Local Distance Correlation For Testing Independence

Cencheng Shen

Joint Work with Joshua T. Vogelstein & Carey E. Priebe

November 19, 2015

Overview

Section 1

Motivations

Motivations

- Given multiple data sets, we would like to test whether they are independent or not.

Motivations

- Given multiple data sets, we would like to test whether they are independent or not.
- Modern data sets may be high-dimensional, nonlinear, noisy, of small sample size; and different data sets may come from disparate spaces.

Motivations

- Given multiple data sets, we would like to test whether they are independent or not.
- Modern data sets may be high-dimensional, nonlinear, noisy, of small sample size; and different data sets may come from disparate spaces.
- One example we have is Brain Connectome vs Personality: the Brain Connectome is measured on 197 brain regions for 194 time steps, and the personality data is a five-factor model. The sample size is 42.

Motivations

- Given multiple data sets, we would like to test whether they are independent or not.
- Modern data sets may be high-dimensional, nonlinear, noisy, of small sample size; and different data sets may come from disparate spaces.
- One example we have is Brain Connectome vs Personality: the Brain Connectome is measured on 197 brain regions for 194 time steps, and the personality data is a five-factor model. The sample size is 42.
- Initially we tried various regression methods on the data in order to predict personality from Connectome, but it does not work...

- So we decide to test whether the two data sets are independent or not: if they are independent, then there is no point to do prediction.

- So we decide to test whether the two data sets are independent or not: if they are independent, then there is no point to do prediction.
- The basic question is: **How to test independence on real data???**

- So we decide to test whether the two data sets are independent or not: if they are independent, then there is no point to do prediction.
- The basic question is: **How to test independence on real data???**
- The Pearson correlation coefficient and RV coefficient are mostly useful for finding linear relationship and may be zero for dependent data sets, while mutual information requires estimating the probability distribution.

- We turn to a great method in the Statistical community: **distance correlation**.

- We turn to a great method in the Statistical community: **distance correlation**.
- It is easy to compute for given data in Euclidean space, and is a consistent test statistic with good simulation results [1].

- We turn to a great method in the Statistical community: **distance correlation**.
- It is easy to compute for given data in Euclidean space, and is a consistent test statistic with good simulation results [1].
- But it does not work well against high-dimensional and non-linear dependency.

- We turn to a great method in the Statistical community: **distance correlation**.
- It is easy to compute for given data in Euclidean space, and is a consistent test statistic with good simulation results [1].
- But it does not work well against high-dimensional and non-linear dependency.
- And its theoretical consistency is not equivalent to good finite-sample testing power: the sample size may be limited due to expensive data collection, while the required sample size to achieve good power for a particular dependency type can be very large.

- We turn to a great method in the Statistical community: **distance correlation**.
- It is easy to compute for given data in Euclidean space, and is a consistent test statistic with good simulation results [1].
- But it does not work well against high-dimensional and non-linear dependency.
- And its theoretical consistency is not equivalent to good finite-sample testing power: the sample size may be limited due to expensive data collection, while the required sample size to achieve good power for a particular dependency type can be very large.
- Most importantly, it fails to detect any relationship in our Connectome vs personality data.

- There exists other methods in the same distance-based framework:

- There exists other methods in the same distance-based framework:
- Modified distance correlation [2] is robust against high-dimensionality, but still not nonlinear data.

- There exists other methods in the same distance-based framework:
- Modified distance correlation [2] is robust against high-dimensionality, but still not nonlinear data.
- The HHG statistic [3] works well for nonlinear data, but falls a bit short in linear and high-dimensional data.

- There exists other methods in the same distance-based framework:
- Modified distance correlation [2] is robust against high-dimensionality, but still not nonlinear data.
- The HHG statistic [3] works well for nonlinear data, but falls a bit short in linear and high-dimensional data.
- The kernel-based independence test developed in [4] is a similar tool in the machine learning community, whose connection to distance correlation is established in [5].

Section 2

Global Distance Correlation

Set-Up

- Given two data sets $\mathcal{X} = [X_1, \cdots, X_n] \in \mathcal{R}^{m_X \times n}$ and $\mathcal{Y} = [Y_1, \cdots, Y_n] \in \mathcal{R}^{m_Y \times n}$,

Set-Up

- Given two data sets $\mathcal{X} = [X_1, \cdots, X_n] \in \mathcal{R}^{m_X \times n}$ and $\mathcal{Y} = [Y_1, \cdots, Y_n] \in \mathcal{R}^{m_Y \times n}$,
- Assume that $X_i, i = 1, \ldots, n$ are identically independently distributed (i.i.d.) as X, similarly $Y_i \overset{i.i.d.}{\sim} Y$.

Set-Up

- Given two data sets $\mathcal{X} = [X_1, \cdots, X_n] \in \mathcal{R}^{m_X \times n}$ and $\mathcal{Y} = [Y_1, \cdots, Y_n] \in \mathcal{R}^{m_Y \times n}$,
- Assume that $X_i, i = 1, \ldots, n$ are identically independently distributed (i.i.d.) as X, similarly $Y_i \overset{i.i.d.}{\sim} Y$.
- For testing independence between X and Y, the null and the alternative hypothesis are

$$
\begin{aligned}
&H_0 : X \text{ is independent of } Y, i.e., f_{XY} = f_X f_Y, \\
&H_A : f_{XY} \neq f_X f_Y,
\end{aligned}
$$

where f_{XY} denotes the joint distribution of $(X, Y) \in \mathcal{R}^{m_X + m_Y}$, and f_X and f_Y are the marginal distributions.

Distance Covariance

- Since the random variables are not directly observed, we test by the sample data $\mathcal{X}$ and $\mathcal{Y}$.

Distance Covariance

- Since the random variables are not directly observed, we test by the sample data $\mathcal{X}$ and $\mathcal{Y}$.
- We first calculate two Euclidean distance matrices $A, B \in \mathcal{R}^{n \times n}$ for $\mathcal{X}$ and $\mathcal{Y}$ respectively, i.e., $A_{ij} = \|X_i - X_j\|_2$.

Distance Covariance

- Since the random variables are not directly observed, we test by the sample data $\mathcal{X}$ and $\mathcal{Y}$.
- We first calculate two Euclidean distance matrices $A, B \in \mathcal{R}^{n \times n}$ for $\mathcal{X}$ and $\mathcal{Y}$ respectively, i.e., $A_{ij} = \|X_i - X_j\|_2$.
- The sample distance covariance is defined as

$$dCov_n(\mathcal{X}, \mathcal{Y}) = \frac{1}{n^2} \sum_{i,j=1}^{n} A_{ij}^H B_{ij}^H, \tag{1}$$

where $A^H = HAH$, $B^H = HBH$ with $H = I_n - \frac{J_n}{n}$.

Distance Covariance

- Since the random variables are not directly observed, we test by the sample data $\mathcal{X}$ and $\mathcal{Y}$.
- We first calculate two Euclidean distance matrices $A, B \in \mathcal{R}^{n \times n}$ for $\mathcal{X}$ and $\mathcal{Y}$ respectively, i.e., $A_{ij} = \|X_i - X_j\|_2$.
- The sample distance covariance is defined as

$$dCov_n(\mathcal{X}, \mathcal{Y}) = \frac{1}{n^2} \sum_{i,j=1}^{n} A_{ij}^H B_{ij}^H, \tag{1}$$

where $A^H = HAH$, $B^H = HBH$ with $H = I_n - \frac{J_n}{n}$.
- Then the sample distance variance is defined as

$$dVar_n(\mathcal{X}) = \frac{1}{n^2} \sum_{i,j=1}^{n} A_{ij}^H A_{ij}^H$$

$$dVar_n(\mathcal{Y}) = \frac{1}{n^2} \sum_{i,j=1}^{n} B_{ij}^H B_{ij}^H.$$

Distance Correlation

The squared sample distance correlation is obtained by normalizing the distance covariance

$$dCorr_n(\mathcal{X}, \mathcal{Y}) = \frac{dCov_n(\mathcal{X}, \mathcal{Y})}{\sqrt{dVar_n(\mathcal{X}) \cdot dVar_n(\mathcal{Y})}}, \tag{2}$$

where all of $dCov_n, dVar_n, dCorr_n$ are always non-negative.

Consistency of Distance Correlation

Theorem 1

As $n \to \infty$, the sample distance correlation converges to 0 if and only if the null hypothesis is true, i.e., X is independent of Y.

Consistency of Distance Correlation

Theorem 1

As $n \to \infty$, the sample distance correlation converges to 0 if and only if the null hypothesis is true, i.e., X is independent of Y.

Therefore distance correlation is a consistent test of independence, i.e., the testing power $\beta \to 1$ as $n \to \infty$.

Consistency of Distance Correlation

Theorem 1

As $n \to \infty$, the sample distance correlation converges to 0 if and only if the null hypothesis is true, i.e., X is independent of Y.

Therefore distance correlation is a consistent test of independence, i.e., the testing power $\beta \to 1$ as $n \to \infty$.

The proof is kind of ingenious by [1], as they show the distance covariance is asymptotically an integral of the joint characteristic function minus the product of the two marginal characteristic functions squared.

Modified Distance Correlation

- It turns out the distance correlation is not robust against high-dimensionality.

Modified Distance Correlation

- It turns out the distance correlation is not robust against high-dimensionality.
- For example, if m_X or m_Y increases with the sample size n, we have $dCorr_n \rightarrow 1$ as $m_X, m_Y \rightarrow \infty$, for two independent Gaussian distributions.

Modified Distance Correlation

- It turns out the distance correlation is not robust against high-dimensionality.
- For example, if m_X or m_Y increases with the sample size n, we have $dCorr_n \to 1$ as $m_X, m_Y \to \infty$, for two independent Gaussian distributions.
- A modified distance correlation is proposed in [2], which is asymptotically the same as the original distance correlation.

Modified Distance Correlation

- It turns out the distance correlation is not robust against high-dimensionality.
- For example, if m_X or m_Y increases with the sample size n, we have $dCorr_n \to 1$ as $m_X, m_Y \to \infty$, for two independent Gaussian distributions.
- A modified distance correlation is proposed in [2], which is asymptotically the same as the original distance correlation.
- So modified distance correlation is also consistent with the additional benefit of being robust against high-dimensional dependency.

Section 3

Local Distance Correlation

Set-Up

- Under the same setting of global distance covariance, we further sort the distance matrix A within column and denote the ranks as $r(A_{ij})$.

Set-Up

- Under the same setting of global distance covariance, we further sort the distance matrix A within column and denote the ranks as $r(A_{ij})$.
- Namely for each $i = 1, \ldots, n$, we always set $r(A_{ii}) = 0$; then set $r(A_{ij}) = k$ if and only if A_{ij} is the kth smallest distance in $\{A_{ij}, i = 1, \ldots, n \;\&\; i \neq j\}$; break ties deterministically.

Set-Up

- Under the same setting of global distance covariance, we further sort the distance matrix A within column and denote the ranks as $r(A_{ij})$.
- Namely for each $i = 1, \ldots, n$, we always set $r(A_{ii}) = 0$; then set $r(A_{ij}) = k$ if and only if A_{ij} is the kth smallest distance in $\{A_{ij}, i = 1, \ldots, n \ \& \ i \neq j\}$; break ties deterministically.
- Similarly sort the distance matrix B within column and denote the ranks by $r(B_{ij})$.

Local Distance Covariance

Then we define the local distance covariance for each $k, l = 1, \ldots, n$ as

$$dCov_{kl}(\mathcal{X}, \mathcal{Y}) = \frac{1}{n^2} \sum_{i,j=1}^{n} A_{ij}^H B_{ij}^H \mathcal{I}(r(A_{ij}) < k) \mathcal{I}(r(B_{ij}) < l), \tag{3}$$

and define the local original distance variance as

$$dVar_k(\mathcal{X}) = \frac{1}{n^2} \sum_{i,j=1}^{n} A_{ij}^H A_{ij}^H \mathcal{I}(r(A_{ij}) < k)$$

$$dVar_l(\mathcal{Y}) = \frac{1}{n^2} \sum_{i,j=1}^{n} B_{ij}^H B_{ij}^H \mathcal{I}(r(B_{ij}) < l),$$

where $\mathcal{I}(\cdot)$ is the indicator function.

Local Distance Correlation

Normalizing local original distance covariance at each k, l yields the local distance correlation:

$$dCorr_{kl}(\mathcal{X}, \mathcal{Y}) = \frac{dCov_{kl}(\mathcal{X}, \mathcal{Y})}{\sqrt{dVar_k(\mathcal{X}) \cdot dVar_l(\mathcal{Y})}}. \tag{4}$$

When $k = l$, we simplify the notations to $dCov_k$ and $dCorr_k$.

Local Distance Correlation

Normalizing local original distance covariance at each k, l yields the local distance correlation:

$$dCorr_{kl}(\mathcal{X}, \mathcal{Y}) = \frac{dCov_{kl}(\mathcal{X}, \mathcal{Y})}{\sqrt{dVar_k(\mathcal{X}) \cdot dVar_l(\mathcal{Y})}}. \tag{4}$$

When $k = l$, we simplify the notations to $dCov_k$ and $dCorr_k$.

Note that our local distance correlation refers to the family of test statistics $\{dCorr_{kl}, k, l = 1, \ldots, n\}$ rather than each individual $dCorr_{kl}$, i.e., the testing power of local distance correlation equals the best power among the family.

Permutation Test

- Given $\mathcal{X}$ and $\mathcal{Y}$, the permutation test is done as follows:

Permutation Test

- Given $\mathcal{X}$ and $\mathcal{Y}$, the permutation test is done as follows:
- Compare $dCorr_{kl}(\mathcal{X}, \mathcal{Y})$ to $\{dCorr_{kl}(\mathcal{X}, \mathcal{Y}P), \forall P\}$ containing the test statistic for all permutations P.

Permutation Test

- Given $\mathcal{X}$ and $\mathcal{Y}$, the permutation test is done as follows:
- Compare $dCorr_{kl}(\mathcal{X}, \mathcal{Y})$ to $\{dCorr_{kl}(\mathcal{X}, \mathcal{Y}P), \forall P\}$ containing the test statistic for all permutations P.
- This yields the p-value. The smaller the p-value, the more significant the dependency.

Permutation Test

- Given $\mathcal{X}$ and $\mathcal{Y}$, the permutation test is done as follows:
- Compare $dCorr_{kl}(\mathcal{X}, \mathcal{Y})$ to $\{dCorr_{kl}(\mathcal{X}, \mathcal{Y}P), \forall P\}$ containing the test statistic for all permutations P.
- This yields the p-value. The smaller the p-value, the more significant the dependency.
- We usually carry out the permutation test for r random permutations rather than all permutations.

Permutation Test

- Given $\mathcal{X}$ and $\mathcal{Y}$, the permutation test is done as follows:
- Compare $dCorr_{kl}(\mathcal{X}, \mathcal{Y})$ to $\{dCorr_{kl}(\mathcal{X}, \mathcal{Y}P), \forall P\}$ containing the test statistic for all permutations P.
- This yields the p-value. The smaller the p-value, the more significant the dependency.
- We usually carry out the permutation test for r random permutations rather than all permutations.
- When the true distribution f_{XY} is known, we may set the type 1 error level α, repeatedly generate the data for some MC replicates, and estimate the testing power.

Consistency of Local Distance Correlation

Theorem 2

Local distance correlation is consistent for testing independence against all alternatives.

Consistency of Local Distance Correlation

Theorem 2

Local distance correlation is consistent for testing independence against all alternatives.

The same consistency holds for local modified distance correlation, which can be similarly defined and is actually the best test statistic in the simulation.

Advantage of Local vs Global

- The idea of k-nearest-neighbor has been long established for unfolding non-linearity within a data set.

Advantage of Local vs Global

- The idea of k-nearest-neighbor has been long established for unfolding non-linearity within a data set.
- In case of nonlinear dependency, a small distance in one data set may corresponds to a large distance in the other data set. Excluding such products in computing the distance correlation may help the finite-sample testing.

Advantage of Local vs Global

- The idea of k-nearest-neighbor has been long established for unfolding non-linearity within a data set.
- In case of nonlinear dependency, a small distance in one data set may corresponds to a large distance in the other data set. Excluding such products in computing the distance correlation may help the finite-sample testing.
- Once the distance matrices are sorted within column, the running time to compute $\{dCorr_{kl}\}$ is always $O(n^2)$. So it is not necessary to pick the optimal neighborhood.

Advantage of Local vs Global

Theorem 3

Suppose $Y = cX$ for a non-zero scalar c, then for any n we always have

$$\beta(dCorr_n) \geq \beta(dCorr_{kl}) \tag{5}$$

for all $k, l = 1, \ldots, n$, where β is the permutation test power at a given type 1 error α.
Thus local distance correlation is no better than global distance correlation under linear dependency.

Theorem 4

There exists f_{XY}, n and α such that

$$\beta(dCorr_n) > \beta(dCorr_{kl}) \tag{6}$$

for some $(k, l) \neq (n, n)$, where β is the permutation test power at the type 1 error α.
Thus local distance correlation can be better than global distance correlation under certain nonlinear dependency.

The joint distribution in its proof corresponds to the quadratic case in the simulation.

Section 4

Experiments

Set-Up

- We consider 20 different distributions f_{XY}, including various dependency types such as linear, quadratic, joint normal, circle, trigonometry, uncorrelated binomial, multiplicative noise, independent clouds, etc.

Set-Up

- We consider 20 different distributions f_{XY}, including various dependency types such as linear, quadratic, joint normal, circle, trigonometry, uncorrelated binomial, multiplicative noise, independent clouds, etc.
- For those joint distributions, we further consider two different scenarios: a dimension 1 with increasing sample size scenario, and a fixed sample size with increasing dimension scenario.

Set-Up

- We consider 20 different distributions f_{XY}, including various dependency types such as linear, quadratic, joint normal, circle, trigonometry, uncorrelated binomial, multiplicative noise, independent clouds, etc.
- For those joint distributions, we further consider two different scenarios: a dimension 1 with increasing sample size scenario, and a fixed sample size with increasing dimension scenario.
- In each scenario, we estimate the testing powers of each distribution, for local distance correlation, global distance correlation, and HHG.

Results

- We will see that local distance correlation is similar to the global one for close to linear dependencies, but significantly improves for nonlinear dependencies.

Results

- We will see that local distance correlation is similar to the global one for close to linear dependencies, but significantly improves for nonlinear dependencies.
- And the local modified distance correlation method is the most reliable test statistic throughout the simulations, due to its robustness against high-dimensionality and non-linearity at the same time.

Results

- We will see that local distance correlation is similar to the global one for close to linear dependencies, but significantly improves for nonlinear dependencies.
- And the local modified distance correlation method is the most reliable test statistic throughout the simulations, due to its robustness against high-dimensionality and non-linearity at the same time.
- Local modified distance correlation also has the lowest p-value for the real data experiment on detecting signal between brain Connectome and personality.

Visualization of Sample Data for Each Joint Distribution

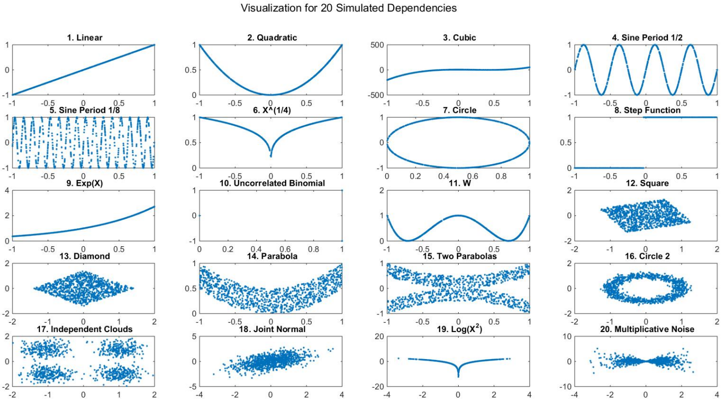

Figure: Visualization of 20 Dependencies at $n = 1000$ and Dimension 1 with No Noise.

Testing Powers at Dimension 1

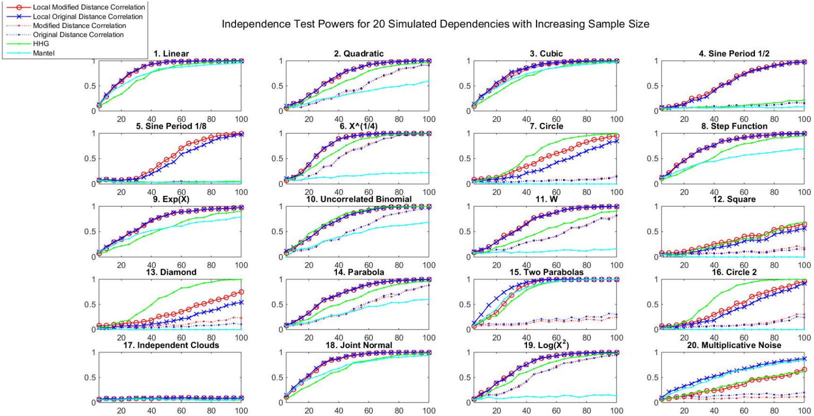

Figure: Testing Powers of Dimension 1 Simulations with Increasing Sample Size

Performance Profiles at Dimension 1

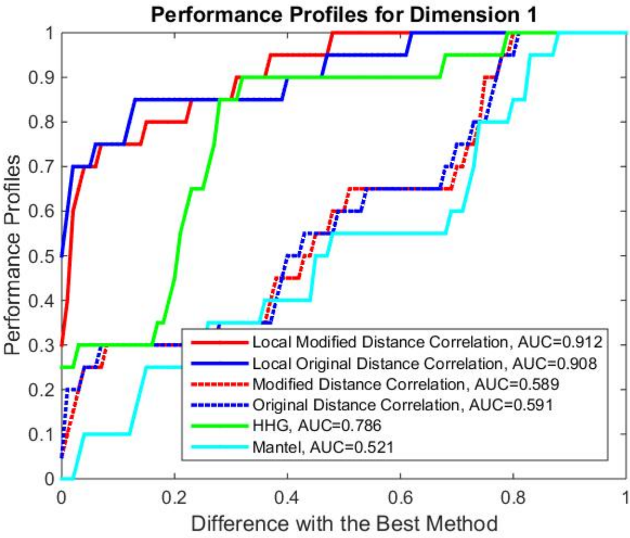

Figure: Performance Profiles of Dimension 1 Simulations

Testing Powers of Increasing Dimension

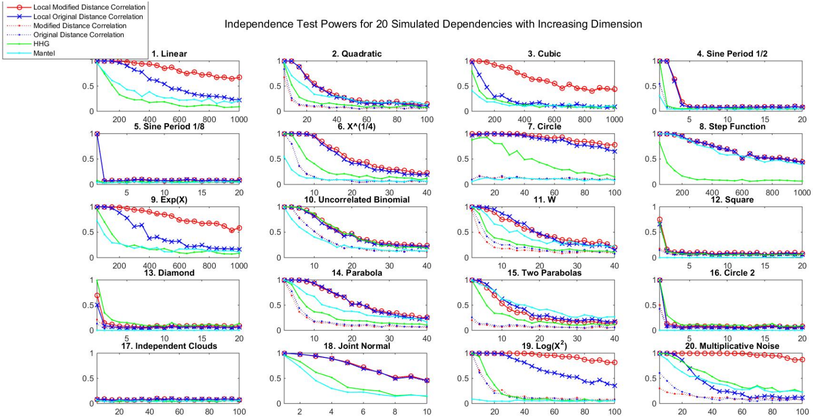

Figure: Testing Powers of Increasing Dimension Simulations with Fixed Sample Size

Performance Profiles of Increasing Dimension

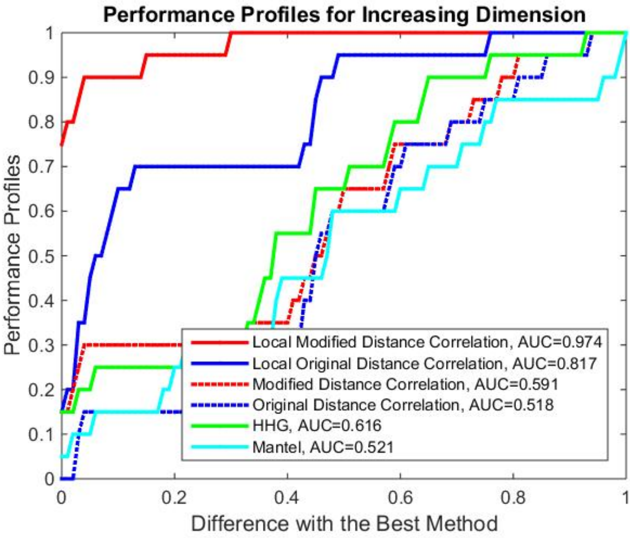

Figure: Performance Profiles of Increasing Dimension Simulations

P-Value of Real Data

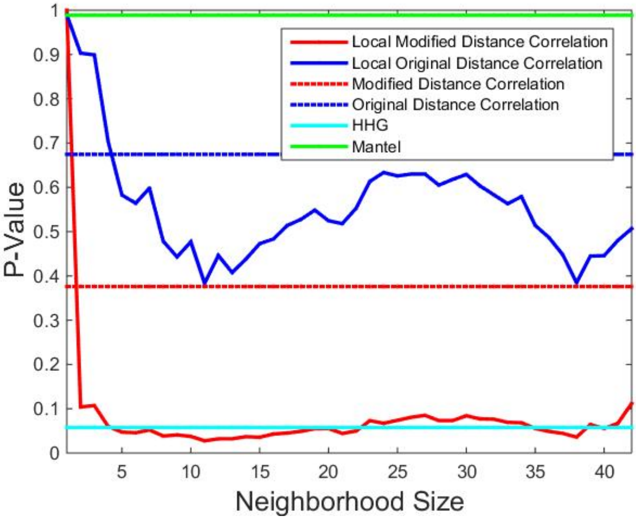

Figure: P-Value of Brain Connectome and Personality

Section 5

Conclusion

- We propose a local distance correlation to test independence.

- We propose a local distance correlation to test independence.
- It is not only theoretically consistent, but also achieves state-of-the-art performance in finite-sample testing of various dependencies.

- We propose a local distance correlation to test independence.
- It is not only theoretically consistent, but also achieves state-of-the-art performance in finite-sample testing of various dependencies.
- Comparing to the benchmarks, it is overall the best method for testing data sets of linearity or non-linearity, high-dimensionality, and small sample-size.

- We propose a local distance correlation to test independence.
- It is not only theoretically consistent, but also achieves state-of-the-art performance in finite-sample testing of various dependencies.
- Comparing to the benchmarks, it is overall the best method for testing data sets of linearity or non-linearity, high-dimensionality, and small sample-size.
- It is able to detect signal in our real data! (Although the underlying truth is unknown)

Potential Works

- The whole framework is applicable to Euclidean data/metric; but what about testing multiple graphs, or graph and other features?

Potential Works

- The whole framework is applicable to Euclidean data/metric; but what about testing multiple graphs, or graph and other features?
- Local distance correlation is potentially related to nonlinear embedding and its k-nearest-neighbor choice.

Potential Works

- The whole framework is applicable to Euclidean data/metric; but what about testing multiple graphs, or graph and other features?
- Local distance correlation is potentially related to nonlinear embedding and its k-nearest-neighbor choice.
- Other theoretical aspects/real-data applications of local distance correlation.

Potential Works

- The whole framework is applicable to Euclidean data/metric; but what about testing multiple graphs, or graph and other features?
- Local distance correlation is potentially related to nonlinear embedding and its k-nearest-neighbor choice.
- Other theoretical aspects/real-data applications of local distance correlation.
- Move from testing independence to classification/prediction!

References I

G. Szekely, M. Rizzo, and N. Bakirov, "Measuring and testing independence by correlation of distances," *Annals of Statistics*, vol. 35, no. 6, pp. 2769–2794, 2007.

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

R. Heller, Y. Heller, and M. Gorfine, "A consistent multivariate test of association based on ranks of distances," *Biometrika*, vol. 100, no. 2, pp. 503–510, 2013.

A. Gretton and L. Gyorfi, "Consistent nonparametric tests of independence," *Journal of Machine Learning Research*, vol. 11, pp. 1391–1423, 2010.

D. Sejdinovic, B. Sriperumbudur, A. Gretton, and K. Fukumizu, "Equivalence of distance-based and rkhs-based statistics in hypothesis testing," *Annals of Statistics*, vol. 41, no. 5, pp. 2263–2291, 2013.