from breast cancer. For the 51 genes chosen by RCDCS, 34 are also selected by JMDC-SIS. With the same threshold value considered, this result is reasonable, and shows that JMDC-SIS is more flexible than RCDCS by selecting out the genes associated with distant metastasis too.

5. Summary and Discussion

In this article, we propose a joint model-free feature screening procedure for ultra-high dimensional semi-competing risks data via distance correlation, and name it JMDC-SIS. The joint approach could pick out the covariates associated with either non-terminal or terminal events, both of which are important in semi-competing risks data analysis. Theoretical properties of JMDC-SIS are established under rather mild assumptions. To determine the number of important features, an adaptive threshold rule is suggested for JMDC-SIS. The JMDC-SIS with the adaptive threshold rule is called aJMDC-SIS. Simulation studies have shown the usefulness of JMDC-SIS and aJMDC-SIS, and the advantages over the existing JCR. In addition, we find surprisingly that, compared with joint screening methods, the marginal screening for just terminal event do not only lose efficiency (see Example 5 in Section 3), but also could fail completely in some cases (see Examples 1 to 4 in Section 3).

To the best of our knowledge, the literature about ultra-high dimensional data analysis for semi-competing risks data is very limited. The feature screening is only the first step to reduce the dimension to a moderate scale. More sophisticated regularized approaches are urgently needed for further data analysis. This guarantees the future investigation for ultra-high dimensional semi-competing risks data.

Acknowledgments

This work was conducted while Xiaolin Chen was visiting Department of Statistics, Rutgers University. Xiaolin Chen wants to express his gratitude to Prof. Min-ge Xie for his hospitality during Xiaolin's visit at Rutgers University. The authors are grateful to the co-editor, anonymous associate editor and two anonymous reviewers for their helpful and insightful comments. Xiaolin Chen's research is supported by the National Natural Science Foundation of China (11771250 and 11501573) and a grant from China Scholarship Council. Sheng Xu's research is supported by The Hong Kong Polytechnic University Ph.D. Studentship. Chunling Liu's research is supported by the General Research Fund of Hong Kong (15327216).

References

Bitouzé, D., Laurent, B., Massart, P., 1999. A dvoretzky-kiefer-wolfowitz type inequality for the kaplan-meier estimator. Annales de l'Institut Henri Poincaré (B) Probability and Statistics 35, 735–763.

Chen, X., Chen, X., Liu, Y., 2019a. A note on quantile feature screening via distance correlation. Statistical Papers 60, 1741–1762.

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.

Chen, X., Zhang, Y., Chen, X., Liu, Y., 2019b. A simple model-free survival conditional feature screening. Statistics and Probability Letters 146, 156–160.

Chen, X., Zhang, Y., Liu, Y., Chen, X., 2019c. Model-free feature screening for ultra-high dimensional competing risks data. Submited to Statistics and Probability Letters .

Dvoretzky, A., Kiefer, J., Wolfowitz, J., 1956. Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. Annals of Mathematical Statistics 27, 642–669.

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

Fan, J., Samworth, R., Wu, Y., 2009. Ultrahigh dimensional feature selection: beyond the linear model. Journal of Machine Learning Research 10, 2013–2038.

Fine, J., Jiang, H., Chappell, R., 2001. On semi-competing risks data. Biometrika 88, 907–919.

Fu, Z., Parikh, C., Zhou, B., 2017. Penalized variable selection in competing risks regression. Lifetime Data Analysis 23, 353–376.

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

Hong, H., Chen, X., Christian, D., Li, Y., 2018. Integrated powered density: screening ultrahigh-dimensional covariates with survival outcomes. Biometrics 74, 421–429.

Kiefer, J., 1961. On large deviations of the empiric d.f. of vector chance variables and a law of the iterated logarithm. Pacific Journal of Mathematics 11, 649–660.

Kong, J., Wang, S., Wahaba, G., 2015. Using distance covariance for improved variable selection with application to learning genetic risk models. Statistics in Medicine 34, 1708–1720.

Lai, P., Liu, Y., Liu, Z., Wan, Y., 2017. Model free feature screening for ultrahigh dimensional data with responses missing at random. Computational Statistics and Data Analysis 105, 201–216.

Lakhal, L., Rivest, L., Abdous, B., 2008. Estimating survival and association in a semicompeting risks model. Biometrics 64, 180–188.

Li, R., Peng, L., 2015. Quantile regression adjusting for dependent censoring from semicompeting risks. Journal of the Royal Statistical Society: Series B 77, 107–130.

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

Lin, D., Ying, Z., 1993. A simple nonparamertic estimator of the bivariate survival function under univariate censoring. Biometrika 80, 573–581.

Lin, H., Zhou, L., Li, C., Li, Y., 2014. Semiparametric transformation models for semicompeting survival data. Biometrics 70, 599–607.

Liu, J., Li, R., Wu, R., 2014. Feature selection for varying coefficient models with ultrahigh-dimensional covariates. Journal of the American Statistical Association 109, 266–274.

Nelsen, R., 2007. An intruduction to copulas. Springer Science & Business Media.

Peng, L., Fine, J., 2007. Regression modeling of semicompeting risks data. Biometrics 63, 96–108.

Peng, M., 2019. Analysis of complex survival data subject to semi-competing risks. Ph.D. thesis. Nanyang Technological University, Singapore.

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

van De Vijver, M., He, Y., van't Veer, L., Dai, H., Hart, A., Voskuil, D., Schreiber, G., Peterse, J., Roberts, C., Marton, M., 2002. A gene-expression signature as a predictor of survival in breast cancer. New England Journal of Medicine 347, 1999–2009.

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

Wen, C., Zhu, S., Chen, X., Wang, X., 2017. Adaptive model-free sure independence screening. Statistics and Its Interface 10, 399–406.

Yan, X., Tang, N., Xie, J., Ding, X., Wang, Z., 2018. Fused mean-variance filter for feature screening. Computational Statistics and Data Analysis 122, 18–31.

Zhang, L., Lu, W., 2007. Adaptive lasso for cox's proportional hazards model. Biometrika 94, 691–703.

Zhao, S., 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., 2014. An iterative approach to distance correlation-based sure independence screening. Journal of Statistical Computation and Simulation 85, 1–15.

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.

Zhou, T., Zhu, L., 2017. Model-free feature screening for ultrahigh dimensional censored regression. Statistics and Computing 27, 947–961.

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

Appendix: Lemmas and Proofs of the Theorems

Lemma 1. *(Bitouzé et al., 1999) Let* $\{T_{2i}\}_{i=1}^n$ *and* $\{C_i\}_{i=1}^n$ *be independent sequences of independently and identically distributed nonnegative random variables with survival functions* S *and* G*, respectively. Let* $\hat{G}$ *be the Kaplan-Meier estimator of* G*. Then there exists a positive constant* d_5 *such that*

$$\Pr\left(n^{\frac{1}{2}} \parallel S(\hat{G} - G) \parallel_\infty > \lambda\right) \leq 2.5 \exp\{-2\lambda^2 + d_5\lambda\},$$

for any positive constant λ.

Lemma 2. *Under Assumption (A1), for any positive constant* $\varepsilon \in (0, \eta/2)$*, there exist positive constants* d_6 *and* d_7 *such that*

$$\Pr\left(\sup_{y\in[0,\tau]} \left|\frac{1}{\hat{G}(y)} - \frac{1}{G(y)}\right| \geq \varepsilon\right) \leq 5 \exp\{-d_6 n\varepsilon^2 + d_7 n^{1/2}\varepsilon\}.$$

Furthermore, if $n^{1/2}\varepsilon \to \infty$ as n goes to ∞, for sufficiently large n, we have

$$\Pr\left(\sup_{y\in[0,\tau]}\left|\frac{1}{\hat{G}(y)}-\frac{1}{G(y)}\right|\geq\varepsilon\right)\leq 5\exp\{-d_8 n\varepsilon^2\},$$

where d_8 is a positive constant.

Proof: Under the event $\{\sup_{y\in[0,\tau]}|\hat{G}(y)-G(y)|\leq\varepsilon\}$, we have $\hat{G}(y)\geq\eta/2$ for any $y\in[0,\tau]$. In addition, by Assumption (A1), it can be obtained that

$$\sup_{y\in[0,\tau]}\left|\frac{1}{\hat{G}(y)}-\frac{1}{G(y)}\right|=\sup_{y\in[0,\tau]}\left|\frac{\hat{G}(y)-G(y)}{\hat{G}(y)G(y)}\right|\leq 2\eta^{-2}\sup_{y\in[0,\tau]}|\hat{G}(y)-G(y)|.$$

Then

$$\begin{aligned}
&\Pr\left(\sup_{y\in[0,\tau]}\left|\frac{1}{\hat{G}(y)}-\frac{1}{G(y)}\right|\geq\varepsilon\right)\\
\leq\ &\Pr\left(\sup_{y\in[0,\tau]}\left|\frac{1}{\hat{G}(y)}-\frac{1}{G(y)}\right|\geq\varepsilon,\ \sup_{y\in[0,\tau]}|\hat{G}(y)-G(y)|\leq\varepsilon\right)\\
&\quad+\Pr\big(\sup_{y\in[0,\tau]}|\hat{G}(y)-G(y)|\geq\varepsilon\big)\\
\leq\ &\Pr\big(\sup_{y\in[0,\tau]}|\hat{G}(y)-G(y)|\geq\varepsilon 2^{-1}\eta^2\big)+\Pr\big(\sup_{y\in[0,\tau]}|\hat{G}(y)-G(y)|\geq\varepsilon\big)\\
\leq\ &2\Pr\big(\sup_{y\in[0,\tau]}|\hat{G}(y)-G(y)|\geq\varepsilon\min\{2^{-1}\eta^2,1\}\big)\\
\leq\ &2\Pr\big(n^{1/2}\sup_{y\in[0,\tau]}|S(y)(\hat{G}(y)-G(y))|\geq n^{1/2}\varepsilon\eta\min\{2^{-1}\eta^2,1\}\big)\\
\leq\ &5\exp\{-2\eta^2\min\{2^{-2}\eta^4,1\}n\varepsilon^2+d_5\eta\min\{2^{-1}\eta^2,1\}n^{1/2}\varepsilon\}\\
\triangleq\ &5\exp\{-d_6 n\varepsilon^2+d_7 n^{1/2}\varepsilon\},
\end{aligned}\tag{A.1}$$

where the fourth inequality is arrived by Assumption (A1) and the last inequality is obtained based on Lemma 1.

Moreover, from the assumption that $n^{1/2}\varepsilon\to\infty$, we could conclude that $d_6-d_7/(n^{1/2}\varepsilon)>d_6/2$ for sufficiently large n. Thus

$$-d_6 n\varepsilon^2+d_7 n^{1/2}\varepsilon=-n\varepsilon^2\{d_6-d_7/(n^{1/2}\varepsilon)\}<-n\varepsilon^2 d_6/2\triangleq -d_8 n\varepsilon^2.$$

The second part is achieved by combing this result with Equation (A.1).

□

Lemma 3. *Suppose that (U,V) is a 2-dimensional random vector with joint survival function $H(u,v)$. Let $\hat{H}(u,v)=n^{-1}\sum_{i=1}^n I(U_i\geq u,V_i\geq v)$ be the empirical estimator of $H(u,v)$ based on an independent and identically distributed sample $\{U_i,V_i\}, i=1,\cdots,n$. For any $\varepsilon>0$, there exist positive constants d_9 and d_{10} such that*

$$\Pr\left(\sup_{u,v}|\hat{H}(u,v)-H(u,v)|\geq\varepsilon\right)\leq d_9\exp\{-d_{10}n\varepsilon^2\}.$$

Proof: It is noted that

$$
\begin{aligned}
& \hat{H}(u,v) - H(u,v) \\
= & \left\{ n^{-1} \sum_{i=1}^{n} I(U_i < u, V_i < v) - \Pr(U < u, V < v) \right\} - \left\{ n^{-1} \sum_{i=1}^{n} I(U_i < u) - \Pr(U < u) \right\} \\
& - \left\{ n^{-1} \sum_{i=1}^{n} I(V_i < v) - \Pr(V < v) \right\} \\
= & \big\{ \hat{F}_{U,V}(u,v) - F_{U,V}(u,v) \big\} - \big\{ \hat{F}_U(u) - F_U(u) \big\} - \big\{ \hat{F}_V(v) - F_V(v) \big\}, \qquad \text{(A.2)}
\end{aligned}
$$

where $F_{U,V}(u,v)$, $F_U(u)$ and $F_V(v)$ are cumulative distribution functions of (U,V), U and V, and $\hat{F}_{U,V}(u,v)$, $\hat{F}_U(u)$ and $\hat{F}_V(v)$ are empirical versions of $F_{U,V}(u,v)$, $F_U(u)$ and $F_V(v)$. From Equation (A.2), it is easy to see that

$$
\begin{aligned}
& \sup_{u,v} |\hat{H}(u,v) - H(u,v)| \\
= & \sup_{u,v} |\hat{F}_{U,V}(u,v) - F_{U,V}(u,v)| + \sup_{u} |\hat{F}_U(u) - F_U(u)| + \sup_{v} |\hat{F}_V(v) - F_V(v)|. \quad \text{(A.3)}
\end{aligned}
$$

According to the well-known Dvoretzky-Kiefer-Wolfowitz inequality (Dvoretzky et al., 1956), there exist positive constants C_1 and C_2 such that

$$
\Pr\left(\sup_{u} |\hat{F}_U(u) - F_U(u)| \geq \frac{\varepsilon}{3} \right) \leq C_1 \exp\left\{ -\frac{2}{9} n \varepsilon^2 \right\} \qquad \text{(A.4)}
$$

and

$$
\Pr\left(\sup_{v} |\hat{F}_V(v) - F_V(v)| \geq \frac{\varepsilon}{3} \right) \leq C_2 \exp\left\{ -\frac{2}{9} n \varepsilon^2 \right\}. \qquad \text{(A.5)}
$$

Applying the multi-dimensional extension of Dvoretzky-Kiefer-Wolfowitz inequality (Kiefer, 1961), we could obtain

$$
\Pr\left(\sup_{u,v} |\hat{F}_{U,V}(u,v) - F_{U,V}(u,v)| \geq \frac{\varepsilon}{3} \right) \leq C_3 \exp\big\{ -C_4 n \varepsilon^2 \big\}, \qquad \text{(A.6)}
$$

where C_3 and C_4 are generic positive constants. Based on Equations (A.3) to (A.6), we finally arrive at

$$
\begin{aligned}
& \Pr\left(\sup_{u,v} |\hat{H}(u,v) - H(u,v)| \geq \varepsilon \right) \\
\leq & \Pr\left(\sup_{u,v} |\hat{F}_{U,V}(u,v) - F_{U,V}(u,v)| \geq \frac{\varepsilon}{3} \right) \\
& + \Pr\left(\sup_{u} |\hat{F}_U(u) - F_U(u)| \geq \frac{\varepsilon}{3} \right) \\
& + \Pr\left(\sup_{v} |\hat{F}_V(v) - F_V(v)| \geq \frac{\varepsilon}{3} \right) \\
\leq & 3C_5 \exp\big\{ -C_6 n \varepsilon^2 \big\} \\
\triangleq & d_9 \exp\big\{ -d_{10} n \varepsilon^2 \big\}
\end{aligned}
$$

where $C_5 = \max\{C_1, C_2, C_3\}$, $C_6 = \min\{2/9, C_4\}$, $d_9 = 3C_5$ and $d_{10} = C_6$.

$\square$

Lemma 4. *Under Assumption (A1), if $n^{1/2}\varepsilon \to \infty$ as n goes to ∞, for sufficiently large n, we have*

$$\Pr\left(\sup_{0\le t_1\le t_2\le\tau} |\hat{S}(t_1,t_2) - S(t_1,t_2)| \ge \varepsilon\right) \le d_{11}\exp\left\{ - d_{12} n\varepsilon^2\right\},$$

where d_{11} and d_{12} are generic positive constants.

Proof: Due to the fact that T_1 and T_2 are rightly censored by C independently, we could obtain

$$\Pr(Y \ge t_1, Z \ge t_2) = \Pr(T_1 \ge t_1, T_2 \ge t_2, C \ge t_2) = S(t_1,t_2)G(t_2),$$

for any $t_1 \le t_2$. Thus $S(t_1,t_2)$ can be expressed as $S(t_1,t_2) = \Pr(Y \ge t_1, Z \ge t_2)/G(t_2)$. And

$$\begin{aligned}
& |\hat{S}(t_1,t_2) - S(t_1,t_2)| \\
= & \left|\frac{n^{-1}\sum_{i=1}^n I(Y_i \ge t_1, Z_i \ge t_2)}{\hat{G}(t_2)} - \frac{\Pr(Y \ge t_1, Z \ge t_2)}{G(t_2)}\right| \\
\le & \left|\frac{n^{-1}\sum_{i=1}^n I(Y_i \ge t_1, Z_i \ge t_2)}{\hat{G}(t_2)} - \frac{n^{-1}\sum_{i=1}^n I(Y_i \ge t_1, Z_i \ge t_2)}{G(t_2)}\right| \\
& + \left|\frac{n^{-1}\sum_{i=1}^n I(Y_i \ge t_1, Z_i \ge t_2)}{G(t_2)} - \frac{\Pr(Y \ge t_1, Z \ge t_2)}{G(t_2)}\right| \\
\le & \left|\frac{1}{\hat{G}(t_2)} - \frac{1}{G(t_2)}\right| + \frac{1}{G(t_2)}\left|n^{-1}\sum_{i=1}^n I(Y_i \ge t_1, Z_i \ge t_2) - \Pr(Y \ge t_1, Z \ge t_2)\right|.
\end{aligned}$$

According to Lemma 2,

$$\Pr\left(\sup_{t_2\in[0,\tau]}\left|\frac{1}{\hat{G}(t_2)} - \frac{1}{G(t_2)}\right| \ge \frac{\varepsilon}{2}\right) \le 5\exp\left\{ - \frac{d_8}{4} n\varepsilon^2\right\}.$$

In addition,

$$\begin{aligned}
& \Pr\left(\sup_{0\le t_1\le t_2\le\tau} \frac{1}{G(t_2)}\left|n^{-1}\sum_{i=1}^n I(Y_i \ge t_1, Z_i \ge t_2) - \Pr(Y \ge t_1, Z \ge t_2)\right| \ge \frac{\varepsilon}{2}\right) \\
\le & \Pr\left(\sup_{t_1,t_2}\left|n^{-1}\sum_{i=1}^n I(Y_i \ge t_1, Z_i \ge t_2) - \Pr(Y \ge t_1, Z \ge t_2)\right| \ge \frac{\varepsilon}{2}\eta\right) \\
\le & d_9\exp\left\{ - d_{10} n\frac{\varepsilon^2\eta^2}{4}\right\},
\end{aligned}$$

where the first inequality is obtained based on Assumption (A1), and the second is from Lemma 3. At last, we have

$$
\begin{aligned}
&\Pr\left(\sup_{0\le t_1\le t_2\le\tau}|\hat{S}(t_1,t_2)-S(t_1,t_2)|\ge\varepsilon\right)\\
\le\ &\Pr\left(\sup_{t_2\in[0,\tau]}\left|\frac{1}{\hat{G}(t_2)}-\frac{1}{G(t_2)}\right|\ge\frac{\varepsilon}{2}\right)\\
&+\Pr\left(\sup_{t_1,t_2}\left|n^{-1}\sum_{i=1}^{n}I(Y_i\ge t_1,Z_i\ge t_2)-\Pr(Y\ge t_1,Z\ge t_2)\right|\ge\frac{\varepsilon}{2}\eta\right)\\
\le\ &5\exp\left\{-\frac{d_8}{4}n\varepsilon^2\right\}+d_9\exp\left\{-d_{10}n\frac{\varepsilon^2\eta^2}{4}\right\}\\
\le\ &d_{11}\exp\left\{-d_{12}n\varepsilon^2\right\},
\end{aligned}
$$

where $d_{11}=\max\{5,d_9\}$ and $d_{12}=\min\{d_8/4,d_{10}\eta^2/4\}$.

□

To facilitate the presentation of the proof, we firstly define an oracle estimator of ω_k as if the empirical survival functions of covariates and the joint survival function of $(T_1,T_2)^T$ are known in advance. Denote this oracle estimator by $\tilde{\omega}_k=\tilde{\rho}_k^2$, where

$$
\tilde{\rho}_k=\frac{\widetilde{\mathrm{dcov}}\{S_k(X_k),S(T_1,T_2)\}}{\sqrt{\widetilde{\mathrm{dcov}}\{S_k(X_k),S_k(X_k)\}}\sqrt{\widetilde{\mathrm{dcov}}\{S(T_1,T_2),S(T_1,T_2)\}}},
$$

where $\widetilde{\mathrm{dcov}}\{S_k(X_k),S(T_1,T_2)\}$, $\widetilde{\mathrm{dcov}}\{S_k(X_k),S_k(X_k)\}$ and $\widetilde{\mathrm{dcov}}\{S(T_1,T_2),S(T_1,T_2)\}$ are defined according to (2) with $S_k(\cdot)$'s and $S(\cdot,\cdot)$ being regarded already known.

Proof of Theorem 1:

Due to the boundness of $S_k(\cdot)$'s and $S(\cdot,\cdot)$, according to the remark of Theorem 1 of Li et al. (2012), it is easily obtained that there exist positive constants C_7 and C_8 such that

$$
\Pr(|\tilde{\omega}_k-\omega_k|\ge 2^{-1}c_1n^{-\kappa})\le C_7\exp\{-C_8n^{1-2\kappa}\}. \tag{A.7}
$$

We now consider $\Pr(|\hat{\omega}_k-\tilde{\omega}_k|\ge 2^{-1}c_1n^{-\kappa})$. Let us deal with the numerator of $\hat{\omega}_k$ and $\tilde{\omega}_k$ firstly. Recall that

$$
\widehat{\mathrm{dcov}}^2\{S_k(X_k),S(T_1,T_2)\}=\hat{S}_{1k}+\hat{S}_{2k}-2\hat{S}_{3k}
$$

and

$$
\widetilde{\mathrm{dcov}}^2\{S_k(X_k),S(T_1,T_2)\}=\tilde{S}_{1k}+\tilde{S}_{2k}-2\tilde{S}_{3k},
$$

where $\hat{S}_{1k}$, $\hat{S}_{2k}$ and $\hat{S}_{3k}$ are defined by replacing U_i, U_j, V_i, V_j and V_l in (3) to (5) with $\hat{S}_k(X_{ki})$, $\hat{S}_k(X_{kj})$, $\hat{S}(Y_i,Z_i)$, $\hat{S}(Y_j,Z_j)$ and $\hat{S}(Y_l,Z_l)$, $\tilde{S}_{1k}$, $\tilde{S}_{2k}$ and $\tilde{S}_{3k}$ are given by replacing U_i, U_j, V_i, V_j and V_l with $S_k(X_{ki})$, $S_k(X_{kj})$, $S(Y_i,Z_i)$, $S(Y_j,Z_j)$ and $S(Y_l,Z_l)$.

For any positive ε satisfying $n^{1/2}\varepsilon \to \infty$ as n goes to ∞,

$$
\begin{aligned}
& \Pr(|\hat{S}_{1k} - \tilde{S}_{1k}| \geq \varepsilon) \\
= \; & \Pr\left(\left| \frac{1}{n^2} \sum_{i,j=1}^{n} |\hat{S}_k(X_{ki}) - \hat{S}_k(X_{kj})||\hat{S}(Y_i, Z_i) - \hat{S}(Y_j, Z_j)| \right.\right. \\
& \quad \left.\left. - \frac{1}{n^2} \sum_{i,j=1}^{n} |S_k(X_{ki}) - S_k(X_{kj})||S(Y_i, Z_i) - S(Y_j, Z_j)| \right| \geq \varepsilon \right) \\
\leq \; & \Pr\left(\frac{1}{n^2} \sum_{i,j=1}^{n} |\hat{S}_k(X_{ki}) - \hat{S}_k(X_{kj})|||\hat{S}(Y_i, Z_i) - \hat{S}(Y_j, Z_j)| - |S(Y_i, Z_i) - S(Y_j, Z_j)|| \geq 2^{-1}\varepsilon \right) \\
& + \Pr\left(\frac{1}{n^2} \sum_{i,j=1}^{n} ||\hat{S}_k(X_{ki}) - \hat{S}_k(X_{kj})| - |S_k(X_{ki}) - S_k(X_{kj})|||S(Y_i, Z_i) - S(Y_j, Z_j)| \geq 2^{-1}\varepsilon \right).
\end{aligned} \tag{A.8}
$$

By the fact that

$$
\begin{aligned}
& ||\hat{S}(Y_i, Z_i) - \hat{S}(Y_j, Z_j)| - |S(Y_i, Z_i) - S(Y_j, Z_j)|| \\
\leq \; & |\hat{S}(Y_i, Z_i) - S(Y_i, Z_i)| + |\hat{S}(Y_j, Z_j) - S(Y_j, Z_j)| \\
\leq \; & 2 \sup_{0 \leq t_1 \leq t_2 \leq \tau} |\hat{S}(t_1, t_2) - S(t_1, t_2)|
\end{aligned}
$$

and $|\hat{S}_k(X_{ki}) - \hat{S}_k(X_{kj})| \leq 1$, we have

$$
\begin{aligned}
& \Pr\left(\frac{1}{n^2} \sum_{i,j=1}^{n} |\hat{S}_k(X_{ki}) - \hat{S}_k(X_{kj})|||\hat{S}(Y_i, Z_i) - \hat{S}(Y_j, Z_j)| - |S(Y_i, Z_i) - S(Y_j, Z_j)|| \geq 2^{-1}\varepsilon \right) \\
\leq \; & \Pr\left(\sup_{0 \leq t_1 \leq t_2 \leq \tau} |\hat{S}(t_1, t_2) - S(t_1, t_2)| \geq 4^{-1}\varepsilon \right) \\
\leq \; & d_{11} \exp\{-16^{-1} d_{12} n \varepsilon^2\},
\end{aligned} \tag{A.9}
$$

where the last inequality comes from Lemma 4. In the similar way, we could prove that

$$
\begin{aligned}
& \Pr\left(\frac{1}{n^2} \sum_{i,j=1}^{n} ||\hat{S}_k(X_{ki}) - \hat{S}_k(X_{kj})| - |S_k(X_{ki}) - S_k(X_{kj})|||S(Y_i, Z_i) - S(Y_j, Z_j)| \geq 2^{-1}\varepsilon \right) \\
\leq \; & \Pr\left(\sup_{x_k \in R} |\hat{S}_k(x_k) - S_k(x_k)| \geq 4^{-1}\varepsilon \right) \\
\leq \; & 2 \exp\{-8^{-1} n \varepsilon^2\},
\end{aligned} \tag{A.10}
$$

where the last inequality is obtained based on Dvoretzky-Kiefer-Wolfowitz inequality (Dvoretzky et al., 1956). From Equations (A.8) to (A.10), it is gotten that

$$
\Pr(|\hat{S}_{1k} - \tilde{S}_{1k}| \geq \varepsilon) \leq C_9 \exp\{-C_{10} n \varepsilon^2\},
$$

where $C_9 = \max\{d_{11}, 2\}$ and $C_{10} = \min\{16^{-1}d_{12}, 8^{-1}\}$.

The same convergence rates could be proved for $\Pr(|\hat{S}_{2k} - \tilde{S}_{2k}| \geq \varepsilon)$ and $\Pr(|\hat{S}_{3k} - \tilde{S}_{3k}| \geq \varepsilon)$. By the techniques used in Lemmas S4 and S5 of Liu et al. (2014), there exist positive constants C_{11} and C_{12} such that

$$\Pr(|\widehat{\mathrm{dcov}}^2\{S_k(X_k), S(T_1, T_2)\} - \widetilde{\mathrm{dcov}}^2\{S_k(X_k), S(T_1, T_2)\}| \geq \varepsilon) \leq C_{11}\exp\{-C_{12}n\varepsilon^2\}.$$

We could achieve the same convergence rates for denominators of $\hat{\omega}_k$ and $\tilde{\omega}_k$ likewise. Utilizing the techniques in Lemmas S4 and S5 of Liu et al. (2014), we have

$$\Pr(|\hat{\omega}_k - \tilde{\omega}_k| \geq \varepsilon) \leq C_{13}\exp\{-C_{14}n\varepsilon^2\}, \tag{A.11}$$

where C_{13} and C_{14} are positive constants. Under Assumption (A2), it is easy to see that $n^{1/2-\kappa} \to \infty$ as n goes to ∞. Thus taking $\varepsilon = 2^{-1}c_1n^{-\kappa}$, Equation (A.11) becomes

$$\Pr(|\hat{\omega}_k - \tilde{\omega}_k| \geq 2^{-1}c_1n^{-\kappa}) \leq C_{13}\exp\{-C_{15}n^{1-2\kappa}\}, \tag{A.12}$$

where $C_{15} = 4^{-1}c_1^2C_{14}$. Combing Equations (A.7) and (A.12), we could conclude that

$$\Pr(|\hat{\omega}_k - \omega_k| \geq c_1n^{-\kappa}) \leq C_{16}\exp\{-C_{17}n^{1-2\kappa}\},$$

where $C_{16} = \max\{C_7, C_{13}\}$ and $C_{17} = \min\{C_8, C_{15}\}$. Furthermore,

$$\begin{aligned} & \Pr\left(\max_{1\leq k\leq p}|\hat{\omega}_k - \omega_k| \geq c_1n^{-\kappa}\right) \\ \leq\ & \sum_{k=1}^{p}\Pr(|\hat{\omega}_k - \omega_k| \geq c_1n^{-\kappa}) \\ \leq\ & C_{16}p\exp\{-C_{17}n^{1-2\kappa}\}. \end{aligned}$$

Let $d_1 = C_{16}$ and $d_2 = C_{17}$. This complete the proof of first part of Theorem.

In the next, let us turn to the proof of the second part. Noting that $\hat{\mathcal{A}} = \{k : \hat{\omega}_k \geq c_1n^{-\kappa}, k = 1, \cdots, p\}$, we can conclude that $\{\mathcal{A} \not\subseteq \hat{\mathcal{A}}\} \subseteq \{|\hat{\omega}_k - \omega_k| \geq c_1n^{-\kappa}$, for some $k \in A\}$. Thus $\{\max_{k\in\mathcal{A}}|\hat{\omega}_k - \omega_k| \leq c_1n^{-\kappa}\} \subseteq \{\mathcal{A} \subseteq \hat{\mathcal{A}}\}$. Consequently,

$$\begin{aligned} & \Pr(\mathcal{A} \subseteq \hat{\mathcal{A}}) \\ \geq\ & \Pr\left(\max_{k\in\mathcal{A}}|\hat{\omega}_k - \omega_k| \leq c_1n^{-\kappa}\right) \\ =\ & 1 - \Pr\left(\max_{k\in\mathcal{A}}|\hat{\omega}_k - \omega_k| \geq c_1n^{-\kappa}\right) \\ \geq\ & 1 - q\Pr(|\hat{\omega}_k - \omega_k| \geq c_1n^{-\kappa}\Big) \\ \geq\ & 1 - d_1q\exp\{-d_2n^{1-2\kappa}\}. \end{aligned}$$

□

Proof of Corollary 1: Define

$$\mathcal{B} = \{k : \omega_k \geq 2^{-1} c_1 n^{-\kappa}, k = 1, \cdots, p\}$$

and

$$\mathcal{C} = \left\{ \max_{1 \leq k \leq p} |\hat{\omega}_k - \omega_k| \leq 2^{-1} c_1 n^{-\kappa} \right\}.$$

On one hand, it is easy to see that, for any $k \in \mathcal{B}$, $2c_1^{-1} n^{\kappa} \omega_k \geq 1$. Thus $|\mathcal{B}| \leq 2c_1^{-1} n^{\kappa} \sum_{1 \leq k \leq p} \omega_k$. On the other hand, we could show that $\mathcal{C} \subseteq \{|\hat{\mathcal{A}}| \leq |\mathcal{B}|\}$. Therefore, we could conclude that there exist positive constants d_3 and d_4

$$\Pr\left\{ |\hat{\mathcal{A}}| \leq 2c_1^{-1} n^{\kappa} \sum_{1 \leq k \leq p} \omega_k \right\} \geq \Pr\{|\hat{\mathcal{A}}| \leq |\mathcal{B}|\} \geq \Pr\{\mathcal{C}\} \geq 1 - d_3 p \exp\{-d_4 n^{1-2\kappa}\},$$

where the last inequality is gotten in the similar way as the proof in Theorem 1.

$\square$

Proof of Theorem 2: For $k \in \mathcal{A}^c$, X_k is independent of (T_1, T_2) according to the Assumption (i) in Theorem 2. Thus $S_k(X_k)$ and $S(T_1, T_2)$ are independent. From Theorem 3 of Székely et al. (2007), we could conclude that $\rho_k = 0$, and furthermore $\omega_k = 0$. For $k \in \mathcal{A}$, from Assumption (A2), we have that $\omega_k \geq 2c_1 n^{-\kappa}$. Therefore, we could draw the conclusion that

$$\max_{k \in \mathcal{A}^c} \omega_k < \min_{k \in \mathcal{A}} \omega_k,$$

and $\omega_k = 0$ if and only if $k \in \mathcal{A}^c$. Thus, the first part of Theorem 2 is proved.

Now, let's deal with the second part. Under Assumption (A2) and the assumptions listed in Theorem 2, we have

$$\begin{aligned}
& \Pr\left(\min_{k \in \mathcal{A}} \hat{\omega}_k \leq \max_{k \in \mathcal{A}^c} \hat{\omega}_k \right) \\
= \; & \Pr\left(\max_{k \in \mathcal{A}^c} \hat{\omega}_k - \max_{k \in \mathcal{A}^c} \omega_k - \min_{k \in \mathcal{A}} \hat{\omega}_k + \min_{k \in \mathcal{A}} \omega_k \geq \min_{k \in \mathcal{A}} \omega_k \right) \\
\leq \; & \Pr\left(\max_{k \in \mathcal{A}^c} |\hat{\omega}_k - \omega_k| \geq c_1 n^{-\kappa} \right) + \Pr\left(\max_{k \in \mathcal{A}} |\hat{\omega}_k - \omega_k| \geq c_1 n^{-\kappa} \right) \\
\leq \; & 2 \Pr\left(\max_{1 \leq k \leq p} |\hat{\omega}_k - \omega_k| \geq c_1 n^{-\kappa} \right) \\
\leq \; & 2 d_1 p \exp\{-d_2 n^{1-2\kappa}\}.
\end{aligned}$$

Finally, we could arrive at

$$\Pr\left(\min_{k \in \mathcal{A}} \hat{\omega}_k > \max_{k \in \mathcal{A}^c} \hat{\omega}_k \right) \geq 1 - 2 d_1 p \exp\{-d_2 n^{1-2\kappa}\}.$$

This finishes the proof of second part.

$\square$

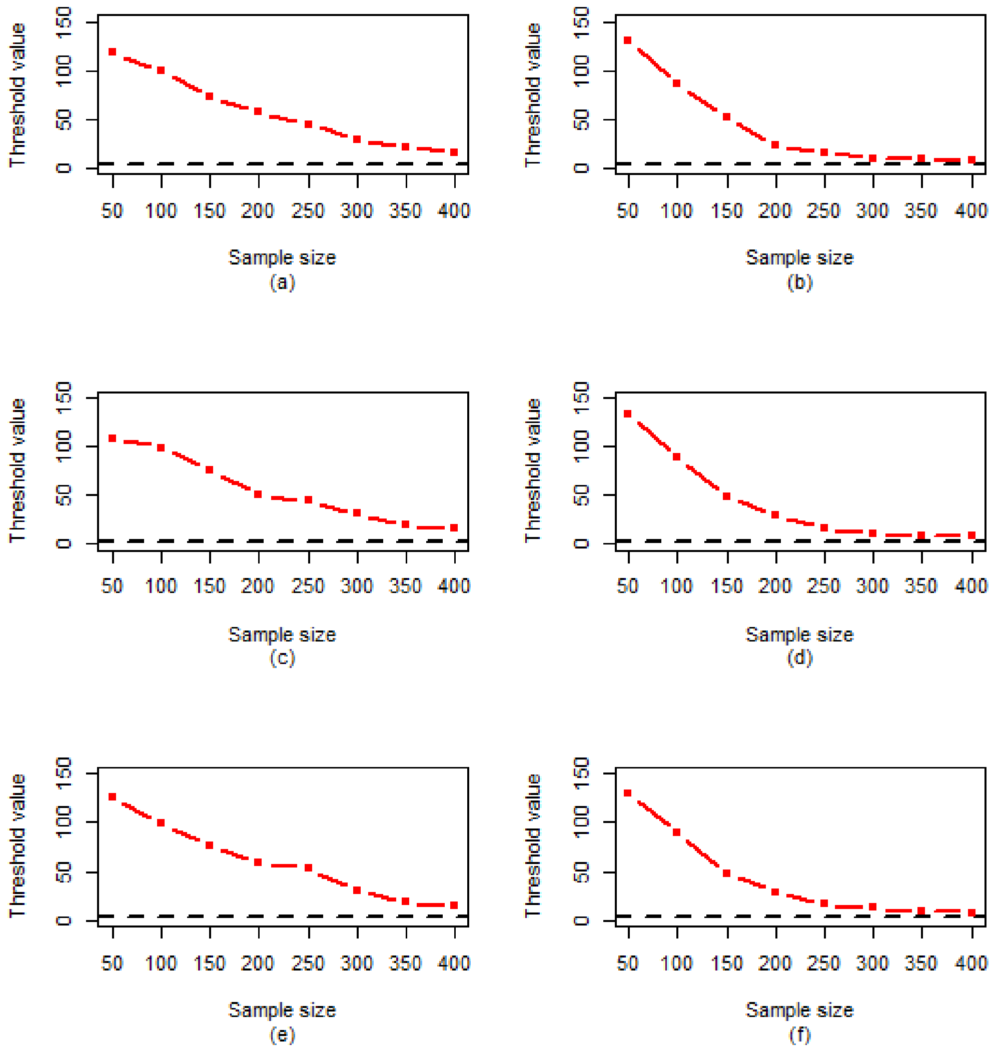


Figure 1: The line charts of adaptive threshold value versus sample size for aJMDC-SIS under different settings in Example 1: (a) $\theta = 0.5$, $\rho = 0.6$; (b) $\theta = 0.5$, $\rho = 0.9$; (c) $\theta = 2$, $\rho = 0.6$; (d) $\theta = 2$, $\rho = 0.9$; (e) $\theta = 8$, $\rho = 0.6$; (f) $\theta = 8$, $\rho = 0.9$. The black dashed lines are the number of truly important covariates, while the red broken lines are the threshold values identified by aJMDC-SIS.

Table 1: $\mathcal{P}_k$, $\mathcal{P}_a$ and threshold value d_0 in Example 1.

θ	n	Method	$\rho = 0.6$					$\rho = 0.9$				
			$\mathcal{P}_1$	$\mathcal{P}_2$	$\mathcal{P}_3$	$\mathcal{P}_a$	d_0	$\mathcal{P}_1$	$\mathcal{P}_2$	$\mathcal{P}_3$	$\mathcal{P}_a$	d_0
0.5	100	aJMDC-SIS	0.920	0.258	0.948	0.248	99	0.966	0.898	0.966	0.892	84
		JMDC-SIS	0.788	0.102	0.872	0.100	21	0.890	0.754	0.892	0.736	21
		RCDCS	0.008	0.558	0.030	0.000	21	0.010	0.040	0.016	0.000	21
		JCR	0.806	0.038	0.900	0.038	21	0.880	0.686	0.890	0.678	21
	200	aJMDC-SIS	0.992	0.534	1.000	0.532	57	1.000	0.992	0.998	0.992	26
		JMDC-SIS	0.994	0.486	1.000	0.484	37	0.998	0.994	0.998	0.992	37
		RCDCS	0.028	0.916	0.084	0.000	37	0.020	0.098	0.020	0.002	37
		JCR	0.992	0.212	0.996	0.212	37	0.998	0.980	0.998	0.980	37
2	100	aJMDC-SIS	0.912	0.260	0.948	0.250	99	0.962	0.902	0.972	0.896	84
		JMDC-SIS	0.772	0.088	0.858	0.084	21	0.886	0.742	0.890	0.728	21
		RCDCS	0.012	0.536	0.028	0.000	21	0.008	0.036	0.014	0.000	21
		JCR	0.814	0.044	0.910	0.044	21	0.876	0.680	0.896	0.676	21
	200	aJMDC-SIS	0.994	0.516	1.000	0.512	55	0.998	0.996	1.000	0.994	28
		JMDC-SIS	0.992	0.480	1.000	0.476	37	0.998	0.988	0.998	0.988	37
		RCDCS	0.022	0.918	0.098	0.002	37	0.014	0.096	0.020	0.004	37
		JCR	0.994	0.196	0.998	0.196	37	0.998	0.982	0.998	0.982	37
8	100	aJMDC-SIS	0.908	0.260	0.956	0.250	98	0.960	0.894	0.968	0.884	88
		JMDC-SIS	0.776	0.084	0.858	0.080	21	0.874	0.716	0.908	0.704	21
		RCDCS	0.012	0.542	0.036	0.000	21	0.006	0.034	0.012	0.000	21
		JCR	0.812	0.044	0.912	0.044	21	0.876	0.674	0.888	0.670	21
	200	aJMDC-SIS	0.996	0.530	1.000	0.528	58	1.000	1.000	1.000	1.000	29
		JMDC-SIS	0.994	0.488	1.000	0.484	37	0.998	0.994	0.996	0.994	37
		RCDCS	0.028	0.916	0.106	0.004	37	0.020	0.108	0.020	0.006	37
		JCR	0.992	0.202	0.998	0.202	37	0.998	0.982	0.998	0.982	37

Table 2: The 5%, 25%, 50%, 75% and 95% quantiles of minimum model size to include all the important covariates in Example 1.

θ	n	Method	$\rho = 0.6$					$\rho = 0.9$				
			5%	25%	50%	75%	95%	5%	25%	50%	75%	95%
0.5	100	JMDC-SIS	10	89	250	576	1376	3	4	6	24	210
		RCDCS	334	949	1371	1712	1937	325	983	1462	1777	1966
		JCR	27	201	579	1252	1865	3	5	9	35	258
	200	JMDC-SIS	4	10	42	126	429	3	3	4	5	9
		RCDCS	170	675	1195	1585	1921	409	1125	1520	1776	1963
		JCR	5	49	212	663	1603	3	4	5	6	15
2	100	JMDC-SIS	13	88	260	600	1424	3	4	6	26	218
		RCDCS	324	955	1374	1680	1942	380	1015	1469	1782	1966
		JCR	25	207	606	1254	1803	3	4	8	32	301
	200	JMDC-SIS	4	11	40	131	498	3	3	4	5	9
		RCDCS	188	734	1137	1572	1938	425	999	1473	1763	1963
		JCR	5	52	202	705	1558	3	4	5	6	16
8	100	JMDC-SIS	13	81	255	614	1351	3	4	6	29	268
		RCDCS	337	890	1331	1730	1953	361	1011	1460	1775	1960
		JCR	24	192	665	1258	1813	3	4	8	37	299
	200	JMDC-SIS	4	11	39	138	527	3	3	4	5	9
		RCDCS	215	661	1158	1575	1919	432	996	1434	1740	1972
		JCR	6	53	222	694	1674	3	4	5	6	19

Table 3: $\mathcal{P}_k$, $\mathcal{P}_a$ and threshold value d_0 in Example 2.

θ	n	Method	$\rho = 0.6$					$\rho = 0.9$				
			$\mathcal{P}_1$	$\mathcal{P}_2$	$\mathcal{P}_3$	$\mathcal{P}_a$	d_0	$\mathcal{P}_1$	$\mathcal{P}_2$	$\mathcal{P}_3$	$\mathcal{P}_a$	d_0
0.5	200	aJMDC-SIS	0.924	0.73	0.848	0.654	61	0.928	0.896	0.896	0.850	38
		JMDC-SIS	0.928	0.652	0.800	0.576	37	0.932	0.886	0.884	0.836	37
		RCDCS	0.328	0.020	0.022	0.006	37	0.470	0.098	0.060	0.032	37
		JCR	0.786	0.428	0.710	0.374	37	0.768	0.712	0.724	0.650	37
	300	aJMDC-SIS	0.990	0.894	0.940	0.854	57	0.992	0.974	0.962	0.952	19
		JMDC-SIS	0.994	0.876	0.952	0.860	52	0.996	0.982	0.976	0.968	52
		RCDCS	0.756	0.046	0.050	0.010	52	0.856	0.196	0.108	0.078	52
		JCR	0.926	0.594	0.844	0.556	52	0.930	0.882	0.902	0.858	52
2	200	aJMDC-SIS	0.938	0.722	0.840	0.646	60	0.924	0.886	0.882	0.836	37
		JMDC-SIS	0.920	0.652	0.804	0.572	37	0.922	0.874	0.878	0.826	37
		RCDCS	0.344	0.022	0.030	0.004	37	0.470	0.080	0.062	0.032	37
		JCR	0.766	0.442	0.696	0.382	37	0.752	0.700	0.716	0.634	37
	300	aJMDC-SIS	0.992	0.896	0.932	0.850	56	0.990	0.964	0.956	0.942	20
		JMDC-SIS	0.988	0.884	0.954	0.862	52	0.996	0.982	0.976	0.968	52
		RCDCS	0.762	0.048	0.048	0.010	52	0.852	0.198	0.116	0.080	52
		JCR	0.916	0.590	0.844	0.554	52	0.934	0.886	0.910	0.862	52
8	200	aJMDC-SIS	0.940	0.718	0.840	0.644	61	0.918	0.864	0.864	0.816	37
		JMDC-SIS	0.932	0.644	0.796	0.572	37	0.920	0.866	0.858	0.804	37
		RCDCS	0.348	0.024	0.026	0.006	37	0.474	0.076	0.064	0.036	37
		JCR	0.758	0.434	0.706	0.382	37	0.754	0.706	0.710	0.640	37
	300	aJMDC-SIS	0.992	0.880	0.950	0.848	56	0.994	0.968	0.966	0.952	20
		JMDC-SIS	0.990	0.874	0.946	0.848	52	0.996	0.984	0.978	0.968	52
		RCDCS	0.772	0.046	0.048	0.012	52	0.854	0.204	0.124	0.086	52
		JCR	0.914	0.592	0.838	0.550	52	0.942	0.892	0.898	0.866	52

Table 4: The 5%, 25%, 50%, 75% and 95% quantiles of minimum model size to include all the important covariates in Example 2.

θ	n	Method	$\rho = 0.6$					$\rho = 0.9$				
			5%	25%	50%	75%	95%	5%	25%	50%	75%	95%
0.5	200	JMDC-SIS	3	6	22	107	532	3	3	5	18	166
		RCDCS	214	709	1100	1441	1858	55	227	499	871	1417
		JCR	3	13	80	393	1517	3	4	13	87	907
	300	JMDC-SIS	3	3	7	24	121	3	3	3	5	31
		RCDCS	215	633	980	1378	1797	30	142	350	579	1088
		JCR	3	5	31	203	1123	3	3	5	16	213
2	200	JMDC-SIS	3	6	24	101	506	3	3	5	20	177
		RCDCS	245	689	1086	1465	1847	54	241	466	826	1440
		JCR	3	14	87	400	1543	3	4	13	100	904
	300	JMDC-SIS	3	3	8	29	133	3	3	3	5	32
		RCDCS	206	625	981	1389	1797	27	146	323	591	1127
		JCR	3	5	33	213	1164	3	3	5	18	236
8	200	JMDC-SIS	3	6	23	108	536	3	3	5	22	220
		RCDCS	269	720	1052	1447	1850	54	242	447	815	1438
		JCR	3	14	80	419	1513	3	4	13	90	1062
	300	JMDC-SIS	3	3	8	26	134	3	3	3	5	36
		RCDCS	220	619	974	1398	1775	31	152	329	597	1069
		JCR	3	5	36	213	1161	3	3	5	17	210

Table 5: $\mathcal{P}_k$, $\mathcal{P}_a$ and threshold value d_0 in Example 3.

θ	n	Method	$\rho = 0.6$					$\rho = 0.9$				
			$\mathcal{P}_1$	$\mathcal{P}_2$	$\mathcal{P}_3$	$\mathcal{P}_a$	d_0	$\mathcal{P}_1$	$\mathcal{P}_2$	$\mathcal{P}_3$	$\mathcal{P}_a$	d_0
0.5	200	aJMDC-SIS	0.920	0.826	0.782	0.678	72	0.982	0.966	0.952	0.942	48
		JMDC-SIS	0.848	0.758	0.716	0.574	37	0.990	0.976	0.944	0.934	37
		RCDCS	0.048	0.046	0.030	0.002	37	0.050	0.058	0.036	0.022	37
		JCR	0.458	0.512	0.600	0.284	37	0.580	0.634	0.688	0.524	37
	300	aJMDC-SIS	0.992	0.944	0.890	0.862	50	1.000	0.994	0.986	0.986	36
		JMDC-SIS	0.992	0.948	0.922	0.896	52	1.000	0.998	0.994	0.994	52
		RCDCS	0.072	0.058	0.040	0.002	52	0.056	0.072	0.066	0.030	52
		JCR	0.692	0.746	0.844	0.558	52	0.786	0.824	0.886	0.760	52
2	200	aJMDC-SIS	0.924	0.816	0.780	0.686	69	0.986	0.976	0.964	0.956	48
		JMDC-SIS	0.862	0.770	0.722	0.586	37	0.986	0.976	0.960	0.942	37
		RCDCS	0.044	0.048	0.022	0.002	37	0.052	0.052	0.032	0.014	37
		JCR	0.454	0.526	0.604	0.282	37	0.584	0.624	0.690	0.528	37
	300	aJMDC-SIS	0.992	0.950	0.892	0.870	49	1.000	0.996	0.984	0.984	35
		JMDC-SIS	0.994	0.942	0.922	0.890	52	1.000	0.998	0.998	0.998	52
		RCDCS	0.072	0.062	0.040	0.002	52	0.064	0.064	0.064	0.030	52
		JCR	0.698	0.744	0.856	0.566	52	0.798	0.852	0.894	0.772	52
8	200	aJMDC-SIS	0.942	0.838	0.772	0.692	70	0.984	0.976	0.964	0.956	48
		JMDC-SIS	0.880	0.768	0.722	0.592	37	0.984	0.982	0.968	0.958	37
		RCDCS	0.042	0.042	0.022	0.000	37	0.044	0.056	0.030	0.016	37
		JCR	0.468	0.534	0.614	0.292	37	0.594	0.656	0.704	0.552	37
	300	aJMDC-SIS	0.996	0.962	0.888	0.876	49	1.000	0.998	0.988	0.988	34
		JMDC-SIS	0.994	0.946	0.920	0.888	52	1.000	0.998	0.996	0.996	52
		RCDCS	0.078	0.066	0.032	0.004	52	0.070	0.070	0.064	0.030	52
		JCR	0.710	0.760	0.864	0.580	52	0.808	0.854	0.896	0.774	52

Table 6: The 5%, 25%, 50%, 75% and 95% quantiles of minimum model size to include all the important covariates in Example 3.

θ	n	Method	$\rho = 0.6$					$\rho = 0.9$				
			5%	25%	50%	75%	95%	5%	25%	50%	75%	95%
0.5	200	JMDC-SIS	3	6	26	100	644	3	3	3	5	46
		RCDCS	375	1009	1459	1723	1957	96	601	1080	1578	1919
		JCR	4	29	145	570	1635	3	7	30	208	1074
	300	JMDC-SIS	3	3	5	15	155	3	3	3	3	9
		RCDCS	346	905	1321	1702	1938	107	487	947	1442	1816
		JCR	3	9	38	184	823	3	4	10	47	391
2	200	JMDC-SIS	3	6	25	93	613	3	3	3	5	43
		RCDCS	387	985	1422	1749	1947	105	602	1088	1571	1902
		JCR	5	29	139	570	1525	3	7	31	192	1074
	300	JMDC-SIS	3	3	5	13	129	3	3	3	3	6
		RCDCS	307	914	1334	1699	1936	115	473	925	1387	1833
		JCR	3	8	32	165	810	3	4	9	41	339
8	200	JMDC-SIS	3	5	24	87	622	3	3	3	5	35
		RCDCS	399	985	1438	1719	1945	102	597	1090	1603	1902
		JCR	4	27	145	533	1586	3	6	28	180	1106
	300	JMDC-SIS	3	3	5	13	118	3	3	3	3	6
		RCDCS	313	929	1343	1676	1923	105	466	926	1450	1861
		JCR	3	7	30	160	756	3	4	8	43	338

Table 7: $\mathcal{P}_k$, $\mathcal{P}_a$ and threshold value d_0 in Example 4.

θ	n	Method	$\rho = 0.6$					$\rho = 0.9$				
			$\mathcal{P}_1$	$\mathcal{P}_2$	$\mathcal{P}_3$	$\mathcal{P}_a$	d_0	$\mathcal{P}_1$	$\mathcal{P}_2$	$\mathcal{P}_3$	$\mathcal{P}_a$	d_0
0.5	200	aJMDC-SIS	0.540	0.852	0.636	0.356	63	0.968	0.984	0.974	0.960	47
		JMDC-SIS	0.430	0.804	0.590	0.294	37	0.960	0.986	0.962	0.940	37
		RCDCS	0.138	0.096	0.026	0.006	37	0.252	0.214	0.124	0.076	37
		JCR	0.066	0.590	0.486	0.044	37	0.562	0.768	0.784	0.544	37
	300	aJMDC-SIS	0.810	0.938	0.802	0.642	59	0.996	0.998	0.994	0.992	22
		JMDC-SIS	0.766	0.956	0.808	0.630	52	0.994	0.998	0.998	0.992	52
		RCDCS	0.268	0.154	0.056	0.012	52	0.626	0.610	0.290	0.246	52
		JCR	0.104	0.768	0.696	0.082	52	0.764	0.928	0.944	0.758	52
2	200	aJMDC-SIS	0.546	0.840	0.638	0.372	64	0.962	0.980	0.972	0.950	48
		JMDC-SIS	0.422	0.806	0.584	0.290	37	0.946	0.980	0.956	0.926	37
		RCDCS	0.134	0.090	0.032	0.012	37	0.246	0.230	0.134	0.076	37
		JCR	0.066	0.582	0.482	0.048	37	0.556	0.772	0.792	0.538	37
	300	aJMDC-SIS	0.814	0.930	0.788	0.636	60	0.996	0.998	0.996	0.994	22
		JMDC-SIS	0.764	0.952	0.806	0.632	52	0.994	0.998	0.996	0.992	52
		RCDCS	0.258	0.152	0.058	0.012	52	0.612	0.600	0.282	0.230	52
		JCR	0.108	0.768	0.702	0.084	52	0.766	0.928	0.944	0.760	52
8	200	aJMDC-SIS	0.558	0.842	0.636	0.382	63	0.964	0.986	0.968	0.952	51
		JMDC-SIS	0.420	0.810	0.588	0.280	37	0.942	0.980	0.958	0.924	37
		RCDCS	0.132	0.086	0.034	0.010	37	0.256	0.228	0.134	0.082	37
		JCR	0.062	0.576	0.484	0.044	37	0.542	0.772	0.790	0.524	37
	300	aJMDC-SIS	0.820	0.930	0.790	0.638	62	0.996	0.998	0.994	0.994	23
		JMDC-SIS	0.764	0.950	0.814	0.630	52	0.994	0.998	0.996	0.992	52
		RCDCS	0.268	0.154	0.052	0.012	52	0.616	0.582	0.286	0.234	52
		JCR	0.112	0.764	0.694	0.086	52	0.770	0.924	0.940	0.762	52

Table 8: The 5%, 25%, 50%, 75% and 95% quantiles of minimum model size to include all the important covariates in Example 4.

θ	n	Method	$\rho = 0.6$					$\rho = 0.9$				
			5%	25%	50%	75%	95%	5%	25%	50%	75%	95%
0.5	200	JMDC-SIS	6	29	109	284	844	3	3	4	7	55
		RCDCS	206	511	859	1339	1828	26	133	268	511	1175
		JCR	41	314	754	1279	1846	3	6	28	165	893
	300	JMDC-SIS	3	9	29	87	392	3	3	3	4	7
		RCDCS	99	359	691	1111	1627	11	54	128	285	667
		JCR	28	206	588	1199	1776	3	5	9	48	307
2	200	JMDC-SIS	5	32	110	289	953	3	3	4	7	62
		RCDCS	191	506	875	1357	1831	26	128	266	519	1211
		JCR	44	322	738	1260	1781	3	6	27	160	903
	300	JMDC-SIS	3	10	30	90	413	3	3	3	4	7
		RCDCS	101	344	685	1121	1680	12	55	130	278	689
		JCR	27	204	595	1198	1784	3	5	9	50	347
8	200	JMDC-SIS	5	32	107	291	949	3	3	4	7	54
		RCDCS	193	488	886	1329	1836	27	122	262	534	1168
		JCR	42	311	742	1271	1835	3	6	29	170	917
	300	JMDC-SIS	3	10	31	97	420	3	3	3	4	7
		RCDCS	120	355	682	1092	1730	13	56	136	279	685
		JCR	23	205	615	1216	1777	3	5	9	46	323

Table 9: $\mathcal{P}_k$, $\mathcal{P}_a$ and threshold value d_0 in Example 5.

θ	n	Method	$\mathcal{P}_1$	$\mathcal{P}_2$	$\mathcal{P}_3$	$\mathcal{P}_a$	d_0
0.5	200	aJMDC-SIS	0.906	0.812	0.936	0.686	71
		JMDC-SIS	0.868	0.736	0.906	0.584	37
		RCDCS	0.988	0.484	0.950	0.456	37
		JCR	0.732	0.280	0.212	0.046	37
	300	aJMDC-SIS	0.982	0.964	0.992	0.940	67
		JMDC-SIS	0.988	0.970	0.990	0.948	52
		RCDCS	1.000	0.892	1.000	0.892	52
		JCR	0.924	0.386	0.340	0.106	52
2	200	aJMDC-SIS	0.906	0.814	0.940	0.694	72
		JMDC-SIS	0.860	0.724	0.890	0.556	37
		RCDCS	0.990	0.482	0.946	0.452	37
		JCR	0.728	0.284	0.212	0.044	37
	300	aJMDC-SIS	0.982	0.966	0.984	0.932	65
		JMDC-SIS	0.986	0.968	0.990	0.944	52
		RCDCS	1.000	0.886	1.000	0.886	52
		JCR	0.924	0.388	0.336	0.112	52
8	200	aJMDC-SIS	0.896	0.804	0.938	0.680	72
		JMDC-SIS	0.850	0.718	0.886	0.540	37
		RCDCS	0.990	0.478	0.938	0.442	37
		JCR	0.724	0.286	0.210	0.044	37
	300	aJMDC-SIS	0.974	0.968	0.988	0.930	66
		JMDC-SIS	0.982	0.970	0.988	0.940	52
		RCDCS	1.000	0.876	1.000	0.876	52
		JCR	0.922	0.386	0.344	0.116	52

Table 10: The 5%, 25%, 50%, 75% and 95% quantiles of minimum model size to include all the important covariates in Example 5.

θ	n	Method	5%	25%	50%	75%	95%
0.5	200	JMDC-SIS	4	11	28	77	273
		RCDCS	7	19	44	99	286
		JCR	40	234	620	1191	1848
	300	JMDC-SIS	3	4	6	15	53
		RCDCS	3	6	11	28	91
		JCR	18	130	415	1037	1779
2	200	JMDC-SIS	4	12	28	77	288
		RCDCS	7	20	42	101	272
		JCR	43	226	643	1184	1823
	300	JMDC-SIS	3	4	6	15	53
		RCDCS	3	5	11	27	88
		JCR	19	128	418	1024	1782
8	200	JMDC-SIS	4	12	29	79	291
		RCDCS	7	20	44	100	286
		JCR	40	222	643	1208	1816
	300	JMDC-SIS	3	4	7	16	61
		RCDCS	3	5	11	27	94
		JCR	19	127	416	1033	1764

Table 11: $\mathcal{P}_k$, $\mathcal{P}_a$ and threshold value d_0 in Example 6.

θ	n	Method	$\rho = 0.6$						$\rho = 0.9$					
			$\mathcal{P}_1$	$\mathcal{P}_2$	$\mathcal{P}_3$	$\mathcal{P}_4$	$\mathcal{P}_a$	d_0	$\mathcal{P}_1$	$\mathcal{P}_2$	$\mathcal{P}_3$	$\mathcal{P}_4$	$\mathcal{P}_a$	d_0
0.5	200	aJMDC-SIS	0.960	0.870	0.756	0.356	0.324	102	0.998	0.990	0.974	0.900	0.898	81
		JMDC-SIS	0.892	0.764	0.646	0.212	0.178	37	0.988	0.972	0.954	0.822	0.816	37
		RCDCS	0.020	0.042	0.028	0.030	0.000	37	0.040	0.056	0.036	0.046	0.016	37
		JCR	0.488	0.512	0.558	0.204	0.108	37	0.596	0.646	0.694	0.596	0.466	37
	300	aJMDC-SIS	1.000	0.962	0.906	0.436	0.422	77	1.000	0.996	0.996	0.946	0.946	46
		JMDC-SIS	0.998	0.948	0.880	0.392	0.376	52	1.000	0.996	0.998	0.950	0.950	52
		RCDCS	0.026	0.072	0.048	0.038	0.000	52	0.028	0.068	0.084	0.040	0.010	52
		JCR	0.744	0.746	0.810	0.392	0.300	52	0.820	0.852	0.890	0.842	0.744	52
2	200	aJMDC-SIS	0.962	0.870	0.768	0.360	0.332	104	0.998	0.984	0.978	0.916	0.910	81
		JMDC-SIS	0.916	0.760	0.652	0.216	0.176	37	0.986	0.974	0.956	0.820	0.816	37
		RCDCS	0.016	0.050	0.030	0.028	0.002	37	0.036	0.052	0.032	0.050	0.016	37
		JCR	0.500	0.538	0.576	0.216	0.116	37	0.600	0.670	0.706	0.592	0.480	37
	300	aJMDC-SIS	1.000	0.970	0.912	0.438	0.434	75	1.000	0.998	0.996	0.968	0.968	46
		JMDC-SIS	1.000	0.952	0.888	0.406	0.394	52	1.000	0.998	0.996	0.954	0.954	52
		RCDCS	0.020	0.076	0.048	0.038	0.000	52	0.034	0.066	0.082	0.036	0.010	52
		JCR	0.766	0.758	0.812	0.388	0.296	52	0.820	0.862	0.900	0.838	0.744	52
8	200	aJMDC-SIS	0.976	0.876	0.776	0.350	0.322	101	0.994	0.990	0.982	0.910	0.908	80
		JMDC-SIS	0.920	0.784	0.654	0.242	0.214	37	0.988	0.976	0.966	0.814	0.810	37
		RCDCS	0.022	0.044	0.032	0.034	0.002	37	0.040	0.046	0.036	0.036	0.010	37
		JCR	0.506	0.532	0.572	0.216	0.110	37	0.602	0.658	0.702	0.604	0.486	37
	300	aJMDC-SIS	1.000	0.970	0.894	0.438	0.426	72	1.000	1.000	0.998	0.968	0.968	46
		JMDC-SIS	0.998	0.964	0.890	0.398	0.386	52	1.000	0.998	0.998	0.958	0.958	52
		RCDCS	0.026	0.084	0.054	0.038	0.000	52	0.046	0.068	0.078	0.032	0.012	52
		JCR	0.782	0.756	0.808	0.384	0.296	52	0.830	0.866	0.896	0.840	0.750	52

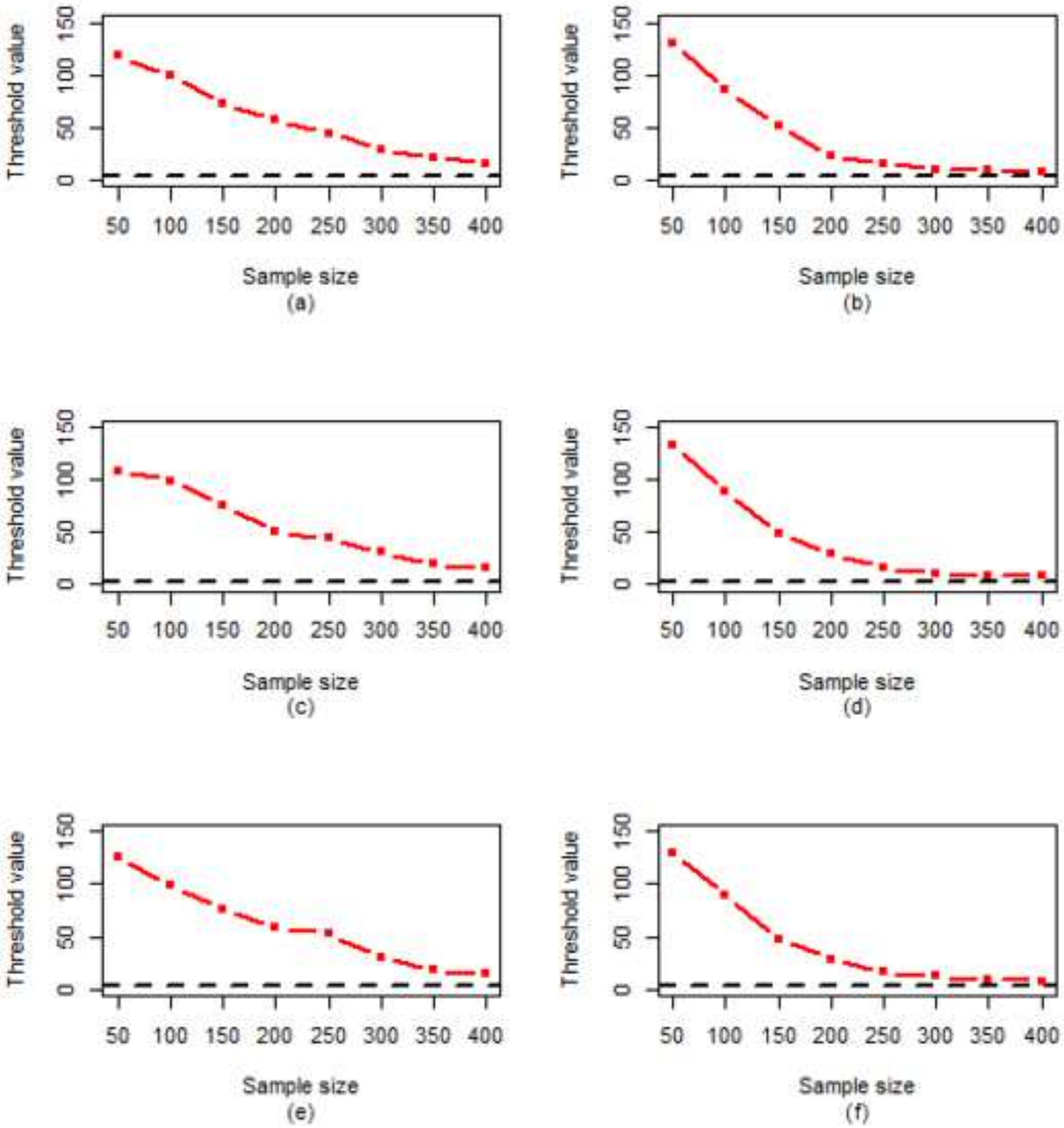
Table 12: The 5%, 25%, 50%, 75% and 95% quantiles of minimum model size to include all the important covariates in Example 6.

θ	n	Method	$\rho = 0.6$					$\rho = 0.9$				
			5%	25%	50%	75%	95%	5%	25%	50%	75%	95%
0.5	200	JMDC-SIS	7	63	249	866	1608	4	4	6	23	155
		RCDCS	578	1211	1562	1808	1966	152	735	1255	1645	1936
		JCR	15	115	452	1092	1804	4	9	49	273	1137
	300	JMDC-SIS	4	20	95	367	1170	4	4	4	6	48
		RCDCS	634	1177	1532	1770	1957	159	642	1136	1544	1916
		JCR	5	37	179	475	1550	4	5	12	55	468
2	200	JMDC-SIS	6	56	235	818	1654	4	4	6	21	144
		RCDCS	578	1256	1602	1830	1954	182	755	1296	1666	1937
		JCR	15	108	429	1037	1793	4	8	44	245	1098
	300	JMDC-SIS	4	19	95	352	1191	4	4	4	6	40
		RCDCS	638	1157	1522	1766	1957	164	655	1123	1502	1910
		JCR	5	35	166	456	1522	4	5	10	54	420
8	200	JMDC-SIS	7	50	229	828	1604	4	4	5	22	147
		RCDCS	579	1228	1603	1810	1969	183	710	1263	1667	1923
		JCR	14	105	419	1028	1817	4	8	41	234	1165
	300	JMDC-SIS	4	18	87	325	1104	4	4	4	6	42
		RCDCS	600	1120	1491	1795	1957	153	630	1127	1517	1910
		JCR	5	34	152	430	1531	4	5	10	52	434

Table 13: Names of genes selected by various approaches.

aJMDC-SIS	JMDC-SIS	RCDCS	JCR
NM_005480	NM_005480	Contig38288_RC	NM_001109
NM_003600	NM_003600	NM_005480	NM_001333
NM_003981	NM_003981	NM_007057	NM_018410
Contig38288_RC	Contig38288_RC	NM_003981	NM_000633
Contig31288_RC	Contig31288_RC	Contig48328_RC	NM_005628
Contig48328_RC	Contig48328_RC	NM_003600	Contig58368_RC
NM_003158	NM_003158	NM_003158	NM_001809
Contig46044_RC	Contig46044_RC	Contig31288_RC	NM_001605
NM_013277	NM_013277	NM_001605	NM_006607
NM_018410	NM_018410	NM_013438	NM_001168
NM_007057	NM_007057	NM_005733	NM_005480
D14678	D14678	NM_004805	NM_020142
NM_003258	NM_003258	Contig33814_RC	NM_005733
NM_001605	NM_001605	NM_018410	NM_004119
NM_004701	NM_004701	D14678	NM_003258
NM_007019	NM_007019	NM_014585	Contig51749_RC
NM_005733	NM_005733	AL117629	Contig56390_RC
Contig41652	Contig41652	NM_013277	NM_014863
NM_004336	NM_004336	NM_004336	AB007916
NM_004217	NM_004217	NM_006607	NM_013277
U74612	U74612	NM_001809	D14678
AB040926	AB040926	Contig51749_RC	NM_004217
NM_001168	NM_001168	NM_006845	NM_003600
NM_001809	NM_001809	NM_004701	Contig38288_RC
NM_004805	NM_004805	Contig34766_RC	NM_000909
	NM_006845	NM_000270	NM_003430
	NM_014585	Contig8818_RC	NM_013299
	NM_006607	U74612	NM_007184
	NM_001109	AB040926	D38553
	AL117629	NM_006819	NM_017702
	NM_014501	NM_003258	AL049265
	Contig45816_RC	NM_014501	NM_019013
	NM_006819	Contig44615_RC	AF007153
	Contig57584_RC	NM_001109	NM_005375
	NM_014176	Contig38726_RC	NM_000125
	NM_007274	NM_007019	Contig31288_RC
	NM_003686	AL137566	AL160131
	Contig8818_RC	NM_004217	NM_006027
	AB024704	Contig45816_RC	NM_007057
	NM_002466	NM_004456	Contig56843_RC
	NM_016359	Contig56843_RC	NM_003686
	D38553	Contig55069_RC	NM_003158
	NM_020974	Contig46044_RC	NM_018455
	NM_004219	Contig39061_RC	NM_005412
	Contig34766_RC	NM_020974	Contig57584_RC
	Contig33814_RC	Contig57584_RC	NM_006819
	AL117530	NM_001333	NM_007019
	NM_001333	NM_001255	NM_005005
	NM_006027	Contig41652	AK001166
	Contig51749_RC	NM_003686	NM_003981
	AL161983	NM_006082	NM_020974

Figure


Threshold value
0
50
100
150
50
100
150
200
250
300
350
400
Sample size
(a)
(b)
(c)
(d)
(e)
(f)