CheMOOCs

Activités des grains 08 et 09

Douglas Rutledge
AgroParisTech, Paris, France

Jean-Michel Roger
IRSTEA, Montpellier, France

V17.10

Table des matières

Première partie

Activités du grain 08.

1 Exercices de reproduction du grain 08.

A partir des jeux de données *pir.tab* et *ags.tab*, il vous est demandé d'établir un modèle PLS basé sur les spectres proche infrarouge pour la quantification de la teneur des huiles en acide oléique C18-1ω9.

Dans ChemFlow, créez un nouvel historique *CheMoocs-exercice-grain08-grain09* puis charger les données du répertoire **chemflow/shared data/data libraries/chemoocs/grain08** ou **/grain09**.

— 1. Un jeu d'étalonnage est obtenu en sélectionnant les 106 premières lignes de *pir.tab* et *ags.tab*. Les fichiers obtenus sont dénommés respectivement *new pir.tab* et *new ags.tab*.

Utiliser **utils/edit files** et dans **enter sample number** entrer *1 :106*. Vous pouvez soit répéter la même opération sur les deux fichiers, soit la faire simultanément sur les deux fichiers. Pour cela, il suffit de cliquer sur l'icône **multiple datasets** représentée par 2 feuilles qui permet de sélectionner puis traiter plusieurs fichiers l'un après l'autre exactement de la même manière.

— 2. Calculez un modèle de régression MLR entre la teneur en acide oléique C18-1ω9 dans *new ags.tab* et *new pir.tab* (laisser les paramètres de validation croisée par défaut).

— 3. Tracer le vecteur de coefficients-b de la MLR.

Utiliser la fonction **scatter plot** avec l'option **plot type → line/multi lines**.

— 4. Calculez un modèle de régression PCR entre la teneur en acide oléique C18-1ω9 dans *new ags.tab* et *new pir.tab*, avec une validation croisée de type stores Venitiens et 4 blocs, 20 variables latentes, données centrées.

Utiliser la fonction **pcr**.

— 5. Tracer les vecteurs de coefficients-b de la PCR pour 1, 3, 5 et 7 variables latentes.

Utiliser la fonction **scatter plot** et les options :

- **plot type → line/multi lines**
- **column for x axis →** *c1 :*
- **column for y axis →** *c2 :lv1 c4 :lv3 c6 :lv5 c8 :lv7*
- **use column names as legends →** *yes*

— 6. Tracer les valeurs prédites avec 1, 3, 5 et 7 variables latentes contre les valeurs observées.
Pour obtenir les valeurs prédites sans validation croisée, utiliser **regressions/apply a regression model to a new dataset**. Les valeurs prédites sont obtenues en appliquant le modèle obtenu (*pcr on (new pir.tab,new ags.tab) : model*) sur les mêmes données *new pir.tab* et *new ags.tab*. Le nombre de variables latentes est fixé à 7 (la plus grande des valeurs 1,3, 5 et 7).

— 7. Calculez un modèle de régression PLS entre la teneur en acide oléique C18-1ω9 dans *new ags.tab* et *new pir.tab*, avec une validation croisée de type stores Venitiens et 4 blocs, 20 variables latentes, données centrées, plus sortie des statistiques des points atypiques (outliers).
Utiliser la fonction **plsr** avec l'option : **compute outlier statistics → yes** .

— 8. Tracer les vecteurs de coefficients-b de la PLSR pour 1, 3, 5 et 7 variables latentes.
On procèdera de la même façon que pour la PCR.

— 9. Tracer les valeurs prédites avec 1, 3, 5 et 7 variables latentes contre les valeurs observées.

2 Exercices de compréhension du grain 08.

L'objectif de cet exercice est de comprendre la démarche de construction d'un modèle de prédiction de la teneur en triglycérides. Nous utiliserons les fichiers *pir.tab* et *trigly.tab*.

— 1. Chargez le fichier *pir.tab*, appliquer le prétraitement SNV, puis effectuer une ACP centrée - non réduite. Représentez les 2 premiers scores issus de l'ACP afin d'étudier la répartition des échantillons dans le plan factoriel 1-2. Les échantillons sont-ils tous répartis uniformément dans l'espace à deux dimensions ?

— 2. A partir des spectres prétraités par SNV, créez deux jeux de données : étalonnage et validation. Les échantillons seront tirés au hasard, 2/3 dans le jeu d'étalonnage et 1/3 dans le jeu de validation.
Utiliser la fonction **calibration-validation/split dataset** avec les options suivantes :
  — **select x data** → *snv(pir.tab)*
  — **select y data** → *trigly.tab*
  — **algorithm choice** → *random*
  — **percent of dataset for the validation dataset** → *0.33*

— 3. Utilisez la régression PLS pour construire un modèle d'étalonnage du triglycéride "OOO" (O représente l'acide oléique) à partir du jeu d'étalonnage (*xcal(pir.tab)*, *ycal(trigly.tab)*). Tracez

les valeurs de RMSEC-RMSECV en fonction du nombre de variables latentes.
Utiliser **plot/scatter plot** puis dans **plot type/ lines and points** cocher *rmsec-rmsecv.*

— 4. Appliquez ensuite ce modèle sur le jeu de validation (*xval(pir.tab)*, *yval(ags.tab)*). Tracez le RMSEP en fonction du nombre de variables latentes.

— 5. Qu'en concluez-vous sur le choix du nombre optimal de variables latentes ?

Deuxième partie

Activités du grain 09.

3 Exercice de reproduction du grain 09.

A partir les jeux de données *pir.tab* et *ags.tab*, il vous est demandé d'établir un modèle PLSR basé sur les spectres NIR prétraités par SNV pour la quantification de la teneur en acide palmitoléique C16-1ω7.

— 1. Sélectionnez les 106 premières observations de *pir.tab*, puis appliquer SNV. Sélectionner aussi les 106 premières observations de *ags.tab*. Etablissez un modèle de régression PLSR pour prédire la teneur en acide gras C16-1ω7, avec les options de validation croisée "stores Venitiens" et 4 blocs, 20 variables latentes, données centrées et statistiques des points atypiques.
N'oubliez pas au champ **compute outliers statistics (T2,Q, yresiduals)** cocher *yes.*

— 2. Tracez le RMSEC, RMSECV et R^2 en fonction du nombre de variables latentes.
Utilisez **plot/scatter plot** puis dans **plot type/ lines and points** cocher *rmsec-rmsecv* et R^2.

— 3. Calculez les statistiques élémentaires : minimum, maximum, moyennes, variances, écarts-types pour les 20 vecteurs de coefficients-b des régressions PLS.
Utilisez **statistics/summary**. Choisir les données de coefficients-b : *nipals-pls on new ags.tab : b-coeffs.* Utiliser **select/unselect** pour sélectionner toutes les colonnes, puis enlever (décocher) la première colonne, c'est à dire *c1 :.*

— 4. Tracez les variances des coefficients-b des régressions PLS en fonction du nombre de variables latentes.
NB : La variance des coefficients-b a pour objectif de mesurer l'augmentation d'amplitude des valeurs des coefficients-b avec le nombre de variables latentes. Un résultat tout à fait

équivalent, mais moins visuel, est obtenu en calculant la norme des vecteurs de coefficients-b plutôt que la variance de leurs valeurs.

Utiliser **scatter plot** avec les options :

- **plot type** → *lines and points*
- **dataset** → *summary on nipals...*
- **column for x-axis** → *c1 :*
- **column for y axis** → *c7 :var*

— 5. Tracez les critères de Durbin-Watson des coefficients-b des régressions PLS en fonction du nombre de variables latentes.

— 6. Tracez les valeurs prédites par PLSR en validation croisée avec 5 variables latentes contre les valeurs observées.

— 7. Tracez les T^2 de Hotelling, puis les T^2 de Hotelling contre les résidus Q pour un modèle avec 5 LVs.

NB : Un synonyme à résidus Q est le terme : variance résiduelle des résidus (residual X-variance). Les résidus Q peuvent aussi être remplacés par un autre critère : DModX qui est la distance d'un point au modèle. DModX est proportionnel à la racine carrée de Q.

Utiliser **scatter plot** avec **plot type** → *lines and points* et **use first column as sample label** → *yes* pour faire apparaitre les noms des observations sur les graphes.

4 Exercice de compréhension du grain 09.

A partir des 106 premières observations des jeux de données *pir.tab* et *ags.tab*, il vous est demandé d'établir un modèle PLS basé sur les spectres PIR prétraités par SNV pour la quantification de la teneur en acide oléique C18-1ω9.

— 1. Construisez un modèle de régression PLS entre la teneur en acide oléique et les spectres infrarouge, avec les options de centrage des données, validation croisée avec 4 blocs.

— 2. Tracez le RMSEC, RMSECV et R^2 en fonction du nombre de variables latentes.

— 3. Calculez les moyennes, écarts-types, variances des 20 vecteurs de coefficients-b des régressions PLS.

— 4. Tracez les variances des coefficients-b des régressions PLS en fonction du nombre de variables latentes.

— 5. Tracez les critères de Durbin-Watson des coefficients-b des régressions PLS en fonction du nombre de variables latentes.

— 6. A partir des figures obtenues aux questions 2, 4 et 5, choisissez le meilleur modèle en argumentant.

— 7. Tracez les valeurs prédites par PLS avec le nombre optimal de variables latentes contre les valeurs observées.

— 8. Tracez les T^2 de Hotelling contre les résidus Q pour un modèle avec le nombre optimal de variables latentes.

— 9. Reconstruisez le modèle d'étalonnage après avoir enlevé les deux observations atypiques, lignes 40 et 66. Qu'en concluez-vous ?