Local Graph Correlation

Cencheng Shen*[1], Joshua T. Vogelstein†[2], and Carey E. Priebe‡[3]

[1]Department of Statistics, Temple University
[2]Department of Biomedical Engineering, Institute of Computational Medicine, Johns Hopkins University
[3]Department of Applied Mathematics and Statistics, Johns Hopkins University

December 21, 2015

Abstract

Understanding and discovering dependence between multiple properties or measurements of our world is a fundamental task not just in science, but also policy, commerce, and other domains. In the past hundred years, people have developed many different measures of dependence that can be applied in a wide variety of settings. An ideal dependence measure would have the following properties. (1) Strong theoretical support, guaranteeing rejecting independence no matter what the dependence structure is. (2) Strong empirical support on a wide variety of low- and high-dimensional simulation settings. (3) Provides insight into the local scale in which dependency is strongest. (4) Detects dependence when it exists, and fails to detect dependence when it does not exist, on real data. No existing test satisfies all of these properties. We develop a novel dependence statistic and test called "Local Graph Correlation". Briefly, we combine the ideas of distance correlation testing with nearest-neighbor testing, to obtain a nearest neighbor distance correlation test. LGC has all four of the above properties, as demonstrated by extensive theory, simulations, and real data examples. We can therefore use this test in a variety of settings in which previous tests failed to detect signal or provide insight.

Keywords: distance correlation, k-nearest-neighbor, independence test, permutation test

Contents

*cshen@temple.edu
†jovo@jhu.edu
‡cep@jhu.edu

1 Introduction

With the increasing type, size, and dimension of modern data sets, detecting dependency among multiple data sets is one of the most important and fundamental tasks in the big data age. The canonical correlation coefficient Hotelling (1936), Spearman's and Kendall's rank correlation coefficient Kendall (1970), RV coefficient Robert and Escoufier (1976), Procrustes coefficient Gower and Dijksterhuis (2004), information theoretic measures Renyi (1959), and the Mantel test Mantel (1967) have been the traditional tools for this task, but each has its limitations when dealing with the increasingly complex modern data sets, e.g., the Pearson correlation coefficient, RV coefficient and the Mantel test are mostly useful for finding linear relationship and may be zero for nonlinear dependent data sets, while mutual information performs poorly for high-dimensional data.

Many recent methods have been proposed to identify the existence of potential relationships between data sets, including Baringhaus and Franz (2004); Taskinen et al. (2005); Gretton et al. (2005); Szekely et al. (2007); Gretton and Gyorfi (2010); Reshef et al. (2011); Heller et al. (2013); Reimherr and Nicolae (2013); Szekely and Rizzo (2013a,b), etc. In particular, the distance correlation method from Szekely et al. (2007) has gained much popularity in the statistical community, due to its theoretical soundness and good numerical performance in testing independence. A similar method from the machine learning point of view is the kernel-based independence test, which is developed in Gretton et al. (2005); Gretton and Gyorfi (2010); Gretton et al. (2012), and connected to the distance-based method in Sejdinovic et al. (2013).

Despite of current progress in the area, it remains a difficult problem to test dependency on real data; and even the best method in theory may suffer from one or more real challenges underlying the data, such as small sample size, high dimensionality, non-linearity, noise, etc. For example, although distance correlation is consistent against all alternatives for testing independence on Euclidean data, the sample distance correlation (dcorr) under-performs in many high-dimensional or non-linear dependencies for finite-sample testing. The modified distance correlation (mcorr) from Szekely and Rizzo (2013a) adjusts the high-dimensional bias, but is still sub-optimal for non-linear dependencies. In comparison, the HHG statistic developed in Heller et al. (2013) performs much better for testing on non-linear data of small sample size, but it may lose some testing power for certain linear and high-dimensional dependency.

In a complementary literature, nearest-neighbor graphs have been used as a key computational primitive in many statistical approaches, ranging from classification and regression Stone (1977) to

data compression to recommender systems Sarwar et al. (2000). More recently, nearest-neighbor has been an invaluable tool in unfolding nonlinear geometry in many recent development of nonlinear embedding algorithms, including Isomap in Tenenbaum et al. (2000); de Silva and Tenenbaum (2003), LLE in Saul and Roweis (2000); Roweis and Saul (2003), and Laplacien eigenmaps in Belkin and Niyogi (2003), among many others. Furthermore, we have successfully applied joint neighborhood to unfold the non-linearity in multiple data sets in Shen et al. (2015), which shows that a good choice of joint neighborhood can better match the nonlinear data sets.

Most relevant to our work, a number of approaches to two-sample and dependence testing have utilizing nearest-neighbor graphs Barton and David (1966); Friedman and Rafsky (1983); Schilling (1986); Dmcke et al. (2014). These approaches have the advantage of naturally operating on any kind of data, including categorical and structured data, as well as strong theoretical guarantees. Perhaps more importantly, they focus only on local distances, rather than global distances, enabling them to be robust to nonlinear and high-dimensional dependence structures. However, none of the previous nearest-neighbor based methods provided an automatic method for choosing the neighborhood size, therefore leaving a crucial tuning parameter unspecified, and impairing its finite sample performance. Moreover, they largely focused on two-sample testing, rather than dependence testing.

In this paper we propose local graph correlation (LGC), in order to better address those challenges from modern data analysis. By marrying ideas from the distance correlation literature to those from the nearest neighbor literature, and adding some of our own special sauce, we obtain a test better than those in either camp. More specifically, the local test statistic naturally inherits the advantages of the distance correlation, such as being consistent, but also inherits properties of graph dependence structures, such as robustness in high-dimensional dependency. YYY not sure about whether the robustness in hd comes from graph dependence or not? YYY

LGC significantly improves the finite-sample testing power over dcorr, for testing on data sets of non-linearity, noise, and/or small sample size. Those advantages make our new test statistic the best method thus far, for detecting dependency on real data and complex dependencies. Indeed in our comprehensive simulation setting, local distance correlation is able to achieve a superior performance comparing to the global distance correlation and HHG; and in the real data experiment, the local test statistic also returns the lowest p-value for testing dependency between human brain and human characteristics, XXX and fails to detect dependence when they are not there in a set of brain imaging experiments. XXX Thus, we expect LGC to find value in a wide

range of applications. To facilitate, we make all of our code open source and incorporate LGC into FlashR.

2 Results

2.1 Intuition

YYY we need to put the algorithm of LGC and dcorr before the next theoretical subsection? Otherwise the context is not enough for theorems. Or just summarize the theorem here and put next section after the testing procedure? YYY

2.2 Theoretical Properties

In this subsection we present the theoretical advantage of local graph correlation. Note that all proofs and additional propositions are provided in the appendix; and we always assume finite second moments of the joint distribution f_{XY}.

First, local graph correlation is consistent since it is built on distance correlation, which is a consistent test statistics.

Theorem 1. *Local graph correlation is consistent for testing independence against all alternatives, i.e., the testing power* $\beta \to 1$ *as* $n \to \infty$.

Although distance correlation is already consistent against all alternatives, it may not always yield a good finite-sample testing power for a particular dependency type; while our local graph correlation is able to improve the testing power by choosing the best neighborhood for a given joint distribution.

The advantage of using k-nearest-neighbor in distance matrices, lies in its capability to exclude the product of small distances in one data set and large distances in the other data set: after double centering, the centered small distance is usually negative and the centered large distance is usually positive, and including such product reduces the magnitude of local distance covariance under the alternative. While excluding such product by k-nearest-neighbor can increase the magnitude of local distance covariance under the alternative, albeit at the cost of increasing its magnitude under the null as well.

For linear dependency, local graph correlation does not benefit from the above trade-off; but for

nonlinear dependency, $gcorr_{kl}$ for $(k,l) \neq (n,n)$ may enjoy a better finite-sample testing power. We characterize the behaviors in the following two theorems.

Theorem 2. *Suppose $Y = cX$ for a non-zero scalar c, then for any n we always have*

$$\beta(dcorr_n) \geq \beta(gcorr_{kl}) \tag{1}$$

for all $k, l = 2, \ldots, n$, where β is the permutation test power at a given type 1 error α.

Thus local graph correlation is no better than distance correlation under linear dependency.

Theorem 3. *There exists f_{XY}, n and α such that*

$$\beta(gcorr_{kl}) > \beta(dcorr_n) \tag{2}$$

for some $(k,l) \neq (n,n)$, where β is the permutation test power at the type 1 error α.

Thus local graph correlation is better than distance correlation under certain nonlinear dependency.

Note that Theorem 2 and the example used in the proof of Theorem 3 correspond to the linear and quadratic relationship in the simulation. Indeed in Figure 1 we observe the distance correlation has the best empirical testing power for linear dependency, while local graph correlation yields a better power for the quadratic relationship.

3 Numerical Experiments

In this section we show the numerical advantage of local graph correlation via simulations and real data experiments. For simulations on known distributions, we carry out the independence test and report the empirical testing power; and for the real data, we report the empirical p-value based on our main algorithm. The benchmarks are dcorr, mcorr, HHG, and the Mantel test.

Overall, we observe that local graph correlation combines the best aspects of dcorr, mcorr and HHG: it performs similarly to dcorr for dependencies that are close to linear, yields similar or better power than HHG in most nonlinear dependencies, and is robust against high-dimensionality throughout all simulations. For real data testing, its superior performance is reflected in the relatively small p-value.

3.1 Simulations

Here we consider 20 different distributions f_{XY} based on the simulations in Szekely et al. (2007); Simon and Tibshirani (2012); Gorfine et al. (2012); Heller et al. (2013). They consist of various polynomial relationships such as linear and quadratic, a variety of complex nonlinear relationship such as circle, trigonometry, and multiplicative noise; we also include two useful benchmark scenarios, the uncorrelated binomial and an independent relationship.

In Figure 6 we offer a visualization of each dependency, by plotting $\mathcal{X}$ against $\mathcal{Y}$ generated by each pair of (X, Y) at dimension 1 and $n = 1000$ with no noise. Clearly type 1, 3, 8, 9, 18 are either linear dependency or very close to linear, while type 2, 4, 5-7, 10-16, 19-20 are nonlinear dependencies. Note that the uncorrelated binomial without noise concentrates on just three points $(0, 0)$, $(1, -1)$ and $(1, 1)$, and the independent clouds does not have any dependency. More details about the simulation set-up and each distribution can be found in the appendix or the simulation code.

We consider two different scenarios for those 20 distributions: a dimension 1 scenario with increasing sample size, and an increasing dimension scenario with fixed sample size. For the first scenario, we always set $m_X = m_Y = 1$ and plot the power with respect to increasing sample size, so as to observe how fast the testing power of each method converges to 1 for various dependencies; for the second scenario, we fix $n = 100$ and plot the power with respect to increasing m_X, so as to determine how robust each method is for increasing dimension of each dependency.

For either scenario, $\mathcal{X}$ and $\mathcal{Y}$ are generated accordingly, then appropriate level of white noise may be added to $\mathcal{Y}$ depending on the distribution (otherwise certain dependency like perfect linear is too easy), and the sample test statistic can be calculated on the sample data. As described in Section E, we carry out the independence test to estimate the testing power for $r = 10000$ Monte-Carlo replicates at $\alpha = 0.05$. The empirical powers are shown in Figure 1 and Figure **??** for the dimension 1 and the increasing dimension scenario respectively.

For the dimension 1 scenario, one may observe that for dependencies that are close to linear, LGC and dcorr always yield similar testing powers, which are better than HHG and Mantel; for the remaining nonlinear dependencies, HHG is usually much better than dcorr and Mantel, while our LGC performs similarly or even better than HHG in most cases due to its significant improvement over the respective global version. Note that for all distributions other than the independent clouds, the empirical powers eventually increase to 1 as the sample size increases, implying that all meth-

ods are consistent (the only exception is the Mantel test, whose powers stay low in many nonlinear dependencies); and for the independent relationship, all testing powers should be exactly the type 1 error level, which approximately holds for the empirical testing powers.

For the increasing dimension scenario, LGC significantly surpasses all other methods: for dependencies that are close to linear, the powers of both dcorr and LGC deteriorate much slower than others; and for the remaining nonlinear dependencies, LGC is much better than all other methods including the mcorr, due to its capability to better handle non-linearity and high-dimensionality at the same time. Note that a quarter of the distributions (e.g. sine period, square, diamond) cannot be detected by any method at dimension higher than 1, since all testing powers quickly degrade to around α.

To intuitively summarize the simulation performance of each method in all settings, we apply the performance profiles introduced by Dolan and More (2002) to the testing powers, which is an evaluation tool to compare different algorithms throughout all given settings. Suppose there are S methods and T different settings, and we denote the respective powers as β_s^t for $s = 1, \ldots, S$ and $t = 1, \ldots, T$. Then the relative performance for each method is defined as follows:

$$performance_s(x) = \frac{1}{T} \sum_{t=1}^{T} I((\beta_*^t - \beta_s^t) \leq x)$$

where $x \in [0, 1]$ and $\beta_*^t = \max_s \beta_s^t$ denotes the best testing power in the tth setting. Namely x stands for the difference with respect to the best power, and the performance profile of each method equals the proportion of simulations that the method is worse than the best method by no more than x. For example, at $x = 0.1$, LGC has a relative performance of 0.75 if and only if there are 15 out of 20 simulations that LGC is worse than the best method by no more than 0.1 in testing power; the relative performance at $x = 0$ stands for the proportion of simulations that the method has the best power; and the performance profile curve always increases to 1 at $x = 1$. The best method should have a similar or higher curve than others; and we also show the area under curve for each profile profile in the legend, which is a numerical way of viewing the advantage of each method.

In Figure 2 we show the performance profiles at fixed dimension and sample size that are determined by a power threshold, for both the dimension 1 and increasing dimension scenarios: for the dimension 1 scenario, the dimension is always fixed at 1, so the sample size is determined by the first sample size that any method has a power of 0.8 (otherwise pick the largest sample size); and for the increasing dimension scenario, the sample size is already fixed at 100, so we

determine the dimension choice by the first dimension that any method has a power that is lower than 0.5 (otherwise pick the smallest dimension). The advantage of LGC does not change much with respect to the threshold choice, which can be seen from the AUC plot of performance profiles against the power threshold.

We can clearly see from Figure 2 that LGC is indeed the most reliable method in finite-sample testing, in accordance with the individual power plots in Figure 1 and Figure 3: for the dimension 1 scenario the performance profiles of LGC is clearly better than others, and for the increasing dimension scenario the advantage is even larger. Note that HHG is slightly better than dcorr in the performance profiles, because there are more nonlinear distributions than linear in the 20 dependencies, and HHG has a larger advantage for nonlinear dependency than its disadvantage in linear dependency when compared to dcorr; and the Mantel test has the lowest performance profile in both scenarios.

Figure 1: Power of different methods on 20 different one-dimensional simulation settings—estimated on the basis of YYY 10000 Monte-Carlo replicates YYY—including those used in Heller et al. (2013) and Reshef et al. (2011). Each panel shows empirical testing power on the absicca, and sample size on the ordinate. Our method empirically achieves as high or higher power than the previous state of the art approaches for nearly all sample sizes on nearly all problems.

(a) (b) (c) (d)

Figure 2: Quantitative comparisons of the power of the various algorithms across all simulations into a single number. (a) Performance profile plots comparing the different algorithms on all 1-dimensional problems at YYY a fixed sample size where any testing power exceeds the power threshold 0.8 YYY. The legend provides the Area-Under-the-Curve (AUC) for each method; larger is better. (b) AUC for each method sweeping over all different power thresholds. (c) Same as (a) but for the high-dimensional simulations, YYY at a fixed dimension where any testing power drops below the power threshold 0.5 YYY. (d) Same as (b) but for the high-dimensional simulations. It is clear that our method outperforms the previous state of the art, regardless of function, sample size, and dimensionality.

Figure 3: Power of different methods on 20 different simulation settings, for dimensionality ranging from 1 to 1000. Details as in Figure 1. Again, our method empirically achieves as high or higher power than the previous state of the art approaches for nearly all sample sizes on nearly all problems and dimensions.

Figure 4: Testing Power Heat-map of Local Graph Correlation for Increasing Dimension. YYY I replaced figure2 and figure6 by the surface map now, so the below description needs to be changed?YYY Understanding how dependence varies with the local scale of the dependence. For each of the 20 panels, the abscissa shows power and the ordinate is the number of neighbors for XXX X XXX (we actually sweep over all pairs of locality for X and Y, so this plot only shows a line in that plane). Each different simulation yields a different curve, highlighting the importance of understanding local scale in terms of understanding the data.

3.2 Real Data

Here we apply LGC to test independence between brain features and personal characteristics from two different experiments, for which the data sets are relatively small in sample size due to the expensive data collection process.

The first experiment is to detect the relationship between the brain connectome and personality from Adelstein et al. (2011). The sample size is $n = 42$, and each person has a 5 dimensional personality data based on questionnaires and the five-factor personality model. Then the brain activity of each person is measured by fMRI for 197 brain regions and 194 time steps. Thus the brain connectome feature is high-dimensional while the personality data is low-dimensional. There seems to exist certain correlation between the brain activity and personality as experimentally shown in Adelstein et al. (2011), but whether the dependency can be detected from the raw data is the question here.

To apply our method, two distance measures are required for the two different data sources: for the personality data, the distance matrix A is formed by the Euclidean distance directly; for the connectome data, we run a spectrum analysis for each region, bandpass and normalize it, then calculate the Kullback-Leibler divergence among regions and use the normalized Hellinger distance as the distance matrix B. Once the distance matrices A and B are obtained, we apply the permutation test in Section E for $r = 10000$ random permutations, and derive the p-values of

LGC, dcorr, mcorr, and Mantel.

The p-value by LGC is 0.0384, for which the estimated neighborhood choice is $k = 12, l = 6$ based on 10000 MC replicates. No other method yields significant (less than 0.05) p-value, although HHG is quite close: dcorr has a p-value of 0.5863, mcorr has a p-value of 0.3217, HHG has a p-value of 0.0619, and the Mantel test has a p-value of 0.9888. We also show the p-value of LGC with respect to all possible neighborhood choice by heat map in the first plot of FigureřeffigReal, and we can clearly see a local structure in the data that yield significant p-values for adjacent neighborhoods.

Note that if we use LGC based on dcorr rather than mcorr, the p-value becomes 0.4348 achieved at $k = 42, l = 33$, which is no longer significant: this implies a high-dimensional structure in the dependency, which is indeed the case for the connectome data. Also note that the distance transformation (especially for the connectome data) may not be the most appropriate for detecting dependency, so it is possible that HHG and mcorr may yield better p-values for a different transformation.

(a) (b) (c)

Figure 5: P-Value Heat-Map of Local Graph Correlation with respect to Different Neighborhood Choice

Next we carry out the same testing procedure on another experiment regarding brain hippocampus shape and major depressive disorder. There are $n = 114$ subjects, and the brain images of each person are obtained by very high resolution MRI scans on the hippocampus; and we also have available a categorical vector containing the disease information, including clinically depressed subject, high-risk subject, and non-affected subject. There has been evidences that relate major depressive disorder to the hippocampus shape in Park et al. (2008) and Posener et al. (2003), and we would like to test the significance of such relationship in the data.

The brain data is transformed into two dissimilarity matrices LML and LMR, representing the left and right hippocampus data based on landmark matching (see Park et al. (2008) for more details on data processing); and the label vector is transformed into its Euclidean distance matrix D, where $D(i, j) = 0$ if and only if the ith subject has a different label from the jth subject. We consider two permutation tests: testing dependency between LML and D, and testing dependency between LMR and D.

For testing between the left brain and the disease, the p-value of LGC is 0.0010 after 10000 random

permutations, and the estimated neighborhood choice is $k = 105, l = 91$ based on 10000 MC replicates; dcorr has a p-value of 0.0448, mcorr has a p-value of 0.0396, HHG has a p-value of 0.0391, and the Mantel test has a p-value of 0.0471, so all benchmarks have larger p-values than LGC but are still significant at 0.05. Note that the heat map of LGC with respect to different neighborhood choices is available in the second plot of Figure 5; and if we use LGC based on dcorr instead, the p-value of LGC is still 0.0010 at $k = 106, l = 92$ with a very similar heat map.

For testing between the right brain and the disease, the p-value of LGC is 0.0018 with the estimated neighborhood choice being $k = 106, l = 91$; dcorr has a p-value of 0.0750, mcorr has a p-value of 0.0758, HHG has a p-value of 0.0859, and the Mantel test has a p-value of 0.0765, so all benchmarks are not significant at 0.05. Note that the heat map of LGC with respect to different neighborhood choices is available in the third plot of Figure 5; and if we use LGC based on dcorr instead, the p-value of LGC becomes 0.0016 at $k = 105, l = 92$.

We can observe from the heat map of LGC that there is a clear threshold in the local structure most neighborhood choices after certain threshold yields significant p-values, and LGC with mcorr performs similarly as LGC with dcorr. They imply that the dependency between the brain data and the disease data is probably cross-region, close to linear and low-dimensional. Furthermore, the left brain seems to be more correlated with the disease than the right brain, as testing between LML and D yields smaller p-values than testing between LMR and D. YYY cross-region means: if we limit our observation within a small region in this data, there won't be any dependency; and there is a significant jump in p-value after certain neighborhood threshold YYY

Note that the p-value is always 0 for any test statistic when testing independence between LML and LMR, implying strong linear dependency between the left and right brain.

4 Conclusion

In short, we propose local graph correlation to test independence between data sets, which has been shown to be perform well for testing independence on data of small sample size, high-dimensionality, linearity or non-linearity. It not only enjoys theoretical guarantee such as being consistent in testing independence, but also exhibits superior numerical performances in a comprehensive simulation setting and real data experiments, comparing to other popular methods.

Acknowledgment

This work was partially supported by National Security Science and Engineering Faculty Fellowship (NSSEFF), Johns Hopkins University Human Language Technology Center of Excellence (JHU HLT COE), and the XDATA program of the Defense Advanced Research Projects Agency (DARPA) administered through Air Force Research Laboratory contract FA8750-12-2-0303.

A Functions

(a)

Figure 6: Visualization of 20 different at dimension 1 and $n = 1000$ with no noise

B Locality

Figure 7: Testing Power Heat-map of Local Graph Correlation for Dimension 1

C Dependence Measures

C.1 Global Distance Correlation

Suppose we are given two data sets $\mathcal{X} = [X_1, \cdots, X_n] \in \mathcal{R}^{m_X \times n}$ and $\mathcal{Y} = [Y_1, \cdots, Y_n] \in \mathcal{R}^{m_Y \times n}$, where n is the sample size, m_X and m_Y are the dimensions for each data set. Under the classical hypothesis testing framework, we assume that $X_i, i = 1, \ldots, n$ are identically independently distributed (iid) acoording to f_X, similarly $Y_i \overset{iid}{\sim} f_Y$. Throughout the paper, we always assume that X and Y have finite second moments, which is a necessary assumption to guarantee the consistency of distance correlation.

For testing independence between X and Y, the null and the alternative hypothesis are

$$H_0 : f_{XY} = f_X f_Y,$$
$$H_A : f_{XY} \neq f_X f_Y,$$

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.

To test independence by distance correlation on sample data, 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

$$dCov_n(\mathcal{X}, \mathcal{Y}) = \frac{1}{n^2}\sum_{i,j=1}^{n} A^H_{ij}B^H_{ij}, \tag{3}$$

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

$$dVar_n(\mathcal{X}) = \frac{1}{n^2}\sum_{i,j=1}^{n} A^H_{ij}A^H_{ij}$$
$$dVar_n(\mathcal{Y}) = \frac{1}{n^2}\sum_{i,j=1}^{n} B^H_{ij}B^H_{ij}.$$

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

$$dCorr_n(\mathcal{X}, \mathcal{Y}) = \frac{dCov_n(\mathcal{X}, \mathcal{Y})}{\sqrt{dVar_n(\mathcal{X}) \cdot dVar_n(\mathcal{Y})}}, \tag{4}$$

where all of $dCov_n, dVar_n, dCorr_n$ are always non-negative. Note that the $dCov_n/dCorr_n$ above is actually the square of distance covariance/correlation in Szekely et al. (2007); but for simplicity we drop the square in the name throughout this paper.

It is shown in Szekely et al. (2007) that as $n \to \infty$, $dCorr_n(\mathcal{X}, \mathcal{Y}) \to dCorr(X, Y) \geq 0$, where $dCorr(X, Y)$ is the population distance correlation of X and Y defined by their characteristic functions. The population distance correlation is 0 if and only if X and Y are independent, so that the sample distance correlation is a consistent test for independence, i.e., the testing power converges to 1 as n increases, at any fixed type 1 error level. Note that in this paper distance correlation always means the sample statistic rather than the population statistic, unless otherwise mentioned.

C.2 Global Modified Distance Correlation

However, in case of high-dimensional data where the dimension m_X or m_Y increases with the sample size n, the original distance correlation $dCorr_n$ is no longer appropriate. For example, even for independent Gaussian distribution, $dCorr_n \to 1$ as $m_X, m_Y \to \infty$, such that it is no longer

a consistent test in high dimension. This problem is solved by the modified distance correlation proposed in Szekely and Rizzo (2013a):

$$mdCov_n(\mathcal{X},\mathcal{Y}) = \frac{1}{n(n-3)}(\sum_{i\neq j}^{n} A_{ij}^{H*}B_{ij}^{H*} - \frac{2}{n-2}\sum_{i=1}^{n} A_{ii}^{H*}B_{ii}^{H*}), \tag{5}$$

where A_{ij}^{H*} adjusts the entries of A^H by

$$A_{ij}^{H*} = \begin{cases} \frac{n}{n-1}(A_{ij}^H - \frac{A_{ij}}{n}), & \text{if } i \neq j \\ \frac{1}{n-1}(\sum_i A_{ij} - \frac{\sum_{i,j} A_{ij}}{n}), & \text{if } i = j \end{cases}$$

XXX why is there a $\frac{n}{n-1}$ both inside the sum, and another right outside the sum? let's put it inside or outside, but not both? XXX YYY explanation: there is only one $\frac{n}{n-1}$ out side the sum? YYY and similarly for B_{ij}^{H*}. Then $mdVar_n(\mathcal{X})$ can be defined by replacing all B_{ij}^{H*} in Equation 5 by A_{ij}^{H*}, similarly define $mdVar_n(\mathcal{Y})$.

XXX i don't see how mdvar can be less than 0? it is a sum of squares? oh, maybe if the diagonal elements are WAY bigger than the off diagonal estimates? XXX YYY it is always non-negative for n¿=3 and no ties; but it can also be zero. So yes, it is unlikely but still may be less or equals 0. YYY If $mdVar_n(\mathcal{X}) \cdot mdVar_n(\mathcal{Y}) \leq 0$, the modified distance correlation is set to 0; otherwise it is defined as

$$mdCorr_n(\mathcal{X},\mathcal{Y}) = \frac{mdCov_n(\mathcal{X},\mathcal{Y})}{\sqrt{mdVar_n(\mathcal{X}) \cdot mdVar_n(\mathcal{Y})}}. \tag{6}$$

It is shown in Szekely and Rizzo (2013a) that $mdCorr_n(\mathcal{X},\mathcal{Y})$ is an unbiased estimator of the population distance correlation $dCorr(X,Y)$ for all m_X, m_Y, n; and $mdCorr_n$ is approximately normal even if $m_X, m_Y \to \infty$. Thus it is also a consistent test of independence, which works better than the original distance correlation in high-dimension.

XXX converges as what goes to what? n to infy? mx and my to infty? both? one relative to the other?XXX YYY mcorr is an unbiased estimator of the population dcorr regardless of m_X,m_Y, which is not true for the sample dcorr. So the convergence should hold regardless of mx and my? But to avoid confusion I rephrased the above paragraph YYY

XXX mdcov is consistent as m_X and m_Y go to ∞, but does it matter at what rate as a function of n? i'd guess yes? XXX YYY yes but we don't know the exact rate YYY

To summarize this subsection, both the distance correlation and modified distance correlation are great for testing independence of Euclidean data due to their theoretical consistency, with the modified test statistic being more robust against high-dimensional dependency. Indeed it is a

flourishing concept by a series of papers Bakirov et al. (2006); Szekely et al. (2007); Szekely and Rizzo (2009); Bickel and Xu (2009); Kosorok (2009); Remillard (2009); Li et al. (2012); Szekely and Rizzo (2013a,b, 2014); and the test statistic is not limited to the Euclidean metric as shown in Lyons (2013). XXX robust against high-dimensionality? robust is with respect to a model. i don't know what this means. XXX YYY robust against high-dimensional dependency sounds right to you or not? YYY

However, the required sample size for achieving a good testing power very much depends on the type of dependency underlying the given data, e.g. for perfect linear relationship, sample distance correlation usually requires less than 10 points for a permutation test to declare significance; but for some nonlinear relationships like circle, sample distance correlation yields no significance even at $n = 100$. Because real data rarely exhibits perfect linear relationship, and in practice large amount of data may not always be available, a better finite-sample method is of tremendous value: it not only yields a better testing power for the same sample size, but may also requires much less sample data for the permutation test to declare significance, which in turn saves the running time and data collection process.

C.3 Heller, Heller & Gorfine (HHG)

There exists another distance-based method that is consistent and works particularly well for non-linear dependencies, which is called the HHG statistics in Heller et al. (2013). It applies Pearson's chi-square test to ranks of distances within each column rather than directly summing up the products of distances, and is shown to be better than distance correlation and other methods for finite-sample testing of many common nonlinear relationships in Gorfine et al. (2012) and Heller et al. (2013). However, in our numerical simulations HHG seems to fall a bit short when testing against high-dimensional or close to linear dependency, but is otherwise a strong competitor of the global distance correlation.

The other method we use as the benchmark in the numerical section is the Mantel test, which simply applies Pearson's correlation to the distance matrices, see in Mantel (1967). Despite its lack of theoretical guarantee (for example, unlike distance correlation and HHG, it is not consistent against all alternatives), it has been a very popular method so far and commonly used in biology and ecology. In our numerical simulations we will observe that the Mantel test is not consistent for many nonlinear dependencies, and is sub-optimal in almost all types of dependency we consider.

Since there has not been a method that perform well against all possible alternatives in finite-sample testing, it motivates us to propose a local distance correlation that is concurrently robust against small sample size, high-dimensionality, linearity or non-linearity in the following subsection.

D Local Distance Correlation

In this subsection we define the local distance correlation, which is based on k-nearest-neighbor and applicable to both the original distance correlation and the modified distance correlation. From now on, we distinguish the local test statistic as either the local original distance correlation or the local modified distance correlation when they have different behaviors in the context, otherwise we call both as local/global distance correlation since they often share the same properties.

Under the same setting and notation as the global distance correlation in Section C.1, we further sort the distance matrix A within column and denote the ranks as $r(A_{ij})$: 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 $\{A_{ij}, i = 1, \ldots, n \ \& \ i \neq j\}$; for ties, we take the minimum rank among them (this is of importance for local modified distance correlation, see explanation below). Similarly sort the distance matrix B within column and denote the ranks by $r(B_{ij})$.

Then for each $k, l = 1, \ldots, n$, we can calculate a "local" version of distance covariance as

$$dCov_{kl}(\mathcal{X}, \mathcal{Y}) = \frac{1}{n^2} \sum_{i,j=1}^{n} A_{ij}^H B_{ij}^H I(r(A_{ij}) < k) I(r(B_{ij}) < l), \tag{7}$$

and calculate a local version of distance variance as

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

where $I(\cdot)$ is the indicator function. After normalizing, we have a family of distance correlation:

$$dCorr_{kl}(\mathcal{X}, \mathcal{Y}) = \frac{dCov_{kl}(\mathcal{X}, \mathcal{Y})}{\sqrt{dVar_k(\mathcal{X}) \cdot dVar_l(\mathcal{Y})}}. \tag{8}$$

When $k = l$, we simplify the notations to $dCov_k$ and $dCorr_k$; and the family of local test statistics $\{dCorr_{kl}, k, l = 2, \ldots, n\}$ includes the original distance correlation $dCorr_n$ as well. Note that in the family we exclude $dCorr_{1l}$ and $dCorr_{k1}$, because it holds that $dCorr_{1l} = dCorr_{k1} = dCorr_{11}$,

which does not consider any neighbor, merely counts the diagonal terms in the distance matrices, and not meaningful for testing independence.

From now on, local distance correlation refers to the family of the test statistics $\{dCorr_{kl}, k, l = 2, \ldots, n\}$ rather than each individual $dCorr_{kl}$; and the testing power of local distance correlation refers to the best testing power achieved in the family. Since k, l are finite, the optimal testing power always exists; and it is clear that the optimal power and the corresponding best local test statistic are dependent on the joint distribution f_{XY} and may not be unique.

In the same manner, we can apply the above to modified distance covariance:

$$mdCov_{kl}(\mathcal{X}, \mathcal{Y}) = \frac{1}{n(n-3)}(\sum_{i \neq j}^{n} A_{ij}^{H*} B_{ij}^{H*} I(0 < r(A_{ij}) < k) I(0 < r(B_{ij}) < l) - \frac{2}{n-2} \sum_{i=1}^{n} A_{ii}^{H*} B_{ii}^{H*}), \tag{9}$$

then $mdCorr_{kl}$ is obtained by normalizing each $mdCov_{kl}(\mathcal{X}, \mathcal{Y})$ by the square root of $mdVar_k(\mathcal{X}) \cdot mdVar_l(\mathcal{Y})$; 0 if the square root is either 0 or not a real number. It follows that local modified distance correlation refers to the family $\{mdCorr_{kl}, k, l = 2, \ldots, n\}$. Note that by using the minimal rank among ties, the summation of Equation 9 never includes repeated points: as the advantage of modified distance correlation mostly lies in its exclusion of the diagonal distance products, when ties occur in the data, there will be multiple zero distance terms other than the diagonal such that including those repeated points (i.e., points of column rank 0) significantly inflate $mdCov_{kl}$ and negates the advantage of the modification; so this change is necessary for local modified distance correlation to function properly in practice, since our bootstrap procedure in the permutation test almost always produce ties in the re-sampled data.

Because $\{mdCorr_{kl}\}$ has the same asymptotic properties as $\{dCorr_{kl}\}$ other than its robustness against high-dimensionality, we will mostly use $\{dCorr_{kl}\}$ for explanation purposes in Section C. Indeed in the numerical section, local modified distance correlation has similar numerical performance as the local original distance correlation at dimension 1, but much more superior at higher dimensions due to its bias adjustment.

Moreover, testing dependency by local distance correlation can be a more elegant solution than nonlinear embedding: choosing an optimal neighborhood size and an appropriate dimension in nonlinear algorithms can be computationally expensive by cross validation, but the family of local test statistics can be easily computed for all possible neighborhoods. In fact, once the distance matrices A and B are sorted within column, computing $\{dCorr_{kl}\}$ has the same running time as computing any individual $dCorr_{kl}$; and the overall computation always takes $O(n^2 \log(n))$, which

comes from sorting.

E Testing Procedure

XXX fix this section XXX

In this subsection we explain how to test independence by local distance correlation, which is necessary for understanding their theoretical properties in Section 2.2, and also used as the testing procedure in our numerical experiment.

E.1 Main Algorithm by Random Permutation Test

To test independence on given sample data $\mathcal{X}$ and $\mathcal{Y}$ by local graph correlation, we use the following algorithm:

Note that neither the sample data nor the distance measure needs to be Euclidean for the purpose of applying the algorithm, but theoretical consistency may not hold for arbitrary distances, see in Lyons (2013); and the algorithm can start with two distance matrices A and B directly.

In the real data experiment, we report the p-values of LGC by the above algorithm. For all the benchmarks used in the experiment, their p-values are derived by the first two steps of the algorithm by replacing the statistic LGC_{kl} with $dcorr$, $mcorr$, HHG, etc.

E.2 Evaluation by Simulations

We evaluate our algorithm by statistical testing powers on known joint distributions. Ideally, a good algorithm should return high testing powers for strong dependency, and has power close to type 1 error α in the absence of dependency.

When the true joint distribution f_{XY} is known, we can repeatedly generate $\mathcal{X}$ and $\mathcal{Y}$ using f_{XY} by r Monte-Carlo replicates, based on which we can estimate the empirical distribution of LGC_{kl} under the alternative; then we generate $\mathcal{Y}^{''}$ independently from the marginal distribution, from which we can estimate the empirical distribution of LGC_{kl} under the null.

From the above empirical distributions, we estimate the testing power of LGC_{kl} at type 1 error level α, and take the maximal power $\hat{\beta}LGC = \max \hat{\beta}(LGC_{kl}), k, l = 2, \ldots, n$ as the empirical power of

Algorithm 1 Test Independence by Local Graph Correlation

Input: Two data sets $\mathcal{X}$ and $\mathcal{Y}$ of same sample size n but possibly different dimensions. A pre-defined significance level α, and a parameter r used for number of Monte-Carlo replicates.

1. **Distance Transformation**:

   Transform $\mathcal{X}$ and $\mathcal{Y}$ into two $n \times n$ distance matrices A and B.

2. **Estimate the p-value of LGC_{kl} by Permutation Test**:

   First calculate $LGC_{kl}(\mathcal{X}, \mathcal{Y})$ for all $k, l = 2, \ldots, n$ by Equation 8.

   Then for each k, l, calculate

$$p_{kl} = Prob(LGC_{kl}(\mathcal{X}, \mathcal{Y}) > LGC_{kl}(\mathcal{X}, \mathcal{Y}Q)), \tag{10}$$

   where Q denotes a random permutation of size n. Since it is usually infeasible to consider all possible permutations, the above probability is estimated by randomly generating r permutations.

3. **Estimate the optimal neighborhood of LGC by Bootstrap**:

   For each Monte-Carlo replicate, generate $\mathcal{X}^{'}$ from $\mathcal{X}$ by re-sampling with replacement, and generate $\mathcal{Y}^{'}$ from $\mathcal{Y}$ use the same re-sampling, and calculate $LGC_{kl}(\mathcal{X}^{'}, \mathcal{Y}^{'})$ for all $k, l = 2, \ldots, n$ by Equation 8. After r MC replicates, this yields the empirical distribution of LGC_{kl} under the alternative for each k, l.

   For each Monte-Carlo replicates, keep the same re-sampling from above, and generate a random permutation Q. Then calculate $LGC_{kl}(\mathcal{X}^{'}, \mathcal{Y}^{'}Q)$ for all $k, l = 2, \ldots, n$, which yields the empirical distribution of LGC_{kl} under the null for each k, l.

   Based on the above empirical distributions, estimate the testing power of LGC_{kl} at type 1 error level α, and denote it as $\hat{\beta}(LGC_{kl})$. Pick the optimal neighborhood choice by maximizing the empirical power:

$$(k^*, l^*) = \arg\max_{k,l} \{\hat{\beta}(LGC_{kl}), k, l \in [2, \ldots, n]\}, \tag{11}$$

Output: Take $p_{k^*l^*}$ as the p-value of LGC; if there are multiple optimal neighborhood choice, use the mean p-value. Reject the null hypothesis of independence if and only if the p-value of LGC is less than the significance level α.

LGC.

We call the above evaluation procedure as the independence test, which approximates the bootstrap procedure in step 3 of the algorithm: because the empirical distribution of the re-sampled data approximates the true distribution for sufficiently large r, the testing power of LGC based on the independence test approximates the testing power of LGC from the permutation test.

Similarly, we can evaluate dcorr, mcorr, HHG, and the Mantel test by the independence test, by replacing LGC_{kl} with the respective test statistic.

Note that we can also use the permutation test to calculate the testing powers for known distribution, which is much slower than directly applying the independence test. Thus for the simulations of known distribution, we report the empirical testing powers of LGC and other benchmarks by the independence test, which approximates their true testing powers.

YYY I think all the subsections below are now obsolete.YYY

E.3 Independence Test

E.4 Permutation Test

E.4.1 Power Calculations

E.4.2 Bias Adjustment

References

Adelstein, J., Z. Shehzad, M. Mennes, C. DeYoung, X. Zuo, C. Kelly, D. Margulies, A. Bloomfield, J. Gray, F. Castellanos, and M. Milham (2011). Personality is reflected in the brain's intrinsic functional architecture. *PLoS ONE 6*(11), e27633.

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

Baringhaus, L. and C. Franz (2004). On a new multivariate two-sample test. *Journal of multivariate analysis 88*(1), 190–206.

Barton, D. and F. David (1966). The random intersection of two graphs. In *Research Papers in Statistics, Wiley, New York*.

Belkin, M. and P. Niyogi (2003). Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation 15*(6), 1373–1396.

Bickel, P. and Y. Xu (2009). Discussion of: Brownian distance covariance. *Annals of Applied Statistics 3*(4), 1266–1269.

de Silva, V. and J. B. Tenenbaum (2003). Global versus local methods in nonlinear dimensionality reduction. *Advances in Neural Informaiton Processing Systems 15*, 721–728.

Dmcke, S., U. Mansmann, and A. Tresch (2014). A novel test for independence derived from an exact distribution of ith nearest neighbours. *PLOS ONE 9*(10), e107955.

Dolan, E. and J. More (2002). Benchmarking optimization software with performance profiles. *Mathematical Programming 91*(2), 201–213.

Friedman, J. and L. Rafsky (1983). Graph-theoretic measures of multivariate association and prediction. *Annals of Statistics 11*(2), 377–391.

Gorfine, M., R. Heller, and Y. Heller (2012). Comment on detecting novel associations in large data sets. *available at http://ie.technion.ac.il/ gorfinm/files/science6.pdf*.

Gower, J. C. and G. B. Dijksterhuis (2004). *Procrustes Problems*. Oxford University Press.

Gretton, A., K. Borgwardt, M. Rasch, B. Scholkopf, and A. Smola (2012). A kernel two-sample test. *Journal of Machine Learning Research 13*, 723–773.

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

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

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

Hotelling, H. (1936). Relations between two sets of variates. *Biometrika 28*, 321–377.

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

Kosorok, M. (2009). Discussion of: Brownian distance covariance. *Annals of Applied Statistics 3*(4), 1270–1278.

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

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

Mantel, N. (1967). The detection of disease clustering and a generalized regression approach. *Cancer Research 27*(2), 209–220.

Park, Y., C. Priebe, M. Miller, N. Mohan, and K. Botteron (2008). Statistical analysis of twin populations using dissimilarity measurements in hippocampus shape space. *Journal of Biomedicine and Biotechnology*, 694297.

Posener, J., L. Wang, J. Price, M. Gado, M. Province, M. Miller, C. Babb, and J. Csernansky (2003). Statistical analysis of twin populations using dissimilarity measurements in hippocampus shape space. *American Journal of Psychiatry 160*(1), 83–89.

Reimherr, M. and D. Nicolae (2013). On quantifying dependence: A framework for developing interpretable measures. *Statistical Science 28*(1), 116–130.

Remillard, B. (2009). Discussion of: Brownian distance covariance. *Annals of Applied Statistics 3*(4), 1295–1298.

Renyi, A. (1959). On measures of dependence. *Acta Mathematica Academiae Scientiarum Hungarica 10*(3), 441–451.

Reshef, D., Y. Reshef, H. Finucane, S. Grossman, G. McVean, P. Turnbaugh, E. Lander, M. Mitzenmacher, and P. Sabeti (2011). Detecting novel associations in large data sets. *Science 334*(6062), 1518–1524.

Robert, P. and Y. Escoufier (1976). A unifying tool for linear multivariate statistical methods: The rv - coefficient. *Journal of the Royal Statistical Society. Series C 25*(3), 257–265.

Roweis, S. T. and L. K. Saul (2003). Think globally, fit locally: Unsupervised learning of low dimensional manifolds. *Journal of Machine Learning Research 4*, 119–155.

Sarwar, B., G. Karypis, J. Konstan, and J. Riedl (2000). Application of dimensionality reduction in recommender system - a case study. In *ACM WebKDD 2000 Workshop*.

Saul, L. K. and S. T. Roweis (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science 290*, 2323–2326.

Schilling, M. (1986). Multivariate two-sample tests based on nearest neighbors. *Journal of the American Statistical Association 81*(395), 799–806.

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

Shen, C., J. T. Vogelstein, and C. Priebe (2015). Manifold matching using shortest-path distance and joint neighborhood selection. *available at http://arxiv.org/abs/1412.4098*.

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

Stone, C. (1977). Consistent nonparametric regression. *Annals of Statistics 4*(5), 595–620.

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

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

Szekely, G. and M. Rizzo (2013b). Energy statistics: A class of statistics based on distances. *Journal of Statistical Planning and Inference 143*(8), 1249–1272.

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

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

Taskinen, S., H. Oja, and R. Randles (2005). Multivariate nonparametric tests of independence. *Journal of the American Statistical Association 100*(471), 916–925.

Tenenbaum, J. B., V. de Silva, and J. C. Langford (2000). A global geometric framework for nonlinear dimension reduction. *Science 290*, 2319–2323.