Rank Distance Correlation

June 12, 2015

Abstract

Keywords:

1 Introduction

In this paper we propose the rank distance correlation, based on the notion of distance correlation in Szekely et al. (2007), Szekely and Rizzo (2009). We show that the rank distance correlation is more robust against high-dimensionality and noise than the original distance correlation, in terms of independence test powers.

The independence test using distance correlation can be found in Bakirov et al. (2006).

A fast computation of distance correlation is recently proposed in Huo and Szekely (2015).

Our numerical setting is similar to Simon and Tibshirani (2012).

To add: more references on distance correlation applications, other independence tests and their applications.

2 Methods

2.1 Rank Distance Correlation

Let $(X_i, Y_i), i = 1, \ldots, n$ be the sample values from a pair of random variables (X, Y); and we calculate the distance matrix $A \in \mathcal{R}^{n\times n}$ for $\{X_i, i = 1, \ldots, n\}$, i.e., $A_{ij} = \|X_i - X_j\|$; as well as the distance matrix B for $\{Y_i\}$.

Define the centering matrix $H = I_n - \frac{J_n}{n}$, where I_n is the matrix of identity and J_n is the matrix of ones. Then doubly center the distance matrices as $A = HAH, B = HBH$, and the distance correlation for the sample data is defined as

$$dCor_n^2(X, Y) = \sum_{i,j=1}^{n} A_{ij}B_{ij} / \sqrt{\sum_{i,j=1}^{n} A_{ij}^2 \sum_{i,j=1}^{n} B_{ij}^2} \tag{1}$$

for the double-centered distance matrices.

The rank distance correlation replaces the distances in Equation by the ranks of distances. Let $A^R, B^R \in \mathcal{R}^{n\times n}$ be the rank matrices of the distance matrix A and B, which is computed as: $A^R_{ii} = 0$, and $A^R_{ij} = k$ if and only if A_{ij} is the kth smallest distance among $\{A_{ij}, i = 1, \ldots, j-1, j+1, \ldots, n\}$; break ties deterministically.

Then doubly center the ranks as $A^R = HA^RH$ and $B^R = HB^RH$. The ranked distance correlation is defined accordingly as

$$RdCor^2_n(X, Y) = \sum_{i,j=1}^{n} A^R_{ij} B^R_{ij} / \sqrt{\sum_{i,j=1}^{n} A^{R2}_{ij} \sum_{i,j=1}^{n} B^{R2}_{ij}}, \tag{2}$$

for the double-centered rank matrices.

2.2 Independence Tests

In this subsection, we compare distance correlation and rank distance correlation via independence tests of a pair of random variables (X, Y), and estimate the testing powers based on the generated sample data $(X_i, Y_i), i = 1, \ldots, n$. The null hypothesis is that they are not independent.

The setting is similar to Simon and Tibshirani (2012): Let $X \sim U(0, 1)^d \in \mathcal{R}^d$, where U is the uniform distribution and d is the dimensionality. Also let $Y = f(CX) + \epsilon \in \mathcal{R}$, where $f : \mathcal{R} \to \mathcal{R}$ is a function of X, C is a fixed $1 \times d$ transformation, and ϵ is random noise. So Y is a one-dimensional variable that is correlated with X.

For the choice of $f(\cdot)$, we use linear, quadratic, cubic, sine period 1/2, sine period 1/8, $X^{0.25}$, circle, step function, exponential, and log function. Please see the code for details. For C, we use $C(i) = 1/i, \forall i = 1, \ldots, d$, so that the entries of C decays as the dimension increases. Note that we may also use a random decay rather than the fixed decay, and the numerical phenomenon is similar.

To calculate the testing power for given function type, we first generate $n = 100$ pairs of (X_i, Y_i) from (X, Y), calculate the distance correlation, and repeat for 1000 Monte-Carlo replicates to obtain the empirical distribution of distance correlation under the null hypothesis. Then we re-generate $n = 100$ pairs of (X_i, Y_i) from (Z, Y) and calculate their distance correlation, where $Z \sim U(0,1)^d \in \mathcal{R}^d$ is independent from X and Y; repeat for 1000 Monte-Carlo replicates to obtain the empirical distribution of distance correlation under the alternative hypothesis.

At last we can estimate the testing power at given type 1 error level for distance correlation; and similarly the power of rank distance correlation.

2.3 Numerical Performance

For each function type, we estimate the testing power at type 1 error level 0.95.

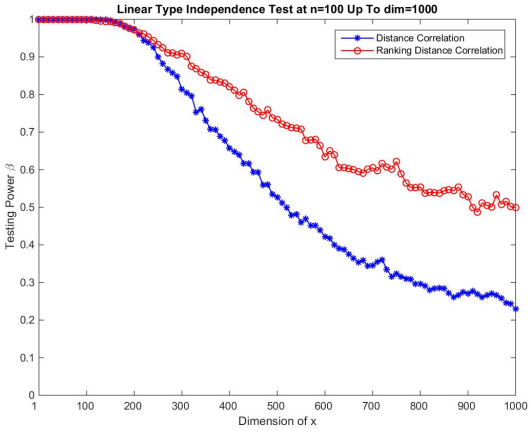
To assess the effect of high-dimensional data, we do the experiment for $d = 1, 10, 20, \ldots, 1000$ at $n = 100$ with $\epsilon = 0$, and show the power plots in Figure 1 and Figure 2.

To assess the effect of errors, we estimate the testing powers with increasing noise level (which is generated by Gaussian) at fixed $d = 500$, and show the power plots in Figure 3 and Figure 4.

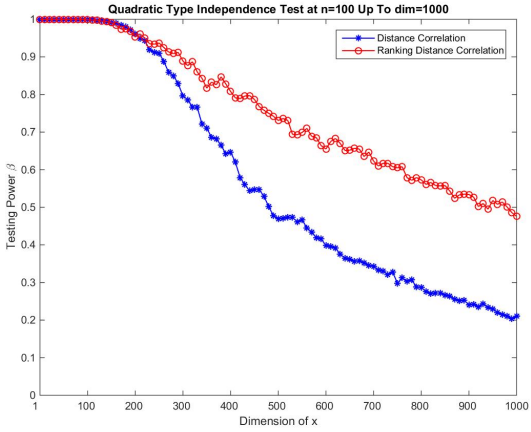
It is clear that rank distance correlation performs almost the same as distance correlation for low-dimensional data; but for high-dimensional data and noisy data, we observe a clear advantage of rank distance correlation in testing power.

3 Theorems

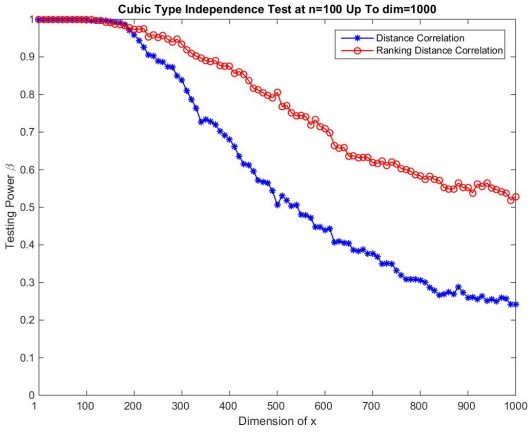
It is proved in Szekely et al. (2007) that distance correlation of (X, Y) is zero if and only if X is independent of Y. Similarly, we have



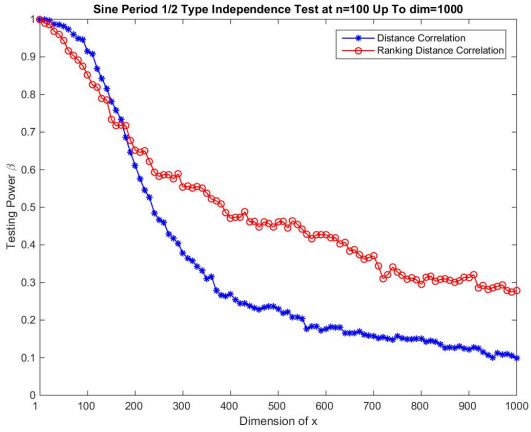
(a)



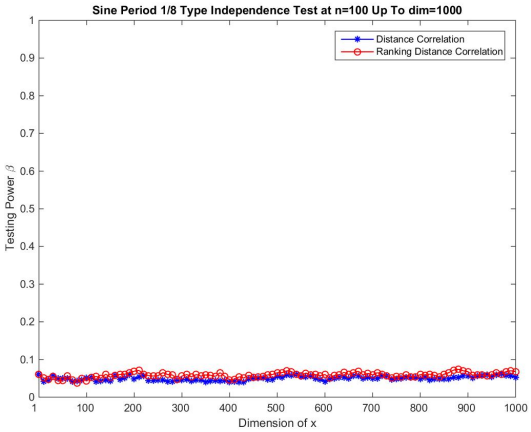
(b)



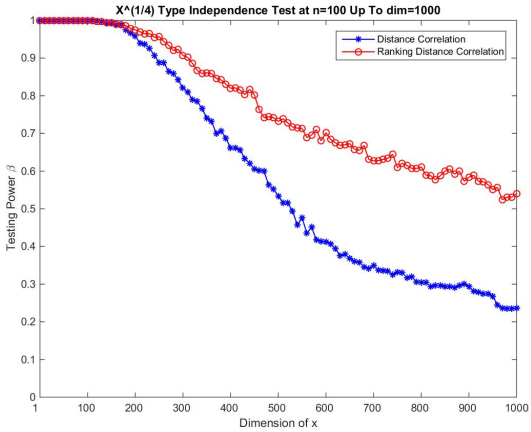
(c)



(d)

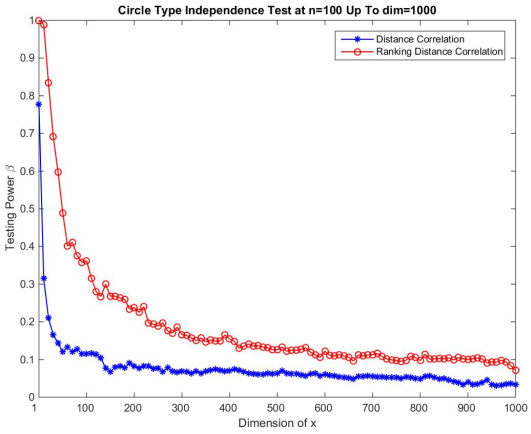


(e)

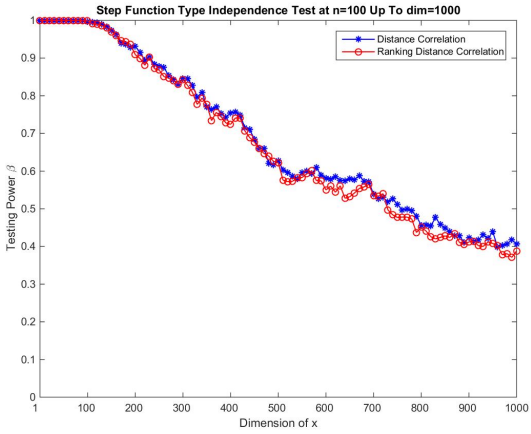


(f)

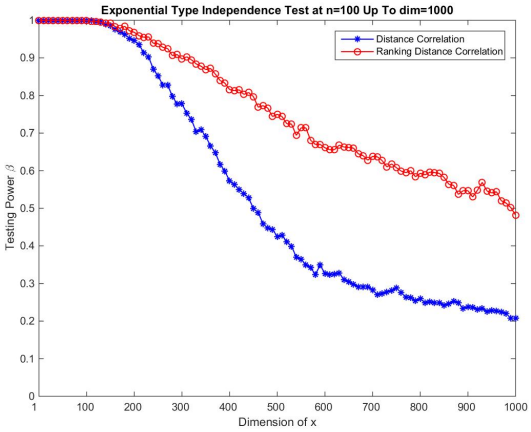
Figure 1: Testing Powers w.r.t. Increasing Dimension



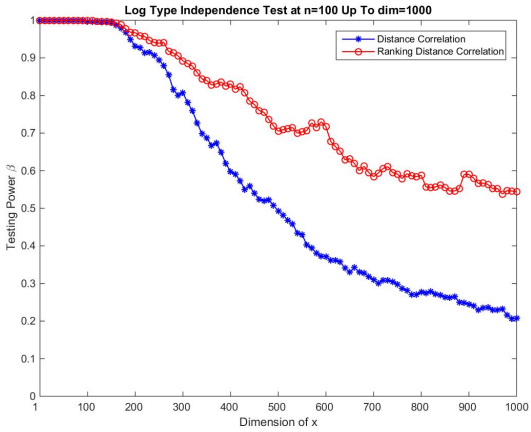
(a)



(b)

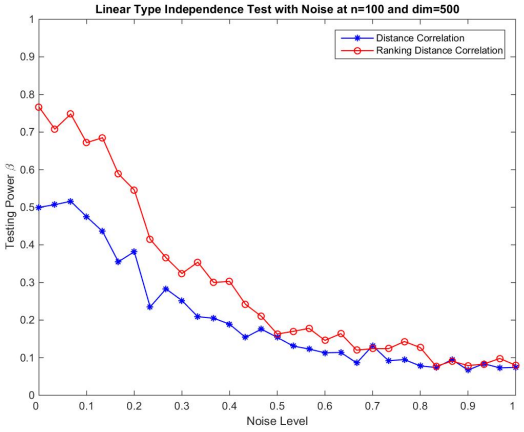


(c)

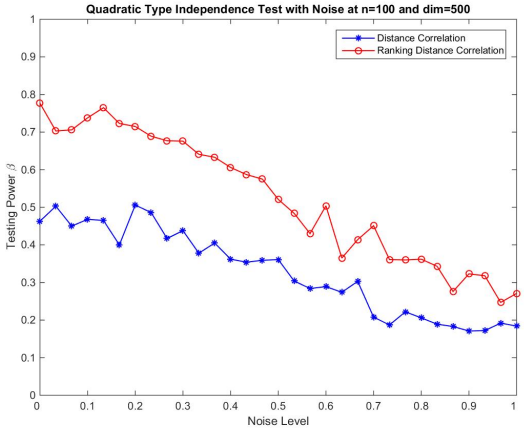


(d)

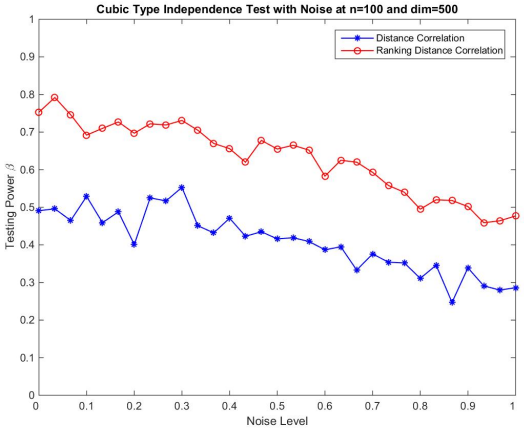
Figure 2: Testing Powers w.r.t. Increasing Dimension



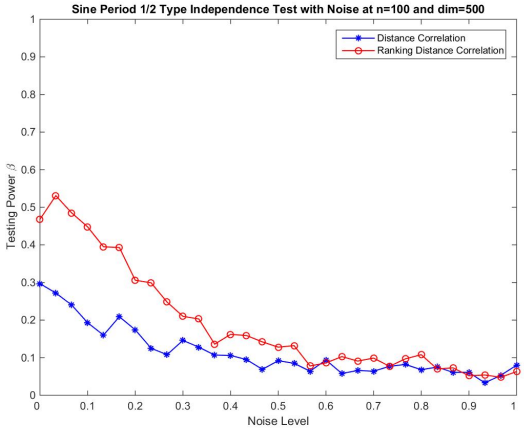
(a)



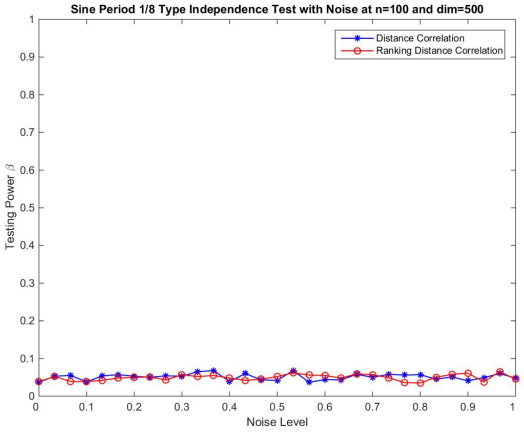
(b)



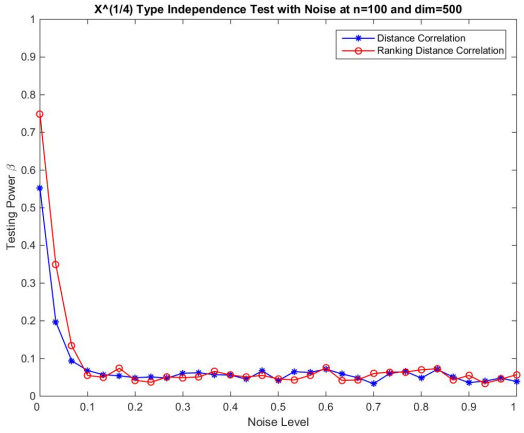
(c)



(d)

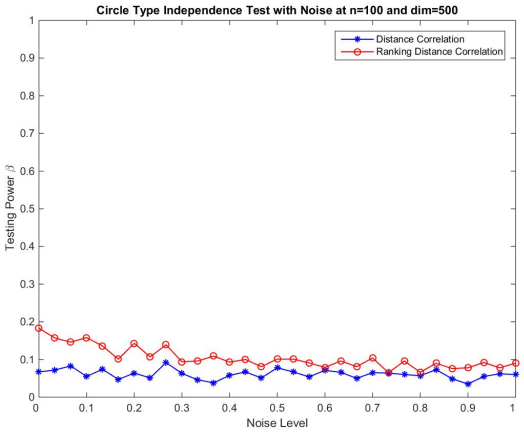


(e)

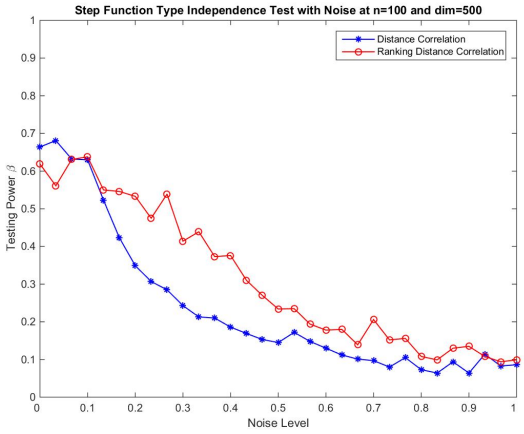


(f)

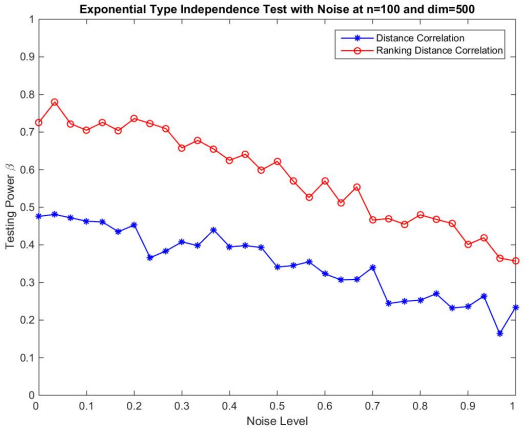
Figure 3: Testing Powers w.r.t. Increasing Noise



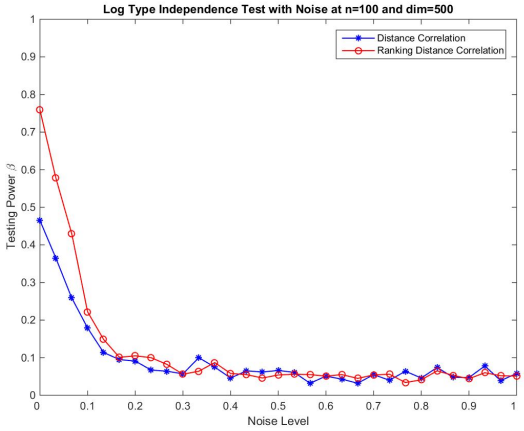
(a)



(b)



(c)



(d)

Figure 4: Testing Powers w.r.t. Increasing Noise

Theorem 1. *The rank distance correlation of* (X_i, Y_i) *converges to zero if and only if* X *is independent of* Y*.*

Also similar to using distance correlation to test independence, we have

Corollary 1. *For fixed* d *and* $n \to \infty$*, the rank distance correlation is a consistent test for testing independence of* $(X_i, Y_i), i = 1, \ldots, n$*.*

In terms of estimating independence, rank distance correlation can be more efficient than distance correlation.

Theorem 2. *For* $(X_i, Y_i), i = 1, \ldots, n$ *generated by independent* (X, Y)*,* $e(RdCor, dCor) = \frac{E(dCor^2)}{E(RdCor^2)} \geq 1$*.*

Furthermore, the relative efficiency is strictly larger than 1 *as* d *increases and* n *fixed.*

tba: adjust the theorems based on the proving progress. We may need constraints on distribution for theorem 1 and corollary 1; theorem 2 empirically holds.

References

Bakirov, N., M. Rizzo, and G. Szekely (2006). A multivariate nonparametric test of independence. *Journal of Multivariate Analysis 97*, 1742–1756.

Huo, X. and G. Szekely (2015). Fast computing for distance covariance. *http://arxiv.org/abs/1410.1503*.

Simon, N. and R. Tibshirani (2012). Comment on detecting novel associations in large data sets. *http://arxiv.org/abs/1401.7645*.

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

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