Statistica Sinica **32** (2022), 2477-2496
doi:https://doi.org/10.5705/ss.202021.0203

SLICED INDEPENDENCE TEST

Yilin Zhang, Canyi Chen and Liping Zhu

Renmin University of China

Abstract: An ideal independence test should possess three properties: it should be zero-independence equivalent, numerically efficient, and asymptotically normal. We introduce a slicing procedure for estimating a popular measure of nonlinear dependence, leading to the resultant sliced independence test simultaneously possessing all three properties. In addition, the power performance of the sliced independence test improves as the number of observations within each slice increases. The popular rank test corresponds to a special case of the sliced independence test that contains two observations within each slice. The sliced independence test is thus more powerful than the rank test. The size performance of the sliced independence test is insensitive to the number of slices, in that the slicing estimation is consistent and asymptotically normal for a wide range of slice numbers. We further adapt the sliced independence test to account for the presence of multivariate control variables. The theoretical properties are confirmed using comprehensive simulations and an application to an astronomical data set.

Key words and phrases: Correlation, measure of association, rank tests.

1. Introduction

Testing for independence between two random variables is a fundamental problem in statistics. Weihs, Drton and Leung (2016); Weihs, Drton and Meinshausen (2018) stated that an independence test should simultaneously possess the following three properties:

1. *Zero-independence equivalent:* At the population level, the dependence metric is equal to zero if and only if the two random variables are independent. This ensures that the independence test is consistent.

2. *Numerically efficient:* The complexity of implementing an independence test is linear or nearly linear in the sample size n, say $O\{n\log(n)\}$. This is almost the minimal computational cost that we have to bear.

3. *Asymptotically normal:* The asymptotic null distribution of an independence test is normal. The asymptotic normality is more desirable for prac-

Corresponding author: Liping Zhu, Center for Applied Statistics and Institute of Statistics and Big Data, Renmin University of China, Beijing 100872, China. E-mail: zhu.liping@ruc.edu.cn.

titioners than is the asymptotically distribution-free property.

Many tests has been proposed in the literature. However, few of them possess these properties simultaneously. For example, the Pearson correlation (Pearson (1895)) and its variations, such as Spearman's rho (Spearman (1906)) and Kendall's tau (Kendall (1938)), are not zero-independence equivalent, although estimating these metrics is numerically efficient. The Hoeffding's index (Hoeffding (1948)) is zero-independence equivalent only if both random variables are continuous. The independence tests based on the correlation of Blum, Kiefer and Rosenblatt (1961) and its variations, such as Zhou and Zhu (2018), are asymptotically distribution free. However, implementing these independence tests has the complexity of a quadratic order of the sample size n, which is typically regarded as numerically inefficient. In general, the distance correlation (Székely, Rizzo and Bakirov (2007)), projection correlation (Zhu et al. (2017)), and binning approach (Heller, Heller and Gorfine (2013)) are not numerically efficient when used to test the independence between two random vectors. In addition, their asymptotic null distributions depend on the parent distribution of the two random vectors. A chi-squared distribution is suggested to approximate the asymptotic null distribution of the distance correlation test, although it is quite conservative (Székely, Rizzo and Bakirov (2007, p.2783)). In general, using bootstrap or random permutations to approximate asymptotic null distributions is regarded as computationally intensive. Several algorithms have been proposed to speed up the calculation of the distance correlation. In particular, Huang and Huo (2022), Huo and Székely (2016), and Chaudhuri and Hu (2019) improved the computational complexity of calculating the distance correlation to the order of $O\{n \log(n)\}$ when both random variables are univariate. Huang and Huo (2022) proposed approximating the asymptotic null distribution of the distance correlation test using a gamma distribution, although this lacks a rigorous theoretical justification (Gao et al. (2021, p.2012)).

Dette, Siburg and Stoimenov (2013), Kong, Xia and Zhong (2019), and Chatterjee (2021) independently introduced a dependence metric that is zero-independence equivalent. It has attracted much attention for its simplicity and implementability (Cao and Bickel (2020); Wiesel (2021)). Dette, Siburg and Stoimenov (2013) and Kong, Xia and Zhong (2019) suggested estimating this metric using a kernel smoother. However, implementing kernel smoothing has complexity of nearly a quadratic order of the sample size, which limits its usefulness when the sample size is extremely large. In addition, its asymptotic null distribution depends upon the kernel function. In contrast, Chatterjee (2021) proposed a rank estimation that is computationally efficient and asymptotically standard normal. This rank estimation satisfies all desirable properties simultaneously, and thus is more appealing than kernel smoothing.

Here, we introduce a slicing procedure for estimating the dependence metric suggested by Dette, Siburg and Stoimenov (2013), Kong, Xia and Zhong (2019), and Chatterjee (2021). This procedure divides the observations into several slices according to the realizations of one random variable, evaluates the local variation of the other within each slice, and aggregates the variations across all slices to form a slicing estimation. The complexity of implementing the resultant sliced independence test is nearly linear in the sample size, which is thus numerically efficient. The asymptotic null distribution is standard normal, and does not depend on the parent distributions of the two random variables. The resultant sliced independence test further improves the popular rank test of Chatterjee (2021), from two perspectives. The rank test corresponds to the sliced independence test when there are only two observations within each slice. We show that the power performance of the sliced independence test improves as the number of observations within each slice increases, even when the total sample size is fixed, making our proposed test more powerful than the rank test. In addition, the slicing estimation is consistent and asymptotically normal for a wide range of the number of slices. Therefore, the size performance of the sliced independence test is, surprisingly, highly insensitive to the number of slices. The concept of this slicing estimation procedure can be readily generalized to the multivariate case using K-means clustering (MacQueen (1967)). These theoretical properties are demonstrated using comprehensive simulations and an application to an astronomical dataset. An R package for implementing the sliced independence test will be available on the Comprehensive R Archive Network.

The remainder of this paper is organized as follows. We propose the slicing procedure and connect it with the rank test in Section 2. We study the asymptotic properties of the sliced independence test in Section 3, and generalize this slicing procedure to the multivariate case using K-means clustering in Section 4. We demonstrate the finite-sample performance of the sliced independence test using comprehensive simulations and an analysis of an astronomical data set in Section 5, and conclude the paper in Section 6. All technical proofs are relegated to the online Supplementary Material.

2. The Slicing Estimation Procedure

2.1. A brief review

Suppose X and Y are two univariate random variables. Define $s(t;X) \stackrel{\text{def}}{=} \text{pr}(Y \geq t \mid X)$. Let T be an independent univariate random variable with probability mass/density and cumulative distribution functions $\omega(t)$ and $\mu(t)$, respectively. The support of T is denoted by $\text{supp}(T) \stackrel{\text{def}}{=} \{t : \omega(t) > 0\}$. We assume throughout that $\text{supp}(Y) \subseteq \text{supp}(T)$. It follows immediately that X and Y are independent if and only if $\text{var}\{s(t;X)\} = 0$, for all $t \in \mathbb{R}$. Dette, Siburg and Stoimenov (2013), Kong, Xia and Zhong (2019) and Chatterjee (2021) independently suggested using the following metric to quantify the degree of deviation from independence:

$$\mathcal{S}(X,Y) \stackrel{\text{def}}{=} \frac{\int \text{var}\{s(t;X)\}\text{d}\mu(t)}{\int \text{var}\{1(Y \geq t)\}\text{d}\mu(t)}. \tag{2.1}$$

The denominator in (2.1) ensure that $\mathcal{S}(X,Y)$ ranges from zero to one. The law of total variance immediately yields that $\mathcal{S}(X,Y)$ is equal to

$$1 - \frac{\int E\big[\text{var}\{1(Y \geq t) \mid X\}\big]\text{d}\mu(t)}{\int \text{var}\{1(Y \geq t)\}\text{d}\mu(t)}. \tag{2.2}$$

Dette, Siburg and Stoimenov (2013), Kong, Xia and Zhong (2019), and Chatterjee (2021) simply set T to be an independent copy of Y. For now, we allow T to be an arbitrary random variable, as long as $\text{supp}(Y) \subseteq \text{supp}(T)$. We revisit this issue in Study 1 of Section 5. We retain the asymmetry between X and Y in $\mathcal{S}(X,Y)$ deliberately in order to study which random variable impacts the other (Zheng, Shi and Zhang (2012); Cui, Li and Zhong (2015); Kong, Xia and Zhong (2019)).

Kong, Xia and Zhong (2019, Lemma 1) and Chatterjee (2021, Thm. 1) show that this metric possesses several desirable properties at the population level. For example, $\mathcal{S}(X,Y) = 0$ if and only if X and Y are independent, and $\mathcal{S}(X,Y) = 1$ if and only if Y is a measurable function of X. If (X,Y) is bivariate Gaussian with correlation coefficient ρ, then $\mathcal{S}(X,Y)$ is strictly increasing in $|\rho|$. In addition, $\mathcal{S}(X,Y)$ remains unchanged if we apply strictly monotone transformations to both X and Y.

2.2. The slicing procedure

Next, we discuss how to estimate $\mathcal{S}(X,Y)$ using a random sample $\{(X_i,Y_i), i=1,\ldots,n\}$. The literature offers two solutions, namely, kernel smoothing and rank estimation. Dette, Siburg and Stoimenov (2013) and Kong, Xia and Zhong (2019) suggest estimating $\text{var}\{1(Y \geq t) \mid X\}$ using kernel smoothing, for each given t. The overall complexity of estimating $\mathcal{S}(X,Y)$ using kernel smoothing is in $O(n^2)$ time, which limits its usefulness when n is extremely large. The asymptotic null distribution depends upon the kernel function, which is not desirable either. Chatterjee (2021) proposed a rank estimation for $\mathcal{S}(X,Y)$ that has complexity in $O(n\log n)$ time. In addition, the rank estimation is asymptotically standard normal. Using the rank estimation is thus more appealing than using kernel smoothing.

We introduce a slicing procedure to estimate $\mathcal{S}(X,Y)$, which proceeds as follows. We first order the random sample $\{(X_i,Y_i), i=1,\ldots,n\}$ according to the values of X_i, which yields $\{(X_{(i)},Y_{(i)}), i=1,\ldots,n\}$, where $X_{(1)} \leq \cdots \leq X_{(n)}$ are the ordered values of X_is, and $Y_{(i)}$ is the concomitant of $X_{(i)}$. Next, we divide the ordered sample $\{(X_{(i)},Y_{(i)}), i=1,\ldots,n\}$ into H slices according to the values of $X_{(i)}$, such that there are c observations within each slice. We assume, for simplicity, that $n = Hc$. We rewrite $X_{(h,j)} = X_{(c(h-1)+j)}$ and $Y_{(h,j)} = Y_{(c(h-1)+j)}$, for $j=1,\ldots,c$ and $h=1,\ldots,H$. The observations in the hth slice are $\{(X_{(h,j)},Y_{(h,j)}), j=1,\ldots,c\}$. Given t, we estimate $\text{var}\{1(Y \geq t) \mid X\}$ within each slice and $E\big[\text{var}\{1(Y \geq t) \mid X\}\big]$ with

$$
\begin{aligned}
&H^{-1}\sum_{h=1}^{H}\left[(c-1)^{-1}\sum_{j=1}^{c}\left\{1(Y_{(h,j)} \geq t) - c^{-1}\sum_{j=1}^{c}1(Y_{(h,j)} \geq t)\right\}^2\right] \\
&= \{n(c-1)\}^{-1}\sum_{h=1}^{H}\sum_{j<l}^{c}\left\{1(Y_{(h,j)} \geq t) - 1(Y_{(h,l)} \geq t)\right\}^2.
\end{aligned}
$$

Suppose $\{T_i, i=1,\ldots,n\}$ is a random sample drawn from $\mu(t)$. Let t run through the values of T_i, which allows us to estimate

$$
\int E\Big[\text{var}\{1(Y \geq t) \mid X\}\Big]\mathrm{d}\mu(t) \tag{2.3}
$$

in (2.2) with

$$
\{n^2(c-1)\}^{-1}\sum_{i=1}^{n}\sum_{h=1}^{H}\sum_{j<l}^{c}\{1(Y_{(h,j)} \geq T_i) - 1(Y_{(h,l)} \geq T_i)\}^2
$$

$$= \{n^2(c-1)\}^{-1} \sum_{h=1}^{H} \sum_{j<l}^{c} |r_{(h,j)} - r_{(h,l)}|, \tag{2.4}$$

where $r_{(h,j)}$ stands for the number of T_i such that $Y_{(h,j)} \geq T_i$, for $i = 1, \ldots, n$. Thus, we have $r_{(h,j)} = \#\{T_i : Y_{(h,j)} \geq T_i, i = 1, \ldots, n\}$.

Next, we turn to the denominator in (2.2). For each given t, we estimate $\text{var}\{1(Y \geq t)\}$ using the standard U-statistic theory van der Vaart (1998, Chap. 12). Specifically, we estimate the denominator in (2.2) as

$$\{n^2(n-1)\}^{-1} \sum_{i=1}^{n} \sum_{j<k}^{n} \{1(Y_j \geq T_i) - 1(Y_k \geq T_i)\}^2 = \{n^2(n-1)\}^{-1} \sum_{i=1}^{n} R_i(n - R_i),$$

where R_i stands for the number of Y_js such that $Y_j \geq T_i$. Then, we combine the above estimate with (2.4) to form a slicing estimation of $\mathcal{S}(X, Y)$, and denote $\widehat{\mathcal{S}}(X, Y)$ as

$$1 - (n-1)(c-1)^{-1} \frac{\sum_{h=1}^{H} \sum_{j<l}^{c} |r_{(h,j)} - r_{(h,l)}|}{\sum_{i=1}^{n} R_i(n - R_i)}. \tag{2.5}$$

The complexity of calculating $\widehat{\mathcal{S}}(X, Y)$ in (2.5) is $O\{n \log(n)\}$.

In the above estimation procedure, we assume implicitly that X is continuous. If X is categorical or discrete, taking H distinctive values, say, $X = 1, \ldots, H$, then we simply divide the random sample $\{(X_i, Y_i), i = 1, \ldots, n\}$ into H slices according to the distinctive levels of X. Observations for X_i that take the same value appear in the same slice. The number of observations within each slice is not necessarily the same. We estimate $\text{var}\{1(Y \geq t) \mid X\}$ within each slice, and aggregate over all H slices to form an estimate of (2.3). We omit the details for the present context.

The notion of a slicing estimation originated from Mardia, Kent and Bibby (1979, Chap. 12) and Li (1991). We adapt this concept to estimate $\mathcal{S}(X, Y)$. If there are only two observations within each slice, namely, $c = 2$, our slicing estimation reduces, in spirit, to the popular rank estimation of Chatterjee (2021). A similar observation is also made by Hsing and Carroll (1992). The slicing estimation introduces an annoying tuning parameter H, or equivalently, c. Thus, it is natural to ask what role H or c plays in the estimation or independence testing. This amounts to studying the theoretical properties of our proposed slicing estimation.

3. The Sliced Independence Test

In this section, we study the asymptotic properties of the slicing estimation by assuming that T is an independent copy of Y. In Study 1 of Section 5, we demonstrate that $\mu(t)$ has little impact on either the estimation or the testing.

We define a family of real-valued functions, $x \mapsto f(t;x)$, for $t \in \mathcal{T}$, to have a uniform total variation of order r over $\mathcal{T}$, if for any finite $B > 0$,

$$\lim_{n\to\infty} n^{-r} \sup_{t\in\mathcal{T},\Pi_n(B)} \sum_{i=1}^{n} |f(t;\widetilde{X}_{(i+1)}) - f(t;\widetilde{X}_{(i)})| = 0, \tag{3.1}$$

where $\Pi_n(B)$ is a collection of all possible n-point partitions of $[-B, B]$ such that $-B \le \widetilde{X}_{(1)} \le \cdots \le \widetilde{X}_{(n)} \le B$. Condition (3.1) is weaker than the uniform bounded variation condition, which requires

$$\lim_{n\to\infty} \sup_{t\in\mathcal{T},\Pi_n(B)} \sum_{i=1}^{n} |f(t;\widetilde{X}_{(i+1)}) - f(t;\widetilde{X}_{(i)})| < \infty.$$

If $f(t;x)$ has bounded first partial derivatives with respect to x on every finite interval, then condition (3.1) holds for any $r > 0$. We further define $x \mapsto f(t;x)$ to be nonexpansive in the metric of $M(x)$ on both sides of B_0 if there exists a nondecreasing real-valued function $M(x)$ and a real number $B_0 > 0$ such that, for any two points, say, $\widetilde{X}_1$ and $\widetilde{X}_2$, both in $(-\infty, -B_0]$ or both in $[B_0, \infty)$,

$$|f(t;\widetilde{X}_1) - f(t;\widetilde{X}_2)| \le |M(\widetilde{X}_1) - M(\widetilde{X}_2)|. \tag{3.2}$$

Let $\varepsilon(t;X) \overset{\text{def}}{=} 1(Y \ge t) - s(t;X)$ and $V(t_1,t_2;X) \overset{\text{def}}{=} \text{cov}\{\varepsilon(X,t_1), \varepsilon(X,t_2) \mid X\}$. We assume the following two conditions on $s(t;x)$ and $V(t_1,t_2;x)$:

(C1) Assume that $x \mapsto s(t;x)$ has a uniform total variation of order $r = 1/2$ and is nonexpansive in the metric of $M(x)$ on both sides of a real number $B_0 > 0$, such that $M^2(x)\text{pr}(X > x) \to 0$ as $x \to \infty$.

(C2) Suppose that $x \mapsto V(t_1,t_2;x)$ has a uniform total variation of order $r = 1$.

These conditions are related to the variation and tail behavior of $s(t;x)$ and $V(t_1,t_2;x)$, and are typically regarded as mild and are popular in the literature; see, for example, Hsing and Carroll (1992), Zhu and Ng (1995), Zhu, Miao and Peng (2006), Li and Zhu (2007), Lin, Zhao and Liu (2018), and Kong, Xia and Zhong (2019).

Let Y, T, T_1, and T_2 be independent copies, and let " $\xrightarrow{d}$ " denote converge in distribution. Define $\theta_1 \overset{\text{def}}{=} E\{V(T_1,T_2;X)^2\}$, $\theta_2 \overset{\text{def}}{=} E\big[\text{var}\{1(Y \ge T) \mid T\}\big]$,

$\sigma^2 \overset{\text{def}}{=} 2E\big[\text{cov}^2\{1(Y_1 \geq T), 1(Y_2 \geq T) \mid T\}\big]/\theta_2^2$, and $\tau^2 \overset{\text{def}}{=} \{\zeta_1 + 2\theta_1/(c-1)\}/\theta_2^2$, where $\zeta_1 > 0$ is defined in (S.1.2) of the online Supplementary Material.

Theorem 1. *Assume the number of observations within each slice, c, is fixed.*

(i) If X and Y are independent, then $\{n(c-1)\}^{1/2}\widehat{\mathcal{S}}(X,Y) \xrightarrow{d} \mathcal{N}(0,\sigma^2)$ *as* $n \to \infty$*. In particular, if Y is a continuous random variable, then* $\sigma^2 = 4/5$.

(ii) If X and Y are not independent, then under Conditions (C1)–(C2), $n^{1/2}\{\widehat{\mathcal{S}}(X,Y) - \mathcal{S}(X,Y)\} \xrightarrow{d} \mathcal{N}(0,\tau^2)$ *as* $n \to \infty$.

Theorem 1 has several important implications. In particular, the slicing estimation is root-n consistent and asymptotically normal for an arbitrary constant $c \geq 2$. The larger c is, the smaller is the asymptotic variance. We reject the null hypothesis H_0: X and Y are independent if $n^{1/2}\widehat{\mathcal{S}}(X,Y)/\sigma \geq z_{1-\alpha}$ at the significance level α, where $z_{1-\alpha}$ is the $(1-\alpha)$th quantile of the standard normal distribution. Let $\Phi(\cdot)$ be the cumulative distribution function of the standard normal distribution. The asymptotic power is $1 - \Phi\big[\{z_{1-\alpha}\,\sigma - n^{1/2}\mathcal{S}(X,Y)\}/\tau\big]$, which is equal to

$$\Phi\left(\theta_2\mathcal{S}(X,Y)\left[\frac{n}{\zeta_1 + 2\theta_1/(c-1)}\right]^{1/2} - \theta_2 z_{1-\alpha}\left[\frac{4}{5(c-1)\zeta_1 + 10\theta_1}\right]^{1/2}\right). \qquad (3.3)$$

This is a strictly monotone increasing function of c. In other words, the larger c is, the more powerful the proposed test becomes. The rank test of Chatterjee (2021) corresponds to the sliced independence test with $c = 2$, indicating that, in general, the sliced independence test is more powerful than the rank test.

The asymptotic power function in (3.3) inspires us to ask whether we can enhance the power performance of the sliced independence test if we allow $c \to \infty$ as $n \to \infty$. To this end, we assume the following conditions:

(C1*) Assume that $x \mapsto s(t;x)$ has a uniform total variation of order $r > 0$ and is nonexpansive in the metric of $M(x)$ on both sides of a real number $B_0 > 0$ such that $M^{2+b}(x)\text{pr}(X > x) \to 0$, for $b > 0$, as $x \to \infty$.

(C2*) Let $c = O(n^{\alpha})$, where $\alpha = 1/2 - \max\{r, 1/(2+b)\}$.

These conditions are even weaker than (C1) and (C2). Letting c diverge to infinity, we relax the smoothness condition on $x \mapsto s(t;x)$ slightly and avoid assuming smoothness conditions on $x \mapsto V(t_1,t_2;x)$. If $x \mapsto s(t;x)$ is L-Lipschitz continuous and X is sub-Gaussian or has a bounded support, then conditions (C1*) and (C2*) hold for any $r > 0$, $b > 0$, $M(x) = Lx$, and $c = o(n^{1/2})$.

Define $\tau_*^2 \overset{\text{def}}{=} \zeta_1/\theta_2^2$, which is smaller than $\tau^2 = \{\zeta_1 + 2\theta_1/(c-1)\}/\theta_2^2$.

Theorem 2. *Assume the number of observations within each slice, c, diverges.*

(i) If X and Y are independent and $c = o(n)$, *then* $(nc)^{1/2}\widehat{\mathcal{S}}(X,Y) \xrightarrow{d} \mathcal{N}(0,\sigma^2)$ *as* $n, c \to \infty$. *In particular,* $\sigma^2 = 4/5$ *if Y is continuous.*

(ii) If X and Y are not independent, under Conditions (C1)–(C2*),* $n^{1/2}\{\widehat{\mathcal{S}}(X,Y) - \mathcal{S}(X,Y)\} \xrightarrow{d} \mathcal{N}(0,\tau_*^2)$ *as* $n, c \to \infty$.

The sliced estimation converges at the faster rate of $(nc)^{-1/2}$ than the rank estimate (Chatterjee (2021)). The asymptotic variance decreases as c increases. At the significance level α, the asymptotic power is

$$\Phi\left[\theta_2\mathcal{S}(X,Y)\left(\frac{n}{\zeta_1}\right)^{1/2} - \theta_2 z_{1-\alpha}\left\{\frac{4}{5c\zeta_1}\right\}^{1/2}\right],$$

which again increases with c. The power improvement of the sliced independence test over the rank test (Chatterjee (2021)) is substantial when c diverges to infinity.

4. An Extension to Multivariate Control Variables

In this section, we generalize the concept of slicing using the K-means clustering procedure (MacQueen (1967)) to account for the presence of multivariate control variables. We use the random vector $\mathbf{x} = (X_1, \ldots, X_p)^{\mathrm{T}} \in \mathbb{R}^p$ to replace the univariate control variable X in $\mathcal{S}(X,Y)$, which leads to

$$\mathcal{S}(\mathbf{x},Y) \stackrel{\text{def}}{=} \frac{\int \mathrm{var}\{s(t;\mathbf{x})\}\mathrm{d}\mu(t)}{\int \mathrm{var}\{1(Y \geq t)\}\mathrm{d}\mu(t)},$$

where $s(t;\mathbf{x}) \stackrel{\text{def}}{=} \mathrm{pr}(Y \geq t \mid \mathbf{x})$. Similarly, we can verify that $\mathcal{S}(\mathbf{x},Y)$ is equal to

$$1 - \frac{\int E\big[\mathrm{var}\{1(Y \geq t) \mid \mathbf{x}\}\big]\mathrm{d}\mu(t)}{\int \mathrm{var}\{1(Y \geq t)\}\mathrm{d}\mu(t)}. \tag{4.1}$$

Both $\mathcal{S}(X,Y)$ and $\mathcal{S}(\mathbf{x},Y)$ share the zero-independence equivalency property at the population level. However, the slicing procedure used to estimate $\mathcal{S}(X,Y)$ cannot be directly used to estimate $\mathcal{S}(\mathbf{x},Y)$, unless the sorting algorithm is delicately adapted to account for multivariate observations.

Suppose a random sample $\{(\mathbf{x}_i, Y_i), i = 1, \ldots, n\}$ is available. Instead of using the slicing procedure, in this section, we propose using the K-means clustering approach (MacQueen (1967)) to partition the random sample into H clusters according to the realizations of the control variables, $\{\mathbf{x}_i, i = 1, \ldots, n\}$. We estimate $\mathrm{var}\{1(Y \geq t) \mid \mathbf{x}\}$ within each cluster, and aggregate the resultant estimates to form an estimate of $\mathcal{S}(\mathbf{x}, Y)$.

We implement the K-means clustering approach according to $\{\mathbf{x}_i, i = 1, \ldots, n\}$ only, which proceeds as follows:

1. Randomly choose H points in $\{\mathbf{x}_i, i = 1, \ldots, n\}$ as the initial centers.
2. For each center, identify the points in $\{\mathbf{x}_i, i = 1, \ldots, n\}$ that are "closer" to it than any other center. Update the centers of all clusters.
3. Iterate the above step until convergence.
4. Delete the clusters with a single data point and repeat all of the above steps.
5. Either (a) absorb the data points in the previously deleted clusters into the cluster with the nearest center and terminate, or (b) terminate without the data points in the deleted clusters.

The last two steps avoid the presence of clusters with a single data point. We implement this K-means clustering approach to partition the whole random sample $\{(\mathbf{x}_i, Y_i), i = 1, \ldots, n\}$ into H clusters, according to the realizations of the control variables, $\{\mathbf{x}_i, i = 1, \ldots, n\}$. The K-means clustering approach cannot guarantee that each cluster contains an equal number of observations. Therefore, we assume that the hth cluster consists of n_h observations, for $h = 1, \ldots, H$. We re-index the random sample as $\{(\mathbf{x}_{(h,j)}, Y_{(h,j)}), j = 1, \ldots, n_h, h = 1, \ldots, H\}$, and estimate

$$\int E\Big[\mathrm{var}\{1(Y \geq t) \mid \mathbf{x}\}\Big] \mathrm{d}\mu(t)$$

in (4.1) using a weighted summation, as follows:

$$n^{-1} \sum_{i=1}^{n} \sum_{h=1}^{H} \frac{n_h}{n} \left[\sum_{j<l}^{n_h} \frac{\{1(Y_{(h,j)} \geq T_i) - 1(Y_{(h,l)} \geq T_i)\}^2}{n_h(n_h - 1)} \right].$$

Recall that $r_{(h,j)}$ stands for the number of T_i such that $Y_{(h,j)} \geq T_i$, for $i = 1, \ldots, n$. It is straightforward to verify that the above is equal to

$$n^{-2} \sum_{h=1}^{H} \sum_{j<l}^{n_h} \frac{|r_{(h,j)} - r_{(h,l)}|}{n_h - 1}.$$

This motivates us to define

$$\widehat{\mathcal{S}}(\mathbf{x}, Y) \stackrel{\mathrm{def}}{=} 1 - \frac{\sum_{h=1}^{H} \sum_{j<l}^{n_h} |r_{(h,j)} - r_{(h,l)}|/(n_h - 1)}{\sum_{i=1}^{n} R_i(n - R_i)/(n - 1)},$$

where R_i is defined in Section 2. We further define

$$c_n^{-1} \stackrel{\text{def}}{=} \sum_{h=1}^{H} \frac{n_h}{\{n(n_h-1)\}},$$

which is equal to $1/(c-1)$ if $n_h = c$, for all $h = 1, \ldots, H$. To study the asymptotic behavior of $\widehat{\mathcal{S}}(\mathbf{x}, Y)$ when $\mathbf{x}$ and Y are not independent, we assume the following two conditions:

(C3) There exist two positive constants, C_1 and C_2, such that $\mathrm{pr}(\|\mathbf{x}\| > t) \leq C_1 \exp(-C_2 t^2)$, for all $t \in \mathbb{R}$.

(C4) There exists a positive constant C_3 such that $|s(t; \mathbf{x}_1) - s(t; \mathbf{x}_2)| \leq C_3 \|\mathbf{x}_1 - \mathbf{x}_2\|$, for all $t \in \mathbb{R}$ and $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^p$.

Condition (C3) requires that $\mathbf{x}$ be sub-Gaussian, and condition (C4) concerns the smoothness of $x \mapsto s(t; x)$.

Theorem 3. *Assume the number of slices, H, diverges.*

(i) If $\mathbf{x}$ and Y are independent, $(nc_n)^{1/2}\widehat{\mathcal{S}}(\mathbf{x}, Y) \xrightarrow{d} \mathcal{N}(0, \sigma^2)$ as $n \to \infty$. In particular, $\sigma^2 = 4/5$ if Y is continuous.

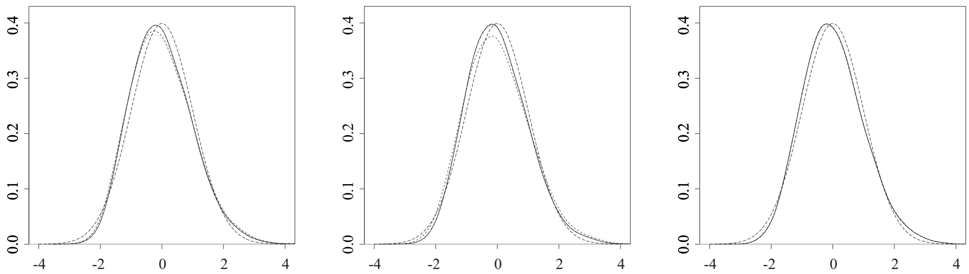
(ii) If $\mathbf{x}$ and Y are not independent and $H = O(n^\delta)$, for some $0 < \delta \leq 1$, under conditions (C3)–(C4), $\widehat{\mathcal{S}}(\mathbf{x}, Y)$ converges in probability to $\mathcal{S}(\mathbf{x}, Y)$ and, accordingly, $(nc_n)^{1/2}\widehat{\mathcal{S}}(\mathbf{x}, Y) \to \infty$ as $n \to \infty$.

5. Numerical Studies

5.1. Simulations

We first demonstrate the finite-sample performance of the slicing estimation and the sliced independence test by means of simulations.

Study 1. The definition of (2.1) involves a probability measure $\mu(t)$. We evaluate the effect of $\mu(t)$ on the asymptotic null distribution. We draw X_i and Y_i independently from uniform, standard normal, and $t(1)$ distributions. We fix $n = 1024$ and $c = 32$. We consider four choices for $\mu(t)$: (i) $T_i = Y_i$; (ii) $T_i \sim \mathcal{N}(0, 1)$; (iii) $T_i \sim t(1)$; and (iv) T_i is a bootstrap sample of Y_i. We replicate each scenario 10,000 times, and draw the kernel density functions of $Z \stackrel{\text{def}}{=} n^{1/2}\widehat{\mathcal{S}}(X, Y)/\sigma$ in Figure 1. All of the kernel densities are relatively close to the reference curve $\mathcal{N}(0, 1)$. This is not surprising, because the indicator functions in the slicing estimation (2.4) vary only at Y_i. The figure also shows that the asymptotic null distribution does not depend on the parent distribution of (X, Y).



(a) $(X,Y) \overset{\text{i.i.d.}}{\sim} \text{unif}(-1,1)$ (b) $(X,Y) \overset{\text{i.i.d.}}{\sim} t(1)$ (c) $(X,Y) \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$

Figure 1. The kernel densities with different choices for $\mu(t)$s: (i) $T_i = Y_i$ (solid), (ii) $T_i \overset{\text{i.i.d.}}{\sim} N(0,1)$ (dashed), (iii) $T_i \overset{\text{i.i.d.}}{\sim} t(1)$ (dotted), and (iv) T_i is a bootstrap sample (dotdash). The density function of the standard normal distribution is used as a reference curve (longdash).

Study 2. Next, we evaluate how the number of observations within each slice, c, affects the resulting slicing estimation. We generate $\varepsilon \sim N(0,1)$ and $X \sim$ uniform$(-1,1)$ independently, and consider the following six dependent structures:

(A) Log: $Y = C_1 \log(X^2) + \lambda\varepsilon$.

(B) Circular: $Y = Z(1-X^2)^{1/2} + \lambda C_2 \varepsilon$, where Z is independent of X and $\text{pr}(Z = \pm 1) = 1/2$.

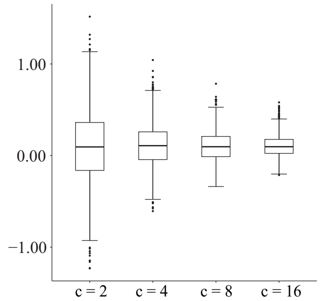
(C) W-shaped: $Y = |X + 0.5| 1(X < 0) + |X - 0.5| 1(X \geq 0) + \lambda C_3 \varepsilon$.

(D) Sinusoid: $Y = \cos(C_4 \pi X) + 3\lambda\varepsilon$.

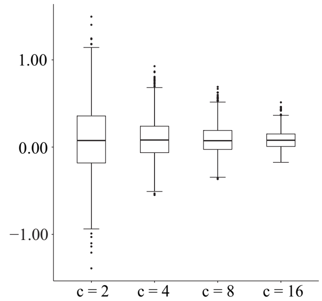
(E) Doppler: $Y = \{X^2(1-X^2)\}^{1/2} \sin(1.05\pi/X^2) + \lambda C_5 \varepsilon$.

(F) HeaviSine: $Y = 4\sin(4\pi X^2) - \text{sign}(X^2 - 0.3) - \text{sign}(0.72 - X^2) + \lambda C_6 \varepsilon$.

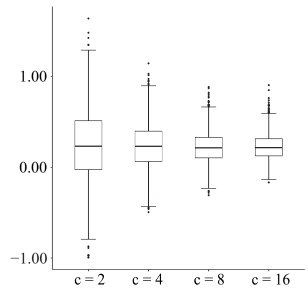
These structures have been used in similar contexts; see, for example, Chatterjee (2021), Heller, Heller and Gorfine (2013), Kong, Xia and Zhong (2019), and Donoho and Johnstone (1995). In this study, we fix $(C_1, \ldots, C_6) = (0.05, 0.9, 0.75, 8, 1.5, 24)$, $\lambda = 0.7$, and $n = 512$, and vary $c \in \{2, 4, 8, 16\}$. We replicate each scenario 10,000 times. Box plots of the resultant slicing estimation with different c values are shown in Figure 2. Clearly, in terms of the median values of the slicing estimates, $\widehat{\mathcal{S}}(X,Y)$ converges to $\mathcal{S}(X,Y)$ across all scenarios. However, the variances of $\widehat{\mathcal{S}}(X,Y)$ decrease substantially as c increases, supporting our theoretical results in Theorems 1 and 2.



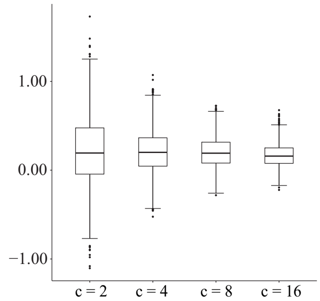
(a) Log



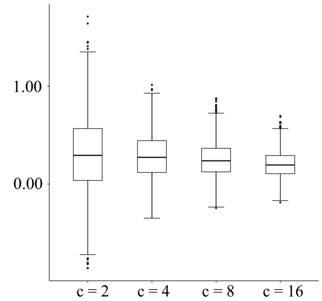
(b) Circular



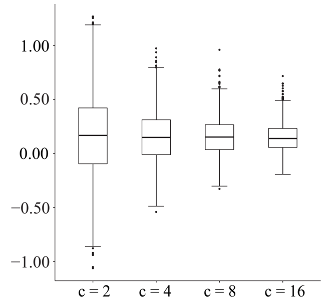
(c) W-shaped



(d) Sinusoid



(e) Doppler

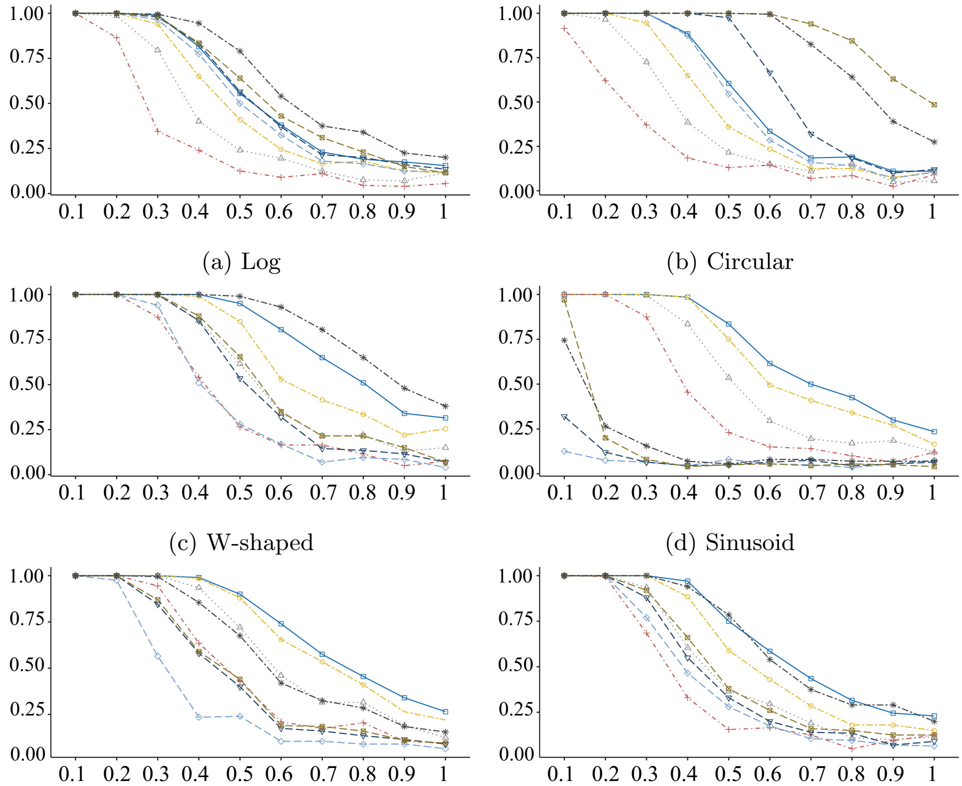


(f) HeaviSine

Figure 2. Box plot of $\widehat{\mathcal{S}}(X, Y)$ with $c \in \{2, 4, 8, 16\}$ in Study 2.

Study 3. We use the dependence structures in Study 2 to compare the power performance of our proposed sliced independence test with that of the modified Blum–Kiefer–Rosenblatt correlation test (Zhou and Zhu (2018)), distance correlation test (Székely, Rizzo and Bakirov (2007)), multivariate test of Heller, Heller and Gorfine (2013), and composite coefficient of determination test of Kong, Xia and Zhong (2019). Note that the composite coefficient of determination is estimated using kernel smoothing, which is computationally intensive. We use 200 random permutations to approximate the asymptotic null distributions for the last three tests. We fix $n = 512$, and vary $c \in \{2, 4, 8, 16\}$ and $\lambda\ = 0 : 0.1 : 1$. We report the empirical power of each test at the significance level $\alpha = 0.05$ in Figure 3. Our proposed tests appears to be superior to its competitors in the oscillatory cases, that is the Sinusoid, HeaviSine, and Doppler structures. Furthermore, as c increases, the empirical power of our proposed test improves accordingly. This again confirms the theoretical results in Theorems 1 and 2.

Study 4. Next, we compare the running times of several popular independence tests in Study 3. We implement the multivariate test of Heller, Heller and Gorfine (2013) using the R package `HHG`, and the composite coefficient of determination



(a) Log (b) Circular

(c) W-shaped (d) Sinusoid

(e) Doppler (f) HeaviSine

Figure 3. The empirical power of four independence tests: the sliced independence test with $c = 2$ ($+$), $c = 4$ ($\triangle$), $c = 8$ ($\circ$), and $c = 16$ ($\square$); the distance correlation test ($\Diamond$); the modified Blum–Kiefer–Rosenblatt correlation test (∇); the multivariate test of Heller, Heller and Gorfine (2013) ($\boxtimes$); and the composite coefficient of determination test of Kong, Xia and Zhong (2019) ($*$). The horizontal axis represents λ, and the vertical axis represents the empirical power.

test with the R code provided by Dr Zhong Wei, one of the authors of Kong, Xia and Zhong (2019). Implementing these two tests is very time consuming. We terminate them when their implementations take more than 30 minutes. We also include three versions of the distance correlation test in the comparison, which are available in the R packages `energy`, `kpcalg`, and `dcov`, respectively. The first is the classic version of the distance correlation test, referred to as DC_1 in Table 1. We refer to the last two versions as DC_2 and DC_3, respectively. For the DC_2 test, the asymptotic null distribution of the distance correlation test is approximated by a gamma approximation in the R package `kpcalg`, where the function `dcov.gamma()` is used. In the DC_3 test, the distance correlation is estimated using

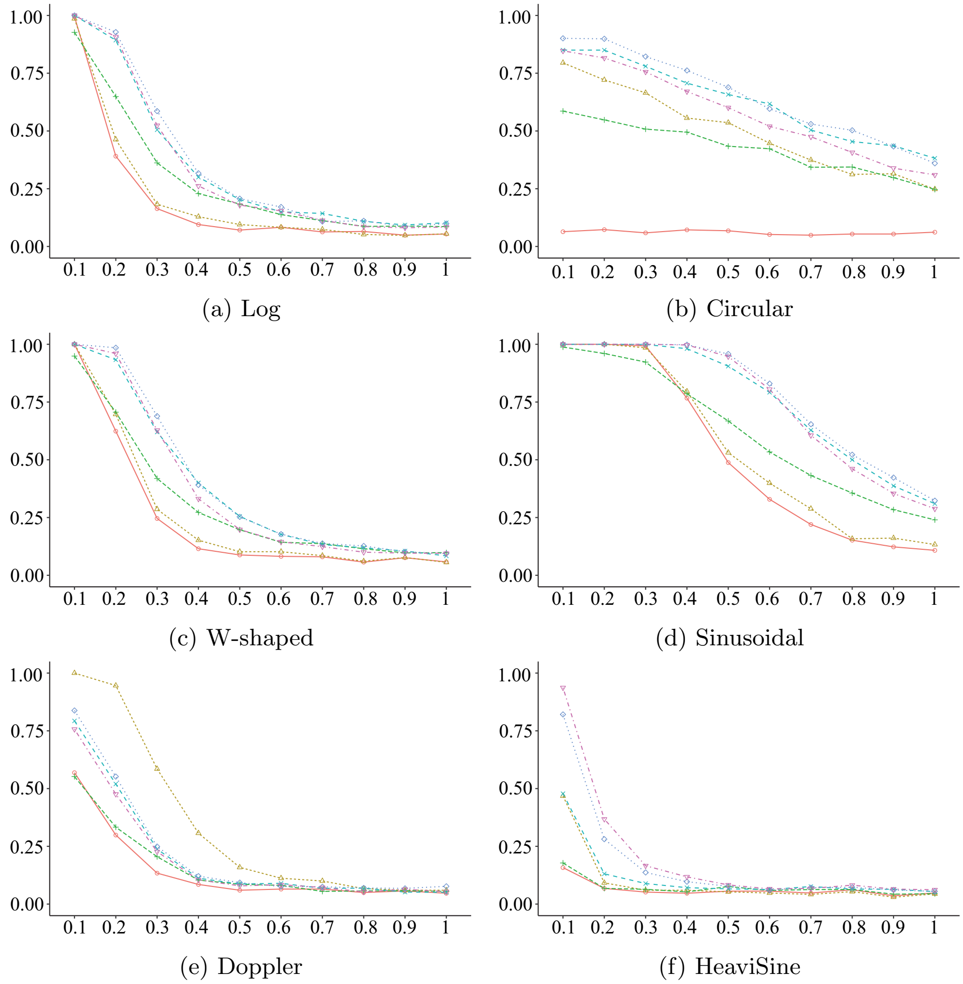
Table 1. The average wall-clock time (in seconds) over 100 replications for three versions of the distance correlation test (DC_1, DC_2, and DC_3), the modified Blum–Kiefer–Rosenblatt correlation test (MBKR), the multivariate test of Heller, Heller and Gorfine (2013) (HHG), the composite coefficient of determination test (CCD), and the sliced independence test (SIT).

n	DC_1	DC_2	DC_3	MBKR	HHG	CCD	SIT
128	0.006	0.032	0.0004	0.004	0.089	1.395	0.00014
256	0.034	0.037	0.0014	0.039	0.248	9.995	0.00017
512	0.109	0.559	0.0034	0.184	0.849	45.934	0.00031
1,024	0.812	0.989	0.0174	1.684	3.834	210.100	0.00056
2,048	4.540	3.502	0.0708	13.253	14.580	575.938	0.00114
4,096	15.823	12.305	0.2249	116.463	$>$ 30mins	$>$ 30mins	0.00215
8,192	63.869	51.024	1.3841	899.942	$>$ 30mins	$>$ 30mins	0.00434

the algorithm proposed by Huo and Székely (2016), which is computationally very efficient. To further speed up the DC_3 test, we also use the gamma approximation in the R package `dcov`. In the sliced independence test, we fix the number of slices as $c = 16$, and vary the sample size $n \in \{128, 256, 512, 1024, 2048, 4096, 8192\}$. We summarize the averages of the wall-clock time in Table 1, based on 100 replications. The sliced independence test runs the fastest, followed by the DC_3 test. These two tests have the smallest order of complexity, and thus are much more efficient numerically than all other competitors.

Next, we conduct a simulation study with multivariate control variables. Instead of using the slicing estimation procedure, we use the K-means clustering approach to classify the observations into H clusters.

Study 5. Let $\mathbf{x} = (X_1, \ldots, X_5)^{\mathrm{T}}$. We generate X_k independently from the uniform distribution defined on the interval $[-1, 1]$, for $k = 1, \ldots, 5$, and ε from the standard normal distribution. Denote $m(\mathbf{x}) \stackrel{\text{def}}{=} (X_1 + \ldots + X_5)/5$. We use the simulated examples used in Study 2, but with X replaced by $m(\mathbf{x})$ throughout. We set $(C_1, \ldots, C_6) = (0.05, 0.05, 0.75, 2, 0.5, 24)$ in this study, and vary $H \in \{8, 16, 32, 64\}$ for our proposal. We include the distance correlation test (Székely, Rizzo and Bakirov (2007) and the multivariate test of Heller, Heller and Gorfine (2013) in our comparison. The sample size is fixed at $n = 512$. We vary $\lambda = 0 : 0.1 : 1$, and replicate each experiment 1,000 times to compare the power performance of the tests. The significance level is fixed at $\alpha = 0.05$, and the simulated results are summarized in Figure 4. In this simulation study, the proposed test is clearly still more powerful than other tests, except for the Doppler case.



(a) Log (b) Circular

(c) W-shaped (d) Sinusoidal

(e) Doppler (f) HeaviSine

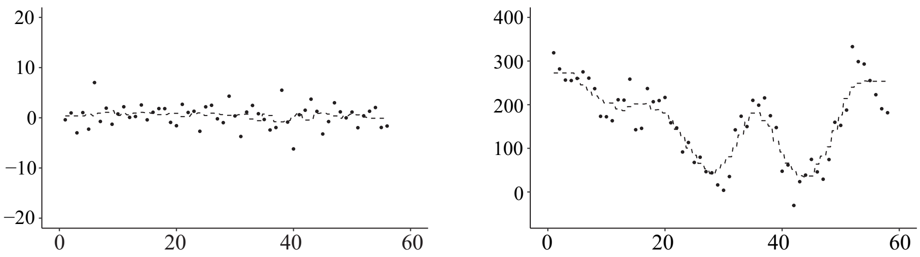
Figure 4. The empirical power of four independence tests: the sliced independence test with $H = 8$ $(+)$, $H = 16$ $(\times)$, $H = 32$ $(\Diamond)$, and $H = 64$ (∇), the distance correlation test $(\circ)$, and the multivariate test of Heller, Heller and Gorfine (2013) $(\triangle)$. The horizontal and vertical axes represent λ and the empirical power, respectively.

5.2. Real-data analysis

We apply the sliced independence test to an astronomical data set. The Photometric LSST Astronomical Time-series Classification Challenge (PLAsTiCC) data set is available on `https://www.kaggle.com/c/PLAsTiCC-2018`. This is a simulated data set and consisting of 15 classes. We consider the r band in classes 65 and 88 only. There are 981 objects in class 65 and 370 objects in class 88. For each object, the number of observations, n, ranges from 10 to 60. We remove objects with fewer than 30 observations, leaving 313 objects in class 65 and 119

Table 2. The number of times that the null hypothesis, H_0: X and Y are independent, is rejected at the significance level $\alpha = 0.05$. The distance correlation test, modified Blum–Kiefer–Rosenblatt correlation test, multivariate test of Heller, Heller and Gorfine (2013), and sliced independence test are denoted by DC, MBKR, HHG, and SIT, respectively.

	DC	MBKR	HHG	SIT	SIT
class				$c = 2$	$c = 4$
65	16	11	17	12	10
88	115	118	119	119	119



(a) class 65 (b) class 88

Figure 5. Scatter plots of the intensity of brightness (on the vertical axis) over time (on the horizontal axis) for one representative object in each class. The dashed line is fitted using a k-nearest neighbor regression ($k = 7$).

objects in class 88. The target is to examine whether the intensity of brightness (Y) varies over time (X) for each object in these two classes.

We apply the sliced independence test with $c = 2$ and $c = 4$, the distance correlation test (Székely, Rizzo and Bakirov (2007)), the modified Blum–Kiefer–Rosenblatt correlation test (Zhou and Zhu (2018)), and the multivariate test of Heller, Heller and Gorfine (2013) to this data set. In Table 2, we report the number of times that we reject the null hypothesis, H_0: X and Y are independent, at the significance level $\alpha = 0.05$. The intensity of brightness does not change over time for more than 95% of the objects in class 65. In contrast, the independence tests all strongly indicate that, for almost all objects in class 88, the intensity of the brightness changes over time.

We present the intensity of the brightness of two representative objects, one from each class, in Figure 5, which echoes the results in Table 2. In class 65, most objects exhibit a similar pattern that the brightness intensity remains unchanged over time. In contrast, for most objects in class 88, the brightness intensity varies over time.

6. Conclusion

We have introduced a slicing procedure for estimating a popular measure of nonlinear dependence. The resultant sliced independence test encompasses the rank test as a special case, has almost the minimal computational complexity, and is asymptotically distribution free. We show that as the number of observations within each class increases, the asymptotic variance of the slicing estimation decreases and the power of the independence test improves. In addition, the size performance of the sliced independence test is insensitive to the number of slices. The slicing estimation is consistent for a wide range of slice numbers. We also generalize the concept of slicing using K-means clustering to account for multivariate control variables. Generalizing this concept further is challenging if both random variables are multivariate, and is left to future research.

Supplementary Material

The proofs of Theorems 1–3 are relegated to the Supplementary Material.

Acknowledgments

Zhu is the corresponding author, and his research was supported by the National Natural Science Foundation of China (12171477, 11731011 and 11931014). Zhang's research is supported by the Fundamental Research Fund for Central Universities and the Research Fund of Renmin University of China (21XNH157).

References

Blum, J. R., Kiefer, J. and Rosenblatt, M. (1961). Distribution free tests of independence based on the sample distribution function. *The Annals of Mathematical Statistics* **32**, 485–498.

Cao, S. and Bickel, P. J. (2020). Correlations with tailored extremal properties. *arXiv: 2008.10177*.

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

Chaudhuri, A. and Hu, W. (2019). A fast algorithm for computing distance correlation. *Computational Statistics & Data Analysis* **135**, 15–24.

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

Dette, H., Siburg, K. F. and Stoimenov, P. A. (2013). A copula-based non-parametric measure of regression dependence. *Scandinavian Journal of Statistics* **40**, 21–41.

Donoho, D. L. and Johnstone, I. M. (1995). Adapting to unknown smoothness via wavelet shrinkage. *Journal of the American Statistical Association* **90**, 1200–1224.

Gao, L., Fan, Y., Lv, J. and Shao, Q.-M. (2021). Asymptotic distributions of high-dimensional distance correlation inference. *The Annals of Statistics* **49**, 1999–2020.

Heller, R., Heller, Y. and Gorfine, M. (2013). A consistent multivariate test of association based on ranks of distances. *Biometrika* **100**, 503–510.

Hoeffding, W. (1948). A non-parametric test of independence. *The Annals of Mathematical Statistics* **19**, 546–557.

Hsing, T. and Carroll, R. J. (1992). An asymptotic theory for sliced inverse regression. *The Annals of Statistics* **20**, 1040–1061.

Huang, C. and Huo, X. (2022). A statistically and numerically efficient independence test based on random projections and distance covariance. *Frontiers in Applied Mathematics and Statistics* **7**, 1–17.

Huo, X. and Székely, G. J. (2016). Fast computing for distance covariance. *Technometrics* **58**, 435–447.

Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika* **30**, 81–93.

Kong, E., Xia, Y. and Zhong, W. (2019). Composite coefficient of determination and its application in ultrahigh dimensional variable screening. *Journal of the American Statistical Association* **114**, 1740–1751.

Li, K.-C. (1991). Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association* **86**, 316–327.

Li, Y. and Zhu, L.-X. (2007). Asymptotics for sliced average variance estimation. *The Annals of Statistics* **35**, 41–69.

Lin, Q., Zhao, Z. and Liu, J. S. (2018). On consistency and sparsity for sliced inverse regression in high dimensions. *The Annals of Statistics* **46**, 580–610.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* (Edited by L. M. Le Cam and J. Neyman), 281–297. Univ. of California Press, Berkeley.

Mardia, K. V., Kent, J. T. and Bibby, J. M. (1979). *Multivariate Analysis*. Probability and Mathematical Statistics. Academic Press, London.

Pearson, K. (1895). Notes on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London* **58**, 240–242.

Spearman, C. (1906). 'Footrule' for measuring correlation. *British Journal of Psychology, 1904-1920* **2**, 89–108.

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.

van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press, Cambridge.

Weihs, L., Drton, M. and Leung, D. (2016). Efficient computation of the Bergsma–Dassios sign covariance. *Computational Statistics* **31**, 315–328.

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

Wiesel, J. (2021). Measuring association with Wasserstein distances. *arXiv:2102.00356 [math, stat]*.

Zheng, S., Shi, N.-Z. and Zhang, Z. (2012). Generalized measures of correlation for asymmetry, nonlinearity, and beyond. *Journal of the American Statistical Association* **107**, 1239–1252.

Zhou, Y. and Zhu, L. (2018). Model-free feature screening for ultrahigh dimensional datathrough a modified Blum-Kiefer-Rosenblatt correlation. *Statistica Sinica* **28**, 1351–1370.

Zhu, L., Miao, B. and Peng, H. (2006). On sliced inverse regression with high-dimensional covariates. *Journal of the American Statistical Association* **101**, 630–643.

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

Zhu, L.-X. and Ng, K. W. (1995). Asymptotics of sliced inverse regression. *Statistica Sinica* **5**, 727–736.

Yilin Zhang

Center for Applied Statistics and Institute of Statistics and Big Data, Renmin University of China, Beijing 100872, China.

E-mail: yzhang97@ruc.edu.cn

Canyi Chen

Center for Applied Statistics and Institute of Statistics and Big Data, Renmin University of China, Beijing 100872, China.

E-mail: chency1997@ruc.edu.cn

Liping Zhu

Center for Applied Statistics and Institute of Statistics and Big Data, Renmin University of China, Beijing 100872, China.

E-mail: zhu.liping@ruc.edu.cn

(Received May 2021; accepted April 2022)