A new measure of dependence: distance correlation

IX Xornada de Usuarios de R en Galicia

María Vidal García

CITMAga

October 20, 2022

Correlation

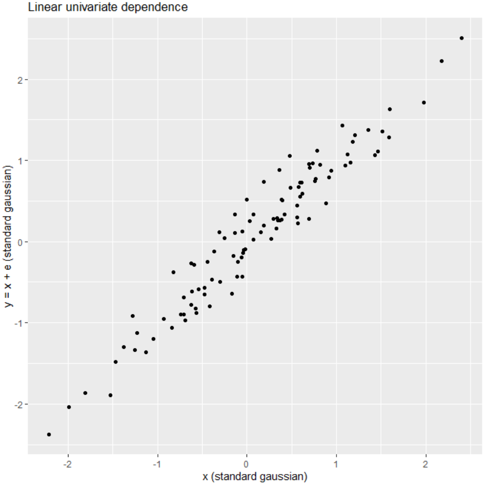
X, Y

Correlation

X, Y $\rightarrow$ Correlation?

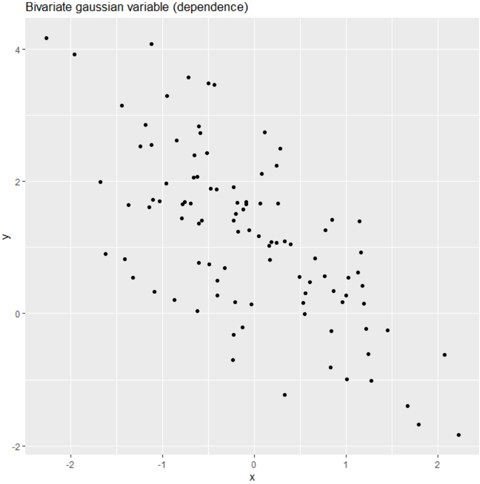
Correlation

X, Y $\rightarrow$ Correlation?



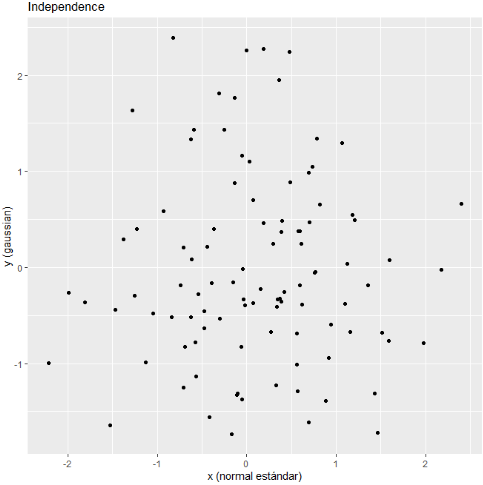
Correlation

X, Y $\rightarrow$ Correlation?



Correlation

X, Y $\rightarrow$ Correlation?



Linear correlation

Linear correlation: $X \in \mathbb{R}$, $Y \in \mathbb{R}$

Linear correlation

Linear correlation: $X \in \mathbb{R}, Y \in \mathbb{R}$

Covariance

$$Cov(X, Y) = \mathbb{E}[(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))]$$

Linear correlation

Linear correlation: $X \in \mathbb{R}, Y \in \mathbb{R}$

Covariance

$$Cov(X,Y) = \mathbb{E}[(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))]$$

Pearson correlation coefficient (Karl Pearson, 1880)

Linear correlation

Linear correlation: $X \in \mathbb{R}, Y \in \mathbb{R}$

Covariance

$$Cov(X, Y) = \mathbb{E}[(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))]$$

Pearson correlation coefficient (Karl Pearson, 1880)

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

Linear correlation

Linear correlation: $X \in \mathbb{R}, Y \in \mathbb{R}$

Covariance

$$Cov(X, Y) = \mathbb{E}[(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))]$$

Pearson correlation coefficient (Karl Pearson, 1880)

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

Properties:

- $\rho \in [-1, 1]$

Linear correlation

Linear correlation: $X \in \mathbb{R}$, $Y \in \mathbb{R}$

Covariance

$$Cov(X, Y) = \mathbb{E}[(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))]$$

Pearson correlation coefficient (Karl Pearson, 1880)

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

Properties:

- $\rho \in [-1, 1]$
- $\rho = 0 \Leftrightarrow X$ e Y are linearly uncorrelated

Linear correlation

Linear correlation: $X \in \mathbb{R}, Y \in \mathbb{R}$

Covariance

$$Cov(X, Y) = \mathbb{E}[(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))]$$

Pearson correlation coefficient (Karl Pearson, 1880)

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

Properties:

- $\rho \in [-1, 1]$
- $\rho = 0 \Leftrightarrow X$ e Y are linearly uncorrelated
- ρ measures the degree of linear correlation between variables

Linear correlation

Linear correlation: $X \in \mathbb{R}, Y \in \mathbb{R}$

Covariance

$$Cov(X, Y) = \mathbb{E}[(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))]$$

Pearson correlation coefficient (Karl Pearson, 1880)

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

Properties:

- $\rho \in [-1, 1]$
- $\rho = 0 \Leftrightarrow X$ e Y are linearly uncorrelated
- ρ measures the degree of linear correlation between variables
- Invariant to scale-location changes

Linear correlation

Linear correlation: $X \in \mathbb{R}, Y \in \mathbb{R}$

Covariance

$$Cov(X, Y) = \mathbb{E}[(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))]$$

Pearson correlation coefficient (Karl Pearson, 1880)

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

Properties:

- $\rho \in [-1, 1]$
- $\rho = 0 \Leftrightarrow X$ e Y are linearly uncorrelated
- ρ measures the degree of linear correlation between variables
- Invariant to scale-location changes
- **Bivariant normal case**: ρ characterizes independence

Linear correlation

Linear correlation: $X \in \mathbb{R}, Y \in \mathbb{R}$

Covariance

$$Cov(X,Y) = \mathbb{E}[(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))]$$

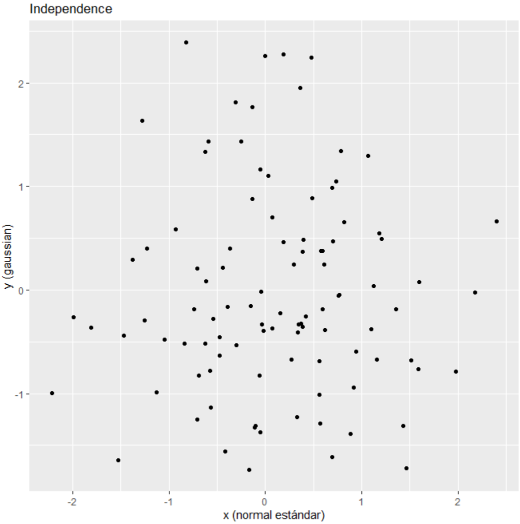
Pearson correlation coefficient (Karl Pearson, 1880)

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}} \quad \leftarrow \quad r = \frac{S_{X,Y}}{\sqrt{S_X^2}\sqrt{S_Y^2}}$$

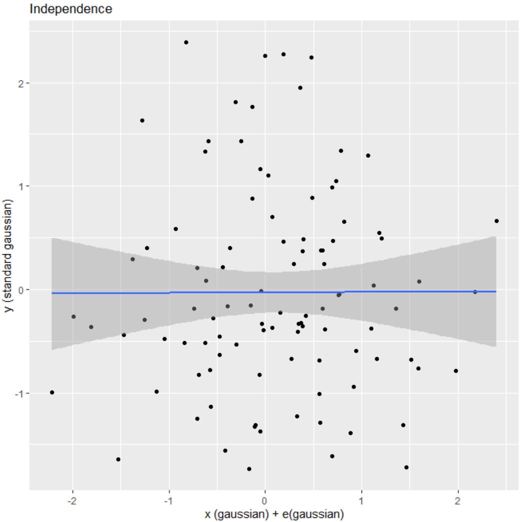
Properties:

- $\rho \in [-1, 1]$
- $\rho = 0 \Leftrightarrow X$ e Y are linearly uncorrelated
- ρ measures the degree of linear correlation between variables
- Invariant to scale-location changes
- **Bivariant normal case**: ρ characterizes independence

Scenario 1: no correlation

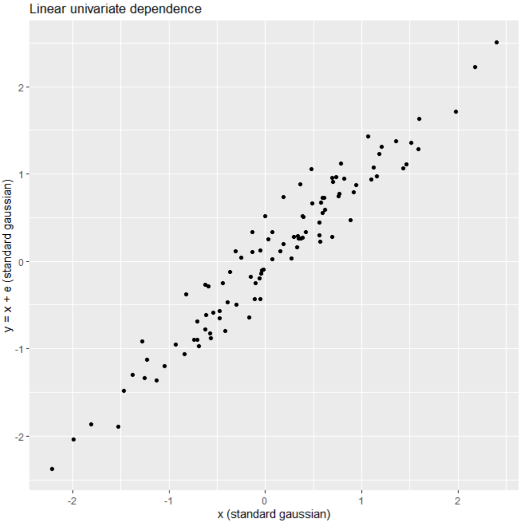


Scenario 1: no correlation

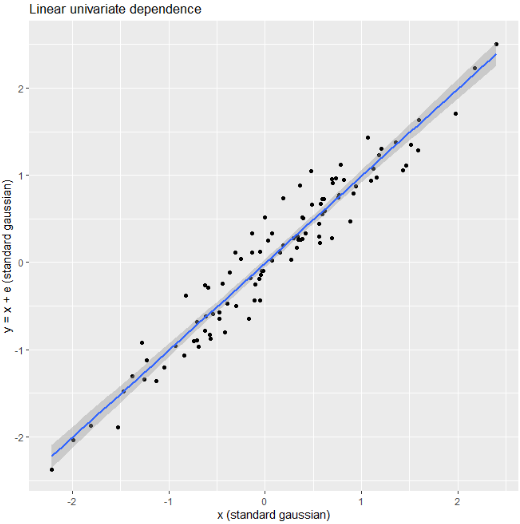


$r = 0.0039$

Scenario 2: positive correlation

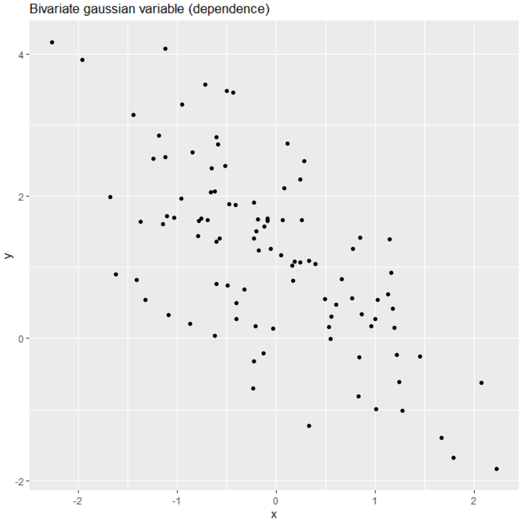


Scenario 2: positive correlation

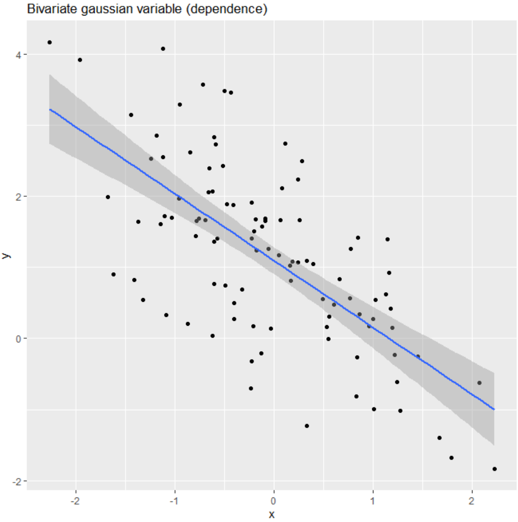


$r = 0.9662$

Scenario 3: negative correlation



Scenario 3: negative correlation

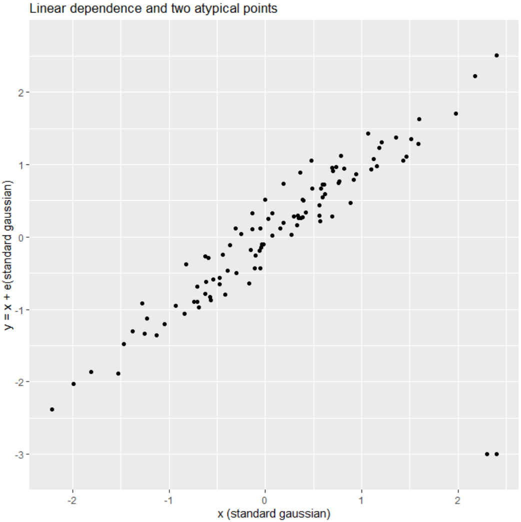


$r = -0.6752$

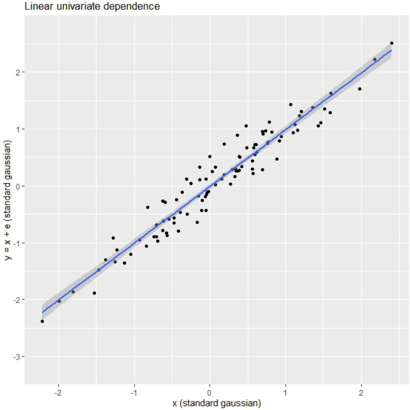
Limitations of Pearson correlation

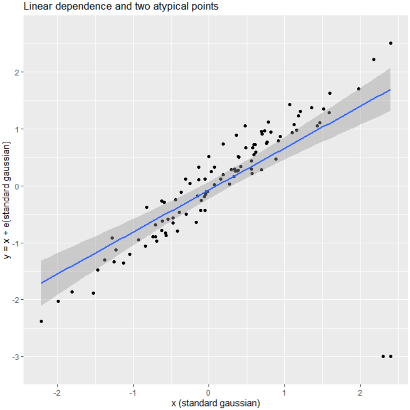
1. Lack of robustness
2. Unable to capture general dependence estructures
3. Non-applicability to multidimensional variables

Scenario 4: Lack of robustness (!)

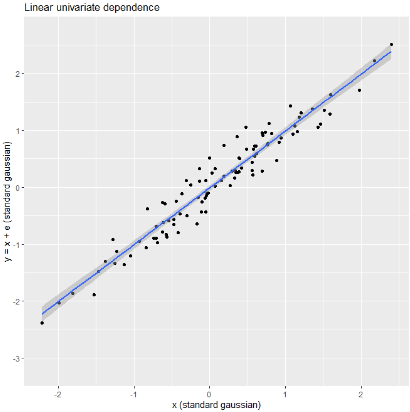


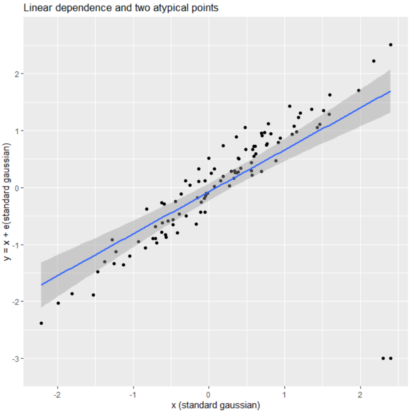
Scenario 4: Lack of robustness (!)





Scenario 4: Lack of robustness (!)





$r = 0.9662$

$r = 0.6845$

Robustified dependence measures

Spearman's rank correlation coefficient r_s (1904)

$$r_s = \rho(R(X), R(Y))$$

Robustified dependence measures

Spearman's rank correlation coefficient r_s (1904)

$$r_s = \rho(R(X), R(Y))$$

Kendall's rank correlation coefficient τ (1938)

$$\tau = \frac{\#\{\text{concordant pairs}\} - \#\{\text{discordant pairs}\}}{\#\{\text{number of pairs}\}}$$

Robustified dependence measures

Spearman's rank correlation coefficient r_s (1904)

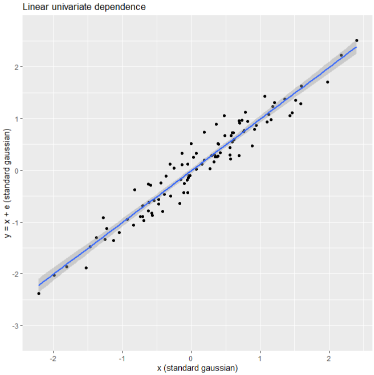
$$r_s = \rho(R(X), R(Y))$$

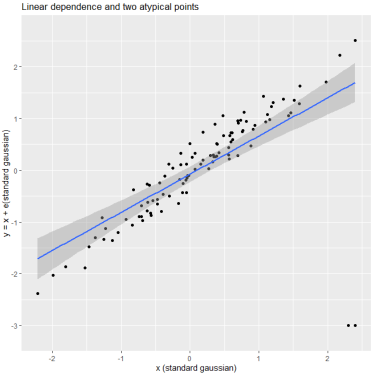
Kendall's rank correlation coefficient τ (1938)

$$\tau = \frac{\#\{\text{concordant pairs}\} - \#\{\text{discordant pairs}\}}{\#\{\text{number of pairs}\}}$$

```
cor(x,y)
cor(x,y, method="spearman")
cor(x,y, method="kendall")
```

Scenario 4: Lack of robustness (!)





$r = 0.9662$

$r = 0.6845$

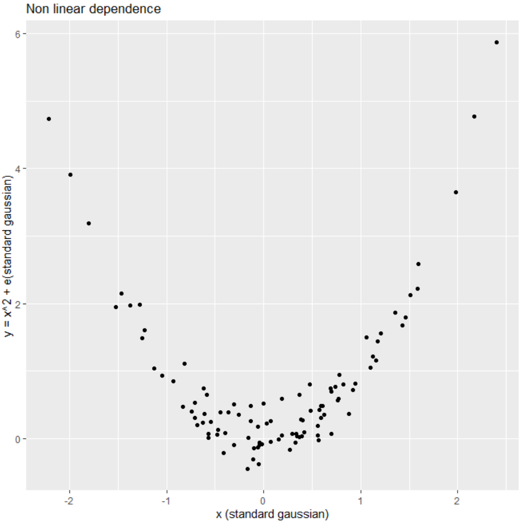
$r_s = 0.8500$

$\tau = 0.7735$

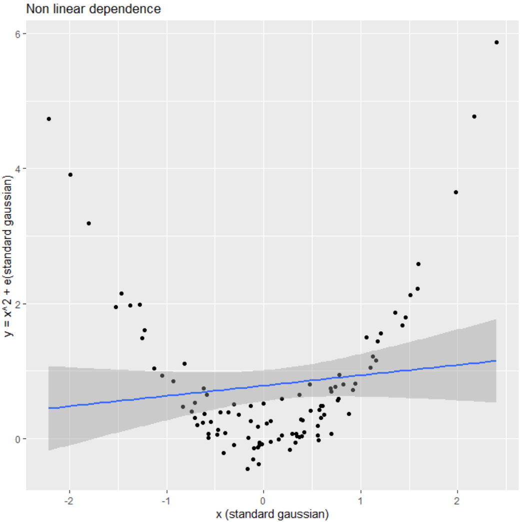
Limitations of Pearson correlation

1. Lack of robustness
2. Unable to capture general dependence estructures
3. Non-applicability to multidimensional variables

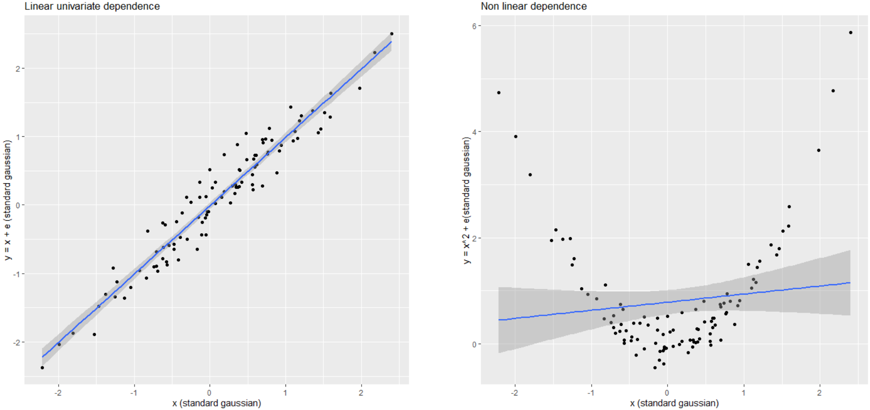
Scenario 5: General dependence (!)



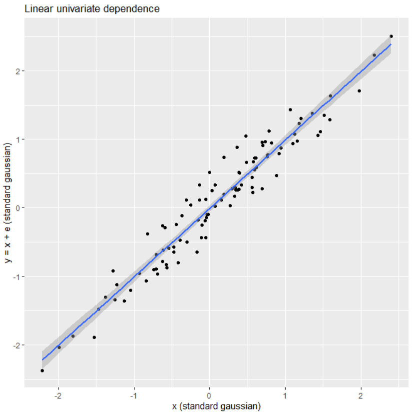
Scenario 5: General dependence (!)

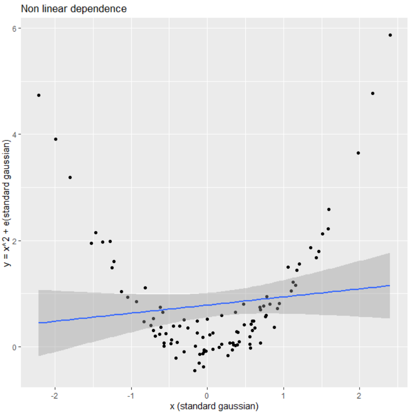


Scenario 5: General dependence (!)



Scenario 5: General dependence (!)





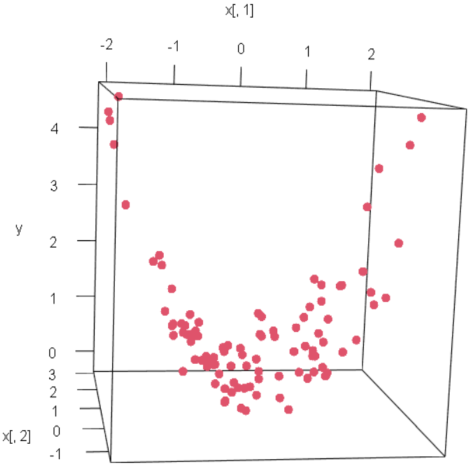
$r = 0.9662$

$r = 0.1210$

Limitations of Pearson correlation

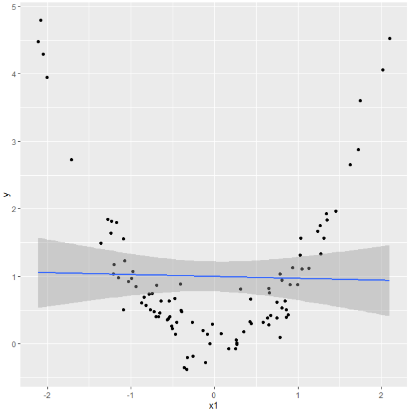
1. Lack of robustness
2. Unable to capture general dependence estructures
3. Non-applicability to multidimensional variables

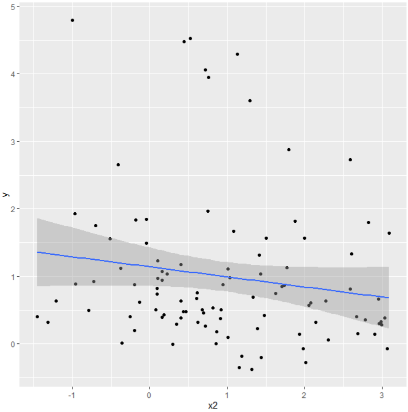
Scenario 6: Multidimensional covariable



$$r(x, y) =?$$

Scenario 6: Multidimensional covariable





$r(X_1, Y) = -0.0251$ $r(X_2, Y) = -0.1561$

Measuring dependence

$X \in \mathbb{R}^p$, $Y \in \mathbb{R}^q$

Measuring dependence

$X \in \mathbb{R}^p, \ Y \in \mathbb{R}^q$

$$\varphi_X(t) = \mathbb{E}\left[e^{itX}\right]$$

Independence condition

$$\varphi_{X,Y} = \varphi_X \varphi_Y$$

Measuring dependence

$X \in \mathbb{R}^p, \ Y \in \mathbb{R}^q$

$$\varphi_X(t) = \mathbb{E}\left[e^{itX}\right]$$

Independence condition

$$\varphi_{X,Y} = \varphi_X \varphi_Y$$

$$\|\varphi_{X,Y} - \varphi_X\varphi_Y\|^2 = \frac{1}{c_p c_q}\int_{\mathbb{R}^{p+q}} \frac{|\varphi_{X,Y}(t,s) - \varphi_X(t)\varphi_Y(s)|^2}{|t|^{1+p}|s|^{1+q}} dtds$$

Székely et al. (2007)

Distance correlation

Distance covariance

Given $X \in \mathbb{R}^p$, $Y \in \mathbb{R}^q$ random vectors with finite first order moments $\mathbb{E}(|X|) < \infty$, $\mathbb{E}(|Y|) < \infty$. Distance covariance between them is defined as the square root of

$$\mathcal{V}^2(X, Y) := \|\varphi_{X,Y} - \varphi_X \varphi_Y\|^2$$

Distance correlation

Distance covariance

Given $X \in \mathbb{R}^p$, $Y \in \mathbb{R}^q$ random vectors with finite first order moments $\mathbb{E}(|X|) < \infty$, $\mathbb{E}(|Y|) < \infty$. Distance covariance between them is defined as the square root of

$$\mathcal{V}^2(X, Y) := \|\varphi_{X,Y} - \varphi_X \varphi_Y\|^2$$

Distance correlation

Given $X \in \mathbb{R}^p$, $Y \in \mathbb{R}^q$ random vectors with finite first order moments $\mathbb{E}(|X|) < \infty$, $\mathbb{E}(|Y|) < \infty$. Distance correlation between them is defined as the square root of

$$\mathcal{R}^2(X, Y) = \frac{\mathcal{V}^2(X, Y)}{\sqrt{\mathcal{V}^2(X)\mathcal{V}^2(Y)}}$$

Distance correlation

Properties of $\mathcal{R}(X, Y)$

- Adimensional and bounded: $\mathcal{R}(X, Y) \in [0, 1]$.
- Non-negative.
- Characterizes independence.
- In bivariate normal case

$$\mathcal{R}(X, Y) \leq |\rho(X, Y)|$$

Empirical statistics

Evaluables on a sample $\{X_i, Y_i\}_{i=1}^n$

Empirical statistics

Evaluables on a sample $\{X_i, Y_i\}_{i=1}^n$

- Distance matrix: $a_{kl} = |X_k - X_l|$.
- Double-centered distance matrix: $A_{kl} = a_{kl} - \bar{a}_{k\cdot} - \bar{a}_{\cdot l} + \bar{a}_{\cdot\cdot}$.

Empirical statistics

Evaluables on a sample $\{X_i, Y_i\}_{i=1}^{n}$

- Distance matrix: $a_{kl} = |X_k - X_l|$.
- Double-centered distance matrix: $A_{kl} = a_{kl} - \bar{a}_{k\cdot} - \bar{a}_{\cdot l} + \bar{a}_{\cdot\cdot}$.

$$\mathcal{V}^2(X, Y) \qquad \leftarrow \qquad V^2 = \frac{1}{n^2} \sum_{k,l=1}^{n} A_{kl} B_{kl},$$

Empirical statistics

Evaluables on a sample $\{X_i, Y_i\}_{i=1}^n$

- Distance matrix: $a_{kl} = |X_k - X_l|$.
- Double-centered distance matrix: $A_{kl} = a_{kl} - \bar{a}_{k\cdot} - \bar{a}_{\cdot l} + \bar{a}_{\cdot\cdot}$.

$$\mathcal{V}^2(X, Y) \qquad \leftarrow \qquad V^2 = \frac{1}{n^2} \sum_{k,l=1}^{n} A_{kl} B_{kl},$$

$$\mathcal{R}(X, Y) \qquad \leftarrow \qquad R = \sqrt{\frac{V^2(X, Y)}{\sqrt{V^2(X) V^2(Y)}}}.$$

Empirical statistics

R package energy (2019): `library(energy)`

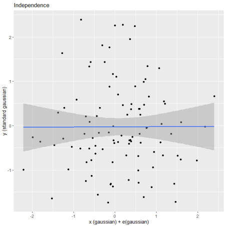
- Cálculo de V^2:

```
a <- dist(x); b <- dist(y)
A <- Dcenter(a); B <- Dcenter(b)
V2 <- sum(A*B)/(n^2)
```

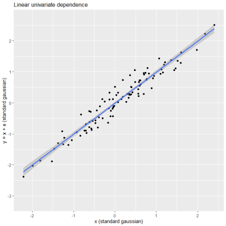
```
dcov(x,y)^2
dcov2d(x,y)
```

- Cálculo de R:

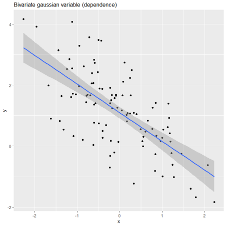
```
dcor(x,y)
```



$$r = 0.0039 \quad \leftarrow \texttt{cor(x,y)}$$
$$R = 0.1474 \quad \leftarrow \texttt{dcor(x,y)}$$

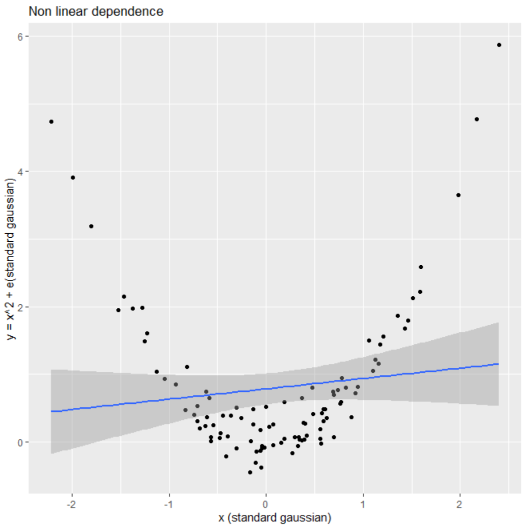


$$r = 0.9662$$
$$R = 0.9479$$



$$r = -0.6752$$
$$R = 0.6186$$

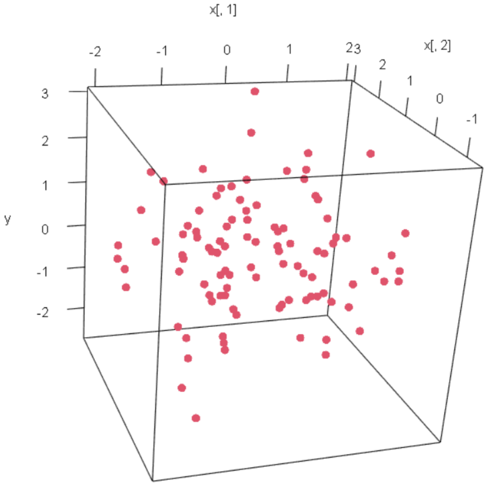
Scenario 5: General dependence (!).



$$r = 0.1210$$

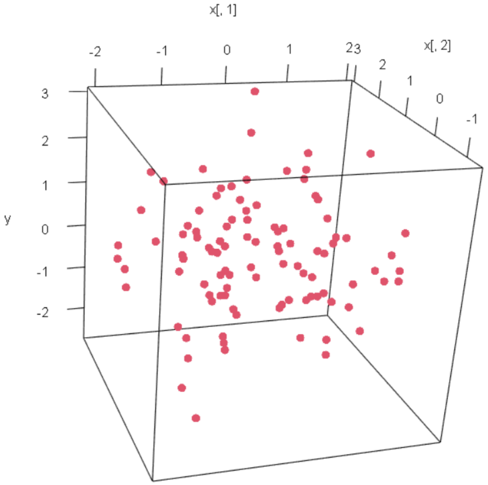
$$R = 0.5250$$

Scenario 6.1: Multidimensional covariable. Independence



$r(x, y) = ?$

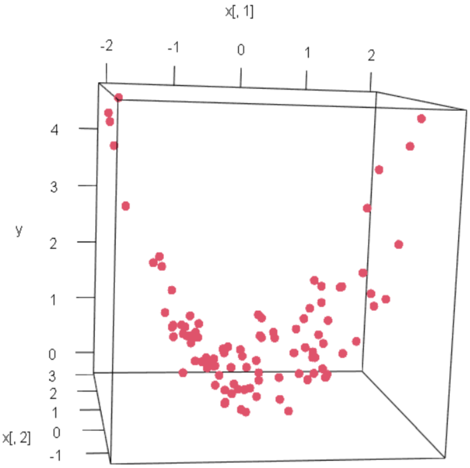
Scenario 6.1: Multidimensional covariable. Independence



$$r(x, y) = ?$$

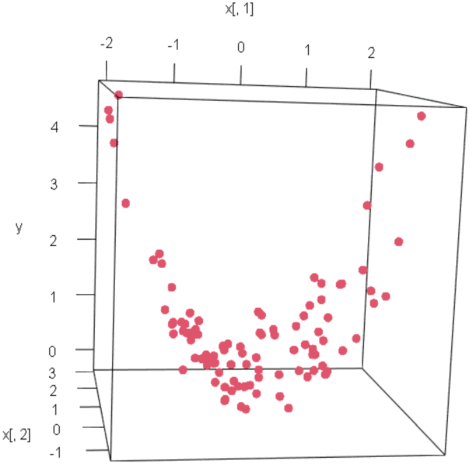
$$R(x, y) = 0.1902$$

Scenario 6.2: Multidimensional covariable. Non-linear dependence with one component



$$r(x, y) = ?$$

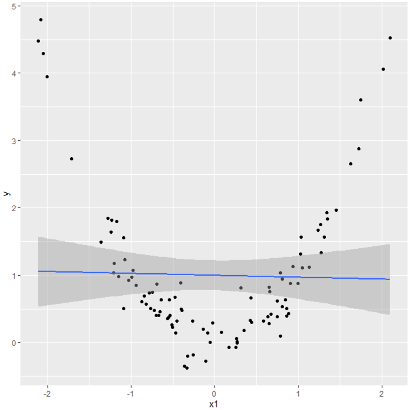
Scenario 6.2: Multidimensional covariable. Non-linear dependence with one component

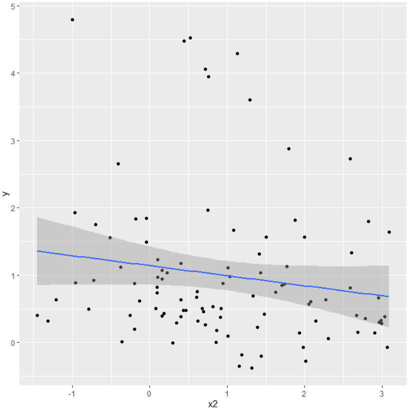


$$r(x, y) = ?$$

$$R(x, y) = 0.3411$$

Scenario 6.2: Multidimensional covariable. Non-linear dependence with one component





$$r(X_1, Y) = -0.0251$$
$$R(X_1, Y) = 0.4525$$

$$r(X_2, Y) = -0.1561$$
$$R(X_2, Y) = 0.1939$$

Scenario 7: High dimension. Linear dependence with every component

High dimension: $p > n$

Scenario 7: High dimension. Linear dependence with every component

High dimension: $p > n$

$$p = 110; \qquad n = 100$$

Scenario 7: High dimension. Linear dependence with every component

High dimension: $p > n$

$$p = 110; \qquad n = 100$$

$$X = \begin{pmatrix} X_1 \\ X_2 \\ ... \\ X_p \end{pmatrix} \sim N_p \left(\begin{pmatrix} 1 \\ 1 \\ ... \\ 1 \end{pmatrix}, \begin{pmatrix} 1 & 0 & 0 & ... & 0 \\ 0 & 1 & 0 & ... & 0 \\ ... & ... & ... & ... & ... \\ 0 & 0 & 0 & ... & 1 \end{pmatrix} \right)$$

$$Y = X_1 + X_2 + ... + X_p + \varepsilon$$

Scenario 7: High dimension. Linear dependence with every component

$r(X, Y) = ?$

$R(X, Y) = 0.4291$

$r(X_1, Y) = 0.0995$

$r(X_2, Y) = -0.0844$

$R(X_1, Y) = 0.2104$

$R(X_2, Y) = 0.1513$

Independence tests

$$\begin{cases} H_0 : \varphi_{X,Y} = \varphi_X \varphi_Y & (\textit{independence}) \\ H_1 : \varphi_{X,Y} \neq \varphi_X \varphi_Y & (\textit{dependence}) \end{cases}$$

Independence tests

$$\begin{cases} H_0 : \varphi_{X,Y} = \varphi_X \varphi_Y & (\mathit{independence}) \\ H_1 : \varphi_{X,Y} \neq \varphi_X \varphi_Y & (\mathit{dependence}) \end{cases}$$

Independence test based on V^2

- Based on asymptotic results.
- Permutation test.

Permutation test

Original sample:

$$\{(X_i, Y_i)\}_{i=1}^{n} \quad \rightarrow$$

Permutation test

Original sample:

$$\{(X_i, Y_i)\}_{i=1}^{n} \quad \rightarrow \quad V^2$$

Permutation test

Original sample:

$$\{(X_i, Y_i)\}_{i=1}^{n} \quad \rightarrow \quad V^2$$

Permutations of the original sample:

$$j \in \{1, ..., B\} : \quad \{1, ..., n\} \rightarrow \{\pi_j(1), ..., \pi_j(n)\}$$

Permutation test

Original sample:

$$\{(X_i, Y_i)\}_{i=1}^{n} \quad \rightarrow \quad V^2$$

Permutations of the original sample:

$$j \in \{1, ..., B\} : \quad \{1, ..., n\} \rightarrow \{\pi_j(1), ..., \pi_j(n)\}$$

$$\{(X_i, Y_{\pi_j(i)})\}_{i=1}^{n} \quad \rightarrow \quad V^2_{\pi_j}$$

Permutation test

Original sample:

$$\{(X_i, Y_i)\}_{i=1}^{n} \quad \rightarrow \quad V^2$$

Permutations of the original sample:

$$j \in \{1, ..., B\} : \quad \{1, ..., n\} \rightarrow \{\pi_j(1), ..., \pi_j(n)\}$$

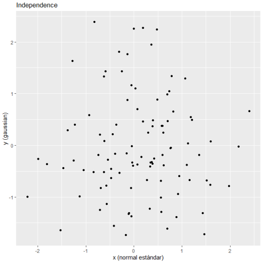
$$\{(X_i, Y_{\pi_j(i)})\}_{i=1}^{n} \quad \rightarrow \quad V_{\pi_j}^2$$

Independence test

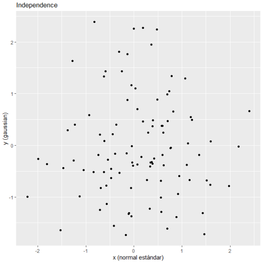
Permutation test

$$p = \frac{\#\{j \in \{1, ..., B\} | V_{\pi_j}^2 \geq V^2\} + 1}{B + 1}$$

Independence tests

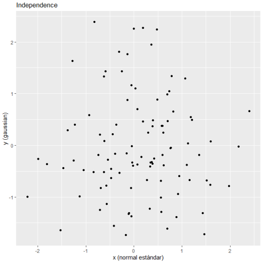


Independence tests




```
dcov.test(x,y,R=499)
```

Independence tests




```
dcov.test(x,y,R=499)

dCov independence test (permutation test)
data:  index 1, replicates 499
```

$nV^2 = 0.76471, p -$ value = 0.77

```
sample estimates:
    dCov
0.08744772
```

Independence tests

Permutation test results				
Scenario	r(X,Y)	R(X,Y)	p-value	Reject H_0
1. Independence	0.0039	0.1474	0.77	NO
2. Linear dependence	0.9662	0.9480	0.002	YES
3. Dependence in bivariate normal	-0.6752	0.6186	0.002	YES
4. Linear dep + atypical obs	0.6845	0.8579	0.002	YES
5. Non-linear dependence	0.1210	0.5250	0.002	YES
6.1. $X \in \mathbb{R}^2$: independence	?	0.1902	0.672	NO
6.2. $X \in \mathbb{R}^2$: non linear dependence	?	0.3411	0.002	YES
7. High Dimension: linear dependence	?	0.4291	0.002	YES

Independence tests

Permutation test results				
Scenario	r(X,Y)	R(X,Y)	p-value	Reject H_0
1. Independence	0.0039	0.1474	0.77	NO
2. Linear dependence	0.9662	0.9480	0.002	YES
3. Dependence in bivariate normal	-0.6752	0.6186	0.002	YES
4. Linear dep + atypical obs	0.6845	0.8579	0.002	YES
5. Non-linear dependence	0.1210	0.5250	0.002	YES
6.1. $X \in \mathbb{R}^2$: independence	?	0.1902	0.672	NO
6.2. $X \in \mathbb{R}^2$: non linear dependence	?	0.3411	0.002	YES
7. High Dimension: linear dependence	?	0.4291	0.002	YES

The proposed independence test detects also the existence of dependence within the covariable.

Take home message

1. Pearson correlation: `cor`
   Easy to interpret
   Lack of robustness, Just linear relations, Just unidimensional

2. Spearman and Kendall: `cor(..., method=" ")`
   Robust
   Just monotone relations, Just unidimensional

3. Distance correlation: `energy::dcor`
   General dependence, multidimensional

   Independence test: `dcov.test(x,y,R=499)`

References

- Rizzo, M. L. & Székely, G. J. (2022). energy: E-Statistics: Multivariate Inference via the Energy of Data. R package version 1.7-10, https://CRAN.R-project.org/package=energy.
- Székely, G. J., Rizzo, M. L. & Bakirov, N. K. (2007). Measuring and testing dependence by correlation of distances. *The annals of statistics*, 35(6), 2769-2794.
- Tjøstheim, D., Otneim, H., & Støve, B. (2022). Statistical Dependence: Beyond Pearson's . *Statistical Science*, 37(1), 90-109.

Thank you.