First Assignment Multivariate Statistical Analysis

Claudio Peroni–Marco Murtinu

Exercise 1

1.1)

We begin by taking the log transformation of the variables x_2, x_3, x_4. From now on we will refer to the logarithm of the variables using the name of the variables themselves.

1.2)

We compute the sample mean, variance and correlation of the variables using the built-in R commands.

$$\mu = \begin{bmatrix} 1.121 \\ 4.263 \\ 3.244 \end{bmatrix} \tag{1}$$

$$S = \begin{bmatrix} 0.525 & 0.003 & 0.086 \\ 0.003 & 0.012 & -0.013 \\ 0.086 & -0.013 & 0.091 \end{bmatrix} \tag{2}$$

$$R = \begin{bmatrix} 1.000 & 0.043 & 0.393 \\ 0.043 & 1.000 & -0.400 \\ 0.393 & -0.400 & 1.000 \end{bmatrix} \tag{3}$$

The first thing we notice in the matrix of the variances and covariances S is that the variance of x_2 is greater by one order of magnitude than the other two variances. By looking at the correlation matrix R we note that the correlation between x_2 and x_4 is almost the same of the one between x_3 and x_4 (in absolute value), while the correlation between x_2 and x_3 is almost negligible.

1.3)

First we plot the boxplot of all the variables together. This graph shows that the range of the variables is quite different. Then we plot the boxplot of each variable separately in order to highlight the outliers. We find that both the second and the third variable have one outlier each. They are respectively the 34th and 47th observation. It appears that there is no univariate outlier for the first variable.

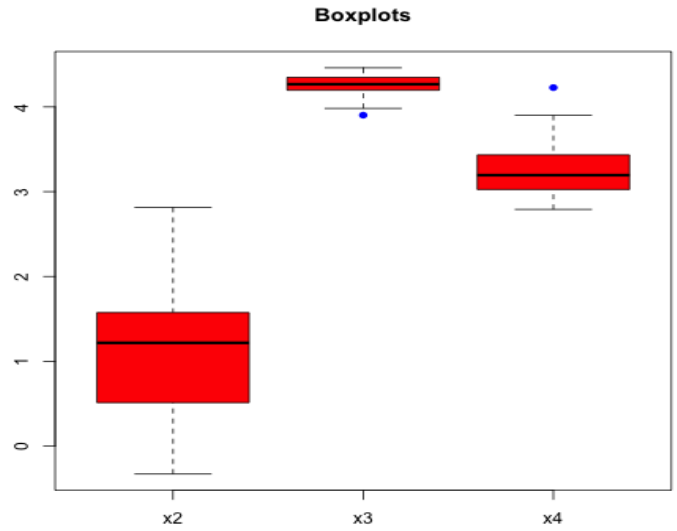

Figure 1: The boxplot of the variables. Outliers in blue.

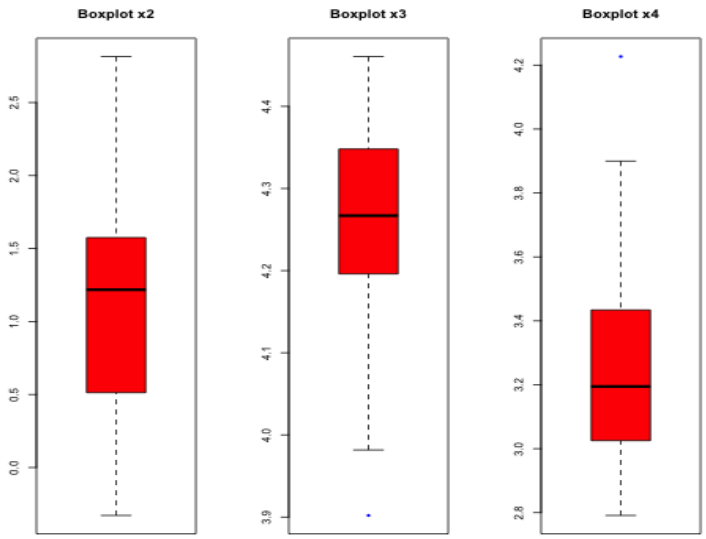

Figure 2: The boxplots for each variable separately. Outliers always in blue.

1.4)

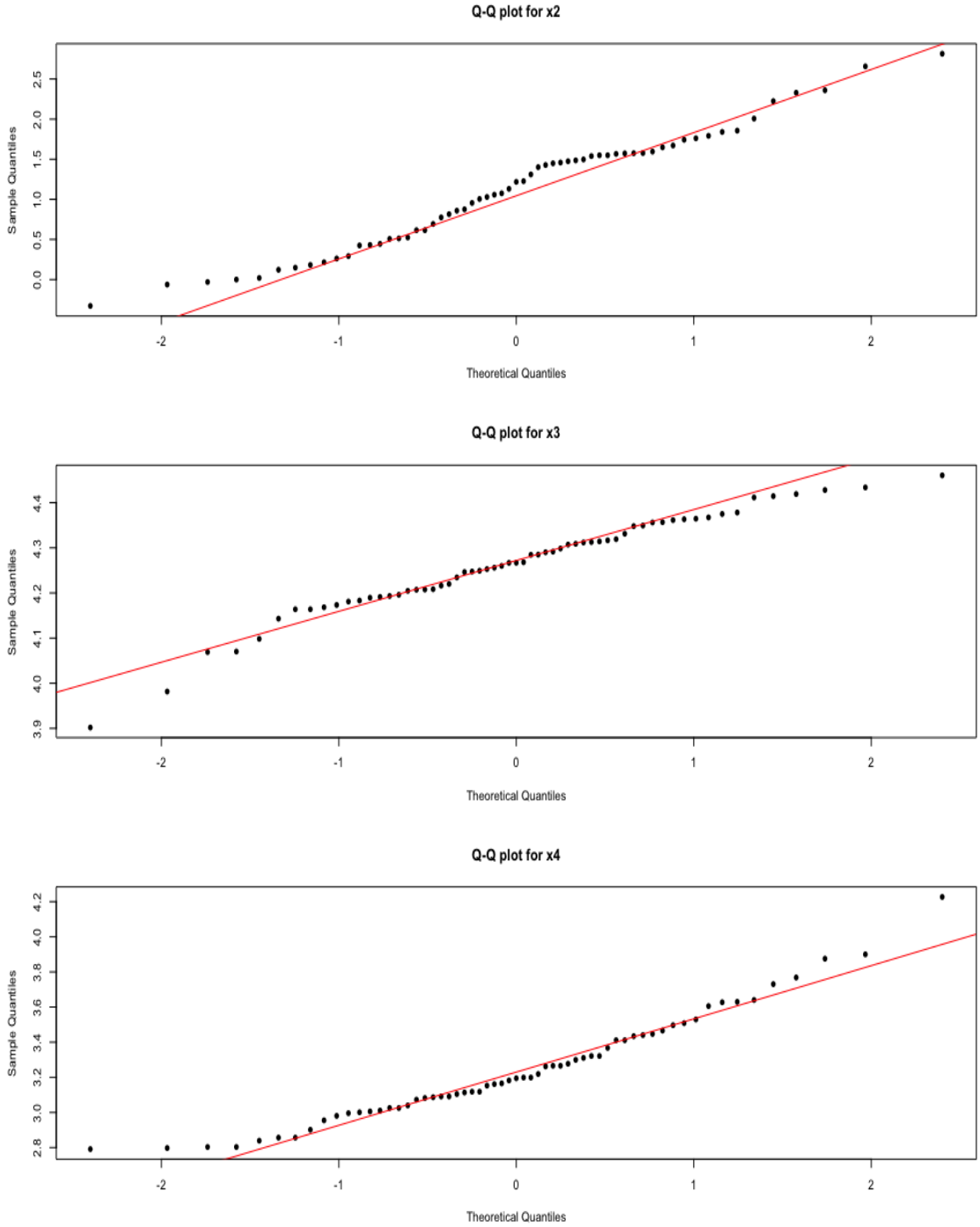

Figure 3: Q-Q plots for the variables.

A visual inspection of the normal Q-Q plots leads us not to reject the univariate normality hypothesis. We notice that there is a slightly unexpected behaviour at the tails.

1.5)

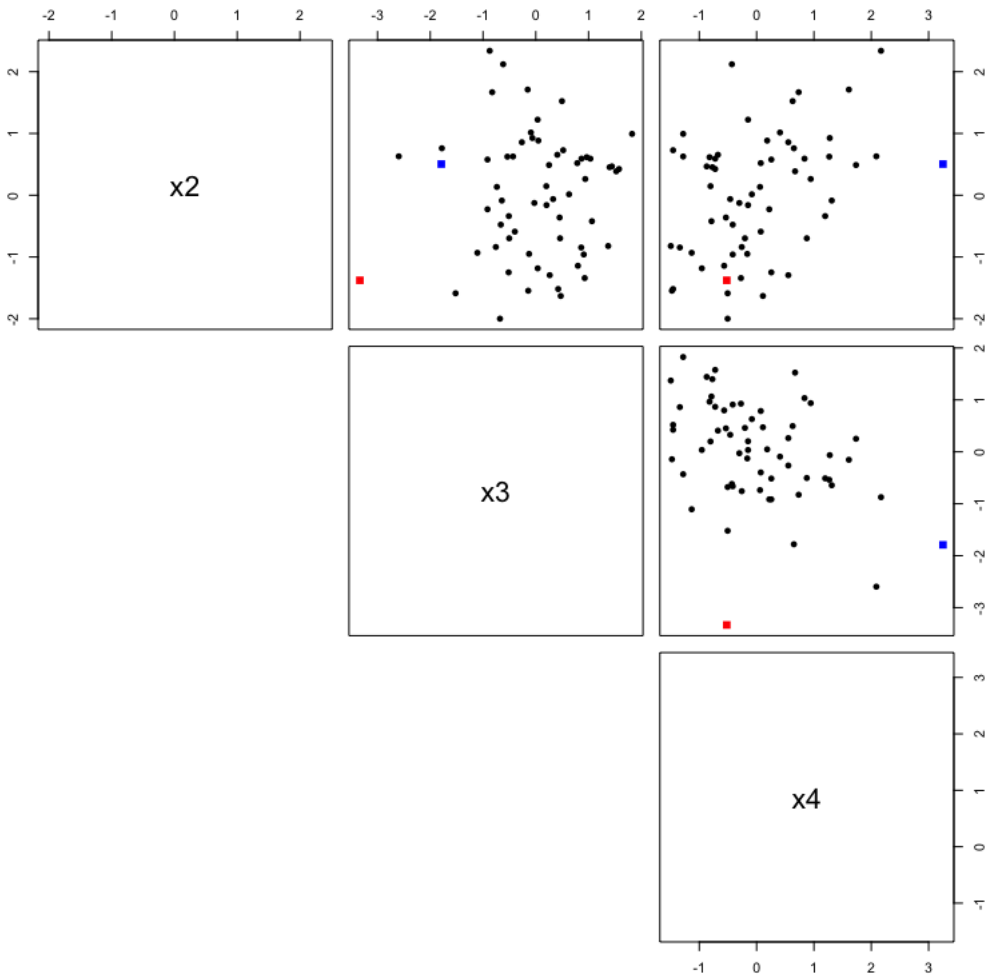

Figure 4: Scatterplots for the variables. In blue the outlier for x_4 and in red the outlier for x_3.

The univariate outliers we found for x_3 and x_4 are still visible in the bivariate representation of the scatterplots when the variable for which they are univariate outliers is plotted against another.

1.6)

To compute the squared Mahalanobis distance d^2 we use the formula:

$$d^2 = (x_i - \bar{x})^T \mathbf{S}^{-1} (x_i - \bar{x}) \tag{4}$$

The result is:

```
 [1]  0.835  4.254  3.673  2.092  2.571  8.031  1.635  2.241  2.544  0.069  1.373  2.302  2.664  3.611
[15]  1.050  1.993  4.039  2.865  2.157  2.065  4.122  0.968  4.303  4.933  2.072  3.992  0.861  3.095
[29]  4.263  0.429  0.824  1.293  0.521 14.951  4.327  4.359  4.939  0.957  2.118  3.154  2.239  0.118
[43]  1.722  3.864  3.195  1.970 11.452  7.480  8.517  0.685  0.990  3.359  0.247  0.890  1.372  0.406
[57]  5.450  1.935  0.818  1.608  1.469
```

Figure 5: the output of the d^2 vector in the R console.

1.7)

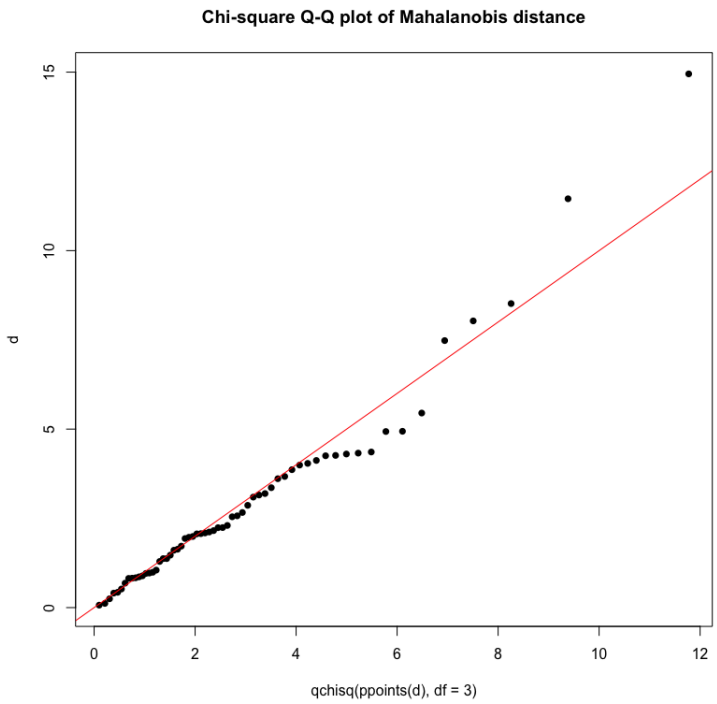

Figure 6: Chi-squared Q-Q plot of the d^2 distance.

This plot shows that the d^2 distances follow the q-q line quite strictly except the last two. Hence we do not reject normality of the data.

1.8)

To find other candidates outliers we sort the d^2 distances and look for unusually large or small values. We see that 5 observations present unusually

```
 [1] 0.069  0.118  0.247  0.406  0.429  0.521  0.685  0.818  0.824  0.835  0.861  0.890  0.957  0.968
[15] 0.990  1.050  1.293  1.372  1.373  1.469  1.608  1.635  1.722  1.935  1.970  1.993  2.065  2.072
[29] 2.092  2.118  2.157  2.239  2.241  2.302  2.544  2.571  2.664  2.865  3.095  3.154  3.195  3.359
[43] 3.611  3.673  3.864  3.992  4.039  4.122  4.254  4.263  4.303  4.327  4.359  4.933  4.939  5.450
[57] 7.480  8.031  8.517 11.452 14.951
```

Figure 7: The sorted d^2 distances.

large distances. They are the 6th, 34th, 47th, 48th, 49th observations. We had already classified as outliers two of them (the 34th and 47th) during the univariate analysis. In order to visualise these 3-dimensional outliers we use a 3D-scatterplot. Here we have coloured all the 5 outliers found.

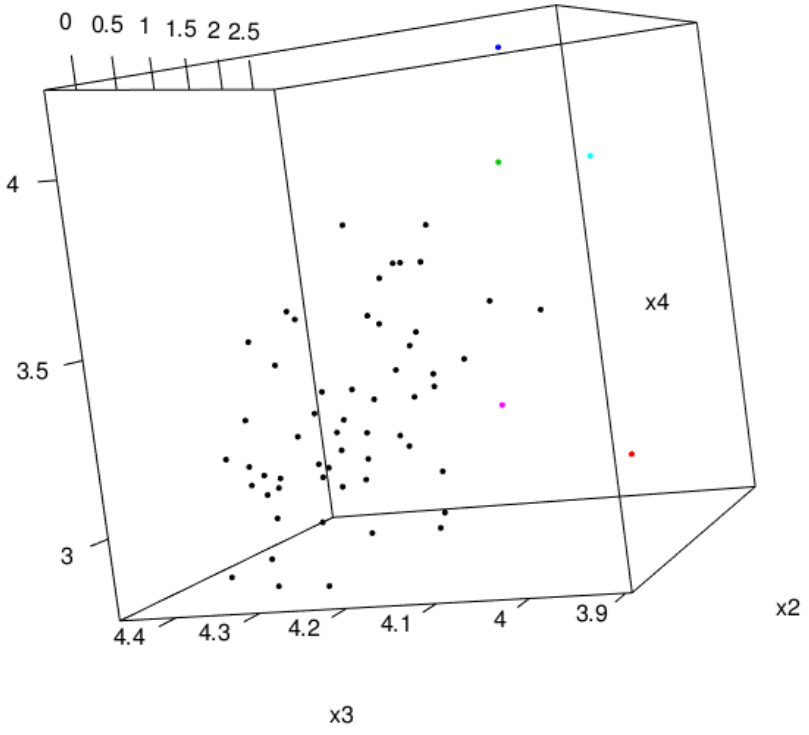

Figure 8: 3D scatterplot of the data.

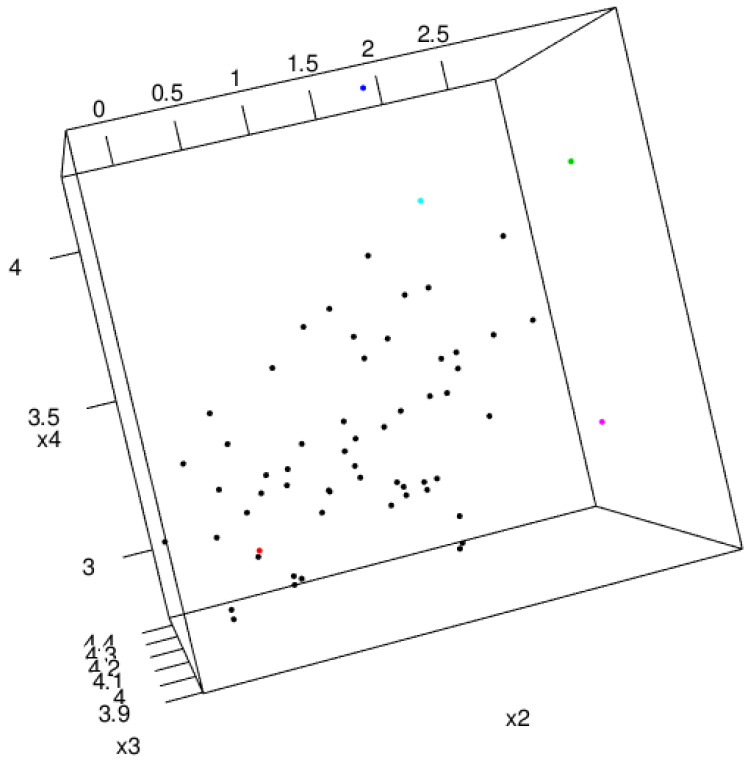

Figure 9: 3D scatterplot of the data from a different perspective.

Exercise 2

2.1)

We have to compute the distribution of $X_3|(X_1, X_2) = (x_1, x_2)$. In order to do so, we apply the following general formulas:

$$\bar{\mu} = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(a - \mu_2) \tag{5}$$

$$\bar{\Sigma} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} \tag{6}$$

to our case (where $a = (x_1, x_2)$). The result is:

$$\bar{\mu}_3 = 2 - \frac{1}{2}x_1 - \frac{1}{2}x_2 \tag{7}$$

$$\bar{\Sigma}_3 = \frac{3}{4} \tag{8}$$

Therefore, we conclude that $X_3|(X_1, X_2) = (x_1, x_2)$ is distributed as a normal with mean $\bar{\mu}_3$ and variance $\bar{\Sigma}_3$.

2.2)

Again we use (5) and (6) to compute the distribution of

$$(X_1, X_2)|X_3 = x_3 = 0$$

We get:

$$\bar{\mu}_{12} = \left(\frac{3}{2}, -\frac{1}{2}\right) \tag{9}$$

$$\bar{\Sigma}_{12} = \begin{bmatrix} \frac{15}{16} & -\frac{9}{16} \\ -\frac{9}{16} & \frac{15}{16} \end{bmatrix} \tag{10}$$

Once more, we conclude that $(X_1, X_2)|X_3 = x_3 = 0$ is distributed as $N(\bar{\mu}_{12}, \bar{\Sigma}_{12})$. Then, we sketch the ellipse

$$(x - \bar{\mu}_{12})^T \bar{\Sigma}_{12}^{-1}(x - \bar{\mu}_{12}) = \chi^2_{2,\alpha}$$

In the 2-dimensional space $x = (x_1, x_2)$ for $\alpha = 0.1$.

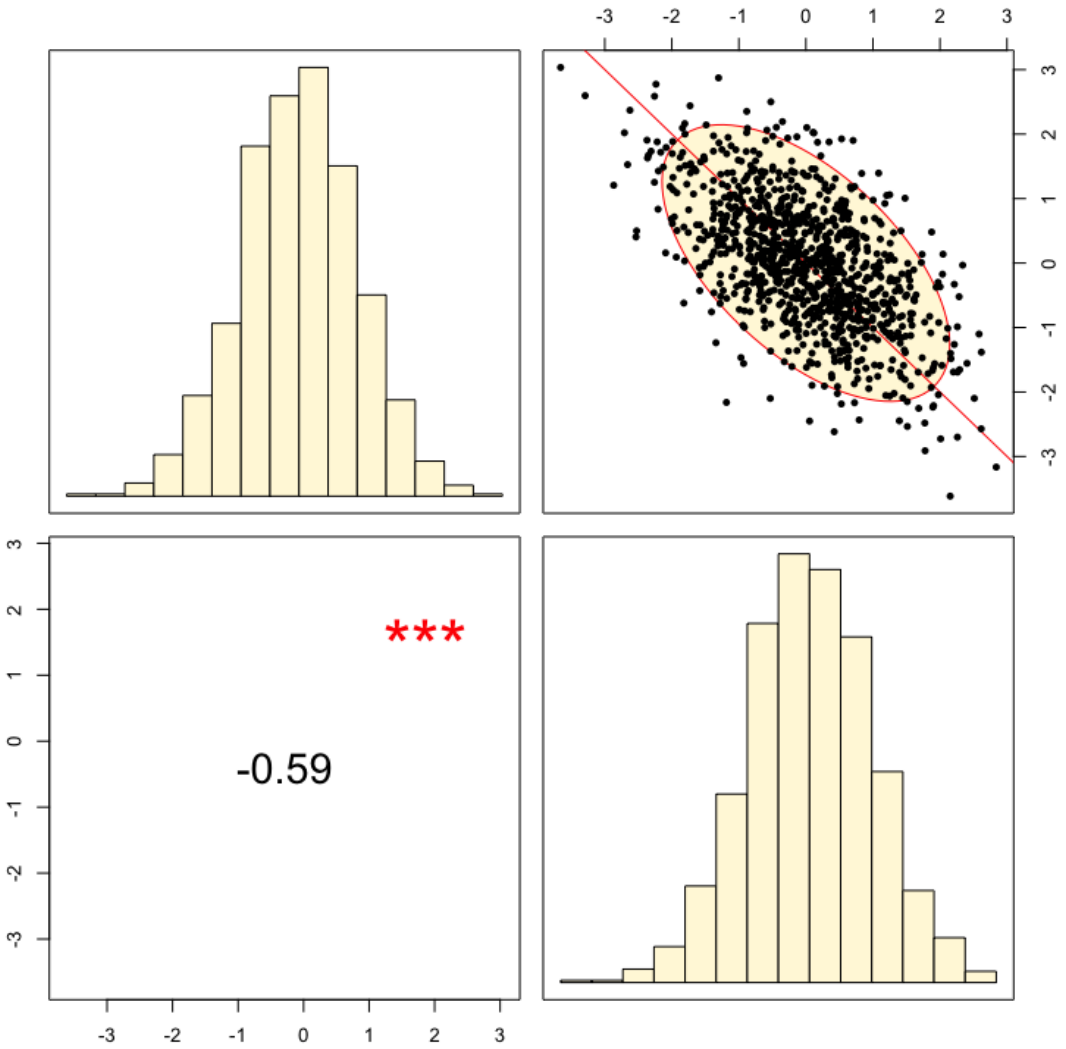

Figure 10: Scatterplot of data with ellipse (of level $\alpha = 0.1$), univariate histograms, and correlation between the variables.

Exercise 3

3.1)

We perform the Principal Component Analysis on the standardised variables of the *socio* dataset. The result is the following:

	PC1	PC2	PC3	PC4	PC5
x1	0.263	-0.463	0.784	-0.217	0.235
x2	-0.593	-0.326	-0.164	0.145	0.703
x3	0.326	-0.605	-0.225	0.663	-0.194
x4	-0.479	0.252	0.551	0.572	-0.277
x5	-0.493	-0.500	-0.069	-0.407	-0.580

Figure 11: Rotation matrix.

Each columns contains the coefficients of the linear combination of the original variables for each Principal Component.

3.2)

	PC1	PC2	PC3	PC4	PC5
Standard deviation	1.411	1.169	0.930	0.731	0.491
Proportion of Variance	0.398	0.274	0.173	0.107	0.048
Cumulative Proportion	0.398	0.672	0.845	0.952	1.000

Figure 12: The *importance matrix* of the principal components.

The cumulative proportion of variance explained by the first two PCs is 0.672.

3.3)

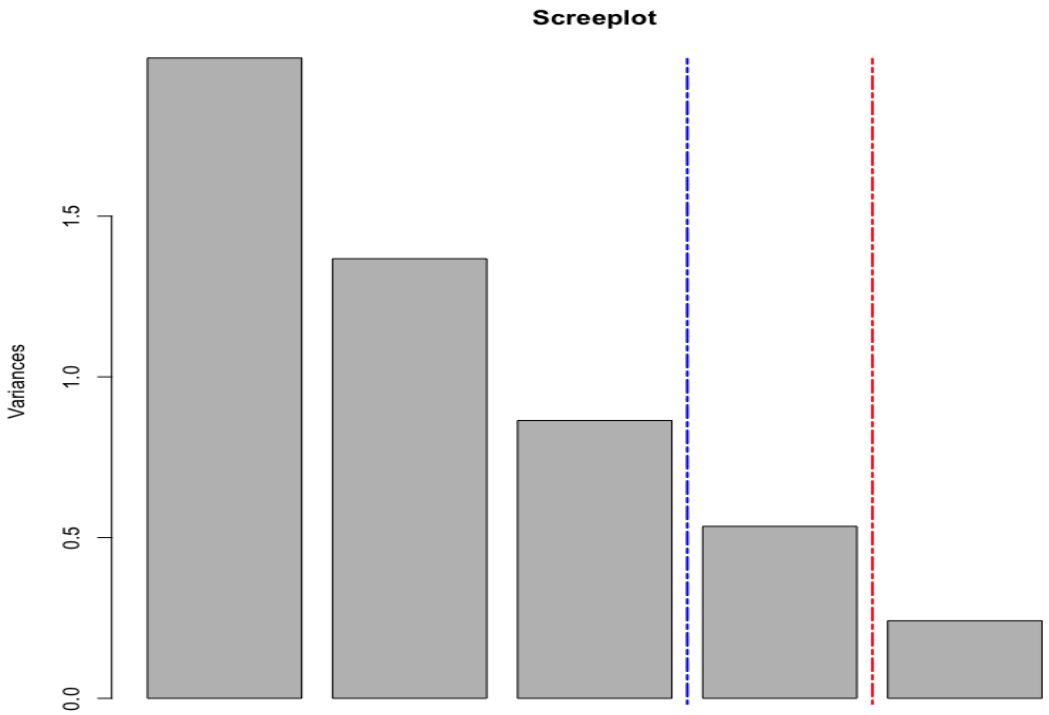

Figure 13: Screeplot (type "barplot") of the PCs. In blue and red the suggested tresholds.

From the screeplot and the importance matrix we can see that 3 PCs are enough to explain more than 80% of the variance. For many uses this is probably enough. If we add one more PC, we are able to account for $\sim 95\%$ of the total variance, so in some cases this might be the best choice.

3.4)

First we retrieve the PCA results from Lecture 4.

	PC1	PC2	PC3	PC4	PC5
x1	0.038	-0.071	-0.182	-0.979	-0.055
x2	-0.104	-0.130	0.962	-0.165	-0.139
x3	0.492	-0.865	-0.047	0.090	0.005
x4	-0.863	-0.480	-0.152	0.029	0.007
x5	-0.009	-0.015	0.126	-0.078	0.989

Figure 14: Principal Component rotation matrix from Lecture 4 as comparison. Here data are *not* standardised.

This rotation matrix suggests that the first component can be seen as an explanatory variable for x_4, while the second principal component for x_3 (and a similar relation can be found between the other 3 PCs with the remaining variables). The link between each PC and one of the original variables is more evident here than in our analysis, see Figure 11.

	PC1	PC2	PC3	PC4	PC5
Standard deviation	10.416	6.352	2.910	1.698	0.395
Proportion of Variance	0.677	0.252	0.053	0.018	0.001
Cumulative Proportion	0.677	0.928	0.981	0.999	1.000

Figure 15: The *importance matrix* of the principal components from PCA in Lecture 4.

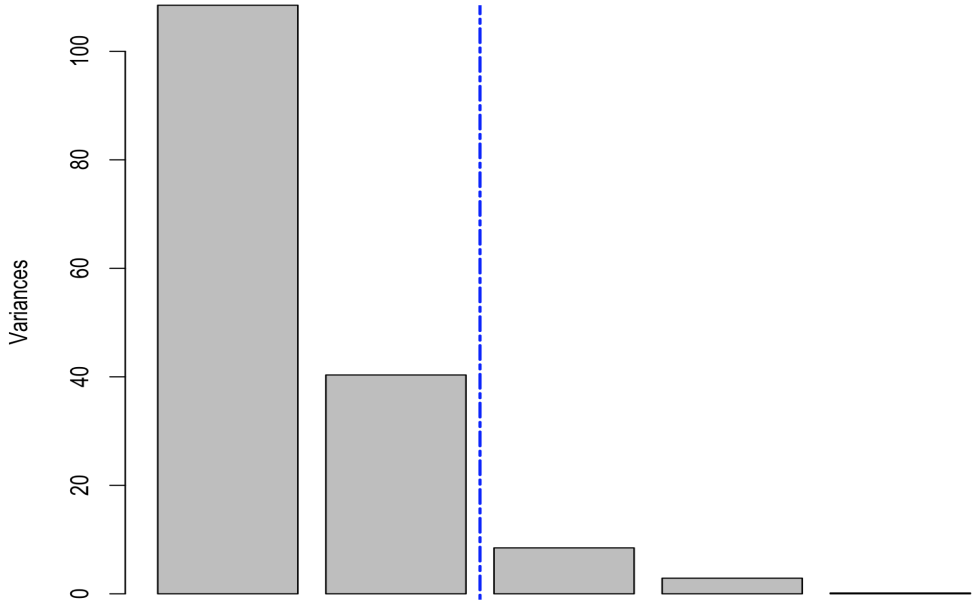

Figure 16: A Graphical visualisation of the cumulative proportion of variance explained bye the PCs (bar screeplot). In blue the obvious cutoff.

Concerning the Cumulative Proportion explained by the Principal Components, we see that the first two PCs are enough to account for more than 90% of the total variance, while in our analysis we need 4 PCs to reach a similar explanation proportion. This difference is due to the standardisation of our data. With non-standardised data the variables which have greater variance tend to *weigh* more in the cumulative proportion.