Measuring Association/Predictive power on Topological Spaces Using Kernels and Graphs

Nabarun Deb

Department of Statistics
Columbia University

New England Statistics Symposium 2021

Joint work with Promit Ghosal (MIT), Zhen Huang (Columbia U) and Bodhisattva Sen (Columbia U)

November 13, 2020

https://arxiv.org/pdf/2010.01768.pdf
https://arxiv.org/pdf/2012.14804.pdf

Formal Introduction: Pearson's Correlation and beyond?

- $(X, Y) \sim \mu$ on $\mathcal{X} \times \mathcal{Y}$ (topological spaces) with marginals μ_X, μ_Y
- **Informal goal**: Construct a measure that can capture the

  strength of association between X and Y

  beyond simply testing for independence.

Formal Introduction: Pearson's Correlation and beyond?

- $(X, Y) \sim \mu$ on $\mathcal{X} \times \mathcal{Y}$ (topological spaces) with marginals μ_X, μ_Y
- **Informal goal**: Construct a measure that can capture the

strength of association between X and Y

beyond simply testing for independence.

Formal Introduction: Pearson's Correlation and beyond?

- $(X, Y) \sim \mu$ on $\mathcal{X} \times \mathcal{Y}$ (topological spaces) with marginals μ_X, μ_Y
- **Informal goal**: Construct a measure that can capture the

  strength of association between X and Y

  beyond simply testing for independence.

Motivation: Pearson's correlation

- Given $(X, Y) \sim \nu \equiv$ bivariate normal, the Pearson's correlation ρ_{XY} measures the strength of association
- $\rho_{XY} = 0$ iff X and Y are independent

Formal Introduction: Pearson's Correlation and beyond?

- $(X, Y) \sim \mu$ on $\mathcal{X} \times \mathcal{Y}$ (topological spaces) with marginals μ_X, μ_Y
- **Informal goal**: Construct a measure that can capture the

  strength of association between X and Y

  beyond simply testing for independence.

Motivation: Pearson's correlation

- Given $(X, Y) \sim \nu \equiv$ bivariate normal, the Pearson's correlation ρ_{XY} measures the strength of association
- $\rho_{XY} = 0$ iff X and Y are independent
- ρ_{XY} approaches its maximum absolute value (i.e., 1) iff one variable looks more and more like a noiseless linear function of the other, i.e., $Y = cX + d$.

Formal Introduction: Pearson's Correlation and beyond?

- $(X, Y) \sim \mu$ on $\mathcal{X} \times \mathcal{Y}$ (topological spaces) with marginals μ_X, μ_Y
- **Informal goal**: Construct a measure that can capture the

  strength of association between X and Y

  beyond simply testing for independence.

Motivation: Pearson's correlation

- Given $(X, Y) \sim \nu \equiv$ bivariate normal, the Pearson's correlation ρ_{XY} measures the strength of association
- $\rho_{XY} = 0$ iff X and Y are independent
- ρ_{XY} approaches its maximum absolute value (i.e., 1) iff one variable looks more and more like a noiseless linear function of the other, i.e., $Y = cX + d$.

What are truly nonparametric analogs of the Pearson's correlation?

Think of nonparametric regression

- This asymmetry is fundamental in even simple regression problems, consider the noiseless version:

$$Y = f(X).$$

- If $f(\cdot)$ is a many-to-one function, predicting X from Y is not possible whereas predicting Y from X is immediate irrespective of $f(\cdot)$.

- Pearson's correlation being symmetric cannot distinguish between the two problems — same is the case for most measures of dependence.

- Design a directional measure that

  1. is "small" for "predicting" X from Y.
  2. but large for "predicting" Y from X.

Introduction

- Want a measure that equals 0 iff $X \perp\!\!\!\perp Y$, equals 1 iff Y is "some function" of X.

- For the past century, most measures of association/dependence only focus on testing for independence, i.e., they equal 0 iff $Y \perp\!\!\!\perp X$; e.g., distance correlation (Székely et al., 2007), Hilbert-Schmidt independence criterion (Gretton et al., 2008), graph-based measures (Friedman and Rafsky, 1983), etc.

Introduction

- Want a measure that equals 0 iff $X \perp\!\!\!\perp Y$, equals 1 iff Y is "some function" of X.

- For the past century, most measures of association/dependence only focus on testing for independence, i.e., they equal 0 iff $Y \perp\!\!\!\perp X$; e.g., distance correlation (Székely et al., 2007), Hilbert-Schmidt independence criterion (Gretton et al., 2008), graph-based measures (Friedman and Rafsky, 1983), etc.

Recent advances

- In Dette et al., 2013, Chatterjee, 2019. When $\mathcal{X} = \mathcal{Y} = \mathbb{R}$, authors propose measures that equal 0 iff $Y \perp\!\!\!\perp X$ and 1 iff Y is a measurable function of X. Extended to the case $\mathcal{X} = \mathbb{R}^{d_1}$ and $\mathcal{Y} = \mathbb{R}$ in Azadkia and Chatterjee, 2019.

Introduction

- Want a measure that equals 0 iff $X \perp\!\!\!\perp Y$, equals 1 iff Y is "some function" of X.

- For the past century, most measures of association/dependence only focus on testing for independence, i.e., they equal 0 iff $Y \perp\!\!\!\perp X$; e.g., distance correlation (Székely et al., 2007), Hilbert-Schmidt independence criterion (Gretton et al., 2008), graph-based measures (Friedman and Rafsky, 1983), etc.

Recent advances

- In Dette et al., 2013, Chatterjee, 2019. When $\mathcal{X} = \mathcal{Y} = \mathbb{R}$, authors propose measures that equal 0 iff $Y \perp\!\!\!\perp X$ and 1 iff Y is a measurable function of X. Extended to the case $\mathcal{X} = \mathbb{R}^{d_1}$ and $\mathcal{Y} = \mathbb{R}$ in Azadkia and Chatterjee, 2019.

- Bottleneck: They rely on the canonical ordering of $\mathbb{R}$.

Introduction

- Want a measure that equals 0 iff $X \perp\!\!\!\perp Y$, equals 1 iff Y is measurable function of X.

- For the past century, most measures of association/dependence only focus on testing for independence, i.e., they equal 0 iff $Y \perp\!\!\!\perp X$; e.g., distance correlation (Székely et al., 2007), Hilbert-Schmidt independence criterion (Gretton et al., 2008), graph-based measures (Friedman and Rafsky, 1983), etc.

Recent advances

- In Dette et al., 2013, Chatterjee, 2019. When $\mathcal{X} = \mathcal{Y} = \mathbb{R}$, authors propose measures that equal 0 iff $Y \perp\!\!\!\perp X$ and 1 iff Y is a measurable function of X. Extended to the case $\mathcal{X} = \mathbb{R}^{d_1}$ and $\mathcal{Y} = \mathbb{R}$ in Azadkia and Chatterjee, 2019.

- Bottleneck: They rely on the canonical ordering of $\mathbb{R}$.

Structure

1. A family of measures of association
   - A measure on $\mathcal{X} = \mathbb{R}^{d_1}$, $\mathcal{Y} = \mathbb{R}^{d_2}$
   - Interpretability and monotonicity
   - Extending to a class of kernel measures
2. Estimating the kernel measure
   - Proposing the estimator
   - Computational complexity
   - Consistency
   - Rate of estimation
   - A central limit theorem when $X \perp\!\!\!\perp Y$

Outline

1. A family of measures of association
   - A measure on $\mathcal{X} = \mathbb{R}^{d_1}$, $\mathcal{Y} = \mathbb{R}^{d_2}$
   - Interpretability and monotonicity
   - Extending to a class of kernel measures
2. Estimating the kernel measure
   - Proposing the estimator
   - Computational complexity
   - Consistency
   - Rate of estimation
   - A central limit theorem when $X \perp\!\!\!\perp Y$

A measure on $\mathcal{X} = \mathbb{R}^{d_1}$, $\mathcal{Y} = \mathbb{R}^{d_2}$

Basic strategy

- Most measures of dependence quantify a "discrepancy" between μ and $\mu_X \otimes \mu_Y$.

A measure on $\mathcal{X} = \mathbb{R}^{d_1}$, $\mathcal{Y} = \mathbb{R}^{d_2}$

Basic strategy

- Most measures of dependence quantify a "discrepancy" between μ and $\mu_X \otimes \mu_Y$.
- We construct a discrepancy between $\mu_{Y|X}$ (regular conditional distribution) and μ_Y.
- When $Y \perp\!\!\!\perp X$, $\mu_{Y|X} = \mu_Y$. When Y is a measurable function of X, $\mu_{Y|X}$ is a degenerate measure.

- Define

$$T \equiv T(\mu) := 1 - \frac{\mathbb{E}\|Y' - \tilde{Y}'\|_2}{\mathbb{E}\|Y_1 - Y_2\|_2}.$$

A measure on $\mathcal{X} = \mathbb{R}^{d_1}$, $\mathcal{Y} = \mathbb{R}^{d_2}$

Basic strategy

- Most measures of dependence quantify a "discrepancy" between μ and $\mu_X \otimes \mu_Y$.
- We construct a discrepancy between $\mu_{Y|X}$ (regular conditional distribution) and μ_Y.
- When $Y \perp\!\!\!\perp X$, $\mu_{Y|X} = \mu_Y$. When Y is a measurable function of X, $\mu_{Y|X}$ is a degenerate measure.

- Define

$$T \equiv T(\mu) := 1 - \frac{\mathbb{E}\|Y' - \tilde{Y}'\|_2}{\mathbb{E}\|Y_1 - Y_2\|_2}.$$

- Generate $Y_1, Y_2 \overset{i.i.d.}{\sim} \mu_Y$.

A measure on $\mathcal{X} = \mathbb{R}^{d_1}$, $\mathcal{Y} = \mathbb{R}^{d_2}$

Basic strategy

- Most measures of dependence quantify a "discrepancy" between μ and $\mu_X \otimes \mu_Y$.
- We construct a discrepancy between $\mu_{Y|X}$ (regular conditional distribution) and μ_Y.
- When $Y \perp\!\!\!\perp X$, $\mu_{Y|X} = \mu_Y$. When Y is a measurable function of X, $\mu_{Y|X}$ is a degenerate measure.

- Define

$$T \equiv T(\mu) := 1 - \frac{\mathbb{E}\|Y' - \tilde{Y}'\|_2}{\mathbb{E}\|Y_1 - Y_2\|_2}.$$

- Generate $Y_1, Y_2 \overset{i.i.d.}{\sim} \mu_Y$.
- $(X', Y', \tilde{Y}')$ is generated as: draw $X' \sim \mu_X$ and then $Y'|X' \sim \mu_{Y|X'}$, $\tilde{Y}'|X' \sim \mu_{Y|X'}$ such that Y' and $\tilde{Y}'$ are conditionally independent given X'.

Some intuition

- Suppose $d_2 = 1$.
- Consider a slight modification:

$$T^* \equiv T^*(\mu) := 1 - \frac{\mathbb{E}|Y' - \tilde{Y}'|^2}{\mathbb{E}|Y_1 - Y_2|^2}.$$

- Plug-in $\mathbb{E}|Y' - \tilde{Y}'|^2 = \mathbb{E}|Y'|^2 + \mathbb{E}|\tilde{Y}'|^2 - 2\mathbb{E}Y'\tilde{Y}'$.
- Do the same for the denominator.
- Simplify $T^*(\mu)$ to get:

$$T^*(\mu) = \frac{\mathrm{Var}(\mathbb{E}[Y|X])}{\mathrm{Var}(Y)} \in [0, 1].$$

- T can be interpreted as the proportion of the variance of Y explained by X.

Back to $T(\mu)$ — More intuition

- Recall $X' \sim \mu_X$ and $Y'|X' \sim \mu_{Y|X'}$, $\tilde{Y}'|X' \sim \mu_{Y|X'}$ such that Y' and $\tilde{Y}'$ are conditionally independent given X'.

$$T = 1 - \frac{\mathbb{E}\|Y' - \tilde{Y}'\|_2}{\mathbb{E}\|Y_1 - Y_2\|_2}.$$

$Y' \sim \mu_Y$, $\tilde{Y}' \sim \mu_Y$ but Y' and $\tilde{Y}'$ are not independent.

Back to $T(\mu)$ — More intuition

- Recall $X' \sim \mu_X$ and $Y'|X' \sim \mu_{Y|X'}$, $\tilde{Y}'|X' \sim \mu_{Y|X'}$ such that Y' and $\tilde{Y}'$ are conditionally independent given X'.

$$T = 1 - \frac{\mathbb{E}\|Y' - \tilde{Y}'\|_2}{\mathbb{E}\|Y_1 - Y_2\|_2}.$$

  $Y' \sim \mu_Y$, $\tilde{Y}' \sim \mu_Y$ but Y' and $\tilde{Y}'$ are not independent.

- Suppose $Y \perp\!\!\!\perp X$, then

$$\mu_{Y|X'} = \mu_Y, Y', \tilde{Y}' \overset{i.i.d.}{\sim} \mu_Y$$

  and so $T = 0$.

Back to $T(\mu)$ — More intuition

- Recall $X' \sim \mu_X$ and $Y'|X' \sim \mu_{Y|X'}$, $\tilde{Y}'|X' \sim \mu_{Y|X'}$ such that Y' and $\tilde{Y}'$ are conditionally independent given X'.

$$T = 1 - \frac{\mathbb{E}\|Y' - \tilde{Y}'\|_2}{\mathbb{E}\|Y_1 - Y_2\|_2}.$$

$Y' \sim \mu_Y$, $\tilde{Y}' \sim \mu_Y$ but Y' and $\tilde{Y}'$ are not independent.

- Suppose $Y \perp\!\!\!\perp X$, then

$$\mu_{Y|X'} = \mu_Y, Y', \tilde{Y}' \overset{i.i.d.}{\sim} \mu_Y$$

and so $T = 0$.

- Suppose $Y = h(X)$ for some measurable $h(\cdot)$, then

$$Y' = \tilde{Y}' = h(X'), \quad \|Y' - \tilde{Y}'\|_2 = 0$$

and so $T = 1$.

A formal result

Theorem

Suppose $\mathbb{E}\|Y_1\|_2 < \infty$. Then

- $T \in [0, 1]$.
- $T = 0$ iff $Y \perp\!\!\!\perp X$.
- $T = 1$ iff Y is a noiseless measurable function of X.

A formal result

Theorem

Suppose $\mathbb{E}\|Y_1\|_2 < \infty$. Then

- $T \in [0, 1]$.
- $T = 0$ iff $Y \perp\!\!\!\perp X$.
- $T = 1$ iff Y is a noiseless measurable function of X.

- The choice $\|\cdot\|_2$ is important. For instance,

$$1 - \frac{\mathbb{E}\|Y' - \tilde{Y}'\|_2^2}{\mathbb{E}\|Y_1 - Y_2\|_2^2}$$

can be 0 even when $Y \not\perp\!\!\!\perp X$.

Monotonicity

What happens in the interval $(0, 1)$?

Monotonicity

What happens in the interval $(0, 1)$?

T for bivariate normal

Suppose μ is the bivariate normal distribution with means μ_X, μ_Y, variances σ_X^2, σ_Y^2 and correlation ρ. Then

$$T(\mu) = 1 - \sqrt{1 - \rho^2}.$$

The above function is strictly convex and increasing in $|\rho|$.

Monotonicity

What happens in the interval $(0, 1)$?

T for bivariate normal

Suppose μ is the bivariate normal distribution with means μ_X, μ_Y, variances σ_X^2, σ_Y^2 and correlation ρ. Then

$$T(\mu) = 1 - \sqrt{1 - \rho^2}.$$

The above function is strictly convex and increasing in $|\rho|$.

Other examples: Let

$$Y = \lambda g(X) + \epsilon$$

where $\lambda \geq 0$, ϵ, X are independent, $\epsilon' \overset{i.i.d.}{\sim} \epsilon$ such that $\epsilon - \epsilon'$ is unimodal. Then $T(\mu)$ is montonic in λ.

Monotonicity

What happens in the interval $(0, 1)$?

T for bivariate normal

Suppose μ is the bivariate normal distribution with means μ_X, μ_Y, variances σ_X^2, σ_Y^2 and correlation ρ. Then

$$T(\mu) = 1 - \sqrt{1 - \rho^2}.$$

The above function is strictly convex and increasing in $|\rho|$.

Other examples: Let

$$Y = \lambda g(X) + \epsilon$$

where $\lambda \geq 0$, ϵ, X are independent, $\epsilon' \overset{i.i.d.}{\sim} \epsilon$ such that $\epsilon - \epsilon'$ is unimodal. Then $T(\mu)$ is montonic in λ.

In nonparametric regression models with additive noise, T turns out to be a monotonic function of the noise variance.

Preliminaries: reproducing kernel Hilbert spaces (RKHS)

- RKHS on $\mathcal{Y}$: linear, complete, inner product space of functions from $\mathcal{Y} \to \mathbb{R}$; non-negative definite kernel; "reproducing property".
- Consider a non-negative definite kernel function on $\mathcal{Y}$ — $K : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ satisfying

$$\sum_{i,j=1}^{m} \alpha_i \alpha_j K(y_i, y_j) \geq 0$$

for all $\alpha_i \in \mathbb{R}$, $y_i \in \mathcal{Y}$ and $m \geq 1$.

Preliminaries: reproducing kernel Hilbert spaces (RKHS)

- RKHS on $\mathcal{Y}$: linear, complete, inner product space of functions from $\mathcal{Y} \to \mathbb{R}$; non-negative definite kernel; "reproducing property".
- Consider a non-negative definite kernel function on $\mathcal{Y}$ — $K : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ satisfying

$$\sum_{i,j=1}^{m} \alpha_i \alpha_j K(y_i, y_j) \geq 0$$

  for all $\alpha_i \in \mathbb{R}$, $y_i \in \mathcal{Y}$ and $m \geq 1$.
- Note $K(y, \cdot) : \mathcal{Y} \to \mathbb{R}$.
- Identify $y \mapsto K(y, \cdot)$ (feature map).
- (Reproducing property) For all $f \in \mathcal{H}$, $y \in \mathcal{Y}$, $\langle f, K(y, \cdot)\rangle_{\mathcal{H}} = f(y)$.

RKHS (continued) - useful identities

As a consequence of the reproducing property:

- $\langle K(y_1, \cdot), K(y_2, \cdot) \rangle_{\mathcal{H}} = K(y_1, y_2)$.

RKHS (continued) - useful identities

As a consequence of the reproducing property:

- $\langle K(y_1, \cdot), K(y_2, \cdot) \rangle_{\mathcal{H}} = K(y_1, y_2)$.
- Using the above,

$$\begin{aligned} & \|K(y_1, \cdot) - K(y_2, \cdot)\|^2_{\mathcal{H}} \\ &= \langle K(y_1, \cdot), K(y_1, \cdot) \rangle_{\mathcal{H}} + \langle K(y_2, \cdot), K(y_2, \cdot) \rangle_{\mathcal{H}} - 2\langle K(y_1, \cdot), K(y_2, \cdot) \rangle_{\mathcal{H}} \\ &= K(y_1, y_1) + K(y_2, y_2) - 2K(y_1, y_2). \end{aligned}$$

Kernel measure of association (KMAc)

- Recall $K(y, \cdot) : \mathcal{Y} \to \mathbb{R}$ for all $y \in \mathcal{Y}$, y identified with $K(y, \cdot)$ and

$$T = 1 - \frac{\mathbb{E}\| Y' - \tilde{Y}' \|_2}{\mathbb{E}\| Y_1 - Y_2 \|_2}.$$

- Replace $Y_1 - Y_2$ with $K(Y_1, \cdot) - K(Y_2, \cdot)$.

Kernel measure of association (KMAc)

- Recall $K(y, \cdot) : \mathcal{Y} \to \mathbb{R}$ for all $y \in \mathcal{Y}$, y identified with $K(y, \cdot)$ and

$$T = 1 - \frac{\mathbb{E}\| Y' - \tilde{Y}'\|_2}{\mathbb{E}\| Y_1 - Y_2\|_2}.$$

- Replace $\| Y_1 - Y_2\|_2$ with $\|K(Y_1, \cdot) - K(Y_2, \cdot)\|_{\mathcal{H}}^2$.
- Define

$$\eta_K := 1 - \frac{\mathbb{E}\|K(Y', \cdot) - K(\tilde{Y}', \cdot)\|_{\mathcal{H}}^2}{\mathbb{E}\|K(Y_1, \cdot) - K(Y_2, \cdot)\|_{\mathcal{H}}^2}$$

Kernel measure of association (KMAc)

- Recall $K(y, \cdot) : \mathcal{Y} \to \mathbb{R}$ for all $y \in \mathcal{Y}$, y identified with $K(y, \cdot)$ and

$$T = 1 - \frac{\mathbb{E}\| Y' - \tilde{Y}' \|_2}{\mathbb{E}\| Y_1 - Y_2 \|_2}.$$

- Replace $\| Y_1 - Y_2 \|_2$ with $\| K(Y_1, \cdot) - K(Y_2, \cdot) \|_{\mathcal{H}}^2$.
- Define

$$\begin{aligned} \eta_K &:= 1 - \frac{\mathbb{E}\| K(Y', \cdot) - K(\tilde{Y}', \cdot) \|_{\mathcal{H}}^2}{\mathbb{E}\| K(Y_1, \cdot) - K(Y_2, \cdot) \|_{\mathcal{H}}^2} \\ &= 1 - \frac{\mathbb{E}K(Y', Y') + \mathbb{E}K(\tilde{Y}', \tilde{Y}') - 2\mathbb{E}K(Y', \tilde{Y}')}{\mathbb{E}K(Y_1, Y_1) + \mathbb{E}K(Y_2, Y_2) - 2\mathbb{E}K(Y_1, Y_2)} \end{aligned}$$

Kernel measure of association (KMAc)

- Recall $K(y, \cdot) : \mathcal{Y} \to \mathbb{R}$ for all $y \in \mathcal{Y}$, y identified with $K(y, \cdot)$ and

$$T = 1 - \frac{\mathbb{E}\| Y' - \tilde{Y}'\|_2}{\mathbb{E}\| Y_1 - Y_2\|_2}.$$

- Replace $\| Y_1 - Y_2\|_2$ with $\|K(Y_1, \cdot) - K(Y_2, \cdot)\|_{\mathcal{H}}^2$.
- Define

$$\begin{aligned}\eta_K &:= 1 - \frac{\mathbb{E}\|K(Y', \cdot) - K(\tilde{Y}', \cdot)\|_{\mathcal{H}}^2}{\mathbb{E}\|K(Y_1, \cdot) - K(Y_2, \cdot)\|_{\mathcal{H}}^2} \\ &= 1 - \frac{\mathbb{E}K(Y', Y') + \mathbb{E}K(\tilde{Y}', \tilde{Y}') - 2\mathbb{E}K(Y', \tilde{Y}')}{\mathbb{E}K(Y_1, Y_1) + \mathbb{E}K(Y_2, Y_2) - 2\mathbb{E}K(Y_1, Y_2)}\end{aligned}$$

Kernel measure of association (KMAc)

- Recall $K(y,\cdot): \mathcal{Y} \to \mathbb{R}$ for all $y \in \mathcal{Y}$, y identified with $K(y,\cdot)$ and

$$T = 1 - \frac{\mathbb{E}\|Y' - \tilde{Y}'\|_2}{\mathbb{E}\|Y_1 - Y_2\|_2}.$$

- Replace $\|Y_1 - Y_2\|_2$ with $\|K(Y_1,\cdot) - K(Y_2,\cdot)\|_{\mathcal{H}}^2$.
- Define

$$\begin{aligned}
\eta_K &:= 1 - \frac{\mathbb{E}\|K(Y',\cdot) - K(\tilde{Y}',\cdot)\|_{\mathcal{H}}^2}{\mathbb{E}\|K(Y_1,\cdot) - K(Y_2,\cdot)\|_{\mathcal{H}}^2} \\
&= 1 - \frac{\mathbb{E}K(Y',Y') + \mathbb{E}K(\tilde{Y}',\tilde{Y}') - 2\mathbb{E}K(Y',\tilde{Y}')}{\mathbb{E}K(Y_1,Y_1) + \mathbb{E}K(Y_2,Y_2) - 2\mathbb{E}K(Y_1,Y_2)} \\
&= \frac{\mathbb{E}K(Y',\tilde{Y}') - \mathbb{E}K(Y_1,Y_2)}{\mathbb{E}K(Y_1,Y_1) - \mathbb{E}K(Y_1,Y_2)}.
\end{aligned}$$

KMAc (continued)

Theorem (informal)

Suppose $K(\cdot, \cdot)$ is characteristic and $\mathbb{E}K(Y_1, Y_1) < \infty$, then:

- $\eta_K \in [0, 1]$.
- $\eta_k = 0$ iff $Y \perp\!\!\!\perp X$.
- $\eta_K = 1$ iff Y is a noiseless measurable function of X.

KMAc (continued)

Theorem (informal)

Suppose $K(\cdot, \cdot)$ is characteristic and $\mathbb{E}K(Y_1, Y_1) < \infty$, then:

- $\eta_K \in [0, 1]$.
- $\eta_k = 0$ iff $Y \perp\!\!\!\perp X$.
- $\eta_K = 1$ iff Y is a noiseless measurable function of X.

- A kernel is characteristic if

$$\mathbb{E}_P[K(Y, \cdot)] = \mathbb{E}_Q[K(Y, \cdot)] \implies P = Q$$

for probability measures P and Q.

Examples

Characteristic kernels — Gretton et al., 2012, Sejdinovic et al., 2013, Lyons 2013, 2014. Some examples include:

- (Distance) $K(y_1, y_2) := \|y_1\|_2 + \|y_2\|_2 - \|y_1 - y_2\|_2$. In this case,

$$\eta_K = T.$$

Examples

Characteristic kernels — Gretton et al., 2012, Sejdinovic et al., 2013, Lyons 2013, 2014. Some examples include:

- (Distance) $K(y_1, y_2) := \|y_1\|_2 + \|y_2\|_2 - \|y_1 - y_2\|_2$. In this case,

$$\eta_K = T.$$

- Bounded kernels: (Gaussian) $K(y_1, y_2) := \exp(-\|y_1 - y_2\|_2^2)$ and (Laplacian) $K(y_1, y_2) := \exp(-\|y_1 - y_2\|_1)$.

- For non-Euclidean domains such as video filtering, robotics, text documents, human action recognition, characteristic kernels constructed in Fukumizu et al., 2009, Danafar et al., 2010, Christmann and Steinwart, 2010.

Outline

1. A family of measures of association
   - A measure on $\mathcal{X} = \mathbb{R}^{d_1}$, $\mathcal{Y} = \mathbb{R}^{d_2}$
   - Interpretability and monotonicity
   - Extending to a class of kernel measures
2. Estimating the kernel measure
   - Proposing the estimator
   - Computational complexity
   - Consistency
   - Rate of estimation
   - A central limit theorem when $X \perp\!\!\!\perp Y$

Estimation strategy

- Suppose $(X_1, Y_1), \ldots, (X_n, Y_n) \sim \mu$.
- $\mathcal{X}$ is endowed with metric $\rho_{\mathcal{X}}(\cdot, \cdot)$.
- Recall

$$\eta_K = \frac{\mathbb{E}K(Y', \tilde{Y}') - \mathbb{E}K(Y_1, Y_2)}{\mathbb{E}K(Y_1, Y_1) - \mathbb{E}K(Y_1, Y_2)}.$$

Estimation strategy

- Suppose $(X_1, Y_1), \ldots, (X_n, Y_n) \sim \mu$.
- $\mathcal{X}$ is endowed with metric $\rho_{\mathcal{X}}(\cdot, \cdot)$.
- Recall

$$\eta_K = \frac{\mathbb{E}K(Y', \tilde{Y}') - \mathbb{E}K(Y_1, Y_2)}{\mathbb{E}K(Y_1, Y_1) - \mathbb{E}K(Y_1, Y_2)}.$$

- From standard U-Statistic theory,

$$\mathbb{E}K(Y_1, Y_1) \approx \frac{1}{n}\sum_{i=1}^{n} K(Y_i, Y_i)$$

and

$$\frac{1}{n}\sum_{i=1}^{n} K(Y_i, Y_{i+1}) \approx \mathbb{E}K(Y_1, Y_2) \approx \frac{1}{n(n-1)}\sum_{i \neq j} K(Y_i, Y_j).$$

- Hardest term to estimate is $\mathbb{E}K(Y', \tilde{Y}')$.

Estimation (continued)

- Suppose X is supported on a finite set. A natural estimator

$$\mathbb{E}[\mathbb{E}[K(Y', \tilde{Y}')|X']] \approx \frac{1}{n}\sum_{i=1}^{n} \frac{1}{|\{j : X_j = X_i\}|} \sum_{j:X_j=X_i} K(Y_i, Y_j).$$

Estimation (continued)

- Suppose X is supported on a finite set. A natural estimator

$$\mathbb{E}[\mathbb{E}[K(Y', \tilde{Y}')|X']] \approx \frac{1}{n}\sum_{i=1}^{n} \frac{1}{|\{j : X_j = X_i\}|} \sum_{j : X_j = X_i} K(Y_i, Y_j).$$

- If X is continuous, replace $X_j = X_i$ with $\rho_{\mathcal{X}}(X_i, X_j)$ being "small".
- Construct a graph G_n on $\{X_1, \ldots, X_n\}$ which joins points that are "close" to each other.
- For example, consider a k-nearest neighbor graph (k-NNG) - join every point to its first k nearest neighbors.

Estimation (continued)

- Replace

$$\frac{1}{n}\sum_{i=1}^{n}\frac{1}{|\{j : X_j = X_i\}|}\sum_{j:X_j=X_i}K(Y_i, Y_j)$$

with

$$\frac{1}{n}\sum_{i=1}^{n}\frac{1}{d_i}\sum_{j:(i,j)\in E(G_n)}K(Y_i, Y_j)$$

where $E(G_n)$ — edge/neighbor set of G_n and d_i — degree of X_i.

Estimation (continued)

- Replace

$$\frac{1}{n}\sum_{i=1}^{n}\frac{1}{|\{j : X_j = X_i\}|}\sum_{j:X_j=X_i}K(Y_i, Y_j)$$

with

$$\frac{1}{n}\sum_{i=1}^{n}\frac{1}{d_i}\sum_{j:(i,j)\in E(G_n)}K(Y_i, Y_j)$$

where $E(G_n)$ — edge/neighbor set of G_n and d_i — degree of X_i.

- Define

$$\hat{\eta}_n := \frac{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{d_i}\sum_{j:(i,j)\in E(G_n)}K(Y_i, Y_j) - \frac{1}{n(n-1)}\sum_{i\neq j}K(Y_i, Y_j)}{\frac{1}{n}\sum_{i=1}^{n}K(Y_i, Y_i) - \frac{1}{n(n-1)}\sum_{i\neq j}K(Y_i, Y_j)}.$$

$$\hat{\eta}_n^{\text{lin}} := \frac{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{d_i}\sum_{j:(i,j)\in E(G_n)}K(Y_i, Y_j) - \frac{1}{n}\sum_{i=1}^{N}K(Y_i, Y_{i+1})}{\frac{1}{n}\sum_{i=1}^{n}K(Y_i, Y_i) - \frac{1}{n}\sum_{i=1}^{N}K(Y_i, Y_{i+1})}.$$

Computational complexity

- Suppose G_n is the k-NNG; computed in $\mathcal{O}(kn\log n)$ time.
- Recall

$$\hat{\eta}_n^{\text{lin}} = \frac{\underbrace{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{d_i}\sum_{j:(i,j)\in E(G_n)} K(Y_i, Y_j)}_{\mathcal{O}(kn\log n)} - \frac{1}{n}\sum_{i=1}^{N} K(Y_i, Y_{i+1})}{\underbrace{\frac{1}{n}\sum_{i=1}^{n} K(Y_i, Y_i)}_{\mathcal{O}(n)} - \frac{1}{n}\sum_{i=1}^{N} K(Y_i, Y_{i+1})}.$$

Computational complexity

- Suppose G_n is the k-NNG; computed in $\mathcal{O}(kn\log n)$ time.
- Recall

$$\hat{\eta}_n^{\text{lin}} = \frac{\underbrace{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{d_i}\sum_{j:(i,j)\in E(G_n)} K(Y_i, Y_j)}_{\mathcal{O}(kn\log n)} - \frac{1}{n}\sum_{i=1}^{N} K(Y_i, Y_{i+1})}{\underbrace{\frac{1}{n}\sum_{i=1}^{n} K(Y_i, Y_i)}_{\mathcal{O}(n)} - \frac{1}{n}\sum_{i=1}^{N} K(Y_i, Y_{i+1})}.$$

- $\hat{\eta}_n^{\text{lin}}$ is computable in near linear time as opposed to $\hat{\eta}_n$ which may be quadratic. In practice, for certain kernels, one may compute $\hat{\eta}_n$ approximately, in near linear time.

Estimation (continued)

Theorem (informal)

Suppose G_n satisfies the "close"-ness condition in the sense that:

$$\frac{\sum_{(i,j)\in E(G_n)} \rho_{\mathcal{X}}(X_i, X_j)}{|E(G_n)|} \xrightarrow{\mathbb{P}} 0$$

and $\mathbb{E}K(Y_1, Y_1)^{2+\epsilon} < \infty$, then

$$\hat{\eta}_n \xrightarrow{\mathbb{P}} \eta_K, \qquad \hat{\eta}_n^{\text{lin}} \xrightarrow{\mathbb{P}} \eta_K.$$

Estimation (continued)

Theorem (informal)

Suppose G_n satisfies the "close"-ness condition in the sense that:

$$\frac{\sum_{(i,j)\in E(G_n)} \rho_{\mathcal{X}}(X_i, X_j)}{|E(G_n)|} \xrightarrow{\mathbb{P}} 0$$

and $\mathbb{E}K(Y_1, Y_1)^{2+\epsilon} < \infty$, then

$$\hat{\eta}_n \xrightarrow{\mathbb{P}} \eta_K, \qquad \hat{\eta}_n^{\text{lin}} \xrightarrow{\mathbb{P}} \eta_K.$$

- Under additional moments, convergence happens almost surely in μ (not required if bounded kernels are used).
- No smoothness assumption needed on $\mathbb{E}K[(\cdot, Y)|X]$.

Examples of graphs (Euclidean)

- Minimum spanning trees, k-nearest neighbor graphs - join every point to its first k nearest neighbors.
- For k-NNG, $\hat{\eta}_n$ is consistent provided $k = o(n/\log n)$.

Examples of graphs (Euclidean)

- Minimum spanning trees, k-nearest neighbor graphs - join every point to its first k nearest neighbors.
- For k-NNG, $\hat{\eta}_n$ is consistent provided $k = o(n/\log n)$.
- Recall

$$\hat{\eta}_n - \eta_K = \underbrace{(\hat{\eta}_n - \mathbb{E}\hat{\eta}_n)}_{\text{Variance term}} + \underbrace{(\mathbb{E}\hat{\eta}_n - \eta_K)}_{\text{Bias term}}$$

. The bias $\uparrow$ with k. However the variances stabilizes because

$$\frac{1}{n}\sum_{i=1}^{n}\frac{1}{d_i}\sum_{j:(i,j)\in E(G_n)} K(Y_i, Y_j).$$

- For consistent estimation, a 1-NNG can be chosen (no tuning required).

Rate of estimation (k-NNG)

Theorem (informal)

Suppose $K(\cdot,\cdot)$ is bounded, $\mathbb{E}[K(Y,\cdot)|X=x]$ is Lipschitz with respect to $\rho_{\mathcal{X}}(\cdot,\cdot)$ and the support of μ_X has intrinsic dimension d_0. Then

$$\hat{\eta}_n^{\text{lin}} - \eta_K = \begin{cases} \mathcal{O}_{\mathbb{P}}((\sqrt{k/n})(\log n)) & \text{if } d_0 \le 2, \\ \mathcal{O}_{\mathbb{P}}((k/n)^{1/d_0}(\log n)) & \text{if } d_0 > 2. \end{cases}$$

Rate of estimation (k-NNG)

Theorem (informal)

Suppose $K(\cdot,\cdot)$ is bounded, $\mathbb{E}[K(Y,\cdot)|X=x]$ is Lipschitz with respect to $\rho_{\mathcal{X}}(\cdot,\cdot)$ and the support of μ_X has intrinsic dimension d_0. Then

$$\hat{\eta}_n^{\text{lin}} - \eta_K = \begin{cases} \mathcal{O}_{\mathbb{P}}((\sqrt{k/n})(\log n)) & \text{if } d_0 \leq 2, \\ \mathcal{O}_{\mathbb{P}}((k/n)^{1/d_0}(\log n)) & \text{if } d_0 > 2. \end{cases}$$

- The rate of estimation adapts to the intrinsic dimension of μ_X (extension of Azadkia and Chatterjee, 2019).
- Recall

$$\hat{\eta}_n - \eta_K = \underbrace{(\hat{\eta}_n - \mathbb{E}\hat{\eta}_n)}_{\text{Variance term}_{\sim n^{-1/2}}} + \underbrace{(\mathbb{E}\hat{\eta}_n - \eta_K)}_{\text{Bias term}_{\uparrow k}}.$$

Rate of estimation (k-NNG)

Theorem (informal)

Suppose $K(\cdot,\cdot)$ is bounded, $\mathbb{E}[K(Y,\cdot)|X=x]$ is Lipschitz with respect to $\rho_{\mathcal{X}}(\cdot,\cdot)$ and the support of μ_X has intrinsic dimension d_0. Then

$$\hat{\eta}_n^{\text{lin}} - \eta_K = \begin{cases} \mathcal{O}_{\mathbb{P}}((\sqrt{k/n})(\log n)) & \text{if } d_0 \leq 2, \\ \mathcal{O}_{\mathbb{P}}((k/n)^{1/d_0}(\log n)) & \text{if } d_0 > 2. \end{cases}$$

- The rate of estimation adapts to the intrinsic dimension of μ_X (extension of Azadkia and Chatterjee, 2019).
- Recall

$$\hat{\eta}_n - \eta_K = \underbrace{(\hat{\eta}_n - \mathbb{E}\hat{\eta}_n)}_{\text{Variance term}_{\sim n^{-1/2}}} + \underbrace{(\mathbb{E}\hat{\eta}_n - \eta_K)}_{\text{Bias term}_{\uparrow k}}.$$

- When $Y \perp\!\!\!\perp X$, bias is always 0 and variance improves with k — useful in independence testing.

Limiting null (general graph)

Theorem (informal)

Suppose $\mu = \mu_X \otimes \mu_Y$, then there exists sequences of random variables $V_n = \mathcal{O}_{\mathbb{P}}(1)$ and V_n^{lin} such that

$$\frac{\sqrt{n}\hat{\eta}_n^{\text{lin}}}{V_n^{\mathrm{lin}}} \xrightarrow{d} \mathcal{N}(0,1), \qquad \frac{\sqrt{n}\hat{\eta}_n}{V_n} \xrightarrow{d} \mathcal{N}(0,1).$$

Limiting null (general graph)

Theorem (informal)

Suppose $\mu = \mu_X \otimes \mu_Y$, then there exists sequences of random variables $V_n = \mathcal{O}_{\mathbb{P}}(1)$ and V_n^{lin} such that

$$\frac{\sqrt{n}\hat{\eta}_n^{\mathsf{lin}}}{V_n^{\mathrm{lin}}} \xrightarrow{d} \mathcal{N}(0,1), \qquad \frac{\sqrt{n}\hat{\eta}_n}{V_n} \xrightarrow{d} \mathcal{N}(0,1).$$

- (Proof) Uses U-statistics projection theory and Stein's method on dependency graphs.
- (General) a uniform CLT holds for a suitable class of graphs $\mathcal{G}_n$, i.e.,

$$\sup_{G_n \in \mathcal{G}_n} \sup_{x \in \mathbb{R}} \left|\mathbb{P}(\sqrt{n}\hat{\eta}_n^{\mathsf{lin}}/V_n \leq x) - \Phi(x)\right| \overset{n\to\infty}{\longrightarrow} 0.$$

- Theorem holds for data driven choices $\hat{G}_n$ provided $\mathbb{P}(\hat{G}_n \in \mathcal{G}_n) \overset{n\to\infty}{\longrightarrow} 1$.

Independence testing

- Consider the testing problem:

$$\mathrm{H}_0 : \mu = \mu_X \otimes \mu_Y \quad \text{vs} \quad \mathrm{H}_1 : \mu \neq \mu_X \otimes \mu_Y .$$

- Recall $\eta_K = 0$ iff $\mu = \mu_X \otimes \mu_Y$, $\eta_K > 0$ otherwise, $\hat{\eta}_n \xrightarrow{\mathbb{P}} \eta_K$.

- A natural test:

$$\text{Reject if } \sqrt{n}\hat{\eta}_n^{\mathrm{lin}} / V_n \geq z_\alpha .$$

- Consistent and maintains level, i.e.,

$$\lim_{n\to\infty} \mathbb{P}_{\mathrm{H}_0}(\text{Reject } \mathrm{H}_0) = \alpha, \quad \lim_{n\to\infty} \mathbb{P}_{\mathrm{H}_1}(\text{Reject } \mathrm{H}_0) = 1.$$

Independence testing

- Consider the testing problem:

$$\mathrm{H}_0 : \mu = \mu_X \otimes \mu_Y \quad \text{vs} \quad \mathrm{H}_1 : \mu \neq \mu_X \otimes \mu_Y .$$

- Recall $\eta_K = 0$ iff $\mu = \mu_X \otimes \mu_Y$, $\eta_K > 0$ otherwise, $\hat{\eta}_n \xrightarrow{\mathbb{P}} \eta_K$.

- A natural test:

$$\text{Reject if } \sqrt{n}\hat{\eta}_n^{\text{lin}} / V_n \geq z_\alpha .$$

- Consistent and maintains level, i.e.,

$$\lim_{n\to\infty} \mathbb{P}_{\mathrm{H}_0}(\text{Reject } \mathrm{H}_0) = \alpha, \quad \lim_{n\to\infty} \mathbb{P}_{\mathrm{H}_1}(\text{Reject } \mathrm{H}_0) = 1.$$

- Near linear complexity.

Summary

- Class of kernel measures of association (KMAc) when $\mathcal{Y}$ admits a non-negative definite kernel.
- Class of graph-based, consistent estimators ($\mathcal{X}$ - metric space) for KMAc without smoothness on the conditional distribution.
- When k-NNG is used, the rate of convergence adapts to the intrinsic dimension of the support μ_X.
- Established a pivotal Gaussian limit uniformly over a class of graphs.
- A linear time estimator $+$ a near linear time test of statistical independence.

Summary

- Class of kernel measures of association (KMAc) when $\mathcal{Y}$ admits a non-negative definite kernel.
- Class of graph-based, consistent estimators ($\mathcal{X}$ - metric space) for KMAc without smoothness on the conditional distribution.
- When k-NNG is used, the rate of convergence adapts to the intrinsic dimension of the support μ_X.
- Established a pivotal Gaussian limit uniformly over a class of graphs.
- A linear time estimator $+$ a near linear time test of statistical independence.
- A wide array of numerical experiments with real and simulated datasets - see https://arxiv.org/pdf/2012.14804.pdf.


The End

Simulations (choice of k)

$(X^{(1)}, X^{(2)}, Y^{(1)}, Y^{(2)}) \sim \mu$ supported on $\mathbb{R}^4$ where $(X^{(1)}, Y^{(1)}), (X^{(2)}, Y^{(2)})$ are i.i.d., where

- (W-shaped)

  $$Y^{(1)} = |X^{(1)} + 0.5| 1(X^{(1)} \leq 0) + |X^{(1)} - 0.5| 1(X^{(1)} > 0) + 0.75\lambda\epsilon,$$

  $\epsilon \sim \mathcal{N}(0, 1)$ with varying λ.

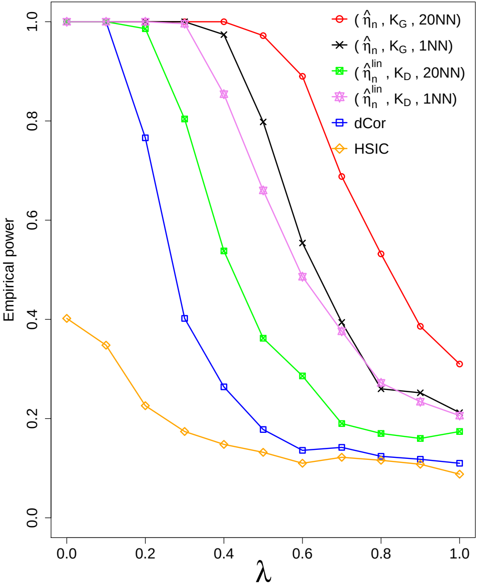
- (Sinusoidal)

  $$Y^{(1)} = \cos(8\pi X^{(1)}) + 3\lambda\epsilon,$$

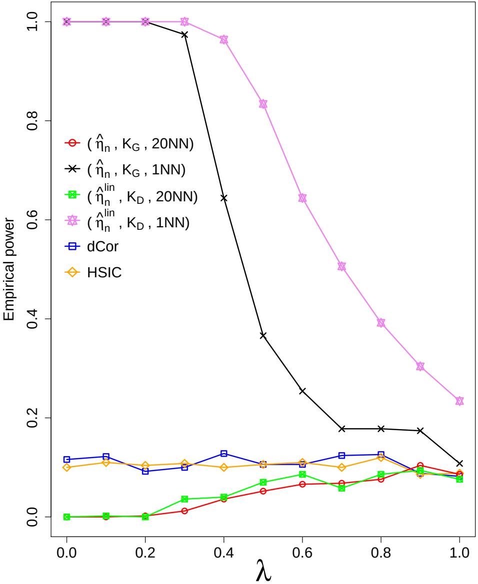
  $\epsilon \sim \mathcal{N}(0, 1)$ with varying λ.

Sample size $n = 300$.

W-shaped (K_G-Gaussian kernel, K_D-Distance kernel)



Sinusoidal (K_G-Gaussian kernel, K_D-Distance kernel)



Conditional association

- Recall

$$\eta_K = \frac{\underbrace{\mathbb{E}K(Y', \tilde{Y}')}_{*\mu_{Y|X}} - \underbrace{\mathbb{E}K(Y_1, Y_2)}_{*\mu_Y}}{\mathbb{E}K(Y_1, Y_1) - \mathbb{E}K(Y_1, Y_2)}$$

  where $X' \sim \mu_X,\ Y', \tilde{Y}'$ are drawn independently from $\mu_{Y|X'}$.
- The surrogate in the numerator show we are comparing $\mu_{Y|X}$ with μ_Y.

Conditional association

- Recall

$$\eta_K = \frac{\underbrace{\mathbb{E}K(Y', \tilde{Y}')}_{*\mu_{Y|X}} - \underbrace{\mathbb{E}K(Y_1, Y_2)}_{*\mu_Y}}{\mathbb{E}K(Y_1, Y_1) - \mathbb{E}K(Y_1, Y_2)}$$

  where $X' \sim \mu_X$, $Y', \tilde{Y}'$ are drawn independently from $\mu_{Y|X'}$.
- The surrogate in the numerator show we are comparing $\mu_{Y|X}$ with μ_Y.
- For conditional association, i.e., how closely is Y associated with Z given X, define:

$$\tilde{\eta}_K := \frac{\underbrace{\mathbb{E}K(Y_2', \tilde{Y}_2')}_{*\mu_{Y|X,Z}} - \underbrace{\mathbb{E}K(Y', \tilde{Y}')}_{*\mu_{Y|X}}}{\mathbb{E}K(Y_1, Y_1) - \mathbb{E}K(Y', \tilde{Y}')}$$

  where $(X', Z') \sim \mu_{XZ}$ and $Y_2', \tilde{Y}_2'$ are drawn independently from $\mu_{Y|(X',Z')}$.

Estimating Conditional association

- Recall

$$T_{1,n} := \frac{1}{n}\sum_{i=1}^{n}\frac{1}{d_i}\sum_{j:(i,j)\in E(G_n)} K(Y_i, Y_j) \approx \mathbb{E}K(Y', \tilde{Y}')$$

where $E(G_n)$ — edge/neighbor set of G_n, the nearest neighbor graph on $(X_1, \ldots, X_n)$ and d_i — degree of X_i.

- Use the estimator

$$\hat{\tilde{\eta}}_K := \frac{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{\tilde{d}_i}\sum_{j:(i,j)\in E(\tilde{G}_n)} K(Y_i, Y_j) - T_{1,n}}{\frac{1}{n}\sum_{i=1}^{n} K(Y_i, Y_i) - T_{1,n}},$$

$\tilde{G}_n$ — edge/neighbor set of G_n, the nearest neighbor graph on $(X_1, Z_1), \ldots, (X_n, Z_n)$ and $\tilde{d}_i$ — degree of (X_i, Z_i).

Estimating Conditional association

- Recall

$$T_{1,n} := \frac{1}{n}\sum_{i=1}^{n}\frac{1}{d_i}\sum_{j:(i,j)\in E(G_n)} K(Y_i, Y_j) \approx \mathbb{E}K(Y', \tilde{Y}')$$

where $E(G_n)$ — edge/neighbor set of G_n, the nearest neighbor graph on $(X_1, \ldots, X_n)$ and d_i — degree of X_i.

- Use the estimator

$$\hat{\tilde{\eta}}_K := \frac{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{\tilde{d}_i}\sum_{j:(i,j)\in E(\tilde{G}_n)} K(Y_i, Y_j) - T_{1,n}}{\frac{1}{n}\sum_{i=1}^{n} K(Y_i, Y_i) - T_{1,n}},$$

$\tilde{G}_n$ — edge/neighbor set of G_n, the nearest neighbor graph on $(X_1, Z_1), \ldots, (X_n, Z_n)$ and $\tilde{d}_i$ — degree of (X_i, Z_i).

- Then

$$\hat{\tilde{\eta}}_K \xrightarrow{P} \tilde{\eta}_K.$$

Also $\tilde{\eta}_K \in [0, 1]$ and $\tilde{\eta}_K = 0$ iff $Y \perp\!\!\!\perp Z | X$ and $\tilde{\eta}_K = 1$ if Y is a measurable function of X, Z.

Local power in independence testing

- Consider the family of alternatives (Farlie):

$$f_{X,Y}(x,y) = (1 - r_n) f_1(x) f_2(y) + r_n g(x,y).$$

- What happens to test based on $\hat{\eta}_n^{\text{lin}}$ as $r_n \to 0$?

Local power in independence testing

- Consider the family of alternatives (Farlie):

$$f_{X,Y}(x,y) = (1 - r_n) f_1(x) f_2(y) + r_n g(x,y).$$

- What happens to test based on $\hat{\eta}_n^{\text{lin}}$ as $r_n \to 0$?
- For $d_1 \leq 7$, power converges to 1 if $r_n \gg n^{-1/4}$ and to 0 if $r_n \ll n^{-1/4}$.
- (Blessing of dimensionality?): For $d_1 \geq 9$, power converges to 1 if $r_n \gg n^{-\left(\frac{1}{2} - \frac{2}{d_1}\right)}$ and power converges to 0 if $r_n \ll n^{-\left(\frac{1}{2} - \frac{2}{d_1}\right)}$.
- For $d = 8$, the power depends on a rather complicated tradeoff.

Illustration of monotonicity

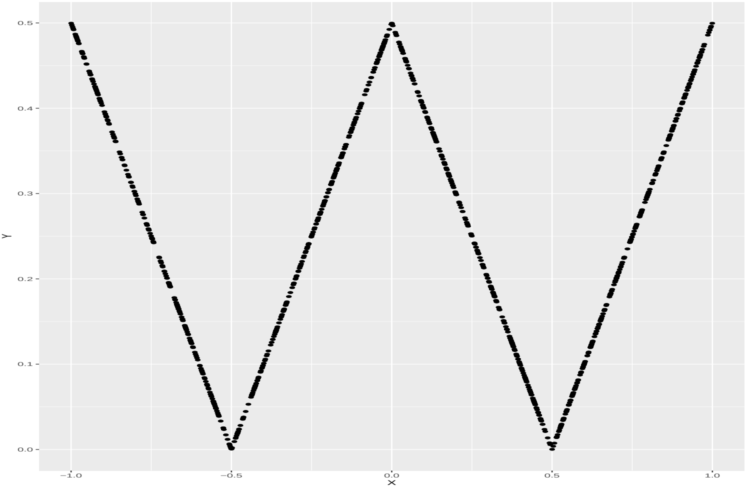
$(X^{(1)}, X^{(2)}, Y^{(1)}, Y^{(2)}) \sim \mu$ supported on $\mathbb{R}^4$ where
$(X^{(1)}, Y^{(1)}), (X^{(2)}, Y^{(2)})$ are i.i.d., where

- (W-shaped)

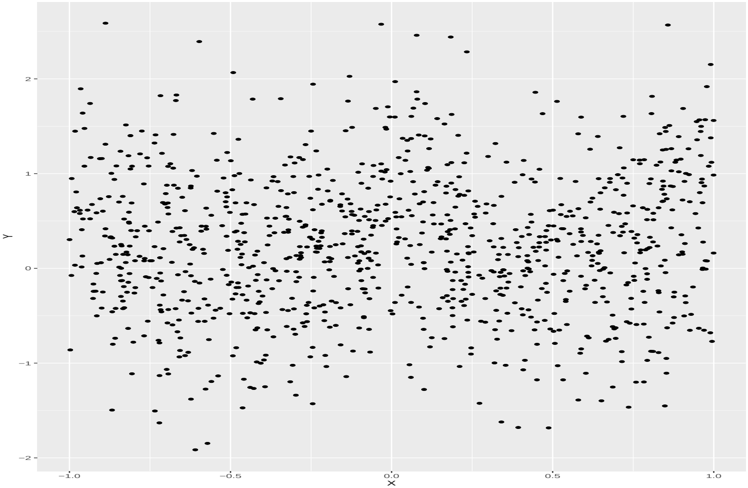
  $$Y^{(1)} = |X^{(1)} + 0.5| 1(X^{(1)} \leq 0) + |X^{(1)} - 0.5| 1(X^{(1)} > 0) + 0.75\lambda\epsilon,$$

  $\epsilon \sim \mathcal{N}(0, 1)$ with varying λ.

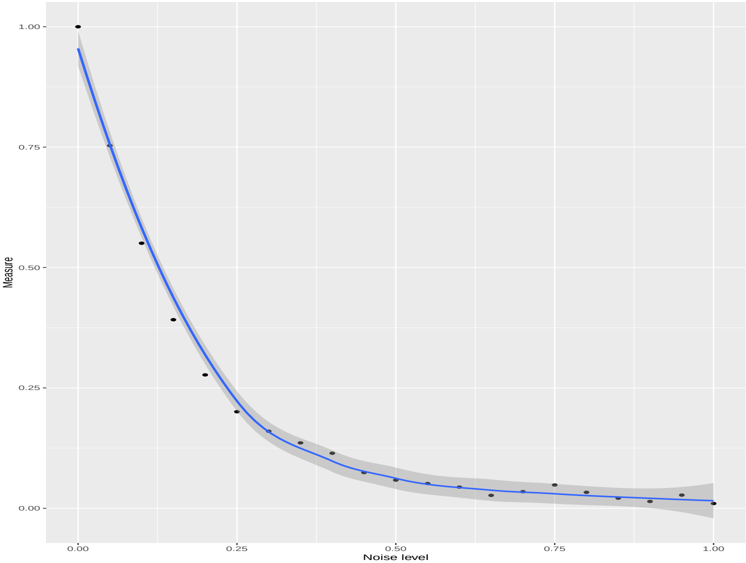
W-shaped (noiseless)



W-shaped (noisy)

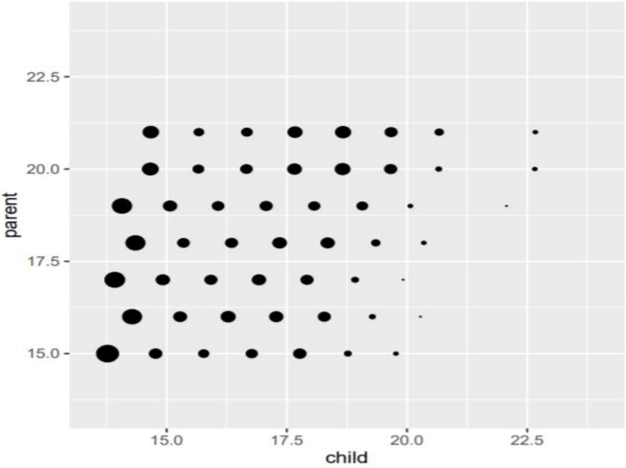


W-shaped (monotonicity)



Galton Peas dataset

- Mean diameters of sweet peas in mother plants and daughter plants (700×2)



Galton Peas (continued)

- 7 unique values for the mother (X) and 52 for the daughter (Y).
- X and Y seem to be associated.
- Pearson's correlation $= 0.35$, p-value $\ll 0.05$.

Galton Peas (continued)

- 7 unique values for the mother (X) and 52 for the daughter (Y).
- X and Y seem to be associated.
- Pearson's correlation $= 0.35$, p-value $\ll 0.05$.
- Can we say something more?

A curious observation (Chatterjee, 2020)

	Parent						
Child	15	16	17	18	19	20	21
13.77	46	0	0	0	0	0	0
13.92	0	0	37	0	0	0	0
14.07	0	0	0	0	35	0	0
14.28	0	34	0	0	0	0	0
14.35	0	0	0	34	0	0	0
14.66	0	0	0	0	0	23	0
14.67	0	0	0	0	0	0	22
14.77	14	0	0	0	0	0	0
14.92	0	0	16	0	0	0	0
15.07	0	0	0	0	16	0	0
15.28	0	15	0	0	0	0	0
15.35	0	0	0	12	0	0	0
15.66	0	0	0	0	0	10	0
15.67	0	0	0	0	0	0	8
15.77	9	0	0	0	0	0	0
15.92	0	0	13	0	0	0	0
16.07	0	0	0	0	12	0	0
16.28	0	18	0	0	0	0	0
16.35	0	0	0	13	0	0	0
16.66	0	0	0	0	0	12	0
16.67	0	0	0	0	0	0	10
16.77	11	0	0	0	0	0	0
16.92	0	0	16	0	0	0	0
17.07	0	0	0	0	13	0	0
17.28	0	16	0	0	0	0	0
17.35	0	0	0	17	0	0	0

A curious observation (Chatterjee, 2020)

	Parent						
Child	15	16	17	18	19	20	21
13.77	46	0	0	0	0	0	0
13.92	0	0	37	0	0	0	0
14.07	0	0	0	0	35	0	0
14.28	0	34	0	0	0	0	0
14.35	0	0	0	34	0	0	0
14.66	0	0	0	0	0	23	0
14.67	0	0	0	0	0	0	22
14.77	14	0	0	0	0	0	0
14.92	0	0	16	0	0	0	0
15.07	0	0	0	0	16	0	0
15.28	0	15	0	0	0	0	0
15.35	0	0	0	12	0	0	0
15.66	0	0	0	0	0	10	0
15.67	0	0	0	0	0	0	8
15.77	9	0	0	0	0	0	0
15.92	0	0	13	0	0	0	0
16.07	0	0	0	0	12	0	0
16.28	0	18	0	0	0	0	0
16.35	0	0	0	13	0	0	0
16.66	0	0	0	0	0	12	0
16.67	0	0	0	0	0	0	10
16.77	11	0	0	0	0	0	0
16.92	0	0	16	0	0	0	0
17.07	0	0	0	0	13	0	0
17.28	0	16	0	0	0	0	0
17.35	0	0	0	17	0	0	0

- Every row has exactly one non-zero element.

Galton Peas (continued)

- Recall X-mother, Y-daughter.
- It is more convenient to predict X from Y (Parent from daughter) than the other way round.
- Pearson's correlation being symmetric cannot distinguish between the two problems — same is the case for most measures of dependence.

Galton Peas (continued)

- Recall X-mother, Y-daughter.
- It is more convenient to predict X from Y (Parent from daughter) than the other way round.
- Pearson's correlation being symmetric cannot distinguish between the two problems — same is the case for most measures of dependence.
- How to design a measure that captures this asymmetry?