Subject

Detecting the right number of factors for a PLS regression using a resampling approach.

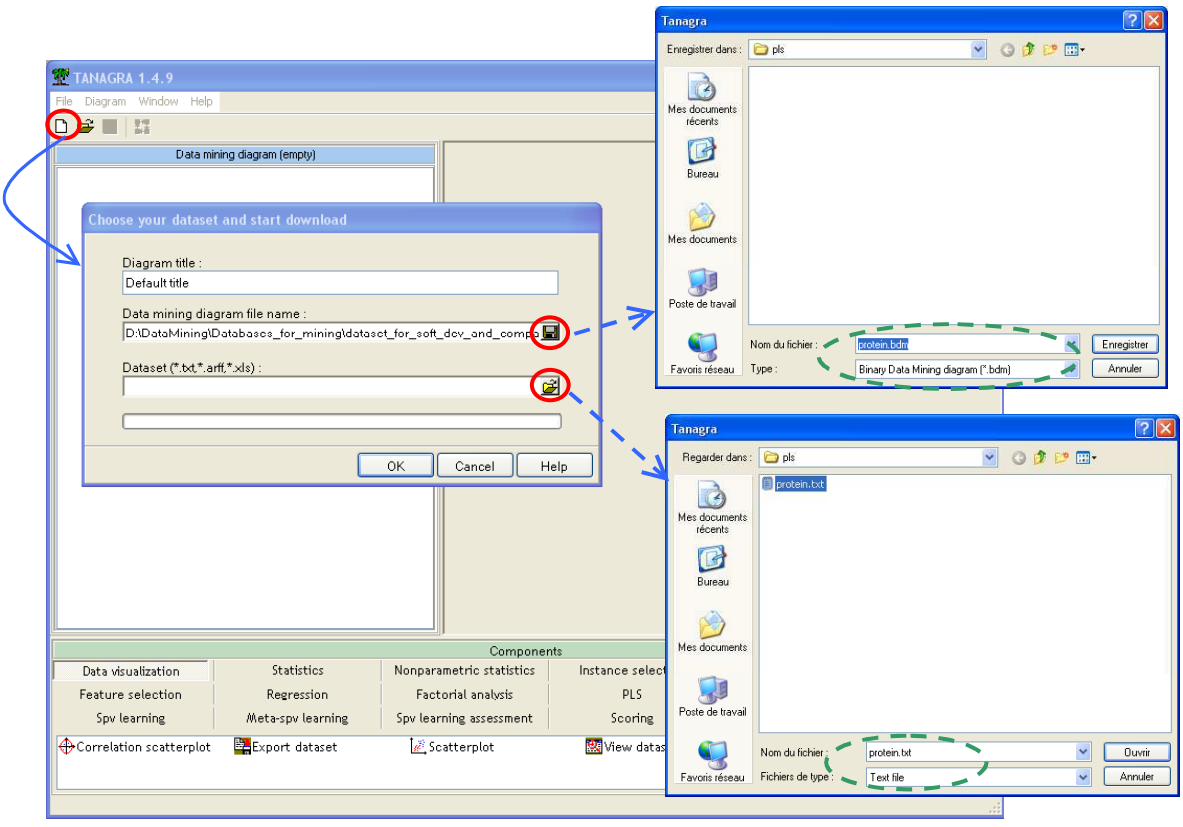
Dataset

In a protein classification task, the descriptors are 3-grams (**7143 attributes**), the dataset contains **101 examples**. We have a binary class attribute; it corresponds to the family membership of the protein sequences.

PLS regression

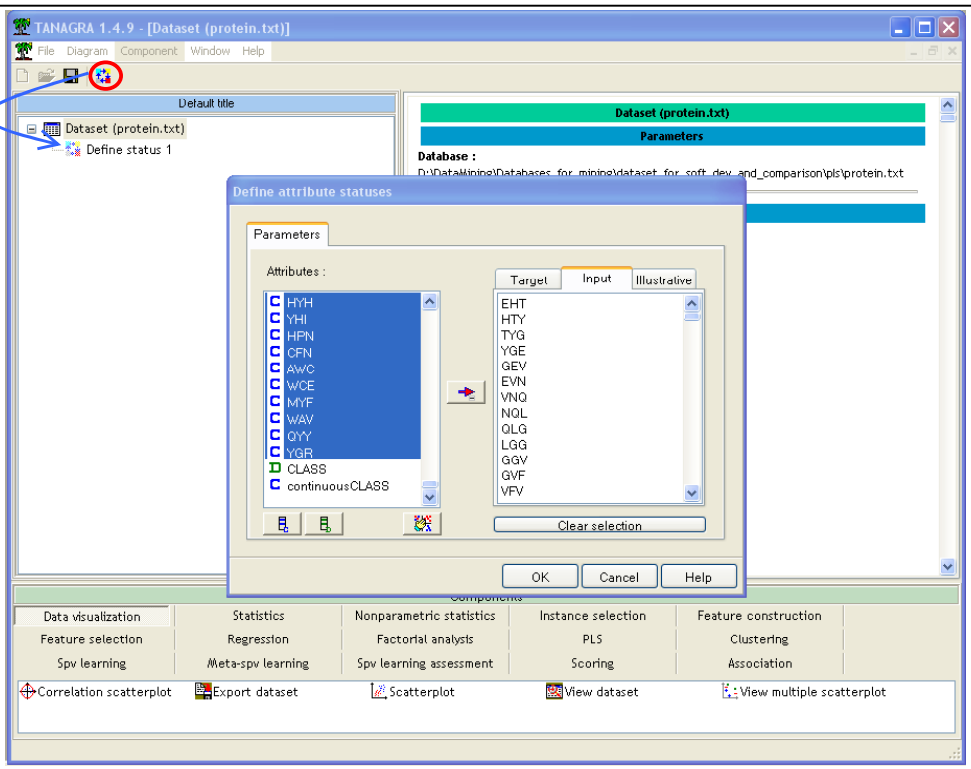
Create a diagram and download the dataset

We create a new diagram (FILE/NEW) and import the dataset. Because there are many attributes, it is more suitable to save the diagram in a binary file format (BDM format of TANAGRA).



The PLS-FACTORIAL component

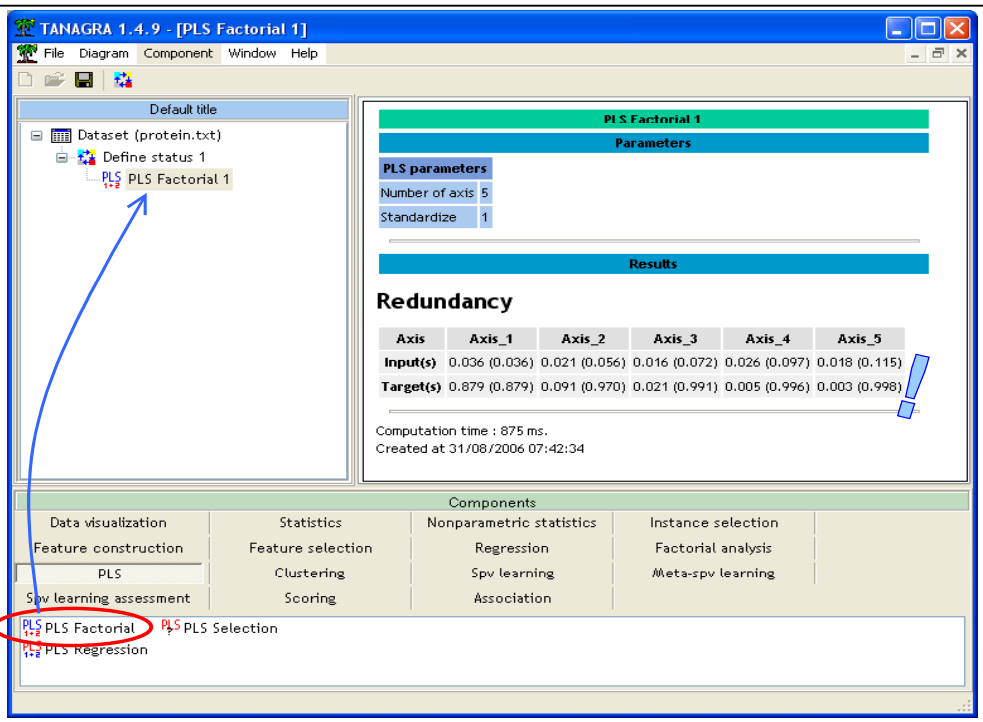
We use the DEFINE STATUS component in order to choose the TARGET attribute (CONTINUOUS_CLASS) and the INPUT attributes (all the others continuous attributes).



Then we insert into the diagram the PLS-FACTORIAL component. The difference of this component in relation to the PLS-REGRESSION component is that it produces new variables that correspond to the factors. The PLS Regression component produces the prediction of the regression.

In our example, we have one target attribute. In the general case, we can select several TARGET attributes.

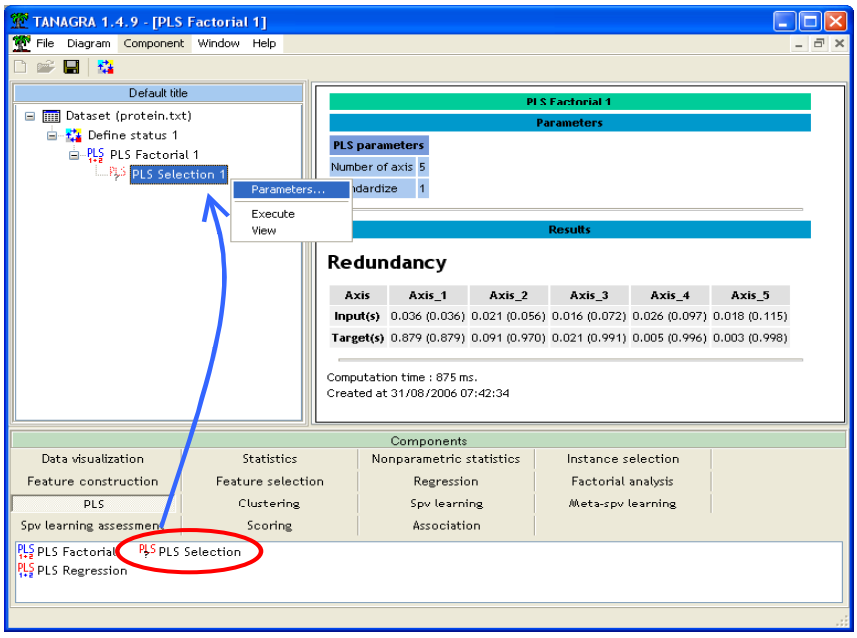
By default, the PLS FACTORIAL computes 5 factors. We can modify this parameter. But we can also detect the "optimal" number of factors with the PLS-SELECTION component.



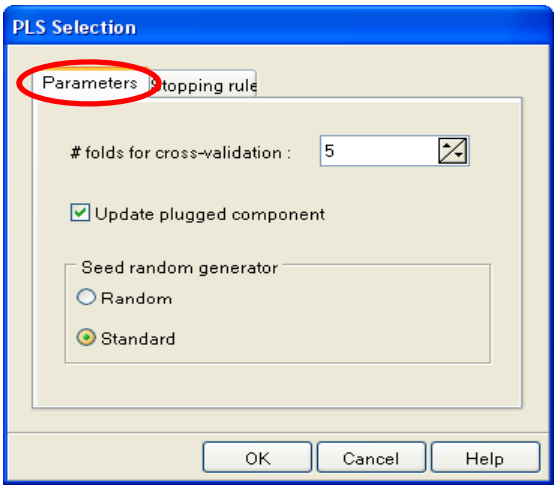
The results show that 99,8% of the variance of the target attribute is explained with 5 factors. But, from the third factor, the increasing of explained variance is unimpressive. In the following section, we use a cross-validation technique in order to detect the relevant factors.

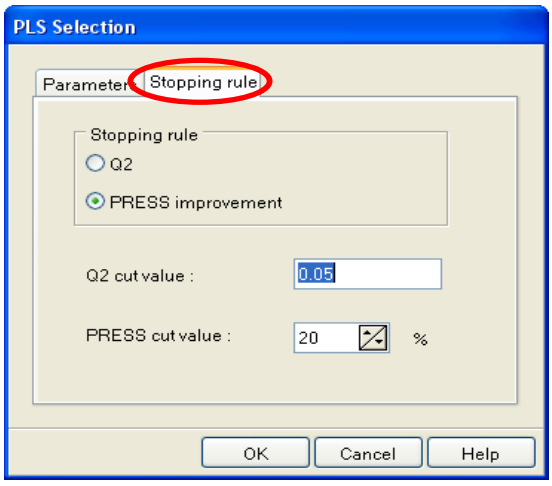
Detecting the "optimal" number of factors

The PLS-SELECTION component is intended to determine automatically the relevant factors of PLS regression. We must insert it under a PLS component (PLS-FACTORIAL or PLS-REGRESSION). In our case, we insert the PLS-SELECTION under PLS FACTORIAL 1 into our diagram. We click on the PARAMTER contextual menu.



There are two tabs in the dialog box.





The first one (PARAMETERS) allows to define the computing parameters. If the UPDATE PLUGGED COMPONENT is checked, the number of factors of PLS FACTORIAL 1 is automatically modified, and the component is updated.

The second one (STOPPING RULE) allows to define the stopping rule of the exploration. There are two approaches: (1) based on the Q2 index, if it is lower that a cut value, the exploration is stopped; (2) based on the PRESS reduction, if the reduction is lower than 20%, the exploration is stopped.

With the default parameter (PRESS reduction), we obtain the following results.

PLS Selection 1

Parameters

Parameter	Value
# folds	5
Rnd	1
Stopping rule	1
Q2 cut value	0.0500
PRESS Reduction cut (%)	20
Update plugged component	1

Results

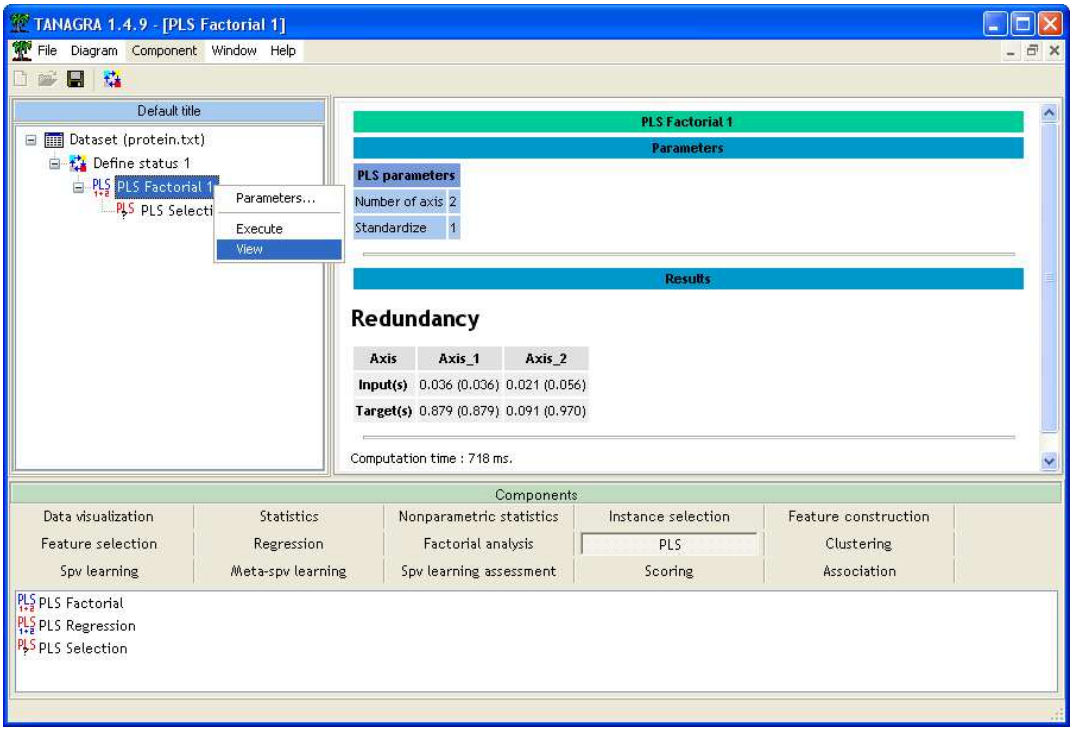
Component selection results

Number of components = 2

Detailed results

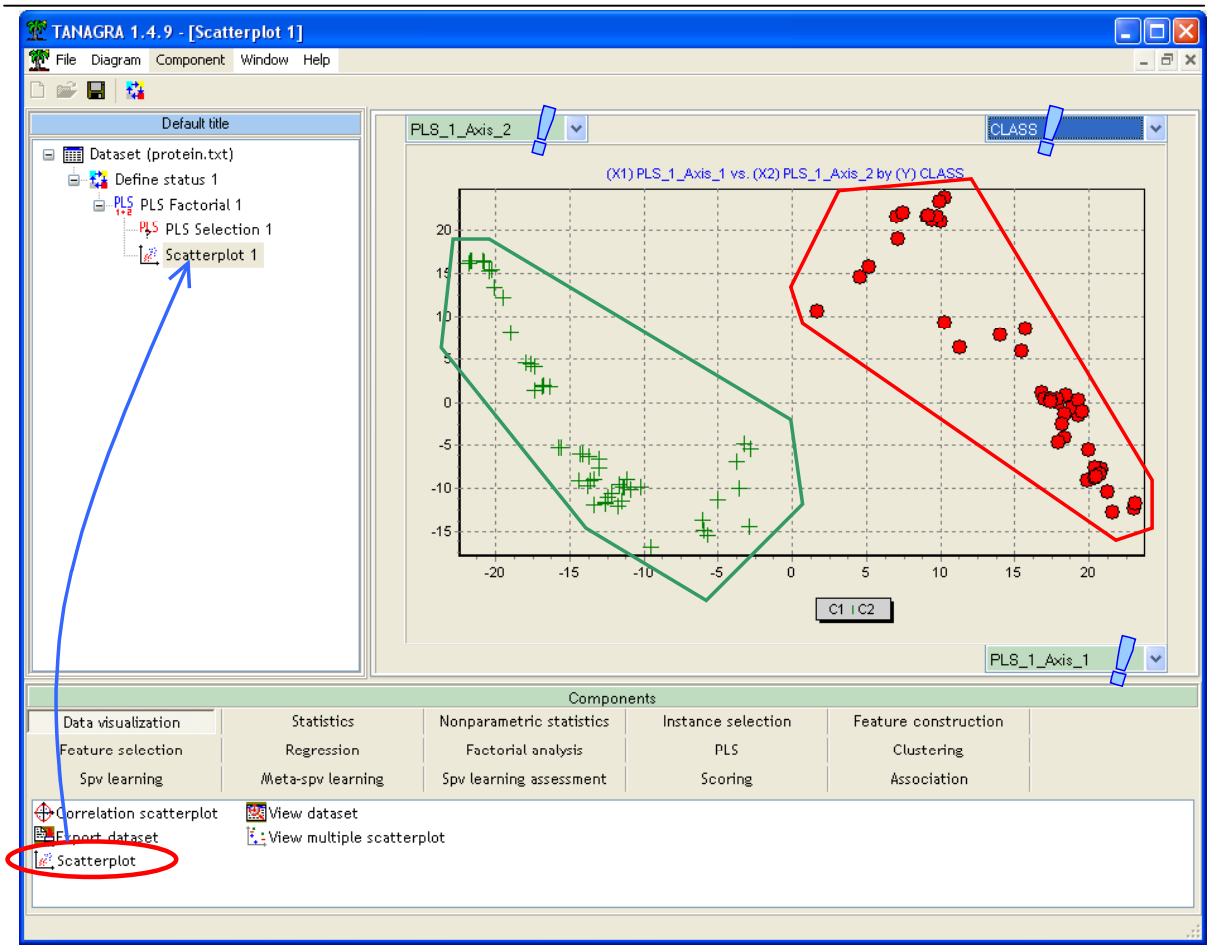
	-		continuousCLASS		
h	Q2	Q2cum	Q2	PRESS	D(PRESS)
1	0.711	0.711	0.711	7.275	71.1 %
2	-0.805	0.477	-0.805	5.481	24.7 %
3	-6.081	-2.701	-6.081	5.274	3.8 %

Two factors seem sufficient in order to predict the values of the target attribute. The PLS FACTORIAL 1 component is automatically updated. We can visualize the new results with its VIEW contextual menu.



Plot the examples

We can plot the example in a new representation space designed by the two PLS factors. We add a SCATTERPLOT component into the diagram. We color the examples with their class membership.



The result is particularly pleasant. We distinguish well the two protein families. A classification process based on these two factors will be probably very powerful.