Politecnico di Milano
Scuola di Ingegneria Industriale e dell'Informazione
Applied Statistics
June 18th, 2021

Problem n.2

Images of 300 grains of 3 different varieties of dry beans were taken with a high-resolution camera. Through a feature extraction software, 8 dimensions were obtained from the grains: area, perimeter, major axis length, minor axis length, eccentricity, convex area (area of the smallest convex polygon that contains the bean), equivalent diameter (diameter of a circle having the same area as the bean) and roundness (calculated with the formula $4\pi\text{area}/\text{perimeter}^2$). The file `beans.txt` collects the data obtained and the indication of the variety of the beans.

a) Perform a Principal Component Analysis of the dataset, by only focusing on the quantitative variables of the dataset; here, evaluate whether it is appropriate to use the original variables or the standardized ones and proceed accordingly.

b) Report a plot of the loadings of the first two principal components and provide an interpretation. Report the scatter plot of the data along the first two PCs. Use the categorical variable `variety` to interpret the results.

c) Considering only the variety cannellini, build a confidence region (level 95%) for the mean of the vector whose components are the first two principal components. Characterize the region by reporting its mathematical expression, its center, the direction of the axes and the length of the semi-axes and provide a plot. Introduce and test the hypothesis of Gaussianity.

Upload your results here:
`https://forms.office.com/Pages/ResponsePage.aspx?id=K3EXCvNtXUKAjjCd8ope612LHtvIHvFEsEi2L6mhPglURFpCRDNJMVpWUTMxQjE3QjZPUjZRQVBXRC4u`