Measuring Association on Topological Spaces Using Kernels and Geometric Graphs

Bodhisattva Sen[1]

Department of Statistics
Columbia University

Applied Statistics Unit
Indian Statistical Institute, Kolkata

Joint work with **Nabarun Deb** (Columbia) & **Promit Ghosal** (MIT)

December 15, 2020

Preprint available at https://arxiv.org/pdf/2010.01768.pdf

[1]Supported by NSF grant DMS-2015376

- $(X, Y) \sim \mu$ on $\mathcal{X} \times \mathcal{Y}$ (topological space) with marginals μ_X & μ_Y
- **Informal goal**: Construct a coefficient that can measure the strength of association/dependence between X & Y beyond simply testing for independence

- $(X, Y) \sim \mu$ on $\mathcal{X} \times \mathcal{Y}$ (topological space) with marginals μ_X & μ_Y
- **Informal goal**: Construct a coefficient that can measure the strength of association/dependence between X & Y beyond simply testing for independence

Motivation: Pearson's correlation coefficient

- Given $(X, Y) \sim$ bivariate normal, Pearson's correlation ρ measures the strength of association between X and Y
- $\rho = 0$ iff X and Y are independent
- $\rho = \pm 1$ iff one variable is a (linear) function of the other

- $(X, Y) \sim \mu$ on $\mathcal{X} \times \mathcal{Y}$ (topological space) with marginals μ_X & μ_Y

- **Informal goal**: Construct a coefficient that can measure the strength of association/dependence between X & Y beyond simply testing for independence

Motivation: Pearson's correlation coefficient

- Given $(X, Y) \sim$ bivariate normal, Pearson's correlation ρ measures the strength of association between X and Y

- $\rho = 0$ iff X and Y are independent
- $\rho = \pm 1$ iff one variable is a (linear) function of the other

- Any value of ρ in [-1,1] conveys an idea of the strength of the relationship between X and Y

- $(X, Y) \sim \mu$ on $\mathcal{X} \times \mathcal{Y}$ (topological space) with marginals μ_X & μ_Y

- **Informal goal**: Construct a coefficient that can measure the strength of association/dependence between X & Y beyond simply testing for independence

Motivation: Pearson's correlation coefficient

- Given $(X, Y) \sim$ bivariate normal, Pearson's correlation ρ measures the strength of association between X and Y

- $\rho = 0$ iff X and Y are independent
- $\rho = \pm 1$ iff one variable is a (linear) function of the other

- Any value of ρ in [-1,1] conveys an idea of the strength of the relationship between X and Y

Question: What are nonparametric analogs of Pearson's correlation?

- **Want**: A measure of association that:
  (a) equals 0 iff $X \perp\!\!\!\perp Y$,
  (b) equals 1 iff Y is a (measurable) function of X, and
  (c) any value in [0,1] conveys an idea of the strength of the relationship between X and Y

- **Want**: A measure of association that:
  (a) equals 0 iff $X \perp\!\!\!\perp Y$,
  (b) equals 1 iff Y is a (measurable) function of X, and
  (c) any value in [0,1] conveys an idea of the strength of the relationship between X and Y

- **Testing for independence**: For the past century, most measures of association/dependence only focus on testing $X \perp\!\!\!\perp Y$, 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.

- **Want**: A measure of association that:
  (a) equals 0 iff $X \perp\!\!\!\perp Y$,
  (b) equals 1 iff Y is a (measurable) function of X, and
  (c) any value in [0,1] conveys an idea of the strength of the relationship between X and Y

- **Testing for independence**: For the past century, most measures of association/dependence only focus on testing $X \perp\!\!\!\perp Y$, 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.

- 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.

- **Want**: A measure of association that:
  (a) equals 0 iff $X \perp\!\!\!\perp Y$,
  (b) equals 1 iff Y is a (measurable) function of X, and
  (c) any value in [0,1] conveys an idea of the strength of the relationship between X and Y

- **Testing for independence**: For the past century, most measures of association/dependence only focus on testing $X \perp\!\!\!\perp Y$, 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.

- 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 $\mathcal{Y} = \mathbb{R}$

We consider the case when $\mathcal{X}$ and $\mathcal{Y}$ are general topological spaces (e.g., metric spaces)

Outline

1. Family of Measures of Association
   - A measure of dependence on Euclidean spaces
   - Extending to a class of kernel measures
2. Estimating the Kernel Measure of Association (KMAc)
   - The estimator
   - Consistency and rate of convergence
   - Central limit theorem
   - Computational complexity
3. Other Applications of Kernels and Geometric Graphs
   - A measure of conditional dependence
   - A measure of dissimilarity between M-distributions

Outline

1. Family of Measures of Association
   - A measure of dependence on Euclidean spaces
   - Extending to a class of kernel measures
2. Estimating the Kernel Measure of Association (KMAc)
   - The estimator
   - Consistency and rate of convergence
   - Central limit theorem
   - Computational complexity
3. Other Applications of Kernels and Geometric Graphs
   - A measure of conditional dependence
   - A measure of dissimilarity between M-distributions

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

$(X, Y) \sim \mu$ on $\mathcal{X} \times \mathcal{Y}$ with marginals μ_X & μ_Y

Basic strategy

- Most measures quantify a "discrepancy" between μ and $\mu_X \otimes \mu_Y$
- We construct a discrepancy between $\mu_{Y|X}$ (regular conditional distribution of Y given X) and μ_Y

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

$(X, Y) \sim \mu$ on $\mathcal{X} \times \mathcal{Y}$ with marginals μ_X & μ_Y

Basic strategy

- Most measures quantify a "discrepancy" between μ and $\mu_X \otimes \mu_Y$
- We construct a discrepancy between $\mu_{Y|X}$ (regular conditional distribution of Y given X) 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: $X' \sim \mu_X$ and then $Y', \tilde{Y}'$ i.i.d. $\mu_{Y|X'}$ (i.e., Y' and $\tilde{Y}'$ are conditionally independent given X')

- Recall $X' \sim \mu_X$, and $Y', \tilde{Y}'|X' \overset{iid}{\sim} \mu_{Y|X'}$, and

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

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

- Suppose $Y \perp\!\!\!\perp X$, then $\mu_{Y|X'} = \mu_Y$,

$$\text{and thus} \quad Y', \tilde{Y}' \overset{i.i.d.}{\sim} \mu_Y \quad \Rightarrow \quad T = 0$$

- Recall $X' \sim \mu_X$, and $Y', \tilde{Y}' | X' \overset{iid}{\sim} \mu_{Y|X'}$, and

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

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

- Suppose $Y \perp\!\!\!\perp X$, then $\mu_{Y|X'} = \mu_Y$,

$$\text{and thus} \quad Y', \tilde{Y}' \overset{i.i.d.}{\sim} \mu_Y \quad \Rightarrow \quad T = 0$$

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

$$Y' = \tilde{Y}' = h(X') \quad \Leftrightarrow \quad \| Y' - \tilde{Y}' \|_2 = 0 \ \text{ a.s.} \quad \Leftrightarrow \quad T = 1$$

- Recall $X' \sim \mu_X$, and $Y', \tilde{Y}'|X' \overset{iid}{\sim} \mu_{Y|X'}$, and

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

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

- Suppose $Y \perp\!\!\!\perp X$, then $\mu_{Y|X'} = \mu_Y$,

$$\text{and thus} \quad Y', \tilde{Y}' \overset{i.i.d.}{\sim} \mu_Y \quad \Rightarrow \quad T = 0$$

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

$$Y' = \tilde{Y}' = h(X') \quad \Leftrightarrow \quad \| Y' - \tilde{Y}' \|_2 = 0 \ \text{a.s.} \quad \Leftrightarrow \quad T = 1$$

- Showing that $T = 0 \ \Rightarrow \ Y \perp\!\!\!\perp X$ is more complicated!

Theorem [Deb, Ghosal and S. (2020+)]

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.

Theorem [Deb, Ghosal and S. (2020+)]

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.

Interpretability and Monotonicity: What happens when $T \in (0, 1)$?

- 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 $|\rho|$.

Theorem [Deb, Ghosal and S. (2020+)]

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.

Interpretability and Monotonicity: What happens when $T \in (0, 1)$?

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

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

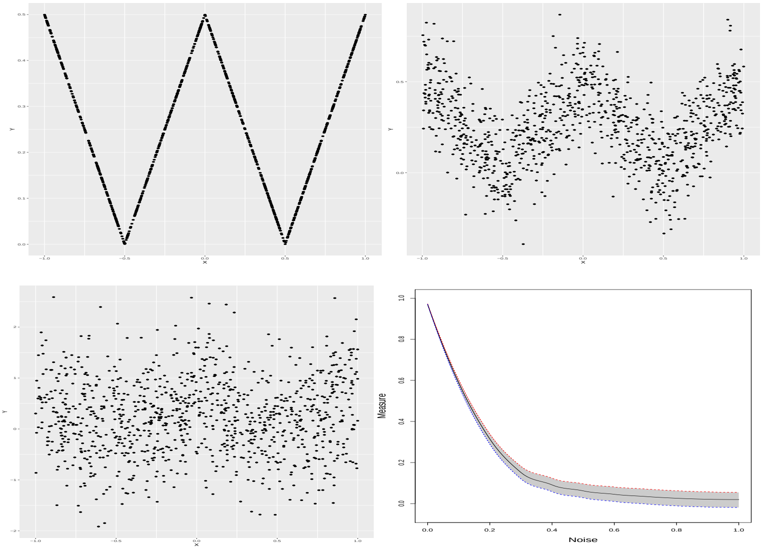
  The above function is strictly convex and increasing in $|\rho|$.

- In many nonparametric regression models, T turns out to be a monotonic function of the degree of dependence between Y and X

T captures the strength of the relationship between Y and X

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

W-shape: $Y^{(1)} = |X^{(1)} + 0.5|\mathbf{1}_{X^{(1)} \leq 0} + |X^{(1)} - 0.5|\mathbf{1}_{X^{(1)} > 0} + 0.75\lambda\epsilon$,
where $\epsilon \sim \mathcal{N}(0,1)$ with varying λ; $\qquad X \sim \text{Uniform}[-1,1]$



Reproducing kernel Hilbert spaces (RKHS)

- $\mathcal{H}$: Hilbert space[2] of functions from $\mathcal{Y}$ to $\mathbb{R}$
- **Kernel function**: A symmetric nonnegative definite function on $\mathcal{Y}$, i.e., $K : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ 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 $m \geq 1$

[2] A Hilbert space is a complete inner product space

Reproducing kernel Hilbert spaces (RKHS)

- $\mathcal{H}$: Hilbert space[2] of functions from $\mathcal{Y}$ to $\mathbb{R}$
- **Kernel function**: A symmetric nonnegative definite function on $\mathcal{Y}$, i.e., $K : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ 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 $m \geq 1$

- For all $y \in \mathcal{Y}$, $\quad K(y, \cdot) \in \mathcal{H}$, $\quad$ (note $K(y, \cdot) : \mathcal{Y} \to \mathbb{R}, \ \forall\, y \in \mathcal{Y}$)
- Identify $y \mapsto K(y, \cdot)$ (feature map)
- **Gaussian kernel**: $k(u, v) := \exp(-\|u - v\|_2^2)$

[2] A Hilbert space is a complete inner product space

Moore-Aronszajn Theorem

Suppose $K(\cdot,\cdot) : \mathcal{Y} \to \mathbb{R}$ is a nonnegative definite kernel. Then there exists a Hilbert space $(\mathcal{H}, \langle \cdot, \cdot \rangle_{\mathcal{H}})$ comprising $\{f : \mathcal{Y} \to \mathbb{R}\}$ such that:

- $K(y, \cdot) \in \mathcal{H}, \quad \forall\, y \in \mathcal{Y};$
- (Reproducing property) For all $f \in \mathcal{H}$, $y \in \mathcal{Y}$,

$$\langle f, K(y, \cdot) \rangle_{\mathcal{H}} = f(y).$$

Moore-Aronszajn Theorem

Suppose $K(\cdot,\cdot) : \mathcal{Y} \to \mathbb{R}$ is a nonnegative definite kernel. Then there exists a Hilbert space $(\mathcal{H}, \langle\cdot,\cdot\rangle_{\mathcal{H}})$ comprising $\{f : \mathcal{Y} \to \mathbb{R}\}$ such that:

- $K(y,\cdot) \in \mathcal{H}, \quad \forall\, y \in \mathcal{Y}$;
- (Reproducing property) For all $f \in \mathcal{H}$, $y \in \mathcal{Y}$,

$$\langle f, K(y,\cdot)\rangle_{\mathcal{H}} = f(y).$$

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

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

- **Idea**: 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

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

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

- Define our kernel measure of association (KMAc) as

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

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

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

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

- Define our kernel measure of association (KMAc) as

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

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

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

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

- Define our kernel measure of association (KMAc) as

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

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

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

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

- Define our kernel measure of association (KMAc) as

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

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

Theorem [Deb, Ghosal and S. (2020+)]

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.

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

Theorem [Deb, Ghosal and S. (2020+)]

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 of Characteristic Kernels

Some examples of characteristic kernels [Gretton et al., 2012, Sejdinovic et al., 2013, Lyons 2013, 2014] 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 of Characteristic Kernels

Some examples of characteristic kernels [Gretton et al., 2012, Sejdinovic et al., 2013, Lyons 2013, 2014] 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. Family of Measures of Association
   - A measure of dependence on Euclidean spaces
   - Extending to a class of kernel measures
2. Estimating the Kernel Measure of Association (KMAc)
   - The estimator
   - Consistency and rate of convergence
   - Central limit theorem
   - Computational complexity
3. Other Applications of Kernels and Geometric Graphs
   - A measure of conditional dependence
   - A measure of dissimilarity between M-distributions

Estimation Strategy

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

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

- By standard U-statistics theory,

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

and

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

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

- Suppose $\mathcal{X}$ is a finite set. Then, $\mathbb{E}K(Y', \tilde{Y}')$ can be handled as

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

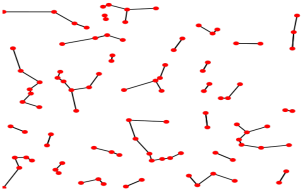
- Suppose $\mathcal{X}$ is a finite set. Then, $\mathbb{E}K(Y', \tilde{Y}')$ can be handled as

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

- If X is continuous, replace $X_j = X_i$ with $\rho_{\mathcal{X}}(X_i, X_j)$ being "small"

Geometric graph

- 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 on $\{X_1, \ldots, X_n\}$ to its first k nearest neighbors



Estimate $\mathbb{E}K(Y', \tilde{Y}')$ by replacing

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

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 $E(G_n)$ is edge set of G_n and d_i is the degree of X_i

Estimate $\mathbb{E}K(Y', \tilde{Y}')$ by replacing

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

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 $E(G_n)$ is edge set of G_n and d_i is the degree of X_i

Geometric graph-based estimator

Now

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

can be estimated by

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

Theorem (informal) [Deb, Ghosal and S. (2020+)]

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

$$\hat{\eta}_n \xrightarrow{\mathbb{P}} \eta_K.$$

- In particular, $\frac{1}{n}\sum_{i=1}^{n} \frac{1}{d_i} \sum_{j:(i,j)\in E(G_n)} K(Y_i, Y_j) \xrightarrow{\mathbb{P}} \mathbb{E}K(Y', \tilde{Y}')$

Theorem (informal) [Deb, Ghosal and S. (2020+)]

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

$$\hat{\eta}_n \xrightarrow{\mathbb{P}} \eta_K.$$

- In particular, $\frac{1}{n}\sum_{i=1}^{n} \frac{1}{d_i} \sum_{j:(i,j)\in E(G_n)} K(Y_i, Y_j) \xrightarrow{\mathbb{P}} \mathbb{E}K(Y', \tilde{Y}')$
- No smoothness assumptions needed on conditional distribution of $Y|X$ — motivated directly from the approach used in Chatterjee, 2019, Azadkia and Chatterjee, 2019.

Theorem (informal) [Deb, Ghosal and S. (2020+)]

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

$$\hat{\eta}_n \xrightarrow{\mathbb{P}} \eta_K.$$

- In particular, $\frac{1}{n}\sum_{i=1}^{n} \frac{1}{d_i} \sum_{j:(i,j)\in E(G_n)} K(Y_i, Y_j) \xrightarrow{\mathbb{P}} \mathbb{E}K(Y', \tilde{Y}')$
- No smoothness assumptions needed on conditional distribution of $Y|X$ — motivated directly from the approach used in Chatterjee, 2019, Azadkia and Chatterjee, 2019.
- For k-NNGs, $\hat{\eta}_n$ is consistent provided $k = o(n/\log n)$
- Thus, for consistent estimation, a 1-NNG can be chosen

Rate of convergence (for k-NNG)

Theorem (informal) [Deb, Ghosal and S. (2020+)]

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

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

- Estimation rate automatically adapts to intrinsic dimension of μ_X

Rate of convergence (for k-NNG)

Theorem (informal) [Deb, Ghosal and S. (2020+)]

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

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

- Estimation rate automatically adapts to intrinsic dimension of μ_X
- Note: $$\hat{\eta}_n - \eta_K = \underbrace{(\hat{\eta}_n - \mathbb{E}\hat{\eta}_n)}_{\text{Variance term}_{\sim n^{-1/2}}} + \underbrace{(\mathbb{E}\hat{\eta}_n - \eta_K)}_{\text{Bias term}_{\uparrow k}}$$
- **When** $Y \perp\!\!\!\perp X$: Bias is always 0, and variance improves with k — useful in independence testing.

1. Family of Measures of Association
   - A measure of dependence on Euclidean spaces
   - Extending to a class of kernel measures

2. Estimating the Kernel Measure of Association (KMAc)
   - The estimator
   - Consistency and rate of convergence
   - Central limit theorem
   - Computational complexity

3. Other Applications of Kernels and Geometric Graphs
   - A measure of conditional dependence
   - A measure of dissimilarity between M-distributions

Limiting Distribution under Independence (general graph)

Theorem (informal) [Deb, Ghosal and S. (2020+)]

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

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

Limiting Distribution under Independence (general graph)

Theorem (informal) [Deb, Ghosal and S. (2020+)]

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

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

- **Result**: A uniform CLT holds for a suitable class of graphs $\mathcal{G}_n$, i.e.,

$$\sup_{G_n \in \mathcal{G}_n} \sup_{x \in \mathbb{R}} \left| \mathbb{P}\left(\frac{\sqrt{n}\,\hat{\eta}_n}{V_n} \leq x \right) - \Phi(x) \right| \xrightarrow{n \to \infty} 0$$

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

- V_n can be computed from the data

Test of Independence

- Consider the testing problem:

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

- **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 level-α test (for $\alpha \in (0,1)$):

$$\text{Reject } H_0 \qquad \text{if} \qquad \frac{\sqrt{n}\,\hat{\eta}_n}{V_n} \geq z_\alpha$$

- Consistent and maintains level, i.e.,

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

Test of Independence

- Consider the testing problem:

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

- **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 level-α test (for $\alpha \in (0,1)$):

$$\text{Reject } H_0 \qquad \text{if} \qquad \frac{\sqrt{n}\,\hat{\eta}_n}{V_n} \geq z_\alpha$$

- Consistent and maintains level, i.e.,

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

- What is the computational complexity?

1. Family of Measures of Association
   - A measure of dependence on Euclidean spaces
   - Extending to a class of kernel measures
2. Estimating the Kernel Measure of Association (KMAc)
   - The estimator
   - Consistency and rate of convergence
   - Central limit theorem
   - Computational complexity
3. Other Applications of Kernels and Geometric Graphs
   - A measure of conditional dependence
   - A measure of dissimilarity between M-distributions

Computational Complexity

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

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

Computational Complexity

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

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

- $\sum_{i,j} K(Y_i, Y_j) = \|\sum_{i=1}^{n} K(Y_i, \cdot)\|_{\mathcal{H}}^2$
- $(\star)$ can be computed in linear time if $\|\cdot\|_{\mathcal{H}}^2$ is exactly computable, e.g., $K(y_i, y_j) = \langle y_i, y_j\rangle$ Otherwise
- Yields a near linear time test for independence; cf. distance correlation (or HSIC) takes $\mathcal{O}(n^2 B)$ time if B permutations are used

Summary

- Class of kernel measures of association (KMAc) when $\mathcal{Y}$ admits a nonnegative 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 automatically adapts to the intrinsic dimension of the support of μ_X
- Established a pivotal Gaussian limit uniformly over a class of graphs
- A near linear time estimator + a near linear time test of statistical independence

Summary

- Class of kernel measures of association (KMAc) when $\mathcal{Y}$ admits a nonnegative 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 automatically adapts to the intrinsic dimension of the support of μ_X
- Established a pivotal Gaussian limit uniformly over a class of graphs
- A near linear time estimator + a near linear time test of statistical independence
- In the paper, when $\mathcal{X}$ and $\mathcal{Y}$ are Euclidean, we propose another class of measures and estimators that is distribution-free under H_0

Simulations (choice of k)

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

- **W-shape**:

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

  where $\epsilon \sim \mathcal{N}(0,1)$ with varying λ; $X \sim$ Uniform[-1,1]

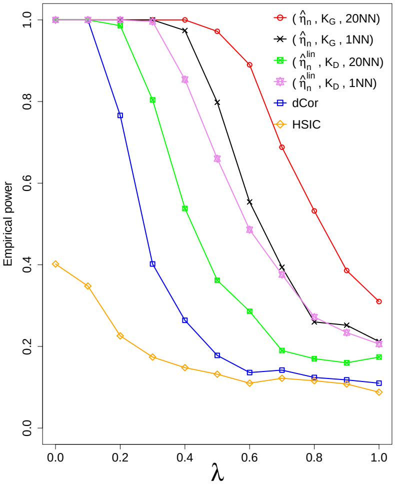
- **Sinusoidal**:

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

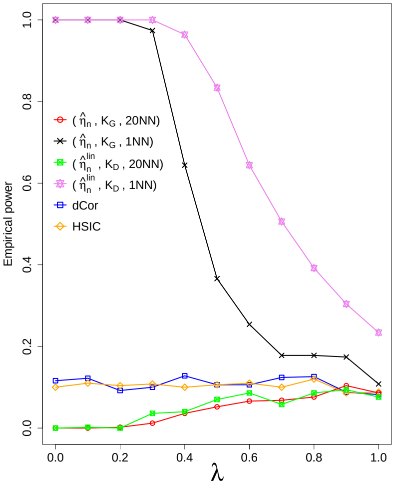
  $\epsilon \sim \mathcal{N}(0,1)$ with varying λ.

Sample size $n = 300$

(K_G-Gaussian kernel, K_D-Distance kernel)



W-shaped



Sinusoidal

Outline

1. Family of Measures of Association
   - A measure of dependence on Euclidean spaces
   - Extending to a class of kernel measures
2. Estimating the Kernel Measure of Association (KMAc)
   - The estimator
   - Consistency and rate of convergence
   - Central limit theorem
   - Computational complexity
3. Other Applications of Kernels and Geometric Graphs
   - A measure of conditional dependence
   - A measure of dissimilarity between M-distributions

Measure of Conditional Dependence[3]

- Suppose $(X, Y, Z) \sim \mu$ on some topological space $\mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$

- **Goal**: Measure the strength of conditional dependence between Y and X given Z

[3] Joint work with Zhen Huang and Nabarun Deb

Measure of Conditional Dependence[3]

- Suppose $(X, Y, Z) \sim \mu$ on some topological space $\mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$

- **Goal**: Measure the strength of conditional dependence between Y and X given Z

- **Question**: Can we define $\tau \equiv \tau(Y, X|Z)$ satisfying:
  (i) $\tau \in [0, 1]$;
  (ii) $\tau = 0$ iff $Y \perp\!\!\!\perp X|Z$;
  (iii) $\tau = 1$ iff Y is a measurable function of X and Z.

[3] Joint work with Zhen Huang and Nabarun Deb

Measure of Conditional Dependence[3]

- Suppose $(X, Y, Z) \sim \mu$ on some topological space $\mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$

- **Goal**: Measure the strength of conditional dependence between Y and X given Z

- **Question**: Can we define $\tau \equiv \tau(Y, X|Z)$ satisfying:
  (i) $\tau \in [0, 1]$;
  (ii) $\tau = 0$ iff $Y \perp\!\!\!\perp X|Z$;
  (iii) $\tau = 1$ iff Y is a measurable function of X and Z.

- **Applications**: Model-free variable selection, modeling causal relations in graphical models, ...

- We propose and study a class of nonparametric yet interpretable measures and their estimates, a sub-class of which can be computed in near linear time

[3] Joint work with Zhen Huang and Nabarun Deb

- Recall KMAc (for measuring dependence between Y and X):

$$\eta_K = \frac{\mathbb{E}K(Y', \tilde{Y}') - \mathbb{E}K(Y_1, Y_2)}{\mathbb{E}K(Y_1, Y_1) - \mathbb{E}K(Y_1, Y_2)} = \frac{\mathbb{E}[\mathrm{MMD}^2(\mu_{Y|X}, \mu_Y)]}{\mathbb{E}[\mathrm{MMD}^2(\delta_Y, \mu_Y)]}$$

  where $X' \sim \mu_X$, $Y', \tilde{Y}'$ are drawn independently from $\mu_{Y|X'}$

- MMD is the maximum mean discrepancy — a distance metric between two probability distributions depending on the kernel $K(\cdot, \cdot)$

- Recall KMAc (for measuring dependence between Y and X):

$$\eta_K = \frac{\mathbb{E}K(Y', \tilde{Y}') - \mathbb{E}K(Y_1, Y_2)}{\mathbb{E}K(Y_1, Y_1) - \mathbb{E}K(Y_1, Y_2)} = \frac{\mathbb{E}[\mathrm{MMD}^2(\mu_{Y|X}, \mu_Y)]}{\mathbb{E}[\mathrm{MMD}^2(\delta_Y, \mu_Y)]}$$

where $X' \sim \mu_X$, $Y', \tilde{Y}'$ are drawn independently from $\mu_{Y|X'}$

- MMD is the maximum mean discrepancy — a distance metric between two probability distributions depending on the kernel $K(\cdot, \cdot)$

Measuring conditional dependence between Y and X given Z

- Kernel partial correlation (KPC) coefficient:

$$\tau_K := \frac{\mathbb{E}[\mathrm{MMD}^2(\mu_{Y|XZ}, \mu_{Y|Z})]}{\mathbb{E}[\mathrm{MMD}^2(\delta_Y, \mu_{Y|Z})]}$$

- Recall KMAc (for measuring dependence between Y and X):

$$\eta_K = \frac{\mathbb{E}K(Y', \tilde{Y}') - \mathbb{E}K(Y_1, Y_2)}{\mathbb{E}K(Y_1, Y_1) - \mathbb{E}K(Y_1, Y_2)} = \frac{\mathbb{E}[\mathrm{MMD}^2(\mu_{Y|X}, \mu_Y)]}{\mathbb{E}[\mathrm{MMD}^2(\delta_Y, \mu_Y)]}$$

where $X' \sim \mu_X$, $Y', \tilde{Y}'$ are drawn independently from $\mu_{Y|X'}$

- MMD is the maximum mean discrepancy — a distance metric between two probability distributions depending on the kernel $K(\cdot, \cdot)$

Measuring conditional dependence between Y and X given Z

- Kernel partial correlation (KPC) coefficient:

$$\tau_K := \frac{\mathbb{E}[\mathrm{MMD}^2(\mu_{Y|XZ}, \mu_{Y|Z})]}{\mathbb{E}[\mathrm{MMD}^2(\delta_Y, \mu_{Y|Z})]} = \frac{\mathbb{E}K(Y_2', \tilde{Y}_2') - \mathbb{E}K(Y', \tilde{Y}')}{\mathbb{E}K(Y_1, Y_1) - \mathbb{E}K(Y', \tilde{Y}')}$$

where: (i) $(X', Z') \sim \mu_{XZ}$ and $Y_2', \tilde{Y}_2'$ are i.i.d. $\mu_{Y|(X', Z')}$,
(ii) $Z' \sim \mu_Z$ and $Y', \tilde{Y}'$ are i.i.d. $\mu_{Y|Z'}$

Can again employ a geometric graph-based estimation strategy

- For example, we can estimate

$$\tau_K = \frac{\mathbb{E}\left[\mathbb{E}[k(Y_2', \tilde{Y}_2')|X, Z]\right] - \mathbb{E}\left[\mathbb{E}[k(Y', \tilde{Y}')|Z]\right]}{\mathbb{E}[k(Y_1, Y_1)] - \mathbb{E}[\mathbb{E}[k(Y', \tilde{Y}')|Z]]}$$

by a 1-NNG by

$$\hat{\tau}_n := \frac{\frac{1}{n}\sum_{i=1}^n k(Y_i, Y_{\ddot{N}(i)}) - \frac{1}{n}\sum_{i=1}^n k(Y_i, Y_{N(i)})}{\frac{1}{n}\sum_{i=1}^n k(Y_i, Y_i) - \frac{1}{n}\sum_{i=1}^n k(Y_i, Y_{N(i)})}$$

where $(X_{\ddot{N}(i)}, Z_{\ddot{N}(i)})$ is NN of (X_i, Z_i) and $Y_{\ddot{N}(i)}$ is the corr. Y-value, and $Z_{N(i)}$ is NN of Z_i and $Y_{N(i)}$ is the corr. Y-value.

- For example, we can estimate

$$\tau_K = \frac{\mathbb{E}\left[\mathbb{E}[k(Y_2', \tilde{Y}_2')|X, Z]\right] - \mathbb{E}\left[\mathbb{E}[k(Y', \tilde{Y}')|Z]\right]}{\mathbb{E}[k(Y_1, Y_1)] - \mathbb{E}[\mathbb{E}[k(Y', \tilde{Y}')|Z]]}$$

by a 1-NNG by

$$\hat{\tau}_n := \frac{\frac{1}{n}\sum_{i=1}^n k(Y_i, Y_{\ddot{N}(i)}) - \frac{1}{n}\sum_{i=1}^n k(Y_i, Y_{N(i)})}{\frac{1}{n}\sum_{i=1}^n k(Y_i, Y_i) - \frac{1}{n}\sum_{i=1}^n k(Y_i, Y_{N(i)})}$$

where $(X_{\ddot{N}(i)}, Z_{\ddot{N}(i)})$ is NN of (X_i, Z_i) and $Y_{\ddot{N}(i)}$ is the corr. Y-value, and $Z_{N(i)}$ is NN of Z_i and $Y_{N(i)}$ is the corr. Y-value.

- **Consistency**: $\hat{\tau}_n \xrightarrow{\mathbb{P}} \tau_K$

- Automatic adaptation to the intrinsic dimensions of μ_X and μ_{XZ}

- Can develop a fully automatic stepwise variable selection algorithm which is provably consistent (cf. Azadkia and Chatterjee, 2019)

A measure of dissimilarity between M-distributions

- M-distributions $P_1, \ldots, P_M$ on $\mathcal{X}$ (e.g., metric space)
- **Data**: $\{X_{ij}\}_{j=1}^{n_i} \overset{iid}{\sim} P_i \quad$ for $i = 1, \ldots, M$
- **Question**: How different are the M distributions?

A measure of dissimilarity between M-distributions

- M-distributions $P_1, \ldots, P_M$ on $\mathcal{X}$ (e.g., metric space)
- **Data**: $\{X_{ij}\}_{j=1}^{n_i} \overset{iid}{\sim} P_i \quad$ for $i = 1, \ldots, M$
- **Question**: How different are the M distributions?
- We want to find a class of measures $\gamma \equiv \gamma(P_1, \ldots, P_M)$ such that:
  (i) $\gamma \in [0, 1]$;
  (ii) $\gamma = 0 \quad$ iff $\quad P_1 = \ldots = P_M$ (i.e., all the distributions are same);
  (iii) $\gamma = 1 \quad$ iff $\quad P_1, \ldots, P_M$ have disjoint supports

A measure of dissimilarity between M-distributions

- M-distributions $P_1, \ldots, P_M$ on $\mathcal{X}$ (e.g., metric space)
- **Data**: $\{X_{ij}\}_{j=1}^{n_i} \overset{iid}{\sim} P_i \quad$ for $i = 1, \ldots, M$
- **Question**: How different are the M distributions?
- We want to find a class of measures $\gamma \equiv \gamma(P_1, \ldots, P_M)$ such that:
  - (i) $\gamma \in [0, 1]$;
  - (ii) $\gamma = 0 \quad$ iff $\quad P_1 = \ldots = P_M$ (i.e., all the distributions are same);
  - (iii) $\gamma = 1 \quad$ iff $\quad P_1, \ldots, P_M$ have disjoint supports
- **Define**: $Y_{ij} = i, \quad$ for $i = 1, \ldots, M$
- Consider $\{\{(X_{ij}, Y_{ij})\}_{j=1}^{n_i}\}_{i=1}^{M}; \quad X \sim \sum_{i=1}^{M} \pi_i P_Y; \quad \pi_i \approx \frac{n_i}{\sum_{\ell=1}^{M} n_\ell}$

A measure of dissimilarity between M-distributions

- M-distributions $P_1, \ldots, P_M$ on $\mathcal{X}$ (e.g., metric space)
- **Data**: $\{X_{ij}\}_{j=1}^{n_i} \overset{iid}{\sim} P_i \quad$ for $i = 1, \ldots, M$
- **Question**: How different are the M distributions?
- We want to find a class of measures $\gamma \equiv \gamma(P_1, \ldots, P_M)$ such that:
  (i) $\gamma \in [0, 1]$;
  (ii) $\gamma = 0 \quad$ iff $\quad P_1 = \ldots = P_M$ (i.e., all the distributions are same);
  (iii) $\gamma = 1 \quad$ iff $\quad P_1, \ldots, P_M$ have disjoint supports
- **Define**: $Y_{ij} = i, \quad$ for $i = 1, \ldots, M$
- Consider $\{\{(X_{ij}, Y_{ij})\}_{j=1}^{n_i}\}_{i=1}^{M}$; $\quad X \sim \sum_{i=1}^{M} \pi_i P_Y$; $\quad \pi_i \approx \frac{n_i}{\sum_{\ell=1}^{M} n_\ell}$
- **Result**: (a) $P_1 = \ldots = P_M \quad$ iff $\quad X \perp\!\!\!\perp Y$;
  (b) $P_1, \ldots, P_M$ have disjoint supports $\quad$ iff $\quad Y$ is a deterministic function of X

A measure of dissimilarity between M-distributions

- M-distributions $P_1, \ldots, P_M$ on $\mathcal{X}$ (e.g., metric space)
- **Data**: $\{X_{ij}\}_{j=1}^{n_i} \overset{iid}{\sim} P_i \quad$ for $i = 1, \ldots, M$
- **Question**: How different are the M distributions?
- We want to find a class of measures $\gamma \equiv \gamma(P_1, \ldots, P_M)$ such that:
  (i) $\gamma \in [0, 1]$;
  (ii) $\gamma = 0 \quad$ iff $\quad P_1 = \ldots = P_M$ (i.e., all the distributions are same);
  (iii) $\gamma = 1 \quad$ iff $\quad P_1, \ldots, P_M$ have disjoint supports
- **Define**: $Y_{ij} = i, \quad$ for $i = 1, \ldots, M$
- Consider $\{\{(X_{ij}, Y_{ij})\}_{j=1}^{n_i}\}_{i=1}^{M}; \quad X \sim \sum_{i=1}^{M} \pi_i P_Y; \quad \pi_i \approx \frac{n_i}{\sum_{\ell=1}^{M} n_\ell}$
- **Result**: (a) $P_1 = \ldots = P_M \quad$ iff $\quad X \perp\!\!\!\perp Y$;
  (b) $P_1, \ldots, P_M$ have disjoint supports $\quad$ iff $\quad Y$ is a deterministic function of X
- "Similar" kernel and graph-based strategy yields a desired measure

Summary

- Measure the strength of association between X and Y on $\mathcal{X}$ and $\mathcal{Y}$
- Class of kernel measures of association (KMAc) when $\mathcal{Y}$ admits a nonnegative definite kernel
- Class of geometric 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 automatically adapts to the intrinsic dimension of the support of μ_X
- Established a pivotal Gaussian limit uniformly over a class of graphs

Thank you very much!

Questions?

Near Linear time Estimator

- Note

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

- Replace with

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

- Define

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

Properties of $\hat{\eta}_n^{\text{lin}}$

- When G_n is k-NNG with $k = \mathcal{O}(1)$, $\hat{\eta}_n^{\text{lin}}$ takes $\mathcal{O}(n \log n)$ time
- $\hat{\eta}_n^{\text{lin}} \xrightarrow{\mathbb{P}} \eta_K$ (the same measure of association)
- $\hat{\eta}_n^{\text{lin}}$ has the same rate of convergence as $\hat{\eta}_n$
- There exists a sequence of random variables $\tilde{V}_n = \mathcal{O}_{\mathbb{P}}(1)$ such that:

$$\frac{\sqrt{n}\,\hat{\eta}_n^{\text{lin}}}{\tilde{V}_n} \xrightarrow{d} \mathcal{N}(0,1)$$

  where $\tilde{V}_n$ can be computed in near linear time

- Yields a near linear time test for independence; cf. distance correlation (or HSIC) takes $\mathcal{O}(n^2 B)$ time if B permutations are used

Properties of $\hat{\eta}_n^{\text{lin}}$

- When G_n is k-NNG with $k = \mathcal{O}(1)$, $\hat{\eta}_n^{\text{lin}}$ takes $\mathcal{O}(n \log n)$ time
- $\hat{\eta}_n^{\text{lin}} \xrightarrow{\mathbb{P}} \eta_K$ (the same measure of association)
- $\hat{\eta}_n^{\text{lin}}$ has the same rate of convergence as $\hat{\eta}_n$
- There exists a sequence of random variables $\tilde{V}_n = \mathcal{O}_{\mathbb{P}}(1)$ such that:

$$\frac{\sqrt{n}\,\hat{\eta}_n^{\text{lin}}}{\tilde{V}_n} \xrightarrow{d} \mathcal{N}(0,1)$$

  where $\tilde{V}_n$ can be computed in near linear time

- Yields a near linear time test for independence; cf. distance correlation (or HSIC) takes $\mathcal{O}(n^2 B)$ time if B permutations are used
- **Price to pay**: Has a higher asymptotic variance than $\hat{\eta}_n$