Food authentication by Proton NMR spectroscopy in combination with chemometric tools

Marion Cuny

N°/__/__/__/__/__/__/__/__/__/__/

THÈSE

pour obtenir le grade de

Docteur

de

l'Institut des Sciences et Industries du Vivant et de l'Environnement (Agro Paris Tech)

Spécialité : Chimie Analytique

présentée et soutenue publiquement par

Marion CUNY

le 7 mars 2008

AUTHENTICITE DES PRODUITS AGROALIMENTAIRES PAR SPECTROSCOPIE RMN ^{1}H ET OUTILS CHIMIOMETRIQUES

APPLICATION AUX JUS DE FRUIT, YAOURTS ET VINAIGRES.

Directeur de thèse : ***Douglas N. RUTLEDGE***

Codirectrice de thèse : ***Michèle LEES***

Travail réalisé dans l'UMR INRA / AgroParisTech, F-75005 Paris

«Ingénierie Analytique pour la Qualité des Aliments» - IAQA

Devant le jury :

Mme **Evelyne VIGNEAU**, Pr., UMR ENITIAA – INRA	..... **Présidente du jury**
Mme **Ana M. GIL,** Pr., Universidade de Aveiro, Portugal	..... **Rapporteur**
M. **Jean-Michel ROGER**, Pr., UMR Cemagref	..... **Rapporteur**
Mme **Michèle LEES,** Dr., Eurofins Scientific Analytics	..... **Examinateur**
M. **Idwin BOUMAN**, Friesland Foods, Pays-Bas	..... **Examinateur**
M. **Ian COLQHOUN**, Dr., Institute of Food Research, Angleterre	..... **Examinateur**
M. **Douglas N. RUTLEDGE,** Pr., AgroParisTech	..... **Examinateur**

i. Remerciements

Je souhaite commencer ces remerciements par les deux personnes qui m'ont suivi pendant trois ans et qui ont défendu ce sujet de thèse et ma candidature. Tout d'abord, je remercie vivement **Dr. Michèle Lees** pour sa confiance, sa disponibilité tout au long de ces trois années, et son soutien au quotidien même dans les moments plus difficiles. Ensuite, mes sincères remerciements vont au **Pr. Douglas N. Rutledge** pour les échanges d'idées "géniales", pour sa patience face à mon impatience, et son regard critique sur mes résultats. Tous deux m'ont permis de produire un travail soigné et méthodique et de publier.

Je remercie les rapporteurs de cette thèse les professeurs **Ana Gil et Jean-Michel Roger** pour l'intérêt qu'ils ont porté à mon travail. Merci également aux autres membres du jury qui ont accepté de juger ce travail : **Evelyne Vigneau, Ian Colquhoun et Idwin Bouman.**

Sans financement, cette thèse n'aurait pu voir le jour et je dois remercier les dirigeants des laboratoires Eurofins, et plus particulièrement **François Vigneau**, responsable du site de Nantes pour avoir permis la mise en place d'une convention CIFRE. De même, je remercie le Ministère de la Recherche pour avoir financer une grande partie de ma thèse.

Pour reprendre un ordre plus chronologique, je voudrais remercier un ami qui a joué un rôle fondamental dans ma décision de faire de la recherche, **Pr. Roland Poss** qui m'a fait découvrir ce petit monde à travers une mission à l'IRD en Thaïlande. Cette expérience m'a rendue enthousiaste pour la science et les voyages pour encore longtemps j'espère.

Durant ces trois années, j'ai eu la chance de croiser beaucoup de personnes qui m'ont apporté leur connaissance, leur expérience et expertise, leur aide, et/ou leur soutien et je souhaite leur exprimer ma reconnaissance.

J'ai eu la chance de bénéficier au début de ma thèse de l'expérience et de cours théorique du **Pr. Maryvonne Martin** qui par la suite a bien voulu consacrer un peu de son temps pour relire en partie mon manuscrit. Je souhaite donc lui adresser ma reconnaissance.

Ce sont **Dr. Gwénaëlle Le Gall** et **Pr. Ian Colquhoun** qui m'ont, les premiers, formé de façon pratique, à la spectroscopie RMN du proton et accueilli à l'IFR de Norwich au début de ma thèse. Par la suite, ils sont restés disponibles pour m'aider dans mes recherches et présenter nos résultats.

J'ai eu également le plaisir de collaborer avec **Pr. Dominique Bertrand** et **Dr. Evelyne Vigneau** de l'ENITIAA qui m'ont aidé dans le raisonnement sur l'analyse des données et l'expérimentation. Grâce à leurs conseils, j'ai pu avancer de façon significative dans la méthodologie et le traitement des signaux.

Je voudrais également remercier **Manfred Spraul, Eberhard Humpfer** et **Monika Moertter**, qui m'ont accueilli à Karlsruhe et m'ont permis d'approfondir mes connaissances en spectroscopie RMN du proton au centre Bruker BioSpin.

J'aimerais par ailleurs souligner la contribution importante du **Pr. Ana Gil**, rencontrée à MR in Food Science, qui m'a aidé, avec son équipe de l'Université d'Aveiro : **Dr. Iola Duarte** et ses étudiants : **Gonçalo** et **Juan**, à entrer dans la complexité de l'attribution des signaux non identifiés…

Mes plus vifs remerciements vont aux **Pr. Gérald Remaud** et **Serge Akoka**, du L A I E M de l'Université de Nantes, qui m'ont fait partager leur expertise et m'ont conseillé pour améliorer l'acquisition des données spectrales.

Mes plus chaleureux remerciements s'adressent à **Idwin Bouman** de Friesland Foods qui m'a envoyé des échantillons de yaourt et de préparation de fruits de qualité et composition connue. Ceci étant très important pour les analyses d'authenticité.

Ce travail de thèse n'aurait pas pu voir le jour sans l'aide répétée des personnes du SAV de Bruker : **Yves Renk, Olivier Assemat**, et leurs collègues qui m'ont, plus d'une fois, dépanné avec le spectromètre.

Je n'oublierai pas, non plus, les aides permanentes reçues de **l'ensemble du personnel du site d'Eurofins Nantes** dans les différents services.

Mes plus sincères remerciements vont également à mes collègues des services Recherche & Développement Chimie : **Eric, Anne-Sophie, Ellen, Lucie, Frédérique** et **Freddy**, et les personnes du service Recherche Collaborative : **Delphine** et **Céline**, ainsi que les gens de passage : **Nuria, Audrey, Pierre-Etienne, Christine, Mélanie, Guénaëlle, Mathieu, Sébastien**… qui en plus de leur aide m'ont permis d'évoluer dans une ambiance sympathique pour échanger des idées, un repas, un café ou une pause festive.

J'ai aussi eu l'opportunité de travailler sur un projet avec des personnes d'autres entités d'Eurofins comme **Martine Peinturier** et **Valérie Perennes** que je remercie chaleureusement pour leur collaboration.

Un grand merci aux personnes du laboratoire d'AgroParisTech que je n'ai certes pas vu souvent mais qui m'ont toujours réservé le meilleur accueil.

Je ne pensais pas avoir autant de personnes à remercier et je me rends compte que la liste exhaustive serait bien trop longue mais je souhaite ajouter une pensée pour ceux qui ont fait avancer cette thèse en me faisant prendre du recul par rapport aux difficultés rencontrées et en m'accompagnant dans les moments heureux : **ma famille** et **mes amis**.

ii. Résumé

La spectroscopie RMN ^{1}H est une technique largement utilisée en analyse et qui se développe en authenticité des produits agroalimentaires. Cependant, l'utilisation des techniques d'analyse multivariée des données RMN ^{1}H n'est pas encore aussi développée que dans le cas d'autres spectroscopies comme le Proche Infrarouge. Dans cette étude, nous avons utilisé différentes méthodes chimiométriques sur différents jeux de données issues de la spectroscopie RMN ^{1}H, ayant pour but de tester le potentiel de cette méthode pour établir l'authenticité des produits.

Tout d'abord, nous avons montré que l'utilisation de l'analyse en composantes indépendantes était un meilleur choix que l'analyse en composantes principales lors de l'analyse des spectres. En effet, les composantes indépendantes par leur nature expliquent mieux les phénomènes physico-chimiques mis en exergue lors de l'analyse spectrale. De plus, il a été possible de retrouver par cette méthode les signaux « purs » de composés, extraits des données brutes.

Ensuite, nous avons testé différents prétraitements tel que l'ourdissage (« warping ») des données qui s'est révélé utile lorsque les données présentées des décalages de pics. De même, la transformation logarithmique des données a montré son intérêt pour l'analyse globale du spectre étant donné les larges variations d'intensité rencontrées. Puis, nous avons testé différentes méthodes de sélection de variables. Les premières méthodes se basaient sur des critères relatifs aux données comme la variance et la covariance pour CLV (Clustering of variables). Le deuxième type de méthode sélectionnait des intervalles de données - ce qui permet de prendre en compte le lien entre variables successives dans un signal. Nous avons comparé une méthode reconnue, Interval_PLS (iPLS), avec des méthodes plus novatrices : Evolving Windows Zone Selection et Interval-PLS_Cluster.

Ces techniques de sélection de variables ont permis de repérer des marqueurs connus de l'authenticité des jus d'orange et de pamplemousse : hespéridine et naringine, mesurées en HPLC par la méthode IFU 58. De plus, les zones sélectionnées sur le vinaigre balsamique traditionnel rendent compte du vieillissement du vinaigre et permettent de discriminer le vinaigre balsamique commun du vinaigre traditionnel qui est un produit à forte valeur ajoutée. Enfin sur les yaourts, ce sont des composés aromatiques ainsi que des solvants de l'arôme qui

ont été détectés et ont permis de séparer les différents types de yaourts aromatisés, aux fruits, à la pulpe de fruit et à différentes concentrations.

Cette étude a montré l'intérêt du développement des techniques d'analyses spectrales en RMN pour prendre en compte ses spécificités par rapport aux autres techniques spectrales : les différences d'intensité dans les différentes zones du spectre, la taille importante des données initiales, ainsi que la redondance de l'information.

iii. Abstract

^{1}H NMR spectroscopy is widely used as an analytical method in different sectors. Its number of applications in the area of food and beverage authenticity is growing. However, the multivariate analysis of the type of data obtained from an NMR spectrum is still not as developed as in Near Infrared spectroscopy. In this work, different chemometric methods have been applied to ^{1}H NMR data in order to assess the potential of the combined techniques to authenticate food and beverages.

First, we have demonstrated that Independent Component Analysis (ICA) was better adapted to the analysis of a ^{1}H NMR spectrum than the commonly used Principal Component Analysis (PCA). Indeed, by its very nature, ICA aims at recovering pure sources from mixed signals. Applied to the spectra, it was shown that it is possible to extract certain component signals, such as the naringin signal from the spectra of a mixture of orange and grapefruit juices.

Different pre-treatments were then tested. Data warping has been found to be useful when the data shows variation in chemical shifts. In addition, as the data also varied considerably in intensity along the spectrum, a logarithmic transformation was performed to produce unbiased results when using other chemometric tools such as PCA and ICA.

Finally, different approaches were investigated to select variables in the spectrum. The first approach was based on criteria related to the variables themselves, such as the total variance and covariance in the Clustering of Variables (CLV) function. The second type of method involved the selection of contiguous variables to take into account the relation between variables in a signal. Interval_PLS (iPLS) was used as a reference to compare other more recently developed methods: Evolving Windows Zone Selection and Interval-PLS_Cluster.

The variable selection techniques used were to highlight known authenticity markers of orange and grapefruit juices: flavonoids hesperidin and naringin that are measured in the standard HPLC method IFU 58. In the case of the balsamic vinegar dataset, the selected zones contained the signal of compounds linked to the product's aging process that differentiate traditional balsamic vinegar from its cheaper, more commonly used counterpart. In the example based on different types of yoghurts, the variable selection procedures focused on

certain aroma compounds and solvents used as flavour carriers. These were used to differentiate the type of yoghurts: flavoured, with fruit, with pulp and at different concentrations.

Through these different applications, this study has shown the importance of using appropriate tools for spectral analysis tools that take into account the specificity of ^{1}H NMR spectroscopy compared to other spectroscopic methods: variation in intensities along the spectrum, the size of the dataset, and redundant information.

iv. Liste des publications

Articles

Cuny, M., Caligiani, A., Palla, G., Lees, M. & Rutledge, D. N. (2008). "Independent Component Analysis (ICA) of ^{1}H NMR spectra as a procedure for discrimination and prediction of balsamic vinegars production methods." submitted to Analytical and Bioanalytical Chemistry.

Cuny, M., Vigneau, E., Le Gall, G., Colquhoun, I. J., Lees, M. & Rutledge, D. N. (2008). "Fruit juice authentication by ^{1}H NMR spectroscopy in combination with different chemometrics tools." Analytical and Bioanalytical Chemistry **390**: 419-427.

Cuny, M., Le Gall, G., Colquhoun, I. J., Lees, M. & Rutledge, D. N. (2007). "Evolving Window Zone Selection method followed by Independent Component Analysis as useful chemometric tools to discriminate between grapefruit juice, orange juice and blends." Analytica Chimica Acta **597**: 203-213.

Cuny, M., Vigneau, E., Lees, M. & Rutledge, D. N. (2006). Dairy product authentication by ^{1}H NMR spectroscopy in combination with different chemometric tools. Magnetic resonance in food science. I. A. Farhat, P. S. Belton and G. A. Webb, RSC Publishing**:** 197-204.

Présentations orales

Chimiométrie 2007 : 28-29 novembre 2007, Lyon **:** Discrimination of balsamic vinegar production methods: Evolving Window Zone Selection and Interval PLS-Cluster on ^{1}H NMR spectra, followed by ICA.

World Juice 2007 : 11 octobre 2007, Barcelone : Fruit or aroma, what is the real taste? An analytical approach using ^{1}H NMR spectroscopy to determine the fruit content in yoghurts.

EurofoodChem 2007 : 29-31 août 2007, Paris : Determination of authenticity and typicality of different varieties of plain yoghurt by ^{1}H NMR spectroscopy.

Chimiométrie 2006 : 30 novembre - 1er décembre 2006, Paris : Comparaison de différentes méthodes de sélection de variables et d'analyses de données. Application à l'authentification de produits alimentaires par RMN ^{1}H.

Magnetic Resonance in Food Science 2006 : 16-19 juillet 2006 : Dairy product authentication by ^{1}H NMR in combination with different chemometric tools.

Posters

Chimiométrie 2006 : 30 novembre - 1er décembre 2006, Paris : Independent Component Analysis combined with ^{1}H NMR spectroscopy as a means to prevent fraud in the food market: application to fruit juices. Strawberry yoghurt classification by ^{1}H NMR spectroscopy in combination with different chemometric tools.

UFoST: International Union of Food Science and Technology World Congress: Food is Life : 17-21 septembre, Nantes : Chemometric method such as Independent Component Analysis combined with NMR spectroscopy used as tools for preventing fraud is the food market: application to fruit juices.

Journées doctorales ABIES 2006 : 8-9 mars 2006, Paris : Comparaison de méthodes chimiométriques : ACP et ACI sur des spectres RMN ^{1}H de jus d'orange et de pamplemousse.

Chimiométrie 2005 : 30 novembre - 1er décembre, Lille : A simplex-lattice design approach to the fruit juice authentication by ^{1}H NMR spectroscopy: discrimination of grapefruit juice, orange juice and pulpwash juices. Fruit juice authentication by ^{1}H NMR: discrimination of grapefruit juice, orange juice and blends.

v. Liste des abréviations

1D	Une dimension
ACI	Analyse en composantes indépendantes
ACP	Analyse en composantes principales
AIJN	Association of the Industry of Juices et Nectars (EU)
ANOVA	Analyse de la variance
Brix	Mesure des sucres solubles
CLV	Clustering of variable
COW	Correlation optimised warping
D_2O	Eau lourde : oxyde de deutérium
dB	Décibel
DE	Temps avant enregistrement de la FID
DMP	Diméthylproline
DR	Résolution digitale (Digital resolution)
DW	Temps de mélange (Dwell time)
EQCS	European Quality Control System
EWZS	Evolving window zone selection
FID	Interférogramme (Free induction decay)
GABA	Acide gamma-amino butyrique
GJ	Jus de pamplemousse (Grapefruit Juice)
IFU	International Federation of Fruit Juice Producers
iPLS	Interval PLS
ISO	International Standards Organisation
LIR	Lointain infrarouge
MIR	Moyen infrarouge
NS	Nombre d'acquisition (Number of scans)
OG	Mélange de jus d'orange et jus de pamplemousse (Orange Grapefruit)
OJ	Jus d'orange (Orange Juice)
PIR	Proche infrarouge
PL	Puissance (Power level)
PLS	Partial Least Squares
R^2	Coefficient de corrélation au carré

RMN	Résonance magnétique nucléaire
RMSECV	Root Mean Square Error of Cross-Validation
S/N	Rapport signal sur bruit
SW	Largeur spectrale (Spectral width)
SWH	Largeur spectrale en Hertz (Spectral width in Hertz)
T_1	Relaxation longitudinale
T_2	Relaxation transversale
TD	Taille des données en domaine temps (Time domain data size)
TSP	Triméthylsilyl [2,2,3,3-$^2H_4$] propionate

vi. Liste des tableaux

vii. Liste des Figures

Sommaire

1. Introduction générale

1.1. Authenticité des produits et contrôle de la filière agroalimentaire

Les problèmes d'authenticité dans la filière agroalimentaire ne sont pas récents et remontent même aux civilisations grecque et romaine [1]. L'authenticité d'un produit peut être définie par sa conformité à sa définition. Encore faut-il qu'il y ait une définition.

Lorsque l'on prend le cas des produits agroalimentaires, plusieurs problématiques d'authenticité se posent : conformité à l'espèce ou à la variété, au contenu présumé, à son origine naturelle ou artificielle, à son mode de culture biologique ou conventionnelle, à son mode de production « sauvage » ou élevé, la non-adultération, la conformité avec le millésime de production, ou encore l'origine géographique. La raison majeure de la fraude alimentaire est économique. Il s'agit de générer des profits supérieurs avec des produits de moins bonne qualité.

Pour des questions de sûreté et d'équité commerciale, la réglementation de l'industrie agroalimentaire évolue en permanence. En juillet 2007, ce sont les allégations nutritionnelles et de santé qui ont été cadrées de façon législative [2]. Les mesures législatives ont pour but d'assurer au consommateur la loyauté de l'information : dénomination, caractéristiques, provenance des produits. La détection et la prévention des fraudes, par les autorités compétentes, utilisent des contrôles adaptés au produit et au paramètre suspect.

Par ailleurs, l'industrie, toujours pressée par les délais, souhaite s'affranchir des contrôles de la qualité de sa filière : de la réception de la matière première au produit fini en sortie d'usine, en passant par la maîtrise de la qualité du procédé, et ce pour se conformer aux standards de qualité tel HACCP ou ISO [3,4]. Pour des questions économiques de rentabilité ces contrôles doivent être menés de manière rapide, simple, sûre et peu coûteuse.

Ainsi, les méthodes d'analyse rapides pour contrôler un nombre important de paramètres sur des produits à différents stades de leur fabrication sont une aide précieuse, génératrice de profits [5].

Les méthodes de criblage ou "profiling", qui permettent de collecter et d'exploiter des données individuelles multiples (un profil), utilisent l'avantage d'une caractérisation plus performante grâce à une description plus complète. Ces techniques de profiling sont particulièrement adaptées à la détection de fraude dans les produits alimentaires car elles prennent en compte différents types d'information rendant compte de la composition de l'échantillon.

Dans ce chapitre, nous allons introduire de façon très concise, les méthodes d'analyse multivariée qui peuvent prendre en compte l'information contenue dans ces profils. Ensuite, nous dresserons un état des lieux des différentes techniques d'analyse qui permettent d'obtenir ce type de données sur les produits agroalimentaires. Enfin, nous étudierons les différents types d'analyses de l'authenticité qui ont été étudiés et développés en spectroscopie RMN du proton.

1.2. Chimiométrie

La chimiométrie est l'application des outils mathématiques, en particulier statistiques, pour optimiser les procédures d'obtention et de traitement de données de la chimie analytique, afin d'en extraire le maximum d'informations pertinentes. Cette science fait l'objet de nombreuses revues générales [6-10]. Elle a trois objectifs principaux [5]. Le premier est la description des données sous forme synthétique, c'est le cas des statistiques élémentaires ou des méthodes descriptives comme l'analyse en composantes principales (ACP) [11]. Le second objectif est la prédiction soit de valeurs continues comme avec la régression PLS [12], soit de classe comme avec une analyse discriminante [13]. Enfin, son dernier objectif est la planification des expériences [14] pour en tirer le meilleur parti.

La chimiométrie se base sur des modèles qui sont soit linéaires, soit non linéaires. Dans un modèle linéaire, l'hypothèse de départ est que l'information cherchée peut être extraite à partir de combinaisons linéaires des variables de départ, le cas le plus simple étant la régression linéaire [15]. Dans les modèles non-linéaires, le modèle prédictif est sous la forme d'un réseau d'unités connectées entre-elles [5]. L'exemple type de ce modèle est le réseau de neurone [16].

La classification des méthodes chimiométriques peut aussi se faire sur le mode de construction du modèle. On parle de modèle supervisé ou non, selon qu'il utilise des connaissances autres que les données acquises de façon expérimentale. L'ACP par exemple est une méthode non-supervisée. Les méthodes supervisées utilisent, en général, des connaissances sur une collection d'échantillons étalons qui permettent de créer le modèle avant de prédire des données sans information annexe. Un exemple classique est la méthode PLS.

1.3. Méthodes multi-analytiques de certification de l'authenticité des produits

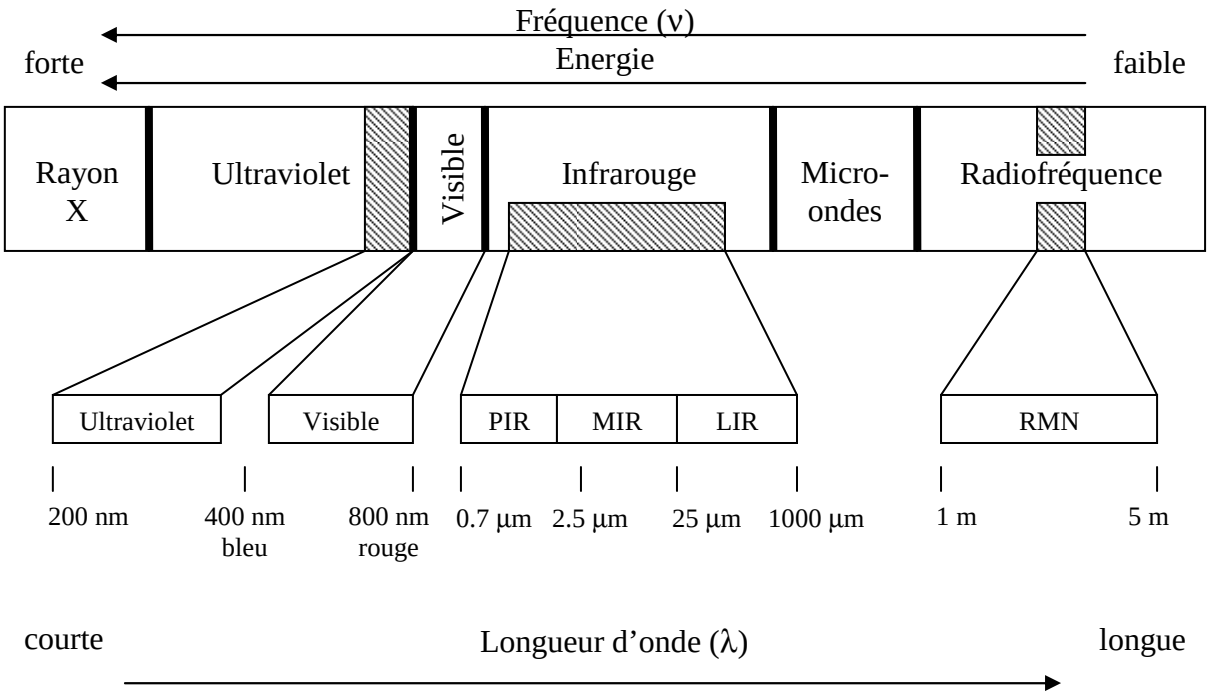
Contrairement à des techniques uni-analytiques comme par exemple la pH-métrie où à chaque mesure on obtient un seul point ou résultat, une méthode multi-analytique peut acquérir de nombreuses données simultanément.

1.3.1. La spectroscopie

Les méthodes spectroscopiques sont largement utilisées pour l'analyse qualitative et quantitative de composés de produits agroalimentaires [5,17]. En général, quand les signaux sont bien définis et non superposés, on peut à partir de l'intensité du signal calculer une concentration. Cependant, lors de l'analyse de matrices complexes, les signaux ne sont pas toujours bien séparés. Il n'est donc pas évident d'identifier et de quantifier les différents composants. Afin de rester dans l'objectif premier de l'analyse spectrale, qui est d'obtenir un résultat d'analyse rapide sur différents paramètres, il est alors nécessaire d'utiliser des méthodes d'extraction d'information rapide telles les méthodes chimiométriques présentées dans la partie précédente.

Les techniques spectroscopiques peuvent être considérées comme des réponses d'absorption d'énergie en fonction de la fréquence du rayonnement concerné [17,18] (cf. Figure 1).

Nous allons approcher les spectroscopies MIR, PIR et RMN et mettre en évidence leurs caractéristiques. Ces trois techniques permettent une acquisition rapide du signal. En effet, l'échantillon nécessite très peu ou pas de préparation, et peut généralement être placé sur un porte-échantillon ou être mesuré en ligne. De plus, l'instrumentation est multi-analytique. Ainsi, plusieurs constituants peuvent être déterminés à partir d'un seul spectre. Ensuite, ces techniques n'utilisent pas ou peu de solvant par rapport aux techniques d'analyse classique, qui peuvent poser des problèmes pour l'environnement. Un échantillon va absorber partiellement et sélectivement le rayonnement auquel il est soumis. Cette absorption est caractéristique des propriétés chimiques et physico-chimiques du produit considéré.



Légende :

PIR : Proche infrarouge

MIR : Moyen infrarouge

LIR : Lointain infrarouge

RMN : Résonance magnétique nucléaire

Figure 1 : Une partie du spectre électromagnétique montrant la relation entre les longueurs d'onde, l'énergie, les fréquences et les techniques spectroscopiques

1.3.1.1.Moyen infrarouge (MIR)

En spectroscopie infrarouge, on s'intéresse aux différentes vibrations des liaisons. La spectroscopie MIR est caractérisée par une résolution fine des pics qui lui confère plusieurs avantages. Tout d'abord, la finesse des raies permet d'identifier facilement le type de liaison rencontré et même de quantifier un composé en fonction de l'intensité des signaux. Un autre avantage important des spectres MIR est qu'ils apportent des informations sur l'état physico-chimique de différents composants comme la cristallinité, l'oxydation, ou encore la phase, qui peuvent être utiles pour étudier les changements causés par le traitement ou le stockage [17].

Le développement des applications MIR en agroalimentaire est récent, suite à l'apparition d'appareils basés sur la transformée de Fourier et de nouveaux dispositifs de présentation des échantillons plus pratiques [19]. Plusieurs problèmes d'authentification ont été étudiés par ce

type de méthode ; par exemple la quantification des adultérants de l'huile d'olive [20], ou encore la discrimination des différents types de fruits dans les purées [21] ou les confitures [22], et enfin la discrimination des différentes variétés de café [23,24]. Au niveau du contrôle qualité de la production, une étude polonaise [25] s'est intéressée à l 'évolution de la teneur en phospholipides lors de la production d'huile de colza.

1.3.1.2.Proche infrarouge (PIR)

En spectroscopie PIR, les raies sont plus larges et donc la résolution des pics est moins bonne. Il n'est donc plus possible de quantifier un composé en fonction de l'intensité d'un pic et l'on fait souvent appel à la chimiométrie pour développer des modèles plus complexes [5,26]. Cette technique étant moins sensible, elle est plus robuste et moins affectée par les problèmes d'humidité que la spectroscopie MIR.

De nombreuses applications ont été développées pour l'analyse des produits agroalimentaires et agricoles. Par exemple, les différentes variétés de blé peuvent être discriminées [27,28] et la détermination de plusieurs paramètres permet même d'en estimer la valeur. De même, les huiles de différentes espèces végétales ont pu être séparées par analyse en composantes principales [29]. Enfin, les techniques PIR permettent d'étudier l'authenticité des produits comme par exemple l'authenticité du jus d'orange [30], ou du ginseng [31], ou encore la pureté de la matière grasse dans le chocolat [32].

L'avantage de la spectroscopie infrarouge par rapport à la RMN, outre l'investissement financier, est que l'analyse est non-destructive. Le même échantillon peut être retenu pour d'autres analyses ou, éventuellement, remis en circulation. Cependant, comme nous allons le voir la RMN est très informative. Alors qu'en IR, l'information se limite souvent à des groupements fonctionnels, la RMN donne des informations sur la structure de la molécule : chaîne carbonée et isomérie.

1.3.1.3.Résonance magnétique nucléaire (RMN)

En acquérant une information quantitative et compositionnelle sur de nombreux types de constituants d'un produit alimentaire [33-35], la spectroscopie RMN se distingue comme un outil performant tout à fait adapté au profiling, sujet de notre étude.

On distingue deux techniques de RMN de haut champ, celle basée sur la détection d'isotopes majeurs comme l'hydrogène-1 (le proton) et celle basée sur des isotopes présents en faible quantité comme l'hydrogène-2 (le deutérium), le carbone 13, etc.

1.4. Applications de la spectroscopie RMN du proton à l'authenticité des produits

Dû à l'abondance naturelle de l'isotope proton de l'hydrogène (99,9855 %) [36], et à la forte teneur en hydrogène dans la matière vivante, la spectroscopie RMN du proton est beaucoup plus sensible et permet donc des acquisitions beaucoup plus rapides que les techniques vues précédemment. Ainsi, aujourd'hui de nombreux problèmes d'adultération sont étudiés par RMN du Proton (^{1}H), utilisant la composition en molécules protonnées des produits et des méthodes chimiométriques adaptées.

1.4.1. Adultération

L'adultération consiste à faire passer un produit qui a subi l'ajout ou la suppression d'un composé, ce qui le rend de moindre valeur, pour un produit authentique. Cette opération frauduleuse a été étudiée en combinant la spectroscopie RMN ^{1}H à des outils chimiométriques.

L'huile d'olive pure a pu être distinguée d'huile d'olive adultérée par ajout d'autres huiles végétales : tournesol ou noisette [37]. De même, la fraude consistant à presser les oranges en rinçant la pulpe, laissant ainsi passer ce que l'on appelle le jus « pulpwash » a pu être repérée par RMN ^{1}H en trouvant des marqueurs spécifiques [38].

1.4.2. Origine botanique

D'autres possibilités d'authentification pour la spectrométrie RMN ^{1}H ont été mises en évidences, notamment en caractérisation des origines botaniques. Ainsi, une étude sur les composés mineurs du vin et en particulier le profil des acides aminés [39] a permis d'identifier les cépages des vins blancs de Slovénie. Une autre étude utilise la RMN ^{1}H pour déterminer la variété des pommes à partir des composés aromatiques du jus [40].

1.4.3. Caractérisation du procédé de fabrication

Pour ce qui est de l'appellation des produits, qui peut, entre autre, dépendre du type de production utilisé, il convient de pouvoir caractériser le procédé. Il a été démontré que ce type de caractérisation était possible par RMN ^{1}H. Dans le cas du vinaigre de vin balsamique, c'est grâce à la présence de composés caractéristiques du vieillissement [41] que le procédé est vérifié comme étant « traditionnel » ou non. De même, à partir des différences dans les signaux obtenus par RMN ^{1}H, les procédés de fermentation des bières [42,43] ou de torréfaction du café [44] ont pu être caractérisés.

1.4.4. Caractérisation de la fraîcheur

L'authenticité des produits est aussi un problème de conformité aux règles sanitaires des produits. La RMN ^{1}H a aussi apporté sa contribution pour le contrôle de la qualité des produits. En effet, l'étude par *Sitter et al.* [45] permet de qualifier l'état de fraîcheur du flétan. Celle par *Belton et al.* [40] caractérise l'état d'oxydation du jus de pomme et apporte donc une information sur la qualité des produits testés. L'article de *Duarte, et al*. [46] aborde la question de la détérioration microbiologique des produits et en particulier du jus de mangue contaminé par *Penicillium expansum*. Le suivi des modifications de compositions du jus a permis de mettre en évidence quelques acides organiques et aminés qui peuvent être des indicateurs précoces de la détérioration.

Toutes ces études, qui exploitent les informations spectrales de RMN ^{1}H, utilisent un ensemble de méthodes statistiques, graphiques ou symboliques permettant d'améliorer leur compréhension. Autrement dit, elles se servent d'outils puissants qui sont des méthodes chimiométriques.

1.5. Positionnement du sujet

Cette thèse a deux objectifs. Le premier est de développer de nouvelles applications de la RMN ^{1}H pour la qualification et quantification simples et rapides de produits agroalimentaires. Ceci répond à la fois aux demandes des industriels (rapidité) et des autorités (fiabilité). Le deuxième objectif est d'étudier les différentes possibilités de traitement des spectres RMN ^{1}H offertes par les outils chimiométriques. Plusieurs études sur différents produits agroalimentaires illustrent ces deux thématiques.

La méthodologie employée dans les différentes études est présentée dans le chapitre 2. Ensuite, une synthèse des résultats obtenus s'appuyant sur les résultats de différentes applications publiées, concernant l'authenticité des produits : les jus de fruits, les yaourts et le vinaigre, est présentée dans le troisième chapitre. Les publications sont insérées à la suite de cette synthèse.

1.6. Références

1. Lees, M. *Food Authenticity : Issues and methodologies*, pp. 311 (Eurofins Scientific Laboratories, Nantes, 1999).
2. Ubifrance. Etiquetage alimentaire : les règles changent en Europe au 1er juillet. http://www.ubifrance.fr/medias/infopresse/document/CP_etiquetage_alim.pdf, 12/01/2008, (2007).
3. ISO 22005:2007 : Traçabilité de la chaîne alimentaire -- Principes généraux et exigences fondamentales s'appliquant à la conception du système et à sa mise en oeuvre. (2007).
4. ISO 22000:2005 : Systèmes de management de la sécurité des denrées alimentaires -- Exigences pour tout organisme appartenant à la chaîne alimentaire. (2005).
5. Bertrand, D. & Dufour, E. *La spectroscopie infrarouge et ses applications analytiques*, pp. 566 (TEC&DOC, Paris, 2006).
6. Brereton, R. G. *Chemometrics applications of mathematics and statistics to laboratory systems*, pp. 307 (Ellis Horwood, Chichester, 1990).
7. Beebe, K., Pell, R. & Seasholtz, M. *Chemometrics: A practical Guide*, pp. 360 (Wiley, New York, 1998).
8. Eriksson, L., Johansson, E., Kettaneh-Wold, N. & Wold, S. *Multi- and Megavariate Data Analysis - Principles and Applications*, pp. 533 (UMETRICS, 2001).
9. Martens, H. & Naes, T. *Multivariate calibration*, pp. 438 (Wiley, New York, 1989).
10. Sharaf, M. A., Illman, D. L. & Kowalski, B. R. *Chemometrics*, pp. 352 (Wiley, New York, 1986).
11. Joliffe, I. T. *Principal Component Analysis*, pp. 271 (Springer-Verlag, New York, 1986).
12. Tenenhaus, M. L'approche PLS. *Revue de Statistique Appliquée* **47**, 2-55 (1999).
13. Bardos, M. *Analyse Discriminante*, pp. 223 (Dunod, Paris, 2001).
14. Goupy, J. *La Méthode des Plans d'Expériences*, pp. 303 (Dunod, Paris, 1996).
15. Daudin, J.-J., Robin, S. & Vuillet, C. *Statistique inférentielle*, pp. 185 (Presses Universitaires de Rennes, Rennes, 1999).
16. Haykin, S. *Neural Networks: A Comprehensive Foundation*, pp. 842 (Prentice Hall, New Jersey, 1998).

17. Wilson, R. H. *Spectroscopic techniques for food analysis*, pp. 246 (VCH, New York, 1994).
18. Pavia, D. L., Lampman, G. M. & Kriz, G. S. *Introduction to spectroscopy*, pp. 416 (Saunders College Publishing, 1996).
19. Downey, G. in *La spectroscopie infrarouge et ses applications analytiques* (ed. Bertrand, D.), 479-504 (TEC & DOC, Paris, 2006).
20. Lai, Y. W., Kemsley, E. K. & Wilson, R. H. Quantitative analysis of potential adulterants of extra virgin olive oil using infrared spectroscopy. *Food Chemistry* **53**, 95-98 (1995).
21. Defernez, M., Kemsley, E. K. & Wilson, R. H. Use of infrared spectroscopy and chemometrics for the authentication of fruit purées. *J. Agric. Food Chem.* **43**, 109-113 (1995).
22. Defernez, M. & Wilson, R. H. Mid-infrared spectroscopy and chemometrics for determining the type of fruit used in jam. *J Science of. Food and Agriculture* **67**, 461-467 (1995).
23. Briandet, R., Kemsley, E. K. & Wilson, R. H. Discrimination of arabica and robusta in instant coffee by Fourier transform infrared spectroscopy and chemometrics. *J. Agric. Food Chem.* **44**, 170-174 (1996).
24. Kemsley, E. K., Ruault, S. & Wilson, R. H. Discrimination between coffea arabica and canephora variant robusta beans using infrared spectroscopy. *Food Chem.* **54**, 321-326 (1995).
25. Szydlowska-Czerniak, A. MIR spectroscopy and partial least-squares regression for determination of phospholipids in rapeseed oils at various stages of technological process. *Food Chem.* **105**, 1179-1187 (2007).
26. Sugiyama, J., McClure, W. F., Hana, M., eds., Murray, I. & Cowe, I. A. in *Advances in Near-Infrared Spectroscopy*, 61-66 (VCH Publishers, 1992).
27. Bertrand, D., Robert, P. & Loisel, W. Identification of some wheat varieties by near infrared reflectance spectroscopy. *JSFA* **36**, 1120-1124 (1985).
28. Davies, A. M. C. & McClure, W. F. Near infrared analysis in the Fourier domain with special reference to process control. *Anal. Proc.* **22**, 321-322 (1985).
29. Sato, T. Application of principal component analysis on near-infrared spectroscopic data of vegetable oils for their classification. *JAOCS* **71**, 293-298 (1994).

30. Evans, D. G., Scotter, C. N. G., Day, L. Z. & Hall, M. N. Determination of the authenticity of orange juice by discriminant analysis of near infrared spectra: a study of pre-treatment and transformation of spectral data. *JNIRS* **1**, 33-44 (1993).

31. Yap, K. Y. L., Chan, S. Y. & Lim, C. S. The reliability of traditional authentication – A case of ginseng misfit. *Food Chem.* **107**, 570-575 (2008).

32. Che Man, Y. B., Syahariza, Z. A., Mirghani, M. E. S., Jinap, S. & Bakar, J. Analysis of potential lard adulteration in chocolate and chocolate products using Fourier transform infrared spectroscopy. *Food Chem.* **90**, 815-819 (2005).

33. Le Gall, G. *Food authenticity and quality assessment using ^{1}H NMR spectroscopy and chemometrics*, pp. 309 (University of East Anglia, 2002).

34. Le Gall, G. & Colquhoun, I. J. in *Food authenticity and tractability*, 131-155 (2003).

35. Mannina, L. & Segre, A. High-resolution NMR: from chemical structure to food authenticity. *Grasas y Aceites* **53**, 22-33 (2002).

36. Rossmann, A. Determination of stable isotope ratios in food analysis. *Food Reviews International* **17**, 347-381 (2001).

37. Fauhl, C., Reniero, F. & Guillou, C. ^{1}H NMR as a tool for the analysis of mixtures of virgin olive oil with oils of different botanic origin. *Magnetic Resonance in Chemistry* **38**, 436-443 (2000).

38. Le Gall, G., Puaud, M. & Colquhoun, I. J. Discrimination between orange juice and pulp wash by ^{1}H NMR spectroscopy: Identification of marker compounds. *J. Agric. Food Chem.* **49**, 580-588 (2001).

39. Kozir, I. J. & Kidric, J. Use of modern NMR spectroscopy in wine analysis: determination of minor compounds. *Analytica Chimica Acta* **458**, 77-84 (2002).

40. Belton, P. S., Delgadillo, I., Gil, A. M., Roma, P., Casuscelli, F., Colquhoun, I. J., Dennis, M. J. & Spraul, M. High-field proton RMN studies of apple juices. *Magnetic Resonance in Chemistry* **35**, 52-60 (1997).

41. Consonni, R. & Gatti, A. ^{1}H NMR studies on Italian balsamic and traditional balsamic vinegars. *J. Agric. Food Chem.* **52**, 3446-3450 (2004).

42. Duarte, I. F., Barros, A., Almeida, C., Spraul, M. & Gil, A. M. Multivariate Analysis of NMR and FTIR data as potential tool for quality control of beer. *J. Agric. Food Chem.* **52**, 1031-1038 (2004).

43. Duarte, C. A., Barros, A., Belton, P. S., Righelato, R., Spraul, M., Humpfer, E. & Gil, A. M. High-resolution NMR spectroscopy and multivariate analysis for the characterization of beer. *J. Agric. Food Chem.* **50**, 2475-2481 (2002).

44. Bosco, M., Toffanin, R., De Palo, D., Zatti, L. & Segre, A. High-resolution ^{1}H NMR investigation of coffee. *J. Sci. Food Agri.* **79**, 869-878 (1999).

45. Sitter, B., Krane, J., Gribbestad, I. S., Jorgensen, L. & Aursand, M. in *Advances in Magnetic Resonance in Food Science* (eds. Webb, G. A., Belton, P. S. & Hills, B. P.), 226-237 (Royal Society of Chemistry, London, 1999).

46. Duarte, I. F., Delgadillo, I. & Gil, A. M. Study of natural mango juice spoilage and microbial contamination with *Penicillium expansum* by high-resolution ^{1}H NMR spectroscopy. *Food Chemistry* **96**, 313-324 (2006).

2. Méthodologie

Dans ce chapitre, nous allons décrire et suivre la méthodologie employée lors de différentes études. Dans chaque étude de cas, une fois les échantillons sélectionnés, ceux-ci étaient tout d'abord préparés pour l'analyse RMN, puis mesurés avec des paramètres d'acquisition spécifiques. Ensuite les signaux recueillis étaient pré-traités puis analysés selon différentes méthodes chimiométriques.

2.1. Collecte des échantillons

Une des grandes difficultés des différentes études présentées est le choix des échantillons. En effet pour traiter de l'authenticité des produits il faut pouvoir attester de leur origine et composition sans doute. Dans le cas contraire, il faut toujours garder à l'esprit qu'un échantillon peut être adultéré.

Dans les études présentées dans cette thèse certains échantillons ont été collectés sur le marché, c'est le cas des jus d'orange et pamplemousse et des yaourts à la fraise. D'autres, ont été fabriqués de manière spécifique pour l'étude comme les yaourts avec différentes concentrations de préparation de fruits. Enfin, certains ont été collectés chez des producteurs comme le vinaigre.

2.2. Préparation des échantillons et influence sur la mesure spectrale

L'échantillon est introduit dans un tube en pyrex de 5 mm. Une fois l'échantillon placé dans la sonde du spectromètre, la mesure s'effectue sur une zone active d'environ 2 cm. La mesure optimum est obtenue lorsque le champ magnétique appliqué sur l'échantillon est homogène. De plus, afin d'établir des comparaisons entre les spectres, il faut que d'un échantillon à l'autre les conditions expérimentales, dont le champ magnétique, soient stables.

2.2.1. Influence de la préparation sur l'homogénéité du champ magnétique

Le champ magnétique peut varier pour différentes raisons. Tout d'abord le vieillissement de l'aimant induit une modification non réversible du champ. De plus le champ peut varier temporairement lors de modifications de son environnement comme le déplacement d'objets

métalliques à proximité ou encore des variations de température dans la pièce. Pour pallier ces phénomènes le spectromètre comporte des petites bobines de correction, les « shims », qui permettent d'ajuster spatialement le champ magnétique. L'homogénéité de l'échantillon est primordiale.

2.2.2. Hauteur de l'échantillon dans le tube

Un facteur souvent sous-estimé lors de la préparation de l'échantillon est la quantité, et donc la hauteur, d'échantillon introduit dans le tube de mesure [1]. L'obtention d'une mesure correcte exige que toute la zone active de la bobine soit remplie. De plus, en raison de variations du champ magnétique au niveau de la surface air/solvant un « effet de bord » est susceptible de perturber les mesures, en particulier lorsque le solvant est l'eau. Il faut aussi noter qu'une hauteur trop importante d'échantillon dans le tube ne permet pas de réguler la température de l'échantillon de manière optimum.

Afin de faciliter le réglage des « shims » dans une série d'expériences, une quantité suffisante, et constante d'échantillon est retenue. Dans ces conditions, un simple ajustement des « shims » sur l'axe vertical est souvent suffisant. Pour le réglage des shims nous avons utilisé la résolution du spectre de manière générale, en vérifiant que les deux satellites de la référence interne, le triméthylsilyl [2,2,3,3-$^{2}H_4$] propionate (TSP), soient bien définis. Dans les différentes études présentées, la quantité d'échantillon n'étant pas limitante nous introduisons environ 1.00 mL d'échantillon ce qui représente une hauteur de 7 cm. Dans la sonde 3 cm et 2 cm sont alors situés respectivement au-dessus et en dessous de la zone active de 2 cm. Dans ces conditions la variabilité inter-échantillons est négligeable et un jeu de shims correspondant à chaque solvant, voire à chaque type d'échantillon peut être mémorisé pour réutilisation.

2.2.2.1.Homogénéité de l'échantillon

La présence de particules en mouvement dans l'échantillon est aussi susceptible de créer des perturbations du champ magnétique qui ne peuvent pas être corrigées. Il est donc conseillé, le cas échéant, de filtrer l'échantillon. Par ailleurs, de façon instrumentale, l'homogénéité de l'échantillon est assurée par la rotation du tube à l'intérieur de l'aimant.

2.2.3. Eléments à inclure dans la préparation

2.2.3.1.Lock

Pour s'affranchir des variations de champ magnétique le système est asservi à une résonance de référence : la fréquence du « lock ». En RMN du proton la référence est en général le deutérium. L'utilisation de solvants deutérés permet d'éviter la présence de raies de solvant intenses. De plus, le signal des protons résiduels fournit une référence de déplacement chimique. Dans notre étude, l'eau lourde (D_2O), qui est un très bon solvant, a été choisie car elle n'apporte aucun signal supplémentaire dans le spectre proton. Une quantité variable d'eau lourde, allant de 10 à près de 100 % a été introduite dans le tube, selon le type d'expérience.

2.2.3.2.Référence interne

Afin de pouvoir comparer les spectres, les fréquences de résonance sont repérées par rapport à une référence interne et exprimées en termes de déplacements chimiques (cf. Annexe 1 sur la théorie de la RMN). Dans notre étude, nous avons utilisé le TSP, δ(TSP) = 0,00 ppm.

De plus, lors de l'utilisation de la RMN dans un but de quantification, il faut introduire dans la préparation, en quantité connue, une substance dont les pics sont distincts de ceux de la molécule à quantifier. De plus, dans une expérience non quantitative l'intensité de ces pics permet de valider la mesure.

Nous avons utilisé la même référence interne pour les déplacements chimiques, le TSP, soit sous forme de cristaux soit sous forme dissoute à concentration connue.

2.2.4. Influence de la préparation de l'échantillon pour l'homogénéité des spectres

Certains facteurs tels que la viscosité, le pH et la température de l'échantillon peuvent influencer le déplacement chimique et la forme des signaux d'une même molécule.

2.2.4.1.Viscosité

La RMN étant une méthode réputée peu sensible, il faut que les composés à quantifier soient présents en quantité suffisante pour obtenir un bon rapport signal sur bruit (S/N). Cependant, afin de limiter la viscosité, mieux vaut qu'il ne soit pas trop concentré.

2.2.4.2.Concentration

La concentration du produit doit être adaptée à l'obtention rapide d'un bon rapport signal sur bruit, tout en tenant compte du risque de dégradation de résolution associé à une trop grande viscosité de l'échantillon.

En effet, la diminution de mobilité des molécules dans l'échantillon entraîne une variation des temps de relaxation [2]. Rappelons que, dans le cas du mécanisme de relaxation dipolaire, souvent dominant pour les protons, les temps de relaxation longitudinale, T_1, et transversale, T_2, sont égaux dans les milieux relativement mobiles. Ils diminuent rapidement lorsque la viscosité augmente. Aux faibles mobilités, T_1 devient différent de T_2 mais celui-ci continue à diminuer en raison des interactions spin-spin. Or la largeur des signaux est gouvernée par la relaxation transversale, T_2, (largeur à mi-hauteur = $1 / \pi T_2$). Lorsque la mobilité décroît les spectres sont donc de moins en moins bien résolus. Dans notre étude sur les jus de fruits par exemple, un des facteurs influençant la viscosité de l'échantillon est son degré Brix (fraction de sucres solubles dans les liquides). Dans un souci d'homogénéité il faut donc éviter les trop grandes différences de concentration en sucre soluble (de l'ordre d'un ou deux degrés). Dans les études présentées dans ce manuscrit nous n'avons pas eu besoin d'homogénéiser les degrés Brix des échantillons car les variations étaient très faibles. Par exemple pour le jus d'orange le degré Brix variait entre 11.4 et 13.2 ce qui était acceptable.

2.2.4.3.pH

Le pH de l'échantillon a une grande importance pour tous les acides. Selon leur pKa, ils sont sous formes plus ou moins protonnées, et le nombre de pics observé dans le spectre peut donc être différent. Une étude [2] portant sur les acides citrique, malique et aspartique à différents pHs a montré que les variations de déplacement chimique et de forme de signaux peuvent être très importantes sur une échelle de pH allant de 1.0 à 4.0.

En conséquence, le pH pouvant être très variable, comme dans le cas des jus de fruits, nous l'avons ajusté par ajout d'acide ou de base forte, très concentrés (soude (NaOH) à 5 et 1 Mol et acide chlorhydrique (HCl) à 2.5 Mol).

2.2.4.4.Température

Tout comme la viscosité, la température a un rôle sur l'agitation moléculaire et donc sur la vitesse de relaxation. Lorsque la température diminue, la largeur de raies augmente [1]. Par ailleurs, l'évaporation associée à une température trop élevée peut perturber la mesure. Le paramètre température est donc ajusté au niveau de l'instrument.

2.3. Acquisition des spectres ^{1}H

Pour pouvoir comparer les signaux de différents spectres il faut que les conditions expérimentales soient comparables; nous verrons donc dans la section suivante les conditions d'acquisition à maîtriser.

2.3.1. Spectromètre

Un spectromètre est en général caractérisé par l'intensité de son champ magnétique B_0. Avec le développement des aimants à supraconductivité, l'intensité des champs magnétiques disponibles a beaucoup augmenté. Le spectromètre utilisé dans ce travail délivre un champ magnétique de 9,4 Teslas environ ce qui correspond à une fréquence de 400 MHz environ pour le proton.

2.3.2. Paramètres d'acquisition

Le réglage des paramètres d'acquisition est très important car ils déterminent la nature et la qualité du spectre. Les différents paramètres à prendre en compte lors de l'acquisition d'un spectre sont les suivants :

2.3.2.1.Temps d'acquisition

Le temps d'acquisition du FID, AQ, doit être assez long pour recueillir le maximum de l'information contenue dans le FID et assurer une bonne résolution digitale, DR. Celle-ci est définie par :

$$DR = \frac{1}{AQ} \qquad \text{Equation 1.}$$

Les composantes sinusoïdales du FID décroissent exponentiellement avec les temps de relaxation T_2. Au bout d'un temps d'environ 5 fois la valeur de T_2 la plus élevée, le FID ne contient pratiquement plus que du bruit.

En résonance protonique, un temps d'acquisition de l'ordre de 3 secondes, qui correspond à une résolution digitale de 0,33 Hz est généralement suffisant dans le cas de produits complexes caractérisés par des T_2 relativement courts.

2.3.2.2.Délai de relaxation

En RMN quantitative, le vecteur aimantation doit être revenu à l'équilibre avant chaque nouvelle séquence d'impulsions. Le temps séparant deux séquences doit donc être suffisamment long pour que tous les noyaux de l'échantillon soient relaxés selon l'axe z. En conséquence, le temps de relaxation longitudinale le plus long, $T_{1(limitant)}$, est le facteur limitant la vitesse d'accumulation des signaux. Lorsque le temps d'acquisition AQ est insuffisant pour assurer un retour quasiment complet à l'équilibre, un délai D_1 est introduit après l'acquisition. L'intervalle de répétition des impulsions doit être tel que :

$$AQ+D_1 = 5 \times T_{1\ limitant} \qquad \text{Equation 2.}$$

Dans notre étude, lors de mesures qui visaient à l'étude de molécules particulières, nous avons utilisé des standards de ces molécules et nous les avons validés dans leur matrice afin de déterminer les D_1 et le temps de répétition minimal nécessaire à la justesse de la mesure. Par exemple nous avons mesuré le T_1 de l'acide lactique dans du D_2O puis vérifié dans la matrice d'extrait de yaourt. Les valeurs sont de 6,48 sec pour le quadruplet à 4,09 ppm et 2,34 sec pour le doublet à 1,33 ppm.

2.3.2.3.Taille de l'interférogramme acquis (TD: Time domain data size)

TD est le nombre de points acquis sur le FID, c'est à dire la taille de mémoire utilisée. En général il est de 16, 32, ou 64 K. Plus le nombre de points est élevé, meilleurs sont les résultats en termes de résolution.

2.3.2.4.Fenêtre spectrale (SW : spectral width, SWH : spectral width en Hertz)

La fenêtre spectrale, SW, fixe la gamme de fréquences observées. Cette fenêtre est conditionnée par la gamme de déplacements chimiques (en Hertz) de l'échantillon. Selon le théorème de Nyquist, pour être correctement identifiée une fréquence doit être repérée en au moins deux points par période. Cette exigence définit le temps théorique entre deux prises de mesure (DW : "dwell time").

$$DW = \frac{1}{2 \times SW} \qquad \text{Equation 3.}$$

En conséquence, la taille de mémoire utilisée, TD, le temps d'acquisition, AQ, et la largeur de la fenêtre spectrale, SW, sont liés par la relation :

$$AQ = \frac{TD}{2 \times SW} = TD \times DW \qquad \text{Equation 4.}$$

La résolution digitale DR (distance en fréquence entre les points recueillis) est donc :

$$DR = 2 \times SW / TD \qquad \text{Equation 5.}$$

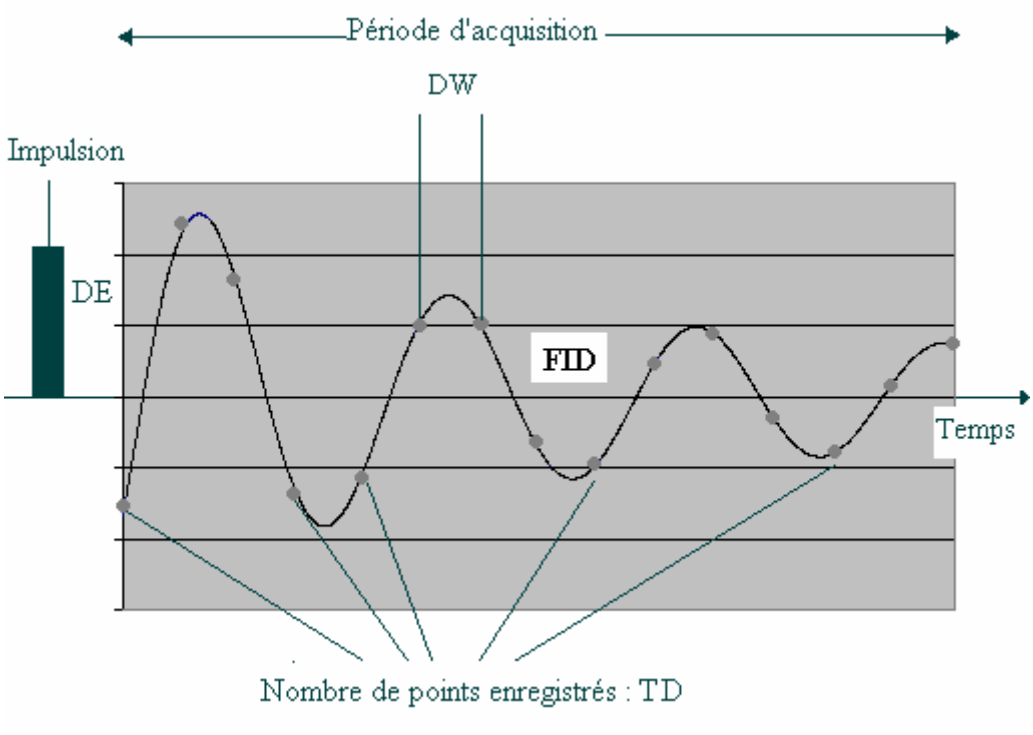
La taille mémoire doit être suffisante pour assurer une bonne résolution digitale.

A priori, la zone d'exploration, SW, peut être diminuée afin d'améliorer la résolution digitale (Equation 5). Cependant, s'il existe des signaux en dehors de la zone observée, ceux-ci sont excités par l'impulsion mais mal reconnus (principe d'échantillonnage). Ils apparaissent alors « réfléchis », et déformés, à l'intérieur de la partie observée. En pratique, la valeur de SW est donc imposée par la taille du spectre. En résonance protonique, la taille maximum est de 20 ppm, soit 10 000 Hz dans un spectromètre à 500 MHz. Avec une taille mémoire de 32 K la résolution digitale obtenue (Equation 4 et 5) ne serait que de 0,6 Hz. Si la gamme de résonances peut être limitée à 10 ppm, la résolution atteint 0,3 Hz. Dans notre étude, la taille de la fenêtre spectrale varie entre 11,8 ppm et 16 ppm et TD varie entre 16, 32 et 64 K. La résolution digitale est donc au minimum de 0,20 Hz soit 0,0005 ppm et au maximum de 0,58 Hz soit 0,0014 ppm.

2.3.2.5.Nombre d'accumulations (NS : number of scans)

NS est le nombre d'accumulations du FID avant de réaliser la Transformée de Fourier. Plus NS est grand, meilleur est le rapport signal sur bruit, S/N. Ce rapport augmente proportionnellement à la racine carrée du nombre d'accumulations. En général, il est recommandé d'utiliser un nombre d'accumulations multiple du nombre de cycles de phase (cf.2.3.2.9) utilisé dans les programmes d'impulsions.

Les différents paramètres que nous venons de décrire joue un rôle dans l'acquisition d'un interférogramme (cf. Figure 2).



DE : temps avant l'enregistrement (*Pre-scan delay*) Ce court délai avant acquisition permet d'éliminer des effets de l'impulsion sur le récepteur. Il permet d'assurer que l'impulsion soit terminée avant le départ de l'acquisition.

Figure 2 : Paramètres d'acquisition d'un FID

2.3.2.6.Cycles de phase

Les cycles de phases permettent de sélectionner les signaux désirés en rejetant ceux qui ne contiennent pas d'information et/ou peuvent masquer de l'information. Ils discriminent le signe des fréquences en dimension f1 et assurent une détection en quadrature optimale dans la dimension f2. En outre, ils compensent, au besoin, les imperfections d'une ou plusieurs impulsions de la séquence. Cette valeur est paramétrée par défaut en fonction de la séquence d'impulsions choisie.

2.3.2.7.Gain

Le gain de réception contrôle l'amplitude de FID. Il doit être adapté à l'intensité du FID. La valeur optimale est la valeur maximale qui permet d'obtenir intégralement l'intensité du signal le plus fort. Pour comparer les intensités de différents spectres cette valeur doit être fixe. En général mieux vaut se placer juste en dessous de cette valeur optimale si les échantillons risquent de ne pas être homogènes.

2.3.2.8.Puissance (PL : Power Level)

Cette variable agit sur l'intensité d'irradiation, B_1. Pour un même temps d'irradiation, plus la puissance est grande, plus l'angle de basculement du vecteur aimantation (« angle d'impulsion ») est grand.

$$\theta = \gamma B_1 PL \qquad \text{Equation 6.}$$

En d'autres termes, plus la puissance augmente plus la durée d'une impulsion 90° est courte. Si la puissance est trop faible, la durée de l'impulsion nécessaire pour obtenir un angle de 90° est longue et la zone d'excitation est réduite (Figure 3). Si PL est trop fort, cela peut endommager la sonde. PL est exprimée en décibel (dB), l'intensité maximale est de –6 dB, et la minimale de 120 dB.

Par ailleurs, quand on utilise une séquence d'impulsions pour faire de la pré-saturation (voir paragraphe suivant), l'intensité de la fréquence appliquée doit être optimisée de manière à ne pas dégrader la qualité des signaux situés de part et d'autre du signal à irradier. Lors de telles mesures nous avons pris soin de choisir une puissance d'irradiation proche de celle capable d'annuler le signal irradié.

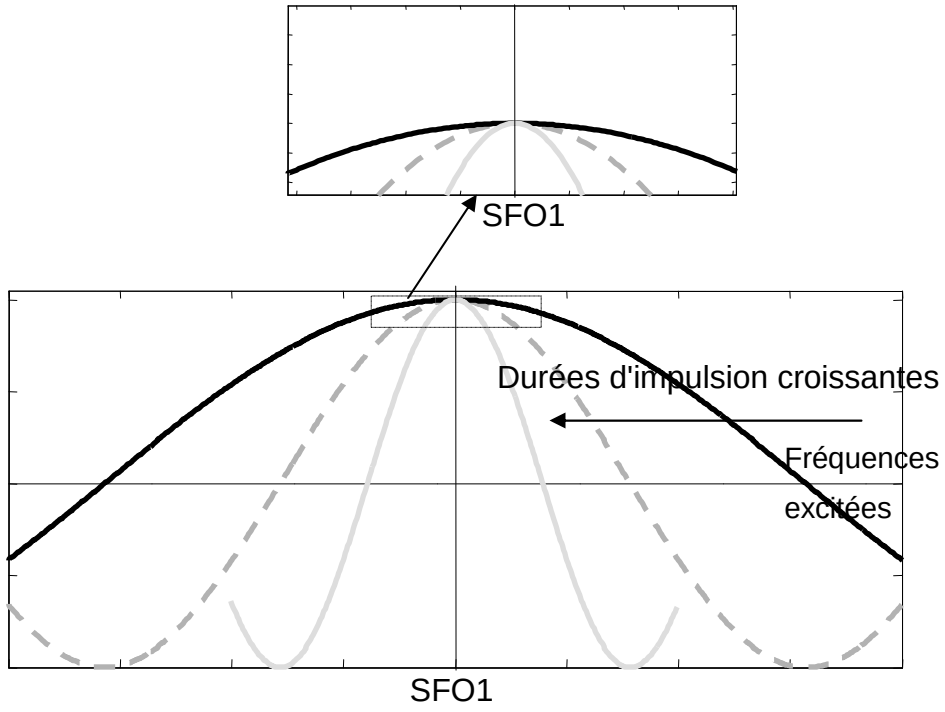
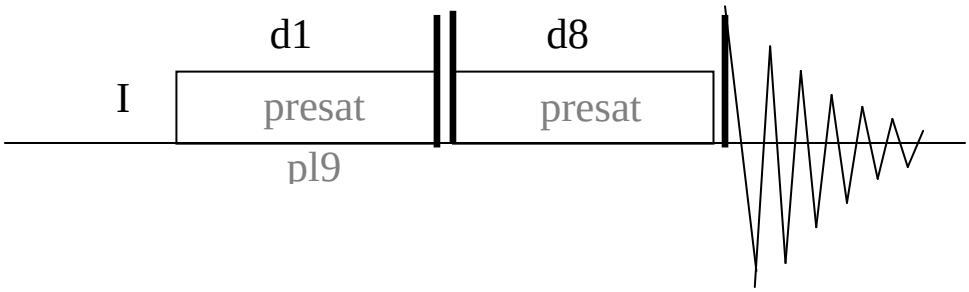


Figure 3 : Stabilité de l'excitation autour de SFO1 en fonction de la durée d'impulsion

2.3.2.9.Séquences d'impulsions

Les séquences d'impulsions ont pour objet de faire subir au vecteur aimantation une série d'orientations et d'évolutions, au cours de périodes de relaxation et de mélange, qui auront des effets particuliers sur les spins et donc sur le FID.

Pour l'acquisition des spectres en une dimension (1D) des échantillons de cette étude, le choix s'est porté sur la séquence « noesypr1d » (Figure 4). En effet, cette séquence permet une pré-saturation de la fréquence de résonance de l'eau, le composé majoritaire des échantillons, et une suppression de l'effet Overhauser nucléaire. Une comparaison (non présentée dans ce rapport) a montré que les résultats étaient meilleurs avec cette séquence qu'avec la séquence plus simple « zgpr » (Annexe 2) qui permet aussi une pré-saturation du signal de l'eau. La séquence choisie comporte une première phase de pré-saturation durant la relaxation, puis elle soumet l'échantillon à deux impulsions-90° successives ce qui permet de renverser le vecteur aimantation selon l'axe-Z. Ensuite, elle comporte une deuxième phase de pré-saturation durant le temps de mélange selon l'axe-Z, puis l'échantillon subit une dernière impulsion -90° pour permettre l'enregistrement de la FID.



pl9 : intensité de la pré-saturation

d1 : temps de relaxation; 1-5 * T1

d8 : temps de mélange

NS : 8 * n, nombre total d'acquisition: NS * TD0

Figure 4 : Séquence "noesypr1d"

2.3.2.10. Température

La température de l'échantillon a, comme nous l'avons vu, une influence sur l'agitation moléculaire et donc sur le phénomène de relaxation. Une sonde de température est placée dans la sonde magnétique. Elle permet de contrôler et d'ajuster en temps réel la température de l'échantillon. En effet, la mesure RMN pouvant produire de la chaleur, il faut maintenir constante la température dans la sonde pour éviter un échauffement de l'échantillon. Un flux d'hélium permet cette régulation. En outre, le spectromètre est placé dans une enceinte dont la

température est contrôlée pour pallier les fortes variations aux cours de la journée et, à plus grande échelle, au cours l'année. Ainsi le flux d'air nécessaire à la régulation reste à peu près constant. Dans nos différentes études de cas, nous avons réglé la température à 29 °C.

2.3.3. Qualité d'un spectre

2.3.3.1. Sensibilité

Tout au long de la chaîne de traitement du signal analogique, un signal indésirable de caractère aléatoire - le bruit gaussien - se superpose au signal intéressant. Son origine se trouve, entre autres, dans le caractère corpusculaire des électrons et dans leur "agitation" thermique (uniquement pour le bruit électronique). Le bruit est donc fonction de la température. Il est considéré comme un signal aléatoire sous-jacent au signal « vrai ». Sa distribution suit une loi normale dont la moyenne est zéro. La tension induite dans la bobine est extrêmement faible mais la perturbation apportée par ce bruit est, dans le cas de noyaux peu sensibles ou peu nombreux, aussi grande voire plus grande que le signal issu de l'échantillon.

La sensibilité est la capacité à distinguer le signal du bruit. Le rapport signal sur bruit, S/N, est défini par la hauteur du pic du signal le plus grand sur la hauteur du bruit. Le niveau de bruit est souvent pris comme l'écart-type du bruit [3]. Le bruit, ε, est défini par une variable aléatoire suivant une loi normale $N(0,\sigma^2)$. Il possède plusieurs propriétés :

$$1/m\sum_{k=0}^{m-1}\varepsilon_k=0 \qquad \text{Equation 7.}$$

$$1/m\sum_{k=0}^{m-1}\varepsilon_k{}^2=\sigma^2 \qquad \text{Equation 8.}$$

$$\sum_{k=0;\, j\neq k}^{m-1}\varepsilon_k\varepsilon_j=0 \qquad \text{Equation 9.}$$

L'écart-type de ce bruit augmente en fonction de la racine carrée du nombre de valeurs cumulées, alors que le signal augmente linéairement. Pour cette raison, le rapport S/N augmente aussi avec la racine carrée du nombre de valeurs.

Les principaux paramètres influant sur le bruit d'un signal sont la température, le bruit dû au pré-amplificateur, les caractéristiques de la bobine réceptrice (volume et géométrie), le remplissage du tube, et le champ magnétique B_0 [1].

Ainsi, lors de l'acquisition des spectres, la température et le remplissage des tubes seront fixés afin de maintenir constante l'influence de ces deux facteurs sur la mesure RMN.

2.3.3.2.Résolution

La capacité à distinguer deux signaux proches en fréquence est un critère important pour évaluer la qualité d'un spectre, c'est la résolution [3]. Elle est fortement influencée par la résolution digitale qui a été définie en 2.2.2.5.

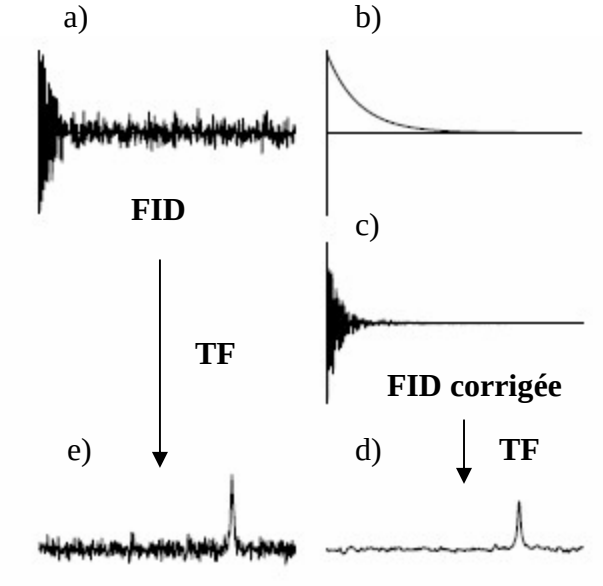
2.4. Méthodes de pré-traitement

2.4.1. Fonction de correction

Pour améliorer le ratio S/N, il est possible après l'acquisition, de traiter la FID. En effet la FID est une somme de sinusoïdes amorties décroissantes. Dans la première partie de l'enregistrement, le signal recueilli comporte beaucoup d'information. En revanche, en fin d'enregistrement, il ne reste quasiment que du bruit qui a été enregistré de manière constante.
On pourrait penser que supprimer cette partie de signal améliorerait la qualité du spectre, mais en fait la troncature du signal introduit des artéfacts dans les signaux après Transformée de Fourier.
La multiplication de la FID par une exponentielle décroissante permet de lisser son profil et de diminuer progressivement l'importance du bruit sans introduire d'artéfacts (cf. Figure 5 a, b et c). En pratique, chaque point i est multiplié par le facteur :

$$\exp\left(\frac{(i-1)\times \mathrm{LB}\times \pi}{2\ \mathrm{SWH}}\right) \qquad \text{Equation 10.}$$

où LB est le facteur d'élargissement Lorentzien (Line Broadening ou Lorentzian broadening factor, en anglais).



a) FID
b) Fonction exponentielle
c) FID corrigée par la fonction exponentielle
d) Signal obtenu après Transformée de Fourier sur la FID corrigée
e) Signal obtenu après Transformée de Fourier sur la FID

Figure 5 : Effet d'une fonction de correction sur la FID et le signal obtenu après transformée de Fourrier

Le signal obtenu après multiplication par une fonction exponentielle est plus clair (cf. Figure 5 d et e) et le ratio S/N est amélioré.

2.4.2. Correction de la phase

L'allure d'un pic doit être une Lorentzienne en absorption de phase. Après la Transformée de Fourier le spectre peut présenter des pics dont l'allure n'est pas celle attendue. Ce défaut peut être éliminé grâce à des corrections automatiques d'ordre zéro et d'ordre 1 (fonctions PHC0 et PHC1 dans le logiciel TOPSPIN de Bruker).

Les données sont composées de valeurs réelles R(i) et imaginaires I(i). La correction de phase se fait selon la formule :

$$R0(i) = R(i)\cos[PHC0 + (i-1)\,PHC1] - I(i)\sin[PHC0 + (i-1)\,PHC1] \quad \text{Equation 11.}$$

$$I0(i) = I(i)\cos[PHC0 + (i-1)\,PHC1] + R(i)\sin[PHC0 + (i-1)\,PHC1] \quad \text{Equation 12.}$$

où R0 et I0 sont les valeurs réelles et imaginaires corrigées et $i > 0$.

Une fois l'option automatique utilisée, une correction manuelle peut être utilisée afin d'améliorer et d'homogénéiser les corrections. En effet, il est important que tous les spectres d'un jeu de données soient "phasés" de la même façon.

2.4.3. Correction de la ligne de base

Le logiciel TOPSIN permet de corriger la ligne de base par un polynôme de degré d, à définir. En général, l'utilisation d'une correction par une constante ou un polynôme de degré 1 est satisfaisante. Par la suite, la ligne de base sera corrigée de la même façon, sous Matlab®.

2.4.4. Alignement des pics

Pour comparer les spectres entre eux, il est important que les déplacements chimiques des pics soient homogènes. Ceux-ci peuvent être décalés en raison, entre autres, de modifications légères du champ magnétique ou du pH. Une première correction sous TOPSPIN permet de positionner le pic de référence. En général, le triméthylsilyl [2,2,3,3-$^{2}H_4$]propionate, noté TSP, est utilisé comme référence interne, son déplacement chimique étant fixé à 0,00 ppm.
Sous Matlab, les autres pics peuvent être réalignés par deux techniques différentes. Tout d'abord, il est possible de découper le spectre en zones dont on peut recaler manuellement les pics un par un. Une alternative consiste à utiliser un algorithme de recalage, Correlation Optimized Warping (COW) [4], utilisé en général pour les chromatogrammes. La Figure 6 illustre l'action de la méthode COW sur une zone spectrale contenant les signaux d'acide citrique et malique de jus de fruits. Le recalage des pics permet d'améliorer la cohésion des données.

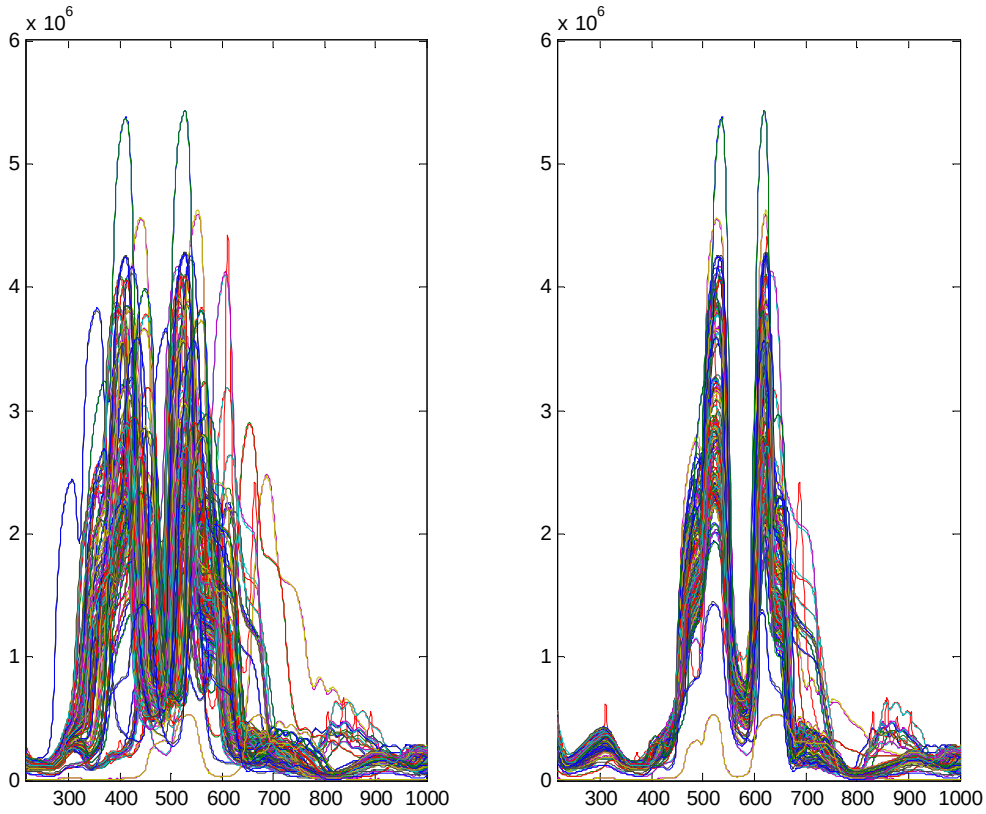


Figure 6 : Effet de la fonction COW sur les signaux de l'acide citrique et malique : a) Données avant le traitement ; b) Données après le traitement.

2.5. La transformation logarithmique

Une des premières observations que l'on peut faire devant un spectre RMN proton (Figure 7 a) est que l'intensité des zones spectrales varie dans les différentes parties du spectre. En général, dans les jeux de données que nous avons étudiés, les intensités les plus importantes se retrouvaient dans la zone « des sucres » entre 3,5 et 5,5 ppm. La différence d'intensité était de l'ordre de 10^3, ce qui pourrait introduire des difficultés d'interprétation dans certaines méthodes chimiométriques comme l'analyse en composantes principales (ACP). En effet, les résultats de l'ACP non normée privilégient la plus grande variabilité et les grandes variations associées aux pics de fortes intensités pourraient cacher des variations moins grandes mais néanmoins intéressantes.

Pour réduire cet effet, nous avons employé une transformation logarithmique des intensités ce qui permet d'obtenir des spectres avec des gammes intensités plus réduites (Figure 7 b).

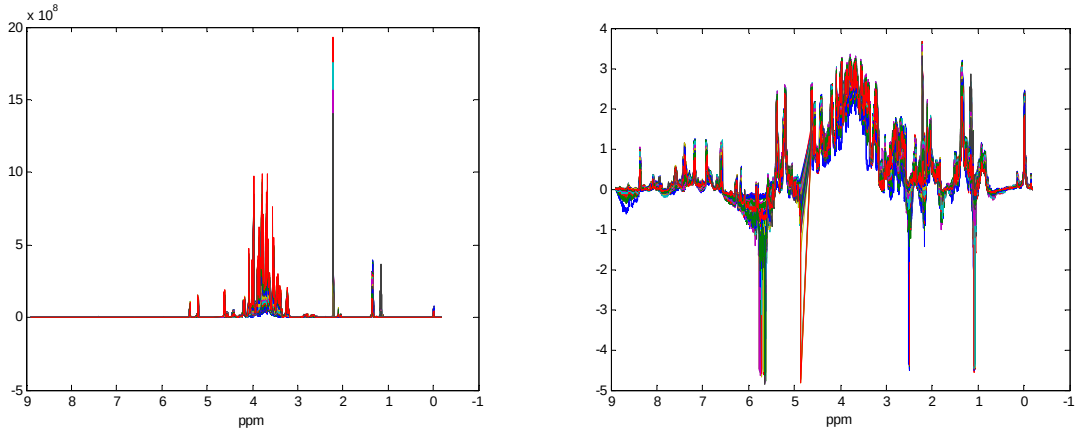


Figure 7 : a) Spectres RMN ^{1}H d'extrait de yaourts aux fruits avant transformation logarithmique ; b) Spectres RMN ^{1}H d'extrait de yaourts aux fruits après transformation logarithmique

2.6. Les méthodes de sélection de variables

Le spectre proton obtenu par RMN est un ensemble de points dont le nombre a été défini par l'utilisateur (2.2.2.5), allant de 16 K à 64 K dans notre étude. Chaque point correspond à un déplacement chimique donné qui est défini comme une variable, *X*. Le nombre, *m*, de variables est donc très important et bien supérieur au nombre d'échantillons analysés, *n*.

Afin de réduire la redondance dans les variables et pour des raisons pratiques de manipulation des jeux de données, nous avons décidé de diminuer la taille des spectres avant de pouvoir appliquer des traitements statistiques. Il s'agit de ne pas éliminer les variables contenant de l'information tout en supprimant le bruit en limitant les redondances. Pour ce faire, plusieurs méthodes de sélection de variables ont été employées.

2.6.1. Variance

La méthode la plus simple pour sélectionner de l'information est de supposer que les variables informatives vont avoir une variation plus importante que les variables non informatives. L'amplitude de la variation d'une variable peut être mesurée par sa variance totale,, donnée par :

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{X}_i - \overline{\mathbf{X}})^2 \qquad \text{Equation 13.}$$

La technique utilisée vise à définir un seuil optimum pour la suppression de variables en vu du résultat escompté. Ainsi, un algorithme de calcul permettait de générer tous les groupes de variables possibles et de retourner les résultats de la méthode testée par la suite, tel qu'un taux de classement correct, en fonction du seuil de sélection. Les meilleurs résultats obtenus pour le seuil de sélection le plus élevé permettent de choisir le jeu à conserver.

2.6.2. iPLS

iPLS [5,6] est une méthode de sélection de variables qui applique l'algorithme PLS sur de petites zones de spectre. La méthode PLS est détaillée en Annexe 3. Il s'agit soit de définir le nombre d'intervalles souhaités, ce qui déterminera la taille des zones spectrales et donc le nombre de variables à prendre en compte, soit de définir manuellement les intervalles à étudier. Sur chacun de ces intervalles, un modèle PLS est calculé avec un nombre croissant de variables latentes jusqu'à un nombre maximal, *f*, défini par l'utilisateur. Il faut aussi introduire un vecteur **y** de valeurs à prédire par régression PLS. Le modèle de référence utilise la totalité du spectre pour la prédiction avec une dimensionnalité déterminée par validation-croisée.
Pour chaque zone spectrale on obtient *f* valeurs de RMSECV qu'il faut alors comparer à la valeur obtenue sur les spectres entiers, $RMSECV_{réf}$ avec un nombre de variables latentes, déterminé par cross-validation (Figure 8).

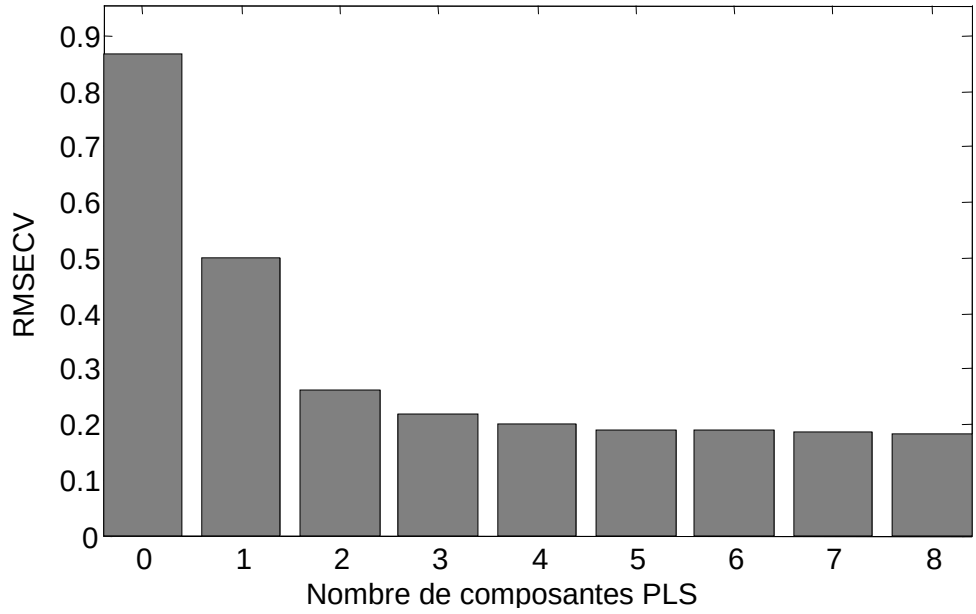


Figure 8 : REMSECV en fonction du nombre de composantes PLS prises en compte pour le modèle global.

Dans cet exemple un modèle à 4 variables latentes permet d'avoir un RMSECV proche du minimum, sans risquer un sur-ajustement ("over-fitting").

Si dans la zone *i* le modèle avec *j* variables latentes donne un RMSECV inférieur à $RMSECV_{réf}$, cette zone sera considérée comme plus prédictive que le spectre complet. Dans la Figure suivante (Figure 9), le jeu de données a été divisé en 31 intervalles. Pour chaque intervalle, les valeurs de RMSECV sont calculées en prenant en compte de 1 à 8 variables latentes. Pour tout le spectre, noté REF sur la Figure 9, la valeur de RMSECV retenue est celle obtenue avec le modèle utilisant 4 variables latentes. Trois intervalles sont en- dessous de ce seuil, les intervalles 9, 12 et 15, tandis que l'intervalle 26 est proche de ce niveau, indiquant que cet intervalle contient de l'information aussi car son pouvoir prédictif est presque aussi grand que celui du spectre entier.

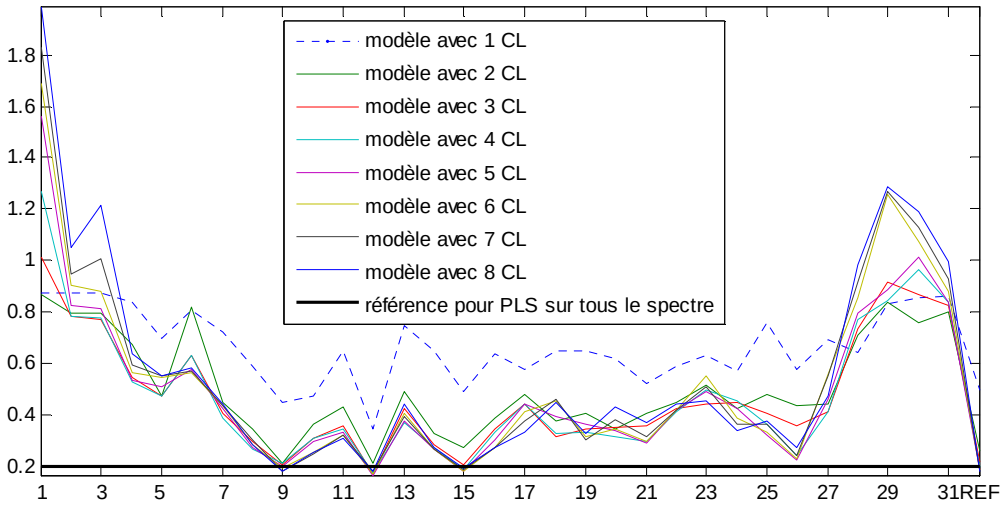
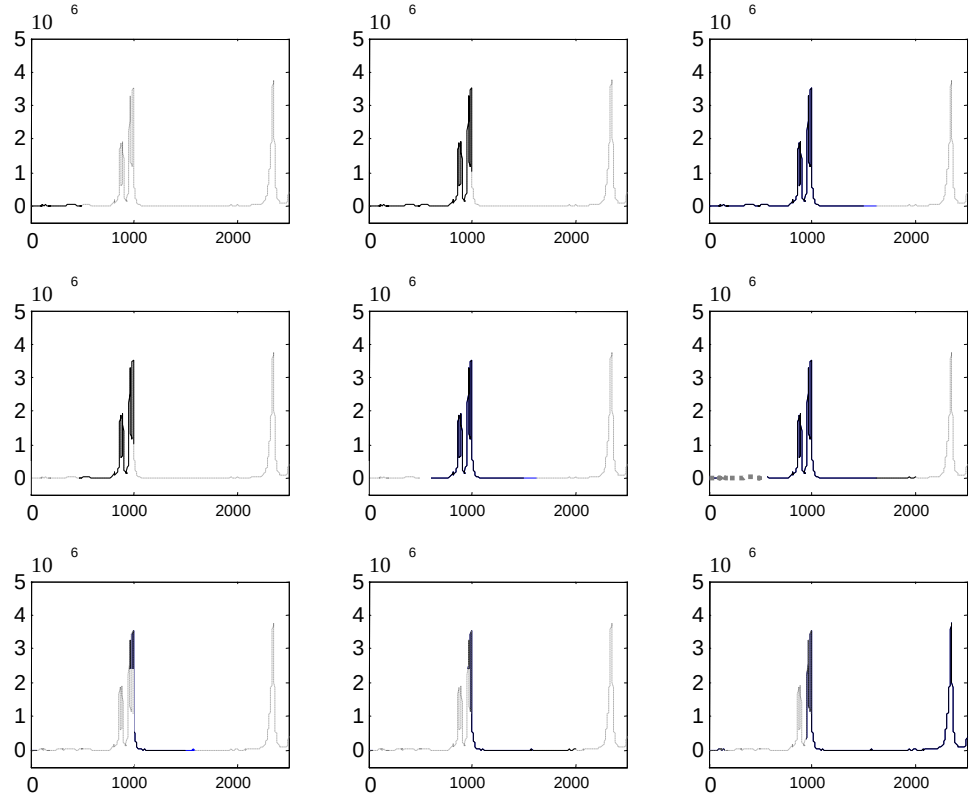


Figure 9 : Valeurs de RMSECV obtenu avec un nombre de composantes variant de 1 à 8, sur les différents intervalles définis par iPLS.

2.6.3. EWZS (Evolving Window Zone Selection)

En général, dans un spectre, plusieurs variables à la suite définissant un signal sont porteuses d'informations similaires. Ainsi, le spectre peut être découpé en zone de variables contiguës et chaque groupe de variables peut être testé par une méthode de régression ou classification. Evolving Window Zone Selection (EWZS) définit une zone spectrale à analyser comme une fenêtre qui se déplace le long du spectre dont la taille peut aussi s'agrandir. L'utilisateur définit une taille minimale et une taille maximale ainsi q'un pas, *p*, définissant le décalage de la fin de la fenêtre. La première fenêtre générée débute à la première variable et prend la largeur minimale. Puis sa taille augmente de p variables jusqu'à la taille maximale. Le point de départ de la fenêtre suivante, est décalé à la $p^{ème}$ variable. Ce processus de déplacement est répété ainsi de suite jusqu'à avoir parcouru le spectre entier (cf. Figure 10).



Légende :

Spectre

Fenêtre d'observation

p =500

taille minimale = 500

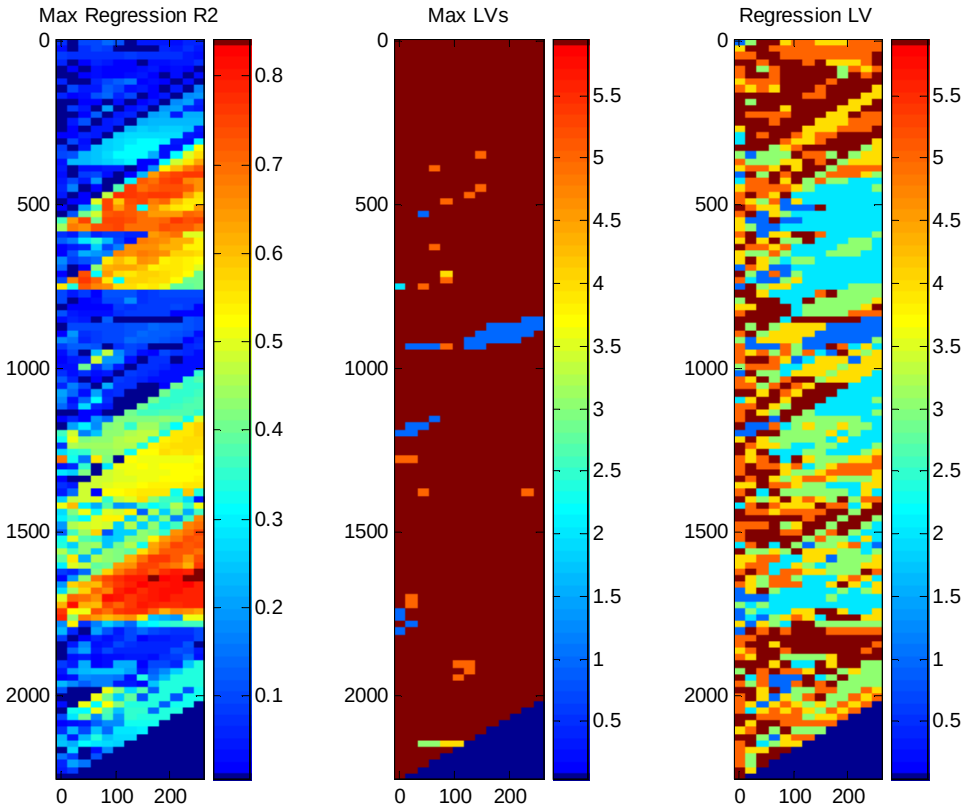
taille maximale = 1500

Figure 10 : Agrandissement de trois fenêtres d'observation le long du spectre

Sur chaque zone, une analyse (ACP, ACI, régression linéaire multiple) est réalisée pour tester l'efficacité des différentes combinaisons de *j* composantes pour la prédiction des résultats attendus. L'indicateur de la meilleure combinaison de composantes (R^2 : coefficient de détermination, ou RMSECV : la moyenne des carrés des écarts entre la valeur observée et la valeur obtenue par validation croisée) est retenu avec le nombre de composantes utilisées. Les résultats de ce calcul sont présentés sous forme de carte de couleur permettant de comparer visuellement les groupes de variable et leur qualité de prédiction.

La Figure 11 présente un exemple de résultats de ce calcul dans un cas où il y a 2241 variables pour prédire les valeurs observées de 65 échantillons. Le type d'analyse choisi est l'ACI. La taille maximale du signal à considérer est 250 variables, le pas est de 20 variables et

la taille minimale de la fenêtre est de 7 variables. Le nombre total de composantes extrait est 6.



<u>Légende :</u>

pas = 20

taille maximale = 250

taille minimale = 7

nombre de composantes à extraire = 6

Figure 11 : Cartes en couleurs des résultats d'une analyse par EWZS

Chaque pixel de la carte représente une fenêtre spectrale dont le point de départ est donné en ordonnée et sa taille en abscisse. Sa couleur est déterminée par la meilleure valeur de R^2 obtenue pour cette zone. Sur la première carte sont présentés les meilleurs résultats de R^2 pour l'ensemble des combinaisons de composantes testées. Ainsi des zones « d'intérêt » peuvent être détectées. La troisième carte présente le numéro de la composante qui a donné la meilleure valeur de R^2 pour cette fenêtre dans la première carte. La carte du milieu indique le nombre total de composantes indépendantes extraites et associées avec la meilleure valeur de

R^2. Ici les résultats présentés correspondent à ceux d'une régression linéaire simple, la même chose est obtenue pour des résultats d'une régression PLS.

2.6.4. CLV (Clustering of variables around Latent Variables)

2.6.4.1.Principe de CLV

Dans un spectre, des variables contiguës ou non, correspondent à des signaux provenant de mêmes molécules ou de mêmes mélanges et varient donc de façon semblable. Pour tenir compte de cette propriété, les variables peuvent être regroupées selon un critère de covariance ou de corrélation.

La technique CLV détaillée en Annexe 4, est très appropriée puisqu'elle a pour but de regrouper des variables autour de variables latentes selon leur covariance ou corrélation carrée. Elle fait une classification hiérarchique des variables en regroupant les variables maximisant le paramètre T qui prend en compte l'un ou l'autre des critères. Les variables latentes sont définies comme la première composante principale de chaque groupe établi.

Dans la Figure 12, ci-dessous, sont représentés les regroupements des variables.

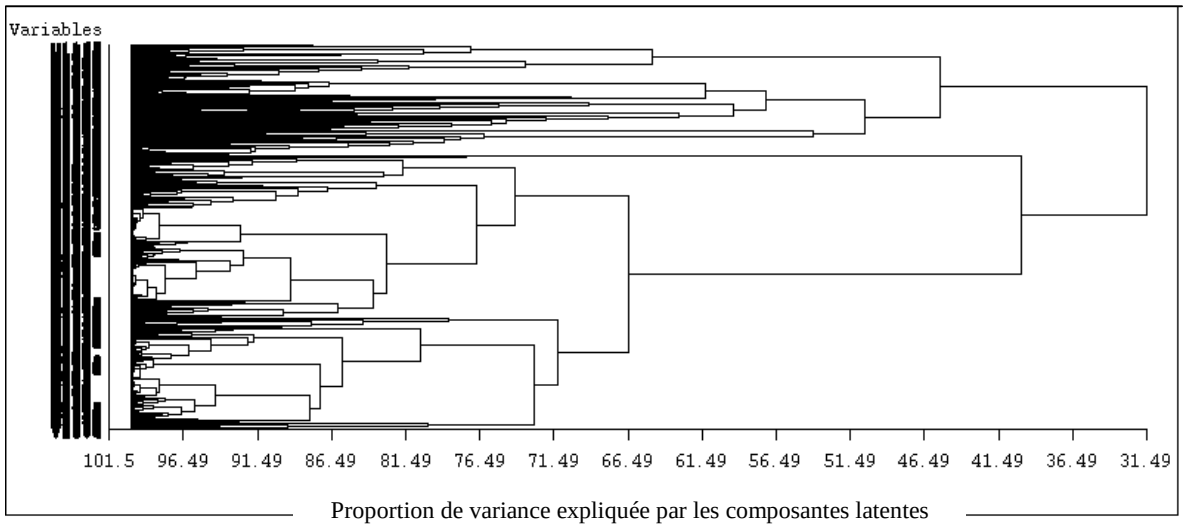


Figure 12 : Résultat de la classification hiérarchique par CLV

Afin de déterminer le nombre de groupes le plus représentatif des données, le suivi de l'évolution du critère T est déterminant (Figure 13). En effet, si la variation de T est importante lorsque le nombre de groupes diminue cela signifie qu'une part importante de l'information contenue dans la partition précédente a été perdue.

Dans cet exemple 6 ou 4 groupes peuvent être envisagés.

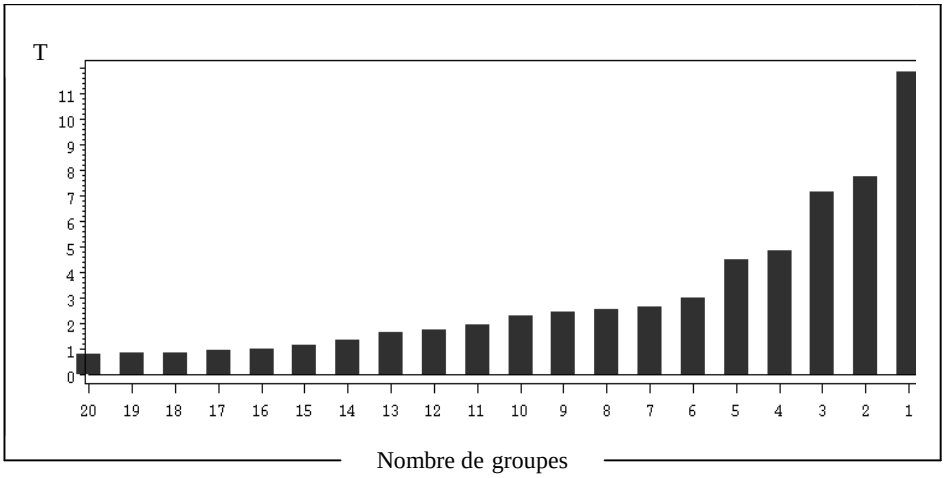


Figure 13 : Evolution du critère T lors de la classification hiérarchique par CLV sur les 20 derniers regroupements

2.6.4.2.Application à la sélection de variables

Une fois le nombre de groupes déterminé, chacun peut être analysé séparément. Dans notre étude, en vue de classer les individus dans différentes catégories connues, une analyse de la variance (ANOVA) sur chacune des variables latentes des groupes de variables a permis de sélectionner ou non un groupe de variables comme explicatif ou non de la séparation des individus.

2.6.5. PLS_Cluster

2.6.5.1.Principe de PLS_Cluster

PLS_Cluster [7, 8] est une méthode récursive qui a pour but de regrouper les individus ayant des caractéristiques semblables, sans *a priori* sur les échantillons et leur véritable proximité.

La première étape de PLS_Cluster sépare les échantillons en G groupes, selon la méthode décrite au paragraphe suivant. Chacun de ces groupes est lui-même soumis à l'algorithme PLS_Cluster, pour être séparé en G_i groupes. La procédure se poursuit jusqu'à ce que chaque échantillon soit dans un singleton et que l'on obtienne ainsi une hiérarchie. A chaque étape de cette classification hiérarchique, le nombre de subdivisions à chaque nœud est déterminé à partir de la structure interne des données.

A partir de la matrice de données (par exemple une matrice spectrale **X**), on crée une variable dépendante pouvant prendre une valeur comprise entre 0 et 1 (un vecteur **y**), de façon à pouvoir construire le modèle PLS qui permettra de grouper les échantillons. Dans l'algorithme

initial, cette variable dépendante était générée de façon aléatoire [7] et cette variable ne pouvait prendre que les valeurs 0 et 1. Dans l'algorithme généralisé, cette variable est initialisée comme étant le premier vecteur "score" normé de la matrice **X,** ce qui permet de tenir mieux compte de la structure des données [8]. Un modèle PLS est construit pour relier **y** à **X**, et grâce à ce modèle, on peut déterminer des valeurs prédites ("ajustées") $\mathbf{y}_p$. On teste la convergence entre **y** et $\mathbf{y}_p$: tant qu'il n'y a pas convergence, on remplace **y** par $\mathbf{y}_p$, et on recalcule un modèle PLS. Une fois la convergence atteinte, les valeurs sont ordonnées et les différences entre valeurs consécutives sont calculées (cf. Figure 14 a) pour déterminer le nombre de groupes présents.

Des échantillons appartenant au même groupe devraient avoir des valeurs de ***y*** similaires. Ainsi, une "grande" différence entre deux valeurs consécutives indique que les deux échantillons n'appartiennent pas au même groupe. On recherche donc les pics "significatifs" sur la Figure 14 b. Un seuil va être établi pour séparer les groupes. Dans l'algorithme utilisé, ce seuil est placé à 3 fois la médiane des différences entre valeurs consécutives. Si aucun groupe n'est trouvé, la valeur seuil sera diminuée de 0,2, jusqu'à ce que l'on détecte au minimum 2 groupes.

On peut ensuite appliquer une fonction d'appartenance aux valeurs de **y**, ce qui permet de classer les échantillons dans les différents groupes.

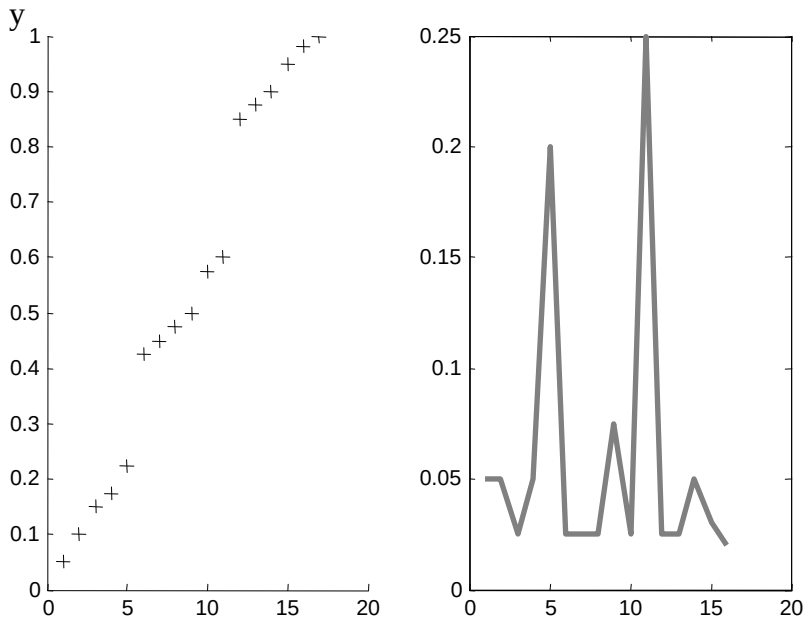


Figure 14 : a) Valeurs de y prédites par PLS_Cluster à l'étape j, ordonnées par valeur croissante ; b) Différence entre deux valeurs consécutives.

La fonction d'appartenance va regrouper les échantillons selon les valeurs

$$i / (p-1) \qquad \text{Equation 14.}$$

où p est le nombre de groupes déterminés et i varie entre 0 et p-1. Dans le cas présenté dans la Figure 14, les valeurs attribuées aux membres des trois groupes sont 0/2 ; 1/2 ; 2/2.

2.6.5.2.Application à la sélection de variables : Interval-PLS_Cluster

Pour appliquer cette méthode à la sélection de variables, on utilise l'algorithme PLS_Cluster sur des fenêtres spectrales d'une taille définie qui se déplacent le long des spectres. La fonction Interval-PLS_Cluster permet de parcourir l'ensemble du spectre en utilisant deux paramètres : la taille de la fenêtre spectrale et l'écart entre chaque fenêtre. Ainsi pour les variables à l'intérieur de chaque fenêtre, un dendrogramme est obtenu. L'analyse de ces dendrogrammes permet de sélectionner ceux qui séparent le mieux les échantillons selon les classes que l'on veut former. Seules les parties de spectre correspondantes à ces dendrogrammes seront retenues.

2.7. Méthodes descriptives multidimensionnelles

Lorsqu'un ensemble de m variables aléatoires doit être réduit pour en simplifier l'analyse, les p premiers axes de l'analyse en composantes principales ou indépendantes, correspond à un meilleur choix de représentation.

2.7.1. Analyse en composantes principales (ACP)

2.7.1.1.Objectif de l'ACP

L'analyse en composantes principales (ACP) consiste à rechercher les directions qui représentent le mieux la dispersion des individus dans l'espace multidimensionnelle des variables de départ, et donc à transformer l'espace de représentation des n individus [9]. Par ailleurs, cette technique permet de repérer des éléments présentant des valeurs aberrantes et/ou une structure sous-jacente aux données sans injecter d'information. Les principes théoriques sont décrits en Annexe 5.

2.7.1.2.En pratique

2.7.1.2.1. Données

Les données peuvent être structurées dans une matrice **X** à *n* lignes et *m* colonnes. Dans le cas des données RMN le nombre d'échantillon est largement inférieur au nombre de points (variables) obtenu par spectre, d'où $n \leq m$.

2.7.1.2.2. Résultats de l'ACP

Le nombre d'axes factoriels qu'il est possible d'extraire est $k = \min(n\text{-}1, m)$ donc n-1 dans le cas des spectres RMN. A partir de $\mathbf{X}_{(n,\ m)}$, il est possible de calculer les coordonnées factorielles des individus $\mathbf{C}_{(n,\ k)}$ avec $k=\min(n\text{-}1,\ m)$, dans la nouvelle base de vecteurs propres $\mathbf{P}_{(k,\ m)}$. Celle-ci permet de tenir compte de la corrélation linéaire entre les *m* variables pour fournir *k* facteurs décorrélés, combinaisons linéaires des *m* variables initiales. Chaque vecteur propre, linéairement indépendant des autres, est lié à un pourcentage τ de la variance totale des données σ^2. τ décroît en fonction de l'ordre des facteurs : le premier facteur contient la plus grande partie de la variance, le second la plus grande partie de la variance résiduelle, et ainsi de suite jusqu'au *k*-ième facteur.

En conservant les k facteurs, τ = 100 % de σ^2 et

$$\mathbf{X}_{(n,m)} = \mathbf{C}_{(n,k)}\,\mathbf{P}_{(k,m)}^{'} \qquad \text{Equation 15.}$$

où **C** est la matrice contenant les coordonnées des individus dans la nouvelle base de vecteurs propres et **P** la matrice contenant les vecteurs propres.

Grâce à différents critères qui permettent de mesurer la signification des facteurs, il est possible de n'en conserver qu'un nombre limité ($j < k$). Ainsi la base de données peut être décrite avec un nombre de facteurs *j* plus petit que le nombre initial *k* sous l'hypothèse que le résidu négligé est peu informatif.

$$\mathbf{X}_{(n,m)} = \mathbf{C}_{(n,j)}\,\mathbf{P}_{(j,m)}^{'} + \mathbf{E}_{(n,m)} \qquad \text{Equation 16.}$$

où **C** est la matrice contenant les coordonnées des individus dans la nouvelle base de vecteurs propres, **P** est la matrice contenant les vecteurs propres et E la matrice de résidus.

2.7.1.2.3. Calcul des coordonnées de nouveaux individus

Une fois la base de vecteurs propres fixée, il est possible de calculer les coordonnées de *nouveaux* individus dans cette base afin de les comparer aux anciens individus.

Cette propriété sert en général à valider un modèle de prédiction en injectant des données de validation.

2.7.2. Analyse en composantes indépendantes (ACI)

2.7.2.1.Objectif de l'ACI

L'analyse en composantes indépendantes est une technique qui découle de la recherche de méthodes de séparation de sources, qui a commencé, il y a une vingtaine d'année [10]. L'objectif de l'ACI est d'estimer *h* signaux "sources", supposés stationnaires, ergodiques et indépendants, en utilisant *k* signaux "observés" *($k \geq h$)* qui sont des mélanges inconnus des signaux "sources" [11-16]. L'exemple le plus courant est celui de la séparation de voix différentes provenant de narrateurs (sources) différents et perçues par des récepteurs situés à différents endroits de la salle. Ces signaux peuvent être supposés, pour des raisons logiques, comme mutuellement statistiquement indépendants. L'ACI permet de retrouver les différents discours initiaux à partir des enregistrements. Les principes théoriques sont décrits en Annexe 6.

2.7.2.2.En pratique

2.7.2.2.1. Données

Les données peuvent être structurées dans une matrice **X** à *n* lignes et *m* colonnes. Dans le cas des données RMN le nombre d'échantillon est largement inférieur au nombre de points (variables) obtenu par spectre, d'où $n \leq m$.

2.7.2.2.2. Résultats de l'ACI

Le nombre de sources recherchées, *h*, est fixé. Dans cette technique le but est d'estimer $\mathbf{S}_{(m, h)}$, la matrice de séparation. L'hypothèse de départ pour résoudre ce type de problème est l'indépendance des sources. Il en existe plusieurs façons de procéder pour séparer les sources. Dans cette étude le choix s'est porté sur l'algorithme JADE (Joint Approximate Diagonalization of Eigen-matrices), développé par Jean-François Cardoso [10] et qui se base sur les cumulants d'ordre 4 [17]. La nouvelle base de sources obtenue, $\mathbf{S}_{(m, h)}$, est constituée de vecteur-colonne correspondant chacun à une composante indépendante.

Il est important de noter que la structure des composantes indépendantes peut varier avec le nombre de composantes calculées, contrairement à l'ACP.

2.7.2.2.3. Calcul des coordonnées de nouveaux individus

De même qu'en ACP, une fois la base de sources trouvée, il est possible de calculer les coordonnées de nouveaux individus dans cette base afin de les comparer aux anciens individus.

Soit $\mathbf{X_1}_{(p,\ m)}$, une matrice de nouveaux échantillons, leurs coordonnées $\mathbf{C}_{(p,\ h)}$ dans la nouvelle base sont déterminées par la formule suivante :

$$\mathbf{C}_{(p,h)} = \mathbf{X_1}_{(p,m)} \,.\, \mathbf{S}_{(m,h)} \,.\, \mathrm{inv}(\mathbf{S}_{(m,h)}' \,.\, \mathbf{S}_{(m,h)}) \qquad \text{Equation 17.}$$

2.7.3. Régression au sens des moindres carrés partiels (Partial Least Squares - PLS)

2.7.3.1. Objectifs de la régression PLS

La régression PLS est une procédure qui permet de modéliser la relation entre *m* variables « explicatives » $\mathbf{X_1}$, …, $\mathbf{X_m}$, prises sur *n* échantillons, dans la matrice $\mathbf{X}_{(n,\ m)}$ et *j* variables « endogènes », $\mathbf{Y_1}$, …, $\mathbf{Y_j}$, dans la matrice $\mathbf{Y}_{(n,j)}$. Cette approche, qui fût introduite pour la première fois par Wold en 1966 [18-20], présente l'avantage, par rapport à la régression linéaire multiple classique, d'accepter plus de variables explicatives que d'échantillons. De plus, lorsqu'il y a beaucoup de variables on risque d'obtenir un modèle sur-ajusté, c'est-à-dire modélisant une partie du bruit. Correctement utilisée, la régression PLS permet d'éviter ce phénomène, car étant une régression séquentielle, on peut arrêter le processus de régression avant de modéliser l'erreur. De plus, elle permet de s'affranchir des problèmes de multicolinéarité qui gênent la régression linéaire multiple.

La régression PLS classique est un cas particulier de la méthode NIPALS (Non-linear Iterative Partial Least-Squares). A chaque étape, l'information expliquée par la variable latente trouvée est soustraite à la matrice **X** jusqu'à ce que la variance liée à l'information soit totalement expliquée et qu'il ne reste plus que le bruit [21,22]. La différence majeure entre la régression PLS et la régression PCR (régression sur Composantes Principales) est que les composantes PLS sont optimisées pour être les plus prédictives de **Y**, contrairement aux composantes principales. Le modèle est le suivant :

$$\mathbf{Y}_{(n,j)=}\, \mathbf{X}_{(n,m)} \,.\, \mathbf{B}_{(m,j)} + \mathbf{E}_{(n,j)} \qquad \text{Equation 18.}$$

où $\mathbf{E}_{(m,\ j)}$ est la matrice des écarts.

2.7.3.2.En pratique

Dans notre étude, la matrice $\mathbf{X}_{(n,\ m)}$ représente les spectres des échantillons et la matrice $\mathbf{Y}_{(n,\ j)}$ représente les concentrations à prédire.

2.8. Les méthodes de classement

Dans le contexte de l'authenticité des produits, une question fondamentale revient systématiquement : le produit est-il authentique ou non ? Pour le prouver ou pour démontrer le contraire des méthodes de classification sont particulièrement bien adaptées. De plus les échantillons authentiques présentent une variabilité intrinsèque importante qu'il faut prendre en compte, ce que permettent ces méthodes.

2.8.1. Classement

Une fois les variables sélectionnées, en utilisant notamment les techniques décrites dans la partie 2.6, et réorganisées, au besoin, en utilisant des techniques de changement de représentation, vues dans la partie 2.7, il s'agit de les combiner pour obtenir un classement des groupes d'individus aussi net que possible.

En général un échantillon est attribué au groupe dont il est le plus proche. Il s'agit donc de quantifier la proximité d'un échantillon à un autre ou à un ensemble d'autres. Une technique utilisée dans cette étude est de calculer des distances par rapport au barycentre de groupes prédéfinis. Il peut s'agir de la distance euclidienne classique où la distance entre deux points est la longueur du segment qui les relie. Soit A (a_1, a_2…a_J) et B (b_1, b_2…b_J), la distance entre A et B est d,

$$d=\sqrt{\sum_{i=1}^{J}(a_i-b_i)^2} \qquad \text{Equation 19.}$$

Par ailleurs, d'autres types de distance existent, telle que la distance de Mahalanobis qui permet de prendre en compte les variances et les covariances entre les variables utilisées. Ainsi, à la différence de la distance euclidienne où toutes les composantes des vecteurs sont traitées de la même façon, la distance de Mahalanobis accorde un poids moins important aux composantes qui présentent la plus grande variance.

Soit a et b, les vecteurs de cordonnées de A et B, la distance de Mahalanobis qui sépare A et B est :

$$d_M=\sqrt{(a-b)^T\sum{}^{-1}(a-b)} \qquad \text{Equation 20.}$$

où Σ est la matrice de variance-covariance entre les J composantes.

2.8.2. Validation et prédiction

Pour valider un modèle de classification ou prédire l'appartenance d'un échantillon à un groupe, une série d'échantillons est utilisée pour calculer le modèle prédictif. Le barycentre de chaque groupe d'échantillons connus est ensuite calculé et la distance entre chaque nouvel échantillon à classer et les différents barycentres est calculée. Le groupe le plus proche de l'échantillon lui est alors attribué.

Lors de cette validation du modèle, la comparaison des groupes attribués et des groupes réels permet de caractériser le modèle par un taux de bons classements.

2.9. Références

1. Hoult, D. I. & Richards, R. E. The signal-to-noise ratio of the nuclear magnetic resonance experiment. *J. of Magnetic Resonance* **24**, 71-85 (1976).
2. Canet Soulas E., Boubel J.-C. & D., C. *La RMN. Concepts, méthodes et applications*, pp. 235 (Dunod, Paris, 2002).
3. Hoch, J. C. & Stern, A. S. *NMR data processing*, pp. 196 (Wiley-Liss, New York, 1996).
4. Tomasi, G., van den Berg, F. & Anderson, C. Correlation Optimized Warping and time warping as pre-processing methods for chromatographic data. *J. Chemometrics* **18**, 231-241 (2004).
5. Norgaard, L. et al. Interval Partial Least-Squares Regression (iPLS): A Comparative Chemometric Study with an Example from Near-Infrared Spectroscopy. *Applied Spectroscopy* **54**, 88A-113A (2000).
6. Borin, A. & Poppi, R. J. Application of mid infrared spectroscopy and iPLS for the quantification of contaminants in lubricating oil. *Vibrational Spectroscopy* **37**, 27-32 (2005).
7. Barros, A. S. & Rutledge, D. N. PLS_Cluster: a novel technique for cluster analysis. *Chemom. Intellig. Lab. Syst.* **70**, 99-112 (2004).
8. Jouan-Rimbaud Bouveresse, D., Barros, A. S. & Rutledge, D. N. Generalised PLS_Cluster: an extension of PLS_Cluster for interpretable hierarchical clustering of

multivariate data. *Sensing and Instrumentation for Food Quality and Safety* **1**, 79-90 (2007).

9. Joliffe, I. T. *Principal Component Analysis*, pp. 271 (Springer-Verlag, New York, 1986).
10. Cardoso, J.-F. Blind separation of real signals with JADE. http://www.tsi.enst.fr/icacentral/Algos/cardoso/ (1995).
11. Shimizu, S., Hyvarinen, A., Kano, Y. & Hoyer, P. O. in *21st Conference of Uncertainty in Artificial Intelligence*, 526–533 (Edinburgh, 2005).
12. Le Borgne, H. & Guérin-Dugué, A. in *ORASIS*, (Cahors, 2001).
13. Hyvarinen, A., Hoyer, P. O. & Inki, M. Topographic independent component analysis. *Neural Computation* **13**, 1527-1558 (2001).
14. Hyvarinen, A. Survey on independent component analysis. http://www.cs.utexas.edu/~kuipers/readings/Hyvarinen-ncs-99.pdf, 05/01/2008, (1999).
15. Durieu, C. & Kieffer, M. Analyse en composantes indépendantes pour la séparation aveugle de sources. http://mfca.ups-tlse.fr/cetsis/Docs/Articles/Durieu_Cecile.pdf, 05/01/2008, (2003).
16. Hoyer, P. O. Independent component analysis in image denoising. http://www.cs.helsinki.fi/u/phoyer/papers/pdf/dippa.pdf, 05/01/2008, (1999).
17. Cardoso, J.-F. Analyse en composantes indépendantes. http://www.tsi.enst.fr/~cardoso/Papers.PDF/jsbl02-long.pdf, 05/01/2008, (2002).
18. Wold, H. in *Encyclopaedia of statistical sciences* (ed. Kotz, S., Johnson, N.L), 581-591 (Wiley, New York, 1985).
19. Wold, H. in *Multivariate analysis*, 391-420 (Academic Press, New York, 1966).
20. Wold, H. in *Systems under indirect observation* (ed. Joreskog, K. G., Wold, H.), 1-55. (Elsevier Science Ltd, Amsterdam, 1982).
21. Tenenhaus, M. L'approche PLS. *Revue de Statistique Appliquée* **47**, 2-55 (1999).
22. Tenenhaus, M. *La régression PLS*, pp. 254 (Editions Technip, Paris, 1998).

3. Synthèse des résultats

3.1. Choix de produits agroalimentaires « à risque » économique

Nous avons choisi de travailler sur des produits qui étaient susceptibles d'être adultérés pour des raisons économiques et dont les méthodes de détermination de l'authenticité sont soit inexistantes soit longues et fastidieuses.

3.1.1. Le jus d'orange

Le jus d'orange est un produit de grande consommation dont la qualité varie selon que le process utilisé : jus à base de concentré ou pur jus. La fraude économique sur ce produit n'est pas un problème récent [1] et les autorités ont mis en place un certain nombre de mesures pour les prévenir. Au niveau européen, l'industrie des jus de fruit tend à mettre en place un système de contrôle de la qualité généralisé, le système EQCS (European Quality Control System). Il aspire à un marché de libre et juste concurrence grâce notamment à la détection des fraudes et une interprétation harmonisée des résultats d'analyse.

L'outil le plus utilisé pour établir la conformité d'un jus sont les valeurs guide de l'AIJN (Association de l'Industrie des Jus et Nectars) qui définit les limites maximales et/ou minimales d'un grand nombre de paramètres entrant dans la composition de 19 jus différents et les méthodes de références à utiliser. Pour vérifier la conformité d'un jus avec cette liste de paramètres un nombre important d'analyses doit donc être réalisé.

Les fraudes majeures qui ont été recensées sont la dilution, l'ajout de sucre, l'ajout d'autres substances pour camoufler la dilution, l'ajout de co-fruits, l'appellation abusive de jus « pur jus » [2].

Dans notre étude nous nous sommes intéressés à la fraude par ajout de pamplemousse (GJ) dans le jus d'orange (OJ) que nous avons modélisé en achetant des jus de fruits mono-fruit ou en mélange (OG). Le choix de produits du marché a été fait pour se rapprocher au plus près des jus commerciaux. En effet le procédé et en particulier le pressage industriel [3] et la pasteurisation, que nous n'aurions pas pu réaliser, modifient la composition des jus.

Cette fraude peut être détectée par la méthode officielle IFU 58 qui établie par chromatographie (HPLC) la quantité d'hespéridine et de naringine.

Ce type d'échantillon a été choisi tout d'abord pour l'intérêt économique qu'il représente pour les industries, et pour valider les potentiels de la spectroscopie RMN ^{1}H pour la détection de

marqueurs et pour étudier les possibilités de discrimination d'échantillons. Les spectres attribués des jus de pamplemousse et d'orange sont présentés en Annexe 7.

3.1.2. Le vinaigre balsamique

Le vinaigre balsamique est une spécialité italienne obtenue à partir de la fermentation acétique de moûts de raisin blanc réduits. La valeur de ce produit dépend de son mode de production. S'il s'agit de vinaigre balsamique produit de façon traditionnelle son prix pourra atteindre les 800 euros / litre pour une qualité supérieure (25 ans d'affinage), alors qu'un bon vinaigre balsamique (3 ans de vieillissement) pourra être trouvé sur le marché à un prix beaucoup plus abordable : 60 euros / litre.

Cette différence de prix se justifie par la faible superficie de la zone de production qui est limitée par une appellation d'origine contrôlée (AOC) aux régions de Modène et d'Emilia ainsi que le procédé particulier datant du XI siècle. En effet, c'est en faisant vieillir du moût réduit de raisin de type Trebbiano au minimum 12 ans en le transférant au fur et à mesure dans des tonneaux de bois différents et de volume décroissant que l'on affine le vinaigre traditionnel balsamique. Le vinaigre traditionnel balsamique de qualité supérieur est vieilli 25 ans et porte l'appellation « extravecchio ». La biochimie de cette maturation du vinaigre est encore peu élucidée.

Le vinaigre balsamique commun est un produit industriel fabriqué à grande échelle et obtenu à partir de moûts réduits auxquels on ajoute du vinaigre de vin et un colorant : le caramel dans la limite de 2 %. La maturation du produit dure 60 jours dans un fût en bois.

Avant d'être mis sur le marché, le vinaigre balsamique traditionnel est seulement soumis à un test sensoriel [4]. Cependant un vinaigre balsamique commun de bonne qualité présente de nombreux caractères organoleptiques similaires : la couleur, la densité, un goût proche. L'absence de méthode officielle plus élaborée et basée sur des paramètres physico-chimiques plus précis ne permet pas d'éviter la contrefaçon.

De nombreuses études ont été menées pour trouver des marqueurs de discrimination des deux types de vinaigre. Certaines ont caractérisé les éléments spécifiques au vieillissement [5-11]. La RMN ^{1}H a été testée pour quantifier des constituants spécifiques [12]. Une autre étude a utilisé la RMN et l'analyse en composantes principales (ACP) [13] pour faire du "profiling" par zone ce qui a donné de bons résultats de discrimination, mais difficiles à interpréter. Cependant une approche profiling globale n'a jamais été menée, ni l'utilisation d'outils chimiométriques plus spécifiques à l'interprétation des signaux. Les spectres attribués des vinaigres balsamiques traditionnels ou non sont présentés en Annexe 10.

Le choix de l'étude du vinaigre balsamique se positionne, comme pour celui du jus d'orange, sur une problématique économique et un manque de solution analytique performante.
Le jeu de données est constitué de 29 échantillons de vinaigre : 15 vinaigres traditionnels balsamiques avec des âges différents, et 14 vinaigres balsamiques (cf. *Publication 4).*

3.1.3. La qualité de la préparation de fruit dans un yaourt

Pour des questions de nutrition et de santé publique les gouvernements émettent des recommandations nutritionnelles. En France notamment, le Programme National Nutrition Santé recommande la consommation de 3 produits laitiers par jour afin de couvrir les besoins en calcium qui varient, selon les tranches d'âge, de 900 à 1200 mg par jour. De plus, il est conseillé de privilégier les produits laitiers à faible teneur en matière grasse et donc du lait ou des produits laitiers frais (PLF) représentant 40 % du marché en 2003 [14].
Parallèlement, le marché mondial des PLF est en forte croissance notamment dans les pays émergents où la consommation moyenne est nettement inférieure à la moyenne de 22 kg en Europe occidentale. Sur ce marché le yaourt occupe une place de très grande importance et ceci en partie grâce à la diversification du marché qui attire de nouveaux consommateurs. Ces vingt dernières années l'offre s'est démultipliée en proposant, en plus des produits classiques, des yaourts aromatisés ou aux fruits, des produits allégés ou au lait entier, des produits avec des morceaux de fruits entiers, des yaourts pulpés, brassés, des probiotiques, des yaourts d'origine animale autre que vache, des yaourts à boire, des produits enrichis en calcium, en oméga-3 ou encore en stérols végétaux.
L'analyse des PLF est complexe, et un des paramètres, le plus difficile à évaluer est la quantité de fruits en présence. Les méthodes existantes [2] requièrent la mesure de nombreux paramètres et donnent des résultats peu précis. De plus, au niveau de l'industrie, la qualité des préparations de fruit n'est pas aisée à vérifier du fait de la stérilité des containers. Les fruits créant la valeur ajoutée du yaourt, des fraudes portant sur ce paramètre pourraient avoir des conséquences économiques importantes.
C'est à cause du manque d'analyses sur un marché d'une telle ampleur et présentant des risques économiques forts, nous avons décidé d'investiguer les possibilités de méthodes de profiling sur les yaourts.
Le jeu de données est constitué d'extraits de yaourts à la fraise soit aromatisé, soit avec des morceaux de fruits, soit avec des fruits mixés ou de la purée de fruit. Pour plus de détails sur les échantillons, se reporter aux articles. Les méthodes de préparation de l'échantillon et

d’acquisition du spectre sont développées en Annexe 11 et le spectre attribué partiellement d’un extrait de yaourt nature est présenté en Annexe 12.

3.2. Acquisition des données spectrales

Après avoir optimisé les paramètres d'acquisition sur les différents types d'échantillons et mesuré les spectres RMN, à l'exception du vinaigre dont les spectres ont été acquis par Augusta Caligiani du département de Chimie Organique et Industrielle de l'Université de Parme, nous avons mis en oeuvre des pré-traitements pour homogénéiser les signaux RMN, et puis procéder à des analyses chimiométriques pour caractériser les échantillons.

Pour des raisons de lisibilité et de cohérence nous avons décidé de présenter en détail ici les résultats obtenus sur un seul jeu de données : les données des jus de fruits orange et pamplemousse et des mélanges industriels. Il y a 92 échantillons et 14999 déplacements chimiques enregistrés. La matrice initiale des spectres a donc pour dimension : 92 x 14999.

3.3. Résultats sur la différence entre ACP et ACI

Pour évaluer la différence des résultats obtenus par ACP et ACI sur les spectres RMN ^{1}H des échantillons, nous avons tout d’abord comparé les résultats de la décomposition des spectres en composantes principales et indépendantes et ensuite leurs contributions. En effet, l’ACI étant une méthode visant à extraire des sources "pures" il semblait possible que la décomposition par ACI de signaux RMN de mélange de produit permette de séparer les signaux originaux.
Nous avons donc comparé tout d’abord visuellement les contributions factorielles de ces deux types de décompositions.

Dans le jeu de données des 92 jus de fruits, nous avons sélectionné la zone des composés flavonoïdes entre 6,00 ppm et 8,05 ppm. Dans cette zone se trouvent les signaux de deux marqueurs de la distinction entre jus d’orange (OJ) et jus de pamplemousse (GJ) selon la méthode officielle de détection IFU 58 : la naringine et l’hespéridine (cf. Figure 15 et Tableau 1). La matrice de cette zone pour les 92 spectres a été décomposée en 3 composantes indépendantes et en 3 composantes principales.

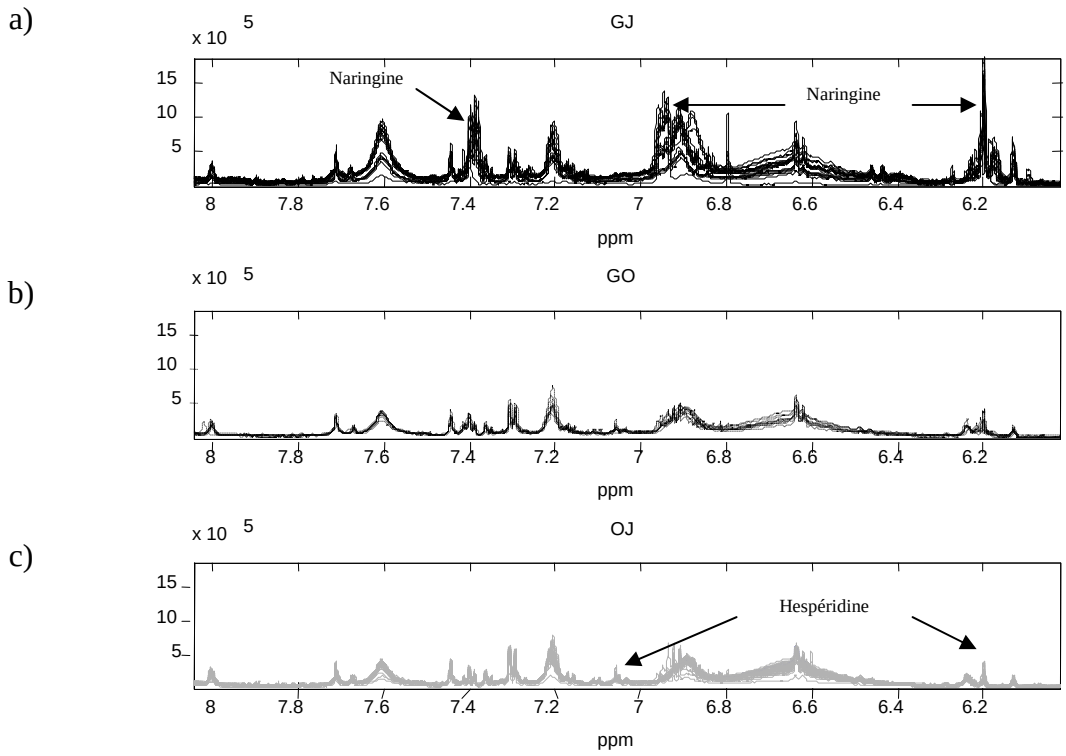


Figure 15 : Zone spectrale de jus de fruits de 6,00 à 8,05 ppm : a) 23 jus de pamplemousse ; b) 10 jus de mélange orange pamplemousse ; c) 59 jus d'orange

Tableau 1. ***Attribution des signaux de la zone 6,00 ppm et 8,05 ppm***

Déplacement chimique (ppm)	**Composé**	**Déplacement chimique (ppm**	**Composé**
6,20	Naringine	7,21	Polyphénols
6,23	Hespéridine	7,30	Tyrosine
6,62	Polyphénols	7,39 (d)	Naringine
6,64	Arginine	7,40	Hespéridine
6,89	GABA	7,60	GABA
6,92 (d)	Tyrosine	7,67	Polyphénols
6,95	Naringine	8,02	Niacine
7,06	Hespéridine		

Les décompositions présentées dans la Figure 16 dans les différents espaces permettent de mettre en évidence la non-significativité physique des composantes principales. En effet, hormis la première composante, les composantes principales comportent des morceaux de spectre positifs et négatifs, ce qui indique qu'ils correspondent à des mélanges de

contributions spectrales différentes. En ICA, au contraire, les composantes ressemblent à de vrais spectres, auxquels une signification physique peut être attribuée.

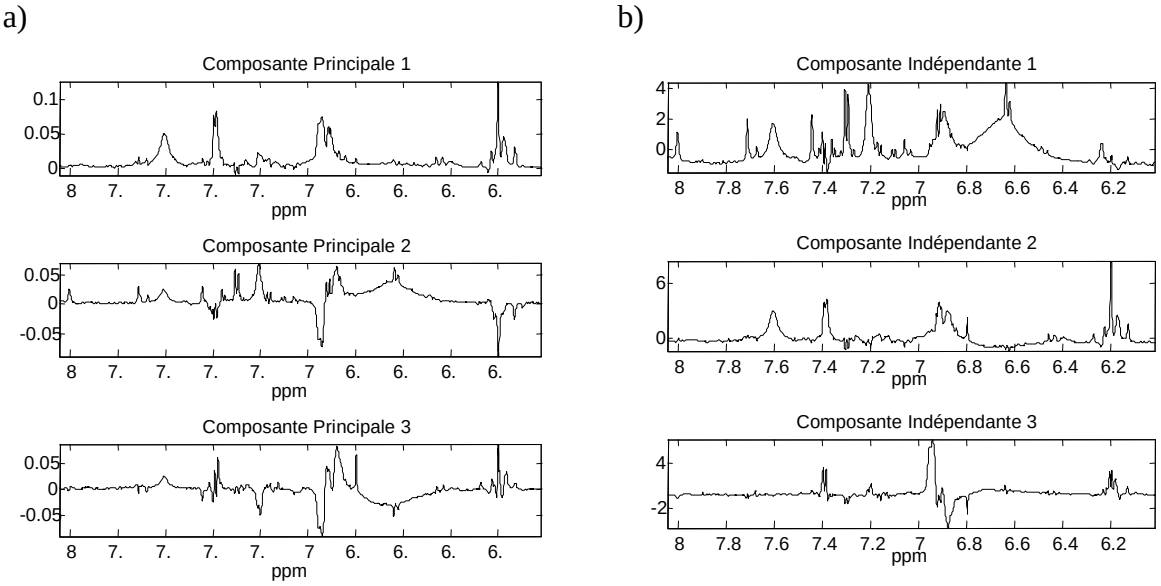
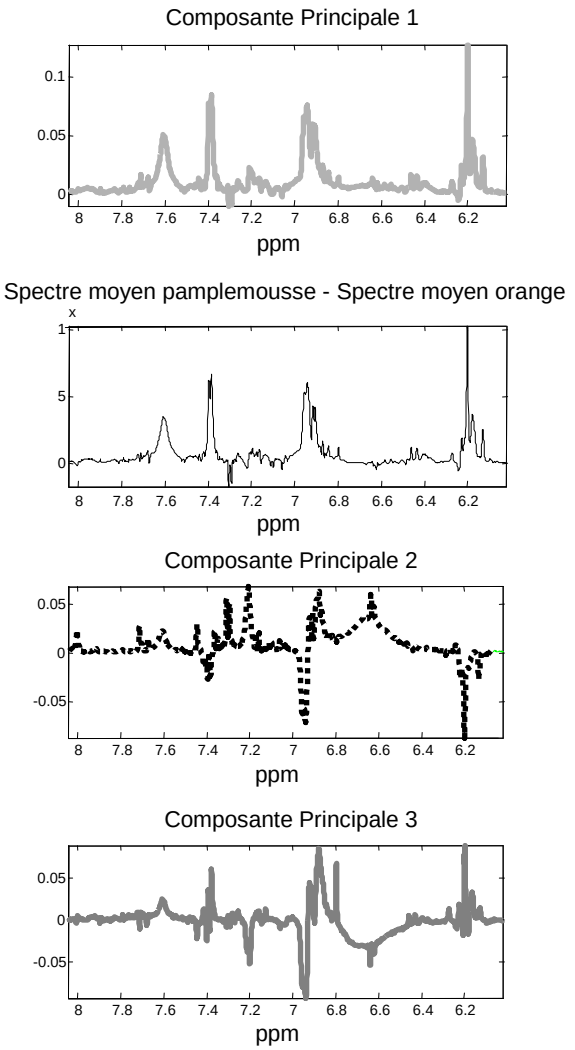


Figure 16 : Décompositions des signaux en composantes factorielles : a) composantes principales ; b) composantes indépendantes

L'ACP recherche les directions de l'espace qui maximisent la dispersion des individus, ce qui fait que, une fois les pics expliquant la différence entre les jus d'orange et les pamplemousses enlevés, (composante 1), les autres composantes ne sont pas faciles à expliquer en termes physico-chimiques (cf. Figure 16 a). Par contre, en ACI, une fois enlevée la première composante représentant le spectre moyen de l'orange, la méthode trouve une deuxième source : la composante indépendante 2 qui ressemble beaucoup au spectre moyen du pamplemousse moins celui de l'orange (Figure 16 b).

La 3ème source en ACI reprend les trois zones où sont localisées les signaux de la naringine uniquement (Figure 17 et tableau 1) ce qui explique que cette composante est très discriminante et la position des jus d'orange proche de l'origine dans le plan IC2-IC3 (cf. Figure 18 b). Comme on l'a vu, l'ACI est une technique d'extraction de sources. Si on enlève la valeur moyenne de IC 2 au signal moyen du pamplemousse moins celui de l'orange on obtient un signal très proche d'IC3.

a)



b)

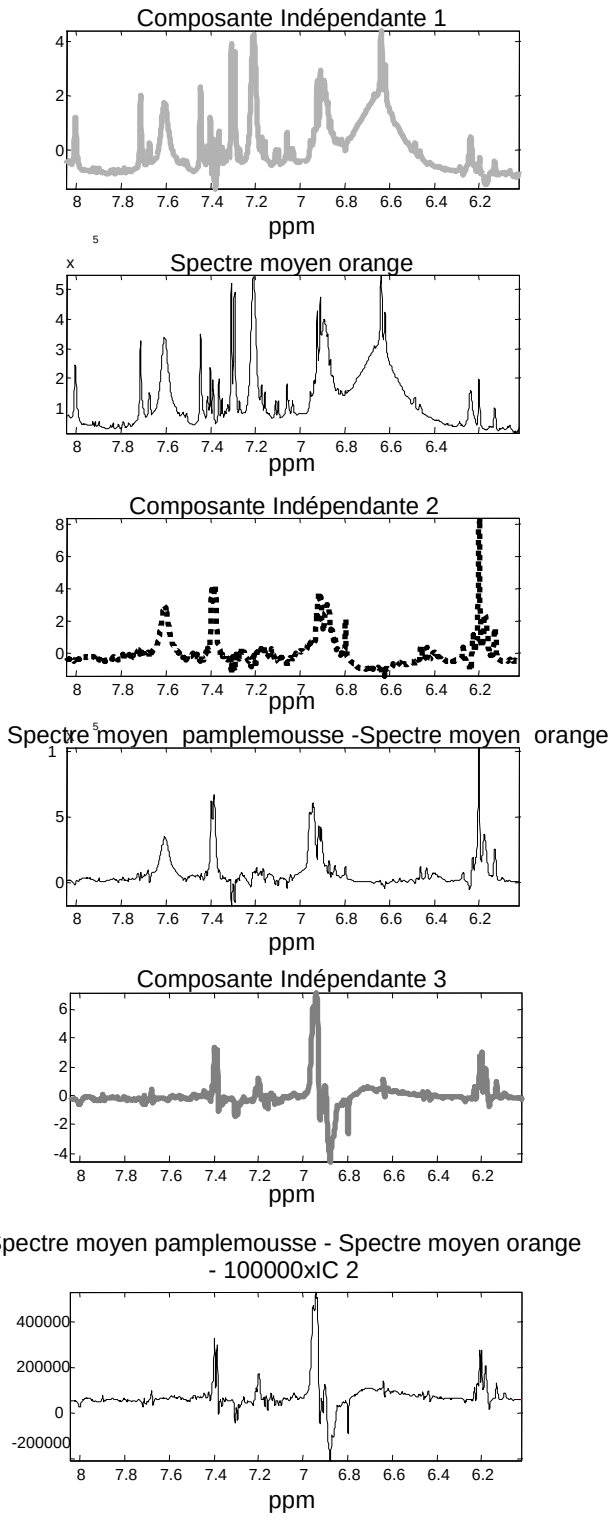


Figure 17 : Relation entre les composantes et les signaux des spectres moyens de l'orange et du pamplemousse : a) composantes principales ; b) composantes indépendantes

Etudions maintenant les coordonnées factorielles des deux méthodes d'analyse. Lorsque les individus sont projetés dans l'espace des composantes principales et indépendantes, une bonne séparation des individus est obtenue sur la composante représentant la différence entre jus de pamplemousse et jus d'orange (cf. Figure 18), respectivement la composante principale 1 et les composantes indépendantes 2 et 3.

a) b)

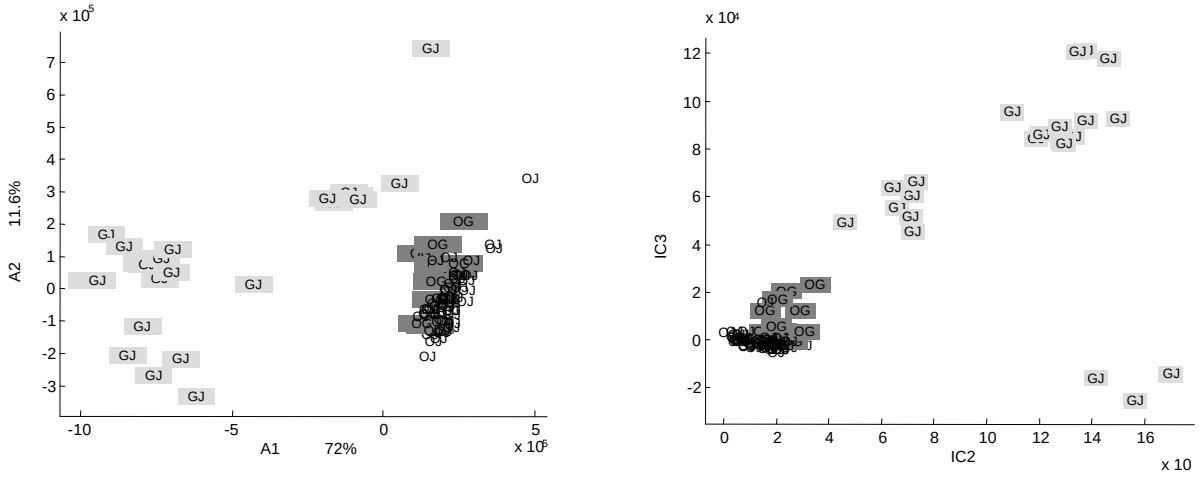
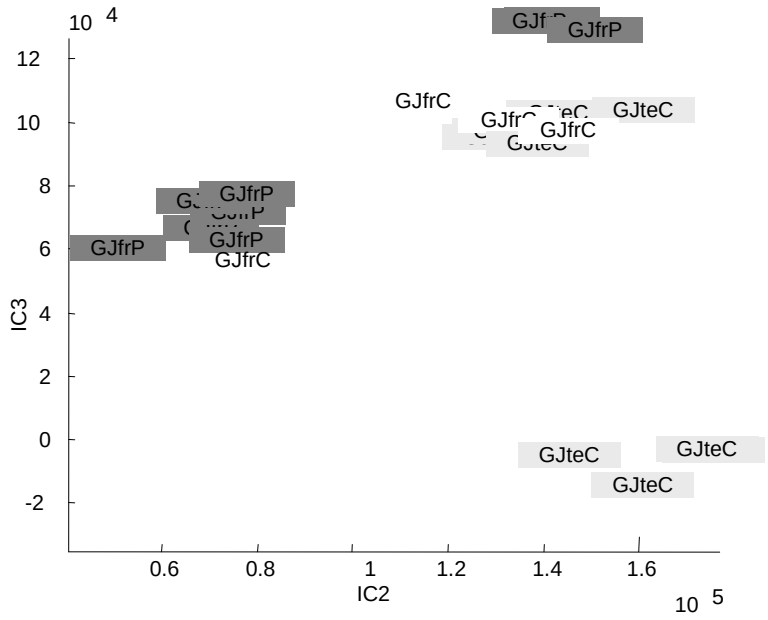


Figure 18 : a) Projections des individus dans l'espace des composantes factorielles : composantes principales ; b) composantes indépendantes

La projection des individus sur le plan des composantes indépendantes 2 et 3 (Figure 18) met en évidence trois sous-populations de pamplemousse. Ayant trois types de process différents pour le pamplemousse : les jus de fruits "pur jus" vendu au rayon frais et donc à courte conservation, les jus de fruits à base de concentré à courte conservation ou longue conservation, nous avons projeté ces sous-populations dans le plan IC2-IC3 dans la Figure 19.

Cette Figure montre que la différence entre les individus ne vient pas des 3 process différents. Cependant on remarque que le groupe « pur jus » se différencie des jus « à base de concentré » sur IC2. En effet, ils présentent majoritairement des valeurs sur les composantes 2 et 3 plus faibles que celles obtenues pour les jus fabriqués « à base de concentré ». La différence que l'on peut constater visuellement est située dans la zone de 6,8 à 7,0 ppm (Figure 20 a et b), les intensités des pics de la naringine (6,95 ppm) et GABA sont plus faibles (6,91 ppm).



Légende :
GJ : jus de pamplemousse
fr : vendu au rayon frais
te : vendu en Tetrapak®
P : jus « pur jus »
C : jus à base de concentré

Figure 19 : Projections des individus dans le plan des composantes indépendantes 2 et 3

Les 4 produits avec des valeurs négatives sur IC3 (en vert sur la Figure 20 a et b) sont des produits « à base de concentré ». Ils présentent des spectres avec une composition des signaux dans la zone 6,70 à 7,00 ppm qui est différente des autres produits « à base de concentré » (en rouge et noir sur la Figure 20 a et b). Tout d'abord, on note la présence d'un pic à 6,79 ppm (pic non attribué) qui est absent dans la majorité des autres jus (exception d'un seul jus « pur jus »). Ensuite, les pics de la naringine à 6,95 ppm dans les autres jus présentent un léger décalage et se retrouvent à 6,92 ppm, il en est de même pour le GABA normalement présent à 6,91 ppm qui est à 6,88 ppm. Ce décalage met en évidence l'intérêt du "warping" pour ces quelques spectres, nous verrons ce prétraitement par la suite. Le décalage de ces échantillons selon l'axe IC3 ne serait donc qu'un artefact.

a) b)

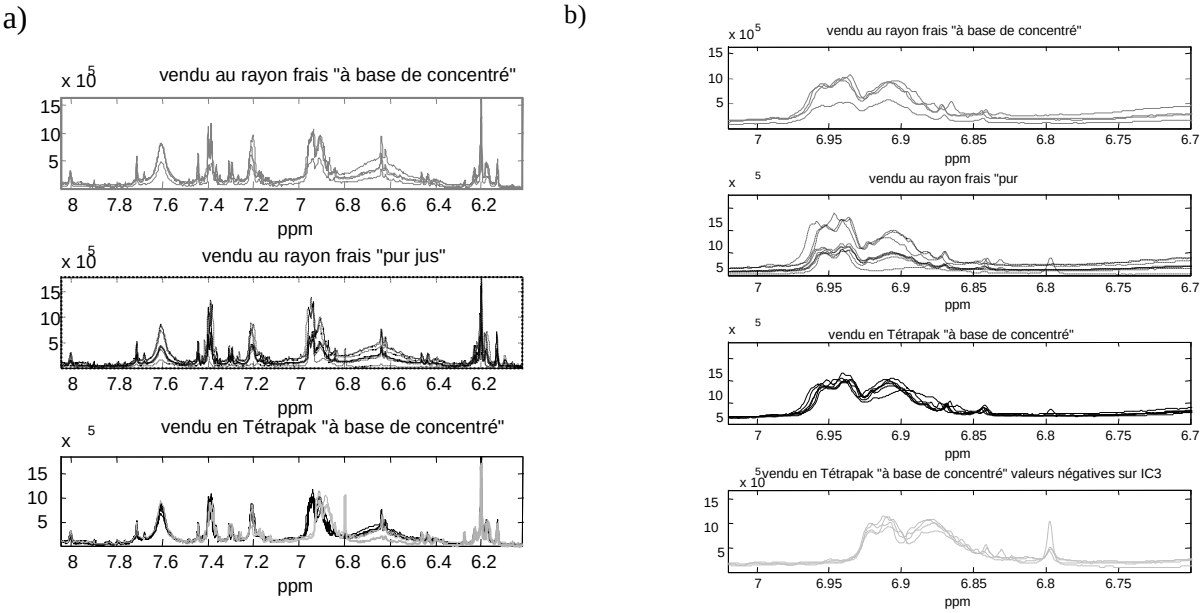


Figure 20 : Séparation des échantillons de pamplemousse en fonction de leur mode de conservation (longue ou courte) et de leur process (à base de concentré ou « pur jus ») : a) sur toute la zone observée ; b) dans la zone de 6,70 à 7,02 ppm

Ensuite nous avons comparé le taux de classement en utilisant les coordonnées des échantillons dans les deux nouveaux espaces de composantes. Nous allons tester la qualité des jeux de données pré-traitées par modélisation et classement "leave-one-out". Chacun des 92 échantillons va être à son tour laissé à part et projeté sur le modèle ACI calculé avec les 91 échantillons restant. L'échantillon laissé à part va être classé en fonction de sa distance des barycentres de chacun des 3 groupes prédéfinis, calculée sur la base des distances de Mahalanobis ou Euclidienne. Les résultats sont présentés dans le tableau 2 ci-dessous.

Tableau 2. Résultats de classements basés sur les coordonnées factorielles de l'ACI et de l'ACP

Traitement	Distance	Meilleur classement des 92 échantillons avec une seule composante	Meilleur classement des 92 échantillons dans les 6 espaces possibles	Composantes utilisées pour le classement
ACP	Mahalanobis	23 ; PC 1, 2 et 3	23	Tous les espaces à 2, 3 et 4 dimensions
	Euclidienne	68 ; PC 3	82	PC 1 à 3
ACI	Mahalanobis	71 ; IC 2	85	IC 2 et 3
	Euclidienne	83 ; IC 3	86	IC 2 et 3

Dans tous les cas, l'ACP donne un moins bon classement que l'ACI que ce soit en distance Euclidienne ou Mahalanobis, cette différence étant plus significative pour la distance de Mahalanobis.

Nous avons donc décidé de ne prendre en compte dans la suite de ce document de synthèse, que les résultats de la décomposition en composantes indépendantes.

3.4. Résultats sur les prétraitements

Dans les différents cas étudiés, la première opération effectuée sur les spectres était de les traiter de manière à rendre les jeux de données homogènes.

En effet, les spectres obtenus présentent plusieurs différences qui perturbent les analyses chimiométriques. Tout d'abord, deux étapes d'homogénéisation ont été réalisées : le phasage des spectres et la correction de la ligne de base qui dépendent de l'acquisition. Ensuite, afin de recaler les pics des spectres et d'avoir des déplacements chimiques homogènes malgré les faibles variations de composition des échantillons et en particulier du pH, une fonction d'ourdissage utilisant une interpolation linéaire (COW) a permis de recaler les pics.

Pour illustrer cette partie, nous avons choisi de travailler dans la zone des signaux des protons des groupements CH_2 de l'acide citrique et de l'acide malique (Figure 21 a). Ces deux composés sont en quantités variables dans les deux types de jus. Bien que le pH ait été homogénéisé, il existe encore des variations de déplacements chimiques. Nous allons voir en quoi le « warping » des données va permettre d'homogénéiser les signaux (Figure 21 b) et l'influence sur les composantes ACI (Figure 22 et 23).

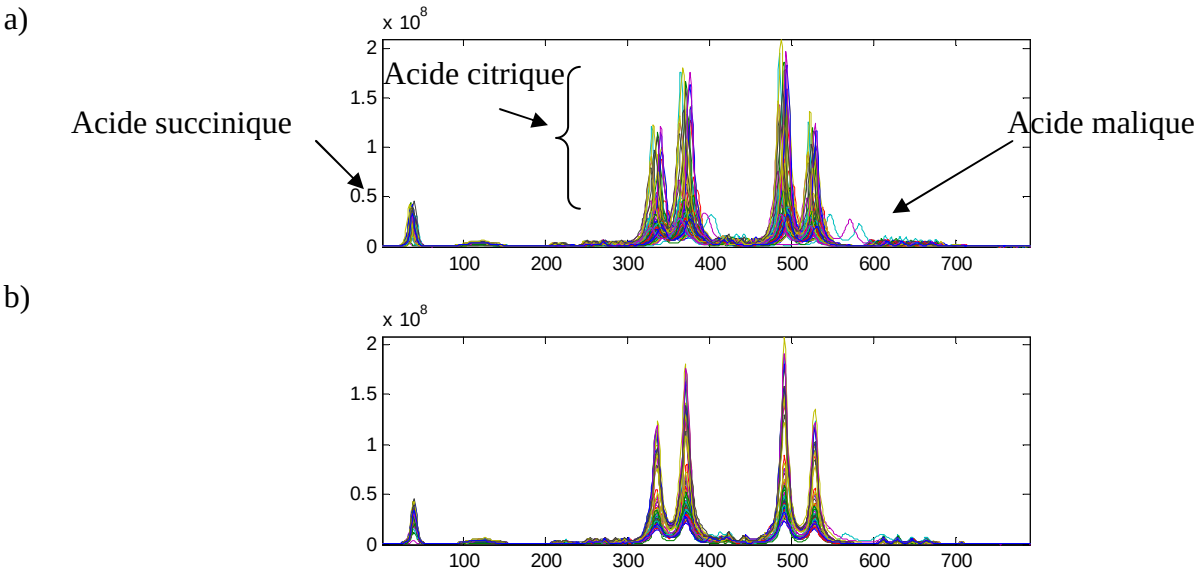


Figure 21 : Zone spectrale de l'acide citrique et de l'acide malique : a) avant warping ; b) après warping

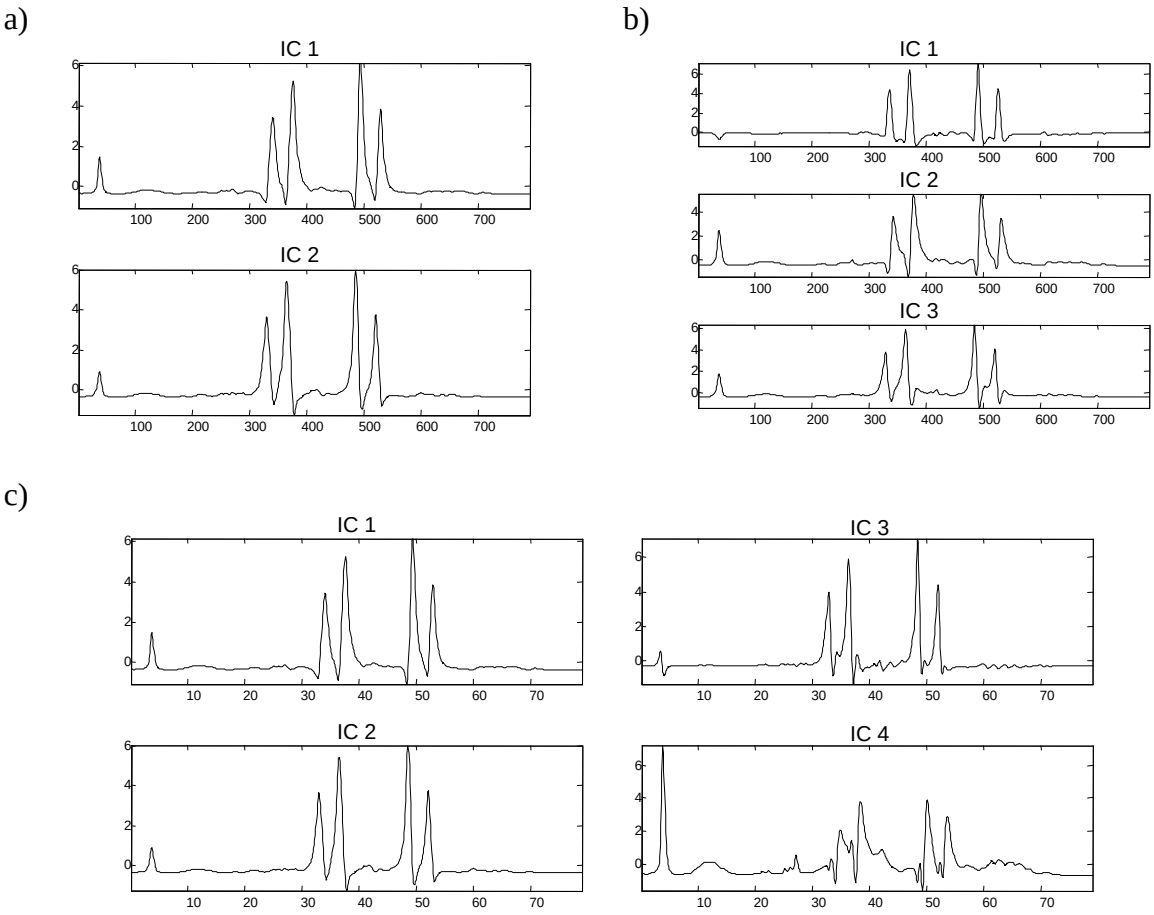
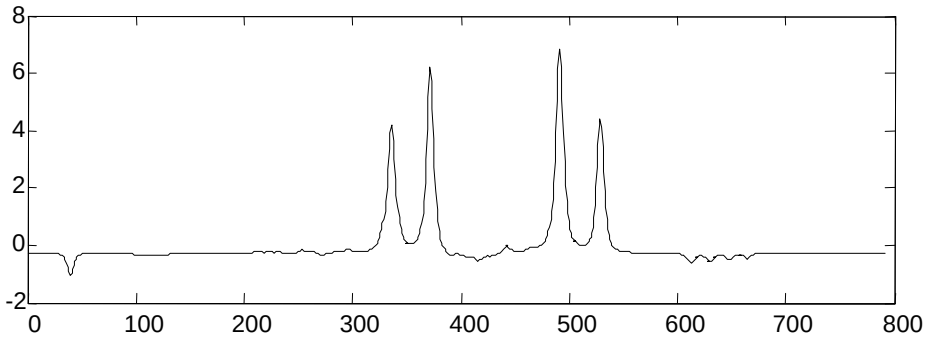


Figure 22 : Décomposition des signaux RMN sans ourdissage en : a) 2 composantes indépendantes ; b) 3 composantes indépendantes ; c) 4 composantes indépendantes

La Figure ci-dessus montre que sans warping des données et quel que soit le nombre de composantes indépendantes choisies, les composantes ressemblent aux signaux de l'acide citrique avec une déformation comme une première dérivée. Cet effet est caractéristique d'un décalage de pic.

Après warping des données, la décomposition en deux composantes indépendantes (Figure 23) fait apparaître sur la composante 1, le signal de l'acide citrique bien défini avec des signaux de l'acide malique et de l'acide succinique très faible et de signe opposé. La composante 2 ressemble à ces deux signaux avec un faible signal de signe opposé pour l'acide citrique. Ces composantes sont beaucoup plus proches des signaux réels et bien moins déformées que dans le cas où il n'y a pas de warping.

a)



b)

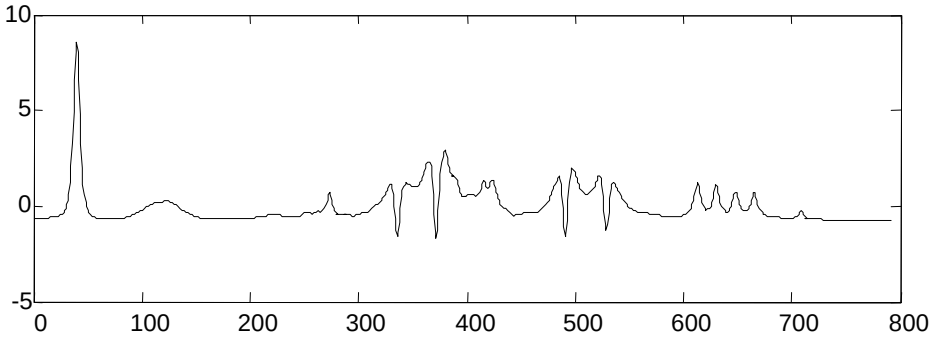


Figure 23 : Décomposition des signaux RMN avec ourdissage en 2 composantes indépendantes : a) IC1 ; b) IC2

3.4.1. Effets des méthodes de modification de l'intensité

Enfin l'intensité des zones spectrales varie en fonction de la composition de l'échantillon. En effet, l'intensité d'un pic RMN correspondant à un type de proton d'une molécule, qui a eu le temps de relaxer complètement, est proportionnelle à la fois à la concentration de la molécule

dans le tube de mesure mais aussi au nombre de protons du même type sur le site de la molécule.

En général, dans les produits que nous avons étudiés, les intensités les plus importantes se retrouvaient dans la zone « des sucres » entre 3,5 et 5,5 ppm. La différence d'intensité était de l'ordre de 10^3, ce qui pourrait introduire un biais dans certaines méthodes chimiométriques comme l'ACP. Afin d'équilibrer les différentes zones, nous avons choisi de transformer notre jeu de données grâce à une transformation logarithmique. En effet, cela permet d'obtenir une variabilité des intensités du même ordre de grandeur dans toutes les régions du spectre, sans trop favoriser les variations aléatoires de la ligne de base, ce qui aurait été le cas si l'on avait centré et réduit les variables comme cela se fait souvent avec d'autres types de transformation. En effet, la transformation centrée-réduite par colonne donne autant de poids à des variables dans la ligne de base qu'à des variables dans les pics, ce qui modifie beaucoup la forme des signaux (Figure 24 b). Avec une transformation logarithmique (Figure 24 c) la forme des signaux est conservée et le rapport des intensités dans les différentes régions est considérablement amélioré.

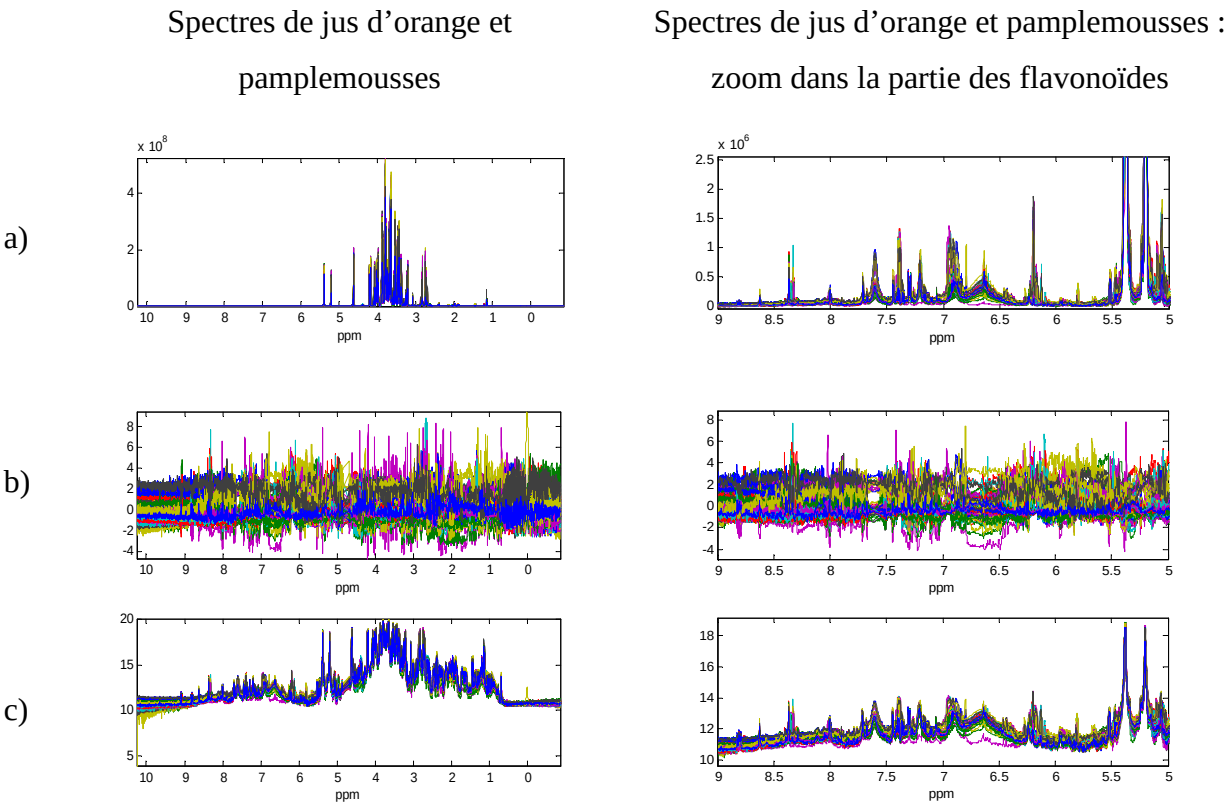


Figure 24 : Données spectrales : a) sans modification ; b) centrées réduites en colonne ; c) avec transformation logarithmique.

Nous allons tester par classement "leave-one-out", la qualité des jeux de données pré-traités. Nous avons trois groupes : OJ, OG, GJ. Chacun des 92 échantillons va être à son tour laissé à part et projeté sur le modèle ACI créé avec les 91 échantillons restant. Par rapport à la proximité aux barycentres de chaque groupe, calculée sur la base des distances de Mahalanobis, l'échantillon laissé à part va être reclassé.

Le tableau, ci-dessous, présente les résultats de classification obtenus après avoir optimisé à 5 le nombre total de composantes ACI. Les composantes utilisées pour faire le classement varient en fonction des prétraitements choisis. En effet, pour chaque modèle, il y a 15 espaces de composantes consécutives possibles (IC 1 seule, IC 2 seule, …, ICs 2 et 3 …).

Tableau 3. Résultats de classements basés sur les coordonnées factorielles de l'ACI

Prétraitements	**Distance**	**Meilleur classement des 92 échantillons avec une seule composante**	**Meilleur classement des 92 échantillons dans les 15 espaces possibles**	**Composantes utilisées pour la classification**
Aucun	Mahalanobis	63 ; IC 4	75	ICs 1 à 4
	Euclidienne	65 ; IC 4	77	ICs 1 à 4
Centrage et	Mahalanobis	71 ; IC 2	79	ICs 1 à 3
réduction en colonne	Euclidienne	49 ; IC 2	64	ICs 1 à 4
Transformation	Mahalanobis	85 ; IC 3	85	ICs 1 à 3
logarithmique	Euclidienne	85 ; IC 3	88	ICs 2 et 3

La référence utilisée est le nombre d'échantillons bien classés sur les 92 échantillons en utilisant le jeu de données traité par ourdissage. Le classement sur une composante donne une approximation de la qualité des sources pures. Sans aucun pré-traitement sauf l'ourdissage, 63 des 92 échantillons sont bien classés sur la composante 4 (cf. Tableau 3).

Les prétraitements par centrage-réduction en colonne des données n'améliorent que peu les prédictions sur une seule composante, en distance de Mahalanobis (65 échantillons sur 92) et donne des résultats bien inférieurs en distance Euclidienne (49 sur 92).

Le meilleur taux de classement sur une composante est obtenu pour la transformation logarithmique avec 85 échantillons bien classés sur 92 (96 %), ce qui indique une source pure discriminante de bonne qualité.

Ainsi pour des questions de représentativité des variables et d'amélioration de la qualité de l'information extraite, nous avons choisi de travailler sur les spectres après transformation logarithmique.

Pour illustrer les résultats ci-dessus, les contributions factorielles et les coordonnées factorielles obtenues suite à la décomposition des signaux avant et après transformation logarithmique sont présentées en Figure 25.

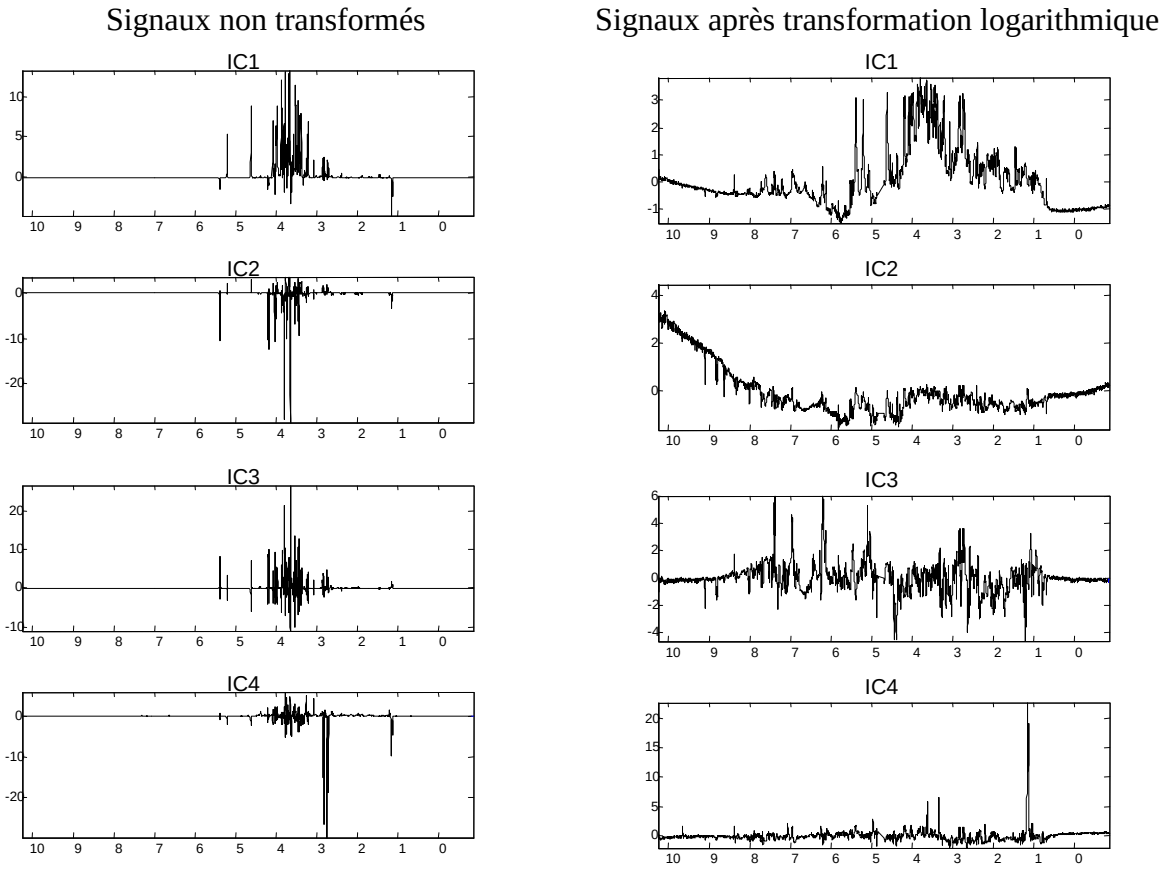
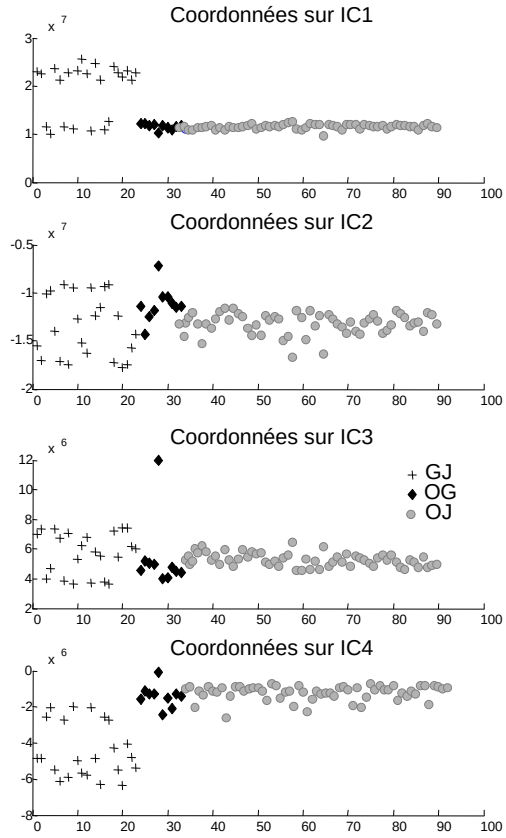


Figure 25 : Décomposition des signaux en 4 composantes

Signaux non transformés



Signaux après transformation logarithmique

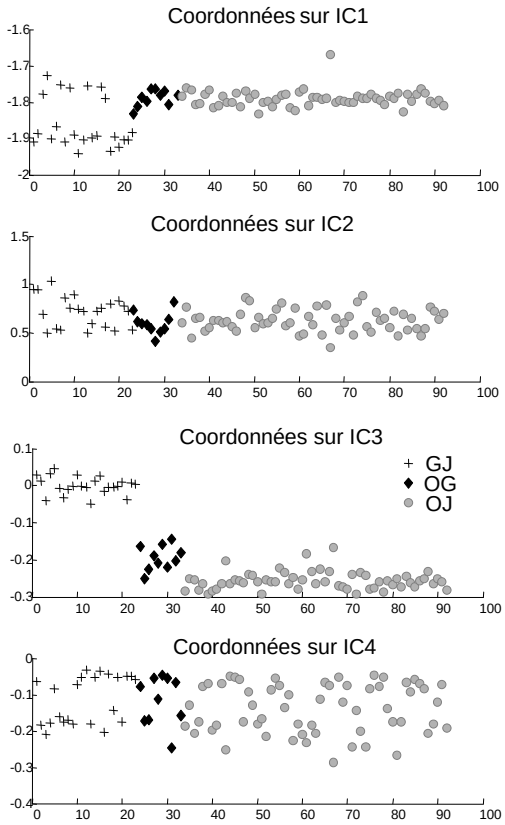


Figure 26 : Projections des échantillons sur chaque composante

Quand les données sont non-transformées, la projection des individus sur les 4 premières composantes principales ne permet pas de distinguer de façon nette les différents groupes (Figure 26). Seule la composante 4 distingue une partie des échantillons de jus de pamplemousse (en noir) des autres. Sur cette seule composante 65 échantillons peuvent être discriminés comme nous l'avons vu dans le paragraphe précédent. Le signal qui a le plus d'importance sur cette composante est le signal de l'acide citrique (2,78 ppm). Ce composant n'est pas un marqueur de la différence entre les jus, bien que la concentration d'acide citrique diffère selon le type de jus de 6,3-17 g/l pour le jus d'orange et de 8-20 g/l pour le jus de pamplemousse selon la FDA.

En transformée logarithmique, la composante 3 - qui utilisée seule permet de classer 85 individus - présente les signaux caractéristiques du pamplemousse dont les pics de la naringine, vus dans la partie précédente (Figure 25). En effet, la projection des individus sur

cette composante (Figure 26) permet de bien séparer le groupe orange qui ne contient pas de naringine (en bleu) et le groupe pamplemousse (en noir), les mélanges (en rouge) ayant bien une position intermédiaire.

3.4.2. Effet de la réduction de la taille des données par moyennage de points

Pour tester les effets des prétraitements tels que la moyenne de points et la transformation logarithmique, nous proposons de soumettre les jeux de données ayant subis différents prétraitements :

- Pas de prétraitement,
- Moyenne de points,
- Transformation logarithmique,
- Transformation logarithmique et moyenne de points,

à un test de discrimination basé sur les composantes indépendantes. Ce test ayant été partiellement réalisé sur les yaourts à la fraise et décrit dans l'article correspondant, nous proposons de l'appliquer de façon complète sur les données des jus de fruits. Nous utilisons les mêmes tests que précédemment.

Tableau 4. Résultats de classements basés sur les coordonnées factorielles de l'ACI

Prétraitements	Distance	Meilleur classement des 92 échantillons avec une seule composante	Meilleur classement des 92 échantillons dans les 15 espaces possibles	Composantes utilisées pour la classification
Aucun	Mahalanobis	63 ; IC 4	75	ICs 1 à 4
	Euclidienne	65 ; IC 4	77	ICs 1 à 4
Moyenne de 7 points	Mahalanobis	69 ; IC 4	73	ICs 1 et 2, ICs 1 à 4
	Euclidienne	67 ; IC 4	78	ICs 1 à 4 et ICs 3 et 4
Transformation logarithmique	Mahalanobis	85 ; IC 3	85	ICs 1 à 3
	Euclidienne	85 ; IC 3	88	ICs 2 et 3
Transformation logarithmique et moyenne de 7 points	Mahalanobis	83 ; IC 3	85	ICs 2 et 3
	Euclidienne	88 ; IC 3	88	IC 3

Sans pré-traitement, seuls 63 des 92 échantillons sont bien classés sur la composante 4. Sur cette même composante, l'effet de la moyenne de point sur les données n'est pas important, seuls quelques échantillons supplémentaires sont bien classés (Tableau 4). Au contraire, l'effet de la transformation logarithmique est important comme nous l'avons vu dans la partie

précédente et ce taux de classement n'est que très peu diminué avec en plus la moyenne de 7 points.

La Figure 27 montre que les composantes factorielles et les composantes indépendantes avec moyenne des données sont similaires à celles sans moyenne des données (Figure 25).

a) b)

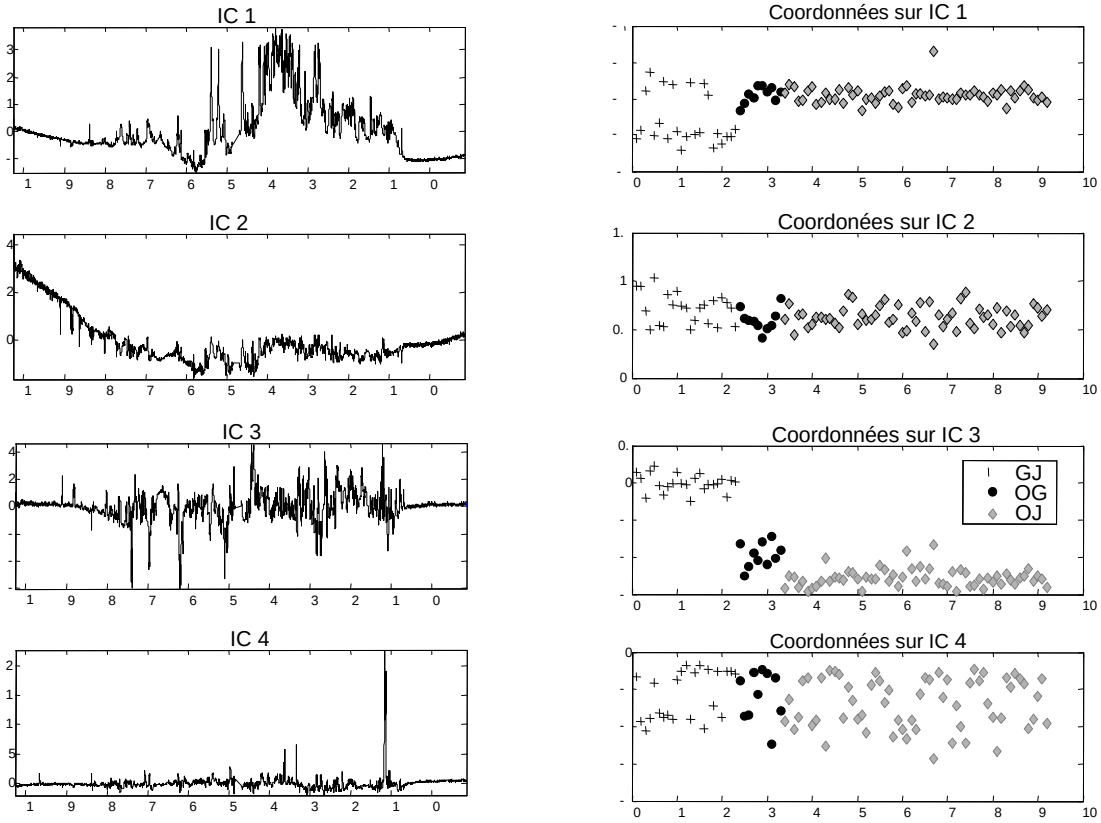


Figure 27 : Modèle ICA 4 composantes indépendantes sur les données après transformation logarithmique et avec moyenne de 7 points : a) Composantes ; b) Coordonnées des individus

En conclusion, le prétraitement des données a un effet non-négligeable sur les résultats d'analyse. Pour les données spectrales RMN ^{1}H, présentant des variations importantes entre les différentes zones, nous avons constaté qu'un traitement logarithmique améliorait de beaucoup la discrimination avec un taux de classement variant de 65 % sans prétraitement à 96 % avec une transformation logarithmique. Ceci confirme les travaux de plusieurs équipes [15-17] qui

ont constaté l'amélioration du traitement des données en utilisant la transformation logarithmique.

La moyenne sur plusieurs points n'a que très peu d'effet sur les résultats mais nous l'utiliserons pour diminuer de façon significative la taille de la matrice de données passant de 92 x 15 001 à 92 x 2 143 et les temps d'analyse des données selon les méthodes utilisées.

3.5. Résultats sur la sélection de variables

Le spectre apportant de nombreuses informations relatives à la composition de l'échantillon, toute l'information n'est pas directement liée au repérage des anomalies liées à une fraude potentielle. Afin de localiser les zones spectrales porteuses de ce type d'information, plusieurs méthodes de sélection de variables ont été testées comme exposé dans la partie *Méthodologie*.

3.5.1. Critères basés sur les variables

3.5.1.1. Variance

La variance apporte de l'information quant à l'amplitude de variation d'une variable. Nous avons d'abord pris comme hypothèse très simplifiée qu'une forte variation au niveau d'une variable était due à une différence entre les groupes prédéfinis d'échantillons.

Connaissant la répartition des individus dans les groupes prédéfinis, il est facile de déterminer un seuil de variance pour sélectionner les variables qui maximisent le taux de bons classements.

Le jeu de données des spectres de jus de fruits après transformation logarithmique et moyennage de 7 points comporte 2143 variables. Une fois la variance de chaque variable calculée (Figure 28 b), un seuil a été déterminé en maximisant le nombre d'échantillons bien classés dans l'espace des coordonnées factorielles de l'ACI par validation croisée « leave-one-out », comme expliqué dans la partie *Méthodologie*. Les résultats sont représentés graphiquement sur la Figure 28 a.

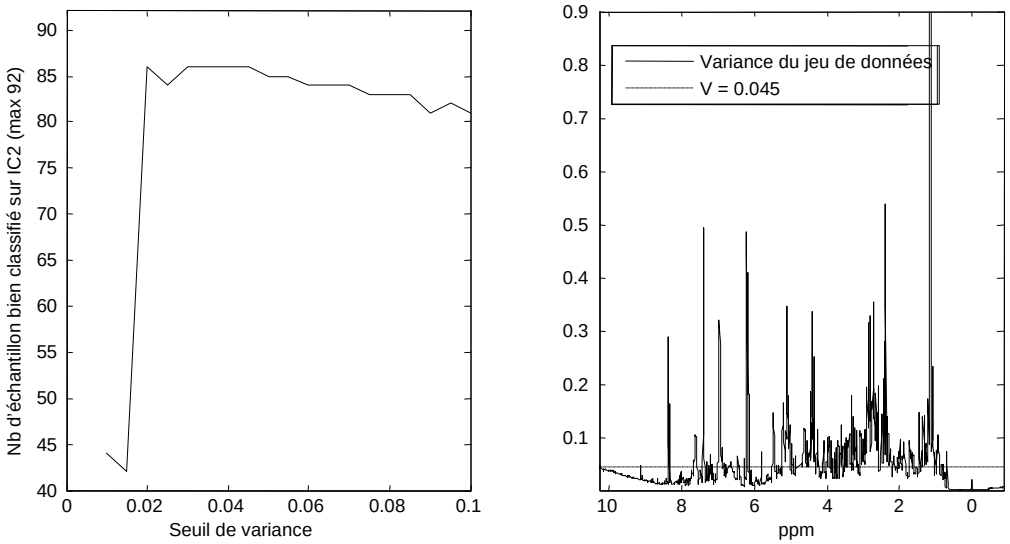


Figure 28 : a) Nombre d'échantillons bien classés en fonction du niveau de variance choisi ; b) Variance du jeu de données et seuil retenu.

Puisque que le but de cette sélection de variables est aussi de réduire au maximum la taille de notre jeu de données, à valeurs de bons classements égales le seuil le plus haut sera favorisé.
Dans cet exemple, le plus grand nombre d'échantillons bien classés a été obtenu sur la seconde composante indépendante pour des seuils de variance de 0,020 puis entre 0,030 et 0,045. Le seuil retenu a donc été 0,045.

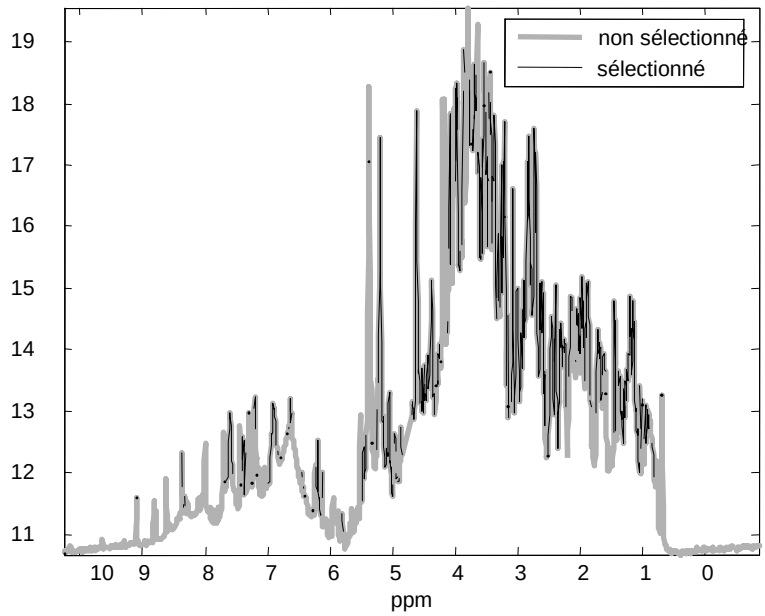


Figure 29 : Zones spectrales sélectionnées ou non par la méthode de sélection sur la variance

Les variables sélectionnées (Figure 29) sont alors évaluées selon les résultats de la classification des individus par ICA comme dans les paragraphes précédents. Les résultats Figurent dans le tableau 5.

Tableau 5. Résultats de classements basés sur les coordonnées factorielles de l'ACI

Prétraitements	Distance	Meilleur classement des 92 échantillons sur une seule composante	Meilleur classement des 92 échantillons dans les 15 espaces possibles	Composantes utilisées pour la classification
Transformation logarithmique et moyenne de 7 points	Mahalanobis	83 ; IC 3	85	ICs 2 et 3
	Euclidienne	88 ; IC 3	88	IC 3
Sélection de variables grâce à la méthode de la variance	Mahalanobis	86 ; IC 2	88	ICs 2 et 3
	Euclidienne	88 ; IC 2	88	ICs 1 et 2 et ICs 1 à 4

Cette sélection d'un tiers des variables (759 sur les 2143 de départ) permet d'obtenir des résultats aussi bons que sur le jeu entier. Les parties informatives ont donc été conservées et une partie du bruit supprimée.

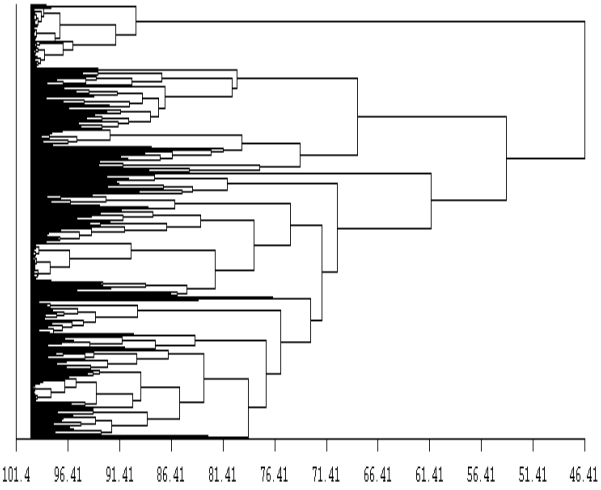
3.5.1.2.CLV

La technique CLV exposée dans la partie *Méthodologie* a été appliquée à ce même jeu de données sur les jus de fruits.

Le critère de la classification hiérarchique est la covariance utilisée au carré ce qui permet de prendre en considération les corrélations négatives entre les variables. A chaque étape le paramètre T est maximisé. Les 2143 variables de notre jeu de données sont, petit à petit, regroupées par paire de classe dont la covariance est la plus grande. Ceci jusqu'à ce que toutes les classes soient regroupées, cf. la Figure 30 a.

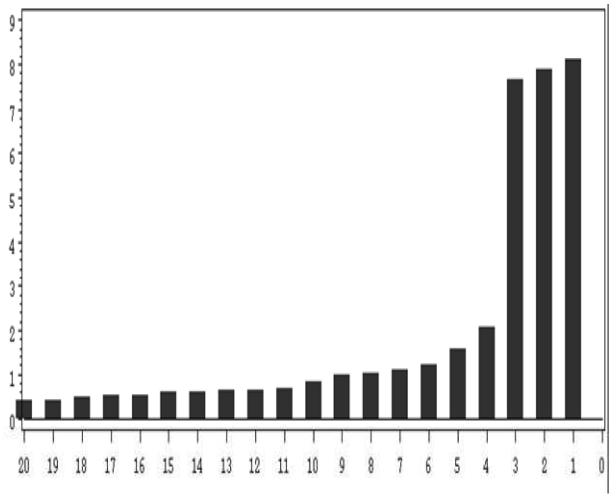
Le but de cette classification hiérarchique est de déterminer à quel moment le regroupement de variables homogènes n'est plus intéressant, c'est-à-dire de déterminer le nombre de groupe le plus représentatif de l'ensemble des données.

Dendrogramme avant consolidation



Part de la variance contenue dans les composantes

Evolution du critère ΔT



Nombre de groupes

Figure 30 : a) Nombre d'échantillon bien classé en fonction du seuil de variance choisi ; b) Variance du jeu de données et seuil retenu.

Le suivi de l'évolution du critère T est déterminant. En effet, si la variation de T est importante lorsque le nombre de groupe diminue cela signifie qu'une part importante de l'information contenue dans la partition inférieure a été perdue. Il faut donc s'arrêter au regroupement juste avant afin de préserver l'information contenue dans les classes.

Dans cet exemple l'évolution du critère T suivi sur la Figure 30b met en évidence deux niveaux de perte d'information - lors du passage de 6 à 5 classes et lors du passage de 4 à 3 classes. Il faut donc envisager de choisir de faire 6 ou 4 groupes de variables.

Chaque groupe de variables est résumé par la première composante principale calculée sur la matrice de l'ensemble des individus et les variables du groupe.

L'hypothèse de départ est que les groupements de variables peuvent être influencés par l'existence de classes d'échantillons.

Grâce à une ANOVA appliquée à la première composante principale de chaque groupe de variables, il est possible de déterminer lesquels sont intéressants pour la discrimination des classes d'individus. Les résultats de la partition en 4 groupes (meilleurs que la partition en 6 groupes) sont présentés dans le Tableau 6.

Tableau 6. Résultats de l'ANOVA sur les premières composantes principales des 4 groupes de variables déterminés par CLV

Source	Sum of Squares	d.f.	Mean Squares	F	Prob>F
Composante 1	0.019	2	0.009	0.85	0.429
Composante 2	0.669	2	0.335	90.02	0
Composante 3	0.010	2	0.005	0.44	0.645
Composante 4	0.014	2	0.007	0.61	0.544

Seule la composante 2 discrimine significativement les classes d'individus. Ainsi les 878 variables qui constituent ce groupe sont retenues (Figure 31).

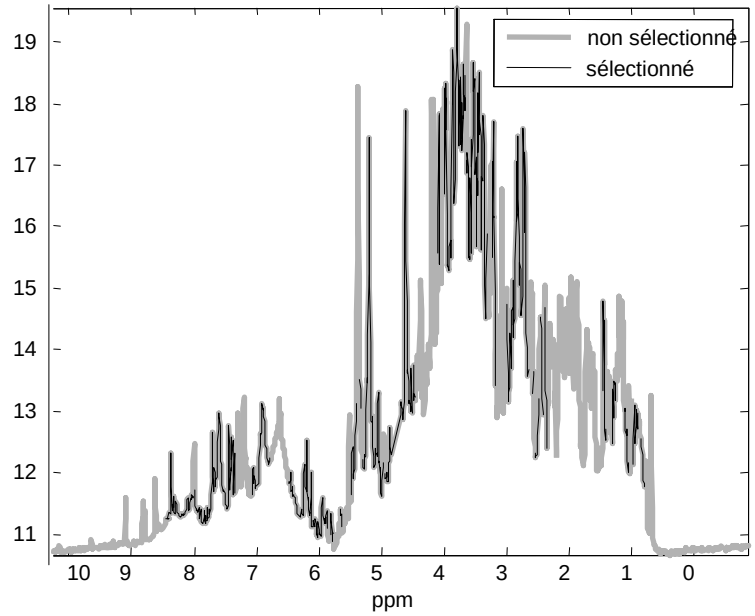


Figure 31 : Variables sélectionnées par CLV

Tableau 7. Résultats de classements basés sur les coordonnées factorielles de l'ACI

Prétraitements	**Distance**	**Meilleur classement des 92 échantillons sur une seule composante**	**Meilleur classement des 92 échantillons dans les 15 espaces possibles**	**Composantes utilisées pour la classification**
Transformation logarithmique et moyenne de 7 points	Mahalanobis	83 ; IC 3	85	ICs 2 et 3
	Euclidienne	88 ; IC 3	88	IC 3
Sélection de variables grâce à CLV	Mahalanobis	83 ; IC 3	88	ICs 2 et 3, ICs 1 à 4
	Euclidienne	86 ; IC 3	89	ICs 1 à 4

Le fait de sélectionner les variables appartenant au deuxième groupe de variables reconnues par CLV permet de faire passer les résultats du classement des individus de 85 à 89, par rapport au jeu de variables complet (Tableau 7). En comparant les tableaux 4 et 6, on peut voir que les résultats avec une seule IC ne sont pas aussi bons que ceux obtenus suite à la sélection de variables basée simplement sur la variance totale, bien que le nombre de variables soit comparable (878 et 759). Avec plusieurs ICs, les résultats sont améliorés d'un échantillon avec la méthode CLV.

3.5.2. Fonctions utilisant des intervalles

Dans un spectre, des variables contiguës sont le plus souvent associées à la même information spectrale. Pour cette raison, il peut être préférable de sélectionner des intervalles de variables ayant un pouvoir prédictif plutôt que des variables individuelles. Nous avons testé trois méthodes de sélection d'intervalles : iPLS [18-20] et Evolving Window Zone Selection (EWZS) [21,22] et Interval-PLS_Cluster [23,24].

3.5.2.1.iPLS

Sur le jeu de données ayant subi comme dans les parties précédentes une transformation logarithmique et une moyenne de point, nous avons choisi différents paramètres disponibles pour l'algorithme Matlab iPLS. Tout d'abord, le nombre maximum de variables latentes a été fixé à 8 et parmi les pré-traitements des données proposés nous avons choisi «la moyenne ». Le spectre est découpé en 31 intervalles pour avoir une taille d'intervalle de 70 variables par souci de comparaison avec l'algorithme Interval-PLS_Cluster.

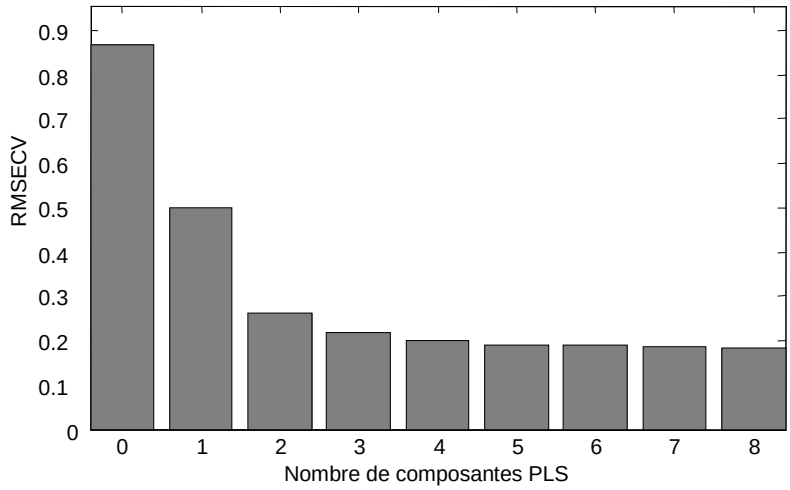


Figure 32 : RMSECV en fonction du nombre de composantes PLS prises en compte pour le modèle global

Le choix du nombre de variable latente pour le modèle complet se fait par validation croisée « leave-one-out ». Le graphique (Figure 32) présentant le RMSECV en fonction du nombre de variables latentes permet de choisir un modèle à 4 variables latentes qui a une valeur proche du minimum, sans risquer un sur-ajustement.

La Figure suivante (Figure 33), permet de choisir parmi les 31 intervalles ceux qui donnent de meilleurs résultats que ceux obtenus sur le jeu de données complet avec un modèle de prédiction à 4 variables latentes. Pour chaque intervalle les valeurs de RMSE sont calculées en prenant en compte de 1 à 8 variables latentes. Pour tout le spectre, noté REF sur la Figure, la valeur de RMSE retenue est celle obtenue avec le modèle utilisant 4 composantes soit 0,2009. Cette valeur est choisie comme référence pour tous les spectres. Dans ce cas, trois intervalles sont en dessous de ce seuil, les intervalles 9, 12 et 15. Ainsi, trois éléments spectraux non contigus pris séparément discriminent mieux les différents groupes de jus. Nous allons aussi tester le pouvoir de discrimination des trois intervalles pris ensembles.

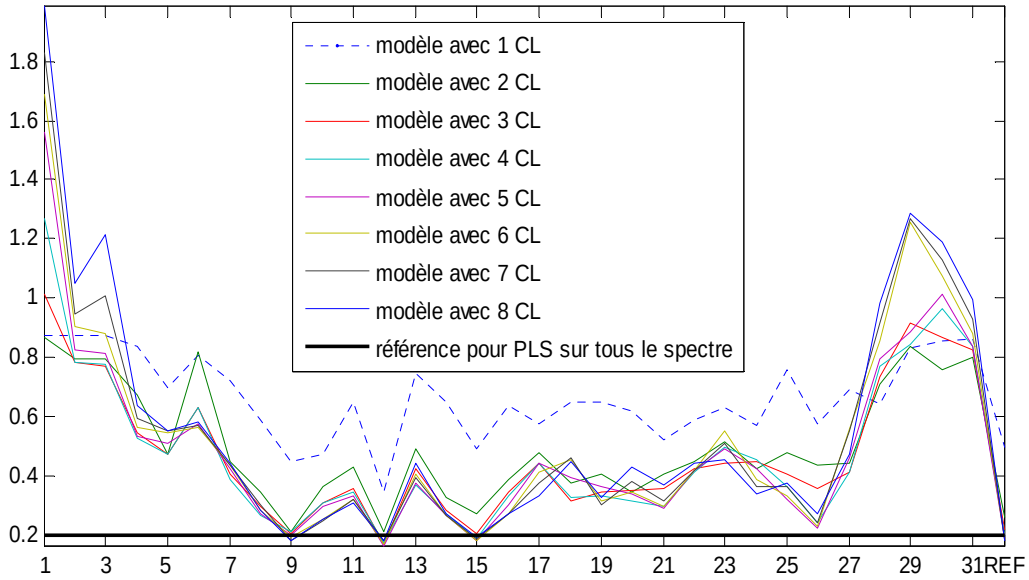


Figure 33 : Valeurs de RMSE obtenu avec un nombre de variables latentes allant de 1 à 8, sur les différents intervalles définis par iPLS.

Le tableau ci-dessous (Tableau 8) présente la correspondance entre les intervalles et les variables sélectionnées ainsi que les valeurs minimales de RMSE obtenues.

Tableau 8. ***Correspondance entre les intervalles et les variables et leurs valeurs de RMSE***

Numéro de l'intervalle	Début de l'intervalle	Fin de l'intervalle	Meilleur RMSE
1	1	70	0.8671
2	71	140	0.7837
3	141	210	0.7733
4	211	280	0.5289
5	281	349	0.4712
6	350	418	0.5660
7	419	487	0.3910
8	488	556	0.2671
9	**557**	**625**	**0.1844**
10	626	694	0.2481
11	695	763	0.3097
12	**764**	**832**	**0.1669**
13	833	901	0.3709
14	902	970	0.2667
15	**971**	**1039**	**0.1824**
16	1040	1108	0.2749
17	1109	1177	0.3327
18	1178	1246	0.3184
19	1247	1315	0.3026
20	1316	1384	0.3177
21	1385	1453	0.2914
22	1454	1522	0.4124
23	1523	1591	0.4406
24	1592	1660	0.3416
25	1661	1729	0.3247
26	1730	1798	0.2253
27	1799	1867	0.4128
28	1868	1936	0.6427
29	1937	2005	0.8308
30	2006	2074	0.7574
31	2075	2143	0.8037
32	**1**	**2143**	**0.2009 (4 LVs)**

Seuls trois intervalles donnent de meilleurs résultats que PLS sur l'ensemble des données (point numéro REF sur la Figure 33). Il s'agit des intervalles 9, 12 et 15 qui représentent

respectivement les variables de 557 à 625, de 764 à 832, et de 971 à 1039 (en gras dans le Tableau 8). L'intervalle 26 semble aussi intéressant et à une valeur de RMSE proche de la référence (cf. Tableau 8) variables n° 1730 à 1798. Les zones spectrales correspondantes sont représentées Figure 34 ci-dessous.

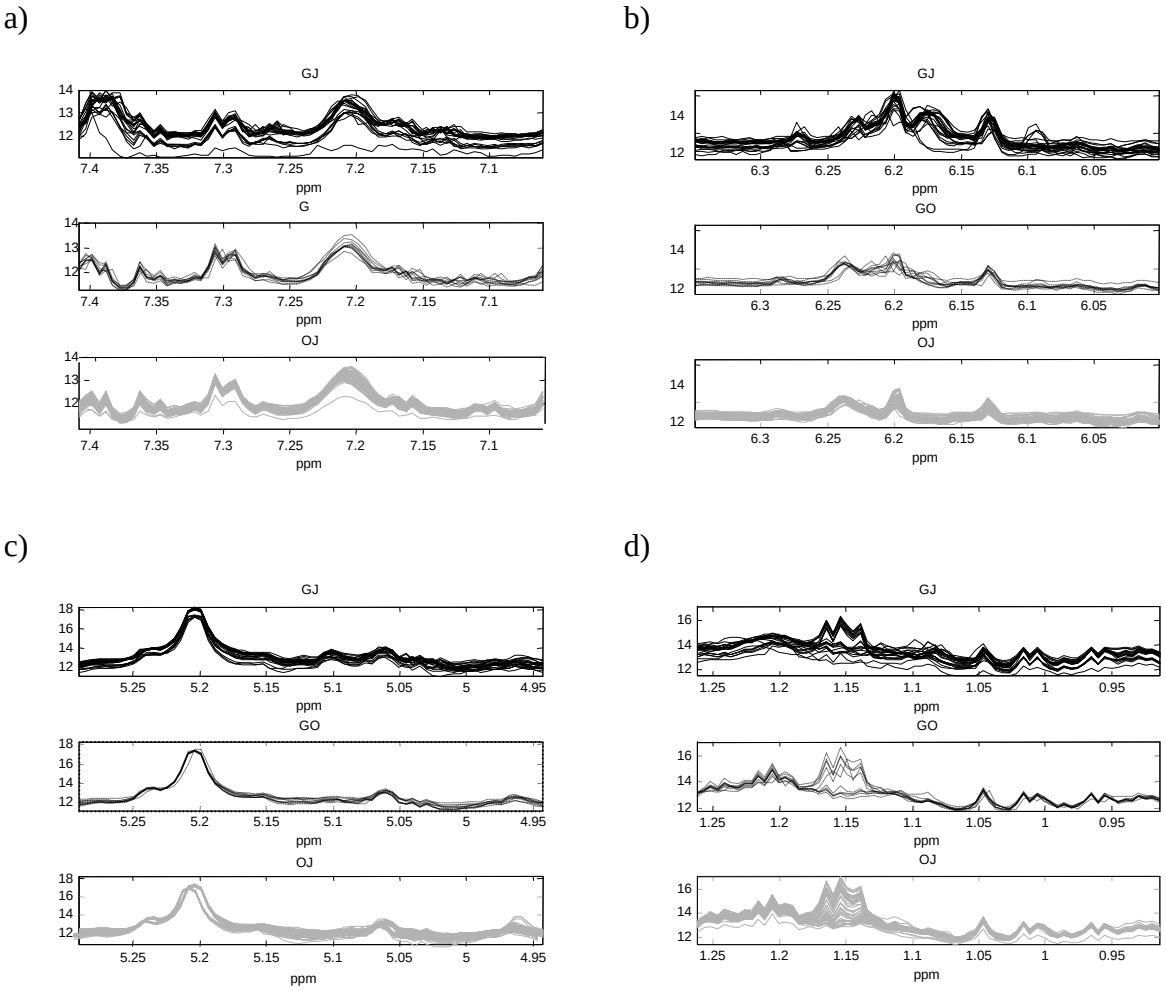


Figure 34 : Intervalles sélectionnés par iPLS : a) N° 9 ; b) N° 12 ; c) N° 15 ; d) N° 26

L’intervalle 9 contient le pic de l’hespéridine à 7,40 ppm et un doublet de la naringine à 7,39 qui sont des marqueurs connus de la distinction entre les jus de type orange et pamplemousse ainsi que des signaux de polyphénols autour de 7,21 ppm. Cependant le pic d’hespéridine est coupé partiellement. Ceci est un inconvénient de cette méthode. La définition de la taille de l’intervalle influe sur la sélection de la zone qui n’est pas en général, optimale.

L’intervalle 12 qui a la plus petite valeur de RMSE, contient à 6,20 le signal de la naringine et à 6,23 celui de l’hespéridine. L’intervalle 15 contient le doublet du glucose à 5,20 ppm et le signal de la naringine à 5,07 ppm (cf. Annexe 7). L’intervalle 26 contient le pic de l’éthanol à

1.21 (cf. Annexe 7), signe de fermentation des jus. Pour évaluer la différence majeure entre les types de jus dans cette zone on calcule la moyenne des groupes orange et pamplemousse (Figure 35).

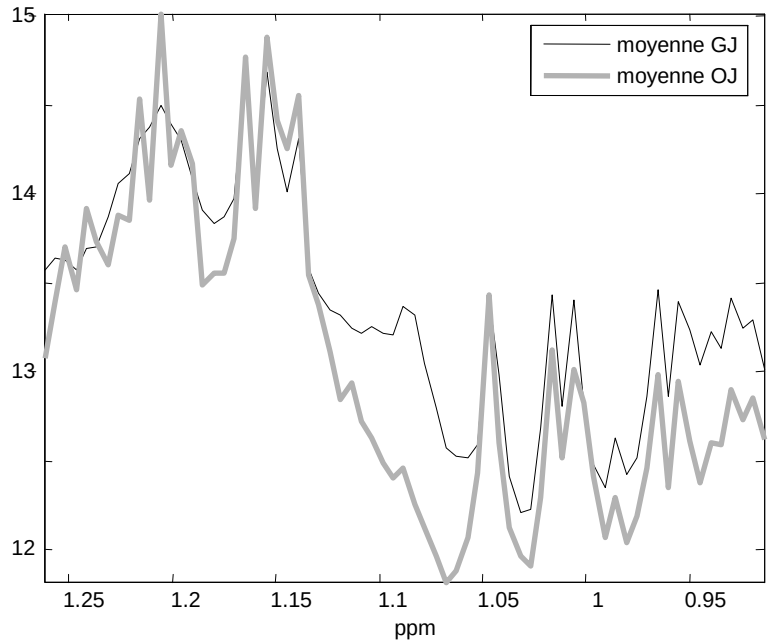


Figure 35 : Moyenne des jus d'orange et pamplemousse dans la zone de l'intervalle 26.

Les signaux qui en moyenne font la différence sont le doublet à 0,93 ppm, appartenant à la leucine, les doublets à 1,01 ppm et 1,09 ppm appartenant à la valine (cf. Annexe 7). Selon les bases de données de l'USDA, la valine est plus abondante dans le jus de pamplemousse que dans le jus d'orange [25]. Pour ce qui est de la leucine, la composition du jus de pamplemousse blanc de Californie, beaucoup utilisé dans les jus de fruits et plus riche que le jus d'orange. Cependant la teneur en leucine dans le jus de pamplemousse varie beaucoup selon l'origine, la variété et sa forme de consommation.

De plus, les résultats du Tableau 9 montrent que la classification est presque parfaite sur les intervalles 9 et 12 (91 sur 92 échantillons). Nous avons choisi de présenter la décomposition des signaux de l'intervalle 12 car à elle seule la troisième composante indépendante obtenue sur cet intervalle, permet de classifier 89 échantillons sur 92 (Figure 36), ainsi que la projection des échantillons dans le plan IC3-IC4 (Figure 38).

Tableau 9. ***Résultats de classements basés sur les zones sélectionnées par iPLS***

Prétraitements	Distance	Meilleur classement des 92 échantillons sur une seule composante	Meilleur classement des 92 échantillons dans les 15 espaces possibles	Composantes utilisées pour la classification
Transformation logarithmique et moyenne de 7 points	Mahalanobis	83 ; IC3	85	IC2 et IC3
	Euclidienne	88 ; IC3	88	IC3
Intervalle 9	Mahalanobis	73 ; IC4	90	IC3 et IC4
	Euclidienne	85; IC3	91	IC3 et IC4
Intervalle 12	Mahalanobis	87 ; IC3	89	IC2 et IC3 ; IC1 et IC3 ; IC3 et IC4
	Euclidienne	89 ; IC3	91	IC2 et IC3; IC1 à IC3
Intervalle 15	Mahalanobis	80 ; IC4	81	IC3 et IC4
	Euclidienne	81 ; IC4	88	IC3 et IC4
Intervalle 26	Mahalanobis	57 ; IC4	86	IC3 et IC4
	Euclidienne	79 ; IC4	89	IC3 et IC4
Intervalles 9, 12 et 15	Mahalanobis	81 ; IC3	84	IC1 à IC4
	Euclidienne	83 ; IC3	91	IC1 à IC4
Intervalles 9, 12 et 15 et 26	Mahalanobis	83 ; IC4	87	IC1 à IC4
	Euclidienne	88 ; IC4	90	IC1 à IC4

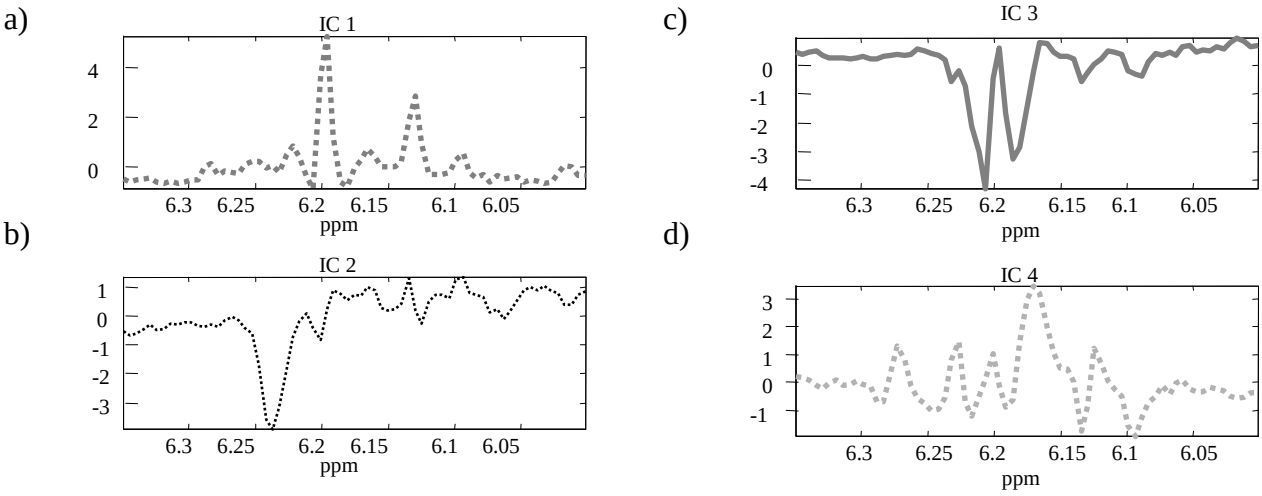


Figure 36 : Composantes indépendantes du modèle ICA 4 composantes sur l'intervalle 12 : a) IC 1 ; b) IC 2 ; c) IC 3 ; d) IC 4

On remarque que les composantes IC 1 et -IC 2 reprennent les signaux qui composent le spectre moyen du jus d'orange. En effet, si on somme ces 2 composantes on obtient un signal proche du spectre moyen du jus d'orange (Figure 37). -IC 3 correspond au doublet de la naringine à 6,20 (cf. Annexe 7 et 8).
Ce sont donc les composantes IC 3 et IC 4 qui séparent les échantillons et en particulier IC 3 qui discrimine 97 % des échantillons.

En effet, la discrimination est très nette sur IC3 (Figure 37), représentant le doublet de la naringine, avec des valeurs proches de 0 (-0,0422 en moyenne) pour les jus d'orange et des valeurs autours de -0,3663 en moyenne pour le pamplemousse.

Sur IC4 (Figure 37), les jus d'orange ont aussi des valeurs proches de 0 (-0,0248), mais les pamplemousses sont plus disséminés avec une moyenne de 0,2374 mais une amplitude de variation de 0,5474.

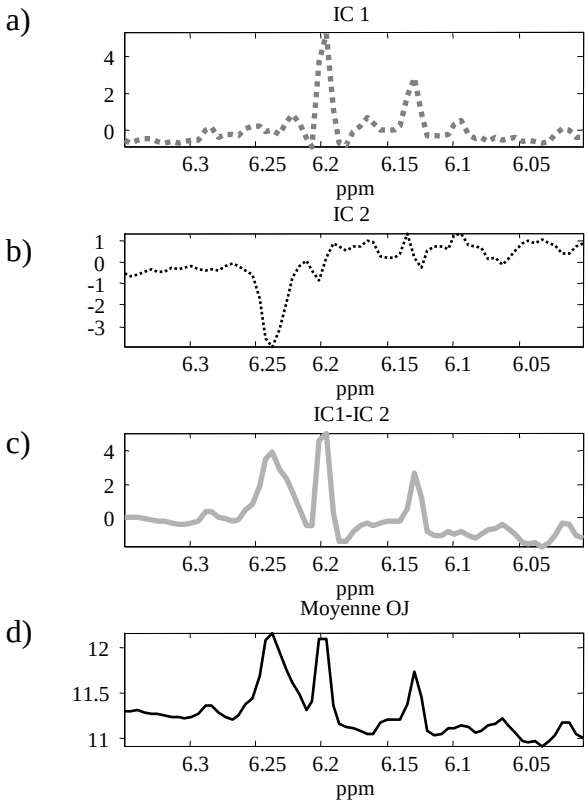


Figure 37 : a) Composantes indépendantes 1 ; b) IC 2 ; c) soustraction de IC2 à IC1 ; d) signal moyen du jus d'orange

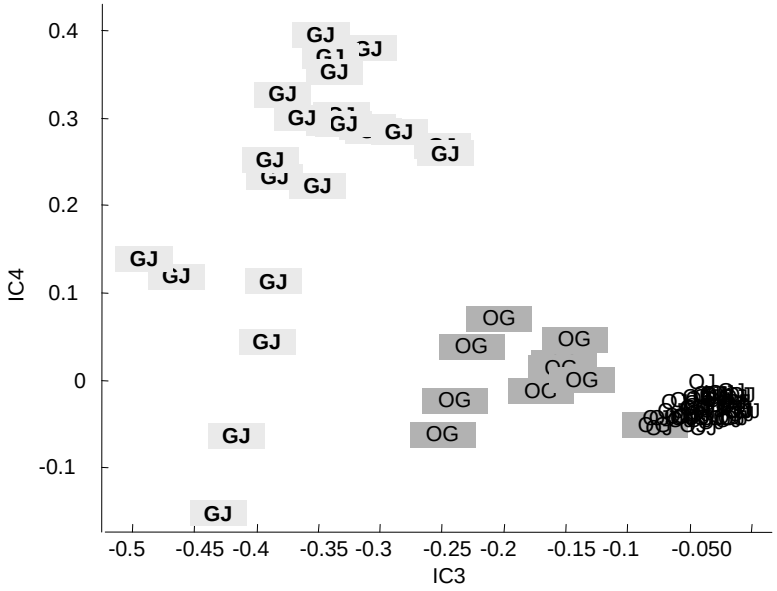


Figure 38 : Projection des individus dans le plan IC3-IC4

Si l'on projette les jus de pamplemousse sur dans le plan IC3-IC4 (Figure 39), on remarque que les jus de pamplemousse « à base de concentré » ont des valeurs proches sur IC4, centrées sur 0,3090 avec une variabilité restreinte (écart-type de 0,0471). Alors que les « purs » jus de pamplemousse ont une valeur sur IC 4 qui varie beaucoup, écart-type de 0,1635 pour une moyenne de 0,1262.

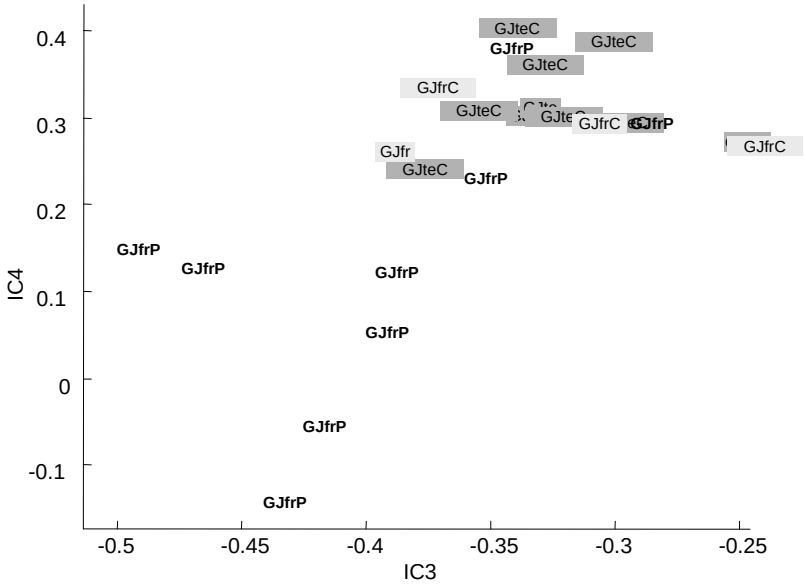


Figure 39 : Projection des jus de pamplemousse dans le plan IC3-IC4

Selon la Figure 40, IC 4 influe majoritairement sur l'intensité du pic à 6,27 ppm ainsi que sur celui à 6,18 ppm. Dans cette zone il peut s'agir d'autres composés avec un cycle aromatique. Cependant, les pics étant superposés dans le signal enregistré, il n'est pas possible d'être plus précis.

a)

b)

c)

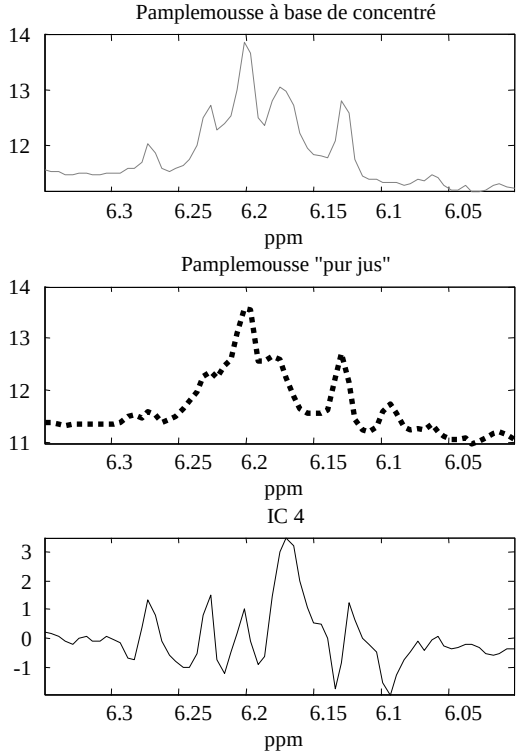


Figure 40 : Zone 6,0 à 6,35 ppm : a) Moyenne du signal des pamplemousses à base de concentré ; b) Moyenne du signal des pamplemousses pur jus

3.5.2.2.EWZS

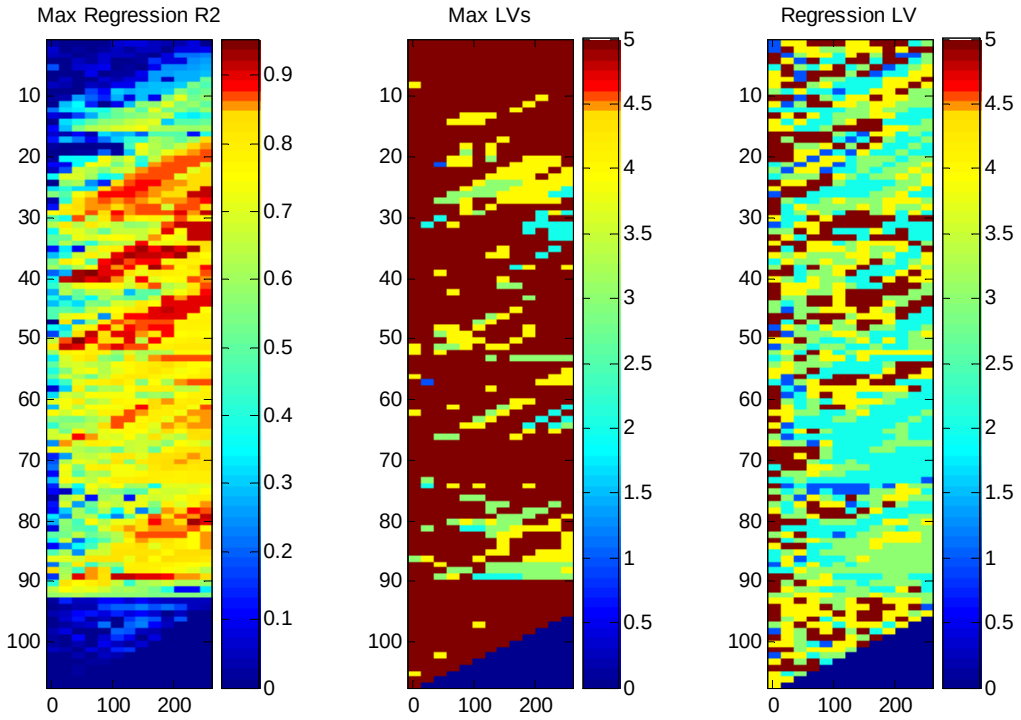
Comme expliqué dans la partie *Méthodologie*, cette méthode permet de faire varier la taille et la position de la fenêtre spectrale testée. Nous avons, dans notre étude, défini une taille minimale de fenêtre de 8 et une taille maximale de 250 ainsi qu'un pas de 20. Nous avons soumis les données de la fenêtre à une réduction de dimensionnalité par ACI avec un nombre de composantes variant de 1 à 5. Ensuite une régression PLS-DA utilisant de 1 à 5 variables latentes a été réalisée sur les différentes fenêtres donnant ainsi un ensemble de valeurs de RMSECV et R^2. Les meilleurs résultats de ces critères en fonction des différentes variations (nombre de composantes indépendantes de l'ACI et variables latentes de la PLS-DA) sont représentés sous forme de cartes de couleur (Figure 41).

La Figure 41 a1 représente la valeur maximale de R^2 pour chaque zone. Elle est obtenue grâce

à une PLS-DA appliquée à un seul vecteur de coordonnées factorielles d'ACI sur la zone sélectionnée. Cette Figure nous permet de déterminer les zones avec des valeurs de R^2 élevées. Ces valeurs ont été reportées dans le tableau 10 ci-dessous. La Figure 41 a2 représente le nombre de composantes indépendantes à prendre dans le modèle pour obtenir la valeur de R^2 retenue dans la Figure 41 a1. La Figure 41 a3 correspond au numéro de la composante indépendante qui seule a donné cette même valeur de R^2.
La série de Figure 41 b représente les mêmes résultats mais pour le critère RMSECV. Figure 41 b1, on visualise la valeur minimale de RMSECV pour chaque zone. Celle-ci est obtenue pour une PLS-DA appliquée à un seul vecteur de coordonnées factorielles d'ACI. C'est à partir de cette Figure que l'on sélectionne les zones d'intérêt ayant une valeur de RMSECV faible. La Figure 41 b2 représente le numéro de la composante indépendante qui seule a donné cette valeur de RMSECV dans cette zone. La Figure 41 b3 donne le nombre de composantes indépendantes extraites pour donner le modèle PLS-DA avec la valeur minimale de RMSECV retenue dans la Figure 41 b1.

Ainsi comme nous l'avons vu, à partir de ces cartes, on peut repérer pour chaque fenêtre la valeur du critère R^2 ou RMSECV et ainsi déterminer les zones qui ont des valeurs fortes de R^2 et/ou faibles de RMSECV. Nous avons reporté dans le tableau 10 les valeurs des critères les meilleures pour chacune des zones mises en évidence sur les cartes, avec leur correspondance en terme d'intervalle. De plus, pour le critère R^2 on note que les meilleures valeurs de R^2 sont obtenues en utilisant un modèle à 4 composantes ICA et la $2^{ème}$ composante donne en général les meilleurs résultats. Pour le critère RMSECV, généralement le modèle comporte aussi 4 composantes indépendantes mais c'est la $3^{ème}$ composante qui donne les meilleurs résultats.

a)



b)

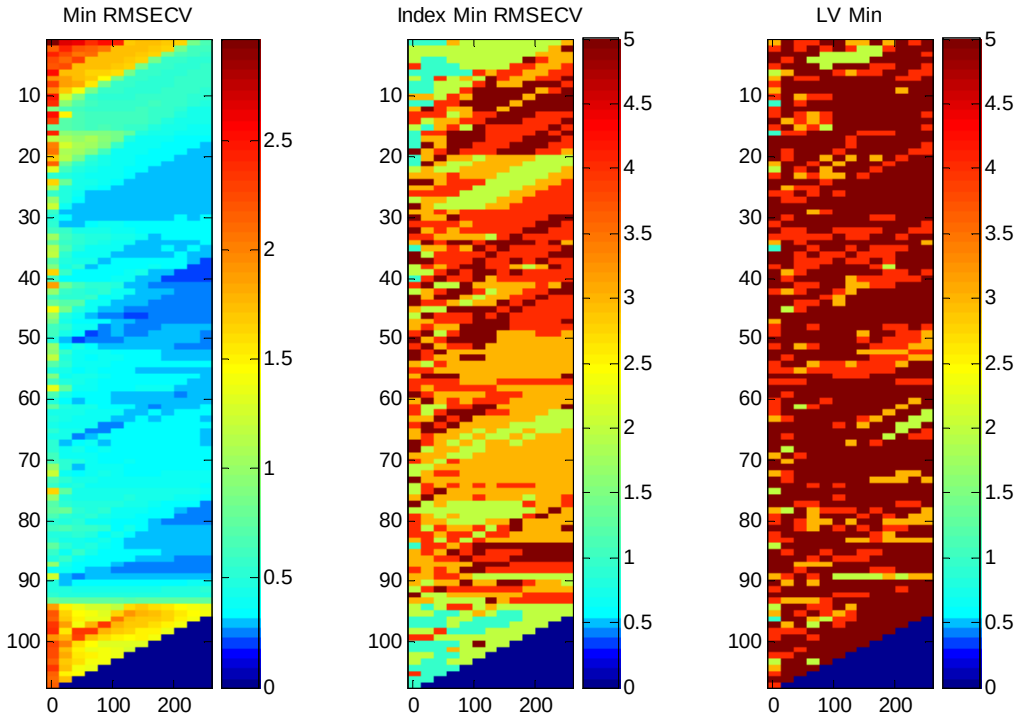


Figure 41 : Cartes couleur présentant les différents résultats obtenus par la méthode EWZS : a) critère R^2 ; b) critère RMSECV

Tableau 10. Valeurs des critères sur les zones sélectionnées

Critère R^2			Critère RMSECV		
Valeur	Début	Fin	Valeur	Début	Fin
0.9174	485	735	0.2788	586	650
0.8892	**506**	**632**	0.1976	748	978
0.9296	**748**	**915**	***0.2061***	***970***	***1034***
0.9441	768	831	**0.2149**	**990**	**1033**
0.9538	990	1033	**0.2719**	**1011**	**1241**
0.8904	1314	1357	0.2588	1294	1358
0.9243	***1597***	***1827***	***0.2584***	***1597***	***1785***
0.9195	***1778***	***1883***	***0.2491***	***1758***	***1946***

Légende :

gras: redondant : non pris en compte ; surligné : meilleur résultat ; *italique* : concaténé

Certaines zones se superposaient, par exemple 485 à 735 et 506 à 632. Le meilleur résultat obtenu pour ces 2 zones est celui de la plus large zone (0,9174 vs. 0,8892). C'est donc la première zone qui a été sélectionnée.

Quand la taille d'une zone était proche du maximum (250), et qu'une autre zone adjacente était plus grande que 50 variables, les zones ont été concaténées comme c'est le cas pour la zone donnant le meilleur résultat de RMSECV, la zone de 748 à 978 et la zone suivante : 970 à 1034.

Selon le critère R^2, cinq zones sont sélectionnées, soit 690 variables avec une zone de 44 variables donnant le meilleur R^2. Pour le critère RMSECV, quatre zones distinctes ont été sélectionnées représentant 767 variables, et la zone donnant le meilleur résultat contenait 231 variables.

La meilleure zone pour le R^2, n'est pas comme on pouvait s'y attendre la zone a2 (Figure 42 a) qui est la zone retenue comme la meilleure avec les autres techniques de sélection d'intervalle mais a3 (Figure 42 a) correspondant à un petit pic de naringine à 5,10 ppm (cf. Annexes 7 et 8).

Pour le critère RMSECV le meilleur résultat est obtenu sur une zone assez large b2 (Figure 42 b) qui contient les signaux de la naringine et de l'hespéridine (cf. Annexes 7, 8 et 9) qui sont retrouvés avec chaque méthode.

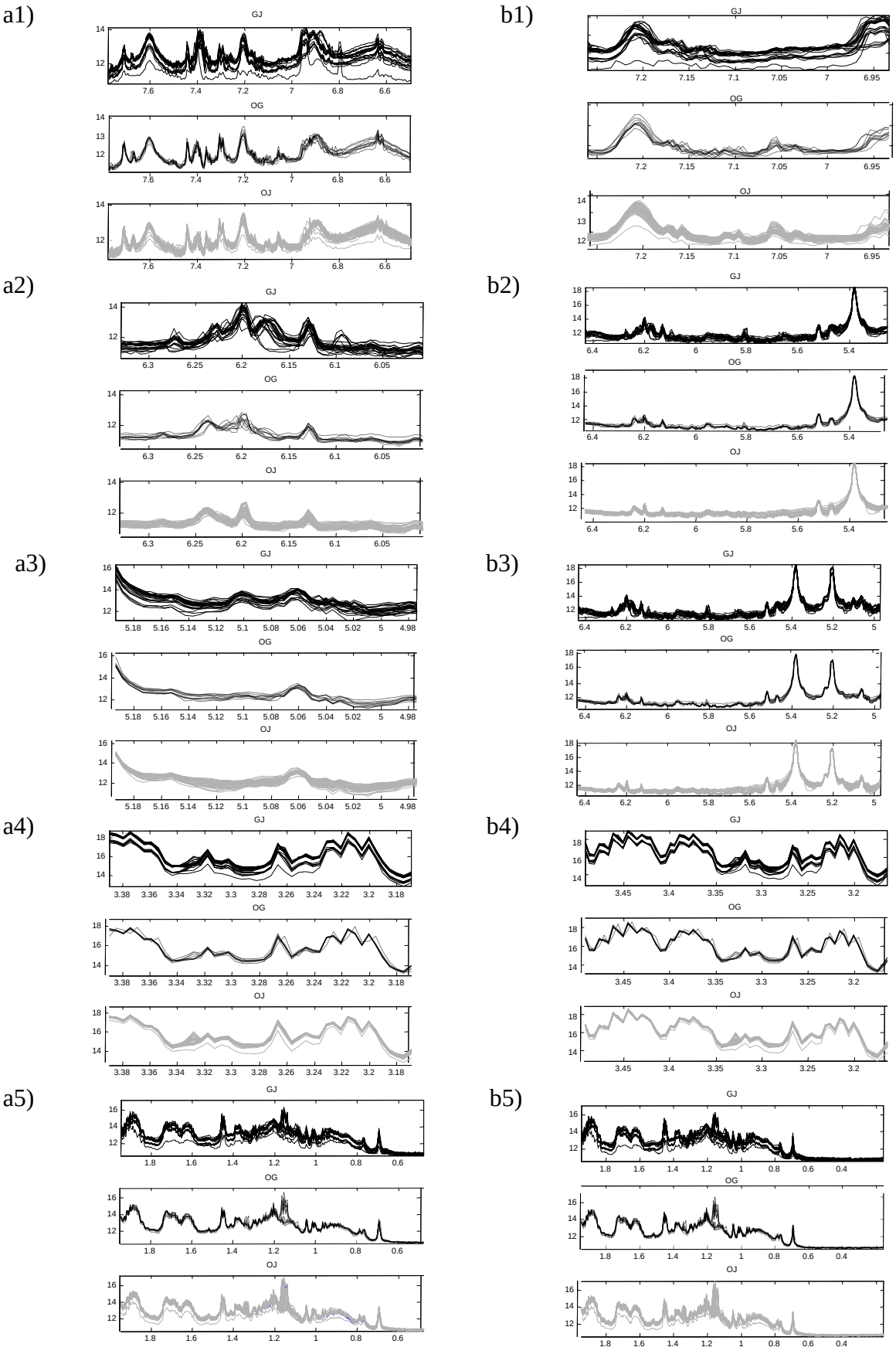


Figure 42 : Zones sélectionnées selon : a) le critère R^2 ; b) le critère RMSECV

Tableau 11. Résultats de classements basés sur les coordonnées factorielles de l'ACI

Prétraitements		Meilleur classement des 92 échantillons sur une seule composante	Meilleur classement des 92 échantillons dans les 15 espaces possibles	Composantes utilisées pour la classification
Transformation logarithmique et moyenne de 7 points	Mahalanobis	83 ; IC 3	85	ICs 2 et 3
	Euclidienne	88 ; IC 3	88	IC 3
a1	Mahalanobis	77 ; IC 3	87	ICs 2 et 3 et ICs 1 à 3
	Euclidienne	82 ; IC 3	89	ICs 3 et 4
a2	Mahalanobis	64 ; IC 1	88	ICs 2 et 3
	Euclidienne	83 ; IC 2	91	ICs 2 et 3 et ICs 1 à 3
a3	Mahalanobis	58 ; IC 3	78	ICs 3 et 4
	Euclidienne	82 ; IC 3	91	ICs 3 et 4 et ICs 1 à 4
a4	Mahalanobis	80 ; IC 1	80	IC 1 et ICs 1 à 3
	Euclidienne	81 ; IC 2	85	ICs 1 à 4
a5	Mahalanobis	81 ; IC 3	83	1 et 3 1 à 3
	Euclidienne	81 ; IC 3	84	1 et 3
a1-a2-a3-a4-a5	Mahalanobis	82 ; IC 4	86	ICs 2 et 4 1 à 4
	Euclidienne	89 ; IC 4	90	ICs 3 et 4
b1	Mahalanobis	78 ; IC 3	83	ICs 1 à 4
	Euclidienne	82 ; IC 3	87	ICs 1 à 4
b2	Mahalanobis	86 ; IC 3	87	ICs 3 et 4
	Euclidienne	90 ; IC 3	91	ICs 3 et 4
b3	Mahalanobis	83 ; IC 4	85	ICs 1 à 4 et ICs 3 et 4
	Euclidienne	86 ; IC 3	88	ICs 3 et 4
b4	Mahalanobis	76 ; IC 1	82	ICs 1 à 4
	Euclidienne	87 ; IC2	87	IC 2
b5	Mahalanobis	80 ; IC3	81	ICs 1 et 3 et ICs 1 à 3
	Euclidienne	81 ; IC3	82	ICs 1 et 3
b1-b3-b4-b5	Mahalanobis	79 ; IC 4	87	ICs 3 et 4
	Euclidienne	86 ; IC 4	86	ICs 3 et 4

a2 et a3 ont finalement des valeurs de classification identiques allant jusqu'à 91 échantillons de bien classés. Faire la concaténation des zones sur le critère R^2 permet d'avoir un taux de classement sur une composante qui augmente jusqu'à 89 en distance Euclidienne. Cette

sélection prend en compte 690 variables. La composante 4 est donc bien représentative de la différence orange / pamplemousse. Ceci peut être vérifié sur la Figure 43 ci-dessous.

a) b)

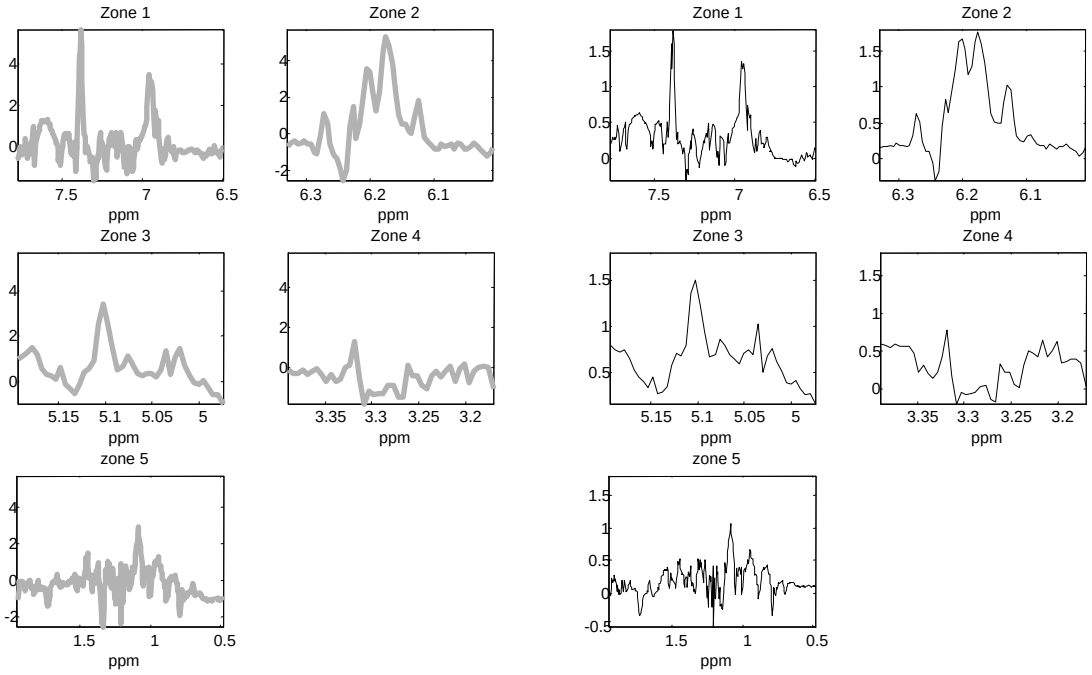


Figure 43 : a) IC 4 sur les 5 zones sélectionnées par EWZS critère R^2 concaténées ; b) Différence entre le spectre moyen de l'orange et du pamplemousse dans les 5 zones sélectionnées par EWZS critère R^2

Pour ce qui est de la sélection sur le critère RMSECV, b2 qui recouvre une zone assez importante : de 5,25 ppm à 6,42 ppm, donne de très bons résultats sur la composante 3. 90 échantillons sont bien classifiés sur 92 et 91 en la combinant avec la composante 4 (distance Euclidienne).

La concaténation des zones sur le critère RMSECV donne de moins bons résultats que sur R^2. Lorsque les zones sont concaténées, on sélectionne 35 % du jeu de données initiales (767 variables) mais l'information est moins discriminante.

Grâce à la méthode EWZS, nous avons pu retrouver les mêmes contributions spectrales que par iPLS, correspondant aux marqueurs connus de la fraude par ajout de jus de pamplemousse dans du jus d'orange (naringine et hespéridine).

De plus la classification est aussi bonne à la fois en zone individuelle et en zones concaténées.

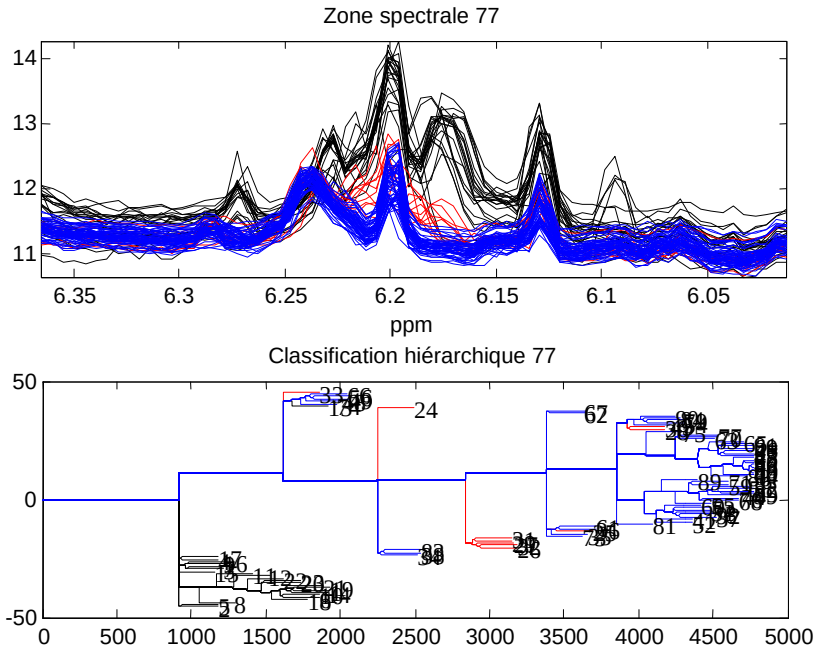
3.5.2.3.Interval-PLS_Cluster

Grâce à la méthode Interval-PLS_Cluster qui réalise une classification hiérarchique non-supervisée pour chaque intervalle, il est possible de sélectionner les zones spectrales donnant un bon regroupement des échantillons en se basant sur les caractéristiques des dendrogrammes. Le jeu de données a été analysé par Interval-PLS_Cluster avec une fenêtre spectrale de taille 70 variables et un pas de 10 variables entre chaque fenêtre. Chacun des 207 intervalles a été étudié et 11 intervalles permettant d'avoir un bon regroupement des échantillons, au vu des dendrogrammes, ont été sélectionnés. Ci-dessous le résultat du regroupement sur une des fenêtres (Figure 44). Les autres résultats sont présentés en Annexe 13.

La taille des fenêtres ici est de 70 comme pour iPLS et leurs positionnements varient comme dans EWZS. Le choix de cette taille est arbitraire. Cependant, nous avons pris en compte la taille moyenne d'un groupe de signaux pour décider de ce paramètre. Nous avons aussi décidé de nous borner à la taille et au positionnement indiqué par les méthodes sans prendre en compte les éventuelles coupures des signaux.

a)

b)



Légende :

- Jus de pamplemousse
- Mélange de jus de pamplemousse et orange
- Jus d'orange

Figure 44 : Résultats de la classification hiérarchique des jus de fruit sur la zone 77 : a) zone spectrale ; b) Classification hiérarchique

Sur la Figure 44 b, les échantillons sont séparés en deux groupes : les jus de pamplemousse en noir et les jus d'orange et mélange de jus en rouge et bleu. Un échantillon de pamplemousse est mal classifié. Il s'agit d'un échantillon « pur jus » vendu au rayon frais.
11 zones (notées a à k) ont donc été sélectionnées (cf. Annexe 13). Cependant les intervalles f, g et h et les intervalles i, j et k étant contigus nous les avons regroupés. Ainsi seules 7 zones ont été retenues au final (Figure 45 a à 45 g).

Dans les 11 zones sélectionnées ci-dessus, on retrouve les pics caractéristiques de cette séparation : naringine et hespéridine ainsi que d'autres composés attribués dans le tableau 1. La seule zone qui permette de séparer les échantillons en 2 groupes distincts est la zone 77 (cf. Figure 44).
Cet intervalle, comme l'intervalle 12 en iPLS et les intervalles A2 B2 et B3 en EWZS, contient à 6,20 le signal de la naringine et à 6,23 celui de l'hespéridine (cf. Annexes 7, 8 et 9).

Les zones sélectionnées sont comme précédemment soumises à validation par classification sur coordonnées ICA et la zone 77 est également testée à part. Les résultats sont présentés dans le Tableau 12.

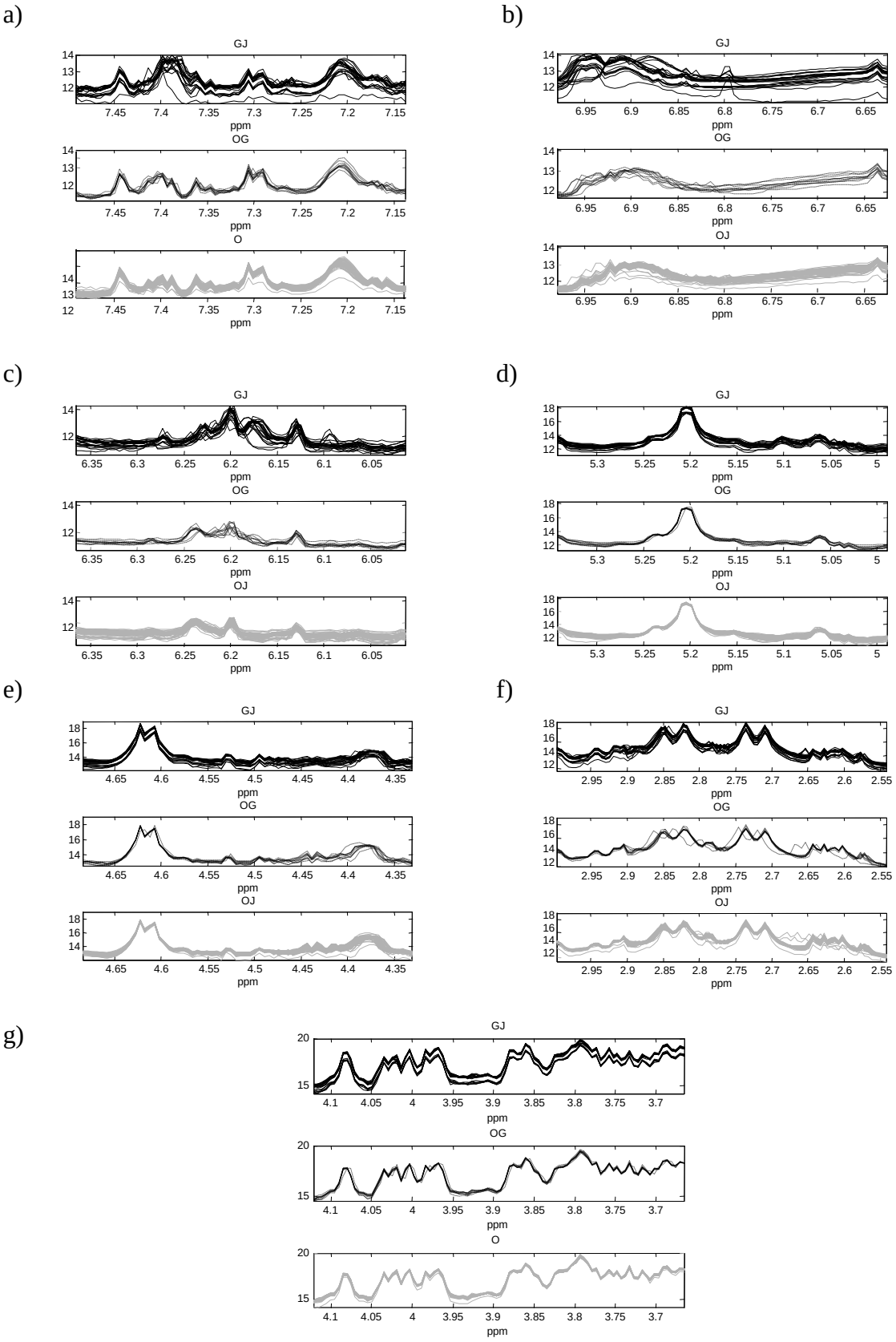


Figure 45 : Zones sélectionnées par Interval-PLS_Cluster : a) zone 55 ; b) zone 65 ; c) zone 77 ; d) zone 97 ; e) zone 107 ; f) zones 118 à 120 concaténées ; g) zones 140 à 142 concaténées

Tableau 12. Résultats de classements basés sur les coordonnées factorielles de l'ACI

Prétraitements		Meilleur classement des 92 échantillons sur une seule composante	Meilleur classement des 92 échantillons dans les 15 espaces possibles	Composantes utilisées pour la classification
Transformation logarithmique et moyenne de 7 points	Mahalanobis	83 ; IC 3	85	ICs 2 et 3
	Euclidienne	88 ; IC 3	88	IC 3
Zone 55	Mahalanobis	79 ; IC 4	90	ICs 3 et 4
	Euclidienne	83 ; IC 3	91	ICs 3 et 4
Zone 65	Mahalanobis	76 ; IC 1	84	ICs 1 à 4
	Euclidienne	79 ; IC 3	81	ICs 3 et 4
Zone 77	Mahalanobis	87 ; IC 3	89	ICs 1 et 3
	Euclidienne	89 ; IC 3	91	ICs1 à 3
Zone 97	Mahalanobis	73 ; IC 3	87	ICs 1 à 4
	Euclidienne	85 ; IC 3	86	ICs 1 à 4
Zone 107	Mahalanobis	58 ; IC 2 et ; IC 3	77	ICs 1 à 4
	Euclidienne	68 ; IC 1	77	ICs 2 et 3
Zones 118 à 120 concaténées	Mahalanobis	80 ; IC 4	82	ICs 3 et 4
	Euclidienne	78 ; IC 4	87	ICs 1 à 4
Zones 140 à 142 concaténées	Mahalanobis	58 ; IC 2	78	ICs 1 à 4
	Euclidienne	75 ; IC 4	77	ICs 3 et 4
Sélection de variables par Interval-PLS_Cluster	Mahalanobis	82 ; IC 2	88	ICs 1 à 3
	Euclidienne	82 ; IC 2	88	ICs 1 à 3 et ICs 1 à 4

Au vu des résultats individuels des zones, on a choisi d'affiner la sélection. Les zones ne donnant pas de meilleurs résultats que les données ayant subi le traitement de référence (Transformation logarithmique et moyenne de 7 points) ont été supprimées : zones 65, 107, 140 à 142 concaténées. De plus, si l'on considère seulement la zone 77, 91 échantillons sont bien classés comme avec l'intervalle 12 trouvé par iPLS. En effet, ces deux zones sont comparables.

Tableau 13. Résultats de classements basés sur les coordonnées factorielles de l'ACI

Prétraitements		**Meilleur classement des 92 échantillons sur une seule composante**	**Meilleur classement des 92 échantillons dans les 15 espaces possibles**	**Composantes utilisées pour la classification**
Transformation logarithmique et moyenne de 7 points	Mahalanobis	83 ; IC 3	85	ICs 2 et 3
	Euclidienne	88 ; IC 3	88	IC 3
Sélection affinée de variables par Interval-PLS_Cluster	Mahalanobis	88 ; IC 3	88	IC 3 ; ICs 1 et 3 et ICs 3et 4
	Euclidienne	90 ; IC 3	91	ICs 1 à 4

Avec cette sélection de 300 variables, on retrouve des valeurs de classification comparables aux précédentes (Tableau 11). La sélection choisie au départ devait contenir des zones pas assez discriminatives. Sur cette dernière sélection on obtient une composante 3 qui permet de discriminer 90 échantillons sur 92. Nous l'avons représenté Figure 46 a.

a) b)

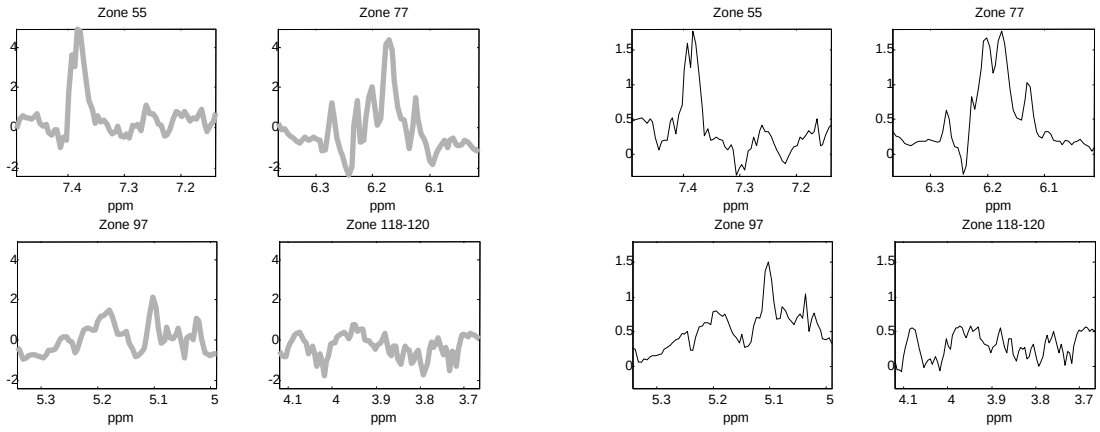


Figure 46 : a) IC 3 sur les 4 zones sélectionnées par Interval-PLS_Clusster concaténées ; b) Différence entre le spectre moyen de l'orange et du pamplemousse dans les 4 zones sélectionnées par Interval-PLS_Cluster

Ici encore la composante qui permet de faire la discrimination ressemble beaucoup à la différence entre le signal moyen des jus d'orange et des jus de pamplemousse (Figure 46 a et b). Ces zones contiennent les signaux discriminants de la naringine : pour la zone 55 à 7,39

ppm, la zone 77 à 6,20 ppm, la zone 97 à 5,10 ppm. La 4ème zone est une zone contenant les signaux des sucres. Elle a été repérée avec la méthode EWZS.

Ces différentes méthodes ont été appliquées sur les autres jeux de données présentés dans notre étude : les vinaigres balsamiques et les yaourts. Dans tous les cas nous avons pu repérer des zones intéressantes. Pour les vinaigres, ce sont principalement des composés marqueurs du vieillissement qui ont été mis en évidence. Pour les yaourts, ce sont majoritairement des composés aromatiques et des additifs de la préparation de fruit qui ont été repérés.

3.6. Conclusion sur les différentes possibilités de traitement des spectres RMN ^{1}H offertes par les outils chimiométriques

Dans nos études nous avons utilisé différents pré-traitements et traitements. Certains pré-traitements nous semblent avoir beaucoup d'influence sur le résultat final et en particulier la transformation logarithmique - quand il y a des différences d'intensité dans les zones étudiées - et le "warping" - quand il y a des variations dans le milieu entraînant des variations de déplacements chimiques des pics.

Pour les traitements des spectres, dans nos études de cas, l'ACI donnait de meilleurs résultats de classification mais surtout permettait une meilleure compréhension et interprétation des résultats que dans le cas d'une ACP.

Pour ce qui est des méthodes de sélection de variables, parmi les quatre méthodes étudiées celles qui sont les plus sélectives sont celles utilisant une sélection par intervalle : iPLS, EWZS et Interval-PLS_Cluster. Dans le cas des jus, la méthode CLV n'a pas amélioré les résultats de façon significative par rapport à la méthode basée sur la variance totale.

Les trois méthodes iPLS, EWZS et Interval-PLS_Cluster nécessitent d'avoir une bonne connaissance des données puisqu'il y a certains paramètres à optimiser (taille des fenêtres, groupes d'échantillons en présence). Cependant elles donnent de meilleurs résultats.

CLV nécessite moins d'information a priori pour le regroupement des variables et permet d'obtenir de plus nombreuses informations sur les différents composés qui peuvent différencier les types d'échantillons. Cependant lorsque l'on applique sur les résultats obtenus, une ANOVA pour sélectionner le groupe de variables à garder, on prend en compte l'information utilisée dans les techniques précédentes.

EWZS permet de ne retenir que les zones ayant un fort pouvoir discriminant. L'avantage de cette méthode est que l'on peut simultanément localiser les zones et optimiser leur taille.

Interval-PLS_Cluster permet de détecter les intervalles présentant des sous-groupes d'intérêt, mais pas forcement uniquement liés à un critère choisi *a priori*. Ce qui dans le cas du vinaigre a permis de repérer des sous-populations.

3.7. Conclusion sur le potentiel de l'utilisation de la RMN ^{1}H pour qualifier les produits agroalimentaires de manière simple et rapide

Les trois différentes études ont permis de montrer qu'il était possible de traiter les spectres afin d'obtenir des informations quant à leur origine ou leur process.

En effet, dans l'étude sur le jus d'orange nous avons montré qu'il était possible de trouver des marqueurs d'identification des jus et que malgré la forte variation intra-spécifique et la variation sur les procédés de fabrication, on pouvait déterminer l'origine botanique des fruits utilisés pour fabriquer le jus. Seuls quelques échantillons ont été mal attribués du fait de leur grande différence par rapport au reste du jeu de données : des échantillons comportant de l'orange sanguine et du pamplemousse rose. Pour ces échantillons, il serait nécessaire d'agrandir la base de données ou de les traiter à part. Dans l'étude sur le vinaigre balsamique, nous avons montré qu'il était possible de discriminer les échantillons en fonction de leur procédé de fabrication et même d'aller plus loin avec le traitement PLS_Cluster et de regrouper les échantillons en sous-classes en fonction de leur origine géographique (région d'Emilia).

Dans l'étude sur les yaourts nous avons montré que la qualité du yaourt : aux fruits, à la pulpe de fruits ou aromatisé pouvait être déterminée en aveugle grâce à l'étude des spectres. De plus, il a été possible de mettre en évidence un gradient sur les différentes concentrations des échantillons.

Cependant bien que les résultats soient très positifs il faut mentionner les difficultés que l'on pourrait rencontrer si le procédé était industrialisé. Tout d'abord, l'automatisation de la méthode poserait quelques difficultés. En effet, bien que la préparation de l'échantillon soit très simple et que le processus d'acquisition des spectres soit lui aussi facile à cadrer par une procédure, le traitement des spectres ne peut pas se faire en boîte noire. Il y a différentes étapes où il faut avoir un esprit critique face aux résultats. Tout d'abord, ces traitements sont spécifiques d'une matrice et d'une fraude. Des phases de développement seraient nécessaires pour chaque extension de la méthode. Ensuite, si le spectre présente des pics non conformes mais non pris en compte dans les zones sélectionnées il faut être capable de les repérer et donc de faire un traitement visuel de chaque spectre. Enfin, au niveau des prétraitements : l'étape la

plus délicate est le « warping » car si les paramètres ne sont pas bien choisis le spectre peut être altéré et les résultats faussés.

Par ailleurs pour des analyses d'authenticité et en particulier de pureté des produits la comparaison d'un spectre RMN de l'échantillon étudié à une base de données de référence est une méthode sûre. Elle permet en outre, de remonter aux données structurales de l'adultérant et donc de l'identifier. Ce type de méthode est donc à envisager plus pour des matrices moins complexes comme des produits pharmaceutiques.

3.8. Références

1. Hammond, D. A. 13c - a refined method to detect the addition of cane/corn derived sugars to fruit juices and purees. *Fruit Processing* (1998).
2. Lees, M. *Food Authenticity : Issues and Methodologies*, pp. 311 (Eurofins Scientific Laboratories, Nantes, 1999).
3. Le Gall, G., Puaud, M. & Colquhoun, I. J. Discrimination between orange juice and pulp wash by ^{1}H NMR spectroscopy: Identification of marker compounds. *J. Agric. Food Chem.* 49, 580-588 (2001).
4. Cocchi, M., Bro, R., Durante, C., Mancini, D., Marchetti, A., Saccani, F., Sighinolfi, S. & Ulrici, A. Analysis of sensory data of Aceto Balsamico Tradizionale di Modena (ABTM) of different ageing by application of PARAFAC models. *Food Qual. Prefer.* 17, 419 (2005).
5. Giudici, P. Gluconic acid: genuineness criterion of the traditional balsamic vinegar. *Ind. Bev.*, 123 (1993).
6. Del Signore, A. Infrared spectra (Mid-IR) classification of balsamic vinegar. *Journal of Commodity Science* 39, 159-172 (2000).
7. Del Signore, A., Campisi, B. & Di Giacomo, F. Characterization of balsamic vinegar by multivariate statistical analysis of trace element content. *Journal of AOAC International* 81, 1087-1095 (1998).
8. Del Signore, A., Stancher, B. & Calabrese, M. Characterization of balsamic vinegars by amino acid content using a multivariate statistical approach. *Italian Journal of Food Science* 12, 317-332 (2000).

9. Cocchi, M., Durante, C., Foca, G., Mancini, D., Marchetti, A. & Ulrici, A. Application of a wavelet-based algorithm on HS-SPME/GC signals for the classification of balsamic vinegar. *Chemometrics and intelligent laboratory systems* 71, 129-140 (2004).
10. Chiavaro, E., Caligiani, A. & Palla, G. Chiral indicator of ageing in balsamic vinegar of Modena. *J. Food Sci* 4, 329 (1998).
11. Anklam, E., Lipp, M., Radovic, B., Chiavaro, E. & Palla, G. Characterisation of Italian vinegar by pyrolysis-mass spectrometry and a sensor device ('electronic nose'). *Food Chem.* 61, 243-248 (1998).
12. Caligiani, A., Acquotti, D., Palla, G. & Bocchi, V. Identification and quantification of the main organic components of vinegars by high resolution ^{1}H NMR spectroscopy. *Analytica Chimica Acta*, 110-119 (2007).
13. Cidilait. La consommation. http://www.cidilait.com/, 07/01/2008, (2007).
14. Consonni, R. & Gatti, A. ^{1}H NMR studies on Italian balsamic and traditional balsamic vinegars. *J. Agric. Food Chem.* 52, 3446-3450 (2004).
15. IFU. *58 : Determination of hesperidin and naringin by HPLC*, pp. (1991).
16. Borin, A. & Poppi, R. J. Application of mid infrared spectroscopy and iPLS for the quantification of contaminants in lubricating oil. *Vibrational Spectroscopy* 37, 27-32 (2005).
17. Noorgard, L. iToolbox Manual. (2002).
18. Steuer, R., Morgenthal, K., Weckwerth, W. & Selbig, J. A Gentle Guide to the Analysis of Metabolomic Data. *Methods Mol Bio.* 358, 105-126 (2007).
19. Raychaudhuri, S., Stuart, J. M. & AltmanY, R. B. in *Pacific Symposium on Biocomputing*, 452-463 (Stanford Medical Informatics, 2000).
20. Han, L. et al. QTL Studies with Microarray Data. http://www.stat.wisc.edu/~yandell/talk/campus/2000.biology.pdf (2000).
21. Cuny, M., Vigneau, E., Le Gall, G., Colquhoun, I. J., Lees, M. & Rutledge, D. N. Fruit juice authentication by ^{1}H NMR spectroscopy in combination with different chemometrics tools. *Analytical and Bioanalytical Chemistry* 390, 419-427 (2008).
22. Cuny, M., Le Gall, G., Colquhoun, I. J., Lees, M. & Rutledge, D. N. Evolving Window Zone Selection method followed by Independent Component Analysis as useful chemometric tools to discriminate between grapefruit juice, orange juice and blends. *Analytica Chimica Acta* 597, 203-213 (2007).

23. Barros, A. S. & Rutledge, D. N. PLS_Cluster: a novel technique for cluster analysis. *Chemom. Intellig. Lab. Syst.* 70, 99-112 (2004).
24. Jouan-Rimbaud Bouveresse, D., Barros, A. S. & Rutledge, D. N. Generalised PLS_Cluster: an extension of PLS_Cluster for interpretable hierarchical clustering of multivariate data. *Sensing and Instrumentation for Food Quality and Safety* 1, 79-90 (2007).
25. USDA. USDA Nutrient database. http://www.nal.usda.gov/fnic/foodcomp/search/, 14/01/2008.

4. Publications

4.1. Publication 1

Dairy product authentication by ^{1}H NMR spectroscopy in combination with different chemometric tools

Cuny, M., Vigneau, E., Lees, M. & Rutledge, D. N. (2006). Magnetic resonance in food science. I. A. Farhat, P. S. Belton and G. A. Webb, RSC Publishing**:** 197-204.

Introduction

As an important dietary source of calcium, protein and vitamins A, B12, and D, the consumption of milk and dairy products is on an upward trend. This increase is mirrored by the growing number of novel fresh dairy products (fat-free, full-fat, set, stirred, drinkable, probiotic) available on the supermarket shelves, very often containing varying proportions of fruit, flavours or both to boost sensory appeal. The authentication of such products generally requires analysis of a large number of parameters linked to the components in the product. Therefore the availability of a rapid analytical method for this purpose would be a great help in ensuring compliance with labelling legislation. As dairy products are rich in fat, few methods are suitable for broad screening techniques, and the measurement of fruit content itself is a notoriously difficult task.

NMR spectroscopy is emerging as a popular technique for food analysis as it can be used for both quantitative compositional analysis and for rapid throughput screening[1]. Combined with chemometric techniques it can be a useful and rapid method to assess food authenticity and it has already been applied to fruit juices and to olive oil in particular to detect various types of adulteration.

The aim of this study was to assess the potential of ^{1}H NMR spectroscopy in analysing fruit-containing dairy products. The measured NMR response is assumed to be a function of the relative proportions of the components in the mixture. The work presented here focuses on different chemometric tools used to extract relevant information from the ^{1}H NMR spectrum. The results of these different methods are assessed in terms of their ability to discriminate between types of yoghurt based on their fruit content.

Method

Sample Preparation

65 strawberry yoghurts (16 flavoured yoghurts, 15 stirred fruit pulp yoghurts, 34 fruit yoghurts) were collected from local supermarkets and frozen prior to preparation (Table 1 Publication / Tableau 14). The samples were assigned to a percentage group (0 %; <10 %; >10 %) according to the fruit content indicated on the label. Sample preparation consisted of 25 min of centrifugation at 4500 rpm, after which 750 µL of supernatant was collected. To this was added 250 µL of D_2O containing 0.75 % of TSP (sodium 3-(trimethylsilyl)propionate-2,2,3,3-d4). No additional treatment was necessary. The D_2O was used as a source of the field frequency lock signal, and TSP for internal referencing of ^1H chemical shifts.

Tableau 14. (Table 1 Publication) : List of samples and associated fruit content group

Sample Type	% Group	Number of samples
Fruit yoghurt	<10 %	16
Fruit yoghurt	>10 %	18
Total:		**34**
Stirred fruit pulp yoghurt	<10 %	14
Stirred fruit pulp yoghurt	>10 %	1
Total:		**15**
Flavoured yoghurt	0 %	**16**

Data Acquisition

^1H NMR spectra were run on a Bruker Advance DPX-400 spectrometer using the noesypr1d sequence with water peak suppression during relaxation delay and mixing time. For each spectrum 520 transients were accumulated with 32 K data points. The acquisition time was about 3.5 sec and the spectral width was 12 ppm. Each experiment was carried out in 45 min.

Data Pre-treatment

The phase was corrected by hand using Bruker Topspin software. All other data treatment was carried out using Matlab version 7. These consisted of base-line correction, warping using the Correlation Optimised Warping function with linear interpolation (COW[2,3]) and data reduction. This data reduction was obtained by using the mean of 11 adjacent points to concentrate the information contained in the ^{1}H NMR spectrum. Finally a logarithmic transformation was applied to the data to reduce the difference in scale between the different parts of the spectrum. In this way the initial spectrum of 32 K points was reduced to 2241 variables, the transformation is shown in Figure 1 Publication / Figure 47.

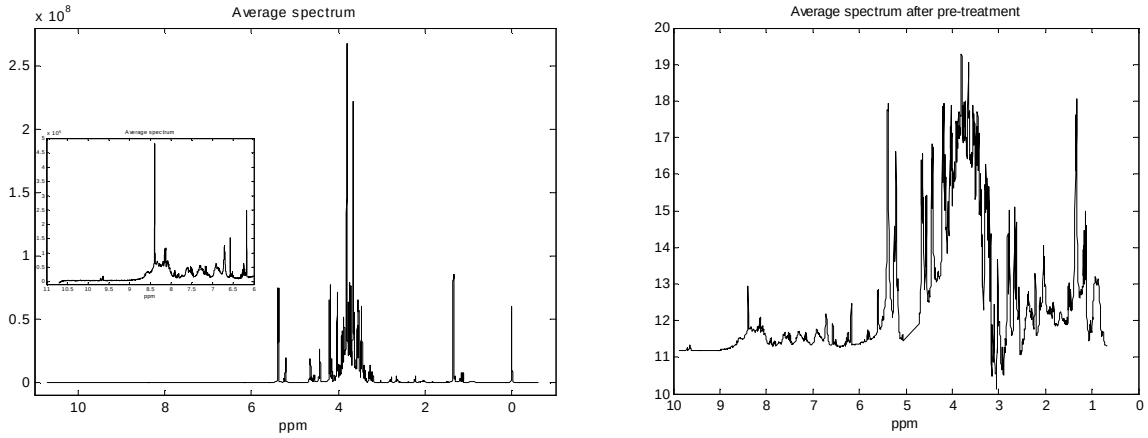


Figure 47 : (Figure 1 Publication) : a) Averaged ^{1}H NMR spectrum of all samples; b) Averaged data set after pre-treatment with logarithmic transformation

Chemometric treatment of data

Variable Selection. Three different approaches were investigated to select NMR variables in order to predict yoghurt type or fruit percentage level.

Selection of variables with high variance (Variance method) – Since the variance of a variable is a measure of the average squared distance between the individual values of the variable for the set of data points and their mean value, this criterion can be used to identify the most discriminant variables in the data set. Indeed the latter are expected to vary more than noisy or non-informative variables. A variance threshold, above which variables were kept in the data set, was defined by the ability of the resulting data to predict membership of

the samples to the correct group (both for yoghurt type or for fruit percentage) using the cross-validation technique.

Selection of clustered variables (CLV[4,5] method) – A cluster analysis around latent variables is applied in order to identify groups among the spectral variables, linked either by their covariance or their correlation. Organising multivariate data into a small number of clusters, each represented by a latent component, makes it possible to reduce dimensionality. The CLV method involves two stages, namely a hierarchical clustering analysis followed by a partitioning algorithm. Partitioning is determined by the value of a quality criterion, in this case T, which is the sum of the first eigenvalues of the data matrices of all the clusters. Then the component most likely to discriminate between samples in terms of yoghurt type or fruit percentage is determined by Analysis of Variance (ANOVA).

Selection of variables defining zones of continuous variables (Evolving Window Zone Selection - EWZS method) – This approach focuses on detecting different continuous parts of the spectrum that are shown to be most discriminant either for yoghurt type or for fruit percentage. This function is used to do a singular value decomposition – based PCA and then both a cross-validation PLS on all the scores and simple linear regressions between each vector of scores and the variable to be predicted (yoghurt type or fruit percentage). The function plots maps of the Root Mean Standard Error of Cross Validation (RMSECV) values calculated by cross-validation PLS and the correlation coefficients (R^2) calculated by linear regressions for each area of the spectrum. EWZS uses windows of varying size, from a small user-defined minimum up to a user-defined maximum (which may be the width of the data set). The first set of windows starts with the first variable. Once the window has increased to the user-defined limit, the starting point of the new set of windows is incremented by a user-defined step. In that way visual examination of these maps is used to identify those zones showing a minimum RMSECV value and a maximum R^2 value which are then selected as informative continuous zones of variables.

Comparison of Variable Selection Methods. Factorial Discriminant Analysis (FDA) was then carried out on the principal components of the parts of the spectrum selecting using the methods described above. To compare the discriminant power of the three approaches the number of correctly classified samples was calculated using a leave-one-out cross-validation method.

Results and discussion

Discrimination of yoghurt types and fruit percentage groups on the dataset (32442 variables) without data reduction and logarithmic transformation

To be able to compare the efficiency of the treatment and variable selection we have taken the dataset constituted of the spectra that were submitted to phase and base-line correction and warping (Figure 1.a / Figure 47a) as a reference. It was tested for discrimination of groups. The results are in Table 2 Publication / Tableau 15.

Tableau 15. (Table 2 Publication) : Summary of results for the dataset without data reduction and logarithmic transformation

Number of variables selected	30442	
FDA on yoghurt types	F1 82 %	F2 18 %
N° of samples correctly-classified		
For 3 groups	32	49 %
For 2 groups	48	74 %
FDA on fruit percentages	F1 94 %	F2 6 %
N° of samples correctly-classified	39	60 %

Only 49 % of samples were well-classified for the three yoghurt types and 60 % for groups of fruit percentages.

Discrimination of yoghurt types and fruit percentage groups on the dataset with data reduction and logarithmic transformation without variable selection

After the data reduction and logarithmic transformation, the set of variables kept was tested for discrimination (results in Table 3 Publication / Tableau 16).

Tableau 16. (Table 3 Publication) : Summary of results for the dataset with data reduction and logarithmic transformation

Number of variables selected	638	28.5 %
FDA on yoghurt types	F1 91 %	F2 9 %
N° of samples correctly-classified		
For 3 groups	43	66 %
For 2 groups	62	95 %
FDA on fruit percentages	F1 93 %	F2 7 %
N° of samples correctly-classified	45	69 %

This selection and transformation enhanced the prediction by 17 % points for the discrimination between the three yoghurt groups, 21 % points for the two groups, and of 9 % points for the fruit percentages. The data size reduction enables the algorithm of classification to be run more rapidly. The logarithmic transformation helped in reducing the variation of intensity in spectra and homogenised the weight of variables in the PCA.

Discrimination of yoghurt types and fruit percentage groups after selection based on high variance criterion

The variance of the variables is shown in Figure 2 Publication / Figure 48. The highest variability is found in the region from 0 to 5.5 ppm of the spectrum. This corresponds mainly to sugar and organic acid content. The number of correctly classified samples using the leave-one-out cross-validation technique was calculated for a variance ranging from 0.01 to 0.1 and the classification carried out for yoghurt types and fruit percentages. The best result of classification was obtained for a threshold of 0.045 for the assignment of yoghurt type. For the fruit percentages the number of correctly classified sample is relatively stable between the threshold values ranging from 0.035 to 0.050, and therefore the value of 0.045 is chosen.

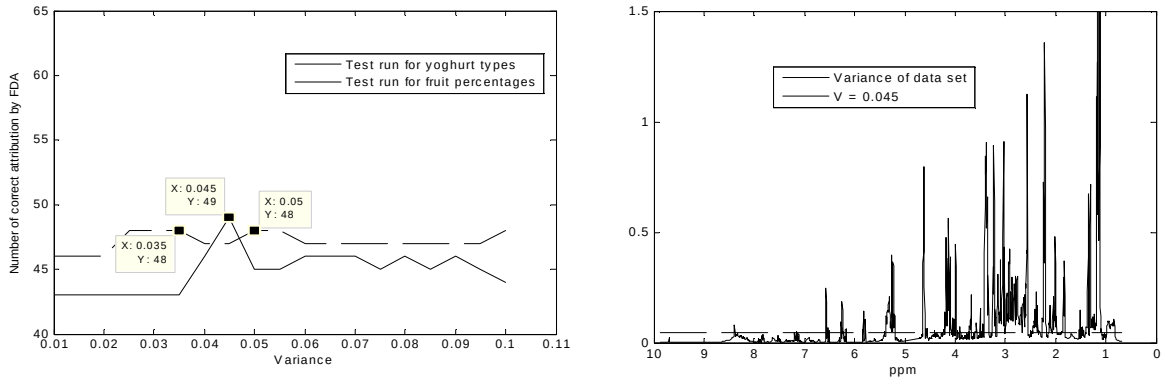


Figure 48 : (Figure 2 Publication) : a) Results of the selection algorithm to define the variance threshold; b) Selected variance threshold on variance data set.

The results for a variance threshold of 0.045 are presented here (Table 4 Publication / Tableau 17). This led to the selection of 638 variables (Figure 5 Publication / Figure 51). The first three principal components of the reduced data set were submitted to FDA to predict either the yoghurt type or the fruit percentage. In the case of yoghurt type, discrimination into two groups (flavoured vs. others) was also tested. The calculation time of this method was very rapid, about 3 minutes.

Tableau 17. (Table 4 Publication) : Summary of results for Variance method

Number of variables selected	638	28.5 %
FDA on yoghurt types	F1 80 %	F2 20 %
N° of samples correctly-classified		
For 3 groups	49	75 %
For 2 groups	62	95 %
FDA on fruit percentages	F1 79 %	F2 21 %
N° of samples correctly-classified	47	72 %
Calculation time:	rapid	

Here the highest increase in correct classification is for the discrimination between yoghurt types, plus 9 % points, thus 75 % of samples were correctly attributed. The two other classifications are similar or better until 95 % of correct classification for the three yoghurt types and 72 % for fruit percentages.

Discrimination of yoghurt types and fruit content with selection based on CLV method

CLV analysis has two options for grouping variables, either by covariance or correlation. Both analyses were carried out but only the first case giving the best results, is presented here (Figure 3 Publication / Figure 49).

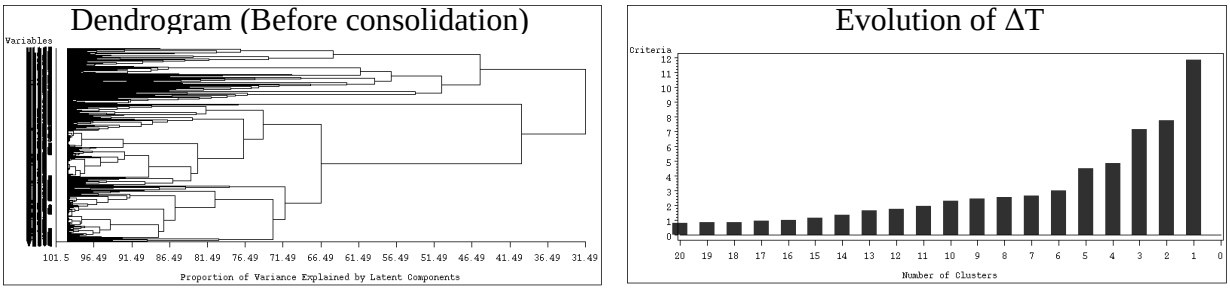


Figure 49 : (Figure 3 Publication) : a) Dendrogram of variables; b) Evolution of the criterion ΔT with the number of clusters

The calculation time of this part of the analysis is rather long, but can be decreased by using other algorithm options.

The graph of ΔT (Figure 3b Publication / Figure 49 b) observed during the hierarchical clustering showed that when passing from a partitioning of 4 to 3 groups the T criterion significantly increased. An ANOVA on the latent variables associated to the four groups of the retained partition was used to identify the significant components for explaining yoghurt type (Table 5 Publication / Tableau 18).

Tableau 18. (Table 5 Publication) : ANOVA results for yoghurt type on the four latent obtained by CLV

Source	Sum of Squares	d.f.	Mean Squares	F	Prob>F
Component 1	0.014	2	0.007	0.44	0.649
Component 2	0.073	2	0.036	2.44	0.0955
Component 3	0.371	2	0.185	18.28	5.74E-07
Component 4	0.840	2	0.420	162.93	0

The latent components of the 3rd and 4th groups were significant at α = 5 %. Therefore the variables of both groups were selected. Next, variables were linked manually to reshape some

of the spectral areas. This was done for both groups. Their discriminatory power was tested both separately and jointly. The best results shown here were obtained with the 275 variables from the 4th group (Figure 5 Publication / Figure 51).

An FDA was carried out based on yoghurt types and fruit percentages. The predictive power was assessed by the number of correctly classified samples in a leave-one-out cross-validation, results in Table 6 Publication / Tableau 19.

Tableau 19. (Table 6 Publication) : Summary of results for CLV method

Number of variables selected	275		12 %	
FDA on yoghurt types	F1	91 %	F2	9 %
N° of samples correctly classified				
For 3 groups	44		68 %	
for 2 groups	64		98 %	
FDA on fruit percentages	F1	79 %	F2	21 %
N° of samples correctly classified	54		83 %	
Calculation time:	long			

With the 275 selected variables representing 12 % of the initial variables, the discrimination of flavoured *versus* pulp fruit stirred and fruit yoghurt is 98 %. In addition the correct classification according to fruit content is of 83 %, representing an increase of 14 points for the dataset after logarithmic transformation. Those results are much better than without or with data reduction and logarithmic transformation, and this method enhanced the classification of 3 % points for two yoghurt types and 11 % points for fruit percentages in comparison with the variance selection method.

When the variables selected from the third component were added to the variables selected from the fourth component the results were not as good but similar to those which could be obtained by using the 2241 initial variables.

Discrimination of yoghurt types and fruit content with selection based on the EWZS method

Applying the EWZS function to the data set gave with a relatively short calculation time, the maps shown in Figure 4 Publication / Figure 50. The minimum window size was set to 11 variables, the maximum window size to 250 while the increment between each series of evolving windows was a step of 20 variables. Visual examination of the R^2 map shows that

variables ranging from 1900 to 2141 along with the zone from 520 to 1000 are the most discriminant for yoghurt type. The minimum RMSECV plot shows that the variables 1900 to 2141 along with the zones from 1220 to 1340, and from 1620 to 1760 are the most discriminant for yoghurt type.

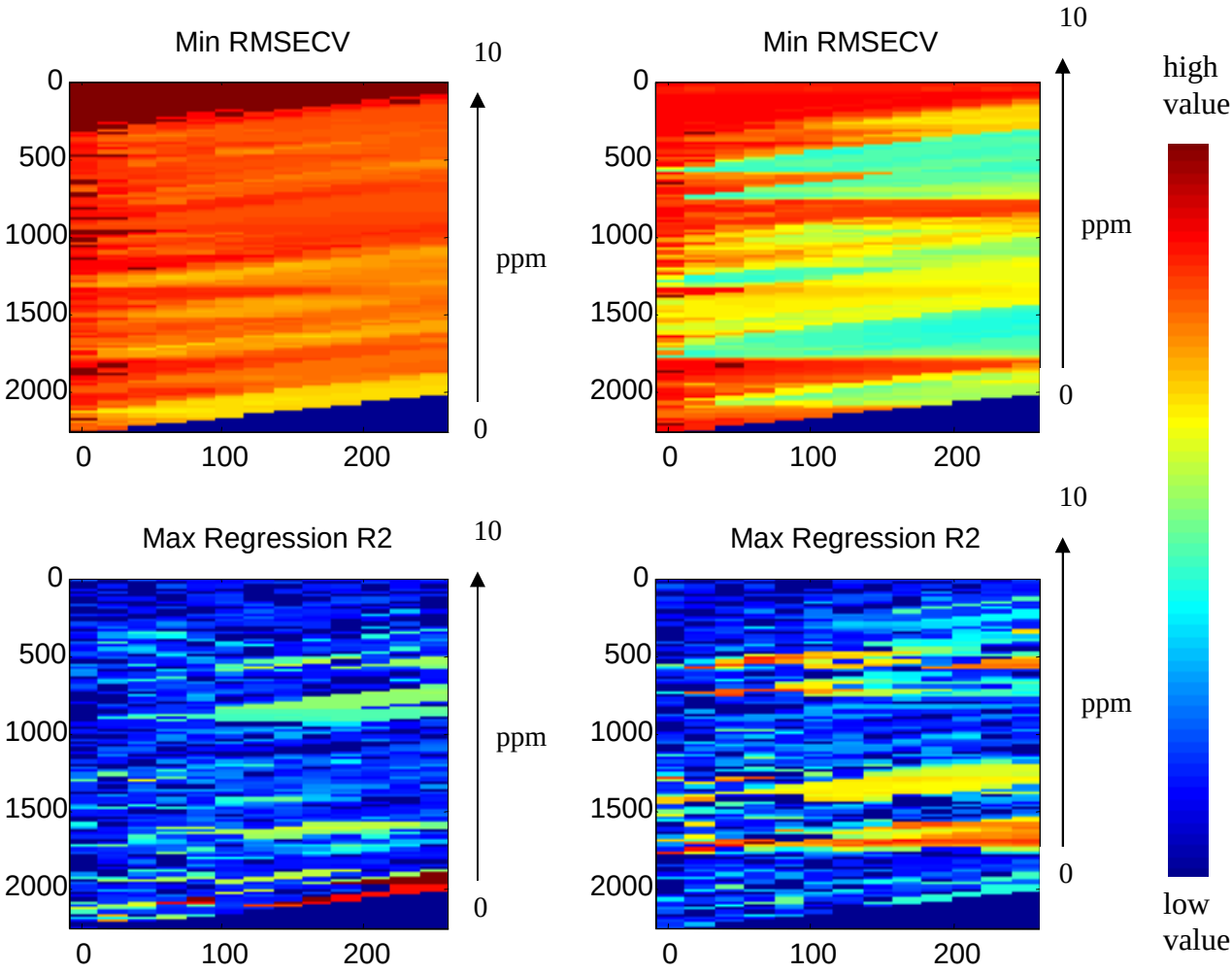


Figure 50 : (Figure 4 Publication) : Maps of minimum RMSECV and maximum correlation coefficients (R^2) for varying size zones up to 250 variables wide and initially starting at the first point: a) prediction of yoghurt type; b) prediction of fruit percentage

The zones selected in this way are shown in Figure 5 Publication / Figure 51 for both yoghurt type and percentage fruit groups. A summary of results for both groups is given in Table 7 Publication / Tableau 20.

Tableau 20. (Table 7 Publication) : Summary of results for EWZS method

Number of variables selected	1105	49 %
FDA on yoghurt types	F1 92 %	F2 8 %
N° of samples correctly-classified		
for 3 groups	46	71 %
for 2 groups	64	98 %
Number of variables selected	722	32 %
FDA on fruit percentages	F1 95 %	F2 5 %
N° of sample correctly-classified	51	78 %
Calculation time:	intermediate	

The results for the discrimination of the two types of yoghurt are similar to the CLV method, for the three types they are better by 3 % points. With 51 samples correctly classified for their fruit contents, that method is not as good as the CLV method but is still better then the variance method.

Conclusion

The three sets of selected variables gave similar results for both predictions, with more than 70 % correctly classified samples. If only two types of yoghurts (flavoured and fruit yoghurts) are taken into account, the best results were obtained using the more complex algorithms for the selection of variables, with 98 % of the samples correctly classified.

All of the approaches investigated enabled a reduction in data set size, ranging from 51 % for the EWZS method to 88 % for the CLV method. Since the EWZS function selects continuous parts of signals, if two relatively close peaks give good predictions then the zone containing both will also be picked as a good predictor. This means that the variables in between will also be included. This is not the case with the CLV and the variance methods, which therefore give smaller data sets. In addition, if the information being used to discriminate samples is not especially close, the result of the EWZS function may not be as good as when only the significant marker peak is retained.

Two main zones containing information that discriminate yoghurt types were found by all three methods: chemical shifts between 2.5 and 3 ppm corresponding to major organic acids, and the zone between 1.1 and 1.3 ppm. To a lesser extent, some parts of the aromatic signals

were selected in the low field region. None of the major sugar components were selected by the CLV method for discrimination (area between 3 and 6 ppm).

The two methods, CLV and EWZS, which involve some multi-dimensional optimisation criteria for selection, provide more information to interpret the results. The major difference between these two methods is in the calculation times – the EWZS methods being significantly much faster. However another way to do CLV is to perform, first, a partitioning algorithm with many groups (more than the number of the individuals), but far less than the number of variables. Then the latent components of these groups are submitted to the classic CLV algorithm. In the first stage, an initial partition is chosen at random and the quality criterion to be optimised is the same criterion as in classical CLV. This method is almost as fast as the variance method although the results will depend on the initial partition.

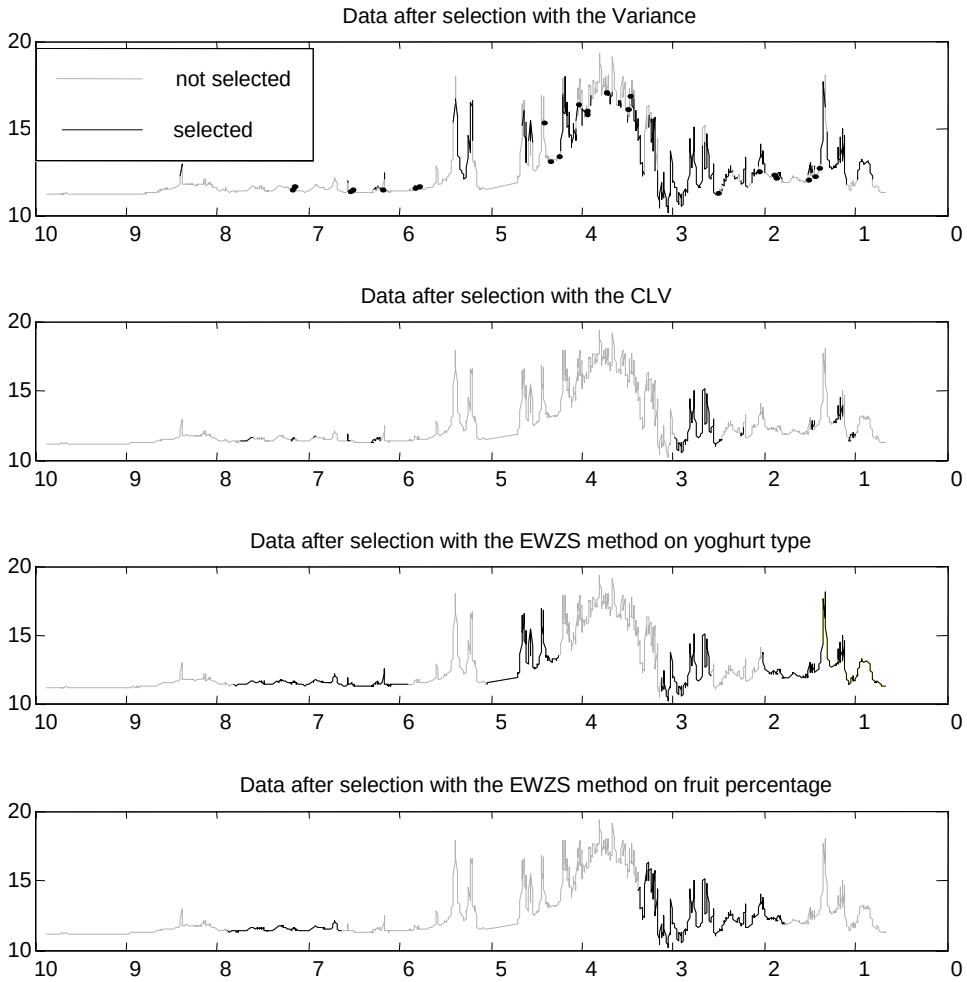


Figure 51 : (Figure 5 Publication) : Selected variables on the averaged spectrum and results of discriminations by applying the different selection methods

Future work will be on testing the feasibility of using the Variance method and/or the EWZS method as input to the slower but more selective CLV method as such a prior reduction in the size of the data set should significantly reduce the calculation time.

The samples used in this study are commercially available products with no guarantee of their authenticity. This work will be further backed up by a study of authentic samples of defined fruit content. These preliminary results, however, show that ^{1}H NMR with suitable data treatment is a promising approach to the quantification of fruit content in yoghurt and other complex matrices.

References

1. Le Gall, G. & Colquhoun, I. J. in *Food authenticity and traceability*, 131-155 (2003).
2. Tomasi, G., van den Berg, F. & Anderson, C. Correlation Optimized Warping and time warping as preprocessing methods for chromatographic data. *J. Chemometrics* **18**, 231-241 (2004).
3. Nielsen, N. P. V., Carstensen, J. M. & Smedsgaard, J. Aligning of single and multiple wavelength chromatographic profiles for chemometric data analysis using correlation optimised warping. *J. Chromatography A* **805**, 17–35. (1998).
4. Vigneau, E., Sahmer, K., Qannari, E. M. & Bertrand, D. Clustering of variables to analyse spectral data. *J. Chemometrics* **19**, 122-128 (2005).
5. Vigneau, E. & Qannari, E. M. Clustering of variables around latent components. *Communications in statistics* **32**, 1131-1150 (2003).

4.2. Publication 2

Evolving Window Zone Selection method followed by Independent Component Analysis as useful chemometric tools to discriminate between grapefruit juice, orange juice and blends

Cuny, M., Le Gall, G., Colquhoun, I. J., Lees, M. & Rutledge, D. N. (2007). Analytica Chimica Acta **597**: 203-213.

Introduction

Orange juice consumption is increasing worldwide. Economic adulteration has always been a concern for the industry and a great deal of time and energy has been expended in ensuring that the market remains fair and free from adulterated product. At the European level, for example, the fruit juice industry operates an advanced quality control system, the EQCS (European Quality Control System) that brings together the majority of national quality control schemes within the European Union.

One of the main tools for ensuring fruit juice authenticity is the Code of Practice established by the AIJN (Association of the Industry of Juices and Nectars). These industry guidelines set acceptable ranges for a large number of compositional parameters (sugars, organic acids, mineral content) that can be used to judge the authenticity of a juice. This implies measuring a large number of parameters with an assortment of analytical techniques. Routine mass screening of food products is a concept that has been much discussed in the fruit juice industry because of the obvious appeal of a one-step acceptance or rejection of an incoming batch. One of the major criticisms of these techniques has been the over-reliance on statistical treatment to reject a sample without being able to identify the cause of the non-compliance.

NMR spectroscopy is emerging as a popular technique for food analysis as it can be used both for quantitative compositional analysis and for rapid throughput screening[1-8]. Although the latter employs chemometric techniques to analyse the data from complex spectra, it is often possible to home-in on specific sections of the spectrum and to make reasonable assumptions about why one sample stands out from the rest.

The overall aim of the study is to assess the potential of ^{1}H NMR spectroscopy in combination with chemometric techniques as a rapid screening method for the determination of adulteration of orange juice with grapefruit juice. This method could provide an alternative to the standard HPLC method [9], which measures the flavonoid glycosides - hesperidin,

naringin and narirutin. The work presented here focuses on the use of a procedure called Evolving Window Zone Selection (EWZS) to select zones that best discriminate between predefined sample groups in a sample set. The latter in this case was composed of orange (OJ), grapefruit (GJ) and blend (OG) juices obtained commercially.

Two multivariate methods - Principal Component Analysis (PCA) and Independent Component Analysis (ICA) - were programmed as options within the EWZS function in order to determine which was better able to detect informative spectral zones.. Once the interesting zones had been selected, PCA and ICA were applied to all possible combinations of these zones to determine their ability to segregate the samples as a function of the type of juice. Since Principal Component Analysis may be considered as the standard method for multivariate data analysis, the reference segregation for comparison purposes was the best of those obtained by applying PCA to the zones selected by PCA. The zones selected by ICA were then submitted to both PCA and ICA. Because it aims to extract the original ("pure") source-signals from compound signals [10, 11] rather than extract vectors corresponding to directions of greatest dispersion of the samples, as is the case for PCA, it is expected that the ICA approach may be a more effective tool.

Experimental

Sample collection.

92 juices were bought in local supermarkets in Norwich, UK, in July 2005. These included 59 orange juices (OJ), 23 grapefruit juices (GJ), of which 15 were white grapefruit juices and 8 pink grapefruit juices, and 10 blends (OG) of which 2 were with pink grapefruit juices. The amount of grapefruit in the juice blends ranged from 10 to 50 % according to the information provided on the label. No % content was specified for one of the samples. In addition to these three categories, the sample set covered three different manufacturing processes: Not From Concentrate (NFC), Chilled From Concentrate (CFC), and Longlife From Concentrate (LFC). There were 24 NFC orange juices, 9 NFC grapefruit juices and 3 NFC blends. For the CFC juices, there were 12 orange and 4 grapefruit. For the LFC juices, there were 22 orange, 10 grapefruit, and 7 blends. Hence a representative dataset of the type of orange and grapefruit juices available on the market was obtained.

For the NMR measurement, the pH of 12 to 14 mL of the juices was adjusted (with vigorous mixing) to 4.00 +/- 0.03 to minimise any fluctuation of the chemical shift of certain

compounds in the spectra [12]. The pH adjusted juice was added to an Eppendorf (~1.5 mL) and centrifuged at 10 000 rpm for 5 minutes. The supernatant was collected. 550 µL of a sample containing 75 % of the treated supernatant and 25 % of D_2O containing TSP (sodium 3-(trimethylsilyl)propionate-2,2,3,3-d_4) as internal reference was transferred to a 5 mm NMR tube. No additional treatment was necessary. The D_2O was used as a source of the field frequency lock signal, and TSP for internal referencing of ^1H chemical shifts.

^1H NMR measurements.

^1H NMR spectra were recorded at 27°C on a 600 MHz BRUKER spectrometer fitted with an auto-sampler (IFR). D_2O was used as the internal lock. Each spectrum consisted of 128 scans of 16384 complex data points with a spectral width of 7183.9 Hz (11.97 ppm). A NOESY pre-saturation sequence suppressed the water signal with low power selective irradiation at the water frequency during the recycle delay (2s) and the mixing time (0.15s). Spectra were Fourier transformed with 1 Hz line broadening, automatically phased and manually baseline corrected using the TOPSPIN software.

Data processing.

Spectra were converted to J-CAMP6 files and transferred to a personal computer for data analysis carried out in Matlab® (The MathWorks Inc, Natick, Massachusetts, USA).

Data pre-treatment.

These consisted in standard NMR signal processing procedures used to eliminate artefacts before multivariate data analysis. The spectra were first base-line corrected, then uncontrolled variations in chemical shifts among the spectra were eliminated by peak alignment using a Correlation Optimised Warping function with linear interpolation (COW [13, 14]). To accelerate the calculations without significant loss of information, the size of the spectra was reduced by eliminating regions without any peaks and then taking the mean of 7 adjacent points. In this way, the initial spectrum of 16 K points was reduced to 2143 variables. This slight reduction in the spectral resolution is negligible compared to the standard "bucketing" procedure used in NMR. Finally, a logarithmic transformation was applied to the data to reduce the difference in scale between the different parts of the spectrum. This pre-treatment is often used to enhance the minor peaks when the dynamic range of the signal is very high. The effect of the transformation is shown in Figure 1 Publication / Figure 52. A group membership vector is created containing 1 if the sample is GJ, 2 if the sample is OG and 3 if

OJ. Although the relation between the group membership values is not linear, the order of the values reflects the relationship between the groups and in particular the intermediate position of the blends (OG). The discrepancy between the "true" unknown membership values, which depend on the exact proportions of orange and grapefruit juice in the mixtures, and the linear sequence used here (1, 2 and 3) is independent of the spectral zone being tested and the multivariate method being used. Therefore, the comparison of the linear regression statistic (R^2) calculated for each zone and using each method is a valid indicator of the information contained in the zone concerning the membership value.

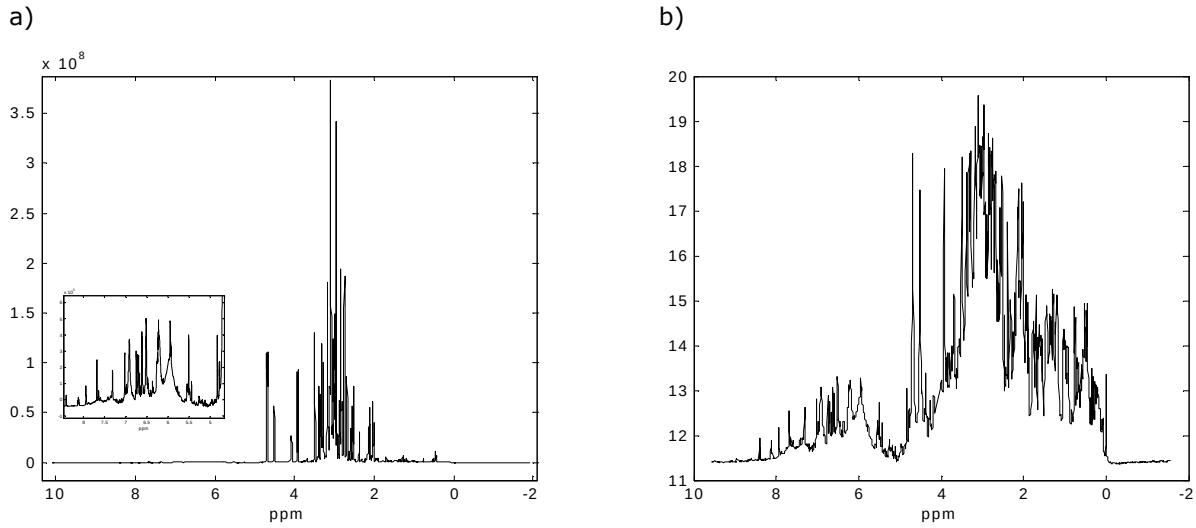


Figure 52 : (Figure 1 Publication) : a) Average ^{1}H NMR spectrum of all samples; b) Average data set after pre-treatment with logarithmic transformation

Independent Component Analysis (ICA).

ICA is a technique developed to extract the pure underlying signals from a set of mixed signals with unknown proportions. A matrix of spectra may be considered as a series of observed signals where the amplitude of spectral features is the result of a weighted sum of "source" spectra and where the weighting coefficients are proportional to the concentrations of the corresponding pure compounds. If only the observed spectra are known then ICA may be able determine both the unknown proportions and the unknown pure spectra. This is possible because of the underlying assumption that the "pure" sources are by definition totally independent. ICA attempts to recover the original signals by estimating a linear transformation, using a criterion related to information theory and entropy and which reflects the statistical independence among the sources. This may be achieved by the use of higher-

order information that can be extracted from the joint probability densities of the data [10, 11]. By finding a demixing transformation that minimizes dependencies between the estimates of the "pure" sources [15], [16] Independent Components can be recovered from a data set of mixed signals. In the present paper, the JADE (Joint Approximate Diagonalization of Eigenmatrices) ICA algorithm [17-19] was applied to the set of fruit juice ^{1}H NMR spectra. The more commonly used PCA method was also applied as a means of evaluating the results from the ICA by comparing the classification rates obtained using the two methods.
The "pure" source signals extracted by ICA may be considered as comparable to the loadings obtained by PCA. Therefore, the coordinates of the samples on these ICA "loadings" may be considered as comparable to PCA scores.

Selection of variables defining zones of contiguous variables (Evolving Window Zone Selection - EWZS method).

This approach focuses on detecting different contiguous parts of the spectrum that are shown to be most discriminant for the predefined sample groups. EWZS is similar to moving-window PLS [20] and spectral window selection [21] in that it uses windows of varying sizes, from a small user-defined minimum up to a user-defined maximum (which may be as large as the width of the complete data set). The first set of windows starts with the first variable. Once the window has increased to the user-defined maximum limit, the starting point of the new set of windows is incremented by a user-defined step. However EWZS differs from the other window selection procedures in that they both use PLS regression to relate the spectra to the dependent variable, whereas EWZS can use a range of multivariate methods. In the present study, ICA and PCA were used to decompose the dataset within each window and an ordinary linear regression was then performed between each vector of ICA or PCA "scores" and the vector of group membership values in order to detect the interesting zones. The function then plots a map of the correlation coefficients (R^2) for each region of the spectrum for the two analyses. A visual examination of these maps is used to identify those zones showing a maximum R^2 value, which may then be selected as informative contiguous zones of variables. As a result, two sets of zones were obtained.

Leave-one-out cross-validation was used to calculate the rate of correct classification for the different PCA and ICA models applied to the different spectral zones. The left-out sample was classed into the nearest group based on its Euclidian distance from each group barycentre. The optimal decomposition of the spectra by ICA resulted in 3 Independent Components.

The number of ICs being limited, it was therefore possible to perform an exhaustive comparison of the classification models based on all possible combinations of the ICs, rather than use a classical stepwise method. The same procedure was also applied to the first 3 Principal Components of the PCA.

Results and discussion

Selection of variables.

Applying the EWZS function described above to the pre-treated (warped and log-transformed) data set, and using the sample type as the criterion to be predicted, gave the R^2 map shown in Figure 2 Publication / Figure 53. The minimum window size was set to 7 variables, the maximum window size to 250, while the increment between each series of evolving windows was a step of 25 variables.

R^2 values were higher when using the ICA option (maximum of 0.9544) than with the PCA option (maximum of 0.9226).

Visual examination of the map from the ICA based decomposition shows that four zones of high R^2 can be identified. Variables ranging from 540 to 600 (corresponding to Zone 1) along with the zones from 700 to 850 (Zone 2) and from 950 to 1050 (Zone 3) and from 1600 to 1850 (Zone 4) are the most discriminant for sample type.

With PCA, the zones are slightly different. First, the zone 4 was not selected. Secondly, the zones were slightly wider for PCA selection - from 520 to 600, from 680 to 880 and from 960 to 1100, respectively.

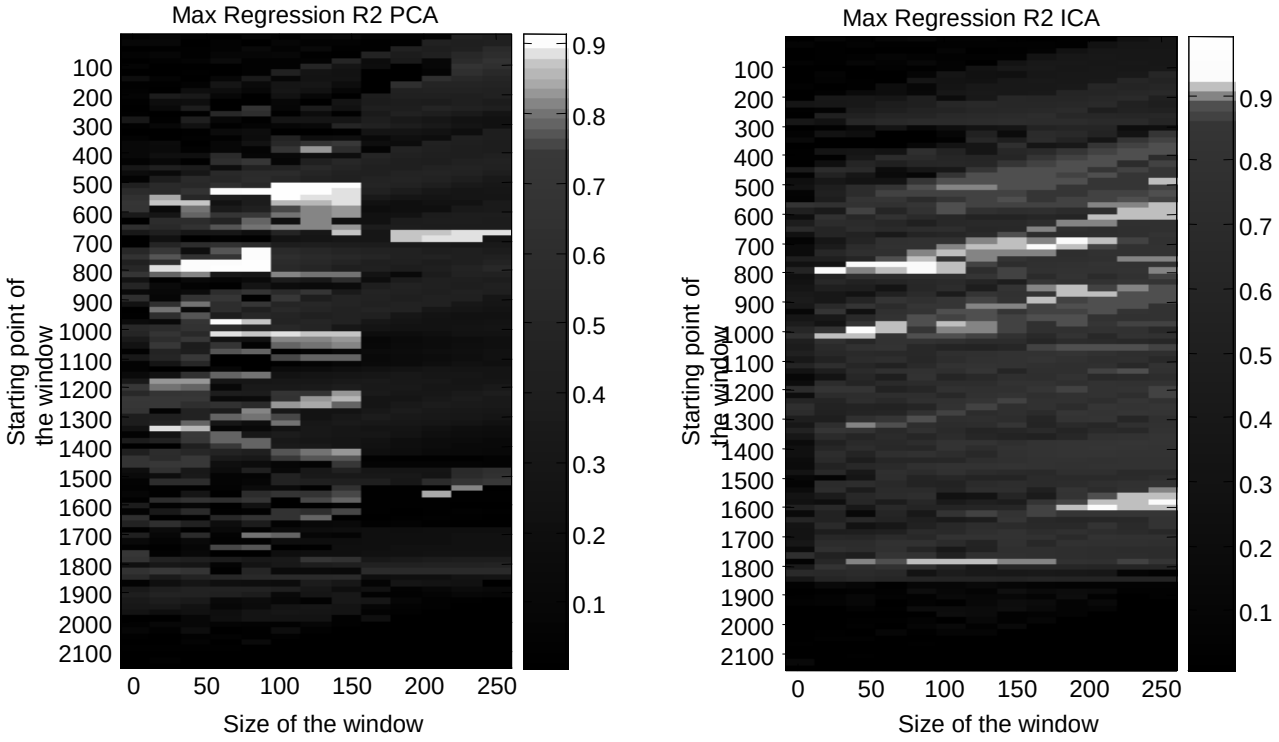


Figure 53 : (Figure 2 Publication) : Results of EWZS on the pre-treated dataset for sample types: a) with the PCA option; b) with the ICA option

Identification of the chemical constituents contributing to the selected zones.

To assign the peaks that are in the selected zones, it is necessary to examine the corresponding zones in the complete dataset without logarithmic transformation. Although the transformations have preserved the information in the spectra, they have changed the shape of signals, making interpretation a little more difficult. Using previously published data [22, 23], the signals were assigned as shown in Figures 3 and 4 Publication / Figure 54 and 55.

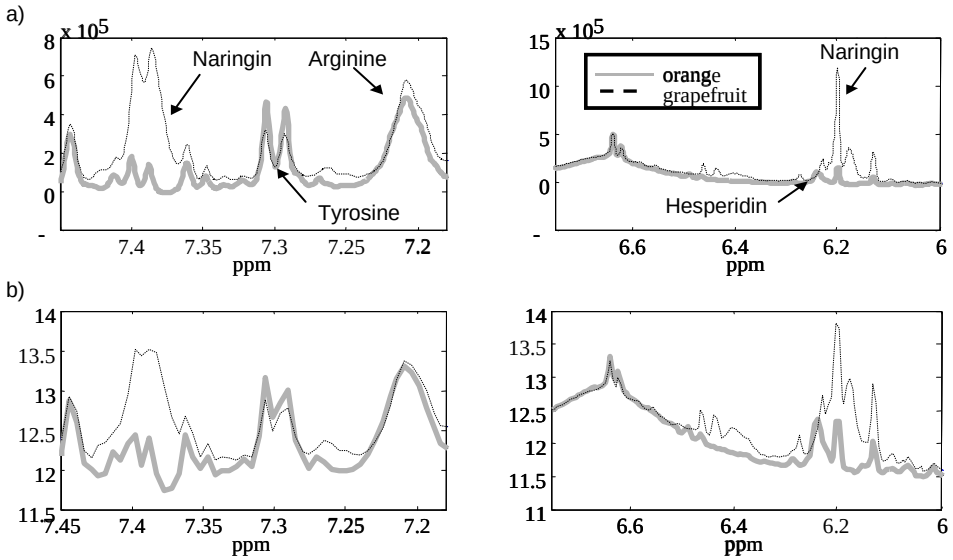


Figure 54 : (Figure 3 Publication) : Zones 1 and 2: a) from the average spectra of orange and grapefruit juice; b) from the average data set after pre-treatment with logarithmic transformation

Zones 1 and 2 contain phenolic compounds such as naringin (peaks at 6.20 and 7.38 ppm) and hesperidin (peak at 6.23 ppm), and amino acids such as arginine (peak at 7.21 ppm) and tyrosine (peaks at 7.30 ppm). It is interesting to note that the EWZS method automatically found the flavonoid glycoside components used in the standard method of orange juice authentication. Some other signals in these zones were not attributed. The average grapefruit spectrum shows higher naringin and arginine peaks, while the average orange spectrum shows a higher tyrosine peak.

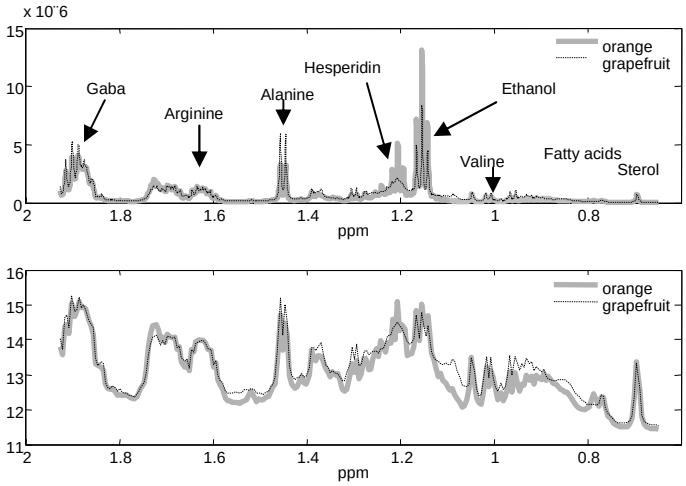


Figure 55 : (Figure 4 Publication) : Zone 3: a) from the average spectra of orange and grapefruit juices; b) from the average data set after pre-treatment with logarithmic transformation

Zone 3 in the mid-field section of the spectrum corresponds primarily to the main sugars. It contains a sucrose peak at 5.39 ppm and the α-glucose peak at 5.20 ppm as well as some smaller signals between 4.85 and 5.15 ppm. Based on previously published data, it is possible that the signal at 5.11 ppm corresponds to the H-1 of the Rha group in naringin [24]. The average grapefruit juice spectrum shows higher naringin and α-glucose peaks.

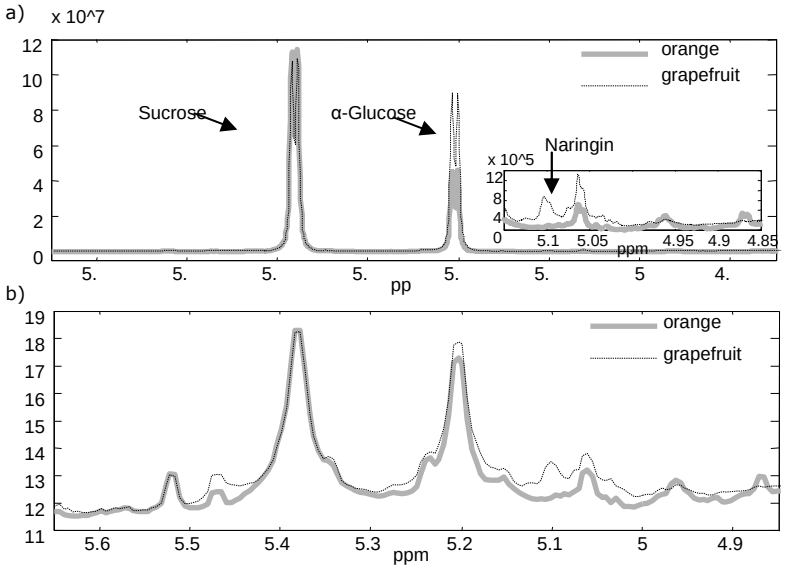


Figure 56 : (Figure 5 Publication) : Zone 4: a) from the average spectra of orange and grapefruit juices; b) from the average data set after pre-treatment with logarithmic transformation

Zone 4 only found by the ICA option, is located between 0.64 and 1.92 ppm (Figure 5 Publication / Figure 56). In this area other parts of the hesperidin and arginine signals can be found at 1.22 ppm and around 1.68 ppm respectively. The most important peak is the one from ethanol at 1.20 ppm. Ethanol appears in the juices when the fermentation starts. Samples containing a lot of ethanol are of low quality. This zone contains also two signals from proteins: alanine (1.45 ppm) and valine (1.08 ppm) and some signals from fatty acids (0.70-1.00 ppm), sterol (0.69 ppm) and GABA (1.88 ppm).

Results of the PCA on the zones selected by PCA.

All the combinations of zones from the PCA selection were tested (Table 1 Publication / Tableau 21) and the best rate was taken as the reference. Using zones 1,2 and 3, 58 samples were correctly classified by leave-one-out classification using the PC1 and PC2 (cf. Table 1 Publication / Tableau 21). Hence, the 63 % of correct classification was taken as the reference for the other classifications in this study.

Tableau 21. (Table 1 Publication) : Results of the leave-one-out classification (Euclidian distance) of 92 samples based on combination of the first 3 PCs for the different combinations of the selected zones from the PCA option in EWZS.

PCA on	Zone 1	Zone 2	Zone 3	Zone 1 & 2	Zone 1 & 3	Zone 2 & 3	Zone 1to 3
PC 1	56	56	39	60	50	51	56
PC 2	44	54	49	53	47	53	49
PC 3	28	18	50	28	46	27	18
PCs 1 & 2	56	58	48	63	49	57	58
PCs 1 & 3	51	54	41	58	47	47	46
PCs 2 &3	32	44	51	34	47	48	43
PCs 1 to 3	53	57	52	58	48	52	50

Underline : best result

Results of the ICA and PCA on the zones selected by ICA.

The ICA was performed on each separate zone and on all 4 zones together.

In each case, a three-independent-component ICA was carried out on the log-transformed zones. The "loadings" extracted by ICA are presented in Figures 6a, 8a, 9a and 10a Publication / Figures 57a, 59a, 60a, 61a, while the coordinates of the samples on these vectors (“scores”) are presented in Figures 6b, 8b, 9b, 10b Publication / Figures 57b, 59b, 60b, 61b.

Zone 1

With the classification based on ICA models, only one orange juice was misclassified as can be seen in Table 1 Publication / Tableau 21 giving the best rate of correct classification 98.9 % of all the zones. It is a better prediction result than the one from the PCA based classification, with 83.7 % of correctly-classified samples. Both results are better compared to the results found using PCA on the zones selected with the PCA option in EWZS (68.5 %). To explain these results the ICA loadings of the 92 samples are shown in Figure 6a.

Indeed, the first IC source IC1, resembles the average orange juice signal in zones 1, shown in Figure 3 Publication / Figure 54 b. The second contains mainly the naringin peaks with a small contribution from arginine. It thus corresponds to the differences between the orange and grapefruit juice in this section of the spectrum. The third component resembles a derivative of the second and is possibly due to residual variations in chemical shifts that were not completely eliminated by the pH adjustment and COW abscissa warping.

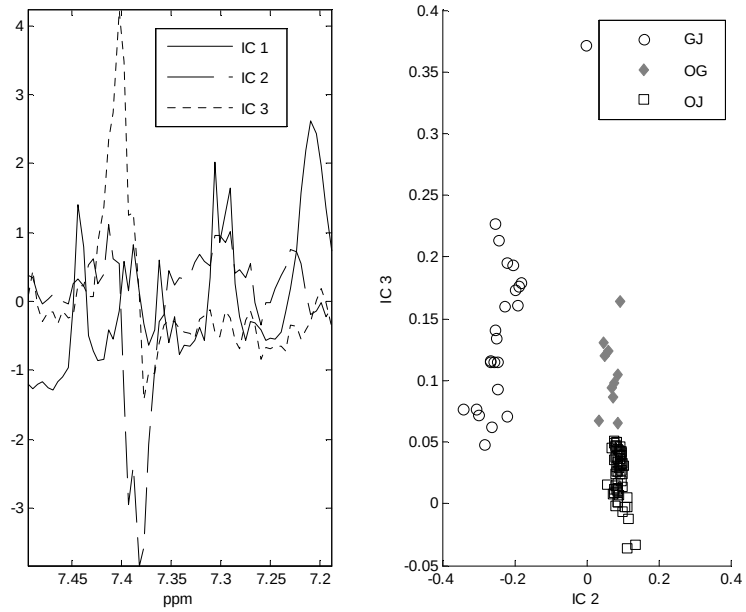


Figure 57 : (Figure 6 Publication) : Results of ICA on the first zone: a) Loadings; b) Scores on the IC2-IC3 plane

IC1 does not appear to be discriminant for the samples, whereas IC2 is (cf. Table 2 Publication / Tableau 22). All the grapefruit juices show more negative values for this component (Figure 6b Publication / Figure 57b), the orange juices have a value around 0 and the mixed juices have slightly negative values. The separation of the sample groups can be clearly seen in the IC2-IC3 plane.

Tableau 22. (Table 2 Publication) : Results of the leave-one-out classification (Euclidian distance) of 92 samples based on combination of the 3 ICs.

	Zone 1 7.19 - 7.50 ppm		**Zone 2 5.91 - 6.68 ppm**		**Zone 3 4.89 - 5.40 ppm**		**Zone 4 0.65 – 1.93 ppm**		**Zone 1 to 4**	
	ICA	**PCA**	**ICA**	**PCA**	**ICA**	**PCA**	**ICA**	**PCA**	**ICA**	**PCA**
1	67	57	58	63	51	40	32	58	29	50
2	**80**	45	**83**	53	41	53	**78**	42	42	18
3	**77**	64	**76**	55	**85**	33	41	54	**85**	33
1 & 2	**81**	60	**81**	64	50	46	**83**	49	45	48
1 & 3	**78**	73	**75**	67	**84**	46	44	62	**85**	56
2 &3	**91**	65	**87**	53	**87**	43	**78**	49	**83**	27
1 to 3	**88**	**77**	**86**	67	**84**	47	**83**	62	**85**	55

Underline : best result

Bold : Result above 75 correctly classified samples.

The spectrum of the misclassified sample shows (Figure 7 Publication / Figure 58) a low content of flavonones [25] in comparison to the average spectrum of orange juice, especially for its arginine content. It was a particular sample labelled ‘organic orange juice’.

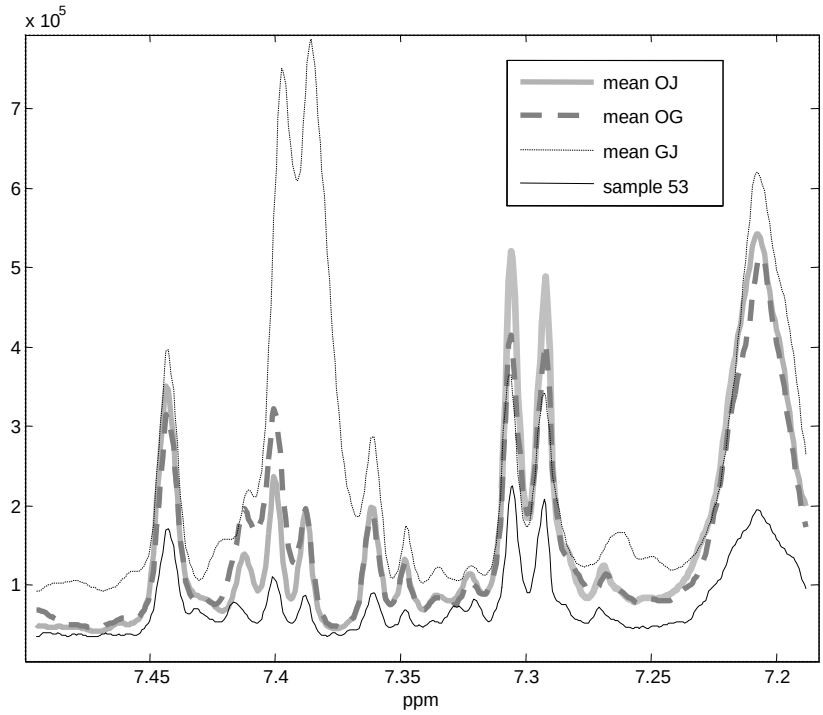


Figure 58 : (Figure 7 Publication) : Zone 1 corresponding part of the average spectra of orange, grapefruit and blend juices and the misclassified sample by ICA on zone 1

Zone 2

The results of classification on zone 2 (Table 2 Publication / Tableau 22), give an advantage to the ICA method with 90.2 % of correct classification versus 72.8 % for PCA. In the same way as for zone 1, the ICA loadings were investigated (Figure 8a Publication / Figure 59a).

As in the first zone, IC1 resembles the average orange juice signal in zones 2. The second source IC 2, contains mainly the naringin peaks at 6.20 ppm with a small contribution from the little signals at 6.43 ppm. It thus corresponds to the differences between the orange and grapefruit juice in this section of the spectrum. The third component resembles a derivative of the second and is possibly due to residual variations in chemical shifts that were not completely eliminated by the pH adjustment and COW abscissa warping.

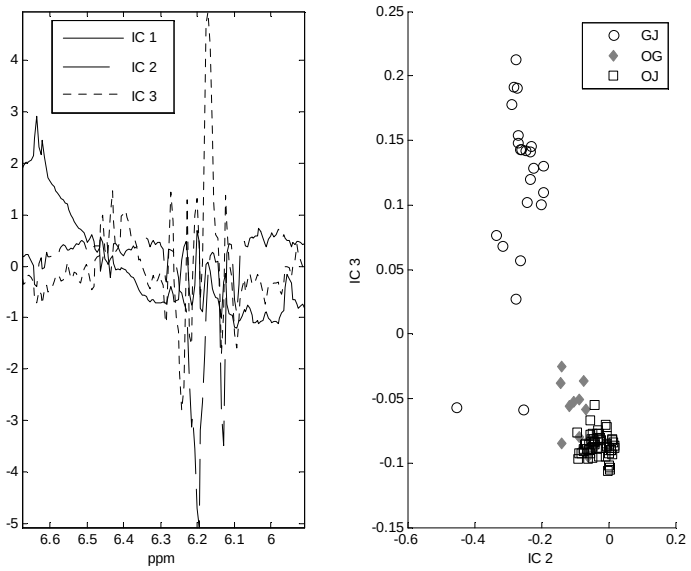


Figure 59 : (Figure 8 Publication) : Results of ICA on the second zone: a) Loadings; b) Scores on the IC2-IC3 plane

Four orange juices were classified among the mixed juices. One of them was a fresh blood orange juice. Blood orange could have a different phenolic composition. The spectrum of this sample shows (Figure not published) lower intensities in the entire region between 6.33 and 6.67 ppm. From the other misclassified orange juices another is a "smooth" juice and a third is a bottom of the range product. The last one is a juice with a geographic origin specified as Florida. The mixed juice not correctly classified, labelled 85 % orange juice and 15 % pink grapefruit juice, contains pink grapefruit juice that tends to be slightly lower in total flavonones [25] and especially in naringin. Contrary to the blood orange juice, this sample shows higher intensities in the 6.33 and 6.67 ppm area.

Zone 3

The third zone was submitted to ICA using three ICs. As can be seen in Figure 9 Publication / Figure 60, the third IC is most discriminant for sample type. It contains a small peak of α-glucose as well as negative signals of the unassigned peaks between 4.9 and 5.00 ppm that represent the differences due to the grapefruit juices and blends. It also contains the naringin signal at 5.11 ppm that is typical of grapefruit containing juices. The first IC is most representative of the orange juice spectra. As with the second zone, 4 orange juices are misclassified, one of them was already misclassified, the one from Florida. The three others are an Italian blood orange juice, a Spanish orange juice and a bottom of the range product.

The mixed juice misclassified is the same as before, its low naringin content explains this misclassification. Here again, ICA is predicting in a better way the juice type with 92.4 % vs. 57.6 % of correct prediction.

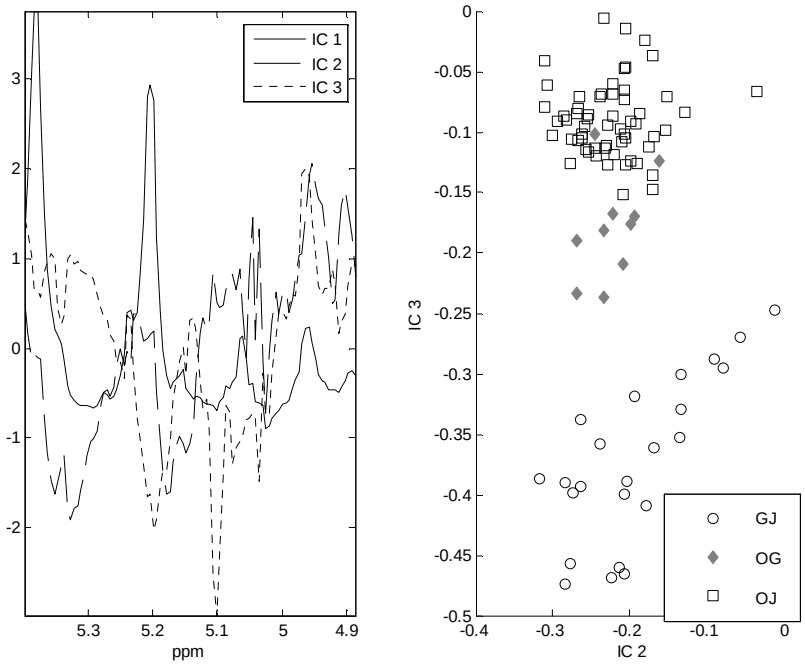


Figure 60 : (Figure 9 Publication) : Results of ICA on the third zone: a) Loadings; b) Scores on the IC2-IC3 plane

Zone 4

Zone 4 is not giving as good results of prediction as the other zones. Indeed this zone is wider and may contain more noise than the others. However 90.2 % of the samples were correctly classified with the ICA vs. 67.4 % for PCA (Table 2 Publication / Tableau 22). The study of the 3 ICs (Figure 10a Publication / Figure 61) shows that IC 1 is representative of the grapefruit juices, when IC 2 is representative of the differences between the orange and grapefruit signals and so contains the signals of hesperidin and a higher peak in the arginine massif and other variations. Finally, IC3 contains only the peak of ethanol. IC 3 is representative of the fermentation of the juice. Hence it is no surprise than the IC 1- IC 2 plane is giving a better separation (Table 2 Publication / Tableau 22 and Figure 10b Publication / Figure 61b).

On IC 3 the scores of the samples range from –0.0019 to 0.5744. The identification of the juice having a score higher than 0.150, gives 42 samples either NFC or CFC out of 53 and

only 8 FC juices out of 39. Juices sold in the chill department are more likely to ferment and contain ethanol than juices sold in pack.

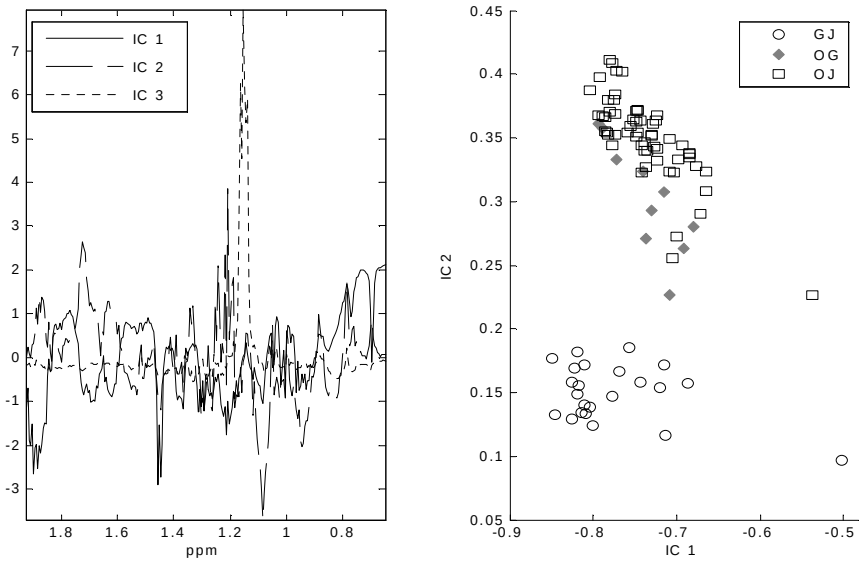


Figure 61 : (Figure 10 Publication) : Results of ICA on the forth zone: a) Loadings; b) Scores on the IC1-IC2 plane

Zone 1 to 4

When all the zones are concatenated together and submitted to ICA, the resulting ICs change (Figure 11 Publication / Figure 62). The comparison of the part of the average spectrum and the "pure" sources found by ICA (Figure 11 Publication / Figure 62) indicates that the negative first independent component (-IC1) resembles the average logarithmic grapefruit spectrum. IC2 reflects the contribution of orange juice. IC3 looks like a compensation of IC1. This IC is the most discriminant one as can be seen in Table 2 Publication / Tableau 22.

The loadings of the ICA are easy to interpret in term of correspondence to the signal. Comparing the 3 loadings of the ICA to the 2 more discriminative loadings of the PCA (Figure 11 Publication / Figure 62), PC1 appears to be similar to IC3. Indeed, PCA usually extracts the most discriminant variable on the first PC. However the third PC, which resembles also a broad version of -IC1, can neither be related to the other ICs, nor to the juice signals.

PCA is here giving its worse result of classification with only 56 correctly classified samples out of 92. On the other hand, ICA still gave good results with 85 correctly classified samples (Table 2 Publication / Tableau 22).

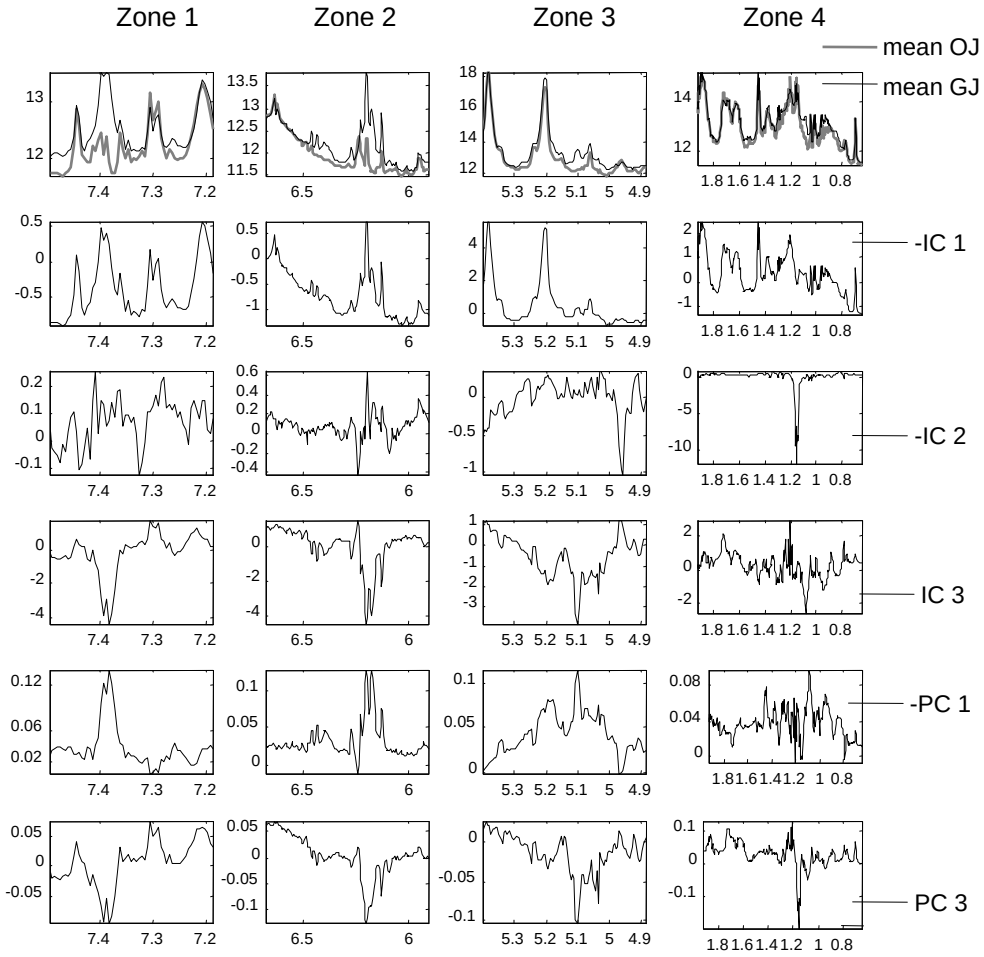


Figure 62 : (Figure 11 Publication) : a) Average logarithmic spectra of orange and grapefruit in the three selected zones; b) The 3 Independent Components loadings for the four concatenated zones ; c) The 2 more discriminant Principal Components loadings for the four concatenated zones

Conclusion

The warping and log-transform pre-treatments used were able to correct the uncontrolled variations due to drifts in chemical shifts and the large difference in the concentration of the juice constituents, respectively.

The EWZS function with the ICA method detected four zones that are discriminant for the juices but only 3 of them with the PCA option. In addition, the linear regression was more significant with the ICA option, as the R^2 values were higher.

The EWZS function was able to identify the well-known markers naringin and hesperidin, which are commonly used in juice authentication using the standard HPLC method. It is in fact the zone (Zone 1) containing those markers that gave the best rate of classification. Other spectral features were also detected, such as higher free α-glucose in grapefruit juice and the presence of unassigned compounds in the 4.85-5.00 ppm area in orange juice, indicating that the NMR technique combined with the EWZS function may be useful to discover new authenticity markers in different matrices.

For the same set of selected zones, the classification rates using ICA were better than with PCA, indicating that ICA may be a better method to extract useful information from complex signals, such as ^{1}H NMR spectra of mixtures. This is confirmed by the fact that the "pure sources" found by performing ICA on the selected zones correspond to the average orange juice signal and the variation in this signal due to the addition of grapefruit juice. This made the interpretation of the results easier than when using the loadings of the PCA.
Suitable chemometric treatment helps the interpretation of how complex ^{1}H NMR data are able to discriminate between different samples of interest: firstly, by selecting variables of interest over the whole spectrum rather than being limited to areas of specific classes of compounds; and secondly by relating the selected information to prior knowledge of the compounds or products under investigation to improve understanding of sample composition and the reasons for the discrimination.

References

1. Belton, P. S., Delgadillo, I., Holmes, E., Nicholls, A., Nicholson, J. K. & Spraul, M. Use of high-field ^{1}H NMR spectroscopy for the analysis of liquid foods. *J.Agric. Food Chem.*, 1483-1487 (1996).
2. Christophoridou, S., Dais, P., Tseng, L.-H. & Spraul, M. Separation and identification of phenolic compounds in olive oil by coupling high-performance liquid chromatography with postcolumn solid-phase extraction to nuclear magnetic resonance spectroscopy (LC-SPE-NMR). *J. Agric. Food Chem.* **53**, 4667-4679 (2005).
3. Colquhoun, I. J. High resolution NMR spectroscopy in food analysis and authentication. *Spectroscopy Europe*, 42583 (1998).

4. Consonni, R. & Gatti, A. ^{1}H NMR studies on Italian balsamic and traditional balsamic vinegars. *J. Agric. Food Chem.* **52**, 3446-3450 (2004).
5. Duarte, C. A., Barros, A., Belton, P. S., Righelato, R., Spraul, M., Humpfer, E. & Gil, A. M. High-resolution NMR spectroscopy and multivariate analysis for the characterization of beer. *J. Agric. Food Chem.* **50**, 2475-2481 (2002).
6. Fauhl, C., Reniero, F. & Guillou, C. ^{1}H NMR as a tool for the analysis of mixtures of virgin olive oil with oils of different botanic origin. *Magnetic Resonance in Chemistry* **38**, 436-443 (2000).
7. Fragaki, G., Spyros, A., Siragakis, G., Salivaras, E. & Dais, P. Detection of extra virgin olive oil adulteration with lampante olive oil and refined olive oil using NMR spectroscopy and multivariate analysis. *J. Agric. Food Chem.* **53**, 2810-2816 (2005).
8. Mannina, L. & Segre, A. High-resolution NMR: from chemical structure to food authenticity. *Grasas y Aceites* **53**, 22-33 (2002).
9. IFU. *58 : Determination of hesperidin and naringin by HPLC*, pp. (1991).
10. Cichocki, A. & Amari, S. (ed. Wiley), (New York, 2002).
11. Hyvarinen, A., Karhunen, J. & Oja, E., (Wiley, New York, 2001).
12. Belton, P. S., Delgadillo, I., Gil, A. M., Roma, P., Casuscelli, F., Colquhoun, I. J., Dennis, M. J. & Spraul, M. High-field proton RMN studies of apple juices. *Magnetic Resonance in Chemistry* **35**, 52-60 (1997).
13. Tomasi, G., van den Berg, F. & Anderson, C. Correlation Optimized Warping and time warping as preprocessing methods for chromatographic data. *J. Chemometrics* **18**, 231-241 (2004).
14. Nielsen, N. P. V., Carstensen, J. M. & Smedsgaard, J. Aligning of single and multiple wavelength chromatographic profiles for chemometric data analysis using correlation optimised warping. *J. Chromatography A* **805**, 17–35. (1998).
15. Astakhov, S. A., Stogbauer, H., Kraskov, A. & Grassberger, P., (2005).
16. Shimizu, S., Hyvarinen, A., Kano, Y. & Hoyer, P. O. Discovery of non-gaussian linear causal models using ICA.
http://www.cs.helsinki.fi/u/ahyvarin/papers/Shimizu05UAI.pdf.
17. Cardoso, J.-F. Blind separation of real signals with JADE.
http://www.tsi.enst.fr/icacentral/Algos/cardoso/ (1995).
18. Cardoso, J.-F. & Souloumiac, A. in *IEE*, 362-370 (1993).
19. Cardoso, J.-F. High-order contrasts for independent component analysis. *Neural Computation* **11**, 157-192 (1999).

20. Jiang, J. H., Berry, R. J., Siesler, H. W. & Ozaki, Y. Wavelength Interval Selection in Multicomponent Spectral Analysis by Moving Window Partial Least-Squares Regression with Applications to Mid-Infrared and Near-Infrared Spectroscopic Data. *Anal. Chem.* **74,**, 3555-3565 (2002).
21. Guthrie, J. A., Walsh, K. B., Reid, D. J. & Liebenberg, C. J. Assessment of internal quality attributes of mandarin fruit. 1. NIR calibration development. *Australian Journal of Agricultural Research* **56**, 405-416 (2005).
22. Le Gall, G., Puaud, M. & Colquhoun, I. J. Discrimination between orange juice and pulp wash by ^{1}H NMR spectroscopy: Identification of marker compounds. *J. Agric. Food Chem.* **49**, 580-588 (2001).
23. Duarte, I. F., Delgadillo, I. & Gil, A. M. Study of natural mango juice spoilage and microbial contamination with *Penicillium expansum* by high-resolution ^{1}H NMR spectroscopy. *Food Chemistry* **96**, 313-324 (2006).
24. Colquhoun, I. J. & Goodfellow, B. J. Chiral recognition by cyclodextrins: the interaction of naringin with-cyclodextrin. *J. Chem. Soc., Perkin Trans 2*, 1803-1807 (1994).
25. Peterson, J. J., Beecher, G. R., Bhagwat, S. A., Dwyer, J. T., Gebhart, S. E., Haytowitz, D. B. & Holden, J. M. Flavonones in grapefruit, lemons, and limes: a compilation and review of the data from the analytical literature. *J. of Food Composition and Analysis* **19**, 74-80 (2006).

4.3. Publication 3

Fruit juice authentication by ^{1}H NMR spectroscopy in combination with different chemometrics tools

Cuny, M., Vigneau, E., Le Gall, G., Colquhoun, I. J., Lees, M. & Rutledge, D. N. (2008). Analytical and Bioanalytical Chemistry **390**: 419-427.

Introduction

Consumers consider fruit and fruit juices as healthy foods and a wide range of public health organisations have promoted them as such. Numerous reports have appeared linking juice consumption to a reduced risk of developing different illnesses such as Alzheimer's disease[1]. Hence, fruit juice consumption is increasing worldwide. To attract consumers, producers tend to be more and more innovative and have been developing the market segment of blends or mixed fruit juices. Juices are valuable commodities and economic frauds in the sector of fruit juice have often been reported [2-10]. One of the most common frauds is to add a co-fruit, a fruit that is less expensive or easier to find, to the authentic juice. For example, grapefruit juice has been reported as being fraudulently added to orange juice in the US [11].

In this study, ways to discriminate orange and grapefruit juices and their blends are evaluated. ^{1}H NMR spectroscopy is a well-known rapid screening method [12] and has been demonstrated to be an efficient method in beverage authentication [13-17]. The information contained in a single spectrum covers a wide range of compounds [18, 19]. The present work focuses on different ways of extracting relevant information from this spectral data.

The treatment of ^{1}H NMR data is not as extensively documented as other spectroscopic techniques [20], and work has still to be done to extract and use all the information a spectrum contains. The work reported here investigates three chemometric tools for the selection of variables containing information relevant to the discrimination of the different types of fruit juice. These include the EWZS (Evolving Window Zone Selection) method, described in more detail elsewhere [21], the CLV (Clustering of Latent Variables) approach [22-25] and a method based on selection using variance. Two different multivariate methods were then used to analyse the data: the commonly used Principal Component Analysis (PCA) [26] and a more recent method, Independent Component Analysis (ICA), usually applied to the processing of complex mixtures of signals [22-25].

Methods

Data collection

Sample collection

The sample set was the same as that described in a previous study [21], details of which are repeated here for the sake of clarity. 92 juices were bought in local supermarkets in Norwich, UK, in July 2005. These included 59 orange juices (OJ), 23 grapefruit juices (GJ), of which 15 were white grapefruit juices and 8 pink grapefruit juices, and 10 blends (OG) of which 2 with pink grapefruit juices. According to the information provided on the label the grapefruit composition in the juice blends varied from 10 to 50 %. One of the samples did not indicate the percentage of grapefruit juice on its label. As will be seen later, one of the NMR markers detected in this study corresponded to a known grapefruit juice constituent, used in the standard HPLC method [27]. The absence of this marker was verified for all the orange juices. As well, 20 of 59 orange juices were also assessed by HPLC. None were found to be adulterated by addition of grapefruit juice.

In addition to the three categories of juice type, the sample set also covered three different manufacturing processes: Not From Concentrate (NFC), Chilled From Concentrate (CFC), and Longlife From Concentrate (LFC). There were 25 NFC orange juices, 9 NFC grapefruit juices and 3 NFC blends. For the CFC juices, there were 12 orange and 4 grapefruit. For the LFC juices, there were 22 orange, 10 grapefruit, and 7 blends. This sample set reflects as much as possible the variability of the available product on the market. For the NMR measurements, the pH of manually homogenised juices (12 to 14 mL) was adjusted to 4.00 +/- 0.03 to limit any fluctuation of the chemical shift of some compounds in the spectra [28]. The pH-adjusted and homogenised juice was added to an Eppendorf (~1.5 mL) and centrifuged at 10 000 rpm for 5 minutes. The supernatant was collected. 550 μL of a sample containing 75 % of the treated supernatant and 25 % of D2O containing TSP (sodium 3-(trimethylsilyl)propionate-2,2,3,3-d4) as internal reference was transferred to a 5 mm NMR tube. No additional treatment was necessary. The D2O was used as the field frequency lock signal, and TSP for internal referencing of ^{1}H chemical shifts.

^{1}H NMR measurements

^{1}H NMR spectra were recorded at 27°C on a 600 MHz BRUKER spectrometer fitted with an auto-sampler. Each spectrum consisted of 128 scans of 16384 complex data points with a spectral width of 7183.9 Hz (11.97 ppm). A presaturation sequence suppressed the residual water signal with low power selective irradiation at the water frequency during the recycle delay. Spectra were Fourier transformed with 1 Hz line broadening, automatically phased and manually baseline corrected using TopSpin® 1.3 software (Bruker BioSpin GmbH, Rheinstetten, Germany).

Data processing

Spectra were converted to J-CAMP6 files and transferred to a personal computer for data analysis using functions written in Matlab® (The MathWorks Inc, Natick, Massachusetts, USA).

Data pre-treatment

Data pre-treatment consisted of base-line correction, peak alignment using Correlation Optimised Warping functions with linear interpolation (COW [29, 30]) (size of segment 120 and slack 20) and data size reduction. The latter was done by taking the mean of 7 adjacent points to improve the signal to noise ratio and reduce the number of data points without significant loss of the information contained in the ^{1}H NMR spectrum. In this way, the initial spectrum of 16 K points was reduced to 2143 variables. The results of the transformation are shown in Figure 1 Publication / Figure 63 . Finally, a logarithmic transformation was applied to the data to reduce the difference in scale between the different parts of the spectrum. This pretreatment is often used to enhance the minor peaks when the dynamic range of the signal is very high. Moreover, peaks with low intensities and high intensities are easily visualised on the same scale. A group membership vector was created containing 1 if the sample is GJ, 2 if the sample is OG and 3 if OJ. Although the relation between the group membership values is not linear, the order of the values is chosen to reflect the intermediate position of the blends (OG).

Chemometric data analysis

Variable Selection

When a lot of variables are collected, as is often the case in chemometrics, especially in spectroscopy, ordinary linear regression gives models of unacceptable predictive ability and difficulties in interpretation. The selection of a subset of the variables is one of the possible ways of improving this problem.. There are many methods used for variables selection such as Stepwise Multiple Regression, methods based on factorial analysis (i.e. Principal Components Analysis [31, 32], Partial Least Squares [33],Projection Pursuit [34], …) or more explanatory methods based for instance on Classification and Regression Trees (CART) [35, 36] or Genetic Algorithms [37-39]. Neural Networks [40] are mainly used in situations of non-linearity while Genetic Algorithms are particularly useful for huge datasets. In our case, there are only 92 samples and no apparent non-linearity. For these reasons, procedures based on more classical linear univariate or multivariate methods were preferred. In this work, three different approaches were investigated : The first one is a very simple approach based on the variance of the variables, the second one is an unsupervised method leading to clusters of variables defined around latent components, the last one is an unsupervised data compression of spectral zones followed by a supervised selection of the most correlated regions.

Selection of variables with high variance (Variance method) – Since the variance of a variable is a measure of the average squared distance between the individual values of the variable for the set of data points and their mean value, this criterion may be able to identify the most discriminating variables in the data set as they are expected to vary more than noisy or non-informative variables. A variance threshold above which variables were retained in the data set, was defined by the ability of the reduced data set to predict membership of the samples to the correct juice groups as reflected in the confusion matrix calculated using the observed and predicted groups (cf. 2.4.3).

Selection of clustered variables (CLV [41, 42] method) – A cluster analysis around latent variables is applied in order to identify groups among the spectral variables, linked either by their covariance or their correlation. Organising multivariate data into a small number of clusters, each represented by a latent component, makes it possible to reduce the dimensionality of the data set. The CLV method involves two stages, namely a hierarchical clustering analysis followed by a partitioning algorithm. Partitioning is determined by the

value of a quality criterion, in this case T, which is the sum of the first eigenvalues of the data matrices of all the clusters. Once the appropriate number of clusters is chosen the discriminative potential of each group of linked variables is assessed. Using analysis of Variance (ANOVA), the components most likely to discriminate between samples in terms of juice type are determined.

Selection of zones of contiguous variables (Evolving Window Zone Selection - EWZS method) – This approach focuses on detecting different contiguous parts of the spectrum that are shown to be most discriminant for the predefined sample groups. EWZS is similar to other zone selection methods such as moving-window PLS [43] and spectral window selection [44, 45] in that it uses windows of varying sizes, from a small user-defined minimum up to a user-defined maximum (which may be as large as the width of the complete data set). The first set of windows starts with the first variable. Once the window has increased to the user-defined maximum limit, the starting point of the new set of windows is incremented by a user-defined step. EWZS differs from the above-mentioned zone selection procedures in that they both use PLS regression to relate the spectra to the dependent variable, whereas EWZS can use a range of multivariate methods. In the present study, both Principal Components Analysis (PCA) and Independent Components Analysis (ICA) were used to decompose the dataset within each window, in order to reduce the dimensionality and concentrate the information in a small number of Principal or Independent Components. A simple linear regression was then calculated between each vector of scores and the vector of percent orange juice. This supervised method facilitates the detection of spectral zones correlated with the dependent variable. For both PCA and ICA, the highest coefficient of determination (R^2) for each region of the spectrum was then plotted as a map. A visual examination of these maps is used to identify those zones showing a maximum R^2 value, which may then be selected as informative contiguous zones of variables.

Independent Components Analysis (ICA).

The ICA approach is based on the assumption of statistical independence of the original ("pure") sources [46-48]. Using this assumption, a range of methods have been developed to recover the pure signals from a data set of mixed signals by finding a demixing transformation that minimizes dependencies between the estimates of the "pure" sources [49]. In this study, considering that in the mixed signals (the set of fruit juice ^{1}H NMR spectra) there are two

pure sources: OJ and GJ and the noise, 3 Independent Components were extracted using the Jade (Joint Approximate Diagonalization of Eigenmatrices) ICA algorithm [50-52].
The "pure" source signal vectors extracted by ICA may be considered as comparable to the loadings vectors obtained by PCA. Similarly, the coordinates of the samples on these ICA "loadings" may be considered as comparable to PCA scores.
The more commonly used PCA is also applied to evaluate whether ICA gave superior classification rates. For PCA model, 3 PCs were also taken into account.

Comparison of Variable Selection Methods: Leave-one-out confusion matrix

To evaluate the efficiency of the different methods used to discriminate samples, a leave-one-out procedure is performed. A confusion matrix [53] containing the actual and predicted group memberships made by each classification procedure was calculated. The classification is based on the Euclidian distances of each sample to the barycentre of scores for each group. A sample is attributed to the group for which the distance to the barycentre is the shortest. For each ICA, all the classification systems made of a combination of the scores on the three loadings were tested. The best model, giving the best result for the confusion matrix, was retained. For the PCA, all the classification models taking into account the scores on the first three principal components were tested, and the best one was retained.
The small number of ICs and PCs being tested made it possible to do an exhaustive comparison of the classification models using all possible combinations of the Components.

Results and discussion

Discrimination of juice types based on the full dataset (14 999 variables) without data reduction and log transformation

To be able to compare the effectiveness of the data pre-treatment and variable selection methods, the dataset consisting of the complete spectra after phasing, base-line correction, suppression of non-informative parts (water peak region and edges), and COW warping (Figure 1.a Publication / Figure 63a) was first tested for discrimination of the groups. Hence the dataset contains 14999 variables. The confusion matrices of classification based on the first three PCs and ICs are presented in Table 1 Publication / Tableau 23.
The dataset gives a reasonable rate, 80.4 %, of correct classification for both procedures.

Tableau 23. (Table 1 Publication) : Summary of the classification results for the data set without data reduction and logarithmic transformation and before variable selection

Number of variables selected			*14999*			
	ICA classification with IC1 & IC2			*PCA classification with PC1, PC2 & PC3*		
	GJ	*OG*	*OJ*	*GJ*	*OG*	*OJ*
GJ	*16*	*7*	*0*	*16*	*7*	*0*
OG	*0*	*8*	*2*	*0*	*8*	*2*
OJ	*0*	*9*	*50*	*0*	*9*	*50*
% of correct classification	*80.4*			*80.4*		

Discrimination of juice types on the dataset with 7-point data reduction and log transformation, but without variable selection

The effect of the data reduction by 7-point averaging, from 14999 to 2413 variables, followed by logarithmic transformation (Figure 1.b Publication / Figure 63b) was then tested. The results are presented in Table 2 Publication / Tableau 24, where it is clear that this pre-treatment has improved the results for ICA but not for PCA.

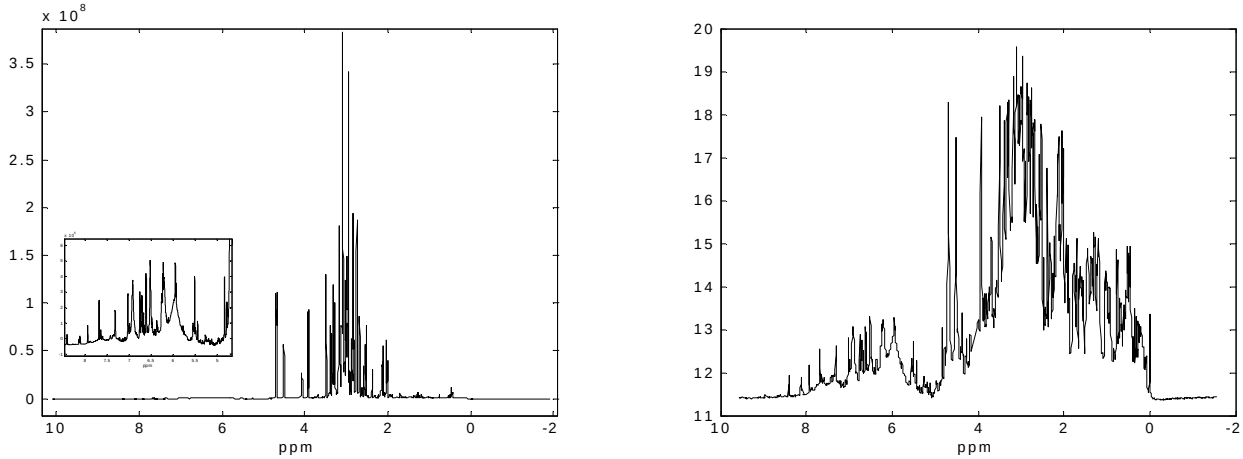


Figure 63 : (Figure 1 Publication) : Mean spectrum for 7-point-averaged : a) ^{1}H NMR spectra of all samples and zoom for the area between 4.5 and 8.5 ppm ; b) data set after logarithmic transformation

Tableau 24. (Table 2 Publication) : Summary of the classification results without variable selection

Number of variables selected			*2143*			
	ICA classification with IC1 & IC3			*PCA classification with PC1, PC2 & PC3*		
	GJ	*OG*	*OJ*	*GJ*	*OG*	*OJ*
GJ	*23*	*0*	*0*	*16*	*7*	*0*
OG	*0*	*8*	*2*	*0*	*7*	*3*
OJ	*0*	*5*	*54*	*0*	*8*	*51*
% of correct classification	*92.4*			*80.4*		

The three different variable selection procedures were then applied to eliminate uninformative variables or noisy variables from the data set as they may reduce the quality of the prediction models.

Discrimination of juice types after selection based on the high variance criterion

The variance of the variables is shown in Figure 2.b Publication / Figure 64b. The best classification results for juice type were obtained for a threshold of 0.02 and for a range of thresholds from 0.03 to 0.045 (Figure 2.a Publication / Figure 64a). As the threshold increases, the number of selected variables decreases, and so to suppress the maximum number of uninformative variables, the highest threshold value of 0.045 was chosen.

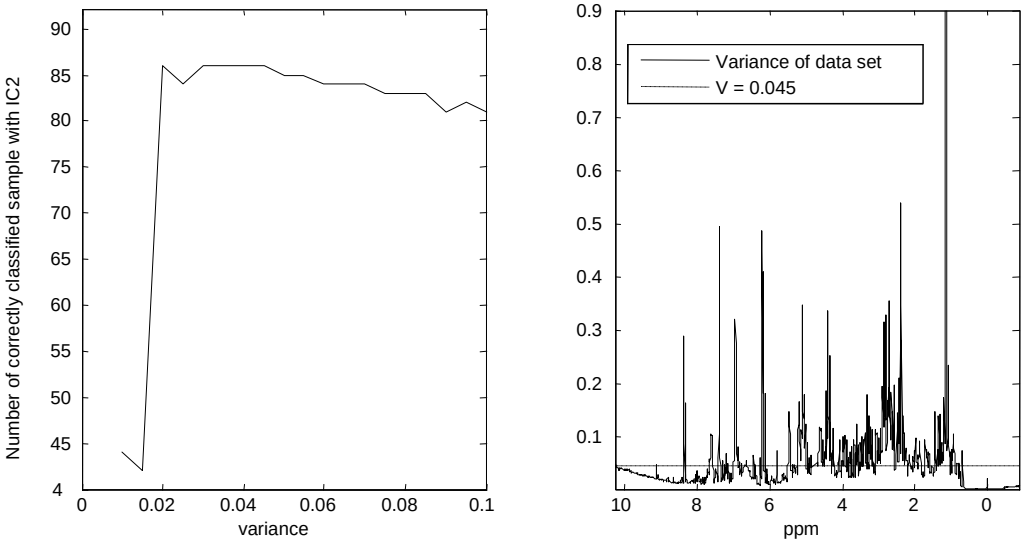


Figure 64 : (Figure 2 Publication) : a) Results of the selection algorithm to define the variance threshold; b) Selected variance threshold on variance data set

The dataset formed with the variables with a variance higher than 0.045 (Figure 5.a Publication / Figure 67a), was submitted to ICA and PCA and the results of classification are shown in Table 3 Publication / Tableau 25.

Both correct classification rates were improved with the selection based on the variance criterion. ICA gives better results on this dataset.

Tableau 25. ***(Table 3 Publication) : Summary of the classification results for variance method***

Number of variables selected			*759*			
	ICA classification with IC2			*PCA classification with PC1 & PC2*		
	GJ	*OG*	*OJ*	*GJ*	*OG*	*OJ*
GJ	*23*	*0*	*0*	*16*	*7*	*0*
OG	*0*	*9*	*1*	*0*	*8*	*2*
OJ	*0*	*4*	*55*	*0*	*7*	*52*
% of correct classification	*94.6*			*82.6*		

Discrimination of juice types after selection based on the CLV method

The results of the hierarchical clustering analysis and partitioning of the 2143 variables are shown in Figure 3 Publication / Figure 65.

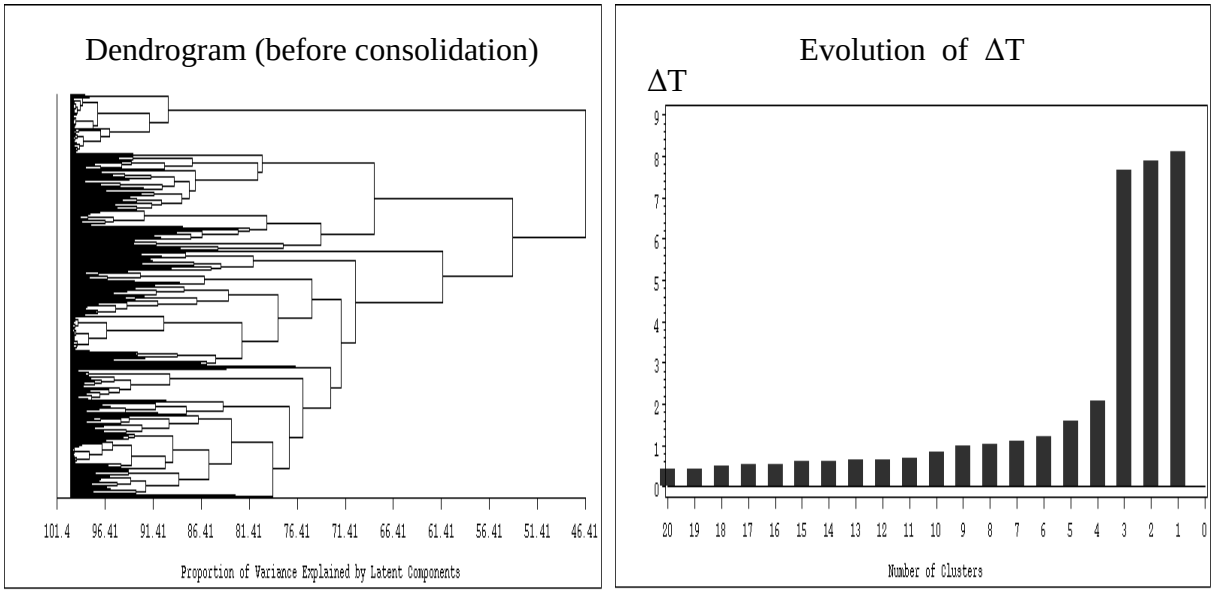


Figure 65 : (Figure 3 publication) : a) Dendrogram of variables; b) Evolution of the criterion ΔT with the number of clusters

The graph of ΔT (Figure 3.b Publication / Figure 65b) observed during the hierarchical clustering showed that when passing from a partitioning of 4 to 3 groups, the T criterion significantly increased indicating that a lot of information has been lost. An ANOVA on the latent variables calculated by PCA on the four groups of the retained partition was used to identify the significant components for explaining juice type (Table 4 Publication / Tableau 26).

The variables belonging to the group represented by the second component are highly significant for explaining the juice type variability. Thus after removing the known non-informative zones or isolated points, corresponding to 231 variables, the remaining variables clustered in the group 2 were submitted to analysis.

Tableau 26. (Table 4 Publication) : ANOVA results for juice type on the four latent components obtained by CLV

Source	Sum of Squares	d.f.	Mean Squares	F	Prob>F
Component 1	0.019	2	0.009	0.85	0.429
Component 2	0.669	2	0.335	90.02	0
Component 3	0.010	2	0.005	0.44	0.645
Component 4	0.014	2	0.007	0.61	0.544

The scores calculated by applying ICA and PCA to these variables are then used to evaluate the ability of the selected variables to classify the samples in their true juice group membership.

Examining the classification results shown in Table 5 Publication / Tableau 27, it is obvious that the correct classification rate for both ICA and PCA has increased with this selection of variables (for ICA: 95.7 % vs. 92.4 % without selection, for PCA: 84.8 % vs. 80.4 % without selection). Moreover the results are also better than those obtained with the variance method (94.6 % for ICA and 82.6 % for PCA).

Tableau 27. (Table 5 Publication) : Summary of the classification results for CLV method

Number of variables selected			*878*			
	ICA classification with IC1, IC2 & IC3			*PCA classification with PC1 to PC3*		
	GJ	*OG*	*OJ*	*GJ*	*OG*	*OJ*
GJ	*23*	*0*	*0*	*16*	*7*	*0*
OG	*0*	*9*	*1*	*0*	*9*	*1*
OJ	*0*	*3*	*56*	*0*	*6*	*53*
% of correct classification	*95.7*			*84.8*		

It is no surprise that this method gives better results, as the choice of a specific group of variables is supervised according to the discriminatory power of its first principal component. Indeed, although the variables are first grouped without using any information on the predefined groups of juices, only those variables associated with the latent variable that is most correlated with the juice groups are then retained to perform the multivariate analyses.

The selected variables (Figure 5.b Publication / Figure 67b) correspond to phenolic compounds in the low field region, among which are the two parts of the naringin signals (between 6.07 and 6.33 ppm and between 7.35 and 7.42 ppm). This component is known to discriminate between OJ and GJ and is in fact the authenticity marker used in the standard HPLC method (IFU 58). The variable selection method also retained a lot of different sugars such as free α-glucose (peaks at 5.23, 3.72, 3.39, 3.42 ppm), and free β-glucose (peaks at 4.64, 3.24, 3.38, 3.90, 3.49 ppm), free fructose (at 4.11 and 3.81 ppm), and other signals in the zones 3.33-3.64, 3.66-4.00, 4.47-5.34, 5.41-5.49. In the aliphatic region, citric acid was selected (2.78 ppm), and several amino acids such as alanine (1.45 ppm) and leucine (0.96 ppm). Table 6 Publication / Tableau 28 shows the content of these components in both juices as given in the AIJN guidelines and in the USDA survey of phenolic compounds observed in *Citrus* [54].

Table 6 Publication / Tableau 28 underlines the fact that, with the exception of naringin, the contents of these different compounds vary over a wide range within each type of juice.

Tableau 28. (Table 6 Publication) : Content of the compounds selected by the CLV method in orange and grapefruit juices

Compound	**Content in orange juice**	**Content in grapefruit juice**
glucose	20-35 g/l	20-50 g/l
fructose	20-35 g/l	20-50 g/l
naringin	0.0 mg/g in Navel and Valencia cultivars (USDA)	maximum 1200 mg/l (AIJN), mean of 0.38mg/g on 9 cultivars (USDA)
alanine	60-205 mg/l	62-180 mg/l
citric acid	6.3-17 g/l	8-20 g/l
leucine	3-15 mg/l	1-10 mg/l

Discrimination of juice types with selection based on the EWZS method

Applying the EWZS procedure to the pre-treated (warped and log-transformed) data set, using the sample type as the criterion to be predicted, gave the R^2 map shown in Figure 4 Publication / Figure 66.

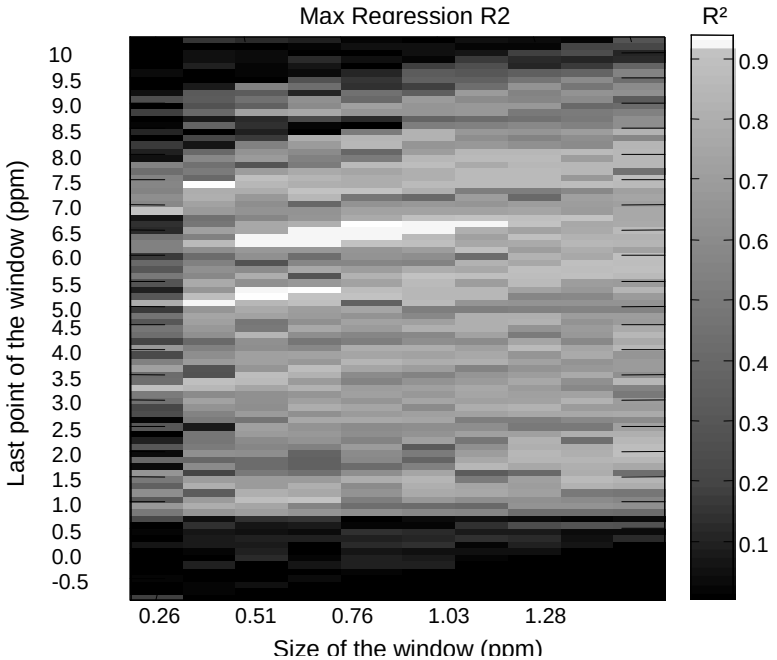
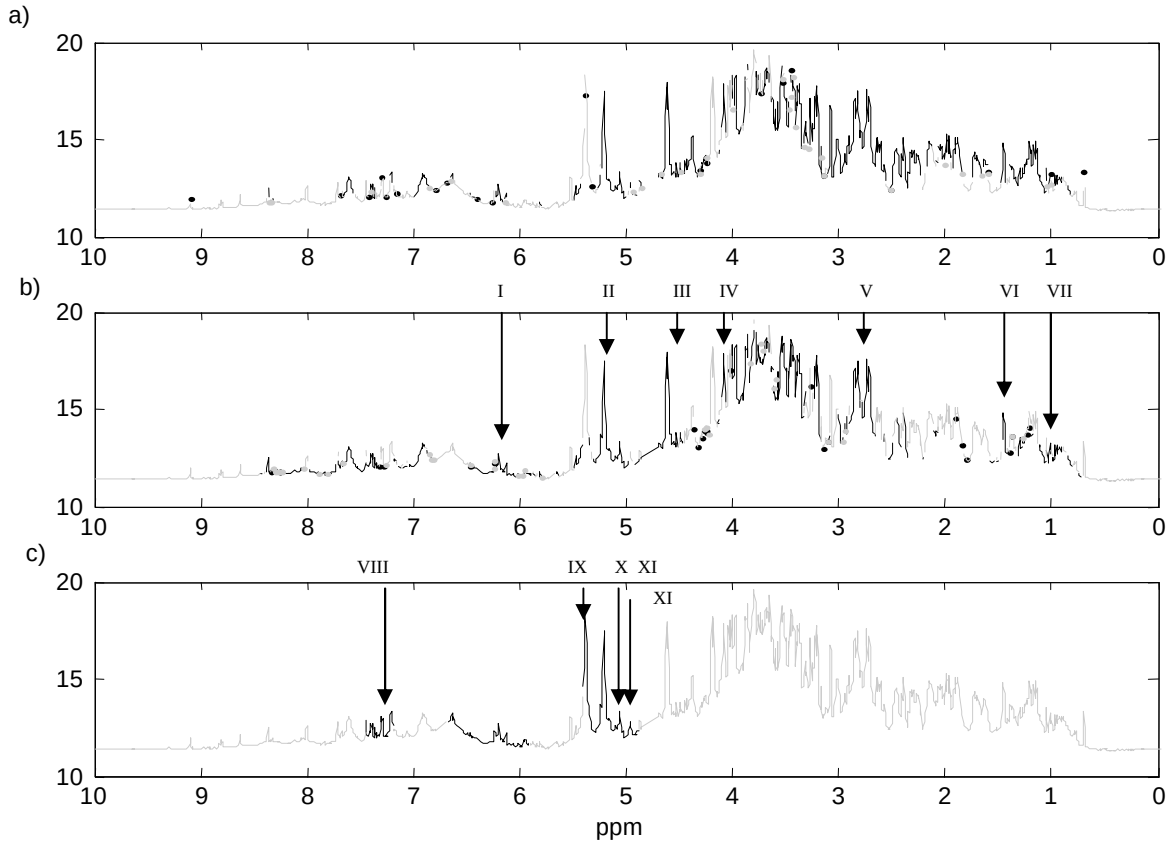


Figure 66 : (Figure 4 Publication) : Map of maximum coefficients of determination (R^2) for varying zone size up to 250 variables wide and initially starting at the first point: prediction of juice type

The minimum window size was set to 7 variables, the maximum window size to 250 variables, while the increment between each series of evolving windows was a step of 25 variables. Since each type of fruit juice may present a large variability on the concentration of individual chemical constituents and therefore individual NMR peaks, it may not be possible to observe strong correlations between individual spectral zones and the separation of juices. However by considering complementary information coming from several weakly correlated zones, it may be possible to discriminate the types of juices. For that reason, all zones with a R^2 above 0.925 were retained. Visual examination of the map shows that three zones of high R^2 can be identified. Variables ranging from 7.19 to 7.44 ppm (Zone 1) along with the zones from 5.91 to 6.68 ppm (Zone 2) and from 4.89 to 5.40 ppm (Zone 3) are the most discriminant for sample type.



I: naringin; II: α-glucose; III: β-glucose; IV: free fructose; V: citric acid; VI: alanine; VII: leucine; VIII: naringin; IX: sucrose; X: naringin; XI: pectic galacturonic acid.

Figure 67 : (Figure 5 Publication) : Variables selected in the log-transformed, 7-point averaged spectrum, using : a) the Variance method; b) the CLV method; c) the EWZS method

Using TOCSY experiments and previously published data [55], signals in zone 1 were assigned as a sucrose peak at 5.39 ppm and an R-glucose peak at 5.20 ppm as well as some smaller signals between 4.85 and 5.15 ppm whose signal at 5.11 ppm may correspond to the to the H-1 of the Rha group in naringin [56]. In the same way and with the help of published data [18, 55], part of the selected variables were attributed to phenolic compounds such as naringin and hesperidin, and amino-acids such as arginine and tyrosine for zones 2 and 3. Hence EWZS automatically found some of the components used in the standard method of orange juice authentication (IFU 58). The dataset corresponding to the values of the samples for the variables in the three zones (Figure 5.c Publication / Figure 67c) were submitted to ICA and PCA.

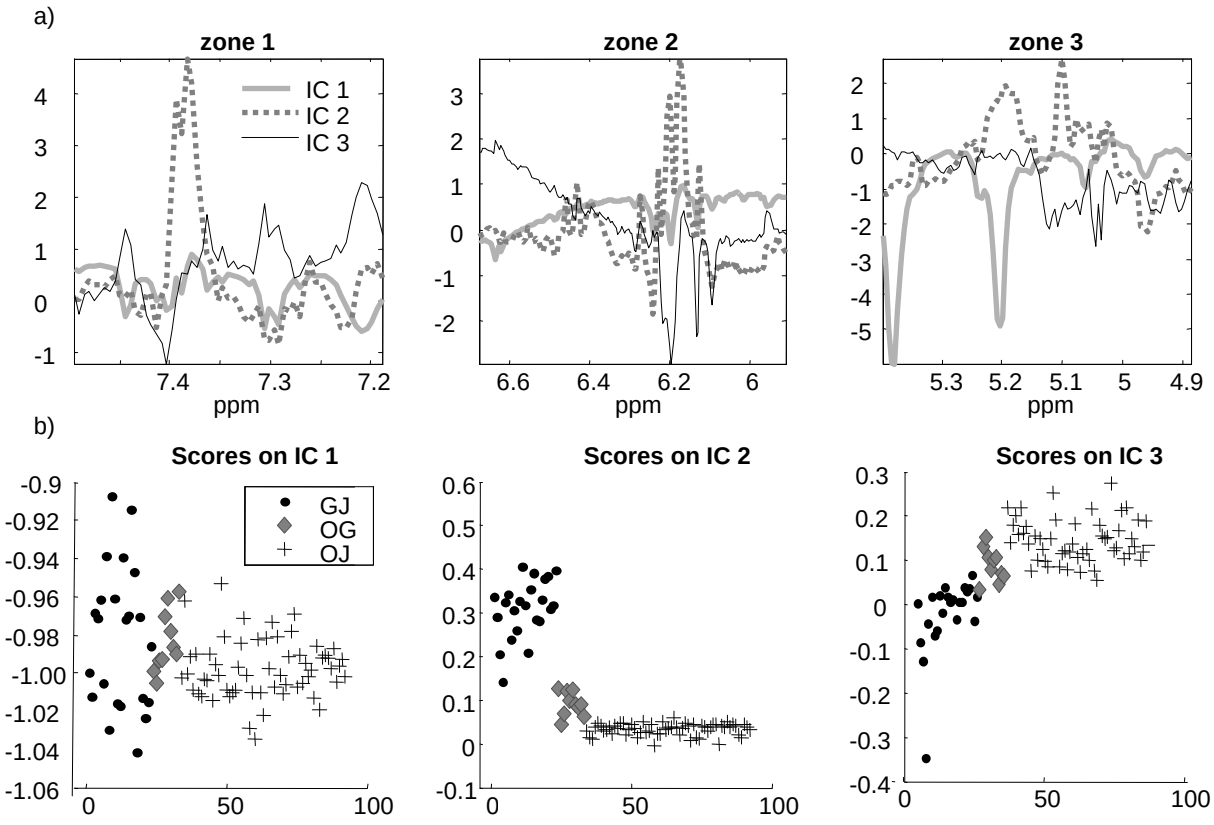


Figure 68 : (Figure 6 Publication) : a) Loadings and b) Scores found applying ICA to the zones selected by the EWZS method

Figure 6 Publication / Figure 68 shows the Loadings (a) and Scores (b) found applying ICA to the zones selected by the EWZS method. The scores of the samples on IC2 (Figure 6b Publication / Figure 68b) enable a clear discrimination between orange and grapefruit juices. Indeed the orange samples have a value around 0. The second IC (Figure 6a Publication / Figure 68a) contains the signals of naringin in all of the 3 zones. Naringin is one of the flavonoid compounds that are markers of adulteration of orange juice by addition of grapefruit juice by the standard HPLC method.

IC 1 is the reverse signal of the average spectrum of orange juices. Therefore the scores on IC 1 for orange samples are more homogeneous than for grapefruit samples. However this component is not discriminative.

IC 3 contains broad signals that look the same as those found in IC 2. They can represent a correction for the variation in shift or ratio of intensity for IC 2. However the discrimination on this axis is not as good as with IC 2.

The selected dataset improves both classification methods with better results for ICA using only IC 2, up to 97.8 % of correct classification (Table 7 Publication / Tableau 29). Only 14 % of the pre-treated dataset was selected.

Tableau 29. ***(Table 7 Publication) : Summary of the classification results for EWZS method***

Number of variables selected			*303*			
	ICA classification with IC 2			*PCA classification with PC 1 and PC 2*		
	GJ	*OG*	*OJ*	*GJ*	*OG*	*OJ*
GJ	*22*	*1*	*0*	*19*	*4*	*0*
OG	*0*	*9*	*1*	*0*	*8*	*2*
OJ	*0*	*0*	*59*	*0*	*6*	*53*
% of correct classification	*97.8*			*87.0*		

Conclusion

With this dataset, ICA consistently gives better results than PCA. Since it is based on the idea of demixing the spectra into a sum of "pure" signals, ICA may be better adapted than PCA, which only determines vectors indicating the direction of greatest dispersion of the samples in the multi-dimensional space of the variables. Calculating the log of the 7-point averages of the intensities is an efficient method of pre-treatment for ^{1}H NMR data for both PCA and ICA.
When selecting significant variables, the results are better, due to the general increase in the signal to noise ratio. Best results were obtained with supervised variable selection procedures, such as in the CLV or the EWZS method. ^{1}H NMR spectroscopy enables a screening of all proton-bearing molecules whereas the standard HPLC method only does a screening for flavonoids. As was seen in this study, ^{1}H NMR spectroscopy was able to detect several other potential markers. It could therefore be used against more subtle potential frauds where the well-known flavonoid marker content has been adjusted. If such a fraud is suspected, several standard methods would be needed which would take more time and possibly cost more than a single ^{1}H NMR spectrum.

References

1. Dai, Q., Borenstein, A. D., Wu, Y., Jackson, J. C. & Larson, E. B. Fruit and Vegetable Juices and Alzheimer's Disease: The Kame Project. *The American Journal of Medicine* **119** (2006).
2. Fry, J., Martin, G. G. & Lees, M. in *Authentification of orange juice*, 1-52 (P.R. Ashurst Blackie Acad., London, 1995).
3. Hammond, D. A. 13c - a refined method to detect the addition of can/corn derived sugars to fruit juices and purees. *Fruit Processing* (1998).
4. Johnston, M. R. & Kauffman, F. L. Orange juice adulteration, detection and the action of the FDA. *J. Food Qual.* **8**, 81-89 (1985).
5. Martin, G. G. SNIF-NMR, a new method to detect added beet sugar to fruit juices and characterize their authenticity. *Fruit Processing*, 123-128 (1992).
6. Martin, G. G., Guillou, C. & Martin, Y. L. SNIF-NMR For Detection of Sugar Addition to Fruit Juices. *Fruit Processing*, 246-254 (1995).
7. Martin, G. G. & Wood, R. G. J. Detection of added beet sugar in concentrated and single strength fruit juices by deuterium nuclear magnetic resonance (SNIF-NMR method) Collaborative study. *Journal of AOAC International*, 917-928 (1996).
8. Patel, T. Real juice or pure fraud ? *New Scientist* **142**, 26-29 (1994).
9. Robards, K. & Antolovich, M. Methods for assessing the authenticity of orange juice. *Analyst* **120**, 1-28 (1995).
10. Wade, R. L. Use of Citrus hybrids in orange juice production. *Fruit Processing*, 358-360 (1995).
11. Blumenthal, D. & Holland, L. in *FDA Consumer*, (1989).
12. Vogels, J. T. W. E., Terwel, L., Tas, A. C., Van Den Berg, F., Dukel, F. & Van Der Greef, J. Detection of adulteration in orange juices by a new screening method using proton NMR spectroscopy in combination with pattern recognition techniques. *J. Agric. Food Chem.* **44**, 175-180 (1996).
13. Fragaki, G., Spyros, A., Siragakis, G., Salivaras, E. & Dais, P. Detection of extra virgin olive oil adulteration with lampante olive oil and refined olive oil using NMR spectroscopy and multivariate analysis. *J. Agric. Food Chem.* **53**, 2810-2816 (2005).
14. Fauhl, C., Reniero, F. & Guillou, C. ^{1}H NMR as a tool for the analysis of mixtures of virgin olive oil with oils of different botanic origin. *Magnetic Resonance in Chemistry* **38**, 436-443 (2000).

15. Brescia, M. A., Caldarola, V., De Giglio, A., Benedetti, D., FanizziI, F. P. & Sacco, A. Characterization of the geographical origin of Italian red wines based on traditional and NMR spectrometric determinations. *Analytica Chimica Acta* **458**, 177-186 (2002).
16. Belton, P. S., Delgadillo, I., Holmes, E., Nicholls, A., Nicholson, J. K. & Spraul, M. Use of high-field ^{1}H NMR spectroscopy for the analysis of liquid foods. *J.Agric. Food Chem.*, 1483-1487 (1996).
17. Colquhoun, I. J. High resolution NMR spectroscopy in food analysis and authentication. *Spectroscopy Europe*, 42583 (1998).
18. Le Gall, G., Puaud, M. & Colquhoun, I. J. Discrimination between orange juice and pulp wash by ^{1}H NMR spectroscopy: Identification of marker compounds. *J. Agric. Food Chem.* **49**, 580-588 (2001).
19. Gil, A. M., Duarte, C. A., Godejohann, M., Braumann, U., Maraschin, M. & Spraul, M. Characterization of aromatic composition of some liquid foods by nuclear magnetic resonance spectrometry and liquid chromatography with nuclear magnetic resonance and mass spectrometric detection. *Analytica Chimica Acta* **488**, 35-51 (2003).
20. Craig, A., Cloarec, O., Holmes, E., Nicholson, J. K. & Lindon, J. C. Scaling and normalization effects in NMR spectroscopy metabonomic data sets. *Anal. Chem.* **78**, 2262-2267 (2006).
21. Cuny, M., Le Gall, G., Colquhoun, I. J., Lees, M. & Rutledge, D. N. Evolving Window Zone Selection method followed by Independent Component Analysis as useful chemometric tools to discriminate between grapefruit juice, orange juice and blends. *Analytica Chimica Acta* **597**, 203-213 (2007).
22. Makeig, S., Jung, T.-P., Ghahremani, D., Bell, A. J. & Sejnowski, T. J. in *Natl. Acad. Sci.*, 10979-10984 (USA, 1997).
23. Vigario, R. Extraction of ocular artefacts from EEG using independent component analysis. *Electroenceph. Clin. Neurophysiol.* **103**, 395-404 (1997).
24. Vigario, R., Särelä, R. & Oja, E. in *International Conference on Artificial Neural Networks*, 287-292 (Skövde, Sweden, 1998).
25. Hyvarinen, A. & Oja, E. Independent component analysis: algorithms and applications. *Neural Networks* **13**, 411-430 (2000).
26. Malinowski, E. R. *Factor Analysis in Chemistry*, pp. (Wiley, New York, 1991).
27. IFU 58 : *Determination of hesperidin and naringin by HPLC* (1991).

28. Belton, P. S., Delgadillo, I., Gil, A. M., Roma, P., Casuscelli, F., Colquhoun, I. J., Dennis, M. J. & Spraul, M. High-field proton RMN studies of apple juices. *Magnetic Resonance in Chemistry* **35**, 52-60 (1997).
29. Tomasi, G., van den Berg, F. & Anderson, C. Correlation Optimized Warping and time warping as preprocessing methods for chromatographic data. *J. Chemometrics* **18**, 231-241 (2004).
30. Nielsen, N. P. V., Carstensen, J. M. & Smedsgaard, J. Aligning of single and multiple wavelength chromatographic profiles for chemometric data analysis using correlation optimised warping. *J. Chromatography A* **805**, 17–35. (1998).
31. Guo, Q., Wu, W., Massart, D. L., Boucon, C. & de Jong, S. Feature selection in principal components analysis of analytical data. *Chemom. Intellig. Lab. Syst.* **61**, 123-132 (2002).
32. Krzanowski, W. J. Selection of variables to preserve multivariate data structure, using Principal Components. *Appl. Statist.* **36**, 22-33 (1987).
33. Hoskuldsson, A. Variable and subset selection in PLS regression. *Chemom. Intellig. Lab. Syst.* **55**, 23-38 (2001).
34. Guo, Q., Wu, W., Massart, D. L., Boucon, C. & de Jong, S. Feature selection in sequential projection pursuit. *Analytica Chimica Acta* **446**, 85-96 (2001).
35. Questier, F., Put, R., Coomans, D., Walczak, B. & Vander Heyden, Y. The use of CART and multivariate regression trees for supervised and unsupervised feature selection. *Chemom. Intellig. Lab. Syst.* **76**, 45-54 (2005).
36. Breiman, L., Friedman, J. H., Olshen, R. A. & Stone, C. J. in *Wadsworth International*, (Monterey, CA, USA, 1984).
37. Abrahamsson, C., Johansson, J., Sparen, A. & Lindgren, F. Comparison of different variable selection methods conducted on NIR transmission measurements on intact tablets. *Chemom. Intellig. Lab. Syst.* **69**, 3-12 (2003).
38. Chtioui, Y., Bertrand, Y., D. & Barba, D. Feature selection by a genetic algorithm. Application to seed discrimination by artificial vision. *J Sience of. Food and Agriculture* **76**, 77-86 (1998).
39. Llobet, E., Gualdron, O., Vinaixa, M., El-Barbri, N., Brezmes, J., Vilanova, X., B. Bouchikhi, Gomez, R., Carrasco, J. A. & Correig, X. Efficient feature selection for mass spectrometry based electronic nose applications. *Chemom. Intellig. Lab. Syst.* **55**, 253-261 (2007).

40. Buydens, L. M. C. & Schoenmakers, P. J. *Intelligent software for chemical analysis*, pp. (Elsevier, Amsterdam, 1993).
41. Vigneau, E. & Qannari, E. M. Clustering of variables around latent components. *Communications in statistics* **32**, 1131-1150 (2003).
42. Vigneau, E., Sahmer, K., Qannari, E. M. & Bertrand, D. Clustering of variables to analyse spectral data. *J. Chemometrics* **19**, 122-128 (2005).
43. Jiang, J. H., Berry, R. J., Siesler, H. W. & Ozaki, Y. Wavelength Interval Selection in Multicomponent Spectral Analysis by Moving Window Partial Least-Squares Regression with Applications to Mid-Infrared and Near-Infrared Spectroscopic Data. *Anal. Chem.* **74,**, 3555-3565 (2002).
44. Guthrie, J. A., Walsh, K. B., Reid, D. J. & Liebenberg, C. J. Assessment of internal quality attributes of mandarin fruit. 1. NIR calibration development. *Australian Journal of Agricultural Research* **56**, 405-416 (2005).
45. Guthrie, J. A., Liebenberg, C. J. & Walsh, K. B. NIR model development and robustness in prediction of melon fruit total soluble solids. *Australian Journal of Agricultural Research* **57**, 1-8 (2006).
46. Cichocki, A. & Amari, S. (ed. Wiley), (New York, 2002).
47. Hyvarinen, A. Survey on independent component analysis. http://www.cs.utexas.edu/~kuipers/readings/Hyvarinen-ncs-99.pdf (1999).
48. Hyvarinen, A., Hoyer, P. O. & Inki, M. Topographic independent component analysis. *Neural Computation* **13**, 1527-1558 (2001).
49. Astakhov, S. A., Stogbauer, H., Kraskov, A. & Grassberger, P., (2005).
50. Cardoso, J.-F. La calibration d'antenne : Identification aveugle aux ordres supérieurs. *Traitement du Signal* **10**, 483-490 (1993).
51. Cardoso, J.-F. Blind separation of real signals with JADE. http://www.tsi.enst.fr/icacentral/Algos/cardoso/ (1995).
52. Cardoso, J.-F. High-order contrasts for independent component analysis. *Neural Computation* **11**, 157-192 (1999).
53. Kohovi, R. & Provost, F. Glossary of terms. *Machine Learning* **30**, 271-274 (1998).
54. USDA. USDA Nutrient database. http://www.nal.usda.gov/fnic/foodcomp/search/, 14/01/2008.
55. Duarte, I. F., Delgadillo, I. & Gil, A. M. Study of natural mango juice spoilage and microbial contamination with *Penicillium expansum* by high-resolution ^{1}H NMR spectroscopy. *Food Chemistry* **96**, 313-324 (2006).

56. Colquhoun, I. J. & Goodfellow, B. J. Chiral recognition by cyclodextrins: the interaction of naringin with-cyclodextrin. *J. Chem. Soc., Perkin Trans 2*, 1803-1807 (1994).

4.4. Publication 4

Independent Component Analysis (ICA) of ^{1}H NMR spectra as a procedure for discrimination and prediction of balsamic vinegars production methods

Cuny, M., Caligiani, A., Palla, G., Lees, M. & Rutledge, D. N. (2008). Submitted to Analytical and Bioanalytical Chemistry.

Balsamic vinegars are typical Italian dressings obtained by alcoholic and acetic fermentation of cooked and concentrated must of white grapes. Traditional balsamic vinegar (TB) is a typical product of the Modena and Reggio Emilia area (provinces of North Italy), since the XI century. In 2000, thanks to its typical production procedures and the well-defined geographical areas of production, it has been certified by a Protected Denomination of Origin (PDO) from the European Union (EU) [1]. TB is obtained using a traditional household preparation, handed down from generation to generation, according to a "disciplinary of production": the only raw material permitted is cooked must, which has to be produced by grapes from the vines traditionally cultivated in the provinces of Modena and Reggio Emilia, in particular from Trebbiano grape varieties. The product has to be aged at least 12 years in barrels of different kinds of wood and of decreasing volume. During ageing, vinegar is transferred periodically from one barrel to another, in a decreasing progression. This procedure is called "topping up". The balsamic vinegar ripening is due to a series of biochemical reactions that are for the most part as yet unknown. The result is the production of a complex mixture of compounds responsible for balsamic vinegar's typical flavour. In this way a valuable product of unique taste is obtained: the product aged at least 12 years is called ''affinato'', whereas ''extravecchio'' is aged at least 25 years and can reach a price of 800 euros/litre. Until now the product suitable for marketing is qualified only by sensory examination [2].

On the contrary, balsamic vinegar (Ba) is a large-scale and industrial product, obtained by adding wine vinegar and caramel (less than 2 % for colour correction) to cooked and concentrated musts. Ba is maturated in wooden recipients for 60 days, although good quality products, aged in wooden barrels for at least 3 years can be found on the market, at a fairly low price (60 euros/litre). The highest quality balsamic vinegars have some organoleptic characteristics (colour, density, base taste etc.) that make it similar to the traditional ones, making food frauds possible.

Many studies have been carried out to distinguish TB from Ba and to detect molecular marker of ageing [3-7]. As NMR has become a popular and effective screening method, recent studies investigated the 1H NMR spectra of vinegars to quantify some selected substances [8] and to characterize balsamic vinegars using Principal Component Analysis [9].

In this study, we used a profiling technique, based on chemometric tools, to select zones in the ^{1}H spectrum that discriminate the samples as a function of production method. This technique combines a methodology of variable selection, called Evolving Window Zone Selection, with Independent Component Analysis. This method is compared to another zone selection method, iPLS and a classification method, PLS. Both models are tested in two steps of discrimination and prediction of vinegar type: the first step tests the robustness of the model by leave-one-out cross-validation on 10 randomly chosen calibration sets of 20 samples. Then the accuracy of the prediction is assessed using the 10 correspondent validation sets of 9 samples.

EXPERIMENTAL

Sample collection.

29 samples of balsamic vinegars were collected of which 15 were traditional coming directly from 4 different Italian producers of the Modena area labelled (TB01- TB15).

Of the non-traditional balsamic vinegars, 7 samples (Ba01-Ba07) were purchased from the market; the others came from 2 vinegar producers of the Emilia Romagna region labelled (Ba08 and Ba14).

A group membership vector was created to predict the type of vinegar with the value 0 for Ba and 1 for TB.

Sample preparation for ^{1}H NMR.

A variable quantity of sample was weighed in order to obtain a final vinegar solution (1 ml of final volume) of about 5% of soluble substances. The weighed vinegar was added to 0.1 ml of 3% TSP (trimethylsilyl [2,2,3,3-$^{2}H_4$]propionate) standard solution, 0.1 ml of deuterated water and diluted to 1 ml of final volume with distilled water. TSP gave two multiplets at about 0.8 ppm and 2.3 ppm that were removed before statistical analysis.

^{1}H NMR conditions.

Spectra were recorded on a VARIAN INOVA-600 MHz spectrometer, operating at 14.1 T, equipped with a 5 mm-triple resonance inverse probe.

Mono dimensional 1H NMR spectra were acquired with low power selective water signal irradiation during 5 s of the relaxation delay.

Data were collected at 298 K, with automatic gradient shimming, without sample rotation, with 32K complex points, using a 45 ° pulse length. 32 scans were acquired with a spectral width of 9650 Hz, an acquisition time of 1.4 s and a recycle delay of 5 s.

Signal Pre-treatment.

The phase and initial baseline corrections were done by hand using MestReC software version 4.5.1. All other signal pretreatments were performed using Matlab routines (Matlab® version 7.14). These pretreatments were base-line correction and alignment of the spectra by Correlation Optimised Warping (COW) [10]. Uninformative peaks such as the water peak (from 4.71 to 4.99 ppm) and the TSP peaks (from 0.00 to 0.86 ppm and from 2.30 to 2.43 ppm) were deleted. In addition, an averaging of 5 consecutive points was used to reduce the initial amount of 32K to 3460 data points at the end of the process. Last the data were submitted to a 10-base logarithmic transformation to reduce the differences in the dynamic range of the high, medium and low fields regions. The final dataset is presented in Figure 1 Publication / Figure 69.

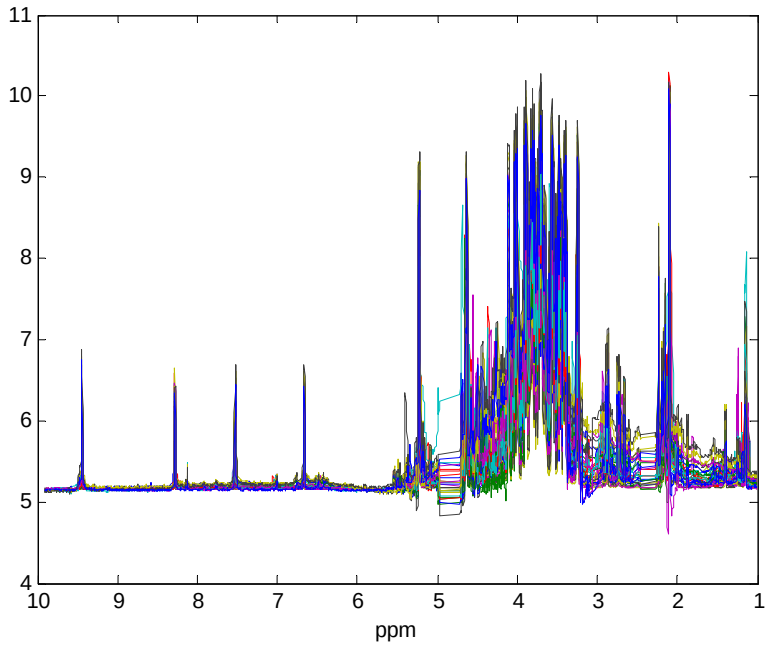


Figure 69 : (Figure 1 Publication) : Dataset of vinegar spectra after all pre-treatment

Reference method.

Interval-Partial Least Squares (iPLS) algorithm.

iPLS is a known method of variable selection [11, 12]. This method applies the PLS algorithm over small parts of the spectra. The user defines a size of interval to study. Within each of them, a PLS model is calculated with an increasing number of components until it reaches the maximum that is defined by the user. The reference model uses the entire dataset to predict the group memberships of the samples. The number of latent variables to use in this model is determined by leave-one-out cross-validation. Then it will be possible to compare the root mean square error of cross-validation (RMSECV) of the reference model and the corresponding value obtained for each interval, using a varying number of components.
The intervals showing a value better or near the RMSECV of the reference model are retained.

PLS.

A PLS model is used to describe the relation between two matrices **X** (the spectra) and **Y** (the group membership). The model tries to explain the maximum multidimensional variance direction in the **Y** space thanks to a linear combination of the variables in **X** - in the present case, the NMR chemical shifts.

If **Y** is a matrix of binary values of group membership, where 0= non-member and 1=member, then PLS can be used for Discriminant Analysis (PLS-DA).

Once a PLS model with **k** components is calculated (Equation 1) it is possible to predict the value of a new sample.

$$\mathbf{X} = \mathbf{T_k}\mathbf{P_k} + \mathbf{E_k} \text{ and } \mathbf{Y} = \mathbf{T_k}\mathbf{Q_k} + \mathbf{f_k} \qquad \textbf{Equation 1}$$

where :

$\mathbf{T_k}$: matrix of factorial coordinates (‘scores’) (n × k)

$\mathbf{P_k}$: loadings (k × m) associated to the prediction of **X** from **T**, the PLS latent variables

$\mathbf{E_k}$: residual matrix associated to X prediction (n × m)

$\mathbf{Q_k}$: loadings (k × p) associated to **Y** prediction from $\mathbf{T_k}$

$\mathbf{f_k}$: residual matrix associated to **y** prediction (n × p)

Indeed, for a new sample $\mathbf{x_{(1,m)}}$ the new scores $\mathbf{T_{(1,k)}}$ are calculated and the predicted value of $\mathbf{Y}$, $\hat{\mathbf{Y}}$ can be assessed (Equation 2).

$$\mathbf{x_{(1,m)}} = \mathbf{T_{(1,k)k}}\mathbf{P_k} \text{ and } \hat{\mathbf{Y}} = \mathbf{T_{(1,k)k}}\mathbf{Q_k} \qquad \textbf{Equation 2}$$

EWZS-ICA model.

Selection of variables defining zones of contiguous variables (Evolving Window Zone Selection - EWZS method).

This approach focuses on detecting different contiguous parts of the spectrum that are shown to be most discriminant for the predefined sample groups. EWZS is similar to Moving-Window PLS [13] and Spectral Window Selection [14] in that it uses windows of varying sizes, from a small user-defined minimum up to a user-defined maximum (which may be as large as the width of the complete data set). The first set of windows starts with the first variable. Once the window has increased to the user-defined maximum limit, the starting point of the new set of windows is incremented by a user-defined step. However EWZS differs from these other window selection procedures in that they both use PLS regression to relate the spectra to the dependent variable, whereas EWZS can use a range of multivariate and univariate methods. In the present study, ICA was used to decompose the dataset within each window. The function then plots a map of the correlation coefficients (R^2) for each region of the spectrum. The R^2 values are obtained on each window, from a PLS-DA applied on a vector of ICA coordinates. A visual examination of this map is used to identify those zones showing a maximum R^2 value. The R^2 values are then identified. Informative contiguous zones of variables may be selected based on these values. Then the zones can be looked at either separately or jointly.

Independent Component Analysis (ICA).

The ICA approach is based on the assumption that original ("pure") signal sources are statistically independent [15-17]. By finding a demixing transformation that minimizes dependencies between the estimates of these "pure" sources [18], independent components, which are often uncorrupted or "pure" signals or noise sources [19], can be recovered from a data set of mixed signals. Here, the JADE (Joint Approximate Diagonalization of

Eigenmatrices) ICA algorithm [20-23] was applied to the set of balsamic vinegars ^{1}H NMR spectra.

The "pure" source signals extracted by ICA may be considered as comparable to the loadings obtained by PCA (Principal Component Analysis). Similarly, the coordinates of the samples on these ICA "loadings" may be considered as comparable to PCA scores. As we have already shown in two previous studies [24, 25] that ICA was better adapted than PCA for the study of spectrum we will not present this work here.

In order to find the best ICA model for the prediction of the vinegar production method using the selected parts of the spectrum, different numbers of ICs were tested (from 2 to 6). The best results were obtained with 3 ICs. Only these results will be shown.

Once an ICA model is defined it is possible to calculate the coordinates of a new sample using equation 3 :

$$B_{(1,H)} = X_{(1,N)} * S_{(N,H)} * \text{inv}(S_{(N,H)}' * S_{(N,H)}) \qquad \textbf{Equation 3}$$

where $B_{(1,H)}$ is the vector of the sample coordinates, N is the number of variables, H is the selected number of sources (here H = 3);

$X_{(1,N)}$ is the vector of coordinates of the sample in the N-variable base; $S_{(N,H)}$ is the matrix of the coordinates of the H pure sources in the N-variable base.

The prediction of the membership of a sample is calculated using the Mahalanobis distance between the sample and the barycentre of the two types of vinegar (Ba and TB). The sample is attributed to the closer group.

As the ICA model taken into account contains 3 ICs, there are 7 different spaces, 3 with one component (IC1, IC2, IC3), 3 with two components (IC1-IC2, IC1-IC3, IC2-IC3), 1 with the three components. All of these spaces are tested. Hence each sample will be classified 7 times for the EWZS/ICA method.

Validation and comparison of the methods

After determining the optimal model dimensionality by cross-validation on a subset of 20 samples, the selected optimal model was validated using the remaining 9 test samples.

In order to avoid having one method giving apparently better performance than the other simply as a result of a particular partitioning of the samples into calibration- and test- sets, the 29 samples were randomly partitioned 10 times and the average of the correct prediction rates used. The composition of the partitions is shown in Table 1 Publication / Tableau 30.

Tableau 30. (Table 1 Publication) : Composition of the 10 partitions of the 29 samples

	1	2	3	4	5	6	7	8	9	10
CAL	Ba01	Ba01	Ba03	Ba01	Ba03	Ba01	Ba01	Ba01	Ba01	Ba02
	Ba02	Ba02	Ba05	Ba02	Ba04	Ba02	Ba02	Ba02	Ba02	Ba03
	Ba04	Ba03	Ba06	Ba03	Ba05	Ba04	Ba03	Ba03	Ba03	Ba04
	Ba05	Ba04	Ba07	Ba04	Ba06	Ba05	Ba05	Ba04	Ba04	Ba06
	Ba06	Ba07	Ba08	Ba05	Ba08	Ba06	Ba06	Ba07	Ba07	Ba08
	Ba07	Ba08	Ba09	Ba06	Ba12	Ba07	Ba09	Ba08	Ba08	Ba09
	Ba08	Ba09	Ba10	Ba08	Ba14	Ba08	Ba10	Ba10	Ba09	Ba11
	Ba10	Ba13	Ba11	Ba11	TB02	Ba10	Ba11	Ba11	Ba11	Ba12
	Ba11	Ba14	Ba13	Ba12	TB03	Ba11	Ba12	Ba14	Ba12	Ba13
	Ba13	TB01	Ba14	Ba14	TB04	Ba12	Ba13	TB01	Ba13	Ba14
	Ba14	TB03	TB02	TB01	TB05	Ba13	Ba14	TB02	Ba14	TB01
	TB01	TB05	TB03	TB02	TB06	Ba14	TB01	TB03	TB01	TB03
	TB02	TB06	TB04	TB03	TB07	TB01	TB03	TB04	TB02	TB05
	TB06	TB07	TB05	TB05	TB09	TB03	TB04	TB05	TB03	TB06
	TB07	TB08	TB06	TB06	TB10	TB05	TB06	TB07	TB06	TB07
	TB09	TB09	TB11	TB07	TB11	TB08	TB08	TB08	TB07	TB09
	TB10	TB10	TB12	TB09	TB12	TB09	TB09	TB10	TB08	TB11
	TB11	TB12	TB13	TB10	TB13	TB10	TB11	TB13	TB09	TB12
	TB12	TB13	TB14	TB14	TB14	TB11	TB12	TB14	TB12	TB13
	TB13	TB15	TB15	TB15	TB15	TB12	TB14	TB15	TB15	TB14
VAL	Ba03	Ba05	Ba01	Ba07	Ba01	Ba03	Ba04	Ba05	Ba05	Ba01
	Ba09	Ba06	Ba02	Ba09	Ba02	Ba09	Ba07	Ba06	Ba06	Ba05
	Ba12	Ba10	Ba04	Ba10	Ba07	TB02	Ba08	Ba09	Ba10	Ba07
	TB03	Ba11	Ba12	Ba13	Ba09	TB04	TB02	Ba12	TB04	Ba10
	TB04	Ba12	TB01	TB04	Ba10	TB06	TB05	Ba13	TB05	TB02
	TB05	TB02	TB07	TB08	Ba11	TB07	TB07	TB06	TB10	TB04
	TB08	TB04	TB08	TB11	Ba13	TB13	TB10	TB09	TB11	TB08
	TB14	TB11	TB09	TB12	TB01	TB14	TB13	TB11	TB13	TB10
	TB15	TB14	TB10	TB13	TB08	TB15	TB15	TB12	TB14	TB15

The iPLS/PLS-DA and EWZS/ICA modelling is performed in two steps.

During the calibration step, the set of 20 samples is used to calculate predictive models of the group membership. To minimise the influence of a particular partitioning of the samples into calibration- and test- sets, the optimal dimensionality for each of the 10 sets of 20 samples is determined by leave-one-out cross-validation.

The validation step uses the optimal iPLS/PLS-DA or EWZS/ICA model for each set of 20 samples, to predict the membership of the 9 samples of the corresponding validation set. Confusion matrices are calculated for both types of model for the 10 partitions to compare the predicted and true membership values.

RESULTS AND DISCUSSION

Selection of variables.

Selection of variables with iPLS.

To be sure to have an entire signal in each interval, the size of an interval was set to 99 variables, and so the spectrum is divided into 35 intervals.

The maximum number of latent variables for an interval was set to 8.

As indicated above, the optimum number of latent variables for the reference model was determined by leave-one-out cross-validation, based on the RMSECV value (Figure 2 Publication / Figure 70).

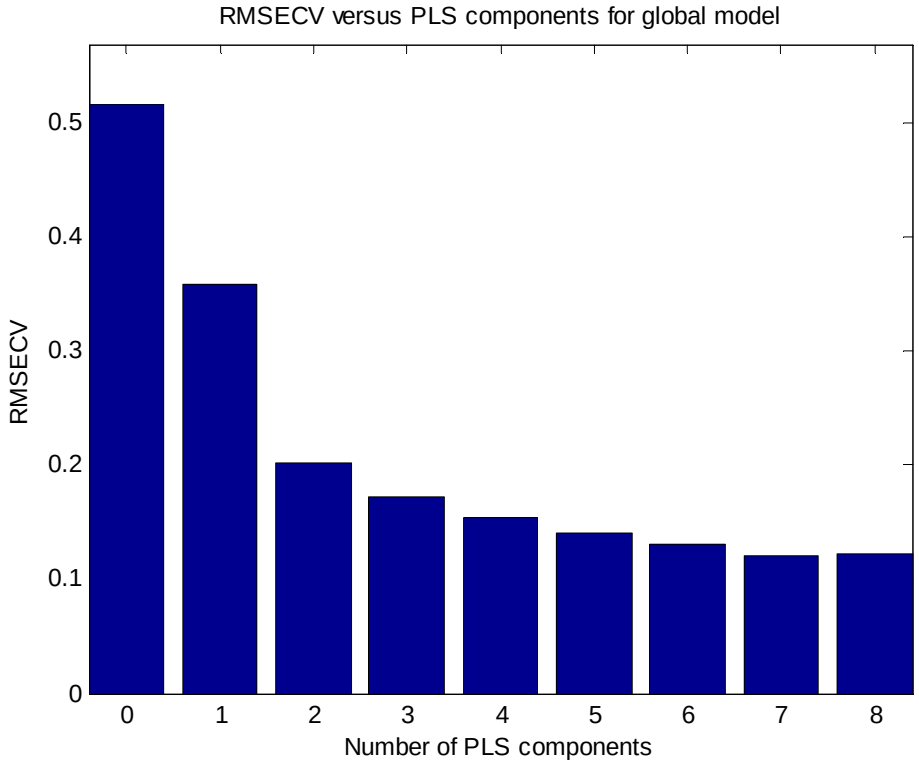


Figure 70 : (Figure 2 Publication) : RMSECV value for the global model as a function of the number of latent variables

The minimum value of RMSECV is found for 7 components (0.12038). However, this is most probably an overoptimistic model, so a more robust model with just 3 latent variables was retained.

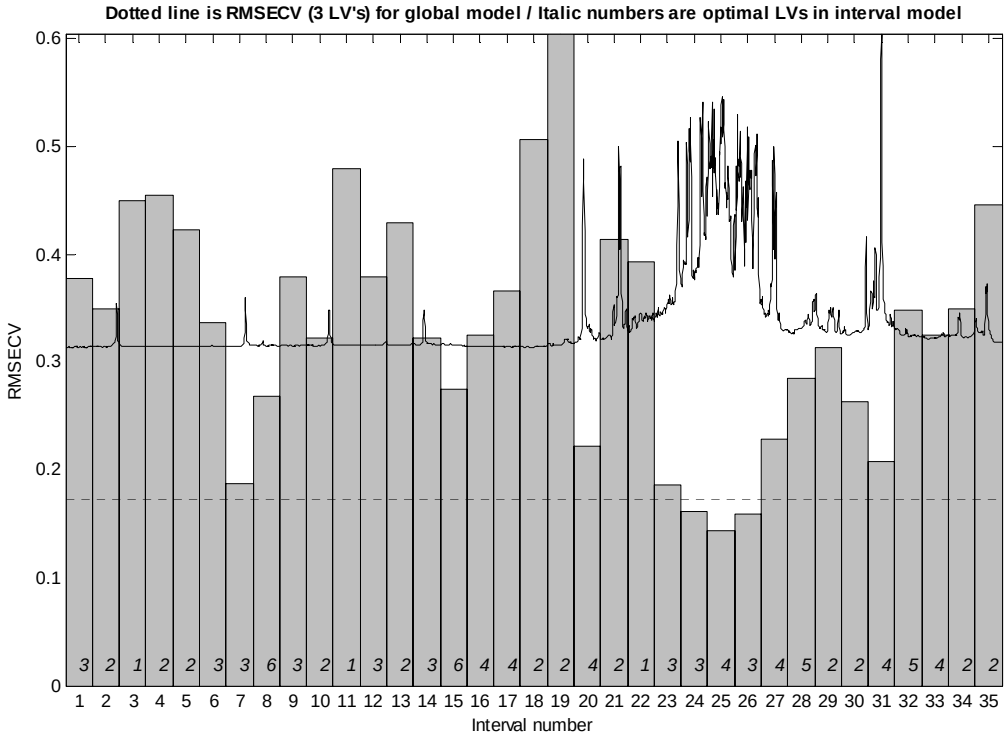


Figure 71 : (Figure 3 Publication) : Minimum RMSECV values for each of the 35 intervals using the number of LVs indicated in italics at the bottom of the corresponding bar.

By comparing the value of RMSECV of the different intervals (Figure 3 Publication / Figure 71) with the reference model value (0.17293), 3 intervals (24, 25 and 26) were selected as they had lower values.
As well, intervals 7, 8, 15, 20, 23, 27, 30 and 31 also seem to be of interest.

Table 2 Publication / Tableau 31gives the correspondences between the interval number, their positions in the spectrum and the lowest RMSECV value.

Tableau 31. (Table 2 Publication) : Correspondences between the intervals and their position in the spectrum and their lowest RMSECV value.

Interval N°	First point number	Last point number	Minimal RMSECV value
1	1	99	0,37784
2	100	198	0,34917
3	199	297	0,43875
4	298	396	0,45524
5	397	495	0,4228
6	496	594	0,33671
7	**595**	**693**	**0,18704**
8	**694**	**792**	**0,26815**
9	793	891	0,37892
10	892	990	0,32198
11	991	1089	0,47946
12	1090	1188	0,37914
13	1189	1287	0,35232
14	1288	1386	0,21496
15	**1387**	**1485**	**0,27462**
16	1486	1584	0,32497
17	1585	1683	0,36693
18	1684	1782	0,50721
19	1783	1881	0,51755
20	**1882**	**1980**	**0,22177**
21	1981	2079	0,33607
22	2080	2178	0,3932
23	**2179**	**2277**	**0,17609**
24	2278	2376	0,15533
25	2377	2475	0,14325
26	2476	2574	0,1584
27	**2575**	**2673**	**0,22807**
28	2674	2772	0,2854
29	2773	2871	0,25306
30	**2872**	**2970**	**0,26299**
31	**2971**	**3068**	**0,20853**
32	3069	3166	0,34777
33	3167	3264	0,32459
34	3265	3362	0,34967
35	3363	3460	0,44666
REF 3 LVs	1	3460	0.1729
REF 7 LVs	1	3460	0.1204

Bold : large selection

Two kinds of selection will be studied : a small range of intervals (24 to 26), apparently very discriminating (corresponding to the 297 variables from 2278 to 2574) as seen in the rectangle in the Figure 4 Publication / Figure 72; a second selection will also include the 8 other intervals, resulting in a total of 1088 variables.

The attributions of the main compounds in these zones are given in Table 3 Publication / Tableau 32.

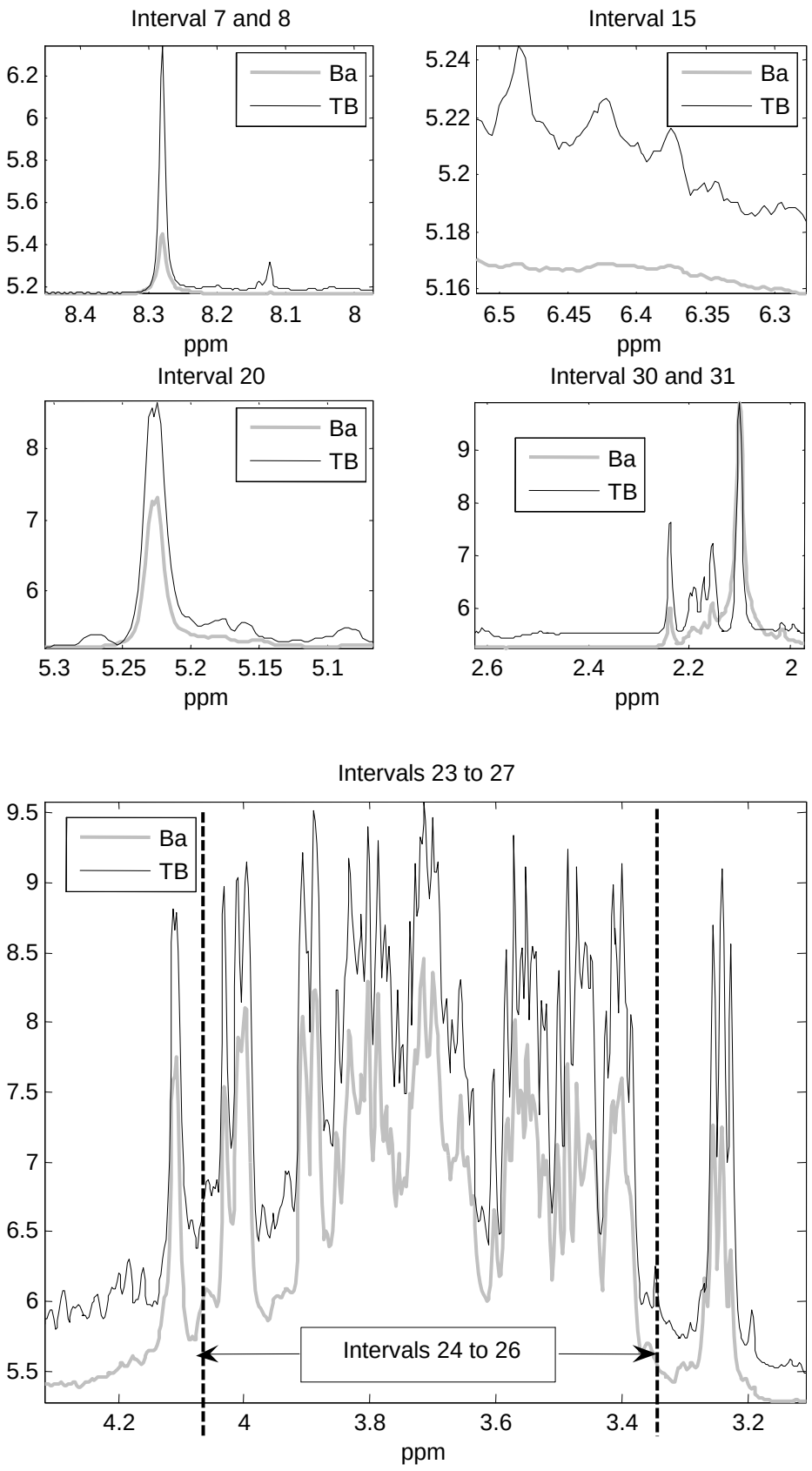


Figure 72 : (Figure 4 Publication) : The 11 intervals selected using iPLS

Tableau 32. (Table 3 Publication) : Characteristics of ^{1}H NMR signals in the selected zones by iPLS of the vinegars (Functional groups are given as according to Fan [26] and Caligiani [8])

Compound	Group	δ (ppm)	Multiplicity[a]	J (Hz)	Hydrogen N°
Acetic acid	$C2H_3$	2.07	s	-	3
β-D-glucopyranose	C2H	3.26	dd	8.03; 9.29	1
β-D-fructopyranose	C5H	4.01	m	-	1
β-D-fructopyranose	C6Hax	4.05	dd	9.29; 8.05	1
β-D-fructofuranose	C3H + C4H	4.10	m	-	2
Formic acid	HC	8.31	s	-	1

[a] Multiplicity: s singleton, d doublet, t triplet, dd doublet of doublets, m multiplet

Selection of variables with EWZS.

Applying the EWZS function described above to the warped and log-transformed data set, and using the sample type (Ba or TB) as the parameter to be predicted, gave the R^2 map shown in Figure 5 Publication / Figure 73. The minimum window size was set to 7 variables, the maximum window size to 250, while the increment between each series of evolving windows was a step of 25 variables. Visual examination of the map from the ICA-based decomposition (Figure 5 Publication / Figure 73) shows that seven zones of high R^2 can be identified. The highest values of R^2 in those areas enable us to detect the intervals giving the best models (Table 4 Publication / Tableau 33).

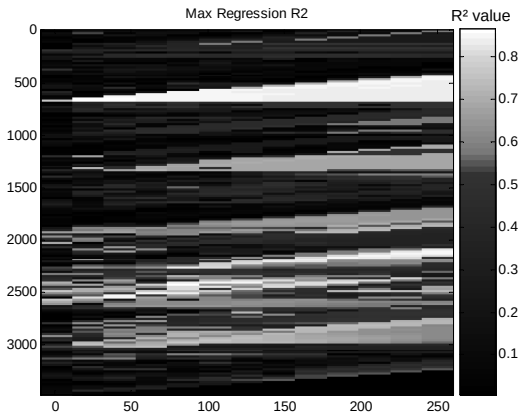


Figure 73 : (Figure 5 publication) : Map of maximum R^2

Tableau 33. (Table 4 Publication) : R^2 values of the selected intervals

	First point number	Last point number	Value
Interval 1	646	686	0.872
Interval 2	966	1050	0.5773
Interval 3	1308	1340	0.7169
Interval 4	1851	2101	0.7311
Interval 5	2253	2342	0.8414
Interval 6	2414	2498	0.8642
Interval 7	2535	2578	0.8572
Interval 8	2957	3041	0.7662

Interval 1 gives the highest R^2 value for the whole dataset.

Interval 2 was not in fact centred on a peak, so we decided to reduce the range of this zone from 966 - 1050 to 966 - 990.

Intervals 5, 6 and 7 give R^2 values above 0.84. Taking into account those 8 zones, it is possible to select a dataset containing only 822 variables instead of the original 3460 variables.

The average spectra of the TB and Ba vinegars within the selected zones are represented in Figure 6 Publication / Figure 74.

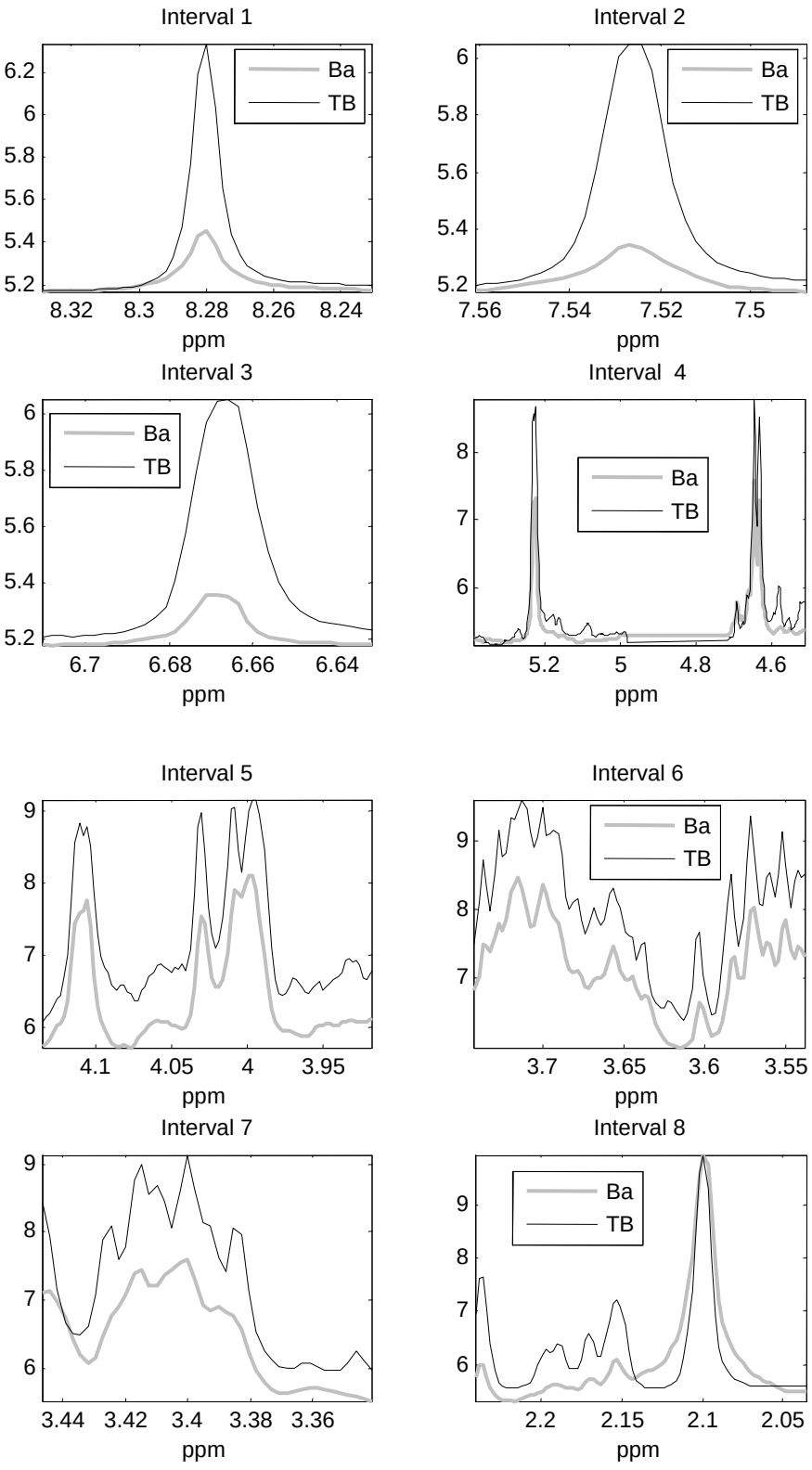


Figure 74 : (Figure 6 Publication) : The eight intervals selected by EWZS function

Identification of the chemical constituents contributing to the zones selected by EWZS.

Tableau 34. (Table 5 Publication) : Characteristics of ^{1}H NMR signals observable in vinegars (Functional groups are given according to Fan [26] and Caligiani [8]) in the selected zone by EWZS

Compound	Group	δ (ppm)	Multiplicity[a]	J (Hz)	Hydrogen N°
Acetic acid	$C2H_3$	2.09	s	-	3
Possible acetic esters		2.15-2.22			
Overlap of glucose and fructose		3.90			
β-D-fructopyranose	C5H	3.99	m	-	1
β-D-fructopyranose	C6Hax	4.04	dd	9.29; 8.05	1
β-D-fructofuranose	C3H + C4H	4.10	m	-	2
β-D-glucopyranose	C1H	4.63	d	7.97	1
Unidentified		5.07			
α-D-glucopyranose	C1H	5.23	d	3.69	1
Hydroxymethylfurfural (HMF)	C4H	6.67	d	3.38	1
HMF	C3H	7.53	d	3.41	1
Formic acid	HC	8.28	s	-	1

[a] Multiplicity: s singleton, d doublet, t triplet, dd doublet of doublets, m multiplet

The presence of acetic acid (Table 5 Publication / Tableau 34) among the most discriminant constituents for balsamic and traditional balsamic vinegars could be explained by the fact that those compounds are ageing indicators. HMF is derived from must cooking and is present in large amounts both in balsamic and traditional balsamic vinegars, but the quantity may vary depending on the length of the ageing process, the storage in wooden barrels and the caramel addition (authorised only for balsamic vinegars). The identification of the signals at 3.95 and 5.08 ppm is not easy because in these spectral zones there are strong signals overlaps, but they probably are due to sugars. The signal at 5.08 ppm could be a sugar derivative such as an ester, which may be slowly formed by reactions between sugars and organic acids (acetic acid

in particular). Using the same hypothesis, the singletons near acetic acid may represent acetic acid esters.

Results on the calibration sets.

The 10 partitions of the samples are submitted to variable selection using either iPLS or EWZS and then the PLS-DA or ICA models dimensionality is determined by leave-one-out cross-validation.

iPLS/PLS-DA.

Three kind of PLS models will be evaluated. The first is the one on the global spectra in order to have a reference for the quality of the prediction without variable selection. Then two variable selections using iPLS were tested, either a small one containing the 3 intervals with value of RMSECV lower than the reference model value, or a larger one with 11 low-RMSECV regions. The results of the leave-one-out cross-validation on the sets of 20 samples are presented in Table 6 Publication / Tableau 35.

Tableau 35. (Table 6 Publication) : Results of leave-one-out cross-validation (using Mahalanobis distance) on the 10 calibration sets using the IPLS/PLS method with up to 3 LVs

Xi = Number of calibration samples out of 20 correctly classified by Cross-validation

	LVs used	*X1*	*X2*	*X3*	*X4*	*X5*	*X6*	*X7*	*X8*	*X9*	*X10*
Whole spectrum	**1**	14	16	16	15	15	15	16	14	16	15
	1 - 2	15	17	16	15	16	16	16	15	16	16
	1 - 3	***19***	***17***	***17***	***20***	***18***	***17***	***17***	***20***	***17***	***16***
Small set of intervals (24, 25 & 26)	**1**	14	16	16	15	15	15	16	14	16	16
	1 - 2	17	18	18	***20***	17	17	18	20	18	18
	1 - 3	***19***	***19***	***18***	19	***19***	***19***	***18***	***20***	***19***	***19***
Selection of all 11 intervals	**1**	14	16	16	15	15	15	16	14	16	16
	1 - 2	16	17	16	17	16	17	17	17	17	17
	1 - 3	***18***	***18***	***18***	***20***	***19***	***18***	***17***	***20***	***18***	***18***

***Italic bold*:** best results

The PLS model on the entire spectrum gives a group membership prediction rate of 89 % of the samples correctly classified by leave-one-out cross-validation.

Using the small set of intervals detected by iPLS enhances the correct classification rate to 94.5 %. However this small area corresponds to the sugar zone in the spectrum. The signals in this zone are superimposed and some interesting peaks may be hidden. In addition sugar composition is easy to modify and a correct authentication of samples may not be possible on such easily imitated constituents. The larger selection of 11 intervals gives a good prediction rate (92 %) by leave-one-out cross-validation. This model is quite stable over the 10 different partitions of the calibration data set. Moreover, it takes into account compounds such as the formic acid and acetic acids that are related to fermentation and aging processes.

EWZS/ICA.

For the ICA-based models, since the order and nature of the "pure signals" extracted by algorithm may vary with the samples in the data set and with the number of ICs being extracted, it is necessary to verify the correspondence of the ICS. For example, with the 6^{th} partitioning of the calibration dataset, IC 2 and IC 3 were sometimes exchanged. Hence the 2^{nd} and 3^{rd} independent components were reordered so as to be able to compare the classifications.

Only the results obtained on the selected zone will be shown. However, the best rate on the entire spectrum was 88.5 % (very close to the result with the PLS model on the whole spectrum) and corresponded to the combination of IC2 and IC3.

Tableau 36. (Table 7 Publication) : Result of leave-one-out cross-validation (Mahalanobis distance) on the 10 calibration sets using the EWZS/ICA method with 3 ICs

Xi = Number of samples correctly classified out of 20

ICs used	*X1*	*X2*	*X3*	*X4*	*X5*	*X6*	*X7*	*X8*	*X9*	*X10*
1	14	16	16	15	14	15	16	14	16	16
2	***19***	18	18	***20***	18	***19***	***19***	***20***	***19***	***19***
3	16	17	18	18	18	11	18	17	17	18
1 & 2	***19***	16	18	19	18	18	***19***	***20***	18	***19***
1 & 3	18	18	18	17	19	15	18	16	18	18
2 & 3	***19***	***19***	***19***	***20***	***19***	17	***19***	***20***	***19***	18
1 to 3	***19***	17	18	19	17	***19***	***19***	***20***	18	***19***

Italic bold: best results

The second IC best discriminates the vinegars sample whereas the first ICs discriminates least well (Table 7 Publication / Tableau 36). Further in our study we will look at this component and try to understand why it gives such a good discrimination. Using the 2nd and 3rd components for the ICA model gives the best results for leave-one-out cross-validation (94.5 %). However for the partitions 6 and 10, the best results are found using all three ICs. For the prediction, ICs 2 and 3 will be used.

Results on the validation sets.

iPLS/PLS-DA.

To validate the models, the predicted memberships of the 9 samples in the validation sets were assessed and compared to the real values. The results of the correct assessment by PLS models are given in Table 8 Publication / Tableau 37.

Tableau 37. (Table 8 Publication) : Correct classification of the validation samples by iPLS/PLS models

	LVs used	*Y1*	*Y2*	*Y3*	*Y4*	*Y5*	*Y6*	*Y7*	*Y8*	*Y9*	Y**10**
Whole spectrum	**1 to 3**	9	8	8	8	9	8	8	8	8	8
Small set of intervals (24, 25 & 26)	**1 & 2**				8						
	1 to 3	9	9	9		9	9	9	8	9	9
Selection of all 11 intervals	**1 to 3**	9	9	9	8	9	9	8	8	9	9

Once again, the model using the small range in the sugar region gives the best classification results (97.8%).

EWZS/ICA.

In the same way as with the iPLS/PLS-DA method, the EWZS/ICA models are validated on the classification of the validation set of 9 samples. The results of the correct assessment of group membership by ICA models are shown in Table 9 Publication / Tableau 38.

For the entire spectrum (results not shown) the best classification rate was to be had on the combination of all 3 ICs.

Tableau 38. (Table 9 Publication) : Correct classification of the validation samples by EWZS/ICA models

ICs used	*Y1*	*Y2*	*Y3*	*Y4*	*Y5*	*Y6*	*Y7*	*Y8*	*Y9*	Y**10**
2 & 3	**9**	**9**	**8**	**8**	**9**	**8**	**8**	**8**	**8**	**9**

Here the correct classification rate is in average of 93.3 % (Table 9 Publication / Tableau 38).

PLS-DA model on the iPLS selection.

In order to understand the reasons for the observed classification results, a PLS model was calculated using all 29 vinegar samples on the large selection of 11 intervals. The 3 latent variables are plotted in Figure 7 Publication / Figure 75 for each of the selected intervals.

In Figure 7 Publication / Figure 75, it can be seen that in some intervals (7, 8, 15, 20, 30 and 31) LV2 looks like a real NMR signal. However the LV1 and LV3 do not. Hence, it may be difficult to interpret the physico-chemical or spectral origin of these LVs.

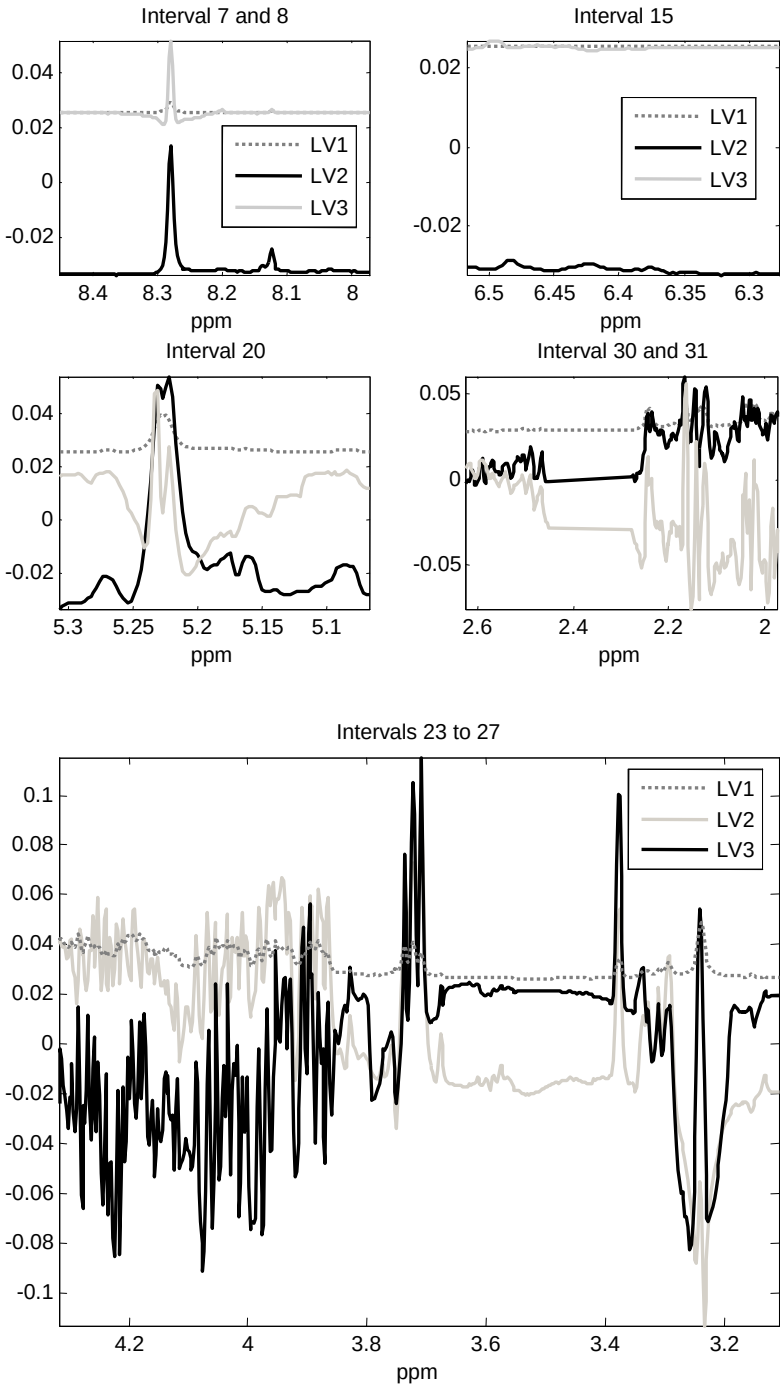


Figure 75 : (Figure 7 Publication) : The first 3 Latent Variables of the PLS-DA model on the 29 samples in the zones selected by iPLS

ICA model with 3 Independent Components on the selected zones by EWZS.

To understand on which spectral features the discrimination is based, it is necessary to look at the IC loadings. These were calculated using all 29 samples.

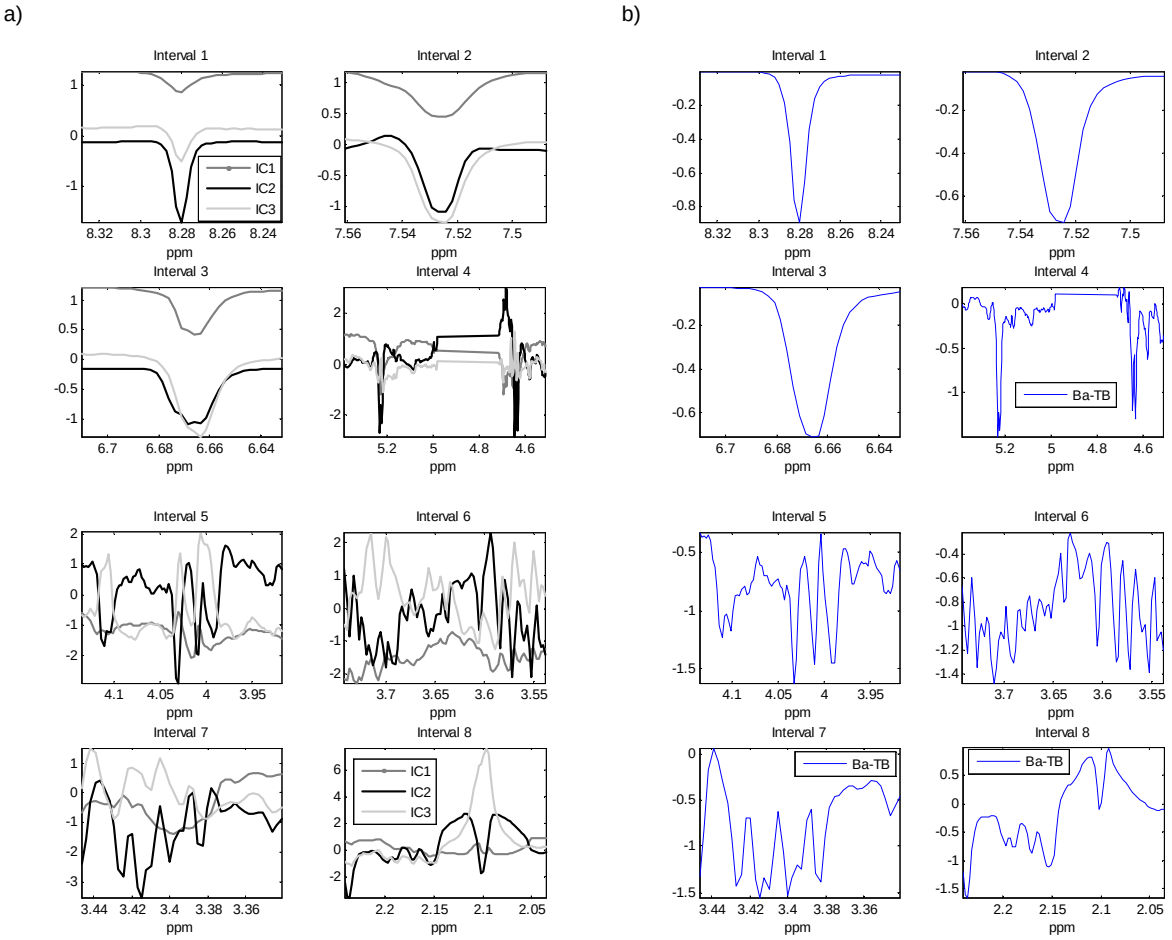


Figure 76 : (Figure 8 Publication) : a) ICA "Loadings" of the ICA model with 3 components in the 8 zones selected by EWZS; b) Difference between average signals of balsamic vinegar and traditional balsamic vinegar in the 8 zones selected by EWZS

The ICA loadings (Figure 8 Publication / Figure 76) should represent three independent signal sources. In fact, IC1 looks like the negative average of the balsamic vinegar signal (Figure 6 Publication / Figure 74).

IC2 is the component that best discriminates the two kinds of vinegar, and looks like the difference between the average spectrum of traditional balsamic vinegar and traditional

balsamic vinegar. TB samples have lower values on this component than Ba samples (Figure 9 Publication / Figure 77).

IC3 also shows the peaks that make the difference between the two sets of vinegar. This would explain the fact that IC3 is also a discriminant component (cf. Tables 7 and 9 Publication / Tableaux 36 and 38).

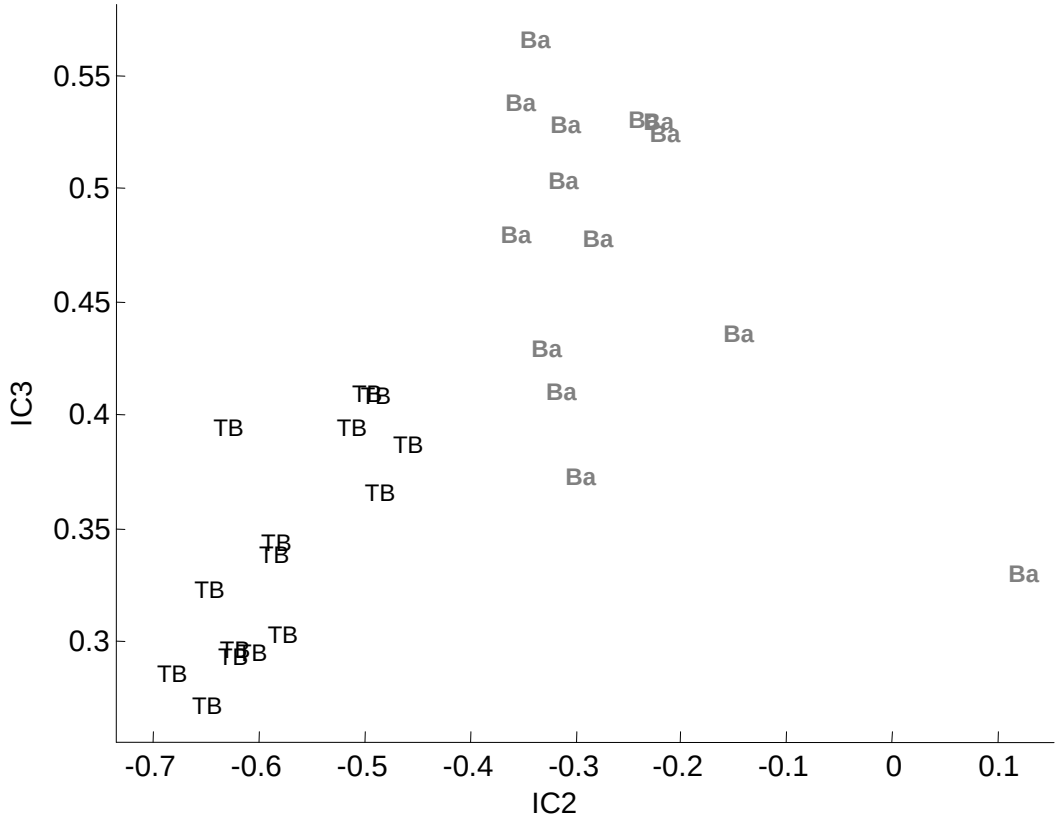


Figure 77 : (Figure 9 Publication) : ICA "Scores" of the vinegar samples in the IC2-IC3 plane

The two kinds of vinegar can be discriminated just using IC2 (Figure 9 Publication / Figure 77). Within each group, it is not possible to find sub-groups clearly related to other sample differences such as ageing or producer. This demonstrates that the method is able to differentiate between traditional balsamic and balsamic vinegars, independently of other characteristics of the sample.

CONCLUSION

By combining two chemometric tools (EWZS, ICA) and a relatively fast and non-invasive analytical method (^{1}H NMR spectroscopy), we have developed an objective method able to discriminate more expensive traditional balsamic vinegar from the ordinary balsamic vinegar.

By comparison with the commonly used iPLS/PLS-DA method, this new method distinguishes itself in that the ICA model facilitates the interpretation of the differences between samples. Here the discriminative component was an approximation of the difference between the average spectrum of Ba and the average spectrum of TB.

In addition, EWZS variable selection can select areas of varying size in one run of the function, whereas iPLS usually requires several runs. Moreover, only one dimensionality reduction method was used here. This function has as options other method such as PLS, PCA and PLS-DA.

The results obtained on this dataset are all the interesting in that the samples in each category were very different in terms of origin, age, and maturation methodology. With the view of validating a protocol for the authentication of traditional balsamic vinegars, it is planned to apply this procedure to a larger set of samples.

The ^{1}H NMR/EWZS/ICA method can be applied to all liquid or extractable food matrices for food characterization and authentication, and presents numerous advantages. First, it uses a relatively fast and automatable instrumental technique, then a mathematical procedure for automatic determination of interesting spectral zones, and finally a robust and accurate discrimination method that can be easily related to samples properties of interest.

References

1. European Council Regulation, E. No 813/2000 (17.04.2000). *Official Journal of European Communities*.
2. Cocchi, M., Bro, R., Durante, C., Mancini, D., Marchetti, A., Saccani, F., Sighinolfi, S. & Ulrici, A. Analysis of sensory data of Aceto Balsamico Tradizionale di Modena (ABTM) of different ageing by application of PARAFAC models. *Food Qual. Prefer.* **17**, 419 (2005).
3. Giudici, P. Gluconic acid: genuineness criterion of the traditional balsamic vinegar. *Ind. Bev.*, 123 (1993).
4. Del Signore, A. Infrared spectra (Mid-IR) classification of balsamic vinegar. *Journal of Commodity Science* **39**, 159-172 (2000).
5. Cocchi, M., Durante, C., Foca, G., Mancini, D., Marchetti, A. & Ulrici, A. Application of a wavelet-based algorithm on HS-SPME/GC signals for the classification of balsamic vinegar. *Chemometrics and intelligent laboratory systems* **71**, 129-140 (2004).

6. Chiavaro, E., Caligiani, A. & Palla, G. Chiral indicator of ageing in balsamic vinegar of Modena. *J. Food Sci* **4**, 329 (1998).
7. Anklam, E., Lipp, M., Radovic, B., Chiavaro, E. & Palla, G. Characterisation of Italian vinegar by pyrolysis-mass spectrometry and a sensor device ('electronic nose'). *Food Chem.* **61**, 243 (1998).
8. Caligiani, A., Acquotti, D., Palla, G. & Bocchi, V. Identification and quantification of the main organic components of vinegars by high resolution ^{1}H NMR spectroscopy. *Anal. Chim. Acta*, 110-119 (2007).
9. Consonni, R. & Gatti, A. ^{1}H NMR Studies on Italian Balsamic and Traditional Balsamic Vinegars. *J. Agric. Food Chem.* **52**, 3446-3450 (2004).
10. Tomasi, G., van den Berg, F. & Anderson, C. Correlation Optimized Warping and time warping as preprocessing methods for chromatographic data. *J. Chemometrics* **18**, 231-241 (2004).
11. Cichocki, A. & Amari, S. (ed. Wiley), (New York, 2002).
12. Hyvarinen, A., Karhunen, J. & Oja, E., (Wiley, New York, 2001).
13. Astakhov, S. A., Stogbauer, H., Kraskov, A. & Grassberger, P., (2005).
14. Shimizu, S., Hyvarinen, A., Kano, Y. & Hoyer, P. O. Discovery of non-gaussian linear causal models using ICA.
http://www.cs.helsinki.fi/u/ahyvarin/papers/Shimizu05UAI.pdf.
15. Cardoso, J.-F. Blind separation of real signals with JADE.
http://www.tsi.enst.fr/icacentral/Algos/cardoso/ (1995).
16. Cardoso, J.-F. & Souloumiac, A. in *IEE*, 362-370 (1993).
17. Cardoso, J.-F. High-order contrasts for independent component analysis. *Neural Computation* **11**, 157-192 (1999).
18. Jiang, J. H., Berry, R. J., Siesler, H. W. & Ozaki, Y. Wavelength Interval Selection in Multicomponent Spectral Analysis by Moving Window Partial Least-Squares Regression with Applications to Mid-Infrared and Near-Infrared Spectroscopic Data. *Anal. Chem.* **74,**, 3555-3565 (2002).
19. Guthrie, J. A., Walsh, K. B., Reid, D. J. & Liebenberg, C. J. Assessment of internal quality attributes of mandarin fruit. 1. NIR calibration development. *Australian Journal of Agricultural Research* **56**, 405-416 (2005).

Annexe 1 : La spectroscopie des liquides par Résonance Magnétique Nucléaire (RMN) du proton

C'est en 1938 qu'Isidore Isaac Rabi découvre le phénomène de résonance magnétique sur des jets moléculaires. Huit ans plus tard, deux équipes américaines, Purcell, Torrey et Pound à Harvard, et Bloch, Hansen et Packard à Stanford précisent la notion de fréquence de résonance.

1. Principes de la spectroscopie RMN

La spectroscopie par RMN exploite l'existence de moments magnétiques nucléaires et leurs variations dans un champ magnétique [1].

1.1. Propriétés des noyaux atomiques

Un noyau peut être caractérisé par un moment angulaire, ou cinétique, $\vec{P}$, et un moment magnétique $\vec{\mu}$, qui lui est proportionnel.

$$\vec{\mu}=\gamma\vec{P} \qquad \textbf{Equation 1}$$

γ étant le rapport gyromagnétique, paramètre caractéristique du noyau.

Dans un champ magnétique $\vec{B}_0$ orienté selon l'axe *z*, la projection de $\vec{P}$ sur z est :

$$P_z=\frac{mh}{2\pi} \qquad \textbf{Equation 2}$$

où *h* est la constante de Planck et *m* est une valeur de spin nucléaire qui ne peut prendre que $2I+1$ valeurs $\|-I;+I\|$. *I* étant le nombre quantique de spin du noyau considéré, les valeurs de *I* sont entières ou demi-entières.

Dans le cas où $I = ½$ (cas du proton), $\vec{P}$ ne peut avoir que deux valeurs sur *z*, proportionnelles à – ½ et + ½, ce qui correspond à deux orientations différentes.

1.2. Le noyau dans un champ magnétique $\vec{B}_0$

En présence d'une induction $\vec{B}_0$, un moment magnétique $\vec{\mu}$ est soumis à un couple et selon les lois de Newton :

$$\frac{d\vec{P}}{dt}=\vec{\mu}\wedge\vec{B}_0 \qquad \textbf{Equation 3}$$

Ainsi l'équation du mouvement de $\vec{\mu}$ est :

$$\frac{d\vec{\mu}}{dt}=\gamma\vec{\mu}\wedge\vec{B}_0 \qquad \textbf{Equation 4}$$

L'intégration de cette équation donne l'équation du mouvement de précession de $\vec{\mu}$ autour de l'axe du champ magnétique $\vec{B}_0$ à la vitesse angulaire $\vec{\omega}_0$:

$$\vec{\omega}_0=-\gamma\vec{B}_0 \qquad \textbf{Equation 5}$$

ou à la fréquence de précession (« fréquence de Larmor »), ν_0 :

$$\nu_0=\left(\frac{\gamma}{2\pi}\right)B_0 \qquad \textbf{Equation 6}$$

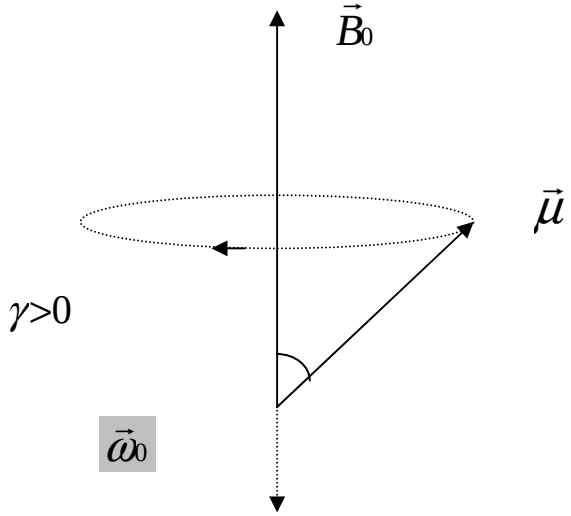


Figure 1 : Précession de $\vec{\mu}$ dans le champ magnétique $\vec{B}_0$

A chacun des états de spin du noyau est associé un niveau d'énergie E_m,

$$E_m=m\,\nu_0 \qquad \textbf{Equation 7}$$

D'un point de vue quantique, le champ magnétique $\vec{B}_0$ lève la dégénérescence des $2I+1$ niveaux de spin avec une énergie d'interaction avec le champ qui peut s'écrire :

$$E=-\vec{\mu}.\vec{B}_0=-\mu_z B_0=\left(\frac{h}{2\pi}\right)mB_0 \qquad \textbf{Equation 8}$$

Dans le cas d'un noyau de spin I = ½, deux niveaux d'énergie sont ainsi définis :

$$E_m=+\frac{\nu_0}{2} \text{ et } E_m=-\frac{\nu_0}{2} \qquad \textbf{Equation 9}$$

Pour le proton, l'état α pour I = +½ est celui de plus basse énergie, l'état β (I = - ½) est celui de plus haute énergie.

L'écart d'énergie entre les deux niveaux est :

$$\Delta E = \gamma\left(\frac{h}{2\pi}\right)B_0 \qquad \textbf{Equation 10}$$

Ceci régit la répartition des noyaux selon la statistique de Boltzmann. Le rapport des populations dans les états + ½ (N_+) et – ½ (N_-) est donné par :

$$\frac{N_-}{N_+} = \exp\left(\frac{-\Delta E}{k_b T}\right) \qquad \textbf{Equation 11}$$

k_0 est la constante de Boltzmann et T la température absolue.

En raison de la faible différence d'énergie entre les deux niveaux, la différence de population est faible, mais crée cependant un léger paramagnétisme nucléaire, $\vec{M}_0$. A l'équilibre, celui-ci est difficile à détecter car il est masqué par le fort diamagnétisme électronique. L'intensité de l'aimantation d'équilibre M_0, associée à un ensemble de N_0 noyaux obéit en général à l'équation :

$$M_0 = \frac{\left[N_0 \mu_2 (I+1) B_0\right]}{3 k_b T I} \qquad \textbf{Equation 12}$$

L'intensité est donc d'autant plus grande que la quantité de noyaux, le rapport gyromagnétique du noyau, et l'intensité du champ magnétique sont plus grands.

1.3. Le phénomène de résonance

La détection du magnétisme nucléaire est réalisée au moyen d'une méthode basée sur le principe de résonance. On utilise pour cela un deuxième champ magnétique B_1 perpendiculaire à B_0. Pour que ce champ B_1 puisse interagir avec les spins qui tournent autour de B_0 à la fréquence de Larmor, ν_0, il faut qu'il tourne dans le même sens à la même vitesse angulaire. Ceci est possible si l'on utilise un champ électromagnétique radiofréquence. Lorsque la fréquence de B_1 est différente de ν_0, aucune interaction n'est observée. Les spins individuels continuent à précesser autour de B_0. Dans ces conditions l'aimantation résultante, M_0, est alignée autour de l'axe 0Z sans autre orientation privilégiée, c'est à dire qu'il n'existe pas de composante dans le plan X-Y perpendiculaire à l'axe 0Z.

Par contre, lorsque la fréquence de B_1 est égale à ν_0, ce champ magnétique est immobile par rapport au référentiel tournant des spins et une interaction devient possible. Alors les spins se mettent à tourner autour de B_1 dans ce référentiel en rotation en même temps qu'ils tournent autour de B_0 dans le référentiel statique. C'est le phénomène de résonance. On dit que les noyaux "résonnent" à la fréquence ν_0. Lorsque cette "condition de Larmor" est remplie, les

spins sont basculés sous l'influence de B_1 et une composante Mxy non-nulle apparaît dans le plan X-Y.

Si le champ B_1 est appliqué pendant un temps suffisamment long, on peut faire basculer les spins de 90° par rapport à leur orientation initiale ; si l'application de B_1 dure deux fois plus longtemps, les spins sont basculés d'un angle de 180°.

1.3.1. Relaxation

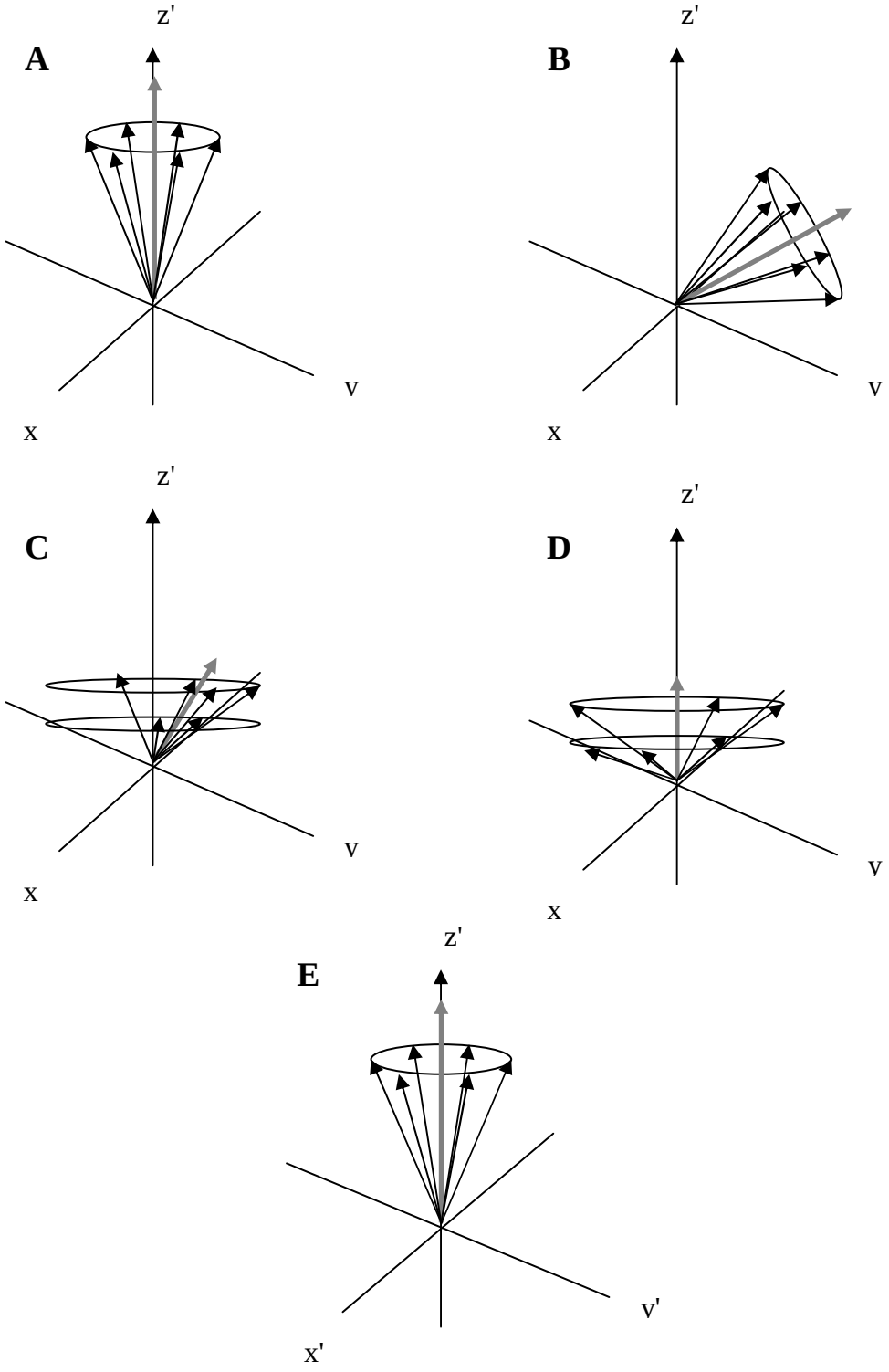
A la résonance, les noyaux changent d'orientation. Lorsque le champ $\vec{B}_1$ n'est plus appliqué à la fréquence de résonance des noyaux, ceux-ci reviennent progressivement vers l'orientation d'équilibre dans le champ magnétique initial B_0. Ces changements se font par des processus de relaxation. L'échange d'énergie peut se faire entre les spins et leur environnement, c'est la relaxation "spin-réseau", ou relaxation "longitudinale" selon la direction de *z,* ou par relaxation "transversale" ou "spin-spin" qui fait intervenir des échanges d'énergie entre les spins eux-mêmes. Ainsi le vecteur aimantation $\vec{M}$ peut retrouver son état d'origine. Le retour à l'équilibre se fait selon les processus décrit par les équations :

$$\frac{dM_z}{dt}=\frac{-(M_z-M_0)}{T_1}=-(M_z-M_0)R_1 \qquad \textbf{Equation 13}$$

$$\frac{du}{dt}=\frac{-u}{T_2}=uR_2 \text{ et } \frac{dv}{dt}=\frac{-v}{T_2}=-vR_2 \qquad \textbf{Equation 14}$$

T_1 est le temps de relaxation longitudinale ou spin-réseau et T_2 est le temps de relaxation transversale, ou spin-spin.

R_1 est la vitesse de relaxation longitudinale ou spin-réseau et R_2 est la vitesse de relaxation transversale, ou spin-spin.



A : Position d'équilibre autour de $\vec{B}_0$, aimantation M_z maximale.

B : Basculement selon *Oy* sous l'influence de $\vec{B}_1$, établissement de *My*.

C : Déphasage des spins par relaxation spin-spin et/ou inhomogénéité de $\vec{B}_0$

D : Relaxation de *My* qui tend vers 0 et retour progressif de M_z à l'équilibre.

E : Retour à la position d'équilibre.

Figure 2 : Phénomène de résonance et de relaxation

1.3.2. Signal de résonance

Chaque noyau présent dans l'échantillon n'est pas soumis exactement au même champ magnétique et ceci à cause de son environnement électronique. Ce sont les faibles variations de champ magnétiques $\Delta\vec{B}_1$ qui justifient des différentes fréquences de résonances et donc d'un signal sous forme de spectre. Pour un noyau *i* de fréquence ν_i, la composante transversale de l'aimantation *v*, détectée en quadrature de phase avec $\vec{B}_1$, a la forme d'une courbe de Lorentz centrée sur ν_i. Une détection en phase avec $\vec{B}_1$ donnerait un signal de dispersion u. En pratique, la représentation du spectre d'absorption, *v*, est préférée.

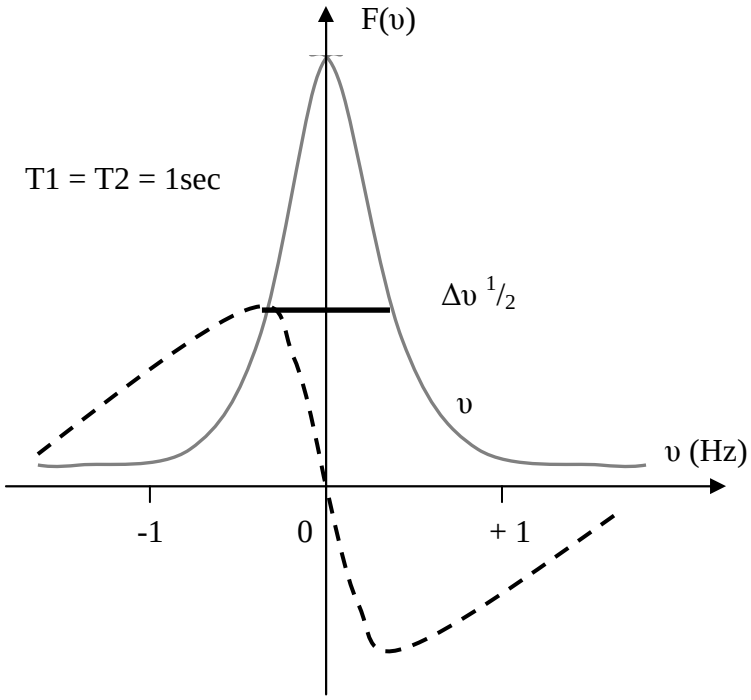


Figure 3 : Signal de résonance

La largeur à mi-hauteur, $\Delta\nu_{1/2}$ (Hz), de la courbe d'absorption *v* est conditionnée par la valeur du temps de relaxation T_2 ou la vitesse de relaxation R_2.

$$\Delta\nu_{1/2} = \frac{1}{\pi T_2} = \frac{R_2}{\pi}$$ **Equation 15**

En pratique, sous l'effet d'une impulsion, dont la durée, notée PW, (Pulse Width) est de l'ordre des microsecondes, un signal, noté *f(t)*, interférogramme des fréquences excitées, est recueilli. Il correspond à la somme des sinusoïdes de fréquences ν_i décroissant avec les temps de relaxation T_{2i}, il est dit "signal d'induction libre" ou FID (Free Induction Decay). Le FID d'un seul type de proton a la forme d'une exponentielle décroissante :

$$f(t) = K M_0 \exp(-\frac{t}{T_2}) = K M_0 \exp(-t R_2)$$ **Equation 16**

La transformée de Fourier de cette fonction donnera un signal contenant un seul pic à la fréquence de Larmor spécifique de ce proton.

La transformée de Fourier (TF) du signal dans le domaine temps permet d'accéder directement à l'information correspondante dans le domaine fréquence - soit la courbe *F(v)* représentée dans la Figure 3. Le signal après TF contient autant de pics que de fréquences différentes.

2. Applications de la spectroscopie RMN

Après traitement du FID par la transformée de Fourier, le spectre *F(v)* est constitué de l'ensemble des raies de fréquence v_i chacune ayant une forme de Lorentzienne de largeur à mi-hauteur $\Delta v_{1/2} = \frac{1}{\pi T_{2i}}$ et une surface proportionnelle au nombre de noyaux résonnant à la fréquence v_i. De manière plus générale, l'intensité du signal est proportionnelle à l'aimantation d'équilibre M_0 qui augmente avec le nombre de noyaux et l'intensité du champ B_0.

2.1. Identification de la structure d'un composé

2.1.1. Champ d'écran

Comme nous l'avons vu précédemment, tous les noyaux d'une molécule ne sont pas soumis au même champ magnétique. En effet, dans un atome, la présence des électrons autour du noyau *i* provoque la formation d'un champ diamagnétique $\Delta\vec{B}_i$ qui s'oppose au champ dû à l'aimant $\vec{B}_0$. Le champ subit par le noyau i, $\vec{B}_{i(subit)}$ est donc plus faible :

$$\vec{B}_{i(subit)} = \vec{B}_0 + \Delta\vec{B}_i$$ **Equation 17**

Il y a un "effet d'écran".

La condition de résonance définit en X doit donc être remplacée par :

$$\nu(\vec{B}_1) = \nu_i = \left(\frac{\gamma}{2\pi}\right)(\vec{B}_0 + \Delta\vec{B}_i) = \left(\frac{\gamma}{2\pi}\right)(1-\sigma_i)\vec{B}_0$$ **Equation 18**

σ_i étant la constante d'écran du noyau *i*

Plus généralement, chaque type d'atome d'une molécule subit donc un champ magnétique résultant différent si les environnements électroniques sont différents.

2.1.2. Déplacement chimique

Pour définir la position des différents signaux de proton, on utilise le déplacement chimique, δ. La mesure absolue de B_0 avec une précision meilleure que 10^{-6} étant impossible, il est plus aisé de mesurer ν par rapport à la fréquence de résonance d'une substance de référence $\nu_{réf}$. En outre, pour permettre une comparaison plus commode entre des spectres enregistrés à l'aide de spectromètres de champ B_0 différent, il s'agit de s'affranchir du paramètre champ magnétique. Ainsi, le déplacement chimique d'un noyau *i*, δ_i, est défini par :

$$\delta_i = \frac{\nu_i - \nu_{ref}}{\nu_{réf}} \qquad \textbf{Equation 19}$$

δ_i est sans dimension en unité 10^{-6} (ppm)

Le déplacement chimique par rapport à une référence choisie est donné par :

$$\nu_i - \nu_{réf} = \left(\frac{\gamma}{2\pi}\right)(\sigma_{réf} - \sigma_i)B_0 \qquad \textbf{Equation 20}$$

Une division par $\nu_{réf}$ permet de faire disparaître la valeur du champ.

$$\delta_i = \frac{\nu_i - \nu_{ref}}{\nu_{réf}} = \frac{\left(\frac{\gamma}{2\pi}\right)(\sigma_{réf} - \sigma_i)B_0}{\left(\frac{\gamma}{2\pi}\right)(1 - \sigma_{réf})B_0} \approx (\sigma_{réf} - \sigma_i) \qquad \textbf{Equation 21}$$

En négligeant $\sigma_{réf}$ de l'ordre 10^{-6} devant l'unité, il vient :

$$\delta_i = (\sigma_{réf} - \sigma_i) \qquad \textbf{Equation 22}$$

L'échelle δ représente donc bien une mesure relative du coefficient d'écran, indépendante de B_0. Elle est fonction de l'environnement chimique du noyau, donc fonction de sa nature. Dans la pratique, δ est déterminé en mesurant la différence de fréquence ($\nu_i - \nu_{ref}$) en Hz, divisée par la fréquence de travail (ν_{ref}) exprimée en Hz.

Comme les premiers spectromètres fonctionnaient par balayage de champs, les spectres étaient représentés sur une échelle de fréquences croissantes de la droite vers la gauche. L'échelle des déplacements chimiques varie dans le même sens et celle des constantes d'écran en sens inverse. Aujourd'hui cette représentation a été conservée.

2.1.3. Constante de couplage

Dans une molécule, les noyaux voisins dotés de spins ont leurs moments magnétiques qui interagissent entre eux. Ces interactions sont transmises par les électrons des liaisons reliant tous les atomes qui se trouvent entre les deux spins considérés.

Pour deux protons voisins, H_1 et H_2, ayant des déplacements chimiques différents dus à la différence d'environnement électronique. Le proton H_2 vient perturber le champ magnétique nécessaire à l'obtention de la résonance de H_1. Chacun des types de proton peut être dans l'état α ou l'état β. Il y a donc quatre possibilités pour l'état global de ces deux protons : $\alpha_1\beta_2$, $\alpha_1\alpha_2$, $\beta_1\alpha_2$, et $\beta_1\beta_2$, correspondant donc à quatre niveaux d'énergie différents et donc comme nous l'avons vu à quatre fréquences de Larmor différentes. L'énergie de couplage mise en jeu est définie par J_{12}, constante de couplage entre le spin de H_1 et celui de H_2 .

L'énergie de l'état $\alpha_1\beta_2$ est :

$$E_{\alpha_1\beta_2} = \frac{1}{2}\nu_{0,1} - \frac{1}{2}\nu_{0,2} - \frac{1}{4}J_{12} \qquad \textbf{Equation 23}$$

et de manière générale :

$$E_{m_1 m_2} = m_1\nu_{0,1} + m_2\nu_{0,2} + m_1 m_2 J_{12} \qquad \textbf{Equation 24}$$

où m_i est la valeur du spin de H_i .

Les fréquences de résonance sont donc, selon les interactions :

Tableau 1. Fréquence de résonance en fonction des interactions

Transition	Etats des spins	Fréquence de résonance
1 -> 2	$\alpha_1\alpha_2$ -> $\alpha_1\beta_2$	$-\nu_{0,2} - \frac{1}{2}J_{12}$
3 -> 4	$\beta_1\alpha_2$ -> $\beta_1\beta_2$	$-\nu_{0,2} + \frac{1}{2}J_{12}$
1 -> 3	$\alpha_1\alpha_2$ -> $\beta_1\alpha_2$	$-\nu_{0,1} - \frac{1}{2}J_{12}$
2 -> 4	$\alpha_1\beta_2$ -> $\beta_1\beta_2$	$-\nu_{0,1} + \frac{1}{2}J_{12}$

Ceci implique un dédoublement des signaux RMN de part et d'autre de la fréquence de résonance, avec une séparation de J_{12} Hz. Cette notion est expliquée par la Figure 4.

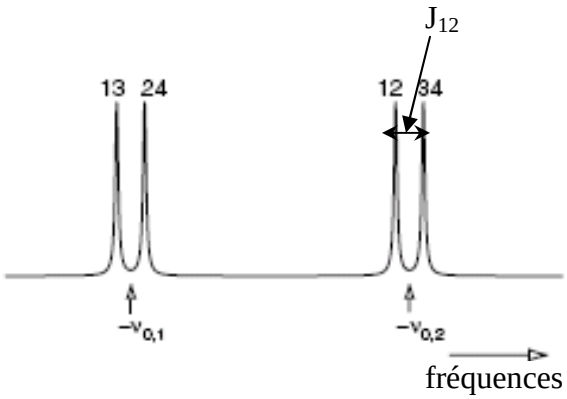


Figure 4 : Signal RMN de deux protons couplés

De manière plus générale, pour *n* protons, soumis à un même environnement électronique, couplés à *m* protons voisins, on obtiendra un multiplet de *(m+1)* pics de par et d'autres de la fréquence de résonance des *n* protons. L'aire du signal total reste proportionnelle à la valeur *n*.

Les constantes de couplage sont indépendantes de la valeur du champ appliqué et donc du spectromètre utilisé contrairement au déplacement chimique.

2.1.4. Détermination de la structure d'une molécule

Une fois le spectre RMN d'un composé acquis, il est possible d'identifier ses déplacements chimiques. Grâce à des tables de correspondance entre déplacements chimiques et groupements atomiques, des morceaux de la molécule peuvent être identifiés. En outre, avec la proportionnalité entre les noyaux et l'intensité du signal chacun des signaux est associé à un nombre de proton, ainsi les groupements atomiques peuvent être quantifiés.

Pour affiner la structure de la molécule, il faut s'intéresser aux phénomènes de couplage qui sont détectables sur le spectre. En effet, quand des noyaux sont voisins et interagissent, le signal présente des multiplets séparés par des constantes de couplages déterminées selon les niveaux d'énergie des noyaux. Ceci permet de situer les différents groupements atomiques les uns par rapport aux autres dans la molécule. Ainsi, la structure de la molécule devient accessible. S'il reste quelques incertitudes le couplage de la RMN proton avec la RMN ^{13}C, peut permettre de les lever.

Chaque composé possède donc une empreinte RMN spécifique. C'est de cette propriété que découlent toutes les applications de profiling**.**

2.2. Quantification

La quantification d'un composé dans le spectre, utilise une propriété de proportionnalité entre l'intensité des signaux, *I*, et le nombre de noyaux, ρ, qui raisonnent et le nombre de mole *N*.

$$I = \rho \times N \qquad \textbf{Equation 25}$$

On utilise une référence dont la concentration et donc le nombre de mole, $N_{réf}$, dans l'échantillon est connue, ainsi que sa pureté $P_{réf}$. Le nombre de mole de l'analyte, N_A, se déduit de la relation suivante, [2] :

$$N_A = N_{réf} \times \frac{I_A}{I_{réf}} \times \frac{\rho_{réf}}{\rho_A} \times P_{réf} \qquad \textbf{Equation 26}$$

où *A* est l'analyte et *réf*, la référence.

2.3. Empreintes spectrales

Bien que la RMN soit réputée peu sensible du fait que son seuil de détection est plus faible que celui d'autres techniques, elle s'avère être une méthode rapide pour balayer toutes les gammes de composés. Sur une matrice contenant de nombreux composés différents, le spectre proton est constitué des signaux de chaque composé. Ces signaux peuvent se superposer si les mêmes groupements sont présents dans différents composés, se chevaucher si les différents déplacements chimiques sont très proches, et avoir des intensités très variables. Il est alors parfois difficile d'attribuer chaque signal à un composé spécifique. Même si seulement une partie du spectre est attribuable, il peut constituer une empreinte caractéristique de l'échantillon entier. L'acquisition de spectre ou "profil" RMN ^{1}H est à la base de différentes études dites de "profiling" utilisées pour par exemple, le suivi de l'évolution de composés et la recherche de voies de synthèse. D'autres applications ont vu le jour dans plusieurs domaines. En médecine ou en botanique, les composés du métabolisme (métabolome) sont étudiés afin de pouvoir dans le premier cas faire des diagnostiques précoces de maladies [3], et dans le deuxième cas étudier les voies de biosynthèse [4-6] ou faire de la taxinomie chimique [7].

En agroalimentaire, le "profiling" permet de classer les échantillons selon des origines géographiques. Ceci a été fait sur l'huile d'olive [8], et sur le vin [9]. De même il permet de discriminer les origines botaniques des produits agroalimentaires comme le café [10] ou encore détecter la présence de composés ajoutés d'origines botaniques différentes comme dans le cas de l'huile d'olive adultérée par d'autres huiles [11].

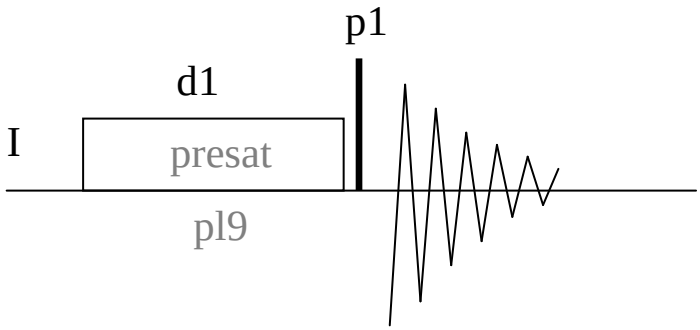
3. Références

1. Martin, M. L. Les isotopes au service de la qualité. *Bull.Union Phys.*, 793-811 (1990).
2. Wells, R. J., Hook, J. M., Al-Deen, T. S. & Hibbert, D. B. Quantitative NMR spectroscopy for assessing the purity of technical grade agrochemicals: 2,4-dichlorophenoxyacetic acid (2,4-D) and sodium 2,2-dichloropropionate (Dalapon sodium). *Agriculture Food Chemistry* **50**, 3366-3374 (2002).
3. Griffin, J. L. Metabolomics: NMR spectroscopy and pattern recognition analysis of body fluids and tissues for characterisation of xenobiotic toxicity and disease diagnosis. *Current opinion in Chemical Biology* **7**, 1-7 (2003).
4. Ott, K.-H., Aranibar, N., Singh, B. & Stockton, G. W. Metabonomics classifies pathways affected by bioactive compounds. Artificial neural network classification of NMR spectra of plant extracts. *Phytochemistry* **62**, 971-985 (2003).
5. Frédérich, M. et al. Metabolomic analysis of *Strychnos nux-vomica*, *Strychnos icaja* and *Strychnos ignatii* extracts by ^{1}H NMR spectrometry and multivariate analysis techniques. *Phytochemistry* **65**, 1993-2001 (2004).
6. Choi, Y. H. et al. Metabolic discrimination of *Catharanthus roseus* leaves infected by phytoplasma using ^{1}H NMR spectroscopy and multivariable data analysis. *Plant Physiology* **135**, 2398-2410 (2004).
7. Choi, Y. H. et al. Classification of Ilex species based on metabolomic fingerprinting using NMR and multivariate data analysis. *J. Agric. Food Chem.* **53**, 1237-1245 (2005).
8. Mannina, L., Patumi, M., Proietti, N., Bassi, D. & Segre, A. L. Geographical characterization of Italian extra virgin olive oils using High-field ^{1}H NMR spectroscopy. *J. Agric. Food Chem.* **49**, 2687-2696 (2001).
9. Kozir, I. J. & Kidric, J. Use of modern NMR spectroscopy in wine analysis: determination of minor compounds. *Analytica Chimica Acta* **458**, 77-84 (2002).
10. Tavares, L. A., Ferreira, A. G., Correa, A. & Mattoso, L. H. in *Magnetic resonance in food science* (ed. Chemistry C, T. R. S. o.), 80-88 (2004).
11. Fauhl, C., Reniero, F. & Guillou, C. ^{1}H NMR as a tool for the analysis of mixtures of virgin olive oil with oils of different botanic origin. *Magnetic Resonance in Chemistry* **38**, 436-443 (2000).

Annexe 2 : Séquence d'acquisition zgpr

Il existe différentes façons de supprimer le pic d'un solvant dans un spectre. La méthode la plus classique est la présaturation.

Il s'agit d'irradier sélectivement une fréquence donnée (celle du solvant), avec une faible puissance avant d'exciter la totalité des protons et d'enregistrer leurs résonances. Cette saturation à basse puissance excite les protons de la molécule d'eau de telle façon qu'aucun signal ne peut entièrement accumuler et être mesurer. Cela permet de limiter la région excitée, et donc altérée, du spectre au voisinage de la fréquence d'irradiation [1-3]. La largeur de la bande d'excitation va dépendre de la durée de l'irradiation pour une puissance donnée. De part sa forme : irradiation à une fréquence constante pour un temps donné, la présaturation est une impulsion rectangulaire.



pl9 : intensité de la pré-saturation
p1 :
d1 : temps de relaxation; 1-5 * T1
NS : 8 * n, nombre total d'acquisition: NS * TD0

Figure 1 : Séquence d'impulsions « zgpr »

1. Alexander, S. Spin-echo method for measuring relaxation times in two-line NMR spectra. *Rev. Sci. Instrum.* **32**, 1066-1067 (1961).
2. Freeman, R. & Wittekoek, S. Selective determination of relaxation times in high resolution NMR. *J. Magn. Reson.* **1**, 238-276 (1969).
3. Redfield, A. G. & Gupta, R. K. Pulsed Fourier-transform NMR spectrometer for use with H2O solutions. *J. Chem. Phys.* **54**, 1418–1419 (1971).

Annexe 3 : PLS (Partial Least Square Regression)

Cette méthode a été introduite par Wold en 1966 [1] et a été révisée depuis notamment par Tenenhaus en 1998 [2].

1. Objectif de PLS

La régression PLS consiste en une régression de l'ensemble des variables à prédire $\mathbf{y_i}$ (i = {1 : p}) sur des variables latentes $\mathbf{t_1}$, $\mathbf{t_2}$, …, combinaisons linéaires de $\mathbf{x_1}$, $\mathbf{x_2}$, …$\mathbf{x_m}$, où m représente le nombre de colonnes de la matrice **X** des variables explicatives. Cependant dans la méthode PLS les variables latentes sont déterminées en tenant compte à la fois des $\mathbf{y_i}$ et des variables prédictives $\mathbf{x_1}$, $\mathbf{x_2}$, …$\mathbf{x_m}$.

Dans la suite nous allons décrire le cas d'une seule variable à prédire : **y**.

Si on représente matriciellement ce que l'on recherche dans le modèle PLS, on a :

$$\mathbf{X} = \mathbf{T_k}\mathbf{P_k} + \mathbf{E_k} \quad \text{et} \quad \mathbf{y} = \mathbf{T_k}\mathbf{q_k} + \mathbf{f_k} \qquad \textbf{Equation 4}$$

Avec :

$\mathbf{T_k}$: la matrice des coordonnées factorielles ('scores') (n × k)

$\mathbf{P_k}$: la matrice des poids ('loadings') (k × m) associés à la prédiction de **X** à partir de **T**, ensemble des composantes PLS

$\mathbf{E_k}$: la matrice des résidus associée à la prédiction de X (n × m)

$\mathbf{q_k}$: le vecteur des poids ('loadings') (k × 1) associés à la prédiction de **y** à partir de $\mathbf{T_k}$

$\mathbf{f_k}$: le vecteur des résidus associé à la prédiction de **y** (n × 1)

L'expression $\mathbf{X} = \mathbf{T_k}\mathbf{P_k} + \mathbf{E_k}$ est similaire à celle vu pour la reconstruction des spectres en ACP. Cependant ce qui est intéressant c'est de prédire **y**.

2. Modélisation

L'algorithme de la régression PLS est itératif. La première étape consiste à calculer $\mathbf{t_1}$, la première composante latente. On estime ensuite les paramètres du modèle à une seule composante soit :

$$\mathbf{X} = \mathbf{T_1}\mathbf{P_1} + \mathbf{E_1} \qquad \textbf{Equation 5}$$

$$\mathbf{y} = \mathbf{T_1}\mathbf{q_1} + \mathbf{f_1} \qquad \textbf{Equation 6}$$

Ainsi chaque ligne de **X** est modélisée par un même « composant pur » dont les éléments forment le vecteur $\mathbf{p_1}$. Les pondérations de ce signal sont données par $\mathbf{t_1}$. De même, en première approximation **y** est modélisé par $\mathbf{t_1}$ pondéré par le nombre q_1.

On peut calculer l'erreur résiduelle, associée à la prédiction de y lors de cette première étape. Si elle est trop importante, on calcule à partir des résidus de $\mathbf{E_1}$ et $\mathbf{f_1}$ une nouvelle composante $\mathbf{t_2}$ et établir les deux nouveaux modèles ci-dessous.

$$\mathbf{X} = \mathbf{T_1P_1} + \mathbf{T_2P_2} + \mathbf{E_2} \qquad \textbf{Equation 7}$$

$$\mathbf{y} = \mathbf{T_1q_1} + \mathbf{T_2q_2} + \mathbf{f_2} \qquad \textbf{Equation 8}$$

Cette procédure est réitérée jusqu'à ce que les composantes $t_1, \ldots, t_k$ expliquent suffisamment **y**, on a, à l'étape k, le modèle suivant à k composantes latentes.

$$\mathbf{X} = \mathbf{T_1P_1} + \mathbf{T_2P_2} + \ldots + \mathbf{T_kP_k} + \mathbf{E_k} \qquad \textbf{Equation 9}$$

$$\mathbf{y} = \mathbf{T_1q_1} + \mathbf{T_2q_2} + \ldots + \mathbf{T_kq_k} + \mathbf{f_k} \qquad \textbf{Equation 10}$$

A l'étape k, la composante t_k est déterminée à partir de $\mathbf{E_{k-1}}$ et $\mathbf{f_{k-1}}$ de l'étape précédente. Dans cette logique de suite on note **y** : **y0** et **X** : **X0**.

On cherche une relation du type :

$$\mathbf{t_k} = \mathbf{X_{k-1}}\,\mathbf{w_k} \qquad \textbf{Equation 11}$$

Dans la régression PLS, le vecteur $\mathbf{w_k}$ est déterminé à chaque étape à partir de $\mathbf{X_{k-1}}$ et de $\mathbf{y_{k-1}}$. Le vecteur $\mathbf{w_k}$ est un vecteur de norme unitaire optimisé tel que la covariance entre $\mathbf{t_k}$ et $\mathbf{y_{k-1}}$ soit la plus grande possible. Soit :

$$\mathbf{Cov}\,(\mathbf{y_{k-1}}, \mathbf{t_k}) = \mathbf{y^T_{k-1}}\,\mathbf{t_k} = \mathbf{y^T_{k-1}}\,\mathbf{X_{k-1}}\,\mathbf{w_k} \qquad \textbf{Equation 12}$$

On a donc que $\mathbf{Cov}\,(\mathbf{y_{k-1}}, \mathbf{t_k})$ est proportionnelle au produit scalaire entre le vecteur $\mathbf{y^T_{k-1}}\,\mathbf{X_{k-1}}$ et le vecteur $\mathbf{w_k}$. Cette quantité est maximum quand l'angle entre ces deux vecteurs est nul. Ils sont alors colinéaires et de même sens. On a alors :

$$\mathbf{w_k} = \lambda\,\mathbf{y^T_{k-1}}\,\mathbf{X_{k-1}} \qquad \textbf{Equation 13}$$

avec $\lambda \geq 0$.

Comme $\mathbf{w_k}$ est supposé normé, on a :

$$\mathbf{w_k} = \frac{\mathbf{y^T_{k-1}}\,\mathbf{X_{k-1}}}{\left\|\mathbf{y^T_{k-1}}\,\mathbf{X_{k-1}}\right\|} \qquad \textbf{Equation 14}$$

Avec cette expression de $\mathbf{w_k}$, on peut exprimer $\mathbf{t_k}$ à partir de l'équation 8. Puis, grâce à une régression linéaire simple entre $\mathbf{y_{k-1}}$ et $\mathbf{t_k}$, on détermine $\mathbf{q_k}$, le coefficient de la régression et, puis $\mathbf{y_k}$, le résidu du modèle selon la relation :

$$\mathbf{y_{k-1}} = \mathbf{q_k}\,\mathbf{t_k} + \mathbf{y_k} \qquad \textbf{Equation 15}$$

On procède de même avec chacune des m colonnes de $\mathbf{X_{k-1}}$ et $\mathbf{t_k}$. Pour une colonne $\mathbf{x_i}$ de $\mathbf{X_{k-1}}$, cette régression s'écrit :

$$\mathbf{x_i} = \mathbf{p}\,\mathbf{t_k} + \mathbf{r_i} \qquad \textbf{Equation 16}$$

Soit sous forme matricielle :

$$\mathbf{X_{k-1}} = \mathbf{t_k p_k} + \mathbf{X_k} \qquad \textbf{Equation 17}$$

La prédiction d'une valeur y connaissant un vecteur **x** peut se faire, en théorie, par la même approche itérative. On calcule les k composantes **t** associés à **x**, puis on applique la relation suivante :

$$\hat{y}=\mathbf{q}_1\mathbf{t}_1+\mathbf{q}_2\mathbf{t}_2+\ldots+\mathbf{q}_k\mathbf{t}_k \qquad \textbf{Equation 18}$$

Mais, en général, on l'exprime directement sur **x** :

$$\hat{y}=b_0+b_1x_1+b_2x_2+\ldots+b_mx_m \qquad \textbf{Equation 19}$$

Où $\mathbf{b_0, b_1, \ldots, b_m}$ sont les coefficients de la régression PLS.

3. Application à la discrimination : PLS-Discriminant Analysis (PLS-DA)

Il peut être intéressant de prédire des variables qualitatives. En effet, cela permet de différencier différents groupes d'individus. C'est en 1986 que l'approche PLS-Analyse Discriminante (PLS-DA) a été développée [3].

La PLS-DA est une méthode de classification supervisée qui utilise la régression PLS pour classifier des échantillons selon leurs propriétés mesurées. On prédit l'appartenance d'un échantillon à différentes classes. **Y** est donc non plus formé de variables quantitatives, mais de colonnes de 1 et de 0 selon que l'échantillon appartient ou non à la classe représentée par la colonne *i*.

Si l'on cherche à différencier n individus selon 3 groupes (groupe 1, groupe 2, groupe 3), on construit la matrice $\mathbf{Y}_{(n,3)}$, chaque échantillon n'appartenant qu'à un seul des groupes.
Chaque ligne de **Y** contiendra un 1 dans la colonne du groupe correspondant à l'échantillon et deux 0 dans les deux autres colonnes.

1. H. Wold, Partial Least Squares, S. Kotz, Johnson, N.L. Vol. 6, Wiley, New York, 1985, pp. 581-591.
2. M. Tenenhaus, *Revue de Statistique Appliquée*, 47 (1999) 2-55.
3. M. Sjöström, S. Wold, and B. Söderström, PLS Discrimination Plots, E.S.G.a.L.N. KANALS, Elsevier, Amsterdam, 1986.

Annexe 4 : Classification de variables autour de composantes latentes (CLV [1-2])

1. Objectif de CLV

Cette méthode a pour but de regrouper des variables qui sont corrélées. Pour se faire, elle regroupe les variables autour de composantes latentes. Pour prendre en compte la corrélation des variables entre elles, la composante latente représentative d'un groupe est définie par la première composante obtenue par analyse en composantes principales.

2. Modélisation

Soient x_1, x_2, …, x_m, m variables centrées mesurées sur n échantillons. Soient G_1, G_2, …, G_k, k groupes de variables et $\mathbf{c_1}$, $\mathbf{c_2}$, …, $\mathbf{c_k}$, leurs k composantes latentes associées.

Pour regrouper des variables corrélées, indépendamment de leur signe de corrélation, le critère T à maximiser est :

$$T = n\sum_{i=1}^{k}\sum_{j=1}^{m}\delta_{ij}COV^2(\mathbf{x_j},\mathbf{c_i}) \text{ avec } \mathbf{c_k^T c_k} = 1, \qquad \textbf{Equation 1}$$

$\delta_{kj} = 1$ si la j^{ème} variable appartient au groupe G_k et $\delta_{kj} = 0$ sinon.

Si $\mathbf{X_i}$ est la matrice formée par les variables appartenant à G_k en colonne, T devient :

$$T = \frac{1}{n}\sum_{i=1}^{k}\mathbf{c}_i^T\mathbf{X}_i\mathbf{X}_i^T\mathbf{c}_i \text{ avec } \mathbf{c}_i^T\mathbf{c}_i = 1 \qquad \textbf{Equation 2}$$

Dans notre étude, le signe des corrélations n'importe pas.

3. Solution

Un algorithme itératif permet de trouver la solution la meilleure en bouclant sur deux étapes. Pour une partition donnée, la première étape transforme les variables de chaque groupe en un repère de composantes principales. Pour le groupe G_i, la première composante principale définit ainsi la composante latente $\mathbf{c_i}$.

A la deuxième étape, chaque variable va être reclassée dans le groupe G_i, pour lequel la composante latente $\mathbf{c_i}$ permet d'obtenir un coefficient de covariance mis au carré la plus élevée. k nouveaux groupes sont ainsi redéfinis.

L'algorithme est relancé jusqu'à ce que les groupes soient stables.

4. Choix du nombre de groupe

Pour trouver une solution au partitionnement des variables, il faut avoir choisi a priori un nombre de groupe *k*. Le problème réside donc dans le choix de ce nombre.

Lorsque le nombre de groupe diminue dans la classification hiérarchique le critère T diminue. La variation ΔT du critère T lors du passage d'une partition en R à R-1 groupes, s'exprime de la façon suivante :

$$\Delta T = \lambda_1^{(A)} + \lambda_1^{(B)} - \lambda_1^{(A \cup B)} \qquad \textbf{Equation 3}$$

où A et B sont les deux groupes de variables qui sont réunis lorsque l'on passe de R à (R-1) groupes. $\lambda_1^{(A)}$ est la plus grande valeur propre associée à la matrice de variance-covariance de l'ensemble de variables A, par exemple.

L'évolution de ΔT au cours de la classification hiérarchique permet de choisir le nombre de groupe k approprié. En effet, plus la variation de T est importante, plus l'information contenue dans les groupes est perdue lors du regroupement.

5. Références

1. Vigneau, E. & Qannari, E. M. Clustering of variables around latent components. *Communications in statistics* **32**, 1131-1150 (2003).
2. Vigneau, E., Sahmer, K., Qannari, E. M. & Bertrand, D. Clustering of variables to analyse spectral data. *J. Chemometrics* **19**, 122-128 (2005).

Annexe 5 : Analyse en composantes principales (ACP)

1. Objectif de l'ACP

L'analyse en composantes principales (ACP) consiste à rechercher les directions dans l'espace des *m* variables qui représentent le mieux la dispersion des *n* individus.

Elle est particulièrement adaptée à l'étude exploratoire des données spectrales. Elle permet de remplacer les variables d'origine (ici, les fréquences de résonances), fortement redondantes, par des variables synthétiques (variables latentes), les composantes principales. Elles contiennent la totalité de l'information, et ont l'avantage d'être orthogonales entre elles. Ces composantes principales sont des combinaisons linéaires des variables de départ. A l'aide de l'ACP, il est souvent possible de condenser la collection spectrale dans des proportions très importantes : 20 composantes au maximum sont en général largement suffisantes pour résumer l'information utile, et la taille de la matrice des données peut être réduite par un facteur de 10 à 100 ou plus.

Dans l'espace original, la collection spectrale peut être représentée par un nuage de points dans un espace à *m* dimensions. L'ACP consiste en une rotation (qui conserve les échelles et les distances entre les spectres). Cette rotation est effectuée de manière à placer les nouveaux axes dans la direction de plus grande dispersion du nuage de points. Le premier axe correspond ainsi à la plus grande dispersion du nuage de points. Le deuxième axe est orthogonal au premier et tient compte de la plus grande dispersion résiduelle, et ainsi de suite. Les composantes principales sont donc ordonnées par ordre décroissant de variance expliquée. Ainsi les composantes principales d'ordre « élevé » expliquent en général très peu de variance, et peuvent donc être éliminées de l'analyse. Les coordonnées des points représentatifs des spectres dans ce nouveau repère sont les coordonnées factorielles. L'intérêt de l'ACP apparaît clairement lorsque les variables d'origine présentent de fortes corrélations.

Les données condensées par ACP peuvent servir de variables de base pour d'autres traitements statistiques tels que la régression ou l'analyse discriminante. Pour de nombreuses méthodes supervisées, l'orthogonalité des variables rend les calculs numériques très simples et plus fiables.

2. Modélisation

Les données peuvent être structurées dans une matrice **X** à *n* lignes et *m* colonnes.

$$\mathbf{X}_{(n,m)} = \begin{bmatrix} X_{1,1} & \cdots & X_{1,m} \\ \vdots & \ddots & \vdots \\ X_{1,n} & \cdots & X_{n,m} \end{bmatrix} \qquad \textbf{Equation 1}$$

Chaque individu *i* est représenté par le vecteur $\mathbf{u_i} = [x_{i,1}, \ldots, x_{i,m}]$. Chaque variable $\mathbf{X_j}$ est représentée par le vecteur $\mathbf{v_j} = (x_{1,j}, \ldots, x_{n,j})'$ a une moyenne $\overline{X_j}$ et un écart-type $\sigma(X_j)$, avec un poids équivalent affecté à toutes les variables, *1/m*. Le vecteur $(\overline{X_1}, \ldots, \overline{X_j}, \ldots, \overline{X_m})$ est le centre de gravité, **G**, du nuage de points. Ce sera l'origine du nouveau repère. Il faut maintenant rechercher les nouveaux axes.

On centre la matrice **X** sur **G** :

$$\overline{\mathbf{X}} = \begin{bmatrix} X_{1,1} - \overline{X}_1 & \cdots & X_{1,m} - \overline{X}_m \\ \vdots & \ddots & \vdots \\ X_{n,1} - \overline{X}_1 & \cdots & X_{n,m} - \overline{X}_m \end{bmatrix} \qquad \textbf{Equation 2}$$

Les moyennes de la matrice centrées sont nulles.

On va effectuer une rotation dans l'espace afin de placer les échantillons dans l'axe de plus grand étirement.

3. Rotation

Nous avons vu qu'il fallait appliquer une rotation à notre repère d'origine. Cette transformation est représentée dans la Figure 1 ci-dessous.

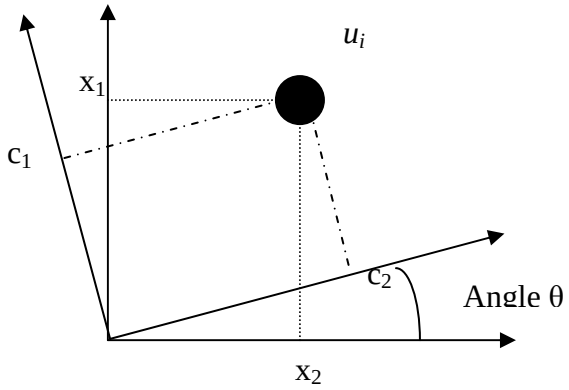


Figure 1 : Changement de repère orthogonal R par rotation d'angle θ

Si on prend un espace orthonormé **R** à 2 dimensions pour simplifier. L'individu $u_{i\ a}$ pour coordonnées dans le premier espace (x_1, x_2). Si on applique une rotation d'angle θ à l'espace **R**, on obtient un espace orthonormé **R'** dans lequel, les cordonnées $c_1$et c_2 sont définis par :

$$\mathbf{c_1 = \cos(\theta)\, x_1 + \sin(\theta)\, x_2} \qquad \textbf{Equation 3}$$

$$\mathbf{c_2 = -\sin(\theta)\, x_1 + \cos(\theta)\, x_2} \qquad \textbf{Equation 4}$$

Sous forme matricielle on obtient :

$$\begin{bmatrix} c_1 \\ c_2 \end{bmatrix} = \begin{bmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \qquad \textbf{Equation 5}$$

On pose :

$$\mathbf{V} = \begin{bmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{bmatrix} = \begin{bmatrix} v_{11} & v_{12} \\ v_{21} & v_{22} \end{bmatrix} \qquad \textbf{Equation 6}$$

On a en introduisant **V** dans l'équation 5 on obtient :

$$\mathbf{c = V\, x} \qquad \textbf{Equation 7}$$

On définit donc que les nouvelles coordonnées sont des combinaisons linéaires des anciennes. La matrice **V** étant une matrice de rotation on a plusieurs propriétés. Tout d'abord les normes des vecteurs formant les lignes de **V** sont égales à 1. Ensuite, les vecteurs ligne sont orthogonaux. Ce qui se résume par :

$$\mathbf{V\, V^T = I} \qquad \textbf{Equation 8}$$

4. Trouver les composantes principales

Nous allons décrire les propriétés que doivent remplir les nouveaux axes afin de les déterminer. Dans le nouveau repère, les coordonnées factorielles des individus (spectres) de la matrice **X** s'écrivent :

$$\mathbf{C = XV} \qquad \textbf{Equation 9}$$

V est une matrice de rotation, de passage d'un repère à l'autre.

Une première contrainte de l'ACP impose l'orthogonalité entre les colonnes de **C**. Cette contrainte permet d'éliminer la redondance dans les nouvelles variables latentes représentées par la matrice **C**. Ainsi, pour remplir cette contrainte, **C** doit avoir des colonnes dont le produit scalaire est nul. On a donc :

$$\mathbf{C^T\, C = L} \qquad \textbf{Equation 10}$$

Avec **L** : matrice dont seuls les éléments diagonaux sont différents de 0. **L** est diagonale.

En remplaçant dans l'équation 10, **C** par son expression donné par l'équation 9 on a :

$$\mathbf{(X\, V)^T\, X\, V = L} \qquad \textbf{Equation 11}$$

Ce qui donne grâce aux propriétés de la transposition :

$$\mathbf{V^T X^T X V = L} \qquad \textbf{Equation 12}$$

En multipliant les deux membres de l'équation précédente par V on obtient :

$$\mathbf{V V^T X^T X V = V L} \text{ d'où } \mathbf{X^T X V = V L} \qquad \textbf{Equation 13}$$

où **L** est diagonale.

La solution de ce système consiste à rechercher les valeurs propres, éléments diagonaux de **L**, et les vecteurs propres, colonnes de **V,** de la matrice $\mathbf{X^T X}$.

L'équation 3 montre que chaque valeur propre est la somme des carrés d'une colonne de **C**. Ainsi la valeur propre d'une des composantes principales est proportionnelle à la variance de cette composante.

Les valeurs propres sont classées dans un ordre décroissant afin d'avoir les composantes principales représentant le plus de variance classées en premiers.

5. Intérêt de l'ACP en spectroscopie

A partir d'un petit nombre de composantes principales k < m, on va pouvoir reconstruire approximativement un spectre. On a donc l'expression de cette reconstruction approximative qui est :

$$\mathbf{x_i = c_{i1} v_1 + c_{i2} v_2 + \ldots + c_{ik} v_k + e_i} \qquad \textbf{Equation 14}$$

Ou $\mathbf{e_i}$ représente l'écart entre le spectre x_i et le spectre reconstruit, c'est donc un vecteur de résidu. Sous forme matricielle on a :

$$\mathbf{X = C_k V_k^T + E_k} \qquad \textbf{Equation 15}$$

Où $\mathbf{C_k}$ et $\mathbf{V_k}$ représentent respectivement les k premières colonnes de **C** et de **V**, et $\mathbf{E_k}$ est la matrice des résidus.

Cette reconstruction s'apparente en une décomposition des spectres en une somme de composantes principales à différentes « concentrations » en fonction des valeurs de c_{ij}.

6. Nouveaux individus

Une fois le modèle ACP construit, de nouveaux individus (spectres) peuvent être projetés dans cet espace. Soit **Z** une nouvelle matrice de données, non-centrée, à *m* colonnes.

Il s'agit de traiter ces spectres comme les précédents : on « centre » la matrice **Z** en fonction de **G**, on obtient **S**.

Et les nouvelles coordonnées sont donc :

$$\mathbf{C_S = S V} \qquad \textbf{Equation 16}$$

Annexe 6 : Analyse en composantes indépendantes (ACI)

1. Objectif de l'ACI

L'analyse en composantes indépendantes est une technique qui découle de la recherche de méthodes de séparation de sources, qui a commencé, il y a une vingtaine d'année [1]. L'objectif de l'ACI est d'estimer *h* signaux "sources", supposés stationnaires, ergodiques [1] et indépendants, en utilisant *n* signaux "observés" ($n \geq h$) qui sont des mélanges inconnus des signaux "sources" [2-7]. L'exemple le plus courant est celui de la séparation de discours différents tenus simultanément dans une même salle provenant de narrateurs (sources) différents et perçus par des récepteurs situés à différents endroits de la salle. Ces signaux peuvent être, pour des raisons physiques, considérés comme mutuellement statistiquement indépendants. L'ACI permet de retrouver les différents discours initiaux à partir des enregistrements. En spectroscopie, on considère que les spectres obtenus sur un ensemble d'échantillons sont des combinaisons linéaires de sources pures que l'on va chercher à estimer.

Pour un ensemble complexe de données où plusieurs signaux différents contribuent aux données observées, l'hypothèse gaussienne de l'analyse en composantes principales (ACP) ne sera pas valable dans tous les cas. En effet, l'objectif de l'ACP est de maximiser la dispersion des individus, et de ce fait, de décorréler les signaux en supposant que toutes les sources suivent une distribution de probabilité gaussienne. Par conséquent, l'extraction de spectres purs et statistiquement indépendants d'un ensemble de mélanges ne peut pas être de manière certaine obtenue par l'utilisation de l'ACP.

Cependant, l'ICA comme l'ACP cherche à transformer l'espace de représentation d'origine. Cette méthode cherche des directions, dans le nouvel espace, telles que les vecteurs résultants soient indépendants, et pas seulement non-corrélés [8]. L'ICA est une méthode de transformation linéaire, dans laquelle la représentation désirée est celle qui minimise la dépendance statistique des composants. L'ICA vise à récupérer les signaux originaux purs en estimant une transformation linéaire, en utilisant un critère lié à la théorie de l'entropie de

[1] Soit un système dynamique évoluant dans le temps. Son évolution est modélisée par la transformation *T* de l'ensemble de tous les états possibles du système. Si le système est dans l'état *x*, à un instant donné, alors *T(x)* représente l'état du système à l'instant suivant.
La théorie ergodique s'intéresse au cas où l'espace d'états est muni d'une loi de probabilité *P* invariante par *T* : si *A* est une partie de l'espace d'états, *P(A)* s'interprète comme la probabilité qu'à un instant donné, l'état du système soit dans *A*. Dire que la probabilité *P* est invariante par *T* signifie que cette probabilité ne dépend pas de l'instant considéré.

l'information, qui fournit l'indépendance statistique entre les signaux en supposant que les données ne suivent pas une distribution gaussienne [9].

2. Modélisation et hypothèses de base

En première approximation, il est supposé que le mélange est linéaire instantané, que le bruit d'observation peut être négligé, et que $n = h$. Soit $\mathbf{s_t}$ et $\mathbf{x_t}$ les vecteurs sources et observations en fonction du temps, respectivement, il existe une relation telle que :

$$\mathbf{x_t = A\, s_t} \qquad \textbf{Equation 1}$$

où A est une matrice dite de "mélange", supposée constante et inversible.

Le but de l'étude est d'estimer **G**, la matrice de séparation, à partir de la seule matrice d'observation $\mathbf{x_t}$, telle que **GA** se réduise à une matrice diagonale à une permutation près. L'hypothèse de départ pour résoudre ce type de problème est l'indépendance des sources.

Il y a plusieurs approches, celle utilisée dans les différentes études présentées dans cette thèse, est de trouver un espace de représentation des observations où les composantes sont aussi indépendantes que possible. Par rapprochement avec le principe de l'ACP, il s'agit aussi de modifier l'espace de représentation des variables. En effet, en ACP, les observations sont représentées dans un espace où la dispersion des individus est maximisée, ici c'est l'indépendance qui est maximisée.

3. Indépendance, et corrélation

Il existe un lien très significatif entre dépendance, corrélation et non gaussienité dans le cadre du modèle linéaire $\mathbf{X} = \mathbf{A\, S}$.

La corrélation entre deux variables met en évidence le rapport entre deux variables qui varient l'une en fonction de l'autre. Il y a un lien, un rapport réciproque.

Deux variables aléatoires sont dites indépendantes lorsque la connaissance de la valeur de l'une n'apporte aucune information sur la valeur de l'autre. L'indépendance se caractérise par le fait que la distribution conjointe est égale au produit des distributions.

L'indépendance de deux variables est une hypothèse plus forte que la non-corrélation. En effet, certaines fonctions peuvent être non-corrélées mais être dépendantes. En effet x et x^2 sont non-corrélées mais dépendantes par définition. Par contre, si deux variables sont indépendantes elles sont non corrélées [10].

Ceci montre que dans la recherche de composantes linéaires, minimiser la dépendance, c'est minimiser la corrélation, mais pas seulement. Comme nous allons le voir au paragraphe suivant la non-gaussienité des variables est aussi un critère important.

4. Mesure de l'indépendance

Pour maximiser l'indépendance ou minimiser la dépendance, il s'agit d'avoir un outil de mesure de l'indépendance. Il existe plusieurs méthodes de calcul, la plus simple est basée sur une propriété du théorème de la limite centrale : "la somme de deux variables aléatoires indépendantes possède une distribution plus proche d'une gaussienne que la distribution de chacune des variables de la somme". Il est donc important d'avoir des sources non-gaussiennes pour mesurer leur indépendance.

Pour mesurer l'éloignement de la distribution d'une variable ***X*** avec une courbe de Gauss, il est fréquent d'utiliser le cumulant d'ordre 4,ou coefficient d'aplatissement (« kurtosis »), défini par :

$$k(\mathbf{X}) = E[\mathbf{X}^4] - 3\,E[\mathbf{X}^2]^2 \qquad \textbf{Equation 2}$$

où E[**X**] est l'espérance de la variable aléatoire **X**.

Cependant, l'estimateur du cumulant d'ordre 4 est relativement sensible au bruit, il est donc préférable d'avoir un nombre d'échantillon choisi en conséquence.

5. Pré-traitement

Le blanchiment des signaux observés permet de réduire le nombre de paramètres à déterminer. Tout d'abord les données sont centrées : x_t est remplacé par $x_t - E[x_t]$, vecteur centré. Puis elles sont décorrélées deux à deux, par une transformation du type :

$$\mathbf{C} = \mathbf{W}\,\mathbf{U}\,\mathbf{R_X}^{-1/2} \qquad \textbf{Equation 3}$$

où **W** est une matrice diagonale quelconque, **U** est une matrice orthogonale quelconque et $\mathbf{R_X}$ est la matrice de covariance de **X.**

Par exemple, si la formule est appliquée à la matrice de covariance $\mathbf{T} = E[\mathbf{x_t}\,\mathbf{x_t}^T]$ de x_t. Soit :

$$\mathbf{T} = \mathbf{P}\,\mathbf{D}\,\mathbf{P}^T \qquad \textbf{Equation 4}$$

où **D** est une matrice diagonale contenant les valeurs propres de **T**, et **P** est la matrice orthogonale des vecteurs propres associés.

Alors :

$$\breve{x}_t = \mathbf{P}\,\mathbf{D}^{-1/2}\,\mathbf{P}^T\,\mathbf{s_t} \qquad \textbf{Equation 5}$$

est telle que :

$$E[\breve{x}_t\,\breve{x}_t^T] = \mathbf{I} \qquad \textbf{Equation 6}$$

I étant la matrice unité. Les différentes composantes de $\breve{x}_t$ sont ainsi décorrélées, $\breve{x}_t$ est dit "blanc".

En conséquence, en injectant cette expression dans l'équation 1, nous obtenons la relation suivante :

$$\breve{x}_t = \breve{A}\, s_t \qquad \textbf{Equation 7}$$

avec $\breve{A} = P\, D^{-1/2}\, P^T\, A$.

En posant comme dernière hypothèse que les différentes composantes des signaux sources sont centrées et ont une variance unité, nous obtenons :

$$E[\breve{x}_t\, \breve{x}_t^T] = \breve{A}\, E[s_t\, s_t^T]\, \breve{A}^T = \breve{A}\, \breve{A}^T = I \qquad \textbf{Equation 8}$$

Le blanchiment des données permet d'obtenir une matrice de mélange $\breve{A}$ qui est orthogonale. Ainsi, la matrice de séparation $\breve{G}$ qui doit être telle que $\breve{G}\breve{A}$ est la matrice unité à une permutation près, d'où :

$$\breve{G}\, \breve{G}^T = I \qquad \textbf{Equation 9}$$

La recherche de la matrice de séparation après avoir blanchi les données est simplifiée puisque nous cherchons maintenant la matrice $\breve{G}$ orthogonale.

6. Matrice de séparation

La matrice de séparation $\breve{G}$ est constituée de vecteurs lignes $\breve{g}^T$ correspondant chacun à une composante indépendante. Comme nous l'avons vu, $\breve{g}^T$ est tel que $\breve{g}^T\, \breve{s}_t$ maximise le critère d'éloignement à une gaussienne sous la contrainte $\|\breve{g}\| = 1$.

Pour résoudre un tel problème, une méthode classique introduit le multiplicateur de Lagrange. Cependant, cette transformation ne permet pas non plus de trouver des solutions analytiques. Pour s'approcher de la solution, il faut donc utiliser des méthodes itératives, comme la méthode du point fixe ou la méthode de Newton, et donc des algorithmes de calcul.

7. JADE

Nous avons utilisé, dans notre étude, l'algorithme JADE (« Joint Approximate Diagonalization of Eigen-matrices »). Il permet d'obtenir une approximation des sources pures S que nous cherchons.

A partir de cette matrice on peut calculer comme pour l'ACP, les coordonnées factorielles (**A**) d'ICA à partir de l'équation suivante :

$$A = X\, S^T (S\, S^T)^{-1} \qquad \textbf{Equation 10}$$

S est la matrice de sources où toutes les colonnes sont statistiquement indépendantes.

Contrairement à l'ACP, les colonnes dans la matrice **A** ne sont pas contraintes d'être orthogonales (Hyvärinen A. *et al*., 2001).

8. Références

1. Cardoso, J.-F. Analyse en composantes indépendantes. http://www.tsi.enst.fr/~cardoso/Papers.PDF/jsbl02-long.pdf (2002).
2. Cardoso, J.-F. Blind separation of real signals with JADE. http://www.tsi.enst.fr/~cardoso/Algo/Jade/jade.m, 12/01/2008, (1995).
3. Durieu, C. & Kieffer, M. Analyse en composantes indépendantes pour la séparation aveugle de sources. http://mfca.ups-tlse.fr/cetsis/Docs/Articles/Durieu_Cecile.pdf, 05/01/2008, (2003).
4. Hoyer, P. O. Independent component analysis in image denoising. http://www.cs.helsinki.fi/u/phoyer/papers/pdf/dippa.pdf, 05/01/2008, (1999).
5. Hyvarinen, A. Survey on independent component analysis. http://www.cs.utexas.edu/~kuipers/readings/Hyvarinen-ncs-99.pdf, 05/01/2008, (1999).
6. Le Borgne, H. & Guérin-Dugué, A. in *ORASIS*, (Cahors, 2001).
7. Shimizu, S., Hyvarinen, A., Kano, Y. & Hoyer, P. O. in *21st Conference of Uncertainty in Artificial Intelligence*, 526–533 (Edinburgh, 2005).
8. Stone, J. V. Independent Component Analysis: An Introduction. *Trends in Cognitive Sciences* **6**, 59-64 (2002).
9. Hyvarinen, A., Karhunen, J. & Oja, E., (Wiley, New York, 2001).
10. Daudin, J.-J., Robin, S. & Vuillet, C. *Statistique inférentielle* (ed. statistique, P. d. l.), pp. 185 (Pesses Universitaires de Rennes, 1999).

Annexe 7 : Attribution des signaux du spectre des jus d'orange et pamplemousse

a)

b)

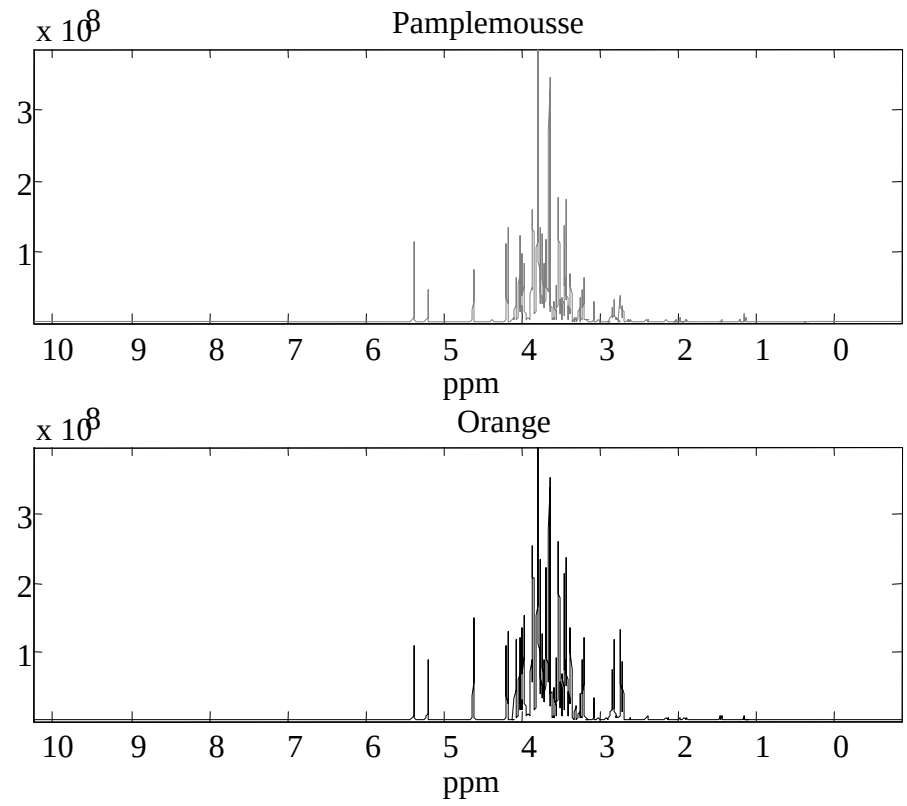


Figure 1 : Spectres moyens : a) Pamplemousse ; b) Orange

a)

b)

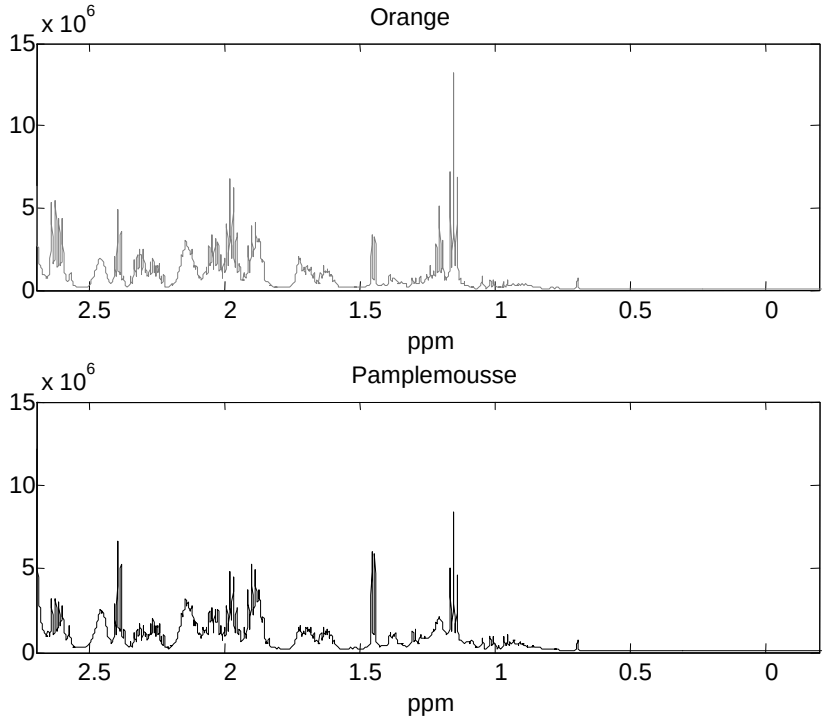


Figure 2 : Spectres moyens, zone spectrale de 0 à 2,7 ppm : a) Pamplemousse ; b) Orange

Tableau 1. Attribution des signaux de la zone spectrale de 0 à 2,7 ppm

Composé	Groupement	δ (ppm)	Multiplicité [a]	J (Hz)	Nombre d' hydrogènes
Leucine	C8	0,93	d	6,00	3
Valine	C8	0,96	d	7,01	3
Leucine	C9	0,98	d	6,00	3
Valine	C7	1,01	d	7,05	3
?		1,05	s		
Ethanol	C3	1,15	t	7,08	3
?		1,21	t		
Hespéridine		127	s		
Thréonine	C8	1,30	d	6,41	2
?		1,38	m		
Alanine	C6	1,45	d	7,14	1
Arginine	C4	1,63	m	-	1
Leucine	C7 ; C6	1,72	m	-	3
Arginine	C3	1,88	dd	9,22 ; 6,53	2
GABA*	C5	1,90	m	-	2
Proline	C3	1,98	m	-	3
Proline	C4	2,04	m	-	1
DMP **		2,15	m		
DMP		2,25	m		
Proline	C4	2,31	m	-	1
GABA	C6	2,40	t	7,39	2
DMP		2,46	m		
Acide succinique	C4 ; C5	2,57	s	-	4

[a] : s : singulet ; d : doublet ; dd : doublet de doublet ; t : triplet ; m : multiplet

* : acide gamma-aminobutyrique ** : Dimethylproline

a)

b)

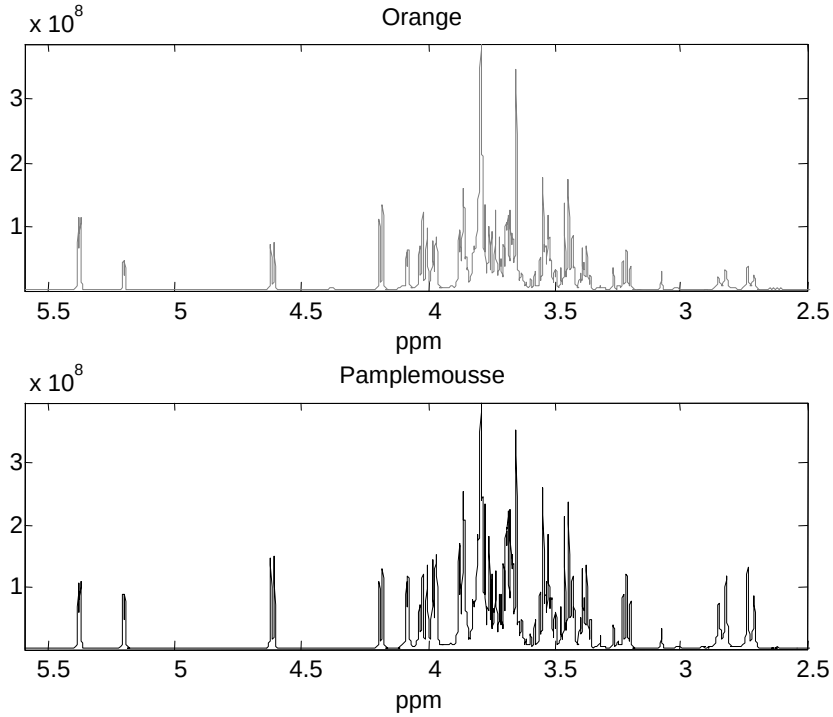


Figure 3 : Spectres moyens, zone spectrale de 2,5 à 5,6 ppm : a) Pamplemousse ; b) Orange

Tableau 2. Attribution des signaux de la zone spectrale de 2,5 à 5,6 ppm

Composé	Groupement	δ (ppm)	Multiplicité [a]	J (Hz)	Nombre d' hydrogènes
Acide malique	C5	2,62	dd	15,38 ; 10,12	1
Acide citrique	C6b ; C3b	2,72	d	-	2
Acide malique	C5	2,79	dd	15,38 ; 2,9	1
Acide citrique	C6a ; C3a	2,83	d	15,16	2
DMP		3,08	s		
Glucose	C2	3,23	m	-	2
DMP		3,27	s		
Glucose	C7 ; C5	3,45	m	-	2
Fructose	C7	3,55	m	-	2
Fructose	C11 ; C7	3,64	m	-	3
Saccharose	C13	3,65	s	-	2
Saccharose	C17 ; C19	3,79	m	-	4
Saccharose	C9	3,86	dd	8,46	1
Saccharose	C9	3,86	dd	3,25	1
Saccharose	C5	3,87	dd	8,50	1
Saccharose	C5	3,87	dd	3,66	1
Fructose	C5	3,97	t	1,70	1
Fructose	C11	4,01	dd	12,74 ; 1,2	1
Fructose	C4 , C3	4,08	d	3,79	1
Saccharose	C3	4,19	d	8,75	1
Glucose	C9	4,61	d	7,90	
Naringine		5,06			1
Glucose	C11	5,20	d	3,73	1
Saccharose	C7	5,38	d	3,89	1

[a] : s : singulet ; d : doublet ; dd : doublet de doublet ; t : triplet ; m : multiplet

a)

b)

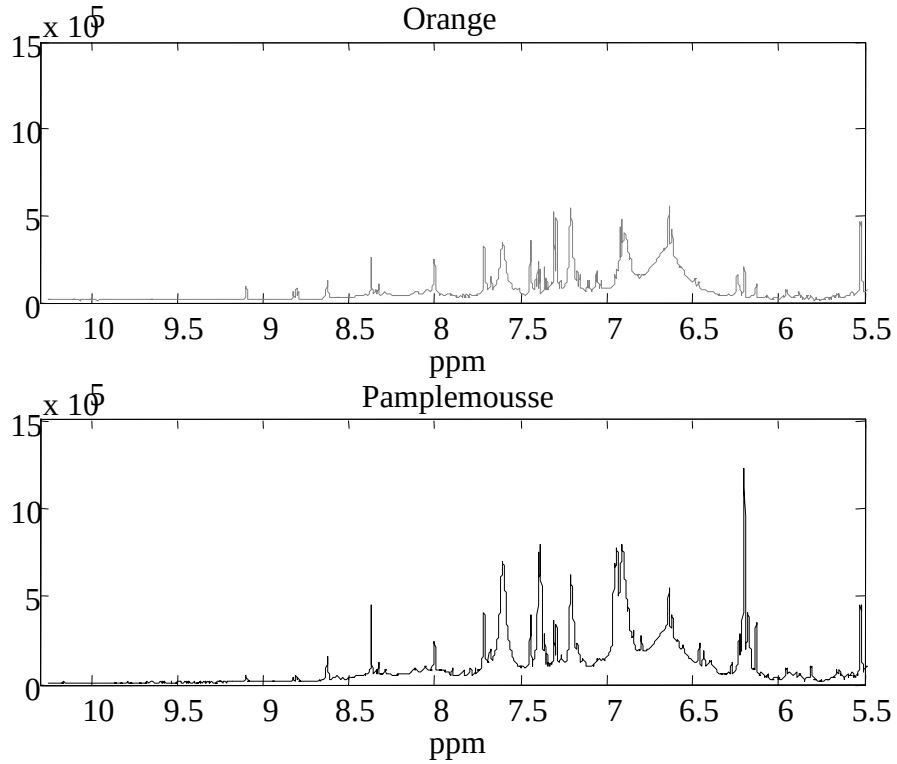


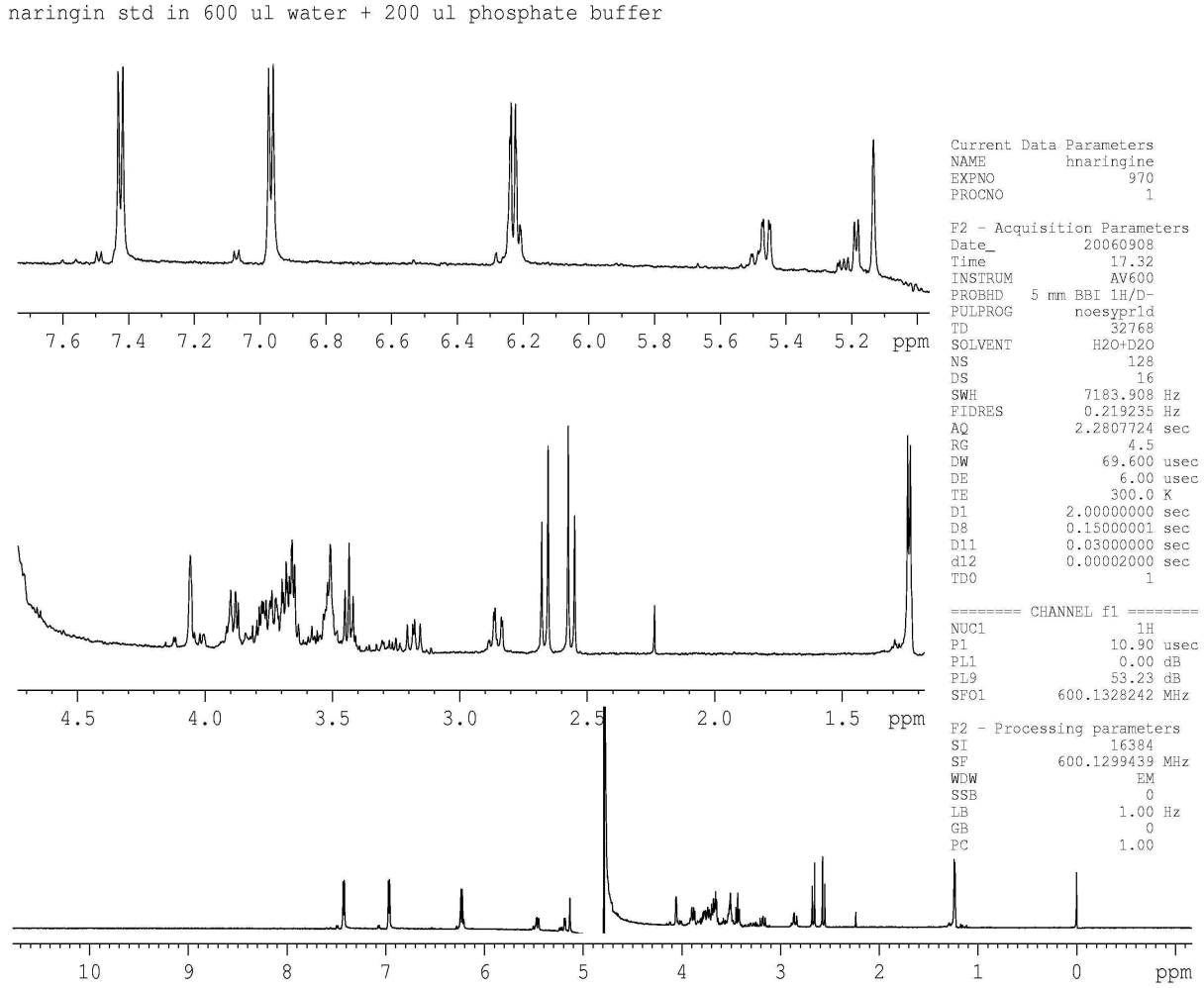
Figure 4 : Spectres moyens, zone spectrale de 5,5 à 10,5 ppm : a) Pamplemousse ; b) Orange

Tableau 3. Attribution des signaux de la zone spectrale de 5,5 à 10,5 ppm

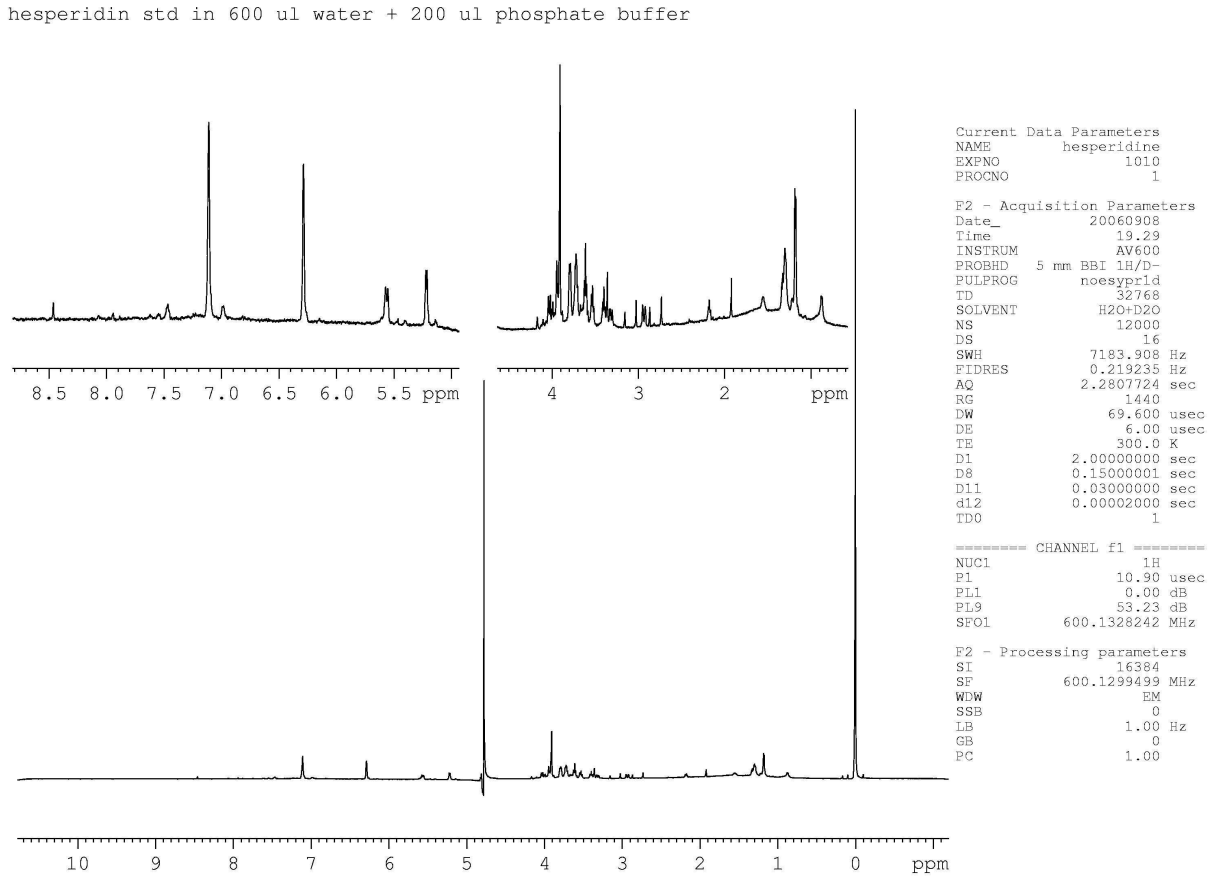
Composé	Groupement	δ (ppm)	Multiplicité [a]	J (Hz)	Nombre d' hydrogènes
Phlorine	C4 ; C6	6,14	s	-	2
Naringine		6,20	d		
Hespéridine		6,23	s		
Composé cinnamique		6,48			
Arginine		6,64			
GABA		6,91			
Tyrosine	C2 ; C6	6,93	d	8,60	2
Naringine		6,95	d		
Hespéridine	C34 ; C38	7,07	s		2
Arginine		7,22			
Tyrosine	C3 ; C5	7,31	d	8,65	
Phénylalanine	C4	7,37	m	-	1
Naringine		7,40	d		2
Phénylalanine	C3 ; C5	7,42	m	-	2
GABA		7,61			
Niacine	C4	8,08	dt	7,92 ; 1,88	1
Adénosine	C12	8,22	s	-	1
Acide formique	C1	8,34	s	-	1
Histidine	C2	8,63	d	1,13	1
Niacine	C6	8,81	dd	4,96 ; 1,54	1
Niacine	C2	9,10	dd	2,15 ; 0,72	1

[a] : s : singulet ; d : doublet ; dd : doublet de doublet ; dt : doublet de triplet ; m : multiplet

Annexe 8 : Spectre de la naringine

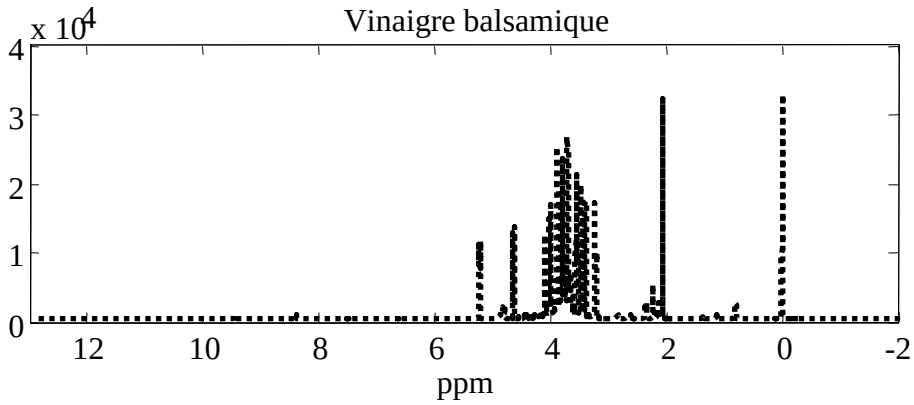


Annexe 9 : Spectre de l'hespéridine



Annexe 10 : Attribution des signaux du spectre du vinaigre balsamique

a)



b)

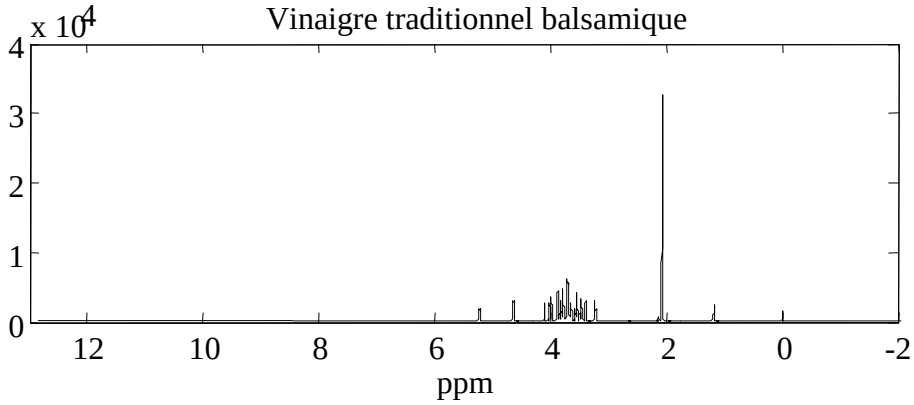
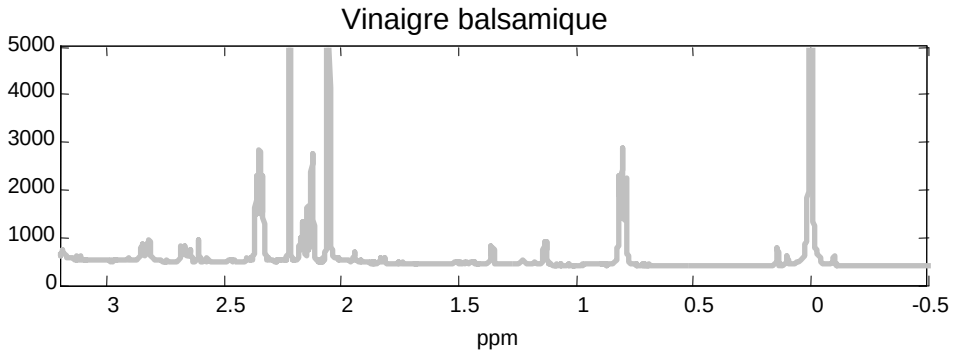


Figure 1 : Spectres moyens : a) Vinaigre balsamique ; b) Vinaigre traditionnel balsamique

a)



b)

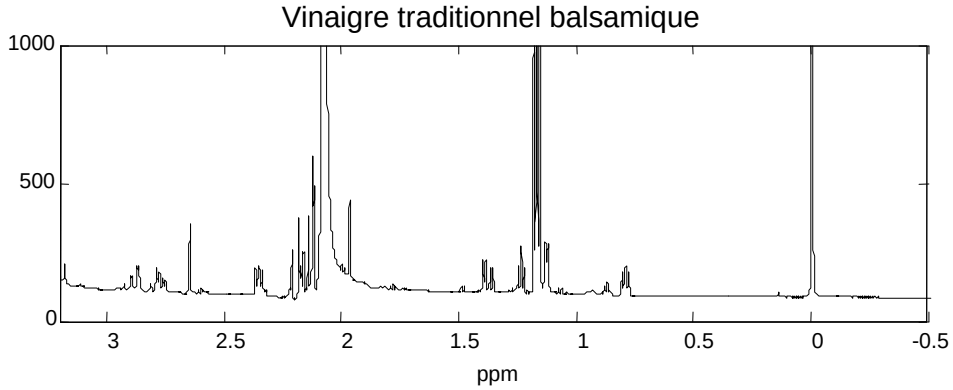


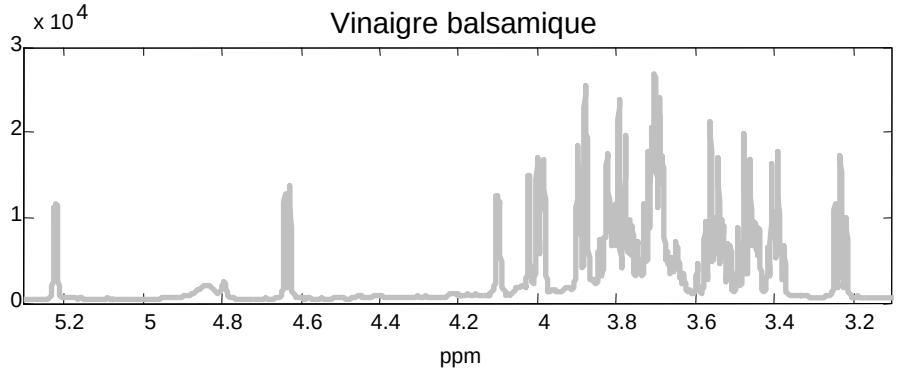
Figure 2 : Zone spectrale de –0,5 à 3,2 ppm : a) Vinaigre balsamique ; b) Vinaigre traditionnel balsamique

Tableau 1. Attribution des signaux de la zone spectrale de –0,5 à 3,2 ppm

Composé	Groupement	δ (ppm)	Multiplicité [a]	J (Hz)	Nombre d' hydrogènes
TSP		0,80			
Valine	$C\gamma H_3$	0,98	d	6,94	3
2,3-butanediol	C1 ; C4	1,14	d	5,67	6
Ethanol	C2	1,17	t	7,09	3
Acétate d'éthyle	C	1,24	t	7,21	3
Acétoïne	C4	1,36	d	7,15	3
Acide lactique	C3	1,42	d	6,84	3
Alanine	Cβ	1,48	d	7,20	3
Acetic acid	C2	2,07	s	-	3
Acide succinique	C2 ; C3	2,67	s	-	4
Acide malique	C3b	2,76	dd	16,30; 7,30	1
Acide malique	C3a	2,85	dd	16,31; 4,27	1
Acide citrique	C2b + C4b	2,87	d	15,83	2
Acide citrique	C2a + C4a	3,02	d	15,83	2

[a] : s : singulet ; d : doublet ; dd : doublet de doublet ; t : triplet

a)



b)

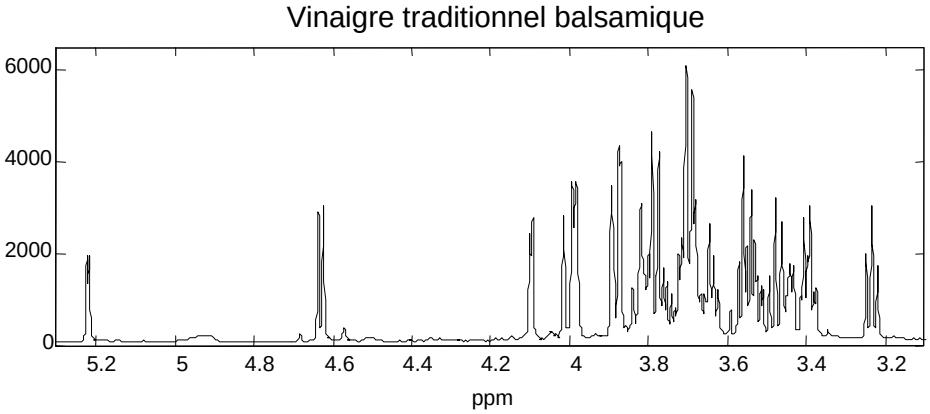


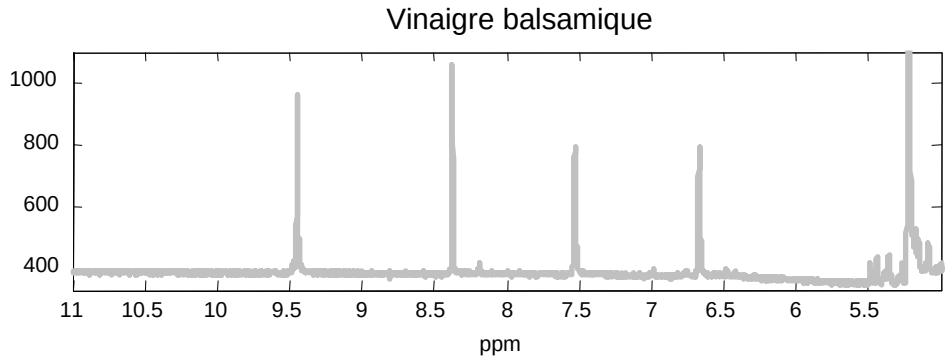
Figure 3 : Zone spectrale de 3,1 à 5,3 ppm : a) Vinaigre balsamique ; b) Vinaigre traditionnel balsamique

Tableau 2. Attribution des signaux de la zone spectrale de 3,1 à 5,3 ppm

Composé	Groupement	δ (ppm)	Multiplicité [a]	J (Hz)	Nombre d' hydrogènes
β-D-glucopyranose	C2H	3,26	dd	8,03; 9,29	1
β-D-fructopyranose	C5H	4,01	m	-	1
β-D-fructopyranose	C6Hax	4,05	dd	9,29; 8,05	1
β-D-fructofuranose	C3H + C4H	4,10	m	-	2
Acide tartrique	C2H +C3H	4,43	s	-	2
β-D-glucopyranose	C1H	4,63	d	7,97	1
α-D-glucopyranose	C1H	5,22	d	3,69	1

[a] : s : singulet ; d : doublet ; dd : doublet de doublet ; m : multiplet

a)



b)

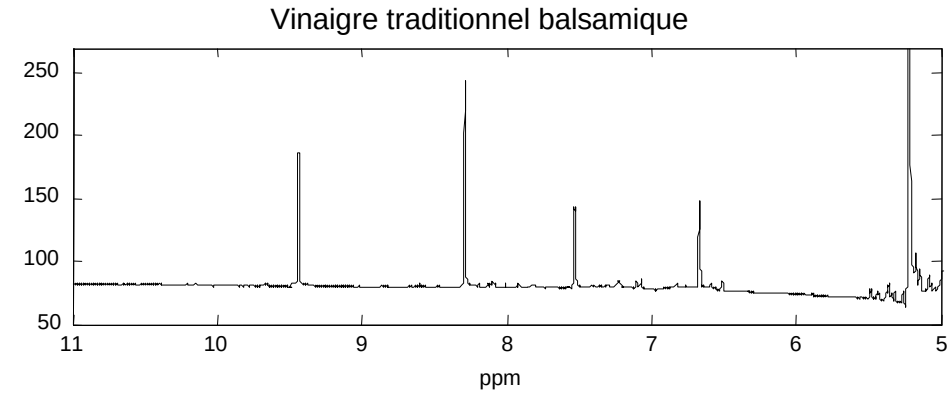


Figure 4 : Zone spectrale de 5 à 11 ppm : a) Vinaigre balsamique ; b) Vinaigre traditionnel balsamique

Tableau 3. Attribution des signaux de la zone spectrale de 5 à 11 ppm

Composé	Groupement	δ (ppm)	Multiplicité [a]	J (Hz)	Nombre d' hydrogènes
Hydroxyméthylfurfural (HMF)	C4	6,68	d	3,38	1
Tyrosine	Ar	6,85	d	8,5	1
Phénylalanine	Ar	7,35	m	-	1
HMF	C3	7,54	d	3,41	1
Acide formique	C1	8,31	s	-	1
Histidine	Ar	8,59	s	-	1
HMF	C1	9,46	s	-	1
AcMF	C1	9,49	s	-	1

[a] : s : singulet ; d : doublet ; m : multiplet

Annexe 11 : Choix du mode de préparation des échantillons de yaourts

1. Types de préparations

Afin d'obtenir les meilleurs résultats possibles, il est nécessaire de choisir les conditions de préparations de l'échantillon pour la mesure par spectroscopie RMN ^{1}H. L'absence de données bibliographiques sur les conditions de préparation des yaourts laisse envisager trois types de préparation : une préparation directe du produit, une préparation avec centrifugation de l'échantillon, et enfin une préparation par lyophilisation du yaourt et dilution dans de l'eau deutérée. Voici les trois modes opératoires adoptés pour cette étude :

- Préparation 1 :
  - Homogénéisation du produit
  - Prélèvement de 750 µL de produit
  - Ajout de 250 µL D_2O contenant 0,75 % de TSP (triméthylsilyl [2,2,3,3-$^{2}H_4$]propionate)
  - Mise en tube
- Préparation 2 :
  - Homogénéisation du produit
  - Centrifugation (15 min, 4,4 rpm)
  - Prélèvement de 750 µL de produit
  - Ajout de 250 µL D_2O contenant 0,75 % de TSP (triméthylsilyl [2,2,3,3-$^{2}H_4$]propionate)
  - Mise en tube
- Préparation 3 :
  - Homogénéisation du produit
  - Mesure du taux d'humidité du yaourt par la mesure de matière sèche présente : 88,5% d'humidité
  - Lyophilisation à –20 °C et 0,120 mbar dans un lyophilisateur Christ Alpha 1-4, pendant 72 h
  - Dilution de 300 mg de résidus déshydratés dans 3 mL de D_2O
  - Homogénéisation dans un bain à ultrasons
  - Prélèvement de 750 µL de produit
  - Ajout de 250 µL D_2O contenant 0,75 % de TSP (triméthylsilyl [2,2,3,3-$^{2}H_4$]propionate)
  - Mise en tube

Nous avons choisi d'analyser un yaourt nature au lait entier contenant entre autres ferments lactiques du Bifidobacterium.

2. Mesure des spectres

Les spectres ont été acquis avec un spectromètre Bruker Advance DPX-400. Pour chaque préparation, les paramètres d'acquisition ont été optimisés pour obtenir la meilleure résolution possible – réglages des shims, et « receiver gain » (RG) – 64 pour la préparation avec lyophilisation et 143,7 pour la préparation par centrifugation.

Pour obtenir le maximum d'information sur les composants du spectre nous avons utilisé une séquence d'impulsion permettant la présaturation du pic de l'eau, la séquence "noesypr1d". La puissance de présaturation (PL9) a été réglée à 60 dB pour affecter le moins possible les signaux alentours.

Chaque acquisition comportait 128 scans acquis en 64 K avec un temps d'acquisition (AQ) de 3,5 sec et un délai de relaxation (D1) de 15 sec tandis que la température a été fixée à 29 °C.

3. Résultats pour la préparation directe du produit

La préparation de ce type d'analyte n'est pas aisée. En effet, malgré la simplicité apparente de cette préparation, sa stabilité n'est pas de bonne qualité et ne permet pas de laisser reposer l'échantillon sur le porte-échantillon. La mesure doit donc être immédiate. Nous avons aussi rencontré des problèmes de stabilité durant la mesure du fait de la rotation du tube dans le spectromètre.

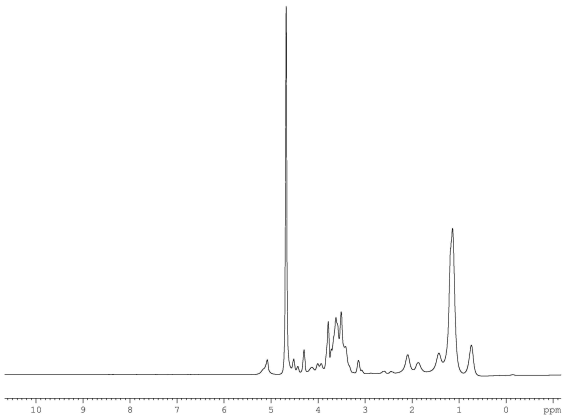


Figure 1 : Spectre de yaourt préparé selon la préparation 1

Par ailleurs, sur un échantillon qui a pu être mesuré nous remarquons que la résolution du spectre du yaourt préparé de manière directe n'est pas bonne (Figure 1). Ceci peut être dû à la

préparation qui malgré son apparence homogène peut être hétérogène à plus petite échelle, à sa viscosité ou encore à la présence de particules dans le tube.

4. Comparaison des résultats pour les spectres d'un yaourt préparé par lyophilisation et par centrifugation

Lors de la préparation de l'échantillon et lors de la mesure des expériences utilisant les préparations 2 et 3, nous n'avons pas rencontré de difficultés. Les analytes étaient stables. De plus, dans les deux cas, la résolution des spectres était bonne pour les deux types de préparation.

Cependant nous pouvons noter (Figure 2) que la présence de matière grasse dans la préparation par lyophilisation entraîne la présence de larges signaux entre 0,5 et 2,5 ppm ainsi qu'autour de 4,3 et 5,35 ppm. Ces larges signaux cachent d'autres signaux plus fins comme le doublet de l'acide lactique à 1,35 ppm. De plus, le retour à la ligne de base n'est pas aussi bon avec la préparation par lyophilisation qu'avec la préparation avec centrifugation.

Nous avons pu attribuer certains signaux grâce à une expérience en RMN du proton 2D TOCSY (présentée dans l'Annexe 12) et des tables de déplacements chimiques publiées sur le lait [1]. Cependant l'expérience TOCSY sur l'échantillon ayant subit une lyophilisation puis une dilution dans de l'eau lourde n'a pas réussi. Ceci peut être dû à l'instabilité de l'échantillon durant l'expérience TOCSY. En effet, bien que l'échantillon soit stable pour la durée d'une expérience 1D le temps important de rotation de l'échantillon au cours de l'expérience 2D (17 h 23 min) a engendré une décantation partielle de l'échantillon. Seuls les composés majeurs ont pu être attribués. Cependant l'expérience ayant réussi sur l'échantillon ayant subi une centrifugation d'autres composés ont été identifiés.

La différence majeure entre les deux spectres est donc la présence de signaux provenant de la matière grasse du lait (acides gras et glycérol) pour l'échantillon préparé par lyophilisation. De plus, la migration différentielle de certains composés entre le surnageant et le culot du yaourt serait à quantifier dans le cas de mesure quantitative.

Nous l'avons fait pour l'acide citrique. Grâce à la mesure enzymatique du L- et D- lactate [2,3] dans le yaourt, le surnageant et le culot nous avons pu quantifier cette migration différentielle. La quantité d'acide lactique dans notre échantillon était de 10,87 g/L d'acide lactique. Dans le surnageant cette quantité avait augmenté (11,27 g/L) et diminué dans le culot jusqu'à 8,87 g/L ce qui montre la tendance de l'acide lactique à migrer dans la phase liquide lors de la centrifugation.

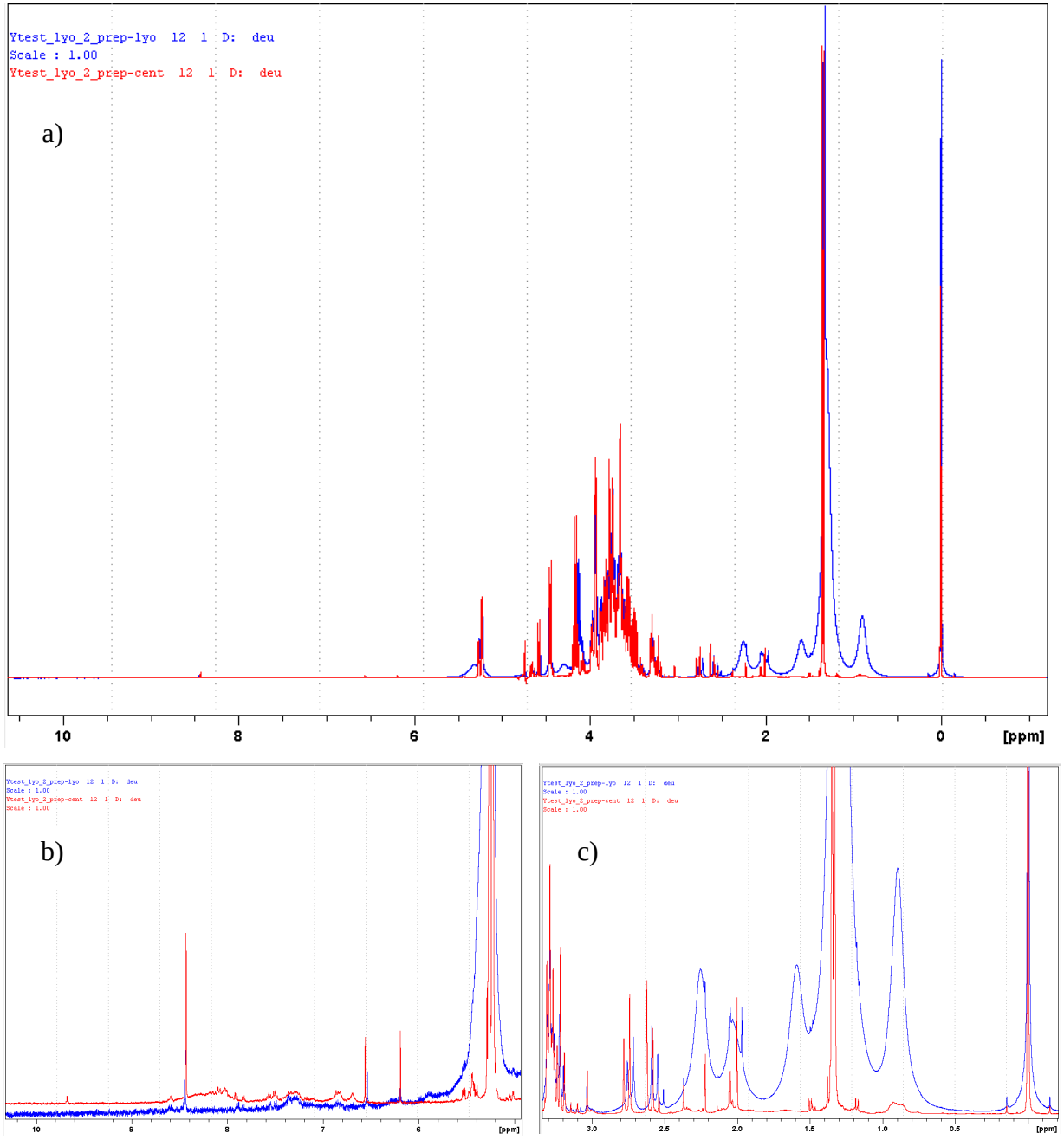


Figure 2 : Spectres des yaourts préparés soit par lyophilisation (bleu), soit par centrifugation (rouge) : a) Spectres entiers ; b) Zoom dans la 5 à 10 ppm ; c) Zoom dans la zone 0 à 3 ppm

5. Conclusions

Pour des raisons pratiques et de reproduction de la préparation d'un échantillon de yaourt, le passage direct du produit en RMN (préparation 1), ne nous semble pas acceptable. De plus, le temps de préparation mis à part, pour obtenir le maximum d'information sur les constituants du yaourt nous avons constaté que la préparation par centrifugation était la mieux adaptée. Ainsi, nous avons choisi d'utiliser cette technique dans nos différentes études.

6. Références

1. Hu, F., Furihata, K., Ito-Ishida, M., Kaminogawa, S. & Tanokura, M. Nondestructive observation of bovine milk by NMR spectroscopy: analysis of existing states of compounds and detection of new compounds. *J. Agric. Food Chem.* **52**, 4969-4974 (2004).
2. Mannheim, B. in *Methods of enzymatic bioanalysis and food analysis*, 74-77 (1995).
3. AFNOR. NF EN 12631: Jus de fruits et de légumes. - Dosage enzymatique des acides D- et L-lactiques (lactate). - Méthode spectrométrique par le NAD (indice de classement : V76142). (1999).

Annexe 12 : Attribution des signaux du spectre d'un extrait de yaourt nature

Les attributions des pics d'un extrait de yaourt nature sont basées sur des données publiées sur le lait [1] et l'analyse du spectre 2D COSY ci-dessous, et de tableaux de couplage de déplacements chimiques [2].

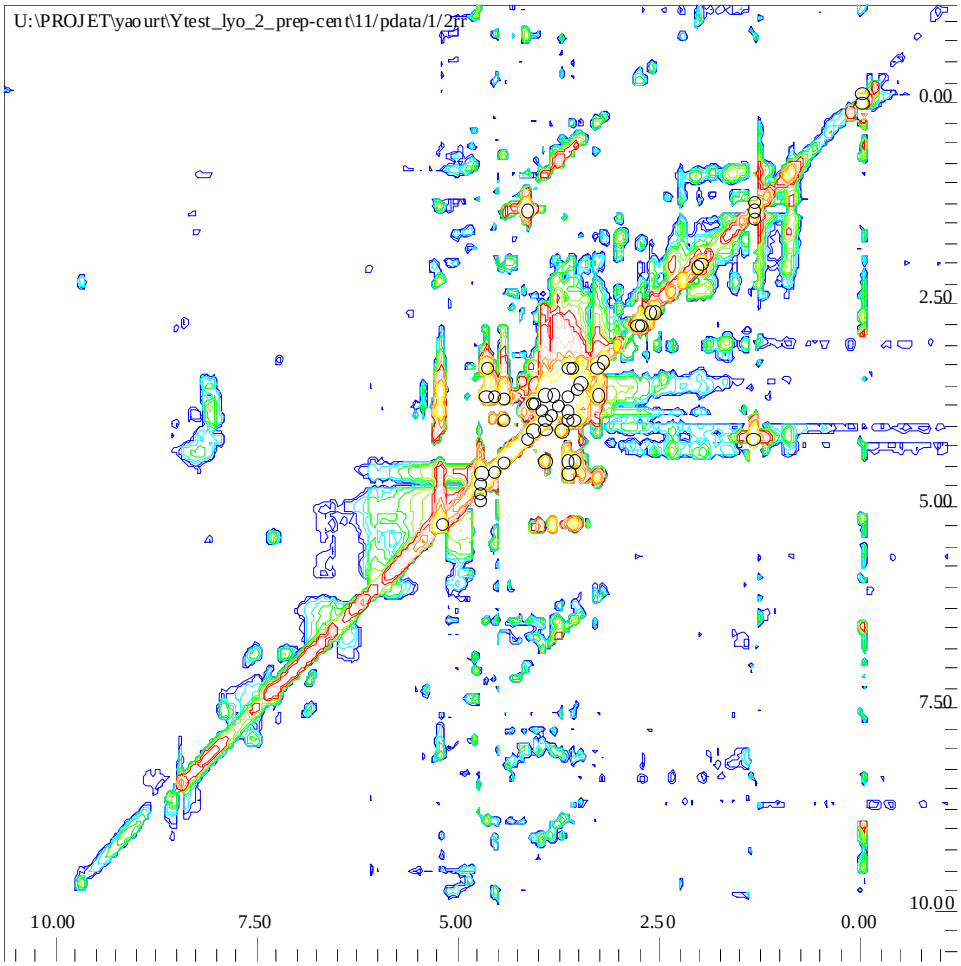


Figure 1 : Spectre COSY d'un extrait de yaourt nature

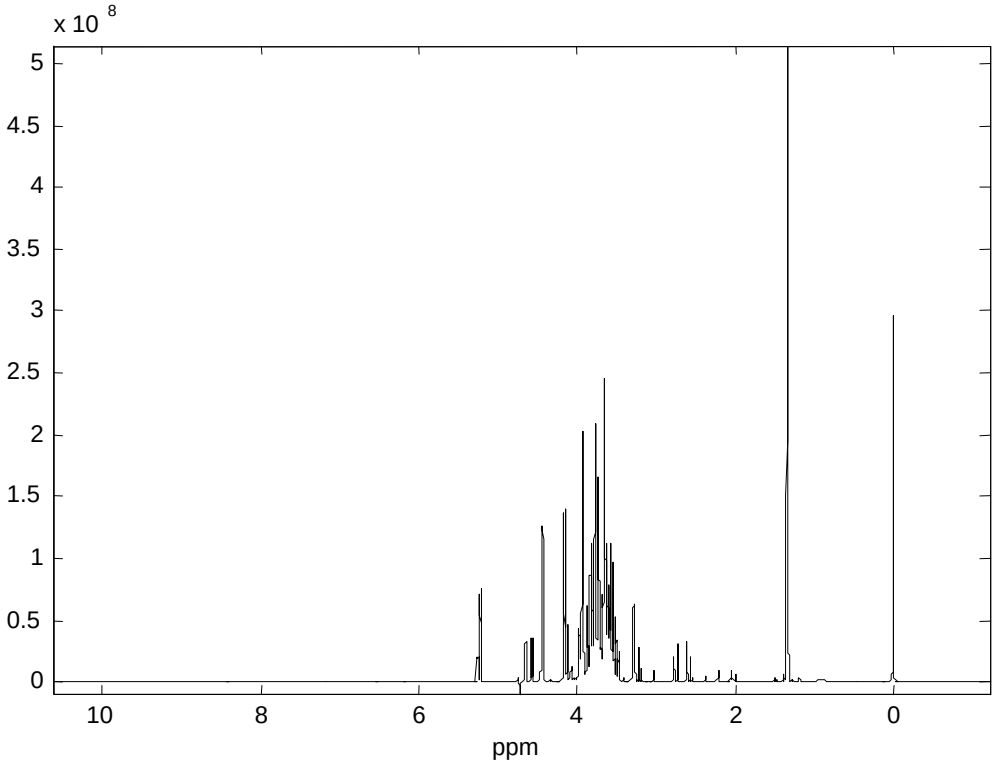


Figure 2 : Spectre 1D d'un extrait de yaourt nature (noesypr1d)

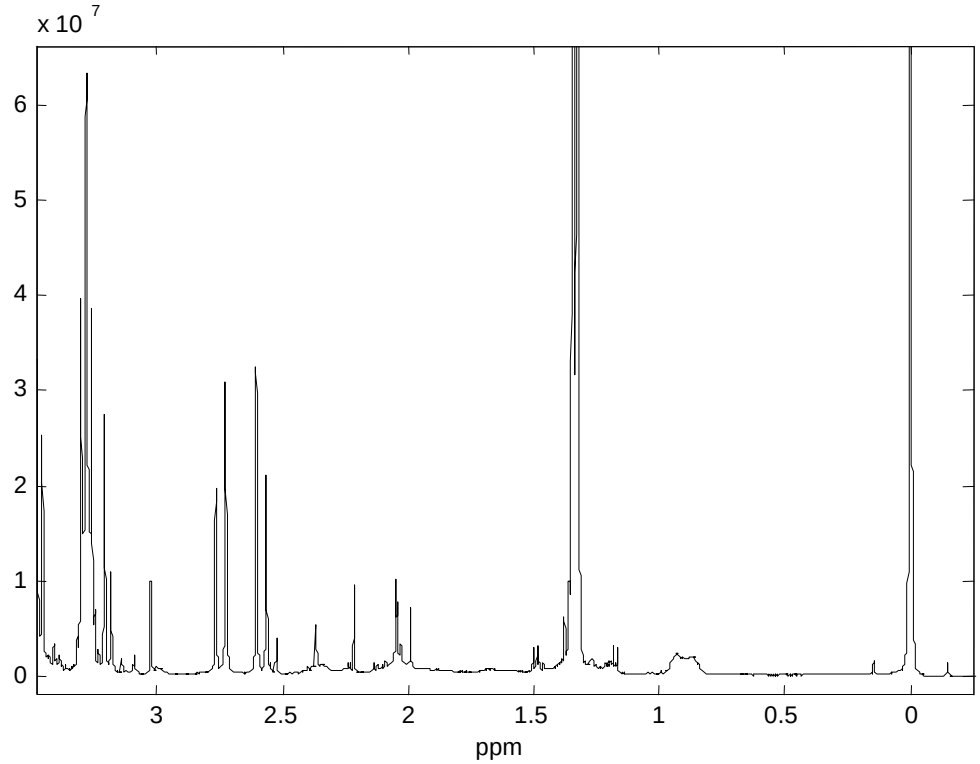


Figure 3 : Spectre 1D d'un extrait de yaourt nature : zone spectrale de –0,30 à 3,50 ppm

Tableau 1. Attribution des signaux de la zone spectrale de –0,5 à 3,50 ppm

Composé	Groupement	δ (ppm)	Multiplicité [a]	J (Hz)	Nombre d' hydrogènes
Acide propionique	C5	1,20	t	7,68	3
Acide lactique	C3	1,33	d	7	3
Lysine	C7	1,43	m	-	2
Alanine	C6	1,47	d	7,14	3
Lysine	C8	1,67	m	-	2
GABA	C5	1,90	m	-	2
Proline	C3	2,02	m	-	3
Proline	C4	2,09	m	-	1
Acétaldéhyde	C2	2,22	s	-	3
GABA	C6	2,34	t	7,39	2
Proline	C4	2,34	m	-	1
Acide propionique	C4	2,41	q	7,66	2
Acide citrique	C6b ; C3b	2,58	d	-	2
Acide citrique	C6a ; C3a	2,76	d	15,16	2
Lysine	C9	2,99	t	7,58	2
GABA	C4	3,01	d	7,61	2
Arginine	C8	3,17	t	6,91	2
Glucose		3,28			
Glucose		3,38			
Glucose		3,41			
Alanine	C4	3,42	q	7,20	1

[a] : s : singulet ; d : doublet ; q : quadruplet ; t : triplet

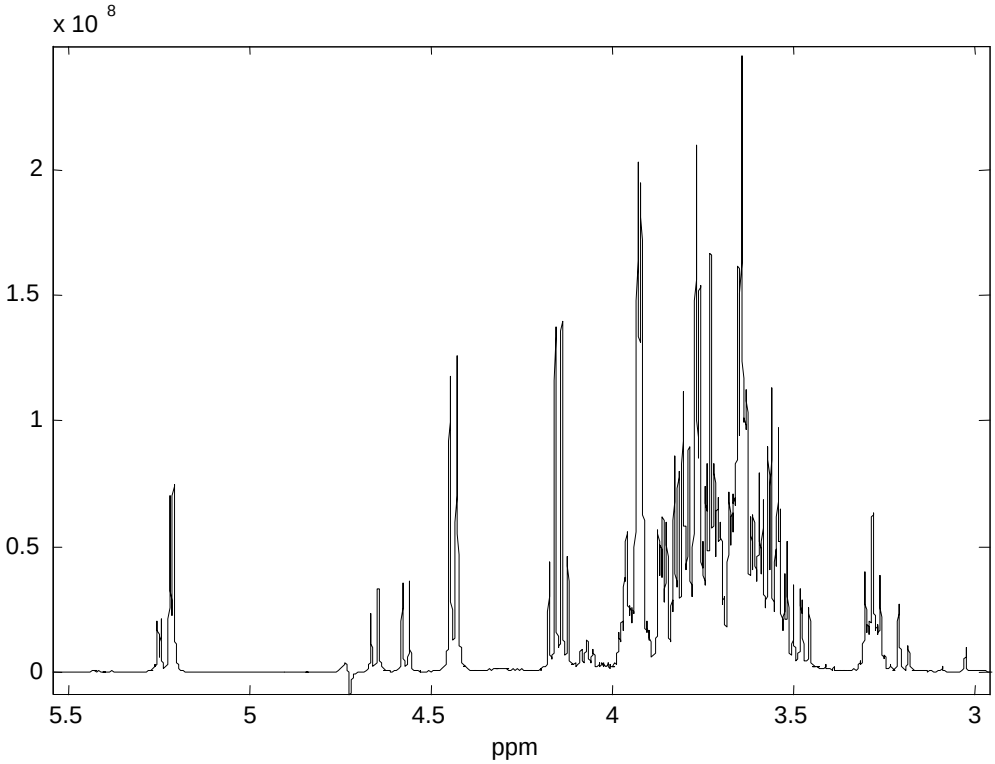


Figure 4 : Spectre 1D d'un extrait de yaourt nature : zone spectrale de 2,95 à 5,5 ppm

Tableau 2. Attribution des signaux de la zone spectrale de 2,95 à 5,5 ppm

Composé	Groupement	δ (ppm)	Multiplicité [a]	J (Hz)	Nombre d' hydrogènes
Glucose		3,51			
Glucose		3,55			
Glucose		3,59			
Glucose		3,6			
Glucose		3,62			
Glucose		3,65			
Glucose		3,66			
Galacto-pyranose		3,70			
Créatine	C4	3,75	s	-	2
Galactose libre alpha		3,76			
Alanine	C4	3,8	q	7,20	1
Glucose		3,92			
Glucose		3,94			
Galactose libre alpha	C5	3,97	d	2,66	1
Galactose libre alpha		4,08			
Proline	C5	4,11	dd	8,63 ; 6,42	1
Acide lactique	C2	4,15	q		1
Glucose		5,21			
Galactose libre alpha		5,25			

[a] : s : singulet ; d : doublet ; q : quadruplet ; dd : doublet de doublet

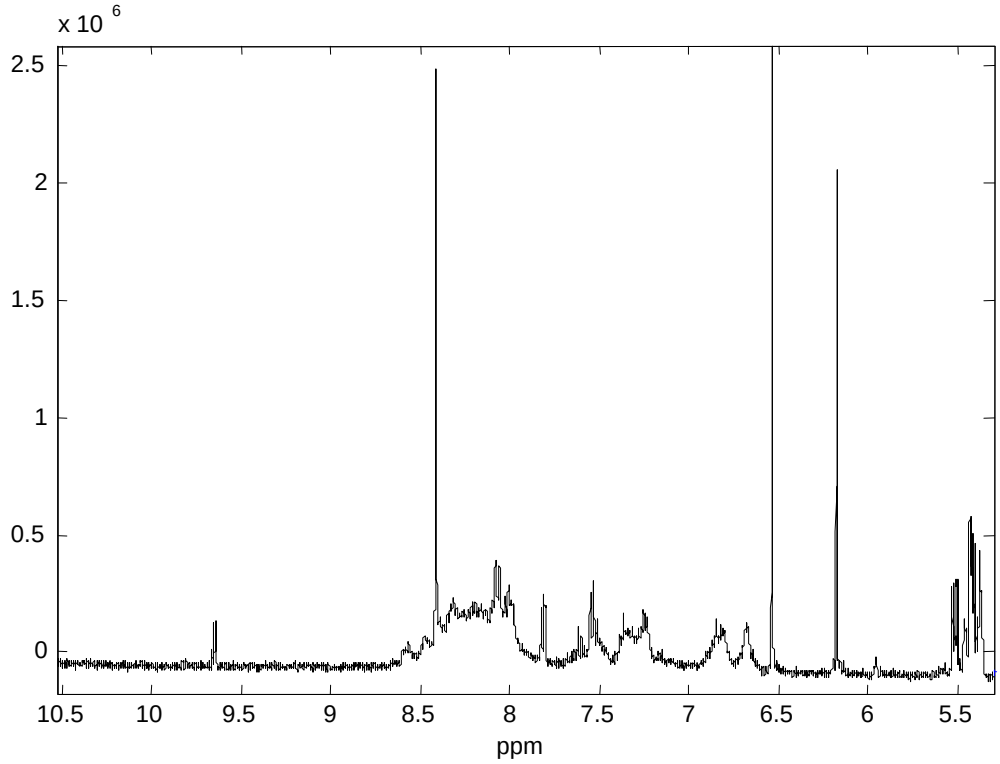


Figure 5 : Spectre 1D d'un extrait de yaourt nature : zone spectrale de 5,31 à 10,5 ppm

Tableau 3. Attribution des signaux de la zone spectrale de 5,31 à 10,5 ppm

Composé	Groupement	δ (ppm)	Multiplicité [a]	J (Hz)	Nombre d' hydrogènes
Acide fumarique	C4a ; C5a	6,55	s	-	2
Tyrosine / Tyrosol	C6 ; C2	6,82	d	8,60	2
Tyrosine / Tyrosol	C3 ; C5	7,12	d	8,65	2
Histidine	C5	7,28	d	0,58	1
Acide formique	C1	8,43	s	-	1
Histidine	C1	8,58	d	1,13	1
Acétaldéhyde	C1	9,65	s	-	1

[a] : s : singulet ; d : doublet

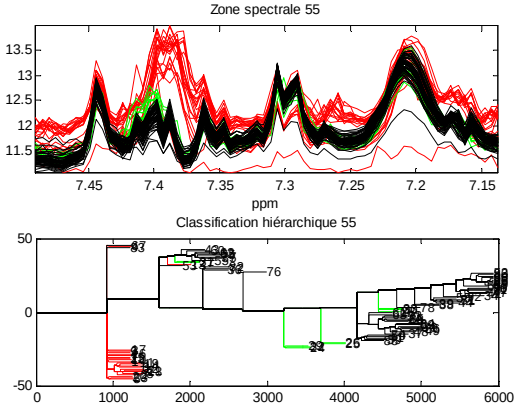
1. Hu, F., Furihata, K., Ito-Ishida, M., Kaminogawa, S. & Tanokura, M. Nondestructive observation of bovine milk by NMR spectroscopy: analysis of existing states of compounds and detection of new compounds. *J. Agric. Food Chem.* **52**, 4969-4974 (2004).
2. Fan, T. W. M. Metabolite profiling by one- and two-dimensional NMR analysis of complex mixtures. *Progress in Nuclear Magnetic Resonance Spectroscopy* **28**, 161-219 (1996).

Annexe 13 : Intervalles retenus en utilisant Interval-PLS_Cluster

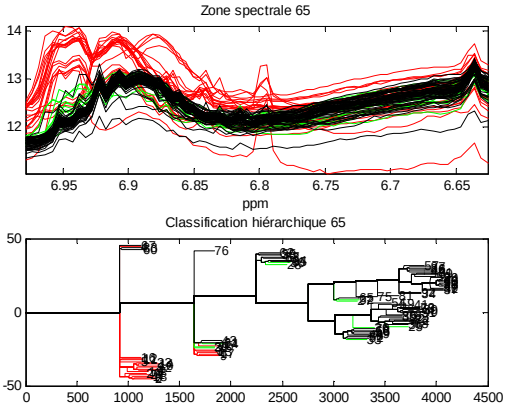
Légende :

Jus de pamplemousse
Mélange orange / pamplemousse
Jus d'orange

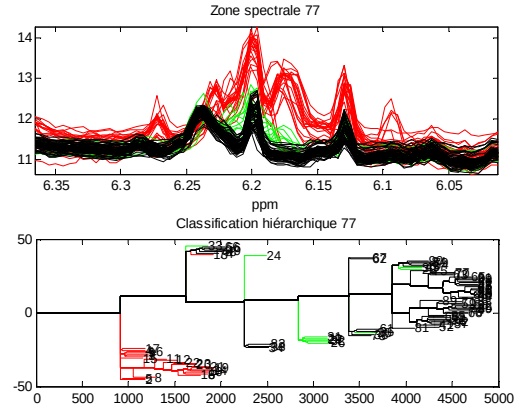
a)



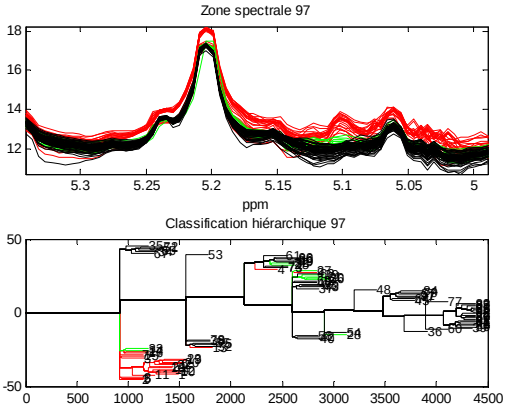
b)



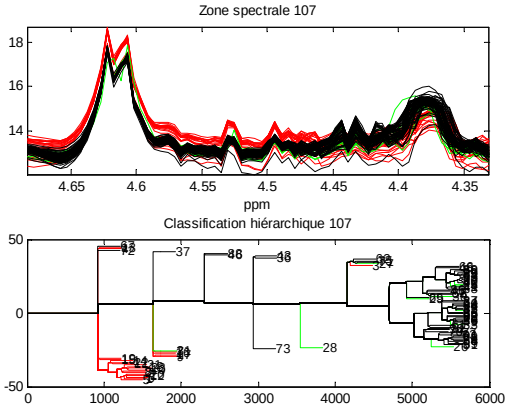
c)



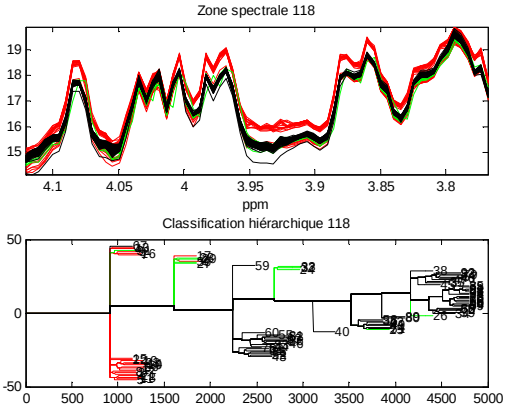
d)



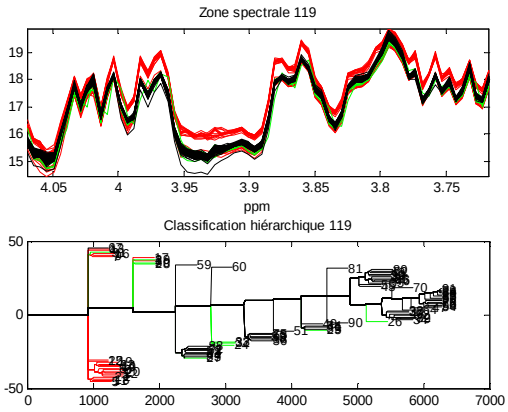
e)



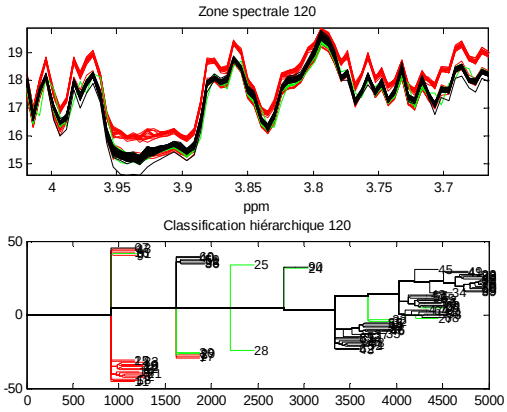
f)



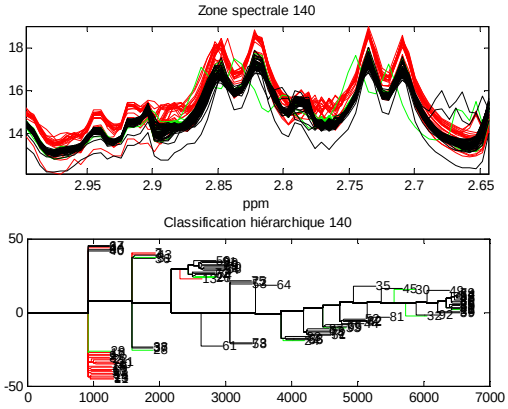
g)



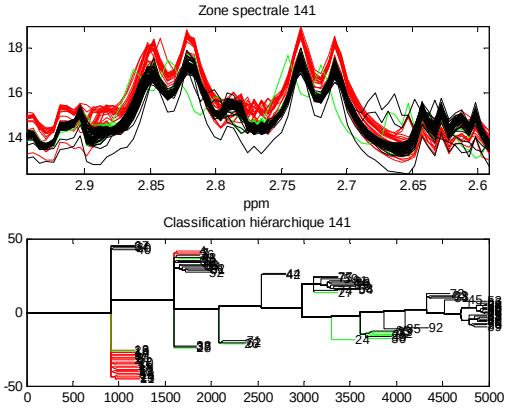
h)



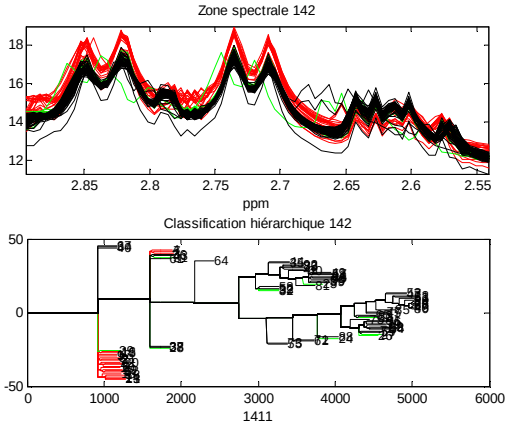
i)



j)



k)



Les intervalles f, g et h et les intervalles i, j et k étant contigus nous avons regroupé les zones. Ainsi 7 zones ont été retenues.