

# The distance correlation $t$ -test of independence in high dimension



Gábor J. Székely<sup>a,b</sup>, Maria L. Rizzo<sup>c,\*</sup>

<sup>a</sup> National Science Foundation, 4201 Wilson Blvd. #1025, Arlington, VA 22230, United States

<sup>b</sup> Rényi Institute of Mathematics, Hungarian Academy of Sciences, Hungary

<sup>c</sup> Department of Mathematics and Statistics, Bowling Green State University, Bowling Green, OH 43403, United States

## ARTICLE INFO

### Article history:

Received 12 August 2010

Available online 27 February 2013

### AMS subject classifications:

primary 62G10

secondary 62H20

### Keywords:

dCor

dCov

Multivariate independence

Distance covariance

Distance correlation

High dimension

## ABSTRACT

Distance correlation is extended to the problem of testing the independence of random vectors in high dimension. Distance correlation characterizes independence and determines a test of multivariate independence for random vectors in arbitrary dimension. In this work, a modified distance correlation statistic is proposed, such that under independence the distribution of a transformation of the statistic converges to Student  $t$ , as dimension tends to infinity. Thus we obtain a distance correlation  $t$ -test for independence of random vectors in arbitrarily high dimension, applicable under standard conditions on the coordinates that ensure the validity of certain limit theorems. This new test is based on an unbiased estimator of distance covariance, and the resulting  $t$ -test is unbiased for every sample size greater than three and all significance levels. The transformed statistic is approximately normal under independence for sample size greater than nine, providing an informative sample coefficient that is easily interpretable for high dimensional data.

© 2013 Elsevier Inc. All rights reserved.

## 1. Introduction

Many applications in genomics, medicine, engineering, etc. require analysis of high dimensional data. Time series data can also be viewed as high dimensional data. Objects can be represented by their characteristics or features as vectors

$$\mathbf{X} = (X_1, \dots, X_p) \in \mathbb{R}^p.$$

In this work, we consider the extension of distance correlation to the problem of testing independence of random vectors in arbitrarily high, not necessarily equal dimensions, so the dimension  $p$  of the feature space of a random vector is typically large.

### 1.1. Overview and background

Distance correlation (dCor) and distance covariance (dCov) (Székely, Rizzo, and Bakirov [19]; Székely and Rizzo [17,18]) measure all types of dependence between random vectors in arbitrary, not necessarily equal dimensions. (See Section 2 for definitions.) Distance correlation takes values in  $[0, 1]$  and is equal to zero if and only if independence holds. It is more general than the classical Pearson product moment correlation, providing a scalar measure of multivariate independence that characterizes independence of random vectors.

\* Corresponding author.

E-mail addresses: [gszekely@nsf.gov](mailto:gszekely@nsf.gov) (G.J. Székely), [mrizzo@bgsu.edu](mailto:mrizzo@bgsu.edu) (M.L. Rizzo).

The distance covariance test of independence is consistent against all dependent alternatives with finite second moments. In practice, however, researchers are often interested in interpreting the numerical value of distance correlation, without a formal test. For example, given an array of distance correlation statistics, what can one learn about the strength of dependence relations from the dCor statistics without a formal test? This is in fact, a difficult question, but a solution is finally available for a large class of problems.

The present work was initially motivated by the observation that the bias of the dCor statistic increases with dimension. We show that, with the help of an unbiased modification of the squared distance covariance, we can construct an unbiased  $t$ -test of independence applicable in high dimension.

In our previous work, we have developed consistent tests for multivariate independence applicable in arbitrary dimension based on the corresponding sample distance covariance. Generally in statistical inference, we consider that, the dimensions of the random variables are fixed and investigate the effect of sample size on inference. In this work, we restrict our attention to the situation where the dimensions of the random vectors are large, relative to the sample size. With the help of a modified distance correlation, we obtain a distance correlation test statistic  $\mathcal{T}$  that has an asymptotic (with respect to dimension) Student  $t$  distribution under independence. Thus we obtain a distance correlation  $t$  test for multivariate independence, applicable in high dimensions. Moreover, the degrees of freedom of the  $t$  statistic are such that the statistic is approximately normal for sample size  $n \geq 10$ . The modified distance correlation statistic for high dimensional data has a symmetric beta distribution, which is approximately normal for moderately large  $n$ , and thus we also obtain an asymptotic (high dimension)  $Z$ -test for independence. The modified distance correlation statistic  $\mathcal{R}_n^*$  converges to the square of population distance correlation ( $\mathcal{R}^2$ ) stochastically. The computing formula and parameters of the  $t$ , beta, and normal limit distributions are simple (linear combinations of Euclidean distances) and the tests are straightforward to apply.

The first high dimensional extension of distance covariance is Kosorok's discussion [8] of Székely and Rizzo [17]. Lyons [10] extended distance covariance to all separable Hilbert spaces which makes our results applicable to functional data. Recent papers that address the problem of testing independence in high dimension include Schott [15] who considers multivariate normal data, Ledoit and Wolf [9] concerning tests for the covariance matrix, and Heer [6]. To date, we are not aware of a test in the literature comparable to the test proposed in this paper, other than the related energy tests of independence already developed by the authors [19,17,1]. A breakthrough in this work and our main result is that, for high dimensional problems, we have derived a modified statistic  $\mathcal{T}_n$  for independence that has a Student  $t$  distribution, so that the statistic is immediately interpretable without any need for Monte Carlo methods. Our methodology also extends to high dimensional problems where the coordinates are not necessarily exchangeable.

1.2. Preliminaries and notation

Suppose that we have random samples of observations

$$(\mathbf{X}_i, \mathbf{Y}_i) \in \mathbb{R}^{p+q}, \quad i = 1, \dots, n,$$

where

$$\mathbf{X}_i = (X_{i,1}, \dots, X_{i,p}), \quad \mathbf{Y}_i = (Y_{i,1}, \dots, Y_{i,q}), \quad i = 1, \dots, n.$$

That is,  $(\mathbf{X}_i, \mathbf{Y}_i) \in \mathbb{R}^{p+q}, i = 1, \dots, n$ , is a random sample from the joint distribution of  $(\mathbf{X}, \mathbf{Y})$ , where  $\mathbf{X}$  and  $\mathbf{Y}$  are assumed to take values in  $\mathbb{R}^p$  and  $\mathbb{R}^q$ , respectively. Throughout this paper we assume that  $E|\mathbf{X}|^2 < \infty$  and  $E|\mathbf{Y}|^2 < \infty$ .

We now summarize certain notation that appears throughout the remaining sections. Notation that is limited in scope to proofs of statements is defined in the Appendix where it is first used. Definitions of dCov and dCor, and computing formulas for statistics follow in Section 2.

A primed symbol denotes an independent copy of the unprimed symbol; that is,  $\mathbf{X}$  and  $\mathbf{X}'$  are independent and identically distributed. The characteristic function of a random vector  $\mathbf{X}$  is denoted by  $\phi_{\mathbf{X}}$ , and the joint characteristic function of random vectors  $\mathbf{X}$  and  $\mathbf{Y}$  is denoted by  $\phi_{\mathbf{X},\mathbf{Y}}$ . The empirical characteristic function of  $\mathbf{X}$  is denoted by  $\phi_{\mathbf{X}}^n$ .

The inner product is denoted with angle brackets and  $|\mathbf{X}| = \langle \mathbf{X}, \mathbf{X} \rangle^{1/2}$  is the Euclidean norm. If the argument of  $|\cdot|$  is complex, then  $|\cdot|$  denotes the complex norm.

The following notation is used for distances, definitions, and computing formulas for the statistics:

$$\begin{aligned} a_{ij} &= |\mathbf{X}_i - \mathbf{X}_j|, \quad i, j = 1, \dots, n, \\ a_{i.} &= \sum_{k=1}^n a_{ik}, \quad a_{.j} = \sum_{k=1}^n a_{kj}, \quad \bar{a}_i = \bar{a}_{i.} = \frac{1}{n} a_{i.}, \\ a_{..} &= \sum_{i,j=1}^n a_{ij}, \quad \bar{a} = \frac{1}{n^2} \sum_{i,j=1}^n a_{ij}, \end{aligned}$$

and the corresponding notation for distances  $b_{ij} = |\mathbf{Y}_i - \mathbf{Y}_j|$ , sums  $b_{i.}, b_{.j}, b_{..}$ , and means  $\bar{b}_i, \bar{b}$  will be used for the second sample. Define  $\alpha = E|\mathbf{X} - \mathbf{X}'|$  and  $\beta = E|\mathbf{Y} - \mathbf{Y}'|$ .

Except for the distances  $a_{ij}, b_{ij}$ , we use upper case letters to denote random variables, and bold letters to denote vectors.

Calligraphic symbols  $\mathcal{V}$ ,  $\mathcal{R}$  are used for population dCov and dCor coefficients, and  $\mathcal{V}_n$ ,  $\mathcal{R}_n$  denote the corresponding sample coefficients as defined in [19,17]. Starred symbols such as  $\mathcal{V}_n^*$  and  $\mathcal{R}_n^*$  are reserved for the modified statistics introduced in this paper.

The paper is organized as follows. Definitions and motivation are covered in Section 2, and the distance correlation  $t$ -test of independence is derived in Section 3. Empirical results are presented in Section 4, followed by a Summary in Section 5. Proofs of statements are given in Appendix.

## 2. Definitions and motivation

For completeness, in Sections 2.1 and 2.2 we restate some important definitions and properties of distance covariance and distance correlation that were first introduced in [19]; modified distance correlation is introduced and defined in Section 2.4.

### 2.1. Distance covariance and distance correlation coefficients

For all distributions with finite first moments, *distance correlation*  $\mathcal{R}$  generalizes the idea of *correlation*, such that:

- i.  $\mathcal{R}(\mathbf{X}, \mathbf{Y})$  is defined for  $\mathbf{X}$  and  $\mathbf{Y}$  in arbitrary dimensions.
- ii.  $\mathcal{R}(\mathbf{X}, \mathbf{Y}) = 0$  if and only if  $\mathbf{X}$  and  $\mathbf{Y}$  are independent.
- iii.  $0 \leq \mathcal{R}(\mathbf{X}, \mathbf{Y}) \leq 1$ .

Distance covariance (dCov) is defined in [19,17], as a measure of the distance between the joint characteristic function  $\phi_{\mathbf{X},\mathbf{Y}}$  of  $\mathbf{X}$  and  $\mathbf{Y}$  and the product  $\phi_{\mathbf{X}}\phi_{\mathbf{Y}}$  of the marginal characteristic functions of  $\mathbf{X}$  and  $\mathbf{Y}$ . In this paper, we focus on the definition corresponding to Euclidean distance, although our results are valid for all powers of Euclidean distance in  $(0, 2)$ . The distance covariance coefficient (for the case of Euclidean distance) is defined by

$$\begin{aligned} \mathcal{V}^2(\mathbf{X}, \mathbf{Y}) &= \|\phi_{\mathbf{X},\mathbf{Y}}(t, s) - \phi_{\mathbf{X}}(t)\phi_{\mathbf{Y}}(s)\|_w^2 \\ &= \int_{\mathbb{R}^{p+q}} |\phi_{\mathbf{X},\mathbf{Y}}(t, s) - \phi_{\mathbf{X}}(t)\phi_{\mathbf{Y}}(s)|^2 w(t, s) dt ds, \end{aligned} \quad (2.1)$$

where

$$\begin{aligned} w(t, s) &= (c_p c_q |t|_p^{1+p} |s|_q^{1+q})^{-1}, \\ c_d &= \frac{\pi^{\frac{1+d}{2}}}{\Gamma(\frac{1+d}{2})}, \end{aligned} \quad (2.2)$$

and  $\Gamma(\cdot)$  is the complete gamma function. This definition is analogous to classical covariance, but with the important property that  $\mathcal{V}^2(\mathbf{X}, \mathbf{Y}) = 0$  if and only if  $\mathbf{X}$  and  $\mathbf{Y}$  are independent.

A standardized version of  $\mathcal{V}(\mathbf{X}, \mathbf{Y})$  is distance correlation, defined as the non-negative square root of

$$\mathcal{R}^2(\mathbf{X}, \mathbf{Y}) = \frac{\mathcal{V}^2(\mathbf{X}, \mathbf{Y})}{\sqrt{\mathcal{V}^2(\mathbf{X})\mathcal{V}^2(\mathbf{Y})}}. \quad (2.3)$$

### 2.2. Distance covariance and distance correlation statistics

The statistics corresponding to the population distance covariance and distance correlation are defined by substituting the empirical characteristic functions in (2.1). Although numerical evaluation of the integral (2.4) appears to be difficult, in fact it can be shown that the resulting statistics are given by an explicit computing formula (2.5) derived in Székely, et al. [19, Theorem 1]. If  $\phi_{\mathbf{X}}^n$ ,  $\phi_{\mathbf{Y}}^n$ , and  $\phi_{\mathbf{X},\mathbf{Y}}^n$  are the empirical characteristic functions of  $\mathbf{X}$ ,  $\mathbf{Y}$ , and  $(\mathbf{X}, \mathbf{Y})$ , respectively, then the *sample distance covariance* of  $\mathbf{X}$ ,  $\mathbf{Y}$  is defined by setting

$$\mathcal{V}_n^2(\mathbf{X}, \mathbf{Y}) = \|\phi_{\mathbf{X},\mathbf{Y}}^n(t, s) - \phi_{\mathbf{X}}^n(t)\phi_{\mathbf{Y}}^n(s)\|_w^2, \quad (2.4)$$

where

$$\|\phi_{\mathbf{X},\mathbf{Y}}^n(t, s) - \phi_{\mathbf{X}}^n(t)\phi_{\mathbf{Y}}^n(s)\|_w^2 = \int_{\mathbb{R}^{p+q}} |\phi_{\mathbf{X},\mathbf{Y}}^n(t, s) - \phi_{\mathbf{X}}^n(t)\phi_{\mathbf{Y}}^n(s)|^2 w(t, s) dt ds,$$

and the weight function  $w(t, s)$  is defined by (2.2). Our original definition [19,17] of the sample distance covariance is the non-negative square root of

$$\text{dCov}_n^2(\mathbf{X}, \mathbf{Y}) = \frac{1}{n^2} \sum_{i,j} A_{i,j} B_{i,j}, \quad (2.5)$$

where

$$A_{i,j} = |\mathbf{X}_i - \mathbf{X}_j| - \frac{1}{n} \sum_{k=1}^n |\mathbf{X}_k - \mathbf{X}_j| - \frac{1}{n} \sum_{l=1}^n |\mathbf{X}_i - \mathbf{X}_l| + \frac{1}{n^2} \sum_{k,l=1}^n |\mathbf{X}_k - \mathbf{X}_l|,$$

$$B_{i,j} = |\mathbf{Y}_i - \mathbf{Y}_j| - \frac{1}{n} \sum_{k=1}^n |\mathbf{Y}_k - \mathbf{Y}_j| - \frac{1}{n} \sum_{l=1}^n |\mathbf{Y}_i - \mathbf{Y}_l| + \frac{1}{n^2} \sum_{k,l=1}^n |\mathbf{Y}_k - \mathbf{Y}_l|,$$

$i, j = 1, \dots, n$ , and  $|\cdot|$  denotes the Euclidean norm. Thus

$$A_{i,j} = a_{ij} - \bar{a}_i - \bar{a}_j + \bar{a}; \quad B_{i,j} = b_{ij} - \bar{b}_i - \bar{b}_j + \bar{b},$$

for  $i, j = 1, \dots, n$ . Theorem 1 [19] establishes the identity  $\mathcal{V}_n^2(\mathbf{X}, \mathbf{Y}) = \text{dCov}_n^2(\mathbf{X}, \mathbf{Y})$ . Sample distance correlation is defined by

$$\mathcal{R}_n^2(\mathbf{X}, \mathbf{Y}) = \frac{\mathcal{V}_n^2(\mathbf{X}, \mathbf{Y})}{\sqrt{\mathcal{V}_n^2(\mathbf{X}, \mathbf{X}) \cdot \mathcal{V}_n^2(\mathbf{Y}, \mathbf{Y})}}. \tag{2.6}$$

For independent random vectors  $\mathbf{X}$  and  $\mathbf{Y}$  with finite first moments (in fixed dimensions),  $n \mathcal{V}_n^2(\mathbf{X}, \mathbf{Y})$  converges in distribution to a quadratic form of centered Gaussian random variables

$$\sum_{i=1}^{\infty} \lambda_i Z_i^2,$$

as sample size  $n$  tends to infinity [19, Theorem 5], where  $Z_i$  are iid standard normal random variables and  $\lambda_i$  are positive constants that depend on the distributions of  $\mathbf{X}$  and  $\mathbf{Y}$ .

In this work, we prove a corresponding limit theorem as dimension tends to infinity; this limit is obtained for a related, modified version of distance correlation defined in Section 2.4.

### 2.3. Motivation

First, let us see why a modified version of sample distance covariance and distance correlation is advantageous in high dimension. We begin by observing that, although dCor characterizes independence in arbitrary dimension, the numerical value of the (original) corresponding statistic can be difficult to interpret in high dimension without a formal test.

Recall that,  $\alpha = E|\mathbf{X} - \mathbf{X}'|$  and  $\beta = E|\mathbf{Y} - \mathbf{Y}'|$ . It is easy to see that for any fixed  $n$ ,

$$E[A_{i,j}] = \begin{cases} \frac{\alpha}{n}, & i \neq j; \\ \frac{\alpha}{n} - \alpha, & i = j; \end{cases} \quad E[B_{i,j}] = \begin{cases} \frac{\beta}{n}, & i \neq j; \\ \frac{\beta}{n} - \beta, & i = j. \end{cases}$$

It can be shown (see Appendix A.1) that for an important class of distributions including independent standard multivariate normal  $\mathbf{X}$  and  $\mathbf{Y}$ , for fixed  $n$  each of the statistics

$$\frac{\mathcal{V}_n^2(\mathbf{X}, \mathbf{Y})}{\alpha\beta}, \quad \frac{\mathcal{V}_n^2(\mathbf{X}, \mathbf{X})}{\alpha^2}, \quad \frac{\mathcal{V}_n^2(\mathbf{Y}, \mathbf{Y})}{\beta^2}$$

converges to  $(n - 1)/n^2$  as dimensions  $p, q$  tend to infinity. Thus, for sample distance correlation, it follows that

$$\mathcal{R}_n^2(\mathbf{X}, \mathbf{Y}) = \frac{\mathcal{V}_n^2(\mathbf{X}, \mathbf{Y})}{\sqrt{\mathcal{V}_n^2(\mathbf{X}, \mathbf{X}) \cdot \mathcal{V}_n^2(\mathbf{Y}, \mathbf{Y})}} \xrightarrow{p,q \rightarrow \infty} 1, \tag{2.7}$$

even though  $\mathbf{X}$  and  $\mathbf{Y}$  are independent.

Here we see that, although distance correlation characterizes independence, and the dCov test of independence is valid for  $\mathbf{X}, \mathbf{Y}$  in arbitrary dimensions, interpretation of the size of the sample distance correlation coefficient without a formal test becomes more difficult for  $\mathbf{X}$  and  $\mathbf{Y}$  in high dimensions. See Example 1 for an illustration of how the corrected statistics and  $t$ -test address this issue.

We propose a modified distance covariance statistic such that under independence, a transformation of the corresponding distance correlation statistic converges (as  $p, q \rightarrow \infty$ ) to a Student  $t$  distribution, which is approximately normal for  $p, q > n \geq 10$ , providing an easily interpretable sample coefficient.

*Numerical illustration.* Table 1 illustrates the original and modified distance correlation statistics with a numerical example. We generated independent samples with iid Uniform(0,1) coordinates and computed  $\mathcal{R}_n^2$ , modified distance correlation  $\mathcal{R}_n^*$  (Section 2.4), and the corresponding  $t$  and  $Z$  statistics (Section 3). Each row of the table reports values for one pair of samples for dimension  $p = q$  and  $n = 30$ . The numerical value of  $\mathcal{R}_n^2$  approaches 1 as dimension increases, even under independence. Without our dCov test, numerical interpretation of original  $\mathcal{R}_n^2$  is difficult. In contrast, we see that the modified statistics  $\mathcal{R}_n^*$  in the table are centered close to zero and stable with respect to dimension.

2.4. Modified distance covariance statistics

A modified version of the statistic  $\mathcal{V}_n^2(\mathbf{X}, \mathbf{Y})$  that avoids (2.7) can be defined starting with corrected  $A_{i,j}$  and  $B_{i,j}$ . Note that in the original formulation,  $E[A_{i,j}] = \alpha/n$  if  $i \neq j$  and  $E[A_{i,i}] = \alpha/n - \alpha$ . Thus in high dimension the difference  $\alpha$  between the diagonal and off-diagonal entries of  $A$  (and  $B$ ) can be large. The modified versions  $A_{i,j}^*$ ,  $B_{i,j}^*$  of  $A_{i,j}$  and  $B_{i,j}$  are defined by

$$A_{i,j}^* = \begin{cases} \frac{n}{n-1} \left( A_{i,j} - \frac{a_{ij}}{n} \right), & i \neq j; \\ \frac{n}{n-1} (\bar{a}_i - \bar{a}), & i = j, \end{cases} \quad B_{i,j}^* = \begin{cases} \frac{n}{n-1} \left( B_{i,j} - \frac{b_{ij}}{n} \right), & i \neq j; \\ \frac{n}{n-1} (\bar{b}_i - \bar{b}), & i = j. \end{cases}$$

One can easily see that  $E[A_{i,j}^*] = E[B_{i,j}^*] = 0$  for all  $i, j$ .

We now define modified distance covariance and modified distance correlation statistics using the corrected terms  $A_{i,j}^*$  and  $B_{i,j}^*$ . Let

$$\mathcal{U}_n^*(\mathbf{X}, \mathbf{Y}) = \sum_{i \neq j} A_{i,j}^* B_{i,j}^* - \frac{2}{n-2} \sum_{i=1}^n A_{i,i}^* B_{i,i}^*. \tag{2.8}$$

**Definition 1.** The modified distance covariance statistic is

$$\mathcal{V}_n^*(\mathbf{X}, \mathbf{Y}) = \frac{\mathcal{U}_n^*(\mathbf{X}, \mathbf{Y})}{n(n-3)} = \frac{1}{n(n-3)} \left\{ \sum_{i,j=1}^n A_{i,j}^* B_{i,j}^* - \frac{n}{n-2} \sum_{i=1}^n A_{i,i}^* B_{i,i}^* \right\}. \tag{2.9}$$

It can be shown that  $E[\mathcal{U}_n^*(\mathbf{X}, \mathbf{Y})] = n(n-3)\mathcal{V}_n^2(\mathbf{X}, \mathbf{Y})$  (see Proposition 2 below), therefore  $\mathcal{V}_n^*(\mathbf{X}, \mathbf{Y})$  is an unbiased estimator of the squared population distance covariance.

In Lemma 3 it is proved that  $\mathcal{U}_n^*(\mathbf{X}, \mathbf{X}) \geq 0$  and  $\mathcal{U}_n^*(\mathbf{Y}, \mathbf{Y}) \geq 0$ , so that  $\sqrt{\mathcal{V}_n^*(\mathbf{X}, \mathbf{X})\mathcal{V}_n^*(\mathbf{Y}, \mathbf{Y})}$  is always a real number for  $n \geq 3$ .

**Definition 2.** The modified distance correlation statistic is

$$\mathcal{R}_n^*(\mathbf{X}, \mathbf{Y}) = \frac{\mathcal{V}_n^*(\mathbf{X}, \mathbf{Y})}{\sqrt{\mathcal{V}_n^*(\mathbf{X}, \mathbf{X})\mathcal{V}_n^*(\mathbf{Y}, \mathbf{Y})}}, \tag{2.10}$$

if  $\mathcal{V}_n^*(\mathbf{X}, \mathbf{X})\mathcal{V}_n^*(\mathbf{Y}, \mathbf{Y}) > 0$ , and otherwise  $\mathcal{R}_n^*(\mathbf{X}, \mathbf{Y}) = 0$ .

While the original  $\mathcal{R}_n$  statistic is between 0 and 1,  $\mathcal{R}_n^*$  can take negative values; the Cauchy–Schwartz inequality implies that  $|\mathcal{R}_n^*| \leq 1$ . Later we will see that  $\mathcal{R}_n^*$  converges to  $\mathcal{R}^2$  stochastically. In the next section, we derive the limit distribution of  $\mathcal{R}_n^*$ .

In the following, we exclude  $|\mathcal{R}_n^*| = 1$ , corresponding to the case when the  $\mathbf{X}$  sample is a linear transformation of the  $\mathbf{Y}$  sample.<sup>1</sup>

3. The  $t$ -test for independence in high dimension

Our main result is that as  $p, q$  tend to infinity, under the independence hypothesis,

$$\mathcal{T}_n = \sqrt{\nu - 1} \cdot \frac{\mathcal{R}_n^*}{\sqrt{1 - (\mathcal{R}_n^*)^2}}$$

converges in distribution to Student  $t$  with  $\nu - 1$  degrees of freedom, where  $\nu = \frac{n(n-3)}{2}$ . Thus for  $n \geq 10$  this limit is approximately standard normal. The  $t$ -test of independence is unbiased for every  $n \geq 4$  and any significance level. As a corollary to our main result Theorem 1, it follows that  $(\mathcal{R}_n^* + 1)/2$  has a symmetric beta distribution. We also obtain as a corollary to Theorem 1 that

$$\sqrt{\nu - 1} \mathcal{R}_n^*$$

is asymptotically standard normal.

Our procedure for testing independence applies distances  $|\mathbf{X}_i - \mathbf{X}_j|$ , and  $E|\mathbf{X}_i|^2 < \infty$ , so without loss of generality we can assume that  $E[\mathbf{X}_i] = 0$ .

<sup>1</sup> To be more precise,  $|\mathcal{R}_n^*| = 1$  implies that, the linear spaces spanned by the sample observations  $\mathbf{X}$  and  $\mathbf{Y}$ , respectively, have the same dimension; thus we can represent the two samples in the same linear space, and in this common space the  $\mathbf{Y}$  sample is a linear function of the  $\mathbf{X}$  sample.

**Table 1**  
 Numerical illustration ( $n = 30$ ) of distance correlation  $\mathcal{R}_n$  and modified distance correlation  $\mathcal{R}_n^*$  statistics in high dimension. The modified statistic  $\mathcal{R}_n^*$  is based on an unbiased estimator of the squared population distance covariance. Each row of the table reports statistics for one sample  $(\mathbf{X}, \mathbf{Y})$ , where  $\mathbf{X}$  and  $\mathbf{Y}$  each have iid standard uniform coordinates. The  $t$  statistic  $\mathcal{T}_n$  and  $Z$  statistic introduced in Section 3 are also reported.

$p, q$	$\mathcal{R}_n$	$\mathcal{R}_n^*$	$\mathcal{T}_n$	$Z$
1	0.4302668	0.0248998	0.5006350	0.5004797
2	0.6443883	0.1181823	2.3921989	2.3754341
4	0.7103136	-0.0098916	-0.1988283	-0.1988186
8	0.8373288	0.0367129	0.7384188	0.7379210
16	0.8922197	-0.0675606	-1.3610616	-1.3579518
32	0.9428649	-0.0768243	-1.5487268	-1.5441497
64	0.9702281	-0.1110683	-2.2463438	-2.2324451
128	0.9864912	-0.0016547	-0.0332595	-0.0332595
256	0.9931836	0.0517415	1.0413867	1.0399917
512	0.9963244	-0.0158947	-0.3195190	-0.3194787
1024	0.9983706	0.0542560	1.0921411	1.0905325
2048	0.9991116	-0.0076502	-0.1537715	-0.1537670
4096	0.9995349	-0.0863294	-1.7417010	-1.7351986
8192	0.9997596	-0.0754827	-1.5215235	-1.5171827
16384	0.9999032	0.0277193	0.5573647	0.5571505

For simplicity we suppose in the main proof that random vectors  $\mathbf{X}$  and  $\mathbf{Y}$  have iid coordinates with finite variance. It will be clear from the proofs that much weaker conditions are also sufficient, because what we really need is that for partial sums of squared coordinates certain limit theorems, like the Weak Law of Large Numbers and Central Limit Theorem (CLT) hold. Corollary 2 below deals with the case when the coordinates are exchangeable. For times series even this condition is too strong. For typical strongly stationary time series we can apply Proposition 3.

The following related statistics will be used in deriving our main result. Let

$$\mathcal{W}_n(\mathbf{X}, \mathbf{Y}) = \frac{2n - 1}{n^3} \sum_{i,j=1}^n a_{ij}b_{ij} - \bar{a}\bar{b} + \frac{2}{n^2 - 2n} \sum_{i=1}^n (\bar{a}_i - \bar{a})(\bar{b}_i - \bar{b}), \tag{3.1}$$

and define

$$\mathcal{U}_n(\mathbf{X}, \mathbf{Y}) \stackrel{\text{def}}{=} \mathcal{V}_n^2(\mathbf{X}, \mathbf{Y}) - \frac{\mathcal{W}_n(\mathbf{X}, \mathbf{Y})}{n}. \tag{3.2}$$

**Lemma 1.** *The following claims hold:*

- i. For all  $\mathbf{X}$  and all  $n \geq 3$ ,  $\mathcal{W}_n(\mathbf{X}, \mathbf{X}) \geq 0$ .
- ii. For all  $\mathbf{X}$  and all  $n \geq 3$ ,  $\mathcal{U}_n(\mathbf{X}, \mathbf{X}) \geq 0$ .
- iii. If  $E|\mathbf{X}_i|^2 < \infty$ ,  $E|\mathbf{Y}_i|^2 < \infty$ ,  $i = 1, 2$ , then for any fixed  $p$  and  $q$ ,

$$\mathcal{W}_n(\mathbf{X}, \mathbf{Y}) \xrightarrow[n \rightarrow \infty]{\mathcal{P}} 2E|\mathbf{X} - \mathbf{X}'||\mathbf{Y} - \mathbf{Y}'| - E|\mathbf{X} - \mathbf{X}'|E|\mathbf{Y} - \mathbf{Y}'|.$$

- iv. If  $E|\mathbf{X}|^2 < \infty$ ,  $E|\mathbf{Y}|^2 < \infty$  and  $\mathbf{X}$  and  $\mathbf{Y}$  are independent, then for any fixed  $p$  and  $q$ ,

$$\mathcal{W}_n(\mathbf{X}, \mathbf{Y}) \xrightarrow[n \rightarrow \infty]{\mathcal{P}} \alpha\beta.$$

Proof of Lemma 1 is given in Appendix A.2.

**Remark 1.** Formula (3.2) and Lemma 1 (iii) imply that

$$\mathcal{V}_n^2(\mathbf{X}, \mathbf{Y}) = \mathcal{U}_n(\mathbf{X}, \mathbf{Y}) + o_p(1), \quad n \rightarrow \infty,$$

where  $o_p(1)$  is a term that converges to zero in probability as  $n \rightarrow \infty$ . Therefore  $\mathcal{U}_n(\mathbf{X}, \mathbf{Y})$  really can be viewed as a modified distance covariance and  $\mathcal{U}_n(\mathbf{X}, \mathbf{X})$  can be viewed as a modified distance variance.

**Lemma 2.** *The following identity holds:*

$$n^2 \mathcal{U}_n(\mathbf{X}, \mathbf{Y}) = \frac{(n - 1)^2}{n^2} \cdot \mathcal{U}_n^*(\mathbf{X}, \mathbf{Y}). \tag{3.3}$$

Proof of Lemma 2 is given in Appendix A.3.

Thus the following decomposition holds for all  $n \geq 3$ :

$$n^2 \mathcal{V}_n^2(\mathbf{X}, \mathbf{X}) = \left(1 - \frac{1}{n}\right)^2 \cdot \mathcal{U}_n^*(\mathbf{X}, \mathbf{X}) + n \mathcal{W}_n(\mathbf{X}, \mathbf{X}),$$

where  $\mathcal{U}_n^*(\mathbf{X}, \mathbf{X}) \geq 0$ ,  $\mathcal{W}_n(\mathbf{X}, \mathbf{X}) \geq 0$ .

Recall that  $\mathbf{X}'$  denotes an independent copy of  $\mathbf{X}$ . In what follows,  $(\mathbf{X}, \mathbf{Y})$ ,  $(\mathbf{X}', \mathbf{Y}')$ , and  $(\mathbf{X}'', \mathbf{Y}'')$  are independent and identically distributed.

**Proposition 1.** *If  $\mathbf{X}$  and  $\mathbf{Y}$  have finite second moments, then*

i.

$$\mathcal{V}^2(\mathbf{X}, \mathbf{Y}) = E|\mathbf{X} - \mathbf{X}'||\mathbf{Y} - \mathbf{Y}'| + E|\mathbf{X} - \mathbf{X}'|E|\mathbf{Y} - \mathbf{Y}'| - 2E|\mathbf{X} - \mathbf{X}'||\mathbf{Y} - \mathbf{Y}''|.$$

ii. *If  $\mathbf{X}$  and  $\mathbf{Y}$  are independent, then*

$$E[\mathcal{V}_n^2(\mathbf{X}, \mathbf{Y})] = \frac{n-1}{n^2} \{E|\mathbf{X} - \mathbf{X}'|E|\mathbf{Y} - \mathbf{Y}'|\}.$$

iii. *If  $\mathbf{X}$  and  $\mathbf{Y}$  are independent, then  $E[\mathcal{V}_n^*(\mathbf{X}, \mathbf{Y})] = 0$ .*

See Appendix A.6 for proof of Proposition 1.

**Proposition 2.** *For all  $n, p, q$ , the modified distance covariance statistic  $\mathcal{V}_n^*(\mathbf{X}, \mathbf{Y})$  is an unbiased estimator of the squared distance covariance  $\mathcal{V}^2(\mathbf{X}, \mathbf{Y})$ , and*

$$E[\mathcal{V}_n^2(\mathbf{X}, \mathbf{Y})] = \frac{(n-1)}{n^3} [(n-2)^2 \mathcal{V}^2(\mathbf{X}, \mathbf{Y}) + 2(n-1)\mu - (n-2)\alpha\beta],$$

$$E[\mathcal{W}_n(\mathbf{X}, \mathbf{Y})] = \frac{n-1}{n^2} [(2n-1)\mu - (n-3)\alpha\beta - 2\delta],$$

$$E[\mathcal{U}_n(\mathbf{X}, \mathbf{Y})] = \frac{(n-1)^2(n-3)}{n^3} \mathcal{V}^2(\mathbf{X}, \mathbf{Y}),$$

$$E[\mathcal{U}_n^*(\mathbf{X}, \mathbf{Y})] = \frac{n^4}{(n-1)^2} E[\mathcal{U}_n(\mathbf{X}, \mathbf{Y})] = n(n-3)\mathcal{V}^2(\mathbf{X}, \mathbf{Y}),$$

where  $\mu = E|\mathbf{X} - \mathbf{X}'||\mathbf{Y} - \mathbf{Y}'|$ ,  $\alpha = E|\mathbf{X} - \mathbf{X}'|$ ,  $\beta = E|\mathbf{Y} - \mathbf{Y}'|$ , and  $\delta = E|\mathbf{X} - \mathbf{X}'||\mathbf{Y} - \mathbf{Y}''|$ .

See Appendix A.7 for proof of Proposition 2.

**Lemma 3.** *If the coordinates of  $\mathbf{X}$  and  $\mathbf{Y}$  are iid,  $0 < E|\mathbf{X}|^2 < \infty$ ,  $0 < E|\mathbf{Y}|^2 < \infty$ , and  $\mathbf{X}$  and  $\mathbf{Y}$  are independent, then for fixed  $n$  there exist independent random variables  $\Omega_{i,j}$ ,  $\Psi_{i,j}$ , such that*

$$(i) \mathcal{U}_n^*(\mathbf{X}, \mathbf{X}) \xrightarrow{p \rightarrow \infty} \sum_{i \neq j} \Omega_{i,j}^2 \stackrel{\mathcal{D}}{=} 2\sigma_X^2 \chi_v^2, \quad (3.4)$$

$$(ii) \mathcal{U}_n^*(\mathbf{Y}, \mathbf{Y}) \xrightarrow{q \rightarrow \infty} \sum_{i \neq j} \Psi_{i,j}^2 \stackrel{\mathcal{D}}{=} 2\sigma_Y^2 \chi_v^2, \quad (3.5)$$

$$(iii) \mathcal{U}_n^*(\mathbf{X}, \mathbf{Y}) \xrightarrow{p, q \rightarrow \infty} \sum_{i \neq j} \Omega_{i,j} \Psi_{i,j}, \quad (3.6)$$

where  $v = \frac{n(n-3)}{2}$ ,  $\chi_v^2$  denotes the distribution of a chisquare random variable with  $v$  degrees of freedom,

$$\sigma_X^2 = \frac{E(\mathbf{X}, \mathbf{X}')^2}{2E|\mathbf{X}|^2}, \quad \sigma_Y^2 = \frac{E(\mathbf{Y}, \mathbf{Y}')^2}{2E|\mathbf{Y}|^2},$$

$\Omega_{ij}$  are iid Normal  $(0, \sigma_X^2)$ , and  $\Psi_{ij}$  are iid Normal  $(0, \sigma_Y^2)$ .

See Appendix A.4 for the proof of Lemma 3. The variables  $\Omega_{i,j}$  and  $\Psi_{i,j}$  are defined in the proof by Eqs. (A.11) and (A.12). For the corresponding correlation coefficient we have

$$\mathcal{R}_n^* = \frac{\mathcal{U}_n^*(\mathbf{X}, \mathbf{Y})}{\sqrt{\mathcal{U}_n^*(\mathbf{X}, \mathbf{X})\mathcal{U}_n^*(\mathbf{Y}, \mathbf{Y})}} \xrightarrow{p, q \rightarrow \infty} \frac{\sum_{i \neq j} \Omega_{i,j} \Psi_{i,j}}{\sqrt{\sum_{i \neq j} \Omega_{i,j}^2 \sum_{i \neq j} \Psi_{i,j}^2}}.$$

Define the test statistic

$$\mathcal{T}_n = \sqrt{v-1} \cdot \frac{\mathcal{R}_n^*}{\sqrt{1 - (\mathcal{R}_n^*)^2}}. \quad (3.7)$$

**Theorem 1.** *If the coordinates of  $\mathbf{X}$  and  $\mathbf{Y}$  are iid with positive finite variance, for fixed sample size  $n \geq 4$  the following hold.*

(i) *Under independence of  $\mathbf{X}$  and  $\mathbf{Y}$ ,*

$$P\{\mathcal{T}_n < t\} \xrightarrow[p, q \rightarrow \infty]{} P\{t_{\nu-1} < t\},$$

where  $\mathcal{T}_n$  is the statistic (3.7) and  $\nu = \frac{n(n-3)}{2}$ .

(ii) *Let  $c_\alpha = t_{\nu-1}^{-1}(1 - \alpha)$  denote the  $(1 - \alpha)$  quantile of a Student  $t$  distribution with  $\nu - 1$  degrees of freedom. The  $t$ -test of independence at significance level  $\alpha$  that rejects the independence hypothesis whenever  $\mathcal{T}_n > c_\alpha$  is unbiased.*

The  $t$ -test of independence rejects the null hypothesis at level  $\alpha$  if  $\mathcal{T}_n > c_\alpha$ , where  $c_\alpha = t_{\nu-1}^{-1}(1 - \alpha)$  is the  $(1 - \alpha)$  quantile of a Student  $t$  distribution with  $\nu - 1$  degrees of freedom. By Theorem 1(i), the test has level  $\alpha$ . See Appendix A.8 for the proof of Theorem 1.

We also obtain a  $Z$ -test of independence in high dimension:

**Corollary 1.** *Under independence of  $\mathbf{X}$  and  $\mathbf{Y}$ , if the coordinates of  $\mathbf{X}$  and  $\mathbf{Y}$  are iid with positive finite variance, then the limit distribution of  $(1 + \mathcal{R}_n^*)/2$  is a symmetric beta distribution with shape parameter  $(\nu - 1)/2$ . It follows that, in high dimension the large sample distribution of  $\sqrt{\nu - 1} \mathcal{R}_n^*$  is approximately standard normal.*

Corollary 1 follows from Theorem 1 and well known distributional results relating Student  $t$ , beta, and normal distributions.

**Corollary 2.** *Theorem 1 and Corollary 1 hold for random vectors with exchangeable coordinates and finite variance.*

Infinite sequences of random variables are known to be conditionally iid with respect to the sigma algebra of symmetric events (de Finetti [4]). If the series is finite, one can refer to Kerns and Székely [7] and Diaconis and Freedman [5]. Further, if we assume that the variance of each of the variables is finite, then in the Corollary we can assume that conditionally with respect to the sigma algebra of symmetric events, the CLT holds. The only factor that might change depending on the condition is the variance of the normal limit. However, in  $\mathcal{R}_n^*$  this variance factor cancels, hence  $\mathcal{R}_n^*$  (and therefore  $\mathcal{T}_n$ ) has the same distribution with or without this condition.

**Example 1.** To illustrate the application of the  $t$ -test of independence in Theorem 1, we revisit an example of the type discussed in Section 2.3. Recall that, in Appendix A.1 it was shown that as dimension tends to infinity, the (uncorrected) distance correlation approaches 1. Let us now apply the corrected statistics and  $t$ -test for independent standard multivariate normal  $\mathbf{X} \in \mathbb{R}^{30}$ ,  $\mathbf{Y} \in \mathbb{R}^{30}$ , with sample size  $n = 30$ . The result of our  $t$ -test, coded in R, is summarized below.

```
> highdim.ttest(X, Y)

dcor t-test of independence

data: X and Y
T = 0.8774, p-value = 0.1904
sample estimates:
R*
0.0436113
```

Here the corrected statistic is  $\mathcal{R}_n^* = 0.0436$  and  $\mathcal{T}_n = 0.8774$ , with 404 degrees of freedom, which is easily interpreted as non-significant without reference to a table or software.

In a simulation of 1000 tests, the Type 1 error rate was 0.100 at 10% significance, and 0.046 at 5% significance. A probability histogram of the replicates is shown in Fig. 1(a).

We repeated the example with a slight modification so that  $\mathbf{X}$  and  $\mathbf{Y}$  are linearly dependent with  $\mathbf{Y} = \mathbf{X} + \boldsymbol{\varepsilon}$ , where  $\boldsymbol{\varepsilon}$  is Gaussian with mean zero and covariance  $2I$ . In this simulation the null hypothesis is rejected for all 1000 samples at level 0.05. The histogram of the simulated test statistics is shown in Fig. 1(b).

#### 4. Application to time series

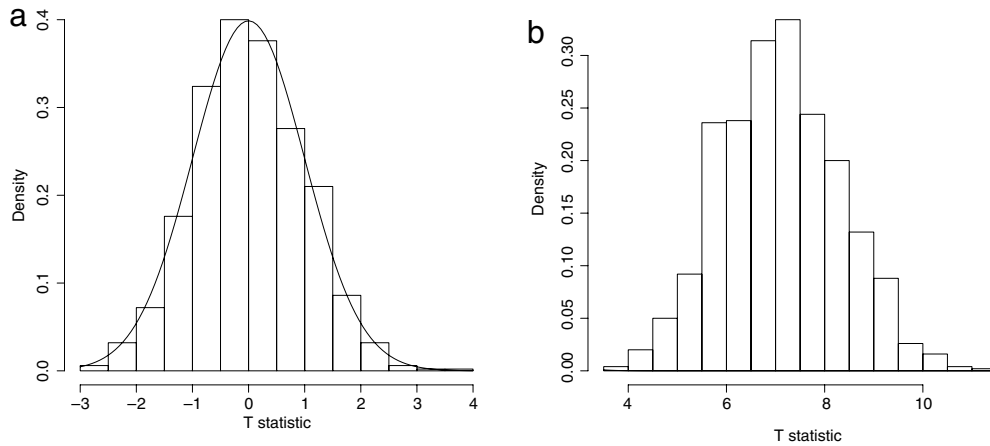
In this section, we discuss the application of the distance correlation  $t$  statistic to test independence of two time series.

Let  $\{X(t), Y(t)\}$  be a strongly stationary times series where for simplicity both  $X(t)$  and  $Y(t)$  are real valued. Strong stationarity guarantees that if we take a sample of  $p$  consecutive observations from  $\{X(t), Y(t)\}$ , then their joint distribution and thus their dependence structure do not depend on the starting point. On the other hand, strong stationarity is not enough to guarantee CLT for partial sums, not even conditionally with respect to a sigma algebra. (The extra condition of  $m$ -dependence of  $\{X(t), Y(t)\}$  would be enough by a classical theorem of Hoeffding, but this condition is too strong.)

In order to apply our  $t$ -test; we need the conditional validity of CLT, conditioned on a sigma algebra. Then the variance of the normal limit becomes a random variable and the possible limits are scale mixtures of normals. Many of these distributions are heavy tailed, which is important in financial applications.

Let us summarize briefly, conditions for CLT that are relevant in this context.





**Fig. 1.** Simulated sampling distribution of  $\mathcal{T}_n$  in Example 1 ( $n = 30, p = q = 30$ ) for standard multivariate normal  $\mathbf{X}, \mathbf{Y}$  (a) under independence, with limiting  $t$  density, and (b) under linear dependence.

- i. If our observations are exchangeable then de Finetti [4] (or in the finite case Kerns and Székely [7]) applies and if the (conditional) variances are finite then the conditional CLT follows. See also Diaconis and Freedman [5].
- ii. For strongly stationary sequences we can refer to Ibragimov's conjecture from the 1960's, and to Peligrad [11] (1985) for a proof of a somewhat weaker claim. (For strongly stationary sequences with finite variance such that  $\text{Var}(S_n)$  tends to infinity, the CLT does not always follow, not even the conditional CLT; thus in general some extra conditions are needed.)
- iii. Stein type of dependence was introduced by Charles Stein [16], who first obtained in 1972 a bound between the distribution of a sum of an  $m$ -dependent sequence of random variables and a standard normal distribution in the Kolmogorov (uniform) metric, and hence proved not only a central limit theorem, but also bounds on the rates of convergence for the given metric.

In order to apply our  $t$ -test we also need iid observations at least conditionally with respect to a sigma algebra. Typically, for time series, only one realization of each series is available for analysis. A random sample of iid observations is obtained (at least conditionally iid) for analysis by the application of the following proposition.

**Proposition 3.** Fix  $p < N$  and let  $T_1, T_2, \dots, T_n$  be integers in  $\{1, \dots, N - p + 1\}$ . Define  $X_j$  to be the subsequence of length  $p$  starting with  $X(T_j)$ , and similarly define  $Y_j$ ; that is,

$$X_j = \{X(T_j), X(T_j + 1), \dots, X(T_j + p - 1)\}, \quad j = 1, \dots, n,$$

$$Y_j = \{Y(T_j), Y(T_j + 1), \dots, Y(T_j + p - 1)\}, \quad j = 1, \dots, n.$$

If  $X \in \mathbb{R}^d$ , then  $X(T_j) \in \mathbb{R}^d$  and  $X_j$  is a vector in  $\mathbb{R}^{pd}$ ; that is,  $X_j = \{X(T_j)_1, \dots, X(T_j)_d, X(T_j + 1)_1, \dots, X(T_j + 1)_d, \dots, X(T_j + p - 1)_1, \dots, X(T_j + p - 1)_d\}$ . If  $T_j$  are drawn at random with equal probability from  $\{1, \dots, N - p + 1\}$ , these vectors  $\{X_j\}$  are exchangeable; thus, conditional with respect to the sigma algebra of symmetric events, they are iid observations. Similarly, the vectors  $\{Y_j\}$  are also conditionally iid, and thus if the variances are finite, we can apply the  $t$ -test of independence conditioned on the sigma algebra of symmetric events. Hence the  $t$ -test of independence is applicable unconditionally.

Corollary 2 and Proposition 3 imply that conditional with respect to a sigma algebra (to the sigma algebra of symmetric events), these vectors are iid with finite variances. Thus, we can apply the  $t$ -test of independence conditioned on this sigma algebra. The variance here is random (thus we can have heavy tailed distributions) but in the formula for  $t$  the variance cancels hence the  $t$ -test of independence is applicable unconditionally.

In summary, our method is applicable for financial time series if we suppose that the differences of the logarithms of our observations (or other suitable transformation) form a strongly stationary sequence whose partial sums conditionally with respect to a sigma algebra tend to normal.

The implementation of the  $t$ -test of independence is straightforward, and all empirical results presented below were performed using R software [12]. See the *energy* package for R [14] for an implementation of our methods available under general public license.

**Example 2 (AR(1) Series).** To illustrate the  $t$ -test of independence implementation using the randomization method of Proposition 3, we applied the test to pairs of AR(1) (autoregressive model order 1) time series. The AR(1) data with total length  $N = 2500$  was generated from the model  $X_t = AX_{t-1} + e_t$ , where  $X_t$  is a vector of length 2,  $A$  is a matrix of autoregressive coefficients, and  $e_t \in \mathbb{R}^2$  is a noise vector with mean 0.<sup>2</sup> The bivariate AR(1) model parameters used in

<sup>2</sup> Here for simplicity  $e_t$  is Gaussian, but the method is applicable for non-normal error distributions as well.

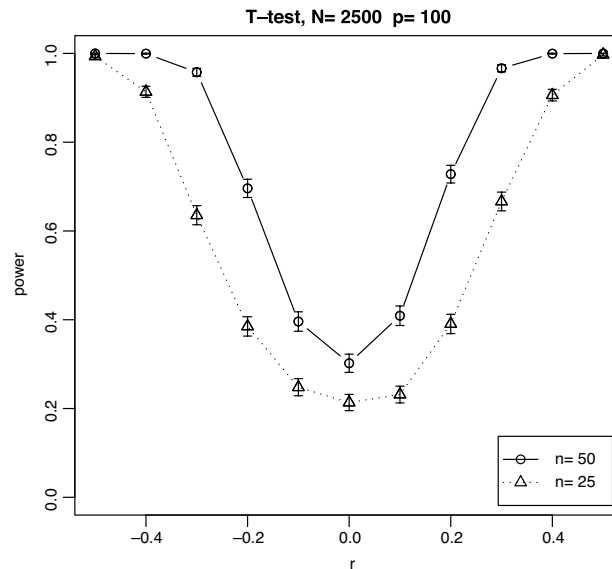


Fig. 2. Proportion of significant tests of independence on dependent AR(1) series. The AR parameter is 0.25 and error correlation is  $r$ .

this simulation are

$$A = \begin{pmatrix} 0.25 & 0 \\ 0 & 0.25 \end{pmatrix}; \quad \text{Cov}(e) = \begin{pmatrix} 1 & r \\ r & 1 \end{pmatrix},$$

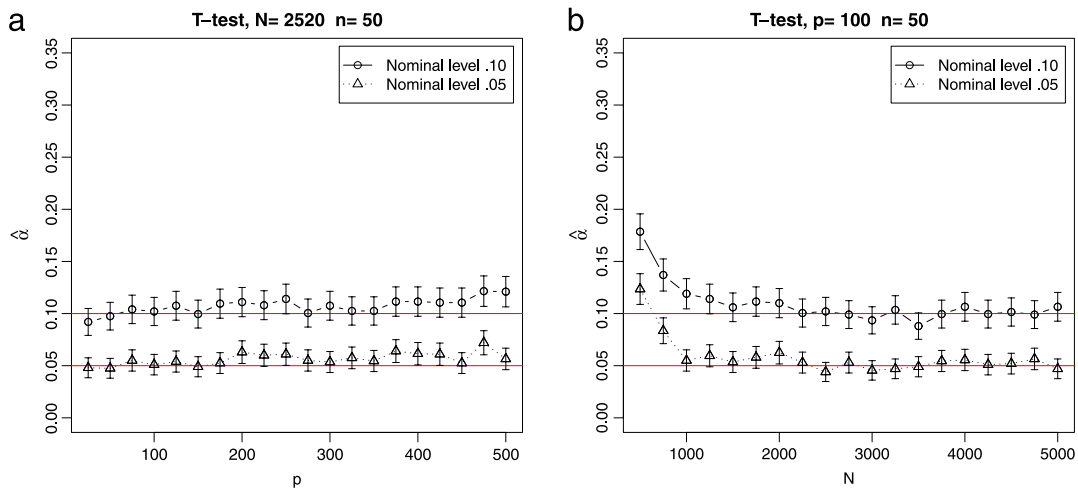
and the length of subsequences is  $p = 100$ . The estimates of power for varying  $r$  are shown in Fig. 2, with error bars at  $\pm 2$  standard errors. These estimates are computed as the proportion of significant  $t$ -tests of independence in 2000 replications (significance level 10%). Fig. 2 summarizes the simulation results for two cases: sample size  $n = 25$  and  $n = 50$ . (The length  $N = 2500$  for this example was chosen to match approximately the length of the DJIA data in Example 4). One can observe that, the empirical power of the test is higher for larger sample size.

**Example 3 (Assessing Type 1 Error Rate).** To assess the Type 1 error rate of this procedure, we compared two white noise series using the same parameters  $N = 2520$ ,  $p = 100$  and  $n = 50$ , as in the DJIA series in Example 4. (Here the white noise is Gaussian; the same experiment with symmetric uniform noise had similar results, not shown.) The results of 2000 Monte Carlo replications are summarized in Fig. 3(a) and (b), with error bars at  $\pm 2$  standard errors. The results of the simulation indicate that the  $t$ -tests on the simulated data have an achieved significance level that is controlled at the nominal significance level of the test, within  $\pm 2$  standard errors, for all cases except when total number sampled  $np$  is large relative to  $N$ . As part of the same simulation, we also computed estimates of Type 1 error for the  $Z$ -test based on  $\sqrt{\nu - 1} \mathcal{R}_n^*$  and found that, the results were essentially identical to those summarized in Fig. 3(a) and (b). We have also repeated the entire study testing individual stocks of the DJIA vs white noise for independence, and the results were essentially the same as those reported in Fig. 3(a) and (b).

**Example 4 (Dow Jones Industrial Index).** The time series data for this example are the daily returns of the 30 stocks of the Dow Jones Industrial Index (DJIA), from August 30, 1993 through August 29, 2003. In this example, pairs of stocks are tested for independence. The 30 stocks correspond to the composition of the index on September 1, 2003. The DJIA closing prices data is provided by Daniyarov [3] with the source attributed to [www.nasdaq.com](http://www.nasdaq.com). Each series is identified by a stock ticker symbol; see the VaR package [3] for the corresponding list of company names. The time series analyzed are  $\log(\text{returns})$ ; that is, the sequence of differences of the logarithms of prices. A plot of  $\log(\text{returns})$  (not shown) suggests that the data approximately satisfy the strong stationarity condition.

The data has 2521 daily prices for each of the 30 stocks, so there are 2520 values of  $\log(\text{returns})$ . For this example we set  $p = 100$  and  $n = 50$  (see Proposition 3). Thus the  $t$  statistics ( $\mathcal{T}_n$ ) here are approximately standard normal under the independence hypothesis, and a statistic greater than 1.645 is significant at the 5% level, indicating dependence. Most of the pairs of stocks had significant  $t$  statistics in this example. The statistics for 20 pairs of stocks are shown in Table 2. Significant values of  $\mathcal{T}_n$  are in the upper tail of the corresponding Student  $t$  distribution; large positive values of  $\mathcal{T}_n$  are significant.

Tabular description of the complete set of 435 statistics is less informative than a graphical summary, so we have utilized a dendrogram to summarize the 435  $\mathcal{R}^*$  statistics in Fig. 4. To obtain this cluster dendrogram, we computed dissimilarities between pairs of stocks as  $1 - C' = 1 - (\mathcal{R}^* + 1)/2$ . Recall that under independence of stock returns,  $C'$  has a symmetric beta distribution with shape parameter  $(\nu - 1)/2$ . With this dissimilarity matrix, we applied hierarchical cluster analysis using complete linkage. In Fig. 4, the clusters that are merged at a lower height are more similar (dependent) than clusters



**Fig. 3.** Empirical Type 1 error rates: proportion of significant tests of independence on white noise series, using the same parameters as applied in the DJIA examples. In (a) the length (dimension) of subsequence  $p$  is on the horizontal axis while  $n = 50$  and  $N = 2520$  are fixed. In (b) the total length of the observed series  $N$  is on the horizontal axis while  $n = 50$  and  $p = 100$  are fixed.

that merge at greater height. One can observe, for example, that the financial stocks (AXP, C, JPM) cluster together, and the technology stocks (HPC, IBM, INTC, MSFT) also cluster together, but these two clusters are well separated and not merged until the second to last step when there are three clusters. One can also observe that, five manufacturers (AA, CAT, DD, GM, IP) cluster together, as do two retail industry (HD, WMT), two drug (JNJ, MRK), and two telecommunications stocks (SBC, T).

Our examples illustrate our theoretical results derived for high dimensional problems. The empirical results above demonstrate that the  $t$ -test of independence (or the corresponding  $Z$ -test of independence) can be applied to stationary time series using the methodology of Proposition 3.

## 5. Summary

The problem of testing independence between random vectors in high dimension is increasingly important as more and more applications arise that do not admit classical analysis. Distance based procedures have the advantage of applicability in arbitrarily high dimension, and distance correlation characterizes independence. In this paper, we addressed the issue that even though original dCov provides a consistent test of independence in arbitrary dimension, nevertheless, the sample dCor coefficient approaches 1 as dimension tends to infinity for fixed sample sizes under independence, making numerical interpretation of the dCor statistic difficult. We have proposed a modified distance correlation statistic and transformation of the modified high dimensional dCor to a  $t$  statistic.

We proved that under independence, the  $t$ -transformation of modified dCor converges in distribution to Student  $t$  as dimension of the random vectors approaches infinity, under standard conditions on the coordinates.<sup>3</sup> This deep theoretical result leads to a quite practical solution to the problem of testing and measuring dependence in high dimensions. Like the original dCor, our modified statistic for high dimension can be evaluated by an explicit computing formula in terms of averages of pairwise Euclidean distances. Our  $t$  (and  $Z$ ) limit theorems provide a test and scalar statistic that is easy to interpret in high dimensional problems, while retaining the good statistical properties of the distance covariance test.

An important application, testing independence of time series, is developed and implemented. This methodology was illustrated for simulated autoregressive time series, and the closing prices of stocks in the Dow Jones Industrials Index.

## Acknowledgments

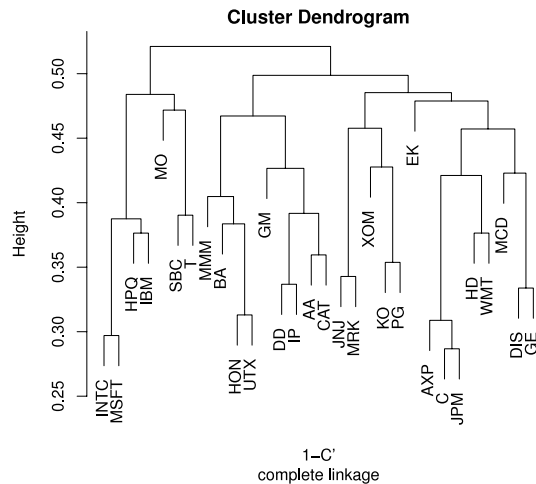
The authors would like to acknowledge the important contributions to this work of our colleague Nail K. Bakirov (Institute of Mathematics, USC Russian Academy of Sciences, Ufa, Russia), who died in May, 2010. The first named co-author is also grateful to his colleague, T. M. Móri, for his valuable advice.

This research was based on work supported by the National Science Foundation, while working at the Foundation.

<sup>3</sup> For example, exchangeability or strong stationarity with application of conditional CLT, discussed in Section 4, and Proposition 3.

**Table 2**  
Selected pairs of stocks of the Dow Jones Industrials Index with largest  $t$ -statistics indicating that  $\log(\text{returns})$  are highly dependent, are shown on the right. Examples of pairs of stocks in the index with non-significant  $\mathcal{T}_n$  statistics are shown on the left.

X	Y	$\mathcal{T}_n$	$p$ -value	X	Y	$\mathcal{T}_n$	$p$ -value
INTC	JNJ	-1.45	0.93	DD	IP	11.83	0.00
INTC	PG	-1.42	0.92	AA	HON	11.97	0.00
IBM	MCD	-0.42	0.66	DIS	GE	12.07	0.00
GM	T	-0.39	0.65	GE	JPM	13.42	0.00
EK	MO	-0.21	0.58	AXP	GE	13.51	0.00
IBM	IP	-0.00	0.50	HON	UTX	13.82	0.00
GM	JNJ	0.09	0.46	AXP	JPM	14.18	0.00
IBM	JNJ	0.19	0.42	AXP	C	14.21	0.00
MO	WMT	0.25	0.40	INTC	MSFT	15.22	0.00
DD	T	0.26	0.40	C	JPM	16.16	0.00



**Fig. 4.** Cluster dendrogram representing the dependence between daily returns of the 30 stocks of the Dow Jones Industrials Index, 1993–2003, as measured by the modified distance correlation statistic for independence.

**Appendix. Proofs of statements**

*A.1. On the bias of distance correlation*

In this section, we show that for important special cases including standard multivariate normal  $\mathbf{X}$  and  $\mathbf{Y}$ , distance correlation of vectors  $\mathbf{X}$ ,  $\mathbf{Y}$  can approach one as dimensions  $p$  and  $q$  tend to infinity even though  $\mathbf{X}$  and  $\mathbf{Y}$  are independent.

The following algebraic identity for the distance covariance statistic was established in [19].

$$v_n^2(\mathbf{X}, \mathbf{Y}) = \frac{1}{n^2} \sum_{i,j=1}^n A_{ij}B_{ij} = \frac{1}{n^2} \sum_{i,j=1}^n a_{ij}b_{ij} + \bar{a}\bar{b} - \frac{2}{n^3} \sum_{i,j,k=1}^n a_{ij}b_{ik}. \tag{A.1}$$

Hence,

$$\frac{v_n^2(\mathbf{X}, \mathbf{Y})}{\alpha\beta} = \frac{1}{n^2} \sum_{i,j=1}^n \frac{a_{ij}b_{ij}}{\alpha\beta} + \frac{1}{n^4} \sum_{i,j=1}^n \frac{a_{ij}}{\alpha} \sum_{k,\ell=1}^n \frac{b_{k\ell}}{\beta} - \frac{2}{n^3} \sum_{i,j,k=1}^n \frac{a_{ij}}{\alpha} \frac{b_{ik}}{\beta}, \tag{A.2}$$

and

$$\frac{v_n^2(\mathbf{X})}{\alpha^2} = \frac{1}{n^2} \sum_{i,j=1}^n \frac{a_{ij}^2}{\alpha^2} + \left( \frac{1}{n^2} \sum_{i,j=1}^n \frac{a_{ij}}{\alpha} \right)^2 - \frac{2}{n^3} \sum_{i,j,k=1}^n \frac{a_{ij}a_{ik}}{\alpha^2}. \tag{A.3}$$

Thus, in (A.2) and (A.3), each nonzero term  $a_{ij}$  or  $b_{ij}$ ,  $i \neq j$ , is divided by its expected value  $\alpha = E|\mathbf{X} - \mathbf{X}'|$  or  $\beta = E|\mathbf{Y} - \mathbf{Y}'|$ , respectively.

Suppose that,  $\mathbf{X} \in \mathbb{R}^p$  and  $\mathbf{Y} \in \mathbb{R}^q$  are independent, the coordinates of  $\mathbf{X}$  and  $\mathbf{Y}$  are iid, and the second moments of  $\mathbf{X}$  and  $\mathbf{Y}$  exist. Then  $X_{i1}, X_{i2}, \dots, X_{ip}$  are iid observations from the distribution of  $X_{i1}$ , and

$$\begin{aligned} \frac{1}{p} \sum_{t=1}^p (X_{1t} - X_{2t})^2 &= \frac{1}{p} \sum_{t=1}^p X_{1t}^2 - \bar{X}_1^2 + \frac{1}{p} \sum_{t=1}^p X_{2t}^2 - \bar{X}_2^2 - \frac{2}{p} \sum_{t=1}^p X_{1t}X_{2t} + 2\bar{X}_1\bar{X}_2 + \bar{X}_1^2 - 2\bar{X}_1\bar{X}_2 + \bar{X}_2^2 \\ &= \widehat{\text{Var}}(X_{11}) + \widehat{\text{Var}}(X_{21}) - 2\widehat{\text{Cov}}(X_{11}, X_{21}) + (\bar{X}_1 - \bar{X}_2)^2, \end{aligned}$$

where

$$\bar{X}_i = \frac{1}{p} \sum_{t=1}^p X_{it}, \quad \widehat{\text{Var}}(X_{i1}) = \frac{1}{p} \sum_{t=1}^p X_{it}^2 - \bar{X}_i^2, \quad i = 1, 2,$$

$$\text{and } \widehat{\text{Cov}}(X_{11}, X_{21}) = \frac{1}{p} \sum_{t=1}^p (X_{1t}X_{2t}) - \bar{X}_1\bar{X}_2.$$

Hence  $\frac{1}{p} \sum_{t=1}^p (X_{1t} - X_{2t})^2$  converges to  $2\theta^2$  as  $p$  tends to infinity, where  $\theta^2 = \text{Var}(X_{11}) < \infty$ . It follows that  $a_{12}/\sqrt{p}$  and  $\alpha/\sqrt{p}$  each converges almost surely to  $\sqrt{2}\theta$  as  $p$  tends to infinity, and  $\lim_{p \rightarrow \infty} a_{ij}/\alpha = 1, i \neq j$ . Similarly for  $i \neq j, \lim_{q \rightarrow \infty} b_{ij}/\sqrt{q} = \sqrt{2}\zeta$ , where  $\zeta^2 = \text{Var}(Y_{11}) < \infty$ , and  $\lim_{q \rightarrow \infty} b_{ij}/\beta = 1$  with probability one. Therefore

$$\begin{aligned} \lim_{p, q \rightarrow \infty} \frac{1}{n^2} \sum_{i, j=1}^n \frac{a_{ij}b_{ij}}{\alpha\beta} &= \frac{n-1}{n}, \\ \lim_{p, q \rightarrow \infty} \left( \frac{1}{n^4} \sum_{i, j=1}^n \frac{a_{ij}}{\alpha} \sum_{k, \ell=1}^n \frac{b_{k\ell}}{\beta} \right) &= \frac{(n-1)^2}{n^2}, \end{aligned}$$

and

$$\lim_{p, q \rightarrow \infty} \frac{2}{n^3} \sum_{i, j, k=1}^n \frac{a_{ij}b_{ik}}{\alpha\beta} = \frac{2n(n-1)}{n^3} + \frac{2n(n-1)(n-2)}{n^3} = \frac{2(n-1)^2}{n^2}.$$

Substituting these limits in (A.2) and simplifying yields

$$\lim_{p, q \rightarrow \infty} \frac{\mathcal{V}_n^2(\mathbf{X}, \mathbf{Y})}{\alpha\beta} = \frac{n-1}{n^2}.$$

By similar steps, substituting limits in (A.3) and simplifying, we obtain

$$\lim_{p \rightarrow \infty} \frac{\mathcal{V}_n^2(\mathbf{X})}{\alpha^2} = \frac{n-1}{n^2}, \quad \lim_{q \rightarrow \infty} \frac{\mathcal{V}_n^2(\mathbf{Y})}{\beta^2} = \frac{n-1}{n^2}.$$

Hence for this class of independent random vectors each of the statistics

$$\frac{\mathcal{V}_n^2(\mathbf{X}, \mathbf{Y})}{\alpha\beta}, \quad \frac{\mathcal{V}_n^2(\mathbf{X})}{\alpha^2}, \quad \frac{\mathcal{V}_n^2(\mathbf{Y})}{\beta^2}$$

converges almost surely to  $(n-1)/n^2$  as dimensions  $p, q$  tend to infinity, and consequently for each fixed  $n$  the distance correlation  $\mathcal{R}(\mathbf{X}, \mathbf{Y})$  has limit one as  $p, q$  tend to infinity.

### A.2. Proof of Lemma 1

**Proof.** (i) Observe that

$$W_n(\mathbf{X}, \mathbf{X}) \geq \frac{1}{n^2} \sum_{i, j=1}^n |\mathbf{X}_i - \mathbf{X}_j|^2 - \bar{a}^2 \geq 0,$$

where the last inequality is the Cauchy–Bunyakovski inequality (also known as the Cauchy–Schwartz inequality). See the end of the proof of Lemma 3 for a proof of statement (ii). Statement (iii) follows from the Law of Large Numbers (LLN) for  $U$ -statistics, and (iv) follows from (iii) under independence.  $\square$

## A.3. Proof of Lemma 2

Lemma 2 establishes the identity:

$$n^2 \mathcal{U}_n(\mathbf{X}, \mathbf{Y}) = \frac{(n-1)^2}{n^2} \cdot \mathcal{U}_n^*(\mathbf{X}, \mathbf{Y}). \quad (\text{A.4})$$

**Proof.** Denote

$$u = \frac{2}{n-2} \sum_{i=1}^n (\bar{a}_i - \bar{a})(\bar{b}_i - \bar{b}).$$

Then

$$\begin{aligned} \frac{(n-1)^2}{n^2} \mathcal{U}_n^*(\mathbf{X}, \mathbf{Y}) &= \sum_{i \neq j} \left( \frac{n-1}{n} \right)^2 A_{i,j}^* B_{i,j}^* - \frac{2}{n-2} \sum_{i=1}^n \left( \frac{n-1}{n} \right)^2 A_{i,i}^* B_{i,i}^* \\ &= \sum_{i \neq j} \left( A_{i,j} - \frac{1}{n} a_{ij} \right) \left( B_{i,j} - \frac{1}{n} b_{ij} \right) - \frac{2}{n-2} \sum_{i=1}^n (\bar{a}_i - \bar{a})(\bar{b}_i - \bar{b}) \\ &= \sum_{i,j=1}^n A_{i,j} B_{i,j} - \sum_{i=1}^n A_{i,i} B_{i,i} + \frac{1}{n^2} \sum_{i,j=1}^n a_{ij} b_{ij} - \frac{1}{n} \sum_{i,j=1}^n A_{i,j} b_{ij} - \frac{1}{n} \sum_{i,j=1}^n a_{ij} B_{i,j} - u. \end{aligned} \quad (\text{A.5})$$

Now

$$-A_{i,i} = 2\bar{a}_i - \bar{a} = 2(\bar{a}_i - \bar{a}) + \bar{a}, \quad -B_{i,i} = 2\bar{b}_i - \bar{b} = 2(\bar{b}_i - \bar{b}) + \bar{b},$$

thus

$$\sum_{i=1}^n A_{i,i} B_{i,i} = n\bar{a}\bar{b} + 4 \sum_{i=1}^n (\bar{a}_i - \bar{a})(\bar{b}_i - \bar{b}) = n\bar{a}\bar{b} + 2(n-2)u.$$

On the other hand,  $\sum_{i,j=1}^n A_{i,j} = 0$ , so

$$\begin{aligned} \frac{1}{n} \sum_{i,j=1}^n A_{i,j} b_{ij} &= \frac{1}{n} \sum_{i,j=1}^n A_{i,j} (b_{ij} - \bar{b}) \\ &= \frac{1}{n} \sum_{i,j=1}^n \{ [a_{ij} - \bar{a}] - [\bar{a}_i - \bar{a}] - [\bar{a}_j - \bar{a}] \} (b_{ij} - \bar{b}) \\ &= \frac{1}{n} \sum_{i,j=1}^n (a_{ij} - \bar{a})(b_{ij} - \bar{b}) - 2 \sum_{i=1}^n (\bar{a}_i - \bar{a})(\bar{b}_i - \bar{b}) \\ &= \frac{1}{n} \sum_{i,j=1}^n a_{ij} b_{ij} - n\bar{a}\bar{b} - (n-2)u, \end{aligned}$$

and similarly,

$$\frac{1}{n} \sum_{i,j=1}^n B_{i,j} a_{ij} = \frac{1}{n} \sum_{i,j=1}^n a_{ij} b_{ij} - n\bar{a}\bar{b} - (n-2)u.$$

Thus the right hand side (A.5) equals

$$\begin{aligned} \sum_{i,j=1}^n A_{i,j} B_{i,j} - (n\bar{a}\bar{b} + 2(n-2)u) + \frac{1}{n^2} \sum_{i,j=1}^n a_{ij} b_{ij} - 2 \left[ \frac{1}{n} \sum_{i,j=1}^n a_{ij} b_{ij} - n\bar{a}\bar{b} - (n-2)u \right] - u \\ = \sum_{i,j=1}^n A_{i,j} B_{i,j} - \left[ \frac{2n-1}{n^2} \sum_{i,j=1}^n a_{ij} b_{ij} - n\bar{a}\bar{b} - u \right] \\ = n^2 \mathcal{V}_n^2(\mathbf{X}, \mathbf{Y}) - n \mathcal{W}_n(\mathbf{X}, \mathbf{Y}) = n^2 \mathcal{U}_n. \quad \square \end{aligned}$$

A.4. Proof of Lemma 3

For the remaining proofs, we introduce the following notation.

For  $p$ -dimensional samples  $\mathbf{X}_i = (X_{i1}, \dots, X_{ip})$ ,  $i = 1, \dots, n$ , define

$$\begin{aligned} \tau &= 2E|\mathbf{X}_i|^2 = 2 \sum_{k=1}^p E(X_{ik}^2), \\ T_i(p) &= \frac{|\mathbf{X}_i|^2 - E|\mathbf{X}_i|^2}{2\sqrt{\tau}} = \frac{1}{2\sqrt{\tau}} \sum_{k=1}^p [(X_{ik})^2 - E(X_{ik}^2)], \quad i = 1, \dots, n, \\ C_{i,j}(p) &= \begin{cases} \frac{\langle \mathbf{X}_i, \mathbf{X}_j \rangle}{\sqrt{\tau}} = \frac{1}{\sqrt{\tau}} \sum_{k=1}^p X_{ik}X_{jk}, & i \neq j; \\ 0, & i = j, \end{cases} \end{aligned}$$

where  $\langle \mathbf{X}_i, \mathbf{X}_j \rangle$  is the dot product in  $\mathbb{R}^p$ . We denote the weak limits, as  $p \rightarrow \infty$  of  $T_i(p)$  and  $C_{i,j}(p)$ , by  $T_i$  and  $C_{i,j}$ , respectively. Note that, the weak limits  $C_{i,j}$  are Gaussian.

**Proof.** Lemma 3 (i) asserts that if  $E|\mathbf{X}_i|^2 < \infty$ , then there exist  $\Omega_{i,j}$  and  $\sigma_X > 0$  such that for a fixed  $n$

$$\mathcal{U}_n^*(\mathbf{X}, \mathbf{X}) \xrightarrow{p \rightarrow \infty} \sum_{i \neq j} \Omega_{i,j}^2 \stackrel{\mathcal{D}}{=} 2\sigma_X^2 \chi_\nu^2,$$

where  $\nu = \frac{n(n-3)}{2}$  and  $\chi_\nu^2$  denotes the distribution of a chisquare random variable with  $\nu$  degrees of freedom.

Observe that, in distribution, the Taylor expansion of the square root implies that we have the limit

$$S_{i,j}(p) = |\mathbf{X}_i - \mathbf{X}_j| - E|\mathbf{X}_i - \mathbf{X}_j| \xrightarrow{p \rightarrow \infty} S_{i,j} = T_i + T_j - C_{i,j}. \tag{A.6}$$

To see this, observe that by Taylor’s Theorem we have

$$\begin{aligned} \sqrt{|\mathbf{X}_i - \mathbf{X}_j|^2} &= \sqrt{\tau} \sqrt{1 + \frac{|\mathbf{X}_i - \mathbf{X}_j|^2}{\tau} - 1} \\ &= \sqrt{\tau} \left( 1 + \frac{1}{2} \left[ \frac{|\mathbf{X}_i - \mathbf{X}_j|^2}{\tau} - 1 \right] + o_p(1) \right), \end{aligned}$$

and

$$\begin{aligned} E\sqrt{|\mathbf{X}_i - \mathbf{X}_j|^2} &= \sqrt{\tau} E \sqrt{1 + \frac{|\mathbf{X}_i - \mathbf{X}_j|^2}{\tau} - 1} \\ &= \sqrt{\tau} \left( 1 + \frac{1}{2} \left[ \frac{E|\mathbf{X}_i - \mathbf{X}_j|^2}{\tau} - 1 \right] + o_p(1) \right) \end{aligned}$$

Thus

$$\frac{|\mathbf{X}_i - \mathbf{X}_j| - E|\mathbf{X}_i - \mathbf{X}_j|}{\sqrt{\tau}} = \frac{|\mathbf{X}_i - \mathbf{X}_j|^2 - E|\mathbf{X}_i - \mathbf{X}_j|^2}{2\tau} + o_p(1).$$

Hence (A.6) is true.

Using (A.6) it can be shown that (in the limit, as  $p \rightarrow \infty$ )

$$\mathcal{U}_n^*(\mathbf{X}, \mathbf{X}) = \sum_{i \neq j} [C_{i,j} - \bar{C}_i - \bar{C}_j + \bar{C}]^2 + \lambda \sum_{i=1}^n (\bar{C}_i - \bar{C})^2, \tag{A.7}$$

where  $\lambda = -2/(n - 2)$ , and

$$\bar{C}_i = \frac{1}{n-1} \sum_{j=1}^n C_{i,j}, \quad \bar{C} = \frac{1}{n^2-n} \sum_{i,j=1}^n C_{i,j} = \frac{1}{n} \sum_{i=1}^n \bar{C}_i.$$

(The identity (A.7) is derived separately below in Appendix A.5.) For finite  $p$ , replace  $C$  by  $S(p)$ , with corresponding subscripts and bars, throughout.

Let us prove now that (A.7) is non-negative (this also completes the proof of Lemma 1). Indeed, for any constant  $\gamma$

$$\begin{aligned} Q(\gamma) &\stackrel{\text{def}}{=} \sum_{i \neq j} [(C_{i,j} - \bar{C}) - \gamma(\bar{C}_i - \bar{C}) - \gamma(\bar{C}_j - \bar{C})]^2 \\ &= \sum_{i \neq j} [(C_{i,j} - \bar{C})^2 + \gamma^2(\bar{C}_i - \bar{C})^2 + \gamma^2(\bar{C}_j - \bar{C})^2 \\ &\quad - 2\gamma(C_{i,j} - \bar{C})(\bar{C}_i - \bar{C}) - 2\gamma(C_{i,j} - \bar{C})(\bar{C}_j - \bar{C}) + 2\gamma^2(\bar{C}_i - \bar{C})(\bar{C}_j - \bar{C})] \\ &= \sum_{i \neq j} (C_{i,j} - \bar{C})^2 - [(n-1)(4\gamma - 2\gamma^2) + 2\gamma^2] \sum_{i=1}^n (\bar{C}_i - \bar{C})^2. \end{aligned}$$

So the right hand side in (A.7) equals

$$\begin{aligned} Q(1) + \lambda \sum_{i=1}^n (\bar{C}_i - \bar{C})^2 &= \sum_{i \neq j} (C_{i,j} - \bar{C})^2 - \left[2n + \frac{2}{n-2}\right] \sum_{i=1}^n (\bar{C}_i - \bar{C})^2 \\ &= Q(\gamma) \geq 0, \quad \text{for } \gamma = \frac{n-1}{n-2}. \end{aligned} \tag{A.8}$$

We have proved that  $\mathcal{U}_n^*(\mathbf{X}, \mathbf{X}) \geq 0$ . To complete the proof of Lemma 3, we need the following Cochran decomposition [2] type lemma. Let  $\chi_n^2$  be a chi-square random variable with  $n$  degrees of freedom.

**Lemma 4.** Let  $Z$  be a Gaussian random vector with zero mean and

$$Q = Q_1 + Q_2, \tag{A.9}$$

where  $Q, Q_1,$  and  $Q_2$  are non-negative quadratic forms of the coordinates of  $Z,$  and

- i.  $Q \stackrel{d}{=} \chi_n^2;$
- ii.  $Q_1 \stackrel{d}{=} \chi_m^2, m \leq n.$

Then  $Q_2$  is independent of  $Q_1,$  and  $Q_2 \stackrel{d}{=} \chi_{n-m}^2$ .

For a proof apply e.g. Rao [13, 3b.4, pp. 185–187]: statement (i) (Fisher–Cochran Theorem) and statement (iv). Set  $Q_2$  equal to the right hand side of (A.8), let

$$Q_1 = \left[2n + \frac{2}{n-2}\right] \sum_{i=1}^n (\bar{C}_i - \bar{C})^2,$$

and

$$Q = \sum_{i \neq j} (C_{i,j} - \bar{C})^2.$$

We proved that  $Q = Q_1 + Q_2.$  The matrix  $\|C_{i,j}\|_{i,j=1}^n$  is symmetric with  $C_{i,i} = 0,$  for all  $i$  and

$$Q = 2 \sum_{i < j} (C_{i,j} - \bar{C})^2,$$

with

$$\bar{C} = \frac{1}{n^2 - n} \sum_{i \neq j} C_{i,j} = \frac{2}{n^2 - n} \sum_{i < j} C_{i,j}.$$

Thus the quadratic form  $Q$  has rank  $(n^2 - n)/2 - 1.$

From classical statistics we know that, if  $X_1, X_2, \dots, X_N$  are iid Normal(0,  $\sigma_X^2$ ), then for  $\bar{X} = \frac{1}{N} \sum_{k=1}^N X_k$

$$\sum_{k=1}^N (X_k - \bar{X})^2 \stackrel{d}{=} \sigma_X^2 \chi_{N-1}^2.$$

Therefore, for  $N = (n^2 - n)/2,$

$$Q \stackrel{d}{=} 2\sigma_X^2 \chi_{N-1}^2,$$

where  $\sigma_X^2 = E(C_{1,2})^2.$



Consider now the quadratic form

$$\frac{n-2}{2(n-1)^2} \cdot Q_1 = \sum_{i=1}^n (\bar{C}_i - \bar{C})^2, \quad (\text{A.10})$$

whose rank is  $n-1$  because there is one single linear relationship between the vectors  $\bar{C}_i - \bar{C}$ ,  $i = 1, 2, \dots, n$ . Here the quadratic form (A.10) is the square of the Euclidean norm of the vector

$$(\bar{C}_1 - \bar{C}, \bar{C}_2 - \bar{C}, \dots, \bar{C}_n - \bar{C})$$

with covariance matrix

$$\Sigma = \begin{pmatrix} d & c & \dots & c \\ c & d & \dots & c \\ \vdots & \vdots & \ddots & \vdots \\ c & c & \dots & d \end{pmatrix}.$$

The moments of  $\bar{C}_i$  and  $\bar{C}$  are

$$E[(\bar{C}_i)^2] = \frac{\sigma_X^2}{n-1}, \quad E[\bar{C}_i \bar{C}_j] = \frac{\sigma_X^2}{(n-1)^2},$$

$$E[\bar{C}_i \bar{C}] = \frac{1}{n} \sum_{j=1}^n E[\bar{C}_i \bar{C}_j] = \frac{2\sigma_X^2}{n(n-1)}, \quad E[\bar{C}^2] = \frac{1}{n} \sum_{j=1}^n E[\bar{C}_i \bar{C}] = \frac{2\sigma_X^2}{n(n-1)},$$

where  $\sigma_X^2 = E(C_{1,2})^2$ . Therefore

$$d = E(\bar{C}_1 - \bar{C})^2 = E(\bar{C}_1)^2 - 2E(\bar{C}_1 \bar{C}) + E(\bar{C})^2 = \frac{\sigma_X^2(n-2)}{n(n-1)},$$

$$c = E(\bar{C}_1 - \bar{C})(\bar{C}_2 - \bar{C}) = -\frac{\sigma_X^2(n-2)}{n(n-1)^2},$$

and  $d + (n-1)c = 0$ . The matrix  $\Sigma$  has the characteristic polynomial

$$f(\lambda) = \det(\Sigma - \lambda I) = (d - \lambda + (n-1)c)(d - \lambda - c)^{n-1} = -\lambda(d - c - \lambda)^{n-1},$$

so one eigenvalue of  $\Sigma$  equals 0 and all other eigenvalues equal

$$d - c = \frac{\sigma_X^2(n-2)}{(n-1)^2}.$$

Therefore

$$Q_1 \stackrel{d}{=} \frac{2(n-1)^2}{n-2} (d-c) \chi_{n-1}^2 = 2\sigma_X^2 \chi_{n-1}^2.$$

Applying Lemma 4 we obtain

$$Q_2 \stackrel{d}{=} 2\sigma_X^2 \chi_\nu^2$$

where  $\nu = \frac{n^2-n}{2} - n = \frac{n(n-3)}{2}$ . Set  $m = (n-1)/(n-2)$  and

$$\Omega_{i,j} = C_{i,j} - \bar{C} - m(\bar{C}_i - \bar{C}) - m(\bar{C}_j - \bar{C}). \quad (\text{A.11})$$

Then

$$\mathcal{U}_n^*(\mathbf{X}, \mathbf{X}) \xrightarrow{p \rightarrow \infty} \sum_{i \neq j} \Omega_{i,j}^2 \stackrel{d}{=} 2\sigma_X^2 \chi_\nu^2.$$

This completes the proof of Lemma 3(i).

Similarly, for the second sample, let

$$\tau_1 = 2E|\mathbf{Y}_i|^2 = 2 \sum_{k=1}^q E(Y_{ik})^2$$

and define

$$D_{i,j}(q) = \begin{cases} \frac{\langle \mathbf{Y}_i, \mathbf{Y}_j \rangle}{\sqrt{\tau_1}} = \frac{1}{\sqrt{\tau_1}} \sum_{k=1}^q Y_{ik} Y_{jk}, & i \neq j; \\ 0, & i = j. \end{cases}$$

Denote the weak limits of  $D_{i,j}(q)$  as  $q \rightarrow \infty$  by  $D_{i,j}$ ,  $\sigma_Y^2 = E(D_{1,2})^2$ , and

$$\bar{D}_i = \frac{1}{n-1} \sum_{j=1}^n D_{ij}, \quad \bar{D} = \frac{1}{n} \sum_{i=1}^n \bar{D}_i.$$

If

$$\Psi_{i,j} = D_{i,j} - \bar{D} - m(\bar{D}_i - \bar{D}) - m(\bar{D}_j - \bar{D}), \tag{A.12}$$

then, by similar arguments as in the proof of Lemma 3 (i), we obtain Lemma 3 (ii) and (iii):

$$\mathcal{U}_n^*(\mathbf{Y}, \mathbf{Y}) \xrightarrow{q \rightarrow \infty} \sum_{i \neq j} \Psi_{i,j}^2 \stackrel{D}{=} 2\sigma_Y^2 \chi_v^2, \quad \mathcal{U}_n^*(\mathbf{X}, \mathbf{Y}) \xrightarrow{p, q \rightarrow \infty} \sum_{i \neq j} \Omega_{i,j} \Psi_{i,j},$$

where  $\{\Omega_{i,j}\}$  are defined by (A.11) and  $\{\Psi_{i,j}\}$  are defined by (A.12).  $\square$

A.5. Proof of identity (A.7) in the proof of Lemma 3

We have from (A.6) that  $S_{i,j} = T_i + T_j - C_{i,j}$ . Notice that  $S_{i,i} = 0$ ,  $S_{i,j} = S_{j,i}$ , and  $E[S_{i,j}] = 0$ , for all  $i, j$ . Denote

$$\bar{S}_i = \frac{1}{n-1} \sum_{j=1}^n S_{i,j}, \quad \bar{S} = \frac{1}{n^2-n} \sum_{i,j=1}^n S_{i,j} = \frac{1}{n} \sum_{i=1}^n \bar{S}_i, \quad \bar{T} = \frac{1}{n} \sum_{i=1}^n T_i.$$

Using (A.6) rewrite

$$\begin{aligned} \bar{S}_i &= \frac{1}{n-1} \sum_{j \neq i} (T_i + T_j - C_{i,j}) = \frac{(n-1)T_i + n\bar{T} - T_i}{n-1} - \bar{C}_i \\ &= \frac{n\bar{T}}{n-1} + \frac{n-2}{n-1} T_i - \bar{C}_i, \\ \bar{S} &= \frac{1}{n^2-n} \sum_{i \neq j} (T_i + T_j - C_{i,j}) = \frac{1}{n^2-n} \sum_{i \neq j} (T_i + T_j) - \bar{C} \\ &= \frac{2n^2\bar{T} - 2n\bar{T}}{n^2-n} - \bar{C} = 2\bar{T} - \bar{C}. \end{aligned}$$

Since  $a_{ij} = S_{ij} + \alpha$ , we have

$$A_{i,j}^* = \begin{cases} a_{ij} - \frac{n}{n-1} (\bar{a}_i - \bar{a}_j + \bar{a}) = S_{i,j} - \bar{S}_i - \bar{S}_j + \bar{S}, & i \neq j; \\ \frac{n}{n-1} (\bar{a}_i - \alpha - (\bar{a} - \alpha)) = \bar{S}_i - \bar{S}, & i = j. \end{cases}$$

Hence, for  $i \neq j$ ,

$$\begin{aligned} A_{i,j}^* &= S_{i,j} - \bar{S}_i - \bar{S}_j + \bar{S} \\ &= T_i + T_j - C_{i,j} - \left( \frac{n\bar{T}}{n-1} + \frac{n-2}{n-1} T_i - \bar{C}_i \right) - \left( \frac{n\bar{T}}{n-1} + \frac{n-2}{n-1} T_j - \bar{C}_j \right) + 2\bar{T} - \bar{C} \\ &= \frac{T_i - \bar{T}}{n-1} + \frac{T_j - \bar{T}}{n-1} - [C_{i,j} - \bar{C}_i - \bar{C}_j + \bar{C}], \end{aligned}$$

with

$$\sum_{j \neq i} [C_{i,j} - \bar{C}_i - \bar{C}_j + \bar{C}] = \bar{C}_i - \bar{C},$$

and for  $i = j$  we have

$$\begin{aligned} A_{i,i}^* &= \bar{S}_i - \bar{S} = 2\bar{T} - \bar{C} - \left( \frac{n\bar{T}}{n-1} + \frac{n-2}{n-1}T_i - \bar{C}_i \right) \\ &= \bar{C}_i - \bar{C} - \frac{n-2}{n-1}(T_i - \bar{T}). \end{aligned}$$

Therefore, setting  $\lambda = -2/(n-2)$  we obtain

$$\begin{aligned} \mathcal{U}_n^*(\mathbf{X}, \mathbf{X}) &= \sum_{i \neq j} A_{i,j}^{*2} + \lambda \sum_{i=1}^n A_{i,i}^{*2} \\ &= \sum_{i \neq j} [C_{i,j} - \bar{C}_i - \bar{C}_j + \bar{C}]^2 + \lambda \sum_{i=1}^n (\bar{C}_i - \bar{C})^2 - \left[ \left( \frac{4}{n-1} + 2\lambda \times \frac{n-2}{n-1} \right) \sum_{i=1}^n (\bar{C}_i - \bar{C})(T_i - \bar{T}) \right] \\ &\quad + \left( 2 \times \frac{(n-2)}{(n-1)^2} + \lambda \left( \frac{n-2}{n-1} \right)^2 \right) \sum_{i=1}^n (T_i - \bar{T})^2 \\ &= \sum_{i \neq j} [C_{i,j} - \bar{C}_i - \bar{C}_j + \bar{C}]^2 + \lambda \sum_{i=1}^n (\bar{C}_i - \bar{C})^2. \quad \square \end{aligned}$$

#### A.6. Proof of Proposition 1

**Proof.** (i): The identity

$$\mathcal{V}^2(\mathbf{X}, \mathbf{Y}) = E[|\mathbf{X} - \mathbf{X}'||\mathbf{Y} - \mathbf{Y}'|] + E|\mathbf{X} - \mathbf{X}'|E|\mathbf{Y} - \mathbf{Y}'| - 2E[|\mathbf{X} - \mathbf{X}'||\mathbf{Y} - \mathbf{Y}''|]$$

is obtained in [19, Remark 3] by applying [19, Lemma 1] and Fubini's theorem.

(ii): The result is obtained by evaluating the expected value of  $\mathcal{V}_n^2(\mathbf{X}, \mathbf{Y})$  using the equivalent computing formula in identity (A.1). Under independence, we have

$$E \left[ \frac{1}{n^2} \sum_{i,j=1}^n a_{ij}b_{ij} \right] = \frac{n-1}{n} \alpha\beta, \quad E[\bar{a}\bar{b}] = \left( \frac{n-1}{n} \right)^2 \alpha\beta.$$

Similarly, evaluate the expected values of each term in (A.1) under independence and simplify. The resulting expression contains the term  $\mathcal{V}^2(\mathbf{X}, \mathbf{Y})$  (applying (i)), which is zero under independence, and the result follows.

(iii): Again, using identity (A.1) we evaluate the expectation of  $n\mathcal{U}_n(\mathbf{X}, \mathbf{Y}) = n\mathcal{V}_n^2(\mathbf{X}, \mathbf{Y}) - \mathcal{W}_n(\mathbf{X}, \mathbf{Y})$ , first combining the terms of  $n\mathcal{V}_n^2(\mathbf{X}, \mathbf{Y})$  and  $\mathcal{W}_n(\mathbf{X}, \mathbf{Y})$  that involve  $a_{ij}b_{ij}$  and  $\bar{a}\bar{b}$ . Under independence, the linear combination of expected values in the resulting expression sum to zero. Now, from Lemma 2 we have that  $n^2\mathcal{U}_n(\mathbf{X}, \mathbf{Y}) = \frac{(n-1)^2}{n^2} \cdot \mathcal{U}_n^*(\mathbf{X}, \mathbf{Y})$ , where  $\mathcal{U}_n(\mathbf{X}, \mathbf{Y}) = \mathcal{V}_n^2(\mathbf{X}, \mathbf{Y}) - \mathcal{W}_n(\mathbf{X}, \mathbf{Y})/n$ . It follows that  $E[\mathcal{U}_n^*(\mathbf{X}, \mathbf{Y})] = 0$  under independence of  $\mathbf{X}$  and  $\mathbf{Y}$ , and therefore  $E[\mathcal{V}_n^2(\mathbf{X}, \mathbf{Y})] = 0$  for independent  $\mathbf{X}, \mathbf{Y}$ .  $\square$

#### A.7. Proof of Proposition 2

As in the proof of Proposition 1(ii), the expected values of  $\mathcal{V}_n^2(\mathbf{X}, \mathbf{Y})$  and  $\mathcal{W}_n(\mathbf{X}, \mathbf{Y})$  are obtained by applying identity (A.1), expanding the sums and products in the statistics, and combining the terms that have equal expected values. The expected values of  $\mathcal{U}_n(\mathbf{X}, \mathbf{Y})$ ,  $\mathcal{U}_n^*(\mathbf{X}, \mathbf{Y})$ , and  $\mathcal{V}_n^*(\mathbf{X}, \mathbf{Y})$  then follow by definition.

#### A.8. Proof of Theorem 1

Introduce

$$Z_{i,j} = \frac{\Psi_{i,j}}{\sqrt{\sum_{i \neq j} \Psi_{i,j}^2}}, \quad \sum_{i \neq j} Z_{i,j}^2 = 1.$$

Under the independence hypothesis the random variables  $\{Z_{i,j}\}$  do not depend on  $\{\Omega_{i,j}\}$ , and for  $\vartheta = \sum_{i \neq j} \Omega_{i,j}Z_{i,j}$  we have

$$\sum_{i \neq j} \Omega_{i,j}^2 - \vartheta^2 = \sum_{i \neq j} (\Omega_{i,j} - Z_{i,j}\vartheta)^2.$$

Now  $\text{Rank}(\vartheta^2) = 1$  and

$$\text{Rank} \left( \sum_{i \neq j} (\Omega_{i,j} - Z_{i,j}\vartheta)^2 \right) = \text{Rank} \left( \sum_{i \neq j} \Omega_{i,j}^2 \right) - 1,$$

so we found one more linear relationship:

$$\sum_{i \neq j} Z_{i,j}(\Omega_{i,j} - Z_{i,j}\vartheta) = 0.$$

By Cochran's theorem,  $\vartheta^2 \stackrel{D}{=} 2\sigma_X^2 \chi_1^2$ , and

$$\sum_{i \neq j} (\Omega_{i,j} - Z_{i,j}\vartheta)^2 \stackrel{D}{=} 2\sigma_X^2 \chi_{\nu-1}^2,$$

which does not depend on  $\vartheta$ .

For any fixed  $\{Z_{i,j}\}$ ,

$$\begin{aligned} P \left\{ \mathcal{T}_n < x \mid \{Z_{i,j}\}_{i,j=1}^n \right\} &= P \left\{ \frac{\vartheta}{\sqrt{\frac{1}{\nu-1} \sum_{i \neq j} (\Omega_{i,j} - Z_{i,j}\vartheta)^2}} < x \mid \{Z_{i,j}\}_{i,j=1}^n \right\} \\ &= P\{t_{\nu-1} < x\}, \end{aligned}$$

where the random variable  $t_{\nu-1}$  has the Student's distribution with  $\nu - 1$  degrees of freedom. Therefore

$$P\{\mathcal{T}_n < x\} = E \left[ P \left\{ \mathcal{T}_n < x \mid \{Z_{i,j}\}_{i,j=1}^n \right\} \right] = P\{t_{\nu-1} < x\}.$$

This proves statement (i) of [Theorem 1](#).

Statement (ii) follows by first observing that  $E[\mathcal{U}_n^*(\mathbf{X}, \mathbf{Y})] = 0$  under independence, and  $E[\mathcal{U}_n^*(\mathbf{X}, \mathbf{Y})] > 0$  under a dependent alternative (see [Proposition 2](#)). The test criterion is to reject the null hypothesis at level  $\alpha$  if  $\mathcal{T}_n > c_\alpha$ , where  $P(\mathcal{T}_n > c_\alpha) = \alpha$  under the null. If  $\mathcal{T}_n > 0$  it is equivalent to reject  $H_0$  if

$$(\mathcal{R}_n^*)^2 > \frac{c_\alpha^2}{\nu - 1 + c_\alpha^2}.$$

Since  $\mathcal{V}_n^*(\mathbf{X})\mathcal{V}_n^*(\mathbf{Y})$  is the same under the null or alternative hypothesis, an equivalent criterion is to reject  $H_0$  if

$$\mathcal{V}_n^*(\mathbf{X}, \mathbf{Y}) > \sqrt{\frac{c_\alpha^2}{\nu - 1 + c_\alpha^2} \mathcal{V}_n^*(\mathbf{X})\mathcal{V}_n^*(\mathbf{Y})}, \tag{A.13}$$

or equivalently if

$$\mathcal{U}_n^*(\mathbf{X}, \mathbf{Y}) > \sqrt{\frac{c_\alpha^2}{\nu - 1 + c_\alpha^2} \mathcal{U}_n^*(\mathbf{X}) \mathcal{U}_n^*(\mathbf{Y})}, \tag{A.14}$$

where  $E[\mathcal{U}_n^*(\mathbf{X}, \mathbf{Y})] = n(n - 3)\mathcal{V}^2(\mathbf{X}, \mathbf{Y})$ .

Now following the proof of part (i) one can show that in (3.6) of [Lemma 3](#)(iii)

$$\mathcal{U}_n^*(\mathbf{X}, \mathbf{Y}) \xrightarrow{p,q \rightarrow \infty} \sum_{i \neq j} \Omega_{i,j} \Psi_{i,j},$$

the random variables  $(\Omega_{i,j}, \Psi_{i,j})$  are bivariate normal, have zero expected value, and under the alternative hypothesis their correlation is positive because we have just shown that  $E[\mathcal{U}_n^*(\mathbf{X}, \mathbf{Y})]$  is a positive constant multiple of  $\mathcal{V}^2(\mathbf{X}, \mathbf{Y})$ .

All we need to show is that if  $(\Omega, \Psi)$  are bivariate normal with zero expected value and correlation  $\rho > 0$ , then  $P(\Omega\Psi > c)$  is a monotone increasing function of  $\rho$ . We can assume that, the variances of  $\Omega$  and  $\Psi$  are equal to 1. To see the monotonicity, notice that if  $a^2 + b^2 = 1, 2ab = \rho$  and  $X, Y$  are iid standard normal random variables, then for  $U = aX + bY, V = bX + aY$  we have  $\text{Var}(U) = \text{Var}(V) = 1$ , and the covariance of  $U$  and  $V$  is  $E(UV) = 2ab = \rho$ . Thus  $(U, V)$  has the same distribution as  $(\Omega, \Psi)$ , and

$$UV = ab(X^2 + Y^2) + (a^2 + b^2)XY = \rho \frac{X^2 + Y^2}{2} + XY.$$

Thus  $P(UV > c)$  is always a monotone increasing function of  $\rho$ . This proves the unbiasedness of our  $t$ -test of independence.  $\square$

## References

- [1] N.K. Bakirov, M.L. Rizzo, G.J. Székely, A multivariate nonparametric test of independence, *J. Multivariate Anal.* 97 (2006) 1742–1756.
- [2] W.G. Cochran, The distribution of quadratic forms in a normal system, with applications to the analysis of covariance, *Math. Proc. Cambridge Philos. Soc.* 30 (1934) 178–191. <http://dx.doi.org/10.1017/S0305004100016595>.
- [3] T. Daniyarov, VaR: Value at Risk Estimation, R package version 0.2, 2004.
- [4] B. De Finetti, La prévision: ses lois logiques, ses sources subjectives, *Ann. l'Inst. Henri Poincaré* 7 (1937) 1–68.
- [5] P. Diaconis, D. Freedman, Finite exchangeable sequences, *Ann. Probab.* 8 (4) (1980) 745–764.
- [6] G.R. Heer, Testing independence in high dimensions, *Probab. Stat. Lett.* 12 (1) (1991) 73–81.
- [7] G.J. Kerns, G.J. Székely, De Finetti's theorem for abstract finite exchangeable sequences, *J. Theoret. Probab.* 19 (3) (2006) 589–608.
- [8] M.R. Kosorok, Discussion of: Brownian distance covariance, *Ann. Appl. Stat.* 3 (4) (2009) 1270–1278.
- [9] O. Ledoit, M. Wolf, Some hypothesis tests for the covariance matrix when the dimension is large compared to the sample size, *Ann. Statist.* 30 (2002) 1081–1102.
- [10] R. Lyons, Distance covariance in metric spaces, *Ann. Probab.* (2012) (in press). <http://arxiv.org/abs/1106.5758>.
- [11] M. Peligrad, An invariance principle for  $\varphi$ -mixing sequences, *Ann. Probab.* 13 (4) (1985) 1304–1313.
- [12] R Development Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, ISBN: 3-900051-07-0, 2010, <http://www.R-project.org>.
- [13] C.R. Rao, *Linear Statistical Inference and its Applications*, second ed., Wiley, New York, 1973.
- [14] M.L. Rizzo, G.J. Székely, Energy: E-statistics (energy statistics). R package version 1.1-1, 2010.
- [15] J.R. Schott, Testing for complete independence in high dimensions, *Biometrika* 92 (4) (2005) 951–956.
- [16] C. Stein, A bound for the error in the normal approximation to the distribution of a sum of dependent random variables, in: *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability* (Univ. California, Berkeley, Calif., 1970/1971), Vol. II: Probability theory, Univ. California Press, Berkeley, CA, 1972, pp. 583–602.
- [17] G.J. Székely, M.L. Rizzo, Brownian distance covariance, *Ann. Appl. Stat.* 3 (4) (2009) 1236–1265.
- [18] G.J. Székely, M.L. Rizzo, Rejoinder: Brownian distance covariance, *Ann. Appl. Stat.* 3 (4) (2009) 1303–1308.
- [19] G.J. Székely, M.L. Rizzo, N.K. Bakirov, Measuring and testing independence by correlation of distances, *Ann. Statist.* 35 (6) (2007) 2769–2794.