Correlation analysis 2: Measures of correlation

Ryan Tibshirani
Data Mining: 36-462/36-662

February 19 2013

Review: correlation

Pearson's correlation is a measure of linear association

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

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

In the sample: for vectors $x, y \in \mathbb{R}^n$,

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

If x, y are have been centered, then

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

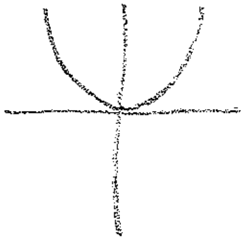
Review: properties of population correlation

Recall: $\mathrm{Cor}(X,Y) = \mathrm{Cov}(X,Y)/(\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)})$

Properties of Cor:

1. $\mathrm{Cor}(X,X) = 1$
2. $\mathrm{Cor}(X,Y) = \mathrm{Cor}(Y,X)$
3. $\mathrm{Cor}(aX+b,Y) = \mathrm{sign}(a)\mathrm{Cor}(X,Y)$ for any $a,b \in \mathbb{R}$
4. $-1 \leq \mathrm{Cor}(X,Y) \leq 1$, with $\mathrm{Cor}(X,Y) > 0$ indicating a positive (linear) relationship, < 0 indicating a negative one
5. $|\mathrm{Cor}(X,Y)| = 1$ if and only if $Y = aX + b$ for some $a,b \in \mathbb{R}$, with $a \neq 0$
6. If X,Y are independent then $\mathrm{Cor}(X,Y) = 0$
7. If $\mathrm{Cor}(X,Y) = 0$ then X,Y need not be independent
8. If (X,Y) is bivariate normal and $\mathrm{Cor}(X,Y) = 0$, then X,Y are independent

$Y = X^2$



Bivariate normal distribution

Assume that the random vector $Z = (X, Y) \in \mathbb{R}^2$ has a bivariate normal distribution, $Z \sim N(\mu, \Sigma)$, where $\mu \in \mathbb{R}^2$ and $\Sigma \in \mathbb{R}^{2\times 2}$,

$$\mu = \begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{pmatrix}$$

Note that we have $\mathrm{E}[X] = \mu_X$, $\mathrm{E}[Y] = \mu_Y$, $\mathrm{Var}(X) = \sigma_X^2$, $\mathrm{Var}(Y) = \sigma_X^2$, $\mathrm{Cov}(X, Y) = \rho\sigma_X\sigma_Y$, and $\mathrm{Cor}(X, Y) = \rho$

The density of $Z = (X, Y)$ is

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

Fact: $\rho = 0$ implies that X, Y are independent (Homework 3)

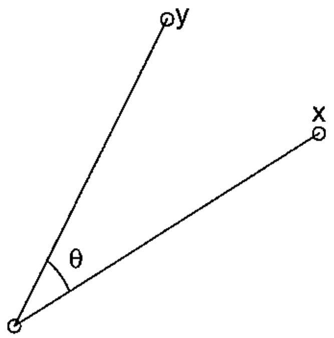
Review: properties of sample correlation

Recall: $\mathrm{cor}(x,y) = (x - \bar{x}\mathbb{1})^T(y - \bar{y}\mathbb{1}) / \big(\|x - \bar{x}\mathbb{1}\|_2 \|y - \bar{y}\mathbb{1}\|_2\big)$

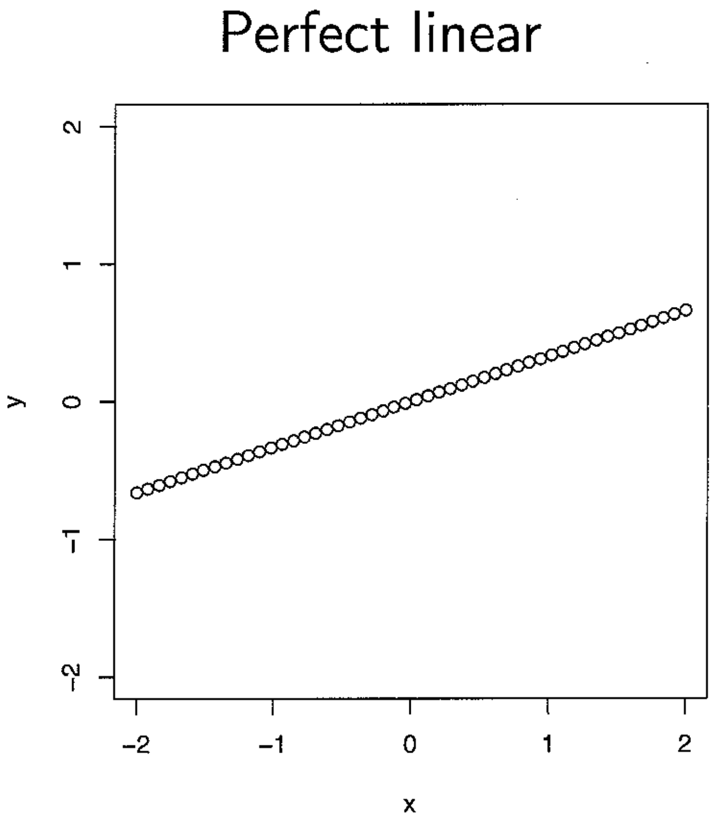
Properties of cor:

1. $\mathrm{cor}(x,x) = 1$
2. $\mathrm{cor}(x,y) = \mathrm{cor}(y,x)$
3. $\mathrm{cor}(ax+b,y) = \mathrm{sign}(a)\mathrm{cor}(x,y)$ for any $a,b \in \mathbb{R}$
4. $-1 \leq \mathrm{cor}(x,y) \leq 1$, with $\mathrm{cor}(x,y) > 0$ indicating a positive (linear) relationship, < 0 indicating a negative one
5. $|\mathrm{cor}(x,y)| = 1$ if and only if $y = ax+b$ for some $a,b \in \mathbb{R}$, with $a \neq 0$
6. $\mathrm{cor}(x,y) = 0$ if and only if x,y are orthogonal centered.
7. If x,y are centered then $\mathrm{cor}(x,y) = \cos\theta$, where θ is the angle between the vectors $x,y \in \mathbb{R}^n$

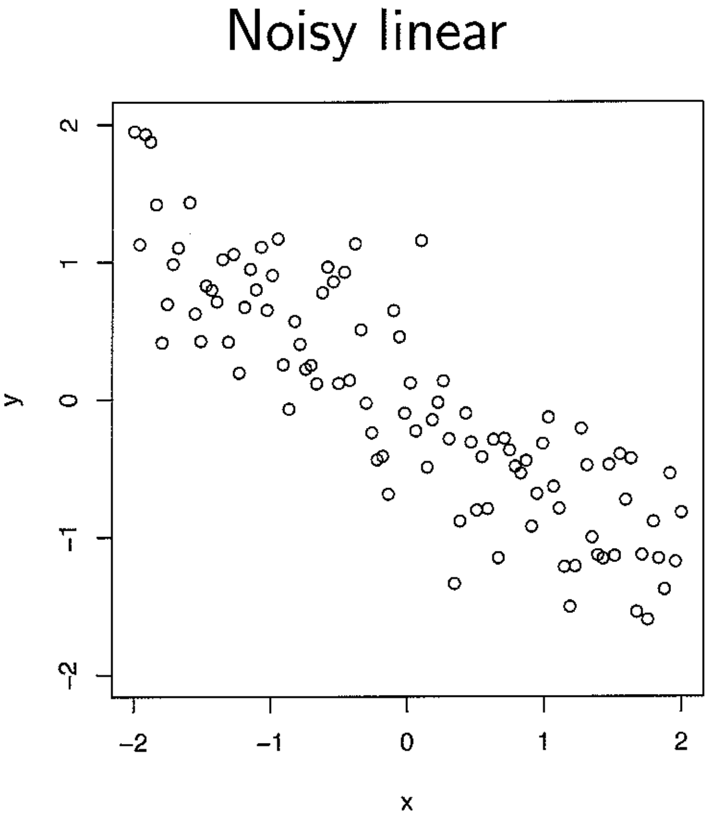
$$\cos\theta = \frac{x^T y}{\|x\|_2 \|y\|_2}$$



Examples: sample correlation

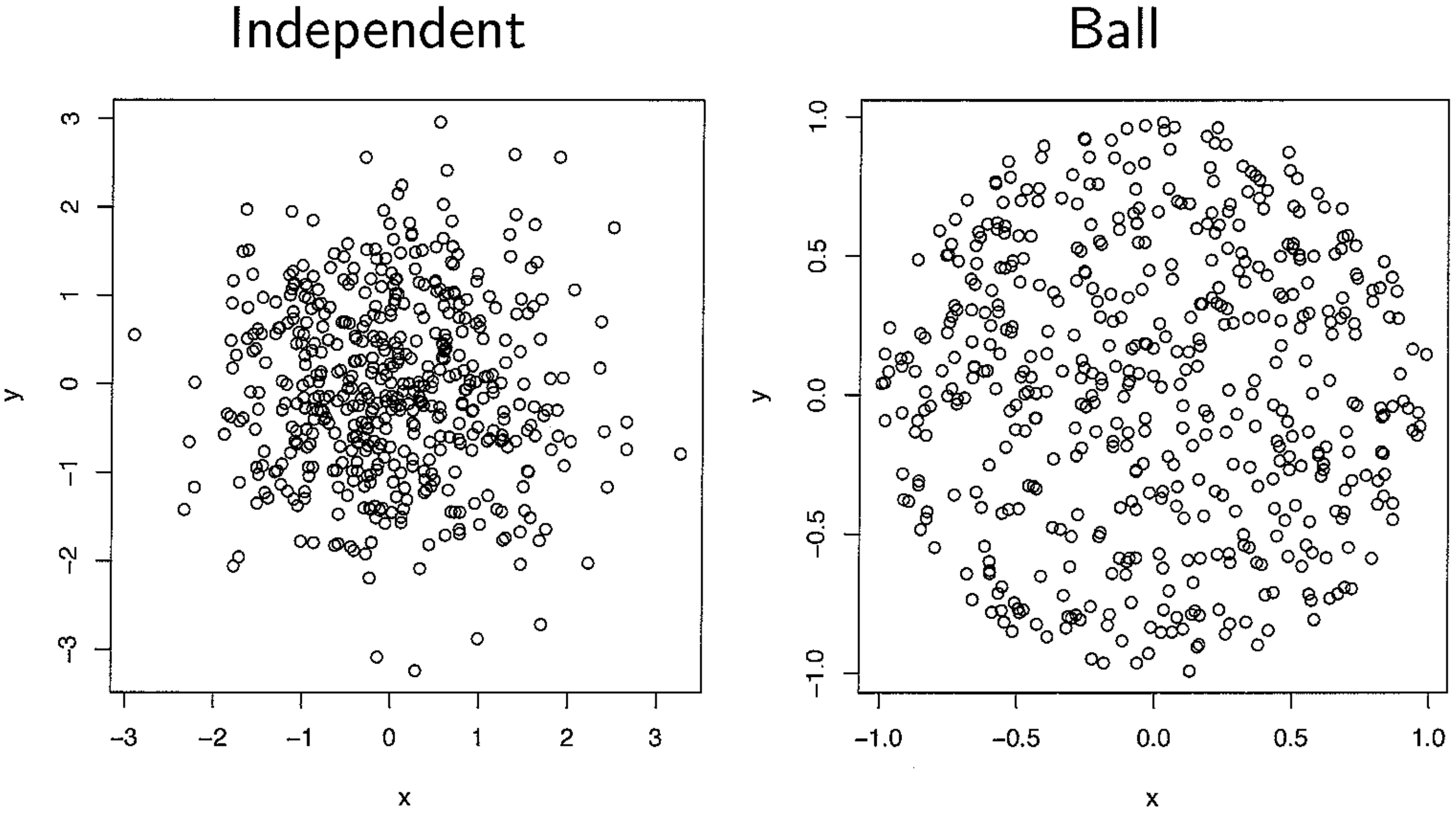


Perfect linear



Noisy linear

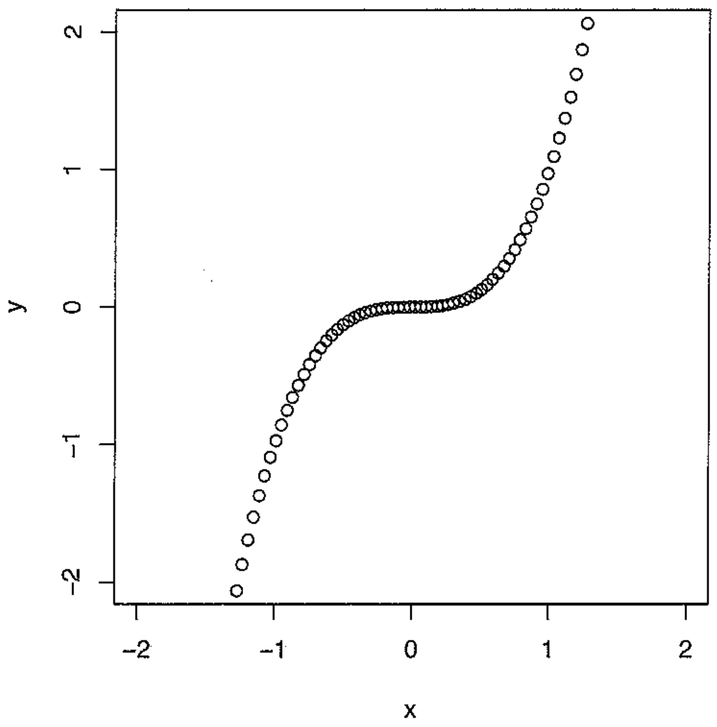
cor = 1.000 -0.866

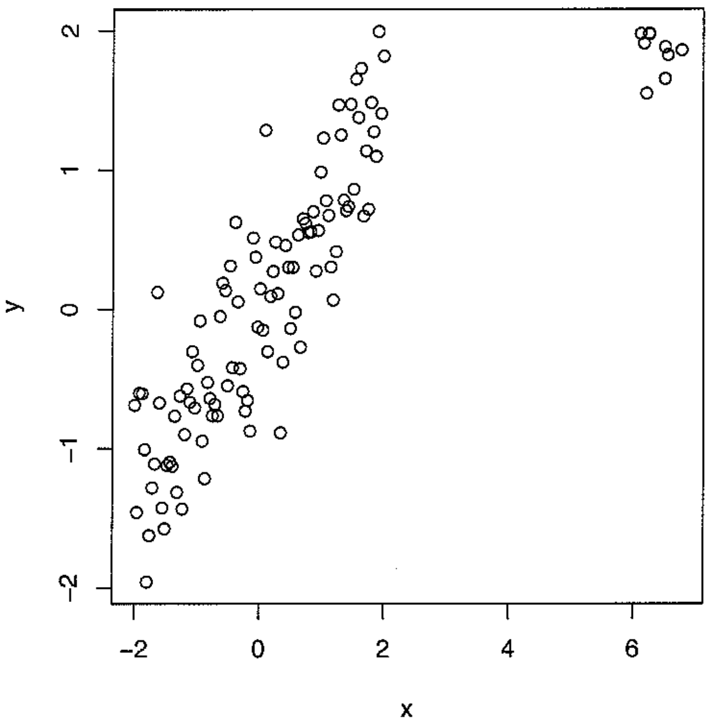


Independent

Ball

cor = -0.023 -0.029


Perfect cubic


Outliers

cor =	0.920	0.834

not robust

Rank correlation

Spearman's rank correlation is only defined in the sample. It goes beyond linearity, and measures a monotone association between $x, y \in \mathbb{R}^n$

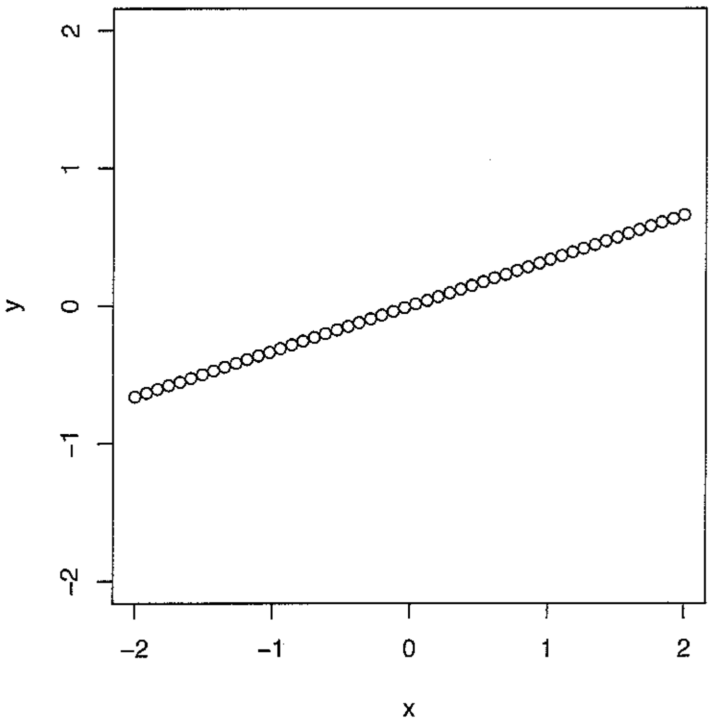
Given vectors $x, y \in \mathbb{R}^n$, we first define the rank vector $r_x \in \mathbb{R}^n$ that ranks the components of x, e.g., if $x = (0.7, 0.1, 1)$ then $r_x = (2, 1, 3)$. We also define the ranks r_y based on y. Rank correlation is now the usual (sample) correlation of r_x and r_y:

$$\underline{\mathrm{rcor}}(x, y) = \mathrm{cor}(r_x, r_y) \qquad \mathrm{cor}\left(\begin{pmatrix}2\\1\\3\end{pmatrix}, \begin{pmatrix}\cdot\\ \cdot\\ \cdot\end{pmatrix}\right)$$

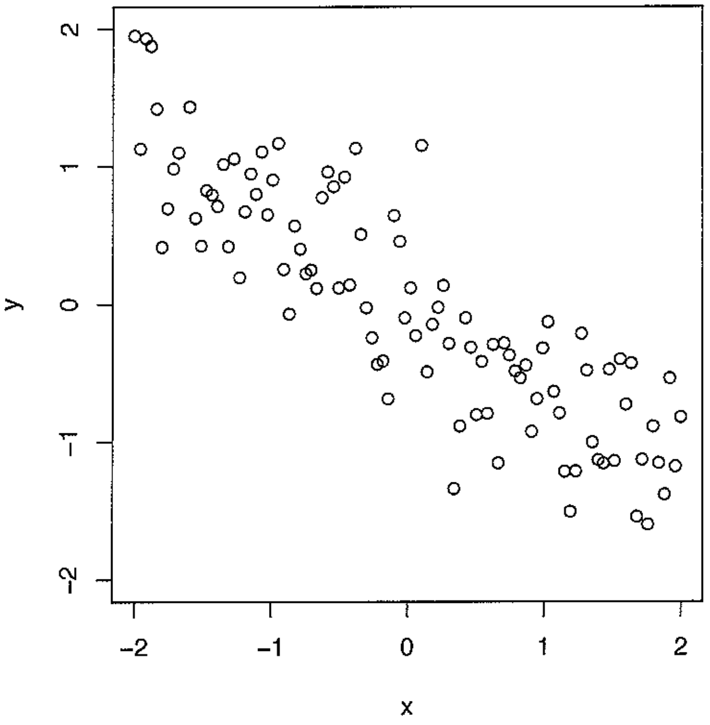
Key property: $|\mathrm{rcor}(x, y)| = 1$ if and only if there is a monotone function $f : \mathbb{R} \to \mathbb{R}$ such that $y_i = f(x_i)$ for each $i = 1, \ldots n$

Examples: rank correlation

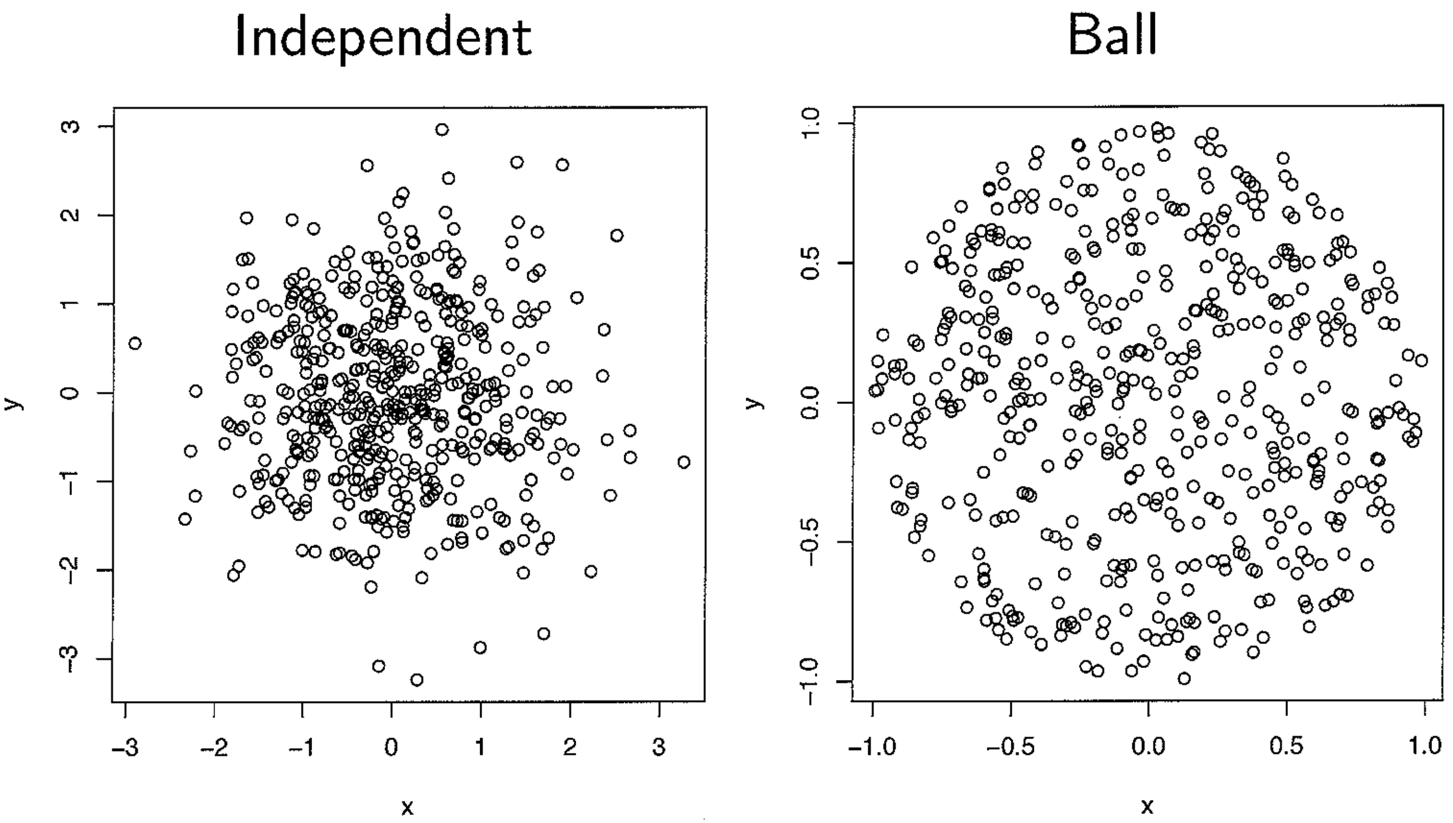
Perfect linear



Noisy linear

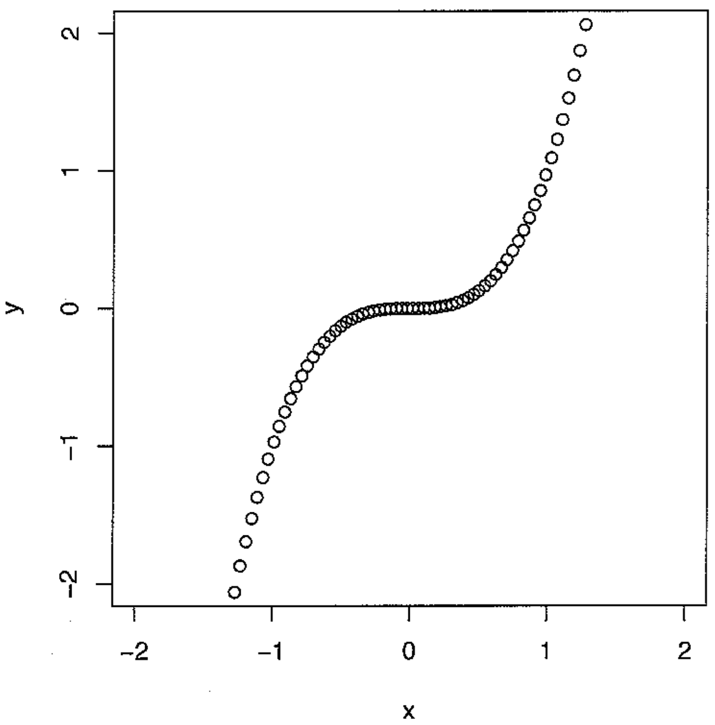


rcor =	1.000	-0.872
cor =	1.000	-0.866

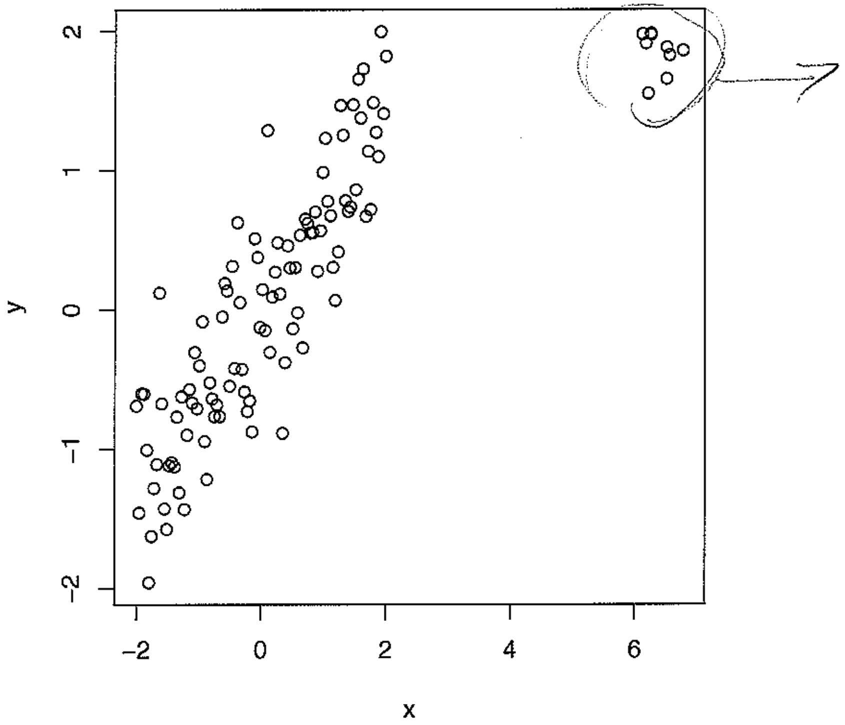


rcor =	-0.021	-0.033
cor =	-0.023	-0.029

Perfect cubic



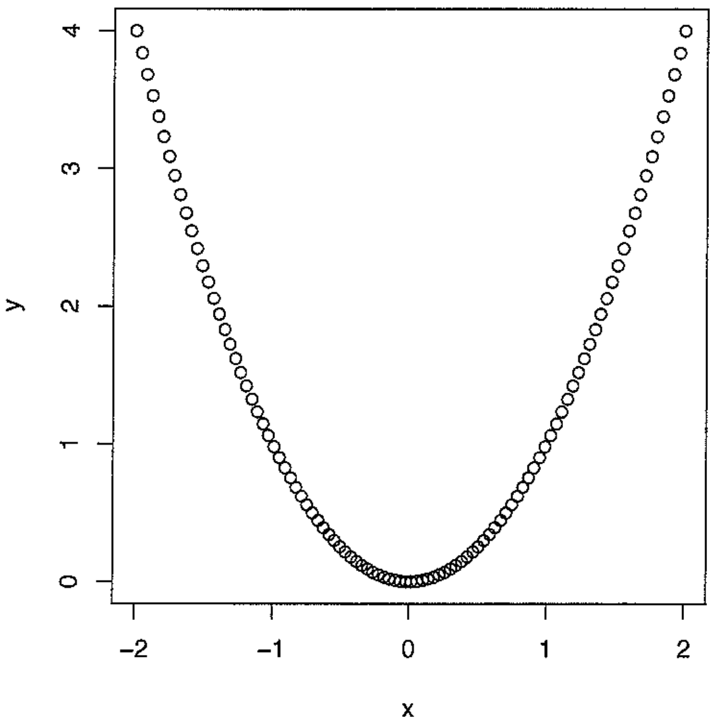
Outliers



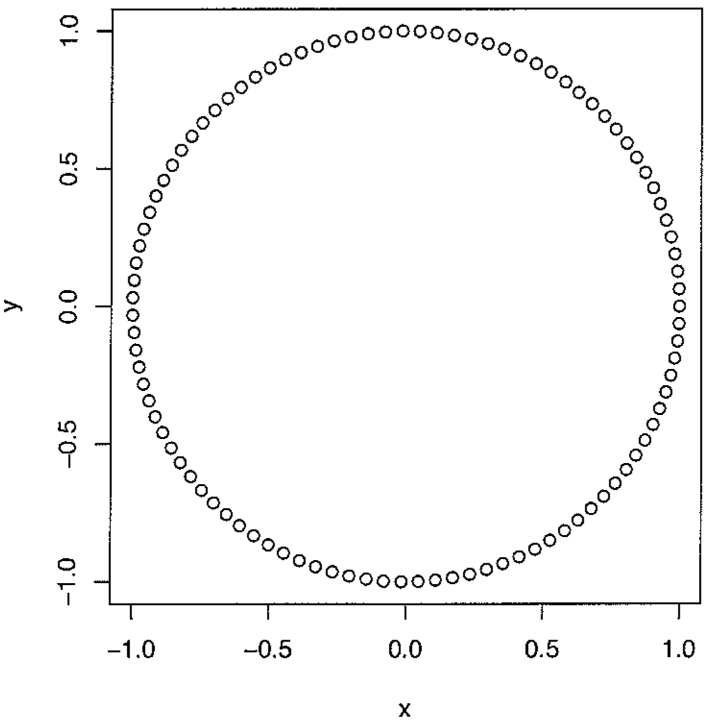
rcor =	1.000	0.905
cor =	0.920	0.834

x	-2	-1	0	1	2
$y = x^3$	-8	-1	0	1	8

Perfect quadratic



Perfect circle



rcor =	0.013	-0.001
cor =	0.000	0.000

Maximal correlation

Maximal correlation[1] is a notion of population correlation. It has no preference for linearity or monotonicity, and it characterizes independence completely

Given two random variables $X, Y \in \mathbb{R}$, the maximal correlation between X, Y is defined as

$$\mathrm{mCor}(X,Y) = \max_{f,g} \ \mathrm{Cor}(f(X), g(Y))$$

if $\mathrm{Cor}(f(X), g(Y)) = -1/2$ then $\mathrm{Cor}(-f(X), g(Y)) = 1/2$

where the maximum is taken over all functions[2] $f, g : \mathbb{R} \to \mathbb{R}$. Note that $0 \le \mathrm{mCor}(X,Y) \le 1$

Key property: $\mathrm{mCor}(X,Y) = 0$ if and only if X, Y are independent

[1] Gebelein (1947), "Das Statitistiche Problem Der Korrelation..."; Renyi (1959), "On Measures of Dependence"

[2] Actually, f, g have to be such that $\mathrm{Var}(f(X)) > 0, \mathrm{Var}(g(Y)) > 0$

can't have $f(X) = c$ $g(Y) = c$

Review: independence

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

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

If X, Y have densities f_X, f_Y, and (X, Y) has a density $f_{X,Y}$, then this is equivalent to

$$f_{X,Y}(s,t) = f_X(s) f_Y(t)$$

for any $s, t \in \mathbb{R}$. In other words, the joint density is the product of the marginal densities

Important fact: if X, Y are independent, then for any functions f, g, we have $\mathrm{E}[f(X)g(Y)] = \mathrm{E}[f(X)]\mathrm{E}[g(Y)]$

Hence X, Y independent implies that $\mathrm{Cor}(f(X), g(Y)) = 0$ for any functions f, g, which means $\mathrm{mCor}(X,Y) = 0$

$$\text{Cov}(f(X), g(Y)) / (\cdot)$$

$$= \frac{\cancel{E[f(X)g(Y)] - E[f(X)]E[g(Y)]}}{(\cdot\ )}$$

Review: characteristic functions

The characteristic function of a random variable $X \in \mathbb{R}$ is the function $h_X : \mathbb{R} \to \mathbb{C}$ defined as

imaginary number i

$$h_X(t) = \mathrm{E}[\exp(itX)] = E[e^{itX}]$$

The characteristic function of a random vector $(X, Y) \in \mathbb{R}^2$ is $h_{X,Y} : \mathbb{R}^2 \to \mathbb{C}$,

$$h_{X,Y}(s,t) = \mathrm{E}\big[\exp\big(i(sX + tY)\big)\big]$$

Characteristic functions completely characterize the distribution of a random variable (hence the name). I.e., if $h_X(t) = h_{X'}(t)$ for every $t \in \mathbb{R}$, then X and X' must have the same distribution

Important fact: X, Y are independent if and only if $\underline{h_{X,Y}(s,t) = h_X(s)h_Y(t)}$ for any $s, t \in \mathbb{R}$. In other words, the joint characteristic function is the product of the marginal ones

Zero maximal correlation implies independence

Suppose that $\mathrm{mCor}(X,Y)=0$. Then taking[3] $f(X)=\exp(isX)$ and $g(Y)=\exp(itY)$ we have

fix s,t

$$\mathrm{Cor}(\exp(isX),\exp(itY))=0$$

This means that

$$\mathrm{Cov}(\underset{u}{\exp(isX)},\underset{v}{\exp(itY)})=0$$

$\mathrm{Cov}(u,v) = E[uv] - E[u]E[v]$

i.e.,

$$\underbrace{\mathrm{E}[\exp(isX)\exp(itY)]}_{\phi_{X,Y}(s,t)}=\underbrace{\mathrm{E}[\exp(isX)]}_{\phi_X(s)}\cdot\underbrace{\mathrm{E}[\exp(itY)]}_{\phi_Y(t)}$$

This holds for each $s,t\in\mathbb{R}$, so X,Y are independent

[3] Strictly speaking, f,g are supposed to take values in $\mathbb{R}$ (not $\mathbb{C}$), but to get around this we can write $\exp(i\theta)=\cos\theta+i\sin\theta$ and then break things up into real and imaginary parts

Maximal correlation in the sample

Given two vectors $x, y \in \mathbb{R}^n$, and functions $f, g : \mathbb{R} \to \mathbb{R}$, write $f(x) \in \mathbb{R}^n$ for the vector with ith component $f(x_i)$, i.e., $f(x) = (f(x_1), \ldots f(x_n))$, and similarly for $g(y)$

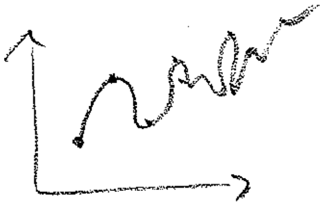
We'd like to define maximal correlation in the sample, analogous to its population definition (so that we can use it in practice!). Consider

$$\mathrm{mcor}(x, y) = \max_{f,g} \ \mathrm{cor}(f(x), g(y))$$

$$\begin{matrix} y_1, \cdots & y_n \\ \shortparallel & \shortparallel \\ f(x_1), \cdots & f(x_n) \end{matrix}$$

There's a big problem here—as defined, $\mathrm{mcor}(x, y) = 1$ for any $x, y \in \mathbb{R}^n$. (Why?)

We'll derive an algorithm to compute mCor in the population. This inspires an algorithm in the sample, and the sample version mcor is then defined as the output of this algorithm



Fixed points of maximal correlation

We define a norm on random variables $Z \in \mathbb{R}$ by $\|Z\| = \sqrt{\mathrm{E}[Z^2]}$

Given any function f, we can always make $f(Z)$ have the following properties:

- $\mathrm{E}[f(Z)] = 0$, by letting $f(Z) \leftarrow f(Z) - \mathrm{E}[f(Z)]$ (centering)
- $\|f(Z)\| = 1$, by letting $f(Z) \leftarrow f(Z)/\|f(Z)\|$ (scaling) $\sqrt{E[f(z)^2]}$

Notice that for such an f, we have $\mathrm{Var}(f(Z)) = \|f(Z)\|^2 = 1$

Therefore, when computing maximal correlation, we can restrict our attention to functions f, g such that $\mathrm{E}[f(X)] = \mathrm{E}[g(Y)] = 0$ and $\|f(X)\| = \|g(Y)\| = 1$. (Otherwise, we simply center and scale as needed, and this doesn't change the correlation.) Hence

$$\mathrm{mCor}(X,Y) = \max_{\substack{\mathrm{E}[f(X)]=\mathrm{E}[g(Y)]=0 \\ \|f(X)\|=\|g(Y)\|=1}} \mathrm{E}[f(X)g(Y)]$$

Now notice that for such f, g,

$$\begin{aligned}\underbrace{\mathrm{E}[(f(X)-g(Y))^2]}_{\nearrow\ \text{min}} &= \overset{1}{\mathrm{E}[f^2(X)]} + \overset{1}{\mathrm{E}[g^2(Y)]} - 2\mathrm{E}[f(X)g(Y)] \\ &= 2 - \underbrace{2\mathrm{E}[f(X)g(Y)]}_{} \leftarrow \text{max}\end{aligned}$$

and hence f, g are optimal functions for mCor if and only if they are also optimal for the minimization problem

$$\min_{\substack{\mathrm{E}[f(X)]=\mathrm{E}[g(Y)]=0 \\ \|f(X)\|=\|g(Y)\|=1}} \mathrm{E}[(f(X)-g(Y))^2] = \int \mathrm{E}[(f(X)-g(y))^2 | Y=y]\, f_Y(y)\,dy$$

Key point: consider just minimizing over the function g (assuming that f is fixed and satisties $\mathrm{E}[f(X)] = 0$ and $\|f(X)\| = 1$). To do so, we can minimize the conditional expectation

$$\rightarrow \mathrm{E}[(f(X)-g(y))^2|Y=y] \leftarrow$$

for each fixed y. And to minimize this over $g(y)$, we simply take $g(y) = \mathrm{E}[f(X)|Y=y]$

$\mathrm{E}[(W-a)^2]$ take $a = \mathrm{E}[W]$ to make the smallest

As a function over Y, this is written as $g(Y) = \mathrm{E}[f(X)|Y]$

Remember though, that we are restricting g to have expectation zero and norm one. Therefore we let

$$g(Y) = \underbrace{\mathrm{E}[f(X)|Y]}/\|\mathrm{E}[f(X)|Y]\|$$

fix f
Solve for g ...

(Check: this satisfies both properties)

The exact same arguments, but in reverse, show that minimizing over f for fixed g yields

$$f(X) = \mathrm{E}[g(Y)|X]/\|\mathrm{E}[g(Y)|X]\|$$

fix g
solve for f ..

These are called the fixed point equations of the maximal correlation problem. That is, if there exists f, g that achieved the maximum $\mathrm{mCor}(X, Y) = \mathrm{E}[f(X)g(Y)]$, then they must satisfy the above equations

Alternating conditional expectations algorithm

The alternating conditional expectations (ACE) algorithm[4] is motivated directly from these fixed point equations. The idea is just to start with a guess at one of the functions, and then iterate the two equations until convergence

ACE algorithm:

- Set $f_0(X) = (X - \mathrm{E}[X])/\cancel{\|X\|}$ $\|X - E[X]\|$
- For $k = 1, 2, 3, \ldots,$
  1. Let $g_k(Y) = \mathrm{E}[f_{k-1}(X)|Y]/\|\mathrm{E}[f_{k-1}(X)|Y]\|$
  2. Let $f_k(X) = \mathrm{E}[g_k(Y)|X]/\|\mathrm{E}[g_k(Y)|X]\|$
  3. Stop if $\mathrm{E}[f_k(X)g_k(Y)] = \mathrm{E}[f_{k-1}(X)g_{k-1}(Y)]$
- Upon convergence, $\mathrm{mCor}(X, Y) = \mathrm{E}[f_k(X)g_k(Y)]$

This has been proven to converge under very general assumptions

[4] Breiman and Friedman (1985), "Estimating Optimal Transformations for Multiple Regression and Correlation"

Back to maximal correlation in the sample

Given $x, y \in \mathbb{R}^n$. As in the population case, if we are considering $\mathrm{cor}(f(x), g(y))$ over all functions f, g, we can restrict out attention to functions f, g with

$$\mathbb{1}^T f(x) = \mathbb{1}^T g(y) = 0 \text{ and } \|f(x)\|_2 = \|g(y)\|_2 = 1$$

where now $\|\cdot\|_2$ is the usual Euclidean norm. For such functions f, g, we have $\mathrm{cor}(f(x), g(y)) = f(x)^T g(y)$. Further,

$$\begin{aligned} \|f(x) - g(y)\|_2^2 &= \|f(x)\|_2^2 + \|g(y)\|_2^2 - 2f(x)^T g(y) \\ &= 2 - 2\underbrace{f(x)^T g(y)}_{cor} \end{aligned}$$

for such functions f, g, so maximizing $\mathrm{cor}(f(x), g(y))$ is the same as minimizing $\|f(x) - g(y)\|_2^2$

We're going to derive a sample version of the ACE algorithm to approximately minimize $\|f(x) - g(y)\|_2^2$, and then we define its output as the sample maximal correlation

The ACE algorithm in the sample

We define the ACE algorithm analogously to its definition in the population case. That is, we repeat iterations of the form

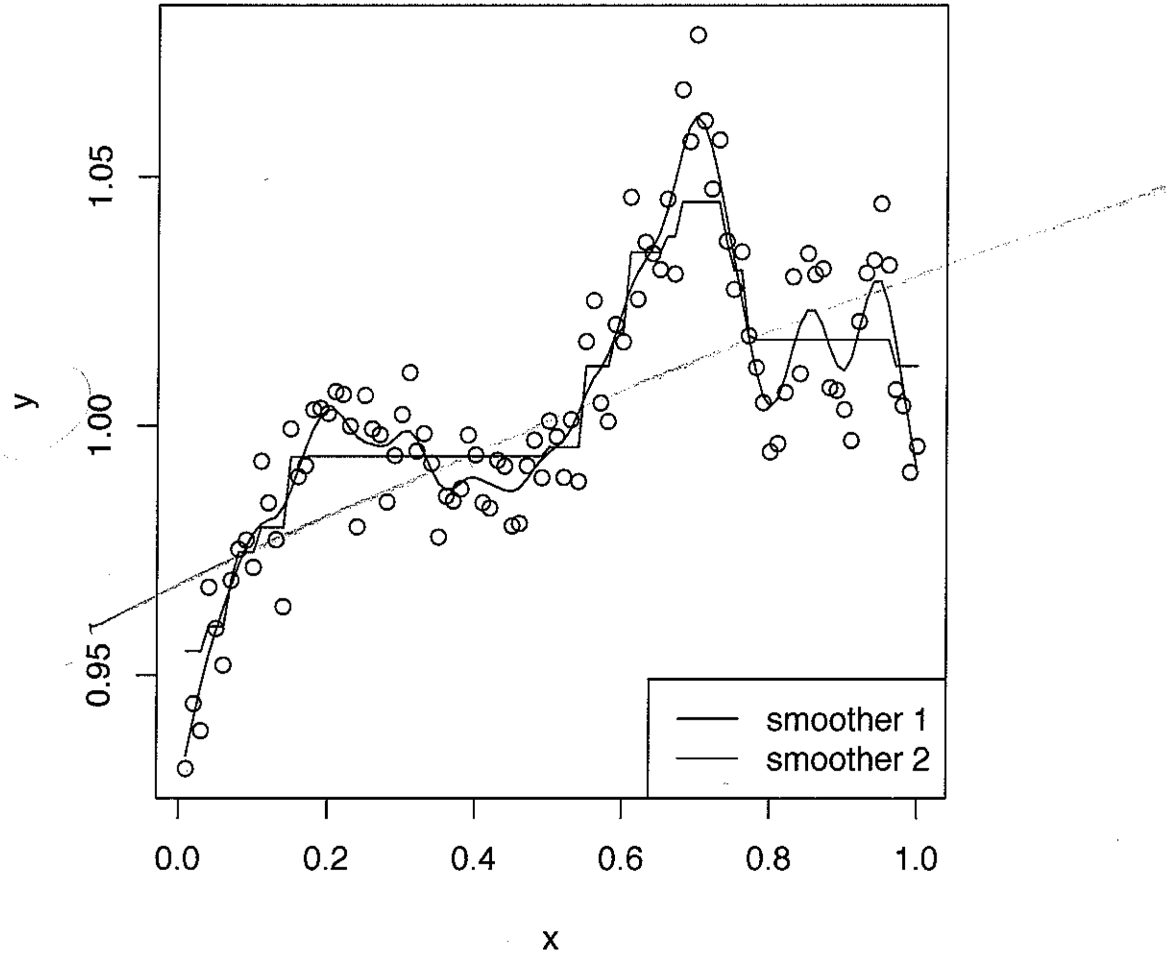
$$g(y) = \mathrm{condexp}[f(x)|y], \text{ center and scale } g(y),$$
$$f(x) = \mathrm{condexp}[g(y)|x], \text{ center and scale } f(x),$$

where $\mathrm{condexp}[\,\cdot\,|x]$ denotes a sample version of the conditional expectation on x, and similary for $\mathrm{condexp}[\,\cdot\,|y]$

A good question is: how do we compute these sample conditional expectations? (There's not a simple convenient sample analog like there is for unconditional expectation, variance, or covariance)

Answer: use a smoother $\mathcal{S}$. We use the notation $\mathcal{S}(y|x)$ to mean that the result is an estimate for $y = (y_1, \ldots y_n)$ as a function of $x = (x_1, \ldots x_n)$

A smoother $\mathcal{S}(y|x)$ could be, e.g., something as simple as a histogram or linear regression. It could also be something more fancy like kernel regression or local linear regression. (We'll see examples of smoothers later in the course)



Assume that we've picked a smoother $\underline{\mathcal{S}}$ to use in the ACE algorithm. (We'd like a smoother that can fit very general shapes of functions, i.e., not just a histogram or linear regression)

ACE algorithm:

- Set $f_0(x) = (x - \bar{x}\mathbb{1})/\|\cancel{x}\|_2$ $\quad \|x - \bar{x}\,\mathbb{1}\|_2$
- For $k = 1, 2, 3, \ldots$
  1. Let $G(y) = \mathcal{S}(f_{k-1}(x)|y)$, and center and scale, $g_k(y) = (G(y) - \overline{G(y)}\mathbb{1})/\|G(y) - \overline{G(y)}\mathbb{1}\|_2$
  2. Let $F(x) = \mathcal{S}(g_k(y)|x)$, and center and scale, $f_k(x) = (F(x) - \overline{F(x)}\mathbb{1})/\|F(x) - \overline{F(x)}\mathbb{1}\|_2$
  3. Stop if $|f_k(x)^T g_k(y) - f_{k-1}(x)^T g_{k-1}(y)|_2$ is small
- Upon convergence, define $\underline{\mathrm{mcor}(x, y)} = \underline{f_k(x)^T g_k(y)}$

Unforunately, the ACE algorithm in the sample is only guaranteed to converge under some restrictive conditions. But in practice, it still tends to perform quite well

Measures of correlation in R

The `cor` function available in the base R distribution can be used to compute (Pearson's) correlation, and also (Spearman's) rank correlation. E.g.,

```
cor(x, y) # default is method="pearson"
cor(x, y, method="spearman") # rank correlation
```

The function `ace` in the package `acepack` implements the alternating conditional expectations algorithm to compute maximum correlation. E.g.,

```
a = ace(x, y)
cor(a$tx, a$ty) # maximal correlation
```

Note: this `ace` implementation doesn't scale the vectors in the way that discussed in class, so the maximal correlation is not simply `sum(a$tx * a$ty)`

Recap: measures of correlation

In this lecture we reviewed the basic facts about correlation, both in the population (random variables) and in the sample (vectors). If random variables $X, Y \in \mathbb{R}$ are jointy normal, then independence and uncorrelatedness are the same thing

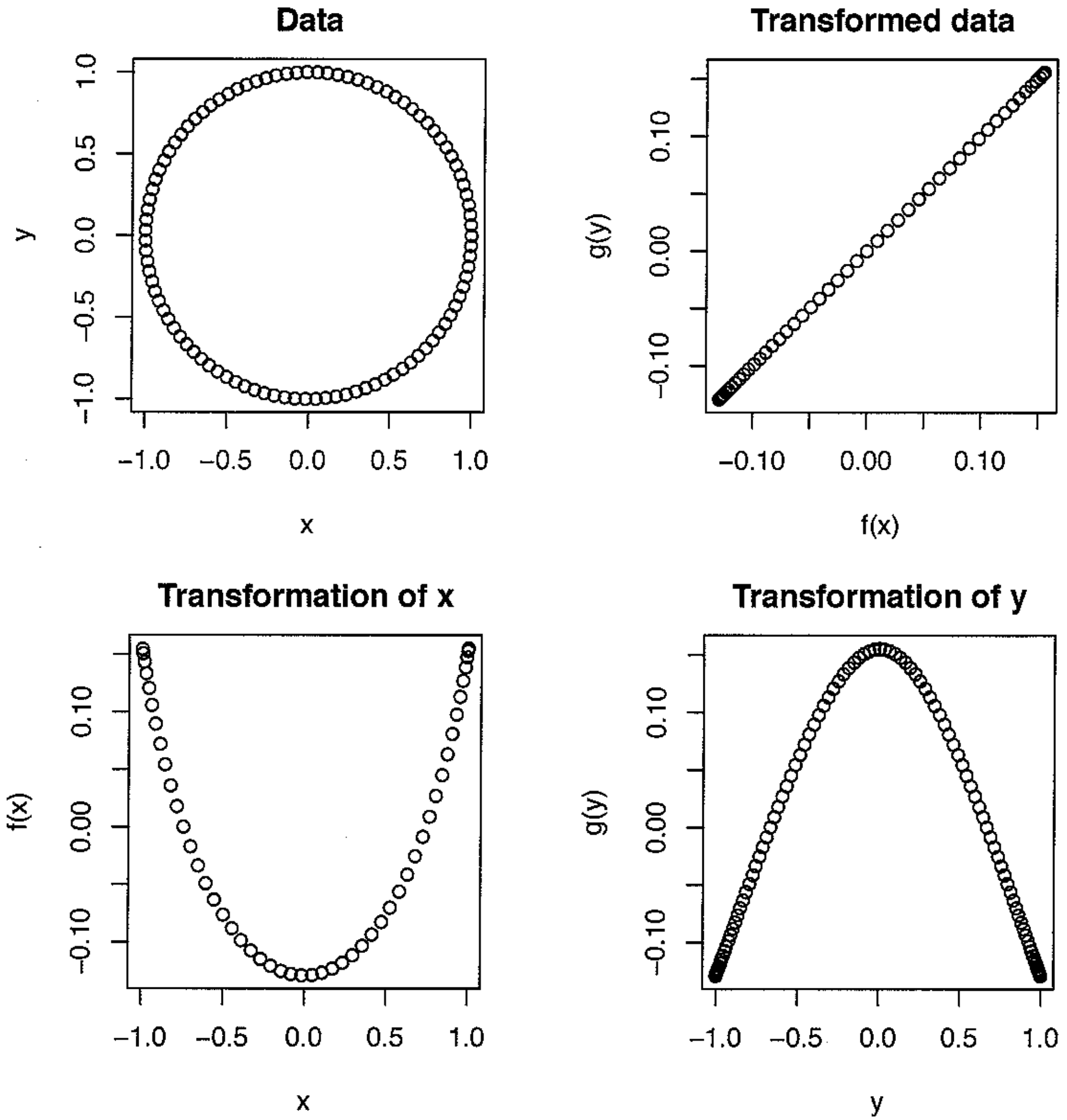
Rank correlation is simply the usual sample correlation, except replacing vectors $x, y \in \mathbb{R}^n$ with the ranks of their components $r_X, r_Y \in \mathbb{R}^n$. This aims to capture monotone associations that are not necessarily linear

Maximal correlation is defined in the population as the maximum correlation over functions of our two random variables. The maximal correlation equals zero if and only if the random variables are independent

The alternating conditional expectations (ACE) algorithm is an elegant way to compute maximal correlation in the population, and furthermore, it allows to define maximal correlation in the sample

Next time: more measures of correlation

More about maximal correlation; distance correlation


Data
y
x
Transformed data
g(y)
f(x)
Transformation of x
f(x)
x
Transformation of y
g(y)
y