Deducit Manual

Deducit is a Latin word that means *"he drives"* or *"he brings to".* The software should drive your data from the real world to new theoretical worlds made by standard transformations. Although many transformations are possible, two are now implemented: Principal Components Analysis (PCA) and Partial Least Square (PLS) regression. There are many Applications that implement these possibilities among many other options, but the aim of ***Deducit*** is to do it very fast and easily. With just few mouse clicks you can move your Data Set in latent spaces and observe their properties with a complete set of graphs. In short, ***Deducit*** can give you PCA and PLS "*a la carte*"!

The limit of data transformation to only PCA and PLS may sound very strong for many data analysts. However, in my experience, there are a lot of things you may do, using these two tools only. Many examples are also given in the research activities of two mentors: Prof. John Mc Gregor (https://en.wikipedia.org/wiki/John_F._MacGregor) in the Multivariate Process Control and Prof. Riccardo Leardi (https://www.researchgate.net/profile/Riccardo_Leardi2) in chemometric field. A short selection of papers is reported in the Reference of the Manual.

Deducit is made as R-Cran derivative (https://cran.r-project.org/) implemented with the strength of Shiny Application (https://shiny.rstudio.com/). The easiness of developing statistical Web Applications is impressive and it takes really few lines of code to achieve surprising results comparing to alternative programming techniques.

So, the points of strength of ***Deducit*** are:

- It is a Web App, so you do not need to installing anything on your pc. The server will send you results.
- You can use with different Web Browsers (Google Chrome, Firefox etc.)
- It is platform independent. I developed on Windows 10 and tested in Linux Mint.
- It is fast with reasonably big Data Set (tested up to 500.000 elements)
- I push a lot of effort to document all the functionalities although they should be already quite evident.
- Statistical routines are base on R project and this guarantees robustness and reliability.

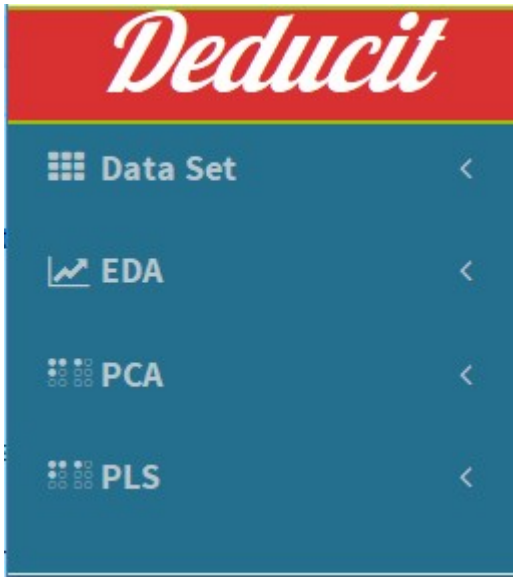
The weak points are instead:

- It is not a professional application, so many unusual situations may drive to some unexpected system errors. The application testing, like the modeling, is quite never complete!
- You have to send you data to the application server and this may raise concerns about security and confidentiality. However, as an open source software, you can download the source and use it in your own machine.
- The manual is complete but do not go deeper on theoretical aspects, at least for this initial version.
- Tools for Data Set management and plot customization are limited

All the above should be consistent with the aim of ***Deducit*** development. **The idea is to develop a software simple and friendly helpful to easily start looking at your Data Set.** The learning curve must be very limited and results should focalize only on the most important topics. ***Deducit*** should not be seen as an alternative to professional and commercial Applications, but a friendly tool to keep on hand.

A lot of work need to be done and I would like to focalize in Data preprocessing (missing values characterization and replacement, dynamic warping and trajectories alignment, etc.) and Graphical interpretation of Results (correlation among score and variables, results sensitivity to variable errors and uncertainty, etc.).

The ***Deducit*** structure is quite simple. It is made by four main chapters like shown in the right-side bar.

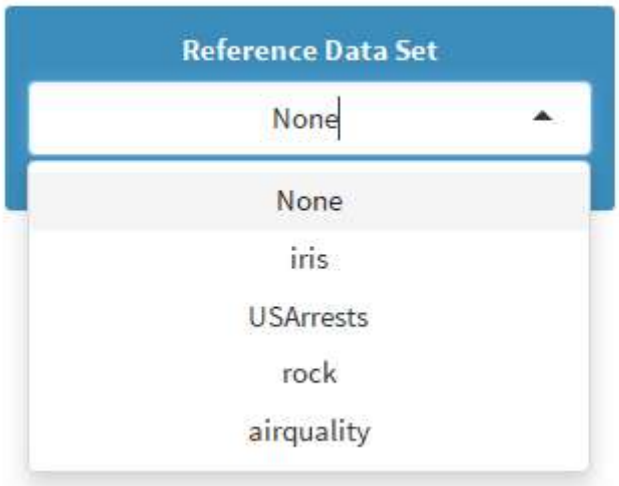


Data Set: here you can fin all the actions on the working Data Set (upload, see and set variables, objects, group, visualize statistics, etc.)	**EDA:** i.e. The Evolutionary Data Analysis where you can plot the Data Set in different ways (2D, 3D and univariate plots)
PCA: i.e. the Principal Component Analysis where you get the required tools for run your analysis and see results	**PLS:** i.e. The Partial Least Square regression where you can develop your multivariate correlations and test them

Each of the four chapters is detailed in this manual. You can find more or less the same information that are available in the software using the tag Help shown in each menu. Just click on it and scroll the pages. For instance, the example Dataset menu has its own Help file with the explanation of what it does.

Example Data Sets

The Data Set is the key element in data analysis. To see how ***Deducit*** works you need to have one loaded in the right format. To facilitate the learning process, I provide some of R standard Data Sets that are already in the right format. In the next menu you will see how to load your own Data Set.
To consider one of the examples, just scroll the combo box and pick up one of the choices. That's it. Data are loaded in the program and you can test the different functions. In the following there are some information about the data loaded as provided by the R-Cran documentation.



Iris - This famous (Fisher's or Anderson's) iris data set gives the measurements in centimeters of the variables sepal length and width and petal length and width, respectively, for 50 flowers from each of 3 species of iris. The species are *Iris setosa*, *versicolor*, and *virginica*. Iris is a data from with 150 cases (rows) and 5 variables (columns) named *Sepal.Length* (numeric), *Sepal.Width (numeric)*, *Petal.Length (numeric)*, *Petal.Width (numeric)* and *Species* (categorial).

USArrests - This data set contains statistics, in arrests per 100,000 residents for assault, murder, and rape in each of the 50 US states in 1973. Also given is the percent of the population living in urban areas. A data frame with 50 observations on 4 variables. *Murder* (numeric) Murder arrests (per 100,000); *Assault* (numeric) Assault arrests (per 100,000); *UrbanPop* (numeric) Percent urban population; *Rape* (numeric) Rape arrests (per 100,000).

rock - Measurements on 48 rock samples from a petroleum reservoir. *area* area of pores space, in pixels out of 256 by 256; *peri* perimeter in pixels; *shape* perimeter/sqrt(area); *perm* permeability in milli-Darcies.

Airquality - Daily air quality measurements in New York, May to September 1973. *Ozone* (numeric) Ozone (ppb); *Solar.R* (numeric) Solar R (lang); *Wind* (numeric) Wind (mph); *Temp* (numeric) temperature (degrees F); *Month* (numeric) Month (1-12); *Day* (numeric) Day of month (1-31)

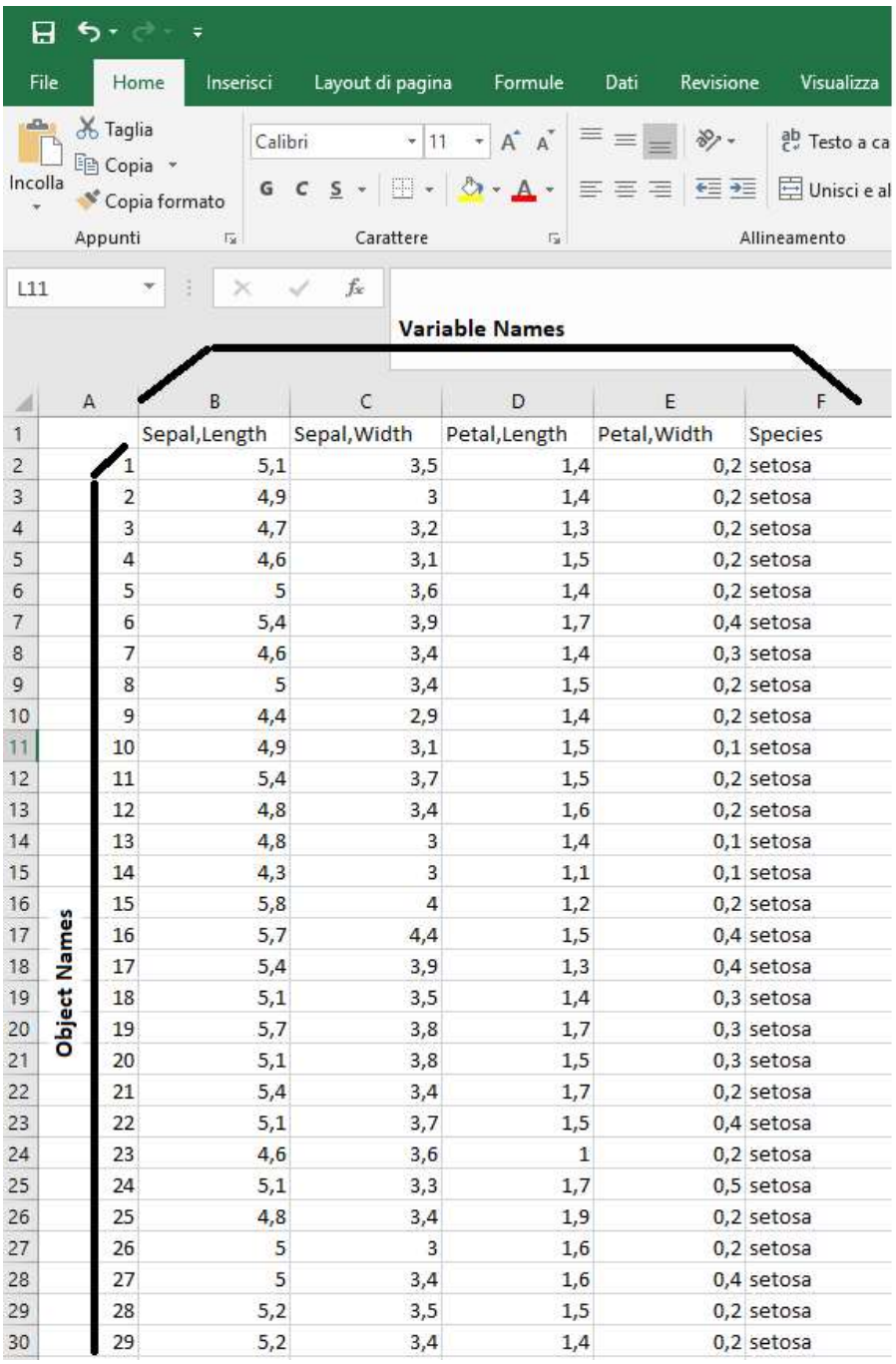
gasoline - NIR spectra and octane numbers of 60 gasoline samples. The NIR spectra were measured using diffuse reflectance as log(1/R) from 900 to 1700 nm in 2nm intervals, giving 401 wavelengths. *octane* (numeric) vector with octane numbers *nir900nm - nir1700nm* the NIR spectrum.

oliveoil - the data set scores 6 attributes from a sensory panel and measurements of 5 physico - chemical quality parameters on 16 olive oil samples. The sensory part is made by the 6 attributes: *yellow, green, brown, glossy, transp and syrup*. The chemical part is made by 5 measurements: *acidity, peroxide, K232, K270, and DK*.

Load CSV Files

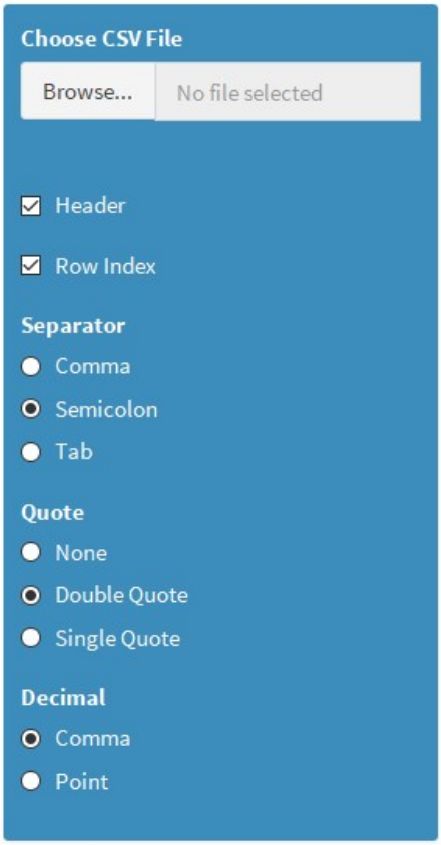
A **comma-separated values** (**CSV**) file is a delimited text file that uses a comma/semi colon to separate values. A CSV file stores tabular data (numbers and text) in plain text and represents a ***Data Set***. Each line of the file is a data record, i.e. a ***population object***. Each record consists of one or more field, i.e. ***factors or variables***, separated by commas in US standard or semi colons in EU standard. The use of the comma as a field separator is the source of the name, however csv is also used with semi colon and this causes many troubles in file import.

Csv file can be read by many software, although MS Excel is the most common. Usually, the name of the column is written as text in the first row, called ***Header***. Names of rows are in the first column as well (this causes an empty position in 1,1 place) with the name ***Row Index***. It is better to choose names short especially for rows because they are used to identify objects in plots and graphs. It is also better do not start names of variables with a number.

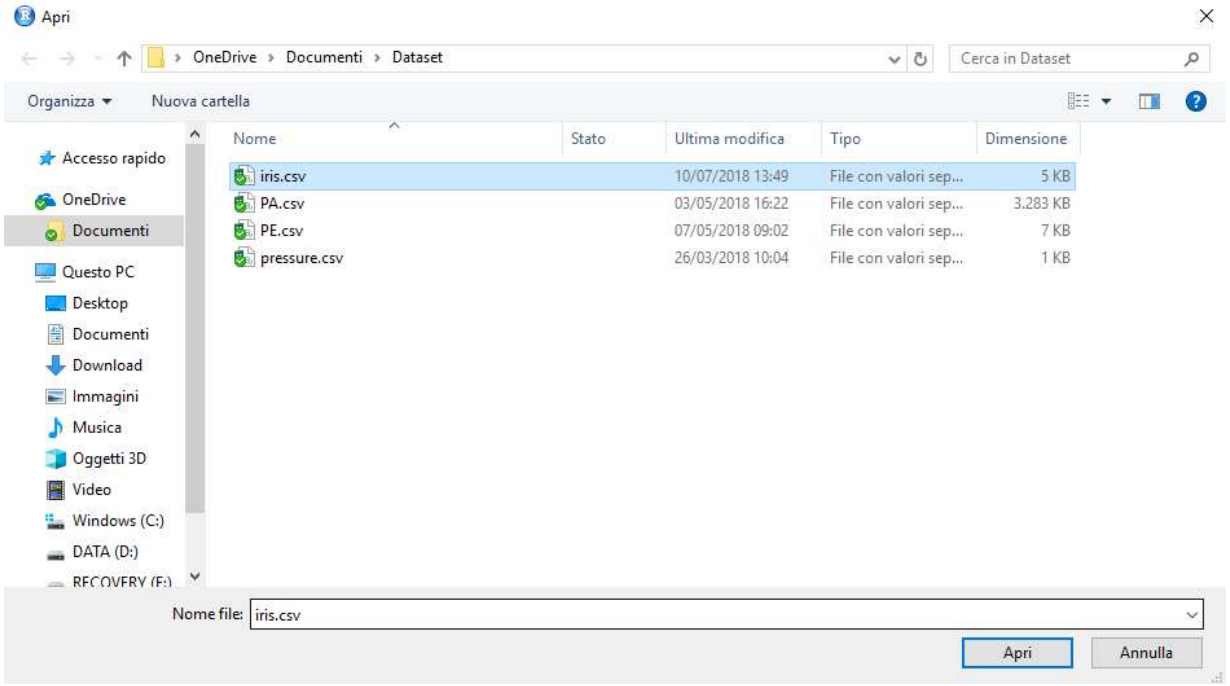


	A	B	C	D	E	F
1		Sepal,Length	Sepal,Width	Petal,Length	Petal,Width	Species
2	1	5,1	3,5	1,4	0,2	setosa
3	2	4,9	3	1,4	0,2	setosa
4	3	4,7	3,2	1,3	0,2	setosa
5	4	4,6	3,1	1,5	0,2	setosa
6	5	5	3,6	1,4	0,2	setosa
7	6	5,4	3,9	1,7	0,4	setosa
8	7	4,6	3,4	1,4	0,3	setosa
9	8	5	3,4	1,5	0,2	setosa
10	9	4,4	2,9	1,4	0,2	setosa
11	10	4,9	3,1	1,5	0,1	setosa
12	11	5,4	3,7	1,5	0,2	setosa
13	12	4,8	3,4	1,6	0,2	setosa
14	13	4,8	3	1,4	0,1	setosa
15	14	4,3	3	1,1	0,1	setosa
16	15	5,8	4	1,2	0,2	setosa
17	16	5,7	4,4	1,5	0,4	setosa
18	17	5,4	3,9	1,3	0,4	setosa
19	18	5,1	3,5	1,4	0,3	setosa
20	19	5,7	3,8	1,7	0,3	setosa
21	20	5,1	3,8	1,5	0,3	setosa
22	21	5,4	3,4	1,7	0,2	setosa
23	22	5,1	3,7	1,5	0,4	setosa
24	23	4,6	3,6	1	0,2	setosa
25	24	5,1	3,3	1,7	0,5	setosa
26	25	4,8	3,4	1,9	0,2	setosa
27	26	5	3	1,6	0,2	setosa
28	27	5	3,4	1,6	0,4	setosa
29	28	5,2	3,5	1,5	0,2	setosa
30	29	5,2	3,4	1,4	0,2	setosa

Before to load a CSV file, it is important to check if the options in the menu are respected by the incoming file. Thick the Header and Row Index check boxes if both they are present in the file. Choose the separator of the data depending of the file convention. Choose the way text data (names, categorial variables, etc.) are indicated. Finally select the way the decimal number are indicated in your data. If you are wrong in these selections, the file is not imported and data cannot be processed. Common mistakes are related to a wrong choice of variable and object names.

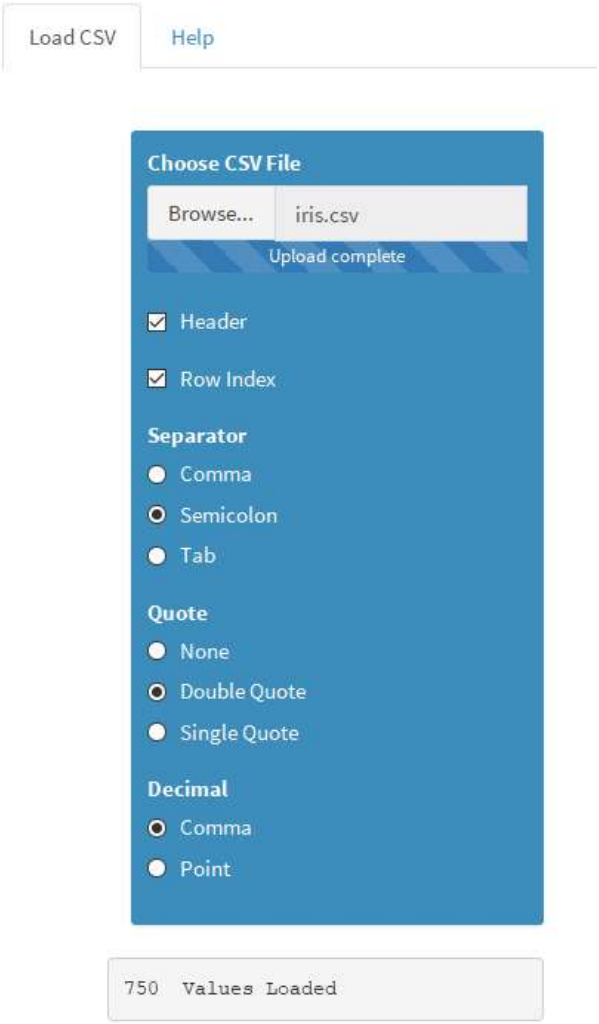


As soon as you made the choices, you can browse and select files (a filter only shows the file with the csv extension). It doesn't matter if your file is resident on the PC or in line. Depending of the size of the file the loading procedure can take few to several seconds.



If everything goes well, you must see the message *Upload complete* and the total number of data loaded in the ***Working Data Set***. The number of values is given by the product of number of columns and number of rows, independently is all the position are full. The number does not consider the row and column names. The number is important because it represent the amount of memory reserved to the process.

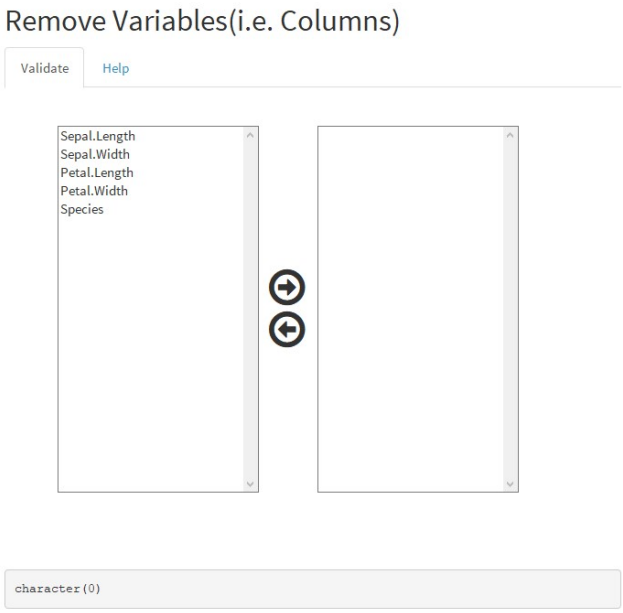
Important: you can load one Working Data Set at the same time. So, consider to include in the Data Set all the data required for your analysis.



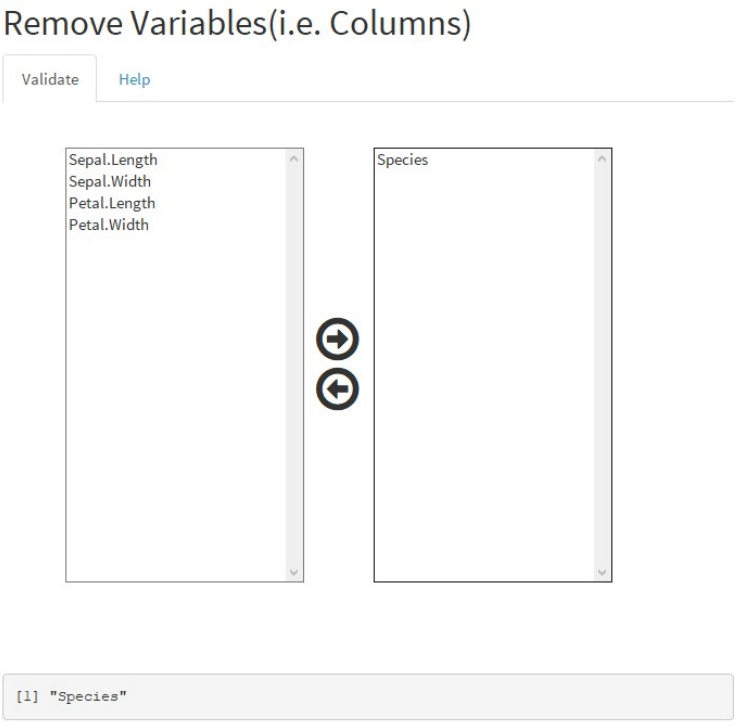
If you load another CSV file the old one is lost and the file constitutes the new Working Data Set. This is done to keep the program functionality as simple as possible. Some suggestions in the way to organize your data and make efficient Data Set will be given in the following. Although the Working Data Set (or simply the Data Set since only one is present at time) contains all the data, not all the data are used for the calculations. Some filters will allow you to choose which data to include in the process before the process starts. The idea to keep all the data in the same file may represent a big restriction, however it is an advantage when you have to store data together with their results. At the moment, I have not the program to change this structure.

Validate Variables

When a Data Set is loaded, it is possible to exclude the data of some Variables. This is done by this menu. On the left box, all the Variable names active are listed (in a just loaded Data Set they are the first-row transpose). To exclude the data of a Variable, e.g. Species, you have to highlight it by right clicking on it and then press the right-hand arrow. The name of the selected variable moves from the left to the right box (box of excluded variables).
Of course, you can re import the variable using the left-hand arrow. It is possible to select more than one variable at the same time and move them back and forward all together.

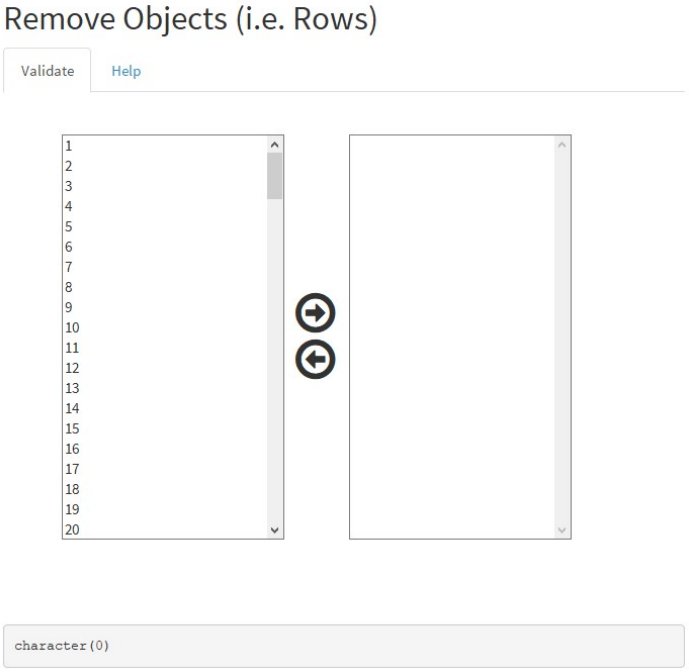


The excluded variables are also listed in the text frame below boxes. Although there are other ways in the software to exclude variables, this menu is the only possibility to re import them back in the Working Data Set.

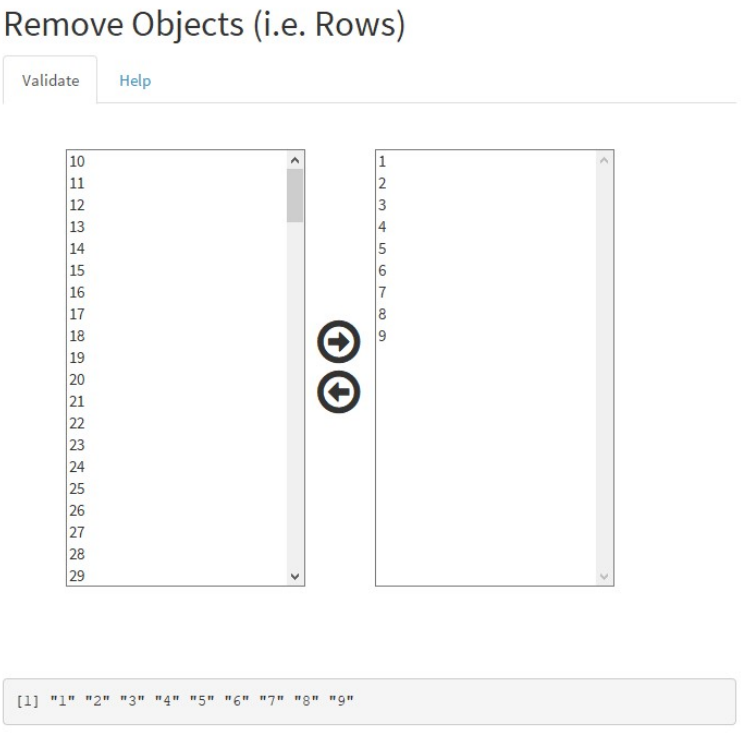


Validate Objects

Quite often some objects need to be excluded by the calculation, for instance when some data are missing or wrong. The procedure to exclude objects, i.e. rows, is similar to that one used for Variables. In fact, the left box lists the Row Index of all the Data Set. Usually these elements are more than those of variables so you can scroll through them by the scrolling bar. When you have identified the row (or rows) you want to exclude, highlight it right clicking on. Then push on the right-hand arrow and the Row index migrates in the right box, i.e. the box of excluded rows. You can select more than one row pressing the Shift and Ctrl keys in case of consecutive or nonconsecutive names. Rows excluded are not deleted and can be re imported again whenever they are necessary.

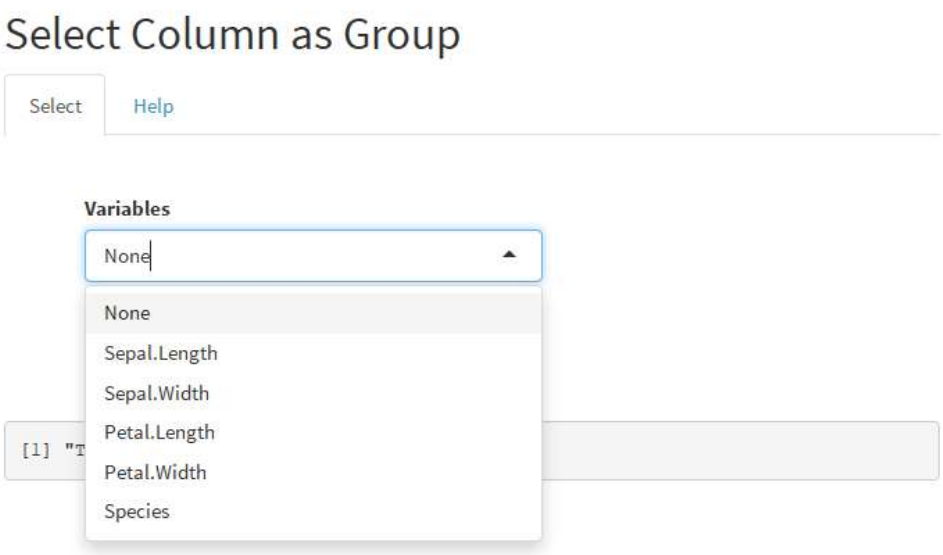


The row excluded are also listed in text frame below the boxes, just to make evident that these data will be not considered.

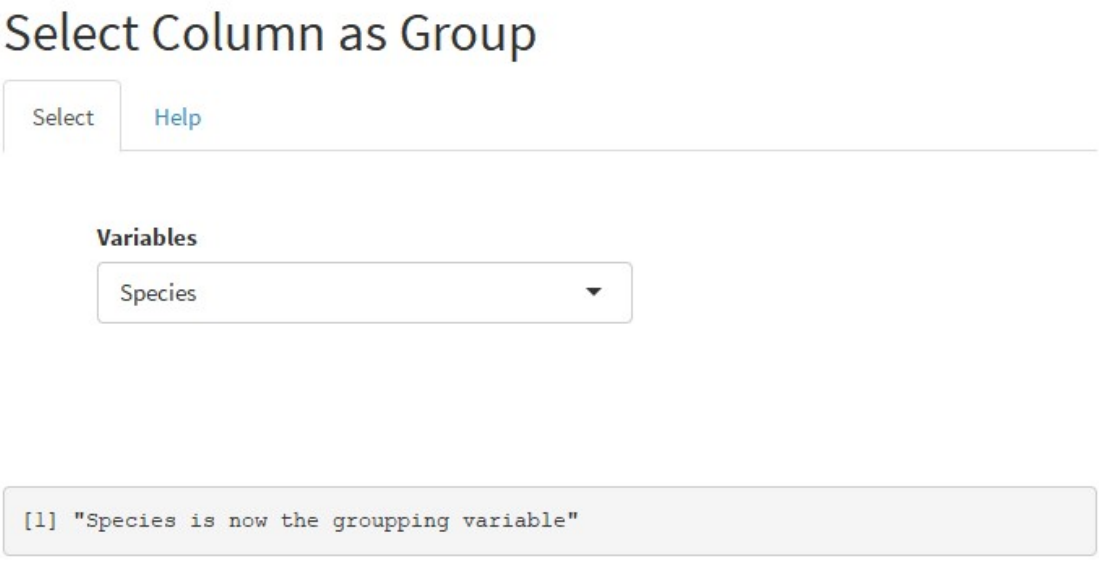


Set Groups

The Group Variable is a categorial variable that is used to group population objects in categories (or groups). The Group Variable can be made by numbers (integer) or text and must be defined for each object of the population. No value can be missing. To let ***Deducit*** know that a variable is a Group Variable, you have to select it from the Combo box. The list of the Combo box is filled and update automatically when the Data Set is loaded or the Variables are validated. If a variable is excluded, it will not be in the list. So, you have to re import it in the Data Set in order to choose it as a Group Variable.

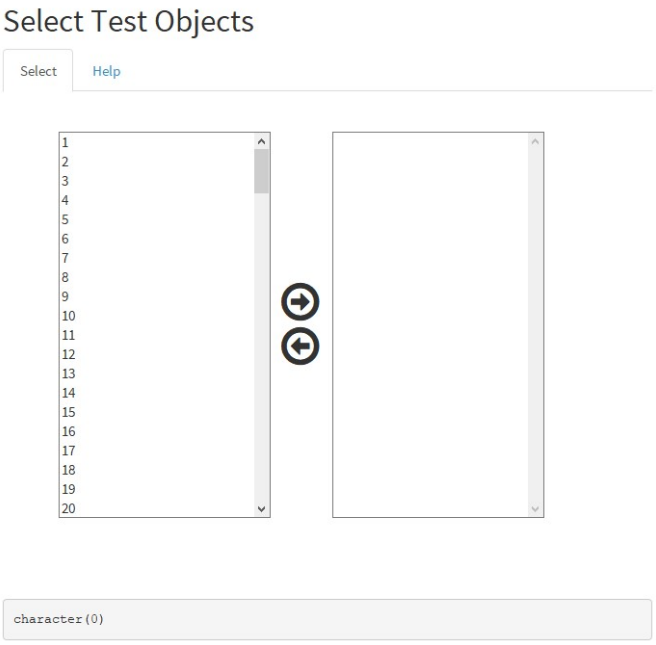


As soon as a Group Variable is selected, its name is clearly indicated in the text frame below. The frame states that the variable is used as a Group Variable. Automatically, the variable is excluded from the calculation although its effect is active in grouping results. If a Group Variable is selected, all the calculations and plots are affected. Results may look very different because groups are indicated with different colors. Consider the effect with care before selecting a variable as a Group Variable.

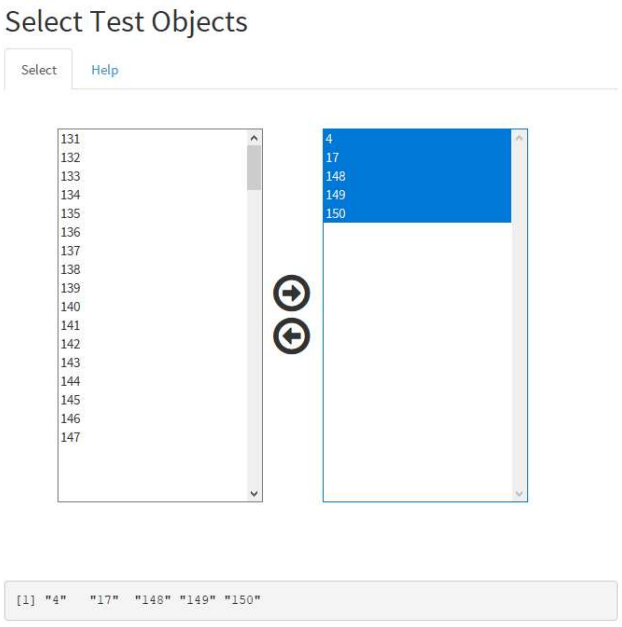


Training/Test Selection

The Training/Test menu allows to split the Data Set objects in two sets: the first, made by the objects used to set up a model (Training Set); the second, made by the objects used to test the model when it is built (Test Set). This menu combines with the menu *Additional Data* that will be described later. The box on the left (Training Set) lists the Row Index of each object in the Data Set. As done in the validation, the object that is considered a Test needs to be highlighted first. Then, pushing the right-hand arrow, its row index is moved from the left to the right box, i.e. in the box of Test Set. Selection and transfer can be done using single or multiple objects and records can be moved in both sides. **Objects in the right side, i.e.in the Test Set, will not be used in model construction such as PCA and PLS.**

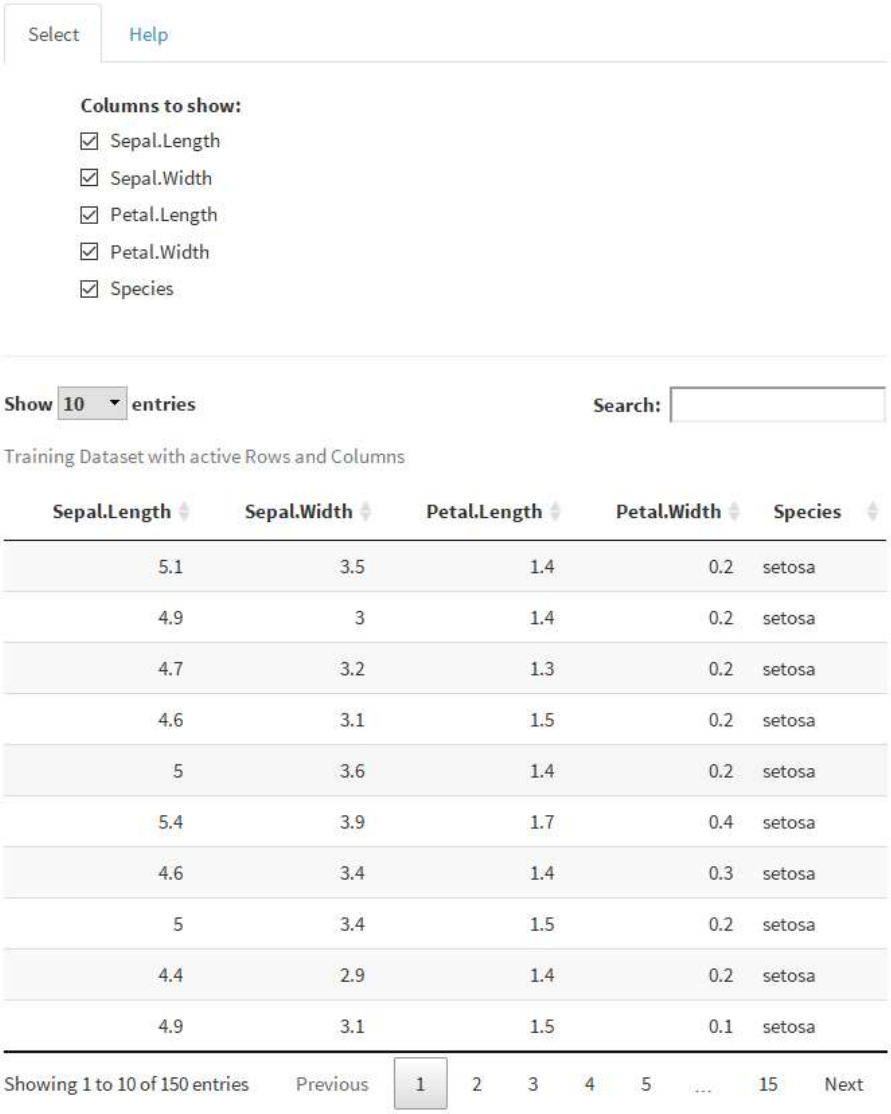


When Objects are moved to the Test Set, they are also listed in the Text Frame below as shown in the following example. In this case objects labelled: 4,17,148,149 and 150, will be used as Test Set, all the other objects stay in the Training Set.



View Data Set

The menu allows the exploration of all the Data Set. Several commands facilitate the task. At the top, the Variable Names are listed with a check box, if one of the thick is removed, the Variable is removed from the view (it is not removed from the Data Set). This is helpful when many Variables are present and only few of them are relevant. In the following example all the variables are shown since the Data Set is very thin. The *Show entries* combo box on the right, fix the number of rows in the next table. Some predefined values are allowed (10,25,50 and 100) and must be selected depending of the Data Set dimension. The bottom line of the table indicates how many objects (entries) are displayed and how many are in total. In the following example 1 to 10 of 150 entries are shown. To see the next ten, the right buttons 1,2,3,4, 5...15, need to be pushed.
The data in the table are listed as in the original Data Set. The name of each Variable is shown at the top of the column. Each row represents a different object. On the right of each Variable name a widget allows to order the column from the minimum to the maximum and vice versa. When this action is made, all the table is reordered following the rule based on the selected Variable. The ordering is done only for the view purpose, data in the Data Set are unchanged. The ordering criterion is valid only for numeric variables, however it may work for text variable too, but results are quite unpredictable.
At the top right, a Search input box is present. It allows the research of a specific value in the data. Just type numbers in the box and hit return. The usefulness of this search is evident for very big Data Set and only when few and defined values are required.



It is important to note there are two tables in the view menu. The top one is for the objects in the Training Set, the bottom one is for the Test Set. The example below shows the case of a non-empty Tests Set. In case this is not true, the second table is empty although is always displayed.

Training Dataset with active Rows and Columns

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
5.1	3.5	1.4	0.2	setosa
4.9	3	1.4	0.2	setosa
4.7	3.2	1.3	0.2	setosa
5	3.6	1.4	0.2	setosa
5.4	3.9	1.7	0.4	setosa
4.6	3.4	1.4	0.3	setosa
5	3.4	1.5	0.2	setosa
4.4	2.9	1.4	0.2	setosa
4.9	3.1	1.5	0.1	setosa
5.4	3.7	1.5	0.2	setosa

Showing 1 to 10 of 145 entries Previous 1 2 3 4 5 … 15 Next

Show 10 entries Search:

Test Dataset with active Rows and Columns

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
4.6	3.1	1.5	0.2	setosa
5.4	3.9	1.3	0.4	setosa
6.5	3	5.2	2	virginica
6.2	3.4	5.4	2.3	virginica
5.9	3	5.1	1.8	virginica

Showing 1 to 5 of 5 entries Previous 1 Next

Summary Report

The summary report is a table that highlights some useful information about the working Data Set. There are: the number of elements (including those missing), the number of positive, negative and null elements, the number of missing (empty) elements, the number of columns and rows.
At the end the split between the number of Training and Test Set elements. All data are based on the original Data Set and they are not affected by user choices, except the number of Training/Test split.

Summary Help

Show 10 entries Search:

Topic	Value
Basic Information on DataSet	
Number of Elements	750
Number of Positives	600
Number of Negatives	0
Number of Null	0
Number of NAs	0
Number of Rows	150
Number of Columns	5
Number of Element in Training	150
Number of Element in Test	0

Showing 1 to 9 of 9 entries Previous 1 Next

Trend Plot

The Trend Plot is done specifically to show Process Time series. Usually, the Data Set is a table with several Variables ordered by columns, different batches are ordered by rows in temporal sequences following a Time Index. A Group Variable highlights where each batch starts and ends using an identification number. The most common request is to show a Variable vs Time Index, superimposing different batches. Other variables may define properties of the bath like: quality, conversion, or different phase (or steps). The next Table shows a typical example of Process Time series Data Set.

Time Index | **Group Variable** | **Batch Properties** | **Process Variables**

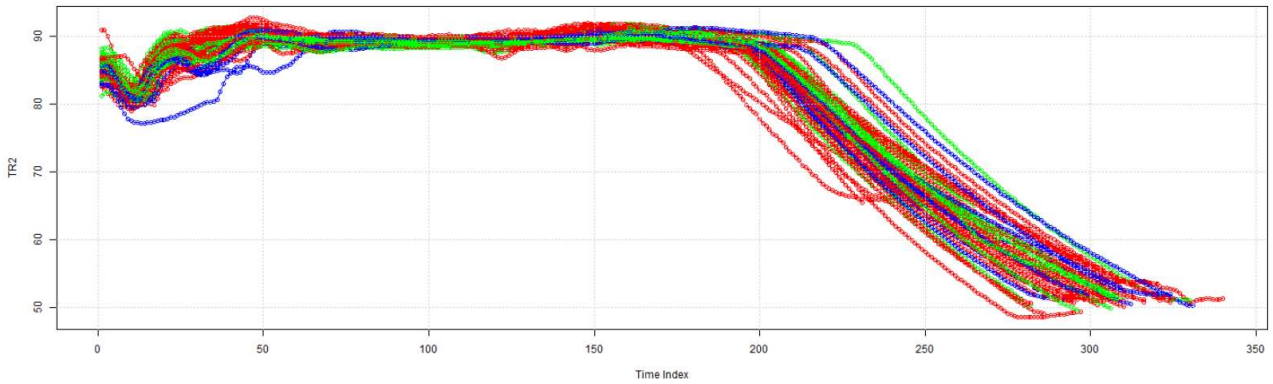
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1		Production	Formula	Step	Performance	W1	T1	TC1	SP1	IP1	SR2	IR2	WR21	TR2	TC21	TC22	WR2
2	1	1	A	1	2	4436,8	21,34	22,12	230	4934,88	53	855	1682	86,7	87,1	72,6	287
3	2	1	A	1	2	4398,5	21,23	22,12	204	3685	53	849	1682	86,6	93,5	72,7	287
4	3	1	A	1	2	4363,4	21,13	22,12	204	3731,75	53	851	1706	86,5	95,4	72,7	286,9
5	4	1	A	1	2	4326,5	21,03	22,12	204	3727,63	53	857	1727	85,9	95,7	72,5	286,9
6	5	1	A	1	2	4290,7	20,92	22,12	204	3748,25	53	866	1782	85,1	87,8	72,3	286,9
7	6	1	A	1	2	4256,9	20,82	22,11	204	3701,5	53	846	1816	84,2	84,1	72	286,9
254	253	1	A	5	2						53	1494	6316	65,5	49,5	57,5	0
255	254	1	A	5	2						53	1564	6372	65,2	50,9	57,3	0
256	255	1	A	5	2						53	1532	6374	64,9	52,4	56,6	0
257	256	1	A	5	2						53	1545	6373	64,7	53,7	56,1	0
258	257	1	A	5	2						53	1579	6367	64,4	55,3	55,7	0
259	258	1	A	5	2						53	1562	6369	64,2	57,3	55,5	0
260	259	1	A	5	2						53	1541	6375	64,1	59,5	55,3	0
261	260	1	A	5	2						53	1539	6364	64	61,9	55,2	0
262	261	1	A	5	2						53	1586	6359	63,9	65,6	55	0
263	262	2	A	1	2	4431	23,33	26,52	229	4991,25	64	966	1621	83,1	98,7	70,3	253,5
264	263	2	A	1	2	4413,6	23,34	26,54	229	4970,63	64	956	1618	83,4	99,1	70,7	253,4
265	264	2	A	1	2	4390,1	23,34	26,55	229	4848,25	64	960	1636	83,6	100	71	253,4
266	265	2	A	1	2	4356,7	23,35	26,57	229	4841,38	64	962	1664	83,6	100,5	71,3	253,3
267	266	2	A	1	2	4326,6	23,35	26,59	229	4877,13	64	945	1682	83,3	100,2	71,4	253,3
268	267	2	A	1	2	4297,9	23,35	26,6	229	4824,88	64	966	1706	82,9	99,8	71,2	253,2
269	268	2	A	1	2	4263,3	23,36	26,62	229	4853,75	64	945	1743	82,5	99,1	71,1	253,2
270	269	2	A	1	2	4229,7	23,36	26,64	229	4798,75	64	956	1777	82,1	98,7	71,2	253,1
271	270	2	A	1	2	4221,3	23,37	26,65	229	4738,25	64	958	1804	81,7	99,3	71,3	253,1
272	271	2	A	1	2	4216,5	23,37	26,67	229	4742,38	64	969	1807	81,5	100	71,4	253
273	272	2	A	2	2	4217,5	23,37	26,68	229	4804,25	54	834	1811	81,6	99,9	71,5	253
274	273	2	A	2	2	4217,8	23,38	26,7	229	4816,63	54	853	1828	81,7	96,5	71,5	252,9
275	274	2	A	2	2	4218,7	23,38	26,72	229	4730	54	825	1849	82,3	94	71,8	252,9

The menu of the Trend Plot allows to select several features by the Combo selectors:

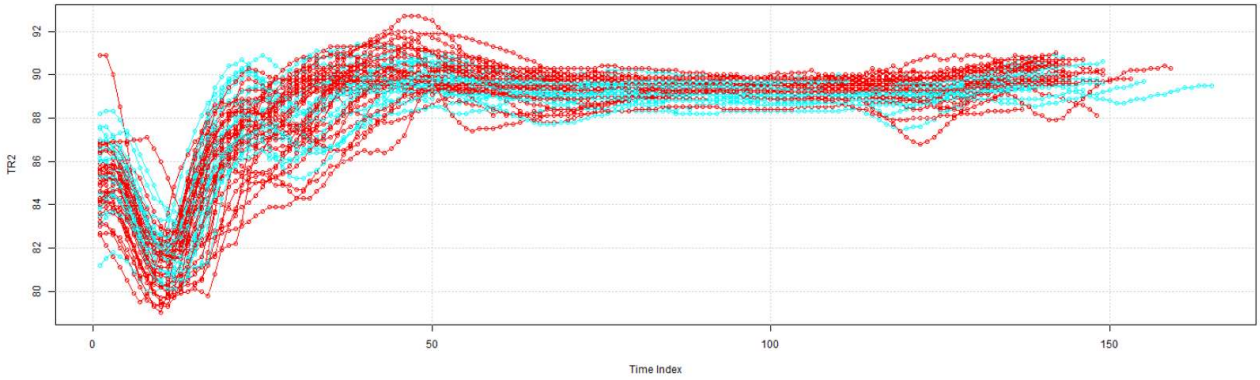
- *Variables* combo enables the selection of the variable that needs to be displayed;
- *Group by* combo allows the selection of the Group variable unless was not selected in the specific menu already;
- *Colored by* combo controls the different color by a user defined variable.

The selection made in the example indicates: the variable TR2 will be drawn vs Time Index (default) for each Production batch and each curve will be colored depending of the Performance parameter. The result is shown in the graph below.

Two other Combo selectors permit a double filtering of the Data displayed. The filtering is clearly convenient only for categorial Variables, i.e. with a finite number of choices. As soon as one of this Variables is selected all the possible values (levels) are displayed as independent check box options.

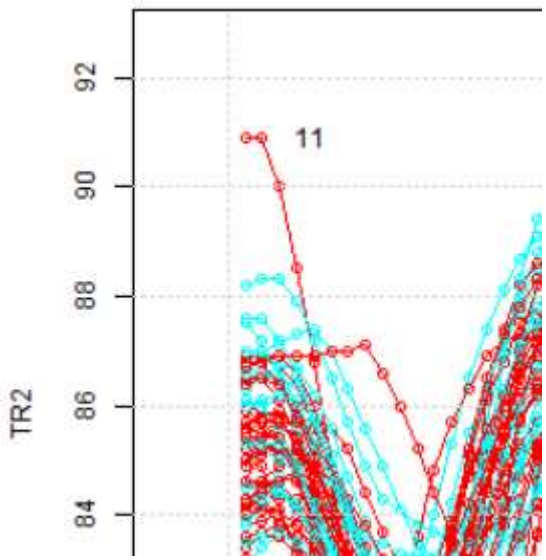


If the variable Step is chosen as first filter and Performance as a second filter, in this example, the menus is automatically update as below. This indicates that the variable Step offers five different options and Performance only three. Selecting or deselecting the Check boxes, you see almost immediately an updating of the graph. Let's suppose to choose only the Step 1,2 and 3 and deselect Performance 3, the results is that in the following figure.

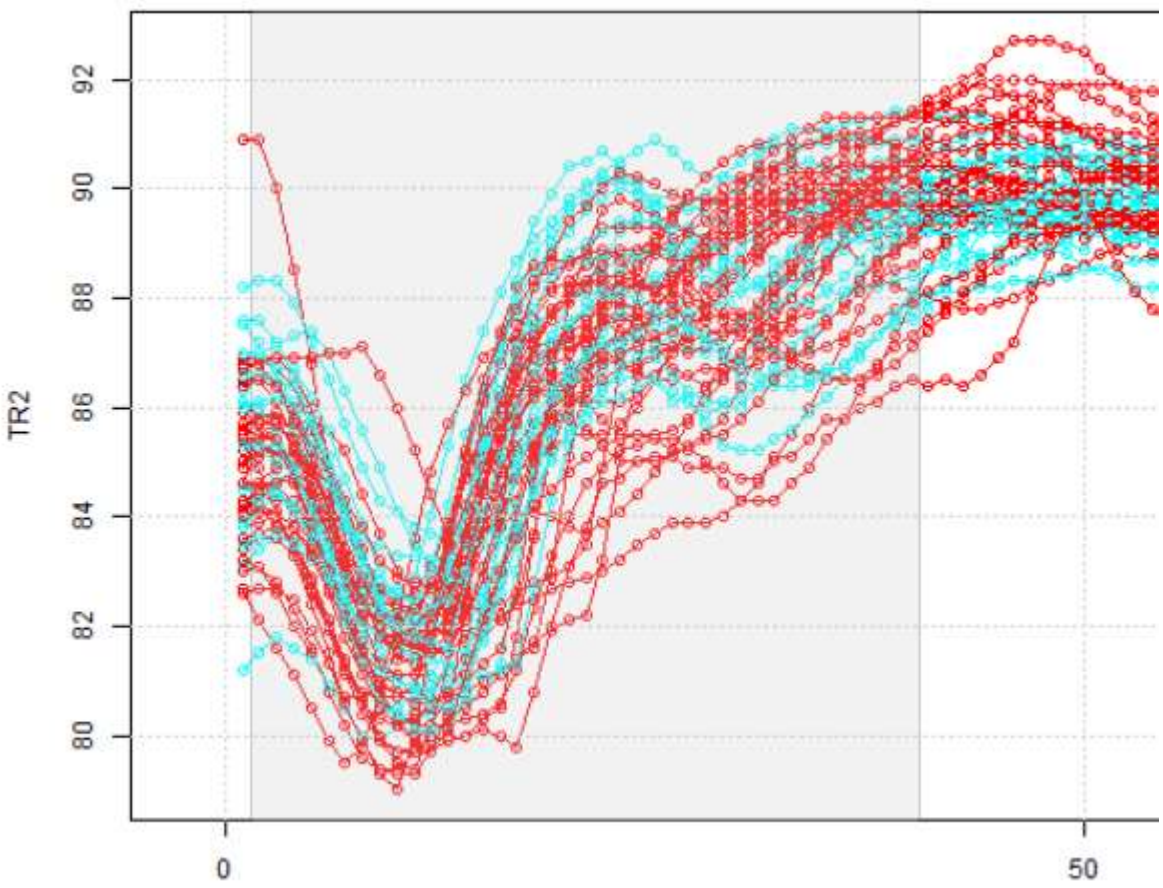


Many possibilities can be experimented in order to visualize what is considered more interesting. When something interesting is found, we would like to know more details. At the moment a useful graph interaction is implemented in order to identify the curves that present anomalies. Let's suppose we want to know which Production batch has a value of TR2 above 90 at Time Index zero. If you right click on the curve, where it looks interesting, a label with the object name is displayed. In our example we find the Production batch number 11 is the right one.

You can right-click wherever on the curves and get how many labels you want. This may happen because graphs are often overcrowded and labels are not always very visible. I suggest you follow the curve of interest and right click exactly where the curve is freer. At some point, you prefer to have a clean plot again and this is the effect of the Delete Labels button.



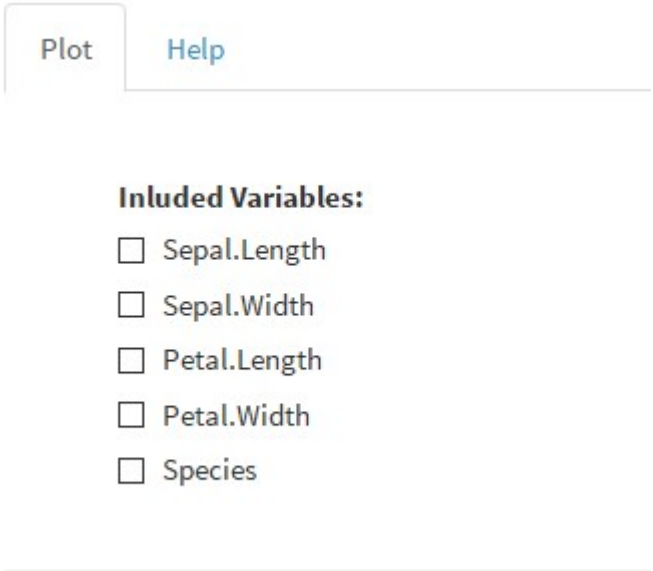
The last but not the least interesting feature of the Trend Plot is **the possibility of graphically filtering the Data Set.** A simple example explains the concept. Referring to the last plot, let's suppose we are only interested to data in the area of minimum, i.e. where the Time Index is between 0 and 50. If brush the mouse over this area we can select it, as shown in the figure:



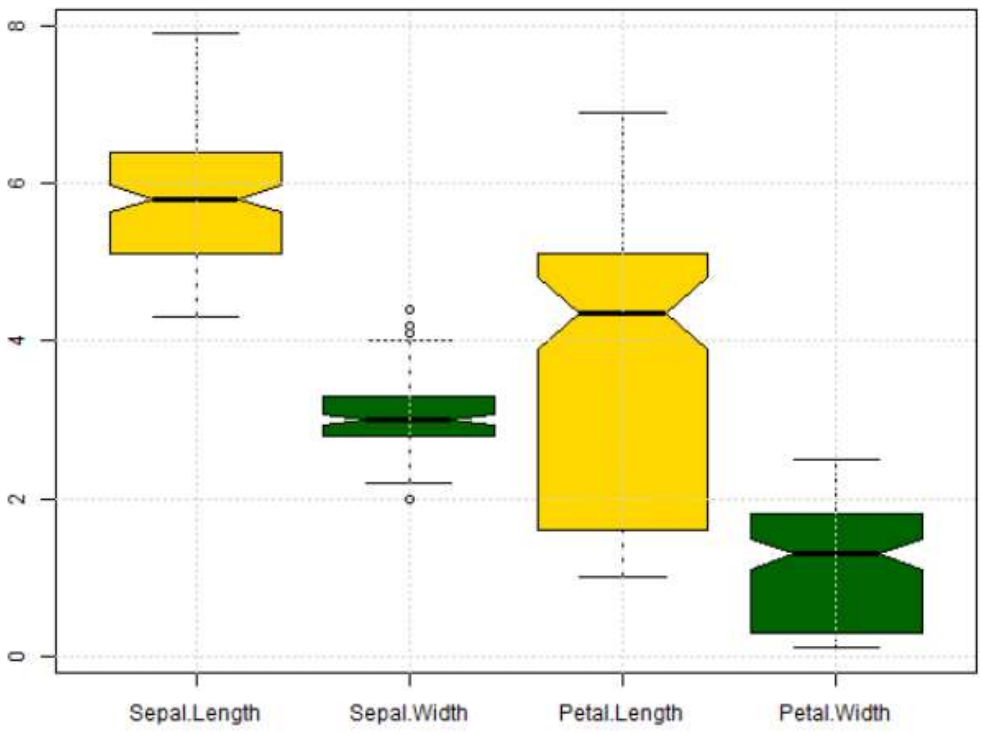
If we push the button *Extract* at the bottom a new CSV file is created with the data we have selected graphically. This is quite useful when Data Set is large and you have to sort many of its parts. The CSV files can be then merged and imported again in ***Deducit*** as a new Data Set.

Box Plot

The Box Plot is a very useful graph when you have to perform a univariate analysis. It is even more useful when you compare boxes of different Variables all together. This easy task is performed by this menu. Consider the iris Data Set and see how the menu will shows up:

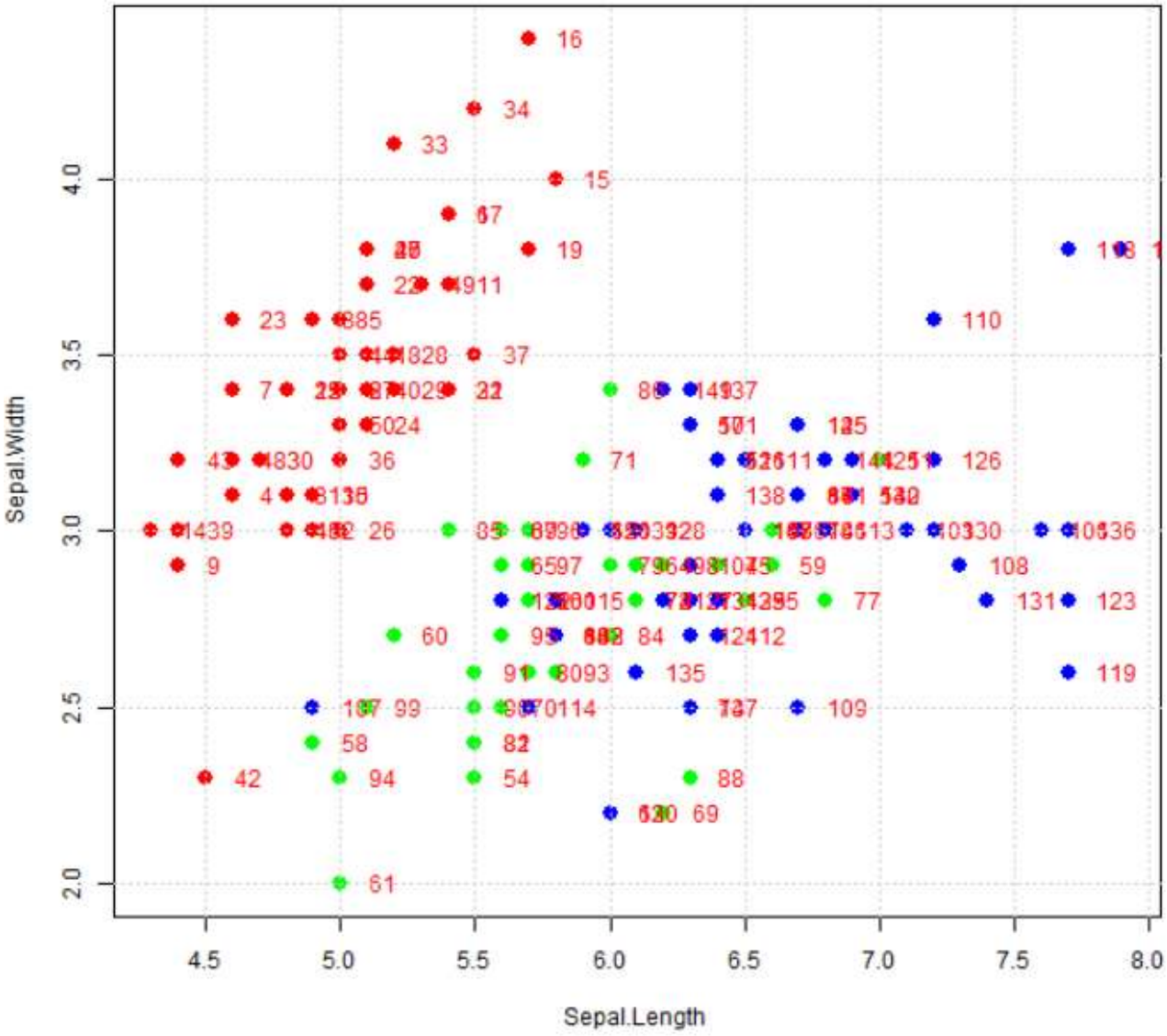


The starting option is to select what Variables to include in the graph. Checking the square at the left of the name will display immediately the corresponding bar in the graph. You can add as many bars you like but make attention that the scale is common and it is always better to work on centered and scaled values. The result of iris example with all the variables selected is below. Points outside the bars are outliers.



2D Plot

A simple and basic graph is the 2D plot where one variable is plotted versus another. The task is done using the corresponding Combo selectors that list the names of all the validated Variables in the Data Set. As soon as two variables are selected (by default the first and the second in the working Data Set are chosen) the plot is made. In case a Group Variable was previously set, different colors are assigned to each group. By defaults also the object names are displayed as shown in the case of iris Data Set.



Some graphical functionalities are available. You can select and deselect objects either by right clicking on the points or dragging an area on them. The names of the selected objects are listed in the Text frame above. The figure shows what happens if we select objects in the right side of the iris plot.

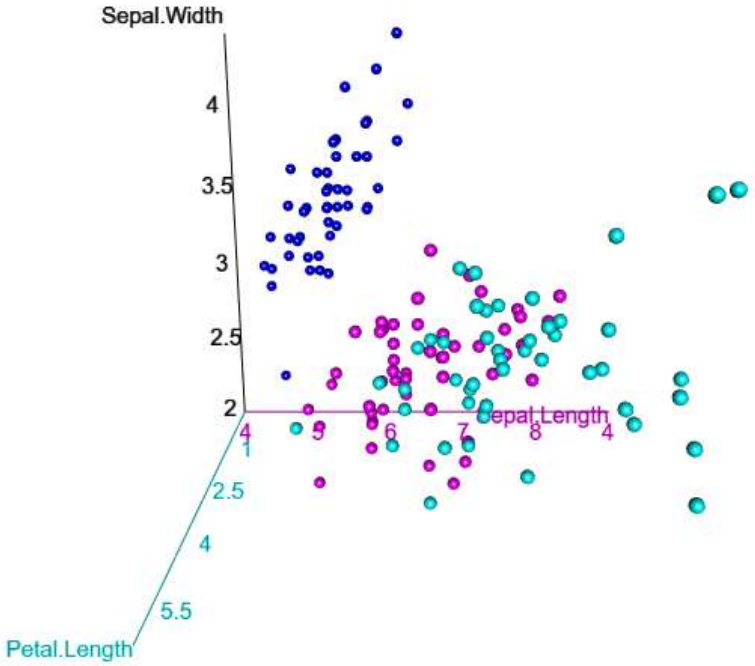
After the selection, you can:

- *Remove* the objects from the plot and move them in the not-validated list. This means data are not deleted by the Data Set but they are excluded in the all calculations (you can reintroduce them again using the Validation menu)
- *New Group*, i.e. set the selected objects as a new group in the Data Set. If the splitting in Groups was already present a new group is created, vice versa two groups are created: one with the new objects the other with all the remaining ones.

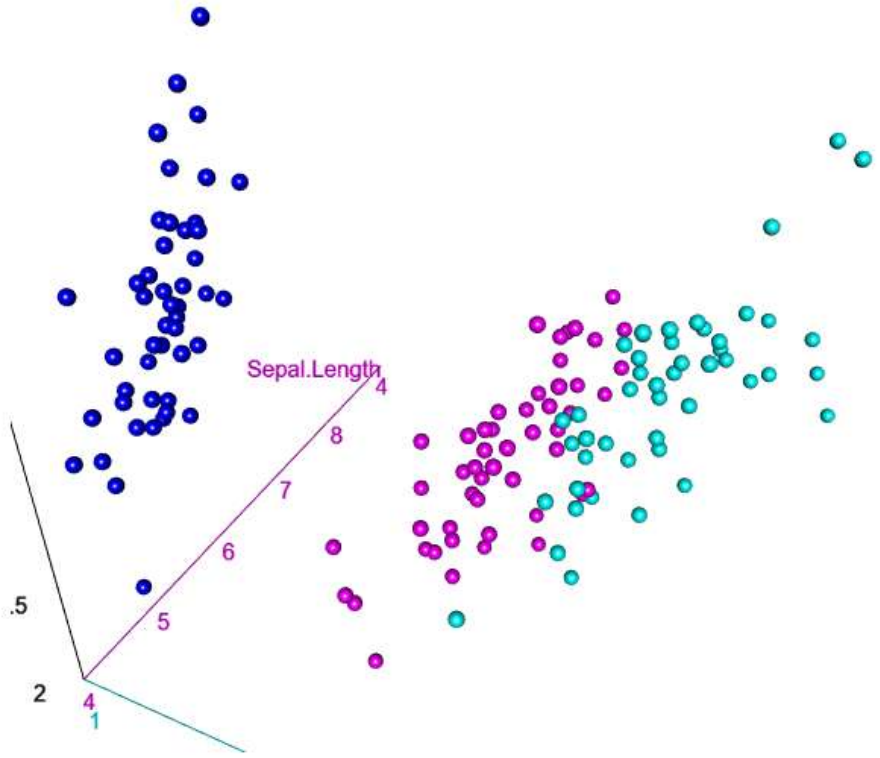
Whatever is your choice, the plot is update accordingly.

3D Plot

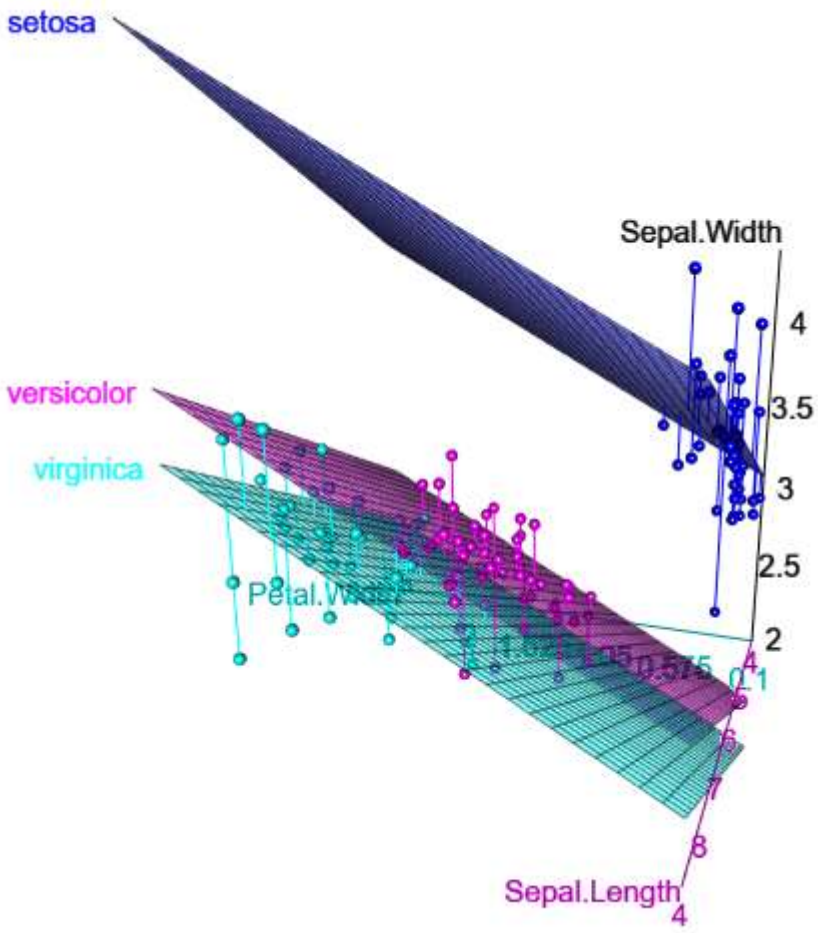
A more sophisticated plot is the 3D plot. This plot is helpful to visualize the spreading and the shape of the Data Set more than making any quantitative evaluation. By default, when the menu is chosen the first three Variables in the Data Set are selected and a scatter plot with a point per object is drawn. See next, what you see in case of the iris Data Set.



The graph is dynamic in the sense that clicking on it and moving the mouse will cause the plot rotation. Scrolling the mouse wheel will zoom in and out on the plot. You can easily arrange the best view you like it.



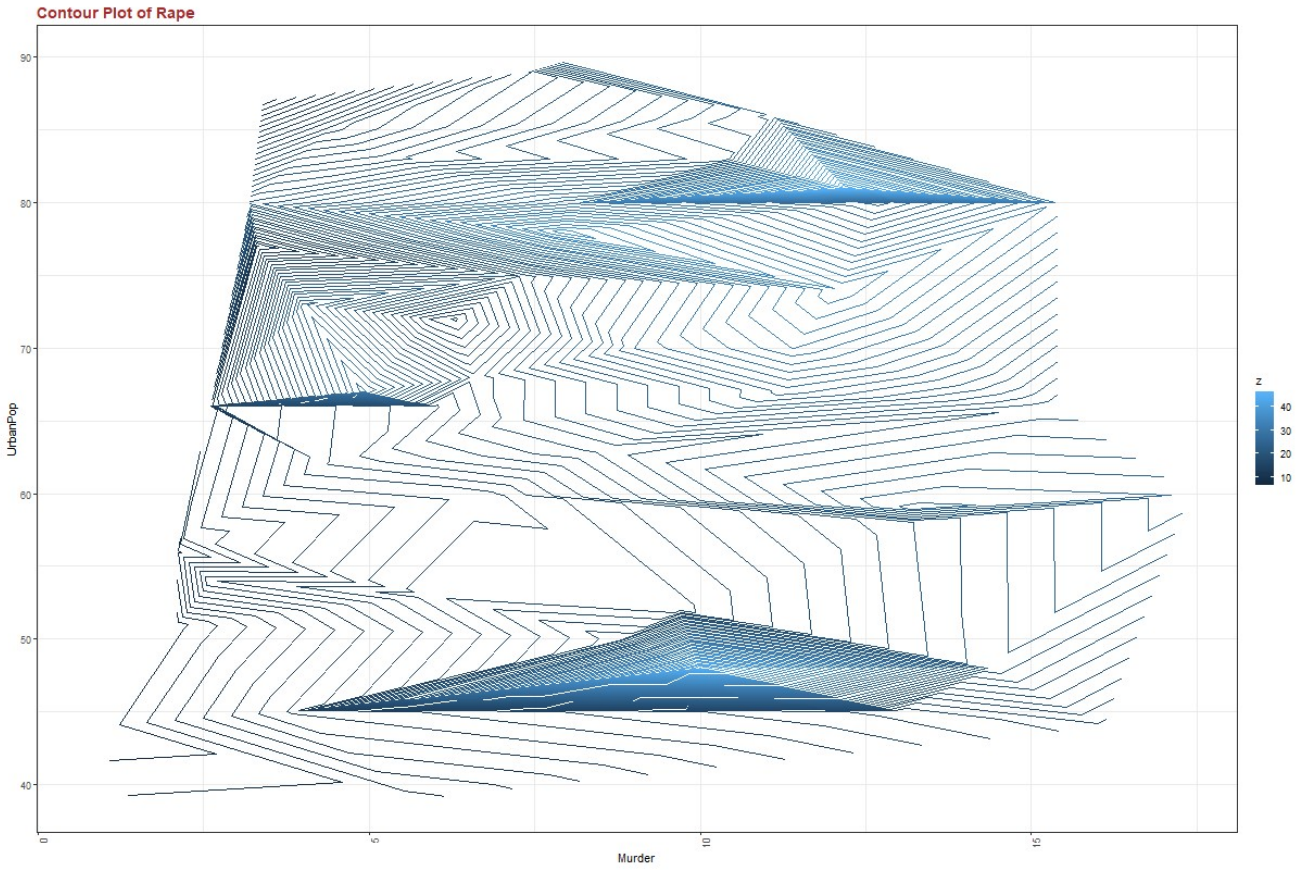
If the Data Set has a Group Variable, as it is the case of iris with the variable Species, you can draw of the best fitting planes for each Group. This is achieved checking the box *Surface* at the top.



Although 3D plots are visually impressive, their use is quite limited and they are prone to software errors and limitation. Make sure that data in the column you would like to show are numeric, none is missing and they have approximately the same dimension.

3D Contour Plot

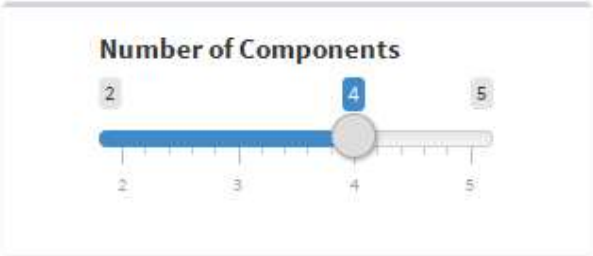
The 3D contour plot helps to map a 3D surface in 2D dimension by interpolation of level surfaces. It is the same method used in geographical maps. If you use a common Data Set with random data, the interpolation could be difficult and sometime impossible. That why it is possible this plot will not be drawn and an error is raised.
The selection of the three variables to plot is made, as usual, by the Combo Selectors. I show later what happens if you use the USArrests example Data Set.



The number of the level lines is managed by the top left Slider, increasing the number makes the surface better defined.

Principal Component Parameter Setting

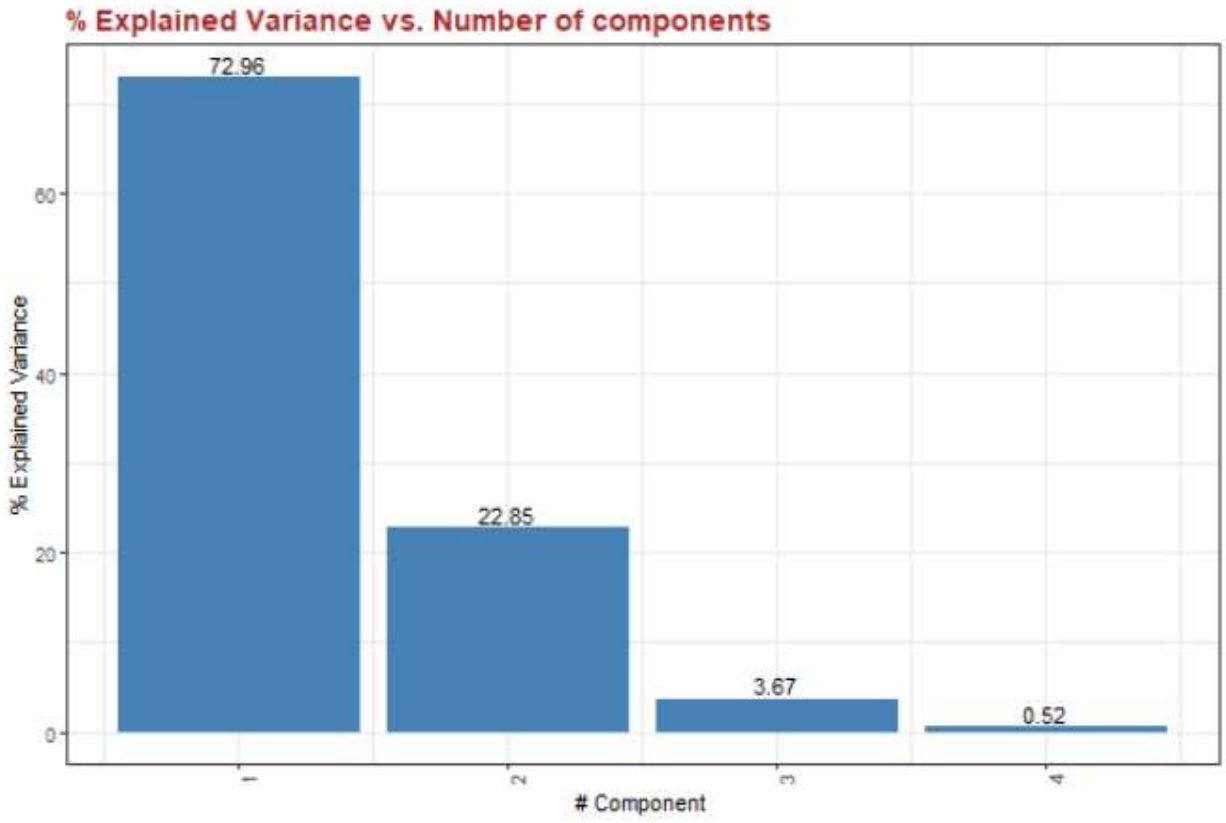
The Principal Component Analysis (PCA) starts setting three simple parameters: the kind of preprocessing of the original data (Centering and Scaling) and the number of components that will be considered. It is almost universal to mean centering the data to avoid the effect of different dimensions. Scaling, i.e. dividing the data by the column (variable) standard deviation is also very common. Scaling is helpful to maintain comparable the deviation (spreading) of all the variables. If both preprocessing methods are performed, Data are said *Autoscaled*. The preprocessing option can be set by the thick boxes on the right panel. By default, autoscaling is set.
More difficult is to fix the number of principal components that need to be considered. The minimum value is 2 and the maximum is given by the minimum value between the number of validated rows and columns (the software considers if you excluded some rows/columns in the validation part). You have to decide between the two limits which value is more suitable for your purpose. If you do not know, fix the value to the maximum, unless this value is not exceptionally high (more than some hundred). To fix the value, you have to use the left slider widget. Consider that is very common changing the value many times before to fix it, and each time the software makes the calculation again. The value of the principal component considered is shown in almost all the results and plots, so you don't need to remember it. Finally, make attention that, if the number of component is the maximum, all the variance is explained and some plots may be useless.



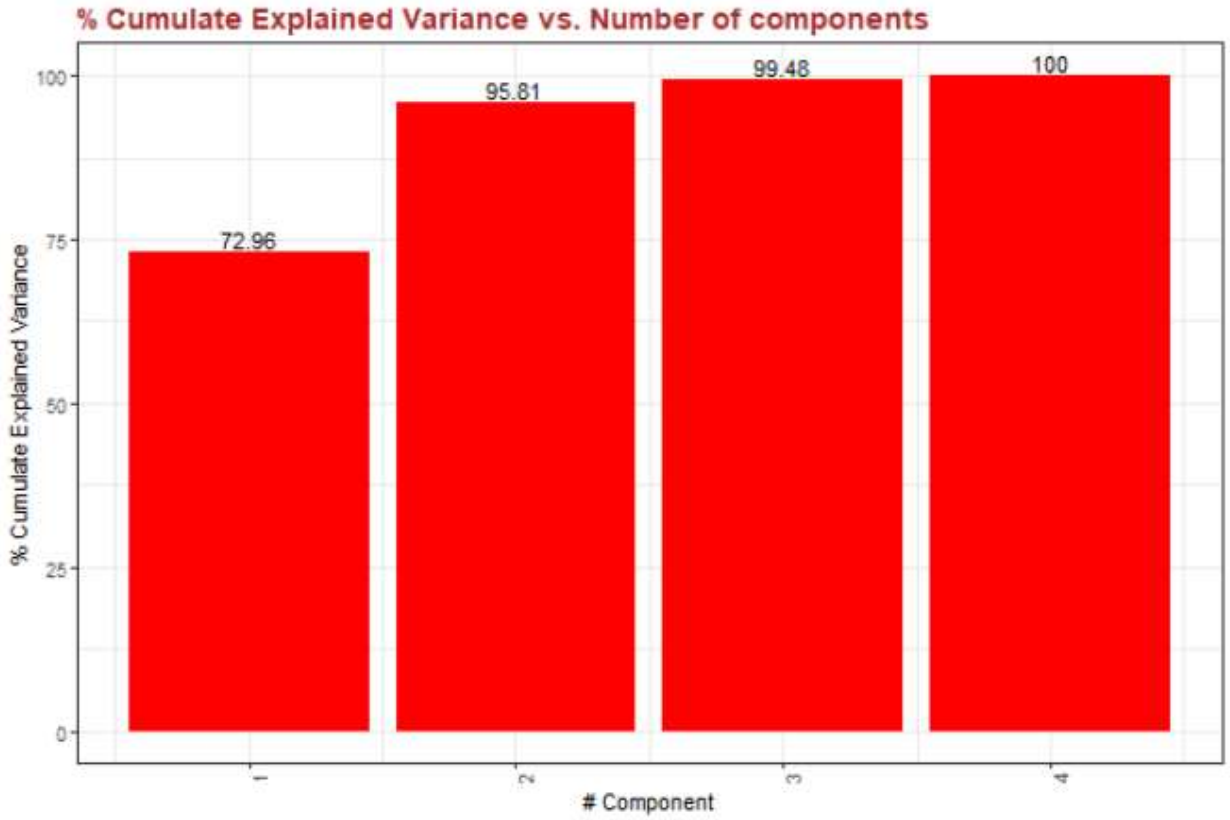
When you fix the number of principal components you have to verify that the number of variables is consistent with your choice and all the variables included in the Data Set are numerically processable. So, if in Data Set, you have some categorial variables, useful for grouping data but not to measure their properties, you have to exclude it before. For variable exclusion, use the menu of variable validation. **Usually the Data Set processable for PCA is a fully numeric Data Set.**

Explained Variance

PCA explains the maximal variance of our Data Set. Remind that the variance of the Data Set is given by the diagonal values of its covariance matrix. The value of the maximum variance is easy the identify but the we need the PCA algorithm to find the direction with the largest variance in the space. The result is a new set of coordinates the first of which runs in the direction of the maximum value. The second moves in the orthogonal direction to the first to find the second largest variance and so on. The covariance matrix between the principal components is diagonal and the values are the eigenvalues of the original covariance matrix. The traces of the two matrices are equal to the total variance of the Data Set (if the Data Set is scaled the variance is equal to the number of variables).
The first plot given by this menu shows the eigenvalues, i.e. the variance associated to each component. Of course, the number of components displayed equals that chosen in the setting. Values are always in decreasing order as made by the research algorithm. In this plot is important to see how fast the variance decreases. The optimal number of components to consider is given by the steepest drop down of the variance values.
In the example of iris Data Set as shown in the picture, the two first components are enough to describe the most part of the variance, since a big drop is observed on the explained variance after the second component.



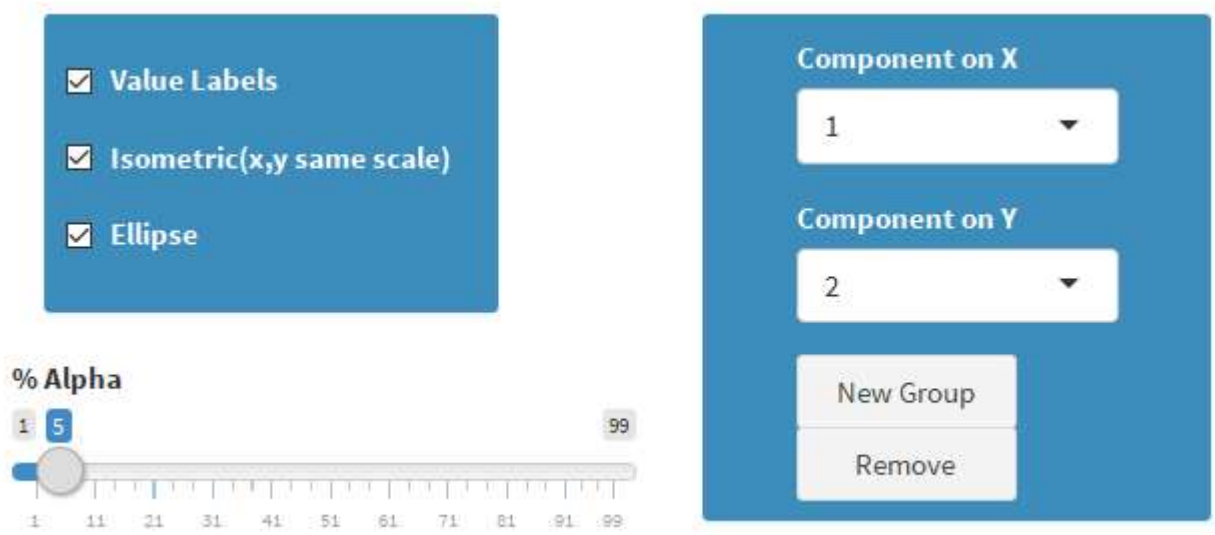
It is important to preserve the highest value of variance in the choice of the number of components but it is important not to include the random error. Since with measured Data Set some random error is always present, the amount of explained variance can never be 100%. For this reason, is helpful see the cumulative plot of the previous one, given by the next picture. If we consider the first two components, we retain almost the 96% of the global variance. The last two components are consequently useless.



The choice given by the Combo box on *Value Labels* enables the option to see on the plot the numerical values of each bar. You can choose to see it or not.
The *Download* button at the bottom allows to get a csv file with the numerical values of the two graphs.

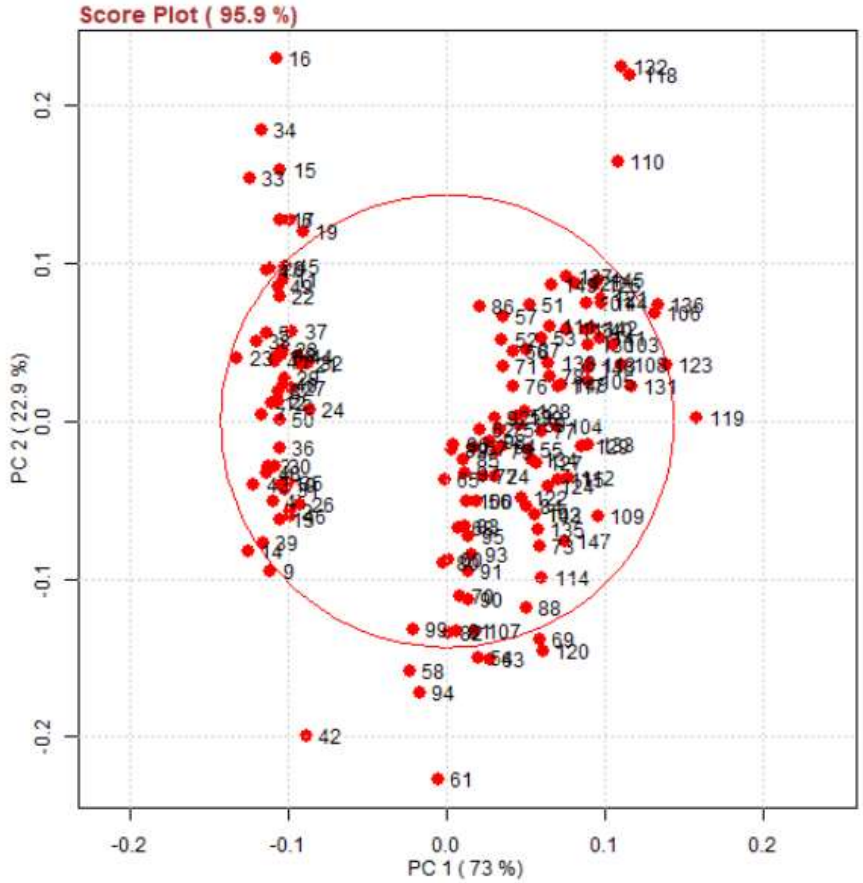
PCA Score Plot

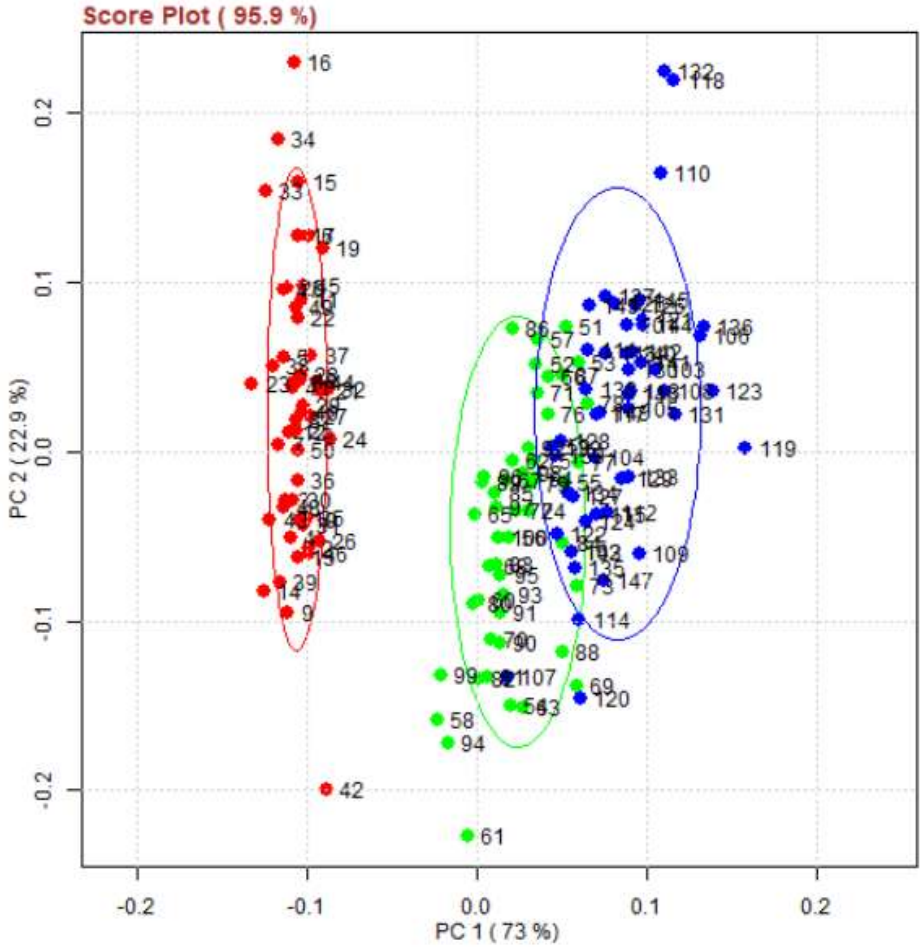
The Score Plot in PCA represents the projection of the Data Set in the new space of the Latent (Principal) Variables. Usually this space is viewed by the use of only two components per time (and quite often two components are the only needed). You can choose the Principal Components on X and Y axes using the combo boxes on the right. The number of components displayed are those selected in the Setting PCA menu and the only used in the calculation. Of course, the components on the two axes must be different in order to get a reasonable plot.



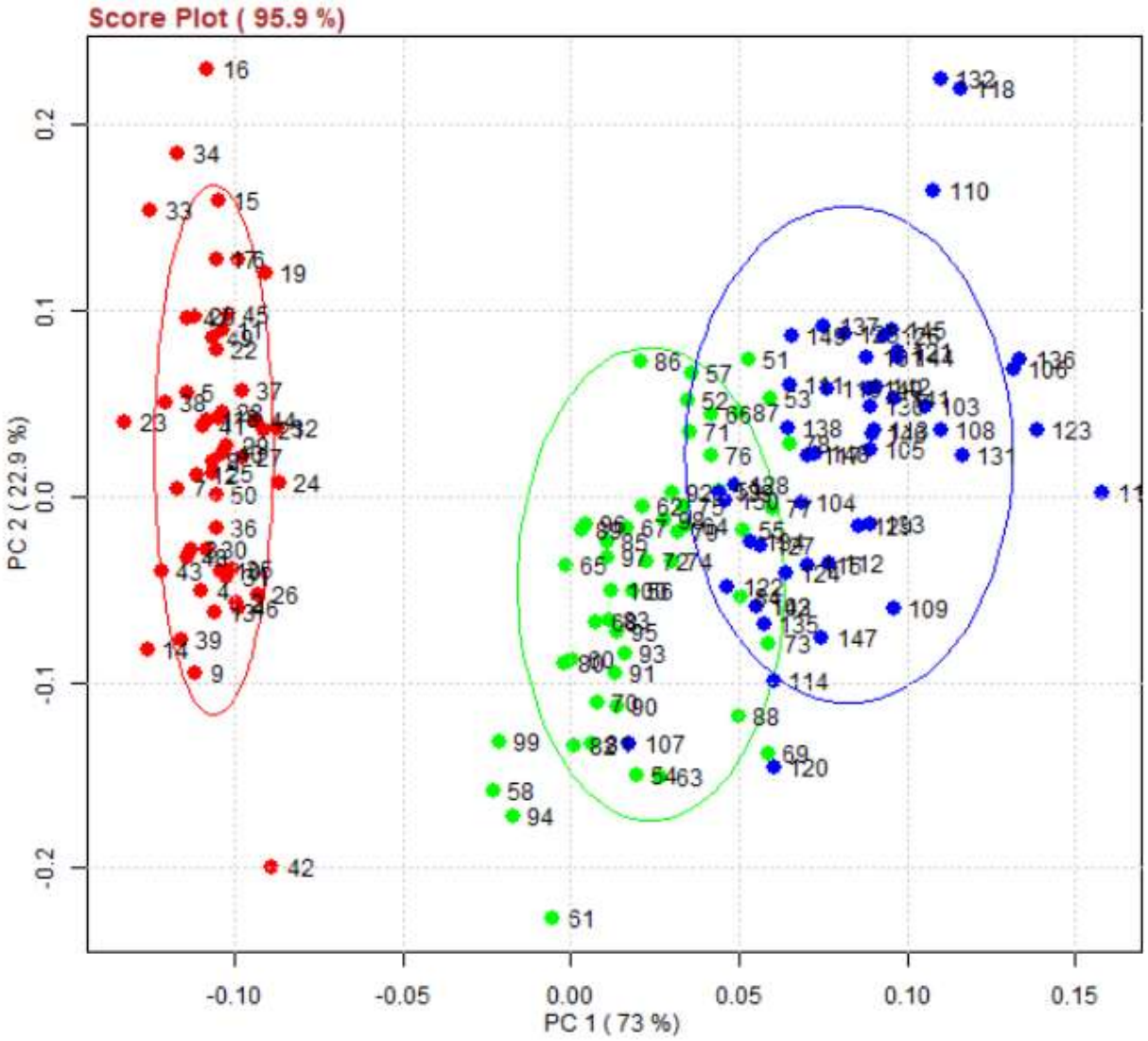
On the left, some plot options are available:

- *Value Labels* is used to show the object name close to each point.
- *Isometric Plot* forces the plot to have the same scale on both axes. This is helpful to assess the real spreading of data; however, it could be needed to switch it off in case of overcrowded clusters of points.
- *Ellipse* option shows the dispersion curve around the average of all the points with Alpha significance. The value of Alpha can be set by the slider below, the default is the usual 5%. The Ellipse is drawn depending of your previous choice too. In fact, if you previously set the splitting of the data in Groups (see Set Group menu) one ellipse for each Group is drawn. In this case, the center and shape of ellipse is related to the parameters (average and spreading) of each single group. Vice versa, if any Group is chosen, the data are considered all together and one ellipse is made. In case of Group splitting also the color of each Group is changed to make the recognition easy. In the next two plots we see the effect of the Score Plot for the Iris Data Set when the objects are selected by Species.

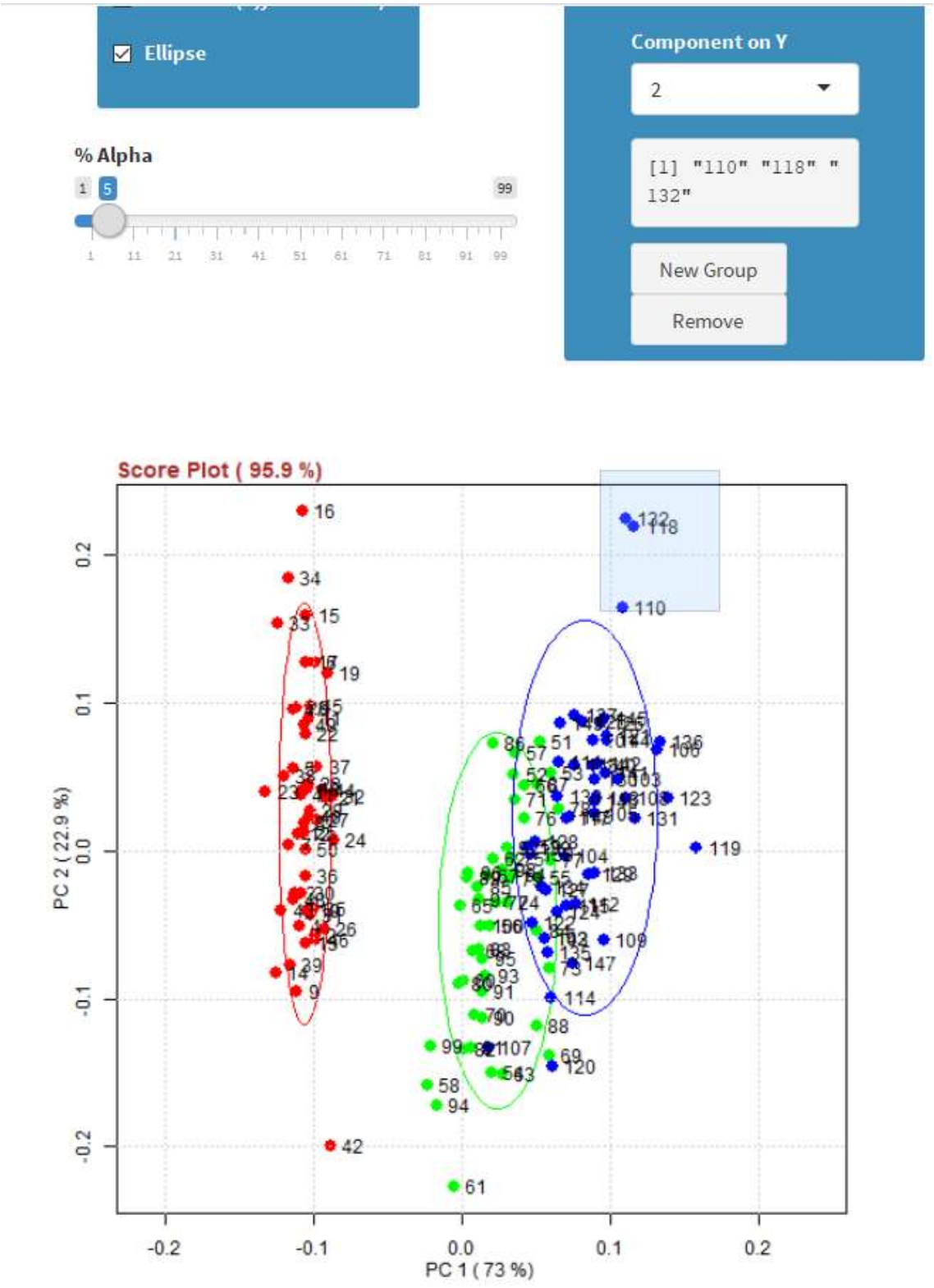

Score Plot (95.9 %)
PC 2 (22.9 %)
PC 1 (73 %)
-0.2
-0.1
0.0
0.1
0.2


Score Plot (95.9 %)
PC 2 (22.9 %)
PC 1 (73 %)
-0.2
-0.1
0.0
0.1
0.2

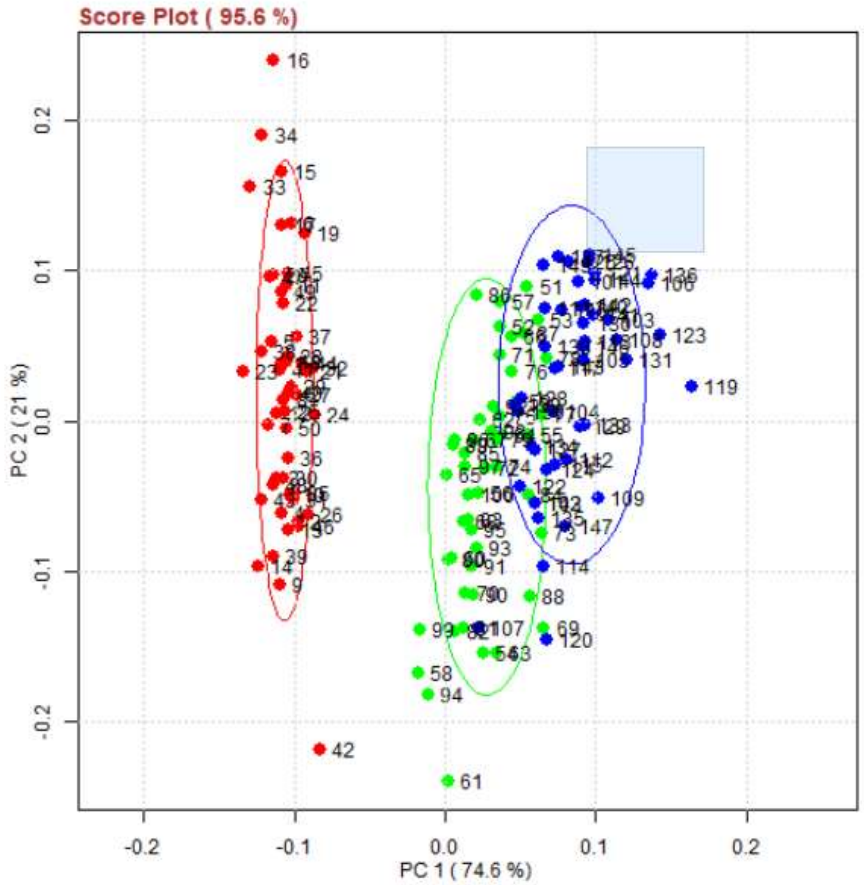
To show the effect of Isometric Scale constrain, I repeat the last plot (where the constrain was active) in the next plot (where the constrain is not active). Of course, this is only a graphical effect (the data are not changed) but you see that the in the second case labels are more easily identified. However, we stress that the correct data distribution is achieved when the same scale is used on all the components, as in the default choice.

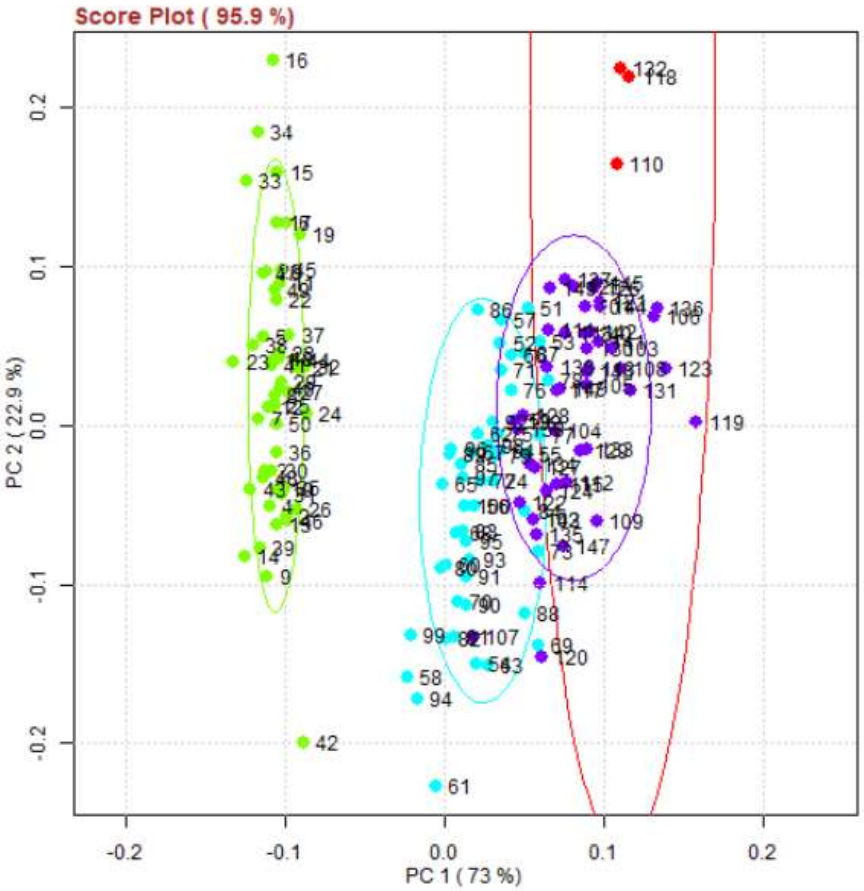


Score Plot is commonly used to detect *Outliers*, i.e. points that are outside some clustering criteria we are considering. For instance, in the Iris plot the Objects 110, 118, 132 are quite far away from the center of the blue Group. As soon as this hypothesis is raised, **it is possible to select them just dragging with the mouse on the plot and pushing the right button**. The selection is confirmed when the labels of the selected objects are written in the Text frame (see picture below). **The selection of a single object can be done also right click on it**. The selection can be erased clicking again on a selected object. Repeating the procedure, it possible to select whatever object in the plot. Since mouse operations are sometime tricky, please check if the selected objects in the Text frame are exactly those wanted.



After a selection of objects is performed, two actions are then possible: *Remove* i.e. deleting the selected objects and excluding them from the PCA calculation; *New Group* i.e. pushing the selected objects in a new additional group. The two effects are shown in the next pictures. Removing objects 110, 118 and 132 do not change to much the shape of the three clusters. Consider that if you want to reintroduce the values again you can do it by the Validation Object menu. The creation of a New Group is more effective because a new color and a new ellipse appear. The creation of a new group cannot be undo unless you reload the original Data Set.

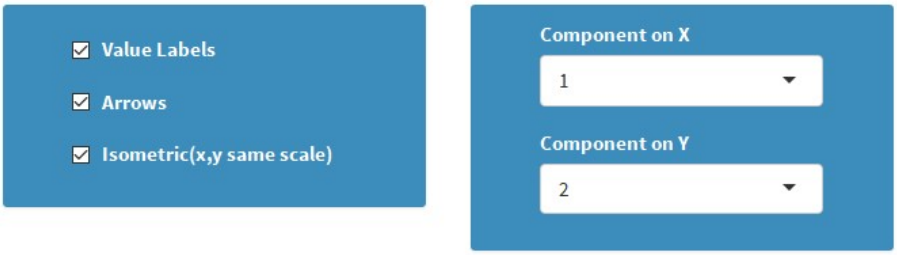




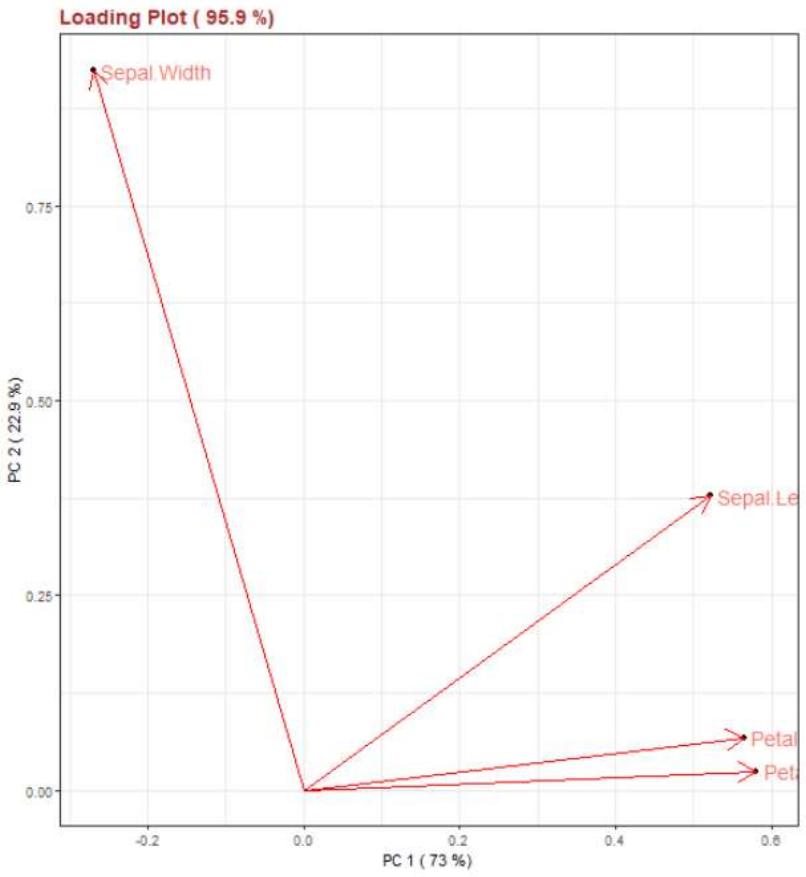
The button *Download* allows the creation of a CSV fie with the Scores Raw Data.

PCA Loadings

Loadings are the correlation coefficients between the original Variables and the unit-scaled components. They are normally represented as arrows (one for each variable) in the latent space. The plot is usually bidimensional choosing which Principal Component put on X and Y axes. This can be easily done by the Combo boxes on the right. *Arrows* can be drawn and they all start from the center. In case there are too many variables and so arrows are not easily identified, uncheck the arrow Check Box in order to have only the points and labels drawn. Even the *Variable Labels* can be omitted by the corresponding Check Box. As for the Score Plot, by default the plot is forced to use the same scale for all the components (Isometric Check Box active).
Use the loading plot to identify which variables have the largest effect on each component. Loadings can range from -1 to 1. Loadings close to -1 or 1 indicate that the variable strongly influences the component. Loadings close to 0 indicate that the variable has a weak influence on the component. Evaluating the Loadings can also help you characterize each component in terms of the variables.



The next figure shows the loading plot for the Iris Data Set. Petal.Length and Width are very much correlated and Sepal.Width and Sepal.Length are instead independent. The *Download* button allows the creation of a CSV file with the numerical values of the calculated Loadings.



Please note as taken from: https://stats.stackexchange.com/questions/2691/making-sense-of-principal-component-analysis-eigenvectors-eigenvalues/35653#35653

I want especially to stress twice here the terminological difference between **eigenvectors** and **loadings**. Many people and some packages (including some of R) flippantly use the two terms interchangeably. It is a bad practice because the objects and their meanings are different. Eigenvectors are the direction cosines, the angle of the orthogonal "rotation" which PCA amounts to. Loadings are eigenvectors inoculated with the information about the variability or magnitude of the rotated data. The loadings are the association coefficients between the components and the variables and they are directly comparable with the association coefficients computed between the variables - covariances, correlations or other scalar products, on which you base your PCA. Both eigenvectors and loadings are similar in respect that they serve regressional coefficients in predicting the variables by the components (not vice versa!11). Eigenvectors are the coefficients to predict variables by raw component scores. Loadings are the coefficients to predict variables by scaled (normalized) component scores (no wonder: loadings have precipitated information on the variability, consequently, components used must be deprived of it). One more reason not to mix eigenvectors and loadings is that some other dimensionality reduction techniques besides PCA - such as some forms of Factor analysis - compute loadings directly, bypassing eigenvectors. Eigenvectors are the product of eigen-decomposition or singular-value decomposition; some forms of factor analysis do not use these decompositions and arrive at loadings other way around. Finally, it is loadings, not eigenvectors, by which you interpret the components or factors (if you need to interpret them). Loading is about a contribution of component into a variable: in PCA (or factor analysis) component/factor loads itself onto variable, not vice versa. In a comprehensive PCA results one should report both eigenvectors and loadings.

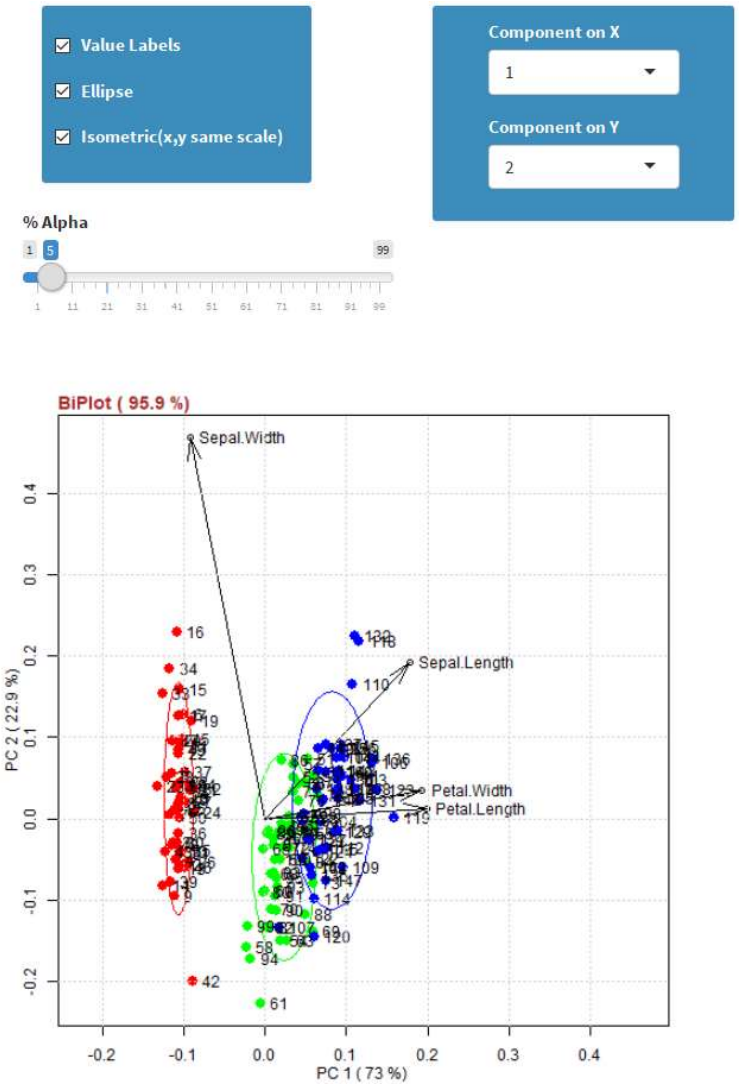
PCA Biplot

As used in Principal Component Analysis, the axes of a biplot are a pair of principal components. These axes are drawn in black and are labeled PC1, PC2, etc. and are selected through the right Combo Boxes. A biplot uses points to represent the scores of the observations on the principal components, and it uses vectors to represent the coefficients of the variables on the principal components. The figure below shows the biplot for the Iris Data Set.

Interpreting Points: The relative location of the points can be interpreted. Points that are close together correspond to observations that have similar scores on the components displayed in the plot. To the extent that these components fit the data well, the points also correspond to observations that have similar values on the variables.

Interpreting Vectors: Both the direction and length of the vectors can be interpreted. Vectors point away from the origin in some direction.

A vector points in the direction which is most like the variable represented by the vector. This is the direction which has the highest squared multiple correlation with the principal components. The length of the vector is proportional to the squared multiple correlation between the fitted values for the variable and the variable itself. The fitted values for a variable are the result of projecting the points in the space orthogonally onto the variable's vector (to do this, you must imagine extending the vector in both directions). The observations whose points project furthest in the direction in which the vector points are the observations that have the most of whatever the variable measures. Those points that project at the other end have the least. Thus, vectors that point in the same direction correspond to variables that have similar response profiles, and can be interpreted as having similar meaning in the context set by the data.

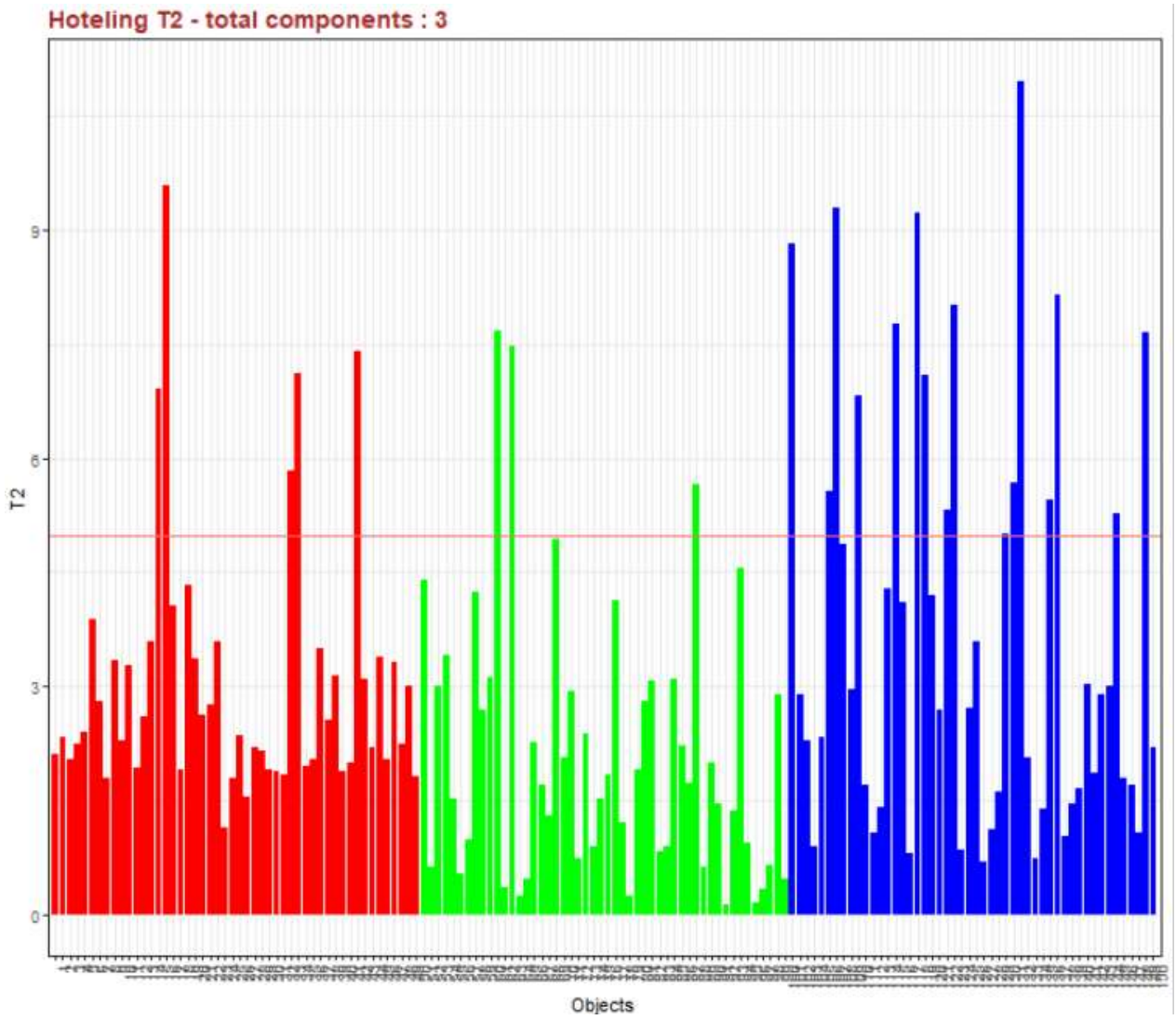


T^2 Hotelling Plot

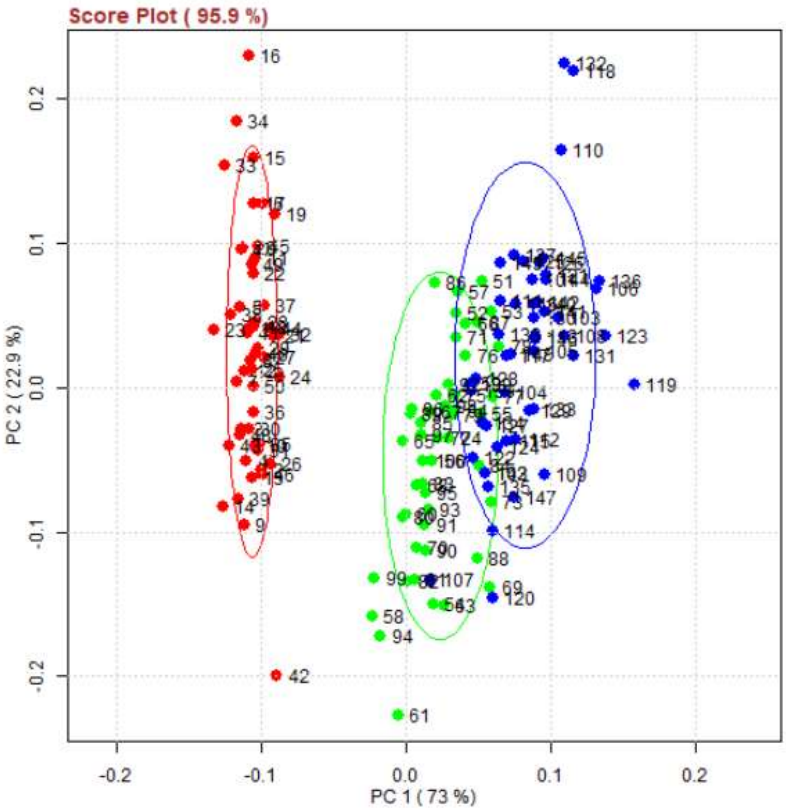
The Hotelling statistic measures the compatibility of one object with the PCA model. It is a positive number, greater than or equal to zero. It is the distance from the center of the hyperplane to the projection of the observation onto the hyperplane. It is calculated by the formula:

$$T^2 = x^T P(\Sigma_a)^{-2} P^T x$$

Where *P* is the loading matrix Σ_a is a diagonal matrix with the non-negative real eigenvalues corresponding to the selected number of components in the PCA model.
One observation that projects onto the model's center has Hotelling T2 null (i.e. the observation where every value is at the mean). The T2 Hotelling is distributed according to the F-distribution and is calculated by the software as soon as the significance Alpha is defined by the Slider. For example, we can calculate the 95% confidence limit for T2, below which we expect, under normal conditions, to locate 95% of the observations. The graph is the calculation for the Iris Data Set.



The T2 Hotelling statistic is closely related to the ellipse shown in the Score Plot. Points inside this elliptical region are within the confidence limit fixed. The numerical values of the T2 Hotelling can be achieved by pressing the *Download* button as CSV file.


Score Plot (95.9 %)
PC 2 (22.9 %)
PC 1 (73 %)
0.2
0.1
0.0
-0.1
-0.2
-0.2
-0.1
0.0
0.1
0.2

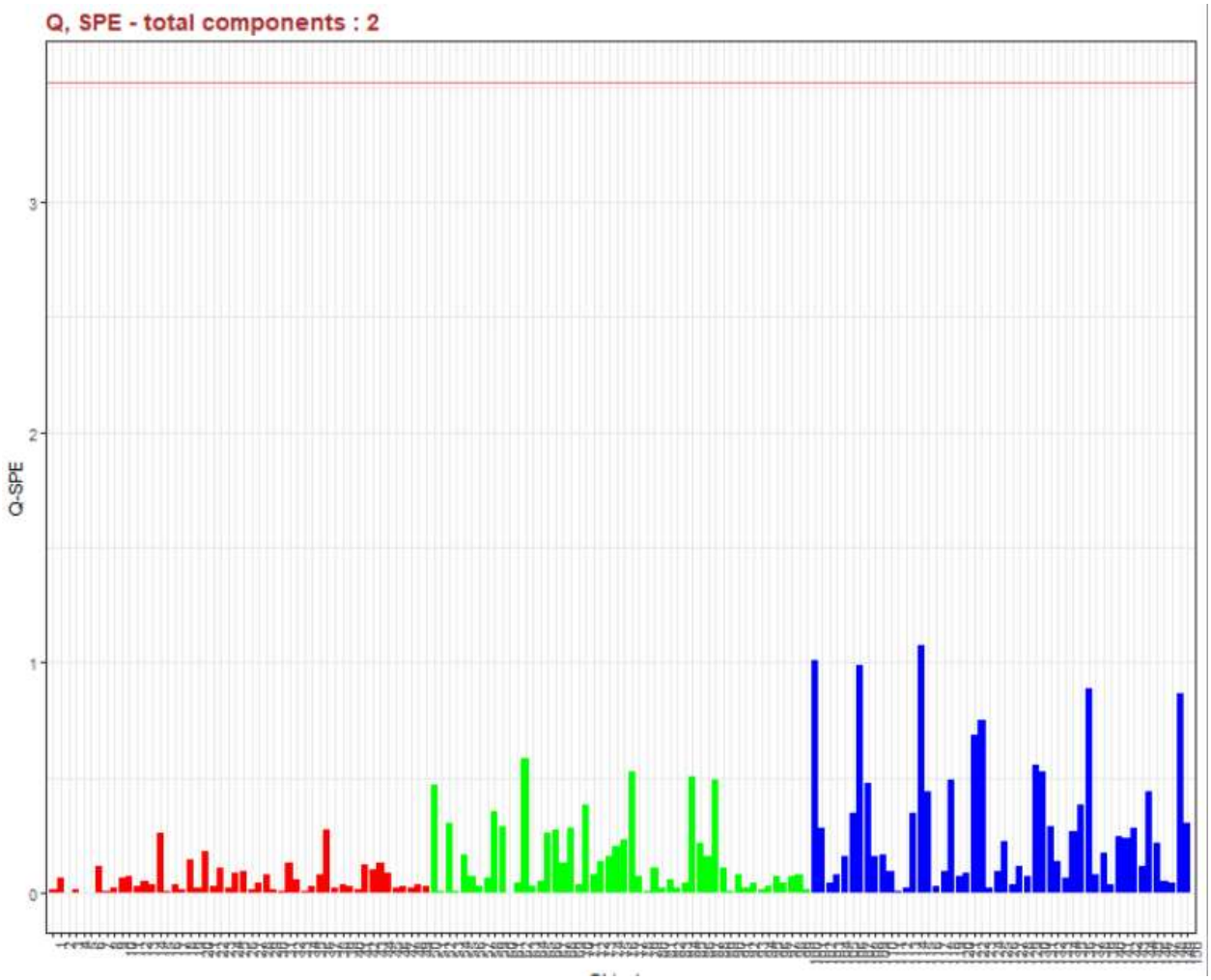
Squared Prediction Error (SPE) or Q Statistic

The portion of measurement space corresponding to the lowest components (eigenvalues) can be monitored using the Squared Prediction Error or Q statistic. The Q statistic does not suffer from over-sensitivity to inaccuracies in the smaller singular values and it is associated with noise measurements. The upper confidence limit of the Q can be computed from its approximate distribution:

$$Q_\alpha = \theta_1 \left(\frac{h_o c_\alpha \sqrt{2\theta_2}}{\theta_1} + 1 + \frac{\theta_2 h_o (h_o - 1)}{\theta_1^2} \right)^{\frac{1}{h_o}}$$

$$\theta_i = \sum_{j=a+1}^{m} \lambda_j^i \qquad h_0 = 1 - \frac{2\theta_1 \theta_3}{3\theta_2^2}$$

where Cα is the value of the normal distribution with α level of significance. A violation of the threshold would indicate that the random noise has significantly changed or unusual event has occurred that had produced a change in the covariance structure of the model.
The significance α can be set as usual with the help of the Slider and the values of the Q statistic are downloaded by the correspondent button. The next figure shows the distribution for each object of the Iris Data Set. Consider that more components are added more the Q values are reduced because little variance is left over**. The Q statistic is a key parameter for outliers (faults) detection.**

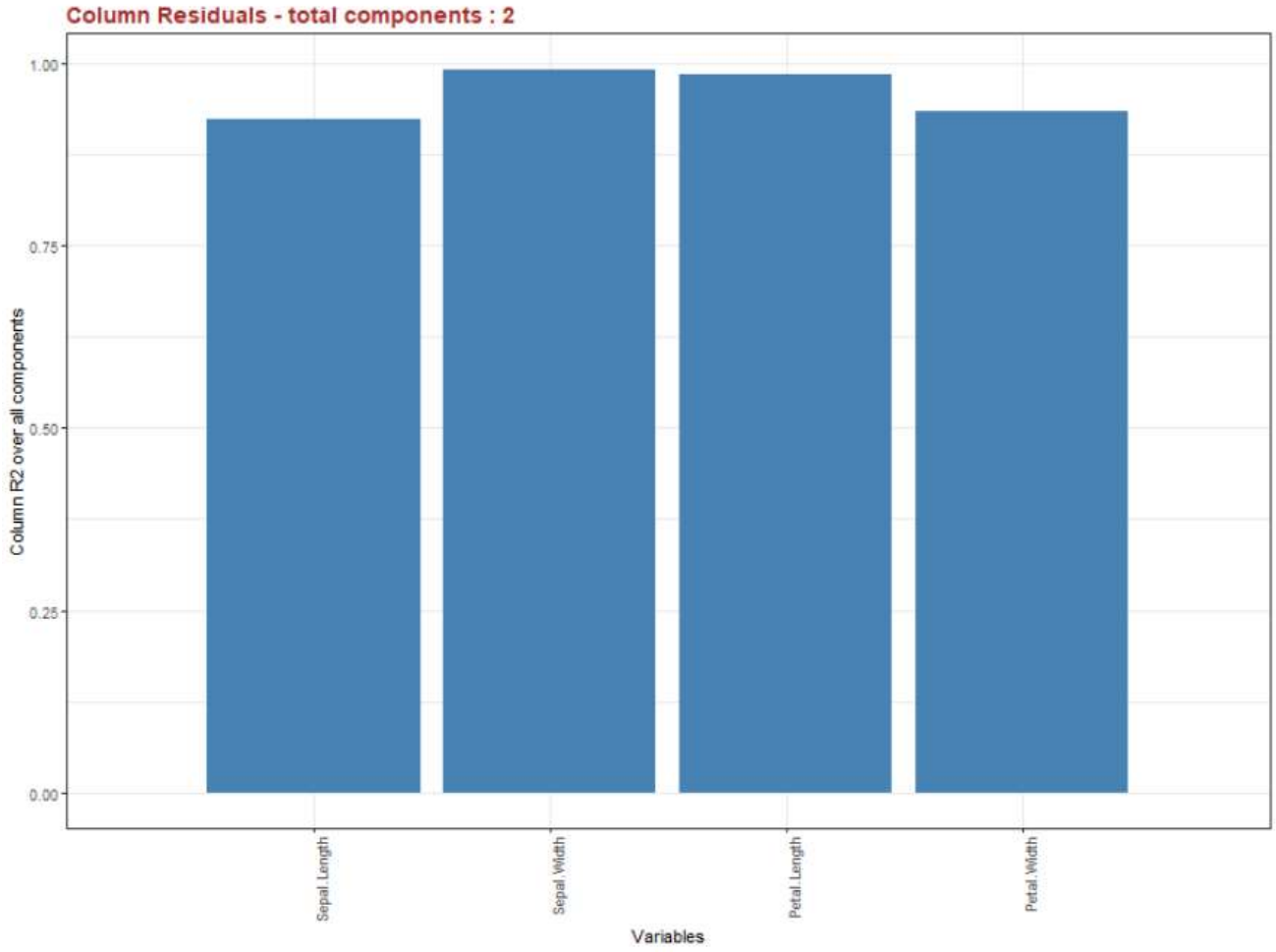


PCA Column Residuals

The Column residuals are calculated by the matrix of residuals and summing elements by column. The formula can be written as follows:

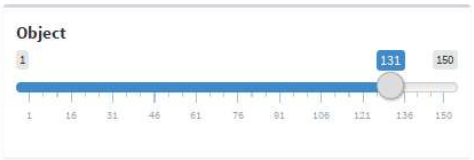
$$R^2_{X,k} = \frac{SS(\underline{x}_k) - SS(\underline{e}_k)}{SS(\underline{x}_k)} = \frac{SS(\text{explained})}{SS(\text{total})} \qquad 0 \le R^2_{X,k} \le 1$$

The plot shows the residual values for each variable. Higher is the value higher is the importance of the variable in the model. Variables with a low level of explained variance are mainly prone to noise. The next plot shows the case of Iris Data Set. Consider that as close we are to a full explanation of the whole variance (number of components equal to the number of the variables) closed we are to the unity for all the variables.

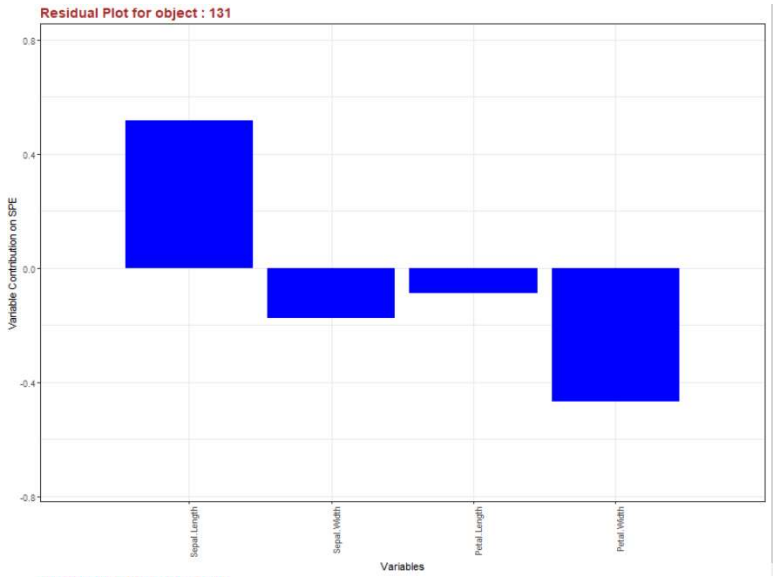


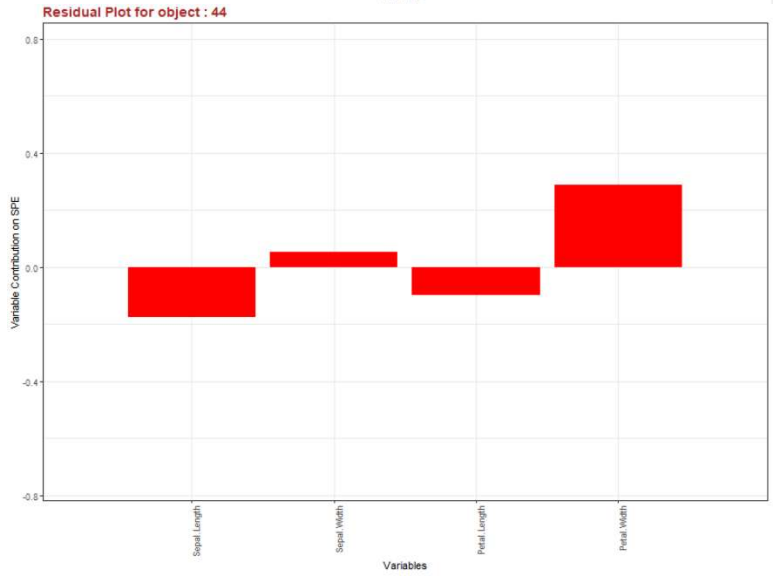
SPE Contribution Plot

The SPE Contribution Plot is useful to split the Q statistic of an object in the effect of each variable. Remember that in case of outliers, that do not stay on the model plane, the Q value is relevant. It is important to understand which Variable contributes most to the high value. It is possible that Variables with the higher contribution plot are those variables that change. The plot depends on the object that is selected by the Slider. The limits of the Slider are the number of objects and it is automatically evaluated by ***Deducit***. Another important option is the Fix Scale. If the check box is selected, the scale of the bar if fixed on the values of all the objects. In this case, plots of different objects can be compared because they are drawn with the same unit. This is key to see if some Variables are differently influencing objects.



The contribution is quite easy to understand and draw as shown by the Iris Data Set for two objects. All the values of the contribution plot can be downloaded as a CSV file by the button.

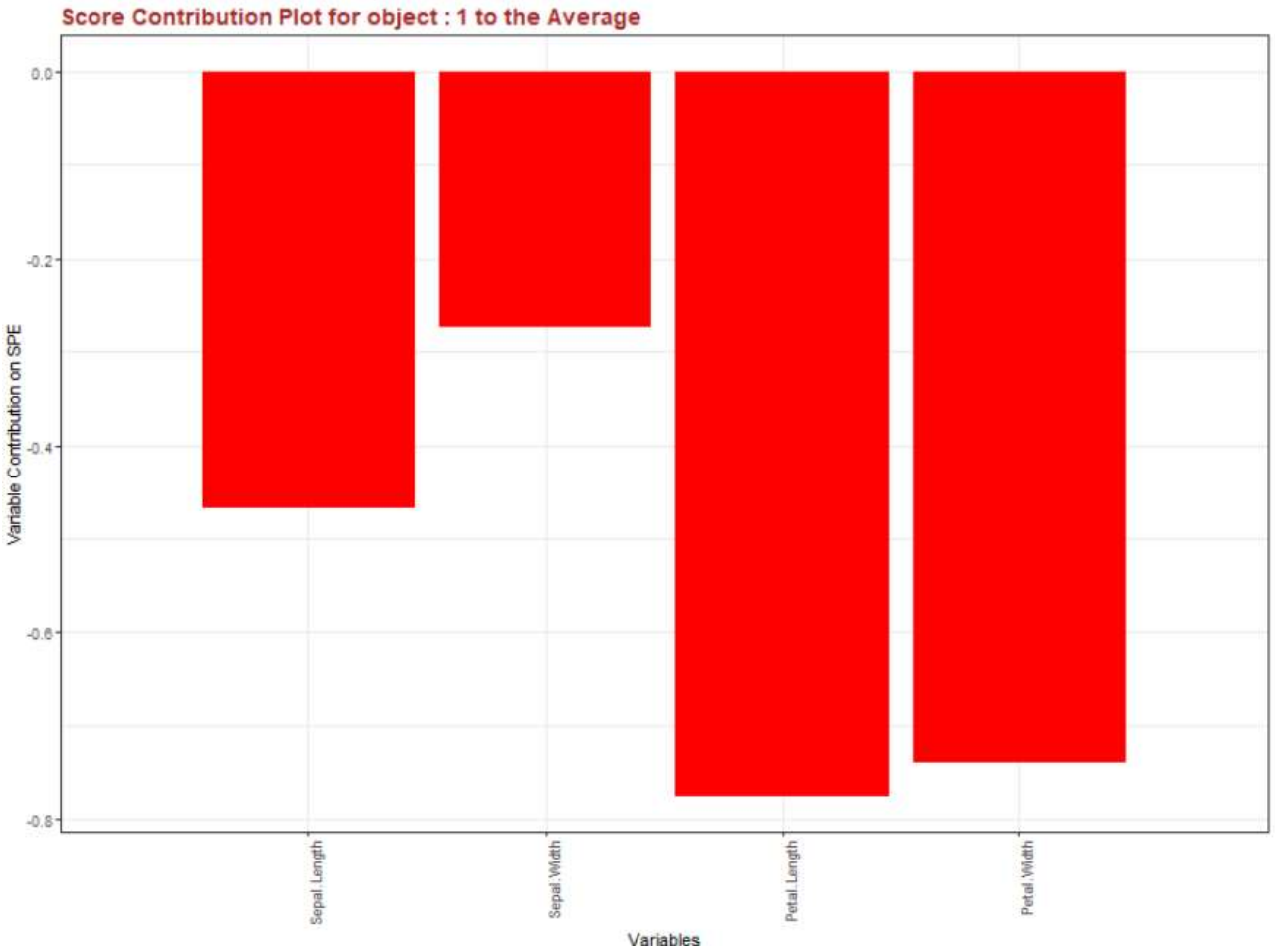




Score Contribution Object to Average

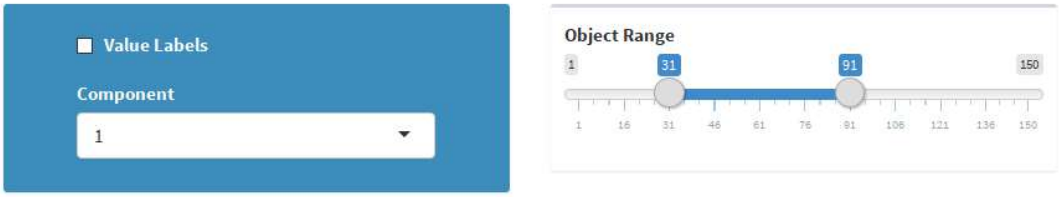
The Score Contribution Plot is required when we would like to understand why one object is far away from the center of the plane (i.e. from the average of variables). Basically, this plot splits the score of one object in the contribution due to all Variables. The object can by chosen by the Slider. Of course, it is also necessary to fix by the left Combo box which latent Variable (score) to analyze. In the next figure, I selected the first object in the direction of Component 1.

The result is a bar plot where the weight of each Variable is shown. Higher is a bar on a variable, higher is the difference of variable vs the average of all the other objects. This highlights why an object can be an outlier in the Data Set. The next figure shows results in the Iris Data Set. The first object is different mainly on the values of Petal both below the average.

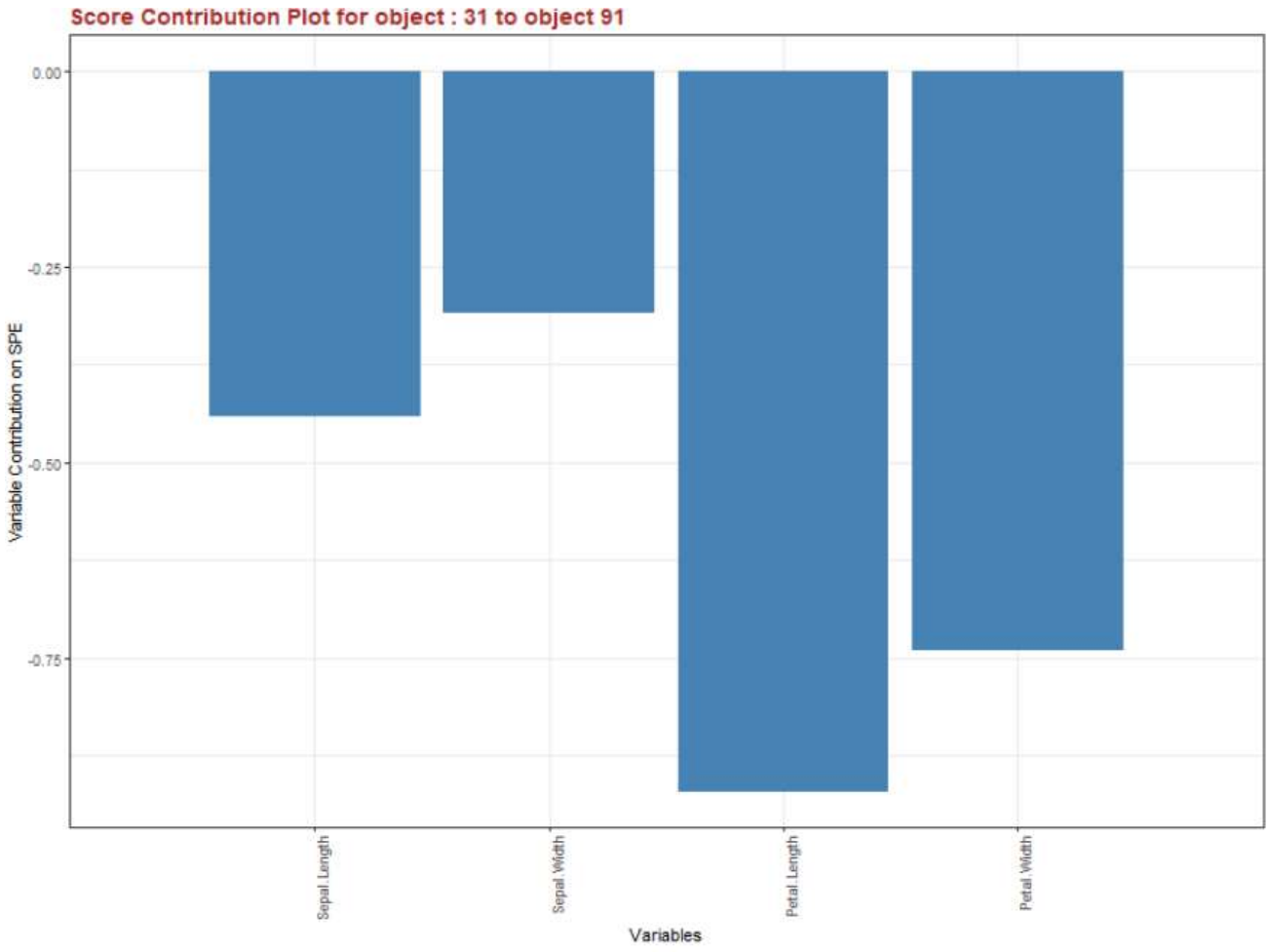


Score Contribution Object to Object

This Score Contribution plot is required when we would like to understand why two objects (points) are different in the latent space. Basically, this plot splits the scores of each object in the components due to the variables and compares values. Objects can by chosen by the Double Slider moving the right and left ends. In this way, it is possible to combine each object to each other's. Of course, it is also necessary to fix which latent Variable (Score) to analyze by the left Combo box. In the next figure, I selected the 31-th object comparing to the 91-th object in the direction of Component 1.

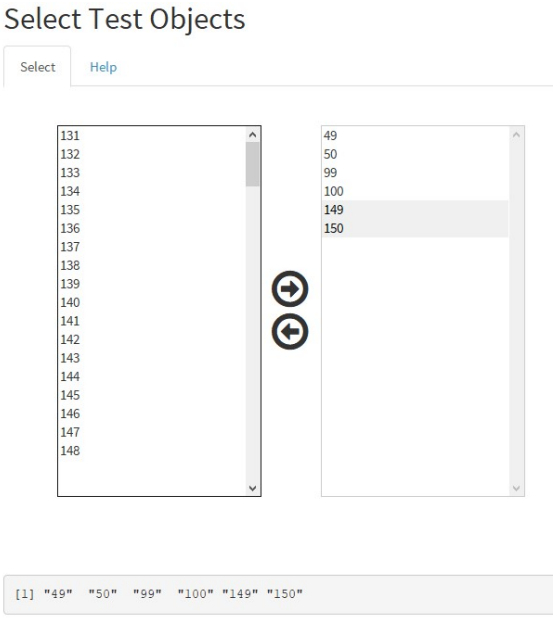


The result is a bar plot where the differences on each variable are shown. Higher is a bar on a variable, higher is the difference between the values of the variable in the two objects. This highlights the object differences especially when one of the two objects is an outlier of the Data Set. The next figure shows the results for objects in the Iris Data Set. The two objects are different mainly on the values of Petal.

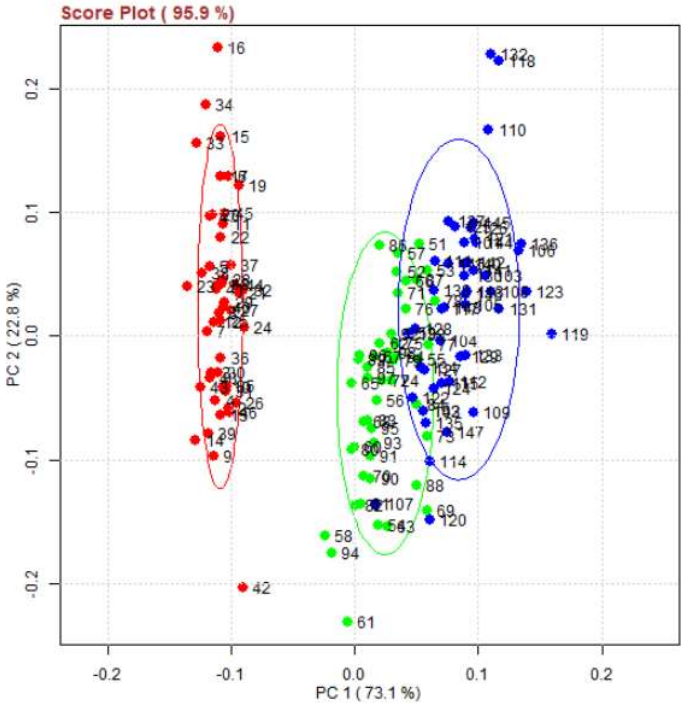


Add Test Data to PCA Space

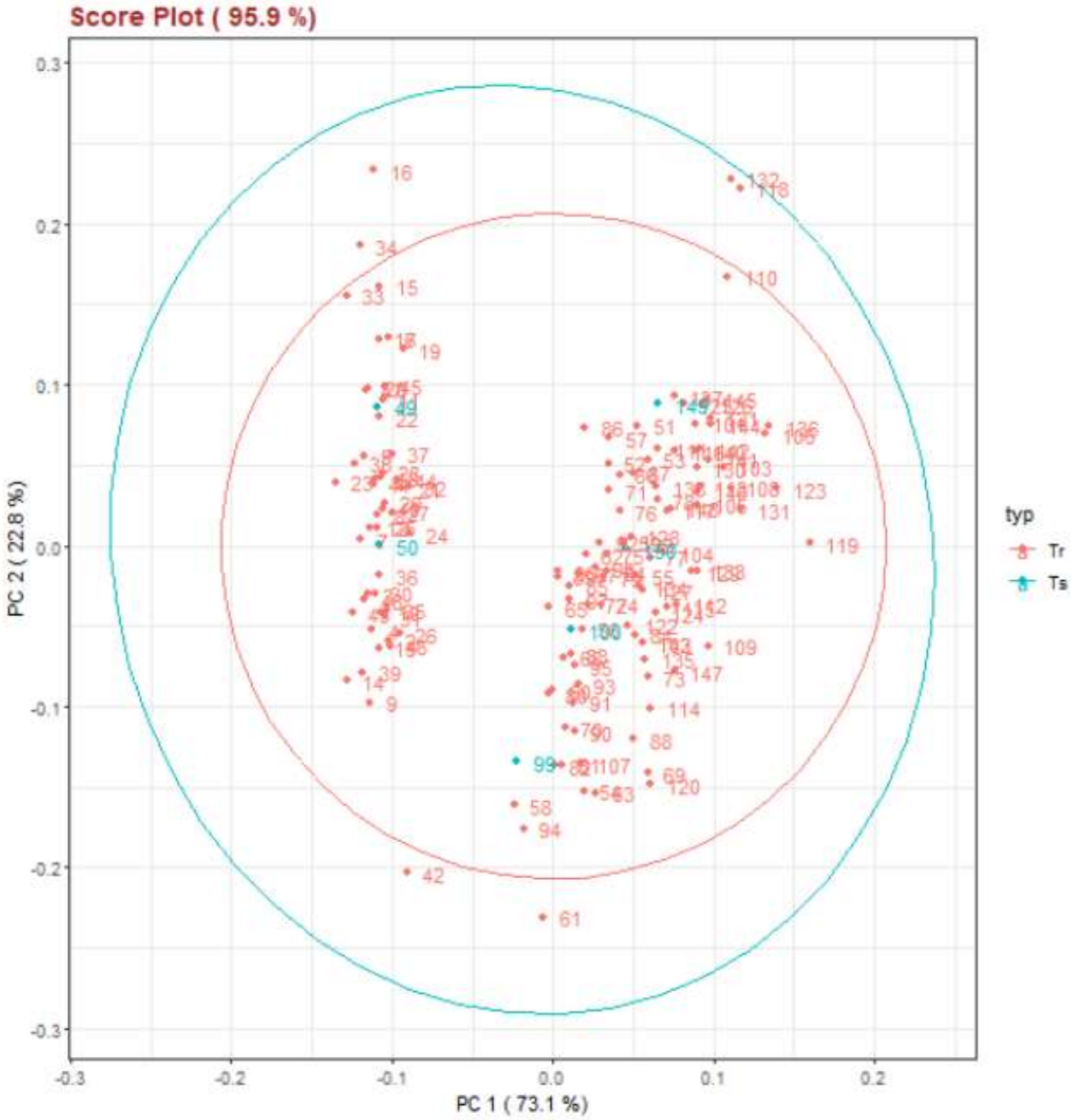
This menu is used when you want to add new data in the PCA space. New data must have the same format of the original Data Set (i.e. the same variables) but they were not previously used in the model construction. This is typical when you have some data that you would like to use as a Test Set (see Training/Test Set menu). **The new data are centered and scaled using the statistic of the original set and they are then projected in the latent space.** You normally expect the coincidence of the new data with the old one, i.e. the Test Set superimposes to the Training Set. Vice versa the model is not representative and some further work is required. As an example, let's consider the Iris Data Set and make Test objects the last two objects of each Species. This is easily done in the Training/Test menu as shown below:



After the selection, you can build the PCA model as usual and get the standard Score Plot. In the plot, Tests objects are not present because they are not used in the calculation. Please note that since we removed just few objects over a big number of elements the Score Plot is almost unchanged



When you select the Additional Data menu in PCA, you get the next plot. The old (Training) values are in red and the new ones (Test) are in blue. Colors are chosen to stress the evidence between new and old, so the different colors for Groups is not available for this plot. However, looking at the position of the Test points in comparison with the unchanged position of the Training points, it is evident that all the new objects are close to the cluster of their Groups. This validate the model at 3 components. On the plot, also the ellipse of the whole Training and Test Sets are drawn. We expected two ellipses are very close and similar as in this case. If you don't like ellipses you can deselect the Check box on the top left. Vice versa you can change the significance of the two ellipses scrolling the alpha Slider.



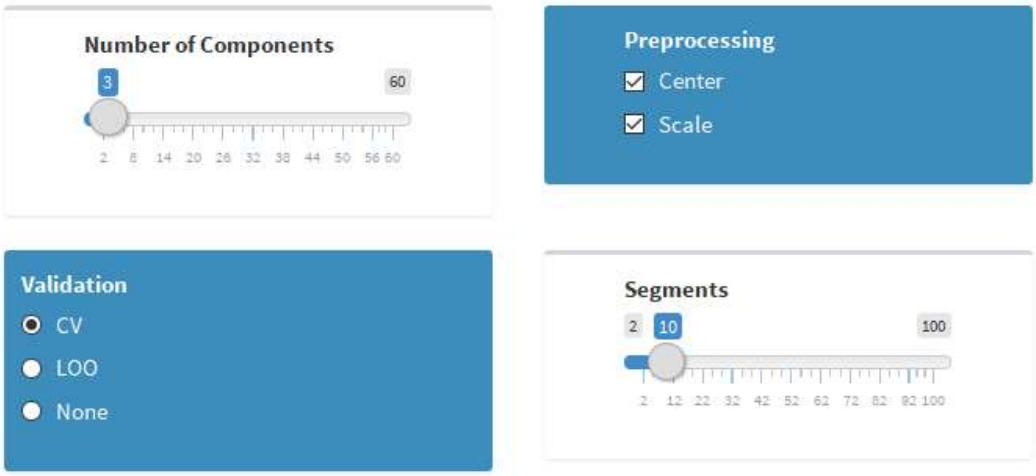
You can get the projection of the Test Set in the CSV file just pushing the *Download* button.

Partial Least Squares Regression

The Partial Least Squares (PLS) regression is used to find fundamental relations between two matrices (X and Y) said Predictor (or Factor) and Response Matrices. The PLS model tries to find the multidimensional direction in the X space that explains the maximum multidimensional variance direction in the Y space. The X space is always multidimensional (more than one variables), the Y space can be univariate (only one response) or multivariate. In case of a single response the PLS model is indicated by the initial PLS1, in case of two or many the model is called PLS2. In any case PLS regression is particularly suited when the matrix of Predictors has more variables than objects and when there is multi collinearity among the Xs.

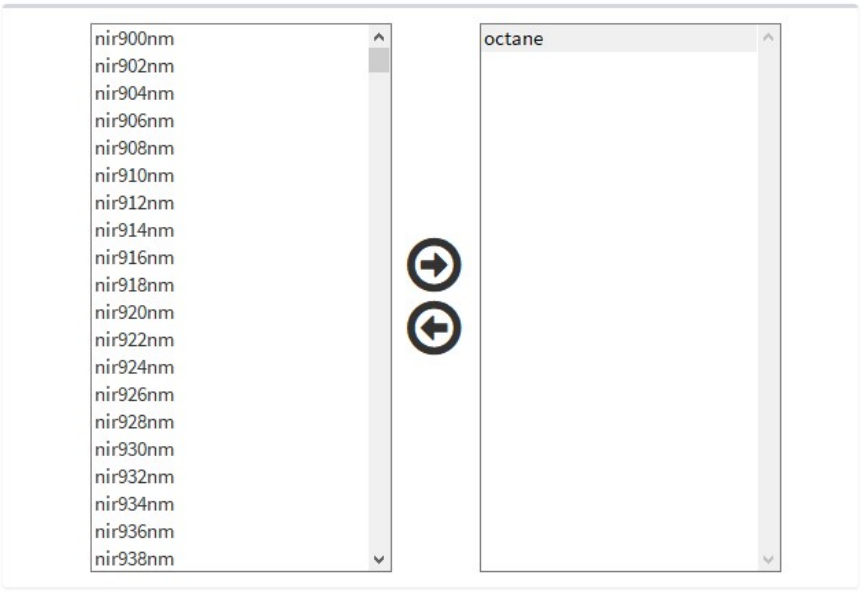
To start modelling with PLS, you have to provide some initial parameters that will be fine-tuned with the prediction results. This is done in the Setting Menu as shown below. You need to fix the number of principal components that need to be considered. The minimum value is 2 and the maximum is given by the minimum value between the number of objects (rows) and Predictors (X part of the columns). You have to decide between the two limits which value is more suitable for your purpose. If you do not know, please go first to the menu PRESS to have a criterion. To fix the value of the principal components you have the Slider on the left. It is very common, changing the value many times before to fix it and each time ***Deducit*** repeats the calculation. The value of the principal component considered is shown in almost all the results and plots, so you don't need to remember it. Finally, if the number of components is maximum, all the variance is explained and some plots are useless.

PLS is quite sensitive to over fitting. I made by default a cross-validation criteria for most of the model outputs. The box on the left lets you select which method chose. The standard is the CV criterion splitting the objects in 10 segments and leaving one at turn out of the calculations. Results are averaged on all the segments. Another criterion is the LOO (Leave-One-Out) were only one object at turn is left out of the calculation, especially useful with few data. Finally, there is also the possibility to do not use validation (None choice). Only in the case you have chosen the standard CV validation you can also modify the number of segments in the splitting part. Choose a number proportional of the number of your objects. The default is 10 and the minimum is normally 3.



After the model parameters are fixed, you have to let ***Deducit*** know what are Responses and what are Predictors. This task is done by the use of the next list boxes. On the left box are listed all the Variable names active because at the beginning they are all considered Predictors. The figure shows the case of gasoline Data Set in the example menu. To move one Variable, e.g. octane, you have to highlight it by right clicking on it

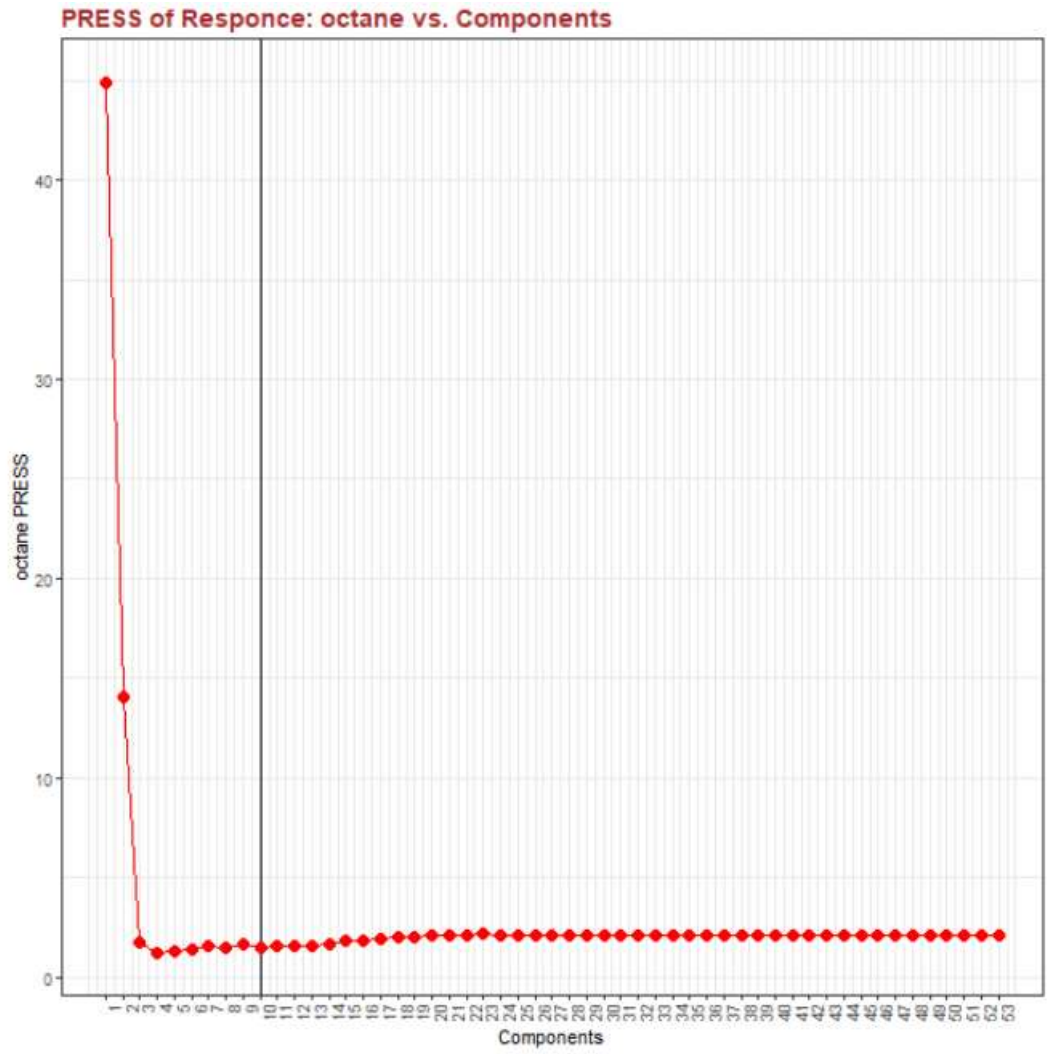
and then press the right-hand arrow. The name of the selected variable moves from the left to the right box (box of the Responses). You can move the variable back to the Predictor list, selecting the variable and clicking on the left-hand arrow. It is possible to select more than one variables at the same time and move them back and forward all together. The selected variables chosen as Response are also printed in the Text frame, just to let you know. **Do not forget to select some Response, because PLS cannot the done without this key information.**




```
[1] "octane"
```

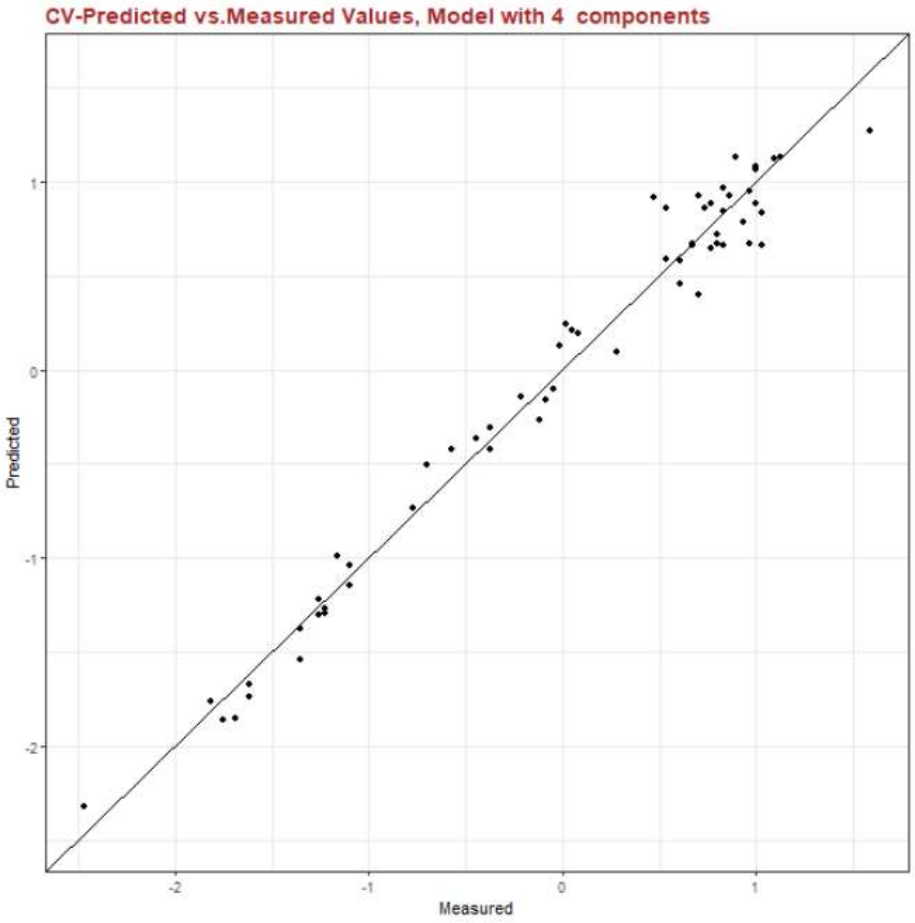
PRESS Plot

The Predicted Residual Error Sum of Squares (PRESS) is a form of cross-validation in regression analysis to provide a summary measure of the fit of a model to a sample of objects that were not themselves used to estimate the model. The PRESS is very useful in PLS to select the number of principal components that need to be considered. In the figure it is shown the PRESS Plot for the gasoline Data Set used to predict the octane vs all the other factors. As usual the plot draws a decreasing curve in red vs the maximum number of components. The vertical line (at 9 in this case) represents the actual choice of the component number now selected. The PRESS is calculated using the component number from the minimum (2) to the maximum (53 in this case). The optimal choice is when the vertical line crosses the red curve at its minimum or close to it. In the picture, the vertical line should be shifted to 3. This means that a model with 9 components over fit the data and a simpler model with only 3 components would be enough.
If your Data Set has more than one Response, a graph for each of them is available. You can select the Response through the Combo selector at the top. In case of multidimensional Responses, the criterion is to take the maximum values among all.
The data of the PRESS plot can be downloaded as CSV file by the button.

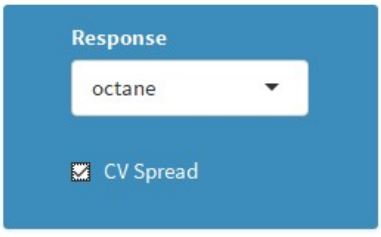


PLS Fit and Prediction

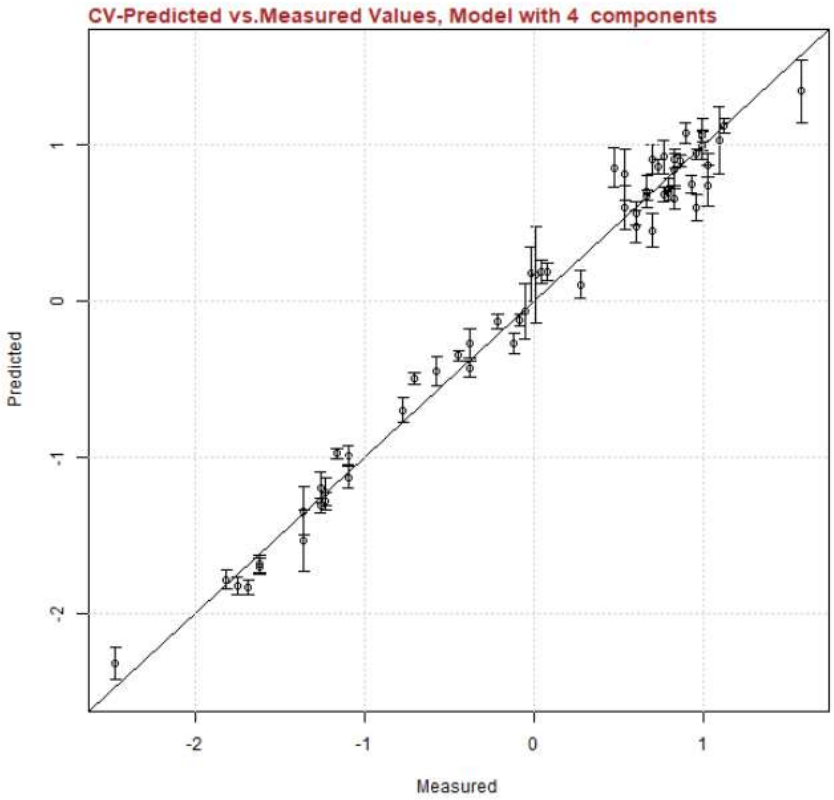
The Prediction of a PLS model is quite straightforward as soon as the model parameters are defined. The prediction is always cross-validated unless the Validation option is set to None in Setting Menu. The prediction data are visualized by a plot as shown in the figure for the gasoline Data Set. A square plot graph Predicted vs Measured values is displayed. The PLS prediction is as good as data are on the plot diagonal, i.e. coincidence between predicted and measured numbers.



Of course, there is a plot for each Response and the plot changes depending of the top Combo selector at the top.

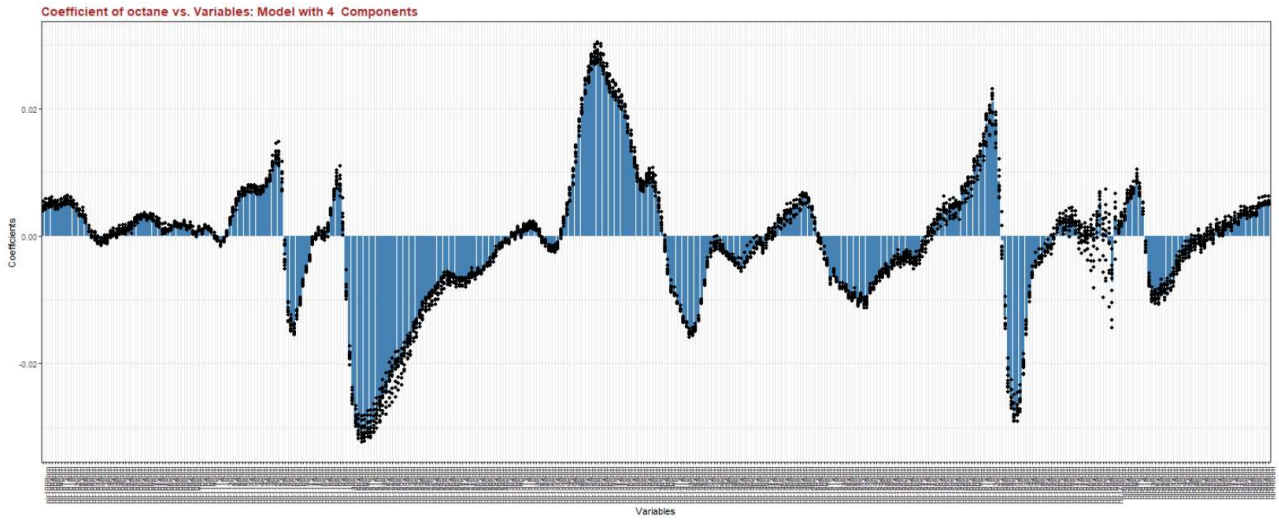


Another interesting possibility is to see the spreading that the Cross Validation has evaluated. Selecting the Check Box of CV Spread, the prediction is repeated several times on random bases and the predictions of each iteration is averaged. In the case of the example the Response is shown in the next figure.
Now, points are replaced with error bars. If the bars cross the diagonal, the prediction could be reliable, if not, the prediction retains some model error. Objects with larger error bars are the most interesting since are the outlier of our regression.

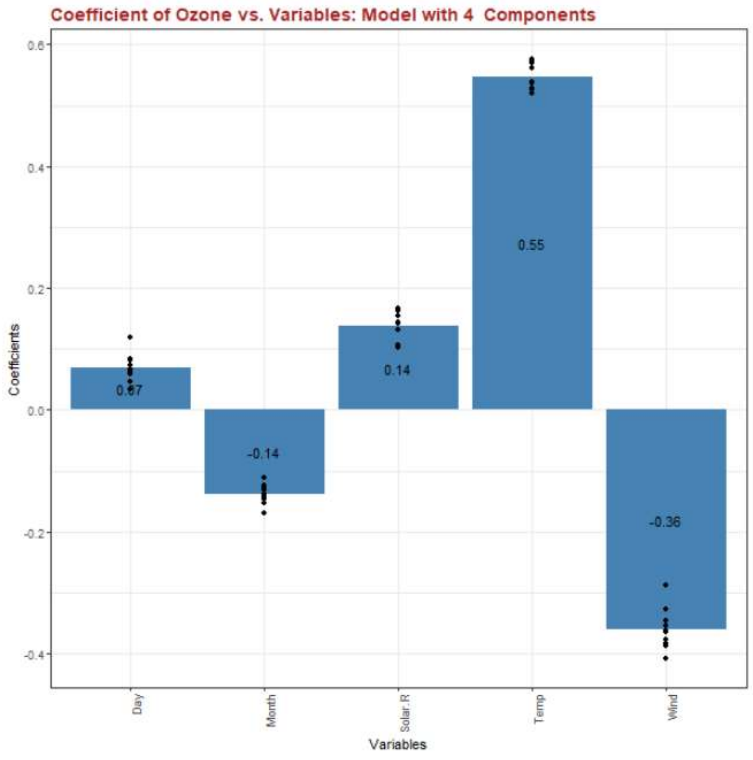

CV-Predicted vs.Measured Values, Model with 4 components
Predicted
Measured
-2
-1
0
1
-2
-1
0
1

PLS Coefficients

The menu of Coefficients allows to visualize the coefficients of the PLS regression. There is one coefficient for each Predictor, that makes a vector and there is a vector for each Response. In case of multi Response Coefficients are organized in a matrix. It is possible to visualize the coefficients of each Response selecting it in the Combo selector at the top. The following graph shows the coefficients of gasoline prediction from the example Data Set. The bars indicate the value of the coefficient referred to the Predictor on the x axis. Black points indicate instead all the values of the same coefficient evaluated with the split of CV. Their spread on the y axis indicate how reliable is the coefficient estimation. As you see, the spreading changes depending on the variable considered.

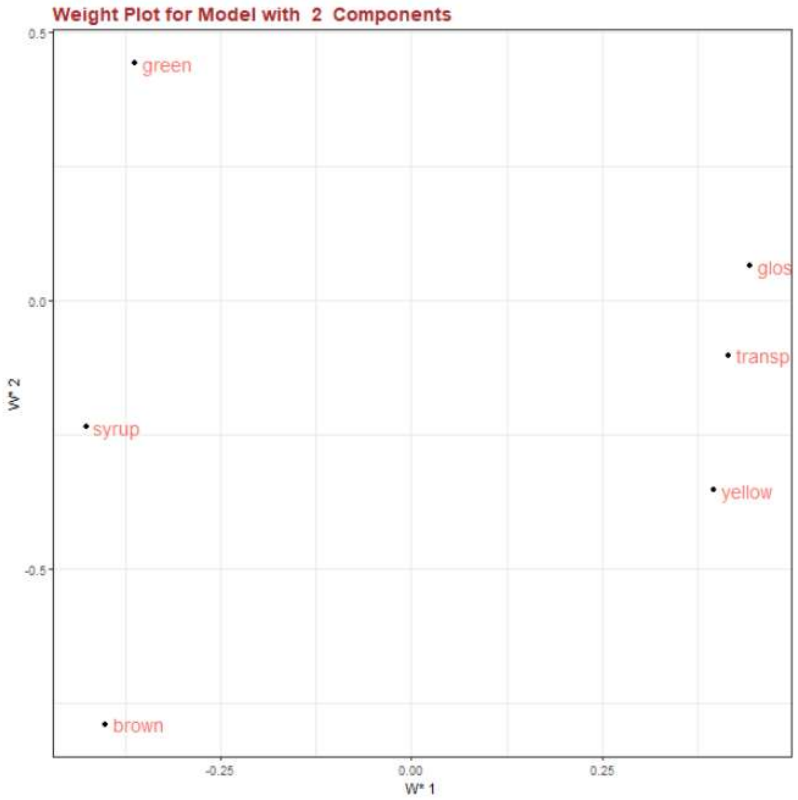


I report the Coefficient plot also for the air quality Data Set were the Ozone is predicted from the other variables. In this graph it is more evident the distribution of the point around the average bar values. Through the *Download* button is it possible to get the CSV with the averaged coefficient for each Response.



PLS Predictor Space Weight

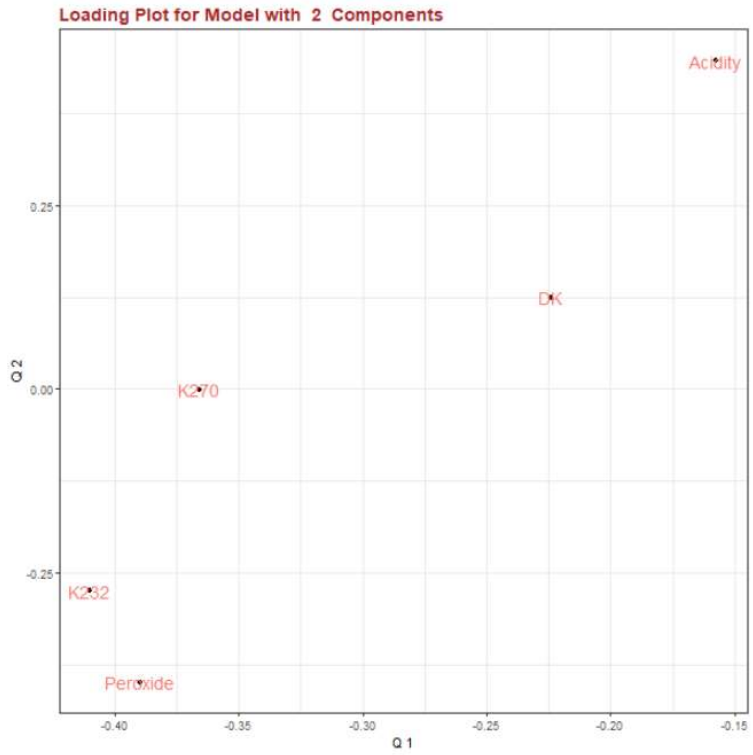
The Predictor Weights Plot describes how the Predictors are related to the scores (just as the Loading of PCA related Variables to Scores). The plot of the Predictor Weight is represented as a point for each Variable in a plane of two Principal Components. The figure shows the case of oliveoil Data Set from the examples. The plot is interpreted in the same way as PCA loading plot. For instance, we see that the first component, clearly discriminate between two groups of Predictors, one on the left and the other on the right side of the axes.



Note: the loadings of the Predictors describe variation in the X space. It is used to calculate the SPE values and to remove explained variation from X after each component is calculated. The loadings are not generally analyzed although they are used. Weights W describe the covariance relationships among X and Y variables in the deflated X matrix. **Understanding the relationships among the original variables in X is more useful, so we prefer to use the calculated weights in W*.**

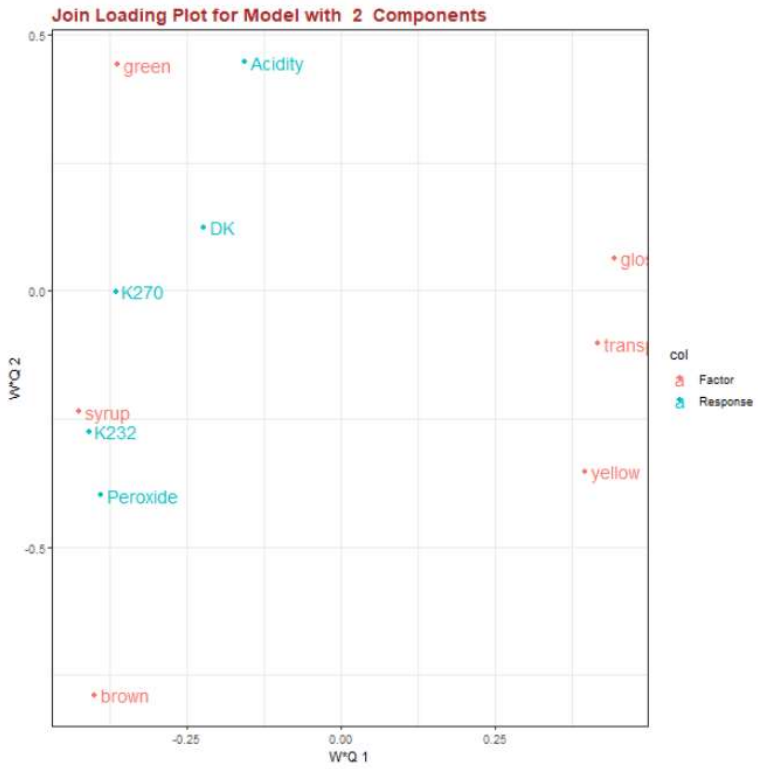
PLS Response Space Weight

The Response Weights Plot describes approximately how the Responses are related to the Scores. This is a little bit different than the Predictor or PCA cases. This plot is helpful only in the case of multi responses (PLS2 case) otherwise a useless single point is drawn. From the plot is possible to see how Responses are correlated with scores and with themselves. This information is very helpful in case of multi-dimensional optimization to identify which targets go together and which are opposites. The graph shows the output of the oliveoil Data Set example. Acidity is inversely correlated to the peroxide content as expected.



PLS W*Q Plot

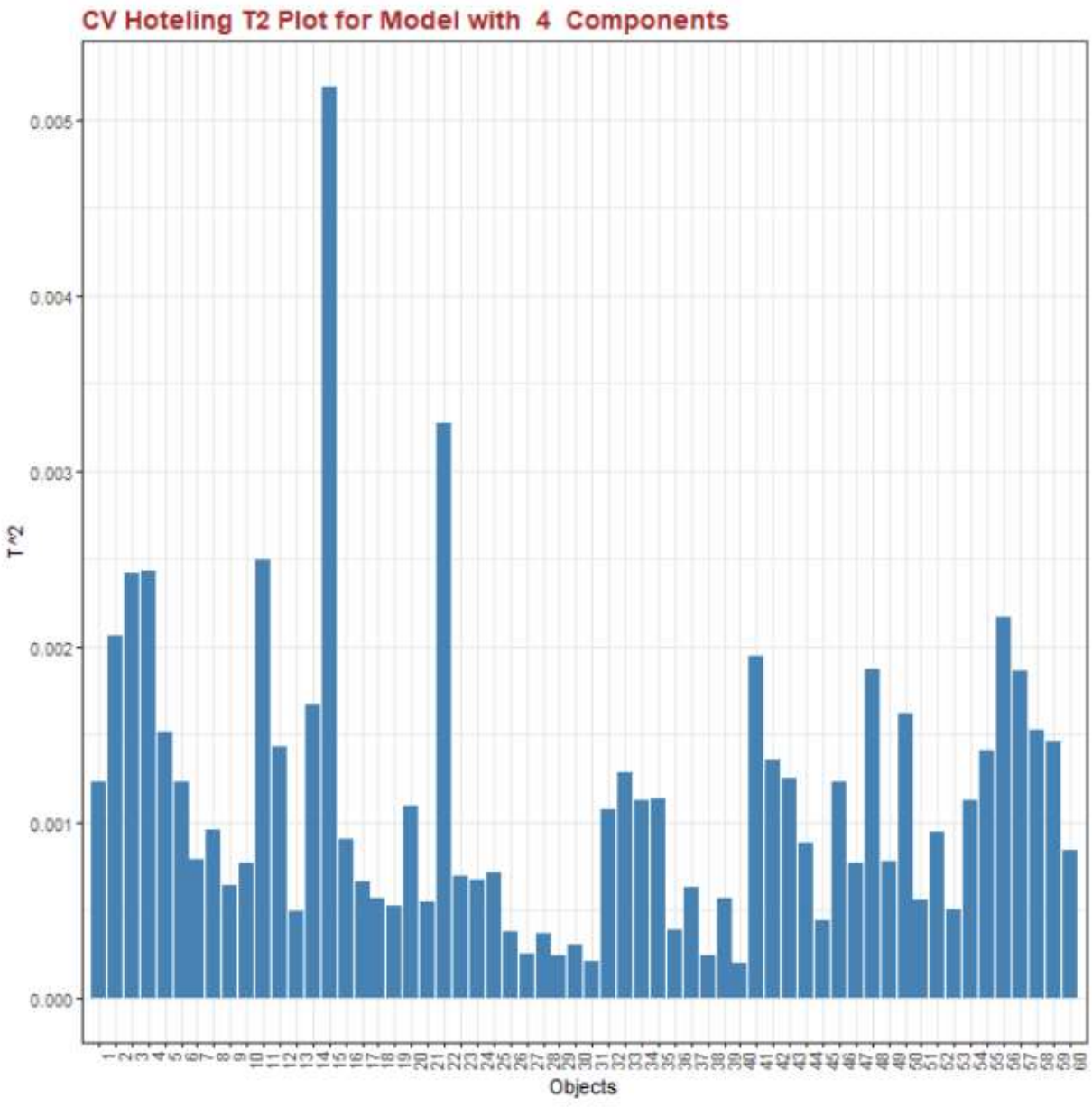
Although this plot has not a specific name, it is very helpful in PLS analysis results. It describes the relationships between Predictor X and Responses Y, and it is made by overlaying the two weight plots for each latent variable. It works because the two weights W and Q share a common link, i.e. the latent variable scores. From the plot it is very easy to understand which Predictor is more closely related to a Response. If the two points are close it means that the two variables are strictly dependent. With many variables this task is not very easy to absolve but it becomes trivial thanks to this plot. The next plot shows the output for the oliveoil Data Set.



It is evident that Syrup and K23 are very much close in the plot and in the data. The data of these plot can be downloaded through the *Download* button as a CSV file.

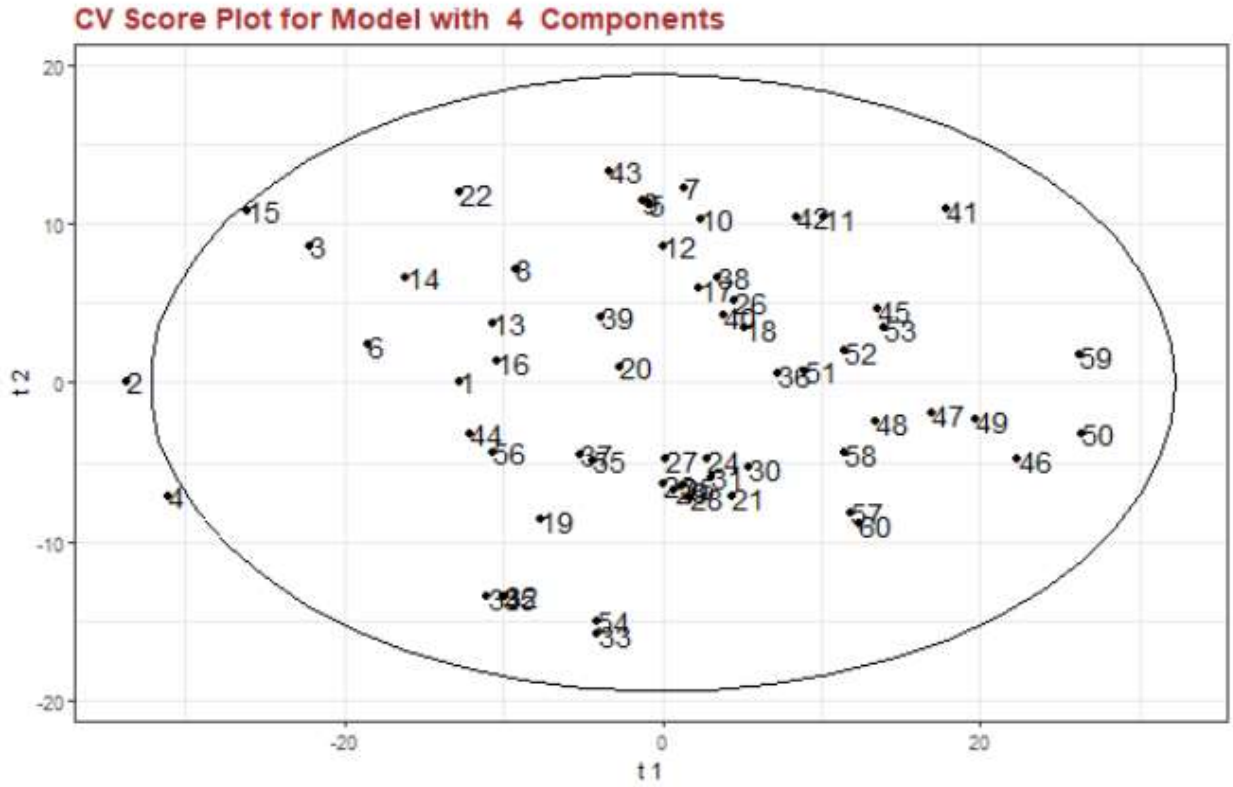
PLS T2 Hotelling Plot

The Hotelling statistic measures the compatibility of one object with the model like in the PCA case (please refers to that menu for details). In case of PLS model the plot is made by a series of bars one for each object used in the Training set. In case of the gasoline Data Set the plot is below. Higher is the bar higher is the probability that the object is outside the structure of the model. The same object could be an outlier also in the prediction plot.



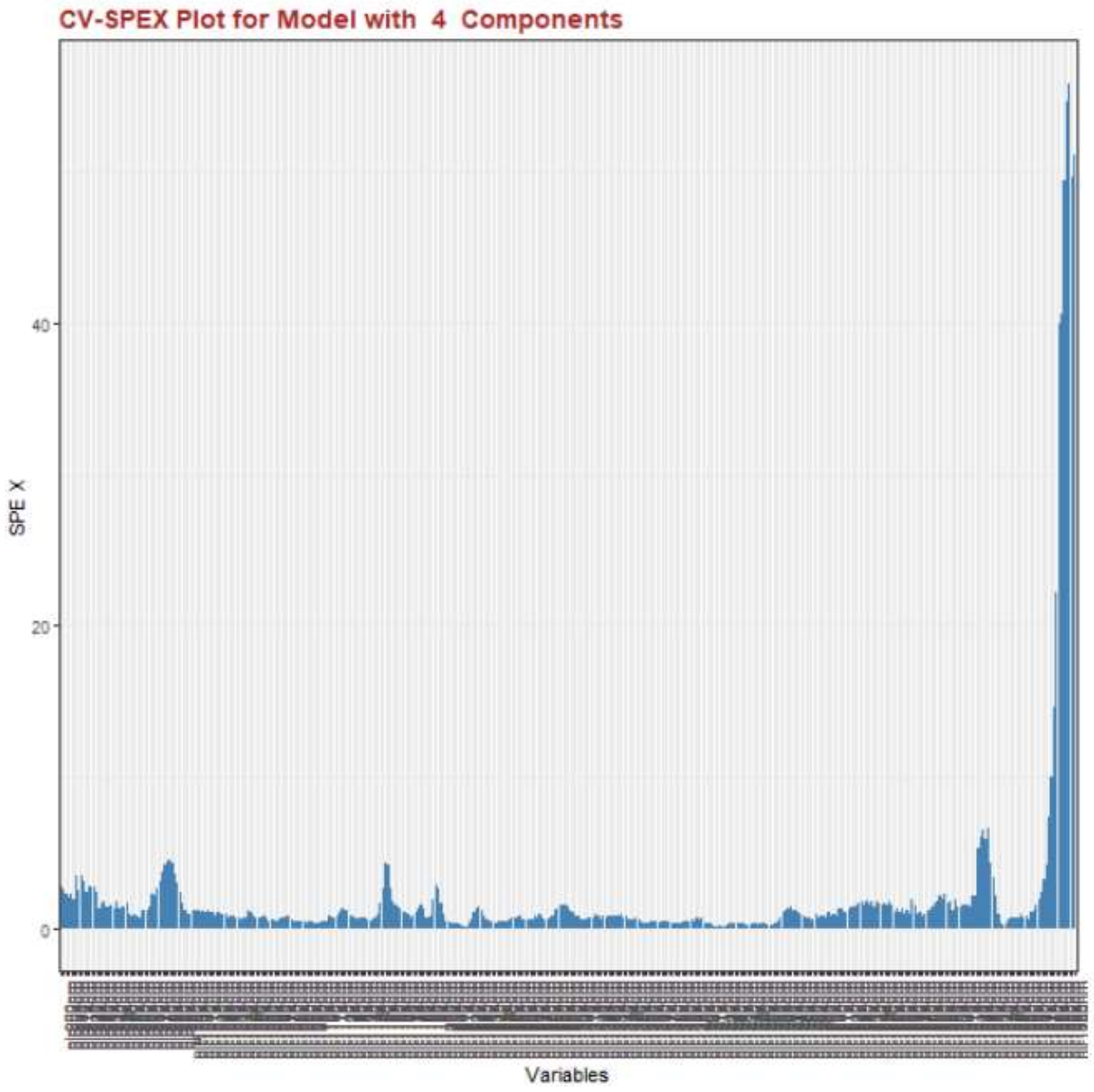
PLS Score Plot

The Score Plot in PLS describes the distribution of the objects in the X -space, however it is not so important like in the PCA models. It is commonly used to check if points do not cluster in some areas and if they are well distributed in the space. As in the PCA case, several combinations of the latent variables can be made selecting them by the two Combo selectors. In case of gasoline Data Set the plot is shown below.



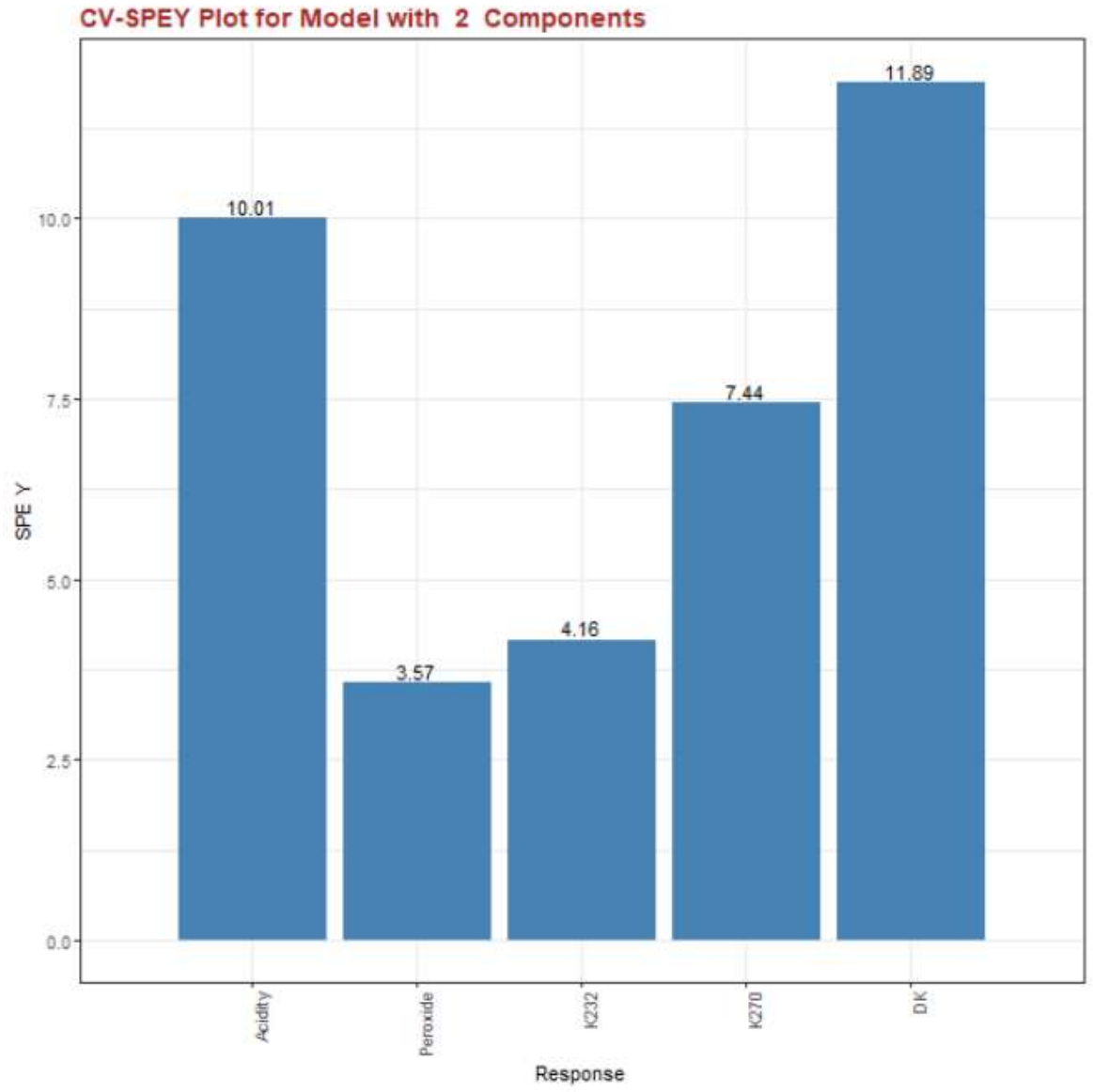
PLS SPEX

The Squared Prediction Error on the Predictor matrix X is a complementary tool for PLS models. The SPE values measure the distance of each object from the plane drawn by the X -space. The SPE is the error off the X -space plane, i.e. it is a multivariate residual complementary to the correspondent score value. The plot is very helpful to highlight which objects do not follow the model correlation and are possible outliers. The next plot shows the results for the gasoline example Data Set.



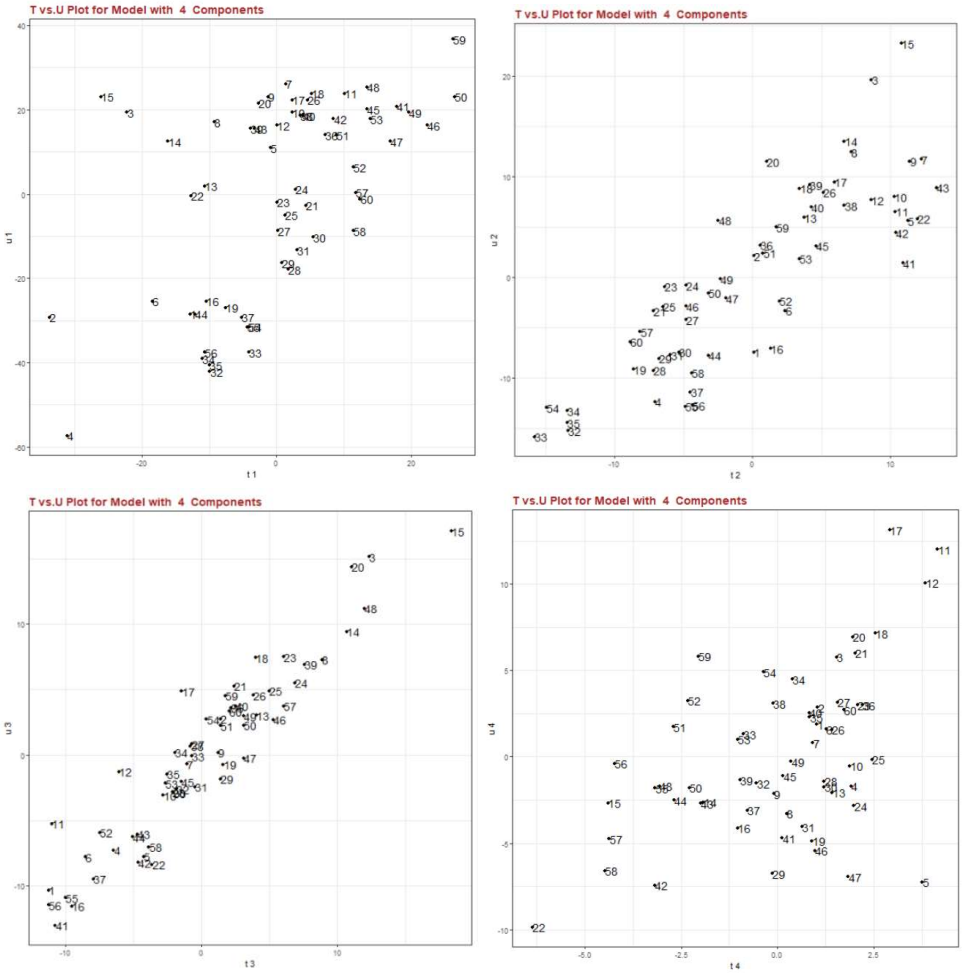
PLS SPEY

The Squared Prediction Error on the Response matrix Y is a complementary tool for PLS models. The SPE values measure the distance of each object from the plane drawn by the Y-space. The SPE is the error off the Y -space plane. The plot is very helpful to highlight which objects do not follow the model correlation and are possible outliers. The next plot shows the results for the oliveoil example Data Set.



PLS T U Space plot

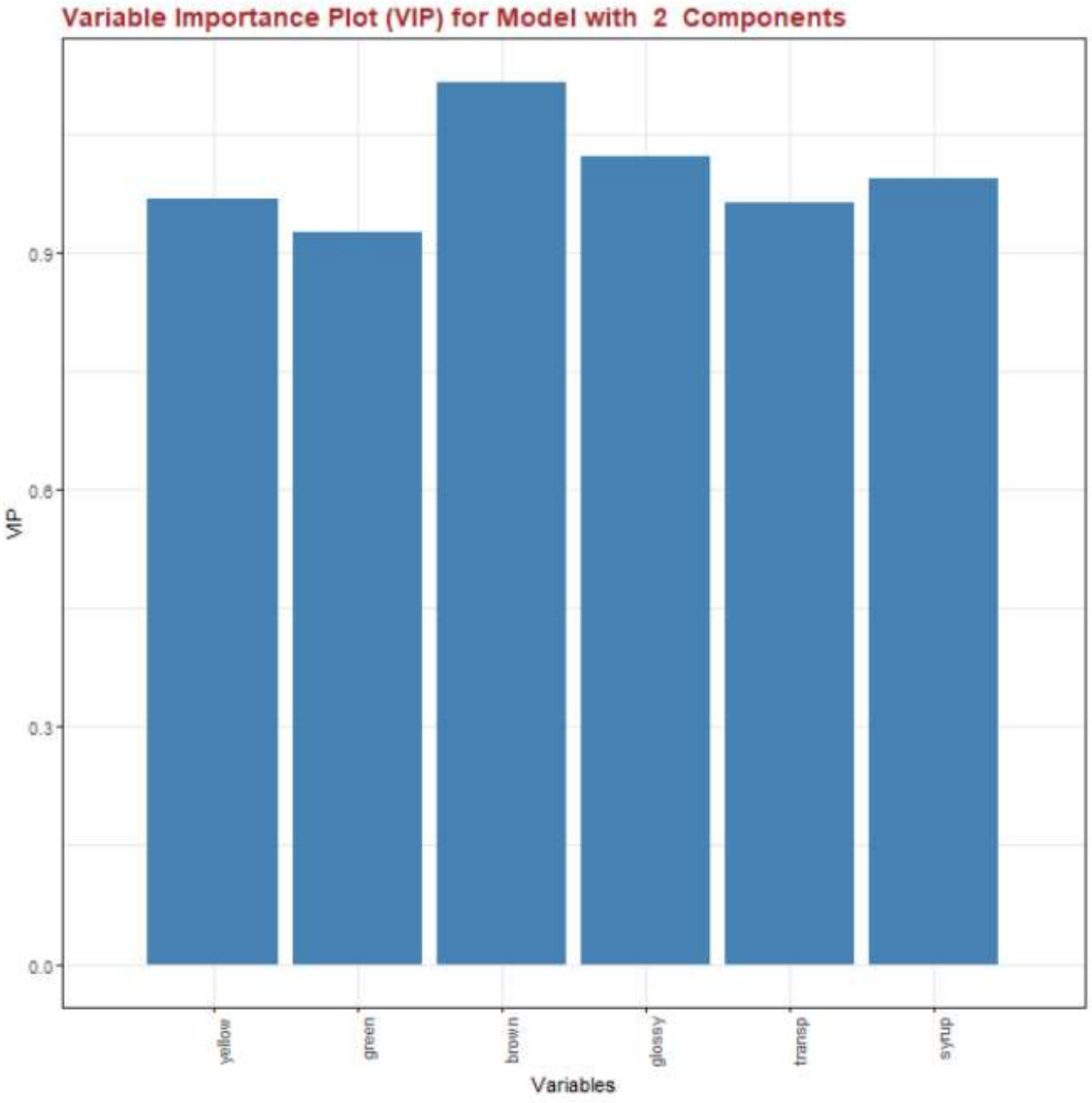
The latent variables for the Predictor are the T scores and summarize the whole X -space. The latent variables for the Responses are the U scores and summarize the Y -space. The PLS model calculate T and U to be maximally correlated. Plots of each column of T vs each column of U show the maximum degree of correlation if they are made on the same component, i.e. T1 vs U1, T2 vs U2, etc. Different components should correlate very poorly instead. Some graphs of the TU plots are shown in the following for the gasoline Data Set.



PLS VIP Plot

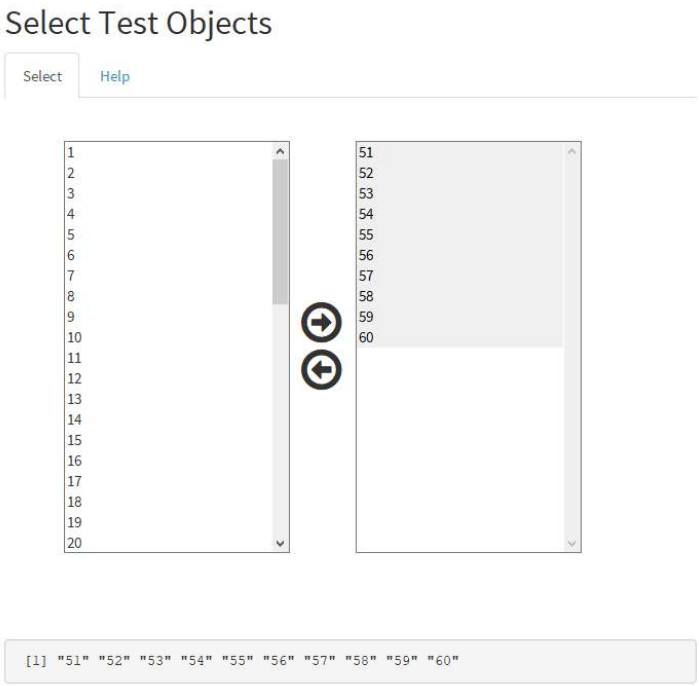
The Variable Importance to Projection (VIP) values show the relative importance of that column in the X Predictor to the overall model. It is the importance to explaining not only the full Y Response space, but also the X space. It looks like a bar plot with a bar for each of the Predictor Variables.

As a rule of thumb: Variables with VIP >1 are important. The following plot shows the example of oliveoil Data Set. The importance of this plot becomes evident as soon as the number of Predictors increases.

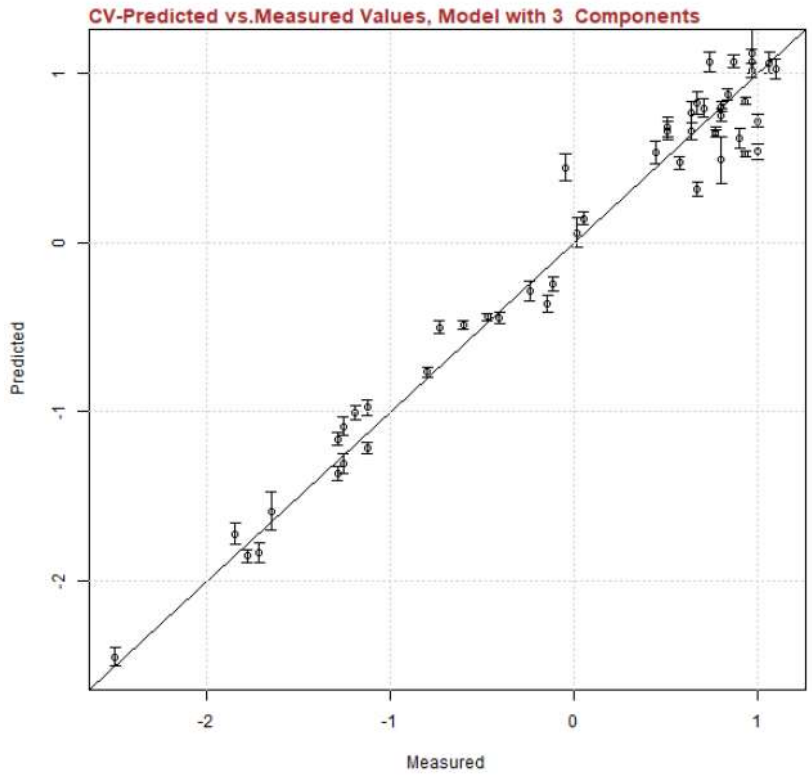


Add Test Data to PLS space

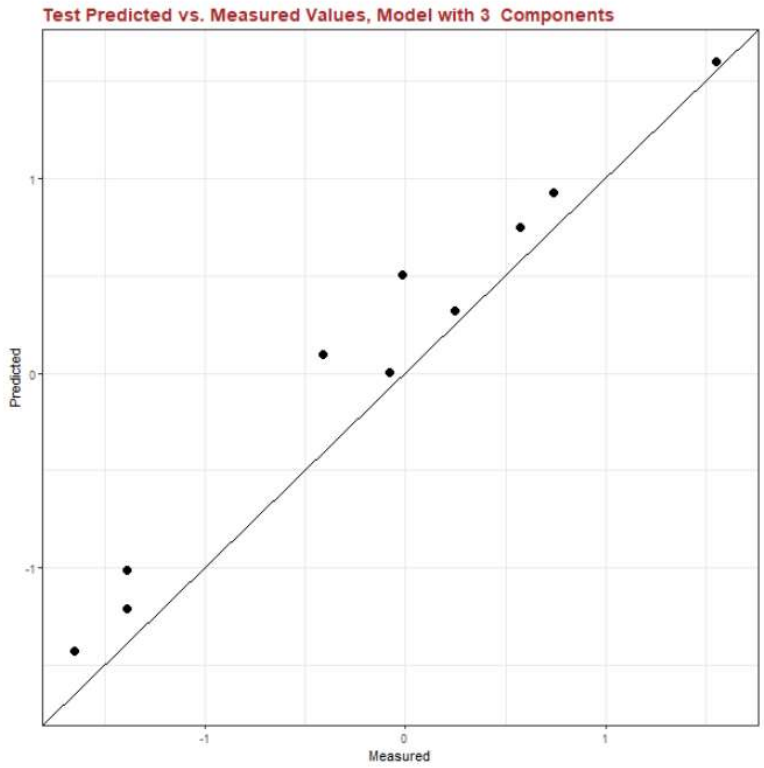
The PLS model is commonly done to predict the behavior of new objects. This is certainly truth when you want to validate your model comparing the prediction with the measured results in a Test Set. This task is easily managed by this menu. Let's take as example the gasoline Data Set and force the last ten objects to be a Test Set for the other that still are considered Training values (see condition below).



The model based on the reduced Training Set is evaluated and does not seem to be reduced in quality by the reduction of the last points. See the prediction on Training Set below.



However, you see that points do not span equally in all the range since they tend to accumulate in the top region. This may cause problem if new values are added in the left region. To test the performance of the model on the Tests Set, it is necessary to invoke this menu and get the prediction plot below.



As usual the plot shows Predicted vs Measured values and the optimal criterion is to have all points on the diagonal. Results are not really bad since all the points are close to the line and they are in a range (-1 , 1) where the model should be well balanced. Therefore, we have to observe that the model always overestimates data. This can be due to the fact that Test points, if returned in the Training Set again, would improve the prediction of the model. Their information results relevant for the model.

Another possibility in the use of this menu is that you want to get the prediction of the model without having any measured value to compare on. In this case, your Data Set will have some "NAs" in the Response columns. See in the following the case of gasoline example Data Set where the octane values of the last 5 objects were removed.

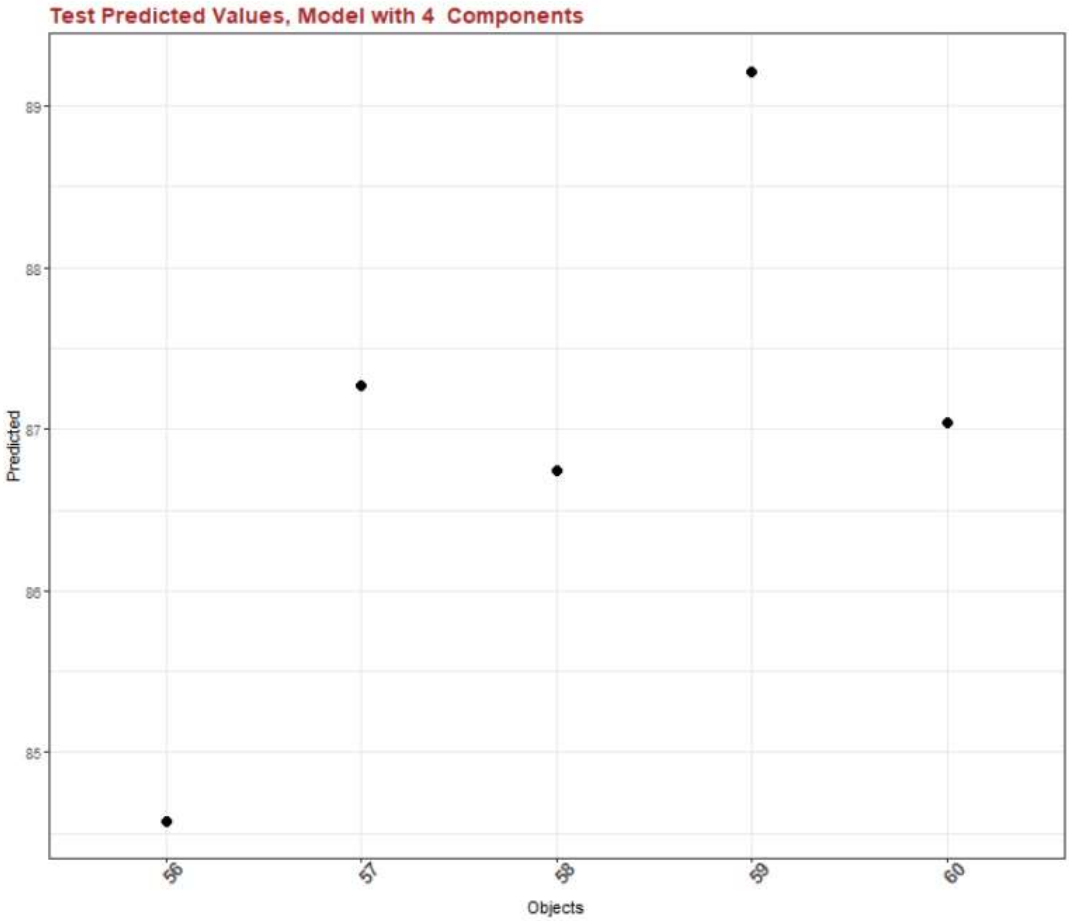
Show 10 entries

Test Dataset with active Rows and Columns

octane	NIR.900.nm	NIR.902.nm	NIR.904.nm	NIR.906.nm
	-0.046884	-0.04236	-0.038683	-0.035291
	-0.055555	-0.049867	-0.045942	-0.042266
	-0.053693	-0.04802	-0.044677	-0.041021
	-0.056311	-0.051231	-0.047483	-0.044605
	-0.058805	-0.053311	-0.049543	-0.045053

Showing 1 to 5 of 5 entries

Again, you have to indicate the last five samples as Test Set, make model and prediction as before. Now the prediction plot changes showing only the predicted values. The predict numbers can be get by the specific button as CSV file.



Useful References

Evolutionary Data Analysis	
1	A.Unwin, M.Theus, H.Hofmann, Graphics of Large Dataset, Springer (2006)
2	D.A:Zighed, S.Tsumoto, Z.W.Ras, H:Hacid, Mining complex data, Springer (2009)
3	V.A.Petrushin, L.Khan,, Multimedia data mining and knowledge discovery, Springer, (2007)
Principal Component Analysis	
1	P.Nomikos, J.F.MacGregor, Monitoring batch processes using multiway PCA. AIChE J., 40,8,1361-1375 (1994)
2	J.V.Kresta, J.F.MacGregor, Multivariate statistical monitoring of process operating performance, Can. J. of Chem. Eng., 69,1, 35-47 (1991)
3	L.H.Chiang, R.Leardi, R.J.Pell, M.B.Seasholtz, Industrial experiences with multivariate statistical analysis of batch process data, Chemometric.and Intelligent Lab. System., 81, 109-119 (2006)
4	G.Luciano, R.Leardi, P.Letardi, PCA of colour measurements of patinas and coating systems for outdoor bronze monuments, J.of Cultural Heritage, 10, 3, 331-337, 2009
5	H.Abdi, L.J.Williams, Principal component analysis,Wires Comp Stat,2,433-459,(2010)
6	R.Bro, A.K.Smilde, Pricipal component analysis, Anal. Methods,6, 2812, (2014)
7	C.B.Y Cordella, Pca: The basic building block of chemometrics, http://dx.doi.org/10.5772/51429
8	I.Jolliffe, Principal Component Analysys, Spring 2nd Ed, 2002
Partial Least Square Regression	
1	A.J.Burnham, J.F.MacGregor, R,Viveros, Latent variable multivariate regression modeling, Chemometric.and Intelligent Lab. System., 48, 167-180 (1999)
2	T.Kourti, J.F.MacGregor, Process analysis, monitoring and diagnosis, using multivariate projection methods, Chemometric.and Intelligent Lab. System., 28, 3-21 (1995)
3	T.Kourti, P.Nomikos, J.F.MacGregor, Analysis, monitoring and fault diagnosis of batch process using multiblock and multiway PLS, J.Proc.Cont.5,4,277-284, (1995)
4	S.Wold, M.Sjostrom, L.Eriksson, PLS regression: a basic tool of chemometrics, Chemometric.and Intelligent Lab. System., 58, 109-130 (2001)
5	S.Yoon, J.F:MacGregor, Incorporation of external information into multivariate PCA/PLS models, On-line Fault detection and supervision in the chemical process industries, Jejudo Island, Korea (2001)
6	S.G.Munoz, J.F. MacGregor, Success stories in the process industries, CEP, 36-40, March 2016
7	R.Leardi, Application of genetic algorithm-PLS for feature selection in spectral data set, J.Chemometrics,14,643-655, (2000)