L'Analyse en Composantes Principales, technique fondatrice de la Chimiométrie

Christophe Cordella

UMR1145 INRA, AgroParisTech,
Laboratoire de Chimie Analytique, 75005 PARIS

L'Analyse en Composantes Principales est une des techniques fondatrices de la chimiométrie particulièrement adaptée à l'étude du comportement des huiles au chauffage.

La Chimiométrie en quelques mots

La chimiométrie est une discipline associant initialement analyse de données et chimie analytique. Elle rassemble et développe un ensemble d'outils mathématiques utilisés pour extraire de l'information structurée et interprétable à partir de données chimiques. Le terme *chimiométrie* vient de l'anglais *chemometrics* et a été inventé il y a près de quarante ans, en 1971, par Svante Wold.

Plusieurs outils très répandus de la chimiométrie tirent leurs origines du domaine de la biométrie avec les travaux de Karl Pearson en 1901 et de l'économie avec ceux de Harold Hotelling. Aujourd'hui, les outils de la chimiométrie intéressent l'ensemble des applications de la chimie, mais aussi de la physique, des sciences de la vie, de l'économie, de la sociologie, des méthodes statistiques et de l'informatique.

En instrumentation, la chimiométrie consiste à **modéliser** les variations d'un certain nombre de variables dont l'obtention est délicate (nécessitant une analyse chimique par exemple) en fonction d'autres variables mesurables "plus facilement" afin de pouvoir se passer ultérieurement de la mesure des premières.

On distingue trois opérations ou catégories d'actions applicables aux données :

- La **description** et la **classification**, étapes importantes dans la visualisation des données, l'exploration des propriétés caractéristiques de l'ensemble de données étudié et la classification des objets qui le composent. La description est une étape primordiale qui permet la mise en évidence des relations entre les variables mesurées (covariation ou corrélation, normalité, anomalie statistique *outlier* (*se dit d'une mesure très à l'écart des autres*).

L'analyse en composantes principales se range dans cette catégorie d'outils.

- L'é**talonnage** en laboratoire (ou **modélisation**) où toutes les mesures de variables doivent être réalisées et où le modèle (ou "prédicteur") est calculé.
- La **prévision**, utilisation courante "sur le terrain", où seules les variables plus facilement accessibles sont mesurées, les autres étant calculées à l'aide du modèle.

Il faut ajouter une étape de **validation des modèles.**

L'Analyse en Composantes Principales (ACP)

L'ACP fait partie du groupe des méthodes descriptives multidimensionnelles appelées méthodes factorielles (descriptives, non supervisées), elle constitue un domaine important de la Chimiométrie. Ces méthodes sont apparues au début du XX^{e} siècle, dans les années 1930, et ont été surtout développées en France dans les années 1960, en particulier par Jean-Paul Benzécri qui a beaucoup exploité les aspects géométriques et les représentations graphiques. Dans la mesure où ce sont des méthodes descriptives, elles ne s'appuient pas sur un modèle probabiliste, mais sur un modèle géométrique. L'ACP propose des représentations géométriques de ces objets et de ces variables organisées sous forme d'un tableau rectangulaire de données comportant les valeurs de $\boldsymbol{p}$ variables quantitatives pour $\boldsymbol{n}$ objets (termes également rencontrés : unités, individus ou observations). Les représentations des objets permettent de voir s'il existe une structure, non connue *a priori*, sur cet ensemble d'objets. De façon analogue, les représentations des variables permettent d'étudier les structures de liaisons linéaires sur l'ensemble des variables considérées.

Pour les variables, on cherchera quelles sont celles qui présentent une corrélation positive plus ou moins forte, et celles qui, au contraire, sont corrélées négativement.

D'un point de vue géométrique, l'ACP recherche les directions de l'espace selon lesquelles les individus sont les plus dispersés, en supposant que ces directions sont les plus intéressantes. Après avoir trouvé une direction, on cherche la suivante avec la contrainte supplémentaire qu'elle soit orthogonale aux précédentes. De cette façon, une fois ces directions déterminées, elle produit $\boldsymbol{n}$ nouveaux axes orthogonaux nommés "composantes principales" ou CP qui sont des combinaisons linéaires des variables initiales . L'ACP est très utile lorsqu'il s'agit de compresser l'information contenue dans un grand nombre de variables, $\boldsymbol{N}$, car les $\boldsymbol{n}$ premiers axes ($\boldsymbol{n} << \boldsymbol{N}$) contiennent souvent la plus grande partie de la variabilité, et, on l'espère, de l'information intéressante.

Enfin, comme pour toute méthode descriptive, réaliser une ACP n'est pas une fin en soi. Elle servira à mieux connaître les données sur lesquelles on travaille, à détecter d'éventuelles valeurs suspectes, et aidera à for-

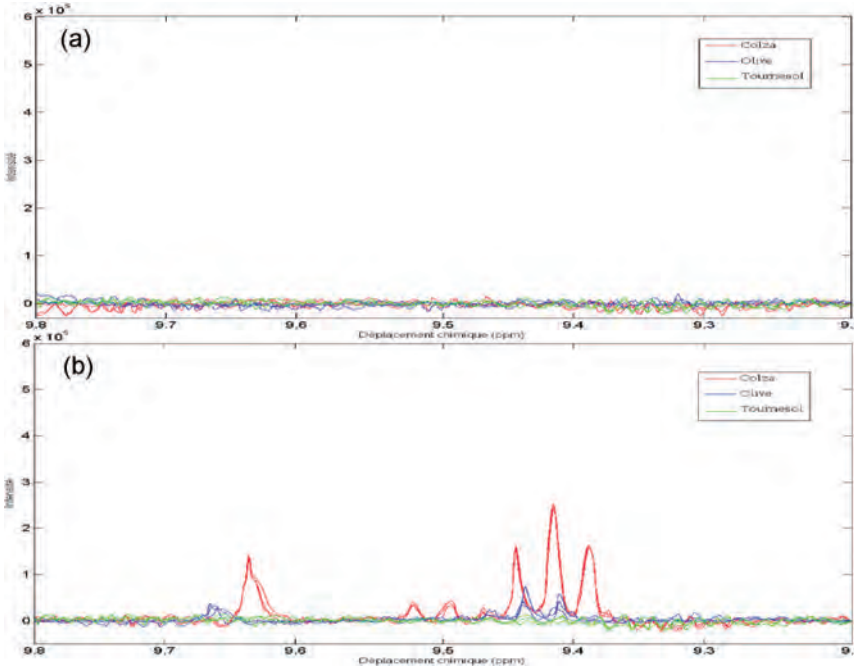
muler des hypothèses qu'il faudra étudier à l'aide de modèles et d'études statistiques inférentielles.

L' ACP pour étudier et comparer l'effet du chauffage d'huiles

Trois types d'huiles (colza, tournesol, olive) sont chauffées selon un protocole consistant à prélever toujours la même quantité d'huile et à la chauffer dans une étuve thermostatée à 170°C, 190°C et 210 °C pendant 0, 30, 60, 90, 120, 150 et 180 min.

Les résultats de l'analyse en composantes principales nous permettent de visualiser l'effet du chauffage en fonction du temps sur la qualité globale des huiles, c'est-à-dire indirectement sur la composition chimique des huiles. Les transformations chimiques se produisant au cours du temps sous l'effet de la température modifient l'intensité et la position des pics sur les spectres RMN du proton. Ces changements chimiques se traduisent par des positions relatives (coordonnées factorielles) différentes des échantillons sur les plans définis par les premières composantes principales.

Sur la figure 1, nous présentons dans un premier temps les spectres RMN-1H des trois huiles chauffées à 190 °C pendant une durée variable, ceci afin de montrer leur comportement vis-à-vis de la chaleur et de mettre en évidence l'effet du chauffage sur les modifications chimiques au sein de l'huile au cours du temps.



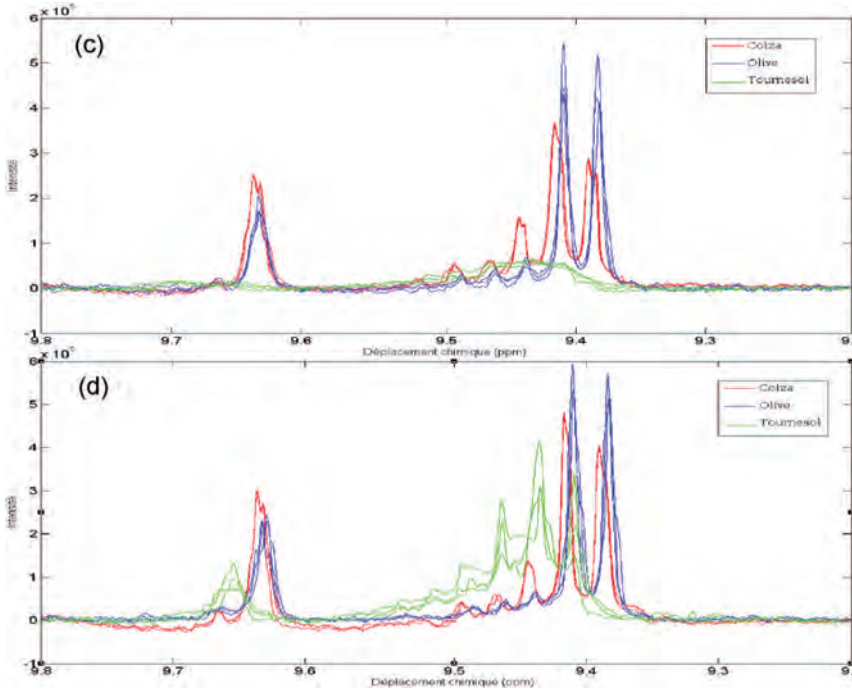


Figure 1
Spectres RMN du proton des huiles
(colza en rouge, olive en bleu, tournesol en vert
(a) froides,
(b) chauffées 30 min à 190 °C,
(c) 120 min à 190 °C,
(d) 180 min à 190 °C.

L'évolution des spectres de RMN du proton des huiles de colza, tournesol et olive, pour la région comprise entre 9,2 à 9,8 ppm correspond au déplacement chimique des aldéhydes. On constate qu'au cours du temps apparaît une quantité d'aldéhyde plus ou moins importante et plus ou moins rapide selon la nature de l'huile. Ces aldéhydes proviennent des réactions d'oxydation des acides gras insaturés naturels.

L'Analyse en Composantes Principales de ces spectres permet de dresser des cartes factorielles. Leur interprétation se fait de la même façon quelle que soit l'huile mais l'effet obtenu diffère d'une huile à l'autre car la répartition des points représentant les échantillons n'est pas identique d'une carte factorielle à l'autre. La figure (2) présente les résultats de l'ACP sur l'huile de colza chauffée à température fixe (190 °C) pendant une durée variable (de 30 à 180 min), et sur cette même huile chauffée à

différentes températures pendant une durée fixe (180 min). On peut visualiser sur le plan formé par les axes CP1 et CP2 , le “chemin thermique” suivi par l’huile au cours du chauffage.

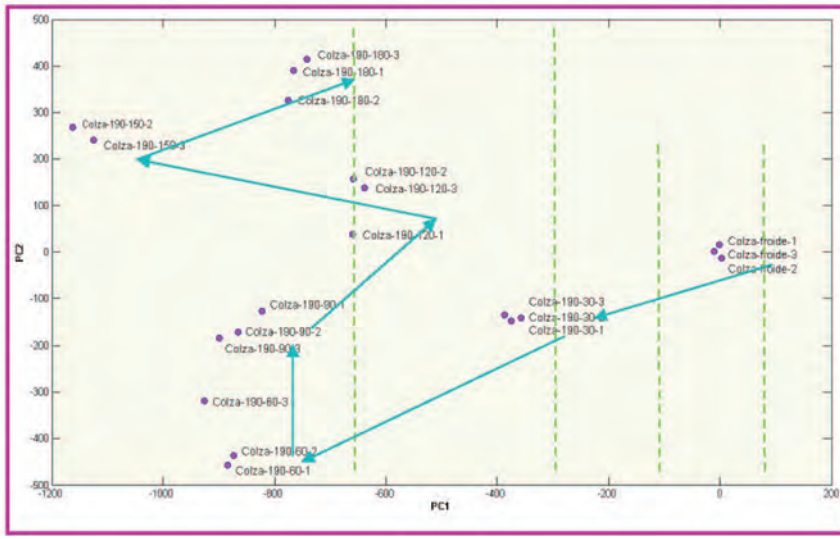


Figure 2
Carte factorielle (plan CP1 x CP2) des échantillons
formée par les deux premières composantes principales.
Huile : colza ; chauffage : 190 °C ; temps = 30 à 180 min.

Ce chemin virtuel rend compte des différences chimiques entre les échantillons d’huile chauffés à 190 °C pendant différentes durées. On voit nettement la différence entre tous les triplets de points, ces derniers étant des répétitions de la même mesure, c’est-à-dire des répétitions de l’acquisition RMN-1H pour chaque échantillon d’huile. On en déduit que chaque température de chauffage engendre des réactions chimiques qui ne sont pas toutes identiques soit d’un point de vue réactionnel, soit d’un point de vue cinétique. On constate que les triplets de points sont bien séparés et donc que les composés chimiques qui caractérisent ces échantillons ne sont pas en quantité égale et en partie sans doute de nature et/ou de structure différente.

De manière générale, sur une carte factorielle, la proximité des groupes d’échantillons indique une ressemblance chimique, l’éloignement est signe de différences.

En résumé, nous visualisons grâce à l’ACP qu’à 190 °C, les échantillons se rangent en cinq groupes chimiques, voir tableau suivant :

Groupe	Echantillons	Durée de chauffage (minutes)	Composante(s) principale(s) utiles(s) à la séparation
Gr1	Colza-froide Colza-190-30	0 ; 30	CP1
Gr2	Colza-190-60 Colza 190-90	60 ; 90	CP1 + CP2
Gr3	Colza-190-120	120	CP2
Gr4	Colza-190-150	150	CP1 + CP2
Gr5	Colza-190-180	180	CP2

Interprétation des cartes factorielles : *scores* et *loadings*

La figure 3 montre le résultat d'une ACP sur les données enregistrées à trois températures de chauffage différentes. L'effet "température" est ainsi mis en évidence. Comme précédemment, la représentation en cartes factorielles constitue une aide à l'interprétation chimique des transformations dont l'huile a été le siège durant ce chauffage. Un résultat similaire est obtenu pour les autres huiles. La carte factorielle synthétise les informations contenues dans les spectres RMN et prend en compte les proportions entre les aires des pics enregistrés. L'interprétation quantitative d'un spectre reste difficile visuellement surtout dans le cas d'une substance naturelle telle que l'huile.

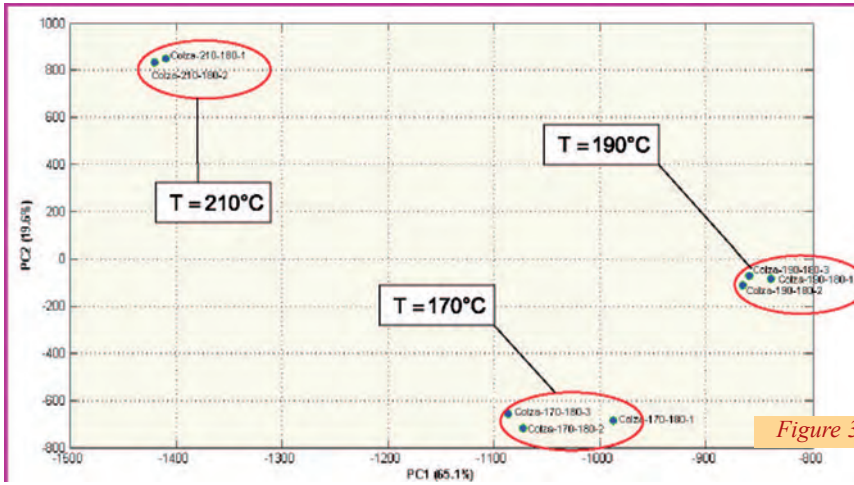


Figure 3

La figure 4 montre effectivement que le chauffage provoque d'importantes modifications chimiques dont la conséquence est l'apparition d'un ou plusieurs composés aldéhydiques pour lesquels l'aire des pics varie notablement. Les transformations sont détectables à l'œil mais leur ampleur ne se mesure pas aisément à partir des spectres. La valeur des coordonnées factorielles de chaque échantillon sur une composante principale donnée en fonction du numéro de l'échantillon dans la matrice est nommée *scores*. Ces *scores* traduisent l'importance des changements chimiques opérés dans les groupes échantillons en matérialisant graphiquement les distances qui séparent ces différents groupes.

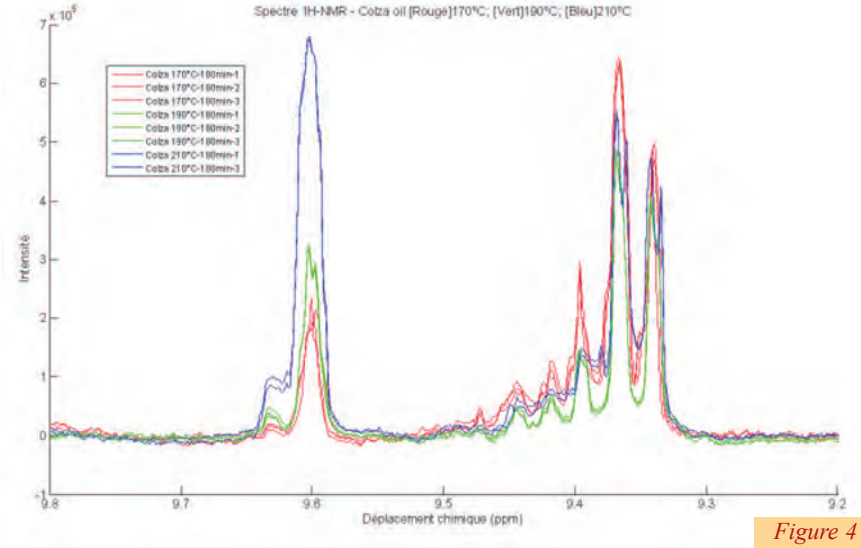


Figure 4

Cette représentation met parfois en évidence un lien particulier entre la composante principale et une des conditions analytiques. Tel est le cas de l'axe CP2 corrélé avec la température ainsi que l'illustre la figure 5a. On y constate aisément que les coordonnées des échantillons sur cet axe augmentent avec la valeur de la température de chauffage.

Il est vrai que l'on s'attendrait à avoir un ordre d'arrangement logique

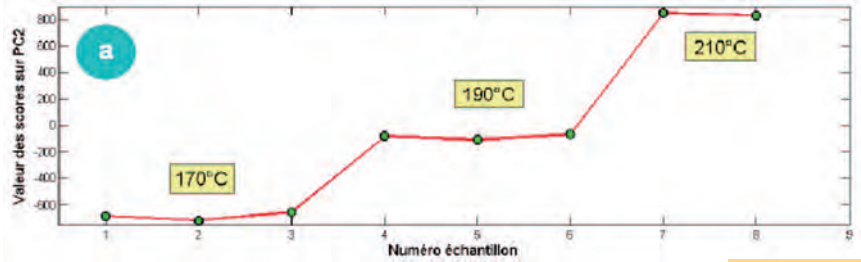


Figure 5a

en fonction de la température sur CP1, car, dans le calcul de l'ACP, cette composante présente toujours le pourcentage de variance le plus élevé. Or, on voit ici que ce n'est pas le cas. Les échantillons chauffés à 170 °C se trouvent situés entre ceux chauffés respectivement à 190 °C (à droite) et à 210 °C (à gauche) le long de CP1. Ceci est dû à la présence des pics entre 9,3 et 9,5 ppm dont l'évolution ne suit pas l'augmentation de la température. Ces pics ont un poids dans le calcul de l'ACP plus important que les pics situés entre 9,5 et 9,7 ppm. C'est la raison pour laquelle on observe leur évolution sur CP1, et l'évolution des pics situés vers 9,5 et 9,7 se retrouve quant à elle sur la seconde composante (CP2).

Une interprétation complète des résultats de l'ACP implique toujours une lecture de l'importance des variables ou "contributions factorielles" dans la construction des composantes principales. Ces contributions factorielles donnent lieu à un graphe appelé *loadings* (figure 5b). Il précise quelles sont les sources de variation qui caractérisent le plus fidèlement les échantillons. Dans le cas de données spectrales, les loadings montrent, par exemple, quelles sont les bandes d'absorption ou les pics de RMN (intensité et déplacements chimiques) qui évoluent le plus pour un échantillon ou un groupe d'échantillons donné.

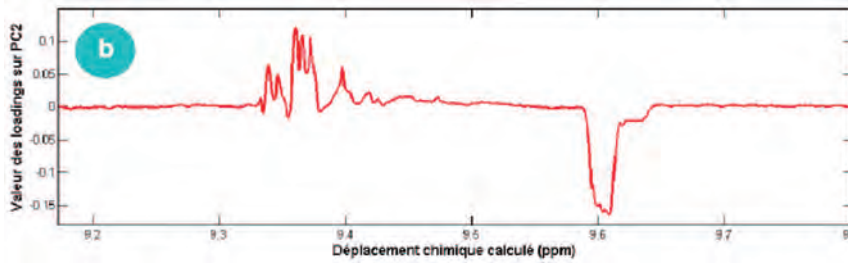


Figure 5b

La figure 5b indique que les pics situés vers 9,6 ppm ont une contribution négative sur la composante CP1 des *loadings*. Tandis que les pics situés entre 9,3 et 9,4 ppm ont des valeurs positives sur ce même axe. Pour comprendre ce que cela signifie, il faut regarder conjointement les scores sur la même composante (figure 5a). On comprend rapidement que les échantillons chauffés à 170 °C sont caractérisés principalement (rarement à 100%) par l'évolution du massif aldéhydique vers 9,6 ppm car le signe des *scores* est le même que celui des *loadings* indiquant une évolution simultanée et dans le même sens. De la même manière, les échantillons chauffés à 210 °C sont fortement corrélés avec l'évolution des pics situés entre 9,3 et 9,5 ppm.

Les *scores* sur PC2 indiquent un comportement intermédiaire des échantillons chauffés à 190 °C. Pour ces derniers, il est plus difficile de connaître les contributions de chaque massif de pics aldéhydiques sur le spectre RMN car l'évolution simultanée des deux types de protons intervient en proportions équivalentes sur CP2 dans la caractérisation de ces échantillons.

En conclusion

L' application de l'ACP à des données issues de l'analyse d'huiles alimentaires par résonance magnétique nucléaire du proton permet une interprétation plus facile et plus riche de l'impact du procédé thermique utilisé qu'une simple intégration des pics RMN 1H enregistrés. Sa capacité à traiter des données multivariées telles que des spectres entiers fait de l'ACP une technique puissante et universelle aboutissant à des sorties graphiques d'une grande simplicité.

De nombreuses techniques s'appuient sur l'ACP, souvent comme étape préliminaire visant à compresser les données. Elle est souvent un point de départ dans l'analyse des données et constitue un outil exploratoire efficace.

C. C.

Pour en savoir (un peu) plus

- **Beebe K.R., Pell R.J., Seasholtz M.B.**, *Chemometrics: a Practical Guide*, Wiley, 1998.

- **Gemperline P**., *Practical Guide to Chemometrics*, 2nd ed., CRC/Taylor & Francis, Boca Raton, 2006.

- **Bertrand D., Dufour E.**, *La spectroscopie infrarouge et ses applications analytiques*, 2e ed., Tec & Doc, 2005.

- **Goupy J.**, *La chimiométrie,* Annales des falsifications, de l'Expertise Chimique et Toxicologiques, 2009, 971, p. 41.