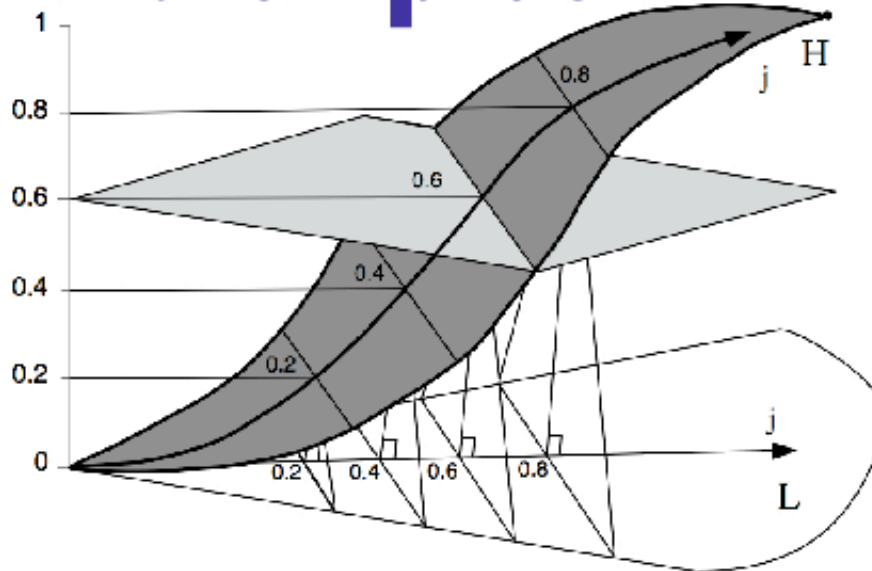


MultBiplot



Multivariate Analysis using Biplots

MultBiplot

Package

José Luis Vicente Villardón
Departamento de Estadística
Universidad de Salamanca
Spain

e-mail: villardon@usal.es

web: <http://biplot.usal.es/classicalbiplot/>

1.- INTRODUCTION

1.1.- Description

Classical Biplot is a program to perform Classical Biplot Analysis. It contains Classical Biplot, HJ-Biplot and Simple Correspondence Analysis of a Contingency Table.

The program is still in Beta version and it is updated frequently. Check the website for updates and new additions.

A more complete version of the program (MULTBILOT – MULTivariate Analysis using BILOT) is available in Alpha version.

The program tries to implement, in a software tool, the experience of the Applied Statistics Group of the Statistics Department at the Salamanca University (Spain) in working with Biplots.

Most of the software available for the construction of Biplots has been developed for very particular applications or as a small part inside general purpose packages. Usually the graphical representations of most packages are not very flexible, producing static pictures that limit the visual interpretation of the results.

CLASSICAL BILOT is conceived not to be “another biplot program”, but conceived to fill the gap between the static pictures and a more dynamic visual interpretation. The package provides the analyst with an interactive approach that uses colour, selection of important features, partial views of the picture and other graphical tools, to interpret biplots. The package also provides some original points of view of the classical techniques based on some original scientific work of the author.

1.2.- System Requirements

The program is designed to work in the Windows environment although Mac versions can be available upon request. The program requires that you have other programs installed to work properly:

- For the compiled version you need the MATLAB Compiler Runtime Library. At the web site a Self-extracting MATLAB Compiler Runtime library utility (platform-dependent file that must correspond to the end user's platform) is provided.
- You can obtain the Matlab source files upon request. To be able to run the source files you need MATLAB with the Statistics Toolbox. Running the source files inside Matlab provides extra capabilities not available in the compiled version.

1.3.- License

The program is citeware. That means you don't have to pay anything for it but if you use it in your work you have to cite it. If you want a custom made adaptation of the program to do something or need my help for the analysis of your data, feel free to contact me. However, please understand that I don't make custom programs or analysis without a full scientific collaboration.

1.4. -Disclaimer

All software available from this website is copyright protected © 2000-2009 by José Luis Vicente Villardón (Departamento de Estadística, Universidad de Salamanca, Spain).

Unless stated otherwise, all software is provided free of charge. If you paid for it, then you got screwed.

As well, all software is provided on an "as is" basis without warranty of any kind, express or implied.

Under no circumstances and under no legal theory, whether in tort, contract, or otherwise, shall the authors be liable to you or to any other person for any indirect, special, incidental, or consequential damages of any character including, without limitation, damages for loss of goodwill, work stoppage, computer failure or malfunction, or for any and all other damages or losses.

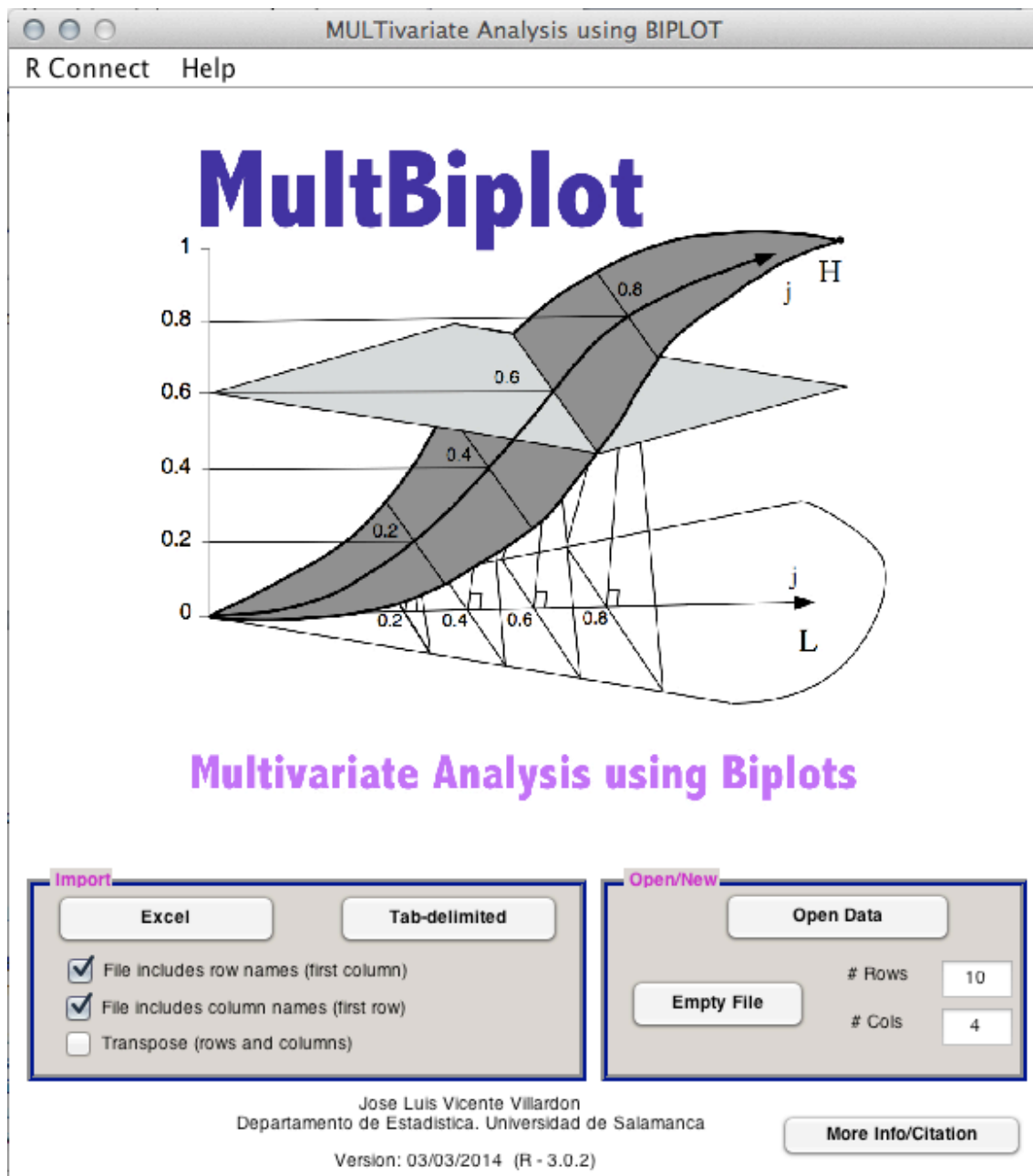
If you do not agree with these terms, then you are advised to not use the software.

Do not link to specific files. If you wish to endorse a product, then please link to the webpage that refers to that product.

2.- RUNNING THE PROGRAM

To run the program, just double click the ClassicalBiplot icon if you are using the compiled version. If you are running the program inside Matlab, type ClassicalBiplot in the command window. For this to work you have to set the folder containing the program and additional files in your Matlab search path (Set Path ... option in the File Menu).

The main window of the program appears in the next figure.



The initial window is simply a bridge between the data files and the main Data Editor and Analysis program.

3.-INITIAL DATA FILES

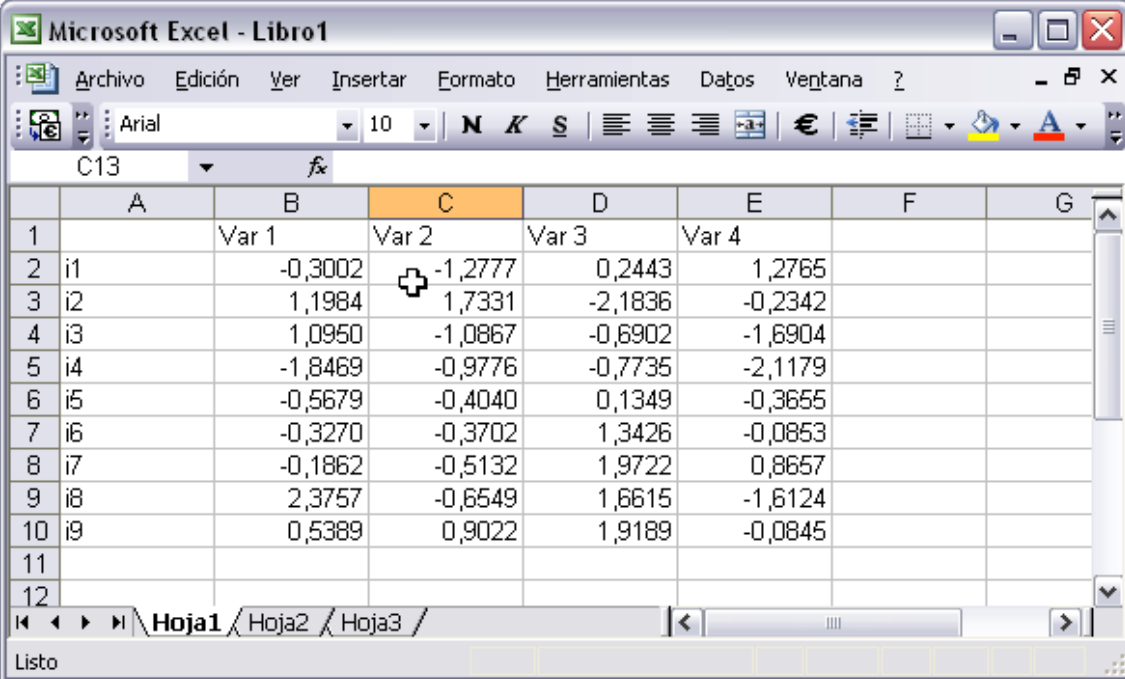
The program can read files from an Excel spreadsheet, delimited text of files previously saved by the program.

The natural way to start a data file is from an Excel spreadsheet. The spreadsheet contains the data and usually the labels for rows and columns. The labels for each row should be in the first column starting in cell A2 and the labels for the columns in the first row starting at cell B1. All the labels should contain alphanumeric strings and can not contain only numbers (in that case the program will not interpret labels correctly).

The numerical data matrix should start at cell B2. Empty cells will be considered as missing data values although you can select any number to code the missing entries (For example, 999 is an usual code for missing values). Empty cells will be read into the program as *NaN* (Not a Number) values in Matlab. Note that some of the modules are not prepared for missing values and will report a Matlab error when *NaNs* are present and incorrect calculations if you use any other numerical code.

The labels can be skipped but you have to remember to tell the program when reading the data selecting the appropriate boxes at the initial window.

The next figure shows the typical Excel spreadsheet used as the input for the CLASSICAL BIPLLOT.



The screenshot shows a Microsoft Excel window titled "Microsoft Excel - Libro1". The spreadsheet has columns A through G and rows 1 through 12. The data is organized as follows:

	A	B	C	D	E	F	G
1		Var 1	Var 2	Var 3	Var 4		
2	i1	-0,3002	-1,2777	0,2443	1,2765		
3	i2	1,1984	1,7331	-2,1836	-0,2342		
4	i3	1,0950	-1,0867	-0,6902	-1,6904		
5	i4	-1,8469	-0,9776	-0,7735	-2,1179		
6	i5	-0,5679	-0,4040	0,1349	-0,3655		
7	i6	-0,3270	-0,3702	1,3426	-0,0853		
8	i7	-0,1862	-0,5132	1,9722	0,8657		
9	i8	2,3757	-0,6549	1,6615	-1,6124		
10	i9	0,5389	0,9022	1,9189	-0,0845		
11							
12							

Make sure that all the cells outside the desired range are empty to avoid undesired errors. The value in cell A1 is ignored by the program.

From the initial program window is also possible to read tab-delimited data with the same structure as the Excel spreadsheet except that a dot has to be used as the decimal separator independently of the local configuration of the computer.

	Var 1	Var 2	Var 3	Var 4
i1	-0.3002	-1.2777	0.2443	1.2765
i2	1.1984	1.7331	-2.1836	-0.2342
i3	1.0950	-1.0867	-0.6902	-1.6904
i4	-1.8469	-0.9776	-0.7735	-2.1179
i5	-0.5679	-0.4040	0.1349	-0.3655
i6	-0.3270	-0.3702	1.3426	-0.0853
i7	-0.1862	-0.5132	1.9722	0.8657
i8	2.3757	-0.6549	1.6615	-1.6124
i9	0.5389	0.9022	1.9189	-0.0845

It is also possible to create an empty data table with the desired numbers of rows and columns or opening a data file previously created by the program. The data file saved by the program is a complex data structure that contains not just the data but also all the information managed by the Data Editor such as colours, markers, labels, measuring scales, etc.

4.- DATA MANAGEMENT and ANALYSIS

4.1.- Data Management Window

When you choose a data file the program opens it in the Data Editor as shown in the next figure. The Data Editor allows the user to change properties of rows and columns to facilitate the visual inspection of the graphical display associated to the biplots and to select the appropriate analysis for the data.

Classical Biplot
Data management

Job Title: C:\Documents and Settings\Usuario\Escritorio\cancer.xls

	Location	Malignancy	EAM103	EAM105	EAM109
STOM	1	1	7.4204	10.8391	6.6463
STOM	1	1	6.9310	11.7231	6.7816
STOM	1	1	7.8532	10.2867	6.4171
STOM	1	1	6.9731	11.8241	7.8499
STOM	1	1	6.3965	11.0672	7.4812
STOM	1	1	6.0197	11.3171	6.5336
COLO	2	1	7.5651	11.5419	6.3749
COLO	2	1	6.5147	10.4559	5.1437
COLO	2	1	6.8552	11.3294	5.6457
COLO	2	1	7.5718	11.5706	6.6991
COLO	2	1	7.5410	11.5891	6.5118
COLO	2	2	7.0296	10.8828	7.3994
COLO	2	2	5.0177	10.3929	6.4165
COLO	2	2	7.2210	11.4748	6.3779
COLO	2	2	6.2457	7.2325	5.8737
COLO	2	2	6.2307	9.9666	6.3536
COLO	2	2	6.5145	8.0910	10.0595
COLO	2	2	5.8233	10.8536	5.7746
COLO	2	2	5.3841	8.6226	5.8841
COLO	2	2	5.8609	9.9004	8.6273
COLO	2	2	5.7960	9.8517	6.4505
PA	3	1	7.1184	10.7908	9.1174
PA	3	2	6.4440	12.0909	9.9793
PA	3	2	6.1251	11.0026	6.7822
PA	3	2	6.8501	11.7178	12.2488

Comments

In the data Editor you will find the Title of the job (that you can change), the name of the file for the data base, the data and the properties of rows and columns. The Analyses supported are in the menus at the top of the window.

4.2.- The Data

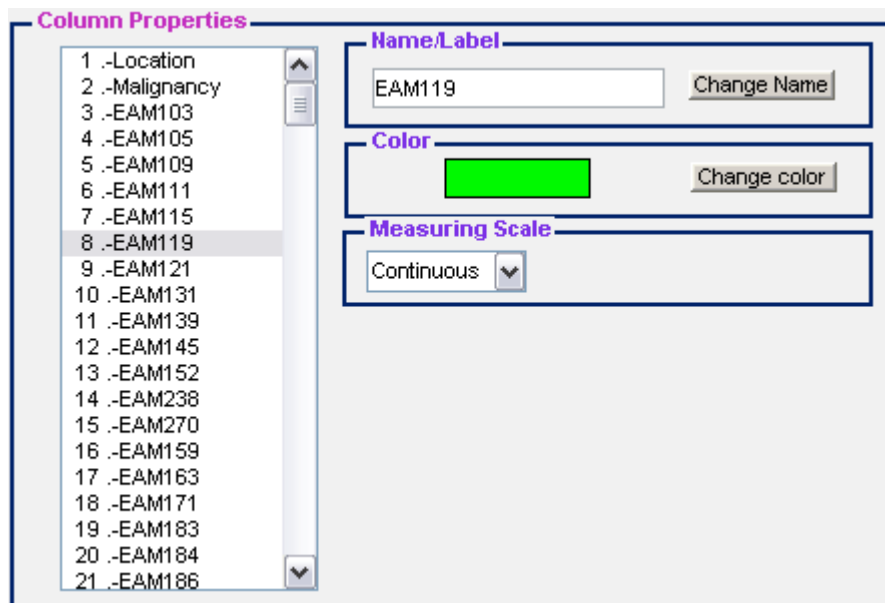
The left side of the window contains the data. The table is editable, that means that you can correct the values when detect an error, or enter new data when you create an empty data table. For the moment, new data can not be added to an existing data table, although this will be implemented in a near future.

	Location	Malignancy	EAM103	EAM105	EAM109	
STOM	1	1	7.4204	10.8391	6.6463	▲
STOM	1	1	6.9310	11.7231	6.7816	
STOM	1	1	7.8532	10.2867	6.4171	☰
STOM	1	1	6.9731	11.8241	7.8499	
STOM	1	1	6.3965	11.0672	7.4812	
STOM	1	1	6.0197	11.3171	6.5336	
COLO	2	1	7.5651	11.5419	6.3749	
COLO	2	1	6.5147	10.4559	5.1437	
COLO	2	1	6.8552	11.3294	5.6457	
COLO	2	1	7.5718	11.5706	6.6991	
COLO	2	1	7.5410	11.5891	6.5118	
COLO	2	2	7.0296	10.8828	7.3994	
COLO	2	2	5.0177	10.3929	6.4165	
COLO	2	2	7.2210	11.4748	6.3779	
COLO	2	2	6.2457	7.2325	5.8737	
COLO	2	2	6.2307	9.9666	6.3536	
COLO	2	2	6.5145	8.0910	10.0595	
COLO	2	2	5.8233	10.8536	5.7746	
COLO	2	2	5.3841	8.6226	5.8841	
COLO	2	2	5.8609	9.9004	8.6273	
COLO	2	2	5.7960	9.8517	6.4505	
PA	3	1	7.1184	10.7908	9.1174	
PA	3	2	6.4440	12.0909	9.9793	
PA	3	2	6.1251	11.0026	6.7822	
PA	3	2	6.8501	11.7178	12.2488	▼

Observe that the first non editable column contains the row labels and the first non editable row the variable names or labels. Those can not be changed directly clicking on them, to do that, you have to change the rows and columns properties on the right side of the table.

4.3.- Column Properties

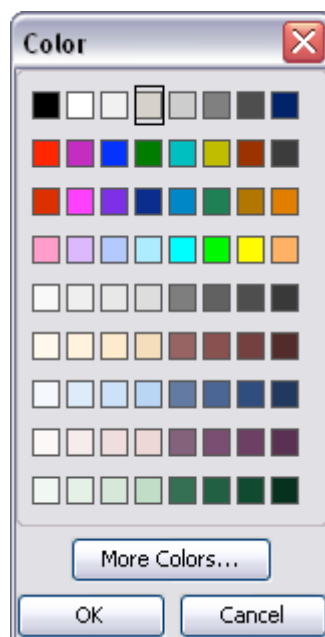
The next figure shows the area for controlling the column properties.



From here you can change the name, colour and measuring scale of each column clicking on the buttons.

To change the name, type the new name in the text box and click on the “Change name” button. The name on the list and the data table will be updated.

To change the colour click the “Change Colour” button and select the new one.



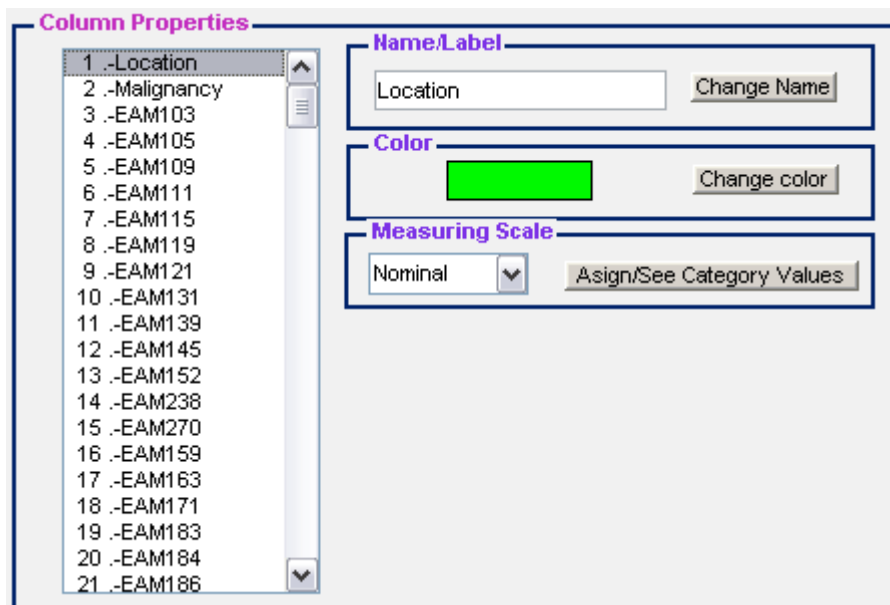
To change the measuring scale select one from the menu.

It is possible to select different measuring scales for each variable. It is important to determine the correct measuring scale for each variable because each technique selects the variables with the right scale for its application. For example, classical biplots will be applied only to the variables with continuous scales and Correspondence Analysis to frequencies or abundances.

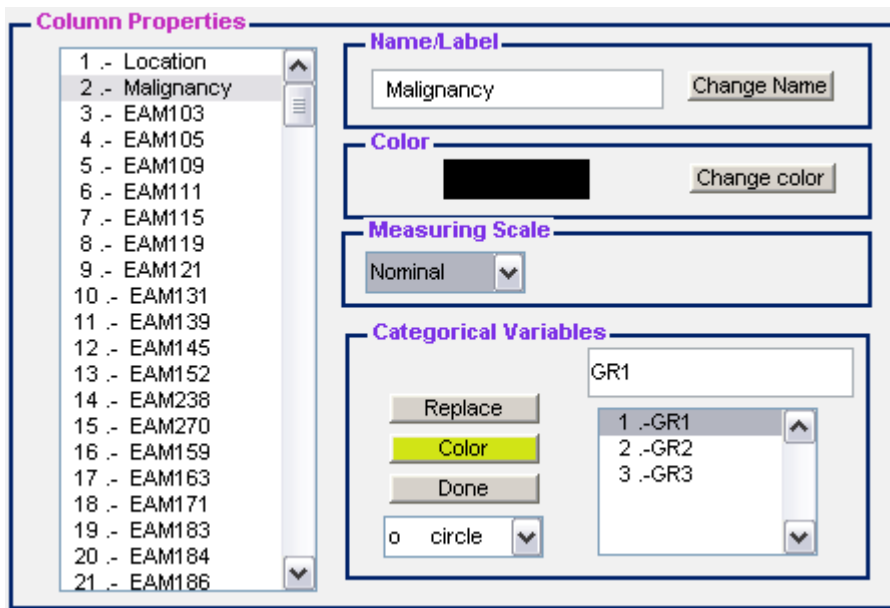
The possible measuring scales are: Continuous, Nominal, Ordinal, Binary, Frequency/Abundance or proximity. There are some simple rules for each type:

- All the initial values have to be numerical.
- Nominal and Ordinal variables should be coded with integer numbers starting with 1 and without any gap between categories.
- Binary variables should be coded using 0 and 1 for absence and presence.
- Frequency/abundance values are non-negative.
- Proximity variables will be used in future versions.

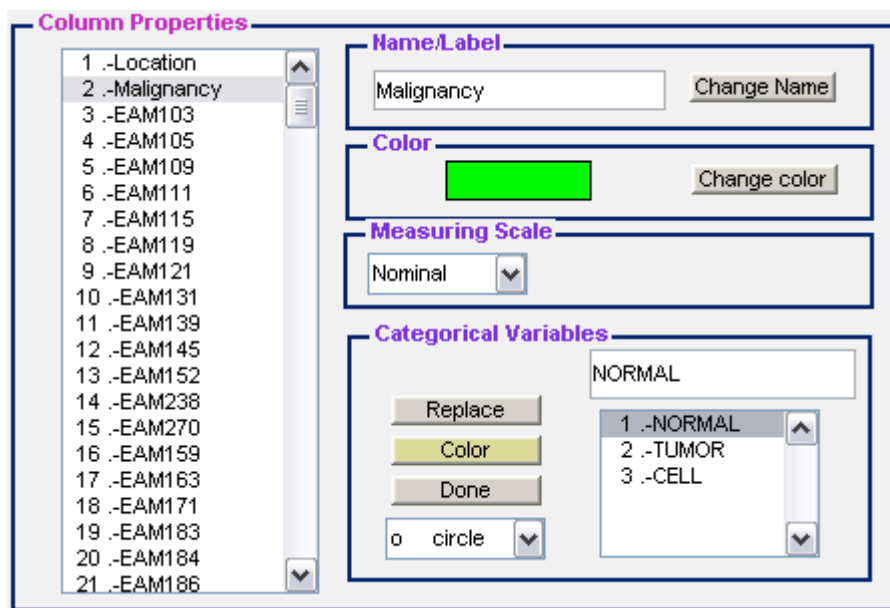
It is possible to assign different labels for each category of a nominal variable. When you select Nominal, a new button to assign the labels appears.



Click on the button “Assign/See Category Values” to assign new labels or to see the set previously defined.



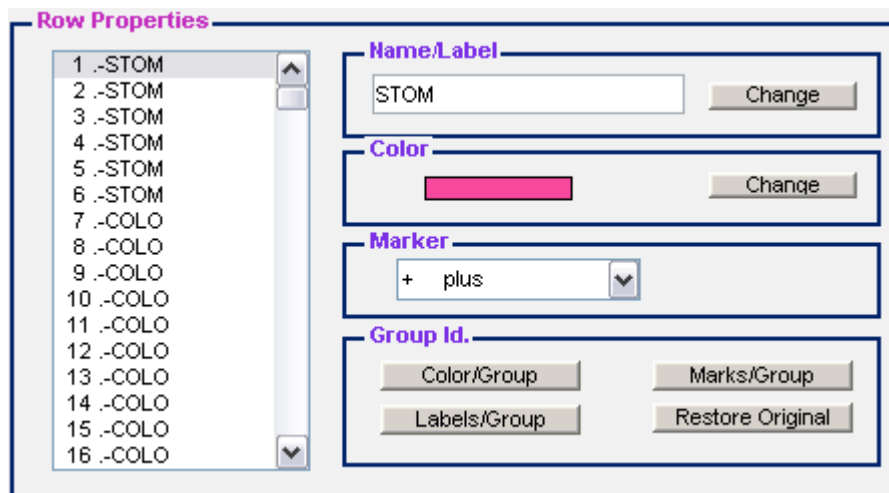
By default the program will assign new labels, colours, and markers for each category. You can change the default values for your own, using the buttons and menus.



When you finish updating the values don't forget to click "Done" for the program to store the modifications.

4.4.- Row properties

The next figure shows the area of the main window used to control the row properties.



For each row you can change the label, colour and marker. The markers and colours are useful in the graphical display to distinguish groups of individuals without using the complete labels that can obscure the representation.

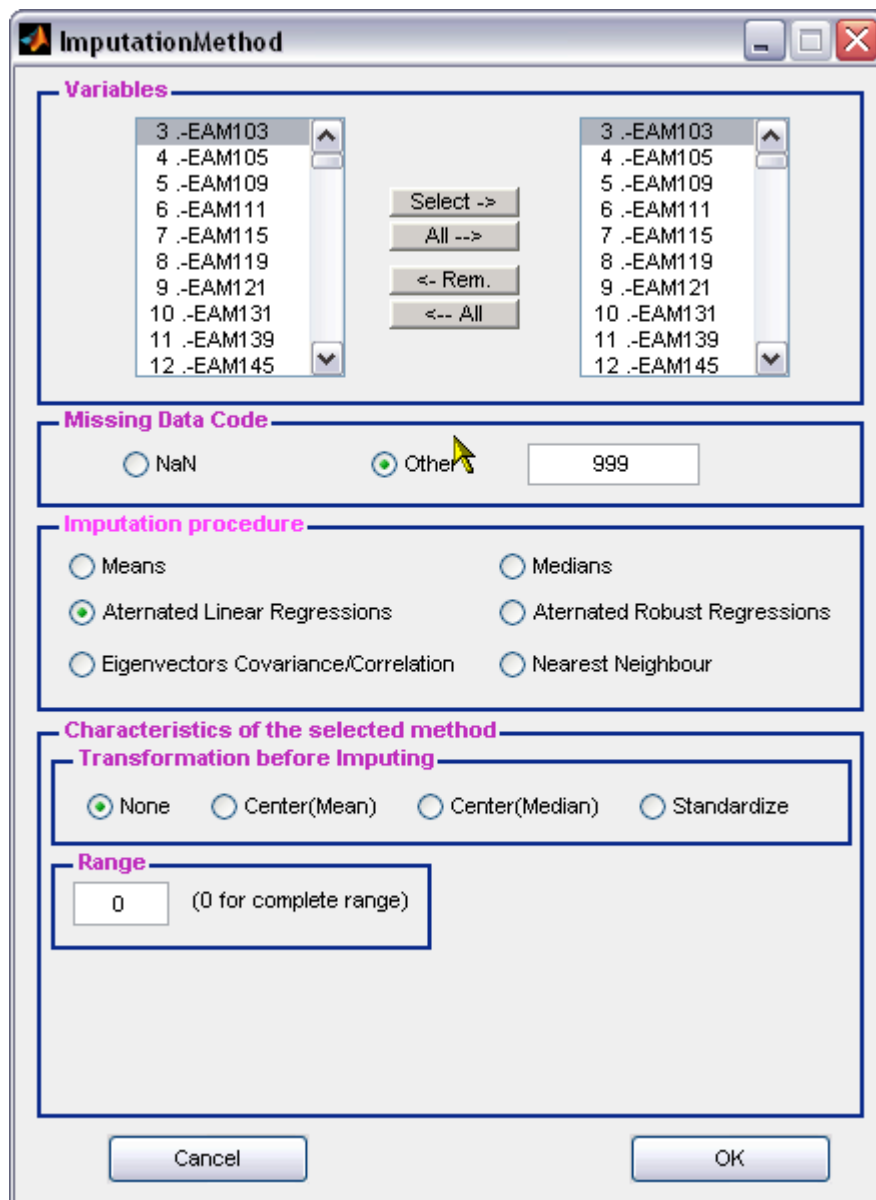
The labels, colours and marks can be assigned according to the values of the nominal variables present in the database, using the buttons in the box “Group Id.”. To do so, select a nominal variable in the variable list and change the category values, colours and marks if need. Clicking on the button “Colour/group” will change the colour of the rows according to the values in the nominal variable selected. Similar actions can be taken for labels and marks using the appropriate buttons.

When you finish the modifications make sure you save the file if you want to use it in a later session.

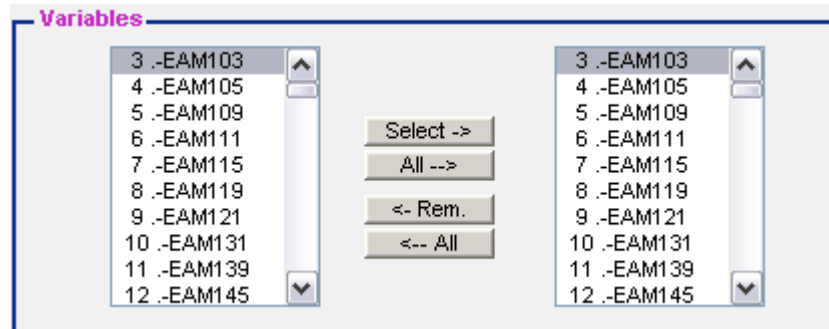
4.5.- Missing Data and Imputation

Most of the analyses have been conceived for complete data matrices and the missing entries can produce a runtime error in the program. The simplest way to solve the problem is eliminating rows and columns to obtain a full matrix. In some practical situations, that implies selecting a very small subset of the original information, even when the number of missing entries is not really high. We have implemented a procedure to complete the data matrix by imputing the missing values, as a previous step to the application of any technique. At the moment, only the continuous variables can be completed although future versions will include the imputations for nominal variables.

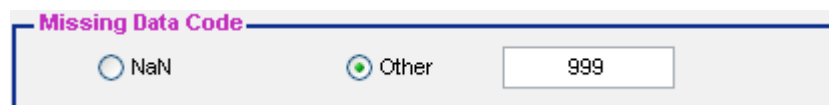
The treatment of missing data can be done using the option “*Missing Continuous Data*” in the “*Impute*” menu. The next figure shows the window that controls the parameters of the imputation.



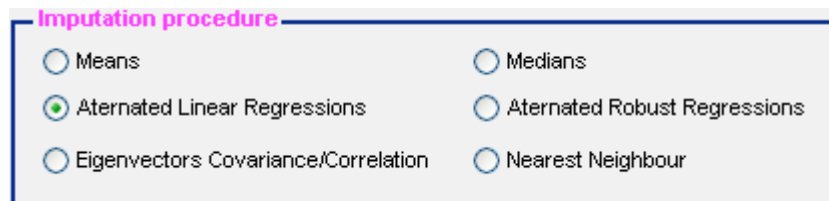
The program selects automatically the continuous variables that can be used for imputation. By default the program selects the whole set to be used in the process. You can add or remove variables using the selection buttons placed between the list of variables and the list of selected variables. The buttons “All →” and “←All” are used to select all variables or remove all respectively. You have to take into account that most imputation techniques are based on the relation among variables to decide what to select.



You can use any numerical code for missing data with the only condition that it has to be the same for the entire database. The Matlab *NaN* (Not a Number) is also accepted as a missing value, actually is the default code when you have a blank cell in the Excel Spreadsheet.



There are several procedures for imputation as shown in the next figure.



- Means and Medians change

The missing values are replaced by the average or the median of its column. The rest of the procedures are more complex and need some additional information that appears in the other boxes inside the window.

Imputation procedure

Means
 Medians
 Alternated Linear Regressions
 Alternated Robust Regressions
 Eigenvectors Covariance/Correlation
 Nearest Neighbour

Characteristics of the selected method

- Alternated Linear Regressions.

The initial data matrix is approximated by a reduced range matrix using alternated rows and columns regressions. The missing values are changed by its corresponding fitted values in the reduced range approximation keeping the non-missing values unchanged. Usually, the higher is the range the better is the approximation; for a complete range approximation the fitted values for the non-missing entries are equal to the actual values and then this is the option that probably should be used unless the observed matrix is too big. A 0 value in the range Box means that the approximation will be done for complete range.

Imputation procedure

Means
 Medians
 Alternated Linear Regressions
 Alternated Robust Regressions
 Eigenvectors Covariance/Correlation
 Nearest Neighbour

Characteristics of the selected method

Transformation before Imputing

None
 Center(Mean)
 Center(Median)
 Standardize

Range

(0 for complete range)

For the alternate procedures we have to decide if we want to center the data before applying the procedure and the range of the approximation. Previous centering or standardization is important if we want to preserve the correlation structure.

- Alternated Robust Regressions

The procedure is the same as ALRs except that the linear regressions are changed by Linear Robust Regressions. For this procedure you have to choose the procedure for the robust regression.

The image shows a software interface with three main sections:

- Imputation procedure:** Contains radio buttons for Means, Medians, Alternated Linear Regressions, Alternated Robust Regressions (selected), Eigenvectors Covariance/Correlation, and Nearest Neighbour.
- Characteristics of the selected method:**
 - Transformation before Imputing:** Contains radio buttons for None (selected), Center(Mean), Center(Median), and Standardize.
 - Range:** Contains a text input field with '0' and the text '(0 for complete range)'.
 - Robust Method:** Contains a dropdown menu with 'Andrews (w = (abs(r)<pi) ...)' selected.

The procedures for robust regressions are those implemented in the Statistics Toolbox in Matlab. The algorithm uses iteratively reweighted least squares with a bisquare weighting function.

Weight Function	Equation	Default Tuning Constant
'andrews'	$w = (\text{abs}(r) < \pi) .* \sin(r) ./ r$	1.339
'bisquare' (default)	$w = (\text{abs}(r) < 1) .* (1 - r.^2).^2$	4.685
'cauchy'	$w = 1 ./ (1 + r.^2)$	2.385
'fair'	$w = 1 ./ (1 + \text{abs}(r))$	1.400
'huber'	$w = 1 ./ \max(1, \text{abs}(r))$	1.345
'logistic'	$w = \tanh(r) ./ r$	1.205
'ols'	Ordinary least squares	None
'talwar'	$w = 1 * (\text{abs}(r) < 1)$	2.795
'welsch'	$w = \exp(-r.^2)$	2.985

- Eigenvectors of the covariance/correlation matrix.

The procedure calculates eigenvalues and eigenvectors of the pairwise Covariance/Correlation matrix and then projects the rows of the data matrix by regression.

Imputation procedure

Means
 Medians
 Alternated Linear Regressions
 Alternated Robust Regressions
 Eigenvectors Covariance/Correlation
 Nearest Neighbour

Characteristics of the selected method

Transformation before Imputing

None
 Center(Mean)
 Center(Median)
 Standardize

Range

(0 for complete range)

- Nearest Neighbour.

The missing values are replaced by a weighted average of the k nearest neighbours of the point with the missing values.

Imputation procedure

Means
 Medians
 Alternated Linear Regressions
 Alternated Robust Regressions
 Eigenvectors Covariance/Correlation
 Nearest Neighbour

Characteristics of the selected method

Nearest Neighbor

Nearest Rows
 Nearest Columns
 Distance
 Number of neighbours

The procedure can be applied using the nearest rows or the nearest columns. Different measures of distance can be chosen.

- 'euclidean' - Euclidean distance (default).
- 'seuclidean' - Standardized Euclidean distance — each coordinate in the sum of squares is inversely weighted by the sample variance of that coordinate.
- 'cityblock' - City block distance.
- 'mahalanobis' - Mahalanobis distance.
- 'minkowski' - Minkowski distance with exponent 2.
- 'cosine' - One minus the cosine of the included angle.

'correlation' - One minus the sample correlation between observations, treated as sequences of values.

'hamming'- Hamming distance — the percentage of coordinates that differ.

'jaccard'- One minus the Jaccard coefficient — the percentage of nonzero coordinates that differ.

'chebychev' - Chebychev distance (maximum coordinate difference).

5.- CLASSICAL BILOT (PRINCIPAL COMPONENTS ANALYSIS)