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

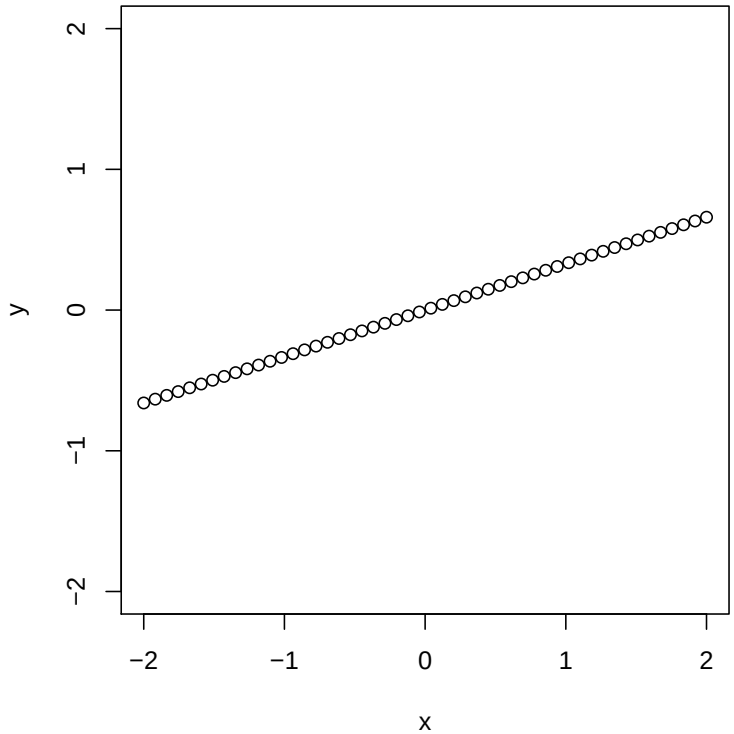
The sample version of the ACE algorithm requires us to pick a smoother $\mathcal{S}$ to approximate the conditional expectations (e.g., kernel regression or local linear regression). Algorithm:

- 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 $\mathrm{mcor}(x, y) = f_k(x)^T g_k(y)$

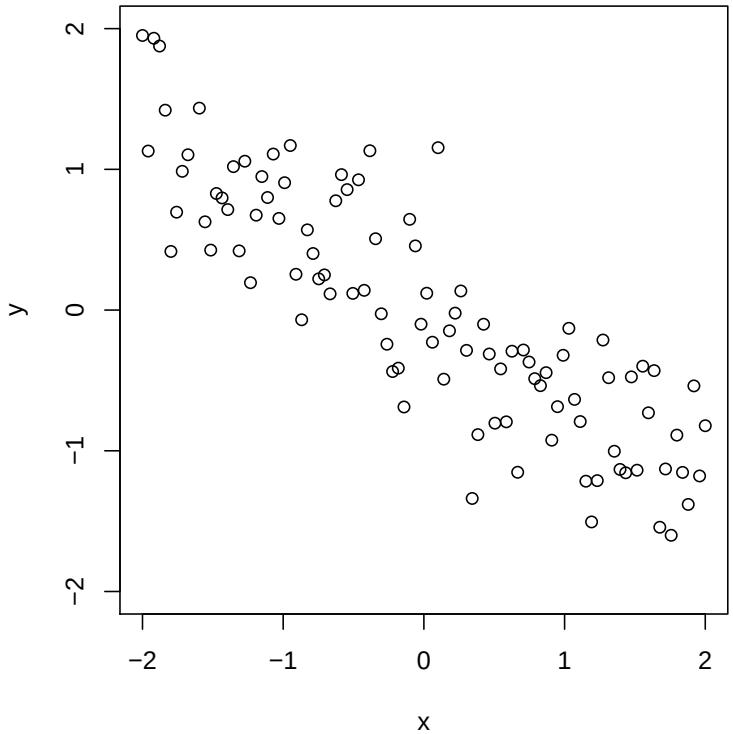
This not always guaranteed to converge, and the answer depends on the smoother. But most of the time it works well in practice

Example: maximal correlation

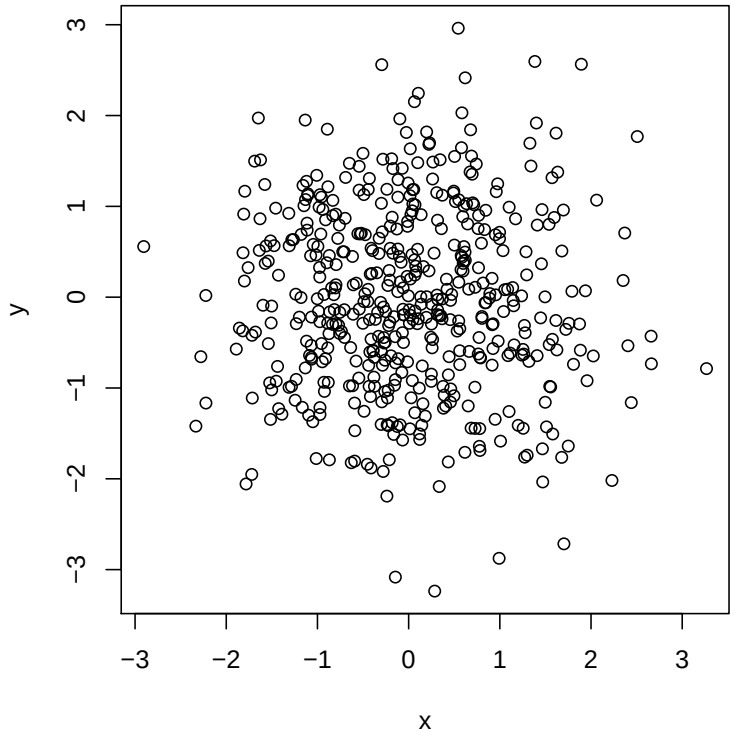
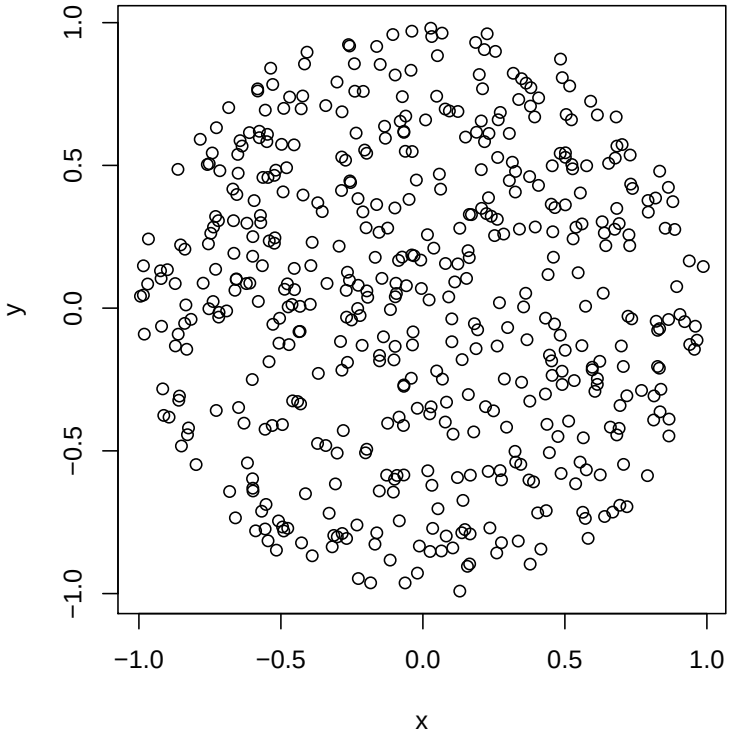
Perfect linear



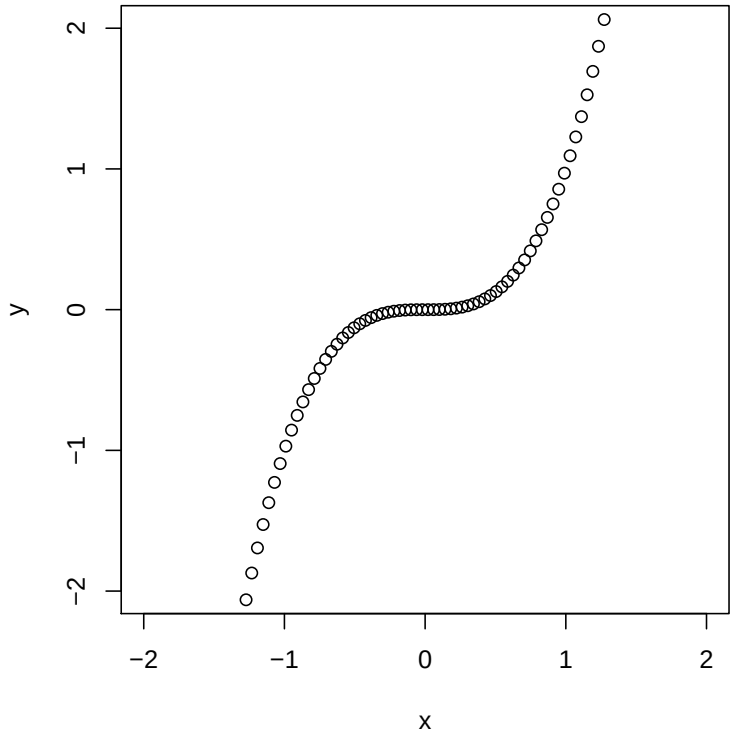
Noisy linear



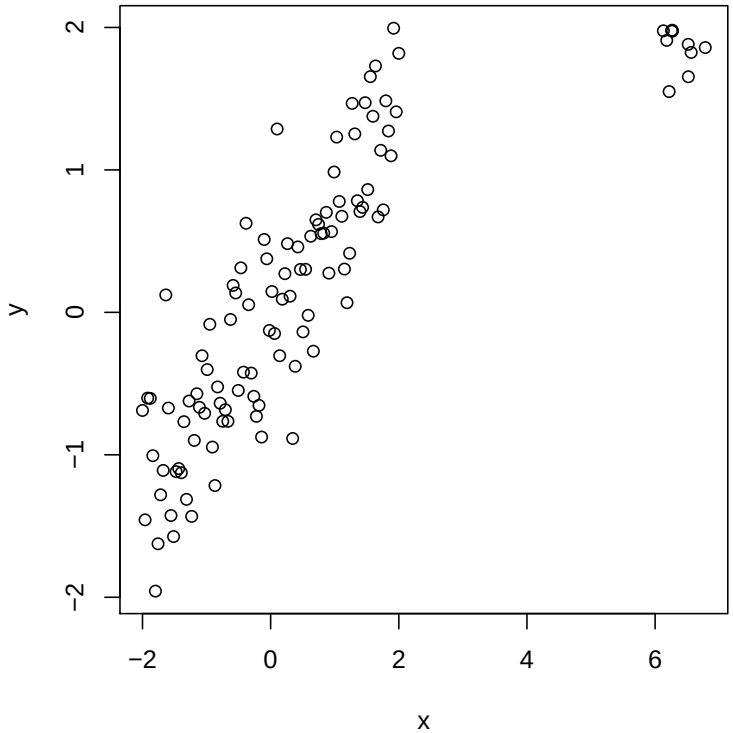
mcor =	1.000	0.896
rcor =	1.000	-0.872
cor =	1.000	-0.866

	Independent	Ball
		
$\mathrm{mcor} =$	0.124	0.316
$\mathrm{rcor} =$	-0.021	-0.033
$\mathrm{cor} =$	-0.023	-0.029

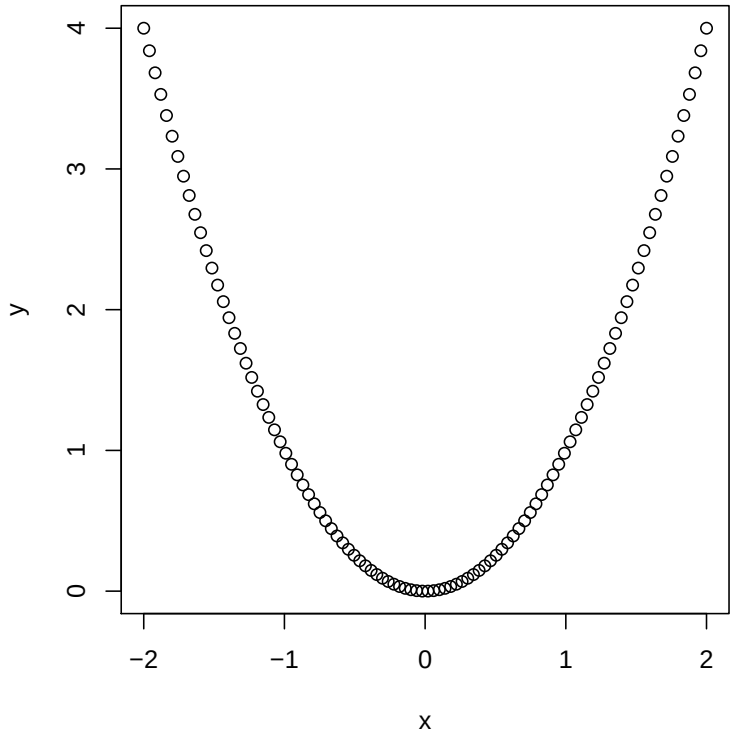
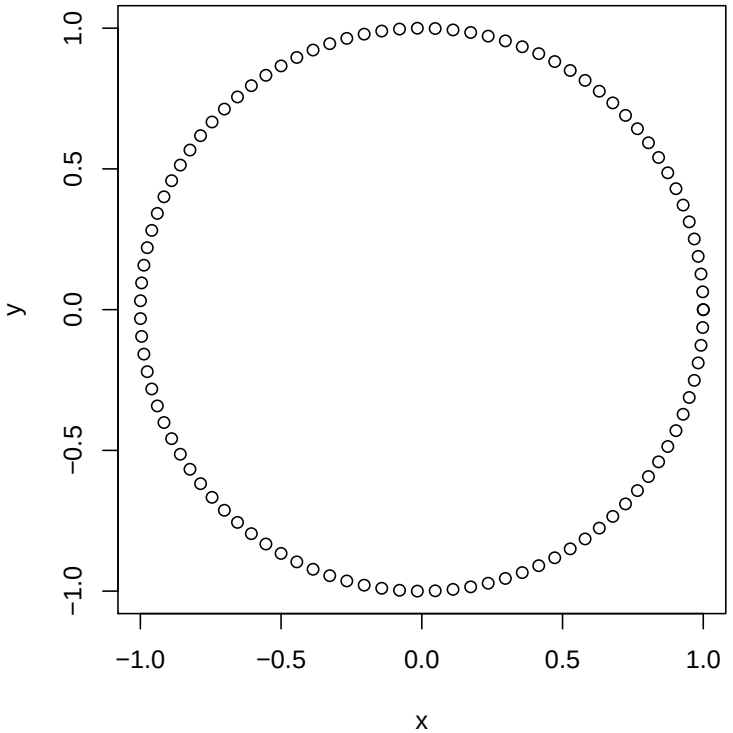
Perfect cubic



Outliers

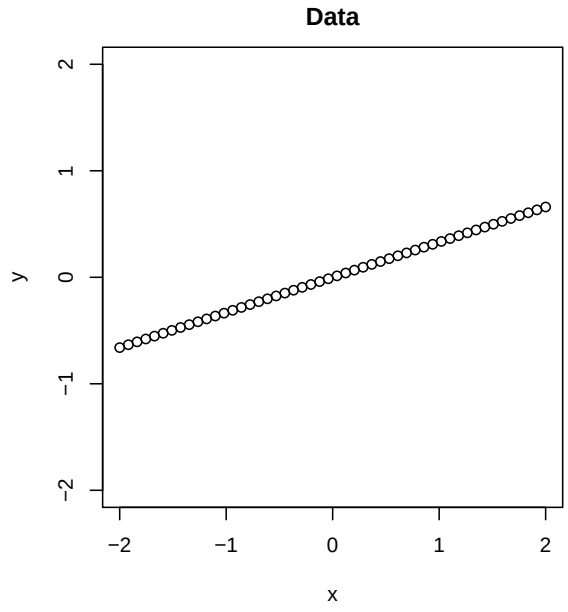


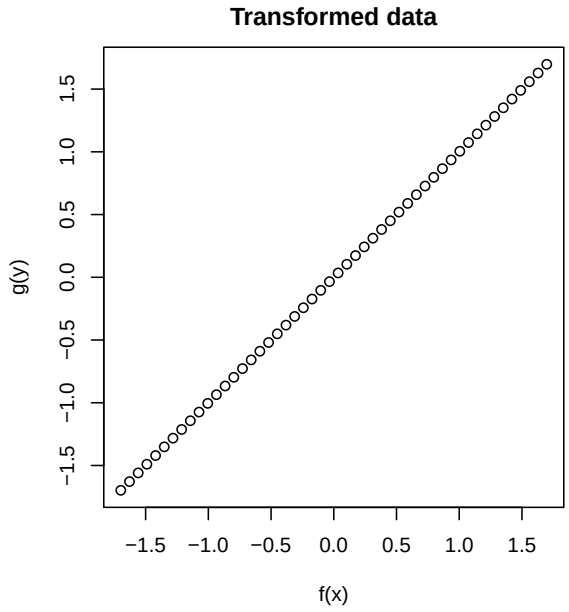
$\mathrm{mcor} =$	1.000	0.913
$\mathrm{rcor} =$	1.000	0.905
$\mathrm{cor} =$	0.920	0.834

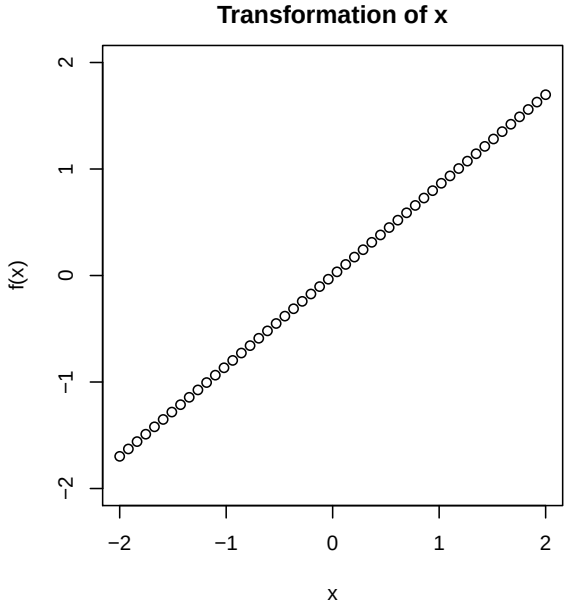
	Perfect quadratic	Perfect circle
		
$\mathrm{mcor} =$	1.000	1.000
$\mathrm{rcor} =$	0.013	-0.001
$\mathrm{cor} =$	0.000	0.000

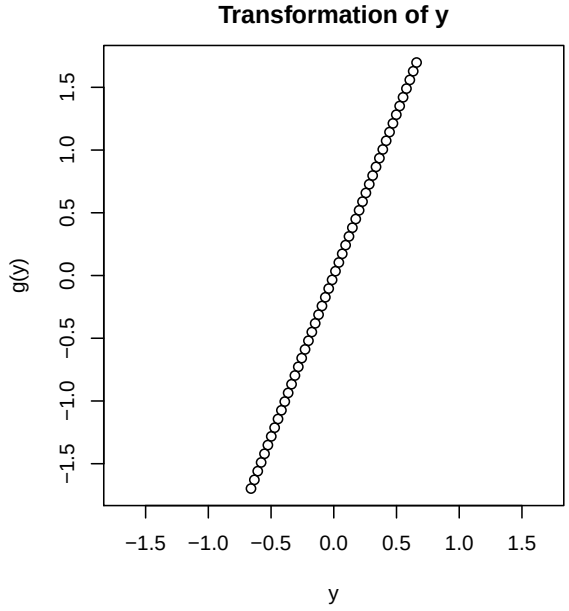
Example: ACE algorithm

Perfect linear

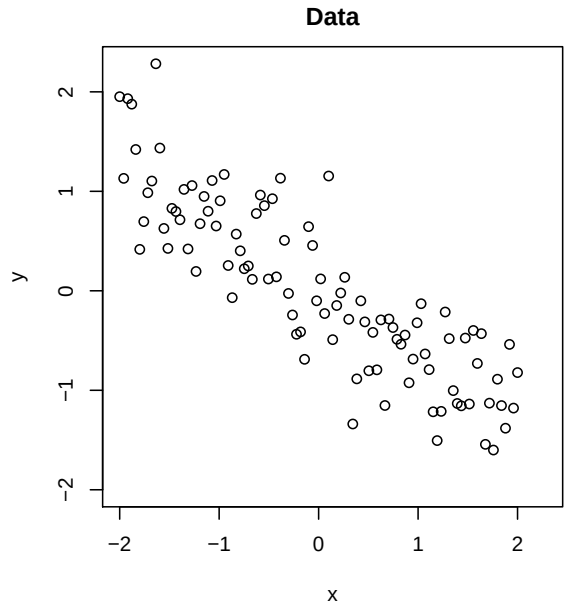


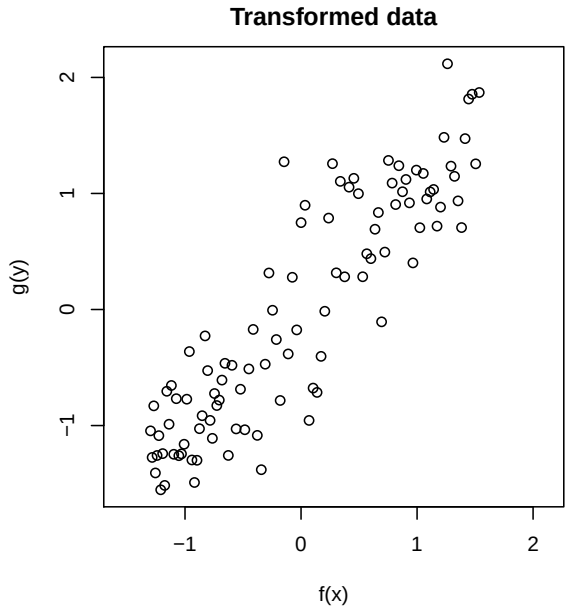


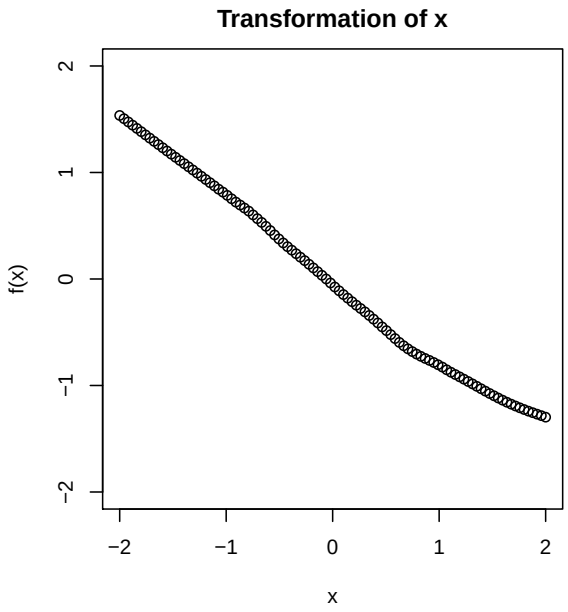


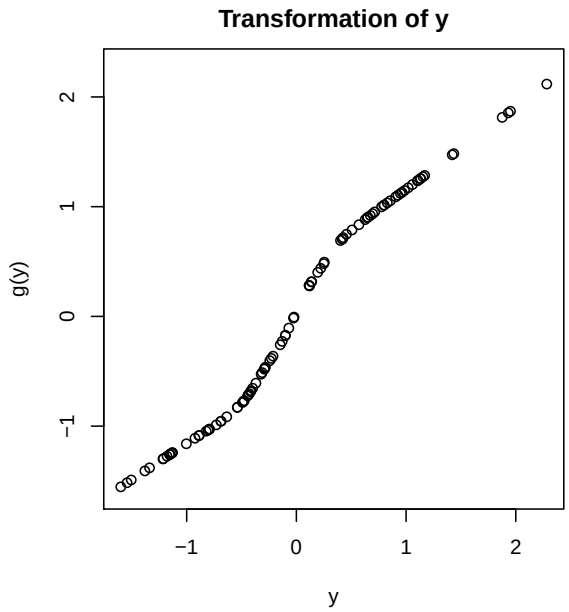


Noisy linear

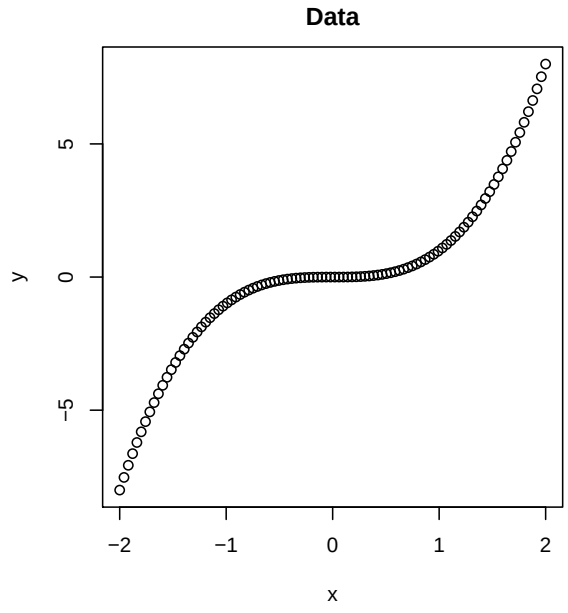


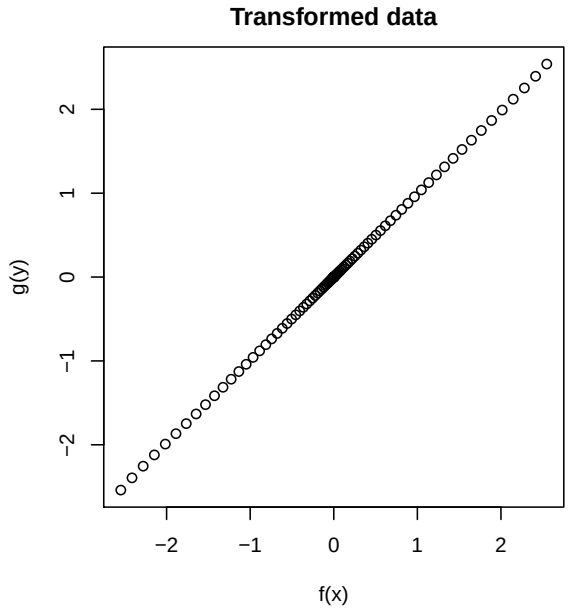


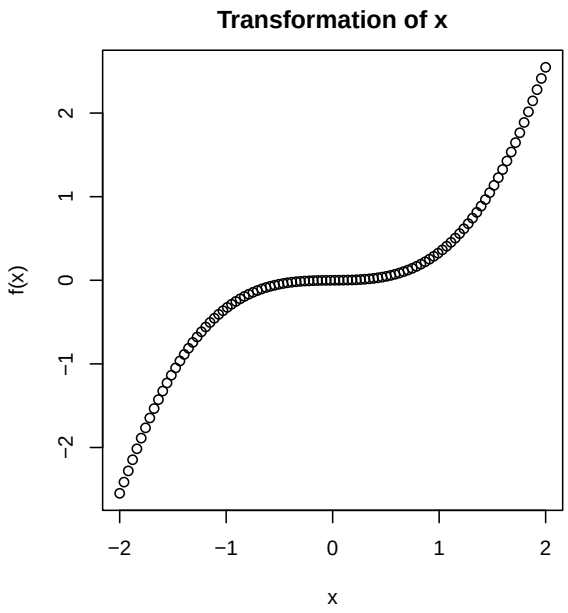


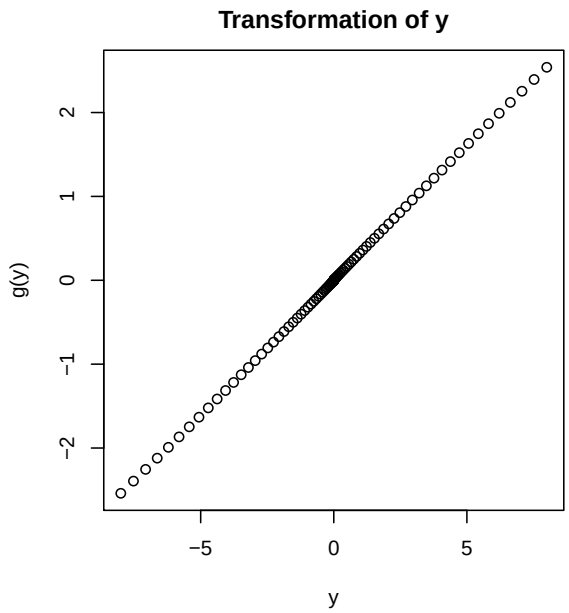


Perfect cubic

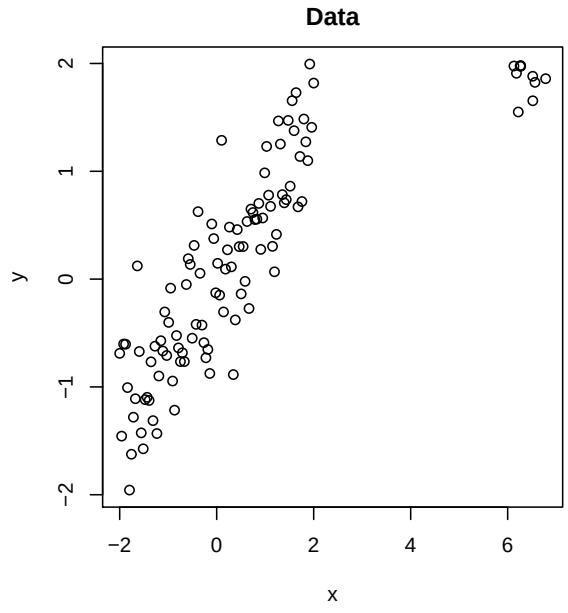


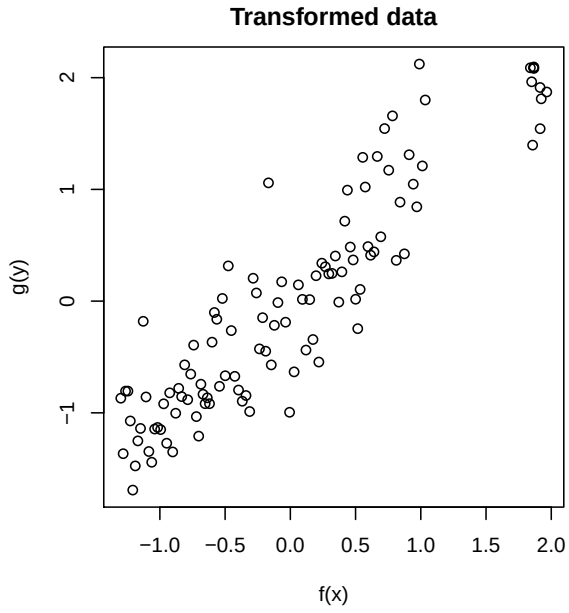


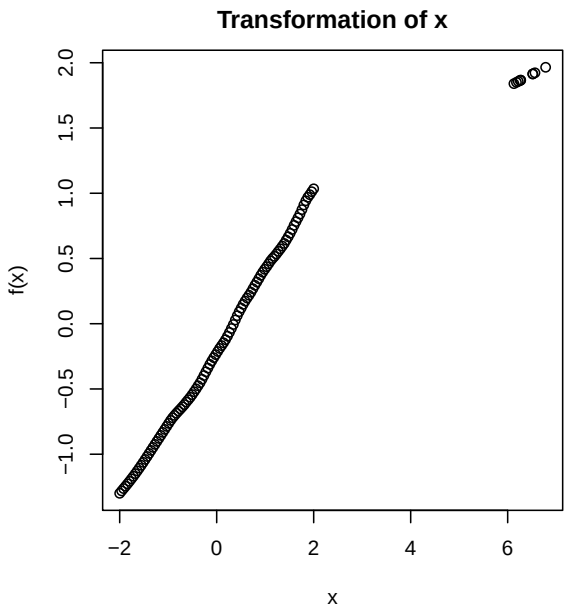


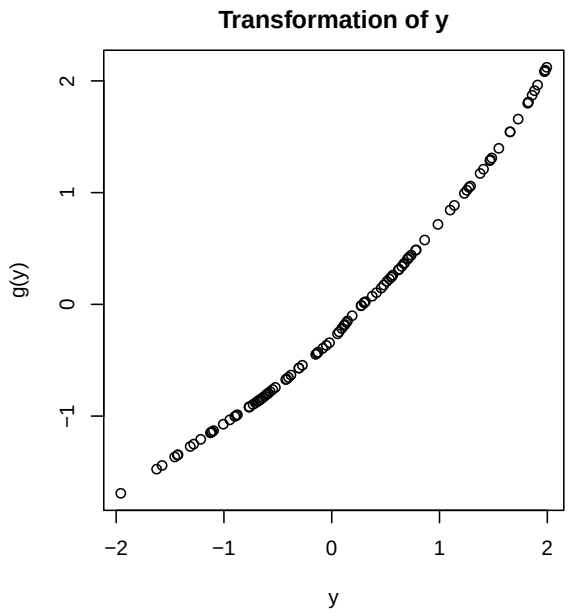


Outliers

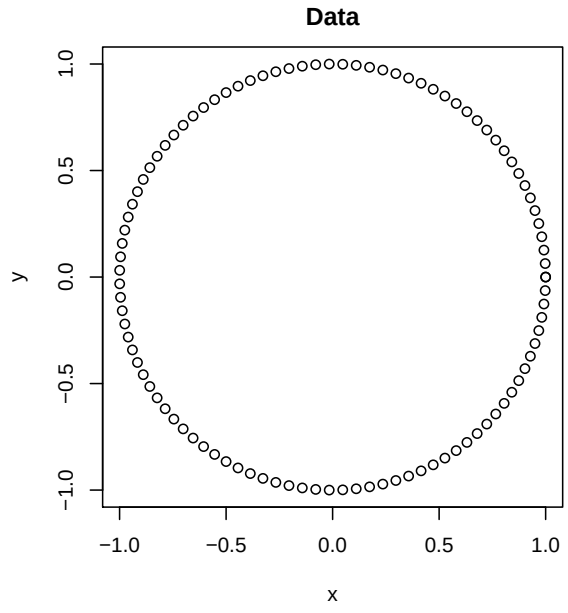

Data
y
x
−2
−1
0
1
2
−2
0
2
4
6

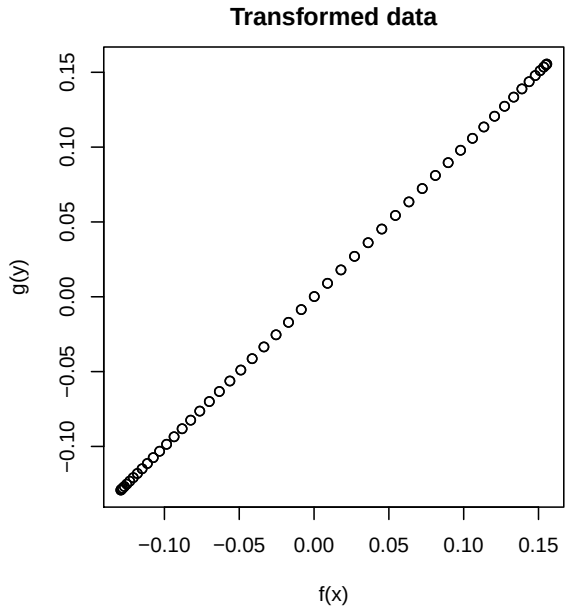

Transformed data
g(y)
f(x)
−1
0
1
2
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0

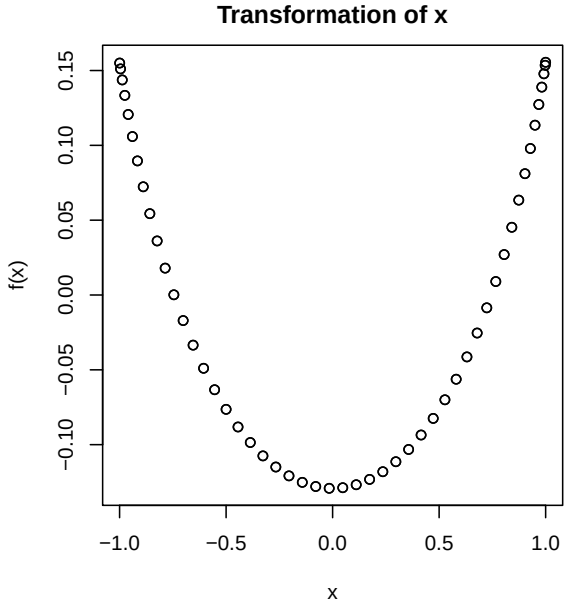

Transformation of x
f(x)
x
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
−2
0
2
4
6

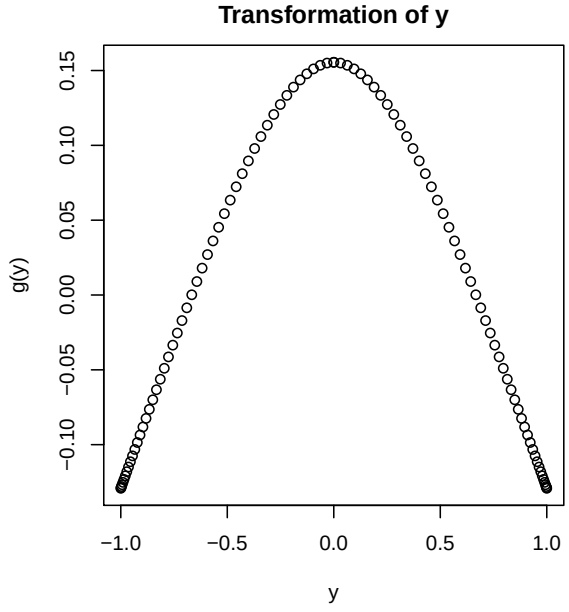

Transformation of y
g(y)
y
−1
0
1
2
−2
−1
0
1
2

Perfect circle

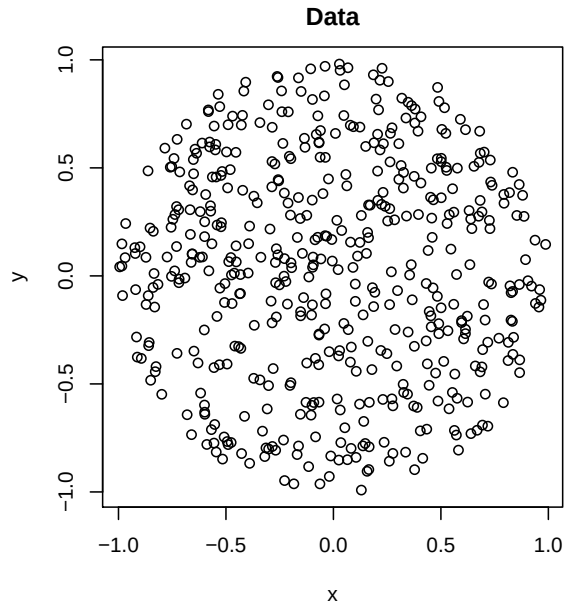


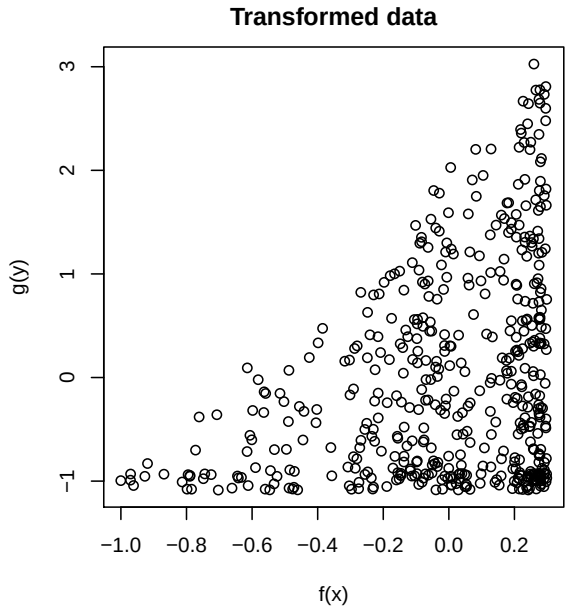


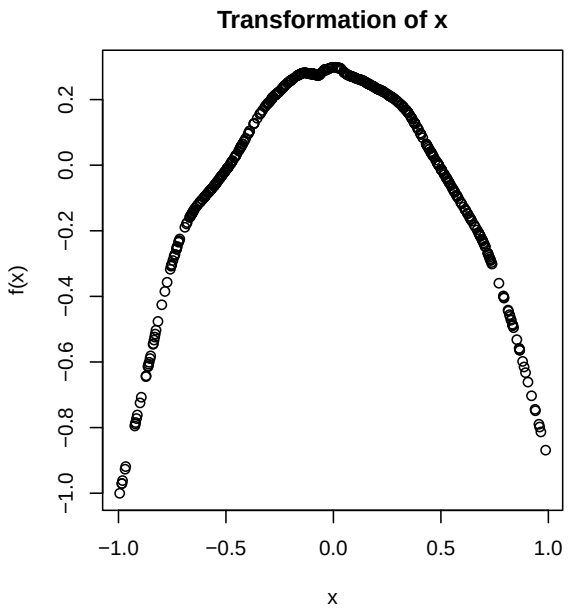


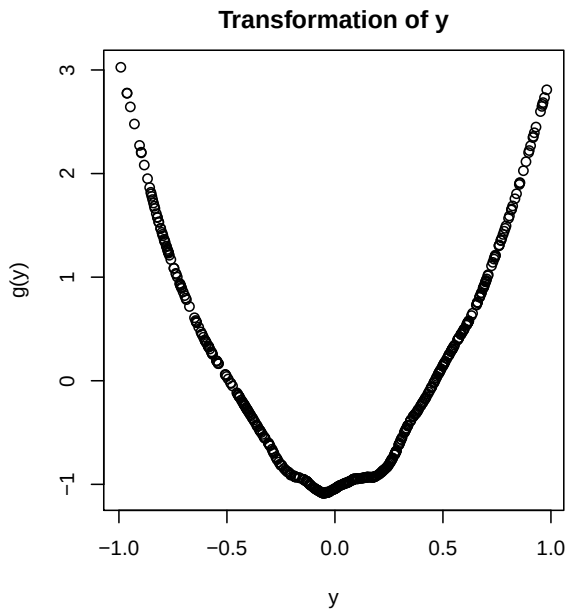


Ball


Data
y
x
1.0
0.5
0.0
−0.5
−1.0
−1.0
−0.5
0.0
0.5
1.0


Transformed data
g(y)
f(x)
3
2
1
0
−1
−1.0
−0.8
−0.6
−0.4
−0.2
0.0
0.2


Transformation of x
f(x)
x
0.2
0.0
−0.2
−0.4
−0.6
−0.8
−1.0
−1.0
−0.5
0.0
0.5
1.0


Transformation of y
g(y)
y
3
2
1
0
−1
−1.0
−0.5
0.0
0.5
1.0

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

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

[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

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

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

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

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

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

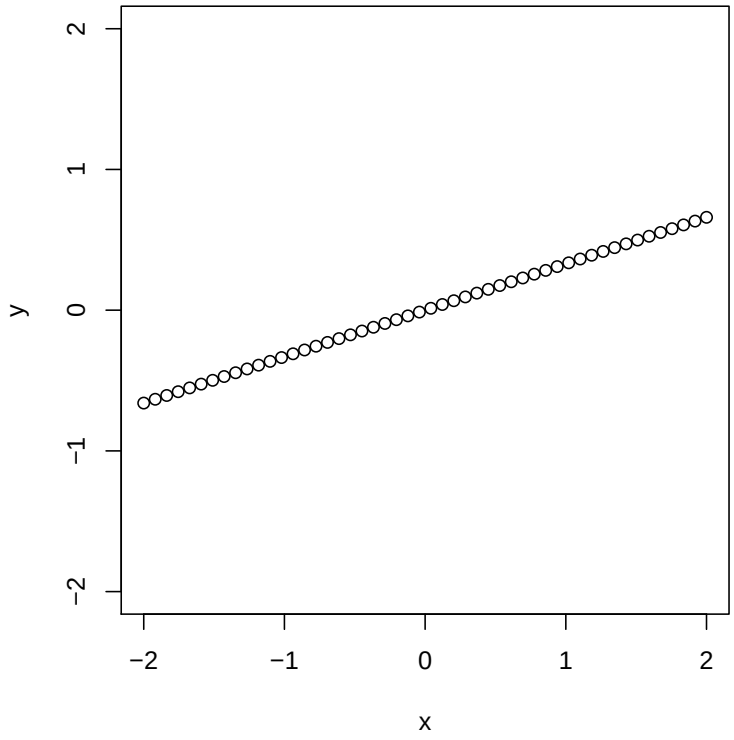
$$\mathrm{dcov}(x,y) = \|h_{x,y} - h_x h_y\|,$$

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

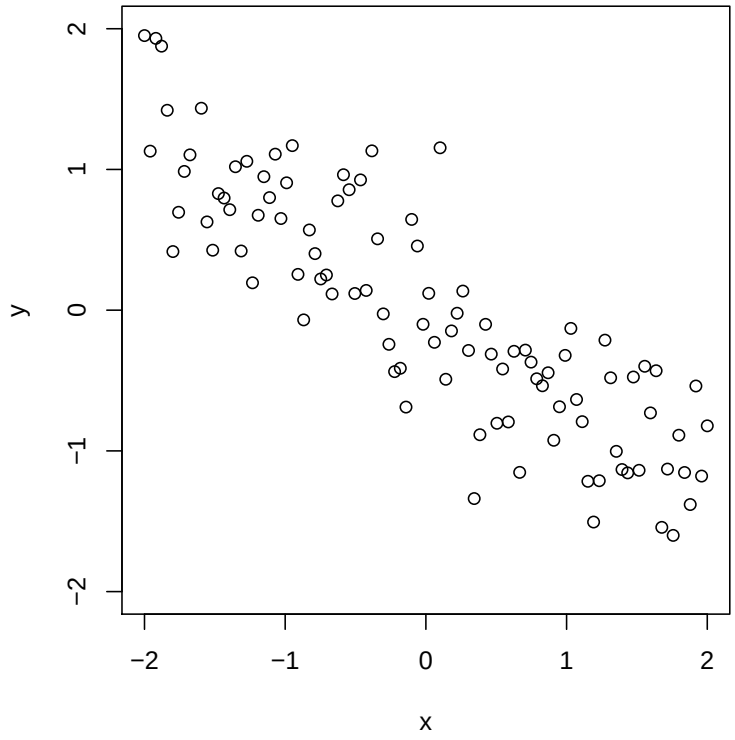
(It's a lot easier to compute dcor using the first definition!)

Example: distance correlation

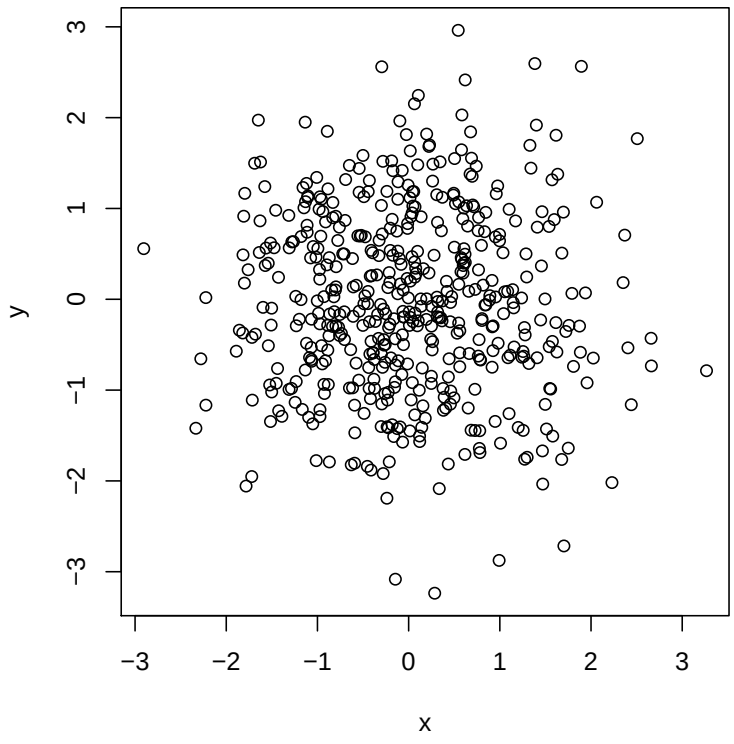
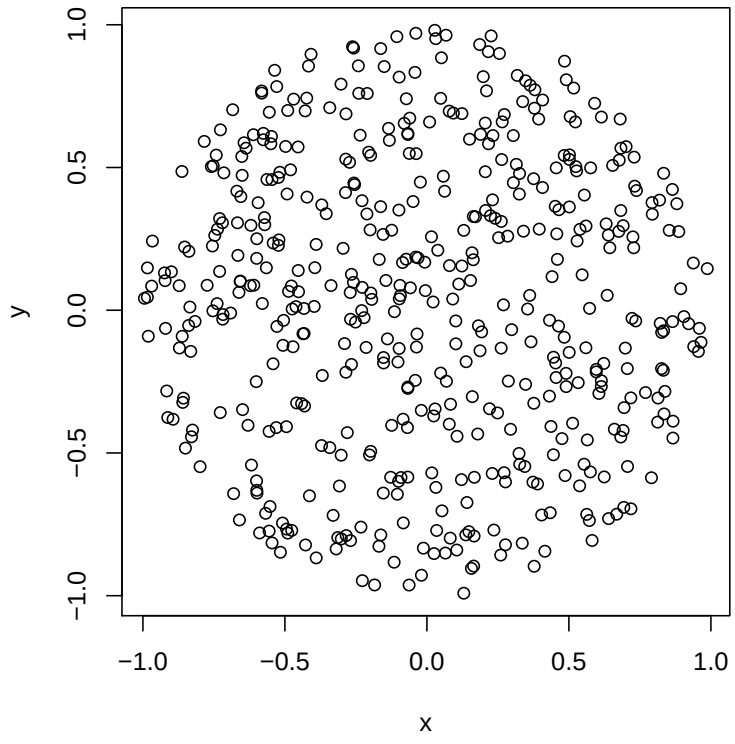
Perfect linear



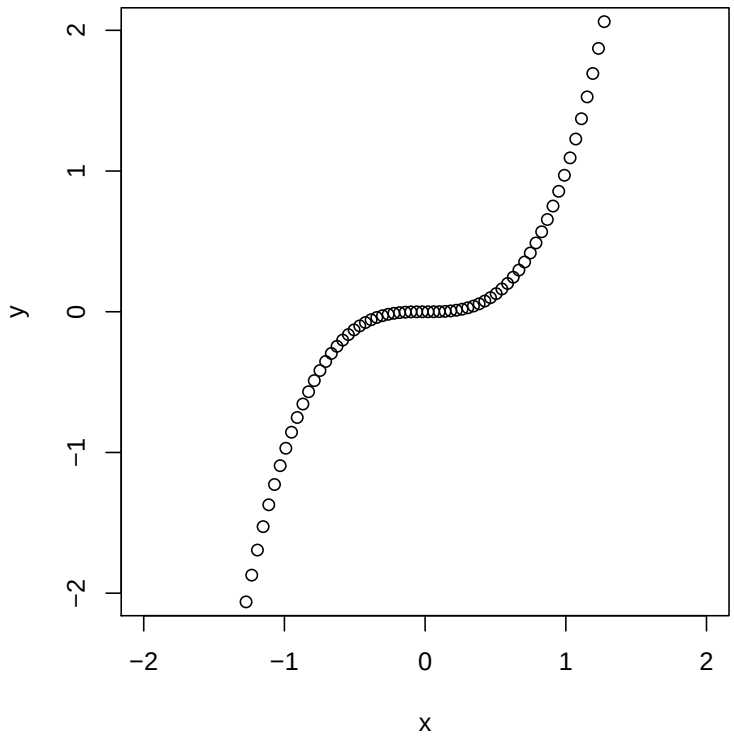
Noisy linear



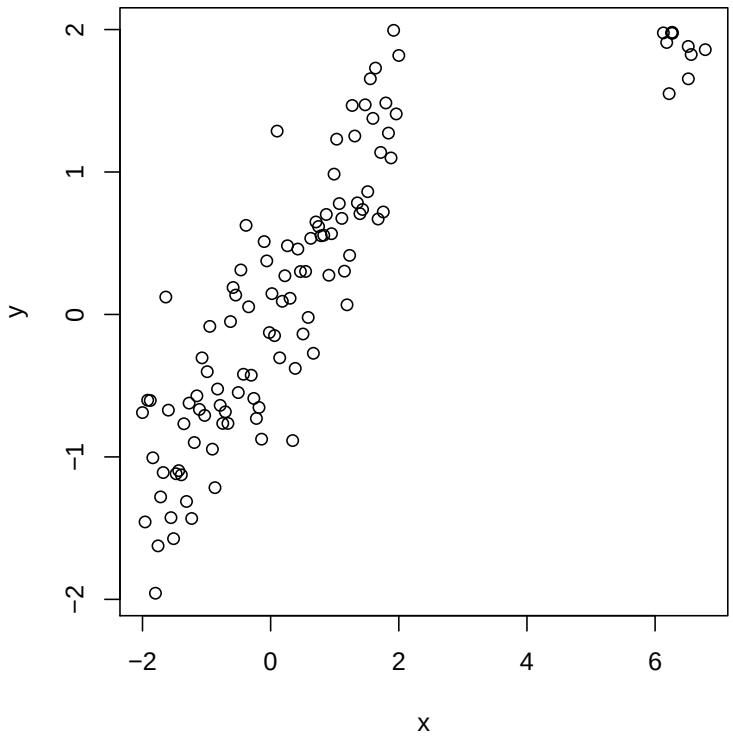
$\mathrm{dcor} =$	1.000	0.867
$\mathrm{mcor} =$	1.000	0.896
$\mathrm{rcor} =$	1.000	-0.872
$\mathrm{cor} =$	1.000	-0.866

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

Perfect cubic

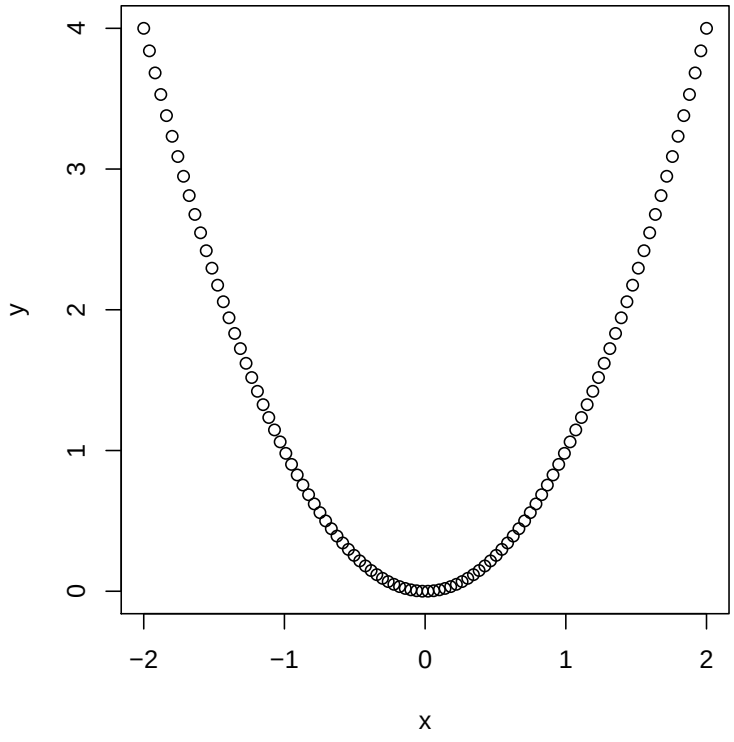


Outliers

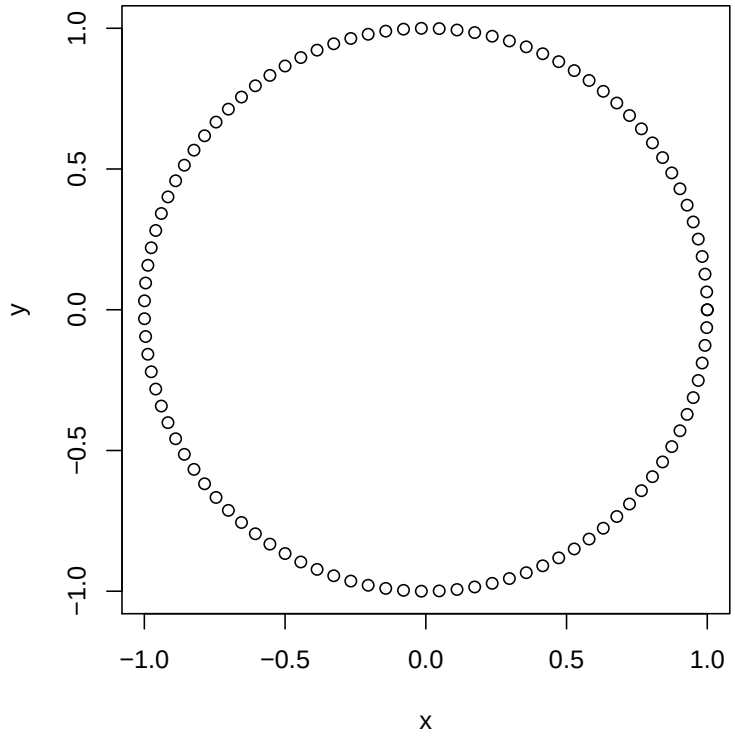


$\mathrm{dcor} =$	0.920	0.854
$\mathrm{mcor} =$	1.000	0.913
$\mathrm{rcor} =$	1.000	0.905
$\mathrm{cor} =$	0.920	0.834

Perfect quadratic



Perfect circle



$\mathrm{dcor} =$	0.492	0.200
$\mathrm{mcor} =$	1.000	1.000
$\mathrm{rcor} =$	0.013	-0.001
$\mathrm{cor} =$	0.000	0.000

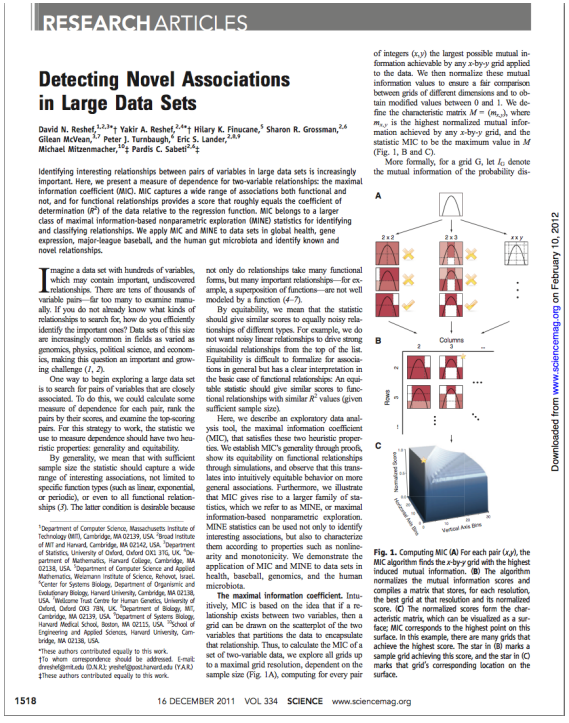
Distance correlation in R

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

David N. Reshef,[1,2,3*†] Yakir A. Reshef,[3,4*†] Hilary K. Finucane,[5] Sharon R. Grossman,[2,6] Gilean McVean,[3,7] Peter J. Turnbaugh,[8] Eric S. Lander,[2,9] Michael Mitzenmacher,[1‡] 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 identify the important ones? Data sets of this size are increasingly common in fields as varied as genomics, physics, political science, and economics, making this question an important and growing challenge (*1*, *2*).

One way to begin exploring a large data set is to search for pairs of variables that are closely associated. To do this, we could calculate some measure of dependence for each pair, rank the pairs by their scores, and examine the top-scoring pairs. For this strategy to work, the statistic we use to measure dependence should have two heuristic properties: generality and equitability.

By generality, we mean that with sufficient sample size the statistic should capture a wide range of interesting associations, not limited to specific function types (such as linear, exponential, or periodic), or even to all functional relationships (*3*). The latter condition is desirable because 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 relationships of different types. For example, we do not want noisy linear relationships to drive strong sinusoidal relationships from the top of the list. Equitability is difficult to formalize for associations in general but has a clear interpretation in 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).

Here, we describe an exploratory data analysis tool, the maximal information coefficient (MIC), that satisfies these two heuristic properties. We establish MIC's generality through proofs, show its equitability on functional relationships through simulations, and observe that this translates into intuitively equitable behavior on more general associations. Furthermore, we illustrate that MIC gives rise to a larger family of statistics, which we refer to as MINE, or maximal information-based nonparametric exploration. MINE statistics can be used not only to identify interesting associations, but also to characterize them according to properties such as nonlinearity and monotonicity. We demonstrate the application of MIC and MINE to data sets in health, baseball, genomics, and the human microbiota.

The maximal information coefficient. Intuitively, MIC is based on the idea that 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. Thus, to calculate the MIC of a set of two-variable data, we explore all grids up to a maximal grid resolution, dependent on the sample size (Fig. 1A), computing for every pair 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-

[1]Department of Computer Science, Massachusetts Institute of Technology (MIT), Cambridge, MA 02139, USA. [2]Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA. [3]Department of Statistics, University of Oxford, Oxford OX1 3TG, UK. [4]Department of Mathematics, Harvard College, Cambridge, MA 02138, USA. [5]Department of Computer Science and Applied Mathematics, Weizmann Institute of Science, Rehovot, Israel. [6]Center for Systems Biology, Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA 02138, USA. [7]Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford OX3 7BN, UK. [8]Department of Biology, MIT, Cambridge, MA 02139, USA. [9]Department of Systems Biology, Harvard Medical School, Boston, MA 02115, USA. [10]School of Engineering and Applied Sciences, Harvard University, Cambridge, MA 02138, USA.

*These authors contributed equally to this work.
†To whom correspondence should be addressed. E-mail: dnreshef@mit.edu (D.N.R.); yreshef@post.harvard.edu (Y.A.R.)
‡These authors contributed equally to this work.

Fig. 1. Computing MIC (**A**) For each pair (x,y), the MIC algorithm finds the x-by-y grid with the highest induced mutual information. (**B**) 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. (**C**) The normalized scores form the characteristic matrix, which can be visualized as a surface; MIC corresponds to the highest point on this surface. In this example, there are many grids that achieve the highest score. The star in (**B**) marks a sample grid achieving this score, and the star in (**C**) marks that grid's corresponding location on the surface.

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!