Correlation analysis 3: Measures of correlation (continued)

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

February 21 2013

Reminder: correlation, rank correlation

Last time we learned about correlation. 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 $x, y \in \mathbb{R}^n$,

$$\mathrm{cor}(x,y) = \frac{\mathrm{cov}(x,y)}{\sqrt{\mathrm{var}(x)}\sqrt{\mathrm{var}(y)}} = x^T y$$

the second equality holding if x, y have been centered and scaled

Rank correlation is only defined in the sample: given $x, y \in \mathbb{R}^n$, let $r_x, r_y \in \mathbb{R}^n$ denote the ranks of x, y, respectively, and then

$$\mathrm{rcor}(x,y) = \mathrm{cor}(r_x, r_y)$$

Reminder: maximal correlation

Maximal correlation is only defined in the population: for random variables $X, Y \in \mathbb{R}$,

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

the maximum taken over all functions f, g. We were able to show that the optimal f, g are also optimal for the problem

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

where $\|Z\| = \sqrt{\mathrm{E}[Z^2]}$. The fixed points of this problem,

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

suggested an alternating algorithm for finding mCor

Reminder: alternating conditional expectations algorithm

The alternating conditional expectations (ACE) algorithm can be used to compute mCor. Algorithm:

- Set $f_0(X) = (X - \mathrm{E}[X])/\|X - \mathrm{E}[X]\|$
- 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, $\mathrm{mCor}(X, Y) = \mathrm{E}[f_k(X)g_k(Y)]$

Maximal correlation isn't well-defined in the sample, because for any vectors $x, y \in \mathbb{R}^n$,

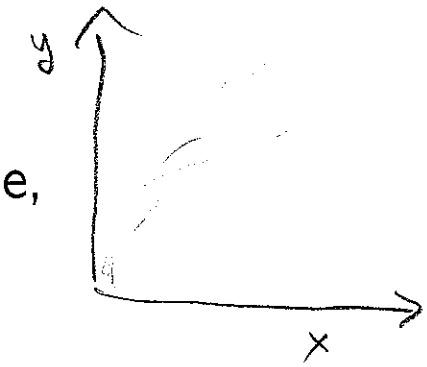
$$\max_{f,g}\ \mathrm{cor}(f(x), g(y)) = 1$$

But the ACE algorithm can be adapted to the sample and we define its output to be the sample maximal correlation mcor

Reminder: ACE algorithm in the sample

The sample version of the ACE algorithm requires us to pick a smoother $\mathcal{S}$ to approximate the conditional expectations (e.g., kernel regression or local linear regression). Algorithm:

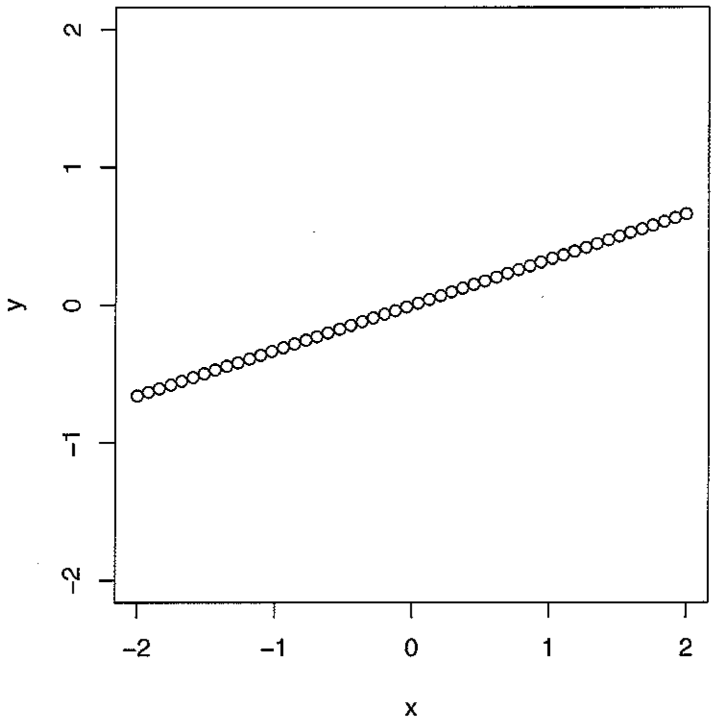
- Set $f_0(x) = (x - \bar{x}\mathbb{1})/\|x - \bar{x}\mathbb{1}\|_2$
- For $k = 1, 2, 3, \ldots$
  1. Let $G(y) = \mathcal{S}(f_{k-1}(x)|y)$, and center and scale,
     $g_k(y) = (G(y) - \overline{G(y)}\mathbb{1})/\|G(y) - \overline{G(y)}\mathbb{1}\|_2$
  2. Let $F(x) = \mathcal{S}(g_k(y)|x)$, and center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



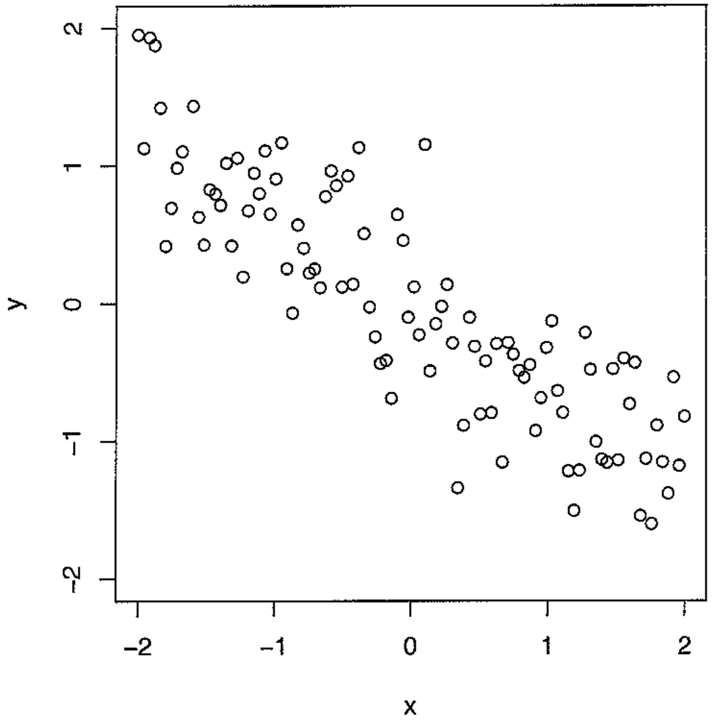
This not always guaranteed to converge, and the answer depends on the smoother. But most of the time it works well in practice

Example: maximal correlation

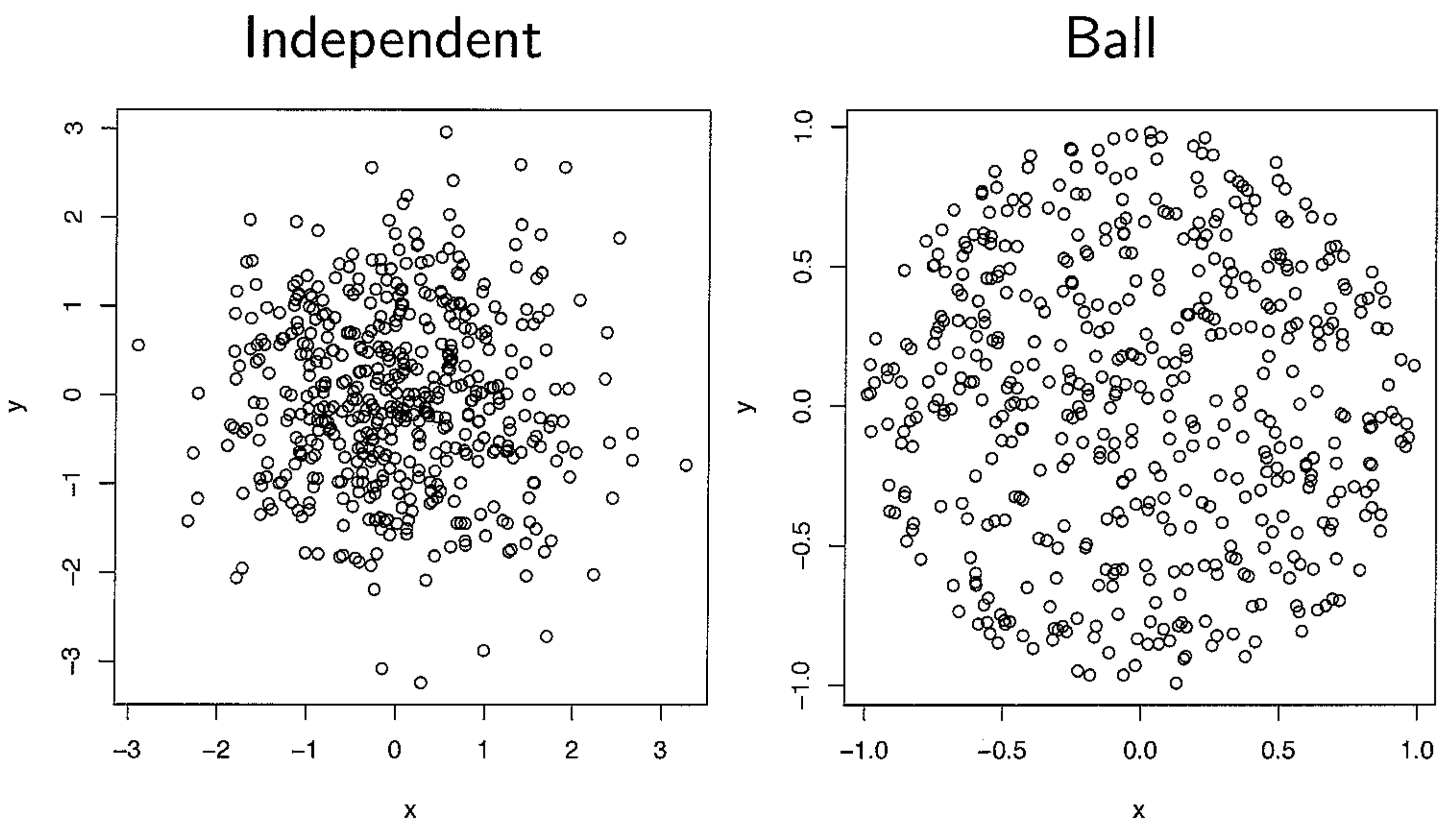
Perfect linear



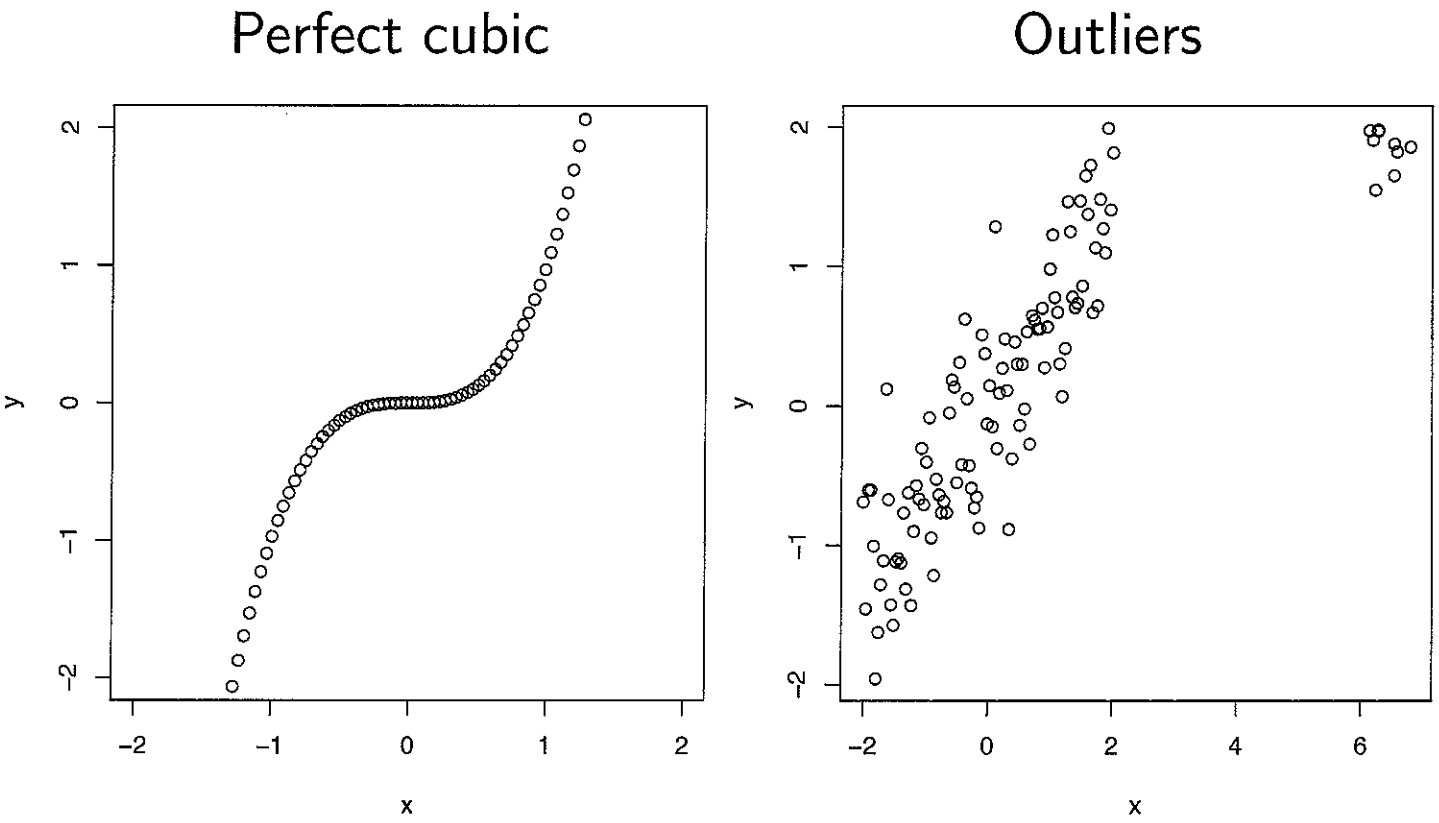
Noisy linear



	Perfect linear	Noisy linear
mcor =	1.000	0.896
rcor =	1.000	-0.872
cor =	1.000	-0.866



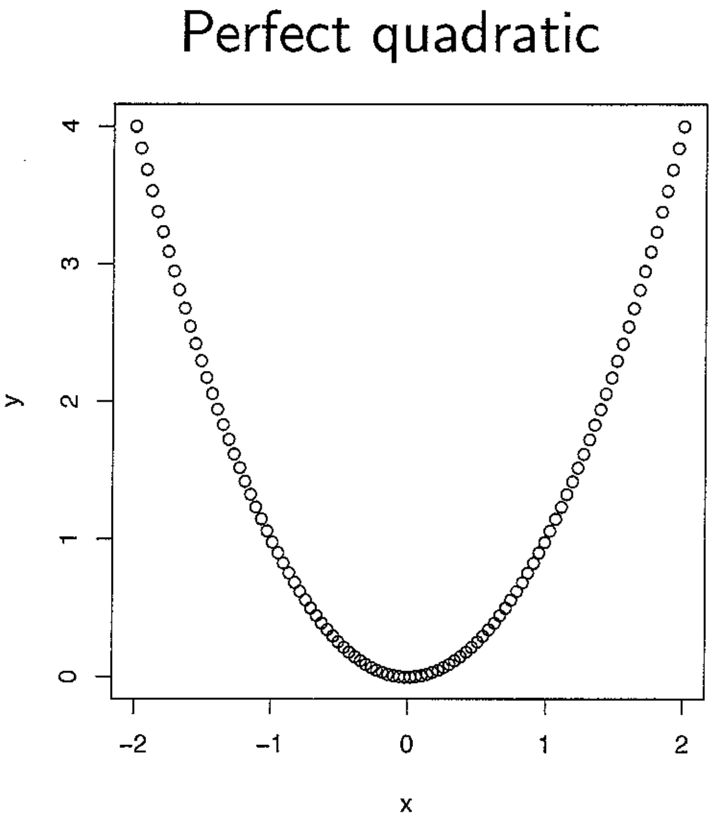
mcor =	0.124	0.316
rcor =	-0.021	-0.033
cor =	-0.023	-0.029

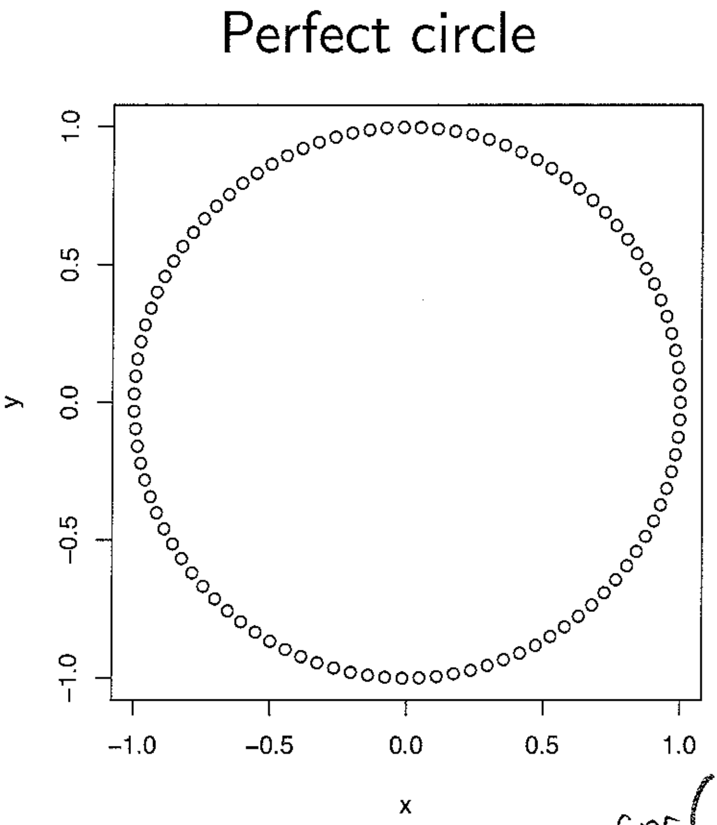


Perfect cubic

Outliers

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


Perfect quadratic

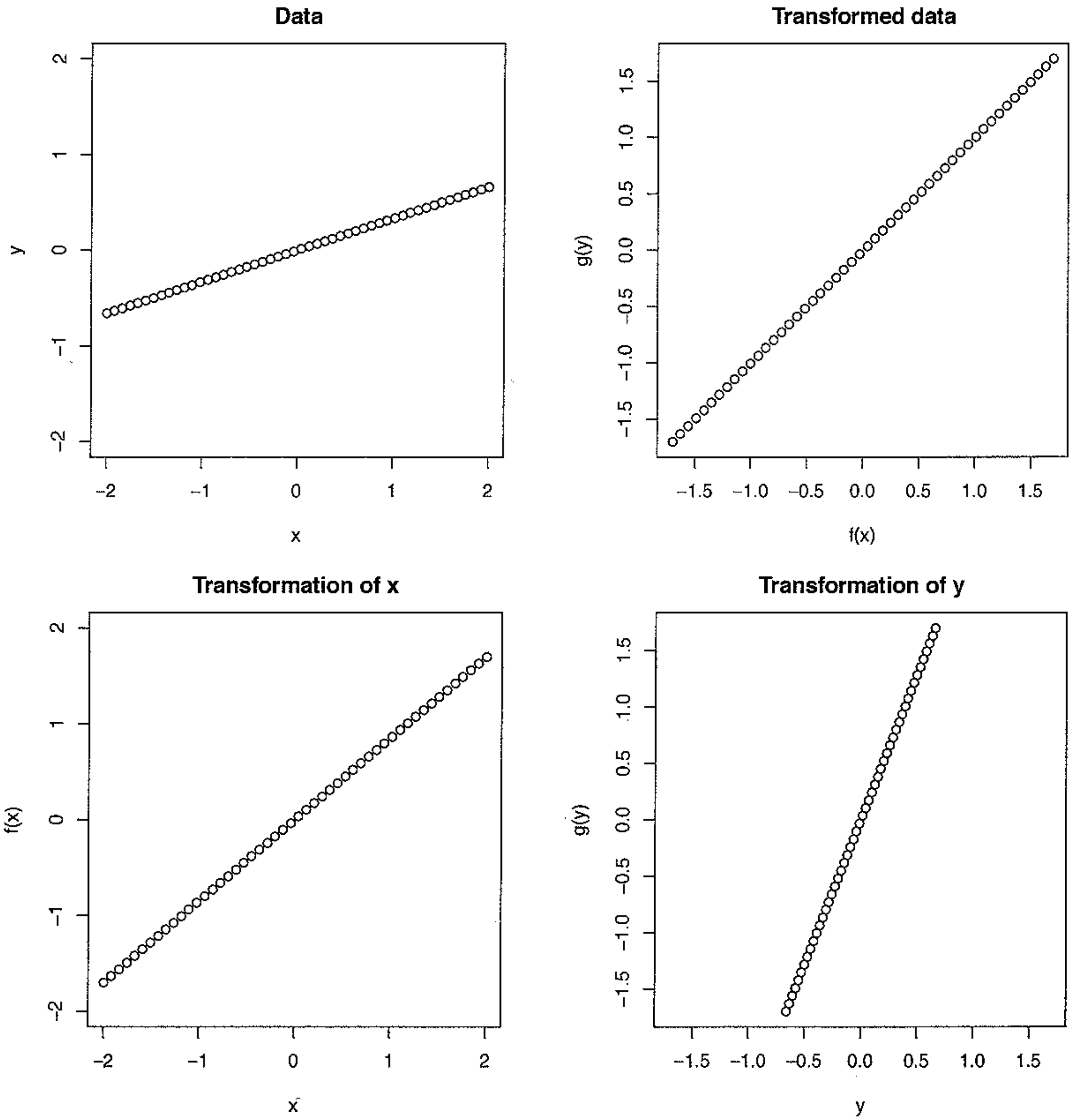

Perfect circle

$cor(f(x), g(y))$

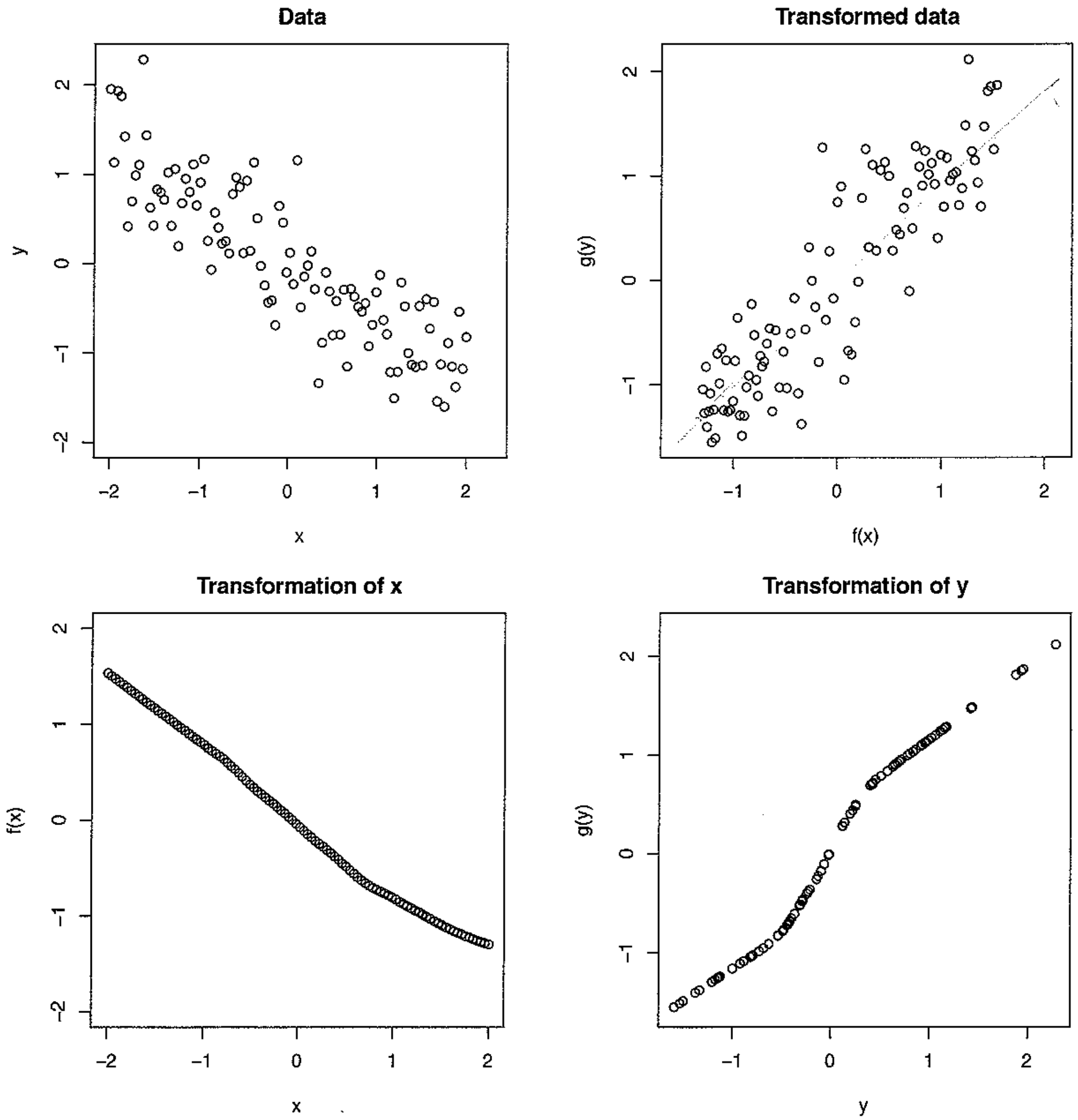
mcor =	1.000	1.000
rcor =	0.013	-0.001
cor =	0.000	0.000

Example: ACE algorithm

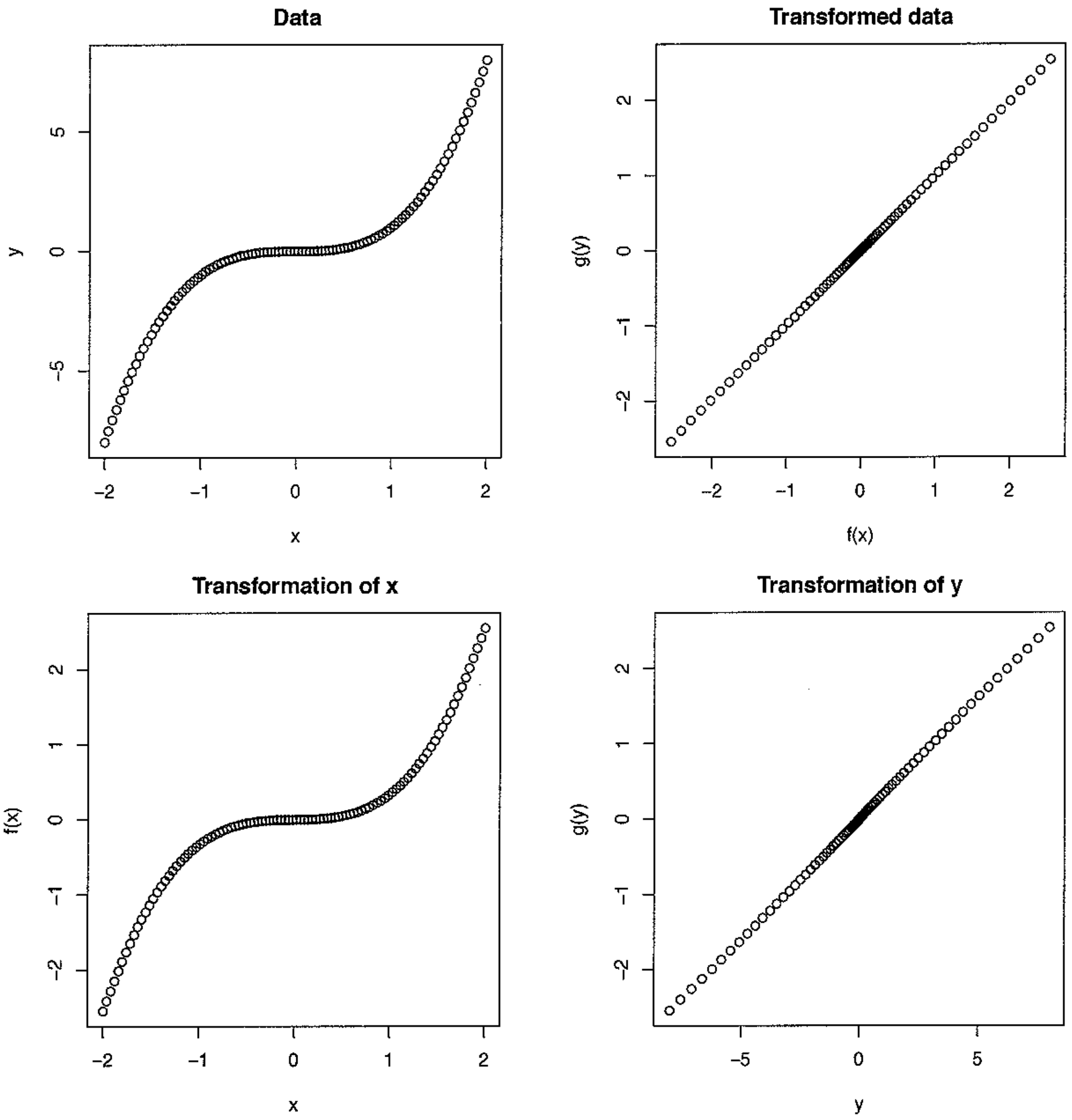
Perfect linear



Noisy linear



Perfect cubic


Data

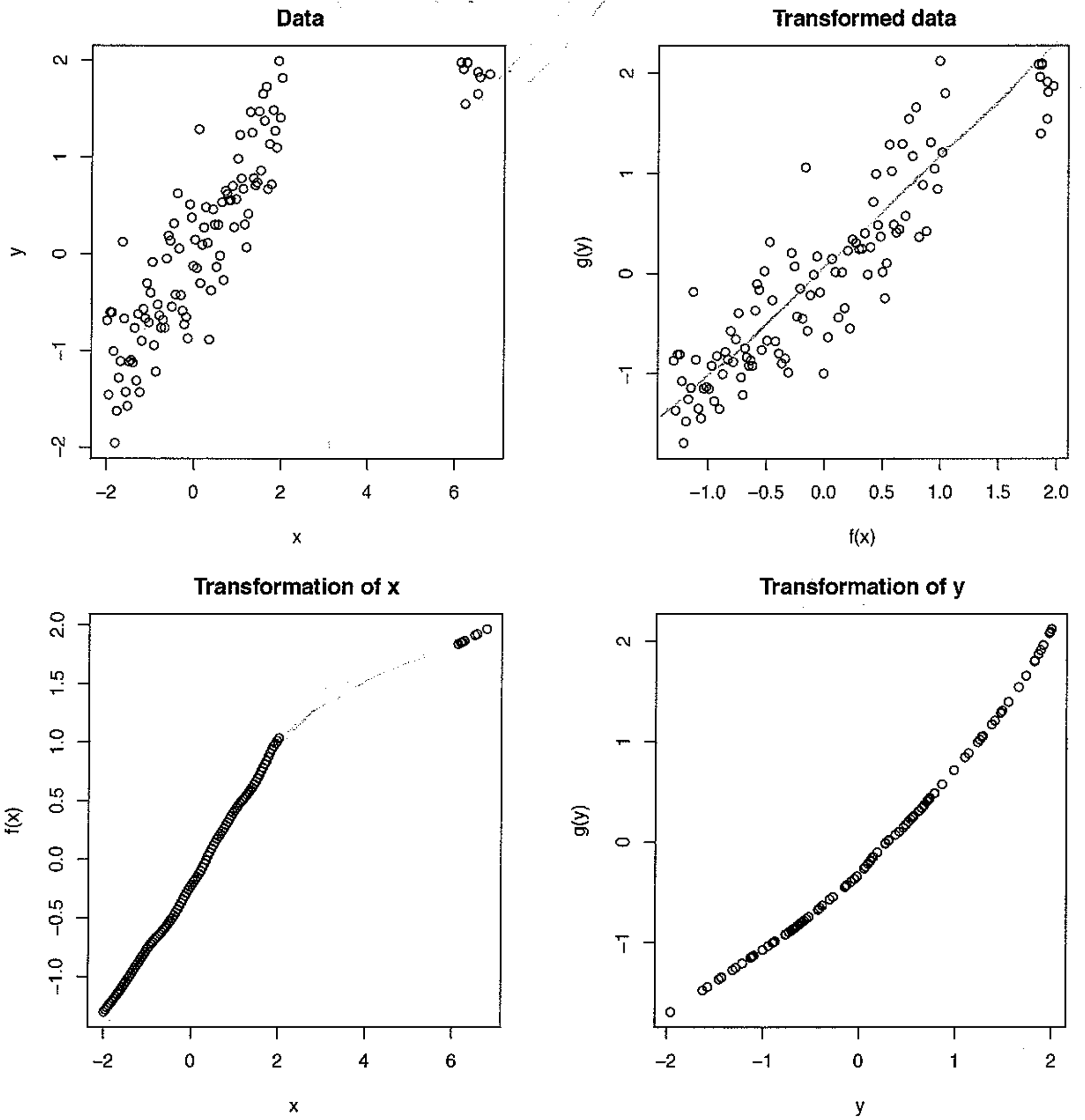
Transformed data

Transformation of x

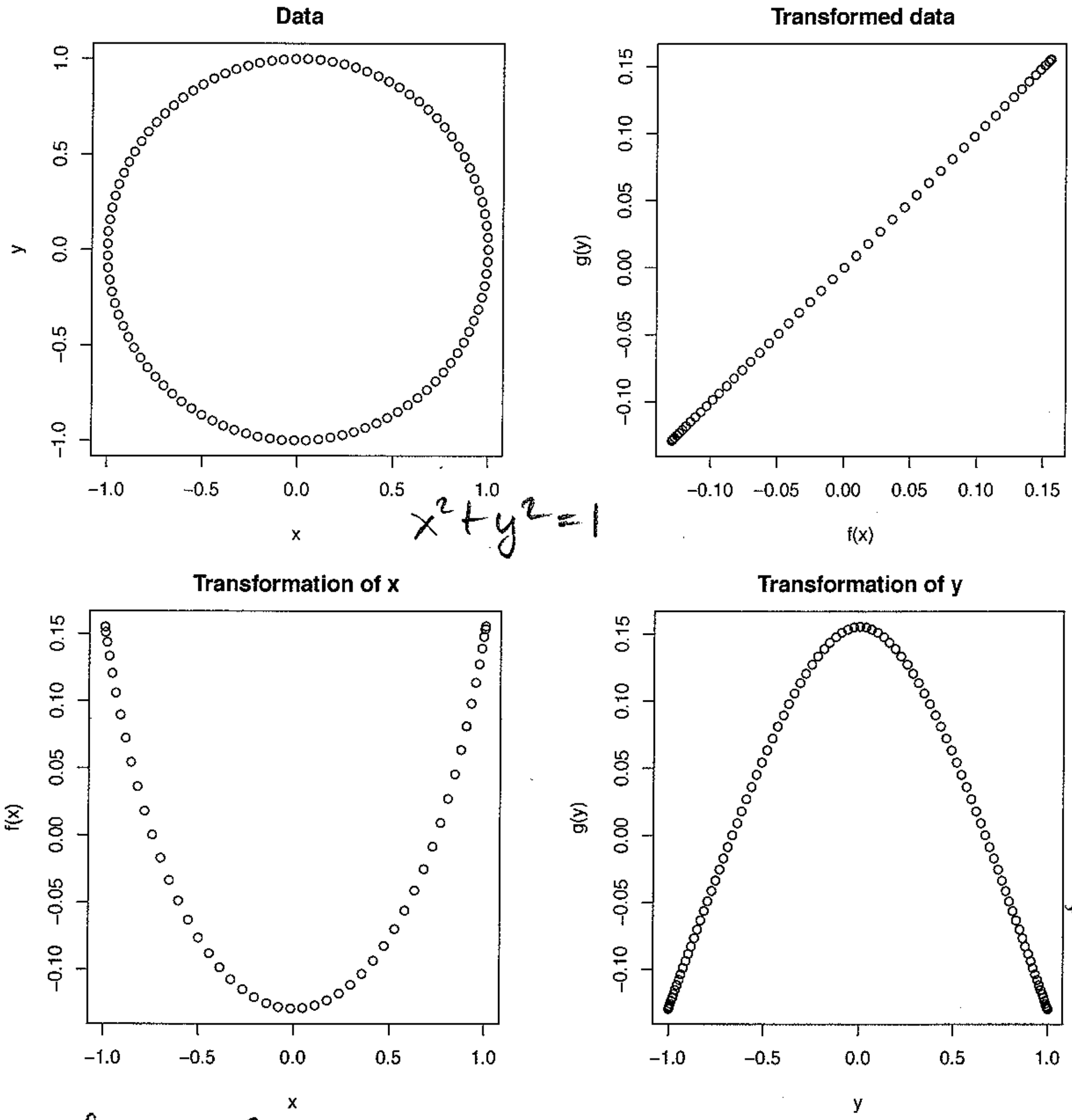
Transformation of y

x^3

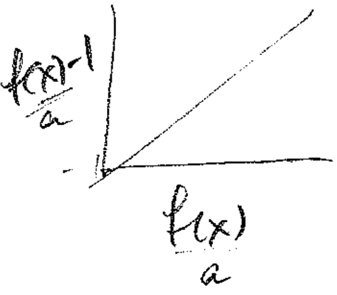
Outliers


Data
Transformed data
Transformation of x
Transformation of y
y
x
g(y)
f(x)
−2
−1
0
1
2
4
6
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0

Perfect circle



$x^2+y^2=1$

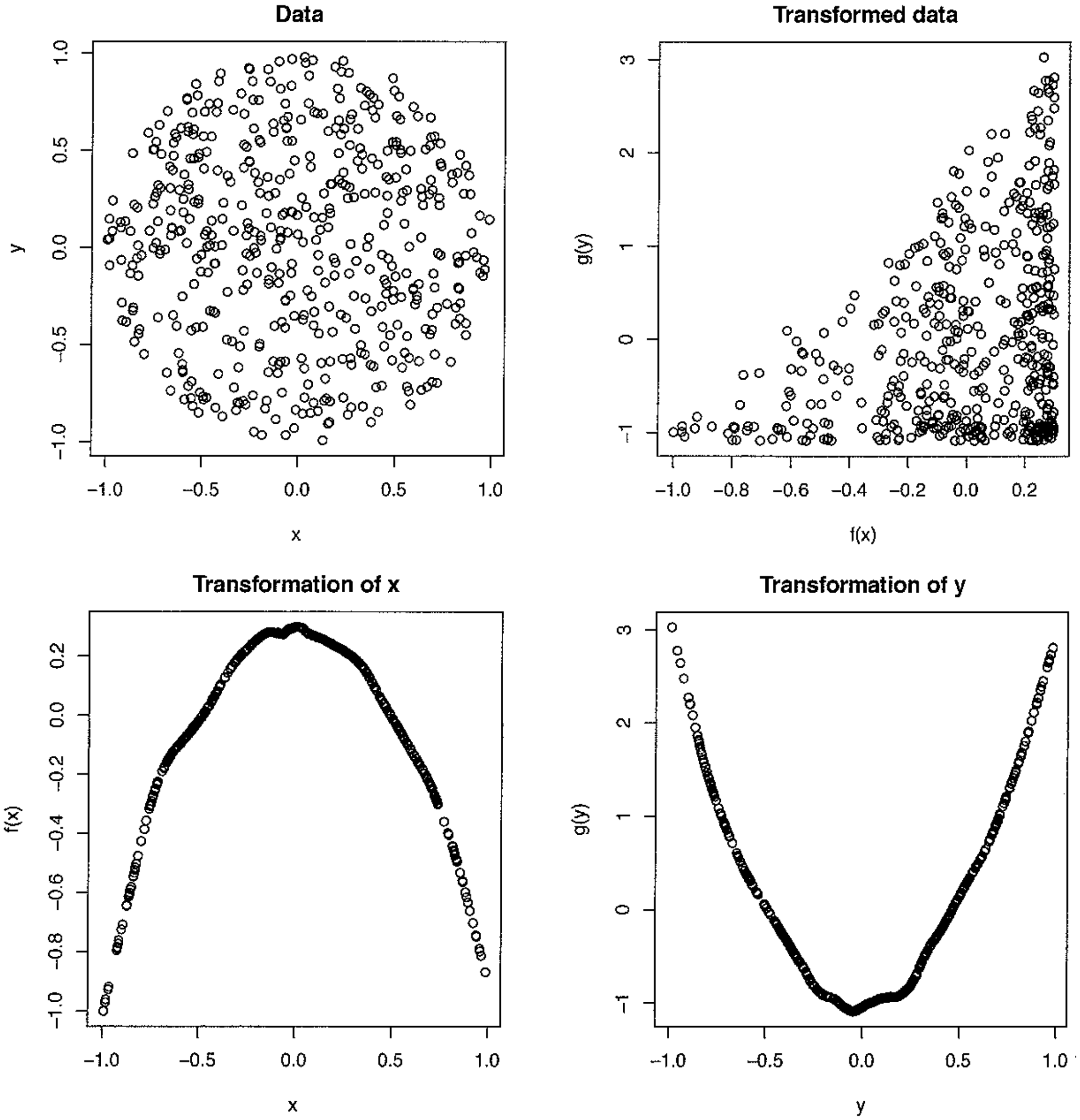


$f(x) = x^2$

$g(y) = -y^2$

$x^2-1 = f(x)-1$

Ball


Data

Transformed data

Transformation of x

Transformation of y

≃ 0.3

Problems with sample maximal correlation and ACE

As mentioned before, a big problem with maximal correlation is that it is not well-defined in the sample

The sample ACE algorithm is a nice tool, but defining the sample maximal correlation in terms of its output is not ideal, because the answer depends on the choice of smoother. It's also hard to compute, compared to the usual correlation (Homework 3)

Aside from these problems, maximal correlation and the ACE algorithm also have some undesirable properties. E.g., there are cases in which $\mathrm{mCor}(X, Y) = 1$ for random variables X, Y that don't exhibit perfect determinism, and similary, cases in which the sample ACE algorithm returns ≈ 1 for vectors x, y whose components are not completely determined from one another (Homework 3)

Distance correlation

Distance correlation[1] is a recent notion of correlation that also characterizes independence completely. It is well-defined in both the population and in the sample, and is easy to compute

In the population: given random variables $X, Y \in \mathbb{R}$, let X', Y' and X'', Y'' be independent pairs of random variables taken from the same joint distribution as that of X, Y. The distance covariance of X, Y is then defined as the square root of

$$\begin{aligned}
\mathrm{dCov}^2(X,Y) &= \mathrm{E}[|X - X'||Y - Y'|] + \mathrm{E}[|X - X'|]\mathrm{E}[|Y - Y'|] \\
&\quad - \mathrm{E}[|X - X'||Y - Y''|] - \mathrm{E}[|X - X''||Y - Y'|] \\
&= \mathrm{E}[|X - X'||Y - Y'|] + \mathrm{E}[|X - X'|]\mathrm{E}[|Y - Y'|] \\
&\quad - 2\mathrm{E}[|X - X'||Y - Y''|] \qquad = 0 \\
&\qquad\qquad \Longleftrightarrow X, Y \text{ indep.}
\end{aligned}$$

[1] Szekely et al. (2007), "Measuring and Testing Dependence By Correlation of Distances"; Szekely et al. (2009) "Brownian Distance Covariance"

Now we can define a notion of distance variance and distance correlation in an analogous manner to their usual definitions. I.e., the distance variance of X is defined by $dCov^2(X,Y)$

$$\mathrm{dVar}^2(X) = \mathrm{dCov}^2(X,X)$$

and the distance correlation of X, Y is defined by

$$\mathrm{dCor}^2(X,Y) = \frac{\mathrm{dCov}^2(X,Y)}{\sqrt{\mathrm{dVar}^2(X)}\sqrt{\mathrm{dVar}^2(Y)}}$$

Properties:

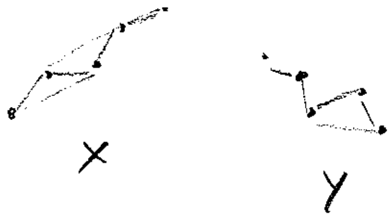
- $\mathrm{dCor}(aX+b,Y) = \mathrm{dCor}(X,Y)$ for any $a,b \in \mathbb{R}, a \neq 0$
- $0 \leq \mathrm{dCor}(X,Y) \leq 1$
- $\mathrm{dCor}(X,Y) = 0$ if and only if X, Y are independent

(There are many more)

Distance correlation in the sample

Now in the sample: given vectors $x, y \in \mathbb{R}^n$, we first define the distance matrices $A, B \in \mathbb{R}^{n \times n}$ by

$$A_{ij} = |x_i - x_j| \quad \text{and} \quad B_{ij} = |y_i - y_j|$$



for all $i, j = 1, \ldots n$. Then we double center A, B to get $\tilde{A}, \tilde{B}$, respectively, i.e., we center both the rows and columns of each matrix. Recall that, letting $M = \mathbb{1}\mathbb{1}^T/n$, this is

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

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

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

Now we follow the same steps as in the population, first defining the distance variance or x by

$$\mathrm{dvar}^2(x) = \mathrm{dcov}^2(x, x)$$

and then defining the distance correlation of x, y by

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

Properties:

- $\mathrm{dcor}(ax + b, y) = \mathrm{dcor}(x, y)$ for any $a, b \in \mathbb{R}, a \neq 0$
- $0 \leq \mathrm{dcor}(x, y) \leq 1$
- $\mathrm{dcor}(x, y) = 1$ if and only if $y = ax + b$ for some $a, b \in \mathbb{R}$, $a \neq 0$

(There are many more)

Why this sample definition?

What's the connection between this sample definition and the population definition?

Let A, B contain the pairwise absolute differences of x, y, and let $\tilde{A}, \tilde{B}$ be their double centered versions. It turns out (Homework 3, bonus):

$$\begin{aligned}\mathrm{dcov}^2(x,y) &= \frac{1}{n^2}\sum_{i,j=1}^n \tilde{A}_{ij}\tilde{B}_{ij} \\ &= \frac{1}{n^2}\sum_{i,j=1}^n A_{ij}B_{ij} - \frac{1}{n}\sum_{j=1}^n A_{\cdot j}B_{\cdot j} - \frac{1}{n}\sum_{i=1}^n A_{i\cdot}B_{i\cdot} + A_{\cdot\cdot}B_{\cdot\cdot}\end{aligned}$$

(where $\cdot$ means sum over that component). Now compare the population version:

$$\begin{aligned}\mathrm{dCov}^2(X,Y) = \mathrm{E}[|X-X'||Y-Y'|] + \mathrm{E}[|X-X'|]\mathrm{E}[|Y-Y'|] \\ - \mathrm{E}[|X-X'||Y-Y''|] - \mathrm{E}[|X-X''||Y-Y'|]\end{aligned}$$

Do these terms match (which ones)?

Distance correlation and characteristic functions

There is an alternate definition in terms of characteristic functions (this was actually the original motivation)

In the population: given random variables $X, Y \in \mathbb{R}$, let $h_X(t) = \mathrm{E}[\exp(itX)], h_Y(t) = \mathrm{E}[\exp(itY)]$ denote their characteristic functions, and let $h_{X,Y}(s,t) = \mathrm{E}[\exp(i(sX + tY))]$ denote their joint characteristic function. Then

$h_{X,Y} = h_X \cdot h_Y$

$\iff X, Y$ indep.

$$\mathrm{dCov}(X,Y) = \|h_{X,Y} - h_X h_Y\|$$

measuring diff b/w marginal & joint char. funs.

where the above $\|\cdot\|$ is a specially defined norm on functions (actually a double integral)

$\| f_1 - f_2 \| \overset{def}{=} \int \left(f_1(x) - f_2(x) \right) dx$

From this, one can show that $\mathrm{dCov}(X,Y) = \|h_{X,Y} - h_X h_Y\| = 0$ if and only if $h_{X,Y}(s,t) = h_X(s) h_Y(t)$ for all $s, t \in \mathbb{R}$, which is true if and only if X, Y are independent

In the sample: given vectors $x, y \in \mathbb{R}^n$, define the marginal and joint empirical characteristic functions,

$$h_x(t) = \frac{1}{n}\sum_{k=1}^{n} \exp(itx_k) \quad \text{and} \quad h_y(t) = \frac{1}{n}\sum_{k=1}^{n} \exp(ity_k)$$

$$h_{x,y}(s,t) = \frac{1}{n}\sum_{k=1}^{n} \exp(i(sx_k + ty_k))$$

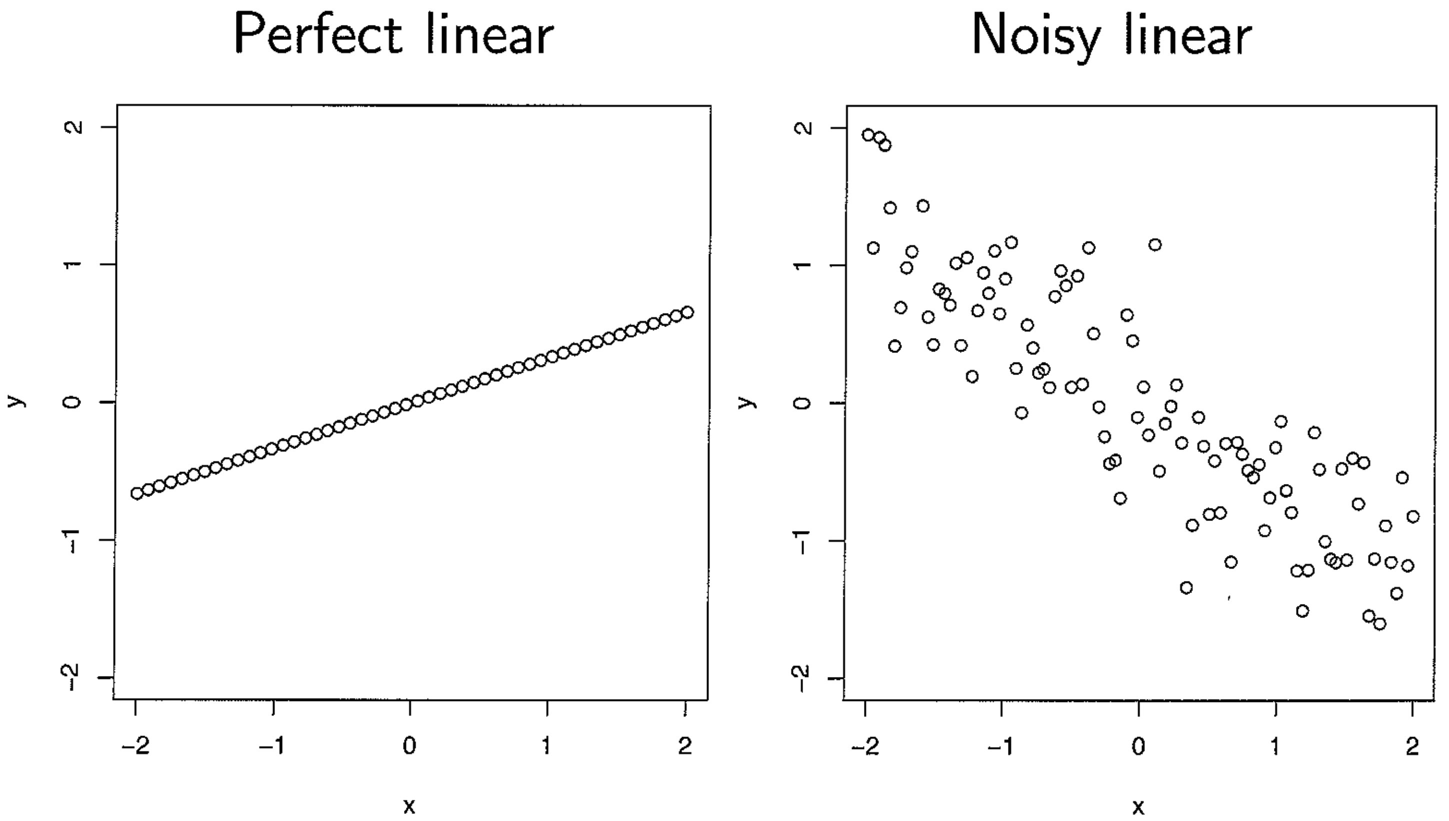
Then it turns out that

$$\mathrm{dcov}(x,y) = \|h_{x,y} - h_x h_y\|, \quad = \frac{1}{n^2}\sum \tilde{A}_{ij}\tilde{B}_{ij}$$

where $\|\cdot\|$ is the same special norm on functions as in the previous slide

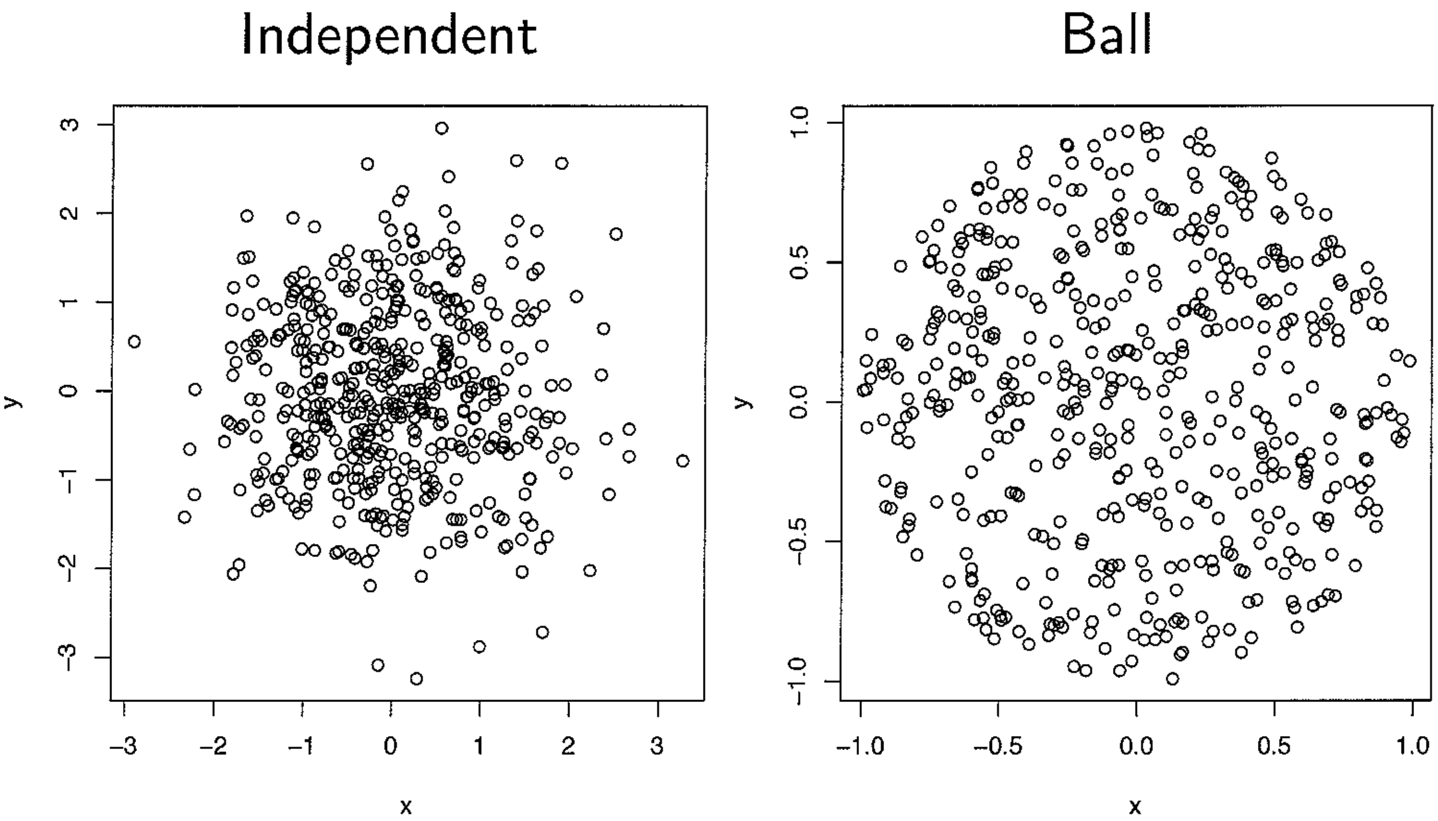
(It's a lot easier to compute dcor using the first definition!)

Example: distance correlation


Perfect linear

Noisy linear

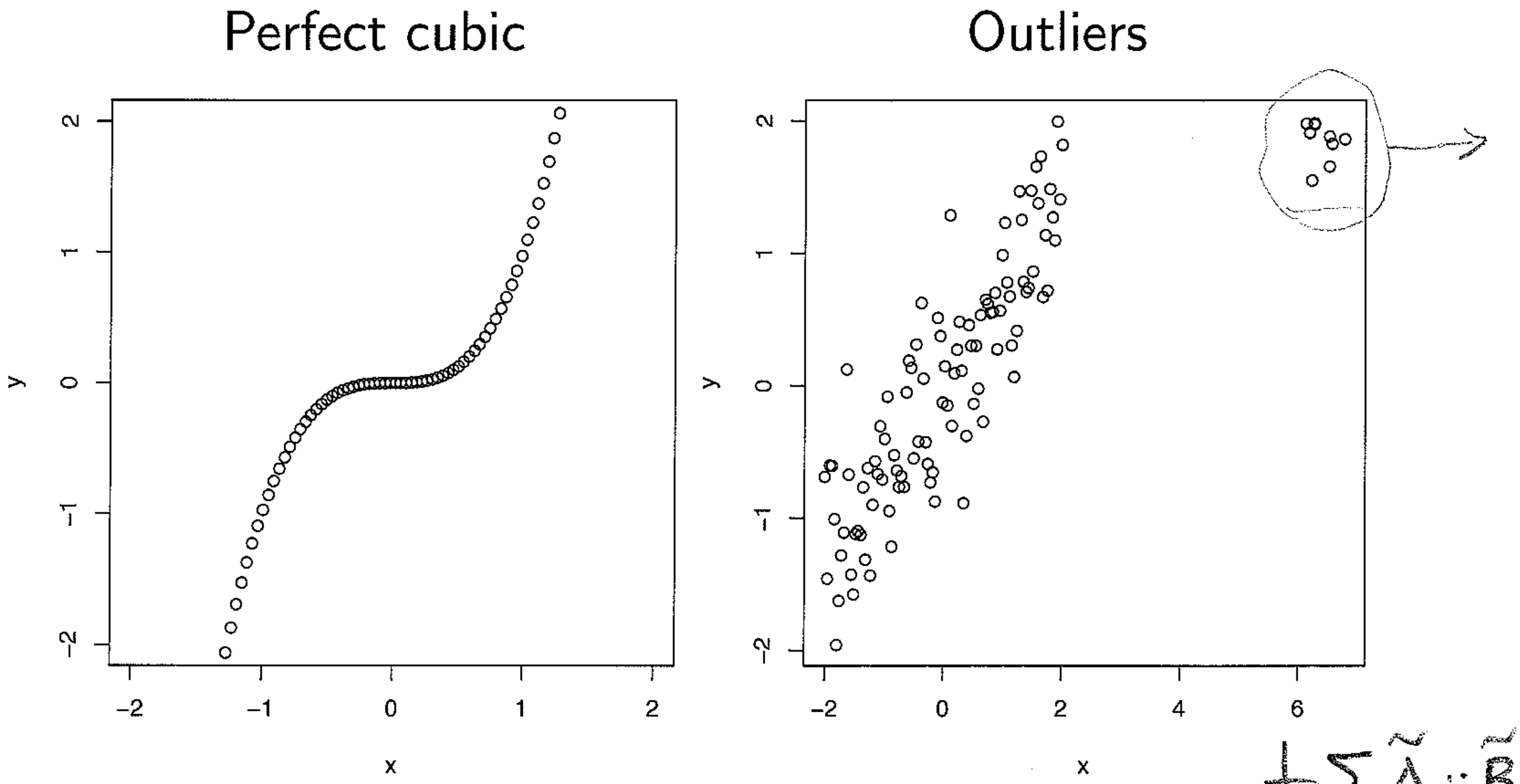
dcor =	1.000	0.867
mcor =	1.000	0.896
rcor =	1.000	-0.872
cor =	1.000	-0.866



Independent

Ball

dcor =	0.078	0.099
mcor =	0.124	0.316
rcor =	-0.021	-0.033
cor =	-0.023	-0.029

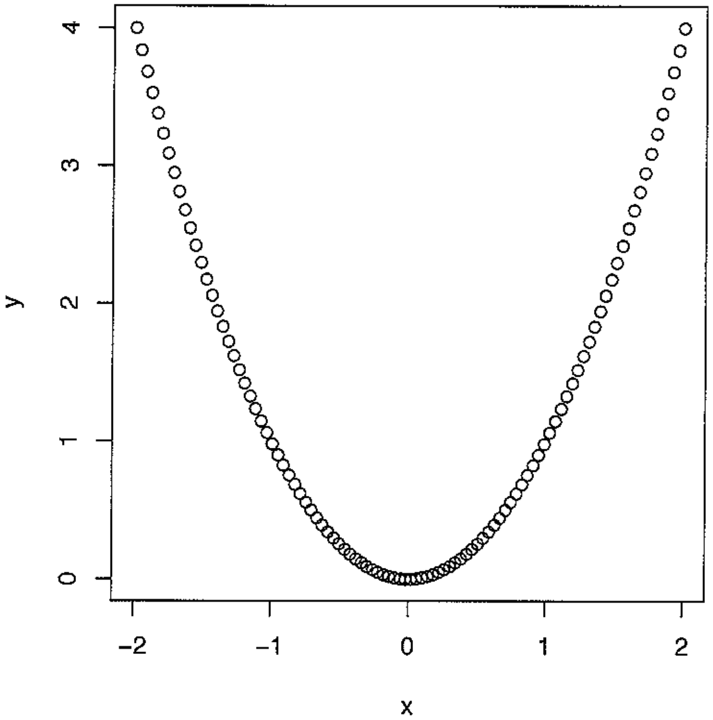


$\frac{1}{n^2}\sum \tilde{A}_{ij}\tilde{B}_{ij}$

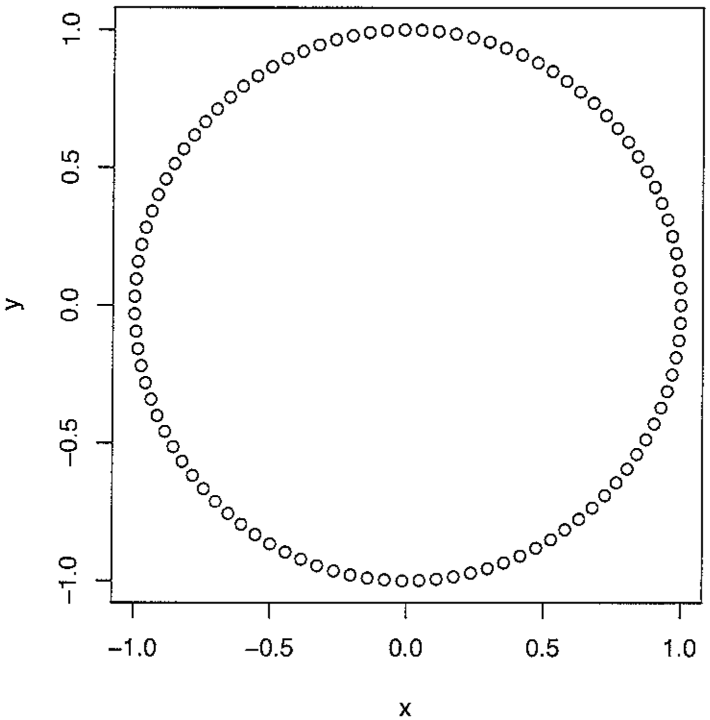
pairwise dist.
between x & y

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

Perfect quadratic



Perfect circle



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

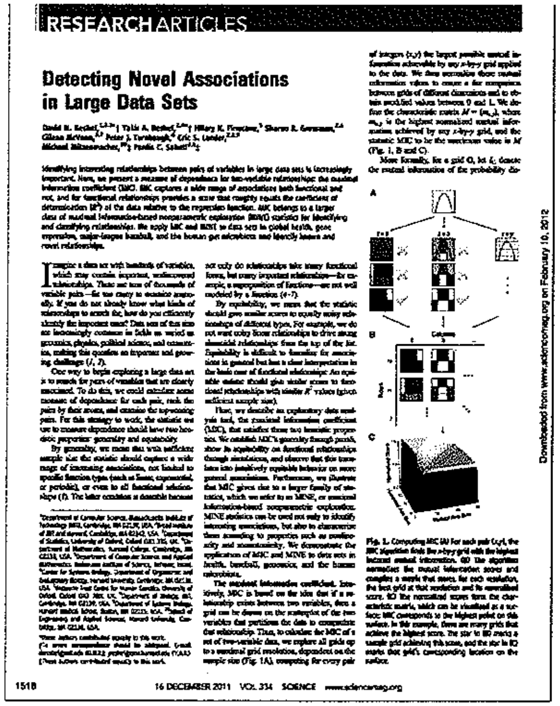
Distance correlation in R

The dcor function in the energy package can be used to compute distance correlation. E.g.,

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

Maximal information coefficient

Reshef et al. (2011), “Detecting Novel Associations in Large Data Sets”



RESEARCH ARTICLES

Detecting Novel Associations in Large Data Sets

[illegible]

1518 16 DECEMBER 2011 VOL 334 SCIENCE www.sciencemag.org

Downloaded from www.sciencemag.org on February 10, 2012

This has been called a “correlation for the 21st century”, and it has also been heavily criticized ... read it yourself and make your own judgements!

No free lunch

Generally speaking, we can't design a measure of correlation that performs well in every situation

This is because methods that are designed to detect very broad notions of associations will tend to give rise to false positives

Recap: measures of correlation (continued)

In this lecture we saw examples of the alternating conditional expectations (ACE) algorithm in practice. The answers depend on the choice of smoother

In the population, distance covariance is defined in terms of independent copies of the pair of random variables. From this we can also define distance variance and distance correlation. Importantly, the latter is zero if an only if the random variables are independent

In the sample, distance covariance is defined by forming two absolute distance matrices, double centering them, and summing the elementwise products. Distance variance and distance correlation follow as before

Both population and sample versions have alternative definitions in terms of characteristic functions

Next time: midterm 1

Good luck!