UNIVERSITY OF LIÈGE

Project 2 - Correlation structure and dimension reduction

MATH2021-1 - High-dimensional data analysis

Robust outlier detection : $5^{75}/7$.
Correlation structure : $4^{75}/6$
2D visualisation : $8^{75}/11$.
Overall quality : 1/1

$\Rightarrow 20^{25}/25$

Yann CLAES (s161317)
Gaspard LAMBRECHTS (s161826)
François ROZET (s161024)

MSc in Data science and engineering
Academic year 2019-2020

Data

We have chosen to use the same data set [1] as in the first project for this second project. As a reminder, this data set describes the evolution of major air pollutants and meteorological variables measured by the air-quality monitoring station of Shunyi, Beijing.

After data handling, our dataset includes $n = 414$ observations of 17 parameters. However only $p = 10$ (PM2.5, PM10, SO_2, NO_2, CO, O_3, temperature, pressure, dew point and wind speed) of these parameters will be considered as quantitative variables in the following. That choice will be explained later on.

1 Robust outlier detection

The robust mean and covariance estimations were performed using the reweighted MCD estimator (which is used by default in `R` when calling the function `cov.rob` with the `method` parameter set to "mcd") with a value of h equal to 214. Indeed, as stated in the paper [2], the MCD estimator is most robust when h, the number of good observations to be considered, is equal to $h = \lfloor \frac{n+p+1}{2} \rfloor$ which, in our case, leads to $h = 214$, i.e. a coverage of 51.7 %. Furthermore, the MCD estimator is able to resist to $n - h$ outliers : in our data set $n - h = 200$ observations.

As was already done for the first part of the project, we decided to remove from the quantitative variables the temporal ones. In fact, we have computed the outlying rate, for both classic and robust distances, for both the quantitative data containing and not containing these time variables. The results are in Table 1.

	Classic rate	Robust rate
With time	0.0942	0.3623
Without time	0.1014	0.3478

Table 1 – Comparison of outlying rates with and without time variables

From these results, it thus seemed that both quantitative sets gave similar results. On Figure 1 is the corresponding DD-plot comparing both robust and non-robust Mahalanobis distances.

From this figure, we can see that the classic outlier detection effectively considers less data points as being outlying with respect to the robust detection, the respective rates being on the second line of Table 1. This clearly illustrates the masking effect where the classical estimates are so influenced by the outlying observations that the detection technique (here, the Mahalanobis distances) cannot spot them. A first thing we can say about Figure 1 is that the observations classified as outlying by both methods can be, for sure, classified as true outliers. In addition, it should be recalled that all the outliers detected by the classic method will inevitably be detected by the robust method as well.

In order to spot the potential influence of each variable on the outlying data points, we drew boxplots (cf. Figure 7) comparing variables distributions in the non outlying and outlying case.

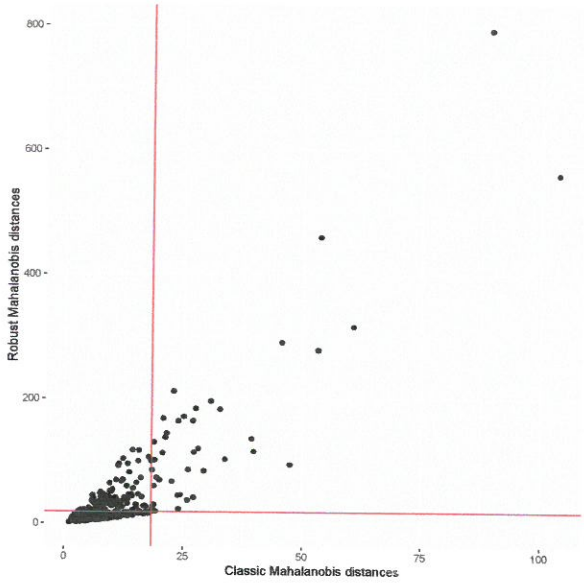


Figure 1 – DD-plot of the Mahalanobis distances.

As far as the atmospheric variables (temperature, dew point, pressure and wind speed) are concerned, it seems that they follow similar yet not identical distributions for both outlying and non outlying data points. Nevertheless, we observe a slight increase of the median temperature and dew point when outliers were removed. From these observations, we could infer that atmospheric variables don't play that important of a role in the outliers detection, at least for pressure and wind speed.

Conversely, for the pollutant concentrations[1], we observe some significant alterations of the distributions for outlying and non outlying data points. Indeed, all pollutants but O_3 have seen their median, third quartile and maximum drop dramatically when we removed the outlying points. For O_3, we observe the opposite behaviour.

We could say that high pollutant concentration values (and low ozone concentration) mostly explain the outlying character of the data and, by looking at Table 2 in Appendix B, we see that these most extreme cases corroborate this hypothesis.

But, it should be noted that this analysis was conducted *variable by variable*, potentially ignoring the impact of one variable on the others. Indeed, we know from project 1 that pollutant concentrations are positively correlated (negatively for O_3). Therefore, it is not surprising to see them increase all at the same time and be considered as outlying by the MCD estimator.

Did you look after the outliers detected by the robust technique but not by the classic one?

2 Correlation structure of the quantitative variables

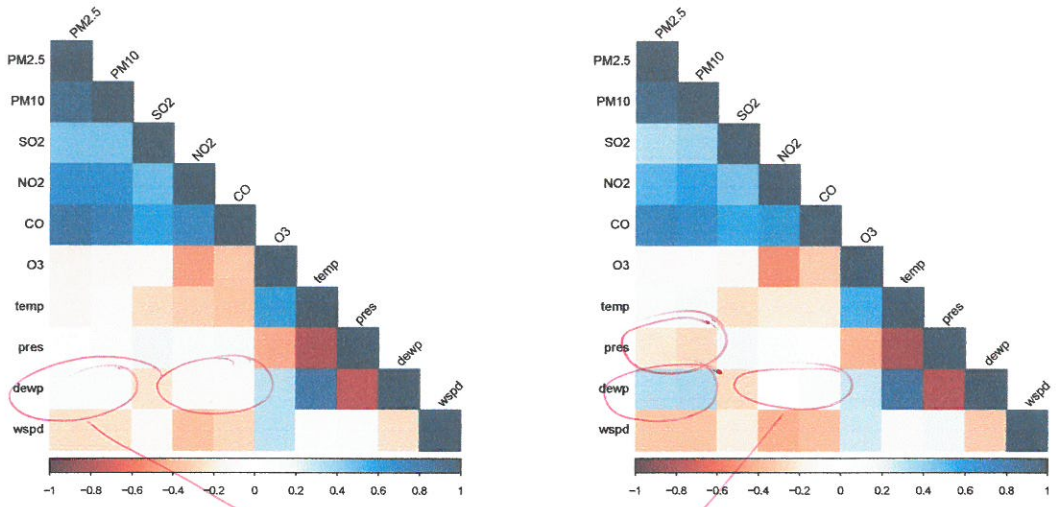
2.1 Robust estimation of the correlation matrix

A robust estimation of the correlation matrix has been computed using the same reweighted MCD estimator as in section 1 ($h = 214$).

As one can see in Figure 2, both matrices are *very* similar. This means that the outliers defined by the (reweighted) MCD estimator don't have much of an influence on the global correlation structure of the data set.

yes but still... it is interesting to further investigate the differences

[1] A more detailed analysis of the boxplots is proposed in appendix A.



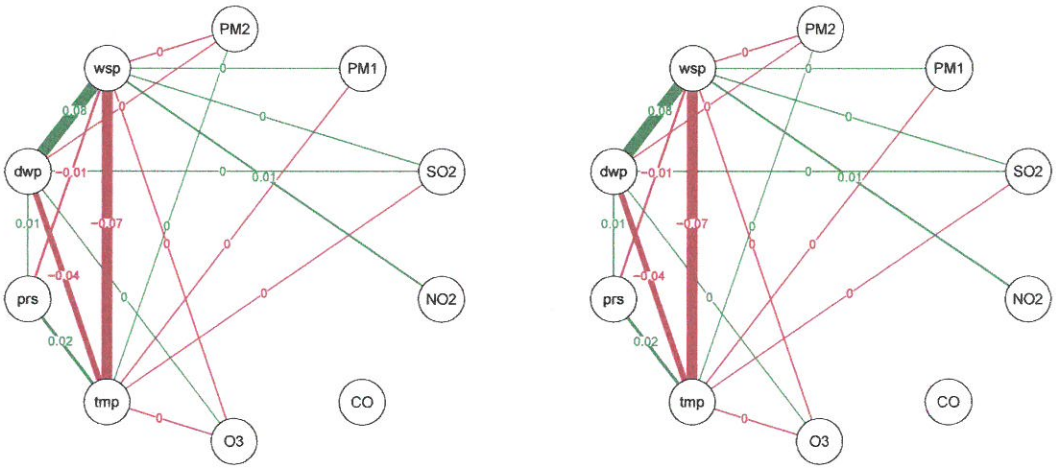
(a) Classic (b) Robust estimation

Figure 2 – Graphical representation of the classic correlation matrix and a robust estimation of it.

Still, there are some discrepancies. Can you explain/quantify them?

2.2 Graphic models

Using the `qgraph` function [3], we have constructed the graphical models[2] of the classical covariance matrix as well as an L_1-regularized estimation of it (cf. Figure 3).



(a) Classic (b) L_1-regularized estimation ($\lambda = 0$)

Figure 3 – Graphical models of the classic covariance matrix and an L_1-regularized estimation of it.

As one can see, the two models are *exactly* the same. Indeed, the penalization parameter used is $\lambda = 0$, therefore the solution to the regularized maximum-likelihood optimization problem is the classic covariance matrix which explains why those models are the same. However, there is yet to explain why we have used such λ.

There exists several proposals in the literature for selecting an appropriate value of λ but, since it is the only one we have properly studied and applied, we decided to find λ whose

[2] All drawn edges have a weight of at least 10^3.

BIC measure is minimal. Therefore, we computed the BIC measure for several values of λ and selected the one associated to the smallest BIC which appeared to be 0 (cf. Figure 8).

In fact, this result is not surprising at all. The goal of the L_1-regularization is to make the concentration matrix (inverse of the covariance matrix), and therefore the graphical model, sparse(-er) but, as we can observe in the Figure 3(a), the graphical model of the classic covariance matrix is *already* quite sparse.

Interpretation

Concerning the actual links drawn in the graphical models, we recognize some that we already had discussed in the project 1.

For example, the relations between the atmospheric variables (temperature, pressure and dew point) we had drawn in the first project are also visible in the graphical models. We also recognize the dependencies between some pollutants and the temperature.

There is also dependencies we hadn't noticed previously : the wind speed seems to have an influence over most other variables. Actually, this could explain why pollutant-to-pollutant relations are missing from the graphical models : if their concentrations are all dependent on the wind speed, pollutants should be conditionally independent (knowing the wind speed).

Discussion

However, these graphical models are based on the multivariate normality assumption which is questionable for our dataset. Indeed, some variables such as temperature, dew point and pressure seem to follow normal distributions but all others don't, as discussed in the first project.

But, this was considering the outliers as part of the dataset. If we don't, we see (cf. Figure 7) that PM2.5, PM10, NO_2 and O_3 are more likely to follow normal distributions. Therefore, the multivariate normality assumption isn't irrelevant, yet questionable.

3 Visualisation of the quantitative data in 2D

From the analysis detailed below, we notice that the principal component analysis seems to be far more informative by giving more interpretable results about the structure of the data while tSNE did not help us highlighting patterns in our data.

3.1 Principal component analysis

In order to apply the principal component analysis, it has been chosen to use the correlation matrix instead of the covariance matrix[3]. Indeed, the units and scales of our data

[3] As a consequence, each time that we talk about variance in the subsections below, it's about the variance of the normalized data

were far to different to use them as is. This can be seen on Figure 9, which can be found in Appendix A.

Then, it has been chosen to use the classic estimation of the correlation matrix instead of the robust one. This can be justified a priori by the small differences between the robust estimation and the classic estimation of the correlation matrix. It has also been verified a posteriori that the PCA using the robust estimation gave quite similar results.

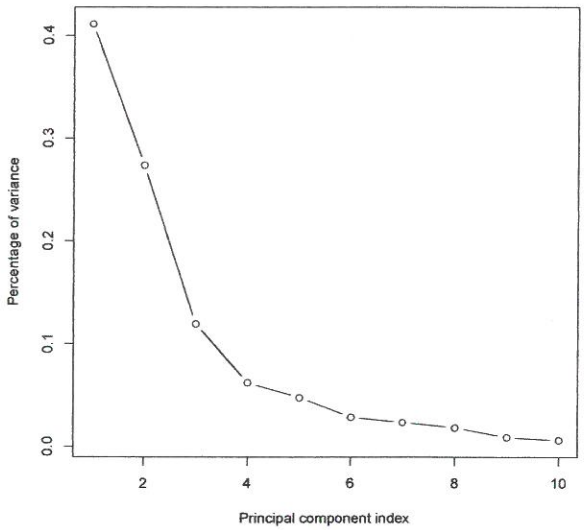


Figure 4 – Scree Plot of the eigenvalues of the correlation matrix.

The principal component analysis corresponds to finding the set of eigenvectors and eigenvalues of the correlation matrix. In the basis of the eigenvectors, the data will thus seem completely uncorrelated. The ordering of the principal components (eigenvectors) is in decreasing order of the eigenvalues, such that the first component is the most explanatory of the correlation in our data. On the scree plot of Figure 4, we can see the percentage of the total variance (of the normalized data, as we use the correlation matrix) for each component. It can be computed that the first two principal components explain 68.49% of the total variance (of the normalized data).

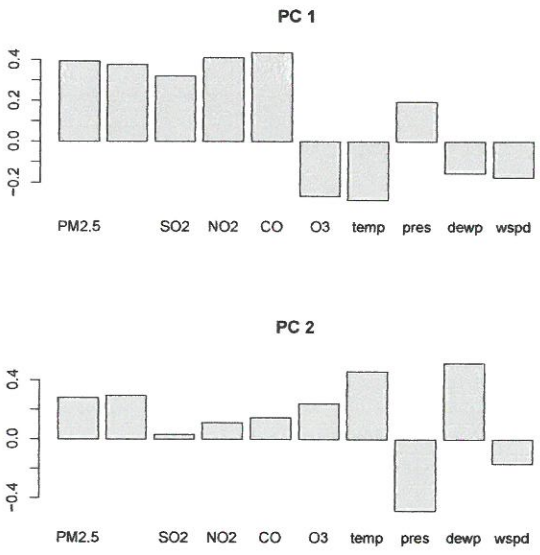


Figure 5 – Loadings of the variables in the first two principal components

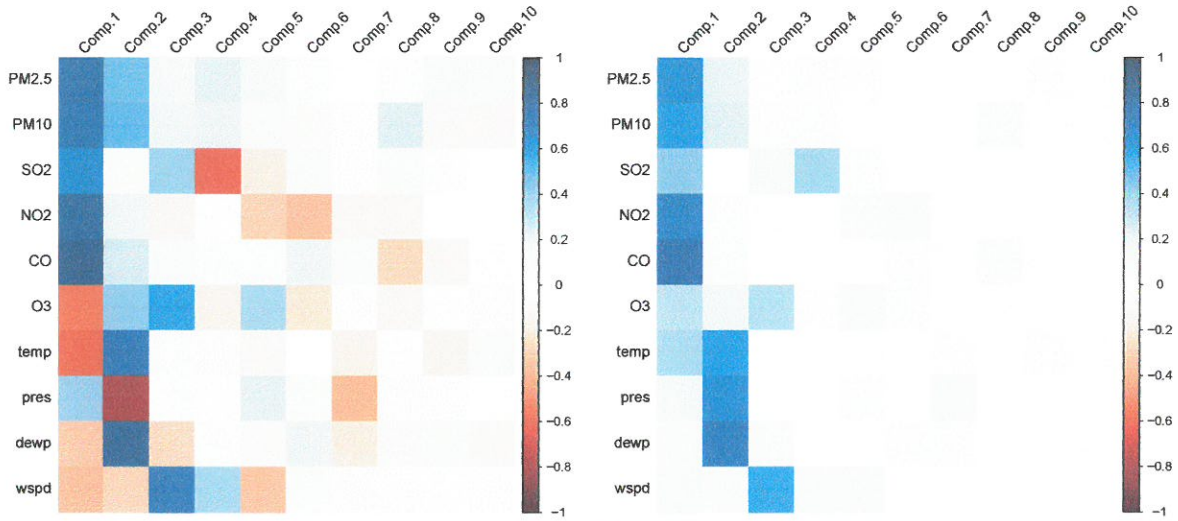
The loadings of the first two principal components can be used in order to get an interpretation of the meaning of these components. By looking at Figure 5 it can been seen that the first component is mainly constituted of all the pollutants (only the O3 having a negative coefficient, accordingly to the correlation structure between the pollutants discussed in the first part of the project), while the second component is mainly composed of the atmospheric variables: the temperature, dew point, and pressure (only the pressure having a negative coefficient, being anti-correlated with the dew point and the pressure, accordingly to the correlation structure between these three variables spotted in the first part of the project).

This is quite interesting to see that 68% of the variance of the data is contained in two independent components, the first corresponding roughly to the pollutants variable, and the second corresponding roughly to the atmospheric variables.

The high eigenvalues of each of these two components tells us about the trend that

have the data to vary according to these two independent components. The columns of the correlation plot of Figure 6a indicates the trend that have certain variables to vary together with the principal components, the first two columns gathering thus 68% of this variation.

On Figure 6b, the squared correlation coefficients between each variable and each principal component have been plotted. Each element represents thus the part of variability of each variable explained by each principal component. We clearly see that every pollutant and atmospheric variable has its variability mainly explained by the two first principal components.



(a) Correlation between each variable and each principal component

(b) Squared correlation between each variable and each principal component

All that is once again illustrated on the correlation circle of Figure 10.

3.2 t-distributed stochastic neighbor embedding

The tSNE algorithm has been used in order to decrease the number of dimension to three or two. No clear pattern has been observed for any perplexity. Moreover, our dataset having very few discrete data, no pattern has been detected by coloring the data points according to the rain, the moment of the day, the wind direction or the season of the year.

It has also been tested to select the 20 most outlying observations according to the robust method, in order to see if the outliers belong to some clusters of data points. Although the outliers were indeed located in a certain number of clusters of the tSNE plot instead of being uniformly distributed, they were not located in very few clusters neither. This could indicate that indeed some outliers are considered to be close in the quantitative data space, ending up in the same cluster in 2 dimensions. But that outliers could also come from very different locations of the quantitative data space, ending up in ~~the same~~ different cluster in 2 dimensions.

The result ~~with~~ of the tSNE algorithm can be seen on Figure 11 for different perplexity values.

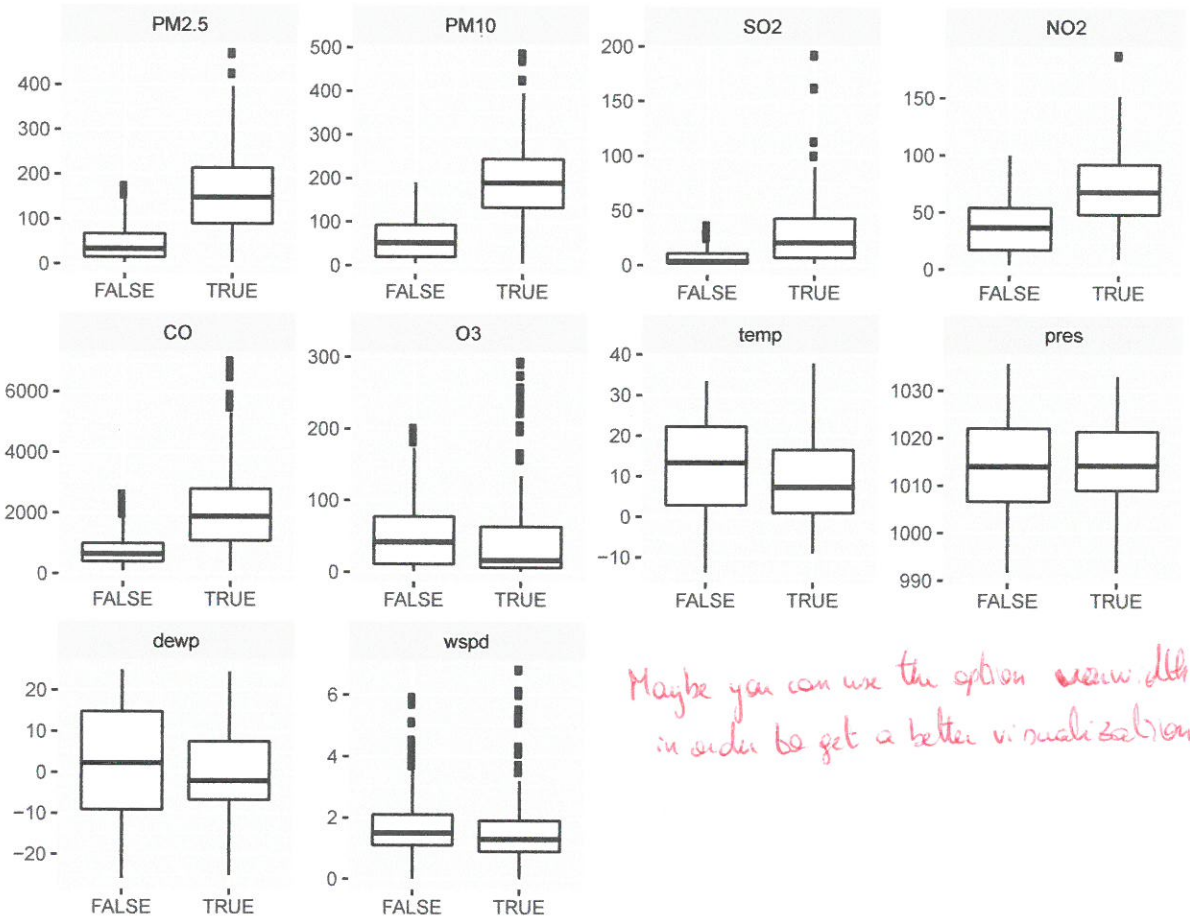
References

[1] *Beijing Multi-Site Air-Quality Data Set.* 2019. URL: https://archive.ics.uci.edu/ml/datasets/Beijing+Multi-Site+Air-Quality+Data.

[2] M. Hubert and M. Debruyne. *Minimum covariance determinant.* 2010. URL: https://wis.kuleuven.be/stat/robust/papers/2010/wire-mcd.pdf.

[3] *qgraph.* 2019. URL: https://cran.r-project.org/web/packages/qgraph/qgraph.pdf.

A Figures



Maybe you can use the option width=TRUE in order to get a better visualization.

Figure 7 – Boxplot of the quantitative variables for the non outlying observations (FALSE) and the robust outlying observations (TRUE).

Detailed analysis of Figure 7

Concerning the pollutant concentrations[4], we can clearly see that low O_3 concentrations are present in both outlying and non outlying observations, while higher values are mainly met in the non outlying ones. Their quartiles look similar, the median of the outlying values being lower as both cases have a similar number of very low values. It should however be noticed that about 7 % of outlying observations present concentration values above 200 $\mu g\,m^{-3}$ while none of the non outlying ones does. These extreme concentration values do not seem to attract the outlying observations to extreme outliers, as can be seen from Table 2.

As far as CO is concerned, 96 % of the non outlying concentrations are below 2000 $\mu g\,m^{-3}$ while only roughly 50 % of the outlying ones do. The other 50 % are thus linked to CO concentrations above 2000 $\mu g\,m^{-3}$ (with 30 % of them being above the maximum non

[4]All the following rate computations can be found in the source code.

outlying CO concentration), while only 4 % of the non outlying observations show values from 2000 to 2600 $\mu g\,m^{-3}$, the maximum concentration observed.

For the NO_2 concentrations, the median of the outlying observations is equal to 68 $\mu g\,m^{-3}$, while the one of the non outlying data points is 36.5 $\mu g\,m^{-3}$ with 12.5 % of these points being above the outlying median. About 20 % of the outlying points have a concentration bigger than the maximum, achieved by only 1 data point, of the non outlying ones.

Concerning the SO_2 concentrations, we can see that most of the non outlying observations are associated to very low values. Furthermore, about 34 % of the outlying concentrations lie above the maximum concentration observed for the non outlying data, equal to 36 $\mu g\,m^{-3}$, again attained by only 1 data point.

Finally, a similar behavior can be noticed for both PM2.5 and PM10 concentrations, which will be analyzed in pair since they are positively correlated, as was shown in the first part of the project. As far as PM10 concentrations are concerned, none of the non outlying data points have a concentration value bigger than the median concentration of the outlying ones, equal to 190.5 $\mu g\,m^{-3}$. It thus also means that 50 % of the outlying data points have concentrations above the maximum value observed in the non outlying data. Quite similar values are obtained for the PM2.5 concentrations: only 2.5 % of non outlying points have values above the outlying median, and 42 % of outlying points have values above the non outlying maximum PM2.5 concentration.

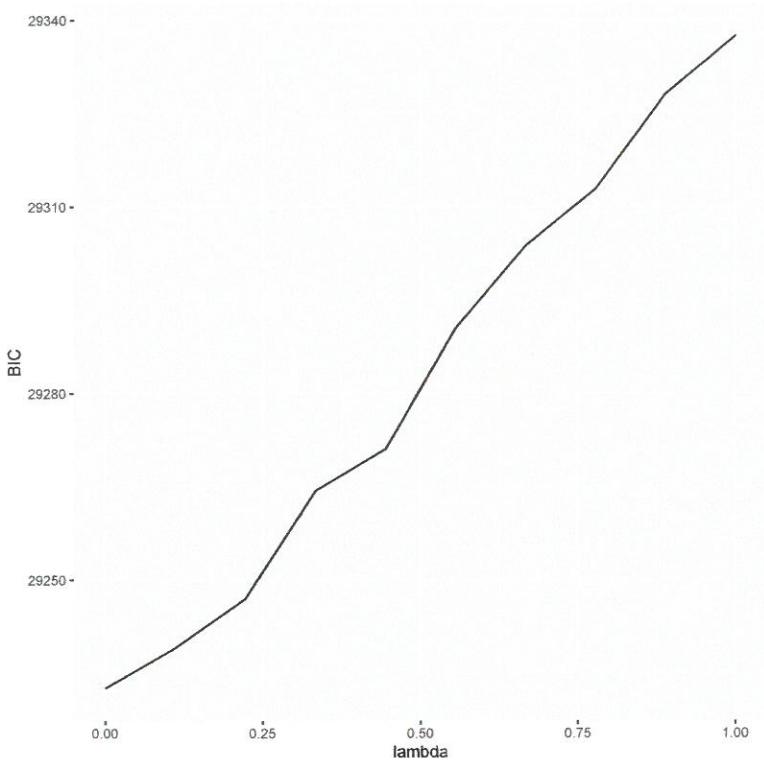


Figure 8 – BIC measure with respect to the λ penalization parameter.

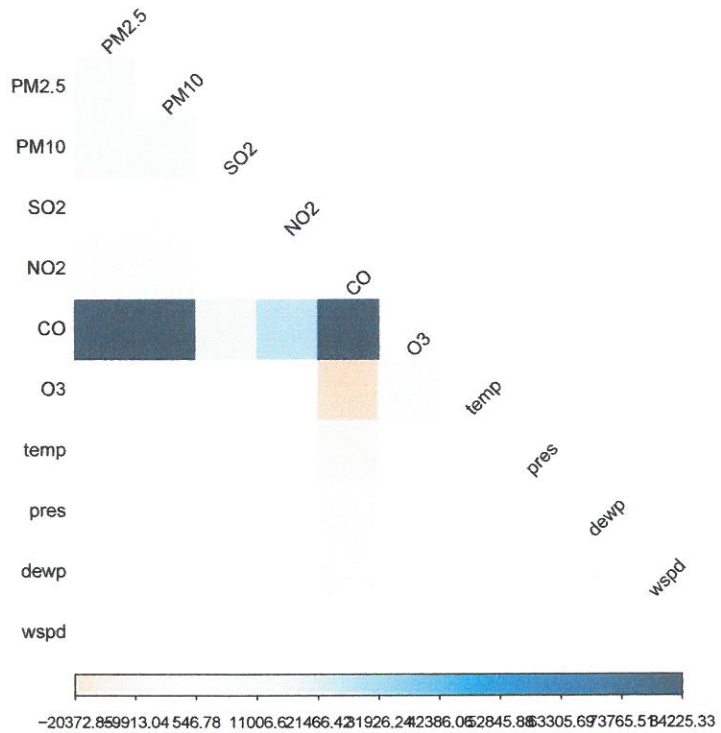


Figure 9 – Heat Map of the classic covariance matrix

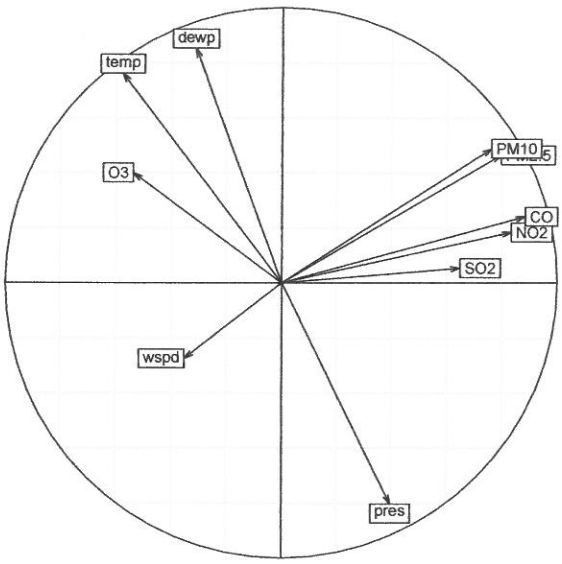
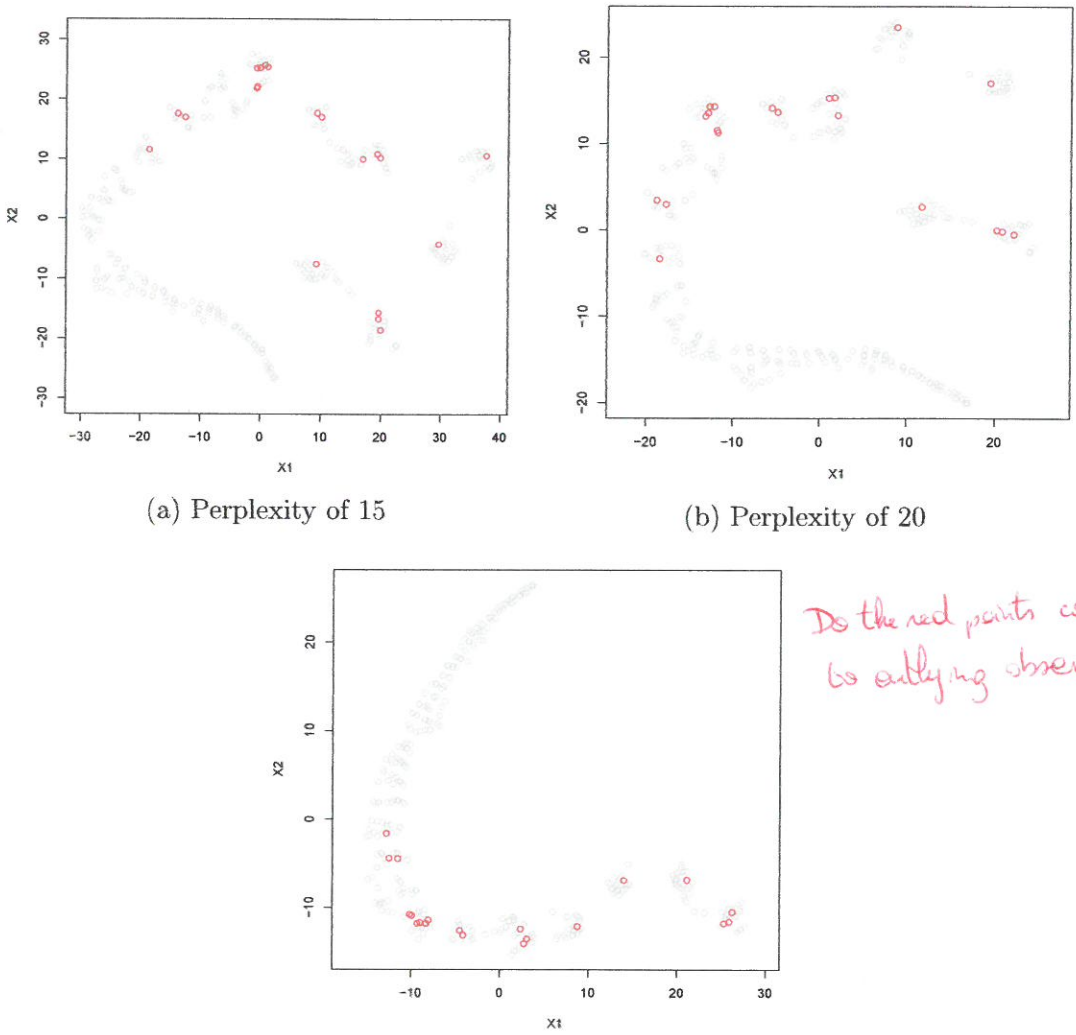


Figure 10 – Correlation circle for the first two principal components



(a) Perplexity of 15

(b) Perplexity of 20

(c) Perplexity of 30

Figure 11 – tSNE of the quantitative data

B Tables

PM2.5	PM10	SO_2	NO_2	CO	O_3	temp	pres	dewp	wspd	distance
199	292	113	115	6700	16	-4.3	1025.0	-11.3	2.4	279.3953
362	362	11	86	7000	8	-0.3	1020.8	-1.1	0.7	291.6911
74	379	3	50	300	64	26.3	1007.4	-3.3	3.1	316.7818
256	321	162	108	5300	2	2.7	1025.8	-2.8	0.5	460.6449
93	485	18	28	800	101	16.4	1017.3	3.4	1.2	560.9199
217	217	192	86	3800	4	2.1	1028.7	-4.0	0.9	792.7060

Table 2 – Values for the 6 observations corresponding to the biggest robust Mahalanobis distances.