Activités des

grains 04 et 05

Sébastien Preys
Ondalys, Montpellier, France

Eric Latrille
INRA, Narbonne, France

Jean-Michel Roger
IRSTEA, Montpellier, France

Martin Ecarnot
INRA, Montpellier, France

V18.10

Table des matières

1 Pré-requis.

— vidéos des grains 3, 4 et 5.

— tutoriels ChemFlow.

2 Durée estimée.

— 1h.

3 Description du jeu de données.

Les données ont été produites par l'Université d'Aix-Marseille, équipe de Nathalie Dupuy. Des analyses proche infrarouge ont été réalisées sur les mêmes 187 huiles d'olives que pour l'activité du grain 03. Elles comprennent :

— un jeu de 187 spectres comprenant 612 longueurs d'onde, 1000 à 2222 nm : *pir.tab* ;

— un codage des 187 échantillons selon les 6 origines géographiques avec des lettres : AP=Aix en Provence, HP=Haute Provence, NI=Nice, NM=Nimes, NY=Nyons, VB=Vallée des Baux de Provence ; plus un autre codage avec des nombres : 1=Aix en Provence, 2=Haute Provence, 3=Nice, 4=Nimes, 5=Nyons, 6=Vallée des Baux de Provence : *labels2.tab*.

4 Chargement des données.

Dans ChemFlow, créez l'historique *CheMoocs-exercice-grain04-grain05* depuis le cadre **history** en haut à droite. Depuis le répertoire **chemflow/shared data/data libraries/chemoocs/grain04** ou **grain05**, importer les spectres, fichier *pir.tab*, ainsi que le fichier codant les origines des observations, *labels2.tab*.

5 Exercice 1 : visualisation des données.

— Visualiser les données sous forme d'un tableau de données.

Dans ChemFlow cliquez sur l'œil.

— Visualiser les spectres sous forme d'un graphe.

Dans ChemFlow, cliquez successivement sur **plots** puis **spectra plot**. Choisir le titre, les noms

des labels des abscisses et ordonnées, ou laisser les valeurs par défaut. Cliquer sur **execute**. Une fois la tâche terminée (en vert), activez **scratchbook** dans la barre du haut puis visualisez les spectres en cliquant sur l'œil. Téléchargez la figure en cliquant sur l'icône de téléchargement en haut à droite de la figure.

— Visualiser les spectres en affectant une couleur à chaque origine. Peut-on différencier les différentes origines à l'aide de cette visualisation ?

Procéder de même que précédemment, mais à l'onglet **use a column of a dataset as spectra column** sélectionnez *yes* et dans l'onglet **dataset** juste en dessous sélectionnez *labels2.tab* puis pour **columns for color** choisir *c2 :code1* ou *c3 :code2*.

6 Exercice 2 : visualisation des points atypiques (outliers) d'une ACP.

— Réaliser une ACP sur le jeu de données avec centrage préalable, sans normalisation.

Depuis ChemFlow, aller dans **exploration** puis dans **pca**. Renseigner le nom du fichier (*pir.tab*) et vérifier que les options **centering option** et **scaling option** sont bien à *yes* et *no* respectivement.

— Visualiser les valeurs numériques du pourcentage de variance expliquée en fonction du nombre de composantes principales.

— Tracer le diagramme de variance expliquée (= éboulis). Qu'observe-t-on ?

Pour obtenir un graphique en barres, utiliser **plot/barplot**, et renseigner :

- **dataset** → *pca on pir.tab : explained variance (%)*
- **column for x axis** → *c2 :explained variance*
- **label for y axis** → *pourcentage de variance*

— Tracer les plans 1-2 et 3-4 de la carte factorielle des individus (score plot). Observe-t-on des outliers, si oui lesquels ? Peut-on différencier les différentes origines à l'aide de cette visualisation ?

Les scores sont dans : *pca on pir.tab : scores*. La carte factorielle est obtenue avec **plot/scatter plot** et renseigner les options suivantes :

- **label for x axis** → *pc 1*
- **label for y axis** → *pc 2*
- **series/plot type** → *points*

- **series/plot type/x-dataset** → *pca on pir.tab : scores*
- **series/plot type/column for x axis** → *c2 : pc1*
- **series/plot type/y-dataset** → *pca on pir.tab : scores*
- **series/plot type/column for y axis**→ *c3 : pc2*
- **series/plot type/add first column of x-dataset as sample label**→ *yes*
- **series/plot type/use a column of a dataset as point color**→ *yes*
- **series/plot type/use a column of a dataset as point color/dataset** → *labels2.tab*
- **series/plot type/use a column of a dataset as point color/column for color** →*c3 : code2*

Faire de même pour pc3 et pc4.

7 Exercice 3 : application de prétraitements.

- Effectuer les pré-traitements suivants sur les spectres :
  - Supprimer la zone spectrale 1000-1148 nm.

    Il faut d'abord connaître les numéros des colonnes correspondant aux longueurs d'onde 1000-1148 nm en visualisant les données des spectres avec **scratchbook** et en utilisant le curseur en bas pour se déplacer vers la droite. Cela donne les n° 2 à 76 - le n°1 est celui des labels des observations-.

    La sélection de variables se fait depuis **utils/edit files**. Renseigner les différents champs :

    - **select dataset** → *pir.tab*
    - **select operation** → *delete*
    - **select operation on** → *columns*
    - **select from** → *column number*
    - **select dataset** → *pir.tab*
    - **enter column number(s)** → *2 :76*

    Bien noter *2 :76* (sans espace entre 2 et 76) dans le dernier champ ; les deux-points indiquent qu'on prend toutes les variables entre la 2eme et la 76eme comprises. Un nouveau fichier apparait dans l'historique : *new pir.tab*.

    Il est aussi possible de supprimer les colonnes selon leurs labels, mais comme beaucoup de colonnes sont à supprimer c'est beaucoup plus long.

  - Appliquez le prétraitement Savitzky-Golay, dérivée première, polynôme de degré 2, fenêtre de 7 points.

    pretreatments/sg

  - Appliquez le prétraitement Standard Normal Variate (SNV) sur les données ayant subi Savitzky-Golay.

    pretreatments/snv

- Réaliser une ACP sur le jeu précédent de données prétraitées, avec l'option centrage, sans l'option réduction.
- Visualiser les valeurs numériques du pourcentage de variance expliquée en fonction du nombre de composantes principales .
- Tracer le diagramme de variance expliquée (= éboulis). Qu'observe-t-on ?.
- Tracer les plans 1-2 et 1-3 de la carte factorielle des individus (score plot). Observe-t-on des outliers, si oui lesquels ? Peut-on différencier les différentes origines à l'aide de cette visualisation ?