



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) alongside a road in a polluted area of Italy. This dataset is available on the UCI repository¹. 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 $\mu\text{g}/\text{m}^3$)
- C6H6(GT) : concentration of benzene in the air (in $\mu\text{g}/\text{m}^3$)
- 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 NO₂ in the air (in $\mu\text{g}/\text{m}^3$)
- PT08.S4(NO2) : average sensor response (nominally NO₂ targeted)
- PT08.S5(O3) : average sensor response (nominally O₃ 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>

2 Information about missing data

We have a total of 2.1% of missing values but this number is overestimated because the binary indicators are taken into account. The real ratio is 1.7% without this indicators. The missing values are due to hardware problems related to the measuring instruments and to the fact that the data was collected in a real environment.

Figure 1: Missing data

3 Exploratory data analysis

3.1 Statistical analysis

	vars	n	mean	sd	median	trimmed	mad	min	max	range
CO(GT)	1	191	2.748691	1.596801	2.50	2.560784	1.33434	0.5	8.1	7.6
PT08.S1(CO)	2	200	1339.000000	255.446559	1332.50	1328.356250	233.50950	831.0	2040.0	1209.0
NMHC(GT)	3	183	160.158470	139.745774	122.00	138.448980	118.60800	7.0	685.0	678.0
C6H6(GT)	4	200	12.254000	8.274006	11.05	11.318750	7.33887	1.0	39.2	38.2
PT08.S2(NMHC)	5	200	1016.950000	281.940276	1017.50	1005.556250	278.72880	501.0	1754.0	1253.0
NOx(GT)	6	191	175.842932	94.999980	161.00	168.830065	85.99080	16.0	478.0	462.0
PT08.S3(NOx)	7	200	1003.195000	278.431170	945.00	976.950000	234.99210	537.0	1918.0	1381.0
NO2(GT)	8	191	115.612565	34.357971	119.00	116.549020	35.58240	28.0	194.0	166.0
PT08.S4(NO2)	9	200	1671.040000	305.901187	1622.50	1641.750000	237.21600	1134.0	2679.0	1545.0
	vars	n	mean	sd	median	trimmed	mad	min	max	range
PT08.S5(O3)	1	200	1233.2450000	389.2906253	1204.5000	1222.0375000	384.734700	384.0000	2359.0	1975.0
T	2	200	15.1965000	5.5702402	14.3000	14.8068750	5.189100	6.1000	29.0	23.0
RH	3	200	49.8030000	15.1352426	53.9000	50.6387500	15.270780	14.9000	81.0	63.0
AH	4	200	0.8085450	0.1059962	0.8125	0.8092338	0.104375	0.5237	1.0	0.3112
HIGH_CO	5	191	0.7225131	0.4489355	1.0000	0.7777778	0.000000	0.0000	1.0	0.0000
HIGH_NMHC	6	183	0.9945355	0.0739221	1.0000	1.0000000	0.000000	0.0000	1.0	0.0000
HIGH_C6H6	7	200	0.6500000	0.4781665	1.0000	0.6875000	0.000000	0.0000	1.0	0.0000
HIGH_NOx	8	191	0.9947644	0.0723575	1.0000	1.0000000	0.000000	0.0000	1.0	0.0000
HIGH_NO2	9	191	0.5968586	0.4918179	1.0000	0.6209150	0.000000	0.0000	1.0	0.0000

3.2 Graphical analysis

3.3 Correlation structure of the data

3.4 Outlying observations using Mahalanobis distance

4 Correlation analysis with data reduction

4.1 Choice between PCA and t-SNE

4.2 2D plot of the data