Basilico Eleonora, Caporali Francesco, Malgieri Luigi Maria
30 March 2023

Multivariate Statistical Analysis
Problem Set 1

Exercise 2

2.1

Let first note that Σ is invertible since $\det(\Sigma) = -2\rho^3 - 3\rho^2 + 1$ which is greater than 0, $\forall p \in \left[-1, \frac{1}{2}\right]$. We compute the inverse of Σ by exploiting the identity

$$\Sigma = (1+\rho)I - \rho aa^T \text{ with } a = (1, 1, -1)$$

and applying the following theorem, known as the Neumann Series Theorem:

Theorem (Neumann Series)**.** *Let T be a linear mapping $T : \mathbb{R}^n \to \mathbb{R}^n$. If the series $\sum_{i=0}^{\infty} T^i$ converges, then $I - T$ is invertible and it holds*

$$(I - T)^{-1} = \sum_{i=0}^{\infty} T^i.$$

First we rewrite

$$\Sigma = (1+\rho)(I - caa^T) \text{ with } a = (1, 1, -1) \text{ and } c = \frac{\rho}{1+\rho}.$$

Let $A \overset{\text{def}}{=} I - caa^T$, it holds

$$\begin{aligned}
A^{-1} &= (I - caa^T)^{-1} = \sum_{i=0}^{\infty} (caa^T)^i = \\
&= \sum_{i=0}^{\infty} c^i (aa^T)^i = I + \sum_{i=i}^{\infty} c^i (\|a\|^2)^{i-1} aa^T = \\
&= I + caa^T \sum_{i=0}^{\infty} (c\|a\|^2)^{i-1} = I + caa^T \sum_{i=i}^{\infty} (c\|a\|^2)^i = \\
&= I + caa^T \frac{1}{1 - c\|a\|^2} = I + \frac{p}{1+p} \frac{1}{1 - 3\frac{p}{1+p}} aa^T = \\
&= I + \frac{p}{1-2p} aa^T.
\end{aligned}$$

Thus

$$\Sigma^{-1} = (1+\rho)^{-1} A^{-1} = \frac{1}{1+\rho}(I + \frac{p}{1-2p} aa^T).$$

We can compute Σ^{-1} also in many other ways, for example we can suppose that Σ^{-1} is of the same form Σ, i.e.

$$\Sigma^{-1} = yI + kaa^t$$

and than find the values for y and k.

2.2

A faster way to find the spectrum (set of eigenvalues, meant with multiplicity) is reported below. We exploit some basic properties of the spectrum.
We denote $\operatorname{Sp}(\Sigma)$ the spectrum of the matrix Σ (as a linear operator).

$$\begin{aligned}\operatorname{Sp}(\Sigma) = \operatorname{Sp}\left((1+\rho)(I - \frac{\rho}{1+\rho}aa^t)\right) &= \\ = (1+\rho)\operatorname{Sp}\left(I - \frac{\rho}{1+\rho}aa^t\right) &= \\ = (1+\rho)\left(1 - \frac{\rho}{1+\rho}\operatorname{Sp}\left(aa^t\right)\right)&.\end{aligned}$$

Observing $(aa^T)a = \|a\|^2 a$ and $\operatorname{rank}\left(aa^T\right) = 1$ it holds

$$\operatorname{Sp}\left(aa^T\right) = \left\{0, 0, \|a\|^2\right\}.$$

Hence

$$\begin{aligned}\operatorname{Sp}(\Sigma) = (1+\rho)\left(1 - \frac{\rho}{1+\rho}\{0,0,\|a\|^2\}\right) &= \\ = (1+\rho)\left\{1, 1, 1 - 3\frac{\rho}{1+\rho}\right\} &= \\ = \{1+\rho, 1+\rho, 1+\rho-3\rho\} &= \\ = \{1+\rho, 1+\rho, 1-2\rho\}&.\end{aligned}$$

where the multiplications and translations of sets are mean component wise.

2.3

We first write the eigenvalues of Σ in ascending order.
We distinguish the following two cases:

1. if $\rho \in \left[0, \frac{1}{2}\right)$ then $1+\rho \geq 1-2\rho$. This leads to

$$\begin{cases}\lambda_1 = 1+\rho \\ \lambda_2 = 1+\rho \\ \lambda_3 = 1-2\rho\end{cases}, \text{ with } \lambda_1 \geq \lambda_2 \geq \lambda_3.$$

Now we find ρ such that the first two principal components (PCs) account for more than 80% of the total variation of Z.
Since λ_i corresponds to the variance of the i-th PC $\forall i \in \{1,2,3\}$ and the variation up to the k-th PC corresponds to the sum of the first k eigenvalues, we just need to find ρ such that

$$\lambda_1 + \lambda_2 > 0.8(\lambda_1 + \lambda_2 + \lambda_3).$$

By solving the inequality we get

$$2(1+\rho) > \frac{4}{5}3 \iff 1+\rho > \frac{6}{5} \iff \rho > \frac{1}{5}.$$

2. if $\rho \in (-1, 0)$ then $1+\rho \leq 1-2\rho$. This leads to

$$\begin{cases}\lambda_1 = 1-2\rho \\ \lambda_2 = 1+\rho \\ \lambda_3 = 1+\rho\end{cases}, \text{ with } \lambda_1 \geq \lambda_2 \geq \lambda_3.$$

By using the same argument we used in the previous poin we obtain that ρ have to satisfy the following condition:

$$(1-2\rho)+(1+\rho) > \frac{4}{5}3 \iff 2-\rho > \frac{12}{5} \iff \rho < -\frac{2}{5}.$$

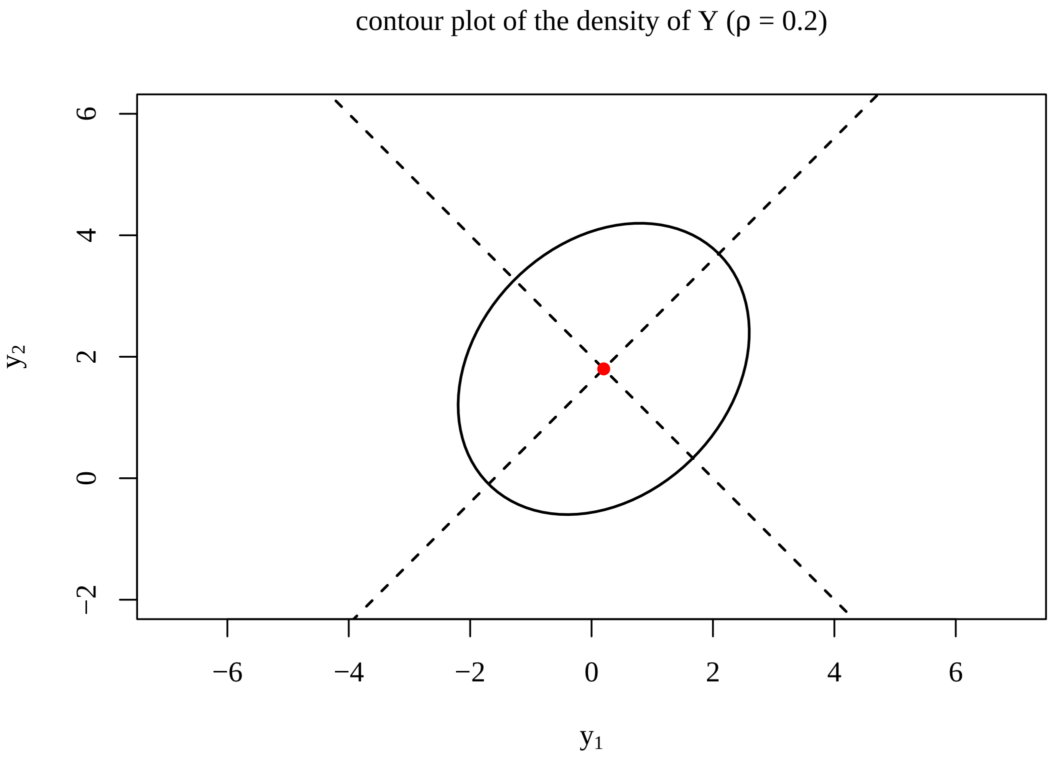
Hence for $\rho \in \left[0, \frac{1}{2}\right)$ it must be $\rho > \frac{1}{2}$ and for $\rho \in (-1, 0)$ it must be $\rho < -\frac{2}{5}$.
So $\forall \rho \in \left(-1, -\frac{2}{5}\right) \cup \left(\frac{1}{2}, 1\right)$ PC1 and PC2 account for more than 80% of the total variation of Z.

2.4

2.5

```
c = sqrt(qchisq(0.95, df = 2))
```

```
rho = 0.2
mu_y = 1 / 5 * c(1, 9)
sigma_y = 6 / 25 * matrix(c(4, 1, 1, 4), nrow = 2)
eig = eigen(sigma_y, symmetric = T)
```



Exercise 3

3.1

First of all we divide each variable by `weight` in order to equalize out the fifferent types of servings of each food.

```
nutritional = read.table("data/nutritional.txt")
nt = nutritional[, -6] / nutritional[, 6]
head(round(nt, 3))
```

fat	food.energy	carbohydrates	protein	cholesterol	saturated.fat
0.133	1.667	0.133	0.000	0.133	0.013
0.375	3.750	0.125	0.000	0.250	0.062
0.035	3.175	0.776	0.141	0.000	0.004
0.000	3.175	0.776	0.106	0.000	0.004
0.000	0.303	0.030	0.030	0.000	0.000
0.035	2.469	0.741	0.141	0.000	0.004

After the standardization of the dataset, carried out with the command `scale` we perform the Principal Components Analysis.

```
nt = scale(nt)
nt_pca = prcomp(nt)
as.data.frame(nt_pca$rotation)
```

	PC1	PC2	PC3	PC4	PC5	PC6
fat	-0.5572394	0.0987008	-0.2750890	0.1304014	-0.4546980	0.6166958
food.energy	-0.5361507	0.3567665	0.1370762	0.0745468	-0.2729547	-0.6974301
carbohydrates	0.0245536	0.6716316	0.5684779	-0.2861681	0.1568663	0.3444441
protein	-0.2352271	-0.3738430	0.6388770	0.5991035	0.1538186	0.1189985
cholesterol	-0.2525045	-0.5213044	0.3256120	-0.7170962	-0.2102965	-0.0029044
saturated.fat	-0.5313507	-0.0192336	-0.2611169	-0.1496468	0.7913619	0.0216043

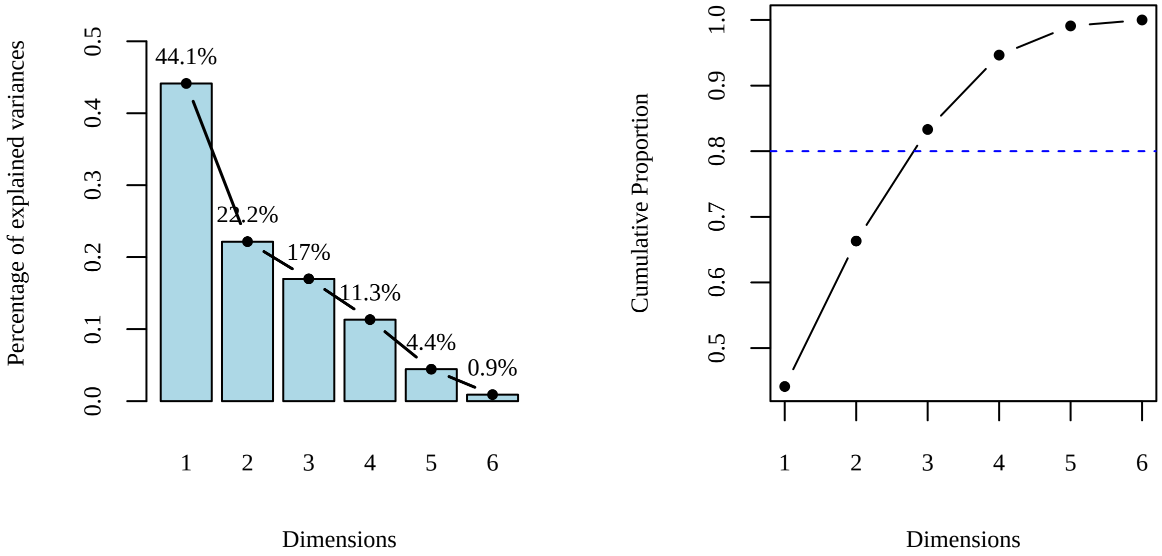
3.2

In order to decide how many components to retain we first observe the proportions and the cumulative proportions of explained variances.

```
nt_sum = as.data.frame(summary(nt_pca)$importance)[-1, ]
nt_sum
```

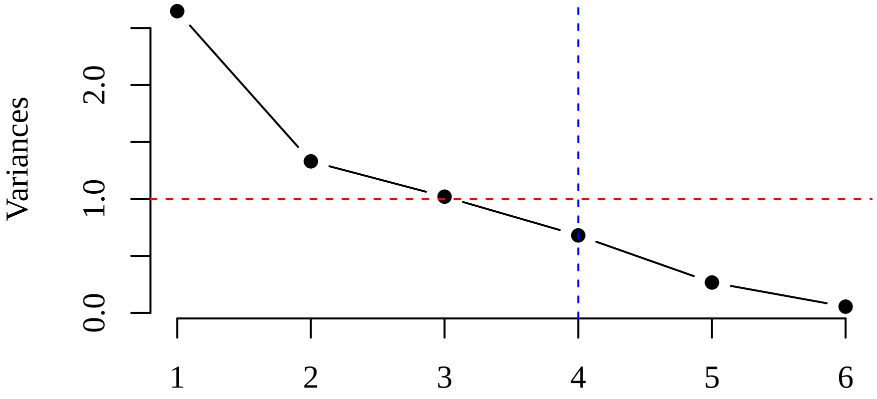
	PC1	PC2	PC3	PC4	PC5	PC6
Proportion of Variance	0.44143	0.22169	0.17002	0.11334	0.04442	0.00909
Cumulative Proportion	0.44143	0.66312	0.83314	0.94649	0.99091	1.00000

For a more immediate visualization we can plot the values reported above.



We can note that the first three PCs explain the 83.3% of the variability of the data, and if we add a fourth PC we arrive near to the 95% which is way too much for our purpose. Hence it seems reasonable to retain only the first three PCs.

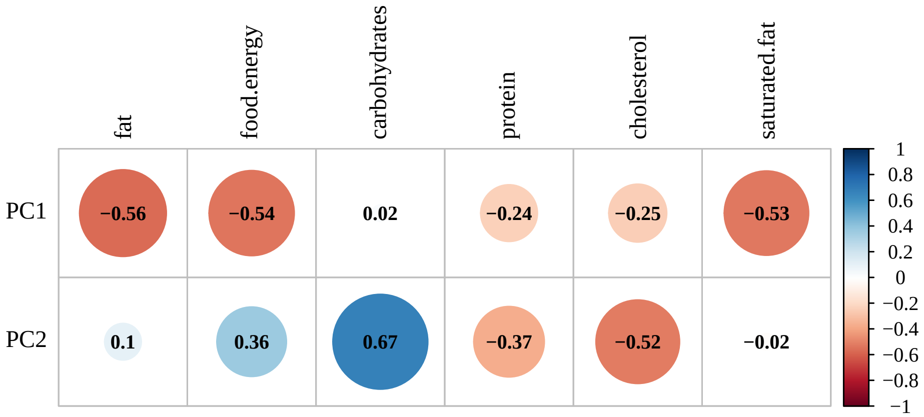
In order to confirm the previous consideration it also useful to look at the eigenvalues associated to each component. We also recall that if the variables are standardized the components whose eigenvalues are greater than 1 capture most of the original variables' variance.



The screeplot is not so easy to interpret, due to the absence of an evident *elbow* in the curve. However only the eigenvalues corresponding to the first three PCs are greater than 1 hence we can keep the choice of retaining only the first three PCs.

3.3

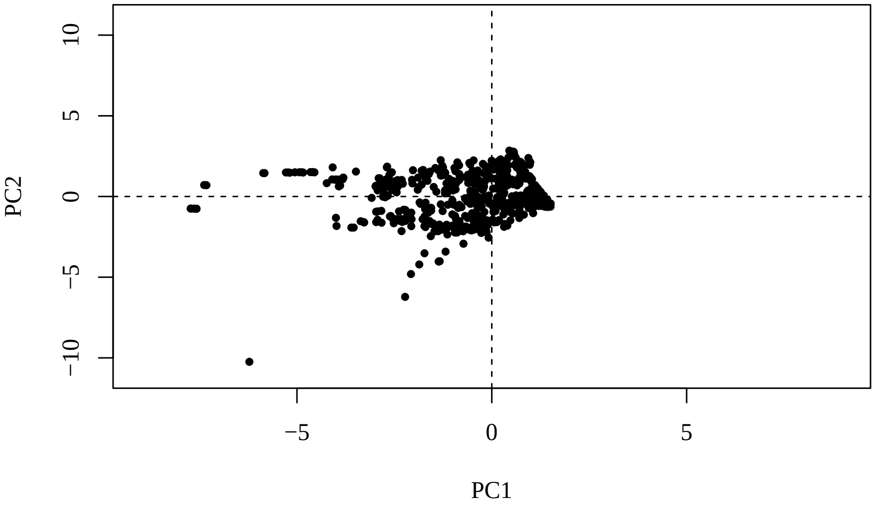
In order to give an interpretation to the first two PCs we look at their loadings (we report them in the following representation).



PC1. Except for the loading associated to the variable *carbohydrates* which is almost negligible, the other loadings are all negative and lies in $[-0.24, -1]$. The most influencial variables are *fat*, *satured.fat* and *food.energy* which are related to how much a food is dietetic (or not). Since those loadings are negative, we choose to interpret the first PC of a food item as a measure of its dietary.

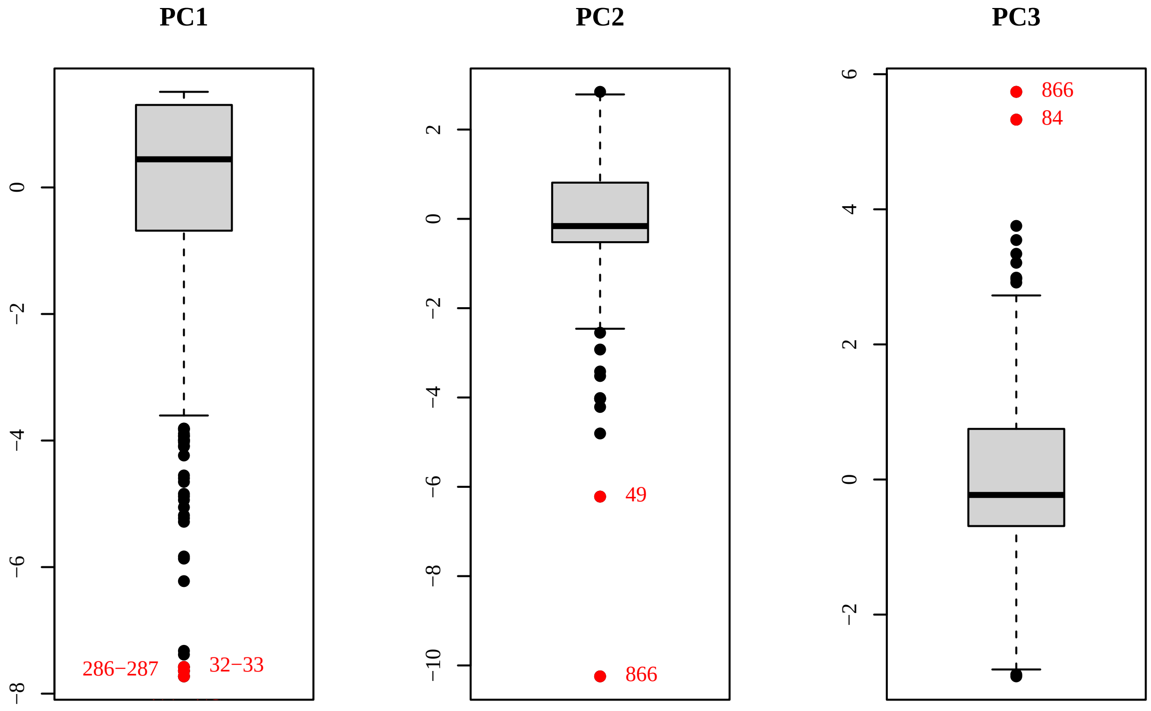
PC2. The loadings of the second PC are more complicated to analyse. The most influencing in a positive sense are the ones corresponding to the variables *carbohydrates* and *food.energy*, while the negative ones are *protein* and *colhesterol*. Hence PC2 identifies how intense in carbohydrates the serving was, and at the same time low in cholesterol. A possible interpretation could be that the component is an indicator of how intensely a food is made of cereals.

Here we report a scatterplot of our dataset projected on the plane PC1 *vs.* PC2.



3.4

In order to identify univariate outliers, we display the boxplots with respect to the first three principal components.



In the figure above we can spot a consistent amount of outliers for each principal component. We identified the most extreme (which are displayed in red) by scaling our PCs and then comparing them with some quantiles (with levels over 0.9999) of a $\mathcal{N}(0,1)$. It is remarkable that some outliers have multiplicity equal to 2: note that in the first boxplot the outliers are actually six.

We now want measure how these outliers score in the original variables. We build a dataframe with *min*, *mean* and *max* of each variable.

	fat	food.energy	carbohydrates	protein	cholesterol	saturated.fat
min	-0.585	-1.164	-0.954	-0.778	-0.379	-0.562
mean	0.000	0.000	0.000	0.000	0.000	0.000
max	4.583	3.498	3.053	8.753	18.170	7.107

We now print the original dataset restricted to the outliers.

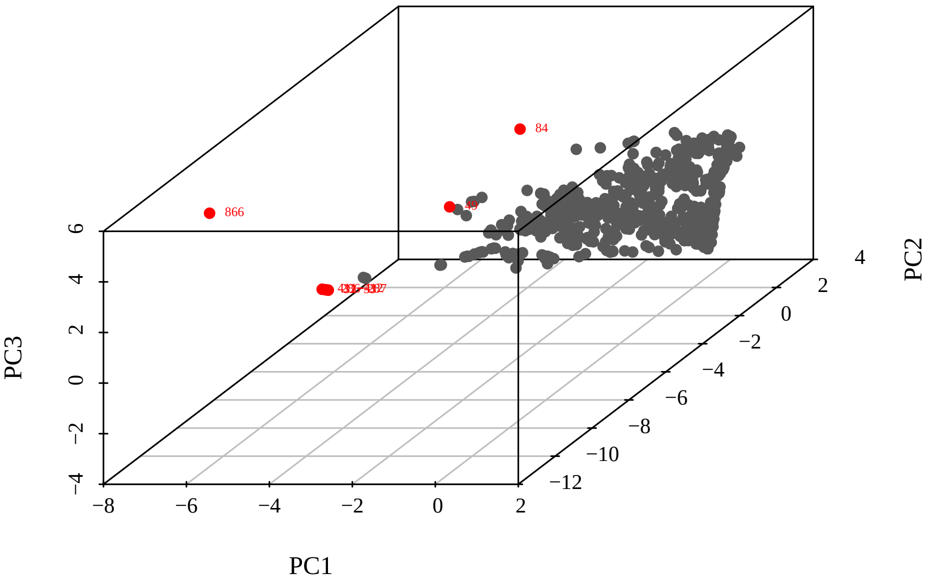
	fat	food.energy	carbohydrates	protein	cholesterol	saturated.fat
32	3.550	2.452	-0.954	-0.778	2.878	6.999
33	3.550	2.452	-0.954	-0.778	2.878	6.999
286	3.476	2.526	-0.954	-0.778	2.899	7.107
287	3.476	2.526	-0.954	-0.778	2.899	7.107
411	3.623	2.539	-0.954	-0.679	2.857	7.079
412	3.623	2.539	-0.954	-0.679	2.857	7.079
49	-0.327	-0.389	-0.954	2.002	8.948	-0.260
866	0.935	0.659	-0.954	1.184	18.170	0.861
84	-0.585	0.681	-0.954	8.753	-0.379	-0.562

We can observe that:

- as for the observations $32-33$, $286-287$, $411-412$, they score very low on *carbohydrates* and *protein*. However, given how low these two variables load on the first PC, this would not be enough to explain their outlier behaviour. The latter is instead due to their high values in *food.energy*, *fat* and *saturated.fat*. If we were to pick only two, they would be these last two;
- as for the observations 49 and 866, they score extremely high in *cholesterol* and attain the minimum in *carbohydrates*;
- finally, the observation 84 reaches the extremes in almost all variables but *food.energy*. If we were to choose only two of them, any random choice would be good.

3.5

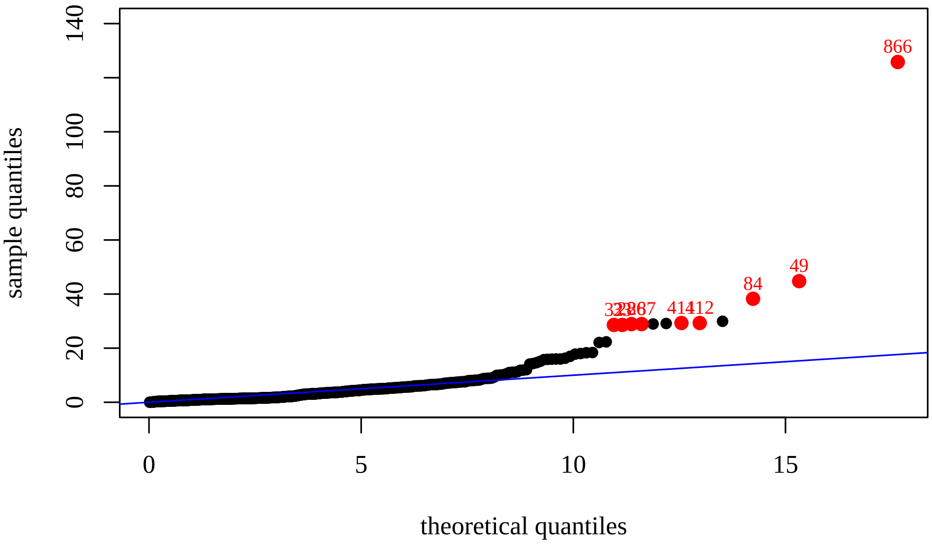
In the figure below we report a 3D scatterplot of the first three principal components.



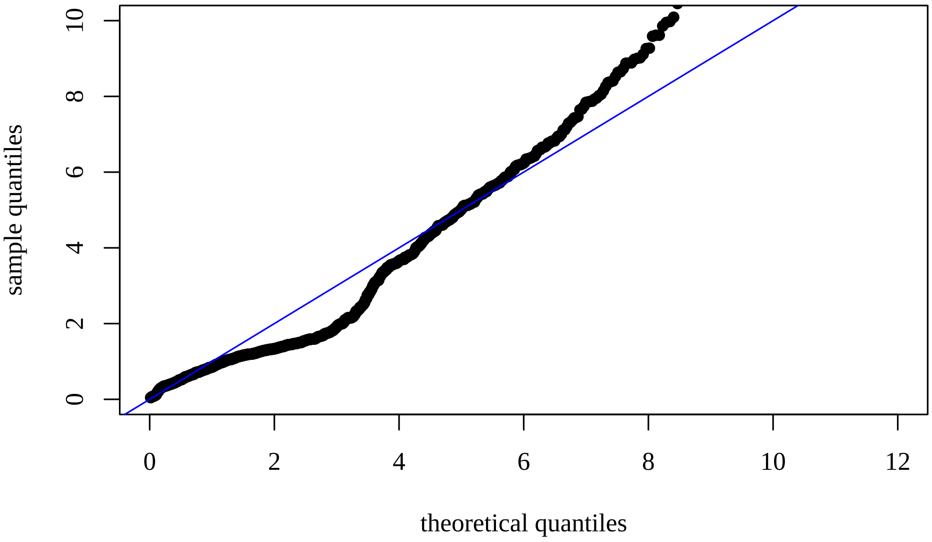
The outliers found in the previous point are labeled and highlighted in red. The plot shows that the joint distribution of these principal components has high variance. Nevertheless, the red points seem to be significantly far from the cloud. We hypotize that they could be multivariate outliers, as it will be discussed in more detail in the following.

3.6

We investigate multivariate normality through the first three principal components by computing the Mahalanobis distances, and comparing them with the quantiles of a χ^2_3 with a Q-Q plot.



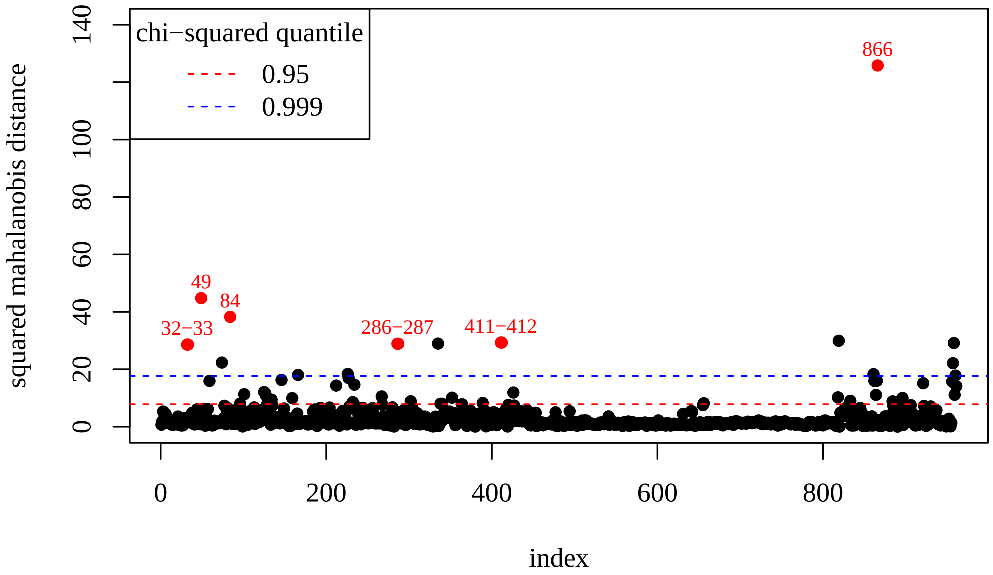
The approximation is rather good for the observations that have a low score in the mahalanobis sense. These points also account for a large part of the overall mass. However, the empirical distribution appears to have a heavy right tail. Hence, the normality assumption is not plausible.
To corroborate this thesis, we display the same plot after removing the outliers and zooming in on the part that seems to fit best.



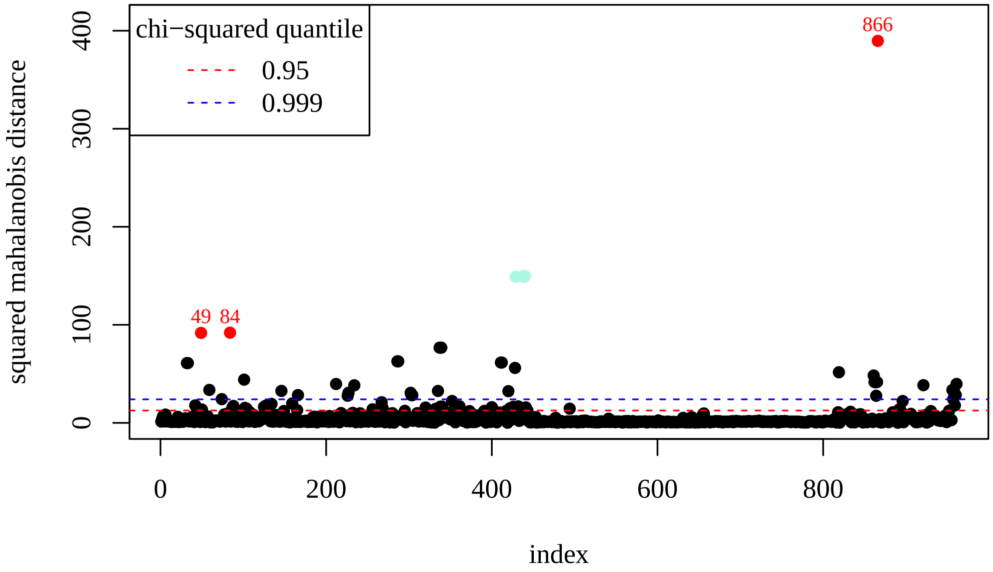
This further advocates against normality. Indeed, what at first sight seems an alignment to the Q-Q line, from a closer perspective reveals a mismatch also for smaller values of mahalanobis distance.

3.7

The Mahalanobis distance is not only of great use to asses the normality of the sample. It can also be employed to detect multivariate outliers. The following plot shows the squared Mahalanobis distances. We also displayed the quantiles of a χ^2_3 for the levels $\alpha = 0.95$ and $\alpha = 0.99$.



We observe that all the extreme univariate outliers we previously identified turned out to be also multivariate outliers indeed (despite not being the only ones). If we were to select only a handful of them, we would choose the most extreme ones, namely being observations 866, 49 and 84.
It can be of interest to asses weather these outliers were also multivariate outliers with respect to the original variables. Therefore, we also plot the corresponding squared Mahalanobis distance, with the quantiles of a χ^2_6 of levels $\alpha = 0.95$ and $\alpha = 0.99$.



Not surprisingly, these three outliers are indeed multivariate outliers also for the joint distribution of the original variables. It is worth to remark that also observations 429, 438, 439, 440 (marked with ●) turn out to be multivariate outliers for the original variables. However, they are not so in the principal components.