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

TP 2 : Data depth and extreme values

- 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.

- Data depth -

- Different notions of statistical depth function are implemented in a number of `R`-packages, such as `ddalpha`, `mrf.Depth`, `DepthProc`, `depth`, `fda.usc`, ...
- You are free to choose any of the implementations or implement certain parts on your own.

1 First examples of data depth

1) Consider the following multivariate normal distributions:
   - $\mathcal{N}(\left[\begin{smallmatrix}0\\0\end{smallmatrix}\right], \left[\begin{smallmatrix}1&1\\1&4\end{smallmatrix}\right])$ which we will call MVN1 in the sequel;
   - $\mathcal{N}(\left[\begin{smallmatrix}2\\2\end{smallmatrix}\right], \left[\begin{smallmatrix}1&1\\1&4\end{smallmatrix}\right])$ which we will call MVN2 in the sequel;
   - $\mathcal{N}(\left[\begin{smallmatrix}2\\2\end{smallmatrix}\right], \left[\begin{smallmatrix}4&4\\4&16\end{smallmatrix}\right])$ which we will call MVN3 in the sequel.

   From each of them, draw a data set containing 25 points and plot it, each on a separate plot.
2) Apply the Mahalnobis depth to each of the data sets from question 1, *i.e.*, calculate the depth of each point of the data set with respect to the data set itself; plot the data sets indicating on the plots depth value (rounded to two digits after the period) next to each point.
3) Repeat question 2 using zonoid and Tukey depth, comment on the differences.

2 Estimation of multivariate location

1) Draw a sample consisting of 250 independent identically distributed points from MVN3. Plot the data.
2) 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.

Compare the plot with the one from question 1.

3) In addition to the bivariate data sets created in questions 1 and 2, use the following (artificially created) three-dimensional data from table 1. For these three data sets, compute depth medians using Mahalanobis, zonoid, and Tukey depth. For a data set $\boldsymbol{X} \subset \mathbb{R}^d$, a **depth median** is any point $\boldsymbol{y} \in \mathbb{R}^d$ such that

$$\boldsymbol{y} \in \arg\max_{\boldsymbol{x} \in \mathbb{R}^d} D(\boldsymbol{x}|\boldsymbol{X})\,.$$

#	x_1	x_2	x_3
1	1	0	0
2	0	1	0
3	0	0	1
4	1.5	1.5	1.5
5	0.309151	0.286697	0.653584
6	0.733359	0.040291	0.316318
7	0.15937	0.304677	0.558091
8	0.056376	0.19044	0.912733
9	0.517479	0.533977	0.19188
10	1.011993	0.058608	0.099067
11	0.117582	0.164475	0.92203
12	0.175112	0.918897	0.221602
13	0.240206	0.454373	0.1701
14	0.906328	0.056292	0.11981

Table 1: An artificial three-dimensional data set.

4) For the data sets from question 3, compute the projection depth median. In addition to the existing (approximate) implementation of the projection depth, use an optimization technique, *e.g.* the Nelder-Mead algorithm or a fine grid, to find the location with the highest depth value.

5) Construct a table which compares the medians found in questions 3 and 4. For the data sets from questions 1 and 2, calculate the distance of the median estimates to the estimated parameters. How close are they for different depths?

3 DD-plot

1) For two data sets $\boldsymbol{X}, \boldsymbol{Y} \subset \mathbb{R}^d$ and a depth function $D(\cdot|\cdot)$, consider a **depth-vs-depth** plot (or DD-plot, shortly):

$$\boldsymbol{Z}\{(z_1, z_2) \,:\, z_1 = D(\boldsymbol{z}|\boldsymbol{X})\,,\, z_2 = D(\boldsymbol{z}|\boldsymbol{Y})\,,\, \boldsymbol{z} \in \boldsymbol{X} \cup \boldsymbol{Y}\}\,,$$

which is itself a bivariate data set.

Draw two i.i.d. samples from MVN1 each containing 250 points. For them, plot two DD-plots, one using Mahalanobis depth and one using Tukey depth. On each DD-plot, draw points belonging to the two different data sets with different colors. Conclude on the obtained visualization.

2) Draw two i.i.d. samples from MVN1 and MVN2 each containing 250 points. Plot the two corresponding DD-plots using Mahalanobis and Tukey depth. Conclude on the obtained visualization.

3) Let MVC2 be a Cauchy distribution with the same parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ as in MVN2. Draw two i.i.d. samples from MVC1 and MVC2 each containing 250 points. Plot the two corresponding DD-plots using Mahalanobis and Tukey depth. Conclude on the obtained visualization and on the difference between the two plots.

- Extreme values -

4 Available tools, packages in R

- Many packages : `ismev`, `extRemes`, `evd`, `fExtremes`, `EVIM`, `Xtremes`, `HYFRAN`, `EXTREMES` , ... All of them available on http://cran.r-project.org/.
- In this tutorial: we use mainly `evd` ("Functions for extreme value distributions"), and `ismev`("Introduction to Statistical Modeling of Extreme Values", from J.S. Coles's book)
- Main functions from pacakge `evd` (*resp.* `ismev`);
  - fitting a GEV model: `fgev` (*resp.* `gev.fit`)
  - fitting a GPD model ('Peaks over threshold') : `fpot` (*resp.* `gpd.fit`)
  - Fitting a Poisson model: `fpot(..., model=''pp'')` (*resp.* `pot.fit`)
  - Direct estimation of quantiles with `evd` : *e.g.* for a probability of exceedance $p = 0.01$) `fgev(..., prob=0.01)` or `fpot(...,prob=0.01)`.
  - Looking for an adequate threshold for the POT or Poisson model: `tcplot(..., model="pp")`, `tcplot(..., model="gpd")` (*resp.* `gpd.fitrange, pp.fitrange`)
  - Graphical diagnostics before fitting a model: Mean residual life plot `mrlplot` (*resp.* `mrl.plot`)
  - Graphical diagnostics after fitting a model: `plot("fitted")` (*resp.* `gev.diag, gpd.diag, pp.diag`)
  - Model comparison *via* likelihood ratios: `anova` (`evd` only), `profile("fitted")` and `plot(profile("fitted"))` (*resp.* `gev.prof, gpd.prof`).

5 Before starting

```
> install.packages("evd")
> install.packages("ismev")
> install.packages("evir")
> install.packages("tseries")
> library(ismev); library(evd); library(tseries)
> help.start()
```

6 Examples

6.1 Fire damage data in Denmark

The dataset `danish` is available in `evir`: it contains the largest insurance claims related to fire damages between 1980 and 1990. To avoid package compatibility issues, do not load the full package `evir`, only the considered dataset.

```
> data(danish, package = "evir")
```

1) Check quickly that a stationary, independent model is suitable and that a gaussian model is not. To do so, plot the time seies, (`plot(ts(danish))` and compute the auto-correlation function (`acf`). You may use the Dickey-Fuller stationarity test.

```
?adf.test
```

For the Gaussianity test, use e.g. a QQ-plot and a Shapiro test

```
?qqnorm
?shapiro.test
```

2) Exploratory analysis of the Peaks-Over-Thresholds. quantile plot in the Pareto model, Mean residual life plot, Hill plot. Is it appropriate to use an extreme values model for this data-set?

```
?mrlplot
?paretoPlot
?hill
```

3) Fit a GPD model on the threshold exceedances (the threshold has to be determined in view of the data). Plot the graphical diagnostics relative to the model fit.

```
?tcplot
?fpot
```

4) Perform an analysis of variance (function `anova`) comparing the POT model and the POT model fitted under the constraint $\xi = 0$, which should involve the following fitted model

```
fpotDanish2 <- fpot(danish,threshold=u, shape=0)
```

On the other hand, plot the profile likelihood by a call to the function `profile` . Does the profile likelihood method lead to the same conclusion as the analysis of variance?

5) Give an estimate for the 50 - years return level based on the estimated parameters and a plug-in method. Compare the answer using an analytic expression derived from the lecture notes and the output of the function `qgpd`.

6) Give an estimator of the probability of occurrence over a year of an excess above

- Twice the maximum observed on the considered period
- the maximum observed on the considered period

Compare with the empirical estimator.

6.2 Annual maxima of sea level at Port Pirie (Australia)

Available in `evd` and `ismev` (NB: in ismev, the data are of matrix type.

```
> data(portpirie)
```

1) Same question as in the previous exercice.

2) How to investigate graphically the adequacy of an extreme value model? What do you think of the hypothesis $\xi = 0$? To answer, use the function `gumPlot`) and an analysis of variance as in Exercise 1. Confirm your diagnostic by a call to the function `profile` which plots the profile likelihood.

3) Give an etimate of the one thousand year return level. Investigate the sensitivity to model choice ($\xi = 0$ versus $\xi \neq 0$)

6.3 Rain data in south England

Available in `ismev` . Daily data between 1914 and 1962.

```
> data(rain); ?rain
```

1) Same question as in the previous exercises.

2) Propose two methods to estimate the 100 year return level based respectively on an analysis of block-maxima and of peaks over threshold.

3) Compare the estimations in both models.