Dependency Discovery via Multiscale Graph Correlation

Cencheng Shen

University of Delaware

Collaborators: Joshua T. Vogelstein, Carey E. Priebe, Shangsi Wang, Youjin Lee, Mauro Maggioni, Qing Wang, Alex Badea.

Acknowledgment: NSF DMS, DARPA SIMPLEX.

R package available in CRAN and `https://github.com/neurodata/MGC/`
Matlab code available in `https://github.com/neurodata/mgc-matlab`

Overview

1. Motivation

2. Methodology

3. Theoretical Properties

4. Simulations and Experiments

5. Summary

Motivation

Motivation

Given paired data $(\mathcal{X}_n, \mathcal{Y}_n) = \{(x_i, y_i) \in \mathbb{R}^p \times \mathbb{R}^q, \quad \text{for } i = 1, \ldots, n\}$,

Motivation

Given paired data $(\mathcal{X}_n, \mathcal{Y}_n) = \{(x_i, y_i) \in \mathbb{R}^p \times \mathbb{R}^q, \quad \text{for } i = 1, \ldots, n\}$,

- Are they related?

Motivation

Given paired data $(\mathcal{X}_n, \mathcal{Y}_n) = \{(x_i, y_i) \in \mathbb{R}^p \times \mathbb{R}^q, \quad \text{for } i = 1, \ldots, n\}$,

- Are they related?
- How are they related?

Motivation

Given paired data $(\mathcal{X}_n, \mathcal{Y}_n) = \{(x_i, y_i) \in \mathbb{R}^p \times \mathbb{R}^q, \quad \text{for } i = 1, \ldots, n\}$,

- Are they related?
- How are they related?

Motivation

Given paired data $(\mathcal{X}_n, \mathcal{Y}_n) = \{(x_i, y_i) \in \mathbb{R}^p \times \mathbb{R}^q, \quad \text{for } i = 1, \ldots, n\}$,

- Are they related?
- How are they related?

X	Y
brain connectivity	creativity / personality
brain shape	health
gene / protein	cancer
social networks	attributes
anything	anything else

Formal Definition of Independence Testing

$$(x_i, y_i) \stackrel{i.i.d.}{\sim} F_{XY}, \quad i = 1, \ldots, n$$

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

Formal Definition of Independence Testing

$$(x_i, y_i) \overset{i.i.d.}{\sim} F_{XY}, \quad i = 1, \ldots, n$$

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

A test is universally consistent if its power converges to 1 as $n \to \infty$ against any dependent F_{XY}.

Formal Definition of Independence Testing

$$(x_i, y_i) \stackrel{i.i.d.}{\sim} F_{XY}, \quad i = 1, \ldots, n$$

$$\begin{aligned} H_0 &: F_{XY} = F_X F_Y, \\ H_A &: F_{XY} \neq F_X F_Y. \end{aligned}$$

A test is universally consistent if its power converges to 1 as $n \rightarrow \infty$ against any dependent F_{XY}.

Without loss of generality, we shall assume F_{XY} has finite second moments.

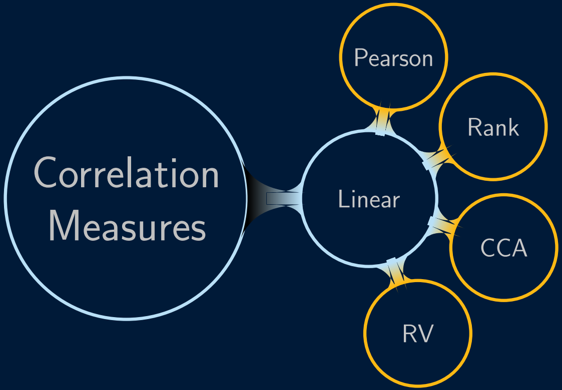
Benchmarks

Correlation
Measures

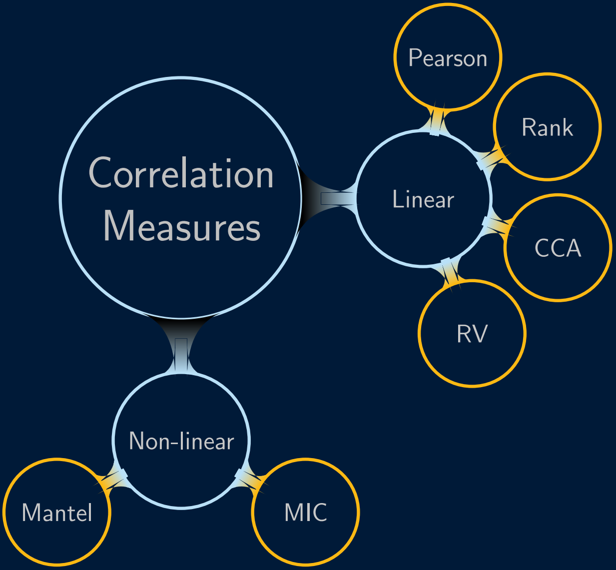
Benchmarks

Correlation Measures

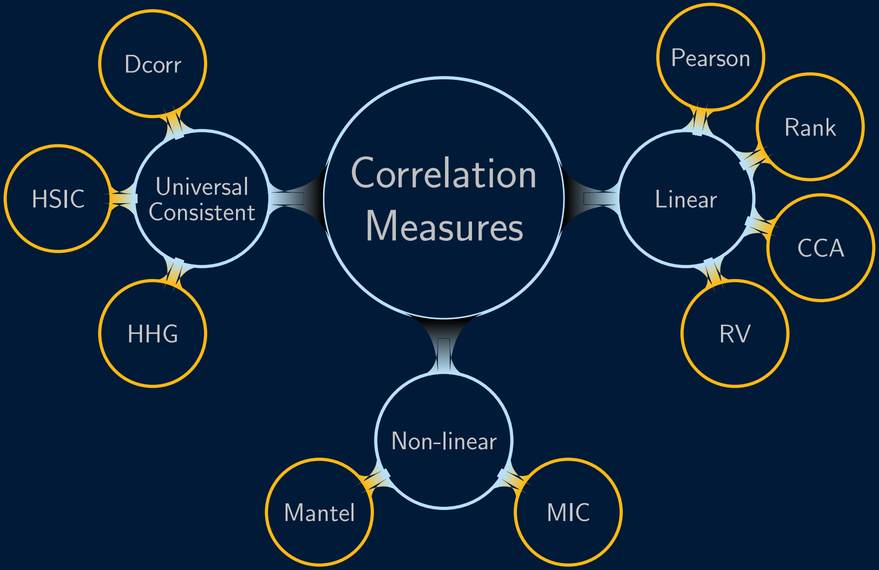
Benchmarks



Benchmarks


Correlation Measures
Linear
Pearson
Rank
CCA
RV
Non-linear
Mantel
MIC

Benchmarks


Correlation
Measures
Universal
Consistent
Dcorr
HSIC
HHG
Linear
Pearson
Rank
CCA
RV
Non-linear
Mantel
MIC

Motivations

Modern data sets may be **high-dimensional, nonlinear, noisy, of limited sample size, structured, from disparate spaces**. Thus we desire a test that

Motivations

Modern data sets may be **high-dimensional, nonlinear, noisy, of limited sample size, structured, from disparate spaces**. Thus we desire a test that

- is consistent against all dependencies;

Motivations

Modern data sets may be **high-dimensional, nonlinear, noisy, of limited sample size, structured, from disparate spaces**. Thus we desire a test that

- is consistent against all dependencies;
- has good finite-sample testing performance;

Motivations

Modern data sets may be **high-dimensional, nonlinear, noisy, of limited sample size, structured, from disparate spaces**. Thus we desire a test that

- is consistent against all dependencies;
- has good finite-sample testing performance;
- is easy to understand and efficient to implement;

Motivations

Modern data sets may be **high-dimensional, nonlinear, noisy, of limited sample size, structured, from disparate spaces**. Thus we desire a test that

- is consistent against all dependencies;
- has good finite-sample testing performance;
- is easy to understand and efficient to implement;
- *provides insights into the dependency structure.

Motivations

Modern data sets may be **high-dimensional, nonlinear, noisy, of limited sample size, structured, from disparate spaces**. Thus we desire a test that

- is consistent against all dependencies;
- has good finite-sample testing performance;
- is easy to understand and efficient to implement;
- *provides insights into the dependency structure.

Motivations

Modern data sets may be **high-dimensional, nonlinear, noisy, of limited sample size, structured, from disparate spaces**. Thus we desire a test that

- is consistent against all dependencies;
- has good finite-sample testing performance;
- is easy to understand and efficient to implement;
- *provides insights into the dependency structure.

To that end, we propose the **multiscale graph correlation** in [*Shen et al.(2018)*].

Methodology

Flowchart of MGC

$(\mathcal{X}_n, \mathcal{Y}_n)$

Flowchart of MGC

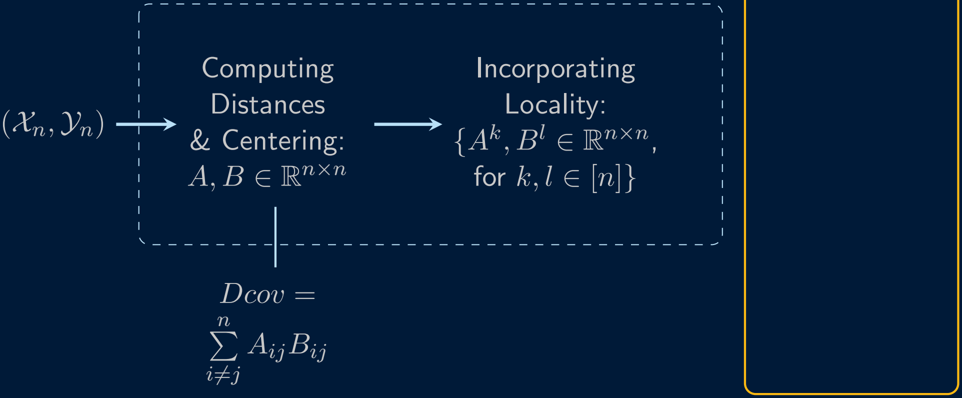
$(\mathcal{X}_n, \mathcal{Y}_n) \rightarrow$ Computing Distances & Centering: $A, B \in \mathbb{R}^{n \times n}$

Flowchart of MGC

$(\mathcal{X}_n, \mathcal{Y}_n) \rightarrow$ Computing Distances & Centering: $A, B \in \mathbb{R}^{n \times n}$

$Dcov = \sum_{i \neq j}^{n} A_{ij} B_{ij}$

Flowchart of MGC



Flowchart of MGC

$(\mathcal{X}_n, \mathcal{Y}_n)$ →

Computing Distances & Centering: $A, B \in \mathbb{R}^{n\times n}$

→

Incorporating Locality: $\{A^k, B^l \in \mathbb{R}^{n\times n},$ for $k, l \in [n]\}$

$$Dcov = \sum_{i\neq j}^{n} A_{ij} B_{ij}$$

$$Dcov^{k,l} = \sum_{i\neq j}^{n} A^k_{ij} B^l_{ij} - \sum_{i\neq j}^{n} A^k_{ij} \sum_{i\neq j}^{n} B^l_{ij}$$

Flowchart of MGC

$(\mathcal{X}_n, \mathcal{Y}_n)$ →

Computing Distances & Centering: $A, B \in \mathbb{R}^{n \times n}$

→

Incorporating Locality: $\{A^k, B^l \in \mathbb{R}^{n \times n}$, for $k, l \in [n]\}$

→

All Local Correlations $\{Dcorr^{k,l}\} \in [-1, 1]^{n \times n}$

$$Dcov = \sum_{i \neq j}^{n} A_{ij} B_{ij}$$

$$Dcov^{k,l} = \sum_{i \neq j}^{n} A^k_{ij} B^l_{ij} - \sum_{i \neq j}^{n} A^k_{ij} \sum_{i \neq j}^{n} B^l_{ij}$$

Flowchart of MGC

$(\mathcal{X}_n, \mathcal{Y}_n)$

Computing Distances & Centering: $A, B \in \mathbb{R}^{n \times n}$

$Dcov = \sum_{i \neq j}^{n} A_{ij} B_{ij}$

Incorporating Locality: $\{A^k, B^l \in \mathbb{R}^{n \times n},$ for $k, l \in [n]\}$

$Dcov^{k,l} = \sum_{i \neq j}^{n} A_{ij}^k B_{ij}^l - \sum_{i \neq j}^{n} A_{ij}^k \sum_{i \neq j}^{n} B_{ij}^l$

All Local Correlations $\{Dcorr^{k,l}\} \in [-1, 1]^{n \times n}$

Smoothed Maximum: MGC $\in [-1, 1]$, and Optimal Scale (k^*, l^*)

Flowchart of MGC

$(\mathcal{X}_n, \mathcal{Y}_n)$

Computing Distances & Centering: $A, B \in \mathbb{R}^{n \times n}$

$$Dcov = \sum_{i \neq j}^{n} A_{ij} B_{ij}$$

Incorporating Locality: $\{A^k, B^l \in \mathbb{R}^{n \times n}$, for $k, l \in [n]\}$

$$Dcov^{k,l} = \sum_{i \neq j}^{n} A^k_{ij} B^l_{ij} - \sum_{i \neq j}^{n} A^k_{ij} \sum_{i \neq j}^{n} B^l_{ij}$$

All Local Correlations $\{Dcorr^{k,l}\} \in [-1, 1]^{n \times n}$

Smoothed Maximum: MGC $\in [-1, 1]$, and Optimal Scale (k^*, l^*)

P-value by Permutation Test

Computing Distance and Centering

Input: $\mathcal{X}_n = [x_1, \ldots, x_n]$ as the data matrix with each column representing one sample observation, and similarly $\mathcal{Y}_n$.

Computing Distance and Centering

Input: $\mathcal{X}_n = [x_1, \ldots, x_n]$ as the data matrix with each column representing one sample observation, and similarly $\mathcal{Y}_n$.

Distance Computation: Let $\tilde{A}$ be the $n \times n$ Euclidean distance matrices of $\mathcal{X}_n$:

$$\tilde{A}_{ij} = \|x_i - x_j\|_2,$$

and similarly $\tilde{B}$ from $\mathcal{Y}_n$.

Computing Distance and Centering

Input: $\mathcal{X}_n = [x_1, \ldots, x_n]$ as the data matrix with each column representing one sample observation, and similarly $\mathcal{Y}_n$.

Distance Computation: Let $\tilde{A}$ be the $n \times n$ Euclidean distance matrices of $\mathcal{X}_n$:

$$\tilde{A}_{ij} = \|x_i - x_j\|_2,$$

and similarly $\tilde{B}$ from $\mathcal{Y}_n$.

Centering: Then we center $\tilde{A}$ and $\tilde{B}$ by columns, with the diagonals excluded:

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

similarly for B.

Incorporating the Locality Principle

Incorporating the Locality Principle

Ranking: 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.

Incorporating the Locality Principle

Ranking: 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.

For any $(k, l) \in [n]^2$, define the rank truncated matrices A^k, B^l, and the joint distance matrix C^{kl} as

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

Incorporating the Locality Principle

Ranking: 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.

For any $(k, l) \in [n]^2$, define the rank truncated matrices A^k, B^l, and the joint distance matrix C^{kl} as

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

When ties occur, minimal rank is recommended, e.g., if Y only takes two value, R^B_{ij} takes value in $\{1, 2\}$ only. We assume no ties for each of presentation.

Local Distance Correlations

Local Distance Correlations

A Family of Local Correlations: 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:

Local Distance Correlations

A Family of Local Correlations: 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:

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

Local Distance Correlations

A Family of Local Correlations: 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:

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

for $k, l = 1, \ldots, n$. If $dVar^k(\mathcal{X}_n) \cdot dVar^l(\mathcal{X}_n) \leq 0$, we set $dCorr^{kl}(\mathcal{X}_n, \mathcal{Y}_n) = 0$ instead.

Local Distance Correlations

A Family of Local Correlations: 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:

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

for $k, l = 1, \ldots, n$. If $dVar^k(\mathcal{X}_n) \cdot dVar^l(\mathcal{X}_n) \leq 0$, we set $dCorr^{kl}(\mathcal{X}_n, \mathcal{Y}_n) = 0$ instead.

There are a maximum of n^2 different local correlations. At $k = l = n$, $dCorr^{kl}(\mathcal{X}_n, \mathcal{Y}_n)$ equals the "global" distance correlation $dCorr(\mathcal{X}_n, \mathcal{Y}_n)$ by *Szekely et al.(2007)*.

Smoothed Maximum

One would like to use the optimal local correlation for testing.

Smoothed Maximum

One would like to use the optimal local correlation for testing.

But directly taking the maximum local correlation

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

will yield a biased statistic under independence, i.e., the maximum is always larger than 0 in expectation even under independent relationship!

Smoothed Maximum

One would like to use the optimal local correlation for testing.

But directly taking the maximum local correlation

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

will yield a biased statistic under independence, i.e., the maximum is always larger than 0 in expectation even under independent relationship!

Instead, we take a smoothed maximum, by finding a connected region in the local correlation map with significant local correlatons – if such a region exists, use the maximum within the region.

Smoothed Maximum

Pick a threshold $\tau \geq 0$ (we choose by an approximate null distribution of $Dcorr$, which is symmetric beta and converges to 0 as $n \rightarrow \infty$), compute the set

$$\{(k,l) \text{ such that } Dcorr^{k,l}(\mathcal{X}_n, \mathcal{Y}_n) > \max\{\tau, Dcorr(\mathcal{X}_n, \mathcal{Y}_n)\}\},$$

Smoothed Maximum

Pick a threshold $\tau \geq 0$ (we choose by an approximate null distribution of $Dcorr$, which is symmetric beta and converges to 0 as $n \to \infty$), compute the set

$$\{(k,l) \text{ such that } Dcorr^{k,l}(\mathcal{X}_n, \mathcal{Y}_n) > \max\{\tau, Dcorr(\mathcal{X}_n, \mathcal{Y}_n)\}\},$$

and calculate the largest connected component R of the set.

Smoothed Maximum

Pick a threshold $\tau \geq 0$ (we choose by an approximate null distribution of $Dcorr$, which is symmetric beta and converges to 0 as $n \to \infty$), compute the set

$$\{(k,l) \text{ such that } Dcorr^{k,l}(\mathcal{X}_n, \mathcal{Y}_n) > \max\{\tau, Dcorr(\mathcal{X}_n, \mathcal{Y}_n)\}\},$$

and calculate the largest connected component R of the set. If there are sufficiently many elements in R $(> 2n)$, take the maximum correlation within R as MGC,

Smoothed Maximum

Pick a threshold $\tau \geq 0$ (we choose by an approximate null distribution of $Dcorr$, which is symmetric beta and converges to 0 as $n \to \infty$), compute the set

$$\{(k, l) \text{ such that } Dcorr^{k,l}(\mathcal{X}_n, \mathcal{Y}_n) > \max\{\tau, Dcorr(\mathcal{X}_n, \mathcal{Y}_n)\}\},$$

and calculate the largest connected component R of the set. If there are sufficiently many elements in R $(> 2n)$, take the maximum correlation within R as MGC, and set the neighborhood pair as the optimal scale (k^*, l^*).

Permutation Test

To get a p-value by MGC for any given data, we utilize the permutation test: randomly permute index the second data set for r times, compute the permuted MGC statistic for each, and compute the percentage that the original MGC is no larger than the permuted MGC statistic, and take it as the p-value.

Permutation Test

To get a p-value by MGC for any given data, we utilize the permutation test: randomly permute index the second data set for r times, compute the permuted MGC statistic for each, and compute the percentage that the original MGC is no larger than the permuted MGC statistic, and take it as the p-value.

This is a standard nonparametric testing procedure employed by Mantel, Dcorr, HHG, HSIC, where the null distribution of the dependency measure cannot be exactly derived.

Computation Complexity

Distance computation takes $\mathcal{O}(n^2 \max(p, q))$, centering takes $\mathcal{O}(n^2)$, ranking takes $\mathcal{O}(n^2 log(n))$, **all local correlations can be iteratively computed in** $\mathcal{O}(n^2)$, and the smoothing step takes $\mathcal{O}(n^2)$.

Computation Complexity

Distance computation takes $\mathcal{O}(n^2 \max(p, q))$, centering takes $\mathcal{O}(n^2)$, ranking takes $\mathcal{O}(n^2 log(n))$, **all local correlations can be iteratively computed in** $\mathcal{O}(n^2)$, and the smoothing step takes $\mathcal{O}(n^2)$.

Overall, MGC can be computed in $\mathcal{O}(n^2 \max(p, q, log(n)))$, which is comparable to DCorr, HHG, and HSIC.

Computation Complexity

Distance computation takes $\mathcal{O}(n^2 \max(p, q))$, centering takes $\mathcal{O}(n^2)$, ranking takes $\mathcal{O}(n^2 log(n))$, **all local correlations can be iteratively computed in** $\mathcal{O}(n^2)$, and the smoothing step takes $\mathcal{O}(n^2)$.

Overall, MGC can be computed in $\mathcal{O}(n^2 \max(p, q, log(n)))$, which is comparable to DCorr, HHG, and HSIC.

The permutation test takes $\mathcal{O}(n^2 \max(r, p, q, log(n)))$ for r random permutations.

Computation Complexity

Distance computation takes $\mathcal{O}(n^2 \max(p, q))$, centering takes $\mathcal{O}(n^2)$, ranking takes $\mathcal{O}(n^2 log(n))$, **all local correlations can be iteratively computed in** $\mathcal{O}(n^2)$, and the smoothing step takes $\mathcal{O}(n^2)$.

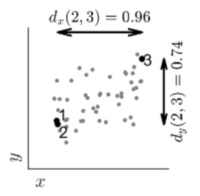
Overall, MGC can be computed in $\mathcal{O}(n^2 \max(p, q, log(n)))$, which is comparable to DCorr, HHG, and HSIC.

The permutation test takes $\mathcal{O}(n^2 \max(r, p, q, log(n)))$ for r random permutations.

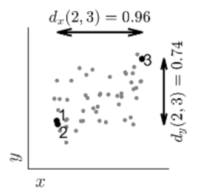
There are a number of ways to speed up the method for big data: faster implementation when $p = q = 1$, null distribution approximation by subsampling and spectral embedding.

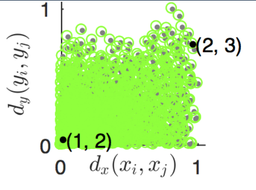
Examples

Examples

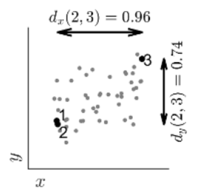


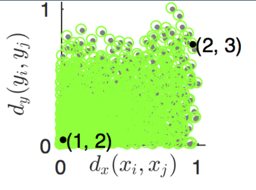
Examples





Examples



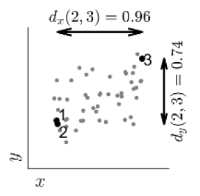


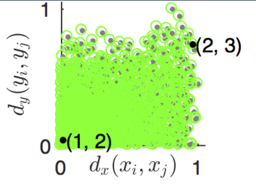
$Dcorr(\mathcal{X}_n, \mathcal{Y}_n) = 0.15$

$MGC(\mathcal{X}_n, \mathcal{Y}_n) = 0.15$

p-vals: < 0.001

Examples

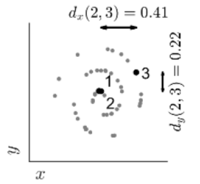




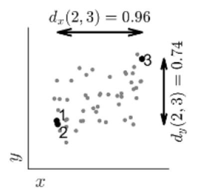
$Dcorr(\mathcal{X}_n, \mathcal{Y}_n) = 0.15$

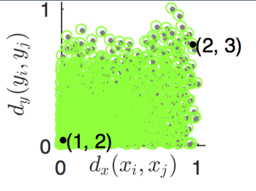
$MGC(\mathcal{X}_n, \mathcal{Y}_n) = 0.15$

p-vals: < 0.001

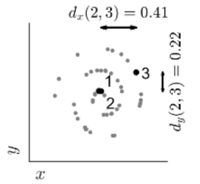


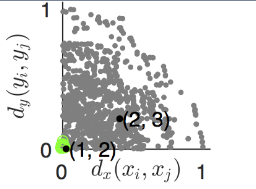
Examples



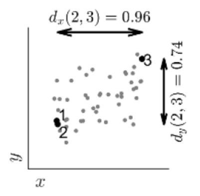


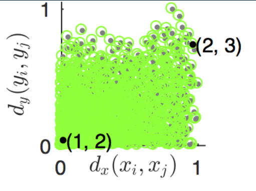
$$Dcorr(\mathcal{X}_n, \mathcal{Y}_n) = 0.15$$
$$MGC(\mathcal{X}_n, \mathcal{Y}_n) = 0.15$$
p-vals: < 0.001



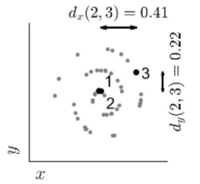


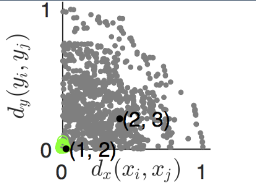
Examples





$Dcorr(\mathcal{X}_n, \mathcal{Y}_n) = 0.15$
$MGC(\mathcal{X}_n, \mathcal{Y}_n) = 0.15$
p-vals: < 0.001





$Dcorr(\mathcal{X}_n, \mathcal{Y}_n) = 0.01$
$MGC(\mathcal{X}_n, \mathcal{Y}_n) = 0.13$
p-vals: 0.3 vs < 0.001

Theoretical Properties

Basic Properties of Sample MGC

Theorem 1 (Well-behaved Correlation Measure)

Basic Properties of Sample MGC

Theorem 1 (Well-behaved Correlation Measure)

1. Boundedness: $c(\mathcal{X}_n, \mathcal{Y}_n) \in [-1, 1]$.

Basic Properties of Sample MGC

Theorem 1 (Well-behaved Correlation Measure)

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

Basic Properties of Sample MGC

Theorem 1 (Well-behaved Correlation Measure)

1. *Boundedness:* $c(\mathcal{X}_n, \mathcal{Y}_n) \in [-1, 1]$.
2. *Symmetric:* $c(\mathcal{X}_n, \mathcal{Y}_n) = c(\mathcal{Y}_n, \mathcal{X}_n)$.
3. *Non-negative Variance:* $c(\mathcal{X}_n, \mathcal{X}_n) \geq 0$ *with equality if and only if* F_X *is non-degenerate.*

Basic Properties of Sample MGC

Theorem 1 (Well-behaved Correlation Measure)

1. Boundedness: $c(\mathcal{X}_n, \mathcal{Y}_n) \in [-1, 1]$.

2. Symmetric: $c(\mathcal{X}_n, \mathcal{Y}_n) = c(\mathcal{Y}_n, \mathcal{X}_n)$.

3. Non-negative Variance: $c(\mathcal{X}_n, \mathcal{X}_n) \geq 0$ *with equality if and only if* F_X *is non-degenerate.*

4. Invariant: $c(\mathcal{X}_n, \mathcal{Y}_n) = c(\{\phi(x_i)\}, \{\delta(y_i)\})$ *for any distance-preserving transformations* ϕ, δ *(i.e., rotation, scaling, translation, reflection).*

Basic Properties of Sample MGC

Theorem 1 (Well-behaved Correlation Measure)

1. Boundedness: $c(\mathcal{X}_n, \mathcal{Y}_n) \in [-1, 1]$.

2. Symmetric: $c(\mathcal{X}_n, \mathcal{Y}_n) = c(\mathcal{Y}_n, \mathcal{X}_n)$.

3. Non-negative Variance: $c(\mathcal{X}_n, \mathcal{X}_n) \geq 0$ *with equality if and only if* F_X *is non-degenerate.*

4. Invariant: $c(\mathcal{X}_n, \mathcal{Y}_n) = c(\{\phi(x_i)\}, \{\delta(y_i)\})$ *for any distance-preserving transformations* ϕ, δ *(i.e., rotation, scaling, translation, reflection).*

5. 1-Linear: $c(\mathcal{X}_n, \mathcal{Y}_n) = 1$ *if and only if* F_X *is non-degenerate and* (X, uY) *are dependent via an isometry for some non-zero constant* u.

Consistency of Sample MGC

Theorem 2 (Consistency)

Consistency of Sample MGC

Theorem 2 (Consistency)

1. 0-Indep: $c(\mathcal{X}_n, \mathcal{Y}_n) \stackrel{n \to \infty}{\to} 0$ *if and only if independence.*

Consistency of Sample MGC

Theorem 2 (Consistency)

1. 0-Indep: $c(\mathcal{X}_n, \mathcal{Y}_n) \stackrel{n \rightarrow \infty}{\rightarrow} 0$ *if and only if independence.*

2. Valid Test: Under the permutation test, Sample MGC is a valid test, i.e., it controls the type 1 error level α*.*

Consistency of Sample MGC

Theorem 2 (Consistency)

1. 0-Indep: $c(\mathcal{X}_n, \mathcal{Y}_n) \stackrel{n \to \infty}{\to} 0$ *if and only if independence.*

2. Valid Test: Under the permutation test, Sample MGC is a valid test, i.e., it controls the type 1 error level α.

3. Consistency: At any type 1 error level α*, testing power* $\beta(c(\mathcal{X}_n, \mathcal{Y}_n)) \stackrel{n \to \infty}{\to} 1$ *against any dependent* F_{XY}.

Defining Population MGC

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

Defining Population MGC

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

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

for $\rho_k, \rho_l \in [0,1]$.

Defining Population MGC

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

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

for $\rho_k, \rho_l \in [0,1]$. Further define

$$d^{\rho_k}_X = (\|X - X'\| - \|X - X''\|)\boldsymbol{I}^{\rho_k}_{X,X'}$$
$$d^{\rho_l}_{Y'} = (\|Y' - Y\| - \|Y' - Y'''\|)\boldsymbol{I}^{\rho_l}_{Y',Y}$$

Defining Population MGC

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

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

for $\rho_k, \rho_l \in [0,1]$. Further define

$$d^{\rho_k}_X = (\|X-X'\| - \|X-X''\|)\boldsymbol{I}^{\rho_k}_{X,X'}$$
$$d^{\rho_l}_{Y'} = (\|Y'-Y\| - \|Y'-Y'''\|)\boldsymbol{I}^{\rho_l}_{Y',Y}$$

The population local covariance can be defined as

$$Dcov^{\rho_k,\rho_l}(X,Y) = E(d^{\rho_k}_X d^{\rho_l}_{Y'}) - E(d^{\rho_k}_X)E(d^{\rho_l}_{Y'}).$$

Defining Population MGC

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

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

for $\rho_k, \rho_l \in [0,1]$. Further define

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

The population local covariance can be defined as

$$Dcov^{\rho_k,\rho_l}(X,Y) = E(d_X^{\rho_k} d_{Y'}^{\rho_l}) - E(d_X^{\rho_k})E(d_{Y'}^{\rho_l}).$$

Normalizing and taking a smoothed maximum yield population MGC.

Sample to Population

Sample to Population

Alternatively, the population version can be equivalently defined via characteristic functions of F_{XY}:

Sample to Population

Alternatively, the population version can be equivalently defined via characteristic functions of F_{XY}:

$$Dcov^{\rho_k=1,\rho_l=1}(X,Y) = \int_{t,s} |g_{XY}(t,s) - g_X(t)g_Y(s)|^2 dw(t,s)$$

with respect to a non-negative weight function $w(t,s)$ on $(t,s) \in \mathbb{R}^p \times \mathbb{R}^q$. The weight function 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.

Sample to Population

Alternatively, the population version can be equivalently defined via characteristic functions of F_{XY}:

$$Dcov^{\rho_k=1,\rho_l=1}(X,Y) = \int_{t,s} |g_{XY}(t,s) - g_X(t)g_Y(s)|^2 dw(t,s)$$

with respect to a non-negative weight function $w(t,s)$ on $(t,s) \in \mathbb{R}^p \times \mathbb{R}^q$. The weight function 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.

Can be similarly adapted to the local correlation.

Theoretical Advantages of MGC

Theorem 3 (Convergence, Mean and Variance)

Theoretical Advantages of MGC

Theorem 3 (Convergence, Mean and Variance)

1. 0-Indep: $c(X, Y) = 0$ *if and only if independence.*

Theoretical Advantages of MGC

Theorem 3 (Convergence, Mean and Variance)

1. *0-Indep:* $c(X, Y) = 0$ *if and only if independence.*
2. *Convergence:* $c(\mathcal{X}_n, \mathcal{Y}_n) \stackrel{n \to \infty}{\to} c(X, Y)$.

Theoretical Advantages of MGC

Theorem 3 (Convergence, Mean and Variance)

1. *0-Indep:* $c(X,Y)=0$ *if and only if independence.*
2. *Convergence:* $c(\mathcal{X}_n,\mathcal{Y}_n) \stackrel{n\to\infty}{\to} c(X,Y)$.
3. *Almost Unbiased:* $E(c(\mathcal{X}_n,\mathcal{Y}_n)) = c(X,Y) + \mathcal{O}(1/n)$.

Theoretical Advantages of MGC

Theorem 3 (Convergence, Mean and Variance)

1. *0-Indep:* $c(X, Y) = 0$ *if and only if independence.*
2. *Convergence:* $c(\mathcal{X}_n, \mathcal{Y}_n) \stackrel{n\rightarrow\infty}{\rightarrow} c(X, Y)$.
3. *Almost Unbiased:* $E(c(\mathcal{X}_n, \mathcal{Y}_n)) = c(X, Y) + \mathcal{O}(1/n)$.
4. *Diminishing Variance:* $Var(c(\mathcal{X}_n, \mathcal{Y}_n)) = \mathcal{O}(1/n)$.

Theoretical Advantages of MGC

Theorem 3 (Convergence, Mean and Variance)

1. *0-Indep:* $c(X, Y) = 0$ *if and only if independence.*
2. *Convergence:* $c(\mathcal{X}_n, \mathcal{Y}_n) \stackrel{n \to \infty}{\to} c(X, Y)$.
3. *Almost Unbiased:* $E(c(\mathcal{X}_n, \mathcal{Y}_n)) = c(X, Y) + \mathcal{O}(1/n)$.
4. *Diminishing Variance:* $Var(c(\mathcal{X}_n, \mathcal{Y}_n)) = \mathcal{O}(1/n)$.

The last three properties also hold for any local correlation by $(\rho_k, \rho_l) = (\frac{k-1}{n-1}, \frac{l-1}{n-1})$.

Theoretical Advantages of MGC

Theorem 4 (Advantages of Population MGC vs Mantel and Dcorr)

Theoretical Advantages of MGC

Theorem 4 (Advantages of Population MGC vs Mantel and Dcorr)

1. There exists dependent F_{XY} such that $Mantel = 0$, thus not universally consistent.

Theoretical Advantages of MGC

Theorem 4 (Advantages of Population MGC vs Mantel and Dcorr)

1. There exists dependent F_{XY} such that $Mantel = 0$, thus not universally consistent.

2. For any dependent F_{XY}, $MGC(X,Y) \geq Dcorr(X,Y)$.

Theoretical Advantages of MGC

Theorem 4 (Advantages of Population MGC vs Mantel and Dcorr)

1. *There exists dependent* F_{XY} *such that* $Mantel = 0$*, thus not universally consistent.*
2. *For any dependent* F_{XY}*,* $MGC(X,Y) \geq Dcorr(X,Y)$*.*
3. *There exists dependent* F_{XY} *such that* $MGC(X,Y) > Dcorr(X,Y)$*.*

Theoretical Advantages of MGC

Theorem 4 (Advantages of Population MGC vs Mantel and Dcorr)

1. *There exists dependent* F_{XY} *such that* $Mantel = 0$, *thus not universally consistent.*
2. *For any dependent* F_{XY}, $MGC(X, Y) \geq Dcorr(X, Y)$.
3. *There exists dependent* F_{XY} *such that* $MGC(X, Y) > Dcorr(X, Y)$.

As MGC and Dcorr share similar variance and same mean under the null, the mean advantage in the alternative is translated to the testing power.

Theoretical Advantages of MGC

Theorem 4 (Advantages of Population MGC vs Mantel and Dcorr)

1. There exists dependent F_{XY} such that $Mantel = 0$, thus not universally consistent.

2. For any dependent F_{XY}, $MGC(X,Y) \geq Dcorr(X,Y)$.

3. There exists dependent F_{XY} such that $MGC(X,Y) > Dcorr(X,Y)$.

As MGC and Dcorr share similar variance and same mean under the null, the mean advantage in the alternative is translated to the testing power.

Theorem 5 (Optimal Scale of MGC Implies Geometry Structure)

Theoretical Advantages of MGC

Theorem 4 (Advantages of Population MGC vs Mantel and Dcorr)

1. There exists dependent F_{XY} such that $Mantel = 0$, thus not universally consistent.

2. For any dependent F_{XY}, $MGC(X,Y) \geq Dcorr(X,Y)$.

3. There exists dependent F_{XY} such that $MGC(X,Y) > Dcorr(X,Y)$.

As MGC and Dcorr share similar variance and same mean under the null, the mean advantage in the alternative is translated to the testing power.

Theorem 5 (Optimal Scale of MGC Implies Geometry Structure)

If the relationship is linear (or with independent noise), the global scale is always optimal and $MGC(X,Y) = Dcorr(X,Y)$.

Theoretical Advantages of MGC

Theorem 4 (Advantages of Population MGC vs Mantel and Dcorr)

1. *There exists dependent F_{XY} such that $Mantel = 0$, thus not universally consistent.*
2. *For any dependent F_{XY}, $MGC(X, Y) \geq Dcorr(X, Y)$.*
3. *There exists dependent F_{XY} such that $MGC(X, Y) > Dcorr(X, Y)$.*

As MGC and Dcorr share similar variance and same mean under the null, the mean advantage in the alternative is translated to the testing power.

Theorem 5 (Optimal Scale of MGC Implies Geometry Structure)

If the relationship is linear (or with independent noise), the global scale is always optimal and $MGC(X, Y) = Dcorr(X, Y)$.

Conversely, the optimal scale being local, i.e., $MGC(X, Y) > Dcorr(X, Y)$, implies a non-linear relationship.

MGC is applicable to similarity / kernel matrix

Theorem 6 (Transforming kernel to distance)

Given any characteristic kernel function $k(\cdot,\cdot)$, define an induced semi-metric as

$$d(\cdot,\cdot) = 1 - k(\cdot,\cdot)/\max\{k(\cdot,\cdot)\}.$$

Then $d(\cdot,\cdot)$ is of strong negative type, and the resulting MGC is universally consistent.

Namely, given a sample kernel matrices $K_{n\times n}$, one can compute the induced distance matrix by

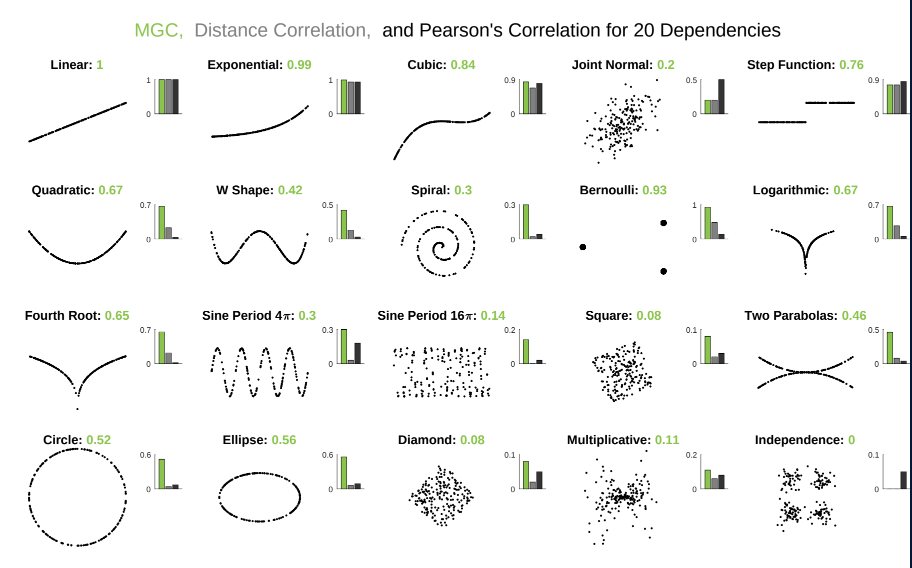
$$D = J - K/\max_{i,j\in[1,\ldots,n]^2}\{K(i,j)\},$$

and apply MGC to the induced distance matrices.

Simulations and Experiments

Visualizations of 20 Simulation Settings

Visualizations of 20 Simulation Settings



Evaluation Criterion

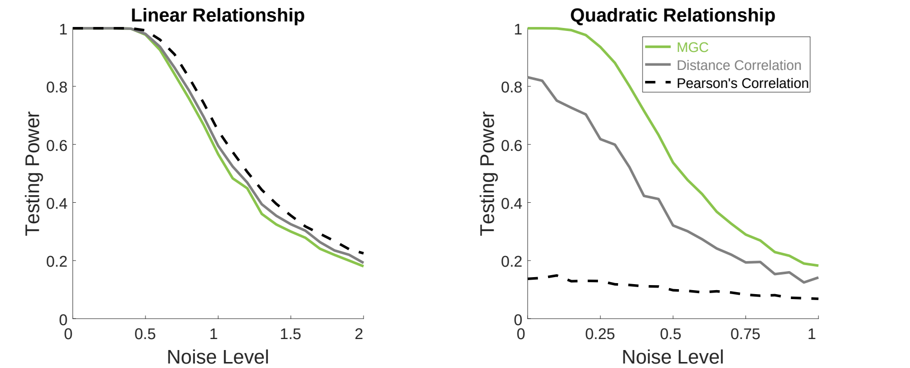
- Power is the probability of rejecting the null when the alternative is true.

Evaluation Criterion

- Power is the probability of rejecting the null when the alternative is true.
- Required sample size $N_{\alpha,\beta}(c)$ to achieve a power of β at type 1 error level α using a statistic c.

Testing Power: Linear vs Nonlinear

Testing Power: Linear vs Nonlinear



$$n = 30, p = q = 1,$$
$$X \sim Uniform(-1, 1),$$
$$\epsilon \sim Normal(0, noise),$$
$Y = X + \epsilon$ and $Y = X^2 + \epsilon.$

Required Sample Size

Required Sample Size

When $noise = 1$, $p = q = 1$, the required sample size $N_{\alpha=0.05,\beta=0.85}(c)$:

Required Sample Size

When $noise = 1$, $p = q = 1$, the required sample size $N_{\alpha=0.05,\beta=0.85}(c)$:

in linear relationship, 40 for all three methods;
in quadratic relationship, 80 for MGC, 180 for Dcorr, and > 1000 for Pearson.

Required Sample Size

When $noise = 1$, $p = q = 1$, the required sample size $N_{\alpha=0.05,\beta=0.85}(c)$:

in linear relationship, 40 for all three methods;
in quadratic relationship, 80 for MGC, 180 for Dcorr, and > 1000 for Pearson.

Next we compute the size for each simulation, and summarize by the median over close-to-linear (type 1-5) and strongly non-linear relationships (type 6-19).

Required Sample Size

When $noise = 1$, $p = q = 1$, the required sample size $N_{\alpha=0.05,\beta=0.85}(c)$:

in linear relationship, 40 for all three methods;
in quadratic relationship, 80 for MGC, 180 for Dcorr, and > 1000 for Pearson.

Next we compute the size for each simulation, and summarize by the median over close-to-linear (type 1-5) and strongly non-linear relationships (type 6-19).

We consider univariate (1D) and multivariate (10D) cases.

Median Size Table

Testing Methods	1D Lin	1D Non-Lin	10D Lin	10D Non-Lin
MGC	**50**	**90**	60	**165**
Dcorr	**50**	250	60	515
Pearson / RV / CCA	**50**	>1000	**50**	>1000
HHG	70	**90**	100	315
HSIC	70	95	100	400
MIC	120	180	n/a	n/a

A Proteomics Experiment: Identify Biomarker

A Proteomics Experiment: Identify Biomarker

The blood samples of 95 individuals are collected. Pancreatic (10), ovarian (24), colorectal cancer (28), and healthy controls (33).

A Proteomics Experiment: Identify Biomarker

The blood samples of 95 individuals are collected. Pancreatic (10), ovarian (24), colorectal cancer (28), and healthy controls (33).

The protein and peptide abundance levels are measured for each person for 318 peptides / proteins.

A Proteomics Experiment: Identify Biomarker

The blood samples of 95 individuals are collected. Pancreatic (10), ovarian (24), colorectal cancer (28), and healthy controls (33).

The protein and peptide abundance levels are measured for each person for 318 peptides / proteins. The target is to find potential biomarkers that best distinguish between pancreatic and non-pancreatic.

A Proteomics Experiment: Identify Biomarker

The blood samples of 95 individuals are collected. Pancreatic (10), ovarian (24), colorectal cancer (28), and healthy controls (33).

The protein and peptide abundance levels are measured for each person for 318 peptides / proteins. The target is to find potential biomarkers that best distinguish between pancreatic and non-pancreatic.

For each feature, we compute MGC and the p-values between the abundance level and the label vector $(n = 95, p = q = 1)$.

A Proteomics Experiment: Identify Biomarker

The blood samples of 95 individuals are collected. Pancreatic (10), ovarian (24), colorectal cancer (28), and healthy controls (33).

The protein and peptide abundance levels are measured for each person for 318 peptides / proteins. The target is to find potential biomarkers that best distinguish between pancreatic and non-pancreatic.

For each feature, we compute MGC and the p-values between the abundance level and the label vector $(n = 95, p = q = 1)$.

Adjusted for multiple testing, MGC uniquely revealed one particular protein, neurogranin, which is exclusively expressed in brain tissue among normal tissues and has not been linked with any other cancer type.

Comparison and Verification

Comparison and Verification

HSIC identifies two peptides, HHG identifies three peptides, and other methods do not identify any peptide as significant. However, there exists strong evidence that the other two peptides are upregulated in other cancers.

Comparison and Verification

HSIC identifies two peptides, HHG identifies three peptides, and other methods do not identify any peptide as significant. However, there exists strong evidence that the other two peptides are upregulated in other cancers.

If we compare kNN (K=3) leave-one-subject-error of these peptides:

Peptides	False Positives	True Positives
Neurogranin (MGC)	**2**	**5**
HSIC Features	4	**5**
HHG Features	**2**	1

Comparison and Verification

HSIC identifies two peptides, HHG identifies three peptides, and other methods do not identify any peptide as significant. However, there exists strong evidence that the other two peptides are upregulated in other cancers.

If we compare kNN (K=3) leave-one-subject-error of these peptides:

Peptides	False Positives	True Positives
Neurogranin (MGC)	**2**	**5**
HSIC Features	4	**5**
HHG Features	**2**	1

Summary

Summary

Summary

MGC builds on distance correlation, the locality principle, and taking a smoothed maximum.

Summary

MGC builds on distance correlation, the locality principle, and taking a smoothed maximum.

- Proper distance transformation ensures the universal consistency.

Summary

MGC builds on distance correlation, the locality principle, and taking a smoothed maximum.

- Proper distance transformation ensures the universal consistency.
- Compute all local correlations iteratively.

Summary

MGC builds on distance correlation, the locality principle, and taking a smoothed maximum.

- Proper distance transformation ensures the universal consistency.
- Compute all local correlations iteratively.
- Identify the optimal local correlation without inflating the sample bias.

Summary

MGC builds on distance correlation, the locality principle, and taking a smoothed maximum.

- Proper distance transformation ensures the universal consistency.
- Compute all local correlations iteratively.
- Identify the optimal local correlation without inflating the sample bias.

Summary

MGC builds on distance correlation, the locality principle, and taking a smoothed maximum.

- Proper distance transformation ensures the universal consistency.
- Compute all local correlations iteratively.
- Identify the optimal local correlation without inflating the sample bias.

They made MGC advantageous in theory and practice.

Advantages of MGC

Advantages of MGC

1. Performant under any joint distribution of finite second moments:

Advantages of MGC

1. Performant under any joint distribution of finite second moments:
   - Equals 0 asymptotically if and only if independence.

Advantages of MGC

1. Performant under any joint distribution of finite second moments:
    - Equals 0 asymptotically if and only if independence.
    - Amplify the dependency signal while mostly avoiding the sample bias.

Advantages of MGC

1. Performant under any joint distribution of finite second moments:
    - Equals 0 asymptotically if and only if independence.
    - Amplify the dependency signal while mostly avoiding the sample bias.
    - Superior finite-sample performance over all benchmarks, against linear / nonlinear / noisy / high-dimensional relationships.

Advantages of MGC

1. Performant under any joint distribution of finite second moments:
   - Equals 0 asymptotically if and only if independence.
   - Amplify the dependency signal while mostly avoiding the sample bias.
   - Superior finite-sample performance over all benchmarks, against linear / nonlinear / noisy / high-dimensional relationships.

Advantages of MGC

1. Performant under any joint distribution of finite second moments:
   - Equals 0 asymptotically if and only if independence.
   - Amplify the dependency signal while mostly avoiding the sample bias.
   - Superior finite-sample performance over all benchmarks, against linear / nonlinear / noisy / high-dimensional relationships.
2. It works for:

Advantages of MGC

1. Performant under any joint distribution of finite second moments:
   - Equals 0 asymptotically if and only if independence.
   - Amplify the dependency signal while mostly avoiding the sample bias.
   - Superior finite-sample performance over all benchmarks, against linear / nonlinear / noisy / high-dimensional relationships.
2. It works for:
   - Low- and high-dimensional data.

Advantages of MGC

1. Performant under any joint distribution of finite second moments:
   - Equals 0 asymptotically if and only if independence.
   - Amplify the dependency signal while mostly avoiding the sample bias.
   - Superior finite-sample performance over all benchmarks, against linear / nonlinear / noisy / high-dimensional relationships.
2. It works for:
   - Low- and high-dimensional data.
   - Euclidean and structured data (e.g., images, networks, shapes).

Advantages of MGC

1. Performant under any joint distribution of finite second moments:
   - Equals 0 asymptotically if and only if independence.
   - Amplify the dependency signal while mostly avoiding the sample bias.
   - Superior finite-sample performance over all benchmarks, against linear / nonlinear / noisy / high-dimensional relationships.
2. It works for:
   - Low- and high-dimensional data.
   - Euclidean and structured data (e.g., images, networks, shapes).
   - Any dissimilarity / similarity / kernel matrix.

Advantages of MGC

1. Performant under any joint distribution of finite second moments:
   - Equals 0 asymptotically if and only if independence.
   - Amplify the dependency signal while mostly avoiding the sample bias.
   - Superior finite-sample performance over all benchmarks, against linear / nonlinear / noisy / high-dimensional relationships.
2. It works for:
   - Low- and high-dimensional data.
   - Euclidean and structured data (e.g., images, networks, shapes).
   - Any dissimilarity / similarity / kernel matrix.

Advantages of MGC

1. Performant under any joint distribution of finite second moments:
   - Equals 0 asymptotically if and only if independence.
   - Amplify the dependency signal while mostly avoiding the sample bias.
   - Superior finite-sample performance over all benchmarks, against linear / nonlinear / noisy / high-dimensional relationships.
2. It works for:
   - Low- and high-dimensional data.
   - Euclidean and structured data (e.g., images, networks, shapes).
   - Any dissimilarity / similarity / kernel matrix.
3. Intuitive to understand and efficient to implement in $\mathcal{O}(n^2 logn)$.

References

1. C. Shen, C. E. Priebe, and J. T. Vogelstein, “From distance correlation to the multiscale graph correlation,” Journal of the American Statistical Association, 2018, *https://arxiv.org/abs/1710.09768*.

2. J. T. Vogelstein, E. Bridgeford, Q. Wang, C. E. Priebe, M. Maggioni, and C. Shen, “Discovering and Deciphering Relationships Across Disparate Data Modalities,” *https://arxiv.org/abs/1609.05148*.

3. Y. Lee, C. Shen, and J. T. Vogelstein, “Network dependence testing via diffusion maps and distance-based correlations,” *https://arxiv.org/abs/1703.10136*.

4. S. Wang, C. Shen, A. Badea, C. E. Priebe, and J. T. Vogelstein, “Signal subgraph estimation via vertex screening,” *https://arxiv.org/abs/1801.07683*.

5. C. Shen and J. T. Vogelstein, “The Exact Equivalence of Distance and Kernel Methods for Hypothesis Testing,” *https://arxiv.org/abs/1806.05514*.