Statistica Sinica Preprint No: SS-2021-0100	
Title	The Binary Expansion Randomized Ensemble Test
Manuscript ID	SS-2021-0100
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202021.0100
Complete List of Authors	Duyeol Lee, Kai Zhang and Michael R. Kosorok
Corresponding Author	Michael R. Kosorok
E-mail	kosorok@unc.edu
Notice: Accepted version subject to English editing.	

Statistica Sinica

The Binary Expansion Randomized Ensemble Test

Duyeol Lee[1], Kai Zhang[2] and Michael R. Kosorok[3]

University of North Carolina at Chapel Hill

Abstract: Recently, the binary expansion testing framework was introduced to test the independence of two continuous random variables by utilizing symmetry statistics that are complete sufficient statistics for dependence. We develop a new test based on an ensemble approach that uses the sum of squared symmetry statistics and distance correlation. Simulation studies suggest that this method improves the power while preserving the clear interpretation of the binary expansion testing. We extend this method to tests of independence of random vectors in arbitrary dimension. Through random projections, the proposed binary expansion randomized ensemble test transforms the multivariate independence testing problem into a univariate problem. Simulation studies and data example analyses show that the proposed method provides relatively robust performance compared with existing methods.

Key words and phrases: Nonparametric inference, Nonparametric test of independence, Binary Expansion, Multiple testing, Multivariate analysis.

[1]Department of Statistics and Operations Research, University of North Carolina, Chapel Hill, NC 27599, USA, E-mail: duyeol@live.unc.edu

[2]Department of Statistics and Operations Research, University of North Carolina, Chapel Hill, NC 27599, USA, E-mail: zhangk@email.unc.edu

[3]Department of Biostatistics, University of North Carolina, Chapel Hill, NC 27599, USA, E-mail: kosorok@bios.unc.edu, Tel: +1-919-966-8107, Fax: +1-919-966-3804

1. Introduction

Nonparametric testing of independence is a fundamental problem in statistics and has been studied carefully by many classical papers such as Hoeffding (1948). This problem has been gaining greater interest recently due to its important roles in machine learning and big data analysis.

Many statisticians have contributed many methods through many approaches. Székely et al. (2007); Wang et al. (2017); Han et al. (2017) generalize the idea of correlation and R-squared. Shapiro and Hubert (1979); Friedman and Rafsky (1983); Azadkia and Chatterjee (2019); Deb and Sen (2019); Deb et al. (2021) related dependence to graphs. Heller et al. (2012); Heller et al. (2016); Heller and Heller (2016) studied the distance matrix of ranks. Berrett and Samworth (2019); Kim et al. (2020); Berrett et al. (2020) considered a classical permutation based statistics. Gretton et al. (2008); Chwialkowski and Gretton (2014); Jitkrittum et al. (2017); Pfister et al. (2018); Zhang et al. (2018); Chakraborty and Zhang (2019) take the advantage of the reproducing kernel Hilbert space to develop the Hilbert-Schmidt independence criterion based statistics. Other recent works include Weihs et al. (2018), Ke and Yin (2019), Bodnar et al. (2019), Shi et al. (2020) and Drton et al. (2018). Zhu et al. (2017) also proposed a projection method related to the distance correlation when testing independence. Ex-

cellent reviews can be found in Jaworski et al. (2010) and Josse and Holmes (2016).

One important problem in nonparametric dependence detection is nonuniform consistency, which means that no test can uniformly detect all forms of dependency, as described by Zhang (2019). This problem is particularly severe for nonlinear relationships, which are common in many areas of science. To avoid the power loss due to nonuniform consistency, Zhang (2019) considers the binary expansion statistics (BEStat) framework; this framework examines dependence with a filtration approach induced by the binary expansion of uniformly distributed variables. Zhang (2019) also proposed testing independence of two continuous variables with the framework of maximum binary expansion testing (BET). Rather than one test of independence, this binary expansion approach utilizes a carefully designed sequence of tests based on a filtration to achieve universality. The BET also achieves uniform consistency and is minimax optimal in power (see section 4.2 in Zhang (2019)). In addition, it provides clear interpretability, and it can be implemented efficiently by bitwise operations.

Although the BET works well in testing independence between two variables, two crucial improvements are needed for greater practical applicability. The first requirement is to improve the power of the BET under certain

cases such as linear dependency. The second requirement is to extend the test for testing independence of random vectors. In this paper, we describe a new approach that solves both of these problems. The first problem is addressed by a novel ensemble approach, and the second problem is solved by using one-dimensional random projecting. Due to utilization of both a random projection and an ensemble approach, we call the new method the binary expansion randomized ensemble test (BERET). We show with simulation studies that the proposed method has good power properties.

Through example datasets, we illustrate how the proposed method provides clear interpretability while maintaining good power properties across various dependence structures, including both linear and nonlinear relationships. In a life expectancy example, our method is able to detect three meaningful and interpretable relationships and provide similar p-values as competing methods. In a mortality rate example, we show that the canonical correlation test can be interpretable but fails to detect nonlinear dependence structure. This is unfortunate since the canonical correlation test is the only other method besides the proposed method that has inherent interpretability. In contrast, our method is able to identify meaningful relationships even when there is a nonlinear relationship. In the house price example, the mutual information test fails to reject independence because

the linear relationship is not sufficiently strong. However, our method rejects independence because of its boosted sensitivity to linear relationships, and it is also able to detect interpretable dependence structures including linear relationships. The canonical correlation test also works here and also provides interpretability. However, our method is the only method that has power for detecting both linear and nonlinear relationships as well as being able to illuminate interpretable dependency structures.

This article is organized as follows. Section 2 describes the ensemble method and the BERET procedure. In Section 3, we present simulation studies on performance of the proposed method. Section 4 illustrates our method with three data examples. Concluding remarks are presented in Section 5. Proofs are given in the supplementary material.

2. Proposed Method

2.1 The Binary Expansion Testing Framework

We briefly introduce the BET and useful notations from Zhang (2019). Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be a random sample from distributions of X and Y. If the marginal distributions of X and Y are known, we can use the CDF transformation so that $U = F_X(X)$ and $V = F_Y(Y)$ are each uniformly distributed over $[0, 1]$. The binary expansions of two random variables U

and V can be expressed as $U = \sum_{k=1}^{\infty} A_k/2^k$ and $V = \sum_{k=1}^{\infty} B_k/2^k$ where $A_k \overset{i.i.d.}{\sim} Bernoulli(1/2)$ and $B_k \overset{i.i.d.}{\sim} Bernoulli(1/2)$. The value of each Bernoulli distributed variable can be found via $A_{k'} = I\{U - \sum_{k=1}^{k'-1} A_k/2^k \geq 1/2^{k'}\}$ or $B_{k'} = I\{V - \sum_{k=1}^{k'-1} B_k/2^k \geq 1/2^{k'}\}$. If we truncate the expansions at depth d, then $U_d = \sum_{k=1}^{d} A_k/2^k$ and $V_d = \sum_{k=1}^{d} B_k/2^k$ are two discrete variables that can take 2^d possible values. We define the binary variables $\dot{A}_k = 2A_k - 1$ and $\dot{B}_k = 2B_k - 1$ to express the interaction between them as their products. We call any products of A_k's and B_k's with at least one A_k and one B_k cross interactions. In other words, cross interactions are defined as the variables of the form $\dot{A}_{k_1} \dots \dot{A}_{k_r} \dot{B}_{k'_1} \dots \dot{B}_{k'_t}$ for some $r, t > 0$. We use the following binary integer indexing. Let $\mathbf{a}$ be a d-dimensional binary vector with 1's at $k_1, \dots, k_r$ and 0's otherwise, and let $\mathbf{b}$ be a d-dimensional binary vector with 1's at $k'_1, \dots, k'_t$ and 0's otherwise. With this notation, the cross interaction $\dot{A}_{k_1,\dots,k_r} \dot{B}_{k'_1,\dots,k'_t}$ can be written as $\dot{A}_{\mathbf{a}} \dot{B}_{\mathbf{b}}$. For example, $\dot{A}_1 \dot{A}_3 \dot{B}_2 \dot{B}_4 = \dot{A}_{\mathbf{a}} \dot{B}_{\mathbf{b}}$ where $\mathbf{a} = 1010$ and $\mathbf{b} = 0101$ when $d = 4$.

Let $\dot{A}_{\mathbf{a},i}$ and $\dot{B}_{\mathbf{b},i}$ be the values of $\dot{A}_{\mathbf{a}}$ and $\dot{B}_{\mathbf{b}}$ of ith observation. We denote the sum of the observed binary interaction variables by $S_{(\mathbf{ab})} = \sum_{i=1}^{n} \dot{A}_{\mathbf{a},i} \dot{B}_{\mathbf{b},i}$ with $S_{(\mathbf{00})} = n$. These statistics are referred to as the symmetry statistics. If U_d and V_d are independent, $(S_{(\mathbf{ab})}+n)/2 \sim Binomial(n, 1/2)$ for $\mathbf{a} \neq \mathbf{0}$ and $\mathbf{b} \neq \mathbf{0}$. If marginal distributions are unknown, we use the em-

pirical CDF transformation and then $(\hat{S}_{(\mathbf{ab})}+n)/4 \sim Hypergeometric(n, n/2, n/2)$ where $\hat{S}_{(\mathbf{ab})}$ is a symmetry statistic with empirical CDF transformation.

If we truncate the expansions at depth $d = d_{max}$, the BET procedure at depth d_{max} can be defined as follows. First, we compute all symmetry statistics with $\mathbf{a} \neq \mathbf{0}$ and $\mathbf{b} \neq \mathbf{0}$ for $d = d_{max}$. For each depth $d = 1, \ldots, d_{max}$, we look for the symmetry statistic with the strongest asymmetry and find its p-value. Finally, we use Bonferroni adjustment to obtain a p-value that considers the family-wise error rate.

The BET has several advantages. The test is minimax optimal under certain regulatory conditions. Moreover, it provides both inferences and clear interpretations. For the BET, rejection of independence implies that there is at least one significant cross interaction. Thus, we can find a potential dependence structure in the sample by investigating the detected cross interaction.

2.2 Univariate Independence Testing Procedure

Although the BET shows good performance in many interesting dependency structures, there is room for improvement. In particular, use of the maximum statistic in the BET testing procedure may introduce loss of power when the sparsity assumption in Zhang (2019) is violated. We consider a

test based on the sum of squared symmetry statistics.

Consider a binary expansion test with specified d_{max}. For each depth $d = 1, \ldots, d_{max}$, we can find a set of symmetry statistics $S_{(\mathbf{ab})}$. Let C_d be a set of corresponding $\mathbf{ab}$ indices of depth d. The sets C_d have a nested structure. Since an interaction has different $\mathbf{ab}$ indices for two different d, to avoid confusion, we use the $\mathbf{ab}$ of depth d_{max}. For example, when $d_{max} = 2$, $C_1 = \{1010\}$ and $C_2 = \{0101, 0110, 0111, 1001, 1010, 1011, 1101, 1110, 1111\}$. Now, for each depth d, we introduce two measures of dependence. Suppose $X \in \mathbb{R}$ and $Y \in \mathbb{R}$ be two continuous random variables. The population measure of dependence is defined as

$$\mathcal{B}_d(X, Y) = \frac{1}{(2^d - 1)^2} \sum_{\mathbf{ab} \in C_d} E(\dot{A}_{\mathbf{a}} \dot{B}_{\mathbf{b}})^2, \tag{2.1}$$

for each depth $d = 1, \ldots, d_{max}$. The joint distribution of (U_d, V_d) with a finite d is not an exact model for the joint distribution of (U, V). Therefore, $\mathcal{B}_d(X, Y) = 0$ does not necessarily indicate the independence between (U, V). When d is large, however, we expect that the dependence in (U_d, V_d) precisely approximates that in (U, V).

Let $\{(X_i, Y_i)\}_{i=1}^n$ be a random sample from the joint distribution of (X, Y). The empirical measure of dependence is defined as

$$\mathcal{B}_{n,d}[\{(X_i, Y_i)\}_{i=1}^n] = \frac{1}{(2^d - 1)^2} \sum_{\mathbf{ab} \in C_d} \left(\frac{S_{(\mathbf{ab})}}{n} \right)^2, \tag{2.2}$$

for each depth $d = 1, \ldots, d_{max}$. The following theorem lists some properties of $\mathcal{B}_d(X, Y)$ and $\mathcal{B}_{n,d}[\{(X_i, Y_i)\}_{i=1}^n]$.

Theorem 1. *Suppose X and Y are continuous random variables. The following properties hold:*

(i) $\mathcal{B}_d(X, Y) = 0$ *if and only if* U_d *and* V_d *are independent.*

(ii) $0 \leq \mathcal{B}_d(X, Y) \leq 1$.

(iii) $\mathcal{B}_{n,d}[\{(X_i, Y_i)\}_{i=1}^n] \xrightarrow{a.s.} \mathcal{B}_d(X, Y)$ *as* $n \to \infty$.

(iv) *If X and Y are independent, then* $(2^d - 1)^2 n \mathcal{B}_{n,d}[\{(X_i, Y_i)\}_{i=1}^n] \xrightarrow{d} \chi^2_{(2^d-1)^2}$ *as* $n \to \infty$.

We define the scaled sum of squared symmetry statistics for each depth $d = 1, \ldots, d_{max}$ as

$$\xi_{n,d} = \sum_{\mathbf{ab} \in C_d} \frac{S^2_{(\mathbf{ab})}}{n}. \tag{2.3}$$

By this definition, each $\xi_{n,d}$ can be used to detect the dependencies up to depth d. Consider a test that rejects H_0: "X and Y are independent" if at least one of $\xi_{n,d}$'s is greater than $\xi_{n,d,1-\alpha_d}$, the $1-\alpha_d$ quantile of $\xi_{n,d}$. Then, by Boole's inequality, the upper bound of the type I error is

$$Pr(reject\ H_0 \mid H_0\ is\ true) \leq \sum_{d=1}^{d_{max}} \alpha_d. \tag{2.4}$$

There are many possible versions of the test based on different choices of the α_d's. We remark here that alternatives in C_d for smaller d reflect more global dependencies with lower resolutions. From this point of view, we propose an exponentially decaying approach for choice of α_d. If we choose $\alpha_d = \alpha\gamma^d / \sum_{d=1}^{d_{max}} \gamma^d$ where $0 < \gamma \leq 1$ then the upper bound of the significance level is

$$Pr(reject\ H_0 \mid H_0\ is\ true) \leq \sum_{d=1}^{d_{max}} \frac{\alpha\gamma^d}{\sum_{d=1}^{d_{max}} \gamma^d} = \alpha, \tag{2.5}$$

guaranteeing a level α test. A natural choice of γ is 1;

$$Pr(reject\ H_0 \mid H_0\ is\ true) \leq \sum_{d=1}^{d_{max}} \frac{\alpha}{d_{max}} = \alpha. \tag{2.6}$$

The correct depth where the dependency may present is not known a priori. An appropriate d_{max} should reflect the desired accuracy in the approximation. Considering $\|(U_d, V_d) - (U, V)\| = O_p(2^{-d})$, however, we believe that $d_{max} = 4$ provides a good approximation in practice.

The power of the proposed test can be improved by a compromise between a distance correlation test and multiple testing over interactions. The binary expansion testing framework loses power from the adverse effect of multiplicity control over depth. This loss of power is particularly severe for linear dependency. See a detailed discussion in Section 1.2 in the supplementary material of Zhang (2019). By considering distance correlation

combined with the proposed test, we can mitigate this power loss. The above test is composed of multiple hypothesis tests, and each test has its own set of dependence structures as its alternative hypothesis. Suppose $d_{max} = 4$, there is only one interaction $\dot{A}_{1000}\dot{B}_{1000}$ in $\xi_{n,1}$. The cross interaction $\dot{A}_{1000}\dot{B}_{1000}$ falls in the first or third quadrant of the unit square $[0,1]^2$ when $\dot{A}_{1000}\dot{B}_{1000} = 1$ and in the second or fourth quadrant when $\dot{A}_{1000}\dot{B}_{1000} = -1$. Therefore, $\xi_{n,1} = S^2_{10001000}/n$ represents the strength of linear dependency. If there is any independence test that performs better than $\xi_{n,1}$ under linear dependency, we can replace the test based on $\xi_{n,1}$ with it while maintaining the performance of the test in other dependence structures. Because we are using a Bonferroni correction for the critical values, this replacement still maintains the targeted level of the test. We call this approach as ensemble method because it combines two testing methods. The independence test with Pearson's correlation can be also combined with the proposed test. However, we choose the distance correlation test as it improves power in a wider range of cases and it is equivalent to Pearson's correlation under normality. The proposed procedure consists of the following steps:

Step 1 : Fix $\alpha_1, ..., \alpha_{d_{max}}$ with $\sum_{d=1}^{d_{max}} \alpha_d = \alpha$.

Step 2 : Find the p-value for the distance correlation test.

Step 3 : For each $d = 2, \ldots, d_{max}$, compute $\xi_{n,d}$ and its p-value.

Step 4 : Reject H_0 if at least one of the p-values is less than respective α_d.

To find p-value for each depth $d \geq 2$, we can use either a permutation approach or the asymptotic distribution given in theorem 1, part (iv). Now we investigate the behavior of our test in large samples.

Theorem 2. *Denote the joint distribution of* (U_d, V_d) *by* $\mathbf{P}_{(U_d,V_d)}$ *and the bivariate uniform distribution over* $\{\frac{0}{2^d}, \ldots, \frac{2^d-1}{2^d}\}^2$ *by* $\mathbf{P}_{0,d}$. *For any fixed* $0 < \delta \leq 1/2$, *denote by* $\mathcal{H}_{1,d}$ *the collection of distributions* $\mathbf{P}_{(U_d,V_d)}$ *such that* $TV(\mathbf{P}_{(U_d,V_d)}, \mathbf{P}_{0,d}) \geq \delta$. *Consider the testing problem,*

$$H_0 : \mathbf{P}_{(U_d,V_d)} = \mathbf{P}_{0,d} \quad v.s. \quad H_1 : \mathbf{P}_{(U_d,V_d)} \in \mathcal{H}_{1,d}.$$

Under H_1, *each* $\xi_{n,d} \to \infty$ *as* $n \to \infty$.

Theorem 2 shows that our test statistics, $\xi_{n,d}$'s, go to infinity as sample size increases. Moreover, the distance correlation test is known to be consistent. Therefore, ensemble method is also statistically consistent against the collection of alternatives described in theorem 2.

2.3 Multivariate Independence Testing Procedure

In this section, we develop a generalized independence test for random vectors. The generalization can be made by converting the independence of random vectors into the independence of univariate random variables. The lemma allowing this conversion is stated below.

Lemma 1. *Let* $\mathbf{X} \in \mathbb{R}^p$ *and* $\mathbf{Y} \in \mathbb{R}^q$ *be two random vectors. Then* $\mathbf{X}$ *and* $\mathbf{Y}$ *are independent if and only if* $\mathbf{s}^T\mathbf{X}$ *and* $\mathbf{t}^T\mathbf{Y}$ *are independent for all* $\mathbf{s} \in \mathbb{R}^p, \mathbf{t} \in \mathbb{R}^q$ *with* $\|\mathbf{s}\| = 1$ *and* $\|\mathbf{t}\| = 1$.

This result shows that, to prove independence of random vectors, it is sufficient to consider independence of arbitrary linear combinations of the components. Therefore, the multivariate independence can be tested by checking all possible combinations of $\mathbf{s}$ and $\mathbf{t}$. Because testing all possible combinations cannot be implemented, we consider an approximation of the test by including a finite but reasonably broad number of combinations. Denote hyper unit spheres in $\mathbb{R}^p$ and $\mathbb{R}^q$ by S_p and S_q respectively. Now, for each depth d, we propose two measures of dependence.

Suppose $\mathbf{X} \in \mathbb{R}^p$ and $\mathbf{Y} \in \mathbb{R}^q$ are two random vectors. For $\mathbf{s} \in S_p, \mathbf{t} \in S_q$, we define a measure of dependence for the multivariate setting by

$$\mathcal{B}_d(\mathbf{X}, \mathbf{Y}) = \frac{1}{c_p c_q} \int_{S_q} \int_{S_p} \mathcal{B}_d(\mathbf{s}^T\mathbf{X}, \mathbf{t}^T\mathbf{Y}) \mathrm{d}\mathbf{s}\mathrm{d}\mathbf{t}, \tag{2.7}$$

where $c_p = \frac{2\pi^{p/2}}{\Gamma(p/2)}$ and $c_q = \frac{2\pi^{q/2}}{\Gamma(q/2)}$.

Let $\{(\mathbf{X}_i, \mathbf{Y}_i)\}_{i=1}^n$ be a random sample from the joint distribution of $(\mathbf{X}, \mathbf{Y})$. The empirical measure of dependence is defined as

$$\mathcal{B}_{n,d}[\{(\mathbf{X}_i, \mathbf{Y}_i)\}_{i=1}^n] = \frac{1}{c_p c_q} \int_{S_q} \int_{S_p} \mathcal{B}_{n,d}[\{(\mathbf{s}^T\mathbf{X}_i, \mathbf{t}^T\mathbf{Y}_i)\}_{i=1}^n] \mathrm{d}\mathbf{s}\mathrm{d}\mathbf{t}. \tag{2.8}$$

The following theorem lists some properties of $\mathcal{B}_d(\mathbf{X}, \mathbf{Y})$ and $\mathcal{B}_{n,d}[\{(\mathbf{X}_i, \mathbf{Y}_i)\}_{i=1}^n]$.

Theorem 3. *Suppose the distributions of* $\mathbf{X}$ *and* $\mathbf{Y}$ *are continuous. Let* $U_d^{\mathbf{s}}$ *and* $V_d^{\mathbf{t}}$ *be truncated binary expansions at depth* d *of* $U^{\mathbf{s}}$ *and* $V^{\mathbf{t}}$ *respectively where* $U^{\mathbf{s}} = F_{\mathbf{s}^T\mathbf{X}}(\mathbf{s}^T\mathbf{X})$ *and* $V^{\mathbf{t}} = F_{\mathbf{t}^T\mathbf{Y}}(\mathbf{t}^T\mathbf{Y})$ *for* $\mathbf{s} \in S_p, \mathbf{t} \in S_q$. *Similarity transformations consist of all Euclidean transformations and all (nonzero) scaling (Móri and Székely (2019)). The following properties hold:*

(i) $\mathcal{B}_d(\mathbf{X}, \mathbf{Y}) = 0$ *if and only if* $U_d^{\mathbf{s}}$ *and* $V_d^{\mathbf{t}}$ *are independent for all* $\mathbf{s} \in S_p, \mathbf{t} \in S_q$.

(ii) $0 \le \mathcal{B}_d(\mathbf{X}, \mathbf{Y}) \le 1$.

(iii) $\mathcal{B}_d(\mathbf{X}, \mathbf{Y})$ *is invariant with respect to all similarity transformations.*

(iv) $\mathcal{B}_{n,d}[\{(\mathbf{X}_i, \mathbf{Y}_i)\}_{i=1}^n] \xrightarrow{a.s.} \mathcal{B}_d(\mathbf{X}, \mathbf{Y})$ *as* $n \to \infty$.

Note that $\mathcal{B}_{n,d}[\{(\mathbf{X}_i, \mathbf{Y}_i)\}_{i=1}^n] = E_{\mathbf{S},\mathbf{T}}(\mathcal{B}_{n,d}[\{(\mathbf{S}^T\mathbf{X}_i, \mathbf{T}^T\mathbf{Y}_i)\}_{i=1}^n] \mid \{(\mathbf{X}_i, \mathbf{Y}_i)\}_{i=1}^n)$ where $\mathbf{S}$ and $\mathbf{T}$ follow uniform distributions on S_p and S_q respectively. This

expectation can be estimated by

$$\widehat{\mathcal{B}}^m_{n,d}[\{(\mathbf{X}_i, \mathbf{Y}_i)\}_{i=1}^n] = \frac{1}{m}\sum_{j=1}^{m} \mathcal{B}_{n,d}[\{(\mathbf{S}_j^T\mathbf{X}_i, \mathbf{T}_j^T\mathbf{Y}_i)\}_{i=1}^n], \tag{2.9}$$

where $\{(\mathbf{S}_j, \mathbf{T}_j)\}_{j=1}^m$ is a random sample generated from uniform distributions on S_p and S_q. We call this statistic BERET measure of dependence. The following theorem shows that the BERET measure of dependence is a consistent estimator of the population measure of dependence.

Theorem 4. *Suppose* $\mathbf{X}$ *and* $\mathbf{Y}$ *are continuous random vectors. Then,* $\widehat{\mathcal{B}}^m_{n,d}[\{(\mathbf{X}_i, \mathbf{Y}_i)\}_{i=1}^n] \xrightarrow{a.s.} \mathcal{B}_d(\mathbf{X}, \mathbf{Y})$ *as* $m, n \to \infty$.

Now, to develop an independence test, we define the statistic

$$\zeta^m_{n,d} = n(2^d - 1)^2 \widehat{\mathcal{B}}^m_{n,d}[\{(\mathbf{X}_i, \mathbf{Y}_i)\}_{i=1}^n], \tag{2.10}$$

for each depth $d = 1, ..., d_{max}$. By computing $1 - \alpha_d$ quantiles of $\zeta^m_{n,d}$, for $d = 1, \ldots, d_{max}$, we can consider the test that rejects H_0 : "$\mathbf{X}$ and $\mathbf{Y}$ are independent" if at least one $\zeta^m_{n,d}$, for $d = 1, \ldots, d_{max}$, is greater than $\zeta^m_{n,d,1-\alpha_d}$. If $\sum_{d=1}^{d_{max}} \alpha_d \leq \alpha$, this procedure provides a level α test. To put the proposed test into practice, we estimate the asymptotic null distribution by a random permutation method.

For better performance, under possible linear dependency, we combine this procedure with the distance correlation test as above. If the scales of

the elements in the random vectors differ greatly, normalization may be helpful to reduce the number of $\mathbf{s}$ and $\mathbf{t}$ values to be sampled when the marginal variance of each entry in the random vector cannot degenerate to 0 or diverge to infinity. The following procedure summarizes the approach.

Step 1 : set $\alpha_1, ..., \alpha_{d_{max}}$ with $\sum_{d=1}^{d_{max}} \alpha_d = \alpha$.

Step 2 : normalize marginally each element of the random vectors.

Step 3 : find the p-value for the distance correlation test.

Step 4 : fix $m \in \mathbb{N}$ and generate random sample $s_1, \ldots, s_m$ and $t_1, \ldots, t_m$ from uniform distributions on hyper spheres, respectivley.

Step 5 : for each $d = 2, \ldots, d_{max}$, compute $\zeta_{n,d}^m$ and its p-value by the permutation method.

Step 6 : reject H_0 if at least one of the p-values is less than respective α_d.

We refer to this procedure as binary expansion randomized ensemble testing (BERET) due to its two aspects of random projection and ensemble structure. Again we investigate the behavior of our test in large samples. Theorem 5 shows that BERET is uniformly consistent against the collection of alternatives in the theorem.

Theorem 5. *For any fixed* $0 < \delta \leq 1/2$, *denote by* $\mathcal{H}_{1,d}^{\mathbf{s},\mathbf{t}}$ *the collection of distributions* $\mathbf{P}_{(U_d^{\mathbf{s}},V_d^{\mathbf{t}})}$ *such that* $TV(\mathbf{P}_{(U_d^{\mathbf{s}},V_d^{\mathbf{t}})}, \mathbf{P}_{0,d}) \geq \delta$. *Consider the testing problem,*

$$H_0 : \mathbf{P}_{(U_d^{\mathbf{s}},V_d^{\mathbf{t}})} = \mathbf{P}_{0,d} \text{ for all } \mathbf{s} \in S_p, \mathbf{t} \in S_q$$

$$\text{v.s. } H_1 : \mathbf{P}_{(U_d^{\mathbf{s}},V_d^{\mathbf{t}})} \in \mathcal{H}_{1,d}^{\mathbf{s},\mathbf{t}} \text{ for some } \mathbf{s} \in S_p \text{ and } \mathbf{t} \in S_q.$$

The following properties hold.

(i) *Under* H_1, $\zeta_{n,d}^m \to \infty$ *as* $m, n \to \infty$.

(ii) *Rejection probability of the permutation test is bounded by* α *under* H_0 *and converges to 1 under* H_1 *as* $m, n \to \infty$ *if* $d_{max} \geq d$.

The BERET has the following advantages. First, the method achieves robust power by a compromise between the distance correlation test and multiple testing over interactions (see the simulation results in section 3). The power loss due to multiplicity control over the depth also exists in the multivariate case. By considering the distance correlation result together with the proposed measure of dependence with $d \geq 2$, we can improve power over a wide range of plausible dependencies.

The second benefit of our method is clear interpretability. The issue of interpretability is particularly important in evaluating multivariate relationships. However, most multivariate independence tests provide only the

results of the tests with no information on potential dependence structures in the sample. In contrast, when the proposed test rejects independence, the $\mathbf{s}$ and $\mathbf{t}$ vectors indicate the linear combinations of the vectors that have strong dependencies (see section 2.3 of the supplementary material). Using these vectors, we can detect the possible dependence structures in the sample. See Figure 1 for a three dimensional double helix structure example for illustration, in which white positive regions and blue negative regions of interaction provide the interpretation of global dependency. It can be seen that the double helix structure is detected by two linear combinations. More interesting interpretation examples are provided in Section 4.

[Figure 1 near here]

The third benefit of our method is "invariance". Móri and Székely (2019) introduced axioms for a measure to be a dependence measure. If a measure Δ satisfies $\Delta(f(X), g(Y)) = \Delta(X, Y)$ where f, g are similarity transformations, it is called invariant with respect to similarity transformations. Because of the random projection and the CDF transformation steps in the proposed method, a translation, an orthogonal linear mapping, and a uniform scaling do not affect the value of the measure of dependence.

Lastly, our method provides useful exploratory information for model selection. A small entry in the unit vector $\mathbf{s}$ or $\mathbf{t}$ may indicate that the

corresponding variable may not be related to the other random vector. See data examples in Section 4 for details.

3. Simulation Studies

3.1 Univariate Independence

For comparison, we consider the Hoeffding's D test (Hoeffding (1948)), the distance correlation test (Székely et al. (2007)), the mutual information test (MINTav, Berrett and Samworth (2019)), Fisher's exact scanning method (Ma and Mao (2019)) and the maximum binary expansion test (Zhang (2019)). We use sample size $n = 128$ as a moderate sample size for power comparison. We set the level of the tests to be 0.1 and simulate each scenario 1,000 times. We adopt $d_{max} = 4$ because this depth provides a good approximation to the true distribution. See a detailed discussion in Section 4.5 in Zhang (2019). The p-values of the proposed method are calculated using the asymptotic distribution of theorem 1, part (iv). We verified that the p-value under the null hypothesis was controlled at level 0.1.

We compare the power of the above methods over familiar dependence structures such as linear, parabolic, circular, sine, checkerboard and local relationship described in Zhang (2019). At each noise level $l = 1, \ldots, 10$,

$\epsilon, \epsilon', \epsilon''$ are independent $\mathcal{N}(0, (l/40)^2)$ random variables. U follows the standard uniform distribution. ϑ is a $U[-\pi, \pi]$ random variable. W, V_1, and V_2 follow *multi-Bern*$(\{1,2,3\}, (1/3, 1/3, 1/3))$, *Bern*$(\{2,4\}, (1/2, 1/2))$, and *multi-Bern*$(\{1,3,5\}, (1/3, 1/3, 1/3))$ respectively. G_1, G_2 are generated from $\mathcal{N}(0, 1/4)$. Table 1 summarizes the details of the setting. These scenarios are visually displayed in the supplementary material.

[Table 1 near here]

Figure 2 shows the performance of the six methods. There are two points to notice. First, except for the proposed test, all the other methods show the lowest power in at least one scenario. The ensemble approach and the BET show similar powers across the scenarios except for the linear and local dependency. The ensemble approach considerably improves power in the linear and local dependency scenarios. As discussed previously, the ensemble approach utilizes the information on dependence remaining in the symmetry statistics that is not reflected in calculation of the maximum binary expansion testing. Therefore, small asymmetries in many symmetry statistics can be combined to provide a significant result in the ensemble approach when the sparsity assumption is violated. This result is related to the second finding that the ensemble approach outperforms Fisher's exact scanning in both global and local dependence structures. Zhang (2019)

reported that maximum binary expansion testing provides better power for global dependence structures, whereas Fisher's exact scanning performs better for local dependence structures. The simulation results suggest that the ensemble approach works better than Fisher's exact scanning even in the local dependency scenario.

[Figure 2 near here]

3.2 Multivariate Independence

Although the proposed method can be applied to arbitrary p and q, we choose $p = 2$ and $q = 1$ for better illustration. We compare the proposed method with the distance correlation test (Székely et al. (2007)), the Heller-Heller-Gorfine test (Heller et al. (2012)), the d-variable Hilbert-Schmidt independence criterion (Gretton et al. (2008)), and the mutual information test (MINTav, Berrett and Samworth (2019)). We again use sample size $n = 128$. We set the level of the tests to be 0.1 and simulate each scenario 1,000 times. For our method, we adopt $m = 30$ because there is no considerable difference in performance compared with larger m's such as $m = 360$. We also use a permutation method with 1,000 replicates to calculate the p-values of the proposed approach. We verified that the p-value under the null hypothesis was controlled at level 0.1.

We compare the power of the methods over dependence structures such as linear, parabolic, spherical, sine, and local dependence structures. These scenarios are generalized from the univariate dependence simulations. In addition, we include an additional interesting relationship, the double helix structure. At each noise level $l = 1, \ldots, 10$, $\epsilon, \epsilon', \epsilon''$ are independent $\mathcal{N}(0, (l/40)^2)$ random variables. U_1, U_2 follow the standard uniform distribution. ϑ follows $\mathrm{U}[0, 4\pi]$. G_1, G_2, G_3 are independent $\mathcal{N}(0, 1/4)$ random variables. I follows the Rademacher distribution. Table 2 summarizes the details of the setting. These three dimensional scenarios are visually displayed in the supplementary material.

[Table 2 near here]

Before we compare the statistical performances, we report the computation time of 100 runs of each method in the following table.

[Table 3 near here]

Figure 3 shows the simulation results. The BERET provides the best power in more complex dependency structures such as sine and double helix dependency. It outperforms the distance correlation test and the d-variable Hilbert-Schmidt independence criterion in at least five scenarios compared with each testing method separately. Moreover, our method provides stable

results across the scenarios considered. It ranks at least third place in all scenarios. The mutual information test performs best in the highest number of scenarios. In linear and sine relationships, however, there is significant loss of power in the mutual information test compared with the proposed method. A point to notice is that our method provides additional insight. Other methods only provide test results of independence, but our method provides potential dependence structures as well as test results. The simulation results show that BERET provides competitive performance while providing a much clearer interpretation.

[Figure 3 near here]

4. Data Examples

4.1 Life Expectancy

We use the proposed method to test independence between geographic location and life expectancy and compare its performance with the performance of other methods, i.e., the distance correlation test (dCor), the Heller-Heller-Gorfine test (HHG), the mutual information test (MINT), and the canonical correlation test (CC). We include the canonical correlation test because it also provides some insight on dependence structure as does the proposed method. For the proposed method, we set $d_{max} = 4$ and $m = 30$.

The p-value of the test is calculated by a permutation method with 1,000 replicates. The dataset is obtained from the life expectancy report released by the World Health Organization in 2016. The dataset includes males and females and total life expectancy of 189 countries and special administrative regions estimated in 2015. The latitudes (X_1), longitudes (X_2), and total life expectancies (Y) are used in the analysis. Table 4 presents the testing results for the five different methods. All five tests provide p-values close to 0, indicating a significant dependence between geographic location and life expectancy.

[Table 4 near here]

To identify the dependence structure, we investigate the symmetry statistics. Figure 4 shows the three largest symmetry statistics and the corresponding $\mathbf{s}$'s. The most asymmetric result is shown in the first row. It is $\dot{A}_2\dot{B}_1$ with $\mathbf{s} = (0.516,\ 0.857)^T$. The horizontal axis is the empirical cumulative distribution function transformation of $0.516X_1 + 0.857X_2$, wherein a smaller value implies the country is located in the southwest and a larger value implies the northeast. There are four different groups, from the first one in the upper left to the fourth group in the lower right. Each blue cell represents a specific region, America, Africa, Europe and Asia from left to right. The countries in America and Europe show higher life

expectancy than countries in Africa and Asia. The four points in the top right corner are Hong Kong, Japan, Macau and South Korea. They can be interpreted as potential outliers distinct from the global pattern.

[Figure 4 near here]

The second row shows that there is a positive relationship between latitude and life expectancy. That is, the countries in North America, Europe and Northeast Asia have higher life expectancy than countries in Africa, South America and the other parts of Asia. The last row shows that a circular dependency can exist, which indicates that countries in America and Asia have a medium life expectancy, whereas countries around the prime meridian have different life expectancies, higher in Europe and lower in Africa. These findings prove clearly that our method detects the dependence structures between geographic location and life expectancy.

The canonical correlation analysis also can be used to find information on dependence structure. The canonical correlation is 0.43, and it is calculated using $0.991X_1 - 0.137X_2$ and Y. The coefficients of X_1 and X_2 are similar to the elements of $\mathbf{s}$ in the result of the proposed method in the second row. However, canonical correlation provides information only on the linear dependence structure, whereas our method provides richer information by considering various nonlinear dependence structures.

4.2 Mortality Rate

The second case is the relationship between mortality rate, birth rate and income level. We use the Central Intelligence Agency's world fact data, estimated in 2018. The dataset includes income level (X_1), birth rate (X_2), and mortality rate (Y) of 224 countries and special administrative regions. The p-values of the five methods are presented in Table 4. Once again, the proposed method and two other methods provide p-values close to 0, which rejects the null hypothesis, whereas the mutual information test and canonical correlation fail to reject it. The poor performance of canonical correlation can be explained by investigating the results of our method. The strongest asymmetry is given in Figure 5, which shows a strong quadratic relationship. This relationship explains the failure of canonical correlation to work for the data we use here. Although the canonical correlation test provides both inference and information on dependence structure, it performs poorly in nonlinear dependency settings.

[Figure 5 near here]

For explanation of the observed quadratic relationship one must point to two conflicting phenomena. The first one is that in developed countries the birth rates are low, but the mortality rates are high due to population

aging. In developing countries, however, the birth rates are high from lack of family planning and the mortality rates are also high due to insufficient public health. Thus, mortality rates are high in countries with low or high birth rates. The BERET detects interesting structure that can be explained by widely recognized relationships between mortality rate and birth rate.

4.3 House Price

The third data example is the market historical dataset of real estate from the University of California, Irvine machine learning repository. The data includes 414 transactions from the Xindan district of Taipei between August 2012 and July 2013. We use these data to detect the relationship between geographic location and house price. The p-values of the five methods are presented in Table 4. All methods except the mutual information test provide p-values close to 0, which is consistent with the commonly assumed relationship between location and house price in a city. The mutual information test fails to reject the independence. Figure 6 presents the two strongest dependencies identified by the proposed method.

[Figure 6 near here]

The symmetry statistic with the strongest asymmetry is $\dot{A}_1\dot{B}_1$, which means that there may be a linear relationship between geographic location

and house price. The corresponding $\mathbf{s}$ for the horizontal axis is $(0.964, 0.268)$. That is, houses have higher values in the north and lower values in the south. It is because the central part of Taipei is above the Xindan district. The symmetry statistic with the second strongest asymmetry is $\dot{A}_1\dot{A}_2\dot{B}_1$. The corresponding $\mathbf{s}$ for the horizontal axis is $(0.215, 0.977)^T$. That is, house prices are high at the center of the district, where two main roads intersect, and prices fall towards the periphery. These results accord closely with the general characteristics of real estate prices in a city. Therefore, we can conclude that the proposed method properly detects the relationships between house price and geographic location.

5. Conclusion

Detection of dependence in a distribution-free setting is an important problem in statistics. Existing methods may have challenges with detecting complicated dependence structures. The distance correlation test, for example, does not detect circular dependency well, whereas it provides good powers in linear, parabolic, and sine settings in simulation studies. The binary expansion testing procedure in Zhang (2019) suggests a novel way to solve this problem. However, it is limited to the independence test of two random variables and there is room for enhancement of power when

the sparsity assumption is violated.

In this paper, we introduce an ensemble approach and a binary expansion randomized ensemble test. The ensemble approach uses both the sum of squared symmetric statistics and the distance correlation test. It shows better power in linear and local settings while maintaining power for other dependence structures. Moreover, it can be easily generalized to an independence test for the multivariate setting, the binary expansion randomized ensemble test. By random projections, the BERET transforms the multivariate independence testing problem into a univariate testing problem. The BERET also maintains the clear interpretability of the maximum binary expansion testing.

Simulation studies suggest that the power of the BERET is advantageous compared with a range of competitors considered in many meaningful dependence structures. Investigation of three data examples shows that the BERET reveals hidden dependence structures from the data while maintaining a level of power similar to the best of the competing methods.

Several improvements are worth studying in the future. For instance, we can choose a different method of combining symmetry statistics for better performance. It is also useful to derive the limiting null distribution of the test statistic for the multivariate setting to avoid a permutation method.

Supplementary Materials

The supplementary material provides technical details and proofs.

Acknowledgements

This research was partially supported by DMS-1613112, IIS-1633212, and DMS-1916237 from the National Science Foundation and a grant P01 CA142538 from the National Cancer Institute.

References

Azadkia, M. and S. Chatterjee (2019). A simple measure of conditional dependence. *arXiv preprint arXiv:1910.12327*.

Berrett, T. B., I. Kontoyiannis, and R. J. Samworth (2020). Optimal rates for independence testing via u-statistic permutation tests. *arXiv preprint arXiv:2001.05513*.

Berrett, T. B. and R. J. Samworth (2019). Nonparametric independence testing via mutual information. *Biometrika 106*(3), 547–566.

Bodnar, T., H. Dette, and N. Parolya (2019). Testing for independence of large dimensional vectors. *The Annals of Statistics 47*(5), 2977–3008.

Chakraborty, S. and X. Zhang (2019). A new framework for distance and kernel-based metrics in high dimensions. *arXiv preprint arXiv:1909.13469*.

REFERENCES

Chwialkowski, K. and A. Gretton (2014). A kernel independence test for random processes. In *International Conference on Machine Learning*, pp. 1422–1430. PMLR.

Deb, N., B. B. Bhattacharya, and B. Sen (2021). Efficiency lower bounds for distribution-free hotelling-type two-sample tests based on optimal transport. *arXiv preprint arXiv:2104.01986*.

Deb, N. and B. Sen (2019). Multivariate rank-based distribution-free nonparametric testing using measure transportation. *arXiv preprint arXiv:1909.08733*.

Drton, M., F. Han, and H. Shi (2018). High dimensional consistent independence testing with maxima of rank correlations. *arXiv preprint arXiv:1812.06189*.

Friedman, J. H. and L. C. Rafsky (1983). Graph-theoretic measures of multivariate association and prediction. *The Annals of Statistics*, 377–391.

Gretton, A., K. Fukumizu, C. H. Teo, L. Song, B. Schölkopf, and A. J. Smola (2008). A kernel statistical test of independence. In *Advances in neural information processing systems*, pp. 585–592.

Han, F., S. Chen, and H. Liu (2017). Distribution-free tests of independence in high dimensions. *Biometrika 104*(4), 813–828.

Heller, R. and Y. Heller (2016). Multivariate tests of association based on univariate tests. In *Advances in Neural Information Processing Systems*, pp. 208–216.

Heller, R., Y. Heller, and M. Gorfine (2012). A consistent multivariate test of association based on ranks of distances. *Biometrika 100*(2), 503–510.

REFERENCES

Heller, R., Y. Heller, S. Kaufman, B. Brill, and M. Gorfine (2016). Consistent distribution-free k-sample and independence tests for univariate random variables. *The Journal of Machine Learning Research 17*(1), 978–1031.

Hoeffding, W. (1948). A non-parametric test of independence. *The annals of mathematical statistics*, 546–557.

Jaworski, P., F. Durante, W. K. Hardle, and T. Rychlik (2010). *Copula theory and its applications*, Volume 198. Springer.

Jitkrittum, W., Z. Szabó, and A. Gretton (2017). An adaptive test of independence with analytic kernel embeddings. In *International Conference on Machine Learning*, pp. 1742–1751. PMLR.

Josse, J. and S. Holmes (2016). Measuring multivariate association and beyond. *Statistics surveys 10*, 132.

Ke, C. and X. Yin (2019). Expected conditional characteristic function-based measures for testing independence. *Journal of the American Statistical Association*.

Kim, I., S. Balakrishnan, and L. Wasserman (2020). Minimax optimality of permutation tests. *arXiv preprint arXiv:2003.13208*.

Ma, L. and J. Mao (2019). Fisher exact scanning for dependency. *Journal of the American Statistical Association 114*(525), 245–258.

Móri, T. F. and G. J. Székely (2019). Four simple axioms of dependence measures. *Metrika 82*(1), 1–16.

REFERENCES

Pfister, N., P. Bühlmann, B. Schölkopf, and J. Peters (2018). Kernel-based tests for joint independence. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 80*(1), 5–31.

Shapiro, C. P. and L. Hubert (1979). Asymptotic normality of permutation statistics derived from weighted sums of bivariate functions. *The Annals of Statistics 7*(4), 788–794.

Shi, H., M. Drton, and F. Han (2020). Distribution-free consistent independence tests via center-outward ranks and signs. *Journal of the American Statistical Association*, 1–16.

Székely, G. J., M. L. Rizzo, and N. K. Bakirov (2007). Measuring and testing dependence by correlation of distances. *The annals of statistics 35*(6), 2769–2794.

Wang, X., B. Jiang, and J. S. Liu (2017). Generalized r-squared for detecting dependence. *Biometrika 104*(1), 129–139.

Weihs, L., M. Drton, and N. Meinshausen (2018). Symmetric rank covariances: a generalized framework for nonparametric measures of dependence. *Biometrika 105*(3), 547–562.

Zhang, K. (2019). Bet on independence. *Journal of the American Statistical Association 114*(528), 1620–1637.

Zhang, Q., S. Filippi, A. Gretton, and D. Sejdinovic (2018). Large-scale kernel methods for independence testing. *Statistics and Computing 28*(1), 113–130.

Zhu, L., K. Xu, R. Li, and W. Zhong (2017). Projection correlation between two random vectors. *Biometrika 104*(4), 829–843.

Table 1: Simulation scenarios for univariate independence test

Scenario	Generation of X	Generation of Y
Linear	$X = U$	$Y = X + 6\epsilon$
Parabolic	$X = U$	$Y = (X - 0.5)^2 + 1.5\epsilon$
Circular	$X = \cos\vartheta + 2\epsilon$	$Y = \sin\vartheta + 2\epsilon'$
Sine	$X = U$	$Y = \sin(4\pi X) + 8\epsilon$
Checkerboard	$X = W + \epsilon$	$Y = \begin{cases} V_1 + 4\epsilon' & \text{if } W = 2 \\ V_2 + 4\epsilon'' & \text{otherwise} \end{cases}$
Local	$X = G_1$	$Y = \begin{cases} X + \epsilon & \text{if } 0 \le G_1 \le 1 \text{ and } 0 \le G_2 \le 1 \\ G_2 & \text{otherwise} \end{cases}$

Table 2: Simulation scenarios for multivariate independence testing

Scenario	Generation of X	Generation of Y
Linear	$\mathbf{X} = \begin{pmatrix} U_1 \\ U_2 \end{pmatrix}$	$\mathbf{Y} = X_1 + X_2 + 7\epsilon$
Parabolic	$\mathbf{X} = \begin{pmatrix} U_1 \\ U_2 \end{pmatrix}$	$\mathbf{Y} = (X_1 - 0.5)^2 + (X_2 - 0.5)^2 + 1.5\epsilon$
Spherical	$\mathbf{X} = \begin{pmatrix} \frac{G_1}{\sqrt{G_1^2+G_2^2+G_3^2}} \\ \frac{G_2}{\sqrt{G_1^2+G_2^2+G_3^2}} \end{pmatrix}$	$\mathbf{Y} = \frac{G_3}{\sqrt{G_1^2+G_2^2+G_3^2}} + 3\epsilon$
Sine	$\mathbf{X} = \begin{pmatrix} U_1 \\ U_2 \end{pmatrix}$	$\mathbf{Y} = \sin(5\pi X_1) + 4\epsilon$
Double helix	$\mathbf{X} = \begin{pmatrix} Icos(\vartheta) + 1.5\epsilon \\ Isin(\vartheta) + 1.5\epsilon' \end{pmatrix}$	$\mathbf{Y} = \frac{\vartheta}{2} + 2\epsilon''$
Local	$\mathbf{X} = \begin{pmatrix} G_1 \\ G_2 \end{pmatrix}$	$\mathbf{Y} = \begin{cases} \frac{X_1}{\sqrt{2}} + \frac{X_2}{\sqrt{2}} + \frac{\epsilon}{2}, & \text{if } 0 \le G_1 + G_2 \le 2 \text{ and } 0 \le G_3 \le 1. \\ G_3, & \text{otherwise.} \end{cases}$

Table 3: Computing time (in seconds) of each method for 100 runs

	BERET	dCor	HHG	d-HSIC	MINT
CPU Time (seconds)	74.89	0.17	510.42	16.96	65.19

Table 4: p-values from five tests of independence

	BERET	dCor	HHG	MINT	CC
Life expectancy	<0.0001	<0.0001	0.0010	0.0010	<0.0001
Mortality rate	0.0040	0.0050	0.0010	0.3077	0.4303
House price	<0.0001	<0.0001	0.0010	0.6204	<0.0001

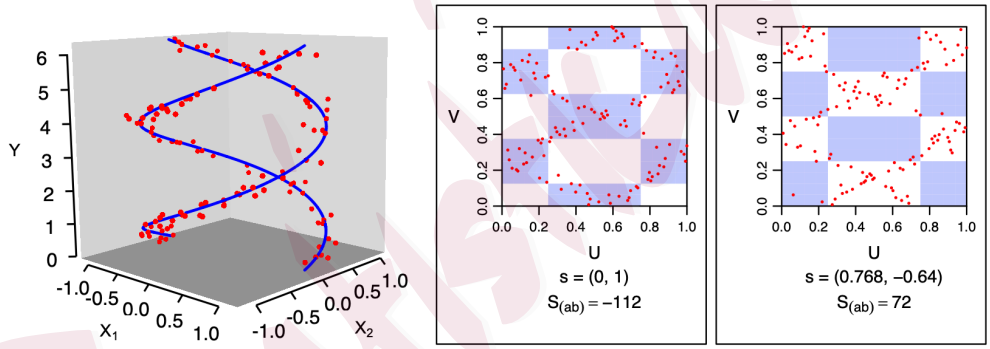

Figure 1: The first plot shows a sample with double helix dependency between a random vector $(X_1\ X_2)^T$ and a random variable Y with $n = 128$. The second and third plots show the linear combinations of X_1 and X_2 with the strongest asymmetries and the corresponding symmetry statistics $(S_{(\mathbf{ab})})$. Positive regions $(\dot{A}_{\mathbf{a}}\dot{B}_{\mathbf{b}} = 1)$ are in white and negative regions $(\dot{A}_{\mathbf{a}}\dot{B}_{\mathbf{b}} = -1)$ are in blue.

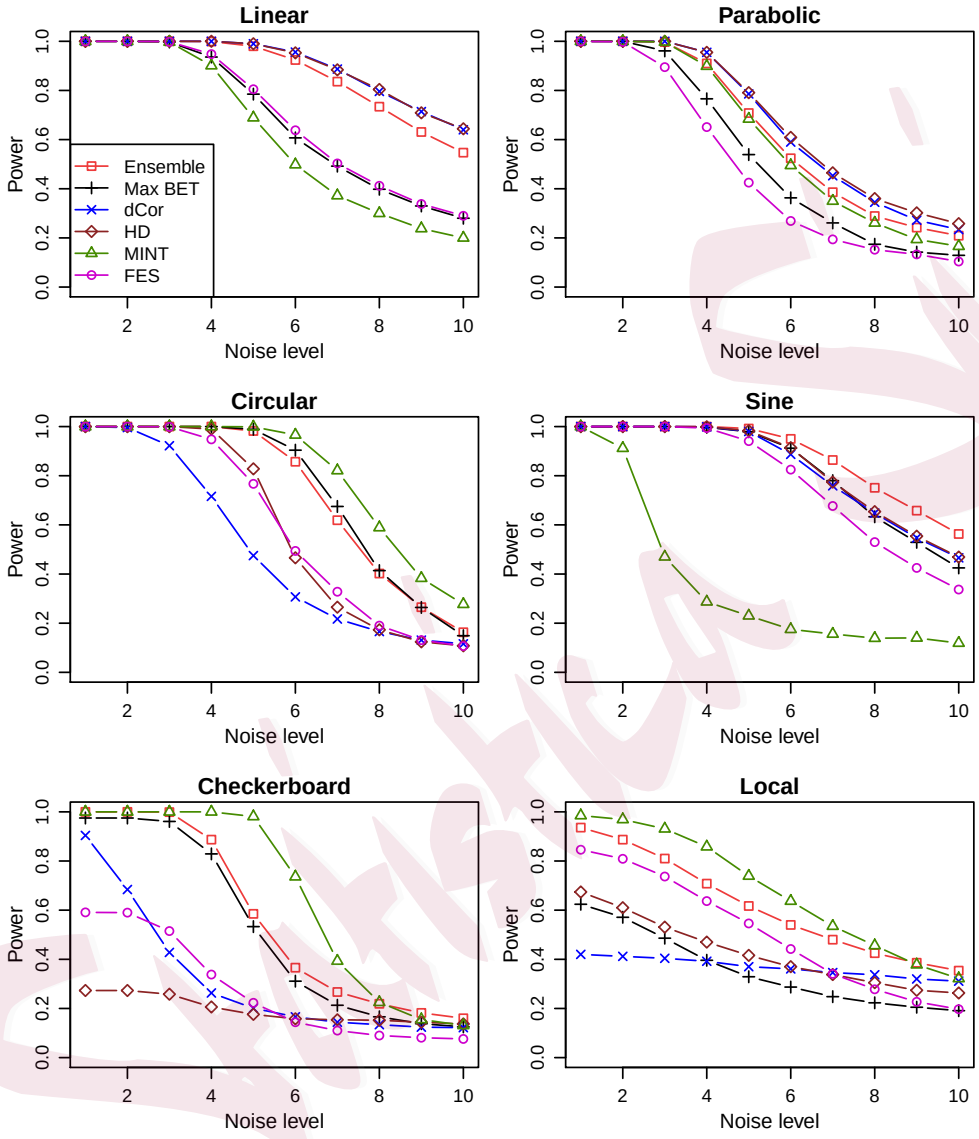

Figure 2: Comparison of powers from six tests of independence: the binary expansion randomized ensemble test with $d_{max} = 4$ (square), the maximum binary expansion test with $d_{max} = 4$ (plus sign), the distance correlation test (cross), Hoeffding's D(diamond), the mutual information test (triangle), and Fisher exact scanning (circle).

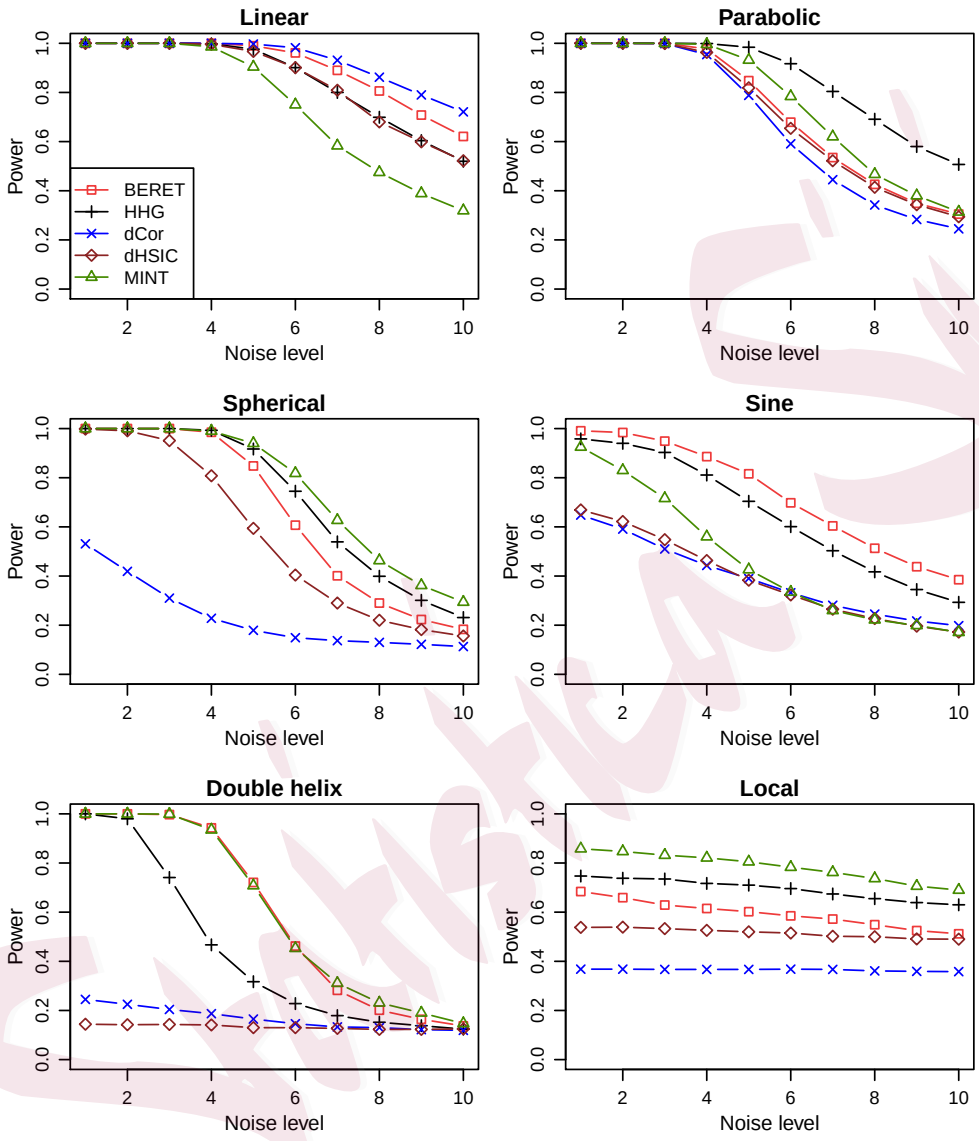

Figure 3: Comparison of powers from five tests of independence: the binary expansion randomized ensemble test with $d_{max} = 4$ (square), the Heller-Heller-Gorfine test (plus sign), the distance correlation test (cross), the d-variable Hilbert-Schmidt independence criterion (diamond), and the mutual information test (triangle).

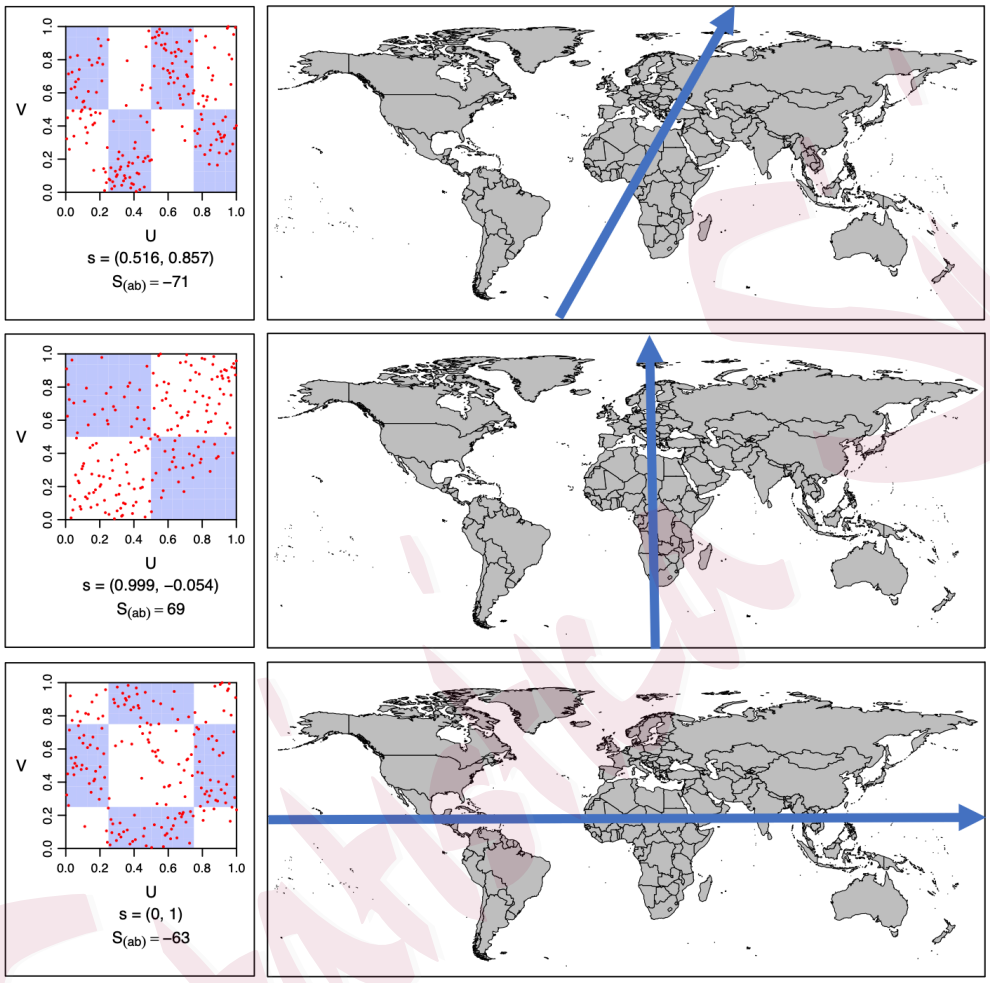

Figure 4: The three strongest dependency structures between geographic location and life expectancy. They also present the corresponding values of the symmetry statistics ($S_{(\mathbf{ab})}$) and the coefficients of the linear combination ($\mathbf{s}$) of X_1 and X_2. The blue arrows in the world maps represent the horizontal axes in the scatterplots.

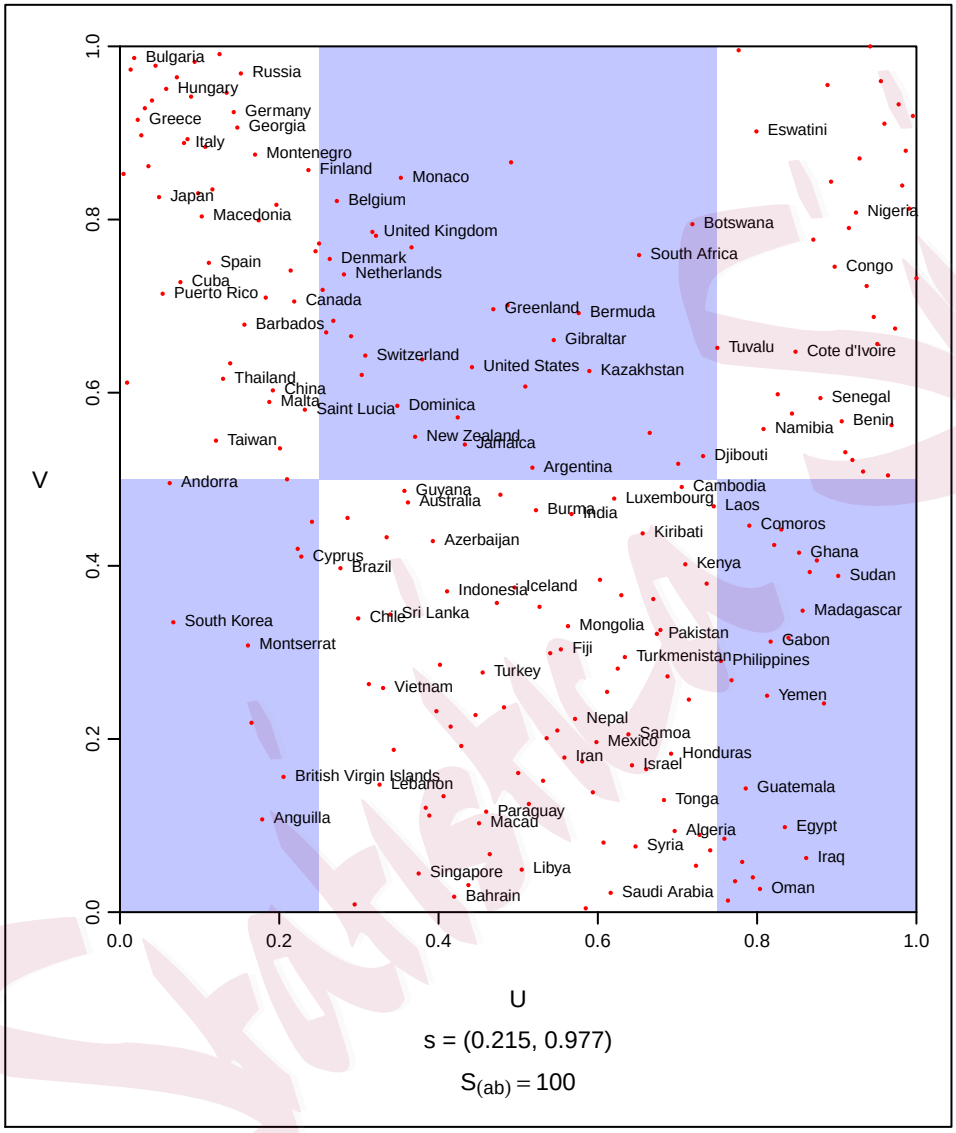

Figure 5: The plot shows the strongest dependency structure between birth rate, income level and mortality rate. It also presents the corresponding value of the symmetry statistic ($S_{(\mathbf{ab})}$) and the coefficients of the linear combination ($\mathbf{s}$) of X_1 and X_2.

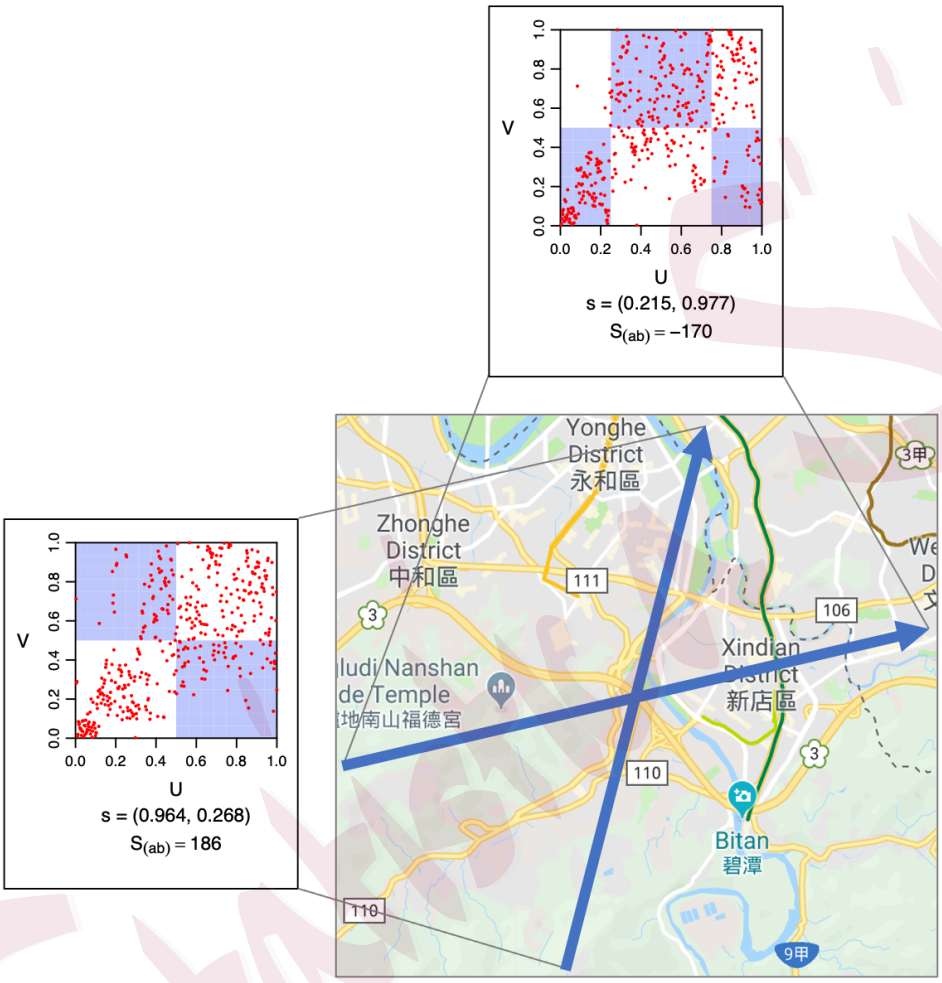

Figure 6: The plots show the two strongest dependency structures between geographic location and house price. The plots also present the values of the symmetry statistics ($S_{(\mathbf{ab})}$) and the coefficients in the linear combinations $\mathbf{s}$ and $\mathbf{t}$. The blue arrows in the map represent the horizontal axes in the scatterplots.