STATISTICAL SUPPORT

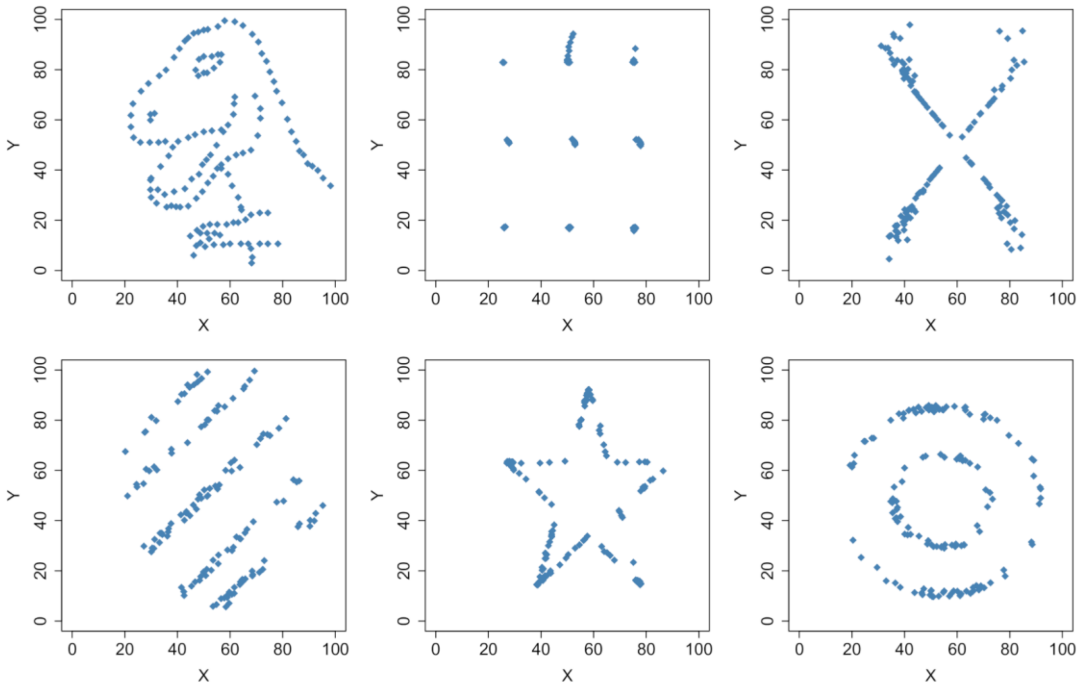
SWISS CENTER FOR AFFECTIVE SCIENCES

Nonlinear correlation

E-mail distributed on 18-10-2022

Dear all,

The Datasaurus Dozen is fast becoming a classic dataset to illustrate the danger of skipping out on visual inspection of your data. These 12 samples were all derived from the same source data (the "datasaurus"), but have identical means, standard deviations, and ≈0 correlation! In their paper, Matejka & Fitzmaurice (2017) proposed an algorithm that can rearrange the datasaurus while preserving the original descriptive values entirely.



The datasaurus highlights not only the importance of plotting, but also the crucial distinction between uncorrelatedness and **non-independence**. That is, X and Y can have 0 linear Pearson correlation and be strongly dependent. Such strong dependence can take on **non-linear** and sometimes **non-functional** forms (e.g., the T-rex and the star in the 6 examples above).

For large data sets, a sweeping visual check may be cumbersome and researchers resort to correlation matrixes for quick inspection. Previously, I have promoted the use of correlation mosaics to facilitate this and uncover structure. This can be extended to non-linear measures, such as **Spearman rank correlation** (RCOR), and distance correlation (DCOR). RCOR is robust against outliers and sensitive to monotone nonlinearity (i.e., strictly increasing or decreasing trend). DCOR is sensitive to non-monotone non-linearity and some non-functional trends. Distance correlation has been implemented in the R package energy (dcor function), and I provide a wrapper function here that generalizes to multivariate data sets (mdcor function; requires the foreach package).

A good standard practice when inspecting bivariate correlations for your data is to contrast the results to the RCOR and DCOR matrixes. Large discrepancies in values between these three statistics may indicate outliers or nonlinearity and should be visually inspected. **Caution:** DCOR is always a non-negative number between 0 and 1. It is best compared to *absolute* PCOR/RCOR!

Best,
Ben

Ben Meuleman, Ph.D.
Statistician
Swiss Center for Affective Sciences
University of Geneva | Campus Biotech
Chemin des Mines 9 | CH-1202 Genève
ben.meuleman@unige.ch | +41 (0)22 379 09 79

```
######################################
## MULTIVARIATE DISTANCE CORRELATION
######################################

library(energy)
library(foreach)
mdcor <- function(V) {
      nms <- colnames(V)
      d <- ncol(V)
      out <- foreach(i=1:d,.combine="c") %:% foreach(j=1:d,.combine="c") %do%
        {
          nm <- complete.cases(V[,c(i,j)])
          dcor(V[nm,i],V[nm,j])
        }
      out <- matrix(out,d,d)
      rownames(out) <- nms
      colnames(out) <- nms
      out
}
```