CheMOOCs

# Activités des grains 06 et 10

Philippe Courcoux
Oniris, Nantes, France

Jean-Michel Roger
IRSTEA, Montpellier, France

Martin Ecarnot
INRA, Montpellier, France

V18.10

L'auteur autorise toute utilisation de l'oeuvre originale (y compris à des fins commerciales) ainsi que la création d'oeuvres dérivées, à condition qu'elles soient distribuées sous une licence identique à celle qui régit l'oeuvre originale présentée.

# Table des matières



L'auteur autorise toute utilisation de l'oeuvre originale (y compris à des fins commerciales) ainsi que la création d'oeuvres dérivées, à condition qu'elles soient distribuées sous une licence identique à celle qui régit l'oeuvre originale présentée.

## Première partie

# Activités du grain 06

L'objectif de cet exercice est de comprendre la démarche de construction d'une classification non supervisée de spectres NIR. Les données sont la collection de spectres NIR d'huiles d'olives déjà utilisée aux grains 03-04-05, avec les identifiants d'origine géographique de chaque huile et les analyses d'acides gras. Parmi les acides gras, seules les valeurs de l'acide linoléique (C18-2$\omega$6) seront utilisées.

— 1- Charger les données.

Dans ChemFlow, créer un nouvel historique : *CheMoocs-exercice-grain06*. Les trois fichiers à importer sont : *pir.tab*, *ags.tab* et *labels2.tab*. Ils se trouvent dans **chemflow/shared data/data libraries/chemoocs/grain06**. Ce sont les mêmes données que celles utilisées pour les grains 03, 04 et 05.

— 2- Sélection d'une plage spectrale.

Sélectionnez les longueurs d'onde entre 1600 et 2000 nm et représenter l'ensemble de ces spectres sur une figure.

Cliquez sur **scratchbook** puis sur l'œil du fichier *pir.tab* afin de le visualiser. Déplacer le curseur vers la droite afin d'afficher les longueurs d'onde demandées, 1600 puis 2000 nm. Noter les numéros de colonnes correspondants : 302 et 502.

La sélection de ces colonnes se fait avec **utils/edit files**.

— 3- Prétraitement des spectres et ACP.

Appliquez un prétraitement SNV, puis effectuer une ACP centrée non réduite des spectres transformés. Représentez ces spectres, puis les valeurs des scores sur le plan formé par les deux premières composantes principales de l'ACP, indiquez le pourcentage de variance des composantes principales sur le graphique et labellisez les échantillons par la variété d'origine.

Le fichier obtenu après SNV s'appelle *snv(new pir.tab)*. Pour la figure des scores sur les axes 1 et 2 de l'ACP, on utilisera la fonction **scatter plot** avec le fichier *pca scores :snv(new pir.tab)* pour les valeurs de scores, et dans **use a color of a dataset as point color** choisir **yes** puis le fichier *labels2.tab* pour l'identification de la variété de chaque échantillon, et dans **column for color** choisir *c2 :code1* ou *c2 :code2*.

— 4- Classification non supervisée.

Faites une classification hiérarchique des spectres SNV avec la méthode Ward, et représentez

L'auteur autorise toute utilisation de l'oeuvre originale (y compris à des fins commerciales) ainsi que la création d'oeuvres dérivées, à condition qu'elles soient distribuées sous une licence identique à celle qui régit l'oeuvre originale présentée.

l'arbre hiérarchique résultant. Coupez l'arbre hiérarchique en 6 groupes, calculez les spectres moyens de ces classes et représentez-les. Représentez les scores des échantillons sur les axes 1-2 de l'ACP en les coloriant par leur classe d'appartenance déterminée après classification ascendante hiérarchique (CAH). Représentez aussi les deux premiers vecteurs-propres de l'ACP obtenue à l'étape 3.

La CAH est obtenue avec la fonction **clustering/hierarchical clustering**. Les options suivantes sont à renseigner :

— **x data** → *snv(new pir.tab)*

— **distance choice** → *euclidian*

— **method option** → *ward*

— **choice of cluster number** → *6*

Le fichier de sortie de la classification nommé *hc on snv(new pir.tab) : cluster number* contient le résultat de la classification, c'est à dire l'attribution de chaque observation à un des 6 groupes qui ont été demandés. Le calcul de la moyenne de chacun de des 6 groupes se fait en utilisant ce fichier. Aller dans la fonction **statistics/mean** et renseigner les options suivantes :

— **dataset** → *snv(new pir.tab)*

— **select all variables of the dataset** → *yes*

— **compute the mean by a column factor** → *yes*

— **dataset** → *hc on snv(new pir.tab) : cluster number*

— **column factor choice for mean** → *c2 :cah.cluster*

Les spectres moyens peuvent être représentés sur une figure, en utilisant la fonction **spectra plot** avec les options suivantes, les autres étant laissées par défaut :

— **plot title** → *spectres moyens des 6 groupes de la CAH*

— **label fox x axis** → *longueurs d onde*

— **label for y axis** → *absorbances*

— **dataset** → *mean on snv(new pir.tab))*

Les scores des échantillons sur les axes 1-2 de l'ACP sont représentés grâce à la fonction **scatter plot** avec le fichier *pca on snv(new pir.tab) : scores* pour les valeurs de scores ainsi que le fichier *hc on snv(new pir.tab) : cluster number* pour l'identification de la classe attribuée à chaque échantillon par la CAH. Enfin, la figure des vecteurs-propres 1 et 2 est obtenue avec la fonction **scatter plot** et les options suivantes :

— **plot type** → *line/multiline*

L'auteur autorise toute utilisation de l'oeuvre originale (y compris à des fins commerciales) ainsi que la création d'oeuvres dérivées, à condition qu'elles soient distribuées sous une licence identique à celle qui régit l'oeuvre originale présentée.

- **dataset** → *pca on new pir.tab : loadings*
- **column(s) for x-axis** → *c1*
- **column(s) for y-axis** → *c2 :pc1  c3 :pc2*
- **line color** → *multicolor*

— 5- Partition k-means.

Faites un partitionnement de type k-means en 6 groupes, initialisé par la moyenne des 6 classes obtenues précédemment par coupure de l'arbre de CAH. Calculez les spectres moyens de ces 6 nouvelles classes issues de k-means et faites-en une représentation graphique.

Pour le partitionnement k-means en 6 groupes, il faut utiliser la fonction **clustering/km** et les paramétrages suivants :

- **x data** → *snv(new pir.tab)*
- **use a file to initialize cluster centers** → *mean on snv(new pir.tab)*
- **choice of iteration number** → *50*

Le fichier *mean on SNV(new pir.tab)* est celui qui a été obtenu à la partie précédente : les 6 moyennes sont celles de la CAH.

La sortie de la partition k-means est le fichier *km on snv(new pir.tab) : cluster number.* Chacune des observations est associée à un groupe. Comme précédemment, ce fichier va être utilisé avec la fonction **mean** pour calculer les moyennes des 6 groupes obtenus par k-means. La représentation graphique des 6 spectres moyens issus de k-means est obtenue comme au paragraphe précédent.

— 6- Représentation des observations par leur groupe issu de k-means.

Représentez les échantillons sur les scores des deux premières composantes principales de l'ACP obtenue à l'étape 3 en les coloriant par leur classe d'appartenance.

On utilisera la fonction **scatter plot**, avec les fichiers *pca on snv(new pir.tab) : scores* pour les valeurs de scores et *km on snv(new pir.tab) : cluster number* pour l'identification des classes issues de k-means.

— 7- Comparaison avec l'acide linoléique.

Visualiser le fichier *ags.tab* et noter à quelle colonne correspond l'acide linoléique.

Faites une représentation de type boxplot de la teneur en acide linoléique par classe d'appartenance. Examinez le lien entre la teneur en acide linoléique des échantillons et la position de ceux-ci sur le premier plan de l'ACP.

La fonction à utiliser est **plot/boxplot by factor level**, avec les paramètres suivants :

L'auteur autorise toute utilisation de l'oeuvre originale (y compris à des fins commerciales) ainsi que la création d'oeuvres dérivées, à condition qu'elles soient distribuées sous une licence identique à celle qui régit l'oeuvre originale présentée.

— **dataset** → *ags.tab*
— **column(s) for x-axis** → *c10 :c18-2ω6*
— **plot title** → *Boxplot de l acide linoleique*
— **plot a boxplot by a column factor** → *yes*
— **column factor choice for boxplot** → *c2 :km.cluster*

## Deuxième partie

# Activités du grain 10

## 1 Exercice de compréhension du grain 10

Les données sont des spectres proche infrarouge réalisés sur des échantillons de sol provenant de deux origines différentes. Dans le fichier « classes » qui correspond à l'origine, l'échantillon appartient, soit à la classe 1 et son nom commence par « B », soit à la classe 2 et son nom commence par « A ».

Créez un nouvel historique : *CheMoocs-exercice-grain10* puis chargez les fichiers *sols_classes.tabular* et *sols_spectres.tabular* depuis **chemflow/shared data/data libraries/chemoocs/grain10_grain11**.

— 1. Tracez les spectres bruts (sans prétraitement). A quel type d'effet (multiplicatif, additif, dérive de ligne de base) est soumis ce jeu de spectres ?
Utilisez la fonction **spectra plot**.
— 2. Réalisez une ACP centrée - non réduite sur les spectres et dessinez la carte factorielle (des scores) des 2 premières composantes, Qu'observez vous ?
Utilisez la fonction **exploration/pca** → *spectres.tabular*, laissez les autres options par défaut (centré-non réduit) et exécutez.
L'édition de la carte factorielle se fait avec la fonction **plots/scatter plot** :
— **series/plot type** → *points*
— **x-dataset** → *pca on sols_spectres.tabular :scores*
— **column for x-axis** → *c2 :pc1 94.1%*
— **y-dataset** → *pca on sols_spectres.tabular :scores*

L'auteur autorise toute utilisation de l'oeuvre originale (y compris à des fins commerciales) ainsi que la création d'oeuvres dérivées, à condition qu'elles soient distribuées sous une licence identique à celle qui régit l'oeuvre originale présentée.

- **column for x-axis** → *c3 :pc2 3.94%*
- **add first column of x-dataset as sample label** → *yes*
- **use a column of a dataset as point color** → *yes*
- **dataset** → *2 :sols_classes.tabular*
- **column for color** → *c2 :classe*

— 3.Appliquez un prétraitement SNV sur les spectres bruts. Puis réalisez une ACP sur les spectres obtenus et dessinez les scores (carte factorielle) des 2 premières composantes. Quelles différences observez-vous par rapport à l'ACP réalisée sur les spectres bruts ?

— 4.Appliquez maintenant un prétraitement Detrend d'ordre 2 sur les spectres bruts. Réalisez à nouveau l'ACP et la carte factorielle. Quelles modifications obtenez-vous par rapport au prétraitement SNV ?

L'auteur autorise toute utilisation de l'oeuvre originale (y compris à des fins commerciales) ainsi que la création d'oeuvres dérivées, à condition qu'elles soient distribuées sous une licence identique à celle qui régit l'oeuvre originale présentée.