# Dependency Discovery via Multiscale Generalized 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.*

# Set-up

# Set-up

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

# Set-up

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?

# Set-up

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

- Are they related?
- How are they related?

# Set-up

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?

# Set-up

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

# Formal Definition of Independence Testing

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

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

# Formal Definition of Independence Testing

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

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

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

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

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

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

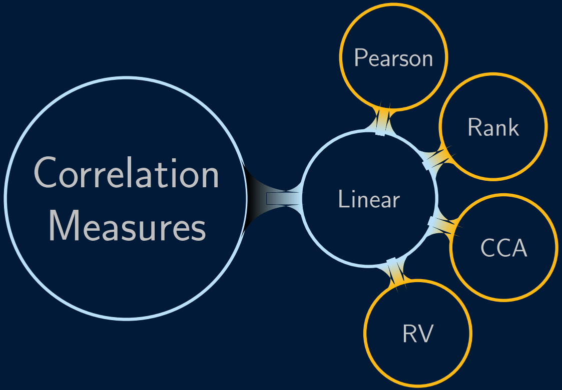
Without loss of generality, we shall assume $F_{XY}$ has finite second moments.

# Benchmarks

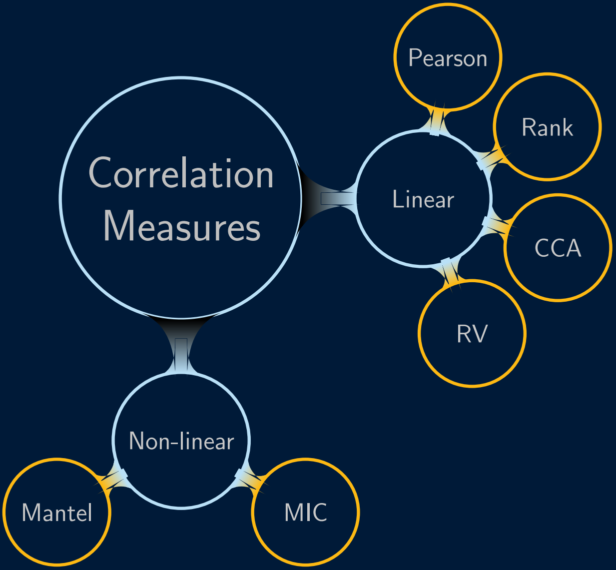
# Benchmarks

Correlation
Measures

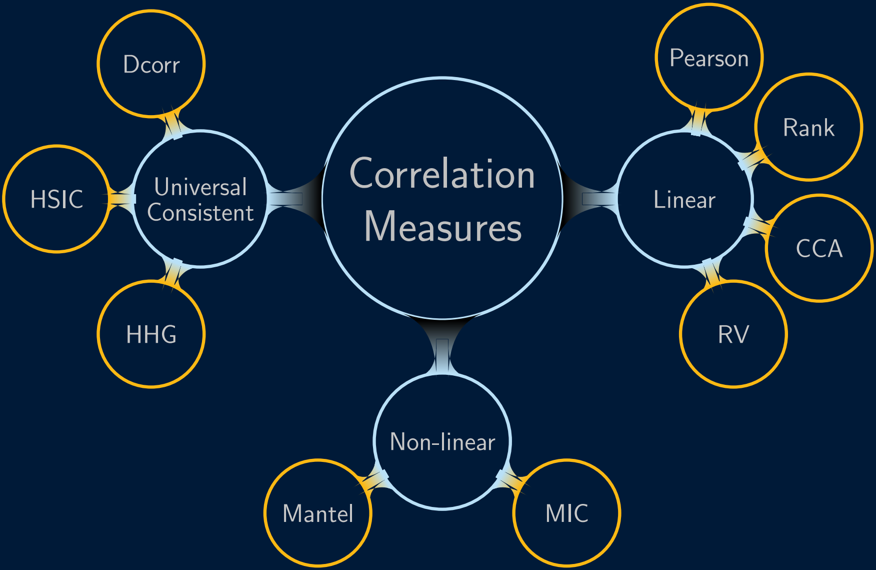
# Benchmarks



# 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

# 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.

Existing method has pros and cons with respect to each point.

# 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.

Existing method has pros and cons with respect to each point.

To that end, we propose the **multiscale generalized correlation** in [*Shen et al.(2017a)*][1].

# Overview

1. Illustration

2. Experiments

3. Theory

4. Summary

# Illustration

# Introducing MGC

# Introducing MGC

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

# Introducing MGC

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

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

# Introducing MGC

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

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

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

# Introducing MGC

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

Computing Distances & Centering: $A, B \in \mathbb{R}^{n \times n}$ $\rightarrow$ Excluding Large Distances: $\{A^k, B^l \in \mathbb{R}^{n \times n},$ for $k, l \in [n]\}$

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

# Introducing MGC

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

Computing Distances & Centering: $A, B \in \mathbb{R}^{n \times n}$ $\rightarrow$ Excluding Large Distances: $\{A^k, B^l \in \mathbb{R}^{n \times n}$, for $k, l \in [n]\}$

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

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

# Introducing MGC

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

Computing Distances & Centering: $A, B \in \mathbb{R}^{n \times n}$ $\longrightarrow$ Excluding Large Distances: $\{A^k, B^l \in \mathbb{R}^{n \times n}$, for $k, l \in [n]\}$ $\longrightarrow$ All Local Correlations $\{Dcorr^{k,l}\} \in [-1, 1]^{n \times n}$

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

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

# Introducing MGC

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

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

Excluding Large Distances: $\{A^k, B^l \in \mathbb{R}^{n\times n}$, for $k, l \in [n]\}$ $\rightarrow$

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

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

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

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

# Introducing MGC

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

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

Excluding Large Distances: $\{A^k, B^l \in \mathbb{R}^{n \times n}$, for $k, l \in [n]\}$

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

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

P-value by Permutation Test

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

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

# Direct Maximum

# Direct Maximum

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

# Smoothed Maximum

Instead, we take a smoothed maximum:

# Smoothed Maximum

Instead, we take a smoothed maximum:
Pick a threshold $\tau \geq 0$, 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

Instead, we take a smoothed maximum:
Pick a threshold $\tau \geq 0$, 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

Instead, we take a smoothed maximum:
Pick a threshold $\tau \geq 0$, 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

Instead, we take a smoothed maximum:
Pick a threshold $\tau \geq 0$, 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 $(k^*, l^*)$.

# Smoothed Maximum

Instead, we take a smoothed maximum:
Pick a threshold $\tau \geq 0$, 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 $(k^*, l^*)$.

$\tau$ is chosen based on an approximate null distribution of $Dcorr$ (symmetric Beta), which converges to $0$ as $n \to \infty$.

# Smoothed Maximum

Instead, we take a smoothed maximum:
Pick a threshold $\tau \geq 0$, compute the set

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

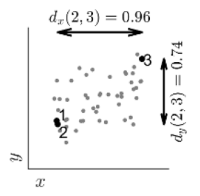
and calculate the largest connected component $R$ of the set. If there are sufficiently many elements in $R$ $(> 2n)$, take the maximum correlation within $R$ as MGC, and set the neighborhood pair as the optimal scale $(k^*, l^*)$.

$\tau$ is chosen based on an approximate null distribution of $Dcorr$ (symmetric Beta), which converges to $0$ as $n \to \infty$.

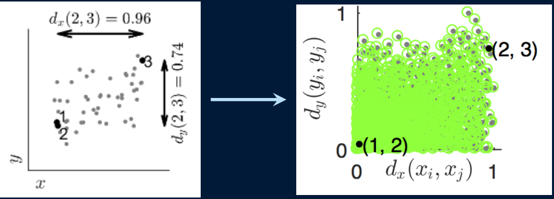
It is a critical step for both the finite-sample performance and certain theoretical properties of MGC.

# Examples

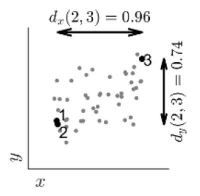
# Examples

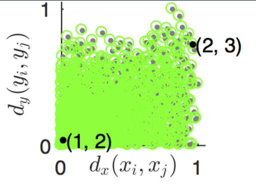


# Examples



# Examples



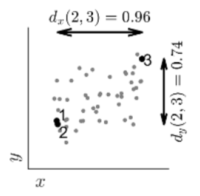



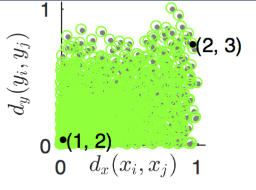

$$Dcorr(\mathcal{X}_n, \mathcal{Y}_n) = 0.15$$

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

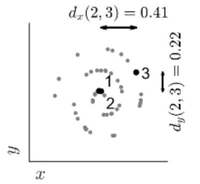
p-vals: $< 0.001$

# Examples

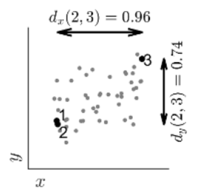


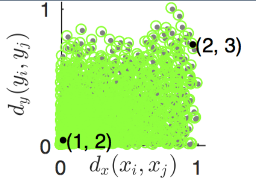




$$Dcorr(\mathcal{X}_n, \mathcal{Y}_n) = 0.15$$
$$MGC(\mathcal{X}_n, \mathcal{Y}_n) = 0.15$$
p-vals: $< 0.001$



# Examples

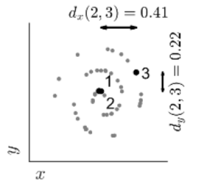


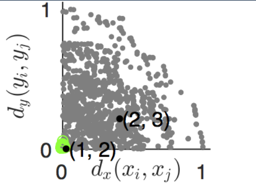




$Dcorr(\mathcal{X}_n, \mathcal{Y}_n) = 0.15$

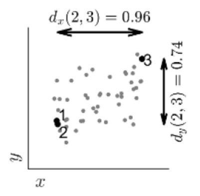
$MGC(\mathcal{X}_n, \mathcal{Y}_n) = 0.15$

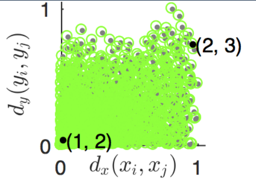
p-vals: $< 0.001$



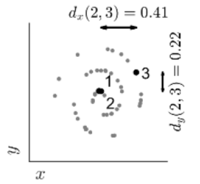



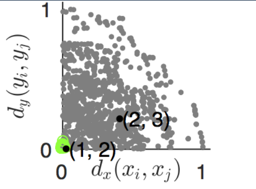
# Examples




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



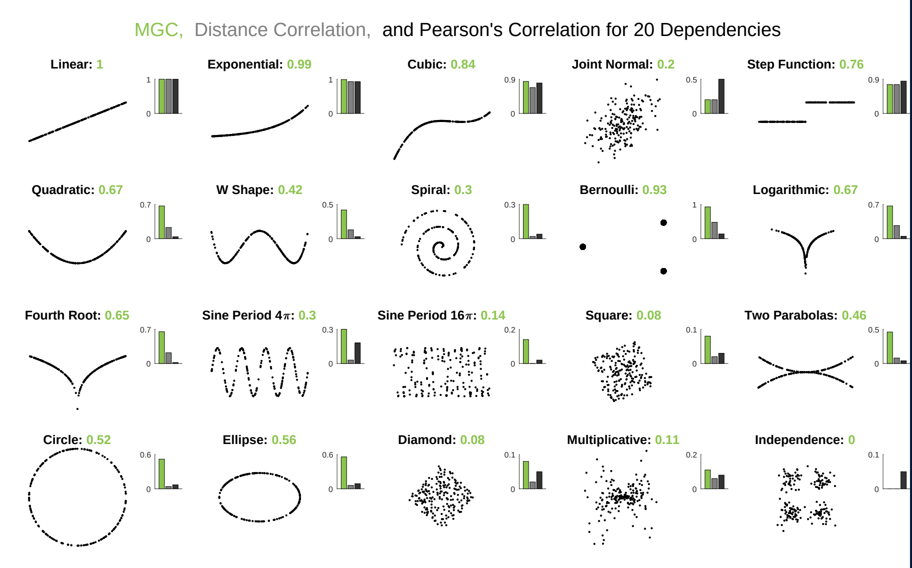




$$Dcorr(\mathcal{X}_n, \mathcal{Y}_n) = 0.01$$
$$MGC(\mathcal{X}_n, \mathcal{Y}_n) = 0.13$$
p-vals: $0.3$ vs $< 0.001$

# Experiments

# Visualizations of $20$ Simulation Settings

# Visualizations of 20 Simulation Settings



# Evaluation Criterion

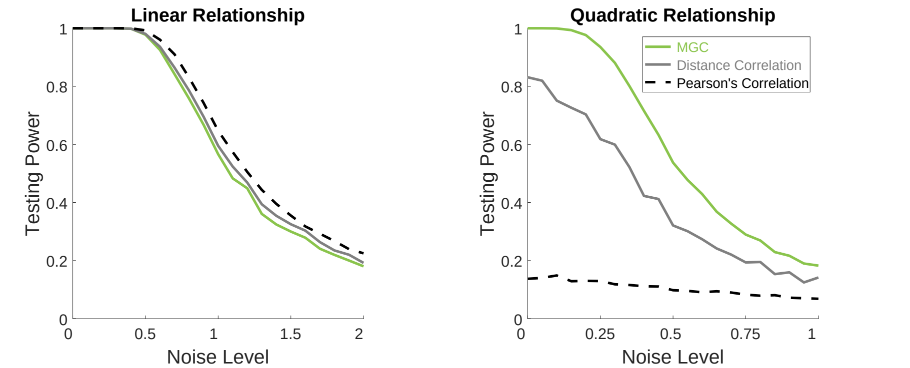
- Power is the probability of rejecting the null when the alternative is true.

# 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 $\beta$ at type 1 error level $\alpha$ 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 |

# Theory

# 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. Invariant:* $c(\mathcal{X}_n, \mathcal{Y}_n) = c(\{\phi(x_i)\}, \{\delta(y_i)\})$ *for any linear transformation* $\phi, \delta$ *(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. *Invariant:* $c(\mathcal{X}_n, \mathcal{Y}_n) = c(\{\phi(x_i)\}, \{\delta(y_i)\})$ *for any linear transformation* $\phi, \delta$ *(i.e., rotation, scaling, translation, reflection).*
4. *1-Linear:* $c(\mathcal{X}_n, \mathcal{Y}_n) = 1$ *under linear relationships.*

# 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. Invariant:* $c(\mathcal{X}_n, \mathcal{Y}_n) = c(\{\phi(x_i)\}, \{\delta(y_i)\})$ *for any linear transformation* $\phi, \delta$ *(i.e., rotation, scaling, translation, reflection).*

*4. 1-Linear:* $c(\mathcal{X}_n, \mathcal{Y}_n) = 1$ *under linear relationships.*

**Theorem 2 (Consistency)**

# 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. Invariant:* $c(\mathcal{X}_n, \mathcal{Y}_n) = c(\{\phi(x_i)\}, \{\delta(y_i)\})$ *for any linear transformation* $\phi, \delta$ *(i.e., rotation, scaling, translation, reflection).*

*4. 1-Linear:* $c(\mathcal{X}_n, \mathcal{Y}_n) = 1$ *under linear relationships.*

**Theorem 2 (Consistency)**

*5. 0-Indep:* $c(\mathcal{X}_n, \mathcal{Y}_n) \stackrel{n\to\infty}{\to} 0$ *if and only if independence.*

# 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. *Invariant:* $c(\mathcal{X}_n, \mathcal{Y}_n) = c(\{\phi(x_i)\}, \{\delta(y_i)\})$ *for any linear transformation* $\phi, \delta$ *(i.e., rotation, scaling, translation, reflection).*
4. *1-Linear:* $c(\mathcal{X}_n, \mathcal{Y}_n) = 1$ *under linear relationships.*

**Theorem 2 (Consistency)**

5. *0-Indep:* $c(\mathcal{X}_n, \mathcal{Y}_n) \stackrel{n\rightarrow\infty}{\rightarrow} 0$ *if and only if independence.*
6. *Consistency: At any type 1 error level* $\alpha$*, testing power* $\beta(c(\mathcal{X}_n, \mathcal{Y}_n)) \stackrel{n\rightarrow\infty}{\rightarrow} 1$ *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

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

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

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

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

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

# 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

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

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

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

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

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

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

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

The population local covariance can be defined as

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

Normalizing and taking a smoothed maximum yield population MGC.

# Sample to Population

# Sample to Population

The population version can be equivalently defined via characteristic functions of $F_{XY}$, e.g.,

# Sample to Population

The population version can be equivalently defined via characteristic functions of $F_{XY}$, e.g.,

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

see [*Shen et al.(2017b)*][2].

# Sample to Population

The population version can be equivalently defined via characteristic functions of $F_{XY}$, e.g.,

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

see [*Shen et al.(2017b)*][2].

**Theorem 3 (Convergence, Mean and Variance)**

# Sample to Population

The population version can be equivalently defined via characteristic functions of $F_{XY}$, e.g.,

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

see [*Shen et al.(2017b)*][2].

**Theorem 3 (Convergence, Mean and Variance)**

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

# Sample to Population

The population version can be equivalently defined via characteristic functions of $F_{XY}$, e.g.,

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

see [*Shen et al.(2017b)*][2].

**Theorem 3 (Convergence, Mean and Variance)**

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

# Sample to Population

The population version can be equivalently defined via characteristic functions of $F_{XY}$, e.g.,

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

see [*Shen et al.(2017b)*][2].

**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:* $E(c(\mathcal{X}_n, \mathcal{Y}_n)) = c(X,Y) + \mathcal{O}(1/n)$.

# Sample to Population

The population version can be equivalently defined via characteristic functions of $F_{XY}$, e.g.,

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

see [*Shen et al.(2017b)*][2].

**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:* $Var(c(\mathcal{X}_n, \mathcal{Y}_n)) = \mathcal{O}(1/n)$.

# Sample to Population

The population version can be equivalently defined via characteristic functions of $F_{XY}$, e.g.,

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

see [*Shen et al.(2017b)*][2].

**Theorem 3 (Convergence, Mean and Variance)**

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

*For a positive definite 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 negative definite, and*

$$Dcorr_d(X,Y) = \textit{normalized } HSIC_k(X,Y).$$

# MGC is applicable to similarity / kernel matrix

**Theorem 6 (Transforming kernel to distance)**

*For a positive definite 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 negative definite, and*

$$Dcorr_d(X,Y) = \textit{normalized } HSIC_k(X,Y).$$

For any sample kernel matrices $K_{n\times n}$, one can transform it to a dissimilarity matrix by

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

and apply MGC / Dcorr to the transformed dissimilarity matrices.

# 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 $\mathcal{O}(n^2 logn)$.

# 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 $\mathcal{O}(n^2 logn)$.

MGC shares the same intrinsic idea as in nonlinear embedding, random forest, multiple kernel learning, deep learning.

# References

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

2. C. Shen, C. E. Priebe, and J. T. Vogelstein, "From distance correlation to the multiscale generalized correlation," *https://arxiv.org/abs/1710.09768*.

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*.