G0O02a: Statistical Data Analysis: Project 1

Prof. Mia Hubert, Can Hakan Dagidir

April 2024

The goal of the project is to analyze the Rice dataset. A computer vision system was created to differentiate between various types of registered rice that share similar characteristics. This dataset comprises 10 distinct morphological attributes, including area, perimeter, and roundness. The final variable, `Class`, denotes the specific type of rice for each observation. Here's a brief overview of the variables:

Area (A)	The area of a rice grain and the number of pixels within its boundaries.
Perimeter (P)	Rice circumference is defined as the length of its border.
Major Axis (L)	The distance between the ends of the longest line that can be drawn on a rice.
Minor Axis (ℓ)	The longest line that can be drawn from the rice while standing perpendicular to the main axis.
Eccentricity (Ec)	Eccentricity of the ellipse having the same moments as the area.
Equivalent Diameter (Ed)	The diameter of a circle having the same area as the rice area.
Convex Area (C)	Number of pixels in the smallest convex polygon that can accommodate the area of the rice grain.
Solidity (S)	The ratio of the pixels in the convex area to those found in the rice grain: A/C.
Aspect Ratio (K)	L/ℓ
Roundness (R)	$4\pi A/P^2$
Class (Cl)	1 (Arborio), 2 (Basmati), 3 (Ipsala), 4 (Karacadag)

Each student draws an individual random data set of 400 observations from a randomly drawn class. You use the following code, where you change 0012345 by your student number.

```
set.seed(0012345)
rice_dataset <- read.csv("Rice_Dataset.csv")
mygroup <- which(rmultinom(1, 1, rep(1/4, 4)) == 1)
mysample <- sample(which(rice_dataset$CLASS == mygroup), 400)
mydata <- data.frame(rice_dataset[mysample, 1:10])
```

You answer the questions by performing an appropriate analysis with `R`. Use the provided R Markdown template to add your code and interpretation. Only report results and interpretations, do not repeat theory from the course!

Your R Markdown file should be named "LastName_FirstName_Project1.Rmd" and uploaded on Toledo on **May 3, 2024** (23h) at the latest. This project is graded on 4 points.

1. Consider the **full** data set (400 observations), perform an exploratory analysis via histograms and scatter plots. Briefly report your main findings.

2. Study the normality of `Eccentricity`, `Solidity` and `Aspect Ratio`. If the distribution of a variable deviates highly from a normal distribution, try to transform it such that its distribution becomes closer to a normal distribution. For the remaining assignments, use the transformed variables if you have transformed any.

3. Make a scatter plot of **Perimeter** and **Minor Axis**. Add the 99% tolerance ellipse based on the classical mean and the classical covariance matrix. Also add the MCD-based tolerance ellipses for the MCD estimator with 50% and 25% breakdown value. Discuss the results, and also compare the robust distances with the Mahalanobis distances.

4. Split your data set into a training data set and a validation set:

   ```
   mytrainingdata <- mydata[1:200, ]
   myvalidationdata <- mydata[201:400, ]
   ```

   Perform a robust PCA on your **training** data set. Argue why you have to standardize your data or not, and how you select the number of principal components. Discuss the outlier map. Remove the orthogonal outliers and the bad leverage points from your training data set.

5. Perform a PCA analysis on your reduced training data set of size $m \leq 200$. Explain how you choose the number of components.

6. Make biplots of the first three scores (or only the first two if you have kept two components). Interpret your results.

7. Compute the root mean squared error:

$$\text{RMSE} = \sqrt{\frac{1}{m}\sum_{i=1}^{m} \|\boldsymbol{x}_i - \hat{\boldsymbol{x}}_i\|^2}$$

   where $\boldsymbol{x}_i$ is the ith observation from the training data and $\hat{\boldsymbol{x}}_i$ its predicted value. Note that this needs to be computed on the standardized data (in case you standardized your data in the PCA analysis). Also compute the median error:

$$\text{ME} = \text{median}_i(\|\boldsymbol{x}_i - \hat{\boldsymbol{x}}_i\|)$$

   Next, compute the RMSE and ME for the validation data. Each $\boldsymbol{x}_i$ then corresponds with one of the 200 observations from the validation set. The predicted values are their fitted values with respect to the PCA subspace computed on the **training** data (so don't apply PCA to the validation data set!).

   Compare and discuss the results.

Good luck!