L'analyse en composantes principales

Une des techniques fondatrices de la chimiométrie

Christophe Cordella

Résumé La chimiométrie est une discipline impliquant initialement analyse de données et chimie analytique. Elle rassemble et développe un ensemble d'outils mathématiques utilisés pour extraire de l'information à partir des données chimiques structurées et interprétables. De nombreuses définitions de la chimiométrie ont été proposées, mais elles ont toutes l'objectif commun de rechercher de nouveaux outils et de nouvelles façons d'exploiter les informations contenues dans les données pour créer de la connaissance. Certains de ces outils sont aujourd'hui de plus en plus appliqués à des données métabolomiques. Ils sont conçus pour montrer ce qu'il serait impossible de voir avec une analyse univariée des données. Le présent article vise à présenter l'une des techniques de base de la chimiométrie : l'analyse en composantes principales (ACP). Après une introduction historique, les principes de base de la technique sont détaillés. Puis un exemple concret d'utilisation de l'ACP est développé. L'interprétation des résultats est présentée avec un souci pédagogique, afin de mieux montrer la puissance de l'outil. Les données analysées dans ce document concernent une étude cinétique de l'oxydation thermique de l'huile comestible suivie par spectroscopie RMN du proton.

Mots-clés **Analyse en composantes principales, ACP, chimiométrie, analyse de données, oxydation de l'huile, Wold, Kowalski.**

Abstract **Principal component analysis: the basic building block of chemometrics**
The chemometrics is a discipline involving analysis of data and analytical chemistry. It brings together and develops a set of mathematical tools used to extract information from structured and interpretable chemical data. Many definitions of chemometrics have been proposed, but all have common goal such as finding new tools and new ways to exploit information contained in the data to create knowledge. Some of these tools are more and more applied today to metabolomic and/or metabonomic data. They are designed to show what it would be impossible to do with a univariate data analysis. This contribution aims to present one of the basic techniques of the chemometrics: principal component analysis (PCA). After a historical introduction and the basic principles of the technique, a practical example of use of PCA is developed. The interpretation of results is shown with a pedagogical concern, to better show the power of the tool. The data analyzed in this paper concern a kinetic study of the thermal oxidation of edible oil monitored by proton NMR spectroscopy.

Keywords **Principal component analysis, PCA, chemometrics, data analysis, oil oxidation, Wold, Kowalski.**

Définitions et historique de la chimiométrie

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. De nombreuses définitions de la chimiométrie ont été proposées. J. Workman en a répertorié plus d'une dizaine [1], mais toutes présentent comme objectifs communs de trouver de nouveaux outils et de nouvelles méthodes pour exploiter l'information contenue dans les données en vue de créer de la connaissance. Une partie de ces outils s'appliquent aujourd'hui de plus en plus à des données métabolomiques et/ou métabonomiques(1). Ils sont conçus pour montrer ce qu'il serait impossible de voir avec un traitement univarié des données(2). La chimiométrie est fondée sur la construction à l'aide d'outils statistiques, puis l'exploitation d'un modèle de comportement applicable à des systèmes complexes et donc souvent multivariables.

Le terme « chimiométrie » vient de l'anglais « chemometrics » et a été inventé il y a près de 40 ans, en 1971, par Svante Wold. Il semble apparaître pour la première fois dans la littérature en 1975 [2]. Ce dernier avec Bruce Kowalski comprirent très tôt l'importance de cette nouvelle discipline et créèrent en 1974 la Société Internationale de Chimiométrie (« International Chemometrics Society ») [1, 3]. Wold [4] et Kowalski [5] sont considérés comme les pères de la chimiométrie. Par la suite, d'autres contributeurs de renom firent leur apparition sur la scène internationale autant pour la qualité scientifique de leurs travaux que pour leur pédagogie. Les citer tous est impossible, cependant profitons de l'occasion qui nous ait donnée pour rendre un hommage au professeur D.L. Massart(3). Ses contributions au développement de la chimiométrie sont nombreuses. Qui n'a pas eu en main l'un de ses ouvrages de chimiométrie tels que le *Handbook of*

Chemometrics and Qualimetrics ou autres *Chemometrics – Tutorials* ? [6].

Plusieurs outils très répandus de la chimiométrie tels que l'analyse en composantes principales (ACP) tirent leurs origines du domaine de la biométrie avec les travaux de Karl Pearson[4] en 1901 et de l'économie avec ceux de Harold Hotelling[5]. 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. Aussi, on lui préfère souvent le terme moins restrictif par rapport à la chimie d'analyse multivariée (ou « multivariate analysis »), voire de reconnaissance de forme au sens large (« pattern recognition »). En instrumentation, la chimiométrie consiste à **modéliser** les variations d'un certain nombre de variables dites « *Y_expliquées* » dont l'obtention est délicate (nécessitant une analyse chimique par exemple) en fonction d'autres variables nommées « *X_explicatives* » mesurables « plus facilement » (mesure de capteurs physiques par exemple, spectres enregistrés sur divers spectroscopes ou spectromètres) 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* »)…). L'analyse en composantes principales se range dans cette catégorie d'outils. La classification peut suivre l'étape de description si l'objectif du chimiométricien est de former des groupes d'objets. Cette catégorie est généralement celle des outils dits de « classification et de reconnaissance de forme »[6]. Plus généralement, il est d'usage de ranger les outils de description et de classification des données selon qu'ils emploient des méthodes faisant appel à une information connue au départ du processus du traitement (méthodes dites supervisées) ou non (méthodes dites non supervisées) [7].
- L'**étalonnage**[7] 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 « *X_explicatives* » sont mesurées, les autres, Y, étant calculées à l'aide du modèle.

Il faut ajouter une étape de **validation des modèles** calculés lors de l'étalonnage qui consiste à utiliser certaines techniques de sélection des échantillons (validation croisée, « leave-one-out », « bootstrap ») dont le but est de prélever un sous-ensemble des échantillons analysés (procédure nommée « échantillonnage » en statistique), afin de les utiliser comme échantillons-tests pour mesurer la fiabilité et la précision du ou des modèles construits.

Les pages qui suivent présentent l'une des méthodes descriptives d'exploration des données les plus utilisées en chimiométrie : *l'analyse en composantes principales (ACP)*. Après un rappel historique et des fondements théoriques de l'outil, nous présentons, à titre d'exemple, une application de l'ACP en vue de visualiser l'effet du chauffage sur la stabilité thermique d'huiles alimentaires au cours du temps. L'objectif de cet article n'étant pas de faire le point sur l'état de l'art de la chimiométrie, vous trouverez quelques éléments généraux récents de bibliographie en fin d'article.

L'analyse en composantes principales (ACP)

Définition de l'outil, historique et objectifs

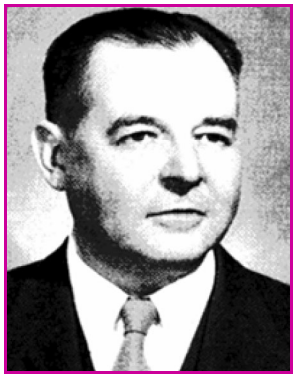

Figure 1 - Harold Hotelling (1895-1973).

L'analyse en composantes principales fait partie du groupe des méthodes descriptives multidimensionnelles appelées méthodes factorielles (descriptives, non supervisées). Ces méthodes sont apparues au début du XXe siècle, dans les années 30, et ont été surtout développées en France dans les années 60, en particulier par Jean-Paul Benzécri[8] qui a beaucoup exploité les aspects géométriques et les représentations graphiques. Elle est aussi connue sous le nom de transformée de Karhunen-Loève ou de transformée de Hotelling (en l'honneur d'Harold Hotelling, *figure 1*). 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 organisés sous forme d'un tableau rectangulaire de données comportant les valeurs de ***p*** variables quantitatives pour ***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. Par exemple, certaines variables peuvent croître ou décroître simultanément ou bien l'une peut croître pendant que l'autre décroît. Ainsi, on cherchera si l'on peut distinguer des groupes dans l'ensemble des objets en regardant quels sont les objets qui se ressemblent et ceux qui se distinguent des autres. 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 ***n*** nouveaux axes orthogonaux nommés « composantes principales » ou « CP » qui sont des combinaisons linéaires des variables initiales (voir *encadré*). L'ACP est très utile lorsqu'il s'agit de compresser l'information contenue dans un grand nombre de variables, ***N***, car les ***n*** premiers axes (***n*** << ***N***) contiennent souvent la plus grande partie de la variabilité et on l'espère de l'information intéressante. On trouve souvent l'expression « *faire le meilleur choix du point de vue de l'inertie expliquée* ». Le terme « inertie » correspond ici à « variabilité » des données et renvoie à des critères statistiques de dispersion tels que la variance et l'écart-type d'une série de données.

L'ACP : matrices et calcul matriciel

L'analyse en composantes principales est réalisée par application d'un des nombreux algorithmes disponibles. Tous ont comme objectif de créer un ensemble de **combinaisons linéaires des variables d'origine** de telle manière que la matrice initiale est décomposée en deux nouvelles matrices, l'une nommée matrice des *scores* (***T***) et l'autre nommée matrice des *loadings* (***P***) selon :

$$ {}_nX^p = {}_nT^q \cdot {}_p[P^t]^q + {}_nE^p $$

$$ \begin{bmatrix} \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & X & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \end{bmatrix} = \begin{bmatrix} \cdot & \cdot \\ \cdot & T \\ \cdot & \cdot \\ \cdot & \cdot \end{bmatrix} \begin{bmatrix} \cdot & \cdot & \cdot & \cdot \\ \cdot & P^t & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \end{bmatrix} + \begin{bmatrix} \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & E & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \end{bmatrix} $$

où ***X*** est la matrice de données originales constituée de ***n*** lignes (les objets) et ***p*** colonnes (les variables) ; ***T*** est la matrice des scores avec ***n*** lignes et ***q*** colonnes (nombre de composantes principales) ; ***P*** est la matrice des loadings avec ***q*** colonnes et ***p*** lignes ; ***E*** est la matrice des résidus (ou « erreurs » ou « écart » entre les données d'origine et leurs projections) constituée de ***n*** x ***p*** éléments ; enfin $\boldsymbol{P^t}$ est la matrice transposée de ***P***.
À noter que si ***q*** est égal au *rang* de la matrice ***X*** (c'est-à-dire au nombre de sources différentes de variabilité – habituellement le plus petit de ***n*** et de ***p***), alors ***E*** est égal à zéro.
Ce formalisme doit être compris de la manière suivante : la projection des éléments de la matrice ***X*** dans l'espace de dimension ***q*** au moyen de la projection de la matrice transposée $\boldsymbol{P^t}$ donne les coordonnées de ces mêmes éléments dans le plan ***T***. Les colonnes de la matrice ***T*** correspondent aux vecteurs scores (les objets) et les lignes dans la matrice $\boldsymbol{P^t}$ correspondent aux vecteurs loadings (les variables) (les vecteurs loadings sont les colonnes de ***P***, donc les lignes de $\boldsymbol{P^t}$). Tous ces vecteurs sont orthogonaux de telle sorte qu'ils respectent les conditions précisées dans l'équation :

$$ p_i^t p_j = 0 \quad \text{et} \quad t_i^t t_j = 0 \quad \text{pour} \quad i \neq j $$

Cette orthogonalité est un moyen pour garantir la non-redondance (du moins *a minima*) de l'information portée par chaque composante principale calculée.

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 formuler des hypothèses qu'il faudra étudier à l'aide de modèles et d'études statistiques inférentielles. On pourra aussi, *a posteriori*, se servir des représentations fournies par l'ACP pour illustrer certains résultats dans un but pédagogique. Dans notre cas, nous avons utilisé l'ACP afin de comparer trois types d'huiles (colza, tournesol, olive) chauffées selon un protocole consistant à prélever toujours la même quantité d'huile et à la chauffer dans une étuve thermostatée à 170, 190 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. Parmi les résultats obtenus, 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. Puis nous présenterons 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). Le premier cas met en évidence l'effet du chauffage sur les modifications chimiques au sein de l'huile au cours du temps et le second reflète les différences chimiques engendrées par une augmentation de la température de chauffage.

Étude des huiles à une température fixe et en fonction du temps de chauffage par RMN du proton

L'effet du temps de chauffage à température fixe (190 °C) est illustré par la *figure 2*. On y observe 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 correspondant au déplacement chimique des aldéhydes. On constate qu'au cours du temps apparait une quantité d'aldéhyde plus ou moins importante et plus ou moins rapide selon la nature de l'huile. Ces aldéhydes proviennent des

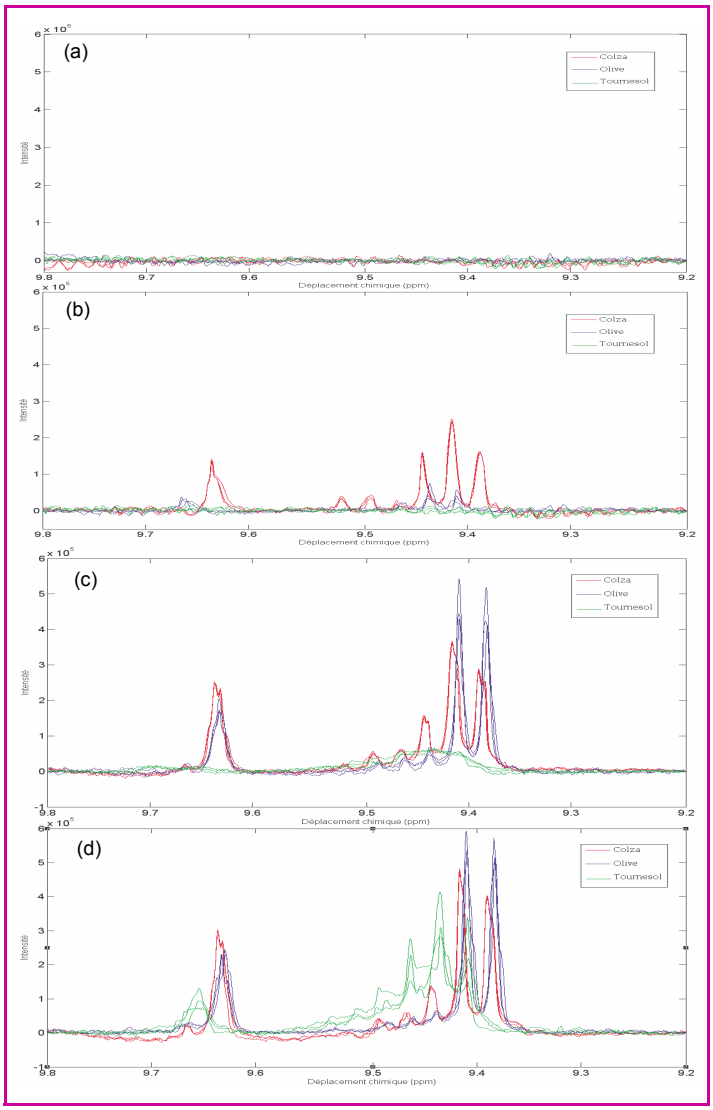

Figure 2 - Spectres RMN du proton des huiles (colza en rouge, olive en bleu, tournesol en vert) : froides (a), chauffées 30 min à 190 °C (b), 120 min à 190 °C (c), 180 min à 190 °C (d).

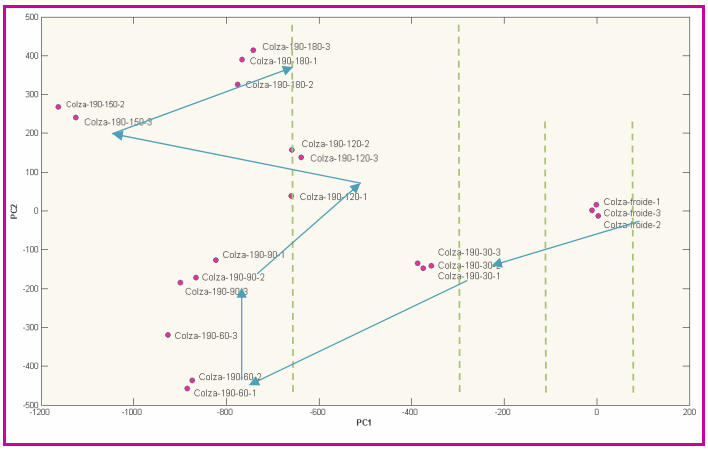

Figure 3 - Carte factorielle (plan CP1 x CP2) des échantillons formée par les deux premières composantes principales.
Huile : colza ; chauffage : 190 °C ; temps = 30 à 180 min.

réactions d'oxydation des acides gras insaturés naturels [8]. Les résultats de l'analyse en composantes principales sur ces spectres sont présentés *figure 3*. L'interprétation des cartes factorielles se fait de la même façon quelle que soit l'huile. Nous présentons ici l'interprétation de celles de l'huile de colza. Ces figures permettent de visualiser l'effet du temps de chauffage sur la composition chimique des huiles. On constate que cet effet diffère d'une huile à l'autre (cartes non présentées) car la répartition des points représentant les échantillons n'est pas identique d'une carte factorielle à l'autre. On peut visualiser sur le plan formé par les axes CP1 et CP2 (*figure 3*), le « chemin thermique » suivi par l'huile au cours du chauffage. 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[9] [9], soit d'un point de vue cinétique.

Une réaction se produit lorsqu'elle atteint l'énergie d'activation nécessaire à la formation des produits de la réaction ; cette barrière énergétique est franchie grâce à la chaleur appliquée au système. Des considérations de géométrie moléculaire entrent également en jeu et influencent la cinétique de la réaction, c'est ce qu'exprime le facteur pré-exponentiel de la loi d'Arrhenius, $k = Ae^{(-Ea/RT)}$. Sur la *figure 3*, nous avons donc le résultat des réactions chimiques engendrées à 190 °C en fonction de la durée du chauffage. 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. Ainsi, on voit que les échantillons d'huile froide (au centre à droite) se distinguent bien des autres échantillons d'huile chauffés et présentent des similitudes avec les échantillons chauffés pendant 30 min. Ce qui, *a priori*, semble normal car après 30 min, les transformations au sein de l'huile ont seulement commencé. En revanche, on voit très bien que le groupe d'échantillons chauffés pendant 150 min est à l'opposé de ceux de l'huile froide (les plus à gauche sur la carte). On note également que les différences entre les échantillons d'huile non chauffés et ceux chauffés 30, 60, 120 et 150 min se font surtout sur l'axe horizontal, c'est-à-dire sur la première composante principale (CP1), tandis que l'axe vertical (CP2) permet la séparation des échantillons chauffés à 60, 90, 120 et 180 min. Il y a donc deux types de variabilité dans les données initiales qui sont exprimées dans les deux premières composantes principales (CP1 et CP2). On peut interpréter ce fait en avançant qu'il y a deux types de protons aldéhydiques concernés car les portions de spectres RMN 1H utilisés ici correspondent à la zone comprise entre 9 et 9,8 ppm : la *figure 2* montre l'évolution de cette zone de déplacement chimique au cours du chauffage. On y voit parfaitement l'apparition de deux pics correspondant à des molécules ayant un groupement aldéhydique dont l'environnement chimique diffère. Le premier groupe de protons possède un déplacement chimique d'environ 9,4 ppm et l'autre d'environ 9,6 ppm.

En résumé, nous visualisons grâce à l'ACP qu'à 190 °C, les échantillons se rangent en cinq groupes chimiques (voir *tableau*).

Classement des échantillons.

Groupe	Échantillons	Durée de chauffage (min)	Composante(s) principale(s) utile(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

Étude des huiles à durée de chauffage fixe et en fonction de la température

Nous montrons ici le résultat obtenu sur l'huile de colza pour une durée de chauffage de 180 min aux trois températures appliquées (170, 190 et 210 °C). L'analyse en composantes principales de la matrice contenant ces échantillons met en évidence l'effet de la température de chauffage sur l'huile de colza. Comme précédemment, la représentation des échantillons sous forme de cartes factorielles permet de visualiser les transformations chimiques dont l'huile a été le siège pendant le chauffage (*figure 4*).

Un résultat similaire est obtenu pour les autres huiles. Ce type de représentation permet de synthétiser 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 est relativement difficile comme le montre la *figure 5* où l'on constate en effet que le chauffage provoque d'importantes modifications chimiques avec comme conséquence l'apparition d'un ou plusieurs composés aldéhydiques dont l'aire des pics varie notablement. Il est difficile de se rendre compte de l'ampleur des transformations à partir des spectres. Les cartes factorielles produites par l'analyse en composantes principales permettent d'évaluer l'importance des changements opérés dans les échantillons en montrant les

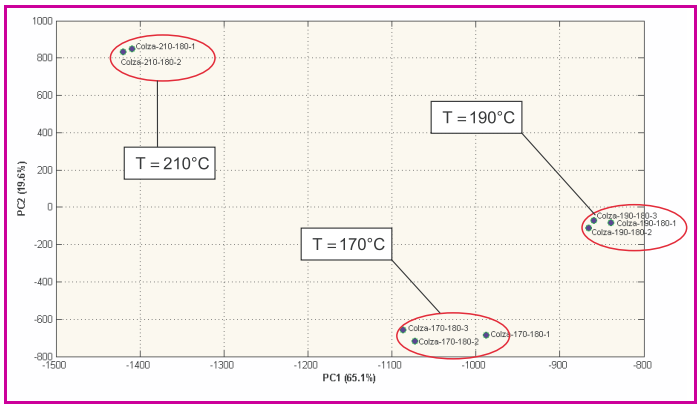

Figure 4 - Scores (CP1 *vs* CP2) pour l'huile de colza chauffée à 170, 190 et 210 °C pendant 180 min.

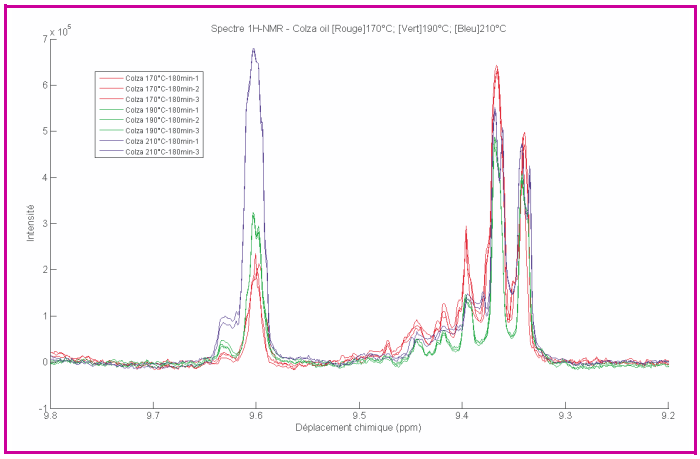

Figure 5 - Spectres des échantillons d'huile de colza chauffés à 170 (rouge), 190 (vert) et 210 °C (bleu) pendant 180 min.

distances qui séparent les différents groupes d'échantillons. On peut utiliser, dans ce cadre là, le centre géométrique des groupes pour calculer la distance qui sépare les groupes et comparer ces distances (calcul non présenté ici).

Enfin, il est souvent utile de représenter la valeur des coordonnées factorielles de chaque échantillon sur une composante principale donnée en fonction du numéro de l'échantillon dans la matrice : cette valeur est nommée « *scores* ». Cette représentation peut parfois mettre en évidence un lien particulier entre la composante principale et une des conditions analytiques. La *figure 6a* montre que l'axe CP2 est corrélé avec la température car les valeurs des coordonnées des échantillons sur cet axe augmentent avec la valeur de la température de chauffage.

On s'attendrait à avoir un ordre d'arrangement logique en fonction de la température sur CP1, car c'est cette composante qui présente toujours le pourcentage de variance le plus élevé. Or, on constate 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).

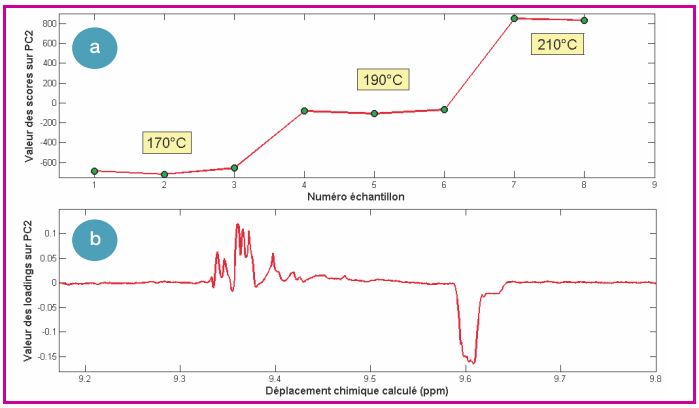

Figure 6 - Représentation des valeurs des loadings (a) et des scores (b) des échantillons d'huile de colza sur CP2.

De façon complémentaire, il est d'usage de représenter l'importance des variables dans la construction des composantes principales. Ce graphe est nommé « *contributions factorielles* » ou « *loadings* » (*figure 6b*). Il précise quelles sont les sources de variation qui caractérisent le mieux 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. La *figure 6b* 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 6a*). 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 un comportement similaire. 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.

Conclusion

Notre but ici était d'évoquer par l'exemple un outil fondamental de la chimiométrie : l'analyse en composantes principales (ACP). Son application 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. L'ACP montre graphiquement les différences et/ou les ressemblances entre des échantillons ayant été soumis à des conditions d'analyse choisies (température, pression, temps, quantité de réactifs, procédés divers…). Les échantillons ayant des caractéristiques semblables se regroupent pour former des agrégats (nommés clusters) plus

ou moins distincts. Les échantillons possédant des caractéristiques chimiques éloignées se positionnent dans des clusters éloignés les uns des autres. L'examen des scores et des loadings nous renseigne sur l'origine chimique des variations présentes dans les données enregistrées et nous indique quelles sont les variables qui contribuent le plus dans la distinction de tel ou tel groupe d'échantillons. 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.

Quelques éléments de bibliographie récents pour en savoir 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.
- Brereton R.G., *Exploratory Data Analysis*, John Wiley & Sons, **2009**.
- Lavine B., Workman J., Chemometrics, *Analytical Chemistry*, **2010**, *82*, p. 4699.
- Miller J.N., Miller J.C., *Statistics and Chemometrics for Analytical Chemistry*, 6th ed., Pearson Prentice Hall, Harlow (England), New York, **2010**.

Notes et références

(1) La *métabolomique* ou *métabonomique* sont rencontrées indifféremment dans la bibliographie mais ont une signification voisine. Il s'agit de l'étude de l'ensemble des métabolites (sucres, acides aminés, acides gras, etc.) présents dans une cellule, un organe, un organisme. Au même titre que la génomique pour l'ADN, la métabolomique utilise la spectrométrie de masse et la résonance magnétique nucléaire pour suivre l'évolution globale de certains éléments clés de la physiologie des organismes. Le but recherché est la mesure globale des perturbations biochimiques causées par les maladies, les médicaments ou les produits toxiques. Une étude métabonomique ou métabolomique se déroule en trois étapes : le prélèvement de biofluides ou de biopsies contenant les marqueurs métaboliques et leur dosage, l'extraction de ces données métaboliques et enfin l'analyse des jeux de données. La dernière étape emploie des outils statistiques tels que l'analyse en composantes principales, les réseaux de neurones, en vue essentiellement d'isoler un petit nombre significatif de paramètres parmi ceux mis en évidence dans l'étude expérimentale.

(2) *Univarié*, dans le sens qui ne se préoccupe que d'une source de variation à la fois, c'est-à-dire en termes statistiques, d'une seule variable.

(3) Professeur de chimie analytique à l'Institut de Pharmacie de l'Université de Bruxelles, puis professeur de chimiométrie à l'Université Catholique de Louvain (Belgique). Voir www.vub.ac.be/fabi/DLM/ (visité le 23 août 2010).

(4) Célèbre par ses contributions à la fondation des statistiques modernes (test du $\chi 2$, régression linéaire, analyse en composantes principales), Karl Pearson (1857-1936), mathématicien anglais, enseigna les mathématiques à l'Université de Londres de 1884 à 1933 et fut l'éditeur de *The Annals of Eugenics* (de 1925 à 1936). En bon disciple de Francis Galton (1822-1911), Pearson poursuit les travaux statistiques de son mentor, ce qui l'amènera à jeter les bases du calcul des corrélations (à la base de l'analyse en composantes principales) et marquera les débuts d'une nouvelle science : la biométrie.

(5) Harold Hotelling : économiste et statisticien américain (1895-1973). Professeur d'économie à l'Université de Columbia, il est à l'origine de contributions importantes en statistiques dans la première moitié du XXe siècle, telles que le calcul de variation, les fonctions de production fondées sur la maximisation des profits, l'utilisation de la distribution de Student pour la validation d'hypothèses qui mènera au calcul de l'intervalle de confiance.

(6) Le terme « pattern recognition » est généralement employé dans la littérature.

(7) Attention, ce terme se traduit en anglais par « calibration » et ne doit pas être confondu avec le « calibrage » d'un instrument qui est le recalage à l'aide d'un ou deux mesurages (« gauging » en anglais).

(8) Statisticien français, ancien élève de l'École normale supérieure, professeur à l'Institut de Statistique de l'Université de Paris, fondateur de l'école française d'analyse des données dans les années 1960-1990, Jean-Paul Benzécri a développé des outils statistiques, notamment l'analyse factorielle des correspondances qui permet de traiter de grandes quantités de données afin de visualiser et hiérarchiser l'information.

(9) Par exemple, la réaction de cycloaddition de Diels-Alder entre un diène et un diénophile fait partie des réactions péricycliques qui peuvent se produire à 1 bar et à température ambiante ou après un léger chauffage ; il peut donc y avoir apparition de composés chimiques issus d'une telle réaction dès que l'on chauffe l'échantillon pourvu que les conditions d'espèces moléculaires mises en présence soient remplies : les recouvrements orbitalaires des systèmes π mis en jeu dans ces réactions demandent un faible apport énergétique et sont favorisés par des considérations de géométrie d'orbitales et de densité électronique. En revanche, à température ambiante, les réactions nommées « transpositions sigmatropiques » se font difficilement et nécessitent un chauffage plus élevé car elles impliquent la destruction et la formation de plusieurs liaisons covalentes et π dans la molécule.

[1] Workman J., The state of multivariate thinking for scientists in industry: 1980-2000, *Chemometrics and Intelligent Laboratory Systems*, **2002**, *60*, p. 13.

[2] Kowalski B.R., Chemometrics – Views and propositions, *Journal of Chemical Information and Computer Sciences*, **1975**, *15*, p. 201.

[3] Gemperline P., Future trends in chemometrics, *Practical Guide To Chemometrics*, 2nd ed., CRC Press, **2006**.

[4] Wold S., Chemometrics, why, what and where to next?, *Journal of Pharmaceutical and Biomedical Analysis*, **1991**, *9*, p. 589 ; Wold S., Chemometrics; What do we mean with it, and what do we want from it?, *Chemometrics and Intelligent Laboratory Systems*, **1995**, *30*, p. 109 ; Wold S., Sjostrom M., Chemometrics, present and future success, *Chemometrics and Intelligent Laboratory Systems*, **1998**, *44*, p. 3 ; Wold S., Trygg J., Berglund A., Antti H., Some recent developments in PLS modeling, *Chemometrics and Intelligent Laboratory Systems*, **2001**, *58*, p. 131 ; Wold S., Berglund A., Kettaneh N., New and old trends in chemometrics. How to deal with the increasing data volumes in R&D&P (research, development and production) – with examples from pharmaceutical research and process modeling, *Journal of Chemometrics*, **2002**, *16*, p. 377 ; Wold S., Analysis of similarities and dissimilarities between chromatographic liquid-phases by means of pattern cognition, *Journal of Chromatographic Science*, **1975**, *13*, p. 525 ; Wold S., Pattern-recognition by means of disjoint principal components models, *Pattern Recognition*, **1976**, *8*, p. 127 ; Wold S., Enius G., Mathematics in service of man, *Chemtech*, **1978**, *8*, p. 770.

[5] Kowalski B.R., Proceedings: Pattern recognition, *CancerChemother Rep 2*, **1974**, *4*, p. 43 ; Kowalski B.R., Chemometrics, *Chemistry & Industry*, **1978**, p. 882 ; Kowalski B.R., Chemometrics, *Analytical Chemistry*, **1980**, *52*, p. R112 ; Kowalski B.R., Bender C.F., Solving chemical problems with pattern recognition, *Naturwissenschaften*, **1975**, *62*, p. 10.

[6] Massart D.L., *Handbook of Chemometrics and Qualimetrics*, Elsevier, **1997** ; Massart D.L., Brereton R.G., Dessy R.E., Hopke P.K., Spiegelman C.H., Wegscheider W., *Chemometrics - Tutorials*, Elsevier, **1990**.

[7] 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**.

[8] Frankel E.N., Lipid oxidation, *Prog. Lipid Res.*, **1980**, *19*, p. 1 ; Frankel E.N., Volatile lipid oxidation products, *Prog. Lipid Res.*, **1983**, *22*, p. 1 ; Frankel E.N., Lipid oxidation - Mechanisms, products and biological significance, *Journal of the American Oil Chemists Society*, **1984**, *61*, p. 1908 ; Frankel E.N., Secondary products of lipid oxidation, *Chem. Phys. Lipids*, **1987**, *44*, p. 73 ; Frankel E.N., Biological significance of secondary lipid oxidation products, *Free Radic. Res. Commun.*, **1987**, *3*, p. 213.

[9] Fosset B., Baudin J.-B., Lahitète F., *Chimie – Exercices et problèmes corrigés, 2e année PC-PC**, Dunod, Collection J'intègre, **2006**.

Christophe Cordella
est ingénieur de recherche INRA à l'Institut des sciences et industries du vivant et de l'environnement, INA-PG AgroParisTech*.

* INRA/INA-PG AgroParisTech, Institut des sciences et industries du vivant et de l'environnement, UMR 1145 GENIAL, Équipe Ingénierie Analytique pour la Qualité des Aliments, 16 rue Claude Bernard, F-75231 Paris Cedex 05.
Courriel : christophe.cordella@agroparistech.fr
www.chimiometrie.fr

Chimiométrie 2010 : 2-3 décembre 2010, Paris
www.chimiometrie.fr/chemom2010