Faculty of Applied Science

MATH2021-1 High-dimensional statistics

Project 1 : Exploratory Data Analysis

Teacher :
Gentiane HAESBROECK

Group :
Romain LAMBERMONT
Arthur LOUIS

October 26, 2022

Contents

List of Figures

List of Tables

1 Presentation of the data

1.1 Discussion on the data

The data we used in this project is a subset of the data collected by the ENEA (National Agency for New Technologies, Energy and Sustainable Economic Development) alonside a road in a polluted area of Italy. This dataset is available on the UCI repository[1]. The data was harvested using a multicensor device and reference analyzers. The data was collected between March 2004 and February 2005.

There are 5 couples of variables in the dataset, each couple is composed of a variable measured by the reference analyzer and the corresponding variable measured by the multicensor device. The values represent the hourly average concentration of each variable. In addition to the variables couples, we have 3 other variables representing the temperature and the humidity (both relative and absolute). The different values are stored in the following columns of our dataset :

- `CO(GT)` : concentration of CO in the air (in mg/m^3)
- `PT08.S1(CO)` : average sensor response (nominally CO targeted)
- `NMHC(GT)` : concentration of non-methane hydrocarbons in the air (in $\mu g/m^3$)
- `C6H6(GT)` : concentration of benzene in the air (in $\mu g/m^3$)
- `PT08.S2(NMHC)` : average sensor response (nominally NMHC targeted)
- `NOx(GT)` : concentration of NOx in the air (in parts per billion)
- `PT08.S3(NOx)` : average sensor response (nominally NOx targeted)
- `NO2(GT)` : concentration of NO_2 in the air (in $\mu g/m^3$)
- `PT08.S4(NO2)` : average sensor response (nominally NO_2 targeted)
- `PT08.S5(O3)` : average sensor response (nominally O_3 targeted)
- `T` : temperature (in °C)
- `RH` : relative humidity (in %)
- `AH` : absolute humidity

On top of that, we created a binary indicator per variable measured by the reference analyzers which is equal to 1 when the measured value is above the median of the variable and 0 otherwise. The binary values are stored in the following columns :

- `HIGH_CO` : binary indicator for CO (above/under median)
- `HIGH_NMHC` : binary indicator for NMHC (above/under median)
- `HIGH_C6H6` : binary indicator for benzene (above/under median)
- `HIGH_NOx` : binary indicator for NOx (above/under median)
- `HIGH_NO2` : binary indicator for NO2 (above/under median)

1.2 Link between the variables

The values for each couple are obviously going to be quite correlated, as they are measuring the same thing. Furthermore, the values for the binary indicators are going to be highly correlated with the values of the corresponding measurements from the reference analyzers (as the binary indicator is equal to 1 when the value is above the median and 0 otherwise).

[1] https://archive.ics.uci.edu/ml/datasets/air+quality