Exam: 2021/06/18

Marco Scarpelli

05 giugno, 2024

Dataset exploration

```
##       Area Perimeter MajorAxisLength MinorAxisLength Eccentricity ConvexArea
## 1 37789.00    766.37          313.57          154.35         0.87   38251.02
## 2 56627.02    922.88          356.20          203.05         0.84   57280.00
## 3 35139.01    734.02          292.62          154.73         0.86   35769.00
## 4 55971.99    897.39          342.49          208.79         0.81   56562.98
## 5 44924.00    817.21          282.14          203.15         0.69   45772.01
## 6 59182.99    934.85          353.58          214.65         0.79   59939.00
##   EquivDiameter Roundness       Type
## 1        219.35      0.81 cannellini
## 2        268.51      0.83     adzuki
## 3        211.52      0.83 cannellini
## 4        266.96      0.87     adzuki
## 5        239.16      0.85 black-eyed
## 6        274.50      0.85     adzuki
```

Removing non-quantitative variables:

```
##       Area Perimeter MajorAxisLength MinorAxisLength Eccentricity ConvexArea
## 1 37789.00    766.37          313.57          154.35         0.87   38251.02
## 2 56627.02    922.88          356.20          203.05         0.84   57280.00
## 3 35139.01    734.02          292.62          154.73         0.86   35769.00
## 4 55971.99    897.39          342.49          208.79         0.81   56562.98
## 5 44924.00    817.21          282.14          203.15         0.69   45772.01
## 6 59182.99    934.85          353.58          214.65         0.79   59939.00
##   EquivDiameter Roundness
## 1        219.35      0.81
## 2        268.51      0.83
## 3        211.52      0.83
## 4        266.96      0.87
## 5        239.16      0.85
## 6        274.50      0.85
```

Point A

Already from the dataset we can see that different variables have different magnitudes, and by a lot. We will now scale the data:

```
##            Area  Perimeter MajorAxisLength MinorAxisLength Eccentricity
## [1,] -1.0333341 -0.8685604     -0.02418799      -1.5137551   1.37974029
## [2,]  1.1573716  1.1239232      1.42756841       0.5995028   0.84310837
## [3,] -1.3415060 -1.2803989     -0.73763619      -1.4972656   1.20086299
```

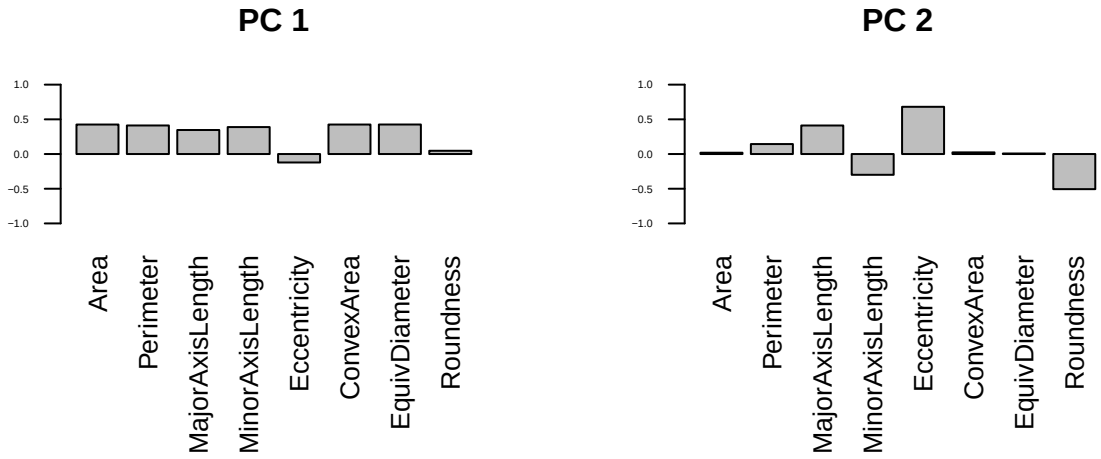
```
## [4,]  1.0811970  0.7994173      0.96067701       0.8485808   0.30647645
## [5,] -0.2035927 -0.2213310     -1.09453057       0.6038421  -1.84005123
## [6,]  1.4546097  1.2763098      1.33834482       1.1028660  -0.05127816
##      ConvexArea EquivDiameter  Roundness
## [1,] -1.0475103    -1.0273023 -0.5931994
## [2,]  1.1252916     1.1335146 -0.1333549
## [3,] -1.3309169    -1.3714682 -0.1333549
## [4,]  1.0434195     1.0653847  0.7863342
## [5,] -0.1887348    -0.1565581  0.3264896
## [6,]  1.4289065     1.3968037  0.3264896
```

Point B

We perform the PCA and report the results:

```
## Importance of components:
##                            Comp.1    Comp.2    Comp.3      Comp.4       Comp.5
## Standard deviation     2.3401911 1.2685723 0.9117951 0.155076229 0.0634409762
## Proportion of Variance 0.6891562 0.2025095 0.1046188 0.003026255 0.0005064712
## Cumulative Proportion  0.6891562 0.8916657 0.9962845 0.999310728 0.9998171997
##                              Comp.6       Comp.7       Comp.8
## Standard deviation     0.0300192097 2.087759e-02 1.075298e-02
## Proportion of Variance 0.0001134001 5.484989e-05 1.455033e-05
## Cumulative Proportion  0.9999305998 9.999854e-01 1.000000e+00
```

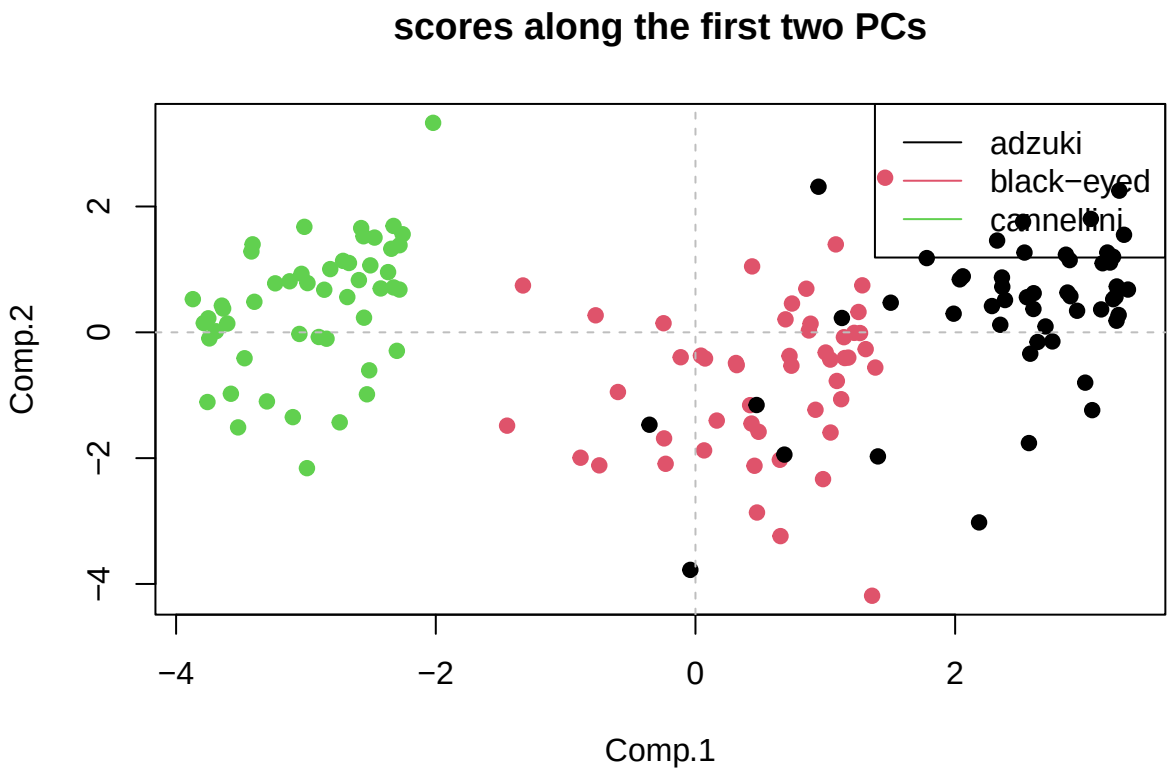
and plot:



The first component seems to be related to the general measurements of the grain of rice, i.e. the component goes in the direction of how big the grain of rice is.

The second component seems to be related to how the grain of rice is ellyptical (or oblong), i.e. it is the contrast between features related to more rounded shapes (i.e. roundess) and more slender ones (i.e. AxisLength and Eccentricity). In fact, we note that contrasted "real-world" features are the opposite of one another, like `MajorAxisLength` and `MinorAxisLength`, and `Eccentricity` and `Roundness`.

We report the scatter plot:



We can see that beans of "Cannellini" type have a low score on the first PC: this means that their dimension is generally smaller than that of the others. Then, beans of the "adzuki" type are generally the biggest even if they do not differ too much from "black-eyed" beans. The three types of beans are generally centered around 0 for the second PC, but the "adzuki" ones tend to be a little higher on average: we can then expect that generally, this type of bean will exhibit more oblong shapes.

Point C

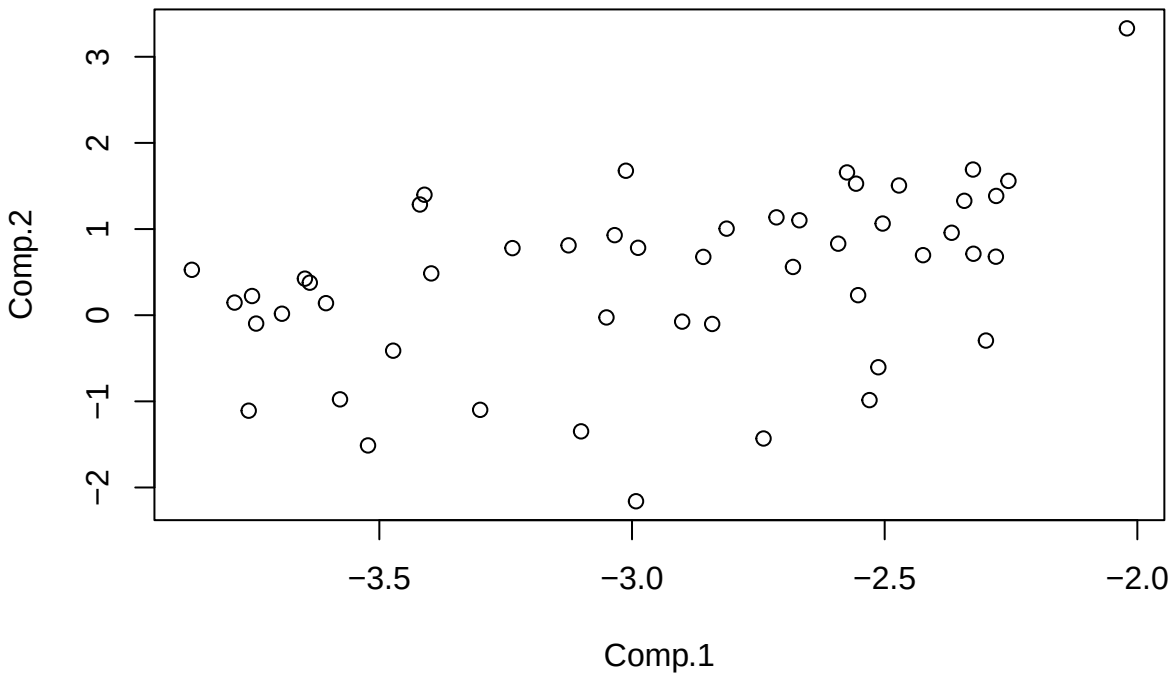
The expression for this ellipse is:

$$\left\{ m \in R^2 \middle| n \left(\overline{X} - m\right)^T \mathcal{S}^{-1} \left(\overline{X} - m\right) < F^* \right\}$$

with F^* equal to:

$$\frac{(n-1)p}{n-p} F(1-\alpha, p, n-p)$$

We check the normality of the resulting data:




```
##            Test       HZ     p value MVN
## 1 Henze-Zirkler 1.365521 0.002290855  NO
```

The resulting data is not Gaussian! It seems that it comes from a uniform distribution.

```
##               inf        center       sup
## Comp.1 -0.4756606 -7.114679e-17 0.4756606
## Comp.2 -0.2578464 -3.608225e-17 0.2578464
```