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?