

Lecture 2: Measures of Correlation and Dependence

Foundations of Data Science:
Algorithms and Mathematical Foundations

Mihai Cucuringu
mihai.cucuringu@stats.ox.ac.uk

CDT in Mathematics of Random System
University of Oxford

Measures of Correlation/Dependence

- ▶ Pearson
- ▶ Spearman
- ▶ Kendall
- ▶ Hoeffding's
- ▶ Maximal Correlation
- ▶ Distance Correlation
- ▶ Mutual Information
- ▶ Maximal Information Coefficient (MIC)

Review of correlation measures

Pearson correlation ρ is a measure of **linear dependence** between variables.

- ▶ In the population: given random variables $X, Y \in \mathbb{R}$

$$\text{Cor}(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)}\sqrt{\text{Var}(Y)}}$$

- ▶ In the sample: given vectors $x, y \in \mathbb{R}$

$$\text{cor}(x, y) = \frac{\text{cov}(x, y)}{\sqrt{\text{var}(x)}\sqrt{\text{var}(y)}} = \frac{(x - \bar{x}\mathbf{1})^T(y - \bar{y}\mathbf{1})}{\|x - \bar{x}\mathbf{1}\|_2\|y - \bar{y}\mathbf{1}\|_2}$$

If x, y are have been *centered*

$$\text{cor}(x, y) = \frac{x^T y}{\|x\|_2\|y\|_2}$$

Properties of the population correlation

$$\rho \stackrel{\text{def}}{=} \text{Cor}(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)}\sqrt{\text{Var}(Y)}}$$

Properties of ρ

- ▶ $\text{Cor}(X, X) = 1$
- ▶ $\text{Cor}(X, Y) = \text{Cor}(Y, X)$
- ▶ $\text{Cor}(aX + b, Y) = \text{sign}(a)\text{Cor}(X, Y)$ for any $a, b \in \mathbb{R}$
- ▶ $-1 \leq \text{Cor}(X, Y) \leq 1$
- ▶ $|\text{Cor}(X, Y)| = 1$ if and only if $Y = aX + b$ for some $a, b \in \mathbb{R}$, with $a \neq 0$
- ▶ If X, Y are independent then $\text{Cor}(X, Y) = 0$
- ▶ If $\text{Cor}(X, Y) = 0$ then X, Y need not be independent!!!
- ▶ If (X, Y) is bivariate normal and $\text{Cor}(X, Y) = 0$, then X, Y are independent

Bivariate normal distribution

Two-dimensional Gaussian distribution

The random vector $Z = (X, Y) \in \mathbb{R}^2$ has a bivariate normal dist.

$$Z \sim N(\mu, \Sigma),$$

$\mu \in \mathbb{R}^2$ is the mean & $\Sigma \in \mathbb{R}^{2 \times 2}$ is the covariance matrix

$$\mu = \begin{bmatrix} \mu_X \\ \mu_Y \end{bmatrix} \quad \Sigma = \begin{bmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{bmatrix} \quad (1)$$

where $E[X] = \mu_X$; $E[Y] = \mu_Y$; $\text{Var}(X) = \sigma_X^2$; $\text{Var}(Y) = \sigma_Y^2$;

$$\text{Cov}(X, Y) = \rho\sigma_X\sigma_Y$$

$$\text{Cor}(X, Y) = \rho$$

- ▶ The probability density function of $Z = (X, Y)$ is given by

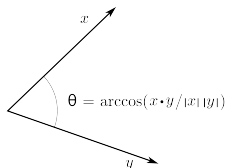
$$f_{X,Y}(z) = \frac{1}{2\pi\sqrt{\det(\Sigma)}} e^{-\frac{1}{2}(z-\mu)^T \Sigma^{-1}(z-\mu)}$$

- ▶ Fact: $\rho = 0$ implies that X and Y are independent rv

Review: properties of sample correlation

$$\text{cor}(x, y) = \frac{\text{cov}(x, y)}{\sqrt{\text{var}(x)}\sqrt{\text{var}(y)}} = \frac{(x - \bar{x}\mathbf{1})^T(y - \bar{y}\mathbf{1})}{\|x - \bar{x}\mathbf{1}\|_2\|y - \bar{y}\mathbf{1}\|_2}$$

- ▶ $\text{cor}(x, x) = 1$
- ▶ $\text{cor}(x, y) = \text{cor}(y, x)$
- ▶ $\text{cor}(ax + b, y) = \text{sign}(a)\text{cor}(x, y)$ for any $a, b \in \mathbb{R}$
- ▶ $-1 \leq \text{cor}(x, y) \leq 1$
- ▶ $|\text{cor}(x, y)| = 1$ iff $y = ax + b$ for some $a, b \in \mathbb{R}$ with $a \neq 0$
- ▶ $\text{cor}(x, y) = 0$ iff x, y are orthogonal
- ▶ If x, y are centered then $\text{cor}(x, y) = \cos \theta$, where θ is the angle between the vectors $x, y \in \mathbb{R}^n$
- ▶ $\cos \theta = \frac{x^T y}{\|x\|_2\|y\|_2}$



Drawbacks of the Pearson Correlation

- ▶ by far the most popular tool in practice for understanding bivariate relationships
- ▶ easy calculation and interpretability means

Pearson ρ is not a useful measure of dependency overall:

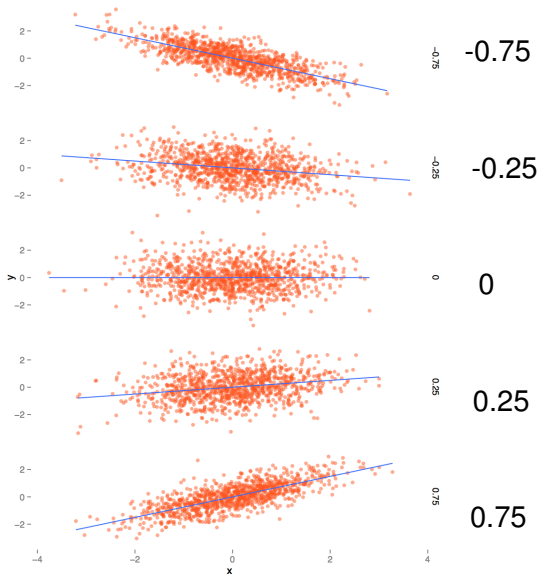
- ▶ does not guarantee a causal relationship
- ▶ a lack of correlation does not even mean there is no relationship between two variables!
- ▶ best suited to continuous, normally distributed data
- ▶ it is easily corrupted by outliers
- ▶ only a measure of linear dependency (Note: most data out there is nonlinear by nature)

R code for data generation and computing the Pearson correlation

```
library(MASS)
cormat = matrix(c(1, 0.25, 0.25, 1), ncol = 2) #.25 population correlation
set.seed(1234)
# empirical argument will reproduce the correlation exactly if TRUE
mydat = mvrnorm(100, mu = c(0, 0), Sigma = cormat, empirical = T)
cor(mydat)

##      [,1] [,2]
## [1,] 1.00 0.25
## [2,] 0.25 1.00
```


Correlation patterns with the regression line imposed



Spearman/Rank correlation

- ▶ defined in the sample
- ▶ goes beyond measuring linearity between $x, y \in \mathbb{R}^n$
- ▶ measures a monotone association between $x, y \in \mathbb{R}^n$
- ▶ given vectors $x, y \in \mathbb{R}^n$, define the rank vector $r_x \in \mathbb{R}^n$ that ranks the components of x

$$r_x(j) = k$$

if x_j is the k^{th} smallest element in x

- ▶ Example: if $x = (0.7, 0.1, 0.5, 1)$ then $r_x = (3, 1, 2, 4)$
- ▶ Similarly define the ranks r_y corresponding to y
- ▶ Rank correlation is given by the (sample) correlation of r_x and r_y

$$rcor(x, y) = cor(r_x, r_y)$$

- ▶ Remark: $|rcor(x, y)| = 1$ if and only if there is a monotone function $f : \mathbb{R} \mapsto \mathbb{R}$ such that $y_i = f(x_i)$ for each $i = 1, \dots, n$

Spearman correlation - R code

```
cor(mydat, method = "spearman") #slight difference

##           [,1]    [,2]
## [1,] 1.0000 0.1919
## [2,] 0.1919 1.0000

cor(rank(mydat[, 1]), rank(mydat[, 2]))

## [1] 0.1919
```

Kendall's τ correlation

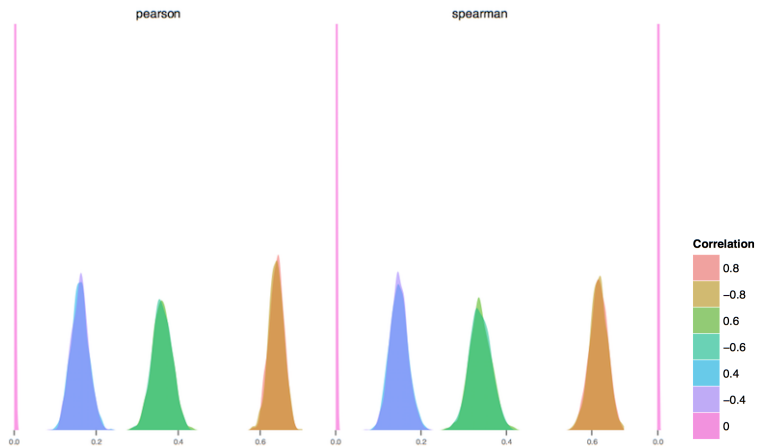
Kendall τ rank correlation coefficient:

- ▶ alternative to Spearman; identifies monotonic relationships

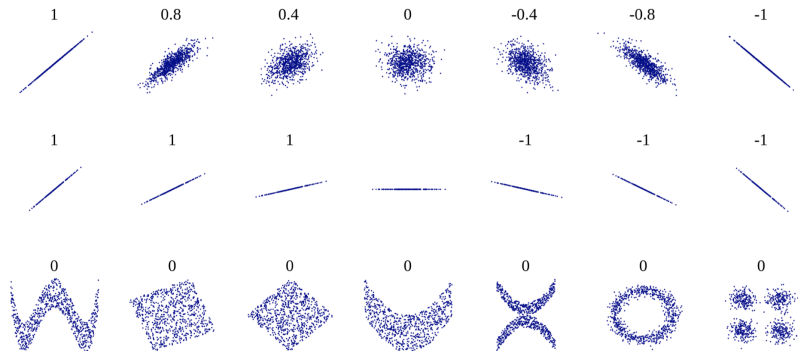
$$\tau(X, Y) = \frac{\# \text{ of concordant pairs} - \# \text{ of discordant pairs}}{\frac{n(n-1)}{2}}$$

- ▶ *concordant* means if the ranks for both elements agree: if both $x_i > x_j$ and $y_i > y_j$ or if both $x_i < x_j$ and $y_i < y_j$
- ▶ otherwise, *discordant*
- ▶ $\tau \in [-1, 1]$, same interpretations as for Spearman's correlation
- ▶ typically used in ranking problems in ML (Lecture 7)
- ▶ R: `cor.test` with parameter `method = 'kendall'` (package `stats`)

Linear Relationships



Pearson Correlation



Nonlinear patterns

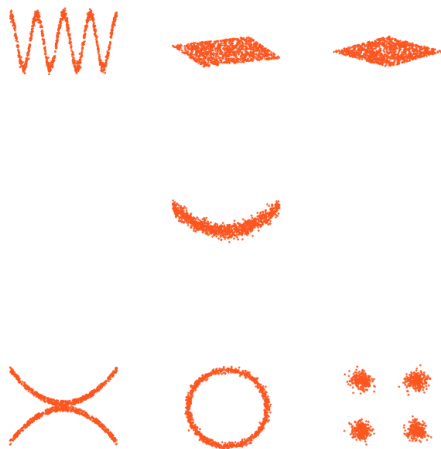


Figure: The patterns will be referred to as TOP: wave, trapezoid, diamond; MIDDLE: quadratic; BOTTOM: X, circle and cluster.

Nonlinear patterns + Noise

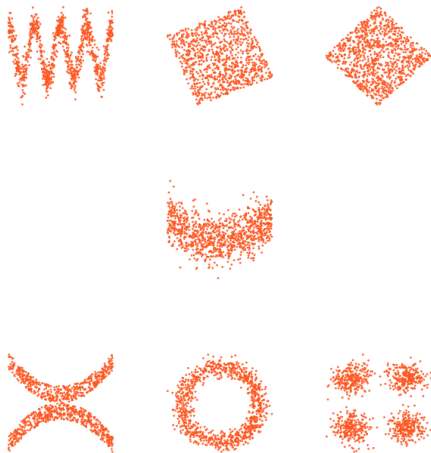


Figure: The patterns will be referred to as TOP: wave, trapezoid, diamond; MIDDLE: quadratic; BOTTOM: X, circle and cluster.

Demand for electricity

- ▶ driven by weather conditions and especially temperature.
- ▶ one could predict the demand for electricity as a function of the temperature
- ▶ weather dynamics could included in pricing (for derivative instruments)

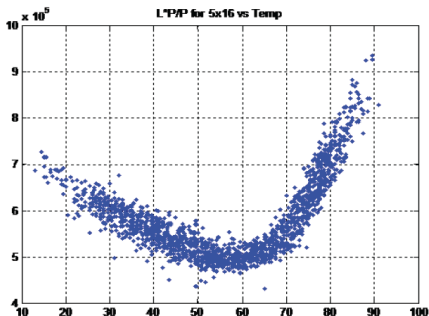


Figure: Source: Carmona, Chapter on Commodity Prices

Hoeffding's-D

- ▶ Hoeffding's D is another rank based since the 1940s
Hoeffding (1948). A non-parametric test of independence
- ▶ a measure of the distance between $F(x, y)$ and $G(x)H(y)$, where $F(x, y)$ is the joint CDF of X and Y , and G and H are marginal CDFs

$$D = \int (F - GH) dF$$

- ▶ measures the difference between the joint ranks of (X, Y) and the product of their marginal ranks
- ▶ it can pick up on nonlinear/non-monotonic relationships
- ▶ lies on the interval $[-.5, 1]$
- ▶ positive/negative signs have no interpretation (D identifies non-monotonic relationships also)
- ▶ the larger the value of D, the more dependent are X and Y (for many types of dependencies)

Maximal correlation

- ▶ a notion of population correlation
- ▶ it has no preference for linearity or monotonicity
- ▶ it characterizes independence completely
- ▶ given two random variables $X, Y \in \mathbb{R}$, the maximal correlation between X, Y is defined as

$$\text{mCor}(X, Y) = \max_{f, g} \text{Cor}(f(X), g(Y)) \quad (2)$$

where the maximum is taken over all functions $f, g : \mathbb{R} \mapsto \mathbb{R}$, with $\text{Var}(f(X)) > 0$ and $\text{Var}(g(Y)) > 0$.

- ▶ Note that $0 \leq \text{mCor}(X, Y) \leq 1$
- ▶ Key property: $\text{mCor}(X, Y) = 0$ if and only if X and Y are independent

Independence of random variables

- ▶ A pair of random variables $X, Y \in \mathbb{R}$ are called independent if for any sets $A, B \subseteq \mathbb{R}$

$$P(X \in A, Y \in B) = P(X \in A)P(Y \in B)$$

- ▶ If f_X , respectively f_Y , denotes the density of X , respectively Y , and (X, Y) has joint density $f_{X,Y}$, independence implies

$$f_{X,Y}(x, y) = f_X(x)f_Y(y), \forall x, y \in \mathbb{R}$$

- ▶ *The joint is the product of the marginals densities*
- ▶ Note: if X, Y are independent, then for any functions f, g it holds true that

$$E[f(X)g(Y)] = E[f(X)]E[g(Y)]$$

- ▶ X, Y being independent implies that $\text{Cor}(f(X), g(Y)) = 0$ for any functions f, g , and thus $\text{mCor}(X, Y) = 0$

Remark: zero mCor implies independence (non-trivial)

The characteristic function

For a random variable $X \in \mathbb{R}$, its characteristic function is defined as

$$\phi_X(t) = \mathbb{E}[e^{itX}]$$

For a pair of random variable $X, Y \in \mathbb{R}$, its joint characteristic function is defined as

$$\phi_{X,Y}(t, s) = \mathbb{E}[e^{i(tX+sY)}]$$

(1) Characteristic functions completely characterize the distribution of a random variable

$$\phi_X(t) = \phi_Y(t), \forall t \in \mathbb{R} \iff X, Y \text{ have the same distribution}$$

(2) X and Y are independent $\iff \phi_{X,Y}(t, s) = \phi_X(t)\phi_Y(s)$

(3) If $a, b \in \mathbb{R}$, then $Z = aX + b$ has characteristic function

$$\phi_Z(t) = e^{ibt} \phi_X(t)$$

Characteristic function - examples

Let $X \sim \text{Bernoulli}(p)$

$$\phi_X(t) = \mathbb{E}[e^{itX}] \quad (3)$$

$$= \mathbb{E}[\cos(tX)] + i\mathbb{E}[\sin(tX)] \quad (4)$$

$$= p \cos(t \cdot 1) + (1 - p) \cos(t \cdot 0) \\ + i(p \sin(t \cdot 1) + (1 - p) \sin(t \cdot 0)) \quad (5)$$

$$= p \cos(t) + (1 - p) + ip \sin(t) \quad (6)$$

$$= (1 - p) + p(\cos(t) + i \sin(t)) \quad (7)$$

$$= (1 - p) + pe^{it} \quad (8)$$

Let $X \sim \text{Exponential}(\lambda)$

$$\phi_X(t) = \frac{\lambda}{\lambda - it}$$

(homework exercise)

"mCor = 0" \implies independence

Assuming mCor = 0

- ▶ let $f_X(t) = e^{itX}$
- ▶ let $g_Y(s) = e^{isY}$
- ▶ mCor = 0 implies

$$\text{Cor}(e^{itX}, e^{isY}) = 0$$

- ▶ which implies

$$\text{Cov}(e^{itX}, e^{isY}) = 0$$

- ▶ using $\text{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$ yields

$$\mathbb{E}[e^{itX} e^{isY}] = \mathbb{E}[e^{itX}] \mathbb{E}[e^{isY}]$$

$$\mathbb{E}[e^{itX+isY}] = \mathbb{E}[e^{itX}] \mathbb{E}[e^{isY}]$$

$$\phi_{X,Y}(t, s) = \phi_X(t)\phi_Y(s)$$

- ▶ which concludes the proof that X and Y are independent.

Maximal correlation in R

- ▶ algorithm to compute $mCor$ in the population
- ▶ fixed points of maximal correlation
- ▶ *Alternating Conditional Expectations*(ACE)
- ▶ adapt the ACE algorithm in the sample

"ace" package in R:

- ▶ $q = ace(x,y)$
- ▶ maximal correlation = $cor(q\$tx, q\$ty)$

● **MAC: Multivariate Maximal Correlation Analysis**, Nguyen et al, ICML 2014

- ▶ genes reveal only a weak correlation with a disease if each gene is considered individually,
- ▶ but, when considered as a group of genes the correlation may be very strong
- ▶ pairwise correlation measures are not sufficient as they are unable to detect complex interactions of a group of genes.

Distance correlation

- ▶ a very recent measure of statistical dependence between two random variables
- ▶ also works for two random vectors of not necessarily equal dimension
- ▶ it characterizes independence completely
- ▶ *Measuring and testing dependence by correlation of distances*, Gabor J. Szekely, Maria L. Rizzo, and Nail K. Bakirov, *Annals of Statistics*, Volume 35, Number 6 (2007), 2769-2794
- ▶ well-defined in both the population and in the sample
- ▶ very computationally easy to calculate

Distance correlation

- ▶ properties of a true dependence measure, like Pearson ρ
- ▶ distance correlation satisfies $0 \leq R \leq 1$, and $R = 0$ only if X and Y are independent
- ▶ In the bivariate normal case, $R \leq |\rho|$ and equals one when $\rho \pm 1$
- ▶ Note that one can obtain a dCor value for X and Y of arbitrary dimension
- ▶ can be computed using the **dcor** function in the **energy R** package
- ▶ one could also incorporate a rank-based version of this metric as well

Distance correlation - sample version

Let $(x_i, y_i), i = 1, \dots, n$ denote a sample from a pair of real/vector-valued r.v. (X, Y)

- ▶ define the distance matrices $A, B \in \mathbb{R}^{n \times n}$ as

$$A_{ij} = |x_i - x_j| \quad \text{and} \quad B_{ij} = |y_i - y_j|, \quad i, j = 1, \dots, n$$

- ▶ for higher (possibly different) dimensions

$$A_{ij} = \|x_i - x_j\|_F \quad \text{and} \quad B_{ij} = \|y_i - y_j\|_F, \quad i, j = 1, \dots, n$$

- ▶ (x, y) could be of different dimensions: $n \times d_1, n \times d_2$)
- ▶ double center the distance matrices A, B to get \tilde{A}, \tilde{B}

$$\tilde{A}_{ij} = A_{ij} - \bar{A}_{i.} - \bar{A}_{.j} + \bar{A}_{..}$$

- ▶ $\bar{A}_{i.}$ denotes the mean of row i
- ▶ $\bar{A}_{.j}$ denotes the mean of column j
- ▶ $\bar{A}_{..}$ denotes the overall mean of A
- ▶ this centers both the rows and columns of A, B
- ▶ all rows and columns of \tilde{A} and \tilde{B} sum to 0

Distance correlation - sample version

- ▶ in short notation

$$\tilde{A}_{jk} = (I - M)A(I - M) \quad \text{and} \quad \tilde{B}_{jk} = (I - M)B(I - M)$$

$$\text{where } M = \frac{1}{n} \mathbf{1}\mathbf{1}^T$$

The distance covariance of x, y is defined as the square root of

$$\text{dcov}^2(x, y) \stackrel{\text{def}}{=} \frac{1}{n^2} \sum_{i,j=1}^n \tilde{A}_{ij} \tilde{B}_{ij}$$

The distance variance is defined as

$$\text{dvar}^2(x) \stackrel{\text{def}}{=} \text{dcov}^2(x, x)$$

The **distance correlation** of the sample is given by

$$\text{dcor}^2(x, y) \stackrel{\text{def}}{=} \frac{\text{dcov}^2(x, y)}{\sqrt{\text{dvar}^2(x)} \sqrt{\text{dvar}^2(y)}}$$

Distance correlation - sample version

Properties:

- ▶ $\text{dcor}(ax + b, y) = \text{dcor}(x, y), \forall a, b \in \mathbb{R}, a \neq 0$
- ▶ $0 \leq \text{dcor}(x, y) \leq 1$
- ▶ $\text{dcor}(x, y) = 0 \iff y = ax + b$ for some $a, b \in \mathbb{R}, a \neq 0$

Distance correlation - sample vs population

$$\text{dcov}^2(x, y) = \frac{1}{n^2} \sum_{i,j=1}^n \tilde{A}_{ij} \tilde{B}_{ij}$$

- ▶ one can show the following holds true

$$\text{dcov}^2(x, y) = \frac{1}{n^2} \sum_{i,j=1}^n A_{ij} B_{ij} - \frac{1}{n} \sum_{j=1}^n A_{.j} B_{.j} - \frac{1}{n} \sum_{i=1}^n A_{i.} B_{i.} + A_{..} B_{..}$$

- ▶ where

- ▶ $A_{i.} = \sum_{j=1}^n A_{ij}$

- ▶ $A_{.j} = \sum_{i=1}^n A_{ij}$

- ▶ $A_{..} = \sum_{i,j=1}^n A_{ij}$ (and similarly for B)

Compare to the (population) distance covariance

$$\begin{aligned} dCov^2(X, Y) \stackrel{\text{def}}{=} & \mathbb{E}[|X - X'| | Y - Y'|] + \mathbb{E}[|X - X'|] \mathbb{E}[|Y - Y'|] \\ & - \mathbb{E}[|X - X'| | Y - Y''] - \mathbb{E}[|X - X''| | Y - Y'|] \end{aligned}$$

Distance correlation - sample vs population

Properties:

$$\text{dcov}^2(x, y) = \frac{1}{n^2} \sum_{i,j=1}^n \tilde{A}_{ij} \tilde{B}_{ij}$$

- ▶ one can show the following holds true

$$\text{dcov}^2(x, y) = \frac{1}{n^2} \sum_{i,j=1}^n A_{ij} B_{ij} - \frac{1}{n} \sum_{j=1}^n A_{.j} B_{.j} - \frac{1}{n} \sum_{i=1}^n A_i. B_i. + A_{..} B_{..}$$

Compare to the (population) distance covariance

$$\begin{aligned} dCov^2(X, Y) \stackrel{\text{def}}{=} & \mathbb{E}[|X - X'| | Y - Y'|] + \mathbb{E}[|X - X'|] \mathbb{E}[|Y - Y'|] \\ & - \mathbb{E}[|X - X'| | Y - Y''] - \mathbb{E}[|X - X''| | Y - Y'|] \end{aligned}$$

Distance correlation via characteristic functions

Recall:

For a r.v. $X \in \mathbb{R}$, its characteristic function:

$$\phi_X(t) = \mathbb{E}[e^{itX}]$$

For a pair of r.v.'s $X, Y \in \mathbb{R}$, their joint characteristic function:

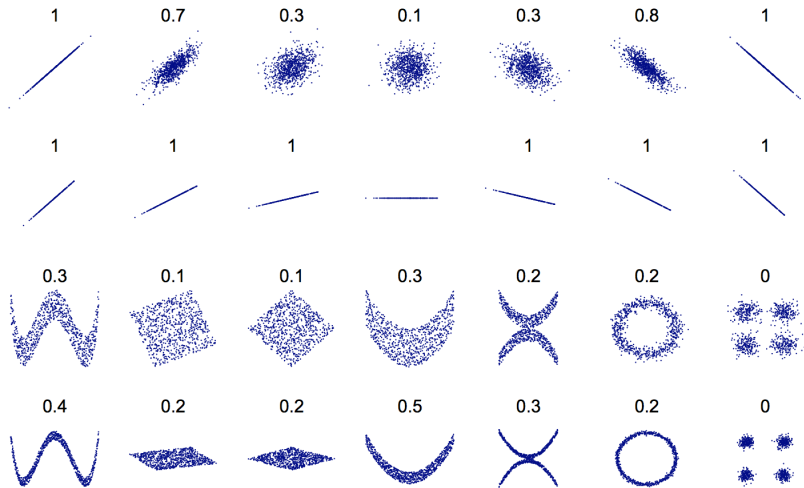
$$\phi_{X,Y}(t, s) = \mathbb{E}[e^{i(tX+sY)}]$$

The initial motivation for dCov:

$$\text{dCov}(X, Y) = \|\phi_{X,Y} - \phi_X\phi_Y\|$$

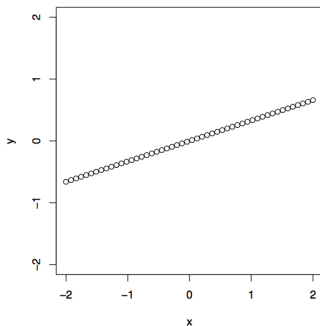
- ▶ $\|\cdot\|$ is a certain norm on functions
- ▶ $\text{dCov}(X, Y) = \|\phi_{X,Y} - \phi_X\phi_Y\| = 0 \iff \phi_{X,Y}(t, s) = \phi_X(t)\phi_Y(s), \forall s, t \in \mathbb{R} \iff X \perp Y$

Distance Correlation (Wiki)

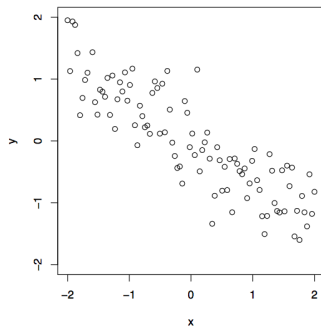


Linear relationship - clean and noisy

Perfect linear



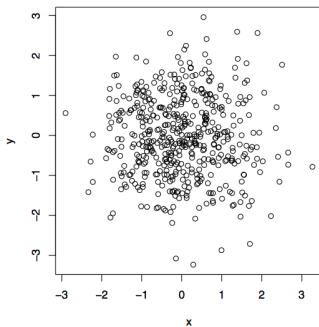
Noisy linear



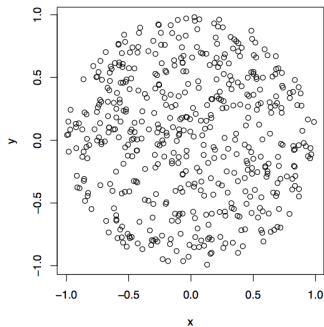
dcor =	1.000	0.867
mcor =	1.000	0.896
rcor =	1.000	-0.872
cor =	1.000	-0.866

Comparison of correlation measures

Independent



Ball

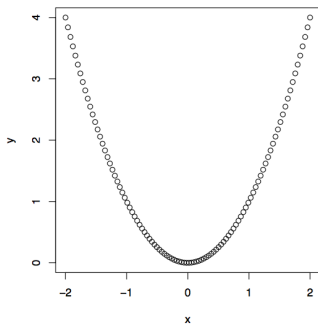


dcor =	0.078
mcor =	0.124
rcor =	-0.021
cor =	-0.023

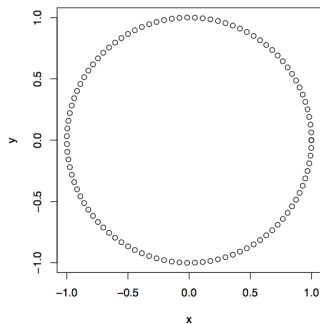
dcor =	0.099
mcor =	0.316
rcor =	-0.033
cor =	-0.029

Comparison of correlation measures

Perfect quadratic

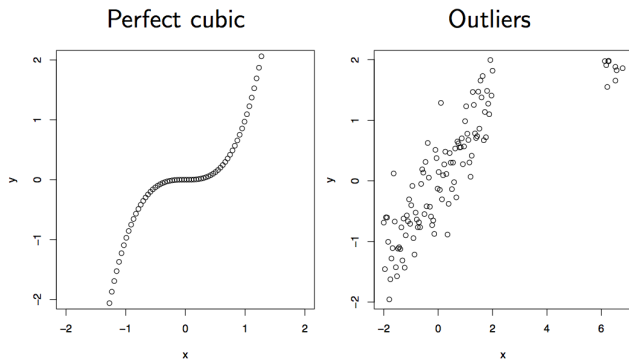


Perfect circle



dcor =	0.492	0.200
mcor =	1.000	1.000
rcor =	0.013	-0.001
cor =	0.000	0.000

Comparison of correlation measures



dcor =	0.920	0.854
mcor =	1.000	0.913
rcor =	1.000	0.905
cor =	0.920	0.834

Linear Relationships

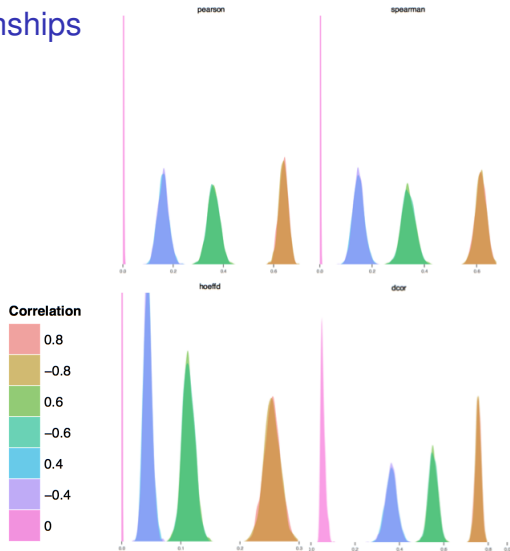


Figure: Pearson (Top Left), Spearman (Top Right), Hoeffding's (Bottom Left), DistCor (Bottom Right)

Clark, M., *A comparison of correlation measures* (2013). For each correlation in the legend, 1000 x,y data sets are created, each of length 1000. Squared values for Pearson and Spearman.

Nonlinear patterns

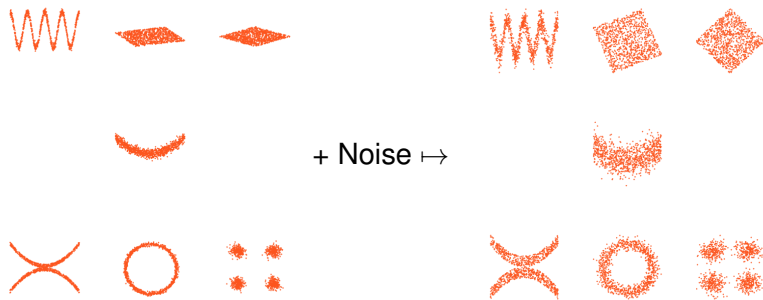


Figure: The patterns will be referred to as TOP: wave, trapezoid, diamond; MIDDLE: quadratic; BOTTOM: X, circle and cluster. Noisy versions on the left.

Nonlinear Relationships: Pearson & Spearman

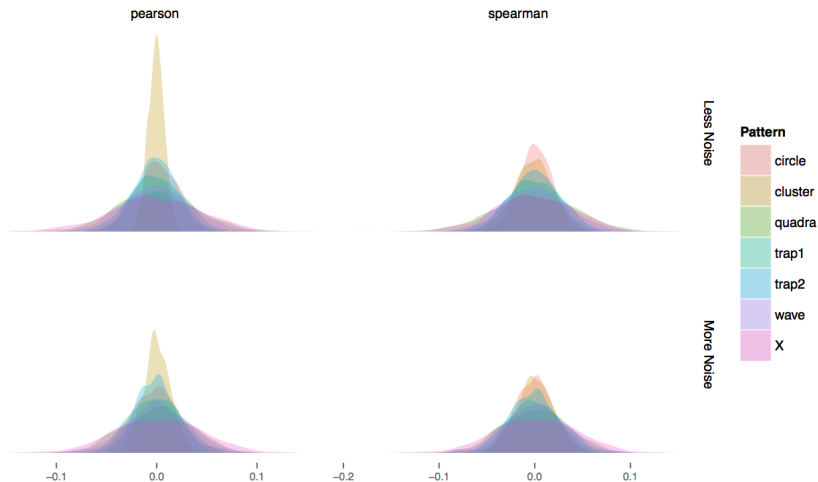
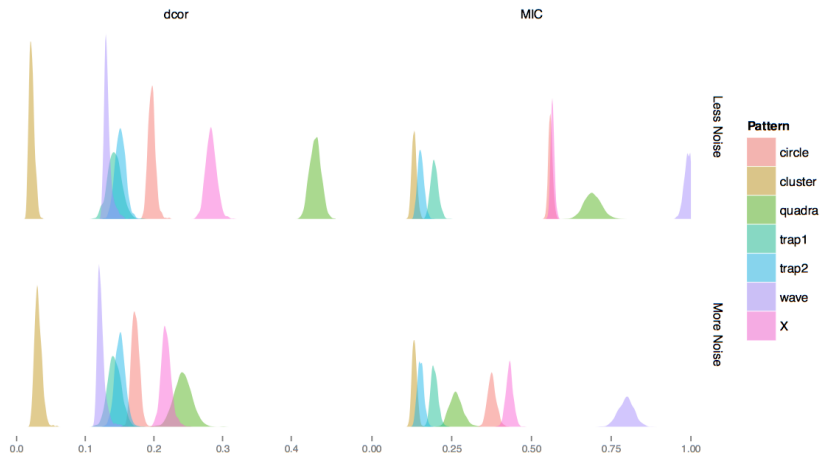


Figure: Neither Pearson nor Spearman are able to find a relationship among any of the patterns regardless of noise level.

dcor vs. MIC: nonlinear patterns

Maximal Information Coefficient (MIC):

- ▶ regarded as a 'correlation for the 21st century'
- ▶ based on concepts from information theory



Both dCor & MIC show significant values for most patterns.

Mutual Information and the Maximal Information Coefficient (MIC)

- ▶ regarded as a 'correlation for the 21st century'
- ▶ based on concepts from information theory
- ▶ entropy as a measure of uncertainty of random variable

The entropy (measured in bits) of a discrete random variable X , with probability mass function $p(x) = P(X = x)$, is given by

$$H(X) = - \sum_{x \in \mathcal{X}} p(x) \log_2 p(x)$$

Equivalent expression

$$H(X) = \mathbb{E}_p \left(\log \frac{1}{p(x)} \right)$$

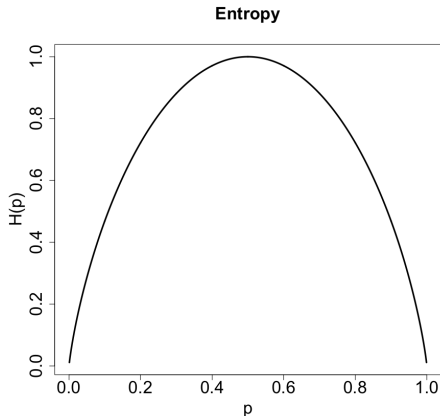
Remark 1: $H(X) \geq 0$

Remark 2: $H_b(X) = \log_b a H_a(X)$

Example: $X \sim \text{Bernoulli}(p)$

$$\begin{cases} 1, \text{ w. p. } p \\ 0, \text{ w. p. } 1 - p \end{cases}$$

$$H(X) = -p \log p - (1 - p) \log(1 - p) := H(p)$$



Recap joint and marginal distributions

Given (X, Y) a pair of discrete r.v.'s taking values in \mathcal{X} and \mathcal{Y}

- ▶ the joint distribution of X and Y is given by

$$p(x, y) = P(X = x, Y = y), x \in \mathcal{X}, y \in \mathcal{Y}$$

- ▶ the (marginal) distribution of X

$$p_X(x) = p(x) = P(X = x) = \sum_{y \in \mathcal{Y}} p(x, y)$$

- ▶ the (marginal) distribution of Y

$$p_Y(y) = p(y) = P(Y = y) = \sum_{x \in \mathcal{X}} p(x, y)$$

Joint Entropy

The joint entropy $H(X, Y)$ of a pair of discrete random variables $(X, Y) \sim p(x, y)$ is given by

$$H(X, Y) = - \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log p(x, y)$$

also written as

$$H(X, Y) = -\mathbb{E}(\log p(X, Y))$$

It measures the uncertainty associated to (X, Y) .

Conditional Entropy

Given $(X, Y) \sim p(x, y)$, define the conditional entropy $H(Y|X)$

$$H(Y|X) \stackrel{\text{def}}{=} \sum_{x \in \mathcal{X}} p(x) H(Y|X = x)$$

where

$$H(Y|X = x) \stackrel{\text{def}}{=} \sum_{y \in \mathcal{Y}} p(y|x) \log p(y|x)$$

and the conditional probability is given by

$$p(y|x) \stackrel{\text{def}}{=} P(Y = y|X = x) = \frac{P(Y = y, X = x)}{P(X = x)} = \frac{p(x, y)}{p(x)}$$

Interpretation: $H(Y|X)$ measures the amount of uncertainty remaining about Y after X is known.

Conditional Entropy

$$\begin{aligned} H(Y|X = x) &= \sum_{x \in \mathcal{X}} p(x) H(Y|X = x) \\ &= \sum_{x \in \mathcal{X}} p(x) \sum_{y \in \mathcal{Y}} p(y|x) \log p(y|x) \\ &= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log p(y|x) \\ &= -\mathbb{E}(\log p(Y|X)) \end{aligned}$$

Using this, one can show that

$$H(X, Y) = H(X) + H(Y|X)$$

Relative Entropy

The relative entropy or Kullback-Leibler distance between $p(x)$ and $q(x)$

$$D(p||q) = \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)} = \mathbb{E}_p \left(\log \frac{p(x)}{q(x)} \right)$$

- ▶ not a proper distance
 - ▶ does not satisfy the triangle inequality
 - ▶ not symmetric
- ▶ a measure of the distance between the two distributions $p(x)$ and $q(x)$

Mutual Information

Given mass functions

- ▶ $X \sim p(x), Y \sim p(y)$
- ▶ $(X, Y) \sim p(x, y)$

The mutual information $I(X; Y)$

- ▶ measure of the variables mutual dependence
- ▶ is the relative entropy between $p(x, y)$ and $p(x)p(y)$

$$\begin{aligned} I(X; Y) &= D(p(x, y) \parallel p(x)p(y)) \\ &= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log \frac{p(x, y)}{p(x)p(y)} \\ &= \mathbb{E}_{p(x, y)} \left(\log \frac{p(X, Y)}{p(X)p(Y)} \right) \end{aligned}$$

Note: $D(p \parallel q) \neq D(q \parallel p)$

Interpretation: $I(X; Y)$ measures the

- ▶ the information that X and Y share
- ▶ average reduction in uncertainty on X that results from knowing Y

Mutual Information

- ▶ measures how much one random variable tells us about another
- ▶ $MI(X, Y) \geq 0$
- ▶ High MI indicates a large reduction in uncertainty
- ▶ Low MI indicates a small reduction in uncertainty
- ▶ $MI(X, Y) = 0 \iff (X, Y)$ are independent

- ▶ various algorithms to estimate MI
- ▶ for discrete data, the density functions $p(x)$, $p(y)$, and $p(x, y)$ can be estimated by simply counting the events

- ▶ R function: *mi.empirical* (package *entropy*)

Mutual Information and entropy

Properties (try to prove on your own):

- ▶ $I(X; Y) = H(X) - H(Y|X)$
- ▶ $I(X; Y) = H(Y) - H(X|Y)$
- ▶ $I(X; Y) = H(X) + H(Y) - H(X, Y)$
- ▶ $I(X; Y) = I(Y; X)$
- ▶ $I(X) = H(X)$

Jensen's inequality: If f is a convex function and X is a random variable

$$\mathbb{E}(f(X)) \geq f(\mathbb{E}(X))$$

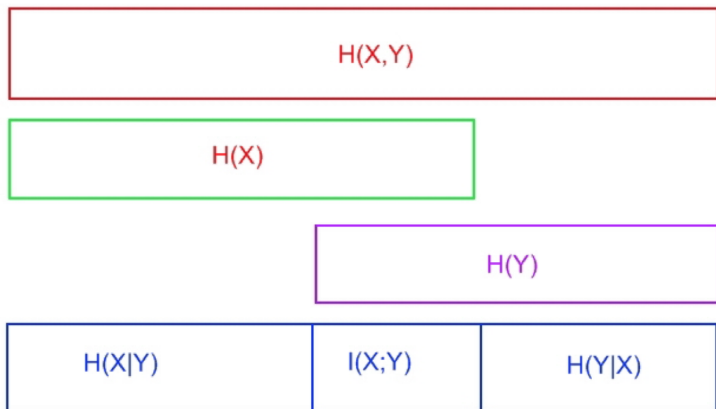


Figure: The relation between entropy and mutual information.

Implications of Jensen's inequality

- ▶ Information/Gibbs inequality inequality:

$$D(p||q) \geq 0$$

with equality if and only if $p(x) = q(x), \forall x \in \mathcal{X}$

- ▶ Corollary:

$$I(X; Y) \geq 0$$

with equality if and only if X and Y are independent

- ▶ Fun fact:

$$H(X) \leq \log |\mathcal{X}|$$

with equality if and only if $X \sim \text{Unif}(\mathcal{X})$ (exercise).

- ▶ Conditioning reduce entropy (additional information cannot hurt)

$$H(X|Y) \leq H(X)$$

(follows from earlier properties)

Motivation for MIC

- ▶ Determine important undiscovered relationships in data sets with lots of variables
 - ▶ move beyond linear and monotonic relationships
- ▶ Have a computationally efficient algorithm that robustly identifies the important relationships

Why is the Pearson correlation not enough?

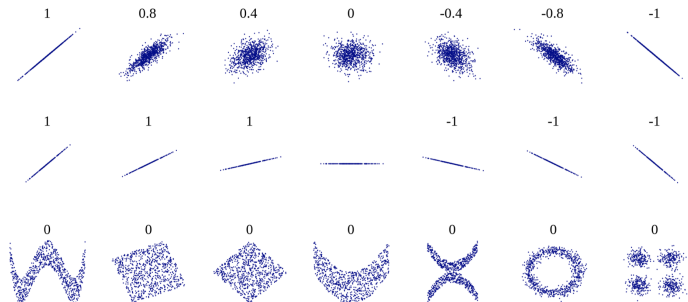


Figure: Pearson is not a viable choice for understanding many dependencies that are ubiquitous in modern data sets.

RESEARCH ARTICLES

Detecting Novel Associations in Large Data Sets

David N. Reshes,^{1,2,3*†} Yakir A. Reshes,^{2,4**†} Hilary K. Finucane,⁵ Sharon R. Grossman,^{2,6} Gilean McVean,^{3,7} Peter J. Turnbaugh,⁶ Eric S. Lander,^{2,8,9} Michael Mitzenmacher,^{10‡} Pardis C. Sabeti^{2,6‡}

Identifying interesting relationships between pairs of variables in large data sets is increasingly important. Here, we present a measure of dependence for two-variable relationships: the maximal information coefficient (MIC). MIC captures a wide range of associations both functional and not, and for functional relationships provides a score that roughly equals the coefficient of determination (R^2) of the data relative to the regression function. MIC belongs to a larger class of maximal information-based nonparametric exploration (MINE) statistics for identifying and classifying relationships. We apply MIC and MINE to data sets in global health, gene expression, major-league baseball, and the human gut microbiota and identify known and novel relationships.

Imagine a data set with hundreds of variables, which may contain important, undiscovered relationships. There are tens of thousands of variable pairs—far too many to examine manually. If you do not already know what kinds of relationships to search for, how do you efficiently

not only do relationships take many functional forms, but many important relationships—for example, a superposition of functions—are not well modeled by a function (4–7).

By equitability, we mean that the statistic should give similar scores to equally noisy rela-

of integers (x, y) the largest possible mutual information achievable by any x -by- y grid applied to the data. We then normalize these mutual information values to ensure a fair comparison between grids of different dimensions and to obtain modified values between 0 and 1. We define the characteristic matrix $M = (m_{x,y})$, where $m_{x,y}$ is the highest normalized mutual information achieved by any x -by- y grid, and the statistic MIC to be the maximum value in M (Fig. 1, B and C).

More formally, for a grid G , let I_G denote the mutual information of the probability dis-

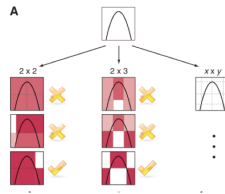


Figure: Citations on Google Scholar: 475 (2015), 1325 (2019)

Exploratory data analysis road-map

Given a huge data set, how do you search for pairs of variables that are closely associated?

- ▶ calculate some measure of dependence for each pair
- ▶ rank the pairs by their scores
- ▶ examine the top-scoring pairs.

Crucial step along the way: the statistic we use to measure dependence should have two heuristic properties

- ▶ generality
- ▶ equitability

MIC - Maximal Information Coefficient

- ▶ Functional relationships: $MIC \approx R^2$
- ▶ Range: from 0 (statistical independence) to 1 (no noise)
- ▶ For linear relationships: $MIC \approx (\text{Pearson } \rho)^2$

Note: the coefficient of determination, denoted R^2 ("R square") $\in [0, 1]$, indicates how well the given data fits a statistical model (line, curve, etc). More on this later, when discussing regression models.

Larger family of statistics:

MINE - Maximal Information-based Nonparametric Exploration:

- ▶ used to identify interesting associations
- ▶ classify associations by properties such as nonlinearity and monotonicity
- ▶ application to data sets in health, baseball, and genomics

Generality

- ▶ with sufficient sample size the statistic should capture a wide range of interesting associations, not limited to specific function types
 - ▶ linear, exponential, or periodic, or even to all functional relationships
- ▶ relationships take many functional forms, but many important relationships (e.g., a superposition of functions) are not well modeled by a function

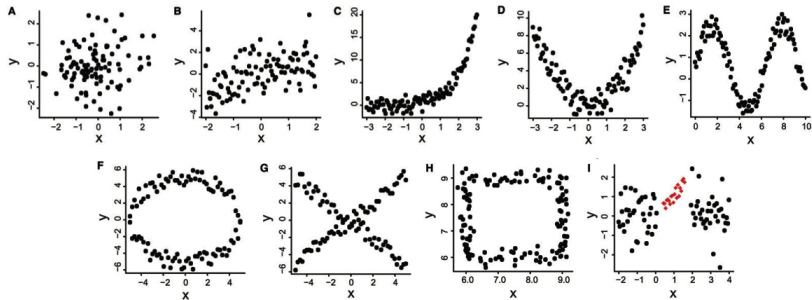


Figure: Simulations. (A) Independent data, (B) linear association, (C) exponential association - non-linear monotonic association, (D) quadratic association - non-linear non-monotonic, (E) sine association: non-linear non-monotonic, (F) circumference: non-functional association, (G) cross: non-functional association, (H) square: non-functional association and (I) local correlation: only part of the data is correlated, which is represented by crosses.

A comparative study of statistical methods used to identify dependencies between gene expression signals, Siqueira et al., BRIEFINGS IN BIOINFORMATICS. VOL 15. NO 6, 2013

Equitability

- ▶ the statistic should give similar scores to equally noisy relationships of different types
- ▶ do not want noisy linear relationships precede strong sinusoidal relationships (when sorting the pairs in terms of the proposed statistic)
- ▶ difficult to formalize for associations in general
- ▶ for the basic case of functional relationships:
 - ▶ An equitable statistic should give similar scores to functional relationships with similar R^2 values (given sufficient sample size).

Controversial result:

- ▶ Kinney & Atwal refute the claim that MIC is "equitable"

MIC - Generality and Equitability

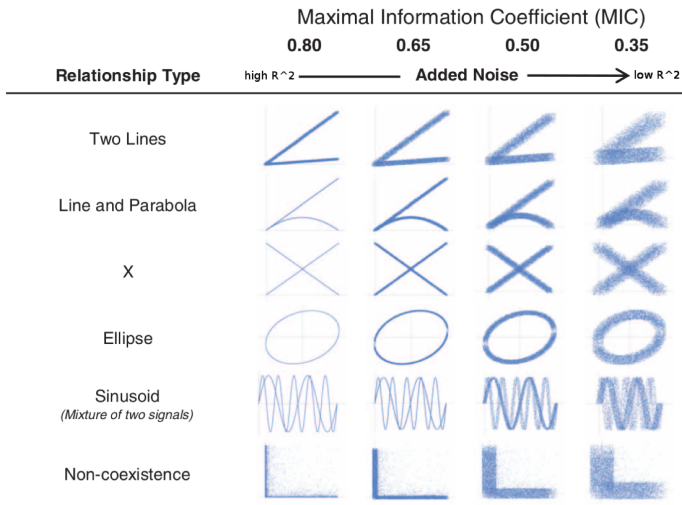


Figure: Performance of MIC on associations that not well modeled by a function (as the noise level varies).

Comparison of MIC to existing methods

A

Relationship Type	MIC	Pearson	Spearman	Mutual Information (KDE) (Kruskal)		CorGC (Principal Curve-Based)	Maximal Correlation
Random	0.18	-0.02	-0.02	0.01	0.03	0.19	0.01
Linear	1.00	1.00	1.00	5.03	3.89	1.00	1.00
Cubic	1.00	0.61	0.69	3.09	3.12	0.98	1.00
Exponential	1.00	0.70	1.00	2.09	3.62	0.94	1.00
Sinusoidal (Fourier frequency)	1.00	-0.09	-0.09	0.01	-0.11	0.36	0.64
Categorical	1.00	0.53	0.49	2.22	1.65	1.00	1.00
Periodic/Linear	1.00	0.33	0.31	0.69	0.45	0.49	0.91
Parabolic	1.00	-0.01	-0.01	3.33	3.15	1.00	1.00
Sinusoidal (non-Fourier frequency)	1.00	0.00	0.00	0.01	0.20	0.40	0.80
Sinusoidal (varying frequency)	1.00	-0.11	-0.11	0.02	0.06	0.38	0.76

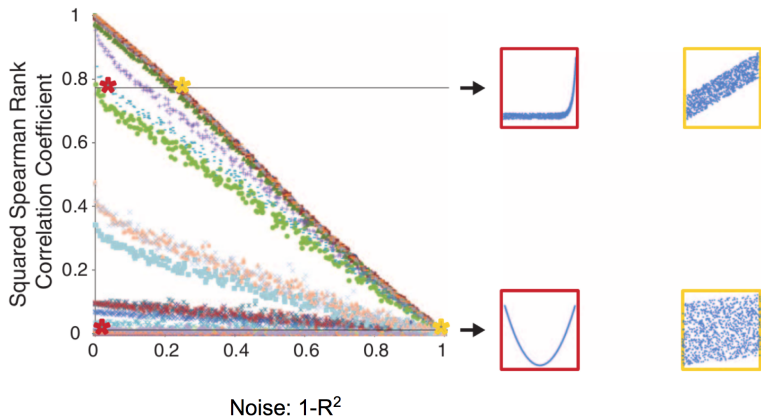
Figure: Scores given to various noiseless functional relationships by several different statistics

Experimental setup

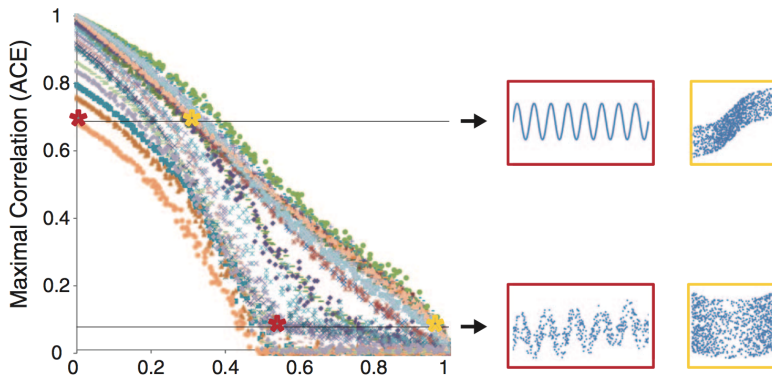
Compare MIC, Spearman correlation coefficient, mutual information, maximal correlation (ACE) on

- ▶ 27 different functional relationships with independent uniform noise added
- ▶ varying the noise level ($1-R^2$ value of the data relative to the noiseless function)
- ▶ each shape and color corresponds to a different combination of function type and sample size
- ▶ in each plot, pairs of thumbnails show relationships that received identical scores
- ▶ for data exploration, we would like these pairs to have similar noise levels

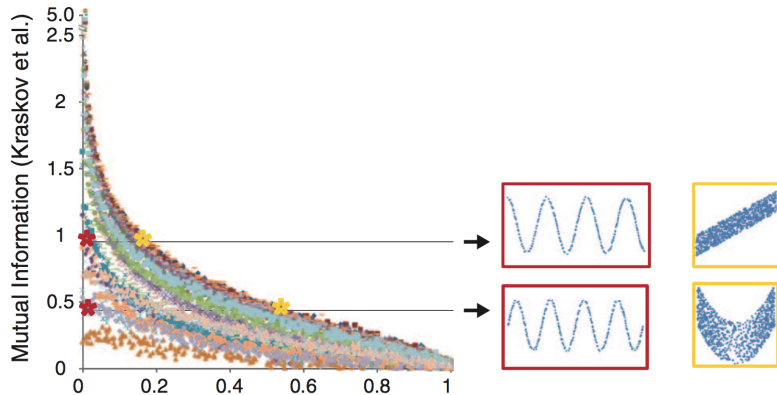
Spearman Rank Correlation vs. Noise



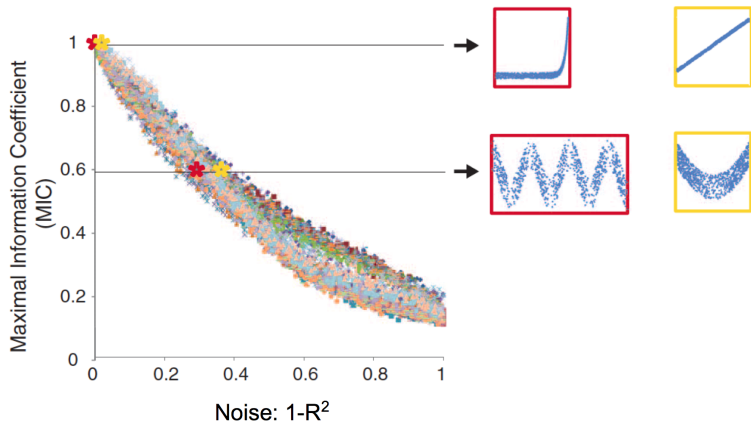
Maximal Correlation (ACE) vs. Noise



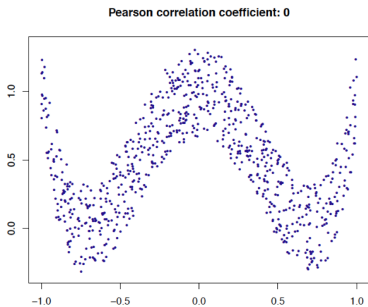
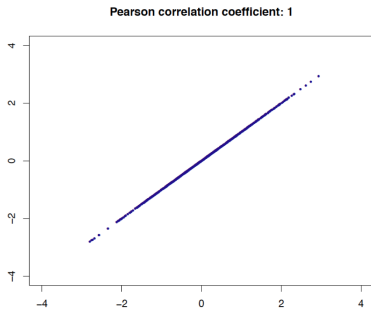
Mutual Information vs. Noise



MIC vs. Noise



Pearson correlation failing



Application to global indicators from the WHO

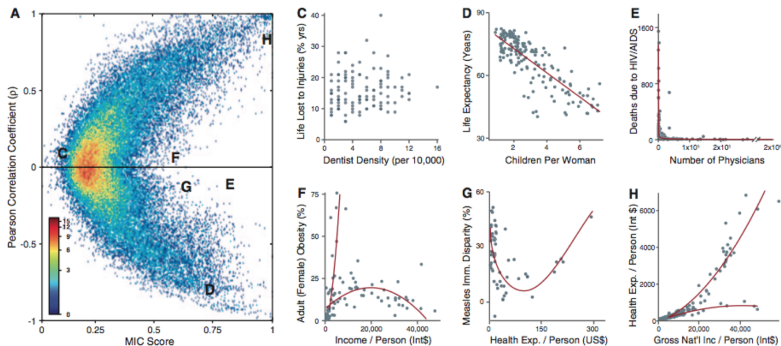


Figure: (A) MIC vs ρ for all pairwise relationships. (C) Both ρ and MIC yield low scores for unassociated variables. (D) Ordinary linear relationships score high under both tests. (E to G) Relationships detected by MIC but not by ρ , because of nonlinearity (E and G) or because more than one relationship is present (F). (H) A superposition of two relationships that scores high under all tests.

Mutual information (WHO data set)

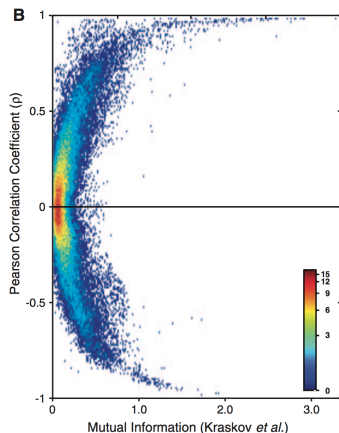


Figure: (B) Mutual information (Kruskov et al. estimator) versus Pearson ρ for the same relationships. High mutual information scores tend to be assigned only to relationships with high ρ , whereas MIC gives high scores also to relationships that are nonlinear.

Mutual information (MI) versus MIC

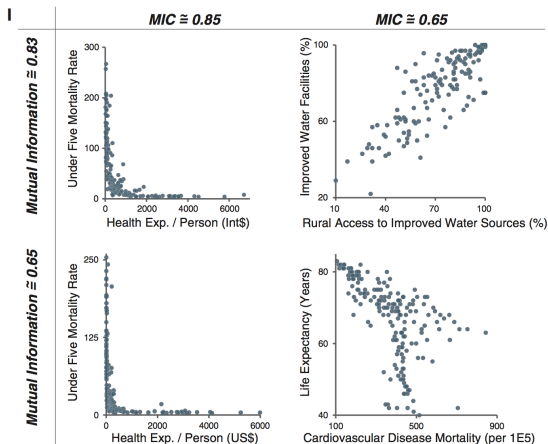


Figure: The relationships on the left appear less noisy than those on the right \Rightarrow MIC assigns higher scores to the two relationships on the left. In contrast, MI assigns similar scores to the top two relationships and similar scores to the bottom two relationships.

Calculating MIC - Central Idea

"If a relationship exists between two variables, then a grid can be drawn on the scatterplot of the two variables that partitions the data to encapsulate that relationship."

Calculating MIC - Central Idea

- ▶ explore all grids up to a maximal grid resolution, dependent on the sample size
- ▶ compute, for every pair of integers (x, y) , the largest possible MI achievable by any x -by- y grid
- ▶ normalize MI values to ensure a fair comparison
- ▶ $\overline{MI}_{xy} \in [0, 1]$
- ▶ define the **characteristic matrix** $M = (m_{x,y})$, where $m_{x,y}$ is the highest \overline{MI} achieved by any x -by- y grid
- ▶ MIC = the maximum value in M

Computing MIC: Scatterplots and Grids

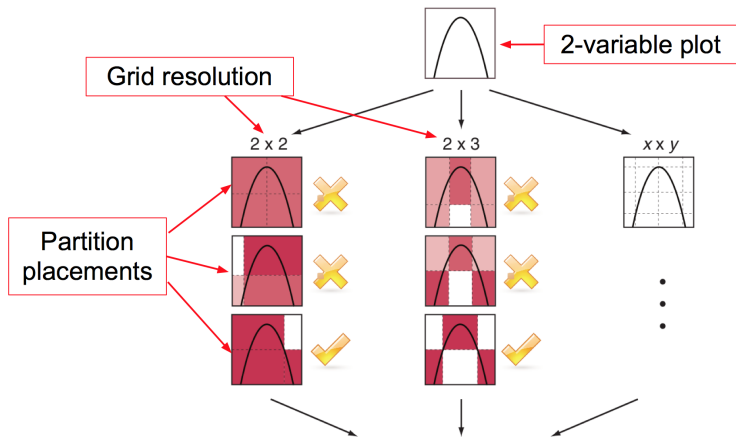


Figure: For each pair (x, y) , the MIC algorithm finds the x -by- y grid with the highest induced mutual information.

Scoring Grids

- ▶ Resolution: MIC tries all resolutions (x, y) where $xy < n^{0.6}$
- ▶ Partitioning: For each resolution (x, y) MIC finds the grid partition placement with highest mutual information MI
 - ▶ Use approximation algorithm to reduce the number of partition placements we consider

(Note: previously, we used $I(X; Y)$ for mutual information)

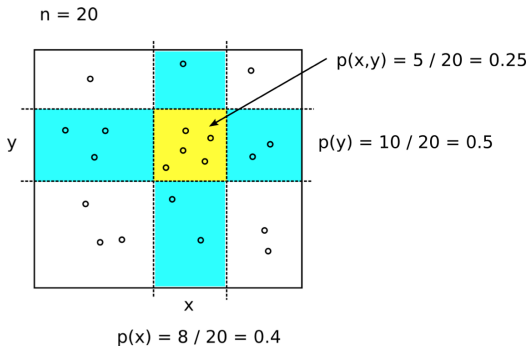
$$MI(X; Y) = D(p(x, y) \parallel p(x)p(y)) \quad (9)$$

$$= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log \frac{p(x, y)}{p(x)p(y)} \quad (10)$$

$$= \mathbb{E}_{p(x, y)} \left(\log \frac{p(X, Y)}{p(X)p(Y)} \right) \quad (11)$$

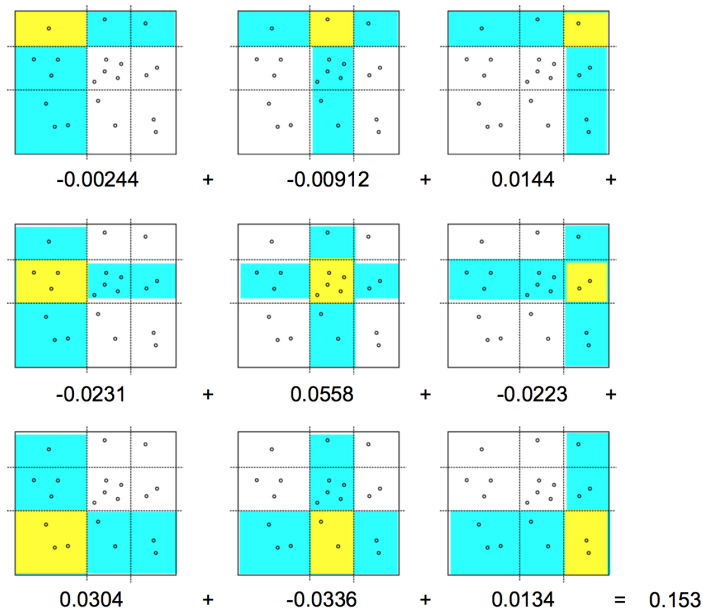
Mutual Information

Probability of a box = # of data points in that box



$$p(x, y) \log \left(\frac{p(x, y)}{p(x) p(y)} \right) = 0.25 \log \left(\frac{0.25}{0.4 \times 0.5} \right) \approx 0.056$$

Mutual Information



Characteristic matrix & Normalization

- ▶ highest mutual information score for each resolution is stored in the characteristic matrix $M_{x,y}$
- ▶ different resolution grids have different maximum mutual information scores
- ▶ normalize

$$M_{x,y} = \frac{\max MI(G)}{\log \min(x, y)}$$

- ▶ maximum is taken over all x -by- y grids G
- ▶ $M_{x,y} \in (0, 1)$

The characteristic M matrix

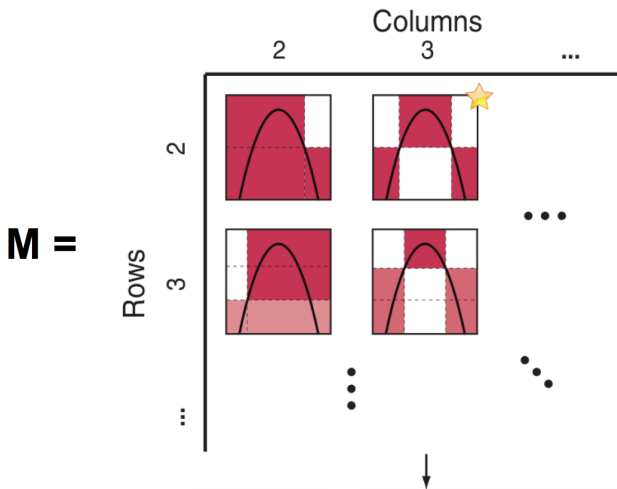


Figure: The algorithm normalizes the mutual information scores and compiles a matrix that stores, for each resolution, the best grid at that resolution and its normalized score.

The surface of the characteristic M matrix

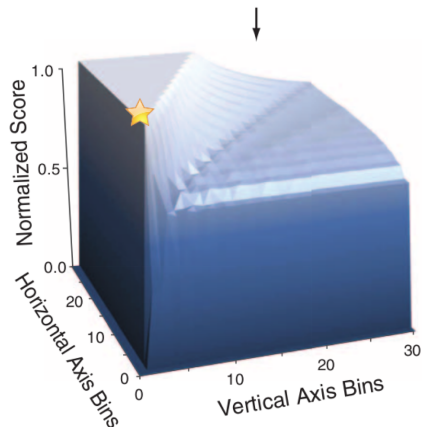


Figure: The normalized scores form the characteristic matrix M , which can be visualized as a surface; MIC corresponds to the highest point on this surface.

Measures based on MIC

Other statistics using MIC and the characteristic matrix M

- ▶ Maximum Asymmetry Score (MAS): Deviation from monotonicity
- ▶ Minimum Cell Number (MCN): Complexity measure
 - ▶ Tells you the minimum number of partitions to get the MIC score

Collection of statistics:








MINE - Maximal Information-based non-parametric Exploration

Mathematical properties of MIC

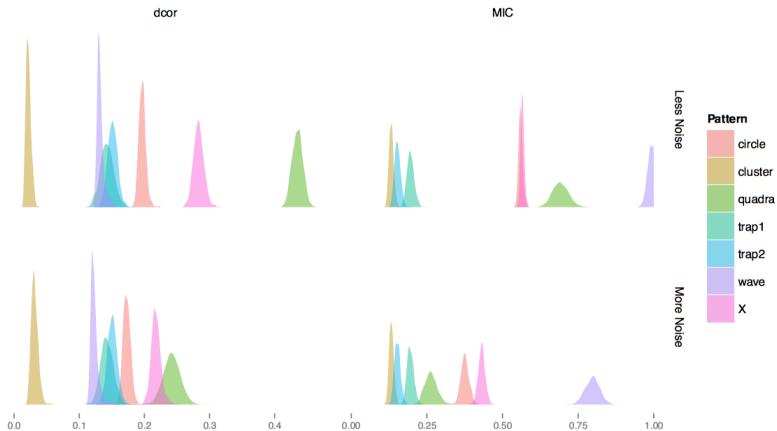
The authors were able to prove that, with probability approaching 1 as sample size grows, the following hold true

- ▶ MIC assigns scores that tend to 1 to all never-constant noiseless functional relationships
- ▶ MIC assigns scores that tend to 1 for a larger class of noiseless relationships (including super-positions of noiseless functional relationships)
- ▶ MIC assigns scores that tend to 0 to statistically independent variables

MINE statistics versus Pearson Correlation

Data	MIC	MAS	MEV	MCN	Pearson
	1.00	0.00	1.00	2.00	1.00
	1.00	0.74	1.00	3.00	-0.09
	1.00	0.89	1.00	4.00	0.01
	1.00	0.69	1.00	2.56	0.61
	0.79	0.16	0.70	6.91	-0.02
	0.71	0.03	0.32	6.87	0.00
	0.46	0.19	0.22	6.98	-0.1

dcor vs. MIC: nonlinear patterns



Criticism of MIC

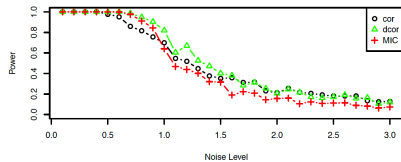
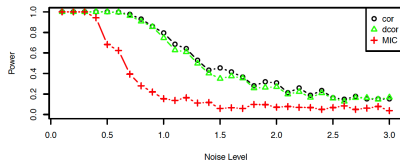
Comment to Science (Simon and Tibshirani 2012)

MIC was shown to have less power than distance correlation (dCor)

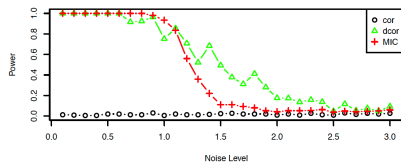
- ▶ simulated pairs of variables with varying amounts of noise added
- ▶ power $\stackrel{\text{def}}{=}$ probability that test will correctly reject H_0 (hypothesis that there is no relationship)
- ▶ lower power = more false positives

MIC vs. Pearson vs. dCor

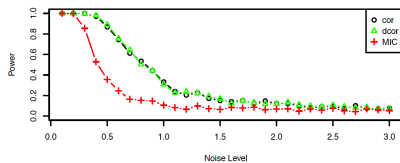
Sine: period 1/2

 $X^{1/4}$ 

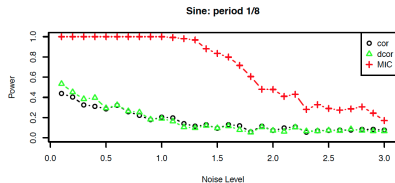
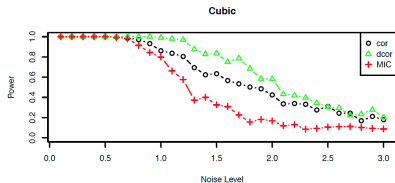
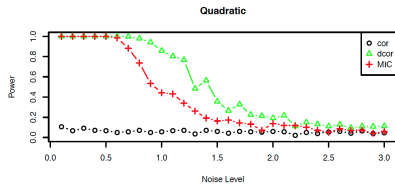
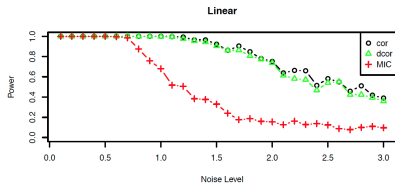
Circle



Step function



MIC vs. Pearson vs. dCor



Conclusion

MIC - useful tool:

- ▶ for mining various types of association rules
- ▶ works well for a variety of data sets
- ▶ for identification and characterization of structure in data

Follow-up references on MIC:

- ▶ *Measuring dependence powerfully and equitably*, by Yakir A. Reshef, David N. Reshef, Hilary K. Finucane, Pardis C. Sabeti, Michael M. Mitzenmacher (arXiv:1505.02213)
- ▶ *An Empirical Study of Leading Measures of Dependence*, D. Reshef, Y. Reshef, P. Sabeti, M. Mitzenmacher (arXiv:1505.02214)