User’s Manual for the Multivariate Exploratory Data Analysis Graphic Interface

Author: Elena Jiménez Mañas

This document is a tutorial of the MEDA_GUI-Toolbox, Version 2.1. Please note that the code is provided “as is” and we do not accept any liability for its use.

Introduction to Exploratory Data Analysis

Exploratory Data Analysis (EDA) is widely used in data analysis in order to extract useful information from analyzed complex data set. In data sets with a large number of variables, projection models based on latent structures, such as Principal Components Analysis (PCA) [1] and Partial Least Squares (PLS) [1], are valuable tools within EDA. The approach of PCA and PLS is to identify a reduced number of new features, referred to as latent variables (LVs) or specifically in PCA as principal components (PCs). These LVs are obtained as a combination of the original features in the data. With the reduced set of features, a number of visualization plots can be obtained. These plots, conveniently used, facilitate the understanding of complex data sets.

Tools in the Graphic User Interface

Consider a data set with a number of observations (rows) measured over a set of features or variables (columns). The MEDA Graphic User Interface includes the following PCA and PLS tools as to analyze data sets:

1. Score plots [1]: this tool allows the user to visualize the distribution of the observations in the reduced set of LVs. This results in a bi-dimensional scatter plot for a pair of LVs selected by the user.

2. Loading plots [1]: this tool allows the user to visualize the distribution of the features or variables, in order to explore the relationship among variables in the data set. Again, this operation will give the user a bi-dimensional scatter plot for a pair of LVs selected by the user.

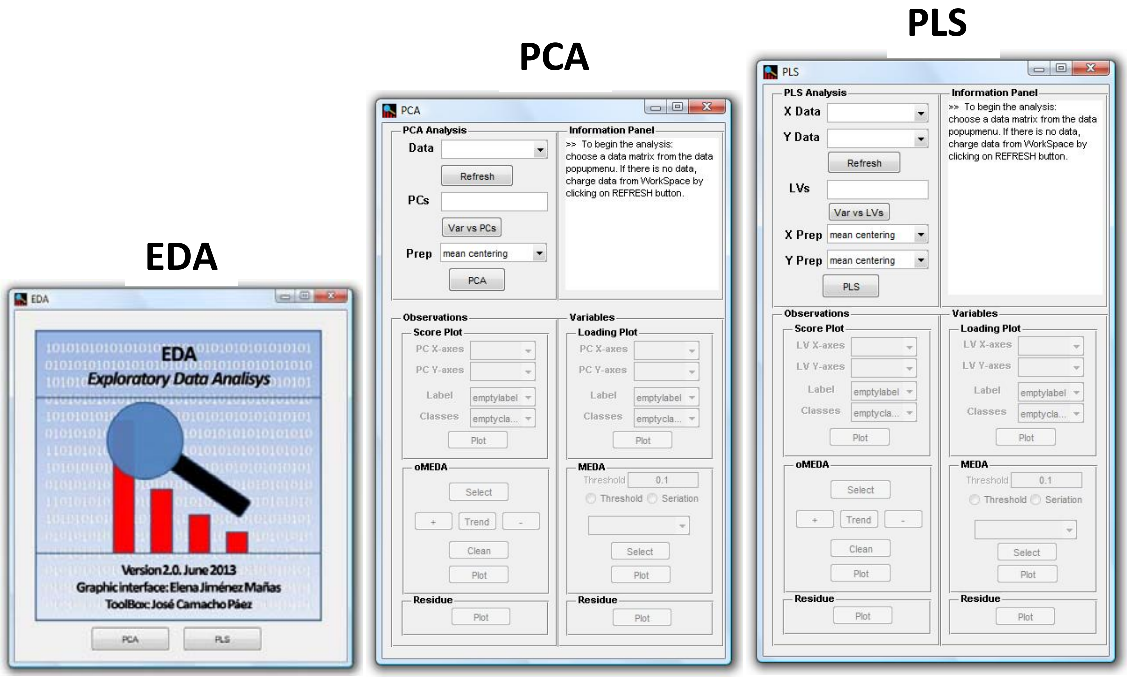
3. MEDA (Missing Data Methods for EDA) [4]: this tool shows the relationship among the features in the data within a simple red-blue grid graphic. The features to display are selected by the user.

4. oMEDA (observation-based MEDA) [5]: is a variant of MEDA to connect observations and features. This is a very powerful tool in order to learn how and what features affect each observation or group of observations. The group of observations to display is selected by the user.

5. Residue: this tool represents the residual variance, in the observations or the variables, of a model with the selected LVs.

Presentation of the Graphic User Interface

MEDA_GUI 2.1 is free software based on Matlab® (v. 2009b). The compatibility with other versions has not been checked and we do not take responsibility for this. The graphic user interface is composed of 3 interfaces shown in the figure below.



>> To begin the analysis: choose a data matrix from the data popupmenu. If there is no data, charge data from WorkSpace by clicking on REFRESH button.

>> To begin the analysis: choose a data matrix from the data popupmenu. If there is no data, charge data from WorkSpace by clicking on REFRESH button.

Fig. 1 Graphical User Interfaces for the EDA Toolbox 2.0

This tutorial is focused on the use of the three user's interfaces. *Important note:* It is necessary to run the MEDA Graphic User Interface with the MEDA Toolbox available onhttp://wdb.ugr.es/~josecamacho/downloads.php

Illustrative example: Comparative of routing algorithms in MANET

This is an example to show the user how to work with the graphic interface. The selected data set, MANET, is available at the CRAWDAD repository (http://crawdad.cs.dartmouth.edu/) and published in [2].It consist of an outdoor experiment for the comparison of three different routing algorithms in a mobile ad-hoc network (MANET) formed by 33 laptops in movement. The evaluated algorithms are: Any Path Routing without Loops (APRL), Ad hoc On-demand Distance Vector (AODV), On-Demand Multicast Routing Protocol (ODMRP) and System and Traffic-dependent Adaptive Routing Algorithm (STARA).

The experiment was carried out in an athletics court of 225x365 meters, in which each laptop user was moving all over the field in a random manner during one hour and a half, approximately. Each laptop generated low rate traffic during this period of time. Each routing algorithm was used during 15 minutes in disjoint time intervals. The laptops positions were captured via GPS, and registered along with the number of the generated packets (TIN packets), the received ones (TOUT), or the retransmitted ones (SIN and SOUT).

A set of statistics listed in table 1 are computed from the original data at regular intervals of time, yielding a total of 100 intervals. The design of this statistics is part of the EDA and should be carried out taking in account the investigation goals. Thus, the first 10 variables are related to the distribution and location of the stations (laptops) in the field, while the remaining 8 variables are related to the network traffic. Among the 100 observations (time intervals), only those in which the four routing algorithms were active are selected (70 observations), and the rest are discarded. Thus, the final data set is formed by 70 observations and 18 variables.

Number	Variable	Description
1	PD	Average distance between laptops
2	mM	Minimum value for max. distances
3	Mm	Minimum value for min. distances
4	cX	X centroid X
5	cY	Y centroid Y
6	cZ	Z centroid Z
7	n1	Amount of laptops with a distance to the centroid lower than 1/32 of the max. distance
8	n2	Amount of laptops with a distance to the centroid between 1/32 and 2/32 of the max. distance
9	n3	Amount of laptops with a distance to the centroid between 2/32 and 3/32 of the max. distance
10	n4	Amount of laptops with a distance to the centroid higher than 3/32 of the max. distance
11	nTI	Number of TIN
12	nTO	Number ofTOUT
13	nSI	Number ofSIN
14	nSO	Number of SOUT
15	vTI	Volumeof TIN
16	vTO	VolumeofTOUT
17	vSI	VolumeofSIN
18	vSO	VolumeofSOUT

Table 1 Features derived from the MANET data set for the example of this tutorial

For the analysis of this data, PLS-DA (PLS discriminatory-analysis) is applied. Since the observations are already classified by the four routing protocols, this means that each one of the 70 observations corresponds to APRL, AODV, ODMRP or STARA. For more information on PLS-DA, see [3].

To begin working with the interface, it is necessary toinclude in the path of Matlab® the directory where the decompressed MEDA_GUI folder is stored. To launch the interface, type in the command line:

```
>>  EDA
```

The window in the figure below pops up. This is the initial interface in the MEDA Graphic User Interface (GUI). From here it is possible to select a PCA or a PLS analysis by clicking over the correspondent button.

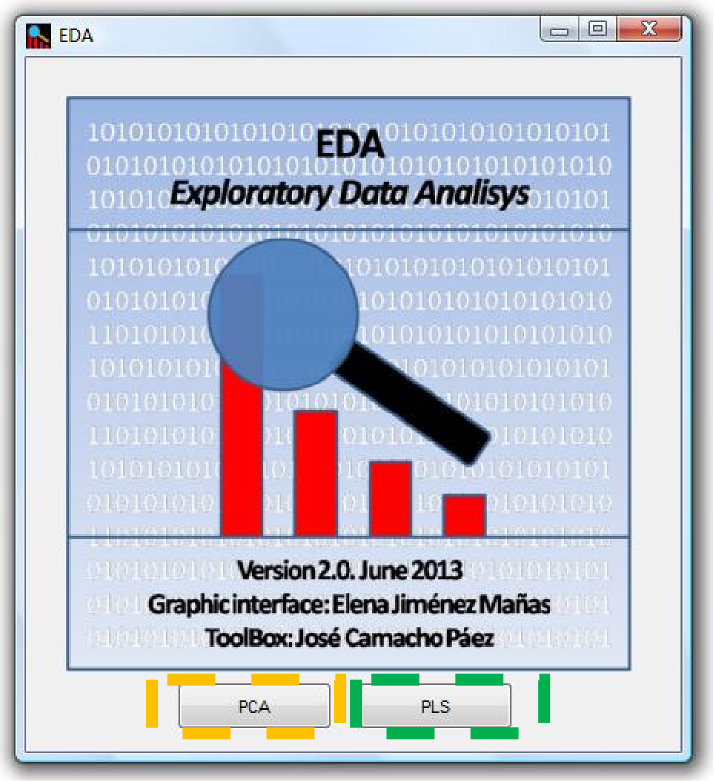
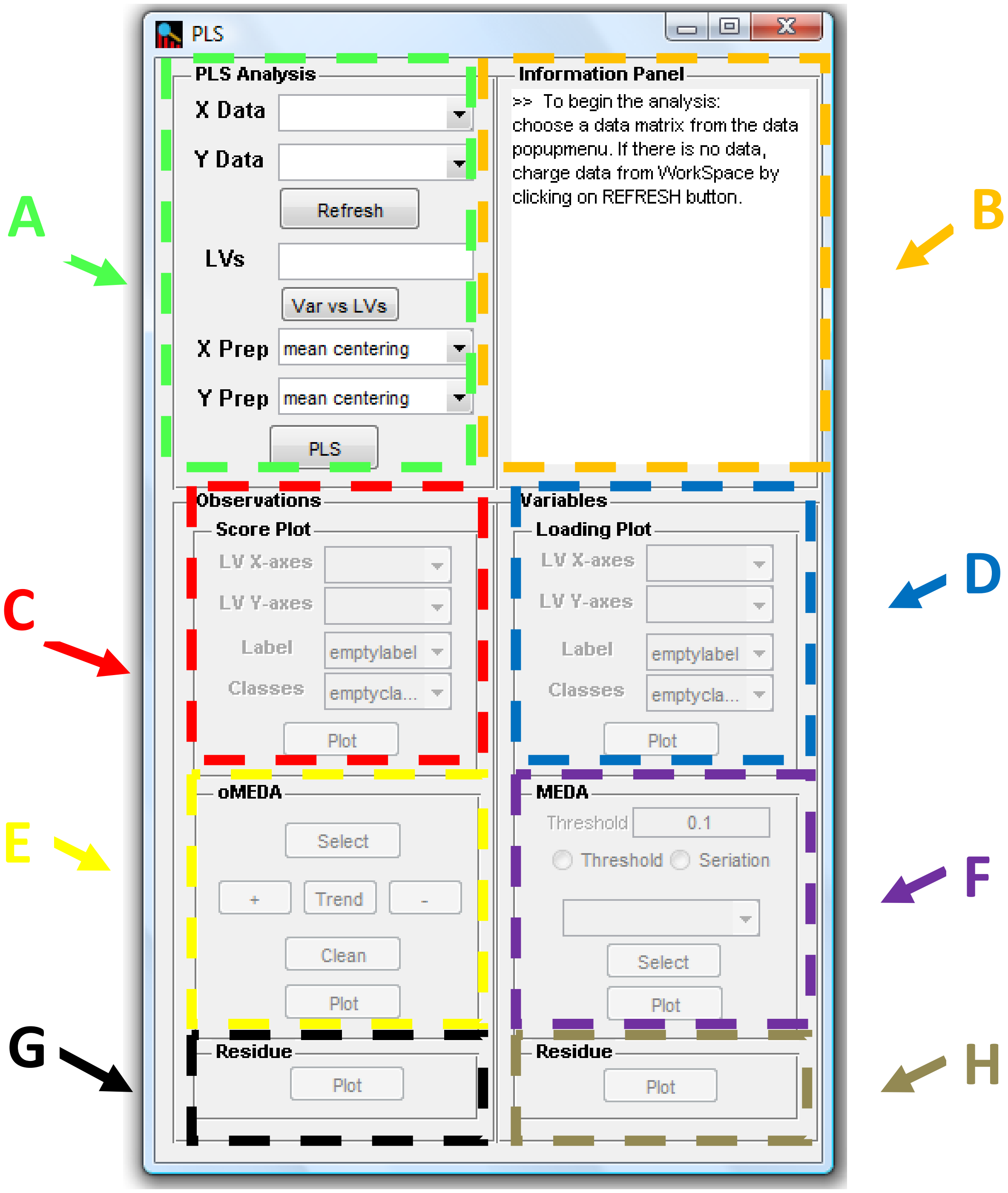

EDA
EDA
Exploratory Data Analisys
Version 2.0. June 2013
Graphic interface: Elena Jiménez Mañas
ToolBox: José Camacho Páez
PCA
PLS

Fig.2 GUI starting window

Because of the nature of the data-set, as explained before, click over the PLS button. *Note:* Although we are only illustrating the use of the PLS graphic interface, the PCA interface works similarly.

After selecting the PLS button, the window in the figure below pops up. This is the PLS interface. The PLS interface is divided in 8 areas: A, B, C, D, E, F, G and H. Areas A and B are related to the PLS model building and Information Panel. Only these two areas will be enabled when initializing the interface. The first step consists on providing the interface with the data to generate the PLS model. These data are: the data set to analyze, *X Data*, the dummy variable for PLS-DA, *Y Data*, the number of LVs to run the model, *LVs*, and the preprocessing method to apply over the matrixes X and Y. By default the preprocessing method is *mean centering*. You can also select no preprocessing or auto-scaling. *Important Note:* Follow the information provided in the *Information Panel* to complete all the data.



X Prep: mean centering
Y Prep: mean centering
Label: emptylabel
Classes: emptycla...
Threshold: 0.1
○ Threshold ○ Seriation

>> To begin the analysis: choose a data matrix from the data popupmenu. If there is no data, charge data from WorkSpace by clicking on REFRESH button.

Fig. 3 PLS GUI

The data of the example in this tutorial is available at the folder named *Examples*, file *MANET.mat*. Load the data matrices in the *Workspace* in Matlab® using command *load*. The following matrices have been loaded, see table 2. Push the *Refresh* button in the GUI in order to load the new data matrices. Now select the data from the *X Data* and *Y Data* popup menus. In this case, it is recommended to use auto-scaling as the preprocessing method due to the different nature of the X-block variables in this data set. An important question arises here; how many LVs are enough to run the model? As an aid in this matter, you can initially select a large number of LVs and use the *Var vs LVs* button on the PLS Analysis area. In this example we are working with 18 variables (listed in Table 1). Let's consider 10 LVs. Write 1:10 on the LVs area and click on the *Var vs LVs* button. The result is shown in the figure below. As it is shown, with just 3 LVs, 70% of the variance is captured, whereas with 5 LVs, 80% of the variance is captured. We will work with 3 LVs. For that, write 3 on the LVs area and press the PLS button. After that new areas on the interface are enabled (Areas C and D) as shown in figure 5.

Name	Size	Observations
x	70x18	Data matrix. Charge on X Data (area A of the GUI).
y	70x4	Data set of predicted variables. Charge on Y Data (area A of the GUI).
clases	70x1	Array containing as many entries as the number of observations in the data ser. This is an optional field that colors the observations according to the value assigned to each of them, classifying the observations in the data set.
label_v	1x18	Array containing as many entries as the number of variables in the data set. This is an optional field that assigns name to each of the variables.
laby	1x70	Array containing as many entries as the number of observations in the data set. This is an optional field that assigns name to each of the observations.

Table 2 Data loaded from the MANET.mat file

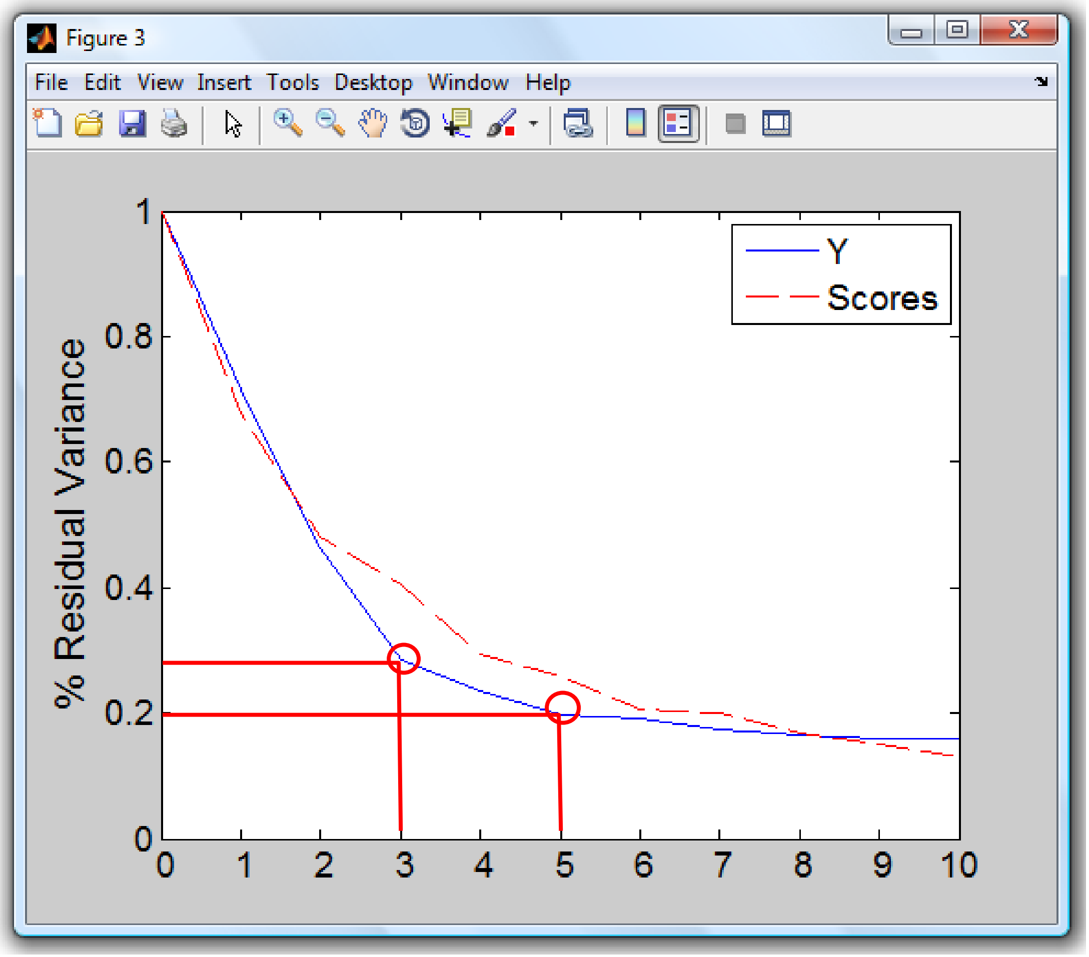
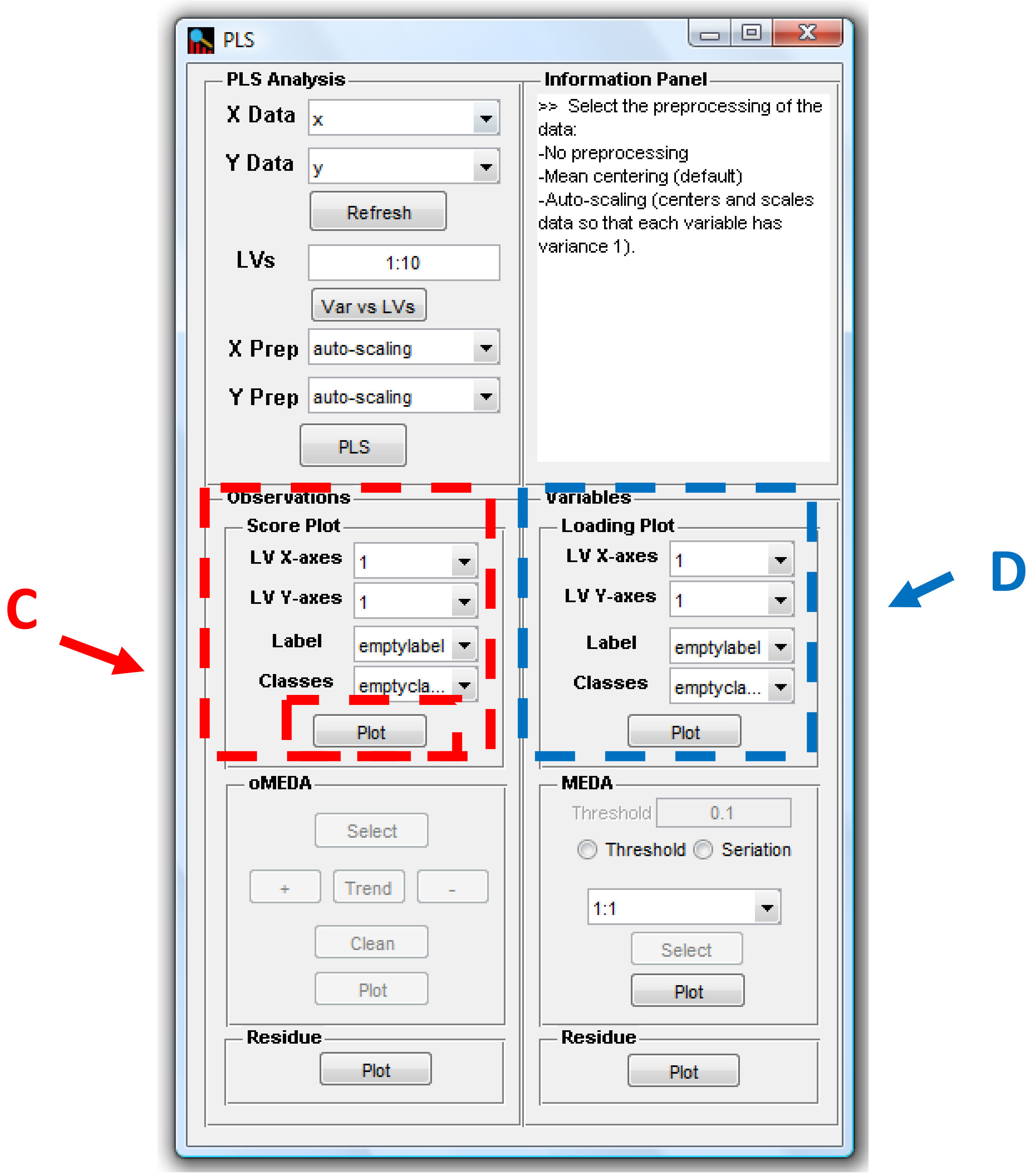


Fig. 4 Var vs LVs plot of the MANET data set



PLS

PLS Analysis

X Data: x

Y Data: y

LVs: 1:10

X Prep: auto-scaling

Y Prep: auto-scaling

Information Panel

>> Select the preprocessing of the data:
-No preprocessing
-Mean centering (default)
-Auto-scaling (centers and scales data so that each variable has variance 1).

Observations

Score Plot: LV X-axes: 1; LV Y-axes: 1; Label: emptylabel; Classes: emptycla...

Variables

Loading Plot: LV X-axes: 1; LV Y-axes: 1; Label: emptylabel; Classes: emptycla...

MEDA: Threshold: 0.1; ○ Threshold ○ Seriation; 1:1

Fig. 5 PLS GUI after model building

Both the *Score Plot* area in the *Observation* menu and the *Loading Plot* area in the *Variables* menu are enabled. It is possible to begin the analysis working with the observations or the variables. In this case we will start with the observations. Select the LVs you want to display on the Score Plot by clicking on the *LV X-axes* and *LV Y-axes* popup menus. Push on the *Plot* button, in area C, to obtain the graphics. In figure 6 you can see 2score plots of the data. We can improve the visualization with colors. For this, we define a *Classes* vector in the GUI. This vector is a row vector, containing 70 values with the assignment of each observation to one class; the assignment is a number from 1 to 4 depending on the routing protocol each observation belongs to. This vector is loaded with the rest of the data from the MANET.mat file. Load this vector on the *Classes* popup menu and repeat the steps to obtain the score plots. The result is shown in figure 7. The colored plot is easier to interpret. In the second plot, the four classes are distributed in different locations. We are working with that score plot from now on. The plot suggests that we can identify the class of an observation from the PLS-DA model with the design variables considered in Table 1. This means that there are significant differences between the algorithms in the data under analysis. We will learn how to unveil the reason for these differences below.

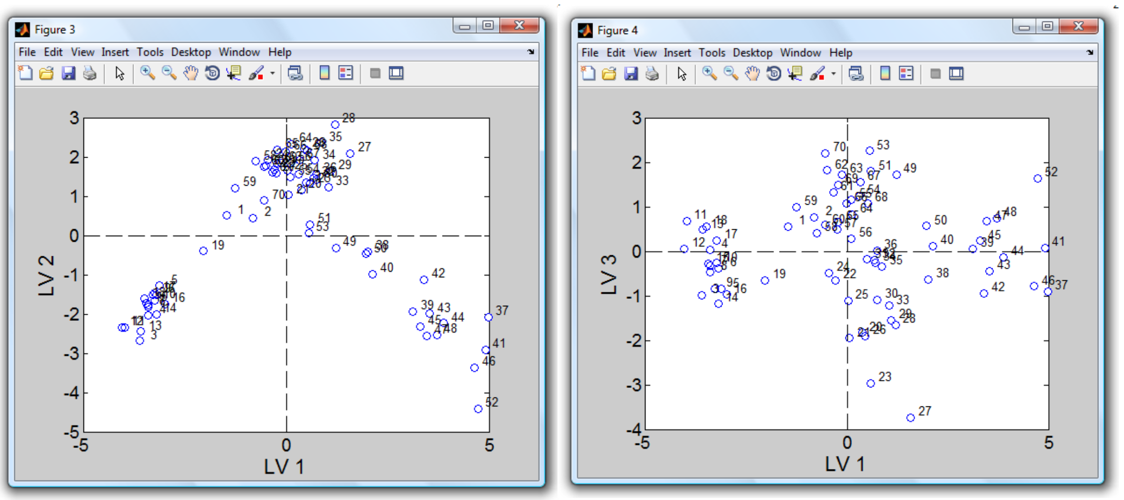


Figure 3
File Edit View Insert Tools Desktop Window Help
LV 1
LV 2
-5
0
5
3
2
1
0
-1
-2
-3
-4
-5
Figure 4
File Edit View Insert Tools Desktop Window Help
LV 1
LV 3
-5
0
5
3
2
1
0
-1
-2
-3
-4

(a) (b)

Fig. 6 Examples of score plots: (a) LV1 vs LV2 and (b) LV1 vs LV3.

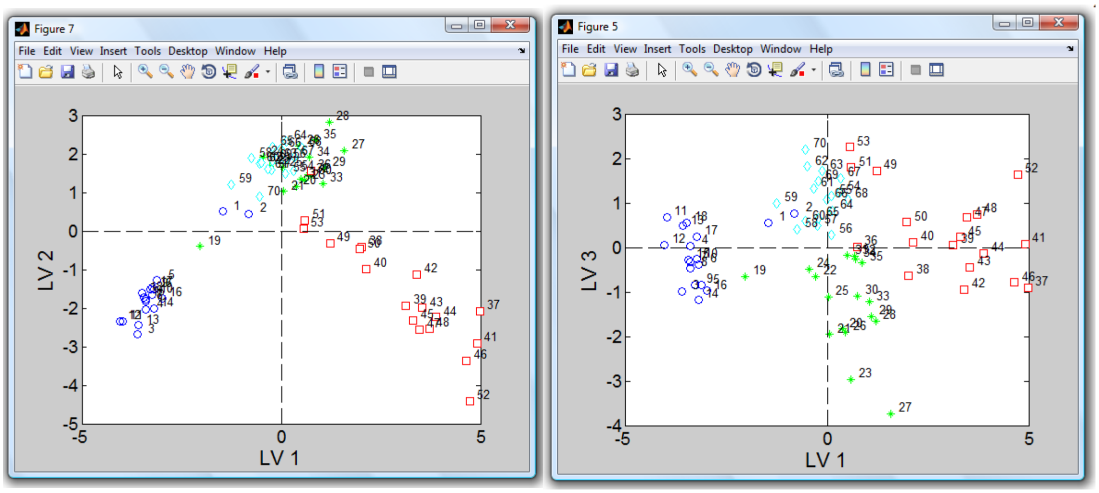
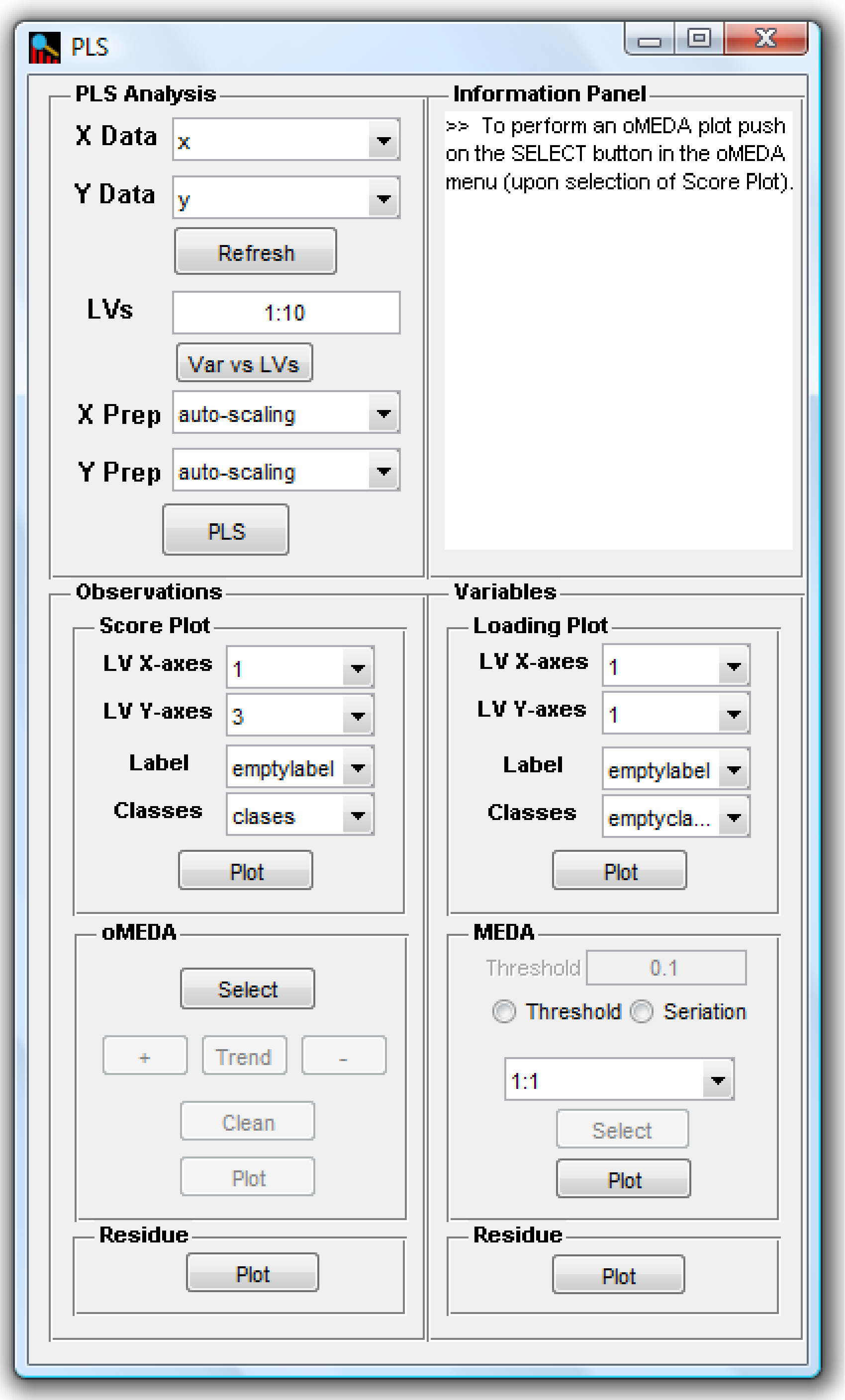


Figure 7
File Edit View Insert Tools Desktop Window Help
LV 1
LV 2
-5
0
5
3
2
1
0
-1
-2
-3
-4
-5
Figure 5
File Edit View Insert Tools Desktop Window Help
LV 1
LV 3
-5
0
5
3
2
1
0
-1
-2
-3
-4

(a) (b)

Fig. 7 Examples of colored score plots: (a) LV1 vs LV2 and (b) LV1 vs LV3.

Maybe you have noticed that after pressing the *Plot* button, the whole interface is enabled, as illustrated in figure 8. At this point, it is possible to work with MEDA, oMEDA and the residues graphics too.



X Data: x
Y Data: y
LVs: 1:10
X Prep: auto-scaling
Y Prep: auto-scaling

>> To perform an oMEDA plot push on the SELECT button in the oMEDA menu (upon selection of Score Plot).

Score Plot: LV X-axes: 1; LV Y-axes: 3; Label: emptylabel; Classes: clases

Loading Plot: LV X-axes: 1; LV Y-axes: 1; Label: emptylabel; Classes: emptycla...

MEDA: Threshold: 0.1; 1:1

Fig. 8 PLS GUI completely enabled

In EDA it is customary to interpret both a score and loading plot of the same subspace together, as done in figure 9. Sometimes this is done in a single plot, the so called bi-plot. In the plots corresponding to a given subspace, it is widely accepted that variables and observations displayed in the same zone or the opposite across the axis of coordinates are related. But this is not a fact. It is important to know that loading and score plots tend to be quite confusing when a high number of points are displayed. For example, take a look at variable cY in figure 9b. Considering its locations, it may be thought that this variable may be related to deviations among the observations in the direction of the arrow in the score plot, figure 9a. This is direction the main discriminant direction between two groups of observations: AODV and STARA. We will see in figure 11 that there is no influence from variable cY over the difference between the groups. To avoid the misinterpretation of the relationship between observations and variables, the tool oMEDA is very useful.

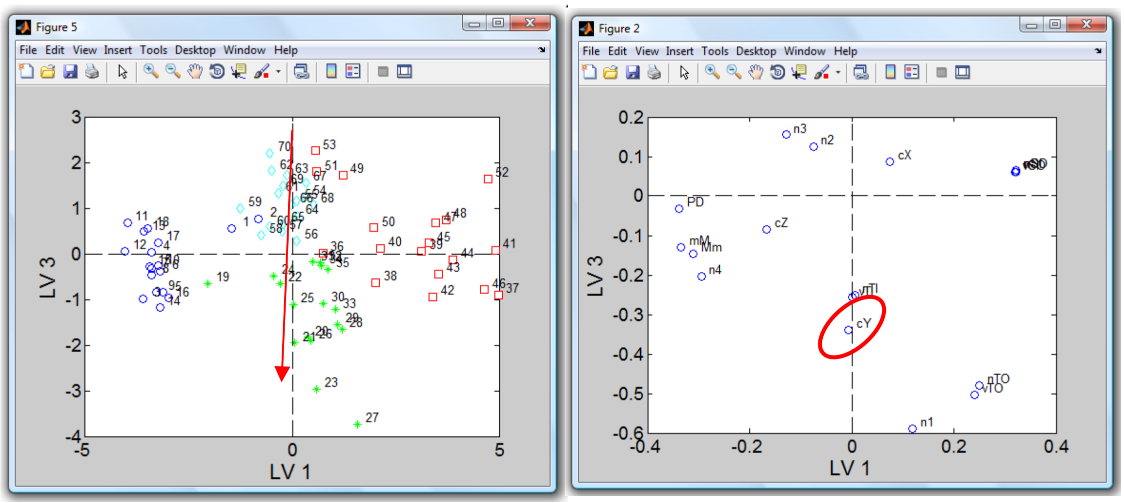
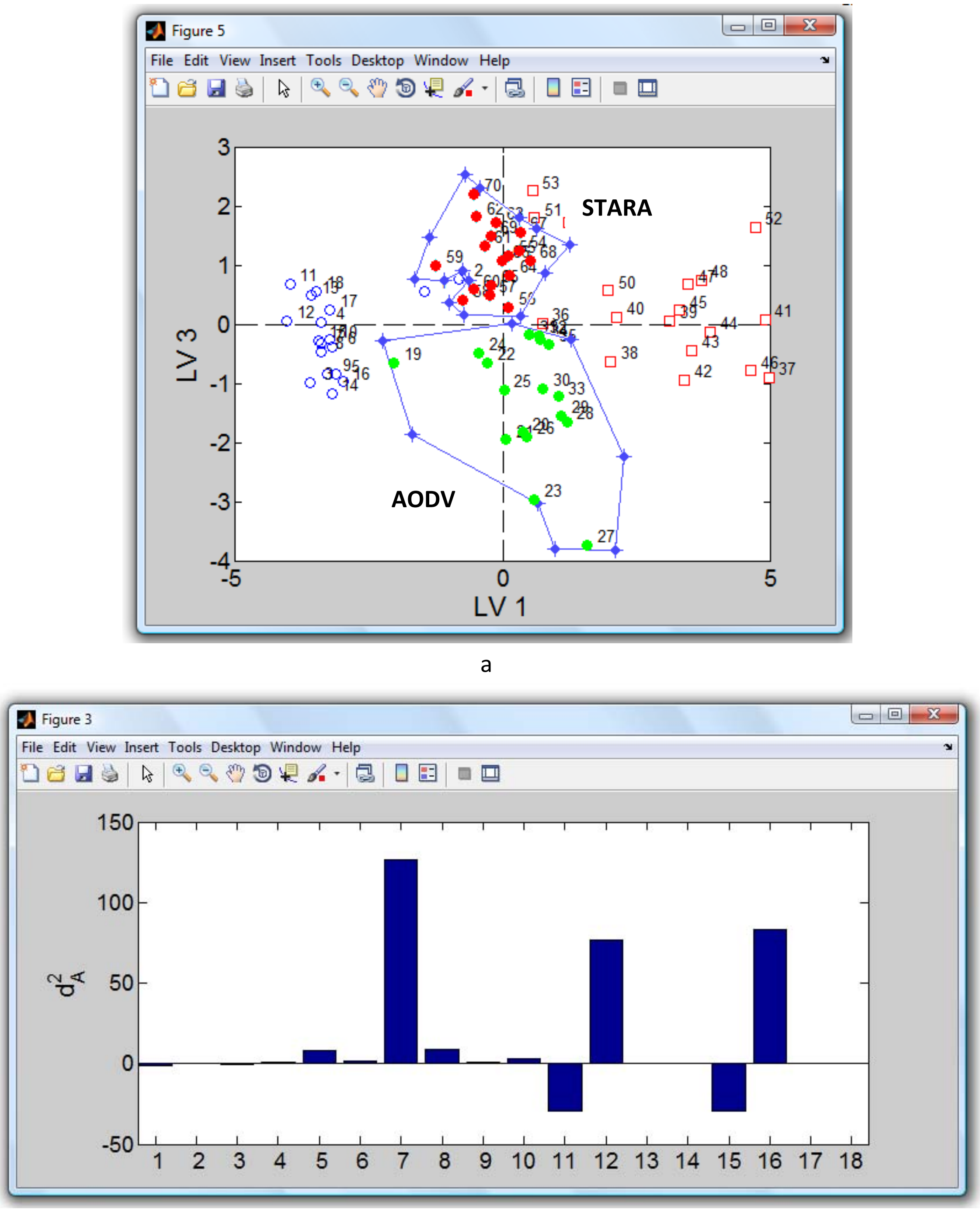


Fig. 9 Score plot (a) and loading plot (b) corresponding to LV1 vs LV3

To use oMEDA in the GUI, a number of steps are followed. Firstly, we need to select one score plot to work with, the one in figure 7(b). We are going to compare two clusters of scores: the blue one (these observations correspond to STARA) vs the green one (these observations correspond to AODV).

Select (click on) the score plot you are going to work with. Go to the oMEDA submenu in PLS interface and click on the *Select* button. If you put your mouse on the score plot, you will see a cross that allows you to draw an irregular polygon around the observations you want to select. Draw the polygon selecting the AODV scores and click on the + (plus) button. By doing this, we assign the value 1 in oMEDA to the selected scores. Repeat the operations, click on the *Select* button and draw a polygon around the STARA scores, and click over the – *(minus)* button. They are assigned the value -1 in oMEDA. At this point our score plot will look like the one in figure 10(a). The unselected scores take the value 0 and are not considered for the oMEDA plot. (*Note:* It is possible to add a trend line to include weights to each of theselected scores but we are not considering that possibility in our example.) Finally click on the *Plot* button. A graphic oMEDA plot like the one shown in figure 10(b) is generated. In an oMEDA plot, there is a bar for each of the variables. Thus, the Y axis in figure 10b represents numbers from 1 to 18, corresponding to the 18 variables in the data set. Alternatively, the name of the variables could be displayed to clarify the plot. To display the name of the variables, first define a vector containing the names, then go to the *Loading Plot* submenu and load that vector in the *Label* popup menu. By doing this the variables' names will be shown. Re-plot the previous oMEDA plot and you will obtain a graphic like the one in figure 11.



a

b

Fig.10 Score plot (a) and oMEDA result (b)

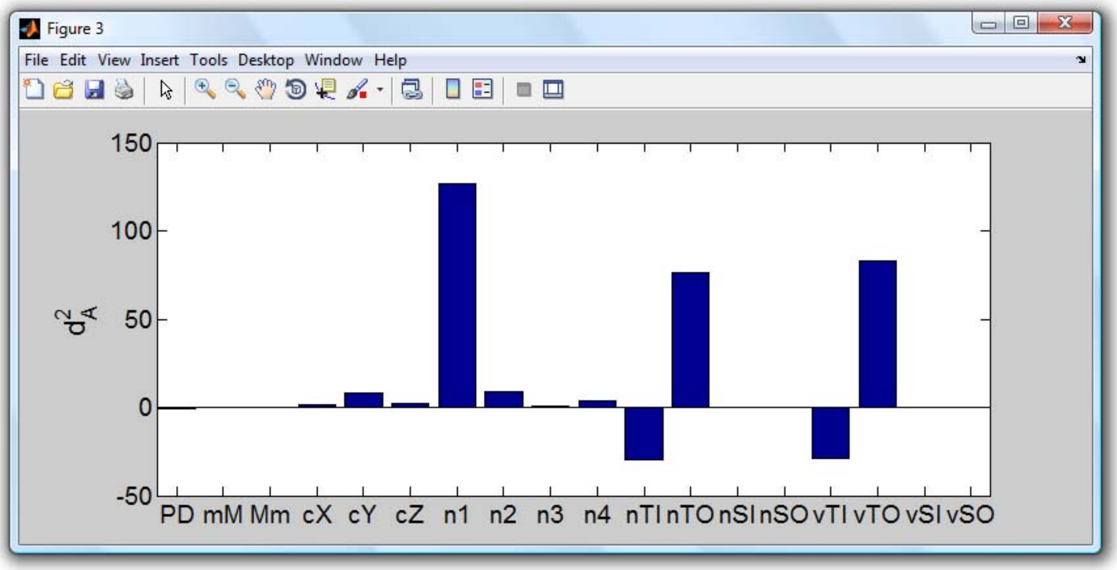


Fig. 11 oMEDA plot of fig. 10(b) with the name of the variables

The oMEDA plot of Figure 11 displays the differences between AODV and STARA. Remember that the value +1 was assigned to the AODV observations, and the value -1 to the STARA observations. Thus, the bars with a large positive value represent variables that take a larger value in the first group (AODV) in comparison to the second group (STARA), while negative values represent the other way round. Do you remember that we were expecting to solve the confusion caused in figure 9? Look at figure 11, variable cY does not take a high value for AODV. Thus, there is not a clear difference in terms of cY in both groups of observations. On the other hand, figure 11 shows some differences in the distribution of the laptops in the experiments of both groups (e.g. n1 is highly positive, where n1 is the amount of laptops with a distance to the centroid lower than 1/32 of the maximum distance, see Table 1), yet there are no differences in the global statistics of that distribution (e.g. PD is close to 0, where PD is the average distance between laptops). Also, it can be seen that STARA implies a low number of packets (e.g. nTO and vTO are positive, see Table 1).

We have already studied the observations (Score Plot) and the relationship between the observations and the variables (oMEDA). Now we are working with the variables alone. To plot the distribution of the variables in the model subspace, select the LVs you want to display on the Loading Plot by clicking on the *LV X-axes* and *LV Y-axes* popup menus, located in the *Loading Plot* submenu (area D). Push on the *Plot* button, to obtain the graphic. Figure 12 depicts the loading plot corresponding to LV1 vs LV3.

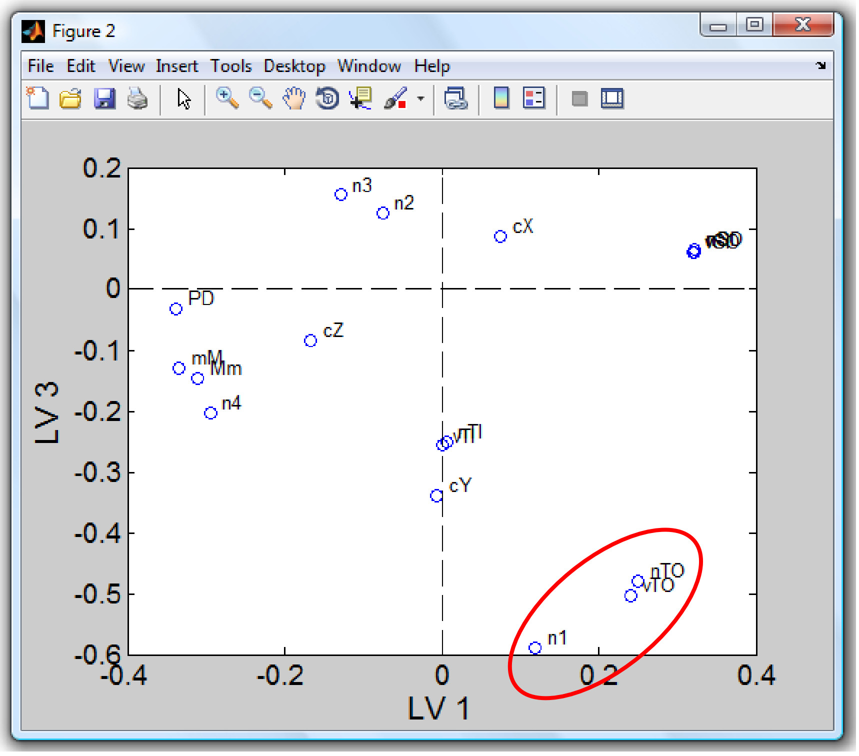


Fig. 12 Loading plot of LV1 vs LV3

Looking at figure 12, we could consider that variables n1, nTO and VTO are highly correlated, since they are quite near from each other in the LV1 vs LV3 subspace. However, this is not the case.

The following tool in the GUI is MEDA. MEDA is used to study the relationships among the features. This tool helps us to know which variables are related to others and to measure the level of that relationship. You will see positive correlations in red and negative correlations in blue. The more intense the color, the stronger the relationship. MEDA represents predictive correlation, which is a stronger measure of correlation, more robust to noise, than the traditional one. Dark red color means correlation 1 between the features and dark blue color means correlation -1. MEDA also reorders the features to simplify the interpretation of the plot, so that higher correlations are closer in the plot. To obtain the MEDA plot, select the LVs on the popup menu in the MEDA submenu (area F), select *Seriation* (to reorder the features) and click on the *Plot* button. A graphic like the one in figure 13 will be displayed.

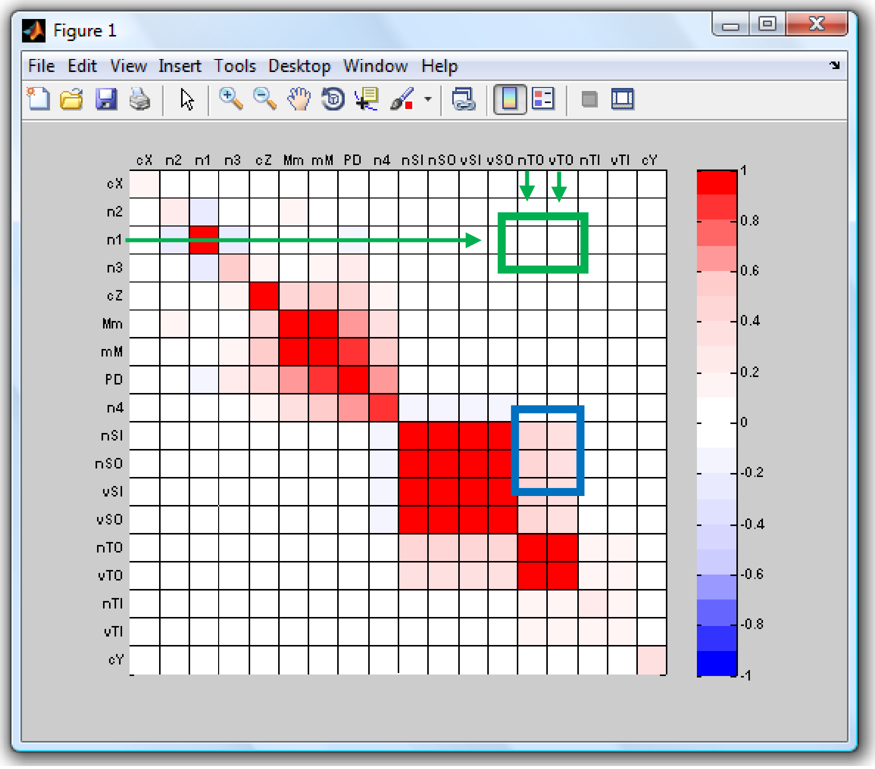


Figure 1
File Edit View Insert Tools Desktop Window Help
cX n2 n1 n3 cZ Mm mM PD n4 nSI nSO vSI vSO nTO vTO nTI vTI cY
cX
n2
n1
n3
cZ
Mm
mM
PD
n4
nSI
nSO
vSI
vSO
nTO
vTO
nTI
vTI
cY
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1

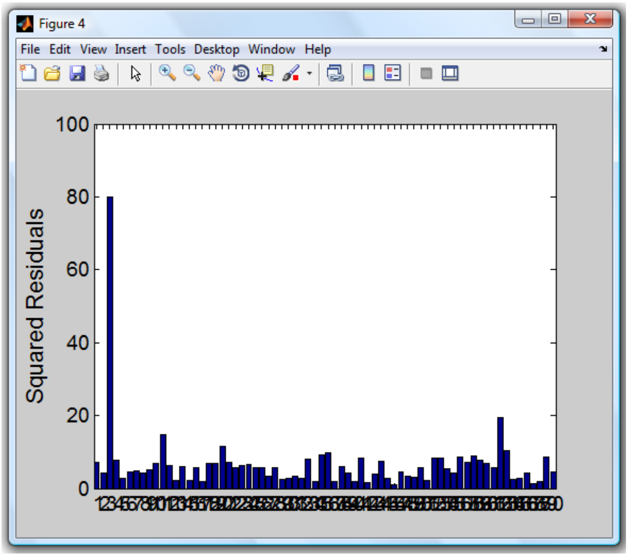
Fig.13 MEDA plot of LVs 1:3

As you can see there is no relationship between n1, and nTO or vTO (see the green square in figure 13). So, a suggestion is to be careful interpreting the Loading Plots and check the conclusions out with MEDA. Also, it can be seen that the number and the volume of packets are related. High packet retransmission (e.g. nSI, nSO) implies more probability of receiving packets (e.g. nTO) (see blue square in figure 13). Note that it is possible to obtain a MEDA plot on a similar way as done before with oMEDA. Plot a loading plot, go to the MEDA submenu and click over *Select*. Draw a polygon over the variables in the loading plot you want to study and you will obtain the MEDA plot. This way you can plot MEDA graphics set of a reduced set of features.

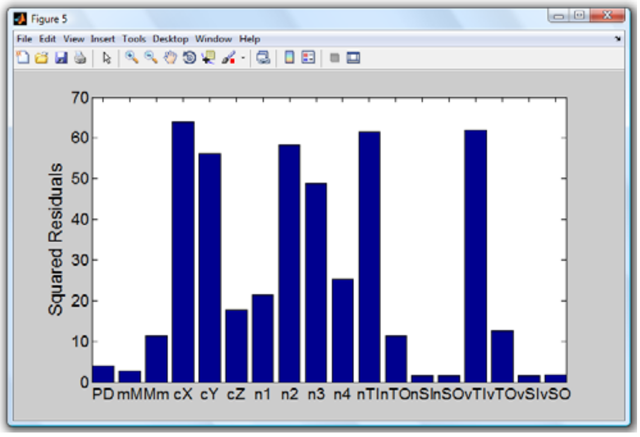
To finish the analysis, it is highly important to have a look at the residuals of the model. It is possible to obtain the residual plot for the observations and the variables in the data. In area G of the PLS interface you can obtain the residuals for the observation just pressing the *Plot* button in that area. To obtain the residue for the variables, repeat this action in area H. Figure 14 shows the residue for the observations (figure 14a) and the variables (figure 14b). This graphics are computed considering the LVs considered in the PLS model (area A).

In figure 14(a), observation number 3 shows up as an outlier. To study this phenomenon we could use oMEDA. That way we could find out which variables are affecting this observation and consequently its behavior.

Finally figure 14 (b) shows the residuals in the variables. Considering auto-scaling was selected as the preprocessing method, the sum of squares of each variable should be N-1, with N=70 observations. This means that there are features with no or very little load in the model. It maybe concluded that these variables are not relevant to differentiate the measured routing protocols.



(a)



(b)

Fig. 14 Residuals plots: (a) observations and (b) variables

References

[1] Kim H. Esbensen. Multivariate Data Analysis: in practice. Camo. ISBN: 82-993330-3-2.

[2] R.S. Gray, D. Kotz, C. Newport, N. Dubrovsky, A. Fiske, J. Liu, C. Masone, S. McGrath, Y. Yuan, Outdoor experimental comparison of four ad hoc routing algorithms, in: Proceedings of the 7th ACM international symposium on Modeling, analysis and simulation of wireless and mobile systems, MSWiM '04, 2004, pp. 220-229.

[3] M.Barker and W. Rayens, "Partial least squares for discrimination," Journal of Chemometrics, vol. 17, pp. 166-173, 2003.

[4] MEDA - Camacho, J. Observation-based missing data methods for exploratory data analysis to unveil the connection between observations and variables in latent subspace models. Journal of Chemometrics, 2011, 25 (11): 592-600.

[5] oMEDA- Camacho, J. Missing-data theory in the context of exploratory data analysis. Chemometrics and Intelligent Laboratory Systems, 2010, 103 (1): 8-18. ScienceDirect TOP25 Hottest Articles