The Chi-Square Test of Distance Correlation

Cencheng Shen[*1], Sambit Panda[2], and Joshua T. Vogelstein [†2,3]

[1]Department of Applied Economics and Statistics, University of Delaware
[2]Institute for Computational Medicine, Department of Biomedical Engineering, Johns Hopkins University
[3]Center for Imaging Science, Kavli Neuroscience Discovery Institute, Johns Hopkins University

May 31, 2021

Abstract

Distance correlation has gained much recent attention in the data science community: the sample statistic is straightforward to compute and asymptotically equals zero if and only if independence, making it an ideal choice to discover any type of dependency structure given sufficient sample size. One major bottleneck is the testing process: because the null distribution of distance correlation depends on the underlying random variables and metric choice, it typically requires a permutation test to estimate the null and compute the p-value, which is very costly for large amount of data. To overcome the difficulty, in this paper we propose a chi-square test for distance correlation. Method-wise, the chi-square test is nonparametric, extremely fast, and applicable to bias-corrected distance correlation using any strong negative type metric or characteristic kernel. The test exhibits a similar testing power as the standard permutation test, and can be utilized for K-sample and partial testing. Theory-wise, we show that the underlying chi-square distribution well approximates and dominates the limiting null distribution in upper tail, prove the chi-square test can be valid and universally consistent for testing independence, and establish a testing power inequality with respect to the permutation test.

Keywords: unbiased distance covariance, centered chi-square distribution, nonparametric test, testing independence

---

[*]shenc@udel.edu
[†]jovo@jhu.edu

1 Introduction

Given pairs of observations $(x_i, y_i) \in \mathbb{R}^p \times \mathbb{R}^q$ for $i = 1, \ldots, n$, assume they are independently identically distributed as F_{XY}. Two random variables are independent if and only if the joint distribution equals the product of the marginal distributions. The statistical hypothesis for testing independence is

$$H_0 : F_{XY} = F_X F_Y,$$

$$H_A : F_{XY} \neq F_X F_Y.$$

Detecting the potential relationships underlying the sample data has long been a fundamental question in theoretical and applied research. The traditional Pearson correlation (Pearson, 1895) has been a valuable tool in quantifying the linear association and applied in many branches of statistics and machine learning. To detect all types of dependence structures, a number of universally consistent methods have been proposed recently, such as the distance correlation (Szekely et al., 2007; Szekely and Rizzo, 2009), the Hilbert-Schmidt independence criterion (Gretton et al., 2005; Gretton and Gyorfi, 2010), the Heller-Heller-Gorfine statistics (Heller et al., 2013, 2016), the multiscale graph correlation (Vogelstein et al., 2019; Shen et al., 2020), among many others. The Hilbert-Schmidt independence criterion (HSIC) can be thought of as a kernel version of distance correlation and vice versa (Sejdinovic et al., 2013; Shen and Vogelstein, 2021), and the multiscale graph correlation is an optimal local version of distance correlation. These universally consistent dependence measures have been applied to two-sample testing (Rizzo and Szekely, 2016; Panda et al., 2021), conditional and partial testing (Fukumizu et al., 2007; Szekely and Rizzo, 2014; Wang et al., 2015), feature screening (Li et al., 2012; Balasubramanian et al., 2013; Zhong and Zhu, 2015; Wang et al., 2019), clustering (Szekely and Rizzo, 2005; Rizzo and Szekely, 2010), time-series (Zhou, 2012; Fokianos and Pitsillou, 2018; Mehta et al., 2019), graph testing (Shen et al., 2020; Xiong et al., 2020), etc.

To populate these methods to big data analysis, a big hurdle is the time complexity. Computing the distance correlation typically requires $\mathcal{O}(n^2)$; and to compute its p-value for testing, the standard approach is to estimate the null distribution of distance correlation via permutation, which

requires $\mathcal{O}(rn^2)$ where r is the number of random permutations and typically at least 100 or more. In comparison, for one-dimensional ($p = q = 1$) data the Pearson correlation can be computed in $\mathcal{O}(n)$, and a Pearson correlation t-test is readily available to compute the p-value in constant time complexity $\mathcal{O}(1)$. The computational advantage makes Pearson extremely fast and attractive in practice. Recent works have successfully expedited the distance correlation computation into $\mathcal{O}(n \log(n))$ under one-dimensional data and Euclidean distance (Huo and Szekely, 2016; Chaudhuri and Hu, 2019). The testing part, however, remains difficult and less well-understood, because the null distribution of distance correlation has no fixed nor known density in general. Some notable works include the distance correlation t-test (Szekely and Rizzo, 2013), HSIC Gamma test (Gretton and Gyorfi, 2010), and the subsampling approach (Zhang et al., 2018), which provided special treatments and very valuable insights into the problem.

In this paper, we propose a chi-square test for the bias-corrected distance correlation. It is simple and straightforward to use, has a constant time complexity without the need to permute nor subsampling nor parameter estimation, and does not rely on any distributional assumption on data. In particular, the bias-corrected distance correlation can be computed in $\mathcal{O}(n \log n)$ under one-dimensional data and Euclidean distance, which renders the method comparable in speed to the Pearson correlation t-test and scalable to billions of observations. The test is applicable to any multivariate data, any strong negative type metric or characteristic kernel, and can be utilized for K-sample and partial testing.

Theory-wise, we prove the chi-square test is universally consistent and valid for sufficiently large n and sufficiently small type 1 error level α (generally $n \geq 20$ and $\alpha \leq 0.05$ suffice), has a similar testing power as the permutation test, and is the most powerful among all valid tests of distance correlation using known distributions (i.e., any test that is based on a fixed distribution, such as using t-test or F-test). In particular, we prove that the underlying chi-square distribution can well-approximate and dominate the limiting null distribution of the bias-corrected distance correlation in upper tail, and establish a testing power inequality among the chi-square test, the distance correlation t-test from Szekely and Rizzo (2013), and the standard permutation test.

The advantages of the chi-square test are supported by simulations and real data experiments.

The code are openly available on Github and implemented in Matlab[1] and Python[2]. The Appendix includes detailed background information, all theorem proofs and intermediate results, and detailed simulation functions.

2   Method

2.1   The Distance Correlation Chi-Square Test

The proposed chi-square test is stated in Algorithm 1: given paired sample data $(\mathbf{X}, \mathbf{Y}) \in \mathbb{R}^{n \times (p+q)}$, first compute the bias-corrected sample distance correlation $C = \mathtt{Dcor}_n(\mathbf{X}, \mathbf{Y})$, then take the p-value as $p = Prob(nC < \chi_1^2 - 1)$. Then at any pre-specified type 1 error level α, the independence hypothesis is rejected if and only if $p < \alpha$. Unless mentioned otherwise, in this paper distance correlation always means the bias-corrected sample distance correlation (see Appendix for the algebraic expression).

As the chi-square distribution is standard in every software package, the p-value computation takes $\mathcal{O}(1)$ regardless of sample size, which is much faster than the standard permutation approach requiring $\mathcal{O}(rn^2)$. The statistic computation and thus Algorithm 1 on its whole require $\mathcal{O}(n^2)$ in general.

Algorithm 1 The Distance Correlation Chi-Square Test for Independence

---

Input: Paired sample data $(\mathbf{X}, \mathbf{Y}) = \{(x_i, y_i) \in \mathbb{R}^{p+q} \text{ for } i \in [n]\}$.
Output: The bias-corrected distance correlation C and its p-value p.
  function FASTTEST$(\mathbf{X}, \mathbf{Y})$
    $C = \mathtt{Dcor}_n(\mathbf{X}, \mathbf{Y})$;   ▷ the bias-corrected distance correlation
    $p = 1 - F_{\chi_1^2 - 1}(n \cdot C)$;   ▷ reject the null when $p < \alpha$
  end function

The chi-square test is well behaved from the following theorem (which follows directly from Theorem 5 in Section 3):

---

[1]https://github.com/neurodata/mgc-matlab
[2]https://github.com/neurodata/mgc

Theorem 1. *The distance correlation chi-square test that rejects independence if and only if*

$$nDcor_n(\mathbf{X}, \mathbf{Y}) \geq F^{-1}_{\chi^2_1 - 1}(1 - \alpha)$$

is a valid and universally consistent test for sufficiently large n and sufficiently small type 1 error level α.

In practice, $n \geq 20$ suffices, and the validity empirically holds for any $\alpha \leq 0.05$.

2.2   Fast Statistic Computation

In the special case of $p = q = 1$ and Euclidean distance, Algorithm 1 can run significantly faster:

Theorem 2. *Suppose $p = q = 1$ and we use the Euclidean distance for the bias-corrected distance correlation. Then Algorithm 1 can be implemented with a computational complexity of $\mathcal{O}(n \log n)$.*

Essentially the bias-corrected sample distance correlation can be computed in $\mathcal{O}(n \log n)$ based on the results from Huo and Szekely (2016) and Chaudhuri and Hu (2019). This makes one-dimensional testing comparable in speed to the Pearson correlation t-test. On a standard Windows 10 machine using MATLAB, we are able to test independence between a million pairs of observations ($p = q = 1$) within 10 seconds with space requirement of $\mathcal{O}(n)$. Previously, the statistic computation and the permutation test need a space complexity of $\mathcal{O}(n^2)$ and a time complexity of $\mathcal{O}(rn^2)$, which would have required external disk storage and days of computation to finish testing the same amount of data. The implementation details can be found in Appendix and Github Matlab code.

2.3   Chi-Square Test for K-Sample

The chi-square test is readily applicable to any inference task using distance correlation, or any statistic involving a similar trace operation with bias-corrected matrix modification. Two immediate extensions are the K-sample test and partial test. Previously, the permutation test has been the standard approach for both of them.

Given K sets of sample data $\mathbf{U}^k = [u_1^k | u_2^k | \ldots | u_{n_k}^k] \in \mathbb{R}^{p \times n_k}$ for $k = 1, \ldots, K$, denote $\sum_{k=1}^K n_k = n$. The K-sample testing problem assumes u_i^k are independently and identically distributed as F_{U^k} for each i and k, and aim to test

$$H_0 : F_{U^1} = F_{U^2} = \cdots = F_{U^K},$$

$$H_A : \text{ there exists at least one } F_{U^k} \text{ that is different from other distributions.}$$

Algorithm 2 shows the K-sample variant of chi-square test.

Algorithm 2 The Distance Correlation Chi-Square Test for K-sample

---

Input: Sample data $\{\mathbf{U}^k \in \mathbb{R}^{p \times n_k} \text{ for } k = 1, \ldots, K\}$.
Output: The distance correlation C and its p-value p.
  function FASTKSAMPLE($\{\mathbf{U}^k\}$)
  (1) Data Transformation:
    $\mathbf{X} = [\ \mathbf{U}^1 \mid \mathbf{U}^2 \mid \cdots \mid \mathbf{U}^K\ ]$; $\triangleright$ concatenate Data
    $\mathbf{Y} = zeros(K, N)$; $\triangleright$ a zero matrix
    for $k = 1, \ldots, K$ **do**
      for $i = 1, \ldots, n_k$ **do**
        $\mathbf{Y}(k, \sum_{j=1}^{k-1} n_j + i) = 1$;
      end for
    end for
  (2) Test:
    $C = \text{Dcor}_n(\mathbf{X}, \mathbf{Y})$;
    $p = 1 - F_{\chi_1^2 - 1}(n \cdot C)$; $\triangleright$ reject the null when $p < \alpha$
  end function

It is shown in Panda et al. (2021) that by concatenating the sample data into $\mathbf{X}$ and forming an indicator matrix $\mathbf{Y}$ via one-hot encoding, $\text{Dcor}_n(\mathbf{X}, \mathbf{Y}) \to 0$ if and only if the null hypothesis is true. Then the next corollary follows from Theorem 1.

Corollary 1. *For sufficiently large n and sufficiently small type 1 error level α and any $K \geq 2$, Algorithm 2 is valid and universally consistent for testing $F_{U^1} = F_{U^2} = \cdots = F_{U^K}$.*

2.4 Chi-Square Test for Partial

Another application is to test whether the partial distance correlation equals 0 or not. Given three sample data $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$ of same sample size n, the partial distance correlation $\text{PDcor}_n(\mathbf{X}, \mathbf{Y}; \mathbf{Z})$

and its population version $\mathtt{PDcor}(X, Y; Z)$ are defined in Szekely and Rizzo (2014). Algorithm 3 shows the chi-square test for fast partial testing,

---

Algorithm 3 The Distance Correlation Chi-Square Test for Partial

---

Input: Paired sample data $(\mathbf{X}, \mathbf{Y}, \mathbf{Z}) = \{(x_i, y_i, z_i) \in \mathbb{R}^{p+q+s} \text{ for } i \in [n]\}$.
Output: The partial distance correlation C and its p-value p.
  function FASTPARTIAL($\mathbf{X}, \mathbf{Y}, \mathbf{Z}$)
    $C = \mathtt{PDcor}_n(\mathbf{X}, \mathbf{Y}; \mathbf{Z})$;   ▷ implementation details in Szekely and Rizzo (2014) or Github Matlab code
    $p = 1 - F_{\chi_1^2 - 1}(n \cdot C)$;   ▷ reject the null when $p < \alpha$
  end function

The partial statistic shares the same trace formulation as bias-corrected distance correlation and operates on the same bias-corrected matrix modification, except using projected distance matrices rather than Euclidean distance matrices. Therefore we have the following result:

Corollary 2. *For sufficiently large n and sufficiently small type 1 error level α, Algorithm 3 is valid and consistent for testing $\mathtt{PDcor}(X, Y; Z) = 0$.*

3   Supporting Theory

In this section we show the theory behind the chi-square test. Note that the results hold for any strong negative type metric or any characteristic kernel for X and Y, e.g., Euclidean distance, Gaussian kernel, Laplacian kernel, etc., which means the chi-square test is also applicable to the Hilbert-Schmidt Independence Criterion. Some results (like Theorem 3 and Theorem 5) are stated in limit or for sufficiently large sample size, for which $n \geq 20$ generally suffices for the limiting null distribution to be almost the same as the actual null.

3.1   The Limiting Null Distribution and the Centered Chi-Square Distribution

Theorem 3. *The limiting null distribution of the bias-corrected distance correlation satisfies*

$$n\mathtt{Dcor}_n(\mathbf{X}, \mathbf{Y}) \xrightarrow{D} \sum_{i,j=1}^{\infty} w_{ij}(\mathcal{N}_{ij}^2 - 1),$$

7

where the weights satisfy $w_{ij} \in [0,1]$ and $\sum_{i,j=1}^{\infty} w_{ij}^2 = 1$, and $\mathcal{N}_{ij}$ are independent standard normal distribution.

This theorem follows from Zhang et al. (2018). For different choice of metric and different marginal distributions, the weights $\{w_{ij}\}$ and the limiting null distribution are different, which is the main obstacle to design a test using known distribution. Nevertheless, from Theorem 3 we observe the mean and variance of $n\texttt{Dcor}_n(\mathbf{X}, \mathbf{Y})$ are always fixed and equal the mean and variance of $\chi_1^2 - 1$, which we shall call the centered chi-square distribution and denote by U from now on.

3.2   Upper Tail Dominance and Distribution Bounds

We aim to show that $U \sim \chi_1^2 - 1$ dominates the limiting null distribution at some upper tail probability α. We denote $F_V(x)$ as the cumulative distribution function of random variable V at argument x, $F_V^{-1}(1 - \alpha)$ as the inverse cumulative distribution function of random variable V at probability $1 - \alpha$, and formally define upper tail stochastic dominance as follows:

Definition 1. *Given two random variables U and V, we say U dominates V in upper tail at probability level α if and only if*

$$F_V(x) \geq F_U(x)$$

for all $x \geq F_U^{-1}(1 - \alpha)$. This is denoted by

$$V \preceq_\alpha U.$$

The next theorem plays a key role in proving the remaining theorems:

Theorem 4. *Assume $U, U_1, U_2, \ldots, U_m$ are independently and identically distributed as $\chi_1^2 - 1$, and the weights $\{w_i \in [0,1], i = 1, \ldots, m\}$ are decreasingly ordered and satisfies $\sum_{i=1}^{m} w_i^2 = 1$. The summation density satisfies*

$$f_{\sum_{i=1}^{m} w_i U_i}(x) = O(e^{-xc/2})$$

where the constant $c = \frac{1}{w_1} \in [1, \sqrt{m}]$.

Namely, the summation density is determined by the largest weight, which equals U if and only if $w_1 = 1$. We can then bound the upper tail distribution of bias-corrected distance correlation as follows:

Theorem 5. *For sufficiently large n, there exists $\alpha > 0$ such that*

$$\mathcal{N}(0, 2) \preceq_\alpha nDcor_n(\mathbf{X}, \mathbf{Y}) \preceq_\alpha U$$

regardless of the metric choice or marginal distributions.

Therefore, despite the null distribution is subject to change without a fixed nor known density form, the centered chi-square distribution is always a valid approximation choice. Theorem 1 in Section 2 (the validity and consistency of the chi-square test) follows trivially from Theorem 5.

3.3   Validity Level and Testing Power

The limiting null distribution have closed-form densities in some special cases, for which the largest validity level α (i.e., the largest type 1 error level so the chi-square test is still valid) can be exactly determined, e.g., 0.05 in Theorem 6 and at most 0.0875 in Theorem 8.

Theorem 6. *Given $m \geq 1$, assume the weights in Theorem 3 satisfy $w_i = \frac{1}{\sqrt{m}}$ for all $i = 1, \ldots, m$ and zero otherwise. It follows that*

$$nDcor_n(\mathbf{X}, \mathbf{Y}) \xrightarrow{D} \frac{\chi_m^2 - m}{\sqrt{m}} \preceq_{0.05} U,$$

where χ_m^2 is the chi-square distribution of degree m.

Corollary 3. *Under the same assumption of Theorem 4. Let $m_1 = \lfloor 1/w_1^2 \rfloor$ and $m_2 = \lceil 1/w_1^2 \rceil$, it always holds that*

$$\frac{\chi_{m_2}^2 - m_2}{\sqrt{m_2}} \preceq_\alpha \sum_{i=1}^{m} w_i U_i \preceq_\alpha \frac{\chi_{m_1}^2 - m_1}{\sqrt{m_1}} \preceq_\alpha U.$$

9

Next, we offer two random variable examples that achieve the known densities. Theorem 7 shows that the limiting null distribution can equal the centered chi-square distribution. Theorem 8 establishes the normal distribution (which is also used in the t-test in Szekely and Rizzo (2013)) with a simplified condition and a simplified proof, i.e., only one random variable is required to have infinite dimension.

Theorem 7. *When X and Y are two independent binary random variables, $n\mathbf{Dcor}_n(\mathbf{X}, \mathbf{Y}) \xrightarrow{D} U$.*

Theorem 8. *Assume X is independent of Y, X is continuous, and each dimension of X is exchangeable with positive finite variance. As $n, p \to \infty$, it holds that*

$$n\mathbf{Dcor}_n(\mathbf{X}, \mathbf{Y}) \xrightarrow{D} \mathcal{N}(0, 2) \preceq_{0.0875} U.$$

As the distance correlation t-test essentially uses a t-transformation of $\mathcal{N}(0, 2)$, we can view the t-test and the chi-square test as two-sides of the permutation test. In particular, the distance correlation t-test is an invalid test that slightly inflates the type 1 error level, while the chi-square test is valid but always conservative in power:

Corollary 4. *At any type 1 error level $\alpha \leq 0.05$, denote the testing power of distance correlation chi-square test, distance correlation t-test, and the permutation test as $\beta_\alpha^\chi, \beta_\alpha^t, \beta_\alpha$ respectively. At any n and α such that Theorem 5 holds, there exists $\alpha_1 \in (0, \alpha]$ and $\alpha_2 \in (\alpha, 0.0875]$ such that*

$$\beta_\alpha^\chi = \beta_{\alpha_1} \leq \beta_\alpha < \beta_{\alpha_2} = \beta_\alpha^t.$$

The actual α_1 and α_2 depend on the metric choice and marginal distributions.

Finally, there is no other valid test of distance correlation that is as fast and as powerful as the chi-square test:

Corollary 5. *At any n and α such that Theorem 5 holds, the chi-square test is the most powerful test among all valid tests of distance correlation using known distributions. Namely, for any valid test z of distance correlation using a fixed distribution, it always holds that $\beta_\alpha^\chi \geq \beta_\alpha^z$.*

4 Numerical Experiments

This section evaluates the numerical advantages of the chi-square test. First, we show that the centered chi-square distribution approximates the true null distribution of bias-corrected distance correlation in simulation. Second, the chi-square test is compared to existing tests on 1D data. Then the comparison is carried out on multivariate and high-dimensional data. And finally, we use a real data application to demonstrate the practical usage and superior performance of the chi-square test. Simulation function details are in Appendix.

4.1 Null Distribution Approximation

The top row of Figure 1 visualizes the centered chi-square distribution U and the normal distribution $N(0, 2)$ (divide each by n), and compare them with the actual null distribution of $\texttt{Dcor}_n(\mathbf{X}, \mathbf{Y})$ in varying dimensions. The centered chi-square distribution and the normal distribution are plotted by a solid line and an dashed line, respectively. We set sample size at $n = 100$, generate independent $\mathbf{X}$ and $\mathbf{Y}$ for $r = 10{,}000$ replicates and different p, q, and plot the null distribution in dotted line. The left panel shows the null distributions at $p = q = 1$, the center panel is for $p = q = 10$, and the right panel is for $p = q = 100$. As expected from the theorems, the upper tail of the null distribution for $\alpha \leq 0.05$ (equivalently y-axis greater than 0.95) always lies between and gradually shifts from U to $N(0, 2)$ as dimension increases.

The bottom row of Figure 1 shows the weights used in the corresponding limiting null distribution, associated with the ordered eigenvalue list $\{\lambda_i, i = 1, \ldots, 10\}$ normalized by $-(\sum_{i=1}^{n} \lambda_i^2)^{0.5}$ (see proof of Theorem 3). When $p = 1$, the leading weight plays a dominating role such that $n\texttt{Dcor}_n(\mathbf{X}, \mathbf{Y}) \overset{D}{\approx} U$. As p, q increase, all weights becoming similar such that $n\texttt{Dcor}_n(\mathbf{X}, \mathbf{Y}) \overset{D}{\to} N(0, 2)$. Therefore, the most conservative the chi-square test can be is to approximate a normal distribution by a chi-square distribution of same mean and variance.

4.2 One-Dimensional Random Variables

Here we evaluate distance correlation chi-square test, permutation test, distance correlation t-test, HSIC Gamma test (only applicable to biased HSIC), and the subsampling method for four

11
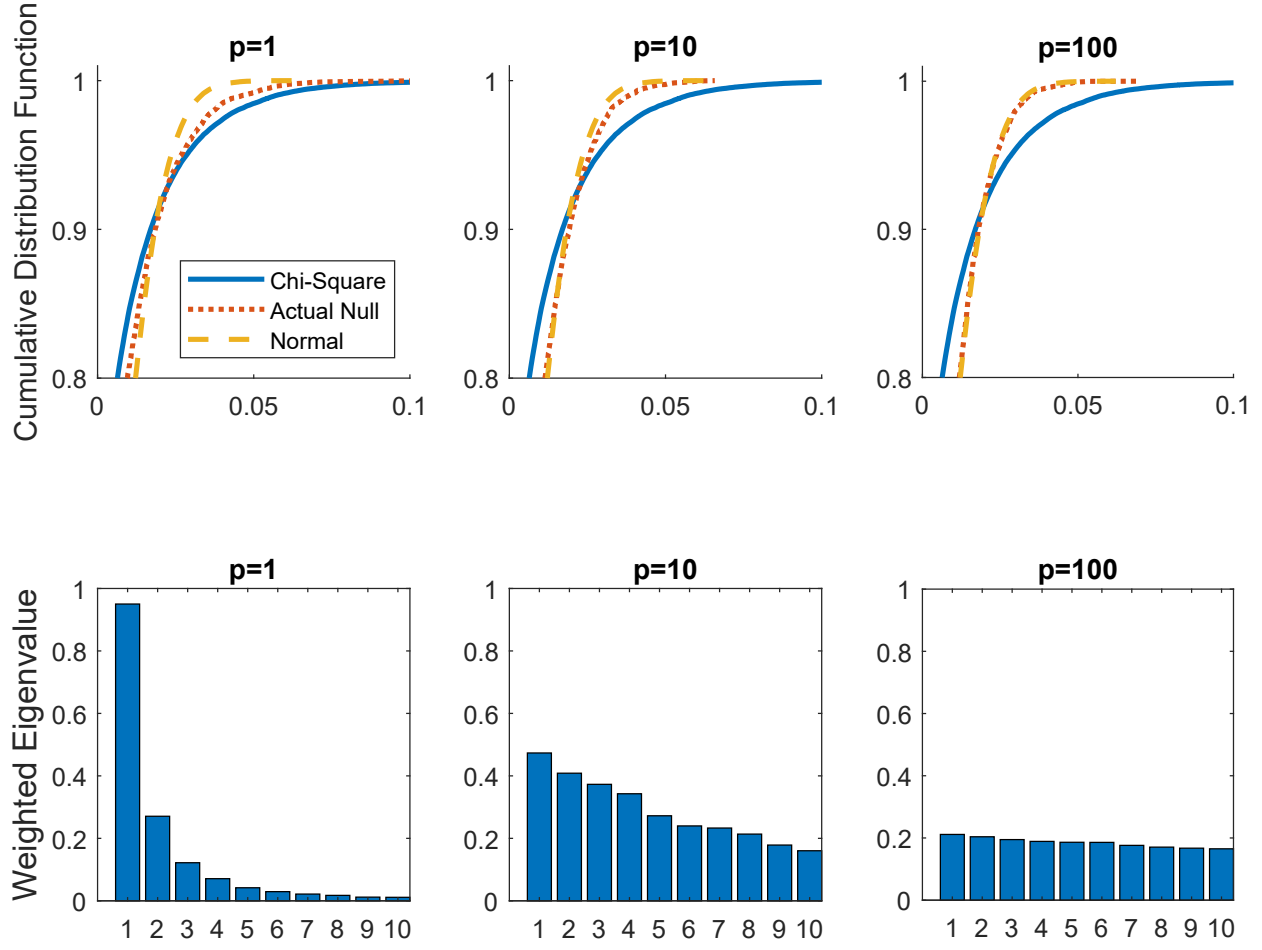
Figure 1: The top row compares the centered chi-square distribution, the normal distribution, and the actual null distribution of distance correlation in case of varying dimensions. The bottom row shows the weights used in the limiting null distribution in each case.
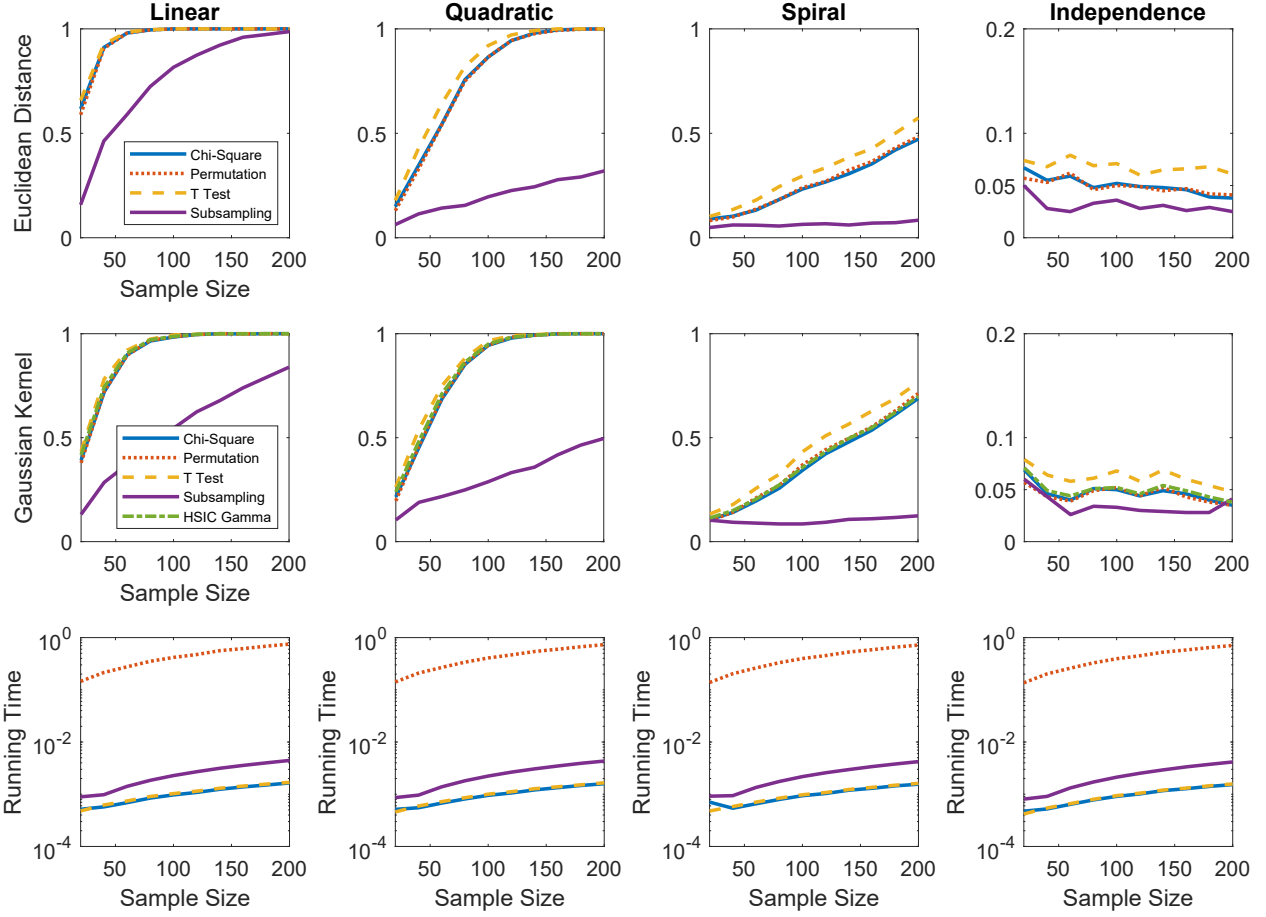
different 1D simulations: linear, quadratic, spiral, and independent. We extensively evaluated the tests on many different dependency types and decided to illustrate four representative simulations only, as the phenomenon is qualitatively similar throughout. All simulations are one-dimensional in this section, that is, $p = q = 1$. In each simulation, we sample $n = 20, 40, \ldots, 200$ points, generate sample data 1,000 times, run each test and reject at $\alpha = 0.05$ level, and compute how often the test is correctly rejected (the testing power), which is then plotted against the sample size.

The top row of Figure 2 shows the power of distance correlation under Euclidean distance, the middle row shows the power of Hilbert-Schmidt independence criterion (equivalently the distance correlation using Gaussian kernel), and the bottom row shows the running time for the top row in log scale. The performance is the same throughout all dependency types and sample size and metric choice: the permutation test is the benchmark for testing power, but significantly slower than the fast alternatives; the distance correlation chi-square test has almost the same testing power as the benchmark permutation test, and the fastest method; the distance correlation t-test is equally fast, but consistently inflates the power in most cases, e.g., it has a power of 0.07 for independence vs about 0.05 of the permutation test; the subsampling method always yields degraded power.

Note that the HSIC Gamma test requires some parameter estimation using mean, variance, and kernel bandwidth, thus only applicable to the middle row for the biased HSIC. Its testing power is almost the same as the chi-square and permutation tests. This does not mean the chi-square test approximates the Gamma test in distribution. Rather, it implies that the Gamma test approximates the null distribution of biased HSIC well; the chi-square test approximates the bias-corrected null distribution well; then the permutation test has almost the same testing power with or without bias-correction.

4.3   Increasing-Dimensional Random Variables

We consider four multivariate settings: equal variance (each dimension is exchangeable with same variance), minimal variance (first few dimensions has same variance while remaining dimensions have very small variance), dependent coordinates (consecutive dimensions are dependent),

Figure 2: Evaluate distance correlation using different tests for linear, quadratic, spiral, and independent simulations. The top row shows the power using the Euclidean distance, the center row shows the power using the Gaussian kernel, and the bottom row shows the running time (in log scale) for each method in the top row.
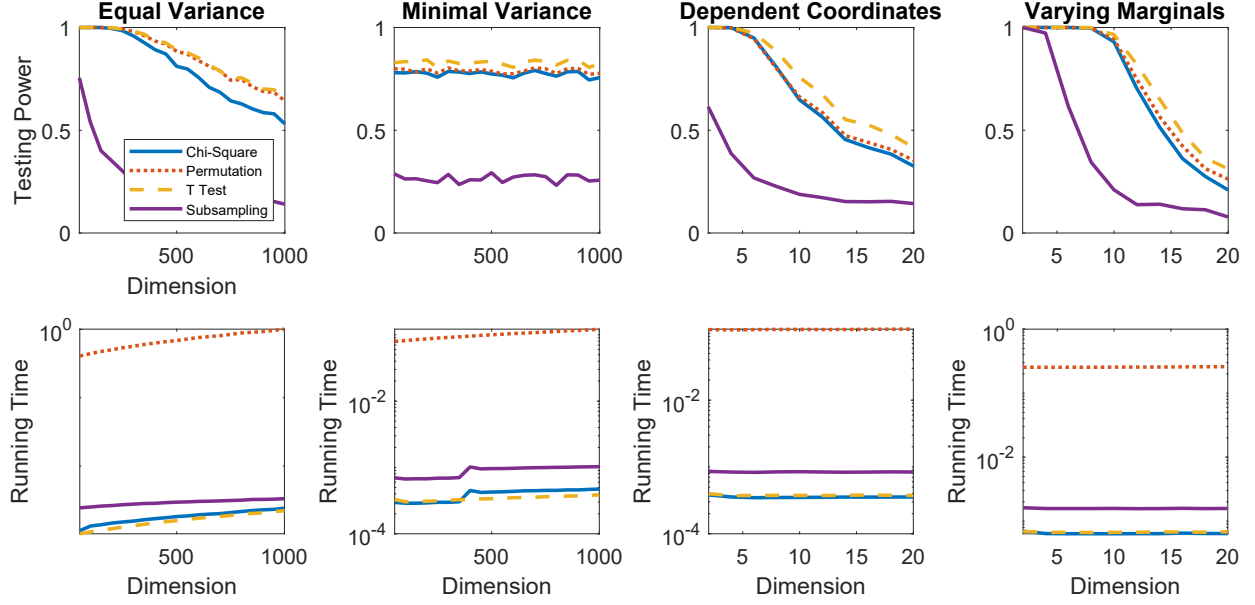
Figure 3: Evaluate distance correlation using different tests for four increasing-dimensional simulations using Euclidean distance. The first row shows the testing power in each simulation, and the second row shows the running time in log scale for each method in the respective first row.

and varying marginals (the marginal distribution of each dimension is different). We fix the sample size and $q = 1$, increase p accordingly in each simulation, and compute the testing power at $\alpha = 0.05$ based on 1000 Monte-Carlo replicates.

The testing power and the running time are plotted against dimension in Figure 3, offering almost the same interpretation as Figure 2. In particular, the equal variance simulation is the only setting here satisfying the assumption of Theorem 8, in which case the t-test only minimally inflate the testing power and the chi-square test exhibits a slightly more conservative testing power vs the permutation test. In the other three high-dimensional settings, the dimensions are no longer exchangeable, and the chi-square test has almost the same power as the permutation test. In terms of running time, the chi-square test and the t-test are the best, which do not increase much as dimension increases.

Finally, we evaluate the testing power for the multivariate simulation in Figure 1, and present the results in Figure 4. Regardless of p, q, n, the chi-square test is always similar in power as the permutation test and slightly conservative.
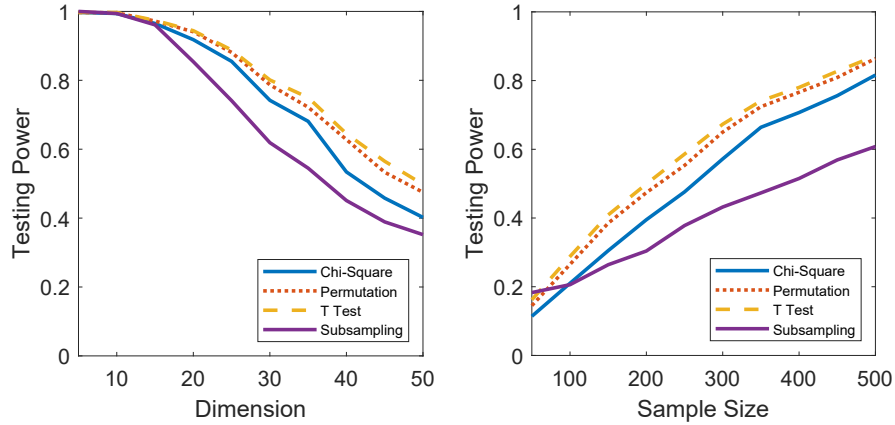
Figure 4: The Testing Power for the simulation in Figure 1. The left panel fix $n = 200$, and let p, q increases; the right panel set $p = q = 50$, and let n increases.

4.4 Real Data Application

Here we apply the tests to do feature selection on a proteomics data (Vogelstein et al., 2019; Wang et al., 2011). The data consist of 318 proteolytic peptides measurements, derived from the blood samples of 98 individuals harboring pancreatic ($n = 10$), ovarian ($n = 24$), colorectal cancer ($n = 28$), and healthy controls ($n = 33$). We would like to identify potential biomarkers for the pancreatic cancer, because it is lethal and no clinically useful biomarkers are currently available (Bhat et al., 2012).

For each biomarker, we first apply the distance correlation permutation test (using 500 replicates) between the blood measurement of each peptide and the class label (1 for pancreatic cancer, 0 otherwise). This yields 318 p-values, and we take all peptides with p-value less than 0.05 as the positive ones (in total 19 of them), and the remaining 299 peptides as the negative ones. Then we apply the chi-square test to each peptide vs the class label, and compute the true positive (i.e., when a peptide has p-value < 0.05 from the chi-square test, whether the peptide also has p-value < 0.05 from the permutation test) and true negative. Repeat for the t-test and subsampling method.

The result is summarized in Table 1. The chi-square test is fast and almost perfect for both true positive and true negative. The t-test is as fast, but inflates one of two errors. The subsampling is actually non-performing (all peptides have p-value > 0.1 so every peptide is a negative) — this

16

is because due to the imbalanced label vector, most subsampled labels are a constant 0, so the subsampled statistics end up being 0 most of the time.

Table 1: Report the running time (in seconds), true positive and true negative for each method. Permutation test is the benchmark for determining the positive and negative peptides. Among all the approaches, the chi-square test is the fastest with almost perfect result.

Method	Running Time	True Positive	True Negative
Permutation Test	149.8	100%	100%
Chi-Square Test	**0.32**	**92%**	**98%**
T Test	0.34	**92%**	68%
Subsampling	0.48	NaN	94%

5 Discussion

This paper proposes a new chi-square test for testing independence. It is very computationally efficient and scalable to big data, valid and universally consistent, applicable to general metric and kernel choices and other tasks like K-sample and partial, achieves similar power as the permutation test, and compares very favorably against all existing tests in simulations and real data. We expect the chi-square test to be widely used in practice due to its computational and performance advantages, and we plan to further investigate the stochastic dominance in theory and its potential applications in other inference tasks.

In particular, by Corollary 4 the chi-square test is always more conservative than the permutation test. But the power loss is negligible in 1D data and minimal in high-dimensional data as shown in Figure 2 and Figure 3. In comparison, the t-test always inflates the testing power. As shown in Figure 1, at the type 1 error level 0.05 the t-test inflates the error to $(0.05, 0.07]$, while the chi-square test has a conservative error within $[0.02, 0.05]$. Viewed in another way, the t-test at $\alpha = 0.05$ is equivalent in power to a permutation test at $\alpha \in (0.05, 0.07]$, while the chi-square test at $\alpha = 0.05$ is equivalent in power to a permutation test at $\alpha \in [0.02, 0.05]$. Inflating the type 1 error slightly may not be a big issue in a single test, but can cause larger deviation in multiple comparison, as evidenced tin the real data application; whereas the chi-square test is not susceptible to this issue.

Acknowledgment

References

Balasubramanian, K., Sriperumbudur, B., and Lebanon, G. (2013). Ultrahigh dimensional feature screening via rkhs embeddings. In *Proceedings of Machine Learning Research*, pages 126–134.

Bhat, K., Wang, F., Ma, Q., Li, Q., Mallik, S., chen Hsieh, T., and Wu, E. (2012). Advances in biomarker research for pancreatic cancer. *Current Pharmaceutical Design*, 18(17):2439–2451.

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

Fokianos, K. and Pitsillou, M. (2018). Testing independence for multivariate time series via the auto-distance correlation matrix. *Biometrika*, 105(2):337–352.

Fukumizu, K., Gretton, A., Sun, X., and Schölkopf, B. (2007). Kernel measures of conditional dependence. In *Advances in neural information processing systems*.

Gretton, A. and Gyorfi, L. (2010). Consistent nonparametric tests of independence. *Journal of Machine Learning Research*, 11:1391–1423.

Gretton, A., Herbrich, R., Smola, A., Bousquet, O., and Scholkopf, B. (2005). Kernel methods for measuring independence. *Journal of Machine Learning Research*, 6:2075–2129.

Heller, R., Heller, Y., and Gorfine, M. (2013). A consistent multivariate test of association based on ranks of distances. *Biometrika*, 100(2):503–510.

Heller, R., Heller, Y., Kaufman, S., Brill, B., and Gorfine, M. (2016). Consistent distribution-free k-sample and independence tests for univariate random variables. *Journal of Machine Learning Research*, 17(29):1–54.

Huo, X. and Szekely, G. (2016). Fast computing for distance covariance. *Technometrics*, 58(4):435–447.

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

Lyons, R. (2013). Distance covariance in metric spaces. *Annals of Probability*, 41(5):3284–3305.

Mehta, R., Shen, C., Ting, X., and Vogelstein, J. T. (2019). Consistent and powerful independence testing for time series. *https://arxiv.org/abs/1908.06486*.

Panda, S., Shen, C., Priebe, C. E., and Vogelstein, J. T. (2021). Multivariate multisample multiway nonparametric manova. *https://arxiv.org/abs/1910.08883*.

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

Rizzo, M. and Szekely, G. (2010). DISCO analysis: A nonparametric extension of analysis of variance. *Annals of Applied Statistics*, 4(2):1034–1055.

Rizzo, M. and Szekely, G. (2016). Energy distance. *Wiley Interdisciplinary Reviews: Computational Statistics*, 8(1):27–38.

Sejdinovic, D., Sriperumbudur, B., Gretton, A., and Fukumizu, K. (2013). Equivalence of distance-based and rkhs-based statistics in hypothesis testing. *Annals of Statistics*, 41(5):2263–2291.

Shen, C., Priebe, C. E., and Vogelstein, J. T. (2020). From distance correlation to multiscale graph correlation. *Journal of the American Statistical Association*, 115(529):280–291.

Shen, C. and Vogelstein, J. T. (2021). The exact equivalence of distance and kernel methods in hypothesis testing. *AStA Advances in Statistical Analysis*.

Szekely, G. and Rizzo, M. (2005). Hierarchical clustering via joint between-within distances: Extending ward's minimum variance method. *Journal of Classification*, 22:151–183.

Szekely, G. and Rizzo, M. (2009). Brownian distance covariance. *Annals of Applied Statistics*, 3(4):1233–1303.

Szekely, G. and Rizzo, M. (2013). The distance correlation t-test of independence in high dimension. *Journal of Multivariate Analysis*, 117:193–213.

Szekely, G. and Rizzo, M. (2014). Partial distance correlation with methods for dissimilarities. *Annals of Statistics*, 42(6):2382–2412.

Szekely, G., Rizzo, M., and Bakirov, N. (2007). Measuring and testing independence by correlation of distances. *Annals of Statistics*, 35(6):2769–2794.

Vogelstein, J. T., Wang, Q., Bridgeford, E., Priebe, C. E., Maggioni, M., and Shen, C. (2019). Discovering and deciphering relationships across disparate data modalities. *eLife*, 8:e41690.

Wang, Q., Chaerkady, R., Wu, J., Hwang, H. J., Papadopoulos, N., Kopelovich, L., Maitra, A., Matthaei, H., Eshleman, J. R., Hruban, R. H., Kinzler, K. W., Pandey, A., and Vogelstein, B. (2011). Mutant proteins as cancer-specific biomarkers. *Proceedings of the National Academy of Sciences of the United States of America*, (6):2444–9.

Wang, S., Shen, C., Badea, A., Priebe, C. E., and Vogelstein, J. T. (2019). Signal subgraph estimation via iterative vertex screening. *https://arxiv.org/abs/1801.07683*.

Wang, X., Pan, W., Hu, W., Tian, Y., and Zhang, H. (2015). Conditional Distance Correlation. *Journal of the American Statistical Association*, 110(512):1726–1734.

Xiong, J., Arroyo, J., Shen, C., and Vogelstein, J. T. (2020). Graph independence testing: Applications in multi-connectomics. *https://arxiv.org/abs/1906.03661*.

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

Zhong, W. and Zhu, L. (2015). An iterative approach to distance correlation-based sure independence screening. *Journal of Statistical Computation and Simulation*, 85(11):2331–2345.

Zhou, Z. (2012). Measuring nonlinear dependence in time-series, a distance correlation approach. *Journal of Time Series Analysis*, 33(3):438–457.

APPENDIX

A   Background

This section contains necessary background for the proofs.

A.1   Biased and Bias-corrected Sample Distance Correlation

Let the paired sample data, which is assumed independently and identically distributed as F_{XY}, be denoted by

$$(\mathbf{X}, \mathbf{Y}) = \{(x_i, y_i) \in \mathbb{R}^{p+q}, i = 1, \ldots, n\}.$$

Given a distance metric $d(\cdot, \cdot)$ such as the Euclidean metric, let $\mathbf{D^X}$ denote the $n \times n$ distance matrix of $\mathbf{X}$ with $\mathbf{D}_{ij}^{\mathbf{X}} = d(x_i, x_j)$, $\mathbf{D^Y}$ denote the distance matrix of $\mathbf{Y}$, and $\mathbf{H} = \mathbf{I} - \frac{1}{n}\mathbf{J}$ denote the $n \times n$ centering matrix where $\mathbf{I}$ is the identity matrix and $\mathbf{J}$ is the matrix of ones. The biased sample distance correlation was proposed in Szekely et al. (2007) with an elegant matrix formulation:

$$\text{Dcov}_n^b(\mathbf{X}, \mathbf{Y}) = \frac{1}{n^2}\text{trace}(\mathbf{H}\mathbf{D^X}\mathbf{H}\mathbf{H}\mathbf{D^Y}\mathbf{H}),$$

$$\text{Dcor}_n^b(\mathbf{X}, \mathbf{Y}) = \frac{\text{Dcov}_n^b(\mathbf{X}, \mathbf{Y})}{\sqrt{\text{Dcov}_n^b(\mathbf{X}, \mathbf{X})\text{Dcov}_n^b(\mathbf{Y}, \mathbf{Y})}} \in [0, 1],$$

where Dcov_n^b denotes the biased sample distance covariance and Dcor_n^b denotes the biased sample distance correlation. The bias-corrected version was later introduced via the following bias correction (Szekely and Rizzo, 2014): compute a modified matrix $\mathbf{C^X}$ as

$$\mathbf{C}_{ij}^{\mathbf{X}} = \begin{cases} \mathbf{D}_{ij}^{\mathbf{X}} - \frac{1}{n-2}\sum_{t=1}^{n}\mathbf{D}_{it}^{\mathbf{X}} - \frac{1}{n-2}\sum_{s=1}^{n}\mathbf{D}_{sj}^{\mathbf{X}} + \frac{1}{(n-1)(n-2)}\sum_{s,t=1}^{n}\mathbf{D}_{st}^{\mathbf{X}}, & i \neq j \\ 0, & \text{otherwise,} \end{cases}$$

and similarly compute $\mathbf{C^Y}$ from $\mathbf{D^Y}$. The bias-corrected sample distance covariance and correlation are

$$\text{Dcov}_n(\mathbf{X}, \mathbf{Y}) = \frac{1}{n(n-3)} \text{trace}\left(\mathbf{C^X C^Y}\right),$$

$$\text{Dcor}_n(\mathbf{X}, \mathbf{Y}) = \frac{\text{Dcov}_n(\mathbf{X}, \mathbf{Y})}{\sqrt{\text{Dcov}_n(\mathbf{X}, \mathbf{X})\text{Dcov}_n(\mathbf{Y}, \mathbf{Y})}} \in [-1, 1].$$

Namely, $\mathbf{C^X}$ always sets the diagonals to 0 and slightly modifies the off-diagonal entries from $\mathbf{HD^X H}$. If $n < 4$ or the denominator term is not a positive real number, the bias-corrected sample distance correlation is set to 0.

As long as the metric $d(\cdot, \cdot)$ is of strong negative type such as the Euclidean metric (Lyons, 2013), distance correlation satisfies the following:

$$\text{Dcor}_n(\mathbf{X}, \mathbf{Y}) \overset{n \to \infty}{\Rightarrow} 0 \text{ if and only if independence,}$$

which guarantees a universally consistent statistic for testing independence. Moreover, it is unbiased in the following sense:

$$\mathbb{E}(\text{Dcor}_n(\mathbf{X}, \mathbf{Y})) = 0 \text{ when } X \text{ is independent of } Y,$$

which is not satisfied by the biased version.

Instead of using a strong negative type distance metric, one may use a characteristic kernel for $d(\cdot, \cdot)$, i.e., $\mathbf{D^X}$ and $\mathbf{D^Y}$ become two kernel matrices, and all above formulation still hold. When one uses the Gaussian kernel with median distance as the bandwidth, the resulting correlation becomes the Hilbert-Schmidt independence criterion. As the theorems hold for any strong negative type metric or any characteristic kernel, we shall consistently use the distance correlation naming regardless of whether a metric or kernel is used for $d(\cdot, \cdot)$.

A.2   Null Distribution of Distance Correlation

The goal for fast testing is to approximate the null distribution of distance correlation via a known distribution. Then given any sample test statistic, one can immediately compute the

p-value, and reject the independence hypothesis when it is smaller than a pre-specified critical level α.

From Zhang et al. (2018), the limiting null distribution of unbiased distance covariance satisfies

$$n\texttt{Dcov}_n(\mathbf{X}, \mathbf{Y}) \xrightarrow{D} \sum_{i,j=1}^{\infty} \lambda_i \mu_j (\mathcal{N}_{ij}^2 - 1), \tag{1}$$

where $\{\lambda_i\}$ are the limiting eigenvalues of $\mathbf{H}\mathbf{D}^X\mathbf{H}/n$, $\{\mu_j\}$ are the limiting eigenvalues of $\mathbf{H}\mathbf{D}^Y\mathbf{H}/n$, $\mathcal{N}_{ij}$ are identically and independently distributed standard normal random variables, and the summation index sums over $i = 1, \ldots, \infty$ and $j = 1, \ldots, \infty$. The limiting null distribution using sample eigenvalues is shown to almost equal the finite-sample null distribution for $n \geq 20$ (see Lyons (2013) and Zhang et al. (2018) for more details), so it suffices to use the limiting null. The eigenvalues $\{\lambda_i\}$ and $\{\mu_j\}$ can vary significantly for different metric or kernel choices $d(\cdot, \cdot)$ and different marginal distributions F_X and F_Y. Therefore, no fixed distribution can perfectly approximate the null all the time.

In a brute-force manner, the sample eigenvalues can be estimated via eigen-decomposition of the sample matrices, then the null distribution can be simulated by generating n^2 independent normal distributions. This method has the best testing power (i.e., almost the same as permutation test for $n \geq 20$) but requires $\mathcal{O}(n^3)$ time complexity thus too costly. Alternatively, one may compute subsampled correlations and take average, which follows a normal distribution via the central limit theorem. However, it is well-known and provable that a subsampled statistic yields an inferior testing power, because the estimated null distribution is a very conservative one with enlarged variance. These two approaches are summarized in Zhang et al. (2018).

The standard permutation test works as follows: for each replicate, permute the observations in $\mathbf{Y}$ (row indices of the matrix) by a random permutation π, denote the permuted sample data as $\mathbf{Y}^\pi$, and compute the permuted statistic $\texttt{Dcor}(\mathbf{X}, \mathbf{Y}^\pi)$. Repeat for r such random permutations, and compute a set of permuted statistics $\{\texttt{Dcor}(\mathbf{X}, \mathbf{Y}^\pi)\}$. Then the p-value is the fraction of times the observed test statistic is more extreme than the permuted test statistics. This is summarized in Algorithm A1. The random permutation effectively breaks dependencies within the sample data and well-approximates the actual null distribution. The permutation test is the default approach

in almost every independence testing methodology, and provably a valid and consistent test with any consistent dependence measure (Shen et al., 2020), not just distance correlation. Also note that distance correlation and distance covariance share the same p-value under permutation test, because the covariance to correlation transformation is invariant to permutation.

Algorithm A1 Permutation Test for Independence

Input: Paired sample data $(\mathbf{X}, \mathbf{Y}) = \{(x_i, y_i) \in \mathbb{R}^{p+q} \text{ for } i \in [n]\}$, and the number of random permutation r.

Output: The test statistic C and the p-value p.

  function PERMUTATIONTEST($\mathbf{X}, \mathbf{Y}, r$)

    $C = \text{STAT}(\mathbf{X}, \mathbf{Y})$;   ▷ can be any dependency measure not just distance correlation

    for $s = 1, \ldots, r$ **do**

      $\pi = \texttt{randperm}(n)$;   ▷ generate a random permute index

      $cp(s) = \text{STAT}(\mathbf{X}(\pi), \mathbf{Y})$;   ▷ cp stores the permuted statistics

    end for

    $p = \sum_{s=1}^{r} \boldsymbol{I}(cp(s) > C)/r$ ▷ the percentage the permuted statistics is larger

  end function

A popular test using a known distribution is the distance correlation t-test (Szekely and Rizzo, 2013), which approximates the null distribution by a normal distribution of mean 0 and variance 2. When X and Y are independent, assume each dimension of X and Y are independently and independently distributed (or exchangeable) with positive finite variance, it was proved that

$$\sqrt{n^2 - 3n - 2} \cdot \texttt{Dcor}_n(\mathbf{X}, \mathbf{Y}) \xrightarrow{D} \mathcal{N}(0, 2) \tag{2}$$

as $n, p, q \to \infty$. The t-distribution transformation and the corresponding t-test follow from the normal distribution. Therefore, the distance correlation t-test has a constant time complexity and enjoys a similar testing power as the permutation test under required condition. However, there has been no investigation on its testing performance out of the high-dimension assumption.

B Proofs

B.1 Proof of Theorem 1

Proof. From Theorem 5, U dominates $n\text{Dcor}(\mathbf{X}, \mathbf{Y})$ in upper tail which the actual null converges to. Therefore, there exists n' and α' such that the test correctly controls the type 1 error level for any $\alpha \le \alpha'$ and sample size $n \ge n'$. For example, when $\alpha' = 0.05$ from Theorem 6, the test is expected to be valid at any type 1 error level no more than 0.05 at sufficiently large n.

For consistency: at any $\alpha < 2\Phi(1) - 1$, $F_U^{-1}(1 - \alpha)$ is a positive and fixed constant. When X is dependent of Y, $\text{Dcor}_n(\mathbf{X}, \mathbf{Y})$ converges to a non-zero positive constant, such that $n\text{Dcor}_n(\mathbf{X}, \mathbf{Y}) \to +\infty > F_U^{-1}(1 - \alpha)$ and the test is always correctly rejected asymptotically.

Therefore, the distance correlation chi-square test is valid and universally consistent for testing independence. $\qquad\square$

B.2 Proof of Theorem 2

As the p-value computation in Algorithm 1 takes $\mathcal{O}(1)$, it suffices to show the bias-corrected distance correlation can be computed in $\mathcal{O}(n \log n)$ for one-dimensional data using Euclidean distance. Denote the distances and centered distances as

$$A_{ij} = \|x_i - x_j\|_2, \quad B_{ij} = \|y_i - y_j\|_2$$

$$A_{i\cdot} = \sum_{j=1}^{n} A_{ij}, \quad B_{i\cdot} = \sum_{j=1}^{n} B_{ij}$$

$$A_{\cdot\cdot} = \sum_{i=1}^{n} A_{i\cdot}, \quad B_{\cdot\cdot} = \sum_{i=1}^{n} B_{i\cdot}.$$

and define

$$T_1 = \sum_{i,j=1}^{n} A_{ij} B_{ij}, \quad T_2 = \sum_{i=1}^{n} A_{i\cdot} B_{i\cdot}, \quad T_3 = A_{\cdot\cdot} B_{\cdot\cdot}.$$

It was shown in Chaudhuri and Hu (2019) that T_1, T_2, T_3 can be computed in $\mathcal{O}(n \log n)$ for one-dimensional data using Euclidean metric. Therefore, it suffices to prove the following lemma:

Lemma 1. *The unbiased distance covariance can be expressed into*

$$Dcov_n(\mathbf{X}, \mathbf{Y}) = \frac{T_1}{n(n-3)} - \frac{2T_2}{n(n-2)(n-3)} + \frac{T_3}{n(n-1)(n-2)(n-3)}.$$

Then unbiased distance covariance and correlation can be computed in $\mathcal{O}(n \log n)$ for one-dimensional data using Euclidean distance.

Proof. The unbiased distance covariance can be decomposed into

$$
\begin{aligned}
\text{Dcov}_n(\mathbf{X}, \mathbf{Y}) = & \frac{1}{n(n-3)} \sum_{i \neq j}^{n} (A_{ij} - \frac{1}{n-2} A_{i\cdot} - \frac{1}{n-2} A_{j\cdot} + \frac{1}{(n-1)(n-2)} A_{\cdot\cdot}) \\
& \cdot (B_{ij} - \frac{1}{n-2} B_{i\cdot} - \frac{1}{n-2} B_{j\cdot} + \frac{1}{(n-1)(n-2)} B_{\cdot\cdot}) \\
= & \frac{T_1}{n(n-3)} - \frac{2T_2}{n(n-2)(n-3)} + \frac{T_3}{n(n-1)(n-2)(n-3)} \\
& - \frac{T_2}{n(n-2)(n-3)} + \frac{(n-1)T_2}{n(n-2)^2(n-3)} + \frac{T_3 - T_2 - T_3}{n(n-2)^2(n-3)} \\
& - \frac{T_2}{n(n-2)(n-3)} + \frac{(n-1)T_2}{n(n-2)^2(n-3)} + \frac{T_3 - T_2 - T_3}{n(n-2)^2(n-3)} \\
& + \frac{T_3}{n(n-1)(n-2)(n-3)} - \frac{2T_3}{n(n-2)^2(n-3)} + \frac{T_3}{(n-1)(n-2)^2(n-3)} \\
= & \frac{T_1}{n(n-3)} - \frac{2T_2}{n(n-2)(n-3)} + \frac{T_3}{n(n-1)(n-2)(n-3)}.
\end{aligned}
$$

To compute the bias-corrected distance correlation, one needs to compute $\text{Dcov}_n(\mathbf{X}, \mathbf{Y})$, $\text{Dcov}_n(\mathbf{X}, \mathbf{X})$, and $\text{Dcov}_n(\mathbf{Y}, \mathbf{Y})$, all of which now take $\mathcal{O}(n \log n)$. Therefore, the bias-corrected distance correlation can be computed in $\mathcal{O}(n \log n)$. □

B.3   Proof of Corollary 1 and Corollary 2

Proof. From Panda et al. (2021), a valid and consistent independence test is also valid and consistent for K-sample testing. Thus Corollary 1 follows immediately from Theorem 1.

From Szekely and Rizzo (2014), the partial distance correlation is also trace of the product of

two modified distance matrix. Therefore $\texttt{PDcor}(X, Y; Z)$ has the same limiting null distribution as in Equation 1. Thus Corollary 2 also follows from Theorem 1. □

B.4 Proof of Theorem 3

Proof. Recall from Equation 1 that the limiting null distribution of distance covariance satisfies

$$n\texttt{Dcov}_n(\mathbf{X}, \mathbf{Y}) \xrightarrow{D} \sum_{i,j=1}^{\infty} \lambda_i \mu_j (\mathcal{N}_{ij}^2 - 1),$$

where $\{\lambda_i\}$ are the limiting eigenvalues of $\mathbf{HD^X H}/n$ and $\{\mu_j\}$ are the limiting eigenvalues of $\mathbf{HD^Y H}/n$. Then the distance variance always satisfies

$$\lim_{n\to\infty} \left(\texttt{Dcov}_n(\mathbf{X}, \mathbf{X}) - \sum_{i=1}^{n} \lambda_i^2\right)$$
$$= \lim_{n\to\infty} \left(\texttt{Dcov}_n(\mathbf{X}, \mathbf{X}) - \texttt{Dcov}_n^b(\mathbf{X}, \mathbf{X})\right)$$
$$\to 0,$$

where the third line follows from the fact that both unbiased and biased statistics converge to a same constant, and the second line follows from

$$\texttt{Dcov}_n^b(\mathbf{X}, \mathbf{X}) = \frac{1}{n^2}\texttt{trace}\left(\mathbf{HD^X HHD^X H}\right) = \sum_{i=1}^{n} \lambda_i^2.$$

Therefore,

$$\texttt{Dcov}_n(\mathbf{X}, \mathbf{X}) \to \sum_{i=1}^{\infty} \lambda_i^2,$$
$$\texttt{Dcov}_n(\mathbf{Y}, \mathbf{Y}) \to \sum_{j=1}^{\infty} \mu_j^2,$$
$$n\texttt{Dcor}_n(\mathbf{X}, \mathbf{Y}) \xrightarrow{D} \sum_{i,j=1}^{\infty} w_{ij}(\mathcal{N}_{ij}^2 - 1).$$

where $w_{ij} = \frac{\lambda_i \mu_j}{\sqrt{\sum\limits_{i=1}^{\infty} \lambda_i^2 \sum\limits_{j=1}^{\infty} \mu_j^2}}$.

A strong negative type metric is always of negative type, and a characteristic kernel is always a positive definite kernel. When the distance metric is of negative type, the two matrices are negative definite and all eigenvalues are all non-positive. When positive definite kernels are used, then these eigenvalues are all non-negative. In either case, the product $\{\lambda_i \mu_j\}$ is always non-negative such that $w_{ij} \in [0, 1]$. Moreover, for any $n \geq 1$ it always holds that

$$\sum_{i,j=1}^{n} w_{ij}^2 = \frac{\sum\limits_{i,j=1}^{n} \lambda_i^2 \mu_j^2}{\sum\limits_{i=1}^{n} \lambda_i^2 \sum\limits_{j=1}^{n} \mu_j^2} = 1.$$

Note that in the special case that either X or Y is constant, all eigenvalues are 0 so the correlation equals 0 instead. This corresponds to a trivial independence case, and all dominance / validity / consistency results hold trivially. $\square$

B.5 Proof of Theorem 4

Proof. We prove this theorem by induction. Lemma 2 proves the initial step at $m = 2$ by analyzing the smaller weight $w_2 \in (0, \frac{1}{\sqrt{2}}]$, and Lemma 3 proves the induction step by considering the smallest weight $w_m \in (0, \frac{1}{\sqrt{m}}]$. As the weights are square summed to 1, the smallest positive weight must satisfy $w_m \in (0, \frac{1}{\sqrt{m}}]$, and the largest weight must satisfy $w_1 \in [\frac{1}{\sqrt{m}}, 1]$. It suffices to consider all positive weights, because it reduces to the $m - 1$ case when $w_m = 0$. Moreover, if $m = 1$, the summation density equals U which satisfies $f_U(x) = O(e^{-x/2})$; and if $w_i = \frac{1}{\sqrt{m}}$ for all i, it corresponds to Theorem 6. $\square$

Lemma 2. *Suppose both U and V are independently and identically distributed as the centered chi-square distribution, and $w \in (0, 1/\sqrt{2}]$. The density of $\sqrt{1 - w^2}U + wV$ decays exponentially at the rate $O(e^{-x/(2\sqrt{1-w^2})})$.*

Proof. The summation density equals

$$f_{wU+\sqrt{1-w^2}V}(x) = \int_{-w}^{+\infty} f_{\sqrt{1-w^2}V}(x-z)f_{wU}(z)dz$$

$$= \frac{b^2}{w\sqrt{1-w^2}} \int_{-w}^{x+\sqrt{1-w^2}} ((\frac{x-z}{\sqrt{1-w^2}}+1)(\frac{z}{w}+1))^{-0.5} e^{-(\frac{x-z}{\sqrt{1-w^2}}+\frac{z}{w}+2)/2} dz$$

$$= e^{-\frac{x}{2\sqrt{1-w^2}}} \cdot \frac{b^2 e^{-0.5}}{\sqrt{1-w^2}} \cdot \int_{-1}^{\frac{x+\sqrt{1-w^2}}{w}} ((\frac{x-wz}{\sqrt{1-w^2}}+1)(z+1))^{-0.5} e^{-((1-\frac{w}{\sqrt{1-w^2}})z+1)/2} dz$$

$$= e^{-\frac{x}{2\sqrt{1-w^2}}} \cdot g(x),$$

where the second to last equality involves a change of variable from z to $\frac{z}{w}$, and the last equality combines every other term into $g(x)$.

The leading exponential term dominates the decay rate of the density, while $g(x)$ is at most $O(x)$: the term $\frac{b^2 e^{-0.5}}{\sqrt{1-w^2}} \le \sqrt{2}b < 1$ is a fixed constant; the upper bound of the integral increases at $O(x)$; the polynomial term of the integral is $O(1)$; and the remaining exponential term in the integral satisfies

$$e^{-((1-\frac{w}{\sqrt{1-w^2}})z+1)/2} \le 1$$

for any fixed $w \in (0, \frac{1}{\sqrt{2}}]$. Therefore, the density of $wU + \sqrt{1-w^2}V$ decays at the rate $O(e^{-xc/2})$, for which $c = (1-w^2)^{-0.5} \in (1, \sqrt{2}]$. When we consider $m = 2$ and let $w_2 = w$ be the smaller weight, $w_1 = \sqrt{1-w^2}$ becomes the larger weight so $c = 1/w_1$. $\qquad \square$

Lemma 3. *Suppose U is the centered chi-square distribution, and V is an $m-1$ weighted summation of U_i using the weights $\{w_i(1-w_m^2)^{-0.5}, i = 1, \ldots, m-1\}$ for any $m > 2$ and $w_m \in (0, \frac{1}{\sqrt{m}}]$. Assume the density of V satisfies*

$$f_V(x) = O(e^{-xc_{m-1}/2})$$

where $c_{m-1} = \sqrt{1-w_m^2}/w_1 \in [1, \sqrt{m-1}]$. Then the density $f_{w_mU+\sqrt{1-w_m^2}V}(x)$ satisfies

$$f_{w_mU+\sqrt{1-w_m^2}V}(x) = O(e^{-xc_m/2})$$

9

where $c_m = 1/w_1 \in [1, \sqrt{m}]$.

Proof. The initial case corresponds to Lemma 2 with $c_1 = 1/w_1 \in (1, \sqrt{2}]$. Moreover, $\{w_i(1 - w_m^2)^{-0.5}, i = 1, \ldots, m-1\}$ is always a valid weighting scheme for $m-1$ summation because

$$\sum_{i=1}^{m-1} w_i^2/(1 - w_m^2) = 1.$$

The density of V must be of the form

$$f_V(x) = e^{-xc_{m-1}/2}g(x),$$

where $g(x)$ is a function that grows at most $o(e^{xc_{m-1}/2})$.

In the following, we let $w = w_m$ to simplify the expression. Then the summation density is

$$
\begin{aligned}
&f_{wU+\sqrt{1-w^2}V}(x)\\
&= \int_{-w}^{+\infty} f_{\sqrt{1-w^2}V}(x-z)f_{wU}(z)dz\\
&= \frac{b}{\sqrt{1-w^2}}\int_{-1}^{\frac{x+\sqrt{m(1-w^2)}}{w}} (z+1)^{-0.5}e^{-(\frac{c_{m-1}(x-wz)}{\sqrt{1-w^2}}+z+1)/2}g\left(\frac{x-wz}{\sqrt{1-w^2}}\right)dz\\
&= e^{-\frac{xc_{m-1}}{2\sqrt{1-w^2}}} \cdot \frac{b}{\sqrt{1-w^2}} \cdot \int_{-1}^{\frac{x+\sqrt{m(1-w^2)}}{w}} (z+1)^{-0.5}e^{-((1-\frac{wc_{m-1}}{\sqrt{1-w^2}})z+1)/2}g\left(\frac{x-wz}{\sqrt{1-w^2}}\right)dz.
\end{aligned}
$$

The only exponential term within the integral satisfies

$$e^{-((1-\frac{wc_{m-1}}{\sqrt{1-w^2}})z+1)/2} \leq 1.$$

This is because

$$\frac{wc_{m-1}}{\sqrt{1-w^2}} \leq \frac{c_{m-1}}{\sqrt{m-1}} \leq \frac{\sqrt{m}}{m-1} < 1$$

for any $m > 2$, so that $(1 - \frac{wc_{m-1}}{\sqrt{1-w^2}})z + 1 \geq 0$ when $z > -1$. Analyzing every other term in the same manner as the base case in the proof of Lemma 2, we conclude that the density is dominated

by the leading exponential term. Therefore, the density decays at $O(e^{-xc_m/2})$, where

$$c_m = \frac{c_{m-1}}{\sqrt{1-w^2}} = 1/w_1 \in (1, \sqrt{m}].$$

$\square$

B.6 Proof of Theorem 5

Proof. From Theorem 4, the decay rate of the summation density is $O(e^{-xc/2})$ with $c \geq 1$. It always decays faster than U such that for sufficiently large x,

$$f_{\sum\limits_{i=1}^{m} w_i U_i}(x) \leq f_U(x)$$

with equality if and only if $m = 1$, leading to upper tail dominance for sufficiently small α. Moreover, $\mathcal{N}(0,2)$ has the same mean and variance as them, with a density decay rate of $O(e^{-x^2/8})$ that is always faster than $O(e^{-xc/2})$. Therefore, there exists $\alpha > 0$ such that $\mathcal{N}(0,2) \preceq_\alpha n\text{Dcor}(\mathbf{X}, \mathbf{Y}) \preceq_\alpha U$. $\square$

B.7 Proof of Theorem 6

Proof. First, the density of U is

$$f_U(x) = b(x+1)^{-0.5}e^{-(x+1)/2}$$

where $b = 2^{-0.5}\Gamma(0.5)^{-1} \approx 0.4$ is the constant from standard chi-square distribution of degree 1. The domain of U is $(-1, +\infty)$, and the density equals 0 otherwise.

When $w_i = \frac{1}{\sqrt{m}}$,

$$W = \sum_{i=1}^{m} w_i U_i = \sum_{i=1}^{m} U_i/\sqrt{m} \sim \frac{\chi_m^2 - m}{\sqrt{m}},$$

11

whose density equals

$$\frac{\sqrt{m}}{2^{\frac{m}{2}}\Gamma(m/2)}(\sqrt{m}x + m)^{\frac{m}{2}-1}e^{-\frac{\sqrt{m}x+m}{2}}.$$

At fixed m, the density of W decays exponentially at the rate $O(e^{-\sqrt{m}x/2})$. This matches Theorem 4, and there must exist x such that $f_U(x') \geq f_W(x')$ for all $x' \geq x$.

As the distribution is known, we can exactly compute the argument x such that $F_W(x) \geq F_U(x)$, which is monotonically decreasing as m increases. In particular, $x = 2.7 < F_U^{-1}(0.95)$ when $m = 2$; $x = 2.5$ when $m = 3$; $x = 2.3$ when $m = 10$; and $x = 2$ when $m = 1000$. Therefore α is at least 0.05 regardless of m, and $W \preceq_{0.05} U$ always holds in the equal weight case. As $m \to \infty$, the validity level converges to $\alpha = 0.0875$ from the proof of Theorem 8. $\qquad\square$

B.8 Proof of Theorem 7

Proof. This theorem follows from Lemma 4: when X and Y are binary random variables, the centered sample matrix $\mathbf{HD^X H}$ only has 1 nonzero eigenvalue, so is $\mathbf{HD^Y H}$. As a result, the eigenvalue products in Equation 1 satisfy $\lambda_1\mu_1 > 0$ and $\lambda_i\mu_j = 0$ otherwise.

Once the eigenvalue products are normalized into w_{ij} (see proof of Theorem 3), it follows that $w_{11} = 1$ and $w_{ij} = 0$ otherwise, and $n\texttt{Dcor}_n(\mathbf{X}, \mathbf{Y}) \overset{D}{\to} U$.

Lemma 4. *Suppose the sample data $\mathbf{X}$ has at most m distinct values. Then the sample matrix $\mathbf{HD^X H}$ has at most $m - 1$ non-zero eigenvalues regardless of n.*

This lemma can be argued via the sample matrix as follows: First, as $\mathbf{H}$ is the centering matrix, the eigenvalues of $\mathbf{HD^X H}$ always equal the eigenvalues of $\mathbf{D^X H}$. Next, observe that $det(\mathbf{H}) = 0$ and thus $det(\mathbf{D^X H}) = 0$, so there exists at least one zero eigenvalue and at most $n - 1$ non-zero eigenvalues. The geometric multiplicity of the zero eigenvalue equals $n - rank(\mathbf{D^X H})$, so the total number of non-zero eigenvalues is $rank(\mathbf{D^X H}) = rank(\mathbf{D^X}) - 1$. When X has m distinct values, the distance matrix only has m distinct rows and $rank(\mathbf{D^X}) = m$. Therefore the centered matrix has $m - 1$ non-zero eigenvalues. $\qquad\square$

B.9   Proof of Theorem 8

Proof. As $p \to \infty$ and X is continuous, the Euclidean distance matrix $\mathbf{D^X}$ converges to $a(J - I)$, so the centered matrix converges to $a(-I) + aJ/n$, where a is a constant depending on the metric choice and marginal distribution F_X. Using Lemma 4, $\mathbf{HD^XH}/n$ has 1 zero eigenvalue and $n - 1$ non-zero eigenvalues that are asymptotically the same. Similarly for $\mathbf{HD^YH}/n$ when $q \to \infty$ and Y is continuous.

This is essentially the asymptotic case of Theorem 6 as $m \to \infty$. By Lyapunov central limit theorem:

$$\sum_{i=1}^{m} w_i U_i \overset{m \to \infty}{\to} \mathcal{N}(0, 2).$$

Evaluating the cumulative distribution of standard normal, it follows that $\mathcal{N}(0, 2) \preceq_\alpha U$ at $\alpha = 0.0875$. Therefore, when either p or q increases to infinity and n also increases to infinity, the limiting null distribution becomes a normal distribution and satisfies $\mathcal{N}(0, 2) \preceq_{0.0875} U$. □

B.10   Proof of Corollary 5

Proof. Given a test z with a fixed distribution Z, it being always valid for testing independence at level α requires

$$U \preceq_\alpha Z$$

because U is the limiting null distribution of distance correlation between binary random variables by Theorem 7. Thus,

$$n\mathtt{Dcor}_n(\mathbf{X}, \mathbf{Y}) \preceq_\alpha U \preceq_\alpha Z$$

when testing independence between arbitrary random variables, and

$$\beta_\alpha^z < \beta_\alpha^X \leq \beta_\alpha$$

13

always holds. Therefore, the chi-square test is the most powerful test among all valid tests of distance correlation using known distributions. □

C   Simulation Details

The simulations in Figure 1 and Figure 4 are generated via the exponential relationship:

$$X[d] \sim Uniform(-3, 1) \text{ for each } d = 1, \ldots, p$$
$$Y = \exp(X) + 0.2 * \epsilon,$$

where $q = p$ and ϵ is an independent standard Cauchy random variable.

The 1-dimensional sample data in Figure 2 are generated via the following:

- Linear (X, Y):

$$X \sim Uniform(-1, 1),$$
$$Y = X + \epsilon.$$

- Quadratic (X, Y):

$$X \sim Uniform(-1, 1),$$
$$Y = X^2 + 0.5\epsilon.$$

- Spiral (X, Y): let $Z \sim \mathcal{N}(0, 5)$, $\epsilon \sim \mathcal{N}(0, 1)$,

$$X = Z \cos(\pi Z),$$
$$Y = Z \sin(\pi Z) + 0.4\epsilon.$$

- Independent (X, Y): let $Z \sim \mathcal{N}(0, 1)$, $W \sim \mathcal{N}(0, 1)$, $Z' \sim Bernoulli(0.5)$, $W' \sim$

$Bernoulli(0.5)$,

$$X = Z/3 + 2Z' - 1,$$

$$Y = W/3 + 2W' - 1.$$

In all four simulations we have $p = q = 1$, and ϵ is an independent standard normal random variable.

The increasing-dimensional simulations in Figure 3 are generated via

- Equal variance:

$$X[d] \sim Uniform(-1, 1) \text{ for each } d = 1, \ldots, p$$

$$Y = X[1];$$

- Minimal variance:

$$X[d] \sim Uniform(-1, 1), \text{ for } d = 1, \ldots, 20,$$

$$X[d] \sim \frac{1}{p} \cdot Uniform(-1, 1), \text{ for } d = 21, \ldots, p,$$

$$Y = X[1];$$

- Dependent coordinate:

$$X[1] \sim Uniform(-1, 1),$$

$$X[d] = 0.5X[d-1] + Uniform(-0.5, 0.5), \text{ for } d = 2, \ldots, p,$$

$$Y = \sum_{d=1}^{p} X^2[d];$$

- Varying marginal:

$$X[d] \sim \mathcal{N}(d, d), \text{ for } d = 1, \ldots, p,$$

$$Y = X[1].$$

In all four simulations we have $q = 1$, sample size n set to 20, 100, 50, 100 respectively, and p increases from 100 to 1000 in the first two simulations and 2 to 20 in the latter two simulations.