Telecom Paris **Master Data-Science**
2019/2020 **Pavlo Mozharovskyi**

TP 1 : Introduction to robust statistics

- Discovering R -

You can use the following links to discover more about the programming language `R`.

- https://cran.r-project.org/doc/manuals/r-release/R-intro.pdf
- https://cran.r-project.org/doc/manuals/r-release/R-lang.html.
- https://cran.r-project.org/doc/manuals/r-release/R-exts.html.

- Nuclear Accidents data set -

1) Load the Nuclear Accidents data set from file "NuclearPowerAccidents2016.csv". Retain only 10 most recent observations with available data on the accident cost. Apply the natural logarithm transform to this data to consider them on an equalized scale.
2) Plot the data, suggest which observation is an outlier.
3) Is the data in general position?
4) Plot two normal quantile-quantile plots (QQ-plots), *i.e.* the data *versus* corresponding theoretical quantiles of the standard normal distribution, one for the entire data set with 10 observations and another one without the outlier (for 9 observations). On both plots, try fit a straight line to the observations. Comment the obtained plots.
5) Compute the mean, the median, the standard deviation, and the interquantile range for both data sets, with and without the outlier. Present the results in a compact form.
6) Program a univariate oulier detection, with both non-robust and robust estimates. Does it detect suggested outlier?
7) For both data sets plot (empirically) the sensitivity curve of the mean, the median, the standard deviation, the interquantile range, and the medcouple. Comment.

- Data on the Length of Stay in the hospital -

1) Load the data set of the length of stay (in days) for 201 patients at the University Hospital of Lausanne during the year 2000; accessible as `los` in `R`-package `robustbase`.
2) Plot its histogram, boxplot and adjusted boxplot.
3) For this data set, calculate the interquantile range and program the computation of the medcouple.
4) For this data set, plot (empirically) the sensitivity curve of the interquantile range and of the medcouple.

- Normal and elliptical data -

1) Consider the following multivariate normal distribution:

$$\mathcal{N}(\left[\begin{smallmatrix}1\\1\end{smallmatrix}\right], \left[\begin{smallmatrix}1 & 1\\1 & 4\end{smallmatrix}\right]),$$

which we will call MVN1 in the sequel.

2) Draw a sample consisting of 250 independent identically distributed points from MVN1.

3) For this sample, compute the mean and the covariance matrix. Comment on the quality of estimation (*i.e.* how far the estimate is from the population parameter).

4) Using these estimates, plot a classic tolerance ellipsoid for this sample.

5) A random vector X in $\mathbb{R}^d$ is said to be generated from a **multivariate Student-t distribution** with ν degrees of freedom with center $\boldsymbol{\mu} \in \mathbb{R}^d$ and scatter matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{d\times d}$ if it can be represented as:

$$X = \boldsymbol{\mu} + \sqrt{\frac{\nu}{W_\nu}} Z,$$

where W_ν is a variable following chi-squared distribution with ν degrees of freedom and Z follows a d-variate normal distribution centered in the origin and with the covariance matrix $\boldsymbol{\Sigma}$, $\mathcal{N}(0_d, \boldsymbol{\Sigma})$, W_ν and Z being independent.

Student-t distribution with 1 degree of freedom is called the Cauchy distribution.

For more information on chi-squared distribution an a related Gamma distribution see: https://en.wikipedia.org/wiki/Chi-squared_distribution and https://en.wikipedia.org/wiki/Gamma_distribution.

Program a function that generates a data set from a multivariate Student-t distribution and takes as arguments center $\boldsymbol{\mu}$, scatter matrix $\boldsymbol{\Sigma}$, number of degrees of freedom ν, and number of points to generate.

Draw a data set from the bivariate Cauchy distribution with the same parameters as MVN1 (let us call this distribution MVC1) containing 250 points and plot it.

6) For this sample, compute the mean and the covariance matrix, plot a classic tolerance ellipsoid, *i.e.* using the Mahalanobis distance. Comment on the quality of estimation.

7) Program the Stahel-Donoho estimator and use it to estimate the center and the scatter for the both samples (use 1000 random directions to estimate the maximum). Plot the tolerance ellipsoid. Comment on comparison of the results with the Mahalanobis distance.

- Animal data -

1) Load the data set on the weight of the body and of the brain for 28 species of animals, which is accessible as `Animals` in `R`-package `MASS`. Plot the data.

2) Compute its mean and covariance matrix using both moment and Stahel-Donoho estimates.

3) Plot the two corresponding tolerance ellipsoids on the same plot. Comment the plots.

4) Compute the center and scatter estimates and plot tolerance ellipsoids as well with the MCD (on the same plot), for three different values of α (*e.g.* 0.5, 0.75, and 0.95). Comment.

- Hawkins-Bradu-Kass data set -

1) Load the Hawkins-Bradu-Kass data set, which is accessible as `hbk` in `R`-package `rrcov`. Plot it using the pairs plot.

2) Calculate the PCA estimates usin the classical PCA.

3) Plot the eigenvalues, the score distances, as well as the scores of the two first principal components.

4) Further, plot the score distances in *versus* to the plane of the first two principal components.

5) Repeat the two previous steps with the MCD-robust PCA, for 3 different values of α. Conclude on outliers.