TESTING DEPENDENCE BY CORRELATION OF DISTANCES

Post-Masters Placement Summer Project 2016

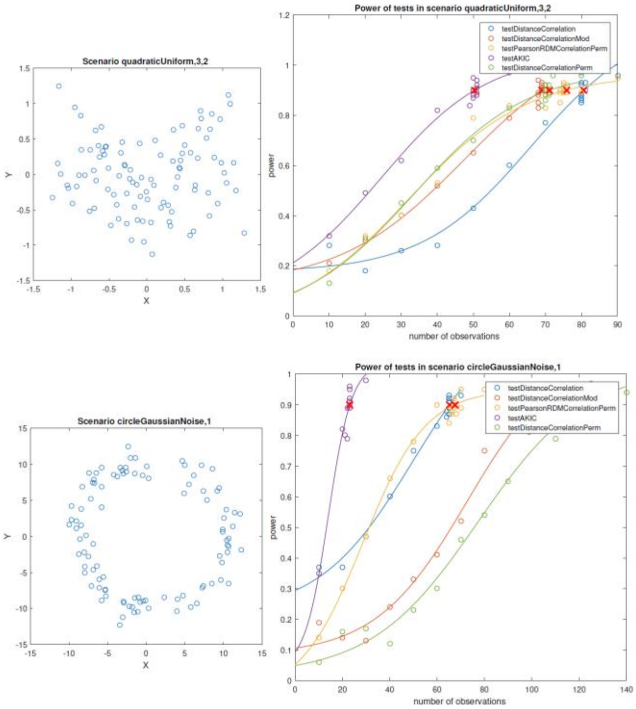
PROJECT MOTIVATION

Given samples from two random variables X and Y, to test whether they are independent or not, we usually compute their Pearson correlation (or a related test statistic). However, the Pearson correlation coefficient is a measure of linear dependence, and if for example $Y = X^2$, then the correlation of X and Y is 0.

In an article[1] from 2007, Szekely, Rizzo and Bakirov introduce the new concept of *distance correlation*, which is equal to 0 if and only if X and Y are independent.

Neuroscientists from the Visual Objects Lab at CBU, motivated by the properties of a neural network, adopted a practice which was strikingly similar to calculating (an estimator of) the distance correlation of two random variables.

The aim of this project is to investigate and explain this similarity, provide neuroscientists in the partnering research group with some insight into the mathematics behind different dependence measures, and build a toolbox for comparing the power of different dependence measures against a variety of data sets.



Figures: Data on the LHS to be tested for dependence, and graph of test power as a function of sample size for various dependence tests.

MATHEMATICS OF DISTANCE CORRELATION

Let X and Y be two random variables taking values in $\mathbb{R}^p$ and $\mathbb{R}^q$ respectively, and let $f_X(s) = \mathbb{E}\left(e^{i\langle s,X\rangle}\right)$ and $f_Y(t) = \mathbb{E}\left(e^{i\langle t,Y\rangle}\right)$ be their characteristic functions. By definition X and Y are independent if and only if $f_{X,Y}(s,t) = f_X(s)f_Y(t)$, where $f_{X,Y}(s,t) = \mathbb{E}\left(e^{i(\langle s,X\rangle+\langle t,Y\rangle)}\right)$ is the joint characteristic function of X and Y. Then we can say X and Y are independent if and only if

$$\left\|f_{X,Y} - f_X f_Y\right\|^2 := \iint \left|f_{X,Y}(s,t) - f_X(s)f_Y(t)\right|^2 w(s,t)\,ds\,dt = 0$$

for any positive weight $w(s,t)$ of our choice.

Given a sample $x_1,\dots,x_n,y_1,\dots,y_n$, we can approximate $f_X(s) \approx f_X^{(n)}(s) := \frac{1}{n}\sum_j e^{six_j}$ and $f_Y, f_{X,Y}$ similarly. Then substituting the sample characteristic functions and choosing weight $w(s,t) = \left(c_p c_q |t|_p^{1+p} |s|_q^{1+q}\right)^{-1}$, where c_p and c_q are normalizing constants, the above integral evaluates to

$$\left\|f_{X,Y}^{(n)} - f_X^{(n)} f_Y^{(n)}\right\|^2 = \frac{1}{n^2}\sum_{i,j}\left(a_{ij} - \overline{a_{i\cdot}}\right)\left(b_{ij} - \overline{b_{\cdot j}}\right)$$

where $a_{ij} := |x_i - x_j|$ and $b_{ij} := |y_i - y_j|$. This quantity is called the (empirical) *distance covariance* of the two samples.

ABOUT THE AUTHOR

My name is Yanitsa Pehova and I am about to start my PhD in Combinatorics. Working on this project was extremely useful to me as a sanity check of what real-world mathematics looks like, and what kind of skills I would need to apply my knowledge to other fields. I found the challenge of communicating mathematics to non-mathematicians highly non-trivial, and I learnt a series of useful practical tricks that aren't necessarily part of the maths Tripos.