Computer Lab Practical n. 5

Today's task

Examples of use of principal components and cluster analysis.

How to start

• Access the course web page on Moodle (/Biology). Download and save the two data files protein.rda and vini.rda.
• **Download also the PDF of Lab. 5** so as to be able to "copy and paste" some of the instructions which will be given below.
• Once prompted by R, type `library(Rcmdr)`.

How was European food in 1973?

• Data ⇛ Load data set
Select the protein.rda data file.
• Inspect the data. The dataset provides protein consumption of various countries divided by the sources of the same. The data were collected in 1973... which explains why some countries no longer exist! The names of the variables should be self-explanatory. Starch ("amido", in Italian) represents largely potatoes. Nuts also includes oils and legumes (and, indeed, is composed almost exclusively of these, not of "nuts").
• Visualize the pairs plot.
Graphs ⇛ Scatterplot matrix
(Omit the nonlinear smoother.) *Your comments?*
• Statistics ⇛ Summaries ⇛ Correlation matrix
Select all variables; ask for pairwise p-values. *Your comments?*
• Statistics ⇛ Dimensional analysis ⇛ Principal-components analysis
Select all variables and check "Analysis based on correlation matrix", as we want to work with standardized variables. Further select "Screeplot" and "Add components". Once prompted, choose to "keep" all components (by moving the "slider" which pops up to its maximum).
Inspect the data. You will see five new variables called PC1, PC2, ..., and PC9. These are the 9 principal components (the one represented by a "Y" in the lecture notes).
Have a look at the screeplot. It shows how important every component is. *Your comments?*
Last but not least, have a look at the output. *Comment it.*

- R Commander does not provide a menu-driven interface for biplots (that is, for a scatterplot of the principal components enhanced by adding the original variables). We may plot it by typing the following instruction into the command line window. Use the following instruction to trace a biplot:

```
biplot( prcomp( protein[,1:9], scale=TRUE ) )
```

(Copy and paste it into the command line window and run it.)

- The above biplot shows the first two principal components. These alone explain between 60% and 70% of the total variability of the original 9 variables. So, not everything, but neither little.

Look at the biplot and answer the following questions:

(a) *With respect to the other countries, which was the most important source of protein in Romania in 1973? Which countries had similar eating habits?*

(b) *How does the diet of the Iberian Peninsula (Spain and Portugal) differ from the rest of Europe?*

(c) *Which are the countries where many proteins are "recovered" by eating meat (white or red) or other products derived from "terrestrial" animals (i.e., eggs and milk)?*

(d) *Is fish consumption more prominent in Norway or Ireland?*

About vines and vineyards

- Data ⇒ Load data set

Select the vini.rda data file.

- Inspect the data. This is the data set we discussed in class.

Actually, this is the "complete" version of the dataset, that is, the one which also includes the variable vintages which identifies the the grape variety. As in class, and it was the case, let us first suppose that we don't know the variety. Our goal is to see which groups get "detected" by the clustering algorithm we discussed in class.

- First of all, let us standardize the variables. Indeed, it's good practice to measure the "distances" between standardized variables.

Data ⇒ Manage variables in active data set ⇒ Standarize variables

Sellect all variables and click OK. Inspect again the data. The "new" variables Z.Original are the standardized version of the original variables called "Original".

- We standardized the dataset by subtracting from each variable its sample mean and dividing this difference by the mean squared error. The standardized variables will all have mean zero and unit variance. Let us check this!

Statistics ⇒ Summaries ⇒ Numerical summaries

Select all variables; in "Statistics" check only "mean" and "standard deviation".

Your comments?

Hierarchical clustering

- Statistics ⇒ Dimensional analysis ⇒ Cluster analysis ⇒ Hierarchical cluster analysis

Select only the standardized variables (those whose names start by Z.). Keep the remaining options as they stand. (We could use them to change a couple of aspects we mentioned during the lecture in by-passing.

Focus on the figure. How many groups are there?

- Let's try and find out in what the groups differ.

Statistics ⇒ Dimensional analysis ⇒ Cluster analysis ⇒ Summarize hierarchical clustering

Select the last "solution". Then ask to work with three clusters. Check "Print summary" and "Biplot".

Comment the biplot. Comment the results in the output window. Answer the following questions:

1. In what does group 1 differ from group 2?

2. How do these two groups differ from group 3?

- As already mentioned, the three groups turned out to identify the grape variety. Let us check this.

Statistics ⇒ Dimensional analysis ⇒ Cluster analysis ⇒ Add hierarchical clustering to data set

Select the last "solution" and work with three clusters. Inspect the data. Look for the new variable called hclus.label which assigns each observation to one of the three groups.

- Let us try and "cross" this variable with the grape variety.

Statistics ⇒ Contingency tables ⇒ Two-way table

Select as row variable hclus.label and as column variable vintages (or viceversa).

Comment upon the table.

Using K-Means

- Remember that you have to select the number of groups before using K-Means. The algorithm is rather fast and allows us to compare the solutions we get with different group numbers. How to choose the number of groups is beyond the scope of this course; let us nonetheless get a glimpse of it.
- Let us start of with two groups which will be formed by K-Means.

Statistics ⇒ Dimensional analysis ⇒ Cluster analysis ⇒ k-means cluster analysis

Select all standardized variables (those whose names start by Z.). Maintain "two clusters" and the remaining options unchanged.

Compare this new biplot with the one obtained by hierarchical clustering. What do you notice?

- The fraction of total variability explained by the presence of the two groups can be expressed as $A/(A+B)$, where A is the "Between Cluster Sum of Squares" and B is the "Total Within Sum of Squares". The two values A and B are provided in the output window. Hence, in the command window run:

```
645.87/(645.87+1642.13)
```

How much variability is explained by the two groups? (Use “copy and paste” to quickly copy something from the output window into the command window.)

- Repeat the analysis by specifying successively a number of 3, 4 and 5 clusters in K-Means. Check the corresponding biplots. Calculate every time the fraction of explained variability.

(You should get that about 45% of total variability is explained by the presence of three groups. That is, increasing the group number from 2 to 3 exploits about 26/27% of information. Successively increasing the group number to 4 and 5 increases the fraction of explained variability to 49% 52%, respectively.

Comment: Going beyond three groups doesn’t imply a high increase in explained variability. An increase of 5% could simply be due to chance. As it was the case for hierarchical clustering, the data pinpoint to the existence of 3 groups, not more.

- Recompute the solution for 3 groups. This time check “Assign clusters to the data set”. Inspect the data and look for the new variable KMeans. This variable assigns each observation to one of the three groups.

Comment the biplot and the numerical results in the output window.

- Compute, as before, the 2-way table using “KMeans” and “vintages”.

Which algorithm performs the best?