SID: a novel class of nonparametric tests of independence for censored outcomes

JINHONG LI[1,a], , JICAI LIU[2,b], and JINHONG YOU[3,c] and RIQUAN ZHANG[4,d]

[1] *School of Statistics, East China Normal University,* [a] *jinhongli950106@gmail.com*
[2] *School of Statistics and Mathematics, Shanghai Lixin University of Accounting and Finance,* [b] *liujicai1234@126.com*
[3] *School of Statistics and Management, Shanghai University of Finance and Economics,* [c] *johnyou07@163.com*
[4] *School of Statistics and Information, Shanghai University of International Business and Economics,* [d] *zhangriquan@163.com*

We propose a new class of metrics, called the survival independence divergence (SID), to test dependence between a right-censored outcome and covariates. A key technique for deriving the SIDs is to use a counting process strategy, which equivalently transforms the intractable independence test due to the presence of censoring into a test problem for complete observations. The SIDs are equal to zero if and only if the right-censored response and covariates are independent, and they are capable of detecting various types of nonlinear dependence. We propose empirical estimates of the SIDs and establish their asymptotic properties. We further develop a wild bootstrap method to estimate the critical values and show the consistency of the bootstrap tests. The numerical studies demonstrate that our SID-based tests are highly competitive with existing methods in a wide range of settings.

Keywords: Characteristic function; counting process; nonparametric independence test; reproducing kernel Hilbert space; survival analysis

1. Introduction

Let T be an event time and $\mathbf{X} \in \mathbb{R}^p$ be a p-dimensional vector of covariates. In this paper, we test the following hypotheses:

$$H_0 : T \text{ and } \mathbf{X} \text{ are independent} \qquad \text{versus} \qquad H_1 : \text{otherwise}. \tag{1}$$

Problem (1) is fundamental in statistics with broad applications. Among many applications, T is frequently subject to censoring. For instance, in clinical trials, patients' survival times to death are often right censored due to the termination of follow-up study or patients' drop out; in employment studies, the unemployment duration may be censored by the cutoff nature of the sampling. When censoring occurs, we only obtain partial information about the subjects' survival times, but do not know the exact time. Most existing methods for testing independence are not designed to deal with censoring, and thus we need to adapt them to censored time-to-event data.

There is a long history of measuring dependence for uncensored data. Pearson's correlation coefficient, Spearman's ρ, and Kendall's τ are probably the three most popular classical measures to quantify dependence between two univariate random variables. However, it is well known that these coefficients can detect only linear or monotone associations. Thus, many other flexible measures have been developed to overcome these difficulties, such as rank-based methods (Chatterjee, 2021, Weihs, Drton and Meinshausen, 2018), kernel-based methods (Gretton et al., 2008, Ke and Yin, 2020), and distance-based methods (Sejdinovic et al., 2013, Székely, Rizzo and Bakirov, 2007, Székely and Rizzo, 2013).

There have been lots of studies on testing indirectly the independence between T and $\mathbf{X}$ for censored data in survival analysis. The classical log-rank test is arguably the most popular approach for a two-sample test (Mantel, 1966) and is the most powerful test against local proportional hazards alternatives. However, the log-rank test cannot directly be used with general covariates, such as discrete or continuous covariates. Alternatively, one may test whether the regression coefficients corresponding to the covariates are equal to zero by fitting a semiparametric regression model, such as the proportional hazards models (Cox, 1972), the accelerated failure time models (Wei, 1992), and the transformation models (Cheng, Wei and Ying, 1995). Notably, a model-based test may also suffer a severe loss of power if the models are violated. Thus, it is urgent to develop model-free methods to problem (1).

Recently, a few nonparametric test methods have been proposed. For example, Edelmann, Welchowski and Benner (2021) developed the inverse probability of censoring-weighted (IPCW) scheme and generalized the distance covariance (Székely, Rizzo and Bakirov, 2007) to right-censored data. Fernández et al. (2021) extended the classic weighted log-rank tests through kernel methods and proposed a kernel log-rank test (KLR). The main challenge to test problem (1) for censored data is that the problem depends on the relationship between the censoring time C and the covariates $\mathbf{X}$. To specify the relationship, Edelmann, Welchowski and Benner (2021) assume that C is independent of $(\mathbf{X}, T)$. Generally, the assumption is strong and may be violated in many applications. In contrast, Fernández et al. (2021) adopted a mild and commonly used censoring mechanism that C is independent of T conditional on $\mathbf{X}$, called the so-called independent censoring scheme in the literature on survival analysis (Fleming and Harrington, 2011).

Denote censoring time as C. For the censoring meachansim, we assume that C is independent of T conditional on $\mathbf{X}$. This assumption combined with (1) is equivalent to testing the independence between T and $(C, \mathbf{X})$. This provides support for the condition commonly used in Stute (1993). ~~In this paper,~~Under this censoring meachansim, we propose survival independence divergence (SID) to test problem (1) using a counting process strategy. Specifically, with the counting process technique, we equivalently transform the original test problem (1) into a tractable test problem (5); see Theorem 3.1. We then propose two types of SIDs to test the equivalent problem. The first type is formulated by the discrepancy between two conditional characteristic functions, and the second is constructed using a Hilbert space distance between two distribution embeddings. We further build the connection between the two types of SIDs through the equivalence of distance-based and kernel-based statistics (Sejdinovic et al., 2013). The proposed SIDs have several appealing properties. They are equal to zero if and only if the right-censored response and covariates are independent, they do not require additional strong assumptions on the censoring mechanism, and they can detect various types of nonlinear dependence.

The idea of the SIDs is different from that of the KLR proposed by Fernández et al. (2021). The KLR can be viewed as a joint method, based on the difference between the joint distribution of T and $\mathbf{X}$, i.e., $\nu_1(t, \mathbf{x})$, and the product of their marginal distributions $\nu_0(t, \mathbf{x})$, see Theorem 3.1 of Fernández et al. (2021). In contrast, the SIDs belong to a class of conditional methods, based on the difference between a conditional distribution and a marginal distribution, see Theorem 3.1. In fact, the two approaches are commonly used to test independence between two variables for uncensored data in the classical statistics.

The rest of the paper is organized as follows. In Section 2, we discuss identifiability of problem (1). In Section 3, we introduce the SIDs and their properties. Sections 4 and 5 develop their sample counterparts and establish their theoretical properties. Section 6 proposes a wild bootstrap method to approximate the null distribution and shows its asymptotic validity. Numerical studies are conducted in Sections 7. We provide a brief discussion in Section 9.

2. Identifiability

In the section, we study identifiability of problem (1). ~~Let T be the continuous survival time and C the censoring time.~~ Denote the observed event time ~~by~~ $Y = \min\{T, C\}$ and the censoring indicator

~~by~~ $\delta = I(T \leq C)$, where $I(\cdot)$ is the indicator function. ~~Throughout the paper, we assume that C is independent of T conditional on $\mathbf{X}$.~~ Let $\tau_{T,\mathbf{x}} = \sup\{t : S_{T|\mathbf{X}=\mathbf{x}}(t) > 0\}$, $\tau_{C,\mathbf{x}} = \sup\{t : S_{C|\mathbf{X}=\mathbf{x}}(t) > 0\}$ and $\tau = \sup\{t : S_Y(t) > 0\}$, for $\mathbf{x} \in \mathbb{R}^p$.

Different from tests of independence for uncensored data, problem (1) may be unidentifiable due to the presence of censoring. For example, assume that the maximum supports of C and T satisfy $\sup\{t : S_C(t) > 0\} = c_0$ and $\sup\{t : S_T(t) > 0\} = c_0 + 1$, for some positive c_0. If $TI(0 \leq T \leq c_0)$ is independent of $\mathbf{X}$ but $TI(c_0 < T \leq c_0 + 1)$ is dependent on $\mathbf{X}$, problem (1) in the case is not identified because we only observe T before c_0. Thus, we need some identifiability constraints to overcome the issue.

We next present a strict definition to characterize the identifiability. Let $p_{Y,\delta|\mathbf{X}=\mathbf{x}}(t,\delta)$ be the conditional density function of (Y,δ) on $\mathbf{X} = \mathbf{x}$. Under the independent censoring scheme, we have

$$p_{Y,\delta|\mathbf{X}=\mathbf{x}}(t,\delta) = \lambda(t|\mathbf{x})^{\delta} S_{T|\mathbf{X}=\mathbf{x}}(t)[1 - S_{C|\mathbf{X}=\mathbf{x}}(t)] p_{C|\mathbf{X}=\mathbf{x}}(t)^{1-\delta},$$

where $\lambda(t|\mathbf{x})$ is the conditional hazard function of T, $S_{T|\mathbf{X}=\mathbf{x}}(t)$ and $S_{C|\mathbf{X}=\mathbf{x}}(t)$ are the conditional survival functions of T and C, and $p_{C|\mathbf{X}=\mathbf{x}}(t)$ is the conditional density of C. Note that T is independent of $\mathbf{X}$ if and only if $\lambda(t|\mathbf{x}) = \lambda(t)$, for almost all $\mathbf{x} \in \mathbb{R}^p$ and $t \geq 0$. Thus, we can use $\lambda(t|\mathbf{x})$ and $p_{Y,\delta|\mathbf{X}=\mathbf{x}}(t,\delta)$ to identify whether $\mathbf{X}$ is independent of T.

Definition 2.1. *Problem (1) is identifiable from $(Y,\delta,\mathbf{X})$ if $L(\lambda_1(t|\mathbf{x}); t, \mathbf{x}) = L(\lambda_2(t|\mathbf{x}); t, \mathbf{x})$ for almost all $(\mathbf{x}, t) \in \mathbb{R}^p \times \mathbb{R}^+$ implies $\lambda_1(t|\mathbf{x}) = \lambda_2(t|\mathbf{x})$, where*

$$L(\lambda(t|\mathbf{x}); t, \mathbf{x}) = \lambda(t|\mathbf{x})^{\delta} \exp\{-\int_0^t \lambda(s|\mathbf{x}) ds\}[1 - S_{C|\mathbf{X}=\mathbf{x}}(t)] p_{C|\mathbf{X}=\mathbf{x}}(t)^{1-\delta}.$$

By Definition 2.1, we provide two identifiable conditions in the following theorem.

Theorem 2.2. *Problem (1) is identifiable from the observed $(Y,\delta,\mathbf{X})$ if and only if $TI(T > \tau)$ is independent of $\mathbf{X}$.*

Throughout the paper, we refer to the condition in Theorem 2.2, where $TI(T > \tau)$ is independent of $\mathbf{X}$, as the identifiability condition. This condition encompasses Assumption 3.1 from Fernández et al. (2021), which states that if $S_{C|\mathbf{X}=\mathbf{x}}(t) = 0$, then $S_{T|\mathbf{X}=\mathbf{x}}(t) = 0$, thereby implying that $\tau_{T,\mathbf{x}} \leq \tau_{C,\mathbf{x}}$. In fact, when $\tau_{T,\mathbf{x}} \leq \tau_{C,\mathbf{x}}$, Proposition C.1 in Fernández et al. (2021), implies that $\tau_{Y,\mathbf{x}} = \tau_{T,\mathbf{x}} \leq \tau$ and thus $TI(T > \tau)$ is independent of $\mathbf{X}$.

3. Methodology

In the section, we equivalently transform problem (1) into a tractable problem by a counting process strategy. To the end, we define the failure counting process $N(t) = I(Y \leq t, \delta = 1)$ and the at-risk process $Y(t) = I(Y \geq t)$. Using the processes, we provide the following equivalent result.

Theorem 3.1. *Let $\Delta N(t) = N((t + \Delta t)^-) - N(t^-)$ be an increment of $N(\cdot)$ over $[t, t + \Delta t)$. Under the identifiability condition and the independent censoring scheme, $\mathbf{X}$ is independent of T if and only if, as $\Delta t \to 0^+$,*

$$\mathrm{P}\{\mathbf{X} \leq \mathbf{x} \mid \Delta N(t) = 1, Y(t) = 1\} = \mathrm{P}\{\mathbf{X} \leq \mathbf{x} \mid Y(t) = 1\}, \tag{2}$$

for almost all $\mathbf{x} \in \mathbb{R}^p$ *and* $t \in [0, \tau)$, *where the limit*

$$\lim_{\Delta t \to 0^+} \mathrm{P}\{\mathbf{X} \le \mathbf{x} \mid \Delta N(t) = 1, Y(t) = 1\} = \mathrm{P}\{\mathbf{X} \le \mathbf{x} \mid T = t, \delta = 1\}.$$

Note that the result in (2) is built upon the two observed processes $N(t)$ and $Y(t)$. Thus, problem (2) can be directly handled from the observed data, which effectively circumvents the difficulty caused by the censoring. We next provide some remarks on Theorem 3.1.

Remark 1. *If the identifiability condition is invalid,* (2) *is still useful, which tests equivalently the independence between* $\min\{T, \tau\}$ *and* $\mathbf{X}$. *In practice, we can use* (2) *to characterize the independence between the truncated time* $\min\{T, t_0\}$ *and* $\mathbf{X}$, *where* $t_0 (< \tau)$ *is a clinically meaningful endpoint.*

Remark 2. *To gain insights into* (2), *we consider the uncensored case that* $C = +\infty$. *In the special case, we immediately obtain that*

$$\mathrm{P}\{\mathbf{X} \le \mathbf{x} \mid \Delta N(t) = 1, Y(t) = 1\} = \mathrm{P}\{\mathbf{X} \le \mathbf{x} \mid t \le T < t + \Delta t\}, \tag{3}$$

$$\mathrm{P}\{\mathbf{X} \le \mathbf{x} \mid Y(t) = 1\} = \mathrm{P}\{\mathbf{X} \le \mathbf{x} \mid t \le T\}. \tag{4}$$

Note that $\{t \le T < t + \Delta t\}$ *and* $\{t \le T\}$ *are not identical as* $\Delta t \to 0^+$. *Thus,* (3) *and* (4) *are not equal, unless* $\mathbf{X}$ *is independent of* T.

For notational convenience, let ~~$P_t^{\mathbf{X}}$~~ $P_{\mathbf{X}|N_t,Y_t}$ and ~~$Q_t^{\mathbf{X}}$~~ $Q_{Y_t}^{\mathbf{X}}$ be the probability measures generated by $\mathrm{P}\{\mathbf{X} \le \mathbf{x} \mid \Delta N(t) = 1, Y(t) = 1\}$ and $\mathrm{P}\{\mathbf{X} \le \mathbf{x} \mid Y(t) = 1\}$ as $\Delta t \to 0^+$, for any fixed $t \in [0, \tau)$. Note that $P_{\mathbf{X}|N_t,Y_t}$ depends on t and Δt, and $Q_{Y_t}^{\mathbf{X}}$ depends on t. By Theorem 3.1, testing problem (1) is equivalent to testing

$$H_0 : P_{\mathbf{X}|N_t,Y_t} = Q_{Y_t}^{\mathbf{X}}, \text{ for any } t \in [0, \tau) \quad \text{versus} \quad H_1 : \text{otherwise}. \tag{5}$$

In the following sections, we develop two new classes of measures to quantify the discrepancy between $P_{\mathbf{X}|N_t,Y_t}$ and $Q_t^{\mathbf{X}}$. One is based on the characteristic function approach, and the other is based on the kernel embedding technique. Using the general framework proposed by Sejdinovic et al. (2013), we further build the connection between the two classes of measures.

3.1. Characteristic Function-based Approach

In this section, we introduce a class of measures based on the discrepancy between the characteristic functions of ~~$P_t^{\mathbf{X}}$~~ $P_{\mathbf{X}|N_t,Y_t}$ and ~~$Q_t^{\mathbf{X}}$~~ $Q_{Y_t}^{\mathbf{X}}$ to test problem (5). For simplicity, we define $\varphi_{\mathbf{X}|Y_t}(\mathbf{u}, t) = E\{\exp\{i\mathbf{u}^T\mathbf{X}\} \mid Y(t) = 1\}$, and $\varphi_{\mathbf{X}|N_t,Y_t}(\mathbf{u}, t) = E\{\exp\{i\mathbf{u}^T\mathbf{X}\} \mid \Delta N(t) = 1, Y(t) = 1\}$ for $\Delta t \to 0^+$, where $i = \sqrt{-1}$ is the imaginary unit. From the one-to-one relation between characteristic functions and distribution functions, we have ~~$P_t^{\mathbf{X}} = Q_t^{\mathbf{X}}$~~ $P_{\mathbf{X}|N_t,Y_t} = Q_{Y_t}^{\mathbf{X}}$ if and only if

$$\varphi_{\mathbf{X}|N_t,Y_t}(\mathbf{u}, t) = \varphi_{\mathbf{X}|Y_t}(\mathbf{u}, t), \text{ for all } \mathbf{u} \in \mathbb{R}^p \text{ and } t \in [0, \tau),$$

which is equivalent to testing whether

$$\int_0^\tau \int_{\mathbb{R}^p} \|\varphi_{\mathbf{X}|N_t,Y_t}(\mathbf{u}, t) - \varphi_{\mathbf{X}|Y_t}(\mathbf{u}, t)\|^2 dw_1(\mathbf{u}) dw_2(t) = 0, \tag{6}$$

where $\|f\|^2 = f\bar{f}$, $\bar{f}$ is the complex conjugate of f, $w_1(\mathbf{u})$ and $w_2(t)$ are two nonnegative weight functions.

It is important to choose the weight functions $w_1(\mathbf{u})$ and $w_2(t)$ in (6) properly. For $w_1(\mathbf{u})$, here we choose the weight function proposed by Székely, Rizzo and Bakirov (2007):

$$dw_1(\mathbf{u}) = \frac{1}{c(p,\beta)\|\mathbf{u}\|^{\beta+p}} d\mathbf{u} \text{ with } c(p,\beta) = \frac{2\pi^{p/2}\Gamma(1-\beta/2)}{\beta 2^{\beta}\Gamma((p+\beta)/2)}, \tag{7}$$

for any $\beta \in (0,2)$. For $w_2(t)$, we choose $dw_2(t) = a(t)d\,\mathrm{P}\{Y \le t, \delta = 1\}$, where $a(t)$ is a given nonnegative function with the support $[0,\tau]$. The choice of $a(t)$ is discussed in Section 4. Using the two weight functions, we define the following divergence.

Definition 3.2. *Let $\mathbf{X}$ be a p-dimensional random vector. The SID based on the characteristic functions between T and $\mathbf{X}$ is defined as*

$$\mathrm{SID}_{\beta}(T,\mathbf{X}) = \lim_{\Delta t \to 0^+} \int_0^{\tau} \int_{\mathbb{R}^p} \|\varphi_{\mathbf{X}|N_t,Y_t}(\mathbf{u},t) - \varphi_{\mathbf{X}|Y_t}(\mathbf{u},t)\|^2 a(t) dw_1(\mathbf{u}) d\nu(t), \tag{8}$$

where $\nu(t) = \mathrm{P}\{Y \le t, \delta = 1\}$, and $a(t)$ is a given nonnegative function.

An important difference between $\mathrm{SID}_{\beta}(T,\mathbf{X})$ and the distance covariance (Székely, Rizzo and Bakirov, 2007) is that the former is built on the conditional characteristic functions and the latter is based on the joint characteristic functions. From this perspective, $\mathrm{SID}_{\beta}(T,\mathbf{X})$ is like that in existing works, such as Wang et al. (2015), and Ke and Yin (2020).

Note that the integrals in (8) may suffer from computational issues due to their intractability. Fortunately, we can apply the following properties of $\mathrm{SID}_{\beta}(T,\mathbf{X})$ to overcome such issues.

Theorem 3.3. *Suppose that $(Y_1,\delta_1,\mathbf{X}_1)$ and $(Y_2,\delta_2,\mathbf{X}_2)$ are independently and identically distributed (i.i.d.) copies of $(Y,\delta,\mathbf{X})$. If $E\{\|\mathbf{X}\|^{\beta}\} < \infty$ for $\beta \in (0,2)$, then we have that*

(i) $\mathrm{SID}_{\beta}(T,\mathbf{X})$ *can be expressed as:*

$$\begin{aligned}\mathrm{SID}_{\beta}(T,\mathbf{X}) = &- \int_0^{\tau} E\{\|\mathbf{X}_1 - \mathbf{X}_2\|^{\beta} \mid Y_1 = t, \delta_1 = 1, Y_2 = t, \delta_2 = 1\} a(t) d\nu(t) \\ &+ 2\int_0^{\tau} E\{\|\mathbf{X}_1 - \mathbf{X}_2\|^{\beta} \mid Y_1 = t, \delta_1 = 1, Y_2(t) = 1\} a(t) d\nu(t) \\ &- \int_0^{\tau} E\{\|\mathbf{X}_1 - \mathbf{X}_2\|^{\beta} \mid Y_1(t) = 1, Y_2(t) = 1\} a(t) d\nu(t).\end{aligned} \tag{9}$$

(ii) $\mathrm{SID}_{\beta}(T,\mathbf{X}) \ge 0$ *and* $\mathrm{SID}_{\beta}(T,\mathbf{X}) = 0$ *if and only if T and $\mathbf{X}$ are independent .*

Theorem 3.3(i) suggests that $\mathrm{SID}_{\beta}(T,\mathbf{X})$ has a closed form and is, thus, easily estimated from the data. Theorem 3.3(ii) implies that $\mathrm{SID}_{\beta}(T,\mathbf{X})$ is an applicable measure for testing independence between T and $\mathbf{X}$. Note that Theorem 3.3(ii) holds only for $0 < \beta < 2$ but may be not true for $\beta = 2$. In fact, for $\beta = 2$, by some algebra, we can easily show that (9) is equal to 0 if and only if $E\{\mathbf{X} \mid \Delta N(t) = 1, Y(t) = 1\} = E\{\mathbf{X} \mid Y(t) = 1\}$ as $\Delta t \to 0^+$, which does not imply $P_{\mathbf{X}|N_t,Y_t} = Q_{\mathbf{X}|Y_t}$.

3.2. Kernel-based Approach

In recent decades, a Hilbert space embedding of a distribution has emerged as a useful tool for many nonparametric hypothesis test problems. For example, it has been developed and successfully used for independence (Gretton et al., 2008), two-sample (Gretton et al., 2012), and goodness-of-fit (Chwialkowski, Strathmann and Gretton, 2016) testing frameworks. The effectiveness of such embeddings prompts us to generalize the embedding method to test problem (5).

We first briefly review the reproducing kernel Hilbert space (RKHS). A RKHS $\mathcal{H}_K$ on $\mathcal{X}$ with a kernel $K(\mathbf{x},\mathbf{x}')$ is a Hilbert space of functions $f(\cdot):\mathcal{X}\mapsto\mathbb{R}$ with inner product $\langle\cdot,\cdot\rangle_{\mathcal{H}_K}$. The kernel $K(\mathbf{x},\mathbf{x}')$ is required to satisfy: (1) $K(\cdot,\mathbf{x})\in\mathcal{H}_K$ for all $\mathbf{x}\in\mathcal{X}$ and (2) $\langle f,K(\cdot,\mathbf{x})\rangle_{\mathcal{H}_K}=f(\mathbf{x})$ for all $\mathbf{x}\in\mathcal{X}$ and $f\in\mathcal{H}_K$. We call the mapping $\phi:\mathbf{x}\to\mathcal{H}_K$ given by $\phi(\mathbf{x})=K(\cdot,\mathbf{x})$ the canonical feature map of $K(\mathbf{x},\mathbf{x}')$.

Let $\mathcal{M}(\mathcal{X})$ be the set of all finite signed Borel measures on $\mathcal{X}$ and $\mathcal{M}_+^1(\mathcal{X})$ the set of all Borel probability measures on $\mathcal{X}$. The basic idea behind the embedding of probability measures is to map measures into a RKHS. Specifically, for $\nu\in\mathcal{M}(\mathcal{X})$, the kernel embedding of v into the RKHS $\mathcal{H}_K$ is $\mu_K(\nu)\in\mathcal{H}_K$ such that $\int f(\mathbf{x})d\nu(\mathbf{x})=\langle f,\mu_K(\nu)\rangle_{\mathcal{H}_K}$, for all $f\in\mathcal{H}_K$. For more details, refer to Sejdinovic et al. (2013).

For a measurable kernel $K(\cdot,\cdot)$ on $\mathcal{X}$ and $\theta>0$, we need to restrict the kernel into the following particular class of measures: $\mathcal{M}_K^{\theta}(\mathcal{X})=\{v\in\mathcal{M}(\mathcal{X}):\int K^{\theta}(\mathbf{x},\mathbf{x})d|v|(\mathbf{x})<\infty\}$. We can see that each element in $\mathcal{M}_K^{\theta}(\mathcal{X})$ is required to have a finite θ-moment with respect to the kernel $K(\cdot,\cdot)$. Using the Riesz representation theorem, it can be shown that $\mu_K(\nu)$ exists if $\nu\in\mathcal{M}_K^{1/2}(\mathcal{X})$; see Lemma 3 in Gretton et al. (2012). Thus, we can impose the assumption that $P_{\mathbf{X}|N_t,Y_t}$, $Q_{\mathbf{X}|Y_t}\in\mathcal{M}_+^1(\mathbb{R}^p)\cap\mathcal{M}_K^{1/2}(\mathbb{R}^p)$ to ensure that the kernel embeddings of $P_{\mathbf{X}|N_t,Y_t}$ and $Q_{\mathbf{X}|Y_t}$ exist. From this assumption, we introduce the following Hilbert space distance.

Definition 3.4. *Let $K(\cdot,\cdot)$ be a measurable kernel on $\mathbb{R}^p$. Assuming $P_{\mathbf{X}|N_t,Y_t}$ and $Q_{\mathbf{X}|Y_t}\in\mathcal{M}_+^1(\mathbb{R}^p)\cap\mathcal{M}_K^{1/2}(\mathbb{R}^p)$, the SID based on $K(\cdot,\cdot)$ is defined as*

$$\mathrm{SID}_K(T,\mathbf{X})=\lim_{\Delta t\to0+}\int_0^{\tau}\|\mu_K(P_t^{\mathbf{X}})-\mu_K(Q_t^{\mathbf{X}})\|_{\mathcal{H}_K}^2a(t)d\nu(t),$$

where $\nu(t)=\mathrm{P}\{Y\le t,\delta=1\}$, and $a(t)$ is a given nonnegative function.

To find the properties of $\mathrm{SID}_K(T,\mathbf{X})$, the kernel has to be a characteristic function. A kernel function is said to be characteristic if the map $\mu_K:\nu\to\mu_K(\nu)$ is injective. The assumption is commonly used in the literature on kernel methods, for example, Gretton et al. (2008, 2012).

Theorem 3.5. *Assume that $P_{\mathbf{X}|N_t,Y_t}$ and $Q_{\mathbf{X}|Y_t}\in\mathcal{M}_+^1(\mathbb{R}^p)\bigcap\mathcal{M}_K^{1/2}(\mathbb{R}^p)$. Then, we have that*

(i) $\mathrm{SID}_K(T,\mathbf{X})$ *can be expressed as:*

$$\begin{aligned}\mathrm{SID}_K(T,\mathbf{X})=&\int_0^{\tau}E\{K(\mathbf{X}_1,\mathbf{X}_2)\mid Y_1=t,\delta_1=1,Y_2=t,\delta_2=1\}a(t)d\nu(t)\\&-2\int_0^{\tau}E\{K(\mathbf{X}_1,\mathbf{X}_2)\mid Y_1=t,\delta_1=1,Y_2(t)=1\}a(t)d\nu(t)\qquad(10)\\&+\int_0^{\tau}E\{K(\mathbf{X}_1,\mathbf{X}_2)\mid Y_1(t)=1,Y_2(t)=1\}a(t)d\nu(t).\end{aligned}$$

(ii) $\mathrm{SID}_K(T,\mathbf{X}) \geq 0$. *If* $K(\cdot,\cdot)$ *is characteristic,* $\mathrm{SID}_K(T,\mathbf{X}) = 0$ *if and only if* T *and* $\mathbf{X}$ *are independent.*

3.3. Connection between $\mathrm{SID}_\beta(T, \mathrm{X})$ and $\mathrm{SID}_K(T, \mathrm{X})$

In this section, we first show that $\mathrm{SID}_\beta(T,\mathbf{X})$ is a special class of negative type semimetric-based measures (Sejdinovic et al., 2013). Through the link between negative type semimetrics and kernels, we then build up the connection between $\mathrm{SID}_\beta(T,\mathbf{X})$ and $\mathrm{SID}_K(T,\mathbf{X})$.

First, we recap a negative-type semimetric and space. Let $\mathcal{X}$ be an arbitrary space endowed with a semimetric ~~$\rho : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$~~ $\rho : \mathcal{X} \times \mathcal{X} \to [0,+\infty)$, satisfying: $\sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j \rho(\mathbf{x}_i, \mathbf{x}_j) \leq 0$, where $\mathbf{x}_i \in \mathcal{X}$ and $\alpha_i \in \mathbb{R}$ with $\sum_{i=1}^n \alpha_i = 0$. Then, ρ is called a semimetric of negative type on $\mathcal{X}$, and $(\mathcal{X}, \rho)$ is called a space of negative type. In addition, like the finite moment condition on kernels, we also need to restrict ρ to a particular class of measures. Specifically, we define a new class of Borel measures: $\mathcal{M}_\rho^\theta(\mathcal{X}) = \{v \in \mathcal{M}(\mathcal{X}) : \exists \mathbf{x}_0 \in \mathcal{X} \text{ s.t. } \int \rho^\theta(\mathbf{x}, \mathbf{x}_0) d|v|(\mathbf{x}) < \infty\}$, for a given negative-type semimetric ρ and $\theta > 0$. Then, we introduce the following divergence.

Definition 3.6. *Let* $(\mathbb{R}^p, \rho)$ *be a semimetric space of negative type. Assuming*

$$P_{\mathbf{X}|N_t,Y_t}, Q_{\mathbf{X}|Y_t} \in \mathcal{M}_+^1(\mathbb{R}^p) \bigcap \mathcal{M}_\rho^1(\mathbb{R}^p),$$

the SID with the negative-type semimetric ρ *is defined as*

$$\mathrm{SID}_\rho(T,\mathbf{X}) = \lim_{\Delta t \to 0+} - \int_0^\tau \int_{\mathbb{R}^p \times \mathbb{R}^p} \rho(\mathbf{X},\mathbf{X}') d([P_{\mathbf{X}|N_t,Y_t} - Q_{\mathbf{X}|Y_t}] \times [P_{\mathbf{X}'|N_t,Y_t} - Q_{\mathbf{X}'|Y_t}]) a(t) d\nu(t),$$

where $\nu(t) = \mathrm{P}\{Y \leq t, \delta = 1\}$, *and* $a(t)$ *is a given nonnegative function.*

Like Theorems 3.3 and 3.5, the following theorem gives a closed-form expression for $\mathrm{SID}_\rho(T,\mathbf{X})$.

Theorem 3.7. *Assume that* $(\mathbb{R}^p, \rho)$ *is a semimetric space of negative type. For*

$$P_{\mathbf{X}|N_t,Y_t} \quad \textit{and} \quad Q_{\mathbf{X}|Y_t} \in \mathcal{M}_+^1(\mathbb{R}^p) \bigcap \mathcal{M}_\rho^1(\mathbb{R}^p),$$

we have that

$$\begin{aligned} \mathrm{SID}_\rho(T,\mathbf{X}) = & - \int_0^\tau E\{\rho(\mathbf{X}_1,\mathbf{X}_2) \mid Y_1 = t, \delta_1 = 1, Y_2 = t, \delta_2 = 1\} a(t) d\nu(t) \\ & + 2 \int_0^\tau E\{\rho(\mathbf{X}_1,\mathbf{X}_2) \mid Y_1 = t, \delta_1 = 1, Y_2(t) = 1\} a(t) d\nu(t) \\ & - \int_0^\tau E\{\rho(\mathbf{X}_1,\mathbf{X}_2) \mid Y_1(t) = 1, Y_2(t) = 1\} a(t) d\nu(t). \end{aligned} \tag{11}$$

By Proposition 3 in Sejdinovic et al. (2013), we obtain that $\rho_\beta(\mathbf{x},\mathbf{x}') = \|\mathbf{x} - \mathbf{x}'\|^\beta$ is a semimetric of negative type for $\beta \in (0,2)$. Thus, we can see that $\mathrm{SID}_\beta(T,\mathbf{X})$ belongs to the family of negative-type semimetric-based measures. We next develop the connection between $\mathrm{SID}_\beta(T,\mathbf{X})$ and $\mathrm{SID}_K(T,\mathbf{X})$ using the following link between negative-type semimetrics and kernels. Specifically, Proposition 3 in

Sejdinovic et al. (2013) suggests that, for any negative-type semimetric ρ, there exists a nondegenerate kernel $K(\cdot,\cdot)$ such that

$$\rho(\mathbf{x},\mathbf{x}') = K(\mathbf{x},\mathbf{x}) + K(\mathbf{x}',\mathbf{x}') - 2K(\mathbf{x},\mathbf{x}') = \|K(\cdot,\mathbf{x}) - K(\cdot,\mathbf{x}')\|^2_{\mathcal{H}_K}. \tag{12}$$

Conversely, if $K(\cdot,\cdot)$ is any nondegenerate kernel, then $\rho(\cdot,\cdot)$ defined by (12) is a valid semimetric of negative type. Whenever the kernel $K(\cdot,\cdot)$ and the semimetric $\rho(\cdot,\cdot)$ satisfy (12), we say that $K(\cdot,\cdot)$ generates $\rho(\cdot,\cdot)$. From (10), (11), and (12), we have the following result.

Theorem 3.8. *Let $(\mathbb{R}^p,\rho)$ be a semimetric space of negative type and $K(\cdot,\cdot)$ any kernel that generates ρ. Assuming $P_{\mathbf{X}|N_t,Y_t}$ and $Q_{\mathbf{X}|Y_t} \in \mathcal{M}^1_+(\mathbb{R}^p)\bigcap\mathcal{M}^1_\rho(\mathbb{R}^p)$, we have that* $\mathrm{SID}_\rho(T,\mathbf{X}) = 2\,\mathrm{SID}_K(T,\mathbf{X})$.

Let $K_\beta(\mathbf{x},\mathbf{x}') = (\|\mathbf{x}\|^\beta + \|\mathbf{x}'\|^\beta - \|\mathbf{x}-\mathbf{x}'\|^\beta)/2$, for $\beta\in(0,2)$. From (12), $\rho_\beta(\mathbf{x},\mathbf{x}') = \|\mathbf{x}-\mathbf{x}'\|^\beta$ is the semimetric generated by $K_\beta(\mathbf{x},\mathbf{x}')$. By Theorem 3.8, we obtain that $\mathrm{SID}_\beta(T,\mathbf{X}) = \mathrm{SID}_{\rho_\beta}(T,\mathbf{X}) = 2\,\mathrm{SID}_{K_\beta}(T,\mathbf{X})$. Thus, $\mathrm{SID}_\beta(T,\mathbf{X})$ is a special case of $\mathrm{SID}_K(T,\mathbf{X})$.

4. Estimation Approaches

In this section, we introduce the empirical estimators of $\mathrm{SID}_\beta(T,\mathbf{X})$ and $\mathrm{SID}_K(T,\mathbf{X})$. We provide details of the derivation of only the estimator of $\mathrm{SID}_K(T,\mathbf{X})$, since the estimator of $\mathrm{SID}_\beta(T,\mathbf{X})$ can be similarly obtained. Two kinds of estimates are proposed. The first is obtained by plugging the empirical functions into (10), and the second is based on a U-statistic.

The closed form in (10) consists of the following three conditional expectations:

$$\begin{aligned}
S_{K,1}(t) &= E\{K(\mathbf{X}_1,\mathbf{X}_2)\mid Y_1=t,\delta_1=1,Y_2=t,\delta_2=1\},\\
S_{K,2}(t) &= E\{K(\mathbf{X}_1,\mathbf{X}_2)\mid Y_1=t,\delta_1=1,Y_2(t)=1\},\\
S_{K,3}(t) &= E\{K(\mathbf{X}_1,\mathbf{X}_2)\mid Y_1(t)=1,Y_2(t)=1\}.
\end{aligned}$$

To estimate $S_{K,1}(t)$ and $S_{K,2}(t)$, we use the Nadaraya–Watson kernel-smoothing methods. Note that the estimates of $S_{K,1}(t)$ and $S_{K,2}(t)$ do not suffer from the so-called curse of dimensionality, because the estimation depends on only the 1-dimensional variable t.

Suppose $\{(Y_i,\delta_i,\mathbf{X}_i), i=1,\ldots,n\}$ are independent random samples drawn from $(Y,\delta,\mathbf{X})$. Let $W(y)$ be a symmetric density function and $W_h(y) = h^{-1}W(y/h)$, where $h>0$ is a bandwidth. In our numerical studies, a Gaussian kernel is used for $W(y)$. The Nadaraya–Watson kernel estimators of $S_{K,1}(t)$ and $S_{K,2}(t)$ are given by

$$\begin{aligned}
\widehat{S}_{K,1}(t) &= \frac{\sum_{i,j=1}^n K(\mathbf{X}_i,\mathbf{X}_j)W_h(Y_i-t)W_h(Y_j-t)\delta_i\delta_j}{\left[\sum_{i=1}^n W_h(Y_i-t)\delta_i\right]^2},\\
\widehat{S}_{K,2}(t) &= \frac{\sum_{i,j=1}^n K(\mathbf{X}_i,\mathbf{X}_j)W_h(Y_i-t)\delta_i I(Y_j\ge t)}{\left[\sum_{i=1}^n W_h(Y_i-t)\delta_i \sum_{j=1}^n I(Y_j\ge t)\right]}.
\end{aligned}$$

And, $S_{K,3}(t)$ can be estimated by

$$\widehat{S}_{K,3}(t) = \frac{\sum_{i,j=1}^n K(\mathbf{X}_i,\mathbf{X}_j)I(Y_i\ge t)I(Y_j\ge t)}{\left[\sum_{i=1}^n I(Y_i\ge t)\right]^2}.$$

Note that the above estimators $\widehat{S}_{K,j}(t)$, $j=1,2,3$, suffer from random denominator issues, which may lead to a large bias near 0. As suggested by Su and White (2007), we can choose a proper weight function $a(\cdot)$ in (10) to overcome such issues. By the definitions of $\widehat{S}_{K,j}(t)$, we choose $a(\cdot)$ as

$$a^*(t) = [f_{Y,\delta}(t,1)E\{I(Y>t)\}]^2, \tag{1}$$

where $f_{Y,\delta}(y,1)$ is the density function of $\mathrm{P}\{Y \le y, \delta = 1\}$. Throughout this paper, we use $a(t) = a^*(t)$ in the definitions of $\mathrm{SID}_\beta(T,\mathbf{X})$, $\mathrm{SID}_K(T,\mathbf{X})$, and $\mathrm{SID}_\rho(T,\mathbf{X})$.

Using the weight function $a^*(t)$, we provide a plug-in estimator:

$$\widetilde{\mathrm{SID}}_K(T,\mathbf{X}) = \int_0^\tau \Big[\widehat{S}_{K,1}(t) - 2\widehat{S}_{K,2}(t) + \widehat{S}_{K,3}(t)\Big]\widehat{a}^*(t)d\overline{N}(t),$$

where $\overline{N}(t) = \sum_{i=1}^n I(Y_i \le t, \delta_i = 1)/n$ and $\widehat{a}^*(t) = [n^{-2}\sum_{i=1}^n W_h(Y_i - t)\delta_i \sum_{j=1}^n I(Y_j \ge t)]^2$. Furthermore, we can easily obtain that

$$\widetilde{\mathrm{SID}}_K(T,\mathbf{X}) = \frac{1}{n}\sum_{k=1}^n \Big[\widetilde{S}_{K,1}(Y_k) - 2\widetilde{S}_{K,2}(Y_k) + \widetilde{S}_{K,3}(Y_k)\Big]\delta_k, \tag{2}$$

where

$$\widetilde{S}_{K,1}(t) = \frac{1}{n^4}\sum_{i,j=1}^n K(\mathbf{X}_i,\mathbf{X}_j)W_h(Y_i - t)\delta_i W_h(Y_j - t)\delta_j \sum_{i,j=1}^n I(Y_i \ge t)I(Y_j \ge t),$$

$$\widetilde{S}_{K,2}(t) = \frac{1}{n^4}\sum_{i,j=1}^n K(\mathbf{X}_i,\mathbf{X}_j)W_h(Y_i - t)\delta_i I(Y_j \ge t) \sum_{i,j=1}^n W_h(Y_i - t)\delta_i I(Y_j \ge t),$$

$$\widetilde{S}_{K,3}(t) = \frac{1}{n^4}\sum_{i,j=1}^n K(\mathbf{X}_i,\mathbf{X}_j)I(Y_i \ge t)I(Y_j \ge t) \sum_{i,j=1}^n W_h(Y_i - t)\delta_i W_h(Y_j - t)\delta_j.$$

After some algebra, the above plug-in estimator satisfies:

$$\widetilde{\mathrm{SID}}_K(T,\mathbf{X}) = \frac{1}{n^5}\sum_{i,j,k,l,r=1}^n [b_{ikr} - b_{kir}]K_{ij}[b_{jlr} - b_{ljr}]\delta_r, \tag{3}$$

where $K_{ij} = K(\mathbf{X}_i,\mathbf{X}_j)$ and $b_{ijk} = \delta_i W_h(Y_i - Y_k)I(Y_j \ge Y_k)$. From (3), we can see that the plug-in estimator $\widetilde{\mathrm{SID}}_K(T,\mathbf{X})$ is essentially a V-statistic. Then, its corresponding U-statistic-based estimator can be obtained as

$$\widehat{\mathrm{SID}}_K(T,\mathbf{X}) = \frac{1}{(n)_5}\sum_{i \ne j \ne k \ne l \ne r} [b_{ikr} - b_{kir}]K_{ij}[b_{jlr} - b_{ljr}]\delta_r, \tag{4}$$

where $(n)_5 = n(n-1)(n-2)(n-3)(n-4)$.

With the same arguments used for (2) and (4), the plug-in and U-statistic estimators for $\mathrm{SID}_\beta(T,\mathbf{X})$ are given by:

$$\widetilde{\mathrm{SID}}_\beta(T,\mathbf{X}) = \frac{1}{n}\sum_{k=1}^n \Big[-\widetilde{S}_{\beta,1}(Y_k) + 2\widetilde{S}_{\beta,2}(Y_k) - \widetilde{S}_{\beta,3}(Y_k)\Big]\delta_k,$$

$$\widehat{\mathrm{SID}}_{\beta}(T,\mathbf{X}) = \frac{1}{(n)_5} \sum_{i\neq j\neq k\neq l\neq r} \|\mathbf{X}_i - \mathbf{X}_j\|^{\beta}[b_{ikr} - b_{kir}][b_{jlr} - b_{ljr}]\delta_r,$$

where $\widetilde{S}_{\beta,1}(t)$, $\widetilde{S}_{\beta,2}(t)$, and $\widetilde{S}_{\beta,3}(t)$ are similarly defined as $\widetilde{S}_{K,1}(t)$, $\widetilde{S}_{K,2}(t)$, and $\widetilde{S}_{K,3}(t)$ by replacing $K(\mathbf{X}_i, \mathbf{X}_j)$ with $\|\mathbf{X}_i - \mathbf{X}_j\|^{\beta}$ for $\beta \in (0, 2)$.

5. Asymptotic Properties

In this section, we first show that the above empirical estimators are consistent with their population counterparts. Thus, we need the following conditions:

(C1) Let ~~$f_{Y,\delta}(y,1)$ and $f_{\mathbf{X},Y,\delta}(x,y,1)$~~ $f_{Y,\delta}(y,\delta)$ and $f_{\mathbf{X},Y,\delta}(x,y,\delta)$ be the joint density functions of ~~$(Y, \delta = 1)$ and $(\mathbf{X}, Y, \delta = 1)$~~ (Y, δ) and $(\mathbf{X}, Y, \delta)$. Assume that ~~$f_{Y,\delta}(y,1)$ and $f_{\mathbf{X},Y,\delta}(x,y,1)$~~ $f_{Y,\delta}(y,1)$ and $f_{\mathbf{X},Y,\delta}(x,y,1)$ have bounded, continuous, twice-order derivatives with respective to y.

(C2) The kernel function $W(u)$ is a bounded and symmetric density function with a bounded derivative and satisfies $\int_{-\infty}^{\infty} |u|^j W(u) dW(u) < \infty, j = 1, 2, \ldots$

(C3) $h \to 0$ and $\sqrt{nh}/\ln(1/h) \to \infty$ as $n \to \infty$.

Conditions (C1) and (C2) are standard for kernel estimation, as they ensure the consistency of the kernel estimators. Condition (C1) imposes some smoothness conditions on the joint density functions. Many kernel functions satisfy condition (C2), for example, the standard Gaussian kernel $W(u) = 1/\sqrt{2\pi}\exp(-u^2/2)$ and the Epanechnikov kernel $W(u) = 3/4(1-u^2)I(|u| \leq 1)$. Condition (C3) requires the bandwidth to be chosen appropriately according to n. It is well known that the choice of the bandwidth in kernel-smoothing methods is a challenging and unsolved problem. In our numerical studies, we choose the optimal bandwidth (Silverman, 1986), given by $h = \{4/3\}^{-1/5}\hat{\sigma}_y n^{-1/5}$, where $\hat{\sigma}_y$ is the sample standard deviation of $\{y_1, \ldots, y_n\}$.

Theorem 5.1. *Assume that conditions (C1)–(C3) hold. Then, we have, as $n \to \infty$,*

$$\widetilde{\mathrm{SID}}_{\beta}(T,\mathbf{X}) \xrightarrow{P} \mathrm{SID}_{\beta}(T,\mathbf{X}), \qquad \widehat{\mathrm{SID}}_{\beta}(T,\mathbf{X}) \xrightarrow{P} \mathrm{SID}_{\beta}(T,\mathbf{X}),$$
$$\widetilde{\mathrm{SID}}_{K}(T,\mathbf{X}) \xrightarrow{P} \mathrm{SID}_{K}(T,\mathbf{X}), \qquad \widehat{\mathrm{SID}}_{K}(T,\mathbf{X}) \xrightarrow{P} \mathrm{SID}_{K}(T,\mathbf{X}).$$

We next develop the asymptotic distribution of $\widehat{\mathrm{SID}}_K(T,\mathbf{X})$ under the null hypothesis. In the proof of Theorem 5.1, we constructed a 5-order U-statistic by symmetrizing $\widehat{\mathrm{SID}}_K(T,\mathbf{X})$. When establishing the asymptotic null distribution, we can simplify the theoretical analysis and approximate $\widehat{\mathrm{SID}}_K(T,\mathbf{X})$ with the following 4-order U-statistic:

$$\widehat{\mathrm{SID}}^{\nu}_{K}(T,\mathbf{X}) = \binom{n}{4}^{-1} \sum_{i<j<k<l} \int_0^{\tau} h_n(\mathbf{Z}_i, \mathbf{Z}_j, \mathbf{Z}_k, \mathbf{Z}_l; t)d\nu(t), \tag{1}$$

where $\mathbf{Z}_i = (Y_i, \delta_i, \mathbf{X}_i)$ and

$$h_n(\mathbf{Z}_i, \mathbf{Z}_j, \mathbf{Z}_k, \mathbf{Z}_l; t) = \frac{1}{4!}\sum_{\pi} P_n(\mathbf{Z}_{i'}, \mathbf{Z}_{j'}, \mathbf{Z}_{k'}, \mathbf{Z}_{l'}; t), \tag{2}$$

with $P_n(\mathbf{Z}_i, \mathbf{Z}_j, \mathbf{Z}_k, \mathbf{Z}_l; t) = [b_{ik}(t) - b_{ki}(t)]K_{ij}[b_{jl}(t) - b_{lj}(t)]$ and $b_{ik}(t) = \delta_i W_h(Y_i - t)I(Y_k \geq t)$. Here, $\sum_\pi$ denotes summation over all the 4! permutations (i', j', k', l') of (i, j, k, l). Some calculations yield that

$$\begin{aligned} h_n(\mathbf{Z}_i,\mathbf{Z}_j,\mathbf{Z}_k,\mathbf{Z}_l;t) = \frac{1}{12}\Big\{ & [K_{ij} - K_{il} - K_{jk} + K_{kl}][b_{ik}(t) - b_{ki}(t)][b_{jl}(t) - b_{lj}(t)] \\ + \; & [K_{ij} - K_{jl} - K_{ik} + K_{kl}][b_{il}(t) - b_{li}(t)][b_{jk}(t) - b_{kj}(t)] \qquad (3) \\ + \; & [K_{ik} - K_{il} - K_{jk} + K_{jl}][b_{ij}(t) - b_{ji}(t)][b_{kl}(t) - b_{lk}(t)]\Big\}. \end{aligned}$$

Let $h_{nc}(z_1, \ldots, z_c) = E\{h_n(\mathbf{Z}_1, \mathbf{Z}_2, \mathbf{Z}_3, \mathbf{Z}_4) | \mathbf{Z}_1 = z_1, \ldots, \mathbf{Z}_c = z_c\}$ be the c-order projection of $h_n(\mathbf{Z}_1, \mathbf{Z}_2, \mathbf{Z}_3, \mathbf{Z}_4)$, for $c = 1, 2, 3, 4$. The following lemma provides closed-form expressions of the first- and second-order projections under H_0.

Lemma 5.2. *Let $\mathbf{z} = (y, \tilde{\delta}, \mathbf{x})$ and $\mathbf{z}' = (y', \tilde{\delta}', \mathbf{x}')$, for any $y, y' \in \mathbb{R}^+$, $\tilde{\delta}, \tilde{\delta}' \in \{0, 1\}$, and $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^p$. Under H_0, we have*

$$h_{n1}(\mathbf{z}; t) = 0; \quad h_{n2}(\mathbf{z}, \mathbf{z}'; t) = \frac{1}{6}U(\mathbf{x}, \mathbf{x}'; t)V(y, \tilde{\delta}; y', \tilde{\delta}'; t),$$

where

$$\begin{aligned} U(\mathbf{x}, \mathbf{x}'; t) &= K(\mathbf{x}, \mathbf{x}') - E\{K(\mathbf{x}, \mathbf{X}_1) \mid Y_1(t) = 1\} - \; E\{K(\mathbf{x}', \mathbf{X}_1) \mid Y_1(t) = 1\} \\ &\quad + \; E\{K(\mathbf{X}_1, \mathbf{X}_2) \mid Y_1(t) = 1, Y_2(t) = 1\}, \\ V(y, \tilde{\delta}; y', \tilde{\delta}'; t) &= [\tilde{\delta} W_h(y - t)S(t) - I(y \geq t)F_h(t)][\tilde{\delta}' W_h(y' - t)S(t) - I(y' \geq t)F_h(t)], \end{aligned}$$

with $S(t) = P\{Y \geq t\}$ and $F_h(t) = E\{\delta W_h(Y - t)\}$.

Lemma 5.2 suggests that $\widehat{\mathrm{SID}}{}_K^\nu(T, \mathbf{X})$ is a degenerate U-statistic under H_0. Using the Hoeffding decomposition (Lee, 1990), we can approximate $\widehat{\mathrm{SID}}{}_K^\nu(T, \mathbf{X})$ with the U-statistic based on the second-order projection $h_{n2}(\mathbf{z}, \mathbf{z}'; t)$. Thus, the closed-form expression of $h_{n2}(\mathbf{z}, \mathbf{z}'; t)$ is key to establishing the asymptotic null distribution. Additionally, note that $h_n(\mathbf{Z}_i, \mathbf{Z}_j, \mathbf{Z}_k, \mathbf{Z}_l; t)$ depends on the bandwidth h and thus, is a variable kernel. We can apply the asymptotic theory for a degenerate variable kernel U-statistic established by Fan and Li (1996) to study the asymptotic null distribution of $\widehat{\mathrm{SID}}_K(T, \mathbf{X})$.

Theorem 5.3. *Assume that conditions (C1)–(C3) hold. Under H_0, we have*

$$nh^{1/2}\widehat{\mathrm{SID}}_K(T, \mathbf{X}) \xrightarrow{d} N\left(0, 2\sigma^2\right),$$

where $\sigma^2 = E\{E\{U^2(\mathbf{X}_1, \mathbf{X}_2; Y) | Y, \delta = 1\}S^4(Y)f^3_{Y,\delta}(Y, 1)\} \int_0^\tau [\int_0^\tau W(u)W(u + v)du]^2 dv$.

Theorem 5.3 suggests that $\widehat{\mathrm{SID}}_K(T, \mathbf{X})$ converges in distribution to $N\left(0, 2\sigma^2\right)$ at the rate of $nh^{1/2}$, under H_0. This is consistent with the existing literature on nonparametric tests, such as Su and White (2007), and Ke and Yin (2020). We next study the asymptotic behavior of $\widehat{\mathrm{SID}}_K(T, \mathbf{X})$ under H_1.

Theorem 5.4. *Assume that conditions (C1)–(C3) hold. Under H_1, we have that $nh^{1/2}\widehat{\mathrm{SID}}_K(T, \mathbf{X}) \xrightarrow{P} \infty$, and $\lim_{n\to\infty} P\{nh^{1/2}\widehat{\mathrm{SID}}_K(T, \mathbf{X}) > \gamma | H_1\} = 1$, for any $\gamma > 0$.*

It can be seen from Theorems 5.3 and 5.4 that $nh^{1/2}\widehat{\mathrm{SID}}_K(T,\mathbf{X})$ is stochastically bounded under the null hypothesis, whereas it diverges to infinity under fixed alternative hypotheses. Thus, $nh^{1/2}\widehat{\mathrm{SID}}_K(T,\mathbf{X})$ is able to detect any type of dependency between T and $\mathbf{X}$. Additionally, under the assumption $E\{\|\mathbf{X}\|^\beta\}<\infty$, $\beta\in(0,2)$, we can derive similar asymptotic properties for $\widehat{\mathrm{SID}}_\beta(T,\mathbf{X})$ as those in Theorems 5.3 and 5.4.

6. Wild Bootstrap

Note that the limiting null distribution of $\widehat{\mathrm{SID}}_K(T,\mathbf{X})$ is unknown. In this section, we propose a wild bootstrap approach to approximate the limiting null distribution. To develop our wild bootstrap approach, a strategy commonly used is to perturb the symmetrized version of $\widehat{\mathrm{SID}}_K(T,\mathbf{X})$ through a zero-mean and unit-variance random variable. However, note that the symmetrical kernel of $\widehat{\mathrm{SID}}_K(T,\mathbf{X})$ in (2) or (3) is computationally intensive and thus, the above strategy is infeasible. To overcome this problem, we use Lemma 5.2 and perturb the second-order projection U-statistic to approximate the null distribution. Specifically, with some abuse of notation, define the wild bootstrap test as

$$\widehat{\mathrm{SID}}^*_K(T,\mathbf{X})=\frac{2}{n(n-1)}\sum_{i<j}e_ie_j\int_0^\tau\widehat{U}(\mathbf{X}_i,\mathbf{X}_j;t)\widehat{V}(Y_i,\delta_i;Y_j,\delta_j;t)d\overline{N}(t),$$

where $\{e_i\}_{i=1}^n$ are i.i.d. samples drawn from a zero-mean and unit-variance variable. Here, $\widehat{U}(\mathbf{x},\mathbf{x}';t)$ and $\widehat{V}(y,\tilde{\delta};y',\tilde{\delta}';t)$ are plug-in estimators of $U(\mathbf{x},\mathbf{x}';t)$ and $V(y,\tilde{\delta};y',\tilde{\delta}';t)$, where $E\{K(\mathbf{x},\mathbf{X}_1)\mid Y_1(t)=1\}$, $E\{K(\mathbf{X}_1,\mathbf{X}_2)\,|\,Y_1(t)=1,Y_2(t)=1\}$, $S(t)$ and $F_h(t)$ are estimated by

$$\widehat{E}\{K(\mathbf{x},\mathbf{X}_1)|Y_1(t)=1\}=\sum_{i=1}^n K(\mathbf{x},\mathbf{X}_i)Y_i(t)/\sum_{i=1}^n Y_i(t),$$

$$\widehat{E}\{K(\mathbf{X}_1,\mathbf{X}_2)|Y_1(t)=1,Y_2(t)=1\}=\sum_{i,j=1}^n K(\mathbf{X}_i,\mathbf{X}_j)Y_i(t)Y_j(t)/\{\sum_{i=1}^n Y_i(t)\}^2,$$

$$\widehat{S}(t)=\frac{1}{n}\sum_{i=1}^n I(Y_i\ge t),\quad \widehat{F}_h(t)=\frac{1}{n}\sum_{i=1}^n\delta_iW_h(Y_i-t).$$

The wild bootstrap test procedure is given by:

1. Generate the bootstrap samples of $\widehat{\mathrm{SID}}^*_K(T,\mathbf{X})$:

$$\widehat{\mathrm{SID}}^{*(b)}_K(T,\mathbf{X})=\frac{1}{n^2}\sum_{i<j}e_i^{(b)}e_j^{(b)}\int_0^\tau\widehat{U}(\mathbf{X}_i,\mathbf{X}_j;t)\widehat{V}(Y_i,\delta_i;Y_j,\delta_j;t)d\overline{N}(t),\qquad(1)$$

   where $\{e_i^{(b)}\}_{i=1}^n$ are i.i.d. samples from a distribution with zero mean and unit variance.
2. Repeat step 1 for B times and obtain $\{nh^{1/2}\widehat{\mathrm{SID}}^{*(b)}_K(T,\mathbf{X})\}_{b=1}^B$.
3. Calculate the $(1-\alpha)$-th quantile of $\{nh^{1/2}\widehat{\mathrm{SID}}^{*(b)}_K(T,\mathbf{X})\}_{b=1}^B$, denoted as $\widehat{Q}^*_{n,(1-\alpha)}$, for a given significance level α.
4. Reject the null hypothesis if $nh^{1/2}\widehat{\mathrm{SID}}_K(T,\mathbf{X})\ge\widehat{Q}^*_{n,(1-\alpha)}$.

We next state the validity of the above wild bootstrap procedure. From the proof of Theorem 5.3, we show that $\widehat{\mathrm{SID}}_K^{\nu}(T,\mathbf{X})$ is a degenerate U-statistic of second order under H_0, and its limiting distribution can be approximated by that of the second-order projection-based U-statistic:

$$\frac{12}{n(n-1)}\sum_{i<j}\int_0^{\tau} h_{n2}(\mathbf{Z}_i,\mathbf{Z}_j;t)d\nu(t). \tag{2}$$

From Lemma 5.2 we can see that $\widehat{\mathrm{SID}}_K^{*}(T,\mathbf{X})$ is essentially a bootstrap approximation to (2). This is confirmed by the following theorem.

Theorem 6.1. *Assume that conditions (C1)–(C3) hold. Under H_0, and conditional on the sample data $\{(Y_i,\delta_i,\mathbf{X}_i),i=1,\ldots,n\}$, we have*

$$nh^{1/2}\widehat{\mathrm{SID}}_K^{*}(T,\mathbf{X}) \xrightarrow{d^*} N\left(0,2\sigma^2\right),$$

where $\xrightarrow{d^*}$ is convergence in distribution with respective to $\{e_i\}_{i=1}^n$.

Theorem 6.2. *Assume that conditions (C1)–(C3) hold. Under H_0, we have*

$$nh^{1/2}\widehat{\mathrm{SID}}_K^{*}(T,\mathbf{X}) \xrightarrow{d^*} N\left(0,2\sigma^2\right),$$

where $\overset{d^*}{\to}$ is convergence in distribution with respective to $\{e_i\}_{i=1}^n$.

Theorem 6.1 implies that $\widehat{\mathrm{SID}}_K^{*}(T,\mathbf{X})$ is consistent with the limiting distribution of $\widehat{\mathrm{SID}}_K(T,\mathbf{X})$ under H_0. The following theorem illustrates the asymptotic behavior of the bootstrap test statistic under H_1.

Theorem 6.3. *Assume that conditions (C1)–(C3) hold. Under H_1, we have that*

$$\lim_{n\to\infty} P\{nh^{1/2}\widehat{\mathrm{SID}}_K(T,\mathbf{X}) \geq \widehat{Q}^*_{n,(1-\alpha)}|H_1\}=1.$$

Theorem 6.3 suggests the the wild bootstrap method can detect a dependency between survival times and covariates asymptotically. However, the U-statistic $\widehat{\mathrm{SID}}_K(T,\mathbf{X})$ is computationally expensive in practice. Instead, we can use its V-statistic version $\widetilde{\mathrm{SID}}_K(T,\mathbf{X})$ to reduce the computational cost. The wild bootstrap statistic of $\widetilde{\mathrm{SID}}_K(T,\mathbf{X})$ is given by

$$\widetilde{\mathrm{SID}}_K^{*}(T,\mathbf{X})=\frac{1}{n^2}\sum_{i,j=1}^{n}e_ie_j\int_0^{\tau}\widehat{U}(\mathbf{X}_i,\mathbf{X}_j;t)\widehat{V}(Y_i,\delta_i;Y_j,\delta_j;t)d\overline{N}(t).$$

Let $\widetilde{\mathrm{SID}}_K^{*(b)}(T,\mathbf{X})$, $b=1,\ldots,B$, be the bootstrap samples of $\widetilde{\mathrm{SID}}_K^{*}(T,\mathbf{X})$. Then, we reject the null hypothesis if $nh^{1/2}\widetilde{\mathrm{SID}}_K(T,\mathbf{X}) \geq \widetilde{Q}^*_{n,(1-\alpha)}$, where $\widetilde{Q}^*_{n,(1-\alpha)}$ is the $(1-\alpha)$-th quantile of $\{nh^{1/2}\widetilde{\mathrm{SID}}_K^{*(b)}(T,\mathbf{X})\}_{b=1}^{B}$.

7. Monte Carlo Simulations

In this section, we study the finite sample performance of the proposed SID-based test methods. We consider the two classes of SID test statistics: characteristic function-based SID tests with the weight function in (7) (SID_β) and kernel-based SID tests (SID_K). For SID_β, we consider the two special choices of β: $\beta = 1$ (SID_1) and $\beta = 1/2$ ($\mathrm{SID}_{0.5}$). For SID_K, we consider the two special kernel functions: the Gaussian kernel ($\mathrm{SID}_{\mathrm{Gauss}}$) and the Laplacian kernel ($\mathrm{SID}_{\mathrm{Lap}}$): $K(\mathbf{x}_1, \mathbf{x}_2) = \exp(-\|\mathbf{x}_1 - \mathbf{x}_2\|^2/\gamma_1^2)$ and $K(\mathbf{x}_1, \mathbf{x}_2) = \exp\left(-\|\mathbf{x}_1 - \mathbf{x}_2\|/\gamma_2\right)$, where $\gamma_1, \gamma_2 > 0$ are tuning parameters. In the numerical studies, we use the median-distance heuristic (Gretton et al., 2012), $\gamma_{\mathrm{med}} = (\text{median } \{\|\mathbf{X}_i - \mathbf{X}_j\|^2 : i \neq j\}/2)^{1/2}$, to select γ_1 and γ_2.

We compare the results from our methods with those from a Cox's proportional hazards model-based test (CPH), the IPCW-based distance covariance (IPCW; Edelmann, Welchowski and Benner, 2021), and the kernel log-rank test (KLR; Fernández et al., 2021). According to Edelmann, Welchowski and Benner (2021) and Fernández et al. (2021), we compute the critical values of the IPCW and KLR test statistics with the permutation and wild bootstrap methods. For KLR, we choose the Gaussian kernels as the kernel functions with respect to time and covariates to conserve space. Here, we repeat each experiment 1000 times and set the permuted and bootstrap sample sizes $B = 2000$. For the wild bootstrap method used by the KLR and SIDs, $\{e_i\}_{i=1}^n$ are generated from the Rademacher distribution, i.e., $P\{e_i = \pm 1\} = 1/2$.

Example 1. *This example examines the type-I error rates of our SID-based tests. We consider the following four cases, which were investigated by Fernández et al. (2021):*

Case 1: $T \mid X \sim \mathrm{Exp}(\lambda)$ *and* $C \mid X \sim \mathrm{Exp}(1)$ *with* $X \sim \mathrm{Unif}[-1, 1]$;
Case 2: $T \mid X \sim \mathrm{Exp}(\lambda)$ *and* $C \mid X \sim \mathrm{Exp}(e^{X/3})$ *with* $X \sim \mathrm{Unif}[-1, 1]$;
Case 3: $T \mid X \sim \mathrm{Exp}(\lambda)$ *and* $C \mid X \sim \mathrm{Weib}(3.35 + 1.75X, 1)$ *with* $X \sim \mathrm{Unif}[-1, 1]$;
Case 4: $T \mid \boldsymbol{X} \sim \mathrm{Exp}(\lambda)$ *and* $C \mid \boldsymbol{X} \sim \mathrm{Exp}(e^{\boldsymbol{1}_{10}^T \boldsymbol{X}/20})$ *with* $\boldsymbol{X} \sim N_{10}(\mathbf{0}, \mathbf{\Sigma}_{10})$.

Here, $\mathrm{Weib}(a, b)$ *is the Weibull distribution with the parameters* a *and* b*, and* $\mathrm{Exp}(\lambda)$ *is the exponential distribution with mean* λ*, where* λ *is used to control the censoring rate approximately. Here,* $\underline{\boldsymbol{1}}_p = (1, \ldots, 1)^T \in \mathbb{R}^p$ *and* $\mathbf{\Sigma}_p = (0.5^{|j-k|})$.

Tables 1 and 2 summarize the empirical type-I error rates for the seven methods: CPH, KLR, IPCW, SID_1, $\mathrm{SID}_{0.5}$, $\mathrm{SID}_{\mathrm{Gauss}}$, and $\mathrm{SID}_{\mathrm{Lap}}$ at the significance levels $\alpha = 0.01$, 0.05, or 0.1 with 30% or 60% censoring and $n = 50$ or 150. It can be seen that the empirical type-I error rates for SID_1, $\mathrm{SID}_{0.5}$, $\mathrm{SID}_{\mathrm{Gauss}}$, and $\mathrm{SID}_{\mathrm{Lap}}$ are very close to the true significance levels, which confirms the asymptotical properties in Theorems 5.3 and 6.1. Additionally, the KLR has similar performance.

As seen from Tables 1 and 2, the empirical type-I error rates for the IPCW method are inflated in cases 2–4, especially when the censoring rate is 60%. This is, perhaps, because IPCW assumes that C is completely independent of $\mathbf{X}$. Specifically, in case 1, for which the completely independent assumption is satisfied, IPCW achieves approximately the true significance levels, whereas it has inflated the empirical type-I error rates in cases 2–4, for which the assumption is invalid. Additionally, the empirical type-I error rates of the CPH method are under reasonable control, except case 4, where p is relatively high.

Example 2. *This example examines the power of our methods in various dependence relations and compares them with IPCW under the completely independent assumption. We generate the data in a similar way as Edelmann, Welchowski and Benner (2021):*

Table 1. Empirical type-I error rate with 30% censoring for Example 1.

						SID_β		SID_K	
α	Case	n	CPH	KLR	IPCW	SID_1	$\mathrm{SID}_{0.5}$	$\mathrm{SID}_{\mathrm{Gauss}}$	$\mathrm{SID}_{\mathrm{Lap}}$
0.01	Case 1	50	0.012	0.009	0.007	0.008	0.007	0.007	0.007
		150	0.007	0.010	0.011	0.007	0.009	0.008	0.009
	Case 2	50	0.012	0.005	0.029	0.006	0.007	0.006	0.007
		150	0.010	0.010	0.014	0.013	0.010	0.011	0.011
	Case 3	50	0.021	0.009	0.017	0.012	0.011	0.011	0.015
		150	0.016	0.010	0.014	0.013	0.014	0.013	0.012
	Case 4	50	0.011	0.010	0.021	0.013	0.013	0.013	0.015
		150	0.033	0.013	0.012	0.011	0.012	0.013	0.012
0.05	Case 1	50	0.040	0.038	0.056	0.041	0.050	0.049	0.054
		150	0.048	0.051	0.052	0.048	0.054	0.047	0.045
	Case 2	50	0.042	0.042	0.105	0.051	0.050	0.048	0.050
		150	0.045	0.056	0.071	0.051	0.049	0.054	0.042
	Case 3	50	0.042	0.043	0.059	0.053	0.049	0.050	0.048
		150	0.056	0.050	0.053	0.061	0.057	0.056	0.058
	Case 4	50	0.121	0.083	0.080	0.066	0.065	0.055	0.056
		150	0.076	0.058	0.064	0.051	0.052	0.051	0.052
0.10	Case 1	50	0.088	0.088	0.103	0.101	0.101	0.097	0.108
		150	0.096	0.110	0.092	0.092	0.091	0.098	0.096
	Case 2	50	0.089	0.086	0.173	0.110	0.107	0.091	0.105
		150	0.099	0.112	0.124	0.101	0.102	0.107	0.106
	Case 3	50	0.091	0.079	0.124	0.109	0.103	0.104	0.109
		150	0.101	0.112	0.106	0.100	0.107	0.105	0.096
	Case 4	50	0.213	0.115	0.148	0.115	0.112	0.105	0.109
		150	0.140	0.146	0.114	0.099	0.101	0.100	0.095

(1) *Log-linear:* $\log(T) = 0.5\eta(X) + \varepsilon$ *with* $\eta(X) = X$*;*
(2) *Log-quadratic:* $\log(T) = 1.2\eta(X) + \varepsilon$ *with* $\eta(X) = X^2$*;*
(3) *Log-cubic:* $\log(T) = \eta(X) + \varepsilon$ *with* $\eta(X) = X^3$*;*
(4) *Log-cosine:* $\log(T) = 0.5\eta(X) + \varepsilon$ *with* $\eta(X) = \cos(3X)$*;*
(5) *Log-twolines:* $\log(T) = 4\eta(X) + \varepsilon$ *with* $\eta(X) = I(A=1)X - I(A=0)X$, $A \sim B(1, 0.5)$, $A \perp X$*;*
(6) *Log-circle:* $\log(T) = \eta(X_1, X_2) + \varepsilon$ *with* $\eta(X_1, X_2) = 1$ *if* $X_1^2 + X_2^2 \leq 0.5$ *and* $\eta(X_1, X_2) = 0$ *otherwise,* $X_1 \perp X_2$.

Here, $X, X_1, X_2 \sim \mathrm{Unif}[-1, 1]$, $\varepsilon \sim N(0, 1)$, *and* $C \sim \mathrm{Exp}(\lambda)$.

Figure 1 shows plots of the power against sample size at $\alpha = 0.05$ with 30% censoring for Example 2. All four of our proposed test methods, SID_1, $\mathrm{SID}_{0.5}$, $\mathrm{SID}_{\mathrm{Gauss}}$, and $\mathrm{SID}_{\mathrm{Lap}}$, performed better than IPCW for the six dependence relations. The results suggest that our methods are all capable of detecting linear and nonlinear dependencies.

Table 2. Empirical type-I error rate with 60% censoring for Example 1.

						SID_β		SID_K	
α	Case	n	CPH	KLR	IPCW	SID_1	$\mathrm{SID}_{0.5}$	$\mathrm{SID}_{\mathrm{Gauss}}$	$\mathrm{SID}_{\mathrm{Lap}}$
0.01	1	50	0.012	0.003	0.013	0.011	0.009	0.007	0.008
		150	0.007	0.008	0.008	0.009	0.010	0.008	0.009
	2	50	0.004	0.007	0.027	0.016	0.015	0.015	0.011
		150	0.014	0.006	0.022	0.009	0.007	0.008	0.009
	3	50	0.014	0.008	0.021	0.015	0.013	0.012	0.015
		150	0.008	0.004	0.011	0.009	0.009	0.011	0.012
	4	50	0.061	0.004	0.006	0.010	0.011	0.011	0.011
		150	0.022	0.010	0.012	0.013	0.013	0.013	0.012
0.05	1	50	0.054	0.042	0.054	0.044	0.050	0.051	0.046
		150	0.051	0.052	0.046	0.051	0.047	0.045	0.050
	2	50	0.053	0.038	0.062	0.054	0.051	0.048	0.051
		150	0.048	0.050	0.041	0.051	0.056	0.053	0.056
	3	50	0.052	0.041	0.035	0.046	0.045	0.048	0.042
		150	0.057	0.050	0.051	0.058	0.052	0.047	0.049
	4	50	0.151	0.047	0.070	0.045	0.044	0.043	0.040
		150	0.084	0.055	0.068	0.063	0.058	0.055	0.057
0.10	1	50	0.099	0.091	0.083	0.099	0.100	0.095	0.097
		150	0.098	0.110	0.107	0.110	0.127	0.112	0.113
	2	50	0.110	0.096	0.152	0.113	0.112	0.111	0.090
		150	0.120	0.121	0.132	0.103	0.101	0.101	0.102
	3	50	0.093	0.121	0.118	0.102	0.106	0.115	0.095
		150	0.118	0.095	0.104	0.120	0.125	0.115	0.102
	4	50	0.256	0.097	0.121	0.110	0.107	0.103	0.108
		150	0.149	0.101	0.113	0.100	0.099	0.100	0.104

Figure 1 also reveals that SID_1 slightly outperformed $\mathrm{SID}_{0.5}$, $\mathrm{SID}_{\mathrm{Gauss}}$, and $\mathrm{SID}_{\mathrm{Lap}}$ for the log-linear and log-cosine cases, while it was inferior to these three methods for the other cases. This is related to the norm $\|\mathbf{x}_1 - \mathbf{x}_2\|$ used by SID_1. In fact, SID_1 works in a similar way to the well-known distance covariance (Székely, Rizzo and Bakirov, 2007) and the energy statistic (Székely and Rizzo, 2013), which also depend on the norm $\|\mathbf{x}_1 - \mathbf{x}_2\|$. We can alleviate the poor performance of SID_β by properly choosing β.

Example 3. *This example compares the empirical power of our methods with that of KLR. The following cases are studied:*

Case 1: $T \mid \boldsymbol{X} \sim \mathrm{Exp}(e^{\boldsymbol{1}_p^T \boldsymbol{X}/10})$. *Case 2:* $T \mid \boldsymbol{X} \sim \mathrm{Exp}(e^{(\beta^T \boldsymbol{X})^2/2})$.

Case 3: $\log(T) = 0.2\beta^T \boldsymbol{X} + 2\varepsilon$. *Case 4:* $\log(T) = -0.5(\beta^T \boldsymbol{X})^2 + 4\varepsilon$.

Case 5: $\log(T) = 0.25\beta_1^T \boldsymbol{X} + 1.5(\beta_2^T \boldsymbol{X})\varepsilon$. *Case 6:* $\log(T) = 2(\beta_1^T \boldsymbol{X})^2 + 0.15(\beta_2^T \boldsymbol{X})\varepsilon$.

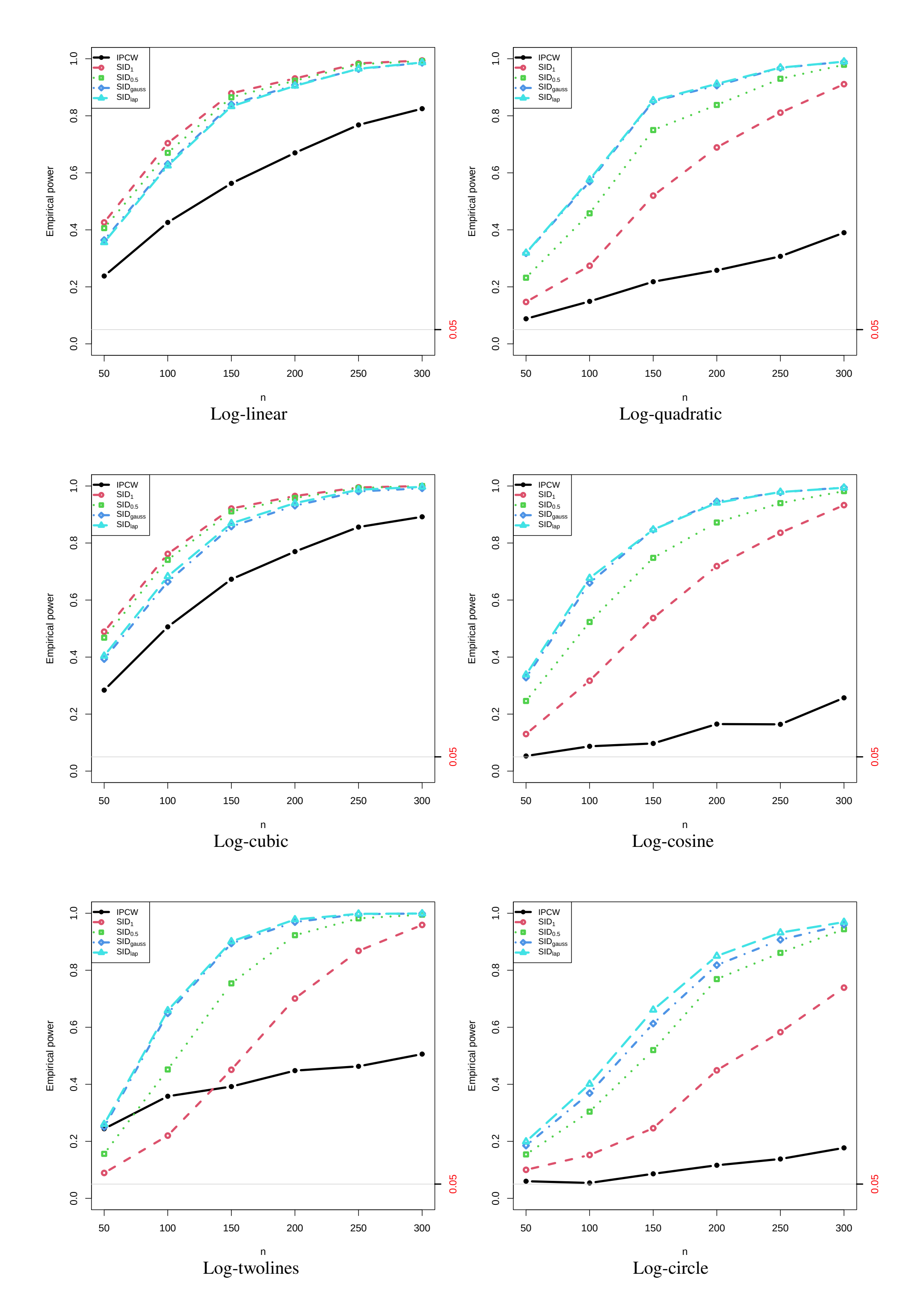


Figure 1. Comparisons of empirical power at $\alpha = 0.05$ with 30% censoring for Example 2.

Here, $\boldsymbol{X} = (X_1, \ldots, X_p)^T$ is generated from $N_p\left(0, \boldsymbol{\Sigma}_p\right)$ with $\boldsymbol{\Sigma}_p = (0.5^{|j-k|})$. Consider that $p = 6$, $\beta = (1,1,1,-1,-1,-1)^T$, $\beta_1 = (0,0,1,-1,0,0)^T$, $\beta_2 = (1,1,0,0,0,0)^T$, $\varepsilon \sim N(0,1)$, and $C \sim \mathrm{Exp}(\lambda)$.

The empirical comparisons of the power for Example 3 are summarized in Figure 2 at $\alpha = 0.05$ with 30% censoring. Note that cases 1 and 2 are Cox models, and cases 3–6 are accelerated failure time models with homogeneous and heterogeneous errors. As expected, CPH has the highest power in case 1 but loses power in the other cases as the CPH assumption is invalid. Our methods, especially $\mathrm{SID}_{\mathrm{Gauss}}$ and $\mathrm{SID}_{\mathrm{Lap}}$, are comparable and even mostly superior to KLR. Additionally, SID_1 has similar performance as in Example 2, as it is powerful in log-linear dependence relations but behaves worse in log-nonlinear or more complex dependence relations.

Example 4. *In the example, we study the empirical power of our methods when the censoring time C depends on the covariates. The data are generated from the models defined in cases 1–6 of Example 3, except that C is generated from $\mathrm{Exp}(e^{\lambda+X_1})$, where λ is used to control the censoring rate.*

As seen from Figure 3, our four methods, as well as KLR, are all capable of detecting the different dependence relations between T and $\mathbf{X}$ when the censoring time C depends on the covariates. Their performances are basically similar to those presented in Figure 2. These results in the two figures indicate that our methods work well, regardless of the association between C and $\mathbf{X}$.

Example 5. *In this example, we further compare the performance of our methods with those of KLR in the following mixture cure rate models:*

Case 1: $T = \eta T^* + (1-\eta)\infty$, *with* $T^* \mid X \sim \mathrm{Exp}(e^{0.5X})$, $X \sim N(0,1)$*;*
Case 2: $T = \eta T^* + (1-\eta)\infty$, *with* $T^* \mid X \sim \mathrm{Exp}(e^{0.5X^2})$, $X \sim N(0,1)$*;*
Case 3: $T = \eta T^* + (1-\eta)\infty$, *with* $\log(T^*) = 0.5\beta^T\boldsymbol{X} + 3\varepsilon$*;*
Case 4: $T = \eta T^* + (1-\eta)\infty$, *with* $\log(T^*) = 0.2(\beta^T\boldsymbol{X})^3 + \varepsilon$.

Here, $\eta \sim B(1,0.6)$. In cases 3-4, $\beta = (1,1,1,-1,1,-1)^T$, $\boldsymbol{X} \sim N_6(0,\Sigma_6)$ with $\boldsymbol{\Sigma}_6 = (0.5^{|j-k|})$, and $C \sim \mathrm{Exp}(\lambda)$.

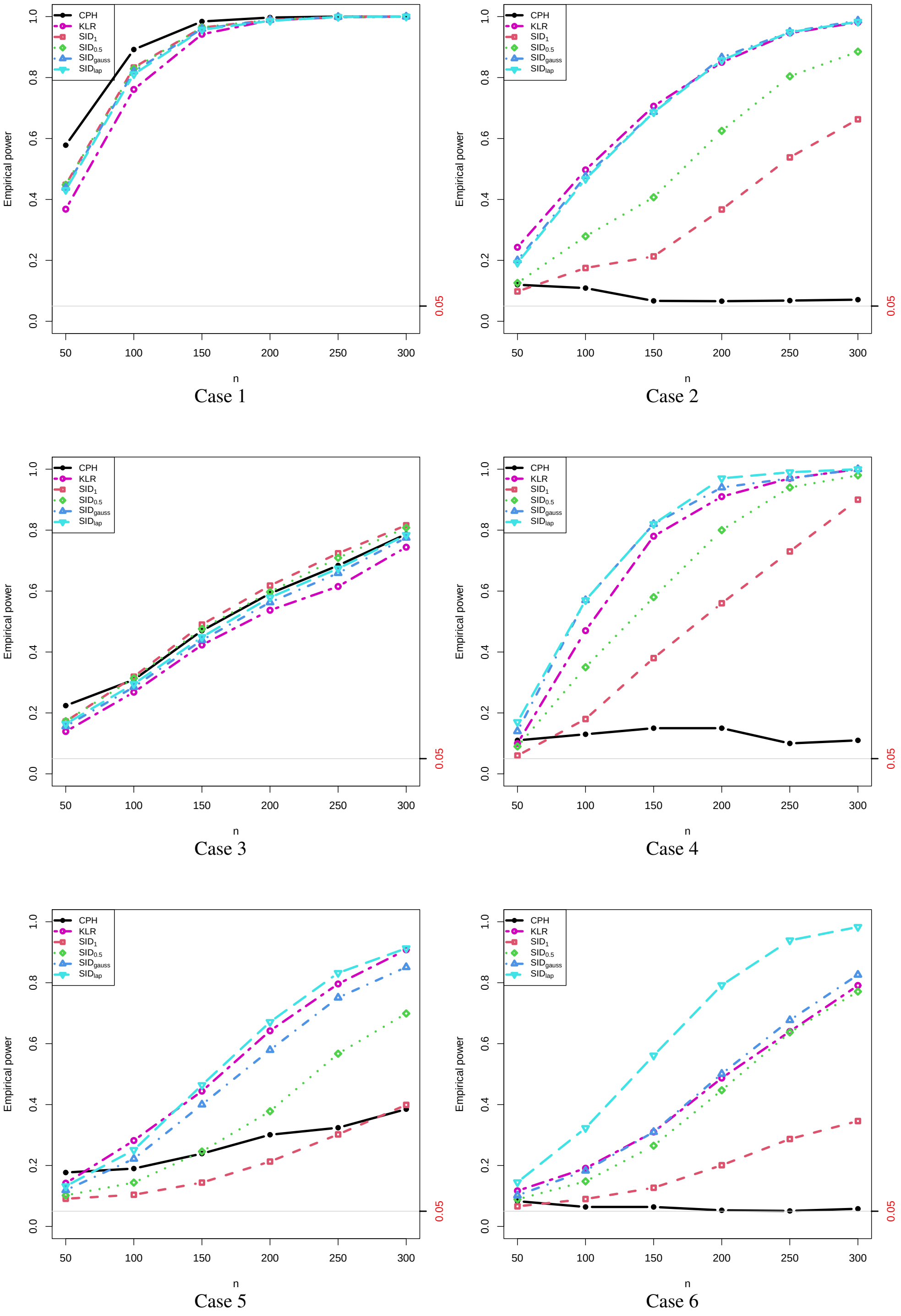
Figure 4 shows that the SID_1, $\mathrm{SID}_{0.5}$, $\mathrm{SID}_{\mathrm{Gauss}}$, and $\mathrm{SID}_{\mathrm{Lap}}$ tests outperformed KLR in the four mixture cure rate models of Example 5. Compared with the results for Examples 3 and 4, KLR was significantly inferior to our methods. This is, presumably, because the SIDs are local and need not to calculate pairwise distances between observations, which have high volatility in the settings of Example 5.

Example 6. *The example investigates the empirical power of our methods for small deviations from the null hypothesis. We generated data for two cases:*

Case 1: $\log(T) = \theta X + \varepsilon$ *and* $C \sim \mathrm{Exp}(3)$.
Case 2: $\log(T) = \theta X^2 + \varepsilon$ *and* $C \sim \mathrm{Exp}(3)$.

Here, X and ε were generated from $N(0,1)$.

Figure 5 displays the power with respect to θ at the significance level $\alpha = 0.05$ and $n = 100$. The power of our methods increases rapidly as θ moves away from 0, so that they are highly competitive with the KLR test. The results indicate that our methods can effectively detect the difference between the null and alternative hypotheses.



Case 1

Case 2

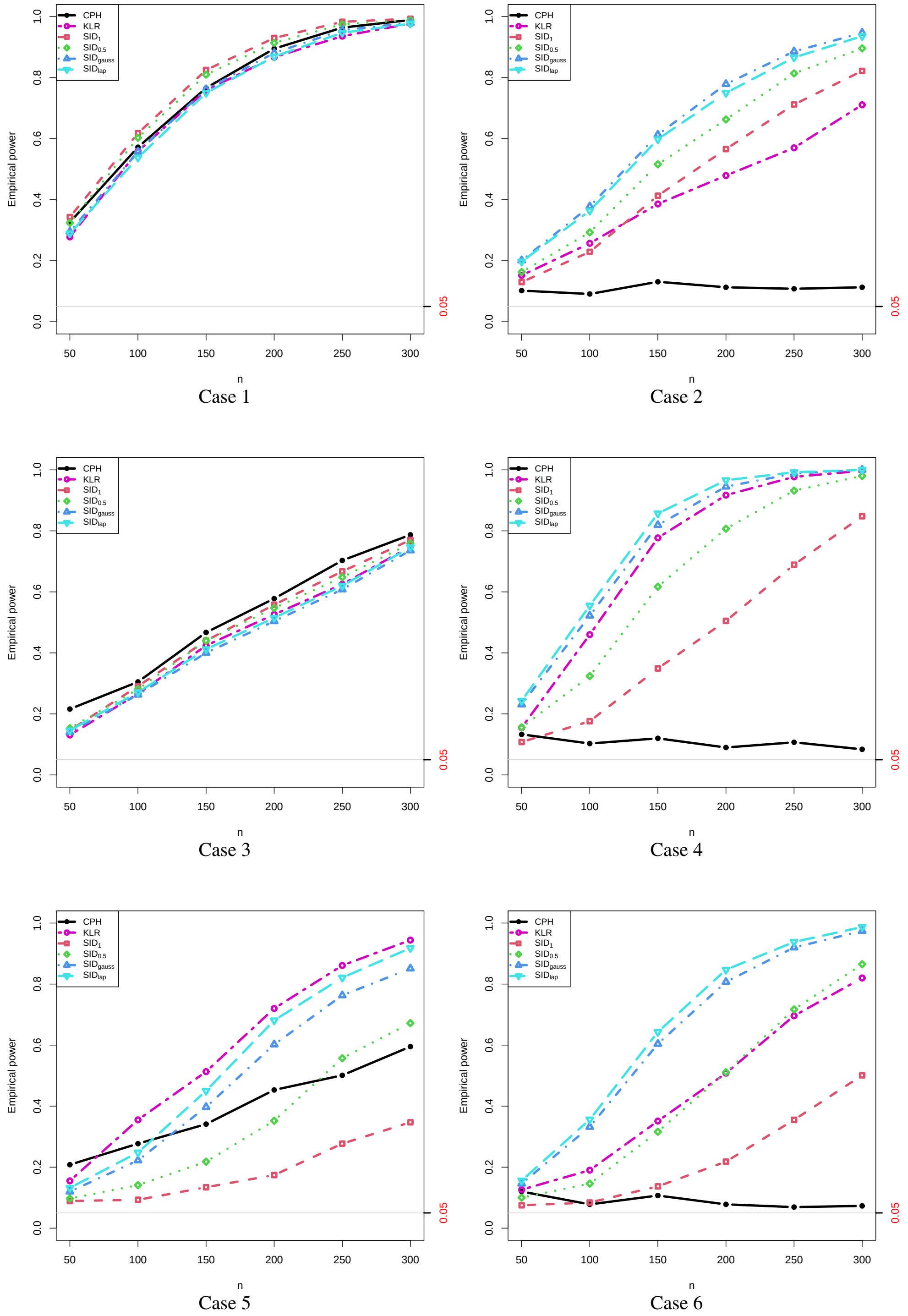
Case 3

Case 4

Case 5

Case 6

Figure 2. Comparisons of empirical power at $\alpha = 0.05$ with 30% censoring for Example 3.



Case 1

Case 2

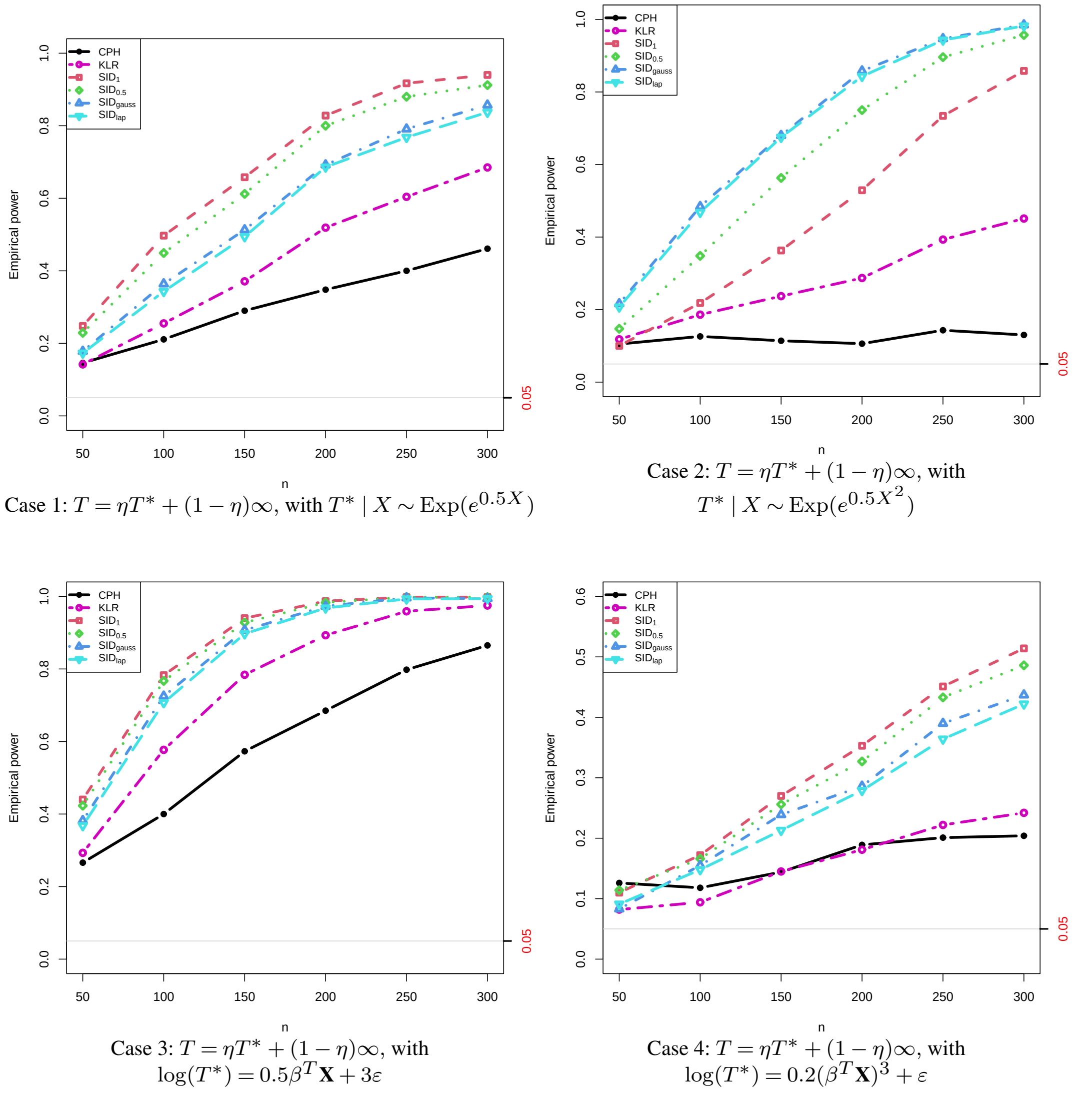
Case 3

Case 4

Case 5

Case 6

Figure 3. Comparisons of empirical power at $\alpha = 0.05$ with 30% censoring for Example 4.



Case 1: $T = \eta T^* + (1-\eta)\infty$, with $T^* \mid X \sim \mathrm{Exp}(e^{0.5X})$

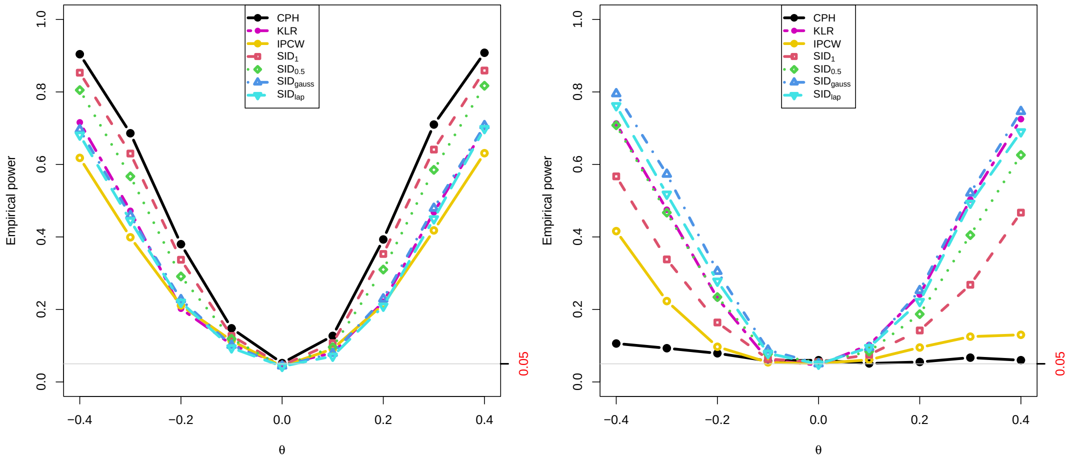
Case 2: $T = \eta T^* + (1-\eta)\infty$, with $T^* \mid X \sim \mathrm{Exp}(e^{0.5X^2})$

Case 3: $T = \eta T^* + (1-\eta)\infty$, with $\log(T^*) = 0.5\beta^T\mathbf{X} + 3\varepsilon$

Case 4: $T = \eta T^* + (1-\eta)\infty$, with $\log(T^*) = 0.2(\beta^T\mathbf{X})^3 + \varepsilon$

Figure 4. Empirical power comparisons at $\alpha = 0.05$ with 50% censoring for Example 5.

From all the numerical results, we can draw the following conclusions: (1) Our proposed SID-based metrics are capable of detecting and testing various dependence relations between the censored outcome and covariates; (2) SID_1 has high power with linear dependence relations but behaves worse with nonlinear dependence relations. With a better choice of β, SID_β could be as powerful as SID_K; (3) In most settings, our test methods are highly comparable to the KLR test and more powerful than IPCW. Overall, our methods do not require strong assumptions on the censoring mechanism.



Case 1:
$\log(T) = \theta X + \varepsilon$

Case 2:
$\log(T) = \theta X^2 + \varepsilon$

Figure 5. Comparisons of empirical power at $\alpha = 0.05$ and $n = 100$ for Example 6.

8. Real Data Analysis

In this section, we illustrate the SID-based tests with empirical analyses of two real datasets. We use 2000 wild bootstrap samples for the KLR and SID-based tests.

Example 7 (BMT data). *This example considers the bone marrow transplant (BMT) data in Klein and Moeschberger (2003), available as the* `bmt` *dataset of the R package* `KMsurv`*. The dataset is for the recovery process of 137 patients after BMT. At the time of transplantation, several risk factors were measured. Risk factors, denoted by z_1 to z_{10}, include the age, gender, and cytomegalovirus immune status for both recipient and donor, and also the waiting time from diagnosis to transplantation, the French-American-British disease grade, hospital stay, and methotrexate dose. In our analysis, we consider the time of death, denoted by t_1, as the event time. About 59.1% of the failure times are censored in the study.*

In the analysis, we are interested in detecting the dependence between t_1 and two of the covariates, namely recipient age in years (z_1) and the waiting time to transplantation in days (z_7). Table 3 lists the p-values for the seven test methods for the BMT data. Our four proposed methods, except SID$_1$, have smaller p-values than the KLR test. In particular, SID$_{\text{Gauss}}$ has the smallest p-values (0.037, 0.018, and 0.035) among the seven methods, which implies that t_1 and z_1/z_7 or both are significantly dependent.

To illustrate the alleged dependence relations, we classify the dataset into four subgroups by the medians of z_1 and z_7. Figure 6 displays the Kaplan–Meier estimates of the survival times for each subgroup, where $Q_{0.5}(z_1) = 28$ (years) and $Q_{0.5}(z_7) = 178$ (days). The survival curves can be divided by the medians of z_1 and z_7, which implies that t_1 depends on these two covariates. These results are consistent with the findings for the SID tests in Table 3. In contrast, KLR has higher p-values in Table 3.

Example 8 (Colon data). *This example considers a colon cancer dataset originally described in Laurie et al. (1989), available as the* `colonCS` *dataset of the R package* `condSURV`*. The dataset consists*

Table 3. p-values for the various tests for the BMT data.

				SID_β		SID_K	
Null hypothesis	CPH	KLR	IPCW	SID_1	$\mathrm{SID}_{0.5}$	$\mathrm{SID}_{\mathrm{Gauss}}$	$\mathrm{SID}_{\mathrm{Lap}}$
$H_0 : t_1 \perp\!\!\!\perp z_1$	0.304	0.101	0.090	0.061	0.068	0.037	0.059
$H_0 : t_1 \perp\!\!\!\perp z_7$	0.888	0.074	0.231	0.180	0.043	0.018	0.018
$H_0 : t_1 \perp\!\!\!\perp (z_1, z_7)$	0.589	0.115	0.238	0.195	0.075	0.035	0.036

Table 4. p-values of the various tests for Colon data.

				SID_α		SID_K	
Null hypothesis	CPH	KLR	IPCW	SID_1	$\mathrm{SID}_{0.5}$	$\mathrm{SID}_{\mathrm{gaus}}$	$\mathrm{SID}_{\mathrm{lap}}$
$H_0 : T \perp\!\!\!\perp \mathrm{Age}$	0.626	0.108	0.552	0.027	0.042	0.047	0.070
$H_0 : T \perp\!\!\!\perp (\mathrm{Age, Perfor, Adhere})$	0.102	0.016	0.543	0.015	0.020	0.025	0.012
$H_0' : C \perp\!\!\!\perp \mathrm{Age}$	0.222	0.566	/	0.799	0.758	0.832	0.842
$H_0' : C \perp\!\!\!\perp (\mathrm{Age, Perfor, Adhere})$	0.555	0.652	/	0.793	0.745	0.752	0.952

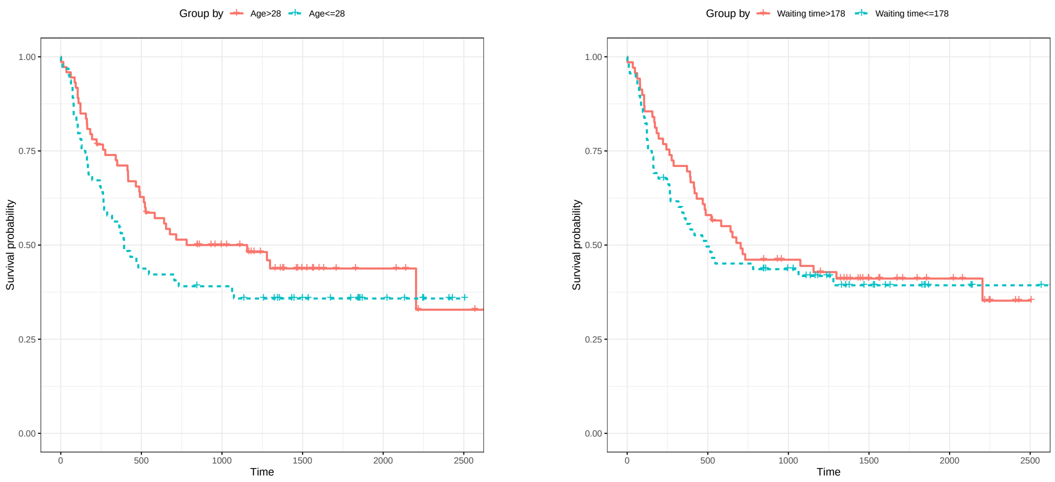


Figure 6. Kaplan–Meier estimates of the survival times grouped by the medians of the recipient age (left) and the waiting time to transplantation (right) for the BMT data.

of information about the recurrence of tumors and survival in 929 patients undergoing treatment for stage B/C colon cancer. Each patient record contains the time to cancer recurrence, the survival time, and 11 covariates. In the example, the survival time is the event of interest.

The dataset has been analyzed by Fernández et al. (2021). Like Fernández et al. (2021), we aim to test independence between the survival time T and the three covariates Age, Perfor, and Adhere. In Table 4, we summarize the p-values for the seven test methods used in our simulation studies. The first two rows show that our proposed tests can detect significant dependence between T and Age or between T and (Age, Perfor, Adhere). Moreover, for these two test problems, our proposed tests have smaller p-values, so it seems that they are more powerful than KLR.

It is a little surprising that IPCW has p-values of 0.552 and 0.543 and seems to fail to detect the above dependence. A direct doubt is whether the condition that the censoring time C is independent of $\mathbf{X}$ is invalid. To answer this problem, we consider the other two test problems: $H_0' : C \perp\!\!\!\perp$Age and $H_0' : C \perp\!\!\!\perp$(Age, Perfor, Adhere). The p-values of all tests except IPCW are summarized in the third and fourth rows of Table 4. All the test methods have high p-values. These indicate that there is no evidence of any correlation between T and Age or (Age, Perfor, Adhere) and thus, the doubt does not hold. The reasons why IPCW does not work for the colon data still require further study, which is beyond the scope of this paper.

9. Discussion

The proposed SID measures focus on continuous covariates. In survival data analysis, we often encounter datasets with categorical and continuous covariates. To extend the SID metrics into such cases, a simple approach is to use kernel functions for the categorical variables. In fact, such kernel functions have been proposed in literatures of support vector machine for classification problems, see, Belanche and Villegas (2013). Additionally, for $\mathrm{SID}_\beta(T, \mathbf{X})$ and $\mathrm{SID}_K(T, \mathbf{X})$, how to choose the optimal β and $K(\cdot, \cdot)$, including $a(t)$, is still an open question. These will be interesting topics for future research.

Supplementary material

Supplementary material includes provides the detailed technical proofs of the main results of the paper.

References

Belanche, L. and Villegas, M. (2013). Kernel Functions for Categorical Variables with Application to Problems in the Life Sciences. In *International Conference of the Catalan Association for Artificial Intelligence* 171-180.

Chatterjee, S. (2021). A new coefficient of correlation. *Journal of the American Statistical Association* **116** 2009–2022.

Cheng, S., Wei, L. and Ying, Z. (1995). Analysis of transformation models with censored data. *Biometrika* **82** 835–845.

Chwialkowski, K., Strathmann, H. and Gretton, A. (2016). A kernel test of goodness of fit. In *International conference on machine learning* 2606–2615. PMLR.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)* **34** 187–202.

Edelmann, D., Welchowski, T. and Benner, A. (2021). A consistent version of distance covariance for right-censored survival data and its application in hypothesis testing. *Biometrics* **78** 867-879.

Fan, Y. and Li, Q. (1996). Consistent Model Specification Tests: Omitted Variables and Semiparametric Functional Forms. *Econometrica* **64** 865-890.

Fernández, T., Gretton, A., Rindt, D. and Sejdinovic, D. (2021). A Kernel Log-Rank Test of Independence for Right-Censored Data. *Journal of the American Statistical Association* 1-12.

FLEMING, T. R. and HARRINGTON, D. P. (2011). *Counting Processes and Survival Analysis*. John Wiley & Sons.

GRETTON, A., FUKUMIZU, K., TEO, C., SONG, L., SCHÖLKOPF, B. and SMOLA, A. (2008). A kernel statistical test of independence. *Advances in Neural Information Processing Systems* **20** 585-592.

GRETTON, A., BORGWARDT, K. M., RASCH, M. J., SCHÖLKOPF, B. and SMOLA, A. (2012). A kernel two-sample test. *The Journal of Machine Learning Research* **13** 723–773.

KE, C. and YIN, X. (2020). Expected Conditional Characteristic Function-based Measures for Testing Independence. *Journal of the American Statistical Association* **115** 985-996.

KLEIN, J. P. and MOESCHBERGER, M. L. (2003). *Survival Analysis: Techniques for Censored and Truncated Data* **1230**. Springer.

LAURIE, J. A., MOERTEL, C. G., FLEMING, T. R., WIEAND, H. S., LEIGH, J. E., RUBIN, J., MCCORMACK, G. W., GERSTNER, J. B., KROOK, J. E. and MALLIARD, J. (1989). Surgical adjuvant therapy of large-bowel carcinoma: An evaluation of levamisole and the combination of levamisole and fluorouracil. The North Central Cancer Treatment Group and the Mayo Clinic. *Journal of Clinical Oncology* **7** 1447–1456.

LEE, J. (1990). *U-statistics: Theory and Practice*. Marcel Dekker, Inc., New York.

MANTEL, N. (1966). Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemotherapy Reports* **50** 163–170.

SEJDINOVIC, D., SRIPERUMBUDUR, B., GRETTON, A. and FUKUMIZU, K. (2013). Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *The Annals of Statistics* **41** 2263–2291.

SILVERMAN, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. CRC Press.

STUTE, W. (1993). Consistent estimation under random censorship when covariables are present. *Journal of Multivariate Analysis* **45** 89–103.

SU, L. and WHITE, H. (2007). A consistent characteristic function-based test for conditional independence. *Journal of Econometrics* **141** 807-834.

SZÉKELY, G. J., RIZZO, M. L. and BAKIROV, N. K. (2007). Measuring and testing dependence by correlation of distances. *The Annals of Statistics* **35** 2769–2794.

SZÉKELY, G. J. and RIZZO, M. L. (2013). Energy statistics: A class of statistics based on distances. *Journal of Statistical Planning and Inference* **143** 1249–1272.

WANG, X., PAN, W., HU, W., TIAN, Y. and ZHANG, H. (2015). Conditional distance correlation. *Journal of the American Statistical Association* **110** 1726–1734.

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

WEIHS, L., DRTON, M. and MEINSHAUSEN, N. (2018). Symmetric rank covariances: A generalized framework for nonparametric measures of dependence. *Biometrika* **105** 547–562.