Data mining and multivariate statistical analysis 2018
1st practical

A report must be uploaded on TEIDE before the deadline.

The report should contain graphical representations, which are very important in statistics. For each graph, axis names should be provided as well as a legend when it is appropriate. Figures should be explained by a few sentences in the text. Answer to the questions in order and refer to the question number in your report.

Computations and graphics should be performed with the software R.

The report should be written using the Rmarkdown format. It is a file format that allows users to format documents containing text, R instructions and the results provided by R when evaluating instructions. The set of R statements is included in the .rmd document so that it may be possible to replicate your analyzes using the .rmd file. From your .rmd file, you are asked to generate an .html file for the final report. The set of .rmd commands and the procedure to generate .html files is explained in the Rmarkdown cheatsheet. In TEIDE, you are asked to submit both the .rmd and the .html files. In the html file, you should limit the displayed R code to the most important instructions.

1st practical: p-values, correlated predictors

1. Simulated data and p-values

a) Set the seed to 0. (`set.seed(0)`)
Simulate 6,000 independent random vectors in dimension 201with independent components, with mean 0 and variance 1. Store them into a matrix, then into a data frame with 6,000 lines and 201 columns. Each of these columns is referred to as a "variable".

b) Define a linear model using the last 200 variables to predict the first one. In the report, write a mathematical equation (do not write R code!) to define this model. Write a mathematical equation defining the true regression model associated with the data. What is the difference between both?

c) Estimate the linear model using the last 200 variables to predict the first one.
Using some R code, compute the number of coefficients assessed as significantly non-zero at level 5%.

Is this result expected? Why? Is this a problem? Why?

Hint: `summary(reg)$coefficients`

2. Correlated predictors and confidence intervals

a) Set the seed to 3 (`set.seed(3)`). Simulate a sample of size n=1000 of the following model:

$$X_{1,i} = \varepsilon_{1,i}; \varepsilon_{1,i} \sim N(0,1)$$
$$X_{2,i} = 3X_{1,i} + \varepsilon_{2,i}; \varepsilon_{2,i} \sim N(0,1)$$
$$Y_i = 2 + X_{1,i} + X_{2,i} + \varepsilon_{3,i}; \varepsilon_{3,i} \sim N(0,1)$$

where the $(\varepsilon_{k,i})_{k,i}$ are independent random variables.
Plot the cloud of points $(X_{1,i}, X_{2,i})_{i=1,\ldots,n}$. What is its shape? Why?
Simulate a new sample of size n =10 that will be used instead of the larger previous sample in the questions below.

b) Estimate both models
$Y_i = \beta_0 + \beta_1 X_{1,i} + \tilde{\varepsilon}_{3,i}$ and $Y_i = \beta_0' + \beta_2' X_{2,i} + \tilde{\varepsilon}_{3,i}'$. Comment on the effect of both predictors.

c) Estimate the model defined in question 2.a). Comment on the effect of both predictors, and provide two different points of view. Are these points of view consistent? What could explain this?

d) On two separate graphs, plot the true probability density functions (pdf) of the distributions of $\hat{\beta}_1$ and $\hat{\beta}_2$ given X_1 and X_2 in the model defined in question 2.c). Represent the quantiles of order 0.025 and 0.975 on the x-axis. What particular result of interest do you observe?

e) Plot ellipses (centered on the true mean) that have probabilities 0.5, 0.9, 0.999 to contain $\hat{\beta}_1$ and $\hat{\beta}_2$ given X_1 and X_2 in the model defined in question 2.c). Add the ellipse associated with probability one minus the p-value of Fisher test. What particular result of interest do you observe? Why does not (0,0) pass through the last ellipse? What to conclude from questions 2.d) and 2.e)?

You may use the code provided on the chamilo webpage of the course.