Multivariate Statistical Analysis

Report for Multivariate Statistical Analysis - problem set 1

Bruno Casella
Department of Computer Science
University of Turin
casella@di.unito.it

April 5, 2022

Abstract

This report describes the results and the implementation of Problem Set 1 of Multivariate Statistical Analysis. All the R code is available at the following GitHub link: https://github.com/CasellaJr/MultivariateStatAnalysis. The exercises has been solved in collaboration with the PhD student Lorenzo Paletto.

Exercise 1

The **air pollution** data consists of 7 measurements recorded at n = 41 cities in the United States. Variables are

- *SO2*: Sulphur dioxide content in micrograms per cubic meter,
- *Neg. Temp*: Average annual temperature in Fo (negative values),
- *Manuf*: Number of manufacturing enterprises employing 20 or more workers,
- *Pop*: Population size (1970 census) in thousands,
- *Wind*: Average annual wind speed in miles per hour,
- *Precip*: Average annual precipitation in inches,
- *Days*: Average number of days with precipitation per year.

We ignore the SO2 variable and concentrate on the remaining 6, two of which relate to human ecology (Manuf. and Pop) and four to climate (Neg. Temp, Wind, Precip, Days).

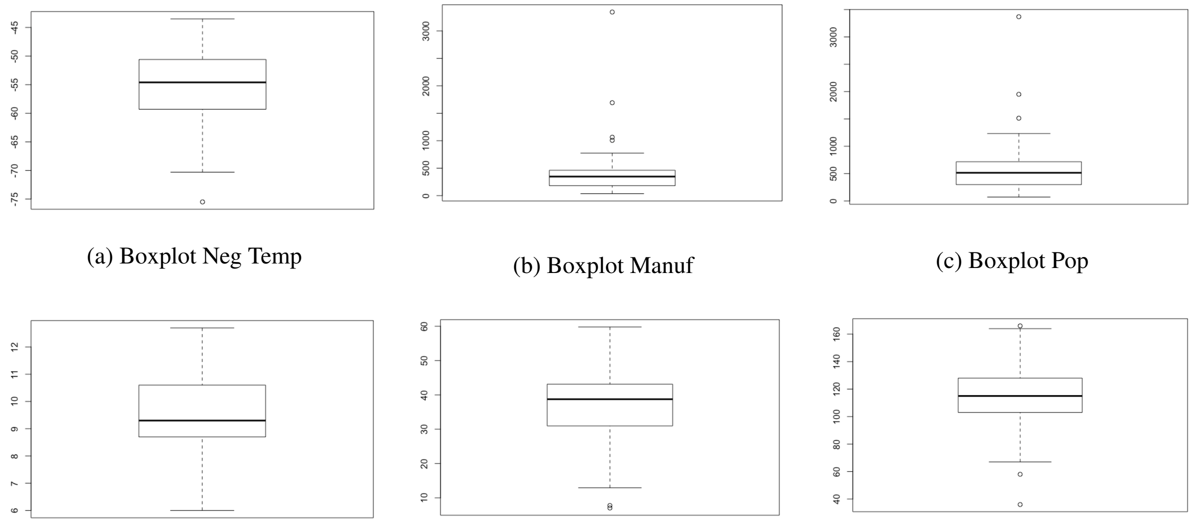
The **sample mean vector** is:

Neg. Temp	Manuf.	Pop.	Wind	Precip.	Days
-55.763415	463.097561	608.609756	9.443902	36.769024	113.902439

The **sample correlation matrix R** is: $\begin{pmatrix} 1.00000000 & 0.19004216 & 0.06267813 & 0.34973963 & -0.38625342 & 0.43024212 \\ 0.19004216 & 1.00000000 & 0.95526935 & 0.23794683 & -0.03241688 & 0.13182930 \\ 0.06267813 & 0.95526935 & 1.00000000 & 0.21264375 & -0.02611873 & 0.04208319 \\ 0.34973963 & 0.23794683 & 0.21264375 & 1.00000000 & -0.01299438 & 0.16410559 \\ -0.38625342 & -0.03241688 & -0.02611873 & -0.01299438 & 1.00000000 & 0.49609671 \\ 0.43024212 & 0.13182930 & 0.04208319 & 0.16410559 & 0.49609671 & 1.00000000 \end{pmatrix}$

Manuf. and Pop. are strongly positive correlated. Days and Neg. Temp have a moderate uphill positive relationship. Days and Precip. are moderately positive correlated. Neg. Temp and Precip. are negatively correlated.

These are the boxplots of each variable.



(a) Boxplot Neg Temp

(b) Boxplot Manuf

(c) Boxplot Pop

(d) Boxplot Wind

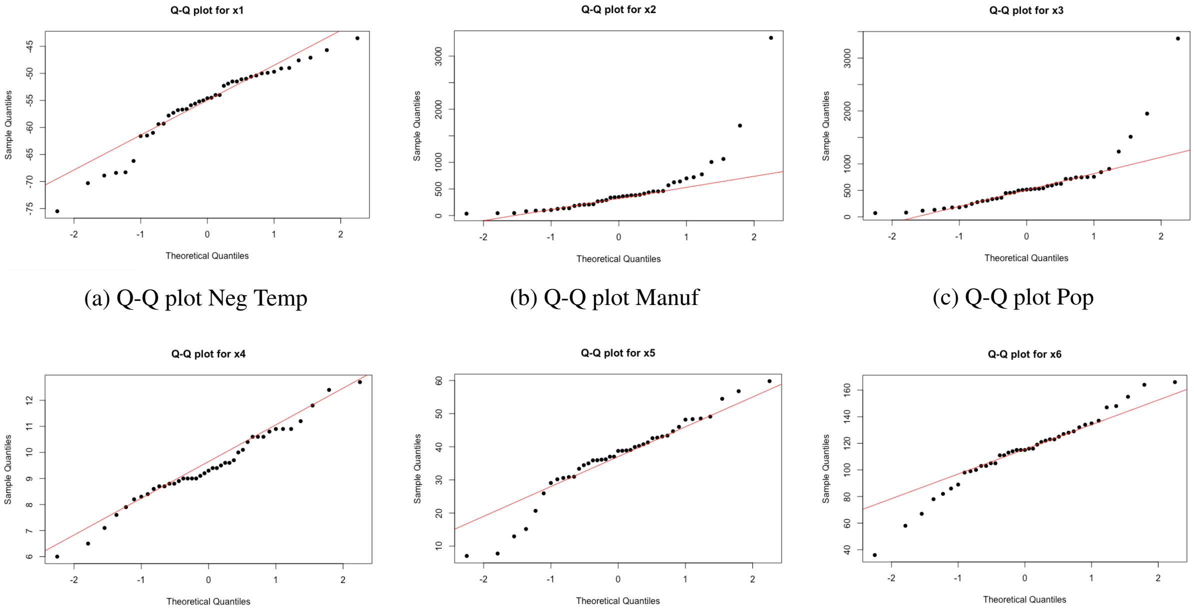
(e) Boxplot Precip

(f) Boxplot Days

Figure 1: Boxplots of all the 6 variables

Boxplot of Neg Temp shows only 1 outlier that corresponds to the 9-th observation. It is about the city of Miami.
Boxplot of Manuf shows 4 outliers. However, the text of the exercise asks for only 2 outliers. So, the most external outliers are the observations 11 (Chicago) and 29 (Philadelphia).
Boxplot of Pop shows 3 outliers. The most external are the same of the previous boxplot.
In boxplot of Wind there are no outliers.
Boxplot of Precip shows 2 outliers that correspond to observation 1 (Phoenix) and 23 (Alburquerque).
Boxplot of Days shows 3 outliers. The most external are the same of the previous boxplot.

Construct a normal Q-Q plot for each variable and comment about normality.



(a) Q-Q plot Neg Temp

(b) Q-Q plot Manuf

(c) Q-Q plot Pop

(d) Q-Q plot Wind

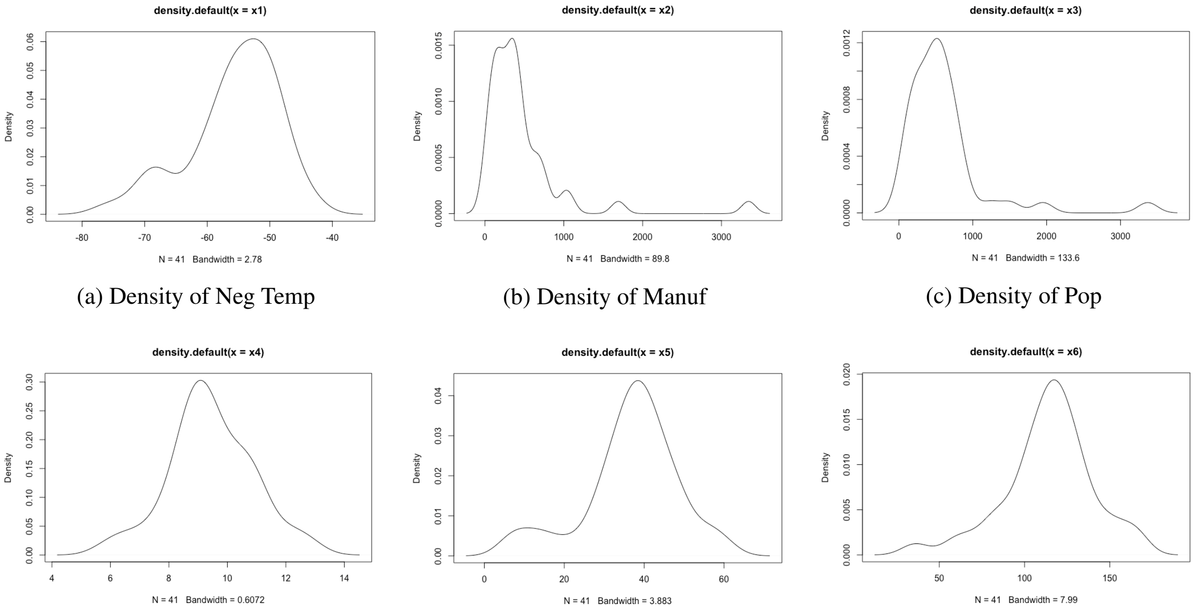
(e) Q-Q plot Precip

(f) Q-Q plot Days

Figure 2: Q-Q plots of all the 6 variables

Q-Q plot of Neg Temp shows a quite normal distribution.
Q-Q plot of Manuf shows a right skewed distribution.
Q-Q plot of Pop shows a right skewed distribution.
Q-Q plot of Wind shows a normal distribution.
Q-Q plot of Precip shows a little left skewed distribution.
Q-Q plot of Days shows a normal distribution.

These interpretations can be confirmed plotting the density or computing the skewness of each variable.



(a) Density of Neg Temp

(b) Density of Manuf

(c) Density of Pop

(d) Density of Wind

(e) Density of Precip

(f) Density of Days

Figure 3: Density of all the 6 variables

Values of skewness:

Neg. Temp	Manuf.	Pop.	Wind	Precip.	Days
-0.8540294	3.616089	3.052242	0.002776073	-0.7186492	-0.5708491

Scatterplots:

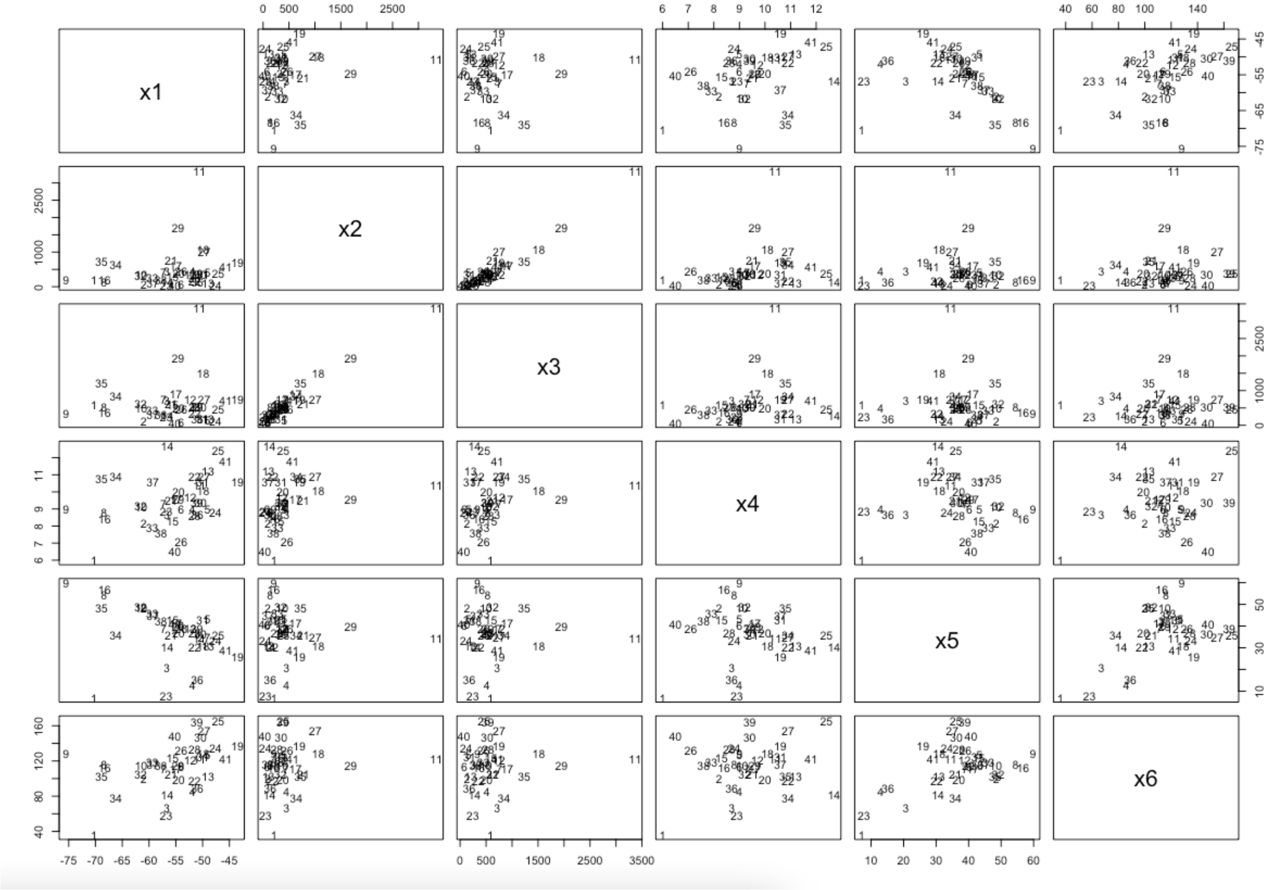


Figure 4: Scatterplots

We can see that the outliers are the observations 11 and 29, that are the same detected with the boxplots of Manuf and Pop.

Now I am going to plot the chi-squared Q-Q plot of the squared Mahalanobis distances.

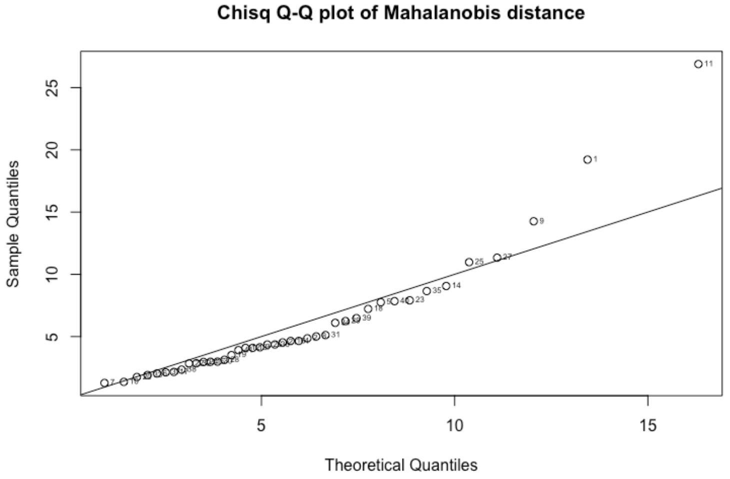


Figure 5: Chi-squared Q-Q plot of the squared Mahalanobis distances

We can see that observations 1, 9 and 11 are outliers. All these 3 observations have been previously detected using the boxplots.
This is the same plot after removing the outliers.

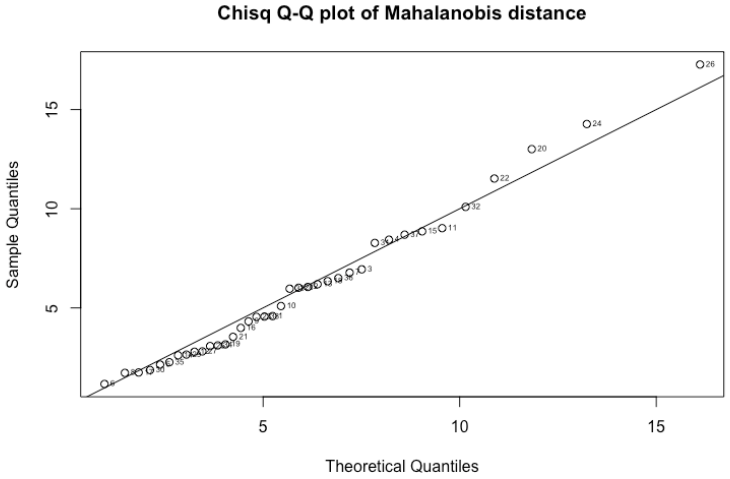


Figure 6: Chi-squared Q-Q plot of the squared Mahalanobis distances without outliers

The distribution without outliers is normal.

Exercise 2

For $X = (X_1, X_2, X_3)$ distributed as $N_3(\mu, \Sigma)$, $\mu = \begin{bmatrix} 1 \\ -1 \\ 2 \end{bmatrix}$, $\Sigma = \begin{bmatrix} 1 & \rho & 0 \\ \rho & 1 & \rho \\ 0 & \rho & 1 \end{bmatrix}$ with $|\rho| < \frac{\sqrt{2}}{2}$.

Let PC1 and PC2 be the first two (population) principal components of X. Find ρ such that they account for more than 80% of total variation of X.
The total variance of the original variables $X_1, ..., X_p$ is equal to the total variance of the PCs: $trace(\Sigma) = s_1^2 + ... + s_p^2 = \lambda_1 + ... + \lambda_p$ where Σ is the sample covariance matrix. The proportion of variance explained by the first k PCs is $\frac{\lambda_1 + ... + \lambda_k}{trace(\Sigma)}$, so in this case, $k = 2$, the formula becomes: $PVE = \frac{\lambda_1 + \lambda_2}{trace(\Sigma)}$.
So, I need to find λ_1 and λ_2. Let's calculate $det(A)$ where $A = \lambda I - \Sigma$. Basically $A = \begin{bmatrix} \lambda - 1 & -\rho & 0 \\ -\rho & \lambda - 1 & -\rho \\ 0 & -\rho & \lambda - 1 \end{bmatrix}$.
$det(A) = (\lambda - 1)[(\lambda - 1)(\lambda - 1) - (-\rho)(-\rho)] - (-\rho)[-\rho(\lambda - 1) - (-\rho \cdot 0)] =$
$= (\lambda - 1)[(\lambda - 1)^2 - \rho^2)] + \rho(\rho - \rho\lambda) =$
$= (\lambda - 1)^3 - \rho^2\lambda + \rho^2 + \rho^2 - \rho^2\lambda =$
$= (\lambda - 1)^3 - 2\rho^2\lambda + 2\rho^2 =$
$= (\lambda - 1)^3 - 2\rho^2(\lambda - 1) =$
$= (\lambda - 1)[(\lambda - 1)^2 - 2\rho^2]$ From the zero product property we have $lambda_2 = 1$ and $\lambda^2 - 2\lambda + 1 - 2\rho^2 = 0$. This can be solved with $\Delta = 4 - 4(1 - 2\rho^2) = 2^2(1 - 1 + 2\rho^2) = 2^2(2\rho^2)$. So, $\lambda_3 = \frac{2 - 2|\rho|\sqrt{2}}{2} = 1 - |\rho|\sqrt{2}$. The same for $\lambda_1 = 1 + |\rho|\sqrt{2}$. Note that λ_1, λ_2 and λ_3 are in increasing order because they are to the first PCs.
Remembering that the trace is $trace(\sigma) = \lambda_1 + \lambda_2 + \lambda_3 = 1 + |\rho|\sqrt{2} + 1 + 1 - |\rho|\sqrt{2} = 3$, then we have that $PVE = \frac{1 + |\rho|\sqrt{2} + 1}{3} = \frac{2 + |\rho|\sqrt{2}}{3}$. We want that this value is bigger than 80%, so $\frac{2 + |\rho|\sqrt{2}}{3} > 0.8; |\rho|\sqrt{2} > 0.4 = \frac{2}{5}; |\rho| > \frac{\sqrt{2}}{5}$.
So, $\rho > \frac{\sqrt{2}}{5}$ is the value needed for getting more than 80% of total variation.

Give an interpretation to PC1 and PC2 in terms of the original variables.

I need to calculate this equation in x: $Ax = 0$ so $\begin{bmatrix} \lambda - 1 & -\rho & 0 \\ -\rho & \lambda - 1 & -\rho \\ 0 & -\rho & \lambda - 1 \end{bmatrix} X = 0$.

We have 3 $lambda$, so let's start from λ_1: $\begin{bmatrix} |\rho|\sqrt{2} & -\rho & 0 \\ -\rho & |\rho|\sqrt{2} & -\rho \\ 0 & -\rho & |\rho|\sqrt{2} \end{bmatrix} \begin{bmatrix} \\ e_1 \\ \\ \end{bmatrix} = 0$.

Solution is:

$$\begin{cases} |\rho|\sqrt{2}x - \rho y + 0z = 0 \\ -\rho x + |\rho|\sqrt{2}y - \rho z = 0 \\ 0x - \rho y + |\rho|\sqrt{2}z = 0 \end{cases} . \tag{1}$$

Looking at the first and the third equations of this system, we can note that $x = z$, so substituting in the second equation, we get: $-\rho x + |\rho|\sqrt{2}y - \rho x = 0$ -> $2\rho x = |\rho|\sqrt{2}y$ which means $y = \frac{2\rho x}{|\rho|\sqrt{2}}$. So, we can get $\begin{bmatrix} \\ e_1 \\ \\ \end{bmatrix} = \begin{bmatrix} 1 \\ \frac{\sqrt{2}\rho}{|\rho|} \\ 1 \end{bmatrix}$

Now with $\lambda_2 = 1$ we have that: $\begin{bmatrix} 0 & -\rho & 0 \\ -\rho & 0 & -\rho \\ 0 & -\rho & 0 \end{bmatrix} \begin{bmatrix} \\ e_2 \\ \\ \end{bmatrix} = 0$. It's easy to check that $y = 0$ and $x = -z$. So, $\begin{bmatrix} \\ e_2 \\ \\ \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix}$

For e_1, if $\rho < 0$ then the second component is the only negative, going in the opposite direction.
Instead, e_2 loads only the first and the third component.

Find the distribution of $Z = \begin{bmatrix} Z_1 \\ z_2 \end{bmatrix} = \begin{bmatrix} X_1 - X_2 \\ X_2 - X_3 \end{bmatrix}$ I am going to find mean vector μ_Z and the covariance matrix Σ_Z of Z.

$\mu_Z = \begin{bmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \end{bmatrix} \mu = \begin{bmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \end{bmatrix} \begin{bmatrix} 1 \\ -1 \\ 2 \end{bmatrix} = \begin{bmatrix} 2 \\ -3 \end{bmatrix}$

$\Sigma_Z = \begin{bmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \end{bmatrix} \Sigma \begin{bmatrix} 1 & 0 \\ -1 & 1 \\ 0 & -1 \end{bmatrix} = \begin{bmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \end{bmatrix} \begin{bmatrix} 1 & \rho & 0 \\ \rho & 1 & \rho \\ 0 & \rho & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ -1 & 1 \\ 0 & -1 \end{bmatrix} = \begin{bmatrix} 1 - \rho & \rho - 1 & -\rho \\ \rho & 1 - \rho & \rho - 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ -1 & 1 \\ 0 & -1 \end{bmatrix} = \begin{bmatrix} 2 - 2\rho & 2\rho - 1 \\ 2\rho - 1 & 2 - 2\rho \end{bmatrix}$

So, finally $Z \sim N(\mu_Z, \Sigma_Z)$.

Let $\rho = -\frac{2}{3}$ and Σ_Z and μ_Z be the corresponding covariance matrix and the mean vector of $Z = (Z_1, Z_2)$. Sketch the ellipse $(z - \mu_Z)^T \Sigma_Z^{-1} (z - \mu_Z)^T = c^2$ in the 2-dimensional space $z = (z_1, z_2)$ by setting the constant "c" such that the ellipse contains 0.95 probability wrt the joint distribution of Z.

So, with $\rho = -\frac{2}{3}$ then $\Sigma_Z = \begin{bmatrix} \frac{10}{3} & -\frac{7}{3} \\ -\frac{7}{3} & \frac{10}{3} \end{bmatrix}$ while μ_Z does not change.

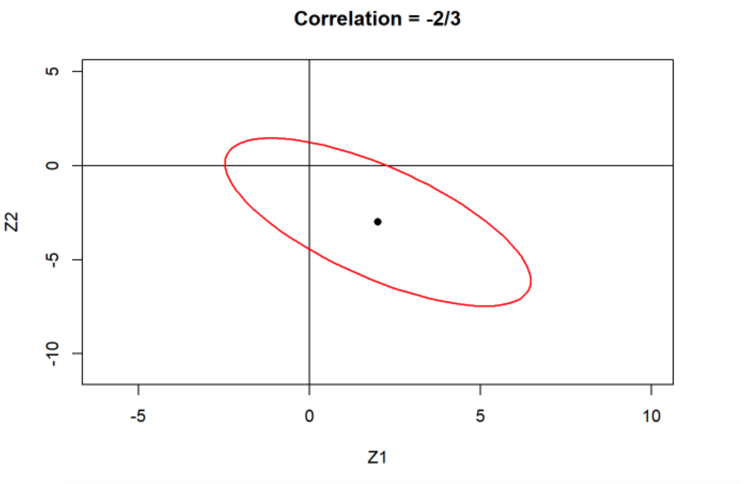


Figure 7: Ellipse on R

Instead, if $\rho = \frac{2}{3}$ then Z_1 and Z_2 will be positively correlated, so the orientation will be the opposite. The ellipse will be also smaller because the values are smaller: $\Sigma_Z = \begin{bmatrix} \frac{2}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{2}{3} \end{bmatrix}$

Exercise 3

The pen digit data set was created by collecting 250 samples from 44 writers. These writers were asked to write 250 digits in random order inside boxes of 500 by 500 tablet pixel resolution. The raw data on each of n = 10992 handwritten digits consisted of a sequence, (x_t, y_t), t = 1, 2, ..., T , of tablet coordinates of the pen at fixed time intervals of 100 milliseconds, where x_t and y_t were integers in the range 0-500. These data were then normalized to make the representations invariant to translation and scale distortions. The new coordinates were such that the coordinate that had the maximum range varied between 0 and 100. Usually x_t stays in this range, because most integers are taller than they are wide. Finally, from the normalized trajectory of each handwritten digit, 8 regularly spaced measurements, (x_t, y_t), were chosen by spatial resampling, which gave a total of p = 16 variables. The data includes a class attribute, column digit, coded 0, 1, . . . , 9, about the actual digit.

Perform a principal component analysis on the standardized variables. Report standard deviations. Decide how many components to retain in order to achieve a satisfactory lower-dimensional representation of the data.

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9
Standard deviation	2.1718	1.7970	1.6052	1.10894	1.03107	0.89294	0.77888	0.74044	0.64096
Proportion of Variance	0.2948	0.2018	0.1610	0.07686	0.06644	0.04983	0.03792	0.03427	0.02568
Cumulative proportion	0.2948	0.4966	0.6577	0.73452	0.80096	0.85079	0.88871	0.92297	0.94865

	PC10	PC11	PC12	PC13	PC14	PC15	PC16
Standard deviation	0.54611	0.45882	0.33511	0.28369	0.24084	0.18511	0.16673
Proportion of Variance	0.01864	0.01316	0.00702	0.00503	0.00363	0.00214	0.00174
Cumulative proportion	0.96729	0.98045	0.98747	0.99250	0.99612	0.99826	1.00000

From the Cumulative proportion we can see that the first 2 components present about 50% of the variance of data. The first 3 present about 65.8% of variance. The first 4 present about 73% of variance. The first 5 present about 80.1% From the 6-th PC we get only a little increase of variance (about +5%). So I would retain only the first 5 PCs in order to achieve a satisfactory lower-dimensional representation of the data.
Another useful tool to decide the number of PCs to retain is the elbow method, through the scree plot:

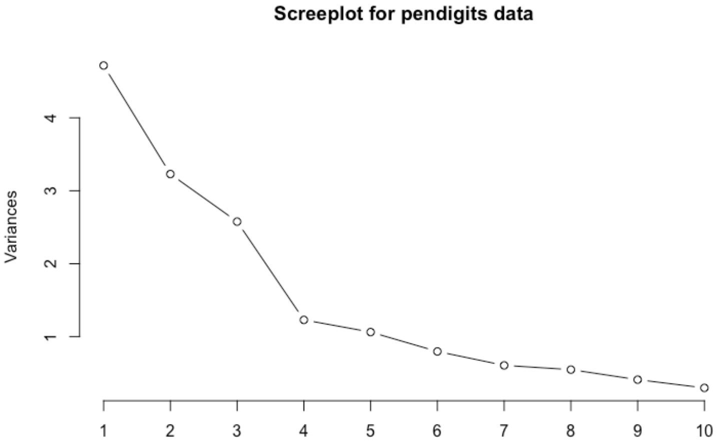


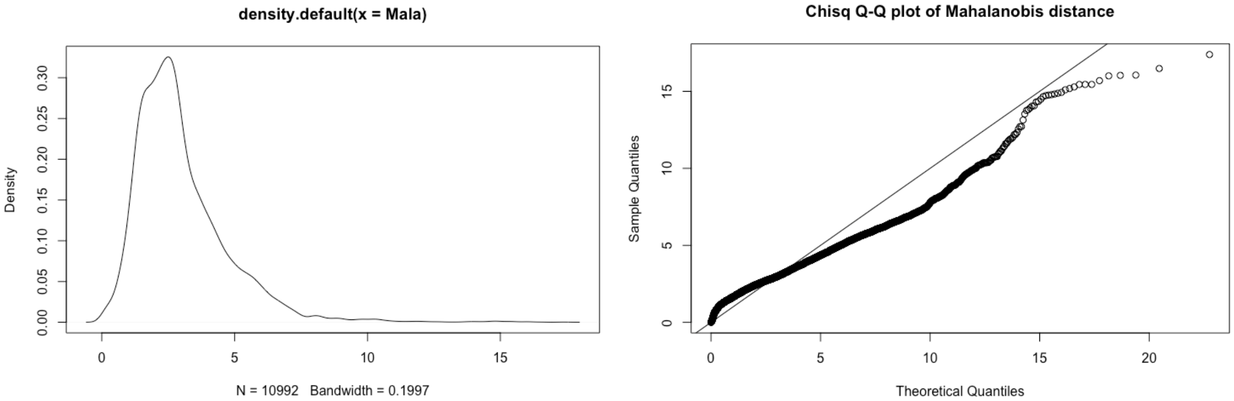
Figure 8: Screeplot

The screeplot shows an evident elbow corresponding to the fourth variable. However I decide to retain 5 PCs because I want to explain at least the 80% of the variability; moreover, another parameter to investigate in order to choose how many PCs to retain is the variance: I use the first 5 PCs because they have variance greater than 1.

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
Variance	4.71658637	3.22911210	2.57680630	1.22974181	1.06309928	0.79733935	0.60665163	0.54825156

	PC9	PC10	PC11	PC12	PC13	PC14	PC15	PC16
Variance	0.41082409	0.29823086	0.21051266	0.11229691	0.08048105	0.05800212	0.03426450	0.02779941

Investigate multivariate normality through the first three principal components.

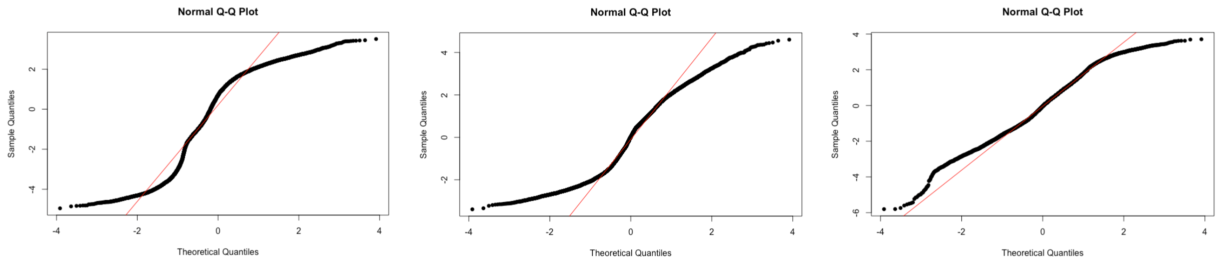


(a) Distribution of the Mahalanobis distances

(b) Chi-Squared Q-Q plot of Mahalanobis distance

Figure 9: Multivariate Normality

It is visible that the distribution is not normal looking at the distribution of the Mahalanobis distances. Moreover, also from the Chi-squared Q-Q plot we can see the tails of the distribution, confirming that it is not normal.
We can look also at the Q-Q plot of the first 3 principal components:



(a) Q-Q plot of 1st PC (b) Q-Q plot of 2nd PC (c) Q-Q plot of 3rd PC

Figure 10: Q-Q plot of the first 3 PCs

It seems that the distributions are heavy-tailed.

Make scatter plots with the first three principal components, while color coding the observations according to the digit class.

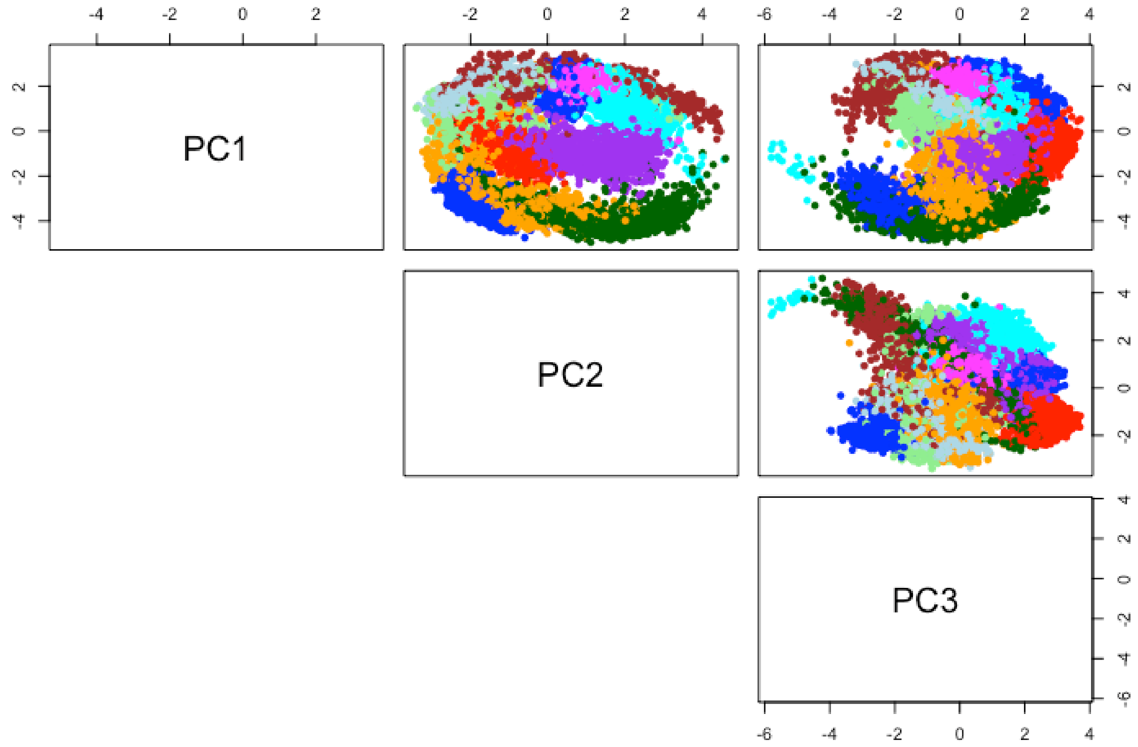
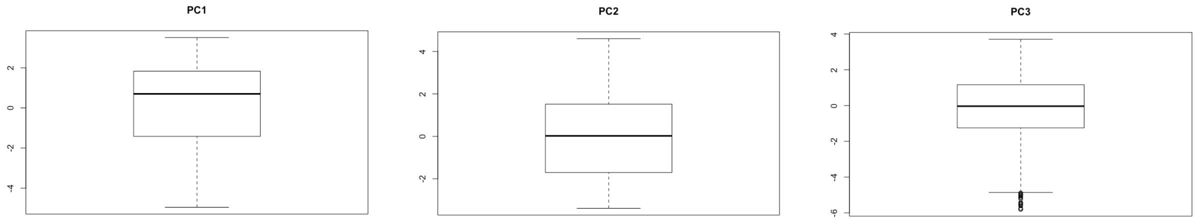


Figure 11: Scatterplot with the first 3 PCs

The first 3 PCs are able to divide the in clusters, according to the different digits.
From these scatterplots we can also note the presence of possible outliers in correspondence of the third PC. In particular I am referring to the light blue datapoints that are apart from the cluster of light blue points.

Comment about outliers with respect to the first three principal components.
I am going to use the boxplots for this point.



(a) Boxplot of 1st PC

(b) Boxplot of 2nd PC

(c) Boxplot of 3rd PC

Figure 12: Boxplots of the first 3 PCs

The first and the second plot do not show outliers, while the third boxplot shows several outliers (the ones mentioned before). These outliers are 26 in total: 2 corresponding to digit 0, 24 corresponding to digit 9. However, considering that in the dataset there are 1143 observations of digit 0 and 1055 of digit 9, they are only a small portion.