From Distance Correlation to Multiscale Graph Correlation

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

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

January 19, 2019

Abstract

Understanding and developing a correlation measure that can detect general dependencies is not only imperative to statistics and machine learning, but also crucial to general scientific discovery in the big data age. In this paper, we establish a new framework that generalizes distance correlation — a correlation measure that was recently proposed and shown to be universally consistent for dependence testing against all joint distributions of finite moments — to the Multiscale Graph Correlation (MGC). By utilizing the characteristic functions and incorporating the nearest neighbor machinery, we formalize the population version of local distance correlations, define the optimal scale in a given dependency, and name the optimal local correlation as MGC. The new theoretical framework motivates a theoretically sound

*shenc@udel.edu
†cep@jhu.edu
‡jovo@jhu.edu

Sample MGC and allows a number of desirable properties to be proved, including the universal consistency, convergence and almost unbiasedness of the sample version. The advantages of MGC are illustrated via a comprehensive set of simulations with linear, nonlinear, univariate, multivariate, and noisy dependencies, where it loses almost no power in monotone dependencies while achieving better performance in general dependencies, compared to distance correlation and other popular methods.

Keywords: testing independence, generalized distance correlation, nearest neighbor graph

1 Introduction

Given pairs of observations $(x_i, y_i) \in \mathbb{R}^p \times \mathbb{R}^q$ for $i = 1, \ldots, n$, assume they are generated by independently identically distributed (*iid*) F_{XY}. A fundamental statistical question prior to the pursuit of any meaningful joint inference is the independence testing problem: the two random variables are independent if and only if $F_{XY} = F_X F_Y$, i.e., the joint distribution equals the product of the marginals. The statistical hypothesis is formulated as:

$$H_0 : F_{XY} = F_X F_Y,$$
$$H_A : F_{XY} \neq F_X F_Y.$$

For any test statistic, the testing power at a given type 1 error level equals the probability of correctly rejecting the null hypothesis when the random variables are dependent. A test is consistent if and only if the testing power converges to 1 as the sample size increases to infinity, and a valid test must properly control the type 1 error level. Modern datasets are often nonlinear, high-dimensional, and noisy, where density estimation and traditional statistical methods fail to be applicable. As multi-modal data are prevalent in much data-intensive research, a powerful, intuitive, and easy-to-use method for detecting general relationships is pivotal.

The classical Pearson's correlation [17] is still extensively employed in statistics, machine learning, and real-world applications. It is an intuitive statistic that quantifies the linear association, a special but extremely important relationship. A recent surge of interests has been placed on using distance metrics and kernel transformations to achieve consistent independence testing against all dependencies. A notable example is the distance correlation (DCORR) [22, 23, 25, 26]: the population DCORR is defined via the characteristic functions of the underlying random variables, while the sample

DCORR can be conveniently computed via the pairwise Euclidean distances of given observations. DCORR enjoys universal consistency against any joint distribution of finite second moments, and is applicable to any metric space of strong negative type [15]. Notably, the idea of distance-based correlation measure can be traced back to the Mantel coefficient [11, 16]: the sample version differs from sample DCORR only in centering, garnered popularity in ecology and biology applications, but does not have the consistency property of DCORR.

Developed almost in parallel from the machine learning community, the kernel-based method (HSIC) [7, 8] has a striking similarity with DCORR: it is formulated by kernels instead of distances, can be estimated on sample data via the sample kernel matrix, and is universally consistent when using any characteristic kernel. Indeed, it is shown in [20] that there exists a mapping from kernel to metric (and vice versa) such that HSIC equals DCORR. Another competitive method is the Heller-Heller-Gorfine method (HHG) [9, 10]: it is also universally consistent by utilizing the rank information and the Pearson's chi-square test, but has better finite-sample testing powers over DCORR in a collection of common nonlinear dependencies. There are other consistent methods available, such as the COPULA method that tests independence based on the empirical copula process [3, 4, 12], entropy-based methods [2], and methods tailored for univariate data [18].

As the number of observations in many real world problems (e.g., genetics and biology) are often limited and very costly to increase, finite-sample testing power is crucial for certain data exploration tasks: DCORR has been shown to perform well in monotone relationships, but not so well in nonlinear dependencies such as circles and parabolas; the performance of HSIC and HHG are often the opposite of DCORR, which perform slightly inferior to DCORR in monotone relationships but excel in various nonlinear dependencies.

From another point of view, unraveling the nonlinear structure has been intensively studied in the manifold learning literature [1, 19, 27]: by approximating a linear manifold locally via the k-nearest neighbors at each point, these nonlinear techniques can produce better embedding results than linear methods (like PCA) in nonlinear data. The main downside of manifold learning often lies in the parameter choice, i.e., the number of neighbor or the correct embedding dimension is often hard to estimate and requires cross-validation. Therefore, assuming a satisfactory neighborhood size can be efficiently determined in a given nonlinear relationship, the local correlation measure shall work better than the global correlation measure; and if the parameter selection is sufficiently adaptive, the optimal local correlation shall equal the global correlation in linear relationships.

In this manuscript we formalize the notion of population local distance correlations and MGC, explore their theoretical properties both asymptotically and in finite-sample, and propose an improved Sample MGC algorithm. By combing distance correlation with the locality principle, MGC inherits the universal consistency in testing, is able to efficiently search over all local scales and determine the optimal correlation, and enjoys the best testing powers throughout the simulations. A number of real data applications via MGC are pursued in [28], e.g., testing brain images versus personality and disease, identify potential protein biomarkers for cancer, etc. And MGC are employed for vertex dependence testing and screening in [13, 29].

The paper is organized as follows: In Section 2, we define the population local distance correlation and population MGC via the characteristic functions of the underlying random variables and the nearest neighbor graphs, and show how the local variants are related to the distance correlation. In Section 3, we consider the sample local correlation on finite-samples, prove its convergence to the population version, and discuss the centering and ranking scheme. In Section 4, we present a thresholding-based algorithm for Sample MGC, prove its convergence property, propose a theoretically sound threshold choice, manifest that MGC is valid and consistent under the permutation test, and finish the section with a number of fundamental properties for the local correlations and MGC. The comprehensive simulations in Section 5 exhibits the empirical advantage of MGC, and the paper is concluded in Section 6. All proofs and detailed simulation setups are in the Appendix, and the code are available on Github [1] and CRAN [2].

2 Multiscale Graph Correlation for Random Variables

2.1 Distance Correlation Review

We first review the original distance correlation in [26]. A non-negative weight function $w(t,s)$ on $(t,s) \in \mathbb{R}^p \times \mathbb{R}^q$ is defined as:

$$w(t,s) = (c_p c_q |t|^{1+p} |s|^{1+q})^{-1},$$

where $c_p = \frac{\pi^{(1+p)/2}}{\Gamma((1+p)/2)}$ is a non-negative constant tied to the dimensionality p, and $\Gamma(\cdot)$ is the complete Gamma function. Then the population distance covariance, variance and

[1] https://github.com/neurodata/mgc-matlab

[2] https://CRAN.R-project.org/package=mgc

correlation are defined by

$$\begin{aligned} dCov(X,Y) &= \int_{\mathbb{R}^p \times \mathbb{R}^q} |E(g_{XY}(t,s)) - E(g_X(t))E(g_Y(s))|^2 w(t,s) dt ds, \\ dVar(X) &= dCov(X,X), \\ dVar(Y) &= dCov(Y,Y), \\ dCorr(X,Y) &= \frac{dCov(X,Y)}{\sqrt{dVar(X) \cdot dVar(Y)}}, \end{aligned}$$

where $|\cdot|$ is the complex modulus, $g_{\cdot}(\cdot)$ denotes the exponential transformation within the expectation of the characteristic function, i.e., $g_{XY}(t,s) = e^{\mathbf{i}\langle t,X\rangle + \mathbf{i}\langle s,Y\rangle}$ (**i** represents the imaginary unit) and $E(g_{XY}(t,s))$ is the characteristic function. Note that distance variance equals 0 if and only if the random variable is a constant, in which case distance correlation shall be set to 0. The main property of population DCORR is the following.

Theorem. *For any two random variables (X,Y) with finite first moments, $dCorr(X,Y) = 0$ if and only if X and Y are independent.*

To estimate the population version on sample data, the sample distance covariance is computed by double centering the pairwise Euclidean distance matrix of each data, followed by summing over the entry-wise product of the two centered distance matrices. When the underlying random variables have finite second moments, the sample DCORR is shown to converge to the population DCORR , and is thus universally consistent for testing independence against all joint distributions of finite second moments.

2.2 Population Local Correlations

Next we formally define the population local distance covariance, variance, correlation by combining the k-nearest neighbor graphs with the distance covariance. For simplicity, they are named the local covariance, local variance, and local correlation from now on, and we always assume the following regularity conditions:

$1)$ (X, Y) have finite second moments,

$2)$ Neither random variable is a constant,

$3)$ (X, Y) are continuous random variables.

The finite second moments assumption is required by DCORR, and also required by the local version to establish convergence and consistency. The non-constant condition is to avoid the trivial case and make sure population local correlations behave well. The continuous assumption is for ease of presentation, so the definition and related properties can be presented in a more elegant manner. Indeed, for any discrete random variable one can always apply jittering (i.e., add trivial white noise) to make it continuous without altering the independence testing.

Definition. *Suppose* $(X, Y), (X', Y'), (X'', Y''), (X''', Y''')$ *are* iid *as* F_{XY}*. Let* $\boldsymbol{I}(\cdot)$ *be the indicator function, define two random variables*

$$\boldsymbol{I}^{\rho_k}_{X,X'} = \boldsymbol{I}(\int_{B(X,\|X'-X\|)} dF_X(u) \leq \rho_k)$$
$$\boldsymbol{I}^{\rho_l}_{Y',Y} = \boldsymbol{I}(\int_{B(Y',\|Y'-Y\|)} dF_Y(u) \leq \rho_l)$$

with respect to the closed balls $B(X, \|X' - X\|)$ *and* $B(Y', \|Y - Y'\|)$ *centered at* X *and* Y' *respectively. Then let* $\bar{\cdot}$ *denote the complex conjugate, define*

$$h^{\rho_k}_X(t) = (g_X(t)\overline{g_{X'}(t)} - g_X(t)\overline{g_{X''}(t)})\boldsymbol{I}^{\rho_k}_{X,X'}$$
$$h^{\rho_l}_{Y'}(s) = (g_{Y'}(s)\overline{g_Y(s)} - g_{Y'}(s)\overline{g_{Y'''}(s)})\boldsymbol{I}^{\rho_l}_{Y',Y}$$

as functions of $t \in \mathbb{R}^p$ *and* $s \in \mathbb{R}^q$ *respectively,* *The population local covariance, variance, correlation at any* $(\rho_k, \rho_l) \in [0,1] \times [0,1]$ *are defined as*

$$dCov^{\rho_k,\rho_l}(X,Y) = \int_{\mathbb{R}^p \times \mathbb{R}^q} \{E(h_X^{\rho_k}(t)\overline{h_{Y'}^{\rho_l}(s)}) - E(h_X^{\rho_k}(t))E(\overline{h_{Y'}^{\rho_l}(s)})\}w(t,s)dtds, \qquad (1)$$

$$dVar^{\rho_k}(X) = dCov^{\rho_k,\rho_k}(X,X),$$

$$dVar^{\rho_l}(Y) = dCov^{\rho_l,\rho_l}(Y,Y),$$

$$dCorr^{\rho_k,\rho_l}(X,Y) = \frac{dCov^{\rho_k,\rho_l}(X,Y)}{\sqrt{dVar^{\rho_k}(X) \cdot dVar^{\rho_l}(Y)}}, \qquad (2)$$

where we limit the domain of population local correlation to

$$\mathcal{S}_\epsilon = \left\{(\rho_k, \rho_l) \in [0,1] \times [0,1] \textit{ that satisfies } \min\{dVar^{\rho_k}(X), dVar^{\rho_l}(Y)\} \geq \epsilon\right\}$$

for a small positive ϵ *that is no larger than* $\min\{dVar(X), dVar(Y)\}$*.*

The domain of local correlation needs to be limited so the population version is well-behaved. For example, when X is a constant or $\rho_k = 0$, $dVar^{\rho_k}(X)$ equals 0 and the corresponding local correlation is not well-defined. All subsequent analysis for the population local correlations is based on the domain $\mathcal{S}_\epsilon$, which is non-empty and compact as shown in Theorem 3. In practice, it suffices to set ϵ as any small positive number, see the sample version in Section 3. Also note that in the indicator function, the two random variables and the distribution $F(u)$ after the differential symbol are independent, e.g., at any realization (x, x') of (X, X'), the first indicator equals $\boldsymbol{I}(\int_{B(x,\|x'-x\|)} dF_X(u) \leq \rho_k)$. Then its expectation is taken with respect to (X, X').

The above definition makes use of the characteristic functions, which is akin to the original definition of DCORR and easier to show consistency. Alternatively, the local covariance can be equivalently defined via the pairwise Euclidean distances. The alternative definition better motivates the sample version in Section 3, is often handy for understanding and proving theoretical properties, and suggests that local covariance is always a real number, which is not directly obvious from Equation 1.

Theorem 1. *Suppose* $(X, Y), (X', Y'), (X'', Y''), (X''', Y''')$ *are* iid *as* F_{XY}*, and define*

$$\begin{aligned} d_X^{\rho_k} &= (\|X - X'\| - \|X - X''\|)\boldsymbol{I}_{X,X'}^{\rho_k} \\ d_{Y'}^{\rho_l} &= (\|Y' - Y\| - \|Y' - Y'''\|)\boldsymbol{I}_{Y',Y}^{\rho_l} \end{aligned}$$

The local covariance in Equation 1 can be equally defined as

$$dCov^{\rho_k,\rho_l}(X, Y) = E(d_X^{\rho_k} d_{Y'}^{\rho_l}) - E(d_X^{\rho_k})E(d_{Y'}^{\rho_l}), \tag{3}$$

which shows that local covariance, variance, correlation are always real numbers.

Each local covariance is essentially a local version of distance covariance that truncates large distances at each point in the support, where the neighborhood size is determined by (ρ_k, ρ_l). In particular, distance correlation equals the local correlation at the maximal scale, which will ensure the consistency of MGC.

Theorem 2. *At any* $(\rho_k, \rho_l) \in \mathcal{S}_\epsilon$*,* $dCov^{\rho_k,\rho_l}(X, Y) = 0$ *when* X *and* Y *are independent. Moreover, at* $(\rho_k, \rho_l) = (1, 1)$*,* $dCov^{\rho_k,\rho_l}(X, Y) = dCov(X, Y)$*. They also hold for the correlations by replacing all the* $dCov$ *by* $dCorr$*.*

2.3 Population MGC and Optimal Scale

The population MGC can be naturally defined as the maximum local correlation within the domain, i.e.,

$$c^*(X, Y) = \max_{(\rho_k,\rho_l)\in\mathcal{S}_\epsilon} \{dCorr^{\rho_k,\rho_l}(X, Y)\}, \tag{4}$$

and the scale that attains the maximum is named the optimal scale

$$(\rho_k, \rho_l)^* = \arg \max_{(\rho_k,\rho_l)\in\mathcal{S}_\epsilon} \{dCorr^{\rho_k,\rho_l}(X,Y)\}. \tag{5}$$

The next theorem states the continuity of the local covariance, variance, correlation, and thus the existence of population MGC.

Theorem 3. *Given two continuous random variables (X, Y),*

(a) *The local covariance is a continuous function with respect to $(\rho_k, \rho_l) \in [0,1]^2$, so is local variance in $[0,1]$ and local correlation in $\mathcal{S}_\epsilon$.*

(b) *The set $\mathcal{S}_\epsilon$ is always non-empty unless either random variable is a constant.*

(c) *Excluding the trivial case in (b), the set $\{dCorr^{\rho_k,\rho_l}(X,Y), (\rho_k,\rho_l) \in \mathcal{S}_\epsilon\}$ is always non-empty and compact, so an optimal scale $(\rho_k,\rho_l)^*$ and $c^*(X,Y)$ exist.*

Therefore, population MGC and the optimal scale exist, are distribution dependent, and may not be unique. Without loss of generality, the optimal scale is assumed unique for presentation purpose. The population MGC is always no smaller than DCORR in magnitude, and equals 0 if and only if independence, a property inherited from DCORR.

Theorem 4. *When X and Y are independent, $c^*(X,Y) = dCorr(X,Y) = 0$; when X and Y are not independent, $c^*(X,Y) \geq dCorr(X,Y) > 0$.*

3 Sample Local Correlations

Sample DCORR can be easily calculated via properly centering the Euclidean distance matrices, and is shown to converge to the population DCORR [23, 25, 26]. Similarly, we show that the sample local correlation can be calculated via the Euclidean distance matrices upon truncating large distances for each sample observation, and the sample version converges to the respective population local correlation.

3.1 Definition

Given pairs of observations $(x_i, y_i) \in \mathbb{R}^p \times \mathbb{R}^q$ for $i = 1, \ldots, n$, denote $\mathcal{X}_n = [x_1, \ldots, x_n]$ as the data matrix with each column representing one sample observation, and similarly $\mathcal{Y}_n$. Let $\tilde{A}$ and $\tilde{B}$ be the $n \times n$ Euclidean distance matrices of $\mathcal{X}_n = \{x_i\}$ and $\mathcal{Y}_n = \{y_i\}$ respectively, i.e., $\tilde{A}_{ij} = \|x_i - x_j\|$. Then we compute two column-centered matrices A and B with the diagonals excluded, i.e., $\tilde{A}$ and $\tilde{B}$ are centered within each column such that

$$A_{ij} = \begin{cases} \tilde{A}_{ij} - \frac{1}{n-1}\sum_{s=1}^{n} \tilde{A}_{sj}, & \text{if } i \neq j, \\ 0, & \text{if } i = j; \end{cases} \qquad B_{ij} = \begin{cases} \tilde{B}_{ij} - \frac{1}{n-1}\sum_{s=1}^{n} \tilde{B}_{sj}, & \text{if } i \neq j, \\ 0, & \text{if } i = j; \end{cases} \tag{6}$$

Next we define $\{R^A_{ij}\}$ as the “rank” of x_i relative to x_j, that is, $R^A_{ij} = k$ if x_i is the k^{th} closest point (or “neighbor”) to x_j, as determined by ranking the set $\{\tilde{A}_{1j}, \tilde{A}_{2j}, \ldots, \tilde{A}_{nj}\}$ by ascending order. Similarly define R^B_{ij} for the y's. As we assumed (X, Y) are continuous, with probability 1 there is no repeating observation and the ranks always take value in $\{1, \ldots, n\}$. In practice ties may occur, and we recommend either using minimal rank to keep the ties or jittering to break the ties, which is discussed at the end of this section.

For any $(k, l) \in [n]^2 = \{1, \ldots, n\} \times \{1, \ldots, n\}$, we define the rank truncated matrices A^k and B^l as

$$A^k_{ij} = A_{ij}\boldsymbol{I}(R^A_{ij} \leq k),$$
$$B^l_{ij} = B_{ij}\boldsymbol{I}(R^B_{ij} \leq l).$$

Let $\circ$ denote the entry-wise product, $\hat{E}(\cdot) = \frac{1}{n(n-1)} \sum_{i \neq j}^{n} (\cdot)$ denote the diagonal-excluded sample mean of a square matrix, then the sample local covariance, variance, and correlation are defined as:

$$
\begin{aligned}
dCov^{k,l}(\mathcal{X}_n, \mathcal{Y}_n) &= \hat{E}(A^k \circ B^{l'}) - \hat{E}(A^k)\hat{E}(B^l), \\
dVar^k(\mathcal{X}_n) &= \hat{E}(A^k \circ A^{k'}) - \hat{E}^2(A^k), \\
dVar^l(\mathcal{Y}_n) &= \hat{E}(B^l \circ B^{l'}) - \hat{E}^2(B^l), \\
dCorr^{k,l}(\mathcal{X}_n, \mathcal{Y}_n) &= dCov^{k,l}(\mathcal{X}_n, \mathcal{Y}_n) / \sqrt{dVar^k(\mathcal{X}_n) \cdot dVar^l(\mathcal{Y}_n)}.
\end{aligned}
$$

If either local variance is smaller than a preset $\epsilon > 0$ (e.g., the smallest positive local variance among all), then we set the corresponding $dCorr^{k,l}(\mathcal{X}_n, \mathcal{Y}_n) = 0$ instead. Note that once the rank is known, sample local correlations can be iteratively computed in $\mathcal{O}(n^2)$ rather than a naive implementation of $\mathcal{O}(n^3)$. A detailed running time comparison is presented in Section 5.

In case of ties, minimal rank offers a consecutive indexing of sample local correlations, e.g., if Y only takes two values, R_{ij}^B takes value in $\{1, 2\}$ under minimal rank, but maximal rank yields $\{\frac{n}{2}, n\}$. The sample local correlations are not affected by the tie scheme, but minimal rank is more convenient to work with for implementation purposes. Alternatively, one can break ties deterministically or randomly, e.g., apply jittering to break all ties. For example, in the Bernoulli relationship of Figure 1, there are only three points for computing sample local correlations and the Sample MGC equals 0.9. If white noise of variance 0.01 were added to the data, we break all ties and obtain a much larger number of sample local correlations. The resulting Sample MGC is 0.8, which is slightly smaller but still much larger than 0 and implies a strong dependency.

Whether the random variable is continuous or discrete, and whether the ties in sample data are broken or not, does not affect the theoretical results except in certain theorem statements. For example, in Theorem 5, the convergence still holds for discrete random variables, but the index pair (k,l) does not necessarily correspond to the population version at $(\rho_k, \rho_l) = (\frac{k-1}{n-1}, \frac{l-1}{n-1})$, e.g., when X is Bernoulli with probability 0.8 and minimal rank is used, $k=1$ corresponds to $\rho_k = 0.8$ instead of $\rho_k = \frac{k-1}{n-1}$. Nevertheless, Theorem 5 and all results in the paper hold regardless of continuous or discrete random variables, but the presentation is more elegant for the continuous case.

3.2 Convergence Property

The sample local covariance, variance, correlation are designed to converge to the respective population versions. Moreover, the expectation of sample local covariance equals the population counterpart up to a difference of $\mathcal{O}(\frac{1}{n})$, and the variance diminishes at the rate of $\mathcal{O}(\frac{1}{n})$.

Theorem 5. *Suppose each column of $\mathcal{X}_n$ and $\mathcal{Y}_n$ are generated* iid *from $(X,Y) \sim F_{XY}$. The sample local covariance satisfies*

$$\begin{aligned} E(dCov^{k,l}(\mathcal{X}_n, \mathcal{Y}_n)) &= dCov^{\rho_k,\rho_l}(X,Y) + \mathcal{O}(1/n) \\ Var(dCov^{k,l}(\mathcal{X}_n, \mathcal{Y}_n)) &= \mathcal{O}(1/n) \\ dCov^{k,l}(\mathcal{X}_n, \mathcal{Y}_n) &\overset{n\to\infty}{\to} dCov^{\rho_k,\rho_l}(X,Y), \end{aligned}$$

where $\rho_k = \frac{k-1}{n-1}$ and $\rho_l = \frac{l-1}{n-1}$. In particular, the convergence is uniform and also holds for the local correlation, i.e., for any ϵ there exists n_ϵ such that for all $n > n_\epsilon$,

$$|dCorr^{k,l}(\mathcal{X}_n, \mathcal{Y}_n) - dCorr^{\rho_k,\rho_l}(X,Y)| < \epsilon$$

for any pair of $(\rho_k, \rho_l) \in \mathcal{S}_\epsilon$.

The convergence property ensures that Theorem 2 holds asymptotically for the sample version.

Corollary 1. *For any (k,l), $dCorr^{k,l}(\mathcal{X}_n,\mathcal{Y}_n) \rightarrow 0$ when X and Y are independent. In particular, $dCorr^{n,n}(\mathcal{X}_n,\mathcal{Y}_n) \rightarrow dCorr(X,Y)$.*

Moreover, one can show that $dCorr^{n,n}(\mathcal{X}_n,\mathcal{Y}_n) \approx dCorr(\mathcal{X}_n,\mathcal{Y}_n)$ for the unbiased sample distance correlation in [25] up-to a small difference of $\mathcal{O}(\frac{1}{n})$, which can be verified by comparing Equation 6 to Equation 3.1 in [25].

3.3 Centering and Ranking

To combine distance testing with the locality principle, other than the procedure proposed in Equation 3, there are a number of alternative options to center and rank the distance matrices. For example, letting

$$
\begin{aligned}
d_X^{\rho_k} &= (\|X-X'\| - \|X-X''\| - \|X'-X''\| + \|X''-X'''\|)\boldsymbol{I}_{X,X'}^{\rho_k},\\
d_{Y'}^{\rho_l} &= (\|Y'-Y\| - \|Y'-Y''\| - \|Y-Y''\| + \|Y''-Y'''\|)\boldsymbol{I}_{Y',Y}^{\rho_l}
\end{aligned}
$$

still guarantees the resulting local correlation at maximal scale equals the distance correlation; and letting

$$
\begin{aligned}
d_X^{\rho_k} &= \|X-X'\|\boldsymbol{I}_{X,X'}^{\rho_k},\\
d_{Y'}^{\rho_l} &= \|Y'-Y\|\boldsymbol{I}_{Y',Y}^{\rho_l},
\end{aligned}
$$

makes the resulting local correlation at maximal scale equal the MANTEL coefficient, the earliest distance-based correlation coefficient.

Nevertheless, the centering and ranking strategy proposed in Equation 3 is more faithful to k-nearest neighbor graph: the indicator $\boldsymbol{I}_{X,X'}^{\rho_k}$ equals 1 if and only if $\int_{B(X,\|X'-X\|)} dF_X(u) \leq$

ρ_k, which happens with probability ρ_k. Viewed another way, when conditioned on $(X, X') = (x, x')$, the indicator equals 1 if and only if $Prob(\|x' - x\| < \|X'' - x\|) \leq \rho_k$, thus matching the column ranking scheme in Equation 6. Indeed, the locality principle used in [1, 19, 27] considers the k-nearest neighbors of each sample point in local computation, an essential step to yield better nonlinear embeddings.

On the centering side, the MANTEL test appears to be an attractive option due to its simplicity in centering. All the DCORR, HHG, HSIC have their theoretical consistency, while the MANTEL coefficient does not, despite it being merely a different centering of DCORR. An investigation of the population form of MANTEL yields some additional insights:

Definition. *Given $\mathcal{X}_n$ and $\mathcal{Y}_n$, the* MANTEL *coefficient on sample data is computed as*

$$M(\mathcal{X}_n, \mathcal{Y}_n) = \hat{E}(\tilde{A} \circ \tilde{B}) - \hat{E}(\tilde{A})\hat{E}(\tilde{B})$$

$$Mantel(\mathcal{X}_n, \mathcal{Y}_n) = \frac{M(\mathcal{X}_n, \mathcal{Y}_n)}{\sqrt{M(\mathcal{X}_n, \mathcal{X}_n)M(\mathcal{Y}_n, \mathcal{Y}_n)}},$$

where $\tilde{A}_{ij}$ and $\tilde{B}_{ij}$ are the pairwise Euclidean distance, and $\hat{E}(\cdot) = \frac{1}{n(n-1)}\sum_{i \neq j}^{n}(\cdot)$ is the diagonal-excluded sample mean of a square matrix.

Corollary 2. *Suppose each column of $\mathcal{X}_n$ and $\mathcal{Y}_n$ are* iid *as F_{XY}, and $(X, Y), (X', Y')$ are also* iid *as F_{XY}. Then*

$$Mantel(\mathcal{X}_n, \mathcal{Y}_n) \to Mantel(X, Y) = \frac{M(X, Y)}{\sqrt{M(X, X)M(Y, Y)}},$$

where

$$\begin{aligned} M(X, Y) &= \int_{\mathbb{R}^p \times \mathbb{R}^q} \{|E(g_{XY}(t, s))|^2 - |E(g_X(t))E(g_Y(s))|^2\} w(t, s) dt ds \\ &= E(\|X - X'\|\|Y - Y'\|) - E(\|X - X'\|)E(\|Y - Y'\|)) \\ &= Cov(\|X - X'\|, \|Y - Y'\|). \end{aligned}$$

Corollary 2 suggests that MANTEL is actually a two-sided test based on the absolute difference of characteristic functions: under certain dependency structure, the MANTEL coefficient can be negative and still imply dependency (i.e., $|E(g_{XY}(t,s))| < |E(g_X(t))E(g_Y(s))|$); whereas population DCORR and MGC are always no smaller than 0, and any negativity of the sample version does not imply dependency. Therefore, MANTEL is only appropriate as a two-sided test, which is evaluated in Section 5.

Another insight is that MANTEL, unlike DCORR, is not universally consistent: due to the integral w, one can construct a joint distribution such that the population MANTEL equals 0 under dependence (see Remark 3.13 in [15] for an example of dependent random variables with uncorrelated distances). However, empirically, simple centering is still effective in a number of common dependencies (like two parabolas and diamond in Figure 3).

4 Sample MGC and Estimated Optimal Scale

A naive sample version of MGC can be defined as the maximum of all sample local correlations

$$\max_{(k,l)\in[n]^2}\{dCorr^{k,l}(\mathcal{X}_n,\mathcal{Y}_n)\}.$$

Although the convergence to population MGC can be guaranteed, the sample maximum is a biased estimator of the population MGC in Equation 4. For example, under independence, population MGC equals 0, while the maximum sample local correlation has expectation larger than 0, which may negate the advantage of searching locally and hurt the testing power.

This motivates us to compute Sample MGC as a smoothed maximum within the largest connected region of thresholded local correlations. The purpose is to mitigate the bias of a direct maximum, while maintaining its advantage over DCORR in the test statistic. The idea is that in case of dependence, local correlations on the grid near the optimal scale shall all have large correlations; while in case of independence, a few local correlations may happen to be large, but most nearby local correlations shall still be small. The idea can be similarly adapted whenever there are multiple correlated test statistics or multiple models available, for which taking a direct maximum may yield too much bias [13]. From another perspective, Sample MGC is like taking a regularized maximum.

4.1 **Sample** MGC

The procedure is as follows:

Input: A pair of datasets $(\mathcal{X}_n, \mathcal{Y}_n)$.

Compute the Local Correlation Map: Compute all local correlations:
$\{dCorr^{k,l}(\mathcal{X}_n, \mathcal{Y}_n), (k,l) \in [n]^2\}$.

Thresholding: Pick a threshold $\tau_n \geq 0$, denote $LC(\cdot)$ as the operation of taking the largest connected component, and compute the largest region R of thresholded local correlations:

$$R = LC(\{(k,l) \text{ such that } dCorr^{k,l}(\mathcal{X}_n, \mathcal{Y}_n) > \max\{\tau_n, dCorr^{n,n}(\mathcal{X}_n, \mathcal{Y}_n)\}\}). \quad (7)$$

Within the region R, set

$$c^*(\mathcal{X}_n, \mathcal{Y}_n) = \max_{(k,l)\in R}\{dCorr^{k,l}(\mathcal{X}_n, \mathcal{Y}_n)\} \quad (8)$$

$$(k_n, l_n)^* = \arg\max_{(k,l)\in R}\{dCorr^{k,l}(\mathcal{X}_n, \mathcal{Y}_n)\} \quad (9)$$

as the Sample MGC and the estimated optimal scale. If the number of elements in R is less than $2n$, or the above thresholded maximum is no more than $dCorr^{n,n}(\mathcal{X}_n, \mathcal{Y}_n)$, we instead set $c^*(\mathcal{X}_n, \mathcal{Y}_n) = dCorr^{n,n}(\mathcal{X}_n, \mathcal{Y}_n)$ and $(k_n, l_n)^* = (n, n)$.

Output: Sample MGC $c^*(\mathcal{X}_n, \mathcal{Y}_n)$ and the estimated optimal scale $(k_n, l_n)^*$.

If there are multiple largest regions, e.g., R_1 and R_2 where their number of elements are more than $2n$ and coincide with each other, then it suffices to let $R = R_1 \cup R_2$ and locate the MGC statistic within the union. The selection of at least $2n$ elements for R is an empirical choice, which balances the bias-variance trade-off well in practice. The parameter can be any positive integer without affecting the validity and consistency of the test. But if the parameter is too large, MGC tends to be more conservative and is unable to detect signals in strongly nonlinear relationships (e.g., trigonometric functions), and performs closer and closer to DCORR; if the parameter is set to a very small fixed number, the bias is inflated so MGC tends to perform similarly as directly maximizing all local correlations.

4.2 Convergence and Consistency

The proposed Sample MGC is algorithmically enforced to be no less than the local correlation at the maximal scale, and also no more than the maximum local correlation. It also ensures in Theorem 4 to hold for the sample version.

Theorem 6. *Regardless of the threshold τ_n, the Sample* MGC *statistic $c^*(\mathcal{X}_n, \mathcal{Y}_n)$ satisfies*

(a) *It always holds that*

$$\max_{(k,l)\in[n]^2}\{dCorr^{k,l}(\mathcal{X}_n,\mathcal{Y}_n)\} \geq c^*(\mathcal{X}_n,\mathcal{Y}_n) \geq dCorr^{n,n}(\mathcal{X}_n,\mathcal{Y}_n).$$

(b) *When X and Y are independent, $c^*(\mathcal{X}_n,\mathcal{Y}_n) \rightarrow 0$; when X and Y are not independent, $c^*(\mathcal{X}_n,\mathcal{Y}_n) \rightarrow$ a positive constant.*

The next theorem states that if the threshold τ_n converges to 0, then whenever population MGC is larger than population DCORR, Sample MGC is also larger than sample DCORR asymptotically; otherwise if the threshold does not converge to 0, Sample MGC may equal sample DCORR despite of the first moment advantage in population. Moreover, Sample MGC indeed converges to population MGC when the optimal scale is in the largest thresholded region R. The empirical advantage of Sample MGC is illustrated in Figure 1.

Theorem 7. *Suppose each column of $\mathcal{X}_n$ and $\mathcal{Y}_n$ are* iid *as continuous $(X,Y) \sim F_{XY}$, and the threshold choice $\tau_n \rightarrow 0$ as $n \rightarrow \infty$.*

(a) *Assume that $c^*(X,Y) > Dcorr(X,Y)$ under the joint distribution. Then $c^*(\mathcal{X}_n,\mathcal{Y}_n) > Dcorr(\mathcal{X}_n,\mathcal{Y}_n)$ for n sufficiently large.*

(b) *Assume there exists an element within the the largest connected area of $\{(\rho_k,\rho_l) \in \mathcal{S}_\epsilon$ with $dCorr^{\rho_k,\rho_l}(X,Y) > dCorr(X,Y)\}$, such that the the local correlation of that element equals $c^*(X,Y)$. Then $c^*(\mathcal{X}_n,\mathcal{Y}_n) \rightarrow c^*(X,Y)$.*

Alternatively, Theorem 7(b) can be stated that the Sample MGC always converges to the maximal population local correlation within the largest connected area of thresholded local correlations. Therefore, Sample MGC converges either to DCORR (when the area is empty) or something larger, thus improving over DCORR statistic in first moment.

4.3 Choice of Threshold

The choice of threshold τ_n is imperative for Sample MGC to enjoy a good finite-sample performance, especially at small sample size. According to Theorem 7, the threshold shall converge to 0 for Sample MGC to prevail sample DCORR.

A model-free threshold τ_n was previously used in [28]: for the following set

$$\{dCorr^{k,l}(\mathcal{X}_n, \mathcal{Y}_n) \text{ s.t. } dCorr^{k,l}(\mathcal{X}_n, \mathcal{Y}_n) < 0\},$$

let σ^2 be the sum of all its elements squared, and set $\tau_n = 5\sigma$ as the threshold; if there is no negative local correlation and the set is empty, use $\tau_n = 0.05$.

Although the previous threshold is a data-adaptive choice that works pretty well empirically and does not affect the consistency of Sample MGC in Theorem 8, it does not converge to 0. The following finite-sample theorem from [23] motivates an improved threshold choice here:

Theorem. *Under independence of (X, Y), assume the dimensions of X are exchangeable with finite variance, and so are the dimensions of Y. Then for any $n \geq 4$ and $v = \frac{n(n-3)}{2}$, as p, q increase the limiting distribution of $(dCorr^{n,n}(\mathcal{X}_n, \mathcal{Y}_n) + 1)/2$ equals the symmetric Beta distribution with shape parameter $\frac{v-1}{2}$.*

The above theorem leads to the new threshold choice:

Corollary 3. *Denote $v = \frac{n(n-3)}{2}$, $z \sim Beta(\frac{v-1}{2})$, $F_z^{-1}(\cdot)$ as the inverse cumulative distribution function. The threshold choice*

$$\tau_n = 2F_z^{-1}\Big(1 - \frac{0.02}{n}\Big) - 1$$

converges to 0 as $n \to \infty$.

The limiting null distribution of DCORR is still a good approximation even when p, q are not large, thus provides a reliable bound for eliminating local correlations that are larger than DCORR by chance or by noise. The intuition is that Sample MGC is mostly useful when it is much larger than DCORR in magnitude, which is often the case in non-monotone relationships as shown in Section 5 Figure 1. Alternatively, directly setting $\tau_n = 0$ also guarantees the theoretical properties and works equally well when the sample size n is moderately large.

4.4 Permutation Test

To test independence on a pair of sample data $(\mathcal{X}_n, \mathcal{Y}_n)$, the random permutation test has been the popular choice [5] for almost all methods introduced, as the null distribution of the test statistic can be easily approximated by randomly permuting one data set. We discuss the computation procedure, prove the testing consistency of MGC, and analyze the running time.

To compute the p-value of MGC from the permutation test, first compute the Sample MGC statistic $c^*(\mathcal{X}_n, \mathcal{Y}_n)$ on the observed data pair. Then the MGC statistic is repeatedly computed on the permuted data pair, e.g. $\mathcal{Y}_n = [y_1, \ldots, y_n]$ is permuted into $\mathcal{Y}_n^{\pi} = [y_{\pi(1)}, \ldots, y_{\pi(n)}]$ for a random permutation π of size n, and compute $c^*(\mathcal{X}_n, \mathcal{Y}_n^{\pi})$. The permutation procedure is repeated for r times to estimate the probability $Prob(c^*(\mathcal{X}_n, \mathcal{Y}_n^{\pi}) > c^*(\mathcal{X}_n, \mathcal{Y}_n))$, and the estimated probability is taken as the p-value of MGC. The independence hypothesis is rejected if the p-value is smaller than a pre-set critical level, say 0.05 or 0.01. The following theorem states that MGC via the permutation test is consistent and valid.

Theorem 8. *Suppose each column of $\mathcal{X}_n$ and $\mathcal{Y}_n$ are generated* iid *from F_{XY}. At any* *type 1 error level $\alpha > 0$, Sample* MGC *is a valid test statistic that is consistent against all possible alternatives under the permutation test.*

4.5 Miscellaneous Properties

In this subsection, we first show a useful lemma expressing sample local covariance in Section 3.1 by matrix trace and eigenvalues, then list a number of fundamental and desirable properties for the local variance, local correlation, and MGC, akin to these of Pearson's correlation and distance correlation as shown in [22, 26].

Lemma 1. *Denote $tr(\cdot)$ as the matrix trace, $\lambda_i[\cdot]$ as the ith eigenvalue of a matrix, and J as the matrix of ones of size n. Then the sample covariance equals*

$$\begin{aligned} dCov^{k,l}(\mathcal{X}_n, \mathcal{Y}_n) &= tr(A^k B^l) - tr(A^k J) tr(B^l J) \\ &= tr[(A^k - tr(A^k J)J)(B^l - tr(B^l J)J)] \\ &= \sum_{i=1}^{n} \lambda_i[(A^k - tr(A^k J)J)(B^l - tr(B^l J)J)]. \end{aligned}$$

Theorem 9 (Local Variances)**.** *For any random variable $X \sim F_X \in \mathbb{R}^p$, and any $\mathcal{X}_n \in \mathbb{R}^{p \times n}$ with each column* iid *as F_X,*

(a) *Population and sample local variances are always non-negative, i.e.,*

$$\begin{aligned} dVar^{\rho_k}(X) &\geq 0 \\ dVar^{k}(\mathcal{X}_n) &\geq 0 \end{aligned}$$

at any $\rho_k \in [0, 1]$ and any $k \in [n]$.

(b) *$dVar^{\rho_k}(X) = 0$ if and only if either $\rho_k = 0$ or F_X is a degenerate distribution;*

$dVar^{k}(\mathcal{X}_n) = 0$ if and only if either $k = 1$ or F_X is a degenerate distribution.

(c) *For two constants $v \in \mathbb{R}^p, u \in \mathbb{R}$, and an orthonormal matrix $Q \in \mathbb{R}^{p \times p}$,*

$$dVar^{\rho_k}(v + uQX) = u^2 \cdot dVar^{\rho_k}(X)$$
$$dVar^{k}(v^T J + u\mathcal{X}_n Q) = u^2 \cdot dVar^{k}(\mathcal{X}_n).$$

Therefore, the local variances end up having properties similar to the distance variance in [26], except the distance variance definition there takes a square root.

Theorem 10 (Local Correlations and MGC)**.** *For any pair of random variable $(X, Y) \sim F_{XY} \in \mathbb{R}^p \times \mathbb{R}^q$, and any $(\mathcal{X}_n, \mathcal{Y}_n) \in \mathbb{R}^{p \times n} \times \mathbb{R}^{q \times n}$ with each column* iid *as F_{XY},*

(a) *Symmetric and Boundedness:*

$$dCorr^{\rho_k,\rho_l}(X, Y) = dCorr^{\rho_l,\rho_k}(Y, X) \in [-1, 1]$$
$$dCorr^{k,l}(\mathcal{X}_n, \mathcal{Y}_n) = dCorr^{l,k}(\mathcal{Y}_n, \mathcal{X}_n) \in [-1, 1]$$

at any $(\rho_k, \rho_l) \in (0, 1]^2$ and any $(k, l) \in [2, \ldots, n]^2$.

(b) *Assume F_X is non-degenerate. Then at any $\rho_k > 0$, $dCorr^{\rho_k,\rho_k}(X, Y) = 1$ if and only if (X, uY) are dependent via an isometry for some non-zero constant $u \in \mathbb{R}$.*

Assume F_X is non-degenerate. Then at any $k > 1$, $dCorr^{k,k}(\mathcal{X}_n, \mathcal{Y}_n) = 1$ if and only if (X, uY) are dependent via an isometry for some non-zero constant $u \in \mathbb{R}$.

(c) *Both population and Sample* MGC *are symmetric and bounded:*

$$c^*(X, Y) = c^*(Y, X) \in [-1, 1]$$
$$c^*(\mathcal{X}_n, \mathcal{Y}_n) = c^*(\mathcal{Y}_n, \mathcal{X}_n) \in [-1, 1].$$

(d) *Assume F_X is non-degenerate. Then $c^*(X, Y) = 1$ if and only if (X, uY) are dependent via an isometry for some non-zero constant $u \in \mathbb{R}$.*

Assume F_X is non-degenerate. Then $c^*(\mathcal{X}_n, \mathcal{Y}_n) = 1$ if and only if (X, uY) are dependent via an isometry for some non-zero constant $u \in \mathbb{R}$.

The proof of Theorem 10(b)(d) also shows that the local correlations and MGC cannot be -1.

5 Experiments

In the experiments, we compare Sample MGC with DCORR, PEARSON, MANTEL, HSIC, HHG, and COPULA test on 20 different simulation settings based on a combination of simulations used in previous works [6, 21, 26]. Among the 20 settings, the first 5 are monotonic relationships (and several of them are linear or nearly so), the last simulation is an independent relationship, and the remaining settings consist of common non-monotonic and strongly nonlinear relationships. The exact distributions are shown in Appendix.

The Sample Statistics

Figure 1 shows the sample statistics of MGC, DCORR, and PEARSON for each of the 20 simulations in a univariate setting. For each simulation, we generate sample data $(\mathcal{X}_n, \mathcal{Y}_n)$ at $p = q = 1$ and $n = 100$ without any noise, then compute the sample statistics. From type $1-5$, the test statistics for both MGC and DCORR are remarkably greater than 0 and almost identical to each other. For the nonlinear relationships (type $6-19$), MGC benefits from searching locally and achieves a larger test statistic than DCORR's, which can be very small in these nonlinear relationships. For type 20, the test statistics for both MGC and DCORR are almost 0 as expected. On the other hand, PEARSON's test statistic is large whenever there exists certain linear association, and almost 0 otherwise. The comparison of sample statistics indicate that DCORR may have inferior finite-sample testing power in nonlinear relationships, but a strong dependency signal is actually hidden in a local structure that MGC may recover.

Finite-Sample Testing Power

Figure 2 shows the finite-sample testing power of MGC, DCORR, and PEARSON for a linear and a quadratic relationship at $n = 20$ and $p = q = 1$ with white noise (controlled by a constant). The testing power of MGC is estimated as follows: we first generate dependent sample data $(\mathcal{X}_n, \mathcal{Y}_n)$ for $r = 10,000$ replicates, compute Sample MGC for each replicate to estimate the alternative distribution of MGC. Then we generate independent sample data $(\mathcal{X}_n, \mathcal{Y}_n)$ using the same marginal distributions for $r = 10,000$ replicates, compute Sample MGC to estimate the null distribution, and estimate the testing power at type 1 error level $\alpha = 0.05$. The testing power of DCORR is estimated in the same manner, while the testing power of PEARSON is directly computed via the t-test. MGC has the best power in the quadratic relationship, while being almost identical to DCORR and PEARSON in the linear relationship.

The same phenomenon holds throughout all the simulations we considered, i.e., MGC achieves almost the same power as DCORR in monotonic relationships, while being able to improve the power in monotonic and strongly nonlinear relationships. The testing power of MGC versus all other methods are shown in Figure 3 for the univariate settings, and we plot the power versus the sample size from 5 to 100 for each simulation. Note that the noise level is tuned for each dependency for illustration purposes.

Figure 4 compares the testing performance for the same 20 simulations with a fixed

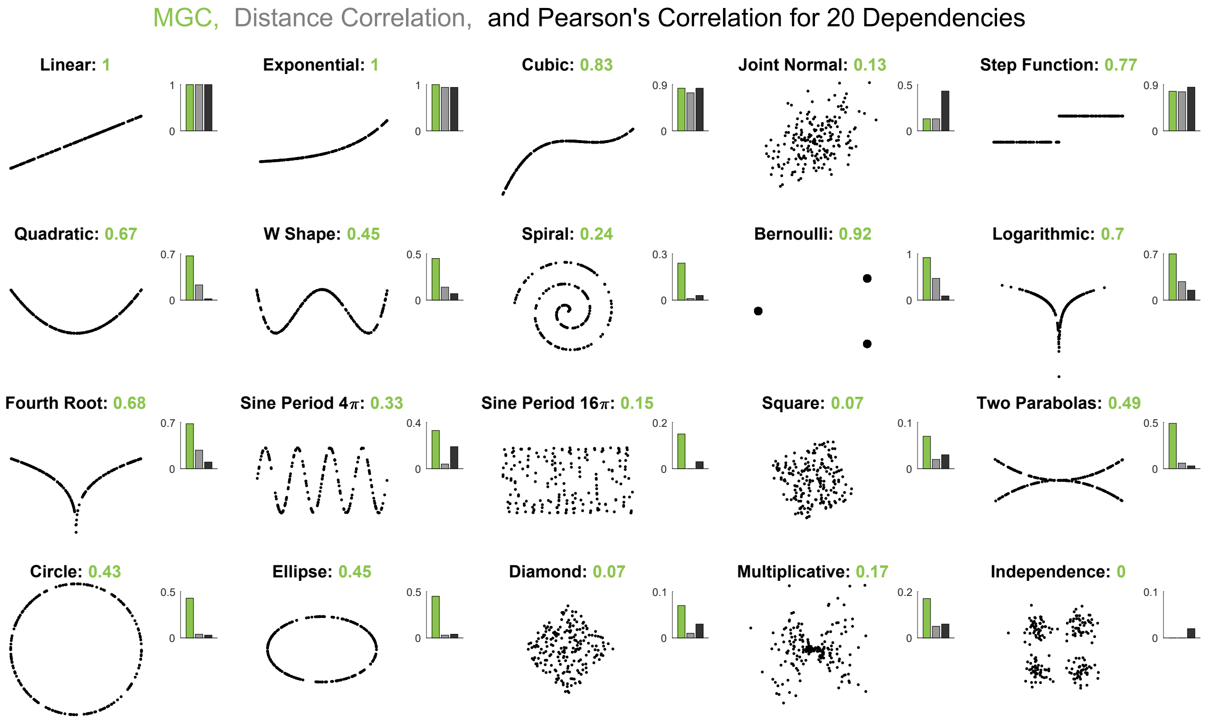

Figure 1: For each panel, a pair of dependent $(\mathcal{X}_n, \mathcal{Y}_n)$ at $n = 100$ and $p = q = 1$ is generated and visualized; the accompanying color bar compares MGC (green), DCORR (gray), and PEARSON in the absolute value (black), all of which lie in the range of $[0, 1]$ with 0 indicating no relationship. MGC yields a non-zero sample correlation for each dependency, while being almost 0 under independence. In comparison, the distance correlation can be close to 0 for common nonlinear dependencies, while the Pearson's correlation only measures linear association and cannot capture nonlinear dependencies. The Sample MGC statistic is shown above each panel.

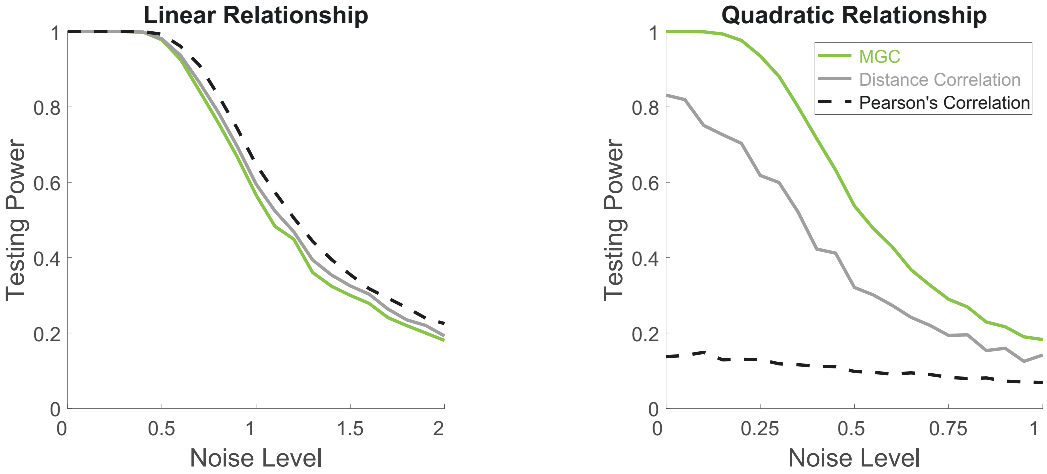

Figure 2: Comparing the power of MGC, DCORR, and PEARSON in noisy linear relationship (left), and noisy quadratic relationship (right). For the linear relationship at $n = 20$ and $p = q = 1$, all three methods are almost the same with PEARSON being slightly higher power; for the quadratic relationship, MGC has a much higher power than DCORR and PEARSON. The phenomenon is consistent throughout the remaining dependent simulations: for testing in monotonic relationships, PEARSON, DCORR, and MGC almost coincide with each other; for strongly nonlinear relationships, MGC almost always supersedes DCORR, and DCORR is better than PEARSON.

sample size $n = 100$ and increasing dimensionality. The relative powers in the univariate and multivariate settings are then summarized in Figure 5. MGC is overall the most powerful method, followed by HHG and HSIC. Since non-monotone relationships are prevalent among the 20 settings, it is not a surprise that DCORR is overall worse than HHG and HSIC, both of which also excel at nonlinear relationships.

Note that the same 20 simulations were also used in [28] for evaluation purposes. The main difference is that the Sample MGC algorithm is now based on the improved threshold with theoretical guarantee. Comparing to the previous algorithm, the new threshold

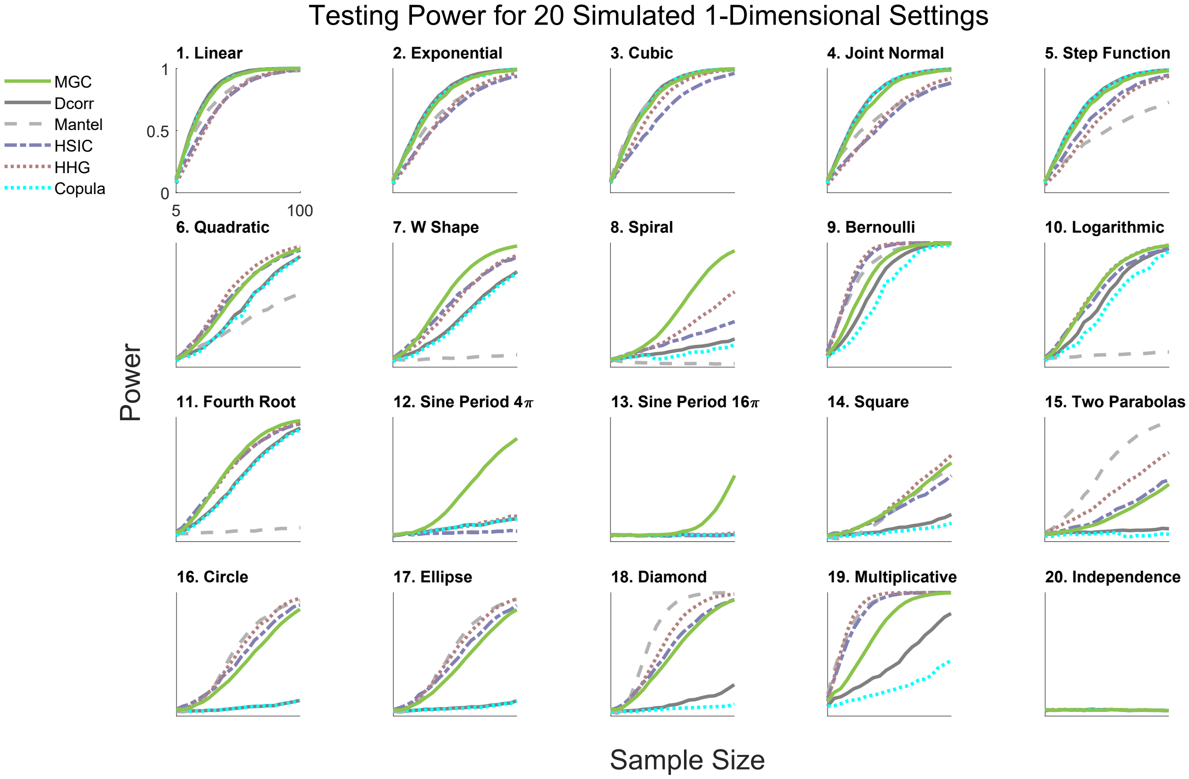

Figure 3: Comparing the testing power of MGC, DCORR, MANTEL, HSIC, HHG, and COPULA. for 20 different univariate simulations. Estimated via $10,000$ replicates of repeatedly generated dependent and independent sample data, each panel shows the estimated testing power at the type 1 error level $\alpha = 0.05$ versus sample sizes ranging from $n = 5$ to 100. Excluding the independent simulation (#20) where all methods yield power 0.05, MGC exhibits the highest or nearly highest power in most dependencies. Note that we only show the ticks for the first panel, because they are the same for every panel, i.e., the x-axis always ranges from 5 to 100 while the y-axis always ranges from 0 to 1.

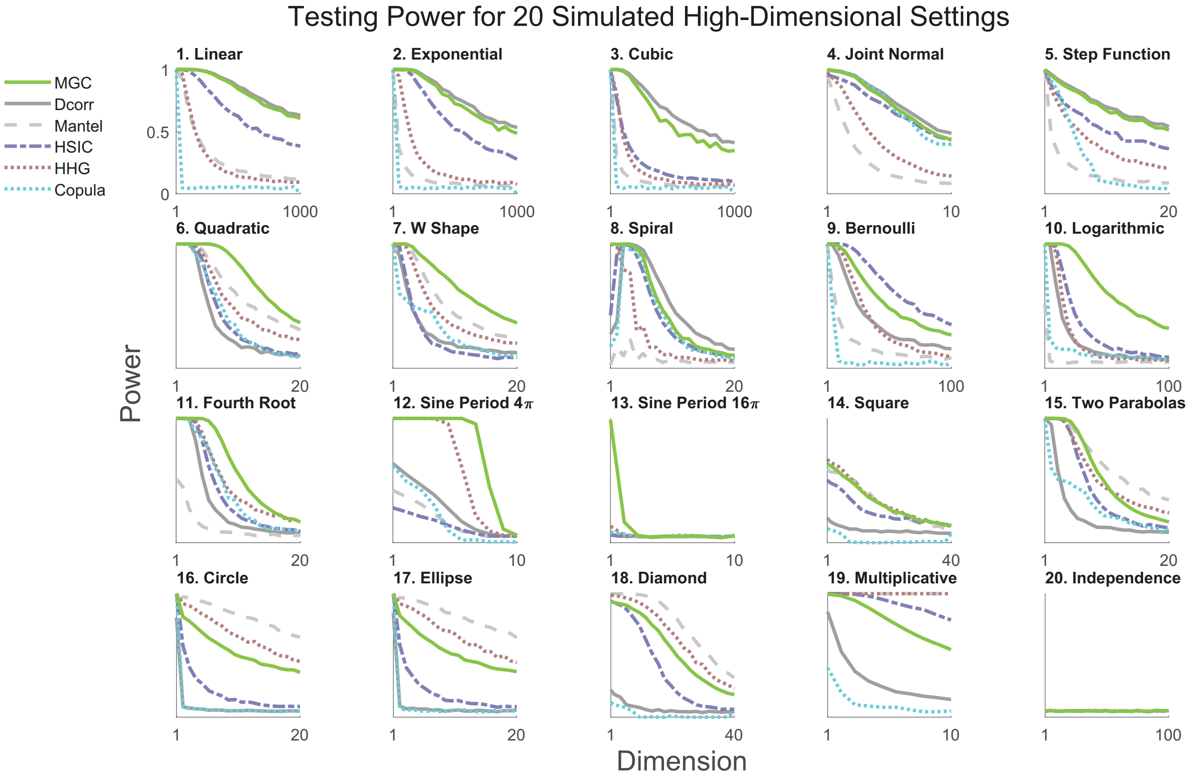

Figure 4: The testing power computed in the same procedure as in Figure 3, except the 20 simulations are now run at fixed sample size $n = 100$ and increasing dimensionality p. Again, MGC empirically achieves similar or higher power than the previous popular approaches for all dimensions on most settings. The ticks for y axis is only shown in the first panel, as the power has the same range in $[0, 1]$ for every panel.

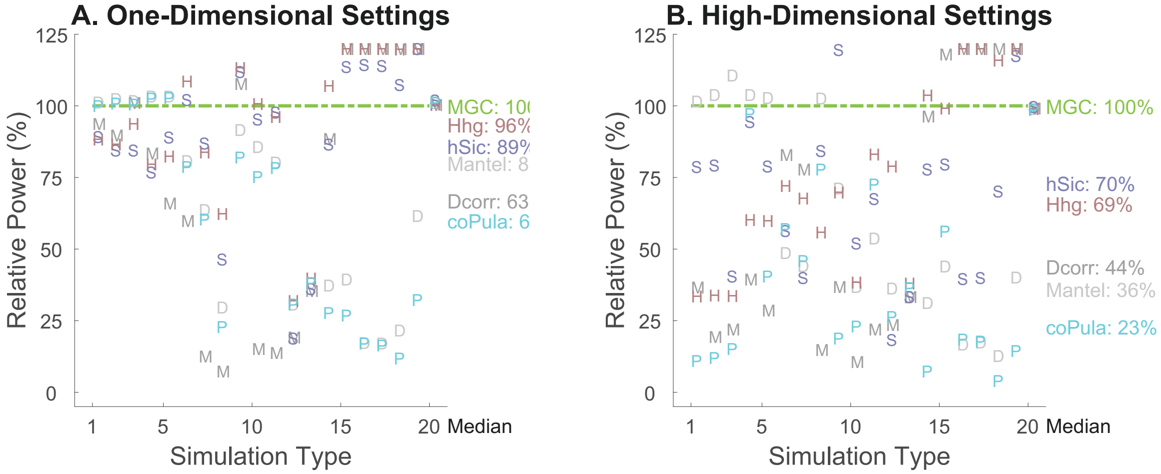

Figure 5: The relative Power of MGC to other methods for testing the 20 simulations under one-dimensional and high-dimensional scenarios. (Left) For each simulation type, we average the testing power of each method in Figure 3 over the sample size, then divide each average power by the average power of MGC. The last column (which also serves as the legend) shows the median power among all relative powers of type $1-19$. The same for the right panel, except it averages over the dimensionality in Figure 4. The relative power percentage indicates that MGC is a very powerful method for finite-sample testing.

slightly improves the testing power in monotonic relationships (the first 5 simulations).

Running Time

Sample MGC can be computed and tested in the same running time complexity as distance correlation: Assume p is the maximum feature dimension of the two datasets, distance computation and centering takes $\mathcal{O}(n^2p)$, the ranking process takes $\mathcal{O}(n^2 \log n)$, all local covariances and correlations can be incrementally computed in $O(n^2)$ (the pseudocode is shown in [28]), the thresholding step of Sample MGC takes $O(n^2)$ as well.

Overall, Sample MGC can be computed in $\mathcal{O}(n^2 \max\{\log n, p\})$. In comparison, the HHG statistic requires the same complexity as MGC, while distance correlation saves on the $\log n$ term.

As the only part of MGC that has the additional $\log n$ term is the column-wise ranking process, a multi-core architecture can reduce the running time to $\mathcal{O}(n^2 \max\{\log n, p\}/T)$. By making $T = \log(n)$ (T is no more than 30 at 1 billion samples), MGC effectively runs in $\mathcal{O}(n^2 p)$ and is of the same complexity as DCORR. The permutation test multiplies another r to all terms except the distance computation, so overall the MGC testing procedure requires $\mathcal{O}(n^2 \max\{r, p\})$, which is the same as DCORR, HHG, and HSIC. Figure 6 shows that MGC has approximately the same complexity as DCORR, and is slower by a constant in the actual running time.

6 Conclusion

In this paper, we formalize the population version of local correlation and MGC, connect them to the sample counterparts, prove the convergence and almost unbiasedness from the sample version to the population version, as well as a number of desirable properties for a well-defined correlation measure. In particular, population MGC equals 0 and the sample version converges to 0 if and only if independence, making Sample MGC valid and consistent under the permutation test. Moreover, Sample MGC is designed in a computationally efficient manner, and the new threshold choice achieves both theoretical and empirical improvements. The numerical experiments confirm the empirical advantages of MGC in a wide range of linear, nonlinear, high-dimensional dependencies.

There are many potential future avenues to pursue. Theoretically, proving when

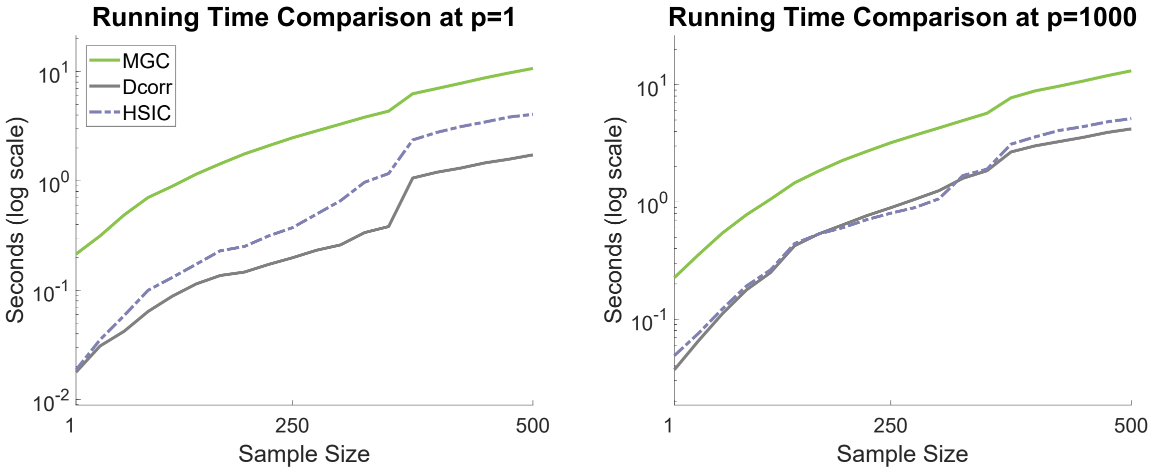

Figure 6: Compute the test statistics of MGC, DCORR, and HSIC for 100 replicates, then plot the average running time in log scale (clocked using Matlab 2017a on a Windows 10 machine with I7 six-core CPU). The sample data is repeatedly generated using the quadratic relationship in Appendix, the sample size increases from 25 to 500, and the dimensionality is fixed at $p = 1$ on the left and $p = 1000$ on the right. In either panel, the three lines differ by some constants in the log scale, suggesting the same running time complexity but different constants. MGC has a higher intercept than the other two, which translates to about a constant of 6 times of DCORR and 3 times of HSIC at $n = 500$ and $p = 1$, and about 3 at $p = 1000$. Note that the increase in p has a relatively small effect in the running time, because the dimensionality p takes part only in the distance matrix computation and is thus relatively cheap.

and how one method dominates another in testing power is highly desirable. As the methods in comparison have distinct formulations and different properties, it is often difficult to compare them directly. However, a relative efficiency analysis may be viable when limited to methods of similar properties, such as DCORR and HSIC, or local statistic and global statistic. In terms of the locality principle, the geometric meaning of the local scale in MGC is intriguing — for example, does the family of local correlations fully characterize the joint distribution, and what is the relationship between the optimal local scale and the dependency geometry — answering these questions may lead to further improvement of MGC, and potentially make the family of local correlations a valuable tool beyond testing.

Method-wise, there are a number of alternative implementations that may be pursued. For example, the sample local correlations can be defined via ϵ ball instead of nearest neighbor graphs, i.e., truncate large distances based on absolute magnitude instead of the nearest neighbor graph. The maximization and thresholding mechanism may be further improved, e.g., thresholding based on the covariance instead of correlation, or design a better regularization scheme. There are many alternative approaches that can maintain consistency in this framework, and it will be interesting to investigate a better algorithm. In particular, we name our method as "multiscale graph correlation" because the local correlations are computed via the k-nearest neighbor graphs, which is one way to generalize the distance correlation.

Application-wise, the MGC method can directly facilitate new discoveries in many kinds of scientific fields, especially data of limited sample size and high-dimensionality such as in neuroscience and omics [28]. Within the domain of statistics and machine learning, MGC can be a very competitive candidate in any methodology that requires a well-defined dependency measure, e.g., variable selection [14], time series [30], etc.

Moreover, the very idea of locality may improve other types of distance-based tests, such as the energy distance for K-sample testing [24].

References

[1] Belkin, M. and P. Niyogi (2003). Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation 15*(6), 1373–1396. 5, 16

[2] Dionsio, A., R. Menezes, and D. A. Mendes (2006). Entropy-based independence test. *Nonlinear Dynamics 44*, 351357. 4

[3] Genest, C., J.-F. Quessy, and B. Rmillard (2006). Local efficiency of a cramer-von mises test of independence. *Journal of Multivariate Analysis 97*, 274–294. 4

[4] Genest, C., J.-F. Quessy, and B. Rmillard (2007). Asymptotic local efficiency of cramer-von mises tests for multivariate independence. *The Annals of Statistics 35*, 166–191. 4

[5] Good, P. (2005). *Permutation, Parametric, and Bootstrap Tests of Hypotheses*. Springer. 22

[6] Gorfine, M., R. Heller, and Y. Heller (2012). Comment on detecting novel associations in large data sets. *Technical Report*. 25

[7] Gretton, A. and L. Gyorfi (2010). Consistent nonparametric tests of independence. *Journal of Machine Learning Research 11*, 1391–1423. 4

[8] 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. 4

[9] Heller, R., Y. Heller, and M. Gorfine (2013). A consistent multivariate test of association based on ranks of distances. *Biometrika 100*(2), 503–510. 4

[10] Heller, R., Y. Heller, S. Kaufman, B. Brill, and M. Gorfine (2016). Consistent distribution-free k-sample and independence tests for univariate random variables. *Journal of Machine Learning Research 17*(29), 1–54. 4

[11] Josse, J. and S. Holmes (2013). Measures of dependence between random vectors and tests of independence. *http://arxiv.org/abs/1307.7383*. 4

[12] Kojadinovic, I. and M. Holmes (2009). Tests of independence among continuous random vectors based on cramr-von mises functionals of the empirical copula process. *Journal of Multivariate Analysis 100*, 1137–1154. 4

[13] Lee, Y., C. Shen, C. E. Priebe, and J. T. Vogelstein (2019). Network dependence testing via diffusion maps and distance-based correlations. *Biometrika*. 5, 18

[14] Li, R., W. Zhong, and L. Zhu (2012). Feature screening via distance correlation learning. *Journal of American Statistical Association 107*, 1129–1139. 34

[15] Lyons, R. (2013). Distance covariance in metric spaces. *Annals of Probability 41*(5), 3284–3305. 4, 17

[16] Mantel, N. (1967). The detection of disease clustering and a generalized regression approach. *Cancer Research 27*(2), 209–220. 4

[17] Pearson, K. (1895). Notes on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London 58*, 240–242. 3

[18] 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. 4

[19] Saul, L. K. and S. T. Roweis (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science 290*, 2323–2326. 5, 16

[20] 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. 4

[21] Simon, N. and R. Tibshirani (2012). Comment on detecting novel associations in large data sets. *http://arxiv.org/abs/1401.7645*. 25

[22] Szekely, G. and M. Rizzo (2009). Brownian distance covariance. *Annals of Applied Statistics 3*(4), 1233–1303. 3, 23

[23] Szekely, G. and M. Rizzo (2013a). The distance correlation t-test of independence in high dimension. *Journal of Multivariate Analysis 117*, 193–213. 3, 11, 21

[24] 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. 35

[25] Szekely, G. and M. Rizzo (2014). Partial distance correlation with methods for dissimilarities. *Annals of Statistics 42*(6), 2382–2412. 3, 11, 15

[26] Szekely, G., M. Rizzo, and N. Bakirov (2007). Measuring and testing independence by correlation of distances. *Annals of Statistics 35*(6), 2769–2794. 3, 6, 11, 23, 24, 25

[27] Tenenbaum, J. B., V. de Silva, and J. C. Langford (2000). A global geometric framework for nonlinear dimension reduction. *Science 290*, 2319–2323. 5, 16

[28] Vogelstein, J. T., E. Bridgeford, Q. Wang, C. E. Priebe, M. Maggioni, and C. Shen (2019). Discovering and deciphering relationships across disparate data modalities. *eLife 8*, e41690. 5, 21, 28, 31, 34

[29] Wang, S., C. Shen, A. Badea, C. E. Priebe, and J. T. Vogelstein (2018). Signal subgraph estimation via vertex screening. *https://arxiv.org/abs/1801.07683*. 5

[30] Zhou, Z. (2012). Measuring nonlinear dependence in timeseries, a distance correlation approach. *Journal of Time Series Analysis 33*(3), 438–457. 34

Acknowledgment

This work was partially supported by the National Science Foundation award DMS-1712947, and the Defense Advanced Research Projects Agency's (DARPA) SIMPLEX program through SPAWAR contract N66001-15-C-4041. The authors thank the anonymous reviewers for a very diligent and constructive review leading to significant improvement of the manuscript; and thank Dr. Minh Tang and Dr. Shangsi Wang for discussions and suggestions.