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

Elad Domanovitz and Uri Erez

July 12th, 2019
2019 International Symposium on Information Theory

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

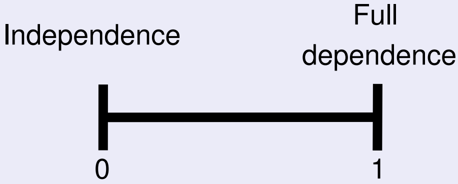
Key application

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

Introduction

What would define a “good” measure?

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



- Person’s corr. coef.
- Person’s corr. ratio
- Ranyi’s max corr.
- Normalized MI
- Linfoot’s NMI

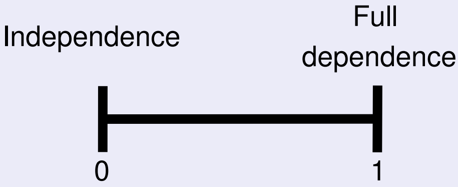
Key points

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

Introduction

What would define a “good” measure?

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



If and only if !

- × Person’s corr. coef.
- × Person’s corr. ratio
- × Ranyi’s max corr.
- × Normalized MI
- × Linfoot’s NMI

Key points

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

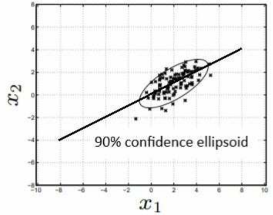
What is known?

Pearson's correlation coefficient (1880)

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

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

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



$$\begin{aligned} LMMSE &= \mathbb{E}[e^2] \\ &= (1 - \rho^2(X \leftrightarrow Y))\mathrm{var}(Y) \end{aligned}$$

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

What is known?

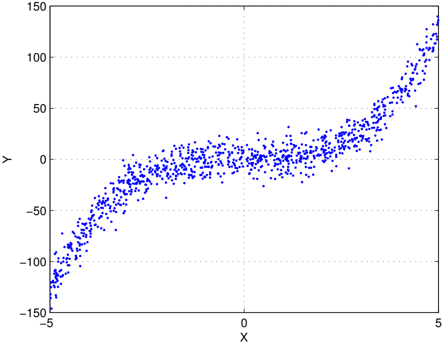
Pearson's correlation coefficient (1880)

- 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)

Person's coefficient

$$LMMSE = \mathbb{E}[e^2]$$

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

Person's correlation ratio

$$\begin{aligned} MMSE &= \mathbb{E}[e^2] \\ &= \mathbb{E}[\mathrm{var}(Y|X)] \end{aligned}$$

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

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

Definition

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

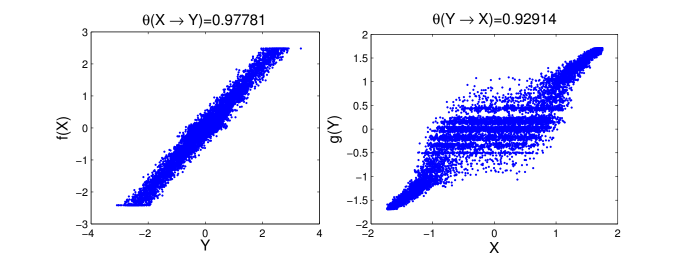
$$\theta(X \to Y) = \sup_f \rho(f(X) \leftrightarrow Y)$$

- 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

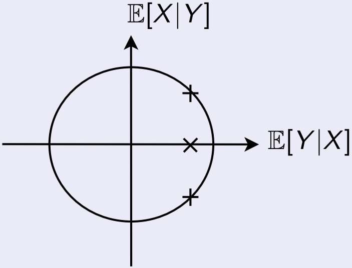
- Non-symmetric



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

Correlation ratio

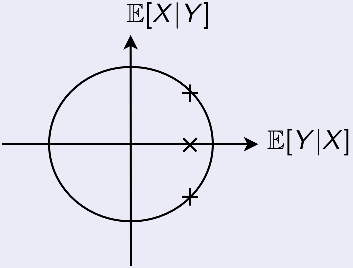
- "Equals zero too easily" $\Longrightarrow$ can vanish even when variables are dependent
  - X and Y are uniformly distributed over a circle of radius 1:



  - Given X, $\mathbb{E}[Y|X] = 0$ for all X !
  - $\Longrightarrow \theta(X \to Y) = \sqrt{\frac{\mathrm{var}(\mathbb{E}[Y|X])}{\mathrm{var}(Y)}} = 0\ldots$

Correlation ratio

- "Equals zero too easily" $\Longrightarrow$ can vanish even when variables are dependent
  - X and Y are uniformly distributed over a circle of radius 1:



  - Given X, $\mathbb{E}[Y|X] = 0$ for all X !
  - $\Longrightarrow \theta(X \to Y) = \sqrt{\frac{\mathrm{var}(\mathbb{E}[Y|X])}{\mathrm{var}(Y)}} = 0$...

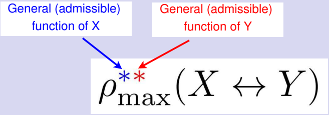
Detects well full dependence, may give false alarms on independence

Another attempt - maximal correlation

Definition

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

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



General (admissible) function of X

General (admissible) function of Y

$$\rho^{**}_{\max}(X \leftrightarrow Y)$$

- In words: measures how well $Y' = g(Y)$ can be approximated (in a mean squared error sense) as a linear function of $X' = f(X)$
- Known as Hirschfeld-Gebelein-Rényi (1935, 1941, 1959) maximal correlation coefficient
- Widely used since readily computable numerically via the alternating conditional expectation (ACE) algorithm

Rényi's axioms

1. $r(X \to Y)$ is defined for any pair of random variables X and Y neither of them being constant with probability 1
2. $r(X \to Y) = r(Y \to X)$
3. $0 \leq r(X \to Y) \leq 1$
4. $r(X \to Y) = 0$ if and only if X and Y are independent
5. $r(X \to Y) = 1$ 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

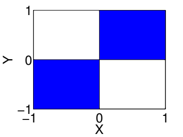
Rényi's axioms (1959)

1. $r(X \to Y)$ is defined for any pair of random variables X and Y neither of them being constant with probability 1
2. $r(X \to Y) = r(Y \to X)$
3. $0 \leq r(X \to Y) \leq 1$
4. $r(X \to Y) = 0$ if and only if X and Y are independent
5. $r(X \to Y) = 1$ 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 the circle but fails on the square...



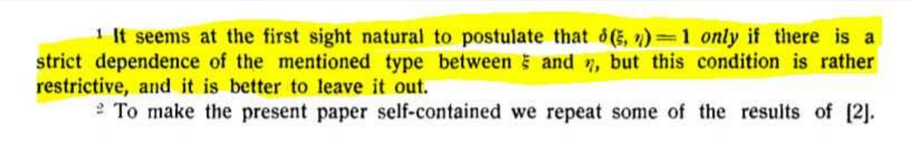
- The maximal correlation coefficient "equals one too easily"
- Another example:
  - Two random variables sharing the LSB

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

  - A_i, B_i, C are mutually independent random variables
  - $f(X) = g(Y) =$ modulo 2 $\Longrightarrow \rho^{**}_{\max}(X \leftrightarrow Y) = 1$, for any value of N

Maximal correlation

Rényi (1959)



[1] It seems at the first sight natural to postulate that $\delta(\xi, \eta) = 1$ *only* if there is a strict dependence of the mentioned type between ξ and η, but this condition is rather restrictive, and it is better to leave it out.

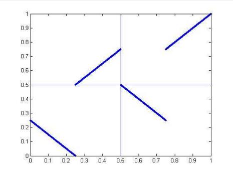
[2] To make the present paper self-contained we repeat some of the results of [2].

Detects well independence, may give false alarms on full dependence

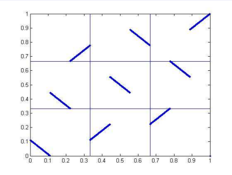
Behaviour at the limit

- There exists a sequence $(X(N), Y(N)) \xrightarrow{d} (X, Y)$ such that:
  - For each element $\rho^{**}_{\max}(X(N) \leftrightarrow Y(N)) = 1$
  - But $\rho^{**}_{\max}(X \leftrightarrow Y) = 0$
- Example 1: Shared LSB while taking $(X(N)/2^N, Y(N)/2^N)$
- Example 2:

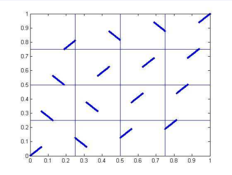
N=2



N=3



N=4



Additional suggestions

- Kimeldorf and Sampson (1978):

Definition

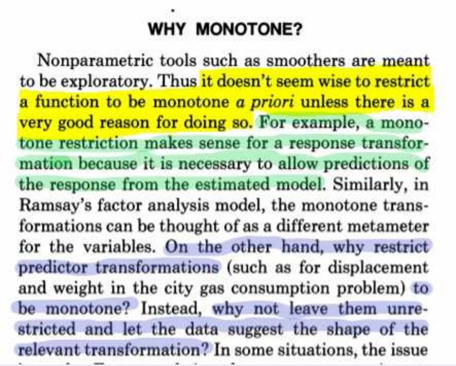
Approximate a monotone function of random variable Y as a monotone function of X: $g(Y) = f(X)$

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

- 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 (fails in the square test...)

Additional suggestions

- Further, this may be too restrictive...
- From "**[Monotone Regression Splines in Action]: Comment**" published in *Statistical Science* by Hastie and Tibshirani (1988):



WHY MONOTONE?

Nonparametric tools such as smoothers are meant to be exploratory. Thus it doesn't seem wise to restrict a function to be monotone *a priori* unless there is a very good reason for doing so. For example, a monotone restriction makes sense for a response transformation because it is necessary to allow predictions of the response from the estimated model. Similarly, in Ramsay's factor analysis model, the monotone transformations can be thought of as a different metameter for the variables. On the other hand, why restrict predictor transformations (such as for displacement and weight in the city gas consumption problem) to be monotone? Instead, why not leave them unrestricted and let the data suggest the shape of the relevant transformation? In some situations, the issue

- 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 $x_2 \geq x_1$:

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

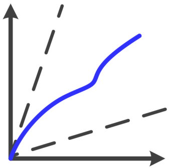
Definition

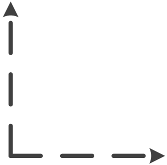
For a given $0 < \kappa < 1$, the semi-κ-monotone maximal correlation measure is defined as

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

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

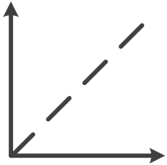
The semi-κ-monotone maximal correlation measure





$\kappa = 0$

Hastie and Tibshirani



$\kappa = 1$

Correlation ratio

The semi-κ-monotone maximal correlation measure

- 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. $r(X \to Y)$ is defined for any pair of random variables X and Y neither of them being constant with probability 1
2. $r(X \to Y)$ may not be equal to $r(Y \to X)$
3. $0 \leq r(X \to Y) \leq 1$
4. $r(X \to Y) = 0$ if and only if X and Y are independent
5. $r(X \to Y) = 1$ 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. $r(X \to Y)$ is defined for any pair of random variables X and Y neither of them being constant with probability 1
2. $r(X \to Y)$ may not be equal to $r(Y \to X)$
3. $0 \leq r(X \to Y) \leq 1$
4. $r(X \to Y) = 0$ if and only if X and Y are independent
5. $r(X \to Y) = 1$ 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

- ✓ $r(X \to Y)$ is defined for any pair of random variables X and Y neither of them being constant with probability 1
- ✓ $r(X \to Y)$ may not be equal to $r(Y \to X)$
- ✓ $0 \leq r(X \to Y) \leq 1$

4. $r(X \to Y) = 0$ if and only if X and Y are independent
5. $r(X \to Y) = 1$ 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

X and Y are independent $\implies r(X \to Y) = 0$

- Since $\rho^{**}_{\max}(X \leftrightarrow Y) \geq \rho^{*m_\kappa}_{\max}(X \to Y)$
- If X, Y are independent $\implies \rho^{*m_\kappa}_{\max}(X \to Y) = 0$ (as so is even $\rho^{**}_{\max}(X \leftrightarrow Y)$)

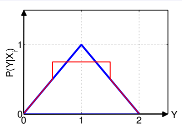
Axiom D: breaking the symmetry

X and Y are dependant $\Longrightarrow r(X \to Y) \neq 0$

- $\rho_{\max}^{*m_\kappa}(X \to Y) \geq \theta(X \to g_{a,\kappa}(y))$
- Focus on the case where $\theta(X \to Y) = 0$ and X, Y are dependent
- $\theta(X \to Y) = 0 \implies \mathbb{E}[Y|X = x] = \int p(y|x)ydy \equiv const$ $\implies \mathrm{var}(\mathbb{E}[Y|X]) = 0$
- We may break the symmetry of $g(y) = y$ by defining, e.g.,

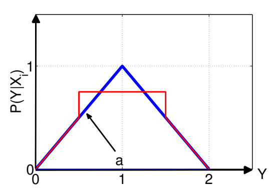
$$g_{a,\kappa}(y) = \begin{cases} y & y \geq a \\ \kappa y & y < a \end{cases}, \qquad \kappa < \mathbf{1} \text{ !!!}$$

- Dependence $\Longrightarrow$ two values x_1 and x_2 such that $p(y|x_1) \not\equiv p(y|x_2)$



Axiom D: breaking the symmetry

X and Y are dependant $\Longrightarrow r(X \to Y) \neq 0$



- Let a be a value such that
$$\int^{a} p(y|x_1) y dy \neq \int^{a} p(y|x_2) y dy$$
- $\Longrightarrow \mathbb{E}[g_a(Y)|X = x_1] \neq \mathbb{E}[g_a(Y)|X = x_2]$
- $\Longrightarrow \theta(X \to g_{a,\kappa}(y)) > 0$

Modified axioms

- ✓ $r(X \to Y)$ is defined for any pair of random variables X and Y neither of them being constant with probability 1
- ✓ $r(X \to Y)$ may not be equal to $r(Y \to X)$
- ✓ $0 \leq r(X \to Y) \leq 1$
- ✓ $r(X \to Y) = 0$ if and only if X and Y are independent

5. $r(X \to Y) = 1$ 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

$Y = f(X) \implies \rho^{*m_\kappa}_{\max}(X \to Y) = 1$

- $Y = f(X)$ (almost surely) $\longrightarrow g(Y) = Y \longrightarrow \rho^{*m_\kappa}_{\max}(X \to Y) = 1$

$\rho^{*m_\kappa}_{\max}(X \to Y) = 1 \implies Y = f(X)$

- For $0 < \kappa < 1$ the supremum is attainable
- $\implies$ there is a *perfect* linear regression between $g(Y)$ and $f'(X)$ (g, f' being maximizing functions of the measure):
    - $\implies g(Y) = af'(X) + b$ where g is an increasing function with slope greater than κ
    - κ is strictly positive $\implies g$ 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)
    - $\implies Y = g^{-1}(af'(X) + b) = f(X)$

Modified axioms

- ✓ $r(X \to Y)$ is defined for any pair of random variables X and Y neither of them being constant with probability 1
- ✓ $r(X \to Y)$ may not be equal to $r(Y \to X)$
- ✓ $0 \leq r(X \to Y) \leq 1$
- ✓ $r(X \to Y) = 0$ if and only if X and Y are independent
- ✓ $r(X \to Y) = 1$ 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

X and Y jointly normal $\longrightarrow r(X,Y) = |\rho(X,Y)|$

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

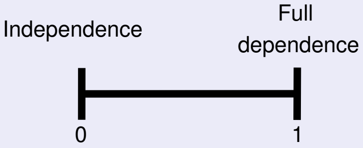
$$\begin{aligned}\rho^{*m_\kappa}_{\max}(X \to Y) &= \rho^{**}_{\max}(X \leftrightarrow Y)\\ &= |\rho|\end{aligned}$$

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 $\mathbf{X} = (X_1, \ldots, X_p)$ 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 $\rho^{*m_0}_{\max}(X \to Y)$, restricting the function applied to the response variable only to be weakly monotone

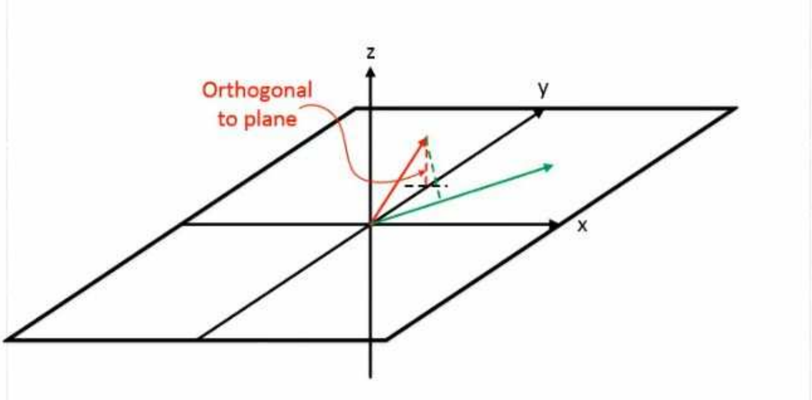
- We do not have an algorithm for computing the semi-κ-monotone maximal correlation measure for $\kappa \neq 0$

- 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 $\Longleftrightarrow$ given $f(X)$, find $g(Y) \in \mathcal{M}_0(Y)$ with smallest distance
- And vice versa...
- $\Longrightarrow$ 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

Calculating semi-0-monotone maximal correlation measure

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

Algorithm 1

```
procedure CALCULATE-SEMI-0-MONOTONE
    Set $g(Y) = Y/\|Y\|$;
    while $e^2(g, f)$ decreases do
        $f'(X) = \mathcal{P}_{\mathcal{H}_2(X)}\left(g(Y)\right)$
        replace $f(X)$ with $f'(X)$
        $g'(Y) = \mathcal{P}_{\mathcal{M}_0(Y)}\left(f(X)\right)$
        replace $g(Y)$ with $g'(Y)/\|g'(Y)\|$
    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

```
procedure REGULARIZED-ACE
    Set $g(Y) = Y/\|Y\|$;
    while $e^2(g, f)$ decreases do
        $f'(X) = \mathcal{P}_{\mathcal{H}_2(X)}(g(Y))$
        replace $f(X)$ with $f'(X)$
        $g'(Y) = \mathcal{P}_{\mathcal{M}_0(Y)}(f(X))$
        replace $g(Y)$ with $g'(Y)/\|g'(Y)\|$
    Apply regularization
    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 ($\kappa = 1$) is insufficient
  - 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:

$$\begin{aligned} Y &\sim \mathcal{U}[0,1] \\ X_1 &= \mathrm{mod}(\mathrm{Y}, 0.2) + \mathrm{N}_1 \\ X_2 &= Y^3 + N_2 \end{aligned}$$

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

Example 1 - multi-variate example

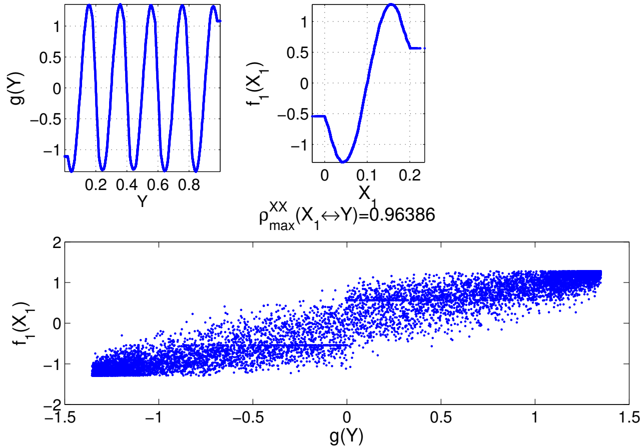


Figure: Example 1: Running ACE on Y and X_1

Example 1 - multi-variate example

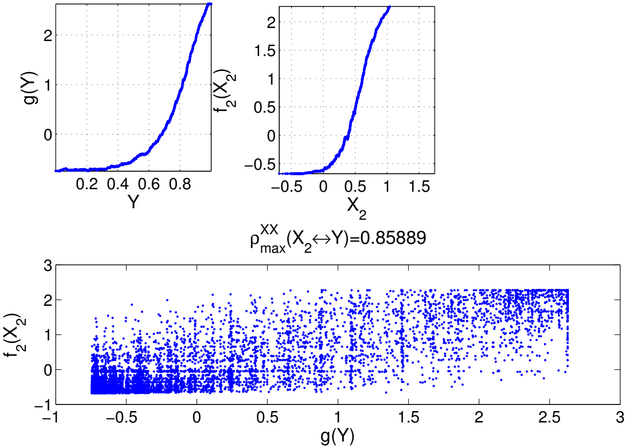


Figure: Example 1: Running ACE on Y and X_2

Example 1 - multi-variate example

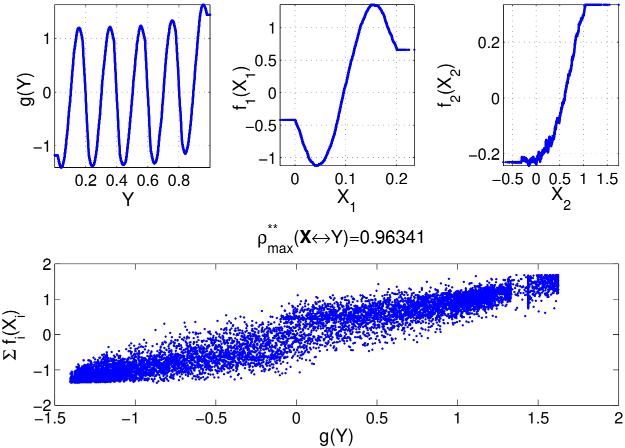


Figure: Example 1: Running ACE on Y, X_1 and X_2

Example 1 - multi-variate example

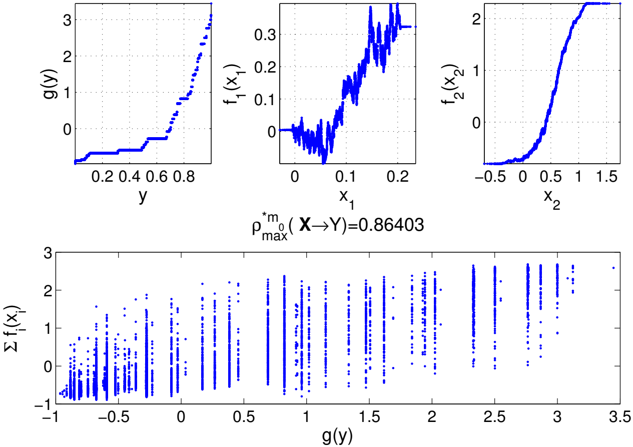


Figure: Example 1: modified ACE (Algorithm 1) on Y, X_1 and X_2 with $\kappa = 0$

Example 2a - correlation ratio fails to detect dependence

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

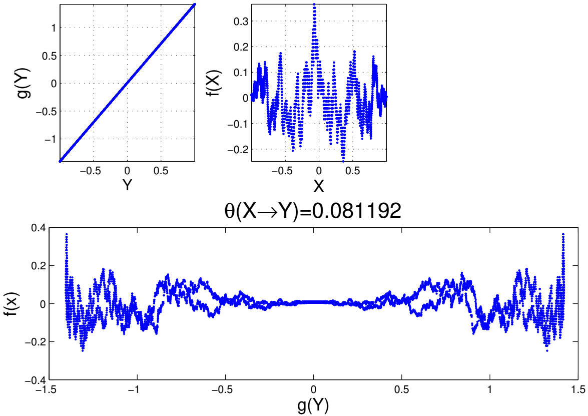


Figure: Transformation corresponding to the correlation ratio

Example 2a - correlation ratio fails to detect dependence

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

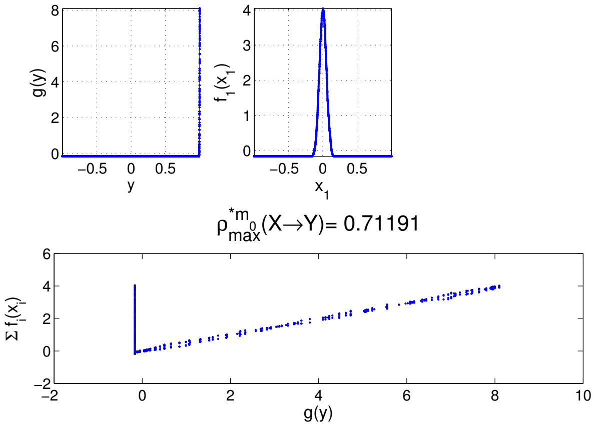


Figure: modified ACE (Algorithm 1) applied to Y and X_1 with $\kappa = 0$

Example 3 - Semi-0-monotonicity is insufficient

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

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

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

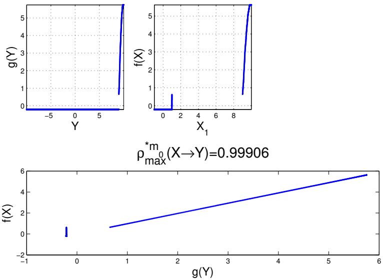


Figure: Modified ACE applied to Y, X with $\kappa = 0$

Example 3 - Semi-0-monotonicity is insufficient

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

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

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

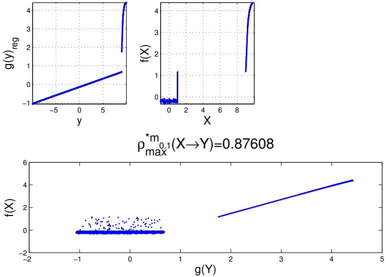


Figure: Regularized ACE applied to Y, X with $\kappa = 0.1$

Thank you for your attention

Backup

Example 2b - preferred parameterization

$$Y \sim \mathcal{U}[0, 10]$$
$$X = log(Y) + Z; \ \ Z \sim \mathcal{N}(0, 1)$$

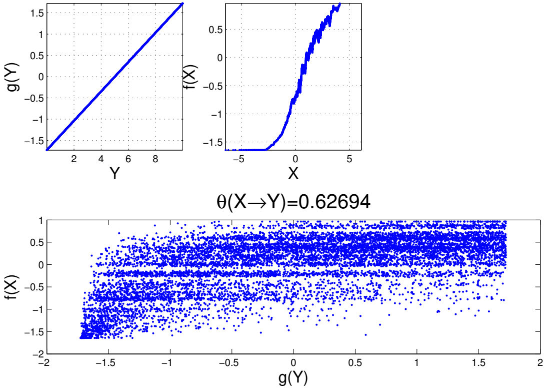


Figure: Transformations corresponding to the correlation ratio

Example 2b - preferred parameterization

$$Y \sim \mathcal{U}[0,10]$$
$$X = log(Y) + Z;\ Z \sim \mathcal{N}(0,1)$$

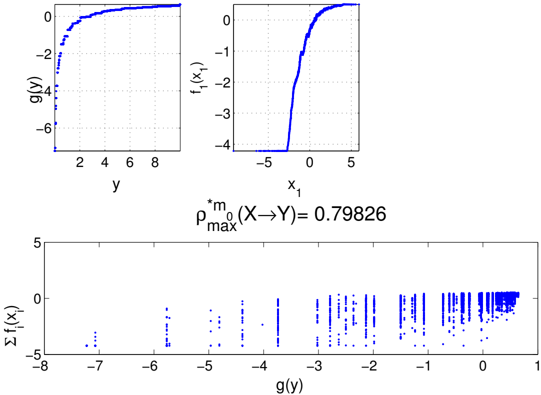


Figure: Modified ACE (Algorithm 1) applied to Y, X and X_2 with $\kappa = 0$

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: $$X \sim \mathcal{U}[0,1]; \quad N \sim \mathcal{U}[0,1]$$ $$Y = \begin{cases} X & 0 < X < \varepsilon \\ N & O/W \end{cases}$$
- $\Longrightarrow$ When $0 < X < \varepsilon$, Y is perfectly known
- Define $$Z = f(X) = \begin{cases} 1 & if\ 0 < X < \varepsilon \\ 0 & O/W \end{cases}$$
- Hence $$\begin{aligned} I(X;Y) &= I(X,Z;Y) \\ &= I(Z;Y) + I(X;Y|Z) \\ &= \underbrace{I(Z;Y) + \Pr(Z=0) I(X;Y|Z=0)}_{>0} \\ &+ \underbrace{\Pr(Z=1)}_{\varepsilon} \underbrace{I(X;Y|Z=1)}_{\infty} \end{aligned}$$

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

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