Problem_Set_1

Pierpaolo De Blasi

deadline 2023-04-05 11.59pm

Exercise 1

The data set `state.x77` (package `dataset`) contains 8 variables recorded to the 50 states of the United States of America in year 1977.

Since it is not a `data.frame` object, we coerce it first into a data frame

```
st<-as.data.frame(state.x77)
head(st)
```

```
##            Population Income Illiteracy Life Exp Murder HS Grad Frost   Area
## Alabama          3615   3624        2.1    69.05   15.1    41.3    20  50708
## Alaska            365   6315        1.5    69.31   11.3    66.7   152 566432
## Arizona          2212   4530        1.8    70.55    7.8    58.1    15 113417
## Arkansas         2110   3378        1.9    70.66   10.1    39.9    65  51945
## California      21198   5114        1.1    71.71   10.3    62.6    20 156361
## Colorado         2541   4884        0.7    72.06    6.8    63.9   166 103766
```

We change a couple of variable names so to avoid spaces, and add a population density variable.

```
names(st)[4] = "Life.Exp"
names(st)[6] = "HS.Grad"
st[,9] = st$Population * 1000 / st$Area
colnames(st)[9] = "Density"
```

For more information on what these variables are, see the help page of `state.x77`.

1. Compute the correlation matrix and comment on the most relevant relationships among variables (up to 10).
2. Find univariate outliers, up to 3 per variable, up to 10 in total.
3. Make a boxplot of any variable plotting the corresponding outliers, if any, found in point 2 in red.
4. Comment about normality of each variable.
5. Make a scatter plot of `Area` vs `Population`, colour-coding the outliers found in point 2 with a different colours. Choose among the following colour names. Can they be considered bivariate outliers?

```
lookup<-c("darkgreen",  "brown", "lightblue",  "magenta", "purple",
                    "blue", "red", "lightgreen", "orange", "cyan")
```

6. Construct a chi-square Q-Q plot of the squared Mahalanobis distances and comment about multivariate normality.
7. Identify multivariate outliers, if any, and compare with the univariate outliers previously found.

Exercise 2

Let $Z = (X, Y_1, Y_2)$ be distributed as $N_3(\mu, \Sigma)$,

$$\mu = \begin{bmatrix} 1 \\ 0 \\ 2 \end{bmatrix} \quad \mathbf{\Sigma} = \begin{bmatrix} 1 & -\rho & \rho \\ -\rho & 1 & \rho \\ \rho & \rho & 1 \end{bmatrix}, \quad -1 < \rho < 0.5$$

1. Find the inverse of $\mathbf{\Sigma}$ *[xx use* $\mathbf{\Sigma} = (1+\rho)\boldsymbol{I} - \rho a\, a^T$ *for* $\boldsymbol{I}$ *identity matrix and* $a = (1, 1, -1)$ *xx]*
2. Find the eigenvalues of Σ.
3. Let PC1 and PC2 be the first two (population) principal components of Z. Find ρ such that they account for more than 80% of total variation of X.
4. Find the conditional distribution of $Y = (Y_1, Y_2)$ given $X = x$.
5. Let $\rho = 0.2$, and $\mathbf{\Sigma}_y$ and μ_y be the corresponding covariance matrix and the mean vector of the distribution of $Y = (Y_1, Y_2)$ given $X = 0$. Sketch the ellipse

$$(y - \mu_y)^T \mathbf{\Sigma}_y^{-1} (y - \mu_y) = c^2,$$

in the 2 dimensional space $y = (y_1, y_2)$ by setting the constant " c " such that the ellipse contains 0.95 probability with respect to the conditional distribution of Y.

Exercise 3

Nutritional data from 961 different food items is given in the file `nutritional.txt`

```
nutritional<-read.table("data/nutritional.txt")
head(nutritional)
```

```
##   fat food.energy carbohydrates protein cholesterol weight saturated.fat
## 1   2          25             2       0           2  15.00           0.2
## 2   6          60             2       0           4  16.00           1.0
## 3   1          90            22       4           0  28.35           0.1
## 4   0          90            22       3           0  28.35           0.1
## 5   0          10             1       1           0  33.00           0.0
## 6   1          70            21       4           0  28.35           0.1
```

For each food item, there are 7 variables: `fat` (grams), `food.energy` (calories), `carbohydrates` (grams), `protein` (grams), `cholesterol` (milligrams), `weight` (grams), and `saturated.fat` (grams).

1. To equalize out the different types of servings of each food, first divide each variable by `weight` of the food item (which leaves us with 6 variables). Next, because of the wide variations in the different variables, standardize each variable. Perform Principal Component Analysis on the transformed data.
2. Decide how many components to retain in order to achieve a satisfactory lower-dimensional representation of the data. Justify your answer.
3. Give an interpretation to the first two principal components
4. Identify univariate outliers with respect to the first three principal components, up to 3 per component. These points correspond to foods that are very high or very low in what variable (up to 2 variables per observation)?
5. Make a 3-d scatter plot with the first three principal components, while color coding these outliers.
6. Investigate multivariate normality through the first three principal components.
7. Find multivariate outliers through the first three principal components, up to 5 in total. Are they the most extreme observations with respect to the 6 original variables?

Problem set 1, notes

Exercise

Let $Z = (X, Y_1, Y_2)$ be distributed as $N_3(\mu, \Sigma)$,

$$\mu = \begin{bmatrix} 1 \\ 0 \\ 2 \end{bmatrix} \quad \Sigma = \begin{bmatrix} 1 & -\rho & \rho \\ -\rho & 1 & \rho \\ \rho & \rho & 1 \end{bmatrix}, \quad -1 < \rho < 0.5$$

2.1

1. Find the inverse of Σ *[xx use $\Sigma = (1+\rho)\boldsymbol{I} - \rho a\, a^T$ for $\boldsymbol{I}$ identity matrix and $a = (1, 1, -1)$ xx]*

Solution

Let first note that Σ is invertible since $\det(\Sigma) = -2\rho^3 - 3\rho^2 + 1$ which is greater than 0 $\forall\ -1 < \rho < 1/2$.

We compute the inverse of Σ by exploiting the identity

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

and applying the following theorem, known as the Neumanns series theorem.

Theorem: Let T be a linear mapping $T: \mathbb{R}^m \to \mathbb{R}^m$. 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

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

Let $A := (I - c\, a a^T)$ it holds

$$A^{-1} = (I - c\, a a^T)^{-1} =$$

$$= \sum_{i=0}^{\infty} (c\, a a^T)^i =$$

$$= \sum_{i=0}^{\infty} c^i (a a^T)^i =$$

$$= I + \sum_{i=1}^{\infty} c^i (\|a\|^2)^{i-1}\, a a^T =$$

$$= I + c \cdot a a^T \sum_{i=1}^{\infty} (c\, \|a\|^2)^{i-1} =$$

$$= I + c \cdot a a^T \cdot \sum_{i=0}^{\infty} (c\, \|a\|^2)^i =$$

$$= I + c \cdot a a^T \cdot \frac{1}{1 - c\|a\|^2} =$$

$$= I + \rho/(1+\rho) \cdot \frac{1}{1 - 3\,\rho/(1+\rho)} \cdot a a^T =$$

$$= I + \rho/(1 + \rho - 3\rho) \cdot a a^T =$$

$$= I + \rho/(1-2\rho) \cdot aa^T$$

Thus

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

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

$$\Sigma^{-1} = y I + k\, aa^T$$

and then find the values for y and k.

computations Elsewhere

2.2

2. Find the eigenvalues of Σ.

Solution

computations shortcut

A faster way to find the spectrum (set of eigenvalues, meant with multiplicity) is reported below. It exploits some basic properties of the spectrum.

We denote $Sp(\Sigma)$ the spectrum of the matrix Σ.

$$Sp(\Sigma) = Sp\left((1+\rho)\left(I - \rho/(1+\rho)\, aa^T\right)\right) =$$

$$= (1+\rho)\, Sp\left(I - \rho/(1+\rho)\, aa^T\right) =$$

$$= (1+\rho)\left(1 - \rho/(1+\rho)\, Sp(aa^T)\right) =$$

Observing $(aa^T)a = \|a\|^2 \cdot a$, and $rank(aa^T) = 1$ it holds

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

Hence

$$Sp(\Sigma) = (1+\rho)\left(1 - \rho/(1+\rho) \cdot \{0, 0, \|a\|^2\}\right) =$$

$$= (1+\rho)\left\{1, 1, 1 - \rho/(1+\rho)\right\} = \left\{1+\rho, 1+\rho, \cancel{(1+\rho)}\, \frac{(1+\rho-\rho)}{\cancel{1+\rho}}\right\} =$$

$$= \{1+\rho, 1+\rho, 1-\rho\},$$

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

2.3

3. Let PC1 and PC2 be the first two (population) principal components of Z. Find ρ such that they account for more than 80% of total variation of X.

Solution

We first write the eigenvalues of Σ in ascending order. We distinguish the following 2 cases :

1. if $\rho \in [0, 1/2)$, 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 PCs account for more than 80% of the total var. of Σ.
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

$$2(1+\rho) > 4/5 \cdot 3 \iff 1+\rho > 6/5$$
$$\iff \rho > 1/5$$

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 point we obtain

$$(1-2\rho) + (1+\rho) > 4/5 \cdot 3 \iff 2-\rho > 12/5$$

$\Leftrightarrow \quad \rho < -2/5.$

Hence for $\rho \in [0, 1)$ it must be $\rho > 1/2$

and for $\rho \in [-1, 0)$ it must be $\rho \leq -2/5$.

So $\forall \rho \in (-1, -2/5) \cup (1/2, 1)$ PC1 and PC2 account for more than 80%

2.4

4. Find the conditional distribution of $Y = (Y_1, Y_2)$ given $X = x$.

Solution

intro Eloomore

Proposition : Let $X = (X_1, X_2) \sim N_p(\mu, \Sigma)$ with

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{12}^T & \Sigma_{22} \end{pmatrix}$$

where the dimension of X_1 is $q < p$.

Then the conditional distribution of $X_2 \mid X_1 = x$ is $N_{p-q}(\tilde{\mu}, \tilde{\Sigma})$ with

$$\tilde{\mu} = \mu_2 + \Sigma_{12}^T \Sigma_{11}^{-1} (X - \mu_1)$$

$$\tilde{\Sigma} = \Sigma_{22} - \Sigma_{12}^T \Sigma_{11}^{-1} \Sigma_{12}.$$

By applying the proposition to

$$z = (x, y_1, y_2) \sim N_3(\mu, \Sigma) \quad \text{with}$$

$$\mu = \mathbb{E}[z] = \begin{pmatrix} \mathbb{E}[x] \\ \mathbb{E}\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \hline \mu_{2,1} \\ \mu_{2,2} \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \quad \text{and}$$

$$\Sigma = \left(\begin{array}{c|cc} \mathrm{Var}(x) & \mathrm{Cov}(x, y_1) & \mathrm{Cov}(x, y_2) \\ \hline \mathrm{Cov}(x, y_1) & \multicolumn{2}{c}{\multirow{2}{*}{Var } y} \\ \mathrm{Cov}(x, y_2) & & \end{array}\right) =$$

$$= \left(\begin{array}{c|c} \Sigma_{11} & \Sigma_{12} \\ \hline \Sigma_{12}^T & \Sigma_{22} \end{array}\right)$$

We obtain

$$(y_1, y_2) \mid X = x \sim N_2(\tilde{\mu}, \tilde{\Sigma}) \quad \text{with}$$

$$\begin{cases} \tilde{\mu} = \mu_2 + \Sigma_{12}^T \cdot \Sigma_{11}^{-1} (x - \mu_1) \\ \tilde{\Sigma} = \Sigma_{22} - \Sigma_{12}^T \Sigma_{11}^{-1} \Sigma_{12} \end{cases}$$

By substituting what we found we get

$$\tilde{\mu} = \left(\mathbb{E}\begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} + \begin{pmatrix} Cov(X, Y_1) \\ Cov(X, Y_2) \end{pmatrix} \cdot \frac{x - \mathbb{E}[X]}{Var(X)} \right) =$$

$$= \begin{pmatrix} 0 \\ 2 \end{pmatrix} + \begin{pmatrix} -\rho \\ \rho \end{pmatrix} \cdot \frac{x - 1}{1} = \begin{pmatrix} 0 - \rho(x-1) \\ 2 + \rho(x-1) \end{pmatrix} =$$

$$= \begin{pmatrix} -\rho(x-1) \\ 2 + \rho(x-1) \end{pmatrix}.$$

$$\tilde{\Sigma} = \left(Var(Y) - \begin{pmatrix} Cov(X, Y_1) \\ Cov(X, Y_2) \end{pmatrix} \cdot \left(Cov(X, Y_1), Cov(X, Y_2) \right) \cdot 1/Var(X) \right) =$$

$$= \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} - \begin{pmatrix} -\rho \\ \rho \end{pmatrix} \begin{pmatrix} -\rho, \rho \end{pmatrix} \cdot 1/1 =$$

$$= \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} - \rho^2 \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} = \begin{pmatrix} 1-\rho^2 & \rho+\rho^2 \\ \rho+\rho^2 & 1-\rho^2 \end{pmatrix}.$$

$$\Rightarrow (Y_1, Y_2) \mid X = x \sim N_2\left(\begin{pmatrix} -\rho(x-1) \\ 2 + \rho(x-1) \end{pmatrix}, \begin{pmatrix} 1-\rho^2 & \rho+\rho^2 \\ \rho+\rho^2 & 1-\rho^2 \end{pmatrix} \right). \square$$

2. 5

5. Let $\rho = 0.2$, and Σ_y and μ_y be the corresponding covariance matrix and the mean vector of the distribution of $Y = (Y_1, Y_2)$ given $X = 0$. Sketch the ellipse

$$(y - \mu_y)^T \Sigma_y^{-1} (y - \mu_y) = c^2,$$

in the 2 dimensional space $y = (y_1, y_2)$ by setting the constant "c" such that the ellipse contains 0.95 probability with respect to the conditional distribution of Y.

Let $\rho = 1/5$, then according to what we found in 2.4 we obtain

$$\begin{cases} \mu_y = \begin{pmatrix} -\rho(x-1) \\ 2 + \rho(x-1) \end{pmatrix} \Big|_{x=0,\ \rho=1/5} \\ \Sigma_y = \begin{pmatrix} 1-\rho^2 & \rho+\rho^2 \\ \rho+\rho^2 & 1-\rho^2 \end{pmatrix}_{\rho=1/5} \end{cases} \iff \begin{cases} \mu_y = 1/5 \begin{pmatrix} 1 \\ 9 \end{pmatrix} \\ \Sigma_y = 6/25 \begin{pmatrix} 4 & 1 \\ 1 & 4 \end{pmatrix} \end{cases}$$

Since Σ_y is invertible ($\det(\Sigma) = 6 \cdot 6/25 > 0$) the random variable $(y - \mu_y)^T \Sigma_y^{-1} (y - \mu_y)$ is well defined. Moreover since $y \sim N_2(\mu_y, \Sigma_y)$ then

$$(y - \mu_y)^T \Sigma_y^{-1} (y - \mu_y) \sim \chi^2_2 .$$

Considering the 2-dimensional space $y = (y_1, y_2)$ and letting $c \in \mathbb{R}^+$ we have that

$$(y - \mu_y)^T \Sigma_y^{-1} (y - \mu_y) = c^2$$

defines a contour line of the density function of y which is an ellipse that contains the following percentage of the mass:

$$\mathbb{P}\left((y - \mu_y)^T \Sigma_y^{-1} (y - \mu_y) \leq c^2 \right).$$

By imposing this probability to 0.95, since $(y - \mu_y)^T \Sigma_y^{-1} (y - \mu_y) \sim \chi^2_2$, we get that

$$c^2 = \chi^2_{2,\,0.95} .$$

* computation

The ellipse is centered in μ_y and its axes have

length $c\sqrt{\lambda_1}$, $c\sqrt{\lambda_2}$ and direction e_1, e_2 where (λ_1, e_1) and (λ_2, e_2) are the eigen-pairs of the matrix Σ_y.

computation

plt

Exercise 1

1.1

Sol

1. Compute the correlation matrix and comment on the most relevant relationships among variables (up to 10).

Matrix (corplot)

- Negative higher:

  1. Cor(murder, life_exp) = -0.78;

     + murder => - life expectancy.

     It is reasonable. more murders implies an overall reduction of life expectancy.

  2. $Cor\begin{pmatrix} a\ frost \\ b\ hs\_grad \\ c\ life\_exp \end{pmatrix}, illiteracy) = \begin{pmatrix} -0,67 \\ -0,66 \\ -0,59 \end{pmatrix}$;

     a. there are not natural consideration to justify it;

     b. reasonable: more graduates => less illiterate citizens, actually still weird that it is slightly less than 2.;

     c. it also means sense indeed if a major percentage of citizens is educated than the overall life expectancy should increase.

- Positive higher:

  1. Cor(murder, illiteracy);

     it is reasonable that n. of murder and perc. of illiterates are pos. corr since the two values are intuitively bind.

  2. $Cor\begin{pmatrix} Income \\ life\_exp \end{pmatrix}, hs\_grad) = \begin{pmatrix} 0.62 \\ 0.58 \end{pmatrix}$

     As we expect the graduates percentage is highly correlated with both income and life_exp which are two important indicators of well-being.

Final notes:

- The variable which have an higher correlation with the other components (both in negative and pos. sense) are hs_grad, illiteracy and murder.
- On the contrary the ones which are less correlated along with the others are population, area and density.
  Note that density is derived from population and area thus it is reasonable that it follows their behaviour in terms of correlations.
  However it is quite surprising that the couples ("density", "area") and ("density", "population") have quite low correlations since, as we said, density is derived from them.

1.2

2. Find univariate outliers, up to 3 per variable, up to 10 in total.

In order to find univariate outliers we first scaled our dataset and we then identified the values x s.t.

$$|x| > \bar{\Phi}^{-1}(0.99) \quad \text{with } \Phi \text{ CDF of a } N(0,1).$$

We use the 99^{th} percentile since by taking lower values the "potential" outliers would have been too many.
However we noticed the presence of another potentially too high (in absolute value) value for the variable used by taking the 98.75^{th} percentile.

metrics

1.3

3. Make a boxplot of any variable plotting the corresponding outliers, if any, found in point 2 in red.

We inserted below the boxplots corresponding to each variable where we highlighted in red the outliers found in point 1.2

chart

- Note that the boxplots generated many other potential outliers but the ones we did not detect in the previous point are not in the outer tails (at least 98.75th perc) of the distribution. So we will not consider them as outliers.
- Moreover note that the outlier identified for the literacy variable does not show in the corresponding boxplot. This is plausible since the distribution seems to have very fat tails. For this reason we choose to not consider it as an outlier.

1.4

4. Comment about normality of each variable.

In order to check whether each variable is normally distributed or not we first examine the relationships between the sample and theoretical quantiles through the corresponding Q-Q plots:

```
# q-q plot
```

- the values corresponding to the variable "income" lie all very near the Q-Q line except for the observation 2, previously identified as an outlier. Also the "life_exp" variable appears to be quite normal: the values are more spread out w.r.t. the ones of the variable income but they still are very close the blue line (which represents the linear relationship between the sample and theoretical quantiles).
- the variables "murder", "hs_grad" and "frost" have all similar behaviour. Most of the points lie near the Q-Q- line but they have a thinner right tail and a heavier left tail.

Note the absence of outliers.

- Also the variables "population", "area" and "density" have very similar shapes which is far for being linear. All of them have heavy tails which is also due to the presence of more them one outlier
- the trajectory of the variable "illiteracy" is very atypical indeed on the lft side of the plot we can observe a consistent percentage of the points which share the same (two) values.

In conclusion we can infer a gaussian behaviour only for the variables "income" and "life_exp"
We can draw the same conclusions looking at the histograms of the single variables.

histograms

Label: blue line: empirical density / red line: theoretical density.

copy from Eleonora's notes.

data.frame

As emerged above, the variables "income" and "life_exp" can be considered normally distributed since they have a p-value greater then 0.05.
Note also that the p-values of the variables variables "murder", "hs_grad" and "frost" are all close to 0.05: the only one for which the null hypotesis is not rejected is "frost".
However its p-value is $\simeq 0.527$ thus it is reasonable to doubt its normality also taking in account the

corresponding Q-Q-plot and histogram presented above. Now it may be interesting to observe how the previous tests change if we remove from each variable the points that we identified as their outliers.
Let us take a look at the qq plots and at the results of the shapiro test.

```
# qq-plot
# shap_test.ot
```

Note that:

- the variables population and density follows the same non-gaussian behaviour observed above. Indeed their p-values are almost negligible;
- for the variable income we can still say that its distribution is gaussian, however we can observe the presence of more pronounced tails. Its p-value is still > 0.05 but it is slightly lower than the previous one;
- finally by removing the outliers of the variable area it becomes more gaussian. Its p-value is very close to 0.06 but there is a huge difference between its current value and the previous one.

1.5

5. Make a scatter plot of `Area` vs `Population`, colour-coding the outliers found in point 2 with a different colours. Choose among the following colour names. Can they be considered bivariate outliers?

We report the scatter plot of the variables Area vs Population colouring all the potential outliers

```
# outlius
```

From the plot we can see that the mass is roughly concentrated in the rectangle $[0, 15 \cdot 10^4] \times [0, 15 \cdot 10^3]$.

Note that unlike the other univariate outliers, the value corresponding to the observations 2, 5, 32, 43 seems to be really far from the other points
Hence we could consider them as bivariate outliers.

1.6

6. Construct a chi-square Q-Q plot of the squared Mahalanobis distances and comment about multivariate normality.

The squared Malhanobis distance is a distance function that quantifies the gap between an observation and the sample mean weighted by the inverse of the covariance matrix.
If the variables are consider jointly normally distributed
$\Rightarrow d = \sum_{i=1}^{g} (N(0,1))^2 \sim \chi^2_g$. Hence we can check the multivariate normality looking at the Q-Q-plot of the squared Mahalanobis distances vs χ^2_g.

```
# Q-Q-plot
```

Note that the majority of the points are close to the Q-Q-line except for the one corresponding to the observation 2, previously detected as univariate outlier, and the second to last observation corresponding to the index 11.
However we can consider the d as χ^2_g-distributed, hence we can say that our variables are jointly distributed as a multivariate gaussian.

1.7

7. Identify multivariate outliers, if any, and compare with the univariate outliers previously found.

In order to identify multivariate outliers we can plot the vector of squared Mahalanobis distances and add some threshold lines corresponding to different levels of the theoretical quantiles of χ^2_g: we used $\alpha_1 = 0.95$ and $\alpha_2 = (n - 0.5)/n$ with $n = \text{nrow}(s6)$.

```
# plot outliers
```

The observations 2 and 11 are above the higher line (corresponding to $\chi^2_g(d_2)$) hence they can be considered multivariate outliers. This confirms what we previously noticed in the Chi-squared QQ-plot.

Some of the observations lies above the $\chi^2_g(d_1)$ line but below the $\chi^2_g(d_2)$ line but we choose to not consider them as multivariate outliers also taking in account what we observed in the previous point

Finally we can observe that the majority of the observations identified as univariate outliers can not be considered multivariate outliers except for the observation 2.

note states: Hawaii, Alaska.

3.1

1. To equalize out the different types of servings of each food, first divide each variable by `weight` of the food item (which leaves us with 6 variables). Next, because of the wide variations in the different variables, standardize each variable. Perform Principal Component Analysis on the transformed data.

First of all we divide each variable by \texttt{weight} in order to equalize out the different types of servings of each food.

codice

After the standardization of the data, carried out with the command \texttt{scale} we perform the Principal Component analysis on th

codice

3.2

2. Decide how many components to retain in order to achieve a satisfactory lower-dimensional representation of the data. Justify your answer.

pagina 3 relazione