Statistica Sinica Preprint No: SS-2017-0538	
Title	A new class of measures for testing independence
Manuscript ID	SS-2017-0538
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202017.0538
Complete List of Authors	Xiangrong Yin and Qingcong Yuan
Corresponding Author	Xiangrong Yin
E-mail	yinxiangrong@uky.edu

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0538

A NEW CLASS OF MEASURES FOR TESTING INDEPENDENCE

Xiangrong Yin* and Qingcong Yuan†

August 14, 2019

Abstract: We introduce a new class of measures that test for independence between two random vectors using the expected difference between conditional and marginal characteristic functions. Based on a selected weight function in the class, we propose a new index for measuring independence and study its properties. To illustrate the use of such an index, two empirical versions are developed: slicing and kernel approaches. Their asymptotic properties and applications are discussed. Lastly, simulation results demostrate the advantages of the proposed method.

Key words and phrases: Categorical variable, distance, independence.

1 Introduction

Measuring and testing independence are important in statistics. The classical Pearson product-moment correlation and covariance measure the linear dependence be-

*Department of Statistics, University of Kentucky, 725 Rose St. Lexington, Kentucky, 40536-0082. E-mail: yinxiangrong@uky.edu.

†Department of Statistics, Miami University, Oxford, Ohio, 45056. E-mail: qingcong.yuan@miamioh.edu.

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0538

tween two random variables. In a multivariate normal setting, a diagonal covariance matrix implies independence, but this does not generalize to all settings. Likelihood-based methods, such as Wilks' lambda (Wilks (1935)) or that of Puri and Sen (1971) cannot be applied if the dimension exceeds the sample size, or if the distributional assumptions do not hold. In the latter case, multivariate nonparametric approaches have been proposed by Taskinen, Oja, and Randles (2005). Furthermore, numerous studies have examined measuring independence, including those of Blomqvist (1950), Blum, Kiefer, and Rosenblatt (1961), Hollander and Wolfe (1999), and Anderson (2003). Székely, Rizzo, and Bakirov (2007) proposed a novel distance covariance (dCov) to test for independence between two random vectors of arbitrary dimensions. This test is very useful because it is nonparametric, but free of tuning parameters. As a result, it is also widely used in other areas, such as variable selection (Li, Zhong, and Zhu (2012)) and dimension reduction (Sheng and Yin (2013)). Huo and Székely (2016) developed a fast algorithm for dCov. Finally, Heller, Heller, and Gofine (2013) proposed a novel multivariate test of association that effectively deals with continuous and discrete random vectors, but may have trouble handling nominal random vectors, owing to its use of ranks.

Dependency measures defined by density-based divergence families can flexibly handle *correlation-type* or *conditional-type* relationships between two vectors, say, X and Y. For example, the ϕ-divergence family (Vajda (1989)) typically involves a term $g(P,Q) = \frac{dP}{dQ}$, where P and Q are distributions, for which we can define the Kullback–Leibler distance (KL-distance) as

$$E\left[\log \frac{f(X,Y)}{f(X)f(Y)}\right] = E\left[\log \frac{f(X|Y)}{f(X)}\right] = E\left[\log \frac{f(Y|X)}{f(Y)}\right]. \tag{1}$$

Here, $f(\cdot)$ is a density, or $g(P,Q) = \frac{f(x,y)}{f(x)f(y)}$, where $dP = f(x,y)$ and $dQ = f(x)f(y)$, which appear in the first term in (1). The discrepancy is calculated using the ratio of the joint distribution and the product of the two marginal distributions (first term in (1)), or the ratio of a conditional distribution and a marginal distribution

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0538

(second and third terms in (1)). Thus, the KL-distance flexibly deals with X and Y as a correlation-type relationship (i.e., equal roles of X and Y), or as a conditional-type relationship. In comparison, dCov is a characteristic function-based divergence measure that calculates the discrepancy as the difference between the joint characteristic function and the product of the marginal characteristic functions; thus it is a correlation-type relationship only. The class of measures we define (using characteristic functions) deals with conditional-type relationships. Therefore, the combination of our proposed measures and dCov form a class that is comparable with those developed for the ϕ-divergence family. In other words, our proposed method fills a gap in the literature by comparing characteristic function-based measures and density-based measures.

We define the proposed class of measures as a conditional class based on characteristic functions, treating one of the random vectors as a response. This is very similar to the approach adopted in classification and discriminant analyses and in inverse regressions. Typical classification and discriminant analysis and in inverse regression methods measure the relations in the inverse mean function (or moments), or dependence that involves densities, whereas our method measures the dependence between two sets of variables using distance. This novel class defines a general collection of new measures by choosing different weight functions in the definition, where the weight function in the class determines the actual measure. For the purpose of illustration, however, we use a weight function similar to that used by Székely, Rizzo, and Bakirov (2007). In general, this weight function leads to an index that can be calculated using Euclidean distance.

The proposed method can flexibly deal with categorical or continuous Y. Regardless of whether Y is categorical or the slicing method is used for continuous Y, our index is a variant of the distance components (DISCO; Rizzo and Székely (2010)) method. The index has a simple population version and requires only the

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0538

calculation of Euclidean distance, while keeping the advantage of nonparametric methods. In comparison, the test defined in DISCO applies to categorical Y only, and is a type of generalized ANOVA from two-samples to k-samples, but using an atypical formulation, namely, differences between groups. Our method, however, is defined for both continuous and categorical Y. In the case of categorical Y, the formulation determines the difference between the group and the complete sample. Further information on the relationship between the proposed method and DISCO can be found in the Supplementary Material. For continuous Y, the slicing method is just one approach, which we connect to several existing approaches. Many other estimation methods can be used, including nonparametric estimations. Here, we employ a kernel method to demonstrate the advantage of our approach.

The rest of the paper is organized as follows. We propose the new class of measures in Section 2. By choosing a particular weight function, we study the resulting index and its properties in Section 3, and obtain formulae for certain distributions in Section 3.1. An empirical version that employs slicing on $\mathbf{Y}$ is proposed in Section 4.1, where we also establish its properties. A smoothing estimation approach using kernel methods is proposed in Section 4.2. A permutation test is outlined in Section 5. Simulations are used to illustrate the usefulness of our proposed measures in Section 6. Section 7 concludes the paper. All derivations and proofs are provided in the online Supplementary Material.

2 The new class of measures

Suppose $\mathbf{X} \in \mathbb{R}^p$ and $\mathbf{Y} \in \mathbb{R}^q$ are random vectors, where p and q are positive integers. If $p = 1$, we use $\mathbf{X} = X$; if $q = 1$, we use $\mathbf{Y} = Y$. The characteristic functions of $\mathbf{X}$, $\mathbf{X}|\mathbf{Y}$ and $(\mathbf{X}, \mathbf{Y})$ are denoted by $f_{\mathbf{X}}$, $f_{\mathbf{X}|\mathbf{Y}}$ and $f_{\mathbf{X},\mathbf{Y}}$, respectively. For a complex-valued function $f(\cdot)$, we denote $\bar{f}$ as the complex conjugate of f. Let $|f|^2 = f\bar{f}$, and the Euclidean norm of $\mathbf{X} \in \mathbb{R}^p$ be $|\mathbf{X}|_p$.

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0538

The hypothesis test of independence between $\mathbf{X}$ and $\mathbf{Y}$ is as follows:

$$H_0 : f_{\mathbf{X}|\mathbf{Y}} = f_{\mathbf{X}} \text{ vs. } H_1 : f_{\mathbf{X}|\mathbf{Y}} \neq f_{\mathbf{X}}.$$

This is because if $\mathbf{X}$ is independent of $\mathbf{Y}$, then $f_{\mathbf{X}|\mathbf{Y}} = f_{\mathbf{X}}$, which implies that $e^{is^T\mathbf{Y}} f_{\mathbf{X}|\mathbf{Y}} = e^{is^T\mathbf{Y}} f_{\mathbf{X}}$, for $s \in \mathbb{R}^q$. Then, by taking the expectation over $\mathbf{Y}$, we obtain $f_{\mathbf{X},\mathbf{Y}} = f_{\mathbf{X}} f_{\mathbf{Y}}$.

Definition 2.1 *The nonnegative measure of the conditional difference using the characteristic function of* $\mathbf{X}|\mathbf{Y}$ *is denoted by* $\mathcal{C}_{w,\mathbf{Y}}(\mathbf{X}|\mathbf{Y})$*, which has a squared value of*

$$\mathcal{C}^2_{w,\mathbf{Y}}(\mathbf{X}|\mathbf{Y}) = ||f_{\mathbf{X}|\mathbf{Y}}(t) - f_{\mathbf{X}}(t)||^2 = \int_{\mathbb{R}^p} |f_{\mathbf{X}|\mathbf{Y}}(t) - f_{\mathbf{X}}(t)|^2 w(t)dt, \tag{2}$$

where $w(t) \in \mathbb{R}^p$ is an arbitrary nonnegative weight function for which the aforementioned integral exists. Note that $\mathcal{C}^2_{w,\mathbf{Y}}(\mathbf{X}|\mathbf{Y}) \geq 0$. The term $\mathcal{C}^2_{w,\mathbf{Y}}(\mathbf{X}|\mathbf{Y})$ is a $\mathbf{Y}$-measurable random variable that depends on w. That is, the subscript w in $\mathcal{C}^2_{w,\mathbf{Y}}(\mathbf{X}|\mathbf{Y})$ indicates that each w may lead to a different index. The expected conditional difference is defined as follows.

Definition 2.2 *The expectation of the conditional difference (ECD) using the characteristic function of* $\mathbf{X}|\mathbf{Y}$ *is denoted by* $\mathcal{C}_w(\mathbf{X}|\mathbf{Y})$*, which has a squared value of*

$$\mathcal{C}^2_w(\mathbf{X}|\mathbf{Y}) = \mathrm{E}_{\mathbf{Y}}[\mathcal{C}^2_{w,\mathbf{Y}}(\mathbf{X}|\mathbf{Y})] = \mathrm{E}_{\mathbf{Y}}[\int_{\mathbb{R}^p} |f_{\mathbf{X}|\mathbf{Y}}(t) - f_{\mathbf{X}}(t)|^2 w(t)dt]. \tag{3}$$

Note again that $\mathcal{C}^2_w(\mathbf{X}|\mathbf{Y}) \geq 0$. Although $\mathcal{C}^2_w(\mathbf{X}|\mathbf{Y})$ depends on the choice of w, we omit the subscript w and write $\mathcal{C}^2_w(\mathbf{X}|\mathbf{Y})$ as $\mathcal{C}^2(\mathbf{X}|\mathbf{Y})$ for simplicity, without ambiguity. The following lemma indicates that $\mathcal{C}^2(\mathbf{X}|\mathbf{Y}) = 0$ is equivalent to the independence of $\mathbf{X}$ and $\mathbf{Y}$. Thus, $\mathcal{C}^2(\mathbf{X}|\mathbf{Y})$ is a measure of independence (see the Supplementary Material for the proof).

Lemma 2.1 $\mathcal{C}^2(\mathbf{X}|\mathbf{Y}) = 0 \Leftrightarrow \mathcal{C}^2_{w,\mathbf{Y}}(\mathbf{X}|\mathbf{Y}) = 0$ *almost surely for* $\mathbf{Y} \Leftrightarrow f_{\mathbf{X}|\mathbf{Y}}(t) = f_{\mathbf{X}}(t)$ *almost surely, for* $\mathbf{Y}$ *and* $t \in \mathbb{R}^p$.

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0538

A direct application of (3) indicates that

$$\mathcal{C}^2(\mathbf{X}|\mathbf{X}) = \mathrm{E}_{\mathbf{X}}[\mathcal{C}^2_{w,\mathbf{X}}(\mathbf{X}|\mathbf{X})] = \mathrm{E}_{\mathbf{X}}[\int_{\mathbb{R}^p} |e^{it\mathbf{X}} - f_{\mathbf{X}}(t)|^2 w(t)dt]. \tag{4}$$

Thus, a correlation coefficient-type statistic can be defined as

$$R_c = R_c(\mathbf{X}|\mathbf{Y}) = \frac{\mathcal{C}(\mathbf{X}|\mathbf{Y})}{\mathcal{C}(\mathbf{X}|\mathbf{X})}. \tag{5}$$

The results below indicate the properties of $\mathcal{C}(\mathbf{X}|\mathbf{X})$, $\mathcal{C}(\mathbf{X}|\mathbf{Y})$, and R_c.

Theorem 2.1 *The following properties hold:*

1. $\mathcal{C}(\mathbf{X}|\mathbf{X}) = 0$ *iff* $\mathbf{X} = \mathrm{E}(\mathbf{X})$, *almost surely.*

2. $\mathcal{C}(\mathbf{W}_1 + \mathbf{W}_2|\mathbf{V}_1 + \mathbf{V}_2) \leq \mathcal{C}(\mathbf{W}_1|\mathbf{V}_1) + \mathcal{C}(\mathbf{W}_2|\mathbf{V}_2)$ *for independent random vectors* $(\mathbf{W}_1, \mathbf{V}_1)$ *and* $(\mathbf{W}_2, \mathbf{V}_2)$. *Equality holds if and only if* $\mathbf{W}_1$ *and* $\mathbf{V}_1$ *are both constant or* $\mathbf{W}_2$ *and* $\mathbf{V}_2$ *are both constant, or if* $\mathbf{W}_1, \mathbf{V}_1, \mathbf{W}_2$, *and* $\mathbf{V}_2$ *are mutually independent.*

3. $\mathcal{C}(\mathbf{X}+\mathbf{Y}|\mathbf{X}+\mathbf{Y}) \leq \mathcal{C}(\mathbf{X}|\mathbf{X}) + \mathcal{C}(\mathbf{Y}|\mathbf{Y})$ *for independent random vectors* $\mathbf{X}$ *and* $\mathbf{Y}$. *Equality holds if and only if at least one of the random vectors* $\mathbf{X}$ *and* $\mathbf{Y}$ *is constant.*

4. $0 \leq \mathcal{C}(\mathbf{X}|\mathbf{Y}) \leq \mathcal{C}(\mathbf{X}|\mathbf{X})$, *and* $0 \leq R_c \leq 1$.

Most of the independence measures in the literature are symmetric; however our measure is asymmetric owing to its conditional setup. If needed, it can be modified to the following symmetric version: $\mathcal{C}^2_s(\mathbf{X}, \mathbf{Y}) = \mathcal{C}^2(\mathbf{X}|\mathbf{Y}) + \mathcal{C}^2(\mathbf{Y}|\mathbf{X})$. Note that the combination of two measures of discrepancies, that is $\mathcal{C}^2(\mathbf{X}|\mathbf{Y})$ and the discrepancy between the joint characteristic function and the product of two marginal characteristic functions (Sejdinvoc et al. (2013)), forms a larger class that is comparable with the ϕ-divergence family.

In the proposed class, different weight functions result in different dependency measures. For example, weight functions such as a Gaussian weight yield a new

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0538

type of measure in the spirit of the Hilbert–Schmidt independence criterion (Gretton et al. (2005)). Hence, the choice of weight function is important. In this paper, we consider only a weight function similar to that of Székely, Rizzo, and Bakirov (2007) which results in a very simple formula.

3 The new index and its properties

Let $\tilde{C}(p,\alpha) = \frac{2\pi^{p/2}\Gamma(1-\alpha/2)}{\alpha 2^{\alpha}\Gamma((p+\alpha)/2)}$, for $0 < \alpha < 2$. For $\alpha = 1$, define $\tilde{c}_p = \tilde{C}(p,1) = \frac{\pi^{(1+p)/2}}{\Gamma((1+p)/2)}$. Suppose that $t \in \mathbb{R}^p$. Let the weight function be $w(t) = (\tilde{c}_p|t|_p^{1+p})^{-1}$. This is a positive weight function, and is very similar to those of Székely, Rizzo, and Bakirov (2007) and Székely and Rizzo (2009). Hereafter, we use this weight function only.

Let $(\mathbf{X}', \mathbf{Y}')$ be an independent and identically distributed (i.i.d.) copy of $(\mathbf{X}, \mathbf{Y})$, $\mathbf{X_Y}$ denote a random variable distributed as $\mathbf{X}|\mathbf{Y}$ (Cook (2007)), $\mathbf{X}'_{\mathbf{Y}'}$ denote a random variable distributed as $\mathbf{X}'|\mathbf{Y}'$, and $\mathbf{X}'_{\mathbf{Y}}$ denote a random variable distributed as $\mathbf{X}'|\mathbf{Y}'$, with $\mathbf{Y}' = \mathbf{y}$ and $\mathbf{Y} = \mathbf{y}$. Throughout the paper, we assume $\mathrm{E}|\mathbf{X}| < \infty$ and $\mathrm{E}|\mathbf{X_Y}| < \infty$, unless otherwise stated. These assumptions guarantee the finiteness of $\mathcal{C}^2(\mathbf{X}|\mathbf{Y})$, and enable us to obtain a simpler, but equivalent formula of (3). The proofs are provided in the Supplementary Material.

Theorem 3.1 *An equivalent form of (3) can be expressed as follows:*

$$\mathcal{C}^2(\mathbf{X}|\mathbf{Y}) = \mathrm{E}|\mathbf{X} - \mathbf{X}'_{\mathbf{Y}}| - \mathrm{E}|\mathbf{X_Y} - \mathbf{X}'_{\mathbf{Y}}| = \mathrm{E}|\mathbf{X} - \mathbf{X}'| - \mathrm{E}|\mathbf{X_Y} - \mathbf{X}'_{\mathbf{Y}}|, \qquad (6)$$

where the expectation is over all random vectors. For instance, the final expectation first takes the conditional expectation given $\mathbf{Y}$*, and then over* $\mathbf{Y}$*.*

Note that, strictly speaking, $\mathrm{E}|\mathbf{X_Y} - \mathbf{X}'_{\mathbf{Y}}| = \mathrm{EE}[|\mathbf{X} - \mathbf{X}'||\mathbf{Y} = \mathbf{y}, \mathbf{Y}' = \mathbf{y}]$. In addition, formula (3) is more general than formula (6). For example, the conditional Cauchy distribution in Section 3.1 can be calculated using (3), but not (6).

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0538

Theorem 3.2 *1.* $\mathcal{C}^2(\mathbf{X}|\mathbf{X}) = \mathrm{E}[\mathcal{C}^2_{w,\mathbf{X}}(\mathbf{X}|\mathbf{X})] = \mathrm{E}|\mathbf{X} - \mathbf{X}'|$.

2. $\mathcal{C}^2(\mathbf{a} + b\mathbf{B}\mathbf{X}|\mathbf{Y}) = |b|\mathcal{C}^2(\mathbf{X}|\mathbf{Y})$ *for a constant vector* $\mathbf{a}$*, a scalar* b*, and an orthonormal matrix* $\mathbf{B}$*.*

3. $R_c = 1$ *iff* $\mathbf{X}$ *is a function of* $\mathbf{Y}$*, i.e.,* $\mathbf{X} = \mathbf{g}(\mathbf{Y})$*, where* $\mathbf{g}$ *is a* $p \times 1$ *vector function.*

3.1 Special distributions

In this section, we describe the connection between this index and several well-known distributions, including the normal, binomial, and Cauchy distributions. The derivations of these relations are provided in the Supplementary Material.

Conditional normal distribution. Suppose $X|Y \sim N(\mu_Y, \sigma_Y^2)$, where $Y \in \{0, 1\}$. For simplicity, assume that $\sigma_Y^2 = \sigma^2 = 1$, and define $\Delta = \mu_0 - \mu_1$. Let p_y be the probability for the class $Y = y$, and let $\mathrm{erf}(z) = \frac{2}{\sqrt{\pi}} \int_0^z e^{-t^2} dt$ be the Gaussian error function. Then, we have:

$$\mathcal{C}^2(X|Y) = 4p_0p_1[\frac{\Delta}{2}\mathrm{erf}(\frac{\Delta}{2}) + \frac{e^{-\Delta^2/4} - 1}{\sqrt{\pi}}].$$

Note that this equivalence indicates that $\Delta = 0$ iff $\mathcal{C}^2(X|Y) = 0$, as expected.

Bivariate normal distribution. Suppose that X and Y follows a standard normal distribution with correlation coefficient ρ. Then, we have that $X|Y \sim N(\rho Y, (1 - \rho^2))$. Our index can be expressed using ρ, as follows:

$$\mathcal{C}^2(X|Y) = \frac{2}{\sqrt{\pi}}(1 - \sqrt{1 - \rho^2}).$$

Once again, we have that $\mathcal{C}^2(X|Y) = 0$ iff $\rho = 0$. In such a case, the difference between this and the distance correlation (Székely, Rizzo, and Bakirov (2007)) becomes evident.

Conditional binomial distribution. Suppose $X|Y \sim Bin(n, q_Y)$, where $Y \in \{0, 1\}$. Let p_y be the probability for the class $Y = y$. For $n = 1$, when it is a

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0538

Bernoulli distribution, we have that

$$\mathcal{C}^2(X|Y) = 4p_0p_1(q_0 - q_1)^2.$$

For $n = 2$, we have that $\mathcal{C}^2(X|Y) = 4p_0p_1(q_0 - q_1)^2[1 + (1 - q_0 - q_1)^2]$. It is clear that in both cases, $\mathcal{C}^2(X|Y) = 0$ iff $q_0 = q_1$. A general formula $\mathcal{C}^2(X|Y) = 0$ for the conditional binomial distribution can be found in the Supplementary Material.

Conditional Cauchy distribution. Although we require finiteness of the conditional means to develop the equivalence formula for $\mathcal{C}^2(\mathbf{X}|\mathbf{Y})$, as in (6), the original definition of our index $\mathcal{C}^2(\mathbf{X}|\mathbf{Y})$ requires only the existence of its respective characteristic functions. It is well known that the Cauchy distribution has a characteristic function, but without finite moments. Nevertheless, we can still perform such a calculation. Suppose that the Cauchy distribution has density $p(x|y) = \frac{q_y}{\pi(q_y^2+x^2)}$, where $y \in \{0, 1\}$. Let p_y be the probability for the class $Y = y$. We then have that

$$\mathcal{C}^2(X|Y) = \frac{4p_0p_1}{\pi}[q_0 \ln(\frac{2q_0}{q_0 + q_1}) + q_1 \ln(\frac{2q_1}{q_0 + q_1})].$$

Again, $q_0 \ln(\frac{2q_0}{q_0+q_1}) + q_1 \ln(\frac{2q_1}{q_0+q_1}) \geq 0$ with strict equality iff $q_0 = q_1$.

4 Estimation approaches

4.1 Slicing estimator

In developing the population version of our index measure, we did not require $\mathbf{Y}$ to be discrete or continuous. We now consider a special sample version of continuous $\mathbf{Y}$, where we slice it into finite categories. Slicing techniques for continuous variables have been used extensively in dimension reduction; see, for example, Li (1991), Cook and Weisberg (1991), Li, Zha, and Chiaromonte (2005), Li and Wang (2007), Wang and Xia (2008), and Cook and Zhang (2014). Slicing is a natural choice for our purposes because it facilitates technical simplicity, owing to the last term in the second equation of (6) in Theorem 3.1. To facilitate the development of our

estimator, we now assume that Y is a categorical variable with H levels; that is, $Y = \{1, \cdots, H\}$.

Slicing in multivariate and high-dimensional situations is an interesting, yet challenging topic. Nevertheless, efforts to address slicing have been made in various areas, including dimension reduction. For example, methods developed by Zhu et al. (2010), Li, Wen, and Zhu (2008) and Cook and Zhang (2014) may be beneficial to our approach.

Let $(\mathbf{X}_k, Y_k)$, for $k = 1, \cdots, n$, be a random sample of $(\mathbf{X}, Y)$. For the purpose of slicing, these n observations can be equivalently written as $(\mathbf{X}_{y,k_y}, Y_{y,k_y})$, where $y = 1, \cdots, H$, $k_y = 1, \cdots, n_y$, where n_y is the number of observations for slice y, and $Y_{y,k_y} = y$ for any k_y. The choice of H is not always independent of the sample size n. If the number of observations in the data set is large, one may choose a larger number of slices. However, if Y is multivariate with a high dimension, then the number of slices for each dimension of Y should be smaller. This strategy helps to ensure that there are enough observations in each slice.

We can now define an empirical measure and establish its corresponding theoretical results.

Definition 4.1 *An empirical measure is defined as the following weighted norm:*

$$\mathcal{C}_n^2(\mathbf{X}|Y) = \sum_{y=1}^{H} \frac{n_y}{n} \mathcal{C}_{w,y,n}^2(\mathbf{X}|Y = y) = \sum_{y=1}^{H} \frac{n_y}{n} ||f_{\mathbf{X}|y}^n(t) - f_{\mathbf{X}}^n(t)||^2. \tag{7}$$

We next establish a different formula for the empirical version that facilitates simple calculations. The proof is provided in the Supplementary Material.

Theorem 4.1 *The empirical measure can be written as*

$$\mathcal{C}_n^2(\mathbf{X}|Y) = \frac{1}{n^2} \sum_{y,y'=1}^{H,H} \sum_{k_y,l_{y'}=1}^{n_y,n_{y'}} |\mathbf{X}_{y,k_y} - \mathbf{X}_{y',l_{y'}}| - \frac{1}{n} \sum_{y=1}^{H} \frac{1}{n_y} \sum_{k_y,l_y=1}^{n_y,n_y} |\mathbf{X}_{y,k_y} - \mathbf{X}_{y,l_y}|. \tag{8}$$

Theorem 4.1 immediately implies the next result.

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0538

Corollary 4.1

$$\mathcal{C}_n^2(\mathbf{X}|Y) = \frac{1}{n^2}\sum_{k,l=1}^{n,n}|\mathbf{X}_k - \mathbf{X}_l| - \frac{1}{n}\sum_{y=1}^{H}\frac{1}{n_y}\sum_{k_y,l_y=1}^{n_y,n_y}|\mathbf{X}_{y,k_y} - \mathbf{X}_{y,l_y}|. \tag{9}$$

$$\mathcal{C}_n^2(\mathbf{X}|Y) \le \mathcal{C}_n^2(\mathbf{X}|\mathbf{X}) = \frac{1}{n^2}\sum_{k,l=1}^{n,n}|\mathbf{X}_k - \mathbf{X}_l|. \tag{10}$$

Based on Definition 4.1, it is easy to see that the following results hold; thus, we omit the proof.

Lemma 4.1 *The following properties hold:*

1. $\mathcal{C}_n^2(\mathbf{X}|Y) \ge 0$.
2. $\mathcal{C}_n^2(\mathbf{X}|\mathbf{X}) = 0$ *iff every sample observation is identical.*

The proof for the following result appears in the Supplementary Material.

Lemma 4.2

$$\lim_{n\to\infty}\mathcal{C}_n^2(\mathbf{X}|Y) = \mathcal{C}^2(\mathbf{X}|Y) \text{ almost surely.}$$

This lemma indicates that our sample version is properly defined and consistent.

Next, we develop asymptotic distributions for the empirical measure.

Theorem 4.2 *(Weak convergence)*

1. *If* $\mathbf{X}$ *and* Y *are independent, and* $\mathrm{E}(|\mathbf{X}|) < \infty$, *then* $n\mathcal{C}_n^2(\mathbf{X}|Y) \xrightarrow[n\to\infty]{D} \mathcal{C}^2(\mathbf{X}|\mathbf{X})Q$, *where* $Q \sim \chi^2_{H-1}$.
2. *If* $\mathbf{X}$ *and* Y *are independent, and* $\mathrm{E}(|\mathbf{X}|) < \infty$, *then* $n\mathcal{C}_n^2(\mathbf{X}|Y)/\mathcal{C}_n^2(\mathbf{X}|\mathbf{X}) \xrightarrow[n\to\infty]{D} Q$, *where* $Q \sim \chi^2_{H-1}$.
3. *If* $\mathbf{X}$ *and* Y *are dependent, then* $n\mathcal{C}_n^2(\mathbf{X}|Y)/\mathcal{C}_n^2(\mathbf{X}|\mathbf{X}) \xrightarrow[n\to\infty]{P} \infty$.

The proof of Theorem 4.2 is provided in the Supplementary Material. We now establish the limiting distribution. If $Q \sim \chi^2_{H-1}$, then

$$P\{Q \ge \chi^2_{H-1}(1-\alpha_0)\} \le \alpha_0, \text{ for all } 0 < \alpha_0 \le 0.215,$$

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0538

where $\chi^2_{H-1}(1-\alpha_0)$ is the $(1-\alpha_0)$-quantile of a chi-square variable with $H-1$ degrees of freedom. This result follows from that of Székely and Bakirov (2003, page 189). Thus, a test that rejects independence if $n\mathcal{C}_n^2(\mathbf{X}|Y)/\mathcal{C}_n^2(\mathbf{X}|\mathbf{X}) \geq \chi^2_{H-1}(1-\alpha_0)$ has an asymptotic significance level of at most α_0. However, the asymptotic test criterion could be quite conservative for many distributions; see Székely, Rizzo, and Bakirov (2007), Székely and Rizzo (2009), and Rizzo and Székely (2010) for further comments.

By using slicing, our measure is equivalent to DISCO, which employs conditional moments directly in a manner similar to that of (6), but for categorical Y only. Hence, in general, DISCO limits certain distributions, such as the conditional Cauchy distribution in Section 3.1. Our theoretical justification also differs from DISCO, but is similar to dCov. Both our measure and dCov are defined using characteristic functions; thus the theoretical justifications for the two are analogous. For continuous $\mathbf{Y}$, we change $\mathbf{Y}$ to a class variable using slicing. In such a case, our index provides an alternative way to specify dCov. However, one does not have to use slicing, because other approaches may be used as well. Thus, our index provides many possible approaches for measuring independence between continuous random vectors, which may lead to new research directions. One such approach is proposed in the next section.

4.2 Kernel estimator

For continuous $\mathbf{Y}$, slicing is just one approach. In fact, even slicing can be improved using techniques such as "moving slicing" (Li, Zha, and Chiaromonte (2005)) or the fused approach (Cook and Zhang (2014)). In this section, we propose a kernel method to estimate (6) (in particular, the last term in (6)), which differs from DISCO.

For simplicity, let $m = \mathrm{E}|\mathbf{X_Y} - \mathbf{X'_Y}|$. Thus, our main goal is to estimate m

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0538

via kernel methods. Write $m = \mathrm{E}_{\mathbf{Y}}\mathrm{E}|\mathbf{X}_{\mathbf{Y}} - \mathbf{X}'_{\mathbf{Y}}| = \mathrm{E}_{\mathbf{Y}} m(\mathbf{Y})$. Then, $m(\mathbf{Y}) = \mathrm{E}_{(\mathbf{X},\mathbf{X}')}(|\mathbf{X} - \mathbf{X}'||\mathbf{Y}) = \mathrm{E}_{\mathbf{X}}[m(\mathbf{X},\mathbf{Y})|\mathbf{Y}]$, where $m(\mathbf{X},\mathbf{Y}) = \mathrm{E}_{\mathbf{X}'}(|\mathbf{X} - \mathbf{X}'||\mathbf{Y})$.

For the kernel estimation, $K_h(t) = h^{-q}K(t/h)$, for $h > 0$, denotes a q-dimensional kernel function. Let $p_0(\mathbf{y})$ be the density function of $\mathbf{Y}$, which has the kernel estimator $\hat{p}_0(\mathbf{y}) = n^{-1}\sum_{k=1}^{n} K_h(\mathbf{y}_k - \mathbf{y})$. Thus, an estimate of $m(\mathbf{X},\mathbf{Y})$ is

$$\hat{m}(\mathbf{X},\mathbf{Y}) = \frac{n^{-1}\sum_{j=1}^{n}|\mathbf{X} - \mathbf{X}_j|K_h(\mathbf{Y} - \mathbf{Y}_j)}{n^{-1}\sum_{j=1}^{n} K_h(\mathbf{Y} - \mathbf{Y}_j)}.$$

Moreover, an estimate of $m(\mathbf{Y})$ is

$$\begin{aligned}\hat{m}(\mathbf{Y}) &= \frac{n^{-1}\sum_{i=1}^{n}\hat{m}(\mathbf{X}_i,\mathbf{Y})K_h(\mathbf{Y} - \mathbf{Y}_i)}{n^{-1}\sum_{i=1}^{n} K_h(\mathbf{Y} - \mathbf{Y}_i)} \\ &= \frac{n^{-2}\sum_{i=1,j=1}^{n}|\mathbf{X}_i - \mathbf{X}_j|K_h(\mathbf{Y} - \mathbf{Y}_i)K_h(\mathbf{Y} - \mathbf{Y}_j)}{n^{-1}\sum_{j=1}^{n} K_h(\mathbf{Y} - \mathbf{Y}_j)n^{-1}\sum_{i=1}^{n} K_h(\mathbf{Y} - \mathbf{Y}_i)}.\end{aligned}$$

Finally, an estimate of m is $\hat{m} = \frac{1}{n}\sum_{l=1}^{n}\hat{m}(\mathbf{Y}_l)$. Hence, the kernel estimator of $\mathcal{C}^2(\mathbf{X}|\mathbf{Y})$ is $\mathcal{C}^2_{n,k}(\mathbf{X}|\mathbf{Y}) = \frac{1}{n^2}\sum_{i,j}|\mathbf{X}_i - \mathbf{X}_j| - \hat{m}$.

We now establish the consistency for the kernel estimator in Theorem 4.3. For such a result, we need the following regularity conditions, taken from Chen, Cook, and Zou (2015):

Condition A_1: The density functions $p(\mathbf{x}|\mathbf{y})$ and $p(\mathbf{y})$ are continuous and bounded away from zero. The support of $\mathbf{y}$ is bounded and compact in $\mathbb{R}^q$.

Condition A_2: The continuous kernel function $K(t)$ is Lipschitz on $[-1, 1]$, and for some $s > q/2$,

$$\int K(t)dt = 1, \int t^i K(t)dt = 0, (1 \le i \le s-1), 0 \ne \int t^s K(t)dt < \infty.$$

Condition A_3: As $n \to \infty$, the bandwidth h satisfies $h \to 0$, $nh^{2q} \to \infty$, and $nh^{2s+q/2}\log n \to 0$.

Condition A_4: We have that $E|\mathbf{X}_{\mathbf{y}}|^4 < \infty$.

Condition A_5: Write $p(\mathbf{x},\mathbf{y}) = p(\mathbf{x}|\mathbf{y})p(\mathbf{y})$, which is s-times differentiable with respect to $\mathbf{y}$, and its sth-order derivative is uniformly bounded by a constant C_0 that does not depend on $\mathbf{y}$.

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0538

Conditions A_1 and A_5 require that the density functions be positive and sufficiently smooth. Condition A_5 facilitates control of the remainder terms in the Taylor expansions. We can relax this condition by assuming local Lipschitz properties for the density functions, which are widely imposed in the literature (Li, Zhu, and Zhu (2011)). Condition A_2 implies that the kernel function is bounded from above, which holds for many well-known kernel functions. Condition A_3 gives conditions on the bandwidth h, and are relatively mild. Condition A_4 requires that certain moments are finite, which is quite typical. To introduce Theorem 4.3, we establish the following lemma, which is a direct application of Lemma S5 of Chen, Cook, and Zou (2015).

Lemma 4.3 *Suppose Conditions A_1–A_5 hold. Then,*

$$\sup_{\mathbf{y}\in\mathbb{R}^q} |\hat{m}(\mathbf{y}) - m(\mathbf{y})| = O(h^s + (nh^q)^{-1/2}\log n), \ \textit{almost surely.}$$

Here, we use the conditions of Chen, Cook, and Zou (2015) directly, for simplicity. The requirement that the density functions be bounded away from zero in Condition A_1 seems restrictive, although our simulations show otherwise. However, one can weaken Condition A_1 by using different conditions, such as those of Härdle and Stoker (1989), Samarov (1993), or Wang et al. (2015). Nevertheless, we need to modify our estimator by incorporating a trim/weight function to deal with the density near zero and to mitigate significant bias. Regardless, for a finite sample, we can ignore this issue, because it is finite. As such, even if the estimates for points near the boundary are small, they will never be zero.

We can now establish the following consistency result.

Theorem 4.3 *Under Conditions A_1–A_5, we have that $\mathcal{C}^2_{n,k}(\mathbf{X}|\mathbf{Y}) \xrightarrow[n\to\infty]{P} \mathcal{C}^2(\mathbf{X}|\mathbf{Y})$.*

Note that the first term in $\mathcal{C}^2_{n,k}(\mathbf{X}|\mathbf{Y})$ is a typical U-statistic, which is root-n asymptotically normal. Using the technicals in Chen, Cook, and Zou (2015), we can establish the asymptotic normality for the second term in $\mathcal{C}^2_{n,k}(\mathbf{X}|\mathbf{Y})$, which has

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0538

rate $nh^{q/2}$. Combining the two terms, we can still manipulate the asymptotic normality at the same rate. However, one of the asymptotic variances in the two terms vanishes at a faster rate. Hence, this is not practically useful when the sample size is large. Furthermore, even if the terms have the same rate of convergence (cf., Székely, Rizzo, and Bakirov (2007), Székely and Rizzo (2009), Rizzo and Székely (2010), Shao and Zhang (2014); and Wang et al. (2015)), permutation or bootstrap tests are usually preferred to asymptotic distributions. We describe the use of a permutation test in the next section. Note that $K_h(t)$ is a q-dimensional kernel function. Therefore, theoretically, the kernel method can be used for $\mathbf{Y}$ with any dimensions. As a result of the high-dimension issue, the kernel method certainly has its own practical restriction. Nevertheless, there exist kernel estimation methods when using (conditional) dCov, as discussed by Wang et al. (2015) and Chen, Cook, and Zou (2015).

5 Testing procedure

To obtain the p-value for our independence test, we implemented a permutation approach (Efron and Tibshirani (1998); Davison and Hinkley (1997)). Based on the preceding discussion, we use R_c as the illustrative test statistic when calculating the p-value. We use R_c in our simulation because it has a value between zero and one and, thus, will not be affected by the unit change of the random vectors. To illustrate the permutation test, we use the slicing method, denoted by $R_c(\text{slice})$, as follows: Let π^b represent one permutation of the sample, for $b = 1, \cdots, B$, where B is the total number of permutations. In our simulations, we set $B = 999$, unless otherwise stated. Let $R_c(\text{slice})^b$ be the test statistic corresponding to the permuted sample π^b, and let $R_c(\text{slice})^0$ be the observed test statistic. Compute the p-value

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0538

using the following formula ($\mathbf{1}(\cdot)$ is the indicator function):

$$\hat{p} = \frac{1 + \sum_{b=1}^{B} \mathbf{1}(R_c(\text{slice})^b \geq R_c(\text{slice})^0)}{B+1}.$$

6 Simulation studies

In this section we provide empirical evidence for our proposed measure using three estimation methods: slicing [R_c(slice)], the Epanechnikov kernel [R_c(epa)], and the Gaussian kernel [R_c(gau)]. We compare our results with those of dCov and DISCO, as well as G_m^2 and G_t^2 of Wang, Jiang, and Liu (2016), and the maximal information coefficient (MIC) of Reshef et al. (2011) and the total information coefficient (TIC) of Reshef et al. (2016). The code for the R packages for dCov and DISCO, G_m^2 and G_t^2, and MIC and TIC are available in Rizzo and Székely (2018), Wang and Bo (2016), and Filosi et al. (2017), respectively.

Example 6.1 Six characteristics of aircraft designs from the 20th century were recorded in the aircraft data of Saviotti (1996). The data are available in the R package *sm* (Bowman and Azzalini (1997, 2007)). Two variables, wing span(m) and speed (km/h), in period 3 (of three brand periods of the 20th century) with $n = 230$ designs were considered. Here, we test the independence of log(Speed) and log(Span).

We apply the slicing method by slicing log(Span) into H groups. The number of observations in each slice is $\lfloor n/H \rfloor$. Table 1 reports the corresponding test statistic and p-value using various numbers of slices and the two kernel methods. With regard to the different numbers of slices, we find that as long as the number is not too small or too big, specifically, the number of data points in each slice is greater than five but not close to $n/2$, then the test results are highly consistent and comparable. In addition, the p-values indicate that all three methods give the same test result as that of dCov of Székely and Rizzo (2009), which has a p-value of 0.001.

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0538

Table 1: Test results using different methods

	R_c(slice)						R_c(epa)	R_c(gau)
	$H=2$	$H=5$	$H=10$	$H=23$	$H=46$	$H=115$		
Test statistic	0.161	0.264	0.328	0.453	0.528	0.752	0.302	0.237
p-value	0.004	0.001	0.001	0.001	0.001	0.007	0.001	0.001

Example 6.2 In this example, we study the type-I error rates for dCov, the kernel methods, slicing on the continuous variable to apply DISCO, and our slicing method. We simulate four models. In model (a), the marginal distributions of $\mathbf{X}$ and that of Y are standard normal, where $p = 5$ and $q = 1$. The elements of $\mathbf{X}$ are independent and are also independent of Y. In models (b)–(d), the dimensions of $\mathbf{X}$ and Y are the same as in (a), except that each individual random variable is independently generated from t_1, χ^2_1, and χ^2_3 distributions, respectively.

We fix the number of slices at $H = 5$ for DISCO and R_c(slice). The total sample sizes are $n = 25, 30, 35, 50, 70, 100$, and we use the number of replicates B $= \lfloor 200 + 5000/n \rfloor$ as suggested by Székely, Rizzo, and Bakirov (2007) to obtain the p-value for each test. We use 10,000 tests to obtain the type-I error rate at a nominal significance level of 0.1. The empirical type-I error rate for each case is recorded in Table 2. It appears that all methods perform similarly, close to the nominal level, and none consistently beat the others. Simulation results for additional models and a nominal level of 0.05 are given in the Supplementary Material. The conclusions remain qualitatively similar.

Example 6.3 Following Example 2 in Székely and Rizzo (2009), we use the model $(X, Y) = (X, \phi(X))$, where X is a standard normal random variable, and $\phi(\cdot)$ is the standard normal density. Our goal is to conduct a power comparison. The power is computed as the proportion of significant tests out of 10,000 at a significance level of 0.1. Again, we use the number of replicates B $= \lfloor 200 + 5000/n \rfloor$ in each

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0538

Table 2: Empirical type-I error rates for 10,000 tests at nominal significance level of 0.1, using B replicates

(a) $N(0,1), p=5, q=1$							(b) $t_1, p=5, q=1$				
n	B	dCov	DISCO	R_c(slice)	R_c(epa)	R_c(gau)	dCov	DISCO	R_c(slice)	R_c(epa)	R_c(gau)
25	400	0.094	0.103	0.100	0.096	0.101	0.104	0.097	0.095	0.094	0.103
30	366	0.102	0.095	0.099	0.100	0.100	0.102	0.100	0.099	0.098	0.097
35	342	0.105	0.099	0.101	0.102	0.099	0.104	0.100	0.102	0.093	0.095
50	300	0.103	0.099	0.100	0.097	0.101	0.100	0.106	0.104	0.097	0.103
70	271	0.103	0.097	0.103	0.100	0.100	0.100	0.098	0.100	0.099	0.098
100	250	0.101	0.098	0.098	0.104	0.098	0.094	0.105	0.103	0.097	0.102
(c) $\chi^2_1, p=5, q=1$							(d) $\chi^2_3, p=5, q=1$				
n	B	dCov	DISCO	R_c(slice)	R_c(epa)	R_c(gau)	dCov	DISCO	R_c(slice)	R_c(epa)	R_c(gau)
25	400	0.096	0.099	0.099	0.099	0.098	0.097	0.099	0.098	0.100	0.098
30	366	0.102	0.094	0.095	0.098	0.098	0.094	0.100	0.100	0.096	0.102
35	342	0.096	0.102	0.104	0.101	0.098	0.101	0.103	0.104	0.102	0.103
50	300	0.102	0.097	0.098	0.103	0.099	0.099	0.102	0.102	0.103	0.100
70	271	0.103	0.099	0.098	0.101	0.100	0.104	0.100	0.102	0.102	0.100
100	250	0.098	0.101	0.098	0.098	0.102	0.099	0.101	0.102	0.100	0.100

permutation test.

Because Y is continuous, we slice it into several categories for both DISCO and the slicing method. Based on Example 6.1, we use three, three, and four slices with sample sizes $n = 10, 15$, and 20, respectively, and five slices for sample sizes greater than 20. Figure 1 is a plot of the power for the different methods as a function of the sample size n. We find that for $n \geq 35$, all five methods are equivalently powerful, with power near one. For $n < 35$, the Gaussian kernel method performs best, followed by the dCov and Epanechnikov kernel methods. As expected, the slicing and DISCO methods lose power for smaller sample sizes. Székely and Rizzo (2009) showed that the power of dCov is much better than that of the Pearson or Spearman correlations. In conjunction with their results, our example demonstrates that characteristic function-based methods outperform density-based methods.

Example 6.4 We next generate multivariate observations from a four-group bal-

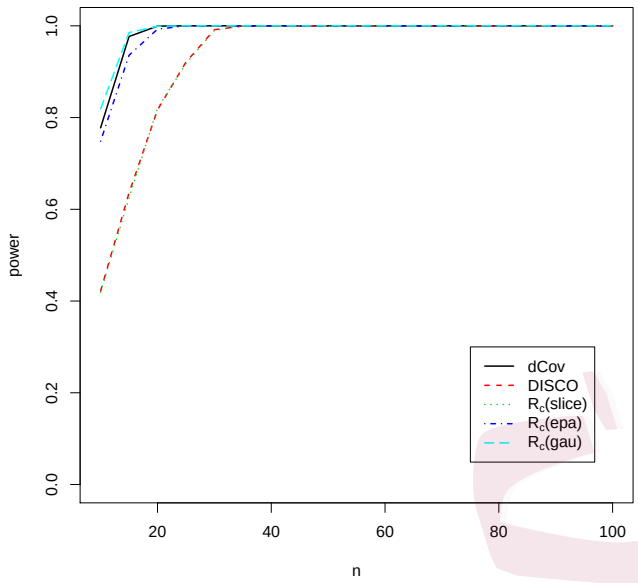


Figure 1: Empirical power comparisons at the 0.1 significance level with different sample sizes, n

anced design with common sample size $n = 30$. The marginal distributions are independent. Group 1 is noncentral $t_4(\delta)$, with noncentrality parameter δ. Groups 2–4 are all central t_4 distributions. The group indicator is Y. This setup is the same as that of Example 3 in Rizzo and Székely (2010). We want to show that for categorical variables, changing values will not change the power of the robust methods.

We first examine the empirical power by fixing the dimension at $p = 10$, but allowing the noncentrality parameter δ to vary. We then consider the empirical power when p varies and δ is fixed at 0.2. The results of the simulations are summarized in Figures 2–3 at a significance level of 0.1. We use $B = 199$ in each test and conduct 10,000 tests.

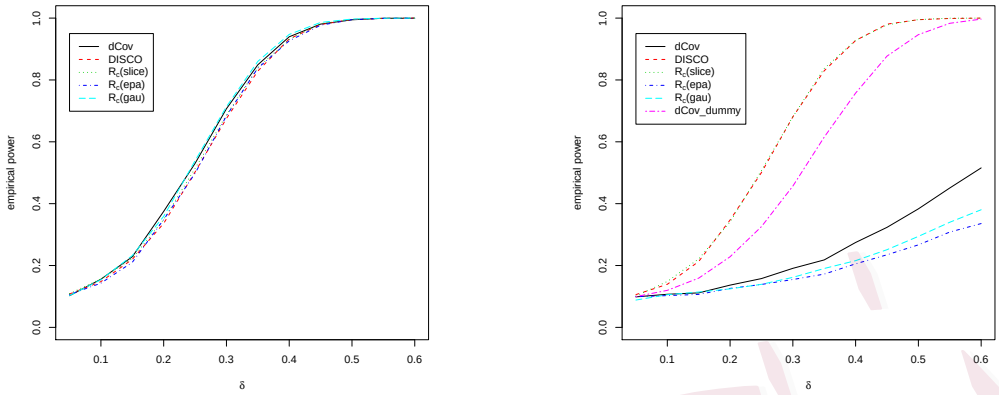
When fixing the dimension p and varying δ, Figure 2 (a) shows that the empirical power when testing the independence of $\mathbf{X}$ and Y is roughly the same when comparing the five methods and the group indicator: 1, 2, 3, and 4. However, when

we change the group indicator Y from 1–4 to 1, 8, 0.5, and 1.2, Figure 2 (b) shows that the power of the DISCO and slicing methods remains the same. The dCov and kernel methods have much smaller empirical power than the other methods do. We also applied dCov with the dummy variables. The dot-dashed line in Figure 2 (b) shows that, although dCov with the dummy variables has greater power than when treating Y as one dimension, with values (1, 0.8, 0.5, 1.2), it still has less power than R_c(slice) or the DISCO method. Figure 3 (a) shows that when the dimension p varies and noncentrality parameter $\delta = 0.2$, the empirical power when testing the independence of $\mathbf{X}$ and Y is again roughly the same when comparing the five methods with the group indicator: 1, 2, 3, and 4. However, after changing the group indicator from 1–4 to 1, 8, 0.5, and 1.2, Figure 3 (b) shows only DISCO and R_c(slice) are robust. Therefore, we believe that, regardless of whether a dummy variable is used, the dCov method has less power than, and is not as stable as DISCO or R_c(slice). Further comparisons with existing density-based methods can be found in Rizzo and Székely (2010).

Example 6.5 The next model considered is $Y = a(\beta^T\mathbf{X})^2\epsilon$, where $\beta = (1, 1, 1, 1, 1, 0, \cdots, 0)^T$, $\mathbf{X} \sim N(\mathbf{0}, \mathbf{\Sigma}_x)$, $\mathbf{\Sigma}_x$ is a $p \times p$ diagonal matrix with the same diagonal element σ_x^2, a is a constant, and $\epsilon \sim N(0, \sigma^2)$ is independent of $\mathbf{X}$.

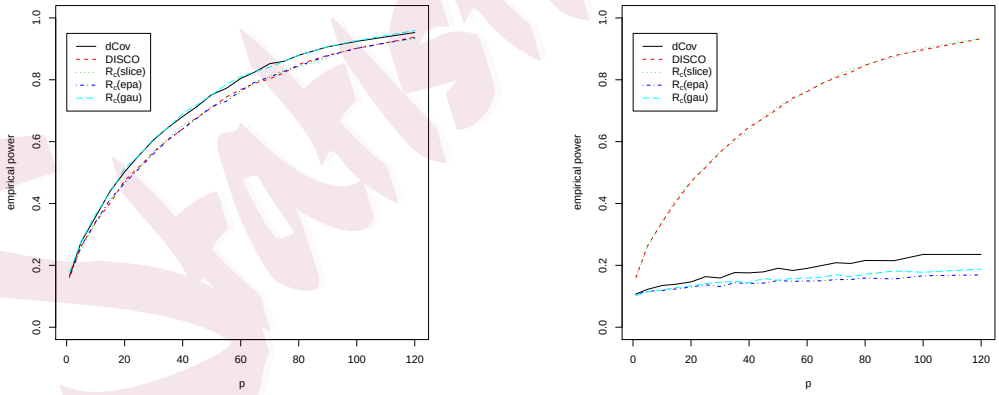
We use the number of replicates B $= \lfloor 200 + 5000/n \rfloor$ in each permutation test, and we use 10,000 tests to obtain the power. We consider different combinations of values of a, p, σ_x^2, and σ^2. Within each combination, we vary the sample size n to determine how the power of the test of the independence of $\mathbf{X}$ and Y changes under the different methods.

In addition to the methods compared previously, we now include the generalized R^2 method of Wang, Jiang, and Liu (2016) (G_m^2 and G_t^2). Figure 4 shows the power change under the four different cases. This figure clearly shows that for such a model with a continuous response, the generalized R^2 method (G_m^2 and G_t^2), DISCO, and



(a) (b)

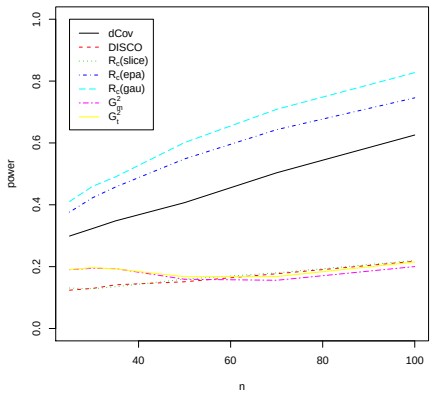
Figure 2: Empirical power when testing the independence of $\mathbf{X}$ and Y using five methods, $n = 30$ per group, dimension $p = 10$, and varying noncentrality parameter δ. The group indicator is (a) 1, 2, 3, 4; (b) 1, 8, 0.5, 1.2, where except for the dot-dashed line, Y is transformed to dummy variables



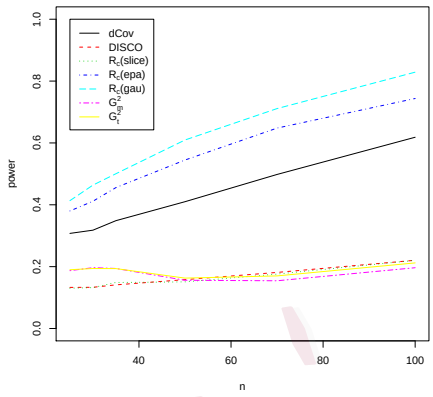
(a) (b)

Figure 3: Empirical power when testing the independence of $\mathbf{X}$ and Y using five methods, $n = 30$ per group, dimension p varies, and noncentrality parameter $\delta = 0.2$. The group indicator is (a) 1, 2, 3, 4; (b) 1, 8, 0.5, 1.2.

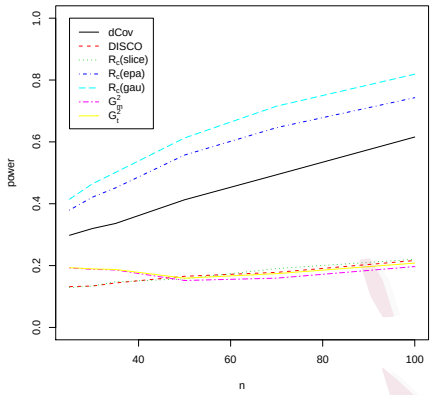
Statistica Sinica: Preprint
doi:10.5705/ss.202017.0538



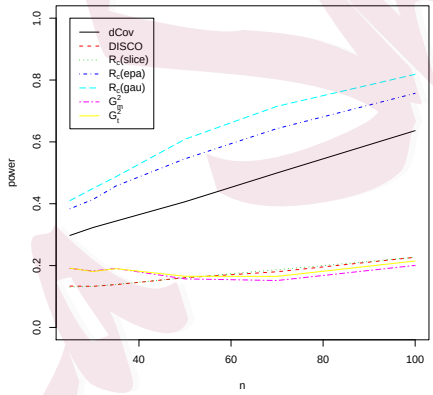
(a) $a = 0.1, p = 10, \sigma_x^2 = 1$, and $\sigma^2 = 1$.



(b) $a = 0.3, p = 10, \sigma_x^2 = 1$, and $\sigma^2 = 1$.



(c) $a = 0.3, p = 10, \sigma_x^2 = 1$, and $\sigma^2 = 4$.



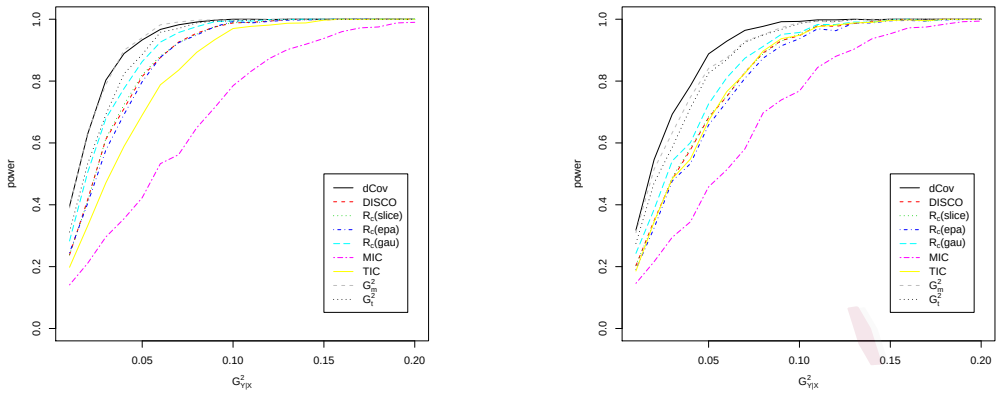
(d) $a = 0.3, p = 10, \sigma_x^2 = 2$, and $\sigma^2 = 1$.

Figure 4: Empirical power with the change of sample size n.

R_c(slice) do not perform well. However, the two kernel methods outperform dCov. This is an example of $\mathbf{X}$ and Y having a nonlinear relationship, and demonstrates the advantage of our method when the dependence is weak, that is, when relations appear in the conditional variance. Additional simulations in the Supplementary Material show similar results.

Example 6.6 The model for this example is taken from Wang, Jiang, and Liu (2016). Let $X \sim U(0,1), Y = f(X) + \epsilon\sigma,$ and $\epsilon \sim N(0,1)$, where $\text{var}\{f(X)\} = 1$. The noise σ^2 changes according to the value of $G^2_{Y|X} = (1+\sigma^2)^{-1}$. We conduct simulations for $f(X) = x$ and $f(X) = sin(2\pi X)$. We use the number of replicates

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0538



(a) (b)

Figure 5: Empirical power when (a) $f(X) = x$, and (b) $f(X) = sin(2\pi X)$.

B $= \lfloor 200 + 5000/n \rfloor$ in each permutation test, and report the results for $n = 225$ and 1,000 tests, as in Wang, Jiang, and Liu (2016).

We compare the performance of the MIC (Reshef et al., 2011) and TIC (Reshef et al., 2016) with that of the other methods. Figure 5 shows these methods do not perform well compared with the other methods. Here, dCov, G_m^2, and G_t^2 perform slightly better than our proposed slicing method and the two kernel methods. This is not surprising, because the relation is strong (in the conditional mean function) where the sample size is reasonable. Note that we did not use the method of Reshef et al. (2011) in Example 6.5 because the measure can only detect the relationship between two continuous variables. If there is a categorical variable, or if we want to examine the relationship between two groups of random vectors, their method does not work properly. Thus, it is not applicable in Example 6.5.

To summarize, for categorical Y, we showed that R_c(slice) is stable and better than dCov. The proposed kernel methods and the dCov method with dummy variables do not have comparable power or stability. This suggests that if Y is categorical, one should use R_c(Slice). For continuous Y, we showed that the proposed kernel methods with R_c(gau) and R_c(epa) outperform dCov, G_m^2, and G_t^2, and the

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0538

discrete methods (DISCO and slicing) when the relationship between X and Y is weak, or when the sample size is small. When the sample size is relatively large, or when the relation is strong, all methods essentially have similar power.

7 Conclusion

We introduced a new class of measures for testing independence that can be used flexibly for continuous and categorical random vectors. We also examined a measure with a particular weight function. Note that our new measure of divergence using characteristic functions differs from those using density functions, each of which possess their own advantages and disadvantages.

We considered a class of density function-based divergence models that, unless they assume a parametric family, need to be estimated nonparametrically. Although a density estimation is usually not difficult, especially when $n > p$ and n is reasonably large, it is perhaps troublesome or impossible when $p > n$ or n is small. The characteristic function-based divergence is not easy to calculate. However, in our case, the index results in a simple Euclidean distance, which provides an alternative choice to existing methods, especially when the accuracy of the density estimation is questionable.

Huo and Székely (2016) discussed a fast computing algorithm for the dCov measure, which reduces the computational complexity to $O(n \log n)$. We believe it is similarly possible to reduce the calculation complexity of the proposed measure. We present a table of computing times for dCov and our current algorithms in the Supplementary Material. Although both use conditional characteristic functions, Wang et al. (2015) developed a conditional independence measure of two random vectors, given a third vector, as a direct extension of dCov, whereas our method determines the independence of two random vectors. A logical direction for future research is to develop a new measure of the conditional independence of two random

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0538

vectors by introducing a third random vector. This possible measure could serve as an alternative to that proposed by Wang et al. (2015).

Although we focused on two random vectors in developing our method, we can extend it to a multi-set of vectors. Ideas for such an extension may be taken from existing methods. For example, the methods of Deheuvels (1981), Genest and Rémillard (2004), Genest, Quessy, and Rémillard (2007), and Kojadinovic and Holmes (2009) could all provide a reasonable framework. In particular, the methods developed by Jin and Matteson (2018), Yao, Zhang, and Shao (2018), Böttcher (2017), and Chakraborty and Zhang (2019) are related to dCov. These methods may prove useful in extending the proposed work on testing mutual independence.

Székely and Rizzo (2013) discussed the bias of the dCov statistic when the dimensions of the random vectors are large. In their work, they constructed an unbiased t-test of independence. Because our measure is defined similarly to theirs, we believe that an analogous calculation will result in a similar unbiased statistic when the dimensions tend to infinity. Dueck et al. (2014) proposed an affinely invariant dCov, and studied its asymptotic properties under a normal distribution when the dimensions go to infinity. Their method could possibly be adopted for our index when the dimensions tend to infinity.

Supplementary Material

The online Supplementary Material contains proofs of the theoretical results, as well as additional numerical studies.

Acknowledgement

The authors thank the editor, associate editor, and two referees for their thoughtful and insightful comments and suggestions. The authors also thank Dr. Derek Young

for reading the final version. Yin's work was supported in part by NSF grant CIF-1813330.

References

Anderson, T. W. (2003). *An Introduction to Multivariate Statistical Analysis*. 3rd Edition. Wiley, New York.

Blomqvist, N. (1950). On a measure of dependence between two random variables. *Ann. Math. Statist.* **21**, 593-600.

Blum, J. R., Kiefer, J. and Rosenblatt, M. (1961). Distribution free tests of independence based on the sample distribution function. *Ann. Math. Statist.* **32**, 485-498.

Böttcher, B. (2017). Dependence structures-estimation and visualization using distance multivariance. *arXiv preprint arXiv:1712.06532.*

Bowman, A. W. and Azzalini, A. (1997). *Applied Smoothing Techniques for Data Analysis: The Kernel Approach with S-Plus Illustrations*. Oxford University Press, Oxford.

Bowman, A. W. and Azzalini, A. (2007). *sm: Nonparametrix Smoothing Methods. R Package (version 2.2).*

Canty, A. and Ripley, B. (2009). *boot: Bootstrap R (S-Plus) Functions. R package (version 1.2-35).*

Chakraborty, S. and Zhang, X. (2019). Distance metrics for measuring joint dependence with application to causal inference. *J. Amer. Statist. Assoc.* DOI: 10.1080/01621459.2018.1513364.

Chen, X., Cook, R. D. and Zou, C. (2015). Diagnostic studies in sufficient dimension reduction. *Biometrika*. **102**, 545-558.

Cook, R. D. (2007). Fisher lecture: dimension reduction in regression. *Statistical Science*. **22**, 1-26.

Cook, R. D. and Weisberg, S. (1991). Discussion of Li (1991). *J. Amer. Statist. Assoc.* **86**, 328-332.

Cook, R. D. and Zhang, X. (2014). Fused estimators of the central subspace in sufficient dimension reduction. *J. Amer. Statist. Assoc.* **109**, 815-827.

Davison, A. C. and Hinkley, D. V. (1997). *Bootstrap Methods and their Application*. Cambridge University Press, Oxford.

Deheuvels, P. (1981). An asymptotic decomposition for multivariate distribution-free tests of independence. *J. Mult. Anal.* **11**, 102-113.

Dueck, J., Edelmann, D., Gneiting, T. and Richards, D. (2014). The affinely invariant distance correlation. *Bernoulli*. **20**, 2305-2330.

Efron, B. and Tibshirani, R. J. (1998). *An Introduction to the Bootstrap*. Chapman and Hall/CRC, Boca Raton, Florida.

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *J. R. Statist. Soc.* B. **70**, 849-911.

Filosi, M., Visintainer, R., Albanese, D., Riccadonna, S., Jurman, G. and Furlanello, C (2017). *minerva: Maximal Information-Based Nonparametric Exploration for Variable Analysis. R Package (version 1.4.7)*.

Genest, C., Quessy, J. F. and Rémillard, B. (2007). Asymptotic local efficiency of Cramër-von Mises tests for multivariate independence. *Ann. Statist.* **35**, 166-191.

Genest, C. and Rémillard, B. (2004). Test of independence and randomness based on the empirical copula process. *Test*. **13**, 335-369.

Gretton, A., Bousquet, O., Smola, A. and Schölkopf, B. (2005). Measuring statistical dependence with Hilbert-Schmidt norms. *International Conference on Algorithmic Learning Theory*. Springer, Berlin, Heidelberg. 63-77.

Härdle, W. (1990). *Applied Nonparametric Regression*. Cambridge University Press, Cambridge.

Härdle, W. H. and Stoker, T. M. (1989). Investigating smooth multiple regression by the method of average derivatives. *J. Amer. Statist. Assoc.* **84**, 986-995.

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

Hollander, M. and Wolfe, D. A. (1999). *Nonparametric Statistical Methods*. 2nd Edition. Wiley, New York.

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

Jin, Z. and Matteson, D. S. (2018). Generalizing distance covariance to measure and test multivariate mutual dependence via complete and incomplete V-statistics. *J. Mult. Anal.* **168**, 304-322.

Kojadinovic, I. and Holmes, M. (2009). Tests of independence among continuous random vectors based on Cramér-von Mises functionals of the empirical copula process. *J. Mult. Anal.* **100**, 1137-1154.

Kullback, S. (1959). *Information Theory and Statistics*. John Wiley and Sons, New York.

Li, B. and Wang, S. (2007). On directional regression for dimension reduction. *J. Amer. Statist. Assoc.* **102**, 997-1008.

Li, B., Wen, S. and Zhu, L. X. (2008). On a projected resampling method for dimension reduction with multivariate responses. *J. Amer. Statist. Assoc.* **103**, 1177-1186.

Li, B., Zha, H. and Chiaromonte, F. (2005). Contour regression: a general approach to dimension reduction. *Ann. Statist.* **33**, 1580-1616.

Li, K. C. (1991). Sliced inverse regression for dimension reduction (with Discussion). *J. Amer. Statist. Assoc.* **86**, 316-342.

Li, R., Zhong, W. and Zhu, L. P. (2012). Feature screening via distance correlation learning. *J. Amer. Statist. Assoc.* **107**, 1129-1139.

Li, L., Zhu, L. and Zhu, L. (2011). Inference on the primary parameter of interest with the aid of dimension reduction estimation. *J. R. Statist. Soc.* B. **73**, 59-80.

Mai, Q. and Zou, H. (2013). The Kolmogorov filter for variable screening in high-dimensional binary classification. *Biometrika*. **100**, 229-234.

Mai, Q. and Zou, H. (2015). The fused Kolmogorov filter: a nonparametric model-free screening method. *Ann. Statist.* **43**, 1471-1497.

Puri, M. L. and Sen, P. K. (1971). *Nonparametric Methods in Multivariate Analysis*. Wiley, New York.

Reshef, D. N., Reshef, Y. A., Finucane, H. K., Grossman, S. R., McVean, G., Turnbaugh, P. J., Lander, E. S., Mitzenmacher, M. and Sabeti, P. C. (2011). Detecting novel associations in large data sets. *Science*. **334**, 1518-1524.

Reshef, Y. A., Reshef, D. N., Finucane, H. K., Sabeti, P. C. and Mitzenmacher, M. (2016). Measuring dependence powerfully and equitably. *J. Mach. Learn. Res.* **17**, 7406-7468.

Rizzo, M. L. and Székely, G. J. (2010). DISCO analysis: A nonparametric extension of analysis of variance. *Ann. Appl. Statist.* **4**, 1034-1055.

Rizzo, M. L. and Székely, G. J. (2018). *energy: E-Statistics: Multivariate Inference via the Energy of Data. R Package (version 1.7-5).*

Samarov, A. M. (1993). Exploring regression structure using nonparametric function estimation. *J. Amer. Statist. Assoc.* **88**, 836-847.

Saviotti, P. P. (1996). *Technological Evolution, Variety and the Economy.* Edward Elgar, Cheltenham.

Sejdinvoc, D., Sriperumbudur, B., Gretton, A. and Fukumizu, K. (2013). Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *Ann. Statist.* **41**, 2263-2291.

Shao, X. and Zhang, J. (2014). Martingale difference correlation and its use in high-dimensional variable screening. *J. Amer. Statist. Assoc.* **109**, 1302-1318.

Sheng, W. and Yin, X. (2013). Direction estimation in single-index models via distance covariance. *J. Mult. Anal.* **122**, 148-161.

Székely, G. J. and Bakirov. N. K. (2003). Extremal probabilities for Gaussina quadratic forms. *Probab. Theory Relat. Fields.* **126**, 184-202.

Székely, G. J., Rizzo, M. L. and Bakirov, N. K. (2007). Measuring and testing dependence by correlation of distances. *Ann. Statist.* **35**, 2769-2794.

Székely, G. J. and Rizzo. M. L. (2009). Brownian distance covariance. *Ann. Appl. Statist.* **3**, 1236-1265.

Székely, G. J. and Rizzo. M. L. (2013). The distance correlation t-test of independence in high dimension. *J. Mult. Anal.* **117**, 193-213.

Taskinen, S., Oja, H. and Randles, R. H. (2005). Multivariate nonparametric tests of independence. *J. Amer. Statist. Assoc.* **100**, 916-925.

Vajda, I. (1989). *Theory of Statistical Inference and Information*. Kluwer Academic Publishers.

Wang, H. and Xia, Y. (2008). Sliced regression for dimension reduction. *J. Amer. Statist. Assoc.* **103**, 811-821.

Wang, X. and Jiang, B. (2016). *Gs: Estimate G-squared via dynamical programming algorithm. R Package (version 1.0).*

Wang, X., Pan, W., Hu, W., Tian, Y. and Zhang, H. (2015). Conditional Distance Correlation. *J. Amer. Statist. Assoc.* **110**, 1726-1734.

Wang, X., Jiang, B. and Liu, J. S. (2016). Generalized R-squared for detecting dependence. *Biometrika*. **104**, 129-139.

Wilks, S. S. (1935). On the independence of k sets of normally distributed statistical variables. *Econometrica.* **3**, 309-326.

Yao, S., Zhang, X. and Shao, X. (2018). Testing mutual independence in high dimension via distance covariance. *J. R. Statist. Soc.* B. **80**, 455-480.

Zhu, L. P., Li, L., Li, R. and Zhu, L. X. (2011). Model-free feature screening for ultrahigh dimensional data *J. Amer. Statist. Assoc.* **106**, 1464-1475.

Zhu, L., Wang, T., Zhu, L. and Ferré, L. (2010). Sufficient dimension reduction through discretization-expectation estimation. *Biometrika.* **97**, 295-304.