SPA_GUI

Matlab graphical user interface allowing signal processing and variable selection for multivariate calibration.

www.ele.ita.br/~kawakami/spa

January/2012

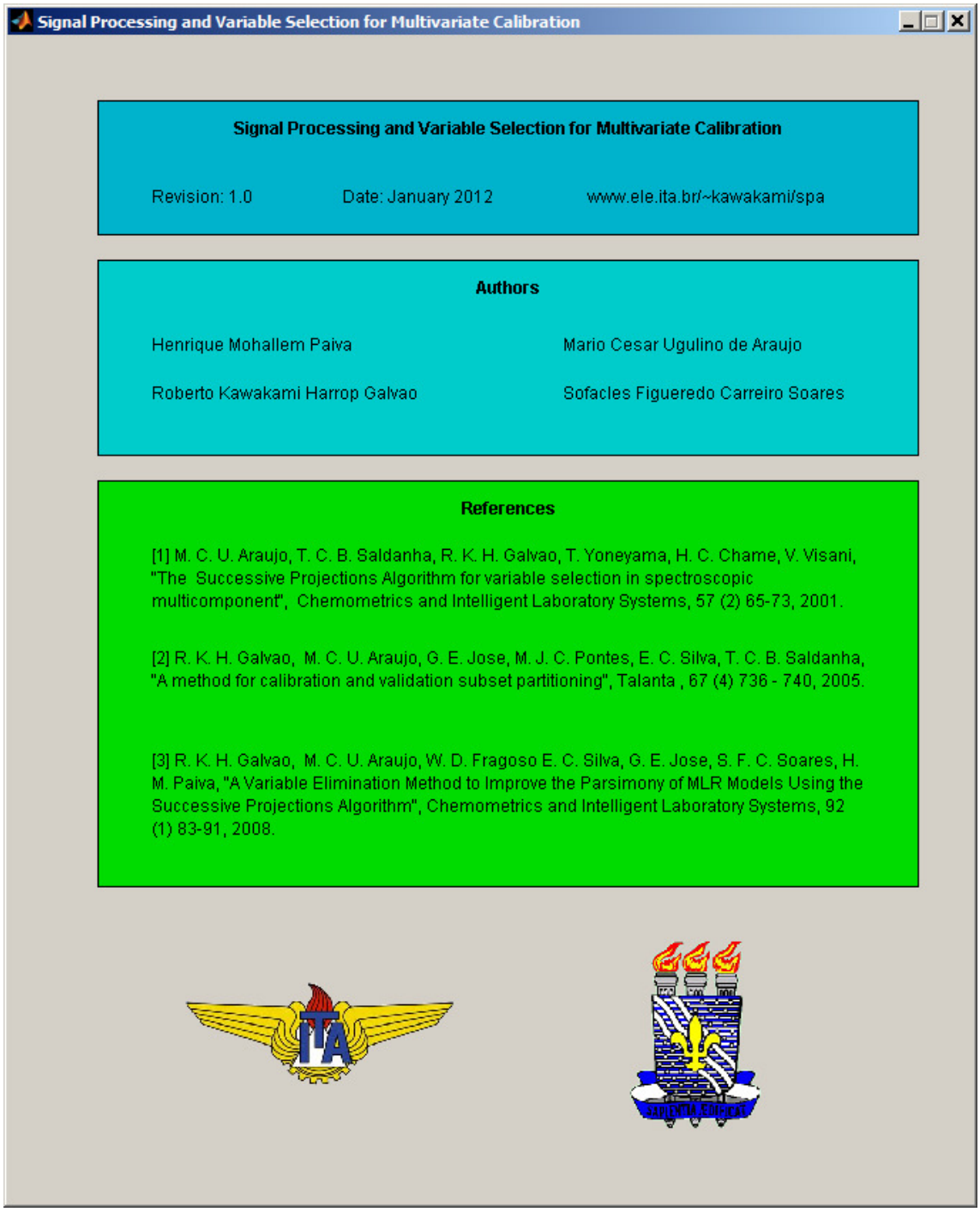
About SPA_GUI

SPA_GUI is a Matlab graphical user interface allowing signal processing and variable selection for multivariate calibration.

It is a freeware intended for academic and non-commercial use. It can be freely used and redistributed, provided that it will not be used for commercial purposes.

No guarantees are given for the quality of the software or for the consequences of its use.

The authors are very interested in feedback from the users.

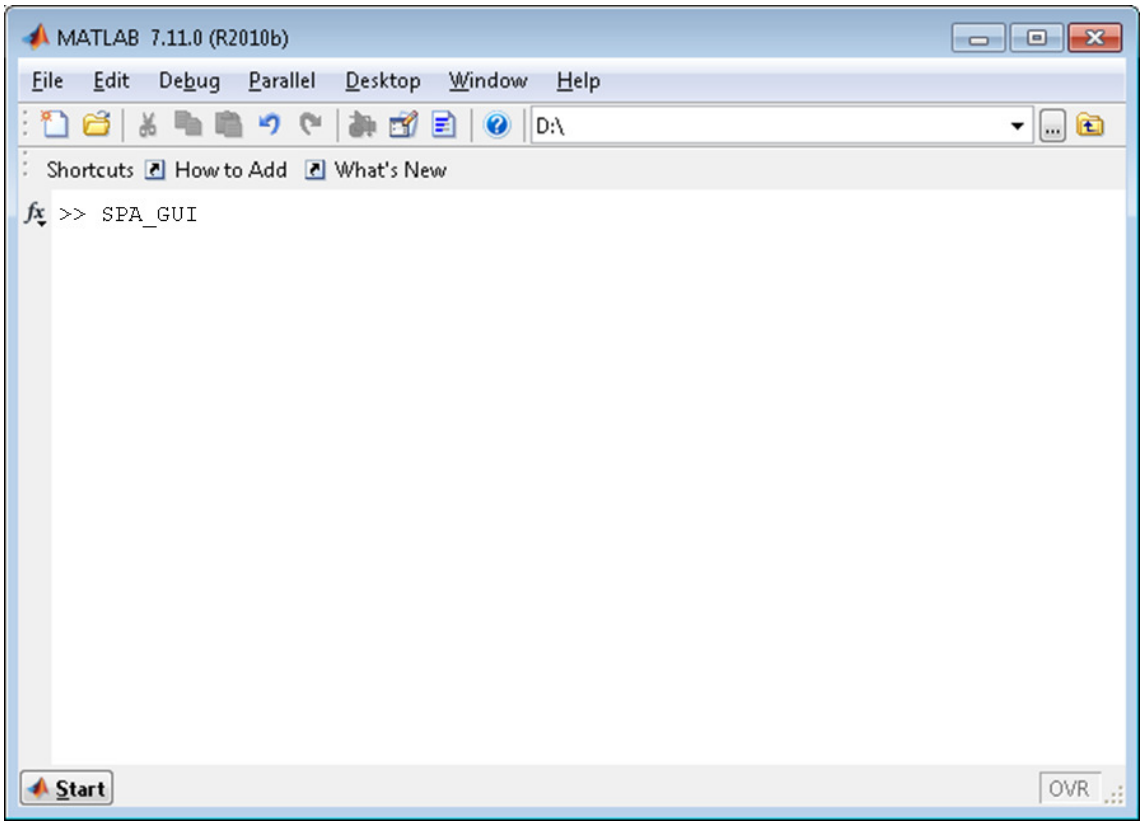


Contents

1. Starting the GUI 4
2. Main Menu 7
2.1. Menu: File 7
2.2. Menu: Workspace 8
2.3. Menu: Figures 8
2.4. Menu: Interface 8
2.5. Menu: About 8
3. "Load Data" page 9
3.1. Data File 10
3.2. Data Matrices in the workspace 10
3.3. Objects 11
3.4. Plot selected matrices and objects 12
4. "Data Pre-Processing" page 13
4.1. Savitzky-Golay Smoothing and Savitzky-Golay Differentiation 14
4.2. Wavelet Denoising 15
5. "Sample Selection" page 16
6. "SPA" page 18
6.1. Successive Projections Algorithm 19
6.2. Prediction 20
7. Example of use 21
7.1. Data set 21
7.2. Loading the data 21
7.3. Preprocessing 25
7.4. Selection of samples 28
7.5. Variable selection for multivariate calibration using SPA 37
7.5.1. Building the model 37
7.5.2. Prediction of samples 39
7.6. Saving your work 41

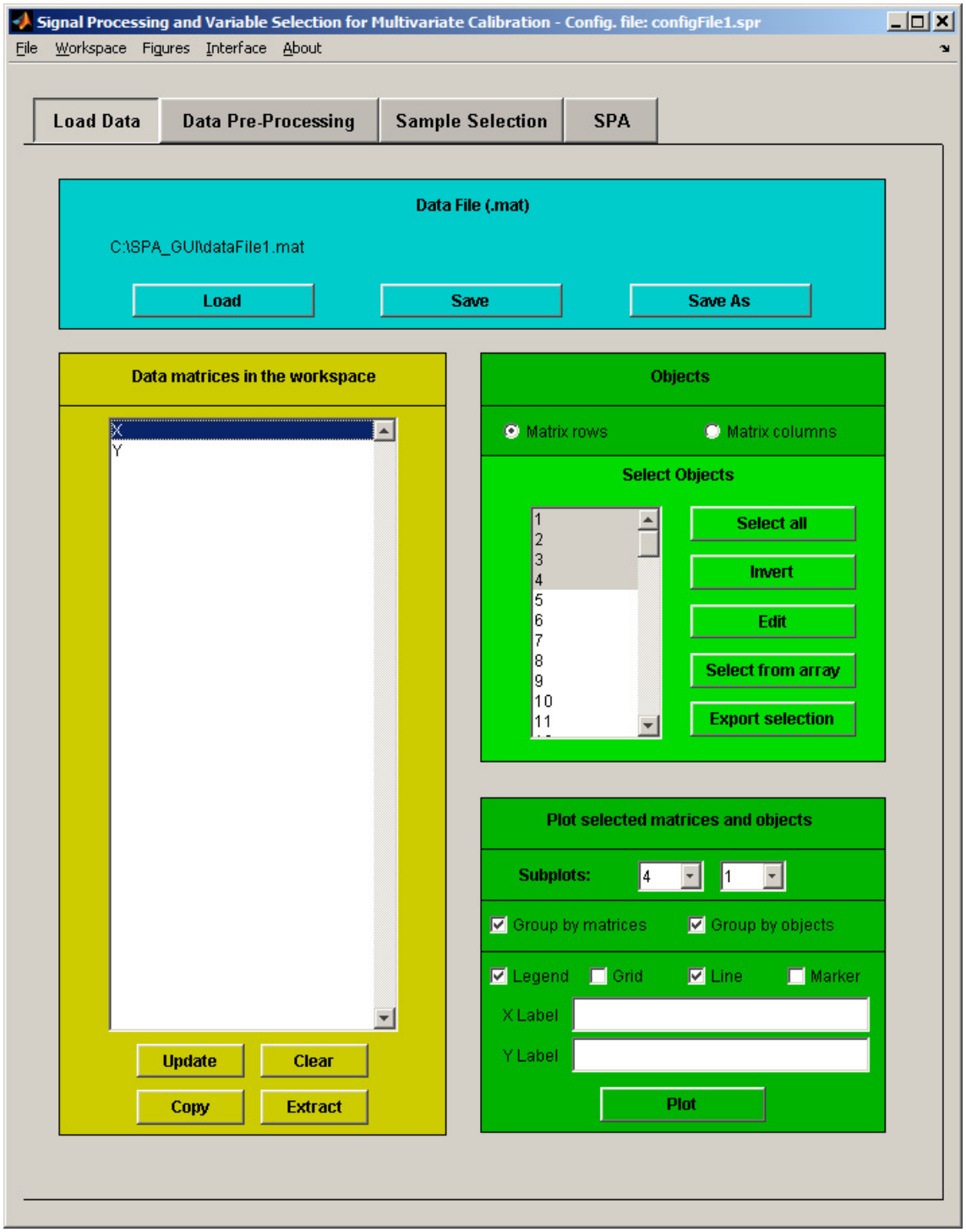
1. Starting the GUI

SPA_GUI consists of a single, self-contained p-file (pre-parsed Matlab m-file). Copy this file to your computer and type `SPA_GUI` at the Matlab Command Window to start the graphical user interface. Note that a p-file is used in the same way as an m-file.

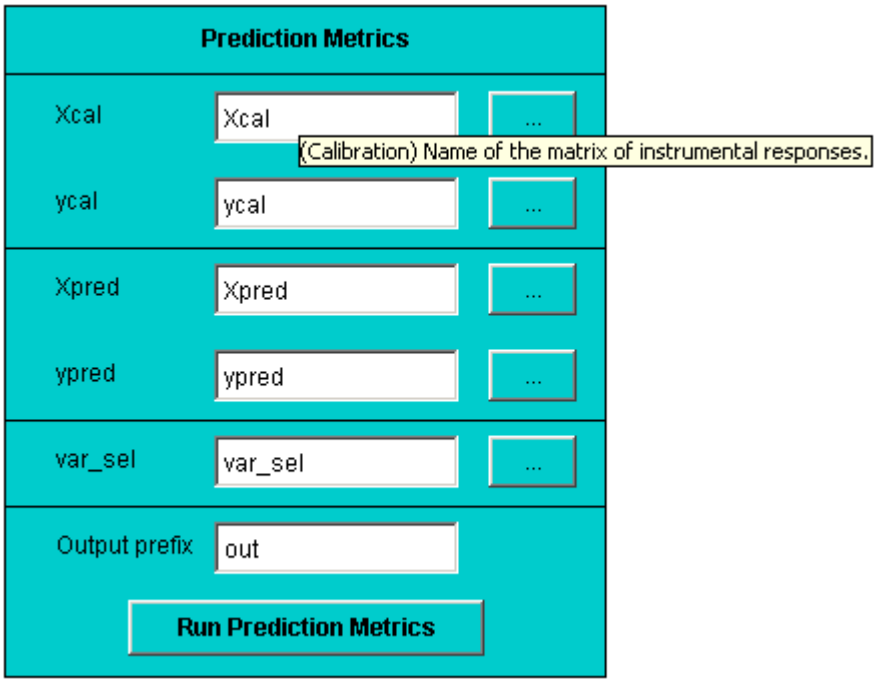


The main screen of the interface is presented on the next page.

The interface provides a main menu and the following modules: “Load Data”, “Data Pre-Processing”, “Sample Selection”, and “SPA”. These elements are described in the sections below. An example of use is presented in Section 7.


Signal Processing and Variable Selection for Multivariate Calibration - Config. file: configFile1.spr
File
Workspace
Figures
Interface
About
Load Data
Data Pre-Processing
Sample Selection
SPA
Data File (.mat)
C:\SPA_GUI\dataFile1.mat
Load
Save
Save As
Data matrices in the workspace
X
Y
Update
Clear
Copy
Extract
Objects
Matrix rows
Matrix columns
Select Objects
1
2
3
4
5
6
7
8
9
10
11
Select all
Invert
Edit
Select from array
Export selection
Plot selected matrices and objects
Subplots:
4
1
Group by matrices
Group by objects
Legend
Grid
Line
Marker
X Label
Y Label
Plot

Tip: For most elements in the interface, help textboxes are presented when the mouse pointer is placed on top of the element.



2. Main Menu

The main menu provides the following options:

- File: “Load”, “Save”, “Save As”, “Save and Quit”, “Quit without Saving”

- Workspace: “Clear All Workspace Variables”, “Close All Simulink Models”, “Clear Screen”, “Add Directory to Path”, “Add Directory and Subdirectories to Path”, ‘Restore Default Path”

- Figures: “Copy to Word as Meta File”, “Copy to Word as BMP”, “Close All”

- Interface: “Colors”, “Increase Font Size”, “Decrease Font Size”, “Expand Screen”, “Contract Screen”, “Default Screen Size’

- About: “About”

Tip: Main menu options can be quickly accessed by using “Alt” combinations. For example, hitting the keys “Alt” “F” “L” will give access to the option “File: Load”. Note that, when pressing the “Alt” key, the corresponding letter of each option in the menu will be underlined. Some menu options have also a “Ctrl” combination.

2.1. Menu: File

This submenu is used to manage a configuration file.

A **configuration file** is a file with .spr extension that is used to store all the parameters specified by the user in the graphical user interface.

The submenu provides the standard options “Load”, “Save”, “Save As”, “Save and Quit”, and “Quit without Saving”. When the latter is selected, the program will ask for confirmation before quitting.

Warning: The “Save”, “Save As”, and “Save and Quit” options in the main menu will save only the configuration file (i.e, the graphical user interface parameters). The data matrices in the workspace will not be saved. To save the data matrices, use the button “Save” in the “Load Data” module (see section 3.1).

Tip: In order to save time, the interface automatically loads the last configuration file (when available).

2.2. Menu: Workspace

This submenu is used to manage Matlab workspace.

The following table shows the menu option and the corresponding Matlab command:

Option	Matlab command
Clear All Workspace Variables	`clear all`
Close All Simulink Models	`bdclose all`
Clear Screen	`clc`

The “Workspace” submenu contains also options to manage the Matlab path.

2.3. Menu: Figures

The “Figures” submenu contains options to copy all opened figures to Word in metafile or bitmap format, and also allows closing all opened figures.

2.4. Menu: Interface

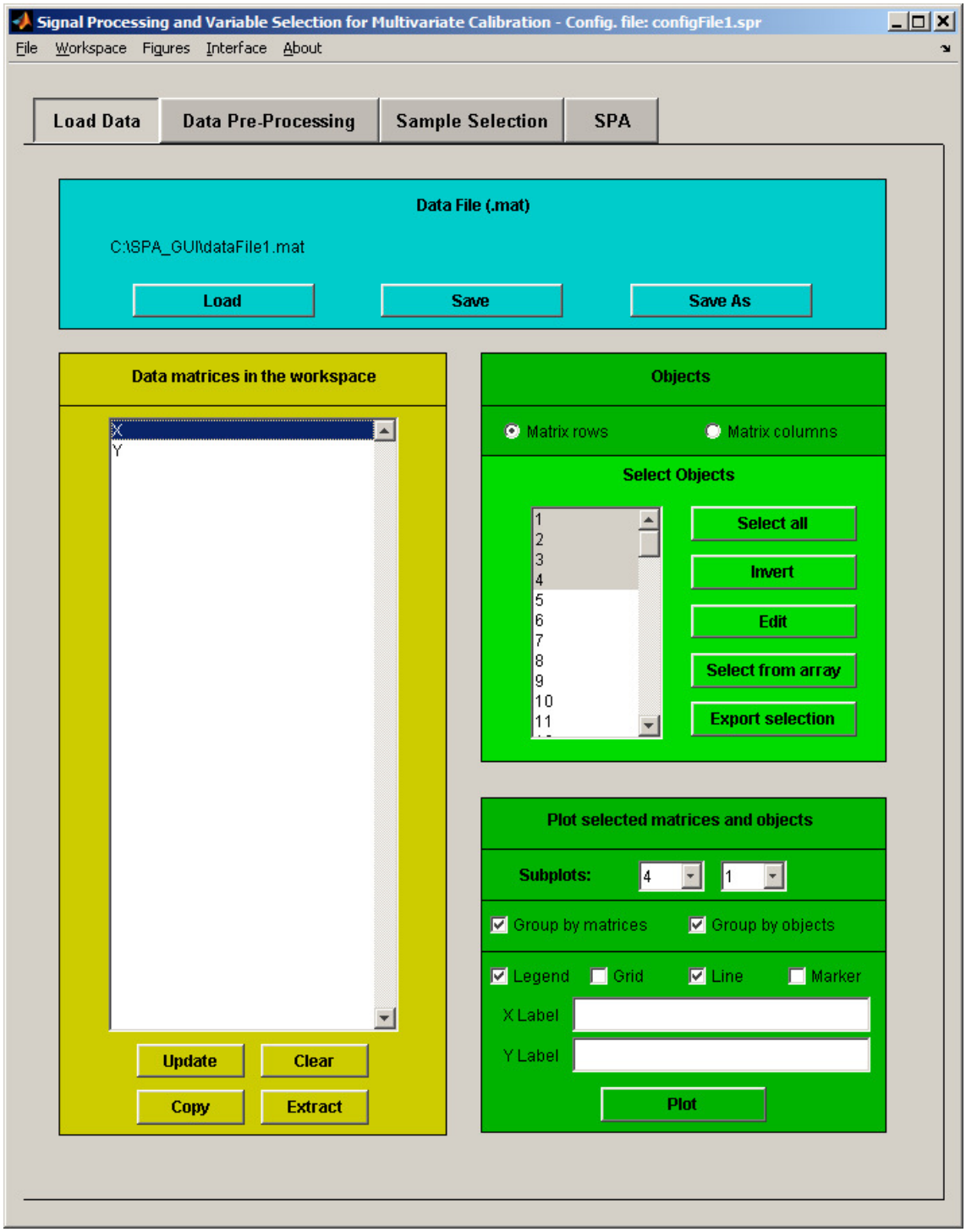
This submenu is used to change the interface, allowing changing its colors, its size, and the font size.

2.5. Menu: About

This submenu shows the “About” screen, with information about the software.

3. “Load Data” module

This module contains the following groups: “Data File”, “Data Matrices in the workspace”, “Objects”, and “Plot selected matrices and objects”.



3.1. Data File

This group is used to manage a data file.

A **data file** is a standard .mat file that is used to store data matrices (numeric Matlab variables).

The group contains the following buttons:

- “Load”: used to load the contents of a data file to the workspace.

- “Save” and “Save As”: used to save all data matrices available in the workspace to the data file.

The name of the selected data file is presented in the screen and is stored in the configuration file.

3.2. Data Matrices in the workspace

This group is used to manage the data matrices available in the Matlab workspace.

A **data matrix** is any numerical Matlab variable.

Note: Other Matlab data types, such as structures, are not considered by this interface as data matrices.

A list of the data matrices available in workspace is presented. The user can select one or more of these matrices.

The group contains also the following buttons:

- “Update”: used to update the list of data matrices available in the workspace.

- “Clear”: used to clear the selected data matrices.

- “Copy”: used to copy the selected matrices to new matrices, whose name will be specified by the user.

- “Extract”: used to extract part of the selected matrices to new matrices, whose name will be specified by the user. Only the selected objects will be copied to the new matrices. For instruction on how to select the objects, see section 3.3.

Example: Suppose that the matrix X is selected, the objects are specified as matrix rows and the objects 1,2,5,7 are selected. Then, by using the extract button, a new matrix, equal to X([1 2 5 7],:), will be created.

Note: The difference between “Copy” and “Extract” buttons is that the first is used to copy the entire matrix, while the last is used to extract only part of the matrix.

3.3. Objects

This group is used to manage the objects, which are rows or columns of the data matrix.

In this group, the user inform if the objects are matrix rows or matrix columns. A list of available objects is then presented. In this list, the user can use the mouse to manually select one or more objects.

The group contains also the following buttons:

- “Select all”: used to select all elements.

- “Invert”: used to invert the selection (that is, objects that were selected are deselected, and vice-versa).

Note: The “Invert” button will not work if all objects are selected.

- “Edit”: used to inform through the keyboard the indices of the objects to be selected.

- “Select from array”: used to select the elements specified in an array of indices.

- “Export selection”: used to export the selected indices to an array of indices (whose name will be specified by the user).

Warning: The “Export selection” button will export only the selected indices. To extract the corresponding objects from the data matrix, use the button "Extract" in the "Data matrices in the workspace" group (see section 3.2).

3.4. Plot selected matrices and objects

This group is used to plot the selected objects of the selected matrices.

The user can use the subplot option to select the number of subplots in each figure.

The plots can be grouped by matrices and/or by objects. When the “Group by matrices” option is checked, all signals of the same matrix are plotted in the same figure. When the “Group by objects” option is checked, signals of different matrices, corresponding to the same object, are plotted in the same figure.

Example: If matrices A, B, C are selected, “Group by objects” is checked, and the objects are matrix rows, then: A(1,:), B(1,:), C(1,:) will be plotted together; A(2,:), B(2,:), C(2,:) will be plotted together, etc.

One typical use of this option is to plot a processed signal against its original signal. For instance, if matrix B was obtained by smoothing matrix A, the user may want to compare each object in B with the corresponding object in A.

Tip: To group all plots in only one figure, check both the options “Group by matrices” and “Group by objects”. To present only one plot per figure, uncheck both these options.

By checking or unchecking the corresponding boxes, it is possible to specify if a legend and a grid shall be used in the plots. It is also possible to specify if a line and/or a marker shall be used in each plot.

Two edits are used to inform the labels to be used in the x and y axes.

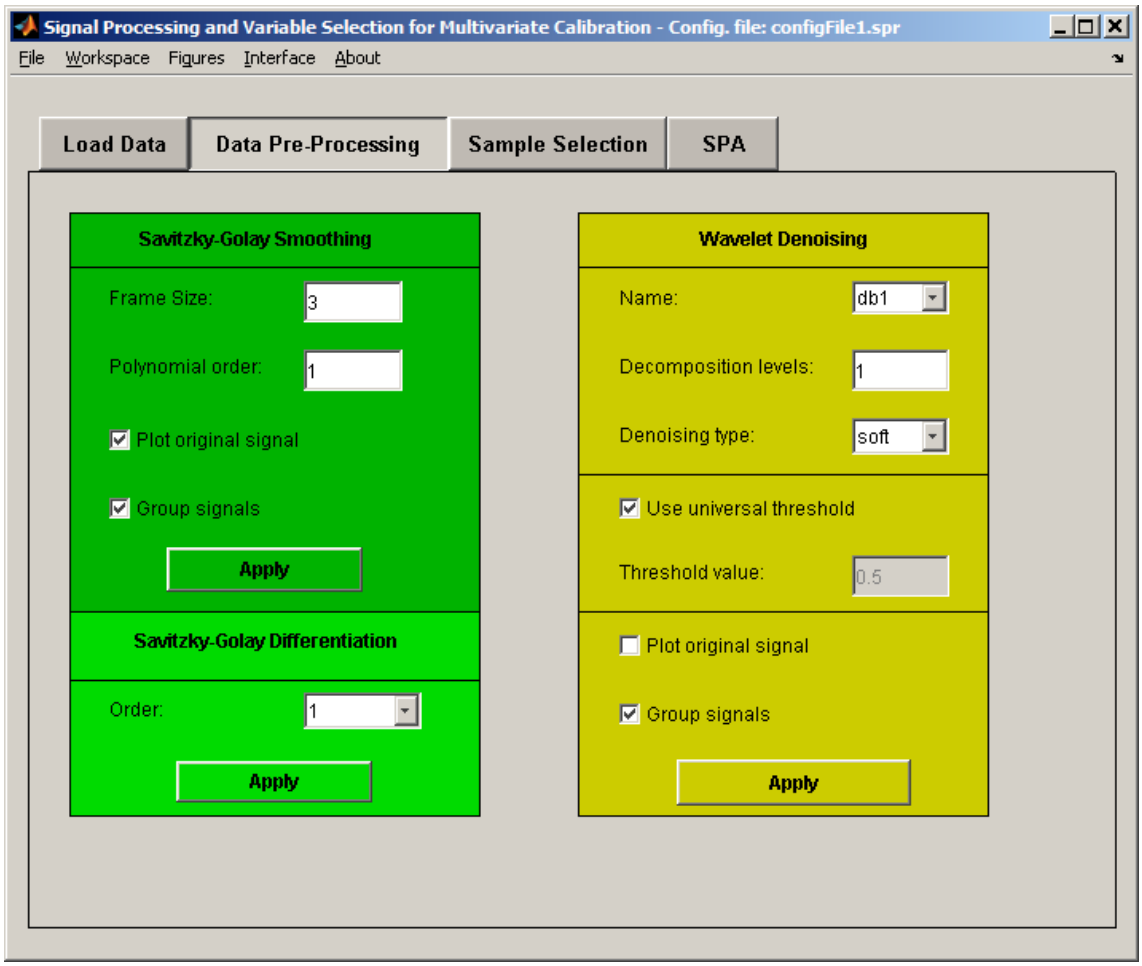
The “Plot” button is used to generate the specified plots.

Note: The “Plot” button will not close previous figures. To close all opened figures, use the option “Figures: Close All” in the main menu.

4. “Data Pre-Processing” module

This module contains the following groups: “Savitzky-Golay Smoothing”, “Savitzky-Golay Differentiation”, and “Wavelet Denoising”.

Note: This module will process the data matrices and objects that were select in the “Load Data” module.



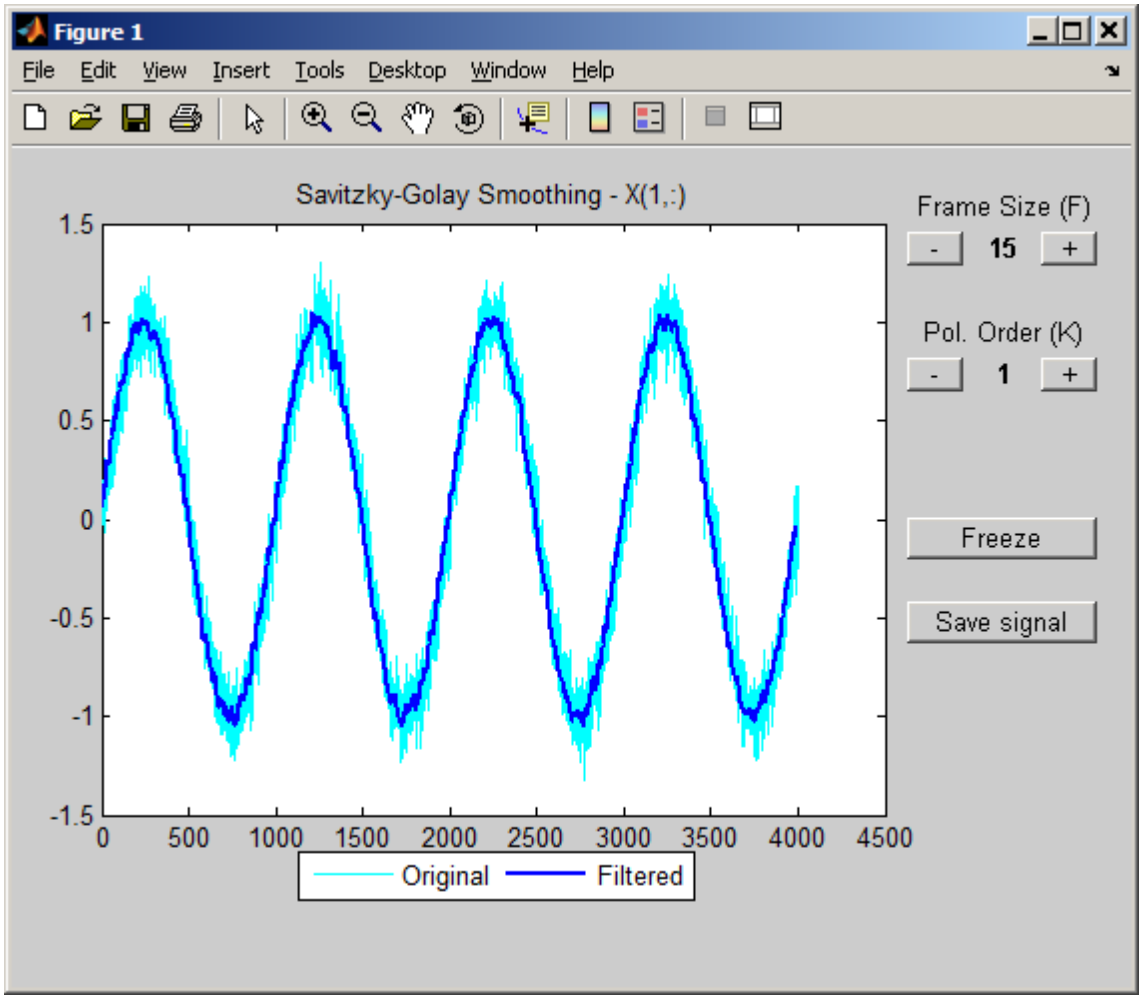
4.1. Savitzky-Golay Smoothing and Savitzky-Golay Differentiation

In these groups, the user shall specify the frame size and the polynomial order for Savitzky-Golay filtering. The frame size must be odd, and the polynomial order must be less than the frame length. If invalid parameters are entered, error messages will appear.

By checking or unchecking the corresponding boxes, it is possible to specify if the original signals shall be plotted. It is also possible to specify if all signals shall be grouped in only one figure or if each signal shall be plotted in a different figure.

The same parameters specified for Savitzky-Golay smoothing will be also used for Savitzky-Golay differentiation. To run the differentiation, it is also necessary to specify the differentiation order (1 or 2, meaning first or second derivative).

The “Apply” buttons are used to run Savitzky-Golay smoothing and Savitzky-Golay differentiation. When these buttons are pressed, the Savitzky-Golay screen is presented.



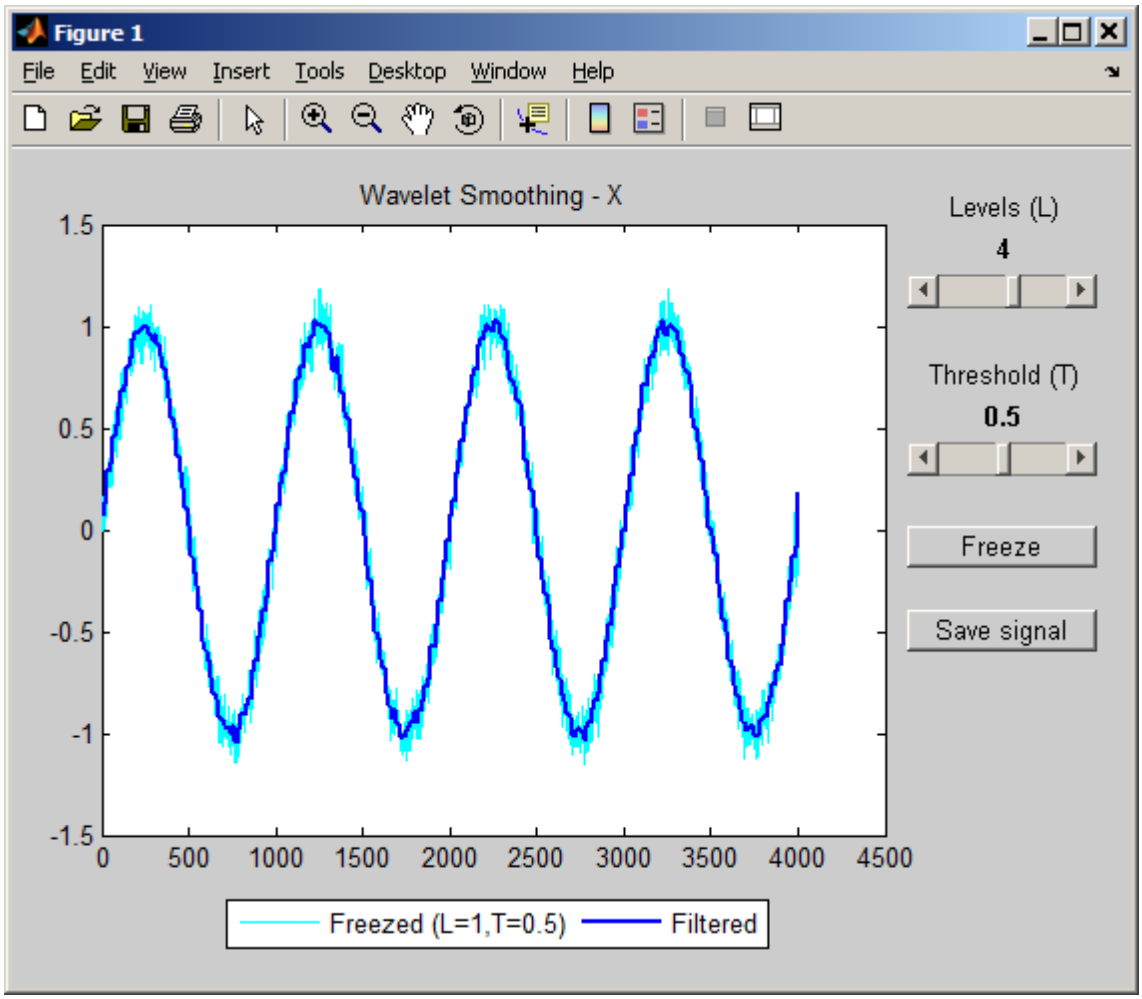
The Savitzky-Golay screen presents buttons “+” and “-”, to increase or decrease the frame size and the polynomial order. The corresponding filtering result is updated according to the selected configuration. The “Freeze” button is used to freeze the current configuration, such that it can be compared with another configuration. The “Save signal” button is used to save the filtered signal with a name specified by the user.

4.2. Wavelet Denoising

In this group, the user shall specify the wavelet filter name (selected from a list), the number of resolution levels, the denoising type (soft or hard) and the threshold for denoising. The threshold may be set as universal or manual, by checking or unchecking the “Use universal threshold” box. If manual threshold is selected, the user must specify the threshold value.

By checking or unchecking the corresponding boxes, it is possible to specify if the original signals shall be plotted. It is also possible to specify if all signals shall be grouped in only one figure or if each signal shall be plotted in a different figure.

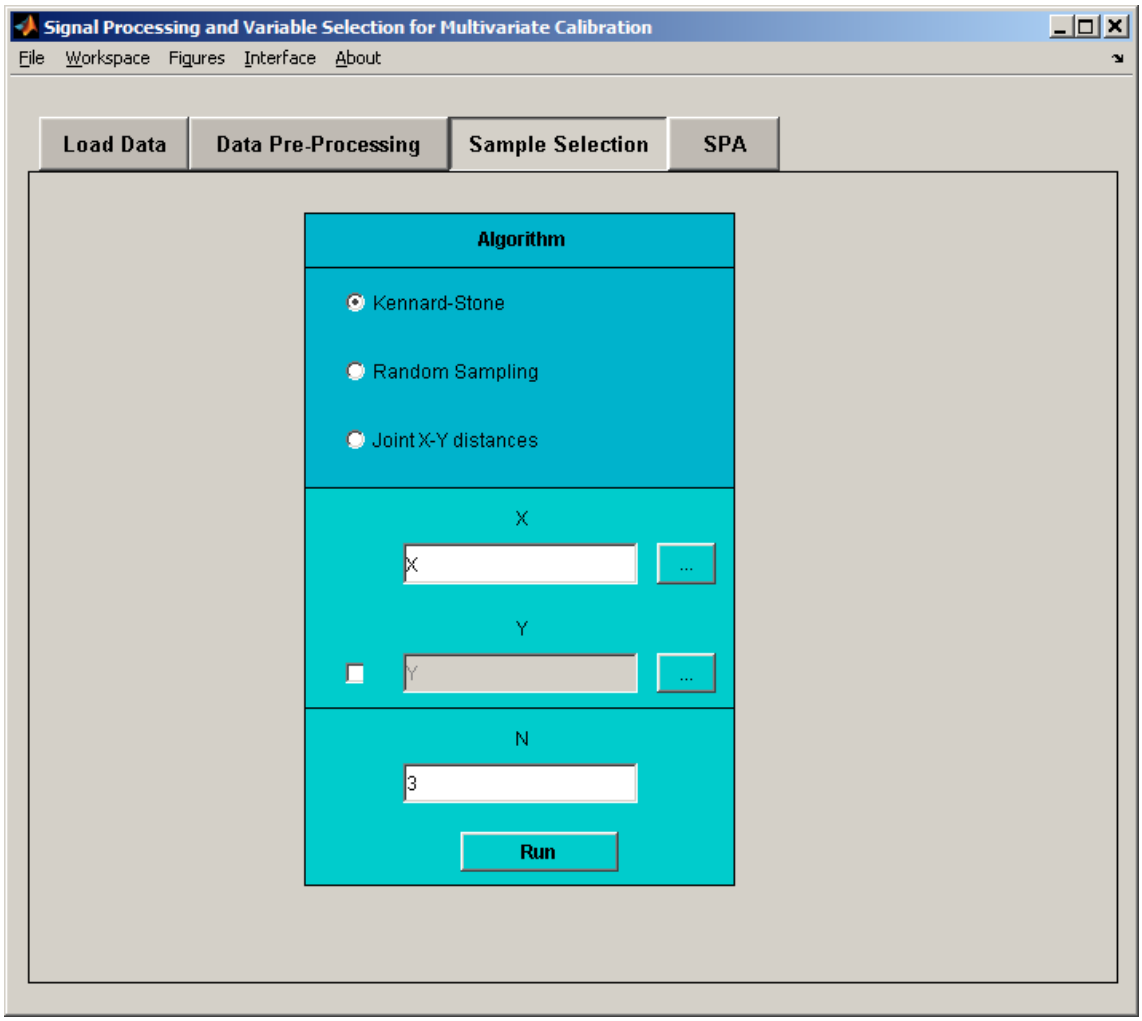
The “Apply” button is used to run Wavelet denoising. When this button is pressed, the Wavelet denoising screen is presented.



The Wavelet denoising screen presents sliders to increase or decrease the number of resolution levels and the threshold (provided that the “universal threshold” option has not been selected). The corresponding filtering result is updated according to the selected configuration. The “Freeze” button is used to freeze the current configuration, such that it can be compared with another configuration. The “Save signal” button is used to save the filtered signal with a name specified by the user.

5. “Sample Selection” module

This module is used to run the following algorithms for sample selection: Kennard-Stone, Random sampling, and Sample set partitioning based on joint X-Y distances (SPXY).



The user shall select the algorithm to be run and the following parameters:

- X: Name of the matrix of instrumental responses.
- Y: Name of the matrix of parameters.
- N: Value specifying the number of samples to be selected (minimum of 2).

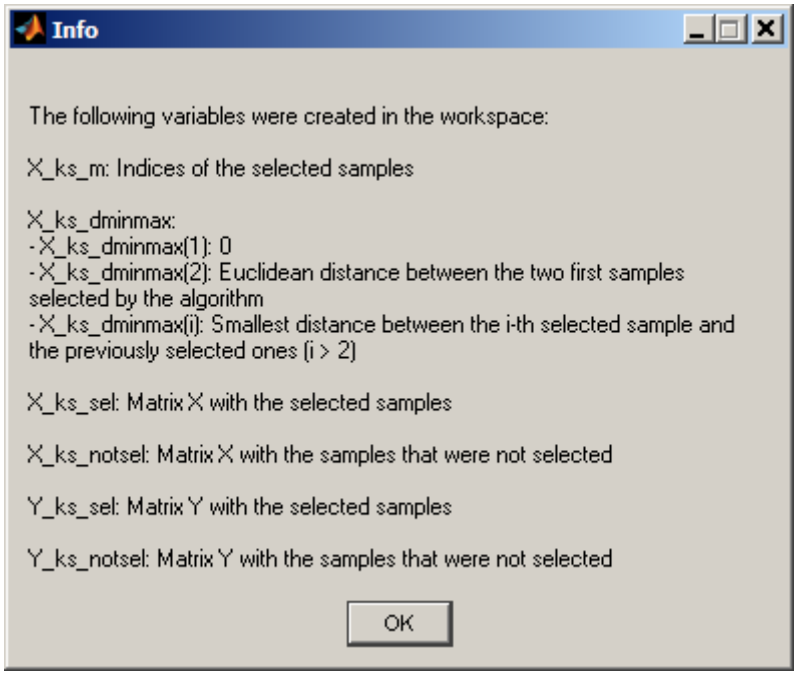
The matrix Y is required by Joint X-Y distances algorithm and opcional for the two other algorithms. If the user wants to specify the Y matrix for these algorithms as well, he shall check the box in the left of the edit specifying the name of the Y matrix.

There are two ways to specify the name of the X and Y matrices:

a) by typing the name of the matrix in the corresponding edit.

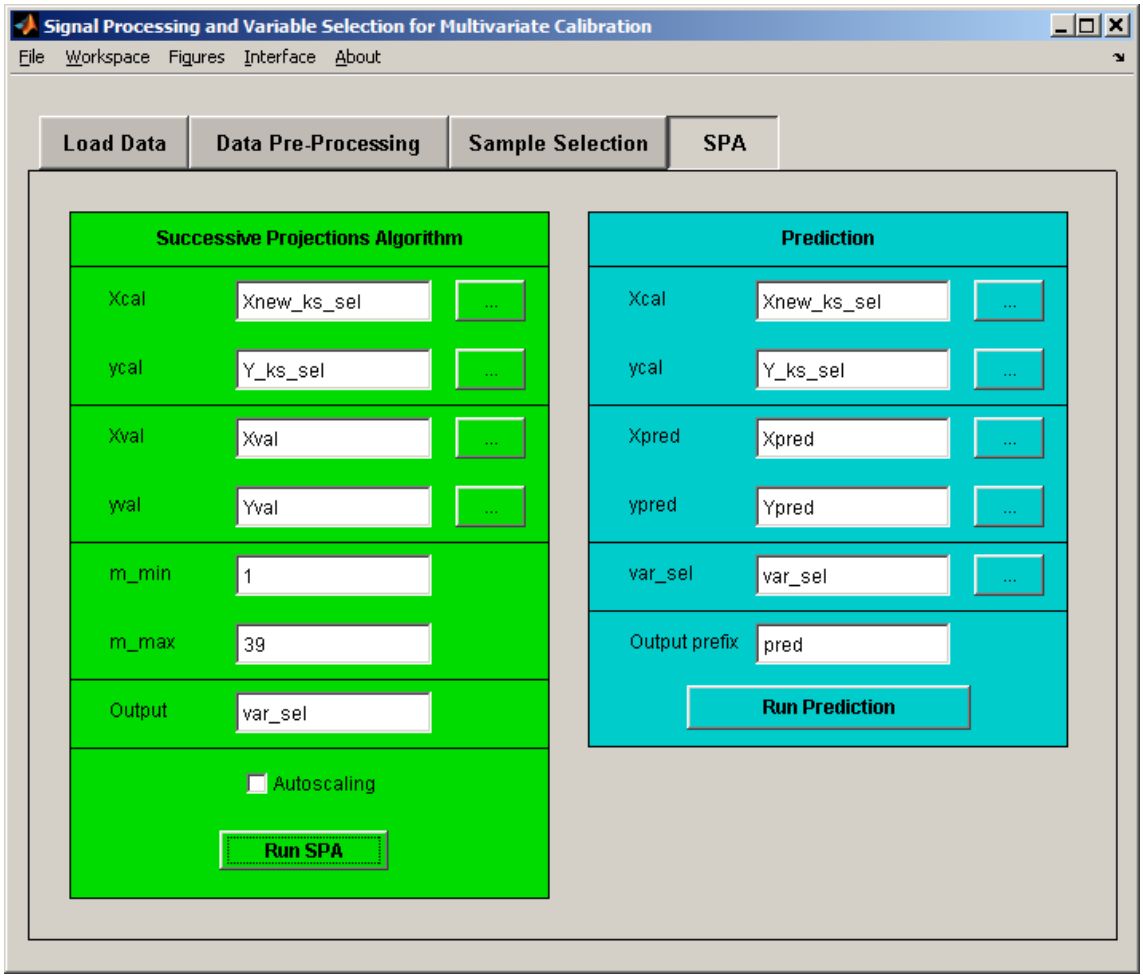
b) by choosing the matrix from a list of available matrices. When the "…" button is pressed, this list is shown, allowing the user to choose the desired matrix.

The "Run" button is used to run the selected algorithm, whose results are saved in Matlab workspace. A message box appears informing the name and meaning of the variables that were created in the workspace. The name of each output variable starts with a prefix specifying the name of the original variable and the selection method.



6. “SPA” module

This module contains the following groups: “Successive Projections Algorithm” and “Prediction”.



6.1. Successive Projections Algorithm

In this group, the user shall specify the X (matrix of instrumental responses) and Y (matrix of parameters) matrices for calibration and validation. If the validation fields are left empty, leave-one-out cross-validation will be carried out.

There are two ways to specify the name of the X and Y matrices:

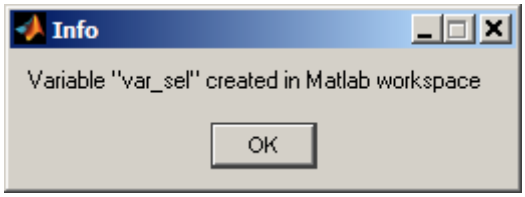
a) by typing the name of the matrix in the corresponding edit

b) by choosing the matrix from a list of available matrices. When the "…" button is pressed, this list is shown, allowing the user to choose the desired matrix.

The fields of minimum and maximum number of variables (m_min and m_max) are optional. If left blank, default values will be assumed. In case of user-assigned values, the maximum number of variables must be less than number of calibration samples minus one.

The data are mean-centered prior to the projection procedures. If the "Autoscaling" checkbox is selected, autoscaling will also be carried out.

The "Run SPA" button is used to run the algorithm. The indexes of the selected variables will be placed in the workspace, in an array whose name is specified in the "Output" edit. A message box appears informing the variable that was created in the workspace.



6.2. Prediction

In this group, the user shall specify the X (matrix of instrumental responses) and Y (matrix of parameters) matrices for calibration and prediction, as well as the name of the "var_sel" array (array containing the indexes of the selected variables - output of SPA algorithm).

If Xpred is informed and ypred is left blank, validation with a separate set with no reference values for y is performed. If both Xpred and ypred are left blank, leave-one-out cross-validation will be carried out.

There are two ways to specify the name of the X and Y matrices and the "var_sel" array:

a) by typing the name of the matrix in the corresponding edit

b) by choosing the matrix from a list of available matrices. When the "…" button is pressed, this list is shown, allowing the user to choose the desired matrix.

The "Run Validation Metrics" is used to obtain a Predicted vs Reference plot with a bisectrix line and one-sigma confidence intervals, as well as the following statistics associated to prediction performance:

- PRESS (Prediction Errors Sum of Squares)

- RMSEP (Root Mean Square Error of Prediction)

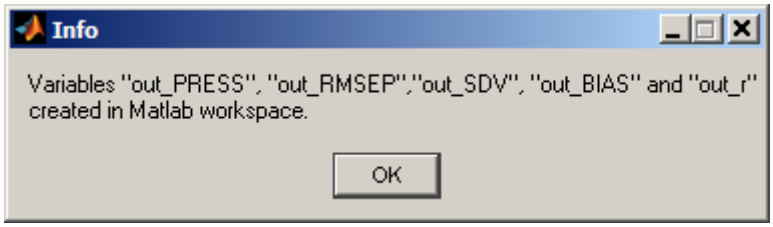
- SDV (Standard Deviation of the Prediction Errors)

- BIAS (Average of the Prediction Errors)

- r (Correlation Coefficient between Predicted and Reference values)

The output arrays will start with the prefix specified by the user in the "Output Prefix" field. For instance, if "Output Prefix" is "out", then the output arrays will be called "out_PRESS", "out_RMSEP", "out_SDV", "out_BIAS" and "out_r".

A message box appears informing the variables that were created in the workspace.



7. Example of use

na verdade, de acordo com as informações do dataset, é até 2498mm

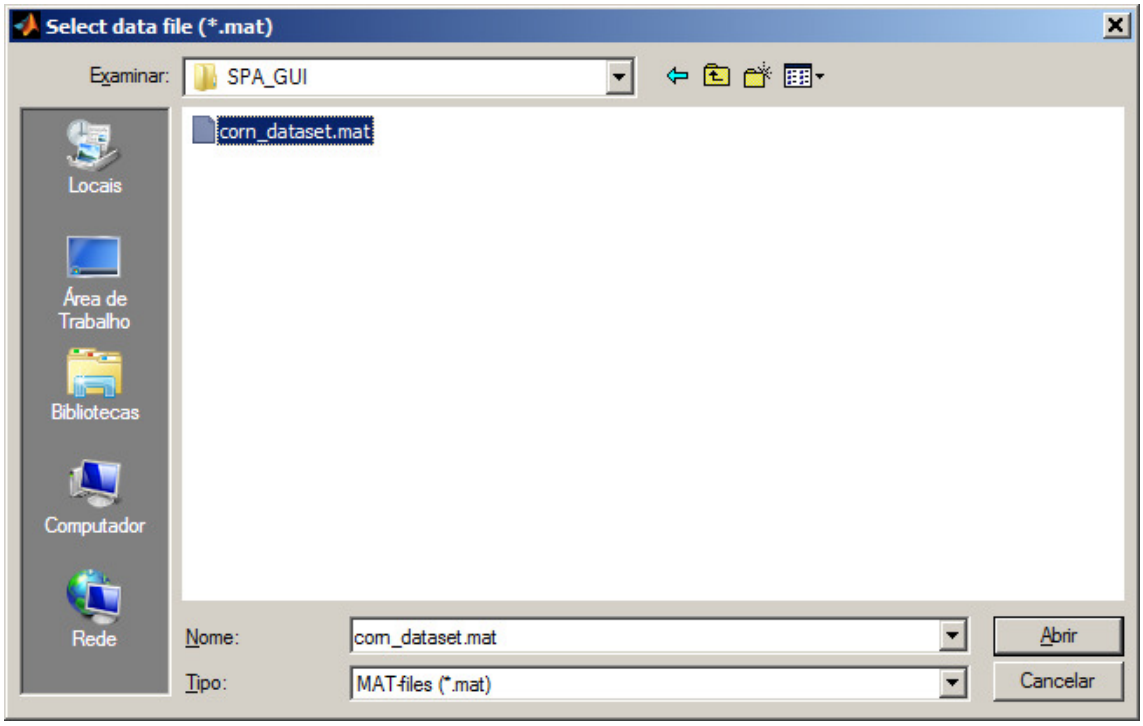
7.1. Data set

The data set used in this example consists of spectra from 80 corn samples, which were acquired in the range 1100–1498 nm, together with moisture content for each sample. The corn data are publicly available at http://www.eigenvector.com/Data/Corn/. In this example, only the equipment “m5” in this data set was used.

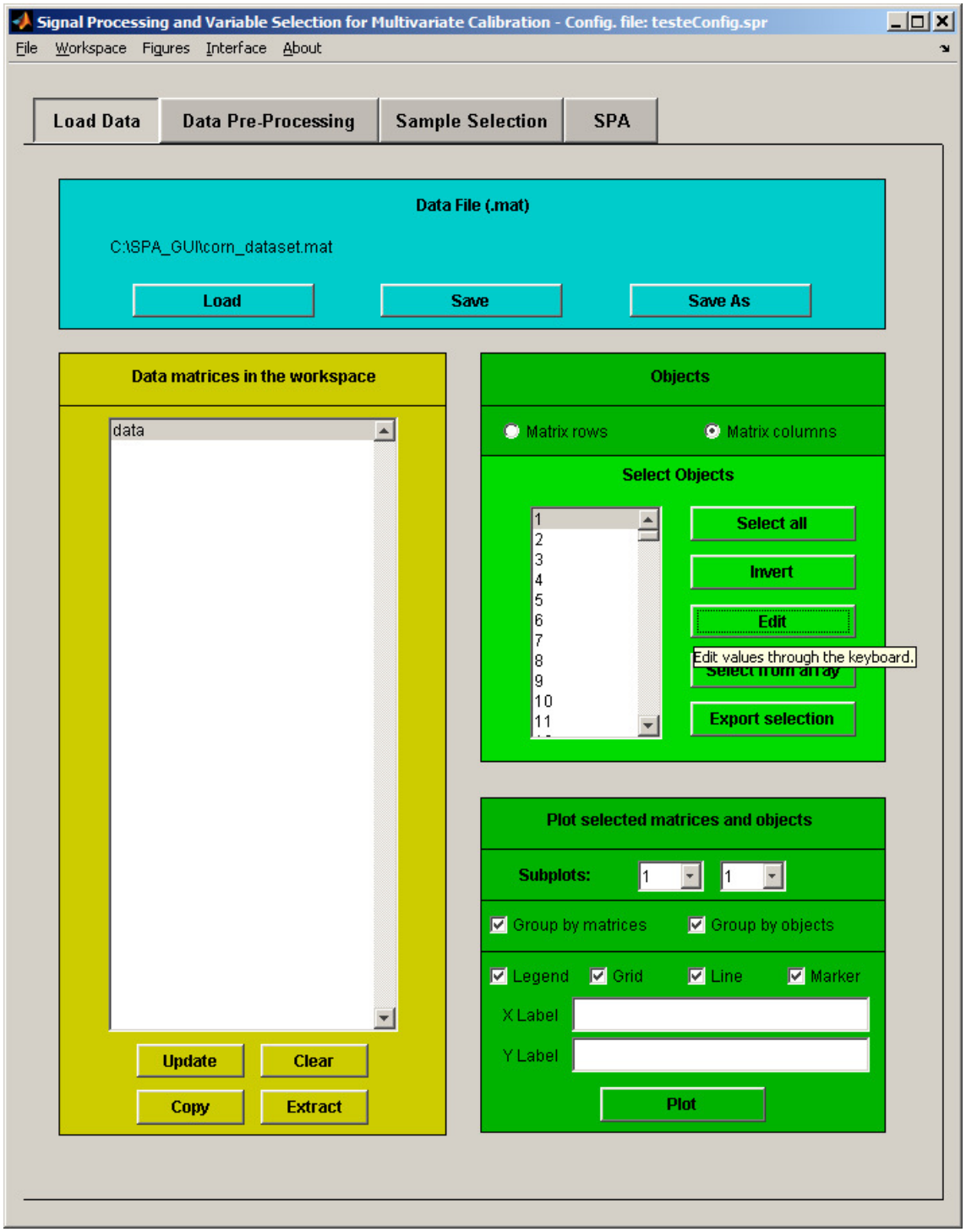
In this interface the data set should not be organized in the matrix format. Therefore, the data were organized in a matrix called “data”. The last column of “data” corresponds to moisture content.

7.2. Loading the data

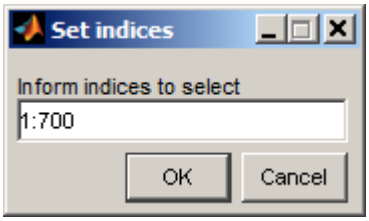
In the “Load Data” module (section 3) use the “load” button to load the contents of data file to the workspace.



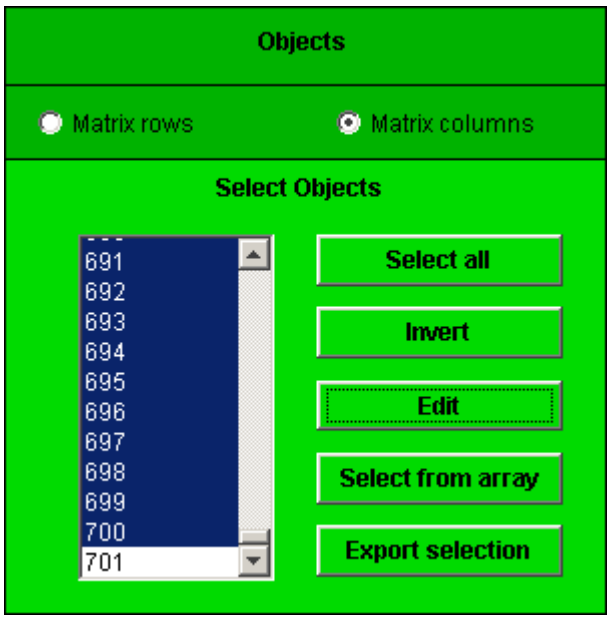
Once the data file is loaded, the matrix "data" is presented at the "Data matrices in the workspace" group.



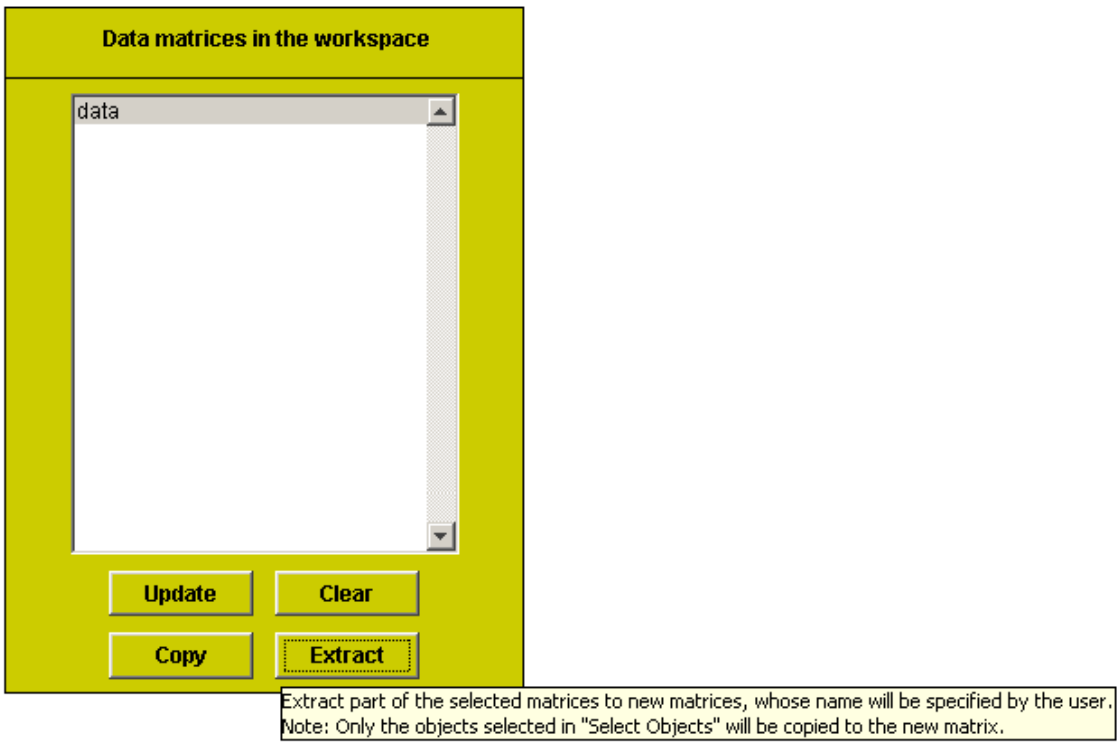
The next step is to split the data matrix will in the matrices X (instrumental responses) and Y (parameter of interesting). In order to do that, first select the “Matrix Columns” checkbox and press the “Edit” button in the “Objects” group. The following window will appear. In this window, select the columns with index of 1 to 700 and press the OK button.



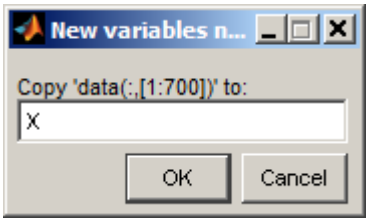
The objects from 1 to 700 will be selected, as illustrated below.



Then, press the “Extract” button in the “Data matrices in the workspace” group.



The following window will appear. In this window, inform the name of the new matrix as “X” and press the OK button.



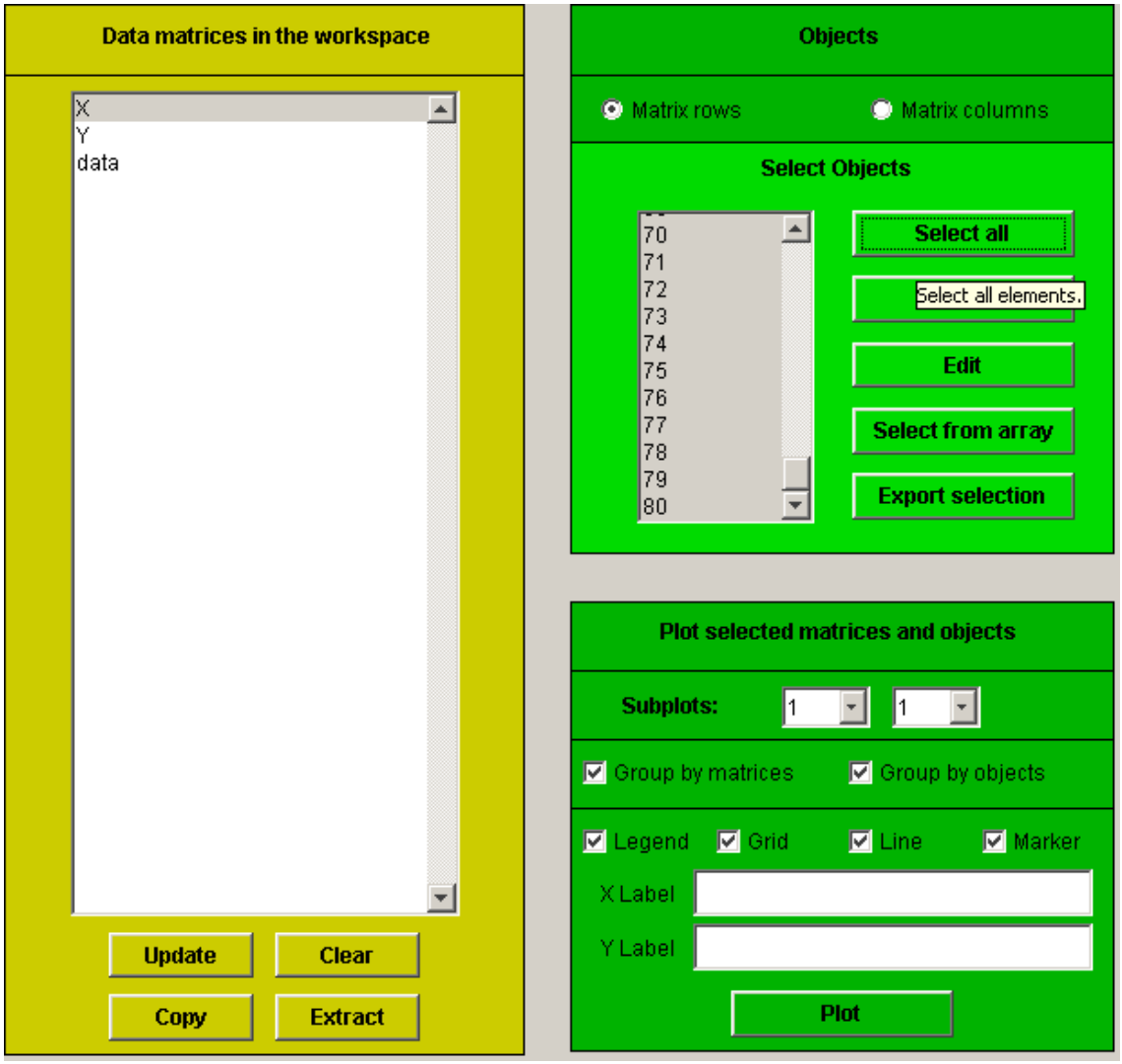
To extract the data to Y, repeat the process selecting only the column 701.

If you desire to plot the spectra of samples, select the “X” matrix at the “Data matrices in the workspace” group and choose the indices of the samples to be plotted in the “Objects” group. Inform the desired plot parameters at the “Plot selected matrices and objects” group and then press the “Plot” button.

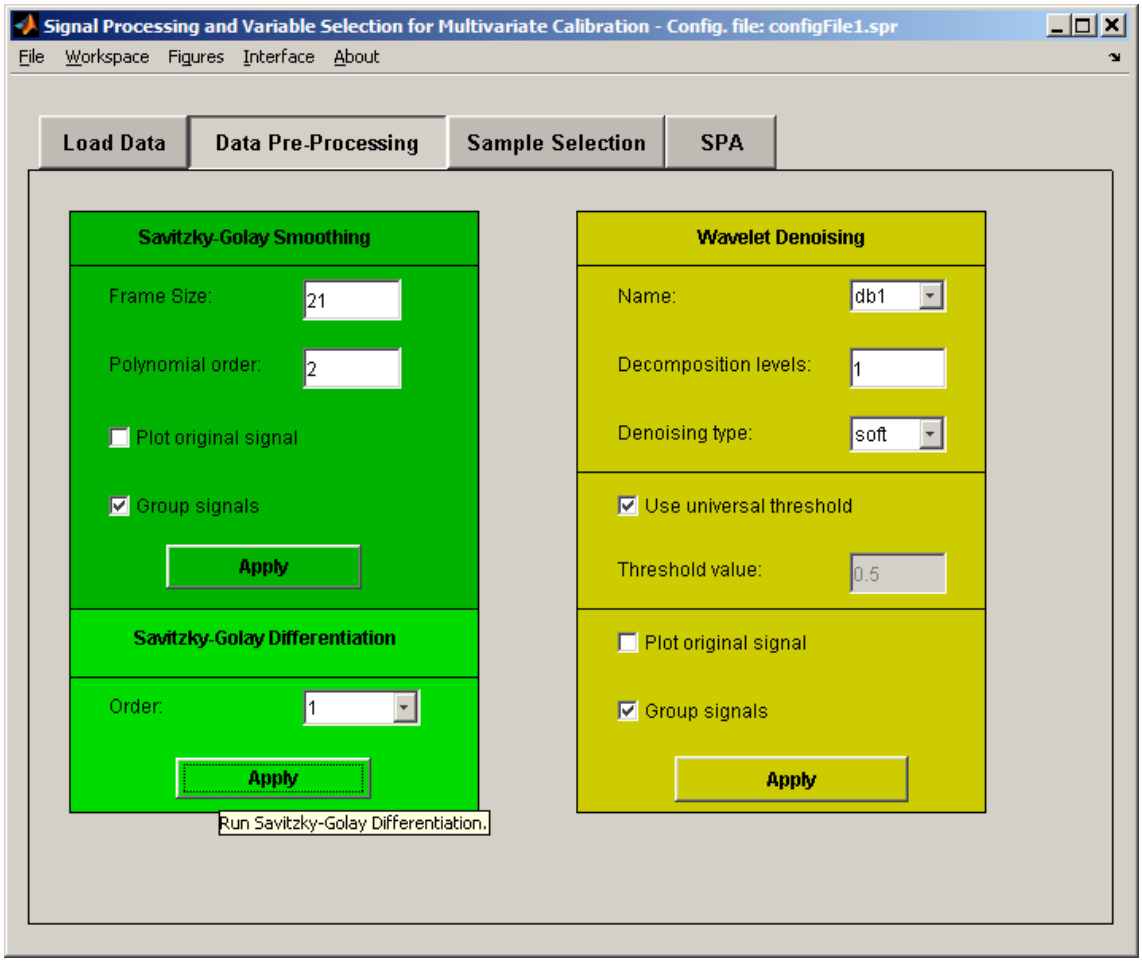
7.3. Preprocessing

In this example, derivative spectra smoothed by a Savitzky–Golay filter are employed, in order to circumvent the problem of systematic baseline variations. A second-order polynomial and a 21-point window are used.

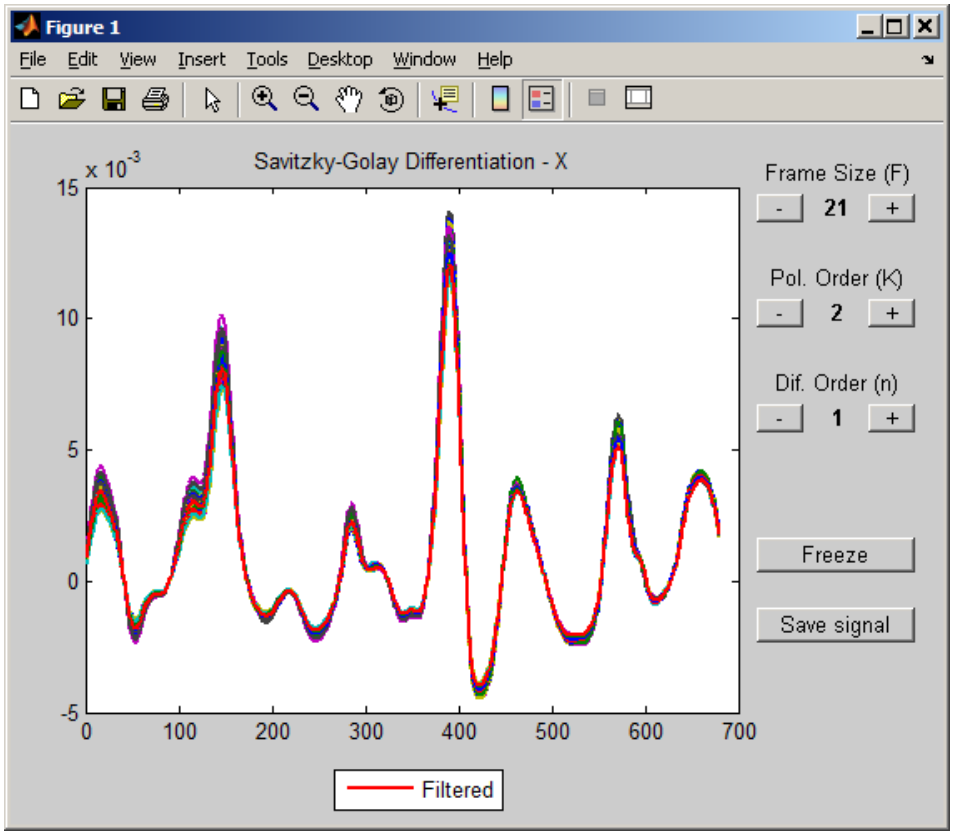
To perform this preprocessing, select the matrix “X” in the “Data matrices in the workspace” group. Then, in the “Objects” group, select the “Matrix rows” option and press the button “Select All” to select all samples.



Switch to the “Data Pre-Processing” module. In this module, inform the Savitzky-Golay filter parameters (the frame size, polynomial order, and differentiation order), according to the figure below.



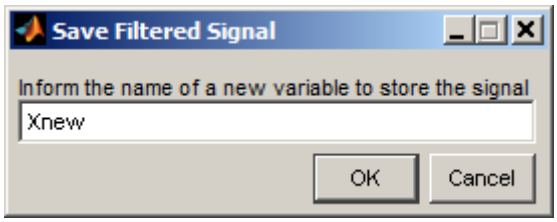
Then, press the “Apply” button in the “Savitzky-Golay Differentiation” group. The Savitzky-Golay screen will appear.



If want to test other preprocessing configurations, press the "+" and "-" buttons to change the frame size and polynomial order.

Note: When testing other configurations using the "+" and "-" buttons, the computational effort may be too high, if there are a high number of samples. In this case, it is recommended to first run the pre-processing using a subset of the samples instead of all samples. After finishing the tests and selecting the best parameters, the Savitzky-Golay processing can be run again with all samples.

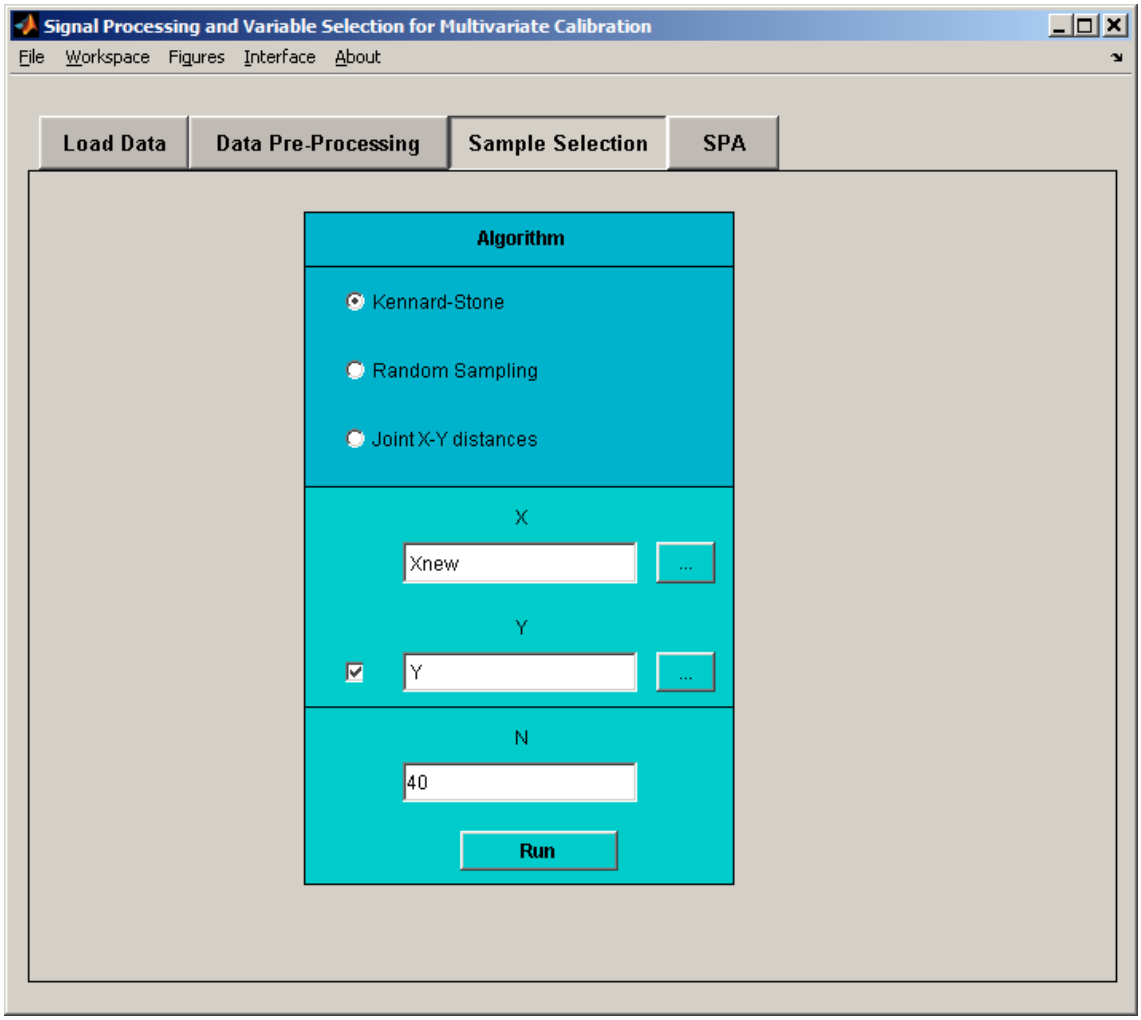
Press the "Save signal" button in the Savitzky-Golay screen to save the processed samples. A window requesting the name of the matrix will appear. In this window, inform the name of the matrix as "Xnew".



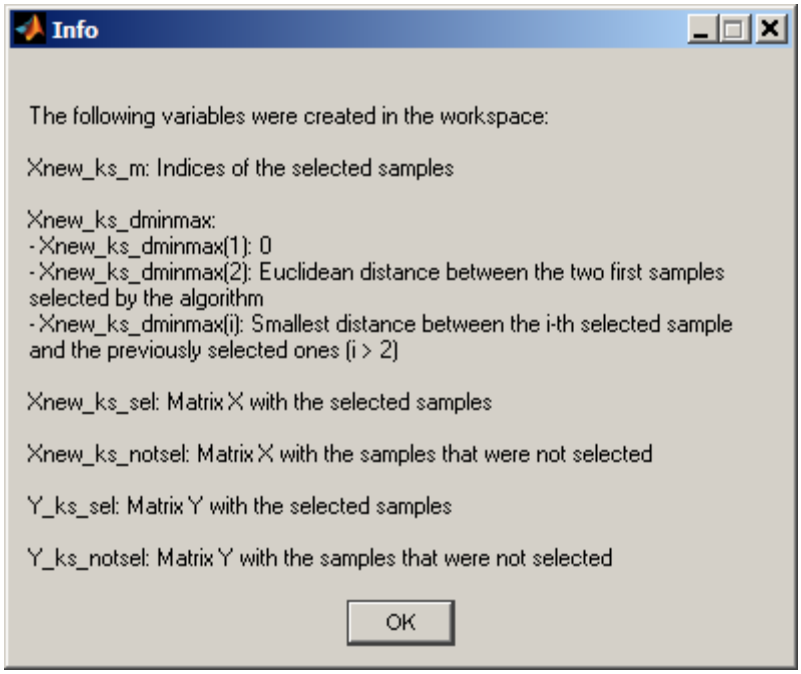
7.4. Selection of samples

The KS algorithm is used to divide the available samples into calibration, validation, and prediction sets. The corn data were divided into 40 samples for calibration, 20 samples for validation, and 20samples for prediction. These sets are used for model-building and performance evaluation.

To select the 40 samples for calibration, switch to the “Sample Selection” module and set the parameters as in the following figure.

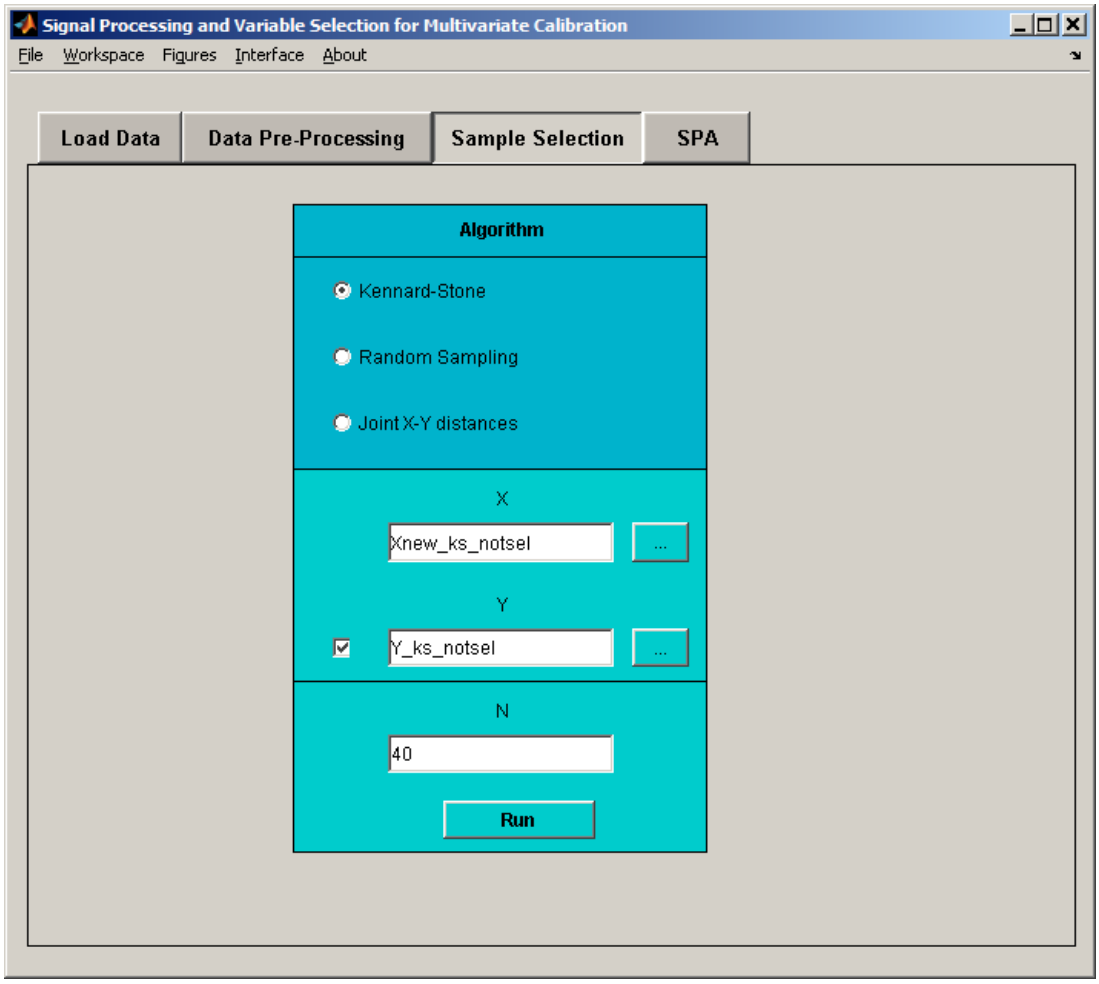


After pressing the “run” button, the following window is shown, informing the name of the new matrices that were created in the workspace.

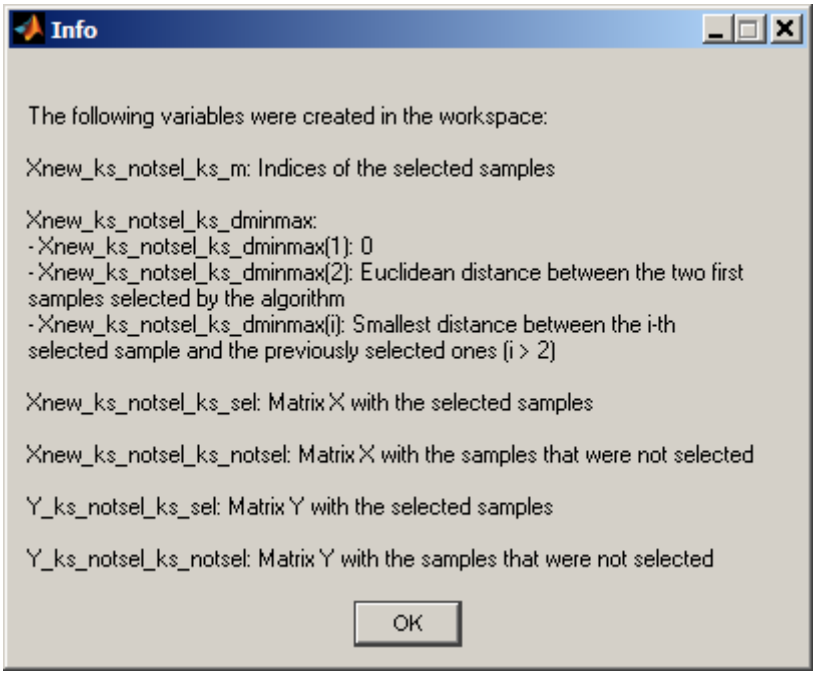


The “Xnew_ks_sel” matrix, which contains the selected samples, is the set to be used for calibration. The “Xnew_ks_notsel’ matrix, which contains the samples that were not selected, will be divided in two sets, for validation and prediction.

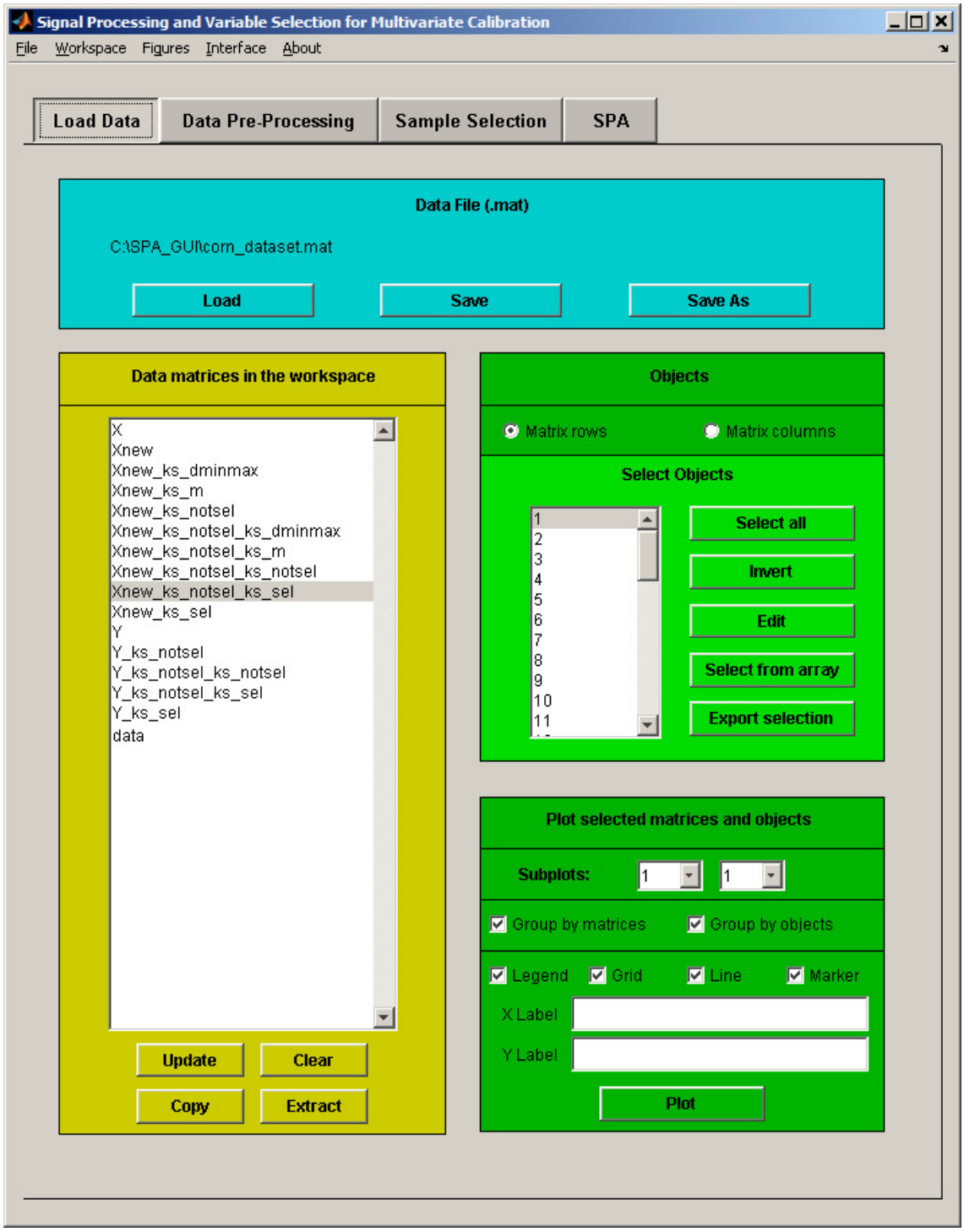
Before dividing the “Xnew_ks_notsel” matrix, it will first be ordered according with Euclidian distances using the KS algorithm. This procedure can be performed by setting the parameters as in the following figure.



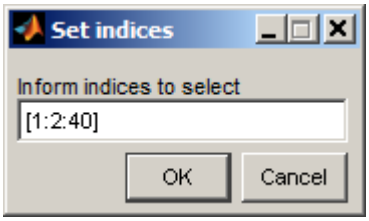
After pressing the "run" button, the following window is shown, informing the name of the new matrices that were created in the workspace.



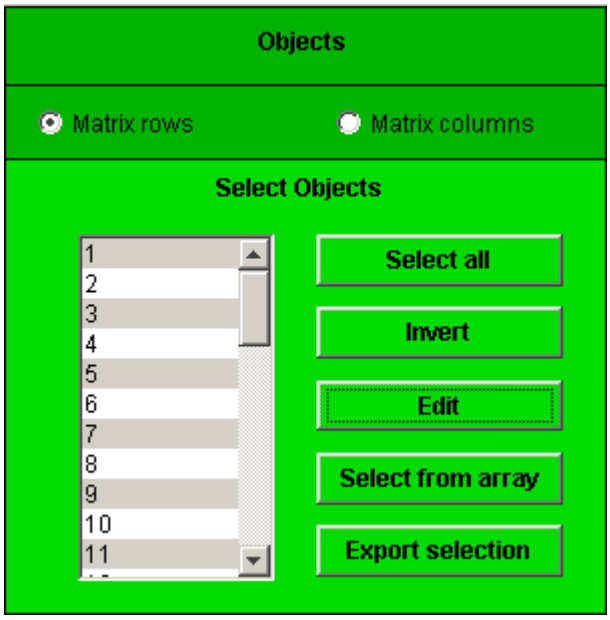
The new matrix “Xnew_ks_notsel_ks_sel” contains the same samples of “Xnew_ks_notsel”, but ordered by distance. In order to split this matrix in the validation and prediction sets, switch to the “Load Data” module and select the “Xnew_ks_notsel_ks_sel” matrix in the “Data matrices in the workspace” group.



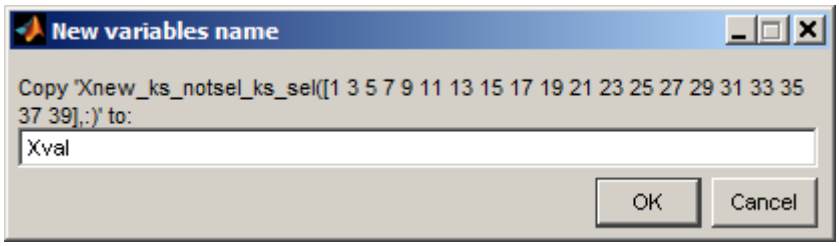
Then, press the “Edit” button in the “Objects” group and select the samples 1:2:40, as illustrated in the figure below.



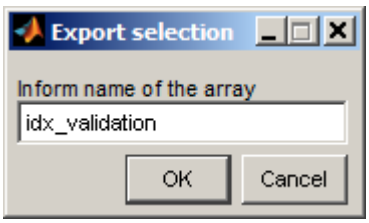
The result will be:



In order to extract the selected samples, press the “Extract” button in the “Data matrices in the workspace” group. The following window will appear. In this window, inform the name of the new matrix as “Xval” and press the OK button.

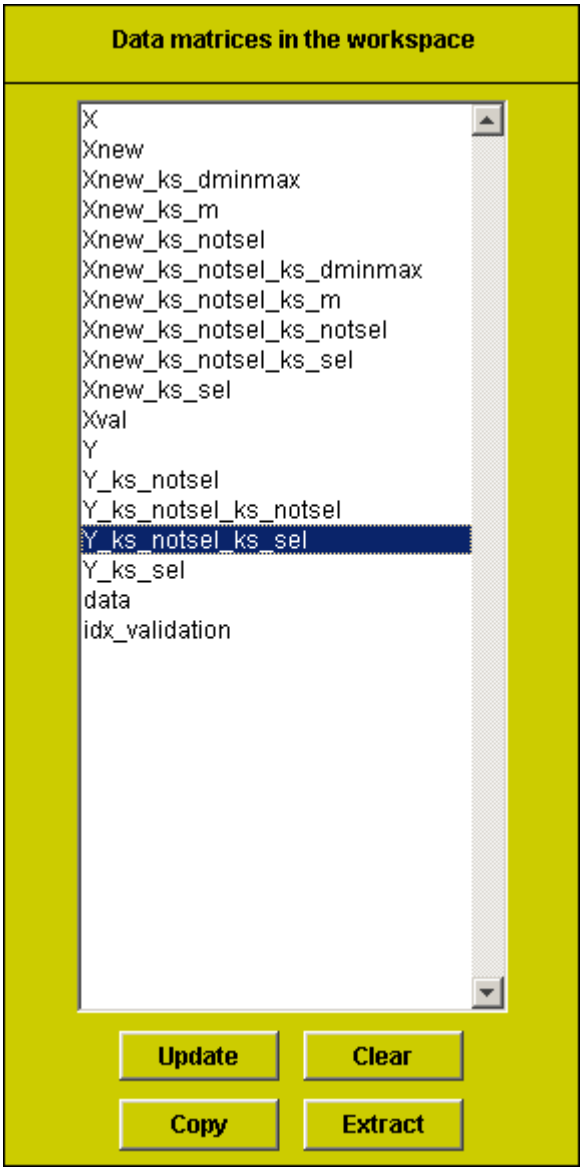


After that, press the “Export Selection” button in the “Objects” group to export the selected indices (for future use). The following window will appear. Inform the name of the array as “idx_validation”.

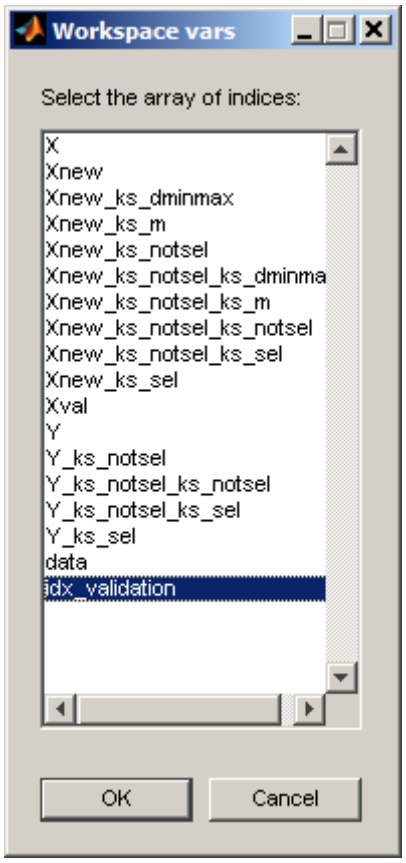


The “idx_validation” array containing the selected indices will be created in the workspace.

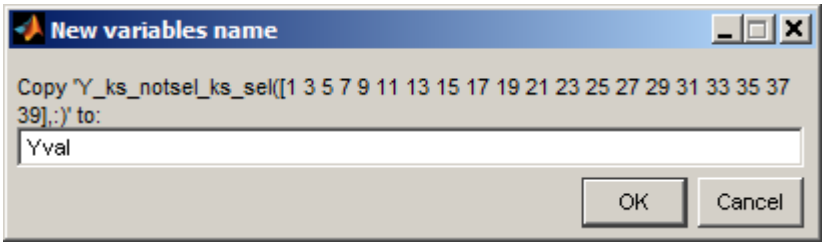
Now, select the “Y_ks_notsel_ks_sel” matrix in the “Data matrices in the workspace” group to extract the moisture content for the validation samples.



The same indices used for matrix Xval must be used for matrix Yval. These indices can be informed by using the same procedure used above. An alternative procedure is to press the “Select from array” button in the “Objects” group. After pressing this button, a list of numeric matrices is presented. In this list, select the “idx_validation” array that was saved before. This will ensure that the same indices are used for Xval and Yval matrix.

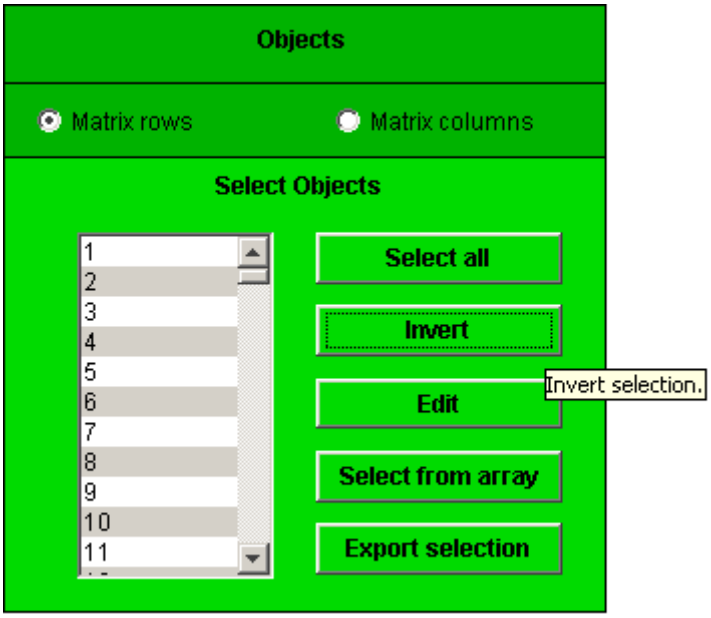


Now, set the matrix Y for validation using the "Extract" button in the "Data matrices in the workspace" group.

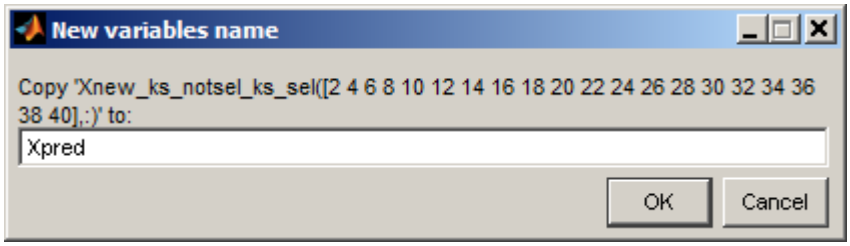


To choose the samples of prediction, select the matrix Xnew_ks_notsel_ks_sel again in the "Data matrices in the workspace" group. Use the "Select from array" button to load again the list of indices available in the "idx_validation" array.

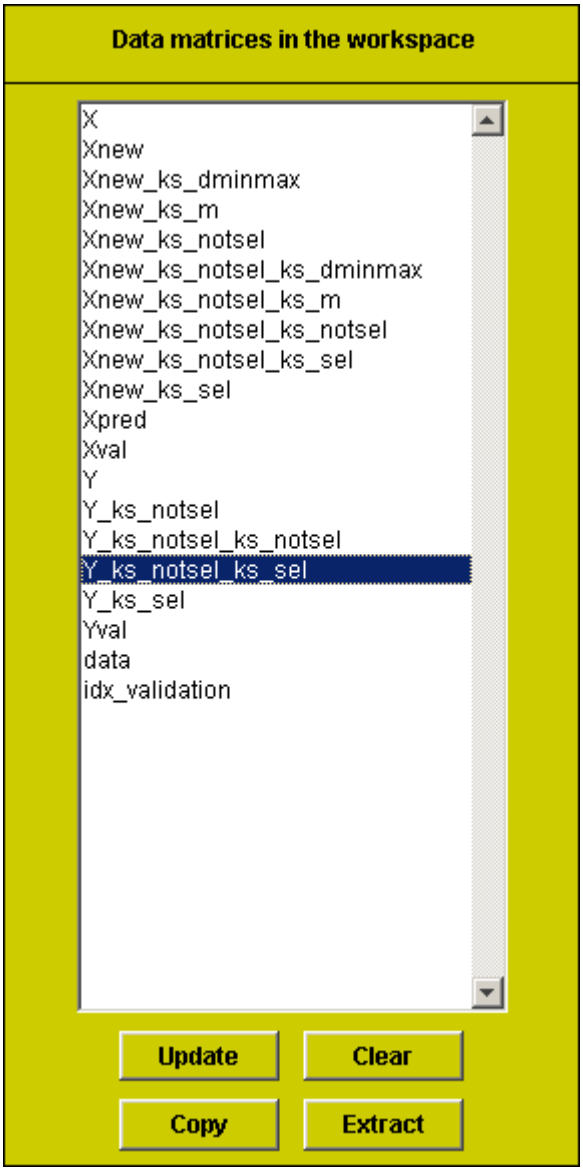
Press the "Invert" button in the "Objects" group to invert the selection, i.e., to select the remaining samples.



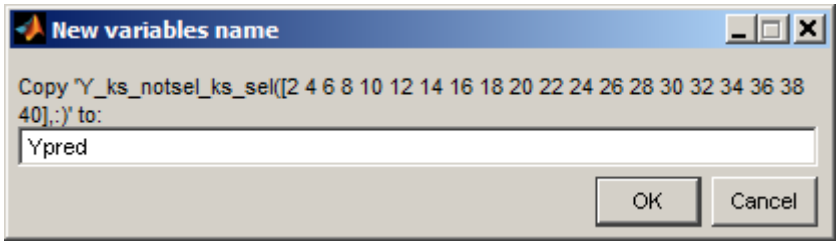
Now, set the matrix X for prediction using the “Extract” button in the “Data matrices in the workspace” group.



Select the matrix Y_ks_notsel_ks_sel in the “Data matrices in the workspace” group to extract the moisture content for the prediction samples.



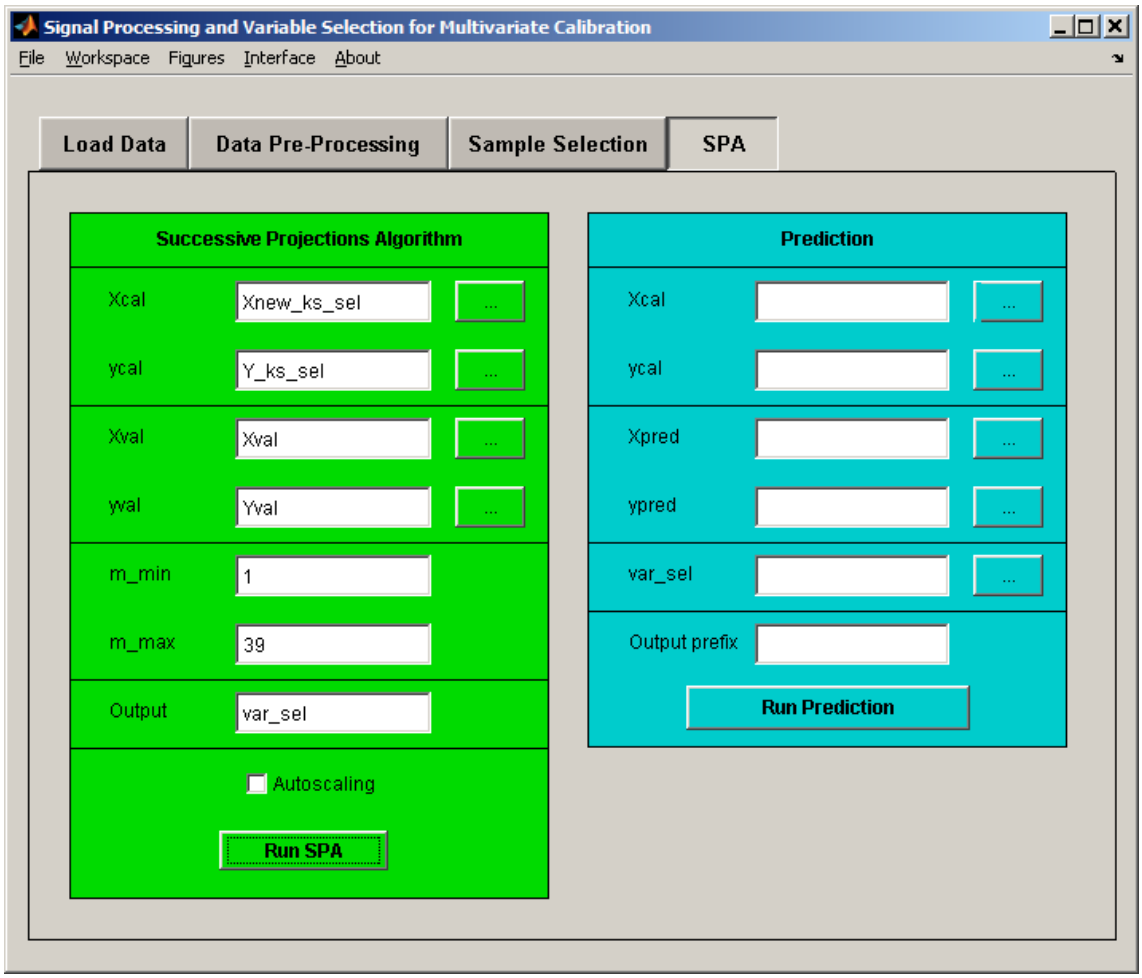
Now, set the matrix Y for prediction using the “Extract” button in the “Data matrices in the workspace” group.



7.5. Variable selection for multivariate calibration using SPA

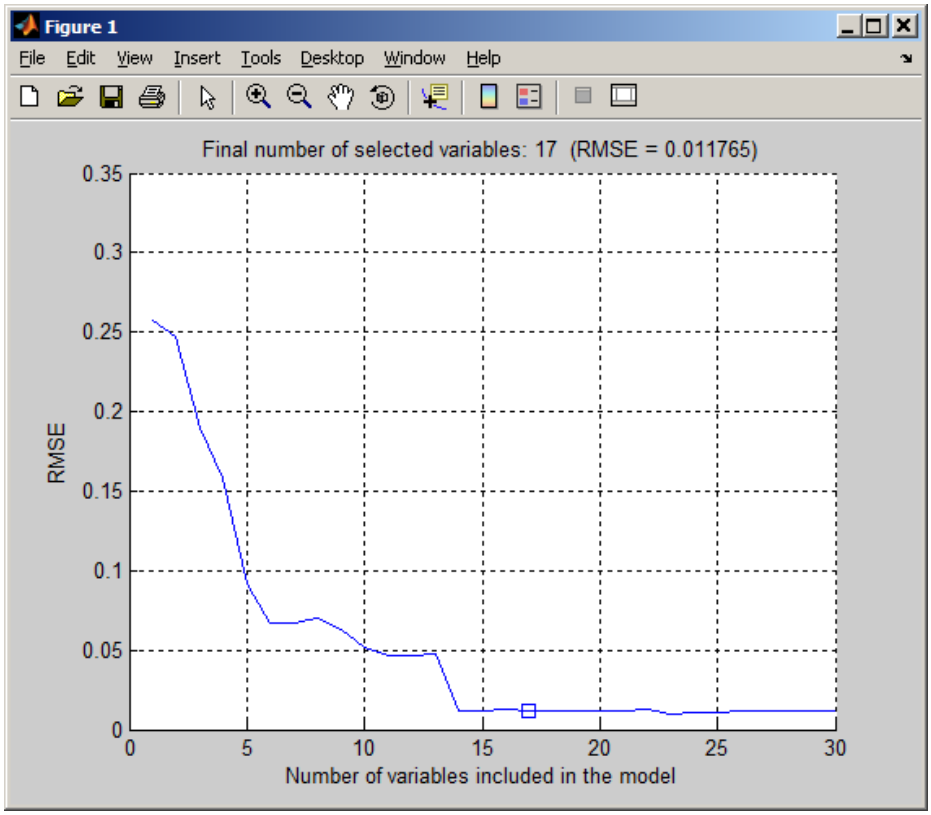
7.5.1. Building the model

To use the SPA algorithm, set the calibration and validation matrices, minimum and maximum number of variables as described in the section 6.1. In this example, specify the parameters as in the following figure and press the “Run SPA” button.

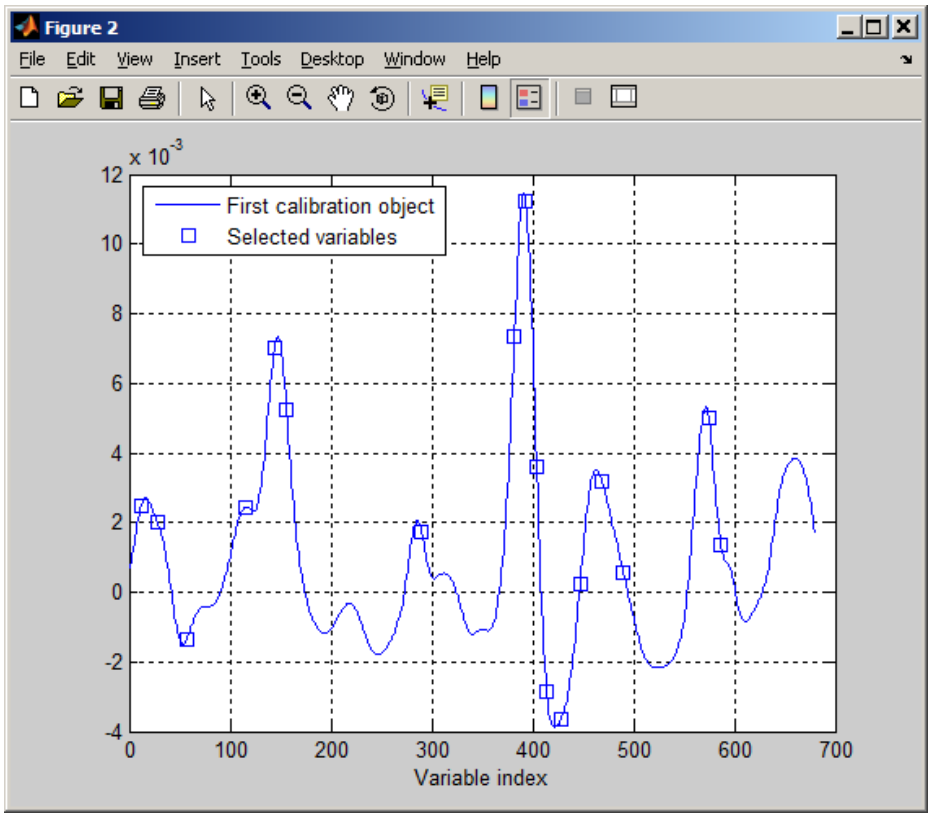


Two figures are presented: scree plot and the variables selected.

The scree plot tends to level off after a certain number of variables is added to the model. The number of variables selected in the third phase of SPA is indicated by square marker. This is the point at which the RMSE is not significantly larger than $RMSE_{min}$ according to an F-test with $\alpha = 0.25$.

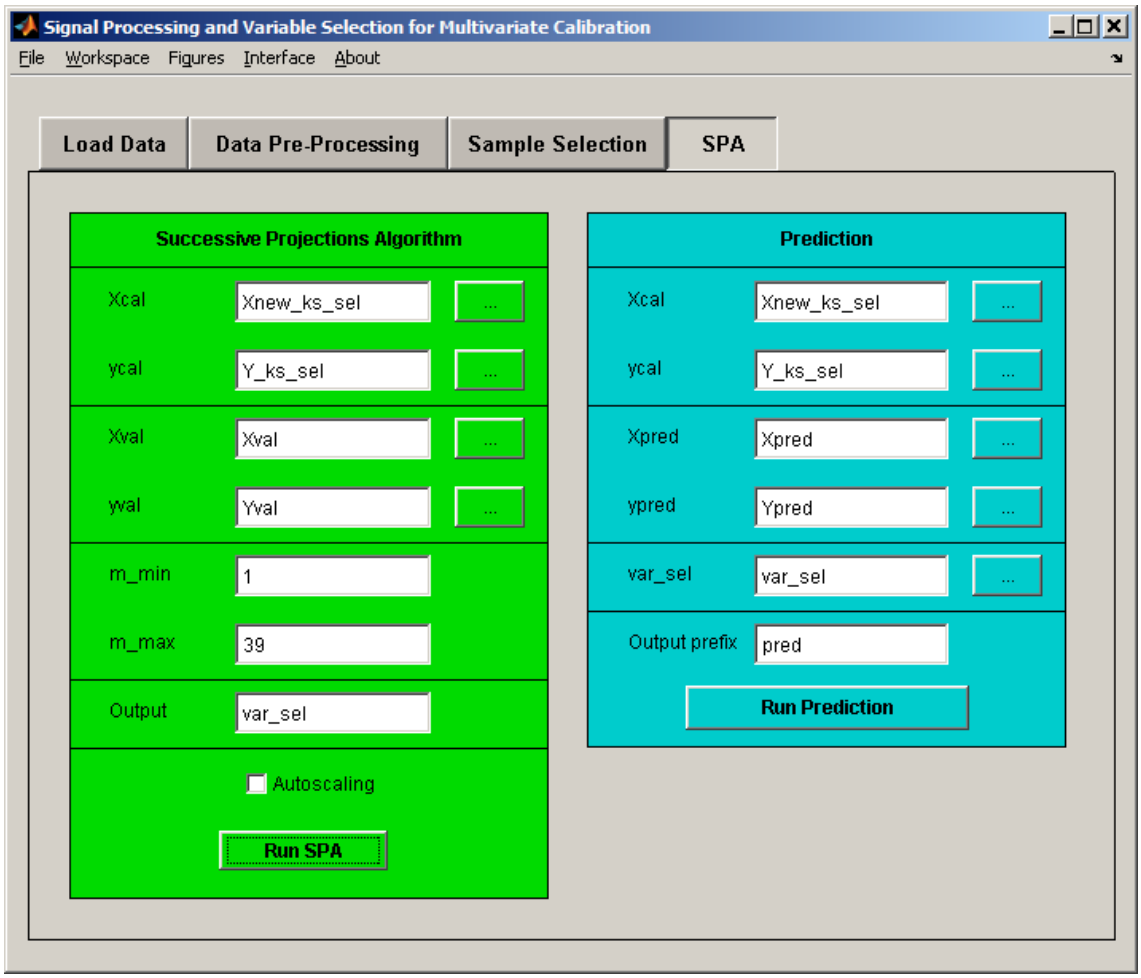


The variables selected by SPA are plotted at the first calibration samples. This figure is presented below.

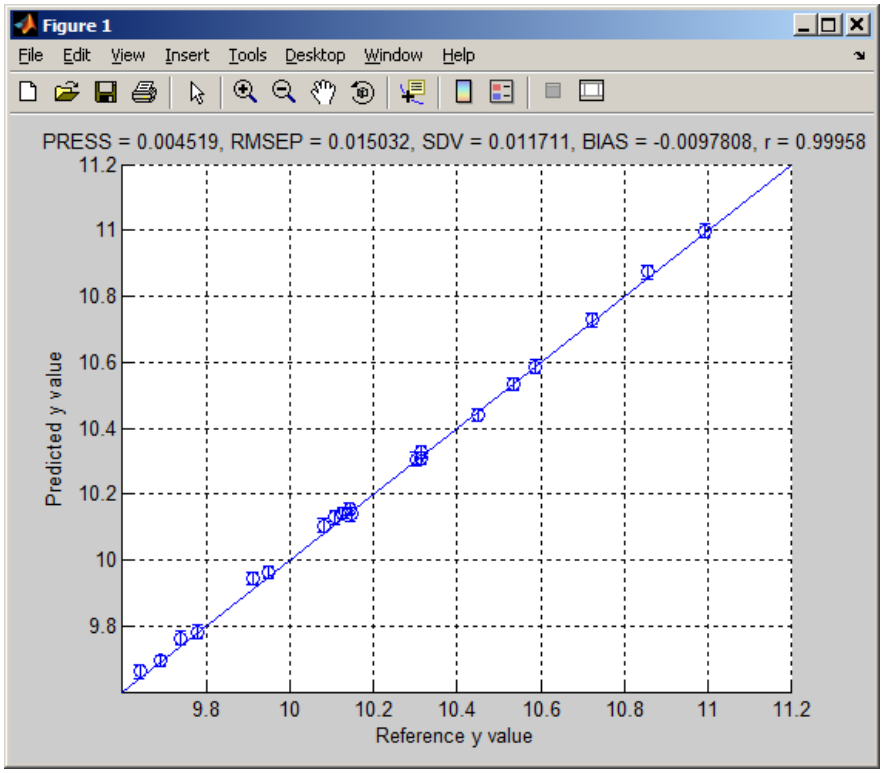


7.5.2. Prediction of samples

In this example, specify the “Prediction” group parameters as in the following figure.

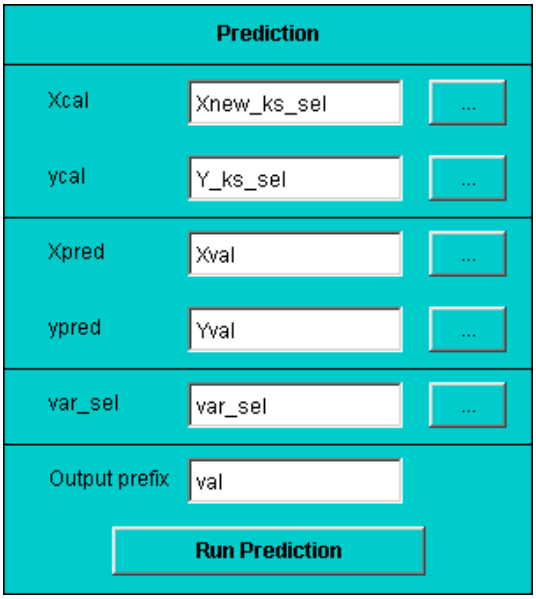


The graph reference versus predicted is presented together with the statistics parameters PRESS, RMSEP, SDV, and r, for the prediction set. The figure below shows the obtained results.

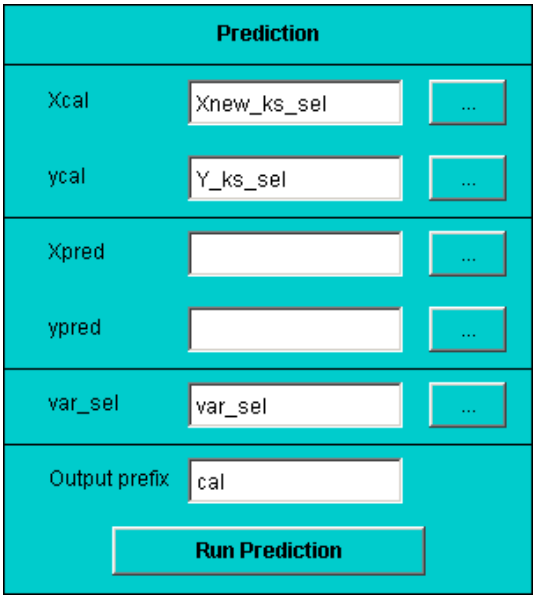


To know the statistics parameter of the validation set, use the validation matrices in the spaces of prediction (Xpred and ypred). If they are left blank, leave-one-out cross-validation will be carried out in the calibration samples.

Prediction for the validation set

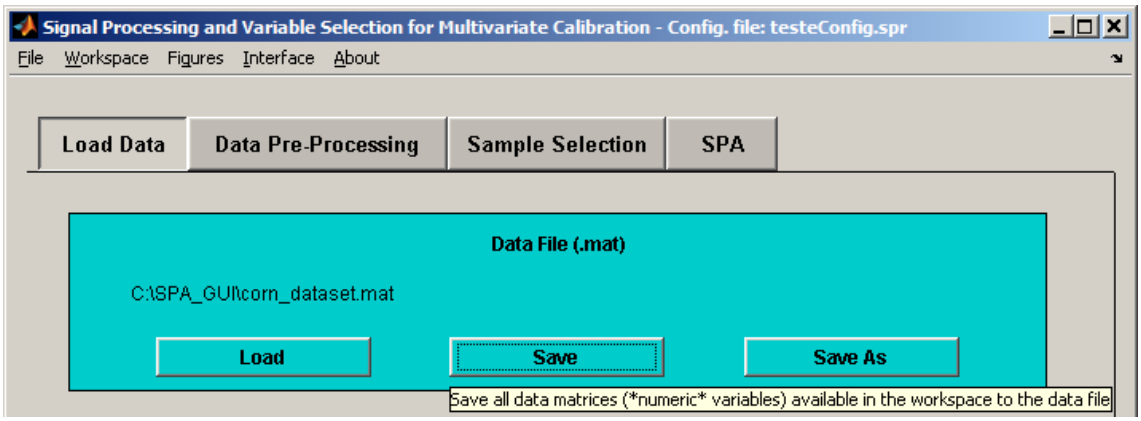
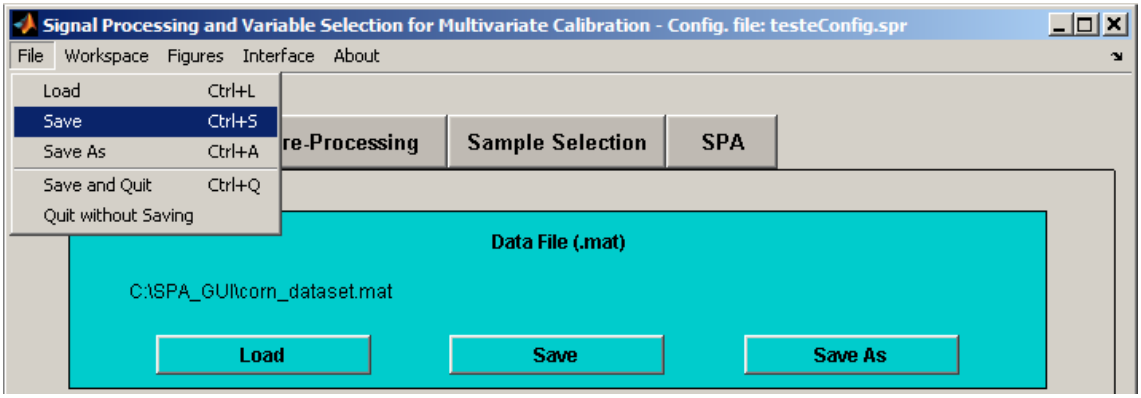


Prediction for the calibration set



7.6. Saving your work

Switch to the “Load Data” module and press either the “Save” or the “Save As” button in the “Data File” group to save the data matrices to the data file (file with .mat extension).



In the main menu, choose the option “File: Save” or “File: Save As” to save the configuration file (file with .spr extension used to store the parameters specified in the graphical user interface).