Objectif

Choisir le nombre adéquat de facteurs dans la régression PLS.

Dans ce didacticiel, nous montrons comment mettre en œuvre le composant PLS-SELECTION pour choisir le nombre adéquat de facteurs dans la régression PLS.

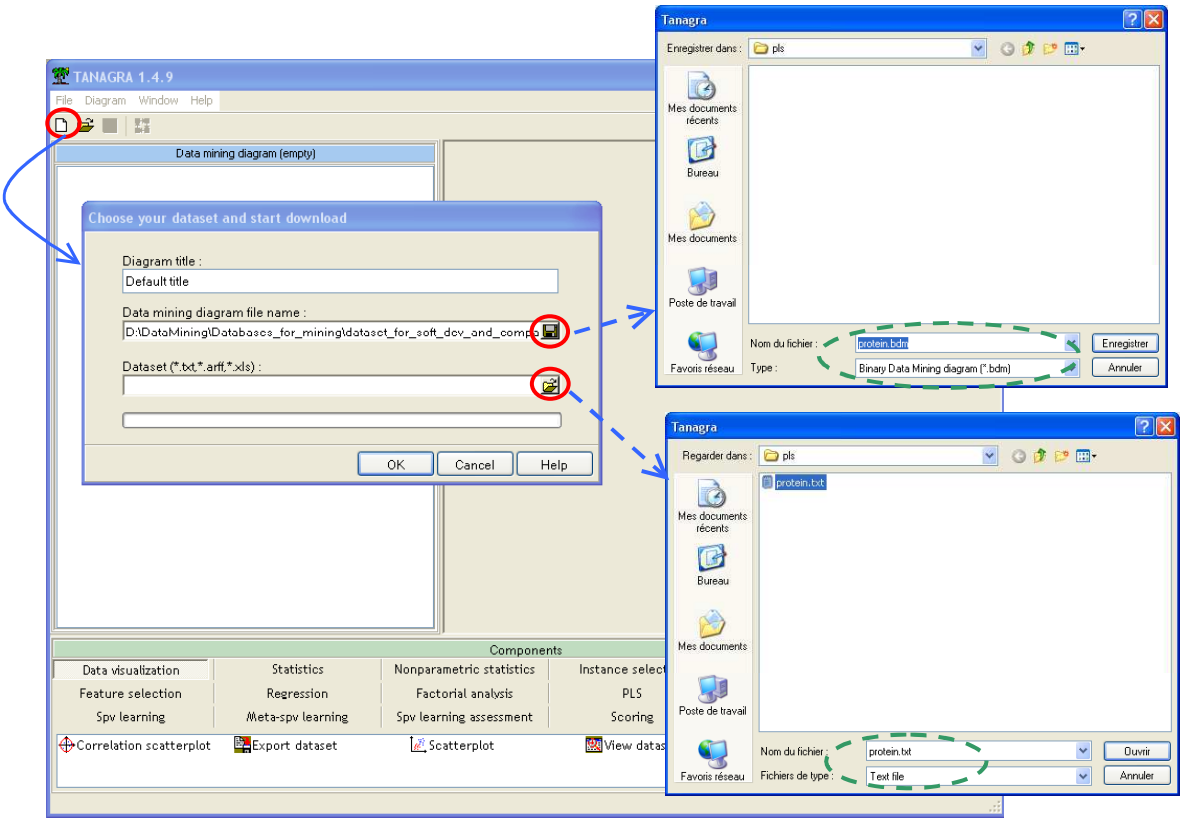
Fichier

Un fichier de séquences de protéines regroupées en deux familles. Les descripteurs sont des 3-grammes extraits à partir de la description primaire des séquences de protéines. Le fichier comporte **101 observations**, **7143 descripteurs**, deux variables indiquent la famille d'appartenance (l'une discrète, l'autre codée 0/1).

Sélection des axes dans la régression PLS

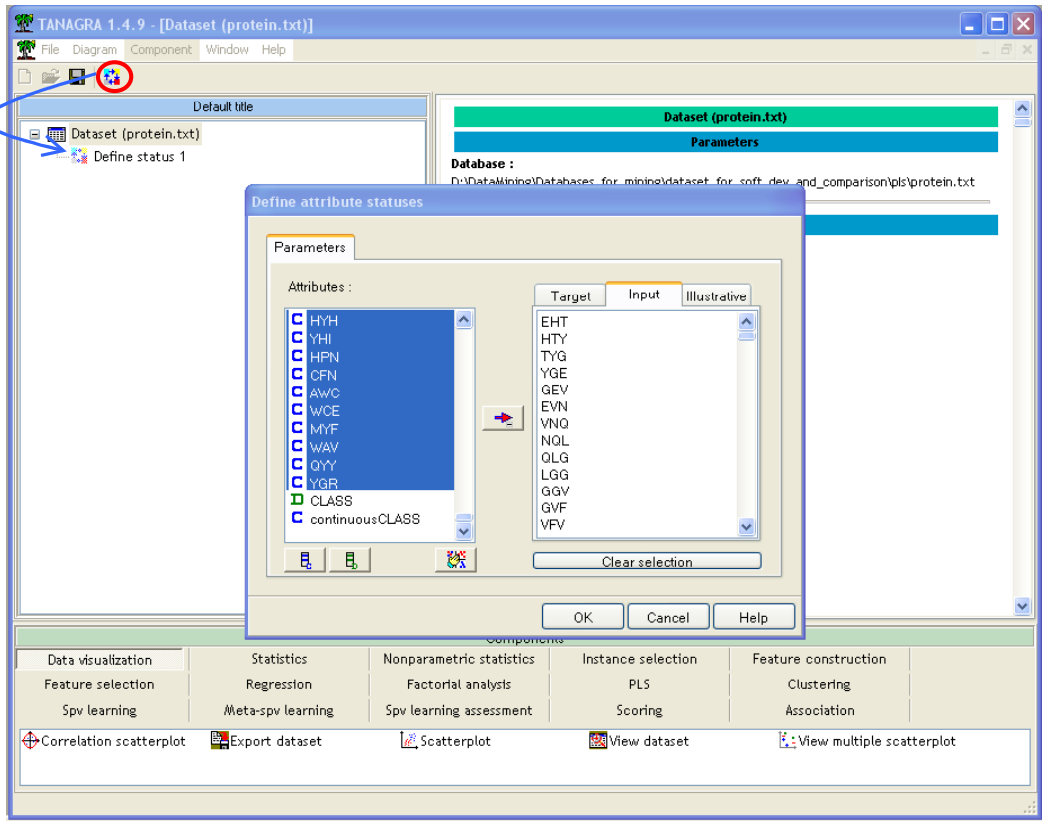
Charger les données

Pour importer les données, nous créons un nouveau diagramme (FILE/NEW). Attention, au vu des caractéristiques du fichier à traiter, le nombre de variables est vraiment très élevé, il est plus judicieux d'enregistrer le diagramme au format binaire (BDM). Les temps de traitements seront optimisés.



Définir la régression PLS avec le composant PLS-FACTORIAL

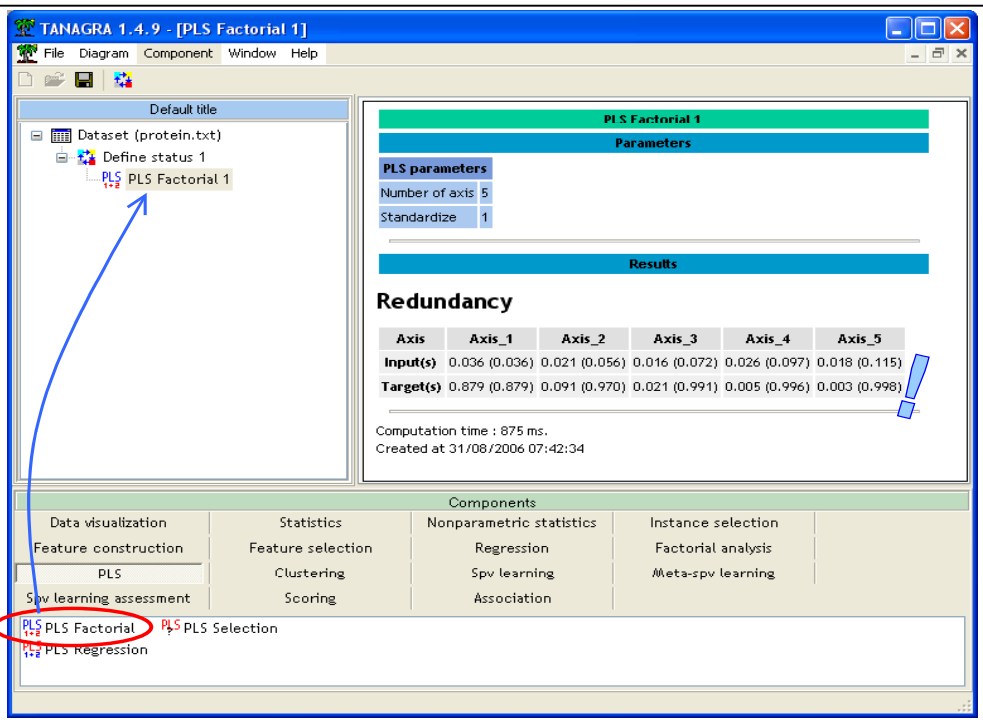
Pour définir la régression PLS, il faut tout d'abord choisir les variables de l'étude. Nous utilisons pour cela le composant DEFINE STATUS. Nous plaçons en TARGET la variable CONTINUOUS_CLASS qui indique l'appartenance aux familles à l'aide des codes 0/1, et en INPUT les variables allant de EHT jusqu'à YGR.



Puis nous insérons dans le diagramme le composant PLS-FACTORIAL (onglet PLS). Sa particularité par rapport au composant PLS-REGRESSION est qu'il crée de nouvelles variables correspondant aux projections sur les axes factoriels, et non pas la prédiction à l'aide de l'équation de régression.

Dans notre exemple, nous n'avons qu'une seule variable à prédire. Dans le cas général, il est possible de définir plusieurs variables TARGET.

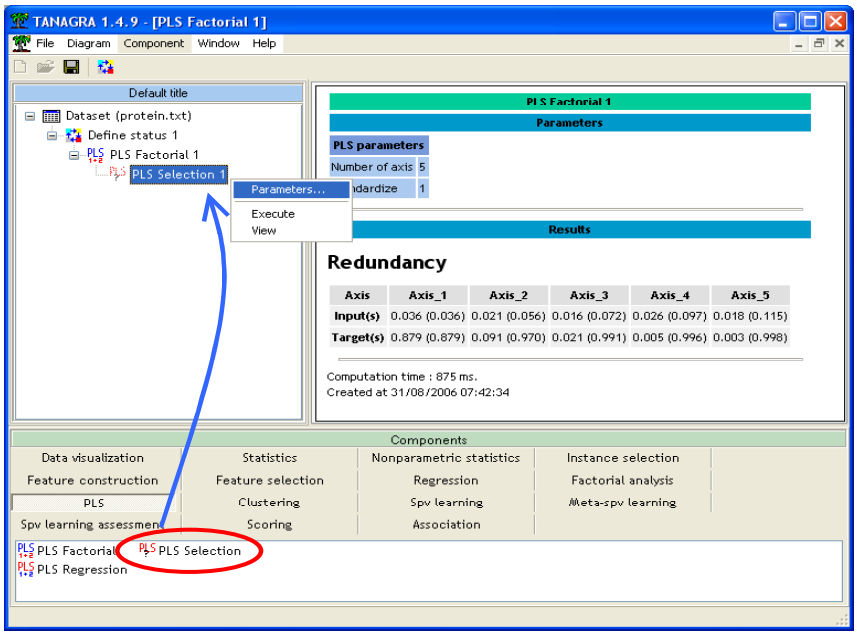
Par défaut, la méthode produit automatiquement 5 facteurs. Nous avons la possibilité de moduler les résultats à afficher, dans le mode par défaut, nous observons principalement les redondances.



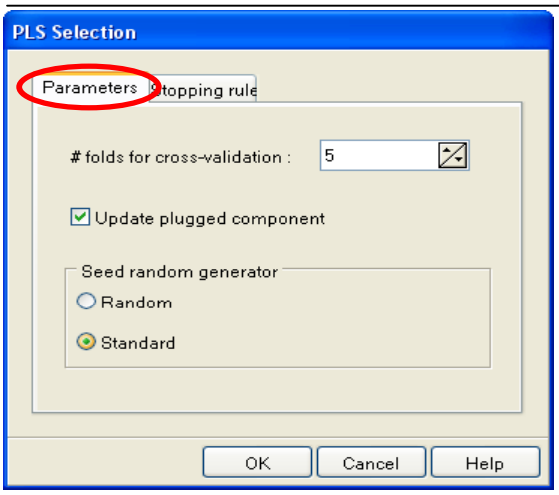
Les résultats montrent que les 5 premiers axes expliquent 99,8% des valeurs de Y. Mais au vu des redondances, à partir du 3ème axe, leur pouvoir explicatif ne semble guère significatif. C'est cette intuition que nous devrons confirmer de manière plus rigoureuse avec la procédure de validation croisée.

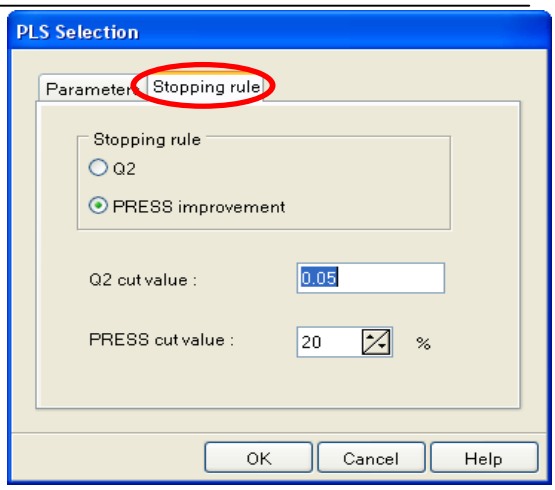
Sélection du nombre d'axes à retenir

Un nouvel outil (PLS-SELECTION) est dédié à la sélection automatique du nombre d'axes. Nous avons préféré élaborer un module à part pour pouvoir le brancher à la suite des différents composants qui effectuent une régression PLS. Dans notre cas, nous l'insérons à la suite de PLS FACTORIAL 1 dans le diagramme puis nous activons le menu PARAMETER.



Nous disposons de deux onglets dans la boîte de paramétrage.





Le premier onglet PARAMETERS permet de définir les paramètres de calcul. Nous pouvons notamment spécifier le nombre de parties dans la validation croisée (FOLDS). Si l'option UPDATE PLUGGED COMPONENT est cochée, le composant PLS associé sera automatiquement recalculé avec le nombre « optimal » de facteurs mis en avant par la procédure de détection.

Le second onglet STOPPING RULE permet de définir la règle d'arrêt dans l'exploration des solutions. A l'origine, notre idée était de reprendre la procédure fondée sur l'indicateur Q2 décrite dans l'ouvrage de Tenenhaus (La régression PLS, Technip, 1998, p.83), qui prend pour référence le logiciel SIMCA-P[1]. Nous retrouvons d'ailleurs le même descriptif dans la documentation du logiciel SIMCA-P. Mais nous n'avons pas pu obtenir les mêmes valeurs. Le mystère est levé par l'article de Chavent et Patouille (Calcul du coefficient de régression et du PRESS en régression PLS1, Revue MODULAD, n° 30) qui, au terme d'un jeu de pistes passionnant, indique la véritable formule utilisée par SIMCA-P. Tout serait pour le mieux si la formule utilisée par SIMCA-P a été modifiée depuis la version 9.0, sans qu'elle ne soit clairement explicitée dans la documentation. Nous n'avons pas retrouvé les résultats de l'ouvrage de Tenenhaus avec la version 11.0 du logiciel.

Dans TANAGRA, nous avons donc introduit deux approches pour la détection de la solution optimale : la première est toujours fondée sur le Q2 conformément (strictement) au descriptif dans l'ouvrage de Tenenhaus[2] ; la seconde est une variante qui teste si la réduction du PRESS (l'erreur quadratique en validation croisée pour chaque variable TARGET) est supérieure ou non à un seuil choisi par l'utilisateur. Le seuil de 20% permet de définir un comportement *raisonnable* sur les données que nous avons pu étudier. Tout cela est à améiorer bien sûr, l'accès au code source vous permettra de modifier à souhait la procédure.

[1] http://www.umetrics.com/default.asp/pagename/software_simcap/c/3

[2] Curieusement, nous retrouvons à peu près les résultats de la version de démonstration 11.0 de SIMCA-P, surtout lorsque les effectifs sont élevés.

Avec les paramètres par défaut, nous obtenons les résultats suivants.

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

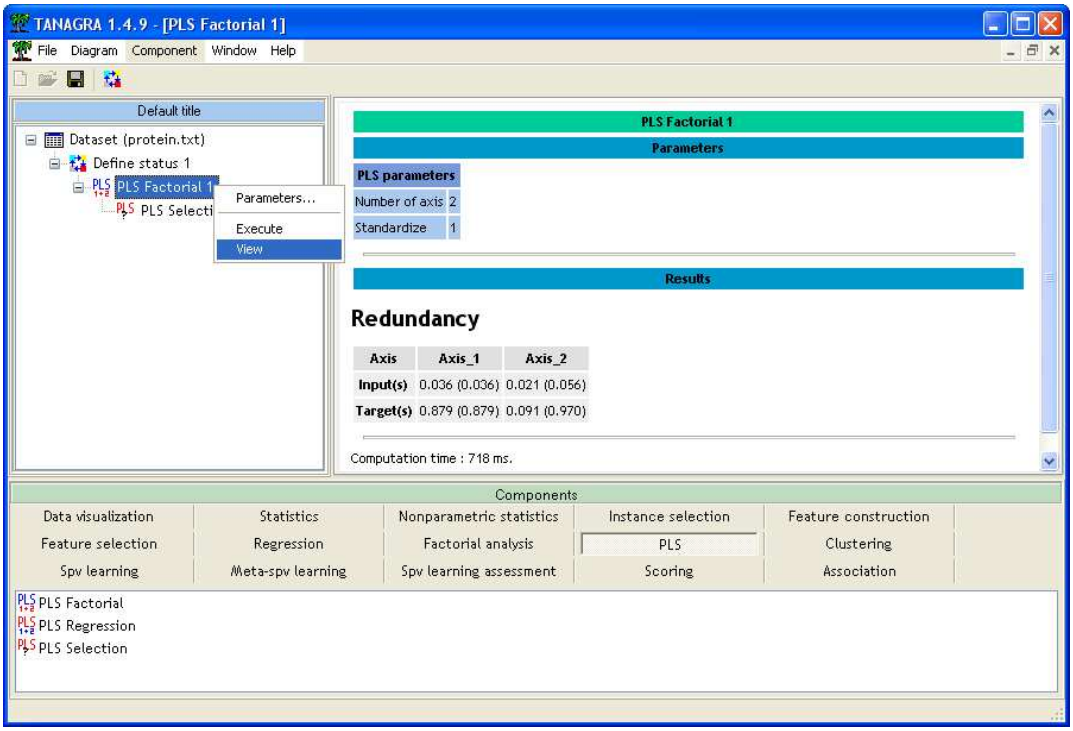
Component selection results

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 %

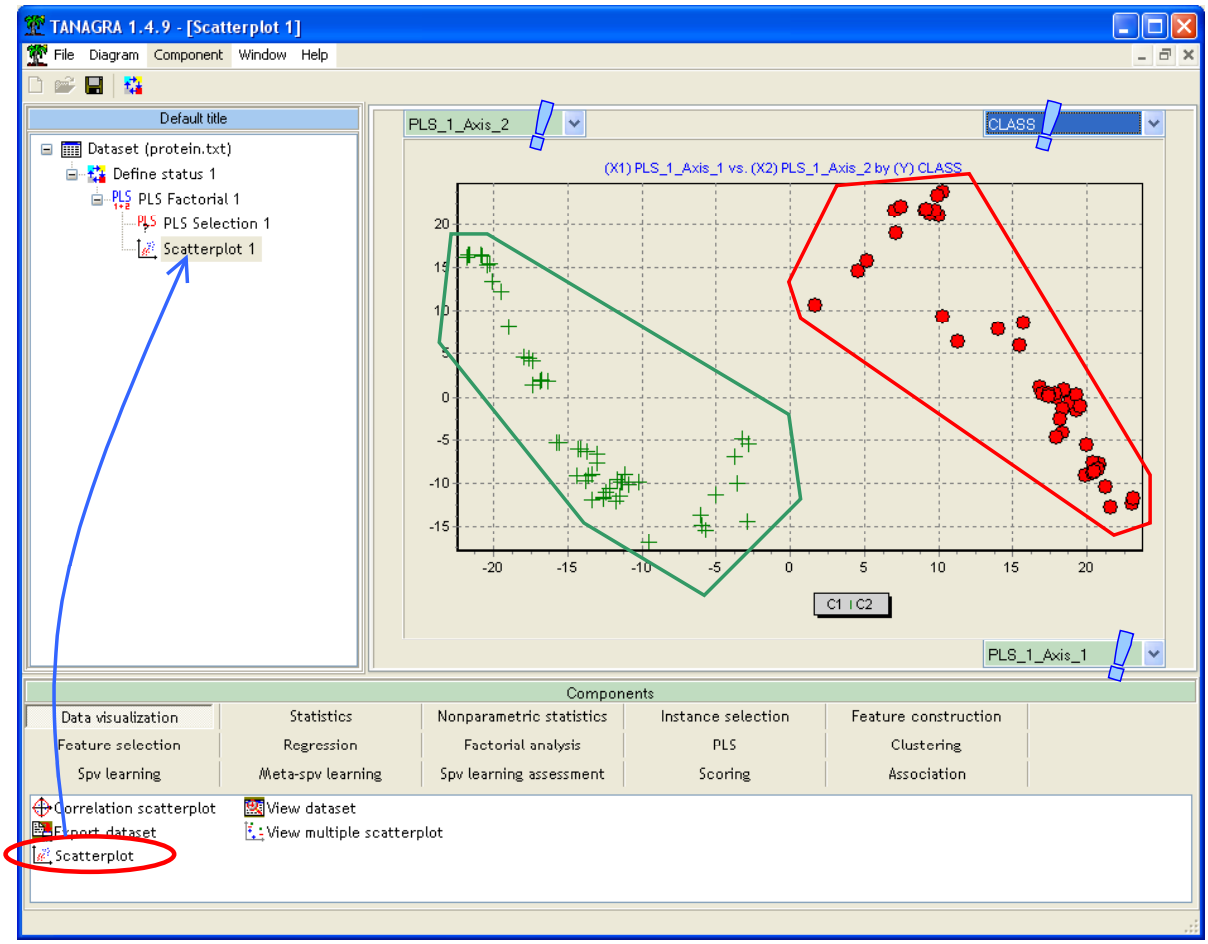
Il semble que deux axes suffisent pour prédire au mieux les valeurs de la variable à prédire. Au terme des calculs, le composant PLS FACTORIAL 1 dans le diagramme est automatiquement mis à jour avec un nombre de facteurs égal à 2. Vous pouvez vous en rendre compte en activant le menu VIEW du composant.



Projection des observations

Puisque nous disposons de deux axes, il est possible de projeter les individus dans le premier plan factoriel, notamment pour visualiser le positionnement des deux classes de protéines.

Pour ce faire, nous ajoutons un composant SCATTERPLOT dans le diagramme, nous sélectionnons le facteur 1 en abscisse, le facteur 2 en ordonnée, et nous illustrons les points à l'aide de la classe d'appartenance.



Le résultat est particulièrement plaisant. Nous distinguons bien les deux familles de protéines sur le premier plan factoriel produit par la régression PLS. Une procédure de classement fondée sur ces deux facteurs sera vraisemblablement très performante.