Rank Distance Correlation

July 18, 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).

rank correlation is very popular, in particular for nonparametric statistics, see in Kendall (1970).

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 $A = AH, B = BH$, 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: $A_{ii}^R = 0$, and $A_{ij}^R = k$ 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

$$RdCor_n^2(X,Y) = \sum_{i,j=1}^{n} A_{ij}^R B_{ij}^R / \sqrt{\sum_{i,j=1}^{n} A_{ij}^{R2} \sum_{i,j=1}^{n} B_{ij}^{R2}}, \tag{2}$$

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.

Once the distances are sorted by $n^2 log(n)$, it took n^2 to calculate all rdcorr from $k = 1, \ldots, n-1$.

3 Numerical Experiments

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 **??**.

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 **??** and Figure **??**.

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.

Table 1: Simulation results at $n = 30$ for independence test at level 0.05. The power is estimated based on 5000 MC replicates.

Distribution	DCorr	MDCorr	HHG	RDCorr
W	0.5998	0.6264	0.9434	**0.9714**
$Square$	0.0582	0.0684	0.1080	**0.1932**
$Diamond$	0.0292	0.0456	0.1604	**0.1770**
$Parabola$	0.5436	0.4964	**0.8202**	0.8180
$TwoParabolas$	0.1290	0.1046	0.6050	**0.6996**
$Circle$	0.0676	0.0818	0.6942	**0.8402**
$IndependentClouds$	0.0486	0.0482	0.0484	**0.0584**

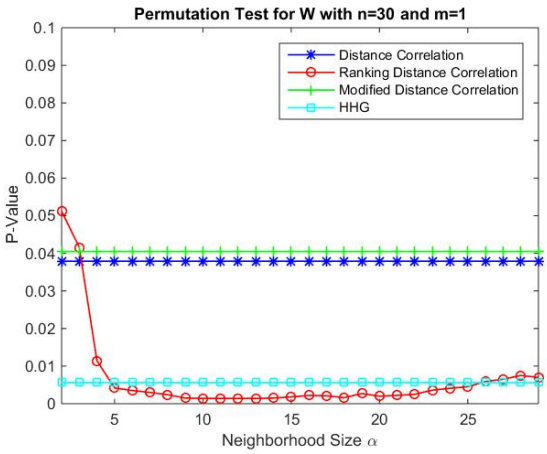
4 Theorems

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

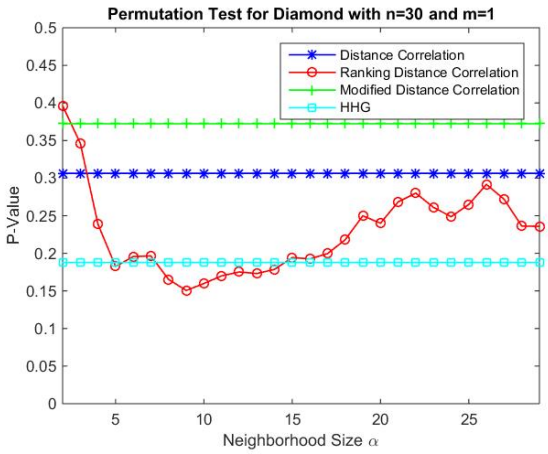
Theorem 1. *The rank distance correlation of (X_i, Y_i) converges to zero if and only if X is independent of Y.*

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

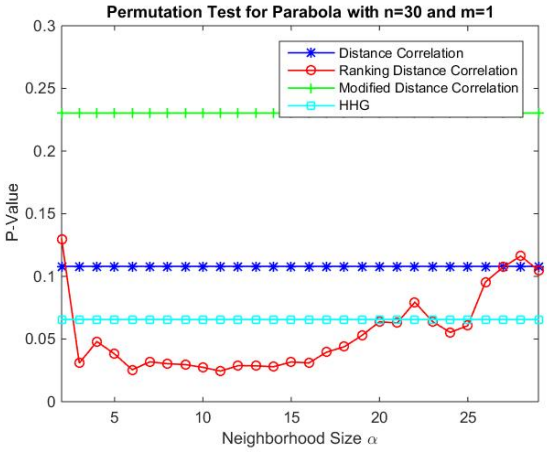
Corollary 1. *For fixed m and $n \to \infty$, the rank distance correlation is a consistent statistic for independence/permutation test of $(X_i, Y_i), i = 1, \ldots, n$.*



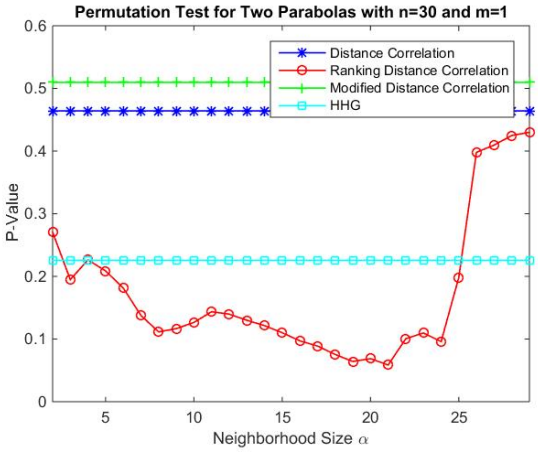
(a)



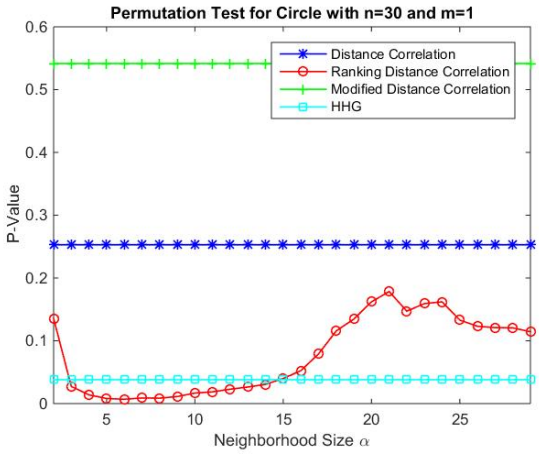
(b)



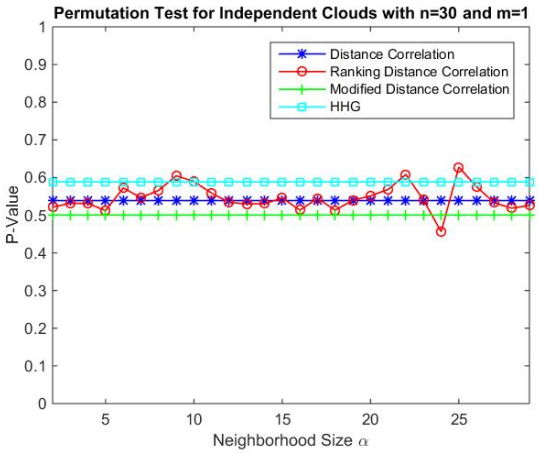
(c)



(d)



(e)



(f)

Figure 1: P-Value

Table 2: Simulation results at $n = 50$ for independence test at level 0.05. The power is estimated based on 5000 MC replicates.

Distribution	DCorr	MDCorr	HHG	RDCorr
W	0.9364	0.9394	0.9990	**0.9998**
$Square$	0.0736	0.0840	0.2112	**0.3014**
$Diamond$	0.0434	0.0770	**0.4874**	0.3616
$Parabola$	0.9306	0.9116	0.9840	**0.9896**
$TwoParabolas$	0.1766	0.1342	0.9064	**0.9512**
$Circle$	0.0950	0.1160	0.9830	**0.9978**
$IndependentClouds$	0.0498	0.0496	0.0470	**0.0568**

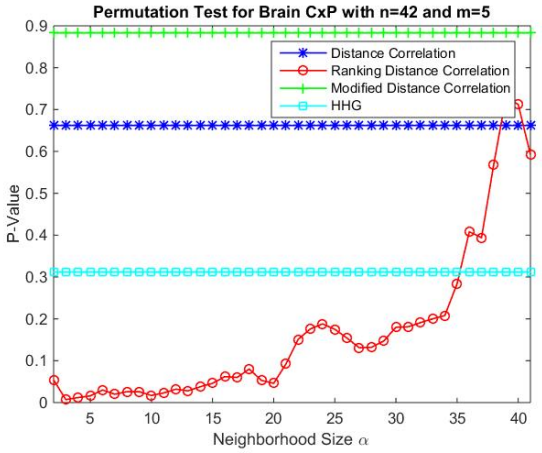


Figure 2: p-value of Brain CxP data

When independent, all test statistic at $k = 1, ...n$ are close to 0.

Facts: By Cauchy Schwarz, we always have

$$E(|RDCov(X,Y)|) \leq \sqrt{E(RDVar(X,X))E(RDVar(Y,Y))}, \tag{3}$$

so $RdCor \in [-1, 1]$.

Note that we always have $E(n^2 RDVar(X,X)) = E(n^2 RDVar(Y,Y)) = \frac{2*n^4-3*n^3+n^2}{4}$.

When independent, we have $E(n^2 RDCov(X,Y)) = \frac{n^2*(n-1)}{4}$; when $X = Y$, we have $E(n^2 RDCov(X,Y)) = E(n^2 RDVar(X,X)) = E(n^2 RDVar(Y,Y))$.

Summary: RdCorr is robust against dimensionality, noise and non-linearity. And its' performance is usually superior or similar to the best of dcorr/mdcorr/HHG in our numerical experiments.

It is robust in hd because each entry in $RDCov$ and $RDVar$ are not affected by dimensionality (similar to mdcorr).

robust to noise because it is based on distance (similar to dcorr),

robust to non-linearity because the test statistic is based on neighborhood info, which can capture local relationship (similar but better than HHG).

Also note that $RdCor$ does not require finite moments in all theorems, which $dCorr$ and $MdCorr$ require.

When the testing power/p-value of $RdCor$ for small neighborhood size is significantly better than large neighborhood, it implies that the functional relationship is likely nonlinear or there exists nonzero curvature.

$dCor/mDCorr$ tends to perform sub-optimal for nonlinear relationship, while HHG tends to perform bad in our high-dimensional setting. On the other hand, $RdCor$ combines the best of $dCorr$, $MdCorr$ and HHG, which exhibits a robust and good performance throughout all our experiments.

tba: adjust all theorems and facts by considering the neighborhood choice k rather than $n-1$.

n vertices per graph

m subjects

D is the maximum rank of the expected adjacency matrices

$X_d \in R^n$ of latent eigenvector

$\lambda_i \in \triangle_D$ is the relative weights of the eigenvectors per subject i

$P_i \in [0,1]^{nxn}$ is the expected adjacency matrix for subject i

A_i is the adjacency matrix for subject i, trial j

$Y_i \in R^p$ is vector of covariates

Some proposed constants:

$n = 100$, $m = 50$, $p = 5$

$X_d = e_d$, that is, X_d is a zero vector with a single 1 in the d-th element

$Y \sim F_Y = Dir(1)$, so, uniform

$\lambda|y \sim F_{\lambda|Y} = Dir(Y)$,

for all i in [n]

$y_i \sim F_Y$

$\lambda_i|y \sim F_{\lambda|y}$

$P_i = \sum_d^D \lambda_{id} \langle X_d, X_d \rangle$

$A_i \sim$ Bern $(\langle X_i, X_i \rangle)$

end

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*.

Kendall, M. G. (1970). *Rank Correlation Methods*. London: Griffin.

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.