

On the Importance of Asymmetry and Monotonicity Constraints in Maximal Correlation Analysis

Elad Domanovitz and Uri Erez

May 13th, 2019
University of Toronto

Introduction

- Let X and Y be random variables on a probability space (neither of them being constant with probability 1)
- **How to characterize by a numerical value the strength of statistical dependence between X and Y ?**
- Fundamental problem that is well studied
- Our focus is on continuous random variables

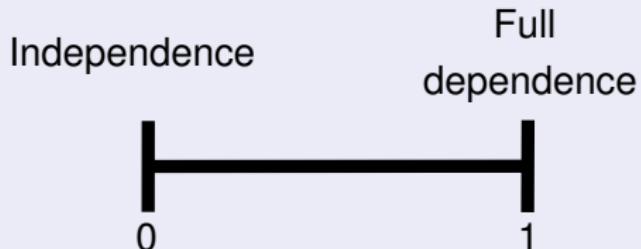
Key application

Approximate (estimate in an average least squares sense) one random variable from the other

Introduction

What would define a “good” measure?

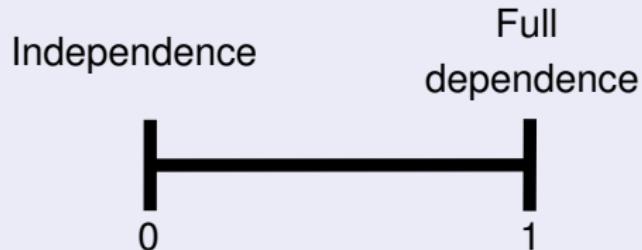
- Serves for comparison \implies range is arbitrary; can take it to be $[0, 1]$
- Value lies between the two extremes:



Introduction

What would define a “good” measure?

- Serves for comparison \implies range is arbitrary; can take it to be $[0, 1]$
- Value lies between the two extremes:

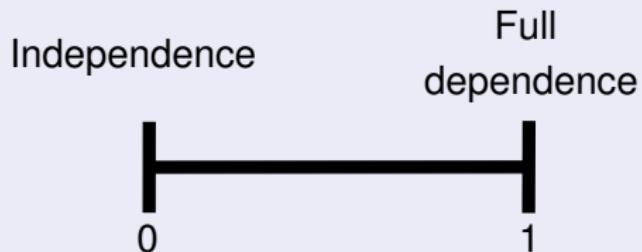


If and only if !

Introduction

What would define a “good” measure?

- Serves for comparison \implies range is arbitrary; can take it to be $[0, 1]$
- Value lies between the two extremes:



If and only if !

Key points

- Exact numerical values can be “negotiated” ... but not the extremes
- Symmetric?

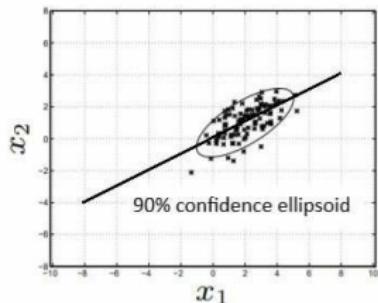
What is known?

Pearson's correlation coefficient (1880)

- Approximate random variable Y as an affine function of random variable X : $Y = aX + b$

$$\rho(X \leftrightarrow Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{var}(X)}\sqrt{\text{var}(Y)}}$$

- $|\rho(X \leftrightarrow Y)|$ works very well for Gaussian vector: extreme cases are captured



$$LMMSE = (1 - \rho^2(X \leftrightarrow Y))\text{var}(Y)$$

$$|\rho(X \leftrightarrow Y))| = \sqrt{1 - \frac{LMMSE}{\text{var}(Y)}}$$

What is known?

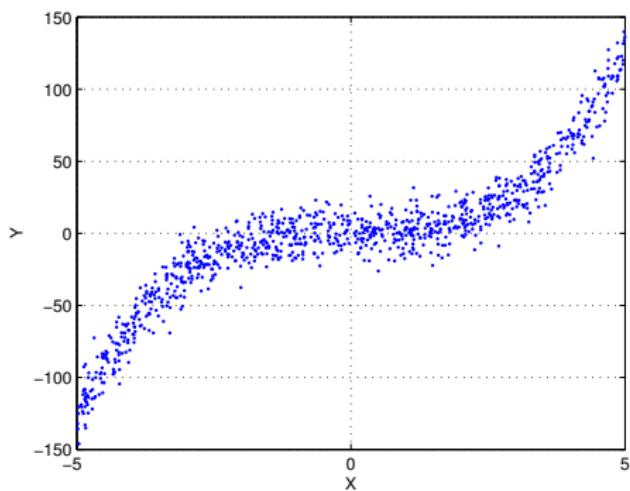
Pearson's correlation coefficient (1880)

- How about non-Gaussian pairs?
- Can still use MMSE estimator to approximate the random variable Y as: $Y = f(X)$ where $f(x) = \mathbb{E}[Y|X = x]$
- For jointly Gaussian case, reduces to the linear estimator

Non Gaussian pair with non-linear MMSE estimator

Example

- $X \sim \mathcal{U}[-5, 5]$
- $Z \sim \mathcal{N}(0, 10)$
- $Y = X^3 + Z$



Natural generalization of Pearson coefficient: correlation ratio, Pearson (1909)

- Approximate random variable Y as a general function of X :
$$Y = f(X)$$

$$\theta(X \rightarrow Y) = \sqrt{1 - \frac{\mathbb{E}[\text{var}(Y|X)]}{\text{var}(Y)}} = \sqrt{\frac{\text{var}(\mathbb{E}[Y|X])}{\text{var}(Y)}}$$

Definition

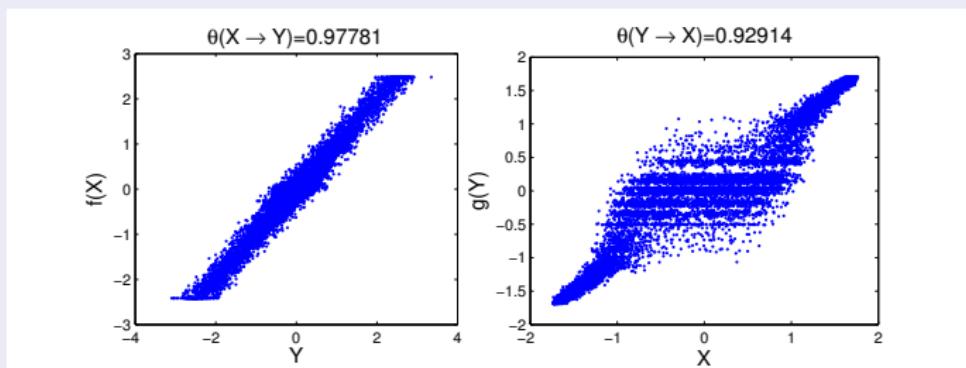
The supremum over all (admissible) functions f of the correlation between $f(X)$ and Y :

$$\theta(X \rightarrow Y) = \sup_f \rho(f(X) \leftrightarrow Y)$$

- In words: measures how well Y can be approximated (in a mean squared error sense) as a linear function of $X' = f(X)$
- Does it satisfy the key requirements?

Correlation ratio

- Non-symmetric

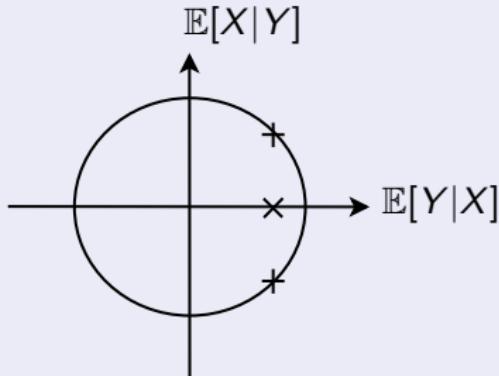


- While it bothered some people, we don't take it as a major drawback...

Correlation ratio

- “Equals zero too easily” \implies can vanish even when variables are dependent

- X and Y are uniformly distributed over a circle of radius 1:

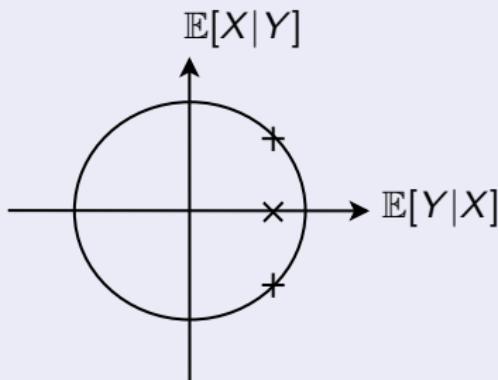


- Given X , $E[Y|X] = 0$ for all X !
 - $\implies \theta(X \rightarrow Y) = \sqrt{\frac{\text{var}(E[Y|X])}{\text{var}(Y)}} = 0\dots$

Correlation ratio

- “Equals zero too easily” \implies can vanish even when variables are dependent

- X and Y are uniformly distributed over a circle of radius 1:



- Given X , $E[Y|X] = 0$ for all X !
 - $\implies \theta(X \rightarrow Y) = \sqrt{\frac{\text{var}(E[Y|X])}{\text{var}(Y)}} = 0\dots$

**Detects well full dependence, may give false alarms
on independence**

Another attempt - maximal correlation

- Known as Hirschfeld-Gebelein-Rényi (1935, 1941, 1959) maximal correlation coefficient
- Approximate a function of random variable Y as a function of X :
$$g(Y) = f(X)$$

Definition

The supremum over all (admissible) functions f, g of the correlation between $f(X)$ and $g(Y)$:

$$\rho_{\max}^{**}(X \leftrightarrow Y) = \sup_{f,g} \rho(f(X) \leftrightarrow g(Y)).$$

- In words: measures how well $Y' = g(Y)$ can be approximated (in a mean squared error sense) as a linear function of $X' = f(X)$
- Widely used since readily computable numerically via the alternating conditional expectation (ACE) algorithm

Rényi's axioms

- ① $r(X \rightarrow Y)$ is defined for any pair of random variables X and Y neither of them being constant with probability 1
- ② $r(X \rightarrow Y) = r(Y \rightarrow X)$
- ③ $0 \leq r(X \rightarrow Y) \leq 1$
- ④ $r(X \rightarrow Y) = 0$ if and only if X and Y are independent
- ⑤ $r(X \rightarrow Y) = 1$ if there is a strict dependence between X and Y , i.e., either $X = g(Y)$ or $Y = f(X)$ where $g(Y)$ and $f(X)$ are Borel-measurable functions
- ⑥ If the Borel-measurable functions $f(X)$ and $g(Y)$ map the real axis in a one-to-one way onto itself, $r(f(X), g(Y)) = r(X, Y)$
- ⑦ If the joint distribution of X and Y is normal, then
$$r(X, Y) = |\rho(X, Y)|$$

Maximal correlation satisfies all axioms

Rényi's axioms

- ① $r(X \rightarrow Y)$ is defined for any pair of random variables X and Y neither of them being constant with probability 1
- ② $r(X \rightarrow Y) = r(Y \rightarrow X)$
- ③ $0 \leq r(X \rightarrow Y) \leq 1$
- ④ $r(X \rightarrow Y) = 0$ if and only if X and Y are independent
- ⑤ $r(X \rightarrow Y) = 1$ if **(but not only if)** there is a strict dependence between X and Y , i. e., either $X = g(Y)$ or $Y = f(X)$ where $g(Y)$ and $f(X)$ are Borel-measurable functions
- ⑥ If the Borel-measurable functions $f(X)$ and $g(Y)$ map the real axis in a one-to-one way onto itself, $r(f(X), g(Y)) = r(X, Y)$
- ⑦ If the joint distribution of X and Y is normal, then
$$r(X, Y) = |\rho(X, Y)|$$

But does it satisfy the key points?

Maximal correlation

- “With great power comes great responsibility...” (Ben Parker)
- Solving one problem results in creating another...
- The maximal correlation coefficient “equals one too easily”
- Example:
 - Two random variables sharing the LSB

$$X = C + \sum_{i=1}^N A_i 2^i; \quad Y = C + \sum_{i=1}^N B_i 2^i$$

- A_i, B_i, C are mutually independent random variables, all taking the value 0 or 1 with equal probability
- Applying modulo 2 to both variables \implies maximal correlation = 1, for any value of N

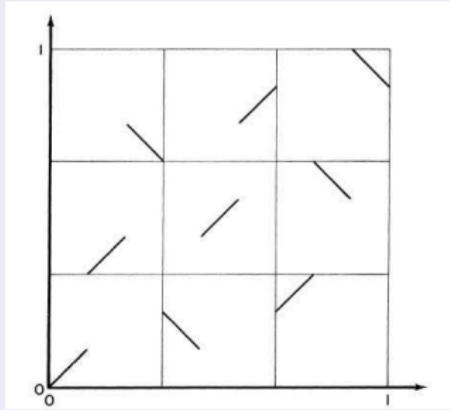
- Rényi (1959):

"It seems at the first sight natural to postulate that $r(X \rightarrow Y) = 1$ only if there is a strict dependence of the mentioned type between X and Y , but this condition is rather restrictive, and it is better to leave it out."

**Detects well independence, may give false alarms
on full dependence**

Additional suggestion

- There exists a sequence $(X(N), Y(N))$ of pairs of random variables which converge in distribution to (X, Y) such that
 - ▶ For each element $\rho_{\max}^{**}(X(N) \leftrightarrow Y(N)) = 1$
 - ▶ But it converges in distribution to (X, Y) such that $\rho_{\max}^{**}(X \leftrightarrow Y) = 0$
- Example 1: Shared LSB while taking $(X(N)/2^N, Y(N)/2^N)$
- Example 2:



Additional suggestions

- Kimeldorf and Sampson (1978):

Definition

Approximate a monotone function of random variable Y as
a monotone function of X : $g(Y) = f(X)$

$$\rho_{\max}^{mm}(X \leftrightarrow Y) = \sup_{f,g} \rho(f(X) \leftrightarrow g(Y))$$

- In words: measures how well a **monotone** $Y' = g(Y)$ can be approximated (in a mean squared error sense) as a linear function of a **monotone** $X' = f(X)$
- The measure *still* equals one too easily
- Specifically, it can be one for a pair of random variables that are not completely dependent.

Additional suggestions

- Further, this may be too restrictive...
- From “[Monotone Regression Splines in Action]: Comment” published in *Statistical Science* by Hastie and Tibshirani (1988):
“A monotone restriction makes sense for a response transformation because it is necessary to allow predictions of the response from the estimated model. On the other hand, why restrict predictor transformations to be monotone? Instead, why not leave them unrestricted and let the data suggest the shape of the relevant transformation?”
- They wave symmetry
- In fact, monotonicity is good, but not good enough...

The semi- κ -monotone maximal correlation measure

- Slightly improve Hastie and Tibshirani's suggestion

Definition

For $0 \leq \kappa \leq 1$, a function f is said to be κ -increasing, if for all $x_2 \geq x_1$:

$$\kappa(x_2 - x_1) \leq f(x_2) - f(x_1) \leq \frac{1}{\kappa}(x_2 - x_1)$$

Definition

For a given $0 < \kappa < 1$, the semi- κ -monotone maximal correlation measure is defined as

$$\rho_{\max}^{*m_\kappa}(X \rightarrow Y) = \sup_{f,g} \rho(f(X) \leftrightarrow g(Y))$$

where the supremum is taken over all admissible functions $f(x)$, and over κ -increasing admissible functions $g(y)$.

- Well, this is not symmetric anymore...
- But, estimation is not a symmetric process
- Obviously this does not satisfy Rényi's axioms
- Well... axioms can be modified...
- We follow Hall (1967) and Liu (2014) and define the following modified set of axioms

Modified axioms

- ① $r(X \rightarrow Y)$ is defined for any pair of random variables X and Y neither of them being constant with probability 1
- ② $r(X \rightarrow Y)$ may not be equal to $r(Y \rightarrow X)$
- ③ $0 \leq r(X \rightarrow Y) \leq 1$
- ④ $r(X \rightarrow Y) = 0$ if and only if X and Y are independent
- ⑤ $r(X \rightarrow Y) = 1$ if and only if there is a strict dependence between X and Y , i. e., either $X = g(Y)$ or $Y = f(X)$ where $g(Y)$ and $f(X)$ are is a Borel-measurable function
- ⑥ If the Borel-measurable functions $f(X)$ and $g(Y)$ map the real axis in a one-to-one way onto itself, $r(f(X), Y) = r(X, Y)$
- ⑦ If the joint distribution of X and Y is normal, then
 $r(X, Y) = |\rho(X, Y)|$

Modified axioms

- ① $r(X \rightarrow Y)$ is defined for any pair of random variables X and Y neither of them being constant with probability 1
- ② $r(X \rightarrow Y)$ may not be equal to $r(Y \rightarrow X)$
- ③ $0 \leq r(X \rightarrow Y) \leq 1$
- ④ $r(X \rightarrow Y) = 0$ if and only if X and Y are independent
- ⑤ $r(X \rightarrow Y) = 1$ if and only if there is a strict dependence between X and Y , i. e., either $X = g(Y)$ or $Y = f(X)$ where $g(Y)$ and $f(X)$ are is a Borel-measurable function
- ⑥ If the Borel-measurable functions $f(X)$ and $g(Y)$ map the real axis in a one-to-one way onto itself, $r(f(X), Y) = r(X, Y)$
- ⑦ If the joint distribution of X and Y is normal, then
 $r(X, Y) = |\rho(X, Y)|$

The semi- κ -monotone maximal correlation measure
satisfy these axioms

Modified axioms

- ✓ $r(X \rightarrow Y)$ is defined for any pair of random variables X and Y neither of them being constant with probability 1
 - ✓ $r(X \rightarrow Y)$ may not be equal to $r(Y \rightarrow X)$
 - ✓ $0 \leq r(X \rightarrow Y) \leq 1$
- ④ $r(X \rightarrow Y) = 0$ if and only if X and Y are independent
- ⑤ $r(X \rightarrow Y) = 1$ if and only if there is a strict dependence between X and Y , i.e., either $X = g(Y)$ or $Y = f(X)$ where $g(Y)$ and $f(X)$ are is a Borel-measurable function
- ⑥ If the Borel-measurable functions $f(X)$ and $g(Y)$ map the real axis in a one-to-one way onto itself, $r(f(X), Y) = r(X, Y)$
- ⑦ If the joint distribution of X and Y is normal, then
 $r(X, Y) = |\rho(X, Y)|$

Axiom D: breaking the symmetry

- If X, Y are independent $\Rightarrow \rho_{\max}^{*m_\kappa}(X \rightarrow Y) = 0$ (as so is even $\rho_{\max}^{**}(X \leftrightarrow Y)$)

Axiom D: breaking the symmetry

- For the other direction it suffices to consider the case where the correlation ratio equals 0 and X, Y are dependent

- $\theta(X \rightarrow Y) = 0 \implies \mathbb{E}[Y|X = x] = \int p(y|x)ydy \equiv const \implies \text{var}(\mathbb{E}[Y|X]) = 0$
- We may break the symmetry of $g(y) = y$ by defining, e.g.,

$$g_{a,\kappa}(y) = \begin{cases} y & y \geq a \\ \kappa y & y < a \end{cases}, \quad \kappa < 1 !!!$$

- Consider two values of x_1 and x_2 for which the functions $p(y|x_i)$ are not identical
- Let a be a value such that

$$\int^a p(y|x_1)ydy \neq \int^a p(y|x_2)ydy$$

- $\implies \mathbb{E}[g_a(Y)|X = x_1] \neq \mathbb{E}[g_a(Y)|X = x_2]$
- \implies the correlation ratio of $Y' = g_a(Y)$ on X is non-zero
- $\rho_{\max}^{*m_\kappa}(X \rightarrow Y) \geq \theta(X \rightarrow g_{a,\kappa}(y)) > 0$

Modified axioms

- ✓ $r(X \rightarrow Y)$ is defined for any pair of random variables X and Y neither of them being constant with probability 1
 - ✓ $r(X \rightarrow Y)$ may not be equal to $r(Y \rightarrow X)$
 - ✓ $0 \leq r(X \rightarrow Y) \leq 1$
 - ✓ $r(X \rightarrow Y) = 0$ if and only if X and Y are independent
- ⑤ $r(X \rightarrow Y) = 1$ if and only if there is a strict dependence between X and Y , i.e., either $X = g(Y)$ or $Y = f(X)$ where $g(Y)$ and $f(X)$ are ~~is a~~ Borel-measurable function
 - ⑥ If the Borel-measurable functions $f(X)$ and $g(Y)$ map the real axis in a one-to-one way onto itself, $r(f(X), Y) = r(X, Y)$
 - ⑦ If the joint distribution of X and Y is normal, then $r(X, Y) = |\rho(X, Y)|$

Axiom E

- If $Y = f(X)$ (almost surely), then $\rho_{\max}^{*m_\kappa}(X \rightarrow Y) = 1$
- If $\rho_{\max}^{*m_\kappa}(X \rightarrow Y) = 1$
- For $0 < \kappa < 1$ the supremum is attainable
- \Rightarrow there is a *perfect* linear regression between $g(Y)$ and $f'(X)$ (g, f' being maximizing functions of the measure):
 - ▶ $\Rightarrow g(Y) = af'(X) + b$ where g is an increasing function with slope greater than κ
 - ▶ κ is strictly positive $\Rightarrow g$ is invertible, and also $g^{-1}(Y)$ has finite variance (since the slope of $g^{-1}(Y)$ is at most $\frac{1}{\kappa}$ and Y has finite variance)
 - ▶ $\Rightarrow Y = g^{-1}(af'(X) + b) = f(X)$

Modified axioms

- ✓ $r(X \rightarrow Y)$ is defined for any pair of random variables X and Y neither of them being constant with probability 1
 - ✓ $r(X \rightarrow Y)$ may not be equal to $r(Y \rightarrow X)$
 - ✓ $0 \leq r(X \rightarrow Y) \leq 1$
 - ✓ $r(X \rightarrow Y) = 0$ if and only if X and Y are independent
 - ✓ $r(X \rightarrow Y) = 1$ if and only if there is a strict dependence between X and Y , i. e., either $X = g(Y)$ or $Y = f(X)$ where $g(Y)$ and $f(X)$ are is a Borel-measurable function
 - ✓ If the Borel-measurable functions $f(X)$ and $g(Y)$ map the real axis in a one-to-one way onto itself, $r(f(X), Y) = r(X, Y)$
- 7 If the joint distribution of X and Y is normal, then
 $r(X, Y) = |\rho(X, Y)|$

Axiom G

- It is well known (Lancaster (1957)) that when X, Y are jointly normal with correlation coefficient ρ , then $\rho_{\max}^{**}(X \leftrightarrow Y) = |\rho|$
- \implies The maximal correlation is achieved taking $g(y) = y$ (a monotone function)
- $\implies f(x) = x$ or $f(x) = -x \implies$

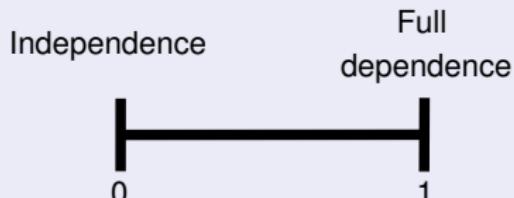
$$\begin{aligned}\rho_{\max}^{*m_\kappa}(X \rightarrow Y) &= \rho_{\max}^{**}(X \leftrightarrow Y) \\ &= |\rho|\end{aligned}$$

Modified axioms

- ✓ $r(X \rightarrow Y)$ is defined for any pair of random variables X and Y neither of them being constant with probability 1
- ✓ $r(X \rightarrow Y)$ may not be equal to $r(Y \rightarrow X)$
- ✓ $0 \leq r(X \rightarrow Y) \leq 1$
- ✓ $r(X \rightarrow Y) = 0$ if and only if X and Y are independent
- ✓ $r(X \rightarrow Y) = 1$ if and only if there is a strict dependence between X and Y , i. e., either $X = g(Y)$ or $Y = f(X)$ where $g(Y)$ and $f(X)$ are is a Borel-measurable function
- ✓ If the Borel-measurable functions $f(X)$ and $g(Y)$ map the real axis in a one-to-one way onto itself, $r(f(X), Y) = r(X, Y)$
- ✓ If the joint distribution of X and Y is normal, then
 $r(X, Y) = |\rho(X, Y)|$

Interim summary

- Key points for the semi- κ -monotone maximal correlation:
 - ▶ Detects well **both** independence and full dependence (QUALITATIVE)



If and only if !

- ▶ Not symmetric, no problem...
- Extreme values of κ :
 - ▶ $\kappa = 0$ results in weak monotonicity (Hastie and Tibshirani)
 - ▶ $\kappa = 1$ results in the correlation ratio
 - ▶ **Do not satisfy the modified axioms**
- The value of κ can be used to control how far we deviate from the correlation ratio (QUANTITATIVE)

What about an efficient algorithm to compute the semi- κ -monotone maximal correlation?

- Well...
- The ACE algorithm suggested by Breiman and Friedman (1985) was shown to calculate the maximal correlation measure
- They showed that
 - ▶ Optimal transformations exist
 - ▶ Each iteration improves the measure
 - ▶ The algorithm converges to the **optimal transformations**
- Before showing modifications, let's discuss the vector observation case

Vector observation case

- Let $\mathbf{X} = (X_1, \dots, X_p)$ be a vector of variables
- The maximal correlation coefficient becomes

$$\rho_{\max}^{**}(\mathbf{X} \leftrightarrow Y) = \sup_{f,g} \rho(f(\mathbf{X}) \leftrightarrow g(Y))$$

- Following Breiman and Friedman, we also consider a simplified (quasi-additive) relationship between Y and \mathbf{X} :

$$g(Y) = \sum_i f_i(X_i)$$

- Breiman and Friedman provide conditions for the existence of optimal transformations $\{f_i\}, g$ such that the supremum is attained are given, and it is shown that under these conditions the ACE algorithm converges to the optimal transformations

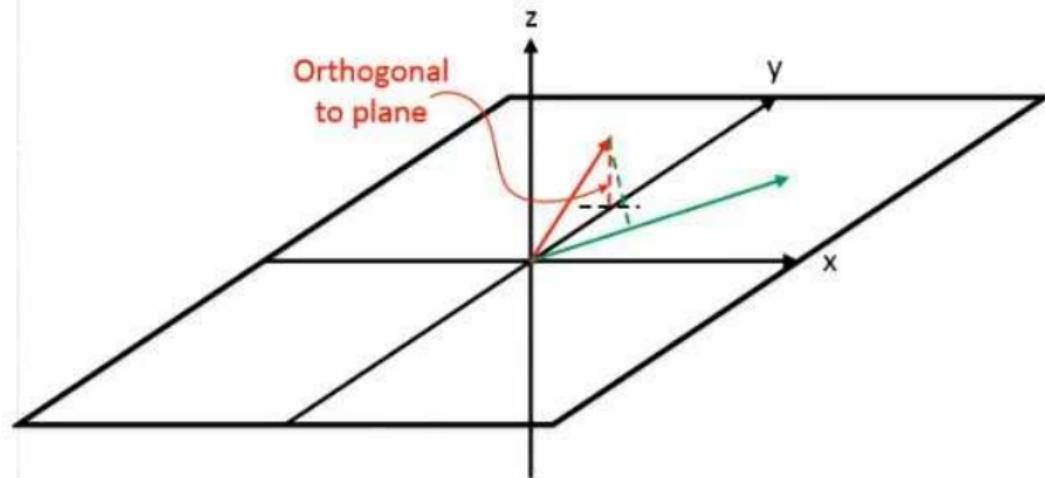
Modified ACE algorithm

- We begin by presenting a modification of the ACE algorithm to compute the semi-0-monotone maximal correlation measure $\rho_{\max}^{*m_0}(X \rightarrow Y)$, restricting the function applied to the response variable only to be weakly monotone
- We do not have an algorithm for computing the semi- κ -monotone maximal correlation measure for $\kappa \neq 0$
- Instead, we present a regularized version of the semi-0-monotone ACE algorithm

Calculating semi-0-monotone maximal correlation measure

- High level overview:
 - ▶ $X, Y \in \mathcal{H}_2$ Hilbert space of finite-variance random variables
 - ▶ $\mathcal{H}_2(X) =$ set of all random variables corresponding to an admissible function of X (subspace of \mathcal{H}_2)
 - ▶ $\mathcal{H}_2(Y) =$ set of all random variables corresponding to an admissible function of Y (subspace of \mathcal{H}_2)
 - ▶ $\mathcal{M}_0(Y) =$ non-decreasing admissible functions of Y (closed and convex subset of $\mathcal{H}_2(Y)$)
- Algorithm goal: minimize the angle between $f(X) \in \mathcal{H}_2(X)$ and $g(Y) \in \mathcal{M}_0(Y)$
- Per iteration: given $f(X)$, find $g(Y) \in \mathcal{M}_0(Y)$ with smallest angle
 \iff given $f(X)$, find $g(Y) \in \mathcal{M}_0(Y)$ with smallest distance
- And vice versa...
- \implies At each iteration angle decreases

Calculating semi-0-monotone maximal correlation measure



- If $g(Y) \in \mathcal{M}_0(Y)$ then $\forall \alpha > 0$, $\alpha g(Y) \in \mathcal{M}_0(Y) \implies$ nearest point satisfies orthogonality
- Angle decreases at each iteration \implies convergence

- Denote by $P_{\mathcal{A}}(Y)$ the orthogonal projection of Y onto the closed convex set \mathcal{A}
 - $\mathcal{P}_{\mathcal{H}_2(X)}(g(Y)) = \mathbb{E}[g(Y) | X]$
 - $\mathcal{P}_{\mathcal{M}_0(Y)}(f(X))$ is called isotonic regression

Algorithm 1

- 1: **procedure** CALCULATE-SEMI-0-MONOTONE
- 2: Set $g(Y) = Y/\|Y\|$;
- 3: **while** $e^2(g, f)$ decreases **do**
- 4: $f'(X) = \mathcal{P}_{\mathcal{H}_2(X)}(g(Y))$
- 5: replace $f(X)$ with $f'(X)$
- 6: $g'(Y) = \mathcal{P}_{\mathcal{M}_0(Y)}(f(X))$
- 7: replace $g(Y)$ with $g'(Y)/\|g'(Y)\|$
- 8: End modified ACE

Regularized ACE

- To limit $g(Y)$ to have (lower and upper) slope κ , we apply:

$$g_1(Y) = g^{-1}(Y) + \kappa \cdot Y$$

$$g(Y) = g_1^{-1}(Y) + \kappa \cdot Y$$

- Results in a slope lower bounded by κ and upper bounded by $1/\kappa + \kappa$

Algorithm 2

```
1: procedure REGULARIZED-ACE
2:   Set  $g(Y) = Y/\|Y\|$ ;
3:   while  $e^2(g, f)$  decreases do
4:      $f'(X) = \mathcal{P}_{\mathcal{H}_2(X)}(g(Y))$ 
5:     replace  $f(X)$  with  $f'(X)$ 
6:      $g'(Y) = \mathcal{P}_{\mathcal{M}_0(Y)}(f(X))$ 
7:     replace  $g(Y)$  with  $g'(Y)/\|g'(Y)\|$ 
8:   Apply regularization
9:   End regularized ACE
```

Examples

- Example 1 - multi-variate example where one of the two observed random variables masks the other even though the latter is more useful for estimation purposes
- Example 2 - Demonstration why correlation ratio is insufficient ($\kappa = 1$)
 - ▶ Example 2a - correlation ratio fails to detect dependence
 - ▶ Example 2b - preferred parameterization (backup)
- Example 3 - Semi-0-monotonicity is insufficient

Example 1 - multi-variate example

- Assume:

$$Y \sim \mathcal{U}[0, 1]$$

$$X_1 = \text{mod}(Y, 0.2) + N_1$$

$$X_2 = Y^3 + N_2$$

where $N_1 \sim \mathcal{N}(0, 0.01)$, $N_2 \sim \mathcal{N}(0, 0.2)$

Example 1 - multi-variate example

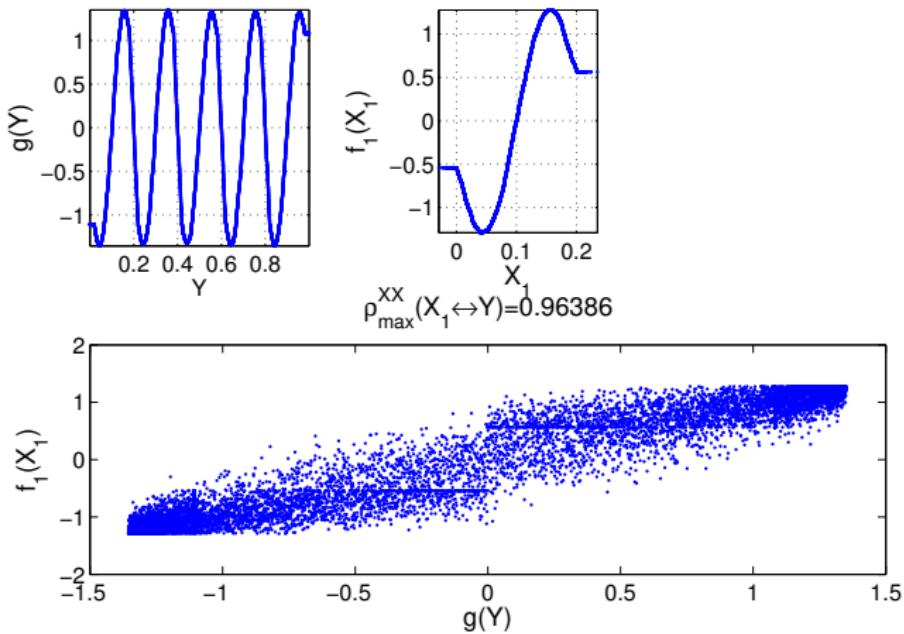


Figure: Example 1: Running ACE on Y and X_1

Example 1 - multi-variate example

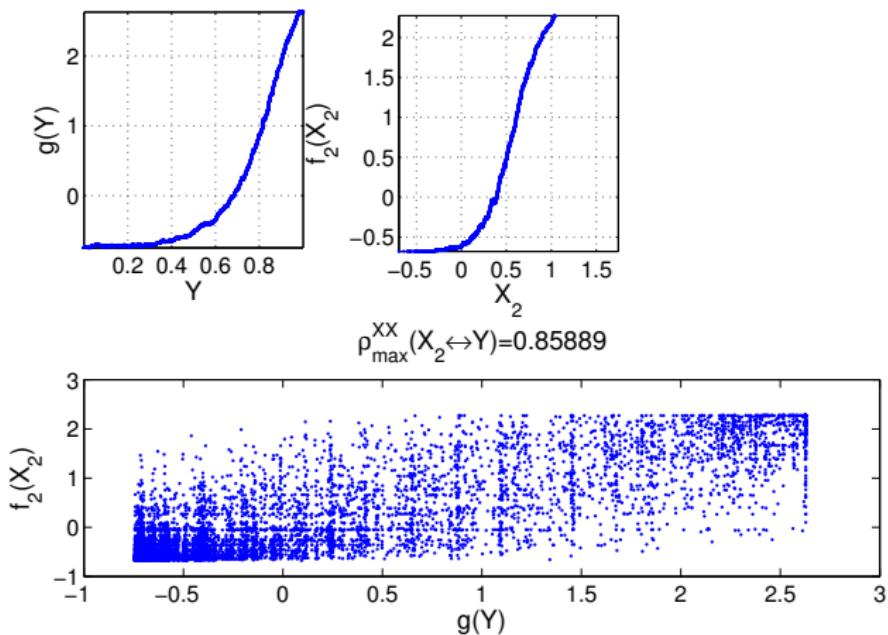


Figure: Example 1: Running ACE on Y and X_2

Example 1 - multi-variate example

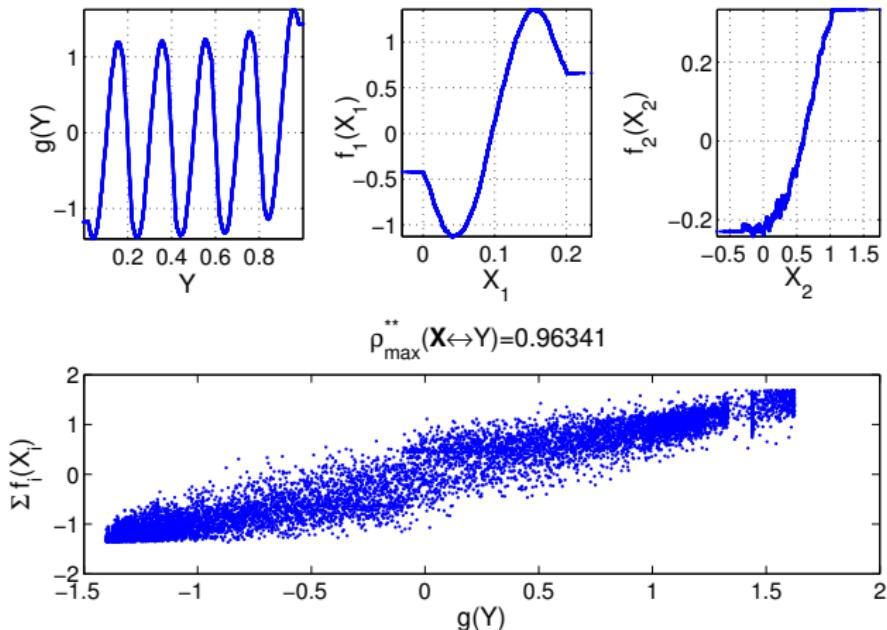


Figure: Example 1: Running ACE on Y , X_1 and X_2

Example 1 - multi-variate example

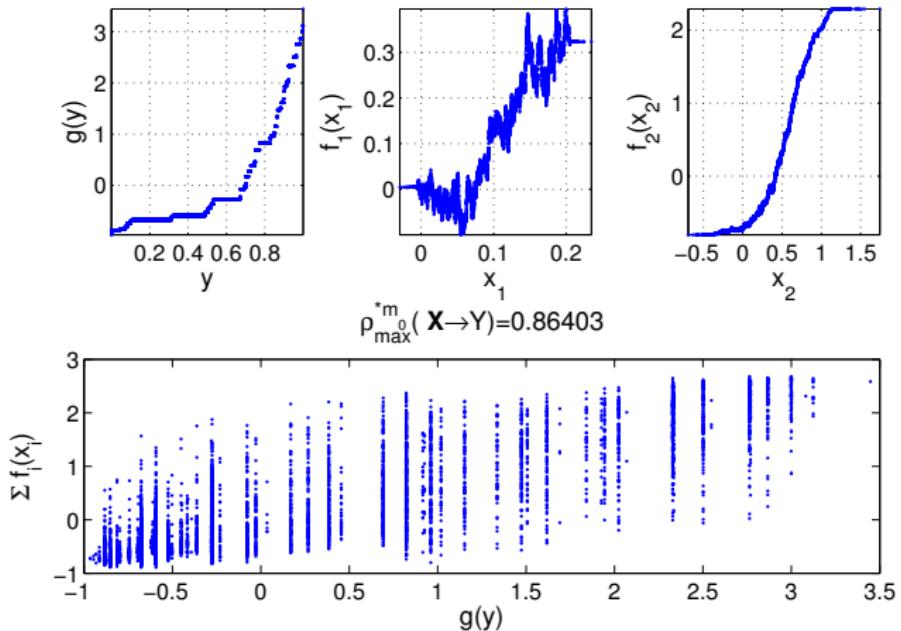


Figure: Example 1: modified ACE (Algorithm 1) on Y , X_1 and X_2 with $\kappa = 0$

Example 2a - correlation ratio fails to detect dependence

- X and Y are uniformly distributed over a circle with radius 1

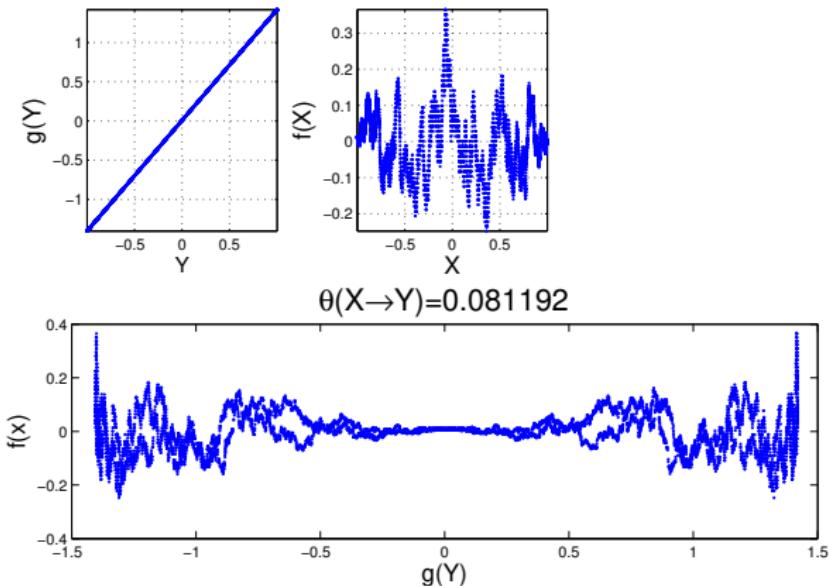


Figure: Transformation corresponding to the correlation ratio

Example 2a - correlation ratio fails to detect dependence

- X and Y are uniformly distributed over a circle with radius 1

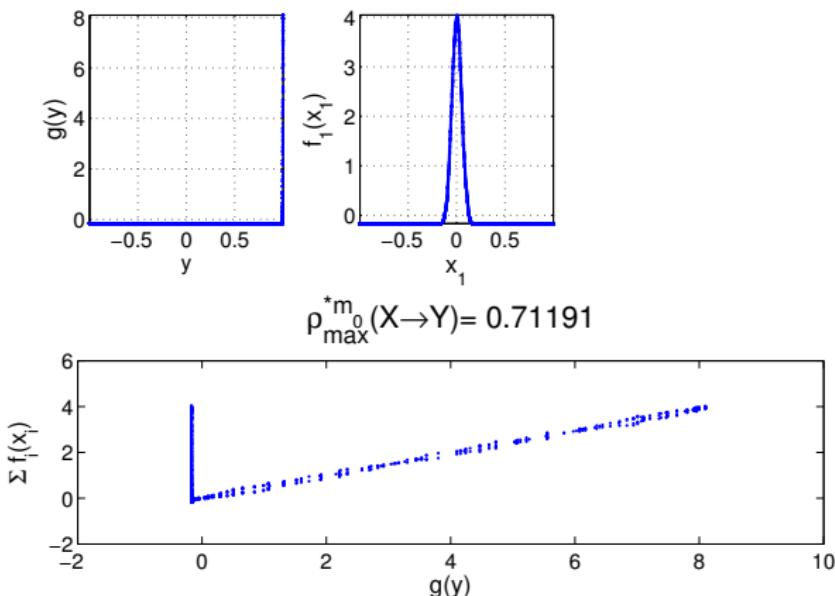
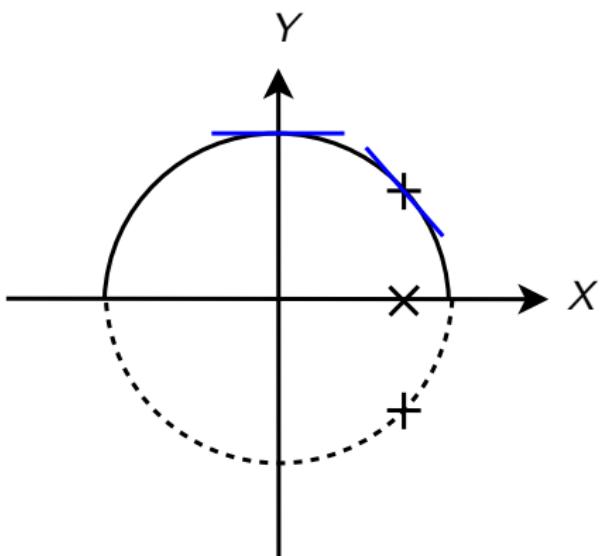


Figure: modified ACE (Algorithm 1) applied to Y and X_1 with $\kappa = 0$

Example 2a - correlation ratio fails to detect dependence

- X and Y are uniformly distributed over a circle with radius 1



- In fact, with infinite precision, $f(x) \rightarrow \delta(x)$ which suggests that $\rho_{\max}^{*m_0}(X \rightarrow Y) \rightarrow 1$.

Example 3 - Semi-0-monotonicity is insufficient

$$Y \sim \mathcal{U}[-10, 10]$$

$$X = \begin{cases} Y & Y > 9 \\ N_1 & \text{otherwise} \end{cases}$$

$$N_1 \sim \mathcal{U}[-1, 1]$$

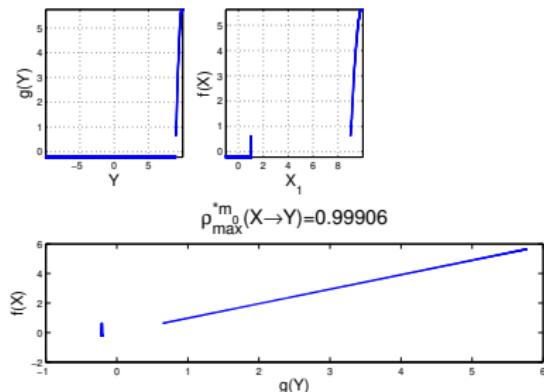


Figure: Modified ACE applied to Y, X with $\kappa = 0$

Example 3 - Semi-0-monotonicity is insufficient

$$Y \sim \mathcal{U}[-10, 10]$$

$$X = \begin{cases} Y & Y > 9 \\ N_1 & \text{otherwise} \end{cases}$$

$$N_1 \sim \mathcal{U}[-1, 1]$$

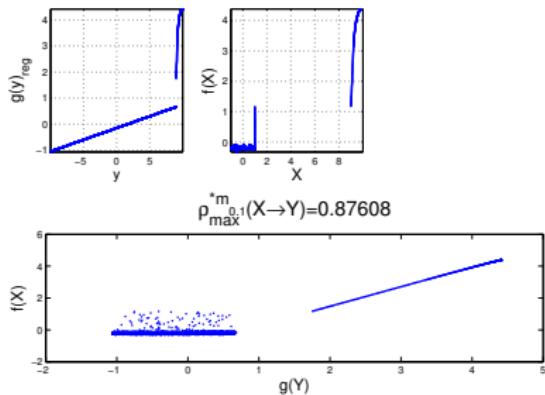


Figure: Regularized ACE applied to Y, X with $\kappa = 0.1$

Thank you for your attention

Backup

Example 2b - preferred parameterization

$$Y \sim \mathcal{U}[0, 10]$$

$$X = \log(Y) + Z; \quad Z \sim \mathcal{N}(0, 1)$$

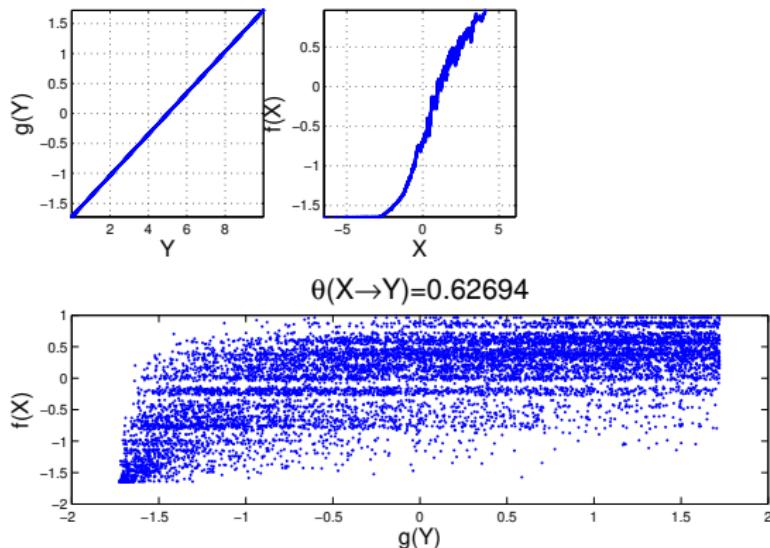


Figure: Transformations corresponding to the correlation ratio

Example 2b - preferred parameterization

$$Y \sim \mathcal{U}[0, 10]$$

$$X = \log(Y) + Z; \quad Z \sim \mathcal{N}(0, 1)$$

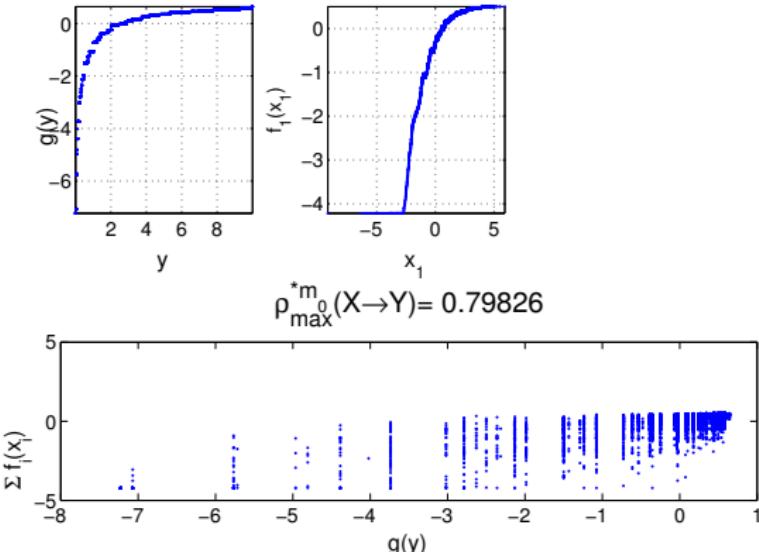


Figure: Modified ACE (Algorithm 1) applied to Y , X and X_2 with $\kappa = 0$

What about mutual information?

- Linfoot (1957):

$$L(X, Y) = \sqrt{1 - e^{-2I(X;Y)}}$$

- Was shown to satisfy all Rényi's axioms
- Suffers from the same deficiencies as the maximal correlation measure...

Problem 1

- Simple example:

$$X \sim \mathcal{U}[0, 1]; \quad N \sim \mathcal{U}[0, 1]$$

$$Y = \begin{cases} X & 0 < X < \varepsilon \\ N & O/W \end{cases}$$

- \implies When $0 < X < \varepsilon$, Y is perfectly known

- Define

$$Z = f(X) = \begin{cases} 1 & \text{if } 0 < X < \varepsilon \\ 0 & \text{O/W} \end{cases}$$

- Hence

$$\begin{aligned} I(X; Y) &= I(X, Z; Y) \\ &= I(Z; Y) + I(X; Y|Z) \\ &= \underbrace{I(Z; Y) + \Pr(Z = 0)I(X; Y|Z = 0)}_{>0} \\ &\quad + \underbrace{\Pr(Z = 1)I(X; Y|Z = 1)}_{\infty} \end{aligned}$$

Problem 2

- Another reason - can't be used for discrete variables since

$$I(X; Y) \leq H(X)$$

$$I(X; Y) \leq H(Y)$$

- Therefore, even in case of full dependence

$$I(X, Y) < \infty \implies L(X; Y) < 1$$