MULTIVARIATE ANALYSIS TECHNIQUES FOR CLASSIFING WHEAT SEEDS

*Chiara Barbi*

*Geoffrey Greene*

20 / 05 / 2019

1. INTRODUCTION\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 3
2. **DATASET\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 3**
3. **METHODS\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_4**
   1. **Principal Component Analysis**
   2. **MANOVA**
   3. **Linear Discriminant Analysis**
4. **DATA ANALYSIS\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_6**
5. **CONCLUSION\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_17**
6. **APPENDIX\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_18**
7. **BIBLIOGRAPHY \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

1. **INTRODUCTION**

The objective of this analysis is to describe and classify three different species of wheat by applying multivariate analysis techniques to a dataset that takes into account the geometric properties of the seed.

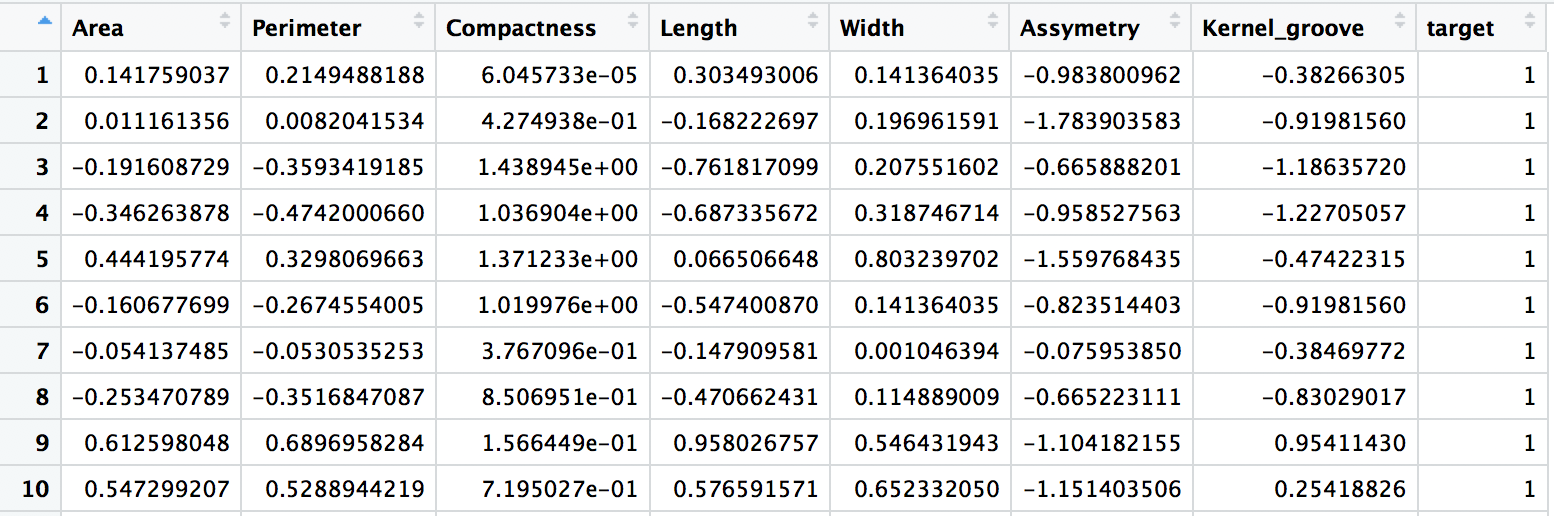
1. **DATASET**

The dataset used in this study contains measurements of geometrical properties of kernels belonging to three different varieties of wheat: *Kama, Rosa* and *Canadian*.

High quality visualization of the internal kernel structure was detected using a soft X-ray technique. It is non-destructive and considerably cheaper than other more sophisticated imaging techniques like scanning microscopy or laser technology. The images were recorded on 13x18 cm X-ray KODAK plates. Studies were conducted using combine harvested wheat grain originating from experimental fields, explored at the Institute of Agrophysics of the Polish Academy of Sciences in Lublin.

The Dataset contains 210 observations and 7 variables representing the various value of the properties of the wheat seeds:  
  
1. area A  
2. perimeter P  
3. compactness C =   
4. length of kernel  
5. width of kernel   
6. asymmetry coefficient   
7. length of kernel groove  
8. Target: one attribute variable differentiating varieties of wheat: Kama (1), Rosa(2), and Canadian(3).

The firsts 10 lines of the data are shown in the following picture.



1. **METHODS**

In the analysis we are going to apply Principal Component Analysis, to preliminary explore the data. Then by using Multivariate Statistical Inference the comparison of groups with hypothesis testing will be done. Eventually, a model for classifying new observations is performed through a Discriminant Analysis technique.

* 1. *Principal Component Analysis*

Principal Component Analysis (PCA) is a multivariate data analysis technique whose main objectives are:

* Data reduction
* Interpretation

Principal components are particular linear combinations of the p random variables present in the dataset. Geometrically, these linear combinations represent the selection of a new coordinate system obtained by rotating the original system with X1, X2,…. , Xp as the coordinate axis. The new axes represent the direction with maximum variability and provide a simpler and more parsimonious description of the covariance structure. The Principal Components are orthogonal to each other and the first principal component is the linear combination with maximum variance.

The principal components are obtained from the Singular Value Decomposition (SVD) of the covariance (or correlation) matrix: indeed, the PC’s correspond to the eigenvectors of the covariance/correlation matrix while eigenvalues measure variance along the principal components.

* 1. *Multivariate Inference: MANOVA*

MANOVA stands for Multivariate Analysis of Variance. It is an extension of the ANOVA, where we examine for statistical differences on one continuous dependent variable with an independent grouping variable. In the multivariate framework, we have one (or more) categorical independent variables (in our case one, that is “type of wheat”) and more than one continuous dependent variable. While ANOVA tests for the difference in means between two or more groups, MANOVA tests for the difference in two or more vectors of means. This technique tests the null hypothesis that the means of the groups of observations are identical.

Manova is a rotational technique designed to maximize variance between groups rather than across an entire data set (like Principal Component Analysis does).

* 1. *Linear Discriminant Analysis*

Linear Discriminant Analysis is a method used in statistics, pattern recognition and machine learning to find a linear combination of features which characterizes or separates two or more classes of objects or events. The resulting combination may be used as a linear classifier or for dimensionality reduction before later classification.

The general idea of LDA is to project the original data matrix onto a lower dimensional space (similiarly to PCA) but in this context the axis will be the direction of maximal separability between clusters.

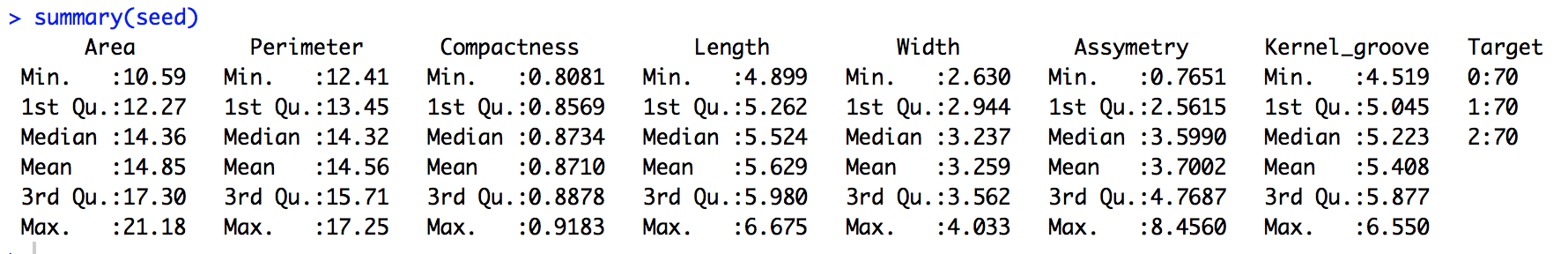
In our study we will use LDA as a classifier (i.e., given a new observation we would like to be able to classify it to one of the existing clusters).

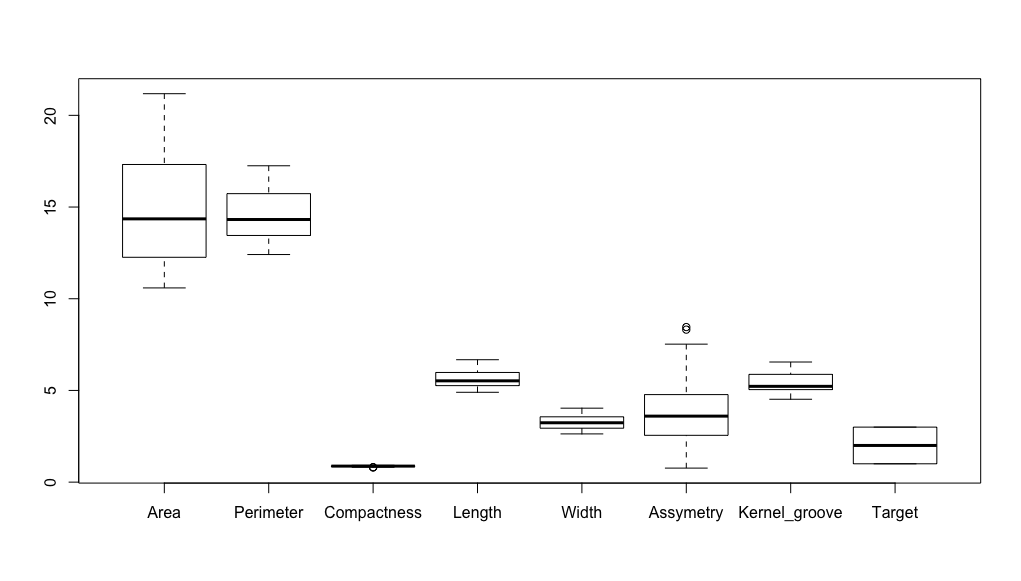
The objective will be to minimize the probability of wrong classification, which is equivalent to maximizing the likelihood that the observation will be assigned to the correct cluster.

1. **DATA ANALYSIS**

4.1 *Exploratory Data Analysis*

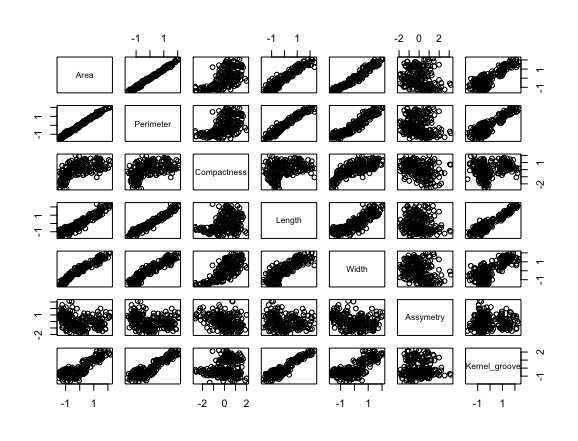
First we briefly explore the dataset by doing an Exploratory data analysis.

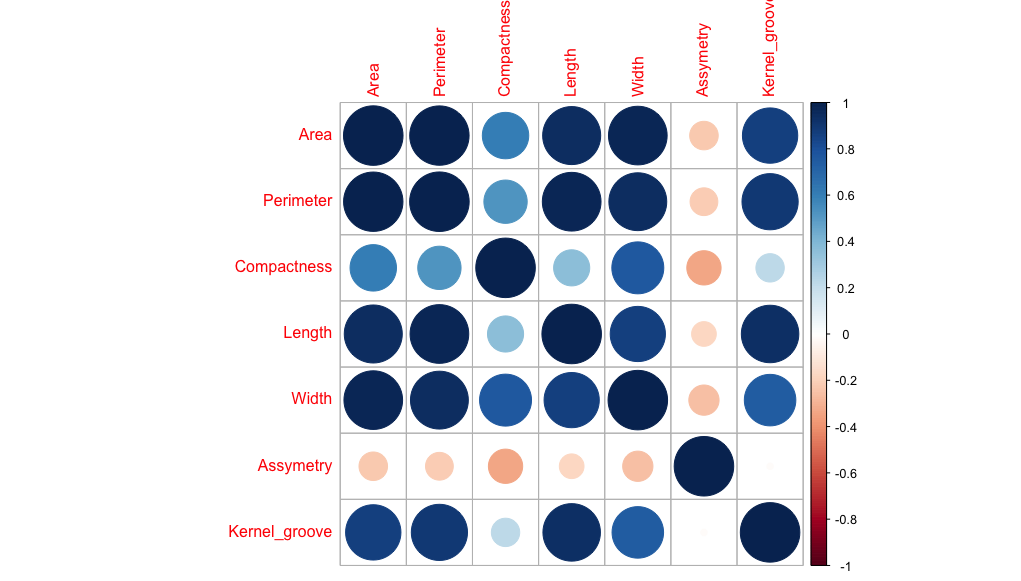




From this boxplot it is clear that there are differences in the measurement scale between many attributes, the scale for the compactness variable is particularly small compared to the others, so it is necessary to standardize the data before the analysis. (for this purpose we used the R function scale(seeds[ , -8], center = TRUE, scale = TRUE).

To investigate the covariance structure, a pairs plot and a correlation plot are performed:





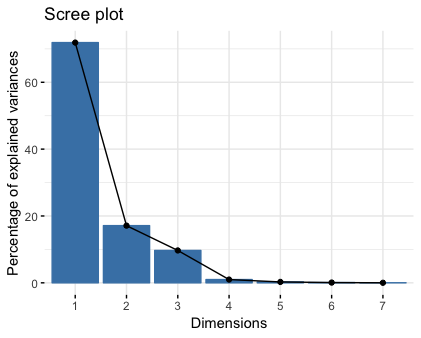
By looking at both of these plots it is clear that a lot of variables are highly, positively correlated. This suggests us that a dimensional reduction of the data could be powerful since it would eliminate redundancy.

* 1. *Principal Component Analysis*

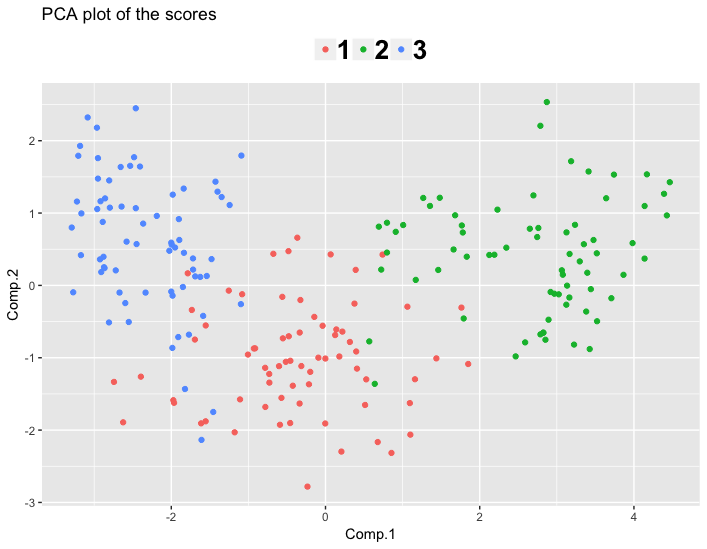
Using the function *princomp* in R we can easily perform a principal component analysis and access the results and the plots.



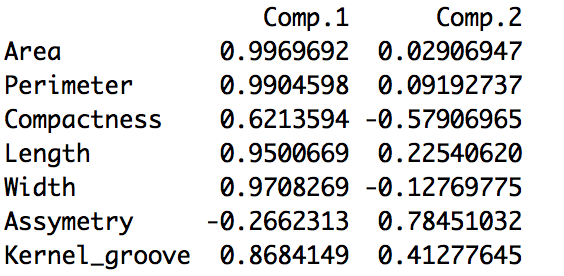
The firsts two components account for 88.98 % of the variance in the data, hence they are enough to properly describe and summarize our dataset. Moreover, the screeplot confirms this intuition. In fact the appropriate number of components is taken to be the point at which the remaining eigenvalues are relatively small and all about the same size (i. e. in our case this point correspond to dimension 2).



Since one of the main purposes of Principal Component Analysis is to represent the data in a low dimensional space and to investigate its structure, a couple of biplots are produced and the correlation between the variables and the principal components is studied.



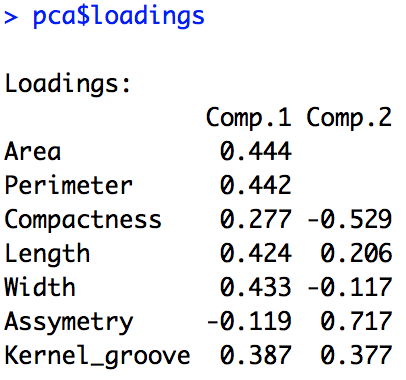
From this plot we can clearly see the separation between the three different varieties of wheat, plotted in three different colors. To characterize the clusters and the axes of this plot we need to understand the correlation between the original variables and the principal components.



The first Principal Component is highly, positively correlated with Area, Perimeter, Length, Width and Kernel\_groove while the second one is highly positively correlated with Asymmetry and negatively correlated with Compactness.

It seems that the first PC mostly takes into account the quantitative geometrical features of the seed, while the second one is somewhat concerned with the overall appearance of the wheat (Asymmetry and Compactness are more general attributes, and also more qualitative than the other variables).

To confirm this assumption, let’s look at the Loadings of the first two principal components:

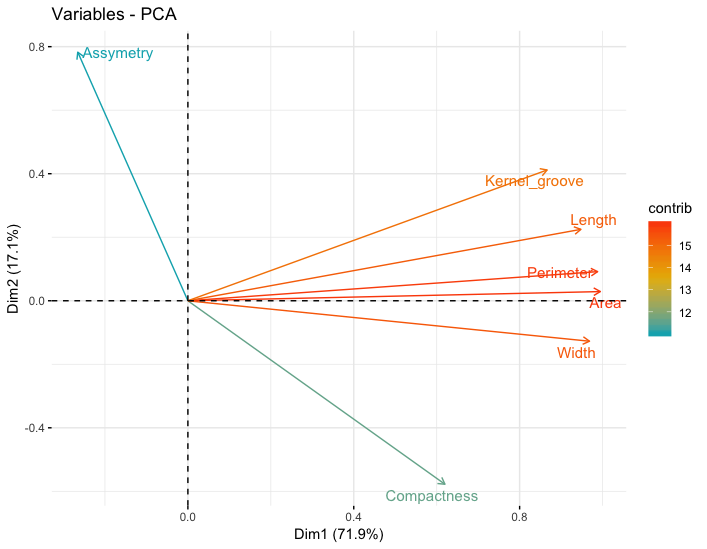


As already stated, comp.1 is like an “average” of the quantitative geometric features of the seed and hence in the plot with respect to the first axis we can distinguish between the seeds that are big (the rightmost points) and small (the leftmost points). Hence, the seeds belonging to the Canadian type (Group 3) are smaller than the ones belonging to the Rosa variety (Group 2).

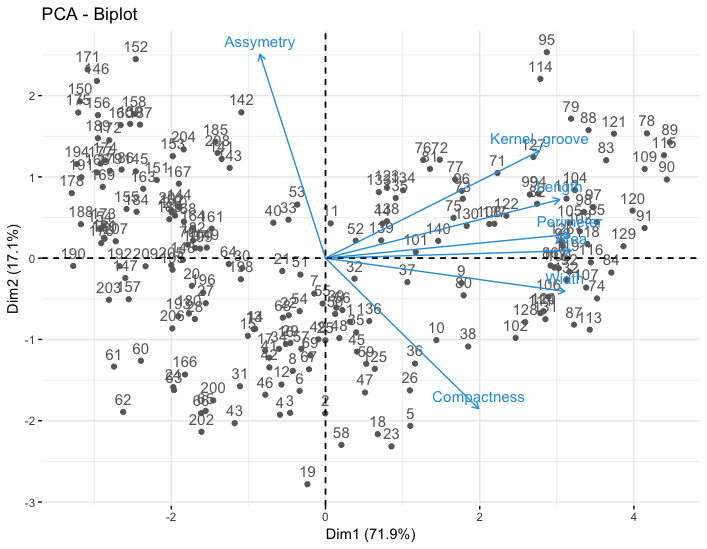
On the other hand, the second component is determined by Assymetry, and also by Compactness to a lesser extent. The loadings indicate that at the top of the plot we can find the seed characterized by high asymmetry and low compactness (which seems to be reasonable since asymmetry and compactness are intuitively inversely proportional). The Kama group is characterized as having higher symmetry and compactness than the others.

To visualize all these assumptions, a biplot of the variables expressed with respect to the principal components is produced.

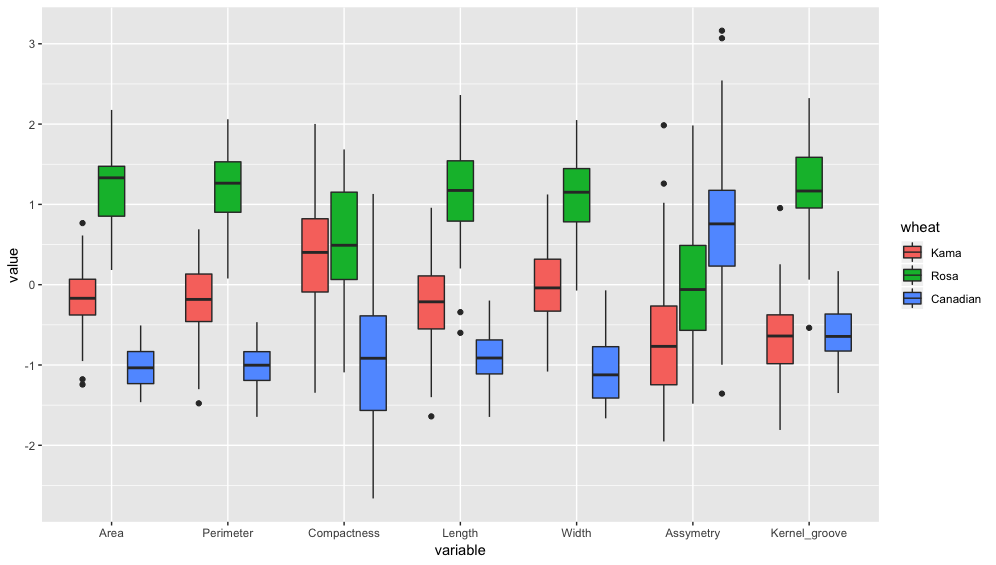
As already stated, the first principal component (which is the one that accounts for the majority of the variance, 71.9% ) takes into account the length, the perimeter, the area and the width of the seeds. Note that these variables are highly correlated with each other, which is visible from the small angles between their vector representations in the plot (this is intuitive, since these variables express connected geometrical features). On the other hand, we can see that Assymetry and Compactness are inversely proportional, due to the fact that the angle between the two arrows is almost 180 degrees.



To conclude this section, a biplot with both the scores and the loadings is produced, graphically summarizing the patterns that have been previously exposed.



*4.2 MANOVA*

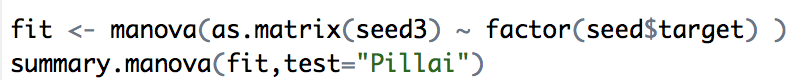
As stated in “Methods”, in this section we want to test for differences between groups. First, the following picture shows the boxplots for every group with respect to each variable. It seems reasonable to assume that there are significant differences between the groups.

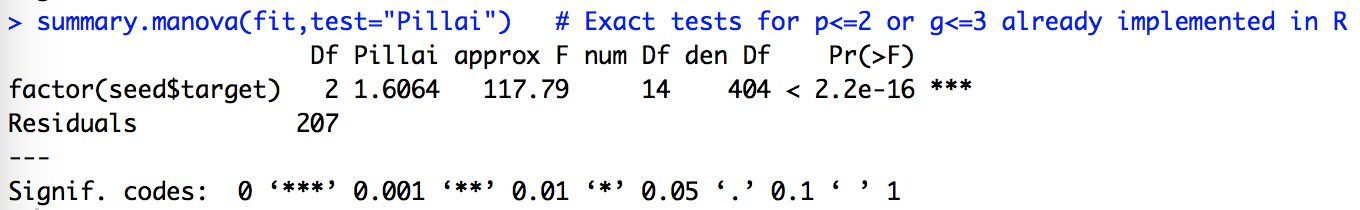
In our problem we have 1 grouping variable on g = 3 levels (“target”) and p = 7 variables.

Basically, the test we want to compute is the following:

* H0: The membership to a wheat species hasn't any significant effect on the mean of (in any direction of
* H1: There exists at least one direction in along which at least two species have some feature significantly different.

The R function to perform Manova is the following (in the stats package):

**



Since the p-value of the test is very low, the null hypothesis is refused and it is stated that there are significant differences between the means. Although the test is rejected, we can test each variable individually with an ANOVA test. The aov() function can output tests on individual variables when wrapped in a summary() call. The output of this function is reported in the appendix and it shows that there are significant differences in the means for all the variables.

Note that to consider this test valid, it is necessary for the following two hypotheses be verified:

* Dependent variables are multivariate normally distributed within each group of the independent variable (in our case “type of wheat”)
* The population covariance matrices of each group are equal.

The check of this assumption is in the appendix.

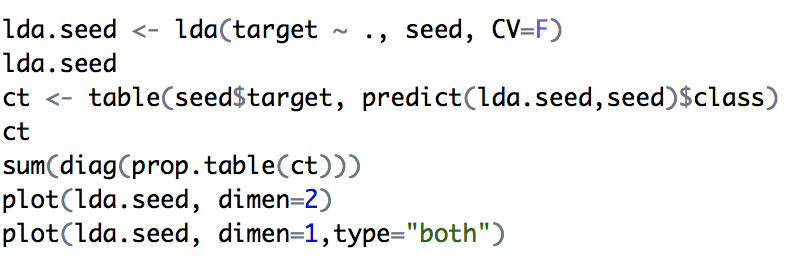
* 1. *Linear Discriminant Analysis*

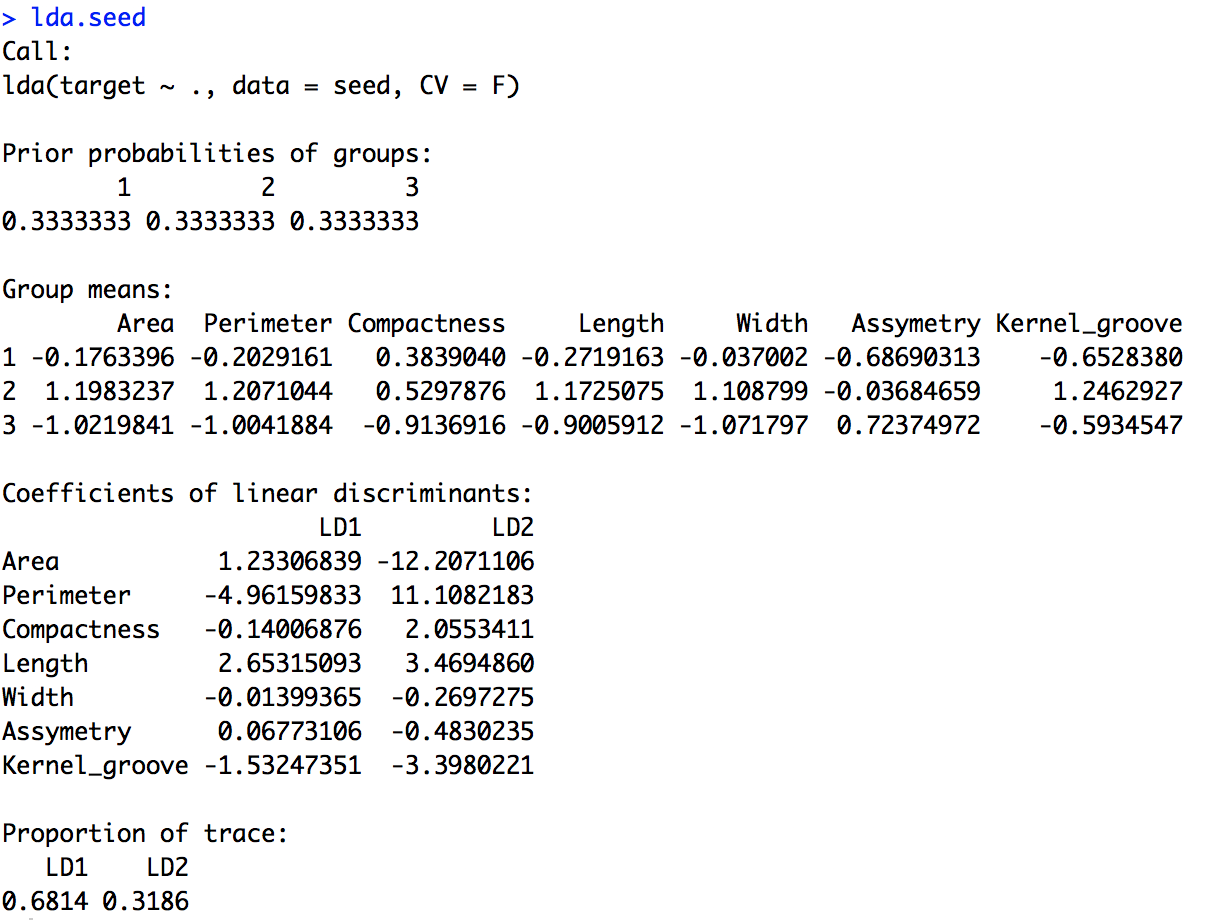
Unlike what happens with MANOVA, for Linear Discriminant analysis it is not needed for the variables to follow a Multivariate Normal Distribution. Moreover, the method is still robust even if the covariance matrices of the variables are not exactly equal.

The guideline of this section will be:

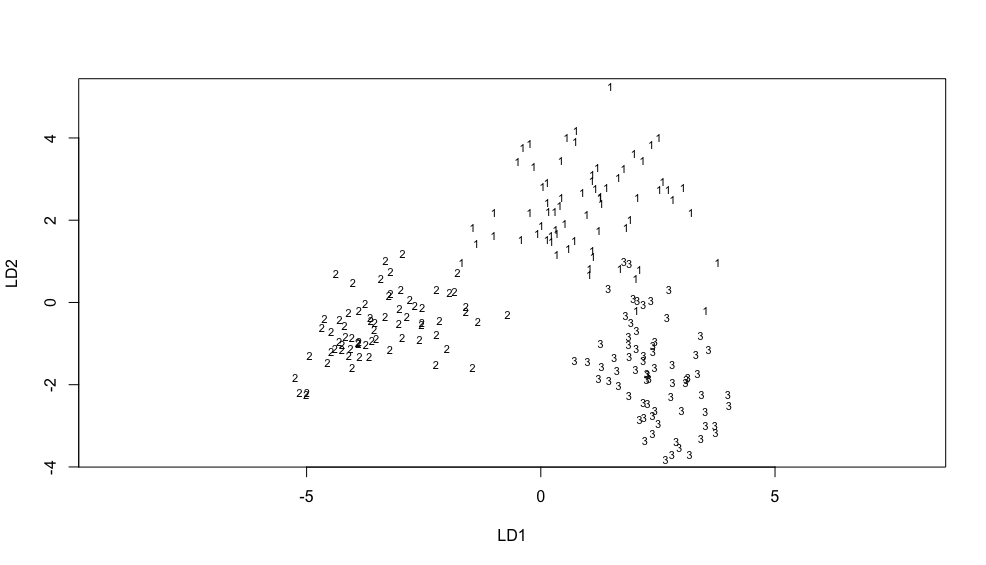
* Using the R function “lda”, present in the package MASS, to construct the model, first without using Leave-one-out Cross Validation, and then adding this option to the function’s parameters.
* Even if our main goal is to classify and not data-visualization, a plot of the partition induced by the LDA will be done.
* Then we will assess the accuracy of the prediction by computing the table of classification. (for both the model with and without Cross Validation)

The following code performs the Linear Discriminant Analysis without Cross Validation.

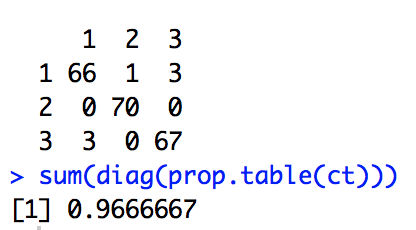




The prior probabilities in this output are calculated as the proportion of each variety of wheat in each cluster. In our case the three different kinds of wheat are equally distributed between the three clusters.

The coefficient of linear discriminants provides the equations for the discriminant function. Indeed, Discriminant analysis works by creating one or more linear combinations of predictors, creating a new latent variable for each function, which are called discriminant functions. The number of functions possible is either *Ng*-1 where *Ng* = number of groups, or *p* (the number of predictors), whichever is smaller (in our case in fact we have 2 functions since Ng=3). The first function created maximizes the differences between groups on that function. The second function maximizes differences on that function, but also must not be correlated with the previous function. Similar to what happens in PCA, we can plot in a reduced space the data and see how each observation is classified by the model. 

For this model we get the following table of classification:

 On the diagonal there is the number of correct classified. The accuracy (or the percentage of correct classified) is 0.97. w

It is important to notice that this table is optimistic one: indeed, due to the fact that we set CV=F, to predict to which category an observation belongs, a discriminant constructed using the same observation was used. intuitively, this method produces an excessively optimistic result.

To overcome this problem, a LDA using leave-one-out cross validation is performed. For each observation, the discriminant is calculated without considering the observation in question. Hence, the following forecast will be more consistent.

In this study the table obtained using the loo cross validation is the same as above, but in general this is not ensured.

1. **CONCLUSIONS**

From the Principal Component Analysis we learned that overall the largest seed type is Rosa, followed by Kama, and the smallest is Canadian. We also saw that Kama has higher compactness and symmetry than the other two types.

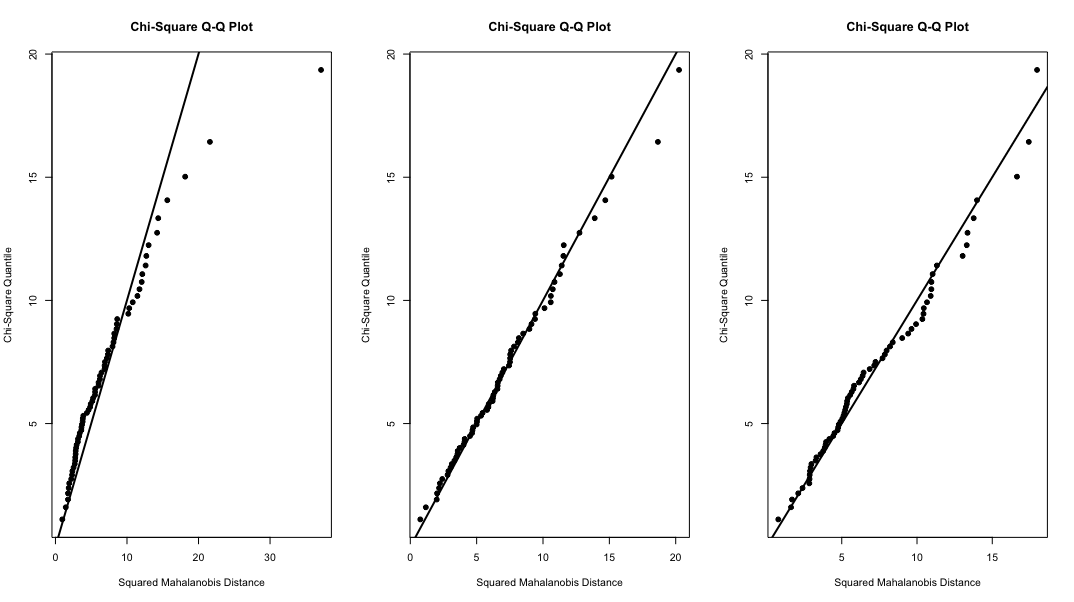
The boxplots that we produced illustrate a clear distinction between the three species with respect to the geometric properties Area, Perimeter, Length and Width. None of the boxes overlap for any of these variables, indicating a significant difference between the groups with respect to seed size. The differences in the variables Compactness, Asymmetry and Kernel Groove between certain groups are not as clear cut as in the other variables, but we then confirmed with the ANOVA test that there are indeed significant differences between the groups.

1. **APPENDIX**

6.1 Manova Hypothesis checking.

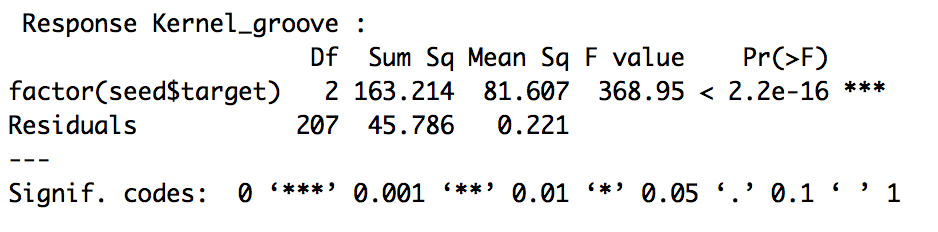
By using the function mvn(group,mvnTest="mardia",multivariatePlot="qq") we can access the results for the multinormality. In the three groups the assumption is not perfectly verified, as can be seen in the qq plot and in the quantitative results of the test (not showed here). In particular, with respect to the “Kurtosis Mardia” test, all the groups but (1) Kama. seems to be multinormal. As further study, outlier detection and data transformation may be considered, such as box-cox transformation to obtain more “normal shape” of the distributions.

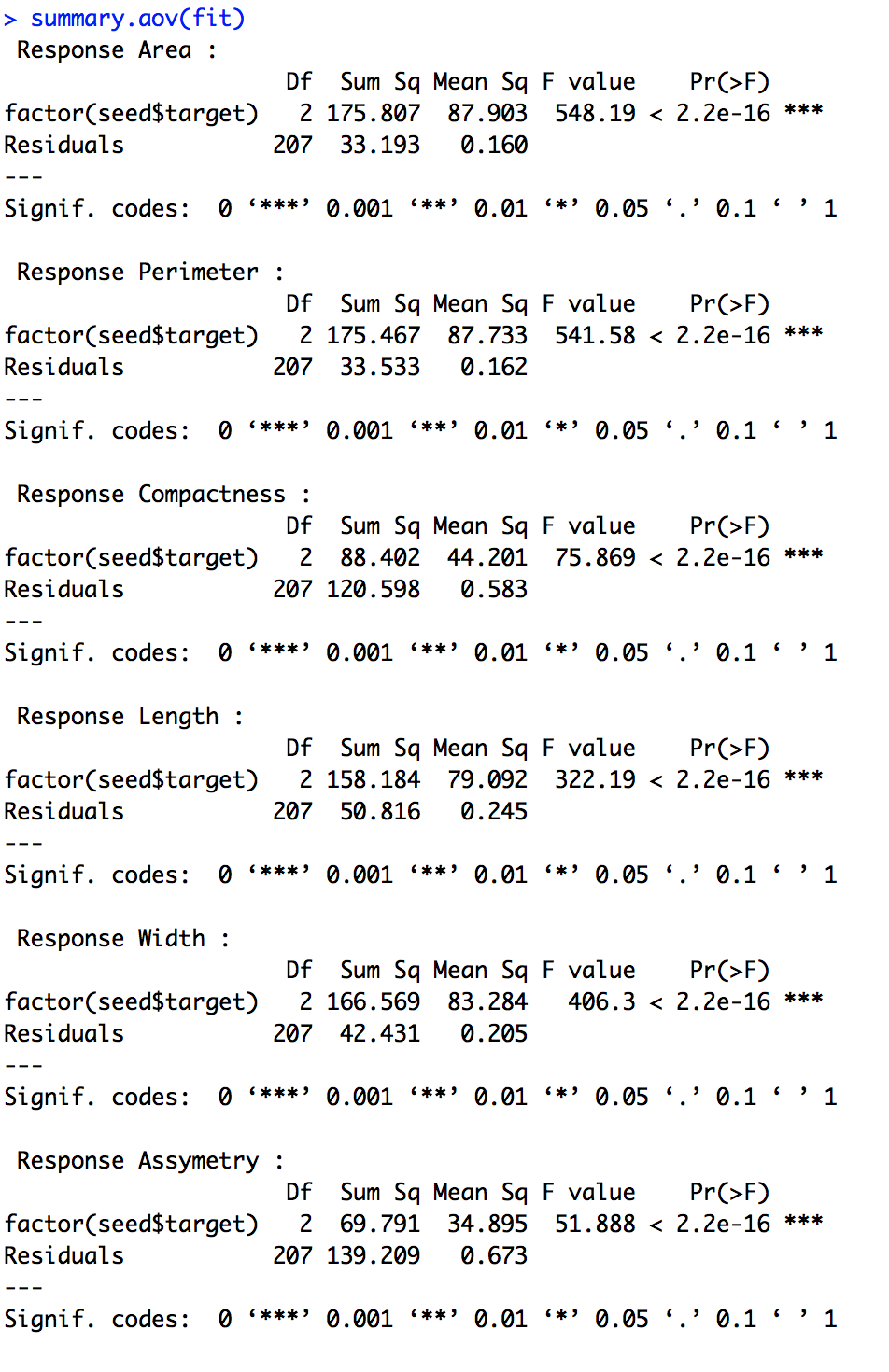
As far as the covariance homogeneity is concerned, we only checked it graphically, by using the function image(), and got these results:





For groups 1 and 2 the hypothesis seems to be verified, while group 3 seems to deviate slightly.

* 1. Anova output for each variable:



1. **BIBLIOGRAHY**

# *Applied Multivariate Statistical Analysis, 6th Edition, Richard A. Johnson, Dean W. Wichern, 2008, Pearson.*

*Multivariate analysis / K. V. Mardia, J. T. Kent, J. M. Bibby*