La chimiométrie appliquée à la spectroscopie

Sylvère André[1], Ludovic Duponchel[1],
Jordane Lallemand[2] et Sylvie Roussel[2]

[1] LASIR, Université de Lille-CNRS UMR 8516, F-59655 Villeneuve d'Ascq, France
[2] ONDALYS, 4 rue Georges Besse, F-34830 Clapiers, France

1 Introduction

La chimiométrie, ou l'analyse de données multivariées, est la science qui applique des méthodes mathématiques et statistiques optimales pour traiter des données issues de mesures au sens large. La chimiométrie comprend la conception d'expériences en amont et l'analyse des données afin d'obtenir des informations pertinentes à partir des mesures effectuées [1,2,3]. Le besoin d'outils chimiométriques puissants provient principalement du développement de l'instrumentation qui offre de grandes quantités de données de plus en plus complexes et non spécifiques.

Le domaine scientifique de la chimiométrie se compose d'une grande variété de méthodes mathématiques, visant à atteindre divers objectifs. Le schéma ci-dessous est une synthèse des diverses approches chimiométriques que tout scientifique devrait suivre face à une problématique d'analyse de données multivariées (Fig. 1).

Même si les principes de chimiométrie sont basés sur les mathématiques et les statistiques, une connaissance approfondie de ces disciplines n'est pas nécessaire pour analyser des données multidimensionnelles. Cependant, la connaissance de l'application, des échantillons et des méthodes de mesure ainsi que du bon sens sont nécessaires pour paramétrer et analyser les résultats des progiciels de chimiométrie et éviter les pièges et les erreurs d'interprétation.

Ce chapitre décrit les méthodes de chimiométrie les plus couramment utilisées et se concentrera principalement sur la compréhension, l'interprétation et l'évaluation de la validité des résultats.

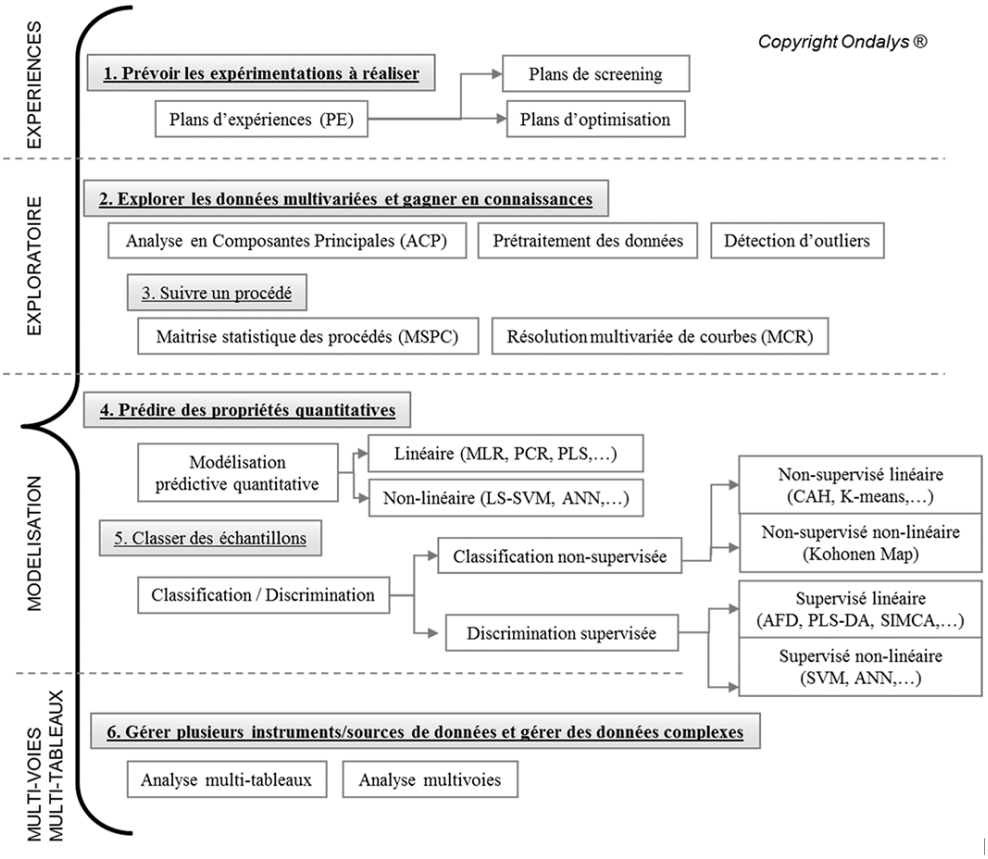


FIGURE 1 – L'approche d'analyse des données multivariées : classification des méthodes chimiométriques.

2 Prétraitements

Les techniques de prétraitement des données sont souvent utilisées avant la modélisation, afin de réduire le bruit et les perturbations indésirables dans le signal. Plusieurs méthodes de prétraitement ont été initialement mises au point en spectroscopie proche infrarouge, en raison de la sensibilité de la mesure aux variations physiques dans les échantillons (granulométrie par exemple), à l'environnement de mesure (variations de température, humidité…) ou encore à la chaine de mesure (changement de spectromètre, de type de préparation d'échantillons…). Néanmoins ces méthodes sont maintenant largement utilisées dans tous les domaines de la spectroscopie vibrationnelle. La technique la plus appropriée de prétraitement dépendra des perturbations à corriger ; ces prétraitements doivent être comparés, voire associés, pour trouver la combinaison optimale pour un jeu de données.

Published in : SIMON G. (dir.) (2020), *Spectroscopies vibrationnelles. Théorie, aspects pratiques et applications*, Editions des archives contemporaines, ISBN : 9782813002556, doi : https ://doi.org/10.17184/eac.9782813002556

2.1 Méthodes de prétraitement par variable (longueur d'onde)

Les méthodes de « prétraitement » les plus largement utilisées consistent à faire une mise à l'échelle des données. Elles peuvent être utilisées pour tous les types de données multivariées : continues, discrètes, spectroscopiques ou pour des variables de procédé.

- **Centrer** chaque variable est le prétraitement le plus commun. Le principe consiste à soustraire à la valeur d'une variable (longueur d'onde) la moyenne de cette variable dans le tableau de données à analyser. Ce centrage des données est quasi-systématique dans les méthodes de projection telles que l'Analyse en Composantes Principales (ACP) ou la régression des moindres carrés partiels (*Partial Least-Squares Regression* – PLS-R). Il est utilisé pour centrer le sous-espace au barycentre de l'ensemble de données d'origine, pour une meilleure visualisation des données en ACP (voir 3.2). Lors de la construction d'un modèle prédictif, le centrage des données a pour effet de rendre le terme constant de l'équation différent de zéro.

- **Normaliser (ou mettre à l'échelle)** est utilisé pour rendre les différentes variables comparables lorsqu'elles sont incluses dans une même analyse multivariée. La technique de normalisation la plus courante divise chaque variable par son écart-type dans le tableau de données. Après cette normalisation, toutes les variables présentent le même écart-type, qui est donc égal à 1. Le prétraitement doit être appliqué systématiquement aux jeux de données contenant des variables d'échelles différentes et donc de variances différentes (par exemple, pH, température et pression) afin de leur donner un poids égal dans le traitement de données ultérieur. En règle générale, la normalisation ne doit pas être appliquée à des données spectroscopiques parce que chaque absorbance d'un spectre est exprimée dans la même unité ; donc, les variations d'intensité des longueurs d'onde constituent les informations importantes. Par exemple, des pics spectraux montreront plus de variabilité qu'une zone de ligne de base où seul le bruit est présent. Il est à noter que d'autres types de normalisation sont possibles, par exemple la normalisation de Poisson qui divise par la racine carrée de l'écart-type (utilisée en spectroscopie SIMS par exemple).

2.2 Méthodes de prétraitements spectroscopiques

Les méthodes de prétraitements spectroscopiques visent à corriger l'influence de différentes perturbations et/ou à mettre en valeur l'information contenue dans les spectres avant l'étape de modélisation.

Dans les données spectroscopiques, des perturbations peuvent être additives, c'est-à-dire qu'une constante est ajoutée au spectre (celle-ci peut être dépendante de la longueur d'onde), ou multiplicatives, c'est-à-dire que chaque élément du spectre est multiplié par une constante. Ces effets sont typiques des phénomènes de diffusion de la lumière, qui induisent des variations de trajets optiques. Les prétraitements visant à corriger les effets de variation de diffusion ne doivent pas être appliqués si le paramètre d'intérêt est de nature physique (taille des particules ou turbidité par exemple).

Published in : SIMON G. (dir.) (2020), *Spectroscopies vibrationnelles. Théorie, aspects pratiques et applications*, Editions des archives contemporaines, ISBN : 9782813002556, doi : https ://doi.org/10.17184/eac.9782813002556

Toutes les méthodes de prétraitements spectroscopiques citées ci-dessous sont des méthodes à appliquer « en ligne », c'est-à-dire que le prétraitement est effectué échantillon par échantillon, contrairement au paragraphe précédent.

Une revue de méthodes de prétraitements spectroscopiques peut être trouvée dans la publication de Rinnan *et al.* [4]. Une application [5] compare l'efficacité des méthodes de prétraitement spectroscopiques pour l'élimination de facteurs de variabilité en vue de la discrimination de crèmes.

- **Correction de ligne de base** : elle soustrait un background spectroscopique à chaque spectre. La méthode la plus classique consiste à soustraire la valeur d'une longueur d'onde précise, ou bien la valeur la plus faible d'un spectre, à toutes les autres valeurs de ce spectre. La correction de ligne de base « Detrend » [6], soustrait une ligne de base polynomiale (généralement un polynôme de degré 2) à chaque spectre afin de pouvoir corriger une ligne de base variant en fonction de la longueur d'onde.

- **Normalisation spectroscopique** : ce type de normalisation est appliqué en « ligne » par spectre, afin de corriger des effets multiplicatifs sur les spectres. Cela permet de mettre l'accent sur le profil du spectre (*fingerprint*) plutôt que sur son intensité globale. La normalisation se fait en divisant chaque spectre par une estimation de son intensité spectrale, par exemple par l'un des critères suivants : son aire, son pic maximum, un pic spectral spécifique, sa longueur (normalisation par vecteur unitaire), ou la somme des valeurs spectrales.

- ***Standard Normal Variate*** **(SNV)** [6] : cette méthode est utilisée, comme la normalisation, pour limiter les problèmes de variation d'intensité spectrale (Fig. 2b). Elle consiste à soustraire la moyenne de chaque spectre et à le diviser par son écart-type.

- ***Multiplicative Signal Correction*** **(MSC)** [7] : cette méthode est aussi une technique très courante pour corriger les effets de diffusion additifs et multiplicatifs. Le principe est de modéliser chaque spectre comme une combinaison linéaire d'un spectre de référence, ayant subi un effet additif et un effet multiplicatif (équation 1). Ce spectre de référence est généralement le spectre moyen de la base de données d'étalonnage, considéré comme le spectre représentant un trajet optique moyen. Puis, chaque spectre est corrigé de son effet additif et/ou multiplicatif, comme indiqué dans l'équation 2. L'inconvénient de cette méthode est que le spectre de référence doit être représentatif de la diffusion moyenne afin d'éviter une correction inappropriée, et doit donc être modifié à chaque mise à jour de la base de données, contrairement au prétraitement SNV précédent.

$$\mathbf{x}_i = a * \bar{\mathbf{x}} + b + e$$

Équation 1

Published in : SIMON G. (dir.) (2020), *Spectroscopies vibrationnelles. Théorie, aspects pratiques et applications*, Editions des archives contemporaines, ISBN : 9782813002556, doi : https ://doi.org/10.17184/eac.9782813002556

$$\mathbf{x}_i^{MSC} = \frac{\mathbf{x}_i - b}{a}$$

Équation 2

Où $\mathbf{x}_i$ *est le spectre i mesuré,* $\bar{\mathbf{x}}$ *est le spectre moyen (ou un spectre de référence), b est l'ordonnée à l'origine, a est la pente, e est l'erreur résiduelle et* $\mathbf{x}_i^{MSC}$ *le spectre i corrigé.*

· **Le lissage** est utilisé pour limiter le bruit sur un spectre. Le principe consiste à remplacer la valeur d'un point par une moyenne calculée sur les points voisins. Par exemple, la méthode de la moyenne mobile (*moving window average*) utilise la moyenne arithmétique sur une fenêtre glissante pour calculer la nouvelle valeur. L'algorithme Savitzky-Golay (SG) [8] utilise un ajustement polynomial sur une fenêtre glissante. Ce dernier est l'algorithme le plus répandu en chimiométrie. Plus la fenêtre est large, plus le lissage est important et donc minimise le bruit mais peut mener à une perte d'information ; *a contrario*, plus le degré du polynôme est élevé, meilleur sera l'ajustement polynomial à des pics spectroscopiques étroits, mais au risque d'augmenter alors le bruit. Ainsi, un compromis entre ces deux paramètres doit être trouvé.

· **Les dérivées** des spectres sont généralement calculées afin de mettre en valeur la structure des spectres, par exemple pour faire mieux apparaitre des pics superposés, et ainsi faciliter l'extraction d'information par la future modélisation. Mais la dérivation a aussi comme intérêt de corriger les effets additifs : une ligne de base constante est retirée par une dérivée première (Fig. 2c) et des effets de décalage et de pente de la ligne de base sont retirés par une dérivée seconde (Fig. 2d). Les algorithmes les plus classiques utilisés pour le calcul des dérivées sont la dérivée *gap-segment* [9] ou l'algorithme de Savitzky-Golay [8]. Étant donné que la dérivation augmente le bruit du spectre en mettant en exergue toutes les petites variations, y compris celles provenant du bruit, ces deux méthodes lissent aussi le spectre, grâce à un ou des paramètres à ajuster [10].

2.3 Méthodes de réduction de la dimensionnalité

Les méthodes de réduction de la dimensionnalité visent à éliminer le signal non informatif, améliorant ainsi l'information et réduisant la colinéarité.

· **La sélection de variables** est souvent utilisée pour supprimer des variables non-informatives, bruitées et de ne garder que celles qui sont les plus pertinentes. Pour les données avec seulement quelques variables (par exemple des données de procédé), la méthode classique est la méthode de régression linéaire multiple pas à pas (*stepwise*-MLR), utilisée pour sélectionner les variables les plus informatives lors de l'élaboration du modèle. Pour des données ayant un plus grand nombre de variables et, de plus, corrélées entre elles (par exemple des données spectroscopiques), d'autres méthodes sont préférables. Une méthode couramment utilisée en spectroscopie, est la méthode de régression *interval-*

Published in : SIMON G. (dir.) (2020), *Spectroscopies vibrationnelles. Théorie, aspects pratiques et applications*, Editions des archives contemporaines, ISBN : 9782813002556, doi : https ://doi.org/10.17184/eac.9782813002556

Partial Least Squares (i-PLS) [11], qui sélectionne les régions spectrales pour tester toutes les combinaisons possibles avec une ou plusieurs fenêtre(s) mobile(s) d'une largeur fixe. Une pléthore de méthodes, très différentes les unes des autres, ont été développées. Par exemple, pour la méthode *Uninformative Variable Elimination* (UVE) [12], l'ajout de variables aléatoires fictives permet l'identification et donc l'élimination des variables spectrales qui sont aussi peu informatives que du bruit. La méthode des Algorithmes Génétiques (AG) [13] est une méthode stochastique utilisant le principe de la théorie de l'évolution pour sélectionner un sous-ensemble de variables. Un grand nombre de paramètres doit être réglé pour appliquer cette méthode. En outre, étant une technique stochastique, chaque essai fournit un résultat différent.
D'autres méthodes de sélection de variables existent, les inventorier de façon exhaustive n'est pas l'objectif de ce chapitre qui vise plutôt à expliquer les bonnes pratiques. Des études comparatives peuvent cependant être trouvées dans la littérature [14,15]. Une publication de 2014 [16] recense les méthodes de sélection de variables les plus récentes pour la modélisation quantitative et qualitative.

- Dans le cadre d'une modélisation prédictive, les prétraitements d'**orthogonalisation**, tels que l'*Orthogonal Signal Correction* (OSC), ont pour objectif de supprimer des variations qui ne sont pas liées à Y [17]. Parmi certaines variantes découlant de l'OSC classique et différentes dans la façon dont la partie non pertinente est modélisée, on peut citer : la *Direct Orthogonalisation* (DO) [18], la *Direct OSC Direct* (DOSC) [19], la *Piecewise OSC* (POSC) [20] et l'*Orthogonal-PLS* (OPLS) [21]. Ces méthodes ont pour objectif d'extraire le *Net Analyte Signal* (NAS), à savoir la partie du spectre liée à la quantité d'intérêt et qui est orthogonale aux autres composés. Il a été démontré [22] que ces prétraitements ne permettent pas forcément d'obtenir de meilleures performances que des modèles PLS classiques sur la base de données brutes. Cependant, un modèle réalisé sur des données orthogonalisées permet une meilleure compréhension du modèle.
Une comparaison des différents algorithmes basés sur l'OSC peut être consultée dans la littérature [22].

- **Les méthodes de compression des données** visent à réduire la dimensionnalité de grands ensembles de données. Des structures latentes sont extraites et utilisées pour reconstruire un signal « nettoyé » de contributions non souhaitées. L'Analyse en Composantes Principales (ACP) est la méthode la plus courante (voir 3.1). D'autres méthodes sont aussi utilisées, comme la transformée de Fourier (FT) travaillant sur le domaine fréquentiel [23,24] ou la transformée en ondelettes (*Wavelet Transform*, WT) travaillant sur le temps et la fréquence des domaines [25,26]. Ces deux méthodes présentent l'avantage de travailler sur un seul spectre à la fois, alors que l'ACP nécessite la modélisation de l'ensemble des spectres. Ces méthodes sont utiles pour reconstruire le signal corrigé ou bien pour extraire des structures latentes (par exemple des scores en ACP, ou des coefficients pour les méthodes FT et WT) et de les utiliser comme entrées dans des modèles prédictifs ou de discrimination.

Published in : SIMON G. (dir.) (2020), *Spectroscopies vibrationnelles. Théorie, aspects pratiques et applications*, Editions des archives contemporaines, ISBN : 9782813002556, doi : https ://doi.org/10.17184/eac.9782813002556

Une comparaison de plusieurs méthodes de prétraitement en vue de l'amélioration de la détermination de l'humidité et de la teneur en protéines par spectroscopie proche infrarouge dans des échantillons de fourrage est donnée par Azzouz *et al.* [27].

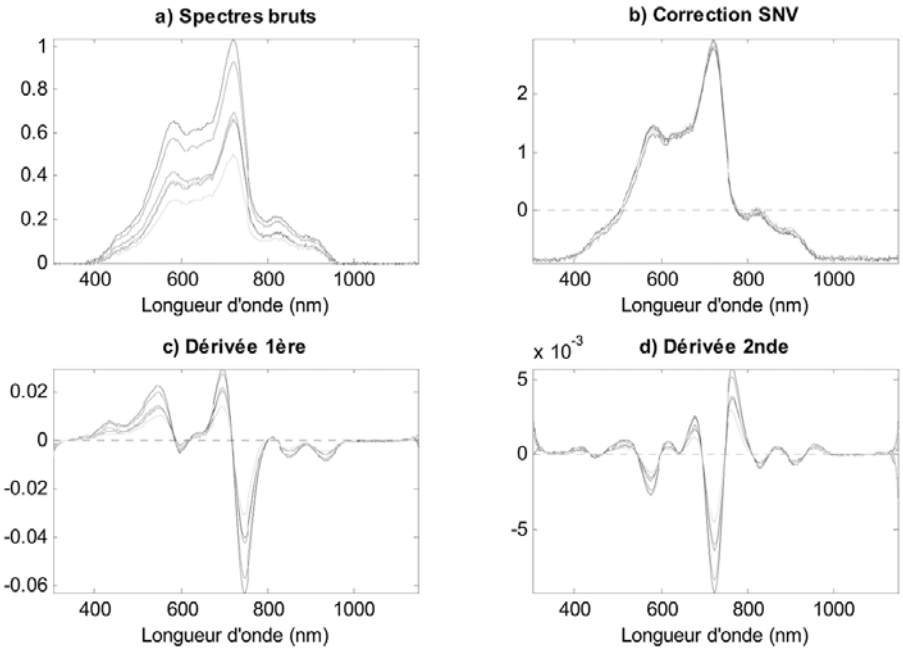


Figure 2 – (a) Spectres bruts ; (b) spectres traités par SNV ; (c) spectres prétraités par une dérivée première ; et (d) par une dérivée seconde, pour un ensemble de données spectrales dans le visible et le proche infrarouge.

3 L'analyse exploratoire : ACP

3.1 Introduction – L'objectif de l'ACP

L'analyse en composantes principales (ACP) [28] est l'outil de travail le plus utilisé en chimiométrie. Cette méthode peut être utilisée pour les objectifs suivants :

- La visualisation des données spectrales X dans l'espace multivarié.
- La détection d'échantillons atypiques ;
- La sélection de variables ;
- La compression des données, en réduisant la dimensionnalité de X et en supprimant le bruit ;
- L'ACP peut être également la première étape pour d'autres méthodes multivariées, telles que la classification non supervisée ou la maîtrise statistique des procédés (MSPC) [29].

Published in : SIMON G. (dir.) (2020), *Spectroscopies vibrationnelles. Théorie, aspects pratiques et applications*, Editions des archives contemporaines, ISBN : 9782813002556, doi : https ://doi.org/10.17184/eac.9782813002556

3.2 Interprétation géométrique

L'ACP peut être considérée comme l'une des meilleures façons de visualiser des échantillons représentés par de nombreuses variables, en projetant leurs coordonnées d'origine dans un nouvel ensemble d'axes, appelés composantes principales (CP). Ces axes répondent à un certain nombre de propriétés, ce qui rend la visualisation des échantillons plus facile. Les graphiques suivants expliquent comment l'ACP fonctionne pour une matrice X simple, composée de seulement trois variables (Fig. 3a) :

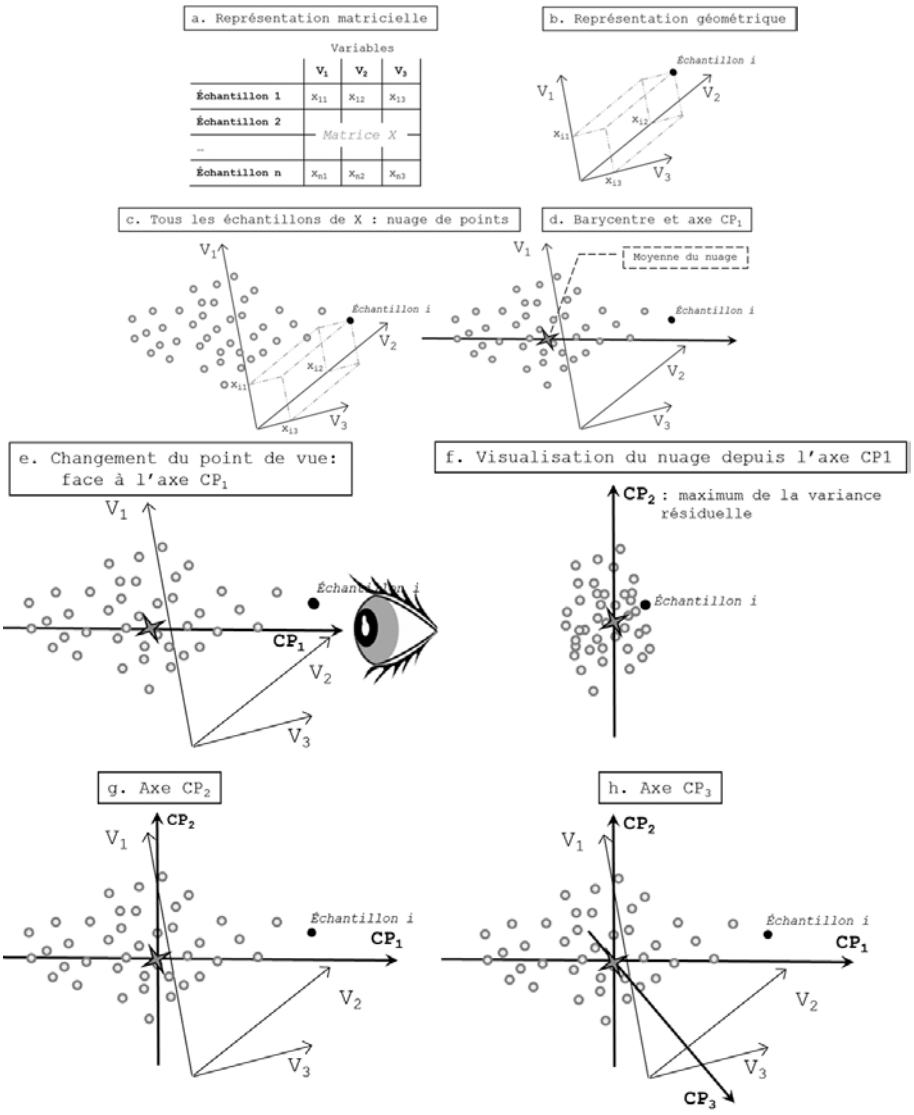


Figure 3 – Visualisation géométrique de l'Analyse en Composantes Principales (ACP).

Published in : SIMON G. (dir.) (2020), *Spectroscopies vibrationnelles. Théorie, aspects pratiques et applications*, Editions des archives contemporaines, ISBN : 9782813002556, doi : https ://doi.org/10.17184/eac.9782813002556

· Chaque échantillon est localisé dans l'espace d'origine par ses 3 coordonnées (Fig. 3b). La matrice X peut alors être visualisée comme un nuage de points dans cet espace 3D (Fig. 3c).

· Le système de coordonnées est translaté au barycentre de nuage de points (étoiles dans la figure 3d), pour une meilleure visualisation, par centrage des variables.

· Un nouvel axe, appelé première composante (CP_1), est construit suivant la direction de la variance maximale des échantillons (Fig. 3d) ; cela permet de mieux visualiser la variabilité maximale de l'ensemble des données. Les nouvelles coordonnées des échantillons sont appelées les scores.

· Un deuxième axe (CP_2), perpendiculaire au premier, est ensuite calculé, pour représenter le maximum de la variance résiduelle des données (Fig. 3g). Cet axe peut être visualisé en plaçant son œil face à la direction de la première composante (Fig. 3e-f).

· Ce processus est effectué de manière itérative pour un nombre de composantes égal au nombre de variables d'origine – 3 dans cet exemple graphique – (ou à $n-1$ si le nombre d'échantillons est inférieur au nombre de variables) (Fig. 3h).

Ainsi, l'ACP peut être considérée comme un changement d'axes, conçus pour mieux visualiser la variabilité des échantillons, tout en maintenant les distances et les échelles entre les échantillons. Pour simplifier la visualisation, les échantillons sont habituellement observés sur un plan 2D, correspondant à la projection des échantillons sur un ensemble de deux axes (voir dans la figure 4, un exemple de plan $CP_1 \times CP_3$).

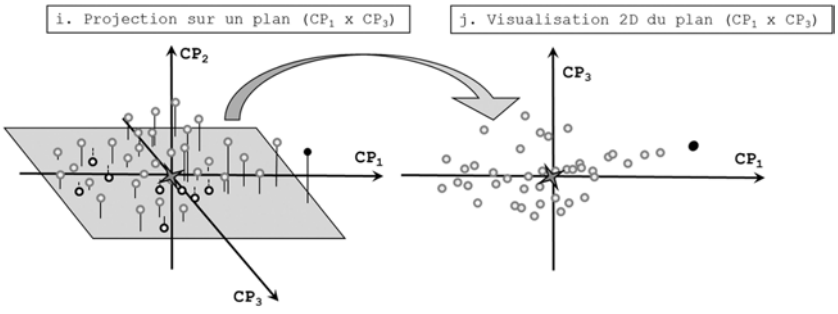


Figure 4 – Visualisation de l'ACP sur un plan 2D.

Published in : SIMON G. (dir.) (2020), *Spectroscopies vibrationnelles. Théorie, aspects pratiques et applications*, Editions des archives contemporaines, ISBN : 9782813002556, doi : https ://doi.org/10.17184/eac.9782813002556

3.3 Calcul mathématique

La propagation (ou l'inertie, ou la variance) du nuage d'échantillons vue dans la section précédente est exprimée mathématiquement par la matrice de variance-covariance de X, notée C. Ainsi, l'ACP décompose l'ensemble de l'espace des échantillons dans la direction de la variance maximale de X. La matrice de variance-covariance est calculée comme représenté dans l'équation 3, à partir de la matrice centrée de X.

$$C_{(p,p)} = \frac{\tilde{\mathbf{X}}^T_{(p,n)} \tilde{\mathbf{X}}_{(n,p)}}{n-1}$$

Équation 3

Où $\tilde{\mathbf{X}}$ *est la matrice centrée de X, n le nombre d'échantillons et l'exposant* ^{T} *correspond à la transposée. Les dimensions des matrices sont notées entre parenthèses.*

Afin de trouver les axes d'inertie maximale, l'algorithme de l'ACP diagonalise la matrice de variance-covariance, c'est-à-dire calcule les valeurs et les vecteurs propres de C. Les valeurs propres sont ensuite classées dans un ordre décroissant, car elles sont directement liées à la variance expliquée par chaque axe. Les vecteurs propres sont appelés *loadings* (P) et correspondent aux poids de chaque variable d'origine pour construire la composante : la CP est donc une combinaison linéaire des variables initiales.

Les données initiales X peuvent être décomposées par l'expression suivante :

$$\mathbf{X}_{(n,p)} = \mathbf{T}_{(n,p)} \mathbf{P}^T_{(p,p)}$$

Équation 4

Où $\mathbf{P}$ *est la matrice des loadings, et* $\mathbf{T}$ *la matrice des scores des nouvelles coordonnées du sous-espace. Les dimensions des matrices sont notées entre parenthèses.*

Puisque les données contiennent une partie d'information et une partie de bruit (erreur), X peut être exprimé avec les k premières composantes, les p-k dernières composantes sont alors incluses dans la matrice des résidus E (voir l'équation 5 et la figure 5) :

$$\mathbf{X}_{(n,p)} = \mathbf{T}_{(n,k)} \mathbf{P}^T_{(k,p)} + \mathbf{E}_{(n,p)}$$

Équation 5

Où $\mathbf{P}$ *est la matrice de loadings des k composantes,* $\mathbf{T}$ *les scores correspondants et* $\mathbf{E}$ *la matrice des résidus. Les dimensions des matrices sont notées entre parenthèses.*

Published in : SIMON G. (dir.) (2020), *Spectroscopies vibrationnelles. Théorie, aspects pratiques et applications*, Editions des archives contemporaines, ISBN : 9782813002556, doi : https ://doi.org/10.17184/eac.9782813002556

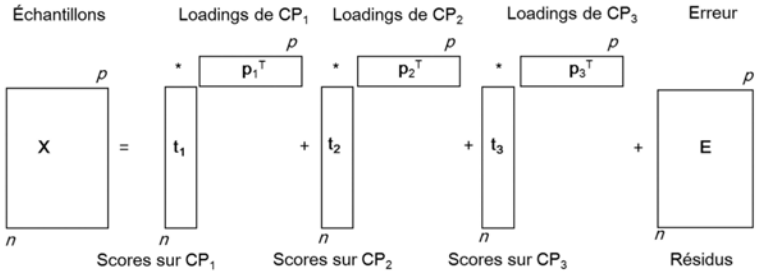


FIGURE 5 – Décomposition matricielle de l'Analyse en Composantes Principales, pour trois composantes.

3.4 Interprétation de l'ACP

Lorsqu'une ACP est effectuée, les scores et les loadings sont visualisés afin de comprendre :

- La variance expliquée cumulée montre la variabilité expliquée par chaque CP (Fig. 6a). Ainsi, l'interprétation des tendances, des échantillons voisins et/ou des variables, doit être pondérée par cette variance expliquée. Lorsque la variance expliquée est proche de 0, les composantes restantes expriment seulement le bruit, et non de l'information structurée. En spectroscopie, ce pourcentage est plus difficilement interprétable, car parfois, l'information d'intérêt peut représenter seulement 1 % de la variance totale, s'il s'agit par exemple d'un composé chimique minoritaire. La première composante exprime généralement plus de 90 % de la variance, et est souvent reliée à des effets de diffusion, qu'il faudra donc minimiser par prétraitement spectroscopique si ce phénomène n'est pas celui recherché. Si l'ACP est utilisée pour compresser l'information (voir 2.3), la détermination de ce nombre optimal de composantes est très importante de façon à ne pas inclure de bruit. Il est également essentiel à des fins de classification (voir le *clustering* [section 4] et SIMCA [section 6.2]), ou bien dans le contrôle de procédés.

- La présence d'échantillons atypiques est détectable par le calcul de statistiques issues de l'ACP (voir section 3.5 et Fig. 6b).

- La relation entre les échantillons est visible sur les graphiques des scores (fig. 6c). Si deux échantillons ont des scores proches (coordonnées voisines dans les différentes CPs), cela signifie qu'ils ont le même comportement, et ont donc des valeurs similaires pour les variables d'origine qui contribuent principalement à la construction de ces composantes. Les graphiques des scores peuvent mettre en évidence des tendances entre échantillons, des clusters...

- La corrélation entre les variables et leur contribution aux CPs sont représentés par les graphiques des *loadings* (Fig. 6d). Les variables sont considérées comme des vecteurs, construisant les directions des CPs avec différents poids. Par conséquent, si deux variables présentent des directions proches, cela signi-

Published in : SIMON G. (dir.) (2020), *Spectroscopies vibrationnelles. Théorie, aspects pratiques et applications*, Editions des archives contemporaines, ISBN : 9782813002556, doi : https ://doi.org/10.17184/eac.9782813002556

fie qu'elles sont fortement corrélées. En outre, si elles ont des valeurs absolues plus élevées sur l'axe (poids élevé du *loadings*), cela signifie qu'elles contribuent beaucoup au calcul de l'axe de cette CP. Cet outil peut donc aussi être utilisé pour la sélection de variables. En spectroscopie, les *loadings* sont observés comme des spectres, il est donc possible de juger visuellement la significativité des composantes en observant les pics et le bruit présent. La figure 6 montre bien que les composantes les plus éloignées contiennent du bruit.

· La relation entre les échantillons et les variables est facilement observable sur des données discrètes avec un graphique, appelé *biplot*, sur lequel les scores et les *loadings* sont représentés de façon simultanée. En spectroscopie, le nombre de variables conséquent, et leurs fortes corrélations, rend ce genre de graphique peu utile. Il faut cependant savoir que si un échantillon a un score élevé sur une CP cela signifie qu'il a des valeurs élevées pour les longueurs d'onde ayant des poids élevés sur le *loadings* correspondant. Il est donc possible de faire le lien entre échantillons et variables de cette manière.

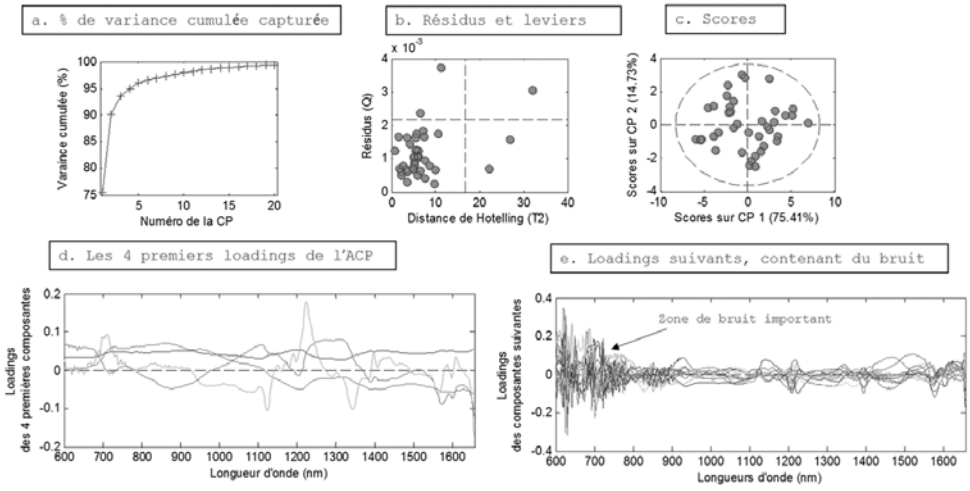


Figure 6 – Graphiques principaux pour l'interprétation de l'ACP.

3.5 Détection et gestion des échantillons atypiques (*outliers*)

Les modèles multivariés classiques exploratoires et prédictifs sont très sensibles aux valeurs aberrantes, à savoir, des échantillons atypiques extrêmes (appelés généralement *outliers*) qui sont très différents des autres échantillons de la population. Par conséquent, il est d'une grande importance d'identifier les valeurs aberrantes et de décider comment gérer ces échantillons, c'est-à-dire soit de les conserver s'ils sont représentatifs de la variabilité qui doit être modélisée, ou bien de les supprimer. Il existe en effet de nombreuses causes de valeurs aberrantes menant à des décisions différentes : étiquetage erroné, erreur de mesure, déviation du procédé, etc.

Published in : SIMON G. (dir.) (2020), *Spectroscopies vibrationnelles. Théorie, aspects pratiques et applications*, Editions des archives contemporaines, ISBN : 9782813002556, doi : https ://doi.org/10.17184/eac.9782813002556

Les valeurs aberrantes peuvent être détectées lors d'une analyse exploratoire (ACP) de la base de données X.

3.5.1 Leviers (T^2) ou distance de Mahalanobis

L'outil de diagnostic le plus couramment utilisé pour identifier les valeurs aberrantes est le levier de X (noté h), calculé pour chaque échantillon i :

$$h_i = \mathbf{t_i}(\mathbf{T_0^T T_0})^{-1}\mathbf{t_i^T}$$

Équation 6

Où $\mathbf{T_0}$ *est la matrice des scores d'étalonnage issue d'une ACP pour un certain nombre de composantes et* $\mathbf{t_i}$ *le vecteur des scores pour l'échantillon i.*

L'effet de levier (similaire à la distance de Mahalanobis [30] ou la distance du $\mathbf{T^2}$ de Hotelling [31]) correspond à la distance d'un échantillon au centre du modèle. Les échantillons avec un fort effet de levier ont une grande influence sur le modèle et doivent donc être bien interprétés (Fig. 7). Ce genre de valeur aberrante est aussi appelée *outlier* « fort ». Une limite critique peut être calculée en fonction d'un test de Fisher (F-test) pour un niveau α de signification donnée (par exemple 5 %) pour détecter les valeurs aberrantes.

3.5.2 Résidus (Q)

L'autre outil de diagnostic également très utilisé dans l'identification des valeurs aberrantes est l'observation des résidus de X (souvent notés Q). Ils correspondent à la partie non modélisée de la variance. C'est la distance par rapport au modèle hyperplan (équation 7). Les échantillons avec un résidu élevé ne sont pas bien modélisés (Fig. 7). Ils sont souvent issus d'un phénomène atypique dans le système exploré. Ce genre de valeur aberrante est aussi appelé *outlier* « faible » ou « modéré ». D'une manière similaire à l'effet de levier, une limite critique peut être calculée à l'aide d'un F-test pour détecter les valeurs aberrantes en regard de cette distance.

$$Q_i = (x_i - \hat{x}_\mathrm{i})(x_i - \hat{x}_\mathrm{i})^\mathrm{T}$$

Équation 7

Où Q_i *est le résidu de l'échantillon i,* x_i *le spectre de l'échantillon i et* $\hat{x}_\mathrm{i}$ *le spectre de l'échantillon i reconstruit à partir de k composantes de l'ACP.*

Le tracé de la valeur des résidus en fonction de l'effet de levier (Fig. 6b) est un outil de diagnostic très utile pour identifier simultanément tous les types de valeurs aberrantes. Sur ce genre de graphique, les échantillons les plus critiques sont ceux ayant un fort levier et un fort résidu, car ces derniers ont une forte influence sur le modèle mais ne sont pas bien modélisés.

Dans tous les cas, un diagnostic des valeurs aberrantes est nécessaire avant toute prise de décision. Retirer des échantillons atypiques mais réels peut diminuer la représentativité des modèles qui ne seront donc pas applicables pour ce genre d'échantillons.

Published in : SIMON G. (dir.) (2020), *Spectroscopies vibrationnelles. Théorie, aspects pratiques et applications*, Editions des archives contemporaines, ISBN : 9782813002556, doi : https ://doi.org/10.17184/eac.9782813002556

- - - - : Distance au centre du modèle . . . . : Distance au modèle

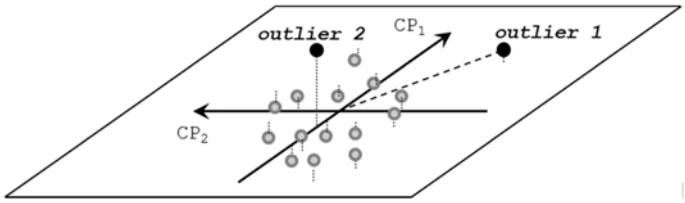


FIGURE 7 – Deux types de valeurs aberrantes dans l'analyse exploratoire : valeurs aberrantes avec un effet de levier élevé (*outlier* 1) et des valeurs aberrantes avec une grande variance résiduelle (*outlier* 2).

L'utilisation classique de l'ACP pour la détection d'échantillons atypiques a été critiquée du fait que les statistiques utilisées dans sa construction (par exemple la covariance) sont fortement influencées par la présence d'*outliers*. Des méthodes basées sur des statistiques robustes ont donc été implémentées [32].

4 Classification non supervisée : CAH, *k-means*

La question de la classification des échantillons est souvent rencontrée. Il existe deux ensembles de méthodes. Le premier regroupe les méthodes de classification non-supervisée (ou *clustering*) qui visent à classer des échantillons similaires sans l'utilisation de connaissances *a priori*. Le second regroupe les méthodes de classification supervisée (ou discrimination), où l'appartenance des échantillons aux différentes classes est utilisée pour construire un modèle. Ce dernier ensemble contient également des méthodes de prédiction quantitatives adaptées aux problématiques qualitatives et sera vu dans la section 6.

4.1 Présentation des techniques de clustering

Les techniques de *clustering* sont également appelées « techniques de classification non supervisée » dans le domaine de la chimiométrie. Ce sont des outils d'exploration qui visent à déterminer les tendances de regroupements naturels de la structure interne des données sans aucune autre connaissance préalable sur l'affectation des échantillons à des classes. Ainsi, les méthodes décrites dans cette section mesurent les similitudes entre les échantillons, uniquement en fonction de leurs valeurs X.

Comme indiqué dans la section 3.1, l'ACP peut être considérée comme une base pour l'analyse de clusters étant donné que les graphiques des scores montrent la répartition des échantillons dans un espace multidimensionnel faible. Ces graphiques peuvent permettre d'identifier des tendances de regroupement d'échantillons pouvant ainsi définir des clusters. Le seul paramètre à régler est le nombre de composantes principales à retenir (k), c'est-à-dire contenant des informations utiles et non pas du bruit résiduel. L'interprétation des résultats de la classification est basée sur les coordonnées des échantillons dans l'ACP réalisée sur ces k composantes informatives. L'ACP est d'ailleurs une étape indispensable avant l'application d'une méthode de classification dans le cas de données de grande dimension.

Published in : SIMON G. (dir.) (2020), *Spectroscopies vibrationnelles. Théorie, aspects pratiques et applications*, Editions des archives contemporaines, ISBN : 9782813002556, doi : https ://doi.org/10.17184/eac.9782813002556

Dans les méthodes de classification, les tâches cruciales sont de régler les paramètres initiaux, de définir un critère d'arrêt et d'évaluer la validité des clusters obtenus. Certains outils ont été développés pour aider l'utilisateur à prendre des décisions. À titre d'exemple, la validité du cluster peut être évaluée par le calcul d'indices ou par des tests statistiques ; cela contribue alors à choisir la meilleure partition [33].

Une revue des différentes techniques de classification peut être consultée dans la littérature [34].

4.2 Classification Ascendante Hiérarchique (CAH)

L'analyse hiérarchique est une méthode qui assemble ou dissocie successivement les ensembles d'échantillons [35]. Dans la Classification Ascendante Hiérarchique (CAH), n classes sont considérées au départ, à savoir une classe par échantillon, et ces classes sont ensuite regroupées successivement jusqu'à constituer une classe unique. Le résultat est donné sous la forme d'un arbre de classification, appelé dendrogramme (Fig. 18), où la longueur des branches représente la distance entre les groupes. Le choix des groupes finaux est décidé en coupant l'arbre à un seuil spécifique ; ainsi le nombre de clusters n'est pas un paramètre à régler à l'avance, contrairement aux techniques des moyennes mobiles (ou *k-means*) (voir 4.3).

La procédure inverse, connue sous le nom de classification hiérarchique descendante considère les échantillons comme un seul groupe unique qui est ensuite divisé successivement en deux groupes, aussi différents que possible, jusqu'à ce que tous les échantillons soient séparés. Les deux approches ne sont pas équivalentes, et cette dernière est beaucoup moins utilisée car elle est informatiquement intensive.

Dans les classifications hiérarchiques, deux paramètres doivent être définis :

- La distance entre les échantillons. Il s'agit généralement de la distance euclidienne (équation 8) ou de la distance de Mahalanobis (équation 9). Il est à noter que la distance peut être calculée sur les résultats d'une ACP si X présente des variables colinéaires.

- Le critère de regroupement. Les critères classiques de classification reposent sur la minimisation de l'une des conditions suivantes : la distance minimale (méthode du saut minimal), la distance maximale (méthode du saut maximal), la distance moyenne (méthode de la moyenne non pondérée), la distance entre les centres de gravité, ou encore l'augmentation de l'inertie intra-classe, favorisant des groupes compacts (critère de Ward).

$$d_e(\mathbf{x}_a, \mathbf{x}_b) = \sqrt{(\mathbf{x}_a - \mathbf{x}_b)(\mathbf{x}_a - \mathbf{x}_b)^T}$$

Équation 8

Published in : SIMON G. (dir.) (2020), *Spectroscopies vibrationnelles. Théorie, aspects pratiques et applications*, Editions des archives contemporaines, ISBN : 9782813002556, doi : https ://doi.org/10.17184/eac.9782813002556

$$d_m(\mathbf{x}_a, \mathbf{x}_b) = \sqrt{(\mathbf{x}_a - \mathbf{x}_b)\mathbf{C}^{-1}(\mathbf{x}_a - \mathbf{x}_b)^T}$$

Équation 9

Où $\mathbf{x}_a$ *et* $\mathbf{x}_b$ *sont les réponses des échantillons A et B, et* $\mathbf{C}^{-1}$ *est l'inverse de la matrice de variance-covariance de l'ensemble des données X.*

En fonction des paramètres choisis, les résultats de classification peuvent être très différents, comme le montre la figure 8.

- - - *Exemple de seuil de coupure de l'arbre pour la définition des groupes.*

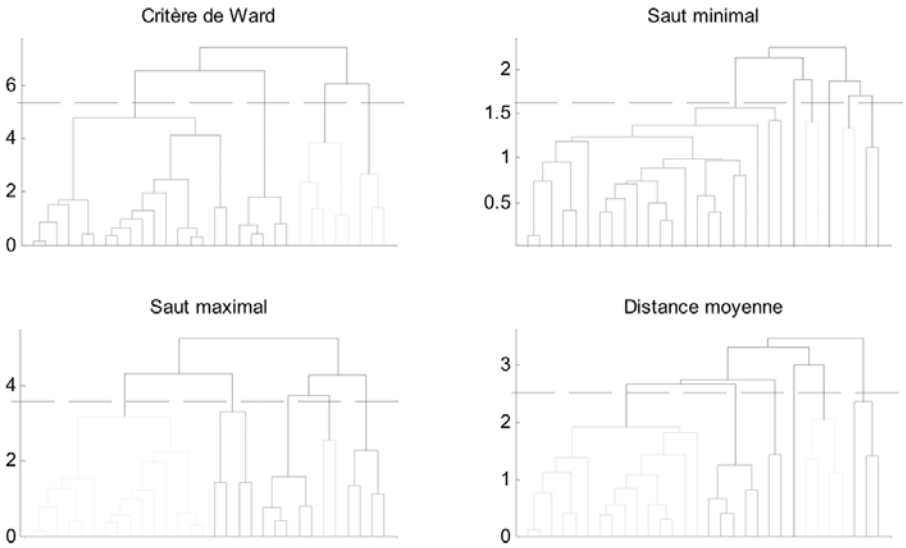


Figure 8 – Dendrogrammes construits avec la distance de Mahalanobis et différents critères de regroupement : (a) critère de Ward ; (b) saut minimal ; (c) saut maximal ; et (d) distance moyenne.

4.3 Méthodes de classifications non hiérarchiques

Les méthodes de regroupement non hiérarchiques visent à construire une partition finale des données. Ainsi, contrairement à l'approche hiérarchique, l'utilisateur doit définir un nombre fixe de groupes *a priori*. Ce qui peut être une forte limitation à l'utilisation de ces techniques.

De nombreuses méthodes de classification non hiérarchiques ont été développées :

- Des méthodes basées sur une mesure de distance pour évaluer les similitudes entre les échantillons, tels que la méthode des centres mobiles (ou *k-means* [36]).

- Des méthodes basées sur la densité des échantillons, comme DB-scan [37,38].

Published in : SIMON G. (dir.) (2020), *Spectroscopies vibrationnelles. Théorie, aspects pratiques et applications*, Editions des archives contemporaines, ISBN : 9782813002556, doi : https ://doi.org/10.17184/eac.9782813002556

· Des méthodes basées sur la relation entre les variables et les échantillons, telles que OPTICS [39] ou les réseaux de neurones artificiels de Kohonen (par exemple *Self-Organizing Map* ou SOM [40,41]).

Cette section décrit la technique la plus couramment utilisée en chimiométrie, à savoir les *k-means*.

La méthode des *k-means* [36] est une procédure itérative qui permet de trouver la partition optimale de k classes (Fig. 9). Comme dans la CAH, le procédé peut être utilisé sur les scores de l'ACP, si les données sont de grande dimension.

· Les conditions initiales : la partition initiale de k groupes est généralement générée au hasard. Les résultats sont très dépendants de cette partition initiale, ainsi que sur le choix du nombre de classes k.

· À chaque itération, le barycentre de chacune des classes est recalculé et les échantillons sont réaffectés au centre le plus proche.

· Le critère d'arrêt : cette procédure est effectuée jusqu'à ce que le critère d'arrêt soit atteint (par exemple : aucun changement d'affectation ou bien un nombre maximum d'itérations atteint).

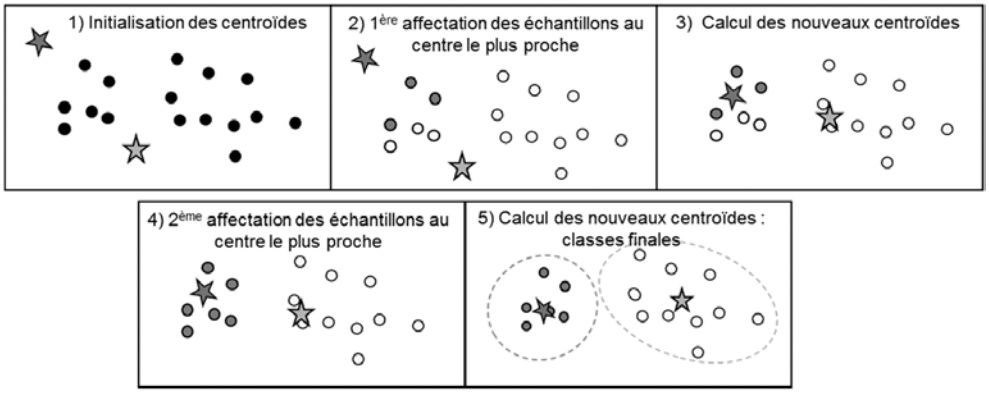


Figure 9 – Principe de l'algorithme des *k-means*.

Une revue des variantes de l'algorithme des *k-means* depuis leur conception peut être trouvée dans la littérature [42] ; cet article mentionne également des critères permettant de trouver le nombre optimal de groupes.

5 Régression : MLR, PCR, PLS

Les régressions multivariées sont certainement les méthodes qui ont permis de démontrer très tôt le potentiel de la chimiométrie dans le milieu industriel. L'objectif d'une régression est de corréler un ensemble de données explicatives $\mathbf{X}$ à une ($\mathbf{y}$) ou plusieurs propriétés d'intérêt ($\mathbf{Y}$). Tous les modèles de régression obéissent finalement

Published in : SIMON G. (dir.) (2020), *Spectroscopies vibrationnelles. Théorie, aspects pratiques et applications*, Editions des archives contemporaines, ISBN : 9782813002556, doi : https ://doi.org/10.17184/eac.9782813002556

à l'équation suivante :

$$\widehat{\mathbf{y}} = f(\mathbf{X})$$

Équation 10

Où $\widehat{\mathbf{y}}$ *est la valeur prédite de la valeur réelle* $\mathbf{y}$ *et* f *la fonction reliant* $\mathbf{X}$ *et* $\mathbf{y}$.

La fonction f est déterminée en minimisant l'erreur entre les valeurs vraies $\mathbf{y}$ et les valeurs prédites $\widehat{\mathbf{y}}$, au moyen de divers algorithmes.

Les modèles sont développés et optimisés sur la base de jeux dits d'étalonnage (calibration) et de test pour lesquels les ensembles $\mathbf{X}$ et $\mathbf{y}$ sont connus. L'objectif final est d'appliquer ces modèles de régression pour prédire leurs valeurs $\mathbf{y}$ d'échantillons totalement inconnus, n'ayant pas contribué à l'élaboration ou l'optimisation des modèles.

Le modèle de régression le plus basique est issu d'une régression linéaire simple (univariée) reliant une quantité d'intérêt $\mathbf{y}$ (réponse) à une variable explicative $\mathbf{x}_1$ suivant l'équation :

$$\widehat{\mathbf{y}}_{(n,1)} = b_0 + b_1 \times \mathbf{x}_{1(n,1)}$$

Équation 11

Où $\mathbf{b}$ *est le vecteur de régression, composé ici de deux constantes :* b_0 *l'ordonnée à l'origine (soit la valeur de y lorsque* x_1 *est nul), et* b_1 *la pente.*

On comprend vite que cette méthode est difficilement applicable en spectroscopie pour des raisons évidentes d'interférences. Lorsque la matrice de données $\mathbf{X}$ est dite multivariée, c'est-à-dire que l'échantillon est décrit par un grand nombre p de variables, un modèle linéaire multivarié relie alors $\mathbf{X}$ et $\mathbf{y}$ selon l'équation :

$$\widehat{\mathbf{y}}_{(n,1)} = b_0 + b_1 \times \mathbf{x}_{1(n,1)} + b_2 \times \mathbf{x}_{2(n,1)} + \cdots + b_p \times \mathbf{x}_{p(n,1)} = b_0 + \mathbf{X}_{(n,p)}\mathbf{b}_{(p,1)}$$

Équation 12

Où $\mathrm{b_1 \ldots b_p}$ *sont les coefficients de régression de chaque variable et réfèrent au vecteur de régression* $\mathbf{b}$ *sur la partie droite de l'équation, et* b_0 *l'ordonnée à l'origine. Les dimensions des matrices sont indiquées entre parenthèses dans l'équation.*

L'objectif final de la modélisation est de parvenir à retrouver une approche optimale de calcul des coefficients de régression $\mathbf{b}$ basée sur un ensemble de données expérimentales. Les différentes sections de cette partie tendent à présenter différentes approches de détermination du vecteur $\mathbf{b}$ de l'équation 12 à partir d'une matrice de données $\mathbf{X}$ et d'un vecteur $\mathbf{y}$ contenant les valeurs de référence.

Published in : SIMON G. (dir.) (2020), *Spectroscopies vibrationnelles. Théorie, aspects pratiques et applications*, Editions des archives contemporaines, ISBN : 9782813002556, doi : https ://doi.org/10.17184/eac.9782813002556

5.1 Régression linéaire multiple (MLR)

La régression linéaire multiple, ou régression multilinéaire (MLR pour *Multiple Linear Regression*) est la méthode de modélisation multivariée la plus simple puisqu'il s'agit d'une extension multivariée de la régression linéaire simple. Afin de calculer les coefficients $\mathbf{b}$, les calculs de pseudo-inverse de $\mathbf{X}$ sont appliqués suivant :

$$\mathbf{b} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}$$

Équation 13

Cette méthode présente un avantage certain dans sa simplicité d'exécution. Toutefois, si les variables explicatives présentent une certaine corrélation (colinéarité), le calcul de la matrice pseudo-inverse ne permettra pas d'obtenir un modèle de régression stable. De plus, la régression MLR ne peut être appliquée si le nombre d'échantillons est inférieur au nombre de variables explicatives ($n < p$), car la matrice $\mathbf{X}^T\mathbf{X}$ ne peut être inversée. Une dernière contrainte reste la sélection d'un sous-ensemble optimal de variables spectrales ainsi que son nombre. Ces trois limitations sont très fréquemment rencontrées en spectroscopie ou en imagerie, domaines dans lesquels un grand nombre de variables très souvent corrélées est mis en jeu. C'est pourquoi d'autres méthodes de régression sont plutôt adoptées pour ce type de données.

5.2 Régression en composantes principales (PCR)

Afin de surmonter la corrélation des variables et d'éviter la problématique de sélection, une régression MLR peut être réalisée, non pas sur les données originales $\mathbf{X}$, mais sur les *scores* $\mathbf{T}$ d'une analyse ACP réalisée sur les données originales (voir partie 3). L'équation 13 peut alors être remplacée par l'équation 14, ci-dessous, modélisation portant alors le nom de régression en composantes principales (PCR pour *Principal Component Regression*).

$$\mathbf{b} = \left(\mathbf{T}^T\mathbf{T}\right)^{-1}\mathbf{T}^T\mathbf{y}$$

Équation 14

Puisque les *scores* d'une analyse ACP sont orthogonaux, il n'existe aucune corrélation entre les variables. De plus, seuls les *scores* des k premières composantes principales informatives sont conservés ($k \ll p$), tel qu'il est expliqué section 3.3 (voir équation 4), afin de contourner les problèmes de forte dimensionnalité ($n > k$).

Néanmoins, le principal inconvénient de cette méthode provient du fait que les *scores* sont calculés afin d'expliquer le maximum de variance de $\mathbf{X}$, ce qui n'est pas nécessairement lié aux réponses contenues au sein du vecteur $\mathbf{y}$. Même si cette méthode de régression sur facteurs a permis de répondre aux contraintes de la régression MLR, elle est largement remplacée aujourd'hui par la régression PLS.

Published in : SIMON G. (dir.) (2020), *Spectroscopies vibrationnelles. Théorie, aspects pratiques et applications*, Editions des archives contemporaines, ISBN : 9782813002556, doi : https ://doi.org/10.17184/eac.9782813002556

5.3 Régression par les moindres carrés partiels (PLS)

La régression par les moindres carrés partiels (PLS pour *Partial Least Squares Regression*) [43] a été proposée afin d'améliorer la modélisation des réponses $\mathbf{y}$. Au-delà de calculer les *scores* seulement sur la matrice de covariance de $\mathbf{X}$ à l'aide de l'analyse ACP (méthode de modélisation employée par la régression PCR), l'algorithme de régression PLS prends également en compte la corrélation existante entre $\mathbf{X}$ et $\mathbf{y}$. Les *scores*, désignés ici par le terme variables latentes (LV pour *Latent Variables*), sont développés afin d'atteindre le maximum d'information possible sur le vecteur des réponses $\mathbf{y}$.

La régression PLS peut également être employée pour la prédiction de plusieurs réponses simultanément même si son utilisation reste anecdotique en spectroscopie. Dans ce cas, un sous-espace latent est également généré pour $\mathbf{Y}$ et la maximisation de la covariance est réalisée entre les *scores* de $\mathbf{X}$ et les *scores* de $\mathbf{Y}$. Cette méthode porte également le nom de régression PLS2.

La régression PLS peut être résumée en deux étapes indépendantes :

- Une réduction de dimensionnalité de $\mathbf{X}$ et $\mathbf{Y}$ (dans le cas d'une régression PLS2) en espaces latents à l'aide de projections des différents *scores*.

- L'intégration d'un vecteur des poids $\mathbf{w}$ reliant les différents sous-espaces en employant la covariance entre les *scores* de $\mathbf{X}$ et $\mathbf{Y}$ comme critère d'optimisation.

Deux algorithmes principaux sont utilisés pour la modélisation PLS : une procédure itérative nommée *Non-Linear Iterative Partial Least Squares* (NIPALS) [43] ou une approche directe plus rapide, l'algorithme SIMPLS (*Straightforward Implementation of the PLS method* ou *Statistically Inspired Modification of the PLS method*) [44].

Même si la problématique de la sélection de variables est résolue avec la régression PLS, il reste néanmoins à sélectionner un nombre optimal de facteurs afin de relier $\mathbf{X}$ à $\mathbf{y}$ de manière optimale. En effet, la sélection d'un nombre trop faible de composantes laisserait des informations pertinentes en dehors du modèle qui serait alors sous-entraîné. D'un autre côté, la sélection d'un nombre trop important de composantes laisserait entrer du bruit dans le modèle réduisant aussi son pouvoir prédictif. Afin de sélectionner ce nombre optimal de composantes, on utilise des méthodes de validation croisée.

6 Discrimination supervisée : FDA, SIMCA, PLS-DA

Au sein des méthodes de discrimination supervisée, les classes des échantillons de l'ensemble d'étalonnage sont connues et employées. L'objectif est de développer un modèle discriminant capable de déterminer la classe d'un nouvel échantillon inconnu. Les performances des modèles sont alors évaluées à l'aide d'une matrice de confusion, exposant le nombre d'échantillons mal classés.

En prenant l'exemple de l'industrie agro-alimentaire, les applications majeures sont l'authentification des produits, le tri de ces-derniers ou encore l'étude de leurs confor-

Published in : SIMON G. (dir.) (2020), *Spectroscopies vibrationnelles. Théorie, aspects pratiques et applications*, Editions des archives contemporaines, ISBN : 9782813002556, doi : https ://doi.org/10.17184/eac.9782813002556

mités par rapport à des standards de qualité. Une revue de ces méthodes et de leur utilisation dans l'industrie agro-alimentaire est présentée dans l'article de Berruta *et al.* [45].

6.1 Analyse discriminante factorielle (FDA)

L'analyse discriminante factorielle (FDA pour *Factorial Discriminant Analysis*) [46] est une méthode de projection en variables latentes qui maximise la variance inter-groupe (B) et minimise la variance intra-groupe (W) au lieu de se concentrer seulement sur la variance globale (matrice variance-covariance) comme dans l'ACP.

Un nouvel échantillon est attribué au groupe pour lequel il présente une distance de Mahalanobis minimale. L'analyse discriminante linéaire (LDA pour *Linear Discriminant Analysis*) est très similaire à l'analyse FDA. L'analyse discriminante quadratique (QDA pour *Quadratic Discriminant Analysis*) est réalisée lorsque l'hétérogénéité des différents groupes doit être prise en compte lors du développement d'une variance interne à chaque groupe utilisé.

L'analyse FDA est couramment employée pour des applications discriminantes. Toutefois, cette méthode requiert plus d'échantillons que de variables. Elle n'est donc pas parfaitement adaptée à diverses applications spectroscopiques si nous restons dans l'espace initial des variables. Afin de contourner cette limitation, l'analyse FDA est souvent appliquée sur des variables spécifiques sélectionnées ou sur des *scores* provenant d'une analyse ACP ou d'une régression PLS. Dans ce dernier cas, les facteurs discriminants peuvent être reconstitués en spectres discriminants pour l'interprétation des classes [47].

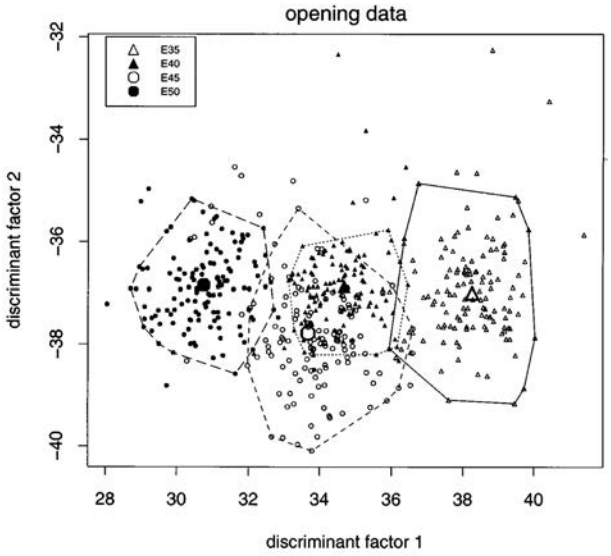


Figure 10 – Exemple d'analyse FDA pour la discrimination de produits de mouture présentant différentes tailles.

Published in : SIMON G. (dir.) (2020), *Spectroscopies vibrationnelles. Théorie, aspects pratiques et applications*, Editions des archives contemporaines, ISBN : 9782813002556, doi : https ://doi.org/10.17184/eac.9782813002556

Les particules de tailles extrêmes (E35 et E50) sont bien discriminées par le premier facteur, tandis que les groupes des particules de taille moyenne (E40 et E45) se recouvrent [48].

Ainsi par exemple, l'analyse FDA a été appliquée aux scores d'une analyse ACP réalisée sur des spectres proche infrarouge pour discriminer quatre types de farine de blé pour ensuite authentifier les produits de l'industrie meunière [49].

6.2 La méthode SIMCA

La méthode SIMCA (pour *Soft Independent Modelling of Class Analogy*) [50] est basée sur l'analyse ACP, ce qui la rend parfaitement adaptée pour les données présentant un grand nombre de variables.

Pour chaque classe k, une analyse en composantes principales est réalisée en ne considérant que les échantillons associés à cette dernière. Dans une deuxième étape, un intervalle de confiance est déterminé pour chaque modèle ACP afin de définir les limites d'appartenance aux classes. Ces limites peuvent être basées sur différents critères tels que la distance Euclidienne des $\mathbf{X}$ résiduels (Q), ou plus souvent sur une combinaison des statistiques Q et du T^2 de Hotelling. Un échantillon inconnu est ensuite considéré comme appartenant à la classe k si celui-ci tombe dans les limites de l'ACP correspondant à cette classe. Si différentes classes sont très proches les unes des autres, ou si celles-ci se recouvrent partiellement, il est possible que l'échantillon inconnu soit associé à plusieurs classes simultanément. Toutefois, spécificité propre à la méthode SIMCA, il se peut que cet échantillon inconnu ne soit attribué à aucune classe ce qui est une bonne chose en soi si ce dernier présente une classe non représentée dans la base d'apprentissage.

En général on utilise le graphique de Cooman permettant de confronter les résidus (ou T^2) de deux modèles (classes). Il présente les limites respectives des classes et permet d'observer d'éventuels recouvrements entre les deux groupes. Le graphique est séparé en quatre régions distinctes délimitées par les intervalles de confiance des $\mathbf{X}$ résiduels (Fig. 11) :

- La partie inférieure droite correspond aux échantillons assignés à la classe A (ici les huiles d'olive).
- La partie supérieure gauche correspond aux échantillons assignés à la classe B (ici les huiles de maïs).
- La partie inférieure gauche présente le recouvrement entre les classes A et B, contenant alors les échantillons attribués aux deux classes.
- La partie supérieure droite contient les échantillons rejetés des deux classes, c'est-à-dire des *outliers* ou des échantillons correspondant à d'autres classes (considérant que l'analyse SIMCA comporte plus de deux modèles).

Published in : SIMON G. (dir.) (2020), *Spectroscopies vibrationnelles. Théorie, aspects pratiques et applications*, Editions des archives contemporaines, ISBN : 9782813002556, doi : https ://doi.org/10.17184/eac.9782813002556

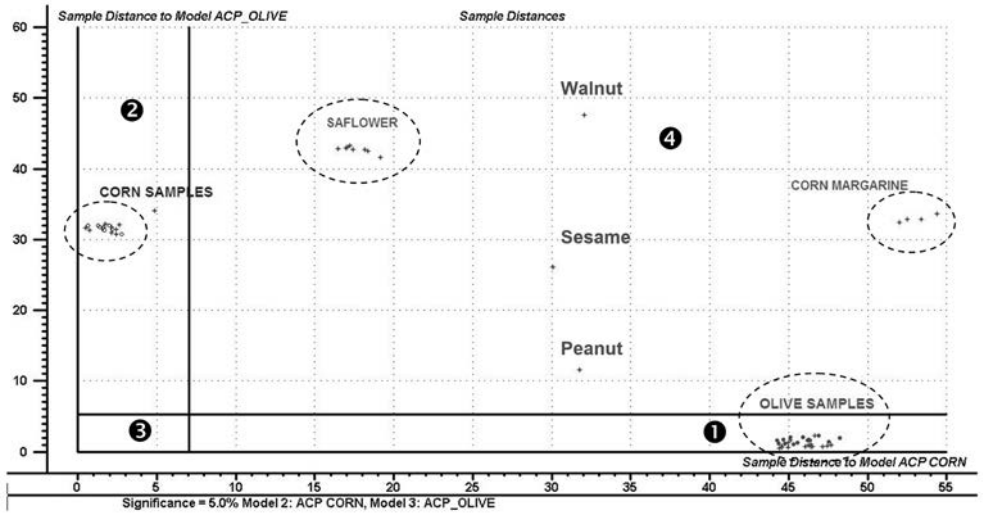


FIGURE 11 – Graphique de Cooman pour une classe A (huile d'olive) et une classe B (huile de maïs). *Remerciements à CAMO, Oslo, Norvège.*

6.3 Analyse discriminante par les moindres carrés partiels (PLS-DA)

L'analyse discriminante par les moindres carrés partiels (PLS-DA pour *Partial Least Squares-Discriminant Analysis*) correspond dans les faits à une régression PLS réalisée sur un jeu $\mathbf{Y}_{(n,k)}$ présentant des réponses binaires (Fig. 12) [51] et décrivant l'appartenance ou non aux k classes.

$$G = \begin{bmatrix} A \\ A \\ B \\ C \\ C \\ C \end{bmatrix} \qquad \mathbf{Y} = \begin{array}{c} \begin{matrix} A & B & C \end{matrix} \\ \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix} \end{array}$$

FIGURE 12 – Conversion d'un vecteur à trois classes (A, B et C) en une matrice binaire à trois entrées $\boldsymbol{Y}$.

Roussel *et al.* ont appliqué la méthode PLS-DA à la classification de moûts de raisin blanc, évaluant la pertinence de différentes techniques analytiques telles qu'un « nez électronique », les spectroscopies UV et FT-IR ainsi que d'autres méthodes de classification [52].

7 Résolution multivariée de courbes ou démélange

L'objectif d'une méthode de résolution est assez simple puisqu'à partir uniquement d'un ensemble de spectres de mélanges, il est possible d'extraire simultanément tous les spectres d'espèces pures présentes dans le système chimique étudié et leurs concen-

Published in : SIMON G. (dir.) (2020), *Spectroscopies vibrationnelles. Théorie, aspects pratiques et applications*, Editions des archives contemporaines, ISBN : 9782813002556, doi : https ://doi.org/10.17184/eac.9782813002556

trations relatives. Nous devons souligner ici que cette décomposition est obtenue sans connaissance a priori du système chimique. Il existe plusieurs méthodes permettant de proposer une telle décomposition. En effet, il n'existe pas de méthode idéale de décomposition mais plusieurs s'adaptant potentiellement aux données acquises ainsi qu'à leurs structures.

7.1 Résolution multivariée de courbes par les moindres carrés alternés (MCR-ALS)

Au cours de cette partie, nous développons la méthode de résolution multivariée de courbes MCR-ALS (*Multivariate Curve Resolution – Alternating Least Squares*). Elle est choisie ici pour la qualité des extractions qu'elle propose et sa grande adaptabilité puisqu'elle est largement appliquée à toutes les spectroscopies moléculaires et les techniques séparatives... tout en considérant différents types d'expérience comme la résolution temporelle, l'imagerie, voire les deux à la fois.

Le père de la méthode MCR-ALS est le D^r Roma Tauler, chercheur à l'IDAEAC-SIC de Barcelone. Depuis plus de vingt ans, il en améliore ses fonctionnalités et ses performances [53,54]. La méthode MCR-ALS décompose les spectres étudiés dans un modèle bilinéaire de contributions pures. En fait cette décomposition exprime le maximum de variance des données spectrales initiales. Plus encore les profils (spectraux et de concentration) extraits par MCR-ALS ont un sens chimique, ce qui est rarement le cas dans les méthodes chimiométrique en général. On dit souvent que la méthode MCR-ALS essaye de résoudre le problème général de l'analyse des mélanges (aussi appelé « démélange » en traitement du signal).

7.1.1 Modèle bilinéaire

L'exemple le plus connu et le plus simple de modèle bilinéaire est la loi de Beer-Lambert. Ainsi un spectre d'un mélange contenant deux espèces A et B est une somme pondérée des deux spectres purs. Sur la figure 13, nous voyons les différentes manières d'exprimer cette bilinéarité. Ainsi en considérant par exemple l'absorbance d'un échantillon i à une longueur j, on montre que x_{ij} est une somme pondérée des absorbances des produits purs à cette même longueur d'onde (Fig. 13a). Ce modèle bilinéaire peut aussi se décliner dans des formes de plus en plus compactes soit composé par composé (Fig. 13b) ou finalement sous une forme matricielle communément utilisée pour la présentation de la méthode (Fig. 13c).

On montre alors que cette matrice $\mathbf{X}$ de données spectrales possédant l'ensemble des spectres suivant ses lignes peut se décomposer en une matrice de concentrations $\mathbf{C}$ et une matrice de spectres $\mathbf{S}$. On retrouve ainsi dans la première colonne de la matrice $\mathbf{C}$ les concentrations de l'espèce A dans tous les mélanges étudiés, la deuxième colonne étant associée à l'espèce B pour notre exemple. Parallèlement à cela, la première ligne de la matrice $\mathbf{S}$ donne le spectre du produit pur A et la deuxième ligne celui de l'espèce B.

Published in : SIMON G. (dir.) (2020), *Spectroscopies vibrationnelles. Théorie, aspects pratiques et applications*, Editions des archives contemporaines, ISBN : 9782813002556, doi : https ://doi.org/10.17184/eac.9782813002556

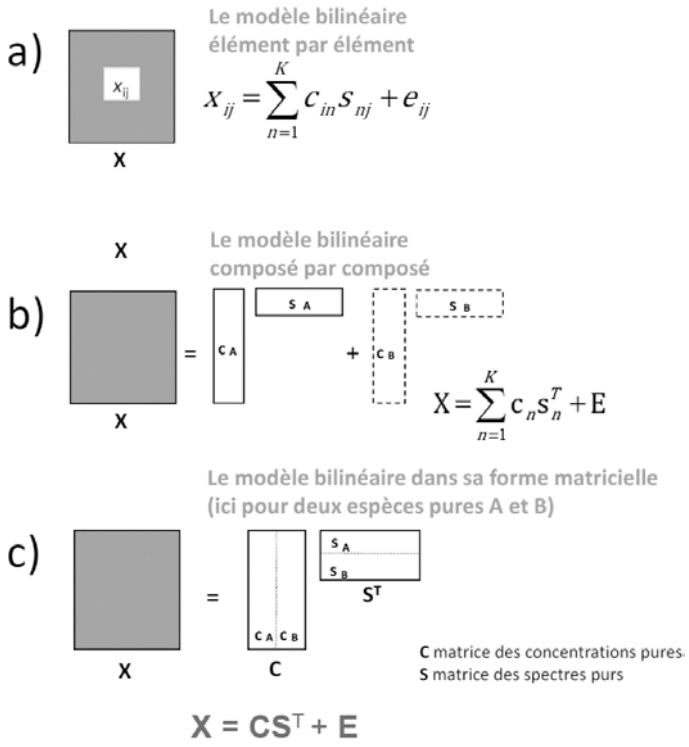


Figure 13 – Représentation du modèle bilinéaire.

7.1.2 Que puis-je faire avec la méthode MCR-ALS ?

L'objectif de cette partie est de donner des exemples concrets d'application de la méthode MCR-ALS et de démontrer le potentiel de la méthode sur différentes structures de données. En premier lieu et même si cela peut paraître trivial, la méthode MCR-ALS travaille sur un ensemble de spectres de mélanges. Nous pouvons d'ailleurs généraliser cette remarque à l'ensemble des méthodes chimiométriques. La figure 14 présente une première application possible de la méthode sur des données dites résolues en temps, c'est-à-dire que nous observons un système par l'acquisition de spectres en fonction du temps. On retrouve ce principe pour des études en photochimie ou comme dans le cas présenté ici pour les techniques séparatives. Dans le cas présenté, les spectres acquis en fonction du temps sont placés suivant les lignes de la matrice **X** de manière ordonnée. L'application de la méthode MCR permet tout d'abord, uniquement sur la base de la matrice **X**, d'évaluer le nombre d'espèces chimiques présentes dans le système chimique (ici 3) puis d'extraire simultanément les matrices **C** et **S**. Chaque colonne de la matrice **C** nous donne ainsi les profils des trois espèces pures alors que la matrice **S** nous donne les spectres des produits purs associés. Cet exemple nous montre d'ailleurs que même si la séparation chromatographique n'est pas assurée au niveau expérimental, l'application de la méthode chimiométrique MCR nous permet de séparer ces informations chimiques. Vous retrouvez des exemples d'appli-

Published in : SIMON G. (dir.) (2020), *Spectroscopies vibrationnelles. Théorie, aspects pratiques et applications*, Editions des archives contemporaines, ISBN : 9782813002556, doi : https ://doi.org/10.17184/eac.9782813002556

cation de MCR-ALS en chromatographie dans les références [55,56] ou encore pour des suivis temporels de réactions chimiques [57].

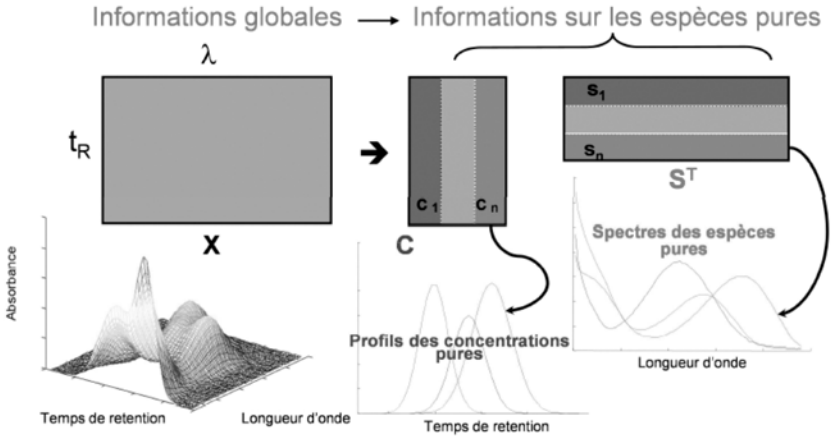


Figure 14 – La méthode MCR-ALS appliquée à une technique séparative.

Néanmoins la méthode MCR-ALS ne nécessite pas nécessairement que les spectres de la matrice **X** soient ordonnés. Il suffit juste que l'ensemble des spectres que nous considérons dans la matrice **X** partage des espèces chimiques. La figure 15 représente ainsi une matrice **X** fruit de la mise en commun de spectres de mélanges préparés en laboratoire. Afin d'imager l'application de MCR-ALS à cette structure de données particulière, je vous invite à lire la publication centrée sur l'analyse des couches d'hydratation par spectroscopie dans le système eau / éthanol [58]. La méthode MCR-ALS a aussi révolutionné la manière d'analyser les données issues d'expériences d'imagerie hyperspectrale. La figure 16 représente tout d'abord l'acquisition d'un cube de données hyperspectrales. Un microscope couplé à un spectromètre permet d'analyser une micro-zone d'analyse pour laquelle nous avons un spectre complet. Si le détecteur est monocanal (constitué d'un seul élément), une table motorisée est utilisée afin d'analyser de manière successive toute la surface d'intérêt de notre échantillon. Un cube de données hyperspectrales est généré à la fin de cette procédure. Lorsqu'un détecteur multicanal est utilisé, le cube est généré en une seule acquisition sans déplacer l'échantillon.

Afin de générer une cartographie chimique du composé d'intérêt, il est ensuite nécessaire d'intégrer le signal à une longueur d'onde spécifique. Cela revient à prendre une « tranche » du cube de données hyperspectrales. Même si cette méthode est largement utilisée pour toutes les spectroscopies, on oublie souvent que cette notion de spécificité spectrale est une hypothèse forte. En effet, elle ne peut normalement être assurée que si nous connaissons tous les produits présents dans l'échantillon. Ainsi si un composé est oublié ou n'est pas attendu, la longueur d'onde sélectionnée peut ne pas être spécifique du produit d'intérêt induisant des cartographies chimiques biaisées présentant souvent des surconcentrations. Cette méthode classique ne permet pas non plus de générer une cartographie pour un composé inconnu. La figure 17 montre que

Published in : SIMON G. (dir.) (2020), *Spectroscopies vibrationnelles. Théorie, aspects pratiques et applications*, Editions des archives contemporaines, ISBN : 9782813002556, doi : https ://doi.org/10.17184/eac.9782813002556

la méthode MCR-ALS appliquée à l'analyse d'un cube de données hyperspectrales permet de supprimer ces contraintes.

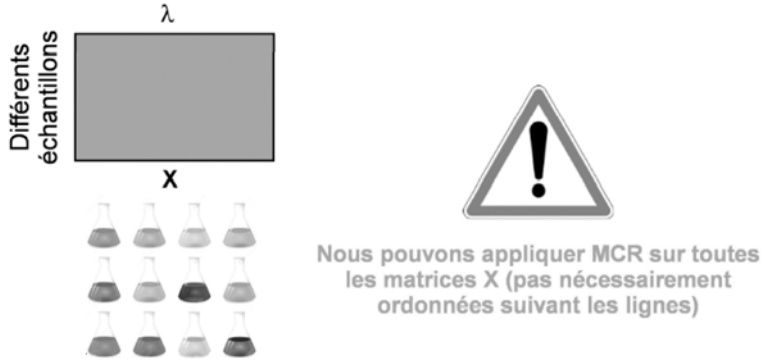


Figure 15 – La méthode MCR-ALS s'applique aussi sur des spectres non-ordonnés.

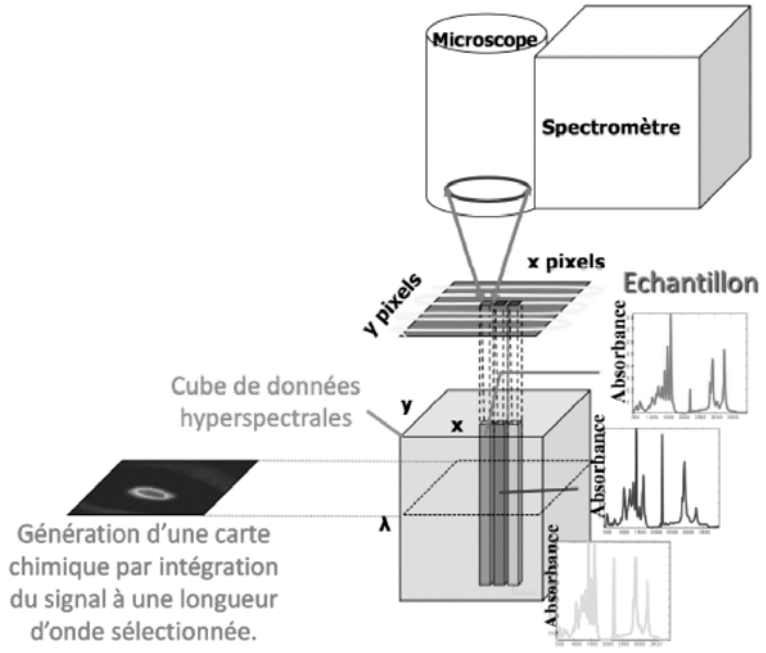


Figure 16 – Génération d'un cube de données hyperspectrales pour l'imagerie spectroscopique.

Published in : SIMON G. (dir.) (2020), *Spectroscopies vibrationnelles. Théorie, aspects pratiques et applications*, Editions des archives contemporaines, ISBN : 9782813002556, doi : https ://doi.org/10.17184/eac.9782813002556

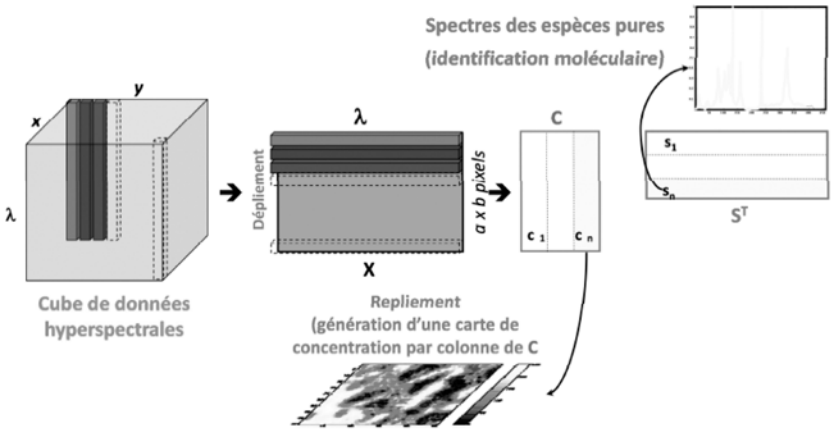


Figure 17 – La méthode MCR-ALS appliquée aux données d'imagerie hyperspectrale.

Dans une première étape, le cube est déplié pour donner notre matrice de spectres **X** habituelle car la méthode MCR ne peut pas gérer directement un tel cube. On retrouve ainsi les spectres des pixels suivant les lignes de la matrice **X**. Après une estimation du nombre d'espèces pures, la méthode MCR extrait les matrices **C** et **S**. Chaque colonne de la matrice **C** peut être ensuite repliée pour générer une cartographie chimique, les spectres des espèces pures correspondant se retrouvant suivant les lignes de **S**. Ils seront utilisés pour l'identification moléculaire.

7.1.3 Les avantages et les limites de la méthode MCR-ALS

L'avantage certainement le plus important pour l'application de la méthode MCR est qu'il n'est pas nécessaire d'avoir un *a priori* sur le système physicochimique étudié. Nous pourrions même explorer des systèmes complètement inconnus. Il n'est donc pas nécessaire d'avoir des informations qualitatives, quantitatives ou même le nombre d'espèces présentes. Par contre si des informations pertinentes sur le système sont disponibles, elles pourront être utilisées par MCR pour améliorer la décomposition (exemple : application de contraintes développées plus loin dans le texte). Enfin cette méthode présente une grande flexibilité puisqu'elle est applicable à de nombreuses matrices, de nombreuses techniques analytiques (au-delà des spectroscopies) et de nombreux domaines.

Une première limite de la méthode MCR pourrait se résumer par la phrase : « Nous pouvons décomposer ce que nous pouvons mesurer. » Néanmoins cette limite n'est pas spécifique à MCR puisqu'elle est généralisable à l'ensemble des techniques chimiométriques. Les constituants chimiques doivent effectivement avoir un signal mesurable pour la technique analytique sélectionnée. La chimiométrie peut souvent nous émerveiller mais ce n'est en aucun cas de la magie. La limite la plus importante est que tous les composés doivent avoir des « dyades » profils de concentration / profils spectraux distincts. Ainsi par exemple, si deux composés chimiques (avec des spectres différents) ont des concentrations corrélées c'est-à-dire par exemple que la concentration du premier composé est toujours deux fois celle du second, la méthode MCR ne pourra pas

Published in : SIMON G. (dir.) (2020), *Spectroscopies vibrationnelles. Théorie, aspects pratiques et applications*, Editions des archives contemporaines, ISBN : 9782813002556, doi : https ://doi.org/10.17184/eac.9782813002556

les extraire séparément. De la même manière, si les concentrations de deux composés chimiques sont toujours différentes mais qu'ils ont des spectres purs identiques, MCR ne pourra pas non plus les extraire séparément.

7.1.4 Pourquoi l'analyse ACP ne peut pas se substituer à MCR-ALS

Précédemment, nous avons vu que l'ACP était un outil puissant d'exploration des données. L'ACP propose aussi une décomposition de la matrice **X** en **T** et **P** respectivement matrices de *scores* et de *loadings* (Fig. 18).

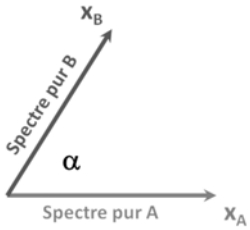


Figure 18 – Les décompositions proposées par l'ACP et MCR-ALS.

Lors de l'interprétation d'une composante principale, on essaye souvent de l'associer aux contributions spectrales d'un composé chimique spécifique. Même si un composé chimique peut effectivement contribuer majoritairement à une composante principale, cette dernière ne peut jamais être pure. Pourquoi l'ACP ne permet donc pas d'extraire des spectres d'espèces pures ? Pour répondre à cette question, considérons simplement le cas de deux spectres de produits purs A et B (Fig. 19).

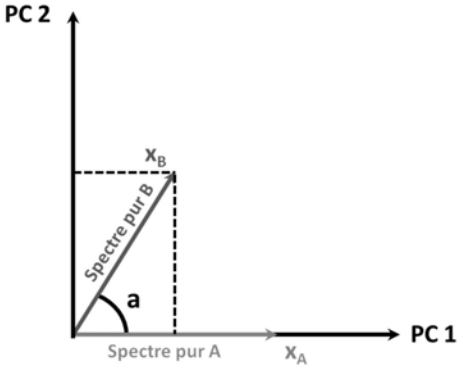


Figure 19 – Représentation vectorielle de deux spectres de composés purs.

Ainsi quels que soit les spectres purs sélectionnés dans une base de données spectrales par exemple, nous avons toujours un angle a différent de 90° puisque les spectres ne sont jamais complètement décorrélés. Or par construction les composantes principales sont toutes deux à deux orthogonales. Ainsi dans l'exemple présenté dans la figure 20 : Les composantes principales et les spectres purs, les contributions spectrales du produit B se retrouvent inévitablement dans les deux composantes principales.

Published in : SIMON G. (dir.) (2020), *Spectroscopies vibrationnelles. Théorie, aspects pratiques et applications*, Editions des archives contemporaines, ISBN : 9782813002556, doi : https ://doi.org/10.17184/eac.9782813002556

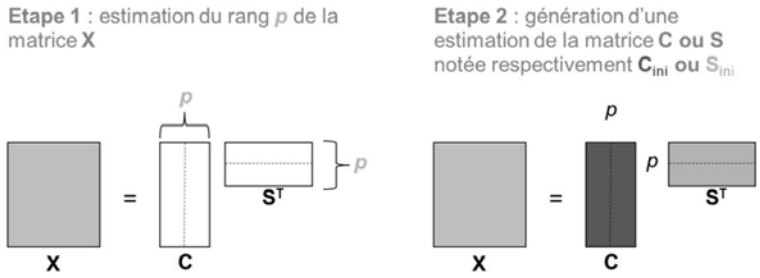


Figure 20 – Les composantes principales et les spectres purs.

7.1.5 Mise en œuvre de la méthode MCR-ALS

Description générale

L'objectif de cette partie est de présenter brièvement les étapes nécessaires à la mise en place d'une analyse MCR-ALS. Elles seront ensuite détaillées dans les paragraphes suivants.

La première étape est l'estimation de ce que l'on appelle le rang de la matrice **X**. C'est en fait le nombre de lignes ou de colonnes indépendantes dans la matrice **X** (Fig. 21).

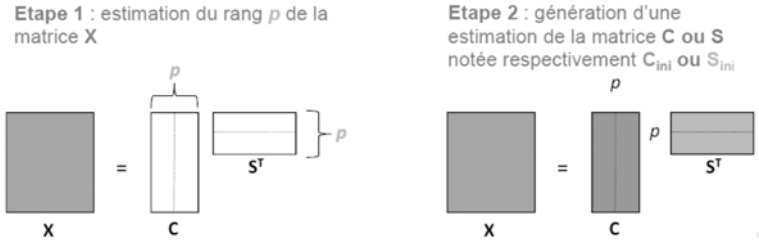


Figure 21 – Les deux premières étapes nécessaires à la décomposition MCR-ALS.

Ainsi par exemple, si une matrice **X**, constituée de vingt spectres (lignes) a un rang de deux, alors chaque spectre de cette matrice est une combinaison linéaire de deux lignes. Dans notre cas, nous utiliserons ce rang p comme une estimation du nombre d'espèces pures dans le système étudié.

La seconde étape est la génération d'une estimation initiale soit de la matrice **C** soit de la matrice **S** appelées respectivement $\mathbf{C_{ini}}$ ou $\mathbf{S_{ini}}$. Cette estimation ne peut pas être aléatoire et se doit d'être le plus proche de la réalité. Nous détaillerons plus tard quelques méthodes à notre disposition.

La troisième étape est le calcul MCR-ALS à proprement parler. Il est représenté par la figure 22 pour laquelle nous considérons que nous avons à disposition la matrice initiale $\mathbf{S_{initiale}}$ (❶). Ayant donc à disposition cette matrice initiale et la matrice de données expérimentales **X**, il est possible de calculer une estimation de la matrice des concentrations $\mathbf{C_{estimée}}$ (❷).

Published in : SIMON G. (dir.) (2020), *Spectroscopies vibrationnelles. Théorie, aspects pratiques et applications*, Editions des archives contemporaines, ISBN : 9782813002556, doi : https ://doi.org/10.17184/eac.9782813002556

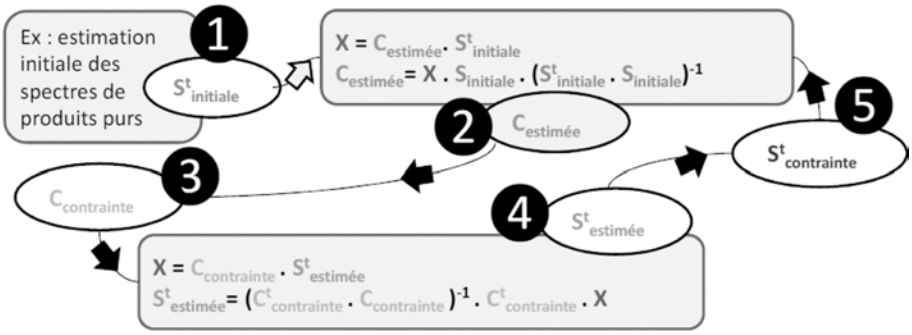


Figure 22 – Le calcul MCR-ALS en détails.

Afin de corriger et améliorer cette matrice, nous appliquons des contraintes pour obtenir une nouvelle matrice $\mathbf{C_{contrainte}}$ (❸). C'est à cette étape que nous entrons réellement dans les régressions alternées (*Alternating Least Squares*). En effet, ayant à disposition cette matrice $\mathbf{C_{contrainte}}$ et notre matrice de données $\mathbf{X}$, il est maintenant possible de calculer une matrice estimée des données spectrales $\mathbf{S_{estimée}}$ (❹). À son tour, cette matrice peut être corrigée en appliquant de nouvelles contraintes pour fournir la matrice $\mathbf{S_{contrainte}}$ (❺). Cette dernière matrice est ensuite réutilisée dans l'étape ❷ à la place de la matrice $\mathbf{S_{initiale}}$. Le calcul itératif de $\mathbf{C}$ puis de $\mathbf{S}$ permet de tendre vers des solutions de plus en plus affinées.

Évaluation du rang

Avant de présenter les méthodes d'estimation du rang, il est important de faire quelques remarques générales. Contrairement à l'ACP, MCR-ALS n'est pas une méthode imbriquée (*nested*). Ainsi dans une ACP calculée sur six composantes principales, les trois premières sont celles que nous obtenons si l'ACP est faite directement sur trois composantes. Il est donc primordial d'évaluer avec précision le rang de la matrice $\mathbf{X}$ lors de l'utilisation de la méthode MCR-ALS au risque d'extraire de mauvais profils pour les matrices $\mathbf{C}$ et $\mathbf{S}$. Néanmoins il n'existe pas de formule mathématique ultime qui permette de trouver le rang d'une matrice bruitée. Voici donc quelques conseils pour cette étape :

- Ne pas choisir le rang d'une matrice sur la base d'un pourcentage de variance exprimée dans une ACP (c'est une pratique assez courante). En effet vous ne connaissez jamais la variance du bruit de vos données expérimentales.

- Ne pas majorer l'estimation du rang en ayant pour objectif d'aller à la « pêche » aux profils potentiellement identifiés dans les extractions à la fin des calculs MCR (ex : « J'ai choisi un rang de 5 et les espèces 2 et 4 me semblent intéressantes »).

- Lorsque l'évaluation du rang est difficile (variance d'une espèce proche de la variance du bruit), il est préférable de sélectionner une fenêtre de rang (ex : 3 à 4). Nous faisons ensuite les calculs MCR pour ces différentes possibilités. La

Published in : SIMON G. (dir.) (2020), *Spectroscopies vibrationnelles. Théorie, aspects pratiques et applications*, Editions des archives contemporaines, ISBN : 9782813002556, doi : https ://doi.org/10.17184/eac.9782813002556

validation finale du rang se fait ensuite sur la base de la pertinence chimique des extractions finales. On trouve effectivement souvent des incohérences sur les profils de concentrations ou spectraux.

La méthode la plus utilisée pour estimer le rang d'une matrice est la décomposition en valeurs singulières (SVD). Cet algorithme est d'ailleurs souvent utilisé pour le calcul de l'ACP. En effet la première valeur singulière élevée au carré est la variance exprimée par la première composante principale, la seconde... Regardons le cas le plus simple, à savoir l'estimation du rang d'une matrice non bruitée. C'est tout simplement le nombre de valeurs singulières non nulles. Dans l'exemple présenté en figure 23, le rang de la matrice $\mathbf{X}$ est donc de 4, c'est-à-dire que seules 4 contributions pures sont observables dans ces données. Chaque ligne de la matrice $\mathbf{X}$ est donc potentiellement une combinaison linéaire de ces quatre contributions.

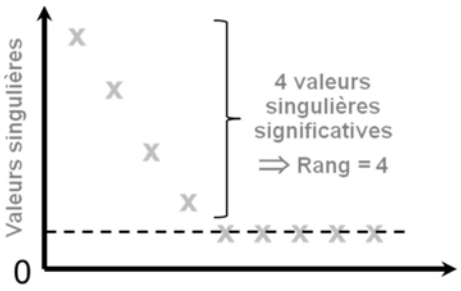


Figure 23 – Décomposition en valeurs singulières d'une matrice non bruitée.

Dans le cas d'une véritable matrice de données expérimentales donc bruitée, il faut trouver un seuil au-dessus duquel les valeurs singulières sont considérées comme significatives. Dans l'exemple présenté en figure 24, on remarque que le rang est toujours de 4, mais que les valeurs singulières supérieures à 4 expriment la variance du bruit.

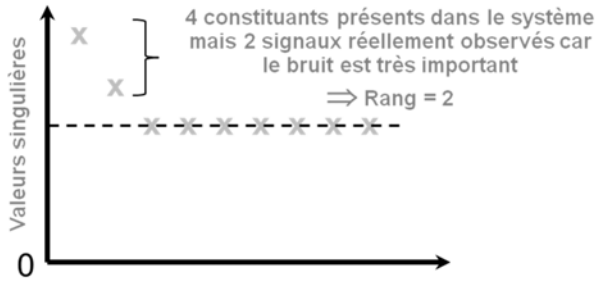


Figure 24 – Décomposition en valeurs singulières d'une matrice de données expérimentales.

Si on considère le cas précédent avec potentiellement quatre espèces pures présentes dans le système mais que les données expérimentales sont encore plus bruitées (issues de mesures plus rapides par exemple), on observe ce qui est présenté en figure 25. Le bruit est ici tellement important que seuls deux signaux indépendants sont observés, les variances de deux autres espèces minoritaires étant maintenant inférieures à la

Published in : SIMON G. (dir.) (2020), *Spectroscopies vibrationnelles. Théorie, aspects pratiques et applications*, Editions des archives contemporaines, ISBN : 9782813002556, doi : https ://doi.org/10.17184/eac.9782813002556

variance du bruit. Cet exemple permet de comprendre qu'il est impossible de fixer un rang sur la base d'une variance totale exprimée. Nous voyons aussi que nous ne pouvons jamais dire : « Je sais que dans le système il y a x composés. »

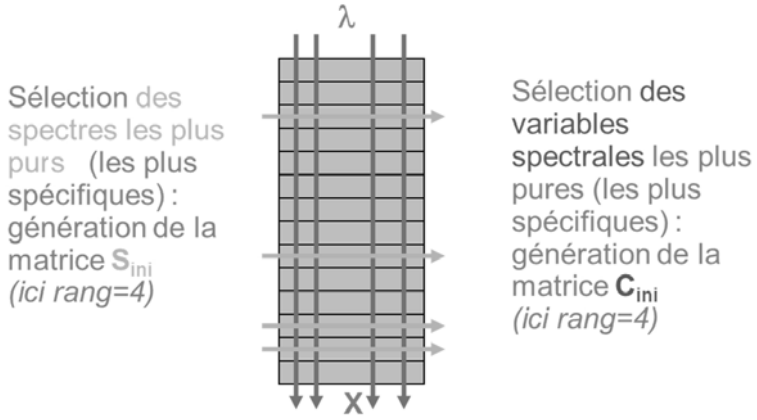


Figure 25 – Influence du bruit sur l'évaluation du rang de la matrice $\boldsymbol{X}$.

Génération d'une estimation initiale de $\boldsymbol{C}$ ou $\boldsymbol{S}$

Avant de rentrer dans les méthodes de génération d'estimations initiales, voici quelques conseils et remarques pour cette étape :

- Ne pas hésiter à tester MCR-ALS en générant des estimations initiales de **C** ou **S** car nous ne savons jamais laquelle donnera la meilleure solution.
- Il n'existe pas de méthode idéale mais un ensemble de méthodes possibles pour cette génération.
- Avoir en tête que plus les estimations initiales sont proches de la réalité, plus les extractions finales issues de MCR sont pertinentes.

La méthode la plus utilisée pour générer des estimations initiales est certainement la méthode SIMPLISMA (*SIMPLe-to-use Interactive Self-modeling Mixture Analysis*) [59]. Elle recherche les variables les plus pures, c'est-à-dire les plus spécifiques dans la matrice de données **X**. La figure 26 montre comment nous pouvons générer des estimations initiales. Ainsi par exemple, à partir de **X**, la méthode SIMPLISMA peut sélectionner les variables spectrales les plus pures pour générer la matrice $\mathbf{C_{ini}}$. Dans le cas présenté, 4 longueurs d'onde sont sélectionnées car le rang est fixé à 4. On retrouve ainsi 4 profils de concentrations suivant les colonnes de $\mathbf{C_{ini}}$. De la même manière, SIMPLISMA peut travailler suivant les lignes afin d'obtenir les 4 spectres les plus purs contenus dans **X** que nous retrouverons dans $\mathbf{S_{ini}}$. Même si cette méthode SIMPLISMA peut répondre à la majorité des cas, il est parfois intéressant d'exploiter une structure ordonnée des spectres (ex : suivi temporel). On utilise alors des méthodes comme l'EFA (*Evolving Factor Analysis*) [60].

Published in : SIMON G. (dir.) (2020), *Spectroscopies vibrationnelles. Théorie, aspects pratiques et applications*, Editions des archives contemporaines, ISBN : 9782813002556, doi : https ://doi.org/10.17184/eac.9782813002556

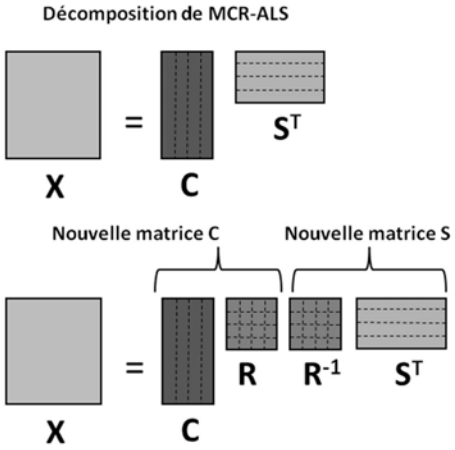


Figure 26 – Extraction de matrices initiales avec SIMPLISMA.

Application des contraintes

Lors de la présentation générale de la méthode MCR-ALS, nous avons vu qu'il était possible d'appliquer des contraintes aux matrices lors des régressions alternées. Une contrainte est tout d'abord une caractéristique que doivent suivre les profils (spectraux ou de concentration) des espèces. C'est aussi une manière de traduire les propriétés chimiques ou physiques de nos signaux en propriétés mathématiques. En fait elle permet de transformer la méthode de résolution mathématique MCR-ALS en outil d'analyse de données centré sur le système chimique. L'application d'une contrainte est optionnelle mais elle doit toujours être justifiée. Il ne faut surtout pas que cela soit une liste de vœux. Nous pouvons appliquer simultanément plusieurs contraintes sur la même matrice. Les deux matrices **C** et **S** peuvent bien entendu être contraintes. L'objectif principal de l'application de contraintes est de réduire le nombre de solutions possibles pour **C** et **S** c'est-à-dire réduire ce que nous appelons l'ambiguïté rotationnelle. De cette manière nous essayons aussi de tendre vers l'unicité des solutions de **C** et **S**. La figure 27 présente la problématique de l'ambiguïté rotationnelle. Ainsi même si vous partez d'une décomposition de **X** d'une certaine qualité, il est toujours possible d'insérer une matrice **R** dite de rotation qui ne changera pas cette qualité mais qui transformera potentiellement les matrices **C** et **S**. Les contraintes sont donc aussi utilisées pour réduire ces rotations. Il existe aussi une ambiguïté dite d'intensité. De la même manière, si vous partez d'une décomposition de **X** d'une certaine qualité, il est toujours possible d'insérer une constante k qui ne changera pas cette qualité. On note donc que seules les contributions relatives des espèces seront extraites par MCR-ALS. Afin de fixer une valeur de k, nous normalisons souvent les spectres extraits par leurs normes.

Les principales contraintes appliquées lors d'un calcul MCR-ALS sont présentées dans la figure 29. Dans nos exemples, elles sont appliquées à la matrice des concentrations **C** mais cela peut l'être aussi à la matrice des spectres **S**. La contrainte la plus utilisée est la non-négativité qui est naturelle pour de nombreux signaux. Comme son nom

Published in : SIMON G. (dir.) (2020), *Spectroscopies vibrationnelles. Théorie, aspects pratiques et applications*, Editions des archives contemporaines, ISBN : 9782813002556, doi : https ://doi.org/10.17184/eac.9782813002556

l'indique, les valeurs négatives dans les profils ne sont pas autorisées et sont donc corrigées. La seconde contrainte est l'unimodalité (*unimodality*), pour laquelle seul un maximum est autorisé par profil. On retrouve souvent cela dans les suivis cinétiques où les espèces apparaissent puis disparaissent sans jamais réapparaitre.

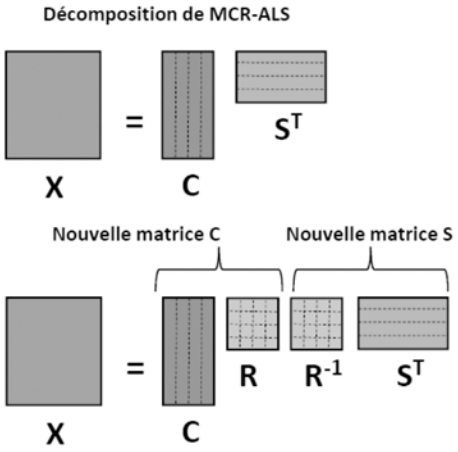


Figure 27 – L'ambiguïté rotationnelle dans MCR-ALS.

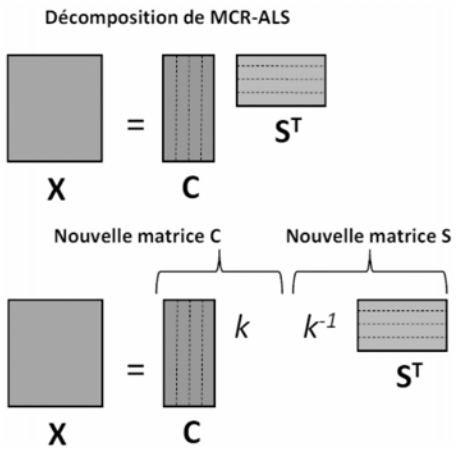


Figure 28 – L'ambiguïté d'intensité dans MCR-ALS.

Published in : SIMON G. (dir.) (2020), *Spectroscopies vibrationnelles. Théorie, aspects pratiques et applications*, Editions des archives contemporaines, ISBN : 9782813002556, doi : https ://doi.org/10.17184/eac.9782813002556

$$\mathbf{X} = \mathbf{C}\mathbf{S}^t + \mathbf{E}$$

- Le lack of fit (%)

$$lof = 100\sqrt{\frac{\sum_{i=1}^{N}\sum_{j=1}^{P} e_{i,j}{}^2}{\sum_{i=1}^{N}\sum_{j=1}^{P} x_{i,j}{}^2}} \qquad \text{avec} \qquad e_{i,j} = x_{i,j} - \hat{x}_{i,j}$$

- Variance exprimée par la décomposition MCR (%)

$$R^2 = 100\frac{\sum_{i=1}^{N}\sum_{j=1}^{P} x_{i,j}{}^2 - \sum_{i=1}^{N}\sum_{j=1}^{P} e_{i,j}{}^2}{\sum_{i=1}^{N}\sum_{j=1}^{P} x_{i,j}{}^2}$$

Figure 29 – Les principales contraintes utilisées dans MCR-ALS.

La troisième contrainte est dite du système fermé (*closure*). Dans ce cas, la somme des concentrations des composés purs est constante (c_{total}) pour tous les spectres de mélanges. Cette contrainte reste difficile à appliquer. La dernière contrainte qui peut être très efficace est la contrainte de sélectivité (*selectivity*). Dans ce cas, notre connaissance du système chimique nous permet d'affirmer que certaines valeurs sont nulles. Ainsi dans le cadre d'une cinétique par exemple, on peut dire qu'une espèce est absente à partir d'un certain temps jusqu'à la fin de la réaction. Cette liste n'est bien entendu pas exhaustive. Vous pouvez même imaginez une nouvelle contrainte s'adaptant particulièrement à vos données expérimentales. En conclusion, l'application de contraintes judicieusement choisies permet d'assurer la pertinence des extractions issues de la méthode MCR-ALS.

Critères d'arrêt du calcul itératif et convergence des calculs

Lors de la présentation du calcul MCR-ALS à proprement parler, nous avons vu que les matrices **C** et **S** étaient calculées de manière alternée puis affinées au cours des itérations. Il est ainsi intéressant de définir des critères d'arrêt du calcul itératif. La manière la plus simple est de convenir au préalable d'un nombre maximum d'itérations fixé en général au-dessus de 50. Parallèlement à cela, on vérifie à chaque itération que le produit $\mathbf{C}\mathbf{S}^T$ calculé reproduit suffisamment la matrice de données **X**. Pour se faire, nous estimons une erreur de modélisation appelée *lack of fit* (LOF) ainsi que le pourcentage de variance exprimée par ces extractions (Fig. 30).

$$\mathbf{X} = \mathbf{C}\mathbf{S}^t + \mathbf{E}$$

- Le lack of fit (%)

$$lof = 100\sqrt{\frac{\sum_{i=1}^{N}\sum_{j=1}^{P} e_{i,j}{}^2}{\sum_{i=1}^{N}\sum_{j=1}^{P} x_{i,j}{}^2}} \qquad \text{avec} \qquad e_{i,j} = x_{i,j} - \hat{x}_{i,j}$$

- Variance exprimée par la décomposition MCR (%)

$$R^2 = 100\frac{\sum_{i=1}^{N}\sum_{j=1}^{P} x_{i,j}{}^2 - \sum_{i=1}^{N}\sum_{j=1}^{P} e_{i,j}{}^2}{\sum_{i=1}^{N}\sum_{j=1}^{P} x_{i,j}{}^2}$$

Figure 30 – Évaluation de la qualité des extractions issues de MCR-ALS.

Published in : SIMON G. (dir.) (2020), *Spectroscopies vibrationnelles. Théorie, aspects pratiques et applications*, Editions des archives contemporaines, ISBN : 9782813002556, doi : https ://doi.org/10.17184/eac.9782813002556

Dans ces formules, $x_{i,j}$ représente la valeur expérimentale à la ligne i et la colonne j de la matrice $\mathbf{X}$, $\hat{x}_{i,j}$ l'élément correspondant calculé par le biais du produit $\mathbf{CS}^T$ et $e_{i,j}$ la différence entre ces deux éléments encore appelé résidu. Malheureusement il n'existe pas de valeur en dessous de laquelle les extractions sont considérées comme pertinentes car elles sont à mettre en relation avec le niveau de bruit des données expérimentales. Néanmoins le calcul itératif peut être arrêté si ces deux paramètres n'ont pas significativement changé depuis plusieurs itérations.

En conclusion, la méthode MCR-ALS permet de faire une décomposition sans *a priori* de la matrice des données $\mathbf{X}$ en $\mathbf{C}$ et $\mathbf{S}$. Ce concept nous permet donc d'explorer des systèmes chimiques inconnus. Nous pouvons aussi souligner que les profils spectraux et de concentration extraits ont un sens chimique ce qui n'est pas le cas de toutes les méthodes chimiométriques. Il est très important de passer du temps à explorer préalablement les données spectrales avec l'ACP même si elle ne propose pas une extraction d'espèces pures. Nous pouvons même généraliser ce conseil à toutes les approches chimiométriques. L'évaluation du rang et l'application des contraintes sont au centre de cette méthodologie. Il convient donc encore une fois de passer du temps sur ces étapes.

7.2 Analyse en composantes indépendantes (ICA)

La méthode MCR-ALS n'est pas la seule permettant de répondre à cette problématique de démélange. L'analyse en composantes indépendantes (ICA pour *Independant Component Analysis*) était initialement appliquée au traitement du signal, notamment en décomposition de signaux sonores. Cette méthode est basée sur la construction de variables latentes extraites de façon à ce qu'elles soient statistiquement aussi indépendantes que possible [61,62]. Ainsi, les variables indépendantes calculées permettent de représenter les « signaux purs » présents dans des lots de spectres de mélanges. Contrairement à l'approche MCR-ALS, l'analyse ICA permet de s'affranchir d'estimations initiales sur les espèces pures ou sur les profils de concentrations. Toutefois, l'inconvénient principal par rapport aux autres résolutions multivariées réside dans l'absence de contraintes pouvant être appliquées pour optimiser les résultats obtenus, induisant donc l'apparition d'aberrations physiques. Toute comme les algorithmes NIPALS et SIMPLS pour la régression PLS, il existe plusieurs algorithmes de calcul pour retrouver les variables indépendantes d'une analyse ICA. Les méthodes les plus couramment rencontrées sont Fast-ICA [63], JADE [64], MILCA ou encore SNICA [65]. Lorsqu'un spectroscopiste non initié à la chimiométrie recherche une méthode de résolution, il favorise souvent ICA car elle semble plus simple à mettre en œuvre. Cela reste un leurre car une mauvaise estimation du rang de la matrice des données spectrales génère des solutions tout aussi discutables.

8 Conclusion

Les outils chimiométriques présentent un fort potentiel pour les travaux portant sur des données multivariées, notamment les données spectrales. Les différentes techniques de prétraitement permettent de corriger ou d'atténuer certains effets indésirables présents sur les spectres. De plus, ces techniques permettent de conditionner les données

Published in : SIMON G. (dir.) (2020), *Spectroscopies vibrationnelles. Théorie, aspects pratiques et applications*, Editions des archives contemporaines, ISBN : 9782813002556, doi : https ://doi.org/10.17184/eac.9782813002556

afin d'améliorer les résultats apportés par les différentes méthodes chimiométriques présentées durant ce chapitre.

Tout d'abord, l'analyse exploratoire des données multivariées, notamment *via* l'analyse en composantes principales, permet de réellement condenser l'information en réduisant la dimensionnalité structurelle des jeux de données. Ainsi, il est plus facile de représenter les données et d'en tirer parti.

Les méthodes de classification permettent ensuite de réaliser différentes explorations. D'une part, les méthodes non-supervisées ont pour but de classer les différents échantillons des ensembles de données sans *a priori*, se basant souvent sur des études d'échantillons de proche en proche. D'autre part, les techniques supervisées vont permettre de réellement développer des modèles multivariés pour séparer différentes classes existantes. Un nouvel échantillon sera alors confronté à ces modèles pour vérifier son appartenance ou non à tel ou tel groupe.

Les techniques de régression ont également une place importante au sein des techniques chimiométriques. En effet, ces méthodes portent sur la détermination d'un paramètre à partir des données multivariées. Ces techniques ont perpétuellement évolué afin de pallier les limitations liées aux données spectroscopiques. Aujourd'hui, la régression PLS fait office de méthode phare, permettant de travailler sur des jeux de données présentant un grand nombre de variables, corrélées ou non.

Enfin, les méthodes de décomposition spectrale permettent aujourd'hui d'extraire, sans connaissance a priori, les spectres des produits purs et leurs concentrations associées au sein de mélanges complexes.

Aujourd'hui, l'amélioration des techniques instrumentales permet de générer toujours plus rapidement de grands volumes de données. L'emploi des outils chimiométriques est donc devenu primordial pour gérer efficacement les données recueillies et étudier de manière plus pertinente les différents procédés et systèmes chimiques ou biologiques mis en jeu. La chimiométrie dépasse maintenant très largement l'exploration des données spectrales en milieu industriel puisqu'elle fait aussi partie des outils des laboratoires de recherche universitaires.

Bibliographie

[1] Wold S. (1995). « Chemometrics ; what do we mean with it, and what do we want from it ? » *Chemometrics and Intelligent Laboratory Systems* 30 : 109-115.

[2] Massart D. L., Vandeginste B. G. M., Buydens L. M. C., De Jong S., Lewi P. J., and J. Smeyers-Verbeke (1997). « Handbook of Chemometrics and Qualimetrics : Part A. » *Data Handling in Science and Technology* Volume 20A. Elsevier : Amsterdam.

[3] Massart D. L., Vandeginste B. G. M., Buydens L. M. C., De Jong S., Lewi P. J., and J. Smeyers-Verbeke (1998). « Handbook of Chemometrics and Qualimetrics : Part B. » *Data Handling in Science and Technology* Volume 20B. Elsevier : Amsterdam.

[4] Rinnan, A., F. v. d. Berg, *et al.* (2009). « Review of the most common pre-processing techniques for near-infrared spectra. » *TrAC Trends in Analytical Chemistry* 28 : 1201-1222.

[5] J. Luypaert, S. Heuerding, Y. Vander Heyden, D. L. Massart (2004). « The effect of preprocessing methods in reducing interfering variability from near-infrared measurements of creams. » Journal of Pharmaceutical and Biomedical Analysis 36 : 495-503.

Published in : SIMON G. (dir.) (2020), *Spectroscopies vibrationnelles. Théorie, aspects pratiques et applications*, Editions des archives contemporaines, ISBN : 9782813002556, doi : https ://doi.org/10.17184/eac.9782813002556

[6] Barnes, R. J., M. S. Dhanoa, *et al.* (1989). « Standard Normal Variate Transformation and De-trending of Near-Infrared Diffuse Reflectance Spectra. » *Applied Spectroscopy* 43 : 772-777.

[7] Geladi, P., D. MacDougall, *et al.* (1985). « Linearization and Scatter-Correction for Near-Infrared Reflectance Spectra of Meat. » *Applied Spectroscopy* 39 : 491-500.

[8] Savitzky, A. and M. J. E. Golay (1964). « Smoothing and Differentiation of Data by Simplified Least Squares Procedures. » *Analytical Chemistry* 36 : 1627-1639.

[9] Norris, K. H. and P. C. Williams (1984). « Optimization of mathematical treatments of raw near-infrared signal in the measurement of protein in hard red spring wheat : I. Influence of particle size. » *Cereal Chemistry* 61 : 158-165.

[10] Zimmermann, B. and A. Kohlerb (2013). « Optimizing Savitzky-Golay Parameters for Improving Spectral Resolution and Quantification in Infrared Spectroscopy. » *Applied Spectroscopy* 67 : 892-902.

[11] Nørgaard, L., A. Saudland, *et al.* (2000). « Interval partial least-squares regression (iPLS) : A comparative chemometric study with an example from near-infrared spectroscopy. » *Applied Spectroscopy* 54 : 413-419.

[12] Centner, V., D.-L. Massart, *et al.* (1996). « Elimination of Uninformative Variables for Multivariate Calibration. » *Anal. Chem.* 68 : 3851-3858.

[13] Leardi, R. and A. L. González (1998). « Genetic algorithms applied to feature selection in PLS regression : how and when to use them. » *Chemometrics and Intelligent Laboratory Systems* 41 : 195-207.

[14] Abrahamsson, C., J. Johansson, *et al.* (2003). « Comparison of different variable selection methods conducted on NIR transmission measurements on intact tablets. » *Chemometrics and Intelligent Laboratory Systems* 69 : 3-12.

[15] Balabin RM1, Smirnov SV. (2011). « Variable selection in near-infrared spectroscopy : benchmarking of feature selection methods on biodiesel data. » *Analytica Chimica Acta* 692 : 63-72.

[16] Michel J. Anzanello, Flavio S. Fogliatto (2014). « A review of recent variable selection methods in industrial and chemometrics applications. » *European J of Industrial Engineering* 8 : 619.

[17] Wold, S., H. Antti, *et al.* (1998). « Orthogonal signal correction of near-infrared spectra. » *Chemometrics and Intelligent Laboratory Systems* 44 : 175-185.

[18] Andersson, C. A. (1999). « Direct orthogonalization. » *Chemometrics and Intelligent Laboratory Systems* 47 : 51-63.

[19] Westerhuis, J. A., S. De Jong, *et al.* (2001). « Direct orthogonal signal correction. » *Chemometrics and Intelligent Laboratory Systems* 56 : 13-25.

[20] Feudale, R. N., H. Tan, *et al.* (2002). « Piecewise orthogonal signal correction. » *Chemometrics and Intelligent Laboratory Systems* 63 : 129-138.

[21] Trygg, J. and S. Wold (2002). « Orthogonal projections to latent structures (O-PLS). » *Journal of Chemometrics* 16 : 119-128.

[22] Svensson, O., T. Kourti, *et al.* (2002). « An investigation of Orthogonal Signal Correction algorithms and their characteristics. » *Journal of Chemometrics* 16 : 176-188.

[23] McClure, W. F., K. H. Norris, *et al.* (1977). « Rapid spectrophotometric analysis of the chemical composition of tobacco. Part 1. Total reducing sugars. » *Beitr. Tabakforsch* 9.

[24] Wu, W., B. Walczak, *et al.* (1996). « Feature reduction by Fourier transform in pattern recognition of NIR data. » *Analytica Chimica Acta* 331 : 75-83.

[25] Daubechies, I. (1990). « The wavelet transform, time-frequency localization and signal analysis. » *IEEE transactions on information theory* 36 : 961-1005.

[26] Alsberg, B. K., A. M. Woodward, *et al.* (1997). « An introduction to wavelet transforms for chemometricians : A time-frequency approach. » *Chemometrics and Intelligent Laboratory Systems* 37 : 215-239.

[27] Azzouz, T., A. Puigdoménech, *et al.* (2003). « Comparison between different data pre-treatment methods in the analysis of forage samples using near-infrared diffuse reflectance spectroscopy and partial least-squares multivariate calibration method. » *Analytica Chimica Acta* 484 : 121-134.

[28] Jackson, J.E. (1981). « Principal Components and Factor Analysis : Part 1-Principal Components. » *Journal of Quality* Technology, 13.

Published in : SIMON G. (dir.) (2020), *Spectroscopies vibrationnelles. Théorie, aspects pratiques et applications*, Editions des archives contemporaines, ISBN : 9782813002556, doi : https ://doi.org/10.17184/eac.9782813002556

[29] MacGregor, J.F. and T. Kourti (1995). « Statistical process control of multivariate processes. » *Control Engineering Practice*, 3 : 403-414.

[30] De Maesschalck, R., D. Jouan-Rimbaud, *et al.* (2000). « The Mahalanobis distance. » *Chemometrics and Intelligent Laboratory Systems* 50 : 1-18.

[31] Hotelling, H. (1931). « The generalization of Student's ratio. » *Annals of Mathematical Statistics* 2 : 360-378.

[32] Rousseeuw, P. J. and M. Hubert (2011). « Robust statistics for outlier detection. » *Data Mining And Knowledge Discovery* 1 : 73-79.

[33] Halkidi, M., Y. Batistakis, *et al.* (2002). « Cluster validity methods : part I. » *ACM SIGMOD Record* 31 : 40-45.

[34] Jain, A. K., M. N. Murty, *et al.* (1999). « Data clustering : a review. » *ACM Comput. Surv.* 31 : 264-323.

[35] Johnson, S. C. (1966). « Hierarchical clustering schemes. » *Psychometrika* 32 : 241-254.

[36] MacQueen, J. (1967). *Some methods for classification and analysis of multivariate observations*. Fifth Berkeley Symposium on Mathematical Statistics and Probability.

[37] Ester, M., H.-P. Kriegel, *et al.* (1996). *A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise*. 2nd International Conference on Knowledge Discovery and Data Mining.

[38] Daszykowski, M., B. Walczak, *et al.* (2001). « Looking for natural patterns in data : Part 1. Density-based approach. » *Chemometrics and Intelligent Laboratory Systems* 56 : 83-92.

[39] Ankerst, M., M. M. Breunig, *et al.* (1999). *OPTICS : Ordering Points To Identify the Clustering Structure*. ACM SIGMOD Int. Conf. on Management of Data (SIGMOD'99), Philadelphia.

[40] Kohonen, T. (1990). *The self-organizing map*. IEEE 78.

[41] Kohonen, T. (1998). « The self-organizing map. » *Neurocomputing* 21(1-3) : 1-6.

[42] Jain, A. K. (2009). « Data clustering : 50 years beyond K-means. » *Pattern Recognition Letters in Press*, Corrected Proof.

[43] Wold, S., Martens, H., *et al.* (1983) « The multivariate calibration-problem in chemistry solved by the PLS method. » *Lecture Notes in Mathematics* 973 : 286-293.

[44] De Jong, S. (1993) « SIMPLS : an alternative approach to partial least squaures regression. » *Chemometrics and Intelligent Laboratory Systems* 18 : 251-263.

[45] Berrueta, L. A., R. M. Alonso-Salces, *et al.* (2007). « Supervised pattern recognition in food analysis. » *Journal of Chromatography* A 1158(1-2) : 196-214.

[46] Fisher, R. (1936). « The use of multiple measurements in taxonomic problems. » *Annals of Eugenics* 7 : 179-188.

[47] Devaux, M. F., D. Bertrand, *et al.* (1988). « Application of Multidimensional Analyses to the Extraction of Discriminant Spectral Patterns from NIR Spectra. » *Applied Spectroscopy* 42 : 941-1132.

[48] Novales, B., S. Guillaume, *et al.* (1998). « Particle size characterisation of in-flow milling products by video image analysis using global features. » *Journal of the science of food and agriculture* 78 : 187-195.

[49] Sirieix, A. and G. Downey (1993). « Commercial wheatflour authentication by discriminant analysis of near infrared reflectance spectra. » *Journal of Near Infrared Spectroscopy* 1 : 187-197.

[50] Wold, S. and M. Sjostrom (1977). « SIMCA : A Method for Analyzing Chemical Data in Terms of Similarity and Analogy – (Chapter Book). » *Chemometrics : Theory and Application* 52 : 243-282.

[51] Barker, M. and W. Rayens (2003). « Partial least squares for discrimination. » *Journal of Chemometrics* 17 : 166-173.

[52] Roussel, S., V. Bellon-Maurel, *et al.* (2003). « Authenticating white grape must variety with classification models based on aroma sensors, FT-IR and UV spectrometry. » *Journal of Food Engineering* 60 : 407-419.

[53] Tauler, R. (1995). « Multivariate curve resolution applied to second order data » *Chemometrics and Intelligent Laboratory Systems* 30 : 133-146.

[54] Tauler, R., Smilde, A., *et al.* (1995). « Selectivity, local rank, three-way data analysis and ambiguity in multivariate curve resolution. » *Journal of Chemometrics* 9 : 31-58.

Published in : SIMON G. (dir.) (2020), *Spectroscopies vibrationnelles. Théorie, aspects pratiques et applications*, Editions des archives contemporaines, ISBN : 9782813002556, doi : https ://doi.org/10.17184/eac.9782813002556

[55] Gargallo, R, Tauler, R., *et al.* (1996). « Validation of alternating least-squares multivariate curve resolution for chromatographic resolution and quantitation. » *TrAC Trends in Analytical Chemistry* 15 : 279-286.

[56] Mallat, E., Barcelo, D., *et al.* (1997). « Degradation study of benomyl and carbendazim in water by liquid chromatography and multivariate curve resolution methods. » *Chromatographia* 46 : 342-350.

[57] Gracia, N., Thomas, S., *et al.* (2011). « Combination of mid-infrared spectroscopy and curve resolution method to follow the antioxidant action of alkylated diphenylamines. » *Chemometrics and Intelligent Laboratory Systems* 106 : 210-215.

[58] Duponchel, L., Laurette, S., *et al.* (2013). « Terehertz microfluidic sensor for in situ exploration of hydration shell of molecules. » *Chemometrics and Intelligent Laboratory Systems* 123 : 28-35.

[59] Windig, W., Lippert, J. L., *et al.* (1990). « Interactive self-modeling multivariate analysis. » *Chemometrics and Intelligent Laboratory Systems* 9 : 7-30.

[60] Maeder, M. (1987). « Evolving factor analysis for the resolution of overlapping chromatographic peaks. » *Analytical Chemistry* 59 : 527-530.

[61] Comon, P. (1994). « Independent component analysis, a new concept. » *Signal Processing* 36 : 287-314.

[62] Hyvärinen, A., Karhunen, J., *et al.* (2001). *Independent Component Analysis.* John Wiley & Sons, Inc.

[63] Hyvärinen, A. (1999). « Fast and robust fixed-point algorithms for independent component analysis. » *Iee Transactions on Neural Networks* 140 : 362-370.

[64] Cardoso, J. F. and Souloumiac, A. (1993). « Blind beamforming for non-gaussian signals. » *Iee Proceedings-F Radar and Signal Processing* 140 : 362-370.

[65] Pasadakis, N., Kardamakis, A., *et al.* (2006). « Identifying constituents in commercial gasoline using Fourier transform-infrared spectroscopy and independent component analysis. » *Analytica Chimica Acta* 578 : 250-255.

Published in : SIMON G. (dir.) (2020), *Spectroscopies vibrationnelles. Théorie, aspects pratiques et applications*, Editions des archives contemporaines, ISBN : 9782813002556, doi : https ://doi.org/10.17184/eac.9782813002556